Lang + Mol 2024

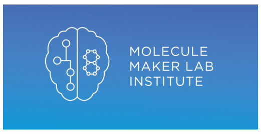
The 1st Workshop on Language + Molecules

Proceedings of the Workshop

August 15, 2024

The Lang + Mol organizers gratefully acknowledge the support from the following sponsors.

Gold



Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-148-3

Introduction

Welcome to Language + Molecules, the inaugural workshop on integrating natural language with molecular structure! The workshop is scheduled to be held on August 15, 2024 in Bangkok, Thailand co-located with ACL 2024.

The world faces an enormous number of problems in the coming decades on scales of complexity never-before-seen, in areas such as climate change, healthcare, and pandemics. To address these issues, we need to discover inventive scientific solutions which are scalable, flexible, and inexpensive. Broadly speaking, many of these problems will require molecular solutions from the chemistry domain, such as developing new drugs, materials, and chemical processes. These solutions exist in extremely large search spaces, which makes AI tools a necessity. Excitingly, the chemistry field is posed to be substantially accelerated via multimodal models combining language with molecules and drug structures. Research in scientific NLP, integrating molecules with natural language, and multimodal AI for science/medicine has experienced significant attention and growth in recent months. This workshop was organized to help connect researchers working in this exciting nascent community.

A natural question to ask is why we want to integrate natural language with molecules. Combining these types of information has the possibility to accelerate scientific discovery: imagine a future where a doctor can write a few sentences describing a patient's symptoms and then receive the exact structure of the drugs necessary to treat that patient's ailment (taking into account the patient's genotype, phenotype, and medical history). Or, imagine a world where a researcher can specify the function they want a molecule to perform (e.g., antimalarial or a photovoltaic) rather than its low level properties (e.g., pyridine-containing). This high-level control of molecules requires a method of abstract description, and humans have already developed one for communication: language. The following key benefits of combining language and molecules were explored:

1. Generative Modeling: One of the largest problems in current LLMs—hallucination— becomes a strength for discovering molecules with high-level functions, abstract properties, and composition of many properties.

2. Bridging Modalities: Language can serve as a "bridge" between modalities (e.g., cellular pathways and drugs) when data is scarce.

3. Domain Understanding: Grounding language models into external real world knowledge can improve understanding of unseen molecules and advance many emerging tasks, such as experimental procedure planning and reasoning, which use LLMs as scientific agents.

4. Automation: Instruction-following, dialogue-capable, and tool-equipped models can guide automated discovery in silico and in robotic labs.

5. Democratization: Language enables scientists without computational expertise to leverage advances in scientific AI.

In particular, this year's workshop focused on the following themes:

- Going beyond language to incorporate molecular structure and interactions into LLMs.

- Addressing data scarcity and inconsistency: new training methodologies and methods for extracting data from scientific literature.

- Language-enabled solutions for discovering new drugs and molecules.

- Incorporating domain knowledge from human-constructed databases into LLMs.

- Instruction-following, dialogue-capable, and tool-equipped LLMs for molecules.
- Sequence representations for molecular structures, including organic molecules, proteins, DNA, and inorganic crystals.

The workshop had 27 total submissions, from which 11 papers and 7 shared task descriptions were accepted. Between these categories, 14 accepted submissions opted to be included in the archival proceedings. The shared task had two tracks: molecule generation and molecule captioning. For captioning, there were 28 participants who had a combined total of 188 submissions. For molecule generation, there were 19 participants and 88 submissions. Submissions achieved improvements over base models of up to 27% absolute metric increase for molecule captioning and 13 absolute for molecule generation. An overview of the shared task and submission results will be given at the workshop. This will include the release of the ensembled captioning results for the "mystery molecules".

The workshop will have 5 keynote speakers: Kyunghyun Cho, Elsa Olivetti, Marinka Zitnik, Huan Sun, and Lei Li. Additionally, a poster session, invited oral talks, and a panel discussion on future research directions will be held.

As a final note, we would like to thank the authors, invited speakers, committee members, and our scientific advisory board for helping make this workshop happen. We would like to thank the NSF Molecule Maker Lab Institute for supporting this initiative.

Organizing Committee

Program Chairs

Carl Edwards, University of Illinois Urbana-Champaign
Qingyun Wang, University of Illinois Urbana-Champaign
Manling Li, Northwestern University
Lawrence Zhao, Yale University
Tom Hope, Hebrew University of Jerusalem and Allen Institute for AI
Heng Ji, University of Illinois Urbana-Champaign

Program Committee

Area Chairs

Carl Edwards, University of Illinois Urbana-Champaign
Qingyun Wang, University of Illinois Urbana Champaign
Lijun Wu, ByteDance
Yaochen Xie, Amazon

Reviewers

Ralph Abboud
Chufan Gao
Giorgio Giannone
Zhihui Guo
Chi Han
Anna Hart
Jeonghwan Kim
Tuan Lai
Weijiang Li
Xuan Liu
Zequn Liu
Ziteng Liu
Ziqian Luo
Thao Nguyen
Siru Ouyang
Qizhi Pei
Brandon Philip Theodorou
Ziqi Wang
Azmine Toushik Wasi
Pengfei Yu

Invited Speakers

Kyunghyun Cho, New York University and Genentech
Elsa Olivetti, Massachusetts Institute of Technology
Marinka Zitnik, Harvard
Huan Sun, The Ohio State University
Lei Li, Carnegie Mellon University

Table of Contents

Program

Thursday, August 15, 2024

09:00 - 09:10	*Opening Remarks*
09:10 - 09:40	*Invited Talk 1*
09:40 - 10:10	*Invited Talk 2*
10:10 - 10:40	*Invited Talk 3*
10:40 - 11:00	*Coffee Break*
11:00 - 11:20	*Shared Task Overview*
11:30 - 12:30	*Oral Presentations*
12:30 - 13:30	*Lunch Break*
13:30 - 14:05	*Invited Talk 4*
14:05 - 14:40	*Invited Talk 5*
14:40 - 15:30	*Panel Discussion*
15:30 - 16:00	*Coffee Break*
16:00 - 17:15	*Poster Session*
17:15 - 17:30	*Closing Remarks*