# Bridging Text and Molecule: A Survey on Multimodal Frameworks for Molecule

*Yi Xiao, Xiangxin Zhou, Qiang Liu and Liang Wang*

University of Chinese Academy of Sciences, Beijing, China

Presented by: Yijingxiu Lu

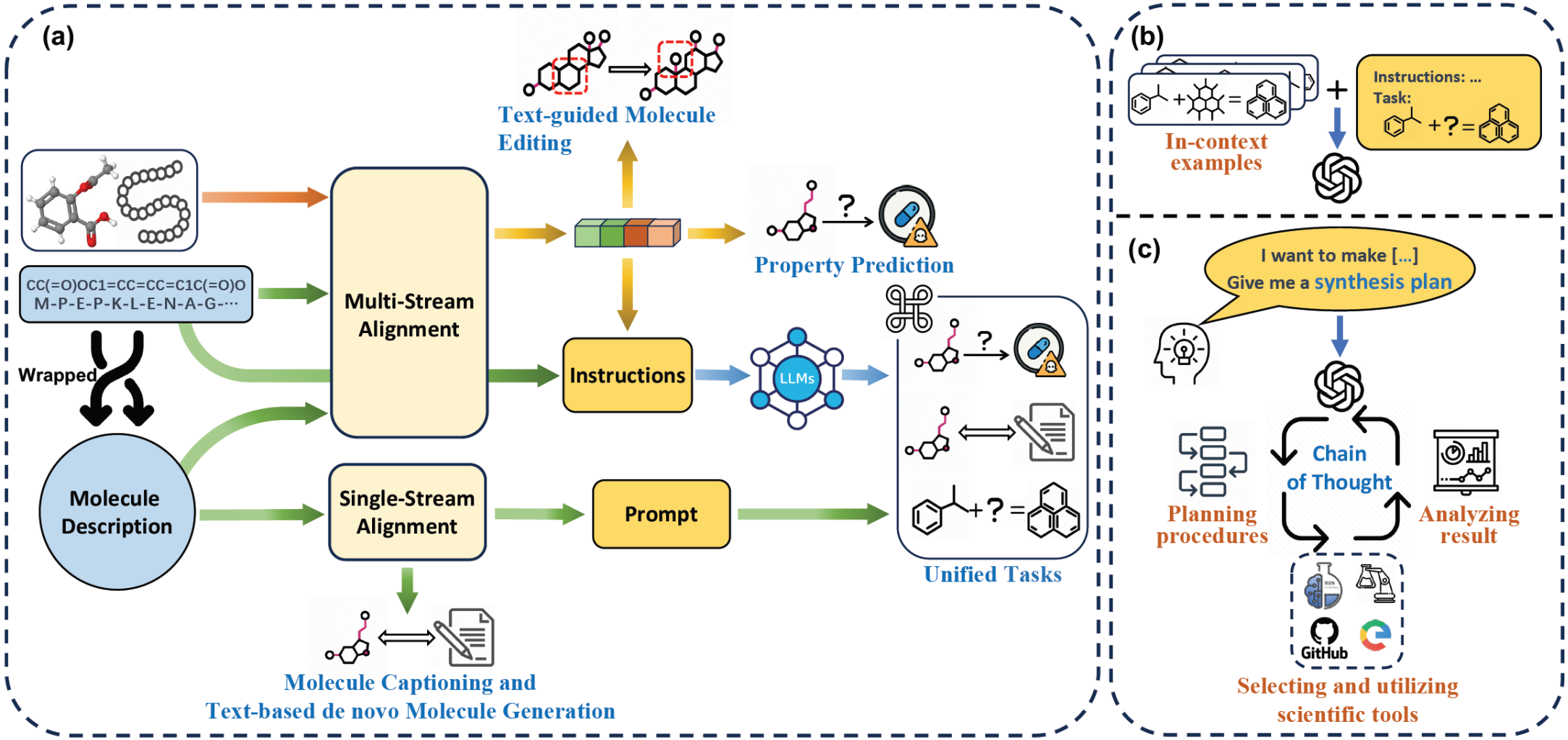
# Summary

- **Encoding Techniques:**
  - 1D molecule sequences,
  - 2D molecular structures
  - 3D molecular structures.

- **Model Architectures:**
  - single-stream architectures
  - multi-stream architectures

- **Pre-training Tasks:**
  - Molecule-Text Contrastive Learning(CL), Molecule-Text Matching (MTM), Conditional Generation (CG), Masked Language Modeling (MLM), and Causal Language Modeling (CLM).

- **Prompting Techniques:**
  - prompting-based fine-tuning
  - instruction tuning (IT), in-context learning (ICL), and chain-of-thoughts (CoT) prompting.


(a)
Text-guided Molecule Editing
CC(=O)OC1=CC=CC=C1C(=O)O
M-P-E-P-K-L-E-N-A-G-···
Wrapped
Molecule Description
Multi-Stream Alignment
Single-Stream Alignment
Instructions
Prompt
LLMs
Property Prediction
Unified Tasks
Molecule Captioning and
Text-based de novo Molecule Generation
(b)
In-context examples
Instructions: ...
Task:
(c)
I want to make [...]
Give me a synthesis plan
Chain of Thought
Planning procedures
Analyzing result
GitHub
Selecting and utilizing scientific tools

# Encoding

## 1D Molecule Sequence

Small-molecule Sequence

*Simplified Molecular Input Line Entry System (SMILES)*
*Self-referencing embedded strings (SELFIES)*
*International Union of Pure and Applied Chemistry (IUPAC)*
*Molecular FP*

Protein Sequence

*Amino Acid Sequence*
*Protein Language Model (PLM)*

## 2D Molecule Structure

2D Graph

*Atom (Node), Bond (Edge)*

## 3D Molecule Structure

3D Geometric Graph

*2D Graph + 3D Coordinates*

Protein Graph

*Residue (Node), Distance (Edge)*

# Methodology

## Model Architecture

Single-Stream Architecture
*Assume the latent space of molecules and text share similar semantic meaning*

Multi-Stream Architecture
*Employ intra-modality processing for text and molecules*

## Pre-training Tasks

Molecule-Text Contrastive Learning (CL)
*Matched pair closer*

Molecule-Text Matching (MTM)
*Pair prediction*

Conditional Generation (CG)
*Token generation*

Masked Language Modeling (MLM)
*Masked token prediction*

Causal Language Modeling (CLM)
*Next Token Generation*

## Prompting Techniques

Prompting-based fine-tuning
*task-specific prompt*

Instruction tuning (IT)
*Multi-task instructions for seamlessly transferring*

In-context learning (ICL)
*prompt with question-answering*

Chain-of-thoughts (CoT)
*reasoning step-by-step*

# Methodology

## Model Architecture

Single-Stream Architecture

*Assume the latent space of molecules and text share similar semantic meaning*

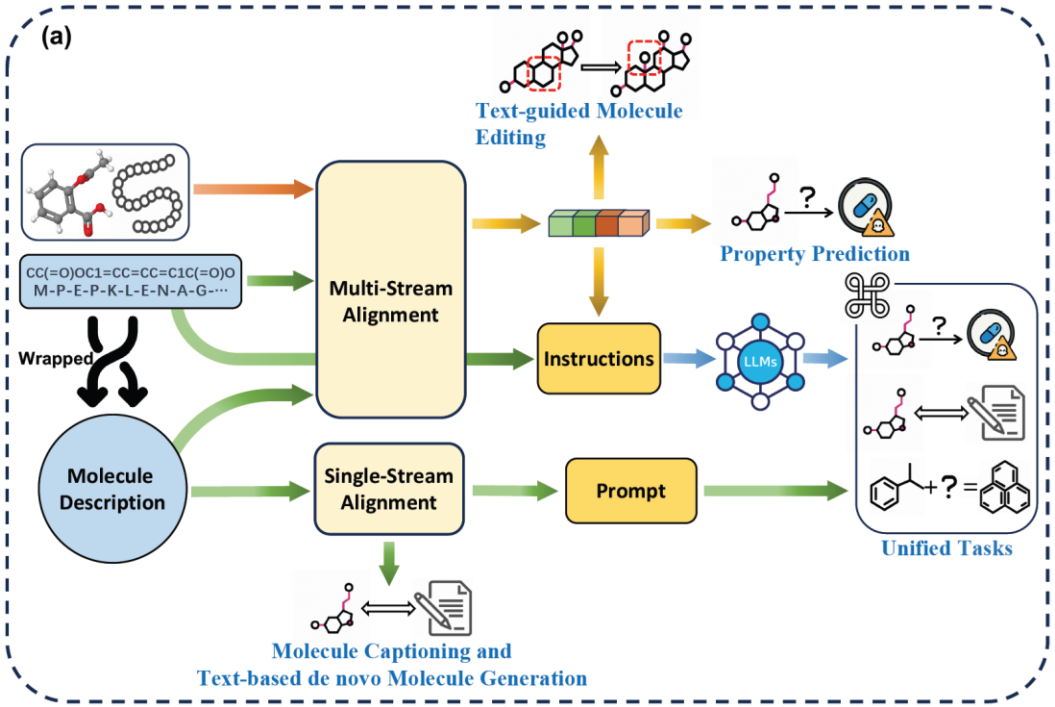
Multi-Stream Architecture

*Employ intra-modality processing for text and molecules*

# Model Architecture

1) **Single-Stream Architecture:**
   - Assume the latent space of molecules and text share similar semantic meaning.

2) **Multi-Stream Architecture:**
   - Employ intra-modality processing for text and molecules.



# Model Architecture | *Single-Stream Architecture*

- **T5 [Raffel et al., 2020]**: A general text-to-text transformation model used for multi-language pretraining.
- **KV-PLM [Zeng et al., 2022]**: Utilizes Byte-Pair Encoding (BPE) to tokenize SMILES and replaces molecular names in sequences with SMILES tokens to create training data.
- **MolXPT [Liu et al., 2023b]**: Also applies BPE for molecular tokenization and embeds SMILES into textual sequences to enhance text-molecule alignment.
- **BioT5 [Pei et al., 2023]**: Creates separate vocabularies for molecules, proteins, and text to prevent token ambiguity arising from different semantic spaces.
- **GIMLET [Zhao et al., 2023a]**: Serializes molecular graphs as node sequences and introduces position embeddings to jointly encode nodes with textual tokens, preserving graph-level inductive bias while avoiding additional graph encoding modules.

| 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 |

***Advantages***:
- *Leverages existing NLP pretraining frameworks.*

***Weaknesses***:
- *Over-reliance on SMILES as a molecular representation may lead to information loss (e.g., structural details).*
- *Assumes molecules and text share the same semantic space, despite molecules having fundamentally different structural characteristics.*

# Model Architecture | *Multi-Stream Architecture*

- **[Abdine et al., 2023]**: Fuses protein sequence and protein graph features using element-wise addition and feeds them into a cross-attention module to adapt to textual information.
- **[Xu et al., 2023]:** Selects both text and protein representations as keys and applies two separate cross-attention modules to generate fused-text and fused-protein representations.
- **Q-Former [Li et al., 2023]:** A vision-language modeling architecture that leverages cross-attention layers to bridge the modality gap.
- **[Li et al., 2024; Liu et al., 2023d; Zhang et al., 2023]**: Adopt Q-Former to connect molecular graphs with text and extract text-related molecular features using a learnable query.
- **GIT-Former [Liu et al., 2023a]**: A variant of Q-Former that incorporates additional input modalities, such as molecular images and sequences, to enhance multimodal information fusion.

| | | | |
|---|---|---|---|
| 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 |

# Methodology

## Model Architecture

Single-Stream Architecture
*Assume the latent space of molecules and text share similar semantic meaning*

Multi-Stream Architecture
*Employ intra-modality processing for text and molecules*

## Pre-training Tasks

Molecule-Text Contrastive Learning (CL)
*Matched pair closer*

Molecule-Text Matching (MTM)
*Pair prediction*

Conditional Generation (CG)
*Token generation*

Masked Language Modeling (MLM)
*Masked token prediction*

Causal Language Modeling (CLM)
*Next Token Generation*

# Pre-training Tasks *To align the fused representation in a unified latent space*

1) **Molecule-Text Contrastive Learning (CL):**
   - pushes the embeddings from matched text and molecules closer in latent space while enlarging the distance between pairs from different molecules.

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

2) **Molecule-Text Matching (MTM):**
   - predict whether a molecule-text pair is matched or not.

$$\mathcal{L}_{\text{MTM}} = -\mathbb{E}\left[\sum_{i} [\log p(m_i, t_i) - \log p(m_i, t_j) - \log p(m_j, y_i)]\right]$$

3) **Conditional Generation (CG):**
   - generate tokens based on given conditions or constraints.

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

4) **Masked Language Modeling (MLM):**
   - predict the masked components using the remaining context.

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

5) **Causal Language Modeling (CLM):**
   - predict the next token in a sequence in a left-to-right direction.

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

# Methodology

## Model Architecture

Single-Stream Architecture
*Assume the latent space of molecules and text share similar semantic meaning*

Multi-Stream Architecture
*Employ intra-modality processing for text and molecules*

## Pre-training Tasks

Molecule-Text Contrastive Learning (CL)
*Matched pair closer*

Molecule-Text Matching (MTM)
*Pair prediction*

Conditional Generation (CG)
*Token generation*

Masked Language Modeling (MLM)
*Masked token prediction*

Causal Language Modeling (CLM)
*Next Token Generation*

## Prompting Techniques

Prompting-based fine-tuning
*task-specific prompt*

Instruction tuning (IT)
*Multi-task instructions for seamlessly transferring*

In-context learning (ICL)
*prompt with question-answering*

Chain-of-thoughts (CoT)
*reasoning step-by-step*

# Prompting Techniques *Multi-modal LLM in molecular science*

1) Prompting-based fine-tuning
   - *to unify the fine-tuning framework among different tasks with task-specific prompts.*
   - *e.g. "We can conclude that the BBBP of <SMILES> is <tag>"*

2) Instruction tuning (IT)
   - *models are trained in multiple tasks which have been unified through task-specific instructions.*
     - *<instruction> that clarifies the task*
     - *<input> which is usually molecular feature*
     - *<output> that embodies the expected outcome*

3) In-context learning (ICL)
   - *usually combines instruction-based prompts with a few molecular Question-Answer examples.*

4) Chain-of-thoughts (CoT)
   - *Few-shot:*
     - *demonstrates the reasoning steps in one or few prompts*
   - *Zero-shot:*
     - *e.g. put "Let's think step by step" at the end of the problem description*

# Prompting Techniques *Multi-modal LLM in molecular science*

1) Prompting-based fine-tuning
   - *to unify the fine-tuning framework among different tasks with task-specific prompts.*
   - *e.g. "We can conclude that the BBBP of <SMILES> is <tag>"*

2) Instruction tuning (IT)
   - *models are trained in multiple tasks which have been unified through task-specific instructions.*
     - *<instruction> that clarifies the task*
     - *<input> which is usually molecular feature*
     - *<output> that embodies the expected outcome*

3) In-context learning (ICL)
   - *usually combines instru*

| | | | |
|---|---|---|---|
| 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 | - |

4) Chain-of-thoughts (CoT)
   - *Few-shot:*
     - *demonstrates the reasoning steps in one or few prompts*
   - *Zero-shot:*
     - *e.g. put "Let's think step by step" at the end of the problem description*

# Dataset Construction

- **Data Processing:**
  1) Collect and construct dataset from multiple datasets (due to imbalance of descriptions in database).
  2) Replace unnecessary annotations.
  3) Remove redundant information.
  4) Rearrange descriptions to ensure consistency.

- **Integrating Generative AI:**
  - GPT-3.5 + PubChem (Li et al. 2024).
  - GPT-3.5 + QA (Fang et al. 2023).
  - GPT4 + molecule captioning (Sakhinana et al. 2023).
  - Fabricate an “artificially-real” database through ChatGPT with retrieval-based few-shot prompting (Chen et al. 2024).

# Applications

1) **Text-molecule Retrieval:**
   - retrieve the corresponding molecule from a given text query.
     - usually use the similarity score to evaluate the distance between text and molecules.

2) **Property Prediction:**
   a) binary classification task achieved by molecular features and simple prediction head.
   b) predict property in a QA format.

3) **Molecule Design:**
   ***a)*** ***De novo* Generation**: molecule captioning, text-guided de novo generation
   **b)** **Molecule Editing**: optimize current molecules with desired properties.
     - e.g. sample a latent representation close to both text and molecule in latent space.

4) **Others:**
   - **Reaction Prediction**: product prediction, reaction condition prediction, retrosynthesis prediction.
   - **Intelligent Agent for Scientific Research:**
     - [Liu et al., 2023b] design a drug editing agent with conversational interaction.
     - [Boiko et al., 2023] develop a “Coscientist” based on GPT-4 similar to ChemCrow [Bran et al., 2023] which can autonomously design and execute chemical research.

# Future Outlooks

1) **Appealing for High-Quality Data and Reliable Benchmarks:**
   - Data scarcity.
   - Authenticity and correlation of retrieved text cannot be guaranteed.
   - How to fairly evaluate the performance among different models.

2) **Extending the Interpretability of Model:**
   - Develop interpretable tools to improve interpretability of latent space.

3) **Improving the Reasoning Ability:**
   - In some cases, models may generate unrealistic predictions or even replicate the values in examples as prediction.

4) **Integration with Foundation Models:**
   - Integrate Foundation models (FMs) in the biomedical domain with LLM.

5) **Learning from Human/AI Feedback:**
   - Reinforcement learning from human/AI feedback (RLHF) + LLM