DYNAMIC FUSION FOR A MULTIMODAL FOUNDATION MODEL FOR MATERIALS

Indra Priyadarsini
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

ABSTRACT

Recent advances in the field of AI and machine learning have revolutionized applications in material science. The rapid advancement has resulted several large scale molecular representation models trained on data across various modalities and domains. Multi-modal learning and fusion approaches attempt to adeptly capture these representations from different modalities to obtain richer insights compared to unimodal approaches. However, traditional multi-modal fusion techniques fail to dynamically adjust modality importance and often lead to suboptimal performance due to redundancy or missing modalities. In this work, we propose a Dynamic Multi-Modal Fusion approach, where a learnable gating mechanism assigns importance weights to different modalities dynamically, ensuring that complementary modalities contribute meaningfully. Our preliminary evaluations on Moleculenet dataset demonstrate that the proposed method improves multi-modal fusion efficiency, enhances robustness to missing data, and leads to superior performance on downstream tasks for property prediction.

1 INTRODUCTION

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have significantly transformed the field of material discovery. These advancements have led to the development of large-scale representation models trained on diverse data modalities such as SMILES, SELFIES, molecular graphs, spectra, molecular properties, etc. Such models span multiple domains, such as polymers, pharmaceuticals, crystalline materials, etc. 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 Shen & Nicolaou (2019); Fang et al. (2022); Wieder et al. (2020); Ahmad et al. (2022); Ross et al. (2022); Soares et al. (2024b); Priyadarsini et al. (2024); Soares et al. (2024a).

Although unimodal models effectively capture domain-specific information from their respective data modalities, a more holistic understanding of materials can be achieved by integrating information from multiple modalities. In various fields such as computer vision, natural language processing, healthcare, and autonomous systems, multimodal models have demonstrated superior performance by leveraging complementary information from different data sources Gong et al. (2023); Baltrušaitis et al. (2018a;b). By integrating and processing information from multiple modalities, multimodal models offer enhanced robustness and improved feature extraction, leading to deeper insights compared to unimodal approaches.

Traditional multimodal fusion strategies often rely on simple concatenation techniques that merge unimodal representations. However, these methods assume paired data availability and fail to address challenges such as data scarcity, missing modalities, and the dynamic relevance of different representations. In this work, we propose a simple dynamic multimodal fusion approach that efficiently combines unimodal representations while adapting to available information to capture a more comprehensive molecular representation.

2 BACKGROUND

Molecular representation learning follows two stages: pretraining and downstream fine-tuning. During pretraining, models learn general representations from large datasets associated with a specific modality (e.g., SMILES, molecular graphs) or domain (e.g., polymers, crystals). These pretrained models are then fine-tuned for specific downstream tasks like molecular property prediction. However, capturing comprehensive molecular representations remains challenging due to data scarcity and molecular complexity.

To mitigate these issues, multimodal models integrate multiple data sources, capturing richer molecular features than unimodal approaches. Traditional multimodal learning employs early fusion and late fusion. Early fusion combines modalities during pretraining, enhancing feature learning but suffering from scalability limitations and requiring large multimodal datasets. Late fusion independently trains unimodal representations before combining them using concatenation, CLIP-based alignment, or attention mechanisms. While more adaptable, late fusion struggles with optimally weighting modality importance and handling missing data.

Several multimodal fusion techniques have been explored in the past. Concatenation-based fusion simply stacks modality representations resulting in very high dimensional feature vectors, assuming equal informativeness. It also relies on the availability of paired data for the modalities and thus struggles with handling missing modality data. More advanced attention mechanisms use self-attention to highlight key intra-modal features and cross-modal attention to enhance interactions between modalities. While effective, these methods introduce significant computational overhead and also overlook handling missing modalities.

Despite advancements in multimodal learning, several challenges persist. Modality redundancy can lead to inefficiencies, while handling missing modalities remains difficult as real-world datasets often contain incomplete information. Modality collapse, where the model over-relies on a dominant modality, can limit insights. Additionally, computational complexity poses scaling challenges, particularly for attention-based fusion models requiring extensive resources. To advance molecular representation learning, developing simple and efficient multimodal fusion strategies is thus crucial.

3 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. The framework has two key components: an intra-modal gating network, and an inter-

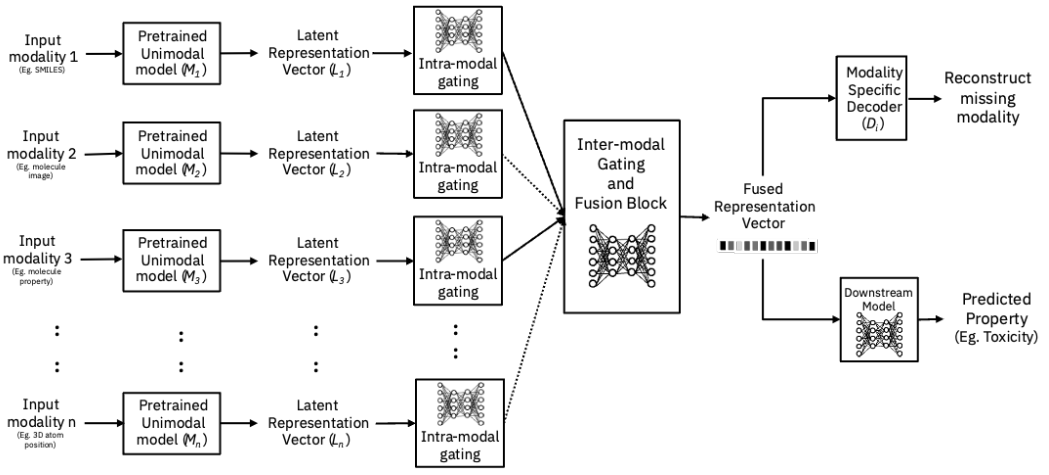

Figure 1: Block diagram of the proposed dynamic fusion model

modal gated fusion block. Our framework effectively weighs and fuses embeddings from multiple modalities through a learnable network.

3.1 INTRA-MODAL GATING

Each modality's representation is derived from its respective unimodal model. Traditionally, these representations are directly employed as features in downstream models for tasks such as property prediction. However, not all dimensions carry equally valuable information. High-dimensional representations necessitate effective feature selection to identify the most salient features. To address this, intra-modal gating employs a soft gating mechanism that dynamically assigns importance weights to each dimension of the latent representation from the unimodal model:

$$g = \text{Softmax}(W_s[X_1, X_2, \ldots, X_N]), \tag{1}$$

where g represents modality selection weights summing to 1, W_s is a learnable weight matrix that maps the latent embeddings X of dimension N to the soft gating space. The softmax function ensures smooth and differentiable weighting. The intra-modality gated feature representation is computed as:

$$\tilde{X}_i = g_i \odot X_i \tag{2}$$

where $\odot$ represents element-wise multiplication, ensuring that only the most relevant features within each modality contribute to the next stage.

3.2 INTER-MODAL GATED FUSION

The second and final stage, inter-modal gated fusion, constructs the fused representation Z_{fused} by dynamically integrating modalities that contain useful information. This is achieved by computing a weighted sum of the embeddings obtained from the intra-modal gating stage:

$$Z_{\text{fused}} = \sum_{i=1}^{N} s_i \tilde{X}_i \tag{3}$$

where s_i denotes the gating weight assigned to modality i, ensuring that the fusion process remains both adaptive and differentiable. This weighting mechanism ensures that all modalities contribute proportionally, rather than strictly selecting a single modality. Also, if there are any missing modalities, the corresponding gating weight ensures eliminating its selection, thereby handling missing modality cases as well.

3.3 TRAINING OBJECTIVE

The proposed dynamic fusion model is trained to reconstruct the original representations from the fused representation. By optimizing a reconstruction loss, the fusion scheme remains self-supervised and does not require labeled data.

A common challenge in gating mechanisms is modality collapse, where the model disproportionately relies on a single dominant modality while disregarding others. To counteract this, we introduce an entropy regularization term that promotes diversity in modality selection:

$$\mathcal{L}_{\text{entropy}} = -\sum_{i=1}^{N} s_i \log(s_i + \epsilon), \tag{4}$$

where ϵ is a small constant added for numerical stability. This term penalizes extreme weight distributions, encouraging the model to maintain a balanced contribution across multiple modalities.

Thus, the overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruct}} + \lambda \mathcal{L}_{\text{entropy}}, \tag{5}$$

where λ is a hyperparameter controlling the strength of the entropy regularization. A higher λ enforces greater modality diversity, while a lower value allows the model to prioritize dominant modalities when necessary.

By integrating soft gating with entropy regularization, our approach ensures that modality fusion remains both adaptive and resilient to missing or noisy modalities.

4 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 Soares et al. (2024b). 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, an encoder-decoder architecture based on BART. This model was trained on molecular representations using the ZINC-22 Tingle et al. (2023) and PubChem Kim et al. (2016) datasets, ensuring effective encoding of SELFIES representations. For the molecular graph modality, we leverage the MHG-GED model Kishimoto et al. (2023), an autoencoder that integrates Graph Neural Networks (GNNs) with Molecular Hypergraph Grammar (MHG), originally introduced in MHG-VAE Kajino (2019). 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 both training and testing phases.

As a preliminary analysis, we evaluate the performance of our proposed fusion method across five 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 both unimodal and naïve concatenation methods in 4 out of 5 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.

	Modalities	**BBBP**	**BACE**	**ClinTox**	**Tox21**	**Sider**
Baseline	Morgan Fingerprint	93.0	88.5	82.8	66.8	68.2
Unimodal	SELFIES (SELFIES-TED)	94.4	85.2	88.7	72.2	63.9
	Graph (MHG-GED)	92.2	86.9	84.6	75.3	65.2
	SMILES (SMI-TED)	91.7	86.5	93.4	69.9	61.1
Naive concatenation	SELFIES $\oplus$ Graph	95.9	86.7	92.3	76.8	64.5
	SELFIES $\oplus$ SMILES	96.6	86.3	88.3	75.1	63.4
	SMILES $\oplus$ Graph	92.2	88.4	93.4	75.3	64.1
	SELFIES $\oplus$ SMILES $\oplus$ Graph	96.2	87.1	89.9	75.0	65.6
Proposed	Dynamic Fusion	95.4	91.0	94.8	80.2	65.7

Table 1: Performance comparison across various modalities on different datasets.

5 CONCLUSION

In this work, we introduced a dynamic multimodal fusion model designed to enhance molecular representation learning by adaptively integrating diverse data modalities. Our approach effectively addresses challenges such as missing modalities, modality redundancy, and computational inefficiency by employing intra-modal gating and inter-modal gated fusion mechanisms. Preliminary analysis demonstrate that our method outperforms both unimodal and naive concatenation-based fusion methods across multiple molecular property prediction tasks. By leveraging adaptive soft gating and entropy regularization, our model ensures robust and flexible fusion while mitigating modality collapse. Our proposed framework provides a scalable and generalizable solution for multimodal learning in molecular representation tasks. Future work will explore expanding the model to incorporate additional modalities and further optimize computational efficiency.

REFERENCES

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

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe 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, 2018a.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018b.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Sheng Gong, Shuo Wang, Taishan Zhu, Yang Shao-Horn, and Jeffrey C Grossman. Multimodal machine learning for materials science: composition-structure bimodal learning for experimentally measured properties. *arXiv preprint arXiv:2309.04478*, 2023.

Hiroshi 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`.

Sunghwan Kim, Jie Chen, Asta Gindulyte, Jane He, Siqian He, Benjamin A Shoemaker, Paul A Thiessen, Evan E Bolton, Gang Fu, Lianyi Han, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Akihiro Kishimoto, Hiroshi Kajino, Masataka Hirose, Junta Fuchiwaki, Indra Priyadarsini, Lisa Hamada, Hajime Shinohara, Daiju Nakano, and Seiji Takeda. Mhg-gnn: Combination of molecular hypergraph grammar with graph neural network, 2023.

Indra Priyadarsini, Seiji Takeda, Lisa Hamada, Emilio Vital Brazil, Eduardo Soares, and Hajime Shinohara. Self-bart: A transformer-based molecular representation model using selfies. *arXiv preprint arXiv:2410.12348*, 2024.

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

Jie Shen and Christos A Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36, 2019.

Eduardo Soares, Nathaniel Park, Emilio Vital Brazil, and Victor Yukio Shirasuna. A large encoder-decoder polymer-based foundation model. In *AI for Accelerated Materials Design-NeurIPS 2024*, 2024a.

Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267*, 2024b.

Benjamin I Tingle, Khanh G Tang, Mar Castanon, John J Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S Moroz, and John J Irwin. Zinc-22 a free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of chemical information and modeling*, 63 (4):1166–1176, 2023.

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*, 37:1–12, 2020.