Graph Multi-Similarity Learning for Molecular Property Prediction

Anonymous Authors[1]

Abstract

Enhancing accurate molecular property prediction relies on effective and proficient representation learning. It is crucial to incorporate diverse molecular relationships characterized by multi-similarity (self-similarity and relative similarities) (Wang et al., 2019) between molecules. However, current molecular representation learning methods fall short in exploring multi-similarity and often underestimate the complexity of relationships between molecules. Additionally, previous multi-similarity approaches require the specification of positive and negative pairs to attribute distinct predefined weights to different relative similarities, which can introduce potential bias. In this work, we introduce Graph Multi-Similarity Learning for Molecular Property Prediction (GraphMSL) framework, along with a novel approach to formulate a generalized multi-similarity metric without the need to define positive and negative pairs. In each of the chemical modality spaces (e.g., molecular depiction image, fingerprint, NMR, and SMILES) under consideration, we first define a self-similarity metric (i.e., similarity between an anchor molecule and another molecule), and then transform it into a generalized multi-similarity metric for the anchor through a pair weighting function. GraphMSL validates the efficacy of the multi-similarity metric across MoleculeNet datasets. Furthermore, these metrics of all modalities are integrated into a multimodal multi-similarity metric, which showcases the potential to improve the performance. Moreover, the focus of the model can be redirected or customized by altering the fusion function. Last but not least, GraphMSL proves effective in drug discovery evaluations through post-hoc analyses of the learnt representations.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. under review at ICML 2024 AI for Science workshop. Do not distribute.

1. Introduction

Graph Neural Networks (GNNs) have emerged as a prominent approach for molecular representation learning, addressing drug-related challenges (Wieder et al., 2020; Zhang et al., 2022; Fang et al., 2022; Wang et al., 2023). However, the generation of task-specific labels for training molecular GNNs is hindered by the resource-intensive and time-consuming nature of chemical synthesis and biological testing experiments. To address this challenge, current research prioritizes self-supervised learning approaches for pretraining molecular GNNs, with a prevalent trend of adopting contrastive learning approaches (Wang et al., 2021; 2022b; Liu et al., 2022a). (See an introduction of pre-training approaches in Section 2)

Contrastive Learning (CL) is a discriminative representation learning approach by bringing similar instances into close proximity within the latent representation space and pushing apart dissimilar instances (Schroff et al., 2015). A fundamental prerequisite of CL is to define positive pairs denoting similarity, and negative pairs representing dissimilarity (Jaiswal et al., 2020). In the CL-based pre-training of molecular GNNs, positive pairs are often established through either *data augmentation* (Sun et al., 2021; You et al., 2020a), such as node deletion, edge perturbation, subgraph extraction, attribute masking, and subgraph substitution, or *domain knowledge*, exemplified by reactant-product pairing (Wang et al., 2022a) or conformer grouping (Moon et al., 2023). However, such a binary characterization of the relationships among molecules, by designating them as either positive or negative pairs, oversimplifies the complex nature of these connections. Moreover, these CL approaches fail to notice relationships among multiple instances simultaneously by adapting simple contrastive loss, thereby hindering the effectiveness and generalizability of representation learning (Wang et al., 2019; Mu et al., 2023; Zhang et al., 2023). (See an illustration of similarity types in CL in Figure Appendix A.1)

Unlike CL, which adapts a binary similarity metric, Similarity Learning (SL) employs a continuous similarity metric for representation learning. It measures the similarity between two instances in the given space (Balcan & Blum, 2006; Wen et al., 2023). This pairwise similarity provides a localized perspective on the relations between two instances

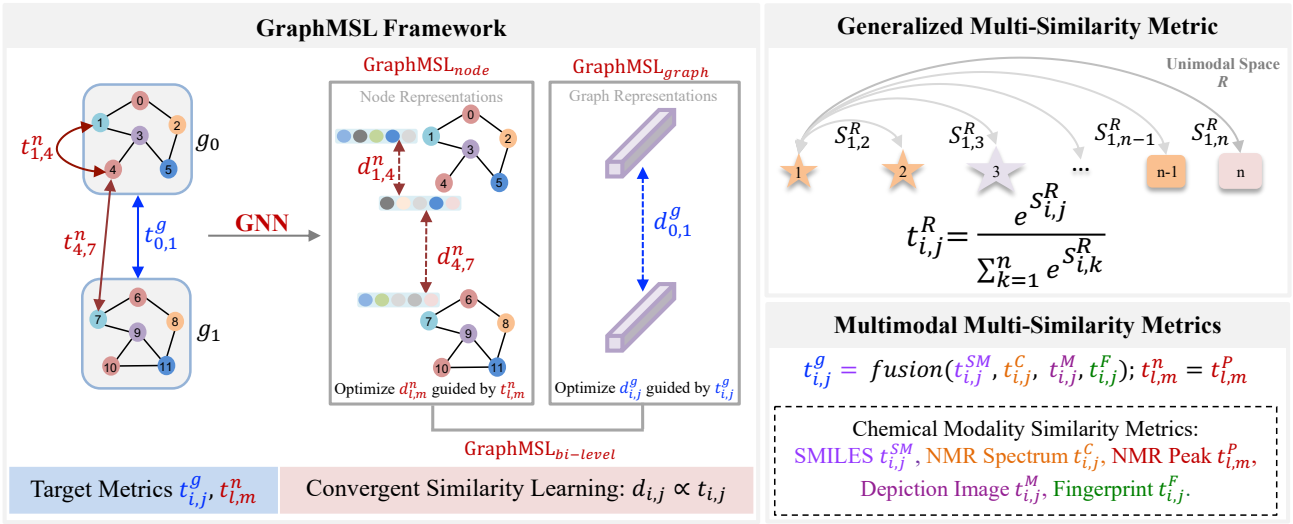

Figure 1. **Graph Similarity Learning for Molecular Property Prediction (GraphMSL).** $t^g_{i,j}$, $t^n_{l,m}$ represent graph-level and node-level target similarity, respectively. $d^g_{i,j}$, $d^n_{l,m}$ represent the similarity for graph-level and node-level embeddings, respectively. Here, i and j are the indices of molecular graphs in the graph pool, and l and m are the indices of atomic nodes in the node pool. Unlike the general contrastive learning framework shown in Appendix Figure A.2, GraphMSL doesn't need to define positive or negative pairs and is capable of learning continuous ordering from target similarity.

in a given instance pool, known as self-similarity (Wang et al., 2019). However, the global relations between two instances can be influenced by their belonging to the instance pool. Therefore, self-similarity alone is inadequate for capturing relationships among multiple instances simultaneously, known as relative similarity (Wang et al., 2019). To address this challenge, Multi-Similarity Learning (MSL) has emerged as a solution, expanding its focus from self-similarity to a global view of relations and encapsulating both self-similarity and relative similarity (Wang et al., 2019; Zhang et al., 2021; Mu et al., 2023; Zhang et al., 2023).

When formulating a similarity metric, a common guiding principle is that an effective similarity function or metric should align tightly with the objectives of specific tasks (Balcan & Blum, 2006; Müller et al., 2018; Yang & Jin, 2006). To target the drug discovery tasks, a robust similarity metric must be adept at discerning molecules based on key properties regarding drug development. Insights drawn from prior research (Xu et al., 2023a) indicate that distinct chemical modalities, such as chemical languages, molecular depiction images, and chemical spectroscopic spectra, possess unique expertise in expressing specific molecular properties. More importantly, a continuous similarity metric for molecules can be projected from each chemical modality space. Thus, a more promising multimodal similarity metric can be formulated by integrating these individual metrics.

In response to the challenges and opportunities in molecular graph representation learning, we propose the Graph Multi-Similarity Learning for Molecular Property Prediction (GraphMSL) framework. This approach aims to advance graph contrastive learning to graph multi-similarity learning by incorporating a continuous multi-similarity metric. A self-similarity metric can be derived from each heterogeneous chemical modality, such as chemical languages, molecular depiction images, and chemical spectroscopic spectra, through representation learning. Then, this self-similarity metric can be transformed into a multi-similarity metric through a pair weighting function. Besides, various unimodal multi-similarity metrics can be fused into a multimodal similarity metric. In addition, GraphMSL framework can be customized for singular or multi-view perspective.

In summary, our contribution comprises three major aspects: ***Conceptually:*** We introduce a generalized multi-similarity metric for graph representation learning, capturing both self-similarity and relative similarity. Our approach doesn't rely on pre-defined negative or positive pairs, and it satisfies the requirement of convergent similarity learning as shown in Section 3.1. To the best of our knowledge, this is the first work to demonstrate such generalized multi-similarity for graph representation learning. ***Methodologically:*** We extract a self-similarity metric from a chemical modality, and transition it into a generalized multi-similarity metric through a pair weighting function, and each modality contributes to a unique multi-similarity metric. Furthermore, we integrate these metrics into a fused multimodal form that has the potential to improve the performance. ***Empirically:*** GraphMSL excels in various downstream tasks, with performance enhancements achieved through multi-level graph

learning. 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. Preliminaries

Directed Message Passing Neural Network (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 $v_1v_2 \dots v_n$, where $v_i = v_{i+2}$ for some i, thereby eliminating unnecessary loops in the message passing trajectory.

Pre-Training for Molecular GNNs. There are two levels of pre-training tasks: node-level (atomic level) and graph-level (molecular level), which enhance the generalization capabilities of molecular GNNs across diverse downstream tasks (Hu et al., 2019; Xia et al., 2022; Fang et al., 2022). Node-level tasks aim to capture local context, often involving the random masking of nodes and subsequent prediction of their properties based on node representations. In contrast, graph-level tasks focus on extracting global information, such as predicting graph properties based on the graph representation.Self-supervised learning (SSL) is a paradigm in which a model is trained on a task by leveraging the data itself to generate supervisory signals, eliminating the need for external annotations (Liu et al., 2021b; 2022c). SSL finds active applications in the pre-training tasks of molecular GNNs by formulating label-free pretext tasks, such as graph reconstructions (Hu et al., 2020; You et al., 2020b; Liu et al., 2021a), context predictions (Hu et al., 2019; Peng et al., 2020), and adopting contrastive learning approaches (Wang et al., 2021; 2022b; Liu et al., 2022a).

Similarity Learning. Given object i and object j, the optimization of object similarity $d_{i,j}$ in the latent representation space is directed by the target similarity $t_{i,j}$ in a given space, a process commonly recognized as similarity learning (Moutafis et al., 2016; Suárez-Díaz et al., 2018). Contrastive learning constrains its similarity metric $t_{i,j}$ to a binary setting, taking on values of either 1 or 0. However, similarity learning allows its similarity metric $t_{i,j}$ to be continuous values. Two distinct types of similarities can be identified, as illustrated in Appendix Figure A.1: *self-similarity* (the pairwise similarity between two objects, typically defined through cosine similarity), *relative similarity* (distinctions in self-similarity with other pairs) (Wang et al., 2019). Contrastive learning, implemented with a binary similarity metric, concentrates on self-similarity and fails to explore the complete relationships between samples (Oh Song et al., 2016; Wang et al., 2019).

3. Methods

We begin by presenting the theorem of convergent similarity learning, followed by introducing the generalized multi-similarity metric and the multimodal multi-similarity metric.

3.1. Convergent Similarity learning

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.

Theorem 3.1 (Theorem of Convergent Similarity learning)**.** *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:*

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

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

$$softmax(d_{i,j}) = t_{i,j} \tag{2}$$

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

3.2. Generalized Multi-Similarity

We formulate a generalized multi-similarity metric from self-similarity by adapting the softmax function as a pair weighting function. The formula for the generalized multi-similarity, denoted as $t^R_{i,j}$ between the i^{th} and j^{th} instances under a given space R, is provided below:

$$t^R_{i,j} = softmax(S^R_{i,j}) = \frac{e^{S^R_{i,j}}}{\sum_{k=1}^{|\mathcal{S}|} e^{S^R_{i,k}}} \tag{3}$$

where $S^R_{i,j}$ represents self-similarity, and $|\mathcal{S}|$ is the size of the instance set. Defined in this manner, the generalized

multi-similarity metric incorporates both self-similarity and relative similarities. Notably, unlike other multi-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, the use of the softmax function ensures that the generalized target similarity $t_{i,j}$ adheres to the principles of Convergent Similarity Learning (refer to Section 3.1).

3.3. Multimodal Multi-Similarity

With a set of generalized multi-similarities $\{t^R\}$ from various modality spaces, we can transform generalized multi-similarities from respective unimodality space 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 $t^M_{i,j}$ between i^{th} and j^{th} objects can be defined as follows:

$$t^M_{i,j} = fusion(\{t^R\}) \tag{4}$$

$$= \sum w_R \cdot t^R_{i,j} \tag{5}$$

where $t^R_{i,j}$ 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 $\sum w_R = 1$. Such that, it still satisfy the requirement of Convergent Similarity Learning (See proof in Appendix SectionB.2).

4. Experiments

In this section, we begin by presenting the design of the multi-similarity metric and the GraphMSL framework. Subsequently, we showcase the results obtained from GraphMSL. Finally, we demonstrate the explainability of the learned molecular representations. (Please refer to the experimental details of pre-training and fine-tuning in the Appendix Section D.)

4.1. The Design of Multi-Similarity Metric

4.1.1. Graph-Level Multi-Similarity Metric

Self-Similarity. By employing representation learning, various chemical modalities, such as NMR spectra, depiction images, and SMILES, can be encoded into latent representation vectors. The cosine similarity between two vectors serves as self-similarity. The unimodal self-similarity for ^{13}C NMR spectrum, denoted as $S^C_{i,j}$, can be defined as follows:

$$S^C_{i,j} = Cos(\mathcal{V}_i, \mathcal{V}_j) = \frac{\mathcal{V}_i \cdot \mathcal{V}_j^T}{\|\mathcal{V}_i\| \cdot \|\mathcal{V}_j\|} \tag{6}$$

where $\mathcal{V}_i, \mathcal{V}_j$ represents the embedding of NMR spectra for two given molecules. Similarly, the uni-modality similarity similarity for depiction images and SMILES can be obtained. The self-similarity of fingerprints adapts a well-established similarity function for molecules, namely the Tanimoto similarity (Bajusz et al., 2015). (See more details in Appendix Section C.2)

Generalized Multi-Similarity Metric. Transitioning from self-similarity to a generalized similarity, we apply the softmax function as our pair weighting mechanism. The formula of generalized similarity is provided below, illustrated with an example of ^{13}C NMR similarity:

$$t^C_{i,j} = softmax(S^C_{i,j})_i = \frac{e^{S^C_{i,j}}}{\sum_{k=1}^{|\mathcal{G}|} e^{s^C_{i,k}}} \tag{7}$$

where $t^C_{i,j}$ represents generalized similarity, $S^C_{i,j}$ denotes the self-similarity, and $|\mathcal{G}|$ is the size of molecular graph pool. For the expressions for SMILES and Image modality, please refer to Appendix Section C.2

Multimodal Multi-Similarity Metric. We employ a simple linear combination to formulate the multimodal multi-similarity $t^M_{i,j}$ between the i^{th} and j^{th} molecules, represented as a graph-level similarity $t^g_{i,j}$, as follows:

$$t^g_{i,j} = t^M_{i,j} \tag{8}$$

$$= w_{SM} \cdot t^{SM}_{i,j} + w_C \cdot t^C_{i,j} + w_I \cdot t^I_{i,j} + w_F \cdot t^F_{i,j} \tag{9}$$

where $t^{SM}_{i,j}$ denotes the similarity based on SMILES, $t^C_{i,j}$ denotes the similarity with respect to ^{13}C NMR spectrum, $t^I_{i,j}$ denotes the similarity regarding images, and f 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 $w_{SM} + w_C + w_I + w_F = 1$.

4.1.2. Node-Level Multi-Similarity Metric

Self-Similarity. The self-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 $S^P_{l,m}$ can be defined as following:

$$S^P_{l,m} = \frac{\tau_2}{|ppm_l - ppm_m| + \tau_1} \tag{10}$$

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.

Generalized Similarity Metric. A generalized multi-similarity $t^P_{l,m}$, as a node-level similarity $t^n_{l,m}$, can be formulated with softmax function, as shown below

$$t^n_{l,m} = t^P_{l,m} = softmax(S^P_{l,m})_l = \frac{e^{S^P_{l,m}}}{\sum_{q=1}^{|\mathcal{N}|} e^{S^P_{l,q}}} \tag{11}$$

where $S^P_{l,m}$ represents self-similarity of NMR peaks, $|\mathcal{N}|$ is the size of atomic node pool.

4.2. The Design of GraphMSL Framework

The GraphMSL framework is versatile, allowing customization for singular or multiple views. In this context, we showcase the customization options for graph-level, node-level, or bi-level GraphMSL. The graph-level GraphMSL, denoted as GraphMSL$_{graph}$, incorporates multi-similarity metrics derived from molecular level chemical semantics. The node-level GraphMSL, denoted as GraphMSL$_{node}$, incorporates multi-similarity metrics derived from atomic level chemical semantics. The bi-level GraphMSL, denoted as GraphMSL$_{bi-level}$, incorporates multi-similarity metrics derived from both molecular and atomic level chemical semantics. In these implementations, the graph encoder adopts DMPNN (Yang et al., 2019) architecture, an interactive message passing scheme considering the interactions. Additionally, neither of these two modules requires additional projection layers.

The loss function of GraphMSL$_{graph}$ model, noted as L_{graph}, can be expressed as:

$$L_{graph} = -\frac{1}{|\mathcal{G}|} \sum_{1 \le j \le |\mathcal{G}|} t^g_{i,j} \log \frac{e^{d^g_{i,j}}}{\sum_{1 \le k \le |\mathcal{G}|} e^{d^g_{i,k}}} \quad (12)$$

where $t^g_{i,j}$ represents graph-level similarity metrics, $d^g_{i,j}$ represents graph-level latent space similarity metrics, i, j, k represent the indices of molecular graphs within a graph pool of size $|\mathcal{G}|$ per batch.

The loss function of GraphMSL$_{node}$ model, noted as L_{node}, can be expressed as:

$$L_{node} = -\frac{1}{|\mathcal{N}|} \sum_{1 \le m \le |\mathcal{N}|} t^n_{l,m} \log \frac{e^{d^n_{l,m}}}{\sum_{1 \le q \le |\mathcal{N}|} e^{d^n_{lq}}} \quad (13)$$

where $t^n_{l,m}$ represents node-level similarity metrics, $d^g_{l,m}$ represents node-level latent space similarity metrics, l, m, q represent the indices of nodes within a node pool of size $|\mathcal{N}|$.

The loss function of GraphMSL$_{bi-level}$ model, noted as $L_{bi-level}$, can be expressed as:

$$L_{bi-level} = L_{graph} + L_{node} \quad (14)$$

4.3. Results

4.3.1. Overall performance of GraphMSL$_{graph}$

The performance of GraphMSL$_{graph}$ is evaluated through a comparative analysis with a range of baselines, the specifics of which are described in Appendix Section D.2.2. We report the performance metrics of GraphMSL$_{graph}$ across 8 classification under ROC-AUC and 3 regression tasks under RMSE from the MoleculeNet benchmark (Wu et al., 2018a), as shown in Tables 1. Within these tables, the best results are denoted in bold, and the second-best are indicated with underlining. From the comparative evaluation, we find that:

1) GraphMSL$_{graph}$ outperforms the baselines in seven of the eight evaluated classification tasks, including BBBP, BACE, SIDER, HIV, MUV, Tox21 and ToxCast.

2) In the regression tasks, GraphMSL$_{graph}$ also achieves the highest performance across all evaluated benchmarks, which include ESOL, FreeSolv, and Lipophilicity.

3) While GraphMSL$_{graph}$ doesn't outperform baselines on the Clintox task, it attains average performance.

In short, these findings underscore the proficiency of GraphMSL$_{graph}$ in learning molecular representations that are effective and impactful, as evidenced by its commendable performance across a suite of diverse tasks.

4.3.2. Ablation Study-Various Mutil,modal Similarity Metrics

We evaluate the performance of GraphMSL$_{graph}$ using a diverse set of similarity metrics, as outlined in Table 3. Each uni-modal similarity metric demonstrates unique strengths across various tasks. For instance, the model guided by Image similarity metrics exhibits outstanding performance in ESOL compared to other uni-modality metrics. Solubility is closely tied to the polarity of molecules. High polar atoms possess the ability to form hydrogen bonds with water, thereby enhancing solubility. In molecular depiction images, nonpolar carbon atoms (C) are typically represented as dots. Conversely, highly polar atoms like oxygen (O), nitrogen (N), and fluorine (F) are depicted more prominently, occupying a significant portion of the image. Consequently, image representations place considerable emphasis on this information, distinguishing between nonpolar and polar atoms by varying pixel density.

Notably, the true strength lies in the flexibility of multimodal similarity metrics achieved through the fusion of multiple unimodal metrics. There are five variations of multimodal similarity metrics, denoted as Fusion$_{\text{Smiles}}$, Fusion$_{\text{NMR}}$, Fusion$_{\text{Image}}$, Fusion$_{\text{Fingerprint}}$, and Fusion$_{\text{Average}}$ (Please refer to their configurations in Appendix Section C.4). Through the comparisons, it becomes evident that a well-designed multimodal similarity metric can significantly enhance model performance compared to unimodal metrics. For example, Fusion$_{\text{Smiles}}$ boosts performance in Tox21 tasks, Fusion$_{\text{NMR}}$ enhances results in MUV task, and Fusion$_{\text{Fingerprint}}$ contributes to improved outcomes in BBBP and BACE tasks.

4.3.3. Ablation Study-Mutil-View Learning

While the GraphMSL$_{graph}$ model demonstrates its efficacy across a variety of downstream tasks, we further investigate

Table 1. Overall performances (ROC-AUC) on classification downstream tasks. The best results are denoted in bold, and the second-best are indicated with underlining. (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	67.2±0.4	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	65.9±1.3	91.9±1.8	80.8±0.3	82.1±1.3	79.6±0.5	69.6±0.1
GROVER	86.8±2.2	82.4±3.6	61.2±2.5	70.3±13.7	68.2±1.1	67.3±1.8	80.3±2.0	56.8±3.4
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
$\text{MOLCLR}_{\text{CMPNN}}$	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
GRAPHMSL_{graph}	93.2±0.8	93.6±2.7	**68.1±1.5**	88.8±4.6	**83.3±1.1**	**84.5±2.9**	86.1±0.6	**71.4±0.2**
GRAPHMSL_{node}	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
$\text{GRAPHMSL}_{bi-level}$	**94.3±0.8**	**94.5±0.7**	67.3±0.6	**93.8±0.8**	83.0±0.7	81.5±3.7	**86.1±0.8**	71.2±1.1

Table 2. Overall performances (RMSE) on regression downstream tasks. The best results are denoted in bold, and the second-best are indicated with underlining.

Data Set	ESOL	FreeSolv	Lipo
AttentiveFP	0.877±0.029	2.073±0.183	0.721±0.001
DMPNN	1.050±0.008	2.082±0.082	0.683±0.016
$\text{N-Gram}_{\text{RF}}$	1.074±0.107	2.688±0.085	0.812±0.028
$\text{N-Gram}_{\text{XGB}}$	1.083±0.082	5.061±0.744	2.072±0.030
GEM	0.798±0.029	1.877±0.094	0.660±0.008
Uni-Mol	0.788±0.029	1.620±0.035	0.660±0.008
GROVER	1.423±0.288	2.977±0.615	0.823±0.010
MolCLR	1.113±0.023	2.301±0.247	0.789±0.009
$\text{MolCLR}_{\text{CMPNN}}$	0.911±0.082	2.021±0.133	0.875±0.003
GraphMSL_{graph}	**0.746±0.060**	**1.437± 0.134**	**0.537±0.005**
GraphMSL_{node}	0.924±0.083	1.707±0.126	0.587±0.021
$\text{GraphMSL}_{bi-level}$	0.843±0.094	1.601±0.057	0.562±0.005

the potential of GraphMSL framework by including both graph (molecule) and node (atom) levels to assess possible performance improvements. Our findings indicate that:

1) For Clintox, $\text{GraphMSL}_{bi-level}$ exhibits superior performance, compared with the baselines and GraphMSL_{graph}. Notably, $\text{GraphMSL}_{bi-level}$ achieves a 5% enhancement in ROC-AUC on the Clintox dataset over GraphMSL_{graph}.

2) For tasks such as BBBP, BACE, HIV, MUV, and TOX21, $\text{GraphMSL}_{bi-level}$ outperforms all compared models, though $\text{GraphMSL}_{bi-level}$ is better than GraphMSL_{graph} by a marginal degree.

3) Conversely, $\text{GraphMSL}_{bi-level}$ underperforms relative to GraphMSL_{graph} in Sider, ToxCast, ESOL, FreeSolv, and Lipo datasets. This discrepancy in performance may stem from small portion of unresolved interactions or slight discord between the graph-level and node-level similarities under the scenarios of these tasks.

4.4. Explainability of Learnt Representations

To demonstrate the interpretability of learnt representations, we present post-hoc analysis for two tasks, ESOL and BACE, as demonstration. The results showcase learnt representations can capture task-specific patterns and offer valuable insights for molecular design.

ESOL. We apply t-SNE to reduce molecule embeddings from 300 to 2, generating a heatmap correlating Log solubility (Figure 3.a). The heatmap visually depicts a smooth transition from high solubility (depicted in red) to low solubility (depicted in blue). The embeddings adeptly capture essential structural information and solubility-related patterns, organizing molecules with analogous solubility in close proximity within the embedding space. Through our investigation, we discovered that molecules with comparable solubility share common graph features, such as recurring motifs or an increased frequency of specific nodes (atoms). For example, in the region of lowest solubility (enclosed by blue dashed line), the molecular graphs all contain biphenyl groups. Biphenyls are nonpolar molecules, and as a consequence, they have limited interaction with water, leading to low solubility in aqueous environments. In the region of highest solubility (enclosed by red solid line), molecules exhibit high polarity, characterized by the prevalence of nitrogen (N) and oxygen (O) atoms. This specific graph configuration facilitates the formation of hydrogen bonds with water molecules, thereby enhancing solubility.

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 and subsequently compute the maximum common subgraph (MCS) for each group. In Figure 3.b, we present the top 10 most frequently occurring

Table 3. Ablation study on the performances of GraphMSL$_{graph}$. The best results are denoted in boldf, and the second-best are indicated with underlining. The first 8 tasks are for classification under evaluation of ROC-AUC, while the last three are for regression with evaluation of RMSE. See the detailed performance of GraphMSL$_{bi-level}$ upon different similarity metrics in Appendix Table E.1.

DATA SET	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	TOX21	TOXCAST	ESOL	FREESOLV	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	85.7±1.2	70.5±2.5	0.811± 0.109	1.623± 0.168	0.539± 0.017
NMR	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	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**
FINGERPRINT	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
FUSION$_{SMILES}$	93.1±1.4	91.4±3.9	66.1±1.0	86.6±6.6	82.7±1.1	82.2±4.1	**86.1±0.6**	71.3±1.3	0.800±0.068	1.505±0.177	0.537±0.145
FUSION$_{NMR}$	93.0±1.6	93.0±2.4	64.3±1.9	83.5±10.6	81.4±3.1	**84.5±2.9**	85.8±1.1	70.9±1.1	0.783±0.105	1.472±0.072	0.552±0.029
FUSION$_{IMAGE}$	92.9±3.4	92.9±2.4	64.3±1.6	88.6±4.6	83.0±0.9	81.6±4.8	85.8±3.8	70.6±1.7	**0.746±0.060**	1.587 ±0.143	0.549±0.025
FUSION$_{FINGERPRINT}$	**93.2±0.8**	**93.6±2.7**	65.8±0.7	85.4±9.4	82.4±3.1	81.6±2.5	85.3±1.1	71.1±1.1	0.818±0.054	1.535±0.080	0.573±0.040
FUSION$_{AVERAGE}$	90.2±6.8	93.4±2.7	67.0±0.6	**88.8±4.6**	80.8±2.2	79.2±5.4	85.4±0.8	**71.4±0.2**	0.781±0.082	1.528±0.180	0.559±0.018

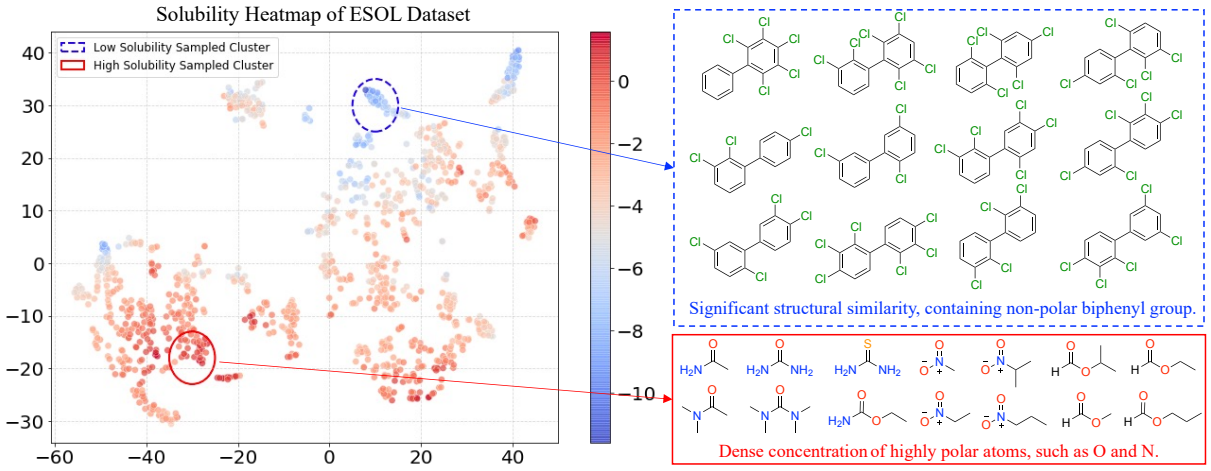

Figure 2. T-SNE visualization depicting the ESOL molecule embeddings alongside molecules within the highlighted region. Each point in the heatmap corresponds to the embeddings of respective molecules in ESOL, with color indicating solubility levels. Red denotes higher solubility, while blue indicates lower solubility.

MPSs, accompanied by their respective groups. As expected, molecular graph embeddings within the same group exhibit proximity after T-SNE reduction. This proximity is a result of shared identical motifs among graphs within each group.

In the development of new inhibitors, it is common to introduce additional functional groups or complex motifs to have more interactions with enzyme, thereby increasing the complexity of the molecules and making synthesis more challenging. However, such modifications do not guarantee a positive contribution or reward to inhibitor design. It is evident by the results of extending molecules from MPS to original molecules, which demonstrate both positive and negative contributions to binding potential. **Negative Contribution:** The complex molecules exhibits lower binding potentials than simple MPS **4**. This disparity can be illustrated by the distinction between MPS **4** and MCS **4a**: in MPS **4**, one of the NH groups, as a secondary amine, establishes a strong binding interaction with the enzyme, but this NH group transforms into an amide group in MCS **4a**, resulting in a significant decrease in binding capability. Therefore, these designs do not receive commensurate rewards considering the increased difficulty in synthesis. **Positive Contribution:** the complex molecules demonstrates higher binding potentials than MPS **9**, indicating successful design. This outcome can be attributed to MCS **9a**, which features more nitrogen and fluorine atoms capable of forming strong bindings with the enzyme. In summary, this experiment can serve as valuable guidelines for advancing inhibitor development.

5. Related work

Contrastive Learning on Molecular Graphs. The primary focus within the domain of contrastive learning applied to molecular graphs centers on 2D-2D graphs comparisons. Noteworthy representative examples: InfoGraph (Sun et al., 2019) maximizes the mutual information between the representations of the graph and its substructures to guide the molecular representation learning; GraphCL (You et al.,

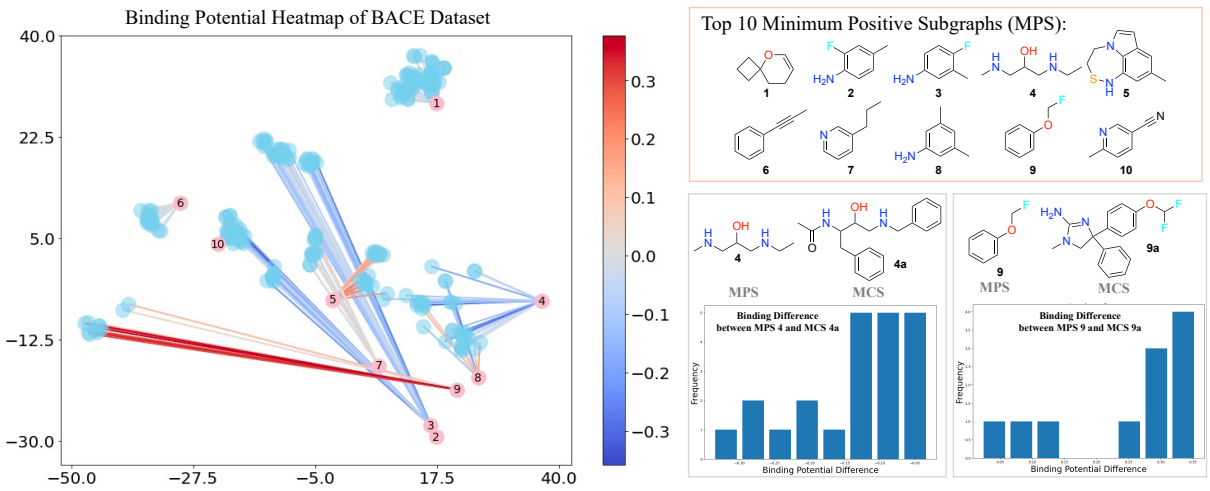

Figure 3. T-SNE Visualization of BACE embedding and clustering based on minimum positive subgraph (MPS). MPS represents minimum positive subgraph of a positive molecule; MCS represents maximum common subgraph of several positive molecules, sharing the same MPS. Pink nodes represent MPS, blue nodes depict molecules, and edge colors indicate binding potential differences. Red edges denote successful designs (original higher than MPS), while blue indicates less efficient designs (original lower than MPS).

2020a), 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 contratsive leraning, as only one graph encoder is employed in these studies. And these approaches often focus on the motif and graph levels, leaving atom-level contrastive learning less explored.

Multi-Similarity Learning. Instance-wise discrimination, a crucial facet of similarity learning, involves evaluating the similarity between instances directly based on their latent representations or features (Wu et al., 2018b). Naive instance-wise discrimination relies on self-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 relative similarities (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). Notably, it is crucial to emphasize that employing these pair weightings requires the manual categorization of negative and positive pairs, as distinct weights are assigned to losses based on their categories.

6. Discussion

In summary, unlike other multi-similarity learning approaches that require explicit categorization of negative and positive pairs, our method enables a straightforward generalization of similarity measures, encapsulating both self-similarities and relative-similarities. Meanwhile, the generalized multi-similarity metrics satisfy the requirement of convergent similarity learning. Notably, our model adeptly integrates chemical semantics from diverse modalities, enhancing its performance across various downstream tasks. Additionally, our framework bridges machine learning and chemical domain knowledge through post-hoc experiment by identifying easily synthesizable and functional substructures, which can be refined into appropriate configurations by experts. 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.

Accessibility

The code and dataset will be made available upon the date of publication.

References

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

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

Clevert, D.-A., Le, T., Winter, R., and Montanari, F. Img2mol–accurate smiles recognition from molecular graphical depictions. *Chemical science*, 12(42):14174–14181, 2021.

Costanti, F., Kola, A., Scarselli, F., Valensin, D., and Bianchini, M. A deep learning approach to analyze nmr spectra of sh-sy5y cells for alzheimer's disease diagnosis. *Mathematics*, 11(12):2664, 2023.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Fang, Y., Zhang, Q., Zhang, N., Chen, Z., Zhuang, X., Shao, X., Fan, X., and Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, pp. 1–12, 2023.

Gerothanassis, I. P., Troganis, A., Exarchou, V., and Barbarossou, K. 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.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Hadsell, R., Chopra, S., and LeCun, Y. 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.

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

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 2023.

Hoffer, E. and Ailon, N. 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.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Hu, Z., Dong, Y., Wang, K., Chang, K.-W., and Sun, Y. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1857–1867, 2020.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Jin, W., Barzilay, R., and Jaakkola, T. Multi-objective molecule generation using interpretable substructures, 2020.

Lambert, J. B., Mazzola, E. P., and Ridge, C. D. *Nuclear magnetic resonance spectroscopy: an introduction to principles, applications, and experimental methods*. John Wiley & Sons, 2019.

Landrum, G. Rdkit: Open-source cheminformatics. 2006. *Google Scholar*, 2006.

Law, M. T., Thome, N., and Cord, M. Quadruplet-wise image similarity learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 249–256, 2013.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the Thirty-Six AAAI Conference on Artificial Intelligence*, pp. 4541–4549, 2022.

Liu, H., Huang, Y., Liu, X., and Deng, L. Attention-wise masked graph contrastive learning for predicting molecular property. *Briefings in bioinformatics*, 23(5):bbac303, 2022a.

Liu, S., Demirel, M. F., and Liang, Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021a.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=xQUe1pOKPam.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021b.

Liu, Y., Jin, M., Pan, S., Zhou, C., Zheng, Y., Xia, F., and Philip, S. Y. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5879–5900, 2022c.

Mahé, P., Ueda, N., Akutsu, T., Perret, J.-L., and Vert, J.-P. Extensions of marginalized graph kernels. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 70, 2004.

Moon, K., Im, H.-J., and Kwon, S. 3d graph contrastive learning for molecular property prediction. *Bioinformatics*, 39(6):btad371, 2023.

Moutafis, P., Leng, M., and Kakadiaris, I. A. An overview and empirical comparison of distance metric learning methods. *IEEE transactions on cybernetics*, 47(3):612–625, 2016.

Mu, E., Guttag, J., and Makar, M. Multi-similarity contrastive learning. *arXiv preprint arXiv:2307.02712*, 2023.

Müller, K.-R., Mika, S., Tsuda, K., and Schölkopf, K. An introduction to kernel-based learning algorithms. In *Handbook of neural network signal processing*, pp. 4–1. CRC Press, 2018.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Peng, Z., Dong, Y., Luo, M., Wu, X.-M., and Zheng, Q. Self-supervised graph representation learning via global context prediction. *arXiv preprint arXiv:2003.01604*, 2020.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.

Steinbeck, C., Krause, S., and Kuhn, S. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of chemical information and computer sciences*, 43(6):1733–1739, 2003.

Suárez-Díaz, J. L., García, S., and Herrera, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges (with appendices on mathematical background and detailed algorithms explanation). *arXiv preprint arXiv:1812.05944*, 2018.

Sun, F.-Y., Hoffmann, J., Verma, V., and Tang, J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

Sun, M., Xing, J., Wang, H., Chen, B., and Zhou, J. 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.

Wang, H., Li, W., Jin, X., Cho, K., Ji, H., Han, J., and Burke, M. D. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=6sh3pIzKS-.

Wang, J., Zhou, F., Wen, S., Liu, X., and Lin, Y. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pp. 2593–2601, 2017.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. 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.

Wang, Y., Min, Y., Shao, E., and Wu, J. Molecular graph contrastive learning with parameterized explainable augmentations. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1558–1563. IEEE, 2021.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b.

Wang, Y., Chen, S., Chen, G., Shurberg, E., Liu, H., and Hong, P. Motif-based graph representation learning with application to chemical molecules. In *Informatics*, volume 10, pp. 8. MDPI, 2023.

Wen, Y., Liu, W., Feng, Y., Raj, B., Singh, R., Weller, A., Black, M. J., and Schölkopf, B. Pairwise similarity learning is simple. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5308–5318, 2023.

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

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018a.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018b.

Xia, J., Zhao, C., Hu, B., Gao, Z., Tan, C., Liu, Y., Li, S., and Li, S. Z. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2022.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., 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.

Xu, H., Wang, Y., Li, Y., and Hong, P. Asymmetric contrastive multimodal learning for advancing chemical understanding. *arXiv preprint arXiv:2311.06456*, 2023a.

Xu, H., Zhou, Z., and Hong, P. Molecular identification and peak assignment: Leveraging multi-level multimodal alignment on nmr. *arXiv preprint arXiv:2311.13817*, 2023b.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Yang, L. and Jin, R. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2(2):4, 2006.

Yang, Z., Song, J., Yang, M., Yao, L., Zhang, J., Shi, H., Ji, X., Deng, Y., and Wang, X. Cross-modal retrieval between 13c nmr spectra and structures for compound identification using deep contrastive learning. *Analytical Chemistry*, 93(50):16947–16955, 2021.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020a.

You, Y., Chen, T., Wang, Z., and Shen, Y. When does self-supervision help graph convolutional networks? In *international conference on machine learning*, pp. 10871–10880. PMLR, 2020b.

Zhang, C., Luo, L., and Gu, B. Denoising multi-similarity formulation: a self-paced curriculum-driven approach for robust metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11183–11191, 2023.

Zhang, L., Shen, S., Li, L., Wang, H., Li, X., and Lang, J. Jointly multi-similarity loss for deep metric learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1469–1474. IEEE, 2021.

Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., and Li, X. Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology*, 73:102327, 2022.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. 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).

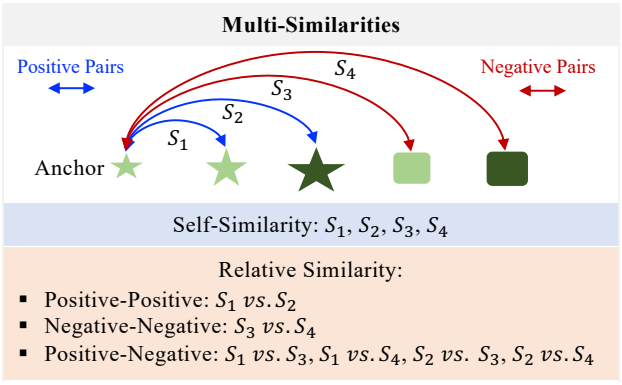

Figure A.1. Illustration of Different Types of Similarities.

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., 2020a), 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).

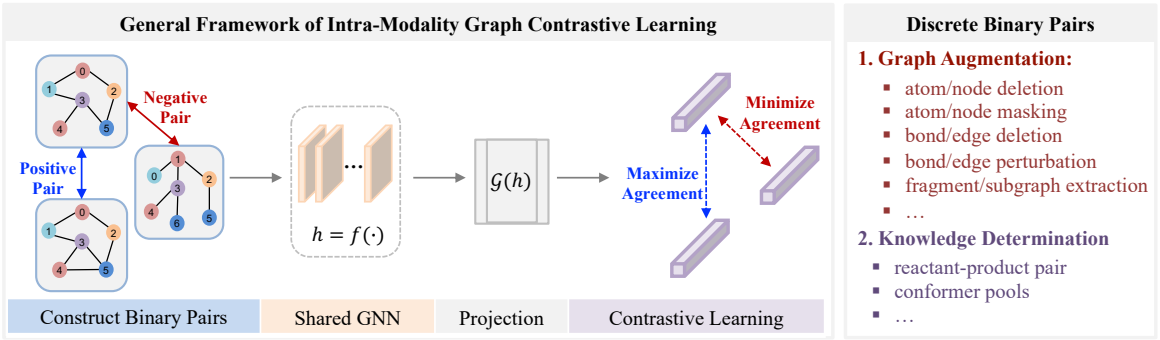

Figure A.2. General framework of Intra-Modality Graph Contrastive Learning. It relies on definition of positive and negative pairs.

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 $\sum_{j=1}^{|\mathcal{S}|} t_{i,j} = 1$, consider the loss function for an instance i defined as follows:*

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

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

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

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 $1 \le j \le |\mathcal{M}|$. 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:

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

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 generosity, suppose $t_{i,j} > t_{i,j'}$:

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

□

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:

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

C. Revisiting Multi-Similarity Settings

C.1. Encoders & Packages

To derive the self-similarities, we need to reply on pre-trained encoders or well-defined packages as follows:

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

C.2. Self-Similarity at Graph Level

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

$$S^F_{i,j} = Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{C.1}$$

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 $S^I_{i,j}$, can be defined as follows:

$$S^I_{i,j} = Cos(\mathcal{V}_i, \mathcal{V}_j) = \frac{\mathcal{V}_i \cdot \mathcal{V}_j^T}{\|\mathcal{V}_i\| \cdot \|\mathcal{V}_j\|} \tag{C.2}$$

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 $S^C_{i,j}$, can be defined as follows:

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

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 $S^S_{i,j}$, can be defined as follows:

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

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

C.3. A Brief Introduction to PPM

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 Fused Multimodal Generalized Similarity Metric

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

$$t_{i,j}^g = t_{i,j}^M = w_{SM} \cdot t_{i,j}^{SM} + w_C \cdot t_{i,j}^C + w_I \cdot t_{i,j}^I + w_F \cdot t_{i,j}^F \tag{C.5}$$

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, and f 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 $w_{SM} + w_C + w_I + w_F = 1$.

For pre-defined weights, denoted as w_{SM}, w_C, w_I, and w_F, we configure them with various settings for ablation studies, as shown in the following table:

Table C.2. The configuration of weight for each unimodality similarity in different Fusion. In particular w_S, w_N, w_M and w_F represents the weight of Smiles, NMR, Image and Fingerprint, respectively.

FUSED MULTIMODAL	w_{SM}	w_N	w_M	w_F
SMILES	1.00	0.00	0.00	0.00
NMR	0.00	1.00	0.00	0.00
IMAGE	0.00	0.00	1.00	0.00
FINGPERPRINT	0.00	0.00	0.00	1.00
FUSION$_{\text{SMILES}}$	0.70	0.10	0.10	0.10
FUSION$_{\text{NMR}}$	0.10	0.70	0.10	0.10
FUSION$_{\text{IMAGE}}$	0.10	0.10	0.70	0.10
FUSION$_{\text{FINGERPRINT}}$	0.10	0.10	0.10	0.70
FUSION$_{\text{AVERAGE}}$	0.25	0.25	0.25	0.25

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 30,000 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 GraphMSL'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. GROVER (Rong et al., 2020) integrates Message Passing Networks into a Transformer-style architecture, creating a class of more expressive encoders for molecules. Moreover, our evaluation encompasses a range of contrastive learning methods, namely InfoGraph (Sun et al., 2019), GraphCL (You et al., 2020a), 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.

E. More Ablation Study

GraphMSL$_{\text{bi-level}}$ demonstrates top performance in BBBP, BACE, CLINTOX, HIV, MUV, and Tox21. Additionally, within the GraphMSLbi-level framework, FusionSmiles enhances performance in Sider, CLINTOX, and MUV. FusionNMR improves results in BBBP and HIV, while Fusion$_{\text{Image}}$ contributes to better outcomes in the Tox21 and ToxCast tasks.

Table E.1. Ablation study on the performances of GraphMSL with bi-level (Graph + Node) and solo-node-level. The best results are denoted in boldf, and the second-best are indicated with underlining.

DATA SET	BBBP	BACE	SIDER	CLINTOX	HIV	MUV	TOX21	TOXCAST	ESOL	FREESOLV	LIPO
SMILES + NODE	93.7±1.3	93.2±3.7	65.9±1.7	87.7±7.8	82.3±1.8	80.9±5.5	84.7±1.7	70.4±1.5	0.873± 0.085	1.658± 0.243	0.594± 0.031
NMR + NODE	91.9±3.1	92.8±1.6	65.7±1.6	89.5±3.4	81.2±1.2	80.6±5.5	85.1±0.2	69.3±0.1	1.052± 0.105	2.391± 0.175	0.654± 0.025
IMAGE + NODE	94.1±1.7	90.8±0.7	63.8±2.8	86.5±8.0	80.3±1.5	76.7±2.7	85.3±1.1	70.8±1.5	**0.843± 0.094**	**1.601± 0.057**	**0.562± 0.005**
FINGERPRINT + NODE	90.2±8.4	**94.5±0.7**	64.3±2.9	91.0±1.5	82.0±2.4	79.2±5.9	85.7±0.5	69.7±1.3	1.170± 0.174	2.801± 0.276	0.607± 0.034
FUSION$_{\text{SMILES}}$ + NODE	91.7±5.3	91.5±1.7	**67.3±0.6**	**93.8±0.8**	82.1±1.7	**81.5±3.7**	85.1±0.1	70.4±1.3	0.965±0.085	2.859±0.281	0.647±0.029
FUSION$_{\text{NMR}}$ + NODE	**94.3±0.8**	93.6±1.9	66.8±1.4	90.4±3.1	**83.0±0.7**	80.1±3.5	85.5±0.6	70.6±1.8	1.009±0.160	2.224±0.368	0.589±0.038
FUSION$_{\text{IMAGE}}$ + NODE	94.2±1.2	93.1±2.5	66.4±1.6	90.7±3.5	82.0±2.4	80.8±3.8	**86.1±0.8**	**71.2±1.1**	0.898±0.098	1.691±0.386	0.579±0.018
FUSION$_{\text{FINGERPRINT}}$ + NODE	91.6±5.0	94.3±2.4	66.4±1.9	85.3±6.8	82.0±2.4	80.6±3.2	85.2±0.2	69.8±1.1	1.037±0.170	2.093±0.090	0.607±0.034
FUSION$_{\text{AVERAGE}}$ + NODE	92.7±1.5	92.6±2.1	65.6±0.7	89.3±4.0	81.8±1.7	81.0±5.0	85.4±1.3	71.2±1.9	1.019±0.118	1.733±0.267	0.593±0.004
NODE	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