

MULTIMODAL FUSION WITH RELATIONAL LEARNING FOR MOLECULAR PROPERTY PREDICTION

Zhengyang Zhou^a, Yunrui Li^a, Pengyu Hong^a, Hao Xu^{b*}

^aDepartment of Computer Science, Brandeis University, Waltham, MA 02453, USA

^bDepartment of Medicine, Harvard Medical School, Boston, MA 02115, USA

ABSTRACT

Graph-based molecular representation learning is essential for predicting molecular properties in drug discovery and materials science. Despite its importance, current approaches struggle with capturing the intricate molecular relationships and often rely on limited chemical knowledge during training. Multimodal fusion, which integrates information from molecular graph and other data modalities, has emerged as a promising avenue for enhancing molecular property prediction. However, existing studies have explored only a narrow range of modalities, and the optimal integration stages for multimodal fusion remain largely unexplored. Furthermore, the reliance on auxiliary modalities poses challenges, as such data is often unavailable in downstream tasks. Here, we present MMFRL (Multimodal Fusion with Relational Learning), a framework designed to address these limitations by leveraging relational learning to enrich embedding initialization during multimodal pre-training. MMFRL enables downstream models to benefit from auxiliary modalities, even when these are absent during inference. We also systematically investigate modality fusion at early, intermediate, and late stages, elucidating their unique advantages and trade-offs. Using the MoleculeNet benchmarks, we demonstrate that MMFRL significantly outperforms existing methods with superior accuracy and robustness. Beyond predictive performance, MMFRL enhances explainability, offering valuable insights into chemical properties and highlighting its potential to transform real-world applications in drug discovery and materials science.

1 INTRODUCTION

Graph representation learning for molecules has gained significant attention in drug discovery and materials science, as it effectively encapsulates molecular structures and enables the effective investigation of structure-activity relationships (Schneider et al., 2020; Wieder et al., 2020; Zhang et al., 2022; Fang et al., 2022; Wang et al., 2023; Chen et al., 2024). In this paradigm, atoms are usually treated as nodes and chemical bonds as edges, effectively encapsulating the connectivity that define molecular behaviors. However, it poses significant challenges due to intricate relationships among molecules and the limited chemical knowledge utilized during training.

Contrastive Learning (CL) is often employed to study relationships among molecules. The primary focus within the domain of contrastive learning applied to molecular graphs centers on 2D-2D graph comparisons. Noteworthy representative examples: InfoGraph (Sun et al., 2019) maximizes the mutual information between the representations of a graph and its substructures to guide the molecular representation learning; GraphCL (You et al., 2020), MoCL (Sun et al., 2021), and MolCLR (Wang et al., 2022b) employs graph augmentation techniques to construct positive pairs; MoLR (Wang et al., 2022a) establishes positive pairs with reactant-product relationships. In addition to 2D-2D graph contrastive learning, there are also noteworthy efforts exploring 2D-3D and 3D-3D contrastive learning in the field. 3DGCL (Moon et al., 2023) is 3D-3D contrastive learning model, establishing positive pairs with conformers from the same molecules. GraphMVP (Liu et al., 2022b), GeomGCL (Li et al., 2022), and 3D Informax (Stärk et al., 2022) proposes 2D-3D view contrastive learning approaches. To conclude, 2D-2D and 3D-3D comparisons are intra-modality contrastive learning, as only one graph encoder is employed in these studies. However, these approaches often focus on the motif and graph levels, leaving atom-level contrastive learning less explored. For example, consider Thalidomide: while the (*R*)- and (*S*)-enantiomers share the same topological graph and differ only at a single chiral center, their biological activities are drastically different—the (*R*)-enantiomer is effective in treating morning sickness, whereas the (*S*)-enantiomer causes severe birth defects. In other words, the (*R*)- and (*S*)-enantiomers are similar in terms of topological structure but dissimilar in terms of biological activities. Thus, a more sophisticated approach is

*Corresponding Author: haxu@bwh.harvard.edu

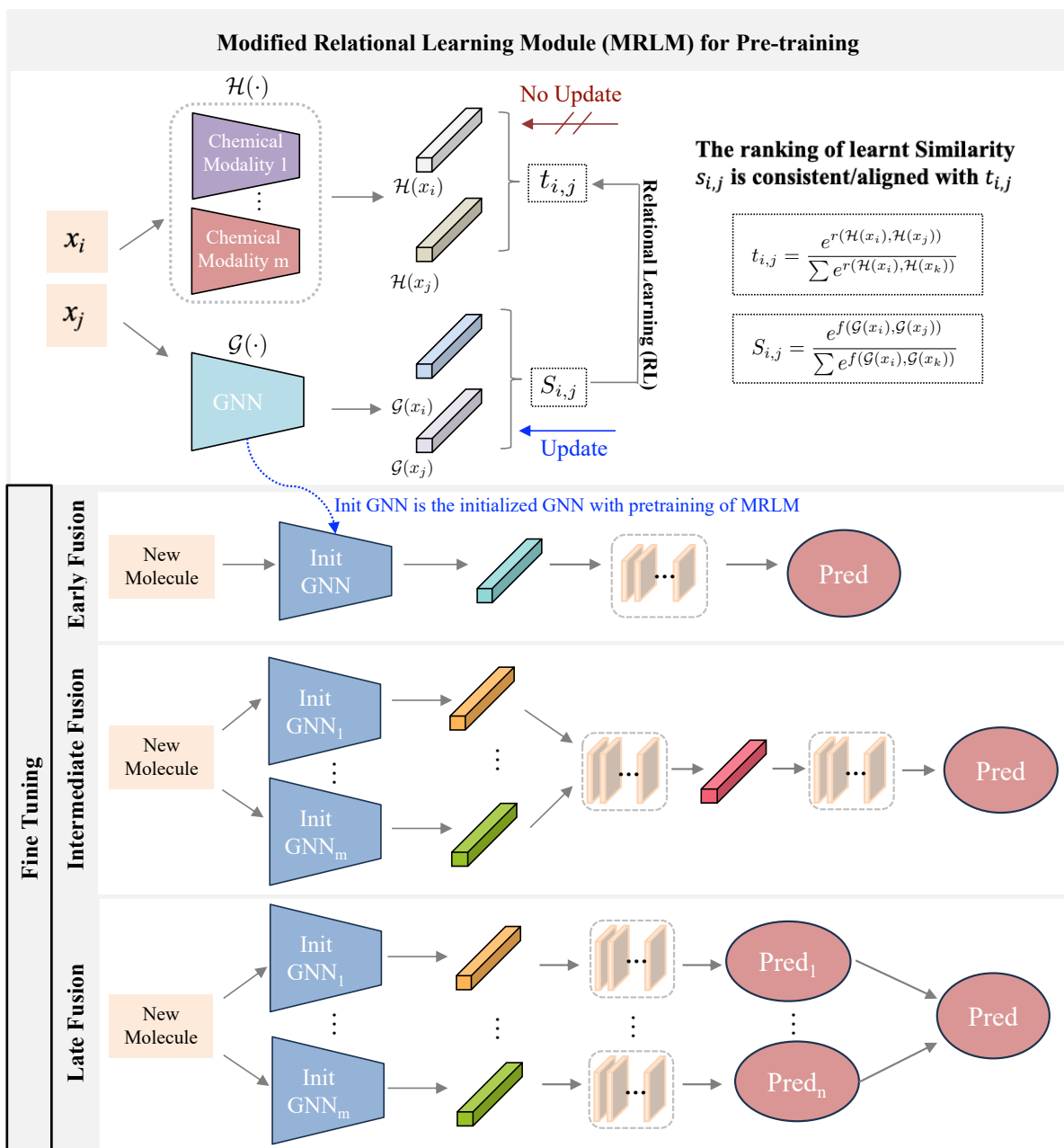


Figure 1: **Multimodal Fusion with Relational Learning for Molecular Property Prediction (MMFRL)**. This figure shows our proposed idea about how to transfer the knowledge from other modalities and use fusion to improve the performance further. Unlike the general contrastive learning framework shown in Appendix Figure A.2, MMFRL does not need to define positive or negative pairs and is capable of learning continuous ordering from target similarity. In Early Fusion, a single Init GNN is created by combining all modality information during pretraining. In Intermediate and Late Fusion, each modality has its own initialized GNN.

required to tackle these scenarios. A potential solution would be to use continuous metrics within a multi-view space, enabling a more comprehensive understanding of these complex molecular relationships.

There are multiple approaches for such similarity learning. One approach of them is instance-wise discrimination, which involves directly assessing the similarity between instances based on their latent representations or features. (Wu et al., 2018b). Naive instance-wise discrimination relies on pairwise similarity, leading to the development of contrastive loss (Hadsell et al., 2006). Although there are improved loss functions such as triplet loss (Hoffer & Ailon,

2015), quadruplet loss (Law et al., 2013), lifted structure loss (Oh Song et al., 2016), N-pairs loss (Sohn, 2016), and angular loss (Wang et al., 2017), these methods still fall short in thoroughly capturing relationships among multiple instances simultaneously (Wang et al., 2019). To address this limitation, a joint multi-similarity loss has been proposed, incorporating pair weighting for each pair to enhance instance-wise discrimination (Wang et al., 2019; Zhang et al., 2021). However, these pair weightings requires the manual categorization of negative and positive pairs, as distinct weights are assigned to losses based on their categories. In this case, we can borrow the idea of Relational Learning (Zheng et al., 2021) from computer vision by using different augmented views of the same instance tasks to similar features, while allowing for some variability. This approach captures the essential characteristics of the instance in a continuous scale, promoting relative consistency across the views without requiring them to be identical. By doing so, it enhances the model’s ability to generalize and recognize underlying patterns in the data.

Besides, in order to enable a multi-view analysis from diverse sources is essential for improving molecule analysis, we can apply the Multi-Modality Fusion (Lahat et al., 2015; Khaleghi et al., 2013; Poria et al., 2015; Ramachandram & Taylor, 2017; Pawłowski et al., 2023; Manzoor et al., 2023; Priessner et al., 2024). It combines diverse heterogeneous data (e.g. text, images, graph) to create a more comprehensive understanding of complex scenarios. This approach leverages the strengths of each modality, potentially improving performance in tasks like sentiment analysis or medical diagnosis. Although challenging to implement due to the need to align different data streams, successful fusion can provide insights that surpass those obtainable from individual modalities, advancing AI and data-driven decision-making. In particular, the way to fuse different modality should also depends on the dominance of each modality (Pawłowski et al., 2023). However, when it comes to multimodal learning for molecules, we often encounter data availability and incompleteness issues. This raises a critical question: how can multimodal information be effectively leveraged for molecular property reasoning when such data is absent in downstream tasks? Recent studies have demonstrated the effectiveness of pretraining molecular Graph Neural Networks (GNNs) by integrating additional knowledge sources (Wang et al., 2021; 2022b; Liu et al., 2022a; Xu et al., 2023a). Building on this foundation, a promising solution is to pretrain multiple replicas of molecular GNNs, with each replica dedicated to learning from a specific modality. This approach allows downstream tasks to benefit from multimodal data that is not accessible during fine-tuning, ultimately improving representation learning.

Facing these challenges and opportunities, we propose MMFRL (Multimodal Fusion with Relational Learning for Molecular Property Prediction), a novel framework features relational learning (RL) and multimodal fusion (MMF). RL utilizes a continuous relation metric to evaluate relationships among instances in the feature space (Balcan & Blum, 2006; Wen et al., 2023). Our major contribution comprises three aspects: **Conceptually**: We introduce a modified relational learning metric for molecular graph representation that offers a more comprehensive and continuous perspective on inter-instance relations, effectively capturing both localized and global relationships among instances. To the best of our knowledge, this is the first work to demonstrate such generalized relational learning metric for molecular graph representation. **Methodologically**: Our proposed modified relational metric captures complex relationships by converting pairwise self-similarity into relative similarity, which evaluates how the similarity between two elements compares to the similarity of other pairs in the dataset. In addition, we integrate these metrics into a fused multimodal representation, which has the potential to enhance performance, allowing downstream tasks to leverage modalities that are not directly accessible during fine-tuning. **Empirically**: MMFRL excels in various downstream tasks for Molecular Property Predictions. Last but not least, we demonstrate the explainability of the learned representations through two post-hoc analysis. Notably, we explore minimum positive subgraphs and maximum common subgraphs to gain insights for further drug molecule design.

2 RESULTS

2.1 THE EFFECTIVENESS OF PRE-TRAINING

We first illustrate the impact of pre-training initialization on performance on DMPNN (Yang et al., 2019). As shown in Table 1, the average performance of pre-trained models outperform the non-pre-trained model in all tasks except for Clintox. The results of various downstream tasks indicate that different tasks may prefer different modalities. Notably, the model pre-trained with the NMR modality achieves the highest performance across three classification tasks. Similarly, the model pre-trained with the Image modality excels in three tasks, two of which are regression tasks related to solubility, aligning with findings from prior literature (Xu et al., 2023a). Additionally, the model pre-trained with The fingerprint method achieves the best performance in two tasks, including MUV, which has the largest dataset.

Table 1: Study on the performances of MMFRL_{Unimodality}. The best results are denoted in bold, and the second-best are indicated with underlining among the five modalities. The first 8 tasks are for classification under evaluation of ROC-AUC, while the last three are for regression with evaluation of RMSE.

DATA SET	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	Tox21	ToxCast	ESOL	FREE SOLV	LIPO
SMILES	92.9±1.5	90.9±3.3	64.9±0.3	78.2±1.9	83.3±1.1	80.1±2.5	<u>85.7±1.2</u>	70.5±2.5	0.811±0.109	<u>1.623±0.168</u>	0.539±0.017
NMR _{SPECTRUM}	91.0±2.0	93.2±2.7	68.1±1.5	87.7±6.5	80.9±5.0	80.9±5.0	85.1±0.4	71.1±0.8	0.844±0.123	2.417±0.495	0.609±0.031
IMAGE	<u>93.1±2.4</u>	92.9±1.8	65.3±1.5	86.2±6.5	82.3±0.6	<u>78.7±1.7</u>	86.0±1.0	<u>71.0±1.6</u>	0.761±0.068	1.648±0.045	0.537±0.005
FINGERPRINT	92.9±2.3	91.7±3.6	<u>65.6±0.7</u>	<u>87.5±6.0</u>	81.2±2.5	82.9±3.1	85.3±1.3	70.0±1.4	<u>0.808±0.071</u>	1.437±0.134	0.565±0.017
NMR _{PEAK}	93.4±2.7	89.3±1.7	62.8±2.1	86.1±5.4	82.1±0.4	75.4±5.2	84.9±1.0	70.6±0.8	<u>0.924±0.083</u>	1.707±0.126	0.587±0.021
AVERAGE	92.8±1.9	91.4±2.7	65.3±2.0	85.0±5.7	81.8±2.2	79.4±4.0	85.4±0.9	70.6±1.3	0.830±0.094	1.766±0.394	0.586±0.048
MAX	93.4±2.7	93.2±2.7	68.1±1.5	87.7±6.5	83.3±1.1	82.9±3.1	86.0±1.0	71.1±0.8	0.761±0.068	1.437±0.134	0.537±0.005
NO PRE-TRAINING	91.9±3.0	85.2±0.6	57.0±0.7	90.6±0.6	77.1±0.5	78.6±1.4	75.9±0.7	63.7±0.2	1.050±0.008	2.082±0.082	0.683±0.016

Table 2: Overall performances (ROC-AUC) on classification downstream tasks. The best results are denoted in bold, and the second-best are indicated with underlining. For early fusion of MMFRL, all the predefined weight of each modality are 0.2. (Note: N-Gram is highly time-consuming on ToxCast.)

DATA SET	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	Tox21	ToxCast
ATTENTIVEFP	64.3±1.8	78.4±2.2	60.6±3.2	84.7±0.3	75.7±1.4	76.6±1.5	76.1±0.5	63.7±0.2
DMPNN	91.9±3.0	85.2±0.6	57.0±0.7	90.6±0.6	77.1±0.5	78.6±1.4	75.9±0.7	63.7±0.2
N-GRAM	91.2±0.3	79.1±1.3	63.2±0.5	87.5±2.7	78.7±0.4	76.9±0.7	76.9±2.7	-
GEM	72.4±0.4	85.6±1.1	<u>67.2±0.4</u>	90.1±1.3	80.6±0.9	81.7±0.5	78.1±0.1	69.2±0.4
UNI-MOL	72.9±0.6	85.7±0.2	<u>65.9±1.3</u>	<u>91.9±1.8</u>	80.8±0.3	82.1±1.3	79.6±0.5	69.6±0.1
INFOGRAPH	69.2±0.8	73.9±2.5	59.2±0.2	75.1±5.0	74.5±1.8	74.0±1.5	73.0±0.7	62.0±0.3
GRAPHCL	67.5±3.3	68.7±7.8	60.1±1.3	78.9±4.2	75.0±0.4	77.1±1.0	75.0±0.3	62.8±0.2
MOLCLR	73.3±1.0	82.8±0.7	61.2±3.6	89.8±2.7	77.4±0.6	78.9±2.3	74.1±5.3	65.9±2.1
MOLCLR _{CMPNN}	72.4±0.7	85.0±2.4	59.7±3.4	88.0±4.0	77.8±5.5	74.5±2.1	78.4±2.6	69.1±1.2
GRAPHMVP	72.4±1.6	81.2±9.0	63.9±1.2	79.1±2.8	77.0±1.2	77.7±6.0	75.9±5.0	63.1±0.4
UNIMODALITY _{avg}	92.8±1.9	91.4±2.7	65.3±2.0	85.0±5.7	81.8±2.2	79.4±4.0	<u>85.4±0.9</u>	70.6±1.3
Unimodality _{Max}	93.4±2.7	93.2±2.7	68.1±1.5	87.7±6.5	83.3±1.1	82.9±3.1	86.0±1.0	71.1±0.8
MMFRL _{early}	91.6±5.0	<u>94.3±2.4</u>	66.4±1.9	85.3±6.8	82.0±2.4	80.6±3.2	85.2±0.2	69.8±1.1
MMFRL _{intermediate}	95.4±0.7	95.1±1.0	64.3±1.2	93.4±1.1	81.2±1.3	83.5±1.6	85.1±0.1	71.9±1.1
MMFRL _{late}	<u>94.7±0.6</u>	91.6±2.6	64.2±1.2	87.0±0.4	<u>82.9±0.2</u>	82.1±1.7	77.7±0.5	70.2±0.3

2.2 OVERALL PERFORMANCE OF MMFRL

As shown in Table 2 and Table 3, MMFRL demonstrates superior performance compared to all baseline models and the average performance of DMPNN pretrained with extra modalities across all 11 tasks evaluated in MoleculeNet. Results in Tables 5 and Table 6 demonstrates our great performance compared to the baseline models on the Dud-E (Mysinger et al., 2012) and LIT-PCBA (Tran-Nguyen et al., 2020) datasets. This robust performance highlights the effectiveness of our approach in leveraging multimodal data. In particular, while individual models pre-trained on other modalities for ClinTox fail to outperform the No-pretraining model, the fusion of these pre-trained models leads to improved performance. Besides, apart from Tox21 and Sider, the fusion models significantly enhances overall performance. In particular, the intermediate fusion model stands out by achieving the highest scores in seven distinct tasks, showcasing its ability to effectively combine features at a mid-level abstraction. the late fusion model achieves the top performance in two tasks. These results underscore the advantages of utilizing various fusion strategies in multimodal learning, further validating the efficacy of the MMFRL framework.

2.3 ANALYSIS OF THE FUSION EFFECT

2.3.1 GENERAL COMPARISON AMONG VARIOUS WAYS OF FUSIONS

Early Fusion is employed during the pretraining phase and is easy to implement, as it aggregates information from different modalities directly. However, its primary limitation lies in the necessity for predefined weights assigned to each modality. These weights may not accurately reflect the relevance of each modality for the specific downstream tasks, potentially leading to suboptimal performance.

Intermediate Fusion is able to capture the interaction between modalities early in the fine-tuning process, allowing for a more dynamic integration of information. This method can be particularly beneficial when different modalities

Table 3: Overall performances (RMSE) on regression downstream tasks. The best results are denoted in bold, and the second-best are indicated with underlining. For early fusion of MMFRL, all the predefined weight of each modality are 0.2.

Data Set	ESOL	FreeSolv	Lipo
AttentiveFP	0.877 \pm 0.029	2.073 \pm 0.183	0.721 \pm 0.001
DMPNN	1.050 \pm 0.008	2.082 \pm 0.082	0.683 \pm 0.016
N-Gram _{RF}	1.074 \pm 0.107	2.688 \pm 0.085	0.812 \pm 0.028
N-Gram _{XGB}	1.083 \pm 0.082	5.061 \pm 0.744	2.072 \pm 0.030
GEM	0.798 \pm 0.029	1.877 \pm 0.094	0.660 \pm 0.008
Uni-Mol	0.788 \pm 0.029	1.620 \pm 0.035	0.603 \pm 0.010
MolCLR	1.113 \pm 0.023	2.301 \pm 0.247	0.789 \pm 0.009
MolCLR _{CMPNN}	0.911 \pm 0.082	2.021 \pm 0.133	0.875 \pm 0.003
Unimodality _{avg}	0.924 \pm 0.083	1.707 \pm 0.126	0.587 \pm 0.021
Unimodality _{Max}	0.761 \pm 0.068	1.437\pm0.134	<u>0.537\pm0.005</u>
MMFRL _{early}	1.037 \pm 0.170	2.093 \pm 0.090	0.607 \pm 0.034
MMFRL _{intermediate}	0.730\pm0.019	1.465 \pm 0.096	0.552 \pm 0.014
MMFRL _{late}	0.763 \pm 0.035	1.741 \pm 0.191	0.525\pm0.018

Table 4: Study on the performances of MMFRL_{Intermediate} with different contrastive loss functions: Contrastive Loss (CL) and Triplet Loss (TL). The best results are denoted in bold. The first 8 datasets are classification tasks evaluated using ROC-AUC, while the last three are regression tasks evaluated using RMSE. Our model outperforms other loss functions in most of the datasets.

DATA SET	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	Tox21	ToxCast	ESOL	FreeSolv	Lipo
MMFRL _{intermediate}	95.4\pm0.7	95.1\pm1.0	64.3\pm1.2	93.4\pm1.1	81.2\pm1.3	83.5\pm1.6	85.1 \pm 0.1	71.9\pm1.1	0.730\pm0.019	1.465\pm0.096	0.552\pm0.014
MMFRL _{intermediateCL}	93.2 \pm 1.5	89.7 \pm 1.3	61.1 \pm 2.6	90.6 \pm 1.7	80.9 \pm 1.7	78.2 \pm 1.3	85.7\pm0.4	70.8 \pm 0.8	0.792 \pm 0.034	2.094 \pm 0.377	0.609 \pm 0.022
MMFRL _{intermediateTL}	93.2 \pm 2.2	91.3 \pm 1.2	61.8 \pm 1.7	91.8 \pm 2.5	80.0 \pm 1.5	78.8 \pm 0.2	85.6 \pm 0.5	70.7 \pm 0.4	0.780 \pm 0.037	2.072 \pm 0.199	0.577 \pm 0.005

provide complementary information that enhances overall performance. If the modalities effectively compensate for one another’s strengths and weaknesses, Intermediate Fusion may emerge as the most effective approach.

In contrast, Late Fusion enables each modality to be explored independently, maximizing the potential of individual modalities without interference from others. This separation allows for a thorough examination of each modality’s contribution. When certain modalities dominate the performance metrics, Late Fusion can capitalize on their strengths by effectively leveraging the most informative signals.. This approach is especially useful in scenarios where the dominance of specific modalities can be leveraged to enhance overall model performance.

In addition, we conduct an ablation study to evaluate the performance of our proposed loss functions against two traditional contrastive learning losses—Contrastive Loss and Triplet Loss—in the context of intermediate fusion. The experimental results as shown in Table 4 demonstrate that our proposed methods outperform the baseline approaches across the majority of tasks in the MoleculeNet dataset, thereby highlighting the superiority of our approach.

2.3.2 EXPLAINABILITY OF LEARNT REPRESENTATIONS

To demonstrate the interpretability of the representations learned by the proposed fusion strategies, we present the post-hoc analysis results on two tasks, ESOL and Lipo, as case studies. The results showcase that the learnt representations can capture task-specific patterns and offer valuable insights for molecular design.

ESOL with Intermediate Fusion. As presented in Table 3, the intermediate fusion method 5.3.2 exhibits superior performance on the ESOL regression task for predicting solubility. To further analyze this performance, we employed t-SNE to reduce the dimensionality of the molecule embeddings from 300 to 2, resulting in a heatmap visualized in Figure 2. The embeddings derived from individual modalities prior to fusion do not display a clear pattern, showing no smooth transition from low to high solubility. In contrast, the embeddings by intermediate fusion reveal a distinct and smooth transition in solubility values: molecules with similar solubility cluster together, forming a gradient that extends from the bottom left (indicating lower solubility) to the upper center (representing higher solubility). This trend underscores the effectiveness of the intermediate fusion approach in accurately capturing the quantitative structure-activity relationships for aqueous solubility.

Additionally, we examined the similarity between the respective embeddings prior to intermediate fusion and the resulting fused embedding, as depicted in Figure 3. Our analysis indicates that the embeddings from each modality

Table 5: Comparison of Average AUC-ROC across DUD-E dataset

Method	Average AUC-ROC
COSP (Gao et al., 2022)	90.10
Graph CNN (Torng & Altman, 2019)	88.60
Drug VQA (Zheng et al., 2020)	97.20
AttentionSiteDTI (Yazdani-Jahromi et al., 2022)	<u>97.10</u>
DrugCLIP _{FT} (Gao et al., 2023)	96.59
MMRFL-Intermediate (Ours)	98.32
MMRFL-Late (Ours)	96.78

Table 6: Overall performances (ROC-AUC) on classification downstream tasks for Lit-PCBA. The performance that are not by us is from the paper Cai et al. (2022)

Task	NB	SVM	RF	XGBoost	DNN	GCN	GAT	FP-GNN	MMRFL_Int	MMRFL_Jat
ADRBZ	55.2	53.4	49.8	50.0	83.3	<u>83.7</u>	76.8	88.6	76.3	76.0
ALDH1	69.3	76.0	74.1	75.0	75.6	73.0	73.9	76.6	78.8	<u>78.7</u>
ESR1_ago	66.1	55.2	44.8	50.0	69.0	58.7	71.3	<u>72.8</u>	76.8	31.5
ESR1_ant	54.3	63.0	53.3	52.8	58.2	67.1	<u>65.6</u>	64.2	54.3	58.8
FEN1	87.6	87.7	65.7	88.8	<u>90.1</u>	89.7	88.8	88.9	90.5	85.4
GBA	70.9	<u>77.8</u>	59.9	83.0	77.7	73.5	77.6	75.1	77.3	75.3
IDH1	88.7	80.7	49.8	50.0	67.8	81.3	<u>86.1</u>	78.7	71.0	41.6
KAT2A	65.9	61.2	53.7	50.0	59.5	62.1	66.2	63.2	71.6	<u>68.7</u>
MAPK1	68.6	66.5	57.9	59.3	70.8	66.8	69.7	77.1	<u>73.0</u>	69.6
MTORC1	59.8	59.1	53.2	63.9	63.4	66.9	61.5	58.3	<u>62.0</u>	59.0
OPRK1	53.8	53.2	49.8	50.0	<u>71.0</u>	64.4	63.6	54.5	72.8	68.6
PKM2	68.4	<u>75.3</u>	58.1	73.7	71.9	63.6	72.4	73.2	77.2	71.4
PPARG	67.5	79.3	66.9	49.0	<u>81.1</u>	78.8	78.4	82.9	69.7	69.5
TP53	64.8	60.2	60.2	64.3	70.6	74.9	70.6	<u>76.3</u>	80.7	57.3
VDR	80.4	69.0	64.4	<u>78.2</u>	79.4	77.3	78.0	77.4	73.9	73.5

exhibit low similarity with the intermediate-fused representation. This observation suggests that the modalities complement each other, collectively enhancing the resulting representation of the intermediate-fused embedding.

Lipo with Late Fusion. As detailed in Table 3, the Late Fusion method (described in Section 5.3.3) demonstrates superior performance on the Lipo regression task for predicting solubility in fats, oils, lipids, and non-polar solvents. According to Equation 11, the final prediction is determined by the respective coefficients (w_i) and predictions (p_i) from each modality.

Figure 4 shows the distributions of the coefficient values, predictions, and their products for each modality. Notably, the SMILES and Image modalities display a wide range of values, highlighting their potential to significantly influence the final predictions. This observation aligns with the strong performance achieved when pretraining using either of these two modalities, as shown in Table 1. In contrast, the NMR_{peak} values display a narrower range, indicating its role as a modifier for finer adjustments in the predictions. Furthermore, we observe that the contributions from NMR_{spectrum} and Fingerprint modalities are minimal, with their corresponding values approaching zero. This outcome highlights the advantages of the Late Fusion approach in effectively identifying and leveraging dominant modalities, thereby optimizing the overall predictive performance.

Substructure analysis with BACE. We explore the binding potential of positive inhibitor molecules targeting BACE and their associated key functional substructures, referred to as minimum positive subgraphs (MPS). To identify MPS, we employ a Monte Carlo Tree Search (MCTS) approach integrated into our BACE classification model, as implemented in RationalRL (Jin et al., 2020). MCTS, being an iterative process, allows us to evaluate each candidate substructure for its binding potential with our model. Following the determination of MPSs, we categorize the original positive BACE molecules based on their respective MPSs. By computing the binding potential difference between the original

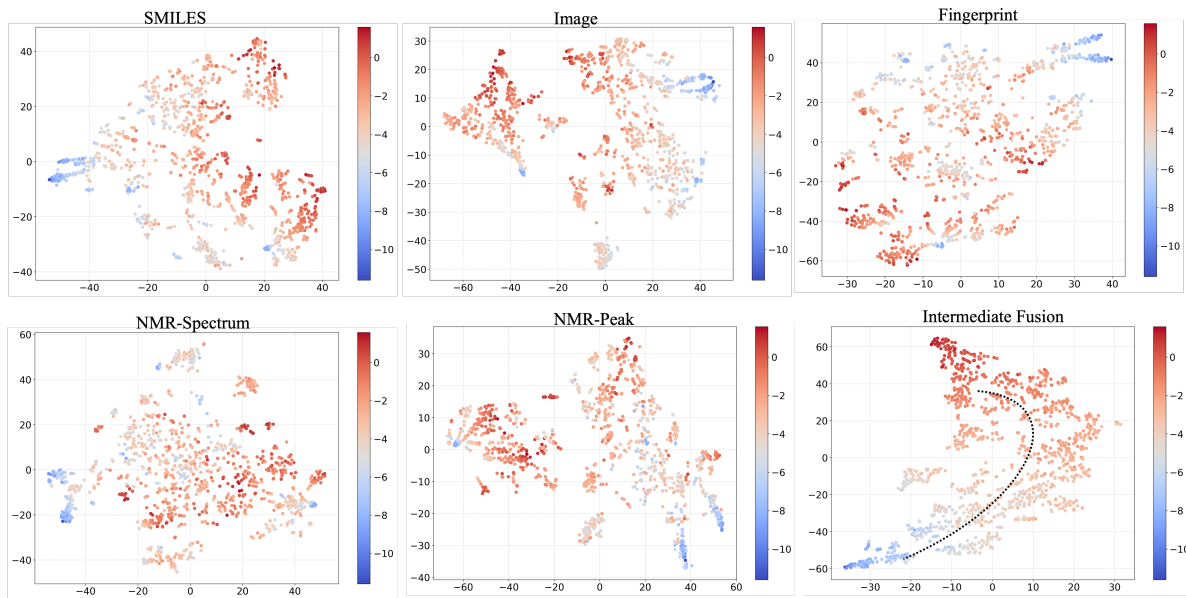


Figure 2: T-SNE visualization depicting the ESOL molecule embeddings for intermediate fusion in Section 5.3.2 alongside molecules within the highlighted region. Each point in the heatmap corresponds to the embeddings of respective molecules in ESOL, with color indicating solubility levels. Red denotes higher solubility, while blue indicates lower solubility. The embeddings derived from individual modalities prior to fusion do not display a clear pattern, the embeddings by intermediate fusion forms a gradient that extends from the bottom left (indicating lower solubility) to the upper center (representing higher solubility).

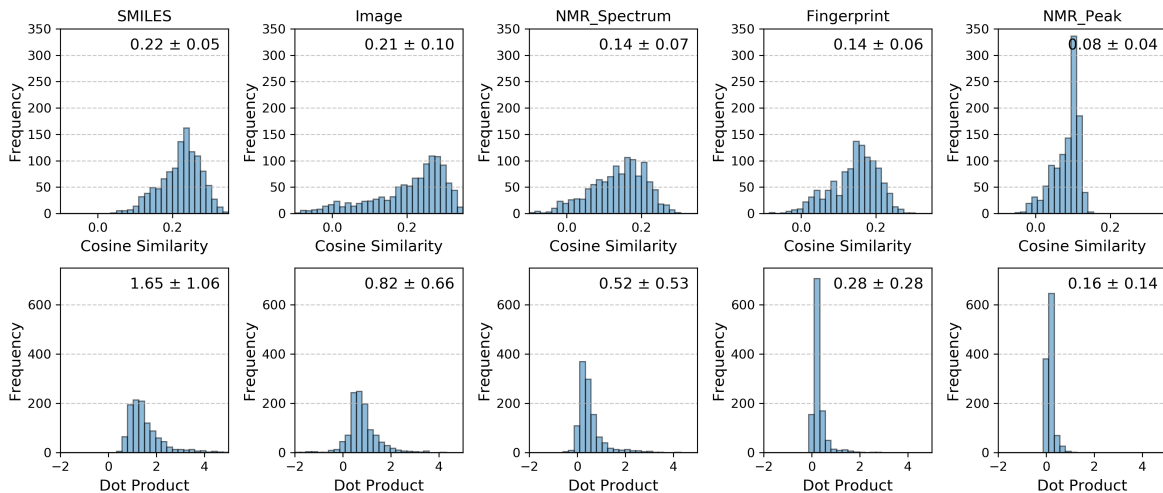


Figure 3: This figure shows the distribution of similarities between each modality and the intermediate fusion embedding for ESOL. In both Cosine Similarity and Dot Product, the embeddings from each modality exhibit low similarity with the intermediate-fused representation.

molecule and its MPS, we can identify structural features that contribute to changes in binding affinity as shown in Figure 5.

In the case of the MPS 5 group, the binding score is heavily influenced by steric effects. The top three high-performing designs (5a–5c) all feature a flexible and compact alkylated pyrazole structure (colored green), which likely facilitates better accommodation within the binding pocket. In contrast, the three lowest-performing designs (5n–5p) incorporate a more rigid and bulky (trifluoromethoxy)benzene moiety (colored red), which may introduce steric hindrance and reduce

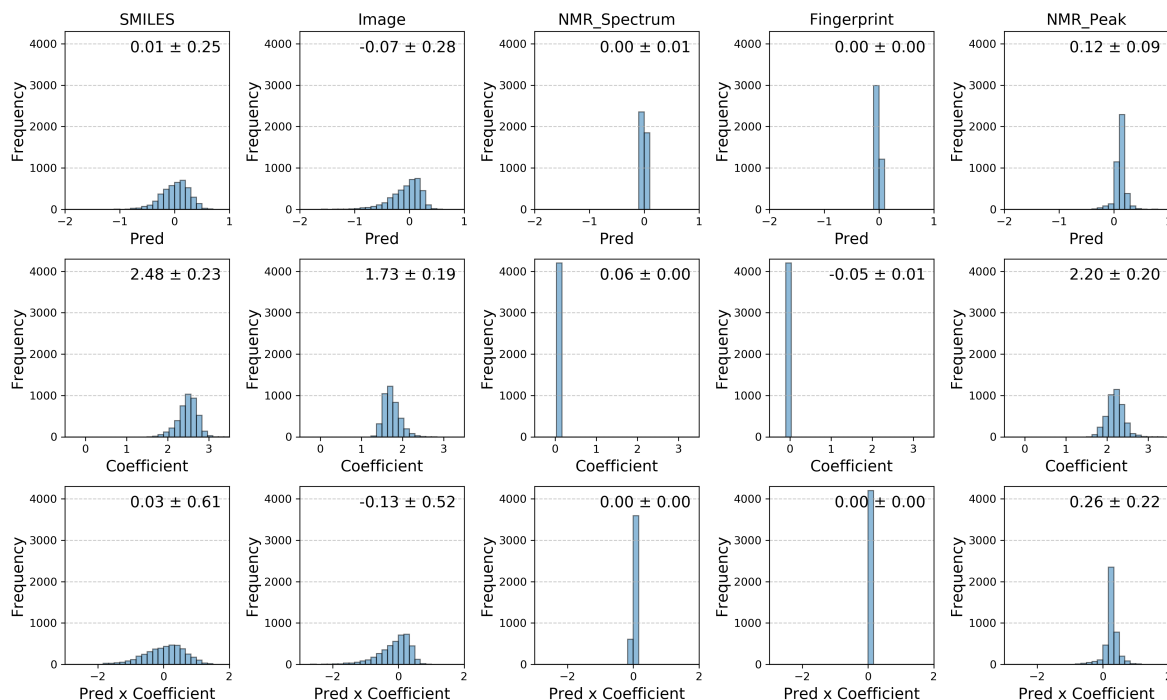


Figure 4: Lipo late fusion contribution analysis reveals that the three primary contributors are SMILES, image, and NMR_{peak}. In contrast, NMR_{spectrum} and fingerprint exhibit negligible contributions.

binding efficiency. Additionally, the pyrazole ring contains two nitrogen atoms, offering more potential for hydrogen bonding interactions with the target protein, whereas the (trifluoromethoxy)benzene group has only one oxygen atom, limiting its capacity for such interactions. This comparison highlights the importance of both molecular flexibility and functional group composition in optimizing binding affinity.

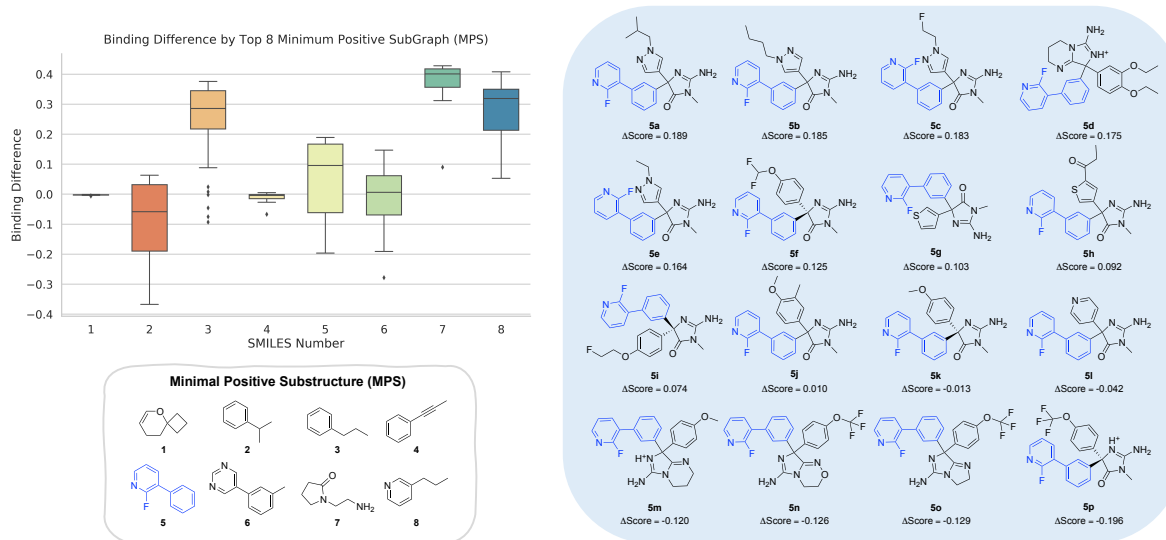


Figure 5: The left sub-figure is the boxplot of the binding difference for the respective groups of molecules by the top 8 most frequent Minimum Positive Subgraph. The right sub-figure shows the detail structure of the 5th MPS.

Table 7: Person correlation of different modalities and chosen fusion strategies across datasets.

Dataset	Smiles	Image	NMR	FP	Peak	Top 1	Concat	Pearson Gain	Strategy
BBBP	0.757	0.765	0.733	0.746	0.484	0.765	0.935	0.170	Intermediate
Bace	0.752	0.759	0.757	0.772	0.474	0.772	0.957	0.185	Intermediate
Sider	0.598	0.579	0.578	0.592	0.403	0.597	0.973	0.376	Intermediate
Hiv	0.305	0.295	0.239	0.287	0.193	0.305	0.420	0.115	Late
MUV	0.159	0.159	0.141	0.147	0.050	0.159	0.315	0.156	Intermediate
Clintox	0.577	0.604	0.546	0.577	0.405	0.639	0.920	0.281	Intermediate
tox21	0.550	0.558	0.578	0.565	0.183	0.578	0.691	0.113	Intermediate
toxcast	0.587	0.590	0.523	0.577	0.333	0.590	0.908	0.318	Intermediate
Lipo	0.782	0.795	0.623	0.780	0.542	0.795	0.920	0.125	Late
ESOL	0.958	0.960	0.893	0.947	0.705	0.960	0.999	0.039	Intermediate
FreeSolv	0.982	0.980	0.915	0.977	0.768	0.982	0.999	0.017	Intermediate

2.3.3 SENSITIVITY ANALYSIS

Choosing the most effective fusion strategy can be empirical. However, our results presented in Table 2, 3, 4, 5, and 6 provide strong evidence that our lightweight fusion strategy (early, intermediate, and late fusion) outperforms existing approaches in the literature. To guide the selection among these strategies, our intuition is as follows: if a modality is highly relevant to the downstream task, earlier fusion is likely to be more effective; otherwise, later fusion may be preferable.

To test this hypothesis, we performed a retrospective analysis to assess the sensitivity of downstream tasks to different fusion strategies. Since early fusion embeddings often lack the flexibility to adapt to individual samples, we excluded them from this analysis. Instead, we used pretrained encoders to extract embeddings for each modality and performed a simple linear regression between the embeddings and task labels. We then computed the Pearson correlation between the predicted values and the ground truth as a measure of each modality’s relevance.

For each dataset, we recorded the highest correlation across all modalities as the "Top 1" score. We then concatenated the embeddings from all modalities and repeated the regression analysis. The improvement in correlation is reported as the "Pearson Gain." A higher Pearson Gain suggests that earlier fusion of multiple modalities is more beneficial. As shown in Table 7, datasets where intermediate fusion performs best generally exhibit higher Pearson Gain compared to late fusion, supporting our intuition. However, for ESOL and FreeSolv, the correlation from a single modality is already high, making them less suitable for this analysis.

3 CONCLUSION

In summary, we introduce a novel relational learning metric for molecular graph representation that enhances the understanding of inter-instance relationships by capturing both local and global contexts. This is the first implementation of such a generalized metric in molecular graphs. Our method transforms pairwise self-similarity into relative similarity through a weighting function, allowing for complex relational insights. This metric is integrated into a multimodal representation, improving performance by utilizing modalities not directly accessible during fine-tuning. Empirical results show that our approach, MMFRL, excels in various molecular property prediction tasks. We also demonstrate detailed study about the explainability of the learned representations, offering valuable insights for drug molecule design. Despite these accomplishments, further exploration is needed to achieve more effective integration of graph- and node-level similarities. Looking ahead, we are enthusiastic about the prospect of applying our model to additional fields, such as social science, thereby broadening its applicability and impact.

4 DATASET

4.1 SELECTED MODALITIES FOR TARGET SIMILARITY CALCULATION

The following modalities are used for target similarity calculation. For details on training the corresponding encoders to obtain fixed embeddings for these modalities, please refer to Appendix Section C.1.

Fingerprint: Fingerprints are binary vectors that represent molecular structures, capturing the presence or absence of particular substructures, fragments, or chemical features within a molecule. In particular, we utilize Morgan fingerprints, which are based on the Extended-Connectivity Fingerprints (ECFP) method introduced by Rogers and Hahn (Rogers & Hahn, 2010). Specifically, we generate fingerprints using `AllChem.GetMorganFingerprintAsBitVect(mol, 2)`,

which corresponds to ECFP4 (radius = 2). Because ECFP4 is one the most effective and Interpretable Molecular Representations (Zhong & Guan, 2023).

SMILES (Simplified Molecular Input Line Entry System): SMILES offers a compact textual representation of chemical structures.

NMR (Nuclear Magnetic Resonance): NMR spectroscopy provides detailed insights into the chemical environment of atoms within a molecule (Bunzel & Ralph, 2006). By analyzing the interactions of atomic nuclei with an applied magnetic field, NMR can reveal information about the structure, dynamics, and interactions of molecules, including the connectivity of atoms, functional groups, and conformational changes. In our experiments, $\text{NMR}_{\text{spectrum}}$ provides the information about the overall information of molecule while NMR_{peak} provides the information about the individual atoms in the molecule.

Image: Images (e.g., 2D chemical structures) provide a visual representation of molecular structures.

All of the similarity calculation from the modalities above are listed in Appendix C.2.

4.2 PRE-TRAINING

NMRShiftDB-2 (Landrum, 2006) is a comprehensive database dedicated to nuclear magnetic resonance (NMR) chemical shift data, providing researchers with an extensive collection of expert-annotated NMR data for various organic compounds with molecular structures (SMILES). There are around 25,000 molecules used for pre-training and no overlap with downstream task datasets. And molecular images and graphs are generated via RDkit (RDK).

4.3 DOWNSTREAM TASKS

Our model was trained on 11 drug discovery-related benchmarks sourced from MoleculeNet (Wu et al., 2018a). Eight of them were classification tasks, including BBBP, BACE, SIDER, CLINTOX, HIV, MUV, TOX21, and ToxCas. The other three are regression tasks, including ESOL, Freesolv, and Lipo. Each dataset was divided into the train, validation, and test subsets in an 80%:10%:10% ratio using the scaffold splitter (Halgren, 1996; Landrum, 2006) from Chemprop (Yang et al., 2019; Heid et al., 2023). The scaffold splitter categorizes molecular data based on substructures, ensuring diverse structures in each subset. Molecules are partitioned into bins, with those exceeding half of the test set size assigned to training, promoting scaffold diversity in validation and test sets. Remaining bins are randomly allocated until reaching the desired set sizes, creating multiple scaffold splits for comprehensive evaluation.

The DUD-E (Directory of Useful Decoys: Enhanced) dataset (Mysinger et al., 2012) is a widely used benchmark for virtual screening, containing 102 protein targets, thousands of active compounds, and carefully selected decoys that resemble actives in physico-chemical properties but differ topologically. In contrast, LIT-PCBA (Low-Throughput Informatics-Targeted PubChem BioAssay) (Tran-Nguyen et al., 2020) offers a more realistic and challenging benchmark, derived from real experimental assays across 15 targets, with no artificial decoys and inherent data noise and imbalance. Together, they represent two ends of the spectrum in virtual screening evaluation—DUD-E with idealized conditions, and LIT-PCBA with real-world complexity. For the fine-tuning setting, We follow the same split and test approach as (Gao et al., 2023) for DUD-E and (Cai et al., 2022) for LIT-PCBA.

5 METHODS

We first explain the preliminaries, and then our proposed modified metric in relational learning to facilitate smooth alignment between graph and referred unimodality. Then, we introduce approaches for integrating multi modalities at different stages of the learning process.

5.1 MOLECULAR REPRESENTATION WITH DMPNN

The Message Passing Neural Network (MPNN) (Gilmer et al., 2017) is a GNN model that processes an undirected graph G with node (atom) features x_v and edge (chemical bond) features e_{vw} . It operates through two distinct phases: a message passing phase, facilitating information transmission across the molecule to construct a neural representation, and a readout phase, utilizing the final representation to make predictions regarding properties of interest. The primary distinction between DMPNN and a generic MPNN lies in the message passing phase. While MPNN uses messages associated with nodes, DMPNN crucially differs by employing messages associated with directed edges (Yang et al., 2019). This design choice is motivated by the necessity to prevent totters (Mahé et al., 2004), eliminating messages

passed along paths of the form $v_1 v_2 \dots v_n$, where $v_i = v_{i+2}$ for some i , thereby eliminating unnecessary loops in the message passing trajectory.

5.2 MODIFIED RELATIONAL LEARNING IN PRETRAINING

Original Relation Learning (Zheng et al., 2021) ensures that different augmented views of the same instance from computer vision tasks share similar features, while allowing for some variability. Suppose z_i is the original embedding for the i -th instance. Then z_i^1 is the embedding of first augmented view for z_i , and z_i^2 is the embedding of second augmented view for z_i . In this case, the Loss of Relational Learning (RL) is formulated as following:

$$s_{ik}^1 = \frac{\mathbb{1}_{i \neq k} \cdot \exp(z_i^1 \cdot z_k^2 / \tau)}{\sum_{j=1}^N \mathbb{1}_{i \neq j} \cdot \exp(z_i^1 \cdot z_j^2 / \tau)}$$

$$s_{ik}^2 = \frac{\mathbb{1}_{i \neq k} \cdot \exp(z_i^2 \cdot z_k^2 / \tau_m)}{\sum_{j=1}^N \mathbb{1}_{i \neq j} \cdot \exp(z_i^2 \cdot z_j^2 / \tau_m)}$$

$$L_{RL} = -\frac{1}{N} \sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq i}}^N s_{ik}^2 \log(s_{ik}^1).$$

We propose a modified relational metric by adapting the softmax function as a pairwise weighting mechanism. Let $|\mathcal{S}|$ denote the size of the instance set. The variable $s_{i,j}$ represents the learned similarity where z_i is the embedding to be trained. On the other hand, $t_{i,j}^R$ defines the target similarity that captures the relationship between the pair of instances in the given space or modality R , where z_i^R is a fixed embedding. The detailed formulation for the Loss of Modified Relational Learning (MRL) is provided below:

$$s_{i,j} = \frac{\exp(\text{sim}(z_i, z_j))}{\sum_{k=1}^{|\mathcal{S}|} \exp(\text{sim}(z_i, z_k))} \quad (1)$$

$$t_{i,j}^R = \frac{\exp(\text{sim}(z_i^R, z_j^R))}{\sum_{j=1}^{|\mathcal{S}|} \exp(\text{sim}(z_i^R, z_k^R))} \quad (2)$$

$$L_{MRL} = -\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} t_{i,j}^R \log(s_{i,j}). \quad (3)$$

Notably, unlike other similarity learning approaches (Wang et al., 2019; Zhang et al., 2021), our method does not rely on the categorization of negative and positive pairs for the pair weighting function. Additionally, our use of the softmax function ensures that the generalized target similarity $t_{i,j}$ adheres to the principles of convergence, which results in better ranking consistency between the graph modality and the auxiliary modality, compared with the original Relational Study, as follows:

Theorem 5.1 (Convergence of Modified Relational Learning Metric). *Let \mathcal{S} be a set of instances with size of $|\mathcal{S}|$, and let \mathcal{P} represent the learnable latent representations of instances in \mathcal{S} such that $|\mathcal{P}| = |\mathcal{S}|$. For any two instances $i, j \in \mathcal{S}$, their respective latent representations are denoted by \mathcal{P}_i and \mathcal{P}_j . Let $t_{i,j}$ represent the target similarity between instances i and j in a given domain, and let $d_{i,j}$ be the similarity between \mathcal{P}_i and \mathcal{P}_j in the latent space. If $t_{i,j}$ is non-negative and $\{t_{i,j}\}$ satisfies the constraint $\sum_{j=1}^{|\mathcal{S}|} t_{i,j} = 1$, consider the loss function for an instance i defined as follows:*

$$L(i) = -\sum_{j=1}^{|\mathcal{S}|} t_{i,j} \log \left(\frac{e^{d_{i,j}}}{\sum_{k=1}^{|\mathcal{S}|} e^{d_{i,k}}} \right) \quad (4)$$

then when it reaches ideal optimum, the relationship between $t_{i,j}$ and $d_{i,j}$ satisfies:

$$\text{softmax}(d_{i,j}) = t_{i,j} \quad (5)$$

For detailed proof, please refer to Appendix Section B.1.

5.3 FUSION OF MULTI-MODALITY INFORMATION IN DOWNSTREAM TASKS.

During pre-training, the encoders are initialized with parameters derived from distinct reference modalities. A critical question that arises is how to effectively utilize these pre-trained models during the fine-tuning stage to improve performance on downstream tasks.

5.3.1 EARLY STAGE: MULTIMODAL MULTI-SIMILARITY

With a set of known target similarity $\{t^R\}$ from various modalities, we can transform them to multimodal space through a fusion function. There are numerous potential designs of the fusion function. For simplicity, we take linear combination as a demonstration. The multimodal generalized multi-similarity $t_{i,j}^M$ between i^{th} and j^{th} objects can be defined as follows:

$$t_{i,j}^M = fusion(\{t^R\}) \quad (6)$$

$$= \sum w_R \cdot t_{i,j}^R \quad (7)$$

where $t_{i,j}^R$ represents the target similarity between i^{th} and j^{th} instance in unimodal space R , w_R is the pre-defined weights for the corresponding modal, and $\sum w_R = 1$. Then we can make $t_{i,j} = t_{i,j}^R$ in equation 3. Such that, it still satisfy the requirement of convergence (See proof in Appendix Section B.2). In this case, the learnt similarity during pretraining will be aligned with this new combined target similarity.

5.3.2 INTERMEDIATE STAGE: EMBEDDING CONCATENATION AND FUSION

Intermediate fusion integrates features from various modalities after their individual encoding processes and prior to the decoding/readout stage. Let $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n$ represent the feature vectors obtained from these different modalities. The resulting fused feature vector can be defined as follows:

$$\mathbf{f}_{\text{fused}} = \text{MLP}(\text{concat}(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)) \quad (8)$$

Where concat represents concatenation of the feature vectors. The fused features are then fed into a later readout function or decoder for downstream tasks prediction or classification. The MLP (Multi-Layer Perceptron) is used to reduce the dimension to be the same as \mathbf{f}_i .

5.3.3 LATE STAGE: DECISION-LEVEL

Late fusion (or decision-level fusion) combines the outputs of models trained on different modalities after they have been processed independently. Each modality is first processed separately, and their predictions are combined at a later stage.

Let p_1, p_2, \dots, p_n be the predictions (e.g., probabilities) from different modalities. The final prediction p_{final} can be computed using a weighted sum mechanism:

$$w_i = T_i(\mathbf{f}_i) \quad (9)$$

$$p_i = \text{readout}_i(\mathbf{f}_i) \quad (10)$$

$$p_{\text{final}} = \sum_{i=1}^n w_i p_i \quad (11)$$

Where w_i are the weights assigned to each modality’s prediction, and they can be adjusted based on the importance of each modality. In particular, w_i is tunable during the learning process for respective downstream tasks.

DATA AVAILABILITY

The pretraining data can be downloaded from NMRShiftDB2. The MoleculeNet dataset is available at MoleculeNet. The DuD-E dataset can be accessed at DuD-E, and the Lit-PCBA dataset can be downloaded from Lit-PCBA.

CODE AVAILABILITY

The code is available in Github: <https://github.com/zhengyjo/MMRFL>

REFERENCES

RDKit: Open-source cheminformatics. <http://www.rdkit.org>.

Maria-Florina Balcan and Avrim Blum. On a theory of learning with similarity functions. In *Proceedings of the 23rd international conference on Machine learning*, pp. 73–80, 2006.

Mirko Bunzel and John Ralph. Nmr characterization of lignins isolated from fruit and vegetable insoluble dietary fiber. *Journal of agricultural and food chemistry*, 54(21):8352–8361, 2006.

H. Cai, H. Zhang, D. Zhao, J. Wu, and L. Wang. Fp-gnn: A versatile deep learning architecture for enhanced molecular property prediction. *Briefings in Bioinformatics*, 23(6):bbac408, November 2022. doi: 10.1093/bib/bbac408.

Yaojia Chen, Jiacheng Wang, Quan Zou, Mengting Niu, Yijie Ding, Jiangning Song, and Yansu Wang. Drugdagt: a dual-attention graph transformer with contrastive learning improves drug-drug interaction prediction. *BMC biology*, 22(1):233, 2024.

Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2mol—accurate smiles recognition from molecular graphical depictions. *Chemical science*, 12(42):14174–14181, 2021.

Filippo Costanti, Arian Kola, Franco Scarselli, Daniela Valensin, and Monica Bianchini. A deep learning approach to analyze nmr spectra of sh-sy5y cells for alzheimer’s disease diagnosis. *Mathematics*, 11(12):2664, 2023.

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.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, pp. 1–12, 2023.

Bowen Gao et al. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614, 2023.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Cosp: Co-supervised pretraining of pocket and ligand, 2022. arXiv preprint arXiv:2210.01776.

Ioannis P Gerothanassis, Anastassios Troganis, Vassiliki Exarchou, and Klimentini Barbarossou. Nuclear magnetic resonance (nmr) spectroscopy: basic principles and phenomena, and their applications to chemistry, biology and medicine. *Chemistry Education Research and Practice*, 3(2):229–252, 2002.

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*, pp. 1263–1272. PMLR, 2017.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 2023.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pp. 84–92. Springer, 2015.

-
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures, 2020.
- Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.
- Dana Lahat, Tülay Adalı, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
- Joseph B Lambert, Eugene P Mazzola, and Clark D Ridge. *Nuclear magnetic resonance spectroscopy: an introduction to principles, applications, and experimental methods*. John Wiley & Sons, 2019.
- Greg Landrum. Rdkit: Open-source cheminformatics. 2006. *Google Scholar*, 2006.
- Marc T Law, Nicolas Thome, and Matthieu Cord. Quadruplet-wise image similarity learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 249–256, 2013.
- Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the Thirty-Six AAAI Conference on Artificial Intelligence*, pp. 4541–4549, 2022.
- Hui Liu, Yibiao Huang, Xuejun Liu, and Lei Deng. Attention-wise masked graph contrastive learning for predicting molecular property. *Briefings in bioinformatics*, 23(5):bbac303, 2022a.
- Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.
- 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*, 2022b. URL <https://openreview.net/forum?id=xQUelpOKPam>.
- Pierre Mahé, Nobuhisa Ueda, Tatsuya Akutsu, Jean-Luc Perret, and Jean-Philippe Vert. Extensions of marginalized graph kernels. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 70, 2004.
- Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.
- Kisung Moon, Hyeon-Jin Im, and Sunyoung Kwon. 3d graph contrastive learning for molecular property prediction. *Bioinformatics*, 39(6):btad371, 2023.
- M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, July 2012. doi: 10.1021/jm300687e.
- Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.
- M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors (Basel)*, 23(5):2381, Feb 2023.
- Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544, 2015.
- Martin Priessner, Richard Lewis, Jon Paul Janet, Isak Lemurell, Magnus Johansson, Jonathan Goodman, and Anna Tomberg. Enhancing molecular structure elucidation: Multimodaltransformer for both simulated and experimental spectra. 2024.
- Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

-
- Petra Schneider, W Patrick Walters, Alleyn T Plowright, Norman Sieroka, Jennifer Listgarten, Robert A Goodnow Jr, Jasmin Fisher, Johanna M Jansen, José S Duca, Thomas S Rush, et al. Rethinking drug design in the artificial intelligence era. *Nature reviews drug discovery*, 19(5):353–364, 2020.
- Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.
- Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnn for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.
- Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of chemical information and computer sciences*, 43(6):1733–1739, 2003.
- Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.
- Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. Mocl: data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3585–3594, 2021.
- Wen Torng and Russ B. Altman. Graph convolutional neural networks for predicting drug–target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019. doi: 10.1021/acs.jcim.9b00628. URL <https://doi.org/10.1021/acs.jcim.9b00628>.
- Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling*, 60(9):4263–4273, 2020.
- Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D. Burke. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022a. URL <https://openreview.net/forum?id=6sh3pIzKS->.
- Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pp. 2593–2601, 2017.
- Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5022–5030, 2019.
- Yifei Wang, Shiyang Chen, Guobin Chen, Ethan Shurberg, Hang Liu, and Pengyu Hong. Motif-based graph representation learning with application to chemical molecules. In *Informatics*, volume 10, pp. 8. MDPI, 2023.
- Yingheng Wang, Yaosen Min, Erzhao Shao, and Ji Wu. Molecular graph contrastive learning with parameterized explainable augmentations. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1558–1563. IEEE, 2021.
- Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b.
- Yandong Wen, Weiyang Liu, Yao Feng, Bhiksha Raj, Rita Singh, Adrian Weller, Michael J Black, and Bernhard Schölkopf. Pairwise similarity learning is simple. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5308–5318, 2023.
- Oliver Wieder, Stefan Kohlbacher, Mélaïne 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.
- Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018a.
- Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018b.

-
- Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.
- Hao Xu, Yifei Wang, Yunrui Li, and Pengyu Hong. Asymmetric contrastive multimodal learning for advancing chemical understanding. *arXiv preprint arXiv:2311.06456*, 2023a.
- Hao Xu, Zhengyang Zhou, and Pengyu Hong. Molecular identification and peak assignment: Leveraging multi-level multimodal alignment on nmr. *arXiv preprint arXiv:2311.13817*, 2023b.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.
- Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.
- Zhuo Yang, Jianfei Song, Minjian Yang, Lin Yao, Jiahua Zhang, Hui Shi, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Cross-modal retrieval between ¹³c nmr spectra and structures for compound identification using deep contrastive learning. *Analytical Chemistry*, 93(50):16947–16955, 2021.
- Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Elayaraja Kolanthai, Craig J. Neal, Sudipta Seal, and Ozlem Ozmen Garibay. Attentionsitedti: An interpretable graph-based model for drug–target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics*, 23(4), July 2022. doi: 10.1093/bib/bbac272. URL <https://doi.org/10.1093/bib/bbac272>.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.
- Li Zhang, Shitian Shen, Lingxiao Li, Han Wang, Xueying Li, and Jun Lang. Jointly multi-similarity loss for deep metric learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1469–1474. IEEE, 2021.
- Zehong Zhang, Lifan Chen, Feisheng Zhong, Dingyan Wang, Jiaxin Jiang, Sulin Zhang, Hualiang Jiang, Mingyue Zheng, and Xutong Li. Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology*, 73:102327, 2022.
- Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Rssl: Relational self-supervised learning with weak augmentation. *Advances in Neural Information Processing Systems*, 34:2543–2555, 2021.
- Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2:134–140, February 2020. doi: 10.1038/s42256-020-0152-y.
- Shun Zhong and Xiaohui Guan. Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants’ activities and properties. *Environmental Science & Technology*, 57(47):18193–18202, 2023. doi: 10.1021/acs.est.3c05050.
- Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*, 2023. URL <https://api.semanticscholar.org/CorpusID:259298651>.

Appendix

A MULTI-SIMILARITY & CONTRASTIVE LEARNING

A.1 MULTI-SIMILARITIES IN CONTRASTIVE LEARNING

Two distinct types of similarities, as illustrated in Appendix Figure A.1, can be identified: *self-similarity* (the pairwise similarity between two objects, typically defined through cosine similarity) and *relative similarity* (distinctions in self-similarity with other pairs) (Wang et al., 2019).

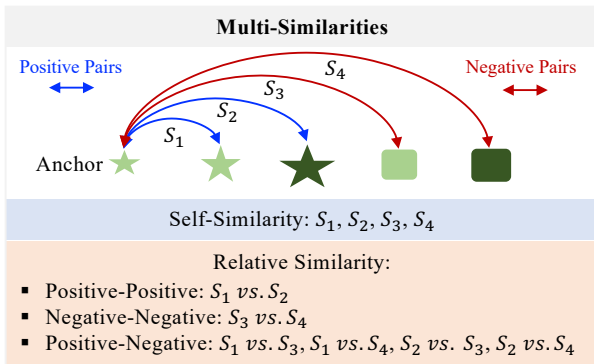


Figure A.1: Illustration of Different Types of Similarities.

A.2 CURRENT MOLECULAR GRAPH CONTRASTIVE LEARNING APPROACHES

In current molecular graph contrastive learning approaches, positive pairs are commonly formed through either *data augmentation* (Sun et al., 2021; You et al., 2020), employing techniques such as node deletion, edge perturbation, subgraph extraction, attribute masking, and subgraph substitution, or *domain knowledge*, as demonstrated by reactant-product pairing (Wang et al., 2022a) or conformer grouping (Moon et al., 2023).

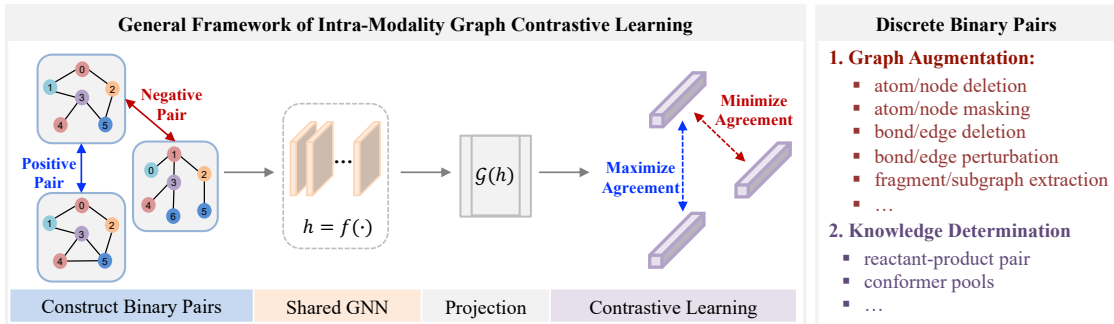


Figure A.2: General framework of Intra-Modality Graph Contrastive Learning. It relies on definition of positive and negative pairs.

B SUPPLEMENTARY PROOF

B.1 REVISITING THEOREM OF CONVERGENT SIMILARITY LEARNING

Let \mathcal{S} be a set of instances with size $|\mathcal{S}|$, and let \mathcal{P} represent the tunable latent representations of instances in \mathcal{S} such that $|\mathcal{P}| = |\mathcal{S}|$. For any two instances $i, j \in \mathcal{S}$, their latent representations are denoted by \mathcal{P}_i and \mathcal{P}_j , respectively. Let $t_{i,j}$ represent the target similarity between instances i and j in a given domain, and $d_{i,j}$ be the similarity between \mathcal{P}_i and \mathcal{P}_j in the latent space.

Theorem B.1 (Theorem of Convergent Similarity learning). *Given $t_{i,j}$ is non-negative and $\{t_{i,j}\}$ satisfies the constraint $\sum_{j=1}^{|S|} t_{i,j} = 1$, consider the loss function for an instance i defined as follows:*

$$L(i) = - \sum_{j=1}^{|S|} t_{i,j} \log \left(\frac{e^{d_{i,j}}}{\sum_{k=1}^{|S|} e^{d_{i,k}}} \right) \quad (\text{B.1})$$

then when it reaches ideal optimum, the relationship between $t_{i,j}$ and $d_{i,j}$ satisfies:

$$\text{softmax}(d_{i,j}) = t_{i,j} \quad (\text{B.2})$$

Proof. In order to optimize the loss $L(i)$, we need to set the following partial derivative to be 0 for each $d_{i,j}$ with $1 \leq j \leq |\mathcal{M}|$. Here are the detailed steps:

$$\begin{aligned} \frac{\partial L(i)}{\partial d_{i,j}} &= \frac{\partial}{\partial d_{i,j}} \underbrace{\left(-t_{i,j} \log \frac{e^{d_{i,j}}}{e^{d_{i,j}} + \sum_{k \neq j} e^{d_{i,k}}} \right)}_{\text{When the numerator includes } e^{d_{i,j}}} + \frac{\partial}{\partial d_{i,j}} \underbrace{\left(\sum_{k \neq j} -t_{i,k} \log \frac{e^{d_{i,k}}}{e^{d_{i,j}} + \sum_{k \neq j} e^{d_{i,k}}} \right)}_{\text{When the numerator does not include } e^{d_{i,j}}} \\ &= -(t_{i,j} - t_{i,j} \cdot \text{softmax}(d_{i,j})) - \sum_{k \neq j} t_{i,k} \cdot \text{softmax}(d_{i,j}) \\ &= - \left(t_{i,j} - \left(t_{i,j} + \sum_{k \neq j} t_{i,k} \right) \cdot \text{softmax}(d_{i,j}) \right) \end{aligned}$$

Since $\sum_{l=1}^{|\mathcal{M}|} t_{i,l} = 1$, we can further simplify it as

$$\frac{\partial L(i)}{\partial d_{i,j}} = -(t_{i,j} - \text{softmax}(d_{i,j}))$$

In order to optimize, we need to see the above partial derivative to be 0:

$$\frac{\partial L(i)}{\partial d_{i,j}} = -(t_{i,j} - \text{softmax}(d_{i,j})) = 0$$

In addition, the corresponding second partial derivative denoted as $\frac{\partial L(i)}{\partial d_{i,j}^2}$ manifests as follows:

$$\frac{\partial L(i)}{\partial d_{i,j}^2} = \text{softmax}(d_{i,j})(1 - \text{softmax}(d_{i,j}))$$

As $\text{softmax}(d_{i,j})$ takes values within the open interval (0,1), it follows that $\frac{\partial L(i)}{\partial d_{i,j}^2}$ is always positive. Consequently, the global optimum is global minimum.

Furthermore, when it comes to optimum:

$$\begin{aligned} t_{i,j} &= \text{softmax}(d_{i,j}) \\ d_{i,j} &= \log(t_{i,j}) + \log \left(\sum_{1 \leq l \leq |\mathcal{M}|} e^{d_{i,l}} \right) \end{aligned}$$

It is easy to show that when it reaches optimum, $d_{i,j}$ is consistent with target similarity metric $t_{i,j}$. Without loss of generality, suppose $t_{i,j} > t_{i,j'}$:

$$\begin{aligned} d_{i,j} - d_{i,j'} &= \log(t_{i,j}) + \log \left(\sum_{1 \leq l \leq |\mathcal{M}|} e^{d_{i,l}} \right) - \left(\log(t_{i,j'}) + \log \left(\sum_{1 \leq l \leq |\mathcal{M}|} e^{d_{i,l}} \right) \right) \\ &= \log(t_{i,j}) - \log(t_{i,j'}) \\ &= \log \left(\frac{t_{i,j}}{t_{i,j'}} \right) > 0 \end{aligned}$$

□

B.2 GUARANTEE OF SUM OF FUSED MULTIMODAL SIMILARITY

Given sets of uni-modal generalized similarity $\{t^R\}$ and $\sum w_{t^R} = 1$, the sum of fused multimodal similarity also equals 1, as demonstrated below:

$$\begin{aligned}\sum (t_{i,j}^R) &= \sum \sum (w_R \cdot t_{i,j}^R) \\ &= \sum (w_R \sum t_{i,j}^R) \\ &= \sum w_R \cdot 1 = 1\end{aligned}$$

C REVISITING TARGET SIMILARITY SETTINGS

C.1 ENCODERS & PACKAGES

To derive the target similarities, we need to rely on pre-trained encoders or well-defined packages as follows in Table C.1:

Table C.1: Encoders and packages used to produce self-similarities

Unimodal	Representation	Encoder/Package	Pre-trained Source
Image	2D image	CNN	Img2mol (Clevert et al., 2021)
SMILES	Sequence	Transformer	CRess (Yang et al., 2021)
¹³ CNMR Spectrum	Sequence	1D CNN	AutoEncoder (Costanti et al., 2023)
¹³ CNMR peak	Scalar	NMRShiftDB2 (Steinbeck et al., 2003)	N/A
Fingerprint	Sequence	RDKit (Landrum, 2006)	N/A

We selected GIN Xu et al. (2018) as the graph encoder for Smile, Image, Fingerprint, NMR, and NMR-Peak, respectively. All modalities share a consistent structure, each with 5 layers, an embedding dimension of 128, and a projection dimension of 512. In addition, during pretraining, contrastive learning is performed on each modality independently. For instance, if molecule A only possesses SMILE, IMAGE, and Fingerprint data, but lacks NMR information, it will be included in the training for contrastive learning on SMILE, IMAGE, and Fingerprint, but not on NMR. In contrast, for early fusion, all modalities must be present for the included molecules.

C.2 TARGET SIMILARITY AT GRAPH LEVEL

Fingerprint. The mathematical formula of fingerprint similarity, denoted as $S_{i,j}^F$, can be viewed as follows:

$$S_{i,j}^F = Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (C.1)$$

where A and B are sets of molecular fragments for molecule i and j , and $|A \cap B|$ and $|A \cup B|$ denote the size of their intersection and union, respectively.

Image. The self-similarity for Image, denoted as $S_{i,j}^I$, can be defined as follows:

$$S_{i,j}^I = Cos(\mathcal{V}_i, \mathcal{V}_j) = \frac{\mathcal{V}_i \cdot \mathcal{V}_j^T}{\|\mathcal{V}_i\| \cdot \|\mathcal{V}_j\|} \quad (C.2)$$

where $\mathcal{V}_i, \mathcal{V}_j$ represents the embedding of Image for two given molecules.

NMR Spectrum. The self-similarity for NMR spectrum, denoted as $S_{i,j}^C$, can be defined as follows:

$$S_{i,j}^C = Cos(\mathcal{V}_i, \mathcal{V}_j) = \frac{\mathcal{V}_i \cdot \mathcal{V}_j^T}{\|\mathcal{V}_i\| \cdot \|\mathcal{V}_j\|} \quad (C.3)$$

where $\mathcal{V}_i, \mathcal{V}_j$ represents the embedding of NMR spectra for two given molecules.

Smiles. The self-similarity for Smiles, denoted as $S_{i,j}^S$, can be defined as follows:

$$S_{i,j}^S = Cos(\mathcal{V}_i, \mathcal{V}_j) = \frac{\mathcal{V}_i \cdot \mathcal{V}_j^T}{\|\mathcal{V}_i\| \cdot \|\mathcal{V}_j\|} \quad (C.4)$$

where $\mathcal{V}_i, \mathcal{V}_j$ represents the embedding of Smiles for two given molecules.

NMR Peak The similarity among nodes (atoms) is derived from the positions of their signal peaks on ^{13}C NMR spectra, measured in parts per million (ppm). The ppm values are continuous, typically ranging from 0 to 200 (see more introduction of ppm in Appendix C.3). The self-similarity of NMR peaks $S_{l,m}^P$ can be defined as following:

$$S_{l,m}^P = \frac{\tau_2}{|ppm_l - ppm_m| + \tau_1} \quad (\text{C.5})$$

where ppm_l and ppm_m are the positions of NMR peaks for the l^{th} , m^{th} Carbon atom, τ_1 and τ_2 are temperature hyper-parameter. Additionally, we conducted an ablation study to examine the impact of different temperature combinations (τ_1 and τ_2) on peak self-similarity and its effect on model performance for *Atom Alignment with Peak Accuracy*. Our findings indicate that the following combinations do not significantly affect performance, as shown in Table C.2. For this analysis, we fixed the GIN depth at 5, set the GIN embedding dimensionality to 128, and kept the projection dimension at 512. The results suggest that the best performance occurs when $\tau_1 = 10^{-5}$ and $\tau_2 = 10^1$.

Table C.2: Ablation study of Effect of different τ_1 and τ_2 combinations on accuracy.

τ_1	τ_2	Accuracy (%)
10^{-1}	10^1	89.6
10^{-1}	10^2	89.8
10^{-1}	10^3	89.6
10^{-1}	10^4	88.9
10^{-1}	10^5	89.3
10^{-2}	10^1	89.8
10^{-2}	10^2	89.8
10^{-2}	10^3	88.8
10^{-2}	10^4	87.2
10^{-2}	10^5	89.4
10^{-3}	10^1	89.2
10^{-3}	10^2	89.1
10^{-3}	10^3	89.0
10^{-3}	10^4	89.7
10^{-3}	10^5	89.4
10^{-4}	10^1	89.8
10^{-4}	10^2	89.7
10^{-4}	10^3	89.8
10^{-4}	10^4	89.5
10^{-4}	10^5	88.4
10^{-5}	10^1	90.0
10^{-5}	10^2	89.5
10^{-5}	10^3	89.6
10^{-5}	10^4	89.7
10^{-5}	10^5	89.7

C.3 A BRIEF INTRODUCTION TO PPM FOR NMR PEAK

In chemistry, ^{13}C NMR stands out as a common technique for structural analysis by revealing molecular structures by elucidating the chemical environments of carbon atoms and their magnetic responses to external fields (Gerothanassis et al., 2002; Lambert et al., 2019). It quantifies these features in parts per million (ppm) relative to a reference compound, such as tetramethylsilane (TMS), thereby simplifying comparisons across experiments. As a result, the continuous peak positions, measured in parts per million (ppm), offer a robust knowledge span—a natural ordering metric that can be employed to derive measures of similarity (Xu et al., 2023b).

C.4 CONFIGURATION OF EARLY FUSION

A simple linear combination is used to formulate the multimodal relational similarity $t_{i,j}^M$ between the i^{th} and j^{th} molecules, represented as follows:

$$t_{i,j}^M = w_{SM} \cdot t_{i,j}^{SM} + w_C \cdot t_{i,j}^C + w_I \cdot t_{i,j}^I + w_F \cdot t_{i,j}^F + w_F \cdot t_{i,j}^F + w_P \cdot t_{i,j}^P \quad (C.6)$$

where $t_{i,j}^{SM}$ denotes the similarity based on SMILES, $t_{i,j}^C$ denotes the similarity with respect to ^{13}C NMR spectrum, $t_{i,j}^I$ denotes the similarity regarding images, F denotes the similarity based on fingerprints, and P denotes the similarity based on fingerprints. w_{SM} , w_C , w_I , and w_F are the pre-defined weights for their respective similarity, and $w_{SM} + w_C + w_I + w_F + w_P = 1$. The ablation study about how the weight combinations influence the performance of early fusion is shown in Table C.3

Table C.3: Performance of different modality weight combinations across datasets in early fusion.

Modality Weight					Dataset											
Smiles	NMR	Image	FP	Peak	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	TOX21	TOXCAST	ESOL	FREESOLV	LIPO	
1	0	0	0	0	92.9±1.5	90.9±3.3	64.9±0.3	78.2±1.9	83.3±1.1	80.1±2.5	85.7±1.2	70.5±2.5	0.811±0.109	1.623±0.168	0.539±0.017	
0	1	0	0	0	91.0±2.0	93.2±2.7	68.1±1.5	87.7±6.5	80.9±5.0	<u>80.9±5.0</u>	85.1±0.4	71.1±0.8	0.844±0.123	2.417±0.495	0.609±0.031	
0	0	1	0	0	93.1±2.4	92.9±1.8	65.3±1.5	86.2±6.5	82.3±0.6	78.7±1.7	86.0±1.0	71.0±1.6	0.761±0.068	1.648±0.045	0.537±0.005	
0	0	0	1	0	92.9±2.3	91.7±3.6	65.6±0.7	87.5±6.0	81.2±2.5	82.9±3.1	85.3±1.3	70.0±1.4	0.808±0.071	1.437±0.134	0.565±0.017	
0	0	0	0	1	93.4±2.7	89.3±1.7	62.8±2.1	86.1±5.4	82.1±0.4	75.4±5.2	84.9±1.0	70.6±0.8	0.924±0.083	1.707±0.126	0.587±0.021	
0.2	0.2	0.2	0.2	0.2	91.6±5.0	94.3±2.4	66.4±1.9	85.3±6.8	82.0±4.2	80.6±3.2	85.2±0.6	69.8±1.1	1.037±0.090	2.093±0.090	0.607±0.034	

D EXPERIMENTAL SETTINGS

D.1 PRE-TRAINING SETTING

During pretraining, we utilized an Adam optimizer with a learning rate set to 0.001, spanning 200 epochs and employing a batch size of 256. The model was trained on around 25,000 data points. The NMR data were experimental data, extracted from NMRShiftDB2 (Steinbeck et al., 2003). Other chemical modalities, such as images, fingerprints and graphs, were produced from SMILES by RDKit Landrum (2006).

D.2 FINE-TUNING SETTING

D.2.1 DATASETS

For fine-tuning, our model was trained on 11 drug discovery-related benchmarks sourced from MoleculeNet (Wu et al., 2018a). Eight of these benchmarks were designated for classification downstream tasks, including BBBP, BACE, SIDER, CLINTOX, HIV, MUV, TOX21, and ToxCast, while three were allocated for regression tasks, namely ESOL, Freesolv, and Lipo. The datasets were divided into train/validation/test sets using a ratio of 80%:10%:10%, accomplished through the scaffold splitter (Halgren, 1996; Landrum, 2006) from Chemprop (Yang et al., 2019; Heid et al., 2023), like previous works. The scaffold splitter categorizes molecular data based on substructures, ensuring diverse structures in each set. Molecules are partitioned into bins, with those exceeding half of the test set size assigned to training, promoting scaffold diversity in validation and test sets. Remaining bins are randomly allocated until reaching the desired set sizes, creating multiple scaffold splits for comprehensive evaluation.

D.2.2 BASELINES

We systematically compared MMFRL’s performance with various state-of-the-art baseline models across different categories. In the realm of supervised models, AttentiveFP (Xiong et al., 2019) and DMPNN (Yang et al., 2019) stand out by leveraging graph attention networks and node-edge interactive message passing, respectively. The unsupervised learning method N-Gram (Liu et al., 2019) employs graph embeddings and short walks for graph representation. Predictive self-supervised learning methods, such as GEM (Fang et al., 2022) and Uni-Mol (Zhou et al., 2023), are specifically designed for predicting molecular geometric information. Moreover, our evaluation encompasses a range of contrastive learning methods, namely InfoGraph (Sun et al., 2019), GraphCL (You et al., 2020), MolCLR (Wang et al., 2022b), and GraphMVP (Liu et al., 2022b), all serving as essential baselines. The baseline results are collected from recent works (Fang et al., 2022; Zhou et al., 2023; Moon et al., 2023; Fang et al., 2023).

D.2.3 EVALUATION

To assess the effectiveness of our fine-tuned model, we measure the ROC-AUC for classification downstream tasks, and the root mean squared error (RMSE) metric for regression tasks. In order to ensure a fair and robust comparisons, we conduct three independent runs using three different random seeds for scaffold splitting across all datasets. The reported performance metrics are then averaged across these runs, and the standard deviation is computed as prior works. In particular, the random selected seeds for respective experiments are drawn from the range between 0 and 20.