Bridging Text and Molecule: A Survey on Language-molecule Models

Anonymous ACL submission

Abstract

Artificial intelligence has demonstrated immense potential in scientific research. Within molecular science, it is revolutionizing the traditional computer-aided paradigm, ushering in a new era of deep learning. With recent progress in multimodal learning and natural language processing, an emerging trend has targeted at building multimodal frameworks to jointly model molecules with textual domain knowledge, known as language-molecule models. In this paper, we present the first systematic survey on language-molecule models. Specifically, we begin with the development of molecular deep learning and point out the necessity to involve textual modality. Next, we focus on recent advances in text-molecule alignment methods, categorizing current models based on their architectures and listing relevant pre-training tasks. Furthermore, we delves into the utilization of large language models and prompting techniques for molecular tasks and present significant applications in drug discovery. Finally, we discuss the limitations in this field and highlight several promising directions for future research.

1 Introduction

Accurately modeling molecules and extracting meaningful features is a primary goal of molecular deep learning. Initially, manual descriptors, such as molecular fingerprints and SMILES, are proposed to describe molecules in strings or sequences. These descriptors can naturally be encoded by language models for feature extraction. Subsequently, graph structures gradually show their superiority in modeling the topology structure within molecules. Graph neural networks (GNNs) are used to learn from molecular graphs by aggregating and propagating information within atoms and chemical bonds (Kipf and Welling, 2017). Simultaneously, numerous works integrate self-supervised pre-training in this process to generate generalized representations. Despite the success in molecular deep learning, two key challenges persistently exist. First, owing to the complexity of chemical space and chemical rules, current deep learning frameworks lack a deep comprehension of chemical domain knowledge (e.g. quantum mechanics rules). Furthermore, both supervised and self-supervised models need to be trained or fine-tuned on labeled molecules, which are typically scarce in real applications due to the high experimental cost. These notorious problems decelerate progress in related areas.

Recently, multimodal learning and Large Language Models (LLMs) have shown impressive competence in modeling and inference. Inspired by the success of vision-language models, it is natural to associate molecules with text description to build language-molecule models (Edwards et al., 2024). Following this idea, a line of works treats molecules as languages with special grammar, and cross-language frameworks, such as T5 (Raffel et al., 2020), are chosen as the backbone to jointly model text and molecules (Edwards et al., 2022; Taylor et al., 2022; Pei et al., 2023, 2024b,a; Jin et al., 2024). At the same time, another line of work explores the alignment of the latent space between text and structured molecular data (Su et al., 2022; Liu et al., 2023a; Xiao et al., 2024a; Huo et al., 2024; Flöge et al., 2024; Su et al., 2024; Liu et al., 2024e), and attempts to integrate LLMs into multimodal frameworks as predictors for cross-modal molecular tasks. Furthermore, prompting techniques are also introduced in the fine-tuning process and yield competitive results in many molecular tasks without large-scale pre-training (Liang et al., 2023; Cao et al., 2023; Zhang et al., 2023; Yu et al., 2024; Jin et al., 2024; Gruver et al., 2024). Recently, some insightful work has attempted to build autonomous agents for chemistry and biology (Boiko et al., 2023; Liu et al., 2024d), bringing a new paradigm for future scientific research.

However, as a prosperous subject, there still lacks a systematic review to summarize recent progress and propose promising outlooks. In this regard, we present the first survey of language-molecule models. We summarize our contributions as follows: (1) We provide an overview of this field with a structured taxonomy that categorizes the framework based on their basic architecture. (2) Our systematic review provides a detailed analysis of training strategies, dataset construction methods and corresponding applications. (3) We analyze the limitations in this field and provide several promising research directions.

2 Molecular Descriptors and Encoding

Molecules need to be transformed into descriptors for the recognition of the model. In this section, we briefly summarize the mainstream descriptors of small molecules and proteins along with their corresponding encoder architectures. Generally, both small molecules and proteins can be described by sequences and graphs.

2.1 1D Molecule Sequence

Small-molecule Sequence Molecules are composed of atoms and connected bonds, allowing the representation of molecules as sequences that describe their components. The Simplified Molecular Input Line Entry System (SMILES) is the most commonly used sequential descriptor, mapping atoms, bonds, and special structures using ASCII symbols. Self-referencing embedded strings (SELFIES) (Krenn et al., 2020) is another string-based descriptor which is recently popular for its robustness and superiority in tokenization. International Union of Pure and Applied Chemistry (IUPAC) is the official name of molecules in the human language, which can serve as a connector for language models to understand chemical expressions. Molecular Fingerprints (Axen et al., 2017; Rogers and Hahn, 2010) are class of binary codes with each position representing a predefined chemical structure. Because of their simplicity and capability to encode structure information, molecular fingerprints have been widely used in chemoinformatics research.

Protein Sequence A protein can be viewed as a combination of 20 types of amino acids, which allows it to be expressed as amino acid sequences in a manner similar to molecules. The amino acid sequence captures the co-evolutionary information and plays a vital role in protein folding and function. Usually, protein sequences are encoded by Protein Language Models (PLMs) and represented as PLM tokens for further processing.

2.2 2D Molecule Structures

2D Graph The topology structure of molecules can be naturally modeled by graph, with atoms as nodes and bonds as edges. The chemical and physical properties of atoms and bonds can also be featurized by molecular graphs. GNNs (Kipf and Welling, 2017) can be used to learn local and global representations of molecules and have shown competitive results in various downstream tasks (Liu et al., 2022).

2.3 3D Molecule Structures

3D Geometric Graph 2D molecular graphs have limitations in capturing spatial information within molecules. For example, chiral molecules cannot be distinguished through most of the 2D graph. The geometry information of the conformers (e.g., torsional angles and bond length) is in direct relation to molecular properties. In 3D geometry, atoms are associated with their coordinates with features expressed in high-order tensors to ensure geometric symmetries and expressiveness. Many studies concentrate on designing equivariant GNNs to accurately model the interaction between atoms (Batzner et al., 2022).

Protein Graph Protein functions are mainly determined by their folded structures (Jumper et al., 2021). To better capture structural information, proteins can be represented as a residue-level relation graph, where nodes are residues with positions of C_α and edges encoding their connectivity or relative distance. GVP (Jing et al., 2020) or EGNN (Satorras et al., 2021) are popular GNNs for protein structure encoding.

3 Latent Space Alignment between Text and Molecule

The encoding stage featurizes text and molecules into a single modality, while these representations still inhabit diverse semantic spaces and cannot interact with each other. To facilitate downstream tasks, different architectures are designed for text-molecule fusion and latent space alignment. In this section, we classify model architectures by the fusion scheme and summarize the corresponding

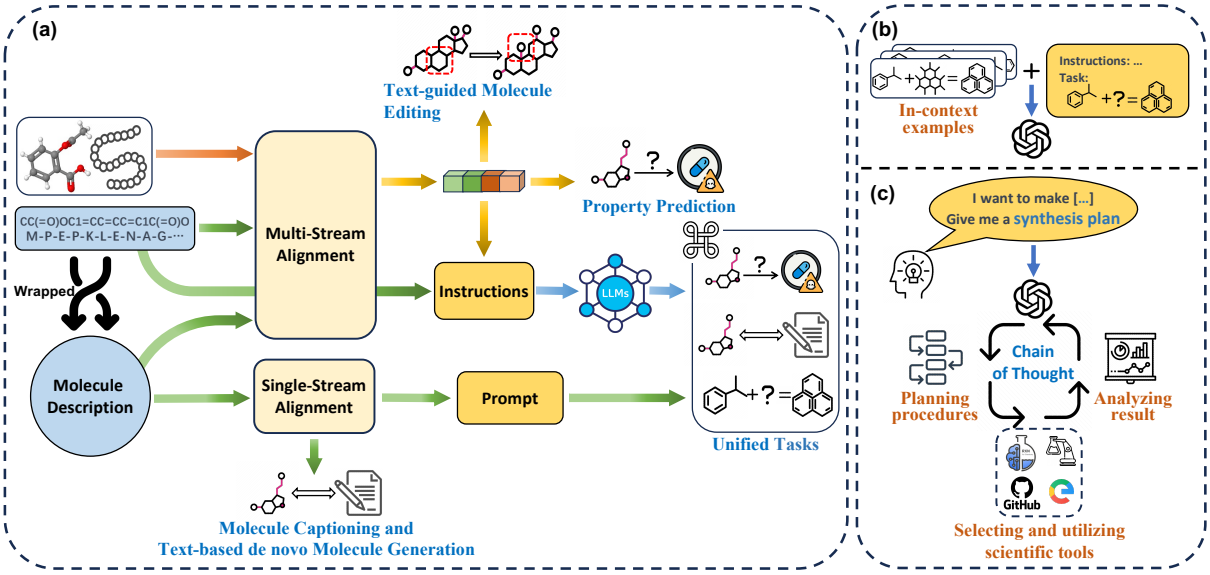

Figure 1: Pipeline of language-molecule models and downstream molecular tasks (a-c). (a) Latent space alignment and adaptation of downstream tasks. The single-stream framework jointly models text and molecules with the same encoder. The downstream tasks are realized with task-specific prompts described in section 4.1; The multi-stream framework involves cross-modal alignment between text and molecules. Features from latent space can be directly used for tasks or be used in instruction-tuning. (b) Building a semi-autonomous agent for molecular research with instructions and in-context examples. (c) Building autonomous agent for chemistry with instructions and chain-of-thought prompting. Equipping agent with external tools and memory largely expand the autonomous level and capabilities.

pre-training tasks. We present a summary of representative works in Table 1.

3.1 Model Architecture

Drawing inspiration from previous works in vision-language pre-training (Du et al., 2022), we categorize models into *single-stream*, and *multi-stream* architecture. The two types of models differ mainly in their understanding of molecular latent space.

Single-Stream Architecture A single-stream architecture assumes that the latent space of molecules and text shares similar semantic meaning. In this circumstance, molecules are treated as a specialized language and expressed by sequential descriptors. Different tokenization strategies are adopted to encode molecules and text, and these tokens will be fed into a language model, such as T5 (Raffel et al., 2020), for multi-language pre-training. As a widely used tokenization method in LLMs, byte-pair encoding (BPE) (Gage, 1994) can also be used to encode molecule sequences (Zeng et al., 2022; Liu et al., 2023b). BioT5 (Pei et al., 2023) optimize this strategy by using separate vocabularies for molecules, proteins, and texts to avoid misunderstanding of tokens that may have the same expression but originate from different semantic spaces. Gruver et al. (2024); Pei et al. (2024a) adopt same numerical tokenization for LLaMA-2 models (Touvron et al., 2023) to improve model performance on arithmetic tasks (Liu and Low, 2023).

Multi-Stream Architecture Models with a multi-stream architecture utilize intra-modality encoding for both text and molecular data. To align multimodal embeddings, approaches such as projection layers (Liang et al., 2023; Cao et al., 2023; Wang et al., 2024a), or pre-training tasks (Tang et al., 2024; Liu et al., 2023a; Flöge et al., 2024; Zhang et al., 2024b) are employed. Another method involves fusing the embeddings into a unified latent space, facilitating integrated representation between modalities(Xu et al., 2023; Liu et al., 2024a; Nguyen et al., 2024; Luo et al., 2024b, 2023a).

A representative architecture for cross modal alignment is Q-Former (Li et al., 2023), which has been widely used in vision-language models. Similarly, Li et al. (2024b); Liu et al. (2023c); Zhang et al. (2023); Luo et al. (2024c) adopt Q-Former to align molecular graph with text embeddings. While Liu et al. (2024e); Wang et al. (2024a) adopt the Q-Former architecture to align text and PLM tokens. Zhang et al. (2024a) introduce causal masks into the Q-Former queries, ensuring that the queries possess the same causal dependency as the text

sequences.

3.2 Pre-training Tasks

The fused representations need to be aligned in a unified latent space to maintain consistent semantic meaning for downstream tasks. In this section, we review the commonly used pre-training tasks for alignment between text and molecules.

Molecule-Text Contrastive Learning The contrastive learning (CL) task between molecules and text aims to align multimodal representations by enhancing the correlation between matched molecule-text pairs. The contrastive learning objective pushes the embeddings of matched text and molecules closer in latent space while enlarging the distance between pairs from different molecules. The CL task will enhance the model with cross-modal retrieval and matching ability. Here, we present the expression of commonly used InfoNCE (van den Oord et al., 2019) loss:

$$\mathcal{L}_{\text{NCE}} = -\sum_{i} \log \frac{\exp(z_i^M \cdot z_i^T/\tau)}{\sum_{j=1}^{N} \exp(z_i^M \cdot z_j^T/\tau)} \quad (1)$$

where τ is the temperature coefficient. In order to facilitate convergence, a trainable linear projector can be used to minimize the modality gap before the contrastive learning (Liu et al., 2023a).

Although contrastive learning is an effective approach for cross-modal molecule-text alignment, the limited number of molecule-text pairs brings negative impacts on the alignment result. Motivated by molecular graph augmentation methods (You et al., 2020), MoMu (Su et al., 2022) introduces two augmented graphs with node drop and sub-graph extraction to extend the number of matched pairs. MolLM (Luo et al., 2024a) introduces two additional augmentations, which are chemical transformation and motif removal, making the alignment process more robust.

Molecule-Text Matching Molecule-text matching (MTM) aims to predict whether a molecule-text pair is matched or not. It is defined as a binary classification task with the following loss function:

$$\mathcal{L}_{\text{MTM}} = \mathcal{L}_{\text{match}} - \mathcal{L}_{\text{unmatch}} \quad (2)$$

where $\mathcal{L}_{\text{match}}$ denotes the cross-entropy loss of matched molecule and text pair (m_i, t_i) and $\mathcal{L}_{\text{unmatch}}$ denotes the loss of unmatched pairs (m_i, t_j) and (m_j, t_i). The MTM task enables the model to have retrieval ability and refines the alignment between text and molecule, usually used in the pre-training stage of Q-former architecture.

Conditional Generation Conditional generation (CG) aims to generate tokens based on given conditions or constraints. Tasks such as molecule captioning and text-based molecule generation all fall into this category. Conditional generation enables models to learn complex mapping rules between text and molecules. It is adaptable for the T5 architecture, where all molecular tasks are transformed into a text-to-text generation format. The objective function can be written as:

$$\mathcal{L}_{\text{CG}} = -\sum_{i}^{n_i} \log P(u_i|C;\theta) \quad (3)$$

where u_i is the i-th token and C denotes the generation condition which may be referred to as a molecule graph or text description depending on the task.

Masked Language Modeling As discussed in Section 3, modeling languages and molecules may share similarities. Under this assumption, masked language modeling as a popular pre-training task for LLMs can also be used for training molecule sequences or wrapped sequences. During the pre-training stage, the models are trained to predict the masked components using the remaining context. The training objective is defined by cross-entropy

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{T\in\mathcal{D}} \sum_{\tilde{m}\in\mathcal{M}} \log p(\tilde{m}|T\backslash\mathcal{M}) \quad (4)$$

where $\mathcal{M}$, $T\backslash\mathcal{M}$, T represent the masked tokens, unmasked tokens, tokenized text and molecules separately. This self-supervised pre-training task can enhance the contextual comprehension of the model, improving performance in many downstream tasks. For MLM, there are two types of masking: **token masking** represented by BERT (Devlin, 2018) and its variants, and **span masking** introduced in T5 (Raffel et al., 2020), which has been shown to be more efficient. Edwards et al. (2022); Pei et al. (2023); Rubungo et al. (2023); Qian et al. (2023) adopt span masking to enhance downstream translation tasks between molecules and text. Xu et al. (2023) introduce MLM to recover fused residue tokens, enhancing the fine-grained connection between descriptions and corresponding residues.

Casual Language Modeling Different from the autoencoder (AE) language models such as BERT and T5, the autoregressive models represented by GPT (Yenduri et al., 2024) are trained with Casual Language Modeling (CLM). The objective of CLM is to predict the next token in a sequence in a left-to-right direction. The objective function can be written as

$$\mathcal{L}_{\text{CLM}} = -\sum_{i}^{n_i} \log P(u_i | u_{i-k}, ..., u_{i-1}; \theta) \quad (5)$$

where n_i and k represent the number of tokens and context length. CLM can seamlessly bridge the pre-training and instruction-tuning stage (Liang et al., 2023; Cao et al., 2023; Zhang et al., 2023). We will discuss the details of instruction-tuning and adaptation of tasks in the following section.

4 Bridging LLMs and Molecular Tasks with Prompting Techniques

With the advancement of multimodal large language models (MLLMs), the cross-modal inference ability of LLMs could be extended to biological research. Compared with traditional cross-modal learning that focuses on modality alignment, MLLMs leverage powerful LLM to process multi-modal information and utilize prompting techniques such as instruction-tuning (IT), in-context learning (ICL) and chain-of-thought (CoT) to realize downstream tasks (Li et al., 2023). As shown in Figure 1, LLMs could conduct multiple molecular tasks with instructions and cross-modal input. In this section, we discuss the prompting techniques in cross-modal molecular research and show the application in building intelligent agents for chemistry.

4.1 Prompt-based Fine-tuning

To bridge the gap between pre-training and downstream tasks, Raffel et al. (2020) transfer all downstream tasks into text-to-text generation format with task-specific prefix. Based on this work, Gao et al. (2021) propose prompt-based fine-tuning that unifies different tasks with task-specific prompts. This strategy can also be applied to cross-modal molecular tasks. For example, the prompt for the property prediction task in MoleculeNet (Wu et al., 2018) can be designed as: "*We can conclude that the property of <SMILES> is <tag>*" where *<tag>* is the predicted "true" or "false" label (Liu et al., 2023b). In this way, we unify all tasks into a text generation format and models are fine-tuned and evaluated with fixed pre-training parameters. Pei et al. (2023) enrich the above-mentioned template with detailed task explanations, which improves the accuracy of property prediction. Liu et al. (2023c) integrate fused feature as a soft prompt and use LoRA (Hu et al., 2022) to improve the efficiency of adaptation. Compared with traditional fine-tuning, prompt-based fine-tuning shows impressive performance in few-shot datasets.

4.2 Instruction Tuning on LLM for Zero-shot Learning Ability

Unlike prompt-based tuning, instruction-tuning (Wei et al., 2022) aims to adapt the model to various tasks. In the tuning process, models are trained in multiple tasks that have been unified through task-specific instructions. This multi-task learning strategy enables models to comprehend instructions and seamlessly adapt to few-shot or zero-shot tasks (Zhao et al., 2023a). A standard instruction entry is typically composed of three main parts: an *<instruction>* that clarifies the task, an *<input>* which is usually the molecular feature, and an *<output>* that embodies the expected outcome (Fang et al., 2024). Liang et al. (2023); Luo et al. (2023b); Cao et al. (2023); Li et al. (2024b); Zhang et al. (2023) use fused feature as a soft prompt to enrich the instructions. During the tuning process, the fusion architecture is fine-tuned solely and LoRA (Hu et al., 2022) can be used to improve efficiency (Li et al., 2024b; Cao et al., 2023).

4.3 In-Context Learning and Chain-of-Thought

Recently, various attempts have been made to integrate LLMs into scientific research as intelligent agents, with applications in autonomous experiment planning (M. Bran et al., 2024; Boiko et al., 2023), conversational drug editing (Liu et al., 2024d), chemical reaction prediction (Shi et al., 2023), etc. These models leverage in-context learning (ICL) or chain-of-thought (CoT) prompting (Wei et al., 2024) which enable LLMs to reason step by step and interact with human experts. In-context learning for molecular tasks usually combines instruction-based prompts with a few molecular Question-Answer examples. Chen et al. (2024); Li et al. (2024a) design few-shot prompts with role definitions, task descriptions, in-context examples and output control to guide the prediction of LLMs. Differently, ReLM (Shi et al., 2023) in-

tegrates LLM as a decision-maker to enhance the reaction prediction results from external model.

The autonomous reasoning of LLM agents can be achieved by chain-of-thought prompting. The CoT method directly demonstrates the reasoning steps in one or a few prompts, and the agent can leverage the emergent ability of LLMs to imitate similar reasoning in the same types of tasks. With effective CoT and access to external knowledge, LLM agents can work semi-autonomously to support experts in scientific research. In StructChem (Ouyang et al., 2025), GPT-4 is guided to solve chemistry problems through formula generation and step-by-step reasoning and self-refinement. ChemCrow (M. Bran et al., 2024) adopts least-to-most prompting (Zhou et al., 2023) (LtM), which can be seen as CoT in an autoregressive manner. The reasoning loop in ChemCrow integrates the decomposition of the task, the selection and use of external tools, and the analysis of the result. The input of the next reasoning loop is built upon the current results until they satisfy the expected format. It is the first LLM agent capable of automatically completing complex planning and synthesis tasks.

5 Dataset Construction

The quality of the training data is crucial for cross-modal alignment and training, significantly influencing the performance of language-molecule models. In this section, we focus on summarizing some common dataset construction methods.

Data Processing To facilitate alignment, pairs of textual and non-textual molecular data are collected from public datasets. However, the content of descriptions in databases is not balanced. Taking PubChem (Kim et al., 2022) as an example, it is very often that some molecules only have a few basic records and lack some detailed properties. To address this issue, many researchers construct training data from multiple datasets or retrieve relevant text from scientific corpus such as S2orc (Lo et al., 2020). Meanwhile, the pre-processing methods are also important. For example, Liu et al. (2023a); Zhang et al. (2023); Cao et al. (2023) first replace all the molecule names in the annotation of PubChem with token '∼' to simplify the comprehension of name in training. Then they remove redundant information in the molecule description, such as origins, sources, and some geographic notation that has no relation to the target tasks. Xu et al. (2023) select four types of key properties from Swiss-Prot (Bairoch and Apweiler, 2000) and use fixed templates to rearrange descriptions, ensuring the consistency of the training data format.

Integrating Generative AI Recent advances in generative AI provide an innovative approach to mitigate the data scarcity challenge. For instance, Li et al. (2024b) use GPT-3.5 to enrich the sparse molecular descriptions in PubChem. Fang et al. (2024); Xiao et al. (2024b) leverage GPT to diversify prompt templates and use them to generate QA pairs for instruction-tuning. Additionally, Sakhinana and Runkana (2023) uses GPT-4 to generate molecule captions for fine-tuning. Chen et al. (2024) fabricate an "artificially-real" dataset for domain adaptation, where molecule descriptions are generated through ChatGPT with retrieval-based few-shot prompting.

6 Applications

This section will showcase applications of the aforementioned methods in drug discovery and chemistry research. Beyond the introduction of tasks, we also emphasize the adaptation between base models and tasks.

6.1 Text-molecule Retrieval

The text-molecule retrieval task is first proposed by Edwards et al. (2021), which aims to retrieve the corresponding molecule from a given text query. This molecule retrieval can be applied in the early stages of drug discovery, where experts need to select potential molecules from the compound database for further design and optimization. The retrieval task can be accomplished by the aligned latent space, from which we can acquire the encoded text descriptions with implicit connection of target molecules. Then we can use the similarity score to evaluate the distance between text and molecules to find the best-matched pair. In KV-PLM (Zeng et al., 2022), descriptions and molecules are encoded by a shared transformer encoder. While MoMu (Su et al., 2022) and MoleculeSTM (Liu et al., 2023a) use separate encoders to extract multimodal features and align the latent space with contrastive learning.

6.2 Property Prediction

One of the important goals of drug discovery is to search for small molecules and proteins with

desired structures and properties. The description of molecules in the scientific literature and databases can serve as knowledge repositories that contain properties, interactions, and structures that can hardly be inferred from current models (Pei et al., 2023). Through molecule-text alignment, text information can act as an additional modality to enhance molecular representation and improve performance in property prediction tasks (Seidl et al., 2023; Xu et al., 2023). The property prediction task is usually in binary classification format and is achieved by fused molecular features and a prediction head. An alternative approach is to leverage powerful generative LLMs with instructions to predict properties in QA format (Zhang et al., 2023; Liu et al., 2024b). As shown in 4.1, property prediction is achieved by the probabilities of "true" or "false" tokens in the generated answer.

6.3 Molecule Design

***De novo* Generation** *De novo* generation in molecule design includes molecule captioning that generates a description of given molecules and text-guided *de novo* generation which generates molecules from scratch with textual guidance. Models with single-stream architecture have the privilege of performing translation between text and molecule, owing to the encoder-decoder structure and text-to-text task format (Raffel et al., 2020). Apart from the translation-based methods. Liu et al. (2024c) propose a protein design framework with a multi-stream encoder. In text-guided protein generation task, the description is first encoded by the aligned text encoder. Then a facilitator module which is parameterized by a multi-layer perception is used to learn the transformation from encoded text to protein representation. The resulting protein representation is then fed into a trained generative decoder to generate protein sequences.

Molecule Editing Molecule editing seeks to optimize current molecules with desired properties. Within the drug discovery pipeline, text-guided editing finds application in lead optimization tasks and proves valuable for decomposing multi-objective lead optimization (Liu et al., 2024c). Drawing inspiration from the success of few-shot text-to-image generation, text description can simplify the complexity of the target chemical space in the generation process. Simultaneously, diversified generation enhances drug editing by introducing high flexibility. As mentioned above, the latent space alignment establishes a unified latent space where features possess semantic meaning in both structure and text. Building upon this approach, Liu et al. (2023a, 2024c); Tang et al. (2024) use latent optimization methods to sample a latent representation close to both text and molecule in latent space. Then, this latent code is fed into a decoder which is usually a trained molecule generation model to produce optimized molecules. Kim et al. (2025) proposes hierarchical textual inversion that introduces intermediate and detail tokens to represent SMILES, with the aim of capturing cluster and molecule-level characteristics. The interpolation sampling can benefit from this hierarchical design with high generation diversity.

6.4 Other Applications

Reaction Prediction Reaction prediction is a challenging but fundamental task in chemistry. The chemical reaction process can be seen as a mapping between a set of reactants and a set of products with specific reaction conditions. Under this framework, there are three main reaction prediction tasks, which are product prediction, reaction condition prediction, and most importantly, retrosynthesis prediction. Text can help to understand complex reaction mechanisms and supply information about reaction templates that GNN-based methods often fail to capture. Qian et al. (2023) retrieve reaction-related text and concatenate with input SIMILES to enhance retrosynthesis prediction. As described in 4.3, we can also involve LLMs in reaction prediction via prompt engineering. For example, Shi et al. (2023) use GPT-4 to predict reaction products with the aid of in-context reaction examples and candidate products from external model.

Intelligent Agent for Scientific Research According to M. Bran et al. (2024), the automation level in chemistry is relatively low compared to other domains. Although LLMs may have difficulties in comprehending chemistry principles, they have demonstrated significant capability in understanding human instructions and organizing information based on extensive training corpora (AI4Science and Quantum, 2023). Consequently, LLMs have the potential to become intelligent assistants to automatically arrange research with the help of professional tools and software. Liu et al. (2024d) design a drug editing agent with conversational interaction. The agent can receive human feedback to retrieve candidate drug molecules from

the database with desired properties. Similarly to ChemCrow (M. Bran et al., 2024), Boiko et al. (2023) develop a "Co-scientist" based on GPT-4 that can independently design and execute chemical research.

7 Conclusions and Future Outlooks

In this paper, we provide a comprehensive review of language-molecule models. After a brief introduction to the background and molecule descriptors, we introduce the model architectures and pretraining tasks for latent space alignment. Then, we summarize the prompting techniques in multimodal large language models which serve as bridge between LLMs and downstream molecular tasks. As an application-oriented domain, we combine the aforementioned methods to exhibit applications in drug discovery and chemistry. Although text-molecule models have made impressive progress, there exist several challenges which appeal to future research.

7.1 Appealing for High-Quality Data and Reliable Benchmarks

According to the neural scaling law, the emergent abilities of LLM in complex molecular tasks have not been shown. The data scarcity challenge still exists for both molecular structures and textual descriptions. In addition to collecting descriptions from databases, many works also automatically retrieve relative text from scientific corpus or using generative tools, while the authenticity and correlation of the retrieved or generated text cannot be guaranteed (Xu et al., 2023; Tang et al., 2024). For the progress of the community, a larger and more qualified molecule-text database is significant. Although language-molecule models exhibit great potential in various molecular tasks, there remains a question of how to fairly evaluate the performance among different models. To address this concern, new benchmarks are necessary to standardize evaluation metrics and settings, providing more reliable and realistic test data (Guo et al., 2024; Fang et al., 2024; Yu et al., 2024).

7.2 Extending the Interpretability of Model

The lack of interpretability prohibits many applications of deep molecular models, since numerical predictions alone may not be convincing enough compared to computational and experimental results. Text-involved frameworks provide an opportunity to enhance the interpretability of the results. By leveraging in-context learning and chain-of-thought prompting in LLMs, models can reasoning and inference, like the human brain, to produce explainable results. Follow-up research can also try to develop interpretable tools to bridge the relation between textural description and molecular structure in latent space (Su et al., 2022).

7.3 Improving the Reasoning Ability

The application of prompting techniques can significantly improve the reasoning ability of LLM-based frameworks. However, it is observed that in some cases, models may generate unrealistic predictions or even replicate the values in examples as prediction (Zhao et al., 2023c). This serves as evidence that LLMs may rely on memorization without truly understanding the molecules and chemical problems. Future studies may integrate successful GNNs into language model architecture (Zhao et al., 2023b), other than simply using GNNs as encoders (Zachares et al., 2023). Designing effective prompts for molecular tasks can also be taken into consideration.

7.4 Integration with Foundation Models

Foundation models (FMs) in the biomedical domain have shown promising performance. For example, AlphaFold (Jumper et al., 2021) can accurately predict protein structures when only protein sequence is available. It is possible to integrate FMs within LLM agents or specially designed frameworks (Wang et al., 2024d). We believe that effective frameworks could unlock the additive power of FMs.

7.5 Learning from Human/AI Feedback

Recent progress in reinforcement learning from human/AI feedback (i.e., RLHF (Ouyang et al., 2024) and RLAIF (Lee et al., 2023)) has achieved promising results in aligning LLMs with human preference. RLHF fits a reward model to human preference dataset and uses RL to optimize LLMs to produce responses assigned with high rewards. This paradigm may pave the way for utilizing LLMs for biomedical applications, especially in scenarios where molecular simulation software can be used as a reward model. Exploring how to fully utilize the power of RLHF at the interaction of text and molecules is an appealing research direction.

Limitations

This work primarily focuses on language-molecule models that connect human language with molecular data. Although other studies integrate additional modalities, such as molecular structures and images (Sanchez-Fernandez et al., 2023), we do not cover these due to space limitations and leave their survey for future work. Similarly, knowledge graphs designed for molecular research or drug discovery may incorporate textual data, but our emphasis is on cross-modal training and the integration of language models. Given the success of LLMs across various domains, we consider this to be a more promising direction. Additionally, dataset sources are not included in this study due to the complexity of data collection and pre-processing, as well as space constraints. Instead, we outline common data processing strategies that are useful for dataset construction.

References

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. 2024. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, pages 10757–10765.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *Preprint*, arXiv:2311.07361.

Seth D Axen, Xi-Ping Huang, Elena L Cáceres, Leo Gendelev, Bryan L Roth, and Michael J Keiser. 2017. A simple representation of three-dimensional molecular structure. *Journal of medicinal chemistry*, 60(17):7393–7409.

Amos Bairoch and Rolf Apweiler. 2000. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Research*, 28(1):45–48.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. 2022. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *Preprint*, arXiv:2311.16208.

Yuhan Chen, Nuwa Xi, Yanrui Du, Haochun Wang, Jianyu Chen, Sendong Zhao, and Bing Qin. 2024. From artificially real to real: Leveraging pseudo data from large language models for low-resource molecule discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):21958–21966.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Fengyuan Dai, Yuliang Fan, Jin Su, Chentong Wang, Chenchen Han, Xibin Zhou, Jianming Liu, Hui Qian, Shunzhi Wang, Anping Zeng, Yajie Wang, and Fajie Yuan. 2024. Toward de novo protein design from natural language. *bioRxiv*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+ m-24: Building a dataset for language+ molecules@ acl 2024. *CoRR*.

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

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*.

Klemens Flöge, Srisruthi Udayakumar, Johanna Sommer, Marie Piraud, Stefan Kesselheim, Vincent Fortuin, Stephan Günneman, Karel J van der Weg, Holger Gohlke, Alina Bazarova, and Erinc Merdivan. 2024. Oneprot: Towards multi-modal protein foundation models. *Preprint*, arXiv:2411.04863.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. 2024. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa, Le Song, and Pengtao Xie. 2024. Multi-modal large language model enables protein function prediction. *bioRxiv*.

Hyosoon Jang, Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2024. Can llms generate diverse molecules? towards alignment with structural diversity. *Preprint*, arXiv:2410.03138.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. ProLLM: Protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. In *First Conference on Language Modeling*.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. 2020. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Seojin Kim, Jaehyun Nam, Sihyun Yu, Younghoon Shin, and Jinwoo Shin. 2025. Data-efficient molecular generation with hierarchical textual inversion. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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. 2022. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024a. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6071–6083.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024b. Towards 3d molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations*.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023. Drugchat: Towards enabling chatgpt-like capabilities on drug molecule graphs. *Preprint*, arXiv:2309.03907.

Nuowei Liu, Changzhi Sun, Tao Ji, Junfeng Tian, Jianxin Tang, Yuanbin Wu, and Man Lan. 2024a. Evollama: Enhancing llms' understanding of proteins via multimodal structure and sequence representations. *Preprint*, arXiv:2412.11618.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024b. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Comput. Biol. Med.*, 171(C).

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2024c. A text-guided protein design framework. *Preprint*, arXiv:2302.04611.

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

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024d. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations*.

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638, Singapore. Association for Computational Linguistics.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024e. ProtT3: Protein-to-text generation for text-based protein understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5949–5966, Bangkok, Thailand. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yizhen Luo, Xing Yi Liu, Kai Yang, Kui Huang, Massimo Hong, Jiahuan Zhang, Yushuai Wu, and Zaiqing Nie. 2024a. Toward unified ai drug discovery with multimodal knowledge. *Health Data Science*, 4:0113.

Yizhen Luo, Zikun Nie, Massimo Hong, Suyuan Zhao, Hao Zhou, and Zaiqing Nie. 2024b. MutaPLM: Protein language modeling for mutation explanation and engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023a. Molfm: A multimodal molecular foundation model. *Preprint*, arXiv:2307.09484.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, Zikun Nie, Hao Zhou, and Zaiqing Nie. 2024c. Learning multi-view molecular representations with structured and unstructured knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2082–2093, New York, NY, USA. Association for Computing Machinery.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023b. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein language model for multi-task protein language processing. *Preprint*, arXiv:2402.16445.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

Long D. Nguyen, Quang H. Nguyen, Quang H. Trinh, and Binh P. Nguyen. 2024. From smiles to enhanced molecular property prediction: A unified multimodal framework with predicted 3d conformers and contrastive learning techniques. *Journal of Chemical Information and Modeling*, 64(24):9173–9195. PMID: 39641280.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sirui Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. 2025. Structured chemistry reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024a. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1216–1240, Bangkok, Thailand. Association for Computational Linguistics.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, and Rui Yan. 2024b. 3d-molt5: Towards unified 3d molecule-text modeling with 3d molecular tokenization. *Preprint*, arXiv:2406.05797.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

Yujie Qian, Zhening Li, Zhengkai Tu, Connor Coley, and Regina Barzilay. 2023. Predictive chemistry augmented with text retrieval. In *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Andre Niyongabo Rubungo, Craig Arnold, Barry P. Rand, and Adji Bousso Dieng. 2023. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *Preprint*, arXiv:2310.14029.

Sagar Sakhinana and Venkataramana Runkana. 2023. Crossing new frontiers: Knowledge-augmented large language model prompting for zero-shot text-based de novo molecule design. In *NeurIPS 2023 Workshop on R0-FoMo*.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. 2023. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

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

Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. 2024. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. 2024. Mollm: a unified language model for integrating biomedical text with 2d and 3d molecular representations. *Bioinformatics*, 40:i357–i368.

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. *Preprint*, arXiv:2211.09085.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. 2024a. Protchatgpt: Towards understanding proteins with large language models. *Preprint*, arXiv:2402.09649.

Runze Wang, Mingqi Yang, and Yanming Shen. 2024b. Graph2token: Make LLMs understand molecule graphs. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2024c. InstructProtein: Aligning human and protein language via knowledge instruction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1114–1136, Bangkok, Thailand. Association for Computational Linguistics.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and RISHITA ANUBHAI. 2024d. Biobridge: Bridging biomedical foundation models via knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kevin E. Wu, Howard Chang, and James Zou. 2024a. Proteinclip: enhancing protein language models with natural language. *bioRxiv*.

Wei Wu, Chao Wang, Liyi Chen, Mingze Yin, Yiheng Zhu, Kun Fu, Jieping Ye, Hui Xiong, and Zheng Wang. 2024b. Structure-enhanced protein instruction tuning: Towards general-purpose protein understanding. *Preprint*, arXiv:2410.03553.

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

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G. Honavar. 2024a. Molbind: Multimodal alignment of language, molecules, and proteins. *Preprint*, arXiv:2403.08167.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. 2024b. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *Preprint*, arXiv:2408.11363.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. Protst: multi-modality learning of protein sequences and biomedical texts. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Gokul Yenduri, M. Ramalingam, G. Chemmalar Selvi, Y. Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G. Deepti Raj, Rutvij H. Jhaveri, B. Prabadevi, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2024. Gpt (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, 12:54608–54649.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*.

Peter A Zachares, Vahan Hovhannisyan, Alan Mosca, and Yarin Gal. 2023. Form follows function: Text-to-text conditional graph generation based on functional requirements. *arXiv preprint arXiv:2311.00444*.

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

Juzheng Zhang, Yatao Bian, Yongqiang Chen, and Quanming Yao. 2024a. Unimot: Unified molecule-text language model with discrete token representation. *Preprint*, arXiv:2408.00863.

Li Zhang, Han Guo, Leah Schaffer, Young Su Ko, Digvijay Singh, Hamid Rahmani, Danielle Grotjahn, Elizabeth Villa, Michael Gilson, Wei Wang, Trey Ideker, Eric Xing, and Pengtao Xie. 2024b. Proteinaligner: A multi-modal pretraining framework for protein foundation models. *bioRxiv*.

Weitong Zhang, Xiaoyun Wang, Weili Nie, Joe Eaton, Brad Rees, and Quanquan Gu. 2023. MoleculeGPT: Instruction following large language models for molecular property prediction. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023a. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jianan Zhao, Le Zhuo, Yikang Shen, Meng Qu, Kai Liu, Michael Bronstein, Zhaocheng Zhu, and Jian Tang. 2023b. Graphtext: Graph reasoning in text space. *Preprint*, arXiv:2310.01089.

Lawrence Zhao, Carl Edwards, and Heng Ji. 2023c. What a scientific language model knows and doesn't know about chemistry. In *NeurIPS 2023 AI for Science Workshop*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. 2024. 3m-diffusion: Latent multi-modal diffusion for language-guided molecular structure generation. In *First Conference on Language Modeling*.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. 2024. ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8950–8963, Bangkok, Thailand. Association for Computational Linguistics.

A Summary Table of Language-molecule Models

Table 1 summarizes molecule descriptors, backbone architectures and pre-training tasks of language-molecule models. We categorize these models by their architectures: single-stream architecture, multi-stream architecture and intelligent agent.

Model	Molecule descriptors	Backbone architecture	Pre-Training task
MolT5 (Edwards et al., 2022)	SMILES	T5	MLM
Galactica (Taylor et al., 2022)	Bio-Sequence	Transformer Decoder	CLM
KV-PLM (Zeng et al., 2022)	SMILES	SciBERT (Beltagy et al., 2019)	MLM
MolXPT (Liu et al., 2023b)	SMILES	GPT	CLM
Text + Chem T5 (Christofidellis et al., 2023)	SMILES	T5	CG
TextReact (Qian et al., 2023)	SMILES	SciBERT	CL + MLM + CG
GIMLET (Zhao et al., 2023a)	Graph	T5	CG
BioT5 (Pei et al., 2023)	SELFIES + Protein Sequence	T5	MLM + CG
3D-MolT5 (Pei et al., 2024b)	SELFIES + Fingerprints	T5	CG+ MLM
BIOT5+ (Pei et al., 2024a)	SELFIES + IUPAC + Protein Sequence	T5	CG+ MLM
ProLLM (Jin et al., 2024)	Protein Sequence	T5	MLM
ProLLaMA (Lv et al., 2024)	Protein Sequence	Llama-2	CLM
LLM-Prop (Rubungo et al., 2023)	Crystal String	T5	MLM
Gruver et al. (2024)	Crystal String	LLaMA-2	MLM
Text2Mol (Edwards et al., 2021)	Graph	Multi-stream + Transformer	CL
MoMu (Su et al., 2022)	Graph	Multi-stream	CL
DrugChat (Liang et al., 2023)	Graph	Multi-stream + Vicuna-13b	CLM
MoleculeSTM (Liu et al., 2023a)	Graph	Multi-stream + Decoder	CL
Graph2Token (Wang et al., 2024b)	Graph	Multi-stream + Vicuna-7B	CG
MV-Mol (Luo et al., 2024c)	Graph	Q-Former+ BioT5	CL + MTM + CLM
3M-Diffusion (Zhu et al., 2024)	Graph	Multi-stream	CL
MolFM (Luo et al., 2023a)	Graph	Multi-stream	CL + MTM + MLM
BioMedGPT (Luo et al., 2023b)	Graph + Protein Sequence	Multi-stream + LLaMA 2	CLM
MOLBIND (Xiao et al., 2024a)	Graph + Geometry + Protein Graph	Multi-stream	CL
GIT-Mol (Liu et al., 2024b)	SMILES + Graph + Image	Q-Former + T5	MTM + CL
MolLM (Tang et al., 2024)	SMILES + Graph + Geometry	Multi-stream	CL
MolCA (Liu et al., 2023c)	SMILES + Graph	Q-Former + Llama 2	MTM + CL + MC + CLM
3D-MoLM (Li et al., 2024b)	SMILES + Geometry	Q-Former + Llama 2	MTM + CL + MC + CLM
MoleculeGPT (Zhang et al., 2023)	SMILES + Graph	Q-Former + Vicuna-7b	CL+CLM
BioBridge (Wang et al., 2024d)	SMILES + Protein Sequence	Knowledge Graph	CL
Nguyen et al. (2024)	SMILES + Geometry	Multi-stream	CLM
UniMoT (Zhang et al., 2024a)	SMILES + Graph	Q-Former + Llama 2	MTM + CL + CG + CLM
InstructMol (Cao et al., 2023)	SELFIES + Graph	Multi-stream + Vicuna-7b	CLM
CLAMP (Seidl et al., 2023)	Fingerprints	Multi-stream	CL
Proteinchat (Huo et al., 2024)	Protein Sequence	Multi-stream + Vicuna-13B	CLM
MutaPLM (Luo et al., 2024b)	Protein Sequence	Multi-stream + LLaMA2-7B	CLM + MLM + CG
ProtST (Xu et al., 2023)	Protein Sequence	Multi-stream	CL + MLM
ProtDT (Liu et al., 2024c)	Protein Sequence	Multi-stream + Decoder	CL
InstructProtein (Wang et al., 2024c)	Protein Sequence	Knowledge Graph + LLMs	CLM
ProteinCLIP (Wu et al., 2024a)	Protein Sequence	Multi-stream	CL
PROTLLM (Zhuo et al., 2024)	Protein Sequence	Multi-stream	CLM
ProtT3 (Liu et al., 2024e)	Protein Sequence	Q-Former + LLMs	MTM + CL + CG
SEPIT (Wu et al., 2024b)	Protein Sequence	Multi-stream + LLMs	CLM
Pinal (Dai et al., 2024)	Protein Sequence	Multi-stream	CLM
OneProt (Flöge et al., 2024)	Protein Sequence + Protein Graph	Multi-stream	CL
EVOLLAMA (Liu et al., 2024a)	Protein Sequence + Protein Graph	Multi-stream + Llama-3	CL
Prot2Text (Abdine et al., 2024)	Protein Sequence + Protein Graph	Multi-stream + Transformer	CLM
ProtChatGPT (Wang et al., 2024a)	Protein Sequence + Protein Graph	Q-Former + Vicuna-13b	MTM + CG + CL + CLM
ProteinAligner (Zhang et al., 2024b)	Protein Sequence + Protein Graph	Multi-stream	CL
ProteinGPT (Xiao et al., 2024b)	Protein Sequence + Protein Graph	Multi-stream + Llama-3	CLM
ProTrek (Su et al., 2024)	Protein Sequence + Protein Graph	Multi-stream	CL + MLM
ReLM (Shi et al., 2023)	SMILES + IUPAC + Graph	ICL + LLMs	-
ChatDrug (Liu et al., 2024d)	SMILES	LLMs	-
MolReGPT (Li et al., 2024a)	SMILES	ICL + GPT-3.5	-
ChemCrow (M. Bran et al., 2024)	-	CoT + LLMs	-
Jang et al. (2024)	-	LLMs + RL	-

Table 1: Summary of representative language-molecule models. “Graph” and “Geometry” denote 2D graph and 3D geometric graph for small molecule respectively.