# MolTC: Towards Molecular Relational Modeling In Language Models

**Junfeng Fang**[1*] **Shuai Zhang**[1*] **Chang Wu**[1] **Zhengyi Yang**[1]
**Zhiyuan Liu**[2] **Sihang Li**[1] **Kun Wang**[1] **Wenjie Du**[1†] **Xiang Wang**[1†]
[1]University of Science and Technology of China [2]National University of Singapore
{fjf,shuaizhang,wuchang0124,yangzhy}@mail.ustc.edu.cn
{acharkq,sihang0520}@gmail.com, {wk520529,duwenjie}@mail.ustc.edu.cn
xiangwang1223@gmail.com

## Abstract

Molecular Relational Learning (MRL), aiming to understand interactions between molecular pairs, plays a pivotal role in advancing biochemical research. Recently, the adoption of large language models (LLMs), known for their vast knowledge repositories and advanced logical inference capabilities, has emerged as a promising way for efficient and effective MRL. Despite their potential, these methods predominantly rely on textual data, thus not fully harnessing the wealth of structural information inherent in molecular graphs. Moreover, the absence of a unified framework exacerbates the issue of insufficient data exploitation, as it hinders the sharing of interaction mechanisms learned across various datasets. To address these challenges, this work proposes a novel LLM-based multi-modal framework for **Mol**ecular in**T**eraction modeling following **C**hain-of-Thought (CoT) theory, termed **MolTC**, which effectively integrate graphical information of two molecules in pair. To train MolTC efficiently, we introduce a *Multi-hierarchical CoT* theory to refine its training paradigm, and conduct a comprehensive **Molecular Interactive Instructions** dataset for the development of biochemical LLMs involving MRL. Our experiments, conducted across twelve datasets involving over 4,000,000 molecular pairs, exhibit the superiority of our method over current GNN- and LLM-based baselines. Our code is available at https://github.com/MangoKiller/MolTC.

## 1 Introduction

Molecular Relational Learning (MRL) (Lee et al., 2023a), aiming to understand interactions between molecular *pairs*, has gained significant interest due to its wide range of applications (Roden et al., 2020). For example, Drug-Drug Interactions (DDIs) are critical in pharmacology and drug development (Lin et al., 2020), while solute-solvent interactions (SSIs) are fundamental in solution chemistry and the design of chemical processes (Varghese and Mushrif, 2019; Chung et al., 2022). However, the exhaustive experimental validation of these interactions is notoriously time-consuming and costly. In response, adopting large language models (LLMs) (Brown et al., 2020; Taylor et al., 2022), known for their vast knowledge repositories and advanced logical inference capabilities, has emerged as an efficient and effective alternative for MRL (Park et al., 2022; Jha et al., 2022a).

Despite their promise, a primary concern of current LLM-based paradigm is the *insufficient data exploitation*. Specifically, they predominantly rely on the textual data such as SMILES (Simplified Molecular Input Line Entry System) and property descriptions, thus not fully harnessing the wealth of structural information inherent in molecular graphs (Sagawa and Kojima, 2023), as indicated in Figure 1 (a). Current studies have indicated that it is challenging for LLMs to fully understand the complex graphs based solely on textual data, hence, it's crucial to explicitly model these structures given their significance in MRL (Park et al., 2022).

Compounding this concern is the absence of a unified framework for LLM-based MRL (Livne et al., 2023; Pei et al., 2023). Concretely, this absence impedes the sharing and integration of interaction mechanisms learned across various datasets, leading to a fragmentation in collective insights. Especially, it poses a catastrophic challenge for tasks with a limited number of labeled pairs (Chung et al., 2022), where LLMs often struggle with due to the high risk of overfitting, as illustrated in Figure 1 (b). Worse still, such limited datasets are prevalent in MRL since the experimental acquisition is often constrained by high costs (Lee et al., 2023a).

To overcome these limitations, in this work, we propose **MolTC**, a unified multi-modal frame-

[*]Equal contribution.
[†]Corresponding author. Xiang Wang is also affiliated with Institute of Dataspace, Hefei Comprehensive National Science Center.



*Findings of the Association for Computational Linguistics: ACL 2024*, pages 1943–1958
August 11-16, 2024 

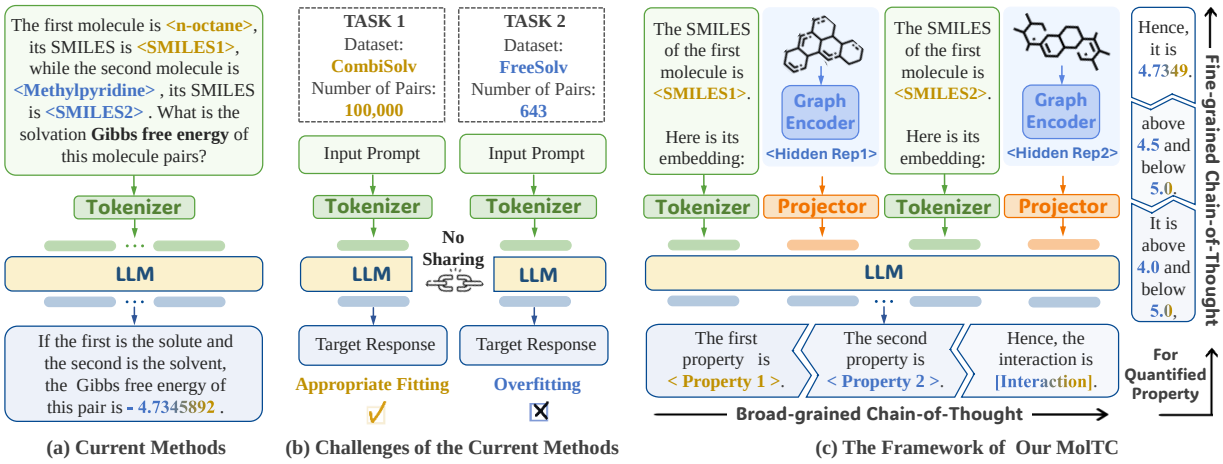


Figure 1: Comparison between the current methods leveraging LLMs to model molecule interactions and our MolTC. (a) The prevailing paradigm of current methods. (b) The challenge of applying the current paradigm to the tasks involving datasets with a small number of samples. (c) The framework of our proposed MolTC, which is enhanced by the principle of CoT. Best viewed in color.

work for **Mol**ecular in**T**eraction modeling following the **Chain**-of-thought theory (Wei et al., 2022). As depicted in Figure 1 (c), MolTC employs the Graph Neural Networks (GNNs) (Kipf and Welling, 2017), known for their proficiency in graph modeling, to explicitly gather graphical information of molecular pairs, and integrates them into the input space of LLMs by two meticulously crafted projectors. In response to empirical findings that LLMs may confuse two input molecules in pair, MolTC incorporates the molecules' SMILES information to reinforce the concept of molecular order.

More importantly, a two-pronged approach is developed to train MolTC efficiently:

**(1) Training Paradigm Refinement:** As shown in Figure 1 (c), we introduce a *Multi-hierarchical CoT* theory to guide the training paradigm of MolTC. Concretely, the broad-grained CoT guides the pre-training stage to identify individual molecular properties before predicting interactions, ensuring an acute awareness of each molecule's unique attribute. For quantitative interaction tasks, which are challenging for LLMs, a fine-grained CoT enables the fine-tuning stage to initially predict a range, and then progressively refining it to a precise value.

**(2) Dataset Foundation Construction:** In sight of the absence of a comprehensive MRL datasets for biochemical LLMs, we construct a **Mol**ecular in**T**eractive instructions dataset, termed **MoT-instruction**. Specifically, we first conduct twelve well-established MRL datasets across various domain, and source their detailed molecular properties from authoritative biochemical databases. Based on this, we meticulously compile these properties and empirically determine their optimal instructions. These process ensures that MoT-instructions can not only enhance the performance of our MolTC, but also contribute to the development of other biochemical LLMs involving MRL.

Our contributions can be summarized as follows:

- We identify the issue of insufficient data exploitation in current LLM-based MRL, and take the first attempt to develop a unified multi-modal framework for LLM-based MRL, named MolTC.
- We introduce the multi-hierarchical CoT theory to enhance the MolTC's training process, especially for quantitative interaction tasks.
- We construct MoT-instructions, the first comprehensive instruction dataset in MRL domain, to enhance the development of biochemical LLMs involving MRL.
- Our experiments, across over 4,000,000 molecular pairs in various domains such as DDI and SSI, demonstrate the superiority of our method over current GNN and LLM-based baselines.

## 2 Methodology

In this section, we detail our MolTC, which harnesses the power of LLMs for comprehending molecular interactions. We begin with the introduction of model framework in Section 2.1. Taking a step further, the training paradigm guided by the principle of Multi-hierarchical CoT is outlined in Section 2.2. Moreover, the dynamic parameter

sharing strategy tailored for MolTC and our developed datasets, MoT-instructions, are elaborated in Section 2.3 and 2.4, respectively.

### 2.1 Framework of MolTC

Here we introduce four key components of MolTC's framework: Graph Encoder, Representation Projector, SIMLES Injector, and the backbone LLM. The specific instantiation details can be found in the experimental section and the appendix.

**Graph Encoder.** The first step of extracting interactions is to precisely encode the molecular graphs. In sight of this, we utilize two GNN-based encoders to capture the embedding of the given molecular pairs, leveraging the GNN's robust capability in aggregating structural information. More formally, let $\mathcal{G}_a = \{\mathcal{V}_a, \mathcal{E}_a\}$ and $\mathcal{G}_b = \{\mathcal{V}_b, \mathcal{E}_b\}$ denote the input pair, where $\mathcal{V}, \mathcal{E}$ represent atomic nodes and the chemical bonds, respectively. The two graph encoders $f_{\texttt{enc1}}$ and $f_{\texttt{enc2}}$ perform aggregating to obtain the atomic embedding:

$$\begin{aligned} \mathbf{H}_a &= [\boldsymbol{h}_a^1, \boldsymbol{h}_a^2, \ldots, \boldsymbol{h}_a^{|\mathcal{V}_a|}] = f_{\texttt{enc1}}(\mathcal{G}_a), \\ \mathbf{H}_b &= [\boldsymbol{h}_b^1, \boldsymbol{h}_b^2, \ldots, \boldsymbol{h}_b^{|\mathcal{V}_b|}] = f_{\texttt{enc2}}(\mathcal{G}_b), \end{aligned} \tag{1}$$

where $\boldsymbol{h}_a^i$ and $\boldsymbol{h}_b^i$ denote to the embedding of the $i$-th atom in molecule $\mathcal{G}_a$ and $\mathcal{G}_b$; $\mathcal{V}_a$ and $\mathcal{V}_b$ represent the number of nodes.

**Representation Projector.** After acquiring molecular pair representations $\mathbf{H}_a$ and $\mathbf{H}_b$, the next step is to map them into the backbone LLM's hidden space using Projectors $f_{\texttt{pro1}}$ and $f_{\texttt{pro2}}$. These projectors serve as pivotal connectors, translating $\mathbf{H}_a$ and $\mathbf{H}_b$ into LLM-comprehensible encodings $\mathbf{M}_a$ and $\mathbf{M}_b$. Drawing inspiration from the state-of-the-art vision-language models, we instantiate $f_{\texttt{pro1}}$ and $f_{\texttt{pro2}}$ by Querying Transformers (Q-Formers) (Li et al., 2023a; Dai et al.). More formally,

$$\begin{aligned} \mathbf{M}_a &= [\boldsymbol{m}_a^1, \boldsymbol{m}_a^2, \ldots, \boldsymbol{m}_a^q] = f_{\texttt{pro1}}(\mathbf{H}_a), \\ \mathbf{M}_b &= [\boldsymbol{m}_b^1, \boldsymbol{m}_b^2, \ldots, \boldsymbol{m}_b^q] = f_{\texttt{pro2}}(\mathbf{H}_b), \end{aligned} \tag{2}$$

where $q$ denotes the number of learnable query tokens of Q-Former's transformer.

In detail, our Projectors, based on the BERT architecture, incorporate an additional cross-attention module positioned between the self-attention and feed-forward modules. This instantiation offers two key benefits. Firstly, it supports seamless integration with conventional BERT-based text encoders, allowing $f_{\texttt{pro1}}$ and $f_{\texttt{pro2}}$ pre-training with extensive molecular graph-text pairs. Secondly, it maintains compatibility with various input dimensions $d$, and allows adjustments in the size of learnable query tokens to align with the LLM's token embedding size. These advantages lay a solid foundation for the thorough interaction of two molecules during the LLM's inference process. Future work will also explore more projector designs, such as streamlining it through specially tailored MLPs (Yang et al., 2023).

**SMILES Tokenization.** When directly analyzing the representations $\mathbf{M}_a$ and $\mathbf{M}_b$ with LLMs, our experiments suggest a potential confusion by LLMs in distinguishing the properties of each molecule in a pair. This observation naturally inspires us to integrate textual information of the molecules to strengthen the concept of their sequential order. Here MolTC employs SMILES due to its ubiquity and specificity. Additionally, SMILES serves as a conduit, linking the task-specific prompts with the corresponding biochemical knowledge stored within the LLM. Therefore, we directly input the SMILES of both molecules into the backbone LLM, utilizing the inherent encoder to acquire their tokens $\mathbf{S}_a$ and $\mathbf{S}_b$.

**Backbone LLM.** MolTC leverages Galactica, a decoder-only transformer built on the OPT framework, as its base LLM. Pretrained on an extensive collection of scientific literature, Galactica demonstrates exceptional proficiency in biochemistry knowledge. This expertise, particularly in parsing molecular sequences such as SMILES and SELFIES strings, enables Galactica to adeptly capture the properties crucial for molecular interactions. Specifically, the goal of MolTC is to harness Galactica's advanced inferential skills to interpret the contextual interactions between two molecular sets of token collections, $\{\mathbf{M}_a, \mathbf{S}_a\}$ and $\{\mathbf{M}_b, \mathbf{S}_b\}$. More formally, we denote an integrated prompt sequence as follows:

$$\begin{aligned} \mathbf{X} = \{\mathbf{P}, \mathbf{M}_a, \mathbf{S}_a, \mathbf{M}_b, \mathbf{S}_b\} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_l] \\ \text{s.t. } \mathbf{P} \sim \mathcal{P}_\mathbf{r}, \end{aligned} \tag{3}$$

where $l$ is the integrated input length, $\mathbf{P}$ denotes the task-specific prompt, and $\mathcal{P}_\mathbf{r}$ represents a collection of various manually designed prompts, each tailored for the molecular interaction task $\mathbf{r}$. The generation process adopts a causal mask to generate a response encapsulating key interactive properties

with length $T$:

$$\hat{\mathbf{X}} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, ..., \hat{\boldsymbol{x}}_T]. \quad (4)$$

Utilizing Galactica's autoregressive framework, the training objective involves regressing the target response based on the input prompt $\mathbf{X}$. Specifically, the output for $i$-th token $\hat{\boldsymbol{x}}_i$, is computed based on its preceding tokens as follows for $t \in (1, T)$:

$$p\left(\hat{\mathbf{X}}_{[1:t]}|\mathbf{X}\right) = \prod_{i=1}^{t} p\left(\hat{\boldsymbol{x}}_i|\mathbf{X}, \hat{\mathbf{X}}_{[1:i-1]}\right). \quad (5)$$

### 2.2 Training Paradigm of MolTC

In this part, we elaborate the training paradigm of MolTC, including pretraining and fine-tuning processes, which is guided by the principle of Multi-hierarchical CoT, as shown in Figure 2.

#### 2.2.1 Broad-grained CoT Guided Pretraining

Given the challenge of directly understanding complex interactions between two input molecules in pair, the broad-grained CoT guides MolTC to initially identify individual molecular properties. By thoroughly understanding each molecule's characteristics, MolTC establishes a solid foundation for accurately predicting their interactions. Specifically, in the pretraining stage, the prompt is uniformly designed as follows:

| Prompt for Pretraining Stage | |
|---|---|
| Input Prompt | <SMILES1>, <GraEmb1>, the front is the first molecule, followed by the second molecule: <SMILES2>. <GraEmb2>. Please provide the biochemical properties of the two molecules one by one. |
| Target Response | The properties of the first molecule are [Property1], and the properties of the second molecule are [Property2]. |

This prompt design enable MolTC to delineate key properties of two molecules sequentially. Based on it, MolTC utilize the generation loss of the backbone LLM to train Graph Encoders, $f_{\text{enc1}}$ and $f_{\text{enc2}}$, as well as the Representation Projectors, $f_{\text{pro1}}$ and $f_{\text{pro2}}$. Notably, during this phase, the backbone LLM remains frozen.

**Dataset Construction for Pretraining.** To ensure backbone LLM can understand the individual characteristics of each molecule, it is pivotal to prepare a comprehensive dataset comprising molecule pairs and their corresponding biochemical properties. To this end, (1) we first conduct an extensive survey of various authoritative biochemical database such as PubChem[1] and Drugbank (Kim et al., 2023), and collect a large amount of molecule-textual properties pairs; (2) then, recognizing the variability in annotation quality within this dataset, we augment and enrich molecular descriptions that were less extensively annotated; (3) subsequently, to simulate diverse molecular interactions, we generated molecular pairs by randomly grouping two distinct molecules from the above database. This random pairing facilitates a broad spectrum of molecular combinations, exposing the pretraining stage to diverse interaction scenarios, thus naturally enhancing the generalizability of our MolTC.

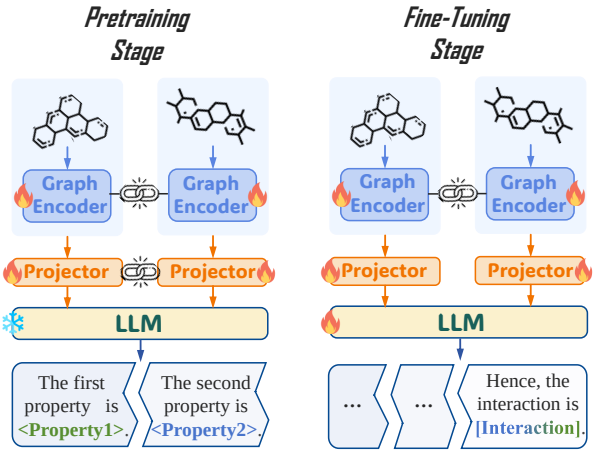


Figure 2: The training process of our MolTC. The flame symbol denotes the parameter update, the snowflake symbol indicates the parameter freezing, and the chain symbol depicts the parameter sharing between two modules. Best viewed in color.

#### 2.2.2 Fine-grained CoT Guided Fine-tuning

During the fine-tuning phase, MolTC is trained to enable the backbone LLM to generate interaction properties based on the properties of individual molecules it initially identifies. To this end, prompts in the fine-tuning stage should be crafted for specific downstream task. For example, in DDI tasks, we construct the following prompt:

[1] https://pubchem.ncbi.nlm.nih.gov

| Prompt for DDI Tasks (Fine-tuning) | |
|---|---|
| Input Prompt | <SMILES1>, <GraEmb1>, the front is the first molecule, followed by the second molecule: <SMILES2>. <GraEmb2>. What are the side effects of these two drugs? |
| Target Response | The property of the first molecule is [Property1], while the property of the second molecule is [Property2]. Hence, the first drug molecule may increase the photosensitizing activities of the second drug molecule. |

Despite the effectiveness of this prompt design, LLMs face notable challenges in quantitative analysis, especially in complex molecular interaction contexts such as SSI and chromophore-solvent interaction (CSI). Our experiments in Section 3 highlight this difficulty, demonstrating that LLMs tend to exhibit indecision regarding the quantitative values in their outputs. To address this, a fine-grained CoT concept is introduced to refine the training paradigm. Specifically, the backbone LLM is guided to initially suggest a range for the target numerical value, then progressively refining it to a precise value. Take a meticulously prompt for SSI tasks as an example:

| Prompt for SSI Tasks (Fine-tuning) | |
|---|---|
| Input Prompt | <SMILES1>, <GraEmb1>, the front is the first molecule, followed by the second molecule: <SMILES2>. <GraEmb2>. What is the solvation Gibbs free energy of this pair of molecules? |
| Target Response | The property of the first molecule is [Property1], while the property of the second molecule is [Property2]. Hence, the solvation Gibbs free energy of these two molecules is above 3.0 and below 3.5, so the accurate value is 3.24791. |

This step-wise refinement process fosters a more accurate and reliable resolution of numerically-intensive challenges. Based on these prompts, in the fine-tuning stage, the parameters in backbone LLM are updated through Low-Rank Adaptation (LoRA) (Hu et al., 2021) strategy, known for its efficiency in tailoring the LLM to the requirements of downstream tasks and minimal memory demands in storing gradients. Meanwhile, to ensure that other modules are optimally adjusted to suit the specifics of the downstream tasks, Graph Encoders $f_{\texttt{enc1}}$ and $f_{\texttt{enc2}}$, as well as Representation Projectors $f_{\texttt{pro1}}$ and $f_{\texttt{pro2}}$ are trained following the generation loss of the backbone LLM.

### 2.3 Dynamic Parameter Sharing Strategy

To implement the above training paradigm effectively, we introduce a novel parameter-sharing strategy, inspired by key biochemical insights:

(1) The Importance of **Role-Playing**: A molecule's role in an interaction crucially influences the outcome. For example, in SSI scenario like the water-ethanol pair, utilizing water and ethanol as solvents, respectively, yields different energy releases (Reichardt, 2021). Sometimes, a reversal of roles can even result in the absence of interaction.

(2) The Importance of **Input Order**: In certain molecular pairs, the sequence of introducing molecules significantly impacts the interactions. For instance, the order of drug introduction can lead to varying therapeutic effects.

(3) The Importance of **Role and Order-Specific** Feature Extraction: The role and input order of molecules determine the relevance of their structural features. For example, a chemical group in a solute-solvent pair may be crucial for the release of Gibbs free energy when in the solute, but less so in the solvent (Reichardt, 2021; J et al., 2022).

These insights inspire MolTC to adaptively prioritize distinct key information, creating unique tokens for the same molecule based on its role and order. To enable this nuanced learning while also capitalizing on the shared aspects of molecular learning, we introduce the following parameter-sharing strategy, as shown in Figure 2:

(1) The GNN-based **Encoders** $f_{\texttt{enc1}}$ and $f_{\texttt{enc2}}$, which focus on extracting molecular graph structures, share parameters during both pretraining and fine-tuning stages to enhance learning efficiency.

(2) The Qformer-based **Projectors** $f_{\texttt{pro1}}$ and $f_{\texttt{pro2}}$, tasked with aligning molecular structures to semantic information, share parameters during pretraining stage to promote generalization and robustness. However, in the fine-tuning stage, we cease sharing

to allow customized semantic mappings tailored to the varying roles and orders.

In summary, this strategy is tailored to balance the need for role and order-based distinctively learning with the efficiency gained from commonalities across molecular pairs.

### 2.4 Construction of MoT-instructions

Given the absence of a comprehensive instruction datasets tailored for LLM-based MRL, we aim to develop a molecular interactive instructions dataset, termed MoT-instructions. This dataset is designed to fulfill several key criteria: (1) it should include extensive molecular pairs capable of interaction, covering a broad spectrum of domains, (2) it should detail important biochemical properties of each molecule within these pairs, and (3) it should elaborate the resultant properties from molecular interactions. Specifically, MoT-instructions are constructed through a three-step process as follows.

(1) We begin by aggregating twelve representative molecular interaction datasets across various widely recognized biochemical tasks, such as DDI, SSI, and CSI. Following this, we engage in a systematic search for textual descriptions of the biochemical properties of each molecule involved in these interactions. Specifically, we source this information from authoritative biochemical databases such as DrugBank and PubChem.

(2) The next critical step is the **experimental determination of the optimal instructions**. Specifically, for all molecular pairs in step (1), we first deconstruct the lengthy molecular properties into a series of questions and answers, a format more comprehensible to LLMs (Taylor et al., 2022). The granularity of this deconstruction is decided based on the performance of our MolTC. For more challenging quantitative tasks, instructions guided by fine-grained CoT are required to provide a numerical range before specifying a concrete value. Given the vast number of possible correct ranges, exhaustive testing is impractical. Therefore, we initially determine the optimal range for a small subset of datasets using a grid search, guided by the predictive performance of MolTC. Subsequently, we derive statistics, such as mean and standard deviation, from these datasets to establish a relationship between statistics and optimal ranges. Finally, for other datasets, we determine their optimal range based on this established rule.

(3) The final step in our dataset construction involved filtering out pairs that lacked sufficient information on molecule properties or interaction data. Specifically, partial properties of a molecular pair are often missing in some datasets. To maximize the utilization of information from these datasets, we consider extracting each property within them as a separate dataset. This approach allows us to naturally omit missing values without wasting other information present in the molecular pair.

## 3 Experiment

In this section, we aim to answer the following research questions:

- **RQ1:** Is MolTC capable of generating the interactive property, involving the *qualitative* knowledge, of the given molecular pair?
- **RQ2:** Does MolTC have the ability to generate the interactive property, involving the *quantitative* property, for a given molecular pair?
- **RQ3:** What is the impact of the proposed strategies, such as the CoT enhancement strategy and SMILES injection strategy, on the inference process of our MolTC?

### 3.1 Experimental Setting

We evaluate MolTC on twelve well-established downstream molecule interaction tasks involving qualitative and quantitative analysis. Here we provide a brief overview of our experimental setup. Detailed descriptions are presented in the appendix.

**Datasets.** We employ 12 datasets across various domains such as DDI, SSI, and CSI. Specifically, we collect *Drugbank* (Version 5.0.3), *ZhangDDI* (Zhang et al., 2017), *ChChMiner* (Zitnik et al., 2018), *DeepDDI* (Ryu et al., 2018), *TWOSIDES* (Tatonetti et al., 2012), *Chromophore* (Joung et al., 2020), *MNSol* (Marenich et al., 2020), *CompSol* (Moine et al., 2017), *Abraham* (Grubbs et al., 2010), *CombiSolv* (Vermeire and Green, 2021), *FreeSolv* (Mobley and Guthrie, 2014) and *CombiSolv-QM* (Vermeire and Green, 2021).

**Baselines.** For a comprehensive evaluation, we conduct various baseline methods encompassing distinct categories such as methods based on: GNNs, DL models other than GNN, and LLMs. Specifically, For DDI task, we employ *GoGNN* (Wang et al., 2020), *MHCADDI* (Deac et al., 2019), *DeepDDI* (Ryu et al., 2018), *SSI-DDI*, *CGIB* (Lee et al., 2023a), *CMRL* (Lee et al., 2023b), *MDF-SA-DDI* (Lin et al., 2022), *DSN-DDI* (Li et al., 2023c)

Table 1: Comparative performance of various methods in qualitative interactive tasks. The best-performing methods are highlighted with a gray background, while the second-best methods are underscored for emphasis.

| Baseline Model | | Drugbank | | ZhangDDI | | ChChMiner | | DeepDDI | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | AUC-ROC | Accuracy | AUC-ROC | Accuracy | AUC-ROC | Accuracy | AUC-ROC |
| GNN Based | GoGNN | $84.78_{\pm 0.57}$ | $91.63_{\pm 0.66}$ | $84.10_{\pm 0.46}$ | $92.35_{\pm 0.48}$ | $91.17_{\pm 0.46}$ | $96.64_{\pm 0.40}$ | $93.54_{\pm 0.35}$ | $92.71_{\pm 0.27}$ |
| | SSI-DDI | $94.12_{\pm 0.33}$ | $98.38_{\pm 0.31}$ | $86.97_{\pm 0.37}$ | $93.76_{\pm 0.34}$ | $93.26_{\pm 0.31}$ | $97.81_{\pm 0.22}$ | $95.27_{\pm 0.25}$ | $98.42_{\pm 0.31}$ |
| | DSN-DDI | $\underline{94.93}_{\pm 0.14}$ | $\underline{99.01}_{\pm 0.12}$ | $87.65_{\pm 0.13}$ | $94.63_{\pm 0.18}$ | $84.30_{\pm 0.17}$ | $94.25_{\pm 0.26}$ | $95.64_{\pm 0.18}$ | $98.01_{\pm 0.16}$ |
| | CMRL | $94.83_{\pm 0.12}$ | $98.76_{\pm 0.10}$ | $\underline{87.78}_{\pm 0.36}$ | $\underline{94.68}_{\pm 0.23}$ | $94.23_{\pm 0.26}$ | $98.37_{\pm 0.12}$ | $\underline{96.37}_{\pm 0.34}$ | $\underline{98.98}_{\pm 0.31}$ |
| | CGIB | $94.68_{\pm 0.34}$ | $98.60_{\pm 0.25}$ | $87.32_{\pm 0.71}$ | $94.18_{\pm 0.60}$ | $\underline{94.25}_{\pm 0.39}$ | $\underline{98.45}_{\pm 0.31}$ | $96.23_{\pm 0.52}$ | $98.45_{\pm 0.64}$ |
| ML Based | DeepDDI | $93.15_{\pm 0.25}$ | $98.06_{\pm 0.54}$ | $83.35_{\pm 0.49}$ | $91.13_{\pm 0.58}$ | $90.34_{\pm 0.62}$ | $95.73_{\pm 0.37}$ | $92.39_{\pm 0.38}$ | $98.11_{\pm 0.42}$ |
| | MHCADDI | $78.50_{\pm 0.80}$ | $86.33_{\pm 0.35}$ | $77.86_{\pm 0.59}$ | $86.94_{\pm 0.68}$ | $84.26_{\pm 0.54}$ | $89.33_{\pm 0.82}$ | $87.01_{\pm 0.77}$ | $88.64_{\pm 0.83}$ |
| | MDF-SA-DDI | $93.86_{\pm 0.31}$ | $97.65_{\pm 0.29}$ | $86.89_{\pm 0.25}$ | $94.03_{\pm 0.23}$ | $93.64_{\pm 0.20}$ | $98.10_{\pm 0.19}$ | $95.12_{\pm 0.30}$ | $97.84_{\pm 0.36}$ |
| LLM Based | Galactica | $79.16_{\pm 0.35}$ | $86.23_{\pm 0.33}$ | $67.20_{\pm 0.46}$ | $78.74_{\pm 0.58}$ | $74.61_{\pm 0.44}$ | $83.51_{\pm 0.63}$ | $71.50_{\pm 0.41}$ | $79.07_{\pm 0.41}$ |
| | Chem T5 | $85.83_{\pm 0.31}$ | $91.97_{\pm 0.38}$ | $72.34_{\pm 0.42}$ | $89.31_{\pm 0.30}$ | $80.79_{\pm 0.52}$ | $85.65_{\pm 0.46}$ | $75.58_{\pm 0.66}$ | $84.42_{\pm 0.43}$ |
| | MolCA | $87.95_{\pm 0.52}$ | $94.00_{\pm 0.37}$ | $68.21_{\pm 0.59}$ | $88.53_{\pm 0.62}$ | $90.15_{\pm 0.43}$ | $92.92_{\pm 0.60}$ | $82.95_{\pm 0.58}$ | $88.52_{\pm 0.77}$ |
| | MolT5 | $89.49_{\pm 0.47}$ | $93.08_{\pm 0.26}$ | $76.46_{\pm 0.30}$ | $89.06_{\pm 0.33}$ | $84.70_{\pm 0.25}$ | $91.18_{\pm 0.32}$ | $86.82_{\pm 0.46}$ | $90.08_{\pm 0.57}$ |
| MolTC (Ours) | | $95.98_{\pm 0.15}$ | $99.12_{\pm 0.31}$ | $89.40_{\pm 0.12}$ | $95.48_{\pm 0.18}$ | $95.59_{\pm 0.20}$ | $98.66_{\pm 0.09}$ | $96.70_{\pm 0.26}$ | $99.05_{\pm 0.32}$ |

Table 2: Comparative performance of various methods in quantitative interactive tasks. The best-performing methods are highlighted with a gray background, while the second-best methods are underscored for emphasis.

| Baseline Model | | FreeSolv | | Abraham | | CompSol | | CombiSolv | |
|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| GNN Based | CIGIN | $0.589_{\pm 0.053}$ | $0.931_{\pm 0.066}$ | $0.314_{\pm 0.004}$ | $0.607_{\pm 0.011}$ | $0.197_{\pm 0.003}$ | $0.349_{\pm 0.005}$ | $0.288_{\pm 0.005}$ | $0.664_{\pm 0.012}$ |
| | D-MPNN | $0.702_{\pm 0.014}$ | $1.231_{\pm 0.029}$ | $0.484_{\pm 0.012}$ | $0.705_{\pm 0.025}$ | $0.205_{\pm 0.006}$ | $0.373_{\pm 0.007}$ | $0.482_{\pm 0.013}$ | $0.895_{\pm 0.055}$ |
| | GEM | $0.598_{\pm 0.018}$ | $1.188_{\pm 0.049}$ | $\underline{0.254}_{\pm 0.004}$ | $0.531_{\pm 0.005}$ | $0.203_{\pm 0.006}$ | $0.337_{\pm 0.007}$ | $0.290_{\pm 0.009}$ | $0.783_{\pm 0.020}$ |
| | CGIB | $\underline{0.541}_{\pm 0.009}$ | $\underline{0.917}_{\pm 0.055}$ | $0.258_{\pm 0.008}$ | $\underline{0.530}_{\pm 0.009}$ | $\underline{0.178}_{\pm 0.004}$ | $\underline{0.301}_{\pm 0.003}$ | $\underline{0.230}_{\pm 0.004}$ | $\underline{0.394}_{\pm 0.009}$ |
| ML Based | GOVER | $0.636_{\pm 0.026}$ | $1.074_{\pm 0.049}$ | $0.347_{\pm 0.005}$ | $0.625_{\pm 0.016}$ | $0.184_{\pm 0.005}$ | $0.371_{\pm 0.014}$ | $0.412_{\pm 0.016}$ | $0.728_{\pm 0.034}$ |
| | SolvBert | $0.602_{\pm 0.029}$ | $1.034_{\pm 0.044}$ | $0.496_{\pm 0.007}$ | $0.693_{\pm 0.014}$ | $0.192_{\pm 0.008}$ | $0.353_{\pm 0.008}$ | $0.418_{\pm 0.018}$ | $0.711_{\pm 0.020}$ |
| | Uni-Mol | $0.575_{\pm 0.060}$ | $1.012_{\pm 0.070}$ | $0.355_{\pm 0.007}$ | $0.602_{\pm 0.024}$ | $0.198_{\pm 0.002}$ | $0.344_{\pm 0.003}$ | $0.267_{\pm 0.005}$ | $0.669_{\pm 0.017}$ |
| | SMD | $0.599_{\pm 0.037}$ | $1.202_{\pm 0.036}$ | $0.400_{\pm 0.022}$ | $0.646_{\pm 0.037}$ | $0.199_{\pm 0.006}$ | $0.348_{\pm 0.007}$ | $0.657_{\pm 0.011}$ | $1.023_{\pm 0.029}$ |
| LLM Based | Galactica | $0.882_{\pm 0.010}$ | $1.438_{\pm 0.066}$ | $0.645_{\pm 0.008}$ | $1.064_{\pm 0.016}$ | $0.594_{\pm 0.006}$ | $0.854_{\pm 0.008}$ | $0.831_{\pm 0.018}$ | $1.486_{\pm 0.035}$ |
| | Chem T5 | $0.802_{\pm 0.036}$ | $1.377_{\pm 0.057}$ | $0.629_{\pm 0.010}$ | $0.910_{\pm 0.017}$ | $0.445_{\pm 0.008}$ | $0.734_{\pm 0.010}$ | $0.882_{\pm 0.015}$ | $1.297_{\pm 0.024}$ |
| | MolCA | $0.760_{\pm 0.033}$ | $1.271_{\pm 0.039}$ | $0.581_{\pm 0.007}$ | $0.897_{\pm 0.008}$ | $0.467_{\pm 0.006}$ | $0.716_{\pm 0.022}$ | $0.648_{\pm 0.033}$ | $1.125_{\pm 0.035}$ |
| | MolT5 | $0.705_{\pm 0.047}$ | $1.135_{\pm 0.069}$ | $0.549_{\pm 0.008}$ | $0.832_{\pm 0.006}$ | $0.476_{\pm 0.003}$ | $0.695_{\pm 0.013}$ | $0.652_{\pm 0.023}$ | $1.124_{\pm 0.027}$ |
| MolTC (Ours) | | $0.502_{\pm 0.011}$ | $0.684_{\pm 0.042}$ | $0.194_{\pm 0.009}$ | $0.388_{\pm 0.010}$ | $0.171_{\pm 0.006}$ | $0.295_{\pm 0.004}$ | $0.172_{\pm 0.004}$ | $0.465_{\pm 0.008}$ |

as the backbone. For SSI and CSI tasks, we utilize *D-MPNN* (Vermeire and Green, 2021), *SolvBert* (Yu et al., 2023a), *SMD* (Meng et al., 2023), *CGIB* (Lee et al., 2023a), *CIGIN* (Pathak et al., 2020), *GEM* (Fang et al., 2022), *GOVER* (Rong et al., 2020), *Uni-Mol* (Zhou et al., 2023) as the backbone. Furthermore, all downstream tasks adopt LLM-based methods, such as Galactica (Taylor et al., 2022), Chem T5 (Christofidellis et al., 2023), MolT5 (Edwards et al., 2022) and MolCA (Liu et al., 2023a) as the backbone.

**Metrics.** For qualitative tasks, we employ prediction *Accuracy* and *AUC-ROC* (Area Under the Receiver Operating Characteristic curve) as comparative metrics, while for quantitative tasks, *MAE* (Mean Absolute Error) and *RMSE* (Root Mean Square Error) are utilized as the standards.

### 3.2 Qualitative Prediction Performance (RQ1)

Table 1 exhibits the performance in qualitative interactive tasks. Due to page width limitations, only a subset of the results is presented, with additional results detailed in the appendix. From Table 1, we deduce the following observations:

**Obs.1:** MolTC consistently outshines its counterparts in qualitative interaction predictions, While GNN-based methods demonstrate commendable performance, maintaining over 90% accuracy across numerous datasets, MolTC transcends these figures in every evaluated scenario. For instance, it marks a notable 1.05% improvement in accuracy on the drugback dataset, a feat attributable to the synergy between the LLMs' reasoning faculties and the GNNs' proficiency in graph modeling.

Table 3: Performance comparison of various models on different datasets.

| Dataset | Metric | w/o SMILES | w/o CoT | |
|---|---|---|---|---|
| | | | Broad | Fine |
| DDI | Accuracy | $6.42_{\pm 0.13}$ | $2.01_{\pm 0.05}$ | – |
| | Rate (↓) | 7.08 % | 2.13 % | – |
| | ACC-AUC | $7.87_{\pm 0.32}$ | $2.98_{\pm 0.08}$ | – |
| | Rate (↓) | 8.22 % | 3.10 % | – |
| SSI | MAE | $0.025_{\pm 0.004}$ | $0.010_{\pm 0.002}$ | $0.036_{\pm 0.007}$ |
| | Rate (↑) | 11.32 % | 4.56 % | 16.40 % |
| | RMSE | $0.045_{\pm 0.007}$ | $0.014_{\pm 0.003}$ | $0.054_{\pm 0.009}$ |
| | Rate (↑) | 9.47 % | 2.95 % | 11.37 % |
| CSI Abs. | MAE | $2.06_{\pm 0.11}$ | $0.51_{\pm 0.03}$ | $2.65_{\pm 0.16}$ |
| | Rate (↑) | 15.03 % | 3.72 % | 19.34 % |
| | RMSE | $3.37_{\pm 0.20}$ | $1.18_{\pm 0.12}$ | $4.84_{\pm 0.29}$ |
| | Rate (↑) | 15.18 % | 5.31 % | 21.80 % |
| CSI Emis. | MAE | $3.10_{\pm 0.17}$ | $0.85_{\pm 0.04}$ | $4.42_{\pm 0.36}$ |
| | Rate (↑) | 16.23 % | 5.23 % | 23.14 % |
| | RMSE | $4.99_{\pm 0.28}$ | $1.47_{\pm 0.12}$ | $7.29_{\pm 0.44}$ |
| | Rate (↑) | 18.34 % | 5.40 % | 26.80 % |
| CSI Life. | MAE | $0.085_{\pm 0.003}$ | $0.026_{\pm 0.002}$ | $0.072_{\pm 0.004}$ |
| | Rate (↑) | 13.70 % | 4.19 % | 11.61 % |
| | RMSE | $0.101_{\pm 0.010}$ | $0.034_{\pm 0.008}$ | $0.093_{\pm 0.010}$ |
| | Rate (↑) | 12.16 % | 4.09 % | 11.20 % |

**Obs.2:** The variability of MolTC's outcomes, as indicated by the standard deviation, is consistently minimal in comparison to other models, On average, the standard deviation for MolTC is 35.41% lower than GNN-based models and 46.86% lower than LLM-based models. The precision in MolTC's performance is largely attributed to the training paradigm enhanced by the multi-hierarchical CoT, which ensures a meticulous and accurate inference process.

### 3.3 Quantitative Prediction Performance (RQ2)

Table 2 shows the performance in a subset of quantitative tasks, with an exhaustive set of results detailed in the appendix. The datasets offer four-dimensional molecular information, comprising atom type, chirality tag, bond type, and bond direction. Key observations from Table 2 include:

**Obs.3:** MolTC continues to lead in quantitative analysis tasks, an area typically challenging for LLMs. Despite the strong baseline set by CGIB, characterized by low MAE and RMSE across datasets, MolTC outperforms it in every metric. For instance, it achieves a 23.98% reduction in RMSE on the CombiSolv dataset relative to CGIB. This underscores the advantage of adeptly leveraging the interaction between SMILE representations and molecular graph structures.

**Obs.4:** LLM-based models, in general, exhibit subpar performance in quantitative tasks compared to traditional DL-based models, attributed to their inadequacy in sharing and transferring learned molecular interaction insights across datasets and the absence of CoT-guided inference.

### 3.4 Ablation Study (RQ3)

Table 3 presents an ablation study aimed at dissecting the influence of SMILE auxiliary analysis and the optimized training paradigms based on Broad-grained and Fine-grained CoT. For the CSI dataset, properties such as the maximum absorption wavelength (Absorption), maximum emission wavelength (Emission), and excited state lifetime (Lifetime) are denoted as Abs., Emis., and Life., respectively. Key observations are as follows:

**Obs.5:** The three studied ablations exhibit significant influence on the results. For example, the collective impact of these three ablations registers an average drop of 12.77%, affirming the substantial enhancement imparted by the proposed strategies.

**Obs.6:** The most pronounced effect is observed with the ablation of the Fine-grained CoT paradigm, which incurs an average accuracy decrement of 18.82%. This underscores the pivotal role of guiding the LLM to deduce a numerical range, a strategy particularly beneficial for quantitative analysis tasks, typically a challenging domain for LLMs.

**Obs.7:** The least pronounced, yet significant, impact stems from the optimization of the Broad-grained CoT training paradigm, with an average accuracy reduction of 4.35%. Its importance is particularly underscored for molecular pairs involving larger and more complex molecules, where directly predicting interactive property by LLMs is arduous.

## 4 Discussion

**Potential Data Leakage Risk.** LLM might have memorized the prediction target during pretraining using a vast amount of text corpus. Hence, we acknowledge that LLM-based methods for biochemistry tasks such as MolT5 (Edwards et al., 2022), MolCA (Liu et al., 2023a) and our MolTC may face potential data leakage risks due to the memory during the pre-training process of their base LLMs. However, we claim that ablation experiments can mitigate the influence of data leakage, demonstrating the effectiveness of methods despite using potentially contaminated LLMs. For instance, in our

ablation experiments, the backbone LLMs are fed with SMILES strings and textual descriptions of molecules solely, and their performance is much worse than that of original methods. This stark contrast leads to two analytical observations:

- Firstly, note that these text inputs, highly recognizable and scattered throughout the LLM's training data, could provide all the necessary information to identify a molecule. Hence, if there is significant data leakage, the LLM could easily recall from memory and make reasonably accurate predictions based on these inputs. Based on this, the poor performance of these ablation experiments can, to some extent, counter the concern of data leakage.
- Secondly, since these ablation experiments and the original experiments utilized the exact same backbone LLM, the significant differences in their outcomes indicate that the original method's design is highly effective, regardless of whether the backbone LLM is subject to data leakage.

In summary, **while these ablation experiments can not completely eliminate the impact of data leakage, they do mitigate it.** Furthermore, it is worth noting that while the risk of data leakage exists, it does not deter researchers from continuously exploring and optimizing LLM-based frameworks for biochemical tasks. In the future, by constructing new biochemical datasets through biochemical experiments, which do not exist in the LLM's training data, researchers can eliminate the data leakage issue in these frameworks thoroughly.

## 5 Conclusion

This work focuses on molecule rationale learning, which plays a pivotal role in predicting molecular interactions. Specifically, we introduce a novel, unified LLM-based framework for predicting molecular interactive properties, termed MolTC. To efficiently train it, we propose a multi-tiered CoT principle to guide the training paradigm. Experiments conducted across twelve varied datasets demonstrate the superiority of our method over the current GNN and LLM-based baselines. This breakthrough sets a new standard for integrating multimodal data in LLM-based MRL.

## Limitations

While this research has undergone extensive testing across a diverse array of datasets covering various domains, it does have certain limitations. Specifically, the study has not been subjected to datasets comprising exceptionally large molecules, which represent extreme cases. Furthermore, the methodologies employed in this research have not yet been adapted or evaluated in contexts requiring few-shot or zero-shot learning scenarios. Future endeavors will focus on expanding the scope of this study to encompass these areas.

## Ethics Statement

This work is primarily foundational in molecular relational learning, focusing on the development of a unified LLM-based paradigm. Its primary aim is to contribute to the academic community by enhancing the understanding and implementation of the molecular relational modeling process. We do not foresee any direct, immediate, or negative societal impacts stemming from the outcomes of our research.

## Acknowledgement

This research is supported by the National Science and Technology Major Project (2023ZD0121102), National Natural Science Foundation of China (92270114).

## References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kexin Chen, Junyou Li, Kunyi Wang, Yuyang Du, Jiahui Yu, Jiamin Lu, Guangyong Chen, Lanqing Li, Jiezhong Qiu, Qun Fang, et al. 2023. Towards an automatic ai agent for reaction condition recommendation in chemical synthesis. *arXiv preprint arXiv:2311.10776*.

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*.

Yunsie Chung, Florence H Vermeire, Haoyang Wu, Pierre J Walker, Michael H Abraham, and William H Green. 2022. Group contribution and machine learning approaches to predict abraham solute parameters,

solvation free energy, and solvation enthalpy. *Journal of Chemical Information and Modeling*, 62(3):433–446.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

Andreea Deac, Yu-Hsiang Huang, Petar Veličković, Pietro Liò, and Jian Tang. 2019. Drug-drug adverse effect prediction with graph co-attention. *arXiv preprint arXiv:1905.00534*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*.

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*.

Junfeng Fang, Xinglin Li, Yongduo Sui, Yuan Gao, Guibin Zhang, Kun Wang, Xiang Wang, and Xiangnan He. 2024. Exgc: Bridging efficiency and explainability in graph condensation. In *WWW*. ACM.

Junfeng Fang, Wei Liu, Yuan Gao, Zemin Liu, An Zhang, Xiang Wang, and Xiangnan He. 2023a. Evaluating post-hoc explanations for graph neural networks via robustness analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Junfeng Fang, Xiang Wang, An Zhang, Zemin Liu, Xiangnan He, and Tat-Seng Chua. 2023b. Cooperative explanations of graph neural networks. In *WSDM*, pages 616–624. ACM.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023c. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023d. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, pages 1–12.

Yuan Gao, Junfeng Fang, Yongduo Sui, Yangyang Li, Xiang Wang, Huamin Feng, and Yongdong Zhang. 2024. Graph anomaly detection with bi-level optimization. In *WWW*, pages 4383–4394. ACM.

Yuan Gao, Xiang Wang, Xiangnan He, Huamin Feng, and Yong-Dong Zhang. 2023. Rumor detection with self-supervised learning on texts and social graph. *Frontiers Comput. Sci.*, 17(4):174611.

Laura M Grubbs, Mariam Saifullah, E Nohelli, Shulin Ye, Sai S Achi, William E Acree Jr, and Michael H Abraham. 2010. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid phase equilibria*, 298(1):48–53.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023a. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In *NeurIPS*.

Taicheng Guo, Changsheng Ma, Xiuying Chen, Bozhao Nan, Kehan Guo, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023b. Modeling non-uniform uncertainty in reaction prediction via boosting and dropout. *CoRR*, abs/2310.04674.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*.

Burrows C J, Harper J B, and et al. Sander W. 2022. Solvation effects in organic chemistry. *The Journal of Organic Chemistry*, 87(3):1599–1601.

Kanchan Jha, Sriparna Saha, and Hiteshi Singh. 2022a. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360.

Kanchan Jha, Sriparna Saha, and Hiteshi Singh. 2022b. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360.

Joonyoung F Joung, Minhi Han, Minseok Jeong, and Sungnam Park. 2020. Experimental database of optical properties of organic compounds. *Scientific data*, 7(1):295.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2023. Pubchem 2023 update. *Nucleic Acids Res.*, 51(D1):1373–1380.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*.

Namkyeong Lee, Dongmin Hyun, Gyoung S. Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. 2023a.

Conditional graph information bottleneck for molecular relational learning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 18852–18871. PMLR.

Namkyeong Lee, Kanghoon Yoon, Gyoung S Na, Sein Kim, and Chanyoung Park. 2023b. Shift-robust molecular relational learning with causal substructure. *arXiv preprint arXiv:2305.18451*.

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

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. 3d-molm: Towards 3d molecule-text interpretation in language models. In *ICLR*.

Tianhao Li, Sandesh Shetty, Advaith Kamath, Ajay Jaiswal, Xiaoqian Jiang, Ying Ding, and Yejin Kim. 2023b. Cancergpt: Few-shot drug pair synergy prediction using large pre-trained language models. *ArXiv*.

Zimeng Li, Shichao Zhu, Bin Shao, Xiangxiang Zeng, Tong Wang, and Tie-Yan Liu. 2023c. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1):bbac597.

Shenggeng Lin, Yanjing Wang, Lingfeng Zhang, Yanyi Chu, Yatong Liu, Yitian Fang, Mingming Jiang, Qiankun Wang, Bowen Zhao, Yi Xiong, et al. 2022. Mdf-sa-ddi: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Briefings in Bioinformatics*, 23(1):bbab421.

Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, volume 380, pages 2739–2745.

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

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024a. ReactXT: Understanding molecular reaction-ship via reaction-contextualized molecule-text pretraining. In *ACL (Findings)*. Association for Computational Linguistics.

Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Rethinking tokenizer and decoder in masked graph modeling for molecules. In *NeurIPS*.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024b. Prott3: Protein-to-text generation for text-based protein understanding. In *ACL*. Association for Computational Linguistics.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, and Alex Zhavoronkov. 2023. nach0: Multimodal natural and chemical languages foundation model. *arXiv preprint arXiv:2311.12410*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. In *NeurIPS*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Aleksandr V Marenich, Casey P Kelly, Jason D Thompson, Gregory D Hawkins, Candee C Chambers, David J Giesen, Paul Winget, Christopher J Cramer, and Donald G Truhlar. 2020. Minnesota solvation database (mnsol) version 2012.

Fanwang Meng, Hanwen Zhang, Juan Samuel Collins-Ramirez, and Paul W. Ayers. 2023. Something for nothing: Improved solvation free energy prediction with learning.

David L Mobley and J Peter Guthrie. 2014. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720.

Edouard Moine, Romain Privat, Baptiste Sirjean, and Jean-Noël Jaubert. 2017. Estimation of solvation quantities from experimental thermodynamic data: Development of the comprehensive compsol databank for pure and mixed solutes. *Journal of Physical and Chemical Reference Data*, 46(3).

Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. Ssi–ddi: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*, 22(6):bbab133.

Gilchan Park, Sean McCorkle, Carlos Soto, Ian Blaby, and Shinjae Yoo. 2022. Extracting protein-protein interactions (ppis) from biomedical literature using attention-based relational context information. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2052–2061. IEEE.

Yashaswi Pathak, Siddhartha Laghuvarapu, Sarvesh Mehta, and U Deva Priyakumar. 2020. Chemically interpretable graph interaction network for prediction of pharmacokinetic properties of drug-like molecules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 873–880.

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*.

C. Reichardt. 2021. Solvation effects in organic chemistry: A short historical overview. *The Journal of Organic Chemistry*, 87(3):1616–1629.

Dan M Roden, Robert A Harrington, Athena Poppas, and Andrea M Russo. 2020. Considerations for drug interactions on qtc in exploratory covid-19 treatment. *Circulation*, 141(24):e906–e907.

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. In *NeurIPS*.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the national academy of sciences*, 115(18):E4304–E4311.

Tatsuya Sagawa and Ryosuke Kojima. 2023. Reactiont5: a large-scale pre-trained model towards application of limited reaction data. *arXiv preprint arXiv:2311.06708*.

Yaorui Shi, An Zhang, Enzhi Zhang, Zhiyuan Liu, and Xiang Wang. 2023. Relm: Leveraging language models for enhanced chemical reaction prediction. In *EMNLP (Findings)*, pages 5506–5520. Association for Computational Linguistics.

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

Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. 2012. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Jithin John Varghese and Samir H Mushrif. 2019. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering*, 4(2):165–206.

Florence H Vermeire and William H Green. 2021. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307.

Hanchen Wang, Defu Lian, Ying Zhang, Lu Qin, and Xuemin Lin. 2020. Gognn: Graph of graphs neural network for predicting structured entity interactions. *arXiv preprint arXiv:2005.05537*.

Kun Wang, Guohao Li, Shilong Wang, Guibin Zhang, Kai Wang, Yang You, Xiaojiang Peng, Yuxuan Liang, and Yang Wang. 2023a. The snowflake hypothesis: Training deep gnn with one node one receptive field. *arXiv preprint arXiv:2308.10051*.

Kun Wang, Yuxuan Liang, Xinglin Li, Guohao Li, Bernard Ghanem, Roger Zimmermann, Huahui Yi, Yudong Zhang, Yang Wang, et al. 2023b. Brave the wind and the waves: Discovering robust and generalizable graph lottery tickets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kun Wang, Yuxuan Liang, Pengkun Wang, Xu Wang, Pengfei Gu, Junfeng Fang, and Yang Wang. 2023c. Searching lottery tickets in graph neural networks: A dual perspective. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. 2023. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487*.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, pages 9240–9251.

Jiahui Yu, Chengwei Zhang, Yingying Cheng, Yun-Fang Yang, Yuan-Bin She, Fengfan Liu, Weike Su, and An Su. 2023a. Solvbert for solvation free energy and solubility prediction: a demonstration of an nlp model for predicting the properties of molecular complexes. *Digital Discovery*, 2(2):409–421.

Xingtong Yu, Zemin Liu, Yuan Fang, and Xinming Zhang. 2023b. Learning to count isomorphisms with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xingtong Yu, Chang Zhou, Yuan Fang, and Xinming Zhang. 2024. Text-free multi-domain graph pre-training: Toward graph foundation models. *arXiv preprint arXiv:2405.13934*.

Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. 2017. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics*, 18:1–12.

Xu Zheng, Farhad Shirani, Tianchun Wang, Wei Cheng, Zhuomin Chen, Haifeng Chen, Hua Wei, and Dongsheng Luo. 2023. Towards robust fidelity for evaluating explainability of graph neural networks. *CoRR*, abs/2310.01820.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: a universal 3d molecular representation learning framework.

M Zitnik, R Sosi, S Maheshwari, and J Leskovec. 2018. Stanford biomedical network dataset collection. *Biosn. Datasets Stanford Biomed. Netw. Dataset Collect*.

## A Related Work

Since exhaustive experimental validation of the molecule interactions is notoriously time-consuming and costly (Lee et al., 2023a), more recently, adopting LLM has emerged as a promising alternative for efficient and effective molecular relational learning, which are known for their vast knowledge repositories and advanced logical inference capabilities. Compared to the prediction of single-molecule properties (Liu et al., 2024b; Li et al., 2024; Liu et al., 2023b; Fang et al., 2023d; Guo et al., 2023a), the field of molecular interaction prediction is still in its early stages. For instance,

- **Protein-protein interactions (PPI)**. In this context, proteins are represented as residue contact graphs, also known as amino acid graphs, where each node is a residue (Park et al., 2022; Jha et al., 2022a,b,b). Notably, Jha et al. (2022b) leverages the superior encoding capabilities of the biochemical LLMs, where the input to the LLM is the protein sequence, and the output is a feature vector for each amino acid in the sequence. This output is then used as node features in the residue contact graph to enhance the prediction of PPI tasks.

- **DDI.** In this context, Pei et al. (2023) focuses on enriching cross-modal integration in biology with chemical knowledge and natural language associations, achieving significant results in multiple drug-target interaction prediction tasks. Recently, few-shot drug pair synergy prediction is gradually gaining attention and entering the spotlight(Li et al., 2023b).

- **Chemical reactions**. For understanding chemical reactions (Chen et al., 2023; Guo et al., 2023b), Shi et al. (2023) selects in-context reaction examples with varying confidence scores closest to the target reaction query, encouraging large models to understand the relationships between these reactions. Sagawa and Kojima (2023) focuses on optimizing low-sample organic chemical applications by pretraining them with extensive compound libraries and fine-tuning with smaller in-house datasets for specific tasks. Livne et al. (2023) introduces a new foundational model, nach0, capable of solving various chemical and biological tasks, including molecular synthesis.

More recently, ReactXT (Liu et al., 2024a) also focuses on modeling molecular interactions using chemical reactions. However, it is unable to leverage CoT for more robust information extraction. Meanwhile, the Mol-instructions dataset has been introduced (Fang et al., 2023c) for effectively training biochemistry LLMs, that can be considered as complementary to our MoT-instruction.

## B Experiments

Here, we provide a detailed experimental setup along with additional results. It is important to note that for aspects such as dataset division and hyperparameter configurations in baselines, we followed the settings established by CGIB (Lee et al., 2023a). Moreover, all settings can be found in our code https://github.com/MangoKiller/MolTC.

### B.1 Datasets

We employ 12 datasets across various domains such as DDI, SSI and CSI.

**Drugbank (version 5.0.3).** This dataset consists of 1704 drugs, 191400 drug pairs, and defines 86 distinct DDI event types. Essential drug information, including DrugBank ID, drug name, molecular SMILES, and target. provided.

**ZhangDDI** (Zhang et al., 2017). It contains 548 drugs and 48,548 pairwise interaction data and multiple types of similarity information about these drug pairs.

**ChChMiner** (Zitnik et al., 2018). It contains 1,322 drugs and 48,514 labeled DDIs, obtained through drug labels andscientific publications.

**DeepDDI** (Ryu et al., 2018). It contains 192,284 labeled DDIs and their detailed side-effect information, which is extracted from Drugbank.

**TWOSIDES** (Tatonetti et al., 2012). It collects 555 drugs and their 3,576,513 pairwise interactions involving 1318 interaction types from TWOSIDES.

**Chromophore** (Joung et al., 2020). It contains 20,236 combinations of 7,016 chromophores and 365 solvents which are given in the SMILES string format. All optical properties are based on scientific publications and unreliable experimental

results are excluded after examination of absorption and emission spectra. In this dataset, we measure our model performance on predicting maximum absorption wavelength (Absorption), maximum emission wavelength (Emission) and excited state lifetime (Lifetime) properties which are important parameters for the design of chromophores for specific applications. We delete the NaN values to create each dataset which is not reported in the original scientific publications. Moreover, for Lifetime data, we use log normalized target value since the target value of the dataset is highly skewed inducing training instability.

**MNSol** (Marenich et al., 2020). It contains 3,037 experimental free energies of solvation or transfer energies of 790 unique solutes and 92 solvents.In this work, we consider 2,275 combinations of 372 unique solutes and 86 solvents following previous work.

**FreeSolv** (Mobley and Guthrie, 2014). It provides 643 experimental and calculated hydration free energy of small molecules in water. In this work, we consider 560 experimental results following previous work.

**CompSol** (Moine et al., 2017). This dataset is proposed to show how solvation energies are influenced by hydrogen-bonding association effects. We consider 3,548 combinations of 442 unique solutes and 259 solvents in the dataset following previous work.

**Abraham** (Grubbs et al., 2010). This dataset is a collection of data published by the Abraham research group at College London. We consider 6,091 combinations of 1,038 unique solutes and 122 solvents following previous work.

**CombiSolv** (Vermeire and Green, 2021). It contains all the data of MNSol, FreeSolv, CompSol, and Abraham, resulting in 10,145 combinations of 1,368 solutes and 291 solvents.

**CombiSolv-QM** (Vermeire and Green, 2021). It is generated with 1 million combinations of 284 commonly used solvents and 11,029 solutes. Those 1 million data points are randomly selected from all possible solvent–solute combinations. Solvents and solutes with elements $H, B, C, N, O, F, P, S, Cl, Br$ and $I$ are included with a solute molar mass ranging from 2.02 g/mol to 1776.89 g/mol.

### B.2 Baselines

We use both specific task conventional deep learning models and current biochemical LLMs as the baselines. Specifically, for qualitative tasks:

**GoGNN** (Wang et al., 2020). It extracts features from structured entity graphs and entity interaction graphs in a hierarchical manner. We also propose a dual attention mechanism that enables the model to preserve the importance of neighbors in both levels of the graph.

**MHCADDI** (Deac et al., 2019). A gated information transfer neural network is used to control the extraction of substructures and then interact based on an attention mechanism.

**DeepDDI** (Ryu et al., 2018). First, the structural similarity profile is calculated between the two input drugs and other drugs, and then prediction is completed based on the deep neural network.

**SSI-DDI** (Nyamabo et al., 2021). It uses a 4-layer GAT network to extract substructures at different levels, and completes the final prediction based on the co-attention mechanism.

**CGIB** (Lee et al., 2023a). Based on the graph conditional information bottleneck theory, conditional subgraphs are extracted to complete the interaction between molecules.

**CMRL** (Lee et al., 2023b). It detects the core substructure that is causally related to chemical reactions by introducing a novel conditional intervention framework whose intervention is conditioned on the paired molecule.

**MDF-SA-DDI** (Lin et al., 2022). It predicts interaction based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism.

**DSN-DDI** (Li et al., 2023c). It employs local and global representation learning modules iteratively and learns drug substructures from the single drug 'intra-view') and the drug pair ('inter-view') simultaneously.

For quantitative task, we employ the following baselines:

**D-MPNN** (Vermeire and Green, 2021). It employs a transfer learning approach to predict solvation free energies, integrating quantum calculation fundamentals with the heightened accuracy of experimental measurements through two new databases, CombiSolv-QM and CombiSolv-Exp.

**SolvBert** (Yu et al., 2023a). It interprets solute and solvent interactions through their combined

Table 4: Comparative performance of various methods in qualitative and quantitative interactive tasks. The best-performing methods are highlighted with a gray background, while the second-best methods are underscored for emphasis.

| Domains | Datasets | Metrics | Baselines | | | | Ours |
|---|---|---|---|---|---|---|---|
| | | | Galactica | Chem T5 | MolCA | MolT5 | MolTC |
| DDI | TWOSIDES | ACC | $82.01 \pm 1.76$ | $84.43 \pm 2.58$ | $90.07 \pm 1.86$ | $\underline{92.73} \pm 1.65$ | $98.42 \pm 0.72$ |
| | | AUCROC | $87.99 \pm 2.41$ | $89.52 \pm 1.64$ | $93.68 \pm 0.83$ | $\underline{94.00} \pm 0.61$ | $99.02 \pm 0.14$ |
| SSI | MNSol | MAE | $0.584 \pm 0.095$ | $0.504 \pm 0.038$ | $0.491 \pm 0.053$ | $\underline{0.449} \pm 0.081$ | $0.324 \pm 0.019$ |
| | | RMSE | $1.002 \pm 0.101$ | $0.973 \pm 0.079$ | $0.930 \pm 0.062$ | $\underline{0.858} \pm 0.069$ | $0.585 \pm 0.023$ |
| CSI | Absorption | RMSE | $43.16 \pm 1.38$ | $38.70 \pm 1.84$ | $\underline{36.53} \pm 2.03$ | $38.01 \pm 2.27$ | $28.28 \pm 2.20$ |
| | Emission | RMSE | $49.85 \pm 2.47$ | $46.18 \pm 2.28$ | $\underline{43.35} \pm 1.94$ | $46.06 \pm 1.65$ | $35.43 \pm 1.88$ |
| | Lifetime | RMSE | $1.951 \pm 0.115$ | $1.633 \pm 0.069$ | $1.480 \pm 0.092$ | $\underline{1.394} \pm 0.145$ | $1.198 \pm 0.073$ |

Table 5: Comparative performance of various methods in CombiSolv-QM. The best-performing methods are highlighted with a gray background, while the second-best methods are underscored for emphasis.

| Baseline Model | | CombiSolv-QM | |
|---|---|---|---|
| | | MAE | RMSE |
| GNN Based | CIGIN | $0.077_{\pm 0.002}$ | $0.176_{\pm 0.004}$ |
| | D-MPNN | $0.116_{\pm 0.006}$ | $0.208_{\pm 0.005}$ |
| | GEM | $0.079_{\pm 0.003}$ | $0.162_{\pm 0.002}$ |
| | CGIB | $\underline{0.074}_{\pm 0.004}$ | $\underline{0.150}_{\pm 0.005}$ |
| ML Based | GOVER | $0.094_{\pm 0.003}$ | $0.277_{\pm 0.005}$ |
| | SolvBert | $0.102_{\pm 0.005}$ | $0.318_{\pm 0.006}$ |
| | Uni-Mol | $0.089_{\pm 0.006}$ | $0.214_{\pm 0.005}$ |
| | SMD | $0.107_{\pm 0.004}$ | $0.341_{\pm 0.003}$ |
| LLM Based | Galactica | $0.303_{\pm 0.004}$ | $0.601_{\pm 0.008}$ |
| | Chem T5 | $0.321_{\pm 0.006}$ | $0.555_{\pm 0.008}$ |
| | MolCA | $0.298_{\pm 0.004}$ | $0.545_{\pm 0.007}$ |
| | MolT5 | $0.214_{\pm 0.004}$ | $0.339_{\pm 0.009}$ |
| MolTC (Ours) | | $0.072_{\pm 0.002}$ | $0.140_{\pm 0.003}$ |

SMILES representation. After pre-training by unsupervised learning with a substantial computational solvation free energy database, SolvBERT is adaptable to predict experimental solvation free energy or solubility by fine-tuning on specific databases.

**SMD** (Meng et al., 2023). It utilizes the quantum charge density of a solute and a continuum representation of the solvent. It breaks down solvation free energy into two components: bulk electrostatic contribution, treated through a self-consistent reaction field using IEF-PCM, and a cavity-dispersion-solvent-structure term, accounting for short-range interactions in the solvation shell based on atomic surface areas with geometry-dependent constants.

**CIGIN** (Pathak et al., 2020). It adopts an end-to-end framework consisting of three essential phases: message passing, interaction, and prediction. In the final phase, these stages are leveraged to predict solvation free energies.

**GEM** (Fang et al., 2022). It exhibits a uniquely designed geometry-based graph neural network architecture, complemented by several dedicated self-supervised learning strategies at the geometry level. That aims to acquire comprehensive molecular geometry knowledge for accurate prediction of molecular properties.

**GOVER** (Rong et al., 2020). It captures rich structural information from extensive unlabeled molecular data through self-supervised tasks, employing a flexible Transformer-style architecture integrated with Message Passing Networks. This allows GROVER to be trained efficiently on large-scale datasets without supervision, addressing data scarcity and bias challenges.

**Uni-Mol** (Zhou et al., 2023). It incorporates two pre-trained models featuring the SE(3) Transformer architecture: a molecular model pre-trained on 209 million molecular conformations and a pocket model pre-trained on 3 million candidate protein pocket data. Additionally, Uni-Mol integrates various fine-tuning strategies to effectively apply these pre-trained models across diverse downstream tasks.

### B.3 Modules

In our experiments, the two graph encoder are instantiated by the five-layer GINE (Hu et al., 2019). We conduct 2 million molecules from the ZINC15 (Sterling and Irwin, 2015) dataset to pretrain them by contrastive learning following (Liu et al., 2023a). Similarly, two projector are initialized with the encoder-only transformer, Sci-BERT, which is pre-trained on scientific publications (Beltagy et al., 2019), while its cross-attention modules are randomly initialized. More detailed pretraining process of our Q-Formors follows the training process

in (Liu et al., 2023a), such as there are 8 query tokens in Q-Formers ($N_q = 8$). Note that for LLM-based baselines, we fine-tune the backbone LLMs on task-specific datasets for fair comparison. Their prediction is considered accurate only if the outputs include words or numbers that correctly depict the interaction in question, without presenting any that describe alternative interactions.

### B.4 Training Epochs

During the fine-tuning phase, the number of epochs varies for different tasks. For example, for the DDI task, we typically fine-tune for 100 epochs. For SSI datasets with more than 3000 molecular pairs, we initially fine-tune on the CombiSolv-QM (Vermeire and Green, 2021) dataset for 100 epochs, followed by an additional 30 epochs on their respective datasets. For SSI datasets with fewer than 3000 molecular pairs, this number is adjusted to 20. Furthermore, both the fine-tuning and pre-training phases employ the same configuration for the optimizer and learning rate scheduler, as detailed in the following section.

### B.5 Training Strategy

We employ the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay set at 0.05. Our learning rate strategy utilizes a combination of linear warm-up and cosine decay, optimizing the training process by initially increasing the learning rate to promote faster convergence, and then gradually decreasing it according to a cosine curve to fine-tune the model parameters. LoRA is implemented using the Open Delta library (Ding et al., 2022), and the PEFT library (Mangrulkar et al., 2022). LoRA's rank $r$ is set to 16, while LoRA is applied to Galactica's modules of `[q_proj, v_proj, out_proj, fc1, fc2]` following (Liu et al., 2023a). This configuration yields a LoRA adapter with 12M parameters which constitutes merely 0.94% of the parametersin the Galactica$_{1.3B}$.

### B.6 More Experimental Results

Table 4 presents the experimental results not shown in the main text due to length constraints. Note that the three datasets in the CSI domain are all derived by splitting the Chromophore dataset. As discussed in Section 3.3, for a fair comparison, we limited the input features to four-dimensional molecular information, comprising atom type, chirality tag, bond type, and bond direction. Given the difficulty of convergence for some DL-based baselines under this setting, we only showcased the performance of the LLM-based baselines. Meanwhile, considering that our SSI tasks are firstly fine-tuned on the CombiSolv-QM dataset, we present the comprehensive results of this dataset, as shown in Table 5. Observations from Table 4 and 5 are largely consistent with those in the main experimental section. That is, across all tasks, our MolTC outperforms the LLM-based baseline methods in a large margin.

## C Future Work

In this paper, we introduce a novel unified framework, leveraging LLM technology to predict molecular interactive properties. The future development directions of this project are twofold. First, we aim to adopt advanced graph encoding techniques to enhance the comprehension of molecular graphs (Gao et al., 2023; Wang et al., 2023a; Yu et al., 2023b, 2024; Gao et al., 2024). Secondly, in light of the extensive input information introduced in multi-molecular input scenarios, we plan to employ techniques such as graph interpretability (Ying et al., 2019; Luo et al., 2020; Zheng et al., 2023; Fang et al., 2024) and sparsification (Wang et al., 2023b,c) to eliminate information redundancy and resist out-of-distribution (Fang et al., 2023a,b).