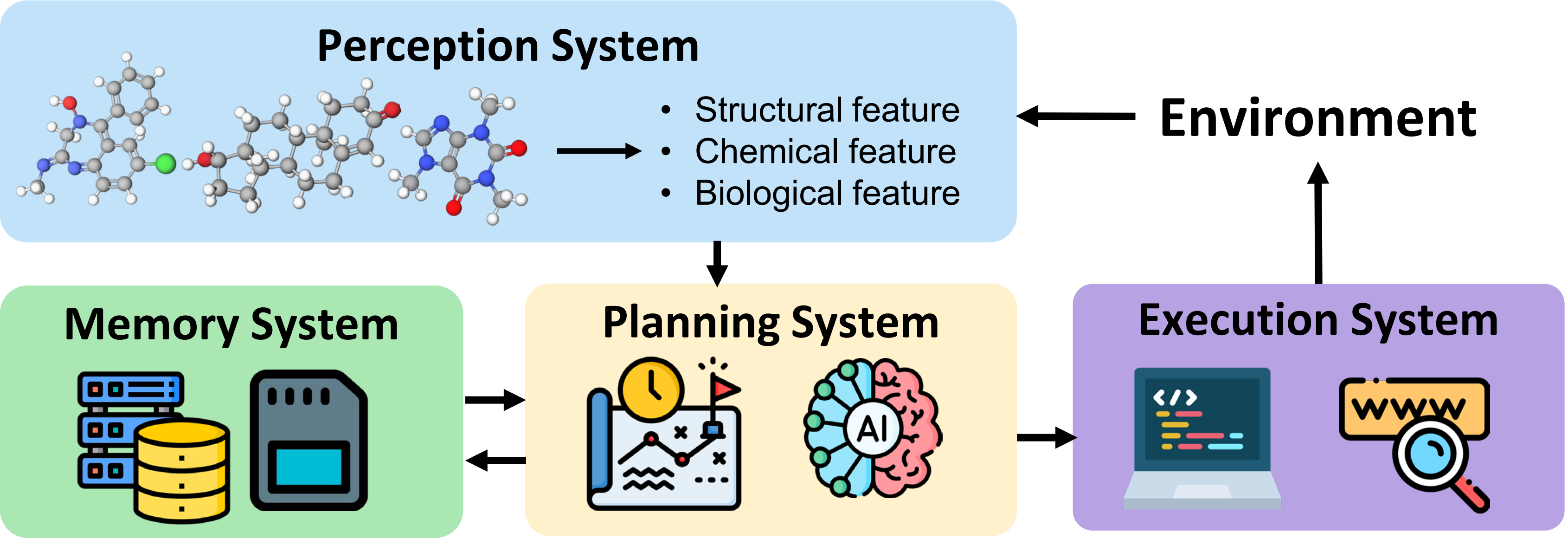
Mol-LLaMA: Towards General Understanding of Molecules in Large Molecular Language Model

Dongki Kim[1], Wonbin Lee[1], Sung Ju Hwang[1,2]
KAIST[1], DeepAuto.ai[2]

NeurIPS 2025

Agent System for Scientific Discovery

Even though the perception system takes a fundamental part in an agent system, LLMs have shortcomings to comprehend molecular structures and their properties.



Perception System

- Structural feature
- Chemical feature
- Biological feature

Environment

Memory System

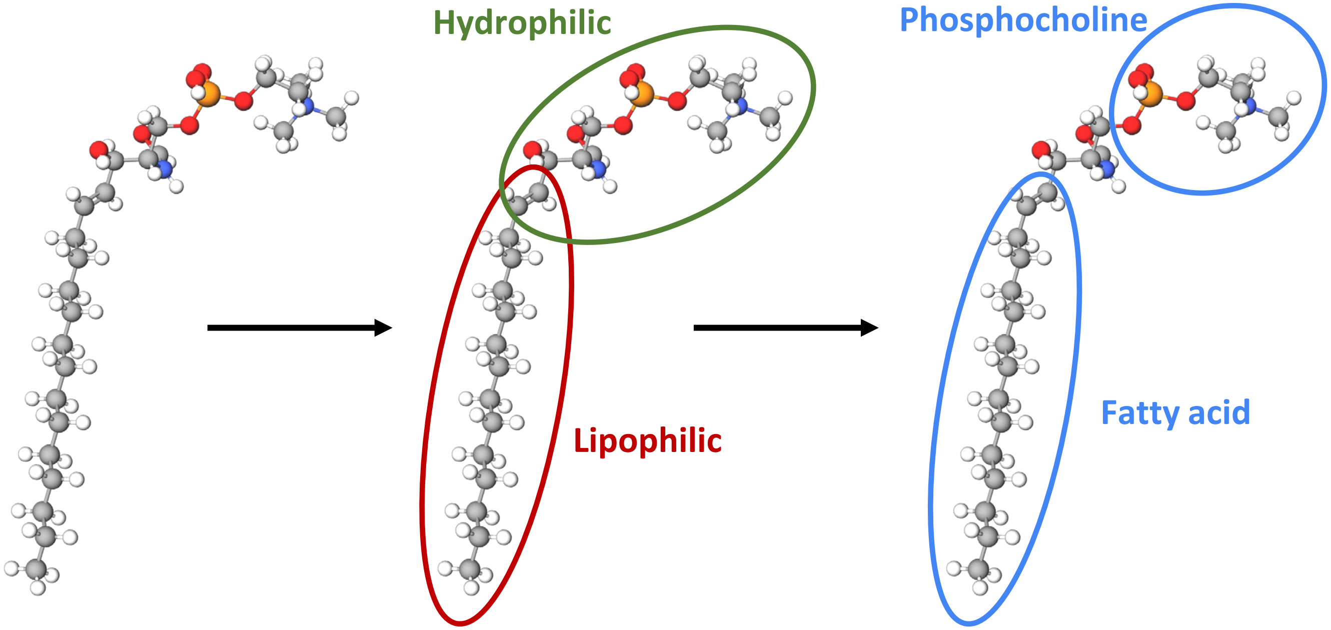
Planning System

Execution System

To overcome this limitation, we aim to develop a molecular LLM, that can be utilized as a general-purpose assistant for molecular analysis.

Molecular Semantics in Chemistry and Biology

We focus on the hierarchical relationship between molecular features: Chemical and biological properties are determined by structural features.

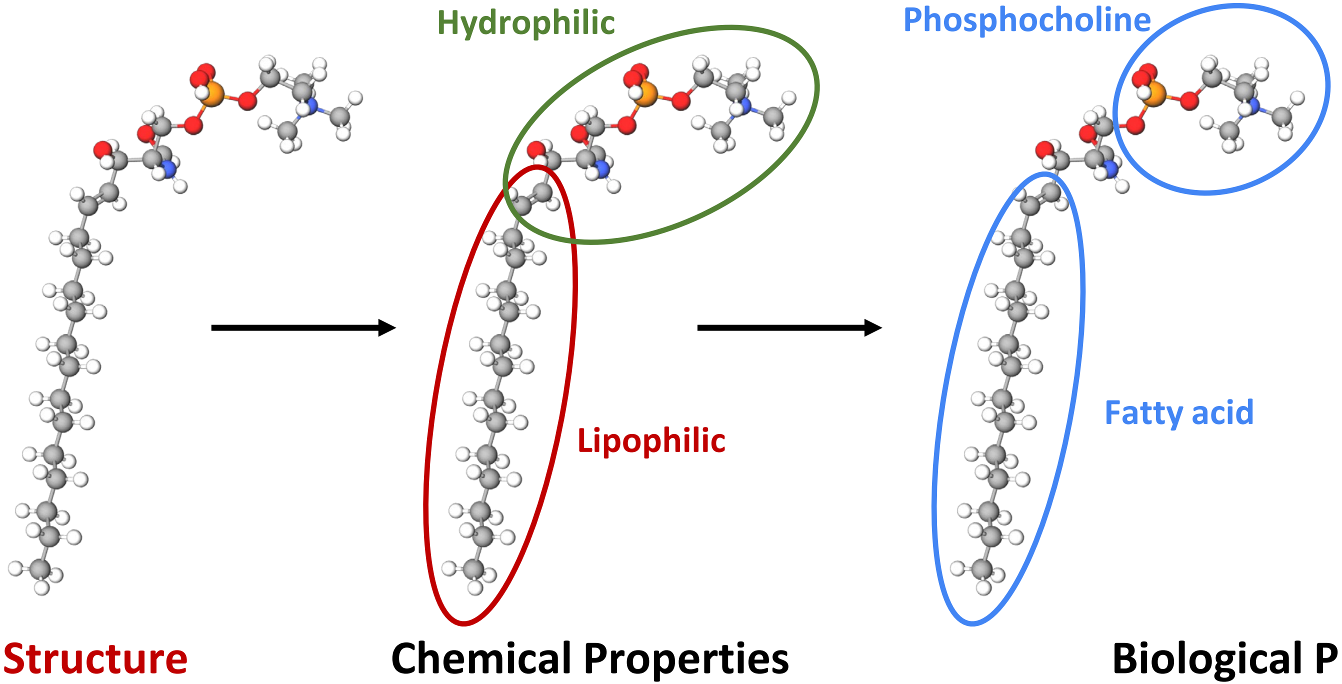


- Type of **Sphingolipid**
- Found in **cell membrane**
- **Cell Signaling**
- Higher concentration in the **nerve tissue**

Structure **Chemical Properties** **Biological Properties**

Molecular Semantics in Chemistry and Biology

We focus on the hierarchical relationship between molecular features: Chemical and biological properties are determined by structural features.



- Type of **Sphingolipid**
- Found in **cell membrane**
- **Cell Signaling**
- Higher concentration in the **nerve tissue**

Our Approach

To overcome the limitations of previous works, we focus on two main objectives.

- **Model Architecture:** Accurate modeling of molecular structures
    - Enhance the foundational understanding of general molecular features.
- **Instruction Dataset:** Training of general knowledge centered on molecules
    - Structures → Chemical features → Biological features

Our Approach

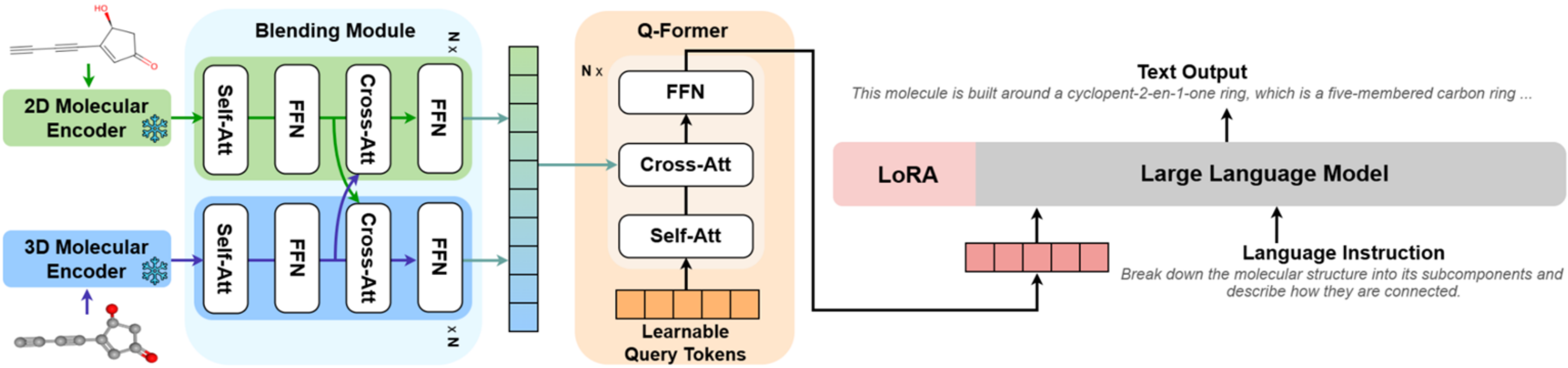
To overcome the limitations of previous works, we focus on two main objectives.

- **Model Architecture:** Accurate modeling of molecular structures
    - Enhance the foundational understanding of general molecular features.

- **Instruction Dataset:** Training of general knowledge centered on molecules
    - Structures → Chemical features → Biological features

Model Architecture

- Blending Module: We propose a blending module to fully leverage the advantages of different molecular encoder.
- Q-Former: To guarantee the permutation invariant, we leverage Q-Former architecture which performs the cross-attention.



This molecule is built around a cyclopent-2-en-1-one ring, which is a five-membered carbon ring ...

Break down the molecular structure into its subcomponents and describe how they are connected.

Our Approach

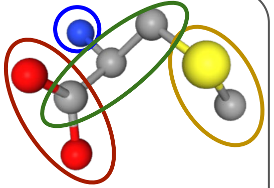
To overcome the limitations of previous works, we focus on two main objectives.

- **Model Architecture:** Accurate modeling of molecular structures
  - ➢ Enhance the foundational understanding of general molecular features.

- **Instruction Dataset:** Training of general knowledge centered on molecules
  - ➢ Structures → Chemical features → Biological features

Data Type: Detailed Structural Descriptions

To upskill the fundamental understanding of molecules and their features, we design detailed description of structures.



Instruction to GPT -4o

Explain which substructures and functional groups are contained and how they are connected.

Input Molecule (IUPAC name): (2R)-2-amino-3-methylsulfanylpropanoic acid

The molecule provided is a chiral amino acid that contains several identifiable substructures.

1. **Amino Group**: This functional group consists of a nitrogen atom bonded to two hydrogen atoms (-NH2). It is attached to the second carbon atom of the main carbon chain...
2. **Carboxylic Acid Group**: Represented by (-COOH), … It is attached to the end of the carbon chain, characteristic of amino acids which bear a carboxylic acid group at the terminal carbon.
3. **Main Carbon Chain**: The molecule has a three-carbon chain serving as the backbone. The central carbon, referred to as the alpha carbon, is a chiral center, meaning it is bonded to four different substituents: the amino group, the carboxylic acid group, a hydrogen atom, and a side chain.
4. **Methylsulfanyl Group**: Attached to the third carbon is a side chain consisting of a sulfur atom single-bonded to a methyl group (-SCH3). This group acts as the side chain substituent and characterizes the molecule as containing a thioether linkage, ...
5. **Chirality**: The molecule is chiral, with the alpha carbon being the stereocenter. Its configuration is designated as (R), indicating the spatial arrangement of its substituents …

Overall, this molecule is a sulfur-containing, optically active alpha amino acid …

Data Type: Structure-to-Feature Relationship

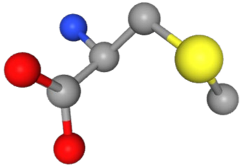
To further learn the chemical and biological features, we design S2F relationship explanations, based on our observation of the hierarchical relationship.

Instruction to GPT-4o

Explain chemical properties by relating chemical properties to its structural information.

Input Molecule (IUPAC name): (2R)-2-amino-3-methylsulfanylpropanoic acid

Description: S-methylcysteine is a **cysteine derivative** that is L-cysteine in which the hydrogen attached to the sulfur is replaced by a methyl group. It has a role as a **human urinary metabolite** and a **plant metabolite**. It is a tautomer of a S-methylcysteine **zwitterion**.



The given molecule is a derivative of cysteine, an amino acid known for containing a thiol group. In the given molecule, this thiol group has one of its hydrogen atoms replaced by a methyl group, making it a methylthioether. ...

Firstly, the presence of the amino group (-NH2) classifies this compound as an amino acid, which contributes to its ability to act as a zwitterion. ...

The sulfur atom with a methyl attachment forms a thioether. This increases the molecule's hydrophobic characteristics compared to cysteine, which contains a more polar thiol group. This methylsulfanyl group can also affect the molecule's reactivity and interaction with biochemical systems, ...

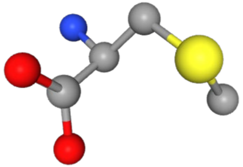
As a human urinary metabolite, the presence of this compound in urine suggests its role in human metabolism where it might play a role in sulfur metabolism or act as a detoxification product. In plants, it could partake in sulfur metabolism and possibly aid in detoxification or defense processes due to environmental stressors.

In summary, the structural modification from cysteine to include a methylthioether enhances its lipophilic properties

Data Type: Comprehensive Conversations

To tailor LLMs to handle diverse requests, we design comprehensive conversations that gradually deepens from structure to chemistry and biology.

Instruction to GPT -4o



Design a sequence of pairs of questions and answers that gradually deepen the level of the conversation, from structural information and chemical properties to biological functionalities.

Input Molecule (IUPAC name): (2R)-2-amino-3-methylsulfanylpropanoic acid

Description: S-methylcysteine is a cysteine derivative that is L-cysteine in which the hydrogen attached to the sulfur is replaced by a methyl group....

User: Can you tell me about the basic structure of the given molecule?
Assistant: The given molecule is a derivative of cysteine, where the hydrogen atom attached to the sulfur group is replaced by a methyl group. ...

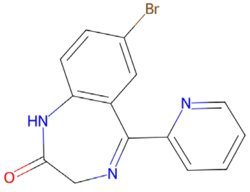
User: Can you tell me more about the specific species of this compound?
Assistant: The compound is known as S-methylcysteine. It is a specific derivative of cysteine with a distinctive methyl substitution on the sulfur atom, differing from the typical cysteine structure.

User: How does the molecule's structure influence its chemical behavior?
Assistant: The structure of the given molecule, particularly the presence of both amino and carboxylic acid groups, contributes to its zwitterionic nature at physiological pH. ...

User: What role does this molecule play biologically?
Assistant: Biologically, S-methylcysteine serves as a metabolite in both humans and plants. ...

Qualitative Evaluation

Mol-LLaMA is able to accurately understand the molecular structures and predict the chemical and biological properties, providing detailed explanation.



Input Molecule

Annotated Properties (Ground Truth)

It is one of the benzodiazepines that is used in the treatment of anxiety disorders Bromazepam binds to the GABA receptor GABA$_A$, causing a conformational change and increasing inhibitory effects of GABA. Other neurotransmitters are not influenced.

User

Explain the features of the given molecule that are most distinct from those of other molecules.

Mol-LLaMA

The given molecule is a member of the benzodiazepine class, which is characterized by a fusion of a benzene ring and a seven-membered diazepine ring. This structure is crucial for its biological activity, particularly its role as a sedative and hypnotic agent.

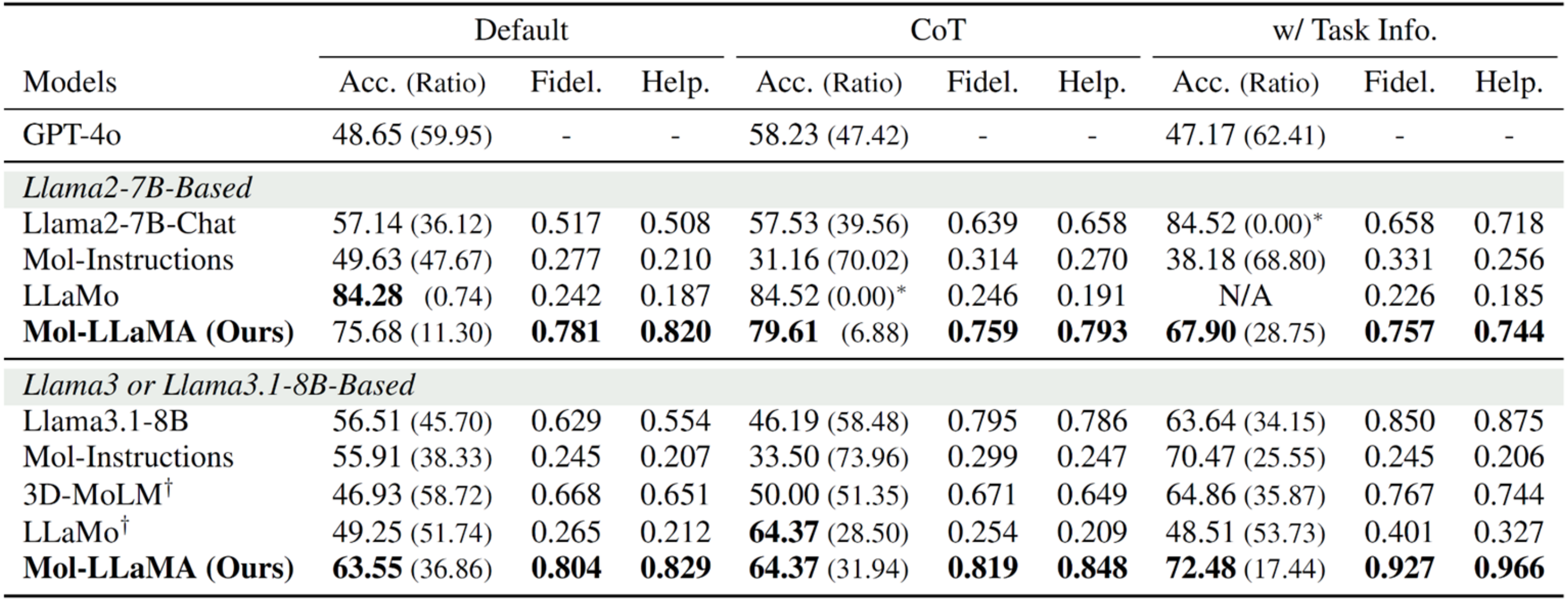
One of the key features of the given molecule is the presence of a pyridine ring attached to the benzodiazepine core. This pyridine moiety is significant because it can influence the molecule's pharmacokinetic properties, such as its ability to cross the blood-brain barrier, which is essential for its function as a central nervous system depressant.

Additionally, the molecule contains a bromine atom attached to the benzodiazepine ring. The presence of this halogen atom can affect the molecule's lipophilicity and its binding affinity to the GABA-A receptor, which is the primary target of benzodiazepines. The bromine atom may also contribute to the molecule's metabolic stability and its duration of action.

The given molecule is also characterized by a ketone group within the diazepine ring, which is a common feature in benzodiazepines. This functional group is important for the interaction with the GABA-A receptor, as it plays a role in the formation of the benzodiazepine-receptor complex, enhancing the inhibitory effects of GABA. …

Molecular Property Prediction

Mol-LLaMA shows the best performance not only on the accuracy but also on the fidelity and helpfulness, implying its utility as a general-purpose assistant.



	Default			CoT			w/ Task Info.		
Models	Acc. (Ratio)	Fidel.	Help.	Acc. (Ratio)	Fidel.	Help.	Acc. (Ratio)	Fidel.	Help.
GPT-4o	48.65 (59.95)	-	-	58.23 (47.42)	-	-	47.17 (62.41)	-	-
Llama2-7B-Based									
Llama2-7B-Chat	57.14 (36.12)	0.517	0.508	57.53 (39.56)	0.639	0.658	84.52 (0.00)*	0.658	0.718
Mol-Instructions	49.63 (47.67)	0.277	0.210	31.16 (70.02)	0.314	0.270	38.18 (68.80)	0.331	0.256
LLaMo	**84.28** (0.74)	0.242	0.187	84.52 (0.00)*	0.246	0.191	N/A	0.226	0.185
Mol-LLaMA (Ours)	75.68 (11.30)	**0.781**	**0.820**	**79.61** (6.88)	**0.759**	**0.793**	**67.90** (28.75)	**0.757**	**0.744**
Llama3 or Llama3.1-8B-Based									
Llama3.1-8B	56.51 (45.70)	0.629	0.554	46.19 (58.48)	0.795	0.786	63.64 (34.15)	0.850	0.875
Mol-Instructions	55.91 (38.33)	0.245	0.207	33.50 (73.96)	0.299	0.247	70.47 (25.55)	0.245	0.206
3D-MoLM†	46.93 (58.72)	0.668	0.651	50.00 (51.35)	0.671	0.649	64.86 (35.87)	0.767	0.744
LLaMo†	49.25 (51.74)	0.265	0.212	**64.37** (28.50)	0.254	0.209	48.51 (53.73)	0.401	0.327
Mol-LLaMA (Ours)	**63.55** (36.86)	**0.804**	**0.829**	**64.37** (31.94)	**0.819**	**0.848**	**72.48** (17.44)	**0.927**	**0.966**

Conclusion

- We present Mol-LLaMA, a large molecular language model with a general understanding of molecular features and reasoning capabilities.

- We design a model architecture that fully leverages the complementary advantages of different molecular representations.

- We construct a large instruction dataset by devising three data types, to cover the fundamental features from structural features to chemical and biological features and cultivate the explainability and reasoning capabilities.

Thank you

Dongki Kim

✉ cleverki@kaist.ac.kr

@DongkiKim95

https://dongkikim95.github.io/

https://mol-llama.github.io