Masking in Molecular Graphs Leveraging Reaction Context

Jiannan Yang [1] **Veronika Thost** [2 †] **Tengfei Ma** [1 †]

Abstract

Token masking has been proven useful for self-supervised learning in various modalities, including the sequential SMILES representation of molecules. Yet, research for masking over molecular graph structures has not received enough attention and existing methods often focus on single molecules. We propose **ReaCTMask** (**Rea**ction **C**on**T**ext-based **Mask**ing), a novel approach that leverages reaction knowledge to provide critical context outside of the molecular structures themselves to guide the graph masking. We show that graph transformers are able to exploit the additional knowledge by applying a unified masking scheme, within and across molecules inside a reaction. Our experiments cover probing and transfer learning, comparing to various baselines, and provide insights into the intricate nature of the task. Overall, the results demonstrate the effectiveness of our approach and, more generally, the usefulness of reaction context in graph pre-training.

1. Introduction

In recent years, people have witnessed the thriving of artificial intelligence in various scientific areas. Self-supervised pretraining models have contributed greatly to its success, as demonstrated by large language models. It also has been playing a crucial role in enhancing model capabilities in learning molecular representations for drug discovery (Liu et al., 2023a; Zhao et al., 2024). By pretraining a backbone deep neural network on some large molecule datasets, the pretrained models can be easily generalized to different datasets and tasks.

Most of the self-supervised learning (SSL) approaches on molecular representation use contrastive learning as the pretraining scheme. For example, GraphCL (You et al., 2020) maximizes the agreement between two different augmentations of the same molecule, such as edge dropping, and subgraph sampling, of a single graph using a contrastive loss, facilitating the model's ability to extract invariant representations; there are also works integrating additional domain knowledge into contrastive learning, for instance, customized topological features (Luo et al., 2024). Another common method for pre-training is mask prediction. AttrMask (Hu et al., 2020) randomly samples nodes in a graph and replaces the original node attributes with a nonexistent token; GROVER (Rong et al., 2020) masks subgraphs within a molecular graph and predicts a designed contextual property constructed from the numbers and types of neighboring atoms and bonds. However, during the pretraining phase, these previous studies focus on the reconstruction of single molecules while overlooking valuable properties inherently presented in the raw data. Among these properties, chemical reactions provide rich contextual information that has the potential to enrich the training set.

Our proposed approach, **ReaCTMask**, integrates chemical reactions in the pre-training process. Specifically, we concatenate the products and reactants in a chemical reaction and leverage the power of a graph transformer, to embed the context information of the chemical reactions. The reactions can either be provided in the dataset or automatically generated using existing open-source retrosynthesis tools, thereby reducing the need for extensive manual annotation or specialized expertise. Moreover, given the significant role that motifs play in chemical reactions, we hypothesize that motif-level masking and reconstruction is an effective pretraining task that can yield substantial performance gains. Our comprehensive evaluations on the MoleculeNet (Wu et al., 2018) benchmarks demonstrate significant improvements in the model's representational power when equipped with ReaCTMask.

2. Background and Related Works

Notation. We denote an input graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X, E)$, where $\mathcal{V}$ represents the set of all nodes (or vertices), and $\mathcal{E}$ represents the set of all edges. Here, X and E are the corresponding node attribute matrix and edge attribute matrix, respectively. Here we only consider undirected graphs to represent the molecules in this paper.

[†]Joint Senior Authors [1]Stony Brook University [2]MIT-IBM Watson AI Lab, IBM Research. Correspondence to: Jiannan Yang <jiannan.yang@stonybrook.edu>.

ICML 2024 AI for Science workshop, Vienna, Austria. Copyright 2024 by the author(s).

Masking in SSL. Previous methods, such as AttrMask (Hu et al., 2020), GraphMAE (Hou et al., 2022) and Mole-BERT (Xia et al., 2023), typically employ node-level attribute masking. This approach involves replacing the attribute of the selected atoms with a non-existent value m. For instance, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X, E)$, we select a subset of nodes $\tilde{\mathcal{V}} \subset \mathcal{V}$ to be masked, we then replace the node attribute matrix X by $\tilde{X}$, where

$$\tilde{x}_i = \begin{cases} m & v_i \in \tilde{\mathcal{V}} \\ x_i & v_i \notin \tilde{\mathcal{V}} \end{cases} \tag{1}$$

Then, the masked molecular graphs are fed into a graph neural network, to reconstruct the masked features.

Motifs in SSL. Besides, motif-level (or subgraph-level) information has been demonstrated to be effective in past studies. MGSSL (Zhang et al., 2021) adopts a motif generation task, where a search algorithm–either depth-first search (DFS) or breadth-first search (BFS)–is used to predict the motif labels of the next motif node auto-regressively, starting from a motif in a motif-level graph. On the other hand, GROVER (Rong et al., 2020) devises a graph-level motif prediction task: they utilize RDKit (Landrum, 2006) to extract a few functional groups and use the graph embeddings to predict the presence of specific motifs within the molecules.

Reaction-Based Pre-training Strategies. While most studies of SSL on molecular data have focused on single molecule data, there are works emerging in recent years that attempt to incorporate chemical reactions into SSL as well. Firstly, since chemical reactions can be represented using SMILES strings, most prior works leverage these representations as inputs for language models. For instance, a BERT-style pre-training scheme can be employed, as demonstrated by (Schwaller et al., 2021). One can also predict the products given the reactants and reagents (Broberg et al., 2022), or conversely, the reactants from the products (Jiang et al., 2023). Nevertheless, methods involving graph representations of chemical reactions are currently under-explored. Recent works have primarily employed contrastive learning frameworks. One approach uses contrastive loss to differentiate reactants across various substrate scopes (Gao et al., 2024). Another method treats the reactants and products on either side of a chemical equation as distinct views, hypothesizing an equivalence relation between them (Wang et al., 2022; Zeng et al., 2023). In particular, MolR (Wang et al., 2022) assumes that:

$$\sum_i R_i = \sum_j P_j$$

Here R_i and P_j are learned graph-level embeddings of a reactant graph and a product graph in a chemical equation, respectively. Under this assumption, a margin-based contrastive loss inspired by (Bordes et al., 2013) is optimized. However, our experiments indicate that the contrastive method is not the most ideal way to incorporate reaction information. Observe that the theoretical hypothesis in MolR, specifically the reaction equivalence, might not always hold, due to structural and property changes in chemical reactions. A subsequent work, ReaKE (Xie et al., 2024), referred to this issue as the **ambiguous embeddings** problem. They proposed alleviating this issue by constructing a knowledge graph that connects products and reactants by the extracted reaction templates using RDChiral (Coley et al., 2019).

3. Methods

3.1. Motif-Masking Strategies in Single Molecules.

Considering the importance of chemical substructures for molecule property prediction and the demonstrated usefulness of motifs in previous approaches, we begin with the motif-masking strategies in a single molecule $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X, E)$. The first step aims to fragment $\mathcal{G}$ to a collection of motifs $\mathcal{M} = \{M_i\}_{i=1}^n$. Each motif $M_i = (\mathcal{V}_i, \mathcal{E}_i, X_i, E_i)$ is defined as a subgraph of the molecular graph, such that

$$\bigcup_{i=1}^n M_i = \mathcal{G}$$

There are various approaches to do so. We adopt the refined BRICS algorithm proposed by (Zhang et al., 2021), which decomposes the motifs from BRICS (Degen et al., 2008) further to finer subgraphs to reduce the number of infrequent large motifs. Based on this, we create a straightforward yet efficient mapping between motifs and atoms, which facilitates the masking of atoms corresponding to motifs, as well as the extraction of motif-level representations. In an abstract way, we can define it as

$$\begin{aligned} F : \mathcal{M} &\to 2^{\mathcal{V}} \\ M_i &\mapsto \mathcal{V}_i \end{aligned}$$

where $2^{\mathcal{V}}$ is the power set of $\mathcal{V}$.

After obtaining the motif-atom correspondence, we uniformly sample the motifs in $\mathcal{M}$ with a ratio $r \in (0, 1)$. Then, we mask out the nodes within the sampled motifs, according to eq. (1). Particularly, we do not mask out all the nodes within a motif to avoid completely losing the information within the motif. Masking all nodes would force the model to rely solely on the information from nodes surrounding the motif, leading to insufficient information and thus poorer learning performance. Therefore, we choose an atom sampling ratio to mask only part of the motifs. See Figure 1 for an example of this process.

Nevertheless, masking motifs in a single molecule might

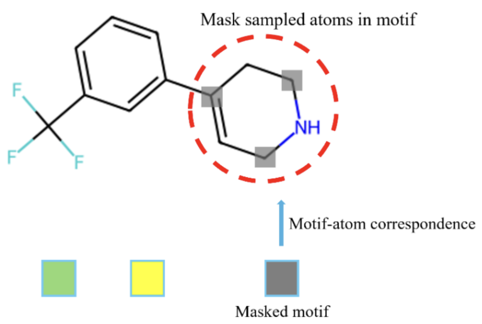

Figure 1. Atom sampling within a motif

overlook the implicit information from the chemical reaction. Take the Fischer esterification as an example:

$$\mathrm{CH_3COOH} + \mathrm{C_2H_5OH} \rightarrow \mathrm{CH_3COOC_2H_5} + \mathrm{H_2O}$$

The ester group $-\mathrm{COO}-$ in ethyl acetate ($\mathrm{CH_3COOC_2H_5}$) is formed by the reaction between a carboxyl group ($-\mathrm{COOH}$) of acetic acid and a hydroxyl group ($-\mathrm{OH}$) of ethanol. In a typical message-passing neural network, if we solely mask and reconstruct the motif $-\mathrm{COO}-$ in the single molecule $\mathrm{CH_3COOC_2H_5}$, the network might overly focus on the information from the methyl group ($-\mathrm{CH_3}$) and ethyl group $-\mathrm{C_2H_5}$. These functional groups, however, do not significantly impact the properties of the resulting ester group. If we can integrate the chemical reaction that produces a molecule into a graph neural network, we will be able to encode information from the reactants, meaning that the model can learn more reliable chemical properties from the masked motif prediction task (the details are outlined in section 3.5). In this example, the carboxyl group and the hydroxyl group in the reactants can guide the model to learn a more chemically meaningful representation of the masked ester group.

Therefore, we devise **ReaCTMask**, which concurrently encodes the products and reactants in a chemical equation, to ensure the model effectively and efficiently learns from the reaction context of each molecule.

3.2. Construction of Reaction Graphs

ReaCTMask is flexible in choosing the pre-training datasets no matter whether the reaction information is ready or not. It can be the dataset that already contains chemical reactions, such as USPTO-30K (Schneider et al., 2016), where each record contains the SMILES representations (Weininger, 1988) of both reactants and products. Alternatively, we can also choose datasets that only contain individual molecules, such as those in the ZINC15 database (Sterling & Irwin, 2015). In this case, we can employ the open-source retrosynthesis tool **AiZynthFinder** (Genheden et al., 2020) to solve the synthetic routes for a target molecule in ZINC. By extracting the last step of a synthetic route, we can determine the SMILES representations of the necessary reactants to synthesize the target molecule.

Based on the data pre-processing pipelines in (Hu et al., 2020), we transform the SMILES of reactants and products to corresponding molecular graphs via RDKit. A pivotal step involves merging these graphs to create a disjoint union of graphs. Within this union, integer labels $\{0, 1\}$ are generated to distinguish atoms belonging to reactants from those belonging to products. We treat this disjoint union as a single entity to be fed into the graph neural networks for pre-training.

Formally, in a simplified chemical reaction

$$R_1 + R_2 \rightarrow P$$

let the reactant graphs be $\mathcal{G}_{R_1} = (\mathcal{V}_{R_1}, \mathcal{E}_{R_1}, X_{R_1}, E_{R_1})$ and $\mathcal{G}_{R_2} = (\mathcal{V}_{R_2}, \mathcal{E}_{R_2}, X_{R_2}, E_{R_2})$, the product graph be $\mathcal{G}_P = (\mathcal{V}_P, \mathcal{E}_P, X_P, E_P)$, we define the **reaction graph** to be the disjoint union of them (see Figure 2):

$$\mathcal{G}_U := (\mathcal{V}_U, \mathcal{E}_U, X_U, E_U) \tag{2}$$

where

$$\begin{aligned} \mathcal{V}_U &:= \mathcal{V}_P \sqcup (\mathcal{V}_{R_1} \cup \mathcal{V}_{R_2}) \\ &= (\mathcal{V}_P \times \{0\}) \cup ((\mathcal{V}_{R_1} \cup \mathcal{V}_{R_2}) \times \{1\}) \\ \mathcal{E}_U &:= \mathcal{V}_P \sqcup (\mathcal{E}_{R_1} \cup \mathcal{E}_{R_2}) \\ &= (\mathcal{E}_P \times \{0\}) \cup ((\mathcal{E}_{R_1} \cup \mathcal{E}_{R_2}) \times \{1\}) \end{aligned}$$

and

$$\begin{aligned} X_U &:= \left[X_P^T, X_{R_1}^T, X_{R_1}^T\right]^T \\ E_U &:= \left[E_P^T, E_{R_1}^T, E_{R_2}^T\right]^T \end{aligned}$$

3.3. Motif-Masking Strategies in Reaction Graphs

It requires minor modifications to adapt the motif-masking scheme from single molecules to reaction graphs: we decompose each molecular graph (e.g. $\mathcal{G}_P$, $\mathcal{G}_{R_1}$ and $\mathcal{G}_{R_2}$) within the reaction graph $\mathcal{G}_U$ and establish their motif vocabulary and motif-atom correspondences. Consequently, the motif sampling is now across the entire reaction graph representing a chemical equation rather than a single molecule. As depicted in Figure 2, products and reactants typically contain some motifs that are either similar or identical. This design aligns with the earlier example, where the functional groups from other molecules in a chemical reaction can be leveraged to learn the representation of the masked motifs.

3.4. Model Architecture

To fully exploit the information from the reaction graphs, the commonly used message-passing graph neural networks

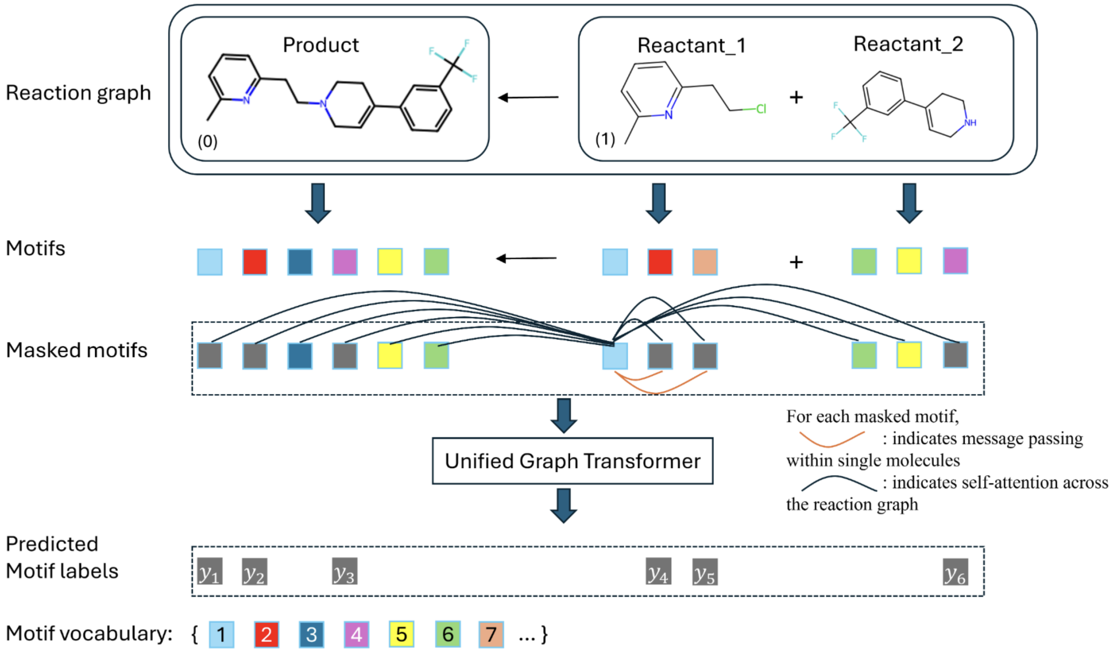
Reaction graph
Product
(0)
Reactant_1
Reactant_2
(1)
+
Motifs
Masked motifs
Unified Graph Transformer
For each masked motif,
: indicates message passing within single molecules
: indicates self-attention across the reaction graph
Predicted Motif labels
y_1 y_2 y_3 y_4 y_5 y_6
Motif vocabulary: { 1 2 3 4 5 6 7 ... }

Figure 2. Overview of the self-supervised learning framework of ReaCTMask.

(MPNNs) such as GCN (graph convolutional networks) (Kipf & Welling, 2016), GIN (graph isomorphism networks) (Xu et al., 2019), have limitations in their standard form. This is because the message-passing mechanism in these networks only allows information transfer among connected nodes, isolating the nodes in the product graphs from those in the reactant graphs. While it is possible to extend MPNNs with techniques like virtual nodes/edges to connect them in a reaction, these adaptations still face challenges in effectively capturing global interactions and dependencies due to the inherent local nature of message passing.

To address this, we employ a graph transformer to uniformly process these reaction graphs and combine their information in one model architecture. Our model, based on GraphGPS (Rampášek et al., 2022), provides several advantages: Each GPS layer consists of a transformer layer and an MPNN layer. In the transformer layer, the self-attention mechanism allows a node to update its embedding by considering all nodes in the disjoint union $\mathcal{G}_U = (\mathcal{V}_U, \mathcal{E}_U, X_U, E_U)$:

$$\tilde{X}_U^{(k+1)} = \texttt{Attn}^{(k)}(X_U^{(k)}) \tag{3}$$

where $\texttt{Attn}^k$ denotes the global attention layer in the k-th GPS layer.

The MPNN layer, on the other hand, integrates the local information from its adjacent nodes and edges via the message-passing mechanism:

$$\hat{X}_U^{(k+1)} = \texttt{MPNN}^{(k)}(X_U^{(k)}, E_U^{(k)}, \mathcal{E}_U) \tag{4}$$

For each node $x_i \in \mathcal{V}_U$ with neighborhood N_i, $\texttt{MPNN}^{(k)}$ updates its k-th embedding as

$$h_i^{(k+1)} = \texttt{MLP}\left(\texttt{Agg}\{h_i^{(k)}, h_j^{(k)}, e_{ij}^{(k)}\}_{x_j \in N_i}\right) \tag{5}$$

where h and e represent the node and edge embeddings, respectively. For instance, in the GINE layer (GIN with edge features) (Hu et al., 2020), the aggregation function is:

$$\texttt{Agg}\{h_i, h_j, e_{ij}\} = (1+\epsilon)h_i + \sum_{j \in N_i} \texttt{ReLU}(h_j + e_{ij}) \tag{6}$$

Then, the node embeddings will be updated by:

$$X_U^{(k+1)} = \texttt{MLP}^{(k)}(\tilde{X}_U^{(k+1)} + \hat{X}_U^{(k+1)}) \tag{7}$$

In addition, we generate positional encodings to distinguish products nodes $(x_p, 0) \in \mathcal{V}_P \times \{0\}$ and reactants nodes $(x_r, 1) \in \mathcal{V}_R \times \{1\}$, using the learnable embedding matrix $W \in \mathbb{R}^{2 \times \text{emb\_dim}}$:

$$\begin{aligned} \texttt{ReactPE}((x_p, 0)) &= [1, 0]W \\ \texttt{ReactPE}((x_r, 1)) &= [0, 1]W \end{aligned} \tag{8}$$

which is added before the GPS layers.

3.5. Learning objective

We use the motif label prediction as our learning objective. Formally, let $\mathcal{V}_M$ be the set of all nodes in a masked motif M.

We have the collection of node embeddings $\{h_i^{(K)}\}_{x_i \in \mathcal{V}_M}$ in the last layer. The motif embedding of M is then obtained by:

$$h_M = \texttt{Pool}(\{h_i^{(K)}\}_{x_i \in \mathcal{V}_M}) \tag{9}$$

where K is the total number of GPS layers, and `Pool` can be any conventional pooling function used in graph-level pooling, such as `mean`, `sum`, or `set2set`. Finally, we use a multi-layer perceptron `MLP` as the prediction head to obtain the logits from h_M and compute the cross-entropy loss with the motif label y_M:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} \texttt{CE}(\texttt{MLP}(h_M), y_M) \tag{10}$$

Here $\mathcal{M}$ is the collection of all masked motifs in a reaction graph $\mathcal{G}_U$.

4. Experiments

4.1. Experimental Settings

In this section, we present the configuration of our experiments and assess the performance of ReaCTMask on molecular property prediction tasks.

Pre-training Datasets and Downstream tasks. Following the settings outlined in (Hu et al., 2020), we used the 2 million unlabeled SMILES representations of molecules from the ZINC-15 database (Sterling & Irwin, 2015), referred to as ZINC-2m. Originally, it contains only the SMILES representation of single molecules. Thus, to extract the reactants that synthesize the molecules in ZINC, we utilized the open-source retrosynthesis tool, AiZynthFinder (Genheden et al., 2020). We extracted one to three reaction routes for each molecule, depending on the solvability of the routes. To reduce pre-training duration, we sampled 500,000 molecules from ZINC-2m, generated their corresponding reactants, and named this subset ZINC-500k. We then combined it with the original ZINC-2m dataset for pre-training. In addition, we used of USPTO-30k dataset (Schneider et al., 2016), which contains 30,000 pre-existing reaction data, to validate ReaCTMask in various settings due to its smaller size.

We evaluate our pre-trained model on the 8 commonly-used datasets with several binary classification tasks from MoleculeNet (Wu et al., 2018). For each dataset, We employ a single-layer MLP as the prediction head, and report the test ROC-AUC's based on the best validation scores. Each experiment is repeated 10 times with different random seeds to ensure robustness. Furthermore, to obtain deeper insights into whether the reaction context improves the pre-training of the model, we conduct extensive probing experiments: specifically, we freeze the parameters of the graph transformer and train only the prediction head.

Model and Training Configuration. We adopt the GraphGPS (Rampášek et al., 2022) model with the Laplacian positional encoder (LapPE), and random-walk structural encoder (RWSE) (Dwivedi et al., 2021). Our model consists of five GPS layers, each with a hidden dimension of 300. Additionally, we construct a reaction positional encoder before the GPS block, which is described in eq. (8). We set the sampling ratio of motifs to 0.3, and the masking ratio of nodes within a motif to be 0.5. See Figure 3 and Figure 4 for more details on how performance changes with different sampling ratios and motif masking ratios.

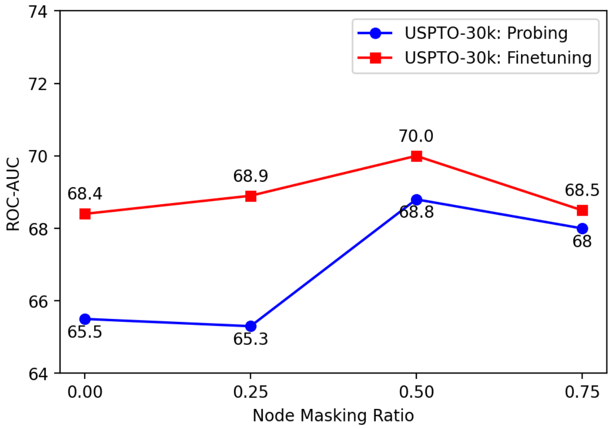

Figure 3. Average ROC-AUC with different atom masking ratio within motifs

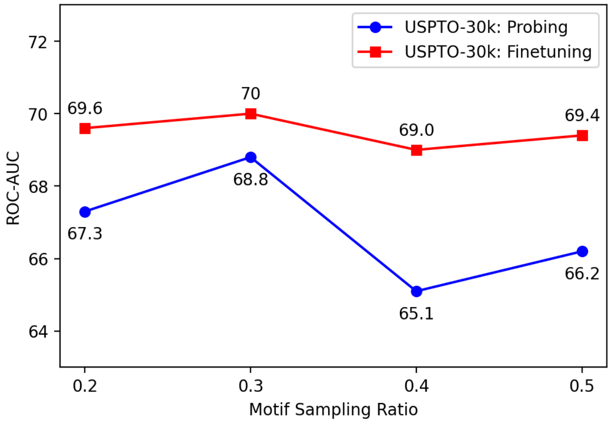

Figure 4. Average ROC-AUC with different motif sampling ratio

In the pre-training phase, we set the total number of epochs to 60 and the batch size to 128, using the Adam optimizer with a learning rate of 0.001. Pre-training is performed on an NVIDIA A6000 GPU, with each epoch taking approximately 30 minutes on ZINC-2m+ZINC500k.

Table 1. Binary classification over MoleculeNet: ROC-AUC; **bold**: best, underline: second best

Num. of Tasks	Tox21 12	ToxCast 617	Sider 27	ClinTox 2	MUV 17	HIV 1	BBBP 1	Bace 1	Average
No pre-train (GINE)	74.6 (0.4)	61.7 (0.5)	58.2 (1.7)	58.4 (6.4)	70.7 (1.8)	75.5 (0.8)	65.7 (3.3)	72.4 (3.8)	67.15
ContextPred (Hu et al., 2020)	73.6 (0.3)	62.6 (0.6)	59.7 (1.8)	74.0 (3.4)	72.5 (1.5)	75.6 (1.0)	70.6 (1.5)	78.8 (1.2)	70.93
GraphCL (You et al., 2020)	75.1 (0.7)	63.0 (0.4)	59.8 (1.3)	77.5 (3.8)	76.4 (0.4)	75.1 (0.7)	67.8 (2.4)	74.6 (2.1)	71.16
JOAO (You et al., 2021)	74.8 (0.6)	62.8 (0.7)	60.4 (1.5)	66.6 (3.1)	76.6 (1.7)	76.9 (0.7)	66.4 (1.0)	73.2 (1.6)	69.71
GraphLoG (Xu et al., 2021)	75.0 (0.6)	63.4 (0.6)	59.6 (1.9)	75.7 (2.4)	75.5 (1.6)	76.1 (0.8)	68.7 (1.6)	78.6 (1.0)	71.56
MGSSL (Zhang et al., 2021)	75.2 (0.6)	63.3 (0.5)	61.6 (1.0)	77.1 (4.5)	77.6 (0.4)	75.8 (0.4)	68.8 (0.6)	78.8 (0.9)	72.28
SimGRACE (Xia et al., 2022)	74.4 (0.3)	62.6 (0.7)	60.2 (0.9)	75.5 (2.0)	75.4 (1.3)	75.0 (0.6)	71.2 (1.1)	74.9 (2.0)	71.15
GraphMVP (Liu et al., 2022)	74.9 (0.8)	63.1 (0.2)	60.2 (1.1)	**79.1** (2.8)	**77.7** (0.6)	76.0 (0.1)	70.8 (0.5)	79.3 (1.5)	72.64
GraphMAE (Hou et al., 2022)	75.2 (0.9)	63.6 (0.3)	60.5 (1.2)	76.5 (3.0)	76.4 (2.0)	76.8 (0.6)	71.2 (1.0)	78.2 (1.5)	72.30
Mole-BERT (MAM) (Xia et al., 2023)	76.2 (0.5)	63.9 (0.3)	61.4 (1.9)	75.1 (3.0)	77.4 (2.1)	77.5 (1.0)	66.8 (1.5)	78.9 (1.1)	72.16
Mole-BERT (TMCL)	74.9 (0.7)	63.2 (0.7)	59.6 (1.4)	77.0 (4.2)	77.2 (0.3)	75.3 (1.1)	67.6 (1.3)	75.1 (1.2)	71.24
Mole-BERT (MAM+TMCL)	76.8 (0.5)	64.3 (0.2)	**62.8** (1.1)	78.9 (3.0)	78.6 (1.8)	**78.2** (0.8)	**71.9** (1.6)	80.8 (1.4)	**74.04**
No pre-train (GPS)	70.0 (0.5)	57.5 (0.7)	57.1 (1.1)	65.7 (2.8)	66.5 (1.9)	66.2 (0.6)	56.7 (1.9)	69.9 (2.7)	63.8
ReaCTMask (ours)	**77.5** (0.4)	**67.3** (0.3)	59.4 (0.3)	78.2 (1.5)	76.7 (1.3)	76.8 (0.1)	68.9 (0.7)	**82.5** (0.7)	73.4

Table 2. Ablation Study - Finetuning: Pre-trained on ZINC

	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	Bace	Average
Node-level mask w/o reaction	74.3 (1.1)	64.8 (0.5)	59.5 (1.1)	70.8 (3.2)	73.5 (1.0)	76.6 (0.5)	67.7 (1.3)	82.6 (1.2)	71.2
Node-level mask w reaction	75.4 (1.0)	65.1 (0.7)	58.9 (0.9)	75.4 (2.0)	74.0 (2.0)	76.5 (1.0)	69.4 (1.8)	82.2 (1.4)	72.1
Motif-level mask w/o reaction	76.5 (0.9)	67.1 (0.6)	59.0 (0.7)	71.0 (0.8)	77.3 (1.3)	76.1 (1.1)	68.5 (0.9)	82.2 (1.1)	72.2
Motif-level mask w reaction (ReaCTMask)	77.5 (0.4)	67.3 (0.3)	59.4 (0.3)	78.2 (1.5)	76.7 (1.3)	76.8 (0.1)	68.9 (0.7)	82.5 (0.7)	73.4
MolR pre-trained on ZINC500k	71.4 (0.6)	61.1 (0.5)	57.7 (2.1)	63.1 (4.7)	71.8 (1.5)	72.0 (1.1)	63.4 (1.5)	77.8 (1.8)	67.3
ReaCTMask pre-trained on ZINC500k	76.3 (0.5)	66.3 (0.5)	57.6 (1.7)	75.7 (3.3)	76.3 (1.3)	76.0 (0.4)	68.5 (0.5)	81.2 (1.2)	72.2

In the finetuning and probing phase on downstream tasks, we set the batch size to 32 and use the scaffold splitting (Ramsundar et al., 2019) to partition our train/validation/test datasets, using the 8:1:1 ratio.

4.2. Results and Analysis

We present the finetuning results of molecular property prediction in Table 1, comparing with other renowned previous works, based on the statistics presented in Mole-BERT (Xia et al., 2023). Note that we do not conduct probing experiments for other methods. Beyond the substantial computational resources and time required for reimplementation and tuning, our decision is also based on the observation that fine-tuning and probing often reflect similar trends, see Figures 3 and 4. For ablation studies, we also evaluate the performance of our method under different settings, see Table 2, Table 3, and 4. We have the following observations:

Observation 1: Reaction context improves model performance. In Table 1, ReaCTMask achieves comparable or superior performance. In Table 2, we observe that the model's performance has improved compared to the one pre-trained without incorporating reaction context. Since other methods utilize GINE as their backbone model while we employ GraphGPS, we also present binary classification results for both GINE and GraphGPS without pre-training. Notably, our method demonstrates a more significant improvement over its respective backbone model. Moreover, in the ClinTox dataset, the performance increases significantly when the reaction context is included in the training data. We can see this more clearly from the probing results in Table 3 and Table 4, where in both pre-training datasets, the reaction context enhances the model's ability to extract more useful features: across the 8 downstream tasks, our method boosts the performance of the majority of tasks while having minimal impact on the remaining tasks.

Observation 2: Motif-label prediction provides a more robust encoding of chemical equations. In Tables 2 and Table 3, we compare ReaCTMask to node-label prediction using the same mask rate for atoms ($0.3 \times 0.5 = 0.15$). The results show that integrating reaction context significantly enhances model performance. However, motif-level prediction consistently outperforms node-attribute masking. This likely occurs because motifs encapsulate richer chemical information compared to the limited atomic types considered in node-attribute masking, leading to more effective and comprehensive molecular representations.

In addition, we also compare ReaCTMask with another reaction-based pretraining method, MolR (Wang et al., 2022), in Table 2 and Table 4. Recall that MolR explicitly hypothesizes the equivalence relation on the graph-level

Table 3. Ablation Study - Linear probing: Pre-trained on ZINC

	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	Bace	Average
Node-level mask w/o reaction	67.9 (0.2)	58.9 (0.1)	51.8 (0.1)	56.8 (0.7)	62.5 (0.6)	63.0 (0.3)	50.2 (0.3)	65.0 (0.6)	59.5
Node-level mask w reaction	71.3 (0.1)	63.1 (0.2)	53.0 (0.3)	65.5 (0.6)	74.9 (0.2)	72.2 (0.1)	55.9 (0.2)	75.4 (0.2)	66.4
Motif-level mask w/o reaction	71.8 (0.1)	63.9 (0.1)	59.8 (1.1)	67.2 (0.1)	75.0 (0.2)	75.0 (0.2)	66.6 (0.2)	80.7 (0.1)	69.8
Motif-level mask w reaction (ReaCTMask)	72.9 (0.1)	64.1 (0.2)	57.2 (0.4)	75.6 (0.5)	77.6 (0.4)	74.4 (0.2)	65.6 (0.2)	80.1 (0.2)	71.0

Table 4. Ablation Study - Linear probing: GPS-based model pretrained on USPTO-30K

	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	Bace	Average
MolR	63.1 (0.1)	52.8 (0.1)	52.9 (0.3)	53.5 (0.1)	56.4 (0.2)	58.1 (0.5)	48.8 (0.2)	63.2 (0.1)	56.1
Motif-level mask + MolR	65.2 (0.3)	58.2 (0.2)	56.6 (0.1)	59.0 (0.5)	65.7 (0.2)	72.1 (0.2)	59.6 (0.3)	74.8 (0.5)	63.9
Motif-level mask w/o reaction	71.4 (0.1)	62.3 (0.1)	60.6 (0.3)	60.8 (0.4)	76.0 (0.2)	69.7 (0.1)	61.9 (0.2)	76.5 (0.3)	67.4
Motif-level mask w reaction (ReaCTMask)	71.1 (0.2)	61.3 (0.2)	60.1 (0.2)	69.5 (0.6)	82.1 (0.4)	70.0 (0.3)	63.7 (0.2)	73.0 (0.2)	68.8

embeddings between reactants and products, as the two views of the same instance. We implement MolR in our own settings, and the results indicate that ReaCTMask is consistently better.

5. Conclusion

In this paper, we introduce a novel approach, ReaCTMask, which leverages reaction context for self-supervised learning in molecular graphs. By exploiting the global attention mechanism and motif-masking strategies, we enable the model to effectively learn from reaction graphs consisting of disjoint reactants and products during the pre-training phase. Our experiments demonstrate significant improvements in molecular property prediction tasks, achieving comparable performance to existing methods. Moreover, the results indicate that masked motif prediction is more favorable than other pre-training tasks used in the reaction context. This study not only highlights the potential of using reaction context in pre-training graph neural networks but also opens new possibilities for joint pre-training schemes for any closely related disjoint graphs.

Future work includes incorporating the proposed method with other empirically validated pre-training techniques, such as the re-mask encoder of node embeddings (Hou et al., 2022) and the first-layer embedding prediction scheme (Liu et al., 2023b). We believe that integrating these techniques could further enhance our method's performance.

References

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Broberg, J., Bånkestad, M. M., and Hellqvist, E. Y. Pre-training transformers for molecular property prediction using reaction prediction. In *ICML 2022 2nd AI for Science Workshop*, 2022.

Coley, C. W., Green, W. H., and Jensen, K. F. Rdchiral: An rdkit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *Journal of chemical information and modeling*, 59(6):2529–2537, 2019.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Dwivedi, V. P., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875*, 2021.

Gao, W., Raghavan, P., Shprints, R., and Coley, C. W. Substrate scope contrastive learning: Repurposing human bias to learn atomic representations. *arXiv preprint arXiv:2402.16882*, 2024.

Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O., and Bjerrum, E. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Jiang, Y., WEI, Y., Wu, F., Huang, Z., Kuang, K., and Wang, Z. Learning chemical rules of retrosynthesis with pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5113–5121, Jun. 2023. doi: 10.1609/aaai.v37i4.25640.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Landrum, G. Rdkit: Open-source cheminformatics. 2006.

Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., Bai, T., Fang, Y., Sun, L., Yu, P. S., et al. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 2023a.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

Liu, Z., Shi, Y., Zhang, A., Zhang, E., Kawaguchi, K., Wang, X., and Chua, T.-S. Rethinking tokenizer and decoder in masked graph modeling for molecules. *Advances in Neural Information Processing Systems*, 36, 2023b.

Luo, Y., Shi, L., and Thost, V. Improving self-supervised molecular representation learning using persistent homology. *Advances in Neural Information Processing Systems*, 36, 2024.

Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., and Beaini, D. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. Deep learning for the life sciences: Applying deep learning to genomics, microscopy. *Drug Discovery, and More*, 1, 2019.

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

Schneider, N., Stiefl, N., and Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12): 2336–2346, 2016.

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, 2021. doi: 10.1038/s42256-020-00284-w.

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Wang, H., Li, W., Jin, X., Cho, K., Ji, H., Han, J., and Burke, M. D. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

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

Xia, J., Wu, L., Chen, J., Hu, B., and Li, S. Z. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference 2022*, pp. 1070–1079, 2022.

Xia, J., Zhao, C., Hu, B., Gao, Z., Tan, C., Liu, Y., Li, S., and Li, S. Z. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023.

Xie, J., Wang, Y., Rao, J., Zheng, S., and Yang, Y. Self-supervised contrastive molecular representation learning with a chemical synthesis knowledge graph. *Journal of Chemical Information and Modeling*, 64(6):1945–1954, 2024. doi: 10.1021/acs.jcim.4c00157. PMID: 38484468.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Xu, M., Wang, H., Ni, B., Guo, H., and Tang, J. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, pp. 11548–11558. PMLR, 2021.

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

You, Y., Chen, T., Shen, Y., and Wang, Z. Graph contrastive learning automated. In *International Conference on Machine Learning*, pp. 12121–12132. PMLR, 2021.

Zeng, L., Li, L., and Li, J. Molkd: Distilling cross-modal knowledge in chemical reactions for molecular property prediction. *arXiv preprint arXiv:2305.01912*, 2023.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

Zhao, Z., Li, Y., Zou, Y., Li, R., and Zhang, R. A survey on self-supervised pre-training of graph foundation models: A knowledge-based perspective. *arXiv preprint arXiv:2403.16137*, 2024.