MOLMIX: A Simple Yet Effective Baseline for Multimodal Molecular Representation Learning

Andrei Manolache[123] **Dragoș Ținţaru**[3] **Mathias Niepert**[12]
[1]Computer Science Department, University of Stuttgart, Germany
[2]International Max Planck Research School for Intelligent Systems [3]Bitdefender, Romania
andrei.manolache@ki.uni-stuttgart.de

Abstract

In this work, we propose a simple transformer-based baseline for multimodal molecular representation learning, integrating three distinct modalities: SMILES strings, 2D graph representations, and 3D conformers of molecules. A key aspect of our approach is the aggregation of 3D conformers, allowing the model to account for the fact that molecules can adopt multiple conformations—an important factor for accurate molecular representation. The tokens for each modality are extracted using modality-specific encoders: a transformer for SMILES strings, a message-passing neural network for 2D graphs, and an equivariant neural network for 3D conformers. The flexibility and modularity of this framework enable easy adaptation and replacement of these encoders, making the model highly versatile for different molecular tasks. The extracted tokens are then combined into a unified multimodal sequence, which is processed by a downstream transformer for prediction tasks. To efficiently scale our model for large multimodal datasets, we utilize Flash Attention 2 and bfloat16 precision. Despite its simplicity, our approach achieves state-of-the-art results across multiple datasets, demonstrating its effectiveness as a strong baseline for multimodal molecular representation learning. Our code is publicly available at `https://github.com/andreimano/MolMix`.

1 Introduction and Related Work

Accurately representing molecular structures is fundamental in computational chemistry and drug discovery [1, 2, 3]. Effective molecular representations enable machine learning models to predict molecular properties, understand chemical behaviors, and accelerate the development of new compounds. Traditional molecular representation methods typically focus on a single modality, such as SMILES strings [4, 5], chemical fingerprints [6], 2D molecular graphs [7], or the 3D geometry of molecules [8, 9]. While effective, these methods overlook important molecular characteristics that can be captured by other modalities [10, 11]. To address this, recent research has introduced multimodal approaches that integrate multiple molecular representations to provide richer representations for machine learning tasks in chemistry. Stärk et al. [12] proposed an information maximization approach to enhance the mutual information between 2D and 3D molecular embeddings. Similarly, Liu et al. [13] used contrastive pre-training to align 2D and 3D representations. Other approaches extract both 2D and 3D features, such as shortest path distances and 3D distance encodings, to build multimodal models [14, 15]. Zhu et al. [16] unified 2D and 3D molecular data in a pre-training framework by predicting either masked 2D structures or 3D conformations. Additionally, language-based models have been integrated with molecular data. Tang et al. [17], Xiao et al. [18], and Srinivas and Runkana [19] leveraged large-scale language models to incorporate textual descriptions of molecules, enhancing molecular property predictions.

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

While these representations capture certain aspects of molecular structures, they may not fully encompass the variability inherent in molecular conformations. Many molecular properties, such as solubility, toxicity, and binding affinity, are influenced by the range of conformations a molecule can adopt in nature [20]. Utilizing a single geometric representation for a molecule, therefore, restricts the effectiveness of machine learning models. Moreover, identifying which conformers most significantly impact the properties of interest remains difficult. Consequently, creating comprehensive multi-modal representations that integrate multiple 3D conformations continues to be a challenge. To address this, Axelrod and Gómez-Bombarelli [21] propose a scheme where 3D conformer embeddings are first extracted using an equivariant backbone and then aggregated with an attention mechanism, but they do not train on multimodal data. Similarly, Nguyen et al. [22] introduce a conformer aggregation approach that leverages Optimal Transport techniques to obtain a single 3D embedding from multiple conformers, which is subsequently combined with 2D embeddings derived from a GNN. Additionally, Zhu et al. [23] propose MARCEL, a conformer aggregation benchmark alongside models that employ set encoders to pool conformer embeddings with 2D structures and SMILES strings for downstream tasks. These approaches illustrate ongoing efforts to develop more holistic and effective molecular representations by incorporating both 2D and multiple 3D conformational data.

Despite these advances, there remains a need for simple yet effective models that can seamlessly integrate multiple modalities and handle multiple conformers without significant computational overhead. Moreover, recent observations [24, 25] suggest that some model design choices might be unnecessary for strong empirical performance, thereby making the added complexity superfluous and inefficient. To address this challenge, we propose MOLMIX, a simple yet effective baseline for multimodal molecular representation learning. We employ modality-specific encoders - a transformer for SMILES strings, a GNN for 2D graphs, and equivariant neural networks for 3D conformers - to extract text and node embeddings from each modality. These embeddings are concatenated into a multimodal sequence, separated by special tokens, and fed into a downstream transformer that predicts molecular properties. By leveraging efficient techniques like Flash Attention [26, 27] and bfloat16 precision, MOLMIX scales to handle large sequences of atom tokens with minimal computational overhead, enabling the direct incorporation of all conformers. Despite its straightforward design, MOLMIX achieves state-of-the-art results across multiple datasets, demonstrating that simplicity can be highly effective in multimodal molecular representation learning, while the modular design allows us to easily exchange the specific modality encoders.

To summarize, our main contributions are:

1. **Simple multimodal molecular framework**: We introduce MOLMIX, which seamlessly combines SMILES strings, 2D molecular graphs, and multiple 3D conformers into a unified sequence for molecular representation learning.
2. **Conformer aggregation**: By incorporating node embeddings from 3D conformers, MOLMIX effectively captures conformational variability.
3. **Scalability**: We utilize Flash Attention and bfloat16 (bf16) precision to scale our model, enabling the processing of large multimodal datasets with minimal computational overhead.
4. **State-of-the-Art performance**: MOLMIX achieves superior results on multiple benchmark datasets, establishing a strong baseline for future research in multimodal molecular representation learning.
5. **Transfer learning capabilities** We show that MOLMIX could potentially be used for pre-training on large molecular datasets.

We make our code publicly available at `https://github.com/andreimano/MolMix`.

2 MOLMIX: A Multimodal Molecular Transformer

1D Encoder We represent molecules using SMILES strings, which encode chemical structures as sequences of characters. Let $S = [s_1, s_2, \ldots, s_n]$ denote the input SMILES string, where each s_i is a character. Each s_i is mapped to an embedding vector $\mathbf{e}_i = \text{Embedding}(s_i)$. To account for sequence order, positional encodings are added: $\mathbf{z}_i = \mathbf{e}_i + \text{PE}(i)$. A transformer encoder [28] then processes these vectors to obtain the hidden representations

$$\mathbf{h}_i^{\text{1D}} = \text{Transformer}(\mathbf{z}_i), \tag{1}$$

for all $i \in \{1, \ldots, n\}$. Each hidden representation $\mathbf{h}_i^{\text{1D}}$ corresponds to the respective input character s_i, effectively capturing the contextual information about the molecule, for each character.

2D Encoder We represent molecules as graphs $G = (V, E)$, where V is the set of atoms and E is the set of covalent bonds. Each atom $v \in V$ and bond $e_{uv} \in E$ are associated with initial feature vectors $\mathbf{x}_v$ and $\mathbf{e}_{uv}$, respectively. We use a message-passing framework with GINE [29, 30] as the backbone to capture the molecular graph's structural information. At each message-passing step j, the hidden representation of atom v is updated as

$$\mathbf{h}_{v,j}^{\text{2D}} = \text{GINE}\left(\mathbf{h}_{v,j-1}^{\text{2D}}, \{\mathbf{h}_{u,j-1}^{\text{2D}} \mid u \in \mathcal{N}(v)\}, \{\mathbf{e}_{uv}\}\right), \tag{2}$$

where $\mathcal{N}(v)$ denotes the neighbors of atom v. This iterative process aggregates information from neighboring atoms and bonds, enabling the model to learn graph representations. The final hidden embeddings $\mathbf{h}_{v,j}^{\text{2D}}$ encode both local and global structural features of the molecule.

3D Encoder To leverage the three-dimensional structural information of molecules, we utilize 3D conformations represented by the spatial coordinates of each atom. Let V denote the set of atoms. Each atom $v \in V$ is associated with a 3D coordinate $\mathbf{r}_v \in \mathbb{R}^3$. To extract meaningful atom embeddings that respect the geometric properties of the molecule, we employ an neural network with 3D inductive biases, such as SchNet [8] or GemNet [9], as the backbone model. These networks process the 3D coordinates $\{\mathbf{r}_v\}_{v \in V}$ along with the initial atom features $\{x_v\}_{v \in V}$ and apply a cutoff function to consider interactions within a specified distance range, generating hidden embeddings

$$\mathbf{h}_v^{\text{3D}} = \text{3DNetwork}(\mathbf{r}_v, \mathbf{x}_v), \tag{3}$$

for all $v \in V$. These atom embeddings $\mathbf{h}_v^{\text{3D}}$ capture both the local geometry and the global spatial arrangement of the molecule.

Downstream Multimodal Transformer To integrate different molecular representations, we design a multimodal transformer that combines three distinct modalities. The SMILES encoder outputs token embeddings $\mathbf{h}_i^{\text{1D}}$, where $\mathbf{h}_i^{\text{1D}}$ corresponds to the i^{th} character in the string. From the 2D MPNN encoder, we extract atom embeddings $\mathbf{h}_{v,j}^{\text{2D}}$ for atom v at layer j. By using embeddings from all layers, the model captures both local and distant atom interactions, mitigating the oversmoothing effect common in deep GNNs. The 3D encoder provides atom embeddings $\mathbf{h}_{v,c}^{\text{3D}}$ for atom v and conformer c, encapsulating spatial geometry. We use multiple conformers by simply adding all the atom embeddings to the multimodal sequence. Modality-specific learnable encodings are added to the embeddings from each modality. These modality-enhanced embeddings are concatenated into a unified sequence, with special tokens included: a classification token $\mathbf{h}_{\text{CLS}}$ is added at the start, and separation tokens $\mathbf{h}_{\text{SEP}}$ are placed between modalities. The resulting input sequence is structured as

$$\mathbf{H} = \left[\mathbf{h}_{\text{CLS}}, \{\mathbf{h}_i^{\text{1D}}\}_i, \mathbf{h}_{\text{SEP}}, \left\{\mathbf{h}_{v,j}^{\text{2D}}\right\}_{v,j}, \mathbf{h}_{\text{SEP}}, \left\{\mathbf{h}_{v,c}^{\text{3D}}\right\}_{v,c}, \mathbf{h}_{\text{SEP}}\right].$$

This sequence is then processed by the downstream transformer, which utilizes the self-attention mechanism to integrate and contextualize information across all modalities. After the transformer layers, the embedding corresponding to the classification token $\mathbf{h}_{CLS}^{out}$ is extracted and sent to a readout MLP to perform downstream tasks such as property prediction and molecular classification:

$$\mathbf{h}_{\text{CLS}}^{\text{out}} = \text{MultimodalTransformer}(\mathbf{H}) \tag{4}$$

$$\hat{y} = \text{MLP}(\mathbf{h}_{\text{CLS}}^{\text{out}}) \tag{5}$$

We reduce memory overhead in our multimodal transformer with bfloat16 precision and Flash Attention 2 [27]. See appendix C for details and a memory comparison with classical attention.

Table 1: Comparison between MOLMIX and other baselines on the *Drugs-75K* and *Kraken* datasets from MARCEL [23]. **1D**, **2D** and **3D** represents training on the SMILES strings and molecule fingerprints, 2D molecular representations and 3D conformers. **Multimodal** represents training on all three modalities. The metric used is the Mean Absolute Error (MAE, ↓). **Bold** indicates the best-performing model, while underline denotes the second-best. MOLMIX obtains the best results on 5 out of 7 properties, with second-best results obtained on two properties (Drugs-75K/χ and Kraken/BurL).

	Model	Drugs-75K			Kraken			
		IP ↓	EA ↓	χ ↓	B5 ↓	L ↓	BurB5 ↓	BurL ↓
1D	RF	0.498±0.003	0.474±0.002	0.273±0.003	0.476±0.004	0.430±0.009	0.275±0.018	0.152±0.014
	LSTM	0.478±0.002	0.464±0.000	0.250±0.005	0.487±0.028	0.514±0.041	0.281±0.004	0.192±0.002
	Transformer	0.661±0.002	0.585±0.003	0.407±0.001	0.961±0.081	0.839±0.043	0.493±0.037	0.278±0.021
2D	GIN	0.435±0.003	0.417±0.003	0.226±0.002	0.313±0.026	0.400±0.034	0.172±0.003	0.120±0.004
	GIN+VN	0.436±0.006	0.417±0.008	0.227±0.000	0.357±0.003	0.434±0.042	0.242±0.003	0.174±0.011
	ChemProp	0.460±0.003	0.442±0.005	0.244±0.001	0.485±0.007	0.545±0.045	0.300±0.009	0.195±0.014
	GraphGPS	0.435±0.005	0.409±0.006	0.221±0.005	0.345±0.032	0.436±0.013	0.207±0.012	0.150±0.014
3D	SchNet	0.439±0.006	0.421±0.002	0.224±0.009	0.329±0.007	0.546±0.034	0.230±0.011	0.186±0.010
	DimeNet++	0.444±0.009	0.423±0.007	0.244±0.008	0.351±0.011	0.417±0.040	0.210±0.016	0.153±0.007
	GemNet	0.407±0.001	0.392±0.002	**0.197±0.004**	0.279±0.013	0.375±0.009	0.178±0.010	0.164±0.006
	PaiNN	0.451±0.004	0.450±0.005	0.232±0.004	0.344±0.039	0.447±0.032	0.240±0.018	0.167±0.009
	ClofNet	0.439±0.008	0.425±0.007	0.238±0.002	0.487±0.009	0.642±0.036	0.288±0.017	0.253±0.005
	LEFTNet	0.417±0.001	0.396±0.001	0.208±0.005	0.307±0.001	0.449±0.026	0.218±0.001	0.149±0.010
Multimodal	SchNet	0.454±0.007	0.438±0.013	0.237±0.010	0.270±0.019	0.432±0.046	0.202±0.018	0.144±0.004
	DimeNet++	0.413±0.008	0.394±0.003	0.227±0.005	0.263±0.012	0.347±0.009	0.178±0.011	0.119±0.011
	GemNet	0.419±0.002	0.400±0.001	0.217±0.004	0.231±0.003	0.339±0.027	0.159±0.007	**0.095±0.001**
	PaiNN	0.447±0.007	0.427±0.003	0.229±0.007	0.223±0.022	0.362±0.019	0.169±0.011	0.132±0.009
	ClofNet	0.428±0.006	0.403±0.002	0.220±0.007	0.323±0.002	0.449±0.005	0.218±0.019	0.155±0.004
	LEFTNet	0.417±0.004	0.395±0.000	0.207±0.002	0.264±0.013	0.364±0.035	0.202±0.003	0.139±0.001
	MOLMIX	**0.405±0.002**	**0.379±0.004**	0.206±0.002	**0.191±0.017**	**0.305±0.020**	**0.146±0.002**	0.121±0.005

Since we use the same positional encoding for each $\mathbf{h}^{\text{2D}}_{v,i}$ and $\mathbf{h}^{\text{3D}}_{v,c}$, we maintain the permutation equivariance property of the 2D and 3D encoders. Another desirable property is for the model to preserve any invariance of the 3D encoder. Indeed, MOLMIX preserves these useful inductive biases:

Theorem 1. *Let S be the SMILES string, G be the 2D graph, and $\{c_1, \ldots, c_k\}$ be a set of k 3D conformers for a molecule. Let $\hat{y} = f_\theta(S, G, \{c_1, \ldots, c_k\})$ be the output prediction obtained as described in eq.* (1) - (5)*. Let our 3D encoder be invariant to the actions of some group $\mathcal{G}$. Then f_θ is also invariant to any $T_1, \ldots, T_k \in \mathcal{G}$, i.e. $f_\theta(S, G, \{T_1 c_1, \ldots, T_k c_k\}) = f_\theta(S, G, \{c_1, \ldots, c_k\})$.*

3 Experimental Setup and Results

In this section, we evaluate how MOLMIX improves predictive performance on real-world datasets by addressing the following questions: ***Q1)*** *How does* MOLMIX*'s performance compare to other sophisticated models?;* ***Q2)*** *Does incorporating multiple modalities enhance downstream performance?;* ***Q3)*** *Are pre-trained weights beneficial for transfer learning?*

Table 2: Comparison between MOLMIX and other approaches as reported in [22]. The metric used is Root Mean Squared Error (RMSE ↓). MOLMIX obtains the overall best scores, significantly improving upon the recently proposed multimodal CONAN-FGW model.

Model	Lipo ↓	ESOL ↓	FreeSolv ↓	BACE ↓
2D-GAT	1.178±0.454	1.513±0.130	2.926±1.160	1.358±0.574
D-MPNN	0.731±0.148	0.961±0.212	2.053±0.261	0.850±0.145
MolFormer	0.701±0.110	0.875±0.249	2.342±0.212	1.045±0.145
SchNet-scalar	0.839±0.179	0.820±0.164	1.268±0.397	0.850±0.316
SchNet-emb	0.767±0.179	0.797±0.239	1.260±0.369	0.832±0.167
ChemProp3D	0.695±0.230	0.825±0.152	1.419±0.427	0.903±0.412
CONAN	0.746±0.114	0.756±0.138	1.223±0.397	0.797±0.226
CONAN-FGW	0.650±0.126	0.727±0.148	1.033±0.288	0.741±0.126
MOLMIX	**0.614±0.022**	**0.639±0.017**	**0.976±0.044**	**0.387±0.041**

To address Question 1, we train MOLMIX on four MoleculeNet datasets [31]—*Lipo*, *ESOL*, *FreeSolv*, and *BACE*—covering various molecular properties, including physical chemistry and biophysics. Conformers are generated using the RDKit chemoinformatics package [32]. We follow the same train/validation/test splits as [22]. We also train models on the newly introduced MARCEL benchmark [23], specifically the *Drugs-75k* and *Kraken* datasets. *Drugs-75K*, a subset of the *GEOM-Drugs* dataset [33], contains 75,099 molecules with conformers generated by Auto3D [34], and labels for ionization potentials (IP), electron affinity (EA), and electronegativity (χ). *Kraken* [35] includes 1,552 monodentate organophosphorus ligands, with conformers generated via DFT, and labels for four 3D ligand descriptors: Sterimol B5 (B5), Sterimol L (L), buried Sterimol B5 (BurB5), and buried Sterimol L (BurL). We follow the same splits as in [23].

For all experiments, we report the mean and standard deviation over five different runs. We use the same hyperparameters for the modality encoders—we encode the SMILES strings using a transformer with two layers and four attention heads, the 2D graph is processed by a GINE [29] network containing 6 layers, and the 3D conformations are processed either using a SchNet [8] or a GemNet [9] model. All of the encoders have a hidden dimension of 128. The modality embeddings are then projected by a linear layer to a 512-dimensional space before we jointly processing them with a downstream Transformer network with 8 heads and 6 layers for the *Lipo*, *ESOL*, *FreeSolv*, *BACE* and *Kraken*, and 12 heads with 8 layers on *Drugs-75K*. We use the Schedule-Free AdamW optimizer [36, 37, 38].

On the MoleculeNet datasets, MOLMIX obtains the overall best results, significantly improving upon the results of [22] on some datasets such as *BACE*, as can be seen in table 2. This highlights that our simple approach can potentially learn better multimodal representations with conformer aggregation than previously proposed methods that contain more sophisticated aggregation techniques.

On the MARCEL datasets, compared to the approaches in [23], we achieve the best results on five out of seven properties and second-best on the remaining two, as can be seen in table 1. This suggests that MOLMIX also consistently performs well when using physically-grounded conformer generation methods like DFT and Auto3D.

To address Question 2, fig. 1 shows that for three of the four *Kraken* descriptors, training on all three modalities yields the best results, while for one property, training on the 3D modality alone performs slightly better. The results indicate that 3D tokens contribute the most to downstream performance, followed by 2D, with SMILES (1D) having the least impact. Notably, SMILES strings significantly improve performance when predicting the Sterimol L descriptor.

Figure 1: Modality ablation study on the Kraken dataset (MAE ↓). We keep the downstream Transformer fixed and train using a single modality or a combination of modalities. Using all three modalities obtains the best results on three out of the four properties, with the second-best results generally being obtained by a configuration that contains 3D conformers. Notably, for the *buried Sterimol L* property, the best results are obtained by a 3D encoder + Transformer model, indicating that the property could mainly depend on the 3D structure.

Modality	B5 ↓	L ↓	BurB5 ↓	BurL ↓
1D	0.499±0.033	0.497±0.025	0.291±0.020	0.187±0.005
2D	0.258±0.017	0.347±0.013	0.176±0.010	0.141±0.004
3D	0.213±0.008	0.337±0.009	0.164±0.004	**0.116±0.003**
1D+2D	0.297±0.015	0.390±0.016	0.180±0.008	0.153±0.006
1D+3D	0.209±0.002	0.337±0.010	0.156±0.013	0.127±0.004
2D+3D	0.202±0.009	0.356±0.026	0.151±0.004	0.122±0.004
1D+2D+3D	**0.191±0.017**	**0.305±0.020**	**0.146±0.002**	0.121±0.005

Figure 2: Transfer learning experiment. We select the best checkpoint of a model trained to predict the electronegativity (χ) on the *Drugs-75K* dataset. We then freeze the model and only train the last linear readout layer on the *Kraken* dataset. We compare with a randomly initialized model. For all descriptors, using the pre-trained weights improve predictive performance. Note that pretraining improves both mean performance and standard deviations.

Modality	B5 ↓	L ↓	BurB5 ↓	BurL ↓
Random init.	0.567±0.010	0.543±0.020	0.334±0.004	0.216±0.003
Pretrain	**0.521±0.003**	**0.509±0.004**	**0.316±0.001**	**0.195±0.003**

To answer Question 3, we select the best checkpoint from a model trained to predict electronegativity (χ) on the *Drugs-75K* dataset. We then freeze its weights and train only the final linear layer to predict descriptors on the *Kraken* dataset, comparing it to a randomly-initialized model. As shown in fig. 2, pre-training improves predictive performance in all cases, suggesting that with sufficient data, MOLMIX could serve as a foundation model for molecular tasks.

4 Conclusions and Further Work

We propose MOLMIX, a simple yet effective multimodal molecular transformer supporting conformer aggregation. MOLMIX preserves inductive biases of modality encoders and achieves state-of-the-art results across multiple datasets. We hint towards MOLMIX being able to support transfer learning, suggesting that it could be used as a molecular foundation model. Finally, we use Flash Attention and bf16 precision to handle longer sequences and multiple modalities efficiently.

We leave three open questions. First, large self-supervised VLMs excel in 0-shot prediction and fine-tuning [39, 40, 41, 42]. Exploring self-supervised pre-training for MOLMIX using signals like masked language modeling [43] and noise-contrastive estimation [44] could be valuable. Second, multiple conformers without pooling may be suboptimal; token merging [45] could improve memory and runtime. Lastly, adding modalities like molecular fingerprints may enhance performance.

Acknowledgments

AM and MN acknowledge funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016, the support by the Stuttgart Center for Simulation Science (SimTech), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). AM and DṬ acknowledge funding by the EU Horizon project ELIAS (No. 101120237).

References

[1] Jun Xia, Yanqiao Zhu, Yuanqi Du, Yue Liu, and Stan Z Li. A systematic survey of chemical pre-trained models. *IJCAI*, 2023.

[2] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37, 2020. ISSN 1740-6749.

[3] W. Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, 54(2), 2021. PMID: 33370107.

[4] Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC Bioinformatics*, 19, 2018.

[5] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 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 '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366663.

[6] Naifeng Wen, Guanqun Liu, Jie Zhang, Rubo Zhang, Yating Fu, and Xu Han. A fingerprints based molecular property prediction method using the bert model. *Journal of Cheminformatics*, 14, 10 2022.

[7] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

[8] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[9] J. Klicpera, F. Becker, and S. Günnemann. Gemnet: Universal directional graph neural networks for molecules. *CoRR*, 2021.

[10] Grzegorz Skoraczynski, P. Dittwald, Blazej Miasojedow, S. Szymkuć, E. Gajewska, Bartosz Grzybowski, and Anna Gambin. Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient? *Scientific Reports*, 7, 06 2017.

[11] Zhen Liu, Yurii Moroz, and Olexandr Isayev. The challenge of balancing model sensitivity and robustness in predicting yields: A benchmarking study of amide coupling reactions, 07 2023.

[12] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D infomax improves GNNs for molecular property prediction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 17–23 Jul 2022.

[13] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

[14] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending, 2023.

[16] Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850.

[17] Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. MolLM: a unified language model for integrating biomedical text with 2D and 3D molecular representations. *Bioinformatics*, 40, 06 2024.

[18] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding, 2024.

[19] Sakhinana Sagar Srinivas and Venkataramana Runkana. Cross-modal learning for chemistry property prediction: Large language models meet graph machine learning, 2024.

[20] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910), 2022.

[21] Simon Axelrod and Rafael Gómez-Bombarelli. Molecular machine learning with conformer ensembles. *Mach. Learn.: Sci. Technol.*, 4(3), September 2023. ISSN 2632-2153.

[22] Duy MH Nguyen, Nina Lukashina, Tai Nguyen, An T Le, TrungTin Nguyen, Nhat Ho, Jan Peters, Daniel Sonntag, Viktor Zaverkin, and Mathias Niepert. Structure-aware e (3)-invariant molecular conformer aggregation networks. *International Conference on Machine Learning*, 2024.

[23] Yanqiao Zhu, Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor W. Coley, Yizhou Sun, and Wei Wang. Learning over molecular conformer ensembles: Datasets and benchmarks. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Yuyang Wang, Ahmed Elhag, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista Martin. Swallowing the bitter pill: Simplified scalable conformer generation. In *ICML*, 2024.

[25] Jan Tönshoff, Martin Ritzert, Eran Rosenbluth, and Martin Grohe. Where did the gap go? reassessing the long-range graph benchmark. 2023.

[26] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv preprint*, 2017.

[29] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[30] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

[31] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 2018.

[32] G Landrum. Rdkit: Open-source cheminformatics http://www.rdkit.org. 2016.

[33] Simon Axelrod and Rafael Gómez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1), 2022.

[34] Zhen Liu, Tetiana Zubatiuk, Adrian Roitberg, and Olexandr Isayev. Auto3d: Automatic generation of the low-energy 3d structures with ani neural network potentials. *Journal of Chemical Information and Modeling*, 62(22), 2022. PMID: 36112860.

[35] Tobias Gensch, Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S. Sigman, and Alán Aspuru-Guzik. A comprehensive discovery platform for organophosphorus ligands for catalysis. *Journal of the American Chemical Society*, 144(3), 2022. PMID: 35020383.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[38] Aaron Defazio, Xingyu Alice Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024. URL https://arxiv.org/abs/2405.15682.

[39] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[40] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[41] Team Chameleon. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[42] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[44] Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. Pre-training transformers as energy-based cloze models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020. Association for Computational Linguistics.

[45] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023.

A Proof for Theorem 1

Theorem 1. *Let S be the SMILES string, G be the 2D graph, and $\{c_1, \ldots, c_k\}$ be a set of k 3D conformers for a molecule. Let $\hat{y} = f_\theta(S, G, \{c_1, \ldots, c_k\})$ be the output prediction obtained as described in eq.* (1) - (5)*. Let our 3D encoder be invariant to the actions of some group $\mathcal{G}$. Then f_θ is also invariant to any $T_1, \ldots, T_k \in \mathcal{G}$, i.e. $f_\theta(S, G, \{T_1 c_1, \ldots, T_k c_k\}) = f_\theta(S, G, \{c_1, \ldots, c_k\})$.*

Proof. Let g_θ be our 3D encoder network, as described in eq. (3). Let V be the set of atoms and $\{x_v\}_{v \in V}$, $\{r_v\}_{v \in V}$ the atom features and their 3D coordinates, such that a conformer can be described as the tuple $c = (\{x_v\}_{v \in V}, \{r_v\}_{v \in V})$. We assume that g_θ is invariant to any action $T \in \mathcal{G}$, therefore we have that, for any conformer c, $g_\theta(Tc) = g_\theta(c) = \mathbf{h}^{\text{3D}}$.

Let h_θ be the downstream Transformer together with the readout layer, as described in eq. (4) - (5). Since we add the same learnable modality encoding to each $h^{\text{3D}}_{v,k}$, we also have that for any permutation $\pi \in Sym(K)$, we have

$$
\begin{aligned}
h_\theta(\{g_\theta(T_1 c_1), \ldots, g_\theta(T_k c_k)\}) &= h_\theta(\{g_\theta(c_1), \ldots, g_\theta(c_k)\}) \\
&= h_\theta(\{\mathbf{h}^{\text{3D}}_{v,1}, \ldots, \mathbf{h}^{\text{3D}}_{v,k}\}_{v \in V}) \\
&= h_\theta(\{\mathbf{h}^{\text{3D}}_{v,\pi(1)}, \ldots, \mathbf{h}^{\text{3D}}_{v,\pi(k)}\}_{v \in V}) \\
&= \tilde{y},
\end{aligned}
$$

therefore, if when we include the 2D graph G and the SMILES string S, we obtain $f_\theta(S, G, \{T_1 c_1, \ldots, T_k c_k\}) = f_\theta(S, G, \{c_1, \ldots, c_k\}) = \hat{y}$. □

B Qualitative attention example

We present the attention scores for each head in a MOLMIX model trained on the Drugs-75k dataset, using a randomly sampled molecule from the dataset for visualization. As shown in Figure 3, distinct patterns emerge across the attention heads. While it remains challenging to assign a definitive interpretation to each individual head, certain sparse or dense patterns are evident in each cross-modality section. This suggests that the model is learning to extract meaningful and potentially useful features from all modalities.

C Attention implementation details

We employ Flash Attention 2 [27] for the self-attention mechanism in our models. Flash Attention 2 is a hardware-optimized implementation that significantly reduces both memory usage and runtime compared to the standard attention algorithm. It achieves these gains by leveraging GPU programming techniques, such as kernel fusion and tiling. Additionally, we utilize the *varlen* implementation, which prevents unnecessary memory and compute consumption on padding tokens.

Table 3 presents the memory savings achieved by using Flash Attention 2 during training.

Table 3: Comparison between fp32 standard attention and bf16 Flash Attention 2 memory usage across different models and batch sizes.

Dataset	Kraken		
Batch Size	16	32	64
Standard Attention (SchNet)	67 GB	OOM	OOM
Standard Attention (GemNet)	80 GB	OOM	OOM
Flash Attention 2 (SchNet)	5.1 GB	11.6 GB	22.3 GB
Flash Attention 2 (GemNet)	22 GB	40 GB	73 GB

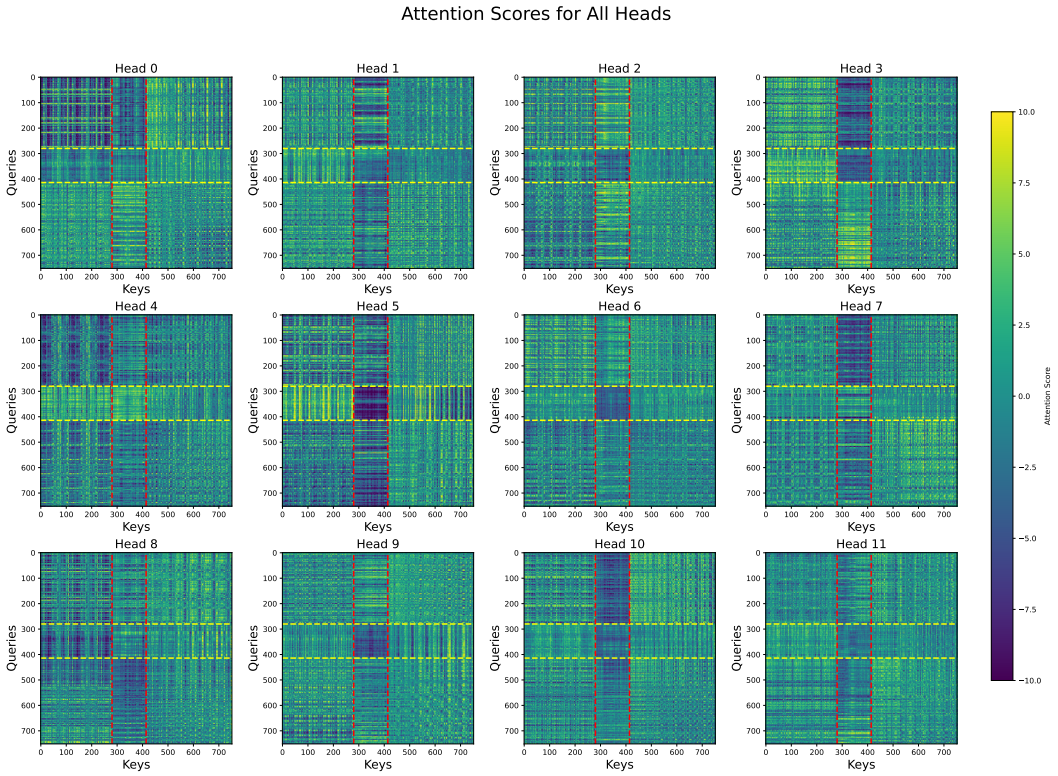

Figure 3: Red lines mark the boundaries between modalities on the key axis (with keys for each token represented by columns), while yellow lines mark the boundaries on the query axis (with queries represented by rows). The modalities are ordered as 3D, SMILES, and 2D. The attention scores are taken from the first layer of the model and clipped to the $[-10, 10]$ range.