

A RELATED WORK

Molecular Learning Based on Graphs. Molecular graphs are commonly employed as inputs for molecular learning. Numerous studies employ various GNNs as base encoders to facilitate molecular representation learning for downstream tasks (Yang et al., 2021; Wang et al., 2021a). These approaches generally require supervised signals and are not applicable to other tasks. Recent advancements in research have introduced self-supervised learning methods. Some methods focus on reconstruction tasks for pre-training. PreGNN (Hu et al., 2020) enhances GNN pre-training through context prediction and node/edge attribute masking. GROVER (Rong et al., 2020) introduces molecular-specific self-supervised techniques, including contextual property prediction and graph-level motif prediction. MGSSL (Zhang et al., 2021) adopts a motif-based graph self-supervised strategy that predicts the topology and labels of motifs. Additionally, some methods leverage contrastive learning for pertaining, and general graph augmentation methods (You et al., 2020) are also applicable to molecular datasets. InfoGraph (Sun et al., 2019) optimizes model training by maximizing the mutual information between the global graph representations and their substructures of varying granularity. Incorporating chemical domain knowledge, MoCL (Sun et al., 2021) employs two novel molecular graph augmentation methods: replacing specific substructures or altering a few carbon atoms. MolR (Wang et al., 2022) aims to maintain the equivalence relation between reactants and products within the embedding space. Despite the effectiveness of pre-trained models, it is still challenging to generalize new categories or tasks without labeled examples or fine-tuning.

Fine-grained Vision-Language models. Various works design fine-grained alignment methods to better align the region visual features (RoIs) into text features of the vision-language model (VLM). OV-DETR (Zang et al., 2022) introduces a transformer-based approach for open vocabulary object detection, using conditional binary matching instead of traditional bipartite matching. VLDet (Lin et al., 2022) aligns objects and language by transforming images into regions and captions into words, employing a set matching strategy. DetCLIPv2 (Yao et al.) leverages ATSS (Zhang et al., 2020) for object detection, trained on a standard detection dataset, a grounding dataset, and an image-text pairs dataset. BARON (Wu et al.) aligns embeddings within groups of related regions, processing them through a text encoder. CoDet (Ma et al., 2024) treats region-word alignment as a co-occurring object discovery problem. F-VLM (Kuo et al., 2022) uses a CLIP vision encoder and a VLM feature pooler for region features, combining detection scores with VLM predictions. RO-ViT (Kim et al.) addresses positional embedding gaps in vision-language pre-training by introducing a cropped positional embedding module, enhancing alignment for downstream tasks. While these methods rely on a detection model to assign labels to the RoI, our approach lacks such a model to provide supervised signals for motifs, making it challenging to align fine-grained information.

B TECHNICAL DETAILS OF TEXT-BASED MOLECULE EDITING TASK

Following the molecule editing framework proposed by (Liu et al., 2023a), we utilize FineMolTex and the generation module, including an encoder and a decoder, to generate molecules. This process is structured into two phases. The first phase is the space alignment, which utilizes two projectors, p_m and p_g , to align the representation space of the generative model with our model into a joint representation space following a contrastive learning strategy as:

$$L_{\text{ali}} = -\frac{1}{2}\mathbb{E}_m[\log \frac{\exp(\cos(\mathbf{z}_{\mathbf{m}_0}, p_g(\mathbf{w}))/\tau)}{\exp(\cos(\mathbf{z}_{\mathbf{m}_0}, p_g(\mathbf{w}))/\tau) + \sum_{w'} \exp(\cos(\mathbf{z}_{\mathbf{m}_0}, p_g(\mathbf{w}'))/\tau)} + \log \frac{\exp(\cos(w, p_m(\mathbf{z}_{\mathbf{m}_0}))/\tau)}{\exp(\cos(w, p_m(\mathbf{z}_{\mathbf{m}_0}))/\tau) + \sum_{\mathbf{z}_{\mathbf{m}'_0}} \exp(\cos(\mathbf{w}, p_m(\mathbf{z}_{\mathbf{m}'_0}))/\tau)}], \quad (10)$$

where m is the molecule, $\mathbf{z}_{\mathbf{m}_0}$ denotes the embedding generated by our model, \mathbf{w} is the embedding produced by the encoder of the generation model, $\mathbf{z}_{\mathbf{m}'_0}$ and w' are the embeddings of negative samples sampled from the same batch.

The second phase is latent optimization. We directly learn the latent embedding \mathbf{w}^* , ensuring it remains closely aligned with the initial molecular embedding while also resembling the embedding of the text prompt. We employ two similarity scores as the objective function. To ensure that the generated molecule is similar to the text prompt, we utilize the projector to transform the latent

embedding of the molecule into the joint representation space, and then calculate its cosine similarity with the embedding of text \mathbf{z}_{t_0} . To ensure the generated molecule is also similar to the initial molecule, we compute the l_2 distance between \mathbf{w}^* and the initial embedding \mathbf{w} .

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}^*} (\cos(p_g(\mathbf{w}^*), \mathbf{z}_{t_0})/\tau) + \lambda l_2(\mathbf{w}^*, \mathbf{w}). \quad (11)$$

Given the optimized latent embedding \mathbf{w}^* , the decoder of the generation module can get the output molecules. For the four prompts aimed at generating molecules with specific motifs, we excluded molecules from the dataset that already contained these motifs.

C REPRODUCIBILITY INFORMATION

C.1 DATASET STATISTICS

C.1.1 PRE-TRAINING DATASET

We employ a dataset known as PubChemSTM (Liu et al., 2023a), derived from the PubChem database (Kim et al., 2021), which consists of chemical structure-text pairs. The dataset is available in two versions: PubChemSTM-raw, which retains the original annotations, and PubChemSTM-extracted, where the names of the molecules are replaced with the generic term "this molecule". In our study, we use the PubChemSTM-extracted version. Examples of PubChemSTM-extracted are illustrated in Figure 6.

C.1.2 RETRIEVAL DATASETS

We utilize three pertinent fields from the DrugBank database (Wishart et al., 2018) for each small molecule drug in our zero-shot retrieval task: the Description field, the Pharmacodynamics field, and the Anatomical Therapeutic Chemical (ATC) classification. Specifically, the DrugBank-Description provides a comprehensive overview of the drug’s chemical characteristics, historical development, and regulatory status. The DrugBank-Pharmacodynamics section details the mechanisms through which the drug influences or alters the organism in which it is used. This includes both desired and undesired effects (commonly referred to as side effects). Lastly, the DrugBank-ATC classification system organizes the molecules into groups based on the organs or systems they target, along with their therapeutic, pharmacological, and chemical properties. The datasets consist of 1154, 1005, and 3007 molecule-text pairs for each field, respectively. To test the generalizability of FineMolTex on unseen molecules and texts, given a molecular graph, we select its corresponding molecular description as the positive candidate, and randomly select molecular descriptions from other molecules as negative candidates, aiming to identify the textual description that best aligns with the molecular graph, following (Liu et al., 2023a).

C.1.3 MOLECULE PROPERTY PREDICTION DATASETS

The MoleculeNet dataset, used for molecular property prediction in downstream tasks, contains two main categories. Some datasets are used for pharmacological property prediction. The Blood-Brain Barrier Penetration (BBBP) (Martins et al., 2012) dataset evaluates whether a molecule can penetrate the central nervous system. The three datasets related to toxicity—Tox21 (Challenge, 2014), ToxCast (Wu et al., 2018), and ClinTox (Gayvert et al., 2016)—assess the toxicity of molecular compounds. The Side Effect Resource (SIDER) (Kuhn et al., 2016) dataset contains information on adverse drug reactions from a marketed drug database. Other datasets are used for biophysical property prediction. The Maximum Unbiased Validation (MUV) (Rohrer & Baumann, 2009) dataset, a subset of the PubChem BioAssay (PCBA), is created using a refined nearest neighbor analysis. The HIV dataset, sourced from the Drug Therapeutics Program (DTP) AIDS Antiviral Screen (Zaharevitz, 2015), focuses on predicting the inhibition of HIV replication. The BACE dataset, included in MoleculeNet (Wu et al., 2018), measures the binding affinity of various inhibitors to β -secretase 1 (BACE-1). The overall dataset statistics of MoleculeNet are shown in Table 5.

C.2 BASELINES

The publicly available implementations of Baselines can be found at the following URLs:

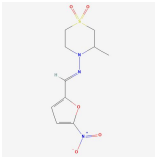
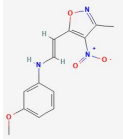
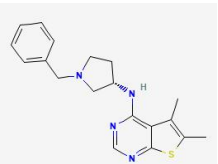
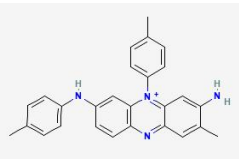
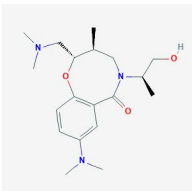
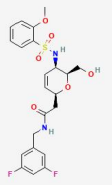
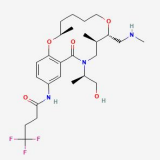
Molecular Graphs	Descriptions
	This molecule is a nitrofurantoin antimicrobial agent used to treat Chagas disease (American trypanosomiasis), a chronic protozoal infection due to <i>Trypanosoma cruzi</i> that can lead to severe disability and death from gastrointestinal and cardiac disease. This molecule is rarely associated with serum aminotransferase elevations during therapy and has not been linked to cases of clinically apparent liver injury.
	This molecule is a substituted aniline and an aromatic ether.
	This molecule is an N-(1-benzylpyrrolidin-3-yl)-5,6-dimethylthieno[2,3-d]pyrimidin-4-amine in which the chiral centre has S configuration. Both enantiomers act as fatty acid synthase inhibitors, although the (S)-enantiomer was found to be more than 4 times as active as the (R)-enantiomer. It has a role as a fatty acid synthesis inhibitor and an EC 2.3.1.85 (fatty acid synthase) inhibitor. It is an enantiomer of a (R)-Fasnall
	This molecule is an organic cation consisting of 7-(4-methylanilino)phenazine carrying additional methyl, amino and 4-methylphenyl substituents at positions 2, 3 and 5 respectively. One of four components of mauvine, a synthetic violet-coloured dye. It has a role as a histological dye. It is an organic cation and a member of phenazines.
	This molecule is a tertiary amino compound and a dialkylarylamine.
	This molecule is a sulfonamide.
	This molecule is a lactam and an azamacrocycle.

Figure 6: Examples on PubChemSTM-extracted

Table 5: Dataset statistics of MoleculeNet for molecule property prediction task.

Dataset	Tasks	Molecules
BBBP	1	2,039
Tox21	12	7,831
ToxCast	617	8,576
Sider	27	1,427
ClinTox	2	1,478
MUV	17	93,087
HIV	1	41,127
Bace	1	1,513

- KV-PLM (MIT license): <https://github.com/thunlp/KV-PLM>
- MoleculeSTM (MIT license): <https://github.com/chao1224/MoleculeSTM>
- MoMu-K and Momu-S (MIT license): <https://github.com/ddz16/MoMu>
- MolCA (MIT license): <https://github.com/acharkq/MolCA>
- AttrMasking (MIT license): <https://github.com/snap-stanford/pretrain-gnns/>
- ContextPred (MIT license): <https://github.com/snap-stanford/pretrain-gnns>
- InfoGraph (MIT license): <https://github.com/sunfanyunn/InfoGraph>
- MolCLR (MIT license): <https://github.com/yuyangw/MolCLR>
- GraphMVP (MIT license): <https://github.com/chao1224/GraphMVP>

C.3 OPERATING ENVIRONMENT

- Operating system: Linux ubuntu 5.15.0-102-generic.
- CPU information: Intel(R) Xeon(R) Platinum 8358 CPU @2.60GHz.
- GPU information: NVIDIA A800 80GB.

C.4 IMPLEMENTATION DETAILS

We use Pytorch to implement our model. For the pre-trained baselines, we directly utilize the checkpoints provided by the authors. For other baselines, we utilize the original codes from their authors and train the models in an end-to-end way. We utilize the BRICS (Degen et al., 2008) to fragment molecules into motifs. As BRICS primarily cleaves bonds according to a predefined set of chemical reactions, often resulting in several large fragments per molecule, we further utilize the post-processing procedure proposed by MGSSL (Zhang et al., 2021), which is designed to minimize the number of ring variants and facilitate the cleavage of side chains. Subsequently, we build a vocabulary of all motif tokens identified in the PubChemSTM dataset, which comprises a total of 30,080 unique motifs. We note that within our dataset, certain motifs are infrequently present, while others are prevalent but lack significant semantic value. This uneven distribution presents a challenge for the model, as it struggles to extract meaningful insights from such disparate occurrences. Thus we have constructed a masking set of 2,457 motifs that excludes motifs appearing fewer than 8 times or more than 80,005 times. Based on this masking set, we selectively mask motifs to enhance the model’s learning efficiency and focus on extracting valuable information from the most informative tokens. To map the embedding of the molecule and the text into the same dimension, we further utilize a projector MLP layer for the graph encoder. We utilize Adam as the optimizer for the GNN encoder, the language model, the two projector MLPs, the multi-modal framework, and the motif classifier with different learning rates, and test the learning rate ranging from $\{1e-3, 1e-4, 1e-5\}$. We set $\ell = 2$, and test ℓ_{trm_M} ranging from $\{1, 2, 3, 4, 5\}$, ℓ_{trm_T} ranging from $\{8, 10, 12\}$. For the coefficient of loss function, we test α and β ranging from $\{0.5, 1, 2\}$. We train our model with a total batch size of 16 for 15 epochs. For fair comparisons, we randomly run 5 times and report the average results for all methods. The code, checkpoints, and optimal parameters can be found in the supplementary material and the anonymous repository <https://anonymous.4open.science/status/FineMolTex-2266>.

Optimal Parameters. The optimal parameters can be found in the config.json file in our code repository. Specifically, we set the dimension of the graph encoder to 300 and the dimensions of the

Table 6: Accuracy ($\% \pm \sigma$) of graph-text retrieval task on DrugBank-Description.

T	Given Molecular Graph			Given Text		
	4	10	20	4	10	20
KV-PLM	73.80 \pm 0.00	53.96 \pm 0.29	40.07 \pm 0.38	72.86 \pm 0.00	52.55 \pm 0.29	40.33 \pm 0.00
MolCA	93.75 \pm 0.09	87.25 \pm 0.06	82.77 \pm 0.12	90.71 \pm 0.04	84.97 \pm 0.16	77.53 \pm 0.15
MoMu-S	76.52 \pm 0.12	61.66 \pm 0.25	50.00 \pm 0.08	77.62 \pm 0.06	61.49 \pm 0.15	52.20 \pm 0.13
MoMu-K	74.15 \pm 0.08	57.18 \pm 0.16	47.97 \pm 0.14	77.79 \pm 0.12	62.33 \pm 0.18	47.97 \pm 0.06
MoleculeSTM	99.15 \pm 0.00	97.19 \pm 0.00	95.66 \pm 0.00	99.05 \pm 0.37	97.50 \pm 0.46	95.71 \pm 0.46
FineMolTex	99.58\pm0.05	97.97\pm0.00	96.45\pm0.16	99.58\pm0.04	97.89\pm0.08	96.11\pm0.12

transformer and cross-attention layer to 768. The hyperparameters are set as $\alpha = 0.5$ and $\beta = 1$. For $\ell = 2$, we set $\ell_{\text{trm}M}$ to 2 for the first round and 3 for the second round, and $\ell_{\text{trm}T}$ to 10 for the first round and 12 for the second round. The learning rates are set as follows: 1e-5 for the graph encoder, text encoder, transformer layers, and cross-attention layer; 3e-5 for the projector MLP of the graph encoder; and 1e-3 for the motif classifier.

D ADDITIONAL RESUTS

D.1 GRAPH-TEXT RETRIEVAL TASK ON DRUGBANK-DESCRIPTION

The results of the graph-text retrieval task on Drugbank-Description can be found in Table 6. We can observe that our model consistently performs better than baselines.

D.2 VISUAL ANALYSIS OF THE OUTPUT MOLECULES ON 4 TEXT-BASED MOLECULE EDITING TASKS

We present additional visual analyses of the output molecules for four text-based molecule editing tasks, as depicted in Figure 8. FineMolTex effectively generates motifs that align with expectations, demonstrating its superior ability to learn fine-grained knowledge.

D.3 VISUALIZATION OF MOTIF AND WORD TOKENS WITH LEGENDS

Figure 7 displays the complete visualization with legends, illustrating that the embeddings of relevant motifs and words are closer in the embedding space.

D.4 EXPLANATION OF PREDICTIONS ON MASKED MOTIFS BASED ON WORD TOKENS

We provide additional explanations for the predictions of FineMolTex in Figure 9, verifying whether the words with the highest interpretive weights are relevant to the masked motifs. In (1), "indanone" is directly the name of the masked motif. In (2), the masked motif is part of "azobenzene". In (3), "azetidine" is part of the masked motif. In (4), the masked motif is a substructure of "trimethylamino" and "butyrobetaine". These explanations demonstrate that FineMolTex has learned fine-grained knowledge, providing insights for the prediction task.

D.5 PRE-TRAINING AND INFERENCE TIMES

We compare the pre-training and inference times of FineMolTex with MoleculeSTM and MolCA. For inference, we conduct experiments on the zero-shot graph-text retrieval task on DrugBank-Pharmacodynamics. Both MoleculeSTM and FineMolTex are tested on one NVIDIA A100 40 GB GPU, while MolCA is tested on two NVIDIA A100 40 GB GPUs. As shown in Table 8, the pre-training and inference times for FineMolTex are comparable to those of the baselines. Despite the longer pre-training time, we believe this trade-off is justified by the SOTA performance of FineMolTex in various downstream tasks.

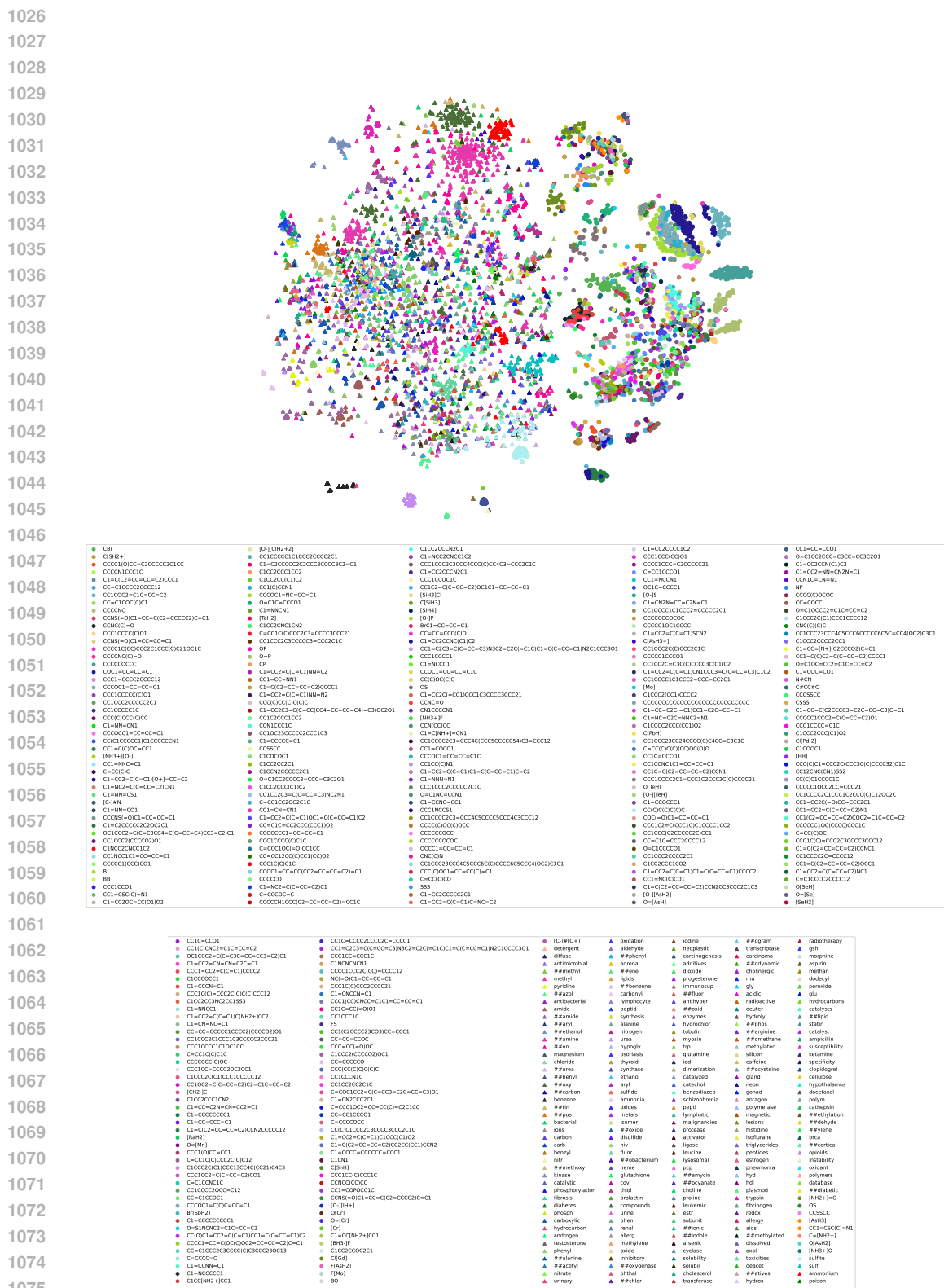


Figure 7: Visualization of motif tokens and word tokens using t-SNE with legend.

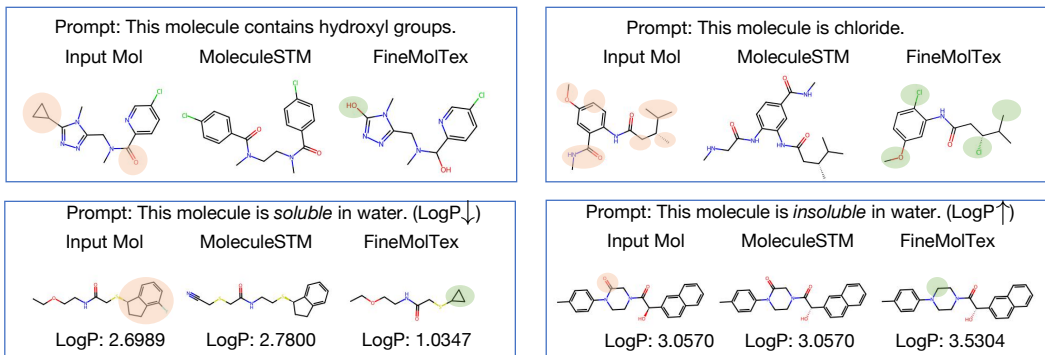


Figure 8: Additional visual analysis of the output molecules of MoleculeSTM and Motif-MolTex on 4 text-based molecule editing tasks. Differences between the input molecule and output molecule of FineMolTex are highlighted in red and green circles. Higher LogP indicates lower water solubility.

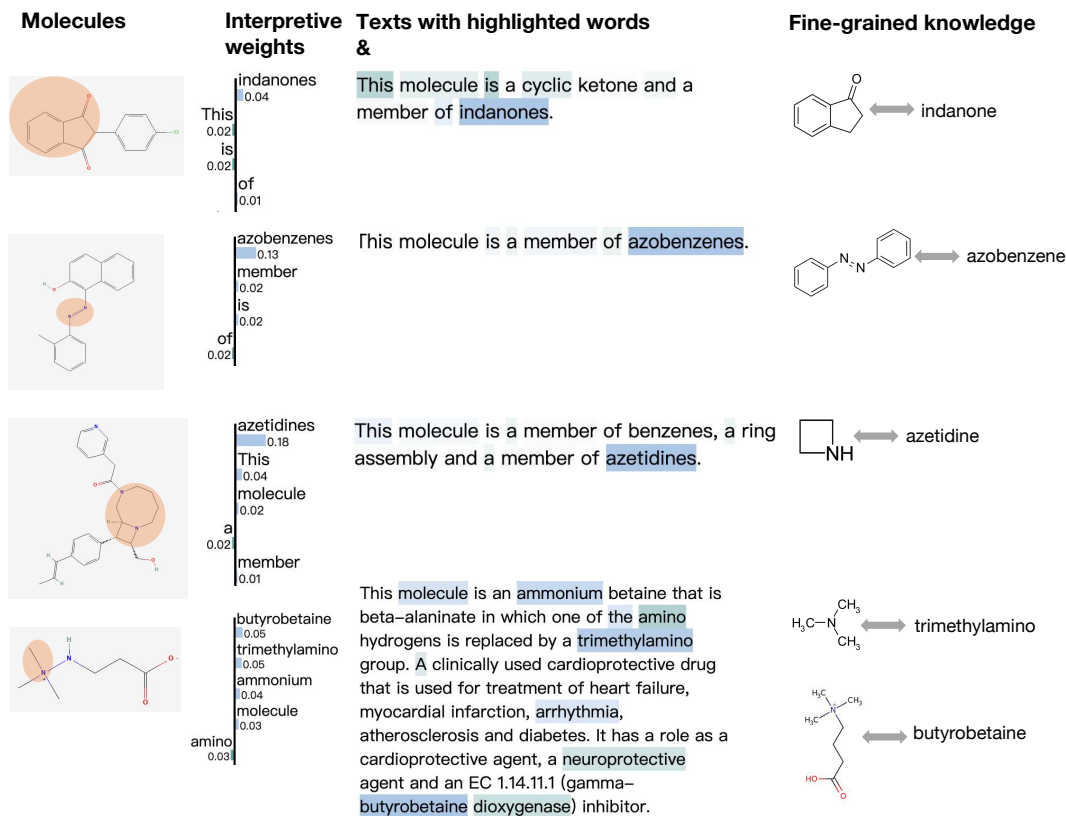


Figure 9: Explanations of prediction of motifs based on word tokens.

D.6 ADDITIONAL ABLATION STUDY

In FineMolTex, all components are trained during the pre-training phase. Although the graph and text encoders are initially pre-trained, it is essential to further train the two encoders to adapt to motif-level pre-training. To validate this necessity, we conduct experiment where FineMolTex is pre-trained with the encoders fixed and then evaluated its performance on the graph-text retrieval task. As demonstrated in Table 7, fixing the graph and text encoders result in a performance decline. This decline occurs because the initial graph and text encoders cannot capture motif-level details.

Table 7: Performance of FineMolTex variants on molecule-ATC and DrugBank-Pharmacodynamics datasets.

	molecule-ATC		DrugBank-Pharmacodynamics	
	Given Molecular Graph	Given Text	Given Molecular Graph	Given Text
FineMolTex (Fixed graph and text encoder)	65.28 \pm 0.03	62.61 \pm 0.05	82.34 \pm 0.10	83.45 \pm 0.06
FineMolTex	75.43\pm0.15	75.22\pm0.12	95.86\pm0.34	95.80\pm0.06

Table 8: Comparison of pre-train and inference times of FineMolTex, MolCa, and MoleculeSTM.

	Device	Pre-train Time	Inference Time
MoleculeSTM	1 A100 40G	65h	47s
MolCA	2 A100 40G	33h	71s
FineMolTex	1 A100 40G	90h	87s