Improving Molecular Property Prediction via Topology-Enhanced Chemical Language Model

Anonymous ACL submission

Abstract

Pre-trained chemical language models (CLMs) excel in the field of molecular property predictions, utilizing string-based molecular descriptors such as SMILES for learning universal representations. However, the one-dimensional format of SMILES can impede the effectiveness of the model because it lacks the topological information necessary for accurate property predictions. In this work, we introduce HINT, a novel framework to enhance the understanding of molecular structures within CLMs with topological fingerprints. HINT enhances molecular representations of CLMs through a molecular substructure prediction task and fingerprint-based contrastive learning. Experimental results on various tasks verify that HINT significantly improves the molecular property prediction performance of CLMs[1].

1 Introduction

In the realms of drug discovery and materials science, the application of deep neural networks to molecular property prediction is increasingly recognized as valuable (Butler et al., 2018). Recently, inspired by the success of the pre-trained language models (Devlin et al., 2019; Liu et al., 2019), chemical language models (CLMs) have been introduced and shown their proficiency in predicting molecular properties (Wang et al., 2019; Honda et al., 2019; Chithrananda et al., 2020; Fabian et al., 2020; Ahmad et al., 2022; Ross et al., 2022). These CLMs are trained on large-scale string-based molecular descriptors to learn universal molecular representations. However, one-dimensional descriptors such as Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, 1988) fall short in providing topological information (Soares et al., 2023; Yüksel et al., 2023). Thus, CLMs trained on SMILES suffer from capturing the relationships between molecular structures and properties (Graff et al., 2023).

In this work, we introduce HINT (*en**H**ancing topological **I**nformation with co**NT**rastive learning*), a novel framework to enhance the topological understanding of CLMs. HINT leverages structural information contained in topological fingerprints, notably Extended-Connectivity Fingerprints (ECFPs) (Rogers and Hahn, 2010), to address the limitation of SMILES. HINT continuously trains pre-trained CLMs with multiple tasks: molecular substructure prediction and topological fingerprint-based contrastive learning. In the molecular substructure prediction task, HINT trains the model to predict the substructure information of molecules hashed in ECFPs. Additionally, in the contrastive learning task, the model learns the representation by contrasting structurally similar and dissimilar molecules that are identified using ECFPs.

We evaluate HINT with two strong CLMs (Ahmad et al., 2022; Ross et al., 2022) on various tasks from MoleculeNet benchmarks (Wu et al., 2018), including six classification and four regression tasks. HINT achieved performance improvements of 4.77% and 4.54% on average for each backbone, demonstrating its effectiveness in molecular property prediction.

2 Methodology

2.1 Molecular Substructure Prediction

To enhance the topological understanding of CLMs, we train the model to predict molecular substructures hashed in ECFPs. ECFPs are the fixed-length binary vectors that hash identified substructures of molecules into fixed-length binary vectors, with 1 representing the presence and 0 for the absence of certain substructures. Through the prediction of ECFPs, the model acquires the capability to detect the presence of substructures, thereby improving its understanding of the topological information of

[1] Our code is available at `https://anonymous.4open.science/r/HINT-0C2D`

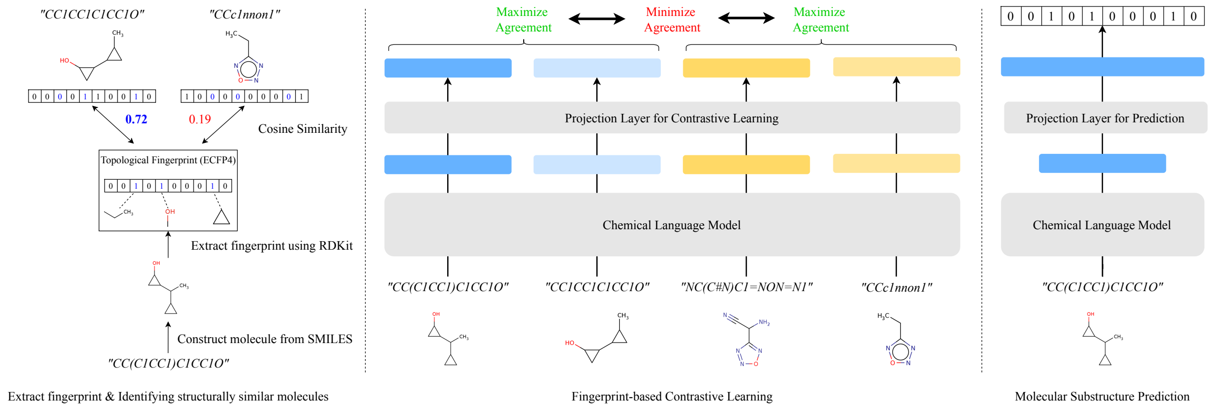

Figure 1: Illustration of HINT. We extract and construct a set of top-k similar molecules by measuring cosine similarity among topological fingerprints. We then predict ECFP4 directly and perform contrastive learning to maximize the agreement between pairs of structurally similar molecules.

molecules.

Specifically, we first extract 2048-dimensional ECFP4 fingerprints from each molecule using RDKit[2]. We then project the molecular representation to match the dimensions of the ECFP4 fingerprint, facilitating the prediction of hashed substructures within it. The molecular representation is obtained by extracting final hidden representation of first token from the CLM. Subsequently, we employ Binary Cross Entropy (BCE) loss to define the substructure prediction tasks.

2.2 Fingerprint-based Contrastive Learning

While molecular substructure prediction effectively incorporates topological information, it may not fully address the challenge of comprehending how these structures correlate with molecular properties. Understanding such relationships is crucial for accurately predicting functional outcomes, such as reactivity, stability, and biological activity (Le et al., 2012).

Hence, we introduce a simple contrastive learning method based on topological fingerprints to further enhance CLMs. This method is rooted in the insight that molecules with similar structures often exhibit similar properties (Martin et al., 2002). HINT trains models to distinguish between structurally similar and dissimilar molecules in a contrastive manner. This approach is expected to facilitate the model's ability to determine properties by recognizing structural differences in molecules.

We first create a set of structurally similar molecules, denoted as H, for each molecule in the dataset. This process involves utilizing the ECFP4 vectors extracted from the molecules. By calculating the cosine similarity between these vectors, we are able to identify the top-k similar molecules. Subsequently, we sample a batch of N molecules and define the contrastive prediction task on pairs of similar molecules. For each molecule in a batch, we randomly select a similar molecule from H to form the positive pair, resulting in $2N$ data points.

We then define the agreement between molecule m and sampled molecule s as follows:

$$\sigma(m, s) = \exp(\text{sim}(M, S)/\tau), \qquad (1)$$

where M and S refer to the molecular representations of m and s, respectively. The τ is the temperature parameter for scaling. We employ the NT-Logistic loss function (Chen et al., 2020) to maximize agreement between positive pairs while minimizing agreement between negative pairs. Instead of explicitly sampling negative examples, we treat the other $2(N-1)$ molecules in a batch as negative examples. The fingerprint-based contrastive loss is as follows:

$$\mathcal{L}_{CL}(m_p, s_p) = -\log \frac{\sigma(m_p, s_p)}{\sum_{i=1}^{2N-1} \sigma(m_i, s_p)}. \qquad (2)$$

Our final objective function is expressed as follows:

$$\mathcal{L}(m_p, s_p) = \mathcal{L}_{BCE}(m_p) + \lambda \mathcal{L}_{CL}(m_p, s_p), \qquad (3)$$

where λ is a non-negative hyper-parameter for balancing the objective functions. To ensure accuracy in learning, contrastive learning is omitted for molecules that are not unique, specifically when there are more than two similar molecules within a batch for a particular molecule.

[2] https://www.rdkit.org

		BBBP ROC	Tox21 ROC	ClinTox ROC	HIV ROC	BACE ROC	SIDER ROC	QM9 MAE	ESOL RMSE	FreeSolv RMSE	Lipop RMSE
Graph	D-MPNN (Yang et al., 2019)	71.2	68.9	90.5	75.0	85.3	63.2	-	0.980	2.180	0.660
	GeomGCL (Li et al., 2022)	-	85.0	91.9	-	-	64.8	-	0.575	0.866	0.541
	MolCLR (Wang et al., 2022)	73.6	79.8	93.2	80.6	**89.0**	68.0	-	1.110	2.200	0.650
	HiMol (Zang et al., 2023)	73.2	76.2	73.7	-	84.6	62.5	3.243	0.833	2.283	0.708
Text	MolBERT (Fabian et al., 2020)	76.2	-	-	78.3	86.6	-	-	0.531	0.948	0.561
	SELFormer (Yüksel et al., 2023)	90.2	65.3	-	68.1	83.2	**74.5**	-	0.682	2.797	0.735
	ChemBERTa-2 (Chithrananda et al., 2020)	70.1	48.1	51.9	74.7	80.9	49.0	2.775	0.949	1.854	0.728
	MoLFormer-XL (Ross et al., 2022)	91.5	84.5	**94.6**	81.3	86.7	65.7	1.628	0.248	0.315	0.518
	HINT$_C$	71.4	49.9	53.5	75.2	82.8	50.9	2.541	0.811	1.806	0.705
	HINT$_M$	**92.4**	**85.4**	94.0	**84.2**	88.7	66.3	**1.445**	**0.212**	**0.301**	**0.508**

Table 1: Main experimental results. **Bold** and Underline indicates best and second-best results, respectively.

	α	C_v	G	gap	H	$\varepsilon_{\text{homo}}$	$\varepsilon_{\text{lumo}}$	μ	$\langle R^2 \rangle$	U_0	U	ZPVE
ChemBERTa-2	0.5164	0.2026	1.2027	0.0057	1.0156	0.0040	0.0041	0.5260	27.3141	1.2618	1.1933	0.0010
MoLFormer-XL	0.3531	0.1594	0.2826	0.0040	0.2864	0.0041	0.0040	0.3691	17.2684	0.4758	0.3291	0.0004
HINT$_C$	0.5051	0.1979	1.2765	0.0055	1.0731	0.0039	0.0041	0.5200	24.4382	1.1632	1.0849	0.0009
HINT$_M$	**0.2786**	**0.1219**	**0.2773**	**0.0033**	**0.2203**	**0.0024**	**0.0028**	**0.3501**	**15.4922**	**0.2961**	**0.2936**	**0.0002**

Table 2: Experimental results for QM9 subtasks.

3 Experimental Settings

Datasets. To evaluate molecular property prediction ability of CLMs, we conduct the experiments on six classification[3] and four regression tasks[4] from the MoleculeNet benchmark (Wu et al., 2018). For evaluation metrics, we report AUC-ROC for classification, MAE for QM9, and RMSE for remaining regression tasks. Task descriptions can be found in Tables 11 and 12 in Appendix.

Training Setup. We use the dataset for each task to train ChemBERTA-2 (Ahmad et al., 2022) and MoLformer-XL (Ross et al., 2022) with HINT framework, naming them HINT$_C$ and HINT$_M$, respectively. We then fine-tune the model on each task. Additionally, we provide the performance of two models without HINT for comparison. Further details are in Appendix B.

4 Experimental Results

Main Results. Table 1 presents our experimental results. Our HINT$_C$ and HINT$_M$ show performance improvements of 4.77% and 4.54% on average for each backbone. Especially, HINT$_M$ surpasses existing CLMs on eight tasks. It also achieves comparable performance on the ClinTox and SIDER datasets, demonstrating its versatility in molecule property prediction.

Among the regression tasks, the QM9 task involves predicting quantum chemical properties, which is particularly challenging in the absence of 3D geometric information. Despite this, HINT$_M$ achieves consistent improvements in performance on the QM9 dataset compared to its baseline. Overall results of QM9 subtasks are shown in Table 2. These results demonstrate the HINT's ability to effectively leverage molecular structures, enhancing prediction accuracy across various chemical properties. For detailed insights, see Appendix C.

[3]BBBP, ClinTox, SIDER, Tox21, HIV, and BACE

[4]QM9, ESOL, FreeSolv, and Lipophilicity (Lipop)

	Source	ESOL	FreeSolv	Lipop
HINT$_M$	QM9	0.236	0.307	0.510
	ESOL	**0.212**	0.328	0.518
	FreeSolv	0.232	**0.301**	0.525
	Lipop	0.228	0.340	**0.508**
	None	0.248	0.315	0.518

Table 3: Evaluation of the transfer of topological information. Source refers to the dataset used to train HINT. Results with None refer to fine-tuning without HINT.

Transferring Topological Information. We evaluate the generalizability of molecular representations obtained by HINT. By training the HINT framework on three different regression tasks, we cross-evaluate each model with unseen data. The results in Table 3 often show improved performance across these tasks, especially for HINT with QM9. This highlights the capability of HINT to effectively transfer topological information, confirming its wide applicability and robustness in boosting performance across various regression tasks.

Topological Analysis. Following Ross et al. (2022), we evaluate the encapsulated topological information of HINT$_M$ by analyzing the resemblance between molecular structures and the attention matrices. We calculate the cosine similarities between average pooled attention matrices and molecular structures. To facilitate this, we randomly select 3,000 molecules from QM9, PubChem (Kim et al.,

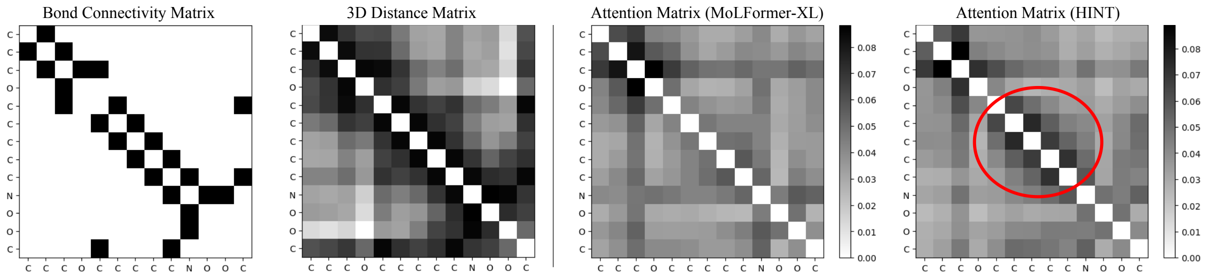

Figure 2: Visualization of attention matrices from MoLFormer-XL and HINT$_M$, accompanied by the corresponding molecular structure for ‘CC[C](O)C1CCCC([N+](=O)[O-])C1’ (ZINC001560407707).

	QM9		PubChem		ZINC	
	Bond	Dist.	Bond	Dist.	Bond	Dist.
MoLFormer-XL	60.99	85.73	45.18	79.68	44.11	77.17
HINT$_M$	**62.27**	**87.44**	**45.76**	**80.67**	**44.31**	**78.89**

Table 4: Evaluation of encapsulated topological information. We use HINT$_M$ trained on QM9 dataset.

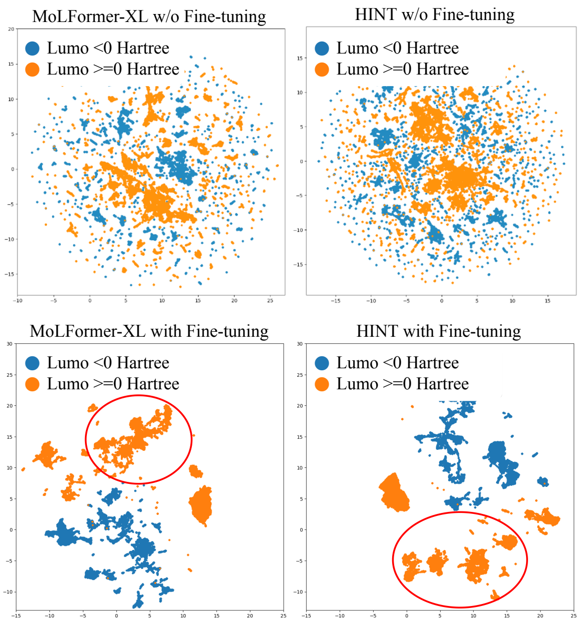

Figure 3: Visualization of embeddings from each model. We use HINT$_M$ trained on QM9 dataset targeted $\varepsilon_{\mathbf{lumo}}$.

2019), and ZINC (Irwin et al., 2012) datasets and extract bond connectivity and 3D distance matrices using RDKit. The results in Table 4 and Figure 2 indicate that HINT can effectively enhance the capability of identifying molecular structures. More examples can be found in Figure 5 and 6 in the Appendix.

Visualization of Molecular Representations. We perform a qualitative analysis by visualizing molecular representations from MoLFormer-XL and HINT$_M$ using the QM9 dataset. Dimensionality is reduced via UMAP (McInnes et al., 2018).

	FCL	MSP	ESOL	FreeSolv	Lipop
HINT$_M$	✓	✓	**0.212**	**0.301**	**0.508**
	✓	-	0.220	0.315	0.524
	-	✓	0.240	0.334	0.526
	-	-	0.248	0.315	0.518

Table 5: Ablation study results. FCL and MSP refer to fingerprint-based contrastive learning and molecular substructure prediction, respectively.

The visualization in Figure 3 indicates minor differences between the two models without fine-tuning. Nonetheless, HINT$_M$ with fine-tuning demonstrates a finer distinction among molecules, proving its ability to differentiate molecules while preserving pre-trained representations. Additional examples are in Figure 4 in the Appendix.

Ablation Study. To assess the distinct contributions of HINT’s components to its enhanced performance, we conduct ablation studies on three regression tasks with HINT$_M$, detailed in Table 5. These demonstrate that the integration of the two objective functions offers advantages over employing either method in isolation. Furthermore, using our contrastive learning method alone resulted in performance gains on ESOL and FreeSolv. This finding implies that understanding the relationships among molecules facilitates the effective integration of topological information.

5 Conclusion

We have introduced HINT, a novel framework that enhances the topological understanding of CLMs to improve property prediction. To do so, HINT continually trains CLMs to predict the molecular substructures and contrast structurally similar and dissimilar molecules. Experimental results have shown that our model better captures topological information of molecules than baselines. Consequently, HINT significantly improves the prediction performance of CLMs on extensive tasks.

Limitations

While we have demonstrated the effectiveness of HINT, a few limitations exist. First, our method for identifying similar molecules leads to quadratic computational complexity $O(N^2)$ as we discussed in Appendix D. Due to this limitation, we utilize relatively small-scale datasets for the HINT framework (<200K) compared to pre-training datasets (>1B). To enable the application of the HINT to large-scale datasets, we will explore the efficient algorithms for identifying similar molecules.

Second, we leave the application of HINT to the state-of-the-art model remains as future work. Due to the unavailability of accessing the full version MoLFormer-XL, our experiments were instead performed with a variant trained on 10% of the pre-training dataset (1.2B) as if MoLFormer-XL. Nevertheless, we have achieved similar or even better performance on many tasks with this variant model using HINT, compared to the full model. Therefore, we believe that HINT will also be effective on the state-of-the-art models based on our comprehensive experimental results.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 1597–1607.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

David E. Graff, Edward O. Pyzer-Knapp, Kirk E. Jordan, Eugene I. Shakhnovich, and Connor W. Coley. 2023. Evaluating the roughness of structure–property relationships using pretrained molecular representations. *Digital Discovery*, 2:1452–1460.

Shion Honda, Shoi Shi, and Hiroki R. Ueda. 2019. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. 2012. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 5156–5165.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan Bolton. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, 47(Database-Issue):D1102–D1109.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. 2022. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588.

Tu Le, V Chandana Epa, Frank R Burden, and David A Winkler. 2012. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews*, 112(5):2889–2919.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. 2022. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual*

Event, February 22 - March 1, 2022, pages 4541–4549.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. 2019. Provably powerful graph networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2153–2164.

Yvonne C Martin, James L Kofron, and Linda M Traphagen. 2002. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mac. Intell.*, 4(12):1256–1264.

Eduardo Soares, Emilio Vital Brazil, Karen Fiorela Aquino Gutierrez, Renato F. G. Cerqueira, Daniel P. Sanders, Kristin Schmidt, and Dmitry Yu. Zubarev. 2023. Beyond chemical language: A multimodal approach to enhance molecular property prediction. *arXiv preprint arXiv:2306.14919*.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Roberto Todeschini and Viviana Consonni. 2010. Molecular descriptors. *Recent Advances in QSAR Studies*, pages 29–102.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019, Niagara Falls, NY, USA, September 7-10, 2019*, pages 429–436.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor W. Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi S. Jaakkola, Klavs F. Jensen, and Regina Barzilay. 2019. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, Gamze Deniz, and Tunca Dogan. 2023. Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. 2023. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34.

Appendix

In this section, we supplement our main content with additional experiments and analysis.

A Related Work

Topological fingerprints, such as ECFPs, have been introduced to encode molecules into binary vectors using rule-based algorithms (Todeschini and Consonni, 2010; Rogers and Hahn, 2010). Earlier machine learning approaches employed neural networks trained on fingerprints in supervised settings for predicting molecular properties.

Recent advancements in natural language processing (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019) have led to the proposal of pre-trained Chemical Language Models (CLMs) trained with textual notations of molecules. ChemBERTa (Chithrananda et al., 2020), a transformer-based model trained on SMILES for molecular property prediction, has demonstrated enhanced predictive capability with masked language modeling (Devlin et al., 2019). Ross et al. (2022) have introduced MoLFormer-XL, which incorporates rotary position embedding (Su et al., 2024) and linear attention (Katharopoulos et al., 2020) with 1.2 billion chemical strings, showcasing superior performance on molecular predictions with scaled-up pre-training data.

Furthermore, efforts have been made to address the fact that SMILES is considered less topologically aware compared to graph-based information. Yüksel et al. (2023) have introduced SELFormer, a CLM based on SELFIES (Krenn et al., 2022), aimed at learning robust molecular representations. Soares et al. (2023) have demonstrated that combining CLMs with physiochemical features improved property predictions. In our work, we focus on incorporating topological information using topological fingerprints into existing transformer-based CLMs.

B Implementation Details

We train the models with the HINT framework on each dataset of downstream tasks before fine-tuning them. For HINT$_C$, we utilize ChemBERTa-2 trained on 250k molecules form ZINC dataset (Irwin et al., 2012)[5]. HINT$_M$ is initialized with a publicly available MoLFormer-XL trained on 10% of its original pre-training dataset. The original version is trained on 1.2 billion molecules from the ZINC and PubChem dataset (Kim et al., 2019). The hyperparameter settings we used for the experiment are shown in Table 6. We attempted to determine the optimal settings for each task to report the highest scores, as shown in Tables 7 and 8. In addition to that information, we use AdamW as our optimizer and we do not apply any learning rate scheduler.

For the fine-tuning, we adhere to the recommended train, validation, and test splits from Wu et al. (2018) and closely follow the experimental settings established by each baseline (Ahmad et al., 2022; Ross et al., 2022). All experiments are conducted on two NVIDIA RTX A6000 GPUs and four NVIDIA RTX A5000 GPUs.

	HINT$_M$	HINT$_C$
Backbone	MoLFormer-XL	ChemBERTa-2
# Pram.	46M	83M
Batch Size	{32, 64, 128, 256}	
Learning Rate	{1e-5, 2e-5, 3e-5, 4e-5, 5e-5}	
λ	{0.1, 0.2, 0.3, 0.4, 0.5}	
# Mols	{5, 10, 50}	
Epoch	{10, 30, 50, 100}	

Table 6: Detailed settings for training HINT framework.

	Epochs	ESOL	FreeSolv	Lipop
HINT$_M$	100	0.234	**0.301**	0.524
	50	0.246	0.332	0.517
	30	**0.212**	0.359	0.522
	10	0.230	0.352	**0.508**
	0	0.248	0.315	0.518

Table 7: Ablation study of contrastive learning. Results with 0 epoch refer to fine-tuning without HINT.

	# Mols	ESOL	FreeSolv	Lipop
HINT$_M$	top-50	0.227	0.316	0.528
	top-10	**0.212**	0.334	0.513
	top-5	0.228	**0.301**	**0.508**
	None	0.248	0.315	0.518

Table 8: Evaluation of number of similar molecules (# Mols) for the fingerprint-based contrastive learning. Results with None refer to fine-tuning without HINT.

C Further Insights from QM9

To clarify the advantage of our HINT across different models, we present all results for the 12 subtasks of the QM9 dataset in Table 9. Comparing models with HINT to those without, models with MoLFormer-XL (M-XL) show significant improve-

[5] https://huggingface.co/seyonec/ChemBERTa_zinc250k_v2_40k

QM9	A-FP	123-gnn	DTNN	MPNN	C-2	HINT$_C$	M-XL	M-XL†	HINT$_M$
α	0.492	**0.27**	0.95	0.89	0.5164	0.5051	0.3327	0.3531	0.2768
C_v	0.252	**0.0944**	0.27	0.42	0.2026	0.1979	0.1447	0.1594	0.1219
G	0.893	**0.0469**	2.43	2.02	1.2027	1.2765	0.3362	0.2826	0.2773
gap	0.00528	0.0048	0.112	0.0066	0.0057	0.0055	0.0038	0.0040	**0.0033**
H	0.893	**0.0419**	2.43	2.02	1.0156	1.0731	0.2522	0.2864	0.2203
ε_{homo}	0.00358	0.00337	0.0038	0.00541	0.0040	0.0039	0.0029	0.0041	**0.0024**
ε_{lumo}	0.00415	0.00351	0.0051	0.00623	0.0041	0.0041	0.0027	0.0040	**0.0028**
μ	0.451	0.476	0.244	0.358	0.5260	0.5200	0.3616	0.4349	**0.3501**
$\langle R^2 \rangle$	26.839	22.90	17.00	28.5	27.3141	24.4382	17.062	17.2684	**15.4922**
U_0	0.898	**0.0427**	2.43	2.05	1.2618	1.1632	0.3211	0.4758	0.2961
U	0.893	**0.111**	2.43	2.00	1.1933	1.0849	0.2522	0.3291	0.2936
ZPVE	0.00207	**0.00019**	0.0017	0.00216	0.0010	0.0009	0.0003	0.0004	0.0002
Avg MAE	2.6355	1.9995	2.3504	3.1898	2.7754	2.5406	1.5894	1.628	**1.445**

Table 9: Evaluation results of SMILES-based methods on QM9 dataset. Baseline results are taken from (Wu et al., 2018; Xiong et al., 2019; Maron et al., 2019; Ross et al., 2022). **Bold** and Underline indicates best and second-best results, respectively. MoLFormer-XL with † refers to the model using 10% of pre-train data.

	# samples	Extraction time (sec)	Identification time (sec)
FreeSolv	642	< 1	11
ESOL	1,128	< 1	12
SIDER	1,427	< 1	12
ClinTox	1,478	< 1	12
BACE	1,513	1	11
BBBP	2,039	1	12
Lipophilicity	4,200	2	13
Tox21	7,831	3	17
HIV	41,127	24	95
QM9	133,885	44	892

Table 10: Time required for extracting ECFP4 fingerprints and identifying similar molecules. We use an NVIDIA A5000 GPU with Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz for this experiment.

ment across tasks. Additionally, HINT$_M$ outperforms the original MoLFormer-XL, even though we utilize its variant that trained on a smaller portion of the pre-training dataset. This demonstrates the efficacy of our framework. However, HINT$_C$ occasionally exhibits lower performance compared to its backbone, ChemBERTa-2 (C-2). This discrepancy could be attributed to the robustness of the model, considering ChemBERTa-2 is trained on a much smaller dataset than MoLFormer-XL. This observation suggests that models with a more robust representation of molecules benefit more from our framework. Moreover, we notice that our framework sometimes yields inferior results on a few tasks compared to models that leverage molecular graphs. Based on this observation, we propose the direct integration of graph information into CLMs as a promising direction for future research.

D Extracting Additional Features

In this work, we utilize topological fingerprints to enhance the topological understanding of CLMs. However, the process of extracting these additional features and identifying similar molecules incurs additional computational costs. Notably, we observed that identifying similar molecules is more time-consuming than the feature extraction process itself. Furthermore, as indicated in Table 10, the duration required for these operations escalates with the increase in dataset size, potentially hindering the application of the HINT framework in the pre-training phase for enhancements. This highlights the necessity for more efficient algorithms for identifying similar molecules as a pivotal consideration, aiming to streamline the application of the HINT framework and optimize pre-training efforts.

	Descriptions	# tasks	# samples
BBBP	Blood brain barrier penetration dataset	1	2,039
Tox21	Toxicity measurements on 12 different targets	12	7,831
ClinTox	Clinical trial toxicity of drugs	2	1,478
HIV	Ability of small molecules to inhibit HIV replication	1	41,127
BACE	Binding results for a set of inhibitors for $\beta-$ secretase 1	1	1,513
SIDER	Drug side effect on different organ clases	27	1,427

Table 11: Classification benchmarks from MoleculeNet.

	Descriptions	# samples
QM9	12 quantum mechanical calculations of small organic molecules with upto nine heavy atoms	133,885
ESOL	Water solubility dataset	1,128
FreeSolv	Hydration free energy of small molecules in water	642
Lipophilicity	Octanol/water distribution coefficient of molecules	4,200

Table 12: Regression benchmarks from MoleculeNet.

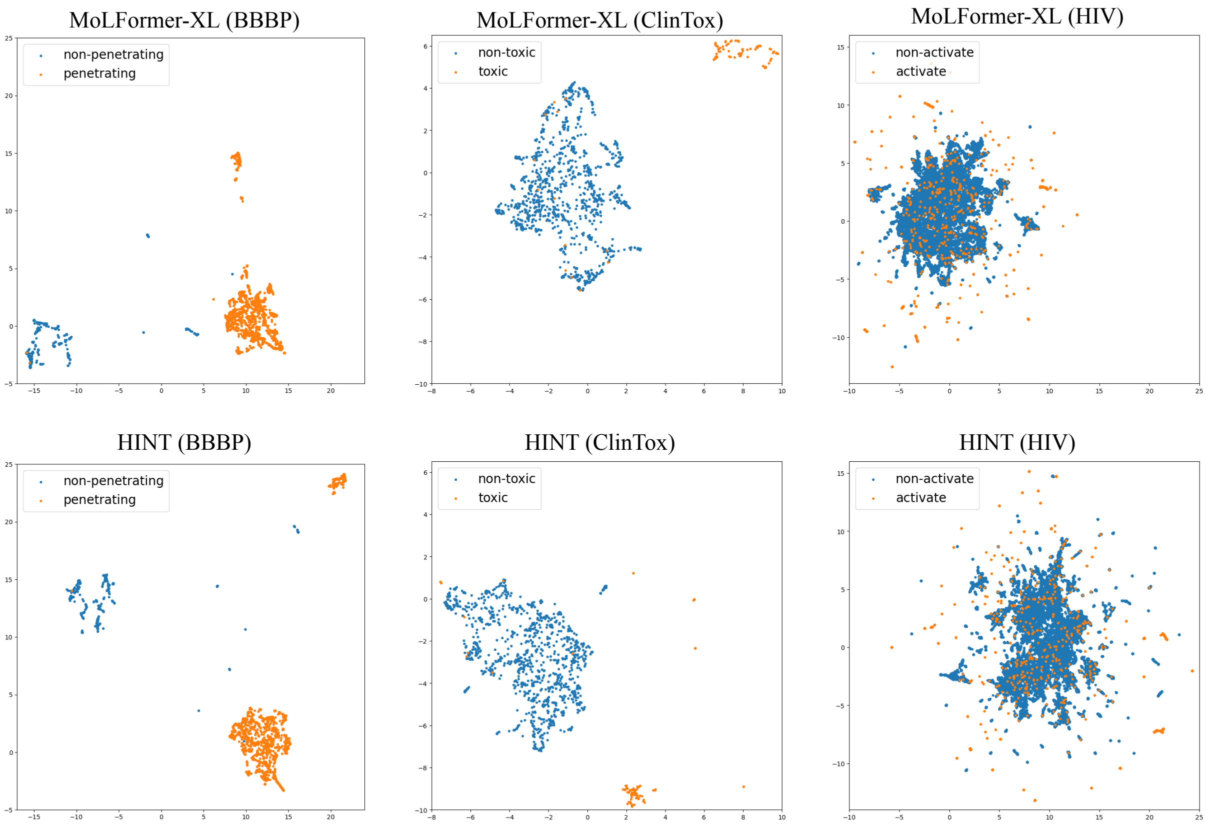

Figure 4: Visualization of embeddings of each model without fine-tuning. We use HINT$_M$ trained on QM9 dataset for this analysis. The name in the bracket refers to the dataset we use to extract embeddings.

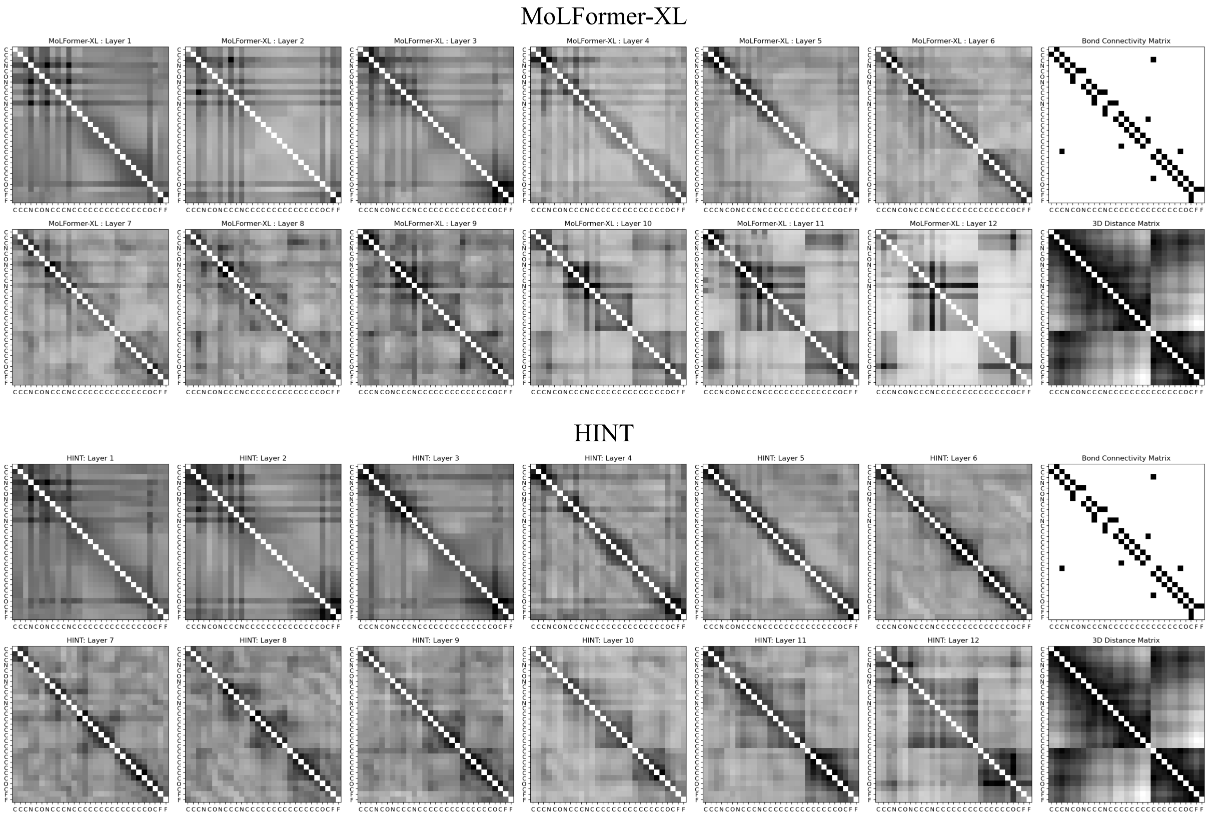

Figure 5: Visualization of attention matrices from MoLFormer-XL and HINT_M with QM9 dataset, accompanied by the corresponding molecular structure for 'CC[C@H](NC(=O)NC[C@H](C)N(C)Cc1ccccc1)c1ccccc1OC(F)F' from PubChem. Both models are not fine-tuned.

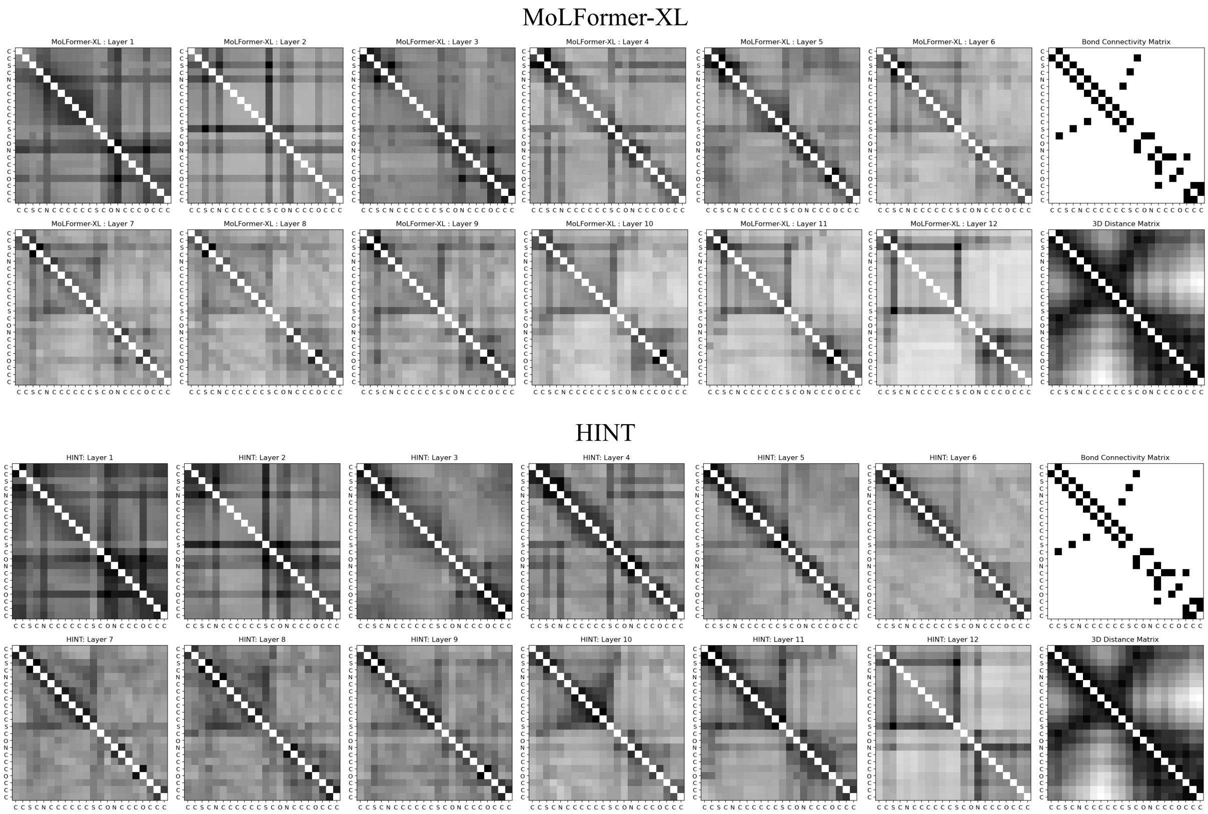

Figure 6: Visualization of attention matrices from MoLFormer-XL and HINT_M with QM9 dataset, accompanied by the corresponding molecular structure for 'CC(Sc1nc2ccccc2s1)C(=O)NC(C)(CO)C1CC1' from ZINC. Both models are not fine-tuned.