nature portfolio

Peer Review File

Molecular Motif Learning as a pretraining objective for molecular property prediction

Corresponding Author: Professor Chaokun Wang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Summary
The authors present a framework, MotiL, for molecular representation learning. The goal of the proposed method is to ensure higher-fidelity modelling of local structures (i.e. motifs) in molecules as well as richer global representations. To achieve this, MotiL consists of three training stages: warmup through employing an edge-based based denoising diffusion model, bi-scaled contrastive training, and finetuning on supervised downstream tasks. The method shows consistent (but marginal) improvements over existing approaches and the relevancy to both the domains of small molecules and macromolecules is a nice aspect of the study. The authors perform subsequent interrogations of the model to evaluate the extent to which the learned representations capture biological/chemical semantics.

Major

For the MoleculeNet-based benchmark datasets, do the authors perform any further processing or standardisation? At this point, it is well known that these datasets contain erroneous labelling and duplicate examples as well as non-standardised structures [1]. I think benchmarking on these types of datasets is fine (if not super meaningful) for a ML conference-type publication but hinders the case for publication in a journal with a broader audience. Furthermore, these benchmarks are in general quite uninformative for real-world applications. Have the authors considered benchmarking on more application-focussed datasets such as [2]?

Overall, I believe the impact is limited. While the proposed method does consistently improve on previous methods, there is only a marginal improvement over KANO in the small molecule tasks. And the benefit on the macromolecular tasks is less clear when compared to pre-trained structural encoders. Taken together with the weakly motivated benchmarks I think this incremental performance improvement does not warrant publication in a journal with such broader readership. As it stands, I think there is little for non-ML researchers to take from this paper and so targeting a venue with a more ML-focussed readership would be more suitable for this particular work.

Minor:
* L45 (+L89) I'd rephrase the use of the word "generate" here to avoid conflation with generative modelling.
* The authors state current methods struggle to effectively capture representations of local motifs. Is there evidence to support this claim?
* Contrastive learning is described twice in the introduction. I would recommend tightening the prose here a little.
* The introduction implicitly assumes structure-based GNN encoders are the only line of research here by not considering alternate representation such as sequence or SMILES that do not require GNNs. These methods also often enable access to much larger quantities of data for training and pre-training which is a trade-off that should be mentioned.
* L136 "Gaussian noise into G in a controlled manner." This is a shallow description. The authors should describe that the adjacency matrix is noised.
* L148. Grammar. "the local feature" -> the local features.
* L149. Typo "encodes" -> encoders
* L245-250 this is quite repetitive of what has been previously stated.
* In Table 2 the authors are comparing to the variants of GearNet without pre-training. This seems like an unfair comparison.

Additionally, the authors should denote the source of the results for the baseline methods if they did not run them themselves.
* L282 EC prediction is a classification task.
* L309. Rephrase messy clustering.
* The authors make frequent to reference to their proposed methods robustness to noise but do not perform any experiments to demonstrate this. Nor do they the describe the source of the noise they are referring to. Surely, given that SE(3) invariance is built into the model, the effect of structural noise is likely to be minimal? As far as I can tell from the description, coordinates are used in to model.
* What is the justification for using t-SNE for the small molecule embeddings and UMAP for the protein embeddings?
* In figure 3a it is a little odd to start the cluster IDs from -1.
* What do the green points along the X axis in the inset panel in Figure 3a indicate?
* I actually disagree with the authors claim that the similarity of the representations of the amino acids reflect their similarity outside of the authors cherry-picked examples. For example, Serine and Tryptophan and the low similarity of Alanine and Glycine.
* L458 "Fig 4." is repeated.
* Figure 4 is a little hard to gain significant insight from. Whilst it is nice to see specific examples where the models disagree, cherry picking favourable cases is second-best to a more quantitative summarisation of the disagreements. I believe such an analysis should replace this figure, and figure 4 could be an inset panel or moved to the supplementary.
* The authors should specify a license for the source code.
* Supplementary L68: I think some evidence (or elaboration) needs to be provided for the claim that "the negative effects of molecular graph augmentation have hindered the advancement of MCL".
* "L120. To the best of our knowledge, we are the first to propose the concept of molecular motif learning.". Could the authors please elaborate on this and clarify what exactly is being claimed? I could find a few works that purport to address this to some degree:

[1] https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html
[2] [https://doi.org/10.1021/acs.jcim.3c00160]https://doi.org/10.1021/acs.jcim.3c00160
[3] https://arxiv.org/abs/2012.12533
[4] https://arxiv.org/abs/2012.12533

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The work developed by Ziyang Liu and co-workers focuses on a new representation learning method, MotiL, which leverages graphs and motifs for representation learning. I agree with the authors that property prediction using ML can significantly advance drug discovery by reducing the need for experimentally demanding wet-lab methods.
The article is well-written, and the description of the algorithm is clear and insightful. The authors provide the necessary information for a general audience. I find it particularly important that both small-molecule and biomolecular benchmarking predictions were included, demonstrating the model's flexibility. However, I struggled to fully agree that this new method stands out in terms of predictive performance compared to existing methods. Nonetheless, I believe this article makes a new and valuable contribution to the field by introducing motif learning. Below are some points for consideration:
1. It would be helpful to clarify why the authors chose MoleculeNet.
2. It was unclear from the main text that the metrics obtained for other methods in Tables 1 and 2 were not calculated by the authors but rather sourced from previous studies ([12], [27], [24]). Since the metrics for other models were taken from past articles, how can the authors guarantee that the comparison is fair and not influenced by differences in training and evaluation methods? Is the cross-validation method identical? Additionally, this approach limits further comparison. Also, it might be harder to answer why RMSE was used to evaluate physical chemistry properties and MAE for quantum properties and why more metrics for regression and classification weren't explored?
3. In the SI, it is mentioned that the standard deviations (std) were calculated from 3 independent runs. Could the authors clarify what this means? Given the std values, is it possible to perform statistical analyses? Since the paper claims that MotiL surpasses other models, is this improvement statistically significant?
4. The study in SI section F, which evaluates the different components of MotiL, is interesting and important. I agree that the results suggest that the full MotiL architecture performs better, especially in regression tasks. It would be beneficial, if possible, to add a sentence to the main text summarizing the key conclusions from these ablation studies.
5. I find Figure 2a very insightful, as it clearly shows that MotiL can distinguish different molecular scaffolds. However, I suggest using more contrastive colors, particularly in the BBBP plots, where some shades (e.g., pinks) are too similar. Additionally, what was the rationale behind choosing the specific molecules, scaffolds, and database for Figure 2a?
6. In Figure 2b, and especially in Figure 3 in the SI, some dashed lines are difficult to see.
7. In not sure if I fully understand the data behind Figure 3a. UMAP and DBSCAN were used to project the representations learned by MotiL for ATP-binding proteins. Why was this specific class chosen? Also, does the shape in the figure indicate whether a prediction is correct in identifying an ATP-binding protein?
8. Figure 3c could be better organized. I prefer the layout used in Figure 5 SI, and I think more GO terms should be written to highlight similarities. For example, in ID 3, the common GO terms, protein kinase activity, Kinase activity, and Protein tyrosine kinase activity, could be included.
9. The authors also make a more in-depth comparison between the 2nd best algorithm (KANO) and MotiL, but I struggle to understand the significance of this section. In fact, improvements between KANO and MotiL appear to be marginal. While I

acknowledge that MotiL correctly predicts BBB permeability and toxicity for key molecules while KANO does not, I find this section unnecessarily long and less relevant. I believe a more challenging comparison could be beneficial. For example, evaluating performance on activity cliffs.
10. I appreciate the author's commitment to improving the model in future work, with already some defined future steps that could further accelerate drug discovery.
11. Finally, I recommend revising the references. For example, reference [1], Diaz, N. et al. Discovery of potent small molecule inhibitors of lipoprotein (a) formation. Nature 1–6 (2024), it's written on the journal's website as Diaz, N. et al. Discovery of potent small-molecule inhibitors of lipoprotein(a) formation. Nature 629, 945–950 (2024). Reference [7] Wan, Z. et al. Unconventional superconductivity in chiral molecule-tas2 hybrid superlattices. Nature 1–6 (2024) it's written on the journal's website as Wan, Z, et al. Unconventional superconductivity in chiral molecule–TaS2 hybrid superlattices. Nature 632, 69–74 (2024).
Given the increasing importance of ML in advancing drug discovery and the innovative methodology presented in this work, I recommend this article be accepted for revision.

(Remarks on code availability)
- The library versions used are indicated
- Databases used for small molecules are available on git-hub. However, for the MUV, QM7, and ToxCast, the #molecules written on the article don't appear to match the CSV file lengths.
- I couldn't access the databases for macromolecules, I was unable to access the Baidu Cloud.
- The pretrained models for small molecule predictions are available on github providing a readme file with a clear description of MotiL, requirements.txt, and instructions on how to run MotiL
- I couldn't access the pretrained models for macromolecule predictions, I was unable to access the Baidu Cloud
- I used codeocean website to run the available code and the example of the bbbp database led to the same output file (Results: 0.97404 +/- 0.00289). It was reproducible, but was it supposed to match Table 1?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors diligence in addressing the great majority of my concerns. Overall, I am satisfied with the the technical quality of the work. My concerns with the impact & relevance to the readership of the journal remain though I, of course, defer to the judgment of the editors.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed my suggestions.
Particularly:
- They clarified their commitment to a fair comparison between methods. It is now clear how the experimental setup enables a fair model comparison.
- I appreciate how they made the colors in Fig. 2 more accessible. However, I am unsure whether it needs to be mentioned in the main text ("where the color palette has been optimized to enhance visual contrast and interpretability", lines 330–332). After this change, I would also suggest harmonizing the colors used throughout the manuscript.
- I believe the authors have more clearly highlighted the strengths of MotiL over KANO.
- I appreciate the authors' commitment to continuing to work on the algorithms in future steps, particularly in molecule generation.
As before, I recommend this article for publication.

(Remarks on code availability)
I reviewed the code before, and they addressed my comments. Particularly, I now have access to the macromolecule benchmark datasets and pretrained models.

Dear Reviewers,

We sincerely thank the reviewers for the thorough evaluation of our manuscript and for recognizing the significance of our work. We are also grateful for the reviewers' constructive comments and insightful suggestions, which have helped us identify areas for improvement and significantly enhanced the quality of our manuscript. In the revised manuscript, we have carefully addressed all comments and highlighted the corresponding changes in blue font for clarity. In this response letter, we provide a point-by-point reply to each of the comments from Reviewer #1 (pages 1 to 24) and Reviewer #2 (pages 25 to 45), respectively.

Response to Reviewer #1:

Comment 1: For the MoleculeNet-based benchmark datasets, do the authors perform any further processing or standardization? At this point, it is well known that these datasets contain erroneous labelling and duplicate examples as well as non-standardized structures [1]. I think benchmarking on these types of datasets is fine (if not super meaningful) for a ML conference-type publication but hinders the case for publication in a journal with a broader audience. Furthermore, these benchmarks are in general quite uninformative for real-world applications. Have the authors considered benchmarking on more application-focused datasets such as [2]? Overall, I believe the impact is limited. While the proposed method does consistently improve on previous methods, there is only a marginal improvement over KANO in the small molecule tasks. And the benefit on the macromolecular tasks is less clear when compared to pre-trained structural encoders. Taken together with the weakly motivated benchmarks I think this incremental performance improvement does not warrant publication in a journal with such broader readership. As it stands, I think there is little for non-ML researchers to take from this paper and so targeting a venue with amore ML-focused readership would be more suitable for this particular work.

[1] https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html
[2] https://doi.org/10.1021/acs.jcim.3c00160

Reply:
We sincerely thank the reviewer for this constructive and detailed feedback. We have carefully addressed each aspect of the comment and have made significant clarifications and additions in the revised manuscript, summarized below:

1. **On the use of MoleculeNet-based benchmark datasets**: We acknowledge the reviewer's concern regarding the known limitations of MoleculeNet datasets, such as erroneous labels, duplicated entries, and lack of molecular standardization [1]. As with many previous studies (e.g., GROVER, MolCLR, KANO), our intention in using MoleculeNet is to ensure fair and directly comparable benchmarking under widely recognized settings. These benchmarks, while imperfect, continue to serve as a de facto standard in the field of molecular machine learning, offering controlled comparisons across methods. We emphasize that we did not apply any additional processing or standardization to these datasets beyond the standard splits provided, in order to strictly follow the established evaluation protocols used by prior work. This design choice facilitates meaningful comparisons with strong baselines and highlights the incremental benefits offered by our method within this common framework.
2. **On evaluating with more application-oriented datasets**: We appreciate the reviewer's excellent suggestion to evaluate MotiL on more realistic, application-focused datasets. To this end, we have conducted new experiments on a benchmark introduced by [2], which focuses on absorption, distribution, metabolism, and excretion (ADME) properties that are central to real-world drug development and pharmacokinetics. There are

six ADME-related prediction tasks in this dataset: HLM clearance, human plasma protein binding (hPPB), MDR1-MDCK efflux ratio, rat liver microsome (RLM) stability, rat plasma protein binding (rPPB), and aqueous solubility. These properties are clinically relevant and span diverse pharmacological mechanisms. We compare our method against strong baselines (CMPNN, MGSSL, MolCLR, KANO), and the RMSE results under three independent runs using three random-seeded are summarized in Table 1 (lower RMSE value indicates better performance):

Table 1: Prediction result comparison of different methods on ADME datasets.

ADME datasets	**CMPNN**	**MGSSL**	**MolCLR**	**KANO**	**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

The results demonstrate that MotiL achieves consistently stable and competitive performance across all six ADME-related prediction tasks. On most tasks, MotiL not only achieves strong predictive performance but also maintains low standard deviations (e.g., ±0.0054 for HLM and ±0.0047 for Solubility), indicating its robustness and reproducibility. In Table 1, we use bold font to denote the best performance on each task and italic font to indicate the second-best. 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. Overall, MotiL consistently ranks among the best-performing methods with minimal fluctuations across diverse pharmacokinetic properties, highlighting its strong generalization ability. Moreover, MotiL handles these tasks within a unified framework without the need for task-specific tuning, demonstrating its practical utility for real-world drug discovery applications.

3. **On the perceived marginal improvement and general impact**:

(1) Statistically significant improvements over KANO: While the numerical differences between MotiL and KANO on small-molecule tasks may seem modest in some cases, we emphasize that these improvements are statistically significant and consistently reproducible. To rigorously verify it, we conducted paired two-tailed t-tests across all 14 MoleculeNet benchmark datasets comparing MotiL with the strongest baseline, KANO. The results, presented in Table 2 and Table 3 (also included as Supplementary Tables 3 and 4), show that **all p-values are below 0.05**, confirming that the improvements are **statistically significant** and unlikely to result from random fluctuations. Moreover, the standard deviations across multiple runs are consistently low (typically <1.5%), indicating that the performance is stable across different random seeds and initializations. Therefore, the observed gains, though sometimes small in absolute value, are consistent, statistically valid, and reproducible, reflecting the robustness of our approach.

(2) Addressing both micro- and macromolecular property prediction: Beyond numerical gains, the key contribution of MotiL is **its unified framework that supports both small-molecule and macromolecule property prediction tasks**, which most prior methods cannot do. For example, models such as KANO and MolCLR are designed specifically for small-molecule graphs and do not generalize to macromolecular graphs (e.g., peptides, nucleotides). Some macromolecular methods **rely on pre-trained protein language models or 3D structural encoders**. However, these methods require expensive and curated protein structures; cannot handle small-molecule graphs effectively; lack a unified representation space for both molecular types. In contrast, MotiL jointly pre-trains on small-molecule and macromolecular graphs using both 1D sequences and 3D geometric information, enabling it to generalize across diverse chemical spaces without task-specific tuning. This versatility meets real-world drug discovery needs, supporting both small-molecule ADME property

prediction and macromolecular interaction modeling, even when structural data are incomplete.

Table 2: ROC-AUC (%) performance comparison on small molecules. The testing results are presented as the mean and standard deviation, calculated from three independent runs. The best result for each dataset is highlighted in bold, and the runner-up result is italicized. All baseline results are directly taken from Ref. [23], which follows the same evaluation protocol as this study: three random-seeded scaffold splits for small molecules.

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

Table 3: Performance comparison on small molecules. We use the metric of RMSE or MAE where a lower value indicates better performance. The testing results are presented as the mean and standard deviation, calculated from three independent runs. The best result for each dataset is highlighted in bold, and the runner-up result is italicized. All baseline results are directly taken from Ref. [23], which follows the same evaluation protocol as this study: three random-seeded scaffold splits for small molecules.

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

[1] https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html
[2] https://doi.org/10.1021/acs.jcim.3c00160

To better reflect these distinctions, we have revised the manuscript (Supplementary Section G) to include: (1) a more explicit discussion of the limitations of MoleculeNet and motivation for ADME evaluation; (2) an expanded section comparing real-world ADME performance; (3) additional clarity in the scope, positioning, and impact of our work across application contexts.

We have also incorporated the above content into Supplementary Section I (Expanded discussion on

benchmark limitations and real-world ADME evaluation) as follows:

Section 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, 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."

Section 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."

Section 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 Table 5. 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 KANO achieves relatively good results on certain tasks, its results often come with larger variances (e.g., ±0.0225 for HLM and ±0.0588 for Solubility), suggesting potential inconsistency. In contrast, MotiL exhibits low standard deviations across most 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 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."

Section 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."

We have also added the following experimental conclusion into the Section of "MotiL boosts the performance of molecular property prediction" in the main text:

"Our experiments on absorption, distribution, metabolism, and excretion (ADME) prediction tasks

(Supplementary Table 5) 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."

The corresponding statistical analysis on the MoleculeNet benchmark dataset has been included in the revised manuscript under Supplementary Section D (Complete performance comparison on small molecules), as presented below:

"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 Table 3 and Table 4. All values are below the 0.05 threshold, confirming that MotiL's improvements are statistically significant and robust across different datasets."

Comment 2: L45 (+L89) I'd rephrase the use of the word "generate" here to avoid conflation with generative modelling.

Reply:

We thank the reviewer for highlighting the potential ambiguity caused by our use of the word "generate" in the context of molecular and protein representation learning (Lines 45 and 89). We agree that the term may inadvertently suggest a connection to generative modeling, which is not the intended meaning in our work.

To avoid this conflation and to align with standard terminology in the field, we have revised the manuscript to replace "generate" with "produce" where appropriate. Specifically, we made the following changes:

1. Line 45: *"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."*
2. Line 89: *"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)."*

In addition, we proactively revised several other instances of the term "generate" to "produce" to ensure consistent and unambiguous language throughout the manuscript. Some examples are as follows:

1. Line 49: *"It is crucial for molecular representation learning to produce representations that accurately reflect the underlying chemical structures and functions."*
2. Caption of Figure 1: *"The pretrained GNN encoders produce representations for unseen molecules, which are then fed into a multilayer perceptron classifier to learn the relationship between these graph representations and the true molecular property labels."*

Comment 3: The authors state current methods struggle to effectively capture representations of local motifs. Is there evidence to support this claim?

Reply:

We thank the reviewer for raising this important question regarding our claim that current methods struggle to effectively capture representations of local motifs. Our motivation for this claim is based on both observations from prior studies and experimental evidence in our work:

1. **Prior work acknowledges limitations in capturing local motifs:** Several recent studies have pointed out that many GNN-based molecular representation models, particularly those relying on global pooling or contrastive strategies over entire graphs, tend to underemphasize fine-grained local substructures. For instance:

   (1) GROVER [1] primarily uses contrastive objectives over whole-molecule views, without explicit

modeling of localized functional groups or motifs.

(2) MolCLR [2] applies contrastive learning on augmented molecular views but does not explicitly model or preserve chemically meaningful local motifs.

(3) KANO [3], while improving representation through domain knowledge, still does not include an explicit mechanism for local motif identification or encoding.

These approaches often rely on the model to implicitly learn local patterns, which may be insufficient for tasks where specific substructures (e.g., pharmacophores, binding pockets) are strongly tied to functional outcomes. To address the reviewer's concern, we have revised the manuscript to more carefully justify this statement (the current methods struggle to effectively capture representations of local motifs) and added necessary citations:

"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 [11, 12, 27] often rely on random subgraph sampling or global-level supervision, which may overlook pharmacologically or structurally meaningful motifs."

2. **Evidence from our experiments**: To provide empirical evidence supporting our claim that current methods struggle to effectively capture representations of local motifs, we conducted activity cliff experiments as part of our evaluation. Activity cliffs refer to pairs of molecules that are highly similar in overall structure (global scaffold), yet exhibit large differences in biological activity due to small variations in local motifs. This experimental setup provides a direct way to test whether a model can capture the impact of subtle local changes on molecular properties. We designed a dedicated activity cliff evaluation protocol as follows:

(1) We first identified molecular pairs from the BBBP and SIDER datasets that exhibit high structural similarity (Tanimoto similarity > 0.85) but substantial differences in biological activity labels.

(2) We then extracted the individual molecules from these pairs to construct the activity cliff test sets, focusing on cases where subtle local differences drive large activity changes.

(3) Finally, for each dataset, we analyzed the prediction disagreements between KANO and MotiL on these challenging examples to assess their ability to capture local motif variations.

As shown in Figure 1, we found that KANO, the best-performing baseline in our main results, exhibited substantial performance degradation on activity cliff molecule pairs. It indicates that **KANO fails to capture the subtle yet critical differences arising from local motifs**. These experimental results provide concrete evidence that current methods, exemplified by KANO, have limitations in learning local motif representations. On the BBBP dataset, among 15 activity cliff disagreement cases, MotiL correctly predicted 10 pairs (66.7%), while KANO only predicted 5 correctly (33.3%). On the SIDER dataset, among 19 disagreement cases, MotiL correctly predicted 11 pairs (57.9%), compared to 8 correct predictions (42.1%) by KANO.

Moreover, Figures 2 and 3 provide concrete examples of activity cliff pairs. These examples include pairs with Tanimoto similarity between 0.85 and 1.0, where only minor differences in functional groups exist. In these examples, **KANO fails to distinguish the structurally similar molecules with distinct activities**, whereas MotiL successfully captures subtle motif-level differences in most cases. In the BBBP dataset (Figure 2), we observe that in the first three molecular pairs, MotiL successfully predicts the correct labels for both molecules in each pair, while KANO misclassifies one molecule per pair. These cases highlight MotiL's stronger sensitivity to local functional group variations, such as substitutions between carboxyl and ester groups, or subtle stereochemical differences on steroid-like scaffolds (e.g., a pair with a Tanimoto similarity of 0.851). In contrast, in the last two pairs, both models incorrectly classify one molecule per pair, indicating that certain fine-grained structural changes still present significant challenges, even for MotiL, despite its overall superior

performance on activity cliff prediction. In the SIDER dataset (Figure 3), several activity cliff pairs with high structural similarity (e.g., 0.956 and 1.0) but opposite toxicity labels are presented. In the first pair (similarity = 0.956), MotiL correctly distinguishes between the two molecules, while KANO produces identical predictions, failing to capture the subtle discriminative features. However, in the remaining three pairs, both models exhibit mixed outcomes: for each pair, one molecule is misclassified by KANO and the other by MotiL. These examples suggest that **both models still face difficulties in fully resolving the subtle chemical cues that define the activity cliff boundary in complex toxicity scenarios**. These results highlight that **MotiL is more robust to subtle substructural changes than KANO**, which are often the root cause of activity cliffs. We attribute this improvement to MotiL's motif-level representation learning framework and the diffusion-based denoising mechanism, which enhance the model's ability to focus on local structural variations that drive biological activity.

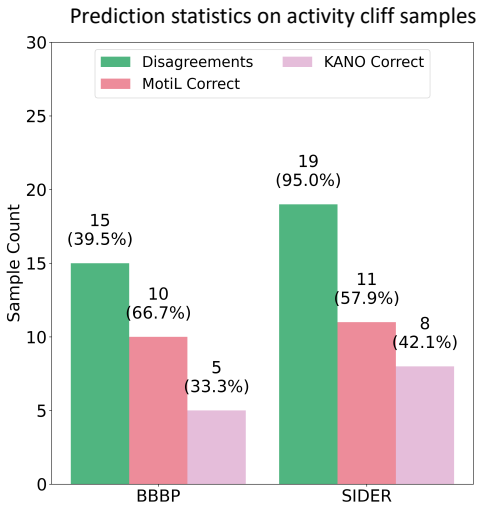


Figure 1: Statistics on disagreements between KANO and MotiL on activity cliff samples.

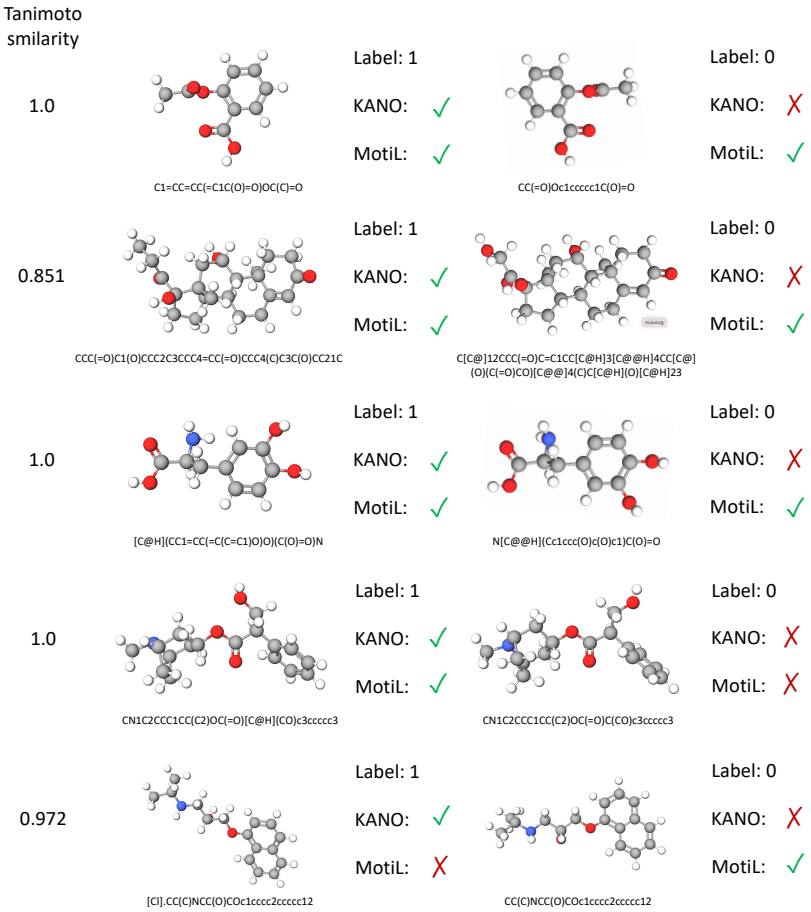


Figure 2: Prediction results for activity cliff samples on BBBP.

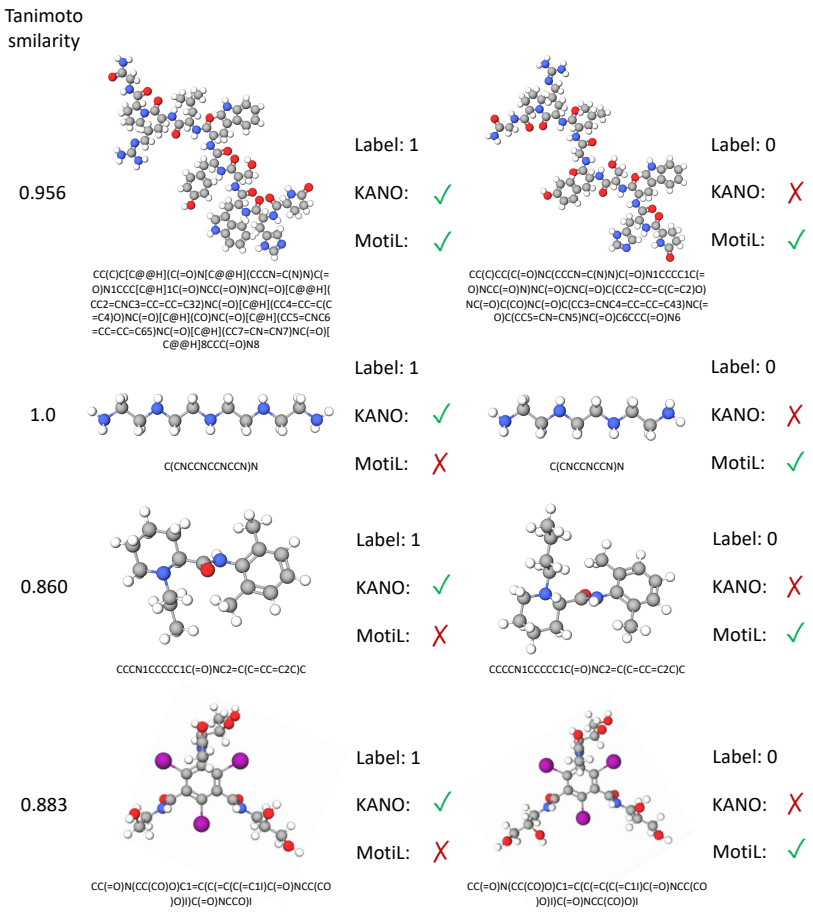


Figure 3: Prediction results for activity cliff samples on SIDER.

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

[2] 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).

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

Comment 4: Contrastive learning is described twice in the introduction. I would recommend tightening the prose here a little.

Reply:

We thank the reviewer for pointing out the redundancy in our introduction regarding the description of contrastive learning. We agree that the prose can be made more concise and have accordingly revised the relevant section to improve clarity and reduce repetition.

Originally, contrastive learning was described both when introducing the main categories of molecular representation learning methods and again when discussing its impact on motif representation. To improve clarity and reduce redundancy, we have refined the exposition. The third paragraph in the Section of Introduction now concisely introduces the contrastive learning mechanism, while the subsequent discussion focuses exclusively on its limitations, specifically its challenges in capturing the representations of chemically meaningful motifs, without reiterating the methodology. This restructuring helps streamline the introduction and emphasizes the motivation

behind our proposed approach.

The revised portion in the Section of Introduction of the main text now reads:

"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 [27–30], 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 [11, 12, 27] 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."

Comment 5: The introduction implicitly assumes structure-based GNN encoders are the only line of research here by not considering alternate representation such as sequence or SMILES that do not require GNNs. These methods also often enable access to much larger quantities of data for training and pre-training which is a trade-off that should be mentioned.

Reply:

We thank the reviewer for this insightful comment regarding alternative molecular representations such as SMILES strings and protein sequences. We agree that these representations constitute an important line of research and offer practical advantages, especially in terms of data availability and scalability. In response, we would like to make the following clarifications:

1. **On the choice of molecular graphs vs. sequences or SMILES**: We acknowledge that SMILES- and sequence-based approaches (e.g., ChemBERTa [1], SMILES-BERT [2], ESM [3]) enable access to massive corpora of molecular or protein data and have shown promising performance in various domains. However, our focus on graph-based representations is motivated by their closer alignment with the underlying physical and chemical structure of molecules. Molecular graphs preserve explicit atom- and bond-level information, spatial topology, and connectivity constraints, which are properties often distorted or omitted in linear representations. In this sense, graph representations can be viewed as a more complete and chemically faithful form of sequence or SMILES data, which makes them particularly suitable for tasks that depend on local substructure integrity, such as motif learning or binding affinity prediction.
2. **Trade-offs between data accessibility and structural fidelity**: We agree with the reviewer that there exists a fundamental trade-off between data scale and structural richness. SMILES and sequence formats are often easier to collect in large quantities from public databases, and thus allow for training models at scale. Graph-based models, while more expressive in encoding chemistry, often suffer from more limited data resources due

to the need for preprocessed structural annotations. We have now explicitly acknowledged this trade-off in the revised introduction to present a more balanced view of current representation learning paradigms. While sequence- and SMILES-based models benefit from greater data availability and computational efficiency, graph-based approaches retain richer chemical semantics by preserving the topological and bonding structures of molecules. In this work, we focus on graph representations as they offer a more direct and faithful encoding of molecular motifs, which are often lost in linearized formats.

3. **Addition of discussion in both the main text and supplementary material**: To better address the reviewer's point and improve the accessibility of our paper for broader audiences, we have made two specific additions: In the section of Introduction, we now briefly mention representative works based on SMILES or sequences (e.g., ChemBERTa [1], SMILES-BERT [2], ESM [3]), highlighting their scalability and complementary role in molecular representation learning:

   "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."

   Furthermore, we have added a dedicated subsection in the appendix (i.e., Supplementary Section A.4, Sequence-based and SMILES-based molecular representation learning) that summarizes prominent SMILES- and sequence-based approaches. This section discusses their relative advantages, limitations, and how they compare to graph-based models in terms of expressiveness, data availability, and applicability to motif-level 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 sequence-based or SMILES -based approaches is their scalability: the simplicity of SMILES and FASTA formats allows access to massive datasets and efficient pretraining pipelines. However, this scalability comes with trade-offs. Unlike molecular graphs, which naturally encode topological and stereochemical information, SMILES strings or amino acid sequences 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."

   Together, these additions clarify the landscape of molecular representation learning and provide context for our graph-based choice, while acknowledging the strengths of alternative input modalities.

[1] Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. NeurIPS Workshop on Machine Learning for Molecules (2020).

[2] 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 429–436 (2019).

[3] 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).

Comment 6: L136 "Gaussian noise into $\mathcal{G}$ in a controlled manner." This is a shallow description. The authors should describe that the adjacency matrix is noised.

Reply:

We thank the reviewer for this valuable suggestion. Based on your suggestions, we have revised the text in the Section of Methods to explicitly state that the noise is added to the adjacency matrix of the molecular graph. The revised sentence now reads:

"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 $\boldsymbol{A}_t$ at each time step t."

Meanwhile, we have revised the content in the Section of Methods as follows:

"The generative diffusion model DiffMoM progressively perturbs the adjacency matrix $\boldsymbol{A}_t$ of the molecular graph $\mathcal{G}$ by adding Gaussian noise, where the level of noise is controlled through a predefined diffusion schedule."

This clarification more accurately reflects the operation performed in the diffusion process and improves the technical precision of our description.

Comment 7: L148. Grammar. "the local feature" -> the local features.

Reply:

We thank the reviewer for their attentive reading and constructive feedback regarding the grammatical issue in Line 148. We have revised the phrase *"we integrate the local feature in a molecule..."* to *"we integrate the local features in a molecule..."* to correct the singular/plural mismatch. In addition, we carefully checked the manuscript for other inaccurate expressions and made further refinements to improve clarity. For example, since Figure 2 presents visualization results of multiple methods, not only MotiL, we have revised the caption of Figure 2a from *"T-SNE visualization of the molecular representations learned by MotiL."* to *"T-SNE visualization of the molecular representations learned by MotiL and other methods."*.

Comment 8: L149. Typo "encodes" -> encoders

Reply:

We appreciate the reviewer's careful attention to detail. We have corrected *"two GNN encodes"* to *"two GNN encoders"* in the revised manuscript. In addition, we have thoroughly reviewed the entire manuscript to identify and correct any typographical errors, ensuring consistency and clarity throughout the paper. In addition, we carefully re-examined the main text and the supplementary material to identify and correct other typographical and formatting issues, ensuring consistency and clarity throughout the paper. For example, in the supplementary material (Line 119) of the originally submitted manuscript, we corrected a missing space in the sentence: *"Lipophilicity[59] is extracted"* -> *"Lipophilicity [59] is extracted."*

Comment 9: L245-250 this is quite repetitive of what has been previously stated.

Reply:

We thank the reviewer for pointing out the redundancy in Lines 245–250. Upon review, we agree that this portion of the manuscript reiterates points already discussed in the Introduction and Methodology sections, particularly the limitations of augmentation-based contrastive learning and the motivation behind our motif-preserving strategy.

To address this concern, we have revised the paragraph to reduce repetition while preserving its function as a transition to the performance evaluation. The updated version now concisely summarizes the contrast between baseline approaches and our method, emphasizing key insights without duplicating previous content. The revised text reads:

"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."

Comment 10: In Table 2 the authors are comparing to the variants of GearNet without pre-training. This seems like an unfair comparison. Additionally, the authors should denote the source of the results for the baseline methods if they did not run them themselves.

Reply:

We thank the reviewer for raising this important concern regarding the fairness and clarity of the comparisons presented in Table 2.

1. **On comparing MotiL with non-pretrained GearNet variants**: We agree with the reviewer that differences in pretraining can influence the validity of performance comparisons. In our case, although MotiL employs a pretraining phase, it is important to clarify that no external protein data were introduced during this phase. As described in the Section of Datasets in the main text, *"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."* That is, all pretraining for MotiL was performed strictly using the proteins in the benchmark datasets themselves, without incorporating additional protein sequences or structures from external sources. In contrast, GearNet's pretraining variant explicitly incorporates a large external protein structure corpus for pretraining. As stated in Section 5.1 of the GearNet paper [1]: *"We use the AlphaFold protein structure database (CC-BY 4.0 License) (Varadi et al., 2021) for pretraining."* This use of an extensive external protein structure database distinguishes GearNet's pretraining setup from ours in both scope and data resources. In Table 2 of our manuscript, we chose to report the non-pretrained versions of GearNet and its variants (e.g., GearNet-Edge, GearNet-IEConv, GearNet-Edge-IEConv), following their original evaluation protocols without pretraining. These results were drawn directly from Ref. [24], and as the original paper does not report performance for the same variants with pretraining, we did not include pretrained counterparts to avoid inconsistencies. We now explicitly clarify this setting in the revised manuscript and have added a note to the Table 2 caption indicating this difference: *"The pretraining phase of MotiL does not incorporate extra protein data, which is consistent with baselines such as GearNet."*
2. **On denoting the source of baseline results**: We also appreciate the reviewer's suggestion to make the provenance of baseline results more transparent. As indicated in the Section of Methods and now explicitly clarified in the caption of Table 2, the results for all baseline models (including GearNet and CDConv) are directly reported from Ref. [24]. The updated caption now reads: *"All baseline results in this table are directly reported from Ref. [24], which follows the same evaluation protocol as this study: 10-fold cross-validation splits for proteins."*

   Therefore, the updated caption for Table 2 is as follows:

"Performance comparison on protein macromolecules. All baseline results in this table are directly reported from [24], which follows the same evaluation protocol as this study: 10-fold cross-validation splits for proteins. The pretraining phase of MotiL does not incorporate extra protein data, which is consistent with baselines such as GearNet."

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

Comment 11: L282 EC prediction is a classification task.

Reply:

We sincerely thank the reviewer for pointing out this important clarification regarding the nature of the EC number prediction task in Line 282. You are correct that EC number prediction is a multi-label classification task, not a regression task. Each protein can be associated with multiple enzyme commission (EC) numbers, corresponding to hierarchical functional annotations based on the type of chemical reactions catalyzed. Thus, the task is best modeled as a classification problem, where the goal is to predict one or more correct EC labels for a given protein.

In our original manuscript, the sentence in Line 282 was imprecise and may have inadvertently implied that EC prediction was a regression task. We have corrected the sentence in the revised manuscript to more accurately reflect the task type. The updated version now reads:

"It suggests that MotiL has a strong advantage in classification tasks like EC number prediction."

We have also reviewed the surrounding text in Section 4.4 to ensure that other task types (e.g., ATP binding classification and GO prediction) are correctly described and consistently referred to using appropriate terminology (i.e., classification vs. regression).

Comment 12: L309. Rephrase messy clustering.

Reply:

We sincerely thank the reviewer for pointing out the use of the informal phrase "messy clustering" in Line 309. To improve clarity and rigor, we have revised the sentence to use more appropriate technical terminology that better reflects the clustering behavior observed in the representation space.

Specifically, we have modified the original sentence from: *"No pretraining yields the worst results, with messy clustering on QM8 (DB index: 4.93), highlighting the necessity of pretraining representation learning for effective molecular representation."* to the following revised version: *"No pretraining yields the worst results, with disorganized and poorly separated clusters on QM8 (DB index: 4.93), highlighting the necessity of pretraining representation learning for effective molecular representation."*

This updated wording provides a clearer and more objective description of the low-quality clustering results observed when pretraining is absent. We also ensure that the interpretation remains directly supported by the quantitative DB index value reported.

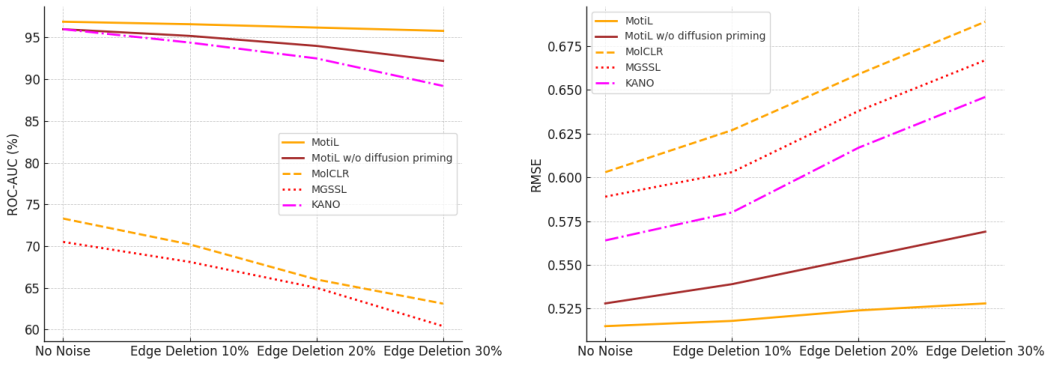
Comment 13: The authors make frequent to reference to their proposed methods robustness to noise but do not perform any experiments to demonstrate this. Nor do they describe the source of the noise they are referring to. Surely, given that SE(3) invariance is built into the model, the effect of structural noise is likely to be minimal? As far as I can tell from the description, coordinates are used in to model.

Reply:

We sincerely thank the reviewer for this insightful and constructive comment. We understand the importance of providing empirical evidence to support our claim of noise robustness and clarifying what we mean by "noise" in our study.

1. **Clarification of noise definition**: In the context of our work, the "noise" refers specifically to structural perturbations in molecular graphs, particularly edge-level noise such as missing chemical bonds, which may arise due to experimental uncertainty in graph construction or preprocessing artifacts [1-4]. Although our model leverages 3D geometric features and is designed to be SE(3)-invariant, graph-based representations (especially the topological structure of the molecular graph) remain central to the encoder. Perturbations to this topology can still significantly impact learned representations in most models.
2. **New experiments to quantify noise robustness**: To address the reviewer's concern, we have added a new

controlled experiment. In this experiment, we simulate graph-level noise by randomly deleting 10%, 20%, and 30% of edges from each molecular graph in the BBBP (from MoleculeNet) and Solubility (from ADME) datasets. We evaluate the ROC-AUC or RMSE performance of three competitive baselines (MolCLR, MGSSL, KANO), our model (MotiL) and an ablation version of MotiL without diffusion priming under each noise level. As shown in Figure 4, MotiL exhibits significantly smaller decreases in ROC-AUC or smaller increases in RMSE as noise levels rise, demonstrating its superior robustness to graph structural perturbations. This robustness is attributed to the diffusion priming phase, which learns denoising-aware representations by modeling and reversing a noise corruption process during pretraining.



(a) BBBP (b) Solubility

Figure 4: Performance of different methods under increasing graph structural noise.

To reflect this improvement, we have revised the manuscript by:

(1) Clarifying our definition of noise as structural perturbations in molecular graphs.

(2) Adding the new experiment and detailed analysis in Supplementary Section G (Robustness evaluation to graph structural noise);

The revised sentence in the main text now reads:

"We propose diffusion priming as the first phase of MotiL to perceive the global structure of a molecule and thus enhance the robustness of GNNs to structural noise perturbations in the molecular graph $\mathcal{G}$ (Fig. 1a)."

"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."

"Also, Supplementary Fig. 8 indicates that the diffusion priming phase used in MotiL effectively mitigates the sensitivity of our method to structural noise perturbations."

The additional content in Supplementary Section G (Robustness evaluation to graph structural noise) is presented as follows:

"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. 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."

[1] Heid E, McGill C J, Vermeire F H, et al. Characterizing uncertainty in machine learning for chemistry[J]. Journal

of Chemical Information and Modeling, 2023, 63(13): 4012-4029.
[2] Ryu S, Kwon Y, Kim W Y. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification[J]. Chemical science, 2019, 10(36): 8438-8446.
[3] Crusius D, Cipcigan F, Biggin P C. Are we fitting data or noise? analysing the predictive power of commonly used datasets in drug-, materials-, and molecular-discovery[J]. Faraday Discussions, 2025, 256: 304-321.
[4] Jiang S, Qin S, Van Lehn R C, et al. Uncertainty quantification for molecular property predictions with graph neural architecture search[J]. Digital Discovery, 2024, 3(8): 1534-1553.

Comment 14: What is the justification for using t-SNE for the small molecule embeddings and UMAP for the protein embeddings?

Reply:

We sincerely thank the reviewer for this insightful question regarding our choice of visualization methods for the learned embeddings. Our decision to use t-SNE for small molecules and UMAP for proteins was guided by both the nature of the underlying data and empirical observations of visualization quality.

1. **For small molecules,** we employed t-SNE because our goal was to inspect whether molecules with similar chemical scaffolds (e.g., naphthalene, barbituric acid) are positioned closely in the embedding space. t-SNE is well-suited for highlighting local neighborhood structures, and is commonly used in molecular representation learning literature for this purpose [1, 2]. Since small molecule graphs tend to be lower-dimensional and structurally simpler, t-SNE is effective in capturing the fine-grained local similarities needed for scaffold-level inspection.
2. **For proteins,** we instead used UMAP due to the higher complexity and dimensionality of protein representations, which encode both 1D sequential and 3D structural features. UMAP has been shown to better preserve both local and global manifold structures, which is critical for interpreting functional similarity across proteins [3, 4]. In our case, UMAP yielded more cohesive and interpretable clusters when visualizing embeddings related to GO terms or EC numbers, which are often multi-label and hierarchically organized.

In summary, our choice reflects a **data-driven and task-sensitive strategy**: t-SNE was preferred for visualizing **local motif-based clustering in small molecules**, while UMAP was more suitable for **global functional embedding structures in proteins**.

We have updated the manuscript to clarify this point. The corresponding revision in the main text reads (added in Section "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).".

[1] Fang, Y. et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nature Machine Intelligence 5, 542–553 (2023).
[2] 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).
[3] Wang J, Zhang S, Qiao H, et al. UMAP-DBP: an improved DNA-binding proteins prediction method based on

uniform manifold approximation and projection[J]. The Protein Journal, 2021, 40: 562-575.
[4] Wang Z, Combs S A, Brand R, et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction[J]. Scientific reports, 2022, 12(1): 6832.

Comment 15: In figure 3a it is a little odd to start the cluster IDs from -1.

Reply:

We thank the reviewer for this helpful observation. In Figure 3a, we used the DBSCAN algorithm to perform unsupervised clustering on the t-SNE embeddings of molecular representations. According to the standard DBSCAN implementation, data points that are identified as outliers or noise (i.e., those not assigned to any cluster) are automatically labeled with cluster ID "-1." We acknowledge that this notation may not be immediately intuitive to all readers. To address this concern and improve clarity, we have revised the manuscript as follows:

We updated the caption of Figure 3a to explicitly clarify the meaning of "-1" as follows:

"UMAP visualization of the representations for ATP binding proteins (Cluster ID -1 represents noise points as defined by DBSCAN, i.e., points not assigned to any cluster)."

Comment 16: What do the green points along the X axis in the inset panel in Figure 3a indicate?

Reply:

We thank the reviewer for this valuable question. The green points shown along the X axis in the inset panel of Figure 3a correspond to data samples from **cluster ID = 9,** which is represented using green in the main t-SNE visualization. These points are valid members of the cluster and are not outliers. However, we acknowledge that their placement along the horizontal axis in the inset panel may have visually suggested they were outliers or unclustered points, which could lead to confusion. This was an unintended artifact of the specific layout and scaling used in the inset. To resolve this and enhance clarity, we have revised Figure 3a in the updated manuscript by replacing the inset panel with a clearer version that eliminates overlap and better preserves spatial relationships. The updated figure is shown below:

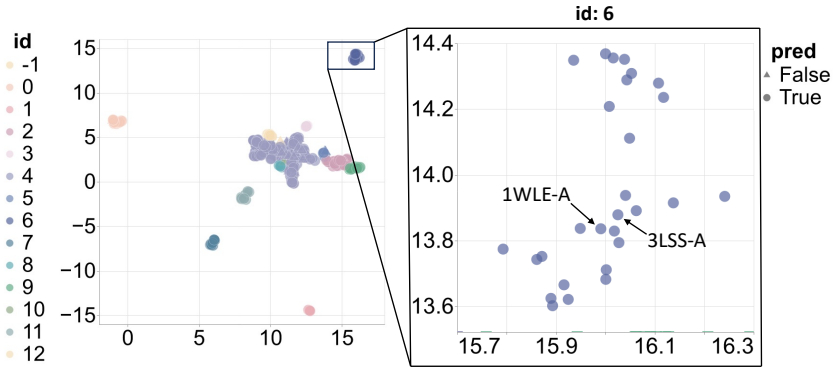


Figure 5: UMAP visualization of the representations for ATP binding proteins (Cluster ID -1 represents noise points as defined by DBSCAN, i.e., points not assigned to any cluster).

Comment 17: I actually disagree with the authors claim that the similarity of the representations of the amino acids reflect their similarity outside of the authors cherry-picked examples. For example, Serine and Tryptophan and the low similarity of Alanine and Glycine.

Reply:

We sincerely thank the reviewer for this insightful comment. We acknowledge the concern that our initial presentation of amino acid similarity was based on a few illustrative examples, which could be perceived as cherry-

picking. We agree that such limited qualitative examples may not be sufficient to support broader claims regarding the fidelity of the learned representations across the full set of amino acids. To address this concern, we would like to clarify the following points:

1. **Systematic analysis provided in submission**: In the originally submitted manuscript, we included a complete pairwise similarity analysis of all 20 standard amino acids based on their learned representations in the pretraining stage. This analysis is visualized as a 20×20 **cosine similarity heatmap** in Supplementary Figure 7 (Appendix Section G), titled "Heatmap of the amino acid's representation similarity". The matrix reflects the global structure of the learned embedding space and provides a holistic view of how the model organizes amino acids.
2. **Interpretation of learned similarities**: We emphasize that the amino acid representations learned by MotiL are **context-sensitive**, i.e., they are shaped by the protein sequence and 3D structural environments encountered during pretraining, rather than fixed based solely on biochemical or physicochemical properties. Therefore, certain similarities or dissimilarities (e.g., Serine–Tryptophan being closer than Alanine–Glycine) arise from **data-driven co-occurrence and structural proximity**, not necessarily canonical chemical similarity.
3. **Specific explanation of the pair of Serine and Tryptophan**: We acknowledge that pairwise examples such as Serine–Tryptophan or Alanine–Glycine may not fully align with conventional biochemical similarity groupings. As the reviewer rightly notes, Serine (a small polar residue) and Tryptophan (a large nonpolar aromatic residue) differ substantially in terms of size, polarity, and side-chain chemistry. Our intent was not to suggest that MotiL's learned representations strictly reproduce canonical biochemical classifications. Instead, the goal was to qualitatively illustrate that the learned motif representations are context-sensitive, i.e., shaped by the local sequence and structural environments encountered during pretraining. For example, Serine and Tryptophan may be encoded similarly if they frequently appear in analogous structural or functional contexts (e.g., surface-exposed residues near active sites or domain boundaries) across the training dataset. This context-dependent behavior is a natural consequence of the self-supervised pretraining framework used in MotiL, where representations are influenced by co-occurrence and neighborhood semantics rather than static chemical properties.
4. **Specific explanation of the pair of Alanine and Glycine**: While Alanine and Glycine have small, nonpolar side chains and are both weakly hydrophobic, their learned embeddings in MotiL are notably dissimilar. This divergence can be attributed to their distinct structural roles: Glycine, due to its minimal side chain, contributes to conformational flexibility and often resides in protein turns or flexible loops, whereas Alanine favors α-helical regions and imparts structural rigidity. Consequently, in context-aware representation learning, such as in MotiL, their embeddings reflect these functional and structural contextual differences rather than their superficial chemical similarity.
5. **Revised wording and figure in the manuscript**: We have revised the relevant descriptions in the manuscript to more carefully reflect this point. Specifically, we now clarify that the examples of amino acid similarities are intended to illustrate localized representation patterns learned by the model rather than to draw broad biochemical conclusions. We explicitly acknowledge that certain pairs, such as Serine and Tryptophan or Alanine and Glycine, may deviate from canonical biochemical expectations due to the context-sensitive nature of the self-supervised learning process. The representations learned by MotiL are influenced by the structural and sequential environments encountered during pretraining, rather than solely by chemical similarity. To provide a more comprehensive and objective analysis, we have revised Supplementary Section J (Detailed results and analysis of protein representations) accordingly and updated Figure 7 (the amino acid similarity heatmap, now shown in Figure 6 below) to enhance visual clarity and interpretability. The updated text in Supplementary Section J "Detailed results and analysis of protein representations" reads:

"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."

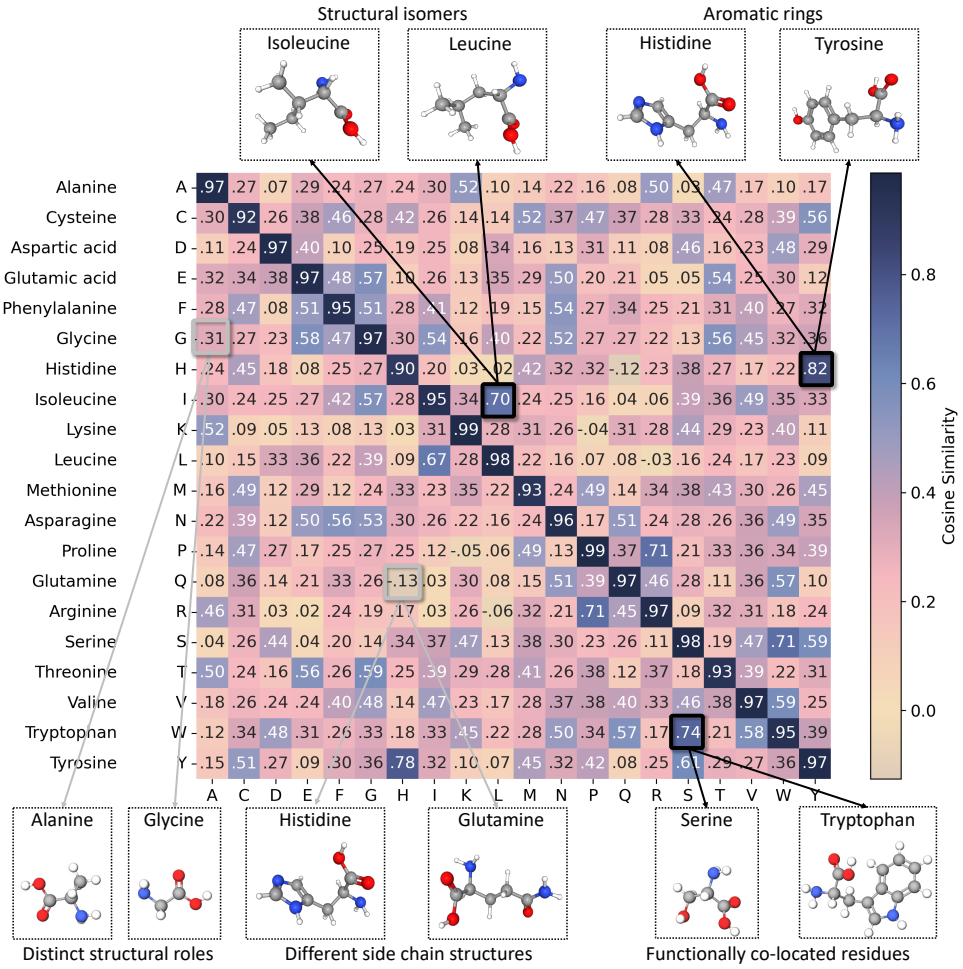


Figure 6: Heatmap of the amino acid's representation similarity.

Comment 18: L458 "Fig 4." is repeated.

Reply:

We thank the reviewer for pointing out the repeated reference to "Fig. 4" in Line 458. We have revised the sentence

to eliminate the redundancy and improve fluency. The updated version in Section "MotiL corrects the erroneous predictions of the strongest baseline" of the main text now reads:

"We present a comprehensive evaluation of prediction disagreements between MotiL and the strongest baseline KANO on two toxicity-related benchmarks: BBBP (blood-brain barrier permeability) and SIDER (drug-induced liver injury). Fig. 4a displays quantitative statistics of disagreement cases between MotiL and KANO on the complete datasets."

"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."

Comment 19: Figure 4 is a little hard to gain significant insight from. Whilst it is nice to see specific examples where the models disagree, cherry picking favorable cases is second-best to a more quantitative summarization of the disagreements. I believe such an analysis should replace this figure, and figure 4 could be an inset panel or moved to the supplementary.

Reply:

We thank the reviewer for this constructive suggestion regarding Figure 4. The intent behind Figure 4 was to provide readers with a concrete, visual comparison between MotiL and KANO by highlighting specific test samples where the two models diverged in their predictions. We agree that although such qualitative examples can be insightful, they are inherently limited in generalizability and may appear as cherry-picking if not supported by broader statistical evidence. In response to the reviewer's valuable feedback, we have taken the following steps to improve the rigor and clarity of this analysis:

1. Addition of a quantitative disagreement analysis: To complement the qualitative examples, we now include a new quantitative evaluation that systematically analyzes prediction disagreements between MotiL and KANO across the entire test sets for BBBP and SIDER. This analysis measures not just how often the models disagree, but also which model is more frequently correct in those disagreement cases. The results are summarized in Figure 7 below. Specifically:
   (1) On the BBBP dataset, the two models produce conflicting predictions on 57 out of 203 test samples (28.1%). Among these disagreement cases, MotiL is correct in 40 instances (70.2%), whereas KANO is correct in only 17 (29.8%).
   (2) On the SIDER dataset, the models disagree on 63 of 295 samples (21.4%). MotiL correctly predicts 43 of these (68.3%), while KANO succeeds on 20 (31.7%).

   These results provide a clear and quantitative indication that MotiL's superiority is not confined to isolated examples, but holds across a sizable portion of challenging cases. It supports the conclusion that MotiL is more robust when facing ambiguous or borderline molecular instances, particularly in the context of real-world biomedical classification.
2. Relocation of Figure 4 to the supplementary materials: In line with the reviewer's suggestion, we have moved the original Figure 4, which presents specific visualized disagreement cases, to the supplementary materials, now labeled as Supplementary Figure 9. We agree that this figure is better suited as a supplemental illustration rather than a main result. In the revised main text, we now reference this figure as a complement to the newly added quantitative analysis, making it clear that it serves only to support, not replace, the broader evaluation.

We have revised the content of Section "MotiL corrects the erroneous predictions of the strongest baseline" in the main text as follows:

"We present a comprehensive evaluation of prediction disagreements between MotiL and the strongest baseline KANO on two toxicity-related benchmarks: BBBP (blood-brain barrier permeability) and SIDER (drug-

induced liver injury). 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 and Supplementary Figs. 10&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."

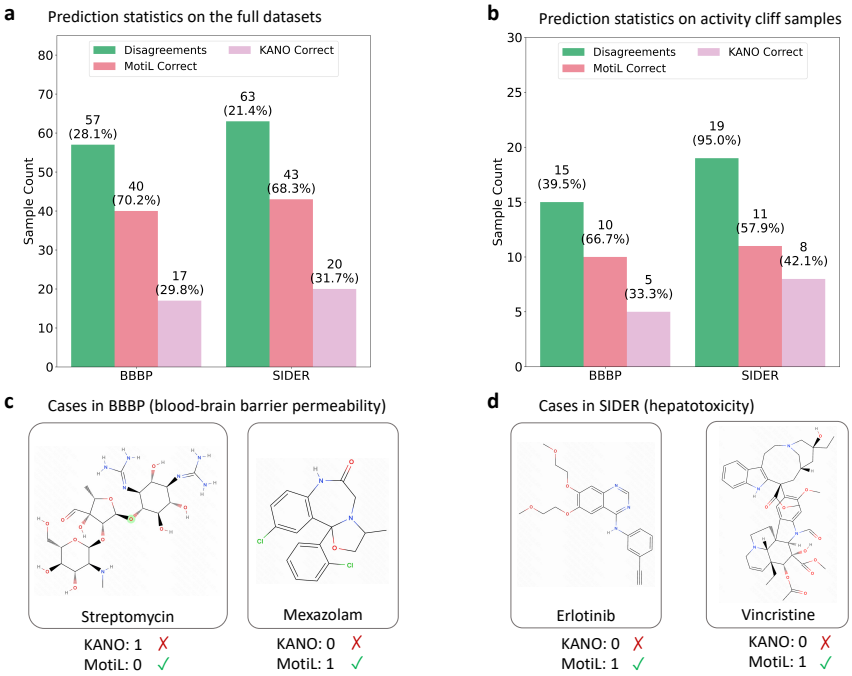


Figure 7: Quantitative and qualitative analysis of prediction differences between KANO [12] and MotiL on BBBP and SIDER. a, Statistics on disagreements between the two models on the full datasets. b, Statistics on disagreements between the two models on activity cliff samples. c, Prediction cases for blood-brain barrier permeability on BBBP. d, Prediction cases for hepatotoxicity on SIDER.

Comment 20: The authors should specify a license for the source code.

Reply:

We thank the reviewer for the suggestion regarding licensing. In response, we have specified the license for our source code. The MotiL codebase is now publicly available at https://github.com/Young0222/MotiL, and is released under the MIT License. This permissive open-source license allows users to freely reuse, modify, and distribute the code with proper attribution. We have included the license information clearly in both the LICENSE file and the README file of the repository. This change ensures transparency, compliance with open-science practices, and ease of use for the research community. As shown in Figure 8, the repository explicitly declares the license (highlighted by red ellipses) on GitHub.

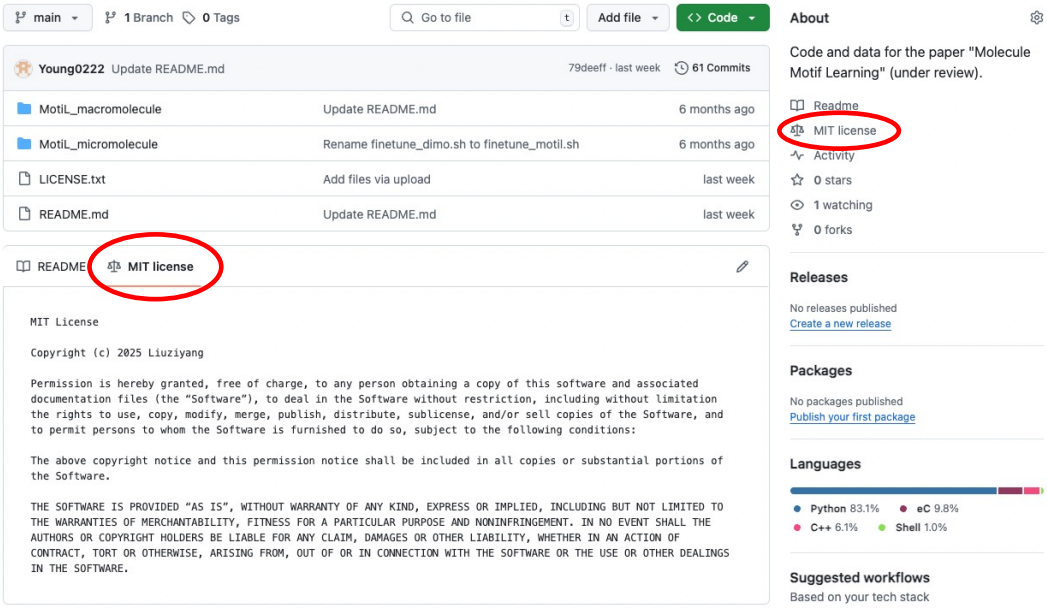


Figure 8: MIT License specified in the MotiL code repository.

Comment 21: Supplementary L68: I think some evidence (or elaboration) needs to be provided for the claim that "the negative effects of molecular graph augmentation have hindered the advancement of MCL".

Reply:

We thank the reviewer for the thoughtful suggestion to substantiate our claim regarding the negative effects of graph augmentation in Molecular Contrastive Learning (MCL). In the revised version, we have provided empirical and theoretical evidence from recent literature to support this point.

Specifically, several studies have pointed out that handcrafted or random augmentations, such as node masking, edge deletion, and subgraph sampling, can significantly distort graph topology and task-relevant information, particularly in molecular domains where subtle structural details are critical. In detail, the work in [1] clearly states: "Handcrafted graph augmented strategies may destroy the graph structure resulting in failing to maintain the task-related information intact… Removing important edges can severely damage graph topology… resulting in poor graph embedding quality". The work in [2] also emphasizes: "The choice of data augmentation for a given task relies heavily on trial and error… Hyperparameters of data augmentations exponentially expand the search space…which is certainly time-consuming and computationally resource-intensive". The work in [3] notes that: "In molecular graphs, dropping a carbon atom from the phenyl ring of aspirin breaks the aromatic system and results in an alkene chain." Collectively, these studies highlight that inappropriate augmentations not only fail to provide consistent

benefits but may even degrade performance, particularly in chemically structured data where topology and function are tightly coupled.

We have revised the supplementary sentence to include supporting evidence and have directly cited the aforementioned references to substantiate our claim. The revised text in Supplementary Section A.2 “Graph representation learning on small molecules” is as follows:

“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.”

[1] Li Y, Zhang H, Yuan Y. Edge Contrastive Learning: An Augmentation-Free Graph Contrastive Learning Model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(17): 18575-18583.

[2] Li H, Cao J, Zhu J, et al. Augmentation-free graph contrastive learning of invariant-discriminative representations[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[3] Lee N, Lee J, Park C. Augmentation-free self-supervised learning on graphs[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(7): 7372-7380.

Comment 22: "L120. To the best of our knowledge, we are the first to propose the concept of molecular motif learning.". Could the authors please elaborate on this and clarify what exactly is being claimed? I could find a few works that purport to address this to some degree.

Reply:

Thank you for this important question. We appreciate the reviewer’s careful reading and effort to connect our work with the related literature. Indeed, molecular motifs have been used in several prior work, but their role in these studies is fundamentally different from the motif representation learning proposed in our work. Below, we summarize five representative motif-related studies and highlight the key distinctions.

1. **Summary of prior work**:
   (1) MGSSL [1]: Leverages chemical fragmentation (BRICS + rules) to predefine a motif vocabulary and formulates a motif generation task. GNNs are trained to autoregressively predict the motif tree structure of a molecule. Motifs are treated as targets for generative prediction, not as semantic representation carriers.
   (2) DGPM [2]: Proposes a dual-level graph pretraining framework where motifs are automatically discovered via edge pooling, and used as intermediate substructures for aligning subgraph kernels. Motifs serve as auxiliary self-supervised targets to enhance graph-level learning.
   (3) MOTOR [3]: Focuses on drug-drug interaction prediction, using BRICS-fragmented motifs as predefined substructures. Motifs are utilized to model intra- and inter-drug subgraph interactions, emphasizing the relational context rather than learning motif semantics.
   (4) HM-GNN [4]: Constructs a heterogeneous graph of motifs and molecules, using motifs as nodes to enable message passing between molecules sharing similar substructures. However, the model learns molecular representations on this augmented graph, not explicit motif embeddings.
   (5) MICRO-Graph [5]: Employs motif-like subgraph clustering to guide subgraph sampling for contrastive learning. Motifs here serve as prototypes for generating informative subgraphs, facilitating better graph-level contrastive pretraining. Motifs are tools for data augmentation, not learned representation units.

2. **Key distinction (motif-based learning vs. motif representation learning)**: From the above, we observe that prior work fall under what we would describe as **motif-based learning or learning with the help of motifs**:

(1) Motifs are used as tools to assist molecular learning, e.g., for subgraph sampling, pretraining tasks, message passing, or generative prediction.

(2) These approaches do not explicitly learn transferable motif representations that can generalize across different molecules.

In contrast, our work is the first to explicitly propose **molecular motif representation learning**, where:

(1) Molecular motifs are the representation units to be learned, not merely tools or targets;

(2) We optimize for motif-level semantic consistency and alignment across molecules, enabling motifs to serve as transferable, interpretable, and task-agnostic representations;

(3) Our framework is designed to work across micro- and macro-molecules, generalizing beyond small molecule property prediction.

To better reflect this contribution, we have revised the last paragraph of Section "Introduction" in the main text as follows:

"To the best of our knowledge, we are the first to 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."

We have also added further discussion to Supplementary Section A.2 "Graph representation learning on small molecules" to compare these distinctions explicitly:

"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 work 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."

[1] Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised learning for molecular property prediction[J]. Advances in Neural Information Processing Systems, 2021, 34: 15870-15882.

[2] Yan P, Song K, Jiang Z, et al. Empowering dual-level graph self-supervised pretraining with motif discovery[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(8): 9223-9231.

[3] Zhang R, Wang X, Liu G, et al. Motif-Oriented Representation Learning with Topology Refinement for Drug-Drug Interaction Prediction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(1): 1102-1110.

[4] Yu Z, Gao H. Molecular representation learning via heterogeneous motif graph neural networks[C]//International Conference on Machine Learning. PMLR, 2022: 25581-25594.

[5] Zhang S, Hu Z, Subramonian A, et al. Motif-driven contrastive learning of graph representations[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(8): 4063-4075.

Response to Reviewer #2:

Comment 1: It would be helpful to clarify why the authors chose MoleculeNet.

Reply:

We thank the reviewer for raising this valuable point regarding our dataset selection. We chose to evaluate our model on the MoleculeNet benchmark suite for several important and well-justified reasons:

1. **Widespread adoption and fair comparison with strong baselines**: MoleculeNet is widely recognized as a standard benchmark for molecular property prediction, and has been extensively adopted in the field. Many recent state-of-the-art methods, including MolCLR [1], GROVER [2], KANO [3], and MGSSL [4], have reported results on MoleculeNet tasks using the same data splits and evaluation protocols. We follow this widely accepted setup to ensure that our results are directly comparable to prior work. This consistency enables us to isolate the effects of our architectural and pretraining innovations, especially our motif-level contrastive design, without introducing confounding factors from dataset differences.
2. **Rich diversity of tasks and molecular properties**: MoleculeNet covers a wide variety of molecular property prediction tasks that test different aspects of model capability, including: (1) Physicochemical properties, such as solubility (ESOL) and hydration free energy (FreeSolv), (2) Biological activity, such as BACE (enzyme inhibition) and HIV (replication inhibition), (3) ADMET-related properties, including blood–brain barrier permeability (BBBP), drug side effects (SIDER), and toxicity. This breadth allows us to evaluate whether our proposed MotiL framework generalizes effectively across different molecular domains and task types (both classification and regression). It also helps assess whether motif-level representation learning can discover meaningful patterns shared across these diverse endpoints.
3. **Standardized data splits and evaluation protocols**: MoleculeNet provides predefined data splits (e.g., scaffold split, which is particularly challenging as it forces the model to generalize to novel chemical scaffolds) and standard metrics (e.g., RMSE, ROC-AUC), facilitating transparent, fair, and reproducible evaluation. This standardization ensures that performance comparisons are not influenced by arbitrary splitting strategies or inconsistent evaluation criteria.
4. **Relevance to pretraining-based methods and contrastive learning**: Our work builds upon the line of research in self-supervised pretraining and contrastive learning for molecular graphs. Most relevant prior work in this area, including MolCLR [1], GROVER [2], and KANO [3], adopt MoleculeNet as their primary evaluation benchmark. Using the same setup not only strengthens the methodological continuity but also ensures that any performance gains of MotiL can be credibly attributed to our design choices (e.g., motif-level alignment and diffusion priming), rather than dataset biases or incompatible experimental setups.

To further clarify our rationale and ensure transparency for all readers, we have added a detailed explanation of our dataset selection in Supplementary Section B (Overview of experimental datasets) as follows:

"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."

[1] 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).
[2] Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems 33, 12559–12571 (2020).
[3] Fang, Y. et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nature Machine Intelligence 5, 542–553 (2023).
[4] Zhang, Z. et al. Motif-based graph self-supervised learning for molecular property prediction. Advances in Neural Information Processing Systems 34, 15870–15882 (2021).

Comment 2: It was unclear from the main text that the metrics obtained for other methods in Tables 1 and 2 were not calculated by the authors but rather sourced from previous studies ([12], [27], [24]). Since the metrics for other models were taken from past articles, how can the authors guarantee that the comparison is fair and not influenced by differences in training and evaluation methods? Is the cross-validation method identical? Additionally, this approach limits further comparison. Also, it might be harder to answer why RMSE was used to evaluate physical chemistry properties and MAE for quantum properties and why more metrics for regression and classification weren't explored?

Reply:

We thank the reviewer for pointing out the need for clearer disclosure regarding the source of baseline results in Tables 1 and 2, and for raising valid concerns about the fairness and consistency of comparisons as well as the choice of evaluation metrics.

1. **On the source and fairness of baseline results:** We clarify that all baseline results presented in Tables 1 and 2 are directly sourced from existing literature, as follows:
   (1) For small-molecule datasets (Table 1), all baseline results, including those for MolCLR, MGSSL, GROVER, and KANO, are taken from the KANO paper [1], which uses a standardized and rigorous evaluation protocol. We have removed the previously included method (denoted as "Zhao et al." in the original manuscript) [2], as its evaluation setup differs significantly from other baselines: the method used linear probing, which introduces supervised information via a frozen encoder and a separate linear classifier. Moreover, the code was not publicly available, making consistent reproduction and evaluation difficult. We acknowledge this inconsistency and have excluded the method from the revised manuscript to ensure fair comparison.
   (2) The remaining baselines all follow the same evaluation setup defined in the KANO paper [1]: "**three independent runs using three random-seeded scaffold splits** for all datasets, except QM9, with a training/validation/testing ratio of **8:1:1**." Our own experiments were conducted following **this exact setup** for all MoleculeNet datasets, thereby ensuring comparability across methods.
   (3) For macromolecule datasets (Table 2), we also ensured identical evaluation conditions for all methods. Specifically, for GO-BP, GO-MF, GO-CC, and EC prediction tasks, we used the same **10-fold cross-validation protocol and dataset splits** defined in the CDConv paper [3]). The split ratios are:
      a) GO-BP, GO-MF, GO-CC: **29,898 / 3,322 / 3,415** (train / val / test)

b) EC: **538 / 1,729 / 1,919** (train / val / test)

We have updated the manuscript to make this provenance and consistency clearer. The updated sentence in the caption of Table 1 now reads:

"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".

The updated sentence in the caption of Table 2 now reads:

"All baseline results in this table are directly reported from Ref. [24], which follows the same evaluation protocol as this study: 10-fold cross-validation splits for proteins."

2. **On the choice of evaluation metrics:** We appreciate the reviewer's inquiry about the rationale for selecting different metrics (RMSE vs. MAE) across datasets. This choice **aligns with standard practices adopted in the respective prior work** that introduced the datasets:
   (1) For physical chemistry datasets (ESOL, FreeSolv, and Lipophilicity), we use RMSE (abbreviation of Root Mean Square Error) following MoleculeNet and prior work such as MolCLR, MGSSL, and GROVER.
   (2) For quantum datasets (QM7, QM8, and QM9), we use MAE (abbreviation of Mean Absolute Error), again following the convention from MoleculeNet and its successors, due to its interpretability and widespread use in quantum property prediction.
3. **On the other evaluation method:** We agree that more evaluation metrics could enrich the analysis. However, to ensure fair and reproducible comparison with existing benchmarks, we followed the most commonly reported metrics in the literature. In addition, we have further improved **visualization-based evaluations** for both small molecules and macromolecules, which provide intuitive insights into the learned representations beyond scalar metrics. Specifically, we chose t-SNE for small molecules and UMAP for proteins based on the data characteristics and visualization needs. For small molecules, t-SNE works well to show whether compounds with similar scaffolds (e.g., naphthalene, barbituric acid) cluster together, as it highlights local structure. Small molecules are simpler and lower-dimensional, making t-SNE a good fit for revealing scaffold-level patterns. For proteins, which have more complex and high-dimensional representations, UMAP better preserves both local and global relationships. It helped reveal clearer functional groupings, such as GO terms and EC numbers, which are often hierarchical and multi-label. For example, in the UMAP visualization for proteins, the representations obtained by MotiL for ATP binding proteins are discriminative and the prediction results are relatively accurate.

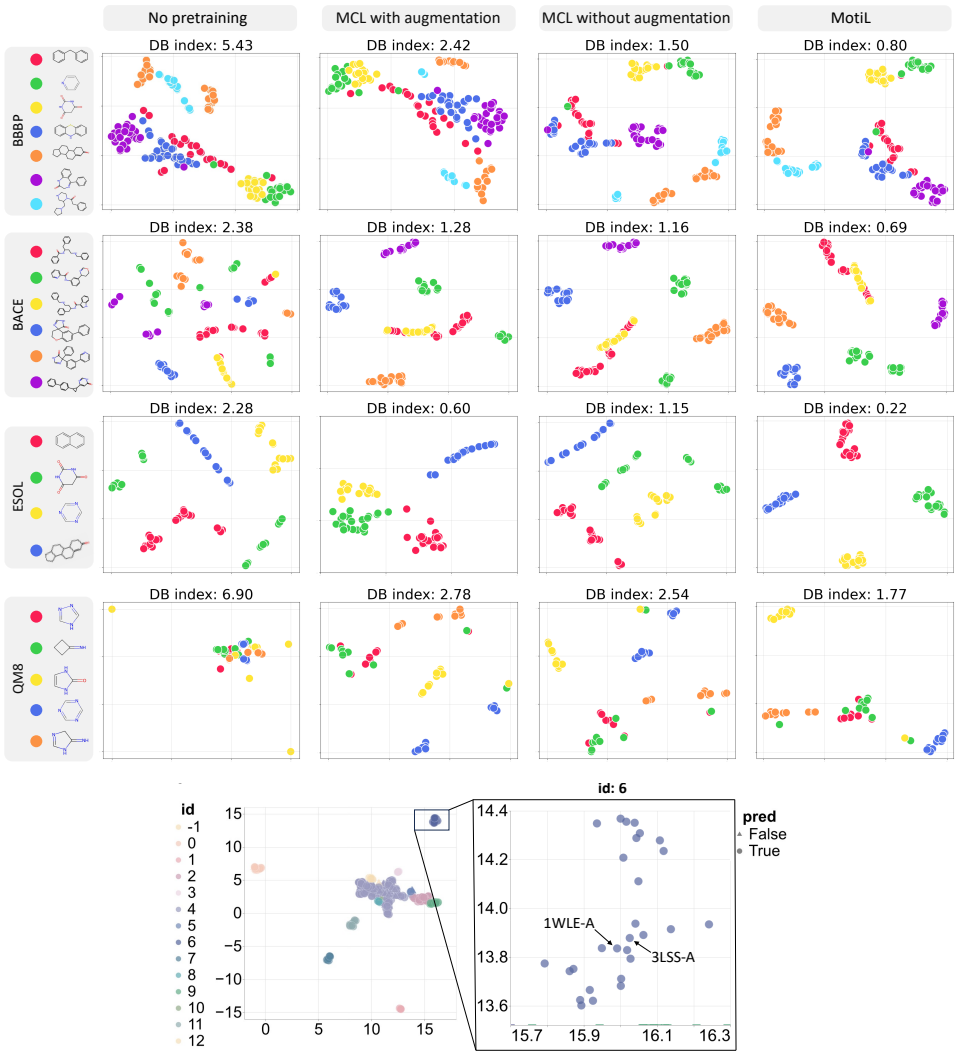


Figure 1: Visualization of representations using t-SNE for small molecules (upper panel) and UMAP for proteins (bottom panel).

4. **Manuscript revisions:** To address these concerns and improve clarity, we have made the following changes in the revised manuscript:

   (1) Clarified in Table 1 caption that all baseline results are taken from the KANO paper [1] using the same experimental protocol; Removed the Zhao et al. baseline [2] due to differences in evaluation methodology and lack of reproducibility. The updated Table 1 and its caption are shown in Table 1.

   (2) Clarified in Table 2 caption that all baseline results are taken from the CDConv paper [3] using the same experimental protocol. The updated Table 2 and its caption are shown in Table 2.

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

[2] 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 (2024).

[3] Fan, H., Wang, Z., Yang, Y. & Kankanhalli, M. S. Continuous-discrete convolution for geometry-sequence modeling in proteins. The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023).

Table 1: Performance comparison on small molecules. We use the metric of ROC-AUC (%) for classification tasks and RMSE or MAE for regression tasks. A lower value of RMSE or MAE indicates better performance. The best result for each dataset is highlighted in bold, and the runner-up result is italicized. 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.

Category	**Physiology**					**Biophysics**			**Physical chemistry**			**Quantum mechanics**		
Metric	ROC-AUC(%)↑					ROC-AUC(%)↑			RMSE↓			MAE↓		
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV	ESOL	FreeSolv	Lip.	QM7	QM8	QM9
# molecules	2,039	7,831	8,597	1,427	1,478	1,513	93,087	41,127	1,128	642	4,200	6,830	21,786	133,885
# tasks	1	12	617	27	2	1	17	1	1	1	1	1	12	3
GCN [34]	71.8	70.9	65.0	53.6	62.5	71.6	71.6	74.0	1.43	2.87	0.71	123	3.66e-2	8.35e-3
GIN [35]	65.8	74.0	66.7	57.3	58.0	70.1	71.8	75.3	1.45	2.77	0.85	125	3.71e-2	8.24e-3
MPNN [36]	91.3	80.8	69.1	59.5	87.9	81.5	75.7	77.0	1.17	1.62	0.67	111	1.48e-2	5.22e-3
DMPNN [37]	91.9	75.9	63.7	57.0	90.6	85.2	78.6	77.1	1.05	1.67	0.68	104	1.56e-2	5.14e-3
CMPNN [38]	95.7	80.4	70.4	63.4	87.6	85.6	76.4	82.5	0.80	1.57	0.61	75.1	1.53e-2	4.05e-3
N-GRAM [30]	91.2	76.9	-	63.2	87.5	79.1	76.9	78.7	1.10	2.51	0.88	126	3.20e-2	9.64e-3
Hu et. al [29]	70.8	78.7	65.7	62.7	72.6	84.5	81.3	79.9	1.10	2.76	0.74	113	2.15e-2	9.22e-3
MGSSL [28]	70.5	76.4	64.1	61.8	80.7	79.7	78.7	79.5	-	-	-	-	-	-
GEM [10]	72.4	78.1	69.2	63.2	90.1	85.6	81.7	80.6	0.81	1.75	0.67	60.0	1.63e-2	5.62e-3
GROVER [27]	86.8	80.3	56.8	61.2	70.3	82.4	67.3	68.2	1.42	2.95	0.82	91.3	1.82e-2	7.19e-3
GraphMVP [39]	72.4	75.9	63.1	63.9	79.1	81.2	77.7	77.0	-	-	-	-	-	-
MolCLR [11]	73.3	74.1	65.9	61.2	89.8	82.8	78.9	77.4	1.11	2.30	0.79	90.9	1.85e-2	4.80e-3
MolCLR$_{CMPNN}$ [11]	72.4	78.4	69.1	59.7	88.0	85.0	74.5	77.8	0.91	2.02	0.88	89.8	1.79e-2	4.75e-3
KANO [12]	*96.0*	*83.7*	*73.2*	*65.2*	*94.4*	*93.1*	*83.7*	*85.1*	*0.67*	*1.14*	*0.57*	*56.4*	*1.23e-2*	*3.20e-3*
MotiL	**96.9**	**84.0**	**73.5**	**65.6**	**95.2**	**93.9**	**85.0**	**85.6**	**0.64**	**1.09**	**0.54**	**50.7**	**1.17e-2**	**2.99e-3**

Table 2: Performance comparison on protein macromolecules. All baseline results in this table are directly reported from Ref. [24], which follows the same evaluation protocol as this study: 10-fold cross-validation splits for proteins. The pretraining phase of MotiL does not incorporate extra protein data, which is consistent with baselines such as GearNet.

Dataset	**GO-BP**	**GO-MF**	**GO-CC**	**EC**
Metric	F_{max}(%) ↑			
# proteins	36,635			19,198
CNN [25]	24.4	35.4	28.7	54.5
ResNet [26]	28.0	40.5	30.4	60.5
LSTM [26]	22.5	32.1	28.3	42.5
Transformer [26]	26.4	21.1	40.5	23.8
GCN [34]	25.2	19.5	32.9	32.0
GAT [40]	28.4	31.7	38.5	36.8
3D CNN [41]	24.0	14.7	30.5	7.7
GraphQA [42]	30.8	32.9	41.3	50.9
GVP [43]	32.6	42.6	42.0	48.9
IEConv (residue level) [44]	42.1	62.4	43.1	-
GearNet [45]	35.6	50.3	41.4	73.0
GearNet-IEConv [45]	38.1	56.3	42.2	80.0
GearNet-Edge [45]	40.3	58.0	45.0	81.0
GearNet-Edge-IEConv [45]	40.0	58.1	43.0	81.0
CDConv [24]	*45.3*	*65.4*	*47.9*	*82.0*
MotiL	**47.9**	**67.6**	**48.9**	**88.1**

Comment 3: In the SI, it is mentioned that the standard deviations (std) were calculated from 3 independent runs. Could the authors clarify what this means? Given the std values, is it possible to perform statistical analyses? Since the paper claims that MotiL surpasses other models, is this improvement statistically significant?

Reply:

We thank the reviewer for this important and insightful comment. We address each point below and have revised the manuscript accordingly for clarity and completeness.

1. **Clarification on "3 independent runs":** When we report mean and standard deviation (std) values for MotiL's performance across different datasets and tasks, these were obtained by:
   (1) Training the MotiL model three times independently, using the same dataset split and model configuration.

(2) Each run was initialized with a different random seed to account for variability from random weight initialization, data loading, and optimizer behavior.

(3) The mean and standard deviation were then computed over these three runs.

We have clarified this procedure in the revised Supplementary Section B (Overview of experimental datasets):

"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."

2. **On statistical significance and interpretability of standard deviations:** On one hand, we observed that the standard deviations across runs are consistently low (typically **less than 1.5%** in absolute terms), suggesting that the performance of MotiL is stable and reproducible under different random initializations. On the other hand, to further address this concern and validate the reliability of our reported improvements, we have conducted additional statistical significance tests. Specifically, we performed paired **two-tailed t-tests** comparing MotiL with the **strongest baseline KANO** across all 14 MoleculeNet benchmark datasets. The resulting *p*-values are summarized in Table 3 and Table 4 (also included in Supplementary Table 3 and Table 4).

These *p*-values demonstrate that the improvements of MotiL over KANO are **statistically significant ($p < 0.05$)** on all tested datasets, confirming that the performance gains are consistent and unlikely to be due to random variation. The corresponding statistical analysis has been included in the revised manuscript under Supplementary Section D (Complete performance comparison on small molecules), as presented below:

"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 Table 3 and Table 4. All values are below the 0.05 threshold, confirming that MotiL's improvements are statistically significant and robust across different datasets."

Table 3: ROC-AUC (%) performance comparison on small molecules. The testing results are presented as the mean and standard deviation, calculated from three independent runs. The best result for each dataset is highlighted in bold, and the runner-up result is italicized. All baseline results are directly taken from Ref. [23], which follows the same evaluation protocol as this study: three random-seeded scaffold splits for small molecules.

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 [37]	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 [65]	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 [66]	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 [67]	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

Table 4: Performance comparison on small molecules. We use the metric of RMSE or MAE where a lower value indicates better performance. The testing results are presented as the mean and standard deviation, calculated from three independent runs. The best result for each dataset is highlighted in bold, and the runner-up result is italicized. All baseline results are directly taken from Ref. [23], which follows the same evaluation protocol as this study: three random-seeded scaffold splits for small molecules.

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 [37]	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 [65]	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 [66]	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 [67]	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

Comment 4: The study in SI section F, which evaluates the different components of MotiL, is interesting and important. I agree that the results suggest that the full MotiL architecture performs better, especially in regression tasks. It would be beneficial, if possible, to add a sentence to the main text summarizing the key conclusions from these ablation studies.

Reply:

We thank the reviewer for highlighting the value of the ablation study presented in Supplementary Section F. We agree that a concise summary of its findings in the main text would improve the clarity and completeness of the manuscript. In response to this suggestion, we have added the following sentence to the main paper to summarize the key conclusions:

"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."

This addition succinctly reflects the main insights derived from the ablation results, particularly the pivotal role of DiffMoM in improving performance, especially in regression tasks such as QM7, QM8, and QM9. We appreciate the reviewer's thoughtful recommendation, which has helped us improve the presentation of our findings.

Comment 5: I find Figure 2a very insightful, as it clearly shows that MotiL can distinguish different molecular scaffolds. However, I suggest using more contrastive colors, particularly in the BBBP plots, where some shades (e.g., pinks) are too similar. Additionally, what was the rationale behind choosing the specific molecules, scaffolds, and database for Figure 2a?

Reply:

We sincerely thank the reviewer for appreciating the value of Figure 2a and for the constructive suggestions regarding its color scheme and design rationale.

1. **Color enhancement:** We agree that some of the original color shades in the BBBP plot, particularly the pink color, were visually similar and may have caused confusion. In the revised version of the manuscript, we have updated Figure 2a by adopting a set of more contrastive and distinguishable colors across different scaffolds, which improves the interpretability. The color palette was carefully selected using colorblind-friendly principles

to ensure clarity for a broad range of viewers. The updated figure has been incorporated into the revised manuscript as Fig. 2a and is also presented as Figure 2 below for reference.

2. **Selection rationale:** We selected the specific molecules, scaffolds, and datasets in this figure to highlight the clustering capability of MotiL and to ensure representative diversity across both chemical and task domains.
   (1) **Dataset selection**: We included four benchmark datasets-ESOL, QM8, BBBP, and BACE-**each representing a distinct task type** (e.g., regression vs. classification). It provides a comprehensive view of how well MotiL generalizes in learning representations across various molecular properties.
   (2) **Scaffold selection**: For each dataset, we selected molecules that belong to chemically distinct scaffolds. It was determined using the Bemis-Murcko scaffold definition to ensure that the scaffolds **differ in core ring systems or frameworks**. Our goal was to visualize whether MotiL can cluster these distinct scaffolds cleanly in the embedding space.
   (3) **Molecule sampling**: Within each scaffold group, we randomly **sampled a small number of molecules that span different activity or property values** (e.g., log solubility in ESOL or BBB permeability in BBBP). It helps evaluate whether structurally related molecules are represented in a locally coherent way.

We have added a short explanatory note in the main text (Section "MotiL learns micromolecular representations consistent with molecular structures") to clarify the rationale behind the molecule and scaffold selection process. The corresponding revision in the manuscript reads:

"For example, we select four chemically diverse scaffolds from ESOL where the color palette has been optimized to enhance visual contrast and interpretability. 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."

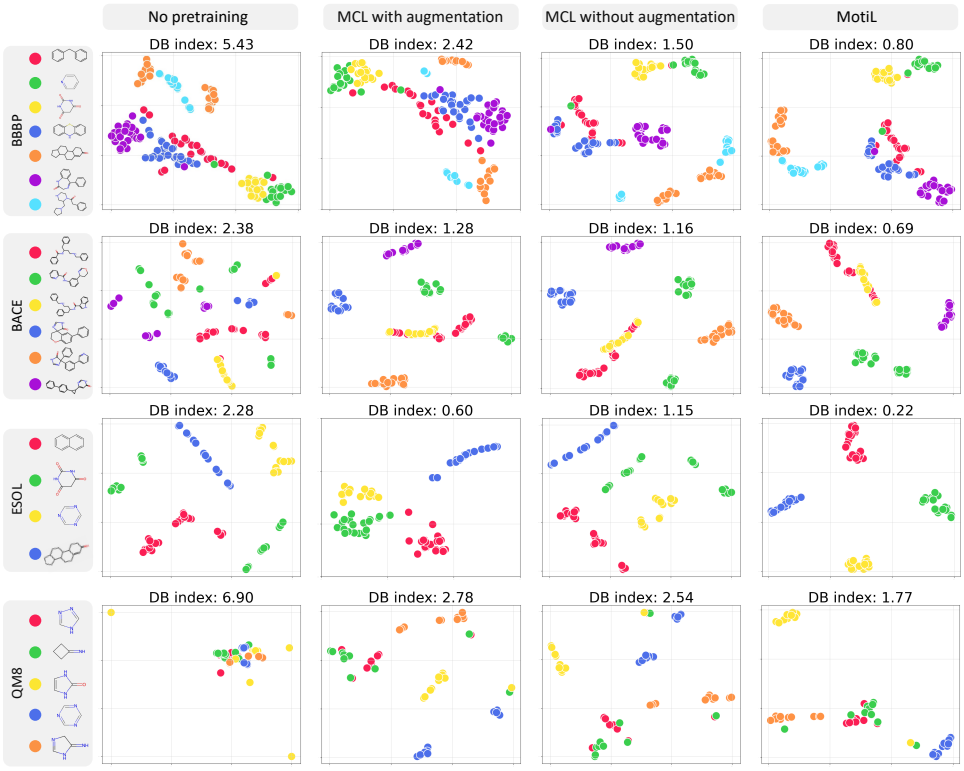


Figure 2: T-SNE visualization of the molecular representations learned by MotiL. Different colors represent different molecular scaffolds, and a lower DB index reflects better separation and clearer distinction between the clusters.

Comment 6: In Figure 2b, and especially in Figure 3 in the SI, some dashed lines are difficult to see.

Reply:

We thank the reviewer for pointing out the visibility issue regarding the dashed lines in Figure 2b and Figure 3 of the Supplementary Information. To address this concern and enhance clarity, we have updated both figures by replacing the original dashed lines with darker and thicker styles. These visual enhancements improve the contrast and ensure that key structural relationships and augmentation paths are clearly distinguishable to readers. The revised figures are now included in the updated manuscript (Figure 2b) and Supplementary Information (Figure 3), respectively. The revised figures are presented below as Figure 3 and Figure 4.

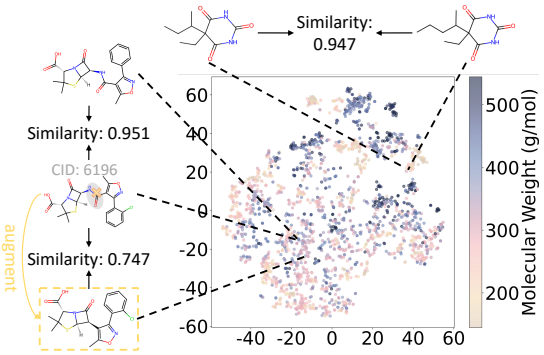


Figure 3: Comparison of positions in 2D space for structurally similar and dissimilar molecular pairs on BBBP.

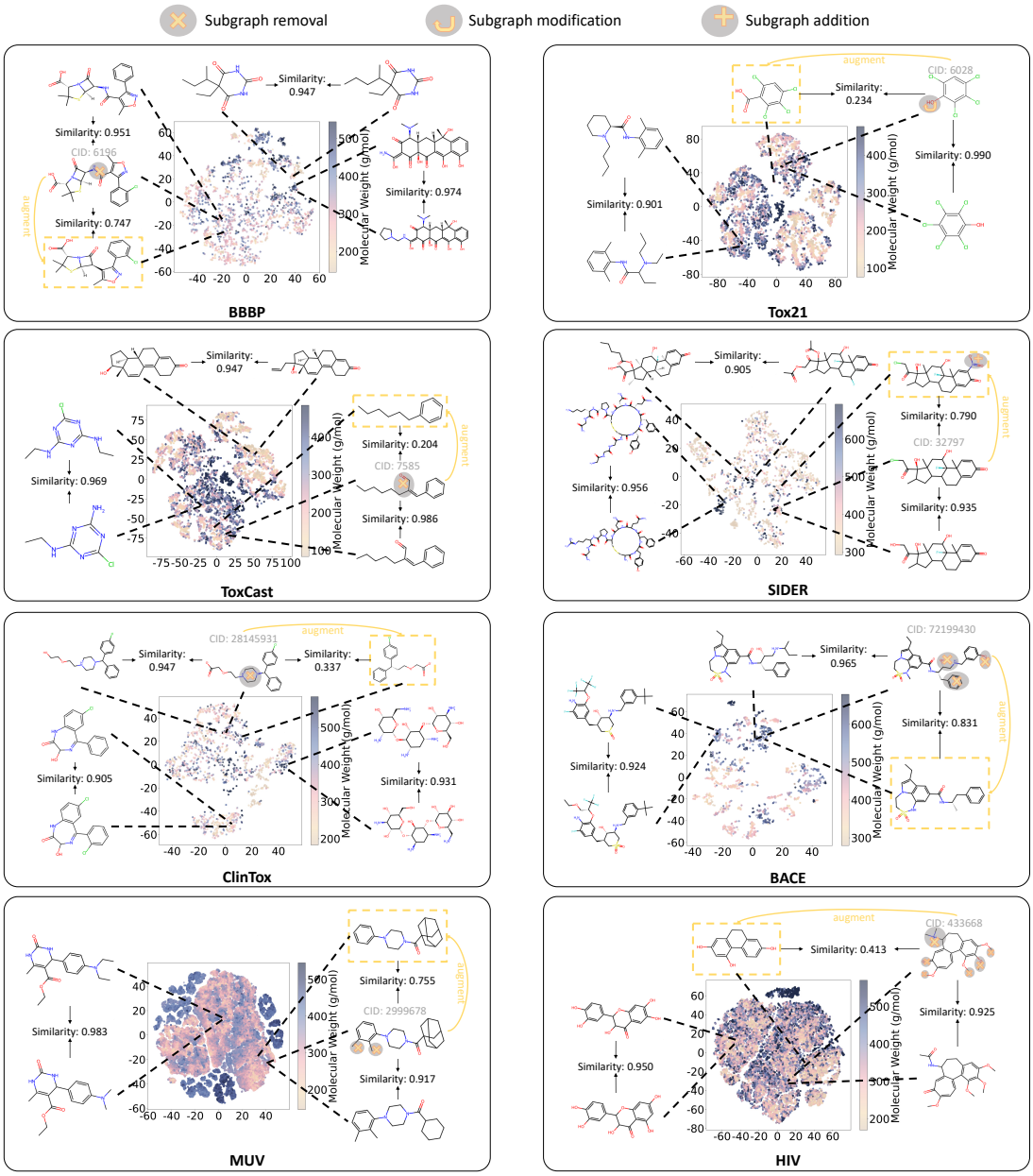


Figure 4: Visualization of micromolecular representations learned by MotiL via t-SNE on multiple datasets.

Comment 7: In not sure if I fully understand the data behind Figure 3a. UMAP and DBSCAN were used to project the representations learned by MotiL for ATP-binding proteins. Why was this specific class chosen? Also, does the shape in the figure indicate whether a prediction is correct in identifying an ATP-binding protein?

Reply:

We thank the reviewer for the thoughtful questions. We selected ATP-binding proteins for this analysis because they represent a biologically important and well-characterized functional class. ATP-binding proteins play critical roles in numerous cellular processes, including energy transfer, signal transduction, and molecular transport, and they are commonly studied in drug discovery as therapeutic targets. Moreover, the functional class of ATP-binding proteins has sufficient diversity and sample size in our dataset, making it suitable for evaluating whether the learned representations capture meaningful biological distinctions.

In Figure 3a, we apply UMAP to visualize the protein representations produced by MotiL's encoder (specifically from the last graph convolution layer), and use DBSCAN to reveal natural clusters in the embedding

space. These clusters reflect how MotiL organizes proteins with similar functional signatures. We chose Cluster 6 as an illustrative example where MotiL groups together ATP-binding proteins such as 1WLE-A and 3LSS-A.

Regarding the visual encoding, we confirm that the **shape of each marker** in the zoomed-in panel indicates **whether the ATP-binding function of that protein was predicted correctly** by MotiL: circular markers denote correct predictions, while triangular markers denote incorrect predictions. As highlighted in the example, all proteins in Cluster 6 are predicted correctly, suggesting that MotiL's embedding space supports functionally consistent clustering.

We have now clarified the visual encoding more explicitly: in the updated figure (as shown in Figure 5 below), circular markers represent correct predictions, while triangular markers indicate incorrect ones. This clearer annotation ensures that readers can easily distinguish prediction correctness based on marker shape.

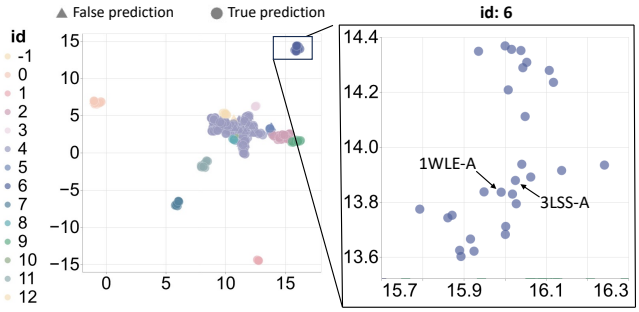


Figure 5: UMAP visualization of the representations for ATP binding proteins (Cluster ID -1 represents noise points as defined by DBSCAN, i.e., points not assigned to any cluster).

Comment 8: Figure 3c could be better organized. I prefer the layout used in Figure 5 SI, and I think more GO terms should be written to highlight similarities. For example, in ID 3, the common GO terms, protein kinase activity, Kinase activity, and Protein tyrosine kinase activity, could be included.

Reply:

We sincerely thank the reviewer for their insightful suggestion regarding the layout and annotation of Figure 3c, and we fully agree that enhancing the clarity and biological interpretability of this figure would improve the presentation of our results.

To address the reviewer's concerns, we have made the following revisions:

1. **Improved figure layout**: We have redesigned the layout of Figure 3c following the structure used in Figure 5 of the Supplementary Information, as suggested. Instead of displaying only a few protein pairs, the updated version (shown in Figure 6 below) presents representative protein structures from the same DBSCAN-identified cluster (id: 3). This revision facilitates a more intuitive and visually systematic comparison of the structural similarities among proteins within the same cluster, helping to better showcase the ability of MotiL to group proteins with related 3D folds.
2. **Expanded GO term annotation for functional similarity**: We now include more comprehensive sets of shared Gene Ontology (GO) terms in each cluster to better reflect the biological consistency of the grouped proteins. As you suggest, we annotate the following common GO terms: "protein kinase activity", "kinase activity", and "protein tyrosine kinase activity" for Cluster ID 3 (e.g., 4AU8-A, 1G3N-A, 2BPM-A, 5XVU-A, and 3VN9-A). These GO terms are extracted based on shared functional annotations from UniProt and are selected to emphasize high-confidence, biologically relevant commonalities. Specifically, protein kinase activity (GO:0004672) refers to the enzymatic function of transferring a phosphate group, typically from ATP, to a protein substrate. This post-translational modification is crucial for regulating protein function, activity,

localization, and interactions in a wide range of cellular processes. Kinase activity (GO:0016301) is a broader term encompassing all enzymes that catalyze the phosphorylation of substrates, not limited to proteins. It includes protein kinases, lipid kinases, carbohydrate kinases, and others involved in signal transduction and metabolic regulation. Protein tyrosine kinase activity (GO:0004713) is a subtype of protein kinase activity that specifically phosphorylates tyrosine residues on proteins. These kinases are key regulators of signaling pathways that control cell growth, differentiation, immune response, and oncogenic transformation.

3. **Clarified caption and in-text description**: To guide interpretation, we revised the caption of Figure 3c in the main text as follows: *"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."*

Furthermore, in the main text, we have revised the description of Figure 3c in the section "MotiL learns macromolecular representations consistent with protein structures and chemical properties" to better reflect the improved layout and the inclusion of representative GO terms that highlight functional similarity among clustered proteins:

"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."

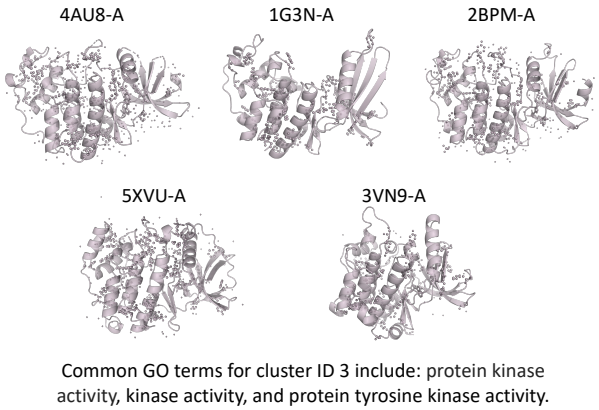


Figure 6: 3D structure comparison of proteins in cluster ID 3.

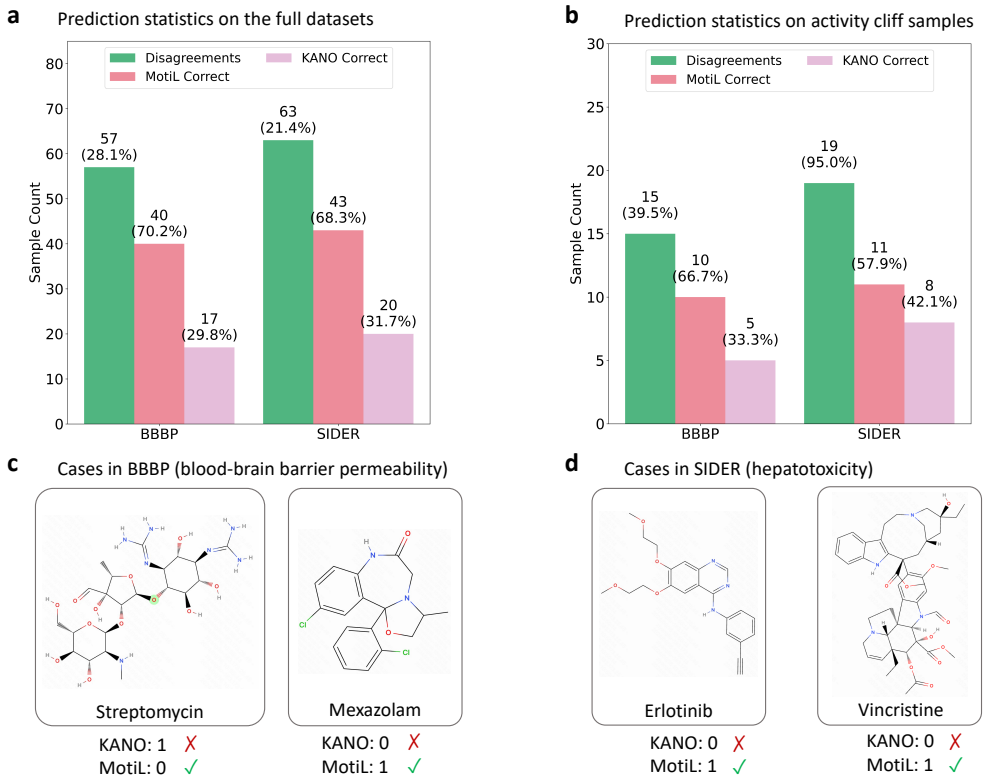


Figure 7: Quantitative and qualitative analysis of prediction differences between KANO and MotiL on BBBP and SIDER. a, Statistics on disagreements between the two models on the full datasets. b, Statistics on disagreements between the two models on activity cliff samples. c, Prediction cases for blood-brain barrier permeability on BBBP. d, Prediction cases for hepatotoxicity on SIDER.

Comment 9: The authors also make a more in-depth comparison between the 2nd best algorithm (KANO) and MotiL, but I struggle to understand the significance of this section. In fact, improvements between KANO and MotiL appear to be marginal. While I acknowledge that MotiL correctly predicts BBB permeability and toxicity for key molecules while KANO does not, I find this section unnecessarily long and less relevant. I believe a more challenging comparison could be beneficial. For example, evaluating performance on activity cliffs.

Reply:

We sincerely thank the reviewer for this insightful and constructive comment. We fully agree that evaluating on more challenging cases, such as activity cliffs, can better highlight the differences between models and is more meaningful than simply reporting average scores. In this revision, we have made the following comprehensive updates to address this concern:

1. **On the significance of the case-level comparison with KANO**: While the average metric improvements between MotiL and KANO may seem modest (e.g., ROC-AUC: +0.9% on BBBP, +0.4% on SIDER), we argue that such summary statistics can obscure critical differences in model behavior, especially on hard-to-predict or chemically ambiguous samples.

   To better highlight these differences, we focused on the subset of disagreement cases, i.e., samples where the two models (KANO vs. MotiL) give different predictions. As shown in Figure 7a, these disagreement sets account for 28.1% of the BBBP dataset and 21.4% of the SIDER dataset. Within these subsets:

   (1) On BBBP, MotiL is correct in 40 out of 57 disagreement cases (70.2%), while KANO is correct in only 17 (29.8%).

   (2) On SIDER, MotiL is correct in 43 out of 63 disagreement cases (68.3%), compared to 20 (31.7%) for

KANO.

The above results demonstrate that MotiL **tends to make the right decision when the models disagree,** implying stronger reliability in high-uncertainty or decision-boundary situations. Also, we provide examples in Figure 7c–d, where MotiL correctly predicts toxicity or permeability for drugs like Streptomycin, Mexazolam, Erlotinib, and Vincristine, which KANO misclassifies.

2. **Activity cliff evaluation – experimental setup and findings**: We fully agree with the reviewer that evaluating on activity cliffs provides a more rigorous and chemically meaningful comparison. Following your suggestion, we designed a specific **activity cliff evaluation protocol** as follows:
   (1) From the BBBP and SIDER datasets, we identified molecular pairs with high Tanimoto similarity (>0.85) but significantly different labels.
   (2) We selected individual molecules involved in such pairs to construct the activity cliff test sets.
   (3) For each dataset, we analyzed prediction disagreements between KANO and MotiL on these challenging samples.

   As shown in Figure 7b, on BBBP, among 15 disagreement cases, MotiL correctly predicted 10 (66.7%), while KANO only predicted 5 (33.3%); on SIDER, among 19 disagreement cases, MotiL predicted 11 correctly (57.9%), compared to 8 (42.1%) for KANO.

   Moreover, Figures 8 and 9 provide concrete activity cliff examples, including: pairs with Tanimoto similarity of 0.85–1.0, with minor differences in functional groups; cases where KANO fails to distinguish structurally similar molecules with different activities, while MotiL succeeds in capturing subtle motif-level distinctions in most cases.

   These results suggest that the motif-level representation learning and diffusion-based denoising mechanism in MotiL better capture fine-grained chemical variations, enabling the model to distinguish subtle functional group differences that often drive activity cliffs. It highlights MotiL's stronger inductive bias toward chemically meaningful substructures.
3. **Clarifying the novelty of MotiL beyond KANO**: While KANO performs well and incorporates chemically motivated augmentations, MotiL offers a conceptually distinct and more expressive framework:
   (1) Motif-level contrastive alignment: MotiL explicitly identifies and aligns representations of functional groups across molecules, but KANO does not support it.
   (2) Bi-scaled training: MotiL simultaneously aligns representations at both the graph and motif level, capturing both global molecular semantics and local substructure invariants.
   (3) Diffusion-based pretraining (DiffMoM): The integration of a denoising diffusion model improves robustness to structural perturbations, a key weakness in most augmentation-based contrastive models including KANO.

   These novel contributions allow MotiL to better generalize in **noisy, or structurally ambiguous regimes, including activity cliffs**.
4. **Revision summary in the manuscript**:
   (1) We revised Figure 4 in the main text to summarize prediction disagreements on the full datasets (BBBP and SIDER). We have added the following text to the Section "MotiL corrects the erroneous prediction of the strongest baseline" to introduce the **activity cliff experiment** and its corresponding results:

   "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."

(2) The specific setting and case study in the activity cliff experiment have been added to Supplementary Section L (Activity cliff evaluation):

Section 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."

Section L.2 Experimental results and case study: "As shown in Fig. 4b (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 (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 (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 = 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."

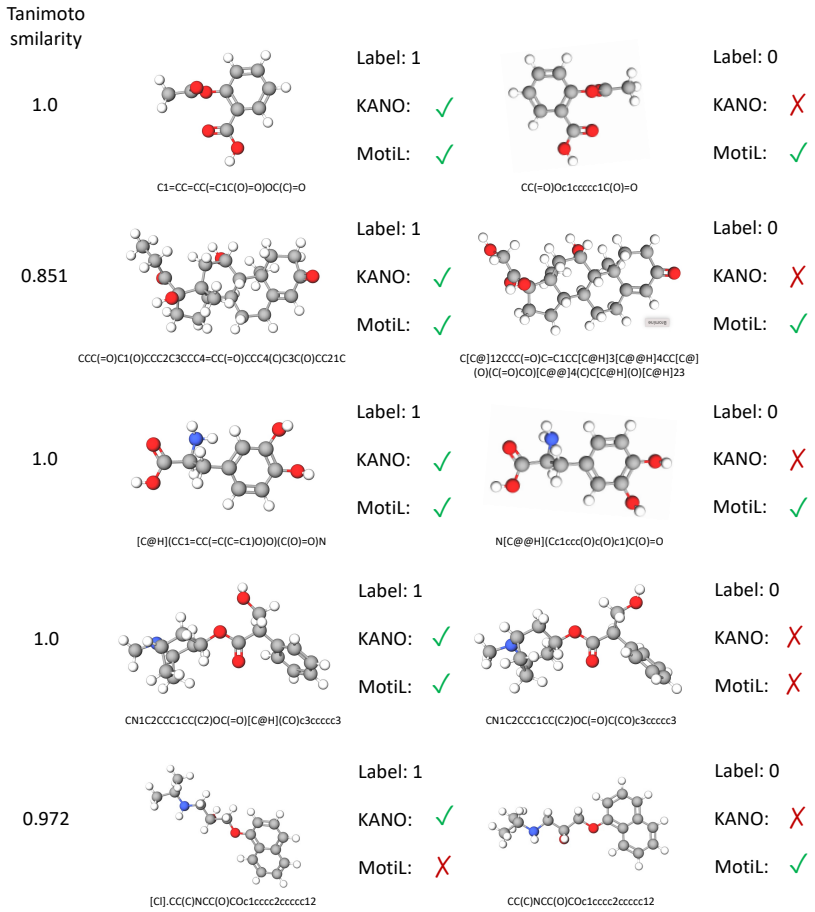


Figure 8: Prediction results for activity cliff samples on BBBP.

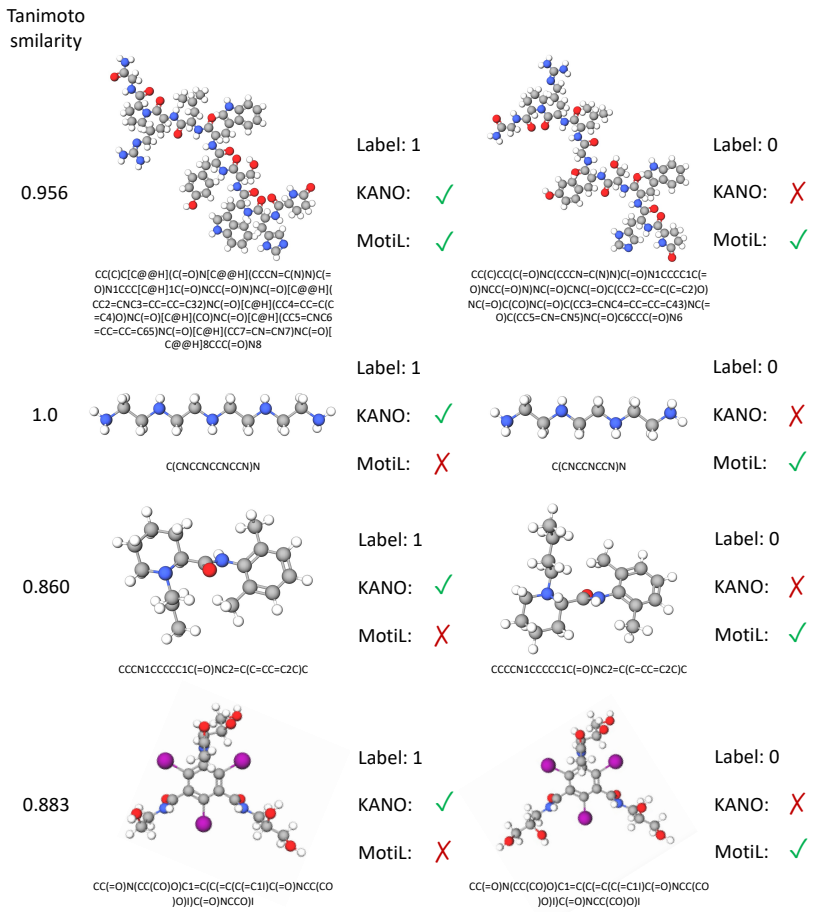


Figure 9: Prediction results for activity cliff samples on SIDER.

Comment 10: I appreciate the author's commitment to improving the model in future work, with already some defined future steps that could further accelerate drug discovery.

Reply:

We sincerely thank the reviewer for their supportive and encouraging feedback. We are pleased that our discussion of future work was found meaningful and aligned with the overarching goal of accelerating drug discovery through machine learning.

As outlined in the manuscript, we see multiple promising avenues for extending our work with MotiL, many of which are aimed at addressing critical challenges in molecular modeling and real-world drug development pipelines. Below, we elaborate on these directions with more technical depth:

1. **Advancing the diffusion priming phase with more expressive generative models**: In the current version of MotiL, the diffusion priming phase utilizes a graph-based generative denoising model (DiffMoM) to improve the encoder's robustness and capture chemically consistent structures. However, more expressive generative models, such as those incorporating equivariant neural networks or continuous-time score-based generative models, may significantly enhance the structural fidelity and diversity of the diffusion process. This improvement could lead to: (1) finer control over local motif-level perturbations, (2) stronger structural priors embedded in representations, and 3) better generalization to novel chemical scaffolds.
2. **Hybridizing representation learning and molecule generation**: We are particularly excited about bridging the gap between representation learning and molecular generation. Given that MotiL captures semantically meaningful motif-level representations, one natural next step is to employ these embeddings as building blocks

in a motif-conditioned generative framework. Such a hybrid approach could: (1) facilitate the design of novel molecules with desired pharmacological properties, (2) enable property-guided optimization (e.g., minimizing toxicity while preserving activity), and (3) support de novo scaffold generation with biologically relevant substructures. This direction aligns with our vision of using representation learning not only for predictive modeling, but also for proactive design and exploration in chemical space.

3. **Accelerating drug discovery and personalized medicine**: By extending MotiL's capabilities into the molecule design space, we foresee broader applications in: (1) targeted therapy development (e.g., designing ligands for specific protein targets), (2) personalized medicine (e.g., generating patient-specific drug variants based on molecular fingerprints or genomic data), and (3) synthetic biology and diagnostic tool design (e.g., engineering molecules with specific reactivity or detectability). The ability to learn from known functional motifs and generalize to unseen yet plausible bioactive compounds may uncover treatments for unmet clinical needs and drive more efficient lead identification, optimization, and validation.

We thank the reviewer once again for acknowledging the importance of these directions. We are committed to advancing MotiL in a way that not only improves learning algorithms but also contributes to **translational impact in biomedical applications**.

Comment 11: Finally, I recommend revising the references. For example, reference [1], Diaz, N. et al. Discovery of potent small molecule inhibitors of lipoprotein (a) formation. Nature 1–6 (2024), it's written on the journal's website as Diaz, N. et al. Discovery of potent small-molecule inhibitors of lipoprotein(a) formation. Nature 629, 945–950 (2024). Reference [7] Wan, Z. et al. Unconventional superconductivity in chiral molecule-tas2 hybrid superlattices. Nature 1–6 (2024) it's written on the journal's website as Wan, Z, et al. Unconventional superconductivity in chiral molecule–TaS2 hybrid superlattices. Nature 632, 69–74 (2024).

Given the increasing importance of ML in advancing drug discovery and the innovative methodology presented in this work, I recommend this article be accepted for revision.

Reply:

We thank the reviewer for carefully identifying inaccuracies in the formatting of our references, and we fully agree that accurate and properly formatted citations are essential for maintaining scientific rigor and clarity. In response, we have conducted a comprehensive revision of all references in our manuscript to ensure correctness, consistency, and compliance with official journal formats.

1. **Specific corrections made**: We have carefully updated the two references mentioned by the reviewer based on their official records from the Nature website: Reference [1] was previously cited as: *"Diaz, N. et al. Discovery of potent small molecule inhibitors of lipoprotein (a) formation. Nature 1–6 (2024)."* It has been revised to: *"Diaz, N. et al. Discovery of potent small-molecule inhibitors of lipoprotein(a) formation. Nature 629, 945–950 (2024)."* Key corrections include: (1) hyphenation of "small-molecule" per journal usage; (2) parenthetical formatting of "lipoprotein(a)" with no added space; (3) insertion of correct volume number (629) and page range (945–950) as published. Reference [7] was previously cited as: *"Wan, Z. et al. Unconventional superconductivity in chiral molecule-tas2 hybrid superlattices. Nature 1–6 (2024)."* It has been revised to: *"Wan, Z. et al. Unconventional superconductivity in chiral molecule–TaS_2 hybrid superlattices. Nature 632, 69–74 (2024)."* Key corrections include: (1) correct casing and formatting of the compound "TaS_2", with a subscript "2"; (2) insertion of official volume number (632) and page numbers (69–74).
2. **Global reference audit**: In addition to the two references explicitly pointed out, we took this opportunity to systematically review all references (total of 21 references in the main text and 23 references in the appendix). We made the following corrections where necessary:
   (1) Ensured that volume numbers and page ranges were accurately included for all journal articles. For

example:

a) *"Du, Y. et al. Machine learning-aided generative molecular design. Nature Machine Intelligence 1–16 (2024)." -> "Du, Y. et al. Machine learning-aided generative molecular design. Nature Machine Intelligence 6, 589–604 (2024)."*

b) "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 (2019)." -> "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)."

(2) Corrected any title formatting inconsistencies. For example:

a) "Hermosilla, P. & Ropinski, T. Contrastive representation learning for 3d protein structures. arXiv preprint arXiv:2205.15675 (2022)." -> "Hermosilla, P. & Ropinski, T. Contrastive representation learning for 3D protein structures. arXiv preprint arXiv:2205.15675 (2022)."

b) "Qiang, B. et al. Coarse-to-fine: a hierarchical diffusion model for molecule generation in 3d. International Conference on Machine Learning 28277–28299 (2023)." -> "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)."

(3) Standardized journal or conference titles to match official publication style. For example:

a) "Nat. Mach. Intell." -> "Nature Machine Intelligence"

b) "Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual" -> "Advances in Neural Information Processing Systems 34"

c) *"Bioinform." -> "Bioinformatics"*

Finally, we sincerely thank the reviewer for the positive assessment of our work and for recognizing both the relevance of machine learning to modern drug discovery and the novelty of our proposed methodology. We greatly appreciate your recommendation for revision and have carefully addressed all comments and suggestions to further improve the clarity, rigor, and completeness of the manuscript. Your feedback has been instrumental in helping us refine our contributions and better position our work for impact in this rapidly evolving field.

Comment 12 (Remarks on code availability):

-The library versions used are indicated

- Databases used for small molecules are available on git-hub. However, for the MUV, QM7, and ToxCast, the #molecules written on the article don't appear to match the CSV file lengths.

- I couldn't access the databases for macromolecules, I was unable to access the Baidu Cloud.

- The pretrained models for small molecule predictions are available on github providing a readme file with a clear description of MotiL, requirements.txt, and instructions on how to run MotiL

- I couldn't access the pretrained models for macromolecule predictions, I was unable to access the Baidu Cloud

- I used codeocean website to run the available code and the example of the bbbp database led to the same output file (Results: 0.97404 +/- 0.00289). It was reproducible, but was it supposed to match Table 1?

Reply:

We sincerely thank the reviewer for their careful inspection of our codebase, data resources, and reproducibility pipeline, as well as for the positive feedback on several key aspects. Below, we address each point in detail:

1. **Typographical errors in molecule counts for MUV, QM7, and ToxCast**: We appreciate the reviewer's attention to the molecule counts for the small molecule datasets MUV, QM7, and ToxCast. After thorough

verification, we confirm that the CSV files hosted on GitHub are correct. The molecule numbers reported in the manuscript contained typographical errors introduced during manuscript preparation. Specifically:

(1) MUV: The actual number of molecules in the dataset is 93,087.

(2) QM7: The correct number is 6,830.

(3) ToxCast: The correct number is 8,597.

We have corrected these numbers in both the main paper (Table 1) and in Supplementary Table 1 to match the true values in the released data files.

2. **Macromolecule dataset and pretrained model checkpoints inaccessible due to Baidu Cloud restrictions**: We acknowledge the reviewer's difficulty in accessing the macromolecular datasets and pretrained models previously hosted on Baidu Cloud, which may be inaccessible in certain regions. To address this and ensure global accessibility without registration or regional limitations, we have made the following improvements:

(1) Re-hosted all macromolecule benchmark datasets on Hugging Face Hub, which is widely accessible internationally and does not require regional authentication. The datasets are available at: https://huggingface.co/datasets/Young0222/MotiL/tree/main.

(2) Uploaded the pretrained model checkpoints for macromolecule prediction tasks to the same Hugging Face repository.

(3) Updated the relevant links in our GitHub repository (e.g., the README file as shown in Figure 10) to reflect this new hosting location.

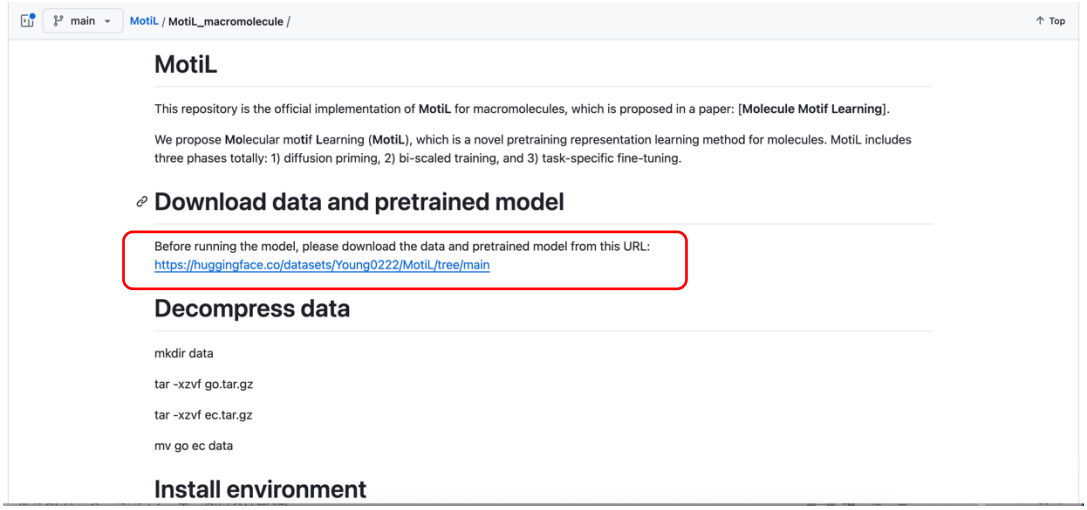


Figure 10: User instructions for macromolecule tasks provided in the README file on GitHub.

The Hugging Face repository now contains all the necessary files to reproduce the experiments in the macromolecule benchmarks, and the pretrained models can be freely accessed worldwide without geographical or technical restrictions.

3. **BBBP result reproduced via Code Ocean**: We have publicly provided the **full training logs** to ensure reproducibility. They are available at: https://github.com/Young0222/MotiL. The BBBP result you obtained on Code Ocean (0.97404 ± 0.00289) is indeed slightly higher than the result reported in our paper (0.969 ± 0.003, Supplementary Table 3). This variation is primarily due to **differences in hardware and software environments** between Code Ocean and our local training setup. Specifically, as shown in Table 5:

Table 5: Environment setup comparison between Code Ocean and our server.

Environment	Our server	Code Ocean
Python	3.7.13	3.8.5
PyTorch	1.12.1	1.13.0
CUDA	11.6	11.7
cuDNN	8302	8500
GPU	RTX 3090	Tesla T4
CUDA driver	12.8	12.0

In this table, for example:

(1) GPU hardware architecture differences: The Tesla T4 (Turing architecture) and RTX 3090 (Ampere architecture) differ significantly in their numerical precision handling, Tensor Core capabilities, and support for mixed-precision computation (float16 vs. float32). The T4 is optimized for efficiency and low-power inference tasks, with a stronger emphasis on float16 computation. These architectural differences can introduce greater numerical drift during training, potentially leading to different convergence paths even under identical seed settings.

(2) PyTorch and cuDNN version differences: PyTorch 1.12.1 and 1.13.0 introduce subtle behavioral differences, particularly in components such as LayerNorm, Dropout, and optimizer initialization (e.g., AdamW). Additionally, cuDNN version 8.3 vs. 8.5 may alter internal precision vs. speed trade-offs for convolution and normalization layers. Some operators in newer versions may favor more aggressive kernel implementations, which can affect numerical stability, especially in smaller datasets or edge-case learning scenarios.

These differences, particularly in GPU architecture and software versions, can introduce slight variations in training dynamics, which lead to minor differences in final results. **Code Ocean recommends using their default environment configuration, and we adhered to that requirement during submission**. To further ensure that our reported results can be exactly reproduced, we have uploaded the trained models to Code Ocean. You can run the prediction step directly with these models and **obtain the same results as reported in our paper** (see Figure 11). The Code Ocean link is: https://codeocean.com/capsule/9672051/tree. We appreciate your attention to reproducibility and hope this clarification resolves your concerns.

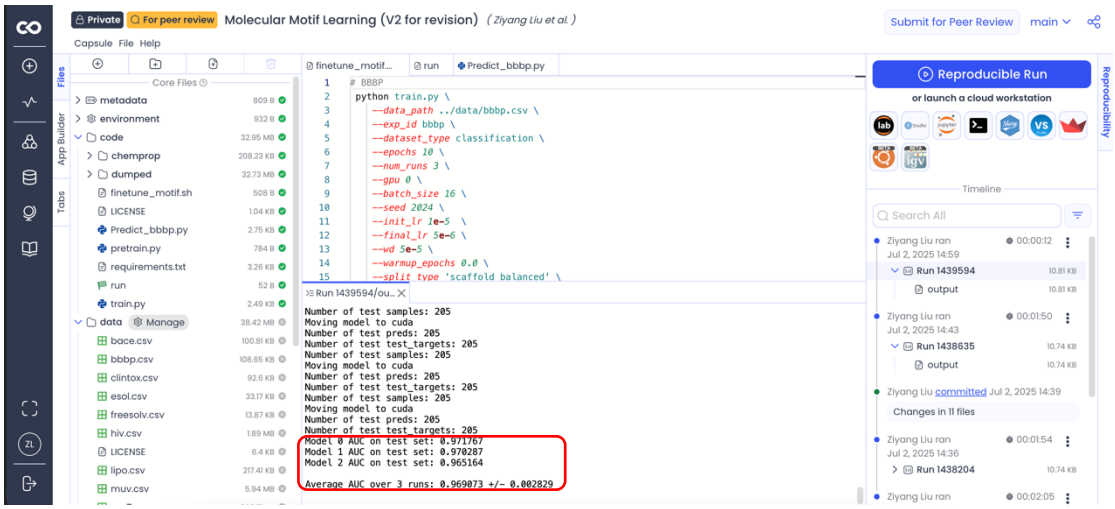


Figure 11: Output results from MotiL's prediction for BBBP on Code Ocean.

4. **Code and models for small molecules – accessibility and clarity**: We are pleased that the reviewer found the resources for small molecule tasks, such as the GitHub code, pretrained models, requirements.txt, and

README instructions, to be clear, functional, and easy to follow. We put considerable effort into ensuring our repository is well-documented and reproducible, and we will continue to maintain and improve the codebase for community use.

We have correspondingly revised the Sections of Data availability and Code availability in the main text:

"The pretraining data and molecular property prediction benchmarks used in this work are available on the Hugging Face Hub at https://huggingface.co/datasets/Young0222/MotiL/tree/main *and the GitHub repository at https://github.com/Young0222/MotiL."*

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