32nd ACM International Conference on Information and Knowledge Management
October 21-25, 2023

MPerformer: An SE(3) Transformer-based Molecular Perceptron

Fanmeng Wang[1], Hongteng Xu[1], Xi Chen[2], Shuqi Lu[2], Yuqing Deng[2], Wenbing Huang[1]

[1]Renmin University of China

[2]DP Technology

October 24, 2023

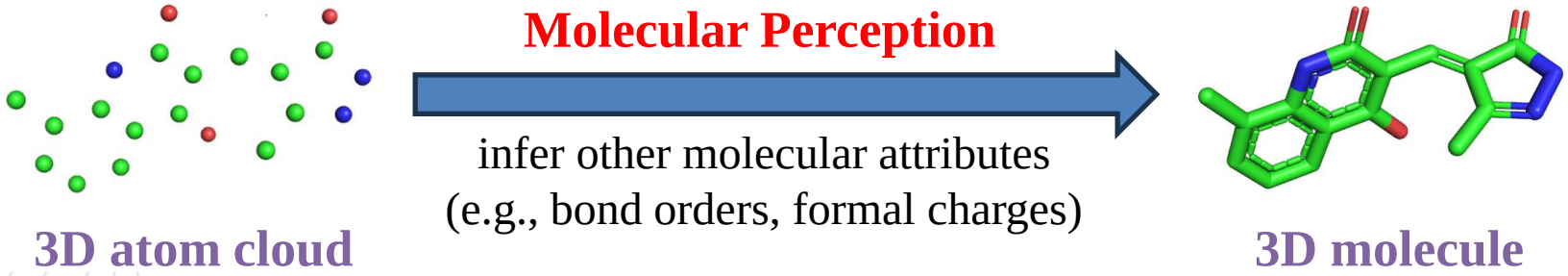
RENMIN UNIVERSITY OF CHINA

DP Technology

Background

■ Molecular Perception

➢ **Task Definition:** based on 3D atom clouds (i.e., atom types and corresponding 3D coordinates) to construct 3D molecules



➢ **Why we need molecular perception?**

- Molecular 3D structure plays a central role in many applications.
- 3D molecular data are often formulated as 3D atom clouds, while other molecular attributes, especially bond orders, are unknown in general.

Background

Rule-based Molecular Perception Methods

- **Practice:** Applying predefined chemical rules to determine molecular attributes
- **Drawbacks:**
  - Sensitive to the precision of atom 3D coordinates
  - Some require additional prior information (e.g., atom connectivity)

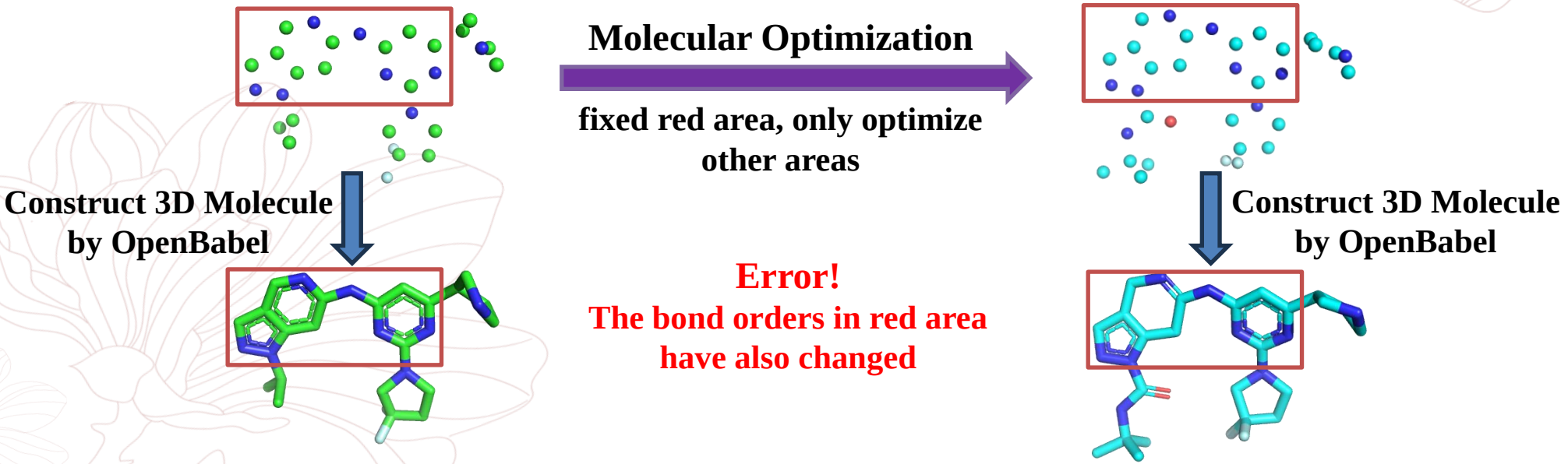
Existing Learning-based Molecular Perception Methods

- **Practice:** Applying machine learning models to predict molecular attributes
- **Drawbacks:**
  - Ignore the multi-task nature of molecular perception
  - Not fully leverage 3D geometric information contained in 3D atom clouds

Background

■ Motivation

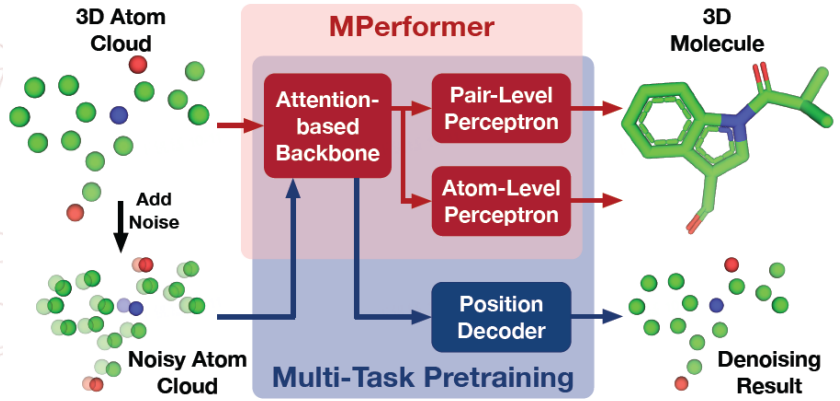
- Currently, **OpenBabel** is the most widely used tool to construct 3D molecules (i.e., sdf file) from corresponding 3D atoms clouds (i.e., xyz file).
- However, the performance of existing methods, including OpenBabel, is still sub-optimal in practice.



Background

- **An Ideal Molecular Perception Method should**
  - SE(3)-invariance.
  - Robustness to position noise.
  - Generalization power.

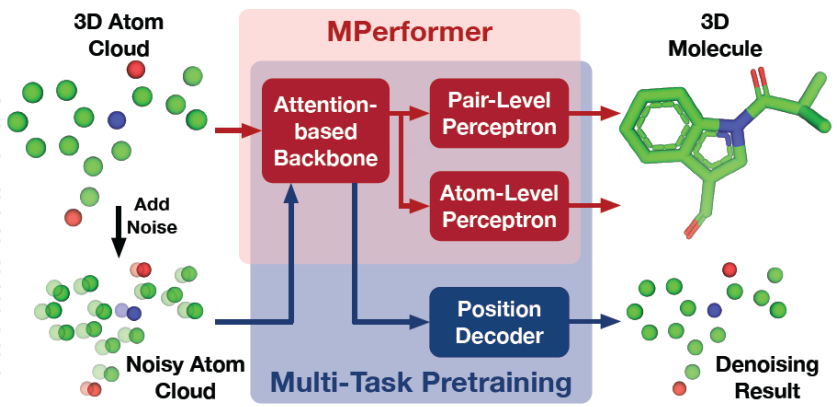
To achieve above properties, we propose **MPerformer**, an SE(3) Transformer-based molecular perceptron, to construct 3D molecules from 3D atom clouds efficiently.

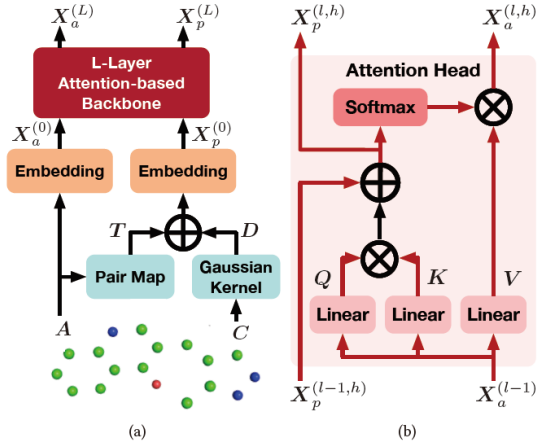


Our Approach: MPerformer

Overview

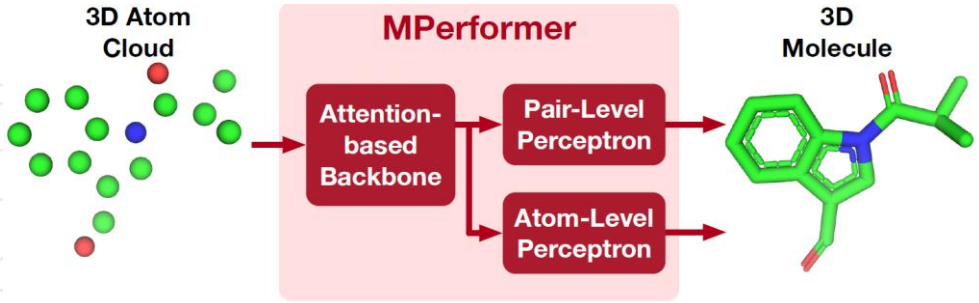
- The **first universal** learning-based molecular perception framework exhibiting **SE(3)-invariance.**
- It purely takes **3D atom clouds** as inputs, predicting atom-level and pair-level **molecular attributes** to construct **3D molecules** accurately and efficiently.
- To further enhance its robustness to position noise, we design a **multi-task pretraining-and-finetuning** learning paradigm.





Our Approach: MPerformer

- **Construct 3D Molecules from 3D Atom Clouds**
  - Given a **3D atom cloud**, we firstly use a attention-based backbone to fully extract atom-level and pair-level **molecular representations**.
  - Considering the **inherent correlations** between various molecular attributes, we further predict atom-level and pair-level **molecular attributes** by corresponding perceptrons simultaneously, thus construct **3D molecules**.



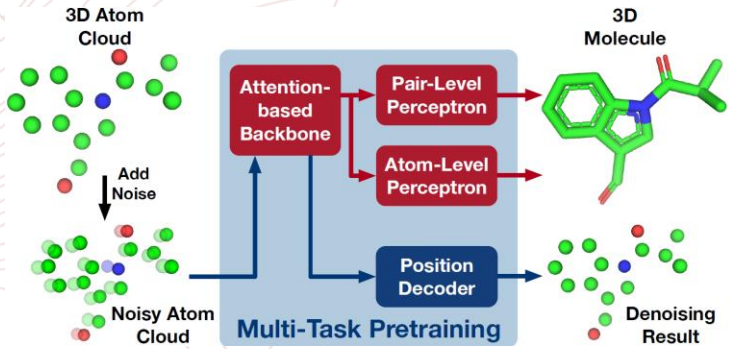
How to achieve high robustness to position noise?

A multi-task pretraining-and-finetuning learning paradigm

Our Approach: MPerformer

■ Multi-Task Pretraining and Finetuning

- In the **pretraining** phase, we introduce a reconstruction task to the learning objective besides the prediction task.
- **Prediction Task:** Predict atom-level and pair-level **molecular attributes** based on the learned molecular representations of the noisy atom cloud.
- **Reconstruction Task:** Reconstruct the **original clear atom cloud** by decoding the learned molecular representations of the noisy atom cloud.
- Finally, we **finetuning** this model purely based on prediction task.



A robust and generalizable model !

Experiments

Experimental Setup

Datasets

- **Experimental datasets:** generated through real-world experiments
- **Idealized datasets:** generated through computational softwares

Table 1: The statistics of datasets

Type	Dataset	Source	#Molecules	#Atoms	#Bonds with different types					
					Null	Single	Double	Triple	Aromatic	Total
Experimental datasets	Epdbccd	CCD	36,733	938,360	26,392,098	1,101,064	156,928	4,186	746,804	28,401,080
	Subpdb	PDBind	4,352	112,722	3,171,012	136,064	22,420	272	81,708	3,411,476
Idealized datasets	Ipdbccd	CCD	37,709	962,711	27,135,388	1,137,172	161,720	4,150	754,778	29,193,208
	Bradley1	UFF in RDKit	28,279	419,422	6,449,805	434,834	66,022	3,804	364,426	7,318,891
	Bradley2	MMFF in RDKit	27,954	416,148	6,415,902	431,658	65,878	3,794	361,312	7,278,544
	Bradley3	ETKDG in RDKit	28,306	420,590	6,511,123	436,732	66,130	3,804	364,802	7,382,591
	3dqsar	3Dqsar study	1,249	30,105	662,068	27,620	3,880	40	33,400	727,008
	Gdb1k	DFT	1,000	6,359	23,222	8,628	1,410	700	1,066	35,026

* Here, the null bond (i.e., disconnected atom pair) is also considered a type of chemical bond.

Experiments

Experimental Setup

Baselines

- **OpenBabel**: the most commonly-used rule-based molecular perception method, which has been widely integrated into many chemical softwares
- **Mamba**: a decision tree-based molecular perception method
- **Schrödinger Maestro**: a popular commercial software, we only test it on some representative molecules

Metrics

- **Overall Molecule-level Perception Accuracy (Overall Acc)**: proportion of molecules in which all molecular attributes are predicted correctly
- **Molecule-level Bond Accuracy (Bond Acc)**: proportion of molecules in which the bond orders of all the atom pairs are predicted correctly

Experiments

■ Performance

➢ Molecular Perception Capacity

Table 2: The molecular perception capacity of different molecular perception methods

Metric	Method	Dataset								Challenging set		
		Epdbccd	Subpdb	Ipdbccd	Bradley1	Bradley2	Bradley3	3dqsar	Gdb1k	Molecules	Rings	Time
Bond Acc	OpenBabel	0.553	0.623	0.551	0.819	0.808	0.799	0.761	0.901	0.580	0.748	51s
	Mamba	0.548	0.834	0.817	0.994	0.931	0.980	0.915	0.948	0.628	0.900	135s
	MPerformer	**0.670**	**0.971**	**0.872**	**0.994**	**0.932**	**0.990**	**0.973**	**0.967**	**0.799**	**0.949**	**25s**
Overall Acc	OpenBabel	0.532	0.424	0.529	0.813	0.803	0.794	0.629	0.836	—	—	—
	MPerformer	**0.616**	**0.700**	**0.821**	**0.898**	**0.899**	**0.892**	**0.772**	**0.962**	—	—	—

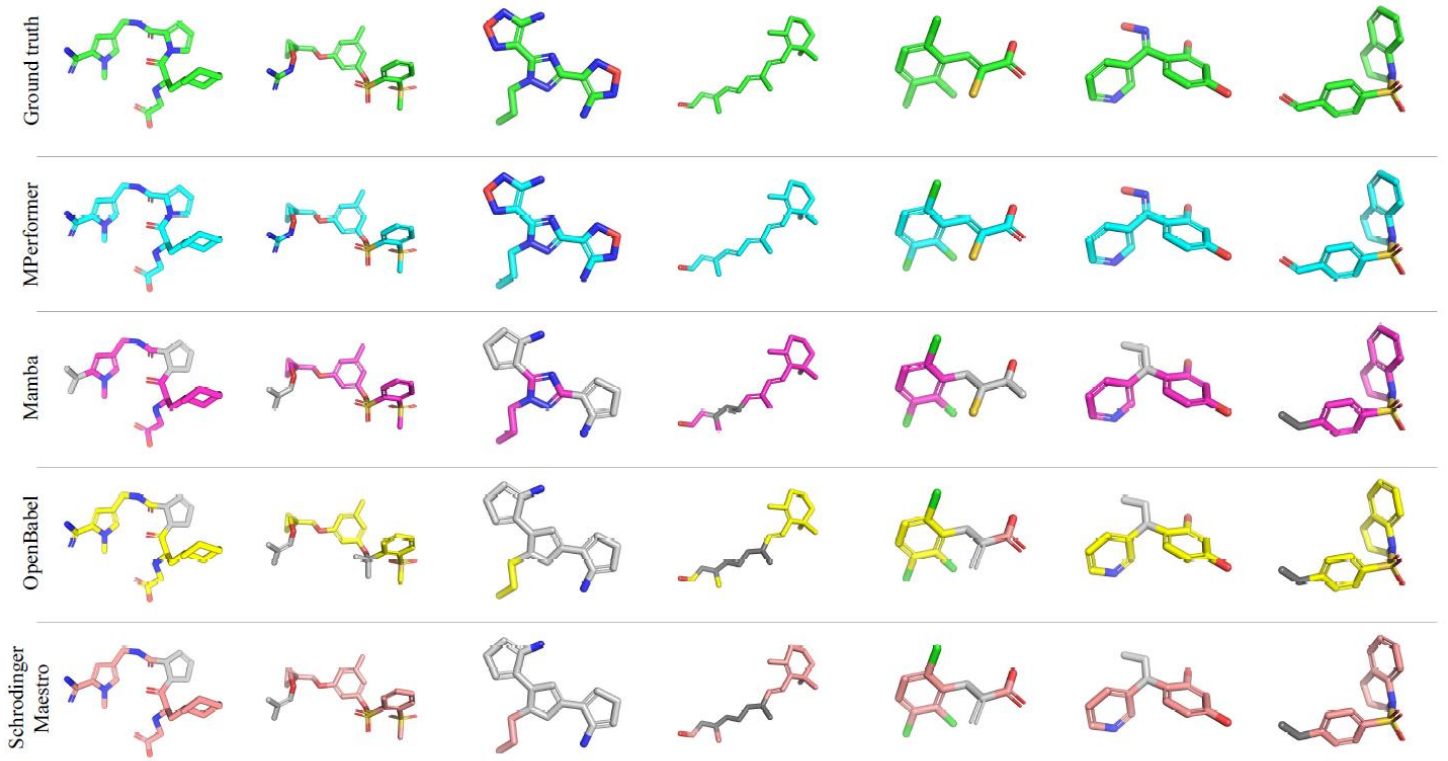
* Mamba can only predict bond orders.

MPerformer achieves **state-of-the-art** performance

Experiments

■ Performance

➢ Molecular Perception Capacity



Experiments

■ Performance

➢ Robustness

Table 3: The robustness of different methods under different noises

Metric	Dataset	Method	No noise	Gaussian noise			Uniform noise		
				0.03	0.05	0.07	0.10	0.15	0.20
Bond Acc	Epdbccd	OpenBabel	0.553	0.528	0.466	0.363	0.532	0.492	0.427
		Mamba	0.548	0.443	0.229	0.118	0.452	0.294	0.168
		MPerformer	**0.670**	**0.654**	**0.601**	**0.525**	**0.654**	**0.624**	**0.576**
	Ipdbccd	OpenBabel	0.551	0.533	0.452	0.338	0.537	0.484	0.403
		Mamba	0.817	0.396	0.127	0.060	0.416	0.174	0.085
		MPerformer	**0.872**	**0.862**	**0.851**	**0.825**	**0.860**	**0.851**	**0.840**
Overall Acc	Epdbccd	OpenBabel	0.532	0.507	0.448	0.351	0.511	0.472	0.413
		MPerformer	**0.616**	**0.603**	**0.549**	**0.470**	**0.602**	**0.571**	**0.522**
	Ipdbccd	OpenBabel	0.529	0.510	0.434	0.324	0.514	0.463	0.386
		MPerformer	**0.821**	**0.808**	**0.798**	**0.775**	**0.807**	**0.800**	**0.785**

MPerformer exhibits strong **robustness** to position noise

Experiments

■ Performance

➢ Ablation Studies

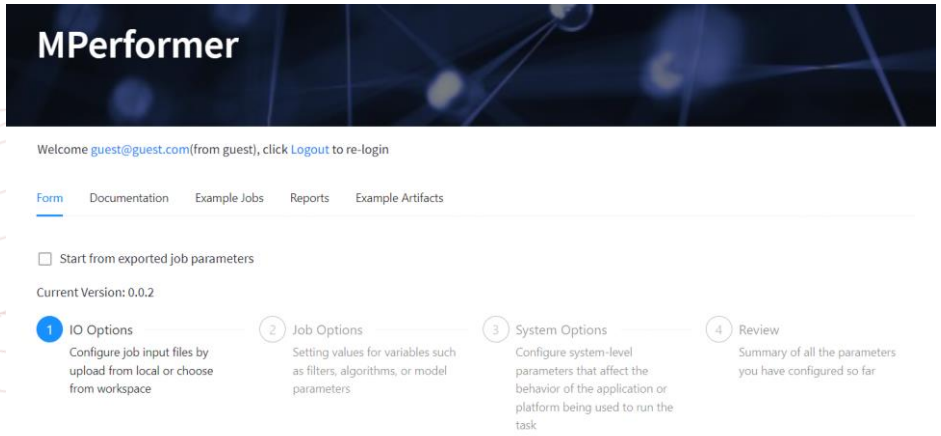
Table 4: The robustness capability of MPerformer under different pretraining settings

Dataset	$\mathcal{L}_{pred}$	$\mathcal{L}_{rec}$	Bond Acc						Overall Acc					
			Gaussian noise			Uniform noise			Gaussian noise			Uniform noise		
			0.03	0.05	0.07	0.10	0.15	0.20	0.03	0.05	0.07	0.10	0.15	0.20
Epdbccd	X	X	0.558	0.457	0.338	0.570	0.501	0.410	0.511	0.408	0.291	0.524	0.451	0.355
	X	✓	0.653	0.585	0.474	**0.662**	0.623	0.550	0.596	0.529	0.413	0.604	0.561	0.493
	✓	X	0.651	0.594	0.507	0.655	0.616	0.572	0.600	0.538	0.447	**0.605**	0.561	0.511
	✓	✓	**0.654**	**0.601**	**0.525**	0.654	**0.624**	**0.576**	**0.603**	**0.549**	**0.470**	0.602	**0.571**	**0.522**
Ipdbccd	X	X	0.302	0.089	0.037	0.319	0.127	0.054	0.266	0.061	0.021	0.281	0.098	0.036
	X	✓	0.813	0.401	0.151	0.836	0.539	0.272	0.725	0.326	0.111	0.753	0.447	0.215
	✓	X	0.842	0.531	0.256	0.850	0.652	0.411	0.763	0.447	0.195	0.775	0.564	0.340
	✓	✓	**0.862**	**0.851**	**0.825**	**0.860**	**0.851**	**0.840**	**0.808**	**0.798**	**0.775**	**0.807**	**0.800**	**0.785**

The effectiveness of our **multi-task pretraining-and-finetuning** learning paradigm

Conclusion

- **MPerformer**: A novel SE(3) Transformer-based molecular perceptron:
  - Construct **high-quality** 3D molecules purely based on 3D atom clouds
  - Do not dependent on any **additional prior knowledge**
  - Achieve **state-of-the-art** performance
  - Exhibit strong **robustness** to position noise



Demo: https://app.bohrium.dp.tech/mperforemr

Thank You & QA

Code: ***https://github.com/FanmengWang/MPerformer***

Demo: **https://app.bohrium.dp.tech/mperforemr**

Fanmeng Wang

fanmengwang@ruc.edu.cn