Deep Interactions for Multimodal Molecular Property Prediction

Patrick Soga
zqe3cg@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

Zhenyu Lei
vjd5zr@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

Camille Bilodeau
cur5wz@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

Jundong Li
jl6qk@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

Abstract

Multi-modal learning by means of leveraging both 2D graph and 3D point cloud information has become a prevalent method to improve model performance in molecular property prediction. However, many recent techniques focus on specific pre-training tasks such as contrastive learning, feature blending, and atom/subgraph masking in order to learn multi-modality even though design of model architecture is also impactful for both pre-training and downstream task performance. Relying on pre-training tasks to align 2D and 3D modalities lacks direct interaction which may be more effective in multimodal learning. In this work, we propose MolInteract, which takes a simple yet effective architecture-focused approach to multimodal molecule learning which addresses these challenges. MolInteract leverages an interaction layer for fusing 2D and 3D information and fostering cross-modal alignment, showing strong results using even the simplest pre-training methods such as predicting features of the 3D point cloud and 2D graph. MolInteract exceeds state-of-the-art multimodal pre-training techniques and architectures on various downstream 2D and 3D molecule property prediction benchmark tasks.

CCS Concepts

• **Applied computing** → **Chemistry**; • **Computing methodologies** → **Neural networks**.

Keywords

Multimodal learning, Molecular representation learning

ACM Reference Format:

Patrick Soga, Zhenyu Lei, Camille Bilodeau, and Jundong Li. 2018. Deep Interactions for Multimodal Molecular Property Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

AI-assisted drug discovery has driven recent research interest in utilizing neural networks for molecule learning, especially for predicting molecular properties for a variety of downstream cheminformatic tasks. Self-supervised learning (SSL) on molecular data has emerged as a prevalent research direction to achieve this, leveraging the 2D graph structures of molecules [3, 12, 17]. In parallel, many SSL strategies for 3D point cloud representations of molecules have also been developed [5]. More recent works demonstrate the effectiveness of multimodal SSL techniques which combine 2D and 3D modalities [6, 11, 14, 22]. These recent approaches consider improving molecular SSL via specific pre-training strategies and tasks, but not the underlying architecture, often resorting to using separate models for encoding 3D and 2D structures and then designing a pre-training task to align their output embeddings. Alternatively, other works take a single modality-agnostic model and task it with predicting 2D and 3D properties. Both of these approaches rely on a chosen pre-training task to align 2D and 3D views of molecules. However, it is not clear to what extent such approaches are able to fully learn cross-modal interactions. To achieve finer-grained multi-modal information with simpler pre-training tasks, we turn our focus to the role of architectural design for more effective SSL. This work introduces MolInteract, a deep learning architecture designed to fuse 2D and 3D modalities of molecules to better foster multimodal performance. MolInteract consists of 2D and 3D message-passing towers whose unimodal embeddings are repeatedly fused and split apart to exchange 2D and 3D information. We pair MolInteract with a set of simple pre-training tasks from the existing literature and show that MolInteract yields strong multimodal performance on various 2D and 3D benchmark tasks.

2 Method

Architecture. MolInteract is composed of two parallel branches, one based on 2D graph message-passing neural networks (MPNNs) and the other based on 3D MPNNs. Specifically, given a molecular graph G_{2D} and one of its nodes i, its 2D node embedding h_i at the $(\ell+1)$th layer of an MPNN is given by

$$h^{(\ell+1,i)} = \text{Update}\left(h^{(\ell,i)}, \underset{j\in\mathcal{N}(i)}{\text{Agg}}\ \psi\left(h^{(\ell,i)}, h^{(\ell,j)}, e_{ij}\right)\right) \quad (1)$$

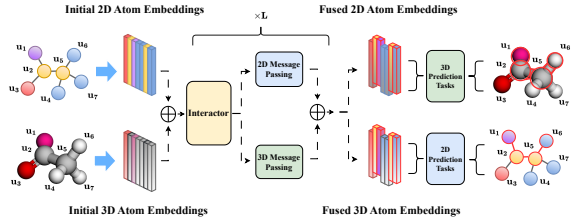

Figure 1: Pipeline of MolInteract. The output of 2 towers of unimodal message-passing are combined per-layer to create a fused embedding.

where Update is a node update function, Agg is a permutation-invariant function on the neighbors of i, and ψ is a function which computes messages between the node i and its neighbor j with the edge between them as context. In our case, we use layers from the GINE [3] architecture, which is a variant of GIN [18] that incorporates edge features during message-passing. To compute 3D atom embeddings, we use the continuous convolutional layers from SchNet [10] which conduct message-passing according to relative distances between atoms, incorporating both geometric and atom information into its embeddings. At the ℓth layer of unimodal message-passing, we take the 2D and 3D atom embeddings $z_{\text{2D}}^{(\ell,i)}$ and $z_{\text{3D}}^{(\ell,i)}$ and pass them to an interaction layer $\phi^{(\ell)}$. Formally, the multimodal embeddings at layer $\ell+1$ are given by

$$z_{\text{2D}}^{(\ell,i)}, z_{\text{3D}}^{(\ell,i)} = w_{:,:d}^{(\ell,i)}, w_{:,d:}^{(\ell,i)} \tag{2}$$

$$h_{\text{2D}}^{(\ell+1,i)}, h_{\text{3D}}^{(\ell+1,i)} = f^{(\ell+1)}\left(z_{\text{2D}}^{(\ell,i)}\right), g^{(\ell+1)}\left(z_{\text{3D}}^{(\ell,i)}\right) \tag{3}$$

where || denotes concatenation. We use a 2-layer MLP with SiLU activation. Then, our atom embeddings are updated accordingly:

$$z_{\text{2D}}^{(\ell+1,i)}, z_{\text{3D}}^{(\ell+1,i)} = w_{:,:d}^{(\ell+1,i)}, w_{:,d:}^{(\ell+1,i)} \tag{4}$$

where $f^{(\ell+1)}$ and $g^{(\ell+1)}$ are per-layer MPNNs and 3D-GNNs. Intuitively, the interaction layer ϕ serves as a exchange pathway between the unimodal towers. As the sole point of contact between the 2D and 3D towers, $\phi^{(\ell)}$ routes cross-modal information relevant to each pre-training task. Figure 1 shows the full pipeline.

Pre-training. To enforce multi-modality, we use the embeddings from the 2D branch of the model to predict 3D quantities and the 3D branch to predict 2D quantities. For 3D quantities, we predict interatomic distances and bond angles using MSE loss. We also predict the dihedral angle bins with width 20 using cross-entropy loss. For 2D quantities, we predict bond-types, shortest-path distances, and node eigenvector centrality ranking using cross-entropy loss. We train using the sum of these loss functions during pre-training with uniform weight for simplicity.

3 Experiments

We pre-train an 8-layer MolInteract with 9M parameters for 50 epochs on PCQM4Mv2 [2] containing over 3.3M molecules. We then fine-tune on QM9 [7] and MoleculeNet [16] and compare against baselines comprehensively reported in [21].

Table 1: Performance on QM9 measured in MAE.

Method	α	$\mathcal{E}_{\text{HOMO}}$	$\mathcal{E}_{\text{LUMO}}$	C_v	R^2	ZPVE
Stock SchNet [10], 8 layers	0.076	33.17	26.53	0.031	0.146	1.605
Distance Prediction [5]	0.065	27.61	23.34	0.033	0.248	1.837
3D InfoGraph [5]	0.062	29.29	24.60	0.030	**0.133**	1.644
3D InfoMax [11]	0.057	25.90	21.60	0.030	0.141	1.670
GraphMVP [6]	0.056	25.75	21.58	0.029	0.136	1.609
MoleculeSDE [4]	0.054	25.74	21.41	0.028	0.151	1.587
MoleBlend [21]	0.060	21.47	19.23	0.031	0.417	1.580
MolInteract	**0.047**	**20.60**	**17.88**	**0.025**	0.136	**1.327**

Table 2: Performance on MoleculeNet measured in ROC AUC. Higher is better.

Method	Tox21	ClinTox	HIV
GraphCL [20]	73.9±0.6	76.0±2.6	78.5±1.2
InfoGraph [12]	73.2±0.4	76.5±1.0	75.1±0.9
GROVER [9]	74.3±0.1	81.2±3.0	62.5±0.9
MolCLR [15]	73.0±0.1	86.1±0.9	76.2±1.5
GraphLoG [19]	75.7±0.5	76.7±3.3	77.8±0.8
GraphMAE [1]	75.5±0.6	82.3±1.2	77.2±1.0
Mole-BERT [17]	76.8±0.5	78.9±3.0	78.2±0.8
3D InfoMax [11]	74.5±0.7	79.9±3.4	76.1±1.3
GraphMVP [6]	74.5±0.4	79.0±2.5	74.8±1.4
MoleculeSDE [4]	76.8±0.3	87.0±0.5	78.8±0.9
MoleBlend [21]	**77.8±0.8**	87.6±0.7	79.0±0.8
MolInteract	77.3±0.5	**88.4±1.0**	**79.5±0.4**

3D performance. Following [13], we finetune on 110K random molecules and use the remaining 10K and 10.8K molecules as validation and test sets and evaluate by MAE. In Table 1, see that MolInteract exhibits a substantial lead in performance compared to baseline 3D pre-training methods, even without pre-training, demonstrating the effectiveness of the deep multimodal interaction layers and validating the use of deep interactions even with simple predictive SSL tasks.

2D performance. For the MoleculeNet datasets, we conduct experiments on Tox21, ClinTox, and HIV and measure performance by ROC AUC. Since 3D information is unavailable, we freeze the 2D embeddings of the 2D branch of MolInteract. We report the mean and standard deviation across three random seeds and use the Bemis-Murcko scaffolds recommended in DeepChem [8]. In Table 2, we see that MolInteract outperforms nearly all baselines with the exception of Tox21, further validating the use of multimodal pre-training and a deep multimodal architecture.

4 Conclusion

In this work, we introduce MolInteract an architectural approach to improving multimodal self-supervised learning that leverages deep interactions to fuse 2D and 3D representations of molecules. Our method is able to access fine-grained cross-modal information without sacrificing rich embeddings from modality-specific backbones, allowing for more effective interplay between 2D and 3D information when paired with even a simple set of predictive pre-training tasks, achieving new state-of-the-art performance on benchmark datasets as a result and contributing to the growing field of multimodal property prediction for small molecules.

References

[1] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *ACM SIGKDD*.

[2] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv preprint arXiv:2005.00687* (2020).

[3] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. Strategies for Pre-training Graph Neural Networks. In *ICLR*.

[4] Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. 2023. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *ICML*.

[5] Shengchao Liu, Hongyu Guo, and Jian Tang. 2023. Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching. In *ICLR*.

[6] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*.

[7] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* (2014).

[8] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. 2019. *Deep Learning for the Life Sciences*. O'Reilly Media.

[9] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *NeurIPS*.

[10] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NeurIPS*.

[11] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 2022. 3D Infomax improves GNNs for Molecular Property Prediction. In *ICML*.

[12] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. 2019. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*.

[13] Philipp Thölke and Gianni De Fabritiis. 2022. Equivariant Transformers for Neural Network based Molecular Potentials. In *ICLR*.

[14] Xu Wang, Huan Zhao, Wei-wei Tu, and Quanming Yao. 2023. Automated 3D Pre-Training for Molecular Property Prediction. In *Proceedings of the 29th ACM SIGKDD*.

[15] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* (2022).

[16] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. 2017. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9 (2017), 513 – 530.

[17] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In *ICLR*.

[18] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *ICLR*.

[19] Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. 2021. Self-supervised Graph-level Representation Learning with Local and Global Structure. In *ICML*.

[20] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. In *NeurIPS*.

[21] Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. 2024. Multimodal Molecular Pretraining via Modality Blending. In *ICLR*.

[22] Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2022. Unified 2D and 3D Pre-Training of Molecular Representations. In *Proceedings of the 28th ACM SIGKDD*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009