Molecular Joint Representation Learning via Multi-Modal Information of SMILES and Graphs

Tianyu Wu, Yang Tang, Qiyu Sun, and Luolin Xiong

Abstract—In recent years, artificial intelligence has played an important role on accelerating the whole process of drug discovery. Various of molecular representation schemes of different modals (e.g., textual sequence or graph) are developed. By digitally encoding them, different chemical information can be learned through corresponding network structures. Molecular graphs and Simplified Molecular Input Line Entry System (SMILES) are popular means for molecular representation learning in current. Previous works have done attempts by combining both of them to solve the problem of specific information loss in single-modal representation on various tasks. To further fusing such multi-modal information, the correspondence between learned chemical feature from different representation should be considered. To realize this, we propose a novel framework of molecular joint representation learning via Multi-Modal information of SMILES and molecular Graphs, called MMSG. We improve the self-attention mechanism by introducing bond-level graph representation as attention bias in Transformer to reinforce feature correspondence between multimodal information. We further propose a Bidirectional Message Communication Graph Neural Network (BMC GNN) to strengthen the information flow aggregated from graphs for further combination. Numerous experiments on public property prediction datasets have demonstrated the effectiveness of our model.

Index Terms—Deep learning, graph neural network, multi-modal, molecular property prediction.

I. INTRODUCTION

In the traditional drug discovery pipeline, designing an entirely new drug from scratch can cost more than one billion USD and take more than 10 years on average [1]. Recently, various artificial intelligence techniques (e.g., Deep learning, Reinforcement learning, etc.) have been applied [2], [3], aiming at accelerating the entire process [4], [5], [6]. One crucial step in drug discovery is the construction of the quantitative structure-activity relationship (QSAR) model. Early QSAR models are generally based on fixed molecular representations, e.g., Extended Connectivity Fingerprints [7] and chemical fingerprints. However, these representations greatly rely on hand-crafted features, which restricts their usage [8]. With the development of deep learning, chemical information can be learned through large amounts of data to construct a data-driven QSAR model. Molecular representations which are generally used in deep learning can be mainly divided into two categories according to the input modal of the data: sequential representation and graph-based representation [9].

Simplified Molecular Input Line Entry System (SMILES) [10] is a kind of sequential representation with specific syntax rules, which has been widely applied for its simplicity [5]. For example, in the SMILES string “c1ccccc1”, each lowercase “c” denotes a carbon atom, and “1” represents the start and end of the ring. Thus, architectures borrowed from the field of natural language processing, for example, BERT and Transformer [11], [12], can be applied [13]. Although SMILES representation can uniquely extract global information of molecules, the loss of spatial chemical structure information still limits the performance of SMILES-based methods [14].

Unlike SMILES which uses syntax rules to encode molecules, molecular graphs are more natural representations of molecular topology. With the quick development on Graph Neural Network (GNN) [15], researchers are intuitively inspired to utilize molecular graphs for molecular representation learning. Numerous graph-based methods have been gradually proposed [16], [17], [18] with surpassing the performance of SMILES-based methods on various downstream tasks. Despite these abundant progresses, most GNN-based methods still face the problem of losing long-range dependencies and high-order properties with a limited receptive field of GNN [19]. This limitation may hurt the performance for molecular representation learning which needs whole-graph understanding. A case in Fig. 1 shows that molecules with alike structures may be projected into adjacent space in corresponding functional space by GNN model, causing problematic final predictions, which restricts their usage in real scenarios.

The main difference between SMILES and molecular graphs is their different feature input tensor encoded by their respective representations, and different explicit chemical information is embedded in them. Even if some GNN-based methods fuse Transformer to learn long-range information [20], [21], their input are still the same to most general GNN methods. Due to these different explicit chemical information, SMILES and molecular graphs have their specific emphases, and thus some
Motivated by the above discussions, we propose a novel framework of molecular joint representation learning via multi-modal information of SMILES and molecular graphs, called MMSG. Focusing on realizing the chemical information correspondence between SMILES and graphs, we make specific improvement on both SMILES and graph pipelines. Differently, bond-level representation from graphs is introduced as the self-attention bias in the Transformer for SMILES to better learn the hidden bond information, helping to fuse SMILES and graphs better. We further propose a Bidirectional Message Communication Graph Neural Network (BMC GNN) to strengthen bond and atom representations aggregated from graphs for further combination. Numerous experiments on different public property prediction datasets are conducted, reaching state-of-the-art results. The main contributions of our work are listed as follows:

- Different from previous attempts on improving fusing strategies [9], [22], [23], our MMSG focuses on realizing the chemical information correspondence between SMILES and graphs by introducing bond-level representation, which is omitted in SMILES, as self-attention bias in Transformer to better fuse multi-modal information.
- To strengthen the information flow aggregated from graphs for further combination, we propose a BMC GNN to update the atom-level and bond-level representations.
- Numerous experiments on public property prediction datasets are conducted, with achieving 3.2% relative improvement compared with current state-of-the-art property prediction baselines.

II. RELATED WORK

A. SMILES-Based Molecular Representation Learning

Previously, lots of RNN models have been proposed for SMILES-based molecular representation learning [25]. In [26], Bi-LSTM is improved by self-attention mechanism for better molecular representation. To further improve the performance of the molecular representation based on SMILES, Transformer-related models have been introduced. Both SMILES-Transformer [27] and SMILES-BERT [28] are first pre-trained on large-scale of molecules. Then the model pre-trained can be generalized to various downstream tasks through fine-tuning. Recently, the message passing operation on SMILES has also been presented to learn the hidden rules in chemical strings [29]. Unfortunately, since SMILES representation lacks the capability to express the spatial chemical structure information and the complex internal connectivity between atoms and bonds, graph-based methods are paid more attention for their better performance on molecular representation learning than SMILES-based methods [30].

B. Graph-Based Molecular Representation Learning

Generally, most of the existing GNN models [31], [32] used for molecular representation learning follow a Message Passing Neural Network (MPNN) framework [33]. The MPNN
framework includes two steps: a message passing phase and a readout phase. To fully use the message embedded in the edges of molecular graphs, some MPNN variants have been proposed, achieving promising results. In Directed Message Passing Neural Network (DMPNN) [34], molecular graphs are seen as directed graphs, and the bond message instead of the atom message has been aggregated for final graph-level representation. In Communicative Message Passing Neural Network (CMPNN) [14], the aggregated bond message has been sent to a communicative kernel to strengthen the atom representation.

Moreover, several recent graph-based methods mainly focus on enhancing information extraction with the sub-structures of molecular graphs or using the attention mechanism. In [24], multi-level information of molecular graphs is considered. In [35], both node and edge information are exploited simultaneously with a cross-dependent message-passing scheme. In [36], an attention-wise graph masking method is developed for graph augmentation. In [37], a fragment-oriented multi-scale graph attention network is proposed. In [38], fingerprints are added with molecular graphs for more information. In [39], evidential message passing network is proposed for out-of-domain examples.

However, local aggregating operations in GNN lead to the suspended animation problem that GNN performance drops dramatically when its depth increases [40], causing the loss of long-range dependencies and high-order properties of molecular graphs. Attempting to alleviate this, Transformer-type GNNs have also been studied to learn the long-range dependencies in graph-structured data. In [20], a self-supervised pre-training strategy has been developed for molecular representation learning. In [21], a Transformer framework is proposed, which uses both the node and the bond attributes for communicative message interaction. In [41], a heterogeneous GNN with Transformer-like attention is proposed for the missed information in different types of atoms or bonds. However, Transformer-type GNNs ignore the inductive bias of local substructures, and thus perform poorly in some tasks, where the topology has an important influence on molecular properties [21].

Our MMSG, which differs from Transformer-type GNNs, focuses on extracting the long-range contextual information from SMILES and the spatial structure information from graphs, helping to learn more comprehensive molecular representations. Moreover, our framework considers the chemical information correspondence between information embedded in SMILES and graphs for further fusion, helping to learning a better molecular representation.

### III. METHODOLOGY

The key idea of our MMSG is to deal with multi-modal representations with different information embedded, so as to improve the performance of molecular representation learning. To reinforce the correspondence between multi-modal information, we modify the self-attention module in Transformer by novelty introducing bond-level representations from graphs. To further strengthen the information flow aggregated from atoms and bonds, we propose a BMC GNN to update nodes and edges embeddings. Our MMSG framework is shown in Fig. 2.

#### A. Preliminary

The main goal of molecular representation learning is to calculate proper molecular representation $H$ through the representation function $R$, according to the given molecular encoding $M$, and it can be formulated as,

$$H = R(M).$$

Then, molecular representation $H$ will be applied for specific downstream tasks, such as property prediction in this work. In this work, SMILES and molecular graphs are used as the molecular encoding input.

**SMILES Encoding.** Generally, a SMILES sequence with length $T$ can be denoted as $S_{1:T} = (S_1, \ldots, S_T)$, $S_i \in D$, where $D$ is the token dictionary used to generate SMILES (e.g., “C”, “N”, “=”, “+”). By mapping these characters in a SMILES sequence into one-hot encoding, we can obtain the input matrix to learn the corresponding molecular representation. In this work, the token dictionary is generated by all the tokens appeared in the datasets.

**Graph Encoding.** A molecular can be naturally described as an attributed graph $G = (V, E)$, including a set of $n$ nodes (atoms) $|V| = n$ and a set of $m$ edges (bonds) $|E| = m$. We use $v_i$ to represent the hidden state of node $v_i$ and $e_{vw}$ to represent the edge feature of edge $e$ between node $v_i$ and $v_w$. Details about the atoms and bonds attributes used to define each node and edge are listed in Tables I and II, following [34].

In MPNN framework, the message passing phase runs for totally $T$ steps iteratively and is defined in terms of an aggregate function [33]. At each time step $t$, node message $m^{t+1}_v$ is aggregated to update the hidden states $h_v^{t+1}$ at each node in the

---

**TABLE I**

| Features     | Size | Description                        |
|--------------|------|------------------------------------|
| Atom type    | 100  | types of atom (e.g., C, H, O)      |
| Degree       | 6    | number of neighbor atoms           |
| Formal charge| 5    | integer electronic charge assigned to atom |
| Chirality    | 4    | unspecified, tetrahedral CW/CCW, or other |
| Number of Hs | 5    | number of bonded hydrogen atoms    |
| Hybridization| 5    | sp, sp2, sp3, sp3d, or sp3d2      |
| Aromaticity  | 1    | whether this atom is part of an aromatic system |
| Atomic mass  | 1    | mass of the atom, divided by 100   |

**TABLE II**

| Features  | Size | Description          |
|-----------|------|----------------------|
| Bond type | 4    | single, double, triple or aromatic |
| stereo    | 6    | none, aryl, E/Z or cis/trans |
| In ring   | 1    | whether the bond is part of a ring |
| Conjugated| 1    | whether the bond is conjugated  |
graph by the update function $U^t$,

$$m_{v}^{t+1} = \text{aggregate}(h_{v}^{t}, h_{w}^{t}, e_{vw}), u \in \mathcal{N}(v),$$  \hspace{1cm} (2)

$$h_{v}^{t+1} = U^t(h_{v}^{t}, m_{v}^{t+1}),$$  \hspace{1cm} (3)

where $\mathcal{N}(v)$ is the set of neighbors of node $v$ in graph $G$, and note that $h_{v}^{0}$ is the input node feature. Moreover, the edge message $m_{e_{vw}}$ and edge hidden state $h_{e_{vw}}$ are considered in DMPNN, and a communicative function is further added in CMPNN. Finally, a readout function [33] is applied to the set of atom representations in $G$ to get the graph-level molecular representation $H$,

$$H = \text{readout}(\{h_{v}^{T}, \forall v \in \mathcal{V} \}). \hspace{1cm} (4)$$

In graph-level molecular representation learning, various readout functions has been researched, including sum, mean, GRU and so on [21], [33], [42].

### B. Bidirectional Message Communication GNN

In MPNN, one key step is to leverage the aggregated message to update the hidden state of nodes or edges, referring to $U^t$ in (3). However, in the near CMPNN [14], only the node states $h_{e_{vw}}$ are updated iteratively, while the atom states $h_{v}$ are only updated by the communicative kernel which is an addition operator. In MMSG, to reinforce the quality of the aggregated information flow from graphs for further combination, we propose a bidirectional message communication (BMC) GNN to make full use of the node message for more effective message interactions. To realize this, we design different operating functions and specific bi-directional message transportation for nodes and edges. The details are described in Algorithm 1, and the underlined are the modifications compared with CMPNN.

The inputs of the algorithm are the graph $G = (\mathcal{V}, \mathcal{E})$ and all of its atom attributes $x_{v}$ and bond attributes $x_{e_{vw}}$. Functions are also designed with specific usage. The initial node feature $h_{v}^{0}$ is simply the atom attributes, while the initial edge feature $h_{e_{vw}}^{0}$ is the bond attributes.

Message aggregation is another crucial step for message passing. Different message aggregators capture different properties of the elements in a molecular graph [32]. The message passing procedure of BMC can be referred to Fig. 3. In our work, we differ the aggregators for node message (line 4) and edge message (line 8). Node message vector $m_{v}^{t+1}$ is aggregated by the representation its incoming edges in $G$. Edge message vector $m_{e_{vw}}^{t+1}$ is aggregated by the representation of its neighbor nodes.

$$\begin{align*}
    m_{e_{vw}}^{t+1} &= \text{MAX}(h_{e_{vw}}^{t}) \odot \text{SUM}(h_{e_{vw}}^{t}), u \in \mathcal{N}(v), \\
    m_{e_{vw}}^{t+1} &= \text{MEAN}(h_{e_{vw}}^{t}, h_{e_{vw}}^{t}),
\end{align*} \hspace{1cm} (5)$$

where MAX [43], SUM [32], MEAN [31] are the corresponding aggregating strategy, $\odot$ is an element-wise multiplication operator.

Then the obtained message vectors of node and edge are concatenated with the corresponding current hidden states to be sent to the communicate function. For simplicity, we still use an addition operator as the communicative kernel (line 5, 9, 20, 22), and more complex calculating operator can be considered in further research. Considering that the molecular graphs are seen as directed graphs, the hidden state of an edge should not rely on its reverse message which should be subtracted in each updating stage (line 12).
Algorithm 1: Bidirectional Message Communication.

Input: Graph $G=(V, E)$; atom and bond attributes $(x_v, \forall v \in V, x_{euv}, \forall e_{uv} \in E)$; aggregate functions $\text{aggregate}_v$ and $\text{aggregate}_e$, communicate functions $\text{comm}_v$ and $\text{comm}_e$, update functions $U_v$ and $U_e$, and readout functions $\text{readout}_v$ and $\text{readout}_e$ for nodes and edges respectively; network depth $T$.

1: Let $h_v^0 \leftarrow x_v, \forall v \in V; h_{e_{uv}}^0 \leftarrow x_{e_{uv}}, \forall e_{uv} \in E$
2: for $t = 0 \cdots T - 1$ do
3: for $\forall v \in V$ do
4: $m_{t+1}^v \leftarrow \text{aggregate}_v(\{h_v^t, u \in N(v)\})$
5: $p_{t+1}^v \leftarrow \text{comm}_v(m_{t+1}^v, h_v^t)$
6: end for
7: for $\forall e \in E$ do
8: $m_{t+1}^e \leftarrow \text{aggregate}_e(h_e^t, h_u^t)$
9: $p_{t+1}^e \leftarrow \text{comm}_e(m_{t+1}^e, h_e^t)$
10: end for
11: for $\forall v \in V$ do
12: $p_{t+1}^v \leftarrow p_{t+1}^v - p_{t+1}^e$
13: $h_{e_{uv}}^{t+1} \leftarrow U_e(p_{t+1}^e, h_{e_{uv}}^0)$
14: end for
15: for $\forall v \in V$ do
16: $h_v^{t+1} \leftarrow U_v(p_{t+1}^v, h_v^0)$
17: end for
18: end for
19: $m_t^v \leftarrow \text{aggregate}_v(\{h_v^{t-1}, u \in N(v)\})$
20: $h_v^t \leftarrow \text{comm}_v(m_t^v, h_v^{t-1})$
21: $m_t^e \leftarrow \text{aggregate}_e(h_e^{t-1}, h_u^{t-1}, x_v)$
22: $h_{e_{uv}}^t \leftarrow \text{comm}_e(m_t^e, h_{e_{uv}}^{t-1}, x_{e_{uv}})$
23: $H_V = \text{readout}_v(\{h_v^t, \forall v \in V\})$
24: $H_E = \text{readout}_e(\{h_e^t, \forall e \in E\})$

To further promote the whole message procedure, following [34], skip connections to the original feature vector of the node and edge have also been added (line 13, 16). The final update function can be written as,

\[
\begin{align*}
\hat{h}_v^{t+1} &= U_v(p_{t+1}^v, h_v^0) = \text{ReLU}(h_v^0 + W_v \cdot p_{t+1}^v), \\
\hat{h}_e^{t+1} &= U_e(p_{t+1}^e, h_{e_{uv}}^0) = \text{ReLU}(h_{e_{uv}}^0 + W_e \cdot p_{t+1}^e),
\end{align*}
\]

where rectified linear unit (ReLU) is the activation function used, and $W_v, W_e$ are learned matrices. $p_{t+1}^v$ and $p_{t+1}^e$ are the message of node and edge after bidirectional communication.

After $T-1$ iteration steps, another iteration is applied, with gathering the initial information, to calculate the set of final node representations $h_v^T$ and the bond representations $h_{e_{uv}}^T$ in $G$ (line 19-22).

Finally, a readout operation is applied to get the graph-level molecular representation from atoms and bonds. Following the previous work [14], we set different gated recurrent units (GRU) [44] parameterized differently for nodes and edges as readout functions, which can be written as,

\[
H_V = \sum_{v \in V} \text{GRU}(h_v^T), H_E = \sum_{e_{uv} \in E} \text{GRU}(h_{e_{uv}}^T),
\]

Comparing with introducing simple permutation invariant functions such as summations as the readout function [32], applying GRU units helps to learn more high-order information [42].

Eventually, we gain graph-level molecular representations for the molecule on atoms $H_V$ and on bonds $H_E$. $H_V$ is used to create final molecular representation and $H_E$ is used to replenish the bond message of SMILES-based molecular representation in Transformer framework.

C. Transformer-Based Message Self-Attention

Since attention mechanism has shown significant performance on extracting various information [45], [46], we further propose a Transformer-based message self-attention module to realize the chemical information correspondence between SMILES and graphs. More precisely, we introduce bond-level representation $H_E$ calculated by our BMC GNN as the self-attention bias, which helps the model accurately capture the dependency of the hidden bond-level information in SMILES.

In our Transformer framework, we do not catch the positional encoding [12] of the SMILES sequence data. Instead, we first use a bidirectional GRU unit to pre-process the one-hot encoding to capture the feature vector of the sequence in SMILES, and this progress can be formulated as,

\[
\begin{align*}
\overrightarrow{h}_{t_i} &= \text{GRU}\left(t_i, \overrightarrow{h}_{t_{i-1}}\right), \\
\overleftarrow{h}_{t_i} &= \text{GRU}\left(t_i, \overleftarrow{h}_{t_{i-1}}\right),
\end{align*}
\]

where $\overrightarrow{h}_{t_i}, \overleftarrow{h}_{t_i}$ are bidirectional hidden states for the $i$th token of a SMILES string embedded by GRU, and a hidden state $h_i$ is obtained to replace token embedding $t_i$ as,

\[
h_i = \left(\overrightarrow{h}_{t_i}, \overleftarrow{h}_{t_i}\right).
\]
Finally, we use $H_s$ to denote the contextual representation of a SMILES string with length $n$ as,

$$H_s = (h_0, h_1, \ldots, h_n).$$

Then $H_s$ is sent to Transformer to learn long-range information hidden in SMILES. The input is mapped into different embeddings $(Q, K, V)$ with corresponding matrices $W$, where $(Q, K, V)$ represents queries, keys and values. The Attention function \cite{vaswani2017attention} for self-attention in Transformer can be defined as:

$$\text{Attention}(H_s) = \text{softmax}(QK^T/\sqrt{d_K})V,$$

where $d_K$ is the dimension of $K$ and softmax is a softmax function. We omit the bias term of multi-head attention here for simplicity since it is a straightforward extension of single-head self-attention.

Considering that in SMILES strings, feature of each atom can be intuitively extracted from word tokens, while most of the single and aromatic bonds are omitted. Following this line, we naturally use bond representations from graphs to reinforce the correspondence on bond-level information learned by SMILES and graphs with the help of attention mechanism, so as to improve the quality of the final molecular representation. We also apply linear transformation for $H_E$ from BMC GNN to deal with the different length of SMILES tokens and bond features, thus they can be processed correctly. The improved self-attention matrix can be written as:

$$\text{Attention}(H_s) = \text{softmax}(QK^T/\sqrt{d_K} + H_E)V,$$

By applying layer normalization (LN) before multi-head self-attention (MHA) blocks and feed-forward network (FFN) blocks \cite{vaswani2017attention}, we get the final sequence representation $H_S$ as:

$$H'_S = \text{MHA}([\text{LN}(H_s)] + H_s),$$

$$H_S = \text{FFN}([\text{LN}(H'_s)] + H'_s).$$

Finally, by combining the representation $H_S$ and $H_V$ from SMILES and graphs, molecular properties $Y$ can be further predicted through an independent feed-forward network (FFN):

$$Y = \text{FFN}(H_S + H_V).$$

IV. EXPERIMENTS

A. Benchmark Datasets

To display the molecular representation capability of our MMSG framework, we conduct experiments on 7 public benchmark datasets from MolecularNet \cite{Jin_2021} across classification and regression tasks. Table III summarizes all the information of the benchmark datasets used, including task types, dataset sizes and the metric used. Among all of them, BBBP \cite{Wright_2019}, Tox21 \cite{Gilbride_2019}, SIDER \cite{Zong_2018} and ClinTox \cite{Zong_2018} are used for binary or multi-binary classification tasks, and FreeSolv \cite{Gibler_2019}, ESOL \cite{Gibler_2019} and Lipophilicity \cite{Yuan_2020} are used for regression tasks.

Details of each dataset are listed as follows. The Blood–brain barrier penetration (BBBP) dataset \cite{Wright_2019} contains compounds on their permeability properties with binary labels. The “Toxicology in the 21st Century” (Tox21) dataset \cite{Gilbride_2019} contains large amounts of experimental compounds for 12 different targets relevant to drug toxicity. The Side Effect Resource (SIDER) dataset \cite{Zong_2018} contains marketed drugs along with side effects and these drugs are grouped into 27 system organ classes. The ClinTox dataset \cite{Zong_2018} contains drugs with FDA approval status and clinical trial toxicity. The Free Solvation (FreeSolv) dataset \cite{Gibler_2019} contains measurements on several compounds with its calculated hydration free energy of in water. ESOL dataset \cite{Gibler_2019} contains water solubility data for molecules. Lipophilicity dataset \cite{Yuan_2020} contains the experimental results of octanol/water distribution coefficient of the molecules.

B. Baseline Models

We compare MMSG with 19 baseline methods, including the models appeared in the MolecularNet and several state-of-the-art approaches, totally numbered 20. All of the baseline methods are summarized in Table IV. These methods can be mainly divided into different types according to the molecular representation and methods used. RNN \cite{Jin_2021}, Transformer \cite{vaswani2017attention} and MPAD \cite{Jin_2021} use SMILES sequence as the input with different network structures. GCN \cite{Defferrard_2016}, Weave \cite{Klicpera_2018}, GIN \cite{Duvenaud_2015}, N-Gram \cite{Vaswani_2017} and Attentive FP \cite{Yuan_2020} are graph convolutional methods.
MPNN [33], DMPNN [34], CMPNN [14], GROVER [20], CoMPT [21] and CD-MVGNN [35] follow the message passing methods operated on directed graphs. MoCL [16] and ATMOL [36] introduces contrastive learning, while HamNet [58] and GeomGCL [17] incorporate 3D geometry of molecules. GraSeq [9] and our MMSG mainly focus on combing multi-modal molecular representations to improve the final performance. Noted that GROVER and ATMOL are utilized without pretrained for fair comparison. In our work, we choose more graph-based methods which have better performance for comparison so that to prove the effectiveness of combing multi-modal representations.

C. Experimental Settings

To comprehensively evaluate the learning ability of our MMSG, we follow MoleculeNet [47] and split all the datasets by random splitting and scaffold splitting. The random splitting separates molecules in datasets randomly. The scaffold splitting, which is a more challenging and realistic setting, aims to separate molecules with different two-dimensional structural frameworks into different subsets. Both of the splitting ways are adopted with a ratio for train/validation/test sets as 0.8: 0.1: 0.1. We perform five independent runs with different random seeds on different splitting methods, and calculate the mean and standard deviation values of area under the receiver operating characteristic (ROC-AUC) or root-mean-squared error (RMSE) metric. The implementation of the model relies on Pytorch and RDKit package. Table V demonstrates the range of hyper-parameters used in experiments for MMSG, including the numbers of different network layers, batch size, learning rate, hidden dimension and so on.

D. Results and Analysis

Quantitative Results on the Scaffold Splitting. Table VI summarizes evaluations on scaffold splitting among parts of the baseline models, since some of the results are not reported previously. Cells in bold indicate the best results on different datasets and the cells in gray rank second. Table VI provides quantitative results on scaffold splitting of different baseline models and our MMSG, and “Trans” means the Transformer model.
the following observations: 1) By combining SMILES and graph representations simultaneously, our MMSG framework achieves the state-of-the-art results on most benchmark datasets on scaffold splitting. For SIDER, our MMSG method also ranks the second, while ATMOL show a significant performance than all of the other methods. The relatively improvement is 3.0% on classification tasks and 3.4% on regression tasks, and the total is 3.2% on all datasets. This fact proves the effectiveness of our multi-modal molecular joint representation learning framework with SMILES and graph. The results show that our MMSG can exactly extract a more comprehensive molecular representation for downstream tasks. 2) Comparing with GraSeq, our MMSG shows big enhancement by improving feature extraction modules of SMILES and graphs with considering the correspondence between them and the relatively improvement is 16.7% on average. 3) Our MMSG can obviously improve the property prediction performance especially on some specific tasks, like BBBP, ClinTox and ESOL. This fact means that with combining multi-modal molecular representation, more hidden chemical information can be learned of specific tasks, helping to improve the final molecular representation.

**E. Ablation Studies**

To further investigate the positive influence of the modules proposed, several ablation experiments are conducted. Noted that all ablation studies are conducted under random splitting.

**The Performance of Different Molecular Representations**. In this study, we testify the potential value of the different molecular representations. Detailed results under random splitting are recorded in Table VIII. “Trans” means the Transformer model which only uses SMILES representation as the input, while “BMC” means our BMC GNN which only uses graph as the input. “MMSG (w/o)” means our MMSG combines both of the molecular representations without adding attention bias, and note that the results of “MMSG (ours)” here is same to Table VII with same random splitting pattern. Table VIII also shows that when only using single-modal representation, SMILES-based model still performs better than graph-based model on ClinTox dataset. This fact furthure proves the fact that using proper molecular representation is helpful to learn specific molecular representation for better multi-modal molecular representation since SMILES and graphs are both 2D molecular features.

From both the results on scaffold splitting and random splitting, it can be obviously found that our MMSG shows significant performance on ClinTox dataset. Similarly, MPAD also performs quite well on ClinTox. An intuitive finding is that it may be easier to capture related chemical information from SMILES to make corresponding predictive tasks on ClinTox, and the great performance of the MPAD on ClinTox which also supports this point. This fact further proves the potential of learning molecular representation from multi-modal representations.
properties. According to the results of “MMSG (w/o)”, with simply combing SMILES and graphs, the improvements are not attractive on all the tasks. With considering the chemical information between multi-modal representation, SMILES and graph in this work, more accurate molecular representation can be learned for various downstream task and the final results are also improved.

The Performance of Different Readout Function in BMC. In this part, we testify the performance of applying various readout functions in BMC under  random splitting. We collect five different patterns used in previous works [14], [21], [34], including “Mean”, “Sum”, “Normalization”, “Summax” and “GRU”. Detailed results under random splitting are recorded in Table IX. Obviously, our BMC framework shows better performance with applying “GRU” as the readout function than others, proving that high-order information can be learned. In this way, we choose using “GRU” with our BMC framework as our final model, and the results of “GRU” here is same to the “BMC” in Table VIII.

The Performance of the Bidirectional Message Communication GNN. To verify the performance of our BMC GNN in promoting the information flow for better molecular representation of nodes and edges, we compare its performance with the original CPMNN (also with GRU readout function) under  random splitting. Results are shown in Table X, and the results re-confirm the effectiveness of our BMC GNN.

F. Molecular Representation Visualization

As mentioned before, most of the GNN models face the problem of losing long-range dependencies which restricts the performance for molecular representation, and a case study is shown in Fig. 1. For molecules with similar structure, GNN model may project them into similar representations which are adjacent in corresponding chemical functional space, causing problematic final predictions. In our work, by realizing the chemical information correspondence between SMILES and graphs, our MMSG framework gains a more comprehensive molecular representation as expected. To prove this, we visualize the latent space by reducing their dimensions to 2 with UMAP. Projections results on FreeSolv, Lipophilicity and BBBP are illustrated in Figs. 4, 5, and 6. Each point corresponds to a molecule and is colored according to its predicted value.
Fig. 4. Visualization of the latent space by UMAP. Molecules are colored with the predicted values on the FreeSolv dataset by Transformer, CMPNN and MMSG. The color bar expresses the correspondence between the value and the color.

Fig. 5. Visualization of the latent space by UMAP. Molecules are colored with the predicted values on the Lipophilicity dataset by Transformer, CMPNN and MMSG. The color bar expresses the correspondence between the value and the color.

Fig. 6. Visualization of the latent space by UMAP. Molecules are colored with the predicted results on the BBBP dataset by Transformer, CMPNN and MMSG. Blue points mean positive results and red points mean negative results.

For large-scale dataset Lipophilicity (Fig. 5) and BBBP (Fig. 6), we can get the same results. From Figs. 5(b) and 6(b), it can be clearly found that molecules with positive properties and molecules have clear boundaries. While in Figs. 5(c) and 6(c), molecules show more delicate difference instead clustering together. In the way, these results suggests that our MMSG model capture better representations of molecules.

V. CONCLUSION

In this paper, we propose a novel molecular joint representation learning framework, called MMSG, via multimodal molecular information (from SMILES and graphs). We
further consider the correspondence between chemical information embedded in multi-modal representations for completely different encoding rules. We modify the self-attention module in Transformer to reinforce the bond-level information correspondence between SMILES and graphs. Further, we realize bidirectional communication between nodes and edges in directed graphs so as to improve the information flow. Numerous experiments on different public datasets prove the effectiveness of our proposed model, with surpassing state-of-the-art results.

Moreover, in this work, we only consider the usage of SMILES and molecular graphs, while our framework has strong portability. More means of molecular representations, e.g., 3D structure of a molecule, can be taken into consideration to gain specific chemical information. More research will be investigated in future work.

REFERENCES

[1] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, “Innovation in the pharmaceutical industry: New estimates of R&D costs,” J. Health Econ., vol. 47, pp. 20–33, 2016.
[2] Y. Qian, Y. Ding, Q. Zou, and F. Guo, “Multi-view kernel sparse representation for identification of membrane protein types,” IEEE/ACM Trans. Comput. Biol. Bioinf., 2022, doi: 10.1109/TCBB.2022.3191325.
[3] J. Guan, R. Li, S. Yu, and X. Zhang, “A method for generating synthetic electronic medical record text,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 18, no. 1, pp. 173–182, Jan./Feb., 2021.
[4] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, pp. 583–589, 2021.
[5] J. Jiménez-Luna, F. Grisoni, and G. Schneider, “Drug discovery with explainable artificial intelligence,” Nature Mach. Intell., vol. 2, no. 10, pp. 573–579, 2020.
[6] Q. Zhao, M. Yang, Z. Cheng, Y. Li, and J. Wang, “Biomedical data and deep learning computational models for predicting compound-protein relations,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 4, pp. 2092–2110, Jul./Aug., 2022.
[7] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010.
[8] B. Tang, S. T. Kramer, M. Fang, Y. Qiu, Z. Wu, and D. Xu, “A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility,” J. Cheminformatics, vol. 12, no. 1, pp. 1–9, 2020.
[9] Z. Guo, W. Yu, C. Zhang, M. Jiang, and N. V. Chawla, “GraSeq: Graph and sequence fusion learning for molecular property prediction,” in Proc. 29th ACM Int. Joint Conf. Inf. Knowl. Manage., 2020, pp. 435–443.
[10] D. Weininger, “Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules,” J. Chem. Inf. Comput. Sci., vol. 28, no. 1, pp. 31–36, 1988.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Nat. Acad. Sci. USA, 2019, pp. 7804–7809.
[12] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, and Y. Yang, “Communicative representation learning on attributed molecular graphs,” in Proc. 29th Int. Joint Conf. Artif. Intell., 2020, pp. 2831–2838.
[13] J. Chen, J. Gao, T. Lyu, B. M. Ouloulade, and X. Hu, “AutoMSR: Auto molecular structure representation learning for multi-label metabolic pathway prediction,” IEEE/ACM Trans. Comput. Biol. Bioinf., 2022, doi: 10.1109/TCBB.2022.3198119.
[14] M. Sun, J. Xing, H. Wang, B. Chen, and J. Zhou, “MoCL: Data-driven molecular fingerprint via knowledge-aware contrastive learning from SMILES,” in Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining, 2021, pp. 3585–3594.
Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, “Gated graph sequence neural networks,” in Proc. 4th Int. Conf. Learn. Representations, 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/gated-graph-sequence-neural-networks/

T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. 5th Int. Conf. Learn. Representations, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYg!

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in Proc. 28th Conf. Neural Inf. Process. Syst. Workshop Deep Learn., 2014.

Z. Cheng, C. Yan, F.-X. Wu, and J. Wang, “Drug-target interaction prediction using multi-head self-attention and graph attention network,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 4, pp. 2208–2218, Jul./Aug. 2022.

Y. Guo, O. Krupa, J. Stein, G. Wu, and A. Krishnamurthy, “SAU-Net: A unified network for cell counting in 2D and 3D microscopy images,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 19, no. 4, pp. 1920–1932, Jul./Aug. 2022.

Z. Wu et al., “MoleculeNet: A benchmark for molecular machine learning,” Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.

I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, “A Bayesian approach to in Silico blood-brain barrier penetration modeling,” J. Chem. Inf. Model., vol. 52, no. 6, pp. 1688–1697, 2012.

NCATS, “Tox21 challenge,” 2017. Accessed: Sep. 07, 2017. [Online]. Available: http://tripod.nih.gov/tox21/challenge/

M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The sider database of drugs and side effects,” Nucleic Acids Res., vol. 44, no. D1, pp. D1075–D1079, 2016.

K. M. Gayvert, N. S. Madhukar, and O. Elemento, “A data-driven approach to predicting successes and failures of clinical trials,” Cell Chem. Biol., vol. 23, no. 10, pp. 1294–1301, 2016.

D. L. Mobley and J. P. Guthrie, “FreeSolv: A database of experimental and calculated hydration free energies, with input files,” J. Comput.-Aided Mol. Des., vol. 28, no. 7, pp. 711–720, 2014.

J. S. Delaney, “ESOL: Estimating aqueous solubility directly from molecular structure,” J. Chem. Inf. Comput. Sci., vol. 44, no. 3, pp. 1000–1005, 2004.

M. Wenlock and N. Tomkinson, “Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds,” 2015, doi:10.6019/CHEMBL3301361.

S. Kearnes, K. McCuskey, M. Berndl, V. Pande, and P. Riley, “Molecular graph convolutions: Moving beyond fingerprints,” J. Comput.-Aided Mol. Des., vol. 30, no. 8, pp. 595–608, 2016.

S. Liu, M. F. Demirel, and Y. Liang, “N-Gram graph: Simple unsupervised representation for graphs, with applications to molecules,” in Proc. 23rd Conf. Neural Inf. Process. Syst., 2019, pp. 8466–8478.

Z. Xiong et al., “Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism,” J. Med. Chem., vol. 63, no. 16, pp. 8749–8760, 2019.

Z. Li, S. Yang, G. Song, and L. Cai, “Conformation-guided molecular representation with hamiltonian neural networks,” in Proc. 10th Int. Conf. Learn. Representations, 2021. [Online]. Available: https://openreview.net/forum?id=sq-cnWaoUTH

Tianyu Wu received the BS degree in automation from the East China University of Science and Technology, Shanghai, China, in 2020, where he is currently working toward the PhD degree in control science and engineering. His fields of interest include molecular representation learning, reinforcement learning.

Yang Tang (Senior Member, IEEE) received the BS and PhD degrees in electrical engineering from Donghua University, Shanghai, China, in 2006 and 2010, respectively. From 2008 to 2010, he was a research associate with the Hong Kong Polytechnic University, Hong Kong. From 2011 to 2015, he was a post-doctoral researcher with the Humboldt University of Berlin, Berlin, Germany, and with the Potsdam Institute for Climate Impact Research, Potsdam, Germany. He is now a professor with the East China University of Science and Technology, Shanghai. His current research interests include distributed estimation/control/optimization, cyber-physical systems, hybrid dynamical systems, computer vision, reinforcement learning and their applications. He was a recipient of the Alexander von Humboldt Fellowship and has been the ISI Highly Cited Researchers Award by Clarivate Analytics from 2017. He is a Senior Board member of the Scientific Reports, an associate editor of the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, IEEE Transactions on Circuits and Systems-I: Regular Papers, IEEE Transactions on Cognitive and Developmental Systems, IEEE Transactions on Emerging Topics in Computational Intelligence, IEEE Systems Journal, Engineering Applications of Artificial Intelligence (IPAC Journal) and Science China Information Sciences, etc. He is a (leading) guest editor for several special issues focusing on autonomous systems, robotics, and industrial intelligence in the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Cognitive and Developmental Systems, IEEE Transactions on Emerging Topics in Computational Intelligence.

Qiuyun Sun received the BS degree in automation from the East China University of Science and Technology, Shanghai, China, in 2019, where she is currently working toward the PhD degree in control science and engineering. Her fields of interest include 3-D scene understanding, domain adaptation.

Luolin Xiong (Graduate Student Member) received the BS degree from the School of Control Science and Engineering, East China University of Science and Technology, Shanghai, China, in 2020. She is currently working toward the PhD degree with the East China University of Science and Technology. Her research interests include reinforcement learning and their applications in smart grid.