





Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction

Ao Shen ^{1,2,†}, Mingzhi Yuan ^{1,2,†}, Yingfan Ma ^{1,2}, Jie Du^{1,2}, Manning Wang ^{1,2,*}

¹Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, 131 Dong'an Road, 200032, Shanghai, China

²Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Fudan University, 131 Dong'an Road, 200032, Shanghai, China

*Corresponding author. Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China.

E-mail: mnwang@fudan.edu.cn

†Ao Shen and Mingzhi Yuan have contributed equally to this work.

Abstract

Self-supervised learning plays an important role in molecular representation learning because labeled molecular data are usually limited in many tasks, such as chemical property prediction and virtual screening. However, most existing molecular pre-training methods focus on one modality of molecular data, and the complementary information of two important modalities, SMILES and graph, is not fully explored. In this study, we propose an effective multi-modality self-supervised learning framework for molecular SMILES and graph. Specifically, SMILES data and graph data are first tokenized so that they can be processed by a unified Transformer-based backbone network, which is trained by a masked reconstruction strategy. In addition, we introduce a specialized non-overlapping masking strategy to encourage fine-grained interaction between these two modalities. Experimental results show that our framework achieves state-of-the-art performance in a series of molecular property prediction tasks, and a detailed ablation study demonstrates efficacy of the multi-modality framework and the masking strategy.

Keywords: multi-modality self-supervised learning; molecular property prediction; molecular representations

Introduction

Accurate prediction of molecular properties is the basis for compound screening [1] and accelerates the drug discovery process [2]. Efficient molecular representation is a priority for predicting molecular properties [3]. Recently, molecular representation learning based on deep learning has been booming [4, 5]. However, since most molecular label data need to be obtained through labor-intensive and costly wet experiments [6], there is a lack of sufficient labeled molecular data, which hinders the development of deep learning methods and can lead to issues like overfitting and poor generalization [7, 8]. Self-supervised learning holds substantial research value in addressing these challenges, which involves pre-training on unlabeled data and fine-tuning with labeled data on downstream tasks. It has shown significant promise in enhancing the performance of molecular representation learning on property prediction tasks [9].

Molecules can be described using various modalities, such as fingerprints, sequences, graphs and more [10–12]. Our work mainly focuses on two widely used modalities: Simplified Molecular-Input Line-Entry system (SMILES) [13] and molecular graph. As depicted in Fig. 1, the same molecule can be represented

using both a SMILES sequence and a graph, with each modality having its unique advantages and disadvantages. SMILES is a compact implicit representation of the molecule that excludes single-bond representation, making it well suited for rapid compound retrieval and identification [14]. Additionally, the SMILES sequence, being a text string, can be processed with Transformer-based networks well developed in the natural language processing (NLP) field for feature extraction, in which the self-attention mechanism weighs and combines information from any position in the input sequence, thereby facilitating the capture of global contextual information [11, 12]. However, SMILES representations only capture the relationships between atoms and bonds. They often struggle to capture the complex structural and topological information of molecules, such as the number and positions of rings, the length of side chains and other intricate details that can be crucial in drug efficacy prediction [15, 16]. Graph representations offer explicit portrayals of atoms, bonds and their interconnections, showcasing the topological structures of molecules [17]. They provide detailed chemical information about molecules, including attributes for each atom such as element type, charge state and stereochemistry, and

Ao Shen is a PhD candidate at Fudan University. Her research interests are AI4Science and drug screening.

Mingzhi Yuan is a PhD candidate at Fudan University. His research interests are 3D computer vision, machine learning, and AI4Science.

Yingfan Ma is a Master degree candidate at Fudan University. Her research interests are computer vision and AI4Science.

Jie Du is a Master degree candidate at Fudan University. His research interests are artificial intelligence and molecular design.

Manning Wang is a professor at Fudan University. His research interests are 3D computer vision, deep learning, and AI4Science.

Received: December 26, 2023. Revised: April 25, 2024. Accepted: May 15, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

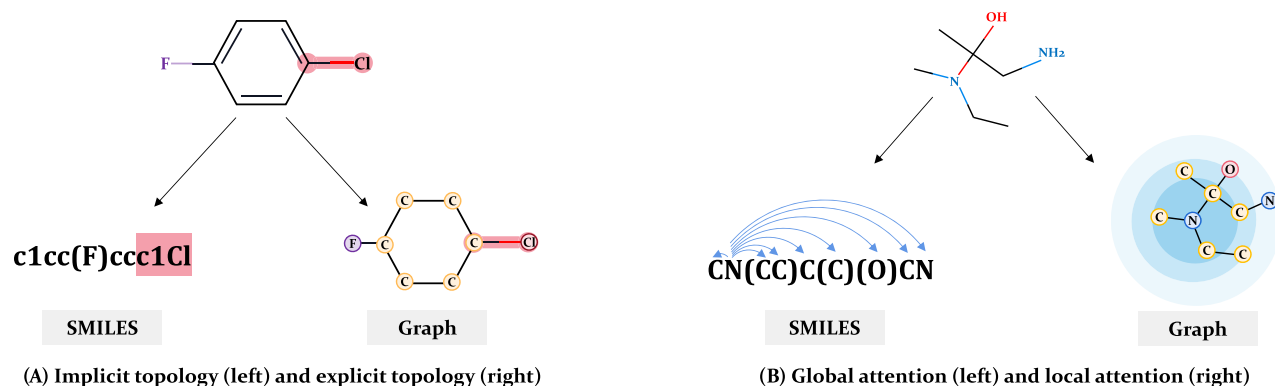


Figure 1. Comparison of two molecular representation modalities, SMILES and graph. (A) Illustration of the topological differences between the two modalities. SMILES represents topology implicitly, while graph displays explicit topology. (B) Difference in receptive field of networks for the two modalities. Global attention is usually used for SMILES, while local attention can be easily implemented for graph.

attributes for each bond, such as bond type and bond length [18]. However, graph neural networks, commonly used to extract features from graphs, primarily rely on message-passing layers to gather information from neighboring nodes, emphasizing the capture of local contextual information. This can lead to a disadvantage in capturing global context information due to information decay when delivering messages between non-adjacent nodes [19]. As a result, for the same molecule, SMILES and graph encode molecular features from different perspectives, offering complementary information. The rational combination of these two modalities holds promise for enhancing molecular representation performance.

There are several existing works on multi-modality molecular pre-training [20–23]. For example, GraphMVP [20] focuses on joint pre-training with 2D graphs and 3D graphs. However, these two modalities exhibit high similarity. Additionally, this study only proved that 3D geometry complements 2D topology in downstream tasks, without proving that 2D topology complements 3D geometry. MoleculeSTM [23] focuses on molecular graphs and text descriptions, using a contrastive learning strategy to learn the consistency between the chemical structure of molecules and their textual descriptions. MOCO [24] and DVMP [21] extract features from SMILES and graph of the same molecule through two specialized encoders and then utilize contrastive learning to minimize the feature distance between different modalities of the same molecule. All these contrastive learning-based methods lack fine-grained cross-modality interactions, resulting in suboptimal performance. Therefore, to achieve better performance, the challenge of more efficiently combining these two modalities with significant differences lies in how to promote information exchange in fine-grain such as at the atom level rather than only achieving contrastive learning at the entire molecule level.

To achieve information interactions in fine-grain, it is necessary to enable the network to understand the fine-grained data in different modalities simultaneously. Recently, a series of Transformer-based methods [25–28] provide a new way to understand multi-modality data. Specifically, they utilize special encoders to map multi-modality data into a universal embedding space, where images/videos can be viewed as foreign languages for language models so that language models understand them. This inspires us to encode multi-modality data into a unified pattern so that a unified Transformer-based network can learn interaction features of different modalities. Specifically, we treat words in SMILES sequences and graph nodes as tokens [29, 30],

and put them into a unified network to perform fine-grained feature interaction between the two modalities. Furthermore, to better learn complementary information, we need to develop strategies to enhance the interaction between the two modalities in pre-training. Inspired by the ‘cloze-type’ generative pre-training task in NLP, which restores the masked part of the information by learning the relationship between the features of the unmasked information, designing an appropriate masking strategy can promote the pre-training network to learn richer features. Intuitively, the information used for reconstructing the masked tokens can come from the context within the same modality, as well as information from the tokens of corresponding structures in the other modality. Therefore, we establish fine-grained correspondences between the SMILES and the graph of a molecule and mask non-overlapping parts in these two modalities to encourage the model to reconstruct the masked part of one modality with the direct information of the corresponding part of the other modality, which strengthens the interactions between the two modalities.

In this paper, we propose MoleSG, a simple yet effective pre-training framework for effectively exploring the complementary information between SMILES and graph in molecular pre-training. Our framework consists of two independent encoders to separately convert masked SMILES and masked graph of an input molecule into token embeddings. Then, we introduce a Transformer-based unified backbone network for jointly processing embeddings from both modalities to facilitate interactions between them. The embeddings from the two modalities are concatenated and inputted into the universal Transformer for joint processing and the output is used to reconstruct the original SMILES and graph by two specific decoders. Our framework is trained by reconstruction losses. Furthermore, we introduce a dedicated non-overlapping masking strategy, in which we establish the atom index correspondence between the SMILES sequence and the graph of a molecule to ensure that regions masked in SMILES and graph do not overlap. To evaluate the effectiveness of MoleSG, we conduct experiments on 14 downstream tasks related to molecular property prediction and MoleSG achieves state-of-the-art (SOTA) performance in all tasks. We also compare it with the same network pre-trained by a single modality, and the experimental results show that multi-modality training learns richer molecular representation knowledge.

Our contributions are as follows: (1) we propose MoleSG, a novel molecular pre-training framework that utilizes the complementary information of SMILES and graph representations, resulting

in improved performance; (2) we introduce an innovative non-overlapping masking strategy and a unified network for handling two distinct modalities, allowing for fine-grained interaction between SMILES and graph representations and achieving better representation learning; (3) MoleSG achieves SOTA performance in a series of molecular property prediction tasks, and detailed ablation study demonstrates efficacy of the multi-modality structure and the masking strategy.

Related work

Molecular representation learning: In recent years, with the development of deep learning, some learning-based molecular representation learning methods have been proposed and achieved great progress, but these learning-based methods are often limited by the availability of labeled data. To solve this problem, a series of pre-training methods [31, 32] have been proposed that can utilize unlabeled data. However, most methods only use single modality data representation such as SMILES [11, 12, 33, 34], but neglect the complementary information between the different modalities, thereby achieving suboptimal performance. As analyzed in Fig. 1, multi-modality pre-training contains more information and tends to achieve better performance, thereby having greater potential.

Molecular multi-modality self-supervised learning: As shown in Fig. 1, the rational combination of SMILES and graph holds promise for enhancing molecular representation performance. Most existing approaches often rely on the contrastive method, such as SMICLR [35], DVMP [21] and MOCO [24], which focus on the same two modalities as we do but they neglect the fine-grained interactions across different modalities. A concurrent work, UniMAP [36], is a generative pre-training based on mask reconstruction, but it only performs simple mask reconstruction without a specific design of masking strategy, so it still cannot fully leverage the complementary information interactions. We introduce a non-overlapping masking strategy to force cross-modality information interaction, thereby having a greater advantage.

Materials and methods

Framework of MoleSG

As shown in Fig. 2, MoleSG learns features jointly from SMILES and graph by performing masked reconstruction on both modalities with a unified feature extraction backbone network. Concretely, for a given molecule, we first convert its SMILES sequence into tokens T_S and calculate features V_G and E_G for nodes and edges in the graph. During pre-training, we randomly mask some node features V_G^M in the graph and then mask a portion of SMILES tokens T_S^M corresponding to the remaining unmasked atoms in the graph, so that we can perform non-overlapping masking to facilitate the interaction of information between the two modalities.

During pre-training, we employ a symmetric joint encoder-decoder framework to perform further feature extraction. The framework consists of two independent branches for the two modalities and a shared backbone for feature fusion. The independent encoder branches encode the data of two different modalities into a unified form i.e. embedding, which is suitable for understanding by a Transformer-based backbone [29, 30]. The shared Transformer-based backbone can learn the dependencies between atoms within and across the modalities and output features for the subsequent independent decoders. Finally, the

SMILES Decoder and the Graph Decoder reconstruct the original SMILES sequence and graph based on the output of the backbone. During fine-tuning, we utilize the pre-trained Graph Encoder as a molecular representation network and add corresponding output heads to predict a series of molecular properties.

Encoder

To facilitate the interaction of fine-grained features across different modalities, we use two independent encoders to convert the data of two entirely different modalities into embeddings of the same dimensions for being further processed by Transformer model.

For the SMILES sequence, we follow ChemBERTa [11] to first convert the masked SMILES tokens T_S^M into a sequence of token ids ID_S^M , and we expand its vocabulary by conducting a comprehensive analysis of all tokens in our dataset, as detailed in Supplementary D. Then, we calculate their corresponding embeddings $F_S \in \mathbb{R}^{N_S \times d}$ by a Transformer model with a series of multi-head attention blocks used in Roberta [37], where N_S represents the number of SMILES tokens, and d is the feature dimension.

For the graph, we feed V_G^M and E_G into the Graph Encoder. We implement CoMPT [38] as our Graph Encoder, which strengthens the message interactions between nodes and edges through a communicative kernel. After the Graph Encoder processing, we obtain the token embeddings $F_G \in \mathbb{R}^{N_G \times d}$ for nodes, where N_G is the number of atoms, and d is the feature dimension.

Unified backbone

We design a unified backbone based on Transformer to promote feature interaction between the two modalities. Through the attention mechanism of Transformer, we enable the model to learn the correlation between different input token embeddings across two modalities. After the processing of the two modality-specific encoders, we add trainable parameters $A_S \in \mathbb{R}^{N_S \times d}$ and $A_G \in \mathbb{R}^{N_G \times d}$, respectively, to $F_S \in \mathbb{R}^{N_S \times d}$ and $F_G \in \mathbb{R}^{N_G \times d}$ and concatenate them. We add learnable parameters just to help the unified backbone to distinguish modalities. Then, the concatenated embeddings $F_{S,G} \in \mathbb{R}^{(N_S+N_G) \times d}$ are then fed into the backbone network. Here, we use the Transformer model employed in Roberta [37] as the backbone network with a series multi-head self-attention blocks. The self-attention mechanism is formulated as

$$\text{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (1)$$

where Q, K and V are the projection vectors of $F_{S,G}$, $Q, K, V = F_{S,G} \times W_Q, F_{S,G} \times W_K, F_{S,G} \times W_V$. Through multi-head self-attention mechanism, we can facilitate information interaction between token embeddings both within the same modalities and across different modalities. Finally, the unified backbone outputs the embeddings $F'_{S,G} \in \mathbb{R}^{(N_S+N_G) \times d}$.

Decoder

After feature extraction in the backbone, we split the output features $F'_{S,G} \in \mathbb{R}^{(N_S+N_G) \times d}$ into features $F'_S \in \mathbb{R}^{N_S \times d}$ for SMILES and features $F'_G \in \mathbb{R}^{N_G \times d}$ for graph. F'_S and F'_G are used for modality-specific mask reconstruction tasks. Specifically, F'_S is fed into the SMILES Decoder, which is the LMhead in Roberta [37], to predict the masked token IDs, while F'_G is inputted into the Graph Decoder, which is a lightweight network GIN [39], after re-masking [40] to reconstruct the masked node features. We calculate the entropy

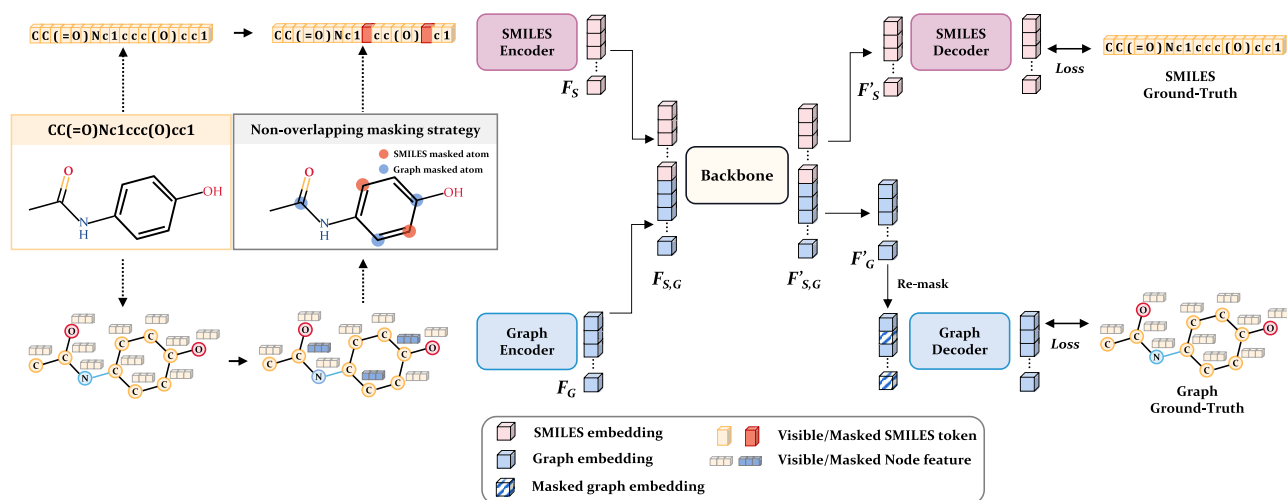


Figure 2. Overview of MoleSG. The SMILES sequence and the graph of a molecule are first randomly masked using the non-overlapping masking strategy. Then they are individually encoded by independent encoders, and the SMILES embeddings and the graph embeddings are concatenated and inputted into a Transformer-based backbone for joint processing. Finally, processed features belonging to each modality are decoded into token IDs and graph nodes for the reconstruction proxy task.

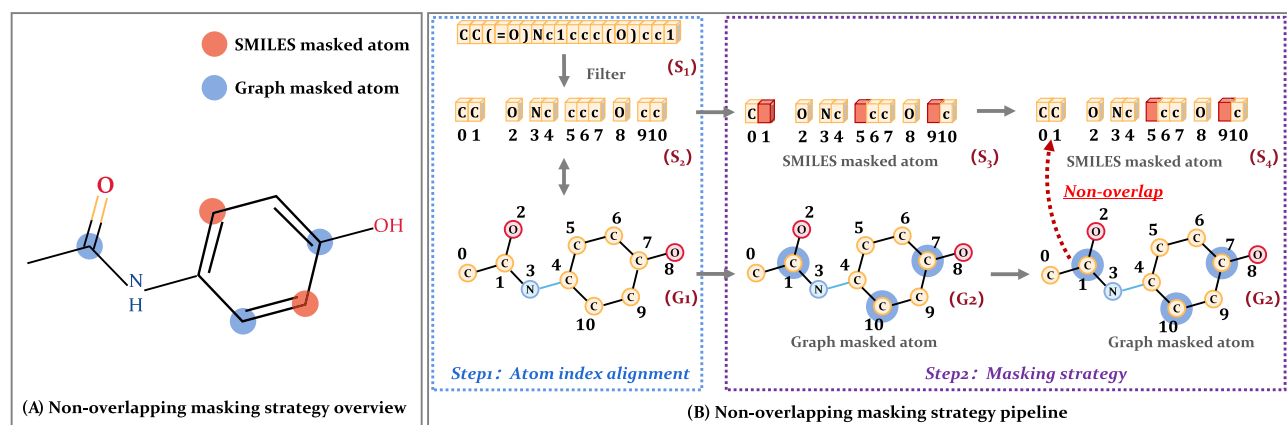


Figure 3. Illustration of non-overlapping masking strategy. (A) Non-overlapping masking strategy: masks in the SMILES sequence and the graph for the same molecule do not overlap. (B) Non-overlapping masking strategy pipeline: first, we establish a correspondence between atom index in both modalities. Then, random masking is applied to the graph, followed by mapping the masked atoms from the graph to the SMILES sequence. Finally, random masking on the SMILES sequence is implemented on the remaining unmasked atoms of the graph.

loss \mathcal{L}_{EN} [37] in SMILES reconstruction and the SCE loss \mathcal{L}_{SCE} [40] in graph reconstruction. The overall loss for the entire task is as follows: $\mathcal{L}_{Total} = \mathcal{L}_{EN} + \mathcal{L}_{SCE}$.

Non-overlapping masking strategy

The non-overlapping masking strategy we propose is illustrated in Fig. 3, which can be divided into two steps, first performing atom index alignment between the two modalities, and then performing non-overlapping masking.

Step 1: Atom index alignment. The SMILES tokens can be categorized into three classes: (1) Atoms, including single-character atoms like C and N, as well as multi-character atoms like Ca and Au, and ions like [Cl-] and [Fe+3]; (2) Chemical bonds, represented by symbols like '#' and '='; (3) Other symbols, such as numbers '1' and '2' indicating the positions of atoms in a ring and parentheses '(' and ')' denoting containing side chains. Given that single bonds are often omitted in SMILES, achieving a one-to-one correspondence between two modalities for chemical bonds is not practical. Therefore, in this paper, we focus on aligning the atom index. Thus, we gather the tokens representing the atoms and assign indexes to them to establish a consistent correspondence

between atoms in graph and those in filtered SMILES tokens, as shown in Fig. 3.

Step 2: Masking strategy. We randomly mask atom features on the graph and atom tokens on the SMILES sequence, where the sets of masked atom indexes on them are denoted as I_G and I_S , respectively. To encourage better interaction between the two modalities, we set the overlap ratio between masked atoms in both modalities to 0. Specifically, based on the one-to-one correspondence of atom index, we localize the positions of the masked atoms in graph onto the SMILES sequence. Through operation $P: I_S - I_G \cap I_S$, we avoid masking atoms on the SMILES sequence that are already masked on the graph.

Fine-tuning

We conduct fine-tuning on the pre-trained Graph Encoder on 14 downstream tasks of predicting molecular properties. Since previous works only utilize a single modality in the downstream tasks, we also take a single modality as input to achieve a fair comparison. We also analyze combining two modalities encodes during fine-tuning; more details can be obtained in Supplementary F. The backbone is not utilized in downstream

Table 1. Performance of different models on eight classification benchmarks in physiology and biophysics. The mean and standard deviation of ROC-AUC (%) from three independent runs are reported (higher values indicate better performance)

Category	Physiology					Biophysics		
Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV
Molecules	2039	7831	8575	1427	1478	1513	93087	41127
Tasks	1	12	617	27	2	1	17	1
MPNN[41]	91.3±4.1	80.8±2.4	69.1±3.0	59.5±3.0	87.9±5.4	81.5±1.0	75.7±1.3	77.0±1.4
DMPNN[42]	91.9±3.0	75.9±0.7	63.7±0.2	57.0±0.7	90.6±0.6	85.2±0.6	78.6±1.4	77.1±0.5
CMPNN[43]	92.7±1.7	80.1±1.6	70.8±1.3	61.6±0.3	89.8±0.8	86.7±0.2	79.0±2.0	78.2±2.2
CoMPT[38]	96.1±0.4	84.5±0.7	72.2±0.8	66.1±0.9	97.3±2.5	94.1±3.6	82.6±1.6	86.4±1.2
GraSeq[44]	92.8±1.8	78.3±1.1	70.3±1.1	65.9±2.7	82.5±2.5	89.4±2.7	72.2±2.7	83.3±2.4
N-Gram[45]	91.2±0.3	76.9±2.7	-	63.2±0.5	87.5±2.7	79.1±1.3	76.9±0.7	78.7±0.4
PretrainGNN[46]	70.8±1.5	78.7±0.4	65.7±0.6	62.7±0.8	72.6±1.5	84.5±0.7	81.3±2.1	79.9±0.7
MGSSL[47]	70.5±1.1	76.4±0.4	64.1±0.7	61.8±0.8	80.7±2.1	79.7±0.8	78.7±1.5	79.5±1.1
GEM[48]	88.8±0.4	78.1±0.4	68.6±0.2	63.2±1.5	90.3±0.7	87.9±1.1	75.3±1.5	81.3±0.3
GROVER[7]	86.8±2.2	80.3±2.0	56.8±3.4	61.2±2.5	70.3±13.7	82.4±3.6	67.3±1.8	68.2±1.1
GraphMVP[20]	72.4±1.6	75.9±0.5	63.1±0.4	63.9±1.2	79.1±2.8	81.2±0.9	77.7±0.6	77.0±1.2
Mole-Bert[49]	92.9±1.9	84.5±3.5	71.5±0.5	63.4±2.3	74.7±10.0	92.6±1.9	84.7±1.1	86.2±1.0
SMICLR[35]	95.7±2.9	82.5±2.6	68.7±1.6	61.3±2.1	77.4±21.5	87.4±3.0	75.2±3.1	78.1±3.9
DVMP[21]	77.8±0.3	79.1±0.4	-	69.8±0.6	95.6±0.7	89.4±0.8	-	81.4±0.4
DVMP _{MoleSG}	80.9±2.1	84.4±1.2	73.3±0.9	66.9±1.2	98.4±2.0	93.5±2.8	80.9±2.1	87.6±1.8
MolCLR[22]	73.3±1.0	74.1±5.3	65.9±2.1	61.2±3.6	89.8±2.7	82.8±0.7	78.9±2.3	77.4±0.6
MolCLR _{CoMPT}	97.2±0.2	82.4±1.8	72.7±0.5	57.1±8.7	77.0±14.5	85.5±0.9	75.8±15.0	81.8±2.2
KANO[50]	96.0±1.6	83.7±1.3	73.2±1.6	65.2±0.8	94.4±0.3	93.1±2.1	83.7±2.3	85.1±2.2
MoleSG	97.9±0.3	85.0±1.2	74.2±0.5	70.0±0.2	99.1±0.9	95.1±2.1	85.1±0.8	87.7±1.9

tasks because it is pre-trained by both modalities and is not suitable for single-modality inputs in downstream tasks.

Datasets setup

During the pre-training stage, we sample 250 000 unlabeled molecules from ZINC15 [51], which is a comprehensive collection of chemical compounds for drug discovery and computational chemistry research. During the fine-tuning stage, we utilize 14 benchmark datasets from MoleculeNet [52], covering molecular data from various domains, including pharmaceuticals, biology, chemistry and physics. These downstream datasets include 678 binary classification tasks and 19 regression tasks. For more detailed information about benchmark datasets, please refer to Supplementary A.

We partition each benchmark dataset into the train, validation and test sets in an 8:1:1 ratio. For all datasets except QM9, we employ scaffold splitting, reporting the mean and standard deviation of results from three random seeds for each benchmark, and we list seed in Supplementary B. Scaffold splitting is a more challenging and realistic data partitioning method [53]. For the QM9 dataset, we follow the approach used in most prior work [22, 50] for random splitting.

Baselines and training details

We compare MoleSG with both supervised (training from scratch) baselines and pre-trained baselines. Supervised methods include MPNN [41], DMPNN [42], CMPNN [43], CoMPT [38] and GreSeq [44]. Pre-training methods include N-gram [45], PretrainGNN [46], MGSSL [47], GROVER [7], GraphMVP [20], MolCLR [22], GEM [48], DVMP [21], KANO [50], Mole-Bert [49] and SMICLR [35]. The specific configurations for these competitors can be found in Supplementary C. Additionally, for a fair comparison, we implement new MolCLR and DVMP by replacing the original encoders in them with the same networks we use, which are

denoted as MolCLR_{CoMPT} and DVMP_{MoleSG}. We also utilize our non-overlapping masking strategy in DVMP_{MoleSG}.

We train MoleSG for 90k iterations using the AdamW optimizer with a base learning ratio of 1e-3. We set the mask ratio for graph at 25% and for SMILES at 15%. The details of the mask ratio setting experiments for the two modalities are shown in Section 4.3. We set a maximum of 150 training epochs, with early stopping applied when the validation set's best value is not improved for more than 20 epochs. We use the AdamW optimizer with a base learning rate of 1e-3 and different warmup factors for different benchmarks. More details of experimental settings can be obtained in Supplementary B.

Results and discussion

MoleSG boost the performance of property prediction

In Table 1 and Table 2, the results of MPNN, DMPNN, CMPNN, N-gram, PretrainGNN, MGSSL, GROVER, GraphMVP, MolCLR, GEM and KANO are taken from the paper of KANO [50], while the results of DVMP is obtained from the original text of its original article [21]. As Mole-Bert [49] uses a different data split setting with KANO, we rerun it with the same data split setting as other baselines. Since the number of experimental repetitions of CoMPT [38] is different, we also rerun it using our experimental settings. For two multi-modality methods for SMILES and graph, GraSeq [44] and SMICLR [35], we fully tune them and achieve their best performance for comparison based on their original codes using our experimental settings.

Table 1 presents the test results in classification tasks. It can be observed that MoleSG consistently outperforms other methods across all eight datasets, demonstrating its effectiveness. It is worth noting that though the Toxcast dataset benchmark with 617 binary classification tasks is challenging, our method still

Table 2. Performance of different models on six regression benchmarks in physical chemistry and quantum mechanics. The mean and standard deviation of root mean square error (RMSE) (for ESOL, FreeSolv and Lipophilicity) or mean absolute error (MAE) (for QM7, QM8 and QM9) from three independent runs are reported (lower values indicate better performance)

Category	Physical chemistry			Quantum mechanics		
Dataset	ESOL	FreeSolv	Lipophilicity	QM7	QM8	QM9
Molecules	1128	642	4200	6830	21786	133885
Tasks	1	1	1	1	12	3
MPNN[41]	1.167±0.043	1.621±0.952	0.672±0.051	111.4±0.9	0.0148±0.001	0.00522±0.00003
DMPNN[42]	1.050±0.008	1.673±0.082	0.683±0.016	103.5±8.6	0.0156±0.001	0.00514±0.00001
CMPT[43]	0.798±0.112	1.570±0.442	0.614±0.029	75.1±3.1	0.0153±0.002	0.00405±0.00002
CoMPT[38]	0.643±0.051	0.970±0.207	0.572±0.058	32.7±7.4	0.0120±0.001	0.00353±0.00067
GraSeq[44]	0.770±0.035	1.266±0.131	0.834±0.009	101.8±2.7	0.0190±0.002	0.01890±0.00010
N-Gram[45]	1.100±0.030	2.510±0.191	0.880±0.121	125.6±1.5	0.0320±0.003	0.00964±0.00031
PretrainGNN[46]	1.100±0.006	2.764±0.002	0.739±0.003	113.2±0.6	0.0215±0.001	0.00922±0.00004
GEM[48]	0.813±0.028	1.748±0.114	0.674±0.022	60.0±2.7	0.0163±0.001	0.00562±0.00007
GROVER[7]	1.423±0.288	2.947±0.615	0.823±0.010	91.3±1.9	0.0182±0.001	0.00719±0.00208
SMICLR[35]	0.883±0.193	1.345±0.132	0.861±0.032	37.5±7.2	0.0164±0.001	0.00560±0.00020
DVMP[21]	0.817±0.024	1.952±0.061	0.653±0.002	74.4±1.2	0.0171±0.004	-
DVMP _{MoleSG}	0.669±0.114	0.942±0.110	0.594±0.018	30.2±3.0	0.0123±0.001	0.00323±0.00006
MolCLR[22]	1.113±0.023	2.301±0.247	0.789±0.009	90.9±1.7	0.0185±0.013	0.00480±0.00003
MolCLR _{CoMPT}	0.849±0.062	1.135±0.163	0.657±0.012	32.7±2.8	0.0141±0.001	0.00350±0.00000
KANO[50]	0.670±0.019	1.142±0.258	0.566±0.007	56.4±2.8	0.0123±0.000	0.00320±0.00001
MoleSG	0.599±0.067	0.932±0.131	0.545±0.014	29.6±2.9	0.0117±0.001	0.00313±0.00006

Table 3. Comparison of our approach with two single-modality pre-training approaches on classification tasks. The mean and standard deviation of ROC-AUC (%) over three independent runs are reported (higher values indicate better performance)

	BBBP	Tox21	ToxCast	SIDER	Clintox	BACE	MUV	HIV
SMILES scratch	63.6±4.3	75.5±0.5	64.2±2.5	54.0±2.4	88.1±6.3	79.2±6.6	63.6±4.3	72.7±3.5
SMILES pre-train	61.5±4.9	77.6±2.5	66.8±0.9	55.0±3.1	93.3±2.8	83.8±0.9	61.5±4.9	75.1±2.5
Ours SMILES	65.3±3.1	77.9±2.5	67.0±0.9	59.6±3.8	94.3±2.0	85.3±1.1	65.3±3.1	77.3±0.7
Graph scratch	96.1±0.4	84.5±0.7	72.2±0.8	66.1±0.9	97.3±2.5	94.1±3.6	82.6±1.6	86.4±1.2
Graph pre-train	96.8±1.8	84.2±0.1	72.6±1.0	66.7±2.2	98.0±0.9	94.9±2.3	82.2±1.4	85.9±2.5
Ours graph	97.9±0.3	85.0±1.2	74.2±0.5	70.0±0.2	99.1±0.9	95.1±2.1	85.1±0.8	87.7±1.9

performs better than the current SOTA method KANO. Complementary information of the two modalities in MoleSG contributes to outstanding results, surpassing methods injecting additional 3D information.

Table 2 shows the test results in regression tasks. We can observe that MoleSG achieves the best scores among both supervised and self-supervised pre-training models, with a relative improvement of 14.4% over KANO across all six regression tasks. MoleSG greatly benefits tasks with limited label information, achieving an 18.4% improvement over KANO on the small dataset FreeSolv, which contains only 642 labeled molecules.

Moreover, it is worth noting that our proposed method still outperforms MolCLR_{CoMPT}, which is a version of the typical contrastive learning method MolCLR with the same encoder as ours, verifying the superiority of our method. We also compare with another contrastive learning competitor DVMP_{MoleSG}, which utilizes the same encoders as ours. In addition, both MolCLR_{CoMPT} and DVMP_{MoleSG} outperform their original counterpart MolCLR and DVMP in most tasks, demonstrating the effectiveness of the corresponding strategies proposed in this paper.

MoleSG outperforms single-modality pre-training only

To further reveal the superiority of our multi-modality method, we compare our multi-modality pre-training with single-modality

pre-training and the results are shown in Table 3 and Table 4. In this experiment, an output head is added to each encoder. 'SMILES scratch' and 'Graph scratch' represent the two networks trained from scratch. The initial weights of encoders in 'SMILES pre-train' and 'Graph pre-train' are obtained from single-modality pre-training using the same MoleSG framework while blocking the other modality. The initial weights of encoders in 'Ours SMILES' and 'Ours graph' are obtained from the corresponding encoders of the multi-modality pre-trained MoleSG. From these results, we can observe that our proposed method achieves the best performance on all downstream tasks. Moreover, it is worth noting that single modality pre-training may cause performance degradation. However, by fully leveraging the complementary information among different modalities, our method can improve performance on all downstream tasks, showing more potential for practical applications.

We present visualization results of our method's feature extraction capability in Fig. 4, which illustrates the strong feature discriminative ability of MoleSG in the classification tasks BBBP and BACE. We compare our proposed model with models trained from scratch (without pre-training), from single-modality pre-training (i.e. graph pre-training), and from contrastive pre-training (DVMP_{MoleSG}). During fine-tuning, these competitors all utilize the Graph Encoder. From Fig. 4, we can observe the superior feature discrimination of our approach compared with single-modality

Table 4. Comparison of our approach with two single-modality pre-training approaches on regression tasks. The mean and standard deviation of RMSE or MAE over three independent runs are reported (lower values indicate better performance)

	ESOL	Freesolv	Lipophilicity	QM7	QM8	QM9
SMILES scratch	0.946 \pm 0.226	2.581 \pm 0.286	1.028 \pm 0.030	160.2 \pm 6.8	0.0146 \pm 0.001	0.01017 \pm 0.00045
SMILES pre-train	1.030 \pm 0.336	1.942 \pm 0.450	1.034 \pm 0.015	159.3 \pm 5.7	0.0141 \pm 0.001	0.01080 \pm 0.00010
Ours SMILES	0.873\pm0.172	1.889\pm0.590	0.964\pm0.036	155.7\pm3.9	0.0139\pm0.001	0.00973\pm0.00059
Graph scratch	0.643 \pm 0.051	0.970 \pm 0.207	0.572 \pm 0.058	32.7 \pm 7.4	0.0120 \pm 0.001	0.00353 \pm 0.00067
Graph pre-train	0.635 \pm 0.104	0.939 \pm 0.225	0.585 \pm 0.031	32.3 \pm 1.6	0.0118 \pm 0.001	0.00323 \pm 0.00012
Ours graph	0.599\pm0.067	0.932\pm0.131	0.545\pm0.014	29.6\pm2.9	0.0117\pm0.001	0.00313\pm0.00006

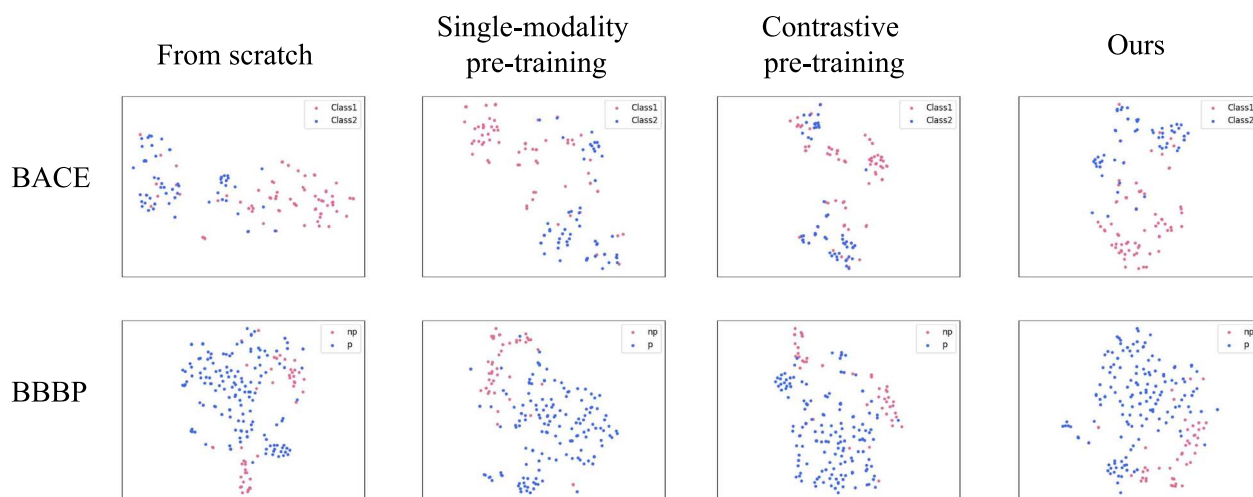


Figure 4. T-SNE visualization of feature separation of four methods on BACE and BBBP benchmark datasets.

pre-training and contrastive pre-training. Learning efficiency analysis can be seen in Supplementary G.

Mask ratio setup

To determine the mask ratio for graph and SMILES modalities, we use a controlled variable approach. We adjust the graph mask ratio while keeping the SMILES mask ratio constant, and vice versa. We conduct experiments across all benchmarks, and the experimental results for the SMILES mask ratio and graph mask ratio are shown in Fig. 5 and Fig. 6, respectively. We observe that a SMILES mask ratio of 15% and a graph mask ratio of 25% are suitable for our purposes. (In classification tasks, a higher ROC-AUC(%) value indicates better performance, while in regression tasks, lower RMSE and MAE values are desirable.)

Ablation experiments

Overlap versus non-overlap

To validate whether our non-overlapping masking strategy benefits pre-training, we conduct experiments on different overlap ratios on all downstream tasks. We define overlap ratio as a metric measuring the proportion of jointly masked atoms in both modality inputs. We conduct experiments at overlap ratios at 0%, 25%, 50%, 75% and 100% across all benchmarks, where our non-overlapping masking strategy is equivalent to setting the overlap ratio to 0. The experimental results shown in Fig. 7 indicate that the performance on downstream tasks is the best when the overlap ratio is 0.

With versus without backbone

As analyzed in Section 3.6, fine-tuning both the encoder and backbone may cause suboptimal performance due to inconsistent

distributions. Therefore, we conduct an experiment to validate it. Specifically, Table 3 and Table 4 show that the performance of Graph Encoder is better than SMILES Encoder. Therefore, we only consider two combinations in this section. The former is fine-tuning a single Graph Encoder, and the other is fine-tuning both the Graph Encoder and the backbone. We perform experiments on all benchmarks, and the results are shown in Table 5 and Table 6. The results show that using only the Graph Encoder achieves higher performance in all tasks.

Conclusion

In this study, we address the challenges of learning fine-grained information from two complementary modalities: SMILES and graph. To better capture rich molecular features from the interaction between these two modalities, we design a simple and efficient multi-modality pre-training framework called MoleSG, which utilizes a unified feature processing network to fuse both modalities. In addition, we propose a non-overlapping masking strategy to facilitate information exchange between the two modalities. Extensive experiments on 14 downstream tasks show that our method achieves new SOTA performance. Our non-overlapping masking strategy has the potential to be used in other masked reconstruction-based multi-modality pre-training studies.

There are two potential directions for future work. (1) Our multi-modality pre-training method can be utilized in the protein representation learning, because proteins also have both sequence and graph representations. (2) Our non-overlapping masking strategies can be extended to other joint pre-training studies of multiple-modality data.

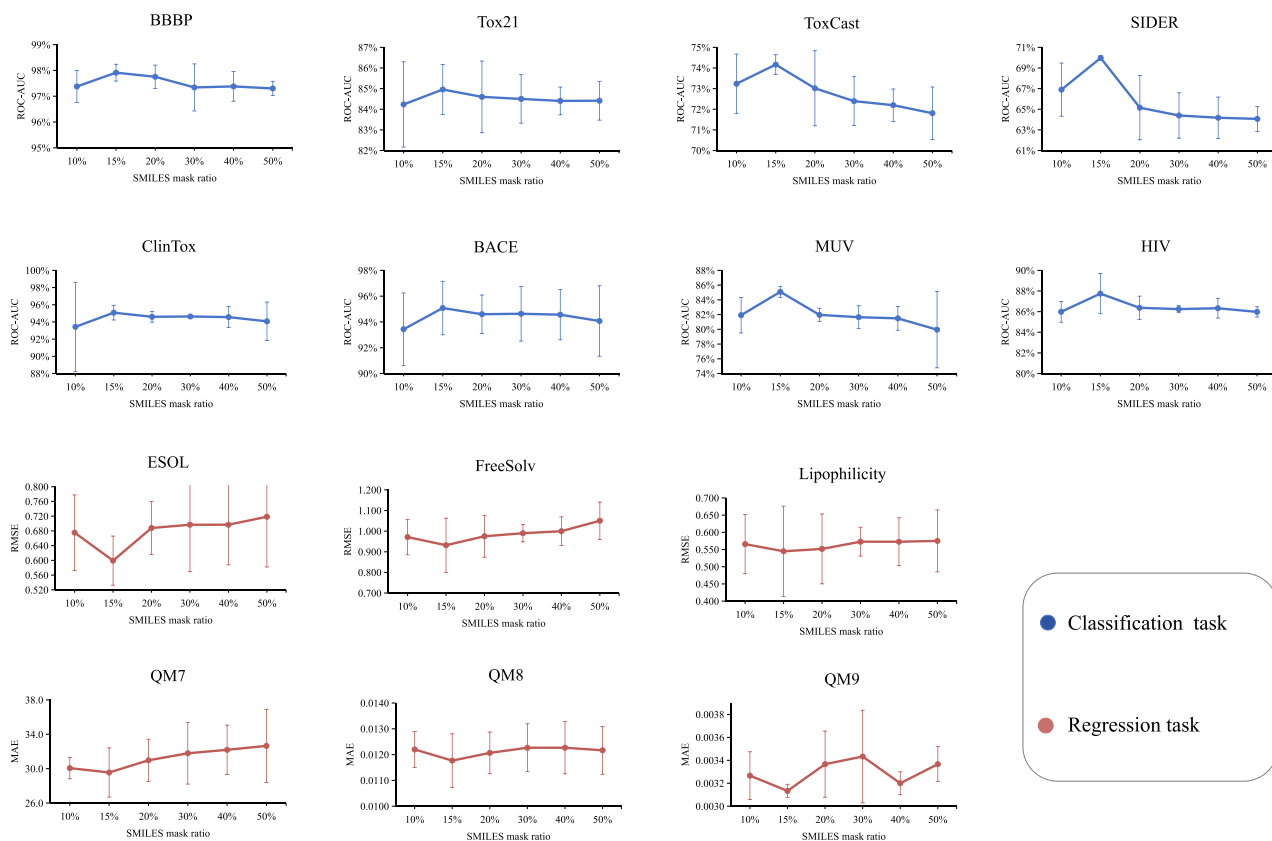


Figure 5. The impact of different mask ratios on downstream task performance on SMILES. The results are reported as mean and standard deviation values on three independent runs.

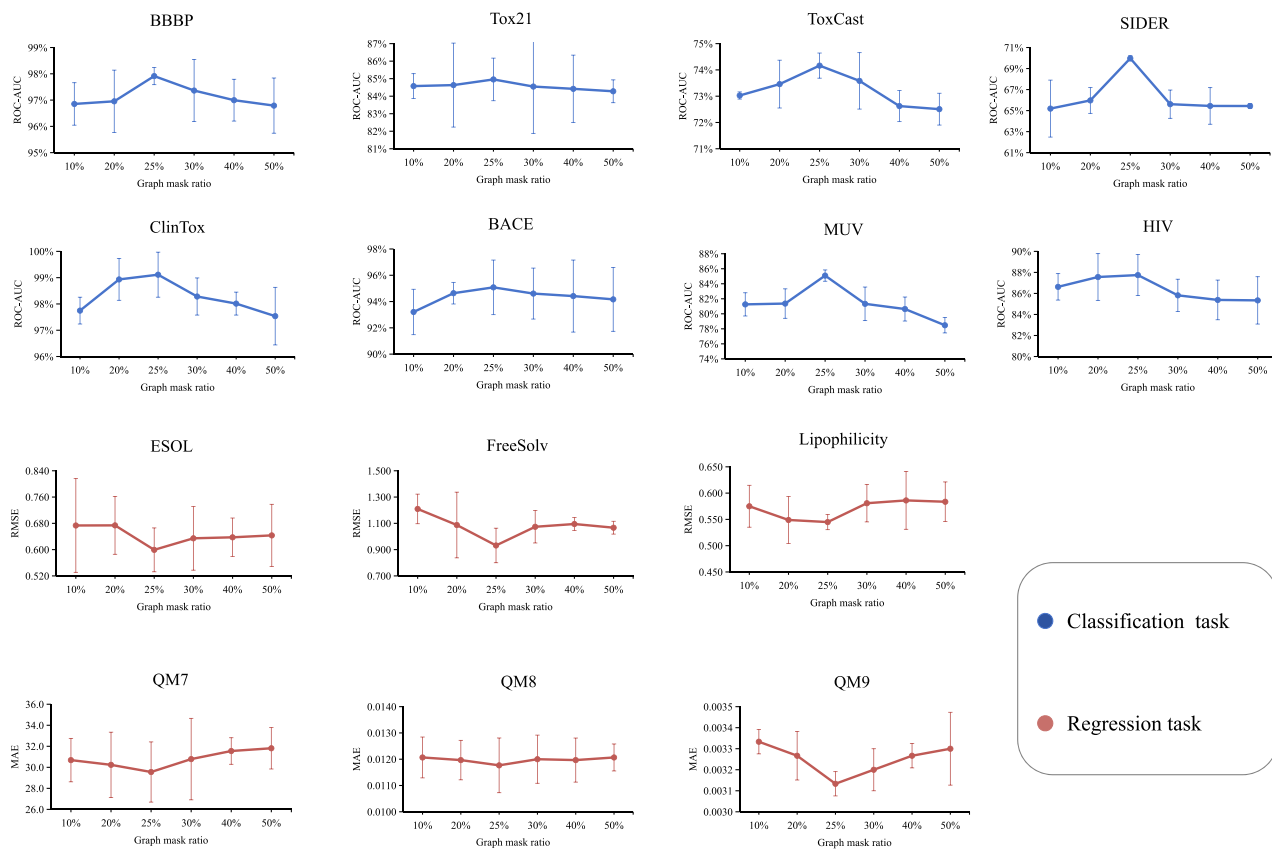


Figure 6. The impact of different mask ratios on downstream task performance on graph. The results are reported as mean and standard deviation values on three independent runs.

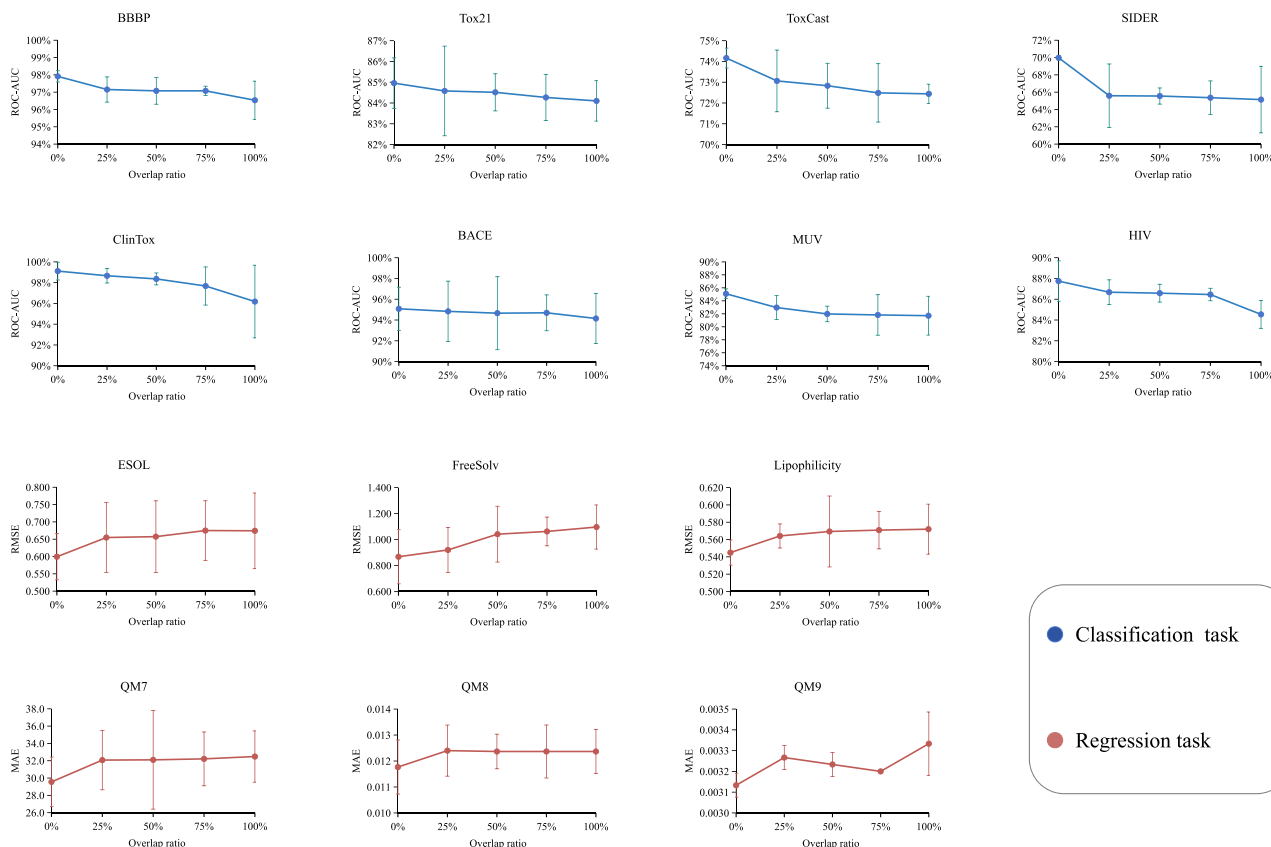


Figure 7. The impact of different overlap ratios on downstream task performance. The results are reported as mean and standard deviation values on three independent runs.

Table 5. Comparison of results on classification tasks with and without the backbone network. The mean and standard deviation of ROC-AUC (%) from three independent runs are reported (higher values indicate better performance)

	BBBP	Tox21	ToxCast	SIDER	ClinTox	BACE	MUV	HIV
Graph	97.2±0.6	84.8±1.8	73.6±0.9	65.6±0.4	98.8±0.6	89.7±5.2	81.9±1.9	85.8±1.4
Encoder+backbone								
Graph Encoder	97.9±0.3	85.0±1.2	74.2±0.5	70.0±0.2	99.1±0.9	95.1±2.1	85.1±0.8	87.7±1.9

Table 6. Comparison of results on regression tasks with and without the backbone network. The mean and standard deviation of RMSE (or MAE) from three independent runs are reported (lower values indicate better performance)

	ESOL	FreeSolv	Lipophilicity	QM7	QM8	QM9
Graph Encoder+backbone	0.661±0.011	0.988±0.250	0.560±0.017	31.9±3.8	0.0119±0.001	0.00353±0.00015
Graph Encoder	0.599±0.067	0.932±0.131	0.545±0.014	29.6±2.9	0.0117±0.001	0.00313±0.00006

Key Points

- MoleSG is a novel molecular pre-training framework that utilizes the complementary information of SMILES and graph representations, resulting in improved performance.
- To achieve information interactions in fine-grain, we design a unified network for handling two distinct modalities, allowing for fine-grained interaction between SMILES and graph representations and achieving better representation learning.

- To better learn complementary information across two modalities, we introduce an innovative non-overlapping masking strategy to encourage the model to reconstruct the masked part of one modality using the direct information of the corresponding part of the other modality, which strengthens the interactions between the two modalities.
- MoleSG achieves SOTA performance in a series of molecular property prediction tasks, and a detailed ablation study demonstrates that our proposed multi-modality

method outperforms single-modality pre-training and the masking strategy promotes performance.

Supplementary data

Supplementary data is available at Briefings in Bioinformatics online.

Funding

This work was supported by the Science and Technology Innovation Plan of Shanghai Science and Technology Commission (Grant No. 23S41900400) and the National Natural Science Foundation of China (Grant No. 62076070).

Data availability

Data and codes in our experiments are released in <https://github.com/ShenAoAO/MoleSG>.

Author contributions statement

A.S. and M.Y. proposed the main idea and the main framework. M.Y. deployed the experimental steps and A.S. performed the experiments. A.S. wrote the manuscript and M.Y. polished the manuscript. Y.M. helped create figures and typeset the manuscript using LaTeX. J.D. helped to organize experimental data. M.W. helped to improve the idea and manuscript. M.W. provided funding.

References

- Patrick Walters W, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. *Acc Chem Res* 2020;**54**(2):263–70.
- Li B, Lin M, Chen T. et al. Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Brief Bioinform* 2023;**24**(6): bbad398.
- Fedik N, Zubatyuk R, Kulichenko M. et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat Rev Chem* 2022;**6**(9):653–72.
- Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. *Nat Mach Intelle* 2021;**3**(12):1023–32.
- Gao J, Shen Z, Xie Y. et al. Transfoxmol: predicting molecular property with focused attention. *Brief Bioinform* 2023;**24**(5): bbad306.
- Brown N, Fiscato M, Segler MHS. et al. Guacamol: benchmarking models for de novo molecular design. *J Chem Inf Model* 2019;**59**(3): 1096–108.
- Rong Y, Bian Y, Tingyang X. et al. Self-supervised graph transformer on large-scale molecular data. *Adv Neural Inf Process Syst* 2020;**33**:12559–71.
- Yuan M, Shen A, Kexue F. et al. Proteinmae: masked autoencoder for protein surface self-supervised learning. *Bioinformatics* 2023;**39**(12): btad724.
- Xie Y, Zhao X, Zhang J. et al. Self-supervised learning of graph neural networks: a unified review. *IEEE Trans Pattern Anal Mach Intell* 2022;**45**(2):2412–29.
- Xia J, Zhu Y, Yuanqi D. et al. A systematic survey of chemical pre-trained models. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*. Macau, SAR, China: Morgan Kaufmann, 2023.
- Chithrananda S, Grand G, Ramsundar B. Chemberta: large-scale self-supervised pretraining for molecular property prediction arXiv preprint arXiv:2010.09885. 2020;
- Wang S, Guo Y, Wang Y. et al. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*. Niagara Falls, NY, USA: ACM, 2019, pages 429–436.
- Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**(1):31–6.
- Quirós M, Gražulis S, Girdzijauskaitė S. et al. Using smiles strings for the description of chemical connectivity in the crystallography open database. *J Chem* 2018;**10**(1):1–17.
- Lim S, Yijingxiu L, Cho CY. et al. A review on compound-protein interaction prediction methods: data, format, representation and model. *Computational and structural. Biotechnol J* 2021;**19**: 1541–56.
- Zhang Z, Chen L, Zhong F. et al. Graph neural network approaches for drug-target interactions. *Curr Opin Struct Biol* 2022;**73**:102327.
- Xiong Z, Wang D, Liu X. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2019;**63**(16):8749–60.
- Hall LH, Mohnney B, Kier LB. The electrotopological state: structure information at the atomic level for molecular graphs. *J Chem Inf Comput Sci* 1991;**31**(1):76–82.
- Zhou J, Cui G, Shengding H. et al. Graph neural networks: a review of methods and applications. *AI open* 2020;**1**:57–81.
- Liu S, Wang H, Liu W. et al. Pre-training molecular graph representation with 3d geometry. In *Proceedings of the Tenth International Conference on Learning Representations*. Virtual Event: Open-Review.net, 2022.
- Zhu J, Xia Y, Qin T. et al. Dual-view molecule pre-training. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach, CA, USA: ACM, 2023.
- Wang Y, Wang J, Cao Z. et al. Molecular contrastive learning of representations via graph neural networks. *Nat Mach Intell* 2022;**4**(3):279–87.
- Liu S, Nie W, Wang C. et al. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nat Mach Intell* 2023;**5**(12):1447–57.
- Zhu Y, Chen D, Du Y. et al. Improving molecular pretraining with complementary featurizations arXiv preprint arXiv:2209.15101. 2022.
- Chen A, Zhang K, Zhang R. et al. Pimae: point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, BC, Canada: IEEE Computer Society Press, 2023, 5291–301.
- Bachmann R, Mizrahi D, Atanov A. et al. Multimaie: Multi-modal multi-task masked autoencoders. In: *European Conference on Computer Vision*. Tel Aviv, Israel: Springer, 2022, 348–67.
- Shah R, Martín-Martín R, Zhu Y. Mutex: learning unified policies from multimodal task specifications. In *Conference on Robot Learning (CoRL)*. Atlanta, Georgia USA: The Robot Learning Foundation, Inc., 2023.
- Kexue F, Yuan M, Liu S. et al. Boosting point-bert by multi-choice tokens. *IEEE Trans Circuits Syst Video Technol* 2024;**34**(1):438–47.
- Fan H, Yishen H, Zhang W. et al. A multimodal protein representation framework for quantifying transferability across

- biochemical downstream tasks. *Advanced. Science* 2023;**10**(22): e2301223.
30. Huang X, Li S, Qu W. et al. Frozen clip model is efficient point cloud backbone. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada: AAAI, 2024.
 31. Linhui Y, Yansen S, Liu Y. et al. Review of unsupervised pretraining strategies for molecules representation. *Brief Funct Genomics* 2021;**20**(5):323–32.
 32. Sun R, Dai H, Yu AW. Does gnn pretraining help molecular representation. *Adv Neural Inf Process Syst* 2022;**35**:12096–109.
 33. Li J, Jiang X. Mol-bert: an effective molecular representation with bert for molecular property prediction. *Wireless Commun Mobile Comput* 2021;**1–7**:2021.
 34. Zhang X-C, Cheng-Kun W, Yang Z-J. et al. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Brief Bioinform* 2021;**22**(6): bbab152.
 35. Pinheiro GA, Da Silva JLF, Quiles MG. Smiclr: contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *J Chem Inf Model* 2022;**62**(17):3948–60.
 36. Feng S, Yang L, Ma W. et al. Unimap: universal smiles-graph representation learning arXiv preprint arXiv:2310.14216. 2023.
 37. Liu Y, Ott M, Goyal N. et al. Roberta: a robustly optimized bert pretraining approach arXiv preprint arXiv:1907.11692. 2019.
 38. Chen J, Zheng S, Song Y. et al. Learning attributed graph representations with communicative message passing transformer. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. Montreal-Themed Virtual Reality: Morgan Kaufmann, 2021, 2242–48.
 39. Keyulu Xu, Weihua Hu, jure Leskovec et al. How powerful are graph neural networks? In *Proceedings of the 7th International Conference on Learning Representations*. New Orleans, LA, USA: OpenReview.net, 2019.
 40. Hou Z, Liu X, Cen Y. et al. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA: ACM, 2022, pages 594–604.
 41. Gilmer J, Schoenholz SS, Riley PF. et al. Neural message passing for quantum chemistry. In: *International conference on machine learning*. PMLR, Sydney, NSW, Australia: ACM, 2017, 1263–72.
 42. Yang Kevin, Swanson Kyle, Jin Wengong. et al. Are learned molecular representations ready for prime time? arxiv:1904.01561, 2019.
 43. Song Y, Zheng S, Niu Z. et al. Communicative representation learning on attributed molecular graphs. In *IJCAI 2020*;2020: 2831–8.
 44. Guo Zhichun, Yu Wenhao, Zhang Chuxu. et al. Graseq: graph and sequence fusion learning for molecular property prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*. New York, NY, USA: ACM, 2020, pages 435–443.
 45. Liu S, Demirel MF, Liang Y. N-gram graph: simple unsupervised representation for graphs, with applications to molecules. *Adv Neural Inf Process Syst* 2019;**32**:
 46. Hu W, Liu B, Gomes J. et al. Strategies for pre-training graph neural networks. In *Proceedings of the 8th International Conference on Learning Representations*. Addis Ababa, Ethiopia: OpenReview.net, 2020.
 47. Zhang Z, Liu Q, Wang H. et al. Motif-based graph self-supervised learning for molecular property prediction. *Adv Neural Inf Process Syst* 2021;**34**:15870–82.
 48. Fang X, Liu L, Lei J. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat Mach Intell* 2022;**4**(2):127–34.
 49. Xia Jun, Zhao Chengshuai, Hu Bozhen. et al. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview.net, 2023.
 50. Fang Y, Zhang Q, Zhang N. et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat Mach Intell* 2023;1–12.
 51. Sterling T, Irwin JJ. Zinc 15–ligand discovery for everyone. *J Chem Inf Model* 2015;**55**(11):2324–37.
 52. Zhenqin W, Ramsundar B, Feinberg EN. et al. Moleculenet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**(2): 513–30.
 53. Ramsundar Bharath, Eastman Peter, Walters Pat. et al. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media, Inc., 2019.