

大模型时代下的AI for Science

王一博, 深势科技

2024.09.21

1.1 AI4S的第一类寻宝图：AI建模高维复杂函数

1.2 AI4S的第二类寻宝图：AI驱动平台化科研

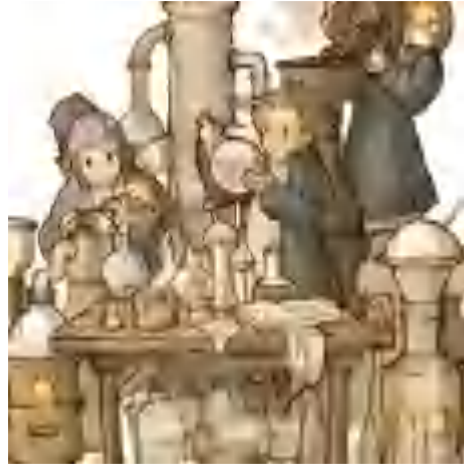
1.3 AI4S该怎么学

前 AI 时代科研体系遇到的困难

“维数灾难”问题



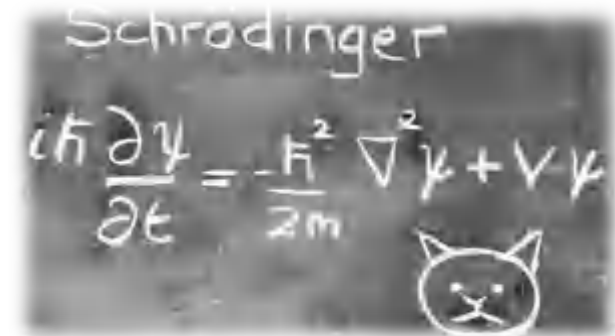
从原始工具到现代科技，从石器时代到AI时代
人类文明的发展离不开工具的变革，而工具的革新，离不开科学技术的发展



第一范式：实验



第二范式：计算



第三范式：理论



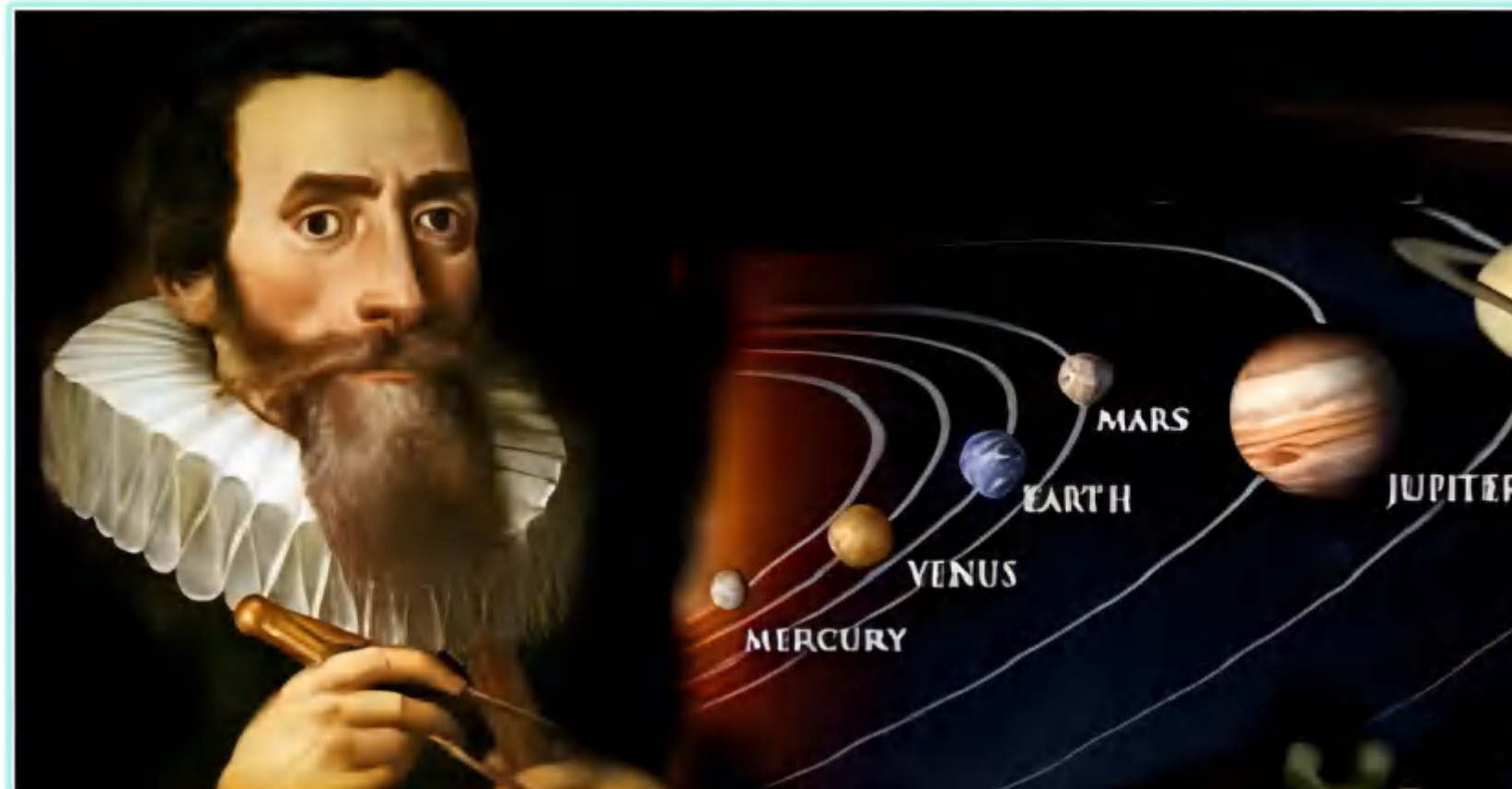
牛顿的三棱镜实验



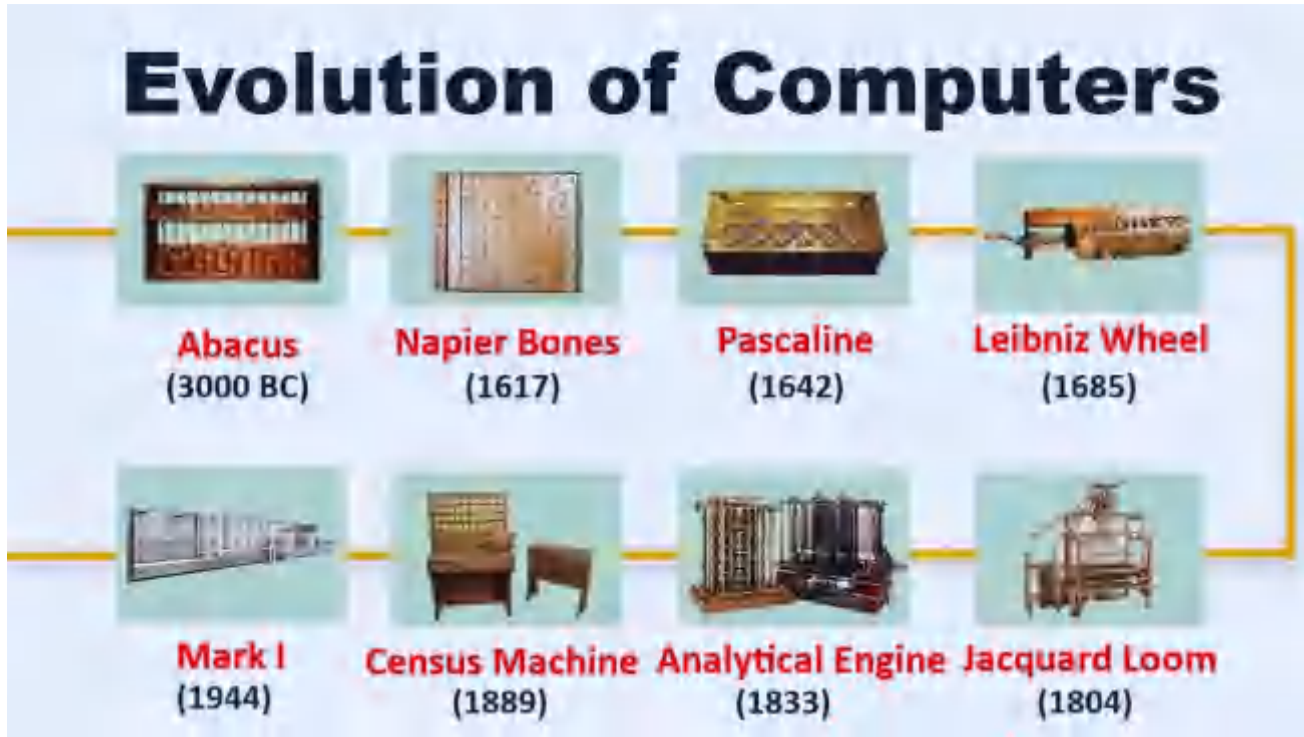
居里夫人发现放射性元素“钋”和“镭”



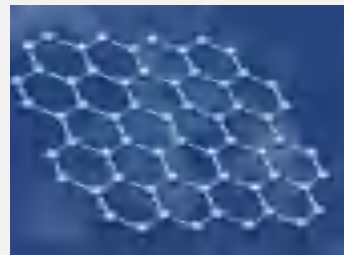
第谷：观测与收集数据



开普勒：分析数据产生科学发现



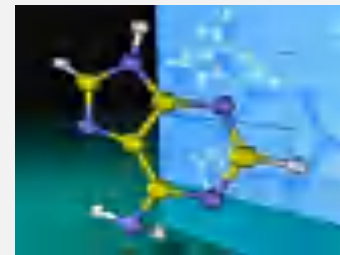
世界上第一台现代电子数字计算机**ENIAC**，诞生于**1946年2月14日**的**美国宾夕法尼亚大学**。占地面积约**170平方米**，重达**30英吨**。它包含了**17468根电子管**，**7200根晶体二极管**。计算速度是**每秒5000次加法**或**400次乘法**，是使用继电器运转的机电式计算机的**1000倍**、手工计算的**20万倍**。用于美国军方弹道轨迹计算和曼哈顿计划。



材料数据



生物数据



药物数据



化工数据

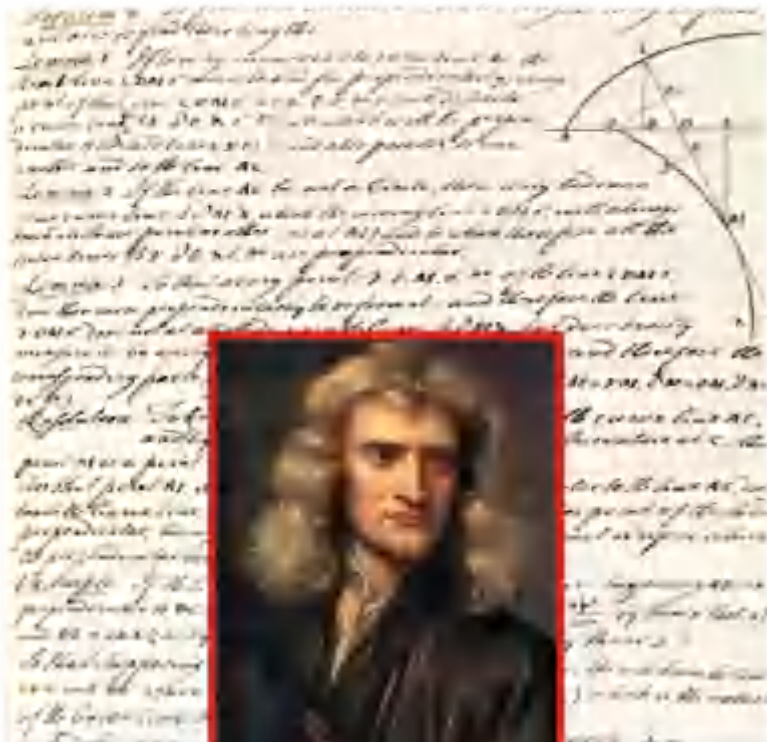
数据收集的效率低下

缺乏有效的数据分析方法

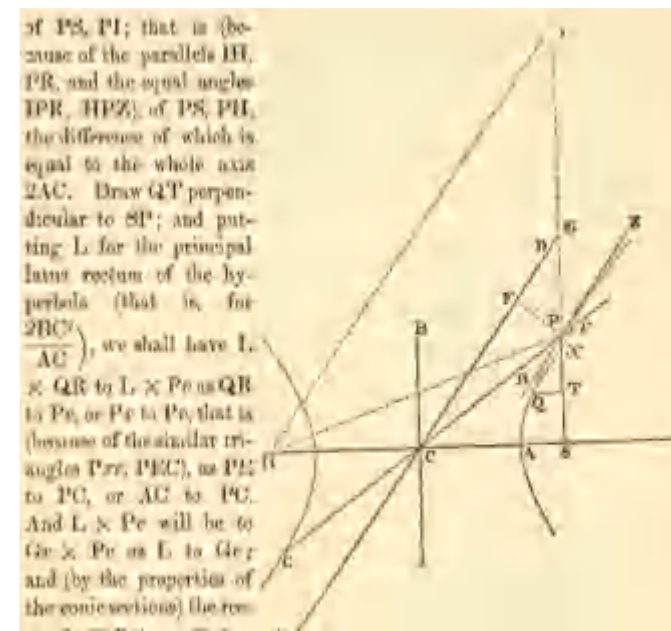
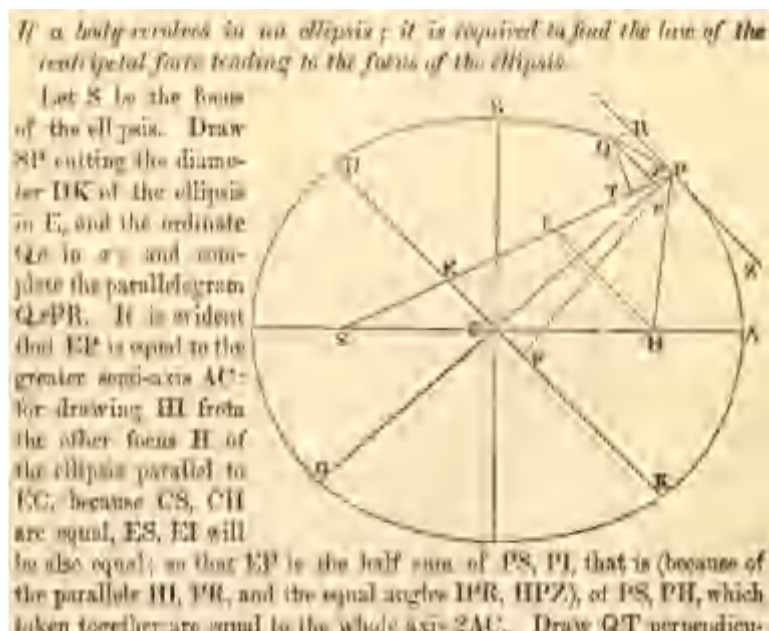
It remains that, from the same principles, I now demonstrate the frame of the System of the World.

现在，我将展示世界体系的框架。

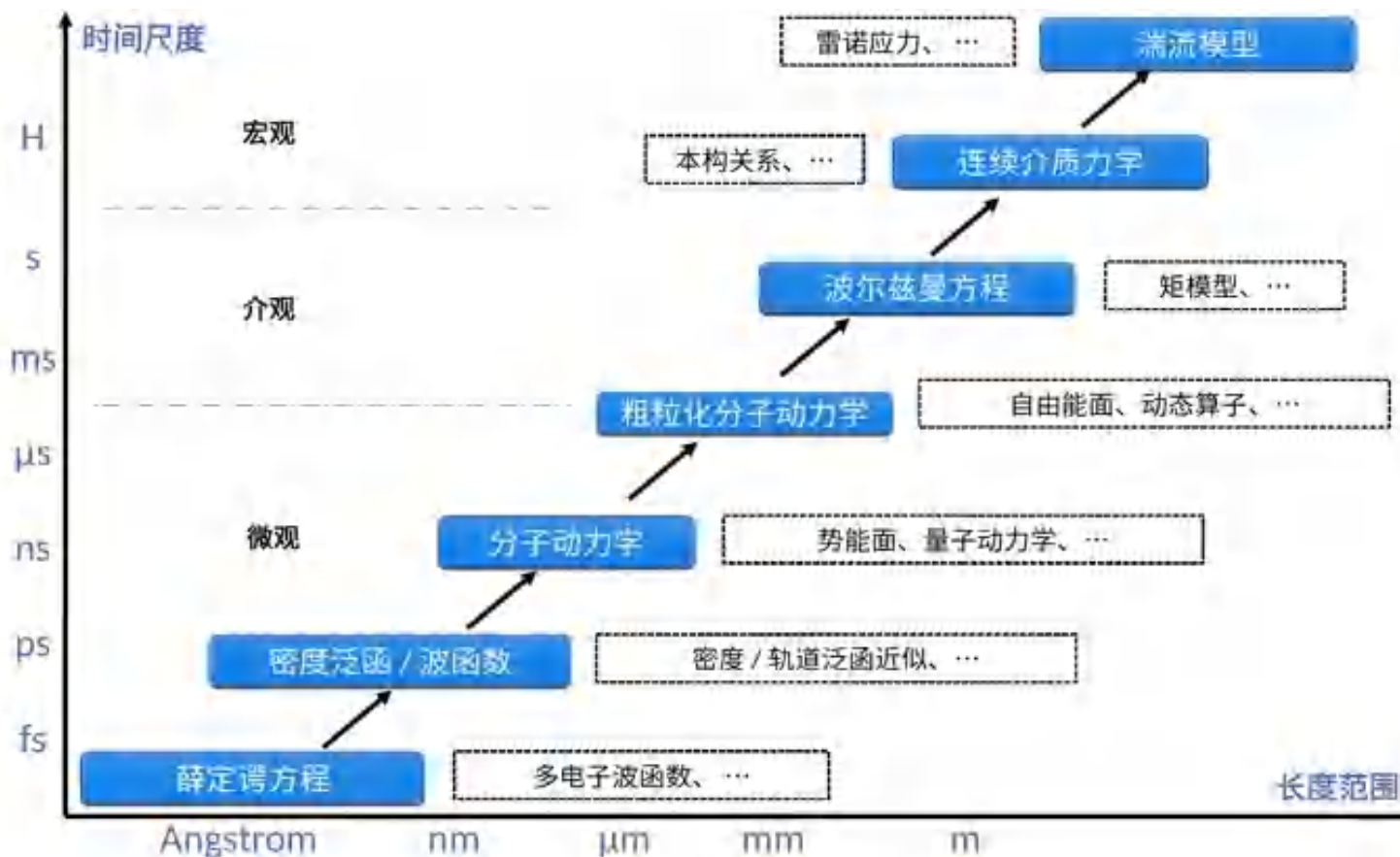
——1687年，艾萨克·牛顿，《自然哲学的数学原理》



Philosophiæ Naturalis
Principia Mathematica
ISAAC NEWTON



多尺度物理模型已经基本完备

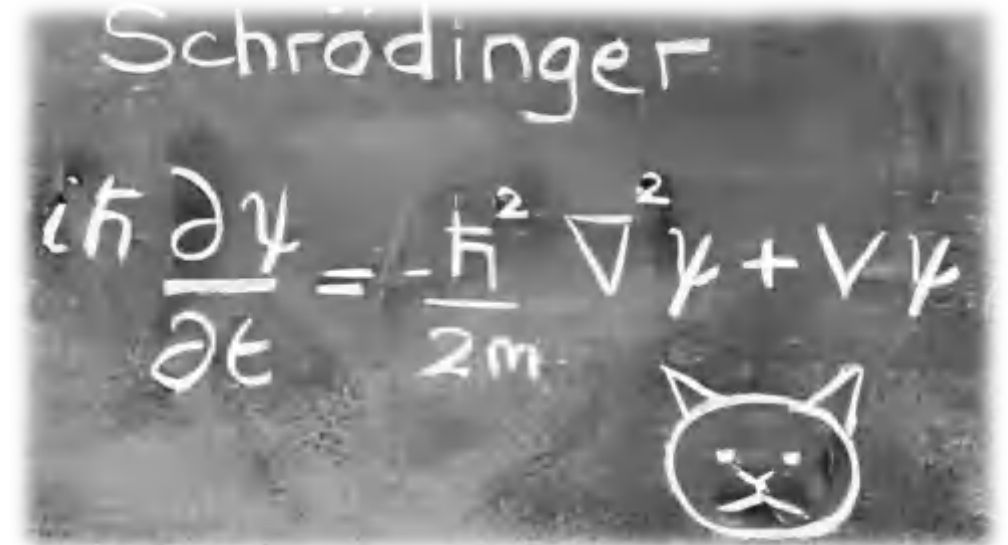
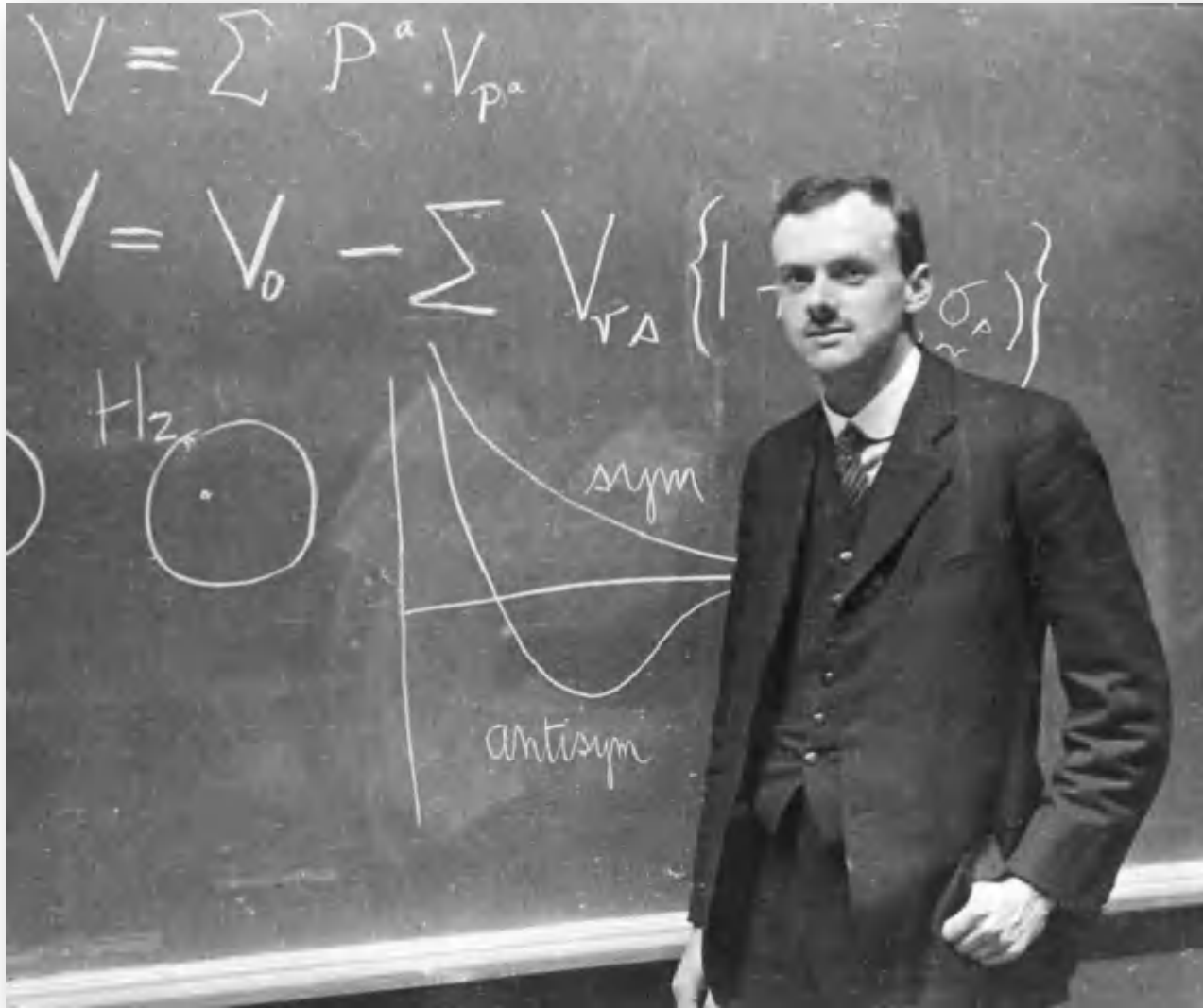


典型应用场景



马尔可夫决策过程

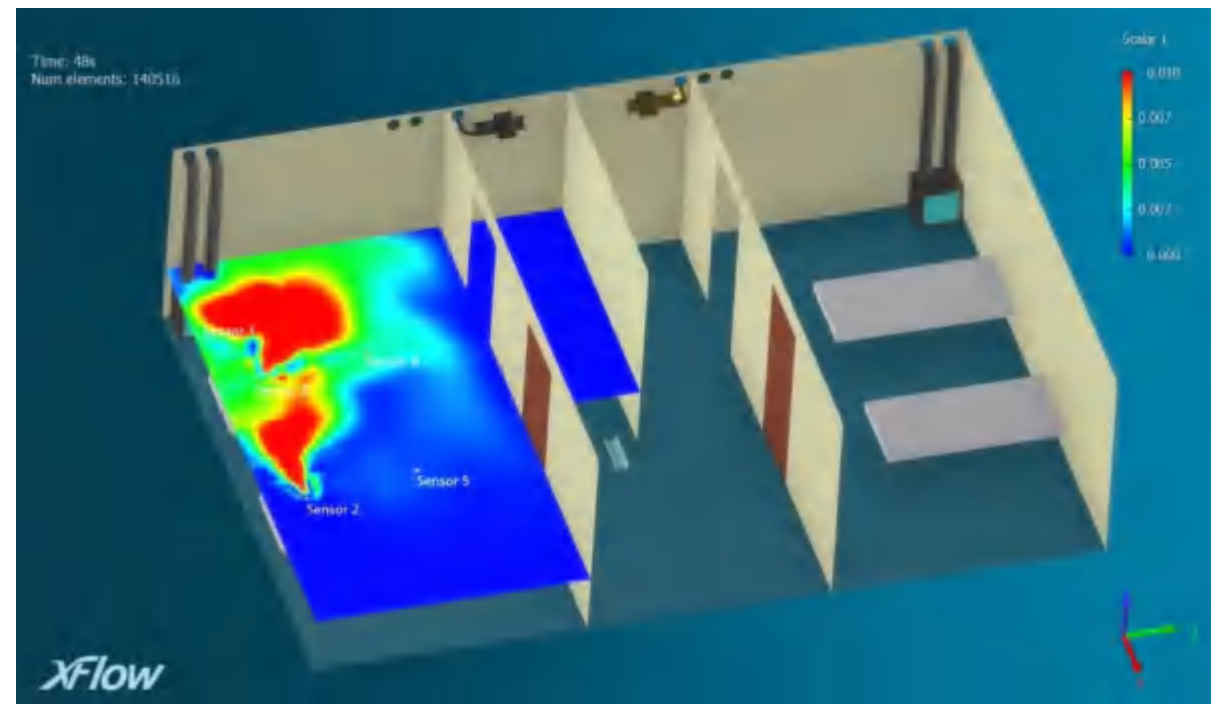
贝尔曼方程



大部分的**物理问题**和**所有的化学问题**在原理上已经**解决**，剩下的问题就是**求解薛定谔方程**。

困难只在于运用这些定律的方程**太复杂**，**无法求解**。

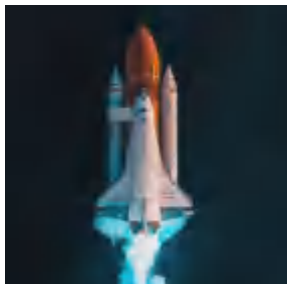
——保罗·狄拉克 《电子的量子理论》



C919研发使用三维数字样机技术

雷神山医院污染扩散仿真

典型场景



火箭模拟



飞机模拟



发动机模拟



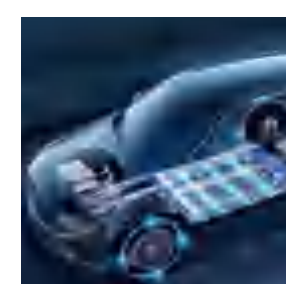
地质模拟



化学反应模拟



药物分子模拟



动力电池模拟



半导体模拟

典型场景

“简单”的问题可以得到解决

结构力学，机械工程，航空航天，电子工程等
这些问题的解决构成了现代工业的基础

复杂的问题

材料性质和材料设计，药物，催化剂等
只能通过经验和试错的办法解决
理论和应用的脱离

简单和复杂问题的分界线：

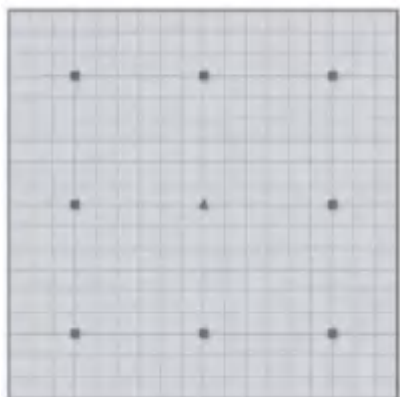
自由度的个数 = 维数

传统方法面临的困难：**维数灾难**

(随着维数的增加，复杂度指数增加)

30秒围棋入门教学

围棋是很简单的一门游戏。孔子曾说，吃饱了没事干就去下围棋吧。以下教程将在半分钟以内教您学会围棋。

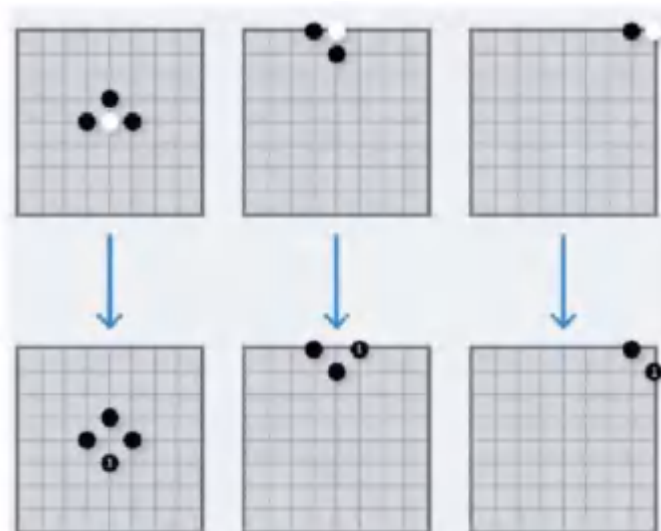


1. 这是围棋的棋盘。上面标记了九个特殊的点，但这只是装饰，并没有什么用。

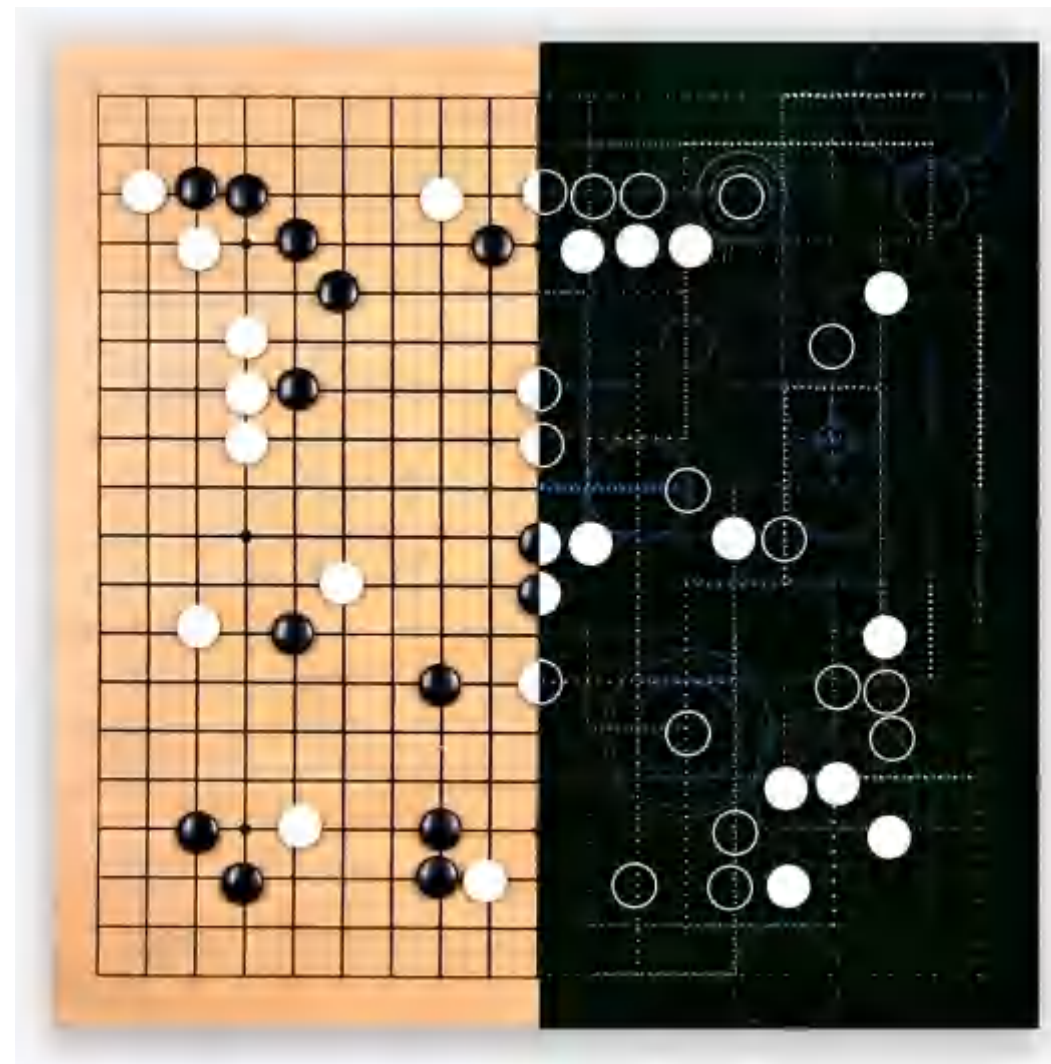
3. 趁对手不注意，吃掉对方的棋子！就像这样，很简单 不是吗？



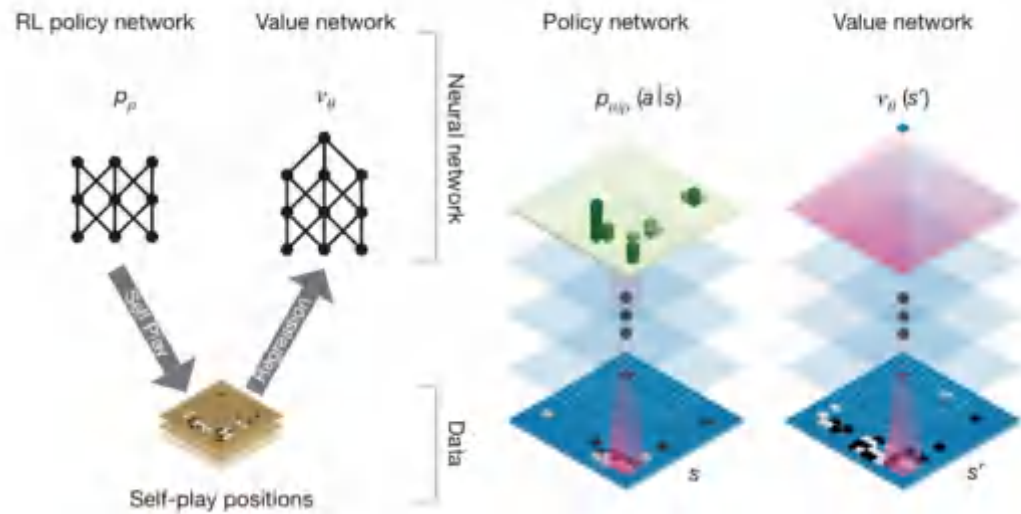
2. 画×的点就是棋子的气。占据所有气就能闷死对面。



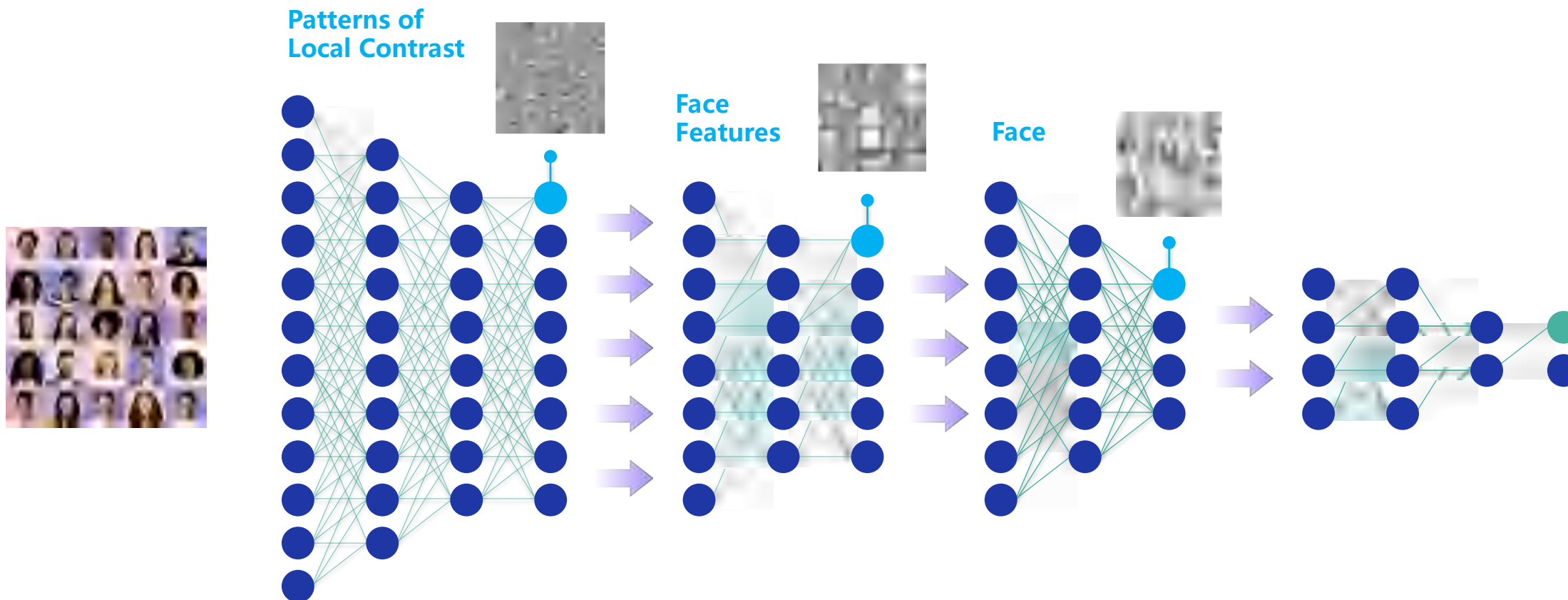
接下来就用你刚学到的知识去和 **AlphaGo** 对线吧



AI解决围棋中的“维数灾难”问题

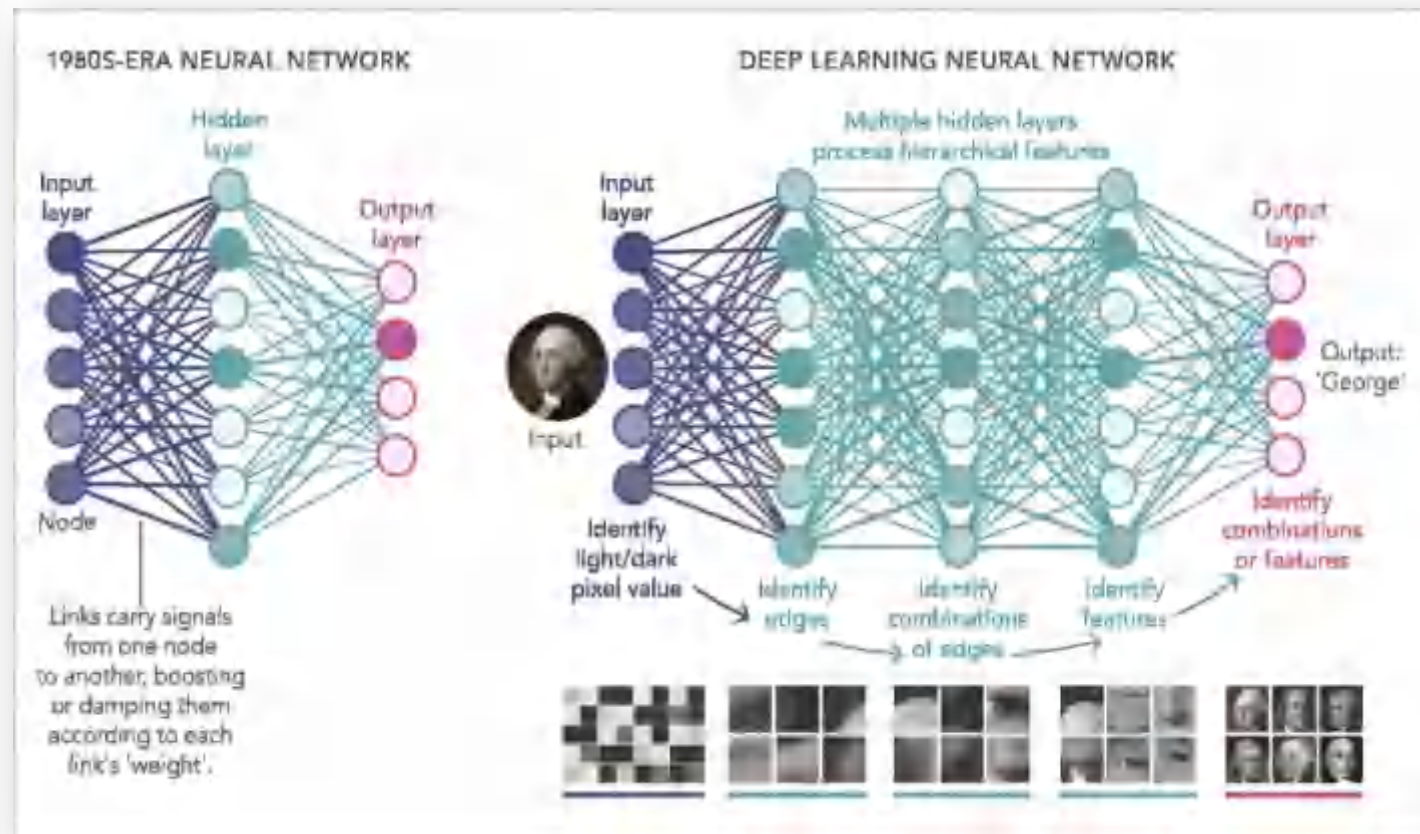
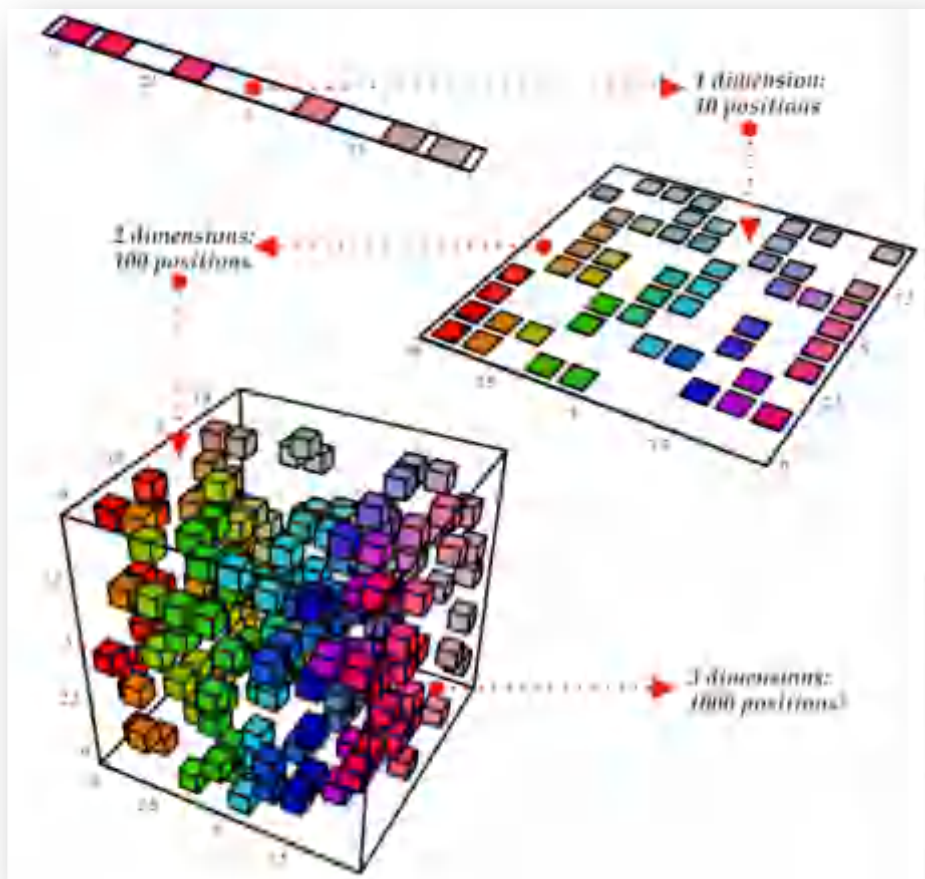


AI解决人脸识别中的“维数灾难”问题



AI带给Science的新机会

AI方法为解决“维数灾难”问题提供了有效手段



维数灾难：从低维问题到**高维**问题

从线性、单体的简化模型到**非线性、多体的复杂模型**

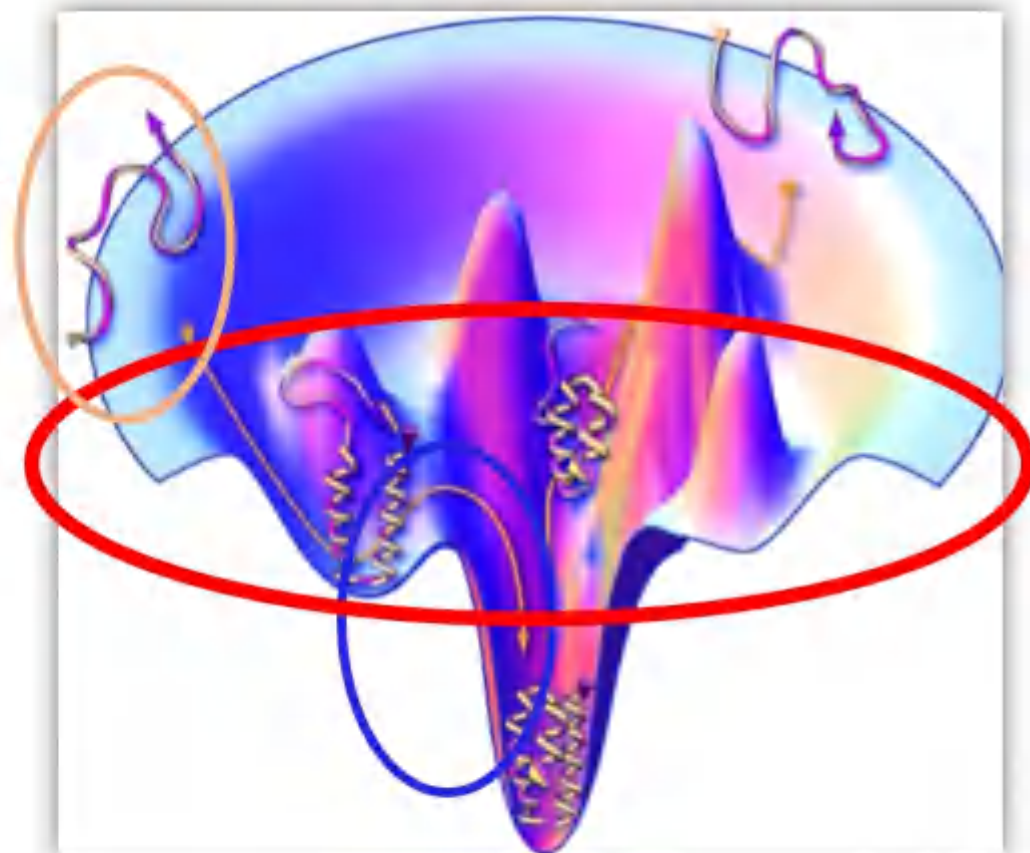
结构、力场与采样是分子模拟的三个关键

$$m_i \ddot{q}_i = F_i \quad (\text{在不和动量同时出现时, 我们也常用 } r \text{ 表示坐标})$$

$$m_i \frac{d^2 r_i}{dt^2} = F_i = -\nabla_{r_i} E, \quad E = E(r_1, \dots, r_i, \dots, r_N).$$

We model and simulate via Newton's Law.

- Structure. $R = (r_1, \dots, r_i, \dots, r_N)$; where to start.
- Forcefield (Potential). $E(R)$; how to represent PES .
- Sampling. $m_i \frac{d^2 r_i}{dt^2} = F_i = -\nabla_{r_i} E$; how to evolve.

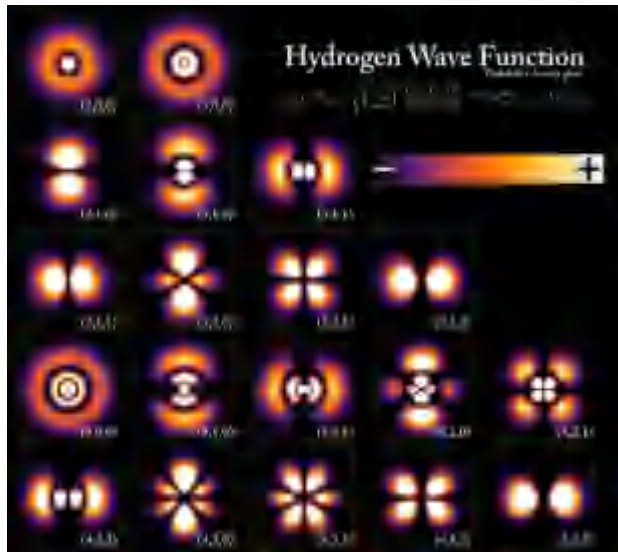


力场：描述原子间相互作用的数学模型

$$E = E(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$$

方法1：第一性原理计算
密度泛函理论 (DFT)

$$i\hbar \frac{\partial}{\partial t} |\Psi\rangle = H |\Psi\rangle$$



准确但计算开销昂贵
计算开销 $O(N^3)$



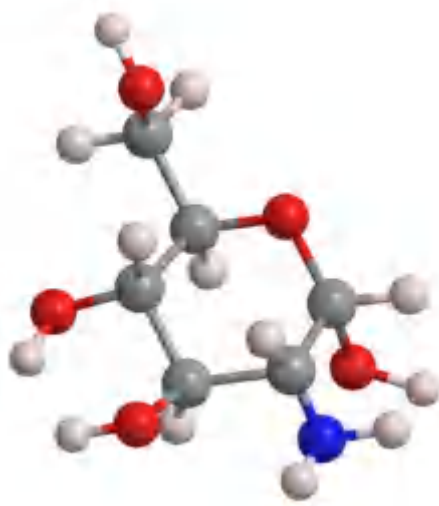
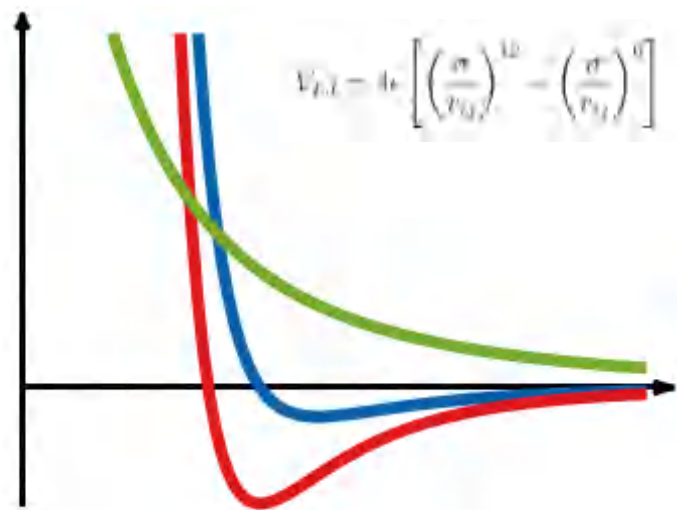
- VASP是目前最流行的电子结构计算和量子力学-分子动力学模拟软件包之一。可以使用赝势和平面波基组，进行从头电子结构和量子力学分子动力学计算。
- CP2K是一个从头算分子动力学软件，可以对固态、液体、分子、周期、材料、晶体和生物系统进行原子模拟。

力场：描述原子间相互作用的数学模型

$$E = E(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$$

方法2：经验力场

例如：LJ, EAM, MEAM,
AMBER, CHARMM



$$\bar{E} = \bar{E}_{\text{bond}} + \bar{E}_{\text{angle}} + \bar{E}_{\text{dihedral}} + \bar{E}_{\text{vdw}} + \bar{E}_{\text{coulomb}}$$

Tunable
parameters

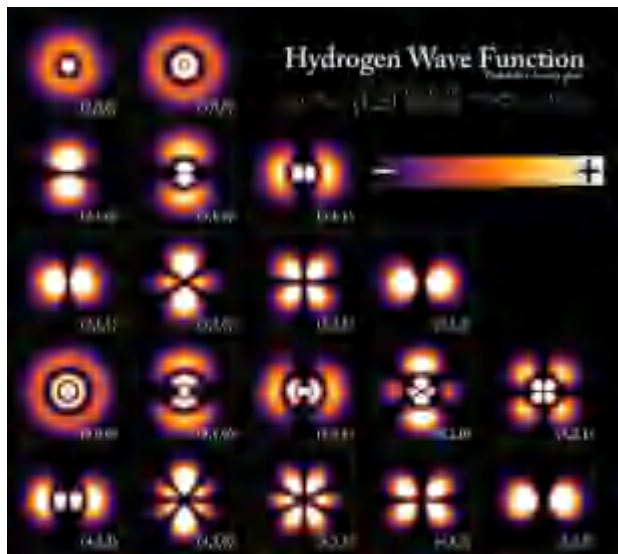
$$k_b, b_0, k_a, \theta_0, k_\xi, \xi_0, C_{ij}, \epsilon, \sigma, q_i$$

计算效率高但**结果不可信**
计算开销 $O(N)$

力场：描述原子间相互作用的数学模型

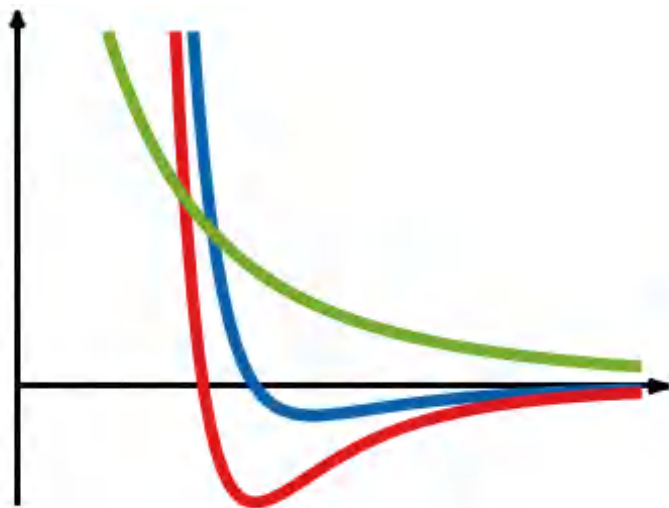
$$E = E(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$$

方法1：第一性原理计算
密度泛函理论 (DFT)



准确但计算开销昂贵
计算开销 $O(N^3)$

方法2：经验力场
例如：LJ, EAM, MEAM,
AMBER, CHARMM



计算效率高但结果不可信
计算开销 $O(N)$

方法3：深度势能
Deep Potential



接近第一性原理计算精度
接近经验力场计算效率
计算开销 $O(N)$

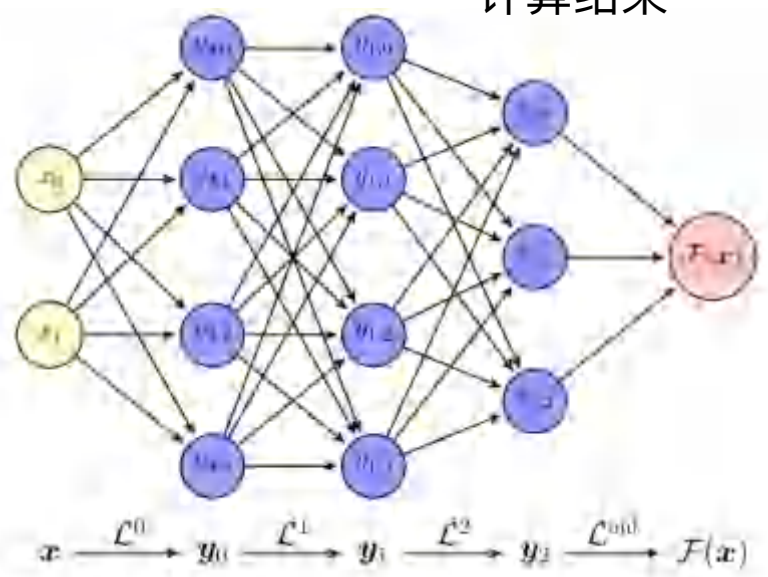
使用机器学习拟合高维势能面

$$E = E(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$$

$$i\hbar \frac{\partial}{\partial t} |\Psi\rangle = H |\Psi\rangle$$

学习第一性原理

计算结果



目标: 数值近似第一性原理定义的能量函数

$$E(\mathbf{R}), \mathbf{R} = \{\mathbf{r}_i\} \quad 3N\text{维}$$

模型:



数据:

$$\{(\mathbf{R}_0, E_0), (\mathbf{R}_1, E_1), (\mathbf{R}_2, E_2), \dots\}$$

$$\mathbf{R}_i, E_i = E(\mathbf{R}_i) \quad \text{由第一性原理计算给出}$$

训练:

$$\min_{\omega} \| E(\mathbf{R}_i, \omega) - E_i \|^2$$

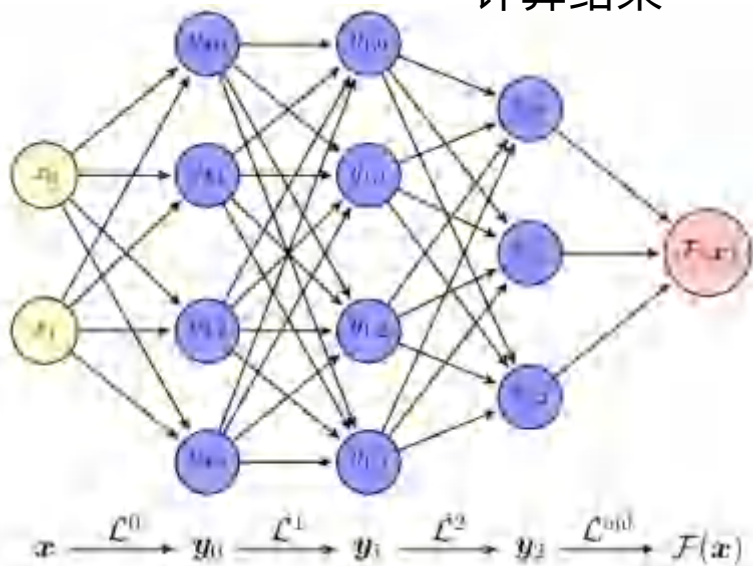
使用机器学习拟合高维势能面

$$E = E(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$$

$$i\hbar \frac{\partial}{\partial t} |\Psi\rangle = H |\Psi\rangle$$

学习第一性原理

计算结果



目标: 数值近似第一性原理定义的能量函数

$$E(\mathbf{R}), \mathbf{R} = \{\mathbf{r}_i\} \quad 3N\text{维}$$

模型:



数据:

$$\{(\mathbf{R}_0, E_0), (\mathbf{R}_1, E_1), (\mathbf{R}_2, E_2), \dots\}$$

$$\mathbf{R}_i, E_i = E(\mathbf{R}_i) \quad \text{由第一性原理计算给出}$$

训练:

$$\min_{\omega} \| E(\mathbf{R}_i, \omega) - E_i \|^2$$

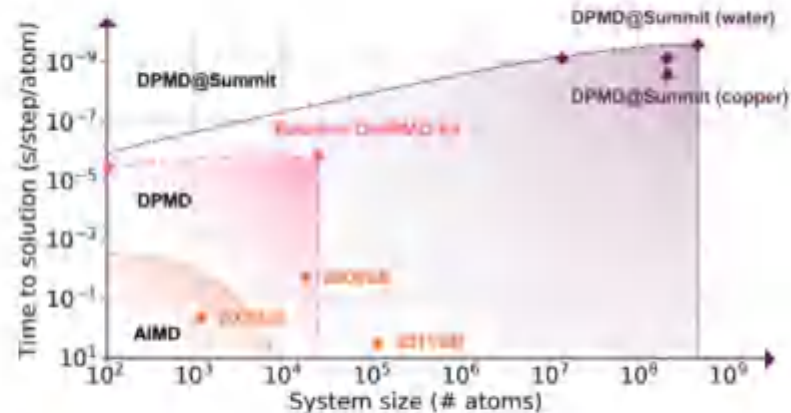
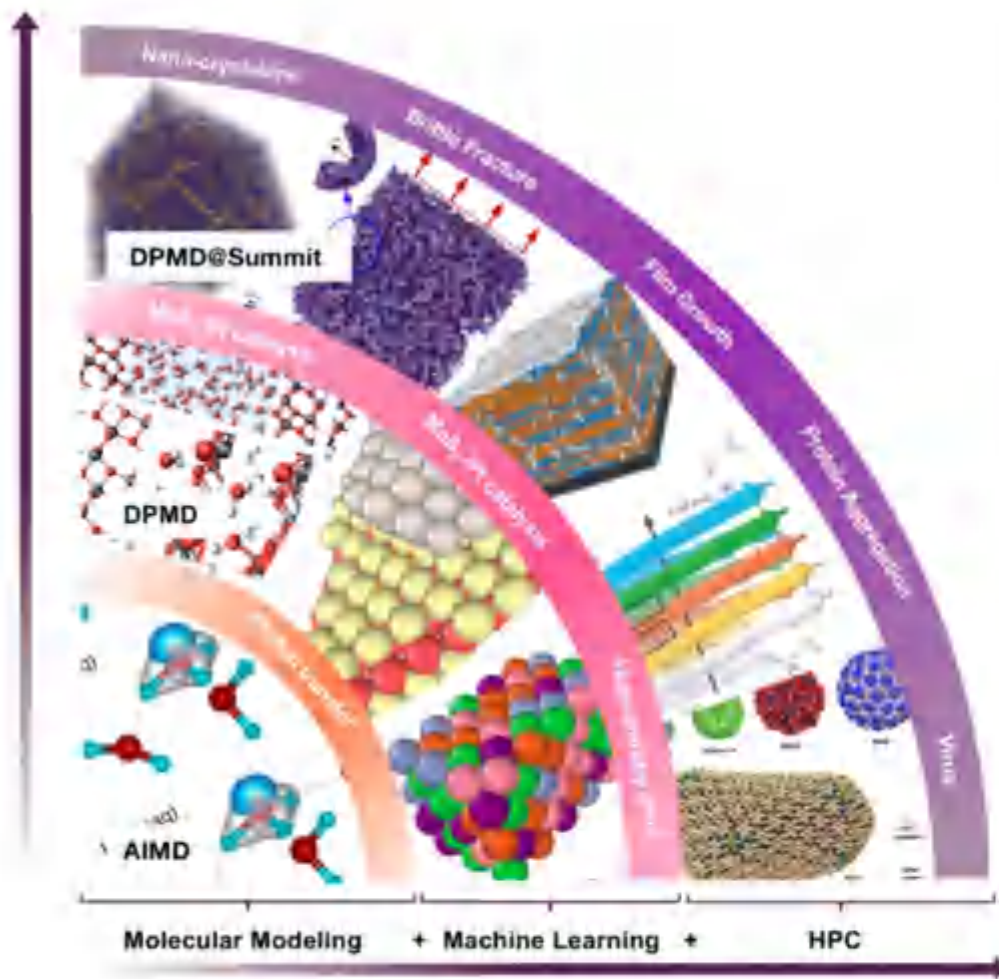
要求:

1. 保证模型可扩展性
2. 保持物理上对称性

最终模型:

$$E = \sum_i N_{\alpha_i} \left(D_{\alpha_i}(\mathbf{r}_i, \{\mathbf{r}_j\}_{j \in n(i)}) \right)$$

深度势能：物理建模+人工智能+高性能计算



ACM GORDON BELL PRIZE - WINNER

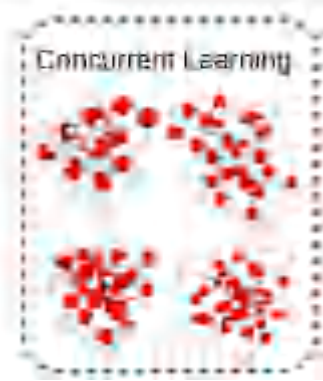
Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning

University of California Berkeley, Institute of Applied Physics and Computational Mathematics, Tsinghua University, Lawrence Berkeley National Laboratory, Princeton University

SC20

Background

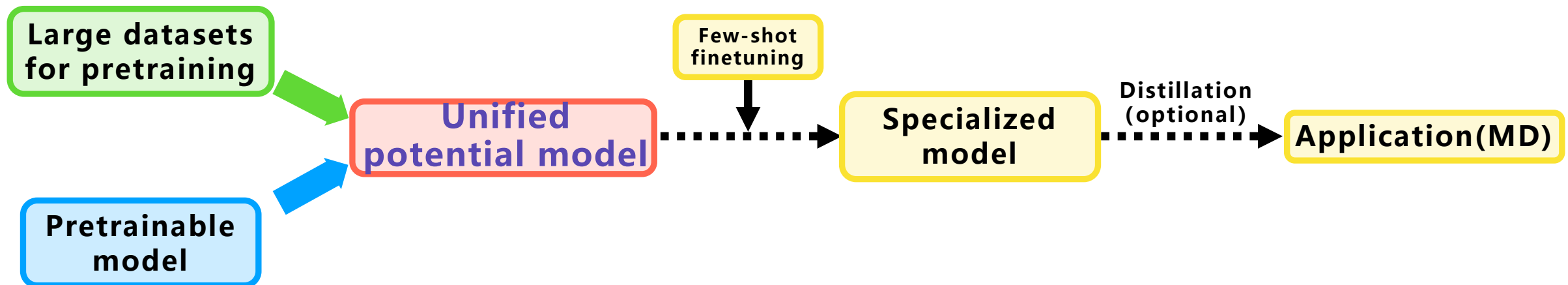
- 训练一个机器学习势函数的开销:



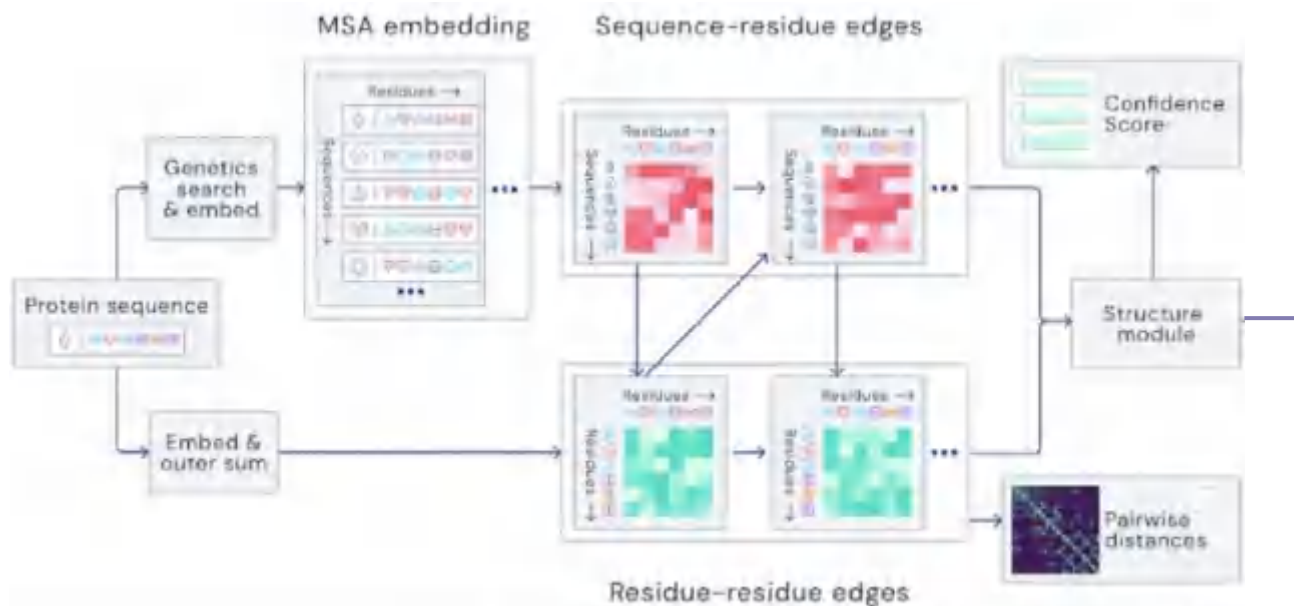
以 $\text{Al}_x\text{Mg}_y\text{Cu}_z$ [*] 势函数为例 ($0 \leq x, y, z \leq 1, x + y + z = 1$):

- **10 million** CPU core-hours 和 **20-30 thousand** GPU card-hours

- New paradigm: **pretraining-then-finetuning**



[*]Jiang, W., Zhang, Y., Zhang, L. and Wang, H., 2021. Accurate Deep Potential model for the Al–Cu–Mg alloy in the full concentration space. Chinese Physics B, 30(5), p.050706.



AlphaFold2 模型训练过程示意图 [Source: DeepMind]



- **AlphaFold2**: 一款数据驱动的解决蛋白质三维结构预测的方法。
- **为什么适合?**
 - 已有数据足够多
 - 建模需求明确
 - 评估标准明确
 - 探索空间大

AlphaFold2的成功给药物设计、结构生物学等许多科学和应用领域带来了根本的改变

1.1 AI4S的第一类寻宝图：AI建模高维复杂函数

1.2 AI4S的第二类寻宝图：AI驱动平台化科研

1.3 AI4S该怎么学



从AlexNet到ChatGPT，AI经历了快速发展和变革的10年 2012 —— 2022

2012年，Ilya Sutskever作为Jeff Hinton的博士生做出**AlexNet**，
2022年，Ilya Sutskever作为OpenAI的首席科学家做出**ChatGPT**；

模型的发展与变革 2012 —— 2022

2012年，**AlexNet**、**ConvNets**；2014年**GAN**；2015年**ResNet**；
2018年，**Transformer**和**注意力机制**；

框架的发展与变革 2012 —— 2022

2015年，Google发布**TensorFlow**，DMLC发布**MXNet**；
2016年，Facebook发布**PyTorch**。

社区与公司的发展与变革 2012 —— 2022

2014年，**Deepmind**被Google收购；2015年，**OpenAI**创立；
2016年，**Hugging Face**社区创立

AI发展迅速的原因：模型、框架、平台完善，快速试错迭代

Science领域的现状：领域知识艰深、软件迭代慢、组织形式落后

2017-2022

2022-2027

2027-未来

以科学家为主导的概念导入期

- 2017年Deep Potential出现，改变了微观模拟领域数十年来的格局
- 2018年鄂维南院士首次提出AI for Science
- 2020年DeepMind 推出的 AlphaFold2 在 CASP14 大赛中轰动世界
- DeepModeling开源社区产生较广泛的影响力，提供了“平台科研”模式的实践经验，成就了学界各方向成规模的研究工作和业界研发

定义出迫切需要也十分适合使用 AI 来求解的关键科学问题，并实现算法领域的0到1突破和应用侧的初步验证

以科学家和工程师协作为标志的大规模基础设施建设期

- 属于AI for Science 领域的chatGPT模型的出现，如 DPA 系列、替代科研文献的AI知识库
- 新一代基础科研软件，如国产密度泛函软件 ABACUS、云原生科学计算 workflow 套件 dflow；
- 新一代工业软件，例如面向药物研发场景的整体方案Hermite；
- 条件成熟的应用场景产生专用硬件和颠覆性进展
- 全面提升创新效能的平台科研体系

数据、模型、算力一体化融合，“科学问题”逐步转变为“计算和工程问题”

全面、系统的应用升级和需求导向的原始创新体系的持续发展迭代

- 新计算工具定义新研发流程
- 软件定义硬件
- 面向细分场景，面向用户开发应用
- 形成新的行业标准和行业生态
- 形成鼓励原始创新的环境

基础设施建设的基本完成后，进入到基础设施“怎么用”和“怎么持续领先”的新阶段

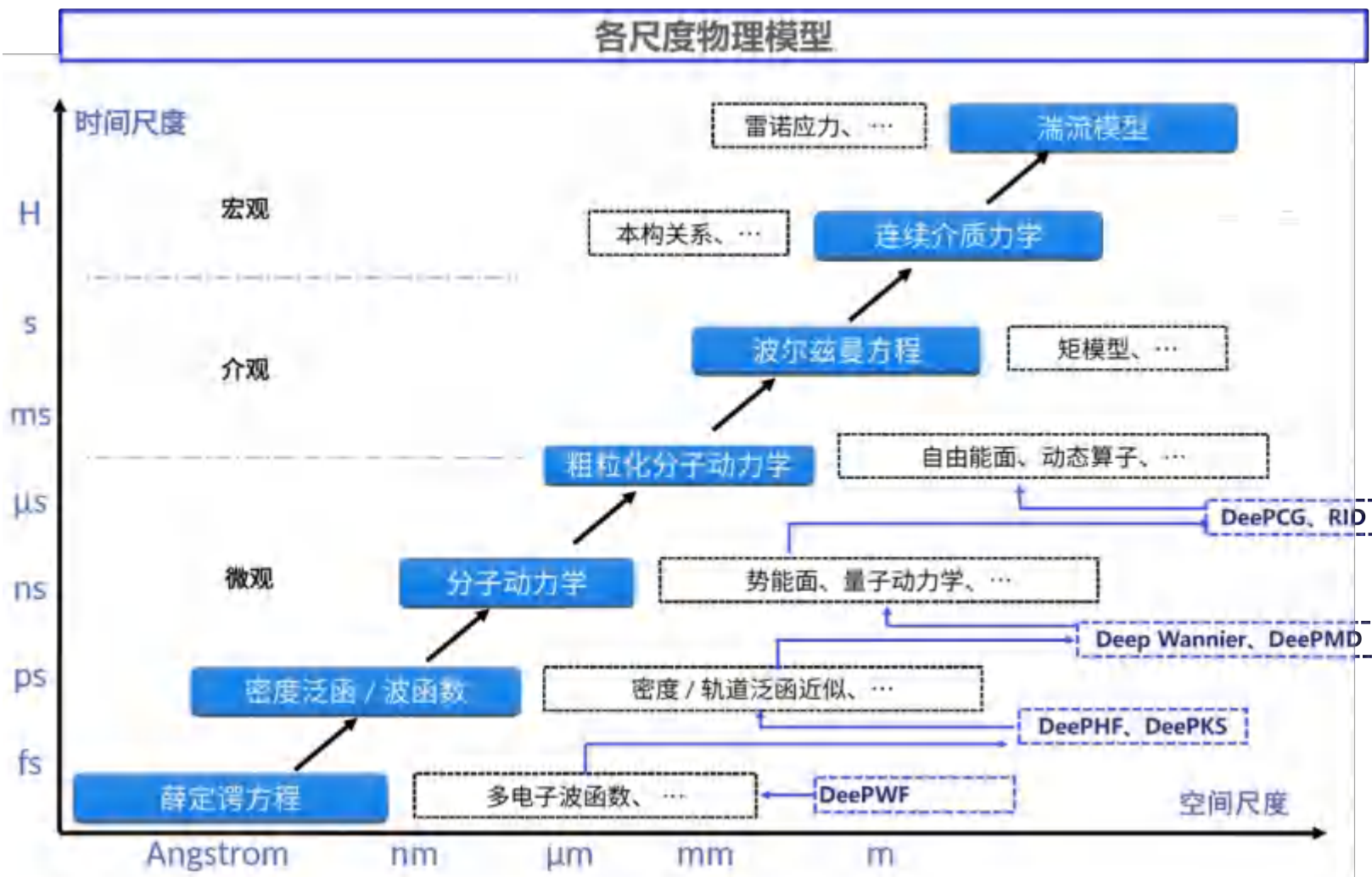
AI4S基础设施：“砖瓦”与“四梁N柱”



基本原理与数据驱动模型算法与软件系统

AI for Science 为从底层构建全新的跨尺度算法与软件带来新机遇

各尺度物理模型



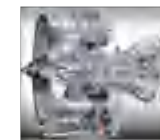
典型应用场景



火箭模拟



飞机模拟



发动机模拟



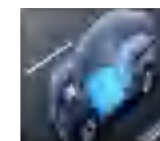
地质模拟



高分子材料模拟



柔性材料模拟



动力电池模拟



半导体模拟



化学反应模拟

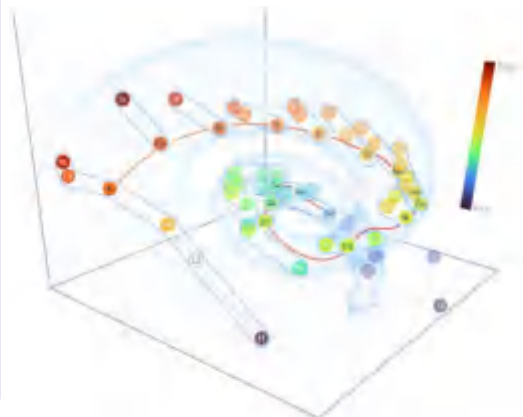
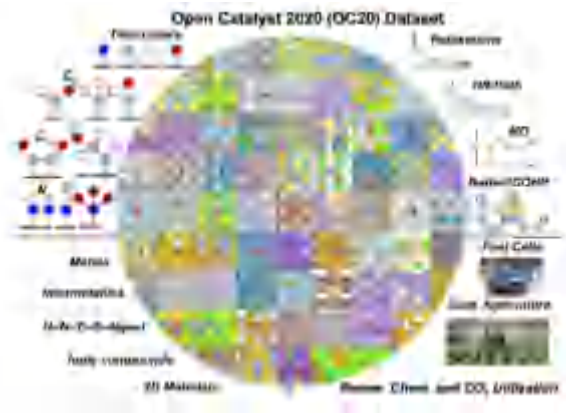


药物分子模拟

构建预训练模型

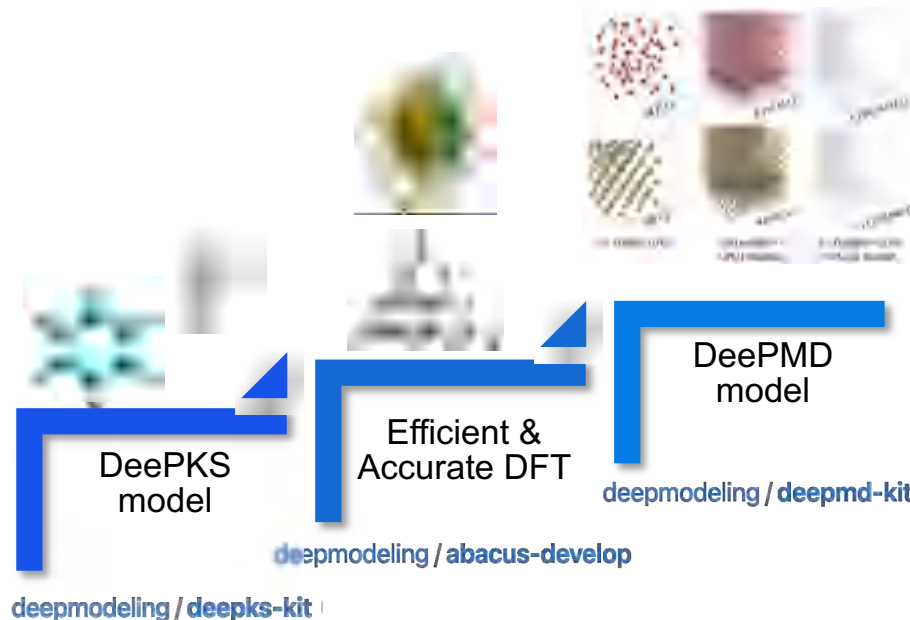
DPA: 打造“自然科学界的ChatGPT”

- 全球首个覆盖元素周期表70种元素的深度势能原子间势函数预训练模型DPA-1
- 模型将成为药物、材料、能源等行业**微观机理探索的重要基础设施**
- 显著降低研究人员使用门槛及研发成本，缩短研发周期**



打造开源平台及软件应用

ABACUS: 国产开源密度泛函软件



作为基础设施

- ABACUS目标: 开源、功能完整、易用易开发
- AI和新硬件: 带来新机遇, 各种密度泛函理论算法仍然持续更新

ABACUS与DeepModeling社区合作

- 从21年初开始, ABACUS在DeepModeling社区下开始践行开源的理念

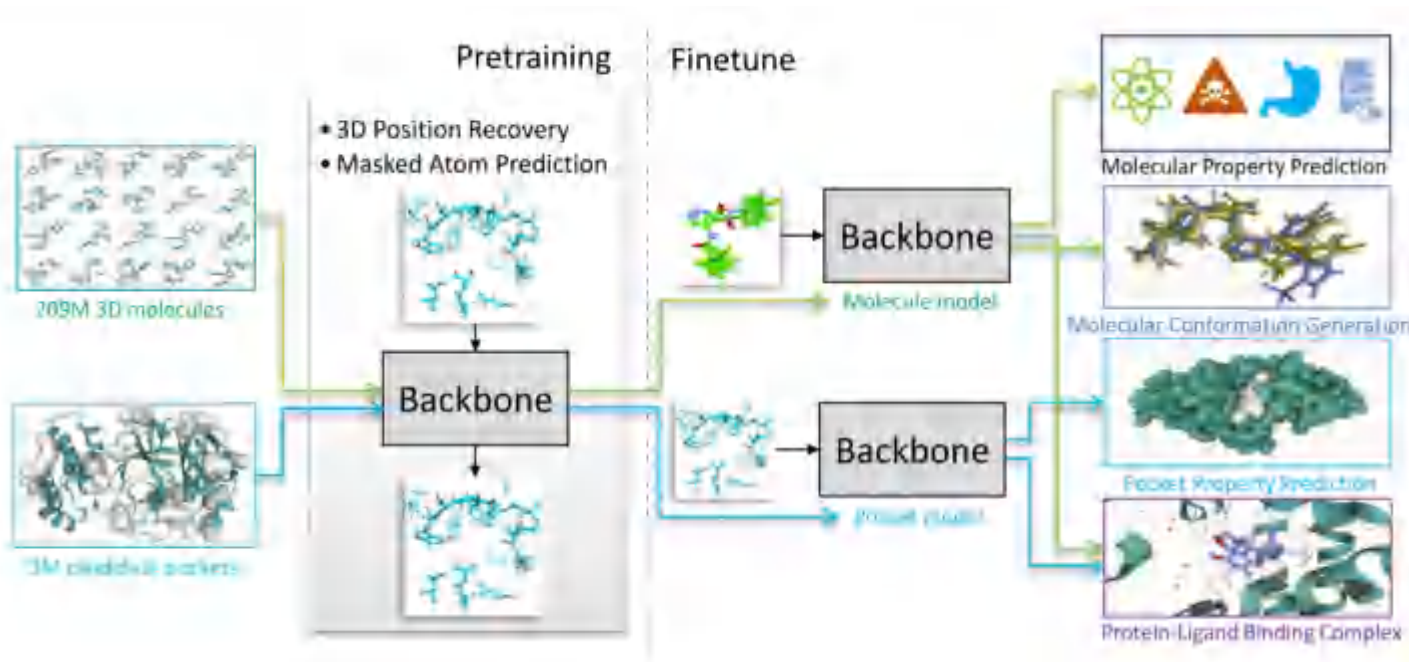
Uni-Mol: 第一个通用分子3D表示学习(MRL)框架和预训练模型

Uni-Mol: A Universal 3D Molecular Representation Learning Framework (ICLR 2023)

<https://openreview.net/forum?id=6K2RM6wVqKu>

通用3D分子表示学习框架:

- **表征:** 原子类型 + 原子坐标
- **模型:** Transformer + 旋转平移不变性 + SE(3)-等变
- **优势:** 同一套表征方法可以同时用以处理小分子、复合物, 可以应用于分子性质预测、3D坐标预测等多种任务。



预训练模型:

- **数据:** 209M小分子构象
- **策略:** 原子类型还原 + 原子坐标还原
- **优势:** 使用海量无标记数据让模型先学习分子结构的特点来获取通用能力

<https://github.com/dptech-corp/Uni-Mol>

Source codes, pretraining data, pretrained models, finetuned model weights, and demos for downstream tasks are all accessible.

提供高通量的正向数据

提供新的反演算法

基于人工智能的实验表征反演算法

自动化实验表征方法与控制软件

实验表征与数据自动化采集与智能化管理

面向实验表征与计算模拟联动的解决方案

电池化学组成表征技术

- (1) X 射线光电子能谱 (XPS)
- (2) 电子能量损失谱 (EELS)
- (3) X 射线吸收近边谱 (XANES, XAS)

电池材料形貌表征技术

- (1) 扫描电镜 (SEM)
- (2) 透射电镜 (TEM)
- (3) 原子力显微镜 (AFM)

电池界面敏感表征技术

- (1) 和频振动光谱 (SFG)
- (2) 表面增强拉曼光谱 (SERS)

电池晶体结构表征

- (1) X射线衍射 (XRD)
- (2) 核磁共振 (NMR)
- (3) X射线吸收谱 (XAS, EXAFS)

电池官能团表征

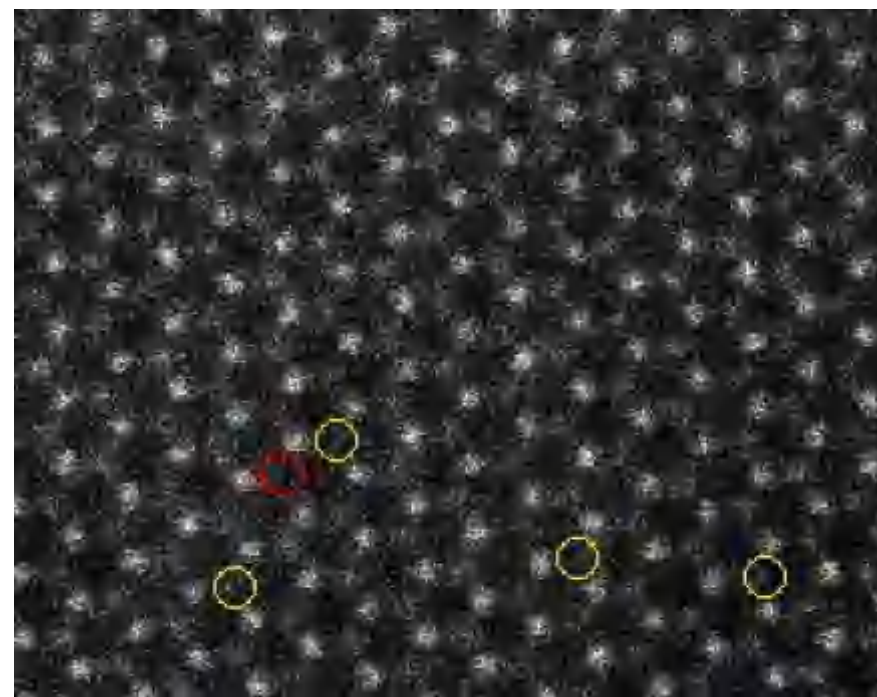
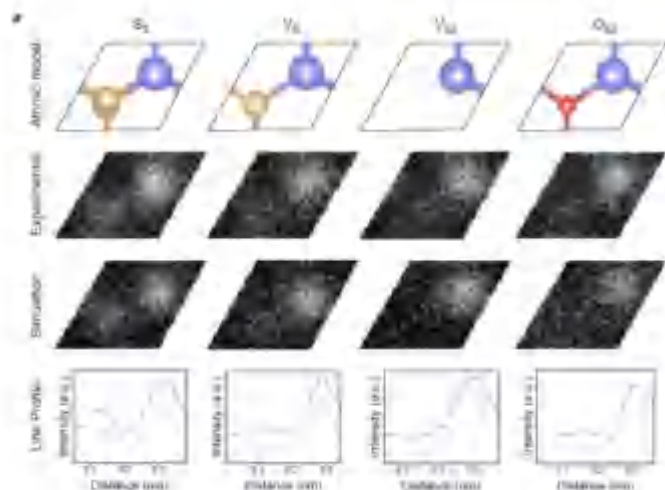
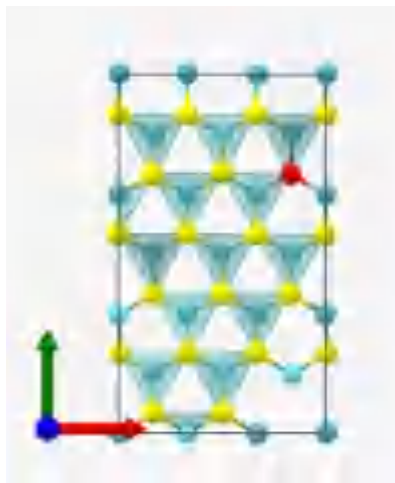
- (1) 红外谱 (IR)
- (2) 拉曼谱 (Raman)

二维材料缺陷检测

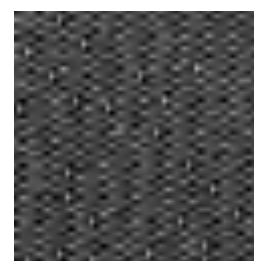
MoS₂中的单硫(S-vacancy), 双硫(2S-vacancy), 氧替代(oxygen substitution)缺陷是指材料晶格中缺少一个或两个硫原子的位置, 或者在晶格中氧原子取代硫原子。这种缺陷的存在对材料的性质有重要的影响。

需求:

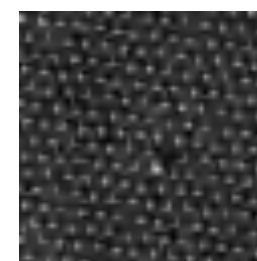
找到STEM图像中的缺失的单/双原子, 氧替代及其位置



MoS₂



WSe₂

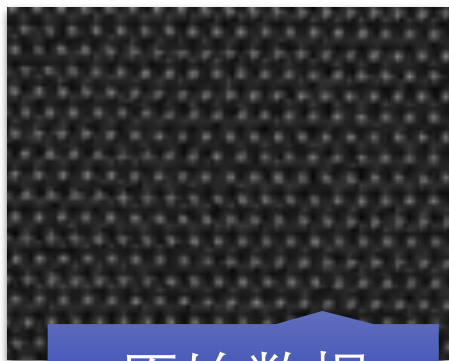
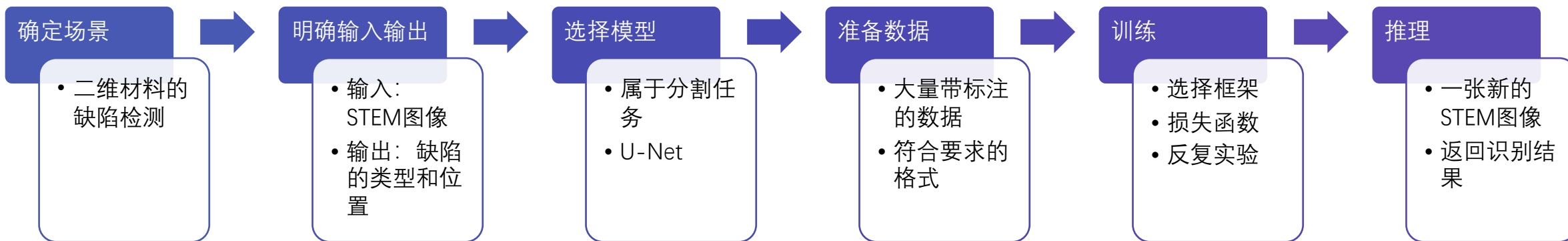


NbS₂

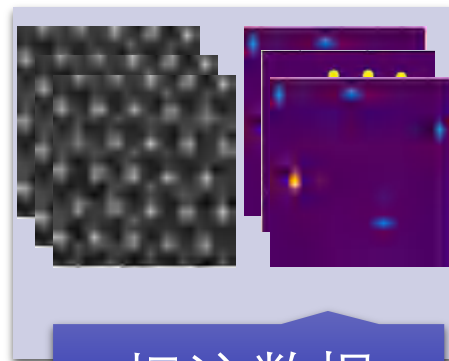


NbSe₂

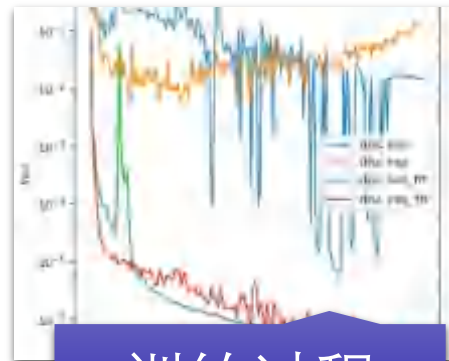
二维材料缺陷检测



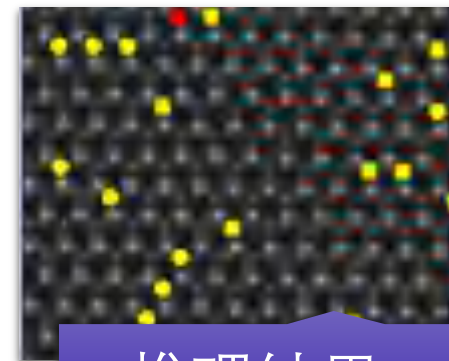
原始数据



标注数据



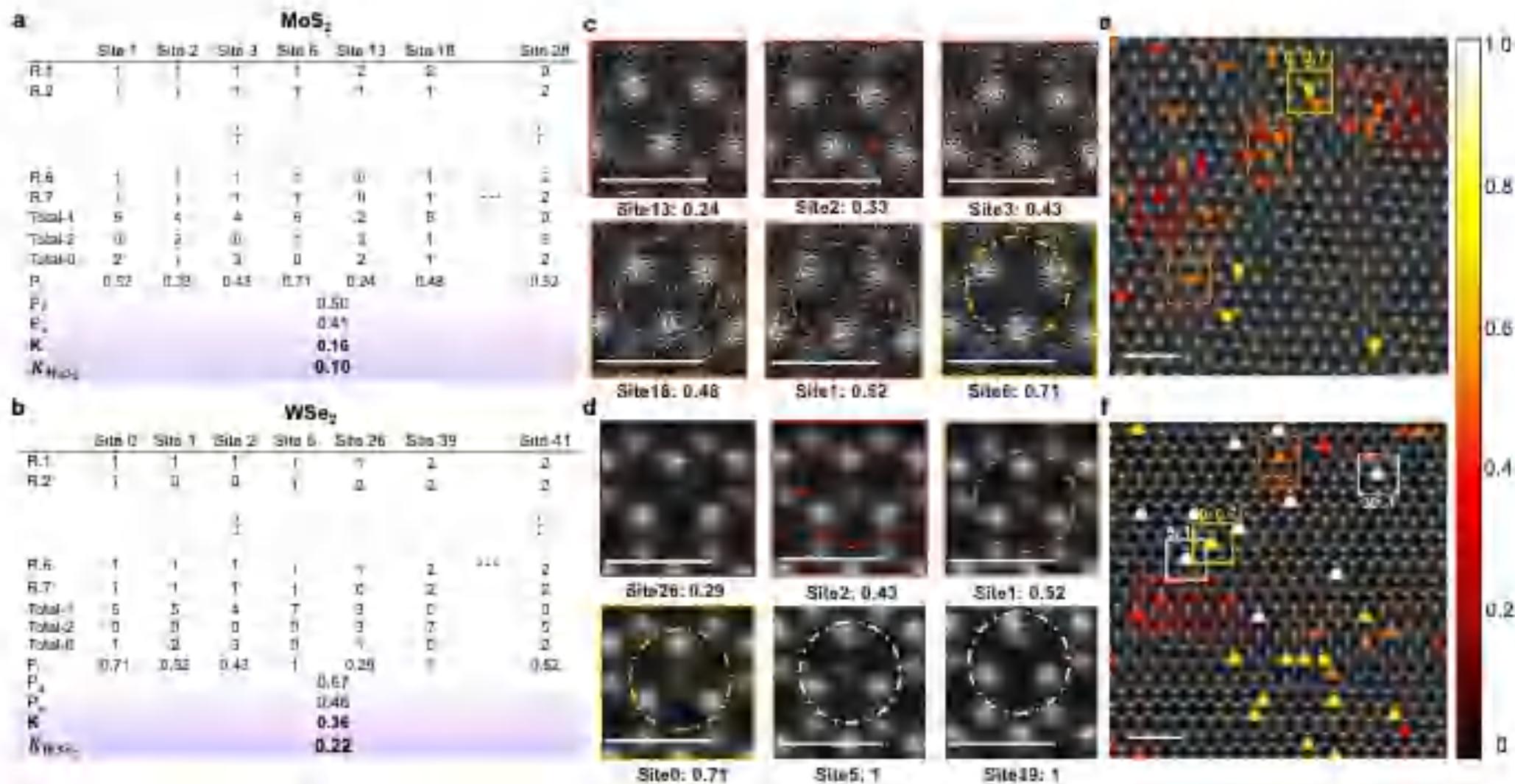
训练过程

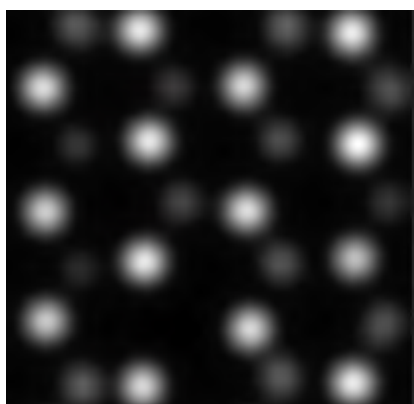
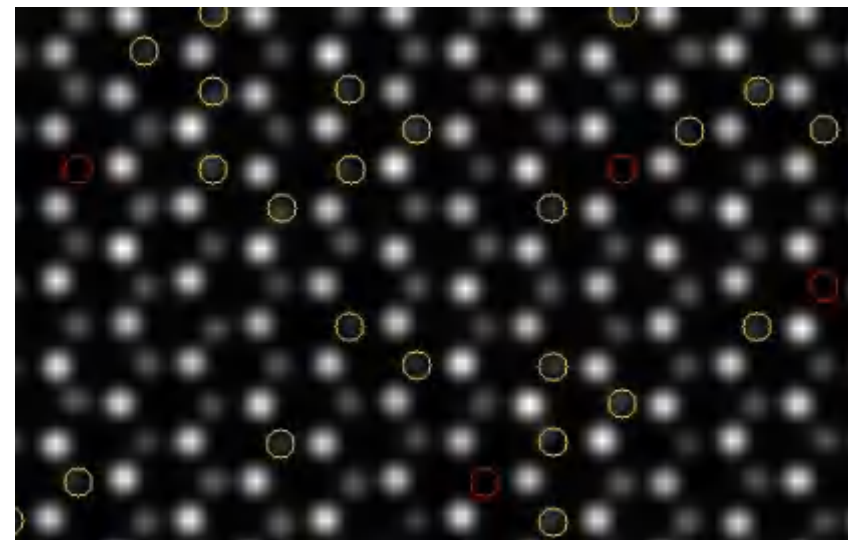
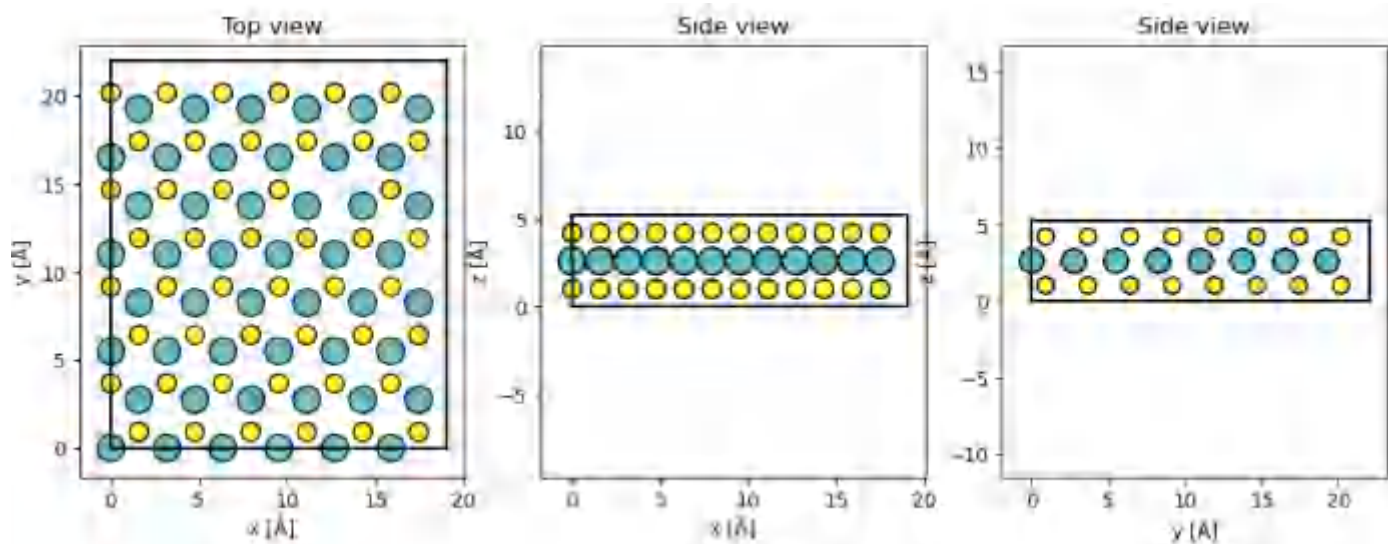


推理结果

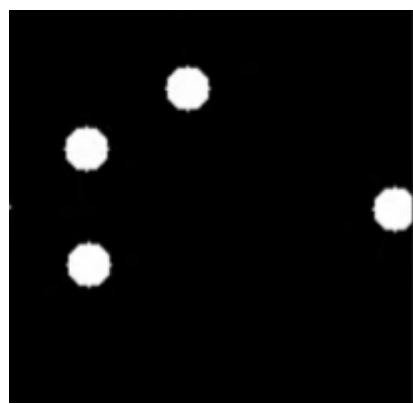
二维材料缺陷检测

MoS₂ 缺陷标注困难

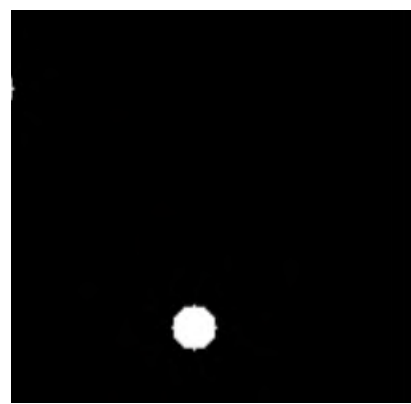


MoS₂-仿真

Patch (256*256)



Single vacancy



Double vacancy

Abtem:

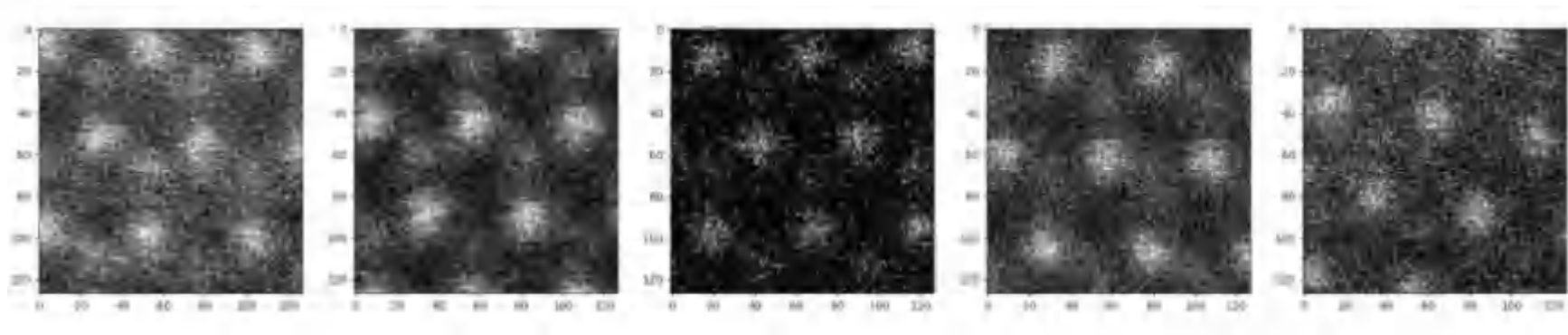
<https://abtem.readthedocs.io/en/latest/intro.html>

Bohrium:

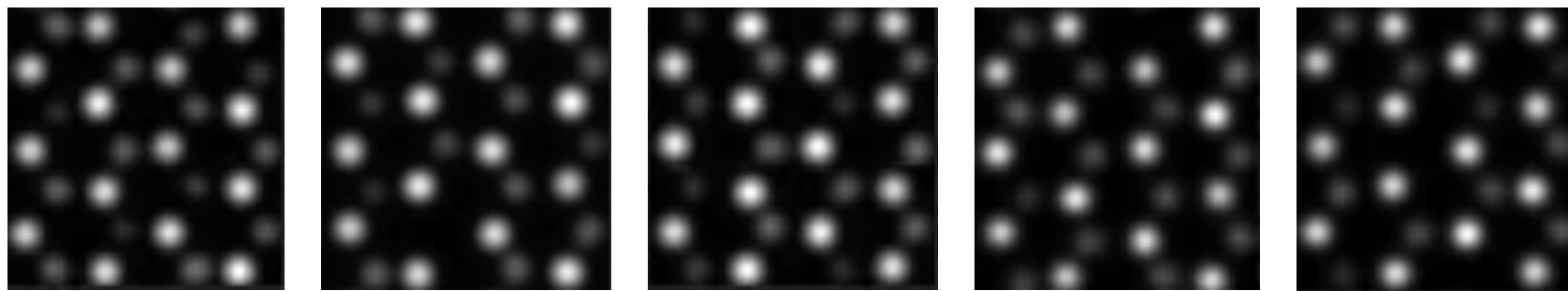
<https://nb.bohrium.dp.tech/detail/6241642088>

二维材料缺陷检测

真实图像



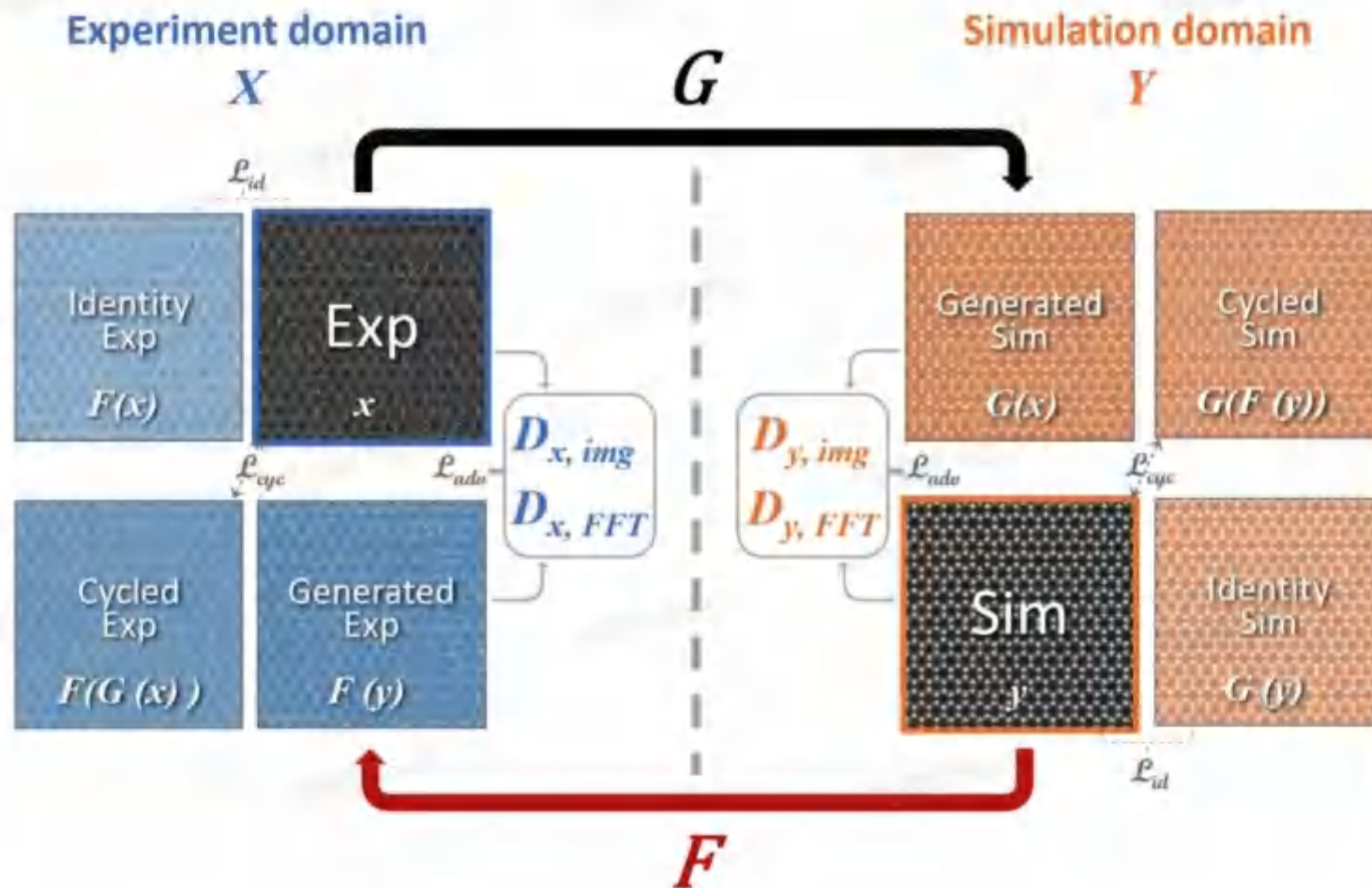
仿真图像



原因：晶格畸变、探测器噪声、样品漂移和扫描畸变、时间依赖的对齐误差、辐射损伤以及表面污染等

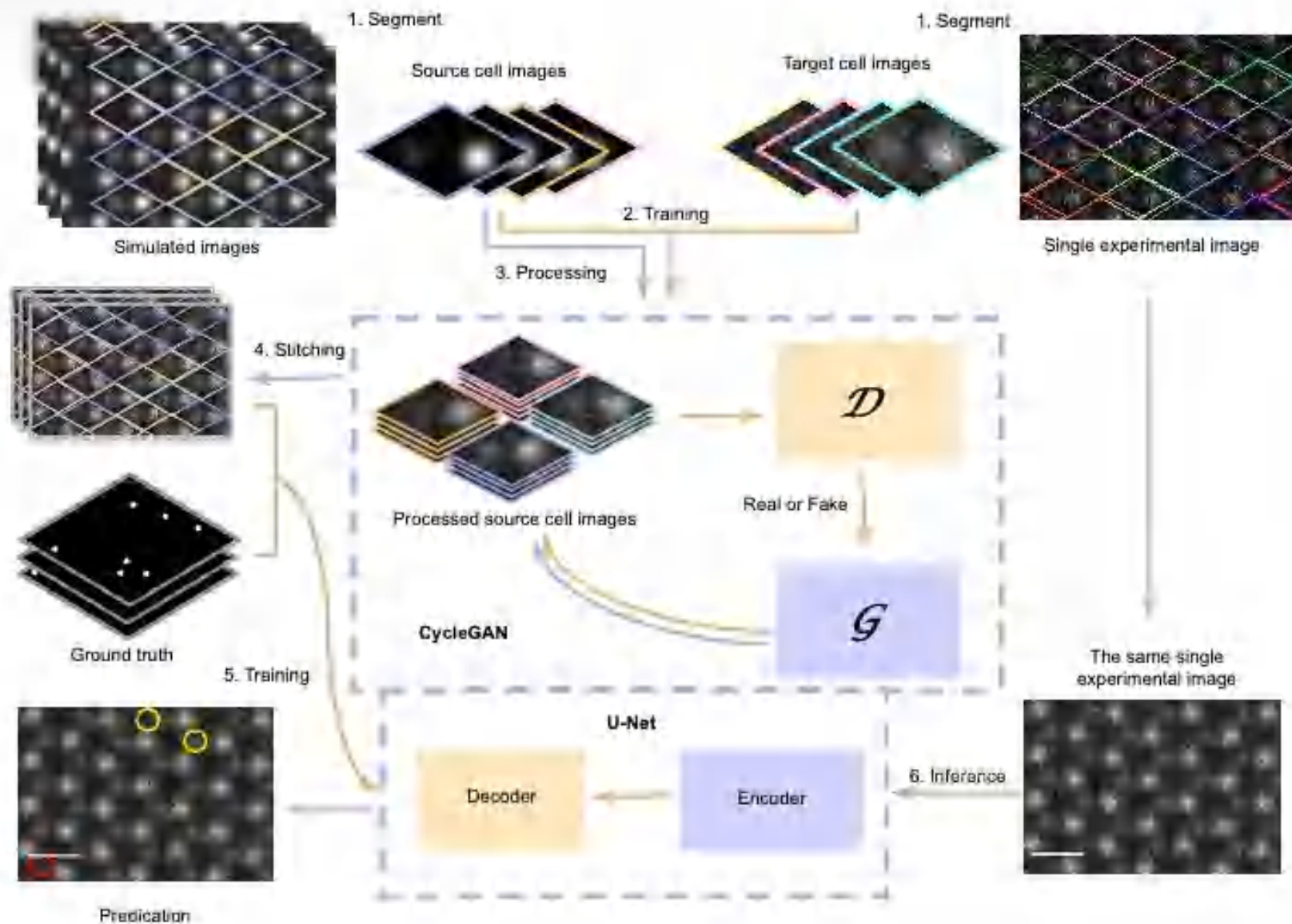
二维材料缺陷检测

MoS2-仿真 + 噪音：生成对抗网络cycle GAN



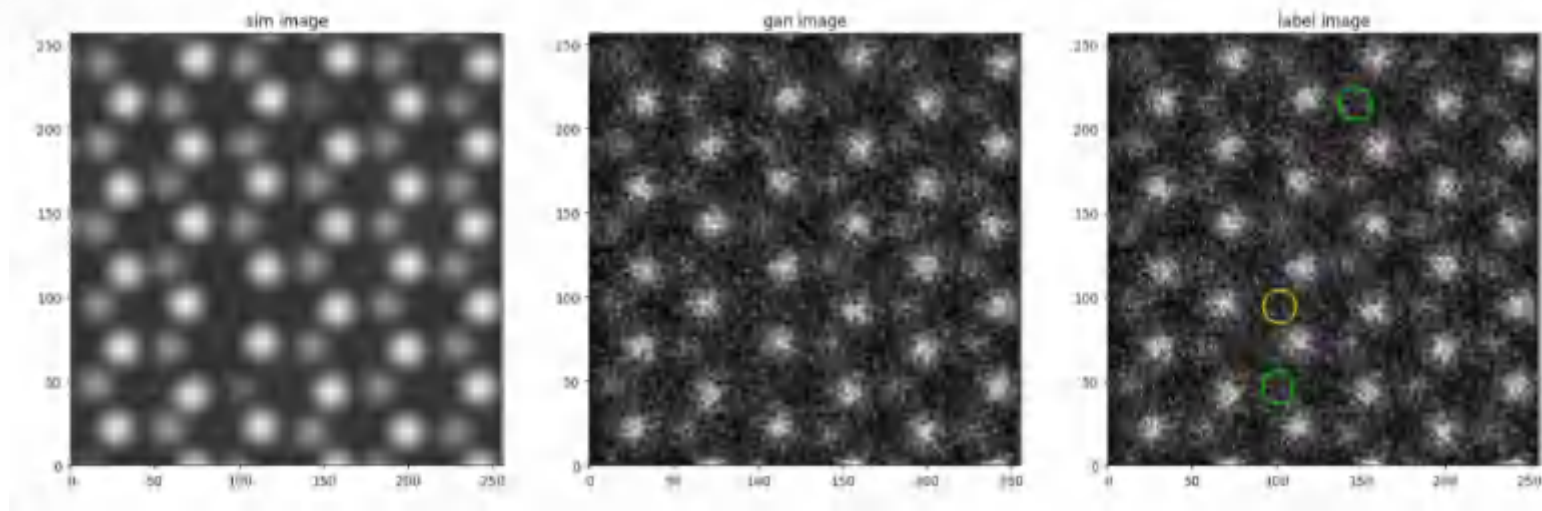
二维材料缺陷检测

整体框架

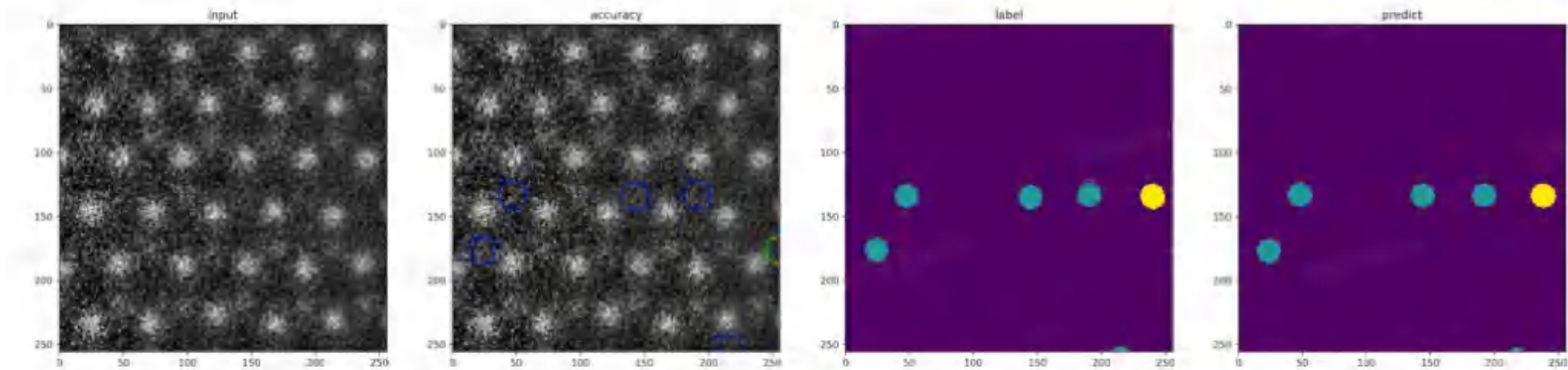


二维材料缺陷检测

生成对抗网络的结果



实验图的缺陷检测

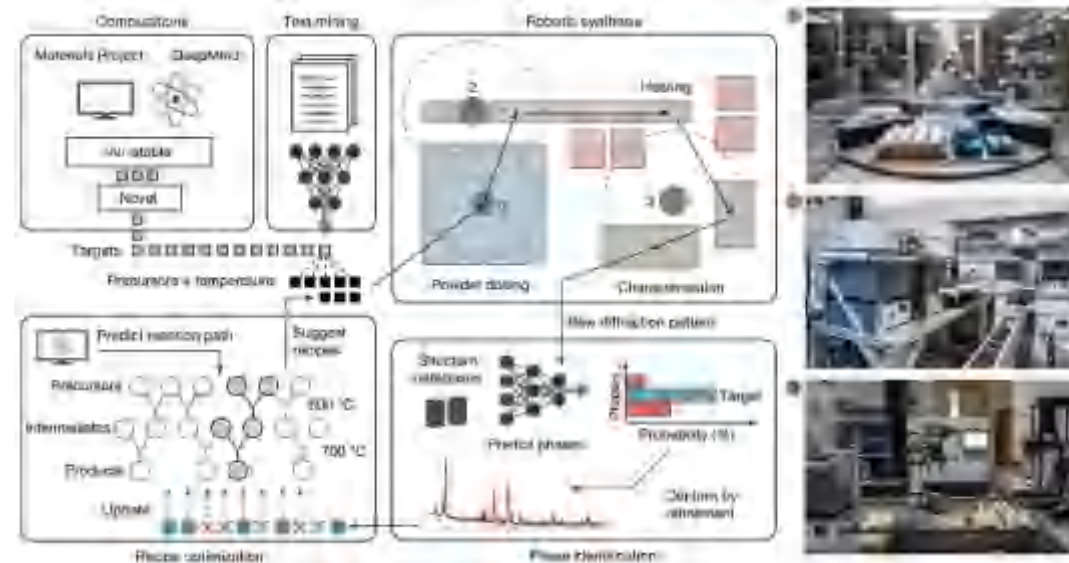




辉瑞的自动化高通量化学反应筛选平台, *Science*, 2018, 359, 429



KUKA机器人发现光解水产氢的光催化剂。 *Nature*, 2020, 583, 237-241



用于无机粉末材料自主固态合成的自动化实验室A-Lab
Nature, 2023, 624, 86-91

“四梁”：高度整合的算力平台

- 通用芯片门槛高、能力增长空间有限
- 最常用的计算只有少数几种
- 专用芯片有很大的发展空间
- 未来的计算能力可能主要靠高度整合的异构架构

开发高效好用的算力服务产品，为科学计算、微尺度工业设计场景助力



Table 5. Calculation time efficiency comparison between the proposed NVNMD and the established MLMD/CMD.

Method	System	Computer architecture	Hardware resource	T_i (s step ⁻¹ atom ⁻¹)	η_i (s step ⁻¹ atom ⁻¹)
NVNMD (this work)	Benzene	vN	FPGA	2.0×10^{-7}	2.0×10^{-7}
	Naphthalene		FPGA	2.0×10^{-7}	
	Aspirin		FPGA	2.0×10^{-7}	
	Antimony (bulk)		FPGA	2.0×10^{-7}	
	Germanium telluride (bulk)		FPGA	2.0×10^{-7}	
	Li-Ge-P-S (bulk)		FPGA	2.0×10^{-7}	
MLMD	H ₂ O ¹²³	vN	GPU	5.6×10^{-5}	2.4×10^{-5}
	SiO ₂ ¹²³		80 CPU cores	3.6×10^{-5}	
	Cu (original DP) ¹²⁴		GPU	2.8×10^{-5}	
	H ₂ O (original DP) ¹²⁴		GPU	9.5×10^{-6}	
	Al-Cu-Mg (original DP) ¹²⁴		GPU	8.7×10^{-5}	
	Cu (compressed DP) ¹²⁴		GPU	2.8×10^{-6}	
	H ₂ O (compressed DP) ¹²⁴		GPU	2.6×10^{-6}	
	Al-Cu-Mg (compressed DP) ¹²⁴		GPU	5.4×10^{-6}	
	GeTe (compressed DP) ²⁵		GPU	3.7×10^{-6}	
	Li-Ge-P-S (compressed DP) ¹²		GPU	9.4×10^{-6}	

保持第一性原理精度的前提下，
相较GPU等“冯·诺依曼”芯片，
第1版NVNMD（基于FPGA）：

- 速度提升1-2个数量级
- 能耗降低2-3个数量级

Table 6. Calculation energy efficiency comparison between the proposed NVNMD and the established MLMD/CMD.

	T (s step ⁻¹ atom ⁻¹)	P (W)	$\eta = T \times P$ (J step ⁻¹ atom ⁻¹)
NVNMD (this work)	2.0×10^{-7}	108	2.1×10^{-5}
MLMD	1 CPU + 1 GPU ^{26,31}	280	1.6×10^{-2}
	Summit supercomputer ³¹	7.6×10^6	2.1×10^{-3}
CMD	$\approx 10^{-7}$	$\approx 10^2$	$\approx 10^{-5}$

T is the calculation time efficiency. P is the power consumption. The NVNMD (this work) uses three CPU cores (Intel i7-10700K) and one FPGA (Xilinx xcvu9p). The MLMD uses CPU (IBM POWER9) and GPU (NVIDIA Tesla V100), as introduced in Ref. ^{26,31}. The CMD results are taken from Table 5.

[1] P.H. Mo et.al. "Accurate and efficient molecular dynamics based on machine learning and non von Neumann architecture", npj Computational Materials, 2022.
<https://www.nature.com/articles/s41524-022-00773-z>

科学知识来自于文献和各类数据库（如PDB）

知识库： 文献内容最有效的表达形式

简洁明了，便于自动化搜索和发展AI算法

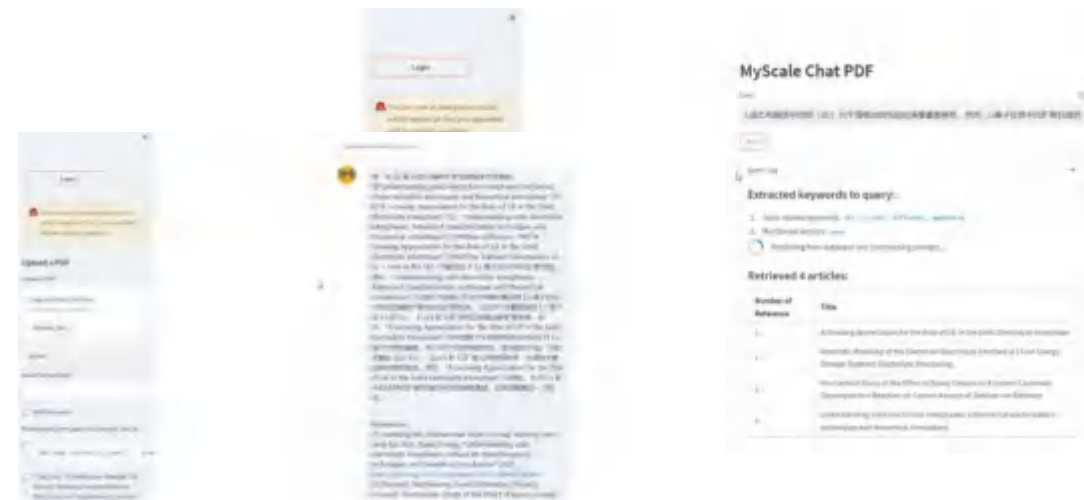
构建智能化的文献知识库

实现科学文献的分类检索与智能推荐

实验表征数据、计算模拟数据、文献数据等
多模态数据库与知识库

Science Navigator 对话式文献知识库

进一步助力科研人员提升科研生产力，
释放更多的时间精力在解决关键问题与创新思考上



文献阅读目标

- 掌握领域现有的基本情况，发展阶段、现有的进展、主要的问题等。主要的研究手段，主要的分析指标、理论或技术的适用范围等。

文献的类型

- **综述类文献**：快速对某个研究领域形成**整体的认识**和了解，把握**领域核心脉络**，找到合理的切入点和方向。
- **进展类文献**：了解文献的**主要结论**，或提出的**新的理论**。以及**研究的方法**，使用的材料和工具，证明结论的逻辑。从而对自己的研究项目形成**支持、启发或对比**，帮助优化研究思路。

文献的构成

- **正文**：文字、公式、化学式、化学反应
- **图片**：示意图 & 数据图（文字、公式、化学式、化学反应）
- **表格**：文字、公式、化学式、化学反应

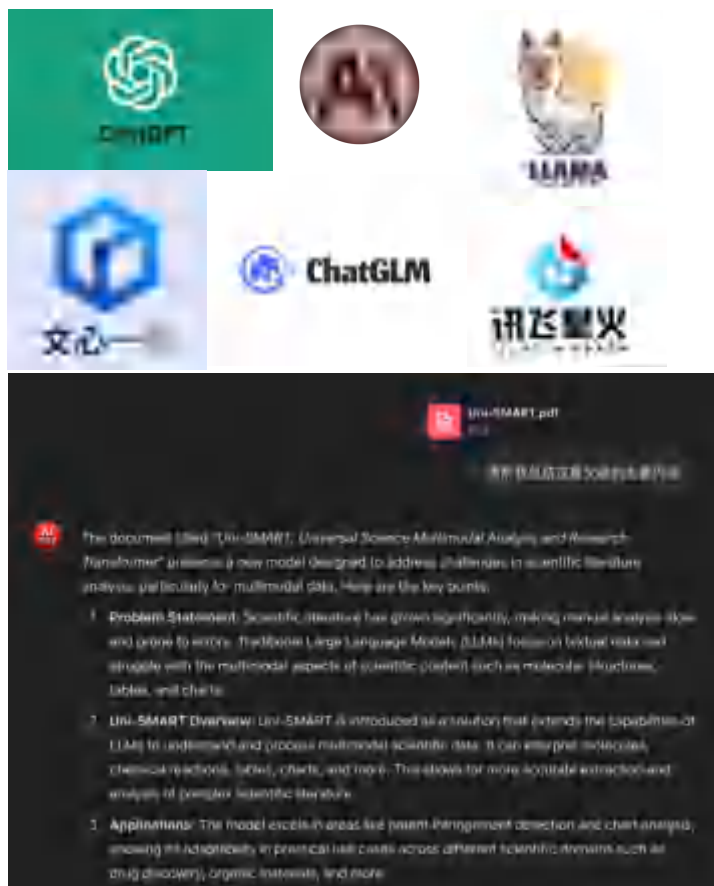


文献阅读与知识提取完全依赖研究人员的手、眼、脑

1 电脑阅读效率低，关键点易遗漏

2 图表数据结构化提取易出错

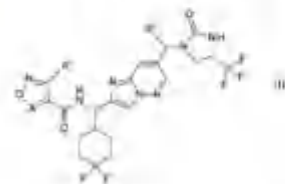
迫切需要智能化工具提升科学信息的获取效率



ChatGPT 4

You

(b) Title: (MIDAZOLAM) (5-SPYRIDAZINONE)-ITA INHIBITORS.



(27) Abstract: The invention provides certain dihydrocyclo-
hex-1-enylmethyl-oxazol-5-one derivatives consisting of hetero-
lytic or 1,3-oxazol-5-one derivatives, pharmaceutical compositions thereof,
and methods of using a composition of treating it to treat certain
symptoms of psychosis, pharmacological systems or analysis systems.

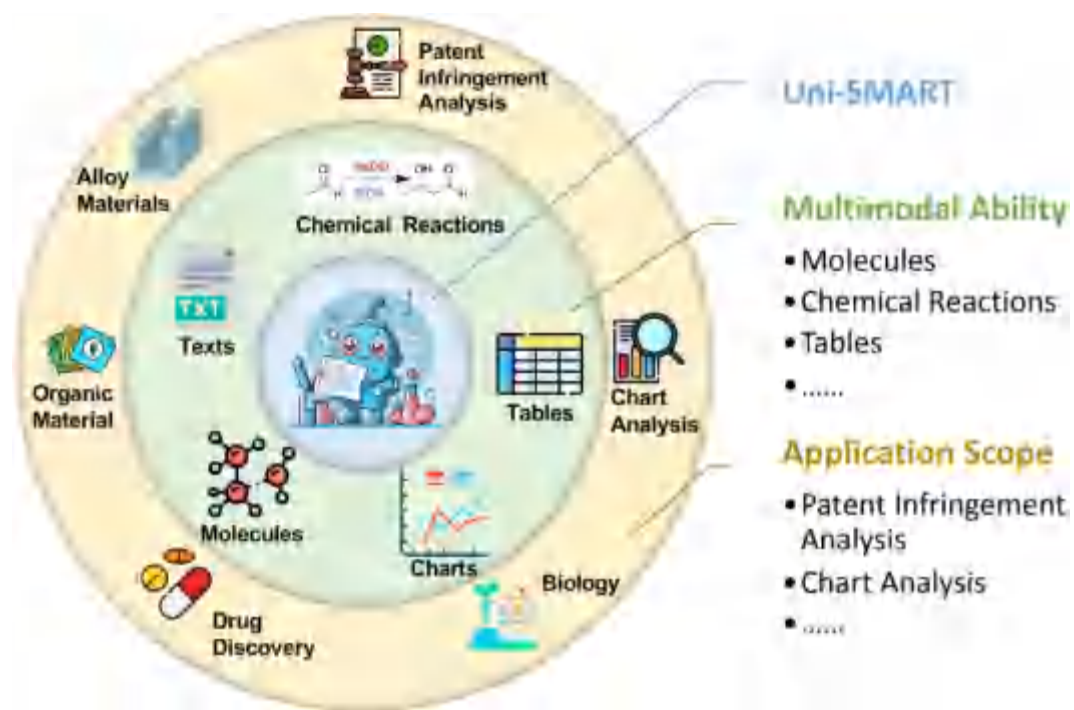
请将专利中的分子结构以smiles形式表达

ChatGPT

将图片中的分子结构转换为SMILES（简化分子输入线索规范）格式是一项挑战性任务，因为它要求将二维图像中的化学结构识别并翻译成文字形式的化学语言。不幸的是，我目前没有直接将图像转换为SMILES的能力。这通常需要专业的化学软件或人工智能模型，这些工具能够识别化学结构并将其转换成相应的文本格式。如果您有化学背景，您可以尝试手动转换，或者您也可以使用化学结构识别软件尝试这样做。如果您需要，我可以帮您理解分子结构的组成，以便您能够手动编写SMILES。

本质上是大语言模型在科学文献领域多模态能力“不足”

然而通用大语言模型在处理科学文献时会“水土不服”





UNI-SMART: UNIVERSAL SCIENCE MULTIMODAL ANALYSIS AND RESEARCH TRANSFORMER

Hengqiang Cai¹, Haochen Cai¹, Shuyen Yang¹, Jinkun Wang², Lin Yao¹, Zhifeng Gao¹,
 Junhao Chang¹, Sihang Li¹, Mingjun Xu¹, Chengxin Wang¹, Honghui Wang¹,
 Yougu Li¹, Mingji Lin¹, Yaqi Li¹, Yuyi Yu¹, Zheng Chang², Zifeng Zhao¹, Linfeng Zhang^{1,3} and Guolin Ke¹
¹IP-Technology, ²AI for Science Institute, Beijing

hengqiangcai, haochencai, yangsh, wangjinkun, yao, gaozf, shuyeh, liyaoh, yongu, wangchengxin, wanghonghui, liyonggu, xumingjun, liyq, yanyuqi, zhanglf, kegl@ip.tech

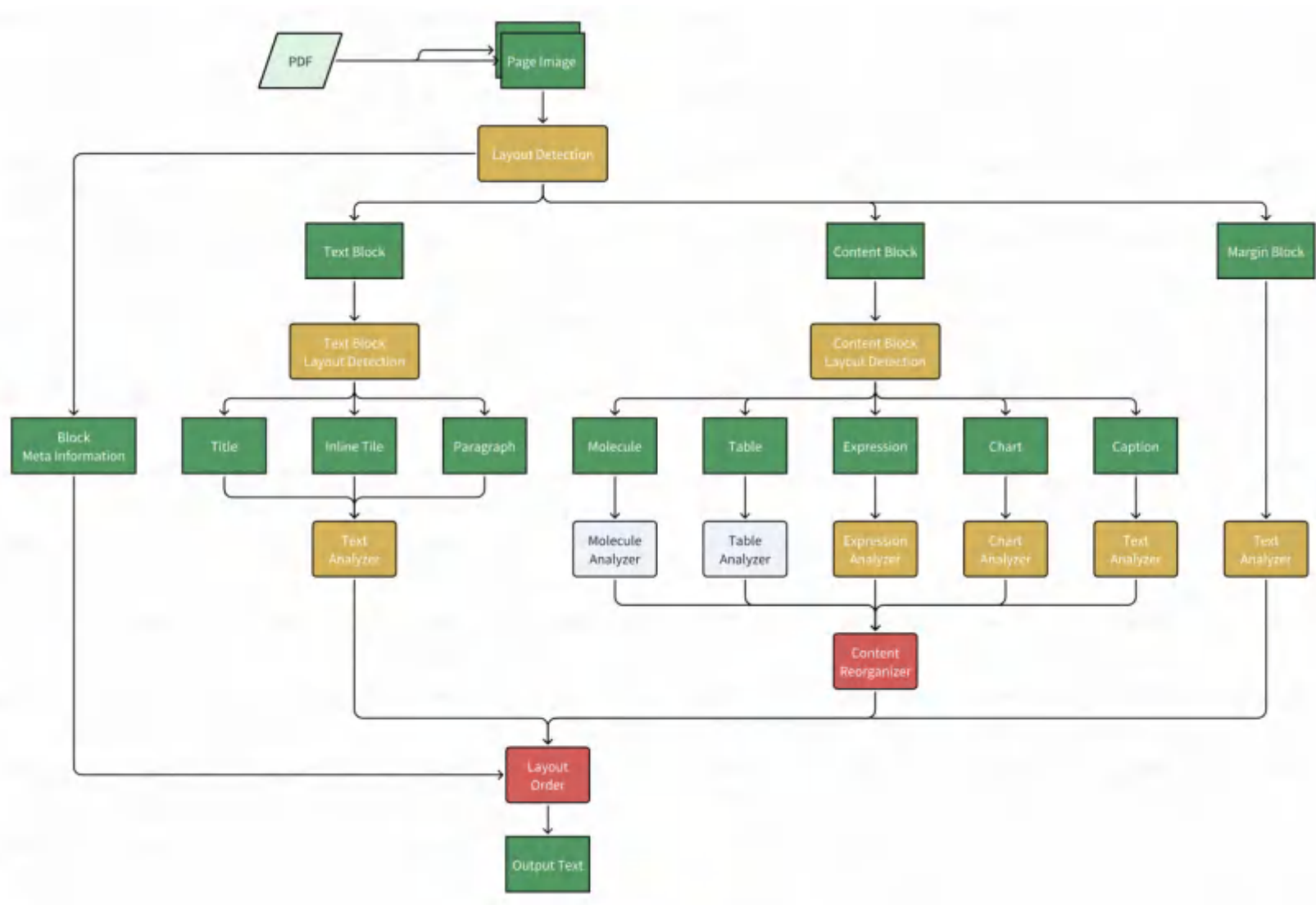
ABSTRACT

In scientific research and its applications, scientific literature analysis is crucial as it allows researchers to build on the work of others. However, the fast growth of scientific knowledge has led to an massive increase in scholarly articles, making in-depth literature analysis increasingly challenging and time-consuming. The emergence of Large Language Models (LLMs) has offered a new way to address this challenge. Known for their strong abilities in summarizing texts, LLMs are seen as a potential tool to improve the analysis of scientific literatures. However, existing LLMs have their own limits. Scientific literature often includes a wide range of multimodal elements, such as tables, charts, and molecules, which are hard for text-focused LLMs to understand and analyze. This issue points to the urgent need for new solutions that can fully understand and analyze multimodal content in scientific literature. To answer this demand, we present **Uni-SMART** (Universal Science Multimodal Analysis and Research Transformer), an innovative model designed for in-depth understanding of multimodal scientific literature. Through rigorous quantitative evaluation across several domains, Uni-SMART demonstrates superior performance over other text-focused LLMs. Furthermore, our exploration extends to practical applications, including patent infringement detection and nuanced analysis of charts. These applications not only highlight Uni-SMART's adaptability but also its potential to revolutionize how we interact with scientific literature.

- 丰富的数据源
- 主动学习
 - 多模态学习
 - 继续预训练
 - 大模型有监督微调
 - 用户反馈
 - 专家标注
 - 数据增强
- 解决各种任务

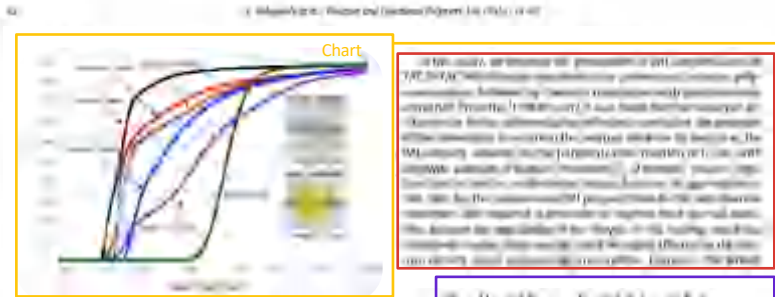


Uni-SMART: multimodal learning



Uni-SMART: multimodal learning

图表解析为文字描述



Convert to Caption

UV-Vis spectra and the digital images of PAI films. Film thickness: Poly(h-TAC-ODA), 66 μm; poly(h-TAC-pPDA), 110 μm; poly(h-TAC-mPDA), 90 μm; poly(h-TAC-BAFL), 100 μm; poly(h-TAC-APB), 85 μm; poly(TAC-ODA), 34 μm.

Convert to Markdown

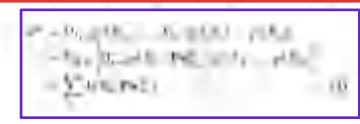
In this study, we reported the preparation of PAI samples based on TAC/h-TAC with diamine monomers in a conventional solution polycondensation, followed by chemical imidization with pyridine/acetic anhydride. From the ^1H NMR study, it was found that the monomer addition order for the polymerization effectively controlled the sequence of the monomeric structure in the polymer skeleton. In particular, the PAI samples obtained by the polymerization reaction of S -TAC with dropwise addition of diamine monomer (T-D method) possess regulated imide-amide-imide linkage fractions of approximately 574%–96%. For the semiaromatic PAI prepared from h-TAC with diamine monomers, this sequence is important to improve their thermal stabilities, because the degradation of the alicyclic h-TAC moiety, which has a relatively smaller bond energy, could be highly affected by the electron density of the surrounding environment. Therefore, the wholly

文本和公式识别为 Markdown/LaTeX

Molecule



Equation



Convert to LaTeX

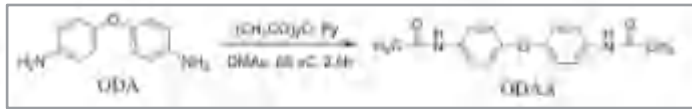
$$\mathcal{M}(\mathcal{G}) = \sum_{i=1}^n \frac{1}{i^2} \mathcal{M}(\mathcal{G}_i)$$

Convert to SMILES

<smiles>Nc1ccc(Oc2ccc(N)cc2)cc1</smiles>

分子识别为 SMILES

Molecule Reaction



Convert to Graph

```
<reaction>
prev_mols: <smiles>Nc1ccc(Oc2ccc(N)cc2)cc1</smiles>
post_mols: <smiles>CC(=O)Nc1ccc(Oc2ccc(NC(=O)O)cc2)cc1</smiles>
condition: [(CH3CO)2O,Py,DMAC,85 oC, 2.5h]
</reaction>
```

反应式识别为 SMILES 关联图

Text

In the ^1H NMR spectra of ODA, the amide hydrogen δ is observed at 8.95 ppm as a doublet signal together with the two clear doublet signals. The δ of 6.95 and 7.6 ppm, corresponding to the aromatic protons, the δ of 7.0 and 7.7 ppm, corresponding to the amide hydrogen δ is observed at 8.95 ppm as a doublet signal together with the aromatic protons δ of 6.95 and 7.6 ppm. The split of the amide signal could be attributed to the effect of the conformation of the cyclohexane ring. Therefore, the amide signal observed for the amide hydrogen in the poly(h-TAC-ODA) sample synthesized via the T-D method seems reasonable, and the two doublet signals for the poly(OOA-h-TAC) sample (D-T method) can be attributed to the amide hydrogen in the polymer structure. In the field region of the ^1H NMR spectra of APCL and the polymer, the aliphatic protons appear as complex signals due to the stereo structures (Fig. S1). The configuration was maintained after the polymerization, and thus the properties of the polymer had arisen from the stereo mixtures.

signals at the early stage of the polymerization at about 6.95 ppm as a doublet signal together with the two clear doublet signals. The δ of 6.95 and 7.6 ppm, corresponding to the aromatic protons, the δ of 7.0 and 7.7 ppm, corresponding to the amide hydrogen δ is observed at 8.95 ppm as a doublet signal together with the aromatic protons δ of 6.95 and 7.6 ppm. The split of the amide signal could be attributed to the effect of the conformation of the cyclohexane ring. Therefore, the amide signal observed for the amide hydrogen in the poly(h-TAC-ODA) sample synthesized via the T-D method seems reasonable, and the two doublet signals for the poly(OOA-h-TAC) sample (D-T method) can be attributed to the amide hydrogen in the polymer structure. In the field region of the ^1H NMR spectra of APCL and the polymer, the aliphatic protons appear as complex signals due to the stereo structures (Fig. S1). The configuration was maintained after the polymerization, and thus the properties of the polymer had arisen from the stereo mixtures.

Text

Convert to Markdown

2). In the ^1H NMR spectra of ODA, the amide hydrogen δ is observed at 8.95 ppm as a doublet signal together with the two clear doublet signals. The δ of 6.95 and 7.6 ppm, corresponding to the aromatic protons. For APCL, the amide hydrogen δ is observed at 8.95 ppm as a doublet signal together with the aromatic protons δ of 6.95 and 7.6 ppm. The split of the amide signal could be attributed to the effect of the conformers of the cyclohexane ring. Therefore, the aliphatic protons observed for the amide hydrogen in the poly(h-TAC-ODA) sample synthesized via the T-D method seems reasonable, and the two doublet signals for the poly(OOA-h-TAC) sample (D-T method) can be attributed to the monomer sequences in the polymer structure. In the field region of the ^1H NMR spectra of APCL and the polymer, the aliphatic protons appear as complex signals due to the stereo structures (Fig. S1). The configuration was maintained after the polymerization, and thus the properties of the polymers had arisen from the stereo mixtures.

表格识别为 LaTeX

Run	Method	h-TAC/ODA	ODA	δ (ppm)	δ (ppm)	δ (ppm)	δ (ppm)	δ (ppm)	δ (ppm)
1	T-D	1.00	1.00	4.00	5.6	4.1	4.7	4.4	4.7
2	T-D	1.00	1.00	4.100	5.6	4.15	4.68	4.4	4.7
3	T-D	1.00	1.00	4.000	5.6	4.1	4.7	4.4	4.7
4	T-D	1.00	1.00	4.000	5.6	4.1	4.7	4.4	4.7
5	T-D	1.00	1.00	4.000	5.6	4.1	4.7	4.4	4.7
6	T-D	1.00	1.00	4.000	5.6	4.1	4.7	4.4	4.7

Convert to LaTeX

```
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline
\text{Run} & \text{Method} & \text{h-TAC/ODA} & \text{ODA} & \delta \text{ (ppm)} & \delta \text{ (ppm)} & \delta \text{ (ppm)} & \delta \text{ (ppm)} & \delta \text{ (ppm)} & \delta \text{ (ppm)} \\
\hline
1 & T-D & 1.00 & 1.00 & 4.00 & 5.6 & 4.1 & 4.7 & 4.4 & 4.7 \\
2 & T-D & 1.00 & 1.00 & 4.100 & 5.6 & 4.15 & 4.68 & 4.4 & 4.7 \\
3 & T-D & 1.00 & 1.00 & 4.000 & 5.6 & 4.1 & 4.7 & 4.4 & 4.7 \\
4 & T-D & 1.00 & 1.00 & 4.000 & 5.6 & 4.1 & 4.7 & 4.4 & 4.7 \\
5 & T-D & 1.00 & 1.00 & 4.000 & 5.6 & 4.1 & 4.7 & 4.4 & 4.7 \\
6 & T-D & 1.00 & 1.00 & 4.000 & 5.6 & 4.1 & 4.7 & 4.4 & 4.7 \\
\hline
\end{tabular}
```


Uni-SMART在科研信息任务中效果显著

- Uni-Finder 拥有极其丰厚的科研大模型数据底座：Uni-Smart科学文献多模态大模型

200篇文献专家评测准确率

Uni-SMART 81 %

GPT-4 62 %

Table 2: Evaluation on table data.

Domain	Task	Metric	Uni-SMART	GPT-4	GPT-3.5	Gemini
Alloy Materials	Composition Enumeration	Value Recall	0.451	0.420	0.437	0.443
Drug Discovery	Affinity Data Extraction	Value Recall	0.320	0.274	0.350	0.175
Organic Materials	Electrolyte Solubility data extraction	Value Recall	0.382	0.365	0.393	0.390
	Electrolyte Table QA	Value Recall	0.674	0.165	0.140	0.235
	Polymer Property Extraction	Value Recall	0.869	0.762	0.286	0.506
	OLED Property Enumeration	Value Recall	0.526	0.520	0.479	0.394

处理表格任务的结果对比

Table 3: Evaluation on chart data.

Domain	Task	Metric	Uni-SMART	GPT-4	GPT-3.5	Gemini
Alloy Materials	Alloy ChartQA	Accuracy	0.667	0.400	0.267	0.333
Biology	Biology ChartQA	Accuracy	0.467	0.333	0.400	0.400
Drug Discovery	Drug ChartQA	Accuracy	0.467	0.400	0.333	0.400
Organic Materials	Polymer ChartQA	Accuracy	0.733	0.667	0.533	0.007

处理图表任务的结果对比

Table 4: Evaluation on molecule data.

Domain	Task	Metric	Uni-SMART	GPT-4	GPT-3.5	Gemini
Drug Discovery	Affinity Data Extraction	Value Recall	0.320	0.274	0.359	0.175
	Tag to Molecule	Mean Similarity	0.275	0.000	0.007	0.000
	Molecule in Document	Accuracy	0.849	0.022	0.489	0.500
	Markush to Molecule	Mean Similarity	0.629	0.545	0.459	0.199
Organic Materials	OLED Property Extraction	Value Recall	0.526	0.520	0.479	0.394

处理分子结构任务的结果对比

Table 5: Evaluation on chemical reactions data.

Domain	Task	Metric	Uni-SMART	GPT-4	GPT-3.5	Gemini
Drug Discovery	Reaction QA	Accuracy	0.400	0.200	0.000	0.133
Organic Materials	Reaction Mechanism QA	Accuracy	0.445	0.352	0.345	0.375

处理化学反应任务的结果对比



知识/信息检索

快速筛选出相关的文献、专利或知识库条目，提高检索效率，提供**个性化推荐**



分子结构的自动提取



交互式解答

支持用户进行**实时多模态交互**，从海量信息中提炼关键点，自动分析和整理数据

主要内容、研究观点、数据依据.....公式、分子式、分子结构图、数据表格.....



数据统计/聚合

对海量数据 (领域、来源、格式) 进行**跨文本统计和聚合**

输入格式包括化合物结构、图片、SMILES、自定义上传的专利和文献

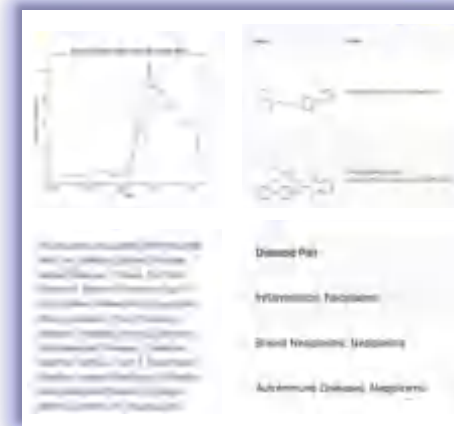


实时更新

订阅式收集/跟踪数据库
通过持续学习和训练，可以提供最新的数据

Uni-Finder 文献阅读和信息提取系统

1. 数据来源包括公开的文献、专利数据
2. 可接入客户内部购买数据库以及支持自定义上传文献。持续接入更多的数据源.....



➤ 支持折线图、表格、文本信息的输出

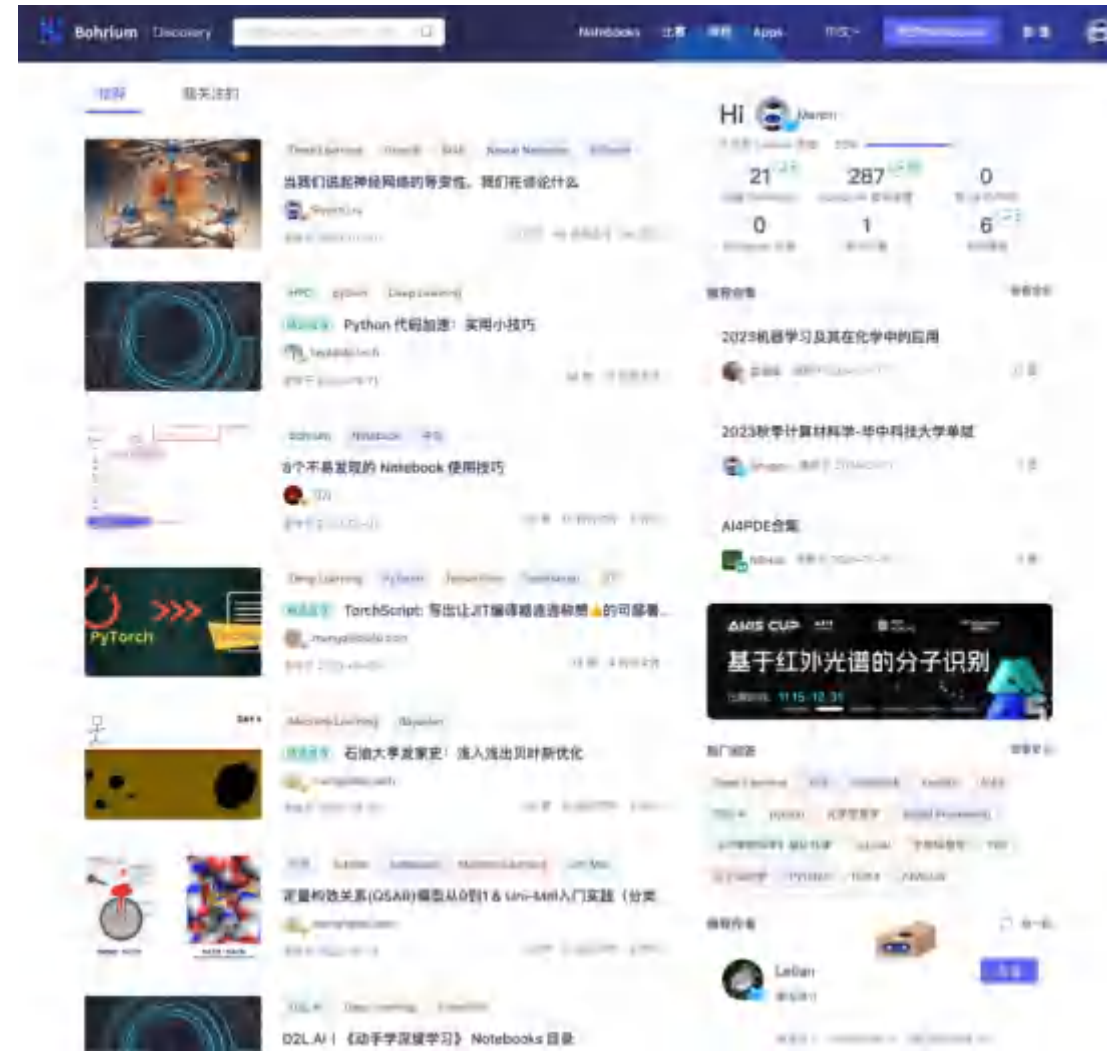
1.1 AI4S的第一类寻宝图：AI建模高维复杂函数

1.2 AI4S的第二类寻宝图：AI驱动平台化科研

1.3 AI4S该怎么学



授之以鱼，不如授之以渔



从实践中来，到实践中去

- 《计算材料学》（单斌老师）
 - 课程视频 + 课件 + 案例 + 教材
 - 学会从头开始编写分子动力学软件

课程内容探索上：系统推进AI4S课程建设

- 《计算材料学》(2023版)
- 课程背景
- 01. 课程背景
- 02. 课程背景-计算材料学简介
- 03. 课程背景-计算材料学简介
- 04. 课程背景-计算材料学简介
- 05. 课程背景-计算材料学简介
- 06. 课程背景-计算材料学简介
- 07. 课程背景-计算材料学简介
- 08. 课程背景-计算材料学简介
- 09. 课程背景-计算材料学简介
- 10. 课程背景-计算材料学简介
- 11. 课程背景-计算材料学简介
- 12. 课程背景-计算材料学简介
- 13. 课程背景-计算材料学简介
- 14. 课程背景-计算材料学简介
- 15. 课程背景-计算材料学简介
- 16. 课程背景-计算材料学简介
- 17. 课程背景-计算材料学简介
- 18. 课程背景-计算材料学简介
- 19. 课程背景-计算材料学简介
- 20. 课程背景-计算材料学简介



- 《计算材料学》(2023版)
- 01. 课程背景-计算材料学简介
- 02. 课程背景-计算材料学简介
- 03. 课程背景-计算材料学简介
- 04. 课程背景-计算材料学简介
- 05. 课程背景-计算材料学简介
- 06. 课程背景-计算材料学简介
- 07. 课程背景-计算材料学简介
- 08. 课程背景-计算材料学简介
- 09. 课程背景-计算材料学简介
- 10. 课程背景-计算材料学简介
- 11. 课程背景-计算材料学简介
- 12. 课程背景-计算材料学简介
- 13. 课程背景-计算材料学简介
- 14. 课程背景-计算材料学简介
- 15. 课程背景-计算材料学简介
- 16. 课程背景-计算材料学简介
- 17. 课程背景-计算材料学简介
- 18. 课程背景-计算材料学简介
- 19. 课程背景-计算材料学简介
- 20. 课程背景-计算材料学简介

5. 周期性边界条件代码实现

计算材料学 课程 华中科技大学 python

starforbehan 更新于 2024-05-21 课程编号: Basic ImageJohann... 讲师信息: cz_ma_gpu

Copyright 2023 @ Authors
作者: 单斌老师, 徐仁杰
日期: 2023-08-09
共享协议: 本作品采用知识共享署名-非商业性使用-相同方式共享 4.0 国际许可协议进行许可。
快速开始: 点击上方 [GitHub](#) 按钮, 选择 [johann-notebook/2023-04-07](#) 镜像及 `v2_m4_gpu` 节点配置, 稍后...

单斌自己总结了这份神奇的计算材料学课件! 这份课件是在熬夜吃泡面不计其数的故事和咖啡后创作出来的。配套的教材网址请到: www.materialsimulation.com/book, 您也可以跟着up主人阿源做没有烦恼的edc课程一起学习, 希望它能给您带来一些比烦恼多一点写代码的大多一样, 当您使用这份课件时, 请:

- 记得告诉大家这份课件是单斌老师写的, 并且在华中科技大学创新中心
- 别用它去赚大钱, 这个课件是免费学习的, 不是用来买股票的
- 保持开放共享的精神

如果您有关于计算材料学的想法, 或者您只是想和我讨论最好吃的披萨口味, 欢迎通过邮件 shiban@mail.hust.edu.cn 联系我!

Particle 类

这个 Particle 类是一个用于表示粒子的类, 以下是该类的详细说明:



玻尔[®] 科研空间站

Bohrium

登录空间站，探索 AI for Science 科研宇宙

➔ 开始体验

教学平台

AI4S的教材与教具

模型库 + 算法库 + 案例库
深势自研 + 社区贡献

科研平台

AI4S科研的生产资料

AI4S算法引擎 + 高性能算力 + 计算模拟环境 + 存储

开发平台

AI4S工程基础设施

预置开发工具 + 模块化开发场景 + Launching



国家关键领域工程硕博核心课程：人工智能辅助科学研究 (AI for Science)

- 课程内容覆盖包括：
 - 数据驱动的AI4S算法
 - 机理驱动的AI4S算法
 - 高效率高精度的实验表征技术
 - AI4S与高性能计算
 - AI4S生态建设



Uni-Mol性质预测实战-回归任务-有机/电解液

© Copyright 2023 @ Authors
作者: 姜海宇 - 已认证
日期: 2023-07-03
开源协议: 本作品采用 [CC BY-NC-SA 4.0](#) 国际版权协议进行许可。
快速开始: 点击上方 [GitHub](#) 按钮, 选择 `unimol-gsar-v0.4` 镜像及任意GPU节点配置, 脚本片刻即可运行。

案例背景

- 熔点 (Melting point) 用于描述一种物质从固态转变为液态所需要的温度。通常在恒压下, 当物质受热而由固态转变为液态时, 物质的温度并不上升, 直到所有的固态都已转变为液态后, 温度才会继续上升。
- 在电池领域, 电解液分子的熔点是一个衡量其稳定性和可用温度范围的重要物理量, 还早的电解液材料要求满足较宽的温度范围, 另外不同的应用场景需要电解液具有特定的熔点, 以满足特定的性能要求。
- 通过对未知分子的熔点进行预测, 将有助于我们从无限的化学空间中定向筛选可用作电解液的材料。

Code Markdown

Step1: 读入数据

- 给定近2万个分子的SMILES式和熔点预测值数据 (TARGET)

