Cross-View Contrastive Fusion for Enhanced Molecular Property Prediction

Yan Zheng[1] , **Song Wu**[1] , **Junyu Lin**[2] , **Yazhou Ren**[1,2*] ,
Jing He[3] , **Xiaorong Pu**[1,2] and **Lifang He**[4]
[1]School of Computer Science and Engineering,University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3]Nuffield Department of Clinical Neurosciences, University of Oxford
[4]Department of Computer Science and Engineering, Lehigh University
{yan9zheng9, linjunyuxx, lotusjing}@gmail.com, songwu.work@outlook.com,
{yazhou.ren, puxiaor}@uestc.edu.cn, lih319@lehigh.edu

Abstract

Machine learning based molecular property prediction has been a hot topic in the field of computer aided drug discovery (CADD). However, current MPP methods face two prominent challenges: 1) single-view MPP methods do not sufficiently exploit the complementary information of molecular data across multiple views, generally producing suboptimal performance, and 2) most existing multi-view MPP methods ignore the disparities in data quality among different views, inadvertently introducing the risk of models being overshadowed by inferior views. To address the above challenges, we introduce a novel cross-view contrastive fusion for enhanced molecular property prediction method (MolFuse). First, we extract intricate molecular semantics and structures from both sequence and graph views to leverage the complementarity of multi-view data. Then, MolFuse employs two distinct graphs, the atomic graph and chemical bond graph, to enhance the representation of the molecular graph, allow us to integrate both the fundamental backbone attributes and the nuanced shape characteristics. Notably, we incorporate a dual learning mechanism to refine the initial feature representations, and global features are obtained by maximizing the coherence among diverse view-specific molecular representations for the downstream task. The overall learning processes are combined into a unified optimization problem for iterative training. Experiments on multiple benchmark datasets demonstrate the superiority of our MolFuse.

1 Introduction

In modern drug research and development (R&D), accurately predicting molecular properties is crucial for speeding up drug launch and enhancing R&D efficiency. Molecular Property Prediction (MPP) technologies offer an innovative approach for estimating properties based on molecular structural data, thereby refining the screening and validation of candidate drug molecules.

Existing MPP approaches can be broadly categorized into single-view and multi-view MPP methods. Single-view MPP methods uncover molecular traits from singular data perspectives,e.g., molecular knowledge graphs [Fang *et al.*, 2022], molecular fingerprints [Kearnes *et al.*, 2016], SMILES representations [Honda *et al.*, 2019], and 2D molecular graphs [Hu *et al.*, 2020]. These data capture different angles and features of molecules to achieve specific downstream tasks, e.g., activity prediction, and drug-target prediction. In contrast, multi-view MPP methods [Zhu *et al.*, 2021; Liu *et al.*, 2021; Ma *et al.*, 2022; Zhou *et al.*, 2023], considering more comprehensive contextual information, have garnered increasing attention in this field. However, existing multi-view MPP approaches generally have not considered quality variations across views, lacking consistency in learning and interaction for the same molecule. Low-quality information could easily misguide the learning of common semantics. For instance, [Zhu *et al.*, 2022] concatenate all views to obtain global features, where the inconsistency between the latent feature spaces is ignored.

To this end, we implement a two-stage feature learning framework for predicting molecular properties. Specifically, we consider different types of molecular data, including biophysical (*e.g.*, BACE) and physiological (*e.g.*, ClinTox) datasets. Biophysical datasets focus on molecule-biomolecule interactions, emphasizing physical traits. On the other hand, physiological datasets spotlight bioactivity, toxicity, and metabolic processes in biological systems. They comprehensively reveal bond types and molecular connectivity, which are particularly critical when certain toxicities link to specific structures and bonds.

Balancing these aspects is complex, as detailed atomic interactions can become intricate, especially in larger molecules, impacting model training. Focusing solely on chemical bond embeddings might inadvertently gloss over granular atomic details, an omission that might be detrimental for biophysical datasets that rely heavily on atomic interaction insights. To bridge this gap, inspired by [Ma *et al.*, 2022], we introduce a novel approach which synergizes the

*Corresponding author.

topological and spatial perspectives of molecules in the first stage, which crafts a holistic portrayal that encompasses both its backbone and shape attributes, thereby streamlining the representational domain while ensuring the nuanced requirements of both dataset types are met.

In the second stage, we design a dual learning loss to utilize view complementarity to refine the feature representation, where the reliable semantic information between different views are shared. Furthermore, we implement cross-view contrastive fusion to obtain the global features. In this module, the mutual information between global features and view-specific refined features is maximized. Specifically, two linear feature transformation layers are constructed for different views to yield their refined features. Then, a nonlinear MLP is used to fuse all views' features to generate global features, thereby exploring view consensus information from different views.

The main contributions of our work are as follows:

- We emphasize the comprehensive learning of molecular information through integrating multiple views, fully considering molecular sequence features, atomic features along with their neighboring attributes, and chemical bond angular features.
- We introduce a multi-view contrastive fusion method, which jointly conduct cross-view interaction and consistency learning on latent features across views, thereby addressing inconsistency in multi-view molecular data while maintaining consistent shared semantics.
- Extensive experiments demonstrate the approach outperforms other state-of-the-art methods on multiple molecular benchmark datasets.

2 Related Work

This section provides a concise overview of molecular machine learning methods, including single-view MPP and multi-view MPP.

2.1 Single-View Molecular Property Prediction

For molecular prediction work, RDKit [Tosco *et al.*, 2014] can process SMILES strings [Anderson *et al.*, 1987], subsequently converting them into feature vectors or molecular fingerprints. These vectors or fingerprints can be used as inputs for machine learning models. Sequence-based MPP methods [Chithrananda *et al.*, 2020; Honda *et al.*, 2019; Ross *et al.*, 2022] decompose SMILES into a series of tokens representing atoms/bonds and then apply deep models on these tokens. In addition to the traditional use of sequence data, graph-based MPP models are one of the dominant models. [Hu *et al.*, 2020] propose a graph-based pretraining model, which pre-train expressive GNNs at the level of individual nodes and entire graphs. [Zhuang *et al.*, 2023] emphasize the relation between each molecule and its multiple properties, and construct a meta-learning framework with scheduling subgraph sampling by contrastive loss. [Xiong *et al.*, 2020] propose to learn molecular characterization from relevant drug discovery datasets, and employ a graph neural network architecture for molecular representation. [You *et al.*, 2020] propose Graph Contrastive Learning (GraphCL), which is a general framework for learning node representations. On this basis, [You *et al.*, 2021] proposes a unified bi-level optimization framework to automatically select data augmentations when performing GraphCL on specific graph data. [Wang *et al.*, 2022] present a self-supervised learning framework for large unlabeled molecule datasets. [Stärk *et al.*, 2022] uses 3D pre-training to provides significant improvements for different properties.

2.2 Multi-View Molecular Property Prediction

Based on the diversity of molecular data, such as molecular fingerprints, SMILES representation, 2D molecular graph and 3D molecular graph, each view captures different angles and properties of molecules, making the technology utilizing multi-view properties gain more attention in this field. [Guo *et al.*, 2022] propose a multilingual molecular embedding generation approach, incorporating molecular SMILES and InChI for pretraining. [Zhu *et al.*, 2021] design a new pretraining algorithm, dual-view molecule pre-training, which constructs an auxiliary contrast loss using molecular map features and sequence features. [Liu *et al.*, 2021] employ self-supervised learning to enhance 2D molecular graph encoders. The enhancement is through correspondence and consistency between 2D topological structures and 3D geometric views. [Fang *et al.*, 2022] combine chemical knowledge graphs with molecular graphs. By integrating domain knowledge into graph semantics, the model can consider the correlation between atoms with shared properties, thereby capturing the features of molecules in the representation of the molecular graph more accurately. [Stärk *et al.*, 2022] utilize existing 3D molecular datasets to pre-train models. It infers the geometry of molecules in advance and then employs a self-supervised learning approach, maximizing mutual information of 3D summary vectors for downstream tasks. [Yang *et al.*, 2023] propose a gradient perturbation-based contrastive Learning to generate positive and negative molecules to optimize the exposure bias problem in translational molecule optimization. [Zhu *et al.*, 2022] present a novel multi-view contrastive learning approach, and a unique pretraining framework which is learned from four molecular representations. In response to molecules are treated as 1D sequential tokens or 2D topology graphs in most MRL methods, [Zhou *et al.*, 2023] propose a universal MRL framework named Uni-Mol. All these valuable works have demonstrated the superior performance of multi-view learning in molecular prediction.

3 Proposed Method

3.1 Multi-View Molecule Feature Learning

Embedding SMILES Strings

First, we use the SMILES to represent molecular sequences, which offers a concise and expressive format. In this way, each atom is uniquely represented using its ASCII symbol and employs distinctive symbols to articulate chemical bonds, branches, and stereochemical configurations. The proposed MolFuse begins by dissecting the SMILES molecular descriptors into a series of tokens e_i. The molecular sequence $\mathbf{S}$ is an ordered list of e_i, expressed as $\mathbf{S} =<$

Notation	Descriptions
$\mathbf{S}$	An ordered list of chemical symbols, representing molecular sequence.
$\mathcal{T} = (\mathcal{V}, \mathcal{E})$	Molecular topological graph.
$\mathcal{M} = (\mathcal{E}, \mathcal{A})$	Molecular spatial graph.
$v \in \mathcal{V}$	Node in the molecular topological graph $\mathcal{T}$, representing atom v .
$(u, v) \in \mathcal{E}$	Edge in the molecular topological graph $\mathcal{M}$, representing the chemical bond (u, v).
$\mathcal{N}(v)$	The neighboring atoms of atom v.
$a_{uvw} \in \mathcal{A}$	Bond angles formed between two consecutive chemical bonds (u, v) and (v, w), unified by the shared atom v.
$\mathbf{H}^{seq}, \mathbf{H}^{gra}$	The latent embedding for the sequence view and graph view, respectively.
$\mathbf{Z}^{seq}, \mathbf{Z}^{gra}$	The refined molecular representations for the sequence view and the graph view, respectively.
$\hat{\mathbf{Z}}^{seq}, \hat{\mathbf{Z}}^{gra}$	The reconstructed features for the sequence view and the graph view, respectively.
$\mathbf{Z}^{fuse}$	The global fused features used to final classification.

Table 1: Nomenclature.

$e_1, e_2, e_3, \cdots, e_n > (1 \leq i \leq n)$. Subsequently, a GRU is applied to these tokens to obtain the molecular embedding $\mathbf{H}^{seq}$ in sequence view:

$$\mathbf{H}^{seq} = \overleftarrow{g_{seq}}(\phi(\mathbf{S})) \oplus \overrightarrow{g_{seq}}(\phi(\mathbf{S})), \tag{1}$$

where the embedding layer ϕ encodes the molecular sequence $\mathbf{S}$, the $\oplus$ operator symbolizes matrix concatenation, and the $g_{seq}(\cdot)$ signifies the bidirectional gated recurrent unit (GRU) [Chung *et al.*, 2014]. This design ensures that the model recognizes and integrates information from both backward and forward states in the sequence, employed to capture contextual information within the molecular sequence. In particular, in the first stage, we train our GRU units under the constraints of the training dataset labels. In order to simplify the computational process and stabilize the pre-trained model, we strategically freeze the GRU units in the second stage.

Embedding Topological and Spatial Graph

Considering the molecular structure represented by the topological graph $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, the individual nodes $v \in \mathcal{V}$ are emblematic of distinct atoms, while each edge $(u, v) \in \mathcal{E}$ encapsulates the chemical bond formed between atoms u and v. We denote the atom attribute as x_v for the atom v, and the bond attribute as $x_{(u,v)}$ for the bond (u, v).

Diving deeper into the spatial intricacies, the molecular spatial graph $\mathcal{M} = (\mathcal{E}, \mathcal{A})$ is presented. Within this depiction, each node $(u, v) \in \mathcal{E}$ is indicative of the chemical bond established between atoms u and v. The edge $a_{uvw} \in [0, \pi]$ captures the bond angles formed between two consecutive chemical bonds (u, v) and (v, w), unified by the shared atom v. This delineation provides an insight into both the molecular configuration and spatial orientation. Specifically, we conduct a comprehensive graph representation by considering both topological graph and spatial graph.

In topological graph $\mathcal{T}$, we leverage the graph neural network (GNN) to get the graph representation. Specifically, the embedding of atom v is initialized as $\mathbf{h}_v^{(0)} = x_v$, and the $\mathbf{m}_v^{(k)}$ represents the hidden state of atom v during the k-th iteration. The GNN produces the embedding $\mathbf{h}_v^{(k)}$ for atom v at the k-th iteration, which can be described by:

$$\mathbf{m}_v^{(k)} = \mathcal{F}^{(k)} \left\{ \mathbf{h}_v^{(k-1)}, \mathbf{h}_w^{(k-1)}, \mathbf{h}_{(v,w)}^{(k-1)} \mid w \in \mathcal{N}(v) \right\}, \tag{2}$$

$$\mathbf{h}_v^{(k)} = \mathcal{C}^{(k)} \left\{ \mathbf{h}_v^{(k-1)}, \mathbf{m}_v^{(k)} \right\}, \tag{3}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{C}(\cdot)$ denote aggregating messages from the neighbours and updating the node representation in the GNN module, respectively. The $\mathcal{N}(v)$ indicates the neighbouring atoms of v at the k-th iteration. Particularly, $\mathbf{h}_{(v,w)}^{(k-1)}$ denotes the embedding of bond (v, w) learned at the $(k-1)$-th iteration in a molecular spatial graph $\mathcal{M}$ as Eq. (5).

Similarly, in spatial graph $\mathcal{M}$, the embeddings of chemical bonds (u, v) and (v, w) are also initialized as $\mathbf{h}_{(u,v)}^{(0)} = x_{(u,v)}$ and $\mathbf{h}_{(v,w)}^{(0)} = x_{(v,w)}$, undergo training via GNN. Let $\mathbf{m}_{(u,v)}^{(k)}$ denote the hidden state of the chemical bond (u, v) during the k-th iteration. Finally, the GNN produces the embedding $\mathbf{h}_{(u,v)}^{(k)}$ for this bond at k-th iteration:

$$\begin{aligned} \mathbf{m}_{(u,v)}^{(k)} = \mathcal{F}^{(k)} \Big\{ &\Big(\mathbf{h}_{(u,v)}^{(k-1)}, \mathbf{h}_{(u,w)}^{(k-1)}, a_{wuv} \mid w \in \mathcal{N}(u) \Big) \\ &\cup \Big(\mathbf{h}_{(u,v)}^{(k-1)}, \mathbf{h}_{(v,w)}^{(k-1)}, a_{uvw} \mid w \in \mathcal{N}(v) \Big) \Big\}, \end{aligned} \tag{4}$$

$$\mathbf{h}_{(u,v)}^{(k)} = \mathcal{C}^{(k)} \left\{ \mathbf{h}_{(u,v)}^{(k-1)}, \mathbf{m}_{(u,v)}^{(k)} \right\}, \tag{5}$$

where $\mathcal{N}(u)$ and $\mathcal{N}(v)$ denote the neighbouring atoms of u and v, respectively. Similarly, the $\mathcal{F}(\cdot)$ and $\mathcal{C}(\cdot)$ denote aggregating messages from the neighbours and updating the node representation in GNN module.

To achieve an advanced molecular representation that adaptively combines both topological and spatial views, we employ $\mathrm{READOUT}$ function to make it integrate local node embeddings into a coherent global graph representation:

$$\mathbf{H}^{gra} = \mathrm{READOUT} \left\{ \mathbf{h}_v^{(K)} \mid v \in \mathcal{V} \right\}, \tag{6}$$

where $\mathbf{H}^{gra}$ is the learned graph representation, and the K is the number of iterations. Notably, with restrictions of the training dataset's label, we can get the optimized molecular representation under the graph view, which will stop gradient propagation in the second stage.

3.2 Cross-View Representation Fusion

After the first learning stage, we obtain both optimized molecular representations under the sequence view and graph view. We design a novel contrastive fusion framework for learning more comprehensive joint representation from two

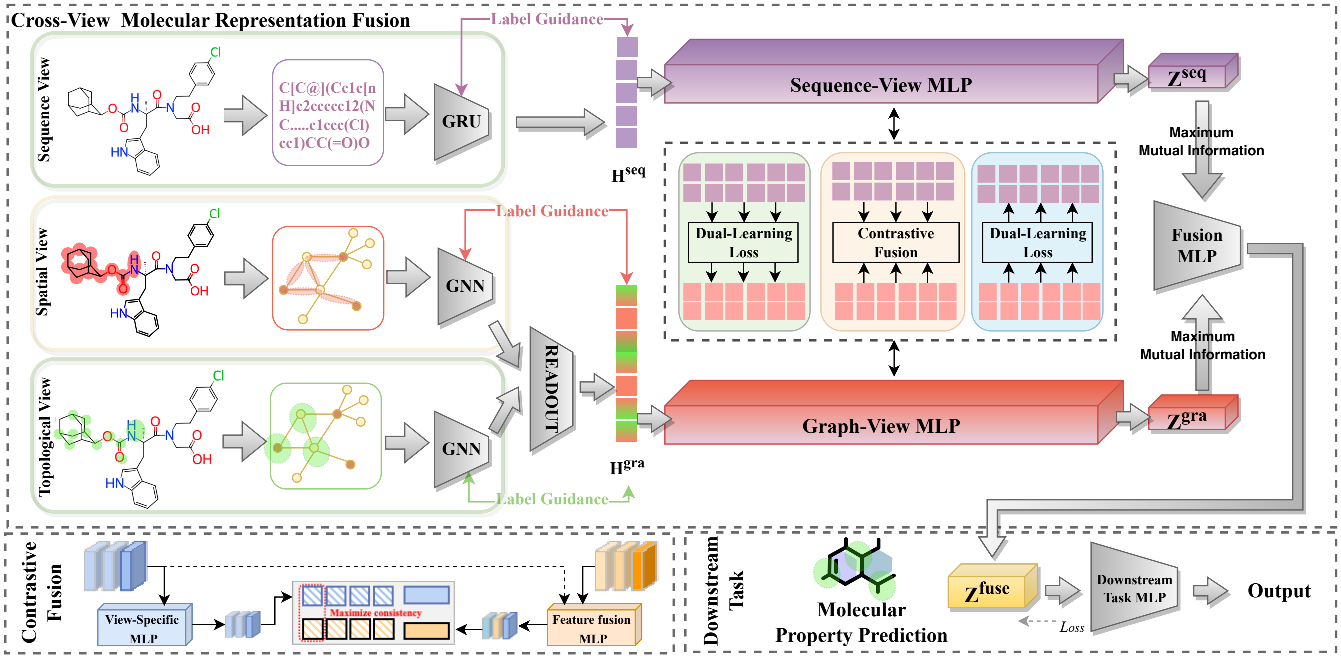

Figure 1: An overview of the proposed MolFuse. (a) Extracting intricate molecular semantics and structures from both sequence and graph views, to optimize molecular representations at distinct granularities, providing a holistic portrayal of molecules that integrates both backbone and shape attributes. (b) Leveraging the linear feature transformation layer $\mathcal{G}^{(\cdot)}$ as parametric function to facilitate the dual learning across the views from $\mathbf{Z}^{seq}$ to $\mathbf{Z}^{gra}$, ensures that the insights from high-quality views are imparted to their low-quality counterparts. Then, we maximize the consistency in the molecular representations specific to each view, to learn a more comprehensive global features.

views. Specifically, we propose a dual learning module, where view-specific features acquire reliable knowledge from another view to refine their own unclear feature expression. Another contrastive fusion module is designed to learn more comprehensive global features via maximizing consistency with view-specific molecular representations. Notedly, in our designed contrastive fusion framework, the consistency and complementarity under different views are fully explored.

Dual Learning Module

Let $\mathbf{h}_i^{seq}$ and $\mathbf{h}_i^{gra}$ denote the i-th molecular representations under the sequence view and the graph view, respectively. Generally, the diversity of molecular features brings inconsistency in molecular representation. We consider that two representations of the same molecule should be consistent in the latent space. To do this, we first stack two view-specific MLPs on the features $\mathbf{h}^{seq}$, and $\mathbf{h}^{gra}$ respectively, to project them into a consistent feature space. Specially, for a molecule i under the different views, the transformation is defined as:

$$\mathbf{Z}_i^{seq} = \sigma\left(\mathbf{W}_s^{\top}\mathbf{h}_i^{seq} + \mathbf{b}_s\right), \tag{7}$$

$$\mathbf{Z}_i^{gra} = \sigma\left(\mathbf{W}_g^{\top}\mathbf{h}_i^{gra} + \mathbf{b}_g\right), \tag{8}$$

where $\mathbf{Z}^{seq}$ and $\mathbf{Z}^{gra}$ denote the refined molecular representations for the sequence view and the graph view, respectively. The $\mathbf{W}_s$, $\mathbf{b}_s$, $\mathbf{W}_g$, $\mathbf{b}_g$ are the learnable parameters in the MLPs, and the σ is an activation function.

Considering that the characteristics and qualities of different views generally vary widely, we hope to utilize view complementarity to refine the feature expression of the low-quality view. Simply put, the low-quality view acquires reliable knowledge from other views. To this end, we design a dual learning module, which aims to achieve 1) different views maintain consistency in latent feature space and 2) reliable semantic information is delivered to each other from different views. Specifically, we construct two linear feature transformation layers $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, where the feature $\mathbf{Z}^{seq}$ is projected to maintain consistency with the feature $\mathbf{Z}^{gra}$, vice versa. Then the reconstructed features $\hat{\mathbf{Z}}^{gra}$ and $\hat{\mathbf{Z}}^{seq}$ can be described as:

$$\hat{\mathbf{Z}}_i^{gra} = \mathcal{G}^{(1)}\left(\mathbf{Z}_i^{seq}\right) = \mathbf{W}_{(1)}^{\top}\mathbf{Z}_i^{seq} + \mathbf{b}_{(1)}, \tag{9}$$

$$\hat{\mathbf{Z}}_i^{seq} = \mathcal{G}^{(2)}\left(\mathbf{Z}_i^{gra}\right) = \mathbf{W}_{(2)}^{\top}\mathbf{Z}_i^{gra} + \mathbf{b}_{(2)}, \tag{10}$$

where $\mathbf{W}_{(1)}$, $\mathbf{b}_{(1)}$, and $\mathbf{W}_{(2)}$, $\mathbf{b}_{(2)}$ denote the parameters on the feature transformation layers $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, respectively.

In this module, We employ a dual learning strategy, where the consistency objective is conducted by forcing view-specific features $\mathbf{Z}^{seq}$ and $\mathbf{Z}^{gra}$ to be consistent with their reconstruction objective $\hat{\mathbf{Z}}^{seq}$ and $\hat{\mathbf{Z}}^{seq}$. In this way, cross-view reliable information is delivered to each other. Then, the dual learning loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{dua} &= \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{Z}_i^{seq} - \hat{\mathbf{Z}}_i^{seq}\right\|_2^2 + \left\|\mathbf{Z}_i^{gra} - \hat{\mathbf{Z}}_i^{gra}\right\|_2^2\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{Z}_i^{seq} - \mathcal{G}^{(2)}\left(\mathbf{Z}_i^{gra}\right)\right\|_2^2 + \left\|\mathbf{Z}_i^{gra} - \mathcal{G}^{(1)}\left(\mathbf{Z}_i^{seq}\right)\right\|_2^2\right).\end{aligned} \tag{11}$$

Contrastive Fusion Module

In this module, we conduct cross-view contrastive fusion to obtain more comprehensive global features from two views. To leverage the global discriminative information, we first concatenate all origin embedding, *i.e.*, $\mathbf{h}^{seq}$ and $\mathbf{h}^{gra}$. Then, a multi-layer perceptron (MLP) is applied to fuse the sequence view representations with the graph view representations adaptively. Then, the global fused feature $\mathbf{Z}^{fuse}$ is obtained by:

$$\mathbf{Z}_i^{fuse} = \sigma\left(\mathbf{W}_{fuse}^{\top}\left(h_i^{seq} \oplus h_i^{gra}\right) + \mathbf{b}_{fuse}\right), \tag{12}$$

where $\oplus$ denotes the matrix concatenation, σ denotes the activation function, and $\mathbf{W}_{fuse}$, $\mathbf{b}_{fuse}$ denote the parameters on fusion MLP.

Motivated by the insight that the representations of the same molecule from different views are typically similar, we conduct contrastive fusion between view-specific features, *i.e.*, $\mathbf{Z}^{seq}$ and $\mathbf{Z}^{gra}$, and global fused feature $\mathbf{Z}^{fuse}$ to implement the fusion objective. In this respect, taking the sequence view as an example, we denote $\left\{\mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{seq}\right\}_{j=i}$ as positive feature pairs, and the rest $\left\{\mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{seq}\right\}_{j\neq i}$ are negative feature pairs. By contrastive learning, the model could achieve that the similarities of positive pairs are maximized, and negative pairs are minimized. Firstly, the cosine distance is used to measure the similarity of feature pairs:

$$sim\left(\mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{seq}\right) = \frac{\left\langle \mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{seq} \right\rangle}{\|\mathbf{Z}_i^{fuse}\| \|\mathbf{Z}_j^{seq}\|}, \tag{13}$$

where $\langle\cdot,\cdot\rangle$ is the dot product operator. Then, the contrastive loss under the sequence view can be expressed as:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim\left(\mathbf{Z}_i^{fuse}, \mathbf{Z}_i^{seq}\right)/\tau\right)}{\sum_{j=1,j\neq i}^{N}\exp\left(sim\left(\mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{seq}\right)/\tau\right)}, \tag{14}$$

where τ is a temperature coefficient. Similarly, we can also compute the contrastive loss under the graph view:

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim\left(\mathbf{Z}_i^{fuse}, \mathbf{Z}_i^{gra}\right)/\tau\right)}{\sum_{j=1,j\neq i}^{N}\exp\left(sim\left(\mathbf{Z}_i^{fuse}, \mathbf{Z}_j^{gra}\right)/\tau\right)}. \tag{15}$$

Then, we combine the contrastive losses under the two views, and the final contrastive fusion loss is:

$$\mathcal{L}_{con} = \mathcal{L}_s + \mathcal{L}_g. \tag{16}$$

After cross-view contrastive fusion, we use the global fused feature $\mathbf{Z}^{fuse}$ as the final molecular representation to predict the molecular property. Specifically, the probability $\hat{y}_i$ of the molecule i classified as a positive instance can be calculated as follows:

$$\hat{y}_i = \sigma\left(\mathbf{W}_{pred}^{\top}\mathbf{Z}_i^{fuse} + \mathbf{b}_{pred}\right), \tag{17}$$

where σ is an activation function, and the $\mathbf{W}_{pred}$, $\mathbf{b}_{pred}$ are learnable parameters on classifier MLP. The supervised loss $\mathcal{L}_{sup}$ based on cross entropy is defined as follows:

$$\mathcal{L}_{\text{sup}}\left(\hat{y}\right) = -\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right), \tag{18}$$

where y_i is the true label of the molecule i. In the end, we jointly optimize the supervised loss, the contrastive fusion loss, the dual learning loss. Specifically, the α and β are hyperparameters, and the overall loss function is:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{dua} + \beta\mathcal{L}_{con}. \tag{19}$$

Category	Dataset	#Compounds	#Task	Split
Physiology	BBBP	2,039	1	Scaffold
	Tox21	7,831	12	Random
	ClinTox	1,478	2	Random
	SIDER	1,427	27	Random
	ToxCast	8,575	617	Random
Biophysics	BACE	1,513	1	Scaffold
	HIV	41,127	1	Scaffold

Table 2: Details of datasets used in this study.

4 Experiments

In this section, we focus on the following three research questions to validate the effectiveness of MolFuse. Due to space limitation, appendix B offers more comprehensive analysis and more details on our experiments.

- **RQ1:** Can MolFuse surpass state-of-the-art on molecular property prediction task?
- **RQ2:** Does MolFuse enable high-quality view guidance through dual learning?
- **RQ3:** Does the multi-view fusion module explore the consistency and complementarity of molecules under different views?

4.1 Experimental Setup

Dataset Description

MoleculeNet [Wu *et al.*, 2017] is a large scale benchmark for molecular machine learning. It curates multiple public datasets and establishes metrics for evaluation. We use seven of these datasets, BBBP, BACE, Tox21, ClinTox, SIDER, ToxCast, HIV, for this experiment. The statistics of these datasets are summarized in Table 2.

Comparison Algorithms

We used three different types of baseline comparisons.

GNN Classifier. We employ a standard classifier that consists of a 5-layer GNN complemented by a single fully connect layer. This encompasses three distinctive GNN architectures: GCN [Kipf and Welling, 2016], GAT [Veličković *et al.*, 2018], and GIN [Xu *et al.*, 2018].

Single-View MMP Methods. We select the representative single-view MPP pretraining models, include AttrMask [Hu *et al.*, 2020], ContextPred [Hu *et al.*, 2020], GraphCL [You *et*

Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	HIV	BACE	Avg.↑
GCN (2016)	65.9 ± 0.9	74.4 ± 0.6	63.6 ± 1.1	60.6 ± 0.8	55.4 ± 3.6	75.2 ± 1.4	71.0 ± 4.6	66.5
GAT (2018)	64.9 ± 1.2	75.0 ± 0.8	63.5 ± 1.6	61.0 ± 1.1	58.9 ± 1.4	75.5 ± 1.7	75.3 ± 2.4	67.7
GIN (2018)	68.9 ± 1.2	74.3 ± 0.6	64.0 ± 1.6	58.1 ± 1.5	58.8 ± 5.7	75.6 ± 1.6	69.0 ± 4.7	66.9
AttrMask (2020)	65.0 ± 2.3	74.8 ± 0.2	62.9 ± 0.1	61.2 ± 0.1	87.7 ± 1.1	76.8 ± 0.5	79.7 ± 0.3	72.5
ContextPred (2020)	65.7 ± 0.6	74.2 ± 0.1	62.5 ± 0.3	62.2 ± 0.5	77.2 ± 0.8	77.1 ± 0.8	76.0 ± 2.0	70.7
GraphCL (2020)	69.7 ± 0.6	73.9 ± 0.6	62.4 ± 0.5	60.5 ± 0.8	76.0 ± 2.6	77.5 ± 1.2	75.4 ± 1.4	70.6
JOAO (2021)	66.0 ± 0.6	74.4 ± 0.7	62.7 ± 0.6	60.7 ± 1.0	66.3 ± 3.9	76.6 ± 0.5	72.9 ± 2.0	68.5
MolCLR (2022)	66.6 ± 1.8	73.0 ± 0.1	62.9 ± 0.3	57.5 ± 1.7	86.1 ± 0.9	76.2 ± 1.5	71.5 ± 3.1	70.5
GraphMVP (2021)	68.5 ± 0.2	74.5 ± 0.4	62.7 ± 0.1	62.3 ± 1.6	79.0 ± 2.5	74.8 ± 1.4	76.8 ± 1.1	71.2
3DInfomax (2022)	69.1 ± 1.0	74.5 ± 0.7	64.4 ± 0.8	60.6 ± 0.7	79.9 ± 3.4	76.1 ± 1.3	79.7 ± 1.5	72.0
MEMO (2022)	71.6 ± 1.0	76.7 ± 0.4	**64.9 ± 0.8**	61.2 ± 0.6	81.6 ± 3.7	78.3 ± 0.4	82.6 ± 0.3	73.8
Uni-Mol (2023)	71.3 ± 0.6	76.1 ± 0.2	63.6 ± 0.1	66.3 ± 0.9	92.0 ± 0.9	77.0 ± 0.8	85.1 ± 0.8	75.9
Ours	**74.3 ± 1.3**	**77.6 ± 0.4**	64.1 ± 0.3	**69.5 ± 1.0**	**95.5 ± 3.3**	**78.6 ± 0.9**	**87.2 ± 1.3**	**78.1**

Table 3: ROC-AUC (%) performance of different methods on seven binary classification tasks from MoleculeNet benchmark. The mean and standard derivation are reported. The best result is shown in bold and the second best result is underlined.

Dataset	BBBP	SIDER	BACE
w/o $\mathcal{L}_{dua}$&$\mathcal{L}_{con}$	59.8 ± 1.0	57.5 ± 1.7	73.9 ± 0.6
w/o $\mathcal{L}_{dua}$	63.8 ± 1.0	60.7 ± 1.0	77.6 ± 0.3
w/o $\mathcal{L}_{con}$	71.7 ± 0.7	62.3 ± 1.6	81.6 ± 3.7
Full modules	**74.3 ± 1.3**	**69.5 ± 1.0**	**87.2 ± 1.3**

Table 4: ROC-AUC (%) performance of ablation studies. All other parameters are kept the same for a fair comparison.

al., 2020], along with the recently introduced methods JOAO [You *et al.*, 2021] and MolCLR [Wang *et al.*, 2022].

Multi-View MMP Methods. We adopt multi-view MPP benchmarks, include GraphMVP [Liu *et al.*, 2021], 3D InfoMax [Stärk *et al.*, 2022], MEMO [Zhu *et al.*, 2022], and the newly revealed Uni-Mol [Zhou *et al.*, 2023].

Dataset Splitting

In addition to dataset selection, how to split the dataset rationally is another important factor in training the model. The dataset is usually split into training set, validation set and test set for benchmarking using random splitting, which is considered an effective and simple method in typical machine learning. However, due to the uniqueness of the molecular data, using random split molecular data is not an appropriate method. We used a splitting method called scaffold splitting, to cope with this situation, which is an attempt to split molecules with different structures into different sets [Bemis and Murcko, 1996]. And in this experiment, we used the dataset splitting method recommended in MoleculeNet.

Evaluation Metrics

Considering that the seven datasets in MoleculeNet used are all related to the classification task, we use the area under the curve of the receiver operating characteristic curve (ROC-AUC) to quantify the classification performance. We use the mean metric value as the final performance results for the datasets with multi-labels.

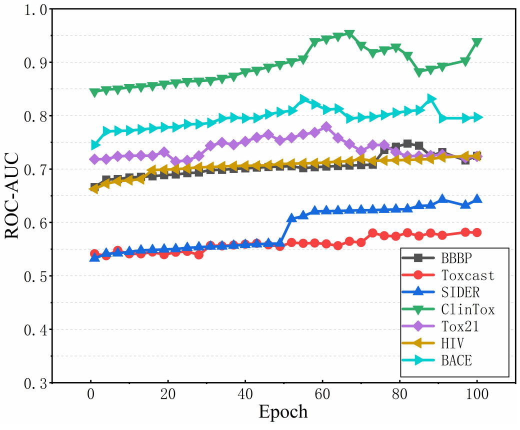

Figure 2: ROC-AUC performance of MolFuse on seven datasets at different epochs.

Implementation Details

Flowing [Ma *et al.*, 2022] and [Wang *et al.*, 2022], each model was trained for up to 100 epochs, with the training procedure halting if there was no increase in the validation ROC-AUC over 15 consecutive epochs. A 1-layer BiGRU is employed as the backbone to extract sequence features and two 5-layer graph isomorphism networks with edge features as the foundation for the graph view representation encoder. All modules undergo training using the Adam optimizer. A grid search based on the validation ROC-AUC was conducted to seek the optimal hyperparameter configuration. Cross-entropy loss was implemented as the classification loss. All simulations are implemented using PyTorch 1.7.1, and the original code of this method will be provided later.

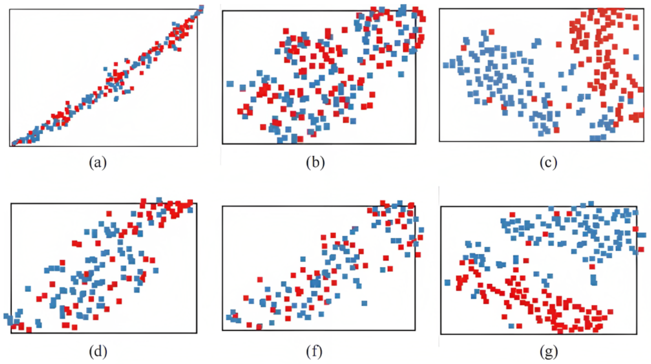

Figure 3: The t-SNE visualization. (a), (b), (c) are the molecular clusters of BBBP in sequence view, topology-spatial view, and MolFuse classified molecular clusters. (d), (f), (g) are the same clusters of BACE.

4.2 Experimental Results and Analysis

Overall Performance (RQ1)

The performance of seven molecular property prediction tasks is summarized in Table 3. Generally, it can be found from the table that MolFuse shows strong empirical performance across all seven downstream datasets, delivering six out of seven state-of-the-art results and acquiring an 2.2% absolute improvement on average. The outstanding results validate the superiority of our proposed model. We further visualize the performance changes during the model's learning process to provide an intuitive understanding in Fig. 2. We observe that as the number of training epochs increased, MolFuse's performance on the BACE, HIV, SIDER, ToxCast, and BBBP datasets first improved and then stabilized. However, for the Clintox and Tox21 datasets, the performance initially improves but subsequently begins to decline. This indicates that MolFuse requires an appropriate number of training epochs to facilitate optimal model iterations.

Molecule Clustering Analysis (RQ1)

We take the BBBP and BACE dataset as the case studies to visualize molecular embeddings at different views and stages. Fig. 3 (a) illustrates that molecules derived solely from the single view SMILES information do not have relatively indistinct separations. This suggests an absence of clear chemical information and atom-level differences. Fig. 3 (b) shows that the molecular differentiation boundaries obtained from 2D information differ from those in Fig. 3 (a), also do not have relatively indistinct separations. This disparity arises due to the inconsistencies in the latent feature space across different views. However, Fig. 3 (c) demonstrates that our MolFuse model distinctly separates different types of molecules with clearer boundaries. This indicates that the comparative fusion strategy of our model effectively deploys the consistency and complementarity across the views.

Ablation Studies (RQ2 & RQ3)

Table 4 investigates the effect of removing different modules on the model. We observe that bypassing the cross-view fusion module and simply concatenating the views produces suboptimal results. This underscores the importance of harnessing the consistency across views. Concurrently, we observe that MolFuse still outperforms baseline methods on the BBBP dataset under study, attesting to the robust capability of the dual learning module in handling view alignment. When the model is without both module, it reduces to a GNN Classifier, and its performance closely mirrors that of the GNN Classifier. Across all tasks, we note that combining the dual learning module with the cross-view fusion module yields superior representations.

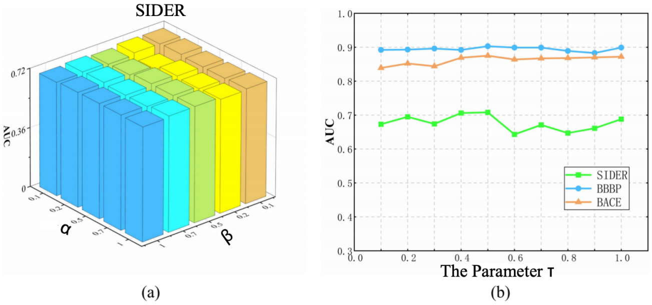

Figure 4: (a) Parameter sensitivity analysis of α and β on SIDER. (b) Parameter sensitivity analysis of τ on three datasets.

Parameter Sensitivity Analysis (RQ1)

In the parameter sensitivity analysis, we investigate two hyperparameters that balance the loss components, specifically: $\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{dua} + \beta\mathcal{L}_{con}$. Additionally, we examine the sensitivity of the temperature coefficient τ in Eqs. (15) and (16).

Fig. 4 (a) displays the average ROC-AUC across 100 independent runs. As depicted in the figure, our method exhibits robustness to the choices of α and β. Selecting both hyperparameters at 0.5 leads to optimal model performance.

The ROC-AUC initially increases with the growth of τ. However, when it surpasses 0.6, there is a noticeable performance drop on the SIDER dataset. This decline might be attributed to the increased entropy causing greater disorder and subsequently deteriorate MolFuse's performance. Overall, MolFuse exhibits relative stability.

5 Conclusion

In this paper, we introduce a novel model MolFuse that amalgamates various molecular features. Particularly, all view information is integrated into a cross-view contrastive fusion framework, where the quality of each view is fully considered. Firstly, the intricate molecular semantics and structures from both sequence and graph views are extracted. Then, we design the dual learning loss to refine low-quality views, which acquire reliable semantic information from different views. Finally, to fully explore view consistency and complementary, we conduct the fusion object on all views' features, where the mutual information between global feature and view-specific refined features is maximized. As evidenced by the results, MolFuse model produces precise chemical representations for a diverse range of molecules, ensuring accurate molecular property predictions.

Acknowledgments

This work is supported in part by Shenzhen Science and Technology Program under grants JCYJ20230807115959041 and JCYJ20230807120010021. Lifang He is partially supported by the NSF grants (MRI-2215789, IIS-1909879, IIS-2319451), NIH grant under R21EY034179, and Lehigh's grants under Accelerator and CORE.

References

[Anderson *et al.*, 1987] EricB. Anderson, GD Veith, and D Weininger. Smiles: a line notation and computerized interpreter for chemical structures. Jan 1987.

[Bemis and Murcko, 1996] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, Jan 1996.

[Chithrananda *et al.*, 2020] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv:2010.09885*, 2020.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeuralIPS Workshop on Deep Learning*, 2014.

[Fang *et al.*, 2022] Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Molecular contrastive learning with chemical element knowledge graph. In *AAAI*, volume 36, pages 3968–3976, 2022.

[Guo *et al.*, 2022] Zhihui Guo, Pramod Sharma, Andy Martinez, Liang Du, and Robin Abraham. Multilingual molecular representation learning via contrastive pre-training. In *ACL*, pages 3441–3453, 2022.

[Honda *et al.*, 2019] Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv:1911.04738*, 2019.

[Hu *et al.*, 2020] W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020.

[Kearnes *et al.*, 2016] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, page 595–608, Aug 2016.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

[Liu *et al.*, 2021] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *ICLR*, 2021.

[Ma *et al.*, 2022] Runze Ma, Yidan Zhang, Xinye Wang, Zhenyang Yu, and Lei Duan. Morn: Molecular property prediction based on textual-topological-spatial multi-view learning. In *CIKM*, pages 1461–1470, 2022.

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

[Stärk *et al.*, 2022] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *ICML*, pages 20479–20502. PMLR, 2022.

[Tosco *et al.*, 2014] Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the mmff force field to the rdkit: implementation and validation. *Journal of Cheminformatics*, Dec 2014.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Wang *et al.*, 2022] 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, 2022.

[Wu *et al.*, 2017] 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*, page 513–530, Oct 2017.

[Xiong *et al.*, 2020] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, page 8749–8760, Aug 2020.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2018.

[Yang *et al.*, 2023] Xixi Yang, Li Fu, Yafeng Deng, Yuansheng Liu, Dongsheng Cao, and Xiangxiang Zeng. Gpmo: gradient perturbation-based contrastive learning for molecule optimization. In *IJCAI*, pages 4940–4948, 2023.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *NeuralIPS*, 33:5812–5823, 2020.

[You *et al.*, 2021] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *ICML*, pages 12121–12132. PMLR, 2021.

[Zhou *et al.*, 2023] 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. 2023.

[Zhu *et al.*, 2021] Jinhua Zhu, Yingce Xia, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Dual-view molecule pre-training. *arXiv:2106.10234*, 2021.

[Zhu *et al.*, 2022] Yanqiao Zhu, Dingshuo Chen, Yuanqi Du, Yingze Wang, Qiang Liu, and Shu Wu. Featurizations matter: a multiview contrastive learning approach to molecular pretraining. In *ICML*, 2022.

[Zhuang *et al.*, 2023] Xiang Zhuang, Qiang Zhang, Bin Wu, Keyan Ding, Yin Fang, and Huajun Chen. Graph sampling-based meta-learning for molecular property prediction. *IJCAI*, 2023.