AI for materials science

周浩

zhouhao.nlp@air.tsinghua.edu.cn

清華大學智能产业研究院

Institute for AI Industry Research, Tsinghua University

大纲

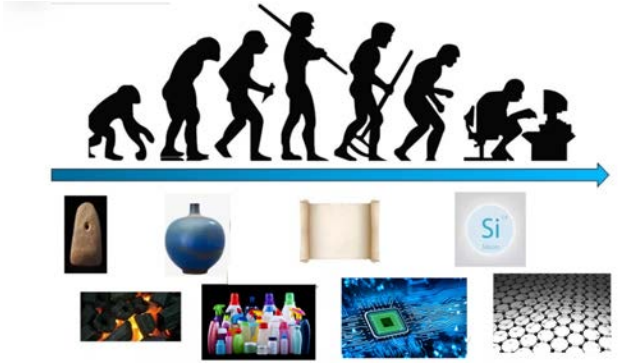
- 背景与动机
- 相关工作
  - ✓ 材料表示学习
  - ✓ 材料生成
- 团队工作
  - ✓ 从生成算法出发，设计适配分子的生成模型
  - ✓ 从基座构建出发，建立富含广袤数据知识的材料基座
- 未来工作
- 总结

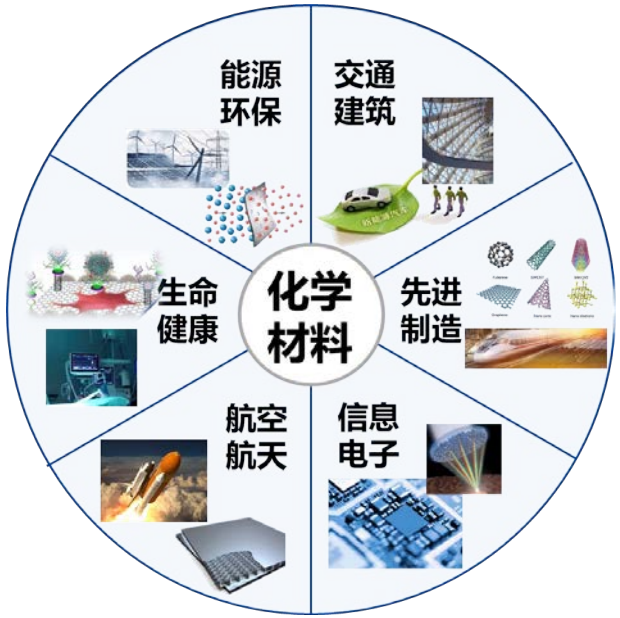
第一部分

背景与动机

材料对人类社会至关重要

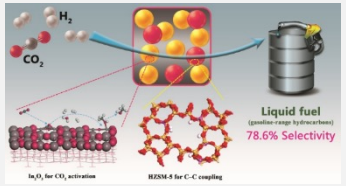
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



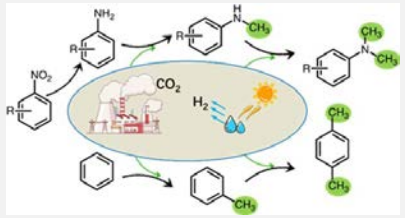


低碳技术

低碳燃料



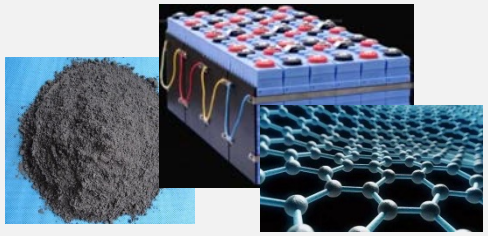
绿色低碳化学品



零碳技术

光伏材料

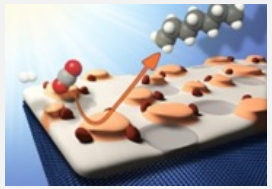
电池材料



负碳技术

碳捕集材料

CO_2转化催化剂



材料研发的难点

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

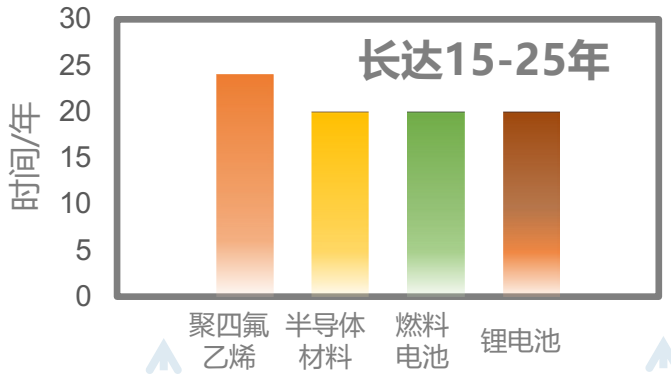
传统模式

化学材料研发

研发周期长、成本高

困境与瓶颈

体系复杂、高度依赖专家经验



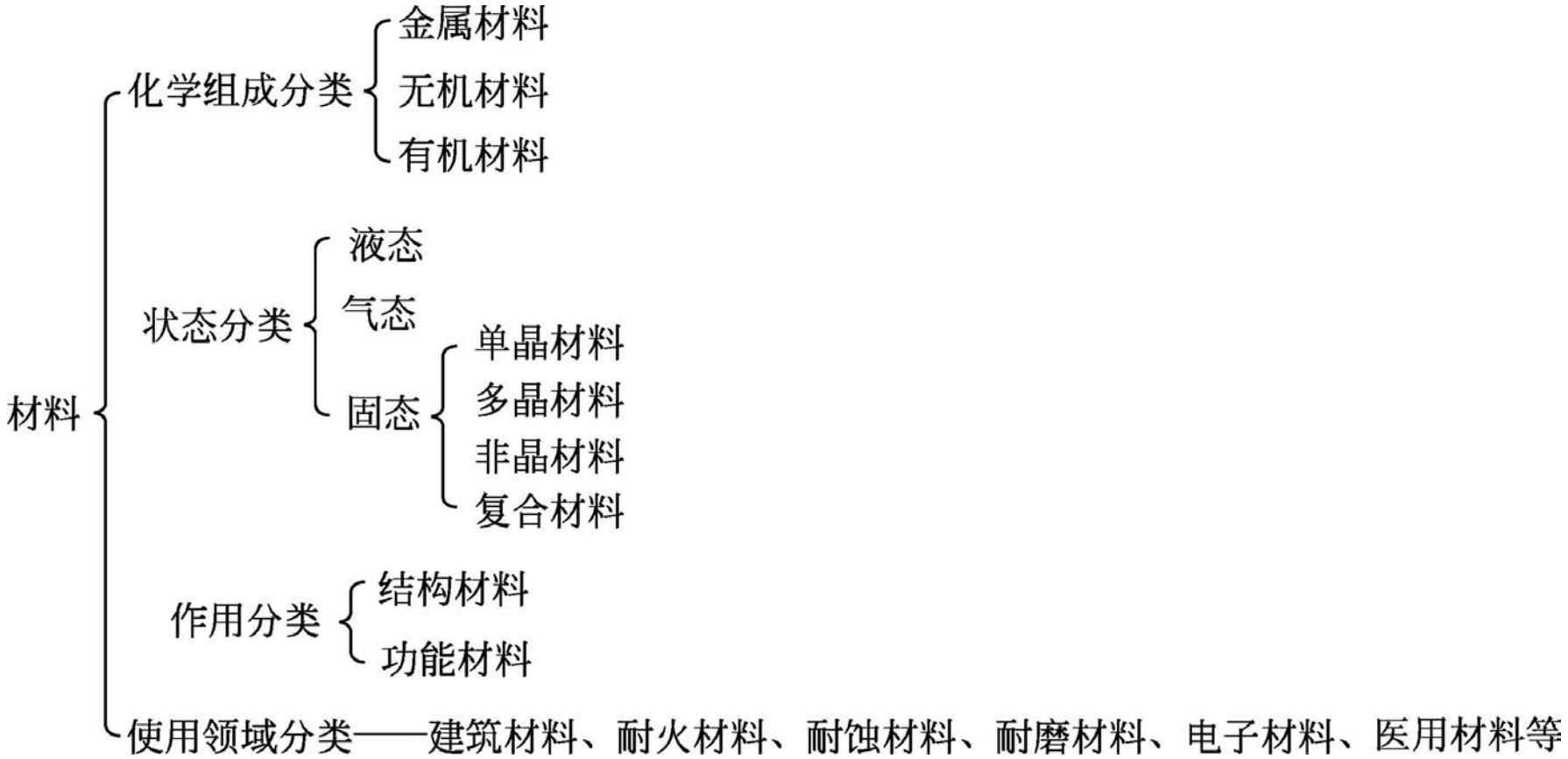
国外巨头
百年研发经验积累，先发垄断
国内研发
配方代差，自研效果有差距

AI赋能材料的机遇

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

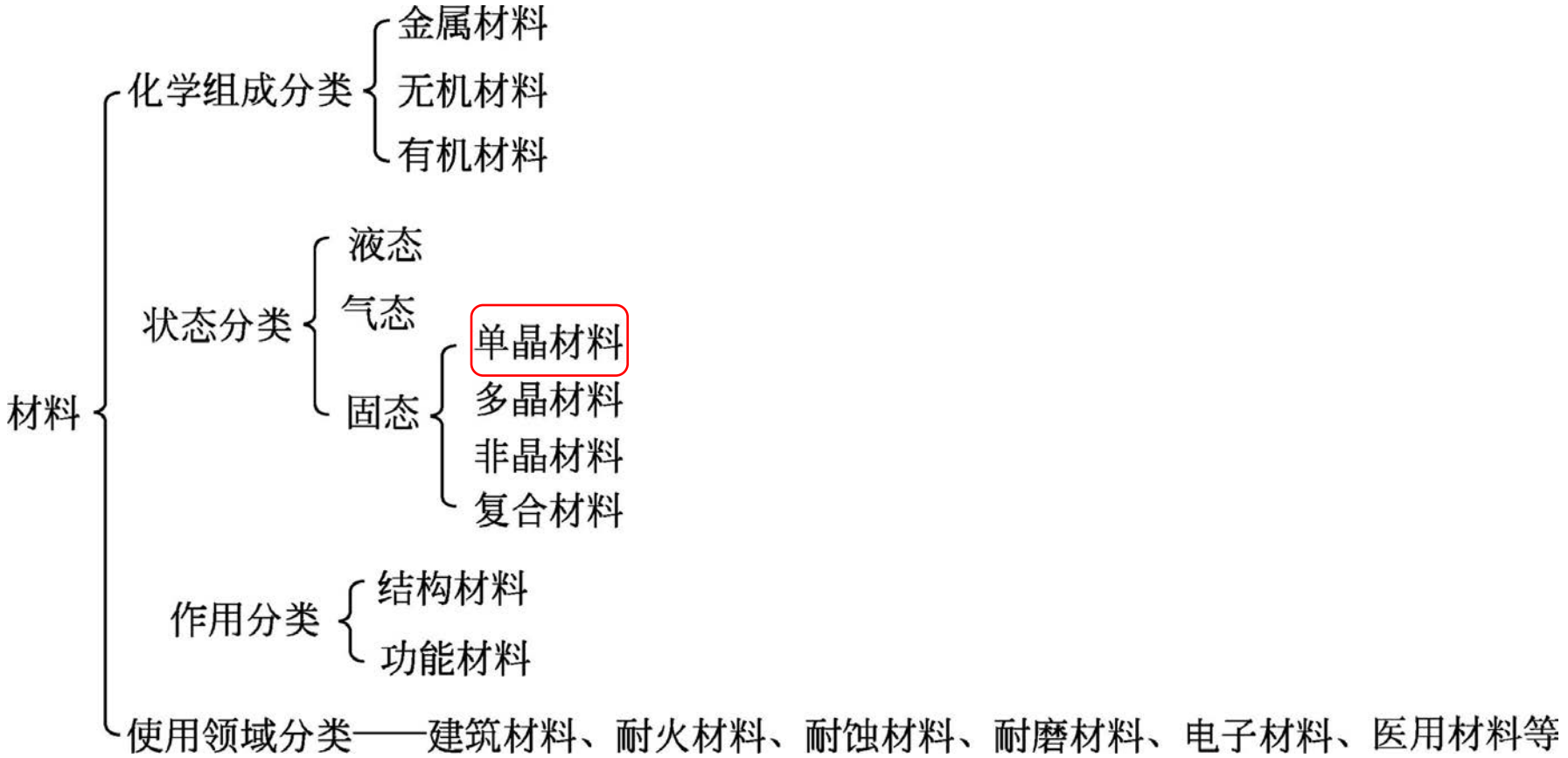
- 两大方向
  - AI 加速材料性质计算：利用神经网络拟合材料属性数据，**更快更准**
  - AI 生成新的候选材料：利用生成模型学习稳定材料分布，**更多更好**
- 两大方向可形成数据飞轮，加速新材料发现

材料类型



《高分子材料概论》，高长有

材料类型

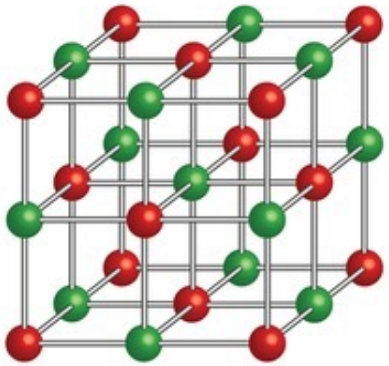


《高分子材料概论》，高长有

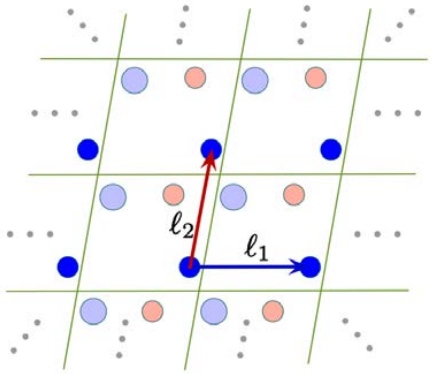
晶体

晶体可以表示为一组原子(unit cell)在3D空间中的**周期性排布**

- 原子类型: $\boldsymbol{A} = (a_0, ..., a_N) \in \mathbb{A}^N$
- 原子坐标: $\boldsymbol{X} = (\boldsymbol{x}_0, ..., \boldsymbol{x}_N) \in \mathbb{R}^{N\times 3}$ (笛卡尔坐标)
- 晶格矩阵: $\boldsymbol{L} = (\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3) \in \mathbb{R}^{3\times 3}$ (晶格矩阵)



3D NaCl



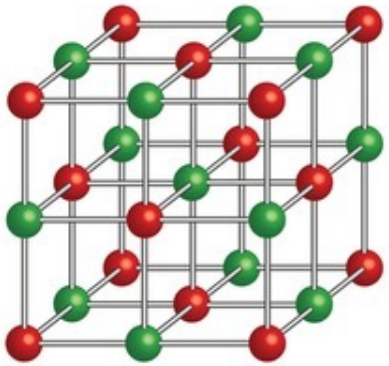
2D toy data

晶体

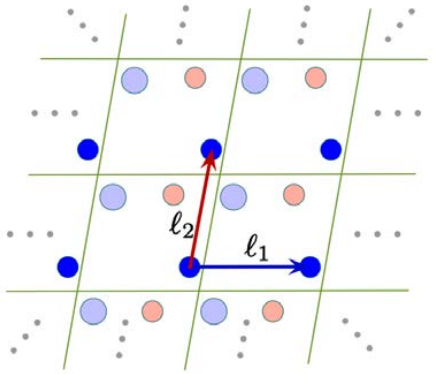
晶体可以表示为一组原子(unit cell)在3D空间中的**周期性排布**

- 原子类型: $\boldsymbol{A} = (a_0, ..., a_N) \in \mathbb{A}^N$
- 原子坐标: $\boldsymbol{X} = (\boldsymbol{x}_0, ..., \boldsymbol{x}_N) \in \mathbb{R}^{N\times 3}$ (笛卡尔坐标) / $[0,1)^{N\times 3}$ (分数坐标)
- 晶格矩阵: $\boldsymbol{L} = (\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3) \in \mathbb{R}^{3\times 3}$ (晶格矩阵)

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F}$$



3D NaCl



2D toy data

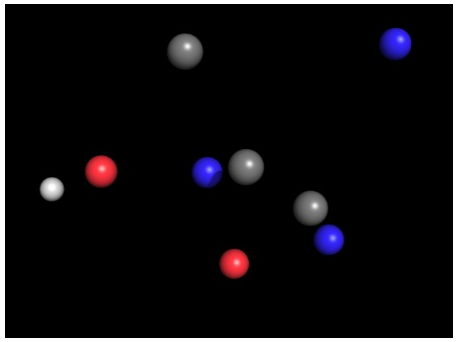
晶体 vs. 其他数据类型

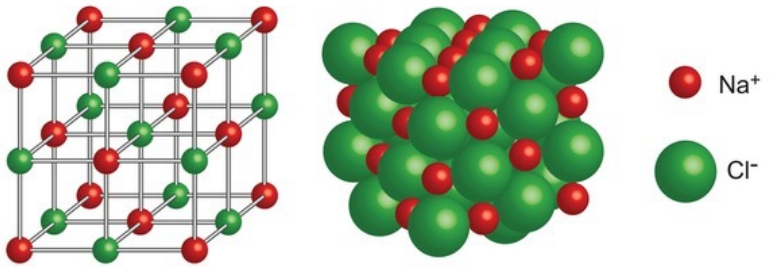
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

文本
Token: 离散数据

图片、视频
Pixel: 连续数据

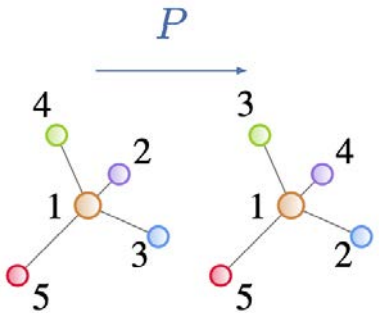
药物分子
原子类型: 离散
原子坐标: 连续



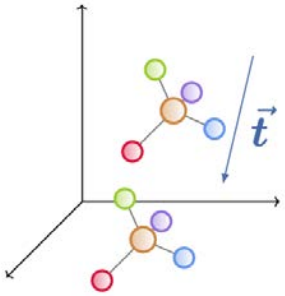


晶体
原子类型: 离散
原子坐标: 连续/分数
晶格矩阵: 连续

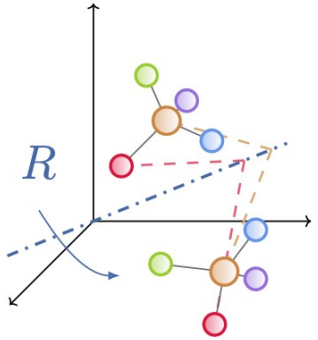
AI赋能材料的建模挑战



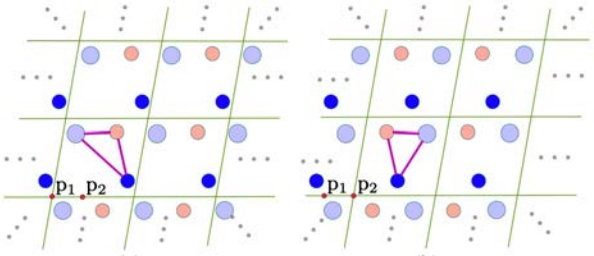
(a) permutation



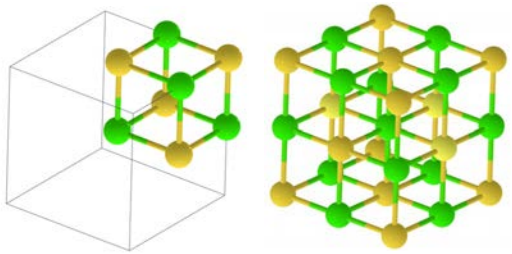
(b) translation



(c) rotation



(d) Periodic cell choice

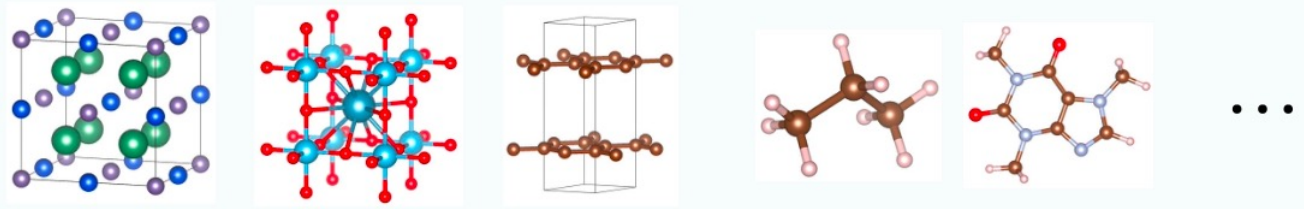


(e) supercell choice

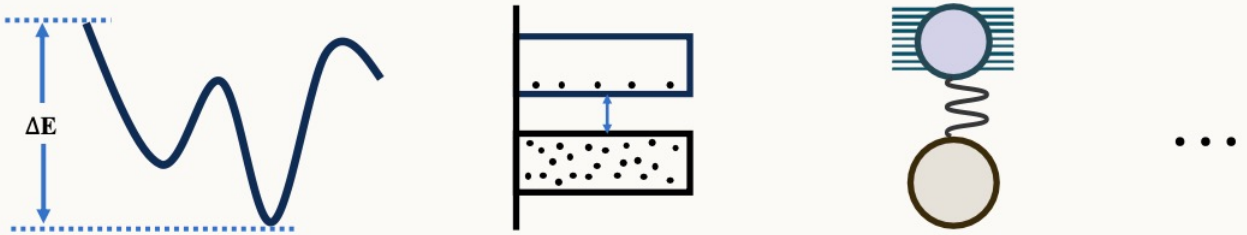
晶体几何约束：不变和等变性

A Survey of Geometric Graph Neural Networks: Data Structures, Models and Applications. FCS

AI赋能材料的数据挑战

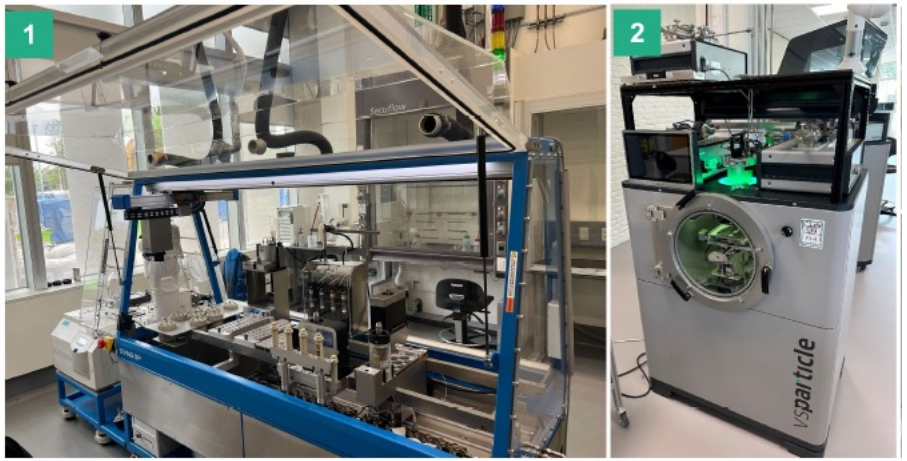


Crystalline Materials　　Organic Molecules



Formation Energy　　Band Gap　　Phonons Prediction

数据、任务体系复杂



湿实验耗时耗力

第二部分

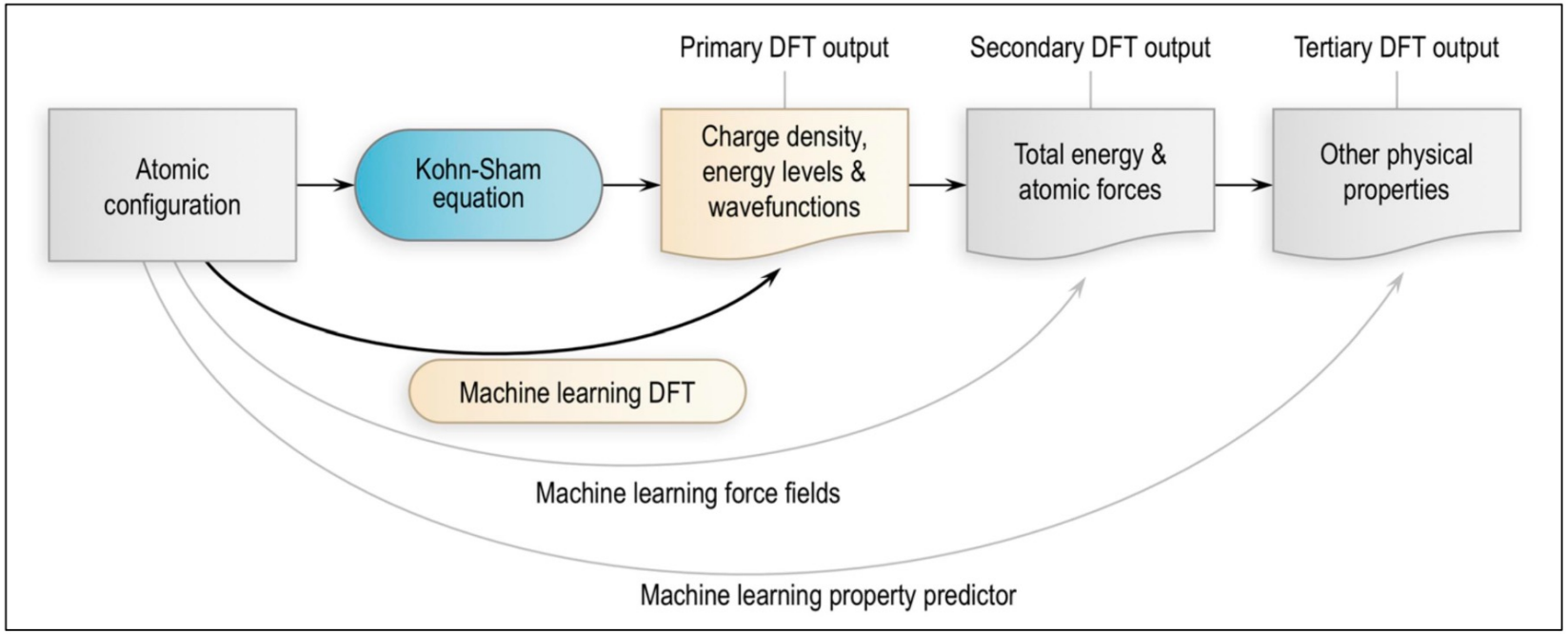
相关工作

大纲

- 背景与动机
- 相关工作
  - ✓ 材料表示学习
  - ✓ 材料生成
- 团队工作
  - ✓ 从数据结构出发，找到本征的数据刻画空间
  - ✓ 从生成算法出发，设计适配分子的生成模型
  - ✓ 从基座构建出发，建立富含广袤数据知识的预训练基座
- 总结

材料表示学习

Material representation learning requires learning a function f to predict the property y of any given material $\mathcal{M}$, and y can be a real number (regression problem) or categorical number (classification problem).



关键物理性质

现有工作

优化模型架构

- 从晶体出发
  - ✓CGCNN[1], Matformer[2], M3GNet[3]…
- 从性质出发
  - ✓eSEN[4], Orb[5,6]…

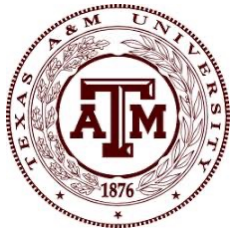
优化预训练策略

- 从有监督出发
  - ✓JMP[7],MoE-18[8]…
- 从无监督出发
  - ✓Orb[5,6], Frad[9]…

[1] Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties
[2] Periodic Graph Transformers for Crystal Material Property Prediction
[3] A universal graph deep learning interatomic potential for the periodic table
[4] Learning smooth and expressive interatomic potentials for physical property prediction
[5] Orb: A Fast, Scalable Neural Network Potential
[6] Orb-v3: atomistic simulation at scale
[7] from molecules to materials: pre-training large generalizable models for atomic property prediction
[8] Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework.
[9] Pre-training with Fractional Denoising to Enhance Molecular Property Prediction

优化模型架构-从晶体出发

Matformer

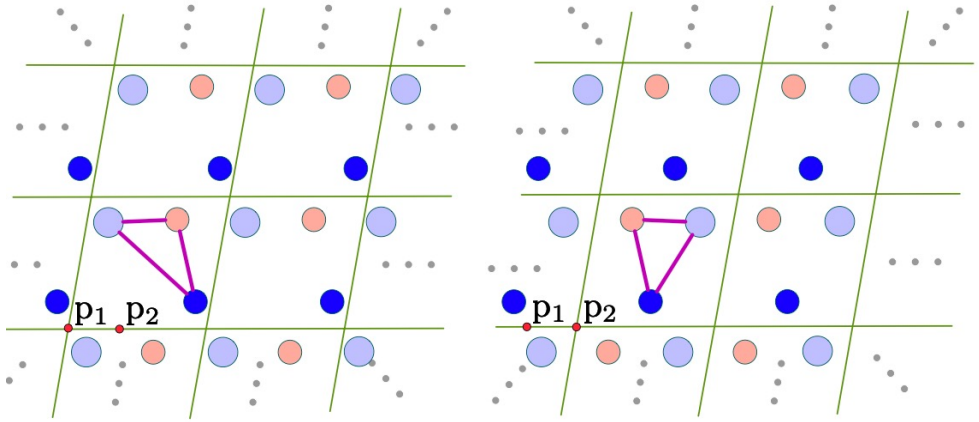


Periodic Graph Transformers for Crystal Material Property Prediction, Keqiang Yan, et al, NeurIPS 2022

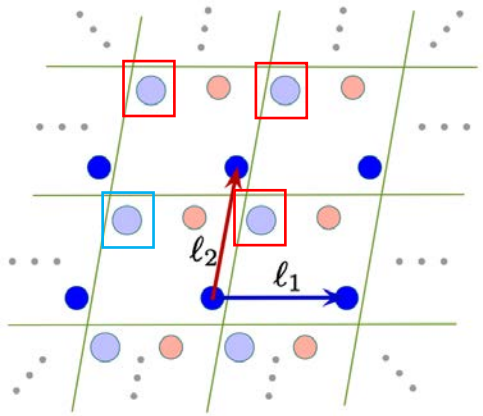
Motivation

多数Encoder只考虑E(3)等变性，仍忽略晶体的以下性质:

(1)不满足周期不变性；(2)忽略晶格矩阵信息



周期不变性

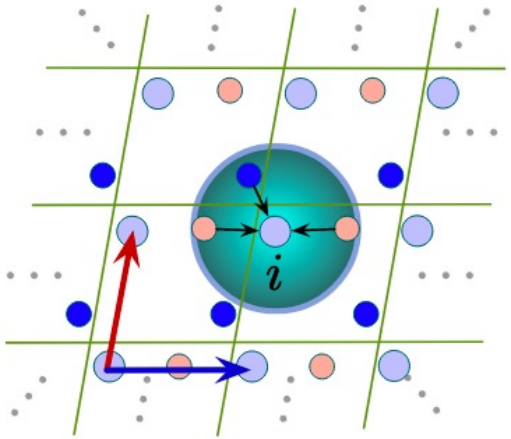


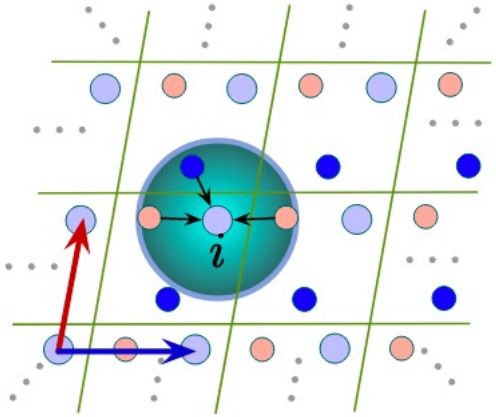
晶格矩阵信息

Method – 新建图方式

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

(1) 无论晶格矩阵如何挑选，原子周围的信息不变的，所以基于半径的图构建可以满足周期不变性

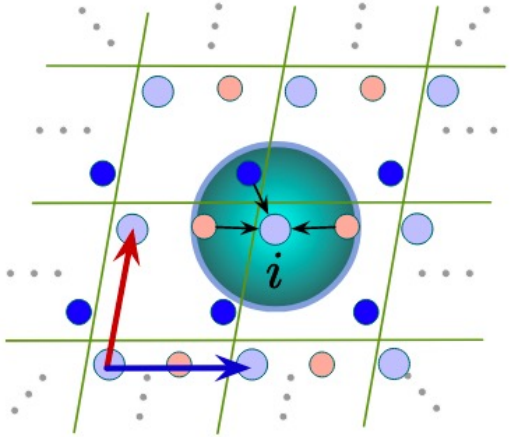


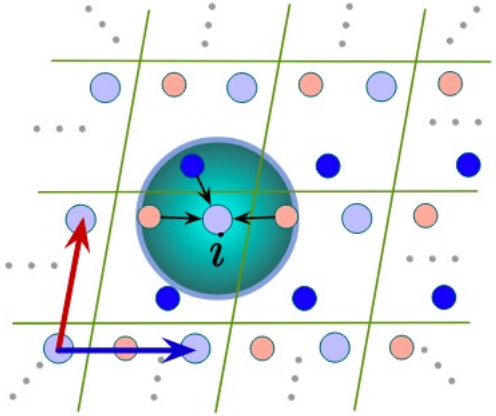


Method – 新建图方式

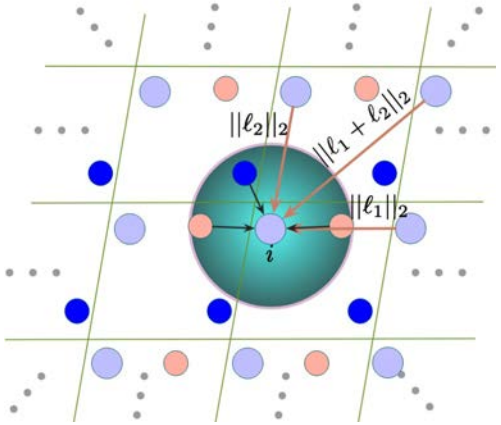
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

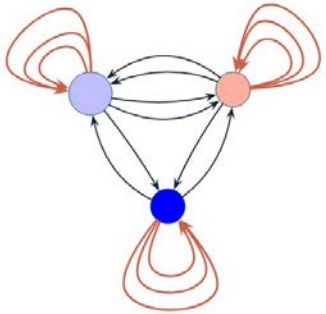
(1) 无论晶格矩阵如何挑选，原子周围的信息不变的，所以基于半径的图构建可以满足周期不变性





(2)同时考虑相邻晶胞的相同原子来建模周期信息

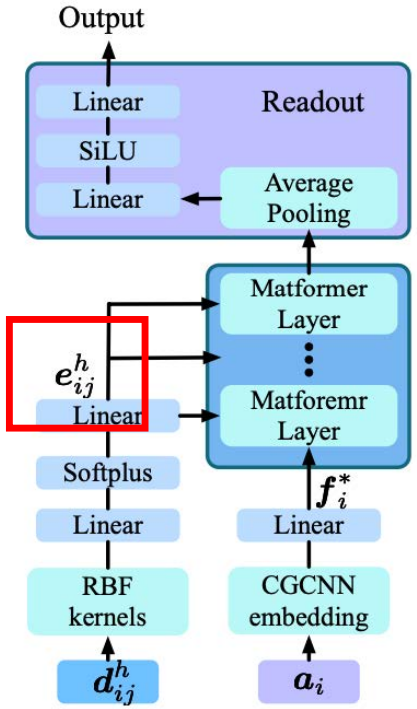




最终建图

Matformer – 新架构

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



$$q_i = \mathrm{LN}_Q(f_i^{\star\ell}),\ k_i = \mathrm{LN}_K(f_i^{\star\ell}),\ k_j = \mathrm{LN}_K(f_j^{\star\ell}),\ e_{ij}^{h'} = \mathrm{LN}_E(e_{ij}^h),$$

$$q_{ij}^h = (q_i|q_i|q_i),\ k_{ij}^h = (k_i|k_j|e_{ij}^{h'}),\ \alpha_{ij}^h = \frac{q_{ij}^h \circ k_{ij}^h}{\sqrt{d_{k_{ij}^h}}},$$

修改注意力机制来充分考虑边信息

Result

Table 1: Comparison in terms of test MAE on The Materials Project dataset. To make the comparison clear and fair, We show results from retrained models using exactly the same training, validation, and test sets. Results from original papers are shown in Appendix A.5. The best results are shown in **bold** and the second best results are shown with underlines.

	Formation Energy	Band Gap	Bulk Moduli	Shear Moduli
Method	eV/atom	eV	log(GPa)	log(GPa)
CGCNN [51]	0.031	0.292	0.047	0.077
SchNet [41]	0.033	0.345	0.066	0.099
MEGNET [4]	0.030	0.307	0.060	0.099
GATGNN [33]	0.033	0.280	0.045	0.075
ALIGNN [6]	0.022	0.218	0.051	0.078
Matformer	**0.021**	**0.211**	**0.043**	**0.073**

相比其他backbone效果显著

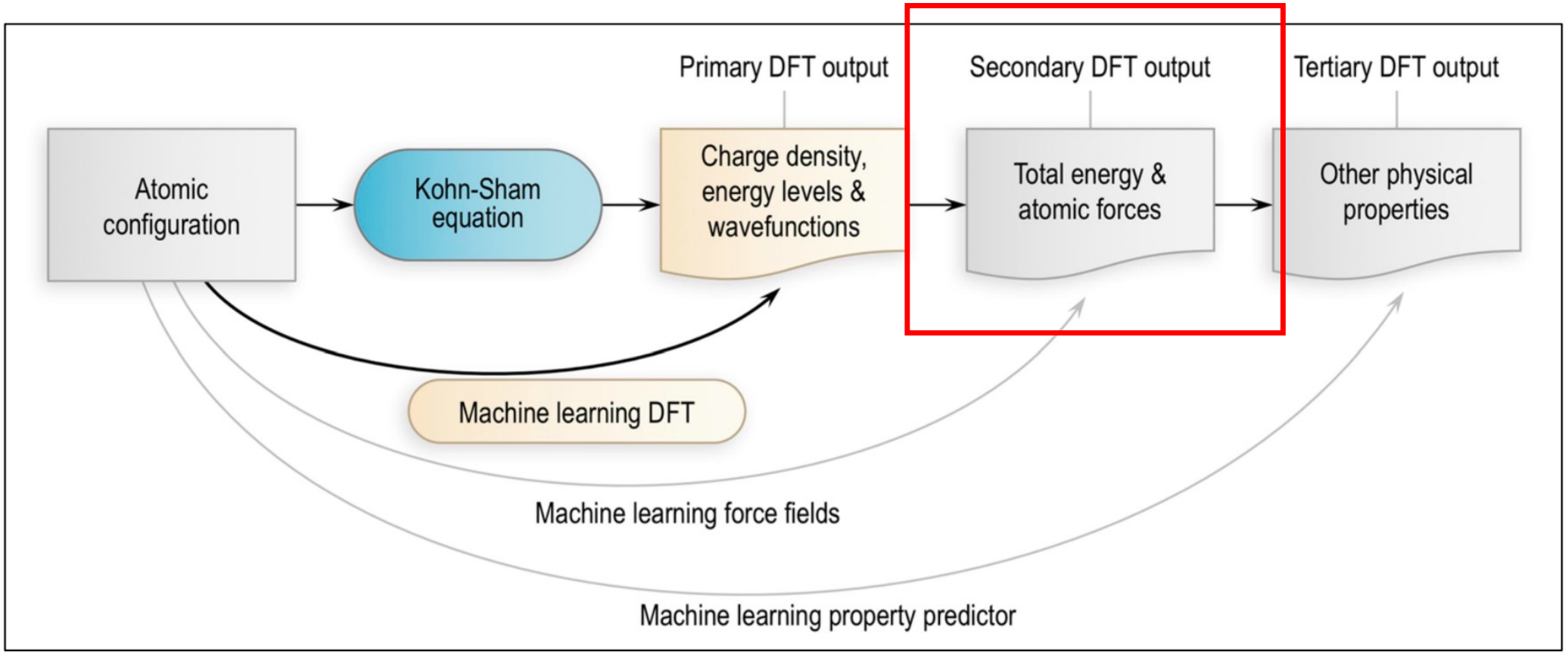
优化模型架构-从性质出发

eSEN

Meta

Meta. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction, arxiv 2025

ML potential



Motivation

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

现有ML potential的模型往往面临 (1) 非保守力；(2) 势能面不连续等问题，导致其不满足能量守恒，在下游任务(MD)中表现不佳

$$\oint \boldsymbol{F} \cdot d\boldsymbol{r} = 0$$

保守力场

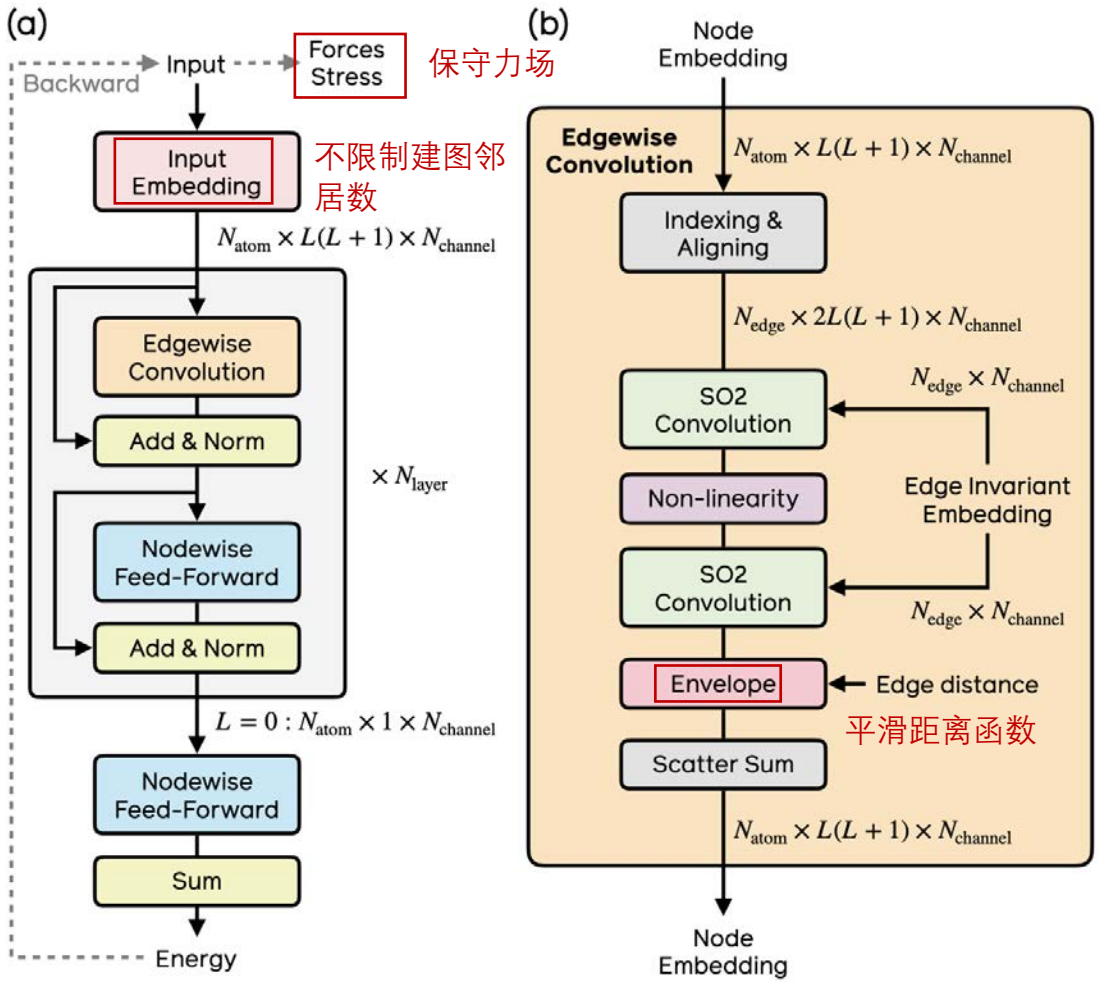
$$|E(\boldsymbol{r}_T, \boldsymbol{a}) - E(\boldsymbol{r}_0, \boldsymbol{a})| \leq C\Delta t^2 + C_N \Delta t^N T$$

势能面导数上界C_N决定能量漂移上界

Meta. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction, arxiv 2025

Method

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



- 保守力场：力由梯度计算

✅ $\hat{\mathbf{F}} = -\nabla_{\mathbf{r}} \hat{E}$ ❌ $\hat{\mathbf{F}} = f_\theta(\mathbf{r}, a)$

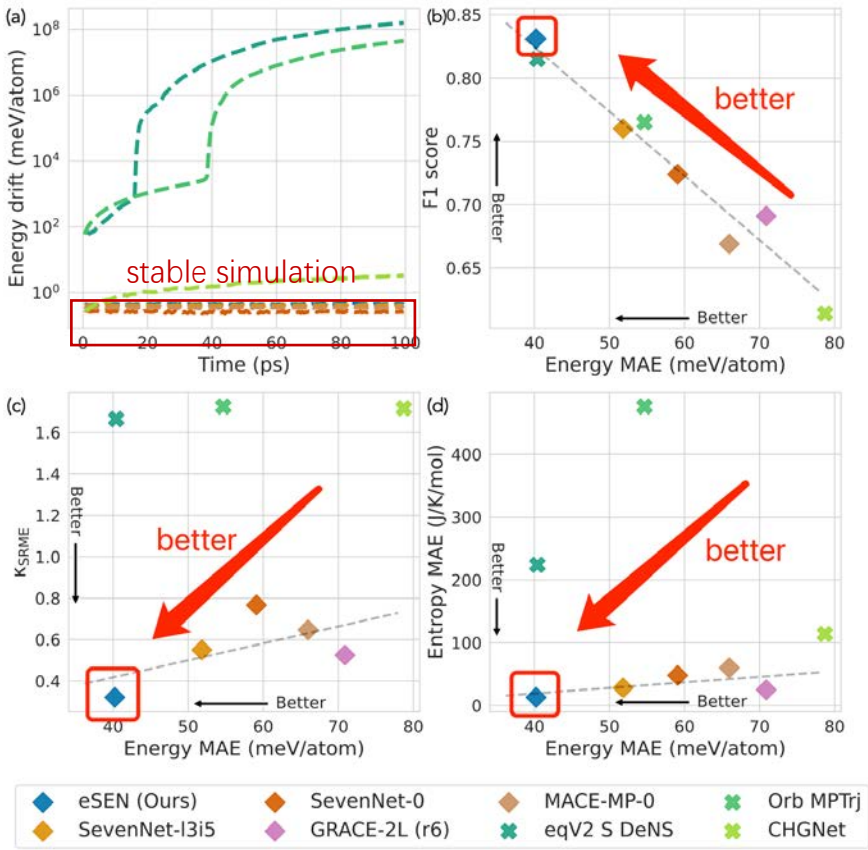
- 势能面平滑：
  - 不限制建图邻居数
  - 平滑的距离函数
  - …

Meta. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction, arxiv 2025

Result

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

Model	CPS ↑	Acc ↑	F1 ↑	DAF ↑	Prec ↑	MAE ↓	R^2 ↑	κ_{SRME} ↓	RMSD ↓
eSEN-30M-OAM	0.888	0.977	0.925	6.069	0.928	0.018	0.866	0.170	0.061
ORB v3	0.861	0.971	0.905	5.912	0.904	0.024	0.821	0.210	0.075
SevenNet-MF-ompa	0.845	0.969	0.901	5.825	0.890	0.021	0.867	0.317	0.064
GRACE-2L-OAM	0.837	0.963	0.880	5.774	0.883	0.023	0.862	0.294	0.067
eSEN-30M-MP	0.797	0.946	0.831	5.260	0.804	0.033	0.822	0.340	0.075
MACE-MPA-0	0.795	0.954	0.852	5.582	0.853	0.028	0.842	0.412	0.073
MatterSim v1 5M	0.767	0.959	0.862	5.852	0.895	0.024	0.863	0.574	0.073
DPA3-v2-OpenLAM	0.762	0.966	0.890	5.747	0.879	0.022	0.869	0.687	0.068
GRACE-1L-OAM	0.761	0.944	0.824	5.255	0.803	0.031	0.842	0.516	0.072
SevenNet-l3i5	0.714	0.920	0.760	4.629	0.708	0.044	0.776	0.550	0.085
MatRIS v0.5.0 MPtrj	0.681	0.938	0.809	5.049	0.772	0.037	0.803	0.861	0.077
GRACE-2L-MPtrj	0.681	0.896	0.691	4.163	0.636	0.052	0.741	0.525	0.090
DPA3-v2-MPtrj	0.646	0.929	0.786	4.822	0.737	0.039	0.804	0.959	0.082
MACE-MP-0	0.644	0.878	0.669	3.777	0.577	0.057	0.697	0.647	0.091
AlphaNet-MPTrj	0.566	0.933	0.799	4.863	0.743	0.041	0.745	1.310	0.107
eqV2 M	0.558	0.975	0.917	6.047	0.924	0.020	0.848	1.771	0.069
ORB v2	0.529	0.965	0.880	6.041	0.924	0.028	0.824	1.732	0.097
eqV2 S DeNS	0.522	0.941	0.815	5.042	0.771	0.036	0.788	1.676	0.076
ORB v2 MPtrj	0.470	0.922	0.765	4.702	0.719	0.045	0.756	1.725	0.101
M3GNet	0.428	0.813	0.569	2.882	0.441	0.075	0.585	1.412	0.112



Rank #1 on Matbench discovery

在几乎所有指标上都能取得最好效果

Meta. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction, arxiv 2025

优化预训练策略-从有监督出发

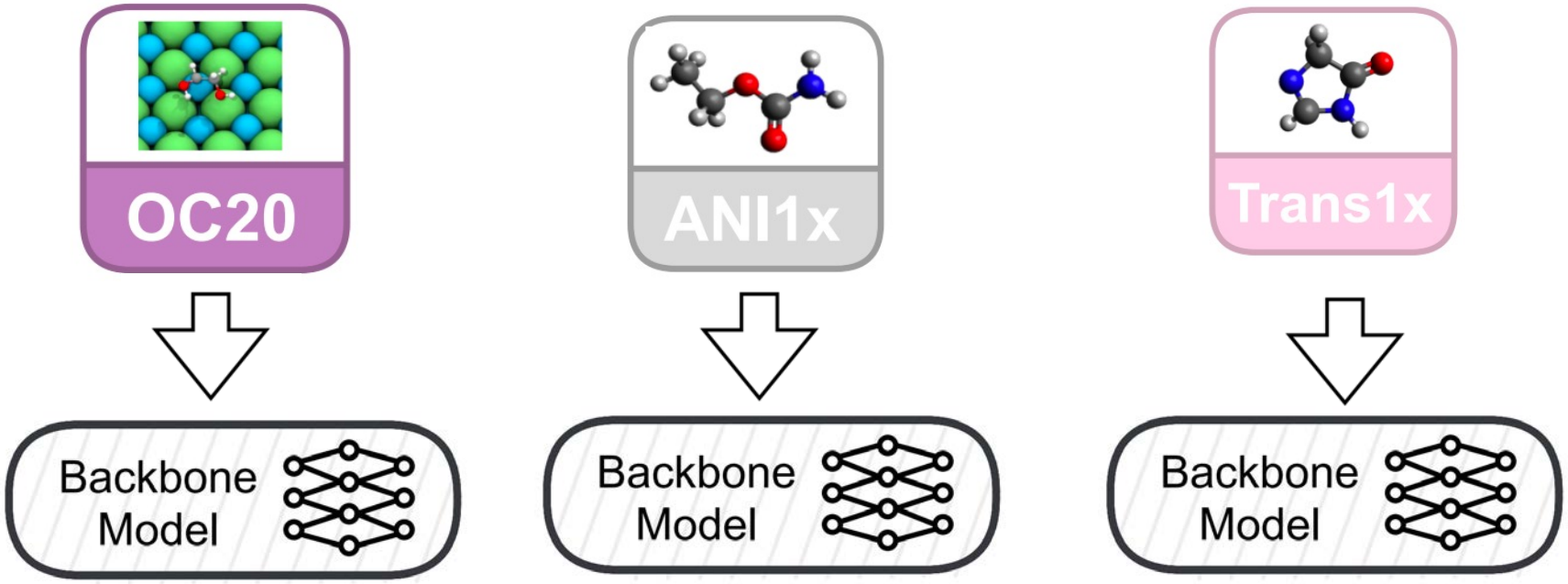
JMP

Meta

from molecules to materials: pre-training large generalizable models for atomic property prediction

Motivation

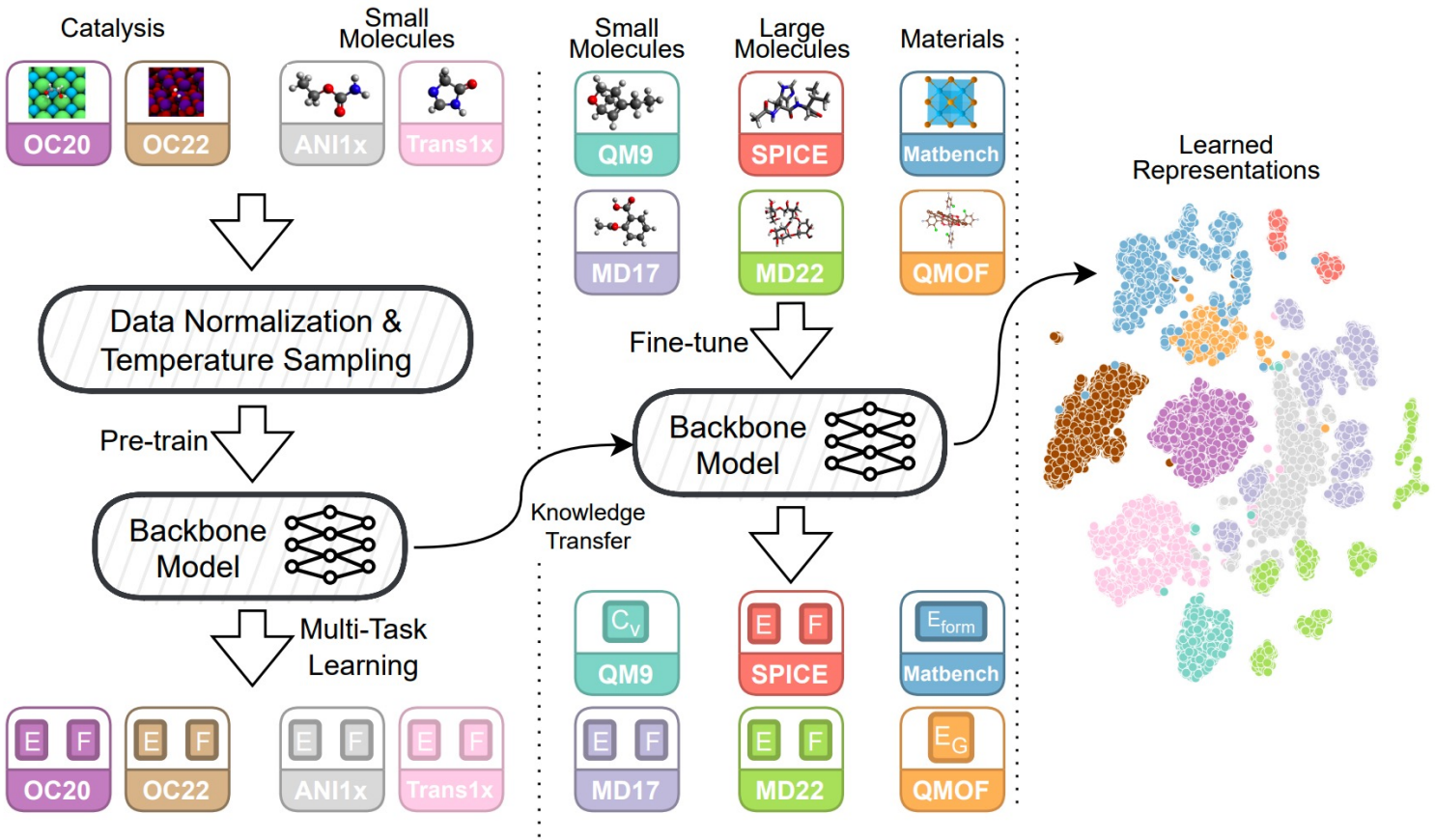
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



每个任务训一个模型不能够实现知识共享

Method

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



先通过多任务学习预训练，后针对每个任务调优

Result

Target (Units)	TorchMD-Net	Equi-former	MACE	Allegro	Pretrained ET-OREO	Pretrained GNS+TAT+NN	GN-OC-S	GN-OC-L	JMP-S	JMP-L
$\mu\,(D)$	0.011	0.011	0.015	-	-	0.016	0.020	0.023	0.010	**0.008**
$\alpha\,(a_0^3)$	0.059	0.046	0.038	-	-	0.040	0.052	0.056	0.037	**0.032**
$\varepsilon_{\text{HOMO}}\,(meV)$	20.3	15.0	22.0	-	16.8	14.9	21.8	22.7	11.1	**8.8**
$\varepsilon_{\text{LUMO}}\,(meV)$	18.6	14.0	19.0	-	14.5	14.7	17.3	18.6	10.8	**8.6**
$\Delta\varepsilon\,(meV)$	36.1	30.0	42.0	-	26.4	22.0	38.5	40.6	23.1	**19.1**
$R^2\,(a_0^2)$	**0.033**	0.251	0.210	-	-	0.440	0.210	0.171	0.200	0.163
ZPVE (meV)	1.8	1.3	1.2	-	-	1.0	1.2	1.2	1.0	**0.9**
$U_0\,(meV)$	6.2	6.6	4.1	4.7	-	5.8	7.2	9.4	3.3	**2.9**
$U\,(meV)$	6.4	6.7	4.1	4.4	-	5.8	6.9	9.7	3.3	**2.8**
$H\,(meV)$	6.2	6.6	4.7	4.4	-	5.8	7.3	8.7	3.3	**2.8**
$G\,(meV)$	8.3	7.6	5.5	5.7	-	6.9	8.1	9.2	4.5	**4.3**
C_ν (Cal/MolK)	0.026	0.023	0.021	-	-	0.020	0.024	0.024	0.018	**0.017**

Table 2: MAE test split results on all targets of the QM9 dataset. SOTA results are bolded.

Materials (Units)	MODNet (fold0 / mean)	coGN (fold0 / mean)	GN-OC-S (fold0)	GN-OC-L (fold0)	JMP-S (fold0 / mean)	JMP-L (fold0 / mean)
JDFT2D $(meV/atom)$	25.55 / 33.20	22.25 / 37.17	26.19	25.34	20.72 / 30.16	23.12 / **29.94**
Phonons (cm^-1)	34.77 / 34.28	32.12 / 29.71	93.45	88.74	26.6 / 22.77	21.28 / **20.57**
Dielectric $(unitless)$	0.169 / 0.271	0.178 / 0.309	0.225	0.211	0.133 / 0.252	0.119 / **0.249**
Log GVRH $(log10(GPA))$	0.073 / 0.073	0.068 / 0.069	0.082	0.082	0.06 / 0.062	0.057 / **0.059**
Log KVRH $(log10(GPA))$	0.054 / 0.055	0.052 / 0.054	0.061	0.063	0.044 / 0.046	0.045 / **0.045**
Perovskites $(eV/unitcell)$	0.093 / 0.091	0.027 / 0.027	0.045	0.045	0.029 / 0.028	0.026 / **0.026**
MP Gap (eV)	0.215 / 0.220	0.153 / 0.156	0.228	0.235	0.119 / 0.121	0.089 / **0.091**
MP E Form $(meV/atom)$	40.2 / 44.8	17.4 / 17	31.4	33.1	13.6 / 13.3	10.3 / **10.1**
	PT CGCNN	**PT MOFTransformer**				
QMOF	0.28	0.27	0.25	0.24	0.18	**0.16**

Table 4: MAE test split results on different targets in the materials domain. SOTA is bolded.

在小分子、材料任务上都得到了显著提升

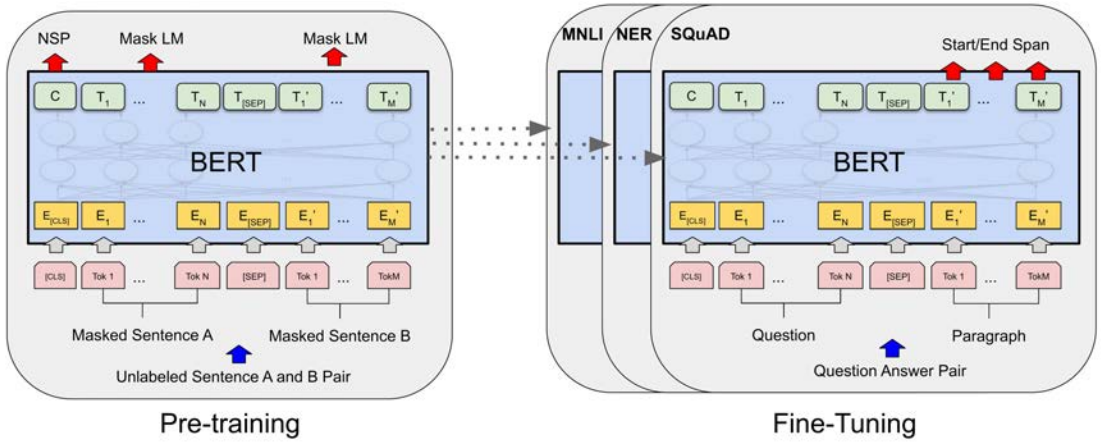
优化预训练策略-从无监督出发

Orb

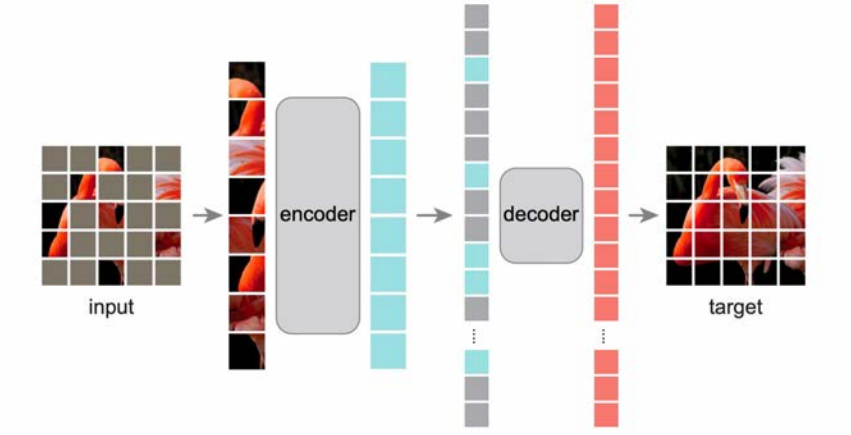
Orb: A Fast, Scalable Neural Network Potential
Orb-v3: atomistic simulation at scale

Motivation

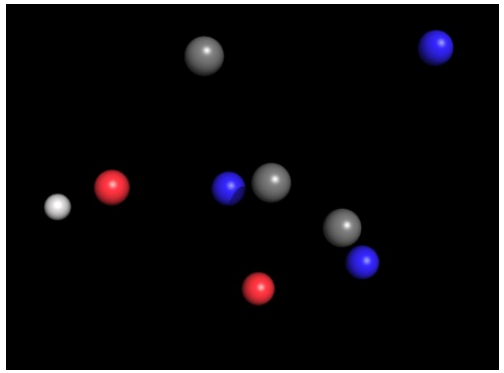
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



NLP



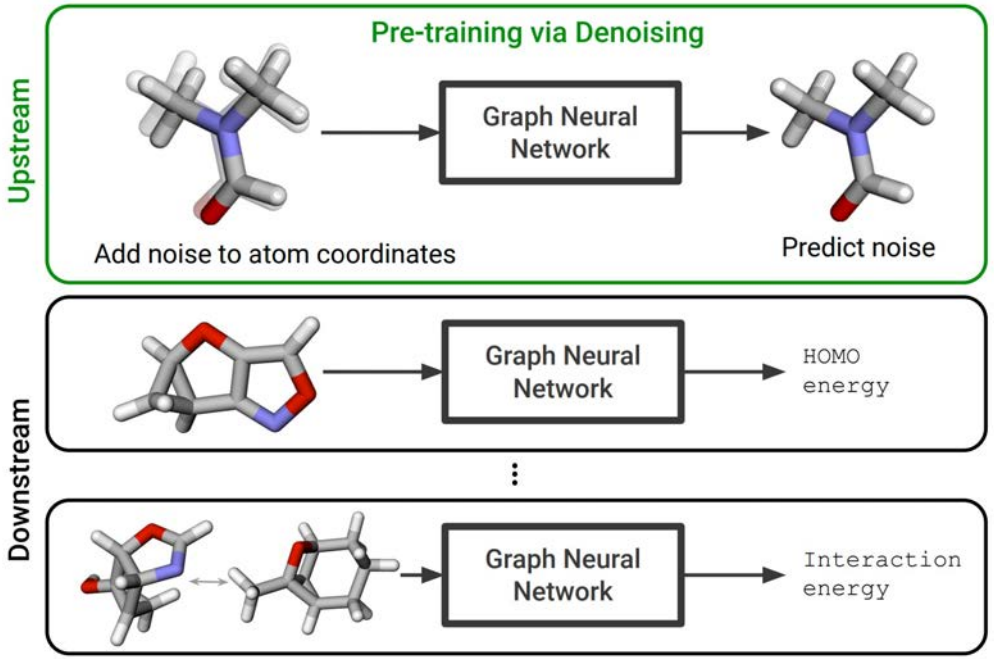
CV



无监督预训练在NLP, CV都取得了显著进展，
在分子领域如何设计预训练策略？

Method

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



Diffusion Pretrain

- add noise: $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\ \mathbf{x}_0, \sigma_t \mathbf{I})$
- estimate noise: $\sum_{t=0}^{T} \lambda'(t)\ \ \mathbb{E}_{N(\boldsymbol{\epsilon};\ 0,1)} \left|\left| \frac{s_{\boldsymbol{\theta}}(\mathbf{x}_t, t)}{\sigma_t} + \boldsymbol{\epsilon} \right|\right|^2$

为大量无标注分子做Diffusion预训练

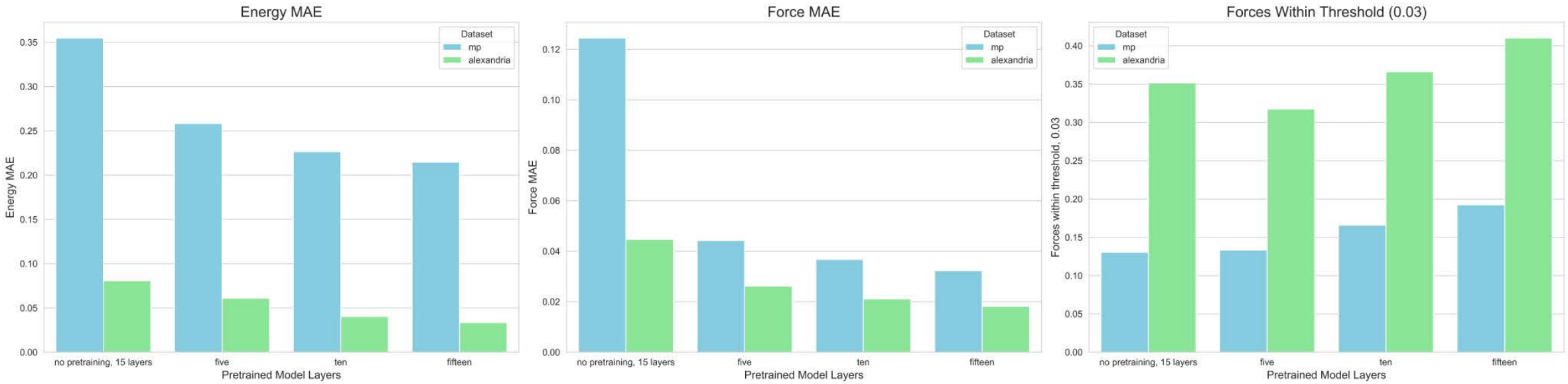
为方便扩充分子尺度只基于普通GNN架构

Result

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

Model	CPS ↑	Acc ↑	F1 ↑	DAF ↑	Prec ↑	MAE ↓	R^2 ↑	κ_{SRME} ↓	RMSD ↓
eSEN-30M-OAM	0.888	0.977	0.925	6.069	0.928	0.018	0.866	0.170	0.061
ORB v3	0.861	0.971	0.905	5.912	0.904	0.024	0.821	0.210	0.075

Rank #2 on Matbench discovery

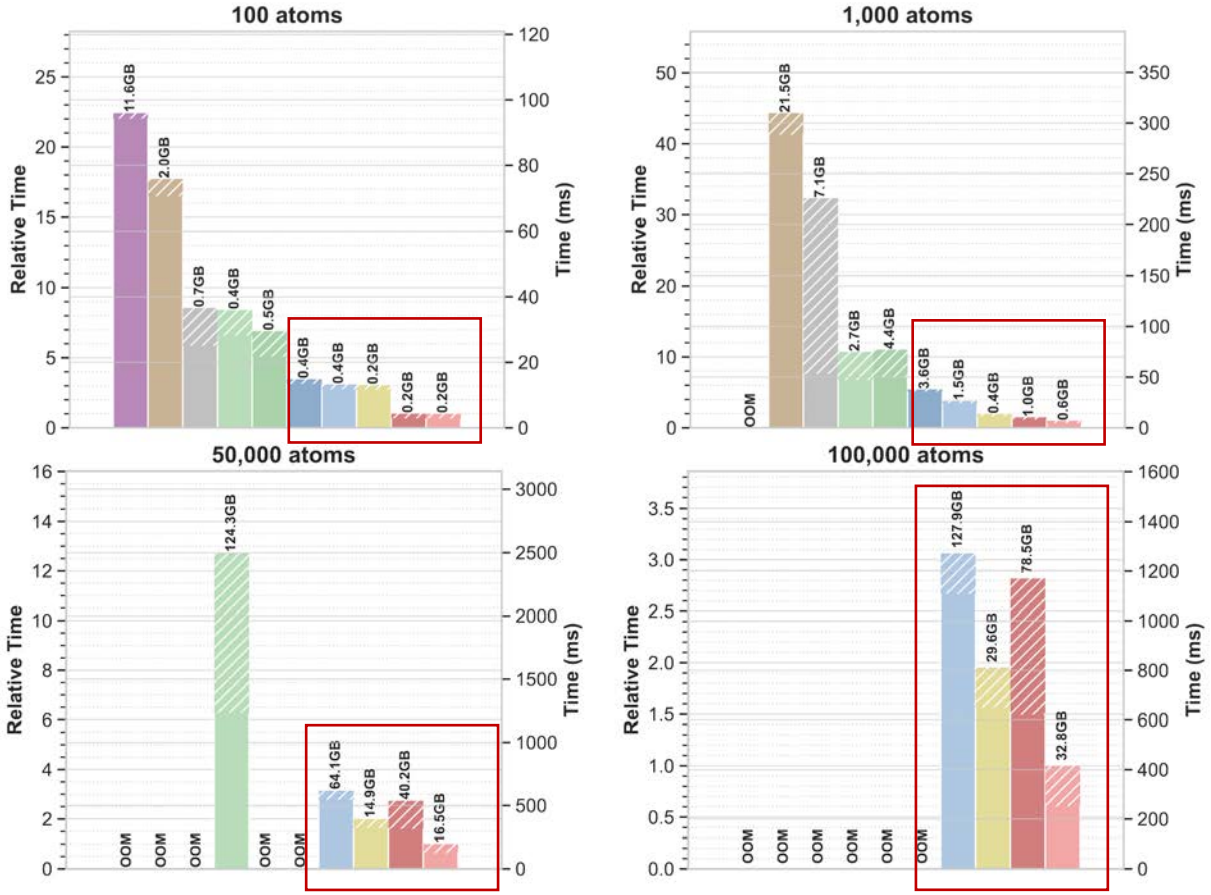


预训练对下游任务能够显著提升

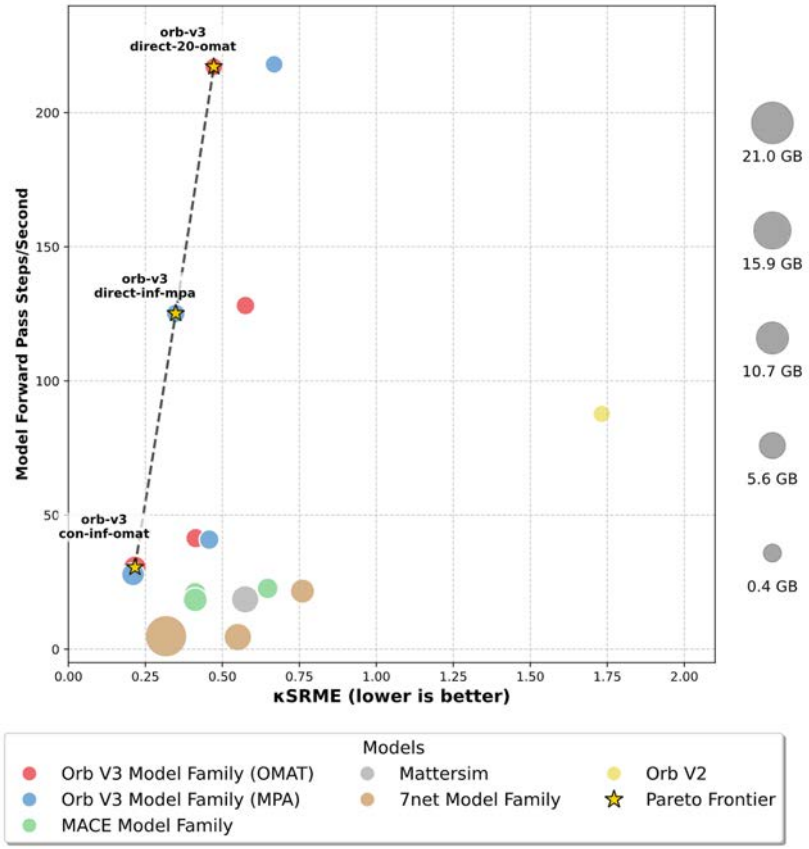
Result

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

Orb能够在更短的时间和更少的GPU内存内运行更大的分子系统。



Orb扩展了性能、速度和内存的帕累托前沿。



大纲

- 背景与动机
- 相关工作
  - ✓ 材料表示学习
  - ✓ 材料生成
- 团队工作
  - ✓ 从数据结构出发，找到本征的数据刻画空间
  - ✓ 从生成算法出发，设计适配分子的生成模型
  - ✓ 从基座构建出发，建立富含广袤数据知识的预训练基座
- 总结

材料生成：学习数据分布

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

$$p_\theta(x) \longleftrightarrow p_{data}(x)$$

生成模型 鸟瞰

共同目标

$$p_\theta(x) \longleftrightarrow p_{data}(x)$$

概率建模

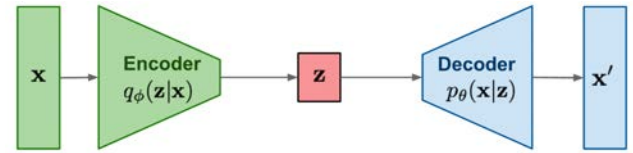
Auto-regressive Model (GPT)

$$x = [x_1, x_2, x_3 \ldots, x_n]$$

$$p_\theta = \prod_{i=1}^{n} p_\theta(x_i | x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} p_\theta(x_i | x_{<i})$$

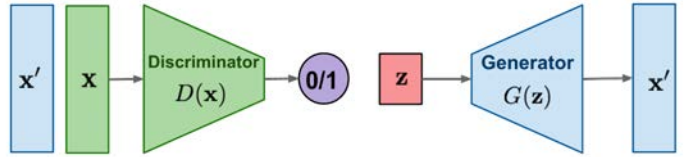
VAE

$$p_\theta(x) = \int_z p(x|z)p(z)$$



GAN

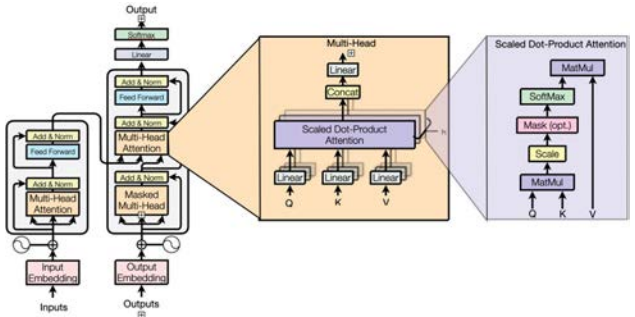
$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$
$$= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x)]$$



Diffusion

$$-\log p_\theta(\mathbf{x}_0) \leq -\log p_\theta(\mathbf{x}_0) + D_{\mathrm{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0))$$

参数化



Transformer、GNN等等

现有工作

面向晶体的生成式建模

- 面向一般晶体
  - ✓Mattergen[1], DiffCSP[2], DiffCSP++[3]…
- 面向大晶体
  - ✓MOFDiff[4], MOFFlow[5]…

基于LLM的通用建模

- NatureLM[6], Crystal LLM[7] …

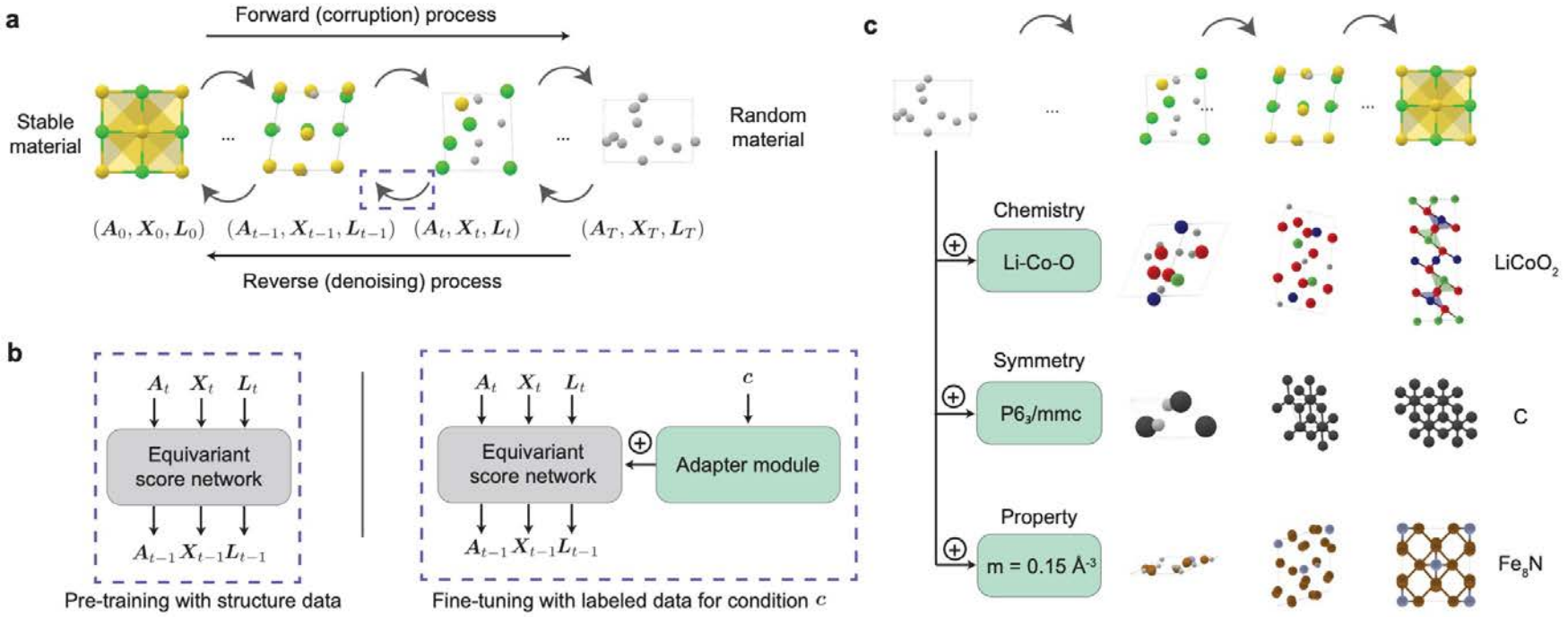
[1] Mattergen: a generative model for inorganic materials design
[2] Crystal structure prediction by joint equivariant diffusion
[3] Space Group Constrained Crystal Generation
[4] Mofdiff: Coarse-grained diffusion for metal-organic framework design
[5] MOFFlow: Flow Matching for Structure Prediction of Metal-Organic Frameworks
[6] Nature Language Model: Deciphering the Language of Nature for Scientific Discovery
[7] FINE-TUNED LANGUAGE MODELS GENERATE STABLEINORGANIC MATERIALS AS TEXT

面向晶体的生成式建模--一般晶体

Mattergen

Mattergen: a generative model for inorganic materials design

Mattergen

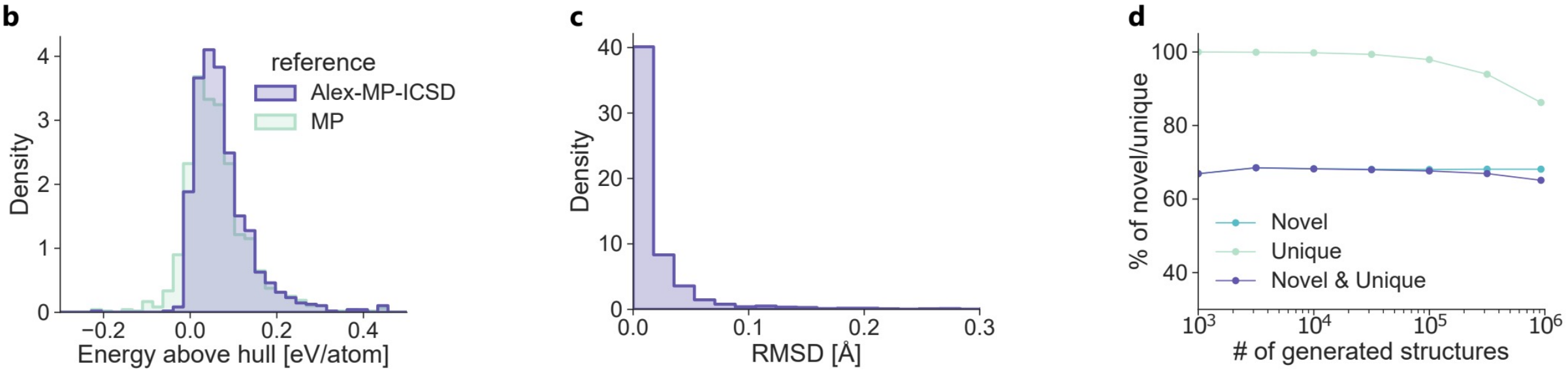


预训练：基于Diffusion做无监督生成

调优：基于adapter做定向生成

Microsoft. Mattergen: a generative model for inorganic materials design[J]. Nature.

Result

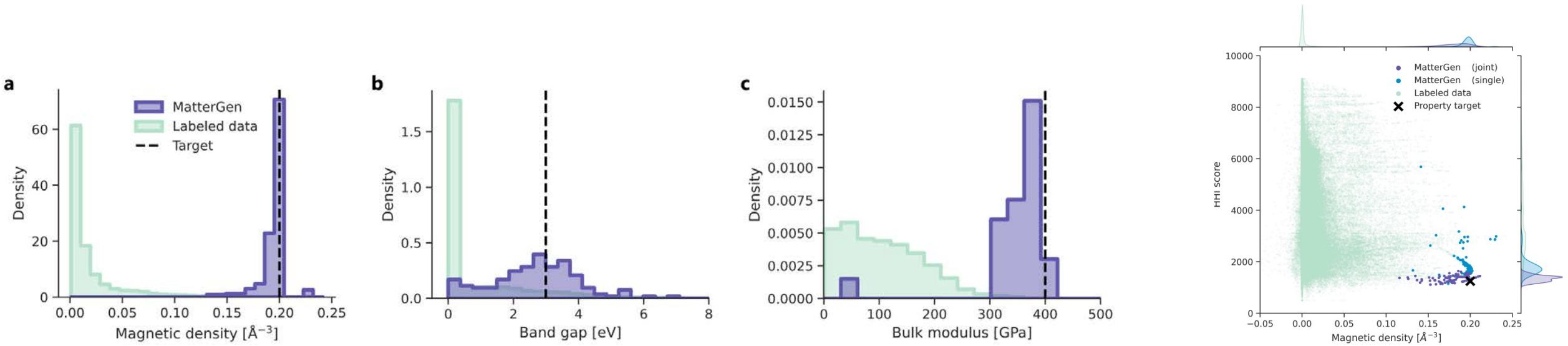


无条件生成：符合数据分布；生成的结构既新颖又独特

Microsoft. Mattergen: a generative model for inorganic materials design[J]. Nature.

Result

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



有条件生成：具有单目标、多目标优化的能力

Microsoft. Mattergen: a generative model for inorganic materials design[J]. Nature.

面向晶体的生成式建模-大晶体

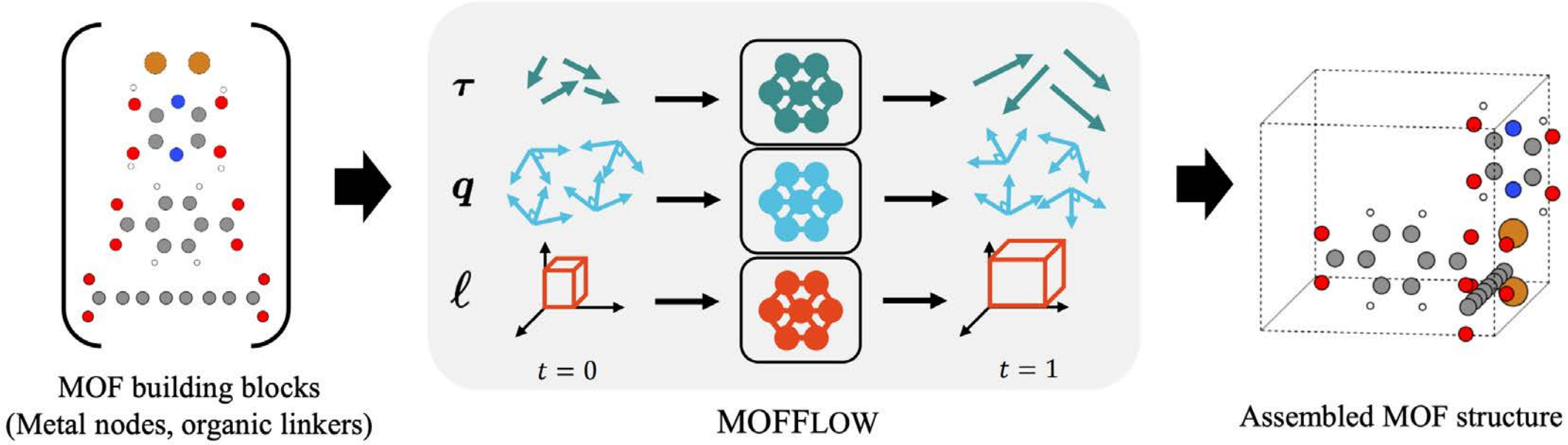
MOFFlow

KAIST

MOFFlow: Flow Matching for Structure Prediction of Metal-Organic Frameworks

MOFFlow

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



引入building blocks(bb)的概念进行层次化建模：
bb的坐标(连续)，转向(SO(3))，晶格矩阵(连续)

Result

	# of samples	stol = 0.5		stol = 1.0		Avg. time (s)↓
		MR (%) ↑	RMSE ↓	MR (%) ↑	RMSE ↓	
RS (Yamashita et al., 2021)	20	0.00	-	0.00	-	332
EA (Yamashita et al., 2021)	20	0.00	-	0.00	-	1959
DiffCSP (Jiao et al., 2024a)	1	0.09	0.3961	23.12	0.8294	5.37
	5	0.34	0.3848	38.94	0.7937	26.85
MOFFLOW (Ours)	1	31.69	0.2820	87.46	0.5183	**1.94**
	5	**44.75**	**0.2694**	**100.0**	**0.4645**	5.69

相比atom-level的方法，性能得到显著提升

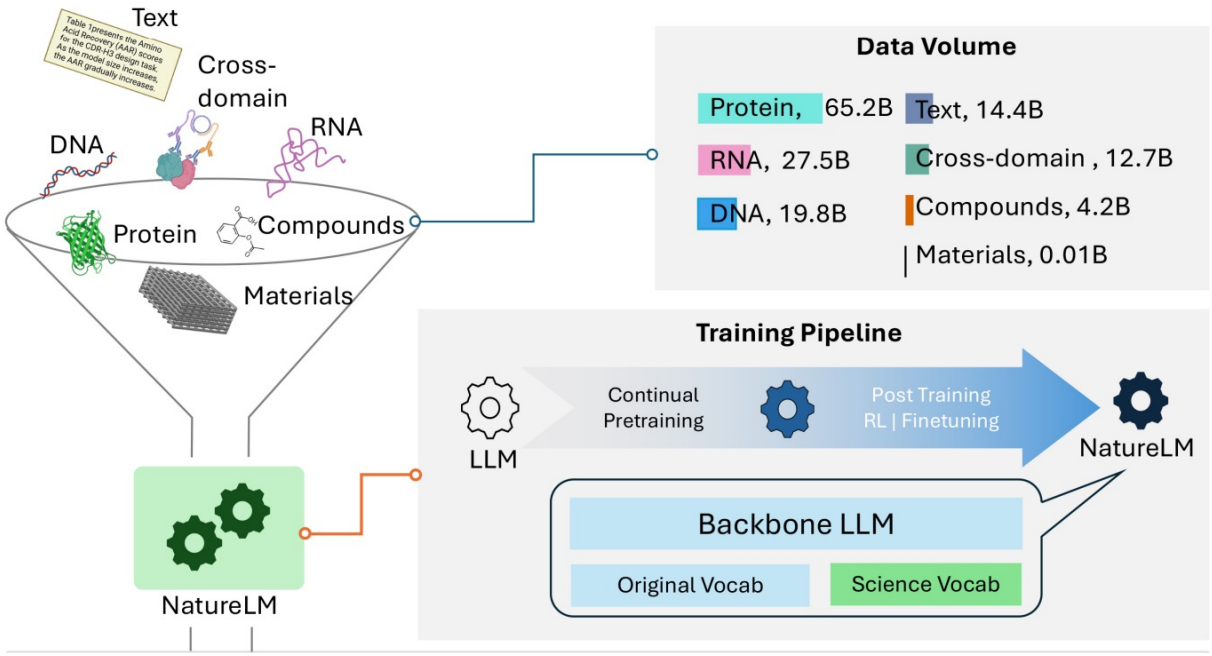
基于LLM的通用建模

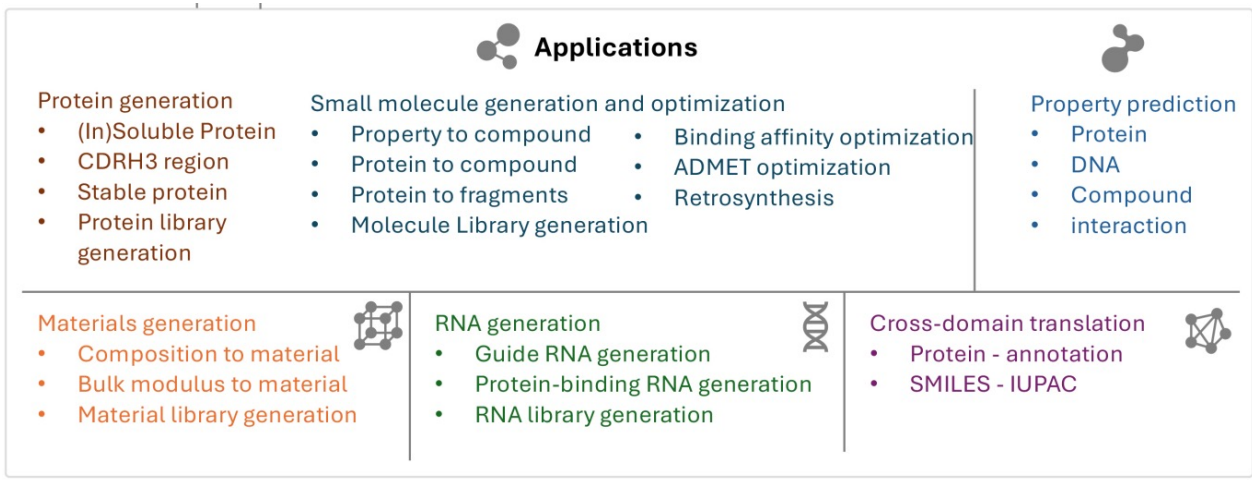
NatureLM

Nature Language Model: Deciphering the Language of Nature for Scientific Discovery

NatureLM

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University





NatureLM是一种类GPT的生成式模型，其训练数据涵盖了广泛的内容，包括小分子化合物、蛋白质、DNA、RNA、材料以及文本

Microsoft. Nature Language Model: Deciphering the Language of Nature for Scientific Discovery. Arxiv

Result

	Perov-5		MP-20		MPTS-52	
	MR (%)	RMSE	MR (%)	RMSE	MR (%)	RMSE
CDVAE	45.31	0.1138	33.90	0.1045	5.34	0.2106
DiffCSP	52.02	**0.0760**	51.49	0.0631	12.19	0.1786
FlowMM	**53.15**	0.0992	61.39	0.0566	17.54	0.1726
NatureLM-Mat3D (1B)	50.78	0.0856	**61.78**	**0.0436**	**30.20**	**0.0837**

Table 15: The match rate (MR) and RMSE on Perov-5, MP-20 and MPTS-52.

材料生成任务上与为材料数据定制的生成模型表现相当，甚至更好

Microsoft. Nature Language Model: Deciphering the Language of Nature for Scientific Discovery. Arxiv

Result

Model	MAE ↓
Dummy[97]	1.1435
gptchem[99]	0.4544
RF-SCM/ Magpie[97]	0.4461
AMMExpress[97]	0.4161
MODNet[100]	0.3327
Ax/SAASBO CrabNet[101, 102]	0.3310
DARWIN[98]	0.2865
NatureLM	**0.2858**

Table 24: Results on matbench_expt_gap.

Model	F1 ↑
Dummy[97]	0.4913
gptchem[99]	0.8953
MODNet[100]	0.9153
RF-SCM/ Magpie[97]	0.9159
AMMExpress[97]	0.9200
DARWIN[98]	0.9599
NatureLM	**0.9630**

Table 25: Results on matbench_is_metal.

Model	F1 ↑
Dummy[97]	0.7127
DARWIN[98]	0.8722
gptchem[99]	0.8782
RF-SCM/ Magpie[97]	0.9278
AMMExpress[97]	0.9043
MODNet[100]	**0.9784**
NatureLM	0.8720

Table 26: Results on matbench_glass.

在属性预测任务上也能取得最好，或有竞争力的结果

Microsoft. Nature Language Model: Deciphering the Language of Nature for Scientific Discovery. Arxiv

第三部分

团队工作

从模型、基座两个角度展开探索

大纲

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

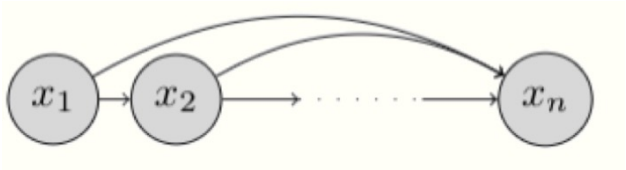
- 背景与动机
- 相关工作
  - ✓ 材料表示学习
  - ✓ 材料生成
- 团队工作
  - ✓ 从生成算法出发，设计适配分子的生成模型
  - ✓ 从基座构建出发，建立富含广袤数据知识的材料基座
- 总结

分子生成挑战

No Intrinsic Orders

Text data

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies, justo ac viverra euismod, justo odio eleifend dolor, a imperdiet quam nibh finibus mauris. Morbi lobortis a lorem id dapibus. Interdum et malesuada fames...



$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|x_1, x_2, \ldots, x_{i-1})$$

Autoregressive Models

The information could be decomposed with the left-to-right order: the meaning of a current word depends on what came before.

ChatGPT

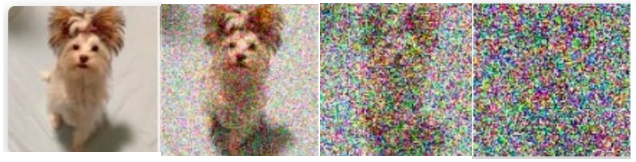
Bard

?

?

分子生成挑战

Image data



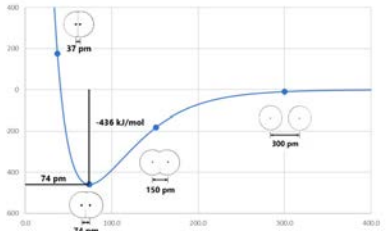
The information in image data could be decomposed by adding multi-level noises which results to a coarse-to-fine modeling order

Structure Constraint

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Diffusion Models

OpenAI Sora

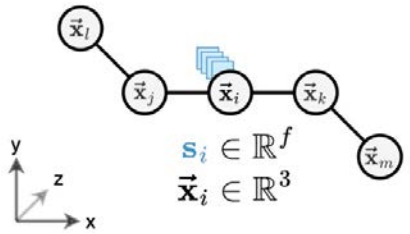


Curves of bond length and energy of H-H

Structure information is very sensitive to perturbation

分子生成挑战

Multi-modality



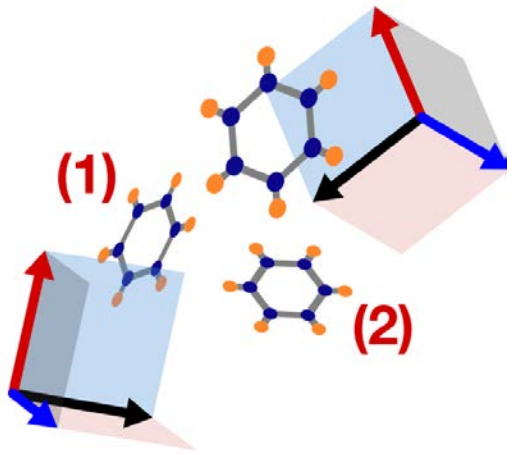
Coordinates variable x is continuous: e.g $\begin{bmatrix} 0.1 \\ -0.1 \\ -0.3 \end{bmatrix}$

The features s could lie in different data types:

Hydrogen ⟶ $[1,0,0,0,0]$ / $[1]$

Discrete atom types　　Discretised Charges

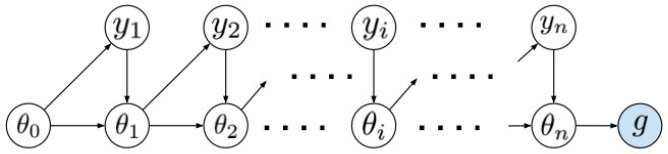
Geometric Symmetry



The learned density function should be roto-translational invariant.

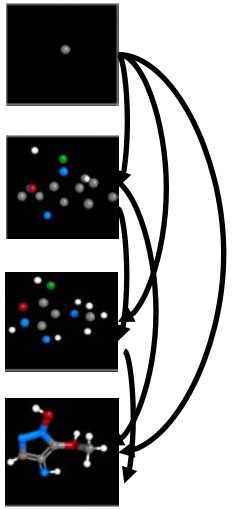
GeoBFN

GeoBFN:

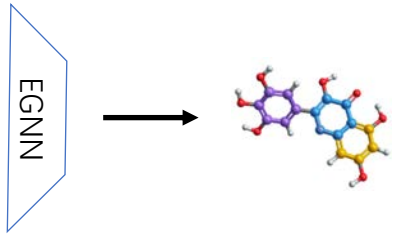


A Latent Variable Model with Non-Markov Autoregressive Prior

Autoregressively in latent variable like LLMs.



Jointly modeling all dimensions like Diffusion models.



Interdependency between all atoms in the molecule are modeled

Specific Advantages for Geometric Graph:

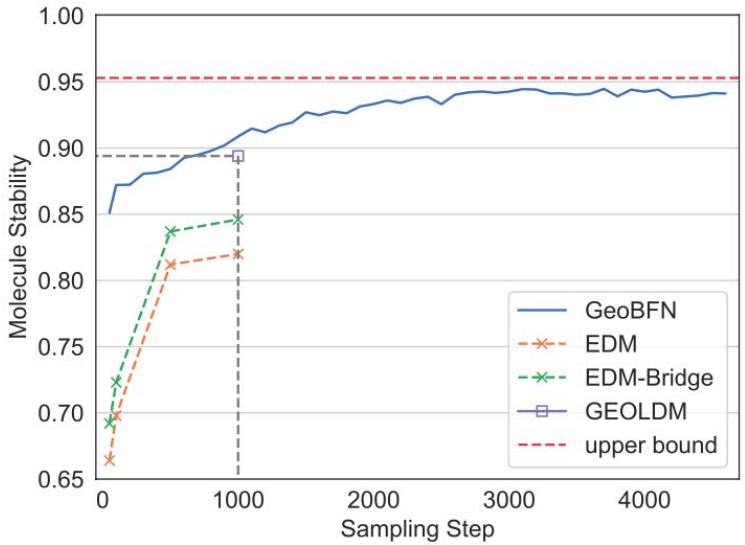
- **Multi-modality:** Different Modalities could all operate in the continuous θ space.
- **Structure Constraint:** the continuous θ space holds much less variance compared to sample space of diffusion models.
- **Geometric Symmetry:** with equivariant networks, the density modeled by GeoBFN is SE(3)-invariant

第一个在分布参数空间而不是数据空间加噪的分子生成模型

GeoBFN结果

# Metrics	QM9 Atom Sta (%)	QM9 Mol Sta (%)	QM9 Valid (%)	QM9 V×U (%)	QM9 Novelty	DRUG Atom Sta (%)	DRUG Valid (%)
Data	99.0	95.2	97.7	97.7	-	86.5	99.9
ENF	85.0	4.9	40.2	39.4	-	-	-
G-Schnet	95.7	68.1	85.5	80.3	-	-	-
GDM-AUG	97.6	71.6	90.4	89.5	74.6	77.7	91.8
EDM	98.7	82.0	91.9	90.7	58.0	81.3	92.6
EDM-Bridge	98.8	84.6	92.0	90.7	-	82.4	92.8
GEOLDM	98.9 ± 0.1	89.4 ± 0.5	93.8 ± 0.4	92.7 ± 0.5	57.0	84.4	**99.3**
GEOBFN $_{50}$	98.28 ± 0.1	85.11 ± 0.5	92.27 ± 0.4	90.72 ± 0.3	72.9	75.11	91.66
GEOBFN $_{100}$	98.64 ± 0.1	87.21 ± 0.3	93.03 ± 0.3	91.53 ± 0.3	70.3	78.89	93.05
GEOBFN $_{500}$	98.78 ± 0.8	88.42 ± 0.2	93.35 ± 0.2	91.78 ± 0.2	67.7	81.39	93.47
GEOBFN $_{1k}$	**99.08** ± 0.06	**90.87** ± 0.2	**95.31** ± 0.1	**92.96** ± 0.1	**66.4**	**85.60**	92.08
GEOBFN $_{2k}$	**99.31** ± 0.03	**93.32** ± 0.1	**96.88** ± 0.1	**92.41** ± 0.1	**65.3**	**86.17**	91.66

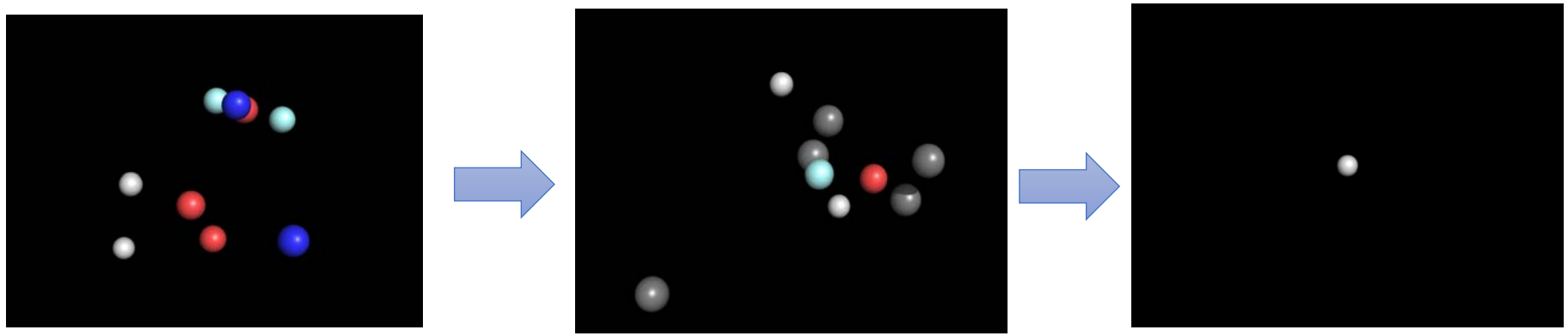
Obtaining the **SOTA** results on Molecule Generation Benchmarks



GeoBFN could sample with any steps:

- Approaching the **empirical limit** with sufficient steps!
- **20 times speed up** for obtaining comparable performance comparing to diffusion models.

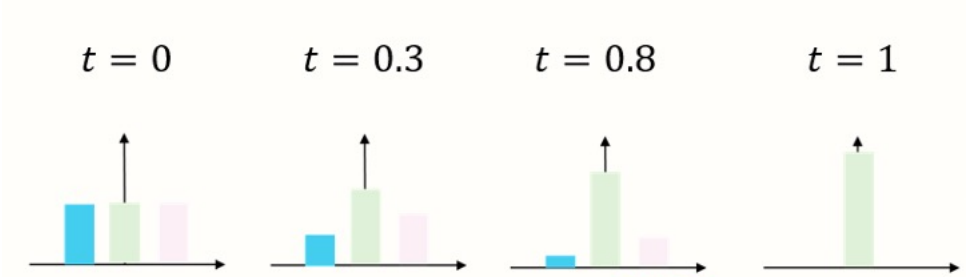
可视化样例



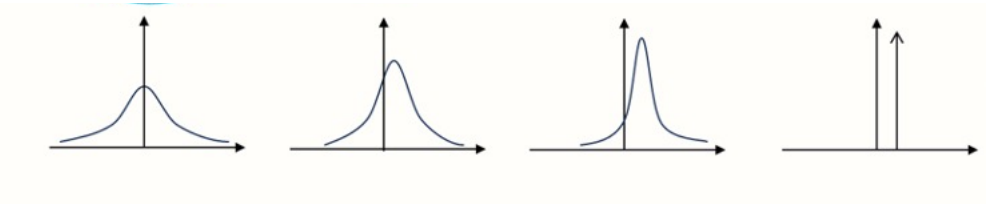
Diffusion Models　　EquiFM　　GeoBFN

CrysBFN

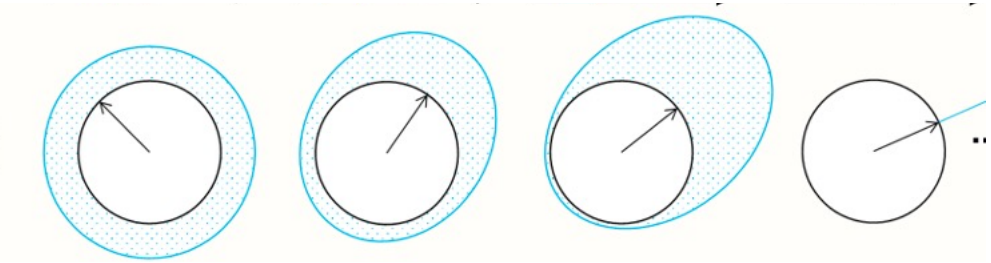
离散原子类型引入类别分布



连续晶格矩阵引入高斯分布



周期原子坐标引入新的冯·米塞斯分布



扩展GeoBFN，使其能够建模分数坐标

CrysBFN效果

Data	Method	Validity (%) ↑ Struc.	Comp.	Coverage (%) ↑ COV-R	COV-P	Property ↓ d_ρ	d_E	d_{elem}
Perov-5	FTCP [6]	0.24	54.24	0.00	0.00	10.27	156.0	0.6297
	Cond-DFC-VAE [9]	73.60	82.95	73.92	10.13	2.268	4.111	0.8373
	G-SchNet [28]	99.92	98.79	0.18	0.23	1.625	4.746	0.0368
	P-G-SchNet [28]	79.63	**99.13**	0.37	0.25	0.2755	1.388	0.4552
	CDVAE [12]	**100.0**	98.59	99.45	98.46	0.1258	0.0264	0.0628
	DiffCSP[13]	**100.0**	98.85	**99.74**	98.27	0.1110	0.0263	0.0128
	CrysBFN	**100.0**	98.86	99.52	**98.63**	**0.0728**	**0.0198**	**0.0098**
Carbon-24	FTCP [6]	0.08	–	0.00	0.00	5.206	19.05	–
	G-SchNet [28]	99.94	–	0.00	0.00	0.9427	1.320	–
	P-G-SchNet [28]	48.39	–	0.00	0.00	1.533	134.7	–
	CDVAE [12]	**100.0**	–	99.80	83.08	0.1407	0.2850	–
	DiffCSP [13]	**100.0**	–	**99.90**	97.27	0.0805	0.0820	–
	CrysBFN	**100.0**	–	**99.90**	**99.10**	**0.0622**	**0.0510**	–
MP-20	FTCP [6]	1.55	48.37	4.72	0.09	23.71	160.9	0.7363
	G-SchNet [28]	99.65	75.96	38.33	99.57	3.034	42.09	0.6411
	P-G-SchNet [28]	77.51	76.40	41.93	99.74	4.04	2.448	0.6234
	CDVAE [12]	**100.0**	86.70	99.15	99.49	0.6875	0.2778	1.432
	DiffCSP[13]	**100.0**	83.25	**99.71**	99.76	0.3502	0.1247	0.3398
	CrysBFN	**100.0**	**87.51**	99.09	**99.79**	**0.2384**	**0.1087**	**0.1905**

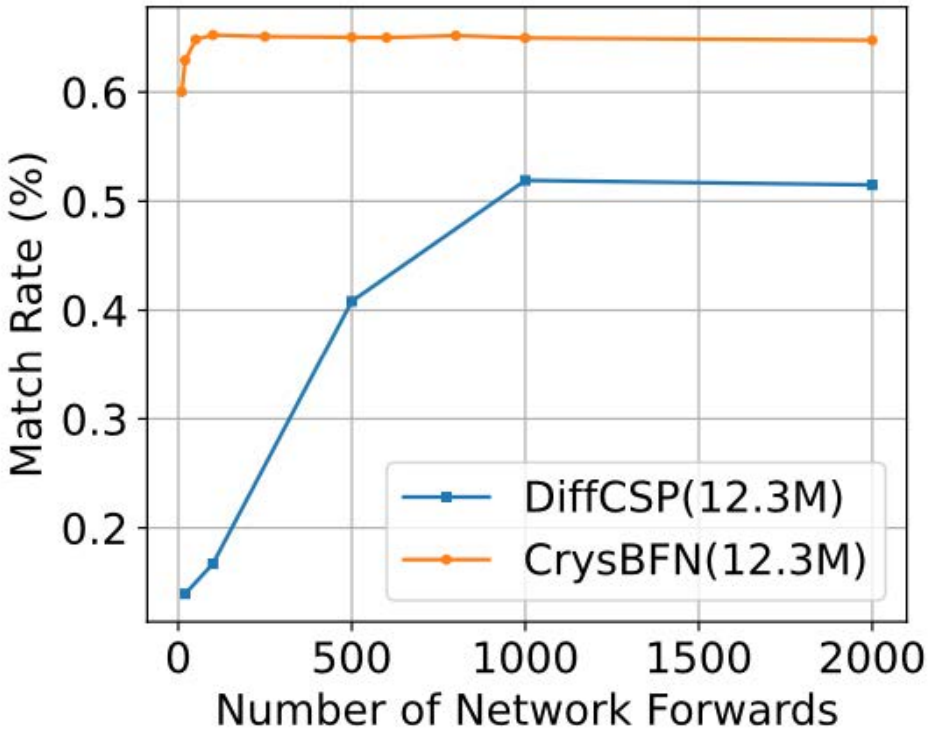
晶体生成任务中，几乎所有指标都达到目前最优效果

CrysBFN效果

	# samples	Perov-5 Match rate↑	Perov-5 RMSE↓	MP-20 Match rate↑	MP-20 RMSE↓	MPTS-52 Match rate↑	MPTS-52 RMSE↓
RS [30]	20	29.22	0.2924	8.73	0.2501	2.05	0.3329
	5,000	36.56	0.0886	11.49	0.2822	2.68	0.3444
BO [30]	20	21.03	0.2830	8.11	0.2402	2.05	0.3024
	5,000	55.09	0.2037	12.68	0.2816	6.69	0.3444
PSO [30]	20	20.90	0.0836	4.05	0.1567	1.06	0.2339
	5,000	21.88	0.0844	4.35	0.1670	1.09	0.2390
P-cG-SchNet [31]	1	48.22	0.4179	15.39	0.3762	3.67	0.4115
	20	97.94	0.3463	32.64	0.3018	12.96	0.3942
CDVAE [12]	1	45.31	0.1138	33.90	0.1045	5.34	0.2106
	20	88.51	0.0464	66.95	0.1026	20.79	0.2085
DiffCSP [13]	1	52.02	0.0760	51.49	0.0631	12.19	0.1786
	20	98.60	0.0128	77.93	0.0492	34.02	0.1749
CrysBFN	1	**52.76**	**0.0695**	**61.06**	**0.0554**	**13.89**	**0.1307**
	20	**98.64**	**0.0116**	**80.72**	**0.0473**	**36.86**	**0.1596**

晶体结构预测任务中，所有指标都达到目前最优效果

CrysBFN效果



仅用10步就能超过之前SOTA方法2000步的性能

BFN for science data

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

GeoBFN

ICLR2024 Oral
Molecular

MolCRAFT

ICML2024 Poster
SBDD

CysBFN

ICLR2025 Spotlight
Material

ProfileBFN

ICLR2025 Oral
Protein Family

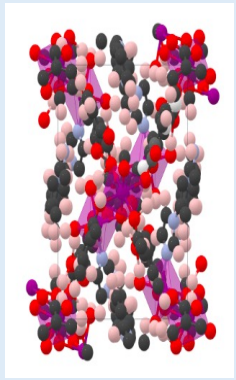
大纲

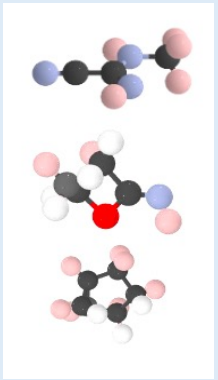
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

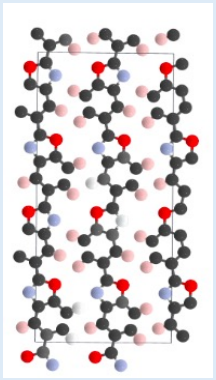
- 背景与动机
- 相关工作
  - ✓ 材料表示学习
  - ✓ 材料生成
- 团队工作
  - ✓ 从生成算法出发，设计适配分子的生成模型
  - ✓ 从基座构建出发，建立富含广袤数据知识的材料基座
- 未来工作
- 总结

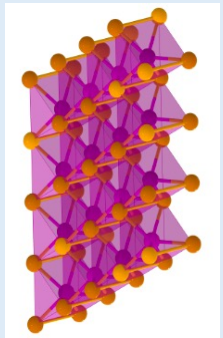
MoMa背景：材料知识的多元异构

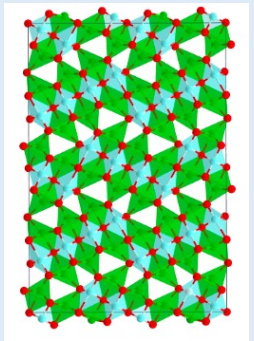
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

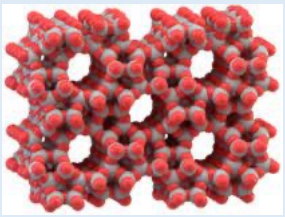

MOF


小分子

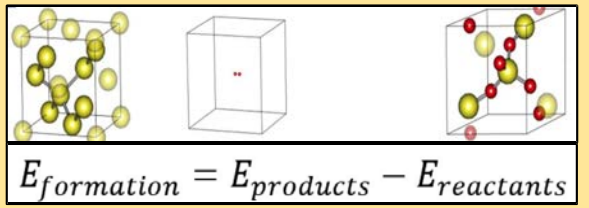

聚合物


2D材料

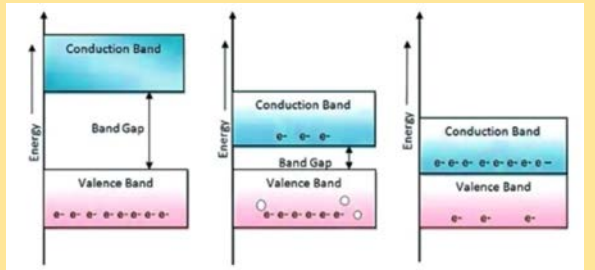

Bulk


催化剂

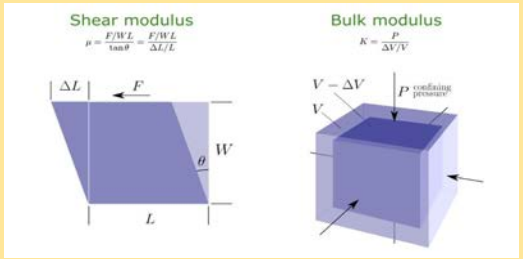
材料数据种类繁杂、结构异构



热力学任务



电学任务



力学任务

材料任务数量众多、领域多元

现有工作

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

现有方法	挑战	具体问题
1. 从头训练	材料标注 少	剥离能 633；形成能 633；介电常数 522；带隙 522 — 训练不充分，知识无法涌现
2. 单任务预训练	材料类型 杂	OLED、催化剂、钙钛矿、MOF … — 模型分散，知识无法共享
3. 多任务预训练	知识共享 难	OLED、MOF — 不加挑选，知识相互冲突

我们提出模块化材料大模型框架，以适配材料知识的多元异构，赋能材料属性预测场景

MoMa架构总览

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

异构材料数据

晶体 | 小分子 | 聚合物 | 催化剂 | ...

提取材料表征

预训练材料表示模型

解耦知识存储

材料知识模块库（>80个模块）

动态装配算法

高效模块装配算法（<30秒）

支撑下游应用

大规模虚拟筛选

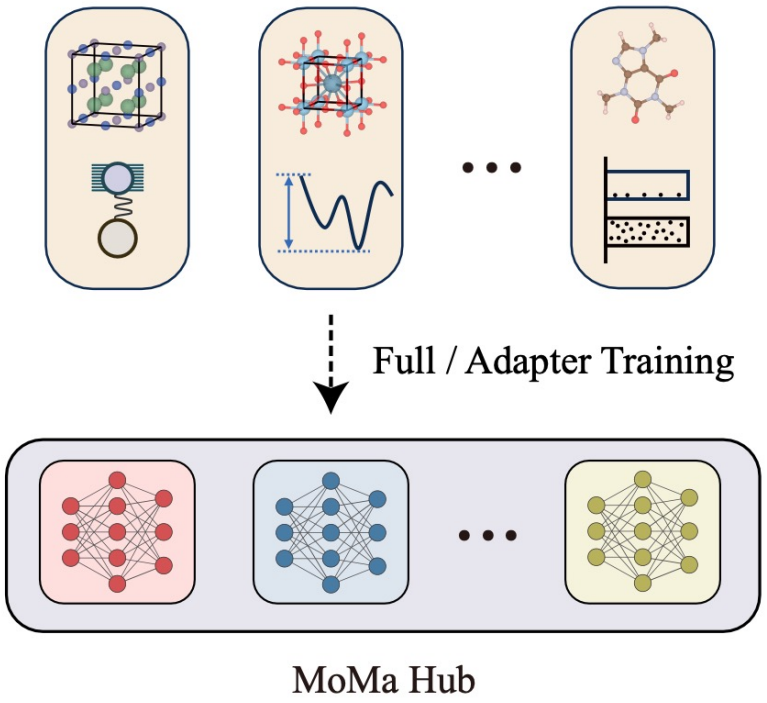
功能材料条件生成

方法总览

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

异构任务解耦训练

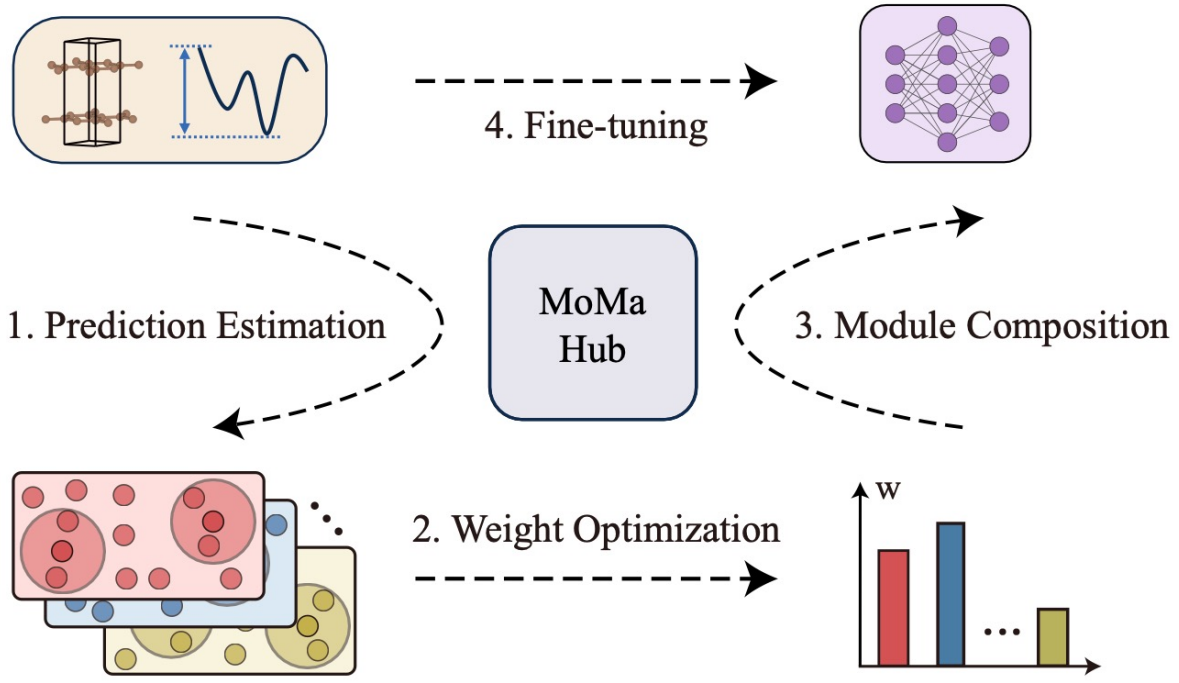
(a) Module Training & Centralization



知识仓库

下游任务适配

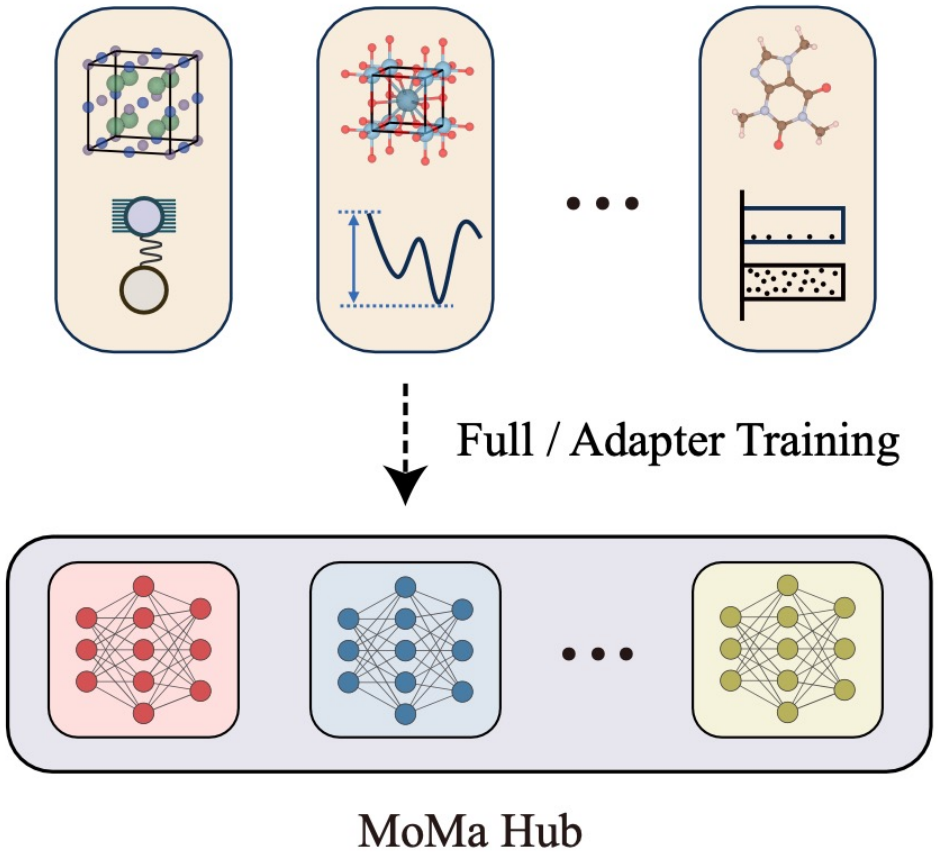
(b) Adaptive Module Composition & Fine-tuning



模块知识解耦

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

以模块化的方式解耦和存储材料知识

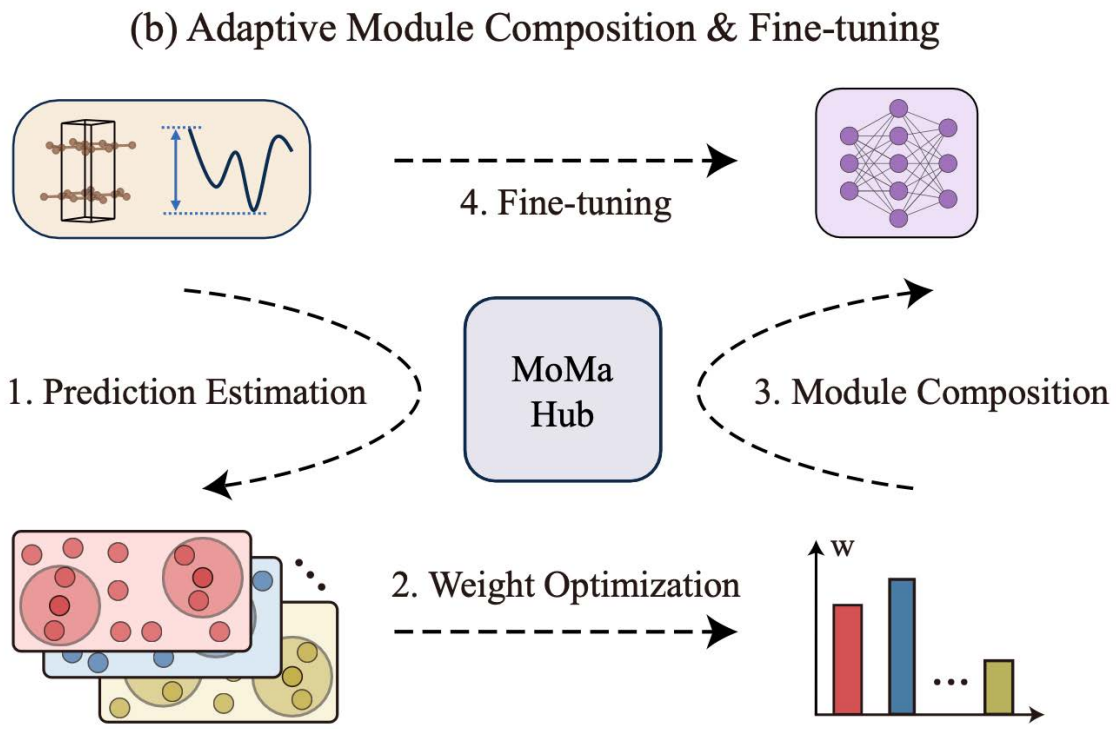


模块化优势

- 细分性：一个任务多个模块
- 轻量性：单个模块参数量<1M
- 广阔性：已解耦>20种材料任务，得到>80个模块
- 可迁移性：所有模块共享编码器
- 可扩展性：为新任务加新模块，不会影响已有模块

模块自适应组装

从知识仓库中自适应挑选模块，服务于下游任务



自适应组装优势

- 自适应性：为不同下游任务自动挑选不同专家
- 高效性：无需训练，30s内完成挑选

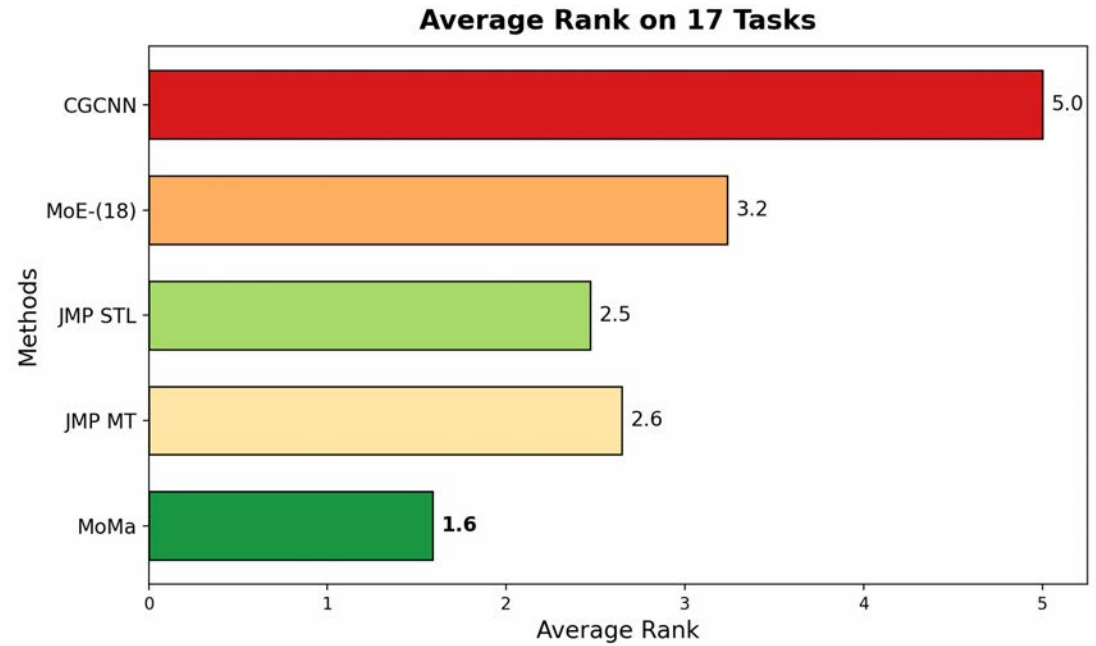
实验

在 Matminer 17个材料属性预测任务上验证了 MoMa 的有效性

MAE of MoMa and Baselines across 17 Tasks

	CGCNN	MoE-(18)	JMP STL	JMP MT	MoMa
expt_bandgap	0.460	0.376	0.383	0.352	0.347
expt_eform	0.194	0.095	0.162	0.161	0.145
jarvis_2d_dielectric_opt	2.710	2.270	2.394	1.866	2.318
jarvis_2d_eform	0.165	0.105	0.140	0.124	0.109
jarvis_2d_exfoliation	62.000	52.800	35.093	45.524	34.287
jarvis_2d_gap_opt	0.693	0.543	0.591	0.467	0.564
jarvis_3d_eps_tbmbj	32.700	28.300	23.564	23.872	23.464
jarvis_3d_gap_tbmbj	0.503	0.353	0.269	0.317	0.247
mp_dielectric	0.086	0.078	0.055	0.061	0.055
mp_elastic_anisotropy	3.690	3.080	2.454	2.703	2.843
mp_eps_electronic	0.170	0.147	0.116	0.115	0.106
mp_eps_total	0.254	0.231	0.172	0.196	0.168
mp_phonons	0.126	0.103	0.076	0.091	0.057
mp_poisson_ratio	0.033	0.029	0.022	0.029	0.022
mp_poly_electronic	2.940	2.700	2.091	2.294	2.199
mp_poly_total	6.400	5.580	4.937	5.328	4.924
piezoelectric_tensor	0.288	0.206	0.179	0.221	0.176

Rank 1
Rank 2
Rank 3
Rank 4
Rank 5

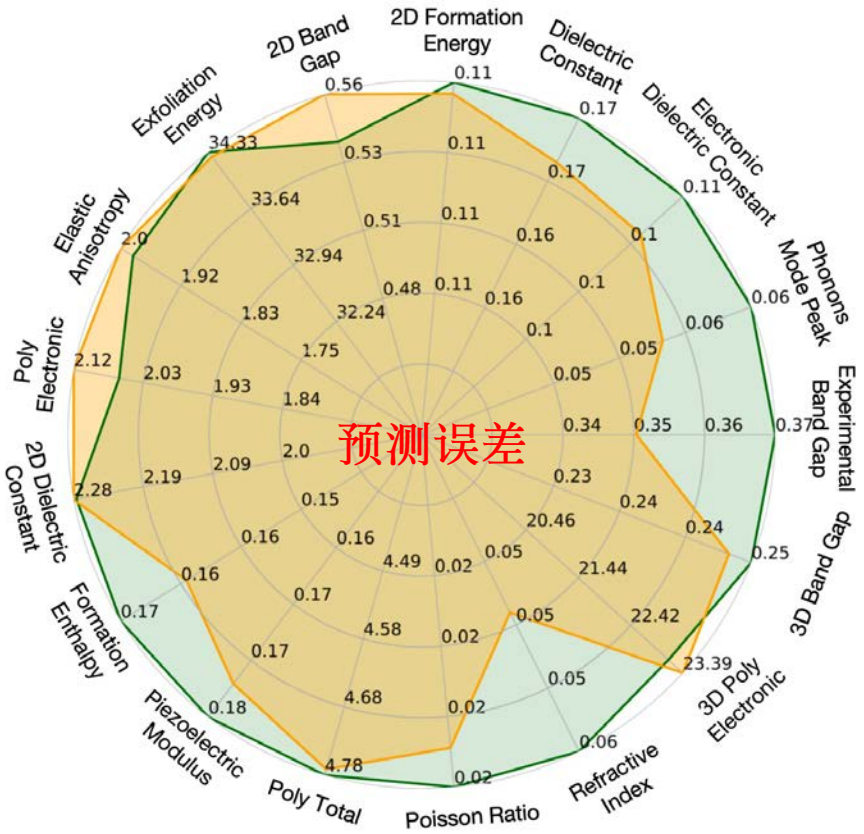


MoMa 在所有任务的平均表现排名最高

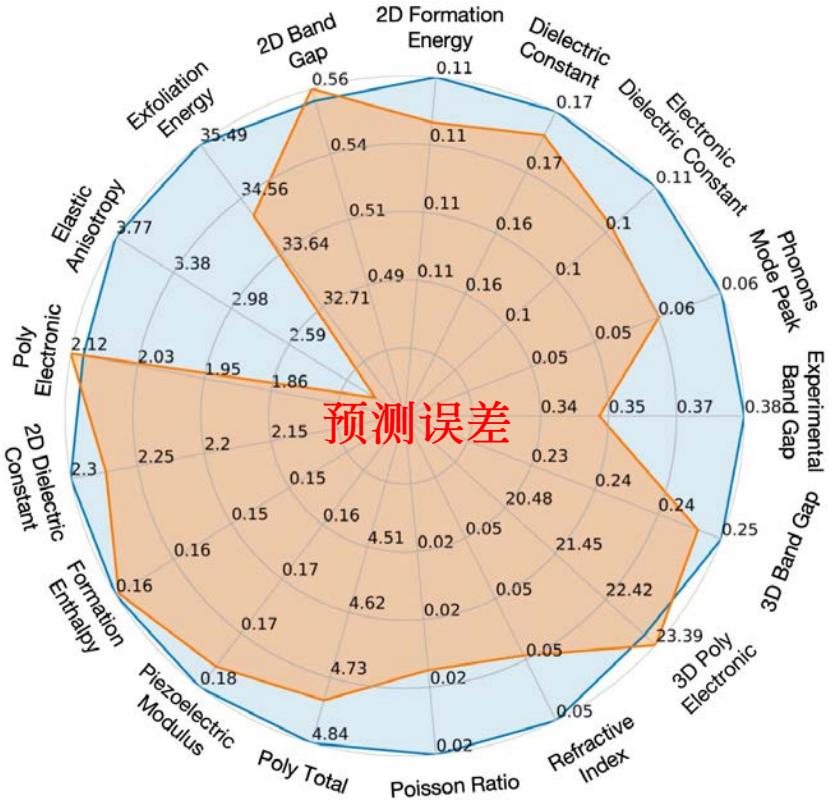
实验效果

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

- 消融实验：模块化与自适应组装均发挥明显作用

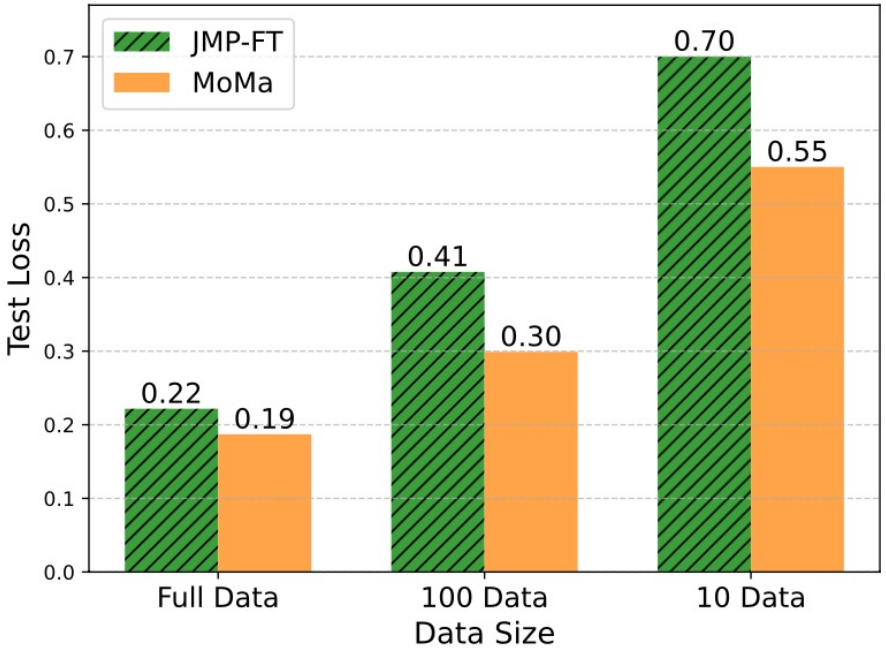


模块化知识解耦相比于无知识解耦，优势明显

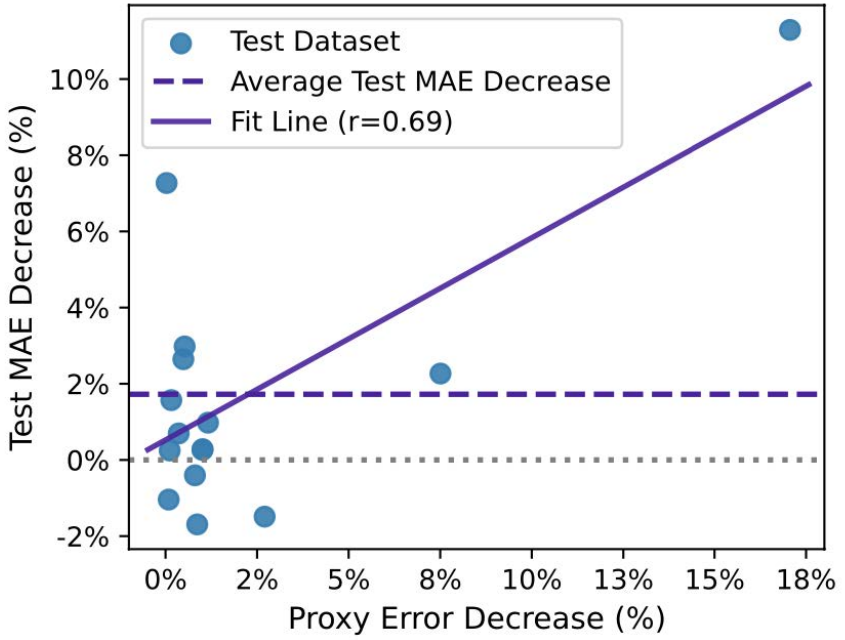


自适应组装相比于随机组装，优势明显

实验效果



低资源场景下MoMa提升更明显



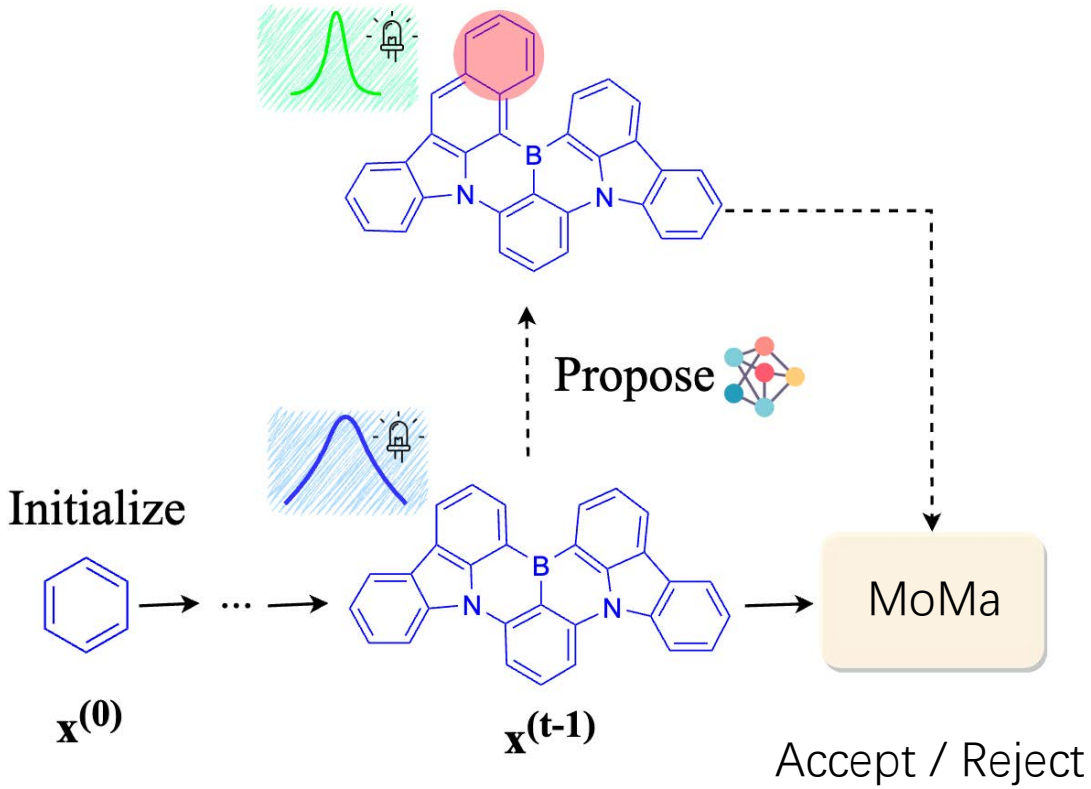
MoMa具有持续学习能力：新加入相关模块后，测试MAE都能得到不同程序降低

适配 OLED 有机发光材料设计

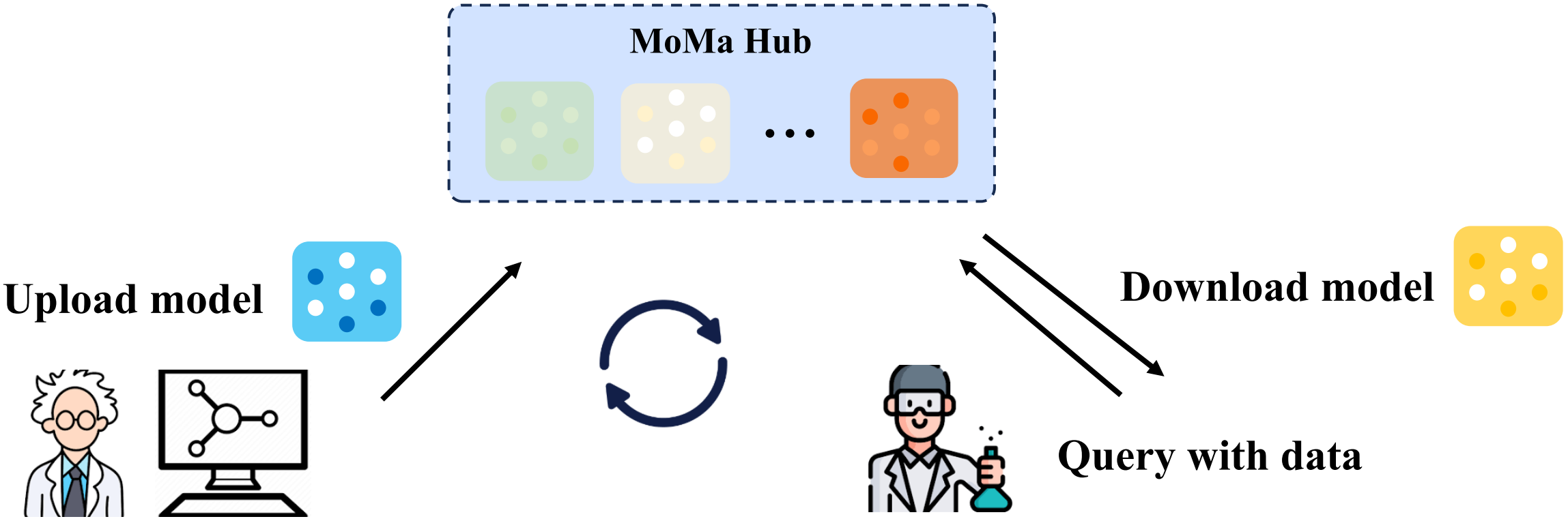
清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

- **下游应用：** 将 MoMa 框架应用到 OLED 材料设计的真实场景

- **方法**
- 基于MCMC框架对OLED分子定向优化
- 使分子波长范围在[500, 550]纳米之间
- **结果**
- 基线方法成功率：19.8%
- 基于 MoMa 的成功率：34.7%



MoMA愿景



我们期望MoMa成为一个知识共享平台，助力材料科学家们共同构建更优质的模型。

第四部分

未来工作

大材料模型

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University

Matbench Discovery

Full Test Set | Unique Prototypes | 10k Most Stable

Click on column headers to sort table rows

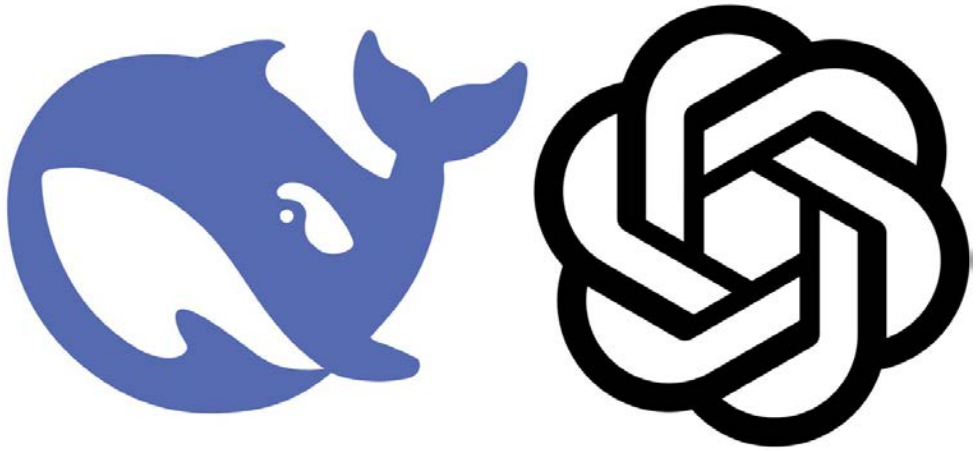
Compliant models | Non-compliant models | Energy-only models | Heatmap | Columns

Model	CPS ↑	Acc ↑	F1 ↑	DAF ↑	Prec ↑	MAE ↓	R^2 ↑	κ_{SRME} ↓	RMSD ↓	Training Set	Params	Targets	Date Added	Links	r_{cut}	Org
eSEN-30M-OAM	0.888	0.977	0.925	6.069	0.928	0.018	0.866	0.170	0.061	6.6M (113M) OMat24+MPtrj+sAl	30.2M	EFS_G	2025-03-17		6 Å	
ORB v3	0.861	0.971	0.905	5.912	0.904	0.024	0.821	0.210	0.075	6.47M (133M) MPtrj+Alex+OMat	25.5M	EFS_G	2025-04-05		6 Å	
SevenNet-MF-ompa	0.845	0.969	0.901	5.825	0.890	0.021	0.867	0.317	0.064	6.6M (113M) OMat24+sAlex+MP	25.7M	EFS_G	2025-03-13		6 Å	
GRACE-2L-OAM	0.837	0.963	0.880	5.774	0.883	0.023	0.862	0.294	0.067	6.6M (113M) OMat24+sAlex+MP	12.6M	EFS_G	2025-02-06		6 Å	ICAMS
eSEN-30M-MP	0.797	0.946	0.831	5.260	0.804	0.033	0.822	0.340	0.075	146k (1.58M) MPtrj	30.1M	EFS_G	2025-03-17		6 Å	
MACE-MPA-0	0.795	0.954	0.852	5.582	0.853	0.028	0.842	0.412	0.073	3.37M (12M) MPtrj+sAlex	9.06M	EFS_G	2024-12-09		6 Å	
MatterSim v1 5M	0.767	0.959	0.862	5.852	0.895	0.024	0.863	0.574	0.073	17M MatterSim	4.55M	EFS_G	2024-12-16		5 Å	
DPA3-v2-OpenLAM	0.762	0.966	0.890	5.747	0.879	0.022	0.869	0.687	0.068	163M OpenLAM	7.02M	EFS_G	2025-03-14		6 Å	
GRACE-1L-OAM	0.761	0.944	0.824	5.255	0.803	0.031	0.842	0.516	0.072	6.6M (113M) OMat24+sAlex+MP	3.45M	EFS_G	2025-02-06		6 Å	ICAMS
SevenNet-l3i5	0.714	0.920	0.760	4.629	0.708	0.044	0.776	0.550	0.085	146k (1.58M) MPtrj	1.17M	EFS_G	2024-12-10		5 Å	
MatRIS v0.5.0 MPtrj	0.681	0.938	0.809	5.049	0.772	0.037	0.803	0.861	0.077	146k (1.58M) MPtrj	5.83M	EFS_GM	2025-03-13		6 Å	
GRACE-2L-MPtrj	0.681	0.896	0.691	4.163	0.636	0.052	0.741	0.525	0.090	146k (1.58M) MPtrj	15.3M	EFS_G	2024-11-21		6 Å	ICAMS
DPA3-v2-MPtrj	0.646	0.929	0.786	4.822	0.737	0.039	0.804	0.959	0.082	146k (1.58M) MPtrj	4.92M	EFS_G	2025-03-14		6 Å	
MACE-MP-0	0.644	0.878	0.669	3.777	0.577	0.057	0.697	0.647	0.091	146k (1.58M) MPtrj	4.69M	EFS_G	2023-07-14		6 Å	
AlphaNet-MPTrj	0.566	0.933	0.799	4.863	0.743	0.041	0.745	1.310	0.107	146k (1.58M) MPtrj	16.2M	EFS_G	2025-03-05		6 Å	
eqV2 M	0.558	0.975	0.917	6.047	0.924	0.020	0.848	1.771	0.069	3.37M (102M) OMat24+MPtrj	86.6M	EFS_D	2024-10-18		12 Å	
ORB v2	0.529	0.965	0.880	6.041	0.924	0.028	0.824	1.732	0.097	3.25M (32.1M) MPtrj+Alex	25.2M	EFS_D	2024-10-11		10 Å	
eqV2 S DeNS	0.522	0.941	0.815	5.042	0.771	0.036	0.788	1.676	0.076	146k (1.58M) MPtrj	31.2M	EFS_D	2024-10-18		12 Å	
ORB v2 MPtrj	0.470	0.922	0.765	4.702	0.719	0.045	0.756	1.725	0.101	146k (1.58M) MPtrj	25.2M	EFS_D	2024-10-14		10 Å	
M3GNet	0.428	0.813	0.569	2.882	0.441	0.075	0.585	1.412	0.112	62.8k (188k) MPF	228k	EFS_G	2022-09-20		5 Å	
CHGNet	0.400	0.851	0.613	3.361	0.514	0.063	0.689	1.717	0.095	146k (1.58M) MPtrj	413k	EFS_GM	2023-03-03		5 Å	Berkeley
GNoME	NaN	0.955	0.829	5.523	0.844	0.035	0.785	n/a	n/a	6M (89M) GNoME	16.2M	EF_G	2024-02-03		5 Å	

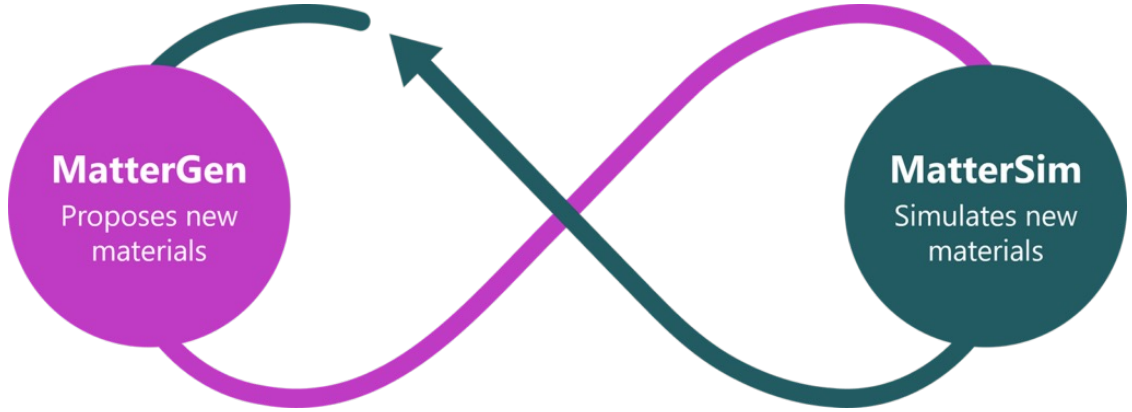
当前材料模型参数量仍比较小，如何scale up？

材料 + LLM

清華大學智能产业研究院
Institute for AI Industry Research, Tsinghua University



LLM：宏观知识



材料模型：微观知识

如何深度结合两者？

材料 + 智能体

清华大学智能产业研究院
Institute for AI Industry Research, Tsinghua University

如何使用Agent助力新材料从提出到落地

第五部分

总结

Takeaways

- 传统材料研发面临周期长、成本高等困境
- AI可加速这一过程
  - ✓ 材料表示学习
  - ✓ 材料生成
- AI面临的挑战
  - ✓ 数据：体系繁杂、标注耗时耗力
  - ✓ 建模：几何约束

- 现有工作
  - ✓ 材料表示学习：优化模型架构、预训练策略
  - ✓ 材料生成：面向晶体的生成式建模、基于LLM的通用建模
- 未来工作
  - ✓ 材料模型的scale up
  - ✓ 材料 + LLM
  - ✓ 材料 + 智能体
  - …

谢谢!