Nature Communications

https://doi.org/10.1038/s41467-025-66685-w

Article in Press

Molecular Motif Learning as a pretraining objective for molecular property prediction

Received: 7 January 2025

Accepted: 6 November 2025

Cite this article as: Liu, Z., Wang, C., Zheng, S. *et al.* Molecular Motif Learning as a pretraining objective for molecular property prediction. *Nat Commun* (2025). https://doi.org/10.1038/s41467-025-66685-w

Ziyang Liu, Chaokun Wang, Shuwen Zheng, Cheng Wu, Hao Feng, Li Xu, Yue Zheng, Liang Rong & Peng Li

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

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;

ARTICLE IN PRESS

Abstract

Molecular property prediction is crucial for drug discovery in biopharmaceuticals since it helps identify promising compounds, optimizing the efficacy of developing new therapies. Despite its importance, existing deep learning-based methods for this task are often incongruous with fundamental chemical properties. Here we show that an unsupervised pretraining approach, Molecular Motif Learning (MotiL), learns molecular representations that preserve both whole-molecule structure and motif-level information directly from native molecular graphs. MotiL produces representations that group small molecules sharing a common core structure (i.e., scaffold) and proteins with related three-dimensional structures and functions. We evaluated MotiL on at least 16 molecule benchmarks, and uncovered that it captures analogous graph representations not only for small molecules with the same scaffold but also for protein macromolecules with similar structures and overlapping chemical functions such as tRNA binding. These informative representations empower MotiL to surpass the accuracy of state-of-the-art contrastive or predictive methods in the prediction of molecular properties like blood-brain barrier permeability.

Keywords: Molecular property prediction, pretraining representation learning, motif, diffusion model

Molecular property prediction is a pivotal area of research in computational chemistry and biophysics, encompassing the property prediction of both small molecules [1, 2] and protein macromolecules [3, 4]. Precise prediction of molecular

ARTICLE IN PRESS

properties can significantly enhance the understanding of chemical and biological systems, leading to advancements in drug discovery [5, 6], materials science [7, 4], and molecular biology [8, 9].

Small molecules, i.e., molecules of low molecular weight, play crucial roles in various biological processes and therapeutic applications [10, 11, 12]. Their properties, such as toxicity [13, 14], blood-brain barrier permeability [15], and adverse drug reactions [16], are pivotal for drug efficacy and safety [17, 18]. Traditional wet-lab experimental methods to determine these properties are time-consuming and costly [19, 20]. It has led to the necessity of developing reliable computational models, particularly machine learning approaches [21, 22]. In addition, proteins play critical roles in many cellular functions, and their properties like folding and stability are essential for understanding biological mechanisms and designing therapeutic interventions [23, 24]. Computational methods, including molecular dynamics simulations and machine learning, have been employed to predict protein properties by bridging the gap between amino acid sequences and biological functions [25, 26]. Understanding these properties allows researchers to anticipate how proteins behave in different environments, which in turn facilitates the rational design of proteins with desired characteristics.

Existing deep learning-based molecular representation learning methods can be divided into two categories: contrastive learning and predictive learning. Contrastive learning methods [11, 12] typically produce two augmented versions of a molecular graph through techniques like subgraph removal or perturbation, and then use them to form positive and negative sample pairs. While these methods can distinguish chemically different molecules, they may inadvertently disrupt chemically meaningful substructures (e.g., functional groups or residues), thereby weakening the preservation of critical motif information. Predictive learning methods [27, 28, 29, 30], in contrast, aim to learn molecular representations by predicting masked subgraphs [27, 28], contexts [29], or attributes [30].

Beyond these graph-based methods, recent methods based on SMILES or sequences have also shown promise by leveraging larger unlabeled corpora (e.g., ChemBERTa [31], SMILES-BERT [32], ESM [33]). While these methods benefit from scalability, they may sacrifice finer chemical granularity, such as stereochemistry or bonding topology, which graph-based methods aim to preserve.

It is crucial for molecular representation learning to produce representations that accurately reflect the underlying chemical structures and functions. However, existing methods struggle to effectively capture the representations of motifs, such as functional groups in small molecules or amino acid residues in proteins, due to their limited ability to model localized chemical semantics or hierarchical substructures. For instance, prior contrastive learning methods [27, 11, 12] often rely on random subgraph sampling or global-level supervision, which may overlook pharmacologically or structurally meaningful motifs. Similarly, predictive molecular representation learning, which involves masking subgraphs, contexts, or attributes, also compromises the integrity of motif information in the original molecule, further hindering the preservation of critical chemical properties.

To address the above limitations, we propose a pretraining representation learning method for molecules, i.e., **Mo**lecular Mo**ti**f **L**earning (MotiL). MotiL includes three phases totally: diffusion priming, bi-scaled training, and task-specific fine-tuning. The three phases achieve the goals of global structure perception, local feature integration, and real-world label alignment, respectively. In the first phase, we design a **Diff**usion **Mo**lecular **M**odel (DiffMoM) to warm up the parameters of GNN by altering and correcting chemical bonds. In the second phase, MotiL learns molecular representations directly from the original molecule graph by the GNN encoder in an unsupervised manner. Both the first and second phases enable MotiL to effectively learn chemically meaningful graph representations and motif representations across various molecular contexts. At the graph level, MotiL produces similar graph representations for small molecules sharing the same scaffold (e.g., naphthalene and barbituric acid) or for proteins exhibiting similar three-dimensional structures and overlapping chemical functions (e.g., protein kinase activity and tRNA binding). At the motif level, MotiL

discerns functional groups with identical chemical bonds (e.g., the carbon-oxygen double bond in carboalkoxy, carbonyl, and carboxyl groups) and assigns them similar motif representations. Moreover, MotiL identifies structural isomers among a wide range of amino acids and assigns them similar representations, while the amino acids with disparate structures (e.g., aromatic rings vs. linear chains) are assigned dissimilar representations. In the final phase, the pretrained GNN parameters are further fine-tuned using backpropagated gradients calculated from the classification or regression loss for each molecule. This phase enables MotiL to accurately predict the micromolecular properties like toxicity and blood-brain barrier permeability, as well as the macromolecular properties like adenosine triphosphate (ATP) binding. After evaluation on public benchmarks, MotiL consistently outperforms the state-of-the-art methods, e.g., MotiL surpasses the runner-up by 10.1% on the QM7 dataset in the quantum mechanics domain. Additional results further highlight MotiL's effectiveness across domains such as physiology, biophysics, and physical chemistry.

In this work, we systematically formulate molecular motif representation learning as a core pretraining objective, rather than using motifs merely as auxiliary tools for molecular learning. Moreover, we successfully applied MotiL to both micro- and macromolecular property prediction, demonstrating its versatility across molecules of different scales.

Results

Overview of MotiL

We propose diffusion priming as the first phase of MotiL to perceive the global structure of a molecule and thus enhance the robustness of GNN to structural noise perturbations in the molecule graph $\mathcal{G}$ (Fig. 1a). DiffMoM consists of two processes: noise addition and noise reduction. In the noise addition process, Gaussian noise is introduced directly into the adjacency matrix of the molecular graph $\mathcal{G}$, producing a perturbed adjacency matrix $\mathbf{A}_t$ at each time step t. Then in the noise reduction process, DiffMoM aims to recover the original bonds in $\mathcal{G}$. It uses the perturbed molecule graph $\mathcal{G}_T$ at the T-th time step as inputs of the parameterized GNN encoder $G_{\boldsymbol{\theta}}$ (where $\boldsymbol{\theta}$ denotes the trainable parameter group) to produce the reconstructed graph $\hat{\mathcal{G}}$. During the optimization of DiffMoM, $\boldsymbol{\theta}$ is adjusted to minimize the reconstruction error between $\mathcal{G}$ and $\hat{\mathcal{G}}$, ensuring that $G_{\boldsymbol{\theta}}$ accurately captures the molecular structure of $\mathcal{G}$.

In the second phase of MotiL, we integrate the local features in a molecule via two GNN encoders where network pruning is applied to $G_{\boldsymbol{\theta}}$ and then results in a new GNN encoder $G_{\boldsymbol{\theta}'}$ with fewer weight connections (Fig. 1b). For a given molecule, $G_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\theta}'}$ produce two different graph representations. Both representations are designed to be as similar as possible for the same molecule, and as dissimilar as possible for different molecules. With this design, we preserve the complete information of graphs and motifs in model training, ensuring that $G_{\boldsymbol{\theta}}$ learns accurate molecular representations and functional group's representations.

Moreover, considering the representation's stability of the same motif (functional group or amino acid) across different molecule graphs, we design the bi-scaled learning mechanism in MotiL, which operates at both the graph and motif scales. First, MotiL aligns the two graph representations derived from the same molecule graph, ensuring the invariance of the overall molecular semantics under different encoders. Second, MotiL identifies key motifs like functional groups and ensures that the representations from the same motif are consistent across different molecules. In the example of Fig. 1b, for the input molecule aspirin, its graph representations $\mathbf{H}_1$ and $\mathbf{H}_2$ should be close to each other, while the graph representations $\mathbf{H}_2$ and $\mathbf{H}_3$ (benzamide's representation) from different molecules should be pushed apart. Meanwhile, the representations of the same functional group (e.g., both $\mathbf{S}_1$ and $\mathbf{S}_2$ for benzene ring) should be similar but the representations of different functional groups (e.g., $\mathbf{S}_1$ for benzene ring and $\mathbf{S}_5$ for carboxyl group) should be dissimilar. The bi-scaled learning mechanism ensures that MotiL not only captures the overall molecular semantics but also maintains the semantic stability of each functional group.

In the third phase of MotiL, $G_{\boldsymbol{\theta}}$ is fine-tuned to align with real-world labels (Fig. 1c). $G_{\boldsymbol{\theta}}$ first extracts the graph representation $\mathbf{H}_5$ from a previously unseen molecule paracetamol. Then $\mathbf{H}_5$ is fed into a multilayer perceptron (MLP) classifier to obtain a logit value. The logit value is compared

ARTICLE IN PRESS

with the true labels, e.g., hepatotoxicity, to calculate the classification loss. During this phase, the optimizer computes the backpropagated gradients based on the classification loss to further update θ. The task-specific fine-tuning phase ensures that G_θ is adjusted to improve its performance on the specific molecular property prediction task.

MotiL boosts the performance of molecular property prediction

The range of molecular properties is vast, covering scales from large-scale impacts on human physiology to minute electronic characteristics [7, 4, 8]. These properties include effects such as drug metabolic stability, target selectivity, toxicity, and the inhibition potential against HIV. To thoroughly evaluate MotiL's effectiveness, we test its performance across diverse small micromolecule datasets categorized into four main areas: physiology, biophysics, physical chemistry, and quantum mechanics. To illustrate MotiL's capability in dealing with complex biological entities, we also include experiments involving protein macromolecules. The details of experimental datasets can be found in Supplementary Section B.

Tables 1 and 2 show the experimental performance of state-of-the-art methods and MotiL on classification and regression tasks related to small molecular property prediction (see Supplementary Tables 3 and 4 for complete results, including exact p-values). Similar to baselines, we use the area under the receiver operating characteristic curve (ROC-AUC) as the metric for classification tasks. For regression tasks, we use root mean square error (RMSE) for ESOL, FreeSlov, and Lipophilicity (abbreviated as Lip.), and mean absolute error (MAE) for QM7, QM8, and QM9. The calculation method for evaluation metrics is provided in Supplementary Section C. From Tables 1 and 2, we can see that MotiL outperforms all baselines on each dataset, achieving an average relative improvement of 3.0% over the runner-up. For instance, MotiL surpasses the runner-up method KANO with a relative improvement of 1.6% in ROC-AUC on MUV and 10.1% in MAE on QM7. Although KANO has demonstrated high performance in Tables 1 and 2, it fails to accurately predict the blood-brain barrier permeability of certain chemotherapeutic drugs or antibiotics, as well as the hepatotoxicity of some molecules (Supplementary Fig. 9a). MotiL corrects KANO's erroneous predictions in specific cases and surpasses KANO in overall prediction accuracy (Supplementary Fig. 9b). The augmentation-based methods like MolCLR, CMPNN, and KANO outperform supervised and predictive self-supervised baselines, but they often overlook the negative impact of graph augmentation on motif integrity. In contrast, MotiL avoids aggressive augmentation, integrates a generative diffusion model to enhance robustness, and achieves consistently superior performance across benchmarks. These results highlight the effectiveness of motif-preserving representation learning. Although augmentation-based methods like $\text{MolCLR}_{\text{CMPNN}}$ and KANO outperform supervised methods like GEM and predictive self-supervised methods like GROVER, they consistently overlook the negative impact of augmentation, thus falling behind MotiL. Our experiments on absorption, distribution, metabolism, and excretion (ADME) prediction tasks (Supplementary Table 6) show that MotiL consistently delivers robust and competitive performance, achieving strong predictive accuracy and lower variance compared to baseline models. These results highlight the potential of MotiL for pharmacologically relevant applications. Our sensitivity analysis (Supplementary Fig. 1) reveals that a retaining probability of $p = 0.7$ in network pruning achieves the best trade-off between representation divergence and convergence stability, leading to consistently strong performance across datasets. Our ablation study (Supplementary Fig. 2) shows that each component of MotiL (DiffMoM, bi-scaled training, and network pruning) contributes to performance gains, with DiffMoM playing the most critical role in enhancing robustness and generalization across diverse molecular tasks.

Next, we evaluate the F_{max} performance of different methods on two protein datasets: gene ontology (GO) term prediction and enzyme commission (EC) number prediction. The results are shown in Table 3 (see Supplementary Table 5 for complete results, including exact p-values). As in previous work, the GO term prediction task is divided into three sub-tasks: biological process (BP), molecular function (MF), and cellular component (CC) [46, 24]. We can observe that

ARTICLE IN PRESS

Table 1 Performance comparison on small molecules in the fields of physiology and biophysics. We use the metric of receiver operating characteristic area under the curve (ROC-AUC, %) for classification tasks. All baseline results are directly taken from Ref. [12], which follows the same evaluation protocol as this study: three random-seeded scaffold splits for small molecules. Results are presented as mean ± standard deviation over three independent runs.

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 [34]	$71.8_{\pm 0.9}$	$70.9_{\pm 0.3}$	$65.0_{\pm 6.1}$	$53.6_{\pm 0.3}$	$62.5_{\pm 2.8}$	$71.6_{\pm 2.0}$	$71.6_{\pm 4.0}$	$74.0_{\pm 3.0}$
GIN [35]	$65.8_{\pm 4.5}$	$74.0_{\pm 0.8}$	$66.7_{\pm 1.5}$	$57.3_{\pm 1.6}$	$58.0_{\pm 4.4}$	$70.1_{\pm 5.4}$	$71.8_{\pm 2.5}$	$75.3_{\pm 1.9}$
MPNN [36]	$91.3_{\pm 4.1}$	$80.8_{\pm 2.4}$	$69.1_{\pm 3.0}$	$59.5_{\pm 3.0}$	$87.9_{\pm 5.4}$	$81.5_{\pm 1.0}$	$75.7_{\pm 1.3}$	$77.0_{\pm 1.4}$
DMPNN [37]	$91.9_{\pm 3.0}$	$75.9_{\pm 0.7}$	$63.7_{\pm 0.2}$	$57.0_{\pm 0.7}$	$90.6_{\pm 0.6}$	$85.2_{\pm 0.6}$	$78.6_{\pm 1.4}$	$77.1_{\pm 0.5}$
CMPNN [38]	$95.7_{\pm 2.8}$	$80.4_{\pm 0.3}$	$70.4_{\pm 0.8}$	$63.4_{\pm 2.0}$	$87.6_{\pm 0.8}$	$85.6_{\pm 3.5}$	$76.4_{\pm 3.9}$	$82.5_{\pm 2.3}$
N-GRAM [30]	$91.2_{\pm 0.3}$	$76.9_{\pm 2.7}$	-	$63.2_{\pm 0.5}$	$87.5_{\pm 2.7}$	$79.1_{\pm 1.3}$	$76.9_{\pm 0.7}$	$78.7_{\pm 0.4}$
Hu et. al [29]	$70.8_{\pm 1.5}$	$78.7_{\pm 0.4}$	$65.7_{\pm 0.6}$	$62.7_{\pm 0.8}$	$72.6_{\pm 1.5}$	$84.5_{\pm 0.7}$	$81.3_{\pm 2.1}$	$79.9_{\pm 0.7}$
MGSSL [28]	$70.5_{\pm 1.1}$	$76.4_{\pm 0.4}$	$64.1_{\pm 0.7}$	$61.8_{\pm 0.8}$	$80.7_{\pm 2.1}$	$79.7_{\pm 0.8}$	$78.7_{\pm 1.5}$	$79.5_{\pm 1.1}$
GEM [10]	$72.4_{\pm 0.4}$	$78.1_{\pm 0.1}$	$69.2_{\pm 0.4}$	$63.2_{\pm 0.4}$	$90.1_{\pm 1.3}$	$85.6_{\pm 1.1}$	$81.7_{\pm 0.5}$	$80.6_{\pm 0.9}$
GROVER [27]	$86.8_{\pm 2.2}$	$80.3_{\pm 2.0}$	$56.8_{\pm 3.4}$	$61.2_{\pm 2.5}$	$70.3_{\pm 13.7}$	$82.4_{\pm 3.6}$	$67.3_{\pm 1.8}$	$68.2_{\pm 1.1}$
GraphMVP [39]	$72.4_{\pm 1.6}$	$75.9_{\pm 0.5}$	$63.1_{\pm 0.4}$	$63.9_{\pm 1.2}$	$79.1_{\pm 2.8}$	$81.2_{\pm 0.9}$	$77.7_{\pm 0.6}$	$77.0_{\pm 1.2}$
MolCLR [11]	$73.3_{\pm 1.0}$	$74.1_{\pm 5.3}$	$65.9_{\pm 2.1}$	$61.2_{\pm 3.6}$	$89.8_{\pm 2.7}$	$82.8_{\pm 0.7}$	$78.9_{\pm 2.3}$	$77.4_{\pm 0.6}$
$\text{MolCLR}_{\text{CMPNN}}$ [11]	$72.4_{\pm 0.7}$	$78.4_{\pm 2.6}$	$69.1_{\pm 1.2}$	$59.7_{\pm 3.4}$	$88.0_{\pm 4.0}$	$85.0_{\pm 2.4}$	$74.5_{\pm 2.1}$	$77.8_{\pm 5.5}$
KANO [12]	$\textit{96.0}_{\pm 1.6}$	$\textit{83.7}_{\pm 1.3}$	$\textit{73.2}_{\pm 1.6}$	$\textit{65.2}_{\pm 0.8}$	$\textit{94.4}_{\pm 0.3}$	$\textit{93.1}_{\pm 2.1}$	$\textit{83.7}_{\pm 2.3}$	$\textit{85.1}_{\pm 2.2}$
MotiL	$\mathbf{96.9}_{\pm 0.3}$	$\mathbf{84.0}_{\pm 0.2}$	$\mathbf{73.5}_{\pm 0.6}$	$\mathbf{65.6}_{\pm 0.9}$	$\mathbf{95.2}_{\pm 0.8}$	$\mathbf{93.9}_{\pm 1.0}$	$\mathbf{85.0}_{\pm 1.0}$	$\mathbf{85.6}_{\pm 0.2}$

The best result for each dataset is highlighted in bold, and the runner-up result is italicized.

Table 2 Performance comparison on small molecules in the fields of physical chemistry and quantum mechanics. We use the metric of root mean square error (RMSE) or mean absolute error (MAE), where lower values indicate better performance, for regression tasks.

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 [34]	$1.431_{\pm 0.050}$	$2.870_{\pm 0.135}$	$0.712_{\pm 0.049}$	$122.9_{\pm 2.2}$	$0.0366_{\pm 0.000}$	$0.00835_{\pm 0.00001}$
GIN [35]	$1.452_{\pm 0.020}$	$2.765_{\pm 0.180}$	$0.850_{\pm 0.071}$	$124.8_{\pm 0.7}$	$0.0371_{\pm 0.001}$	$0.00824_{\pm 0.00004}$
MPNN [36]	$1.167_{\pm 0.430}$	$1.621_{\pm 0.952}$	$0.672_{\pm 0.051}$	$111.4_{\pm 0.9}$	$0.0148_{\pm 0.001}$	$0.00522_{\pm 0.00003}$
DMPNN [37]	$1.050_{\pm 0.008}$	$1.673_{\pm 0.082}$	$0.683_{\pm 0.016}$	$103.5_{\pm 8.6}$	$0.0156_{\pm 0.001}$	$0.00514_{\pm 0.00001}$
CMPNN [38]	$0.798_{\pm 0.112}$	$1.570_{\pm 0.442}$	$0.614_{\pm 0.029}$	$75.1_{\pm 3.1}$	$0.0153_{\pm 0.002}$	$0.00405_{\pm 0.00002}$
N-GRAM [30]	$1.100_{\pm 0.030}$	$2.510_{\pm 0.191}$	$0.880_{\pm 0.121}$	$125.6_{\pm 1.5}$	$0.0320_{\pm 0.003}$	$0.00964_{\pm 0.00031}$
Hu et.al [29]	$1.100_{\pm 0.006}$	$2.764_{\pm 0.002}$	$0.739_{\pm 0.003}$	$113.2_{\pm 0.6}$	$0.0215_{\pm 0.001}$	$0.00922_{\pm 0.00004}$
GEM [10]	$0.813_{\pm 0.028}$	$1.748_{\pm 0.114}$	$0.674_{\pm 0.022}$	$60.0_{\pm 2.7}$	$0.0163_{\pm 0.001}$	$0.00562_{\pm 0.00007}$
GROVER [27]	$1.423_{\pm 0.288}$	$2.947_{\pm 0.615}$	$0.823_{\pm 0.010}$	$91.3_{\pm 1.9}$	$0.0182_{\pm 0.001}$	$0.00719_{\pm 0.00208}$
MolCLR [11]	$1.113_{\pm 0.023}$	$2.301_{\pm 0.247}$	$0.789_{\pm 0.009}$	$90.9_{\pm 1.7}$	$0.0185_{\pm 0.013}$	$0.00480_{\pm 0.00003}$
$\text{MolCLR}_{\text{CMPNN}}$ [11]	$0.911_{\pm 0.082}$	$2.021_{\pm 0.133}$	$0.875_{\pm 0.003}$	$89.8_{\pm 6.3}$	$0.0179_{\pm 0.001}$	$0.00475_{\pm 0.00001}$
KANO [12]	$\textit{0.670}_{\pm 0.019}$	$\textit{1.142}_{\pm 0.258}$	$\textit{0.566}_{\pm 0.007}$	$\textit{56.4}_{\pm 2.8}$	$\textit{0.0123}_{\pm 0.000}$	$\textit{0.00320}_{\pm 0.00001}$
MotiL	$\mathbf{0.644}_{\pm 0.014}$	$\mathbf{1.097}_{\pm 0.001}$	$\mathbf{0.548}_{\pm 0.000}$	$\mathbf{50.7}_{\pm 1.0}$	$\mathbf{0.0117}_{\pm 0.000}$	$\mathbf{0.00299}_{\pm 0.00001}$

MotiL consistently outperforms baselines, including state-of-the-art graph-based methods like GearNet and CDConv, highlighting its potential for broader applications in computational biology. On GO-BP, GO-MF, and GO-CC, MotiL achieves the highest F_{max} performance and outperforms the runner-up CDConv by 2.2%, 2.2%, and 6.9%, respectively. Also, on EC, MotiL achieves the most significant improvement, reaching an F_{max} of 87.9%, surpassing the runner-up CDConv by 1.8% (with a p-value of 3.7e-2 in a two-sided t-test). These consistent gains across diverse tasks suggest that the generative diffusion model DiffMoM allows the GNN encoder in MotiL to capture intricate structural information from the original protein graphs, which is crucial for accurate protein property prediction.

Table 3 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 [46], 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.

Dataset	**GO-BP**	**GO-MF**	**GO-CC**	**EC**
Metric	F_{max}(%) ↑			
# proteins	36,635			19,198
CNN [25]	$26.3_{\pm 0.5}$	$23.0_{\pm 0.6}$	$33.7_{\pm 1.7}$	$35.9_{\pm 2.2}$
ResNet [26]	$26.4_{\pm 0.6}$	$22.4_{\pm 0.7}$	$33.3_{\pm 2.0}$	$19.4_{\pm 0.7}$
LSTM [26]	$24.2_{\pm 0.3}$	$23.6_{\pm 0.4}$	$31.3_{\pm 0.8}$	$21.2_{\pm 0.5}$
Transformer [26]	$24.9_{\pm 1.4}$	$22.9_{\pm 0.9}$	$35.2_{\pm 4.0}$	$20.0_{\pm 0.3}$
GCN [34]	$27.2_{\pm 0.3}$	$23.7_{\pm 0.1}$	$38.5_{\pm 0.5}$	$31.6_{\pm 1.0}$
GAT [40]	$26.0_{\pm 0.3}$	$21.8_{\pm 0.5}$	$37.2_{\pm 0.6}$	$11.0_{\pm 2.9}$
3D CNN [41]	$25.9_{\pm 0.2}$	$21.0_{\pm 1.1}$	$32.5_{\pm 0.8}$	$23.4_{\pm 2.7}$
GraphQA [42]	$25.9_{\pm 0.2}$	$23.8_{\pm 0.3}$	$34.1_{\pm 0.1}$	$16.3_{\pm 1.0}$
GVP [43]	$26.5_{\pm 0.3}$	$20.7_{\pm 0.1}$	$33.4_{\pm 2.2}$	$16.9_{\pm 1.2}$
$\text{IEConv}_{\text{residue level}}$ [44]	$29.5_{\pm 1.0}$	$38.6_{\pm 0.8}$	$36.7_{\pm 0.8}$	$45.5_{\pm 1.4}$
GearNet [45]	$31.3_{\pm 0.6}$	$51.5_{\pm 0.6}$	$34.6_{\pm 0.6}$	$64.0_{\pm 1.6}$
$\text{GearNet}_{\text{IEConv}}$ [45]	$34.6_{\pm 0.5}$	$52.4_{\pm 0.6}$	$37.4_{\pm 0.5}$	$64.6_{\pm 0.7}$
$\text{GearNet}_{\text{Edge}}$ [45]	$35.5_{\pm 0.5}$	$53.2_{\pm 0.3}$	$37.5_{\pm 0.4}$	$66.2_{\pm 0.8}$
$\text{GearNet}_{\text{Edge-IEConv}}$ [45]	$36.2_{\pm 0.5}$	$55.0_{\pm 0.8}$	$41.0_{\pm 0.6}$	$69.3_{\pm 0.5}$
CDConv [24]	*42.2*$_{\pm 0.7}$	*61.8*$_{\pm 0.5}$	*42.0*$_{\pm 0.3}$	*86.1*$_{\pm 0.4}$
MotiL	**44.4**$_{\pm 0.2}$	**64.0**$_{\pm 0.3}$	**48.9**$_{\pm 0.3}$	**87.9**$_{\pm 0.3}$

MotiL learns micromolecular representations consistent with molecular structures

We use *t*-distributed stochastic neighbor embedding (t-SNE) [47] to project the micromolecular representations learned by MotiL into a 2D space (Fig. 2a), where each point represents a molecule, and the color indicates the molecule's scaffold. A scaffold represents the basic structure of a molecule, and compounds sharing the same scaffold typically exhibit similar chemical and biological properties. For example, we select four chemically diverse scaffolds from ESOL. The selected scaffolds, such as Naphthalene, are frequent and structurally distinct within ESOL, making them suitable for evaluating the model's ability to differentiate solubility-relevant chemotypes. A lower Davies-Bouldin (DB) index [48] indicates tighter grouping within clusters and greater separation between them. We compare MotiL with three methods: No pretraining (i.e., not using diffusion priming and bi-scaled training), Molecular contrastive learning (MCL) with augmentation (i.e., MolCLR [11]), and MCL without augmentation (removing the diffusion priming phase from MotiL). The following conclusions can be drawn: 1) No pretraining yields the worst results, with disorganized and poorly separated clusters on QM8 (DB index: 5.42), highlighting the necessity of pretraining representation learning for effective molecular representation. 2) The performance of MCL with augmentation and MCL without augmentation is comparable. On one hand, proper molecule graph augmentation helps the GNN identify similar and dissimilar molecules, but sometimes it may have a misleading effect. On the other hand, MCL without augmentation removes the reliance on molecule graph augmentation, but meanwhile ignores the GNN encoder's robustness to structural noise perturbations in molecule graphs and thus does not yield better results than MCL with augmentation. 3) MotiL achieves clear clustering results on the four datasets. For instance, on ESOL, the molecules with four different scaffolds, such as those containing naphthalene (colored yellow) and barbituric acid (colored red), are clearly separated in the representation space. Also, Supplementary Fig. 8 indicates that the diffusion priming phase used in MotiL effectively mitigates the sensitivity of our method to structural noise perturbations.

Next, we plot the t-SNE representation visualization results (Fig. 2b, Supplementary Fig. 3) of all molecules on multiple datasets. For a pair of molecules, we measure their structural similarity by calculating the RDKit fingerprint (RDKFP) similarity [11] based on the Tanimoto coefficient [49] between their fingerprints. From Fig. 2b, we can see that molecular pairs with high RDKFP similarity (above 90%) tend to be closely positioned in the representation space on the BBBP dataset. In addition, we insert an augmented molecule to verify the effect of molecule graph augmentation. The augmented molecule is obtained from the molecule with the compound identifier (CID) of 6196 through subgraph removal and is highlighted with the brown dashed box. Since the removed amide group is a key functional group of the original molecule, the RDKFP similarity between the augmented molecule and the original one is less than 75%. Correspondingly, the augmented molecule is projected farther from the original one in the representation space than another molecule with a similarity

ARTICLE IN PRESS

of 0.951. Interestingly, the polarity of the amide group, its ability to form hydrogen bonds, and its potential to increase molecular rigidity generally reduce a molecule's permeability through the blood-brain barrier [15], which is the molecular property predicted in BBBP. Such results demonstrate that MotiL captures the structural and functional properties of small molecules, and can distinguish between easily confusable molecule pairs constructed via augmentation. More visualization results and analysis can be found in Supplementary Section G.

Finally, we analyze the relationships among 10 common functional group representations by calculating their cosine similarity (Fig. 2c). The results are presented in the form of heatmap where the x-axis and y-axis correspond to the different functional group representations from $G_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\theta}'}$, respectively. We observe that the representations of the same functional group remain similar although they are from different encoders, while those of different functional groups are mostly dissimilar. Such results reflect the different roles that different functional groups play in forming molecule graph representations. Also, some functional groups exhibit higher similarity due to overlapping atoms and bonds in their chemical structures. For example, carboalkoxy, carbonyl, and carboxyl groups all contain carbon-oxygen double bonds, resulting in pairwise representation similarities of no less than 80%, forming a 3×3 block (framed in gray) in the heatmap. Similarly, the carboxamide, amidine, and secondary ketimine groups, which share carbon-nitrogen single or double bonds, exhibit representation similarities of no less than 70% (framed in black). The above analysis demonstrates that the bi-scaled learning mechanism in MotiL enables the learned motif representations to align with the structural properties of functional groups.

MotiL learns macromolecular representations consistent with protein structures and chemical properties

For visualization, we use t-SNE for small molecule embeddings due to its strength in preserving local structure, which aligns with our goal of examining scaffold-level clustering. For protein embeddings, we employ UMAP, which better maintains both local and global relationships in high-dimensional data. This choice improves interpretability when analyzing functional similarity in multi-label classification tasks such as GO and EC prediction. Specifically, we use a non-linear dimension reduction technique named uniform manifold approximation and projection (UMAP) [50] to project the ATP binding protein representations learned by MotiL onto a 2D coordinate space (Fig. 3a). These representations are obtained from the last convolution layer of $G_{\boldsymbol{\theta}}$ in MotiL. We apply the density-based spatial clustering of applications with noise (DBSCAN) [51] clustering algorithm to identify and visualize the cluster IDs of individual proteins. For example, when we focus on Cluster 6, where the shape indicates whether the ATP binding of a protein is predicted correctly, it is evident that proteins 1WLE-A and 3LSS-A are grouped together, and all proteins in this cluster are predicted correctly.

Next, we plot the heatmap of Jaccard similarity for the GO terms related to the molecular function within the clusters containing 10 to 40 proteins, specifically Clusters 1, 6, 7, 9, 10, and 11 (Fig. 3b). We observe that the similarity is generally high within each cluster and relatively low between clusters (Fig. 3b, Supplementary Fig. 4), indicating that the learned representation of each protein is consistent with its molecular functions. We then focus on specific cases, examining the three-dimensional (3D) structures of proteins (Fig. 3c, Supplementary Fig. 5). Fig. 3c illustrates that within Cluster 3, proteins 4AU8-A, 1G3N-A, 2BPM-A, 5XVU-A, and 3VN9-A display significant structural similarity to each other. Additionally, these proteins in the same cluster are linked by shared GO terms, such as protein kinase activity. Moreover, we perform GO term analysis for multiple clusters (Supplementary Fig. 6) and find that different clusters are associated with distinct molecular functions. Finally, we calculate the similarity scores among the representations of 20 amino acids and present them in the similarity heatmap (Fig. 3d, Supplementary Fig. 7) where the x-axis and y-axis correspond to the different amino acid representations from $G_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\theta}'}$, respectively. We can see that the representations of the same amino acid across different encoders are highly similar, whereas different amino acids show lower similarity. Also, MotiL accurately identifies molecular structures with similar aromatic

ARTICLE IN PRESS

ARTICLE IN PRESS

rings, structural isomers, and different side chain structures as highly similar, similar, and dissimilar, respectively (Supplementary Fig. 7).

MotiL corrects the erroneous prediction of the strongest baseline

We present a comprehensive evaluation of prediction disagreements between MotiL and the strongest baseline KANO on two toxicity-related benchmarks: blood-brain barrier permeability (BBBP) and drug-induced liver injury (SIDER). Fig. 4a displays quantitative statistics of disagreement cases between MotiL and KANO on the complete datasets. On the BBBP dataset, the two models produce conflicting predictions on 57 out of 203 test samples (28.1%). Among these, MotiL correctly predicts 40 instances (70.2%), while KANO is correct in only 17 (29.8%). On the SIDER dataset, the models disagree on 63 of 295 samples (21.4%), with MotiL accurately predicting 43 samples (68.3%) and KANO only 20 (31.7%). Fig. 4b, Supplementary Fig. 10, and Supplementary Fig. 11 present the results of the activity cliff experiment, designed to evaluate model robustness on challenging molecular pairs with high structural similarity but divergent labels. We identified activity cliff samples from BBBP and SIDER by selecting molecular pairs with a Tanimoto similarity above 0.85 and differing binary labels. Among the disagreement cases on these activity cliffs, MotiL achieves higher accuracy than KANO: on BBBP, MotiL correctly predicts 10 out of 15 samples (66.7%) compared to KANO's 5 (33.3%); on SIDER, MotiL is correct on 11 of 19 samples (57.9%), while KANO predicts 8 correctly (42.1%). The results in Fig. 4a and Fig. 4b demonstrate that MotiL outperforms KANO in handling ambiguous or challenging samples. Fig. 4c and Supplementary Fig. 9a illustrate molecular structures from selected cases in the BBBP dataset where MotiL outperforms KANO. Streptomycin, a large and highly polar aminoglycoside antibiotic, has poor BBB permeability due to its molecular weight (581 Da) and multiple hydrophilic groups. While KANO misclassifies it as permeable, MotiL correctly predicts its non-permeability. Mexazolam, a benzodiazepine derivative, exhibits high lipophilicity and low molecular weight, making it BBB-permeable. MotiL correctly identifies this molecule's property, while KANO fails. Fig. 4d and Supplementary Fig. 9b show representative cases from the SIDER dataset. Erlotinib, a tyrosine kinase inhibitor, contains aromatic rings and is known to form reactive metabolites, contributing to hepatotoxicity. MotiL successfully predicts this risk. Vincristine, with its complex alkaloid structure and nitrogenous rings, has documented hepatotoxic potential due to metabolism-induced stress. Again, MotiL correctly classifies this molecule, while KANO does not. These qualitative results complement our statistical findings and provide insight into MotiL's decision-making process, especially in chemically intricate or borderline cases.

Discussion

In this work, we present MotiL, a pretraining representation learning method for molecules. MotiL operates in three phases: diffusion priming, bi-scaled training, and task-specific fine-tuning. We integrate the generative diffusion model, network pruning, and bi-scaled training loss into MotiL to enable molecular motif learning. Experimental results demonstrate that MotiL achieves superior performance on 14 small molecule benchmarks and 2 protein macromolecule benchmarks. More importantly, the representations of small molecules, proteins, functional groups, and amino acids learned by MotiL align well with their structural and chemical properties. Our results advance molecular representation learning by yielding embeddings that more faithfully capture underlying chemical structure, motif-level information, and molecular function. This improvement supports more accurate property prediction across small molecules and macromolecules, with potential utility in pharmaceutical research and bioinformatics.

There are many promising directions to explore in future work. For instance, exploring more advanced generative models for the diffusion priming phase could further enhance the quality of molecular representations. Also, exploring the hybrid use of generative diffusion models and representation learning for new molecule design can accelerate drug discovery, enhance targeted therapies, and advance personalized medicine. The capability to design molecules may result in more effective treatments, a deeper understanding of

biological processes, and new opportunities in synthetic biology and disease diagnosis.

Methods

In a small molecule graph $\mathcal{G}$, the atom set is represented by $\mathcal{V}$ and the chemical bond set by $\mathcal{E}$. In this context, an atom $u{\in}\mathcal{V}$ is treated as a node of the graph, while a chemical bond $(u,v){\in}\mathcal{E}$ is considered an edge connecting atoms u and v. The initial node feature matrix is denoted by $\mathbf{X}$ and the initial features of a node u are denoted by $\mathbf{x}_u$. We denote the adjacency matrix of $\mathcal{G}$ as $\mathbf{A}$, where each entry is either 0 or 1: 1 if there is an edge between two nodes, otherwise 0. Therefore, we can represent $\mathcal{G}{=}(\mathbf{A},\mathbf{X})$. For a protein graph, the nodes correspond to amino acid residues, and the edges correspond to peptide bonds. In the following method description, the operations on small molecule graphs are elaborated, while similar operations on protein graphs are not detailed for the sake of brevity.

Diffusion priming

Let the perturbed adjacency matrix and perturbed graph at the t-th time step be $\mathbf{A}_t$ and $\mathcal{G}_t$, respectively. In the diffusion priming phase, the perturbed molecule graph $\mathcal{G}_T{=}(\mathbf{A}_T,\mathbf{X})$ at the T-th time step is produced in the noise addition process, and then fed into $G_{\boldsymbol{\theta}}$ in the noise reduction process (Fig. 1a). The details are as follows:

1. Noise addition process: The generative diffusion model DiffMoM progressively perturbs the adjacency matrix $\mathbf{A}_t$ of the molecular graph $\mathcal{G}$ by adding Gaussian noise, where the level of noise is controlled through a predefined diffusion schedule. At each time step t $(t{\in}[1,2,\cdots,T])$, the graph $\mathcal{G}_{t-1}$ from the previous step is perturbed to produce $\mathcal{G}_t$ as follows:

$$\begin{aligned}\mathcal{G}_t &= (\mathbf{A}_t,\mathbf{X}),\\ \mathbf{A}_t &= \sqrt{\alpha_t}\cdot\mathbf{A}_{t-1}+\sqrt{1-\alpha_t}\cdot\mathbf{s}_t,\\ \mathbf{A}_0 &= \mathbf{A},\end{aligned}\tag{1}$$

where α_t is a noise scheduling coefficient and $\mathbf{s}_t\sim\mathcal{N}(0,\mathbf{I})$ represents Gaussian noise. The noise scheduling coefficient α_t is typically defined as a predefined sequence using the following formula:

$$\alpha_t = 1-\Big(\frac{t}{T}\cdot(\beta_{\max}-\beta_{\min})+\beta_{\min}\Big),\tag{2}$$

where $\beta_{\min}$ and $\beta_{\max}$ represent the minimum amount of noise introduced in the early time step $(t=0)$ and the maximum amount of noise introduced in the final time step $(t=T)$, respectively.

2. Noise reduction process: In the noise reduction process, the parameterized GNN encoder $G_{\boldsymbol{\theta}}$ is trained to reverse noise perturbations and reconstruct the original graph $\mathcal{G}$ from $\mathcal{G}_T$. Specifically, the combination of $G_{\boldsymbol{\theta}}$ and an MLP encoder $M_{\boldsymbol{\Phi}}$ first predicts the noise $\mathbf{s}'_T$ at the T-th time step:

$$\mathbf{s}'_T = M_{\boldsymbol{\Phi}}(G_{\boldsymbol{\theta}}(\mathbf{A}_T,\mathbf{X})).\tag{3}$$

We use the reparameterization trick to avoid redundant computations [52, 53], i.e., the predicted noise $\mathbf{s}'_T$ is directly used to calculate the reconstructed adjacency matrix $\hat{\mathbf{A}}$ of $\hat{\mathcal{G}}$ in one step:

$$\hat{\mathbf{A}} = \frac{\mathbf{A}_T-\sqrt{1-\bar{\alpha}_T}\cdot\mathbf{s}'_T}{\sqrt{\bar{\alpha}_T}},\tag{4}$$

where $\bar{\alpha}_T=\alpha_T\alpha_{T-1}\cdots\alpha_1$.

3. Training objective: The overall objective is to denoise $\mathcal{G}_T$, eventually reconstructing the original molecule graph from pure noise. Therefore, the loss function $\mathcal{L}_{\text{DiffMoM}}$ of DiffMoM is designed to measure how well $G_{\boldsymbol{\theta}}$ can recover $\mathbf{A}$:

$$\mathcal{L}_{\text{DiffMoM}} = \left\|\mathbf{A}-\hat{\mathbf{A}}\right\|^2.\tag{5}$$

Bi-scaled training

The phase of bi-scaled training includes two main parts: bi-scaled network encoding and bi-scaled representation optimization.

Bi-scaled network encoding

In $G_{\boldsymbol{\theta}}$, the representation of a node is updated through the message passing mechanism as follows:

$$\begin{aligned}\mathbf{h}_v^{(l+1)}{=}\sigma\Big(\sum_{u\in\mathcal{N}(v)}\mathbf{W}^{(l)}\mathbf{h}_u^{(l)}{+}\mathbf{b}^{(l)}\Big),\\ l\in\{0,1,\cdots,L-1\},\end{aligned}\tag{6}$$

ARTICLE IN PRESS

where $\mathbf{h}_u^{(l)}$ is the representation of node u at the l-th layer ($\mathbf{h}_u^{(0)}$=$\mathbf{x}_u$), $\mathcal{N}(v)$ is the neighbor set of node v, $\mathbf{W}^{(l)}$ is the weight matrix at the l-th layer, $\mathbf{b}^{(l)}$ is the bias vector at the l-th layer, and σ is an activation function like the rectified linear unit (ReLU). For an L-layer GNN encoder $G_{\boldsymbol{\theta}}$, we define the parameter group $\boldsymbol{\theta}=\left\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(L)}\right\}$.

To prune GNN, we apply random masking to its weight matrix $\mathbf{W}^{(l)}$. Given the retaining probability p, the weight matrix $\mathbf{W}_{\text{prune}}^{(l)}$ after masking is defined as follows:

$$\mathbf{W}_{\text{prune}}^{(l)} = \mathbf{W}^{(l)} \odot \mathbf{M}_p, \tag{7}$$

where $\odot$ denotes element-wise Hadamard product. $\mathbf{M}_p$ is a binary masking matrix with the same shape as $\mathbf{W}^{(l)}$. Each entry in $\mathbf{M}_p$ is 1 with probability p (indicating that the corresponding weight is retained) and 0 with probability $1-p$ (indicating that the corresponding weight is dropped). Thus, in $G_{\boldsymbol{\theta}'}$, the update formula for node representation $\tilde{\mathbf{h}}_u^{(l)}$ after applying masking becomes:

$$\begin{aligned}\tilde{\mathbf{h}}_v^{(l+1)}=\sigma\Big(\sum_{u\in\mathcal{N}(v)} \mathbf{W}_{\text{prune}}^{(l)}\tilde{\mathbf{h}}_u^{(l)}+\mathbf{b}^{(l)}\Big),\\ l\in\{0,1,\cdots,L-1\}.\end{aligned} \tag{8}$$

where $\tilde{\mathbf{h}}_u^{(0)}$=$\mathbf{x}_u$. Also, we define the parameter group $\boldsymbol{\theta}'=\left\{\mathbf{W}_{\text{prune}}^{(1)}, \mathbf{W}_{\text{prune}}^{(2)}, \cdots, \mathbf{W}_{\text{prune}}^{(L)}\right\}$. Finally, the output node representations of two encoders are defined as $\mathbf{h}_v$=$\mathbf{h}_v^{(L)}$ and $\tilde{\mathbf{h}}_v$=$\tilde{\mathbf{h}}_v^{(L)}$.

Bi-scaled representation optimization

After obtaining node representations, given the graph $\mathcal{G}$ and motif $\mathcal{G}_{\text{motif}}$, we can further derive the graph representation $\mathbf{H}$ and motif representation $\mathbf{S}$ using a summing-average graph readout function:

$$\mathbf{H}=\frac{1}{|\mathcal{V}|}\sum_{k\in\mathcal{V}}\mathbf{h}_k, \quad \mathbf{S}=\frac{1}{|\mathcal{V}_{\text{motif}}|}\sum_{k\in\mathcal{V}_{\text{motif}}}\mathbf{h}_k, \tag{9}$$

where $\mathcal{V}$ and $\mathcal{V}_{\text{motif}}$ are the node sets of $\mathcal{G}$ and $\mathcal{G}_{\text{motif}}$, respectively. Another encoder, $G'_{\boldsymbol{\theta}}$, produces graph and motif representations in the same manner as described in Eq. (9). For brevity, the details are omitted here.

Then the loss $\mathcal{L}_{\text{bi-scale}}$ during the bi-scaled training phase is defined as:

$$\mathcal{L}_{\text{bi-scale}} = \mathcal{L}_{\mathcal{G}} + \alpha \cdot \mathcal{L}_{\mathcal{G}_{\text{motif}}}, \tag{10}$$

where α is a balance coefficient between $\mathcal{L}_{\mathcal{G}}$ and $\mathcal{L}_{\mathcal{G}_{\text{motif}}}$. Given two cross-correlation matrices $\mathbf{C}$=$[c_{i,j}]$ and $\mathbf{C}'$=$[c'_{i,j}]$, the graph-level training loss $\mathcal{L}_{\mathcal{G}}$ and motif-level training loss $\mathcal{L}_{\mathcal{G}_{\text{motif}}}$ are defined based on the Barlow Twins loss [54, 55]:

$$\begin{aligned}\mathcal{L}_{\mathcal{G}} &= \sum_{k=1}^{D}(1-c_{k,k})^2+\frac{1}{D}\sum_{i=1}^{D}\sum_{j=1,j\neq i}^{D}c_{i,j}^2,\\ \mathcal{L}_{\mathcal{G}_{\text{motif}}} &= \sum_{k=1}^{D}(1-{c'}_{k,k})^2+\frac{1}{D}\sum_{i=1}^{D}\sum_{j=1,j\neq i}^{D}{c'}_{i,j}^2,\end{aligned} \tag{11}$$

where D is the representation dimension. Let the number of graphs and motifs be N_{graph} and N_{motif}, respectively, $\mathbf{C}$ and $\mathbf{C}'$ are calculated as follows:

$$\begin{aligned}\mathbf{C} &= \frac{1}{N_{graph}}\mathbf{H}_{\text{norm}}^T\mathbf{H}_{\text{norm}},\\ \mathbf{C}' &= \frac{1}{N_{motif}}\mathbf{S}_{\text{norm}}^T\mathbf{S}_{\text{norm}},\end{aligned} \tag{12}$$

where $\mathbf{H}_{\text{norm}}$ and $\mathbf{S}_{\text{norm}}$ represent the normalized representations of graphs and motifs, respectively. We denote the normalized representations of the i-th graph and i-th motif as $\mathbf{H}_{i,\text{norm}}$ and $\mathbf{S}_{i,\text{norm}}$ in $\mathbf{H}_{\text{norm}}$ and $\mathbf{S}_{\text{norm}}$, respectively. They are derived from the i-th graph representation $\mathbf{H}_i$ and i-th motif representation $\mathbf{S}_i$ (without loss of generality, $\mathbf{H}_i$ and $\mathbf{S}_i$ are defined similarly to Eq. (9)):

$$\mathbf{H}_{i,\text{norm}}=\frac{\mathbf{H}_i-\mu_i}{\sigma_i+\epsilon}, \quad \mathbf{S}_{i,\text{norm}}=\frac{\mathbf{S}_i-\mu'_i}{\sigma'_i+\epsilon}, \tag{13}$$

where μ_i, μ'_i are mean values, σ_i, σ'_i are standard deviation values, and ϵ is a small positive constant added to prevent overflow due to division by zero.

Task-specific fine-tuning

In the third phase, we apply the pretrained $G_{\boldsymbol{\theta}}$ to the molecular property prediction task to fine-tune its parameter group $\boldsymbol{\theta}$.

For the classification task, given a new molecule graph $\mathcal{G}'$=$(\mathbf{A}', \mathbf{X}')$, we first obtain its logit vector $\hat{\mathbf{y}}$=$[\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_C]$ (C is the number

of classes) by using $G_{\boldsymbol{\theta}}$ in combination with an MLP-based classifier $M_{\boldsymbol{\psi}}$:

$$\hat{\mathbf{y}} = \text{Softmax}(M_{\boldsymbol{\psi}}(G_{\boldsymbol{\theta}}(\mathbf{A}', \mathbf{X}'))). \quad (14)$$

Then we minimize the cross-entropy loss $\mathcal{L}_{\text{CE}}$ between $\hat{\mathbf{y}}$ and the true label's one-hot vector $\mathbf{y} = [y_1, y_2, \cdots, y_C]$:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c. \quad (15)$$

For the regression task, given a new molecule graph $\mathcal{G}'=(\mathbf{A}', \mathbf{X}')$, we first obtain its predicted value $\hat{y}$ by using $G_{\boldsymbol{\theta}}$ in combination with an MLP-based classifier $M_{\boldsymbol{\omega}}$:

$$\hat{y} = M_{\boldsymbol{\omega}}(G_{\boldsymbol{\theta}}(\mathbf{A}', \mathbf{X}')). \quad (16)$$

Then we minimize the mean squared error loss $\mathcal{L}_{\text{MSE}}$ between $\hat{y}$ and the true score y of $\mathcal{G}'$:

$$\mathcal{L}_{\text{MSE}} = (\hat{y} - y)^2. \quad (17)$$

Datasets

1. Small molecule datasets: For MotiL pretraining, which includes diffusion priming and bi-scaled training, we sample 250,000 unlabeled molecules from ZINC15 [21], a publicly accessible database of purchasable drug-like compounds. For task-specific fine-tuning, we utilize 14 benchmark datasets from MoleculeNet [22], encompassing 678 binary classification tasks and 19 regression tasks across domains such as drugs, biology, physics, and chemistry. Following the methods in previous work [11, 12], we conduct three independent runs using three random-seeded scaffold splits for all datasets, except QM9, with a training/-validation/testing ratio of 8:1:1. Scaffold splitting [56] is a more challenging task that partitions molecules based on their scaffolds, offering a better assessment of model generalization on out-of-distribution data. QM9 follows the random splitting strategy commonly used in related works [30, 11, 12] for comparison.

2. Protein macromolecule datasets: The task in the Gene Ontology (GO) dataset focuses on predicting protein functions across multiple GO terms. Following ref. [46], proteins are divided 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 GO dataset consists of 29,898 proteins for training, 3,322 for validation, and 3,415 for testing. The task in the Enzyme Commission (EC) dataset focuses on predicting three-level and four-level EC numbers, following the splits in ref. [46], with 15,550 proteins for training, 1,729 for validation, and 1,919 for testing. EC number prediction is also a multi-label classification task. We use 1D sequential orders and 3D geometric coordinates of amino acids from the protein datasets for MotiL pretraining, and incorporate the label information of GO terms or EC numbers during the task-specific fine-tuning phase. Further details about the experimental datasets are provided in Supplementary Section B.

Baselines

For the small molecular property prediction task, we compare MotiL with 6 supervised learning methods and 10 self-supervised learning methods. The supervised learning methods include two widely used GNNs (GCN [34] and GIN [35]), one geometry-based GNN (GEM [10]), and three message passing neural networks (MPNN [36], DMPNN [37], and CMPNN [38]). The self-supervised learning methods include four predictive self-supervised learning methods (N-GRAM [30], Hu et al. [29], MGSSL [28], and GROVER [27]) and five contrastive self-supervised learning methods (GraphMVP [39], MolCLR [11], $\text{MolCLR}_{\text{CMPNN}}$ [11], and KANO [12]). For the protein macromolecular property prediction task, we compare MotiL with 4 sequence-based methods, 3 graph-based methods, and 8 hybrid methods. Specifically, sequence-based methods include CNN [25], ResNet [26], LSTM [26], and Transformer [26]; graph-based methods include GCN [34], GAT [40], and 3D CNN [41]; hybrid methods include GraphQA [42], GVP [43], $\text{IEConv}_{\text{residue level}}$ [44], GearNet [45], $\text{GearNet}_{\text{IEConv}}$ [45], $\text{GearNet}_{\text{Edge}}$ [45], $\text{GearNet}_{\text{Edge-IEConv}}$ [45], and CDConv [24]. Further details about baselines and related work are provided in Supplementary Section A.

ARTICLE IN PRESS

Training details

We use CMPNN [38] and CDConv [24] as the specific GNN encoders in MotiL for micro- and macromolecule tasks, respectively. CMPNN enhances graph representations by strengthening the message interaction between atoms and bonds. CDConv applies discrete and continuous convolutions to model regular and irregular structures, respectively. In the small molecule dataset, the initial features of atoms are determined by eight attributes (e.g., chirality, hybridization, and atomic mass), while bonds are represented by four attributes (e.g., bond type and conjugation). For the protein macromolecule dataset, we use the sequence and geometric structure information of proteins as inputs to CDConv.

During the diffusion priming phase, we optimize the loss $\mathcal{L}_{\text{DiffMoM}}$ using the Adam optimizer [57] with a learning rate of 3e-5. The iteration number is set to T=1,000, and the MLP encoder $M_{\boldsymbol{\Phi}}$ is a 3-layer network with hidden sizes of 300, 300, and 50. In the bi-scaled training phase, the Adam optimizer with a learning rate of 3e-5 is used to optimize the bi-scaled training loss $\mathcal{L}_{\text{bi-scale}}$. $G_{\boldsymbol{\theta}}$ is trained with a batch size of 1,024 over 50 and 5 epochs for micro- and macromolecules, respectively. The balance coefficient α in Eq. (10) is set to 1.0. In the task-specific fine-tuning phase, we add a two-layer MLP on top of the pretrained $G_{\boldsymbol{\theta}}$ for molecular property prediction. We fine-tune $G_{\boldsymbol{\theta}}$ for N_e epochs ($N_e \in [10, 200]$) during each independent run. The Adam optimizer is used to update $G_{\boldsymbol{\theta}}$ with a learning rate ranging from 1e-6 to 1e-3 for all datasets. Fine-tuning is performed only on the training set, with hyperparameter tuning on the validation set to achieve optimal results. MotiL was implemented using PyTorch (version 1.12.1) and other open-source libraries including PyTorch Geometric (version 2.3.0), RDKit (version 2022.9.1), and scikit-learn (version 0.24.2). All experiments were conducted on an Ubuntu server equipped with NVIDIA GeForce RTX 3090 GPUs.

Data availability

The pretraining data and molecular property prediction benchmarks used in this study are available on the Hugging Face Hub at https://huggingface.co/MotiL [58] and in the GitHub repository at https://github.com/Young0222/MotiL [59]. Source data are provided with this paper.

Code availability

The source code of this work is freely available in the Code Ocean capsule at https://codeocean.com/capsule/9672051/tree [60] and the GitHub repository at https://github.com/Young0222/MotiL [59].

ARTICLE IN PRESS

ARTICLE IN PRESS

References

[1] Diaz, N. *et al.* Discovery of potent small-molecule inhibitors of lipoprotein(a) formation. *Nature* **629**, 945–950 (2024).

[2] An, L. *et al.* Binding and sensing diverse small molecules using shape-complementary pseudocycles. *Science* **385**, 276–282 (2024).

[3] Tong, Y. *et al.* Programming inactive rna-binding small molecules into bioactive degraders. *Nature* **618**, 169–179 (2023).

[4] Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence* **6**, 170–179 (2024).

[5] Wang, X. *et al.* Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. *Nature Machine Intelligence* **5**, 445–456 (2023).

[6] Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).

[7] Wan, Z. *et al.* Unconventional superconductivity in chiral molecule–tas2 hybrid superlattices. *Nature* **632**, 69–74 (2024).

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

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

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

[11] 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).

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

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

[14] 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).

[15] 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).

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

[17] Baumgartner, C. K. *et al.* The ptpn2/ptpn1 inhibitor abbv-cls-484 unleashes potent anti-tumour immunity. *Nature* **622**, 850–862 (2023).

[18] Bianco, S. *et al.* Mechanical release of homogenous proteins from supramolecular gels. *Nature* **631**, 544–548 (2024).

[19] Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nature Biotechnology* **32**, 40–51 (2014).

[20] Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery* **18**, 495–496 (2019).

[21] Sterling, T. & Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling* **55**, 2324–2337 (2015).

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

[23] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

[24] 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).

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

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

[27] Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing*

ARTICLE IN PRESS

ARTICLE IN PRESS

Systems **33**, 12559–12571 (2020).

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

[30] 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).

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

[32] 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).

[33] 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).

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

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

[36] 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).

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

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

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

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

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

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

[43] 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).

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

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

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

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

[48] Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**, 224–227 (1979).

[49] Rácz, A., Bajusz, D. & Héberger, K. Life beyond the tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of Cheminformatics* **10(1)**, 48:1–48:12 (2018).

[50] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *Journal of Open*

ARTICLE IN PRESS

ARTICLE IN PRESS

Source Software **3(29)**, 861 (2018).

[51] 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).

[52] Chung, H., Kim, J. & Ye, J. C. Direct diffusion bridge using data consistency for inverse problems. *Advances in Neural Information Processing Systems* **36**, 7158–7169 (2023).

[53] 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**, 80730–80743 (2023).

[54] Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *Proceedings of Machine Learning Research* **139**, 12310–12320 (2021).

[55] Bielak, P., Kajdanowicz, T. & Chawla, N. V. Graph barlow twins: A self-supervised representation learning framework for graphs. *Knowledge-Based Systems* **256**, 109631 (2022).

[56] Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More* (O'Reilly, 2019).

[57] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *The 3rd International Conference on Learning Representations* (2015).

[58] Liu, Z. *et al.* Molecular motif learning as a pretraining objective for molecular property prediction. motil. *DOI:10.57967/hf/6302* (2025).

[59] Liu, Z. *et al.* Molecular motif learning as a pretraining objective for molecular property prediction. motil: Initial release. *DOI:10.5281/zenodo.16916467* (2025).

[60] Liu, Z. *et al.* Molecular motif learning as a pretraining objective for molecular property prediction. molecular motif learning (initial release). *DOI:10.24433/CO.9672051.v2* (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62372264 (Z.L., C.K.W., and S.Z.), Grant No. 61872207 (Z.L., C.K.W., C.W., and H.F.), Grant No. 92357302 (L.X., Y.Z., L.R., and P.L.), and Grant No. 92157107 (L.X., Y.Z., L.R., and P.L.). This work was also sponsored by Sina Inc. (C.K.W.).

Author contributions

Z.L. proposed the idea, designed the algorithm, developed the code, conducted the experiments, analyzed the results, and drafted the initial manuscript. C.K.W. and P.L. supervised all aspects of the project. C.K.W. revised the manuscript. S.Z. and Z.L. jointly prepared Figure 1. C.W. and H.F. assisted with data preprocessing and experimental evaluation. L.X. contributed to experimental design and technical discussions. Y.Z. supervised and guided the case analysis. L.R. and Z.L. jointly conducted the GO term analysis experiments. All authors reviewed and commented on the manuscript.

Competing interests

The Authors declare no competing interests.

ARTICLE IN PRESS

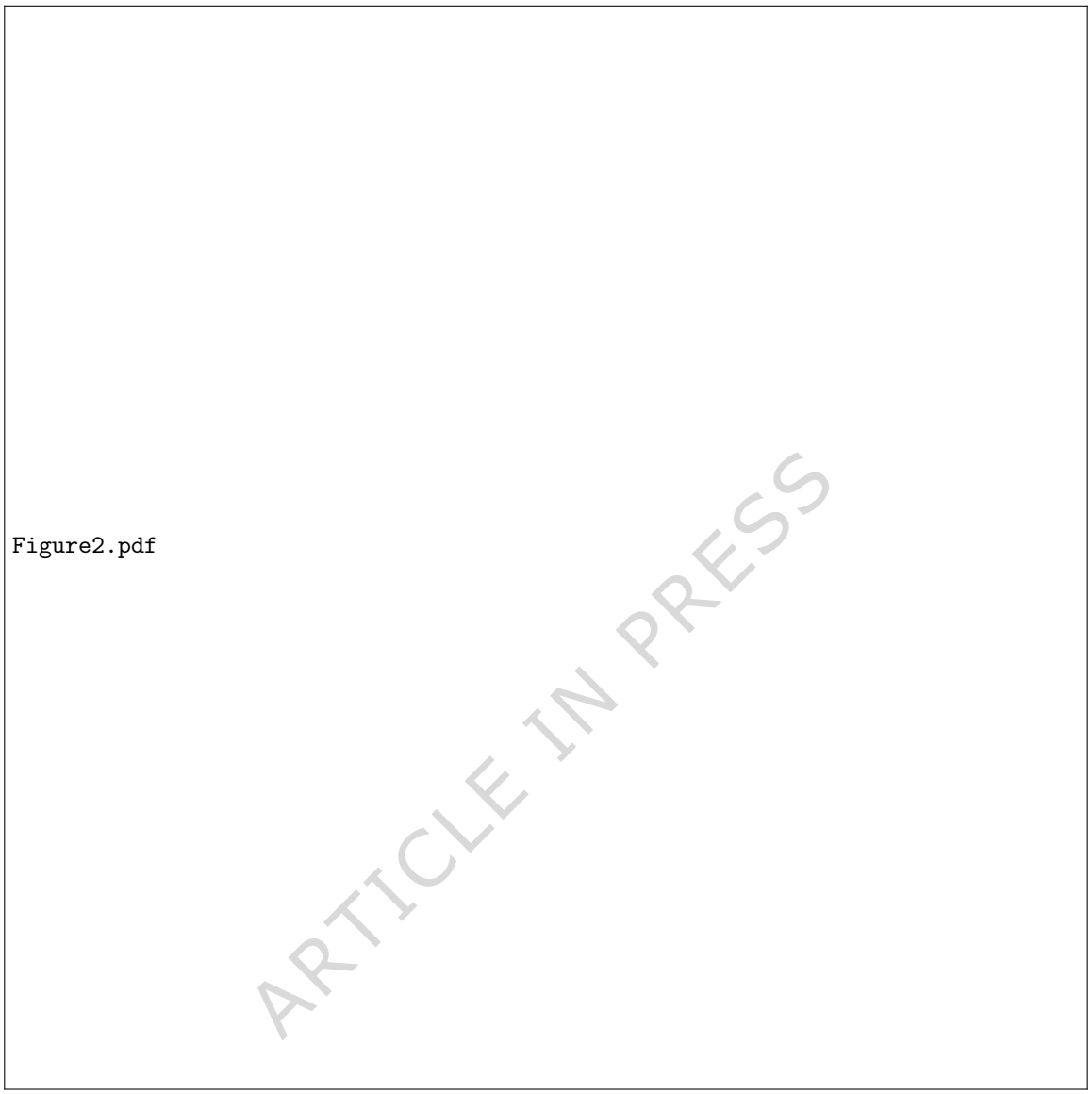
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig. 1 Overview of molecular motif learning (MotiL). a, Design of MotiL. MotiL includes three goals: global structure perception, local feature integration, and real-world label alignment. To achieve these goals, we design diffusion priming, bi-scaled training, and task-specific fine-tuning in MotiL, respectively. **b,** Diffusion priming. The diffusion molecular model (DiffMoM) is represented by the downward arrow indicating the noise addition process (from the original graph $\mathcal{G}$ to the perturbed graph $\mathcal{G}_T$, i.e., the lower part of panel b) and the upward arrow indicating the noise reduction process, where the parameterized graph neural network (GNN) encoder $G_{\boldsymbol{\theta}}$ reconstructs the original graph $\mathcal{G}$ from $\mathcal{G}_T$, with $\boldsymbol{\theta}$ denoting the trainable parameters. DiffMoM enables $G_{\boldsymbol{\theta}}$ to accurately capture the molecular structure before representation learning. **c,** Bi-scaled training. We construct positive and negative pairs based on the original and pruned networks. Additionally, we design the bi-scaled representation alignment mechanism to ensure consistency between representations from the same graph (i.e., molecule) or motif (i.e., functional group). Since the molecular representation (e.g., Aspirin's representation $\mathbf{H}_1$) or functional group's representation (e.g., Benzene ring's representation $\mathbf{S}_1$) is a vector composed of multiple values, the lines on the molecular representations in panels c and d are used to separate different values. **d,** Task-specific fine-tuning. The pretrained GNN encoders produce representations for unseen molecules, which are then fed into a multilayer perceptron (MLP) classifier to learn the relationship between these graph representations and the true molecular property labels.

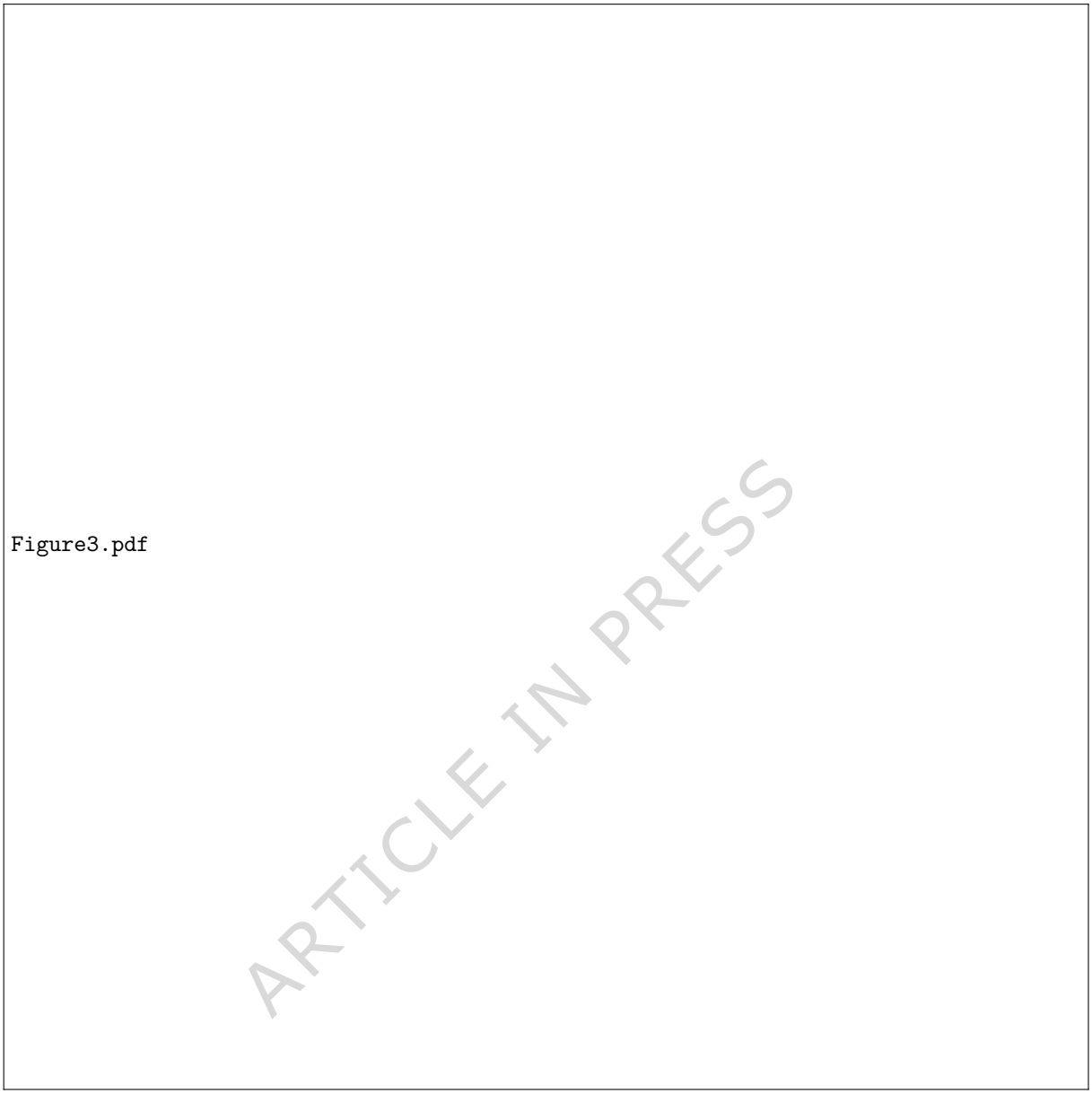
ARTICLE IN PRESS



ARTICLE IN PRESS

Fig. 2 Analysis of micromolecular representations learned by MotiL. a, *t*-distributed stochastic neighbor embedding (t-SNE) visualization of the molecular representations learned by MotiL and other methods. Different colors represent different molecular scaffolds, and a lower Davies-Bouldin (DB) index reflects better separation and clearer distinction between the clusters. **b,** Comparison of positions in 2D space (obtained using t-SNE representation) for structurally similar and dissimilar molecular pairs on the dataset of blood-brain barrier permeability (BBBP). **c,** Heatmap of the functional group's representation similarity. The highlighted gray block corresponds to carboalkoxy, carbonyl, and carboxyl groups, all of which contain carbon–oxygen double bonds and exhibit similarities of no less than 80%. The highlighted black block corresponds to carboxamide, amidine, and secondary ketimine groups, which share carbon–nitrogen single or double bonds and exhibit similarities of no less than 70%. Source data are provided as a Source Data file.

ARTICLE IN PRESS



ARTICLE IN PRESS

Fig. 3 Analysis of macromolecular representations learned by MotiL. a, Uniform manifold approximation and projection (UMAP) visualization of the representations for adenosine triphosphate (ATP) binding proteins. Different colors indicate different clusters identified by the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm. Cluster ID -1 represents noise points as defined by DBSCAN, i.e., points not assigned to any cluster. **b,** Heatmap of the Jaccard similarity calculated based on the protein's gene ontology (GO) terms among multiple clusters. **c,** 3D structure comparison of proteins in cluster ID 3. Proteins within the same cluster show similar 3D conformations and share consistent GO annotations such as protein kinase activity. **d,** Heatmap of the amino acid's representation similarity. Source data are provided as a Source Data file.

ARTICLE IN PRESS

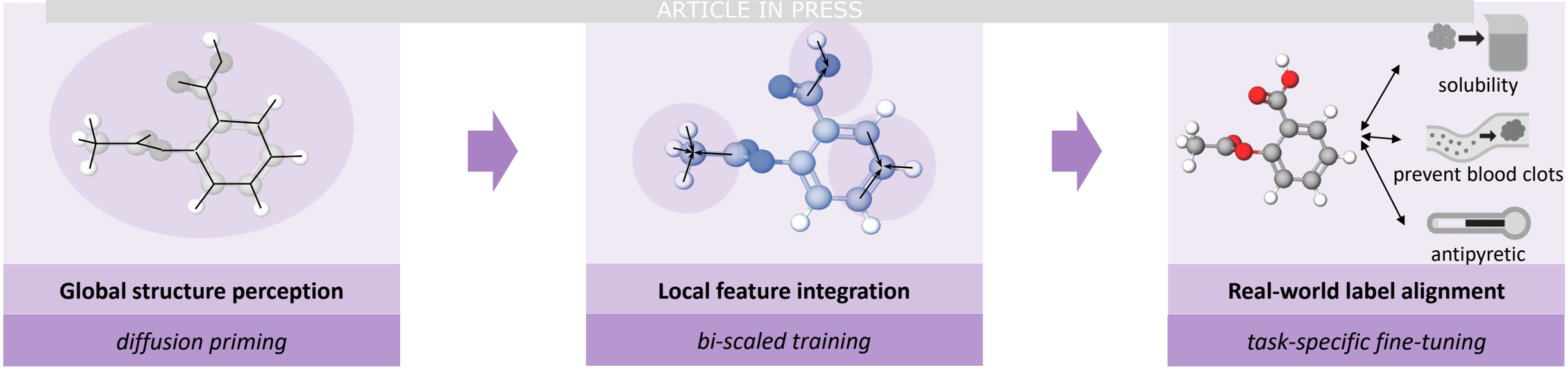


ARTICLE IN PRESS

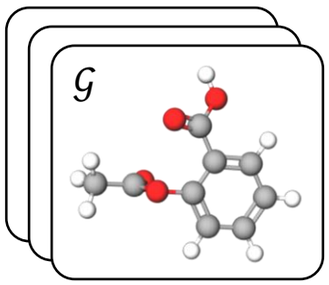
Fig. 4 Quantitative and qualitative analysis of prediction differences between KANO [12] and MotiL on BBBP and SIDER. a, Statistics on disagreements between the two models on the full datasets. **b,** Statistics on disagreements between the two models on activity cliff samples. **c,** Prediction cases for blood-brain barrier permeability on BBBP. **d,** Prediction cases for hepatotoxicity on SIDER. Source data are provided as a Source Data file.

ARTICLE IN PRESS

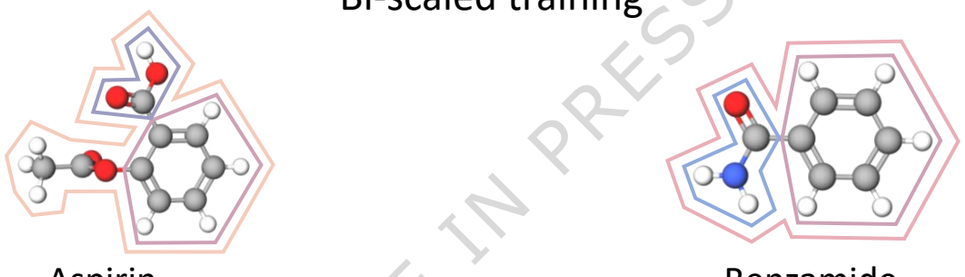
a MotiL



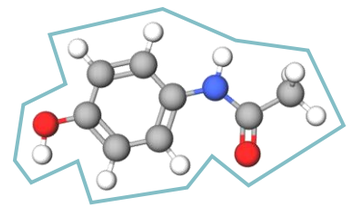
b Diffusion priming

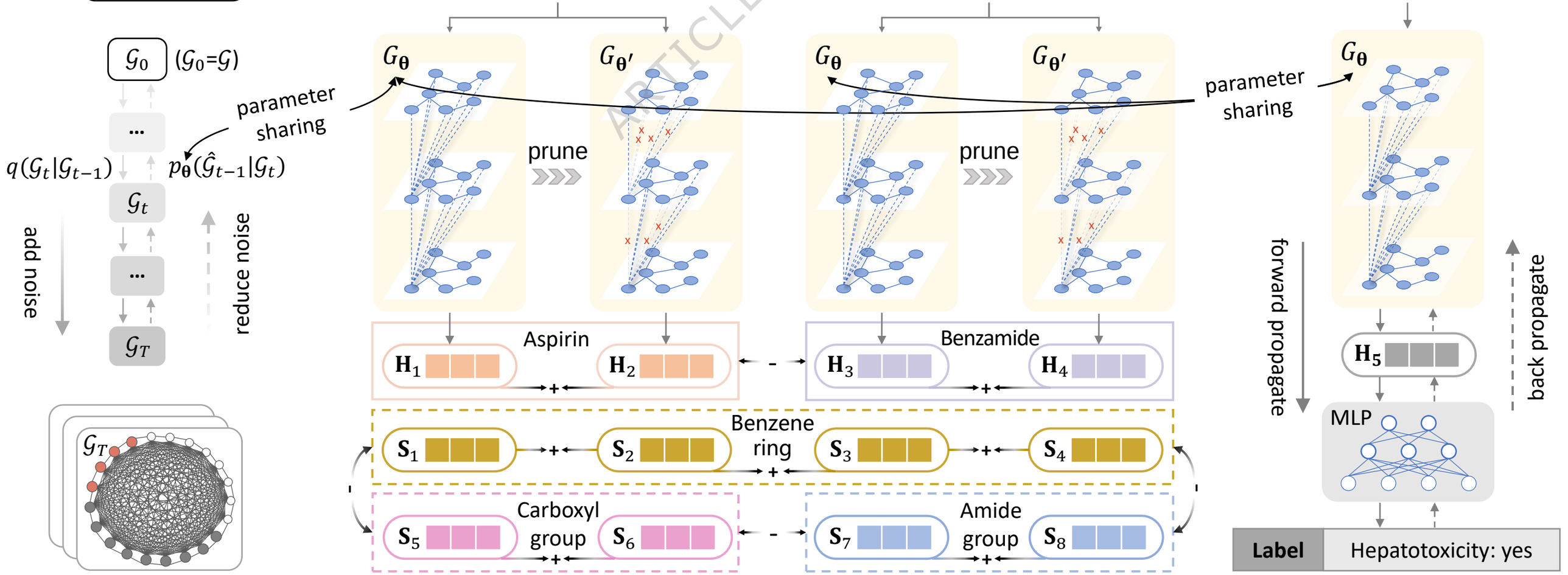


c Bi-scaled training



d Task-specific fine-tuning





ARTICLE IN PRESS

a

No pretraining | MCL with augmentation | MCL without augmentation | MolIL

BBBP

DB index: 3.99 | DB index: 2.46 | DB index: 1.52 | DB index: 0.79

BACE

DB index: 3.51 | DB index: 1.34 | DB index: 1.15 | DB index: 0.69

ESOL

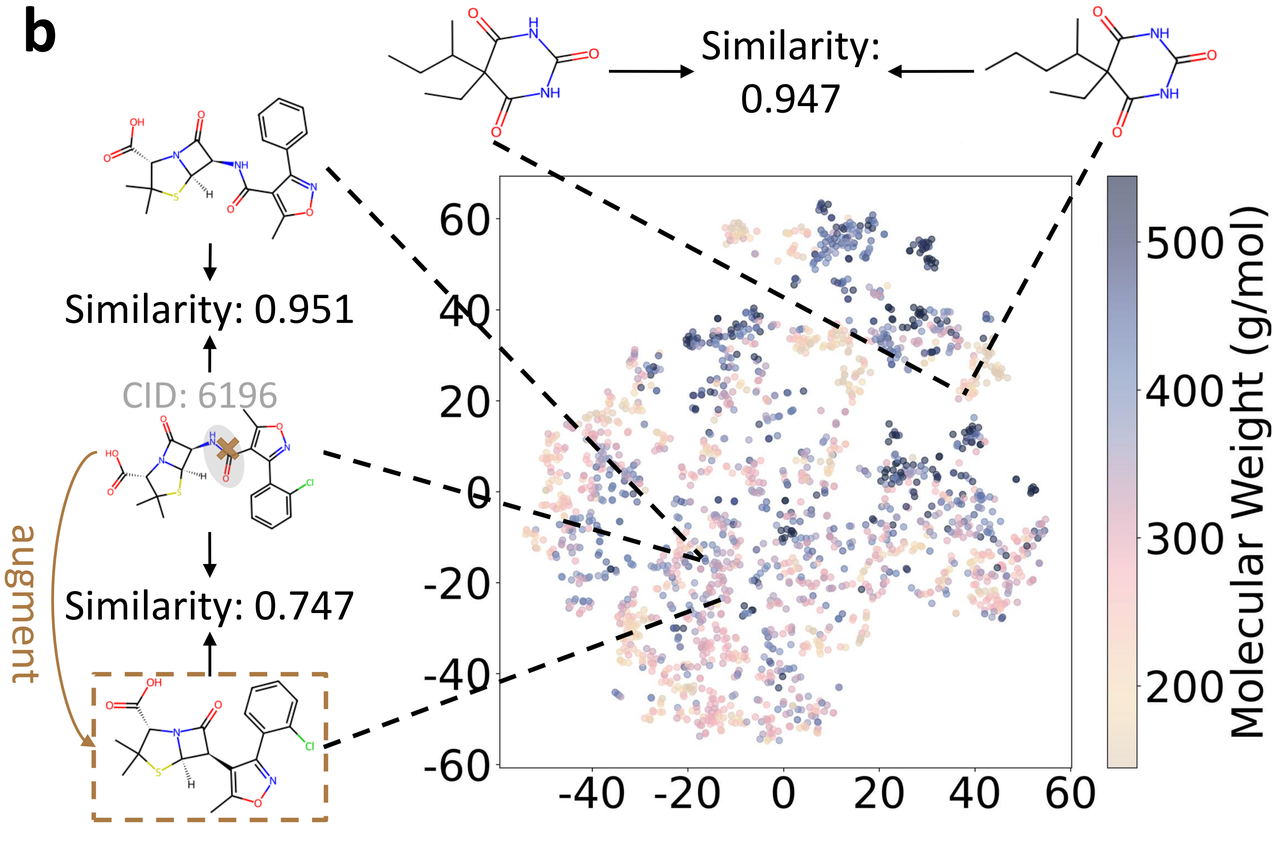
DB index: 1.91 | DB index: 0.60 | DB index: 0.61 | DB index: 0.21

QM8

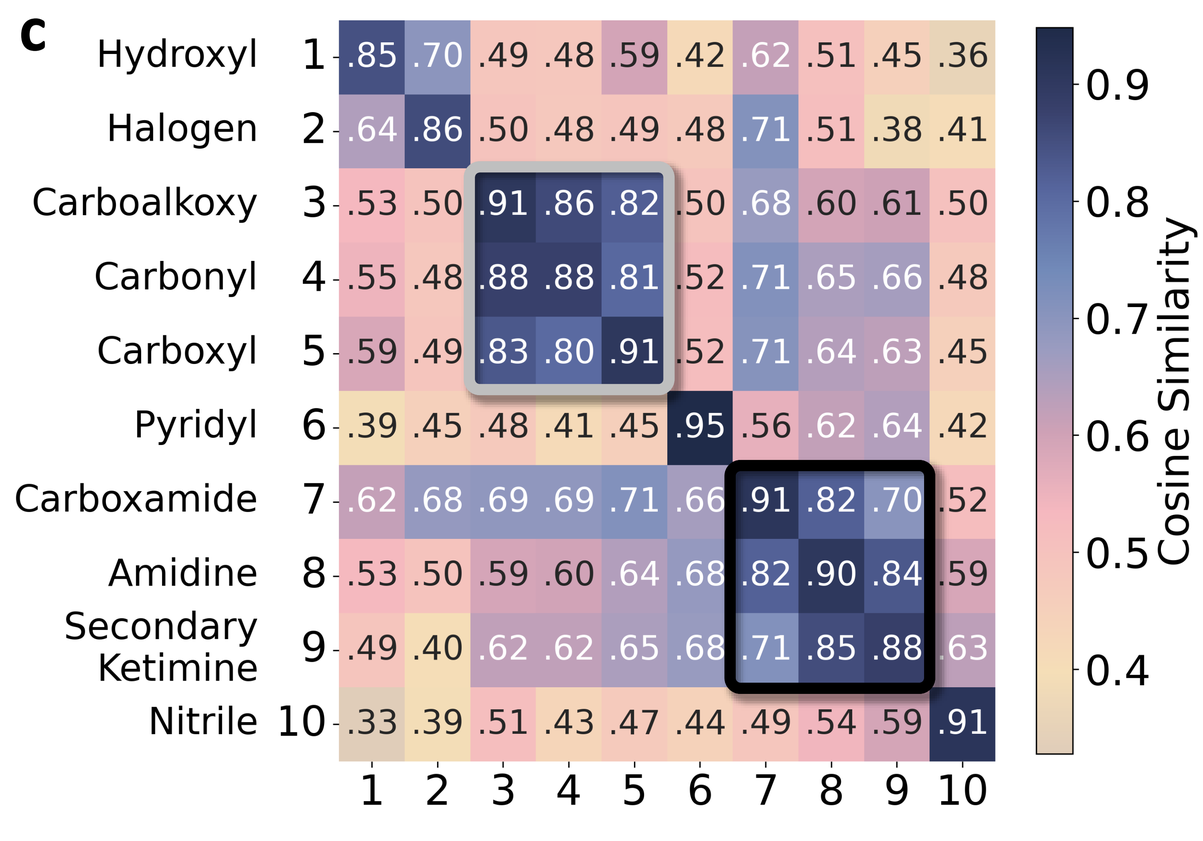
DB index: 5.42 | DB index: 2.73 | DB index: 2.54 | DB index: 1.25

ARTICLE IN PRESS

b

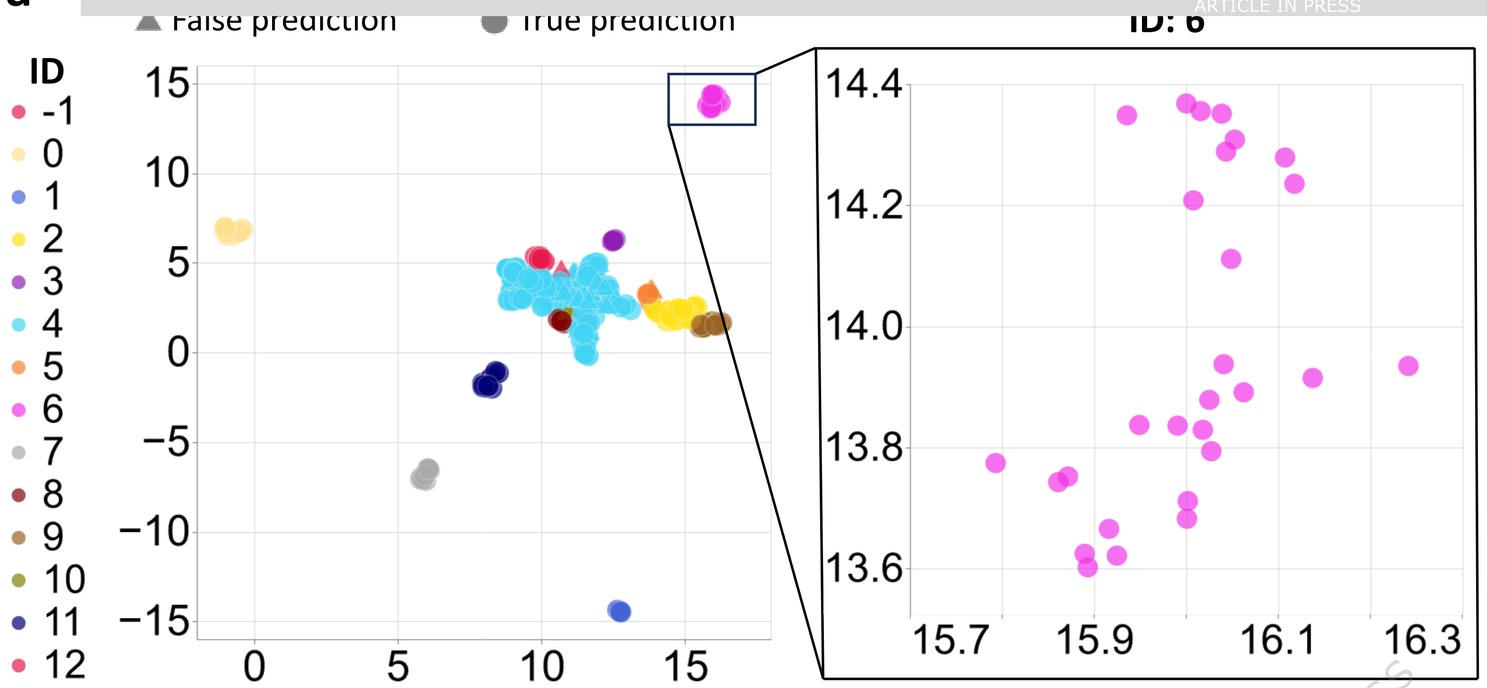


c

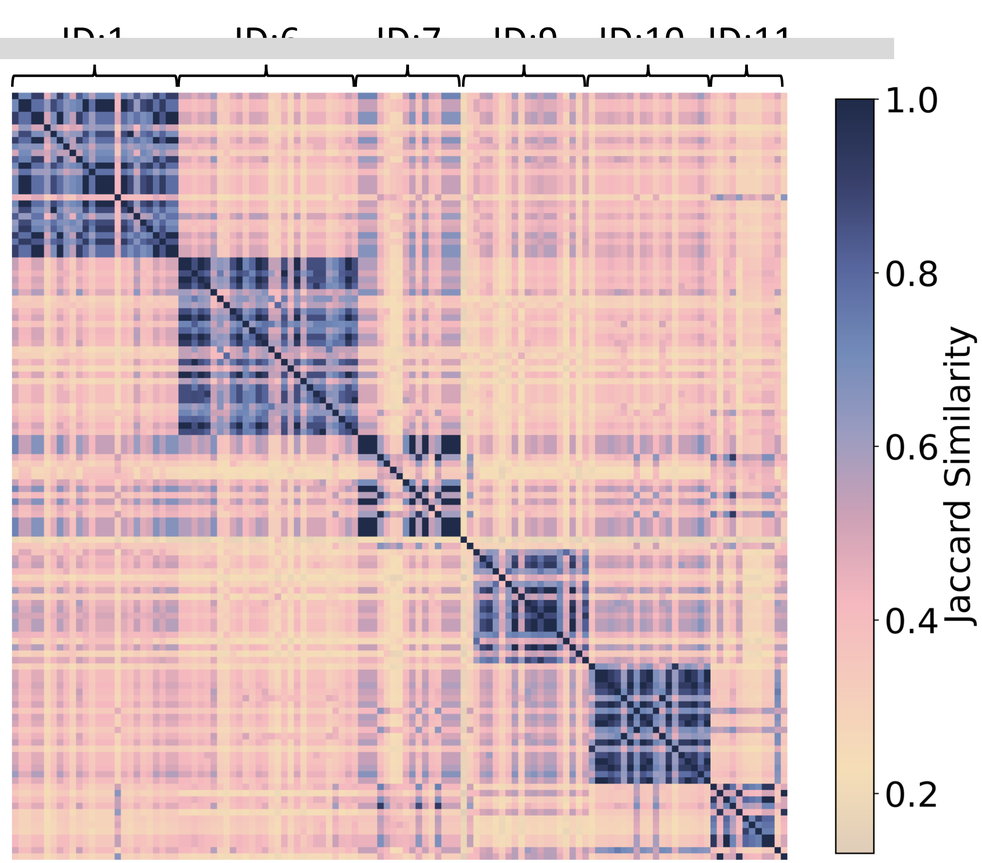


ARTICLE IN PRESS

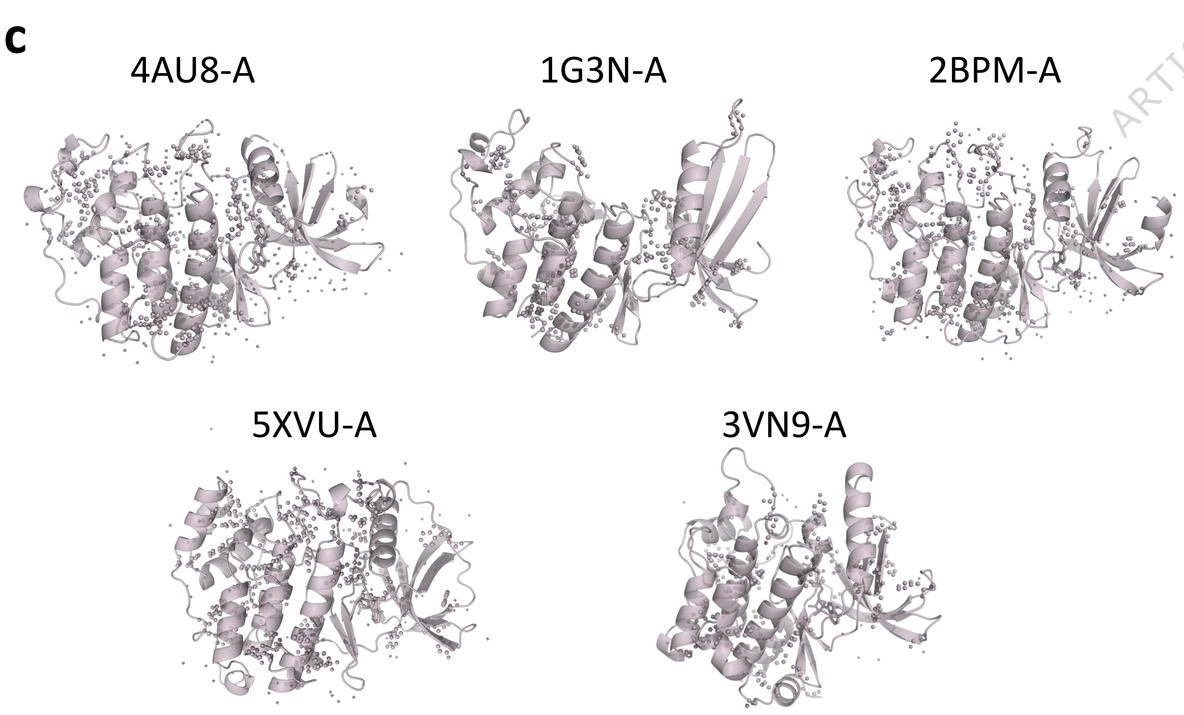
a



b

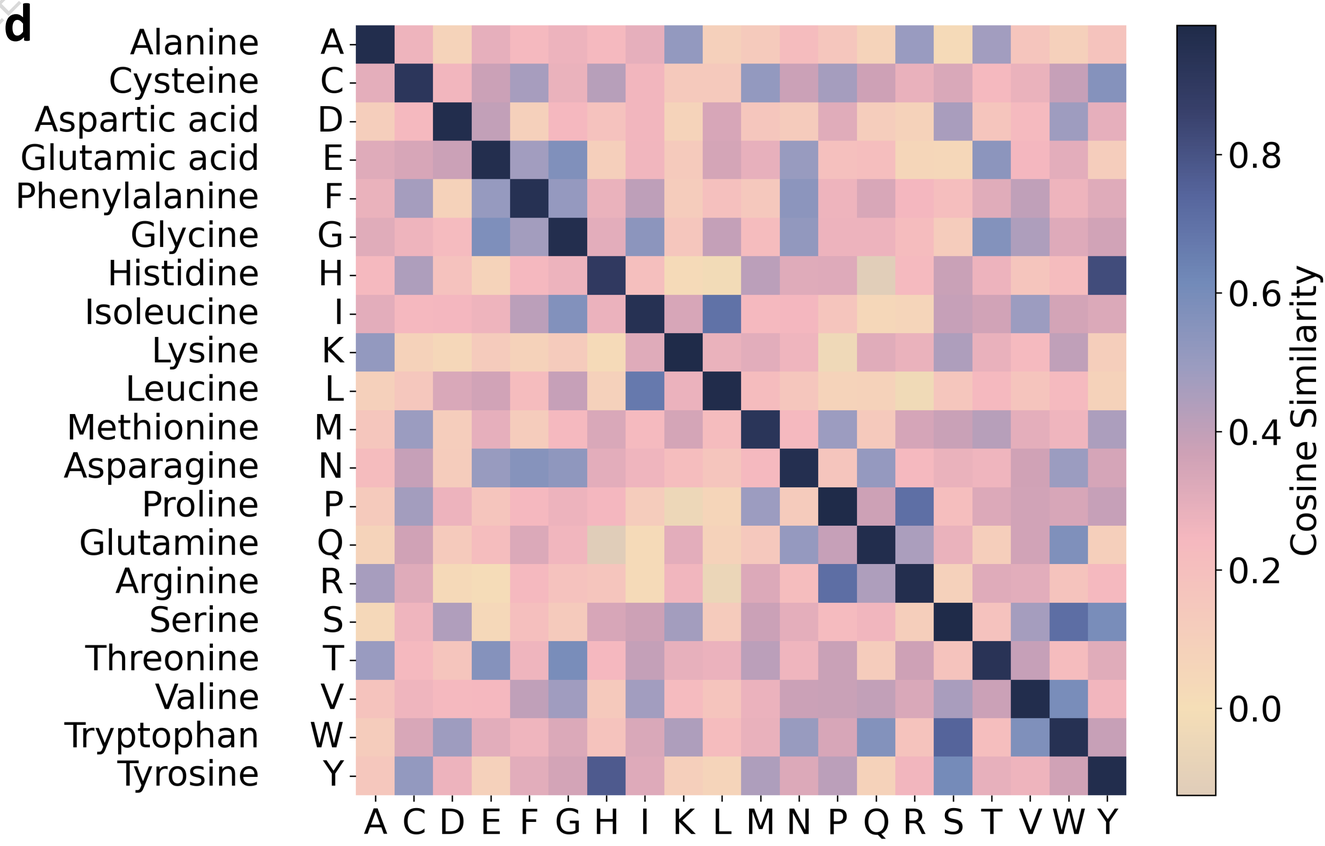


c



Common GO terms for cluster ID 3 include: protein kinase activity, kinase activity, and protein tyrosine kinase activity.

d

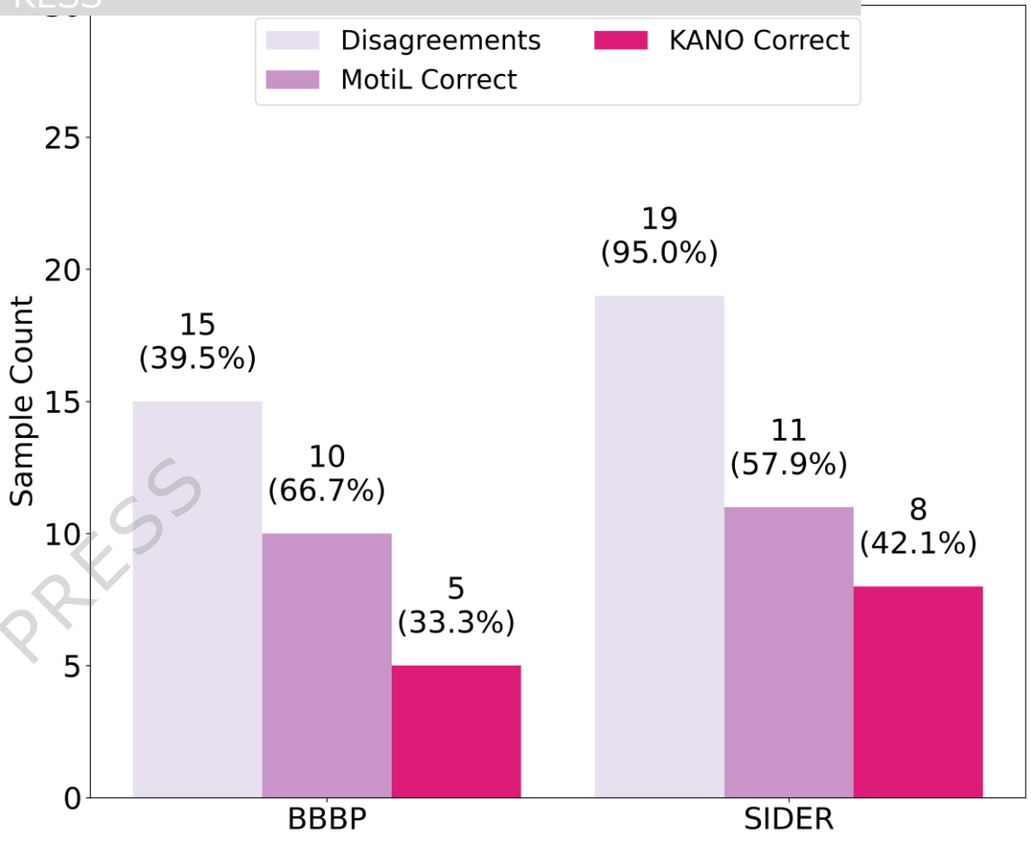


ARTICLE IN PRESS

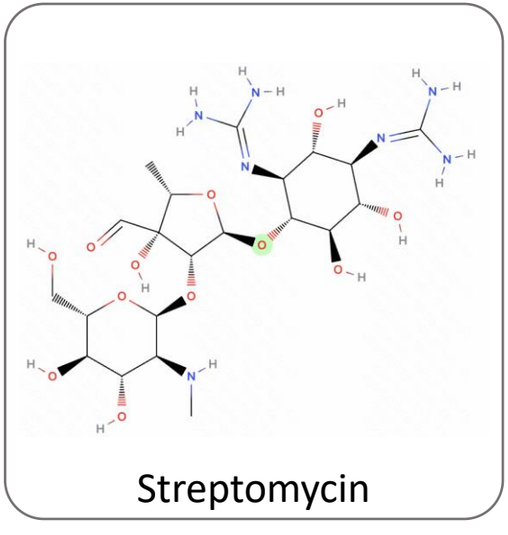
a Prediction statistics on the full datasets



b Prediction statistics on activity cliff samples

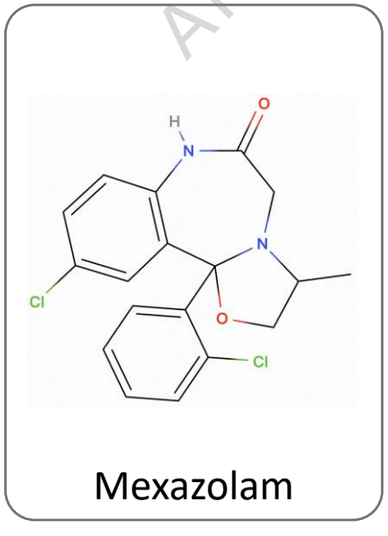


c Cases in BBBP (blood-brain barrier permeability)



Streptomycin

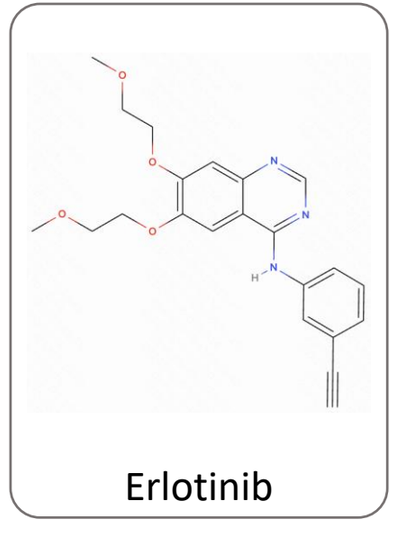
KANO: 1 ✗
MotiL: 0 ✓



Mexazolam

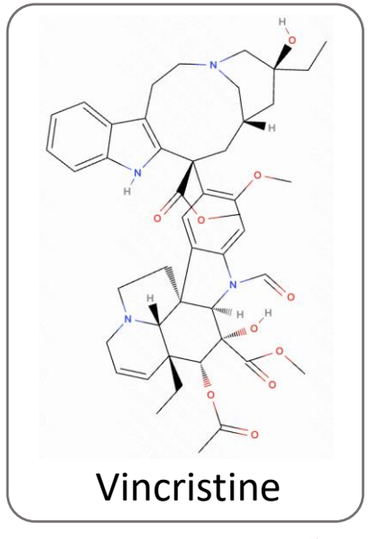
KANO: 0 ✗
MotiL: 1 ✓

d Cases in SIDER (hepatotoxicity)



Erlotinib

KANO: 0 ✗
MotiL: 1 ✓



Vincristine

KANO: 0 ✗
MotiL: 1 ✓