# PharmaQA: Prompt-Based Molecular Representation Learning via Pharmacophore-Oriented Question Answering

**Chengwei Ai[1], Qiaozhen Meng[2], Mengwei Sun[1], Ruihan Dong[3], Hongpeng Yang[4], Shiqiang Ma[5*], Xiaoyi Liu[6*], Cheng Liang[7*], Fei Guo[1*]**

[1]School of Computer Science and Engineering, Central South University
[2]School of Computer Science, Xiangtan University
[3]Academy for Advanced Interdisciplinary Studies, Peking University
[4]Department of Computer Science and Engineering, University of South Carolina
[5]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[6]School of Chinese Materia Medica, Beijing University of Chinese Medicine
[7]School of Information Science and Engineering, Shandong Normal University
ai_chery@163.com, qiaozhenm@xtu.edu.cn, smw010207@163.com, dongruihan@stu.pku.edu.cn, hongpeng@email.sc.edu, sq.ma@siat.ac.cn, xiaoyi.liu@bucm.edu.cn, alcs417@sdnu.edu.cn, guofei@csu.edu.cn

## Abstract

Molecular representation plays a central role in computational drug discovery. Pharmacophores, functional groups responsible for molecular bioactivity, have been widely studied in cheminformatics. However, their incorporation into molecular representation learning, particularly in a context reasoning or generalization, remains relatively limited. To address this gap, we propose PharmaQA, a pharmacophore oriented question answering framework that formulates tailored prompts to extract context-aware molecular semantics. Rather than encoding pharmacophore features, PharmaQA learns to answer pharmacophore related queries. This design enables flexible reasoning across diverse tasks, including molecular property prediction, compound-target interaction prediction, and binding affinity estimation. Experimental results on benchmark datasets demonstrate that PharmaQA achieves competitive performance. In a ligand discovery case study using FDA-approved compounds, the framework identified potential inhibitors for three therapeutic targets, with strong docking performance. As a generalizable and modular solution, PharmaQA incorporates pharmacophoric knowledge into molecular embeddings, enhancing both predictive accuracy and interpretability in drug discovery applications.

## Introduction

Effective molecular representation is essential for drug discovery, where predictive performance relies on capturing both structural and functional information. Traditional machine learning models relied on handcrafted descriptors or fingerprints, which limited scalability and adaptability (Dong et al. 2018; Butler et al. 2018). Deep learning techniques, including convolutional, recurrent, and graph neural networks (CNN, RNN and GNNs) have enabled automated feature extraction from SMILES strings and molecular graphs (Xu et al. 2017; Gilmer et al. 2017; Li et al. 2023), yet often struggle to capture long range dependencies and functional semantics essential to molecular bioactivity (You et al. 2020; Sun, Dai, and Yu 2022). Although recent pretraining strategies have improved generalization (You et al. 2020, 2021; Hu et al. 2020), these models still lack mechanisms to integrate structured domain knowledge into molecular embeddings that are sensitive to context.

Meanwhile, identifying molecules with specific pharmacological properties remains a central challenge in drug discovery, mainly due to the high costs and time-consuming nature of experimental screening (Dickson and Gagnon 2004; Mullard 2014). Pharmacophores, spatially arranged functional groups responsible for molecular bioactivity, offer a powerful abstraction to capture molecular interactions with biological targets (Jiang et al. 2023a). They have long played a central role in ligand based virtual screening and pharmacophore modeling (Yu et al. 2025; Li et al. 2022; Zhu et al. 2023). While recent studies have begun to incorporate pharmacophoric information into molecular modeling, these approaches often rely on static encodings or predefined pharmacophore graphs, which limit their ability to model contextual and task-specific semantics (Jiang et al. 2023b). As a result, existing methods may fall short in expressiveness, interpretability, and generalization, particularly in multiple property learning or unseen target binding scenarios.

Recent advances in Question-Answering (QA) methods, especially those enabled by large language models, demonstrate that structured prompting can elicit rich, context-sensitive representations across diverse domains (Liu et al. 2025; Park et al. 2024; Li et al. 2024; Zhong et al. 2022). These approaches offer a modular mechanism for injecting external knowledge and supporting semantic reasoning, moving beyond static, manually designed features. Inspired by this, we reinterpret molecular representation learning as a pharmacophore-centric QA task: the model is prompted with domain-specific questions about functional groups that underlie molecular bioactivity.

*Corresponding author.

To realize this idea we propose PharmaQA, a pharmacophore-oriented question answering framework that formulates structured prompts to extract semantic, context-aware molecular features (Figure 1). Firstly, we construct a set of 27 natural language questions and explanatory descriptions targeting common pharmacophoric substructures, identified via SMARTS patterns in RDKit. These prompts capture the spatial and functional roles of key molecular motifs, serving as semantic queries to guide molecular interpretation.

To integrate semantic signals with structural information, we introduce a multi-path knowledge guided attention mechanism that jointly encodes pharmacophore-centric questions, descriptive knowledge, and molecular graph features. Each question is processed in parallel with its associated substructure and chemical context through dedicated attention branches, enabling alignment between function and structure. The resulting fused representation serves dual roles: it is used both to answer the pharmacophore query and to act as a semantic prompt injected into downstream molecular embedding modules. This design allows PharmaQA to encode task-specific functional knowledge into molecular representations in a contextual and adaptive manner. We evaluate PharmaQA across multiple public benchmarks involving classification, regression, and compound-target interaction and affinity prediction. Our method consistently outperforms state-of-the-art baselines. Furthermore, in retrospective validation, our model prioritized FDA-approved compounds with high overlap to literature reported actives (e.g., 13/20 for FGFR1), while docking studies further identified novel candidates with strong binding potential. Our contributions are summarized as follows:

- We propose PharmaQA, a pharmacophore-oriented QA framework for interpretable molecular representation.
- We design a multi-path knowledge guided attention module that integrates semantic prompts with graph structure.
- We validate our approach across multiple tasks, demonstrating improved performance over baselines.

## Related Work

**Molecular Representation Learning.** Molecular representation methods typically model molecules as graphs composed of atoms and bonds. MoleBERT (Xia et al. 2023) introduces a triplet masked contrastive learning strategy to capture both local and global structure. Recent multimodal approaches, such as MoleculeSTM (Liu et al. 2023) and SPMM (Chang and Ye 2024), enhance representations by integrating textual or biochemical property features, but often lack structured domain knowledge such as pharmacophores. Knowledge guided models like KPGT (Li et al. 2023) incorporate knowledge nodes and use the LineGraphTransformer to improve structural encoding. PharmHGT (Jiang et al. 2023a) introduces pharmacophore constraints but relies on static representations. These limitations highlight the need for a more flexible and contextual paradigm, motivating our pharmacophore-oriented QA approach.

## Proposed Framework: PharmaQA

### Molecular Representation

We construct an augmented molecular graph $\mathcal{G}_{aug} = (\mathcal{V}_{aug}, \mathcal{E}_{aug})$ by abstracting each bond as a node connected to its two atoms, and adding virtual edges between bond nodes sharing an atom, following line graph construction (Chen et al. 2024; Li et al. 2023). This formulation enriches interaction path representation and bond context. To embed $\mathcal{G}_{aug}$, we adopt a LineGraphTransformer encoder that integrates multi-head self attention with structural priors. Node features are initialized using atom and bond level attributes extracted via RDKit. In addition, two structural bias matrices are introduced into the attention computation: a path adjacency matrix $\boldsymbol{A}^p$ capturing topological proximity, and a distance matrix $\boldsymbol{A}^d$ reflecting interatomic distances (Li et al. 2023). At each layer $l$, the hidden states $\boldsymbol{H}^{(l)} \in \mathbb{R}^{n \times d}$ are updated as:

$$\begin{aligned} \boldsymbol{A} &= \phi\left(\frac{\boldsymbol{H}^{(l-1)}\boldsymbol{W}_Q(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_K)^\top}{\sqrt{d}} + \boldsymbol{A}^p + \boldsymbol{A}^d\right), \\ \boldsymbol{H}^{(l)} &= \text{Residual}(\boldsymbol{H}^{(l-1)}, \boldsymbol{A}(\boldsymbol{H}^{(l-1)}\boldsymbol{W}_V)), \end{aligned} \quad (1)$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d \times d_g}$ are learnable projection matrices, $\phi$ is the softmax function, and $Residual(\cdot)$ includes normalization and skip connections. The final node level representation $\boldsymbol{H}_G = \boldsymbol{H}^{(L)}$ serves as the graph embedding input for downstream pharmacophore reasoning.

### Question Representation

To capture pharmacophore level semantics, we define a set of $P$ natural language question–description pairs $\{(\boldsymbol{q}_1, \boldsymbol{d}_1), \cdots, (\boldsymbol{q}_P, \boldsymbol{d}_P)\}$, each targeting a specific pharmacophoric substructure. These pairs are encoded using PubMedBERT (Gu et al. 2021), a domain-specific pretrained language model tailored for biomedical semantics. Given tokenized question $\boldsymbol{q}_i$ and description $\boldsymbol{d}_i$ with lengths $l_q$ and $l_d$, their contextual embeddings are computed as:

$$\boldsymbol{H}_{\boldsymbol{x}} = \text{PubMedBERT}(\boldsymbol{x}), \quad \boldsymbol{x} \in \{\boldsymbol{q}_i, \boldsymbol{d}_i\}, \quad (2)$$

where $\boldsymbol{H}_{\boldsymbol{x}}$ is the contextual embedding of question $\boldsymbol{q}_i$ or description $\boldsymbol{d}_i$ with dimension $d_l$. Each question–description pair serves as a semantic embedding prior knowledge of pharmacophores, contributing to a contextual representation that supports functional reasoning.

### Pharmacophore Knowledge Alignment

Inspired by VTQA (Chen and Wu 2024), which introduces external descriptive text to enhance cross modal reasoning, we integrate pharmacophore question and description embeddings with molecular graph features using a multi-path knowledge guided attention module. To enhance structural grounding during pharmacophore alignment, the embedding of the SMILES is incorporated into the embeddings of each description input. These SMILES-augmented embeddings serve as guidance, enabling the model to better associate functional descriptions with relevant atomic regions in the molecular graph. For notational simplicity, we

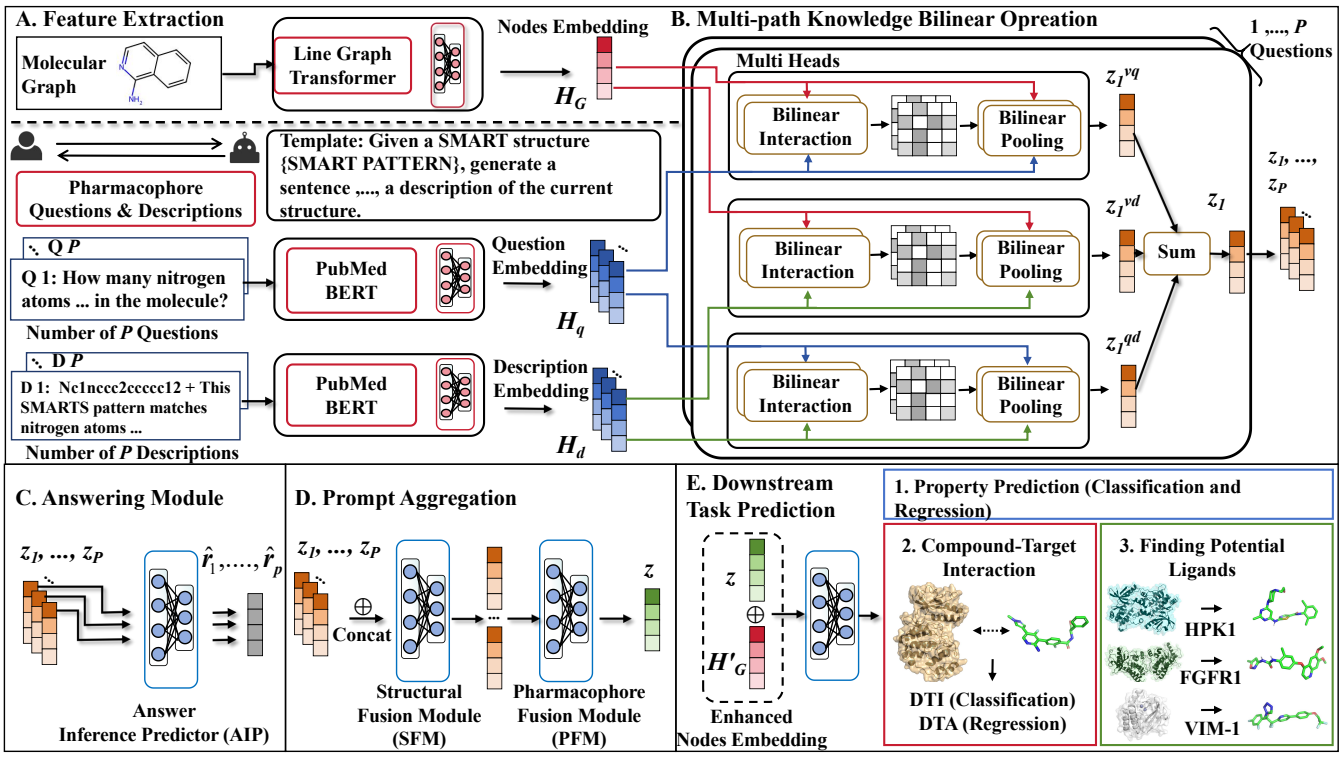



Figure 1: Overall architecture of PharmaQA. **A** Molecular graph and pharmacophore related questions are encoded. **B** Multi-path knowledge guided bilinear attention aligns graph–text modalities. **C-D** Answer prediction and prompt aggregation. **E** Prediction across downstream tasks.

also denote the fused representation as $\mathbf{H}_d$ in our formulation. For each pharmacophore query $i = 1, \ldots, P$, we compute three attention maps: $\boldsymbol{G}_i^{vq} \in \mathbb{R}^{n \times l_q}$ (graph-question), $\boldsymbol{G}_i^{vd} \in \mathbb{R}^{n \times l_d}$ (graph-description), and $\boldsymbol{G}_i^{qd} \in \mathbb{R}^{l_q \times l_d}$ (question-description), as follows:

$$
\begin{aligned}
\boldsymbol{G}_i^{vq} &= \phi\left((\mathbf{1} \cdot \boldsymbol{u}_i^{vq\top}) \odot \sigma(\boldsymbol{H}_G \boldsymbol{U}_i) \cdot \sigma(\boldsymbol{H}_{\boldsymbol{q}_i} \boldsymbol{V}_i)^\top\right), \\
\boldsymbol{G}_i^{vd} &= \phi\left((\mathbf{1} \cdot \boldsymbol{u}_i^{vd\top}) \odot \sigma(\boldsymbol{H}_G \boldsymbol{U}_i) \cdot \sigma(\boldsymbol{H}_{\boldsymbol{d}_i} \boldsymbol{W}_i)^\top\right), \\
\boldsymbol{G}_i^{qd} &= \phi\left((\mathbf{1} \cdot \boldsymbol{u}_i^{qd\top}) \odot \sigma(\boldsymbol{H}_{\boldsymbol{q}_i} \boldsymbol{V}_i) \cdot \sigma(\boldsymbol{H}_{\boldsymbol{d}_i} \boldsymbol{W}_i)^\top\right),
\end{aligned} \tag{3}
$$

where $\mathbf{1} \in \mathbb{R}^n$ is an all one vector, $\phi$ is the softmax function, $\sigma$ is the ReLU activation, and $\odot$ denotes elementwise multiplication. Matrices $\boldsymbol{U}_i, \boldsymbol{V}_i, \boldsymbol{W}_i$ and vectors $\boldsymbol{u}_i^*$ are learnable parameters. Next, we use the attention maps to derive joint embeddings for each pharmacophore query $i$, capturing interactions between molecular structure, query, and descriptive knowledge. For each attention type $p \in \{\text{vq}, \text{vd}, \text{qd}\}$ and channel $k = 1, \ldots, K$, we compute the bilinear interaction embedding as:

$$
\begin{aligned}
f_{i,k}^{vq} &= \sigma(\boldsymbol{H}_G \boldsymbol{U}_i)_k^\top \boldsymbol{G}_i^{vq}\, \sigma(\boldsymbol{H}_{\boldsymbol{q}_i} \boldsymbol{V}_i)_k, \\
f_{i,k}^{vd} &= \sigma(\boldsymbol{H}_G \boldsymbol{U}_i)_k^\top \boldsymbol{G}_i^{vd}\, \sigma(\boldsymbol{H}_{\boldsymbol{d}_i} \boldsymbol{W}_i)_k, \\
f_{i,k}^{qd} &= \sigma(\boldsymbol{H}_{\boldsymbol{q}_i} \boldsymbol{V}_i)_k^\top \boldsymbol{G}_i^{qd}\, \sigma(\boldsymbol{H}_{\boldsymbol{d}_i} \boldsymbol{W}_i)_k,
\end{aligned} \tag{4}
$$

we then concatenate the $K$ dimensional outputs across all channels:

$$
\boldsymbol{f}_i^p = [f_{i,1}^p, \ldots, f_{i,K}^p], \quad p \in \{\text{vq}, \text{vd}, \text{qd}\}. \tag{5}
$$

To generate a compact representation from each interaction branch, we apply a bilinear pooling operation followed by sum pooling with stride $s$. This compresses the $K$ dimensional vectors into lower dimensional embeddings. To capture diverse relational patterns, we use a multi-head strategy, aggregating $M$ separate heads for each attention type $p \in \{\text{vq}, \text{vd}, \text{qd}\}$. The output for each attention type is defined as:

$$
\boldsymbol{z}_i^p = \sum\nolimits_{m=1}^{M} \text{SumPool}_m(\boldsymbol{f}_i^p, s), \quad p \in \{\text{vq}, \text{vd}, \text{qd}\}, \tag{6}
$$

where $\text{SumPool}_m(\cdot, s)$ denotes sum pooling on the $m$-th head with stride $s$, and $\boldsymbol{z}_i^p \in \mathbb{R}^{K/s}$. Finally, we obtain the integrated representation for the $i$-th pharmacophore query by summing the three pooled vectors:

$$
\boldsymbol{z}_i = \boldsymbol{z}_i^{\text{vq}} + \boldsymbol{z}_i^{\text{vd}} + \boldsymbol{z}_i^{\text{qd}}. \tag{7}
$$

The resulting embedding $\boldsymbol{z}_i$ integrates graph and textual semantics for the $i$-th pharmacophore query, serving as a semantic prompt for downstream tasks. Despite lacking explicit labels, the attention maps enable alignment between functional queries and molecular substructures in a differentiable manner.

## Prompt Based Prediction

To effectively leverage the functional group hierarchy, we first group these $P$ pharmacophore types into 8 categories (e.g., Donors, Acceptors, etc.) by applying the Substructure Fusion Module (SFM), implemented as an Multi-Layer Perceptron (MLP), to fuse embeddings within each group. Subsequently, the resulting 8 group level embeddings are further integrated using the Pharmacophore Fusion Module (PFM), which is also designed as an MLP, to obtain a holistic pharmacophore-aware molecular representation. This representation, concatenated with the graph encoding, is used for final classification or regression prediction. Specifically, we obtain the final prompt vector $\boldsymbol{z}$:

$$\boldsymbol{z} = \mathrm{PFM}(\mathrm{SFM}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_P)), \tag{8}$$

each $\boldsymbol{z}_i$ can be viewed as the model's answer to the pharmacophore-related question $\boldsymbol{q}_i$ posed about the input molecule, representing how strongly the molecule exhibits the queried functional characteristic. The learned embeddings encode the molecular response to the query in a continuous vector space. In downstream tasks, we integrate this semantic prompt $\boldsymbol{z}$ with molecular embeddings obtained from a secondary encoder, denoted as $\boldsymbol{H}'_{\mathcal{G}} = \mathrm{Encoder}'(\mathcal{G})$. This produces a unified embedding for final prediction:

$$\hat{y} = \mathrm{MLP}(\mathrm{concat}(\boldsymbol{z}, \boldsymbol{H}'_{\mathcal{G}})), \tag{9}$$

where $\mathrm{concat}(\cdot)$ denotes vector concatenation. This design allows PharmaQA to inject functional knowledge into molecular representations in a flexible and modular fashion. To formalize the answer in our QA framework, each pharmacophore-aware embedding $\boldsymbol{z}_i$ is additionally passed through an Answering Inference Predictor (AIP), a regression head that outputs a score $\hat{r}_i$ corresponding to the presence or strength of the pharmacophore feature in the molecule:

$$\hat{r}_i = \mathrm{AIP}(\boldsymbol{z}_i). \tag{10}$$

During training, the AIP outputs are supervised by ground truth scores $r_i$ obtained via substructure matching with RDKit, which enforces interpretable and biologically meaningful predictions. This pharmacophore answering loss encourages each $\boldsymbol{z}_i$ to capture specific functional group information beyond the aggregated molecular embedding, as described in the *Model Training* section.

## Model Training

PharmaQA is trained using a composite loss function that balances three objectives: task-specific prediction, pharmacophore QA, and structural alignment.

**Task Prediction Loss.** We denote the task loss as $L_p$, which varies depending on the task type. For the classification task, the Binary Cross Entropy (BCE) loss function is adopted, which is suitable for handling binary or multi-label classification problems. For regression tasks, we adopt Mean Squared Error (MSE) loss.

**Pharmacophore Answering Loss.** To enhance the model's ability to reason over pharmacophoric prompts, we introduce an auxiliary loss $L_{\text{ph}}$ that supervises the AIP outputs $\hat{r}_i^j$, which estimate the answers for each pharmacophore-aware embedding $\boldsymbol{z}_i^j$ of molecule $j$. The ground truth scores $r_i^j$ are obtained by applying RDKit pharmacophore substructure matching to identify pharmacophore-related nodes in the molecule and aggregating presence and count information over the matched regions. The loss is computed as the mean squared error between predictions and ground truth:

$$L_{\text{ph}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{P}\sum_{i=1}^{P}(r_i^j - \hat{r}_i^j)^2, \tag{11}$$

where $N$ is the number of molecules and $P$ is the number of pharmacophore queries.

**Structural Alignment Loss.** To improve interpretability and localize pharmacophore-relevant substructures, we propose an alignment loss $L_{\text{align}}$ based on attention maps. For each molecule, we define a binary matrix $\boldsymbol{O}^{(j)} \in \mathbb{R}^{n\times P}$ indicating whether node $k$ belongs to functional group $i$. We derive predicted node-level response by aggregating $\boldsymbol{G}_i^{\text{vq}}$ across the question dimension, yielding $\boldsymbol{v}_i \in \mathbb{R}^n$ per query. Stacking all $P$ outputs gives predicted $\hat{\boldsymbol{O}}^{(j)} \in \mathbb{R}^{n\times P}$:

$$L_{\text{align}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{nP}\left\|\boldsymbol{O}^{(j)} - \hat{\boldsymbol{O}}^{(j)}\right\|_F^2. \tag{12}$$

This loss encourages consistency between predicted attention scores and known functional substructures, acting as a supervision signal for the alignment between pharmacophore prompts and molecular regions.

**Total Loss.** The full training objective is a weighted combination of the three components:

$$L = L_p + \alpha L_{\text{ph}} + \beta L_{\text{align}}, \tag{13}$$

where $\alpha$ and $\beta$ are hyperparameters that control the trade-off among prediction, reasoning, and alignment objectives.

# Experiments

## Datasets

To comprehensively evaluate the effectiveness of PharmaQA, we employ several benchmark datasets spanning multiple drug discovery tasks. For molecular property prediction, we use the MoleculeNet dataset (Wu et al. 2018), which includes six classification tasks and three regression tasks. Following (Li et al. 2023), molecules are split using scaffold splitting to ensure structural diversity between training, validation, and test sets. To assess compound–target interaction prediction, we adopt two datasets curated from BindingDB by PMF-CPI (Song et al. 2023): a classification dataset with binary interaction labels, and a regression dataset with experimentally measured binding affinities. We follow their default data splitting protocol to construct the training and test sets. For ligand affinity prediction, we evaluate on three ligand datasets: FGFR1 and HPK1 from (Li et al. 2023), representing kinase targets, and VIM-1 from BindingDB, a comprehensive protein-ligand binding affinity database.

| Methods | BACE ↑ | BBBP ↑ | ClinTox ↑ | SIDER ↑ | Tox21 ↑ | ToxCast ↑ | AVE ↑ |
|---|---|---|---|---|---|---|---|
| GROVER | 0.840±0.030 | 0.887±0.006 | 0.874±0.048 | 0.638±0.005 | 0.838±0.017 | 0.696±0.014 | 0.796 |
| GraphLoG | 0.830±0.014 | 0.846±0.008 | 0.667±0.021 | 0.615±0.013 | 0.796±0.025 | 0.677±0.019 | 0.739 |
| MolCLR | 0.796±0.057 | 0.914±0.015 | 0.869±0.048 | 0.615±0.018 | 0.773±0.038 | 0.622±0.010 | 0.765 |
| MoleBERT | 0.843±0.031 | 0.851±0.022 | 0.797±0.074 | 0.615±0.010 | 0.832±0.021 | 0.720±0.009 | 0.776 |
| KPGT | 0.855±0.011 | 0.908±0.010 | 0.946±0.022 | 0.649±0.009 | 0.848±0.013 | **0.746±0.002** | 0.825 |
| MoleculeSTM | 0.812±0.008 | 0.880±0.013 | 0.875±0.031 | 0.615±0.018 | 0.813±0.023 | 0.730±0.013 | 0.788 |
| SPMM | 0.834±0.016 | 0.914±0.015 | 0.897±0.014 | 0.620±0.010 | 0.821±0.020 | 0.708±0.011 | 0.799 |
| MolTailor | - | 0.857±0.034 | 0.846±0.048 | - | 0.790±0.026 | - | - |
| **PharmaQA** | **0.880±0.011** | **0.936±0.016** | **0.970±0.008** | **0.667±0.017** | **0.850±0.010** | **0.746±0.008** | **0.842** |

Table 1: AUC performance on MoleculeNet classification datasets (mean ± std).

## Experimental Setup

**Implementation Details.** We employ pretrained encoders with selective finetuning, balancing efficiency and performance. To ensure fair comparisons with baseline methods, we follow the same data splitting protocols and evaluation settings as established in previous works (Li et al. 2023; Song et al. 2023). For most experiments, each setting is repeated three times with different random seeds, reporting the mean and standard deviation to demonstrate result stability. However, for the BindingDB experiments, we conduct a single run to align with the original evaluation protocol.

All experiments conducted in this study are executed utilizing the PyTorch deep learning framework, leveraging a single GPU which is NVIDIA GeForce RTX 4090. The training process is designed with 50 epochs, with an early stopping criterion of 20 epochs to prevent overfitting. The batch size is set to 12. In the AIP, SFM and PFM, each of these components adopts a two-layer architecture with GELU activation. In the multi-path knowledge guided bilinear attention module, we set the parameter head $M$ to 2, the stride s to 3. The hidden layer dimension is 768 and the parameters $\alpha$ and $\beta$ are both set to 0.1.

**Pharmacophore Question-Description Generation.** The ChatGPT-4o is used to generate natural language question–description pairs for each of the $P = 27$ pharmacophore types, based on domain-specific SMARTS definitions extracted from RDKit. These pairs are further grouped into 8 pharmacophore categories (e.g., Donors, Acceptors), covering the major types of molecular interactions.

**Evaluation Metrics.** We adopt task-specific evaluation metrics tailored to each dataset. For the moleculeNet dataset, we follow the evaluation protocol of (Li et al. 2023), using the Area Under ROC Curve (AUC) for classification datasets, and Root Mean Square Error (RMSE) for regression datasets. For the BindingDB dataset, we follow the setup in (Song et al. 2023), using AUC and Area Under the Precision-Recall curve (AUPR) for classification, and mean squared error (MSE) as well as Pearson correlation for regression. For the ligand prediction dataset, we evaluate the predicted binding affinities using Pearson and Spearman correlation coefficients, consistent with prior studies (Li et al. 2023). These metrics together enable a comprehensive assessment of both classification performance and regression accuracy.

## Evaluation of Molecule Property Prediction

We conducted a comprehensive evaluation of PharmaQA on MoleculeNet benchmark (Wu et al. 2018), which includes six classification and three regression datasets. Our baselines cover two categories: (1) molecular graph pretraining methods, including GraphLoG (Xu et al. 2021), GROVER (Rong et al. 2020), MolCLR (Wang et al. 2022), MoleBERT (Xia et al. 2023) and KPGT (Li et al. 2023); (2) multimodal approaches such as MoleculeSTM (Liu et al. 2023), SPMM (Chang and Ye 2024), and MolTailor (Guo et al. 2024), which leverage multiple modalities for enhanced molecular representation learning. To ensure fair comparison, we reproduced baseline results under identical experimental settings. For GraphLoG and GROVER, we adopted the results from (Li et al. 2023), where experiments were conducted with settings consistent with ours.

As shown in Table 1 and Table 2, PharmaQA consistently achieves the best performance across all molecular property prediction tasks, attaining the highest average AUC in classification and the lowest average RMSE in regression among the compared methods. While the AUC improvement over KPGT is modest (1.7%), the RMSE reduction is more notable (13.6%), reflecting stronger predictive accuracy. These results indicate that PharmaQA effectively leverages pharmacophore questions and textual description information to capture complex molecular relationships, leading to stable improvements in predictive performance.

| Methods | Lipo ↓ | Esol ↓ | Freesolv ↓ | AVE ↓ |
|---|---|---|---|---|
| GROVER | 0.752±0.010 | 0.928±0.027 | 2.991±1.052 | 1.557 |
| GraphLoG | 1.104±0.024 | 2.335±0.073 | 4.174±1.077 | 2.538 |
| MolCLR | 0.729±0.052 | 1.249±0.082 | 2.741±0.408 | 1.573 |
| MoleBERT | 0.690±0.023 | 1.185±0.083 | 2.801±0.602 | 1.559 |
| KPGT | 0.600±0.010 | 0.803±0.008 | 2.121±0.837 | 1.175 |
| MoleculeSTM | 0.706±0.032 | 1.161±0.078 | 3.244±0.634 | 1.704 |
| SPMM | 0.690±0.029 | 0.872±0.054 | 2.131±0.790 | 1.231 |
| MolTailor | 0.859±0.021 | 0.982±0.040 | 2.645±0.590 | 1.495 |
| **PharmaQA** | **0.598±0.026** | **0.794±0.047** | **1.726±0.521** | **1.039** |

Table 2: RMSE performance on MoleculeNet regression datasets (mean ± std).

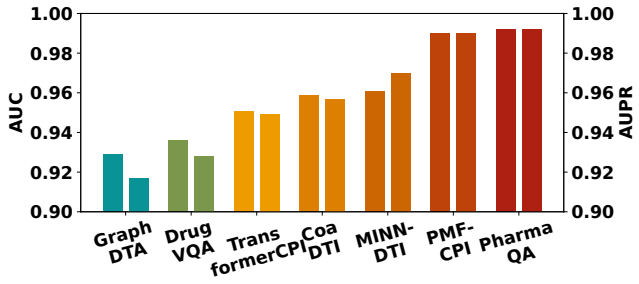



Figure 2: Performance Comparison on the BindingDB Classification Dataset (AUC, AUPR).

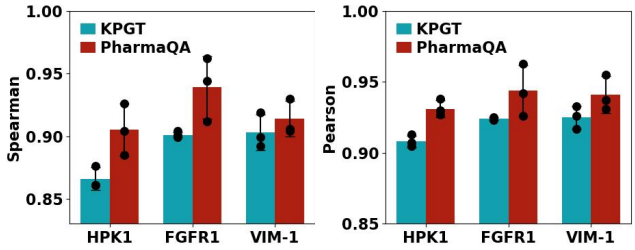



Figure 4: Affinity prediction comparison between PharmaQA and KPGT on three ligand datasets (Spearman, Pearson).

## Evaluation of Compound-Target Interaction and Affinity Prediction

We evaluate PharmaQA on the BindingDB classification and regression benchmarks. Baseline results are adopted from the PMF-CPI benchmark (Song et al. 2023), which includes several state-of-the-art models. As shown in Figure 2, PharmaQA achieves the best performance on the BindingDB classification task with an AUC of 0.992 and AUPR of 0.992, outperforming all baseline methods, including recent advanced models such as PMF-CPI (AUC 0.990, AUPR 0.990). This demonstrates the strong capability of PharmaQA in capturing discriminative interactions between compound and protein representations. For the BindingDB regression task (Figure 3), PharmaQA again surpasses all competitors, achieving the lowest MSE (0.465) and the highest Pearson correlation coefficient (0.887), highlighting the effectiveness of PharmaQA's reasoning ability over continuous binding affinity values. These results demonstrate that incorporating pharmacophore derived prompts improves performance in both classification and regression tasks on BindingDB, leading to a more accurate modeling of compound–target interactions.

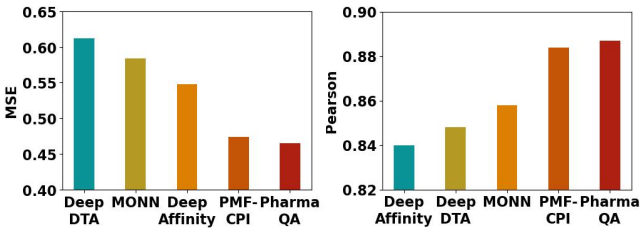



Figure 3: Performance Comparison on the BindingDB Regression Dataset (MSE, Pearson).

## Evaluation of Binding Affinity Prediction on Key Ligand Datasets

To validate the representation capability of our model on important target molecules, we conducted an experiment in which we trained our model on binding affinity datasets for three key targets:FGFR1, HPK1 and VIM-1. Fibroblast growth factor receptor 1 (FGFR1) is a transmembrane receptor tyrosine kinase that is often overexpressed or mutated in various diseases, including myeloproliferative syndromes and multiple cancers (Acevedo et al. 2007). Hematopoietic progenitor kinase 1 (HPK1) plays a critical role in the negative regulation of immune functions (Si et al. 2020). Lastly, Verona integron-encoded metallo-$\beta$-lactamase 1 (VIM-1) is capable of hydrolyzing carbapenem $\beta$-lactam antibiotics, leading to serious drug-resistant infections (Boyd et al. 2020). These targets are particularly significant as they are involved in immune regulation, cancer progression, and the development of drug-resistant infections, making them highly relevant for advancing therapeutic research and drug discovery. The prediction results were compared with the leading KPGT method, as shown in Figure 4. Our model consistently outperforms KPGT across all three ligand datasets in both Spearman and Pearson correlations. Specifically, average improvements for FGFR1 were 0.038 (Spearman) and 0.020 (Pearson); for HPK1, 0.039 and 0.023; and for VIM-1, 0.018 and 0.015. These results highlight the robustness and superior ability of our model to capture binding affinity for key therapeutic targets.

To further assess the practical utility of our model in target-specific molecular discovery, we conducted a retrospective validation using a curated set of FDA-approved compounds. For each of the three targets, we ranked compounds based on predicted binding affinity and validated the top-20 candidates through literature evidence. Notably, 13 of the top-20 molecules for FGFR1 and 11 for HPK1 were previously reported to exhibit bioactivity against the respective targets. For VIM-1, which poses a greater challenge due to zinc ion coordination, 8 of the top-ranked molecules were associated with zinc-binding proteins, supporting their potential relevance. These findings suggest that our model effectively prioritizes biologically meaningful compounds and exhibits virtual screening capability, even within a constrained chemical space.

## Ablation Studies

To assess the impact of pharmacophore prompts, we perform ablation studies on MoleculeNet classification and regression tasks, comparing PharmaQA with the following variants: removing all prompt inputs and using only the molecular graph (W/o prompt), following the KPGT (Li et al. 2023) baseline, which does not incorporate pharmacophore information; replacing structured prompts with a semantically irrelevant sentence ("To be, or not to be, that is the

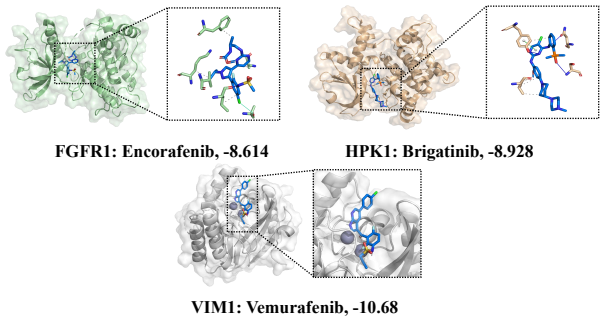



Figure 5: Docking visualization of three identified ligands with their respective targets.

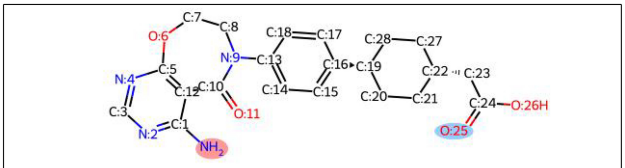
question.") to disrupt meaningful pharmacophore guidance (Noise prompt); and substituting learned prompts with handcrafted pharmacophore features from RDKit as a static alternative (Pharma_RDKit). The Table 3 presents the average performance of different model variants on MoleculeNet datasets. The W/o prompt baseline excludes all pharmacophore prompt inputs and relies solely on molecular structure encoding, achieving 0.825 AUC and 1.175 RMSE. Replacing meaningful prompts with random noise (Noise prompt) leads to a noticeable performance drop in both classification (↓3.4%) and regression (↑3.2%), highlighting the necessity of pharmacophore guided features. Using static RDKit derived prompts without further adjustment (Pharma_RDKit) also degrades AUC (↓2.2%) and RMSE (↑7.6%), suggesting that fixed patterns may introduce noise or task-irrelevant signals. In contrast, our full model (PharmaQA) outperforms all variants with 0.842 AUC (↑1.7%) and 1.039 RMSE (↓13.6%), showing that learned and integrated pharmacophore reasoning modules can enhance both classification and regression tasks.

| Variant Methods | Classification (AUC) | Regression (RMSE) |
|---|---|---|
| w/o prompt | 0.825 | 1.175 |
| Noise prompt | 0.791 (↓3.4%) | 1.207 (↑3.2%) |
| Pharma_RDKit | 0.803 (↓2.2%) | 1.251 (↑7.6%) |
| **PharmaQA** | **0.842** (↑1.7%) | **1.039** (↓13.6%) |

Table 3: Ablation Results on MoleculeNet datasets (Average across tasks).

## Case Studies

We further evaluate the molecular representations our model through two case studies. Firstly, we investigates the model's ability to identify potential ligands. Building on previous ligand discovery experiments, we examined additional top-ranked molecules that lacked direct literature evidence. Three representative candidates were selected from the top-20 sets and docked to their respective targets (FGFR1: 5A4C, HPK1: 7SIU, and VIM-1: 5N5H) using AutoDock Vina. As shown in Figure 5, all three achieved docking scores below $-7$ kcal/mol, a commonly accepted threshold for favorable binding affinity (Trujillo-Correa et al. 2019; Ahmad et al. 2021). These results provide additional structural evidence that our model can identify novel ligands with strong binding potential, even in the absence of prior experimental confirmation. Moreover, we analyzed attention maps from the final Attention layer to evaluate the model's identification of key pharmacophore information. Using the Lipo training set, we generated pharmacophore-specific attention maps for the test set, focusing on SingleAtomDonor and SingleAtomAcceptor pharmacophores. Attention weights from the graph-question and graph-description channels were extracted and ranked to identify the Top-10 tokens linked to pharmacophore atoms. As shown in Figure 6, for SingleAtomDonor at atom 0, the model highlights 'hydrogen' in the graph-question and both 'hydrogen' and 'nitrogen' in the graph-description. For SingleAtomAcceptor at atom 25, although the graph-question map lacks key tokens, the graph-description map emphasizes 'oxygen', demonstrating the model's ability to capture critical pharmacophore features.



**SingleAtomDonor Question on atom number 0:**
a, in, the, bonding, hydrogen, [CLS], atoms, atom, present, how.
**SingleAtomDonor Discription on atom number 0:**
this, or, hydrogen, hydrogen, and, description, smarts, atoms, positive, nitrogen.

**SingleAtomAcceptor Question on atom number 25:**
as, present, molecule, question, a, to, connected, such, nitrogen, charged.
**SingleAtomDonor Discription on atom number 25:**
nitrogen, with, with, but, non, oxygen, aromatic, atoms, [CLS], charge.

Figure 6: Visualization under SingleAtomDonor (atom 0, red background) and SingleAtomAcceptor (atom 25, blue background) queries. Middle and bottom: Top-10 tokens per atom with key terms color-highlighted.

## Conclusion

We propose PharmaQA, a pharmacophore-oriented question answering framework for molecular representation learning. Instead of directly encoding pharmacophore features, PharmaQA employs structured question–description pairs to guide semantic reasoning and aligns them with molecular graphs features via a multi-path knowledge guided bilinear attention module. This enables the model to generate context-aware embeddings that capture both structural and functional semantics. Extensive evaluations on molecular property prediction, compound–target interaction, and binding affinity tasks show improvements over baselines. Target-specific screening and docking validation further highlight PharmaQA's potential for identifying bioactive compounds. Attention analysis confirms its ability on pharmacophore-relevant regions, providing interpretability. Overall, PharmaQA offers a generalizable and interpretable framework for pharmacophore-based molecular learning.

## Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (NSFC 62322215, 62532017, 62402488, 62502050, 62372279), and the Natural Science Foundation of Shandong Province (ZR2025QB62, ZR2023MF119). This study was also supported in part by the High-Performance Computing Center of Central South University.

## References

Acevedo, V. D.; Gangula, R. D.; Freeman, K. W.; Li, R.; Zhang, Y.; Wang, F.; Ayala, G. E.; Peterson, L. E.; Ittmann, M.; and Spencer, D. M. 2007. Inducible FGFR-1 Activation Leads to Irreversible Prostate Adenocarcinoma and an Epithelial-to-Mesenchymal Transition. *Cancer Cell*, 12(6): 559–571.

Ahmad, S.; Waheed, Y.; Abro, A.; Abbasi, S. W.; and Ismail, S. 2021. Molecular screening of glycyrrhizin-based inhibitors against ACE2 host receptor of SARS-CoV-2. *Journal of molecular modeling*, 27(7): 206.

Boyd, S. E.; Livermore, D. M.; Hooper, D. C.; and Hope, W. W. 2020. Metallo-$\beta$-Lactamases: Structure, Function, Epidemiology, Treatment Options, and the Development Pipeline. *Antimicrobial Agents and Chemotherapy*, 64(10): 10.1128/aac.00397–20.

Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; and Walsh, A. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715): 547–555.

Chang, J.; and Ye, J. C. 2024. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1): 2323.

Chen, K.; and Wu, X. 2024. VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27218–27227.

Chen, S.; Chen, J.; Zhou, S.; Wang, B.; Han, S.; Su, C.; Yuan, Y.; and Wang, C. 2024. SIGformer: Sign-aware Graph Transformer for Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 1274–1284. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.

Dickson, M.; and Gagnon, J. P. 2004. Key factors in the rising cost of new drug discovery and development. *Nature reviews Drug discovery*, 3(5): 417–429.

Dong, J.; Wang, N.-N.; Yao, Z.-J.; Zhang, L.; Cheng, Y.; Ouyang, D.; Lu, A.-P.; and Cao, D.-S. 2018. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *Journal of cheminformatics*, 10: 1–11.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for Quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1263–1272. JMLR.org.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1).

Guo, H.; Zhao, S.; Wang, H.; Du, Y.; and Qin, B. 2024. Moltailor: Tailoring chemical molecular representation to specific tasks via text prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18144–18152.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations*.

Jiang, Y.; Jin, S.; Jin, X.; Xiao, X.; Wu, W.; Liu, X.; Zhang, Q.; Zeng, X.; Yang, G.; and Niu, Z. 2023a. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Communications Chemistry*, 6(1): 60.

Jiang, Y.; Jin, S.; Jin, X.; Xiao, X.; Wu, W.; Liu, X.; Zhang, Q.; Zeng, X.; Yang, G.; and Niu, Z. 2023b. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. *Communications Chemistry*, 6(1): 60.

Li, F.; Yin, J.; Lu, M.; Mou, M.; Li, Z.; Zeng, Z.; Tan, Y.; Wang, S.; Chu, X.; Dai, H.; Hou, T.; Zeng, S.; Chen, Y.; and Zhu, F. 2022. DrugMAP: molecular atlas and pharma-information of all drugs. *Nucleic Acids Research*, 51(D1): D1288–D1299.

Li, H.; Zhang, R.; Min, Y.; Ma, D.; Zhao, D.; and Zeng, J. 2023. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1): 7568.

Li, J.; Wang, J.; Zhang, Z.; and Zhao, H. 2024. Self-Prompting Large Language Models for Zero-Shot Open-Domain QA. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 296–310. Mexico City, Mexico: Association for Computational Linguistics.

Liu, J.; Li, L.; Rao, S.; Gao, X.; Guan, W.; Li, B.; and Ma, C. 2025. Union Is Strength! Unite the Power of LLMs and MLLMs for Chart Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 5487–5495.

Liu, S.; Nie, W.; Wang, C.; Lu, J.; Qiao, Z.; Liu, L.; Tang, J.; Xiao, C.; and Anandkumar, A. 2023. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12): 1447–1457.

Mullard, A. 2014. New drugs cost US $2.6 billion to develop. *Nature reviews drug discovery*, 13(12).

Park, J.; Jang, K. J.; Alasaly, B.; Mopidevi, S.; Zolensky, A.; Eaton, E.; Lee, I.; and Johnson, K. 2024. Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models. *ArXiv*, abs/2408.12763.

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; WEI, Y.; Huang, W.; and Huang, J. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12559–12571. Curran Associates, Inc.

Si, J.; Shi, X.; Sun, S.; Zou, B.; Li, Y.; An, D.; Lin, X.; Gao, Y.; Long, F.; Pang, B.; Liu, X.; Liu, T.; Chi, W.; Chen, L.; Dimitrov, D. S.; Sun, Y.; Du, X.; Yin, W.; Gao, G.; Min, J.; Wei, L.; and Liao, X. 2020. Hematopoietic Progenitor Kinase1 (HPK1) Mediates T Cell Dysfunction and Is a Druggable Target for T Cell-Based Immunotherapies. *Cancer Cell*, 38(4): 551–566.e11.

Song, N.; Dong, R.; Pu, Y.; Wang, E.; Xu, J.; and Guo, F. 2023. PMF-CPI: assessing drug selectivity with a pretrained multi-functional model for compound–protein interactions. *Journal of Cheminformatics*, 15(1): 97.

Sun, R.; Dai, H.; and Yu, A. W. 2022. Does gnn pretraining help molecular representation? *Advances in Neural Information Processing Systems*, 35: 12096–12109.

Trujillo-Correa, A. I.; Quintero-Gil, D. C.; Diaz-Castillo, F.; Quiñones, W.; Robledo, S. M.; and Martinez-Gutierrez, M. 2019. In vitro and in silico anti-dengue activity of compounds obtained from Psidium guajava through bioprospecting. *BMC complementary and alternative medicine*, 19: 1–16.

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

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

Xia, J.; Zhao, C.; Hu, B.; Gao, Z.; Tan, C.; Liu, Y.; Li, S.; and Li, S. Z. 2023. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In *The Eleventh International Conference on Learning Representations*.

Xu, M.; Wang, H.; Ni, B.; Guo, H.; and Tang, J. 2021. Self-supervised Graph-level Representation Learning with Local and Global Structure. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 11548–11558. PMLR.

Xu, Z.; Wang, S.; Zhu, F.; and Huang, J. 2017. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, 285–294.

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

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

Yu, J.-L.; Zhou, C.; Ning, X.-L.; Mou, J.; Meng, F.-B.; Wu, J.-W.; Chen, Y.-T.; Tang, B.-D.; Liu, X.-G.; and Li, G.-B. 2025. Knowledge-guided diffusion model for 3D ligand-pharmacophore mapping. *Nature Communications*, 16(1): 2269.

Zhong, W.; Gao, Y.; Ding, N.; Qin, Y.; Liu, Z.; Zhou, M.; Wang, J.; Yin, J.; and Duan, N. 2022. ProQA: Structural Prompt-based Pre-training for Unified Question Answering. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4230–4243. Seattle, United States: Association for Computational Linguistics.

Zhu, H.; Zhou, R.; Cao, D.; Tang, J.; and Li, M. 2023. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications*, 14(1): 6234.