Molecule Language Model with Augmented Pairs and Expertise Transfer

Namkyeong Lee[1,2†], **Siddhartha Laghuvarapu**[2], **Chanyoung Park**[1*], **Jimeng Sun**[2*]
[1] Korea Advanced Institute of Science and Technology
[2] University of Illinois Urbana-Champaign
{namkyeong96, cy.park}@kaist.ac.kr
{sl160, jimeng}@illinois.edu

Abstract

Understanding the molecules and their textual descriptions via molecule language models (MoLM) recently got a surge of interest among researchers. However, unique challenges exist in the field of MoLM due to 1) a limited amount of molecule-text paired data and 2) missing expertise that occurred due to the specialized areas of focus among the experts. To this end, we propose AMOLE, which 1) augments molecule-text pairs with structural similarity preserving loss, and 2) transfers the expertise between the molecules. Extensive experiments on various downstream tasks demonstrate the superiority of AMOLE in comprehending molecules and their descriptions, highlighting its potential for application in real-world drug discovery. The source code for AMOLE is available at https://github.com/Namkyeong/AMOLE.

1 Introduction

Recently, Language Models (LMs) have gained traction in molecular science, fueled by the abundance of literature in this domain. Notably, the conventional method of representing molecules as strings, such as SMILES strings (Weininger, 1988), facilitates the integration of two different modalities, i.e., text and molecules, into a single LM. Then, following the masked language modeling (Devlin et al., 2018), Zeng et al. (2022) train the model on masked SMILES and text from the scientific literature. Moreover, inspired by T5 models (Raffel et al., 2020), training the Molecule Language Model (MoLM) with multiple tasks and fine-tuning the downstream task has been proposed (Edwards et al., 2022; Pei et al., 2023; Christofidellis et al., 2023). However, all these models require molecules to be represented in a string format, like SMILES, to be understood by language models.

As an alternative approach, one can treat molecules and language as multiple modalities, following the recent success of the vision language model (VLM). For instance, several recent works (Edwards et al., 2021; Liu et al., 2023a; Seidl et al., 2023) have proposed using separate encoders for each modality and training the paired modalities to be similar in the representation space using contrastive learning, inspired by the CLIP model (Radford et al., 2021). These models are effective in capturing complementary information between different modalities by learning a joint representation space that can be utilized for various downstream tasks such as cross-modal retrieval and molecular property prediction.

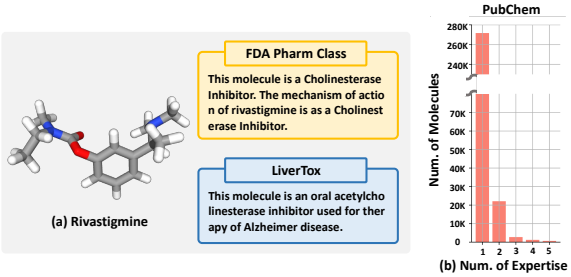

Figure 1: (a) Rivastigmine's textual descriptions from various experts. (b) The majority of molecules in the PubChem database have only one description provided by an expert.

Despite promising early strides, the progress of MoLM lags far behind its VLM counterparts due to the scarcity of molecule-text paired data, both in *quantity and expertise*. First, in terms of quantity, the VLM community largely follows the viewpoint that scale is everything, as image-text pairs are widely available on the web (Wang et al., 2023). Pre-training VLM on the massive amounts of crawled image-text pairs, ranging from tens of thousands to billions, consistently leads to significant performance gains in various downstream tasks (Jia et al., 2021; Yu et al., 2022). In contrast, MoLM faces a bottleneck due to the *limited amount of molecule-text paired data* available, which demands costly domain knowledge and significant time investment for wet lab experiments.

* Corresponding Author. † Work done while the author was a visiting Ph.D. student in UIUC.

Moreover, each molecule has various descriptions provided by different experts, each focusing on its unique areas of expertise. For example, in the case of a molecule named Rivastigmine (Figure 1 (a)), FDA explains its function as a cholinesterase inhibitor, whereas LiverTox (Hoofnagle, 2013) focuses on its effectiveness on Alzheimer's disease. However, due to the specialized areas of the experts, they typically restrict their knowledge to a selective group of molecules, resulting in numerous molecules having *missing expertise* across various experts. Specifically, our examination of the largest molecule database, PubChem (Kim et al., 2021), as illustrated in Figure 1 (b), reveals that only 27K out of 299K molecules have descriptions from multiple experts. Consequently, the remaining 272K molecules are each documented with a single expert description, which may lack comprehensive expertise about the molecule.

In this paper, we propose AMOLE which addresses the unique challenges faced in MoLM by Augmenting MOLecule-text pair and transferring Expertise between the molecules. To overcome the lack of molecule-text paired data, we utilize the idea that molecules with similar structures have similar properties, which is grounded in the well-established biochemical principle (Martin et al., 2002). Specifically, we propose to augment molecule-text pairs by sharing descriptions among structurally similar molecules. However, this approach can lead to false positives, as the descriptions may not be specifically written for structurally similar molecules. To address this issue, we introduce a novel loss function that preserves structural similarity, guiding the model to align the augmented molecule-text pairs more closely based on the structural similarity of the molecules.

Furthermore, to address the missing expertise issue, we utilize the fact that different areas of expertise are interrelated, allowing us to infer additional expertise based on one known area. As an example, in Figure 1 (a), since the Cholinesterase inhibitor is widely known to improve communication between nerve cells by increasing levels of Acetylcholine in the nervous system (Grossberg, 2003), one could infer Livertox's expertise about Alzheimer's disease from FDA descriptions. To this end, we propose to transfer the expertise acquired from molecules with extensive descriptions to those with less description. Specifically, given a molecule that possesses descriptions from multiple experts, we train the model to reconstruct one description from another, thereby enhancing the ability to deduce expertise from one expert using information from another. With our proposed training strategy, the model behaves as if it has access to abundant expertise, even if the model faces the molecule with missing expertise.

We make the following contributions:

- To increase the *limited amount* of molecule-text paired data, we propose to selectively share descriptions among molecules with a novel loss function based on their structural similarity.
- To address the issue of *missing expertise*, we propose to transfer the expertise between molecules by enhancing the model's ability to reconstruct one description from another.
- Extensive experiments including two novel and practical tasks; zero-shot question and answering, and zero-shot virtual screening, demonstrate the superiority and potential applicability of AMOLE in real-world drug discovery process.

2 Related Works

MoLM with a Single Language Model. Thanks to the wealth of literature and the traditional string-based representation of molecules, such as SMILES, LMs have been applied to the domain of molecular science. Drawing inspiration from the masked language model approach used in BERT training (Devlin et al., 2018), KV-PLM (Zeng et al., 2022) proposes to train LMs by reconstructing masked SMILES and textual data. Moreover, MolT5 (Edwards et al., 2022) proposes to pre-train the LMs with the "replace corrupted spans" objective (Raffel et al., 2020) on both SMILES string and textual data, followed by fine-tuning for tasks such as molecule captioning and generation. Pei et al. (2023) and Christofidellis et al. (2023) extend MolT5 with various pre-training tasks, such as protein FASTA reconstruction and chemical reaction prediction. However, these models rely on string-based representations of molecules, e.g., SMILES, which are broadly recognized for their lack of topology awareness (Rong et al., 2020). Furthermore, merging two modalities into a single model prevents the adoption of existing pre-trained models tailored for each modality (Liu et al., 2023a).

MoLM with Multi-Modal Contrastive Learning. Inspired by the recent success of VLM, researchers have started to conceptualize molecules and text

as separate modalities, particularly by adopting separate encoders for each modality. As a pioneering work, Text2Mol (Edwards et al., 2021) proposes training separate encoders for molecular graph and textual description with cross-modal contrastive learning. Following this, CLAMP (Seidl et al., 2023) suggests employing contrastive learning for predicting activities based on the textual description of the task. Furthermore, MoleculeSTM (Liu et al., 2023a) develops the largest multi-modal dataset sourced from the PubChem database for cross-modal contrastive learning applications. It is essential to recognize that the distinction between these models is based on their architectural design and the data used for training rather than the training loss itself. On the other hand, MoMu (Su et al., 2022) introduces an intermolecular contrastive loss along with random molecular augmentations like node dropping, which may lead to chemically invalid structures (Lee et al., 2022a). Unlike previous works, AMOLE concentrates on overcoming the specific hurdles encountered in MoLM: the *lack of abundance and expertise in molecule-text pairs*, through innovative training loss strategies.

MoLM with Other Multi-Modal Learning. On the other hand, there have been other approaches for integrating molecule text through a novel model architecture. GIMLET (Zhao et al., 2023) suggests encoding graph structure and instructional text directly, without separate graph encoding modules, by utilizing generalized position embeddings. Furthermore, inspired by BLIP-2 (Li et al., 2023) in VLM, MolCA (Liu et al., 2023b) introduces the alignment of two modalities through the Q-Former, allowing the language model to understand 2D molecular graphs. Note that these works are not directly related to our research, as we focus on introducing novel training loss designed for molecule-text pair augmentation and expertise transfer.

3 Preliminaries

3.1 Problem Statement

Notations. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent a molecular graph with atoms $\mathcal{V}$ as nodes and the edges $\mathcal{E}$ given by covalent bonds. Moreover, we have a set of textual descriptions $\mathcal{T} = \{t^1, \ldots, t^N\}$ regarding the molecule $\mathcal{G}$ from N different experts, such as FDA and LiverTox, in Figure 1, each of which details various attributes of the molecule. Note that the number of descriptions for each molecule varies, i.e., N_i depends on molecule $\mathcal{G}_i$.

Task Description. Given a molecule graph $\mathcal{G}$ with its textual description $t \in \mathcal{T}$, we aim to train encoders f_{mol} and f_{text} that produce a molecule representation $z_{\mathcal{G}} \in \mathbb{R}^D$ and a textual representation $z_t \in \mathbb{R}^D$, respectively. We aim to obtain the encoders that produce a generalized representation of molecules and text, that can be utilized for a wide range of downstream tasks.

3.2 Tanimoto Similarity

One traditional way of representing a molecule is through fingerprints, a series of binary bits indicating the presence or absence of specific substructures (Rogers and Hahn, 2010). The Tanimoto similarity is a widely accepted criterion for calculating the similarity between two molecules (Bajusz et al., 2015) based on the fingerprints. Specifically, for the molecules $\mathcal{G}_i$ and $\mathcal{G}_j$ that are represented with the fingerprints $\texttt{fp}_i$ and $\texttt{fp}_j$, respectively, the Tanimoto similarity is calculated as follows:

$$s_{ij} = \frac{|\texttt{fp}_i \cap \texttt{fp}_j|}{|\texttt{fp}_i| + |\texttt{fp}_j| - |\texttt{fp}_i \cap \texttt{fp}_j|}. \tag{1}$$

Intuitively, the Tanimoto similarity takes both common and distinct substructures between two molecules into account, thereby offering an assessment of their structural similarity.

3.3 Molecule-Text Contrastive Learning

Previous works (Liu et al., 2023a; Su et al., 2022) have introduced multi-modal contrastive learning to obtain encoders that establish qualified joint space between a molecule $\mathcal{G}$ and its corresponding text t. This approach ensures that paired molecules and texts are closely aligned in the representation space, while unpaired ones remain distant. Specifically, given a molecule $\mathcal{G}$ and its corresponding text t, we first obtain the molecule and text representations as follows: $z_{\mathcal{G}} = f_{\text{mol}}(\mathcal{G})$ and $z_t = f_{\text{text}}(t)$. Then, the model is trained with the following Noise-Contrastive Estimation (InfoNCE) (Oord et al., 2018) loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2}\Big\{\log \frac{\exp(\texttt{sim}(z_{\mathcal{G}}, z_t)/\tau)}{\sum_{t'} \exp(\texttt{sim}(z_{\mathcal{G}}, z_{t'})/\tau)} + \log \frac{\exp(\texttt{sim}(z_t, z_{\mathcal{G}})/\tau)}{\sum_{\mathcal{G}'} \exp(\texttt{sim}(z_t, z_{\mathcal{G}'})/\tau)}\Big\}, \tag{2}$$

where t' and $\mathcal{G}'$ represent all the texts and molecules within the batch, respectively, $\texttt{sim}(\cdot, \cdot)$ indicates the cosine similarity between the representations, and τ denotes the temperature hyperparameter.

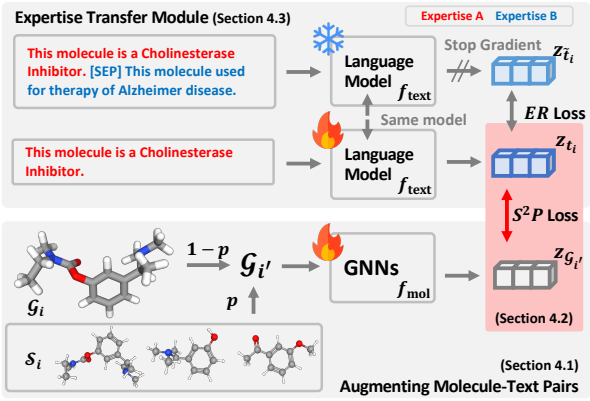

Figure 2: Overall model architecture of AMOLE.

4 Methodology

In this section, we introduce our proposed method, called AMOLE, a novel molecule-text contrastive learning approach that addresses the two unique challenges faced in MoLM, i.e., scarcity of molecule-text paired data in quantity and expertise. Figure 2 presents the overall model architecture.

4.1 Augmenting Molecule-Text pairs

While image-text pairs can be easily generated by the general public, molecule-text pairs are typically produced by specialized experts, making it more challenging to augment the data with external sources such as Web. To this end, we propose to augment the molecule-text pair by sharing the textual description among molecules within the data. Building on the well-established biochemical principle that structurally similar molecules often display analogous biological activities (Martin et al., 2002), we propose to share textual descriptions among molecules with structural resemblances. Specifically, we begin by computing all the pairwise structural similarity, i.e., Tanimoto similarity (Eq. 1), between the molecules in the training set. Then, given the similarity information of a molecule $\mathcal{G}_i$ to other molecules, we identify the top k molecules that exhibit the highest similarity to $\mathcal{G}_i$. We represent these molecules as the set $\mathcal{S}_i$. During training, a molecule $\mathcal{G}_{i'}$ is randomly selected from the set $\mathcal{S}_i$ and substituted for the original molecule $\mathcal{G}_i$ according to a predetermined probability p; otherwise, it is kept identical to the original molecule $\mathcal{G}_i$. By doing so, the textual description t_i of molecule $\mathcal{G}_i$ is shared among k molecules within the set $\mathcal{S}_i$, generating a new molecule-text pair ($\mathcal{G}_{i'}$, t_i). As a result, this approach effectively expands the initial n molecule-text pairs to $n \cdot k$ pairs.

4.2 Structural Similarity Preserving Loss

While we increase the number of pairs by sharing the descriptions, training the model with augmented pairs is challenging since the shared description t_i is not specifically written for the substituted molecule $\mathcal{G}_{i'}$. Consequently, employing traditional InfoNCE loss in Eq. 2 for model training could make the model prone to false positives.

To address this issue, we propose structural similarity preserving (S^2P) loss, designed to preserve molecules' structural similarity in molecule-text joint space. That is, given a text t_i and a molecule $\mathcal{G}_{i'}$, instead of directly defining the pair as positive, we propose to utilize the Tanimoto similarity $s_{ii'}$ between $\mathcal{G}_i$ and $\mathcal{G}_{i'}$ as a pseudo label for contrastive learning. In particular, for a given molecule $\mathcal{G}_i$, we compile a set of $s_{ij'}$ values, where $j' = 1, \ldots, N_{\text{batch}}$ represents the range of molecules within a batch. Then, the pseudo label is calculated as follows:

$$y_{ij'}^{t\to m} = \frac{\exp(s_{ij'}/\tau_1)}{\sum_{k'=1}^{N_{\text{batch}}} \exp(s_{ik'}/\tau_1)}, \tag{3}$$

where τ_1 is a temperature hyperparameter. Then, we make a prediction on the pseudo label based on the similarity between text t_i and molecule $\mathcal{G}_{j'}$ in representation space as follows:

$$\hat{y}_{ij'}^{t\to m} = \frac{\exp(\texttt{sim}(z_{t_i}, z_{\mathcal{G}_{j'}})/\tau_2)}{\sum_{k'=1}^{N_{\text{batch}}} \exp(\texttt{sim}(z_{t_i}, z_{\mathcal{G}_{k'}})/\tau_2)}, \tag{4}$$

where τ_2 is a temperature hyperparameter. It's worth noting that in order to approximate the representation similarity $\texttt{sim}(z_{t_i}, z_{\mathcal{G}_{j'}})$ between a molecule and text to structural similarity $s_{ij'}$, we generate pseudo labels by applying softmax normalization to structural similarity. Then AMOLE is trained with the following cross-entropy loss:

$$\mathcal{L}_{S^2P}^{t\to m} = -\frac{1}{N_{Batch}} \sum_{i=1}^{N_{batch}} \sum_{j'=1}^{N_{batch}} y_{ij'}^{t\to m} \log \hat{y}_{ij'}^{t\to m}. \tag{5}$$

By doing so, the structural similarity between molecules $\mathcal{G}_i$ and $\mathcal{G}_{i'}$ instructs the model on how to align the text t_i and molecule $\mathcal{G}_{i'}$ closely in the representation space, enabling the selective sharing of descriptions based on structural similarities. For the symmetricity, we compute $\mathcal{L}_{S^2P}^{m\to t}$ and derive our final training loss: $\mathcal{L}_{S^2P} = \mathcal{L}_{S^2P}^{t\to m} + \mathcal{L}_{S^2P}^{m\to t}$.

4.3 Expertise Transfer Module

Experts in the field often have a specific area of focus, which results in some molecules lacking comprehensive expertise coverage from different specialists. To address this issue of missing expertise,

we suggest transferring the expertise gained from molecules with extensive descriptions to those with less description. Specifically, we use a molecule $\mathcal{G}_i$ with a set of descriptions $\mathcal{T}_i$ obtained from various experts. We train the language model f_{text} to reconstruct the description $t_{i'} \in \mathcal{T}_i$ from a given description t_i. This approach allows the language model f_{text} to become skilled at inferring expertise from one institution based on another. It behaves as if it has access to abundant expertise even when dealing with molecules with missing expertise.

However, reconstructing the description $t_{i'}$ in text space poses a challenge, as our model is designed to learn a qualified representation space without a decoder structure. To this end, we propose a novel expertise reconstruction (ER) loss, which guides the model to reconstruct the description within the representation space rather than directly in the text space. More formally, when we have a textual description t_i and another associated description $t_{i'}$ for a molecule $\mathcal{G}_i$, we formulate the reconstruction target $\tilde{t}_i$ by concatenating the two descriptions with [SEP] token, i.e., $\tilde{t}_i = t_i$ [SEP] $t_{i'}$. Then, the model is trained to minimize the L2 distance as follows:

$$\mathcal{L}_{ER} = -\frac{1}{N_{Batch}} \sum_{i=1}^{N_{batch}} \|f_{\text{text}}(t_i) - \text{SG}(f_{\text{text}}(\tilde{t}_i))\|_2^2, \tag{6}$$

where SG denotes the stop-gradient operation, which halts the propagation of gradients when inputs are structured as $\tilde{t}_i$, thereby preventing the model from acquiring degenerate solutions by disregarding the text following the [SEP] token. With the novel ER training loss, the model simulates having extensive expertise available, even when specific expertise is lacking.

4.4 Model Training

Finally, AMOLE is trained by jointly optimizing two losses, i.e., structural similarity preserving loss and expertise reconstruction loss, as follows:

$$\mathcal{L} = \mathcal{L}_{S^2P} + \alpha \cdot \mathcal{L}_{ER}, \tag{7}$$

where α denotes the hyperparameter for controlling the weight of the expertise reconstruction loss.

5 Experiments

5.1 Experimental Setup

Pretraining Dataset. We pre-train AMOLE with PubChem database (Kim et al., 2021), which is one of the most extensive public molecular databases available. Our pre-training dataset is compiled using the preprocessing script provided in the repository of the previous work (Liu et al., 2023a). However, due to regular updates to the PubChem database, our dataset varies slightly from the previous work, comprising a total of 299K unique molecules and 336K molecule-text pairs. We provide further details on PubChem database and diverse datasets utilized across four distinct downstream tasks in Appendix C.

Implementation Details. One key benefit of treating molecules and text as distinct modalities is the opportunity to leverage powerful pre-trained models specifically designed for each modality. Following a previous work (Liu et al., 2023a), we employ a GraphMVP (Liu et al., 2021) pre-trained checkpoint for graph isomorphism network (GIN) model (Xu et al., 2018) for our molecule encoder f_{mol}. Moreover, for our text encoder f_{text}, we utilize a pre-trained SciBERT (Beltagy et al., 2019), which has been trained on a vast corpus of textual data from the biochemistry domains. We provide further implementation details in Appendix A.

Baseline Methods. We evaluate AMOLE against three single encoder models: MolT5 (Edwards et al., 2022), BioT5 (Pei et al., 2023), and KV-PLM (Zeng et al., 2022), all of which represent molecules as 1D SMILES strings. While T5-based models are not designed for learning representations of molecules and text, we assess their capabilities by examining the hidden representations produced by the encoder model following a previous work (Seidl et al., 2023). Additionally, we compare AMOLE with two separate encoder models: MoMu (Su et al., 2022) and MoleculeSTM (Liu et al., 2023a). While MoMu depicts molecules as 2D graph structures, MoleculeSTM introduces two models that utilize both SMILES and 2D graph representations for molecules. For the single encoder models, we utilize the checkpoints provided by the authors of the original papers. However, for models with separate encoders, we independently pre-train models with the same pre-training data and model architecture for fair comparison. We provide further details on baseline methods in Appendix B.

5.2 Zero-Shot Cross-Modal Retrieval

Task Description. This task requires choosing an appropriate description from several alternatives for a specific molecule (**Given Molecule**) or re-

	SMILES	Graph	Given Molecule @ 20			Given Text @ 20		
			Descr.	Pharma.	ATC	Descr.	Pharma.	ATC
Single Encoder								
MolT5	✓	✗	5.06 (0.44)	6.80 (0.28)	6.48 (0.25)	6.66 (2.02)	6.02 (0.57)	6.10 (0.09)
BioT5	✓	✗	6.47 (0.13)	7.42 (0.52)	7.71 (0.16)	6.02 (0.42)	7.36 (0.13)	6.78 (0.45)
KV-PLM	✓	✗	42.28 (3.29)	36.84 (0.53)	30.21 (0.40)	45.64 (2.51)	37.93 (0.62)	33.22 (0.40)
Separate Encoder								
MoMu	✗	✓	**97.39** (0.19)	77.82 (0.54)	51.34 (0.37)	96.84 (0.17)	77.05 (0.28)	47.68 (0.34)
MoleculeSTM	✓	✗	96.70 (0.35)	77.28 (0.94)	52.36 (0.29)	96.22 (0.29)	75.01 (0.49)	50.01 (0.40)
MoleculeSTM	✗	✓	95.87 (1.87)	79.21 (0.75)	52.70 (0.75)	95.82 (0.37)	77.15 (0.74)	48.54 (0.49)
AMOLE	✗	✓	96.48 (2.94)	**81.46** (0.60)	**54.76** (0.57)	**97.20** (0.26)	**80.11** (0.42)	**51.47** (0.56)

Table 1: Model accuracy (%) in zero-shot cross-modal retrieval task. The value within the brackets indicates the variance observed across five trials.

	Aug-ment	S^2P Loss	ER Loss	Descr.	Pharma.	ATC
MoleculeSTM	✗	✗	✗	95.85	78.18	50.62
Ablation 1	✓	✗	✗	96.48	78.65	51.46
Ablation 2	✓	✓	✗	96.65	80.47	51.55
AMOLE	✓	✓	✓	**96.84**	**80.79**	**53.12**

Table 2: Ablation studies results.

	SMILES				Graph		
	MolT5	BioT5	KV-PLM	Molecule STM	MoMu	Molecule STM	AMOLE
Descr.	24.84	27.54	30.07	36.11	38.31	37.97	**39.26**
Pharma.	22.49	27.03	26.68	29.60	29.85	29.52	**31.58**

Table 3: Model accuracy (%) in zero-shot question and answering task results.

trieving the molecule that aligns with a particular description (**Given Text**). Following a previous work (Liu et al., 2023a), our experiments are carried out with varying numbers of choices, specifically 4, 10, and 20 options. We report the models' performance on 20 options in Table 1 of the main manuscript, while the complete set of experimental results is available in the Appendix E.1. We also provide details on the datasets and evaluation scheme in Appendix C.2 and D.1, respectively.

Empirical Results. In Table 1, we have the following observations: **1)** As T5-based models (i.e., MolT5 and BioT5) are not intended to build joint representation space between molecule and text, their performance was the worst among the models. This suggests that T5-based models struggle to learn suitable representations and necessitate a costly fine-tuning process after the pre-training step. **2)** Additionally, despite the SMILES-based MoleculeSTM having a significantly higher number of parameters compared to the graph-based model, [1] SMILES and graph structure representations of molecules yield comparable results. This underscores the efficacy of graph-based representations in capturing molecular properties. **3)** Within the graph-based models, it is evident that MoMu generally exhibits the lowest performance across all but one dataset. This is largely due to the random augmentation of molecules, such as by dropping nodes, without taking the chemical validity of the molecules into account (Lee et al., 2022b), which makes the model perform even worse than MoleculeSTM. On the other hand, in the upcoming Ablation studies, we will demonstrate that augmenting molecule-text pairs with chemically valid molecules can consistently improve MoleculeSTM performance. **4)** Overall, we observe that AMOLE performs the best on five out of six datasets, demonstrating its ability to comprehend and bridge two heterogeneous modalities.

Ablation Studies. Now, we empirically evaluate the impact of each component in AMOLE by sequentially removing them one at a time. We assess the effectiveness of each component by calculating the average performance across hard cross-retrieval tasks, including retrieving among 20 texts given a molecule and retrieving among 20 molecules given a text. In Table 2, we have the following observations: **1)** By comparing MoleculeSTM and Ablation 1, we observe that augmenting molecule-text consistently brings performance gain, indicating that "scale-is-everything" perspective might also be applicable and beneficial within the MoLM community. **2)** Additionally, a comparison between Ablation 1 and Ablation 2 reveals that incorporating a S^2P loss can further enhance the performance. This indicates that aligning molecules and text based on the structural similarity among molecules can effectively mitigate issues related to false positives. **3)** Lastly, it is noted that the ER loss uniformly enhances performance across all datasets, with a particularly notable improvement in the ATC dataset. This improvement can be attributed to the nature of the ATC dataset, which comprises labels for a molecule classification system and is inherently abstract. Consequently, reconstructing missing expertise from such abstract descriptions proves advantageous for model performance. We provide further model analysis and statistical significance tests in Appendix E.1.

In summary, our model effectively augments

[1] For MoleculeSTM SMILES, we use ChemBERTa, which contains 83,450,880 parameters, and for MoleculeSTM Graph, we use GIN, which contains 1,885,206 parameters.

	BBBP	Tox21	ToxCast	Sider	Clintox	MUV	HIV	Bace	Avg. Rank
GraphSSL									
AttrMask	68.92 (1.68)	74.86 (0.64)	64.12 (0.22)	59.56 (1.51)	86.52 (0.53)	75.68 (1.71)	75.79 (1.06)	78.71 (2.47)	5.50
ContextPred	66.77 (1.39)	73.95 (0.42)	61.77 (0.67)	54.51 (2.11)	81.45 (3.78)	72.88 (1.80)	66.51 (3.74)	74.94 (6.34)	7.63
GPT-GNN	60.74 (0.32)	72.14 (0.55)	59.55 (0.51)	54.69 (0.31)	55.87 (1.92)	71.74 (1.01)	71.20 (0.52)	73.23 (2.63)	8.63
InfoGraph	66.28 (2.12)	73.12 (0.49)	61.52 (0.56)	57.82 (1.86)	89.14 (3.30)	73.94 (3.72)	77.14 (1.32)	69.41 (0.39)	6.38
GraphMVP	69.86 (0.90)	75.37 (0.20)	**65.21** (0.26)	60.12 (0.60)	87.98 (1.46)	76.61 (0.91)	76.12 (1.04)	79.30 (1.17)	3.25
Mole-BERT	65.50 (1.19)	74.05 (0.52)	64.75 (0.71)	57.09 (1.05)	**92.03** (1.06)	73.95 (1.41)	76.26 (0.67)	76.93 (0.76)	5.50
MoLM w/ Graph									
MoMu	69.70 (0.47)	75.16 (0.34)	64.90 (0.26)	60.21 (0.76)	86.20 (0.97)	76.63 (1.02)	77.10 (1.00)	78.78 (0.90)	3.75
MoleculeSTM	69.08 (0.54)	75.47 (0.29)	64.94 (0.51)	59.60 (0.51)	88.46 (0.99)	75.77 (1.19)	77.96 (0.63)	80.10 (1.16)	3.13
AMOLE	**69.94** (0.84)	**76.19** (0.27)	65.03 (0.27)	**60.69** (0.70)	89.94 (0.96)	**76.76** (0.96)	**78.42** (0.71)	**80.26** (1.80)	**1.25**

Table 4: ROC-AUC performance in molecular property prediction task. The value within the brackets indicates the variance observed across five trials. Bold text denotes the top performance, while an underline highlights the second best. ”Avg. Rank" represents the average ranking across all datasets.

molecule-text pair by incorporating chemically viable molecules and employing a S^2P loss. Additionally, transferring the knowledge through ER loss consistently enhances AMOLE performance, particularly with brief and abstract descriptions.

5.3 Zero-Shot Question and Answering

Task Description. While earlier studies have assessed the retrieval capabilities of MoLM by contrasting the correct description with randomly chosen descriptions from other molecules (Liu et al., 2023a; Su et al., 2022), we propose a novel task, termed the “Zero-shot Question and Answering” task. Specifically, given a textual description of a molecule, we instruct GPT-4 (Achiam et al., 2023) to generate a multiple-choice question comprising five options, all derived from the given textual description. Then, with a generated question and its five corresponding options, we merge the question with each option to form a single input, i.e., $\texttt{input}_i = \text{Concat}(\texttt{question}, \texttt{option}_i)$, where $i = 1, \ldots, 5$. Given a molecule and these combined inputs, we then select the one that includes the correct answer from the options. In Appendix C.3, we offer details for generating and validating questions and answer datasets, along with an evaluation scheme for the task in Appendix D.2.

Empirical Results. In Table 3, we have the following observations: **1)** Comparing to Table 1, we observe that despite having far fewer options for retrieval, most models learning representation space (i.e., KV-PLM, MoMu, MoleculeSTM, and AMOLE) perform much worse in the task. This is because the model must discern based solely on the subtle differences between the options provided, requiring the model to have a more fine-grained understanding of molecules than cross-modal retrieval. **2)** On the other hand, AMOLE consistently outperforms baseline models in this task, showcasing its ability to extract more intricate information from the slight variations in options through the inference of related expertise. In conclusion, we posit that AMOLE offers benefits in tasks that demand a more intricate understanding of molecules, achieved by integrating additional expertise. In Appendix E.2, we further showcase the effectiveness of the expertise transfer module by integrating modified inputs as AMOLE into the baseline methods.

5.4 Molecular Property Prediction

Task Description. In this task, we assess the potential benefits of incorporating external knowledge, i.e., textual descriptions, into the molecule encoder f_{mol} as done in AMOLE, in enhancing the prediction of molecular properties. We mainly compare to recent graph self-supervised learning (**GraphSSL**) approaches, i.e., AttrMask (Hu et al., 2019), ContextPred (Hu et al., 2019), GPT-GNN (Hu et al., 2020), InfoGraph (Sun et al., 2019), MolCLR (Wang et al., 2022), GraphMVP (Liu et al., 2021), and Mole-BERT (Xia et al., 2022), and MoLMs which represent molecules with graph structure (**MoLM w/ Graph**). Following previous works (Liu et al., 2021, 2023a), we pre-train the molecule encoder f_{mol} using each of the proposed strategies, and fine-tuning on MoleculeNet benchmark (Wu et al., 2018). We provide further details on the dataset and evaluation scheme for the task in Appendix C.4 and D.3, respectively.

Empirical Results. In Table 4, we have following observations: **1)** We observe that incorporating external textual descriptions into the pre-training phase uniformly enhances the prediction of molecular properties, as evidenced by improved overall

Dataset	Prompt
HIA	Human intestinal absorption (HIA)
	The molecule is positive w.r.t. a property that is defined as the ability of the body to be absorbed from the human gastrointestinal system into the bloodstream · · ·
Pgp Inhibition	P-glycoprotein Inhibition
	This molecule is known to inhibit P-glycoprotein, which is an ABC transporter protein involved in intestinal absorption, drug metabolism, and brain · · ·
DILI	Inducing liver injury
	This molecule induces liver injury that is most commonly caused by Amoxicillin clavulanate isoniazid, and nonsteroidal anti-inflammatory drugs.
VDR	Vitamin D receptor
	This molecule is active w.r.t. Vitamin D receptor. The best pharmacophore hypothesis contains one hydrogen bond acceptor (A), one hydrogen bond · · ·

Table 5: Examples of abstractive and detailed prompts for each dataset.

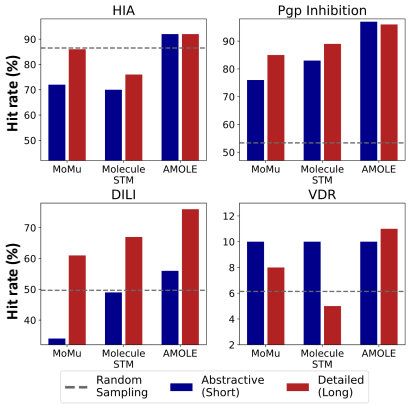

Figure 3: Hit rate (%) in virtual screening task.

performance (i.e., averaged across all eight tasks). This improvement is credited to the implicit influence of external domain knowledge, i.e., the textual descriptions of molecules, which is proved to be beneficial for property prediction. (Liu et al., 2023a). **2)** Among the MoLM, AMOLE outperforms baseline methods on six out of eight tasks, demonstrating that the knowledge can be more efficiently transferred to molecular property prediction through our strategy. In summary, our methodology is advantageous not just for cross-modal tasks, but also for tasks within a single modality, demonstrating the versatility of AMOLE across a wide range of downstream applications.

5.5 Zero-Shot Virtual Screening

Task Description. Virtual screening is a computational technique to search large libraries of compounds quickly to identify those structures most likely to have desired properties, emerging as a principal technique in the drug discovery process (Mehta et al., 2021). Therefore, in this paper, we propose a novel task named "Zero-shot Virtual Screening," where we assess the model's proficiency in virtual screening drugs by supplying a textual description of a desired property. Specifically, our evaluation of the model's capability in virtual screening is conducted by providing two distinct prompts for each property as shown in Table 5: one brief and abstract, and the other longer and more detailed, offering comprehensive information about the property. For each description, we identify the top 100 molecules nearest to the prompt in the representation space and compute the hit rate to evaluate the model's performance. We provide further details on each dataset and used prompts for virtual screening in Appendix C.5 and D.4.

Empirical Results. In Figure 3, we have the following observations: **1)** While previous studies have shown comparable results on various downstream tasks, their effectiveness in virtual screening leaves room for improvement. Notably, these models often yield results inferior to the mere random selection of molecules (i.e., gray dashed line in Figure 3), and their performance significantly fluctuates based on the specific textual description used. **2)** However, AMOLE consistently performs the best in various datasets and different types of descriptions, further demonstrating its ability to learn a more qualified joint space between molecules and text. **3)** One interesting observation is that, AMOLE demonstrates notably robust performance compared to the baseline methods, regardless of whether the description is abstract or detailed. This can be attributed to the expertise transfer module, which equips the model with the ability to deduce related information even when only an abstract level of detail is provided. In conclusion, AMOLE can consistently screen the qualified molecules given any textual description, highlighting the adaptability of AMOLE for real-world drug discovery process. We provide additional experimental results on more datasets, how the expertise transfer module affects the model performance with various levels of α in Appendix E.3.

6 Conclusion

In this paper, we propose AMOLE, addressing the two unique challenges in MoLM, i.e., scarcity of molecule-text paired data in both quantity and expertise. Our extensive testing on four downstream tasks, notably including two innovative and practical tasks such as zero-shot question and answering and zero-shot virtual screening, demonstrate the efficacy of AMOLE in grasping the nuances of molecules and their textual descriptions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George T Grossberg. 2003. Cholinesterase inhibitors for the treatment of alzheimer's disease:: getting on and staying on. *Current Therapeutic Research*, 64(4):216–235.

Jay H Hoofnagle. 2013. Livertox: a website on drug-induced liver injury. In *Drug-induced liver disease*, pages 725–732. Elsevier.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. 2020. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1857–1867.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2021. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2021. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395.

Namkyeong Lee, Junseok Lee, and Chanyoung Park. 2022a. Augmentation-free self-supervised learning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7372–7380.

Namkyeong Lee, Junseok Lee, and Chanyoung Park. 2022b. Self-supervised graph representation learning via positive mining. *Information Sciences*, 611:476–493.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023a. Multimodal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*.

Yvonne C Martin, James L Kofron, and Linda M Traphagen. 2002. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358.

Sarvesh Mehta, Siddhartha Laghuvarapu, Yashaswi Pathak, Aaftaab Sethi, Mallika Alvala, and U Deva Priyakumar. 2021. Memes: Machine learning framework for enhanced molecular screening. *Chemical science*, 12(35):11710–11721.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. *arXiv preprint arXiv:2303.03363*.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*.

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. 2020. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60(9):4263–4273.

Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. 2023. Too large; data reduction for vision-language pre-training. *arXiv preprint arXiv:2305.20087*.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

David Weininger. 1988. 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.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. 2022. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *bioRxiv*, pages 2023–05.

A Implementation Details

In this section, we provide implementation details of AMOLE.

Text Encoder f_{text}. Following previous work (Liu et al., 2023a), we use BERT architecture (Devlin et al., 2018) as the text encoder f_{text}, and adapt the SciBERT (Beltagy et al., 2019) checkpoint to initialize the model parameters [2]. Text encoder contains a total 109,918,464 number of parameters.

Molecule Encoder f_{mol}. We use GIN (Xu et al., 2018) architecture as a molecular encoder f_{mol}, which has been widely used as the backbone model in recent graph self-supervised learning works (Hu et al., 2019; Liu et al., 2021). Additionally, we initialize the encoder parameters using the GraphMVP (Liu et al., 2021) checkpoints provided by the original authors [3]. Molecule encoder contains a total 1,885,206 number of parameters.

Training Details. Our method is implemented on Python 3.7.16, PyTorch 1.10.1, and Torch-geometric 2.0.3. All experiments are conducted using an 80GB NVIDIA A100 GPU. It takes 90 minutes per epoch for training, a total of 2700 minutes for training.

Hyperparameters. We list the key hyperparameters used during training in Table 6.

Hyperparameter	Value
Training epochs	30
Learning rate for text encoder f_{text}	1e-5
Learning rate for molecule encoder f_{mol}	1e-5
Temperature for pseudo label τ_1	0.1
Temperature for model prediction τ_2	0.1
Number of similar molecules k	50
Replacement ratio for original molecule p	$\{0.2, 0.5\}$
Weight of expertise reconstruction loss α	$\{0.1, 1.0\}$

Table 6: Hyperparameter specifications for AMOLE pre-training.

B Baseline Methods

In this section, we elaborate on baseline methods compared during the experiments.

Single Encoder Models. For single encoder models, we mainly compare AMOLE with MolT5, BioT5, and KV-PLM. While T5-based models were not originally created for capturing the representations of molecules and text descriptions, we evaluate their performance by leveraging the hidden representations, aligning with approaches used in prior studies (Seidl et al., 2023).

- **MolT5** (Edwards et al., 2022) stands out as a trailblazer in the field of molecule language models. It introduces a novel approach of pre-training the model on an extensive dataset of unlabeled natural language texts and molecular strings (SMILES), using a denoising objective. Subsequently, the model undergoes fine-tuning for tasks such as molecule captioning and generation. For evaluation, we utilize the checkpoints provided by the authors, which are available in the Huggingface repository [4].
- **BioT5** (Pei et al., 2023) extends MolT5 by incorporating a diverse array of pre-training tasks, such as denoising molecule SELFIES, protein FASTA sequences, general text, wrapped text, as well as translating between bio-sequences and structured text descriptions. We also utilize checkpoints provided by the authors, which are available in the Huggingface repository [5].
- **KV-PLM** (Zeng et al., 2022) introduces a pre-training approach for language models that incorporates masked language modeling on a specialized corpus featuring inserted SMILES strings. We utilize the checkpoint available on the author's Github repository [6].

Separate Encoder Models. For separate encoder models, we mainly compare AMOLE with MoleculeSTM and MoMu. To isolate the impact of the training loss on model performance, we ensure that all models are trained under uniform conditions, using the same training data and model architecture. This approach guarantees that any differences in performance can be attributed to the distinct training losses employed by each model.

- **MoleculeSTM** (Liu et al., 2023a) suggests developing representations for molecules and texts through contrastive learning, where a molecule and its corresponding text are considered a positive pair, and all other combinations within the

[2] https://huggingface.co/allenai/scibert_scivocab_uncased

[3] https://huggingface.co/chao1224/MoleculeSTM/tree/main/pretrained_GraphMVP

[4] https://huggingface.co/laituan245/molt5-large-caption2smiles

[5] https://huggingface.co/QizhiPei/biot5-base-text2mol

[6] https://github.com/thunlp/KV-PLM?tab=readme-ov-file

same batch are treated as negative pairs. Additionally, the authors introduce the extensive dataset, named PubChemSTM, which includes 250K distinct molecules and 281K molecule-text pairs. They introduce two distinct models, one utilizing SMILES representations of molecules as model input and the other as graph representations. For the SMILES model, they initially use the MegaMolBART (Irwin et al., 2022) checkpoint, and for the graph model, the GraphMVP (Liu et al., 2021) checkpoint. However, due to issues with the CUDA environment, we utilize ChemBERTa (Chithrananda et al., 2020) [7] as the initial checkpoint for the SMILES model training and GraphMVP [8] for the graph model in MoleculeSTM.

- **MoMu** (Su et al., 2022) introduces a method of contrasting among molecules themselves with an additional loss function besides contrasting molecules and texts.

Graph Self-Supervised Learning Methods. In Section 5.4, we compare AMOLE to representative graph self-supervised learning approaches, which will be briefly introduced here.

- Attribute Masking (**AttrMask**) (Hu et al., 2019) introduces a technique of randomly masking the attributes of nodes or edges in the input and pre-training a GNN to predict these masked attributes.
- Context Prediction (**ContextPred**) (Hu et al., 2019) proposes to pre-train GNN to ensure nodes situated in analogous structural contexts are represented by proximate embeddings by using subgraphs to predict their surrounding graph structures.
- **GPT-GNN** (Hu et al., 2020) introduces a method for pre-training GNN through the generation of attributed graphs, which is achieved by dividing the generation process into two distinct phases: the generation of node attributes and the formation of edges between nodes.
- **InfoGraph** (Sun et al., 2019) proposes to maximize the mutual information between graph-level representation and node-level representation via contrastive learning.
- **GraphMVP** (Liu et al., 2021) introduces a hybrid training approach that merges generative and contrastive methodologies. In the contrastive setting, 2D molecular graphs and their corresponding 3D structures are considered positive pairs, while all other combinations are treated as negative pairs. On the generative side, the model aims to reconstruct the representation of a 3D structure from its 2D molecular graph counterpart and vice versa, facilitating a comprehensive understanding of molecular structures from both dimensions.

[7] https://huggingface.co/seyonec/ChemBERTa-zinc-base-v1

[8] https://huggingface.co/chao1224/MoleculeSTM/tree/main/pretrained_GraphMVP

C Datasets

C.1 Pre-training

We pre-train AMOLE with the PubChem database, which is one of the most extensive public molecular databases available. PubChem database consists of multiple data sources including DrugBank, CTD, PharmGKB, and more. Please refer to the following URL for more details: https://pubchem.ncbi.nlm.nih.gov/sources/.

The PubChem database we used during training comprises a total of 299K unique molecules and 336K molecule-text pairs. During preprocessing, we consolidate each expertise into a unified description. Thus, each description for an individual molecule originates from a distinct database (expertise). On average, molecules are associated with 1.115 descriptions, with a maximum of 17 descriptions and a minimum of one. We provide a histogram and boxplot on the number of descriptions per molecule in Figure 4. As shown in Figure 4, it is evident that the majority of molecules have a singular description, indicating a lack of comprehensive expertise across various experts for numerous molecules. Moreover, each description in training data consists of 17.62 words on average, with a maximum of 874 words and a minimum of one. We also provide a histogram on the number of words per description in Figure 5.

C.2 Zero-Shot Cross-Modal Retrieval

Following previous work (Liu et al., 2023a), we use the datasets extracted from the Description field, the Pharmacodynamics field, and the anatomical therapeutic chemical (ATC) field in the DrugBank database (Wishart et al., 2018). Each field contains the following information:

- **Description (Descr.)** field provides an overview of **1,154** drugs, including their chemical characteristics, development history, and standing in terms of regulatory approval.

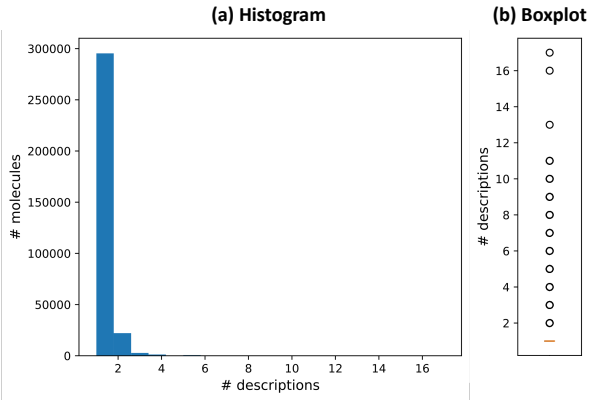

Figure 4: (a) Histogram and (b) Boxplot on the number of descriptions per molecule.

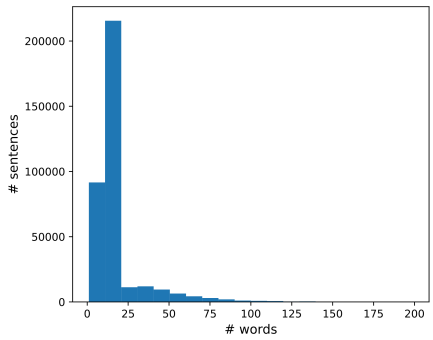

Figure 5: Histogram on the number of words per description.

- **Pharmacodynamics (Pharma.)** field explores the effects and mechanisms of **1,005** drugs on the body, detailing the biochemical and physiological interactions and responses induced by the drug within the organism.
- **ATC** field represents a classification framework that organizes **3,007** molecules based on the organ or system they target and their therapeutic, pharmacological, and chemical characteristics.

We use the datasets provided in the repository from MoleculeSTM [9] for our evaluation purposes. It is worth noting that, we ensure that molecules appearing in the training dataset with identical canonical SMILES are excluded to avoid data leakage. Moreover, for the ATC dataset, exclusion criteria also consider high similarity between textual descriptions in addition to identical canonical SMILES.

C.3 Zero-Shot Question and Answering

We generate questions based on the textual descriptions used for the cross-modal retrieval task in Section 5.2. Please note that we restrict our question generation to descriptions and pharmacodynamics datasets since the ATC dataset consists of brief labels for molecules.

Question and Answer Generation. In Section 5.3, we assess the MoLM's capacity to discern the correct answer from options with minor differences, aiming to evaluate a more nuanced understanding of molecules and their textual descriptions. To achieve this, we employ GPT-4 to craft a multiple-choice question with five options, each based on the textual descriptions of molecules, by providing specific prompts to guide its generation. We provide the precise prompts used for generating questions and answers (QAs) in Figure 7 (a). Consequently, we acquire a sum of 8,215 QA pairs for the description dataset and 7,300 QA pairs for the pharmacodynamics dataset.

Question and Answer Validation. While GPT-4 proficiently generates questions and answers based on the textual descriptions of molecules, there are cases where it produces questions with wrong answers. Consequently, we refine the generated questions and answers by assessing if GPT-4 accurately identifies the same answers as GPT-4 provided using the original textual descriptions and the generated questions. As an example, given the context of the original description and the GPT-4 generated questions, there are cases where the answers generated by GPT-4 and the answers validated by GPT-4 are different, as shown in Figure 8. Since GPT-4 fail to produce consistent answers based on the original context, we consider these instances as unsuccessful question and answer generation, excluding them from our analysis. We provide the specific prompts for validating QAs in Figure 7 (b). After filtering out invalid QA pairs, we obtain a sum of **7,986 QA pairs** for the description dataset and **7,184 QA pairs** for the pharmacodynamics dataset.

C.4 Molecular Property Prediction

For the molecular property prediction task, we use the MoleculeNet benchmark (Wu et al., 2018) for evaluation, which is commonly used for evaluating the machine learning methods for molecular property prediction (Liu et al., 2021). The MoleculeNet benchmark encompasses a diverse array of datasets, each characterized by distinct properties

[9] https://huggingface.co/datasets/chao1224/MoleculeSTM/tree/main/DrugBank_data/raw

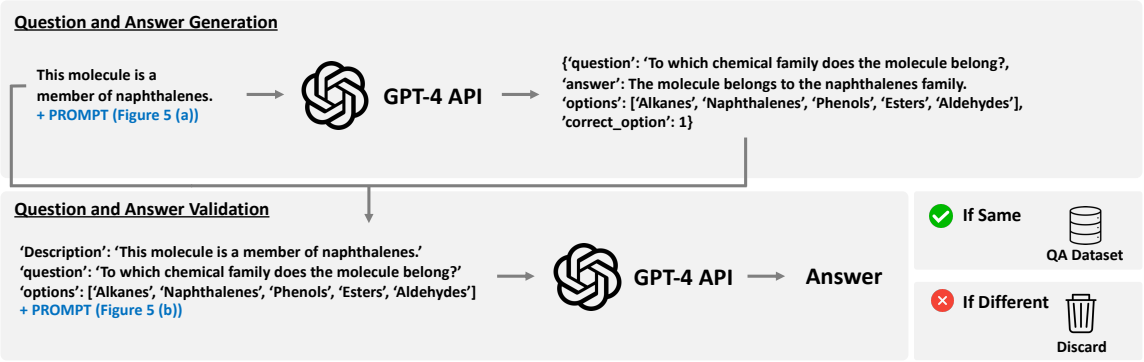

Figure 6: Overall pipeline for generating and validating questions and answering datasets.

(a) Question and Answer Generation

Generate a set of chemically relevant questions from the paragraph. Do not include the name of the molecule in the question. Refer to it as 'the molecule'. The questions should have two k inds of answers. One is a short descriptive sentence answering the question. The second is a set of five options, of which one is correct. Make sure that the questions are chemically relevan t, and of good quality, and chemically/biologically relevant content in the paragraph are covered in the questions. The answers should also be discriminative and should contain chemically relevant options. Question/answer should focus on a specific characteristic. Only one of the options should be correct. Questions and answers should be sourced from the paragraph only.

The output should be JSON formatted and should include only the QA and nothing else. Include as many questions as possible. The JSON should have the following format:

```
{ "questions": [
                { "question": "Description of the question",
                "answer": "Description of the answer",
                "options": [ "optionA", "optionB", "optionC", "optionD", "optionE" ],
                "correct_option": 1 },
                { "question": "Description of the question",
                "answer": "Description of the answer",
                "options": [ "optionA", "optionB", "optionC", "optionD", "optionE" ],
                "correct_option": 2 }
                ]
}
```

Strictly adhere to the format and DO NOT include any additional text with the response.

(b) Question and Answer Validation

You are given a description about molecule and a list of questions following with a list options. For each question, provide the index of the correct option. If the answer cannot be inferred f rom the description or the correct option is not available, output 0 for that question. The output be a list of integers and strictly do not include any other text before or after the answer. Str ictly adhere to the output format.

Example Output: [answer_idx_for_q1, answer_idx_for_q2, ...].

Figure 7: Prompts for (a) generating and (b) validating question and answer from textual description of molecule.

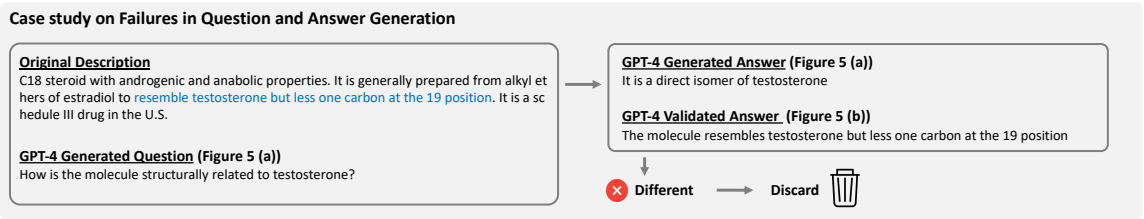

Figure 8: Failure cases in generating questions and answers.

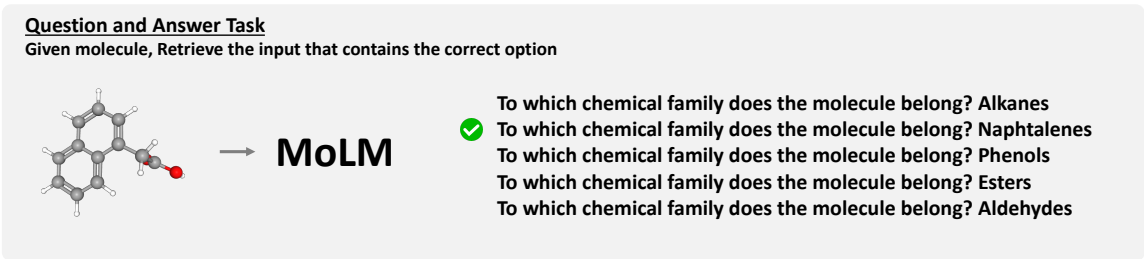

Figure 9: Evaluation scheme for question and answering task.

as follows:

- The blood-brain barrier penetration **(BBBP)** dataset comprises binary labels for **2,039** compounds regarding their barrier permeability, addressing a critical challenge in the development of drugs aimed at the central nervous system.
- The toxicology in the 21st Century **(Tox21)** dataset provides qualitative toxicity measurements for **7,831** compounds across 12 distinct targets.
- The **ToxCast** dataset offers toxicological data collected from more than 600 experiments on **8,577** compounds.
- The side effect resource **(Sider)** dataset categorizes the side effects of **1,427** approved drugs into 27 different organ system classes.
- The **Clintox** dataset comprises two classification tasks for **1,477** drug compounds, focusing on 1) toxicity during clinical trials and 2) FDA approval status.
- The **MUV** dataset features 17 demanding tasks for **93,087** compounds, curated from the PubChem BioAssay database.
- The **HIV** dataset, created by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, assesses the capacity of more than **41,127** compounds to block the replication of HIV.
- The **BACE** dataset offers qualitative binding outcomes for a collection of inhibitors targeting human β-secretase 1, encompassing **1,513** compounds.

C.5 Zero-Shot Virtual Screening

For the zero-shot virtual screening task, we utilize datasets from the Therapeutics Data Commons (TDC) (Huang et al., 2021)[10] and LIT-PCBA (Tran-Nguyen et al., 2020)[11], containing the drugs that exhibit a range of desirable properties. Among the various datasets, we utilize the following datasets:

- The human intestinal absorption **(HIA)** dataset comprises **578** drugs and their capacity for absorption from the human gastrointestinal tract into the bloodstream.
- The P-glycoprotein inhibition **(Pgp Inhibition)** dataset includes **1,212** drugs, detailing their Pgp inhibitory activities, which can significantly affect a drug's bioavailability and safety profile.
- The drug-induced liver injury **(DILI)** dataset includes **475** drugs, annotated with information regarding their potential to induce liver damage.
- The Vitamin D Receptor **(VDR)** dataset initially contains 263,303 drugs, of which 655 are active. Considering the significant imbalance between active and inactive drugs, we sample a subset of 10,000 drugs from the inactive category for analysis, i.e., a total of **10,655** drugs.

In addition to the previously mentioned datasets, we explore the virtual screening capabilities of the methods detailed in Appendix E.3 on the following datasets:

- The oral bioavailability **(Bioavailability)** dataset encompasses **640** drugs, each labeled to indicate the extent to which the drug's active ingredient is absorbed into the systemic circulation and becomes accessible at the intended site of action.
- The blood-brain barrier **(BBB)** serves as a protective shield, preventing most foreign substances from entering. This dataset encompasses **1,975** drugs, each annotated with their ability to penetrate the BBB, posing a significant challenge in developing drugs for the central nervous system.
- The human ether-a-go-go related gene **(hERG)** plays a vital role in regulating the heart's rhythm. This dataset includes 648 drugs, each evaluated for their potential to block hERG, a condition that may result in significant adverse effects.
- The **HIV** dataset contains **41,127** drugs and the label about its ability to inhibit HIV replication.
- The **SARS-CoV-2** dataset contains **1,480** drugs and the label about their activity against SARSCoV2.

D Experimental Setups

D.1 Zero-Shot Cross-Modal Retrieval

In this section, we provide further details on experimental setups for the zero-shot cross-modal retrieval task. Following previous work (Liu et al., 2023a), we evaluate the task performance in two distinctive settings 1) given molecule to retrieve the textual description, and 2) given texture description to retrieve the molecule. In each scenario, we conducted experiments with a range of options, specifically 4, 10, and 20 choices. Within these options, one is the matching counterpart, while the others are randomly selected from the dataset. Following this, the model performance is determined

[10] https://tdcommons.ai/
[11] https://drugdesign.unistra.fr/LIT-PCBA/

by its capacity to pinpoint the correct counterpart from the options provided, such as correctly matching the description to the given molecule or vice versa. For the evaluation, we conducted five separate experiments, each featuring a distinct random selection of options, and we present both the mean and standard deviation of these experiments.

D.2 Zero-Shot Question and Answering

In this section, we provide further details on the zero-shot questions and answering task. We provide an overall pipeline for generating a dataset for the question and answering tasks in Figure 6. After the creation of a set of validated questions and answers, we design the question and answering task as a retrieval task. Specifically, for a given question and its five options, we concatenate the question with each option to generate a singular input, i.e., $\texttt{input}_i = \text{Concat}(\texttt{question}, \texttt{option}_i)$, for $i = 1, \ldots, 5$. The MoLM then determines the correct answer by selecting from these inputs the one that correctly matches the question as shown in Figure 9. The only difference between the options comes from $\texttt{option}_i$, makes the task much harder compared to the cross-modal retrieval task in Section 5.2.

D.3 Molecular Property Prediction

In this section, we provide details on experimental setups for the molecular property prediction task. In this task, we evaluate how the pre-training methods affect the prediction of various molecular properties. To achieve this, we initially divide the dataset according to scaffold information using an 8:1:1 ratio for the training, validation, and test sets, respectively. That is, the molecules in the training, validation, and test sets possess distinct scaffolds. Subsequently, we fine-tune the molecular encoder f_{mol} using the training data across 100 epochs. Following previous work (Liu et al., 2023a), the model's performance is then evaluated on the test set where the hyperparameters achieve optimal performance on the validation set. We ran five individual experiments and report the average and standard deviation of the results.

D.4 Zero-Shot Virtual Screening

In this section, we provide further details on the zero-shot virtual screening task. In this task, we assess the model's capability to conduct virtual screening for drugs with specific properties described in a language model prompt. Given a dataset with binary labels, we evaluate the model performance by identifying and counting the number of molecules tagged with a positive label among those that have been screened. We start by encoding the prompt into its representation, followed by identifying the top 100 molecules whose representations are nearest to that of the prompt in the representation space. For instance, in the case of the HIA dataset, we initially identify the top 100 molecules that are closest to the prompt "Human intestinal absorption (HIA)" within the representation space among a total of 578 molecules and count the molecules with positive labels. Additionally, for every dataset, we employ two types of prompts with distinct characteristics, i.e., one with short and abstract descriptions and one with long and detailed descriptions. In Tab. 7, we provide the specific prompts used for each dataset.

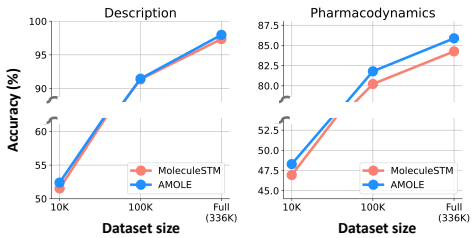

Figure 10: Model performance depending on the size of training data.

E Additional Experimental Results

E.1 Zero-Shot Cross-Modal Retrieval

Complete Set of Experimental Results. In this section, we provide a complete set of experimental results for the zero-shot cross-modal retrieval task. In Table 8, we provide empirical results on the (a) description, (b) pharmacodynamics, and (c) ATC dataset. We observe that as tasks get harder, i.e., increase in the number of options from 4 to 20, the performance disparity between baseline methods and AMOLE broadens.

Statistical Significance Test. To demonstrate the statistical significance of the improvement in Table 1, we performed a paired t-test in the representation learning context for each dataset. Since we evaluate models on five independent trials, we compared the mean performances of MoleculeSTM, ablations, and AMOLE across each trial. The p-values presented in the table are the outcomes of these paired t-tests, based on the following hypothesis: $H_0 : \mu_{\text{AMOLE}} = \mu_{\text{Baseline}}$ and $H_1 : \mu_{\text{AMOLE}} \geq$

Dataset	Prompt
HIA	Human intestinal absorption (HIA)
	The molecule is positive w.r.t. a property that is defined as 'the ability of the body to be absorbed from the human gastrointestinal system into the bloodstream of the human body'
Pgp Inhibition	P-glycoprotein Inhibition
	This molecule is known to inhibit P-glycoprotein, which is an ABC transporter protein involved in intestinal absorption, drug metabolism, and brain penetration, and its inhibition can seriously alter a drug's bioavailability and safety
DILI	Inducing liver injury
	This molecule induces liver injury that is most commonly caused by Amoxicillin clavulanate isoniazid, and nonsteroidal anti-inflammatory drugs.
VDR	Vitamin D receptor
	This molecule is active w.r.t. Vitamin D receptor. The best pharmacophore hypothesis contains one hydrogen bond acceptor (A), one hydrogen bond donor (D) and two hydrophobic regions (H).
Bioavailability	Oral Bioavailability
	The molecule is positive w.r.t. a property that is defined as 'the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action'
BBB	Blood-Brain Barrier penetration
	The molecule is able to penetrate the Blood-Brain Barrier, which is the protection layer that blocks most foreign drugs as a membrane separating circulating blood and brain extracellular fluid, to deliver to the site of action.
hERG	hERG Blocker
	This molecule blocks the hERG, which is crucial for the coordination of the heart's beating.
HIV	Active against HIV virus
	This molecule is active against HIV virus. These drugs typically possess one or more of the following properties: Reverse Transcriptase Inhibitors, Protease Inhibitors, Integrase Inhibitors, Fusion Inhibitors, CCR5 Antagonists, Post-Attachment Inhibitors.
SARS-CoV-2	Active against SARS-CoV-2 virus
	This molecule is active against SARS-CoV-2 virus. These drugs typically possess one or more of the following properties: Viral Entry Inhibitors, RNA Polymerase Inhibitors, Protease Inhibitors, Immune Modulators, Interferon Inducers, Antibody-based Therapies.

Table 7: The abstractive and detailed prompts used for each dataset during the virtual screening task.

			(a) Description						(b) Pharamacodynamics						(c) ATC					
	SMILES	Graph	Given Molecule			Given Text			Given Molecule			Given Text			Given Molecule			Given Text		
			4	10	20	4	10	20	4	10	20	4	10	20	4	10	20	4	10	20
Single Encoder																				
MolT5	✓	✗	25.39 (2.93)	10.08 (0.30)	5.06 (0.13)	27.94 (4.15)	11.81 (1.98)	6.66 (0.42)	25.88 (0.55)	11.70 (0.18)	6.80 (0.28)	25.70 (0.90)	11.29 (0.70)	6.02 (0.57)	25.80 (0.45)	11.56 (0.21)	6.48 (0.25)	25.22 (0.24)	10.84 (0.11)	6.10 (0.09)
BioT5	✓	✗	27.42 (2.34)	11.48 (0.30)	6.47 (0.13)	27.95 (4.15)	12.73 (1.98)	6.02 (0.42)	27.43 (0.52)	12.79 (0.37)	7.42 (0.52)	27.74 (0.80)	12.65 (0.92)	7.36 (0.13)	29.26 (0.57)	13.05 (0.31)	7.71 (0.16)	28.01 (0.53)	12.08 (0.42)	6.78 (0.45)
KV-PLM	✓	✗	72.42 (2.76)	53.73 (2.71)	42.28 (3.29)	73.77 (2.51)	56.47 (2.60)	45.64 (2.51)	68.61 (1.06)	47.85 (0.64)	36.84 (0.53)	67.86 (0.98)	49.46 (0.44)	37.93 (0.62)	59.13 (0.42)	40.28 (0.40)	30.21 (0.40)	60.91 (0.40)	43.07 (0.70)	33.22 (0.40)
Separate Encoder																				
MoMu	✗	✓	99.22 (0.08)	**98.48** (0.22)	**97.39** (0.19)	99.07 (0.58)	98.04 (0.21)	96.84 (0.17)	89.50 (0.69)	82.98 (0.56)	77.82 (0.54)	88.59 (0.60)	82.06 (0.46)	77.05 (0.28)	73.27 (0.43)	59.99 (0.48)	51.34 (0.37)	69.71 (0.46)	56.13 (0.26)	47.68 (0.34)
MoleculeSTM	✓	✗	99.17 (0.10)	98.08 (0.18)	96.70 (0.35)	98.95 (0.16)	97.61 (0.25)	96.22 (0.29)	90.38 (0.66)	83.27 (0.86)	77.28 (0.94)	88.98 (0.41)	81.69 (0.92)	75.01 (0.49)	73.90 (0.49)	60.75 (0.41)	52.36 (0.75)	71.86 (0.10)	58.56 (0.54)	50.01 (0.49)
MoleculeSTM	✗	✓	99.14 (0.13)	97.06 (1.85)	95.87 (1.87)	98.92 (0.09)	97.29 (0.34)	95.82 (0.37)	91.11 (0.67)	84.84 (0.61)	79.21 (0.75)	90.01 (0.56)	83.26 (0.74)	77.15 (0.74)	75.39 (0.49)	61.83 (0.41)	52.70 (0.75)	71.98 (0.10)	57.79 (0.54)	48.54 (0.49)
AMOLE	✗	✓	**99.30** (0.04)	97.52 (1.89)	96.48 (2.94)	**99.16** (0.13)	**98.13** (0.32)	**97.20** (0.26)	**91.93** (0.28)	**85.84** (0.54)	**81.46** (0.60)	**90.67** (0.52)	**85.24** (0.53)	**80.11** (0.42)	**75.85** (0.34)	**63.03** (0.57)	**54.76** (0.57)	**73.52** (0.22)	**60.48** (0.45)	**51.47** (0.56)

Table 8: Zero-shot cross-retrieval task results on (a) description, (b) pharmacodynamics, and (c) ATC dataset.

	Augment	S^2P Loss	ER Loss	Descr.	Pharma.	ATC
Ablation 2	✓	✓	✗	96.65	80.47	51.55
Ablation 3	✓	✗	✓	96.02	78.83	51.85
Ablation 4	✗	✓	✓	96.69	79.78	50.35
AMOLE	✓	✓	✓	**96.84**	**80.79**	**53.12**

Table 9: Additional ablation studies.

μ_{Baseline}. In Table 11, we observe that AMOLE's improvement over previous works is statistically significance at a p-value of 0.05.

Moreover, we also conduct a statistical significance test in ablation studies results in Table 2 to demonstrate the effectiveness of each component in our model. The p-values detailed in the table are derived from these paired t-tests, underpinning the hypothesis: $H_0 : \mu_{\text{AMOLE}} = \mu_{\text{Ablation}}$ and $H_1 : \mu_{\text{AMOLE}} \geq \mu_{\text{Ablation}}$. From Table 12, we observe that each component of AMOLE statistically significantly contributes to the model performance in the most cases.

Model Performance on Dataset Scale. In Figure 10, the impact of training data volume on model effectiveness is depicted. It's evident that enlarging the training dataset size invariably enhances model efficacy, underscoring the significance of data scale not just in the realm of VLM but equally in MoLM. We also note that AMOLE uniformly surpasses MoleculeSTM across all data volumes, showcasing AMOLE's superior data efficiency.

Additional Ablation Studies. In Table 9, we provide further ablation studies to evaluate the effectiveness of each component in AMOLE. As done in Section 5.2, we assess the effectiveness of each component by calculating the average performance across hard cross-retrieval tasks, i.e., retrieving among 20 texts given a molecule and retrieving among 20 molecules given a text. The significance of individual components is observed to change across datasets due to the distinct characteristics of the descriptions they contain. Nonetheless, when all elements are integrated, namely in AMOLE, it uniformly surpasses a range of simplified models in performance. This underlines AMOLE's capability to effectively amalgamate different modules for enhanced outcomes.

Sensitivity analysis on k. In Figure 11, we illustrate how model performance fluctuates depending on k, the hyperparameter that specifies the number of molecules sharing the same description. We note that as k is reduced ($k = 10$), the count of

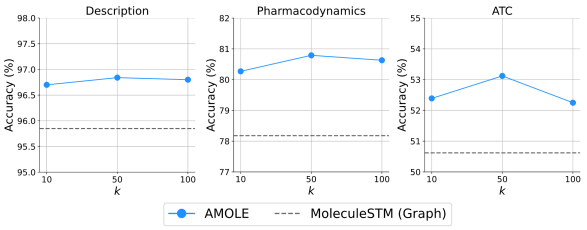

Figure 11: Sensitivity analysis on k.

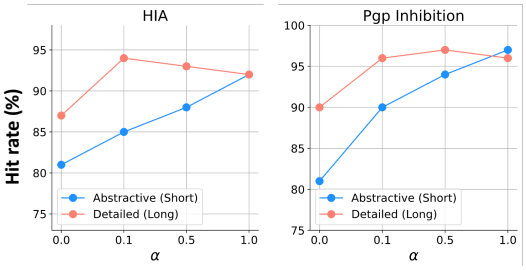

Figure 12: Sensitivity analysis on α.

molecule-text pairs diminishes, leading to a slight decline in model performance. Conversely, as k increases ($k = 100$), the issue of false positives intensifies, suggesting that an optimal level of k needs to be determined during training. Despite these variations, we observe that AMOLE consistently surpasses MoleculeSTM across all values of k, showcasing the robustness of AMOLE against changes in the hyperparameter k.

E.2 Zero-Shot Question and Answering

In this section, we explore how changes in the input format impact the performance of graph-based baseline models on Zero-Shot question-and-answer tasks. Specifically, during training, we randomly replace the input textual description t_i (Original Input) with $\tilde{t}i = ti$ [SEP] $t_{i'}$ (Modified Input), which is incorporated for the ER loss in AMOLE. According to Table 13, we see a decline in the performance of baseline models when the input textual description is altered in this manner. This suggests that simply concatenating textual descriptions into $\tilde{t}_i$ does not aid the Zero-Shot Question and Answering task, despite their similar input format. It underscores the need for a more advanced approach, such as expertise reconstruction module in AMOLE, to effectively handle QA tasks.

E.3 Zero-Shot Virtual Screening

Additional Experimental Results. In this section, we present supplementary experimental outcomes for the zero-shot virtual screening task across multiple datasets. As shown in Table 10, it is evident that AMOLE uniformly surpasses the baseline ap-

Dataset	Prompt	Model			Random
		MoMu	MoleculeSTM	AMOLE	
HIA	Human intestinal absorption (HIA)	72.00	70.00	**92.00**	86.50
	The molecule is positive w.r.t. a property that is defined as the ability of the body · · ·	86.00 (+19.44%)	76.00 (+8.57%)	**92.00** (-)	
Pgp Inhibition	P-glycoprotein Inhibition	76.00	83.00	**97.00**	53.36
	This molecule is known to inhibit P-glycoprotein, which is an ABC transporter · · ·	85.00 (+11.84%)	89.00 (+7.22%)	**96.00** (-1.03%)	
DILI	Inducing liver injury	34.00	49.00	**56.00**	49.68
	This molecule induces liver injury that is most commonly caused by Amoxicillin · · ·	61.00 (+79.41%)	67.00 (+36.73%)	**76.00** (+35.71%)	
VDR	Vitamin D receptor	**10.00**	**10.00**	**10.00**	6.14
	This molecule is active w.r.t. Vitamin D receptor. The best pharmacophore · · ·	8.00 (-20.00%)	5.00 (-50.00%)	**11.00** (+10.00%)	
Bioavailability	Oral Bioavailability	82.00	93.00	**88.00**	76.88
	The molecule is positive w.r.t. a property that is defined as 'the rate and extent · · ·	81.00 (-1.22%)	66.00 (-29.03%)	**82.00** (-6.82%)	
BBB	Blood-Brain Barrier penetration	94.00	90.00	**96.00**	76.40
	The molecule is able to penetrate the Blood-Brain Barrier, which is the protection layer · · ·	85.00 (-9.57%)	86.00 (-4.44%)	**88.00** (-8.33%)	
hERG	hERG Blocker	80.00	82.00	**87.00**	68.86
	This molecule blocks the hERG, which is crucial for the coordination of the heart's beating.	68.00 (-15%)	71.00 (-13.41%)	**78.00** (-10.33%)	
HIV	Active against HIV virus	2.00	1.00	**3.00**	3.51
	This molecule is active against HIV virus. These drugs typically possess · · ·	1.00 (-50%)	3.00 (+200.00%)	**4.00** (+33.33%)	
SARS-Cov-2	Active against SARS-CoV-2 virus	5.00	5.00	**6.00**	5.93
	This molecule is active against SARS-CoV-2 virus. These drugs typically possess · · ·	4.00 (-20%)	5.00 (-)	**7.00** (+16.67%)	

Table 10: Additional zero-shot virtual screening task results. The numbers in the bracket indicate the performance increase with detailed (long) prompts compared to abstract (short) prompts.

	Given Molecule @ 20			Given Text @ 20		
	Descr.	Pharma.	ATC	Descr.	Pharma.	ATC
MoMu	7.8e-01	**5.2e-04**	**6.5e-06**	**1.6e-03**	**3.9e-05**	**5.2e-05**
MoleculeSTM (SMILES)	5.7e-01	**5.9e-04**	**4.3e-04**	**5.1e-04**	**4.8e-05**	**3.4e-03**
MoleculeSTM (Graph)	**4.3e-03**	**2.4e-03**	**3.1e-04**	**1.6e-03**	**2.2e-03**	**1.4e-05**

Table 11: Statistical significance test results. Each number indicates the p-value for the test. Bold indicates the p-values below 0.05.

	Given Molecule @ 20			Given Text @ 20		
	Descr.	Pharma.	ATC	Descr.	Pharma.	ATC
MoleculeSTM	**4.3e-03**	**2.4e-03**	**3.1e-04**	**1.6e-03**	**2.2e-03**	**1.4e-05**
Ablation1	3.6e-01	**4.7e-03**	**4.7e-03**	**9.1e-04**	**1.3e-02**	**1.9e-02**
Ablation2	7.3e-01	**6.6e-03**	**3.2e-03**	**2.9e-02**	6.7e-01	**3.9e-03**

Table 12: Statistical significance test on ablation studies results. Each number indicates the p-value for the test. Bold indicates the p-values below 0.05.

proaches across a range of datasets. Additionally, it is often noted that the model's efficacy declines when provided with long and detailed descriptions as opposed to short and abstract ones. This suggests that detailed prompts are not invariably beneficial, underscoring the significance of the model's resilience across different types of descriptions.

Sensitivity Analysis on α. We further investigate the impact of the expertise reconstruction (ER) loss weighting factor, α, on virtual screening performance. It is found that while the model's performance remains largely consistent when provided with a detailed and lengthy prompt, its performance on abstract and brief prompts significantly varies with the choice of α. Specifically, the model's effectiveness on abstract and concise prompts improves as α increases. This suggests that the expertise transfer module effectively allows the model to infer related information, thereby enhancing stability and performance, even when faced with brief and abstract prompts.

	Original Input		Modified Input		AMOLE
	MoMu	Molecule STM	MoMu	Molecule STM	
Descr.	36.62	38.21	36.05	37.09	**38.61**
Pharma.	30.73	30.72	29.57	30.58	**31.44**

Table 13: Model Performance comparison when the baseline models trained with modified input.

E.4 Effect of ER Loss during training

As illustrated by our data analysis in Appendix C.1, molecules described by more than two texts are rare, and this scarcity could reduce the effective-

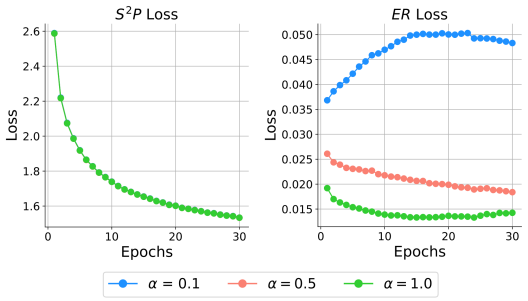

Figure 13: Training curve in AMOLE.

ness of the ER Loss. To analyze the effect of ER loss, we have visualized the change of the S^2P loss and ER loss during training with various levels of α. Figure 13 demonstrates a significant decrease in ER loss as α increases, signifying that the module is capable of efficiently performing expertise reconstruction. However, given our main objective is to align molecule and text representations effectively, it is appropriate for the S^2P Loss to play a more substantial role than the ER loss. To sum up, we contend that our model adeptly learns the representations of molecules and texts based on structural similarity and enriches the language model through the reconstruction of expertise.

F License for the Datasets

Dataset	License URL
PubChem	https://www.nlm.nih.gov/web_policies.html
MoleculeNet	https://opensource.org/license/mit/
DrugBank	https://creativecommons.org/licenses/by-nc/4.0/legalcode.en
TDC	https://opensource.org/license/mit/

Table 14: Licenses for the datasets used in the paper

In Table 14, we detail the sources and data rights for all data components used in this paper. All data sources underwent thorough examination to confirm that their licensing agreements allow for the type of research we conducted and its further applications.

Throughout the paper, we believe we have properly attributed the creators of the scientific artifacts cited. We affirm that all data utilized in this study adhere to the conditions of the CC BY 4.0 License. We accept the obligation to promote transparent and equitable data usage, acknowledging the original creators' efforts. Furthermore, we guarantee that our dataset is free of personally identifiable or privacy-sensitive information.