A Self-Supervised Framework for Robust Multi-Modal Molecular Representation Learning

Indra Priyadarsini S
IBM Research - Tokyo
indra.ipd@ibm.com

Seiji Takeda
IBM Research - Tokyo
seijitkd@jp.ibm.com

Lisa Hamada
IBM Research - Tokyo
lisa.hamada@ibm.com

Sina Klampt
IBM Research - Tokyo
sina.klampt@ibm.com

Takao Moriyama
IBM Research - Tokyo
moriyama@jp.ibm.com

Abstract

Molecular property prediction has greatly benefited from learned embeddings such as SMILES-based, SELFIES-based, and graph-derived representations. However, existing approaches often rely on a single modality or naïvely concatenating multiple modalities, limiting robustness and failing under missing-modality conditions. In this work, we propose a novel self-supervised fusion framework - dynamic fusion, that dynamically integrates multiple molecular embeddings. The proposed framework employs intra-modal gating for feature selection, inter-modal attention for adaptive weighting, and cross-modal reconstruction to ensure information exchange. Through progressive modality masking during training, the dynamic fusion approach learns to generate fused embeddings resilient to missing modalities. We conducted preliminary evaluations of the proposed approach on MoleculeNet benchmarks, and demonstrate a superior performance in reconstruction, modality alignment, and downstream property prediction tasks compared to unimodal baselines. Our findings highlight the importance of feature-level gating, entropy-regularized attention, and cross-modal reconstruction in achieving robust fusion.

1 Introduction

Molecular property prediction is central to drug discovery, materials science, and computational chemistry. Traditional cheminformatics relied heavily on handcrafted descriptors, but recent advances in deep learning have shifted the field towards learned molecular embeddings. String-based representations such as SMILES and SELFIES, along with graph-based models that encode molecular structures as graphs, have proven highly effective.These representation models are extensively used for tasks such as molecular property prediction, where their ability to capture and encode crucial molecular features has demonstrated remarkable efficacy (1; 2; 3; 4; 5; 6; 7; 8). These modalities, however, emphasize different facets of molecular structure: SMILES models capture sequential dependencies, SELFIES enforces chemical validity via a robust grammar, and graph encoders reflect topological organization. This complementarity motivates multi-modal fusion, an approach that has delivered state-of-the-art performance in vision, language, healthcare, and autonomous systems by leveraging signals from heterogeneous sources (9; 10; 11).

Realizing the benefits of multimodality in chemistry and material science, however, presents two persistent challenges. First, missing modalities are common: data pipelines are often incomplete and not all molecules have all representations due to computational cost or preprocessing failures. Second, noisy or redundant features arise within and across modalities, which can dilute signal. Conventional fusion strategies such as naive concatenation, averaging, or simple pooling—presume complete inputs, treat modality-specific noise uniformly, and cannot adjust per-molecule contributions, leading to brittle performance when data are sparse, imbalanced or missing.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To address these challenges, in this work, we propose a self-supervised dynamic fusion framework that learns a single, compact molecular embedding by dynamically selecting informative features, adapting modality weights on a per-sample basis, and reconstructing missing embeddings during training. By incorporating a curriculum of progressive modality masking, the proposed dynamic fusion approach is trained to remain robust under missing-modality conditions, making it particularly useful in real-world molecular applications.

2 Proposed Dynamic Fusion Framework

In this section, we outline the methodological framework of our proposed approach. Fig. 1 illustrates the schematic of the proposed dynamic fusion approach. The core objective of our dynamic multimodal fusion model is to enhance robustness and performance by adaptively tailoring the fusion process to inputs from distinct unimodal models and efficient handling of missing or scarce paired data.

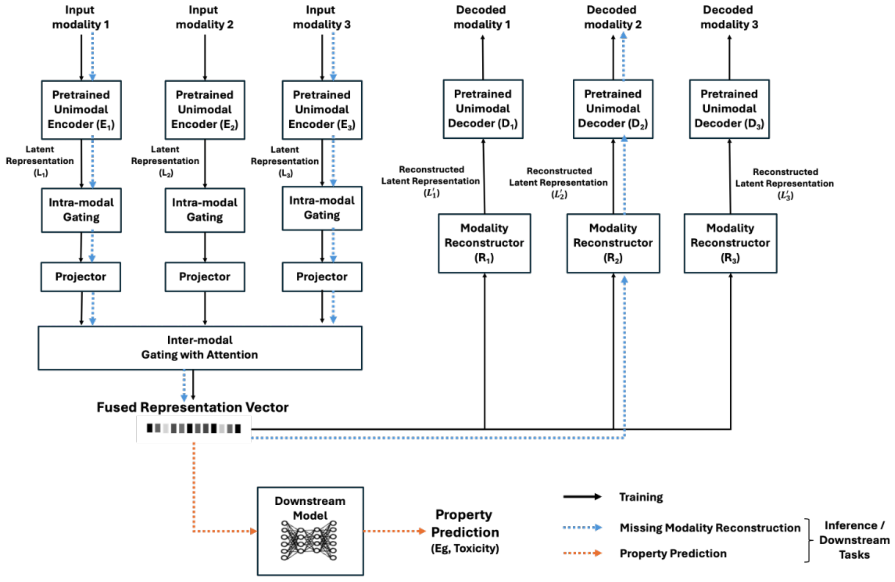

Figure 1: Block diagram of the proposed dynamic fusion model

We assume that for each molecule, multiple pretrained embeddings are available by means of their corresponding unimodal models: for example, SMILES-based transformer embeddings (768 dimensions) (12), SELFIES-based embeddings (1024 dimensions)(13), and graph neural network embeddings from MHG-GED (1024 dimensions) (14). Each embedding is treated as a modality. The task is to produce a fused embedding of fixed dimension (512 in our experiments) that integrates information across modalities and generalizes well to downstream supervised tasks.

The framework first applies intra-modal gating networks, small MLPs that output sigmoid weights to select informative features within each modality. The gated embeddings are then linearly projected into a shared latent space, enabling direct comparison across heterogeneous dimensions. Fusion is achieved through an inter-modal attention mechanism, which assigns adaptive weights to modalities per sample. Missing inputs are masked during softmax normalization, ensuring that absent modalities do not influence the fused representation. The final fused embedding is a weighted sum of projected embeddings. To improve robustness, our framework incorporates cross-modal reconstruction: decoder networks attempt to reconstruct masked embeddings from the fused representation. Training randomly masks modalities and requires their reconstruction, enforcing redundancy and making the fused embedding informative even with incomplete inputs.

The proposed dynamic fusion framework is trained with a multi-component objective function comprising of (i) reconstruction loss ensures recovery of masked modalities; (ii) alignment loss keeps the fused embedding consistent with modality projections; (iii) diversity loss reduces redundancy between gating patterns; (iv) entropy regularization prevents attention collapse; and (v) sparsity

regularization promotes compact feature selection. A key aspect of the training strategy is employing progressive masking, by gradually increasing the masking probability during training. This curriculum forces the model to handle increasingly difficult reconstruction tasks, leading to fused embeddings that are accurate, balanced, and resilient to missing modalities.

At inference, the fused embedding can be directly employed for downstream property prediction tasks or for reconstructing missing modalities, making the framework broadly applicable across diverse molecular learning scenarios.

3 Results and Discussions

To evaluate our proposed dynamic multimodal fusion approach, we consider three modalities : SMILES, SELFIES, and molecular graphs. Each modality's latent representation is derived using pretrained open-source models, ensuring a robust and scalable feature extraction process. For the SMILES modality, we utilize the encoder of the SMI-TED foundation model (6). This large-scale, open-source encoder-decoder model was pre-trained on a meticulously curated dataset of 91 million SMILES samples from PubChem, encompassing a total of 4 billion molecular tokens. For the SELFIES modality, we employ the SELFIES-TED model (13), an encoder-decoder architecture based on BART. This model was trained on molecular representations using the ZINC-22 (15) and PubChem (16) datasets, ensuring effective encoding of SELFIES representations. For the molecular graph modality, we leverage the MHG-GED model (14), an autoencoder that integrates Graph Neural Networks (GNNs) with Molecular Hypergraph Grammar (MHG), originally introduced in MHG-VAE (17). MHG-GNN encodes molecular structures as graphs, employing a Graph Isomorphism Network (GIN) that incorporates edge information to generate meaningful latent embeddings. To simulate real-world scenarios with missing data, modality-specific embeddings were randomly omitted during training phases.

As a preliminary analysis, we evaluate the performance of our proposed fusion method across six classification tasks from the MoleculeNet dataset. The evaluation includes comparisons between the respective unimodal, multimodal by naive concatenation and our proposed dynamic fusion method. The results are summarized in Table 1. As observed, multimodal by naïve concatenation generally outperforms unimodal approaches. However, its performance varies significantly based on the combination of modalities, and as the number of modalities increases, so does the computational overhead associated with identifying optimal modality combinations. Additionally, naïve concatenation leads to increased feature dimensionality, which can introduce redundancy and inefficiency. In contrast, our dynamic fusion approach surpasses unimodal cases and is competitive to naïve concatenation methods in majority of the tasks. By incorporating intra- and inter-modal gating mechanisms, our approach adaptively selects and fuses the most informative features while effectively handling missing modalities. Furthermore, unlike naïve concatenation, which requires paired data for training, our method remains robust even in scenarios where certain modalities are absent, making it a more flexible and scalable solution for multimodal molecular representation learning.

To further evaluate the robustness of the proposed framework, we investigated performance degradation under missing-modality conditions. Figure 2 illustrates the relative percentage drop in predictive accuracy when one or more modalities were masked during inference. Unlike naïve concatenation approaches, which cannot handle incomplete inputs, our dynamic fusion framework degrades gracefully as modalities are removed. For most datasets, removing a single modality leads to only marginal reductions in performance, confirming that the fused embedding preserves redundant information across representations. Interestingly, in some cases performance slightly improves when a modality is absent, suggesting that dynamic gating and attention effectively suppress noisy or less informative features, and raising promising directions for further study on modality reliability. When multiple modalities are simultaneously masked, performance decreases more noticeably, yet the model still avoids catastrophic collapse and maintains usable accuracy without retraining. The "all missing" case, where up to 25% of modality inputs are randomly masked, demonstrates that the framework can still preserve meaningful predictive capability despite substantial input degradation. Overall, these findings highlight that adaptive gating, attention, and reconstruction not only mitigate the effects of missing modalities but also reduce redundancy compared to static fusion strategies, making dynamic fusion a robust and generalizable solution for molecular property prediction.

Finally, we evaluated reconstruction quality using cosine similarity (Figure 3). The results demonstrate that the fused embedding retains sufficient information to recover missing modalities with high fidelity,

achieving similarity scores above 0.9 in most cases. This reinforces the effectiveness of cross-modal reconstruction in maintaining alignment across modalities and providing resilience under incomplete data conditions.

4 Conclusion

In this work, we introduced a self-supervised dynamic multimodal fusion framework that combines intra-modal gating, inter-modal attention, and cross-modal reconstruction to generate robust molecular embeddings. Our approach consistently outperforms unimodal baselines and is competitive to naïve concatenation strategies, while offering greater flexibility in handling missing modalities. Extensive evaluations demonstrated that dynamic fusion degrades gracefully when modalities are absent, avoids redundancy, and can even reconstruct missing embeddings with high accuracy.

Table 1: ROC-AUC on classification benchmarks (bold = best per column).

	bace	bbbp	clintox	hiv	tox21	sider
SMILES (SMI-TED)	0.863	0.907	0.907	0.789	0.768	0.665
SELFIES (SELFIES-TED)	0.862	0.940	0.878	0.797	0.717	0.642
Graph (MHG-GED)	0.860	0.927	0.839	0.820	0.770	0.666
SMILES $\oplus$ SELFIES	0.855	0.952	**0.921**	0.811	0.759	**0.674**
SELFIES $\oplus$ Graph	0.882	0.949	0.904	0.804	**0.792**	0.667
SMILES $\oplus$ Graph	**0.887**	0.915	0.906	**0.824**	0.760	0.672
SMILES $\oplus$ SELFIES $\oplus$ Graph	0.862	**0.954**	0.917	0.814	0.770	0.670
Dynamic Fusion	0.875	0.939	0.895	0.815	0.749	0.673

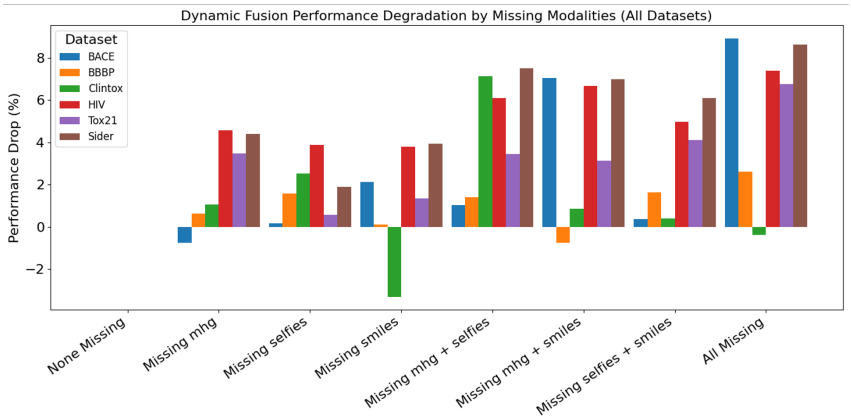

Figure 2: Dynamic fusion performance degradation due to missing modalities

Figure 3: Missing Modality Reconstruction Quality (Cosine Similarity)

References

[1] J. Shen and C. A. Nicolaou, "Molecular property prediction: recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, vol. 32, pp. 29–36, 2019.

[2] X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu, and H. Wang, "Geometry-enhanced molecular representation learning for property prediction," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 127–134, 2022.

[3] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, and T. Langer, "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, 2020.

[4] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta-2: Towards chemical foundation models," *arXiv preprint arXiv:2209.01712*, 2022.

[5] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[6] E. Soares, V. Shirasuna, E. V. Brazil, R. Cerqueira, D. Zubarev, and K. Schmidt, "A large encoder-decoder family of foundation models for chemical language," *arXiv preprint arXiv:2407.20267*, 2024.

[7] I. Priyadarsini, S. Takeda, L. Hamada, E. V. Brazil, E. Soares, and H. Shinohara, "Selfbart: A transformer-based molecular representation model using selfies," *arXiv preprint arXiv:2410.12348*, 2024.

[8] E. Soares, N. Park, E. V. Brazil, and V. Y. Shirasuna, "A large encoder-decoder polymer-based foundation model," in *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

[9] S. Gong, S. Wang, T. Zhu, Y. Shao-Horn, and J. C. Grossman, "Multimodal machine learning for materials science: composition-structure bimodal learning for experimentally measured properties," *arXiv preprint arXiv:2309.04478*, 2023.

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Challenges and applications in multimodal machine learning," *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pp. 17–48, 2018.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[12] E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira, and K. Schmidt, "An open-source family of large encoder-decoder foundation models for chemistry," *Communications Chemistry*, vol. 8, no. 1, p. 193, 2025.

[13] I. Priyadarsini, S. Takeda, L. Hamada, E. V. Brazil, E. Soares, and H. Shinohara, "Selfies-ted: A robust transformer model for molecular representation using selfies,"

[14] A. Kishimoto, H. Kajino, M. Hirose, J. Fuchiwaki, I. Priyadarsini, L. Hamada, H. Shinohara, D. Nakano, and S. Takeda, "Mhg-gnn: Combination of molecular hypergraph grammar with graph neural network," 2023.

[15] B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz, and J. J. Irwin, "Zinc-22 a free multi-billion-scale database of tangible compounds for ligand discovery," *Journal of chemical information and modeling*, vol. 63, no. 4, pp. 1166–1176, 2023.

[16] S. Kim, J. Chen, A. Gindulyte, J. He, S. He, B. A. Shoemaker, P. A. Thiessen, E. E. Bolton, G. Fu, L. Han, *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2016.

[17] H. Kajino, "Molecular hypergraph grammar with its application to molecular optimization," in *ICML*, pp. 3183–3191, 2019. Also see the supplementary material available at `http://proceedings.mlr.press/v97/kajino19a/kajino19a-supp.pdf`.