LlaSMol: Advancing LLMs for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset

Botao Yu, Frazier N. Baker*, Ziqi Chen*, Xia Ning, Huan Sun (* equal contribution)
The Ohio State University
{yu.3737, baker.3239, chen.8484, ning.104, sun.397}@osu.edu

THE OHIO STATE UNIVERSITY

SCAN ME

Awesome dataset, SoTA LLMs for chemistry tasks, and more insights!

Introduction

Problems

- While LLMs (e.g., GPT-4) show remarkable capabilities on NLP, their performance of existing on chemistry tasks is discouragingly low.
- Existing deep learning models for chemistry tasks are usually task-specific models, which neglect shared chemistry knowledge across tasks and can hardly be adapted to different tasks.

Our Solution

- We construct a large-scale, comprehensive, and high-quality dataset, **SMolInstruct**, for instruction tuning and evaluation. It has 14 tasks illustrated in the figure below.

🤗 Dataset: osunlp/SMolInstruct

Name Conversion

IUPAC to Molecular Formula (NC-I2F)
Query: What is the molecular formula of the compound with this IUPAC name <IUPAC> 2,5-diphenyl-1,3-oxazole </IUPAC> ?
Response: <MOLFORMULA> C15H11NO </MOLFORMULA>

IUPAC to SMILES (NC-I2S)
Query: Could you provide the SMILES for <IUPAC> 4-ethyl-4-methyloxolan-2-one </IUPAC> ?
Response: Of course. It's <SMILES> CCC1(C)COC(=O)C1 </SMILES>

SMILES to Molecular Formula (NC-S2F)
Query: Given the SMILES representation <SMILES> S=P1(N(CCCl)CCCl)NCCCO1 </SMILES>, what would be its molecular formula?
Response: It is <MOLFORMULA> C7H15Cl2N2OPS </MOLFORMULA> .

SMILES to IUPAC (NC-S2I)
Query: Translate the given SMILES formula of a molecule <SMILES> CCC(C)C1CNCCCNC1 </SMILES> into its IUPAC name.
Response: <IUPAC> 3-butan-2-yl-1,5-diazocane </IUPAC>

Molecule Description

Molecule Captioning (MC)
Query: Describe this molecule: <SMILES> CCOC(=O)C1=CN=CN1[C@H](C)C1=CC=CC=C1 </SMILES>
Response: The molecule is an imidazole derivative with short-acting sedative, hypnotic, and general anesthetic properties. Etomidate appears to have gamma-aminobutyric acid (GABA) like effects, mediated through GABA-A receptor. The action enhances the inhibitory effect of GABA on the central nervous system by causing chloride channel opening events which leads to membrane hyperpolarization.

Molecule Generation (MG)
Query: Give me a molecule that satisfies the conditions outlined in the description: The molecule is a member of the class of tripyrroles that is a red-coloured pigment with antibiotic properties produced by Serratia marcescens. It has a role as an antimicrobial agent, a biological pigment, a bacterial metabolite, an apoptosis inducer and an antineoplastic agent. It is a tripyrrole, an aromatic ether and a ring assembly.
Response: Here is a potential molecule: <SMILES> CCCCCC1=C(C)NC(/C=C2\N=C(C3=CC=CN3)C=C2OC)=C1 </SMILES>

Property Prediction

ESOL (PP-ESOL)
Query: How soluble is <SMILES> CC(C)Cl </SMILES> ?
Response: Its log solubility is <NUMBER> -1.41 </NUMBER> mol/L.

LIPO (PP-LIPO)
Query: Predict the octanol/water distribution coefficient logD under the circumstance of pH 7.4 for <SMILES> NC(=O)C1=CC=CC=C1O </SMILES> .
Response: <NUMBER> 1.090 </NUMBER>

BBBP (PP-BBBP)
Query: Is blood-brain barrier permeability (BBBP) a property of <SMILES> CCNC(=O)/C=C/C1=CC=CC(Br)=C1 </SMILES>?
Response: <BOOLEAN> Yes </BOOLEAN>

ClinTox (PP-ClinTox)
Query: Is <SMILES> COC[C@@H](NC(C)=O)C(=O)NCC1=CC=CC=C1 </SMILES> toxic?
Response: <BOOLEAN> No </BOOLEAN>

HIV (PP-HIV)
Query: Can <SMILES> CC1=CN(C2C=CCCC2O)C(=O)NC1=O </SMILES> serve as an inhibitor of HIV replication?
Response: <BOOLEAN> No </BOOLEAN>

SIDER (PP-SIDER)
Query: Are there any known side effects of <SMILES> CC1=CC(C)=C(NC(=O)CN(CC(=O)O)CC(=O)O)C(C)=C1Br </SMILES> affecting the heart?
Response: <BOOLEAN> No </BOOLEAN>

Chemical Reaction

Forward Synthesis (FS)
Query: <SMILES> NC1=CC=C2OCOC2=C1.O=CO </SMILES>
Based on the reactants and reagents given above, suggest a possible product.
Response: A possible product can be <SMILES> O=CNC1=CC=C2OCOC2=C1 </SMILES>

Retrosynthesis (RS)
Query: Identify possible reactants that could have been used to create the specified product. <SMILES> CC1=CC=C(N)N=C1N </SMILES>
Response: <SMILES> CC(C#N)CCC#N.N </SMILES>

▲ 4 types of 14 tasks in the proposed SMolInstruct dataset, determined with advice from domain experts for their real-world practicality and benefits for building strong foundation models.

- We build a series of LLMs for chemistry named **LlaSMol** by fine-tuning four open-source LLMs with LoRA.

Galactica | Llama 2 | Code Llama | Mistral

🤗 Our best model: osunlp/LlaSMol-Mistral-7B

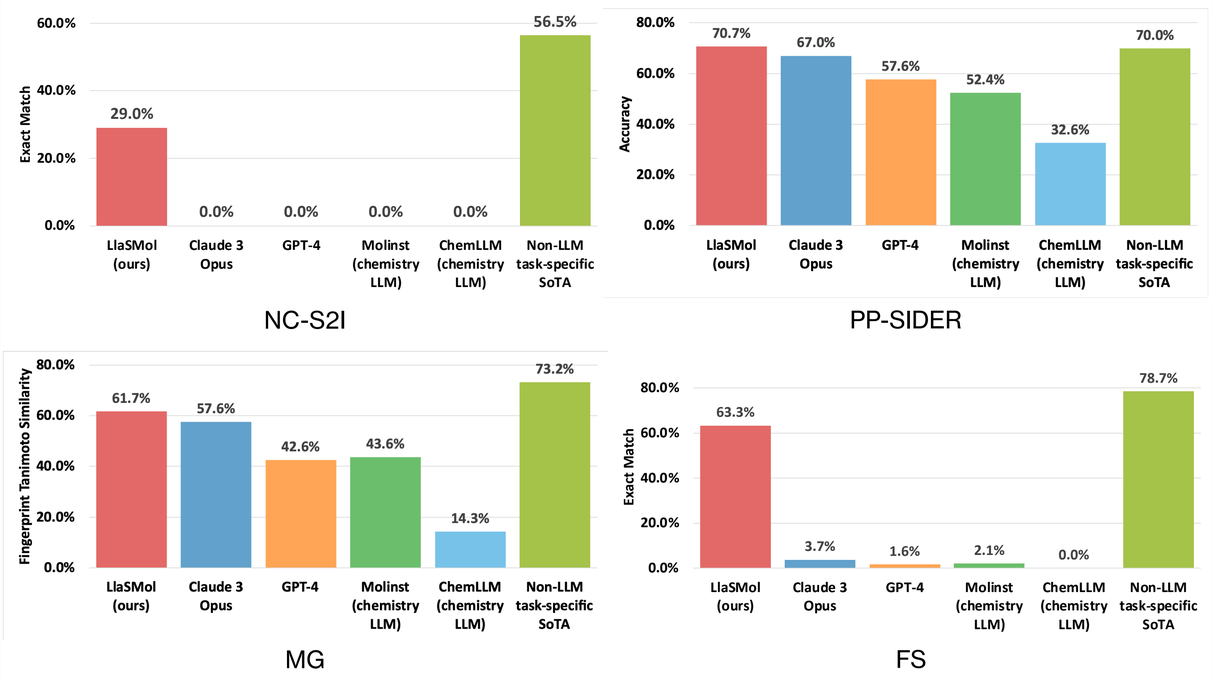
- We conduct comprehensive experiments and provide insights.

Dataset Construction

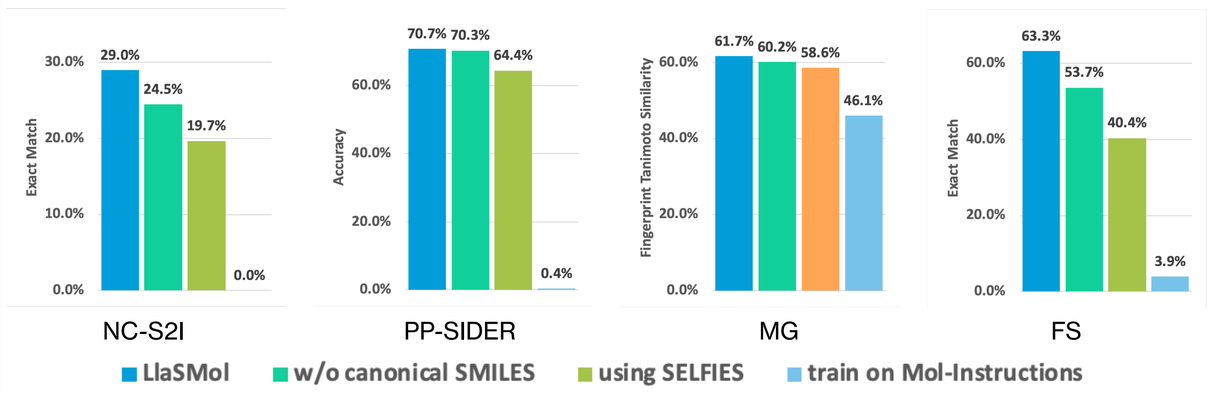
- **Data Collection**: Collect the original data from multiple large-scale sources such as PubChem, USPTO-full, etc.
- **Quality Control**: Apply rigorous scrutiny to remove 1) chemically invalid SMILES, 2) wrong or inaccurate information, and 3) duplicated samples.
- **Data Splitting**: It requires careful handling for multi-task datasets to avoid data leakage across tasks and to compare with other datasets. We carefully ensure:
  - Related samples for related tasks (like FS and RS) are placed in the same split.
  - Samples with identical input (and different outputs) are placed in the same split to avoid biased evaluation.
  - For fair comparison, the split is compatible with existing datasets.
- **Instruction Creation**: Manually craft several templates and apply GPT-4 to rephrase. Use special tags to encapsulate core information to inform models and facilitate answer extraction, such as:

`<SMILES> ... </SMILES>`, `<IUPAC> ... </IUPAC>`.

Results



▲ Partial results of overall comparison with Claude 3 Opus, GPT-4, Molinst, ChemLLM, and non-LLM task-specific SoTA models. Full results can be found in the paper.



▲ Partial results of ablation study. Full results can be found in the paper.

Task	Metric	Single-Task	Multi-Task	Improv.
NC-I2F	EM (%)	86.8	87.9	1.1
NC-I2S	EM (%)	67.6	70.1	2.4
NC-S2F	EM (%)	93.2	93.2	0.0
NC-S2I	EM (%)	27.4	29.0	1.5
PP-ESOL	RMSE↓	20.616	1.150	19.466
PP-Lipo	RMSE↓	1.241	1.010	0.231
PP-BBBP	Acc (%)	68.5	74.6	6.1
PP-Clintox	Acc (%)	79.9	93.1	13.2
PP-HIV	Acc (%)	96.7	96.7	0.0
PP-SIDER	Acc (%)	64.3	70.7	6.4
MC	METEOR	0.299	0.452	0.153
MG	FTS (%)	33.1	61.7	28.6
FS	EM (%)	62.6	63.3	0.7
RS	EM (%)	31.5	32.9	1.4

▲ Results of single-task vs multi-task training. Orange cells represent better positive improvements.

Model	NC-I2F EM (%)	NC-I2S EM (%)	NC-S2F EM (%)	NC-S2I EM (%)	PP-ESOL RMSE	PP-Lipo RMSE	PP-BBBP Acc (%)	PP-Clintox Acc (%)	PP-HIV Acc (%)	PP-SIDER Acc (%)	MC METEOR	MG FTS (%)	FS EM (%)	RS EM (%)
w/o NC	-	-	-	-	1.520	1.090	76.1	93.1	96.8	70.6	0.436	54.9	63.2	33.5
w/o PP	87.9	70.7	93.5	28.7	-	-	-	-	-	-	0.447	62.3	64.2	33.1
w/o MC	87.6	71.0	93.5	27.8	1.133	1.057	74.1	93.1	96.8	70.9	-	64.1	63.3	33.4
w/o MG	87.8	69.6	93.4	27.8	1.231	0.982	77.2	93.1	96.8	70.9	0.445	-	63.4	34.0
w/o FS	87.9	70.4	93.8	29.5	1.278	1.288	70.6	93.1	96.8	70.8	0.452	63.2	-	33.1
w/o RS	88.0	71.1	93.7	29.7	1.203	1.048	72.1	93.1	96.8	70.6	0.450	62.6	61.9	-
$\text{LlaSMol}_{\text{Llama 2}}$	87.9	70.1	93.2	29.0	1.150	1.010	74.6	93.1	96.7	70.7	0.452	61.7	63.3	32.9

▲ Results of removing certain tasks. Orange cells represent better results than $\text{LlaSMol}_{\text{Mistral}}$ while blue cells represent worse results.

Takeaways

- **LlaSMol models fine-tuned on our SMolInstruct achieve SoTA among LLMs.**
- Regarding molecular representations, Canonicalizing SMILES helps, and SMILES is generally better than SELFIES.
- Multi-task training is better than single-task training overall.
- Removing one certain task from the training set does not consistently influence the performance on other tasks, showing a degree of independence among the tasks.

Acknowledgement: The authors would thank colleagues from the OSU NLP group and the OSU Ning Lab for constructive feedback. This research was supported in part by NSF IIS-2133650, NIH 1R01LM014385-01, and NSF CAREER #1942980, as well as Ohio Supercomputer Center (Ohio Supercomputer Center, 1987). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.