ICLR

Molecular Generation with Chemical Feedback

https://openreview.net/forum?id=9rPyHyjfwP

Yin Fang, Ningyu Zhang†, Zhuo Chen, Lingbing Guo,

Xiaohui Fan, Huajun Chen†

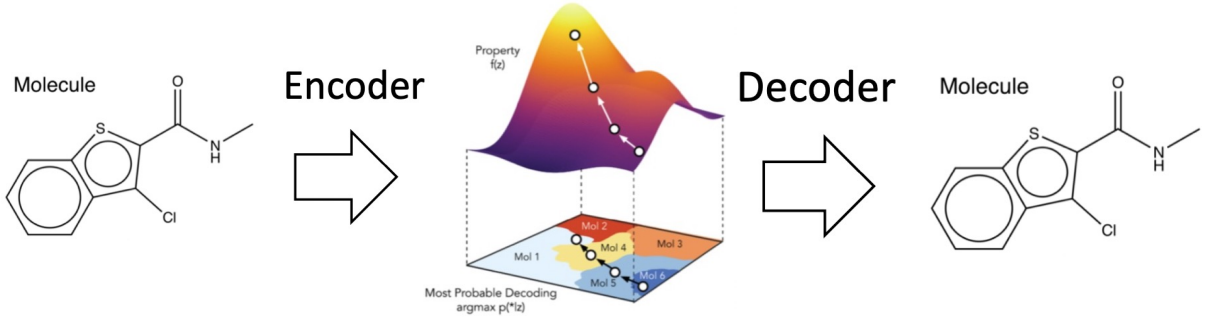
CONTENT

CONTENT

Molecule Generation

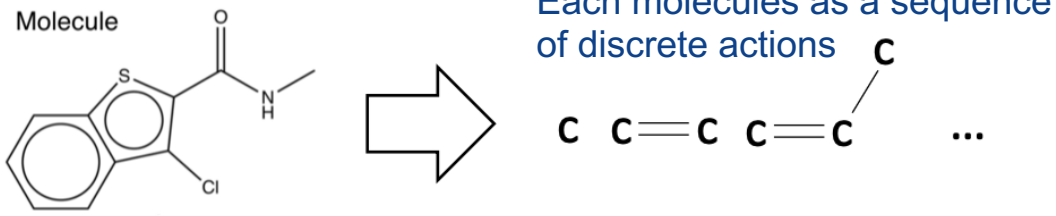
❑ Molecule Generation: Finding novel molecular structures with desired properties

❑ Search in ***continuous*** hidden space



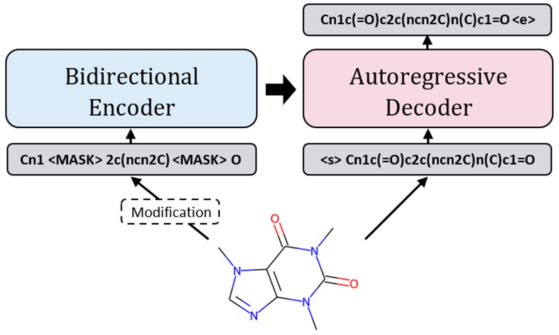
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules (2018)

❑ Search in ***discrete*** chemical space



Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation (2019)

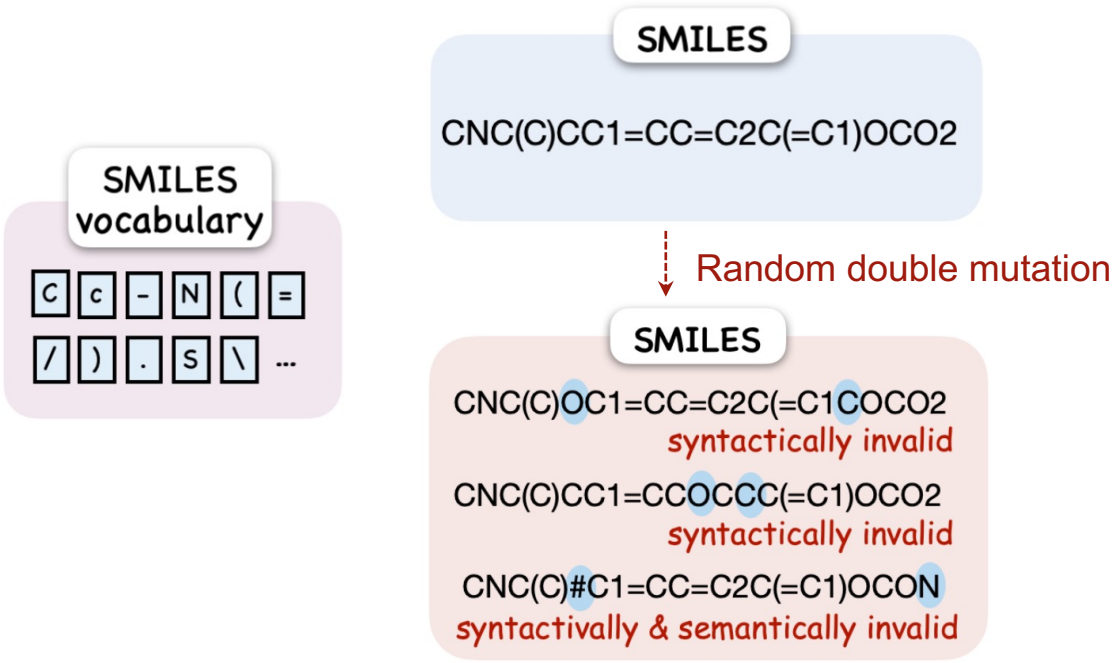
❑ Search in ***molecular language*** space



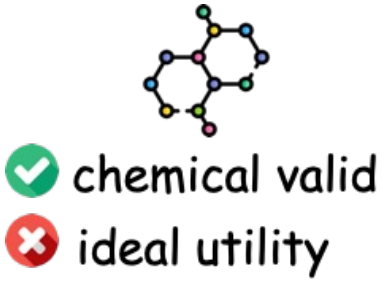
Chemformer: a pre-trained transformer for computational chemistry (2022)

Challenges in Molecular Language Models

- SMILES-based language models have a certain probability of producing ***invalid*** molecules

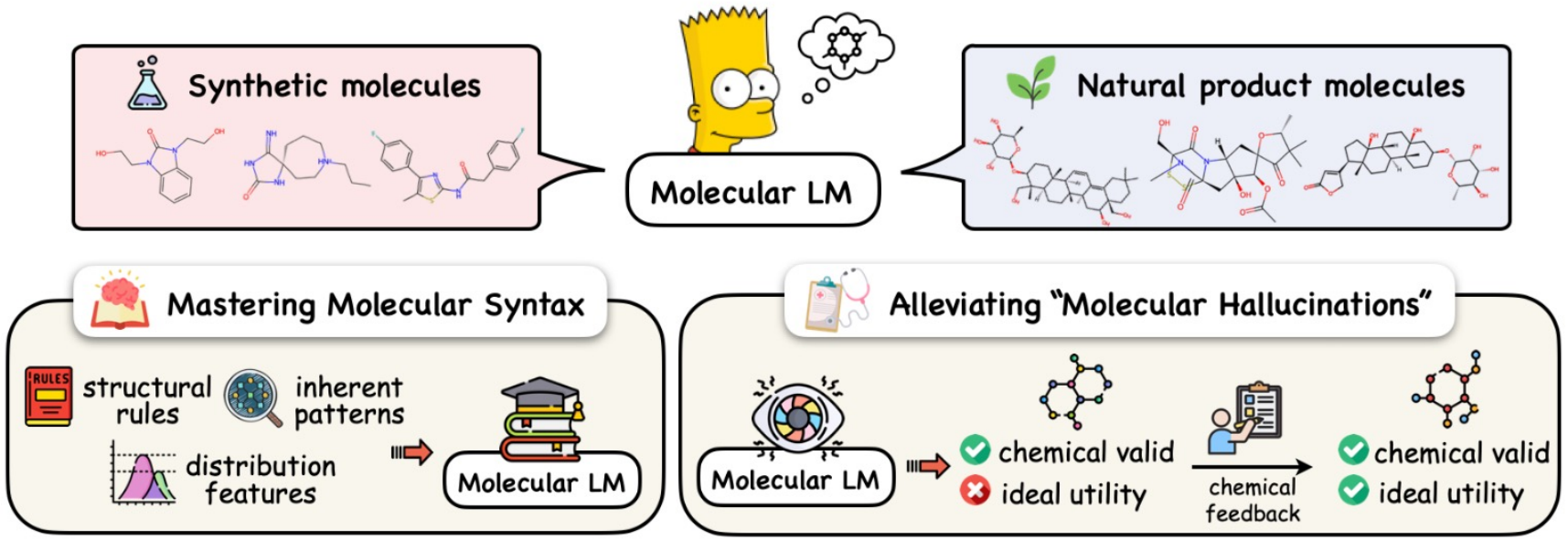


- Molecular language models often suffer from "***molecular hallucinations***"



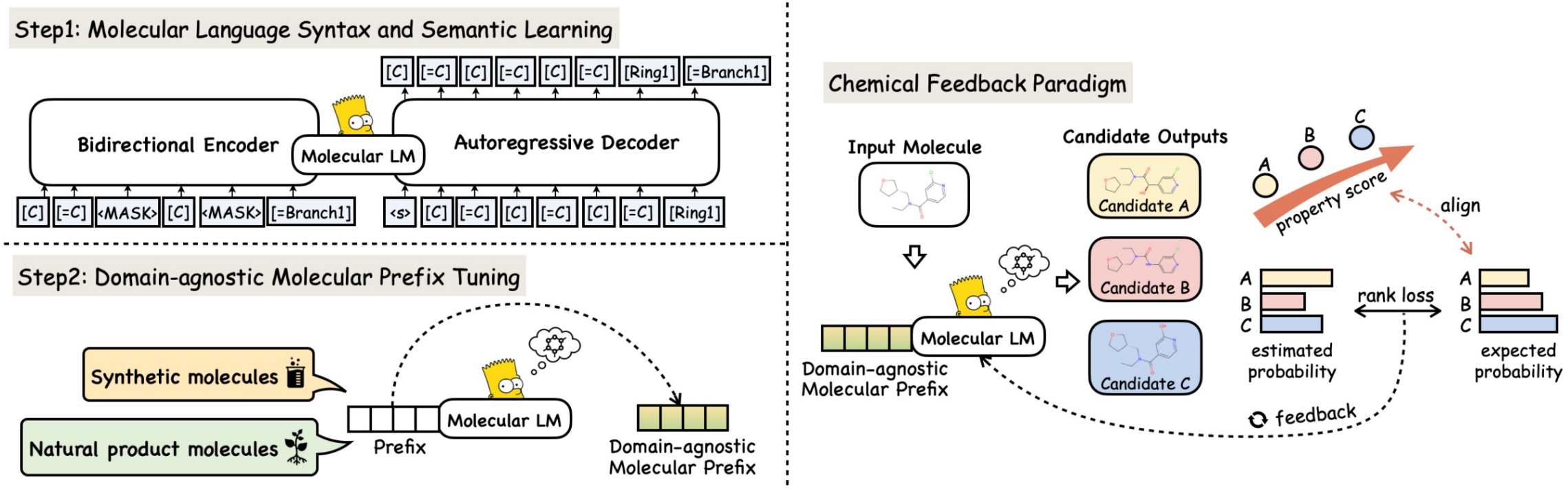
Proposal

Aligning pre-trained molecular language model with chemical preferences



CONTENT

MolGen's Framework



Domain-agnostic Molecular Pre-training:

- *Stage 1:* Understand the molecular structure, grammar, and intrinsic semantics.
- *Stage 2:* Harness knowledge transferable across diverse domains.

Self-feedback Paradigm - align PLM with chemical preference:

- Align the probabilistic rankings with chemical preference rankings.
- Learn to evaluate and rectify its molecular outputs.

CONTENT

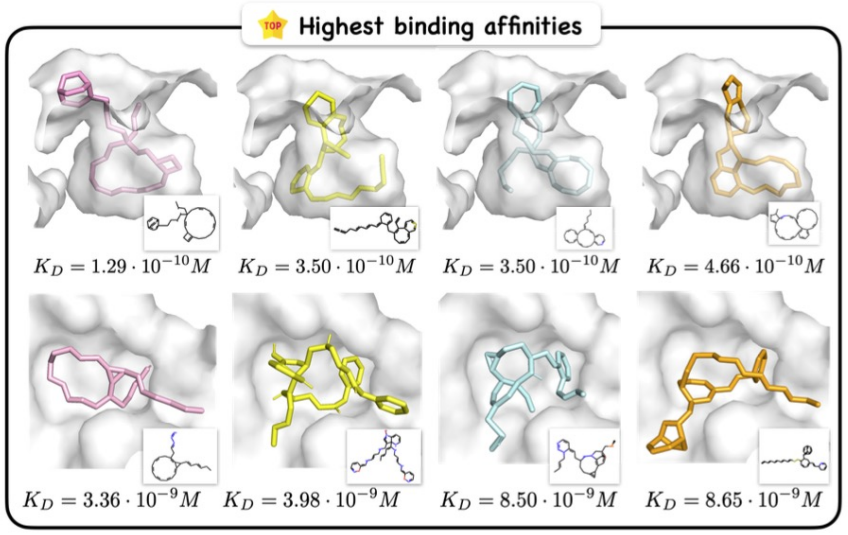
Molecular Distribution Learning

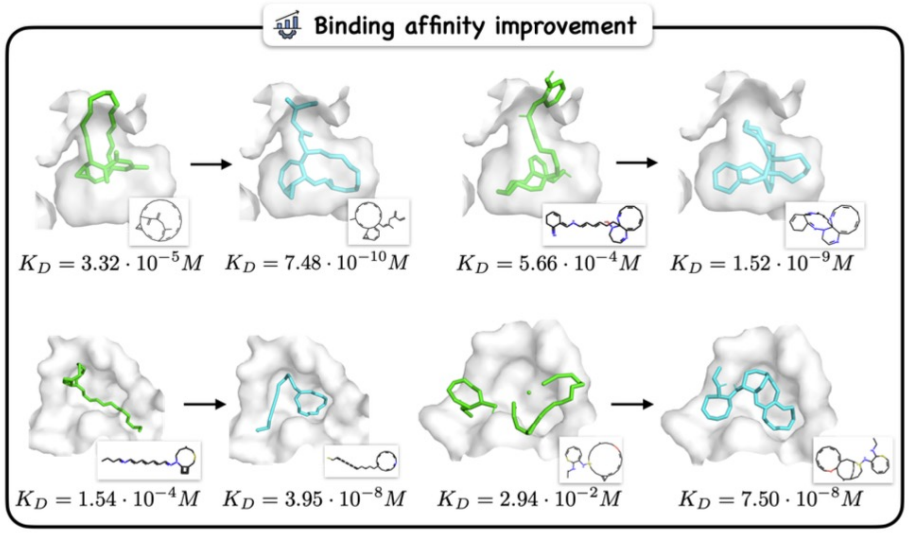
Reflects real-world molecular distributions

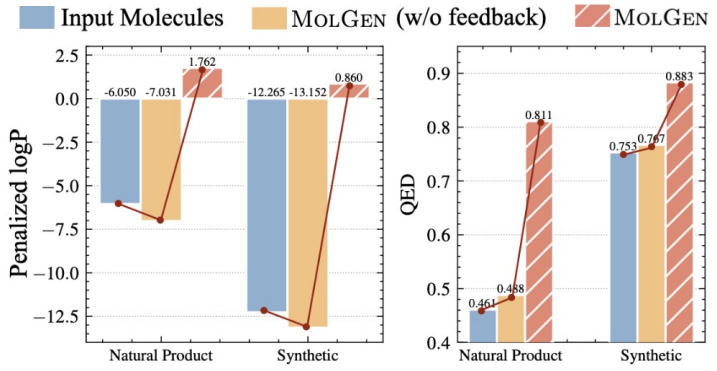
Model	Synthetic Molecules							Natural Product Molecules						
	Validity↑	Frag↑	Scaf↑	SNN↑	IntDiv↑	FCD↓	Novelty↑	Validity↑	Frag↑	Scaf ↑	SNN↑	IntDiv↑	FCD↓	Novelty↑
AAE	.9368	.9910	.9022	.6081	.8557	.5555	.7931	.0082	.9687	.2638	.3680	.8704	4.109	.9943
LatentGAN	.8966	.9986	.8867	.5132	.8565	.2968	.9498	.9225	.2771	.0884	.5321	.6009	45.53	.9949
CharRNN	.9748	.9998	.9242	.6015	.8562	.0732	.8419	.7351	.8816	.5212	.4179	.8756	2.212	.9792
VAE	.9767	.9994	.9386	.6257	.8558	.0990	.6949	.2627	.8840	.4563	.3950	.8719	4.318	.9912
JT-VAE	**1.000**	.9965	.8964	.5477	.8551	.3954	.9143	**1.000**	.8798	.5012	.3748	.8743	12.03	.9957
LIMO	**1.000**	.9562	.1073	.6125	.8544	.1532	.8956	**1.000**	.7242	.0005	.3416	.7726	31.84	.9962
Chemformer	.9843	.9889	.9248	.5622	.8553	.0061	.9581	.9825	.9826	.4126	.5875	.8650	.8346	.9947
MolGen	**1.000**	**.9999**	**.9999**	**.9996**	**.8567**	**.0015**	**1.000**	**1.000**	**.9994**	**.8404**	**.8148**	**.8878**	**.6519**	**.9987**

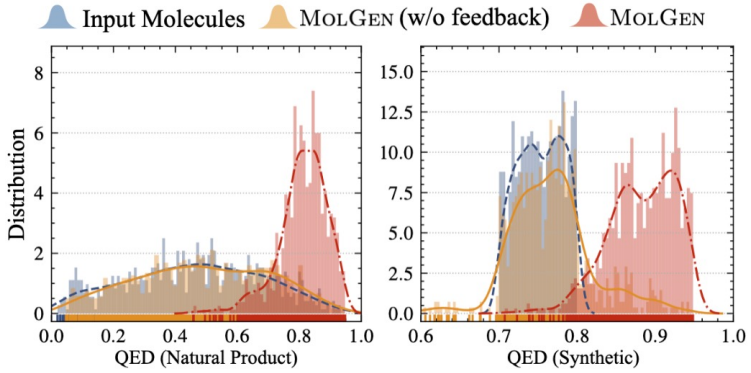
Molecular Optimization

Alleviate molecular hallucinations



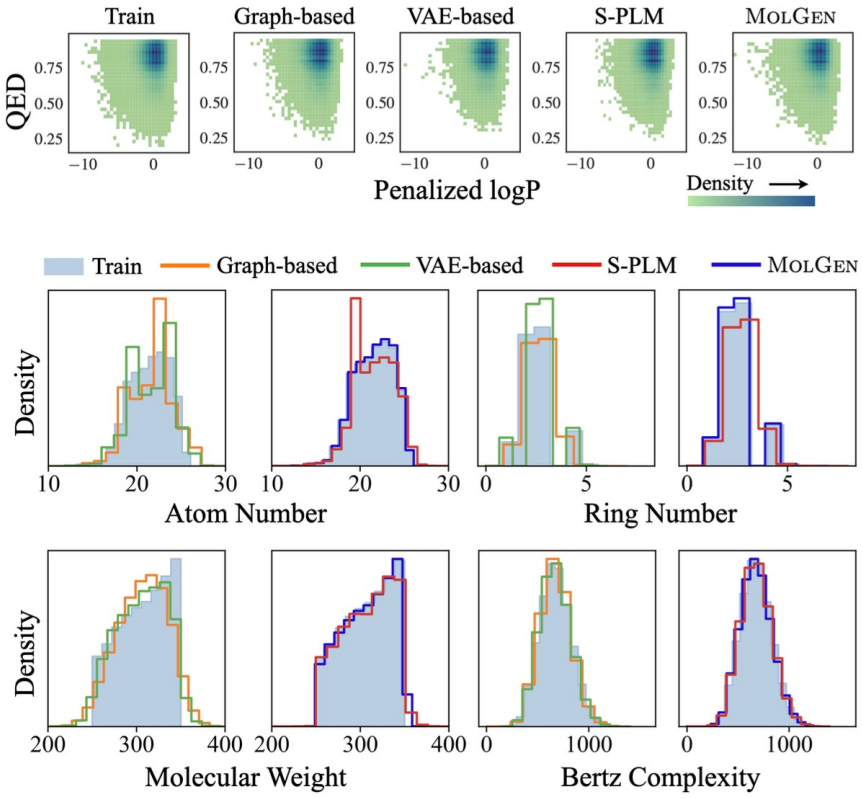




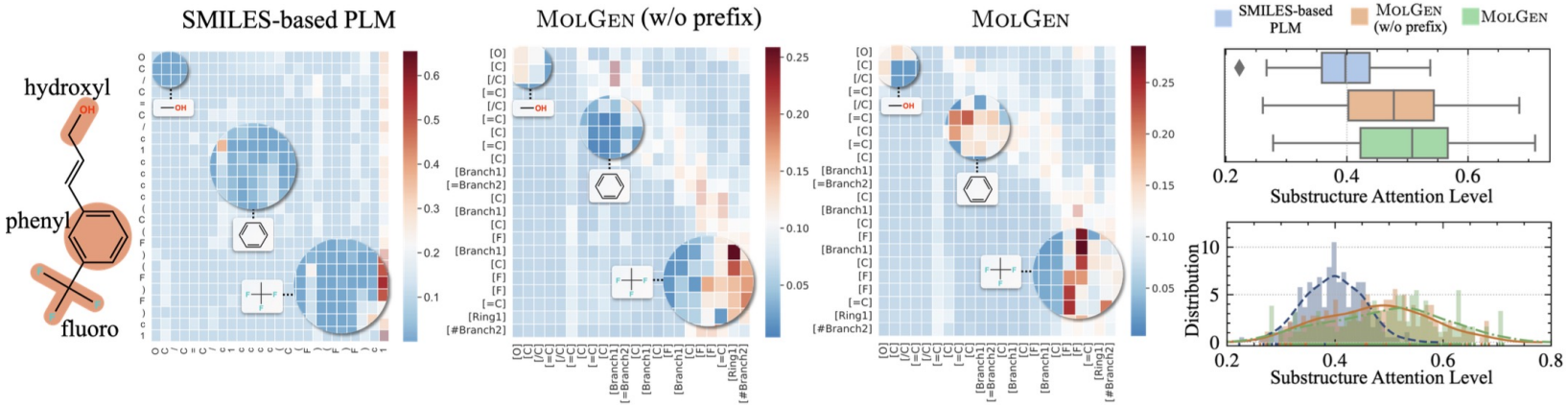


Analysis

Captures molecular characteristics



Recognizes meaningful substructures



CONTENT

Take Away

- This study proposes a pre-trained molecular language model tailored for molecule generation:
  - generating valid molecules while avoiding “molecular hallucinations”
  - identifying essential molecular substructures

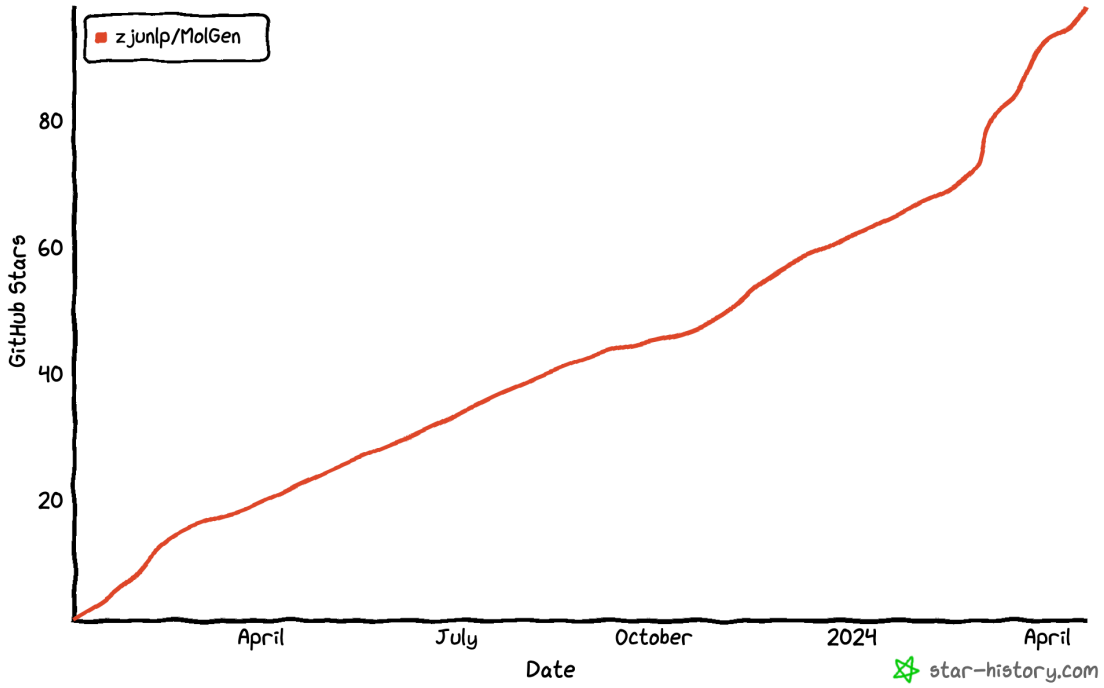
Future Work

- Apply to other tasks such as retrosynthesis and reaction prediction
- Explore multimodal pre-training
- Incorporate additional sources of knowledge

Open Source

Github

github.com/zjunlp/MolGen



Hugging Face

zjunlp/MolGen-large

Total downloads

12,632 (all time, tracked internally since January 2021)

zjunlp/MolGen-large-opt

Total downloads

1,726 (all time, tracked internally since January 2021)

zjunlp/MolGen-7B

Total downloads

1,568 (all time, tracked internally since January 2021)

Thank you!

Code

Model

浙江大学 ZHEJIANG UNIVERSITY