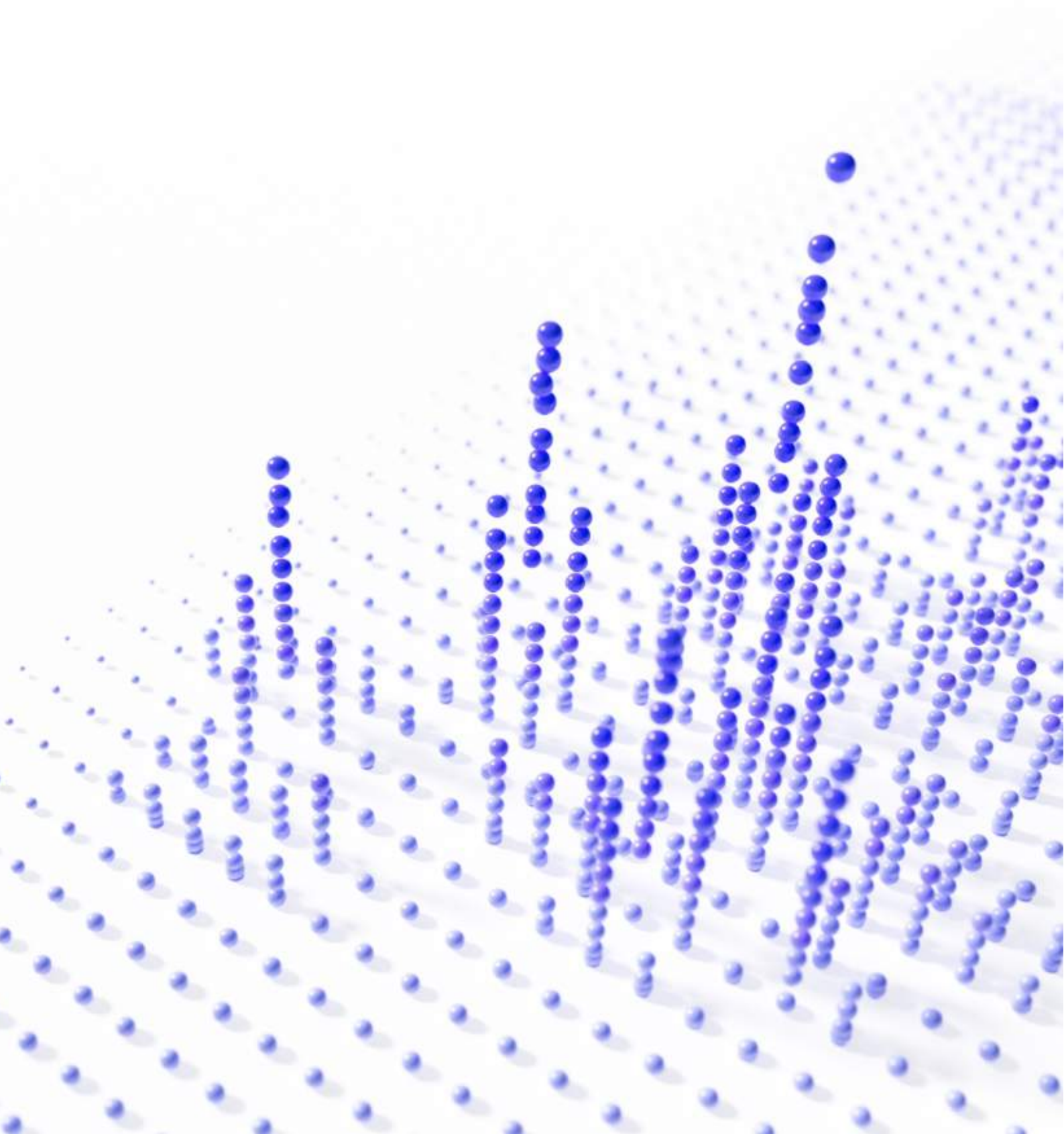
科学智能 (AI4S)
全球发展观察与展望

2023 版

AI for Science Global Outlook
2023 Edition

This version is updated on Aug 9, 2023



前言：AI for Science 已进入加速涌现期

2022 年底，ChatGPT 横空出世，并以超过 iPhone, 微信, 抖音等科技产品的速度在短短数周的时间内达到 1 亿用户。在随后的 2023 年，GPT 和其他生成式 AI 技术品类占据了科技讨论的绝对中心。从国内到国外，从几个人的初创企业到十万人的科技巨头，AI 的赛道上摩肩接踵。

2022 年的秋天，我们发布的第一版《AI for Science 全球发展观察与展望》（简称《展望》），获得了来自各界的认可和支持。在 AI 大发展的 2023 年，各界也在敦促我们对过去一年的发展进行追踪并阐述其如何影响 AI for Science 的发展路径。响应读者的热情，我们决定对《展望》进行重构，新增一倍的新内容，并对已有内容进行大幅更新，以反映相关技术、产品、产业、政策的演化，并以全新的框架重新梳理 AI for Science 的要素、沿革、展望，并详尽描述其在各行各业的实践。

10 年前的耶鲁大学礼堂上，Peter Thiel 说到："We wanted flying cars, instead we got 140 characters." 10 年后的今天，这句话依然成立。千禧年以来，海量的科学人才涌入互联网科技行业，带来生产和协作方式的变革，创造了巨量的财富。而随着互联网热潮渐渐褪去，人们重新把目光聚焦回科学本身，以及它们所映射的实体经济领域上。

科学技术是第一生产力。科技革命的历史波澜壮阔。过去三百年间，科学行进经历过数次系统性危机，正是这些危机的解决才带来了科学的突破，新科学经由新工具的应用和普及，进而带来生产力的大幅提升。时至今日，随着数字化时代的到来，科技创新转化为直接生产力的速度越来越快。面对纷繁复杂的现实世界，虽然数据越来越丰富，但是寻求简单、漂亮的洞见却变得越来越难；各行业的技术专家也深受困扰：控制和设计的自由度越来越多，"炒菜式"的试错和喊口号式的攻坚也越来越难以解决实际问题 ……

此时，AI 来了。从艾伦·图灵的系统思考开始，伴随着算法、算力、数据的融合发展，AI 在计算机视觉、自然语言处理，自动驾驶等领域大放异彩。但 AI 若想从一套"数据处理"工具，走向更加通用的"智慧"，则无法绕开"科学"这一人类智慧结晶中最精华的一部分。于是，一群人先行者开始探索用 AI 学习科学原理，解决科学问题的路径。他们发现，当下 AI 取得成就的本质原因是在算力和数据基础之上算法对高维函数处理能力的大幅提升，这一能力是应对当下系统性危机的关键；他们发现，AI 是驱散 Science 各领域的乌云的法宝，AI for Science（AI4S）会是 AI 的下一个主战场，它将极大地拓展 Science 和 AI 的边界；他们发现，AI4S 将赋能技术和工业的方方面面，帮助我们加快走完科学研究和技术创新之间的最后一公里，也将帮助科学家从纷乱的自然和社会特征之中抽丝剥茧，发现事物背后作用着的关键规律。

AI4S 的未来正在走向流行。AI 求解薛定谔方程、AI 求解控制论方程、AI 加速分子模拟、AI 预测蛋白结构、AI 赋能药物和材料设计…… 在 2022 年版《展望》发表不到一年的时间中，AI for Science 的发展依然超出了我们的预期：

在国内，2023 年科技部会同自然科学基金委启动"人工智能驱动的科学研究"（AI for Science）专项部署工作，布局"人工智能驱动的科学研究"前沿科技研发体系。科技创新 2030—"新一代人工智能"重大

项目也将 AI for Science 作为人工智能的重要发展方向进行安排。在指南中，部署了"重大科学问题研究的 AI 范式"任务，面向地球科学、空间科学、化学和材料科学、生物医药科学等领域重大科学问题开展创新研究。同时，面向国际竞争激烈的蛋白质结构预测领域，支持国内优势团队开展科研攻关。在平台支撑方面，科技部正在加快推动公共算力开放创新平台建设，将为 AI for Science 发展打造智能算力基座。

国际上， Elon Musk 官宣 x.ai，其理念为"建立理解自然规律的人工智能系统 (understand the true nature of the universe)"；前谷歌掌门人 Eric Schmidt 宣布捐出 1.48 亿美元 成立 AI for Science 博后奖学金，已布局 9 所大学；微软宣布成立专门的 AI4science 部门；英伟达联合 IIT 发布 AI for Science 公开课程；龙头药企赛诺菲宣布 "all-in" AI for (life) science；美国能源部联合 5 大国家实验室发布 AI for Science, Energy & Security 先进科研课题指引；OECD 面向全球政策制定者发布 AI in Science 的综述与政策建议 ..... 从学界到业界，从产业到政府，从生命科学的 RFDiffusion、到化工领域的 Open Catalyst、到材料科学的 Uni-Mol，各行各业的优秀 AI for Science 应用正在加速涌现，AI for Science 已成燎原之势。

AI4S 发展路上也将充满挑战。它呼唤各行各业的人们打破壁垒、凝聚共识、创造连接、形成迭代。拨云见日的路上，真理与泡沫、洞见与偏见差别，均在毫厘之间。站在这个科技革命的时代转角，北京科学智能研究院（AISI）、深势科技、全球最大 AI4S 开源社区 DeepModeling 的核心开发者与 AI4S 产业实践的先行者，和 AI4S 传播者络绎科学一道，整理了数十家领先企业和科研机构的经验方法，编写成册。道阻且长，行则将至。希望这份 2023 版《展望》将陪伴每一个关心与关注 AI for Science 的人度过每个春秋，见证这场发生在当下的科技革命，从花开花落走向硕果累累。

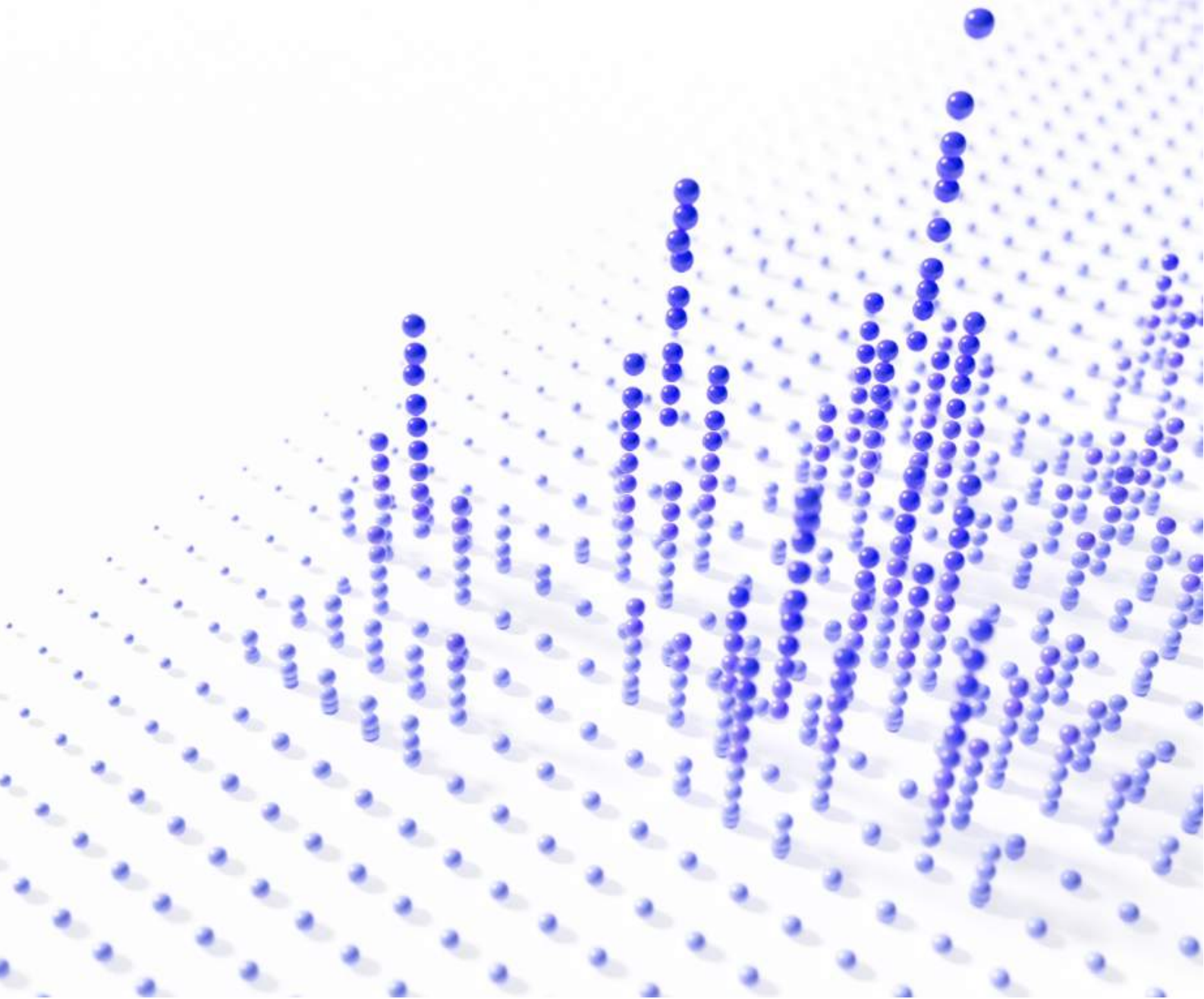
目录

出品团队

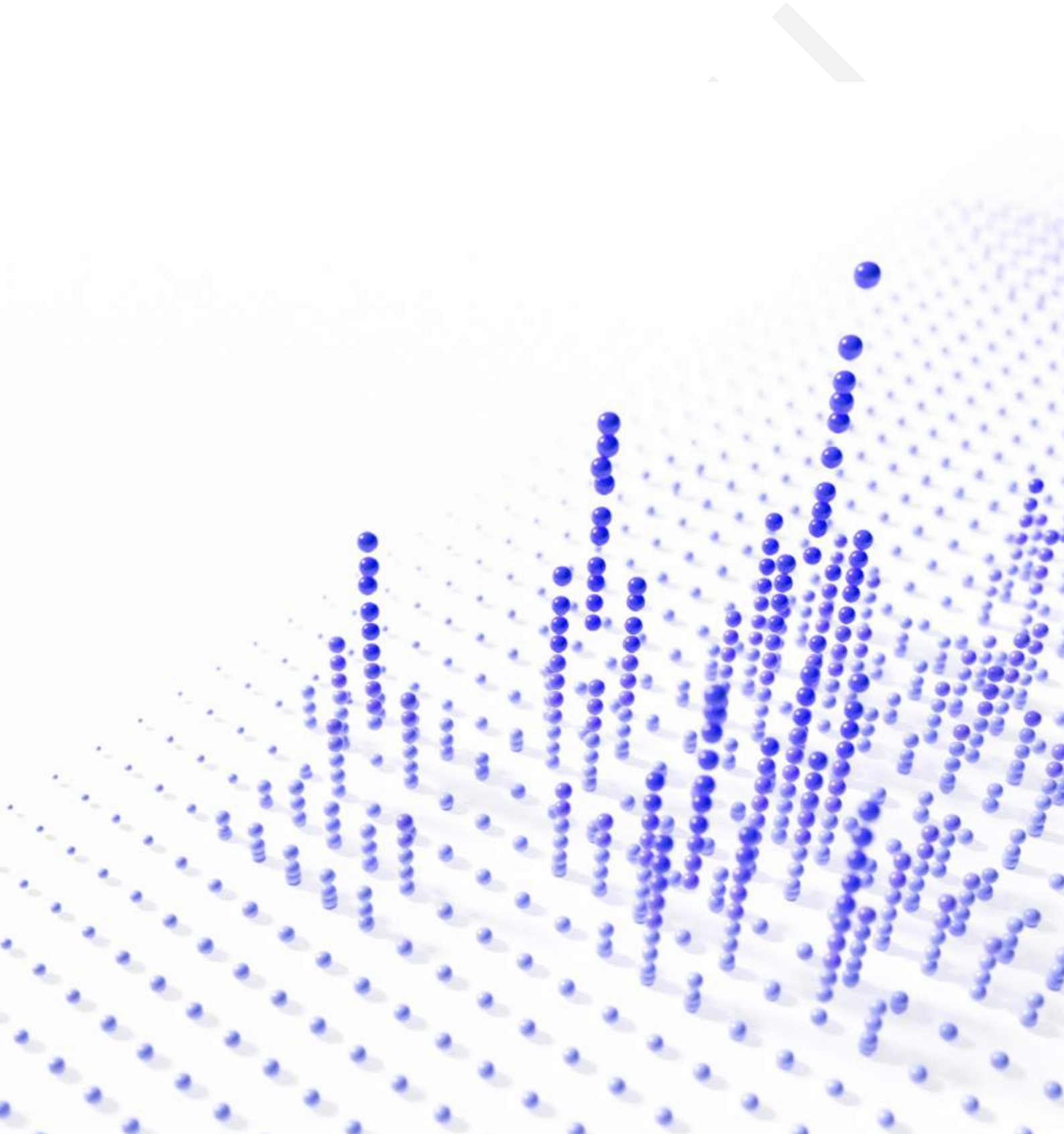
联席主编	张林峰 孙伟杰 李鑫宇 王小佛*
科学顾问	鄂维南
内容团队	白晓矿 陈帜 戴付志 邓杰 高志峰 胡太平 李航 刘杰 欧琪 许审镇 王一博 王晓旭 王冬冬 王宇航 王涵 王沁蕊 文通其 温瀚 向上 谢莹莹 宋宁 孙晓琦 张天汉 张与之 朱正诞
联合发布	北京科学智能研究院 深势科技 络绎科学
首席发布媒体	新华网

特别鸣谢	北京市科学技术委员会 北京大学

PART I:
AI for Science 原理与发展框架



第一章：AI for Science 的“四梁 N 柱“



1.1 什么是 AI for Science (AI4S)

2023 年这个夏天，全球遭受前所未有热浪、野火和洪水的极端天气攻击。面对这些挑战，英伟达采用 AI 技术，构建名为地球 2 号的"数字地球模型"，以更精确地预测这些极端事件。地球 2 号依赖于 FourCastNet AI 模型，利用了数十 TB 的地球系统数据，能以数千倍速度提高预测准确性，预测未来两周的天气状况。与一般只能生成大约 50 种未来一周预测的天气预测系统相比，FourCastNet 能预测出成千上万种可能性，准确捕捉罕见而致命的灾难风险，从而给弱势群体争取宝贵的准备和疏散时间。

事实上，气象科学仅是受益于 AI 发展的众多科学学科之一。AI 的出现正在带动科学研究的激动人心的转变，并且影响正在扩散到实验室之外，深入到我们所有人的生活中。如果我们能明智地采取行动，制定合适的监管措施，并适当支持 AI 在解决科学最紧迫问题方面的创新应用，AI 就有可能彻底改变科学过程。

这样的愿景，我们称之为 AI for Science。我们期待一个由 AI 驱动的未来，在这个未来，AI 工具可以解放我们从繁琐乏味和耗时的劳动中，同时引导我们进行创新性的发明和发现，促使本应需要几十年的突破提前实现。

近期，AI 的讨论几乎等同于大型语言模型(LLM)的讨论。随着 GPT 在各行各业的爆发，"是否能将 LLM 用于科研场景"成为了一个水到渠成的问题。当 ChatGPT 超越大部分人类在高考、SAT、美国法考、医考等领域取得令人咋舌的高分后，人们对于 LLM 驱动科研的兴趣愈发高涨。一方面，LLMs 使得知识的提取和综合变得高效、便捷。通过解密和呈现复杂的科学信息，LLM 大大降低了学者进入新领域的门槛，推动交叉学科的发展。另一方面，LLMs 可以加速并改进知识贡献的过程。利用 LLMs 进行多步推理和决策的能力，研究人员可以在科学文献的迷宫般的广度中找到最相关的论文。同时，LLMs 能提供语言方面的帮助，帮助构建逻辑叙述并确保连贯性，使得研究人能更从容的驾驭复杂的观点表述，从而促进世界范围科学的异步交流效率和规模。

然而，对 AI for Science 的讨论远不止步于 LLM 在科学领域的应用。究其根本，LLM 面向的是一维的字符串数据结构，而科学领域的数据类型纷繁多样，即有一维的基因序列，也有二维的分子图、三维的分子坐标、N 维的波函数。因此，在具体的科学领域中，使用专门的模型架构很可能比使用基于 LLM 的迁移模型更为直接有效。在过去的十年中，科学领域的大部分进步都源自于针对特定问题的模型。最近，人们开始使用融合专业领域知识和深度学习/预训练策略来构建更强大的领域专用模型。举例来说，McMaster 和 MIT 的科学家利用 AI 模型成功识别出了一种抗生素，该抗生素能够对抗世界卫生组织认为是对住院患者最危险的抗生素耐药细菌之一的致病菌。谷歌 DeepMind 的一个模型成功控制了核聚变反应中的等离子体，为清洁能源革命的到来更近一步。在医疗保健领域，美国 FDA 已经批准了 523 种使用 AI 的设备，其中 75%用于放射学。[1]

这些令人兴奋的研究，并不是无源之水，更不是"拿着锤子找钉子"的 AI 万能论。首先，将复杂的科学问题表述为 0101 的计算机语言本身就是极难的任务，需要能融合 "基本原理与数据驱动的算法模型和软件系统"；同

时，为了给 AI 提供高质量的训练数据，我们也需要 "高效率、高精度的实验表征系统"；第三，我们需要最大化利用 LLM 给科研效率带来的提升，建立 "替代文献的数据库与知识库系统"；第四，以上的智能系统都需要运行在 "高度整合的算力平台系统" 之上。以上的考量，我们将其概括称为 AI for Science 的"四梁"，而将 AI for Science 落地于各个学科和交叉学科领域的系统性工程，我们讲其统称为 "N 柱"。后续的章节会围绕着 "四梁 N 柱" 进行详尽的讨论。而完成"四梁 N 柱"的系统建设，一来要面临着高度抽象化的领域知识门槛，二来要摆脱"作坊模式"推动科研想"平台模式"转变，这其中科学问题与工程问题相互交织，相互影响，因此推动科学家与工程师的充分协作是高效实现 AI for Science 时代科研基础设施建设的关键因素。

	前计算机时代 (400BC - 1946)	计算机时代 (1946-2020)	AI4S 时代 (2020-)
主要科研方式	数学推演（纸笔） "假设" -- "实验"	将部分复杂科学问题转换为相对简单的计算问题实现粗粒度建模，在此基础上进行大量实验验证 [2]	利用 AI 求解高维函数的优势实现高精度高效建模、高通量筛选，并有针对性的进行实验验证
主要成就	经典物理模型、 量子力学的雏形	微观世界的初步探索、 宏观尺度科学成果的大规模应用（航空、汽车、能源、通讯等）	微观世界的多尺度探索、 宏观+微观尺度科学成果的应用（新材料、新能源、生化、信息）
主要瓶颈	缺少高效计算手段	维度灾难	"四梁"

Source:

[1] source: https://www.technologyreview.com/2023/07/05/1075865/eric-schmidt-ai-will-transform-science/

[2]: Mahoney, Princeton University, https://www.princeton.edu/~hos/Mahoney/articles/mathnat/mathnatfr.html

1.2 AI 与科学研究范式的变迁

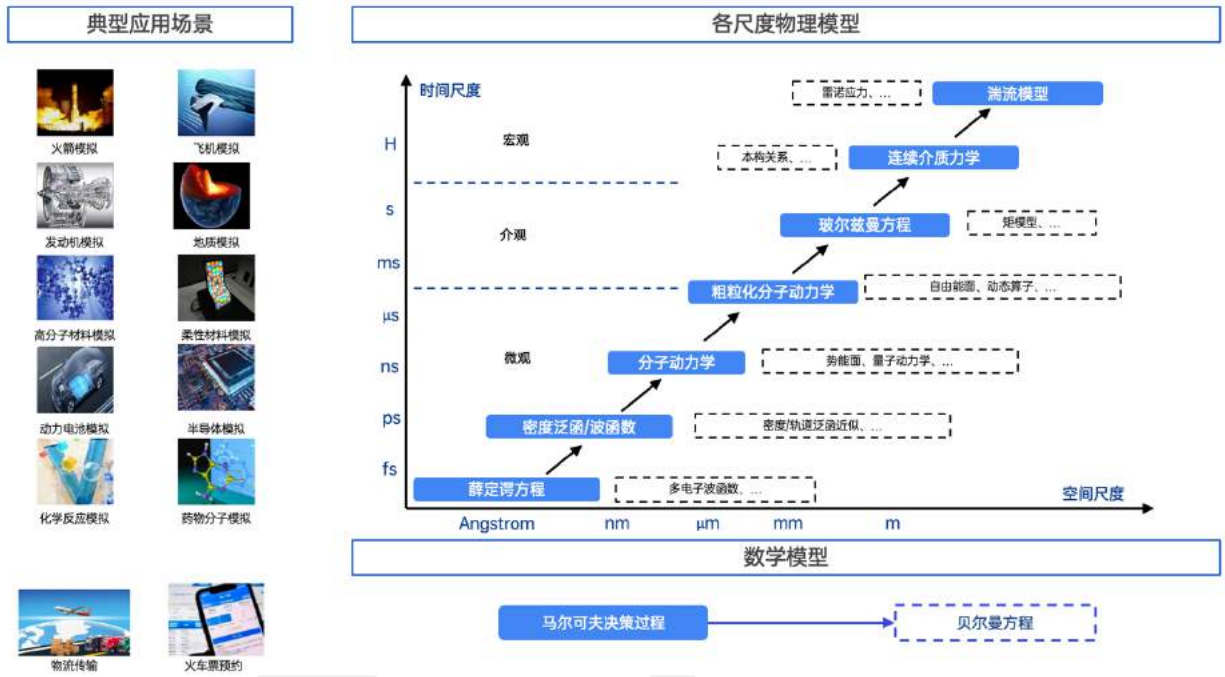
自文艺复兴以来，科学研究基本上是按照"开普勒范式"和"牛顿范式"这两种不同的范式展开：

- 开普勒范式是一种数据驱动的研究方式，通过对数据的分析寻找科学规律并解决实际问题，其经典案例是行星运动的开普勒定律；随着统计方法和机器学习的发展，数据驱动的"开普勒范式"研究方法已经成为一种非常强大的工具。数据驱动的研究方法能有效地帮助我们在缺乏明确原理的场景解决具体问题。但是该类方法的可解释性较弱，很难解释结论背后的原因。
- 牛顿范式是一种基于第一性原理的研究方式，其目标是发现物理世界的基本原理，其经典案例包括牛顿、麦克斯韦、玻尔兹曼、爱因斯坦、薛定谔等人的理论工作。[1] 对第一性原理的追求很大程度上驱动了物理学的发展。1929 年，随着量子力学的建立，这条道路出现了一个重大转折点：正如狄拉克[1] 所宣称的那样，有了量子力学，除一些极端尺度下的情形以外，我们已经掌握了大多数工程和自然科学所需要的基本原理。即便如此，当人们希望使用这些原理来求解真实场景的复杂物理模型时，往往发现所需的计算量过大，而陷入"空有原理缺无法有效使用"的境地。

从启蒙运动到工业革命到如今，上述两种泛式支撑了人类文明的演进并形成了今日丰富灿烂的经济社会。而在未来的发展中，AI 能扮演的角色，即是进一步推进科学在这两个泛式下的发展速度和高度。

I. 模型驱动：AI 加速计算求解

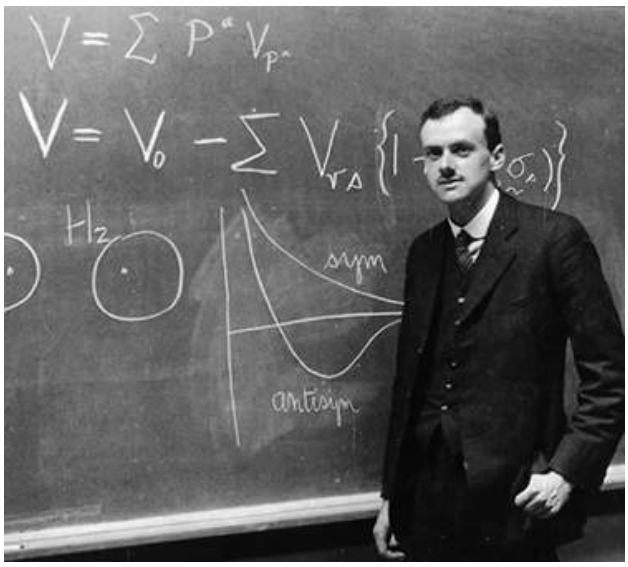
"牛顿范式"中，基于第一性原理的研究方法旨在从最基本的层面理解事物。对第一性原理的追求很大程度上驱动了物理学的发展。1929 年，随着量子力学的建立，这条道路出现了一个重大转折点：正如狄拉克[2]所宣称的那样，有了量子力学，除一些极端尺度下的情形以外（如核物理），我们已经掌握了大多数工程和自然科学所需要的第一性原理。



然而，也正如狄拉克所指出的那样，描述量子力学基本原理的数学问题异常复杂。困难之一在于它是一个多体问题：每加上一个电子，问题的维数便增加了三。事实上，第一性原理方法经常面临的困境是：尽管它很深刻，但它不太有用。因此，在实践中，我们常常不得不放弃严格优雅的理论，而采取经验的、非系统的近似方法。我们为此付出的代价不仅仅是丢失了严格和优雅，还有结果的可靠性和普适性。

二十多年前，多尺度、多物理建模的想法曾经给该问题的解决带来一线希望：通过将小尺度下的无关紧要的自由度整合起来，人们应当能够直接使用更可靠的微观尺度模型，为我们感兴趣的宏观尺度过程提出更为有效的算法。然而不同微观尺度模型本身也并不总是可靠，同时虽然多尺度方法能够大幅减少微观模拟所需时间，但仍然超出了目前的能力。这就意味着我们仍需要处理物理模型的新方法来应对"维度灾难"问题。

量子力学的奠基人之一 ——1933 年与薛定谔一起获得诺贝尔物理学奖的 Paul Dirac 曾这样表述科学研究的困境："The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."



Paul Dirac, Picture credit: AIP Emilio Sergè Visual Archives

不严谨的比喻一下，就是

"我们有了打开科学大门的钥匙，却没有力气去把门推开"

而"推不动"的原因，就是"维度灾难"

"维度灾难"是指在某些问题的求解中，随着维数的增加，计算代价会呈指数增长[2]。例如使用密度泛函理论（Density functional theory，DFT）求解势函数的计算代价会随着体系规模的增加而指数增长[3]。因此密度泛函理论的方法虽然准确，但难以应用到大规模体系的问题求解中。

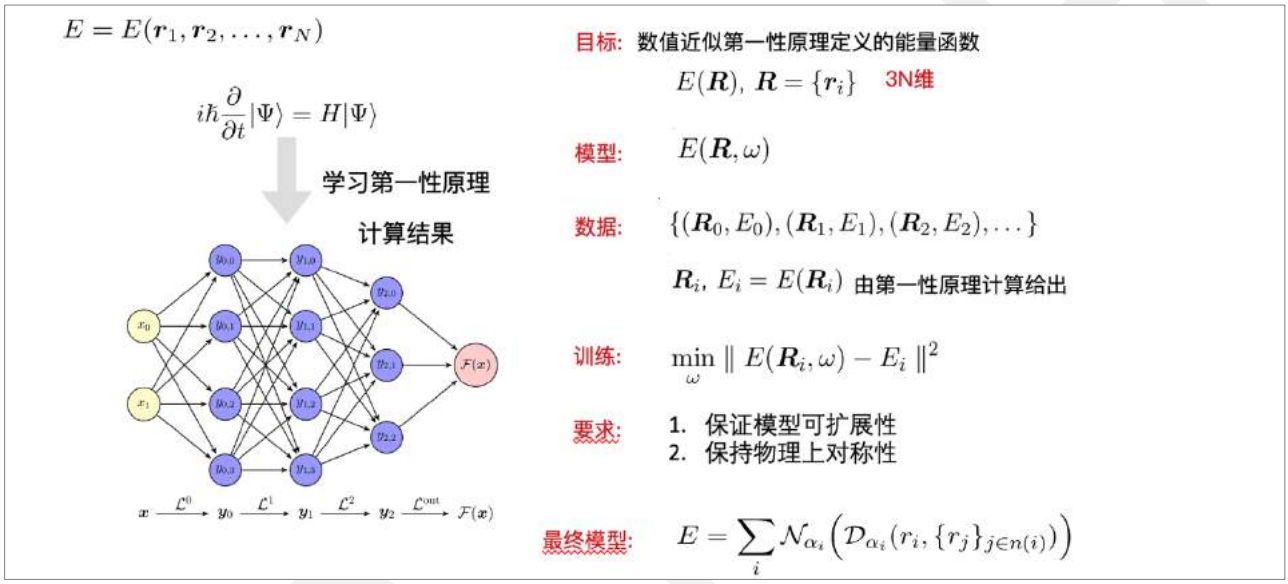
物理学中的基本原理不仅广泛适用，而且简洁优雅。薛定谔方程就是一个很好的例子。不幸的是，正如前面所指出的那样，使用这些模型来解决实际问题是一项极其困难的任务。因此，寻求简化模型一直是物理学乃至所有科学领域的一个永恒的主题。然而，正如我们在湍流模型中所经历的那样，如果不采取经验近似，我们通常很难提出这样的简化模型。

机器学习即将大大提高我们开发这种物理模型的能力。这其实已经以三种不同的方式发生了。第一，它提供了可以帮助我们把多尺度建模的梦想变为现实的工具。这个工具正是以前缺乏的。第二，它提供了直接从数据开发模型的框架。第三，顺着数据同化的思路，它将提供一个整合物理模型与观察数据的非常强大的工具。

然而，拟合数据是一回事，构建可解释且真正可靠的物理模型则是另外一回事。让我们首先讨论可解释性的问题。众所周知，机器学习模型有着"黑箱子"的名声，这为使用机器学习来帮助开发物理模型带来了心理障碍。为克服这一障碍，首先我们需要注意到可解释性并不是绝对的。以空气动力学中的欧拉方程为例。这些方程本身具有很清晰的解释，因为它们仅代表质量，动量和能量的守恒。但是，能否解释状态方程的细节则是另外一回事。事实上，复杂气体的状态方程可能是由一些实验数据经样条插值得到的，它以一个子程序的形式呈现。我们并不真正关心这些样条函数的系数是否可解释。相同的原则应当也适用于基于机器学习的模型。我们的目标应该是：这些模型的基本出发点和基本结构是可解释的，这些模型中代表本构关系的一些函数的具体形式未必都得可解释。

现在来谈谈可靠性问题。理想情况下，我们希望基于机器学习的模型和普通物理模型（如纳维 - 斯托克斯方程）一样可靠。要做到这一点，有两点至关重要。第一点是基于机器学习的模型必须满足所有物理约束，例如来自对称性和守恒律的约束。第二点是，我们用于训练模型的数据必须能充分代表实际中遇到的所有物理状态。由于对数据进行标记几乎总是非常昂贵的，因此选择一个既尽可能地小又具有充分代表性的优质数据集是此类模型开发过程中的一个非常重要的组成部分。我们将在下一节中对此做更多阐述。

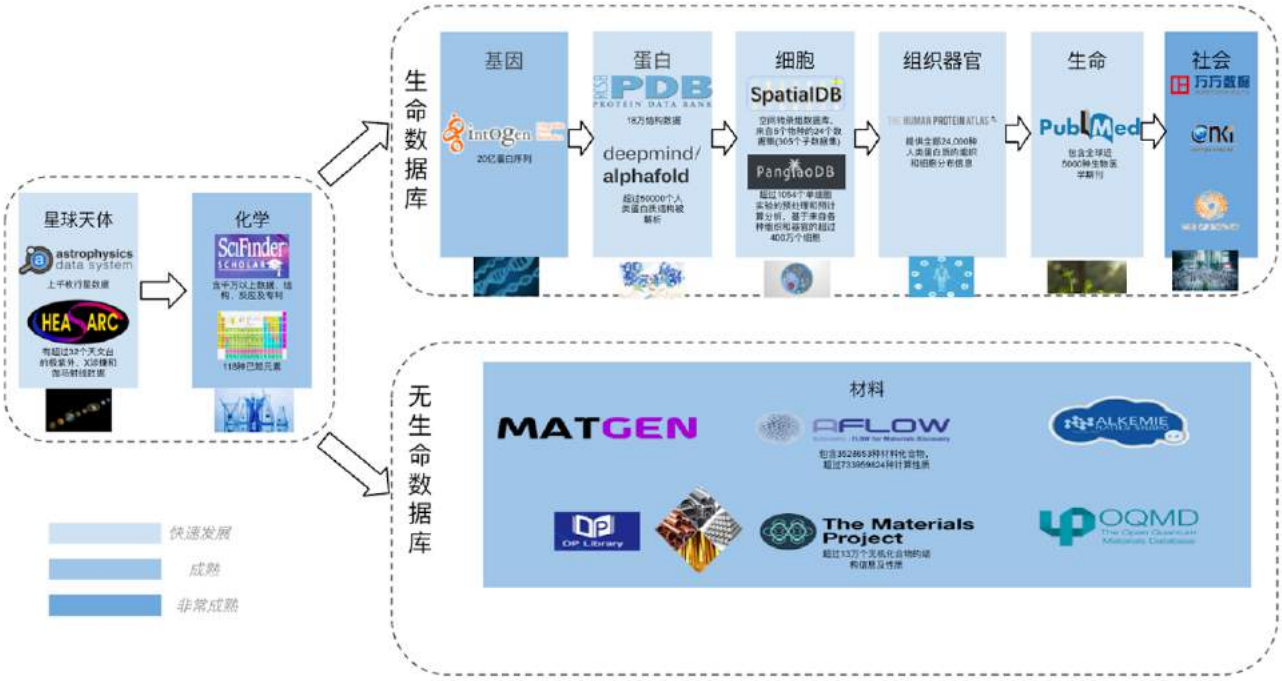
这些想法已经被成功地应用到许多问题，包括分子动力学和稀薄气体动力学。以分子动力学为例，通过将机器学习与高性能计算相结合，我们能够以从头计算（ab initio）的精度来模拟数亿个原子的系统，这相比此前有五个数量级的提升。分子动力学应用非常广泛，比如如果我们想要预测一种新型纳米材料的性质，就需要计算其海量原子间的相互作用。传统分子动力学在计算势函数的时候依赖经验力场，导致结果不准确；第一性原理的方法通过量子力学模型计算，虽然可靠但是效率低，难以大规模使用。而基于机器学习的分子动力学方法，依靠量子力学模型提供训练数据，用深度神经网络对高维势函数进行拟合，就可以同时保证算法的准确性和高效性。这种将物理模型、机器学习和高性能计算深度结合的方法，为我们展示了非常巨大的想象空间。



深度势能训练过程示意图 [Source: Deep Modeling]

II. 数据驱动：AI 处理科学大数据

传统数据处理方法主要是针对小规模数据，以统计模型为基础寻找数据中的规律。然而基于小规模数据所建立的模型，其表达能力受限于数据规模，只能进行粗粒度的模拟与预测，在精度要求比较高的情况就不再适用。如果想要进一步提升模型精度就需要利用海量数据生成相关模型。



近来各个领域可获取数据种类和数量都有显著提升，为这个问题的解决提供了数据基础。然而随着数据量的提升，数据噪声逐渐增大，信噪比越来越低。而传统数据处理方式面对海量数据时会遭遇"维度灾难"问题，即无法有效在可控时间内利用海量数据建立高精度的模型。这就意味着我们需要全新的数据处理方法来应对维度灾难问题。这个方向目前最成功的例子是 AlphaFold2 [5]。蛋白折叠问题是一个典型的高维问题，AlphaFold2 通过 AI 的方式彻底改变了蛋白折叠的技术路线，有效的解决了这个问题。

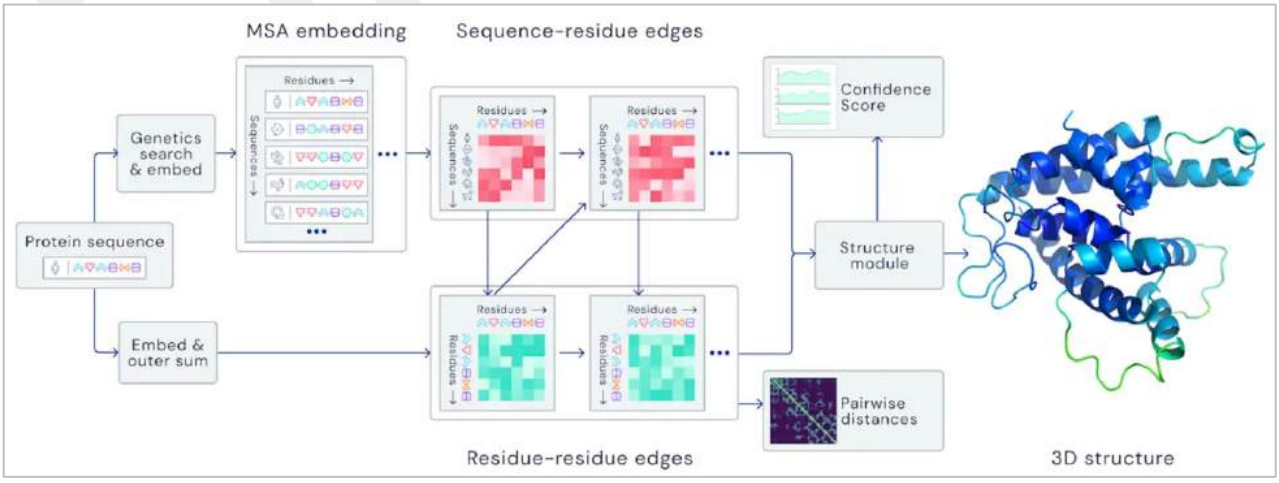


图. AlphaFold2 训练策略

III. 模型与数据的融合：AI for Science 的系统工程

AI for Science 的第三条实现途径是将模型驱动和数据驱动的方法深度融合。在科学领域，从"数据"中可以提炼出经验性"原理"，也可以使用"原理"来仿真模拟出"数据"。因此可以说，科学领域的"数据"和"原理"一定程度上是可以接近无损转化的，这一点是 AI for Science 相比于语言大模型(LLM)等其他领域的独特优势。

这个领域的主要挑战很多，比如"数据同化"、"观测和模型的同步学习"、"强化学习"、"理性实验设计"等。参考语音模型领域中 Langchain 的成功经验，AI for Science 在模型与数据融合的过程也更像是一个系统化的工程，不仅需要原理层面的创新，也需要从基础设施到产品到具体场景交互的全方面变革。每一个场景可能都需要一个庞大的团队来完成，当然这也意味着巨大的空间和机会。

Figure credit: DP Technology

1.3 大语言模型（LLM）：AI 与 Science 共生的桥梁

2023 年，随着 GPT 在各行各业的爆发，“是否能将 GPT 用于科研场景”成为了一个水到渠成的问题。当 ChatGPT 超越大部分人类在高考、SAT、美国法考、医考等领域取得令人咋舌的高分后，人们对于 GPT 驱动科研的兴趣愈发高涨。截止本报告发布时间（2023 年 8 月），LLM 在科研领域的实践已出具端倪，应用生态愈发丰富。

在人类的历史长河中，科学以其神秘的魅力吸引着我们，不断引领着我们探索未知的世界，解答生命的奥秘，推动社会的进步与文明的发展。而如今，当人工智能技术正在改变人们日常生活的同时，它也开始影响人类对科学边缘的探索。我们正处在这样一个交汇点 —— 人工智能与科学探索开始相互碰撞并融合。在这个交汇点的中心，大语言模型（Large Language Models，LLMs）正作为全新的媒介，重塑我们与知识本身的互动方式。在本篇中，我们将系统地探讨 LLMs 在科学研究中的变革潜力和挑战。首先，我们将阐述 LLMs 作为人类与知识双向接口的观点；然后，我们将深入探讨 LLM 处理科学内容的能力边界和改善方法；最后，我们将开放式的讨论当下 LLM 的最强潜力 —— 思维链（Chain of Thoughts），以及其革新人类科研方法的可能性。

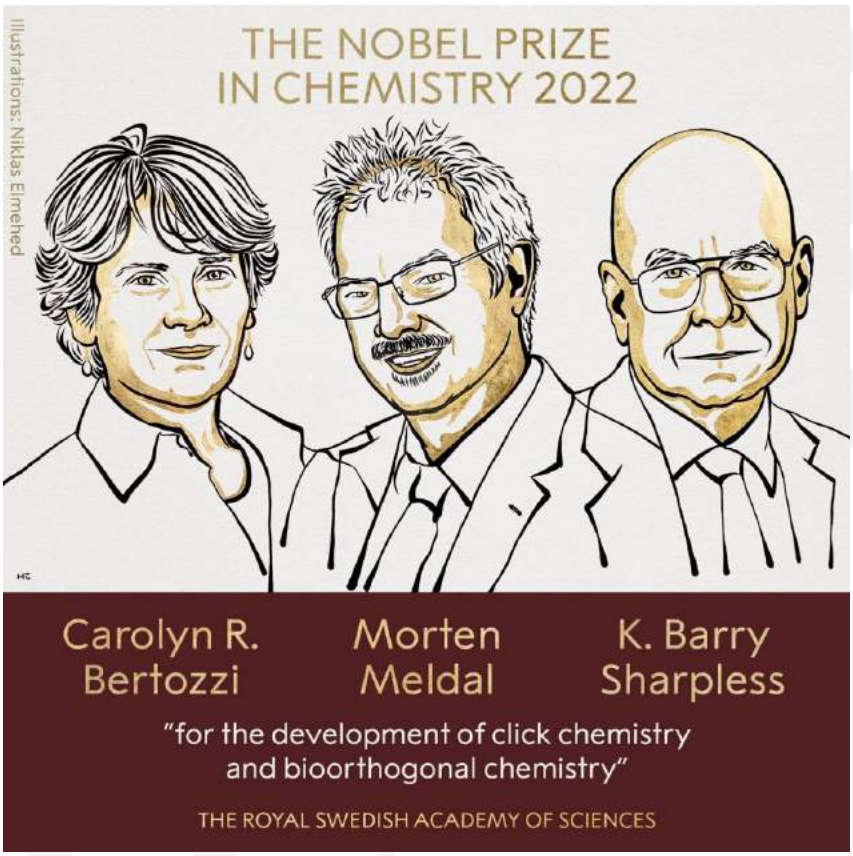
I. AI 作为人与知识交互接口的可能性

印刷术跻身人类的四大发明，实至名归。每一个人的生命有限，然而，将每个人在有限的生命中创造的知识蒸馏沉淀并以文字的形式记录在文本上代代相传、迭代更新，这一范式可以说是人类文明存续的基础。从原始人刻蚀甲骨、到今天人们在指尖就能调用复杂的数字数据库，人类永远在追求更高效的知识获取、存储和传播方法。正是由于人类以语言作为知识的“溶剂”，像 GPT-4 这样的大型语言模型（LLMs）才有潜力彻底改变我们与科学知识的互动方式。这场革命的核心在于将 LLM 视为双向接口的概念，LLM 是一种动态媒介，人们可以通过它提取和贡献知识。这种双重功能重新定义了人类与无边无际的科学文献之间的关系。这在学科不断交叉，知识不断交汇的今天尤为重要。

科学的进步和发展无疑在推动我们的社会向前发展，但在这个过程中，一个主要矛盾越来越显著：即愈发浩繁的知识量与有限人脑容量之间的矛盾。

科学的各个领域正以前所未有的速度相互交织和交叉。如同浩瀚的宇宙中的星系碰撞和融合，这些交叉和碰撞产生了全新的领域和深远的影响。最为明显的例子莫过于 2022 年诺贝尔奖的“生物正交反应”，它结合了化学反应的设计和优化与生物体系的复杂性和特殊性，发出具有选择性、效率和生物相容性的反应体系，以满足生物研究、医学诊断和治疗等领域的需求。又如，生物信息学，这一新兴领域就是计算机科学与生物学的

结合，正是通过对生物学数据的计算和分析，使得我们能更深入地理解基因，细胞和生态系统。亦如改变微尺度科研的"从头算分子动力学"，正是电子结构研究与分子动力仿真的融合；再比如量子计算，这一领域将量子物理学和计算机科学完美结合，有可能彻底改变我们对计算的理解和应用。这种趋势的背后也带来了一个挑战：科研人员需要了解的知识面越来越广，而要获取、理解并应用这些跨领域的知识，无疑是一项艰巨的任务。正是在这样的背景下，大语言模型（LLM）的出现给我们带来了希望。



2022 年诺贝尔化学奖授予的 click chemistry 和 bioorthoganal chemistry 正是学科交叉的优秀成果

一方面，LLMs 使得知识的提取和综合变得高效、便捷。通过解密和呈现复杂的科学信息，它们大大降低了进入新领域的门槛。这确保了即使是新手，也能够导航和理解科学话语，即使他们可能会被专业文献的复杂性所吓倒。例如，ChatPDF 允许用户上传 PDF 文件（通常是一篇科学论文），然后向 AI 代理提问有关这篇论文的问题，而 AI 则可以基于其本身的知识理解用户 PDF 的内容并做出回答 —— 就像教授向学生的传道、授业、解惑一样。这样的能力对于科研人员来说无疑是一种解放。他们无需再浪费大量时间去处理无关的信息，而可以直接专注于自己的研究领域，或者探索新的研究领域。这使得科研人员能够更快地接触并进入新的领域，更好地实现跨领域的研究。

在未来，当进入科学的门槛被大幅降低后，不仅是刚进入科学殿堂的年轻人会大幅受益；已在专业领域深耕数十年的科学家也将得以轻松跨越越来越细分的学科壁垒，在他山找到可以攻玉之石。在世界领先的大学

中，常举行跨学科的非正式交流活动，以促使不同领域知识与智慧的碰撞。LLM 的出现，可以使得这样的机制在全球范围打破时空限制异步进行，推倒学科之墙，让知识再次自由流淌。

另一方面，LLMs 可以加速并改进知识贡献的过程。例如，利用 LLMs 进行多步推理和决策的能力，研究人员可以在科学文献的迷宫般的广度中找到最相关的论文。这不仅加快了文献回顾的过程，还通过确保对现有知识的全面和相关审查，提高了科学论文的质量。这个过程并非简单的信息检索，而是一个深度学习、多步骤的推理过程，正是这样的推理能力，才能帮助研究人员在繁杂的信息中寻找到最有价值的知识。此外，LLMs 在撰写科学论文中也可以起到关键的作用。通过协助语言生成，它们可以为学者，尤其是英语为第二语言的学者提供必不可少的帮助。LLMs 不仅能提供语言方面的帮助，还能帮助构建逻辑叙述并确保连贯性，这在科学话语中是至关重要的，因为复杂的观点和发现必须被精心组织和呈现。

然而，LLM 的意义远不止于此。在更大的范围内，LLM 让更多人有了贡献知识的机会。知识不再是高墙之内的象牙塔，不再是少数人的专利。任何人，无论他们的语言、背景或地位，都可以通过 LLM 参与到知识的探索和创新中来。这无疑是一种巨大的进步，是人类文明进入下一个阶段的必要工具。如果没有这样的工具，知识的发展可能只会被少数的传统机构所垄断，导致创新的停滞。

在科技迅猛发展的今天，旧有的知识获取和创新方式已经无法满足我们对进步的渴望。深度学习和人工智能的出现，以及 DeepMind 和 OpenAI 等机构的巨大成功，都向我们证明，新的科研方法能够更高效、更有创新性地推动知识的发展。为什么 AlphaFold、ChatGPT （以及其背后的 Transformer）这些震撼世界的成果没有诞生在大学？这一灵魂拷问也许值得大学管理者们思考。这些新的科研方法并非要取代大学，而是希望通过更开放、更公平的方式，让更多的人参与到知识的探索和创新中来。

AI 不会取代大学，但是拥抱 AI 的大学会超越抵触 AI 的大学。

人类对知识的渴求是无尽的。LLM 的出现不仅为我们解决了获取和处理跨领域知识的难题，也使得科学的交叉和碰撞变得更为自由和流畅。我们有理由期待，这种变革将为我们带来更多的科学突破和创新，这一点，确实令人充满期待和激动。

II. 如何评估和提高 AI 对科学知识的处理能力

我们也要考虑 LLMs 和人类科学知识性质之间的基本兼容性。我们对宇宙的理解是以语言编码的，无论是自然语言还是数学语言。科学的本质在于其叙事性，即假设的提出，方法论的解释，结果的分享，以及通过语言媒介得出的结论。科学知识的叙述性使得它成为 LLMs 处理的合适任务。

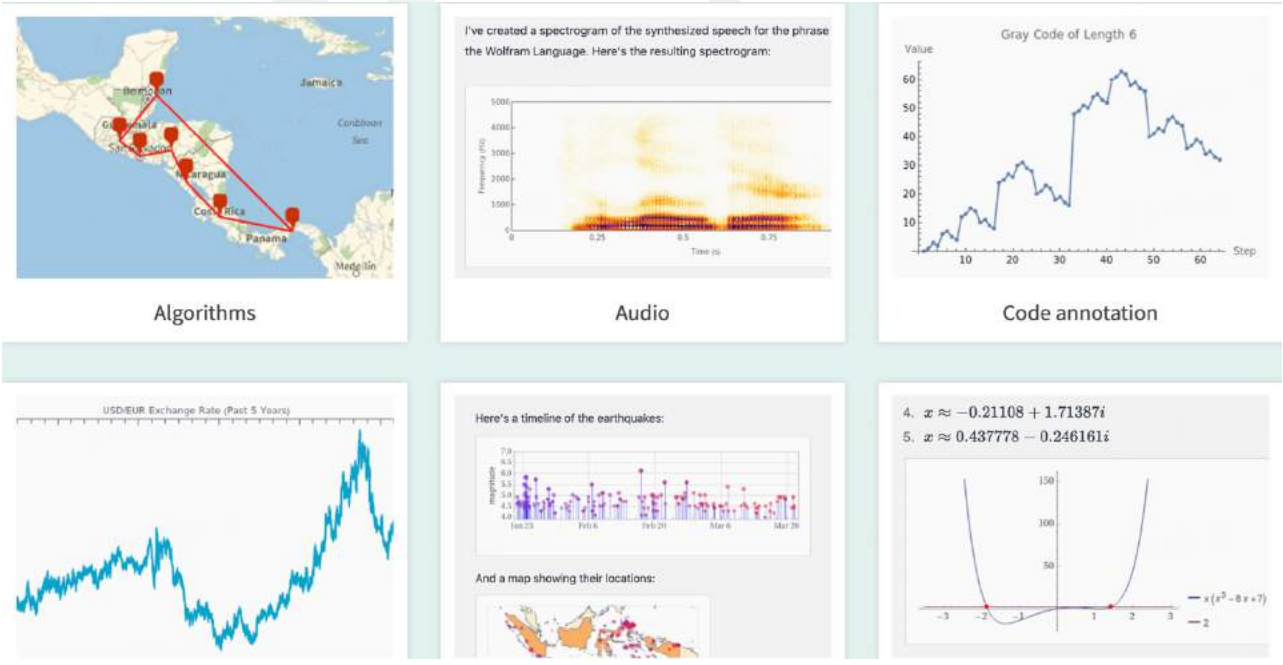
联想像 David Hilbert、Kurt Gödel 和 Alan Turing 这样的杰出思想家的作品，人们可能会找到他们的哲学和理论探索与 LLMs 的功能之间的相似之处。Hilbert 对完整和一致系统的追求与我们对 LLMs 的期望相呼应——创建一个能全面准确代表我们科学知识的模型。本质上，这些形式证明作为结构化的、数学的叙述，与自然语

言并无二致。从这个角度看，经过大量自然语言数据训练的 LLMs 理论上应具备处理科学概念的某种能力。然而，正如 Godel 的不完备性定理所示，没有系统可以既一致又完整。即使我们有了最好的工具，也会有一些问题超出我们的范围。图灵对不可判定问题的工作进一步加强了这一点，同时也提出了一个挑衅的反对观点 - LLMs 可能帮助我们确定知识系统中的这种边界。

可以说，阻碍人类科学文明进步的另一个重要矛盾，就是人类语言(包括自然语言和数学语言)的局限性和自然本身无限奥秘间的矛盾。而 LLM 的出现，有可能帮助我们更高效的评估这一矛盾的边界，并探索潜在的解决方案。

大型语言模型（LLMs）用于科学研究的领域，也面临着一些挑战。评估和提高 LLMs 在处理科学问题上的熟练程度，已成为科研界一个重要的议题。我们需要发展一套标准，以此来评估 LLMs 理解和处理科学问题的能力。我们希望通过这个标准，能够详细了解 LLMs 对科学概念的理解深度、对科学问题的分析能力、以及对科学方法的应用能力。这需要我们从理论和实践两个方面，深入探索和理解 LLMs 的工作机制。在实践方面，我们需要发展出能够衡量 LLMs 生成能力的标准。我们需要找出让 LLMs 从已有知识中生成新的推断和联想的方法，同时保证其生成内容的质量和准确性。这是一项艰巨的任务，但也是一项充满可能性的挑战。

在这个过程中，我们将会遇到各种挑战，但同时也会有许多令人兴奋的机遇。例如，将 WolframAlpha 集成到 ChatGPT 中就是一个有趣的步骤。通过融合 Mathematica 的海量知识，WolframAlpha 无疑可以极大地扩展 ChatGPT 处理复杂科学问题的能力。

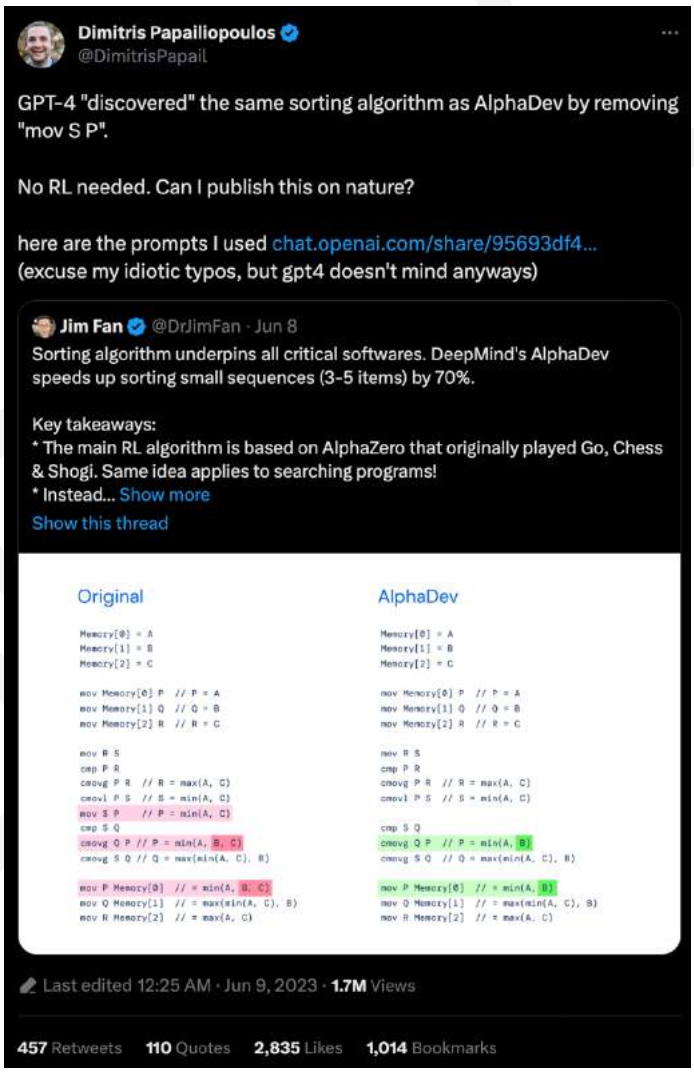


2023 年 5 月，ChatGPT 接入了 WolframAlpha 插件

III. 科学哲学引导我们对 AI 的理解和使用

在 2023 年的夏天，人工智能科学先驱 DeepMind 在 Nature 杂志上发表了一篇文章，论证了其 AI 代理 AlphaDev 利用强化学习发现了更高效的排序算法——超越了科学家和工程师几十年的努力。排序每天被数十亿人使用，即使他们并未意识到。它支持着从在线搜索结果和社交帖子的排名，到计算机和手机上数据的处理等所有事物，因此这个层面的改进具有系统性的重要性。

令人震惊的是，仅仅几天后，威斯康星大学麦迪逊分校的副教授 Dimitris Papailiopoulos 在 Twitter 上宣布，他成功地引导了 GPT-4 发现了 Alphadev 所做的同样的突破。这在社交媒体平台上引起了一场狂潮，最终引起了其亿万富翁实控人埃隆·马斯克的关注。两个不同的 AI 都发现这种新的"科学"，令人兴奋。因为这可能是自启蒙时代以来，人类首次找到了一条"批量生产"新科学的范式 -- AI for Science。



这两次 AI 实现的突破表明，像 GPT-4 这样的大型语言模型（LLMs）和科学方法以及科学哲学之间有深刻的相关性。上述案例中，Papailiopoulos 所使用的 LLM 具有模仿"思维链"或进行多步推理和决策的能力，这对应了科学家在科学研究过程中的关键思维方式。这些模型可以进行持续的互动，超越简单的响应，并理解和生成上下文相关的反馈。这种能力模拟了科学工作中的研究过程，为科学研究提供了新的可能性。例如，Auto-GPT 可以解析复杂的查询，记住过去的交互，理解上下文，并生成适当的响应。这种能力使它能够模拟科学工作中的调查过程，从而开辟了新的可能性。

We tend to define ourselves by what we DO. And this is crazy if you think about it. What if AI is able to do FOR us. How do we then define ourselves? By what we CREATE!

~ Toran, creator of Auto-GPT

科学方法与科学哲学指导着人类所有的科学活动。其基础根植于经验主义和迭代学习，包括生成假设，设计和进行实验，分析数据。这一哲学领域的集大成者，莫过于卡尔波普。波普的关键原理之一是"可证伪性"，他主张科学理论必须能够被实验测试并有可能被证明是错误的。这就需要生成假设，然后通过实证检验来证明或证伪。与此类似，LLMs 通过它们的"思维链"能力，可以模仿这种生成假设并对其进行逻辑推理的过程。

也就是说，LLM 的 "思维链" 使得它并不局限于现有的知识，而是能深度参与新知识的诞生。

例如，一个 LLM 可以生成一个新的科学假设，然后通过分析现有的科学文献和数据来检验这个假设的有效性。虽然 LLM 本身无法进行实证检验，但它可以辅助人类科学家进行此类工作。想象一下，一个 LLM 通过汲取大量的科学文献来生成假设或设计实验，提出新的组合或视角。或者考虑 LLM 作为一个不知疲倦的研究者，筛查大量的数据，寻找可能被人类忽视的规律。

波普的科学哲学强调了批判性思维和逻辑推理的重要性，这与 LLMs 所表现出来的"思维链"有着深刻的共鸣。尽管 LLMs 并非真正的"思考"实体，但它们的能力在某种程度上模拟了人类的思维过程。在这个意义上，LLMs 可以被视为是实现波普科学方法的有力工具。这些影响是重大的。科学家的角色可能会发生变化，因为曾经只有人类才能完成的任务现在可能与先进的 AI 工具共享。这引发了关于科学调查的性质以及人工智能与人类伙伴关系的深刻问题。在这种新范式下，科学方法将如何适应或扩展？我们应该如何确保科学的严谨性和完整性，因为我们越来越依赖这些工具？最后，人类和机器的探索合作将如何催生下一波科学发现？

大语言模型 LLM 为科学研究提供了新的可能性，AI 和人类科学家将成为密切的合作伙伴，共同推动科学的进步。在这个共生的过程中，我们不仅需要积极利用 AI 的能力，同时也要勇于面对和解决由此带来的各种挑战。

首先，AI 模型由现有数据集训练而来，而这些数据集可能包含偏见，因此在科研中使用 AI 可能会无意间延续系统性问题。其次，大语言模型可能会对复杂的科学概念产生误解，产生看似合理但实际上不准确或误导的

答案。此外，AI 在科学中的作用也带来了一些伦理问题，比如在 AI 的建议基础上得出错误的科学结论，不仅会导致资源的浪费，而且也可能造成严重的影响。如何界定相关行为结果的主体责任和义务是亟待澄清解决的挑战。这就要求我们在充分利用 AI 的同时，积极研究如何识别和减轻这些偏见，管理 AI 的误解风险，并探讨相关的伦理问题。此外，随着 AI 在科研中的应用日益广泛，我们更需要透明和可解释的 AI 决策过程。这些努力都是为了在科学和 AI 的融合进程中，保证我们的步伐坚实而稳健，使我们能够在未来的探索旅程中取得更大的成功。

在这个旅程中，我们的目标不是让 AI 成为科学的主导者，而是让 AI 成为科学的助手，为我们提供更深入、更全面的科学理解。这是令人兴奋的时代。AI 的出现引领我们进入了一个前所未有的人机科学合作的时代。当我们试图解决这些深刻的哲学问题和影响时，我们不仅仅是观察者，更是定义科学探索未来的积极参与者。

小结：可预见的未来，LLM 无法取代自然科学大模型

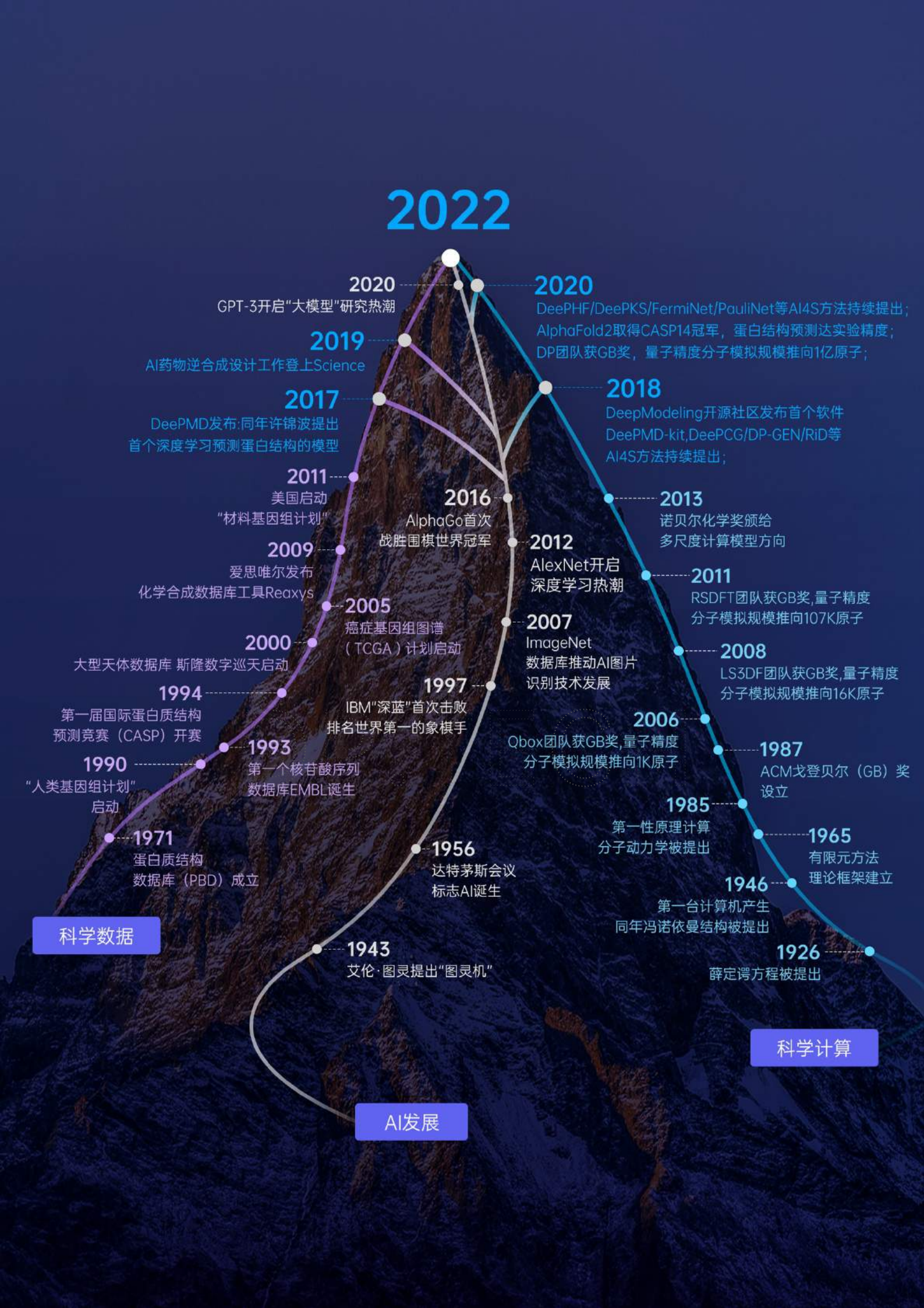
尽管近年来语言模型 LLM 在自然语言处理方面取得了巨大进步,但它难以取代专门为科学建模而设计的 AI 系统。究其根本，LLM 面向的是一维的字符串数据结构，而科学领域的数据类型纷繁多样，即有一维的基因序列，也有二维的分子图、三维的分子坐标、N 维的波函数。因此，在具体的科学领域中，使用专门的模型架构很可能比使用基于 LLM 的迁移模型更为直接有效。在过去的十年中，科学领域的大部分进步都源自于针对特定问题的模型。AlphaFold、DeePMD、PINN 等 AI for Science 模型成功的关键在于融入了生物学、物理学等领域知识，而非单纯文本预测。

虽然 LLM 拥有表征知识的潜力,但实际应用中其准确度和适用范围仍有限。相比之下，依托科学理论设计的独立 AI 系统，能更好地针对实际需求进行建模、仿真和预测。 基于产业实际应用的角度来看，开发自然科学领域的专业 AI 仍然是未来发展的主流方向。我们需要深入理解客观世界的运行规律，用 AI 来辅助人类进行创造，而非简单化约为语言学问题。

把 LLM 视为可以解决一切问题的万能工具，有些类似心理主义者(psychologism)的认知偏差。心理主义者试图用心理学的原理来解释各学科的基础 —— 即认为逻辑、数学等也可以化约为心理过程。这种观点忽视了各学科有其独立的体系和规律。类似地，过度期待 LLM 可以解决所有问题，也有把各领域过度简化的风险。各学科都有其专业性，不仅需要文本理解，还需要对具体领域的理论体系和方法的掌握。

将 LLM 视为一个通用工具确实有利于知识的迁移和连接。但同时我们也要注意方法论的区分，理解不同领域的复杂性，结合专业知识对 LLM 给予合理引导，避免陷入一种语言主义或机器中心主义的思维定势。保持开放和反思的态度对科技发展都是有益的。

Source：本节 (1.3) 原文发于 www.intelli-science.com/p/the-symbiosis-of-ai-and-science-unraveling


2022
2020
GPT-3开启“大模型”研究热潮
2019
AI药物逆合成设计工作登上Science
2017
DeePMD发布:同年许锦波提出
首个深度学习预测蛋白结构的模型
2011
美国启动
“材料基因组计划”
2009
爱思唯尔发布
化学合成数据库工具Reaxys
2005
癌症基因组图谱
（TCGA）计划启动
2000
大型天体数据库 斯隆数字巡天启动
1994
第一届国际蛋白质结构
预测竞赛（CASP）开赛
1993
第一个核苷酸序列
数据库EMBL诞生
1990
“人类基因组计划”
启动
1971
蛋白质结构
数据库（PBD）成立
科学数据
2020
DeePHF/DeePKS/FermiNet/PauliNet等AI4S方法持续提出;
AlphaFold2取得CASP14冠军，蛋白结构预测达实验精度;
DP团队获GB奖，量子精度分子模拟规模推向1亿原子;
2018
DeepModeling开源社区发布首个软件
DeePMD-kit,DeePCG/DP-GEN/RiD等
AI4S方法持续提出;
2016
AlphaGo首次
战胜围棋世界冠军
2012
AlexNet开启
深度学习热潮
2007
ImageNet
数据库推动AI图片
识别技术发展
1997
IBM“深蓝”首次击败
排名世界第一的象棋手
1956
达特茅斯会议
标志AI诞生
1943
艾伦·图灵提出“图灵机”
AI发展
2013
诺贝尔化学奖颁给
多尺度计算模型方向
2011
RSDFT团队获GB奖,量子精度
分子模拟规模推向107K原子
2008
LS3DF团队获GB奖,量子精度
分子模拟规模推向16K原子
2006
Qbox团队获GB奖,量子精度
分子模拟规模推向1K原子
1987
ACM戈登贝尔（GB）奖
设立
1985
第一性原理计算
分子动力学被提出
1965
有限元方法
理论框架建立
1946
第一台计算机产生
同年冯诺依曼结构被提出
1926
薛定谔方程被提出
科学计算

1.4 AI4S 的相关要素

AI for Science 的发展即包含 AI 行业的要素，也包含科学领域的要素，更需要来自产业和公共管理侧的发展要素。这些要素相互交织影响，共同促成 AI for Science 发展的正反馈。本节，我们将分别考察其中各个要素自身是发展历程，以及其在 AI for Science 中扮演的角色，和相应的机会。

I. 机器学习算法 / 预训练模型

在过去 10 年，AI 取得了巨大的进展。在近两年，以大规模预训练模型为代表的 AI 技术发展速度急剧加快，诞生了 GPT、Diffusion 等一系列影响深远的模型序列。不完全统计，从 2010 年 ImageNet 开始，每年 AI 的标志性的技术发展事件包括（摘录）：

2010：ImageNet 以及相关 AI 大赛

2021：IBM Watson 在 Jeopardy 知识竞赛中获胜

2012：AlexNet 发布；Google Knowledge Graph 发布；Word2Vec 发布

2013：DeepMind 开发 DoN (AlphaGo 的技术前身)；Caffe 发布

2014：Google Brain 发表生成式对抗学习网络 GAN；Adam 优化器被引入；Tesla Autopilot 上线

2015：ResNet 赢得 ImageNet ；AlphaGo 发布；TensorFlow 发布

2016：Google 发表 Transformer 论文；WaveNet 声音合成技术出现；Google 阐述 AutoML

2017：AlphaGo Zero 自博弈训练性能超过初代 AlphaGo；Face ID 上线 iPhone X

2018：BERT 发表；AlphaFold 发表；深度势能发表

2019：GPT-2 发表；AlphaStar 打败人类星际争霸 II 玩家

2020：GPT-3 发表；AI 被用于抗击疫情的众多方面（药物研发；流行病学研究；轨迹追踪等）

2021：AlphaFold2 发表；Telsa FSD 发布

2022：StableDiffusion, Midjourney,Dall•E 上线，AI 出图走红；ChatGPT 上线，掀起全球 LLM 热

2023：GPT-4 上线；GPT-4 启动联网；AlphaTensor 发表；RFDiffusion 发表

AI 在各领域的发展为其在科学领域的发展打下了基础。比如，AI for Science 的核心成果之中的蛋白折叠预测 AlphaFold 和深度学习分子势能 Deep Potential 均是基于 2016 年发表的语言模型 —— Transformer 模型骨架；又比如，2022 年在图像生成领域大放异彩的 Diffusion 模型，正在被运用于生命科学领域药物分子对接的预测，如 MIT 学者所开发的 DiffDock。AI 的发展是网络式的，也是非线性的，下一个 AI for Science 的惊艳成果，也很可能源自于意想不到的领域。

II. 算力基础设施：异构计算 / 云计算 / 超算中心

科学计算的发展史伴随着计算硬件的发展。在同一个算法体系下，更强大的算力、更高效的算力利用，意味着精度更高、时空维度更大、体系更复杂的模型。

高性能计算 (High Performance Computing, HPC)泛指通过聚合计算能力来提供比传统计算机和服务器更强大的计算性能，HPC 或超级计算环境可以使多个节点（计算机）以集群（互联组）的形式协同作业，在短时间内执行海量计算，从容应对这些规模庞大而又极其复杂的负载挑战。

HPC 由多要素组成。其各自发展的轨迹简述如下：

计算 (compute)：CPU 在过去几十年内基本按摩尔定律发展，每个周期推出单核性能更高，能耗更低的芯片；

存储 (storage)：与计算类似，存储领域也有类似摩尔定律的规则。头部企业不断提高存储介质的密度，让 HPC 对数据的使用能力不断提高；

网络 (interconnect)：由于 HPC 通常涉及多节点的协同，其节点间的网络至关重要。尤其在 CFD，AI 等算法中，网络速度经常成为计算效率的瓶颈。以 MSFT Infiniband, AWS EFA, Nvidia NVLink 等为代表的技术将网络速度提高至 100Gbps (甚至 900Gbps) 以上，极大提高了 HPC 系统的整体效率。

算法和算力的强耦合，是 AI4S 的必要条件。算法对算力的高效利用与算力本身的理论性能同样重要。

分布式计算将计算任务拆解成无需高频的单节点任务并分发给 HPC 系统内互联的节点、或分散在世界各地的闲置资源，以实现更大规模的计算。斯坦福大学的 Folding@HOME，加州大学伯克利的 BOINC 开放网络计算系统等都是其中优秀的代表。再如，当前工业仿真领域的算法软件（如 CFD, CAE, EDA）在商业客户的使用中，通常都需要进行软-硬件适配调优，以达到最大产出。

云计算 (cloud computing) 泛指通过网络交付计算等服务。高性能计算系统建设成本昂贵，更新周期长，且大部分科研场景对计算的需求有波峰波谷，因此，全球科研人员已开始逐渐转向云上的高性能计算解决方案。根据 Rescale 2022 State of Computational Engineering Report，53%的被访机构(欧美日韩地区)已将科学计算任务迁移至云上。[15]

异构计算泛指非传统 CPU 的计算体系，其中包括 GPU，FPGA，SoC 等方案。以 Nvidia 为首的 GPU 厂商，为 AI 的发展做出了重大贡献。GPU 的计算范式和 AI 算法完美契合。FPGA 让硬件可编程化，进一步加深了算法和算力耦合的可能性。SoC 通过将部分功能直接集成在硬件上，实现具体场景中效率的极致化，让"通过硬件设计解决软件难题"的思路成为可能。

现如今，计算正在成为前沿科研工作者群体最主要的科研手段。但是，过去几年，计算辅助科研正在面临进一步发展的瓶颈。而 AI for Science 为此提供了新的动力。

我们先来简单重温下算力体系在性能和规模两个维度上的发展。从性能上来看，过去 50 多年来算力性能提升的节奏基本被"摩尔定律"概括。近几年随着芯片处理器尺寸逼近物理极限，这一经验定律正在第一次经历挑战。通用硬件架构正在走向专用化，软硬件协同设计可能是未来；从规模上来看，互联网出现前主要是超算与个人电脑的天下。互联网的发展推动了信息的连接，推动的分布式的算力体系的建设，并最终推动了云计算的兴起。

AI for Science 与云原生时代前的很长时间里面，计算辅助科研整体依托的算力主题词是"超算"、"摩尔定律"、"通用硬件架构"等。在这个基础上，上一代计算方案的发展已经逐渐放缓。这表现在主要的数值、统计方法和底层软件包已经基本成熟。更重要的是，各方面科学计算方案的边界——特别是不同尺度物理模型求解的精度、效率，以及统计模型对科学数据的处理能力——已经基本确定。一些相对成熟的解决方案已经逐渐固化为工业软件；而一些面临瓶颈的模型算法则很难取得突破性发展。尽管算力和数据规模仍在快速提升，但由于"维数灾难"，这些方案解决问题的程度并不能显著提升。

AI for Science 与云原生时代的到来意味着什么？意味着算法和硬件这两个维度上发展的规模和迭代效率正在发生质变。深度神经网络作为本轮 AI 发展的核心推动力，所带来的高维复杂函数的逼近能力和高维复杂数据的处理能力恰恰能系统攻克计算辅助科研的瓶颈——无论是以 DeePMD 为代表的系统升级物理模型求解能力的方案、以 AlphaFold 为代表的系统处理大规模科学数据的方案，还是在天气预报、药物设计场景中物理模型与数据融合的方案，对从前的方案都提供了数个量级的提升。

算法和底层硬件性能规模的快速发展都对科学计算软件体系带来了巨大的挑战。过去的科学计算软件架构显然没有考虑到过算法和硬件发展带来的新需求。而且，大量科学计算生产任务依赖的是已经发展了数十年的代码，不仅历史包袱重，也缺乏合适的人才体系做重构和更新。看起来整个产业还相当地庞大、复杂，但其进一步的发展已经几乎到了不破不立的地步。

在这个背景下，AI for Science 将成为计算辅助科研全面升级的核心推动力。这不光是因为 AI 结合物理模型、科学数据的算法能大大突破过去方案的瓶颈、全面升级计算辅助科研的能力，更是因为过去 AI 发展过程中在数据、模型、软件、硬件、社区等方面的平台化实践为 AI for Science 进入平台模式指明了道路。在未来，像电子结构、分子动力学、有限元等引擎类算法均应发展出像 TensorFlow 这样的体系，方便上层算法与底层硬件两侧的快速迭代；像蛋白结构等场景的大规模数据库一方面可能与实验表征体系联动，面向目标自动化、智能化地生成数据，另一方面也应发展出相应的算法、模型；对于行业落地来说，也应该形成一套 MLOps 这样的系统，来支撑从研发到生产过程中整个算法功能生命周期面临的系统工程需求，并逐渐形成面向不同场景的具体解决方案。这些体系组合起来，终将成为科研工作者的新一代基础设施平台，推动形成新的分工体系，也推动"创新-落地"链条的持续迭代。

III. 软硬件数据基础设施

过去十年中，人工智能(AI)的瞩目增长与数据基础设施的同样显著演变密切相关。从存储技术的革新，大规模采用云计算，到数据湖和数据仓库的出现，以及向量数据库的日益利用，每个组成部分都在 AI 繁荣中发挥了重要角色。然而，在这些快速进步的背后，新的挑战已经出现——数据引力的挑战。

1. 内存芯片和存储技术的革命: 存储技术的进步与 AI 的崛起同步。动态随机存取内存(DRAM)和 NAND 闪存等内存芯片的发展已经扩展到支持复杂 AI 模型的需求。如 3D NAND 技术和英特尔 Optane 内存等创新，推动了存储密度和速度的增加，使得对于 AI 操作至关重要的大量数据集的快速访问和处理成为可能。AI 在训练和推理的过程中都需要调用大量的数据，而调用的速度直接决定了 AI 的性能和使用体验，因此新的存储芯片也备受 AI 开发者的关注。

2. 云存储: 人工智能（AI）快速发展的当下，数据的存储和处理需求正在迅速增加。这时，云存储作为一种灵活、经济且可扩展的数据存储方式，其重要性越发凸显。云存储为大规模 AI 应用提供了基础设施支持。随着数据规模的扩大和复杂度的提升，传统的本地存储方式面临着显著的空间和处理能力的限制。而云存储不仅能提供近乎无限的存储空间，同时，依托于云计算的强大计算能力，能够进行快速、高效的数据处理和分析，为 AI 算法提供源源不断的数据输入。此外，云存储极大地提高了数据的可获取性和可用性。借助云存储，企业和个人可以随时随地访问和利用数据，这对于 AI 的实时学习和决策能力具有重要的推动作用。此外，云存储还可以实现数据的跨平台和跨设备同步，提升了数据使用的便捷性。

3. 数据湖和数据仓库: 在人工智能（AI）时代，数据不再仅仅是被动接收和存储的对象，而是一种可以被深度挖掘和利用的资源。数据湖和数据仓库在这个过程中起着至关重要的作用。首先，数据湖和数据仓库的出现，大大改变了我们对数据的处理和存储方式。数据湖允许企业在一个中心化的平台上存储所有形式和类型的数据，从而使得数据的获取、整合和分析变得更加方便快捷。数据仓库则以预先定义的结构存储经过清洗、转换、集成和对齐的数据，为后续的数据分析提供高质量的数据支持。其次，数据湖和数据仓库的设立，为 AI 应用提供了重要的基础设施。AI 算法的训练和优化需要大量、多元的数据，数据湖和数据仓库的出现，使得这些数据得以集中存储和管理，大大提升了数据的可用性和可获取性，同时也为保护数据安全和隐私提供了可能。最后，随着 AI 的发展，数据湖和数据仓库的重要性将会更加凸显。因为 AI 的发展将进一步推动数据的多元化和规模化，未来的数据湖和数据仓库需要具备更强的数据处理和分析能力，以适应日益复杂和庞大的数据环境。

4. 向量数据库: 在人工智能（AI）的快速发展过程中，我们正在见证数据处理和存储需求的质变。尤其是在预训练模型（Pretrained Models）被广泛应用到私有知识领域的背景下，向量数据库（Vector Database）的重要性越来越明显。首先，向量数据库能有效处理高维度向量数据。在预训练模型中，我们常常需要处理由文本、图像或者语音转换而来的高维度向量数据，这在传统关系型数据库中会遇到诸多困难。而向量数据库，尤其是基于近似最近邻搜索的向量数据库，能够以极高的精度和效率处理这种类型的数据。其次，向量数据

库为个性化的知识域提供支持。对于私有知识领域，预训练模型需要处理具有高度个性化和特异性的数据，这些数据的特征抽取和相似性匹配需要大量的向量运算。而向量数据库正是专门为此类向量运算设计的，它的存在让 AI 模型能够更高效地处理和理解私有知识领域。再者，向量数据库也有助于 AI 的持续学习和优化。通过向量数据库，我们可以方便地实现新的向量插入、删除和更新，这为 AI 模型的迭代优化提供了基础。而且，由于向量数据库的高效性，这些操作可以在极短的时间内完成，从而提高 AI 的学习效率和反应速度。

在 AI for Science 的产业实践中，很多企业对实验数据的密保要求非常高，如何高效合规的使用这些私有知识领域的数据是一个新的挑战 —— 那就是"数据引力"（Data Gravity）。

首先，数据引力的概念是指数据在其所在的环境中产生的一种"引力"，使得应用、服务和其他数据都倾向于向这些数据集中的地方聚集。这是由于数据迁移的代价巨大，包括时间、成本和数据安全等因素，导致数据在其生成的位置呈现出"稳定"状态，使得相关应用和服务必须围绕这些数据来进行设计和部署。

在私有知识领域的预训练模型中，数据引力的问题尤其突出。这是因为这些私有领域中的数据通常具有高度的敏感性和独特性，它们在源地生成和存储，难以轻易进行迁移。这就使得相关的 AI 模型必须在数据的源地进行训练和优化，这一限制可能影响到模型的通用性和扩展性。

为了应对数据引力的挑战，AI 领域需要进行一系列的技术和策略上的创新。例如，研发更高效的数据迁移技术，通过压缩、封装等手段，减少数据迁移的时间和成本；同时，也可以探索在数据源地建设边缘计算设施，使得数据处理和分析的任务能够就近进行，降低数据迁移的需求。此外，通过构建分布式的 AI 训练和推理框架，使得模型能够在数据分散的环境中进行学习和优化，也是应对数据引力挑战的一种可能途径。

数据基础设施的进步和伴随的挑战如数据引力，证明了 AI 旅程的复杂性。未来无疑将见证更复杂的数据基础设施创新，因为我们努力满足 AI 的不断提升的需求。在我们进入这个令人兴奋的未来，AI 和数据基础设施之间的相互依赖不断加深，也塑造了我们科技进步的轨迹。

IV. 科学计算与工业仿真软件

千百年来，科技发展翻天覆地，但在大部分时间里，理论和实验是科研工作者们唯二的科研手段。计算成为科研新手段，是从二十世纪中叶计算机的发明开始。在这里我们广义地将计算机算力、网络、存储的整合利用统称为计算。计算用于科研通常有两种方式，一种是演绎式计算，例如基于对理论公式的数值求解、进行仿真模拟，或者 AI 中的模型推理；另一种是归纳式计算，例如对实验数据和模拟数据的处理和分析，或者 AI 中的模型训练。

当下，计算正在成为前沿科研工作者群体最主要的科研手段：不光一系列以"计算"作为前缀的学科（计算物理/化学/生物学/天文学等）在走向成熟、有很多计算方案沉淀成为了工业软件；甚至对于主要依赖实验的群体，"计算"的渗透率也正在变的越来越高——表征数据分析、自动化，等等，无不依赖计算。

在这个数字化时代，科学计算和工业仿真模拟软件的指数级增长已经深刻地影响了人类的经济发展和科技进步，标志着我们正朝着尼采所述的"第二自然"演进。在航空早期，飞机的设计和测试大多依赖于反复的试验和调整。例如，莱特兄弟的首次成功飞行，就是在无数次失败、调整和重新测试之后实现的。然而一个世纪后，计算工具的出现彻底改变了这一过程。今天，在生产原型之前，飞机设计就会经过严格的计算模拟，预测其在各种条件下的性能。多物理模拟的应用使我们能全面分析材料性质、强度、热行为、撞击动力学和空气动力学。这种从物理试错到计算预测和优化的转变清楚地证明了科学计算和工业模拟软件的价值和重要性。

多尺度的算法基础：从纳米尺度到星球尺度

科学计算和模拟软件的应用范围跨越了广大的时空尺度，涵盖了从微观到宏观的各个层次。科学计算的力量和深度可以通过它们所模拟的各种物理现象的广泛性和多样性来深刻体现。

在微观尺度上，我们遇到的是量子力学现象。这些现象由薛定谔方程描述，并且在原子和分子之间的相互作用中起着至关重要的作用。VASP 等软件工具以薛定谔方程为基础，为我们模拟和理解原子和分子之间的相互作用提供了强大的工具。在分子尺度，我们面对的是牛顿运动定律，这是描述微观粒子群体行为的基础。LAMMPS 等分子动力学软件利用牛顿运动定律，可以模拟大量微观粒子的集体行为。这使得我们可以预测和优化各种材料和化学反应的行为，从而推动了材料科学、药物设计、催化剂开发等许多前沿领域的进步。

介观尺度上，模拟考虑了个体分子的相互作用,同时也考虑了物质的整体性质。这种方法在许多领域得到广泛应用,特别是在材料科学和化学工程方面。例如,相场理论是一种介观尺度的方法,用于计算多组分体系平衡状态下的相图和相平衡。它整合了类统计力学和量子力学,可以定量预测混合物的相分离和化学反应过程。相场理论可以模拟不同条件下材料的微结构演变,指导材料合成和加工工艺的优化。另一个例子是粗粒化技术,它将一些不相关的原子或分子聚集成"粗粒",以简化计算复杂体系的相互作用。这大大降低了计算量,使得可以在较长

的时间和长度尺度上仿真材料的结构和性能演化。粗粒化方法应用广泛,可以研究诸如相变动力学、晶体生长、裂解以及材料在外界载荷下的塑性变形等过程。

在更大的尺度上，固体力学原理和有限元方法成为主导。这些原理和方法是 ANSYS 和 LS-DYNA 等软件的基础，这些软件可以帮助我们预测材料在各种工况条件下的行为。无论是桥梁、大楼、飞机、汽车，还是微观的集成电路和微机电系统，我们都可以用这些工具进行设计和优化，以确保它们在实际环境中的性能和可靠性。

进一步向上，到达宏观层面，我们会遇到液体和气体流动的问题。这些问题由 Navier-Stokes 方程描述，并且是计算流体动力学（CFD）的基础。ANSYS、DS 等公司开发的 CFD 软件使得我们能够模拟和优化各种流体流动问题，从飞机和汽车的空气动力学，到能源系统的热流和流体流动，再到环境和气候系统的复杂动力学。

在这之上，我们关注的是地球系统的动力学，这包括大气、海洋、冰川、陆地生物地理等各个子系统的相互作用。借助 WRF 模型等工具，我们能够模拟和预测复杂的大气动力学，进行天气预报和气候建模。这不仅对我们的日常生活和经济活动至关重要，而且对我们理解和应对全球气候变化问题也具有深远的意义。

总的来说，科学计算和模拟软件的应用已经深深地渗透到了我们理解和操作世界的各个方面。从微观到宏观，从分子到星系，科学计算都在助力我们探索和改造自然，解决各种科技和社会问题。

工业设计的数字化转型

模拟软件的发展极大地改变了工业研发过程，尤其是引发了实验设计(Design of Experiment)和自动化数字设计工作流等概念。DoE 方法的实施，通过系统地调整各种输入参数来确定其对输出的影响，为设计者提供了一种系统而高效的方法来探索设计空间。此外，R&D 的数字化转型正在促进更有效的工作流程，使设计团队能够利用自动化，促进创新，并加速上市时间。

设计实验的概念可以追溯到 1920 年代的英国，当时罗纳德·费希尔首次在农业研究中引入了这种方法。然而，直到电子计算机的发明和软件工具的开发，DoE 才能广泛应用在各种工业领域。传统上，工业产品设计和优化主要依赖于经验和直觉。然而，由于产品的复杂性和工艺要求的增加，这种方法已经不能满足需求。通过 DoE 和自动化设计流程，工程师可以系统地改变输入参数，并通过模拟或实验来研究它们对产品性能的影响。这大大减少了需要测试的实验设计数量，加快了产品开发速度，并提高了产品质量和可靠性。

例如，在汽车工业中，通过 DoE，工程师可以同时考虑多个设计因素，如材料、形状、重量等，并找到最佳组合，以达到最优的燃油效率、安全性能和驾驶体验。此外，借助自动化设计流程，可以将这个过程完全数字化，从而进一步提高效率和精确度。

在半导体行业中，自动化设计流程的应用尤其重要。随着集成电路的尺寸不断缩小，芯片设计变得越来越复杂。在设计现代芯片时，需要考虑的因素数以千计，包括晶体管的尺寸、材料、布局等。通过使用 DoE 和自

动化设计流程，工程师可以在虚拟环境中进行大规模的实验和优化，从而在更短的时间内开发出更高性能、更可靠的芯片。

总的来说，设计实验和自动化设计流程的发展为我们提供了一种全新的方法来开发和优化复杂的产品。这些方法已经成为现代工程实践中不可或缺的部分，并将继续推动我们的创新能力和效率。

工业仿真软件的行业整合

在过去的几十年中，科学计算和工业模拟软件市场经历了大规模整合。像 ANSYS、达索系统（DS）和西门子等先驱公司通过提供能让研究人员和工程师分析和预测产品在真实应用中的表现的软件解决方案，确立了自己的市场地位。

ANSYS 已经显著扩大了其产品组合，收购了业内领先的解决方案，提供了涵盖结构力学、流体动力学、电磁等的多学科套件。LSTC，LS-DYNA 的开发者，这是一款全面的有限元软件，最近也被 ANSYS 收购，显示了持续的整合趋势。DS 也是如此，收购了 SolidWorks 和 SIMULIA 等公司，以增强其 3D 建模和模拟能力。这种整合趋势在电子设计自动化（EDA）领域也有所体现。Cadence 和 Synopsys 等公司已经构建了广泛的工具库，利用收购和开发提供从头到尾的半导体行业设计和模拟工具。

总结来说，科学计算和工业模拟软件的演进与工业研发的数字化转型以及在设计中寻求更高效率和创新的追求密切相关。从量子力学的纳米尺度到天气系统的全球尺度，这些工具以前所未有的方式促进了我们对物理世界的探索和操控，有效地推进人类工业发展的理性化。

在 AI for Science 的时代，无论是 DeepMind 的 AlphaFold，还是 DeePMD 等新兴工具，它们的出现和发展都离不开前人在科学计算和模拟软件方面的开创性工作。例如，AlphaFold 的成就，离不开 Gromacs 等在模拟生物分子动力学方面的深入研究；DeePMD 的成功，也离不开 CP2K 等软件在量子力学和分子动力学模拟方面的积累。值得一提的是，DeePMD 来自普林斯顿大学 Roberto Car 组，而 CP2K 的名字，正是来自于 Car 教授和他的合作者 Parrinello 教授，以及两者在 1985 年发表的里程碑成果 "Car-Parrinello 分子动力学方法"

这些人工智能工具在科学领域的革新，实际上是"站在巨人的肩膀上"的结果。而这些"巨人"，就是长年累月为我们提供精确模拟和深入理解自然现象的科学计算和模拟软件。今天，我们有幸见证并参与到这场由人工智能推动的科学研究革新。但我们也应记住，正是那些科学计算和模拟软件的开发者们，他们的智慧和努力，才为我们铺就了今天的道路。

V. 先进表征手段 / 科学数据集

在我们探讨机器学习的效果时，无法忽略的一点是高质量数据的重要性。在 AI 行业中，一直有着 "垃圾进、垃圾出 (garbage in, garbage out)" 的共识 —— 即，如果没有高质量的数据输入，则无论多么先进的算法、多么庞大的算力都无法带来高质量的成果。AI for Science 的数据，按照来源做划分，常见类型包括观测数据、实验数据、记录数据、调查数据、模拟数据等。每种类型的数据都有其独特的应用领域和获取方式，综合利用这些数据对于科学研究有着重要的意义。

不仅如此，近年来科学表征技术的快速发展也极大地推动了数据质量的提升。扫描透射电子显微镜（STEM）、原子力显微镜（AFM）、冷冻电子显微镜（Cryo-EM）以及 X 射线衍射（XDS）等表征工具在物理、化学、生物等领域中发挥着重要的作用，为我们提供了大量高精度、高分辨率的实验数据。例如，STEM 可以帮助我们直接观察到纳米材料的结构，Cryo-EM 则能提供高分辨率的生物大分子结构信息。这些高质量的实验数据大大促进了 AI for Science 领域的发展，让机器学习算法能更准确地预测和理解复杂现象。

在过去，由于数据采集、储存和处理方式的限制，人类只拥有小规模数据。然而近年来随着数据的重要性得到广泛认可，以及科学表征技术的发展，数据的数量和质量都有非常大的提升。在数量上，大量领域的数据集实现了几个数量级的增长，部分领域数据集甚至实现了从 KB 级别到 PB 级别的飞跃；在质量上，高质量、高精度、细粒度的数据取代了原来粗粒度的数据。

除了质与量的提升以外，数据获取成本的大幅降低也是近年来一个非常显著的特征。过去数据的获取需要高昂的代价，比如耗时耗力的实验、昂贵的模拟计算等。现在一方面专业的商用数据库凭借商业力量推广普及，另一方面开源数据库（比如 ImageNet, PDB, Material Project）依托开源力量建设蔚然成风。相关从业者可以用非常低的成本（甚至免费）获取海量、高质量的数据。

海量高质的数据意味着获取海量信息、发现未知规律的潜力；此外，获取成本的降低减小了相关研究与生产的成本。但在 AI4S 领域中，数据要素依然存在一系列的挑战。尽管整体上数据的质量和数量相比十年前有极大提升，但在某些领域，数据获取成本依然很高，现有的数据量和模型训练所需要的数据量仍然存在差距。同时，如何有效地整合和利用这些高质量的实验数据，如何克服各种表征工具之间的数据差异性等问题，也是需要我们进一步探索的。

总的来说，AI for Science 的发展需要数据、算法、计算资源等多方面的配合。在面对数据相关的挑战时，我们应不仅仅关注数据的数量和质量，还应关注数据的来源、数据的可用性以及如何更好地利用数据等问题。同时，表征技术的进步和各种新型的实验设备的出现，为我们提供了更多优质数据的来源，这无疑为 AI for Science 的发展提供了更多可能性。

VI. AI for Science 算法核心：实现物理约束的强弱形式

长期以来,科学计算和应用数学中的许多难点都是高维问题。这些高维性主要来源于多尺度效应。比如在研究量子多体系统时,随着加入的电子数增加,问题的维数呈指数增长;在多尺度建模中,微观尺度模型中的自由度数目往往很大。由于维数过大而导致的困难，也叫"维度灾难"

传统的平滑性正则化方法在高维情况下明显失败,无法有效描述函数的复杂性。近年来研究表明,用特定的神经网络结构逼近函数的误差是更好的复杂性度量。基于这一 观察，形成了一系列与特定机器学习模型相关的函数空间,如 RKHS、Barron 空间等。这为分析函数的高维逼近奠定了新的基础。另一方面,机器学习模型本身已展现出逼近高维函数的强大能力。这使得许多此前因维数灾难无法解决的控制论和 PDE 问题成为可能。

各类 AI 算法，包括强化学习、迁移学习等，通过不断的优化和改进，不仅拓宽了 AI4S 模型的应用范围，还降低了模型的训练成本。在计算机时代，我们常常通过把物理模型映射到计算机进行数学模拟来解决实际问题。然而，更微观的模型，虽然更接近第一性原理，因此更精确，但由于其复杂度更高，所涉及的自由度更大，面临的"维度灾难"也就更严重。而对比之下，更宏观的模型则更简单、效率更高，但精度低。在这种情况下，我们需要找到一种方式，能够在保持精度的同时，降低计算复杂度。

在统计物理学中，蒙特卡洛法是解决这种"维数灾难"的方法（方程见下），其在采样过程中收敛的效率与函数的方差和样本有关，但与输入的维数不相关，因此是一个维数不依赖的算法。

$$E(I - I_m)^2 = \mathrm{var}/m$$

注：使用基于网格的求积法则，如梯形规则(Trapezoidal rule)，I 是要计算的量，Im 是求积法得到的近似值，m 是使用评估函数的数量。

在 AI4S 时代，神经网络函数（见下述公式）与蒙特卡洛方法类似，对于我们所关心的函数空间，通过近似，参数的量的需求与函数更内蕴的方差相关的性质有关，与函数的维度并不相关。

$$f^*(x) \sim \frac{1}{m}\sum_{j=1}^{m} a(\omega_j)e^{i(\omega_j, x)},$$

注：ωj 是欧几里得空间维数上概率分布的独立同分布样本，等式右侧是具有一个隐藏层的神经网络函数的示例，Activation function σ and independent variable z are defined by σ(z) = eiz。

利用 AI 算法的这一特性，我们可以将其作为多尺度物理模型之间的桥梁，有效地逼近微观模型的解。这使得微观模型成为数据的生成器，然后在 AI 的学习之下，将相应的模型融合在更宏观的模型中。通过这种迭代性的交互，我们能够找到一个兼顾微观模型精度和宏观模型效率的解决方案。使用神经网络逼近动态规划中的

策略函数,可以解决百维甚至更高维的 stochastic 控制问题。解决高维 Hamilton-Jacobi-Bellman 方程也成为期待。

在具体的实践中，如何构建 AI - Science 之间的桥梁是核心的创新点。目前而言，对于不同的科学场景其思路也各不相同。在控制论问题中,需求找出一个策略函数,使某一目标函数(通常是状态轨迹的积分)达到最优。传统的动态规划方法需要解一个维数与状态空间维数相同的 Bellman 方程,因此面临维数灾难。针对这一问题,研究者提出将策略函数用神经网络表示,将控制问题的目标函数视为损失函数,控制问题中的动力学系统相当于构造深度残差网络。然后可以用随机梯度下降的方式训练网络参数。这种深度学习控制算法已经可以处理上百维甚至更高维的确定性和随机控制问题。Hamilton-Jacobi-Bellman (HJB)方程在控制论中也起着核心作用。近年来ㄖ有研究采用自适应深度学习算法解决高维 HJB 方程并取得进展。这为更好地处理连续状态空间下的复杂控制问题提供了可能。

在多尺度建模问题中,微观模型往往包含太多细节,无法直接用于实际问题的模拟。而传统的均匀空间离散化算法又难以应对如此高维的微观模型。机器学习方法可以在微观模拟的基础上识别出宏观变量或者集合变量,从而建立起不同尺度之间的联系,实现直接的多尺度耦合。最近的研究显示,这种方法可以使得原子级模拟扩展到包含上亿原子的系统。

对于从数据出发构建物理模型,机器学习也提供了新的思路。仅用数据拟合是不够的,需要结合物理约束条件,并使用具有代表性的数据进行训练,才能使模型可解释性强并保证外推性。这为许多难以从一般物理原理出发建模的问题带来了希望。只要满足物理约束,这类模型可以获得与一般物理模型可比的可靠性。

鉴于机器学习的 "黑箱" 属性，我们将如何才能实现物理约束?

- Level 0 无约束: 从"经典 AI" 的角度看，提高模型的物理水平主要来自提高其数据的质量：这是最弱的形式，仅仅将物理知识嵌入到训练数据本身中。例如,生成满足守恒定律或已知对称性的数据进行训练。这种方式端到端地利用数据进行建模,但无法保证模型在推断时会遵守物理定律。直接套用大语言模型 LLM 执行科学计算（比如要求 LLM 生成符合一定特征的分子）即属于此类 。
- Level 1 弱约束: 通过损失函数间接地加强物理定律。例如 (PINN)将微分方程作为损失函数的一部分,以此引导参数优化。PINNs 通过将微分方程等物理约束嵌入到神经网络结构中,以实现对包含物理定律的系统进行函数逼近。PINN 的输入包含自变量 x 和参数 λ,输出为方程的解 u(x,λ)。网络参数通过最小化损失函数进行训练。损失函数包含两部分: 1. 数据拟合项 - 使用网络输出 u(x,λ)逼近给定数据; 2. 物理约束项 - 对 u(x,λ)求导,带入微分方程中,要求满足微分方程;通过自动微分实现。训练完成后,损失函数降低意味着找到了既符合数据又满足物理约束的函数 u(x,λ)。变分方法也属于这一类,其中优化一个泛函(如经典力学中的作用量或统计力学中的自由能)。这种形式可部分约束模型,但不会硬性要求其遵循物理学原理。
- Level 2 强约束: 直接在模型架构中建立物理基本定律。例如,保留问题对称性的描述子就是一个典型的例子。另一个例子是构建哈密顿或拉格朗日神经网络,其结构尊重某些物理守恒定律,如能量守恒。这种形式

可最大程度地保证模型行为符合物理学原理,但也限制了模型的表达能力。在 DeePMD 中，研究者设计巧妙地加入了一些关键的物理约束:1. 平移不变性:物理系统的总能量不依赖于原子的绝对位置。这是通过将系统能量表示为原子能量之和来实现的,其中每个原子的能量由该原子的局域环境决定。局域环境是相对于原子定义的,因此确保了系统的平移不变性。2. 旋转不变性:当系统发生旋转时,总能量不应改变。DeePMD 通过使用旋转不变的局域环境描述子来实现旋转不变性。3. 置换不变性:如果调换原子的索引,总能量不应改变。DeePMD 同样通过使用原子能量之和来确保置换不变性。这些不变性是 DeePMD 设计的关键组成部分。该模型通过学习从量子力学计算中获得的势能面进行训练,这些不变性帮助确保学习到的模型可以很好地推广到新的系统。如果没有这些不变性,模型就难以预测不同于训练数据的系统的性质。

必须强调的是，实现物理约束并非越强越好，而是取决于具体场景中本章所讨论的各种要素的具体情况。比如 AlphaFold 可以算是介于 L0-L1 之间的算法，由于其所使用的 PDB 数据集质量极高，因此其训练效果出色，同时推理性能和效率优异。另外，在训练中引入 L2 强约束有时会降低模型的训练效率，对时间和成本产生影响。在构建具体场景的 AI for Science 算法时，需要通盘考虑各种因素，以实际效果为导向。

AI for Science 的进步不仅仅取决于 AI 算法的应用，还取决于大量经典算法的改进和提升。随着高质量和大量数据获取门槛的降低，如何将不同来源、不同尺度、不同频率甚至不同类型的数据更好的融合在一起成为了一个重要的挑战。例如，在天气预报的预测中，不同观测指标所对应的时空尺度差异很大，如何利用这些有差异的数据进行更精准高效的预测就是一个实际的算法挑战。这就要求我们改进经典的数据融合算法以适应新的需求。AI for Science 的巨大想象空间存在于如何更好地利用 AI 算法将科学计算和物理模型相连接，进而指导科学和产业创新。AI 的力量在于其具有解决复杂问题，从而推动科学研究和技术发展的巨大潜力。此时，科研的瓶颈就不仅是"如何解决问题"，也是"如何定义问题，如何选择工具"。

例如同样是在求解 PDE 时，如果主要难点是解决高维问题，则 PINN 会非常有效，但若所面临的挑战不在维数灾难上，而在几何复杂、多尺度等方面，则神经网络的优势不容易发挥出来，反倒因为非线性优化等方面的困难，面临着求解效率低、误差不可控、难以系统改进等挑战。 这些场景中选择随机特征方法（The Random Feature Method, RFM）就可避免网格生成、容易处理复杂几何区域。

对问题的深刻认识是解决问题的第一步。AI for Science 算法的原始创新，不仅来自日新月异的 AI 模型，更来自科学家们对具体科学挑战的剖析、拆解、分诊，如此方能最大化 AI 在科学领域的效能。

Source: 本节 1.4(VI)原文发于 www.intelli-science.com/p/the-weak-and-strong-form-of-ai-for

VII. 高通量实验 / 自动化实验室

在过去的十年里，科学研究领域正在悄然发生一场革命性的变革。由机器人技术、人工智能以及先进传感器技术推动的高通量实验（HTE）基础设施，正在逐步改变实验室工作的格局。这些进步，统称为"自动化实验室"，有望重新定义科研的未来。

1. 机器人实验室的崛起： 在过去的十年里，自动化实验室设备被设计出来，用以执行常规任务，例如液体处理、样本准备，甚至更复杂的步骤如色谱或细胞培养。像 Tecan 和 Beckman Coulter 这样的公司已经开发出了能够连续不断执行实验工作流的先进机器人系统，这不仅增加了吞吐量，还降低了人为错误的可能性。

2. 高通量筛选（HTS）的普及： HTS 是一种科学实验方法，特别适用于药物发现领域。它利用机器人技术和数据处理软件，快速进行数以百万计的生化、遗传或药理学测试。能够快速地对大量化合物进行生物靶标测试，为药物发现和基因组学等领域开辟了新的途径。

3. AI 和机器学习的融入： AI 和机器学习已经改变了高通量实验的数据分析阶段。通过处理高通量筛选生成的大量数据，AI 可以识别模式，得出洞见并进行预测，极大地加速了发现过程。像 Zymergen 和 BenevolentAI 这样的公司已经成功地使用这些方法来加速新材料和药物的发现。

4. 先进传感技术的发展： 在过去的十年里，传感器的精度和微型化技术已经突飞猛进。这些进步已经被集成到实验室基础设施中，实现了实时数据采集，并为实验结果提供了前所未有的详细级别。

5. 云基础实验室管理软件的崛起： 自动化实验室的管理产生了大量的数据。像 Labguru 和 Benchling 这样的云实验室管理平台实现了实时跟踪、数据管理和分析，使得大量数据得以管理和获取。

6. 远程实验室的到来： 像 Emerald Cloud Lab 和 Transcriptic（现为 Strateos）这样的公司引领了一种新的模式 - 远程实验室。这些平台为研究人员提供了通过互联网访问完全自动化的实验室设施，使高通量实验功能平民化。

自动化实验室的持续发展正在为科研新时代的到来做好准备。通过使研究人员在前所未有的规模上生成和分析数据，高通量实验基础设施正在为未来 AI 在加速研究和开发步伐中发挥核心角色的场景奠定基础。

VIII. 从"小作坊"到"开放式平台"：跨学科复合能力人才与大规模协作

当今科研面临一个重大转型,那就是从过去封闭的"小作坊"模式,向开放协作的"平台"模式转变。这场变革的主因之一,是人工智能技术在各个科学领域的深入应用。将 AI 与科学结合,既需要大规模的跨学科人才,也需要新的组织形式来支撑高效的协作。

当下，习惯了"小农作坊"和"分散式"科研模式的科研工作者正在面临全新的局面：寻求基本规律之路遭遇瓶颈，更广泛的主题是用更前沿的工具挖掘、利用已知基本规律，以及寻求及利用复杂体系的规律；知识爆炸推动了学科分工，但学科惯性，无论是其对研究对象和研究手段的限定、还是其激励机制的固化，正在成为创新的阻碍；软件正在吞噬世界，软件是知识的沉淀、是可执行的知识，是连接算力与算法、应用需求的桥梁。仿真模拟、数据分析、团队协作，无不需要软件；硬件的摩尔定律即将终结，从性能出发的专用化发展趋势和规模化带来的云原生实践，结合 AI 的发展带来的数据、模型、软件的协同发展，正在倒逼软件系统翻天覆地的更新换代。

生产力的发展终将推动生产关系的变革。以开拓创新为标签的科研工作者面临的主要矛盾是：先进计算技术发展带来的无限创新可能与日益限制创新效能的"小农作坊"式的科研模式之间的矛盾。这可能不适用于纯数学、理论物理等基础学科，但适用于大量越来越离不开计算的应用学科。例如：科研工作者需要软件，而非科研文章或理论公式，来实现计算技术发展带来的新可能。但是，写软件对大多数课题组来说，一不被鼓励、二也不擅长——有志于此的老师往往被上课写本子带学生占据大量时间；有志于此的学生也独木难成林，很难有好的持续性。软件难发展、难维护，也制约了人才体系的培育和计算红利的释放。相比于互联网等行业的实践，科研软件生态的形成所面临的，是一个"小农作坊"模式下的负反馈死循环。

AI for Science 需要大规模的人才协作。那么，有怎样的关键词，能够最大程度地凝结所有相关的力量，最大程度地、最快地推动"小农作坊"模式走向"平台模式"呢？我认为，应该回归到科研工作者面临的主要矛盾的讨论——创新效能。持续做解放科学家、提升创新效能的事，既有利于科学研究以及业界研发，也能推动"平台模式"的达成。

有人会说，冲破瓶颈是前提。对于科学家们和工业场景下的研发人员来说，使用新工具可能会面临一定的迁移瓶颈，但只要迁移带来的创新效能的提升足够大，新工具体系就一定会替代旧的。某种意义上，过去机器学习平台的发展，对这样一个体系成长起来的时间尺度以及不同层面上的具体实践都已经提供了极为丰富的素材。但是，AI for Science 所涉及的知识体系、算法体系、软硬件架构体系和用户体系都更为复杂。多多少少有人会同时是计算机体系架构、软件工程、AI 算法的专家；但再加上数理化和各个应用领域，相应的人才体系是需要被重构的。

这需要很大甚至更大范围的协同努力，在相关的实践中，开源社区，以及围绕开源社区的基础设施建设正不断完善。在过去的科研组织形态下，由于成本、管理效率、激励机制等挑战，很难聚合多领域的人才进行科研协作。近年来，开源社区形式的出现让问题出现了转机。以 WordPress 为例，WordPress 除了核心开发

者，还有设计师，咨询师等提供各类解决方案或服务。又如，MongoDB, Couchbase 等 NoSQL 项目极大拓展了关系型数据库的专业度和适用场景。开源的成果已经深入我们生产生活的方方面面。Linux, Apache, Android 等社区的成功，证明了开源协作的可行性和规模效应。我们认为，AI4S 由于其对多领域协作的要求，非常适合开源社区的形式。如 DeepModeling 等社区的初步成功也佐证了这一点。

DeepModeling 社区目标是打造一系列 AI for Science 的软件基础设施平台、共同定义科学计算的未来。在此，一系列 AI for Science 的基础设施都是在做提升创新效能的事情。以 DeePMD 为代表的算法软件致力于推动第一性原理精度的分子动力学效率数个量级的提升；以 ABACUS 为代表的基础软件致力于将电子结构软件框架做的像机器学习框架一样，能快速适应硬件、算法发展的需求和实际生产的需求；以 dflow 为代表的云原生科学计算工作流套件致力于科学计算流程实现的可维护性、可扩展性以及高效的自动化、规模化；等等。

此外，正如 CV/NLP 领域发生的那样，数据规模的持续提升催生了像 GPT-3 这样的大模型，基于大模型的小样本数据精调和越来越多生成型能力的释放正在带来一系列新的可能性。在 AI4S 的一些方向上，数据规模化程度也达到了量变向质变的交界点上，由此我们开始了像覆盖元素周期表的 DPA-1 这样的预训练模型的探索；同时，如同 NLP 领域催生 Hugging Face 的逻辑一样，逐渐地我们将不仅需要代码的协同开发与测试、使用，还需要模型、数据、工作流在更大范围上的协同开发与共享，这催生了科学智能广场 AIS-Square 的实践。这些方面都是最新的探索，我们相信它们能够推动这场科研变革更快地发生、更有效地惠及所有科研人员。

最后，在这场科研范式大变革中，习惯"小作坊"式的科研工作者们有哪些需要关注的地方？这是有些微妙的。前沿研发的创新很难是"集体运动"，很多具体的科研创新可能还是得由一个个相对敏捷高效的课题组来完成——恰如安卓/iOS 平台激发了智能手机应用创新的爆发。因此，"小作坊"本身还是应该长期存在的。我们希望倡导的是推动一个个课题组形成对 AI for Science 与云原生底层逻辑和具体内涵的自觉意识，成为平台化科研模式的共建者，并在共建的过程中及时享受到创新效能的提升带来的好处——即做出更好的科研工作。在这个过程中，一系列创新成果还是"小作坊"做的，但做的过程中，软件、模型、数据、工作流将以开源社区作为载体，不断沉淀——这是课题组除自身创新外值得被鼓励做的事情。

AI4S 作为一个新兴领域，对人才的需求尤其迫切。AI4S 希望将人工智能技术应用到科研实践中，客观上要求相关从业者既要懂人工智能相关技术，又要有足够的科学学科背景，还要具备将科学突破转化为产业创新的工程能力。如 AlphaFold2，要求开发者既要了解蛋白质结构相关的生物学知识，又要了解深度学习等 AI 技术，还要有能力将算法软件化并推广到各大药厂与科研机构。

AI for Science 作为大量学科交叉，产学研紧密协作的新兴业态，更依赖 "涌现"式的发展逻辑和要素积累。这值得借鉴我国改革开放的实践 —— 对大目标、大原则无比坚定的同时，最大程度地解放大家的思想。在具体问题上不争论、先尝试，见效就推广。

IX. 长期主义的产业政策与产业资本

在人工智能（AI）的世界中，成功往往是毅力、探索的愿望以及投入资源实现长期目标的结果。 NVIDIA 作为 AI 基础设施领域的先驱，就是一个恰当的例子。如今，其图形处理器（GPU）已成为全球机器学习和深度学习框架的重要组成部分。然而，NVIDIA 在 AI 领域的旅程始于一个完全不同的领域 - 游戏。通过不断推动图形处理的极限以满足日益增长的游戏行业需求，NVIDIA 无意中为如今的 GPU 加速 AI 奠定了基础。

同样，AI for Science 领域的先驱者 DeepMind 和 OpenAI，并未在一夜之间取得了令人印象深刻的成就。在 DeepMind 的 AlphaFold 彻底改变蛋白质折叠预测，以及在 GPT-3 展示自然语言处理的力量之前，这两个机构都在进行一些其他的项目，乍一看，这些项目可能与有意义的科学发现无关。例如，DeepMind 的 AlphaGo 和 OpenAI 的 Dota 2 游戏智能 OpenAI Five，并没有直接关注科学问题，但它们为即将到来的事情奠定了基础。

这些例子突显了 AI 发展和实施科学进步的一个重要方面 - 即政策制定者和资本配置者都需要长期视角。AI 在科学中取得有意义突破的路径常常是间接和不可预测的，这需要塑造其未来的人理解并接受这一长期旅程的本质。

从政策角度看，这就意味着需要培养一种鼓励探索和奖励耐心的环境。政策应鼓励在广泛的领域进行 AI 研究，而不仅仅是那些具有立即明显的实际应用的领域。它还应为将培养下一代 AI 科学家和工程师的教育和研究机构提供强有力的支持。同样重要的是，确保一个能够促进创新同时负责解决道德、隐私和安全问题的监管环境。

对于资本配置者来说，长期视角意味着投资 AI 企业，理解到显著回报可能会在多年后或者来自意想不到的地方。这可能涉及到支持从事基础研究、建设 AI 基础设施或在非传统领域应用 AI 的公司和项目。

总的来说，只有政策制定者和投资者的长期承诺，才能实现 AI 在推动科学突破方面的变革潜力。正如谚语所说，“罗马不是一天建成的”。我们开始见证的这场 AI for Science 驱动的科学革命也是如此。

1.5 AI4S 的发展阶段

我们相信任何一个领域的发展历程都是张连续谱，尤其是身处行业内的人士，可以灵敏地感知到行业任何一丝细微的突破。但当我们把时间周期拉得更长，或者以一个更宏观的视角总结和预测 AI4S 行业的发展，按照行业总体上需要解决的问题，可以把 AI4S 的历史和未来十年可预见的发展大致分为三个阶段：

- 以科学家为主导的概念导入期；
- 以科学家和工程师协作为标志的大规模基础设施建设期；
- 以工程师为主导的应用期。

三个阶段的递进，即是人类对 AI4S 的开发程度的不断加深和使用范围的不断扩大。在 AI4S 的时代，我们认为相关算法会经历从"简单模拟"到"智能化搜索"3 个阶段。

AI4S 1.0 阶段的关键词是"模仿"。如果以纯粹实验手段作为 0.0 阶段，那么计算模拟算法的 1.0 阶段就是"基于实验的思路，在实验基础上进行简单的外推和扩大"。例如药物虚拟高通量筛选就是通过某个标准（打分函数）让计算机自动的对实验结果进行选择。该阶段相比于纯粹实验阶段效率有所提升，但是这个阶段中计算模拟仅仅是辅助的手段，整体流程依然围绕实验展开。因此该阶段依然需要耗费大量的时间与成本进行实验探究。

AI4S 3.0 阶段的关键词是"搜索"。即，算法可以非常准确的对真实场景进行建模，并在此基础上根据特定需求设计并返回所需结果。例如在材料设计领域，3.0 阶段的算法可根据所需的材料特性（如密度、硬度等）自主地设计并返回材料的组分、结构，实现真正的"de novo design"。

在讨论 AI4S 时，人们常常会陷入上述二分式的极端，要么追求"完美预测"，要么停留在"简单模仿"，而忽略了这二者之间存在着的巨大地带，以及通过这片地带的潜在路径。事实上，在二者中间，还有一个关键的 2.0 阶段：即，算法虽然无法十分精准地复现现实世界，但人们对算法的能力有清晰的认知，算法有可预测的误差范围和置信区间。这样的算法实际上已经有了在科研实践与工业研发中的使用价值。在这个阶段，人们已经可以在算法预测结果基础上进行有针对性的实验验证，极大降低实验的时间与资源消耗。

我们现在所处的正是 AI4S 2.0 阶段，其关键词是"预测"，有边界地预测，有明确、可验证的置信区间。

AI4S 时代之前，绝大多数科学研究与生产实践都处于纯粹实验的 0.0 阶段和简单"模仿"的 1.0 阶段。随着 AI 技术的广泛应用，越来越多的算法正在向 2.0 阶段过渡。我们预测未来几年内，AI4S 的相关领域都会完成 2.0 阶段的算法升级，而之后将逐渐进入到智能化设计的 3.0 阶段，最终实现 AI4S 广泛普及。

I. 概念导入期（2016-2021）

AI4S 思想的首次出现，可追溯到 2016 年附近，就在人们为 AlphaGo 首次战胜人类冠军而或欢呼或惊叹之时，科学家们已经在尝试将机器学习等 AI 工具用于科学问题的求解之中。芝加哥大学丰田中心的许锦波教授尝试用深度神经网络预测蛋白质的三维结构，普林斯顿大学的鄂维南和 Roberto Car 团队也开始着手运用机器学习构建原子间相互作用的势函数。

2020 年前后，AI4S 正式走入人们视野。DeepMind 推出的 Alphafold2 在 CASP14 大赛中轰动世界，深度势能团队的 DeePMD 以"AI+物理模型+HPC"新范式获得 Gordon Bell Prize，而 FermiNet、DeePKS、DM21 等优秀开源工具也于同年前后相继诞生。这些有些工具与成果的出现，意味着 AI4S 已经得到真正意义的重视，科学家们在很多领域不用再"造轮子"，而是可以直接使用新方法、新工具进行科学研究。

因此我们认为 2016 年至 2021 年这个阶段是 AI4S 的"概念导入期"：AI4S 概念被初步证明，AI4S 思想的潜力被核心圈层认可，但更多处在萌芽阶段。这个阶段的主要目标是 AI 在各个科学场景应用的 POC（Proof of Concept），主要工作是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 0-1 突破。

在这个阶段，由于 AI 领域的工具和方法已经初步成熟，而科学研究的实际门槛更高，使得 AI4S 的参与者更多来自科学界，是以科学家为主导的阶段。由于算法的突破更像是散点状的格局，还没有产生系统性的工程化需求，因此软件工程等方面的力量还没有深度参与，来自下游工业界用户的声音和力量也相对较少。

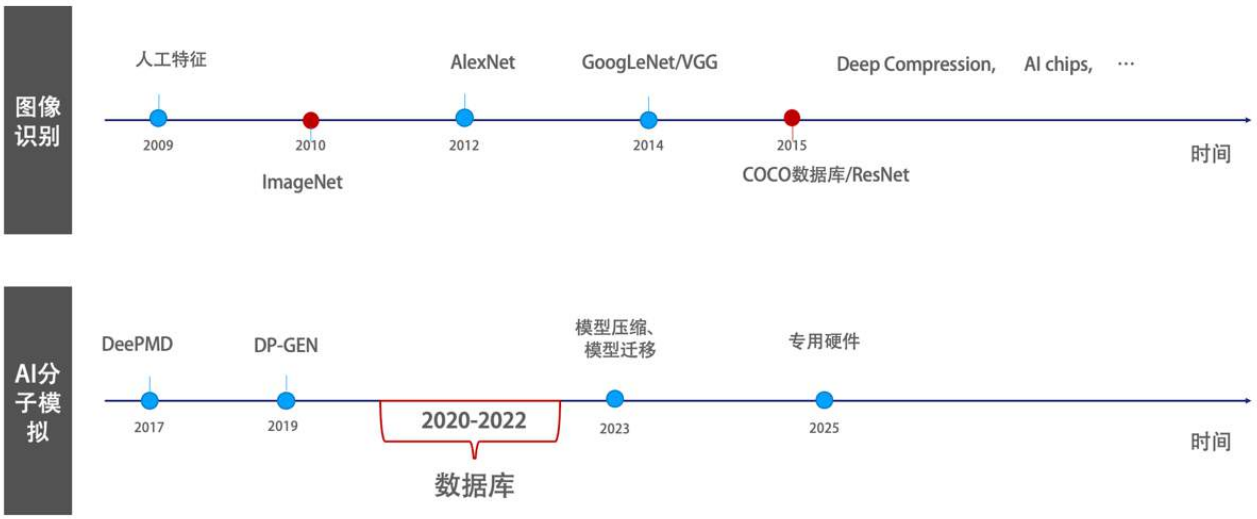


表 1. 2016-2021 AI4S 代表性成果（摘选）

	AlphaFold (DeepMind)	Modulus (PINN) (Nvidia)	DeePMD (AISI / 深势科技 / Princeton)
产业问题	蛋白质结构是理解生命机理，进行药物设计的基础	物理仿真（如流体力学），是工业设计，乃至游戏引擎的底层支撑性技术	微观粒子的仿真是理解物理世界运行原理的基础，也是药物、材料设计的基础
科学困难	蛋白质具有非常复杂的四级结构，其组成基础氨基酸序列存在 10 亿种以上组合，而组成蛋白质的过程又是非常复杂的变化过程； 使用冷冻电镜的方式获取蛋白质结构的实验成本较高	以流体力学为例：实验手段（如风洞）昂贵； 用计算的方式预测流体的运动轨迹的传统计算手段极度依赖大规模计算集群，时间长，且无法有效解决湍流等复杂问题	通过计算手段模拟微观粒子的相互作用依赖势函数的求解：使用传统经验力场计算结果不准确； 使用密度泛函理论计算虽然准确但是计算效率低，难以应用到大规模复杂体系中
AI4S 新范式	DeepMind 团队用特殊的网络结构设计，充分利用数据，使得蛋白质结构预测达到前所未有的精度。 该成果一经面世便登上《Nature》封面，引起全球轰动，是各界公认的"诺奖级"成就	Nvidia 开发了"基于物理的人工智能引擎"Modulus，底层架构包括 PINN，其同精度级别的计算速度比传统仿真快 1000-100,000 倍。基于其显著的速度优势，伯克利劳伦斯国家实验室与加州理工团队实现对复杂气象的实时仿真（0.25s 计算出 7 日预测数据）； 包括西门子能源在内的科研团队得以使用 Modulus 引擎对复杂系统进行实时仿真，实现高质量的工业研发数字孪生[7]	深度势能团队在前世界第一超算 Summit 上，保持第一性原理计算精度的前提下，成功对数亿原子的物理体系进行了分子动力学模拟，将超大系统的分子动力学模拟带入了全新时代。 本研究不仅将模拟尺度的记录提高了几个数量级，其速度也比 AIMD 提高了万倍。本工作获得了 2020 年计算界"诺贝尔奖" 戈登贝尔奖。

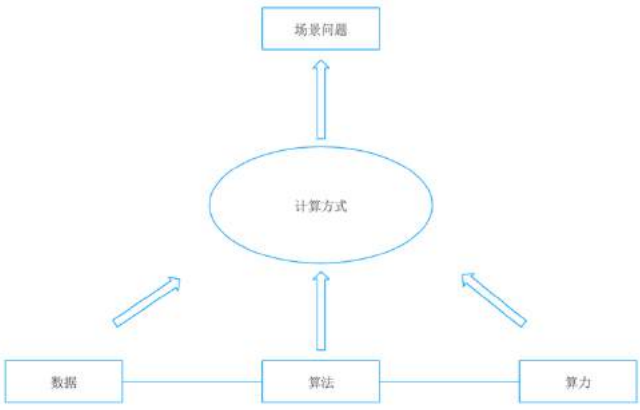
II. 大规模基础设施建设期（2021-2026）

有了 Alphafold2、DeePMD 等工具作为先行者，AI4S 的可行性已经无需质疑，其巨大潜力也逐渐为世人所公认，AI4S 从萌芽期，迅速转为基础设施建设期。如果说导入期的工作像是单点突破，基础设施建设期则是定义关键问题，在平台打造层面更强调"通用性"，主要任务是建设 AI4S 领域的"TensorFlow"和"GPT3"，即针对主流任务的训练器和通用模型。做个不严谨的参照，对比 AI 在其他应用场景的发展：

- 图像识别领域从 2011 年 ImageNet 数据库的诞生，标志着基础设施建设的开始，到 2015 年 ResNet 发布，基本解决了 CV 领域的通用能力建设问题；
- 在 NLP 领域，从 2013 年起 Mikolov 等人提出 word embedding 方法，并把神经网络系统应用于自然语言处理，到 2018 年通用的预训练模型 GPT 诞生，此后 NLP 开始大规模应用；

因此，我们预计 AI4S 的基础设施建设期同样会持续数年。

计算方式是数据、算法、算力的”三位一体“



基础设施的建设涉及到的不仅仅是"原创性算法"的单点突破，而是走完"被证明"到"被使用"这段关键距离。从这个意义上讲，AI4S 带来的计算革命，不仅是算法革命问题，更是面向下游场景问题，由 AI4S 算法驱动，数据、算法、算力"三位一体"的计算方式的变革。因此面向应用场景进行更良好的问题定义与抽象，确定哪些部分是基础性的，哪些部分是需要针对具体应用领域单独设计的，以及新的计算方式是个系统性工程，需要对快速变化的算力、数据、模型进行全新的统一架构设计。

面向这样的任务，这个阶段除了持续面向下游场景定义关键问题，还需要进一步定义全新计算方式的形态，这需要多领域人才的深度融合，算法开发、算力工程、性能工程、数据工程、软件工程等多领域人才将会发生密切互动。

我们有很大机会看到属于 AI4S 领域的 TensorFolw、GPT3 模型的出现，以及类似 Databricks 这样的数据、模型、算力一体化计算平台的出现，也会看到药物设计、材料设计、湍流模拟等此前被"维数灾难"封印的领域，诞生"AI 原生"的工业级软件。当这个阶段任务完成后，AI4S 将不再停留于少数先行者的范围。AI 将不再是互联网公司的专属武器，科学研究也不再是科学家们的专属工作。

正如基于固体力学、流体力学和电磁学的工业设计和仿真软件为汽车、飞机等工业门类带来的变革，经过 AI4S 的武装，人们对药物、材料、能源、化工等相关领域的许多关键问题的理解将更加理性化，"科学问题"逐步转变为"计算和工程问题"。

III. 成熟应用期（2026 年及以后）

在 2025 年基础设施建设的任务基本完成后，由于行业的主要问题已经被较好定义，通用层面的工具也趋于成熟，行业更多面临的将是这些基础设施"怎么用"的问题。由此进入大规模应用期，面向行业研发用户的需求，由工程师主导的阶段。

在这个阶段，有两个机会极其值得我们关注，一是新计算工具定义新研发流程，二是软件定义硬件。

对于实验学科而言，计算模拟手段的最终意义是替代一部分实验，将"盲目试错"的研发流程，改造为"大规模计算设计-定向化实验验证"的理性研发流程。参照汽车、飞机等工业门类的发展，时至今日，工业设计与仿真软件已实现对宏观物理场景的高度还原，工程师得以在计算机上对飞机设计进行充分验证之后再制造样品，进行实物实验。以汽车安全验证（crashing）为例，Ls-Dyna 等软件的大规模应用不只是极大的降低了制造样车进行碰撞的昂贵成本；更是将汽车安全性的验证前置到了设计环节，大幅规避了因安全不过关而将全部设计推倒重来的风险。

在技术主导的行业中，新范式的使用者有机会通过提高自身 idea-to-product 的开发效率，在市场竞争中占得先机。新的范式让部分企业抓住机会，促成市场格局的改变，并逐渐形成新的行业标准（best practice）。

但是，药物和材料等领域的研发模式仍旧停留在上个世纪，甚至近乎"炼金术"。爱迪生发明灯泡时为了选定合适的材料进行了数万次重复实验。如今的药物设计，也经常需要进行数十万次甚至百万次实验，才能得到一种潜在的候选化合物。AI4S 带来的一系列高效高精度的计算工具，有望将单个药物和材料研发所需的湿实验降低至现在的百分之一。

工业设计和仿真软件的发展目标，是驱动工业的数字孪生和智能制造。AI4S 算法将带来有了充足的模拟仿真能力，结合自动化机器人，将进一步改变药物、材料、化工、能源等行业研发和生产的载体。实验机器人和工业机器人的落地正在发生：比如在规则定义相对明确的实验和生产领域，机器人已经能够替代很多人类的工作。但在当下，药物研发、化工合成、材料研发等算法能力尚且不足，仍在持续迭代的领域，软硬件之间的碰撞和磨合势必会面临着诸多挑战。而 AI4S 基础设施的完善，势必将有更强的能力对硬件设施的发展产生更良好的定义能力。

IV. AI4S 的长期愿景是发现新的科学原理

过去 10 年中，AI 在工业和商业领域的应用取得了巨大成就，但其"黑箱"属性也一直被学界诟病——能知其然却不能知其所以然。由于 science 本身的客观存在性，将 science 与 AI 融合为 AI 提供了绝佳的"验证"步骤，让 AI 在特定领域内能出产"可解释"的成果。这不亚于为人类发现新的科学原理！

通过将"科学方法与实验设计（scientific method and experimental design）"的理念嵌入 AI，国际前沿学者正在开发 AI4S 发现新科学原理的可能性，其成果令人振奋。

2022 年 7 月，诺贝尔奖得主 Ronald Hoffman 受邀于北京科学智能研究院（AISI）分享其对 AI4S 的理解和展望，其中提到麻省理工 Jeffrey C. Grossman 团队通过机器学习揭示氢化纳米晶和非晶硅中 Staebler-Wronski 效应起源，并称赞该工作验证了 AI4S 发现新科学原理的可行性：

- 氢化非晶硅是薄膜光伏的主要材料，具有成本低、毒性小、光吸收大的优势。然而，它的光电转换效率受到光致衰减——Staebler-Wronski 效应的阻碍。为了解决这个问题，科学家们提出了很多意在消除或降低 Staebler-Wronski 效应影响的方法，其中通过对氢化非晶硅进行纳米晶化处理形成纳米晶硅被证明可以明显地抑制 Staebler-Wronski 效应。[14] 过去关于氢化非晶硅中 Staebler-Wronski 效应的研究已经有不少的积累，但是对于改善 Staebler-Wronski 效应的纳米晶硅技术仍然缺乏相应的研究。
- 麻省理工 Jeffrey C. Grossman 团队利用机器学习遗传规划算法，对纳米晶硅中的结构集合进行第一性原理计算，确定纳米晶硅中对因 Staebler-Wronski 效应而产生的空穴陷阱影响最深的结构特征，结果证明空穴在非晶区域比在晶体区域更容易被俘获，另外通过结果分析突破性地发现了 Si-H-Si 桥键是导致 Staebler-Wronski 效应的最可能因素，这也是过去的研究中没有发现到的部分。
- 本工作中，AI4S 方法被证明在材料结构探索和研究中发挥重要作用，帮助加深科学家们对复杂物理过程的定性理解，发现新的科学规律。相关工作于 2014 年发表在《Physical Review B》中 [13]。

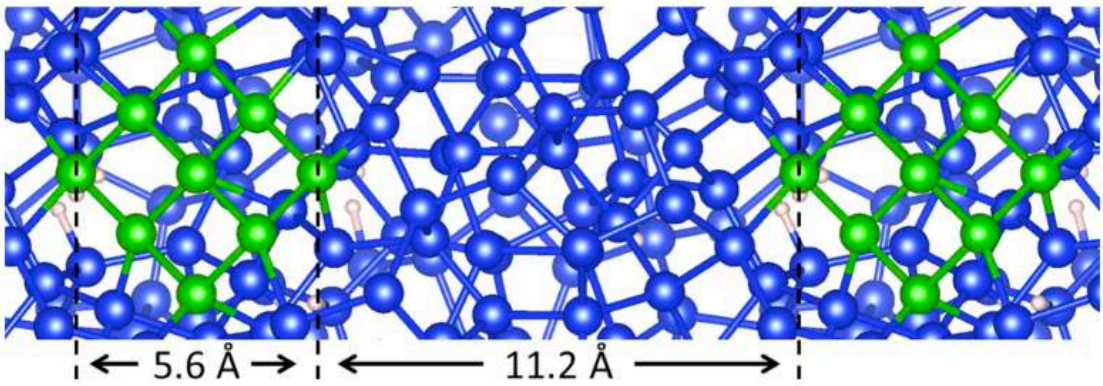


图. 纳米晶硅结构图

2023 年夏天，DeepMind 公司的 AlphaDev 人工智能"发现"了新核心算法登上《Nature》。AlphaDev 通过强化学习发现了增强的排序算法，超越了科学家和工程师几十年来磨练的算法。排序每天被数十亿人使用，却往往被人们忽视。它支撑着从在线搜索结果和社交帖子的排名，到计算机和手机上的数据处理的各个方面，因此这个层面的改进具有系统性的重要性。

令人惊讶的是，仅几天后，威斯康星大学麦迪逊分校的副教授 Dimitris Papailiopoulos 在 Twitter 上宣布，他成功利用 GPT-4 先进的 "chain of thoughts"能力，引导其复现了同样的科学发现过程，发现了与 AlphaDev 相同的突破性成果。这在社交媒体平台上掀起了一场热潮，最终引起首席推特官 Elon Musk 的点赞关注。两种不同的人工智能能够发现这种新的"科学"，使得这一发现变得更加令人兴奋，因为它可能意味着人类文明终于发现了 "批量生产新科学" 的潜在方式 —— AI for Science!

Source：

1. Weinan E. The dawning of a new era in applied mathematics[J]. Notices of the American Mathematical Society, 2021, 68(4): 565-571.
2. Richard Bellman, Dynamic programming, Princeton University Press,Princeton, N. J., 1957. MR0090477
3. Hohenberg P, Kohn W. Inhomogeneous electron gas[J]. Physical review, 1964, 136(3B): B864.
4. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
5. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.
6. Wang H, Zhang L, Han J, et al. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics[J]. Computer Physics Communications, 2018, 228: 178-184.
7. NVIDIA DEVELOPER 官网
8. DeepMind 官网
9. Car R, Parrinello M. Unified approach for molecular dynamics and density-functional theory[J]. Physical review letters, 1985, 55(22): 2471.
10. Zhang Y, Wang H, Chen W, et al. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models[J]. Computer Physics Communications, 2020, 253: 107206.
11. LAMMPS Molecular Dynamics Simulator. Sandia National Laboratories. [2010-10-03].
12. Georg, Kresse. VASP Group, Theoretical Physics Departments, Vienna. March 31, 2010 [February 21, 2011].
13. Mueller T, Johlin E, Grossman J C. Origins of hole traps in hydrogenated nanocrystalline and amorphous silicon revealed through machine learning[J]. Physical Review B, 2014, 89(11): 115202.
14. 李志. 氢化非晶硅薄膜的晶化处理研究[D]. 电子科技大学, 2010.
15. Rescale, https://rescale.com/resources/2022-state-of-computational-engineering-report/
16. Google, Why Opensource https://opensource.google/documentation/reference/why

1.6 2023 版《展望》核心观点：AI4S “四梁 N 柱” 的发展框架与新基建思路

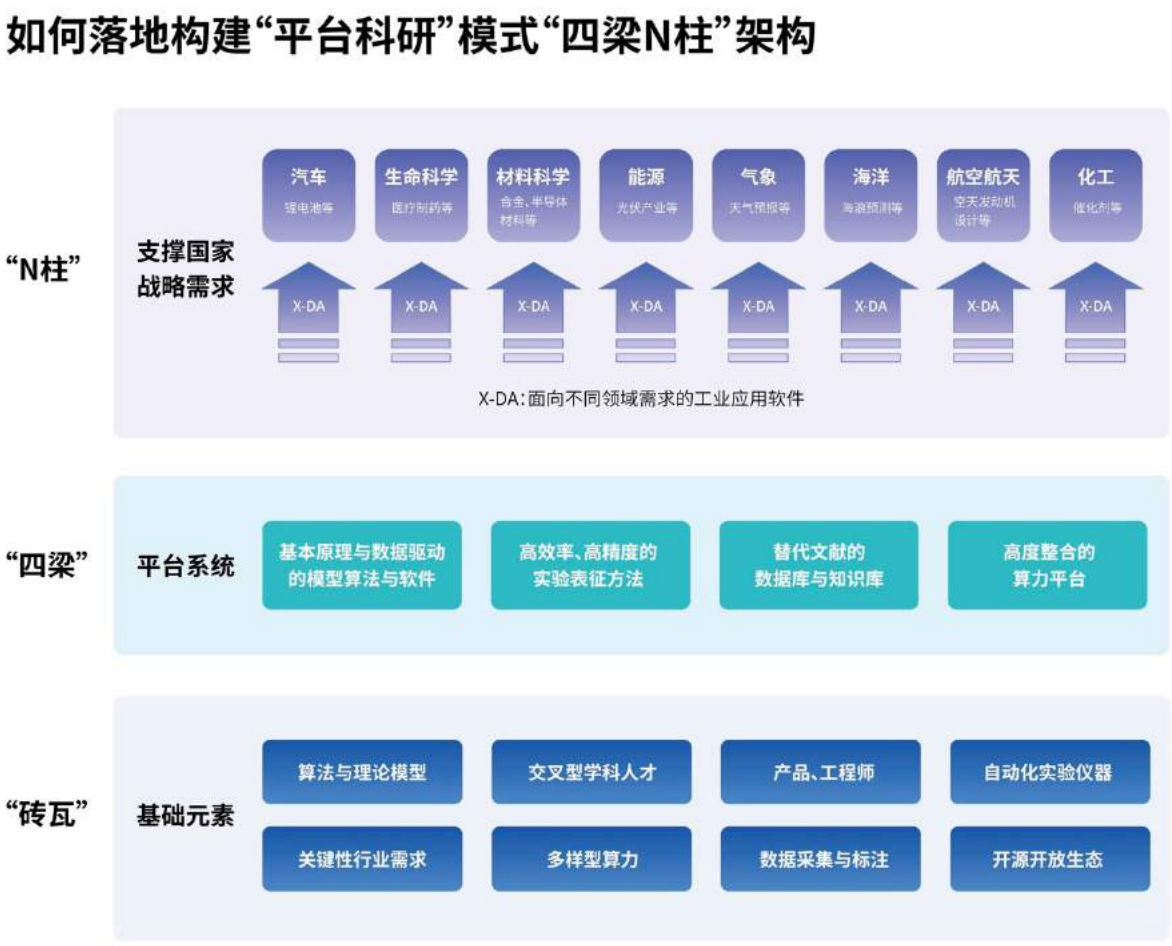
在本报告的前面章节，我们对 AI for Science 作为一种新的科学研究范式做了较为详细的介绍与分析；详述了 AI 和 Science 发展中的各要素；解析了当前备受关注的大语言模型在 AI for Science 浪潮中的机会；同时也给出了 AI for Science 正处于科学家与工程师协作完成大规模基础设施建设时期的判断。本节我们将与大家共同探讨属于 AI for Science 时代的科研基础设施的样子。

当我们在讨论科学研究的基础设施时，通常我们在指代大型的科学装置、实验仪器与设备、科学软件工具以及与之相关的数据系统、信息系统等。科学研究的基础设施为科学家和研究人员提供了必要的实验环境和条件，以便其更有效地进行研究、实验和创新，对于推动科学研究的原始创新，提升科学研究的创新效能具有重要意义。

科学研究范式与基础工程能力的不断发展是科研基础设施迭代演化的重要推动力。在古代，科研基础设施主要是观察和实验工具,如日晷、测量仪器等。这些设施简单且粗糙,但为人类提供了对自然现象的初步认识。随着文艺复兴时期对科学的重视，科研基础设施开始向实验室方向发展。这时期的实验室主要用于研究化学、生物学等领域,设施相对简陋，但科学家们开始借助实验室进行系统化研究，如伽利略的自由下落实验台。工业革命时期，科学与工程相互融合，科研基础设施发展为工程实验室。这时期的设施开始变得更为复杂，如牛顿的光学实验室，以及用于蒸汽机实验的实验室。蒸汽机的发展极大推动了工业革命的进行。电磁学的发展也推动了发电、电信等领域的进步。这些设施为工程科学的发展提供了有力支持。进入 20 世纪初，科学研究的复杂性和跨学科性越来越明显，科研基础设施进入大型化、集成化阶段。例如，粒子加速器、核反应堆等大型设施应运而生，为物理学、材料学等领域的突破提供了关键支持。图像技术的进步也推动了医学影像学的发展。随着信息技术的发展，科研基础设施开始向虚拟和网络化方向演进。如高性能计算机、云计算、远程实验室等，在很大程度上降低了研究成本，提高了研究效率。促进了全球范围内的科研合作。

而今天，以 AI for Science 新科研范式驱动下，我们在对基本物理原理规律的求解方面、对实验仪器中复杂数据的生成和处理方面、对于知识型文本如文献、专利等的理解和利用方面、对于新的计算软硬件设施的使用方面都有着全面的能力提高。而这些能力的提高目前多数还体现在点状的科研突破上，并没有全面的赋能整个科学研究，因此 AI for Science 时代的科研基础设施的打造成为了当下的重要工作。在原有的科学基础设施上，AI for Science 时代下的基础设施将将有以下四个方面的全面突破，我们将其概括称为“四梁”：其一，基本原理与数据驱动的算法模型和软件系统；其二高效率、高精度的实验表征系统；其三，替代文献的数据库与知识库系统；其四，高度整合的算力平台系统。而完成以上四个方面的系统建设，一来要面临着高度抽象化的领域知识门槛，二来要摆脱“作坊模式”推动科研想“平台模式”转变，这其中科学问题与工程问题相互交织，相互影响，因此推动科学家与工程师的充分协作是高效实现 AI for Science 时代科研基础设施建设的关键

因素。下面我们围绕"四梁"，就 AI for Science 时代下基础设施发展的机遇与挑战，重点建设内容展开讨论。在后续的第二至七章，我们将详述 AI for Science 在广泛场景中的具体实践。我们称这些垂直领域的应用和技术为"N 柱"。



来源: 北京科学智能研究院

I. 基本原理与数据驱动的算法模型和软件系统

机遇与挑战分析

在计算机诞生之后的几十年里，基于物理学基本原理的算法模型和软件发展取得了显著的成就，涉及到多个科学领域。在这个过程中，数学方法和计算技术的进步起到了关键作用。

在量子力学领域，薛定谔方程和密度泛函理论为研究分子结构和性质提供了理论基础。在这个领域，研究人员发展了许多数值计算方法，如 Hartree-Fock 理论、多体摄动论等。Gaussian（1970 年发布）、VASP（1990 年发布）等量子化学计算软件在材料科学、化学和生物学等领域具有重要的应用价值。

在分子动力学领域，牛顿运动方程和分子间相互作用力为基本原理，研究人员发展了广泛应用于生物分子、材料科学等领域的算法和软件。GROMACS（1991 年发布）、LAMMPS（1995 年发布）等软件为研究人员提供了强大的计算工具，帮助他们揭示了许多重要的科学现象。

在电磁学领域，以麦克斯韦方程为基本原理，研究人员发展了许多数值计算方法，如有限差分时域法(FDTD)、有限元法等。这些方法广泛应用于天线设计、微波器件、光通信等领域。相应的软件如 CST（1992 年发布）、HFSS（1990 年发布）等在工程实践中具有广泛的应用价值。

在固体力学领域，弹性力学和塑性力学等基本原理指导着有限元方法的发展。有限元方法成为了结构分析、材料性能研究等领域的核心计算方法。FEM 方法始于 NASA 的登月计划，其软件化成果 Nastran 于 1969 年由 MSC 进行商业化。在随后几十年内越来越多的厂商开始进入该领域并推出相应产品。诸如 ABAQUS（1978 年发布）、ANSYS（1970 年发布）等专业软件在汽车制造、桥梁建设、土木工程等行业中发挥着重要作用。

在流体力学领域，基于诸如欧拉方程和纳维-斯托克斯方程的基本原理，研究人员发展了一系列数值计算方法，如有限元方法、有限体积法和格子玻尔兹曼法等。这些方法被广泛应用于天气预报、航空航天、水利工程等领域。相应的，许多专业软件如 ANSYS Fluent（1983 年发布）、OpenFOAM（2004 年发布）等也应运而生，为工程师和研究人员提供了实用的计算工具。

这些基于基本原理的算法模型和软件的发展，极大地推动了科学研究和工程实践的进步。但我们也清晰地认识到，在真正解决实际问题时，除了流体力学取得了巨大成功以外，其他领域的进步更多是增量性的，而不是革命性的。造成这一局面的重要的原因就是这些领域的问题通常是多尺度的，而多尺度问题的求解中几乎全部的核心困难来源于高维问题，维数灾难使得如果不使用一些既不系统又十分繁琐的简化手段，计算将是一件在有限的时间尺度内不靠谱的方案，也因此计算总在精度和效率之间做着权衡取舍，且即使有时获得了一些好的解决方案，也因为简化方案的非系统性难以在更多使用场景下推广。

而在这个方面，以机器学习为代表的人工智能技术拯救了我们。以处理多变量能力见长的机器学习方法让众多以前被认为是不可能的事情现在正在变得高度可行。例如在处理"维数灾难"最早诞生地-针对高维控制问题的贝尔曼方程求解，基于深度学习算法人们可以毫不费力地处理高维的随机控制问题。在辅助物理建模方

面，机器学习工具在跨尺度建模方面提供了重要的支持，同时也进一步打开了通过观测数据优化物理方程的可行性，在这方面以 DeePMD（2017 年发布） 为代表的机器学习势函数助力分子动力学全面升级就是一个极好的例子。这些能力的补充，也为我们寻找新的科学规律提供了重要助力。

上面我们分析了基于基本原理的算法模型和软件系统的演化过程和重要机遇。接下来我们再看一下数据驱动的软件工具的演化历程以及新的机遇。在人工智能模型出现之前，数据驱动的软件发展经历了几个重要阶段。最早，依赖统计方法的分析软件逐渐应用于各个领域，例如最小二乘法、线性回归以及主成分分析等。随着计算机技术的发展，统计分析软件如 SPSS（1968 年发布）和 SAS（1976 年发布）开始被广泛使用，成为统计学家和数据分析师的重要工具。

随后，为了更高效地处理和分析大量数据，数据库管理系统（DBMS）开始被广泛应用。早期的数据库管理系统如 IBM 的 IMS（1968 年发布）以及关系型数据库管理系统 Oracle（1979 年发布）等，它们促进了数据驱动软件的发展。

进入 20 世纪 90 年代，随着数据量的不断增长，数据挖掘和知识发现领域开始兴起。数据挖掘是从大量数据中发现有用信息和知识的过程，其方法包括聚类分析、关联规则挖掘、分类和回归树等。这些方法在商业智能、市场分析、生物信息学等领域得到了广泛应用。同时，一些数据挖掘软件如 Weka（1997 年发布）和 RapidMiner（2001 年发布）等也应运而生。

而随着互联网和移动设备的普及，数据量呈爆炸式增长，传统的数据处理和分析方法难以应对这种海量、多样性以及实时性的数据。因此，大数据技术应运而生，以解决这些挑战，以 Hadoop(2006 年发布）为代表的分布式计算框架、以 MongoDB（2009 年发布）为代表的 NoSQL 数据库等成为大数据技术的重要代表。而机器学习技术也是在处理这类型大数据问题中逐步成熟，并诞生了 TensorFlow（2015 年发布）、PyTorch（2016 年发布）等框架，为机器学习和深度学习模型的开发提供了强大的支持。

随着机器学习技术的深入发展，科研人员处理海量高维观察数据的能力获得了极大提升。Alphafold2（2020 年发布）在蛋白质结构预测方面的成功就是一个非常好地例证。与此同时，基于大量无标注数据的预训练技术也给基于数据进行理论建模这一技术路线提供了重要帮助，比如深势科技推出的 UniMol（2022 年发布）。

以上的分析让我们较为清晰地看到，人工智能技术快速发展，让我们在解决实际问题上所迎来的全新机遇，也为我们发展基本原理与数据驱动的算法模型和软件系统提供了重要窗口期。但与此同时我们也要认识到无论从人才禀赋、协作机制上也都蕴藏着诸多挑战。

具体包括的建设内容

算法模型发展是核心动力引擎。有效地利用机器学习辅助物理建模，将基本原理原理与数据驱动模型相结合，是扩展算法能力边界，充分利用人工智能发展红利的重要措施。这一过程中有三点值得特别注意。

第一、机器学习模型必须满足所有的物理约束，例如守恒律与不变性，这对于构建可解释且可靠的模型具有重要意义。这里提到了"可解释性"这一非常有争议的话题，就简单展开一点。机器学习模型通常被大家称之为"黑箱"，而物理基本原理通常又被人们认为是人类对于这个世界最深刻的洞见。二者的结合是否令人信服，一直是备受争议的。但其实可解释性本身就是一个相对概念。我们现在在不同研究尺度上广为使用的基本原理如复杂气体的状态方程，我们对其的可解释性也通常停留在对欧拉方程守恒律的解释上，而对于状态方程的细节，如一些具体参数，其得到过程本身就是经验拟合的，只是我们习惯了使用这样的方程，所以我们大部分人对于利用这类型方程解决问题称之为可解释的，而对于保有物理约束的机器学习模型则认为是不可解释的黑箱，这个观点是有失偏颇的。鄂维南教授在 The Dawning of a New Era in Applied Mathematics 一文中对于我们构建可解释性的算法模型的目标是这样描述的"这些模型的基本出发点和基本结构是可解释的，这些模型中代表本构关系的一些函数的具体形式未必都得可解释。"

第二、用于模型训练的数据要具有代表性，这对我们的采样算法发展以及利用不同类型数据的能力都提出了更高的要求。在基础科研领域的数据与日常生活中的数据最大的不同是，基础科研领域相较于其所要研究的问题，数据往往是稀少且昂贵的，部分实验数据还存在噪声大、规范不统一等问题。因此主动高效地选择具有代表性的数据，广泛利用已有数据是算法模型发展的另一个值得注意的问题。

第三、积极探索底层算法框架，使其更适合科学研究本身。今天我们在研究人工智能技术助力基础科学研究时，通常将人工智能的算法模型框架如 transformer model、diffusion model 当做已知工具，进一步向其中加入基本原理或者科研数据，使其满足我们的使用需求。但是这类型的人工智能算法模型框架诞生之初并不是为了解决我们所面临的科学研究问题。要想在科研领域获得更好、更优质的解决方案就需要在底层算法框架层面做更深入的研究和持续性的尝试。

协同开发机制与评测机制是效率的基本保障。它们为各类创新项目提供了稳定的基础和保障。在这个时代，我们需要高效且成规模地解决一系列抽象困难问题，这就要求科学家与工程师有更深入的协作。同时，AI for Science 的快速发展让我们看到了利用基本原理解决实际问题的可行性。面对实际问题时，学科交叉是最为普遍的情形，这就要求我们能够跨领域地进行交叉合作。此外，我们需要更加重视评测机制，使得算法和软件可以在更高效的反馈下快速迭代。

首先，协同开发机制在科学家与工程师的合作中起到了关键作用。科学家通常擅长解决抽象的、困难问题，而工程师则擅长解规模和效率的问题。在现代科技项目中，这两者的结合变得越来越重要。科学家需要借助工程师的技能将理论应用于实际问题，而工程师也需要理解科学家的理论框架，以便为项目提供更为精确的解决方案。好协同开发机制如开源社区的组织模式、新型研发机构的组织模式等正是为了解决这种需求，它鼓励科学家和工程师紧密合作，共同解决问题，实现资源共享和知识共享。

其次，AI for Science 的发展让学科交叉合作变得尤为重要。在解决现实问题时，我们往往需要多个领域的专家共同参与。例如，在生物医学领域，我们需要生物学家、化学家、物理学家、计算机科学家等多个领域的专家共同解决问题。跨学科的合作可以使项目从不同的角度获得支持，为问题的解决提供更多可能性。同时

需要认识到，跨学科合作不是简单的物理组合，而是需要每一个参与合作的个体都充满了对与其紧密合作伙伴所擅长领域的强烈好奇心和基本能力素养，这往往是跨学科合作的前提也是最为困难的点。

再次，软件评测机制在新时代的算法模型与软件系统建设中具有关键作用。过去，我们的基础科研软件经过了长时间的打磨才发展到今天。然而，在 AI for Science 时代，我们需要快速构建起新的算法模型和软件系统。为了实现这一目标，我们需要建立更为完善的评测机制。通过构建自动化评测系统使得下游应用方可以方便的为算法模型与软件添加测试样例。算法模型与软件可以在快速迭代的过程中保证其鲁棒性。通过有效的评测机制，我们可以在短时间内实现算法和软件的高质量发展。

II. 高效率、高精度的实验表征系统

机遇与挑战分析

实验是科学研究的重要工具之一，实验过程中需要使用一系列的仪器、设备和技术手段，以得到准确、可靠和有意义的数据。我们将通过对实验系统进行定量或定性的测量和分析，以描述和理解实验系统中的物理、化学或生物现象的研究方法称为实验表征。实验表征可以采用各种方法，如光谱分析、物理测量、化学试验、生物学检测等，以获得实验数据并对其进行分析和解释，从而得出结论和推断。以电化学里面电池为例，用于表征电池结构的所有实验表征的工具目前有：电解质-体态结构探测振动光谱：红外、拉曼光谱，电解质/电极-界面结构探测振动光谱：和频光谱 (SFG)、表面增强拉曼光谱，X 光谱：XRR (X-ray reflectivity, X 光反射谱)，电极-体态结构探测局域结构：XAS (X-ray absorption, X 光吸收谱)、NMR 长程结构：XRD (X-ray diffraction, X 光衍射谱)。然而，实验表征系统也存在一些制约因素，这些因素可能会影响实验结果的准确性和可靠性，甚至可能导致误解和错误的结论，比如：

（1）仪器和设备的精度和稳定性：实验表征需要使用各种仪器和设备，如电子显微镜、质谱仪、红外光谱仪等。这些仪器和设备的精度和稳定性直接影响实验结果的准确性和可靠性。如果仪器和设备的精度和稳定性不够高，可能会导致误差的累积，从而影响实验结果的准确性。

（2）样品准备和处理的影响：实验表征需要对样品进行准备和处理，以使其符合实验要求。然而，样品准备和处理的过程可能会引入一些变异性和不确定性，从而影响实验结果的准确性和可靠性。例如，在样品制备过程中可能会引入杂质或污染物，从而影响实验结果的准确性。

（3）实验条件的控制和稳定性：实验条件的控制和稳定性对实验结果的准确性和可靠性有着至关重要的影响。例如，在电子显微镜观察样品时，需要控制环境温度和湿度，以避免样品受到环境因素的影响。如果实验条件不够稳定或不能有效地控制，可能会导致实验结果的不确定性。

（4）数据分析和处理的误差：实验结果的数据分析和处理也可能会引入误差。例如，在质谱分析中，需要进行数据解释和峰识别，这可能会引入一些主观性和误差。如果数据分析和处理的误差不加以纠正，可能会影响实验结果的准确性和可靠性。

（5）样品的复杂性和变异性：一些样品可能非常复杂，例如生物样品和环境样品。这些样品的复杂性和变异性可能会导致实验结果的不确定性。例如，在生物样品中，可能存在多种成分和变量，这些成分和变量可能会相互影响，从而影响实验结果的准确性和可靠性。

为了解决以上问题，人们主要做了以下方面的尝试，同时也遇到了一些瓶颈问题：

第一、制造可以观察更精细的仪器设备。例如，同步辐射环是一种粒子加速器，它产生的高强度、高亮度的同步辐射光源被广泛应用于多种科学领域，如物理学、化学、生物学、材料科学等。同步辐射光源具有高亮度、连续光谱、高时间分辨率、高空间分辨率等优点，使得科学家可以在原子和分子水平上研究材料的结构

和性质。然而，高精度设备通常成本昂贵、技术复杂，并需要专业的操作和维护。随着测量精度的提高，测量误差和噪声控制也变得越来越困难。

第二、不断提升实验自动化水平，在药物研发领域，高通量筛选技术可以自动化地测试大量化合物，快速找到具有潜在治疗价值的候选药物。此外，实验室自动化系统，如液体处理机器人，可以在短时间内完成大量实验操作。实验自动化水平的提升极大地解决了实验的标准化程度和数据的高通量获取的难题。然而，这一过程也面临着两个核心困难，一个是高维控制算法的开发，这个决定了我们可以在做多么复杂的自动化实验，另一个是多种设备、传感器和控制器可能出自不同的制造商，使用不同的通信协议和数据格式，甚至不少仪器设备的控制系统根本无法对外开放，在将这些设备集成到一个自动化系统中时，会不断遇到统一和兼容性的问题。

第三、持续推进反演算法的研究。反演算法是一种通过将实验数据与理论模型相结合，来估计物质系统的内在参数和状态的计算方法。因此，反演算法在实验表征设备中起着至关重要的作用，它们为科学家提供了一种从实验数据中获取深入洞察的手段。自 20 世纪初以来，随着实验技术的不断发展，实验表征设备和反演算法也在不断进步。在早期，科学家们主要依靠简单的理论模型和人工计算来进行反演。然而，随着实验数据的复杂度和数量不断增加，这种方法变得越来越不适应。从上世纪 50 年代开始，随着计算机技术的发展，科学家们开始使用计算机辅助的反演方法。这些方法通常基于数值优化和迭代求解等技术，可以处理更复杂的实验数据和模型。地震学家利用反演算法将地震波数据转化为地球内部结构的信息，从而揭示地壳和地幔的结构、地球动力学过程和地震活动。在医学成像领域，磁共振成像（MRI）和计算机断层扫描（CT）等成像技术使用反演算法将原始数据转换为对病变组织和器官的清晰可视化图像。然而，这些传统方法在处理非线性、多尺度和多物理场问题时仍然面临挑战，这也就是说，我们有时看到的由实际仪器设备生成的“图”、“谱”并不是他真实的样子，这也是限制实验表征精度的重要环节。

第四、研究更高效的数据收集、整理、分析的方法。这一点非常的直观，对于实验数据的处理一直以来就是我们科学家主要的进行科研探索的环节。2010 年以后，大数据技术得到了广泛的发展，通过大数据技术可以高效地存储、管理和分析这些数据。这也是近些年来推动像生物信息学快速发展重要的原因，例如，在基因测序中，研究人员需要处理数以亿计的基因数据。通过使用大数据技术和分布式计算平台，如 Apache Hadoop 和 Spark，研究人员可以快速完成基因序列分析，加速科学研究进程。而最近，以机器学习和人工智能技术的诞生，就进一步加速了这部分的进展。然而，我们实验数据的数量、标准化程度等问题依旧是限制领域发展的关键因素，如何在小样本场景下最大程度的实现数据价值的挖掘，构建起可靠的模型也是新的挑战。

在推动实验表征技术高速发展的路上，我们可以看到，除了制造可以创造更理想实验条件的仪器设备外，在 AI for Science 新科研范式的加持下，无论是反演算法的发展、还是自动化仪器系统的构建、还是更高效的数据收集整理和分析的方法的升级，都将迎来新的重大突破。这也是今天我们将构建高效率、高精度的实验表征系统列为 AI for Science 时代下基础设置“四梁”之一的重要原因。

具体包括的建设内容

AI for Science 新的科研范式正在引领反演算法研究进入一个全新的阶段。反演算法的开发通常要包含以下的一些过程：

- 明确问题的背景和目标，这个过程主要是对于用于输入的探测信号以及最终要得到的具体任务的明确。以地球物理领域的地震波反演问题为例，其目标是根据地震波观测数据来推断地下介质的速度结构和地质特性。
- 构建数学模型：基于问题背景和目标，建立描述数据与未知参数之间关系的数学模型。这个模型一般包括正问题（已知参数求观测数据）和逆问题（已知观测数据求未知参数）。在建立模型时，需充分考虑实际问题的物理、化学或生物学原理，以确保模型的合理性和准确性。
- 分析问题的适定性。（适定问题是指问题具有存在性、唯一性和确定性，不如果一个问题不满足上述三个条件中的任何一个，那么它就被认为是一个不适定问题。）检查数学模型的适定性，判断问题是适定问题还是不适定问题。若问题不适定，可以通过引入先验信息、正则化技术等方法，提高问题的稳定性和可靠性。在实际应用中，不适定问题很常见，特别是在逆问题领域。由于观测数据可能包含噪声、不完整或受到各种约束条件的影响，因此逆问题往往具有不适定性，如何在算法中有效地处理噪声和数据不完整性，以提高反演结果的准确性和稳定性，是一个重要的难点。
- 选择或开发求解算法并不断验证与优化：通常情况下，我们所建立的可以既包含基本原理的约束又满足适定性要求的模型，大多为非线性模型。在选择模型求解方案时，要同时考虑算法的收敛性、稳定性、计算复杂性等因素。这也是在反演算法真正运用到实际问题求解过程中面临权衡取舍的关键。

经过以上的介绍不难看出，通过构建基本原理驱动的人工智能算法模型，既可以满足模型对于物理、化学或生物等基本原理的要求，又在很大程度上实现了求解过程中精度与效率的统一，为算法求解提供了重要选项，例如，在材料科学领域，研究人员利用深度学习方法对电子显微镜图像进行分析，成功识别出纳米级的晶体结构，从而提高了实验表征的精度。在生物医学领域，科学家们使用深度学习方法对生物组织的光学成像数据进行实时分析，成功检测到细胞和组织的异常变化，为临床诊断和治疗提供了重要依据。与此同时，在解决适定性问题方面，人工智能技术也提供了重要帮助，例如物理信息神经网络（Physics-Informed Neural Networks，PINN）或许它还并不是一个解决正问题的好方案，但是一种有效解决适定性问题的方案，特别是在解决偏微分方程（Partial Differential Equations，PDEs）类的逆问题时。

AI for Science 反演算法的发展提供了强大的助力。通过引入深度学习和其他 AI 方法，科学家们可以实现更准确、稳定、高效的实验表征，从而推动科学研究和实际应用的发展。未来，我们有理由相信，随着 AI 技术的不断创新和发展，有了更优质的反演算法，实验表征技术将更好地服务于科学发现和技术进步。

智能化实验室将成为未来实验室发展的主流。实验方案的智能化生成在现代化实验中扮演着至关重要的角色。基于通过利用大型语言模型，从文献、专利、教科书等数据源中挖掘的化学数据，建立化学知识图谱，

根据实验设计目的，智能化生成实验方案，我们可以大幅提高实验效率，降低实验成本，同时也能提高实验的准确度和可重复性。这种方法已经成为实验的首要步骤，将是大部分现代实验的必备技术。这部分内容我们将在"四梁"的下一章更充分地展开。

控制算法和理论模型的全面升级将使得构建智能化调优实验系统成为现实。在上一章关于算法模型发展的部分已经分析了，以机器学习为代表的人工智能技术对高维复杂函数的拟合能力使我们可以轻松实现上百维度控制问题的求解，这为我们构建自动化实验系统提供了最为重要的助益。与此同时，以 Alphafold 为代表的数据驱动的人工智能模型的成功，让我们在构建实验和积累数据时会更加侧重标准化这一维度。高通量自动化的实验工具将为我们提供大量标准统一的实验数据，这将为构建数据驱动的算法模型打下最为坚实的基础。而一旦我们同时掌握了先进的算法模型和可以自动化执行复杂实验任务的实验系统时，我们就可以在此基础上构建起更为智能化的可以自主调优的实验系统。算法模型可以对实验结果进行实时分析，并指导实验向更优方向调整，从而改变了盲目试错的方式，大幅度提高实验效率，降低实验开销。

III. 替代文献的数据库与知识库系统

机遇与挑战分析

随着科学的不断进步和发展，很多关键学科领域的重要理论被一一创造或发现。现在，前沿理论的出现往往存在于一些交叉学科上，是不同研究成果的互相启发带来了新的理论突破，这不仅得益于产业的融合、学科的交互，更是因为有了信息技术的加持。数字图书馆的出现提供了大量数字化学术文献，学术搜索引擎如 Google Scholar、CNKI 等可以帮助用户快速检索到相关的学术文献，DOAJ、PLOS 等提供了免费的学术文献下载服务，用户可以更便捷地获取学术资料，学术社交网络 ResearchGate、Academia.edu 等平台可以帮助用户快速找到相关领域的专家和研究成果。人们可以更好地获取学术资料，今天的问题在于海量文献不断出现，EndNote、Zotero 等可以帮助用户整理和管理学术文献。但信息越多，人们似乎越焦虑。首先，科研人员无法穷尽全部文献，在众多文献中给出判断和选择；另外，整理、阅读、辨析、总结文献的过程，占据了科研人员太多精力和时间，也让他们在一定程度上无法专注于科学研究。而这给了 AI for Science 一个让科研效率提升突破的机会。

a) 大语言模型

大语言模型是一种能够自动生成自然语言文本的人工智能技术，它主要基于自然语言处理、机器学习和深度学习等技术。它们可以生成高质量的文本内容，为人们提供更为便捷的自然语言交互方式。大语言模型是目前自然语言处理领域的热点之一，很多人认为，未来随着技术的不断发展，大语言模型将会越来越普及和应用。

大语言模型是基于自然语言处理、机器学习、深度学习等技术实现的。自然语言处理（Natural Language Processing，NLP）是大语言模型的前提技术，它是一种将人类语言与计算机交互的技术。自然语言处理包括分词、词性标注、命名实体识别、句法分析、语义分析等多个子任务，用于将自然语言转化为计算机能够理解和处理的形式。在大语言模型中，自然语言处理技术主要用于分析文本数据，提取文本特征，为后续的机器学习和深度学习提供数据基础。机器学习（Machine Learning）是大语言模型中的一项核心技术。它是一种以数据为基础，通过统计学和算法实现预测和决策的技术。在大语言模型中，机器学习主要用于文本分类、文本聚类、特征选择、模型训练等任务。其中，最常用的机器学习算法包括朴素贝叶斯、决策树、支持向量机、最大熵模型等。深度学习（Deep Learning）是大语言模型中的另一项核心技术。它是一种基于神经网络的机器学习技术，能够自动提取特征、学习模型，实现从输入到输出的端到端映射。在大语言模型中，深度学习主要用于文本生成、文本分类、文本情感分析、文本摘要等任务。其中，最常用的深度学习算法包括卷积神经网络、循环神经网络、长短时记忆网络等。

大语言模型在机器翻译、语音识别、自动文摘、问答系统等领域有广泛的应用。它可以通过学习不同语言之间的语言规则和语境，实现将一种语言自动翻译成另一种语言。其中，最常用的方法是序列到序列模型，它

通过将输入序列映射到一个固定长度的向量，再将该向量映射到输出序列，从而实现机器翻译任务。大语言模型在语音识别领域也有广泛的应用。它可以通过学习语音信号和文本之间的对应关系，实现从语音信号到文本的自动转换。其中，最常用的方法是基于深度学习的语音识别方法，它通过将语音信号分帧、提取特征，然后通过多层神经网络进行训练，最终实现语音识别任务。此外，大语言模型还能够对文本进行概括和摘要，它可以通过学习文本的主题和关键信息，实现自动提取文本摘要的功能。其中，最常用的方法是基于文本分类的自动文摘方法，它通过将文本分为不同的类别，然后选择最相关的类别和关键信息，最终生成文本摘要。大语言模型在问答系统领域也有应用。它可以通过学习自然语言问答的语言规则和语境，实现从用户提出的问题中自动提取答案的功能。其中，最常用的方法是基于深度学习的问答系统方法，它通过将问题和答案表示为向量，然后计算它们之间的相似度，最终选择最相关的答案。

自动化程度高、具有较高的准确率和效率极大解放了人类工作中的重复性劳动，但是它也有一些缺点。大语言模型对数据的数量规模要求很高，需要大量的语料库对模型进行训练，如果数据不足或数据质量不好，可能会影响模型的性能。此外还会在交互过程中出现延时，导致误差传递问题，影响交互者获得信息的准确性。

那么，对于自然科学研究来说，GPT 是否就够了？答案显然是不够的。在 AI for Science 领域，ChatGPT 也不完全适用于学术性问答的交互。经过各种广泛的测试，GPT 本身的局限性是在学界内有广泛共识的。科学文献大多为非结构化数据，且专业属性极强，各个领域的文献呈现爆炸式增长，更新速度快。即使大语言模型（LLM)技术的兴起深刻影响着众多行业领域，科研场景也不例外，但大语言模型本身仍存在着不可解释性、对于较为细分的领域知识难以快速覆盖等问题。具体来说，科学研究要求严谨性，需要依靠准确的知识体系和知识系统作为支撑，这就要求语言模型的背后应当存在一个准确的知识库文献库数据库作为"智囊"。此外，一些数据不仅局限于文献本身，实验数据也是重要的数据来源，而这些真实的实验数据恰恰是 ChatGPT 所缺少的。第三，来自于数据的滞后性。由于大模型的结构特点，每一次重新训练的成本都是非常高昂的，怎样让最新的数据高效的覆盖进来也是目前面临的重大问题。第四是输入的缓存限制，是 Transformer 结构本身决定的，单词输入和输出的 Token 总数总是限制在 4K 或者 32K 数字之内，局限了每一次问答的长度和问答的精细度。

b) 将大语言模型与数据库相结合

大语言模型与数据库结合是可能的。大语言模型可以通过对大量文本数据进行训练来生成语言模型，然后可以将这个模型与数据库结合使用，以便对数据库中的文本数据进行自然语言处理和语义分析。这种结合可以帮助提高数据库的查询和分析能力，从而更好地满足用户的需求。例如，可以使用大语言模型来自动化生成数据库查询语句，或者使用大语言模型来识别和提取数据库中的实体、关系和事件等信息，以便给用户更加智能化和个性化的查询和分析服务。

大语言模型和数据库结合的技术原理主要是将大语言模型作为一个预测模型，用于预测数据库中的数据，从而实现快速的数据查询和检索。具体来说，大语言模型可以通过学习数据库中的数据，提取其中的关键信息，然后将这些信息用于预测数据库中的其他数据。传统的数据库查询方式是通过 SQL 语句进行查询，但是这种查询方式存在一些不足，比如需要手动编写 SQL 语句，查询结果不够精准等。而大语言模型和数据库结合后，可以通过自然语言进行查询，这样不仅可以提高查询效率，还可以减少查询错误的可能性。例如，一个用户可以通过自然语言询问"我想知道最近一个月内被提及最多次的技术是什么"，数据库就可以根据用户的需求返回相应的结果，提高了查询的准确性和效率。

这种结合方式可以使得数据库的查询速度大大提高，同时也可以提高数据的精确度和可靠性。在数据分析的过程中，数据的质量和准确性非常重要。但是，传统的数据分析方法很难对数据进行准确的描述和分析。而大语言模型和数据库结合后，可以通过自然语言进行数据分析，这样可以减少数据分析过程中的错误，提高数据分析的精度。例如，一个用户可以通过自然语言询问"我想知道最近一个月内人工智能领域发表的论文中出现频率最高的技术是什么"，数据库就可以根据用户的需求返回相应的结果，这样可以帮助用户更好地了解技术发展情况，提高数据分析的精度。

大语言模型和数据库结合还可以提高数据管理的效率。在传统的数据库管理中，需要手动编写 SQL 语句进行数据的管理，这种方式效率低下，容易出错。而大语言模型和数据库结合后，可以通过自然语言进行数据管理，这样可以提高数据管理的效率和准确性。例如，一个用户可以通过自然语言进行数据删除操作，"删除最近一个月内指标低于 XX 的实验数据"，数据库就可以根据用户的需求进行相应的操作，这样可以减少人工操作的时间和错误。

在实际应用中，大语言模型和数据库结合已经得到了广泛的应用。其中，最为典型的应用是搜索引擎。搜索引擎可以通过大语言模型来预测用户的搜索意图，并将这些意图与数据库中的数据进行匹配，从而提供用户最相关的搜索结果。此外，大语言模型和数据库结合还可以用于推荐系统。推荐系统可以通过学习用户的历史行为和偏好，然后将这些信息用于预测用户可能喜欢的物品或服务，并将这些预测结果与数据库中的数据进行匹配，从而提供用户最符合其需求的推荐结果。

除了搜索引擎和推荐系统之外，大语言模型和数据库结合还可以用于智能客服。智能客服可以通过学习用户的问题和需求，然后将这些信息用于预测用户可能的解决方案，并将这些预测结果与数据库中的数据进行匹配，从而提供用户最合适的解决方案。此外，大语言模型和数据库结合还可以用于智能家居、智能医疗、智能交通等领域，以提高系统的智能化程度和数据处理效率。

具体包括的建设内容

文献"管家"的诞生：根据美国国家科学基金会统计数据，科研人员花费在查找和消化科技资料上的时间，需占全部科研时间的 51%，计划思考占 8%，实验研究占 32%，书面总结占 9%。检索从过去到现在经历了三个阶段大的演进，从最早的眼查手翻查阅式检索，到后来基于搜索引擎以及互联网搜索式检索，再到现在 GPT 横空出世，首次看到大语言模型在问题的理解和问答能力上接近人类智能的水平，我们认为检索的模式将进入对话时代。如果我们将大语言模型与数据库相结合，是否可以设计出一个具有学术背景的 ChatGPT?答案是可能的，并且已经照进了现实。

简单说，就是将所有文献、实验数据等整合成一个知识库或数据库。在此基础上，通过简单的自然语言对话或其他搜索、查询方法，我们可以较为明确、简单地找到所需的知识，无需再花费更多时间阅读和总结文献。此外，我们还可以利用人工智能算法反过来对数据进行优化，例如在合成化学领域，建立合成化学文献的附录，描述已有实验的步骤和结果。通过结构化查询，我们可以将其转化为合成化学实验知识库或数据库，开发人工智能算法进行对比分析，自动给出最佳合成化学路径的建议。

此外，还有 AI 数据库，它能够为整个人工智能的发展带来新的方法。目前，大部分人工智能模型都是从原始数据开始发展，尽管人工智能目前已经相当成功，但仍然不能避免在原始数据基础上直接发展模型，例如人脸识别等等。AI 数据库提供了一种新的方法，即通过加入中间层 AI Data，将数据转化为一张巨大的表格。这只是一个比喻，但是可以想象，如果我们的所有数据都可以分析在一个表格中，我们就不仅仅可以使用机器学习这一方法来进行模型构建，还可以进行查询和简单的表格处理，这是一个能够改变我们人工智能方法的工具。此外，它也可以从数据和知识库的角度进行应用，如文献检索等，提高检索的效率，减少占用的计算资源。

一个合格的文献"管家"应该有哪些功能？大语言模型的科学文献数据库（文献"管家"），不仅可以实现科学文献的分类检索和智能推荐等基础功能，还可以智能抽取文献信息，生成文档知识，并构建相应的数据库。为了实现这一目标，我们将研究大语言模型的关键技术，包括文献信息智能抽取、文档知识智能生成以及文献数据的智能化管理等。同时，我们也将实现数据可查询、可比对、可分析等功能，以便用户能够快速地找到自己想要的文献信息。此外，我们还将研究实验表征数据、计算模拟数据、文献数据等多模态数据统一管理、统一查询、统一分析的下一代数据库技术，并完成相关数据库的功能建设，以便用户可以在同一个平台上获取不同来源的数据。最终，我们的研究成果将能够为科学研究提供更加便捷、高效的文献查询和数据管理服务。

构建过程中要注意以下几个方面：

a) 数据库建设。文献数据库的建设是核心，因此必须建立高质量、全面、准确的文献数据库。数据来源上，文献数据来源应该尽量广泛，包括各种学术期刊、会议论文、学位论文、专利、技术报告等。数据质量：建

立文献数据库时，需要对数据进行清洗、去重、标准化等处理，以确保数据的质量和准确性。数据格式：文献数据库应该采用统一的数据格式，方便数据的管理和处理。

b) 大语言模型。大语言模型是核心技术之一，它能够提供高质量的文献推荐和检索服务。模型训练上，大语言模型需要通过大规模的文献数据进行训练，以获取语言模型的参数。模型优化中，为了提高模型的准确性和效率，需要对模型进行优化，包括模型结构优化、参数调整等。此外，在模型集成上，为了获取更好的效果，可以采用多个模型进行集成，例如使用多种模型进行推荐和检索，融合它们的结果。

c) 系统架构。文献管家需要考虑到系统的性能和可扩展性，因此需要设计合理的系统架构。分布式架构：为了提高系统的性能和可扩展性，可以采用分布式架构，将系统拆分成多个模块，分别运行在不同的服务器上。缓存设计：为了提高系统的响应速度，可以采用缓存技术，将常用的数据存储在缓存中，减少数据库的访问次数。负载均衡：为了避免单个服务器过载，可以采用负载均衡技术，将用户请求均衡地分配到不同的服务器上。

d) 安全性。文献管家需要考虑到用户数据的安全性，因此需要采取相应的安全措施。安全措施需要考虑以下几个方面：数据加密，为了防止用户数据被泄露，需要采用数据加密技术，将用户数据加密存储在数据库中。权限管理，为了保护用户数据的安全，需要采用权限管理技术，根据用户角色和权限限制用户的操作。防止攻击，为了避免系统被攻击，需要采用防火墙、入侵检测等技术，保护系统的安全。

IV. 高度整合的算力平台系统

算力平台的现状分析

数据、算法、算力，一直以来被誉为人工智能发展的三大要素。算力的发展除了依赖计算芯片、网络、存储、服务器等硬件本身的发展以外，硬件之上的软件编译层、机器学习层、任务调度层等同样是将算力充分释放的不可忽略的因素。在讨论构建面向 AI for Science 的高度整合的算力平台系统时，我们将针对人工智能计算和传统科学计算的计算特点，分别就硬件和软件发展方面的现状和挑战展开分析。

典型的人工智能计算如深度学习涉及大量的矩阵运算，特别是在卷积神经网络（CNN）和循环神经网络（RNN）中。其计算特点从整体来看，具有计算密集、可并行程度高、精度要求低的特点，在模型训练过程中还具有参数占存储空间大、训练数据大、多计算单元并行计算时对通讯的网络带宽要求高等特点。

典型的科学计算场景主要涉及数值计算、离散化、误差分析等，其计算过程通常需要高精度的浮点运算如双精度（FP64）以实现运算收敛，从而保证计算结果的准确性。同时，科学计算对内存的需求也较高，主要是因为需要存储大量的数据和中间结果，大容量的内存有助于提升科学计算应用的性能。

从计算芯片发展的视角来看，在本世纪初的十年，计算芯片领域主要以中央处理器（CPU）为主导，其中英特尔（Intel）和 AMD 两家公司长期保持竞争。伴随着摩尔定律的发展，计算机芯片的性能不断提高，科学计算软件在并不需要进行额外的高性能优化工作的情况下，便获得了不错的效率提升。2007 年英伟达推出了 CUDA（Compute Unified Device Architecture）编程模型，将 GPU 的计算能力扩展到了通用计算领域，使得科学家和研究人员可以利用 GPU 强大的并行计算能力解决复杂的科学计算问题。

2010 年后，原本面向科学计算市场打造的 GPU 与人工智能发展的浪潮相遇。正如前文对人工智能计算特点的分析中提到的，神经网络计算中所需的大量矩阵运算恰好是英伟达显卡所擅长的领域，自此英伟达的 GPU 与人工智能开始了相互促进的发展历程。与此同时，伴随着人工智能发展对于计算芯片需求的高速增长，计算芯片市场也得到蓬勃发展。在国际上如 AMD、Intel，在国内如华为、寒武纪、壁仞科技、摩尔线程、燧原科技、天数智芯等企业也纷纷投入到了支持人工智能计算需求的芯片生产市场中。而英伟达从科学计算逐步扩展到人工智能的发展历程是其相比其他后入局的厂商在 AI for Science 领域更具吸引力的主要原因之一。

说到用户对硬件的选择，就不得不涉及软件开发者如何开发出可以在硬件上奔跑的代码这一过程。机器学习框架的发展和与硬件相配合的编程模型的发展是其中影响最大的两个环节。随着计算能力的提升和大量数据的积累，机器学习技术逐渐成为人工智能领域的核心技术之一。机器学习算法的设计、实现和优化需要具备相当高的专业知识，对于许多研究人员和开发者来说，这些过程非常复杂和耗时的。为了解决这些问题，许多专家开始尝试创建更高级别的抽象工具，以便更容易地构建、训练和部署机器学习模型。这就是机器学习框架诞生的背景。其主要目的是，简化模型的构建和训练过程，降低机器学习的门槛；提供预先实现的算法和模型，使开发者能够快速实现自己的任务；提供自动优化和硬件加速功能，以提高模型的性能；促进模型的共享和复用，便于研究人员和开发者之间的合作。2010 年，谷歌发布了 DistBelief，这是一种用于构建、训

练和部署深度神经网络的分布式系统，为后来的框架发展奠定了基础。2015 年，谷歌推出了 TensorFlow，以其灵活性、可扩展性和广泛的社区支持迅速成为最受欢迎的机器学习框架之一。2015 年，微软推出 CNTK（Microsoft Cognitive Toolkit），它具有高性能和分布式计算能力，适用于大规模深度学习任务。2016 年，PyTorch 发布，这是一个基于 Python 和 Torch 的深度学习框架，以其动态计算图和易用性受到研究人员的欢迎。而在国内，2016 年百度也推出了自己的机器学习框架 PaddlePaddle，2020 年华为推出了 Mindspore。可以说今天 99%以上的机器学习代码是利用机器学习框架构建的。

因此机器学习框架的发展对 AI for Science 的整体发展也起到了推动作用。

相较于机器学习框架提供的自动硬件性能优化的能力，编程框架的影响更为直接。比较为人所熟知的编程框架有

- CUDA（Compute Unified Device Architecture）：英伟达推出的并行编程框架，它使得开发者能够充分利用 NVIDIA GPU 的强大计算能力，以解决各种复杂数学和计算密集型问题。在很多领域，如深度学习、图形处理和科学计算等，CUDA 已经成为了事实上的并行计算标准。
- OpenCL（Open Computing Language）：这是一个开放标准的并行计算框架，由 Khronos Group 开发。它允许使用多种处理器（如 CPU、GPU 和其他加速器）进行通用计算。与 CUDA 不同，OpenCL 支持各种硬件厂商，包括 NVIDIA、AMD 和 Intel。
- ROCm（Radeon Open Compute）：这是 AMD 开发的一种开源的 GPU 计算平台，用于 AMD GPU。它包括一个运行时系统、编程语言、工具链和库，支持各种编程模型，如 OpenCL、HIP（用于将 CUDA 代码移植到 AMD GPU）和其他。
- OpenACC：这是一个编程标准，用于简化在加速器上（如 GPU 和其他并行处理器）进行通用计算的编程。与 CUDA 和 OpenCL 不同，OpenACC 使用编译器指令和运行时库来自动处理并行性，无需显式地管理数据移动和线程同步。
- OpenGL Compute Shader：这是一种在 OpenGL 中执行通用计算任务的方法。虽然主要用于图形处理，但它也允许开发人员利用 GPU 进行通用计算。
- Intel oneAPI：这是 Intel 推出的统一编程模型，旨在简化在 Intel CPU、GPU 和其他加速器上进行通用计算的开发。它包括 DPC++（Data Parallel C++）编程语言、编译器和运行时库。

必须要说明的是，与其他编程模型相比，CUDA 具有成熟而被广泛使用的优势，自 2006 年推出以来，CUDA 已经成为 GPU 计算的事实标准，并得到了广泛的支持和采用。许多软件库和框架（如 TensorFlow、PyTorch 等）针对 CUDA 进行了优化，使英伟达 GPU 在这些应用中具有更好的性能。与此同时，CUDA 为英伟达 GPU 提供了许多高级优化和调整功能，使开发人员能够充分利用 GPU 的计算能力。这使得 CUDA 程序通常比使用其他编程模型的程序具有更高的性能。在此基础上，英伟达投入了大量资源来支持和发展 CUDA 生态系统，

包括编译器、调试器、性能分析工具、库和文档。这使得开发人员能够更容易地学习和使用 CUDA 进行高性能计算。而换个角度来看，CUDA 也成为其他芯片生产厂商与用户建立链接最先需要突破的一关。

简要谈完了计算芯片和中间软件层的发展历程后，再简单分析一下两种不同的硬件组织形式，一种是超级计算机，另一种是云计算。

超级计算机的诞生，缘起于 20 世纪 60 年代计算机性能的提升。在那个时代，单台计算机的处理能力已逐步提高，但仍难以满足日益增长的科学研究对计算资源的需求。譬如气候模拟、空间技术等都对计算机提出了更高要求。为了提供更强大的计算能力，超级计算机作为一种通过大规模并行来获得超高速度的计算机应运而生。1964 年面世的 CDC 6600 就是第一台超级计算机，开启了"算力求真"的新纪元。在那以后，IBM、Control Data、Cray 等公司都投入到超级计算机的研发与创新中去，推动它们在速度、规模等方面的不断进步。可以说,超级计算机满足了科学家对工具的需求，极大地推动了高能物理、空间科技、气候科学等领域的发展。换句话说，超级计算机为推动科学计算做出了巨大的贡献。

而云计算的出现，也是对时代需求的回应。20 世纪 90 年代互联网的高速发展产生了海量数据。单一的企业服务器已经难以存储和计算这些数据。为了提供可扩展的存储和计算资源，云计算概念开始兴起。它实现通过服务器集群来提供各种服务，使企业和用户可以按需获取所需的计算资源。以亚马逊的 AWS 云服务为例，它从 2006 年推出至今，已经提供了非常丰富的云服务，包括计算、存储、数据库、网络、分析、机器学习等等。通过虚拟化、容器化等技术，大幅抽象掉了底层计算硬件的差异，降低了运算环境维护的成本。用户只需要根据需要调用 API，就可以获得充足弹性的资源。云计算极大地降低了获取计算资源的门槛，促进了大数据应用和互联网服务的繁荣。

而今天伴随 AI for Science 能解决的实际问题范围的逐步变大以及计算模拟的普及，超级计算机和云计算正在以相互走近的方式开始了新的融合。无论是超算云化，还是云上超算，都为推动算力基础设施朝着更好的方向做出了积极地尝试。

具体包括的建设内容

a) AI for Science 的专用芯片

随着通用芯片开发难度的逐步增大，尤其是当我们考虑到 2 纳米和 1 纳米等越来越小的尺寸。然而，在大部分科学计算领域，我们通常只使用有限的几种计算形式，伴随着 AI for Science 在算法上的逐步成熟，这些有限的计算形式也将在接下来 2-3 年内趋于稳定。因此，如果我们可以为每种计算开发高效且高质量的专用芯片，这可能会成为计算芯片更优质的解决方案。在过去，相比于通用处理器，专用芯片在功耗、性能和成本等方面具有优势，因此被广泛应用于各种领域，如通信、图像处理、嵌入式系统等。

在科学计算场景中，存在着大量从内存中读写数据的过程，传统的计算机芯片在处理这样的操作时显得非常缓慢，业界称之为"内存墙"。而专用芯片则可以规避这样的操作，在短时间内完成大量的计算任务。另外，科学研究中往往需要进行长时间的计算，因此功耗成为一个重要的问题。由于专用芯片针对特定任务进行了优化，它可以在更低的功耗下实现更高的性能。例如，专用芯片可以用于加速分子动力学模拟，从而在保证计算精度的同时降低计算机的能耗。

目前，国内已经可以看到在 AI for Science 专业芯片领域较为前沿和成熟的研究。2022 年，湖南大学电气与信息工程学院刘杰教授课题组自主研制出了"存算一体"非冯·诺依曼类脑芯片架构，并基于 FPGA 研制出了基于新型非冯·诺依曼芯片架构的分子动力学计算系统"NVNMD"（第一版），实现了从传统冯·诺依曼芯片架构向新型非冯·诺依曼芯片架构的"范式转移（paradigm shift）"，用于加速分子动力学高性能科学计算。相较主流 Intel CPU、NVIDIA GPU 芯片，在保持计算高精度前提下，实现了约 2 个数量级提速，并将计算功耗降低大约 3 个数量级。研究成果以"Accurate and efficient molecular dynamics based on machine learning and non von Neumann architecture"为题，发表在《npj Computational Materials》期刊。现在已经在 FPGA 这个技术上开展了概念测试，这只是花了几年时间，和较少的资源，已经可以明显看到它的效果，说明 AI for Science 领域专用芯片发展空间十分巨大。

b) 异构算力的统一调度平台

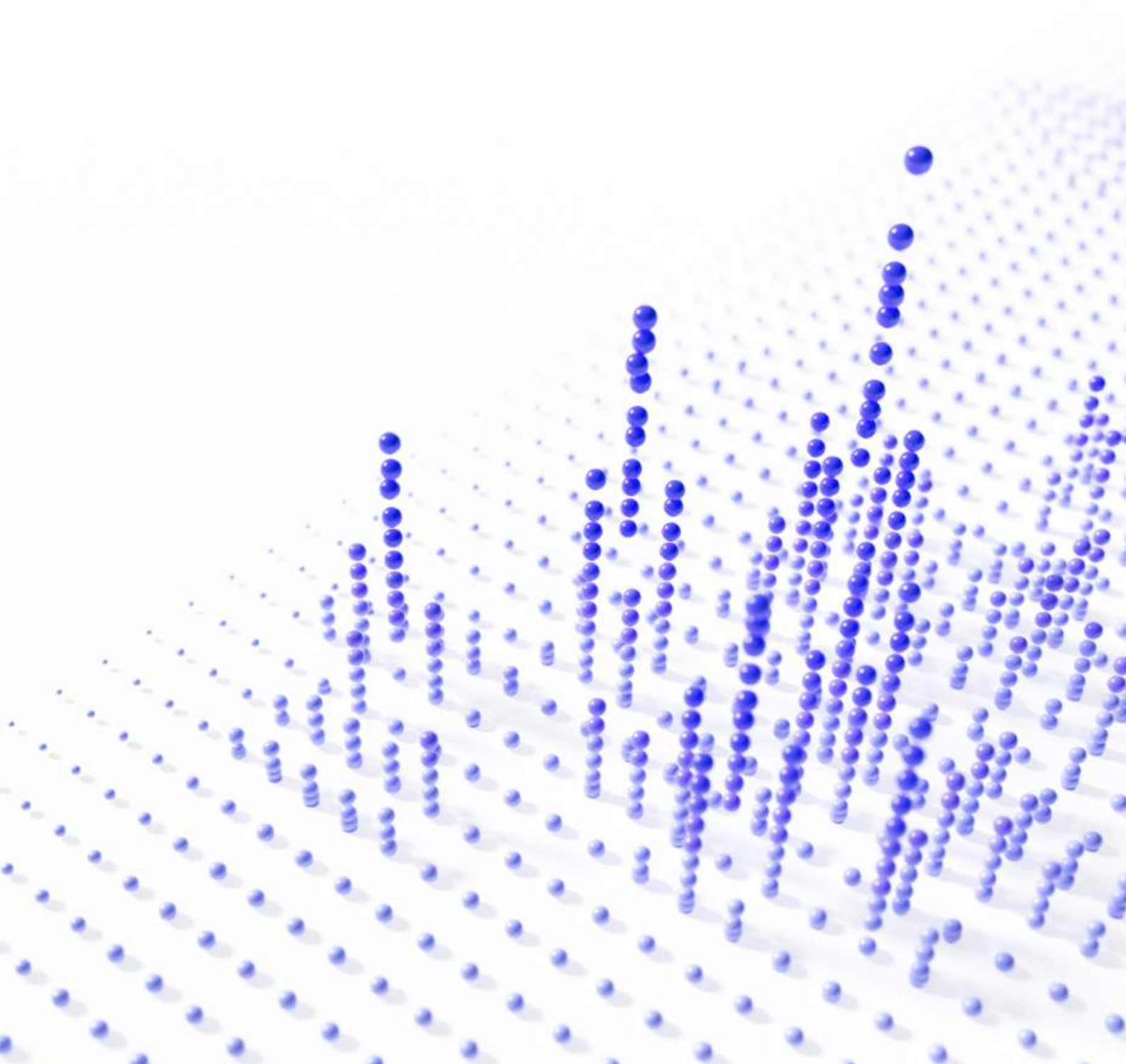
AI for Science 的蓬勃发展已经让我们看到了越来越多领域的关键性问题可以通过计算模拟的方式进行解决。同时，我们也看到各种计算硬件，如 CPU、GPU、FPGA 和 ASIC 等，为 AI for Science 提供了极大的计算能力。然而，针对 AI for Science 的不同计算需求，如何在多种算力资源之间高效切换和调度任务，实现计算资源的最大化利用，成为了一个值得关注的问题。

针对这一问题，构建一个针对各种异构算力的统一任务调度平台至关重要。这样的平台可以根据任务的计算特点和需求，在不同的计算硬件上完成计算，实现资源的高效利用。

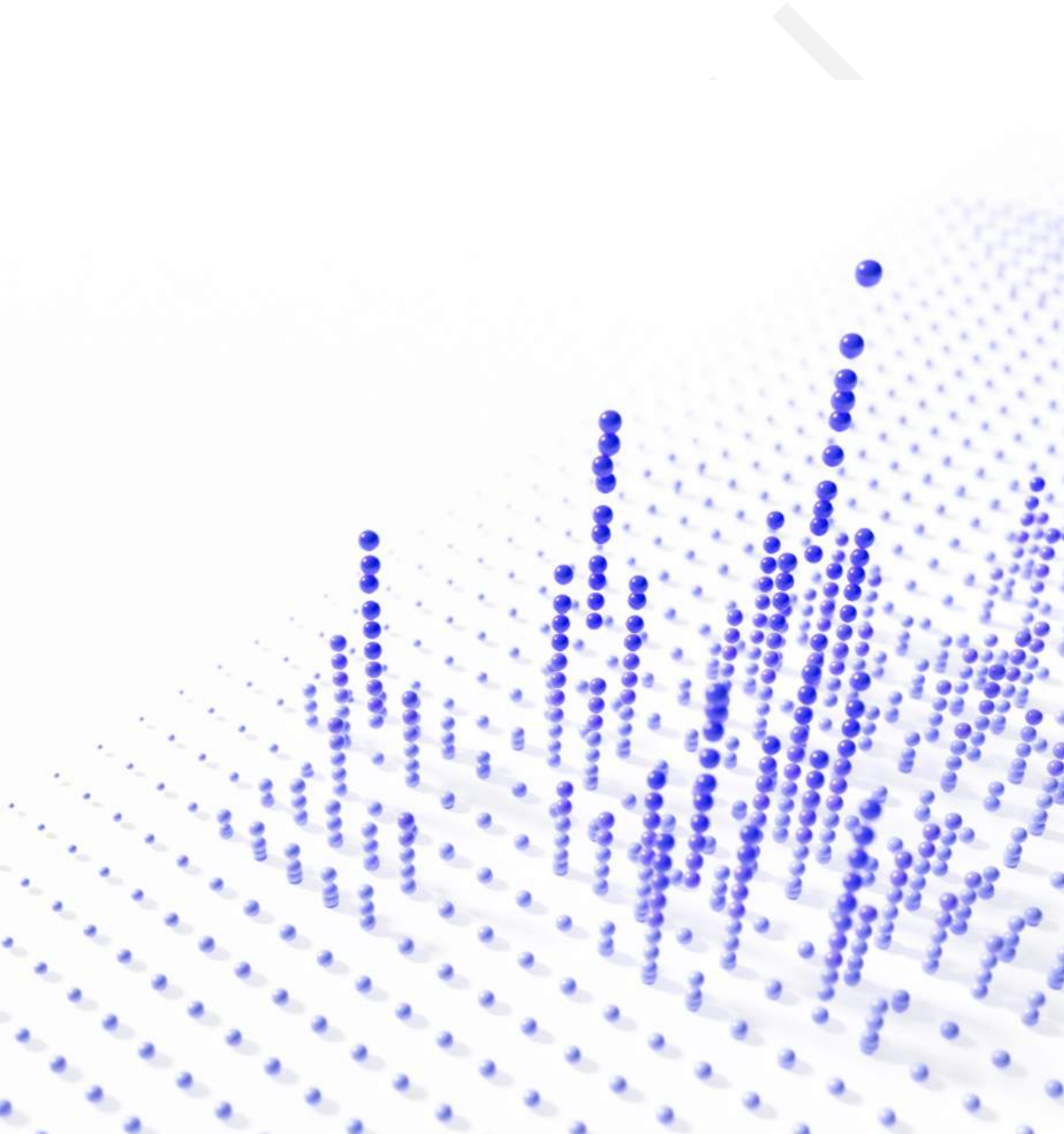
1. 满足多样化的计算需求：AI for Science 涵盖了多个领域，如生物学、物理学、化学、地球科学和天文学等。这些领域的研究任务往往具有多样化的计算需求，例如，需要大量并行计算的模拟任务，适合在 GPU 上运行；而对于高度优化的专用计算任务，则更适合使用 ASIC。有些任务需要跨多节点进行并行计算，因此对网络带宽要求较高，适合在有高速网络连接的环境如超级计算机环境下运行；有些任务适合只对同时并发的算力数有需求，而对网络没有要求，适合弹性资源丰富的云环境下运行。通过针对不同任务类型的在不同异构资源下进行灵活资源调度，一个方面可以降低功耗和成本，另一方面也能提高任务效率与资源利用率。

2. 实现灵活的资源分配：AI for Science 中的任务具有不同的优先级和紧迫性，如预测自然灾害的任务可能需要比其他任务更快地完成。统一任务调度平台可以根据任务的优先级和紧迫性，灵活地分配计算资源，确保关键任务能够得到优先处理。

PART II:
AI for Science 的产研实践



第二章：AI for Life Science 原理与实践



2.1 生命科学中的 AI4S

2.1.1 生命科学走入 AI 时代

1865 年奥地利帝国遗传学家格里哥·孟德尔提出孟德尔遗传定律，1928 年《基因论》发表，美国进化生物学家托马斯·亨特·摩尔根系统阐述遗传学的基本原理。从 1928 年格里菲思的肺炎双球菌转化实验，到 1952 年赫尔希和蔡斯的噬菌体侵染实验，历经 24 年的探索科学家们确认 DNA 为遗传物质。随后的十年中，奠定现代生物学基础的诸多重大发现喷涌而出：沃森和克里克提出了 DNA 的双螺旋结构模型（下图），梅瑟生-史达实验确认了 DNA 复制机制，弗朗西斯·克里克于 1958 年提出分子生物学的中心法则。随着人类对生命之谜的不断探索，对遗传密码的突破发现，生命科学迎来了分子生物学篇章。

图：DNA 的结构[1]

1958 年至今，科学家探索生命密码的脚步迈入测序时代，弗雷德里克·桑格于 60 年代发明了蛋白质氨基酸序列测定方法，又于 1977 年提出双脱氧链终止法，标志着第一代 DNA 测序技术的诞生，为生命科学研究的基因组时代打下基础。

1990 年，“人类基因组计划”正式启动，全球各地的科学家投入到这一浩大工程中，旨在测定人类染色体所包含的三十亿个碱基对的核苷酸序列、破译人类遗传信息。2003 年，这项被誉为生命科学领域的“阿波罗登月计划”终于在距 DNA 双螺旋结构提出半个世纪后宣告完成。

虽然人类全基因组测序完成，但测序技术的发展带来的范式革命才刚刚开始。

时至今日的 20 多年间，测序技术推进了基因组，转录组，蛋白质组等高通量数据的产生，大量基于“组学”的生命科学研究计划接棒人类基因组计划，进一步推进着人类对于生命奥秘的探索：1000 Genome 计划为基于人群的疾病关联性研究提供数据基础，了解不同种族之间遗传作用的区别 。GTEx 计划搜集了不同生物组织中遗传多样性和基因表达水平的信息 ，ENCODE 计划旨在通过收集多组学数据，研究不同水平上的基因调控机制。针对疾病，美国也开启了 TCGA 和 Moonshot 等国家计划，为全面了解癌症的分子生物学机制提供充分的数据。

从 1958 年到今天，我们对中心法则的理解不断深化、细化，多组学方法不断丰富着我们对 DNA->RNA->蛋白质中复杂调控过程的理解。随着高通量测序能力的指数增长，生物学正在以前所未有的速度产生海量数据，基因组学、生物信息学等结合了生命科学和信息科学的交叉学科应运而生。信息基础设施的不断建设使得数据的获得、处理、储存和分享变得越发完备。与此同时，测序成本以突破摩尔定律的速度减低，高通量组学技术逐渐成为生命科学研究的新基石。

传统的单位点、单基因的研究方法逐渐被替代，高通量测序技术推进生命科学研究进入“组学时代”。

就像贺福初院士在"大发现时代的'生命组学'"一文中说的："组学的发展引领了 20 世纪末至今的生命科学大发现"，一种在全基因组尺度开展的，多组学大数据驱动的研究图景正在展开[3]。

伴随着组学时代的来临和生命科学数据爆炸而日益凸显的，是传统研究方法的不足。

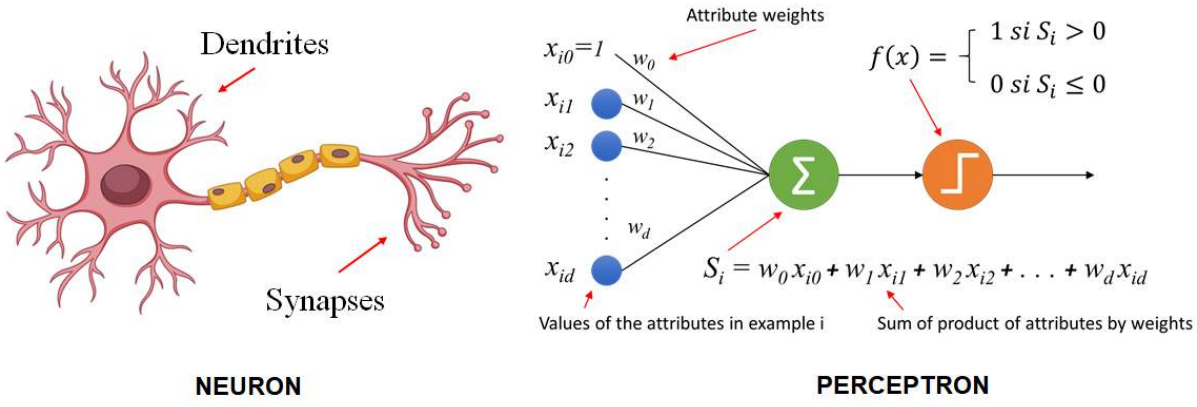
低通量的方式无法最大化挖掘跨尺度、多模态的信息，人类对于复杂生物系统的了解程度可能仍远小于我们对于宇宙终极规律的认知。分子生物学数据中存在的低信噪比问题也困扰着很多科研者。

如何从大量的信息中挖掘出有意义的规律，并提出可验证的新的猜想，成为新时代生物学研究的一个重要课题。

而 AI 似乎正是最合适的方式。1957 年美国神经学家弗兰克·罗森布拉特提出由生物神经细胞启发而设计的感知机（perceptron）（下图），为神经网络模型和现代深度学习算法打下基础，生物学仿佛现代 AI 的"启蒙老师"。不久前，AlphaFold2 在 CASP 比赛拔得头筹，宣布突破困扰了生物学家 50 年之久的蛋白质折叠问题，AI 仿佛成为了生物学问题的"最优解法"。半个世纪以来，AI 的发展与生物学研究似乎一直有着千丝万缕的联系。2021 年，DeepMind 创始人 Demis Hassabis 成立并出任 isomorphic labs CEO，在公司愿景中，他这样说："如果说数学是物理的语言，那么生物可能是 AI 语言最完美的描述对象" [6]。

在过去的十几年中，大量基于机器学习和深度学习的理解基因调控的方法被开发出来，AI 也在逐渐驱动药物研发的和个性化医疗中的新突破。

人工智能正在生物学机制研究、疾病的筛查、检测和治疗中发挥着越来越重要的作用

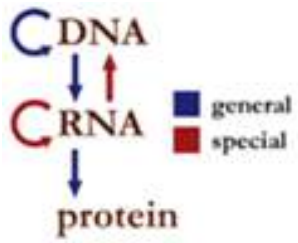


图：神经元细胞启发的感知机 [5]

2.1.2 AI4S 推动生物机理的探索

战胜疾病关键在于对生命机制规律的了解，分子生物学研究的中心为探索从基因到表型中关键过程与机制。基因是人类生存的密码，基因组计划为人类绘出基因图谱。从 DNA 双螺旋结构的提出，人们开启了对遗传物质的定量研究，使遗传的研究深入到分子层次。除了朊病毒以外，目前的生物都遵守中心法则。

中心法则的第一个元素是 DNA，生命信息的传递由 DNA 开始（下图）。生物所携带的遗传信息以 DNA 的形式被储存在细胞内，DNA 分子中有功能的单位被称为基因。DNA 序列指四种不同含氮碱基 A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)、G(鸟嘌呤）的线性排列顺序，这一排列顺序对生命过程十分重要。分子生物学的中心法则即是描述 DNA 中包含的信息如何传递进而指导蛋白质的合成。



图：中心法则[7]

人体内 2%左右的 DNA 序列可编码蛋白质，在制造蛋白质的过程中，这些编码区域的 DNA 以三个单核苷酸为单位形成一组密码子，每个密码子对应一个氨基酸或终止信号。通过密码子，蛋白编码区域的 DNA 序列中所包含的遗传信息得以发挥功能。

人类基因组计划的推动，帮助人类绘制了一幅人类基因地图，测定了组成人类染色体（单倍体）所包含的 30 亿个碱基对的核苷酸序列，除此之外还对大肠杆菌、酵母、拟南芥、线虫、果蝇和小鼠等进行基因组测序，不同生物演化阶段的物种测序可帮助认识人类的基因结构与功能，对人类在基因上的探索是里程碑的事件。

但仅仅知道 DNA 序列是不够的，我们对于基因到表型的生物机制了解还不够清晰。

储存在 DNA 顺序中遗传信息经过转录和翻译，转变成具有生物活性的蛋白质分子，这一过程被称为基因的表达。基因的表达水平，指的是基因表达量的高低。

为什么不同种类的细胞拥有完全相同的 DNA 序列和基因组成，然而却体现出完全不同的性状和功能？一个 DNA 位点的突变如何导致不可控的癌症？这些都与基因的表达水平有关。

基因表达水平的调控，指的是细胞通过一系列机制，使基因组的表达水平在时间和空间上处于特定状态的复杂过程。了解基因表达的调控机制，是解开基因型到表型之谜的重要一环。 探究这些科学家尚未能完全回答的问题能帮助我们更好了解疾病发生的原因，优先选择潜在的可针对的靶标，进而提出预防、检测、和治疗疾病的方法。

基因的表达水平与基因中的序列同样重要。基因的正常表达维系着人体的健康。

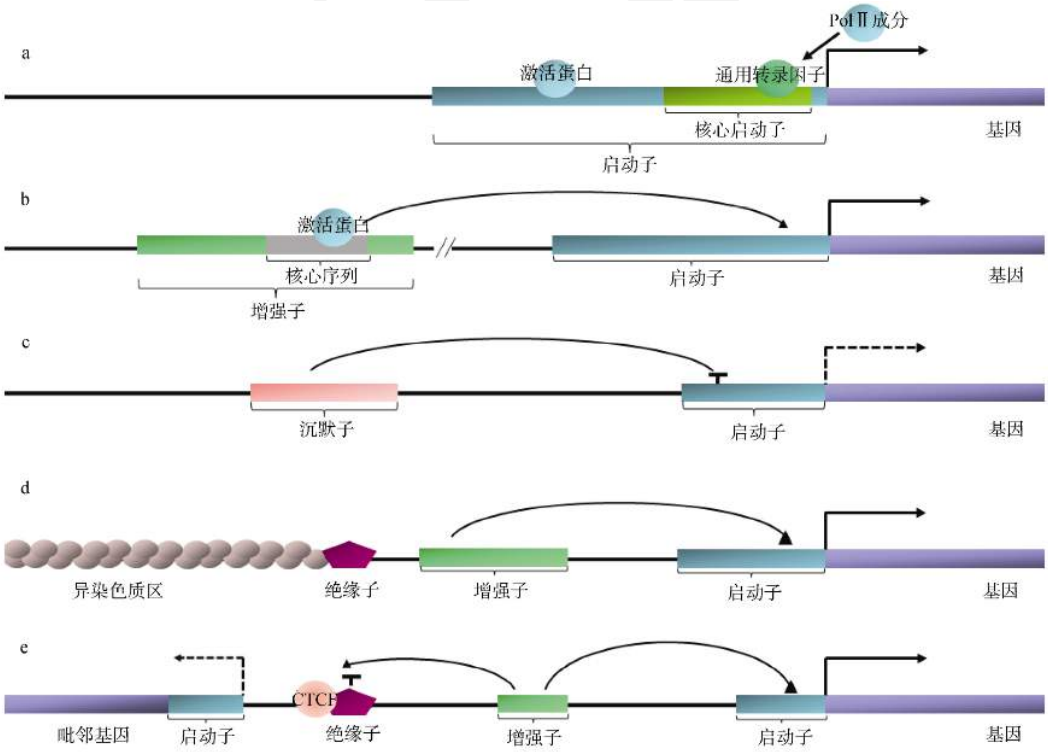
科学家不断发现越来越多的重大疾病是由于某些基因的表达水平异常而造成的。人体内几乎所有的细胞都拥有相同的基因组，决定了它们的差异的，是它们对基因的差异化表达。在不同种类细胞中，不同的基因会被表达，基因表达的水平也有所差异。

Dennis Slamon 博士在 1987 年在对乳腺癌的研究中发现，189 名患者中，30%的肿瘤存在 HER2 基因的扩增或过量表达，是健康人的 2 到 20 倍[8]。

基因表达水平的调控机制仍是科学家们研究的重点。人体内约 98%不参与编码蛋白质的碱基序列，在调控过程中发挥着重要作用。据 DNA 元件百科全书 (Encyclopedia of DNA Elements, ENCODE)研究计划总结，基因组中约 80%的序列都具有功能性意义[9]。比如这些序列可以作为顺式调控元件 Cis-Regulatory Elements（CRE）（如启动子、增强子、沉默子、绝缘子等），通过与其他蛋白或染色质间相互作用等方式，实现对临近基因表达水平的调控作用（下图）。

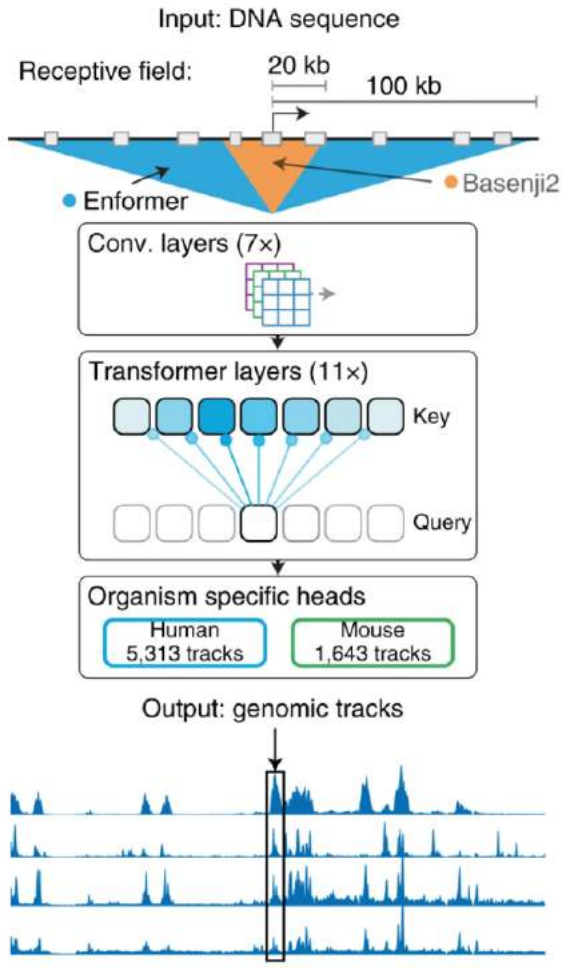
在经典的序列-功能对应研究中，科学家们通常需要对基因上游的临近序列进行 DNA 片段切除或诱发位点突变，再观察基因表达水平的变化，从而对序列功能进行推断和注释。然而许多重要调控元件，可以对距离在 20k 碱基对之外的基因产生影响。如何高效全面地整合及解释包含在如此长区段内的 DNA 信息，给传统的实验和计算方法带来了巨大的挑战。

随着深度学习的发展，尤其是自然语言处理模型的不断完善，AI 逐渐成为基于 DNA 序列建模的一件利器。



图：非编码 DNA 序列的功能[10]

2021 年，Calico Life Sciences 公司联合 DeepMind 团队，开发了可整合远至 100k 碱基范围序列进行基因表达和染色质状态预测的深度学习模型 Enformer，发表于 Nature Methods[11]。该模型仅利用 DNA 序列作为输入，通过结合远端序列相互作用的信息，极大的提高了从序列到基因表达量的预测精度，并且能准确预测单位点突变带来的基因表达后果，为疾病-位点映射研究带来了新的可能。（下图）

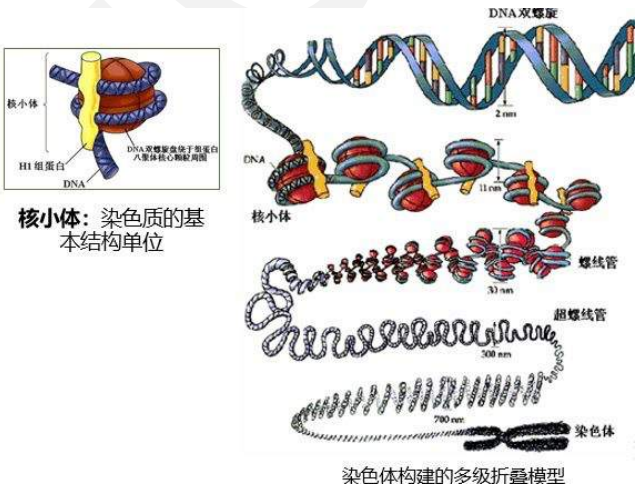


图： Enformer 模型结合长距离 DNA 序列，对基因表达水平进行准确预测[11]

除 DNA 序列外，DNA 在细胞核内的组织方式及折叠状态也影响着基因的表达。

人类的 DNA 长达约两米，被储存在直径仅为约 5 微米的细胞核中，为了确保基因调控的精确性，DNA 并非像一团无序的毛线球一样存在于细胞核中，而是通过精密的组织方式压缩于微小的空间内。DNA 缠绕在组蛋白上形成核小体，核小体的密集及疏离一定程度上决定了核小体间 DNA 上的基因是否可被读取。

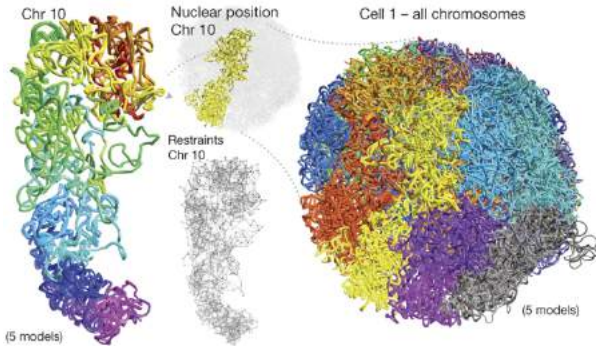
核小体的有序排列形成染色质，染色质进一步折叠成为染色体。我们所熟知的"X 形"染色体，便是染色质多级折叠后所呈现的状态（下图）。



图：染色体构建的多级折叠模型

DNA 通常不是以"X 形"存在于细胞核中，而是会以多种层级折叠方式混合的形态存在。近年来科学家们发现，DNA 的折叠方式对基因的表达也发挥重要的作用：因为特定的折叠方式，有些基因不可读取，有些基因则可以被其他蛋白机制结合，从而进行基因表达。

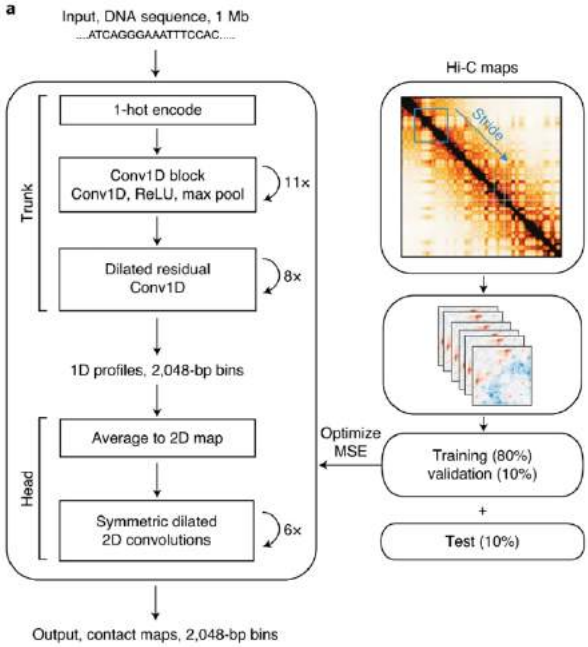
科学家们结合 Hi-C 这一组学技术，于 2017 年在 Nature 上发表了通过计算模拟得到的哺乳动物单个细胞基因组的物理结构，展示了 DNA 在细胞中的包装方式。（下图）。



图：计算模拟所得哺乳动物单个细胞基因组的物理结构[12]

来自 Los Alamos 国家实验室 Sanbonmatsu 团队在 Trinity 超级计算机上进行了突破性模拟，对 DNA 基因完成了 10 亿原子的建模。这进一步推动了人类对于染色质折叠的了解。

AI 也被应用于 DNA 折叠的预测，2020 年，Katherine Pollard 研究组于 Nature Methods 上发表基于卷积神经网络模型的 Akita（下图），Akita 使用 DNA 序列为输入，准确地预测了基因组的三维空间结构，并且学习到了对 3D 基因组折叠至关重要的 CTCF 蛋白与 DNA 结合位点方向性的重要性。



图：Akita 模型[13]

表 2：AI4S 在多组学中的应用

AI4S 对科学的助力不仅局限于上述几个问题中，新的研究范式体现在现代生物学研究的方方面面。随着生命科学来到多组学和数据加速猜想的时代，AI 被广泛运用在基因组学的所有环节。表 2 沿着中心法则的顺序，对多种组学中的 AI 应用场景进行了总结，供有兴趣的读者深入了解。[14]

研究对象	组学类型	组学技术	AI4S 实践
DNA 序列	基因组学	鸟枪法测序，二代测序，三代测序	• 变异位点检测（如 DeepVariant[15]） • DNA 序列功能注释（如 Basset[16]） • 从 DNA 序列预测特定性状（如 DeepBind[17], Enformer[11] ）
DNA 的化学修饰、DNA 开放程度、组蛋白的分布和化学修饰等	表观基因组学	BS-Seq，DNase-Seq，MNase-Seq，ATAC-Seq，ChIP-Seq 等	• 表观遗传标记预测及染色质状态判断（如 ChromeInpute[18]） • 数据降噪 • 基因组三维空间结构预测（如 Akita [13]）
RNA 及 RNA 的转录后修饰	转录组及表观转录组学	Microarray，RNA-Seq，scRNA-Seq，GRO-Seq 等	• 编码蛋白及非编码蛋白转录本分类 • 基因表达量分析 • RNA 预训练大模型（如 Uni-RNA）
蛋白质	蛋白质组学	MS,MS/MS, LC-MS,MALDI-TOF 等	• 蛋白设计（如 RFDiffusion） • 蛋白质结构预测（如 AlphaFold2） • 蛋白功能预测（如 DeepFunc[19]） • 蛋白-蛋白相互作用 • 蛋白药物相互作用
代谢通路及代谢产物	代谢组学	MS, LC-MS 等	• 代谢产物分析

2.1.3 基因+AI4S 在靶标发现和精准医疗中的利用

I. 基因组学数据驱动的靶标发现

靶标发现和筛选是药物开发的第一步，AI 可以从三个层面对基因组学驱动的靶标的发现过程提供助力：

1. 基因调控的原理发掘，通过对大高通量数据的分析，辅助扩展已知的分子通路和应答机制，或者发现潜在的全新分子通路，应答机制。
2. 通过将病人和健康人的多组学数据进行对比，从中挖掘基因的差异表达或异常调节通路，探寻基因与疾病的相关性。
3. 利用自然语言处理(NLP)技术，对于已知疾病机制进行深度挖掘，学习和分析科研文献、专利和临床报告等文档，挖掘疾病与靶点的相互联系，发现靶标相关的更多疾病，另一方面也可以辅助发掘潜在"老药新用"的机会。

近年来基于 AI 技术进行靶标发现的工作逐渐发生商业转化：Benevolent AI 公司在 2020 年发表了他们所建立的靶点发现 AI 模型 Rosaline，通过整合临床数据，基因研究，生化实验数据，科研文献和化合物库挖掘，加速靶标的发现。In silico 的 AI 新药靶点发现平台 PandaOmics，旨在结合 AI 赋能的多组学数据分析及基于自然语言处理的文献挖掘，寻找潜在靶点。

Exscientia 公司的 Centaur Biologist 平台，可将文献提取数据以及结构化数据源（基因组学、转录谱和路径数据库）相结合，构建知识图谱，将靶点和疾病进行匹配。并且还有人工智能预测模型可对药物分子的药理毒性、毒理学和药代动力学进行预测，同时还可对人体活体组织识别预测患者临床有效性和预后情况。

II. 精准医疗、基因治疗和核酸药物

随着人类基因组计划的完成，科学家们对于疾病与基因关系的理解更加深入，AI 也被应用于精准医疗和基因治疗的各个环节中。

精准医疗，或个性化医疗，是一种新兴的医疗概念，指的是根据患者个体的基因组学信息，结合遗传学和生活方式等因素，定制针对性的精确诊断、用药及医疗方案。测序成本平民化使多组学数据在很多场景下开始被作为辅助医疗检测指标，丰富的数据也催生了 AI 助力的个性化诊疗。

越来越多的地方用到了基因筛查来进行疾病早筛，期望能够在早期就检测出疾病进行介入治疗；或者提前预防，最大可能性地降低此遗传病的发生率。基因检测和 AI 辅助筛查目前最常应用于产前检测、遗传风险检测和癌症早筛等场景。

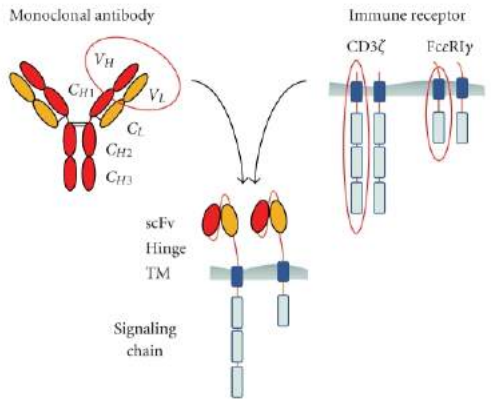
基因检测领域最著名的事件是好莱坞影星安吉丽娜·茱莉，由于有乳腺癌家族遗传史，她在做基因筛查时，发现乳腺癌相关基因 BRCA1 发生了突变，这有可能会引起乳腺癌的发生，她决定采取双侧乳房切除，来预防乳腺癌的发生。

在筛查之外，AI 也在临床上也有广泛应用。计算机视觉算法可以辅助肿瘤组织病理分型，哈佛大学 Wyss 研究所和麻省理工学院的研究院开发出的一种对黑色素瘤的辅助分析算法可自动识别可疑色素病变（SPL），其诊断结果与 3 位皮肤科医生共识一致性高达 88%[21]。AI 也可辅助临床治疗时靶向药物的选择，上海交通大学附属胸科医院肿瘤科陆舜教授团队开发了一种深度学习算法，可用于预测晚期非小细胞肺癌单药 PD-1/PD-L1 免疫治疗的临床获益情况。[23]

另一个与基因紧密相关的医学突破是基因治疗。基因治疗或基因疗法的基本原理是将外源基因导入靶细胞或组织，替代、补偿、阻断、修正特定基因，以达到治疗和预防疾病的目的。基因药物在治疗遗传病、癌症、糖尿病，预防传染病等方面正不断取得突破性进展。

基因治疗可以大致分为基于 DNA 与基于 RNA 两大类。基于 DNA 的方法包含离体基因治疗，如 CAR-T 细胞、使用 CRISPR 技术对离体细胞进行基因编辑，和在体基因治疗，如利用质粒和病毒等载体进行基因片段递送。基于 RNA 的方法包含使用反义寡核苷酸（ASO）、RNA 干扰（使用小核苷酸如 siRNA、miRNA）和 mRNA 的药物等。

在 CAR-T 相关疗法中，CAR 分子的胞外结构域（下图）中识别抗原的单链抗体片段 (scFv)十分重要，AI 的技术可被应用于学习抗体片段规律，对抗体亲和力或人源化性质进行预测和推荐。



图：CAR-T 的结构[25]

AI 也被应用于优化 CRISPR 基因编辑技术。Cas9 是 CRISPR-Cas9 技术的重要组成部分，AI 算法可用于寻找毒性更弱的 Cas9 酶[29]。同时，我们可以借助 AI 在酶设计中的应用，尝试对已知的 Cas9 酶进行优化和改造，这个我们在第二章也会再次详述。

载体的设计和使用是在基因药物设计中的重要环节，AI 可辅助载体的设计和筛选。

Dyno therapeutics 利用机器学习算法和生物学实验的结合，在对 AAV 载体的设计过程中，根据 5 大重要参数（递送效率、靶标的特异性，免疫系统反应，包装大小和制造特性）以及实验数据来训练预测模型，以期达到对不同序列的预测对应的参数的效果，最终期望达到根据参数来理性设计衣壳序列的愿景[27]，而剂泰则主要利 AI 来帮助对 LNP 载体设计的筛选中，减少这个过程中湿实验的时间和成本投入[28]。

另一个不可忽视的领域是核酸药物，尤其是 RNA 药物。RNA 药物在过去的几十年中逐步走向临床，mRNA 疫苗更是在对抗 COVID-19 新型冠状病毒疫情中发挥了至关重要的作用。AI 被广泛运用在 RNA 研究中，如斯坦福大学研究团队开发的几何深度学习算法可用于 RNA 的构象预测[31]。

香港中文大学团队基于蛋白表面结构与局部特征进行了高精度 RNA-蛋白质结合偏好预测[32]。这些研究方向都是 RNA 药物的成药性的关键决定因素。

Source:

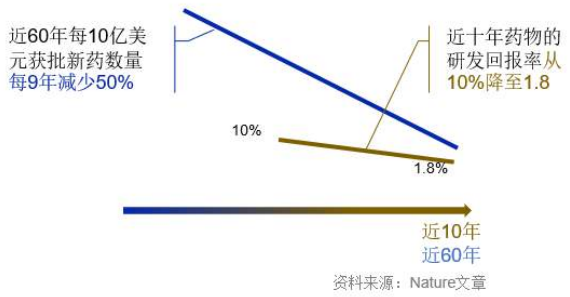
1. Andrey Prokhorov/E+/Getty Images，DNA Definition: Shape, Replication, and Mutation
2. National Human Genome Research Institute (NHGRI)， https://www.flickr.com/people/142595545@N03
3. 光明日报，贺福初：大发现时代的"生命组学"
4. National Human Genome Research Institute (NHGRI), DNA Sequencing Costs: Data
5. Deep neural networks,or Perceptron vs dogs and cats
6. Isomorphic Labs announces first phase of management team
7. 中心法则图片引用
8. Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science (New York, N.Y.), 235(4785), 177–182. https://doi.org/10.1126/science.3798106
9. Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J. W., Tanaka, F. Y., Adenekan, P., O'Neill, E., … Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic acids research, 48(D1), D882–D889.
10. 秦丹， 徐存拴（2013 年）非编码 DNA 序列的功能及其鉴定，遗传 HEREDITAS，DOI: 10.3724/SP.J.1005.2013.01253
11. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods, 18(10), 1196–1203.
12. Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M., Lehner, B., Di Croce, L., Wutz, A., … Laue, E. D. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature, 544(7648), 59–64.
13. Fudenberg, G., Kelley, D. R., & Pollard, K. S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nature methods, 17(11), 1111–1117.
14. Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A., & Colombo, T. (2021). AI applications in functional genomics. Computational and structural biotechnology journal, 19, 5762–5790.
15. Poplin, R., Chang, P. C., Alexander, D., (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature biotechnology, 36(10), 983–987.
16. Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990–999. https://doi.org/10.1101/gr.200535.115
17. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831–838.
18. Ernst, J., & Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature biotechnology, 33(4), 364–376. https://doi.org/10.1038/nbt.3157
19. Zhang, F., Song, H., Zeng, M., Li, Y., Kurgan, L., & Li, M. (2019). DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions. Proteomics, 19(12), e1900019.
20. 在线 TCGA 数据库挖掘网站汇总
21. 医药魔方，准确率近 90%，更早捕获皮肤癌 哈佛新型 AI 系统成功辅助诊断
22. Chiu Y C, Zheng S, Wang L J, et al. Predicting and characterizing a cancer dependency map of tumors with deep learning[J]. Science Advances, 2021, 7(34): eabh1275.

23. Yang,Y., Yang, J., Shen, L., et al. A multi-omics-based serial deep learningapproach to predict clinical outcomes of single-agent anti-PD-1/PD-L1immunotherapy in advanced stage non-small-cell lung cancer. Am J Transl Res. ;13(2): 743–756.
24. web of science，中银证券，新药研发（AIDD）行业系列报告—洞鉴行业发展，把握投资先机：（一） AIDD 概览篇
25. Cartellieri M, Bachmann M, Feldmann A, et al. Chimeric antigen receptor-engineered T cells for immunotherapy of cancer[J]. Journal of Biomedicine and Biotechnology, 2010, 2010.
26. 高中生物必修 2 知识点总结:基因突变及其他变异。
27. AI 第一视角，药物递送，人工智能（AI）布局药物研发的新赛道_Dyno
28. 金融界，顶级机构重押，AI 破局药物递送的新尝试
29. Choi GCG, Zhou P, Yuen CTL., Combinatorial mutagenesis en masse optimizes the genome editing activities of SpCas9. Nat Methods. 2019 Aug;16(8):722-730.
30. 知乎专栏：揭因解遗：揭基因神秘面纱，解遗传繁杂密码。焦老师讲遗传 2：染色体组成与命名。
https://zhuanlan.zhihu.com/p/118349815
31. [Townshend, R., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., & Dror, R. O. (2021). Geometric deep learning of RNA structure. Science (New York, N.Y.), 373(6558), 1047–1051. https://doi.org/10.1126/science.abe5650],
32. Wei, J., Chen, S., Zong, L., Gao, X., & Li, Y. (2022). Protein-RNA interaction prediction with deep learning: structure matters. Briefings in bioinformatics, 23(1), bbab540. https://doi.org/10.1093/bib/bbab540

2.2 AI4S 驱动的药物研发

2.2.1 药物研发的现状与挑战

人类生命长度和质量的提升离不开新药的研发，但新药研发的投入产出比近年来呈现出越来越低的趋势。自 1950 年以来，每 10 亿美元研发投入所产出的获批新药数量几乎每 9 年减少一半。对新药的研发而言，投入越来越高，回报率越来越低 （下图）。如何打破创新药研发的"反摩尔定律"成为当下行业关注的核心问题之一。



图：药物研发回报率越来越低[2]

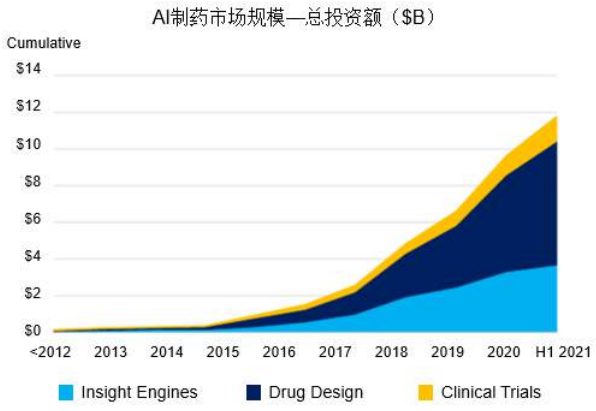
目前药物研发的主流范式，是基于靶标与疾病的关系，开发能够干预靶标生物功能的药物分子。

药物研发的成功率低，主要来自于靶标与疾病的关系探索难度大，靶标成药难度高，和药物分子的开发难度大等几个方面。

生命科学众多悬而未决的问题进展缓慢；更关键的是，生命科学领域的基本原理尚未系统性建构，"新通路"和"新靶标"的发现仍具有极大的经验性和偶然性。此外，对于已经被证明与疾病的发生与治疗有强关联的重要靶标，依靠过去的药物研发认知和技术手段，其中的"低垂之果"也几乎被采摘殆尽；高价值的靶标哪怕仅仅只是获得初步验证或突破，就有大量跟风者涌入，导致了当前在研药物的靶点高度同质化的局面。

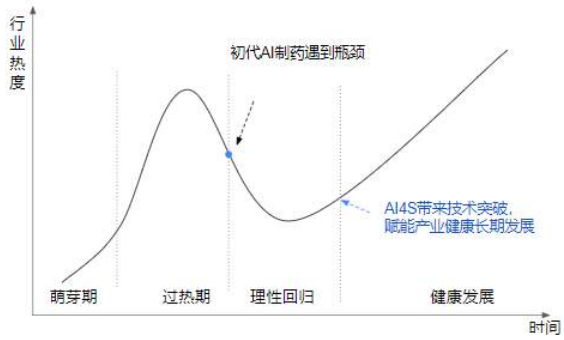
按照不同的统计口径，潜在有望成为药物靶标的蛋白质中，被开发出药物的仅占 5%-20%，其余的 80%-95%均为"难成药"靶标，这类靶标的成药性问题也成为了行业发展的重要瓶颈。虽然科学家们为了攻克难成药靶标，在研发手段和药物形态等方面不断取得突破，比如蛋白降解技术之于需要开发抑制剂的难成药靶标，但每个特定药物形态的药物分子的成药性难题依旧客观存在。

正是由于上述问题的严峻性，在过去几年中，人们把极高的热情和期望投入到"AI+药物研发"中来（下图），国内外头部的药厂也纷纷与"AI 制药"公司合作，并在生命机制研究-药物发现-临床前研究-临床研究-审批上市一整条链中的多个细分环节，诞生了一系列 AI 的不同程度的应用。



图：AI 在药物研发中受到资本追捧[3]

初代 AI 制药已经经历了资本和行业的过热期，正逐渐回归理性。过去的 AI 制药仍面临一些技术瓶颈，尤其是纯粹数据驱动的 AI 在药物研发中的准确性较低。将 AI 与底层生物机理结合的新范式 —— AI for Life Science，正在从底层技术的突破为整个行业注入崭新的活力，其系统性的发展有机会为行业带来新的机会（右图）。

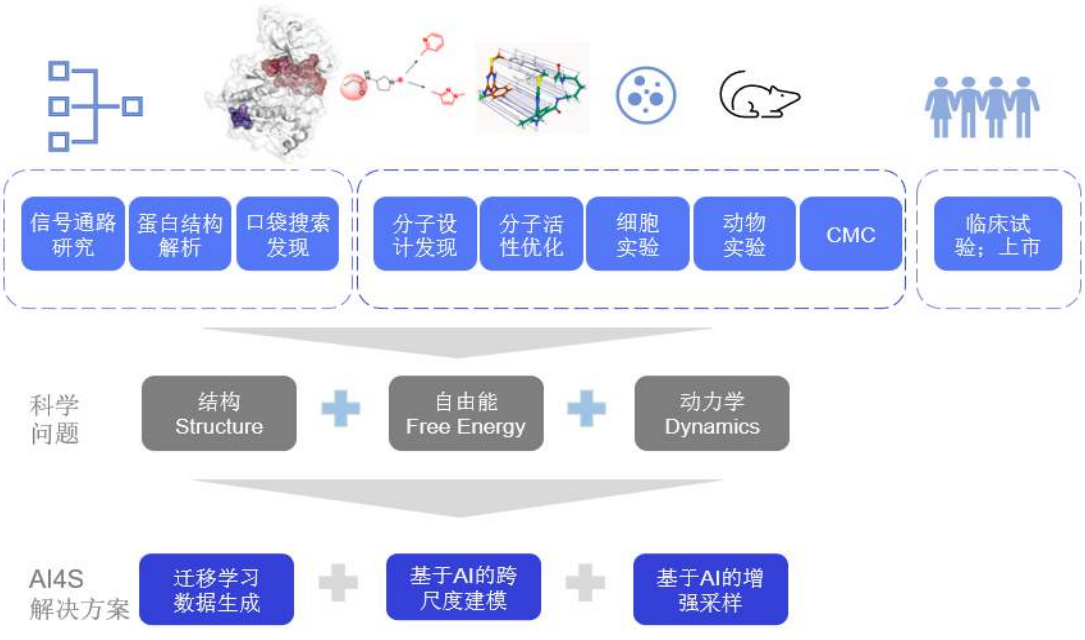


图：AI 制药行业的技术发展周期，AI4S 可能带来新的突破当前药物研发的流程和挑站

药物研发是个多环节、漫长且昂贵的流程，每一环境的效率提高都有巨大的商业价值。

简化来看，整个流程从前至后大致会涉及到早期生物学研究、药物的发现与优化、临床前验证以及临床研究。（下图）

- 早期生物学研究的阶段，主要目标是通过对疾病信号转导通路的研究，建立疾病与靶标的相关性，明确可供开发的蛋白质靶标、相关的评估模型和生物标志物，作为后续药学验证的基础；
- 蛋白质结构解析和微观机理研究环节，主要目标是获得可靠的蛋白质三维结构，并更进一步地，通过对序列-结构-动力学的进一步分析，探究蛋白质的微观性质，以及微观性质跟序列-结构-动力学三者的相关关系，为后续的药物开发提供基础；
- 药物作用位点探索，主要目标是找到能影响靶蛋白在与疾病相关的信号转导通路中发挥功能的、且适合药物分子结合的关键位点，基于此，我们才能根据位点的微观环境以及靶标蛋白的整体性质进行药物设计；
- 苗头化合物发现的阶段，核心目标是在前期建立的药效学评价中，找到能靶标蛋白生物活性的化合物，在此基础上通过系列的评估与优化，确定进一步推进的先导化合物；
- 先导化合物优化阶段，核心目标是对化合物从药效学、理化性质与药代动力学性质、毒理学



图：药物研发的主要环节

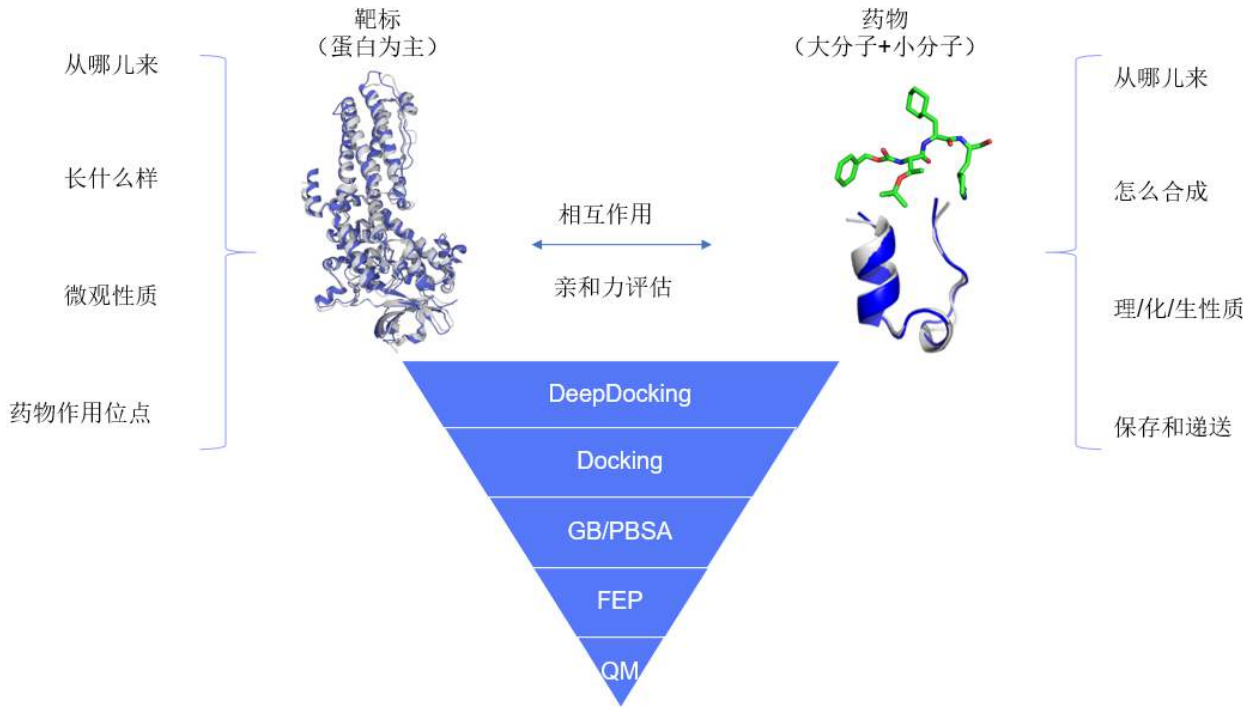
性质等角度进行多参数优化，并基于此进行取舍和平衡，从而确定值得继续推进的 PCC；

- 临床前验证阶段，核心目标是获得符合监管和临床注册需求的评价数据，使得药物可以被顺利推进上临床，因此，在 PCC 确定前后，会考虑药物制备工艺的优化与杂质含量的控制、晶型和剂型的确证及针对不同动物种属的药效学、药代动力学和毒理性评价等。

整体来看，药物研发的主要环节和关键问题可以归纳为靶标，药物分子侧，以及二者间相互作用力三个方面的问题（下图）

- 靶标方面的问题：包括了药物靶标从哪来，长什么样，微观性质如何，药物的作用位点在哪；
- 药物方面的问题：包括了药物从哪来，如何合成，药物的理化生性质是怎样的，药物的保存和递送应该如何设置。

联系两个方面之间的最主要问题是要评估药物靶点和药物之间的相互作用力，也就是亲和力的计算。



图：药物发现的主要环节与关键问题 (credit, DPTechnology)

2.2.2 AI4S 药物研发新范式

表 3：药物研发流程中的各步骤的挑战和 AI4S 范式

研发环节		研发目标	科学难点	AI4S 实践
早期生物学研究与评价体系建立		建立疾病与靶标的相关性，明确可供开发的蛋白质靶标、相关的评估模型和生物标志物	实验手段时间长、成本高；实验设计需要技巧；现有知识对生物基本原理缺乏系统认识，寻找靶标和生物标志物一定程度依赖运气；	• 利用组学数据，在 AI 辅助下寻找规律，初步筛选出潜在可选靶标和生物标志物；
蛋白质结构解析与可药性位点确证	靶蛋白建模与解析	获得可靠的蛋白质三维结构或解析其动态行为	实验方法成本高、周期长，部分蛋白难以纯化和结晶，需要摸索实验条件；	• AI 蛋白质折叠、电镜图分类、基于电镜密度图的分辨率优化；
	结合位点确证	找到能影响靶蛋白在与疾病相关的信号转导通路中发挥功能的、且适合药物分子结合的关键位点	实验方法成本高、周期长； 突变实验盲目性大，结构生物学手段需要面对蛋白纯化和结晶等问题；	• AI 学习已知蛋白结合位点特征，对新的靶蛋白的潜在位点进行预测； • 通过 AI 提高分子动力学采样效率（寻找蛋白质结合位点），以缩小实验验证范围并降低盲目性；
候选化合物的发现与优化	苗头化合物发现	找到能影响靶标蛋白生物活性的化合物，在此基础上通过系列的评估与优化，确定进一步推进的先导化合物	实验：成本较高，筛选需要有实体分子，筛选范围受限于实体库； 基于结构的策略：打分函数不准，对接程序计算效率无法满足超大规模库筛选需求； 基于已有配体的策略：分子表示不合理，搜索容易自带偏好性；	• 利用 AI 技术，将主动学习、底层程序高性能优化与传统对接软件结合，加快计算筛选药物的速度； • 基于 AI 预训练模型提高对分子的表示能力，改进搜索效果，最终富集阳性分子供实验验证；
	活性优化	优化分子的亲和力	实验手段成本高、周期长；传统药物化学方法具备一定的盲目性且过度依赖人为经验； 经典打分函数精度不足，无法普适性地产生指导性结论；	• 利用自由能微扰等方法，结合 AI 优化其力场或电荷表示方式以优化精度； • 结合 AI 主动学习策略提升计算通量；

研发环节		研发目标	科学难点	AI4S 实践
	成药性优化	在分子的药效、理化性质与药代动力学性质、毒理学性质中找到平衡，确定 PCC	分子设计-验证的实验流程成本搞、周期长、依赖人为经验且存在一定盲目性； 计算手段依赖于大量数据，整体预测效果不佳；	• 利用 AI 预训练模型提高对分子的表示能力和基于小样本的学习能力，实现更好的预测效果； • 结合自由能微扰（FEP）技术，找到亲和力没有显著变化的情况下成药性得到优化的分子；
临床前验证	工艺优化	优化制备工艺、控制杂质含量以满足监管和注册的要求；	依赖丰富的人为经验	• AI 合成路线设计和优化；
	晶型确证	确证分子晶型以满足成药性要求和监管要求；	依赖于人经验实验成本高，耗时长	• 第一性原理晶型预测
	剂型确证	确证分子剂型以满足成药性要求和监管要求；	依赖于人经验实验成本高，耗时长；	• AI+分子动力学模拟预测稳定剂型
	多种属验证评价	在不同动物种属上做药效学、药代动力学和毒理学评价以满足监管和注册的要求；	实验成本高，耗时长	• AI 预测药物在人体可能的药效学、药代动力学和毒理学等，乃至预测临床实验的首次给药剂量，最大化发挥多种属验证评价获得的宝贵实验数据的价值；
临床阶段上市阶段		设计临床方案和入组标准，基于人体实验确证药物的安全性、有效性、剂量和周期等信息；	工作量大、耗时长	• 借助 AI 对病人筛选招募，处理临床数据和方案设计，流程管理

I. 物理模型驱动的药物研发

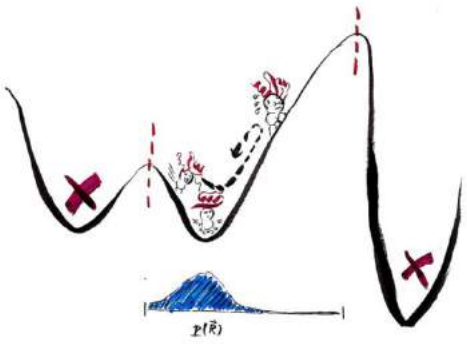
"物理模型驱动的研发范式"应用于原理清晰的场景，套用牛顿方程、薛定谔方程等清晰定义的物理学原理来描述体系内在的规律和外在的表现。以期望获得接近真实世界计算精度的结果。

药物研发的流程中，在结合位点确证环节，计算方法的底层逻辑是基于对蛋白质运动过程的充分采样，理解蛋白质运动的过程及其与蛋白质功能的关系，从而找到与蛋白质发挥功能相关的可药性空腔作为分子靶向的位点；在化合物活性优化环节，计算方法的底层逻辑是模拟蛋白质-配体复合物的动力学行为，并在特定的势函数（力场）下评估不同的配体与蛋白质的亲和力变化；在药物晶型预测环节，计算方法的底层逻辑是准确描述不同分子间的相互作用，及不同分子组合方式的稳定性。这些场景的背后都对应着明确的物理问题，因此，都能利用基于物理模型的计算得到较好的解决。在这些场景中，需要考虑的共性因素是对体系动力学行为的采样效率和势函数的准确性。

物理模型精度的核心是采样问题和力场问题。

在采样问题上，传统的采样方法效率较低，在可承受的计算代价下，能同时描述的反应坐标数量较少，在有限的模拟时间内，容易陷入势能面上的局部极小值（右图），从而难以系统观测体系的动力学全景。

AI 技术在快速处理复杂数据和变量时具有天然的优势，也给采样带了了多种思路和方法。比如通过基于神经网络的偏置势，在模拟过程中跳出势能面局部极小值，从而实现对多个反应坐标的同时描述和对体系动力学状态的系统探索。



图：传统采样难以跨越局部能量最低点

在力场/势函数问题上，传统的方法主要有两种策略，第一种策略是直接进行量子化学就算，这类方法准但不够快、计算成本过高；第二种策略是拟合量子化学计算和相关实验数据的结果获得经验性力场，这类方法通常快但不够准。神经网络势方法的出现打破了快和准的内在矛盾。深度势能等方法依靠量子力学模型提供训练数据，用深度神经网络对高维势函数进行拟合，实现了以分子力学的计算代价，获得量子力学精度势能面的效果，极大地推动了第一性原理方法的模拟边界。

因此，基于近年 AI 技术的发展，传统物理模型的采样效率和力场精度都分别得以有了较大的突破，在可期的未来，类似技术的结合和发展或许也将为药物分子和蛋白质行为的模拟带来准确度的提升和模拟边界的突破，从而为下游的药物研发问题提供有力的帮助和新的视角。

II. 数据驱动的药物研发

"数据模型驱动的范式"的应用场景是：底层规律不够清晰，无法用物理模型良好定义，却有大量数据的情况。此时，可以依托已有数据的积累，通过对数据进行整合提炼分析，从而得到规律来预测相关问题在真实世界的表现或行为。在过去 AI 技术的发展中，人脸识别、自然语言处理等问题，都是经典的数据模型驱动的场景。

同样，在药物研发的流程中，在靶蛋白结构建模的环节中，依赖于多序列比的结果，人们得以学习蛋白质进化上的行为和关键氨基酸对的匹配情况，从而衍生出蛋白质折叠技术，而如果基于物理模型计算的思路，为了达成同样的效果则需要付出成百上千甚至更大的计算成本；在成药性优化环节，受限于人类对生命科学领域运行原理的认知，大量的不确定因素导致 ADMET 性质预测等问题无法抽象为物理模型并进行计算，而对历史探索积累的大量数据的总结和推演则成了这一环节的有效驱动力；在合成路线预测环节，对于既有反应规则的学习和推演也有效地演绎了逆合成分析的艺术。

但如前文所提，数据模型驱动计算往往受限于数据的质量和数量。同样的实验，在不同的机构、研究组和仪器下会产生不同的系统误差，更不用说有相当一部分公开文献的数据不可重复，足够规模的高质量数据成为了数据模型驱动计算的瓶颈。

而依靠 AI 的迁移学习、预训练模型的策略，人们得以通过大量数据进行学习后，再利用小样本的数据集再次训练，来实现将有限的数据价值最大化的目的。

III. 物理模型和数据驱动的融合

在数据缺乏的时候，可利用高精度物理模型生成的仿真数据反过来优化和纠正 AI 数据模型，从而达到数据驱动和物理驱动相互促进的发展。

同时，AI 还可以学习先验知识和科学原理，将"模型"、"数据"和"知识"有机结合起来，促进科学原理的进步。

在新范式下，药物研发相当一部分实验内容也可以像飞机、汽车等领域实现仿真模拟工业化，通过更多的计算手段进行测试和筛选，最后再通过真实实验去进一步的验证和筛选，减少真实实验带来的时间和经济成本的消耗。

展望未来，或许随着生命科学的探索和发展，更多底层的科学原理和规则可能产生重大突破，细胞、动物、人体的作用变化对人类来说可能不再纯是黑箱问题。很多现在只能通过数据驱动去解决的问题可能会变得不那么神秘。

这导致的或许不仅仅是对目前视为常规验证性实验需求的降低，而是有可能从底层对科学带来跨越式发展。届时很多疾病的通路和影响因素也将不再神秘，有可能做到早期识别和阻断疾病。甚至基于对底层原理的理解，对生态系统、海洋系统进行建模计算，模拟类似"蝴蝶效应"的发生过程，从而介导人类健康乃至整个生命科学领域的系统进步和重大突破。

I. 靶点蛋白结构解析、功能机理探索和理性设计

a. 蛋白质结构的解析曾困扰人类 50 年

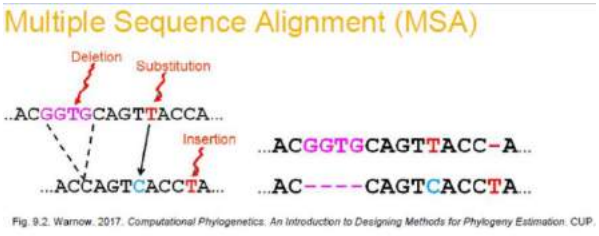
通过测序手段我们已知的蛋白质有数十亿条序列，人们基于此预测会有 2 亿个对应的蛋白质结构。而过去几十年通过 X 射线衍射（X-Ray）、核磁共振（NMR）、冷冻电镜（Cryo-EM）技术等实验手段只解析了约 17 万个蛋白质结构，且其中还有大量的重复数据，这大大限制了我们基于蛋白质结构的探索。

2016 年之前，人们尝试使用计算来辅助预测蛋白质结构的手段主要分成两类，即基于物理模型计算和基于结构特征统计信息。前者通过采样方法根据序列随机生成蛋白质构象，再基于分子模拟来获得能量最小的构象；后者则预设序列相近的蛋白质具有相似的结构，在此假设下，通过寻找当前序列的同源序列来模建蛋白的结构；这种策略通常在模板蛋白序列与目标蛋白序列的同源性高于 30%的情况下才能得到高精度的预测结果，在序列等同度较低时则难以进行三级结构的预测[5]。

这两种方法均难以扩大应用，此时人们对超过 150 个氨基酸序列的蛋白质结构预测都很困难，基本只能依靠实验室方法解析蛋白[5-9]。

b. AI4S 使得蛋白质结构预测成为现实

2014 年，人们开始尝试机器学习的方法来预测蛋白的结构，提出了（右图）多序列比对 MSA 的方法，将蛋白切割成不同的区域，对局部氨基酸采样后在蛋白序列库中寻找与原序列接近的序列，找到其中共进化序列来预测原子之间的相互作用关系。这种方法称为共进化分析方法，即前置于整体蛋白结构的预测，会先在空间上对氨基酸距离的预测采取局部取样计算，但这样的策略对很多蛋白质没有很好的效果。



图：多序列比对（MSA）

2017 年，芝加哥大学丰田计算技术研究所的许锦波教授利用深度学习首次解决了蛋白结构氨基酸之间空间距离的预测问题。同样是基于多序列比对（MSA）的方法寻找共进化序列，区别于传统方案中只考虑每对氨基酸距离预测，许锦波通过对空间中的所有对氨基酸距离的同步预测，基于深度卷积残差网络（ResNet，一种算法）对蛋白序列进行卷积变换，准确地预测蛋白质氨基酸之间的相互作用关系，获得二维图像，而后再通过相互关系重构其三维坐标[7]。

这种方法后来得到了 DeepMind 的重视，其核心思想也被运用到 AlphaFold 的构建中：首先使用 ResNet 从 MSA 多序列比对来预测蛋白氨基酸距离地图（distance map），同时预测蛋白质中的二面角的旋转折叠情况，最终输出的是蛋白的二维图像。AlphaFold 已是数据驱动的较为成功的蛋白结构预测方法，2020 年，进一步进化的 AlphaFold2 已可端到端获得蛋白三维结构，且其预测的蛋白结构已接近实验精度。

第一代 AlphaFold 的出现，证明人工智能预测蛋白结构可行，加速了人类对蛋白结构的解析进程。著名的结构生物学家施一公认为 AlphaFold 的出现是

对结构生物学领域颠覆性的突破[17]。此时 AlphaFold 对蛋白结构的预测还未达到实验精度。

随后，来自华盛顿大学的生物化学家 David Baker 在《Science》上发表了一种算法 RoseTTAFold，他们结合了 AlphaFold 的算法，加入了退火机制（用于取能量最小值），虽然预测精度略逊于 AlphaFold，但对算力的需求更低。

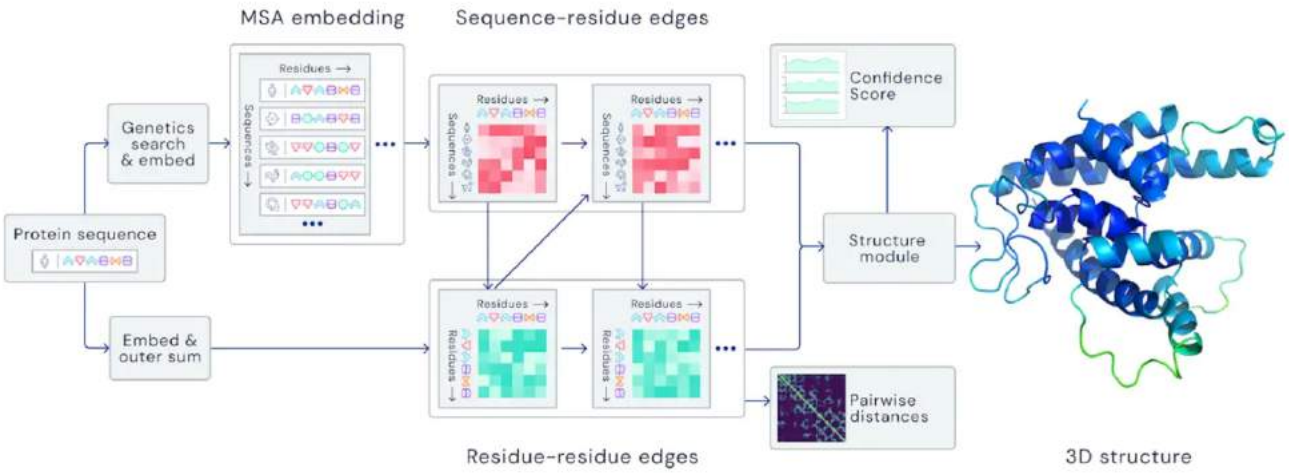
2020 年 AlphaFold2 在原有基础上进行了技术迭代（下图），赢得了国际蛋白质结构预测顶级竞赛 CASP 14 比赛。AlphaFold2 单体蛋白质结构预测结果平均 GDT-TS 得分达到了 92.4，达到了一定实验精度，这个成果可以说是轰动世界。

AlphaFold2 带来的一个前所未有的改变是：可以直接端到端生成原子的三维坐标。这主要是因为他们引入了 structure module（结构模块），并引入了新的算法——注意力机制（Attention 架构，可以进行自监督和相互监督），同时他们能够对三维结构进行原子水平的优化[19]；

更强大的 MSA 序列比对和三维模板搜索提升了整体预测准确度的上限，这主要是 AlphaFold2 在 MSA 序列比对中引入了 BFD、MGnify 和 UniRef 90 这三个更大的蛋白序列数据库；同时在三维结构的模板搜索中引入了 PDB 和 PDB 70 聚类蛋白结构数据库，同时在 MSA 序列比对、三维结构模板搜索中都引入了注意力机制，可以自监督和相互监督[19]；

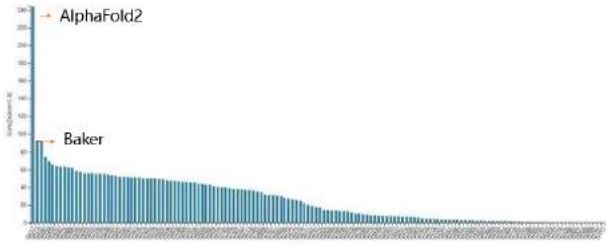
AlphaFold2 架构中引入了 Recycling 多轮迭代机制，使得在结构预测、Structure Module 以及多输出这几个模块进行多轮迭代，使得整个模型的深度更深，能够通过迭代使得结构预测更精确，能预测到更为复杂的结构，比如有些蛋白需要优化到第 4 轮才能折叠到正确的结构；

同时 AlphaFold2 有多输出的功能，构建了自己的打分机制，能够在输出蛋白结构的同时还输出了 Ca 局部距离差异测试（IDDT-Ca）的打分，用于反映自己对蛋白质三维结构预测的准确程度。



图：AlphaFold2 算法框架[20]

这些改变，使得 AlphaFold2 赢得了 CASP 14 大赛，并且能够实现鲁棒性（能够得到很稳定、很好的结果），达到了一定的实验精度。[21]



图：AlphaFold2 在预测精度上取得重大突破

AlphaFold2 虽然开源了模型的推理代码，但训练代码未公布，同时不可商用，研究者难以通过训练或调整模型进一步提升 AlphaFold2 的表现，或迁移更多的应用场景，并且高度依赖于 Google 的生态系统。与此同时，市场上也有大量的折叠预测方法推出，比如天壤 XLab 的 TR-Fold，华深制药的 helixonAI，华为的 m-fold，JOHNS HOPKINS 大学的 IgFold 等，都致力于推动蛋白结构预测的发展。

在开源方案中，深度势能团队开发维护的 Uni-Fold 是首款完全开源且成功复现大规模训练的工具，不仅成功复现了 AlphaFold2 的全规模训练，并且克服了 AlphaFold2 硬件支持单一的难题，在训练和推理环节进行了 NVIDIA GPU 上的适配、性能优化及功能完善的工作。Uni-Fold 脱离了对 Google 所提供的的 TPU Pod 的依赖，并且开源了训练代码，模型可商用，为更多人参与推动领域进一步发展提供了基础。

在相同的测试条件下，Uni-Fold 预测的单体蛋白质结构平均 Cα-lDDT 高达 86.1，预测精度已赶超了 AlphaFold2 和 RoseTTAFold。在复合物方面的预测，Uni-Fold 的 DockQ 平均分数达到 75.7%，平均 TMScore 达到 74.9%，超越了 AlphaFold-Multimer，且完整训练时间消耗降低了一半。

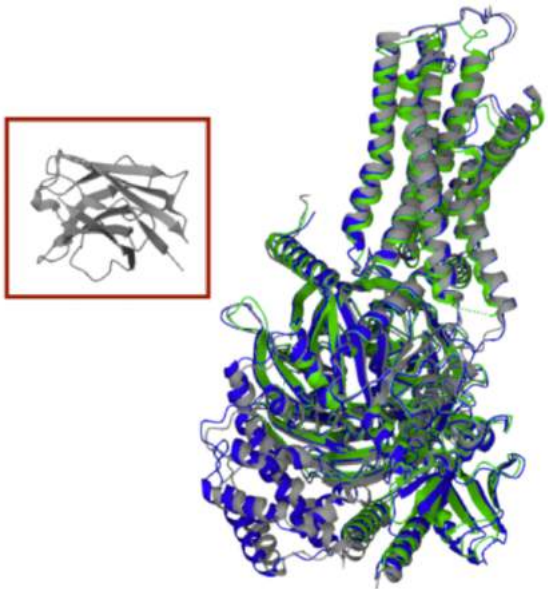


图. 蛋白质结构预测结构比对：蓝：Uni-Fold v2.0.0 预测结果；绿：实验结果（异源五聚体，PDB-ID 7V9L）；灰：AlphaFold-Multimer 预测结果，红框中的单链结构出现较大偏离。[36]

c. AI4S 助力冷冻电镜数据三维结构的自动化搭建

随着冷冻电镜技术的发展，通过实验手段获得复杂生物体系的真实蛋白质结构变得相对容易，尤其是对于药物研发比较重要的膜蛋白体系。药物研发中需要高质量的蛋白质原子尺度结构模型（RMSD<2 Å，RMSD 为均方根差，数字越小，说明与真实实验结构之间误差越小，精度越高），但如果想直接从电镜拍摄的原始数据获得原子尺度的蛋白质三维结构模型，则需要在模型构建时消耗大量的人工、时间和算力。

在 AI4S 的新范式中，可通过蛋白序列及电镜数据结果，直接端到端地获得相应的原子三维结构模型，实现准确且高通量的蛋白结构解析，赋能结构生物学研究和下游的新药研发。

例如，深势科技开发的 Uni-EM 实现自动搭建分辨率不同的多种生物大分子结构。原本需要通过同源建模，手动搭建调整与冷冻电镜密度图不相吻合的结构，经过 Uni-EM 的快速优化，最终获得的结构模型与文章所发表的基于领域专家经验搭建的结构十分接近，且时间可由原来的数周降低至一小时。

进一步展望，未来还有可能利用 AI4S 更好地处理低性噪比的数据，降低对电镜图分辨率的精度要求，从而获得降低电镜拍摄成本、提升各型号电镜的使用效率。且在此基础之上，甚至还有可能借助 RiD 的增强采样，进一步提高冷冻电镜结构解析的精准度（见下一节）。

d. 强化动力学用于蛋白质结构精修

目前蛋白结构预测有一定局限性，蛋白质的较长环区预测结果差，难以达到接近实验的精准性，这需要进行蛋白质精修，其中使用分子动力学从原始结构出发寻找能量极小点的精修方法效果最好。

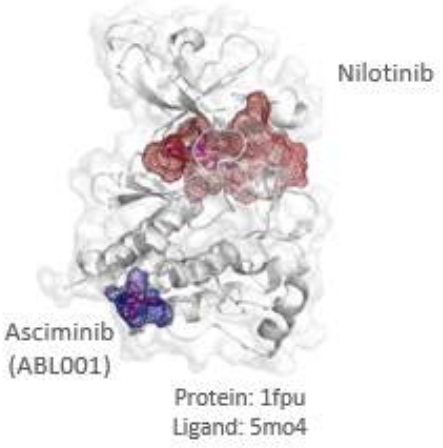
此外，蛋白质处于动态三维结构，需要动力学模拟才能寻找更多隐藏靶点以及别构靶点。然而，传统采样方式难以跨越局部的能量最低点，陷入这个局部极小值势阱，对整个体系的变化不能有效采样，无法有效探索蛋白的潜在结合位点，并且对蛋白结构优化有限。所以以上问题都需要高效的采样工具。

正因如此，在 CASP14 比赛中，仍有 47%的蛋白结构 AlphaFold2 的预测结果打分小于 75。这是由于通过 MSA 多序列比对所预测的蛋白结构的某些区域缺少可参考的同源结构，预测结果需要进一步优化。

而优化的方法中，与图神经网络方法相比，借助分子动力学模拟从初始结构出发对局域相空间采样，寻找最合理的构象优化表现最好[10]。

如前文所述，分子动力学等基于物理模型方法的两大难题就是力场精度和采样效率。力场精度的解决方法我们前文已提及；采样效率的提升也有多种方法，但难以处理几十个集合变量以上的采样数据。

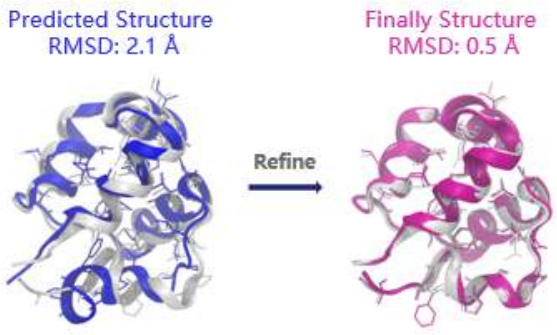
蛋白构象的动态变化直接影响功能调控，同时与药物在生物体内的活性高度相关。基于 AI4S 的底层逻辑，融合机器学习的增强采样算法助力结构生物学发挥更大价值。相较传统分子动力学能处理时间尺度问题，RiD（强化动力学）具有独特优势找到别构位点以及进行诱导契合对接；



图：BCR-ABL1 别构靶点探索

在针对 BCR-ABL1 的案例中（上图），蛋白质构象变化由 RiD 采样，而后计算可能的口袋及 correlation，从而找到与正构口袋 correlation 高的别构口袋；

此外，RiD 助力动态构象采样与结构精修。在实践中，使用 AlphaFold 预测的蛋白，RMSD 为 2.1Å，蛋白质侧链结构预测还需进一步优化，使用 RiD 优化后，蛋白质的 RMSD 降低至 0.5Å（下图）。



图： RiD 应用下，RMSD 由 2.1Å 降低至 0.5Å

第三，RiD 可帮助开发"难成药"的固有无序蛋白(IDP)。IDP 生理状态下没有固定三维结构，常规分子动力学方法难以模拟 IDP 的结构，通常被认为难成药靶点，人们发现固有无序蛋白在紧凑的无序状态有助于配体的识别[23]，加深了人们对固有无序蛋白相互作用机制的理解，并为 IDP 的理性药物设计提供了重要理论基础。

RiD 使得对构象空间充分采样成为可能，这有助于对 IDP 这种生理状态下不断动态变化而难以寻找有效药物的蛋白的成药性探索。

e. DNA/RNA 折叠结构探索，助力药物靶点的开发

体内的遗传信息表达不仅和碱基序列相关，遗传物质的三维结构的折叠方式也会影响基因的表达，所以对遗传物质三维结构的解析意义重大。

由于 DNA 的折叠和表达的原理不够清晰，以及 RNA 结构稳定性低，目前很难用实验方法解析它们具有功能机制的结构。

来自 Howard Hughes Medical Institute 的研究组在《Nature》上展示了 4D Nucleome 项目，他们尝试使用物理建模计算的方法解释实验观察结果并构建原子核的动态空间模型，在此基础上再对功能进行验证。[12]

RNA 的结构解析受 AlphaFold 成功案例的启发，来自斯坦福大学的研究团队在《Science》发表的结果表示其可预测 RNA 三维结构，这是基于对已知的 18 个 RNA 结构深度学习训练预测的结果。来自深圳湾实验室周耀旗课题组在《Nature Communications》上的结果表示，其已可基于深度学习实现对 RNA 近天然态结构的高精度优化修正[13-15]。

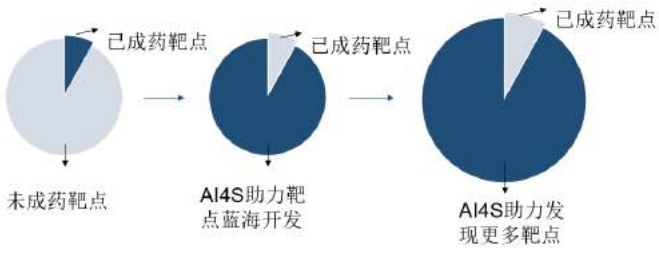
随着 AI4S 的发展，遗传物质 DNA/RNA 三维结构的解析，或许也能达到接近 AlphaFold2 的水平，不仅能够接近现实世界实验精度，并且也拥有自己的打分机制来评判其置信区间，这将给生命科学领域带来巨大的进步。

同时，DNA 和 RNA 都可以作为药物的靶点来进行药物发现的研究[16]，或许对这些遗传物质结构的预测，会给药物研发打开一扇新的大门。

f. 基于结构和动力学的蛋白质功能机理探索

靶标发现和筛选是药物开发的第一步，传统对于靶标蛋白的机理研究，是基于生物学实验手段进行验证，多是通过对靶点的敲除或增强来验证是否作用在此靶点蛋白上，再借助其他实验手段来检测下游的细胞因子或炎症因子的释放情况，从而验证蛋白功能。

从 2.1 的部分我们可以得知 AI 可以从三个层面对靶标的发现提供助力：对分子通路和应答机制进行规



律挖掘；探寻病人的异常的基因表达和调节通路；利用 NLP 探索疾病和靶点的对应关系等。已有商业化公司在这一领域探索，此处不再赘述。

g. 强化动力学助力蛋白动力学的发展，开发更多药物靶点

想要得知蛋白序列与功能对应关系背后的原理规则，还需要探索从蛋白的序列到蛋白结构再到功能之间的对应关系。

由于组成蛋白质原子在生理状态下，比如蛋白和配体结合，通道蛋白激活前后以及蛋白和药物分子的结合时，蛋白原子在不同尺度上存在运动，所以在这个对应关系的研究过程中还需考虑到蛋白的动态构象，这需要涉及到蛋白质结构动力学，也需使用分子动力学进行模拟计算[24]。

如前文所述，分子动力学主要受限于力场精度和采样效率，力场可以通过前述 AI4S 方法提高力场精度来解决，采样效率则可借助 RiD（reinforced dynamics/RiD）覆盖更大的构象空间，这些解决了对蛋白动态构象的探索问题。

基于这些动态构象，才能更好地探索从结构到功能的对应关系，研究蛋白信号通路的规律及与疾病的对应关系。

在此之外，RiD 还有助于对难成药蛋白质靶点不同构象变化中隐藏口袋和别构口袋的探索。比如固有无序蛋白（Intrinsically disordered proteins, IDP）是一种在天然状态下缺乏稳定三维结构的蛋白，无序蛋白在溶液中始终在快速变化，其所具有的动态特性使得传统的分子动力学模拟难以探索其构象空间，RiD 可帮助探索 IDP 的可药性。届时药物研发不仅是能够开发难成药靶点，还可发现更多生物机制的规律，扩大可药靶点的范围。

h. AI4S 助力蛋白-蛋白的相互作用（PPI）预测，推动生物大分子药物研发进程

在有精确蛋白质结构的基础上，我们可以更轻松、更准确地使用计算方法来预测蛋白-蛋白相互作用（protein-protein interaction, PPI）。

PPI 的计算主要困难有三个方面：一是蛋白与蛋白之间的接触面积较大，对采样要求较高；二是蛋白与蛋白结合过程中会产生动态构象变化，难以用传统分子动力学方法进行描述；再一个是结合过程中蛋白原子与溶剂的相互作用变化受限于算力而难以描述。基于以上的困难，计算方法难以对抗原抗体结合，MHC 结合等 PPI 相关的问题进行预测，主要是依靠实验手段，停留在计算 0.0 的重复实验阶段，试错成本较高，一旦实验结果不太好，则需要重新优化。

随着 AI4S 的发展，AI 较为擅长高维空间的搜索，可以提高搜索效率，有效地解决 PPI 计算的第一个问题；同时 RiD 和力场精度的提升给采样带来很大的改善，解决了第二个难题；基于第一性原理的计算精度来快速计算大量原子相互作用的能力又给第三个问题带来解决方案。

在科研实践中，David Baker 团队尝试在设计蛋白质时，根据 AlphaFold 和 RoseTTAFold 的预测结果，尝试模拟蛋白质之间的相互作用，并发现了以前未知的 100 多个多聚体 [25]

i. 生物大分子药物的定向优化

随着小分子药物研发的竞争激烈，人们开始探索更多基于生物大分子的疗法，如抗体，多肽，酶，通道蛋白，mRNA，siRNA 等。类似 PPI 预测的难题，在这些计算难题解决之前，人们只能停留在计算 0.0 时代，借助实验的方法来测试他们的性质或验证设计的优劣，试错成本偏高。

随着 AI4S 的出现，前文已我们提到了应用 Folding 方法来解析蛋白结构，David Baker 团队尝试在设计蛋白质时，根据 AlphaFold 和 RoseTTAFold 的预测结果，不断优化氨基酸序列来获得更优的蛋白质序列 [26]。

抗体优化中，主要是需要优化其亲和力或人源化改造；酶主要是优化其稳定性和活性；在高通量测序纳米孔设计中，主要是希望模仿通道蛋白特性。

Moderna 在开发新冠 mRNA 疫苗时，已经开始尝试使用计算模拟病毒结构，并从同源序列中挖掘最优的 mRNA 序列[27]。

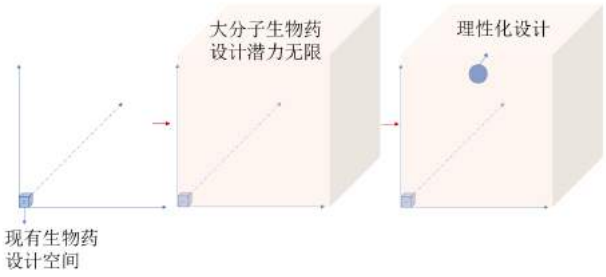
利用点突变与深度学习结合，可以帮助我们探索蛋白序列与亲和力、稳定性、人源化、通道蛋白等功能性质对应的规律，并达到根据序列预测功能的能力，完善前文所述序列-结构-功能对应关系的探索。

j. 基于性质/功能的大分子从头设计进程

在未来，有可能基于前述所说的对蛋白-蛋白相互作用关系的探索，以及蛋白/多肽定向优化的探索，使人们更深刻地理解序列与功能之间的对应关系。就有可能实现基于目标功能从头设计所需要的抗体药物，多肽药物，抗体试剂以及酶。甚至在材料应用上也可使用，比如纳米孔或者人工骨骼上的蛋白分子。

目前已有一些迷你蛋白的设计，可精准结合到蛋白特定的活性位置[28]。

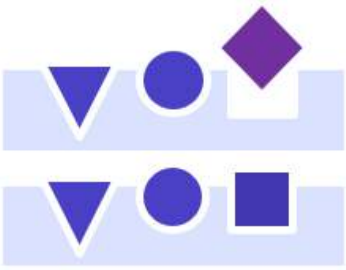
基于前述，就可能实现下图所描述的图景：基于深度学习的蛋白序列设计和结构预测，扩大了生物大分子药物的设计潜能，在基于对序列-结构-功能对应关系的探索上，又可能达到基于理性化设计的理念，缩小需要验证的药物范围。



图：AI4S 最终实现大分子药物的理性设计

II. AIGC：基于靶点空间构象的分子设计

在对药物分子设计的过程中，除了可从现有的药物数据库中筛选，还可通过人为方法也可通过计算手段来设计小分子化合物，但人的经验和人脑的想象空间终究是有限制的，所以出现了基于机器学习生成分子的方法。



图：基于靶点的分子生成示意图

生成的药物分子也会有两种用途：一种是大量生成分子，形成数据库（比如 ZINC20 数据库等），再去筛选出所需的分子；另一种是边生成边修正生成模型，使得生成的分子更符合筛选条件的定向分子生成（上图）。

两种方案均面临以下挑战：生成的药物分子不够符合人们的需求：或是生成的分子无法合成；或是能合成但属于很浅显的科研经验即能设计出对应分子。

AI4S 已经基本解决蛋白质结构的问题，在力场精度和采样效率的问题上也有了一些突破，结合已出现的基于靶点蛋白的分子生成理念，可在探索到的靶点口袋中，通过片段连接、片段替换、骨架跃迁、片断生长等方法来设计药物分子，改善分子的理化性质。

蛋白靶点和分子的结合姿势及对应能产生的相互作用决定了它们的结合强度，基于蛋白靶点结构生成三维分子（即生成由特定类型和相应三维位置的原子组成的分子）在药物发现领域具有重要意义。近年来，基于深度学习的三维分子生成模型取得了显著进展，以往的研究主要分为两类：

1）三维密度网格生成方法：此类方法将蛋白质口袋和分子转换为具有粗粒度位置信息的三维密度网格，然后利用生成模型预测每个网格中的原子密度，最后将网格密度图转换为原子。然而，由于这些方法生成的模型仅能产生网格级别的坐标位置，因此无法获得高质量细粒度的三维分子结构。

2）自回归三维生成方法：此类方法逐个生成带有三维位置和原子类型信息的原子。然而，在训练过程中，确定首先生成哪些原子具有挑战性，而生成原子的前后顺序往往对生成结果有所影响，因此这些模型通常表现不佳。

2023 年，深势科技发表 VD-Gen 可以生成与给定蛋白靶点具有高结合亲和力的分子 [1]。VD-Gen 受

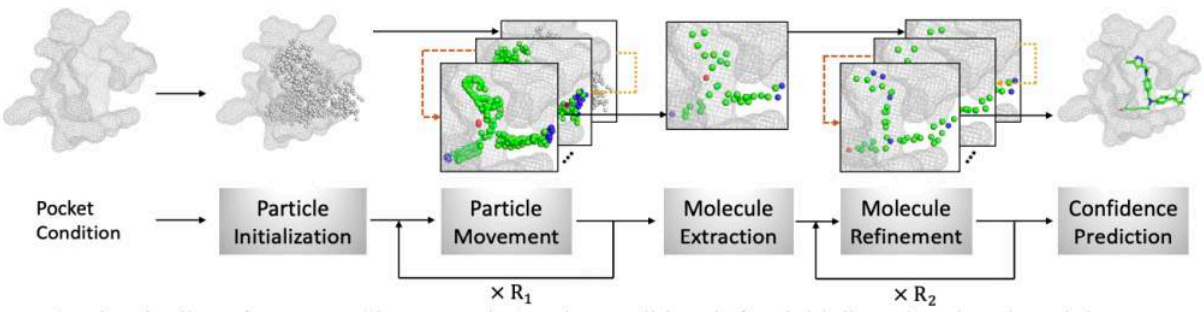


图. VD-Gen 通过五个阶段实现端到端的 3D 分子生成

启发于 AIGC 领域的 inpainting 模式以及分子动力学，其核心思想是通过模型学习高亲和力分子中的原子在靶点空腔中的分布，然后通过让虚拟粒子运动来逼近这个分布。相较于其他分子生成方法，VD-Gen 可以基于给定蛋白靶点结构，直接生成细粒度的蛋白靶点-小分子结合构象。大量实验结果表明，VD-Gen 可以生成新的具有高结合亲和力的 3D 分子以填充口袋空腔，显著优于先前的基线。

具体而言，VD-Gen 通过五个阶段实现端到端的 3D 分子生成，如下图所示。

（1）粒子初始化（Particle Initialization）：在给定蛋白靶点的情况下，VD-Gen 在三维蛋白靶点空腔中初始化多个具有空类型和随机位置的虚拟粒子；

（2）粒子移动（Particle Movement）：迭代地移动虚拟粒子位置并预测虚拟粒子类型，直至达到平衡，理想情况下，平衡状态下的虚拟粒子在三维空间上的分布将接近目标分子的原子分布；

（3）分子提取（Molecule Extraction）：模型从平衡状态的虚拟粒子分布中提取原子组成分子；

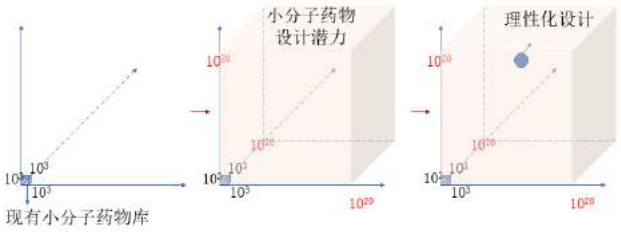
（4）分子优化（Molecule Refinement）：在提取原子后，VD-Gen 再次迭代移动原子，使原子移动到更精确的位置；

（5）置信度预测（Confidence Prediction）：预测生成分子的置信度分数，以进行选择或排序。

与基于 3D 网格的生成模型相比，VD-Gen 可以生成具有细粒度坐标的高质量 3D 分子。与自回归生成模型相比，VD-Gen 能够高效地同时生成所有原子，因此性能更优，不受生成顺序的影响。

AI4S 算法正快速发展。分子表征方法与亲和力评估模型结合，若能高精度探索出构效对应关系，则可根据各性质的预测对应调整修正所生成的分子结构，更进一步地可根据所需性质来理性生成对应的分子结构。

如下图所示，AI4S 在药物发现中的应用，不仅能扩大可筛的药物分子库，未来还可更进一步，直接根据科学家的搜索所需"理性设计"分子，大大压缩需要实验验证的环节。



图：AI4S 助力小分子药物设计

Source: [1] Shuqi Lu, Lin Yao, Xi Chen, Hang Zheng, Di He, and Guolin Ke. "3D Molecular Generation via Virtual Dynamics." arXiv preprint arXiv:2302.05847 (2023).

III. 从 Docking 到 FEP：AI 增强"靶点-药物配体" 亲和力评估与高通量筛选

在众多理化性质中，与靶点蛋白有较高的亲和力是配体分子能够成药的必要条件，而亲和力可以通过结合自由能表征。

过去二十年来，粗略计算自由能方法的分子对接（Docking）[30] 得到了广泛的应用，其有能力进行千万级别分子数据库的筛选，但打分函数本身计算方法近似较多，以及对分子构象的采样不足限制了这种方法的精度，因此其多应用于药物设计的粗筛阶段以得到苗头化合物，但有可能漏筛活性较高的分子。

传统的分子对接过程如下：用蒙特卡洛方法（Monte Carlo method，MC）对蛋白-配体复合物构象开展全局搜索，进行配体可旋转键二面角和配体位置的组合；计算当前复合物构象下配体的能量和受力，使用 BFGS 算法梯度下降开展局部优化，获取局部能量最低的复合物构象；重复若干次如是过程，直至搜索步骤耗尽；将所有低能复合物构象比对，返回最低能的蛋白-配体复合物构象和对应的能量。在行业中，广泛使用的是 AutoDock Vina 等国外软件。2022 年， 自然杂志子刊《Nature Protocols》报道了 UBC 研究组 Deep Docking 的虚拟筛选效果。在 Deep Docking 工作流程中，仅使用需要对接的数据库的很小一部分（1%）来训练深度神经网络（DNN）模型，从而从二维分子指纹预测对接分数。然后推断出剩余分子的分数类别（高分或低分），跳过实际对接显式计算。最后，只有预测出的得分最高的分子需要进行传统的对接，而不利的分子则被过滤掉[1]。

2023 年，知名国际期刊 JCTC 在封面报道深势科技最新成果 Uni-Dock[2]。通过充分利用 GPU 并行计算性能和显存空间，结合细致的流程设计调优，在保持与传统分子对接可比较的精度的前提下，Uni-Dock 在 NVIDIA V100 GPU 上实现了对比 AutoDock Vina 单核计算超 1600 倍的加速比率，是其他 GPU 加速的分子对接引擎的 10 倍以上，以 0.1s/ligand 的分子对接效率，仅花费不到 12 小时就能完成超 3820 万分子数据库的虚拟筛选工作。

与 Docking 相比，MM-PB/GBSA 方法将结合自由能近似为分子力学与溶剂化能项，精度相比 Docking 更高，可以用于近千个分子的亲和力评估工作。

结合以上两种办法可以实现多级打分来更精准地筛选药物分子，比如首先进行刚性对接筛选，然后是半柔性分子对接，再是柔性分子对接，最后再通过 MM-PB/GBSA 来进一步打分筛选，提高 docking 的计算精准度。

此外，自由能微扰方法(Free Energy Perturbation, FEP) [31] 严格基于统计物理理论，并通过设计一条转换路径并沿路径进行分子动力学模拟，可以对结合位点构象变化进行有效采样，从而精准评估蛋白与配体的结合自由能。Schrödinger 研发的 FEP+与深势科技研发的 Uni-FEP 等自由能微扰计算软件对活性的预测与实验结果相比，可达化学精度（活性差异约 6 倍之内）。

先导化合物通常还存在着某些缺陷，如活性不够高，化学结构不稳定，毒性较大，选择性不好，药代动力学性质不合理等，需要对先导化合物进行化学修饰，进一步优化使之发展为理想的药物。此时药物化学家经过对先导化合物的侧链基团或部分骨架进行改造，设计出数十/百个活性可能增强的分子，下一步目标是挑选出活性最高的分子，而将这些分子全部实验合成并实验验证的方法成本过高、周期过长。

由于自由能能够与活性（IC_{50}）换算，可根据自由能的换算得出药物分子的活性数据。所以已达到实验精度的 FEP 多应用于先导化合物优化阶段中的活性计算或预测，节省了实验合成到活性验证的成本和时间。

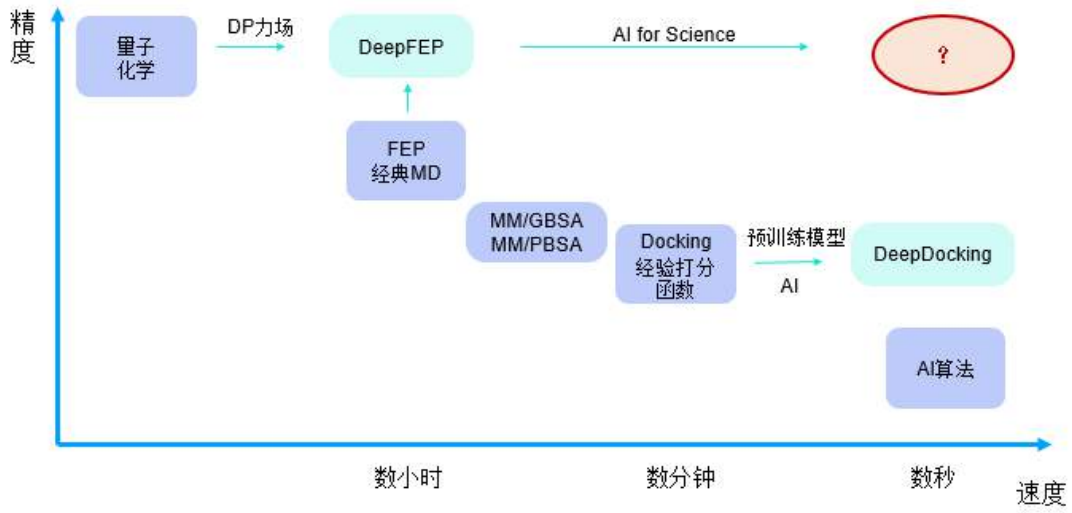
FEP 方法需要进行分子动力学模拟，对 FEP 的计算精度影响的要素包括描述原子间相互作用的力场精度、作为模拟起点的蛋白与配体的结构合理性，以及对分子构象空间采样的遍历性。机器学习力场、AlphaFold2/Uni-Fold 等结构预测工具和以 RiD 为代表的增强采样算法等 AI4S 方法均有望助力 FEP 方法的精度提升。

如下图所示，从 Docking 到 MM-PB/GBSA 再到 FEP，这些方法的精度虽然在逐渐提升，但由于计算量的增加，这些方法的计算速度也逐步降低，从而限制了亲和力评估的通量。但这个问题有望被 AI4S 所突破：如果将 FEP 等高精度方法标注出的亲和力数据，结合 3D 坐标分子表征方法作为机器学习模型的数据集，我们则可以期待根据分子表征准确预测亲和力的新时代！

Source:

[1] Gentile, F., Yaacoub, J.C., Gleave, J. et al. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nat Protoc 17, 672–697 (2022). https://doi.org/10.1038/s41596-021-00659-2

[2] Yu, Y., Cai, C., Wang, J., Bo, Z., Zhu, Z., & Zheng, H. (2023). Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. Journal of Chemical Theory and Computation.



图：AI4S 给亲和力计算带来新的可能性

IV. 预训练大模型驱动 ADME/T 等药物分子性质预测

进入药物优化的阶段，这一步则可根据想要的性质进行分子的定向生成，借助 FEP 来计算或预测分子的活性性质，借助 ADMET 预测来计算分子的药代动力学和毒性性质；在此之后依然需要对分子的可合成性进行预测，甚至自动合成分子，最后再通过湿实验来进一步的验证或筛选。想要向计算机描述小分子的特性，则需要一种能让计算机识别的语言，使用类似于像素表示图片，声波表示声音的通用且完备的方法来描述分子结构信息的数值指标，这称为分子表征。

目前市面上对 ADMET 性质预测已有众多软件，目前的障碍，主要是由于实验数据噪音较大，实验精度不统一，数据格式无标准，或大量的数据积累于各大药企和 CRO 中，共享数据库较少质量不齐。

所以 ADMET 的预测面临的困难主要是数据数量和质量的问题，此外还有分子表征的问题。

预训练模型为分子表征带来了新的思路，那么也依靠其对小样本数据学习的优势解决数据数量和质量的难题，给 ADMET 的预测带来突破。那么对于药物发现的阶段，则可以形成这样一个循环：通过分子生成方法扩充可筛选的药物数据库，从数据库通过分级打分方法实现虚拟筛选出需要的分子；从分子库生成或筛选之后的到的分子，则通过逆合成的方法来筛选出可合成的药物分子，甚至可自动合成药物分子；最后再通过湿实验的方法来验证或进一步筛选。

业界目前主要使用的是一维序列（如 SMILES）和二维拓扑图来表示小分子，但这两种方法都有其不完备和限制之处。寻找到一种好的分子表征方法是人们在探寻的方向。如果没有一个好的分子表征方法，那么所有基于分子表示，依托数据驱动所探索的一切规律，从底层来说是有缺陷的。比如对亲脂性、物化性质，水化自由能，ADMET 等性质的探索，基于此所有归纳的构效关系都不会足够准确。

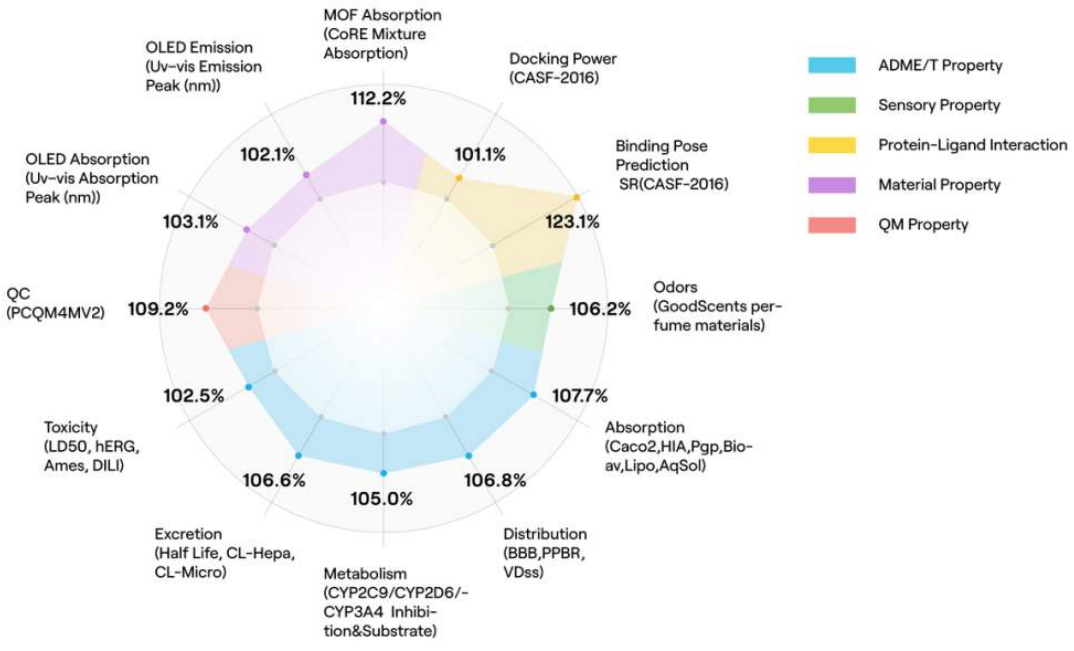


图. Uni-Mol 在各类任务上都表现优异，超越之前的最好方法。图中内部灰色区域为之前的最好方法的效果，外部多种颜色区域描述的是 Uni-Mol 在多种任务上超出之前最好方法的百分比。

化合物分子本身在现实世界中就是三维结构，所以如果能够从三维信息出发去探索分子表征的方法，将会是一个很大的突破口。

2023 年 AI 领域国际 AI 顶会 ICLR 收录了深势科技所开发了三维分子预训练模型 Uni-Mol [29]，其可直接将分子三维结构的坐标信息作为模型的输入/输出。这种直接从三维信息出发，通过预训练模型（下图）和自监督策略的方法，在几乎所有与药物分子和蛋白口袋相关的下游任务中，计算结果都达到了 SOTA。

在与 3D 结构强相关的回归任务上，例如水化自由能（ESOL，FreeSolv），亲脂性（ Lipo），物化性质（QM 系列）等，相对于之前的 SOTA 平均有 21% 的效果提升。在打分函数的 docking power 评测上也超越了主流 Docking 工具以及基于 AI 的打分函数模型。在结合位点的预测中，对于 CASF-2015 基准数据集（RMSD<2.0，一般认为是可接受的 pose 预测结果）预测准确度超过目前主流 docking 工具约 35%。

这种纯基于分子 3D 坐标的分子表征方法，可能成为最优的分子表征方法，在此基础上，结合位点、BBB（血脑屏障）、ADMET 性质预测，构效关系的探索，逆合成预测，亲脂性，物化性质的预测等都有机会取得巨大突破。Uni-Mol 后续推出 "+" 升级版。相较于之前的方法，Uni-Mol+ 在整体框架、模型结构和训练策略等方面都进行了创新。在整体框架方面，Uni-Mol+ 基于低成本的方法如 RDKit/Openbabel 生成初始构象，并通过迭代优化这些构象，使其逼近 DFT 方法得到的高精度稳态构象。这样一来，可以通过基于模型优化后的构象来获得更精确的量子化学性质预测结果。在模型结构方面，Uni-Mol+ 进一步加强了 Uni-Mol 的双分支 Transformer 结构，以更好地捕捉三维空间的信息。而在训练策略方面，Uni-Mol+ 提出了一种新的方法，即线性轨迹注入，可以更有效地学习 DFT 构象的优化。

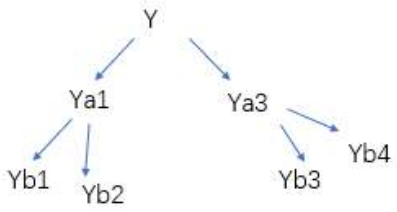
Uni-Mol+ 在 OGB-LSC 的量子化学性质预测任务上夺得了冠军。OGB-LSC 是一项由斯坦福大学发起的学术竞赛，旨在评估机器学习在大规模图数据上的表现。该竞赛首次在 KDD CUP 2021 上举办，吸引了来自 DeepMind、微软、NVIDIA、UCLA 等顶尖企业和高校的 500 多个参赛队伍，备受业界关注。可以说，OGB-LSC 已成为公认的检验图机器学习模型性能的最佳试金石，类似于 ImageNet 在图像领域的影响力。

V. 合成预测及自动化实验

研发过程中，药物分子还是需要通过实验测定各种性质，达到筛选的目的；且前述可知，我们也需要依据可合成性判断所生成或筛选的分子是否可用。

以往化合物分子的合成很依赖于化学家的经验，如果合成路线设计不好，合成则会反复失败，需要消耗数周甚至数月的宝贵时间[32] 。

逆合成由 Corey 在 20 世纪 60 年代提出，用于描述如何通过断键将一个复杂的目标分子还原为简单前体的迭代过程[33]。如下图，对于计算机而言，关键就是预测一些反应，能用相对简单的分子合成目标分子。



图：逆合成思路

逆合成反应预测主要可以分成两类方法：

- 第一是将人的知识总结成公式，这种方法符合化学家的经验，很快就能总结出大量的公式。但实际应用中发现，这些公式会互相影响，所以化学反应不仅是局部反应，也需要考虑整体。同时化学本身很复杂，公式本身如果足够精确，那么公式数量就会巨大，难以总结且应用困难，且不具有广泛的通用性。
- 第二种就是用 AI 的方法，不再拘泥于人能理解的原理，而是根据大数据学习探索规律。这种方法有两个难点，一是数据的获得，目前反应的数据库稀缺；二是缺乏好的分子表征方法。

现在已有众多推出的逆合成软件，比如 ASKCOS，Chemical.AI，IBM 的软件，实际应用过程中，还是会出现一些不尽如人意的表现，比如反应会有明显错误，或将反应路线复杂化等。

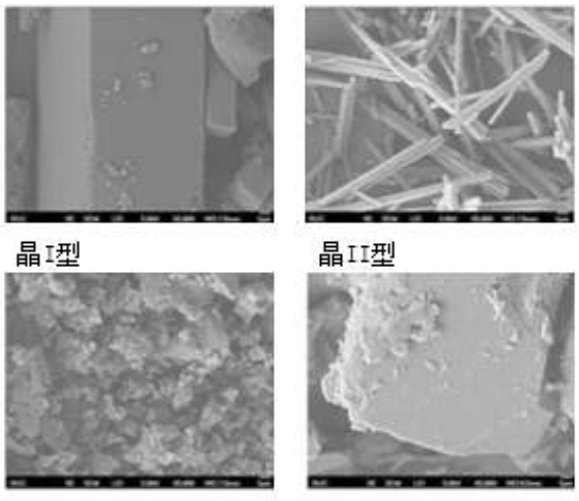
AI 的预训练模型和自监督训练也可以应用于化学合成预测上，基于预训练模型解决数据样本少的问题来解决逆合成的预测。而化学反应条件预测则可以借鉴 AI 辅助计算化学的思路。

VI. CMC 药剂学优化

药物分子的溶解、沉淀或结晶一直是药剂学的基本问题。传统主要通过实验方法来进行研究，试错成本高，一旦不符合需求则需重新进行设计。随着数据的积累，以及计算的发展，可以借助计算机来帮助预配方研究、配方筛选、临床精准用药。

a. AI4S 应用于晶型预测

晶型是药物分子内部质点三维空间中周期性排列的固体，相同的药物分子会因其晶型不同而具有不同的理化性质。如下图所示，药物不同晶型所带来的



图：多晶型药物毗罗昔康

差异可能会影响到药物的专利、稳定性、安全性、有效性以及生物利用度等方面，最终又会影响到药效、给药形式和计量。

晶型预测是通过计算模拟获得其所有可能的稳定晶型，主要分为晶体搜索，能量排位和室温稳定性计算三个问题。晶体搜索需构造关于晶体的所有自由度的能量函数，求解其全局或部分区域最小点；能量排位主要是对于自由能的计算；而室温稳定性需要基于热力学和构型采样；所以这三个问题主要涉及的问题依然是物理模型的力场和采样问题。

目前晶泰科技整合了晶型搜索算法、XForce Field 小分子通用力场，量子动力学计算与晶体自由能计算等技术来预测晶型热力学的相对稳定性，带来一定预测思路的启示。

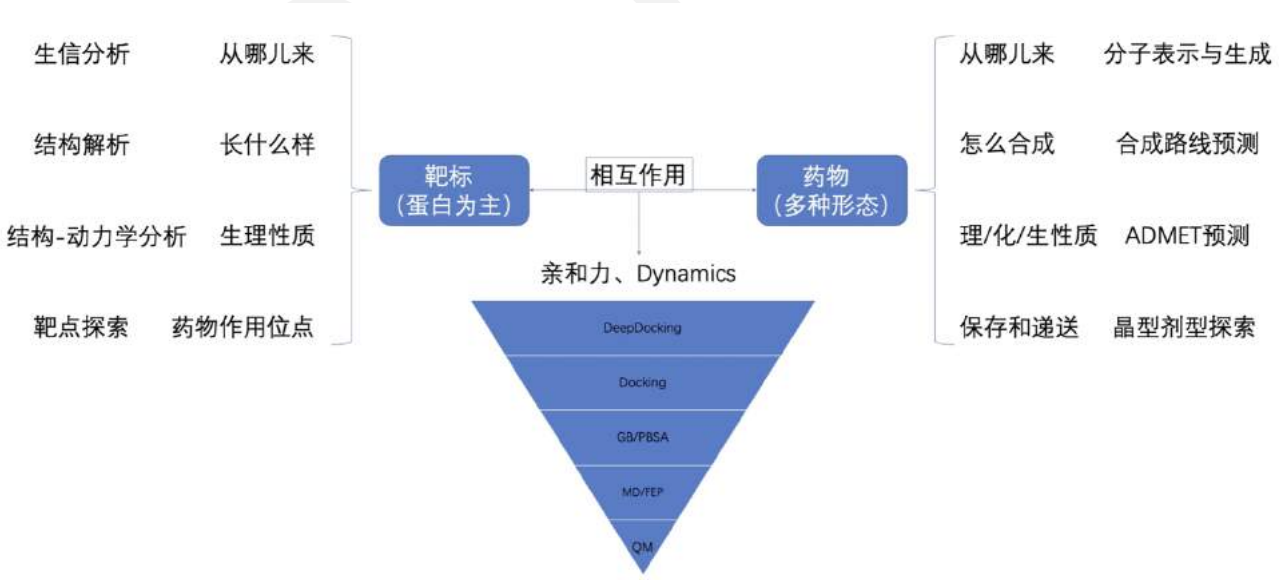
b. AI4S 在剂型预测中的使用和推动

药剂的剂型设计中需选择辅料来决定配方配比，可选择空间约为 10^{25}-10^{30}，仅靠经验很难预测和优化配方。并且由于辅料数量多，分子间作用计算维度过大，目前只能够通过数据驱动的方法预测剂型选择，主要是利用深度学习基于大数据来帮助预测辅料优化和筛选。

目前药物剂型预测产品主要有 Protheragen 公司开发的 MedAI[34]和 Coformula 公司的人工智能工具剂泰医药也在推动剂型预测产品[35]

所以在 CMC 阶段则可以先通过剂型预测、晶型预测，以及前文所述的逆合成和反应条件辅助的合成路径预测来筛选最合适的晶型、剂型和合成路径，然后再进入湿实验进行验证，节省这个环节中的人力、物力成本。

那么药物研发流程就可以在下图中所示的 AI4S 实践模块的帮助下获得效率的提升。



AI4S 实践（1）：AIGC 推动蛋白理性设计，David Baker 团队发表 RFDiffusion

2023 年，基于 diffusion 模型的图像生成软件爆红，从国外的 midjourney, dall-e 到国内的妙鸭等，AI 生成图像展现了巨大潜力。基于类似思路，美国华盛顿大学 David Baker 教授及其团队开发了 Rfdiffusion，它可以"凭空"设计出各种功能性蛋白质。2023 年 7 月 11 日，相关论文以 De novo design of protein structure and function with RFdiffusion 为题在 Nature 上发表。

基于 AI 的蛋白质设计工具出现之后，从氨基酸序列中准确预测蛋白质结构成为可能。这些工具通常使用被称为"幻觉"和"修复"的方法，前者先创建一串随机的氨基酸，并通过 AI 模型进行优化，后者则先获取蛋白质序列或结构的特定片段，并使用 AI 模型在其周围构建分子的剩余部分。不过，这两种方法都存在不足之处。用"幻觉"方法设计出的结构，无法在实验室中稳定地形成折叠良好的蛋白质，而采用"修复"方法补全蛋白质的效果也欠佳。

现在，利用 RFdiffusion 模型，研究人员能够生成与真实蛋白质类似、但在自然界中却从未出现过的蛋白质。

据了解，RFdiffusion 模型由真实的蛋白质结构数据训练而成，这些结构被存储在蛋白质数据库中。用该模型设计蛋白质，最初会显示一些噪音，即 AI 系统随机排列的氨基酸，经过几轮去噪后，氨基酸才逐步变形为复杂的形状，并开始具备与真实蛋白质类似的特征，比如 α- 螺旋和 β- 折叠片。

"与一年前相比，现在的设计过程几乎完全改变了。"论文共同一作大卫·尤尔根斯表示。这种神经网络解决了此前许多效率低下或使用其他方法无法完成的蛋白质设计挑战，比如，蛋白质结合物设计、对称寡聚体设计、无条件和拓扑约束的蛋白质单体设计、酶活性位点支架设计等。

"一个机器学习工程师可以设计出足够 100 个生物学家忙上几个月的设计。"微软研究院生物医学机器学习研究员凯文·杨（Kevin Yang）如是评价。

在进行了大量实验以后，贝克团队发现，用 RFdiffusion 设计的蛋白质，有 10-20% 具有很强的与预期目标结合的能力。相比之下，此前的 AI 方法仅能生成少于 1% 的符合该标准的蛋白质。不过，需要说明的是，用 RFdiffusion 模型生成的蛋白质只是一个三维结构模型，后续还需要用其他的 AI 工具绘制氨基酸序列，以及预测出与设计匹配的蛋白质折叠结构。

Source: DeepTech,
https://mp.weixin.qq.com/s/gRYyGmtpbbONmXss12PJSg

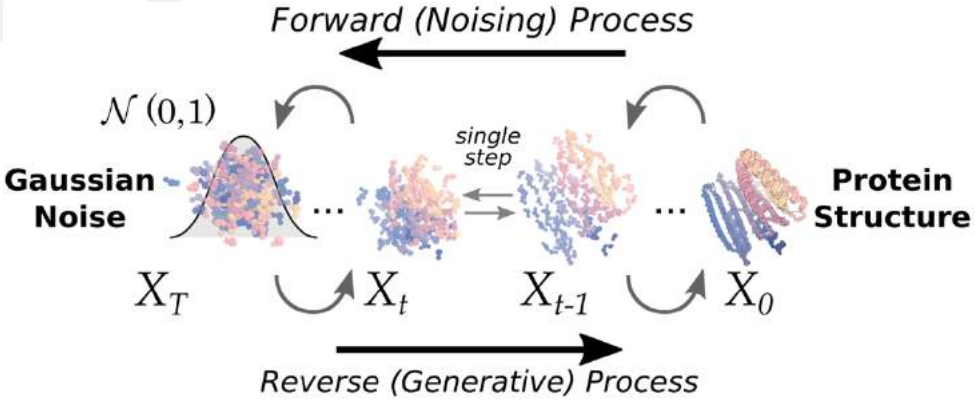


图. RFDiffusion 的模型训练过程

AI4S 实践（2）：Uni-RNA 预训练大模型在广泛下游任务达到 SOTA 性能

RNA 体系复杂，相关疗法开发缓慢

自从 1984 年第一款 ASO 药物得到 FDA 批准以来，RNA 疗法一直被寄予着巨大希望，有助于解决疾病位点蛋白不可成药的难题，极大的拓展可药靶点的选择并从中心法则更上游的环节进行调控，从而有望开发出更有效的新型药物。然而复杂且庞大的序列和结构空间使得科学家对 RNA 空间的探索十分局限，RNA 体系的复杂性也使得相关实验数据可重复性差，整体数量少，整合程度低。因此科学家急切需要一种能够高效全面探索描述 RNA 空间的计算工具，以实现对 RNA 研究的数字化革新。

从 GPT 到 RNA，科学专属大模型技术路线潜力得到充分展现

自 ChatGPT 发布以来，预训练通用大模型以惊人的速度不断迭代和更新，展现了前所未有的能力。但对于具体的科学数据和信息，我们往往需要深入结合学科底层特性，以其特殊的"语言" 作为载体，如在生命科学中以蛋白序列和核酸序列这样有别于一般自然语言的方式来承载信息。

在这样的背景下，Uni-RNA 应运而生。Uni-RNA 利用约 10 亿条高质量 RNA 序列进行了大规模的预训练，几乎涵盖了所有 RNA 空间，充分挖掘了 RNA 序列的潜在信息。通过在广泛的下游任务中微调模型，Uni-RNA 在 RNA 结构预测、mRNA 序列性质预测和 RNA 功能预测等三个 RNA 领域的七个主要任务中全部取得了领先的结果，更是为未来 RNA 领域研究的深度革新提供了无限可能。（见下图）

为了利用大规模无标签数据进行预训练，我们构造了一个大规模的 RNA 序列数据集，通过预训练模型框架，结合有效的预训练任务策略，在大规模分布式集群上进行了预训练模型的训练。

RNA 序列数据集基于现有公开的 RNA 序列相关数据库构建。经过长度截断和去冗余得到高质量的 RNA 序列相关数据，并将其统一映射到 DNA 字母表中，最终数据集包含约 10 亿条 RNA 序列，几乎覆盖了目前所有 RNA 序列相关的数据，为预训练提供了广阔的样本空间，充分挖掘 RNA 序列的潜在信息。

Uni-RNA 全面针对 RNA 优化了经典 BERT 预训练网络框架，使用自研 CUDA 算子和 Flash Attention 等先进技术。相比于传统 Transformer 模型，Uni-RNA 的训练速度提高 5 倍以上，并能自适应 RNA 序列空间的特点，使得预训练更加有的放矢，行之有效，从而在所有下游应用中的表现都取得突破性提升。

Uni-RNA 的出现深入挖掘了 RNA 序列中的信息，为 RNA 相关领域提供了一个基于预训练的新颖的基础设施，也将为 RNA 研究领域提供新的研究范式，构建了 AI for RNA 的"安卓系统"。将为下游任务"App"，如 mRNA 疫苗设计、RNA 结构预测、ASO 开发、SiRNA 疗法创新、靶向 RNA 小分子开发、Aptamer 研发等众多关键难题提供全新的解决方案。目前，Uni-RNA 文章已发布在预印本网站上[1]

Source:

[1] https://www.biorxiv.org/content/10.1101/2023.07.11.548588v1

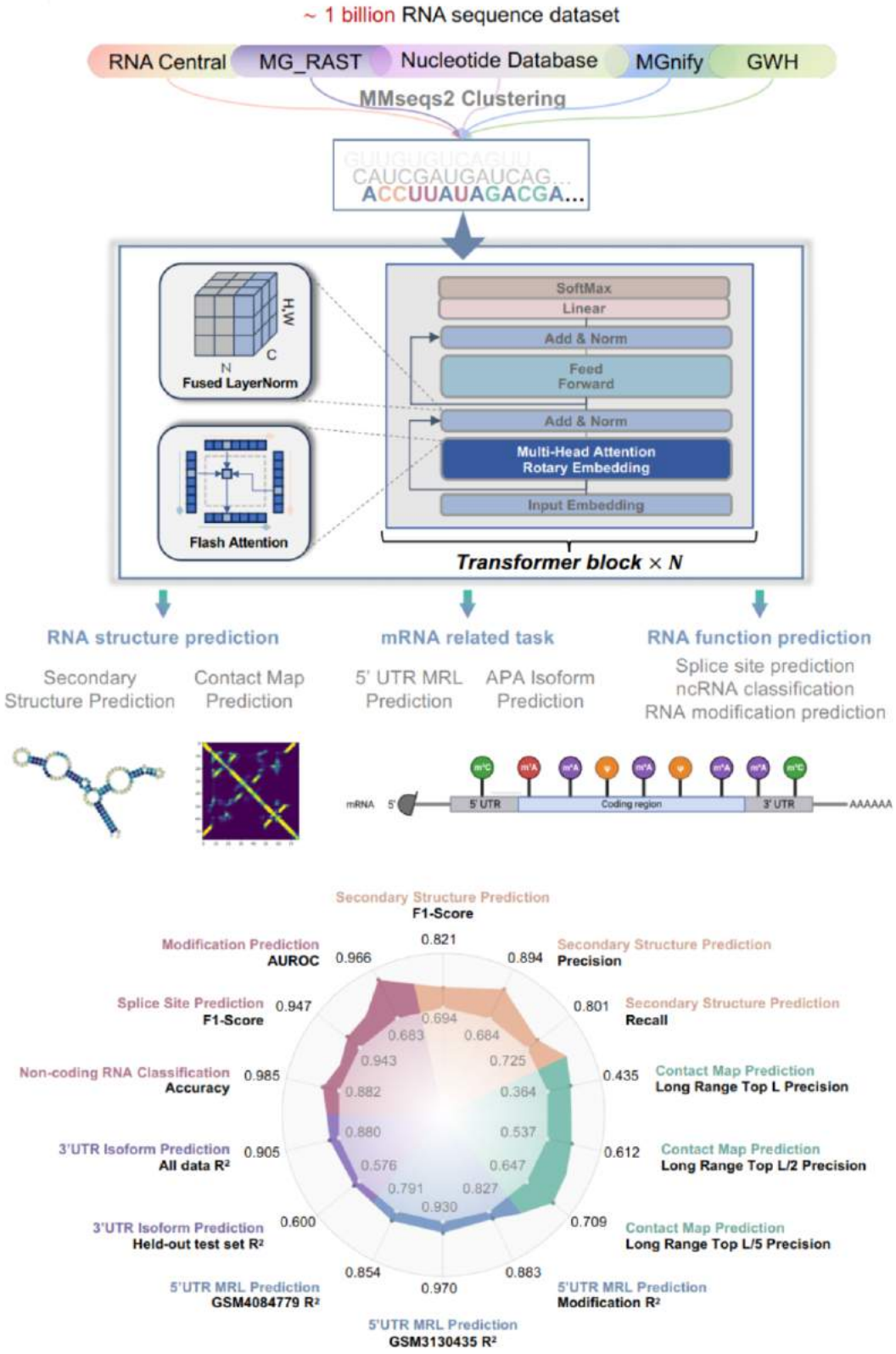


图. Uni-RNA 在结构、性质预测等各项下游任务中达到 SoTA

2.2.3 AI4S 驱动制药行业的 De Novo Design

当我们畅想未来，我们会发现，上述所有 AI4S 工具和范式的落实及其在药物研发中的深度应用或许有朝一日能对药物研发的效率和投入产出比带来系统性提升。

对小分子药物研发而言：

- 蛋白质折叠（Uni-Fold）、电镜结构解析（Uni-EM）及在此基础上的强化动力学模拟（RiD）将系统解决靶蛋白结构建模和功能性结合位点预测问题。基于此，基于 AI 的预训练模型系统解决"分子从哪来的问题"；
- 经 AI 优化的高精度的自由能计算（Uni-FEP）和预训练的成药性预测模型系统回答"分子如何评估的问题"；
- 基于逆合成的方法再系统回答"实体分子如何获取"甚至"如何确定制备工艺"的问题；
- 继而再进入 AI 赋能的晶型和剂型预测，获得高质量的临床候选分子，从而以 AI4S 工作流的形式，实现真正意义上的"理性药物设计"。

而对生物药的药物研发而言：

靶蛋白结构建模和功能性结合位点的逻辑不变。

另外生物大分子的研发也发生了改变：

- 前述 Diffusion, Folding、Uni-EM 以及 RiD 的出现不仅仅是对蛋白结构和结合位点发现有帮助，并且也有助于蛋白相关药物的设计；
- 对于 DNA/RNA 等遗传物质的折叠预测，则可以促进如 RNA 疫苗或药物的设计，影响最终的蛋白；
- 蛋白相关药物的设计则可通过深度学习来预测 PPI（蛋白-蛋白相互作用），可以认为是 Protein FEP，促进 MHC 的预测和抗原抗体结合的预测；
- 另外点突变可以探索蛋白/多肽的亲和力、人源化、稳定性等性质的设计，改造后的蛋白可以通过 Folding 的方法进行折叠，通过 Protein FEP 来计算与靶点的亲和力作用情况。

在此过程中可探索蛋白序列-结构-功能的对应关系，最终可达到根据靶点，以及科学家想要的功能来设计对应的序列，将少量大分子候选药物进行实验验证，降低了试错的实验成本，并且能够拓宽科学家对大分子设计和优化的天花板。另外，也可以探索诸如"反向找靶"等设计思路，充分利用数据和原理的结合，对现有药物进行再开发。

再进一步，如果能够通过 AI4S 的应用探索，对生物体底层代谢通路和相关蛋白探索清楚，那么 ADMET 及更多药效实验则有望通过计算蛋白与药物分子的相互作用来计算，对大分子药物的生产以及合成生物学的应用带来潜在的突破性进展。

Source：
1. Frost & Sullivan, Pharmaceuticals market research reports by Frost & Sullivan
2. 亿欧智库，2021 中国 AI/计算制药产业报告：药物发现篇
3. Emersion Insghts, Investors Double Down on AI Drug Developmen8
4. Paul, Steven M.; Mytelka, Daniel S.,(2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, (0), 540–540. doi:10.1038/nrd3078
5. Lindorff-Larsen K, Piana S, Dror R O, et al. How fast-folding proteins fold[J]. Science, 2011, 334(6055): 517-520
6. 李雨晨、刘海涛，四大顶级专家谈 AlphaFold2：关于技术细节、局限与意义，医健 AI 掘金志
7. 文猛，芝加哥丰田计算所教授许锦波：深度学习已颠覆了蛋白质结构预测，新浪科技
8. 孙卫涛. 蛋白质结构动力学研究进展[J]. 力学进展, 2009, 39(2): 129-153.
9. Lindorff-Larsen K, Piana S, Dror R O, et al. How fast-folding proteins fold[J]. Science, 2011, 334(6055): 517-520
10. Heo, L. & Feig, M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proc. Natl Acad. Sci. USA 115, 13276–13281 (2018) .
11. Wang D, Wang Y, Chang J, et al. Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics[J]. Nature Computational Science, 2022, 2(1): 20-29.
12. Dekker J, Belmont A S, Guttman M, et al. 4D Nucleome Network[J]. The 4D nucleome project. Nature, 2017, 549: 219-226.
13. Townshend R J L, Eismann S, Watkins A M, et al. Geometric deep learning of RNA structure[J]. Science, 2021, 373(6558): 1047-1051.
14. Nature Communications RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning, Nature Communications 10, 5407 (2019).
15. Xiong P, Wu R, Zhan J, et al. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement[J]. Nature Communications, 2021, 12(1): 1-11.
16. Phone , 知乎，alphafold
17. 施一公：AlphaFold 蛋白结构预测是本世纪最重要的科学突破之一，DeepTech
18. Ewen Callaway，Nature，'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures
19. Jumper J , Evans R , Pritzel A , et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021:1-11.
20. Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021)
21. AlphaFold 的原理和展望 - 钟博子韬 | 钰沐菡 公益公开课）
22. Deepmind 官网
23. Wang D, Wu S, Wang D, et al. The importance of the compact disordered state in the fuzzy interactions between intrinsically disordered proteins[J]. Chemical science, 2022, 13(8): 2363-2377.
24. 孙卫涛. 蛋白质结构动力学研究进展[J]. 力学进展, 2009, 39(2): 129-153.
25. Matthaei J F, DiMaio F, Richards J J, et al. Designing two-dimensional protein arrays through fusion of multimers and interface mutations[J]. Nano letters, 2015, 15(8): 5235-5239.
26. Ewen Callaway, Nature 新解读，AlphaFold 和 AI 蛋白质折叠革命的下一步是什么？水木未来资讯公众号
27. EMILY WALTZ, What AI and can't–do in the race for a coronavirus vaccine, IEEE Spectrum.
28. 苏尚，卢培龙：人工设计蛋白质将怎样改变人类生活？返朴的博客
29. Zhou G, Gao Z, Ding Q, et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework[J]. 2022.
30. Lengauer, Thomas, and Matthias Rarey. "Computational methods for biomolecular docking." Current opinion in structural biology 6.3 (1996): 402-406.
31. Chipot, Christophe, and Andrew Pohorille. "Free energy calculations." Springer series in chemical physics 86 (2007): 159-184.
32. Yuning Shen, Julia E. Borowski., Automation and computer-assisted planning for chemical synthesis, Nat. Rev. Methods Primers, 2021, 1, 23, DOI: 10.1038/s43586-021-00022-5
33. 张弈. 科里教授和逆合成分析法[J]. 化学教育, 1992 (1): 4-6.
34. 芸徽, 药物剂型预测, 博客.
35. Wang W, Ye Z, Gao H, et al. Computational pharmaceutics-A new paradigm of drug delivery[J]. Journal of Controlled Release, 2021, 338: 119-136.
36. 深势科技, 蛋白复合物结构预测上线 Hermite™ Uni-Fold：从头训练，精度超越 AlphaFold, https://mp.weixin.qq.com/s/kHFWtYHcnZEKHwiHWXokcw

2.3 合成生物学与现代农业

2.3.1 AI4S 在合成生物学中的应用实践

合成生物学是一门基于对生物机制的理解，通过工程化修改和重设计生物（通常是微生物）而使之产生新的功能的学科。合成生物学家合理的组合和利用自然界中已存在的生物机制来解决药物，生产，和农业中的诸多重要问题[1]。

合成生物学目前已应用于医疗健康、视频、化学品、环境监控及农业领域，包括利用糖类合成医药产品，人造食品，生物肥料，工业酶，生物燃料等用途，随着技术的发展，应用场景也在不断地拓宽：比如材料和医疗。人工合成胰岛素、角鲨烷、皮革和真丝替代物推进了动物保护。开发抗疟疾药物、以及通过基因驱动法来消除携带登革热的蚊子等，都是合成生物学对医疗行业的助力。

合成生物学被中国"十四五"规划列为科技前沿攻关方向、被美国国防部誉为未来"重点关注的六大颠覆性基础研究领域"。

产业应用以微生物细胞工厂为核心，建立"原料输入—菌株培育—发酵控制—提取纯化—产品输出"的工艺路线。在 20 世纪 90 年代之前，主要通过非理性诱变及筛选技术获得目标产物高产菌株，"以时间（人力） 换水平"。20 世纪 90 年代以来，代谢工程学科逐步创立，利用重组 DNA 技术对生物体中已知的代谢途径进行有目的的设计，构建具有特定功能的细胞工厂。但由于微生物代谢网络及其调控机制的复杂性，仍然需要耗费大量的时间和精力[2]。

从工程的角度讲，合成生物学家们遵循工程设计原则，旨在以工程的可预测性原理为指导，设计并控制复杂的生物系统。

合成生物学的工程设计原则包含：使用准化的遗传元件，以及遵循"设计-构建-测试-学习循环"（the DBTL cycle），从而实现预期结果。

合成生物学的 DBTL 循环的四个阶段如下所示[4]：

1. 设计：给定一个特定的预期设计目的，提出 (hypothesize) 一段 DNA 序列或者一系列细胞层面的操作
2. 构建：在生物系统上实施这些设计。这主要包括合成此 DNA 序列及将其转化进细胞的基因组内。
3. 测试：（等待工程转化好的"人造"细胞进行生物学过程，比如表达基因或对外界刺激作出反应）收集数据以检查测量出的表现型于预期之间的统一度，并评估任何脱靶现象和未预见的副作用。
4. 学习：利用测试数据来学习，检查设计组是否比随机搜索的方式能更有效地达到预期目标。人工智能 (AI) 可用于为下一循环的设计提供信息，从而更快的达到预期目的。

生物机制的研究是合成生物学发展的基石，AI4S 在解决生物机制问题上有着独特的优势（见 2.1 节）

下一代测序技术的发展和提高所带来的高通量组学对生物机制研究方法，使科学家对于基因到功能的了解不断细化。我们对可利用生物原件的知识库在不断增加，这一点作为驱动力推进了合成生物学近年来的飞速发展奠定了科学基础。

与此同时，生命科学领域的基本原理尚未系统性建构，我们对于生物机制的全貌仍不了解，我们只能通过对黑盒的干扰-观察的方式来对这个复杂的编码调控系统进行学习。合成生物学提供了探测黑盒的工具，而 AI 恰好可以通过以巨大参数空间对复杂系统的输入和输出进行拟合的方式，战略性地设计探测实验，更加高效的探索高维参数空间，从加速我们对生物系统的了解和应用。

科学家们希望通过像蛋白质那样建立数据库，来积累更多的可用数据，用来辅助探索包括基因到功能、代谢机制网络的内在科学原理，以及希望通过构建全细胞模型来模拟细胞内的代谢情况，AI 对于大规模数据的处理能力使得帮我们探索基因到功能，再到代谢网络，以及到表型的内在科学原理变得更易实现。

AI4S 的设计与合成生物学的工程模式相辅相成，共同造就了假设、构建、测试、学习的闭环（DBTL）

当然还有可能基于自动化实验室利用 AI 对于实验的定量设计，来反向补充数据库，以及探索更多的规律，突破理性设计的瓶颈。最终可实现全基因组水平定制化工厂的设想，完成实验室的小规模理性化设计到大规模制备工艺的生产。甚至有可能实现从底层对于微生物系统的模拟搭建，到宏观的海洋系统甚至生态系统的模拟，实现对物种分布和梯度影响的生态过程有所模拟和了解，最终我们可以真正推演出类似"蝴蝶效应"的过程。

研发环节	产业问题	问题难点	科学问题	AI4S 应用
生物设计	根据设计需求，选择基因-产物的生成路径和生物机理	生物元件自然界获得和设计基因组及基因路线的确定	生物元件序列与功能对应关系不清晰	可基于生物测序和信息挖掘，获得或设计启动子或终止子等生物元件资源[5]。 DNA 序列的基因表达预测变异效应预测增强子-启动子相互作用[6]
底盘细胞	选择保证生物设计目标表达效果的底盘细胞	原有的代谢网络与调控机制对于组装入的基因组表达的影响不清晰	代谢网络和调控机制原理不清晰	通过收集整理的大量数据可以利用生物信息学和人工智能机器学习等相关技术进行分析以及构建数学模型，如基因组尺度的代谢网络模型和全细胞模型[7]。

在实践中，AI4S 已经在合成生物学的一些领域展现其独特的潜力

I. AI4S 与酶改造

酶是一种具有催化功能的蛋白质。随着高性能计算技术、量子力学和分子动力学、深度学习理论和方法的发展，以及 AI4S 为力场精度和采样效率带来的改变，蛋白质设计的技术在酶催化方面发挥出巨大作用，可以通过设计合成的方法来寻找更多的核心元件，或者通过点突变方法来探索序列与酶的活性、稳定性等性质的对应关系，使用 Folding 方法对最终的结构预测来调整对酶的设计路线。

比如加州理工学院的 Stephen L.Mayo 课题组利用计算迭代方法从非活性蛋白质支架 HG-1 为起始点计算，获得的 8 种设计酶均表现出显著的催化活性，大幅提高了计算设计酶分子的成功率[8]。

如果在 AI4S 助力下，对底盘细胞的代谢途径和调控机制，以及所插入的基因和基因路线对底盘细胞影响规律探索到一定程度以后，对酶的改造将更加趋向"理性设计"，在全基因组范围内搜索或设计改造基因型，通过 Folding 预测酶蛋白的结构，来调整基因设计获得所需的性质的酶，再通过建模高精度模拟酶促反应来验证酶活性和稳定性等性质，甚至在生产工艺中也可以有 AI4S 的介入来优化生产工艺，最终减少人员和成本上的消耗。

II. 生物基化学品的生成

生物基化学品是利用可再生的生物质（淀粉、葡萄糖、木质纤维素）为原料生产的大宗化学品和精细化学品。可以替代以石油、天然气和煤等为原料生产的化学品，减少对石油等资源的依赖以及对温室气体的排放。AI4S 可帮助寻找微生物基因型与生产化学品分子结构的对应关系，从而为理性化设计生产所需生物基化学品奠定基础

III. 生物燃料开发助力能源问题

化石燃料这种非可再生能源，无法满足全球 70 亿人口的需求，且会造成对大自然的污染。大自然中有一些能源植物，分泌的汁液或提取液与石油成分相似，不仅能够提供新的能源，还能降低对环境的污染。合成生物学能够对能源作物的基因和遗传物质进行改良，并且也能够在优化微生物转化过程，也可根据需求理性化设计获得所需要的生物燃料或化学产物。AI4S 可以通过深度学习辅助探索基因与最终产品的对应关系，最终达到真正的理性化设计来获得人们所需要的更优质的生物燃料产物。

IV. AI4S 与人造食品

Hoxton Farms 正在利用计算生物学和机器学习来生产细胞培养的动物脂肪，使用"数字孪生"迭代模拟和改进人造肉的成本和风味。

Imagindairy 利用机器学习和分子进化计算建模，将牛蛋白语言翻译成酵母的语言，尝试使用酵母生产牛奶蛋白，并通过这种方法编码风味、营养价值和香气等特征的表达。[10]

Source:

1) Guoping,Zhao. Synthetic Biology: Unsealing the Convergence Era of Life Science Research[J].Bulletin of Chinese Academy of Sciences,2018,33(11):1135-1149.

2) 中信证券，王**喆**，合成生物行业深度报告：合成生物学蓬勃发展，市场空间广阔

3) 华安证券，刘万鹏，合成生物学——属于未来的生产方式。

4) 生物细胞工厂的设计构建：从诱变育种到全基因组定制化创制

5) 赵国屏. 合成生物学:开启生命科学"会聚"研究新时代. 中国科学院院刊, 2018, 33(11): 1135-1149

6) Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions[J]. Nature methods, 2021, 18(10): 1196-1203.

7) 杨永富, 耿碧男, 宋皓月, 等. 合成生物学时代基于非模式细菌的工业底盘细胞研究现状与展望[J]. 生物工程学报, 2021, 37(3): 874-910.

8) Privett H K, Kiss G, Lee T M, et al. Iterative approach to computational enzyme design[J]. Proceedings of the National Academy of Sciences, 2012, 109(10): 3790-3795.

9) 贤集网，玮龙，Clara Foods：人造鸡蛋白的好处可谓数不胜数

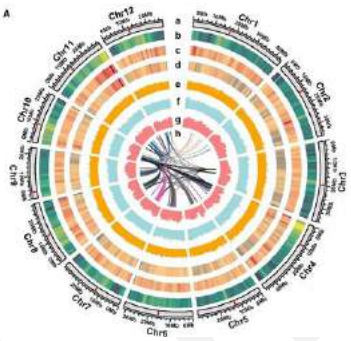
10) Atypical AI food，Let's examine some abnormal artificial intelligence applications within the world of food

2.3.2 AI4S 在现代农业的应用

I. 科学育种

从古至今，人们一直想要培育出集抗性强、优质、高产等性状为一体的作物品种，而以往想要培育出想要的品种表型，需要进行大规模的田间试验，才能预测基因变异后的作物表达生长情况，不仅容易受环境、气候的影响，还效率低，成本高，田间管理难度大。极端天气出现的频率不断变高，影响到粮食的产量，对全球食物保障造成了重大挑战。

我国利用基因组学、遗传学和分子生物学，在国际上率先构建水稻全基因组序列框架图（下图），克隆了一批调控株型、氮高效利用、耐低温、抗旱、耐盐碱、抗病、新型抗除草剂等具有重大育种价值的新基因，并在育种中逐步加以利用。但依然存在着比如：重大原创基础理论研究不够深入、关键技术和核心技术掌握不够、企业研发投入不够、自助创新产业化需要突破等问题[1]。



图：111 份水稻种质的水稻泛基因组特征

想要研发出抵抗恶劣环境且产量高特性的作物，一般研发周期会长达几年或数十年，传统的科学育种经历了主要依赖经验和统计学、数量遗传学和杂交水稻的阶段，再到引入基因工程和分子标记的阶段，这期间依然会出现大量的数据，依靠人脑和实验的力量难以抵抗目前所面临的问题。

随着 AI4S 发展，一方面可以对植物的基因型与表型的大数据进行快速积累，另一方面可借助深度学习的方法对作物基因与表型对应关系的探索，有可能实现作物调控基因的快速挖掘，以及对表型的精准预测。AI4S 将人工智能的引入，以及合成生物学的发展将有望给科学育种带来更大的突破和进展。

在实践中，Yield10 公司实现了用 AI 挖掘与作物代谢相关的目标基因，再用 CRISPR 技术进行目标基因改造，提升作物产量。

他们开发了 Gene Ranking Artificial Intelligence Network (GRAIN)算法，挖掘出了四个可能让 camelina 籽含油量提高的基因，且在使用 CRISPR 技术提高该基因活性后的确使 camelina 籽含油量提升 10%。注明的农业公司孟山都，通过人工智能筛选后，只需对最具开发潜力的品种分子进行田间测试，即可帮助农民增收[2]。

但这种基于数据的探索依然还有着一些障碍，比如不同作物表型数据差异性较大，不同人采集的数据的真实可靠性与准确性也很难统一控制。拥有大量数据的研究单位或公司对数据并不共享，可比较的数据量少[3]。而深度学习中预处理模型对于小数据集的处理具有一定的优势，可能可以利用少量数据的学习解决这个难题。也有越来越多的高科技初创公司加入了育种行业的发展中，比如英国的 Phytoform Labs 和 Tropic biosciences，美国的 Inscripta，以色列的 Seed-X，以及中国的古奥基因和博睿迪等。AI4S 可能是解决科学育种和食物短缺的必经之路。

II. 化肥研发

化肥是粮食的“粮食”，是现代农业的物质支撑及科技成果，打破了传统农业依赖自然恢复的瓶颈。[4]但化肥生产中仍然存在较多的问题，如如何提高化肥效能，如何减少化肥生产过程中污染，同时提高化肥营养含量等，都限制了化肥在农业生产中的进一步应用，AI4S 成为化肥研发、生产转型中重要工具。

氮肥是化肥中的重要肥料之一，也是我国主要应用的肥料之一，在农业生产中发挥重要的作用，其生产化学方程式如下：

$$2NH_3\text{（液氨）} + CO2 \rightarrow NH_2COONH_4\text{（尿素）} \rightarrow CO(NH_2)_2 + H_2O$$

现有挑战	AI4S 展望
氮过度使用造成的染	AI4S 可以结合环境信息、土壤信息和作物自身生长等信息，进行肥料的合理供应管理。伦敦帝国理工学院借助 AI 技术开发了一个可以分析土壤数据的模型，可以预测土壤中氮水平的变化趋势，并结合气候预测，从而准确确定和预测两者对施肥计划的影响，根据作物需求调整安排，减少过度施肥，并提高作物产量[5]
在水溶液体系中中，水参与反应及对反应的影响具有较高的复杂性	借助机器学习势和高性能计算，可以在微观尺度研究氮肥生产过程中水溶液里的反应，探索水对于生产氮肥的微观影响（如水溶液中尿素分解），找到潜在的提升效率的方式
氮素利用率不高，合成路径中损失较多	AI4S 可研究掺杂其他成分对现有氮肥氮素利用率的影响，探索减少氮素损失的路径，提升作物的质量。如有团队对氮肥掺杂纳米镁，提升了氮肥的利用率（~ 5%），未来 AI4S 有望帮助科研人员进一步探索微观机制，从而找到更多好的添加剂，进而促进农业生产[6]
氮肥生产水资源浪费严重	AI4S 能探究如电化学处理、研究新的渗透膜/吸附剂等新的处理氮肥生产废水的方式，从而赋能绿色氮肥生产[7]

III. 植保研究

农药是确保农业稳产、丰收、保证全球粮食供应必不可少的重要生产资料，其对农业的发展具有重大的意义：减少杂草、害虫对农作物的伤害，提高产量。然而由于农药对人体和环境存在一定的不良影响，新农药的开发与应用成为农药发展的主流。

农药通常分为化学农药和生物农药，农药市场的主体一直以来都是以化学农药为主导，其占比大约是93%。研发农药除了要考虑基本用途之外，还要照顾到安全性和环境友好性。现阶段农药研发的核心问题是：耐药性，因自然环境而导致的药物沉降、蒸发以及失效等。发展高效、低毒、低残留的新型环保农药将成为农药研发的焦点和主流趋势。

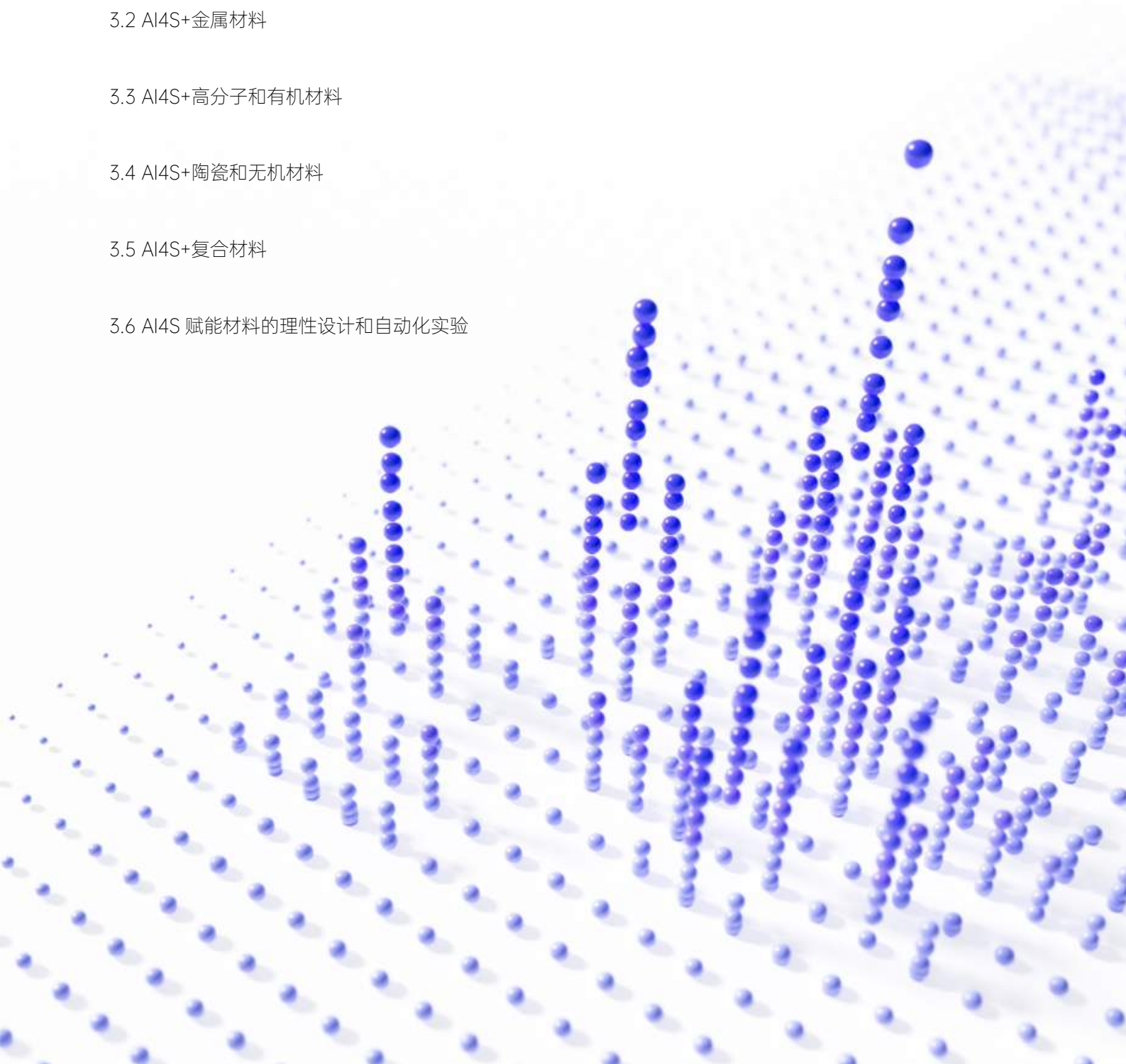
AI4S 潜在研发方向展望 [8]:

- 研究新的农药剂型和特殊剂型。农药制剂的研究开发应从粗略到精确转变，控释技术，纳米技术等都是新的研发热点
- 从害物靶标出发经仿生合成开发新药物
- 以具有农药活性的生物源物质为先导进行仿生合成
- 对已知天然物结构修饰，开发活性更高或更安全的农药。如通过改造提高药性、降低毒性、化废为宝
- 实施农药与医药的双向开发，由于农药与医药的开发应用有很多相通性，因此在新农药或医药品研发中，往往借助双方的化学结构和理论进行双向开发
- 运用新的筛选方法进行农药开发，提高新药开发水平和效率。如运用高通量筛选、构效关系研究、虚拟设计、 新颖剂型；运用代谢组学的高效生物测试方法；运用杀虫剂离体筛选的组织水平法、细胞水平筛选法、分子水平法等

Source:

1. 经济日报，有序推进生物育种产业化应用
2. Peter Rogers, FreeThink, How artificial intelligence is boosting crop yield to feed the world
3. Washburn JD, Mejia-Guerra MK, Ramstein G., Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci U S A. 2019 Mar 19;116(12):5542-5549.
4. 张福锁，中国农业大学，科学认识化肥的作用，
5. Max Grell, Giandrin Barandun, et al. Point-of-use sensors and machine learning enable low-cost determination of soil nitrogen[J]. Nature Food, 2021, 11(2):981-989.
6. 赵灿, 刘光明, 戴其根, 许轲, 高辉, 霍中洋. 氮肥对水稻产量、品质和氮利用效率的影响研究进展[J]. 中国稻米, 2022, 28(1): 48-53.
7. 汪家铭，氮肥生产循环冷却水排污水双膜法处理工艺及应用，上海化工，2014
8. 于春雷, 张全. 农药研发趋势研究[J]. 农业科学（2630-4678）, 2019, 2(1):
9. Mohammed Eslami, Aaron Adler, Rajmonda S. Caceres, Joshua G. Dunn, Nancy Kelley-Loughnane, Vanessa A. Varaljay, and Hector Garcia Martin. 2022. Artificial intelligence for synthetic biology. Commun. ACM 65, 5 (May 2022), 88–97. https://doi.org/10.1145/3500922

第三章：AI for Material Science 原理与实践



3.1 材料研发的核心是建立准确的构效关系

材料的演进驱动着历史的发展，对材料的利用能力是人类生产力和生产方式的标志性体现。人们约定俗成，通过材料进行历史阶段的划分：如石器时代、铜器时代、铁器时代到如今的电子信息时代[1]。人类的衣、食、住、行、娱乐及我们日常生活的每个部分也都离不开各种材料。材料被称为"工业骨骼"，在工业发展中起到基础性和先导性作用，材料科学与技术的发展为各行业的技术进步都形成了巨大的助力，如医疗、交通、能源、信息、国家安全、航空航天等；另一方面，各工业领域的发展也对更高性能材料产品的持续需求，反过来推动了材料科学和技术的进一步发展。

材料常被划分为：有机材料和无机材料两大类，其中无机材料又常被分为金属材料和无机非金属材料两大类。此外复合材料是近些年来逐渐发展起来的新材料体系。有机材料中存在两类典型化学键：分子内强共价键和分子间的范德华键，两者的作用都不可忽略。无机材料中范德华作用一般可以忽略，主要键合是原子间的金属键、离子键或共价键。

从使用性能分类，材料可以分为结构材料和功能材料。结构材料的主要作用是作为承力构件，主要关心材料的力学性能；功能材料则主要是利用材料特殊的物理性质，如声、光、电、磁、热等各类特性。

结构材料主要为金属材料和复合材料（树脂基复合材料、陶瓷基复合材料、碳基复合材料），主要用于大型结构件如飞机、火箭、汽车、建筑、桥梁、风力发电、核能等。相比金属材料而言，复合材料具有较好的比强度、高韧性、轻质、耐高温等各种优异性能。在一些高端结构件领域，金属材料正在被各种高性能复合材料所替代。如飞机中已经大规模采用复合材料。

功能材料主要利用材料的特殊物理性能，使用领域主要包括半导体、光电转换（太阳能电池、OLED等）、铁电、铁磁、介电、压电、热电、电容器、传感器、隔热、导热、光导等。很多领域需要结构功能一体化材料，如不锈钢、可降解植入器件、防热结构一体化（高超声速飞行器外壳）、波导防隔热一体化（火箭天线罩）、导热承载一体化（大功率功率器件基板）等。

材料的研发的核心是对"构效关系"的研究

更具体地说，是对"组分-结构-工艺-性能"之间的关系的研究，虽然材料科学家们对于"结构决定性能"这一点有高度的共识，但组分、加工工艺都对材料微观结构产生着复杂的影响，几者之间又存在着十分复杂的关联，这为材料的研发带来了巨大困难。

此外，材料的真实服役环境也往往非常复杂，通过理论研究其细节过于困难，需要做各种假设和简化。在服役环境下，往往涉及非均相多外场耦合的情况，如热、力、电、辐照同时作用于固体、液体混合的体系。因此，需要同时对材料和环境的物理、化学及力学性质进行耦合建模。

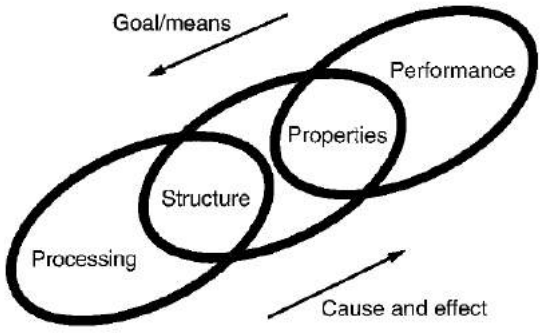


图. 材料的成分、工艺、微观结构和性能之间关系

先进制造、新能源、消费电子、航空航天等行业的飞速发展对新材料的研发提出了愈发严苛的要求。当前以实验为主的材料研发速度远远落后于终端产品的更新迭代速度

多样的需求 x 多样的技术栈 == 多样的研发需要

从终端产品的性能参数出发，对基础材料优异性质提出了越来越高的要求。粗略来讲，材料的性质一般分为三类：物理性质、化学性质和力学性质。物理性质包括材料的光学、电学、磁学、声学和热学性质，研究的是材料结构稳定前提下对外场的响应关系。化学性质包括电化学、燃烧、腐蚀等性质，是指材料与环境的相互作用导致的不可逆化学反应（物质改变）。力学性质包括弹性、塑性、强度、韧性、硬度、疲劳、蠕变等各种性质，既有可恢复状态下的性质（如弹性），也有材料不可恢复的性质（如塑性）。一般而言，物理性质和化学性质对功能材料至关重要，如耐烧蚀、耐腐蚀、铁电、铁磁、隔热、光电转换等。力学性质对结构材料至关重要，如疲劳、蠕变、强度等。

对于材料性质研究而言，所有对外场的响应都有其特征时间和空间尺度。因此理解材料的性质（即材料对外场的响应）需要采用不同尺度的模拟方法，如电子尺度、原子尺度、粗粒化及介观尺度、宏观尺度。一般而言，空间尺度与时间尺度呈现一定的正相关性，即空间尺度越大的行为其对应的特征时间尺度也会越大。因此，研究材料不同的性质可以采用不同尺度的模拟方法。例如，DFT、MD、CGMD、PF、FEM 等。

从微观到宏观，材料的设计和工艺环节都面临着一系列关键科学问题，但由于算法、平台的不成熟，数据库的不完善和计算资源的限制，传统材料研发难以真正实现对传统方法的突破，很大程度上依旧依靠经验乃至运气。历史上，许多材料是通过偶然发现的——如著名的不粘聚合物聚四氟乙烯[2]。现阶段，新材料的研发仍主要依靠实验试错。材料科学是跨学科领域，需要结合物理、化学、工程学等知识，从原子和分子出发，根据需求优化和设计材料，但是：

材料空间结构复杂，涉及多个尺度的结构特征；

材料组分复杂，且在多个尺度存在不均匀分布；

工艺复杂，工艺特异性要求多，成本高、时耗高；

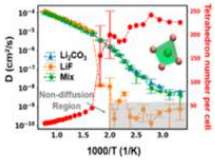
因此，需要将计算和实验组合，缩小材料创新的搜索空间，才能加速材料创新进程。AI4S 可以从第一性原理出发，预测材料成分分布和微观结构的多尺度特征；并整合现有材料的数据库、专家知识和人工智能方法。在不少领域，已有科研人员成功利用 AI4S 方法进行了关键问题的攻坚

（如下页图、表所示）

Source:

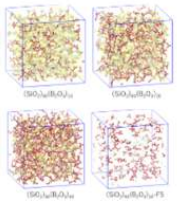
[1] Materials Science, https://en.wikipedia.org/wiki/Materials_science

[2] University of Liverpool, De Novo Materials Design, https://www.liverpool.ac.uk/materials-innovation-factory/research/de-novo-materials-design/



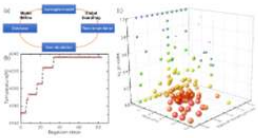
Solid Electrolyte Interphases Composition: of LiF – Li_2CO_3

J. Am. Chem. Soc. 2022. 2c11521



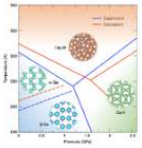
Codoping: Boron and Fluorine in silica glass

J. Phys. Chem. C 2022, 126, 2264−2275



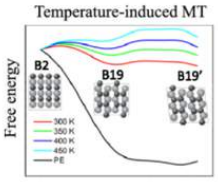
Refractory Material Search: witin Hf-Ta-C-N

arXiv:2211.03103v1



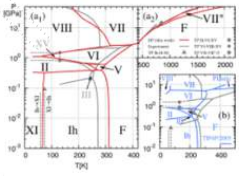
Galium: ab initio phase diagram

Nat. Commun. (2020) 11:2654



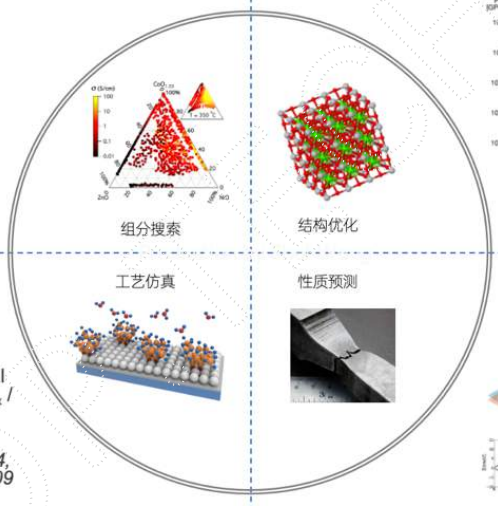
Shape Memory Alloy: martensitic transformations of NiTi

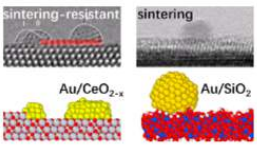
Acta Materialia 2022.118217



Water: ab initio phase diagram

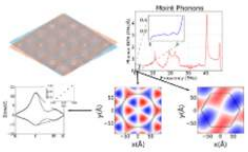
Phys. Rev. Lett. 126, 236001





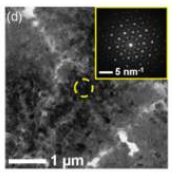
Sintering: Metal affinity of CeO_{2-x} / Au catalyst

J. Am. Chem. Soc. (2022), 144, 45, 20601–20609



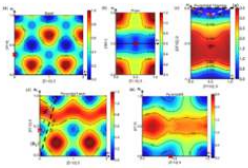
Graphene: Phonon band structure of magic-angle twisted bilayer graphene

Nano Lett. 2022, 22, 19, 7791–7797



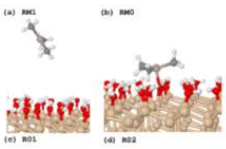
Laser Processing: Dendritic Quasicrystal Growth of Al-Cr thin film

Physical Review Letters 125, 195503 (2020)



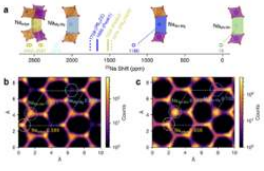
Titanium: defect and mechanical properties of Titanium

Npj Comput. Mater (2021) 7:206



Atomic Layer Deposition: $Al_2(CH_3)_6$ deposition on OH/Si(111)

ACS Appl. Mater. Interfaces 2022, 14, 22, 26116–26127



Battery Cathode: Paramagnetic shift of Na+ distribution in Na-Mg-Mn-O

Angew. Chem. Int. Ed. 2021, 60, 12547 – 12553

图. AI for Science 沿材料科学 "组分、工艺、结构、性能"的应用实例摘录

图表 4：材料研发领域的多尺度问题和 AI4S 示例

问题尺度	科学问题	AI4S 实践示例
电子层面	电子耦合；载流子输运；电子激发态；强关联作用	• 解二维纳米结构系统的电子输运问题； • 构建更精确电子密度和相互作用图
原子/分子层面	相变；离子迁移/扩散；多尺度缺陷；原子间相互作用；能量密度；力学性能；能量输运	• 模拟非晶态物质的复杂相变过程； • 研究电极反应机理； • 研究晶体缺陷形成机理； • 定量预测材料强度等性能； • 揭示材料力学性能的微观机理
微介观层面	微纳尺度力学；流体流变问题；界面优化和设计；固化过程；缺陷表征	• 高精度预测界面热传输； • 构筑设计材料微观结构； • 模拟材料微介观结构的演变； • 材料对对外场（如电场、磁场、应力场等）的响应
宏观层面	工业仿真；良率提升；工序优化；溶液渗透；	• 建设 IIoT（工业物联网）； • 多物理场耦合分析

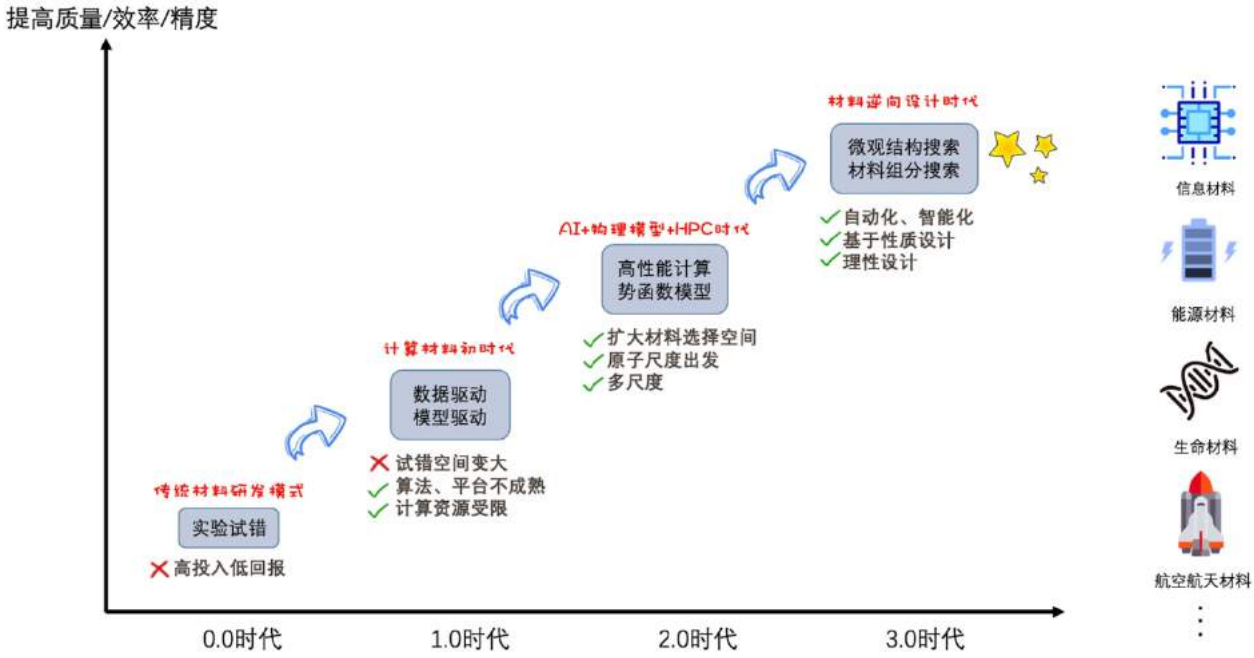


图. AI4S 驱动材料研发范式的不断演进

AI4S 实践（3）：深势团队荣获领域最高荣誉"Gordon Bell Prize"；并不断突破极限，实现 170 亿原子的第一性建模，将分子动力学带入新时代

在材料科学、生命科学及信息科学等领域，分子动力学是主流的研究手段，而如何高效的对真实场景的复杂体系进行分子动力学计算是决定研究质量的关键因素之一。对于纳米晶粒、储存材料、薄膜生长、蛋白聚合及病毒机理研究等体系，传统手段由于计算资源的限制，无法在快速且保持精度的前提下实现大体系模拟，只能依靠低效率的实验手段进行试错，而高性能计算的出现为大体系的计算模拟带来了解决手段。对于二维材料、催化材料、高熵合金等体系，传统的有限数据和经验函数为主的计算手段无法对缺陷、晶体结构、原子排布等复杂问题有效研究。

著名物理学家理查德·费曼曾说过："人类总是在不断地尝试理解生命，在这条道路上，一个最强大的假设就是：'所有的事物都是由原子组成的，生物所做的一切都可以通过原子的振动和摆动来理解'"。分子动力学方法在物理、化学、生物及材料等科学领域有着广泛的应用。在高性能计算的快速发展下，分子动力学方法已成为阐释复杂物理化学现象的重要工具。然而，计算效率和精度不可兼得是分子动力学方法长期面临的困扰。

深势团队在前世界第一超算 Summit 上，在保持第一性精度的前提下成功模拟了亿级原子的运动轨迹，将超大系统的分子动力学模拟带入了全新时代。此前最大规模的研究是 100 万个原子的体系，而深势团队则成功模拟 6.79 亿个水分子及 1.27 亿个铜原子，速度也比之前模拟体系提高了几个数量级，使原来可能需要 60 年才能完成的第一性原理精度模拟一个具有 1 亿个原子的体系缩短到了 1 天。在模拟中，打破了 35 年来的从头算分子动力学（AIMD）局限于小型系统（几千个原子的系统）的困境，效率为之前的万倍。

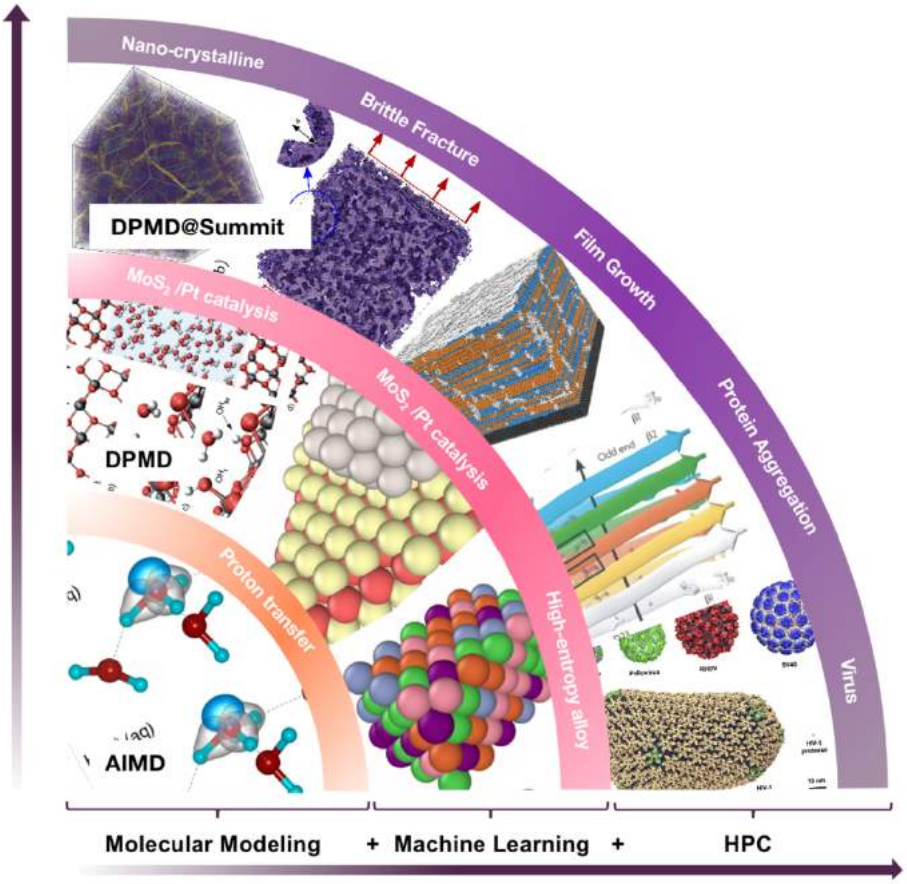


图. 深度势能实现微观尺度建模，解决材料、生命难题

团队充分利用美国能源部橡树岭国家实验室的 Summit 超级计算机，将超算的计算能力充分发挥，在双精度下实现了 91PFLOPS 的速度，在混合单/半精度下实现了 162/275PFLOPS 的速度，实现了全球首次亿级原子高精度模拟和高性能计算，是人类在原子模拟上迈出的“一大步”。同时，团队开启了“人工智能（AI）+高性能计算（HPC）+物理模型”研究新范式，将三个方法“并轨”，能实现从最本质的物理原理出发，高精度、高效率的解决困扰人们的科学问题、应用问题。一扫学界和业界对 AI、HPC 和分子材料模拟的阴霾和焦虑，是人类在理解世界运行原理上迈出的重要一步。该成果对于更好地集成机器学习和物理建模的下一代超级计算机也提出了新的挑战，实现用技术引领硬件发展。[1]

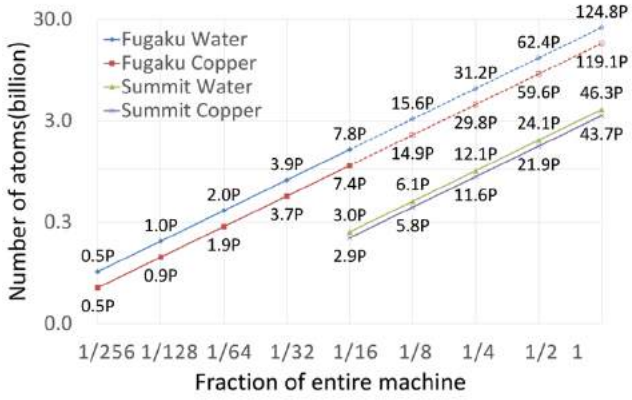


图. 深度势能将 Summit 和 Fugaku 超算发挥至极限 [2]

2022 年，深势团队再次报告在世界第一超算 Fugaku 上将计算规模再次扩大 134 倍，达到 170 亿原子规模。成果发表于第 27 届国际计算机学会 SIGPLAN 学术会议上，受到广泛关注。[2]

2023 年 4 月，来自 37 个机构的 47 名研究者共同完成了题为 DeePMD-kit v2: A software package for Deep Potential models 的论文，并在 arXiv 上预发表[3]。v2 的架构调整让多硬件平台支持成为了一件容易的事情，并与百度 PaddlePaddle 架构合作，促进了双方在开发者生态上的双向融合。此外，v2 最丰富的更新和调整，还是在于对模型架构的支持。例如，对于偶极、极化率等张量型物理量的训练和推断， v2 进行了系统重构，使得用户无需对算法有深入了解即可直接使用。这项功能使得若干光谱计算成为可能，例如 IR 光谱和 Raman 光谱。在此之外， v2 加入了原子类型嵌入功能，通过将原子类型嵌入特征空间，将嵌入网络和逼近网络对原子类型的依赖解放出来，达到仅需一个嵌入网络和一个逼近网络即可实现对多组元体系建模的目的。

关于“戈登•贝尔奖”

戈登贝尔奖设立于 1987 年，是国际高性能计算领域最高的学术奖项 ，也被誉为“超级计算应用领域的诺贝尔奖”。每年 11 月，美国计算机协会（ACM）将该奖授予最前沿的并行计算研究成果团队，以表彰他们在高性能计算（HPC）领域的杰出成就。

Source:

[1]. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. (SC '20). IEEE Press, Article 5, 1–14.

[2] Extending the limit of molecular dynamics with ab initio accuracy to 10 billion atoms. (PPoPP '22). Association for Computing Machinery, New York, NY, USA, 205–218. https://doi.org/10.1145/3503221.3508425

[3] DeePMD-kit v2: A software package for Deep Potential models. 2023. doi:10.48550/arXiv.2304.09409.

AI4S 实践（4）： DPA 原子间势能预训练大模型驱动性质预测和新科学发现

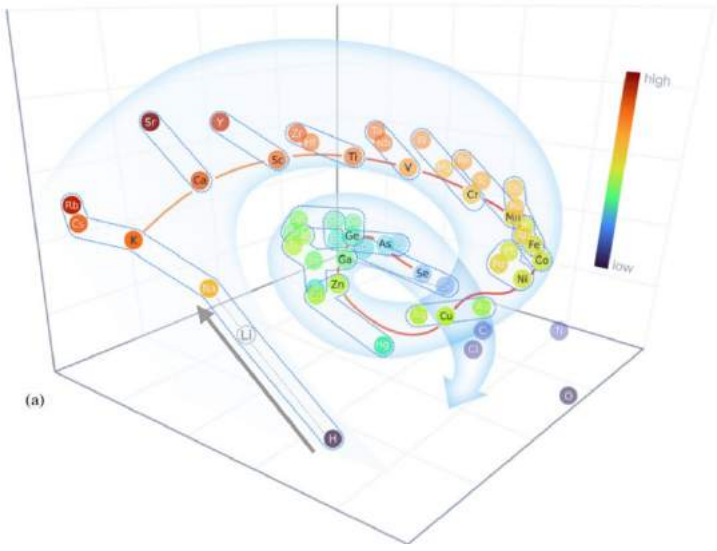
对于诸如量子力学、材料科学、化学和生物学等领域，势函数对于理解和解决问题至关重要。这种函数描述了一个物理系统的状态，以及物质粒子受到的力。但是，建立精确且高效的势函数是一项非常复杂的任务，需要对物理、数学和计算科学有深入的理解。

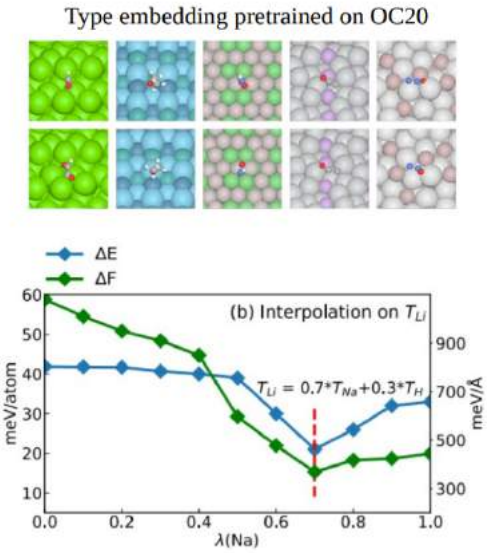
2022 年底，北京科学智能研究院联合深势科技，发布了首个覆盖元素周期表近 70 种元素的深度势能原子间势函数预训练模型—— DPA-1[1]，可模拟原子规模高至 100 亿，目前已在合金、半导体等应用场景中证明了其领先性和优越性，并展现出了在人工智能领域难得可贵的"可解释性"潜力。

研究团队指出，该模型确实是学到了一些元素的本质的知识，使得它能够获得更好的精度。比如一个简单的例子，就是镁铝酮这样一个例子。训练集是里面所有的单组元跟两组元的合金。而验证集不在训练集里面，验证集是三元的组分。研究者发现，在这样的设置下，DPA-1 的精度是远远好于标准 DP 模型。看上去只学单组元跟两组元，标准 DP 实际上是不太知道三组元合金里面会发生什么，DPA-1 实际上是已经通过单组元跟两组元的学习，知道了很多这些元素的构型和化学知识，在三组元的合金预测上已经达到了非常好的精度。这个 all 就是将所有的作为训练集。可以看到，相比于 baseline 来讲，其实它已经离得不太远。比如它已经学到了三组元里面的绝大多数的值。

这里面还有一个更有趣的观察，就是 OC20 里面，在碱金属这一组上面是没有锂的。如果它有锂，它应该是坐落在氢跟钠之间。研究者做了一个非常有意思的考察，就是对氢跟钠的隐空间的表示，做一个差值，看看能不能"发现"锂。事实表明它是能够一定程度上实现的。怎么考察？研究这看差值的系数，在 1 的时候是钠，在 0 的时候是氢，纵轴是它在一个独立的锂的测试集上面去直接去进行测试的精度。这基本就意味着如果锂在表示之中，那锂应该大概是在 0. 7 个钠加上 0. 3 个氢的这样的一个位置，说明做了这样的一个 embedding 以后，不仅是形状上看着像元素周期表，确实是也能够在这样的一个隐空间的表示中，正确的获取未知元素的预测。这种能驱动新科学发现的强大能力令人十分兴奋！

Source: [1] arXiv:2208.08236





3.2 金属材料中的 AI4S 应用

3.2.1 合金材料

合金材料也是人类社会发展的重要物质基础，小到螺丝钉，大到航母、飞船，每一个都需要各类型的合金材料。合金材料从方方面面影响着人类的历史发展进程和生产生活。为了更好地了解合金材料，可以粗略地将常见金属材料按照其母体合金元素特性分为两大类。一类是可以产生结构相变的金属元素，如 Fe、Ti、Zr 等，一类是自身无结构相变的元素，如 Al、Mg、Ni 等。一般而言，自身具有结构相变的材料将带来更丰富的可能性，微观组织的可设计性非常强。例如，钢的种类就非常多，铁素体不锈钢、奥氏体不锈钢等。调节金属材料的微观结构是提升现有金属材料及设计新合金的重要手段，其主要方式有 “相变调节” 和 “压力加工”调节：通过这些手段改变材料内部的微观结构。

合金研究的主要关注点是“成分+加工工艺”，目标是“强度、塑性、抗蠕变性能”等材料的真实服役性能。从科学层面上抽象，主要分析工具是热力学、动力学及各种模拟工具，最终目标是落实到微观组织。实际上，仅仅是微观组织并不足以完全确定材料的力学性能，还涉及很多更低层次的问题，比如元素在界面的偏聚、在位错的偏聚、原子尺度的局部有序等。这些因素介观模型无法考虑，更不可能预测，但是对材料性能有着本质性影响。例如、高温抗蠕变性能、屈服应力、服役过程中的脆化（如 H 脆、S 脆、P 脆）等各种问题。因此，需要从微观尺度和原子尺度均进行设计才能做好高性能材料，尤其是一些高端金属材料。这些问题的处理主要依赖 DFT+MD 的原子尺度方法。AI4S 有望从微观出发，为金属材料的研发和生产提供全新解决方案与思路。

参考资料：
[1] 苏娅、郭建波，《金属材料在人类社会中的作用及应用》
[2] Determination of γ-γ' lattice misfit in a single-crystal nickel-based superalloy using convergent beam electron diffraction aided by finite element calculations, DOI:10.1016/j.micron.2011.10.009
[3] Segregation of tungsten at γ′(L12)/γ(fcc) interfaces in a Ni-based superalloy DOI:10.1063/1.3026745

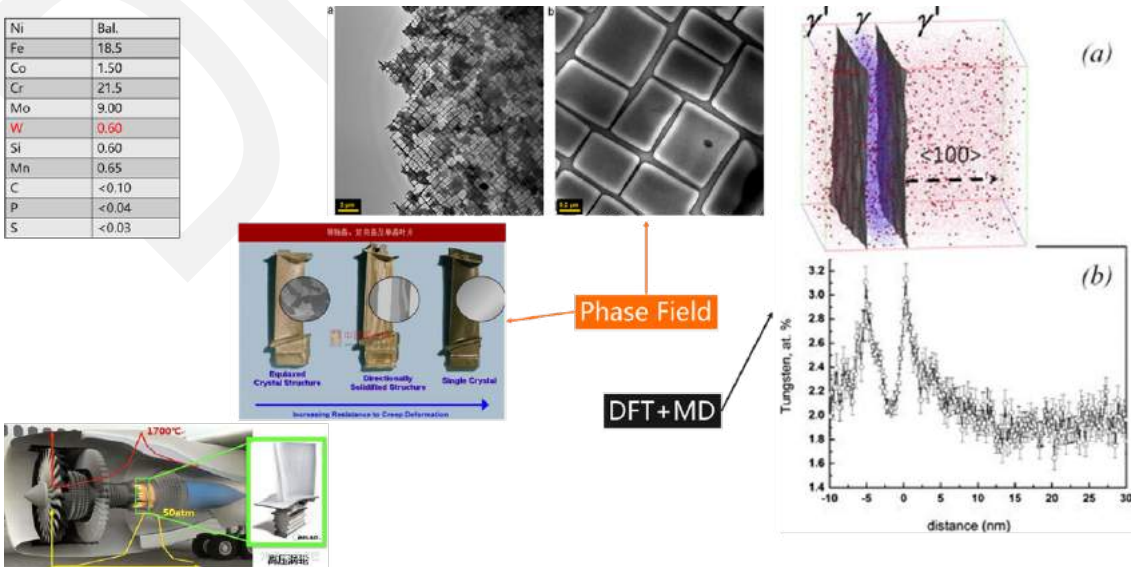


图. 金属材料中的物理建模 [1,2,3]

AI4S 实践（5）：《Nature》 正刊报道 AI4S 助力高熵合金纳米颗粒的设计与工艺仿真

2023 年，来自武汉大学与南方科技大学的研究团队，在液态金属领域取得重要进展，其论文以 “Liquid metal for high-entropy alloy nanoparticles synthesis”（《液态金属用于高熵合金纳米颗粒的合成》）为题目发表在 Nature（《自然》）。

高熵合金是一种由五种或五种以上主元金属组成的新型合金，在极端条件下结构力学、能源转换与存储、医疗器械等领域具有重要的应用前景。实现高熵合金的原子级精准制造是其应用的基础。不同元素的物理化学性质差异会限制元素间的均匀混溶，不仅理想的高熵态难以获得，元素的选择也备受限制。根据吉布斯函数，高熵合金的合成通常依赖苛刻的高温反应条件（增大熵因子的贡献）来克服原子间不混溶性，并通过淬火等方式保持高熵态。在温和条件下实现高熵合金的多组元原子混溶有利于其可规模化、可定制化的精准制造，而这个目标极具挑战。

研究团队以“混合焓”为切入点，降低反应吉布斯自由能变，采用兼具负混合焓特性和流动性的液态金属，实现了温和条件下各类高熵合金体系的原子制造。液态金属（如镓）与大多数金属间亲和性好，混合焓为负值；且流动性良好，可加速传质，促进元素的均匀分散和合金化反应的进行。由此，在液态金属反应体系中，可在温和条件下实现高熵合金的多组元原子混溶，极大拓展了高熵合金的组分选择空间，有望促进其在更多关键领域的应用。

原位实验结合理论计算探究液态金属原子制造高熵合金的机制：作者通过对反应过程进行原位环境球差校正透射电镜和原位同步辐射表征探究了合金纳米颗粒的原子制造的机制。在高温和还原气氛下进行原位环境透射电镜观察，纳米颗粒展现出流动性，发生“融合”与“裂分”。在原位同步辐射 XRD 实验中，样品的特征衍射峰在高温及降温过程中一直保持，这表明样品晶化行为的存在。同时，作者使用基于机器学习势函数的分子动力学模拟对降温过程中样品的结构进行了模拟，验证了液态金属在原子制造高熵合金的重要作用。

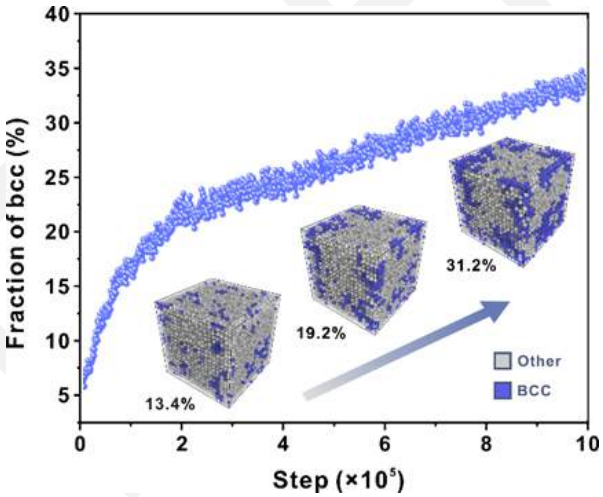


图. GaFeMnNiCu 高熵合金冷却过程的动力学模拟。bcc 结构的体积相分数随着模拟步骤而演变，其中另一相为非晶相结构。

为了进一步探索合金形成过程，团队使用 DeePMD 构建了五元高熵合金体系势函数。基于密度泛函理论计算，对 128 个原子的体系进行 10 ps 的不同温度和组分下的从头算分子动力学模拟以获取训练数据集。使用 DeePMD-kit sel_a 描述符构建机器学习势函数，随后使用 LAMMPS 对 16,000 个原子 GaMnFeNiCu 的 1000K 到 300K 的冷却过程进行 10 ns 的分子动力学模拟，结果表明在冷却过程中该体系可以保留其晶体结构。

Source: Cao, G., Liang, J., Guo, Z. et al. Liquid metal for high-entropy alloy nanoparticles synthesis. Nature 619, 73–77 (2023). https://doi.org/10.1038/s41586-023-06082-

AI4S 实践（6）：DP+CALYPSO 自主方案将结构搜索能力提高万倍，助力合金理性研发

Al-Mg 二元合金因其轻巧型和优异的机械性能广泛应用于汽车、航空航天和电子设备行业。然而现有的合金或材料数据库（如美国金属学会、合金相图数据库、无机晶体结构数据库、开放量子材料数据库和材料项目数据库等）中只记录了有限数量的 Al-Mg 二元系统合金间化合物，因此如何搜索更多镁铝二元合金材料成为获得高性能合金材料的关键。

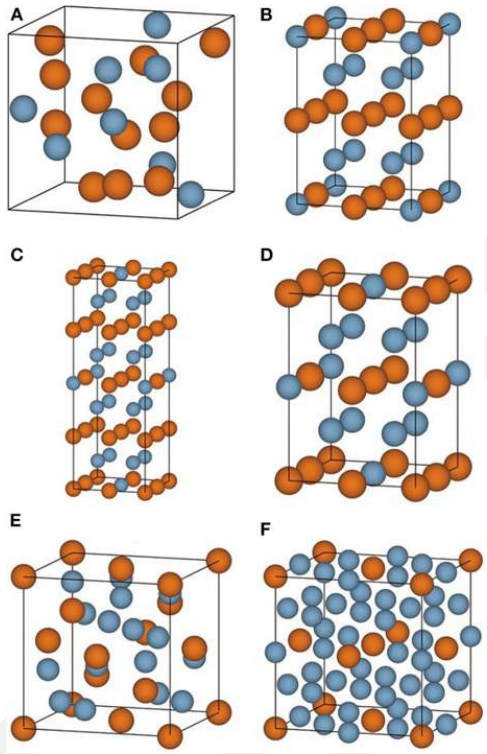


图. 镁铝合金结构示意图

来自合肥科技大学和北京大学的研究者利用深度势能结合 CALYPSO 来搜索潜在的稳定的 Mg-Al 金属间结构。借助 CALYPSO 结构搜索所有结构中的最低能量，从约 3200 个结构中，筛选出 458 个在小于 20meV per atom 的能量范围内。基于此训练集，借助深度势能进行 DFT 计算加以完善，得到有 Mg3Al3、Mg2Al2、Mg1Al1 和 Mg3Al，与材料库中已知结构 mp-1038779、mp-1094987、mp-1038934 和 L12 阶段一致，证实了方法的有效性。

此后，研究人员运用深度势能和 CALYPSO 方案系统构建新 Al-Mg 合金材料筛选体系（见下图），将有效筛选空间从过去的 10^3 级提升至 10^7 级。由热力学稳定性、动态稳定性和机械稳定性作为先决条件。筛选得到 Mg12Al8、Mg7Al9、Mg14Al18、Mg6Al10、Mg8Al16 和 Mg5Al27 这几种稳定的美铝合金材料。其中 Mg12Al8 表现出出色的延展性，Mg5Al27 具有很高的杨氏模量。

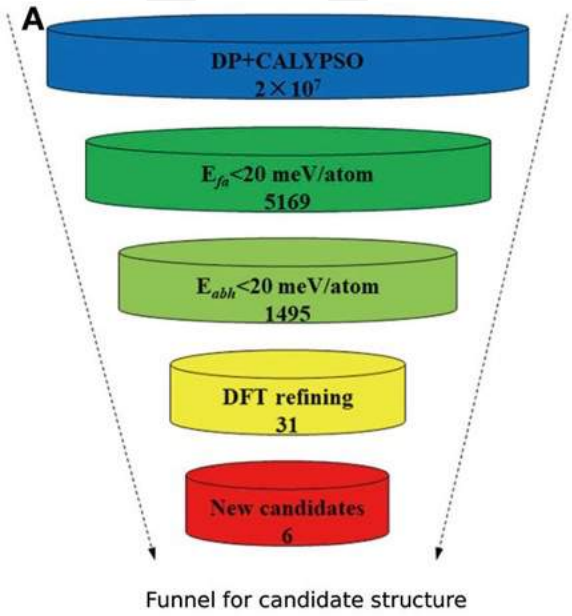


图. 深度势能+CALYPSO 生成新合金流程

Source：

Haidi Wang, Yuzhi Zhang, Linfeng Zhang, Han Wang, Crystal Structure Prediction of Binary Alloys via Deep Potential, ORIGINAL RESEARCH article Front. Chem., 26 November 2020 Sec.Theoretical and Computational Chemistry https://doi.org/10.3389/fchem.2020.589795

3.2.2 催化材料

催化剂可以提高生产率并降低相关反应的外场要求，显著降低实施特定工艺和/或生产目标产品所需的能耗。催化剂是现代工业中不可或缺的材料。所有工业制品中超过 80% 都涉及催化剂，主要包括化工合成催化剂、石油炼化催化剂、高分子及石油化工催化剂和环保催化剂四大类。然而，现有催化剂面临着反应过程中动态微观结构演变不清晰、反应路径不明确、催化剂数据库缺失、全流程"炒菜试错"现象严重等问题，极大制约了催化剂的发展。[1]

AI4S 依靠量子力学精度的建模和高性能计算，有望实现催化剂微观模型设计，揭示活性位点、表面吸脱附反应过程，深入理解催化剂构效关系等本质原理；借助机器学习势和第一性原理计算，模拟电子特征、原子间相互作用等催化关键"特征描述符"，生成高质量催化材料数据，辅助高通量筛选手段和先进的表征技术，分析催化机理和影响催化剂性能因素，筛选新的高效催化剂，解决工业难题。[2]

Source：

[1] Thermo Fisher, 催化剂研究, https://www.thermofisher.cn/cn/zh/home/materials-science/catalysis-research.html

[2] Bryan R. Goldsmith, Jacques Esterhuizen, et al. Machine learning for heterogeneous catalyst design and discovery, AIChE Journal[J], 2018, 5. doi.org/10.1002/aic.16198

[3] Catalysis and Synthesis Research Group, NWU, https://natural-sciences.nwu.ac.za/catalyst-and-synthesis

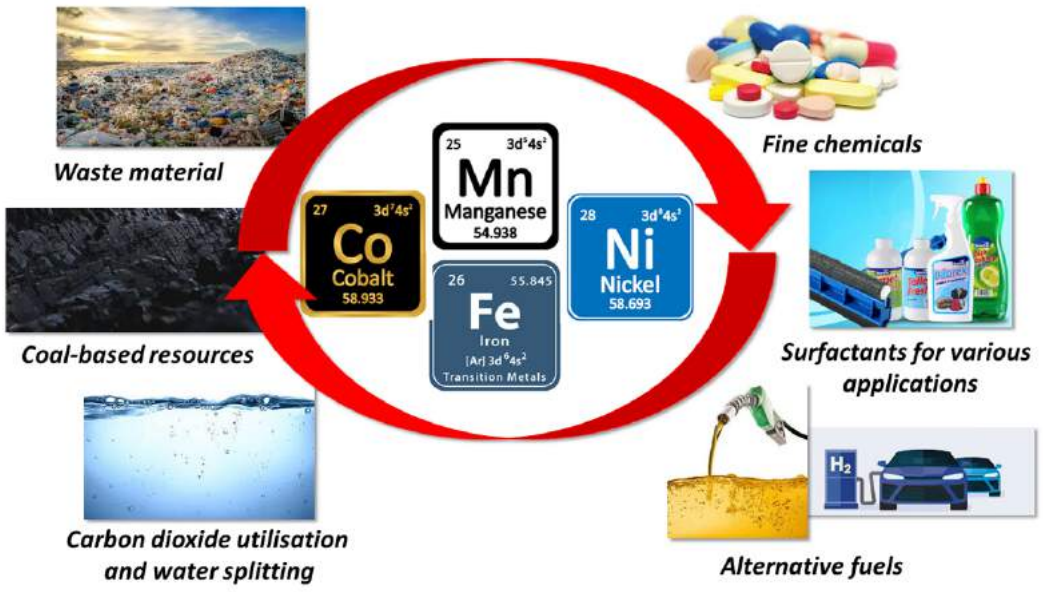


图. 典型催化金属和应用场景 [3]

AI4S 实践（7）：Parrinello 团队使用 AI4S 对百年化学工艺"铁催化-哈勃法"进行探究

催化剂被誉为现代化学工业的"芯片"，90%以上的化工过程与催化技术有关，从能源侧的石油精炼，到农业侧的化肥合成，到环保侧的尾气无害化，到医疗侧的药物合成生产，催化主导的化学反应覆盖生产生活的方方面面。催化剂的研发过程异常复杂，不仅要考虑催化效率，还要考虑寿命，服役条件，成本，毒性，生产放大等一系列问题。

2023 年 2 月，世界物理学大师 Parrinello 与化工巨头科莱恩的研究员等合作，发表重磅成果，利用深度势能等方法对化工中最基础的反应过程——铁催化哈伯法制氨，进行了完整的数字化复现，打破了过去几十年的行业默认假设，影响深远。

过去，受限于实验和计算手段的局限，学界只能对低温情况下的哈伯法反应进行研究。然而，实际生产过程中的哈伯法需要高温条件。学界通过低温中的研究成果进行推演，得到高温条件下的预测，与实际情况并不相符，使得长久以来，对哈伯法的调控和优化仍很大程度处于"盲人摸象"的"炼丹"模式。

在此研究中，Parrinello 等深入研究了哈伯-博斯赫氨合成过程的动态性质，尤其是在 Fe(111)表面上的 N2 三键的切割，这被认为是这个过程的限速步骤。文章指出，温度对此过程以及 Fe(111)表面形态的影响显著。

哈伯-博斯赫过程长期以来一直是氨合成的基础，然而这项研究表明，传统的静态理解可能无法全面捕捉到这个过程，尤其是在实际操作条件下。随着温度的升高，Fe(111)的表面结构会动态变化，使催化位点不断出现和消失，从而影响氮气分子的吸附和解离。这种在高温下的行为挑战了基于低温模型的传统观点，表明只有充分考虑动态效应，才能完全理解催化过程。

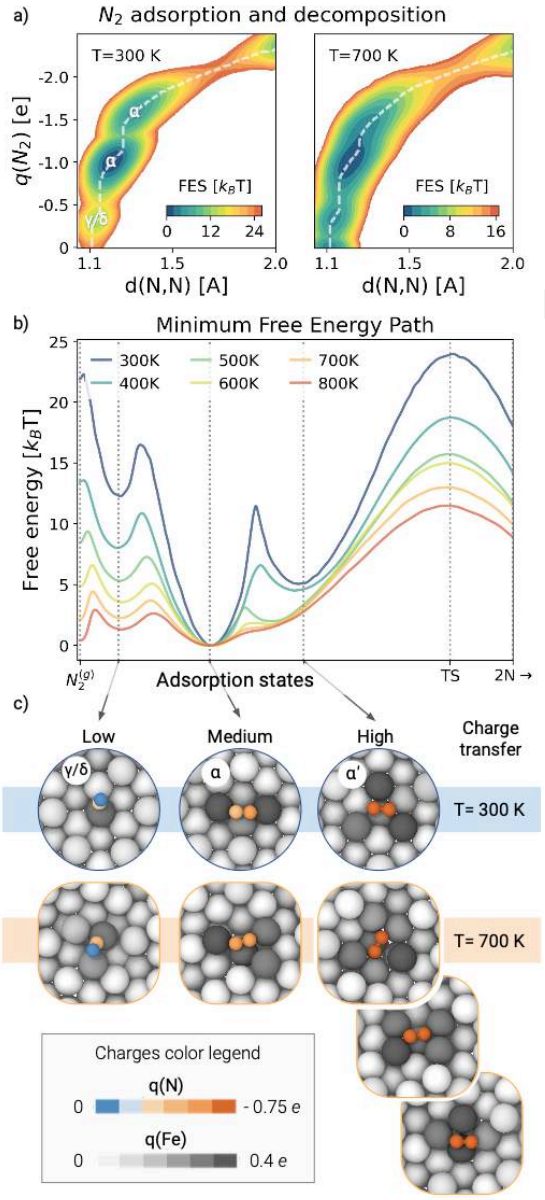


图. 吸附和分解机制。(a) 自由能作为 N-N 距离和 N2 部分电荷的函数。局部最小值表示亚稳态，白色虚线表示在这个平面上的最小自由能路径。 (b) 沿最小自由能路径在 d—q 空间内，从气相 (NS?) 到吸附态到解离态 (2N) 计算的自由能。每个温度下的自由能通过热能 kpT 进行重新标度，其中 kp 是玻尔兹曼常数。查看图 |S9|以获取以能量单位表示的自由能。只显示自由能最多到 N-N 距离为 2x 的地方，从那里开始应用谐波限制。 (c) 基于传输电荷量的吸附态的代表几何图像的快照，对于 T=300K 和 T=700K。原子根据神经网络模型预测的电荷进行着色。

此外，研究强调催化并非仅由静态的原子排布引发，而是涉及催化位点的持续形成和破裂。尽管这种动态性质看似有害，但它实际上可能通过防止形成可能中毒催化剂的稳定氮化物，从而提高整体催化性能。此外，研究表明，任何干扰铁表面动态重组的共吸附物都将作为强烈的毒素；催化剂对氧气、水或硫种类的极度敏感性，使得在氧化物或硫化物相形成的浓度远远低于毒性浓度，都会导致性能下降。

该研究最后指出，铁和氮之间的电荷重新分配对于成功的氨合成至关重要，需要双功能催化剂才能有效地解离和重组分子。这也可能有助于解释"助剂"在过程中的关键作用，可能通过形成氮化物-金属酸根中间体，促进部分负离子氢物种的氢化反应。

AI for Science 使得这项研究成为可能。文章中主要使用了 metadynamics (稀有事件采样) 以及深度势能两个 AI for Science 方法：

1）文章使用了 well-tempered metadynamics。这是 metadynamics 的一种变种,可以更好地控制 bias 的积累。

- 选取的集体变量(CV)是 N-N 距离 d(N-N)和 Fe-N 配位数 CN,Fe。这两个变量可以描述 N2 分子与 Fe 表面的相互作用。
- 在 DFT 模拟中,使用这两个 CV 进行 well-tempered metadynamics。每 50 步添加一个高斯 bias,以加速 N2 的吸附和分解过程。
- 通过 metadynamics,可以采样到更多反应过程中的中间状态,而不仅仅是平衡状态。这为训练势提供了重要的数据。

2）深度势能分子动力学 DeePMD

- 使用深度神经网络,训练了一个势函数来预测体系的能量和力。
- 训练数据集来源于 DFT 的静态计算和 metadynamics 轨迹。总计约 3 万组数据。
- 4 个模型组成一个集成,以评估预测的不确定性。最终模型的力预测误差只有 31 meV/A。
- 获得 deePMD 势之后,使用 LAMMPS 进行大规模的分子动力学模拟。
- 再次使用 well-tempered metadynamics,研究 N2 在 Fe(111)表面的反应机制。
- 通过解析自由能面,识别出不同温度下反应的状态和过渡态。

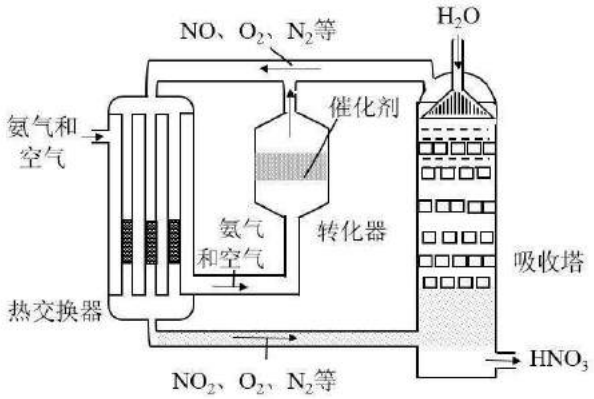


图. 哈伯法合成氨示意图

Parrinello 的成果展示了如何用神经网络和增强采样方法来数字化复现哈伯反应的真实过程，并保留其原子尺度的反应细节。metadynamics 提供稀有事件数据,DeePMD 提供运算效率,两者互补的联合使用使这项复杂的计算成为可能。在此基础上，通过不断改变数字模型中的反应条件来观察其预测结果的变化，就可以实现理性的工艺优化，也就是常说的 in silico design 或 design automation。

Source: Bonati L, Polino D, Pizzolitto C, Biasi P, Eckert R, Reitmeier S, et al. Non-linear temperature dependence of nitrogen adsorption and decomposition on Fe(111) surface. ChemRxiv. Cambridge: Cambridge Open Engage; 2023; This content is a preprint and has not been peer-reviewed.

AI4S 实践（8）：Meta AI + CMU Open Catalyst 项目提供 AI4S “四梁”支柱

面向可再生能源储能的开放式催化剂项目 Open Catalyst Project 是一个由 Meta AI 的基础 AI 研究部门(FAIR)和卡内基梅隆大学化工系联合进行的合作研究项目。该项目旨在利用 AI 来模拟和发现新的催化剂,以应用于可再生能源储能,从而帮助应对气候变化。随着我们对可再生能源(如风能和太阳能)的依赖性不断增强,这些间歇性发电的可再生能源迫切需要储能技术。其中一个具有扩展潜力的解决方案是将可再生能源转化为其他燃料,例如氢气。为了被广泛采用,这一转化过程需要高效低成本的催化剂。

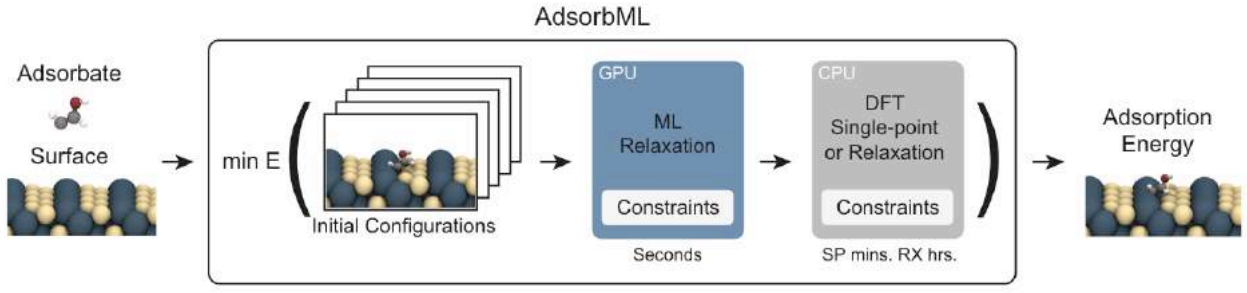
一个公开的挑战是找到低成本的催化剂来高速驱动这些反应。通过量子力学模拟(密度泛函理论),可以测试和评估新的催化剂结构。不幸的是,这些模拟的高计算量限制了可以测试的结构数量。人工智能或机器学习技术可以提供一种高效逼近这些计算的方法,以发现新的有效催化剂。

开放催化剂 2020(OC20)和 2022(OC22)数据集可用于训练机器学习模型。这些数据集共包含了 130 万个分子松弛的结果,这些结果来自于超过 2.6 亿个密度泛函理论计算。

2023 年 7 月，Open Catalyst 发布网页 demo，允许用户通过网页交互搜索吸附分子在催化剂表面的最佳吸附位点,返回吸附能量和对应的结构。这些输出可以用于设计具有特定性能的材料,例如其催化 CO2 还原反应或绿色 H2 生成的能力。在搜索这些全局优化的吸附分子配置时,示范首先在催化剂表面生成数十个吸附分子的配置。然后,它对每个吸附分子-表面配置进行局部优化,使用模型预测的力直到找到一个局部能量最小点。这个局部优化也称为松弛。

这些松弛由先进的图神经网络驱动,这些网络是在开放催化剂数据集上开发的——GemNet-OC 和 EquiformerV2——它们提供了快速准确的能量和力的预测。预测的力被用来驱动结构松弛,直到达到一个局部能量极小,这类似于 DFT 的过程,但快了几个数量级。目前 demo 支持 11,427 种催化剂材料和 86 种吸附分子。假设每个催化剂大约有 100 个表面,这就可以计算探索 11,427 x 86 x 100 = 98,272,200 = 约 1 亿种表面-吸附分子组合。再假设每个组合大约有 100 个初始配置,这意味着需要运行约 10 亿次结构松弛。这种规模的探索对 DFT 来说是不可能的——DFT 松弛在 GPU 上需要几小时到几天,而 AI4S 模型只需要几分钟。

Source: https://opencatalystproject.org/ ; https://arxiv.org/abs/2211.16486



Open Catalyst Demo 所使用的 AI4S 模型流程示意图

3.3 高分子材料的 AI4S 应用

聚合物，包括塑料、橡胶、纤维、涂料等常见材料，在我们的日常生活中发挥着重要的作用，涵盖了从家庭用品、纺织品到交通运输、航空航天等人类世界的方方面面。高分子聚合物微观高度复杂的链结构和聚集态结构决定了这些重要工业制品的整体性能，因此必须在相应的尺度对其进行研究和优化。

基于 AI4S 的第一性原理计算和物理模型驱动，有望实现从高分子单体到链结构再到大分子团聚结构的多尺度建模，探索聚合物材料构效关系；同时借助机器学习不断产生新的高分子聚合物数据，填补目前数据缺乏的问题；最终实现高效解决聚合物材料微观结构、组成、杂质分析及失效分析等问题，对聚合物玻璃化温度等重要性质进行预测，实现更高性能材料合成。[1]

Source:

[1]. Bryan R. Goldsmith, Jacques Esterhuizen, et al. Machine learning for heterogeneous catalyst design and discovery, AIChE Journal[J], 2018, 5. doi.org/10.1002/aic.16198

材料	分类	举例	应用
塑料	通用材料	聚乙烯、聚丙烯、环氧树脂、聚氯乙烯等	汽车、器械、日用品、电子电器等
	工程塑料	聚酰胺、聚甲醛、聚碳酸酯等	工程材料、代替金属制造机器零部件等材料
橡胶	通用橡胶	顺丁橡胶、丁苯橡胶等	汽车轮胎
	特种橡胶	氯丁橡胶、丁晴橡胶等	用于某种特性的特殊场合，如密封件、输油管道等
纤维		粘胶纤维、聚酯纤维、聚酰胺纤维、高性能纤维等	纺织、服装等
涂料		汽车涂料、船舶涂料、飞机涂料、建筑涂料等	

AI4S 实践（9）：聚合物结构和溶液中动态性能数据驱动粗粒度建模

分子动力学（MD）模拟已被广泛应用于研究溶液中的聚合物和生物分子。MD 跟踪每个分子链和溶剂分子的相互作用，这种模拟方式虽然准确但却昂贵。对于大型聚合物溶液系统，聚合物的介观性能和整体的动力学相对更重要。因此，没有必要模拟系统的所有原子细节，而可以适当消除或平均一定程度的自由度，以降低模拟成本，即所谓的粗粒度（CG）建模。[1]

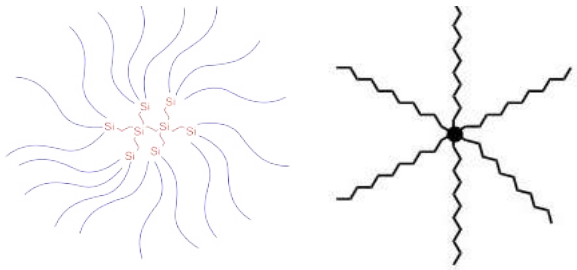


图. 常见星型聚合物 [1]

在实践中，威斯康辛麦迪逊大学的研究人员利用 AI4S 方法对 Star polymers（星形聚合物）计算建模 [2]。模型可以准确再现聚合物熔化中熔体特征（见右图），模型预测的结果与 MD 模拟结果具有高度一致性。这两个模型还可以复现聚合物溶液（浓溶液和稀溶液均成功复现）中原子系统的所有结构特征。两个模型的计算效率远比 MD 要高，如 CG1 模型计算效率比传统 MD 高 2 个数量级。

该研究成功的降低了计算模拟成本，加快了速度。为机器学习势函数在高分子聚合物中的应用做了初步验证。

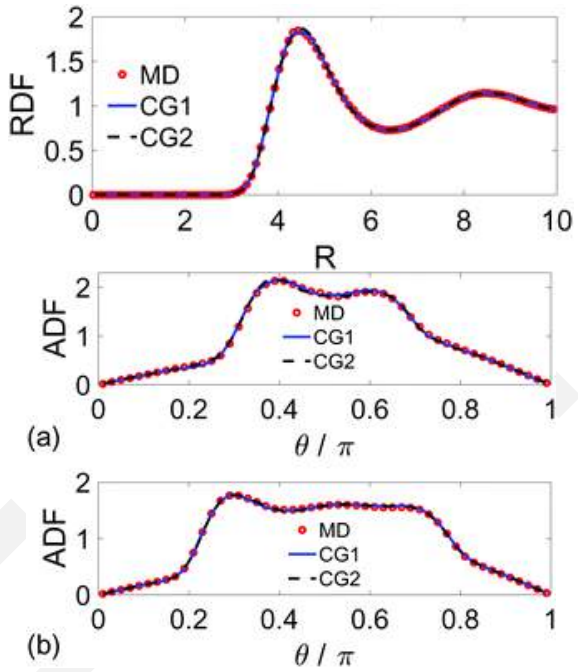


图. 聚合物熔体结构传统 MD 方法和机器学习势函数方法对比

在另一项有密歇根州立大学团队的研究中，研究者探讨了在建立粗粒度（Coarse-Grained，简称 CG）模型时如何克服界面液体的多重性质和尺度依赖性的挑战。界面能量在分子和连续尺度上表现出不同的特性，给模型的构建带来困难。粗粒度模型需要能够真实地编码未解析的原子间相互作用产生的多体相互作用，并且需要考虑到界面上的不均匀密度分布 [3]。

为了解决这个问题，研究人员同样使用了 DeePCG 方案，构建了单组分和双组分聚合物液体系统的粗粒度模型。在这种模型中，每一个聚合物分子被建模为一个粗粒度粒子。只需要使用热平衡状态下的瞬时力的训练样本，就可以构建出粗粒度模型。

研究人员发现，他们构建的模型能够准确地再现出在体积中的空位形成的概率密度函数，以及液体界

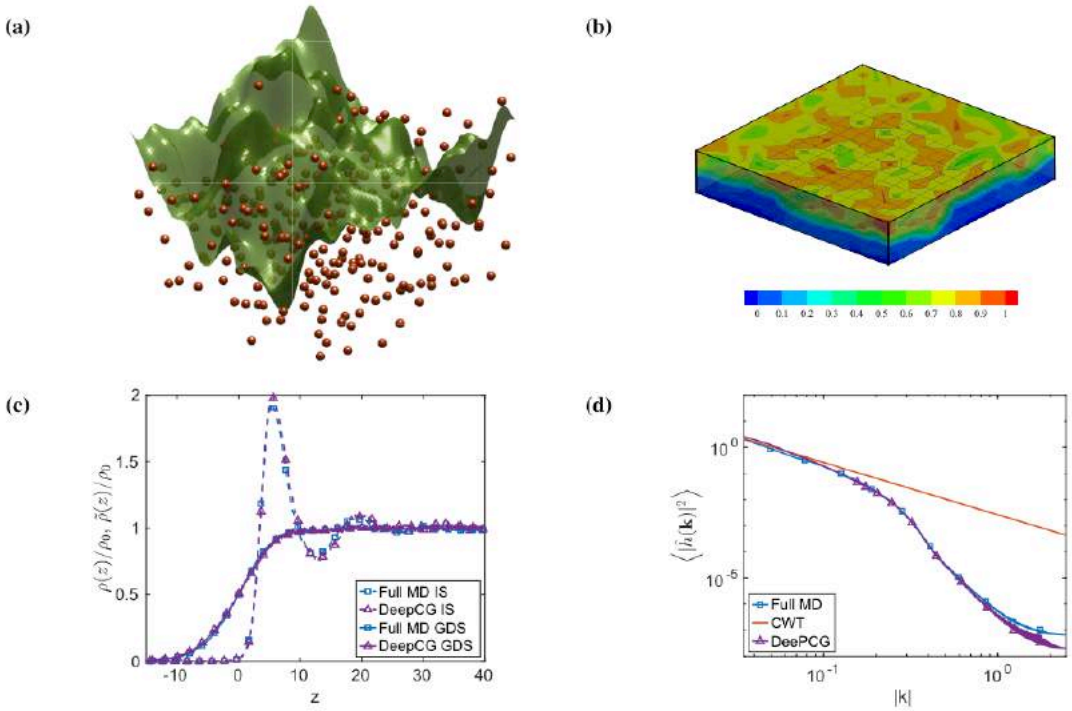


图. 单组分界面流体系统的流体密度和波动界面。 (a) 由公式(14)定义的界面，其中分子为红色，界面为绿色。 (b) 根据公式(13)定义的瞬时密度场的草图。 (c) 通过公式(14)定义的吉布斯分割面（GDS）和瞬时界面（IS）的平均密度剖面图。 (d) 波动界面的毛细波谱的系综平均。红色实线代表在低波数下使用公式(16)进行的 CWT 拟合。

面上的毛细波谱。更重要的是，这些模型可以准确地预测出非极性溶剂化能的体积-面积缩放转变，证明了该模型在探测分子级别编码的中尺度集体行为的有效性。

Source:

[1]. https://en.wikipedia.org/wiki/Star-shaped_polymer

[2]. Shu Wang, Zhan Ma ,Wenxiao Pan, Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved, Soft Matter, 2020,16, 8330-8344, https://doi.org/10.1039/D0SM01019G

[3] Pei Ge, Linfeng Zhang, Huan Lei; Machine learning assisted coarse-grained molecular dynamics modeling of meso-scale interfacial fluids. J. Chem. Phys. 14 February 2023; 158 (6): 064104. https://doi.org/10.1063/5.0131567

3.4 陶瓷和无机材料的 AI4S 应用

3.4.1 陶瓷

陶瓷材料是经过成形、烧结制成的一类无机非金属材料。历史上，陶瓷及其制造技巧成为各个国家的重要科技成果。除了在食器、装饰的使用上，在科学、技术的发展上同样也扮演着重要角色。陶瓷主要分为传统陶瓷材料和新型陶瓷材料。传统陶瓷材料常见的陶瓷材料原料有粘土、氧化铝、高岭土等。新型陶瓷材料是比传统陶瓷更加优异的新一代陶瓷材料，主要以高纯、超细人工合成的无机化合物为原料，采用精密控制工艺烧结制成，其成分主要为氧化物、氮化物、硼化物和碳化物等。

从应用上划分，陶瓷可以分为特种陶瓷材料（如功能陶瓷、结构陶瓷、生物陶瓷）、玻璃、水泥和化工陶瓷。随着人们对于材料的认识不断深入，陶瓷材料的发展也不断进化，现如航空航天、汽车、电子电器等诸多方面都见到陶瓷材料的身影，因此开发高性能陶瓷是关乎国家发展的重要课题。

人工智能（AI）在陶瓷材料研发中有很多潜在的应用。以下是一些主要的应用和示例：

- 材料设计和合成：AI 可以通过预测新材料的性质和性能来帮助研究人员设计新的陶瓷材料。例如，通过使用机器学习算法对大量的实验数据进行分析，AI 可以找出影响陶瓷材料性能的关键因素，然后为最优设计提供建议。此外，AI 还可以预测材料合成过程中的最佳条件，从而提高产量和质量。AI 还可以用于搜索已有的材料库（如 Materials Project），找到具有特定性质的陶瓷材料。例如，可以使用类似于推荐系统的算法，根据用户的需求，推荐最适合的材料。
- 掺杂配方设计、组分优化：在陶瓷材料中，不同的掺杂元素或组分会极大地影响材料的性质，如电导率、强度、耐热性等。通过 AI，我们可以对大量可能的掺杂元素或组分组合进行快速高效的筛选和优化，这比传统的试错法更高效。陶瓷材料通常由多种组分组成，不同的组分比例会对材料的性能产生重要影响。AI 可以通过优化算法，如遗传算法、蚁群算法等，来寻找最优的组分比例，以获得最佳的性能。有了预测模型和优化算法，AI 还可以自动设计实验方案，以尽快找到最优的掺杂元素和组分比例。例如，可以使用强化学习算法来自动调整实验参数，从而快速找到最优解。
- 质量控制：AI 可以用于自动检测陶瓷材料的缺陷，如裂纹、气孔等。例如，可以使用深度学习算法对成像数据进行分析，以识别出细微的缺陷。这比人工检查更准确，也更快。
- 工艺参数优化：AI 可以通过对生产过程中的数据进行实时分析，来提高生产效率和减少浪费。例如，AI 可以预测炉温、压力等参数对最终产品质量的影响，然后自动调整这些参数，以保持最优状态。此外，AI 也可以帮助优化陶瓷 3D 打印过程。例如，AI 可以预测打印参数如打印速度、喷嘴温度等对打印结果的影响，然后自动调整这些参数以提高打印质量。

这些只是 AI 在陶瓷材料研发中的一些应用，而且随着 AI 技术的进步，其在这个领域的应用肯定会变得更加广泛和深入

AI4S 实践（10）：《ACS Nano》收录哈工大团队陶瓷结晶过程模拟仿真算法

结晶是无机、有机和生物晶体材料的一个基本和必要的物理过程。理解结晶过程背后的机制可以推动科学研究和工业应用中晶体材料的设计和合成。 原子模拟一直被广泛地应用于材料科学、化学和物理学,以帮助我们在微观水平上更好地理解结晶过程。但是,经典的分子动力学模拟通常被限制在微秒时间尺度,远远小于结晶的实际时间尺度,即毫秒甚至更长。相关材料相之间的高能垒障碍了结晶过程在分子动力学模拟的时间尺度上实现的难度。 为了解决这个挑战,人们在增强抽样方法上做了很多工作。但是这些方法中的许多都依赖于选择适当的集体变量来区分不同的材料相。对于复杂的多组分材料系统,选择适当的集体变量仍然是一个关键的挑战。

2023 年，领域顶刊 ACS Nano 收录哈工大团队基于 AI4S 的复杂结晶过程仿真。这篇文章中使用了增强采样分子动力学(MD)模拟来研究复杂结晶材料的相变过程。由于多组分引起的多相问题,使用 MD 模拟复杂结晶材料的结晶过程仍面临巨大挑战。

研究者通过使用各向异性集体变量(CVs)和机器学习(ML)势实现了对复杂陶瓷的从头计算准确性 MD 结晶模拟。各向异性 X 射线衍射强度 CVs 可以精确识别含详细晶体学信息的复杂晶体结构,而 ML 势使得进一步进行具有从头计算精度的增强抽样模拟成为可能。研究者通过三种具有代表性结构的复杂陶瓷(即 MAX 结构的 Ti3SiC2,矿物结构的锆石和钛酸铅锆酸钛酸盐的钙钛矿结构)验证了这种方法的普适性和准确性。它表现出异常高效的从头计算质量,实现了所有这些陶瓷的结晶和自由能曲面的生成,有助于复杂结晶材料的分析和设计（下页图）。

技术思路:

1、使用各向异性 X 射线衍射强度作为集体变量。与 Debye 衍射方程式相比,各向异性 X 射线衍射包含额外的晶体学信息,可以更准确地区分复杂的各向异性晶体结构。

2、使用机器学习势。针对许多多组分或新发现的材料,缺乏经典势障碍了分子动力学模拟。机器学习势可以提供缺失的准确原子间相互作用。文章采用主动学习策略高效获得准确的机器学习势。

3、整合各向异性集体变量和机器学习势与几种前沿的分子动力学技术,实现了对复杂陶瓷的从头计算准确性的结晶模拟,验证了该策略的有效性和普适性。

AI4S 所驱动的结晶仿真方法可以计算复杂陶瓷的自由能曲面,揭示相变机理,预测熔点等热力学参数。这为设计和优化复杂陶瓷材料的组成和微结构提供了理论指导。此外，模拟结果可用于研究点缺陷等对材料性能的影响,指导材料优化设计。例如文章中提到的钛酸铅锆酸钛酸盐(PZT)的点缺陷会影响其粒长行为和光电性能。该方法可扩展至高熵合金、钴基超耐蚀合金等其他复杂材料的材料设计。

对于工业上普遍的工艺优化需求：文章准确预测了 MAX 相 Ti3SiC2 先结晶成 TiC 类型结构的竞争性核化行为,解释了实验中难以获得纯 Ti3SiC2 的原因,为优化合成工艺提供了理论依据。方法还可用于研究烧结、淬火等工艺参数对材料性能的影响,指导工艺优化。该计算方法可扩展到更多类型的复杂材料,为材料科学研究和工业提供强有力的理论工具。

Source: ACS Nano 2023, 17, 14, 14099−14113

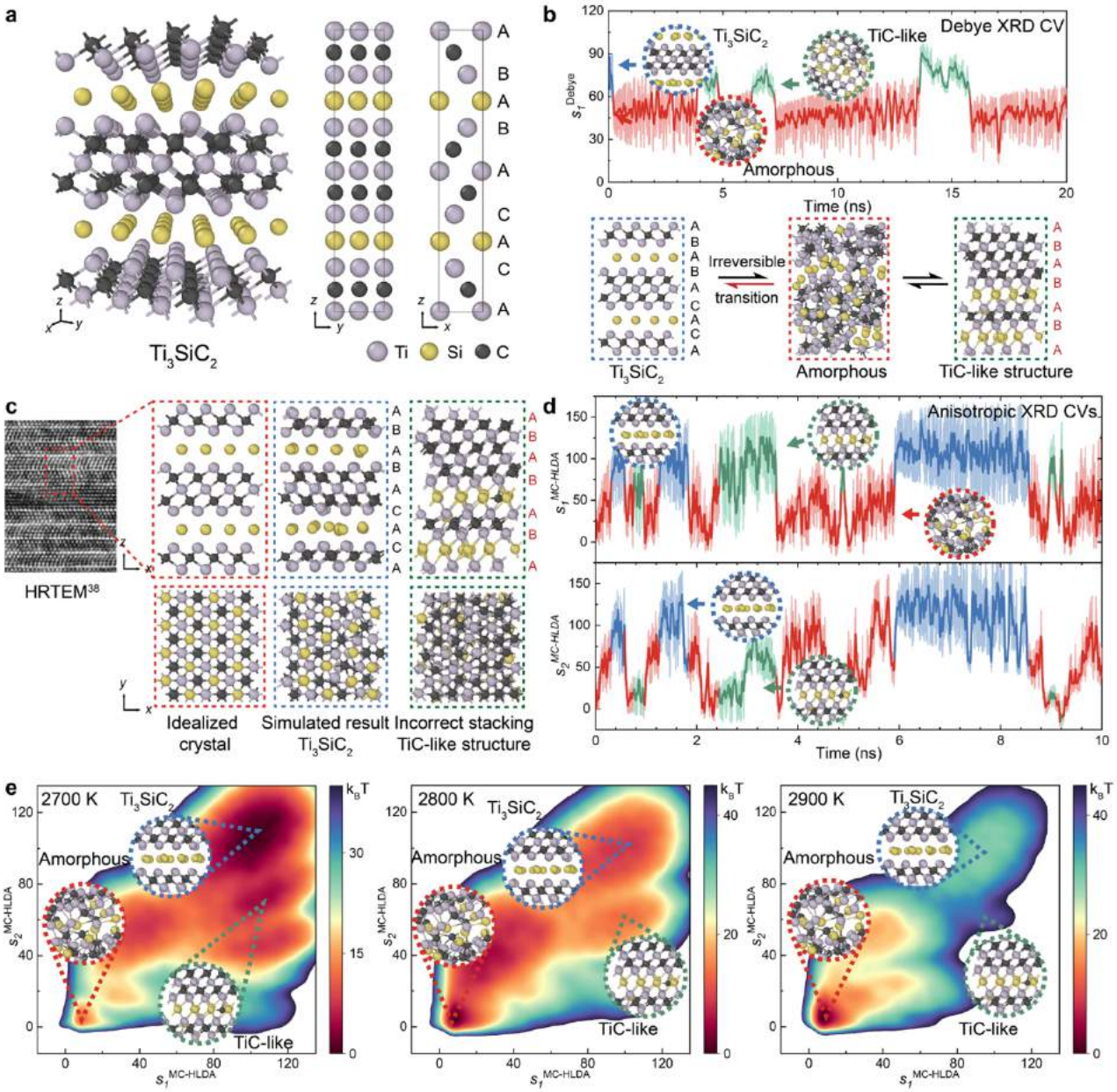


图 3. MAX 相的结晶。(a) Ti3SiC2 的理想化晶体结构。(b) 在使用 Debye XRD CV 的情况下,随着模拟时间的推移,CVs sD1ebye 的变化。Ti3SiC2 不可逆地熔融,然后无定形结构结晶成叠置不正确的 TiC 类结构。(c) 用模拟获得的 Ti3SiC2 晶体和 TiC 类结构与 HRTEM 结果和理想化晶体结构的比较。部分重印自 ref38。版权 2003 Elsevier。(d) 在 2900K 温度下,192 原子 Ti3SiC2 体系中,通过 MC-HLDA 方法获得的 CVs sM1C-HLDA 和 sM2C-HLDA 随着模拟时间的变化。CV sM1C-HLDA 可以很好地区分晶体类和无定形结构,而 CV sM2C-HLDA 可以区分 Ti3SiC2 和其他晶体类结构。粗线为运行平均(50 ps 间隔)的数据,以引导观察。(e) 在 2700K、2800K 和 2900K 温度下,Ti3SiC2 的自由能曲面。

3.4.2 水泥

水泥是全球广泛使用的结构材料。硅钙水凝胶作为混凝土材料的添加剂，很大程度上决定混凝土的强度、体积稳定性和耐久性，尤其是纳米 C-S-H 可以显著提高硅酸盐水泥混凝土早期的抗压/抗折强度。

但由于 C-S-H 是一种无定形相，有多尺度特征，受分辨率限制，其纳米结构尚未完全阐明，因此如何通过微观设计提高纳米晶核对水泥水化过程的促进作用仍是未解问题。由于计算成本限制，传统第一性计算很难实现大规模模拟，而现有 MD 模拟（如 ClayFF、ReaxFF 力场等）的精度依赖于经验立场，精度低。[1]

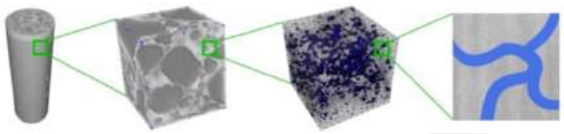


图. 水泥结构从宏观到微观[2]

为解决这个问题，一些研究者把目光转向 AI4S。其中，东南大学与江苏省建筑科学研究院高性能土木工程材料国家重点实验室的研究人员使用 AI4S 建模，对三种晶格尺度（9Å、11 Å、14 Å）的托贝莫来石（Tobermorite）进行了原子坐标进行构建，之后对其晶格常数、径向，角度分布函数、位移、弹性等进行计算，结果高度准确且耗时很短（计算效率与 ReaxFF 类似，且比 AIMD 快 2-3 个数量级）

基于上述三个训练模型，研究人员整合后训练了一个新 AI4S 模型，模拟了更大的体系，与实验观测的结果十分相似（见右图）。但随着混合结构的增加，精确度下降，且预测存在一些偏差，未来仍需考虑更多相关矿物结构和水结构，提高训练的准确性和可靠性[2]。

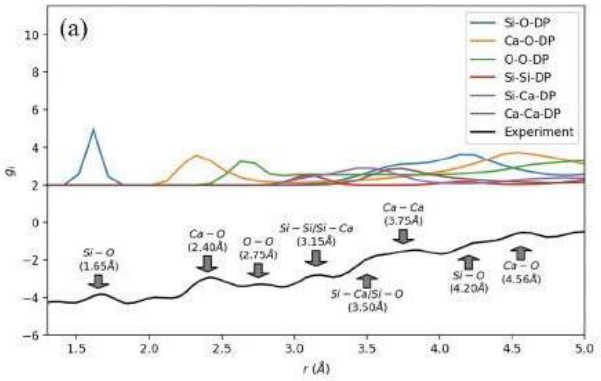


图. 整合模型对于更大体系的模拟

澳门大学团队使用深度势能、增强采样分子动力学研究不同温度下，从水/二钙硅酸盐界面开始的钙溶解的自由能表面、最小自由能反应路径和动力学速率[3]。研究发现，钙的溶解是一个自发反应，并且在不同的温度下遵循不同的最小自由能反应路径。常温下，五配位钙离子的溶解时间大约在几百秒的数量级，而加热后溶解时间增加到纳秒级。与每个基础反应低的自由能障碍相比，溶解动力学相对较慢，这归因于溶解反应的多方向和多步骤特性。

这一新的原子级洞察为我们更好地理解二钙硅酸盐和水泥水化提供了新的理解。这意味着我们可以更准确地理解并控制水泥硬化过程，从而优化水泥的性能，减少生产成本，提高效率。此外，这种新的原子级别的理解也为发展更环保、更耐用的水泥材料提供了可能性。

Source：

[1]. A deep learning potential applied in tobermorite phases and extended to calcium silicate hydrates, Cement and Concrete Research, doi.org/10.1016/j.cemconres.2021.106685.

[2]. A microscale model for concrete failure in poro-elasto-plastic media, Theoretical and Applied Fracture Mechanics 107:102517

[3] Unravelling the dissolution dynamics of silicate minerals by deep learning molecular dynamics simulation: A case of dicalcium silicate, Cement and Concrete Research, doi.org/10.1016/j.cemconres.2023.107092

3.4.3 纳米材料

纳米材料，如碳纳米管和石墨烯，具有卓越的物理和化学性质，例如高强度，高导电性和高热导率等，因此它们在电子学，光子学，能源和环境科学等许多领域有广泛的应用。AI 可以帮助科研人员理解和优化这些材料的性质，提高其生产效率，以及开发新的应用。纳米材料在各行业中的应用越来越广。在医学中，纳米粒子技术正被用于开发新一代药物、疫苗和生物传感器，如成熟的纳米复合材料、纳米磁性颗粒等。在能源方面可用于提高太阳能和风能的效率。由于其比表面积大，在工业领域里还可以用于制备优异催化剂等物品。

对纳米粒子的成分、尺寸和形状的认知是进一步了解其独特性质的关键第一步。纳米粒子的多功能性是因为原子和分子在纳米级具有基本不同的性质，有些在传热和电方面较好。有些更坚固，具有不同的磁性或更好的反射光能力。但正因为其材料尺寸小，量子效应明显，因此传统计算模拟手段受限于计算能力低和建模尺度大，无法精确描述。而 AI 在碳纳米材料开发中由广泛的应用思路。

材料设计：AI，特别是机器学习和深度学习，可以帮助科研人员理解和预测材料的性质。例如，机器学习算法可以根据材料的化学组成和结构预测其电子和力学性质。这可以帮助科研人员设计具有特定性质的新材料，例如高导电性或高强度的碳纳米材料。

制程优化：AI 也可以用于优化碳纳米材料的生产过程。例如，机器学习算法可以分析生产数据，如温度，压力和时间等参数，以确定最优的生产条件。这可以提高生产效率，降低生产成本，并提高产品的质量。

新应用开发：AI 还可以帮助科研人员开发新的碳纳米材料应用。例如，通过机器学习分析材料的性质和应用数据，科研人员可以发现材料的新用途，如新的能源存储或环境清洁应用。

虚拟筛选：AI 能够通过计算机模拟来预测新材料的性质，从而筛选出最有前景的候选材料进行实验研究，大大减少了实验成本和时间。

AI4S 通过第一性原理的建模、分析，保持 DFT 精度的前提下，从原子、分子出发，实现对微观场景如原子相互作用、成键、运动轨迹及层间作用的高精度模拟，突破量子效应，进而扩展对纳米材料认知，增加对纳米材料调优的手段和工具；有望成为优化材料、发展新材料的新方法。

图. 典型碳纳米材料：石墨烯，碳纳米管，碳炔 [6,7]

I. 人造钻石

自上世纪 50 年代以来，人造钻石技术的发展一直是科技领域的一项重要突破。初期的高压高温（HPHT）方法，到如今更加先进的化学气相沉积（CVD）技术，人类的知识和技术在不断地突破钻石的制造界限。今天，这些人造钻石不仅在珠宝行业有所应用，还广泛用于工业、科研以及先进的电子设备领域。而随着科技的进步，我们对于钻石的理解和制造能力也在不断提升，但其形成的深层次机理仍有待挖掘。

在人造钻石的制造过程中，理解碳原子如何在极端条件下自组织并形成稳定的晶体结构是一项重要的科学挑战。这涉及到复杂的物理化学过程，包括碳源的分解、碳氢物种的分离以及钻石核的形成和生长。因此，研究这些过程不仅可以提高我们对于行星科学、地球内部结构，甚至聚变能源技术的理解，还能为人造钻石工艺的发展提供理论指导。

国防科大戴佳玉等研究者在预印平台发布了一项研究[1]。论文研究了氢化碳系统中钻石的动态形成过程，并利用深度学习模型揭示了钻石形成的三步骤：分解，物种分离，和核化过程。研究者发现，钻石的生长过程与一个关键的核大小有关，其中动态能量障碍起到了关键作用。

首先，研究者构建模型来模拟和理解在高压和高温环境下，碳和氢的化合物（比如烃类）是如何分解和重组形成钻石的。这涉及到大规模的分子动力学模拟，这是一个非常计算密集的任务。为了处理这个问题，作者们利用深度学习技术来优化这个过程。

通过运行这个深度学习模型，作者们能够模拟出在极端条件下（如 125 GPa 的压力和 4590 K 的温度）

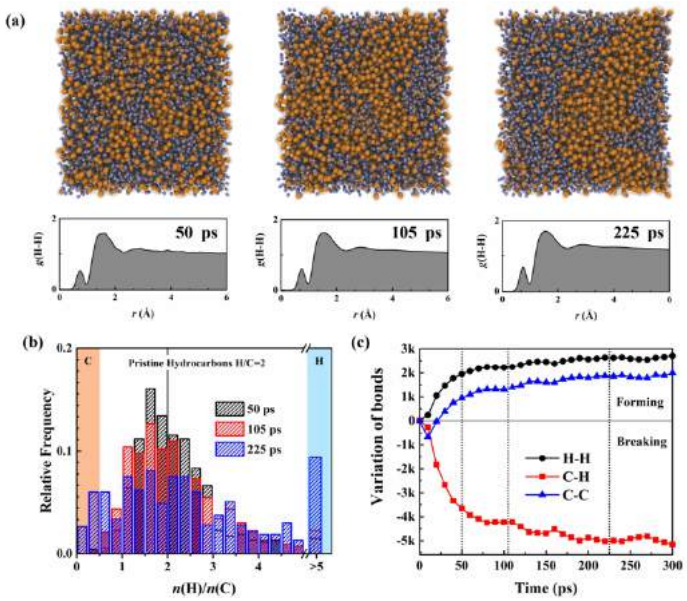


图. 冲击波高压下 C - H 的演化

钻石的形成过程。这个过程包括三个阶段：（1）碳和氢的化合物分解，产生氢气和低分子量的烷烃（如甲烷和乙烷）；（2）这些产物从碳链中逃逸出来，导致碳/氢物种的分离；（3）剩下的碳原子积累并核化形成钻石结构。

模型的结果显示，钻石的生长是一个动态过程，与动态能垒有关，并需要达到一个临界的核化半径。作者们发现，只有在达到这个临界半径时，钻石才会自发生长。这提供了关于如何优化人造钻石生产过程的重要线索。

作者们利用深度学习模型，通过大规模的分子动力学模拟，揭示了在极端条件下钻石形成的动态过程，提供了关于钻石如何在这些条件下自我组织形成的全新视角，这为人造钻石工艺的发展提供了理论依据。这篇文章仍在预印阶段，其准确度等待检验，但其已可为其他研究者提供思路和启发。

Source: arXiv:2208.01830

II. 石墨烯

石墨烯是一种由碳原子组成的六角型呈蜂巢晶格的平面薄膜，是有一个碳原子厚度的二维纳米材料，在常温下可观测到量子霍尔效应。其是世界上最薄但却十分坚硬的纳米材料，几乎完全透明，只吸收2.3%的光，导热系数高达 5300W/(m·K)，常温下电子迁移率超过 15,000cm2/(V·s)，且电阻率仅为 10-6 Ω·cm，为目前世上电阻率最小的材料。因此石墨烯被期待用来开发更薄、导电速度更快的新一代电子元件，同时其也适合做触控荧幕、光板及太阳能电池。[1]

在相关研究团队中，上海科技大学团队在国际权威期刊 Nano Letters (纳米通讯) 发表对石墨烯"魔角"的研究[8]。团队利用深度势能对双层石墨烯魔角(Magic-angle twisted bilayer graphene，简称 TBG) 的声子性质和电学性质进行第一性原理建模，建立了 DFT 精度的 11000 原子体系，对以往实验中发现的各种难以解释的石墨烯"神迹"提供了潜在的理论解释。研究中发现了一些在实空间中可以展示偶极子、四极子和八极子振动模式的软模式，以及一些时间反转破缺的手性声子模式。他们进一步研究了通过冻结某些软声子模式对电子结构的影响。

在文章中，他们通过冻结特定类型的软声子模式，并研究其对电子带结构的影响。研究发现，如果假设冻结了一种软的四极子声子模式，系统将表现出与最近的实验完全一致的电荷序。此外，一旦冻结了一些低频的 C2z 破坏模式，零电荷中性点的狄拉克点将会出现能隙，这为零电荷中性点的关联绝缘体状态的起源提供了另一种视角。比如，研究者通过建模发现，当冻结 soft stripe-type phonon mode 时，体系的电荷序会呈现 STM 观察结果；此外，AI4S 模型显示，冻结特定 C2z-broken phonon mode 可打开 CNP （电荷中性点位）的间隙，帮助为 CNP 点位出现关联绝缘态的原因提出假说。

石墨烯被视为未来电子和光子器件的理想材料。石墨烯的性质在一定程度上取决于其结构，而结构可以通过改变层间的相对旋转角度来调整。特别地，当两层石墨烯的相对旋转角度为"魔术角度"时，石墨烯会表现出一些非常有趣的电子性质，如超导和关联绝缘状态。因此，研究魔术角度的双层石墨烯对于理解石墨烯的电子性质和开发基于石墨烯的新型器件具有重要意义。

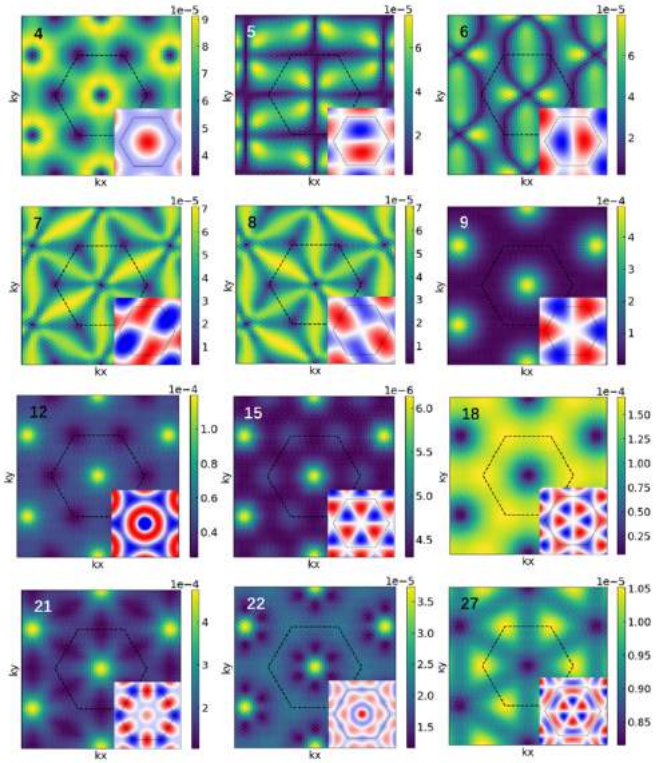


图. pPmn |gmnν(k, 0)|2 作为 k 函数的热图，对于第 4(0.043 THz)、第 5(0.046 THz)、第 6(0.046 THz)、第 7(0.063 THz)、第 8(0.063 THz)、第 9(0.073 THz)、第 12(0.116 THz)、第 15(0.132 THz)、第 18(0.163 THz)、第 21(0.179 THz)、第 22(0.239 THz)和第 27(0.253 THz)个 Γ 点声子模式，单位为电子伏特(eV)。虚线六边形标记了第一布里渊区，插图显示了模式的面外（

III. 碳纳米管

碳纳米管（CNT）是一种管状的碳分子。管子的半径方向非常细，只有纳米尺度，几万根碳纳米管并起来也只有一根头发丝宽，碳纳米管的名称也因此而来。碳纳米管的分子结构决定了它具有一些独特的性质。由于巨大的长径比，碳纳米管表现为典型的一维量子材料，具有超常的强度、热导率、磁阻，且性质会随结构的变化而变化、可用于制造高性能复合材料、生物材料及半导体晶体管等。[2]

复旦大学研究团队利用深度势能对低维碳结构数据进行学习并训练了模型，其用于预测碳纳米管的晶体结构及原子能量，与实验和 DFT 计算数据有较好的拟合。证明了此描述碳体系模型的可靠性和好的预测能力。[3]

在实际工业生产中，CNT 依赖 CVD (化学沉积) 等复杂工艺，制造过程复杂。CNT 成品质量取决于其沉积/生长过程，而控制其生长以形成特定的结构、尺寸和纯度是一个挑战。对此，韩国基础科学研究所 (Institue of Basic Science) 团队在 NT22 (国际纳米管应用大会) 提出使用 AI4S 模拟和理解 CNT 生长过程。 [9]

该团队指出：从干净的铁催化剂到完全生长的管子是一个挑战，因为 CNT 的生长速度/缺陷修复速度很慢。这使我们无法使用 DFT 等精确技术来模拟这个过程。为了解决这个问题，研究者训练了一个先进的深度神经网络势，开发一个铁-碳机器学习力场（MLFF），在原子级别研究 CNT 的生长动力学。研究者表示使用单个 Nvidia V100 GPU，可以模拟每天大约 70 ns（70 000 000 步）的 CNT 生长。这项研究发现，为了无缺陷地生长 CNT，足够的刻蚀过程是关键，而碳的添加速率应约为刻蚀速率的 1/5 到 1/10。这种发现对于优化 CNT 的制造工艺来说至关重要。研究还发现，单壁碳纳米管（SWCNT）的边缘在生长过程中非常动态，最常见的边缘并不一定与管子的螺旋度相匹配。这意味着，我们需要更深入地了解和控制 CNT 生长过程中的动态行为，以得到具有特定结构和性质的 CNT。此外，该研究显示，“产率”相当低，只有 37/100 的管子具有明确定义的螺旋度（即没长歪）。这意味着，我们需要进一步优化生长条件，以提高 CNT 的产率和质量。

图. AI for Science 仿真预测, 在无特殊工艺控制时，CNT 的"良率"仅为 37%

IV. 碳炔

碳炔是碳原子聚集在一起形成的链，这些碳原子通过双键或者交替的单键和三键连接在一起。碳炔被认为是世界上最强韧的材料，强度比钢高 200 多倍，超过钻石 40 倍，是石墨烯抗拉强度的 2 倍。碳炔在未来超高强度设备的发展中将有很重要的用途。[4]

北京大学的研究团队借助深度势能对其波恩奥本海默电子结构进行了基于第一性原理的建模分析。研究人员用该模型对碳炔的动态、静态电子性质进行了模拟预测，结果与 DFT 计算得到光谱准确性一致。基于此，还对光学电导率等进行了计算预测。该方法为与模拟电子结构结合的哈密顿量提供了新的手段，也扩大了对更宽温度区间及更多动力学现象中电子结构和过程建模的范围。[5]

V. MXenes 二维过渡金属碳化物等衍生材料

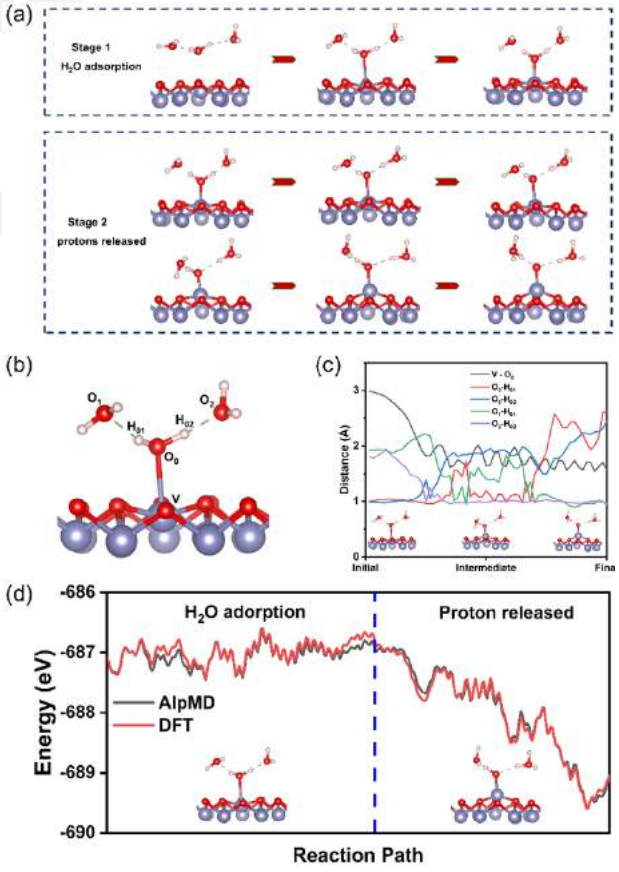
由于碳原子独特性质，其在各种不同工艺和环境的影响下可以形成繁复多样的衍生物，并具有令人眼花缭乱的奇特性质。

过去十年,一类被称为 MXenes 的二维过渡金属碳化物、氮化物和碳氮化物材料受到广泛关注。其通式为 Mn+1XnTx(M 为过渡金属,X 为 C 或 N,Tx 为表面基团)。由于其优异的性能,MXenes 在能源存储、催化剂、传感器等领域展现出广阔的应用前景。

MXene 具有层状结构、高导电性、大比表面积等优异性质,被认为是理想的锂电池电极材料。但其稳定性是影响实际应用的关键问题。但是 MXenes 极易在含氧、含水、光照环境下氧化降解为相应的过渡金属氧化物,严重影响其应用。水溶液环境下的 MXenes 氧化尤其突出。虽然早期研究显示溶解氧是氧化的主要原因,但即使去除溶解氧,MXenes 在水溶液中也会发生氧化。然而 MXenes 在水溶液环境下的氧化机制尚不清楚。

2023 年，国际顶刊 Angrew Chem 收录吉林大学新型电池物理与技术教育部重点实验室研究。研究采用神经网络势方法获得了 V2CO2-H2O 体系的势能面,并进行了长时间尺度的分子动力学模拟。结果发现,V2CO2 在水中快速降解,氧化率会随水层厚度增加和模拟时间延长而降低。还提出了质子运动和钒氧化物保护层的新机制来解释模拟结果[12]。

研究显示，MXene 材料在含氧、含水等环境下极易氧化降解,严重影响其实际应用,但氧化机制尚不明确。该研究首次在原子尺度上清晰展示了 MXene 与水分子的氧化过程,为理解其水解过程提供理论指



(a) V2CO2-H2O 体系中的氧化过程。(b) 过渡状态结构。(c) 相关原子之间键长的变化。(d) 用 DFT 计算的氧化过渡状态能量与 AIpMD 计算的比较。

导。并利用神经网络势进行长时间大系统模拟,获取 MXene 氧化动力学信息,为设计保护策略提供支持。

研究成果对提高 MXene 的储存稳定性,扩大其应用领域具有重要的参考意义。研究揭示了 MXene 在水溶液环境下的氧化行为及机制,为改善 MXene 在电池电解质中的稳定性提供了理论依据。

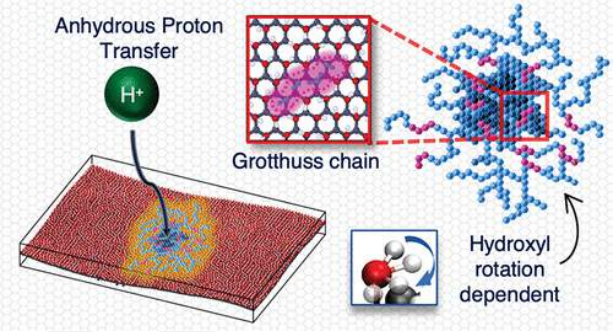
在另一材料体系中，2010 年哈佛大学 Kaxiras 等发现 一种名为"Graphanol"的二维结晶水合物，这种物质是由石墨烯通过水合过程得到的。在这种情况下，石墨烯形成了一种新的二维结构，其特点是石墨烯基底 C 原子平面的一侧自发形成了鱼骨状的羟基结构，而另一侧则发生了质子的化学吸附。这个新的石墨烯衍生物有一些独特的性质。例如，水合后，原本是半金属的石墨烯转化为了具有显著直接带隙的绝缘体。由于电荷转移和排列，垂直于 C 原子平面产生了一个宏观的净偶极矩。水合对振动光谱也有显著影响，出现了新的峰值，石墨烯的键伸展模式转移到了较低的能量。 [10]

2023 年，匹兹堡大学 Karl Johnson 组使用 AI for Science 方法对 Graphanol 进行探究[11]。Graphanol 能够在缺水的情况下导电，因此对于提高质子交换膜燃料电池的性能，降低成本以及扩大操作条件至关重要。

本研究中，作者详细地进行了原子级的模拟，展示了 graphanol 在无水的情况下可以以非常低的扩散障碍进行质子传导。作者开发了一个针对 graphanol 的深度势能模型，计算了作为温度函数的质子自扩散系数，估计了质子扩散的总障碍，并描述了热波动对系统尺寸的影响。作者提出并测试了一个详细的质子在 graphanol 表面的传导机制。

研究发现，质子可以沿着 Grotthuss 链（由几个排列对齐的羟基团组成的链）快速跃迁，这些羟基团之间的氢键允许质子在链的前后方向上进行传导，而不需要羟基团旋转。长距离的质子传输仅在新的 Grotthuss 链通过链中一个或多个羟基团的旋转形成时发生。因此，总的扩散障碍由内在的质子跃迁障碍和内在的羟基旋转障碍的卷积组成。

这项研究为开发新的无水质子传导膜提供了一套设计规则，这些膜的扩散障碍更低。这为燃料电池技术的进步提供了新的可能性，可望实现更高效、更便宜且更广泛应用的质子交换膜燃料电池。



Source:

[1]. https://zh.m.wikipedia.org/zh-hk/石墨烯

[2]. https://zh.wikipedia.org/wiki/碳纳米管

[3]. Jinjin Wang, et al, A deep learning interatomic potential developed for atomistic simulation of carbon materials, Carbon, Volume 186, 2022, Pages 1-8, ISSN 0008-6223, https://doi.org/10.1016/j.carbon.2021.09.062.

[4]. https://www.baike.com/wiki/碳炔/10442949?view_id=3vw7cs3oror6yo

[5]. Qiangqiang Gu, Linfeng Zhang, Ji Feng, Neural network representation of electronic structure from ab initio molecular dynamics, Science Bulletin, Volume 67, Issue 1, 2022, Pages 29-37, ISSN 2095-9273, https://doi.org/10.1016/j.scib.2021.09.010.

[6] Industrial-Scale Graphene Nanoplatelets & Dispersions Jan 10, 2018 | ACS MATERIAL LLC,

[7] C.S Casari, A.Milani, Carbyne: from the elusive allotrope to stable carbon atom wires, MRS Communications , Volume 8 , Issue 2 , June 2018 , pp. 207 – 219, DOI: https://doi.org/10.1557/mrc.2018.48

[8] Moiré Phonons in Magic-Angle Twisted Bilayer Graphene, Xiaoqian Liu, Ran Peng, Zhaoru Sun, and Jianpeng Liu, Nano Letters 2022 22 (19), 7791-7797 DOI: 10.1021/acs.nanolett.2c02010

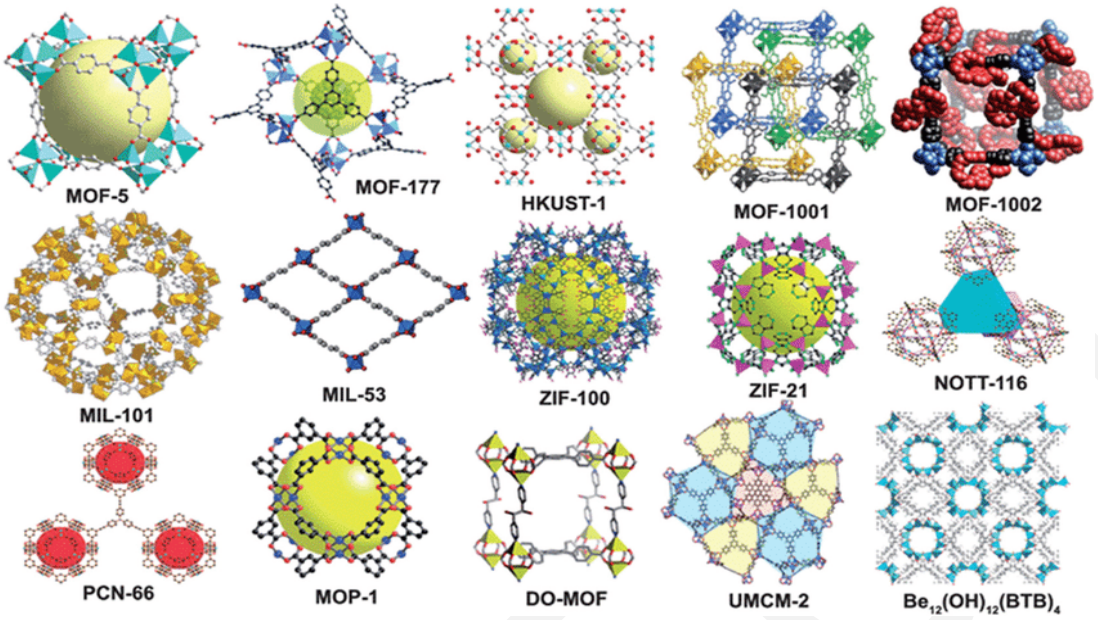
[9] Hedman, Daniel & Mclean, Ben & Larsson, Andreas & Ding, Feng. (2022). Atomistic simulations of carbon nanotube growth using machine learning force fields - From a clean Fe cluster to a fully grown tube. 10.13140/RG.2.2.34637.92642.

[10] Wei L Wang and Efthimios Kaxiras 2010 New J. Phys. 12 125012

[11] ACS Appl. Mater. Interfaces 2023, 15, 21, 25873–25883

[12] Angew. Chem. Int. Ed. 2023, e202304205

3.4.4 金属有机框架（MOF）



金属有机框架（MOF）是一类具有重大科技价值的晶体材料。由于其具有永久孔隙和可调整的孔隙大小和孔隙化学环境，MOF 具有广泛的应用前景，包括催化、非线性光学、气体分离、气体存储和传感器。一方面，MOF 的微孔和纳米多孔通常为分子尺度，与其他材料相比，会表现出独一无二的特性。另一方面，MOF 可以使用无机砖或金属氢氧化物二次构筑单元（SBU）和有机配体（链接）实现广泛的定制，以创建具有高孔隙体积、表面积大和量身定制的化学孔隙环境的高多孔三维结构。

由于 MOF 结构的复杂性，DFT 计算时间、资金成本高，且应用有一定局限性，所以需要新的研究方式。在各类 MOF 中，UiO-66 是一种重要的捕获和销毁有毒化学品的物质，已经有较为广泛的研究。但其孔隙结构较窄，会存在一些运输限制，影响其反应动力学及应用的有效性。因此，研究其内部的分子扩散性会对未来其应用产生积极作用。

匹兹堡大学的研究人员用 AI4S 方法训练了针对 UiO-66 的模型，对比了该 MOF 的结构、力学性能与 DFT 计算结果，具有较高的一致性。研究者随后通过混合势方法精确计算扩散率。研究人员用该模型计算了 UiO-66 中 Ne 的扩散系数，与 AIMD 模拟的结果非常吻合，而且该方法可以对包含数千个原子的系统进行计算，而传统 DFT 计算方法无法实现。基于此，研究人员预测了 Xe 的扩散系数，结果比 Ne 低一个数量级，表明 Xe 穿过孔隙时有明显的空间位阻。通过混合势可以模拟多孔材料内的吸附和扩散，包括吸附物引起的框架变化。这种混合方法不仅提高了模型的准确性，也极大地扩大了其适用范围，对于理解多孔材料中的吸附和扩散过程具有重要的理论意义和实践价值。这项研究验证了机器学习模拟多孔材料的吸附和扩散的可行性。

Source:
[1]. J Chem Theory Comput. doi: 10.1021/acs.jctc.2c00010
[2]. Kampouraki, et al. 2019. "Metal Organic Frameworks as Desulfurization Adsorbents of DBT and 4,6-DMDBT from Fuels" Molecules 24, no. 24: 4525.

AI4S 实践（11）：IBM Research 使用 AI4S 研究 "MOF 捕获 CO_2" 命题，助力绿色未来

金属有机框架（MOFs）在气体储存和分离方面具有显著的潜力，其中包括二氧化碳（CO2）的捕获和储存。由于二氧化碳是主要的温室气体，其排放量的控制与减少在全球应对气候变化和推动绿色经济转型中起着关键作用。实现碳中和，即实现"双碳"目标（即碳达峰和碳中和）的一部分，需要我们找到有效的方法来捕获和储存二氧化碳。

2023 年，IBM Research 的研究者使用深度势能对 Mg-MOF-74 吸附 CO2 问题进行了探究[3]。作者展示了基于 QMLFF 的 CO2 在 Mg-MOF-74 中的分子动力学模拟可以预测接近实验值的结合自由能景观和扩散系数。机器学习和原子模拟的结合帮助实现了更准确和有效的 MOFs 中气体分子的化学吸附和扩散的理论评估。此外，该方法的一个有趣的应用是使用机器学习势在蒙特卡洛模拟中获得每个温度和压力下 MOFs 中的 CO2 吸附。可以在大正则蒙特卡洛（GCMC）模拟中使用能够准确预测能量的 ML 势，以产生气体吸附的准确等温线。

这项研究证明了结合机器学习和量子力学的深度是能，能有效进行原子 MD 模拟，可以产生不仅与 QC 模拟直接相当，而且接近实验测量值的结果。这种方法可以应用于任何在具有开放金属位点的 MOFs 中的小型气体分子的化学吸附。最重要的是，这种方法不需要人为干预可以在高性能集群或超级计算机上自动化，用于对大约一百万种 MOFs 进行计算机模拟筛选。

研究者期望这种低成本和高精度的方法可以大大促进任何固体吸附剂中小分子的化学吸附和扩散的计算机模拟建模，让我们更有效地筛选和设计 MOFs，以实现更好的二氧化碳捕获和储存性能。

Source: ACS Nano 2023, 17, 6, 5579–5587

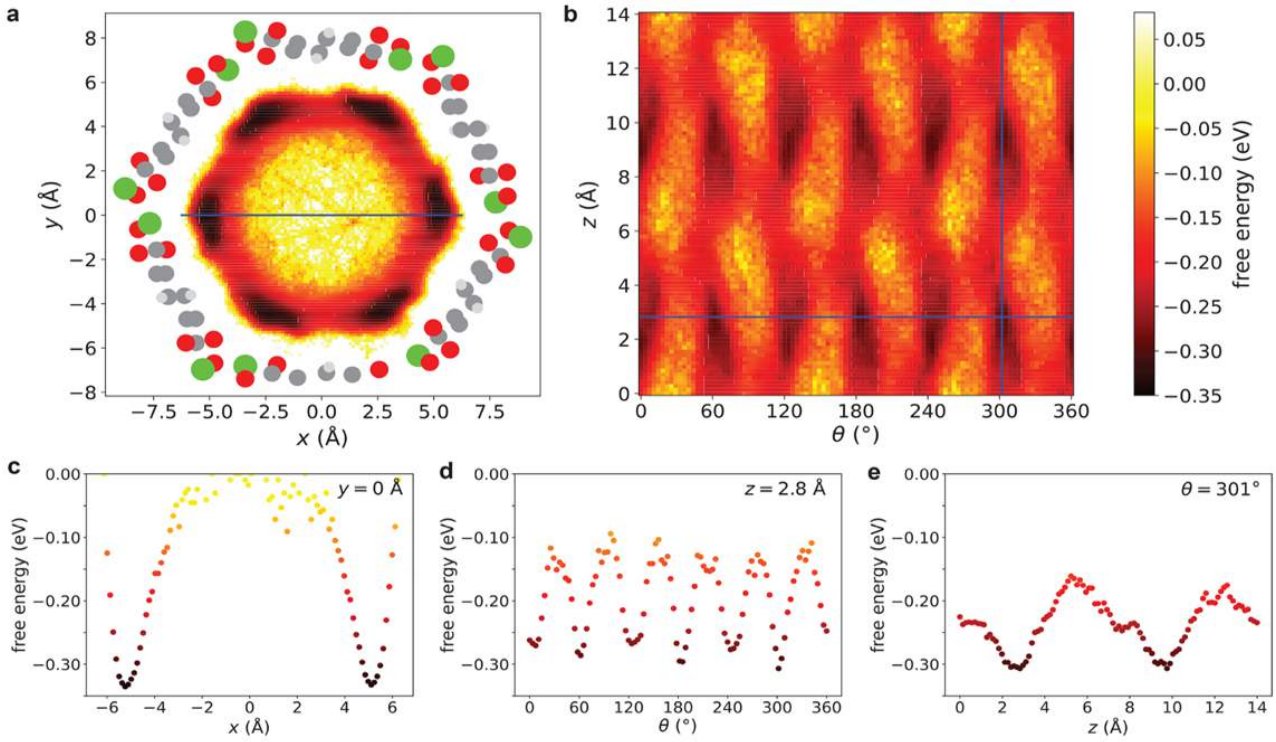


图. 二氧化碳在一个 Mg-MOF-74 通道内的自由能面。（a）在 x-y 平面上的自由能。为了参考，Mg-MOF-74 框架的位置也被绘制在图中。（b）在 θ-z 平面上的自由能。（c）沿着 y = 0 Å 从面板 a 切割的一维自由能图。（d，e）沿着（d）z = 2.8 Å 和（e）θ = 301°从面板 b 切割的一维自由能图。所有图都使用相同的色标刻度。

3.5 复合材料的 AI4S 应用

上世纪四十年代，随着航空行业对于高性能材料的要求，复合材料应运而生。复合材料是由金属材料、陶瓷材料或高分子材料等两种或两种以上的材料经过复合工艺而制备的多相材料，各种材料在性能上互相取长补短，产生协同效应，使复合材料的综合性能优于原组成材料而满足各种不同的要求。 复合材料由连续相的基体和被基体包裹的增强体组成。复合材料的基体材料分为金属和非金属两大类。金属基体常用的有铝、镁、铜、钛及其合金。非金属基体主要有合成树脂、橡胶、陶瓷、石墨、碳等。增强材料主要有玻璃纤维、碳纤维、硼纤维、芳纶纤维、碳化硅纤维、石棉纤维、单晶晶须、金属丝和硬质细粒等。**随着各领域的发展和材料科学的进步，复合材料的研发也正在由宏观复合形式向微观复合形式发展。**

由于人们对材料应用的要求日益严格，微观复合材料成为当下研究热点：均质材料在加工过程中内部析出的增强相和剩余的基体相构成的原位复合材料或纤维增强复合材料，也包括用纳米级增强体的复合材料以及刚强棒状分子增强的分子复合材料等。人工智能（AI）在复合材料领域的应用具有巨大的潜力，尤其是在材料设计、制造、测试和优化等环节。AI 可以通过机器学习（ML）算法自动筛选和优化复合材料。使用深度学习和其他 AI 技术，可以构建预测模型，预测新材料的特性和性能。例如，通过训练复合材料的微观结构与宏观性能之间的关系，可以预测新设计的材料是否能达到预期性能。这种方法可以显著减少繁琐的实验过程，加速新材料的研发。同时，AI 可以研究材料界面：材料界面是决定复合材料性能的重要因素。AI 可以对材料界面进行深入研究，例如识别和分类材料界面的微观结构，预测界面特性，以及研究界面性质与复合材料宏观性能之间的关系。此外，AI 也能通过优化算法设计更优的材料界面，以提高复合材料的性能。

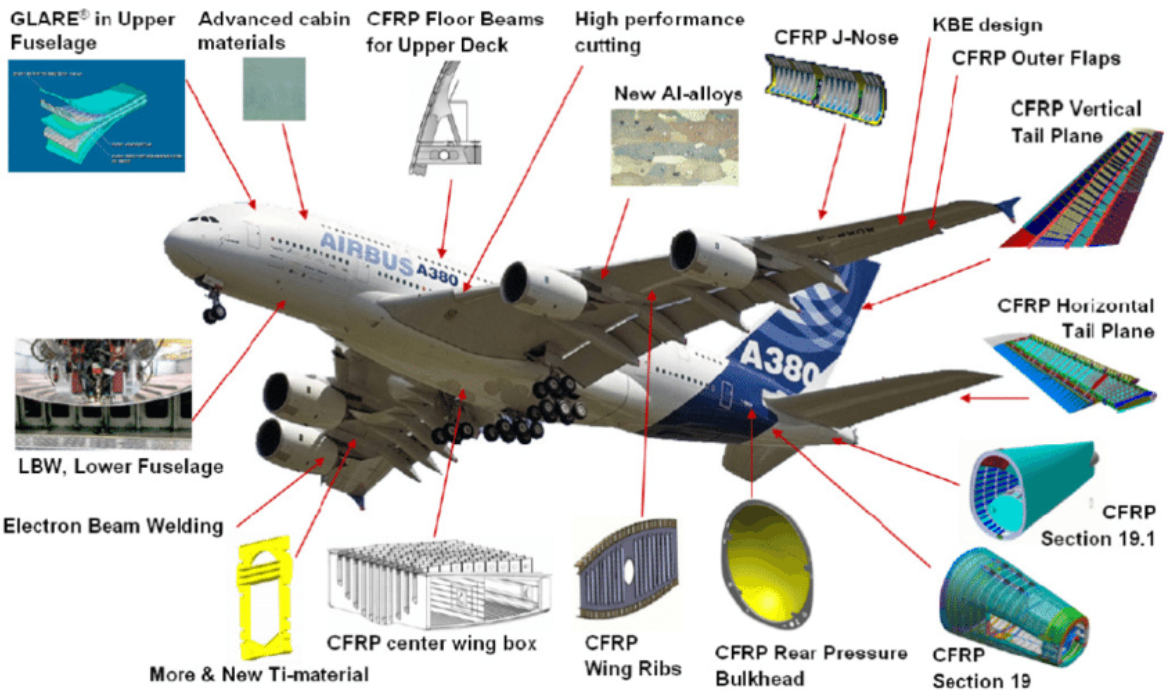


图.空客 380 中的复合材料应用 [picture credit: airbus]

AI4S 实践（12）：《自然•通讯》收录 Monash 大学复合材料原位纳米析出机理研究

金属材料加工过程中内部常常形成第二相析出（一般为金属间化合物或者陶瓷相），形成原位纳米颗粒增强复合材料。

目前，原位纳米析出增强的原子尺度的机理研究还不够深入，存在许多未理解清楚的问题，从而阻碍了材料的理性设计。热处理过程中的纳米析出过程，不仅涉及局部成分的起伏、局域结构的演变，还涉及与缺陷的相互作用（例如位错、空位等），过程十分复杂，需要全新的研究方法来攻克这一难题。

澳大利亚蒙纳士大学研究团队结合原子尺度成像、深度势能分子动力学模拟及经典成核理论计算等方法，研究了铝铜合金（Al-1.7at.%Cu）中的强化析出相 θ′的成核机理。

研究结果表明，空位可以显著促进 θ′依附于先析出的 GPII 相通过模板定向固态形核（TDN）。样品表面以及电子束辐照均可以在合金内产生显著的空位，从而在透射电镜中很容易产生 θ′相的 TDN 形核生长。相比而言，正常的体材料内部空位浓度则难以促使这种 TDN 的发生。

这种形核新机理的发现显著提升了人们对合金析出相演变过程的理解，对原位纳米颗粒增强复合材料的设计提供了理论参考，有利于进一步提升此类材料的研发效率。

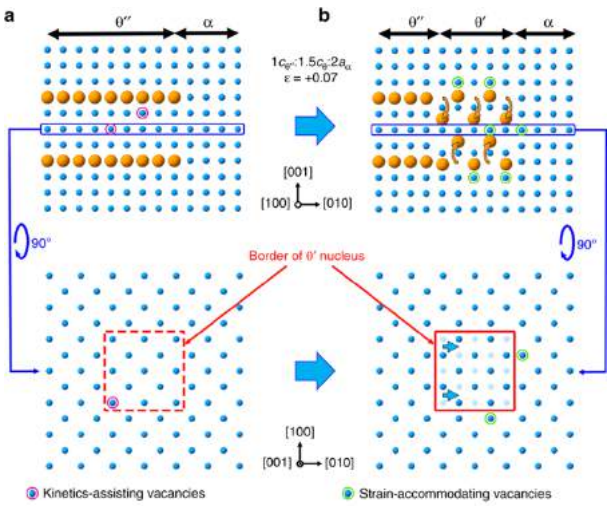


图. 成核途径及机理推测

Source：

Bourgeois, L., Zhang, Y., Zhang, Z. et al. Transforming solid-state precipitates via excess vacancies. Nat Commun 11, 1248 (2020). https://doi.org/10.1038/s41467-020-15087-1

3.6 AI4S 赋能材料研发的 De Novo Design

随着 AI4S 技术的进一步发展，材料研发将进入下一个时代。得益于 AI4S 算法的进步，算力发展和数据、模型库等方面基础设施的建设，会有一批材料设计的算法应用进入产业落地。AI4S 算法结合自动化实验手段，有望将材料研发由过去的高通量试错，转变为"大规模计算设计-定向实验验证"的干湿实验闭合范式。

如前文所说，材料的正向预测是指，通过成分和工艺参数确定微观组织，进一步通过微观组织预测材料性能；而更理想的材料发现模式 —— "逆向设计"，则是求解正向预测的反问题，即根据材料的性能需求逆向推导材料的组分、结构和工艺，也可称为 de novo design / 理性设计。

由于实验数据注定稀疏、非标且昂贵，通过大量而离散的数据积累来实现"逆向设计"的可能性极低。相比之下，物理模型更加稳定且连续，从性质需求出发，推导到物理空间，在物理空间中搜索出潜在的材料组分和结构，在理论可行性上则高出很多。这也是 AI4S 在更长期内有望实现的突破。

由于材料研发面临着多尺度耦合问题，若想在物理空间内对新材料的组分和结构进行"搜索"，则 AI4S 必须在几个算法模块实现突破：

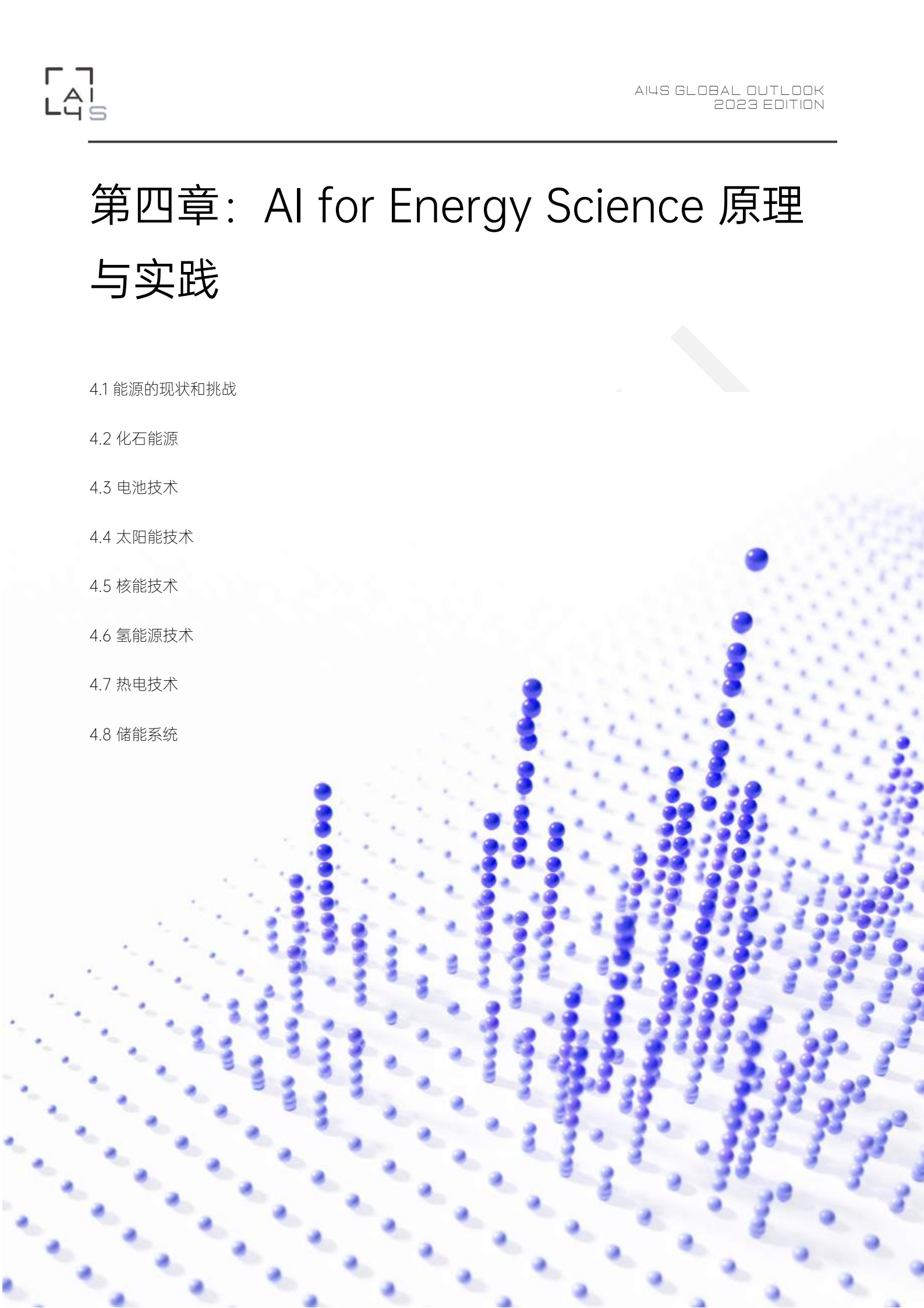
- 基于跨尺度建模思想的材料设计"大模型"，尤其是从微观到介观尺度范围内，主要包括能够有效描述研究对象涉及的所有元素和构型的高精度 DFT、势函数（如 DPA）、介观模型；
- 与实验表征对接的性质计算（如 dflow）；
- 不同尺度模型的有机联动，微观模型作为宏观模型的输入参数和搜索空间，通过可微分模型实现尺度间模型联动的自动化；
- 基于对微观分布的高效采样算法设计的反问题解法，也就是组分和结构搜索算法；
- 对于未知分子/材料的结构与性质预测的预训练模型（如 Uni-Mol）；

在这几个领域里，目前科学家们已经进行了诸多积累和尝试，形成了一系列高价值的算法和工具基础，如 DeePMD（模型构建），ABACUS（跨尺度联动），RiD（强化采样）。此外结构搜索领域也出现了诸如 CALYPSO 这样的工具。而根据相关开发者的表述，上述各领域算法的结合目前也已经迈出了第一步。

最后，随着设备智能化程度和机器控制技术的不断提高，由 AI4S 辅助的自动化实验室正走进现实。AI4S 有望通过高效的边缘计算和高精度物理建模能力实时处理实验产生的数据进行分析并引导实验设计的智能实时调参，从传统网格化的 Design of Experiment 演进为更理性的实验设计，减少无效实验，大幅提高实验效率和投入产出比。

结合 AI+多尺度物理模型+自动化实验，AI4S 有望逐渐将材料的理性设计、甚至生成式设计变为现实。届时，人们有望从需求端、应用端出发设计材料，进而在众多相关行业建立新的技术标准和商业模式。

第四章：AI for Energy Science 原理与实践



4.1 能源行业的现状和挑战

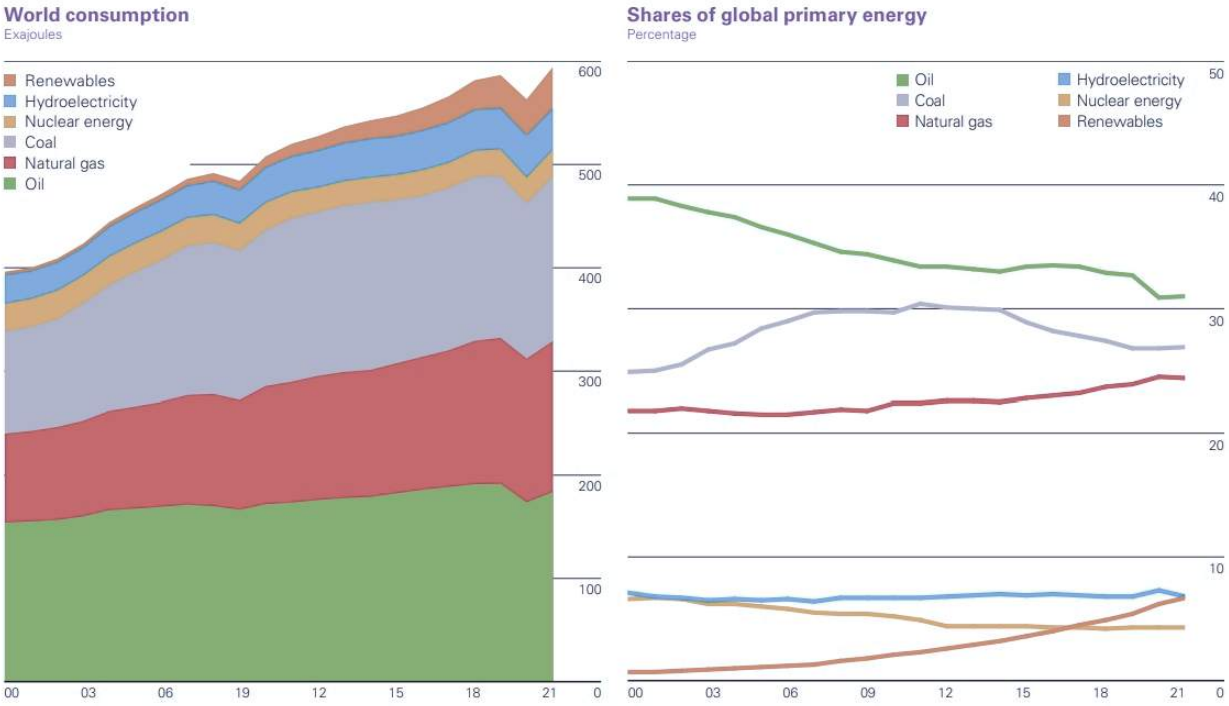


图. 世界能源使用情况和比例 （source: BP[2]）

能源是人类社会发展的物质基础，是现代社会的血液。十八世纪以后，煤炭、石油、电力的广泛使用，先后推动了第一、第二次工业革命，使人类社会从农耕文明迈向工业文明，能源从此成为世界经济发展的重要动力。当今世界，化石能源大量使用，带来环境、生态和全球气候变化等领域一系列问题，加快能源转型发展、大力发展清洁低碳能源是未来趋势所在。加快能源新技术与现代信息、材料和先进制造技术深度融合，随着太阳能、新能源汽车等技术的不断成熟，大规模储能、氢能源、核能等技术有望进一步突破，丰富能源利用新模式。[1]

除了发展新能源之外，推进化石能源的清洁利用也是能源革命战略中的重要一环。加强对化石能源使用过程中的污染控制，推动其清洁高效的开发利用。针对上述的发展要求，除了要从产业链整体出发外，还需要从科学的角度突破当前能源材料的技术难点。大规模数据的分析和处理，跨尺度物理模型的构建等等，高效地探索能源材料"组分-结构-工艺-性能"之间的关系，加速能源科学研发，让高效利用清洁、安全、多样能源的未来不再遥远。

Source:

[1]. 发改委, 能源生产和消费革命战略（2016—2030）

[2] BP: Statistical Review of World Energy 2022

4.2 化石能源与 AI4S

现阶段全球范围内，能源消耗状况由不可再生的化石能源主导。2020 年，化石能源消耗占总能源的 83%。[1] 其中，燃烧是最常见的能耗方式，各类交通运输（航运、陆运、空运等）、发电、日常生活中取暖烹饪等都与之息息相关。

燃烧是十分复杂的化学过程，由于对燃烧的研究和认识的不充分，导致燃烧效率的提高面临瓶颈期。以最常见的汽轮发动机举例，自 2000 年以来，能量转化效率一直在 40%左右，难以有显著提高 [4]。近年来，均质压燃发动机、爆轰发动机等新型的发动机技术成为了燃烧领域的热点问题，致力于提高燃烧效率，并降低污染物排放。

从基础科学研究的角度来说，各类燃烧器中流体力学、传热学和燃烧化学等方面的研究对于燃料的高效利用有着较大的影响。

从研究尺度的方面来说，燃烧是典型的多尺度问题，涉及量子尺度的基元反应和物性参数计算、分子尺度的分子碰撞和化学反应路径、宏观尺度的流体流动，以及各尺度效应在实际燃烧器中的耦合

I. AI4S+流体力学/燃烧流体仿真

在航空发动机等尖端设备的设计中，对燃烧中流体力学的研究至关重要。传统的数值方法对于湍流等复杂模型的处理效果和效率均不能令人满意。利用 AI 负责流体力学的思路正在蓬勃发展。

这当中开创性的工作来自国际燃烧学会主席 Thierry Poinsot 团队。大涡模拟（Large-Eddy Simulation，LES）是常用的湍流模型，为平衡模拟准确性以及计算复杂度，模型定义了一个截止尺度，用于判断"多大的涡才算是大涡"。小于截止尺度的涡对流场结构的影响通过子网格尺度（Sub-Grid Scale，SGS）贡献项进行体现。LES 模型的可靠性取决于对 SGS 贡献的准确建模，然而截止尺度的降低会产生大量计算成本，因此在火焰表面密度估计中快速准确地对 SGS 贡献进行建模的能力将非常有价值。[4] 2019 年他们提出将卷积神经网络用于湍流燃烧建模中，通过借助该 AI4S 模型详细研究 SGS 对预混合湍流火焰反应速率的贡献，预测 SGS 的皱折。该模型在 150 个 epoch（时期）内收敛，极大地降低了计算成本，同时在火焰拓扑性质的提取上，表现出优于经典代数模型的结果。[5]

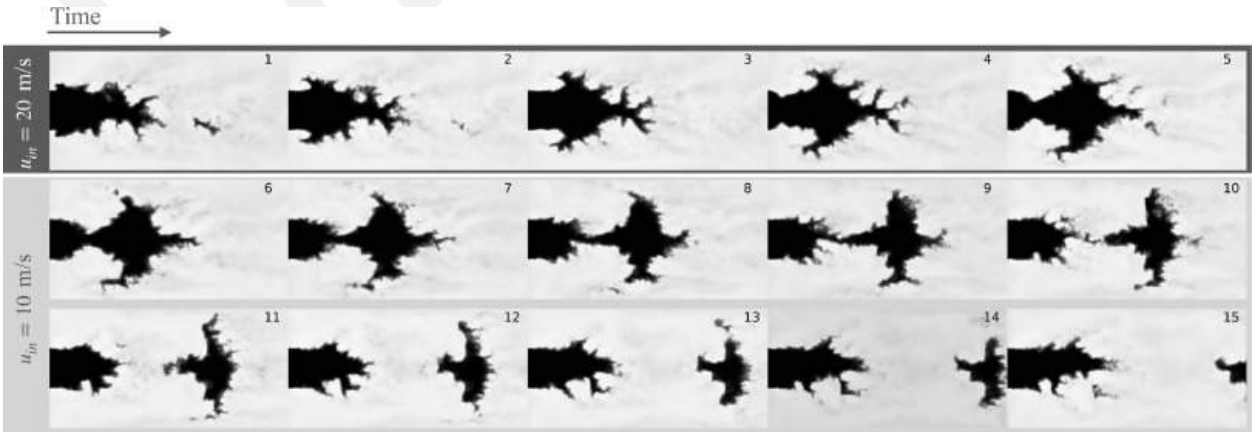


图. 基于 CNNs 进行湍流建模，模型模拟从未燃烧气体(黑色)到燃烧气体(白色)的转变过程[5]

据王建春博士于《力学者说》论坛中所报告，目前机器学习在流体领域已经广泛应用于：雷诺平均模型（如基张量神经网络模型，AI+Spalart-Allmaras 模型，局部神经网络模型）；大涡模拟模型（隐式，半显式，反卷积模型，AI 能流约束模型）；湍流快速模拟（傅里叶神经算子，注意力机制）。结果非常喜人，在效率和精度的提高上均实现了对传统数值方法的成本提升。[6]

与之同时，在业界，先行者也已开始使用 AI 处理复杂的流体问题。西门子能源和英伟达合作，对其大型热回收蒸汽发生器进行数字孪生建模，并通过收集各位置的压力，温度，流速数据，并使用英伟达 physics-ML 模型来高保真度的实现对蒸汽/水在管道中的实时仿真。研究者称，过去这样精度的计算需花费 8 周，现在只需要数小时，该成果可将维修导致的停转时间（downtime）降低 70%，为行业节约每年 17 亿美元的损失。[7]

"燃烧流体"或"燃烧流体动力学"是一种非常复杂的科学问题，因为它涉及到大量的化学反应、流体动力学、热传导和质量传递等多个交织在一起的物理过程。其难点在于 1）复杂的化学反应：燃烧是一种化学反应，涉及到大量的燃料和氧化剂分子在极高温度下的化学反应，这些反应是非线性的，并且经常有大量的反应参与，需要精确模拟和控制。2）流体动力学：燃烧通常发生在流动的气体中，需要考虑到流体动力学的影响，这也是一个非常复杂的问题，涉及到湍流、涡旋、冲击波等多种复杂的流体现象。3）热传导和质量传递：燃烧会产生大量的热量，需要考虑到热量的传导和扩散，同时，燃烧过程中会产生新的物质，需要考虑到物质的扩散和传递。

传统的科学计算方法难以处理其复杂的耦合过程，导致仿真的效率和精度均不理想。如前文所述，AI 也可以通过学习流体动力学的基本规律，来预测和模拟流体的行为，提高模拟的准确性。同时，AI 也可以通过深度学习等方法，来处理高维度的数据，处理燃烧过程中的热传导和质量传递问题。

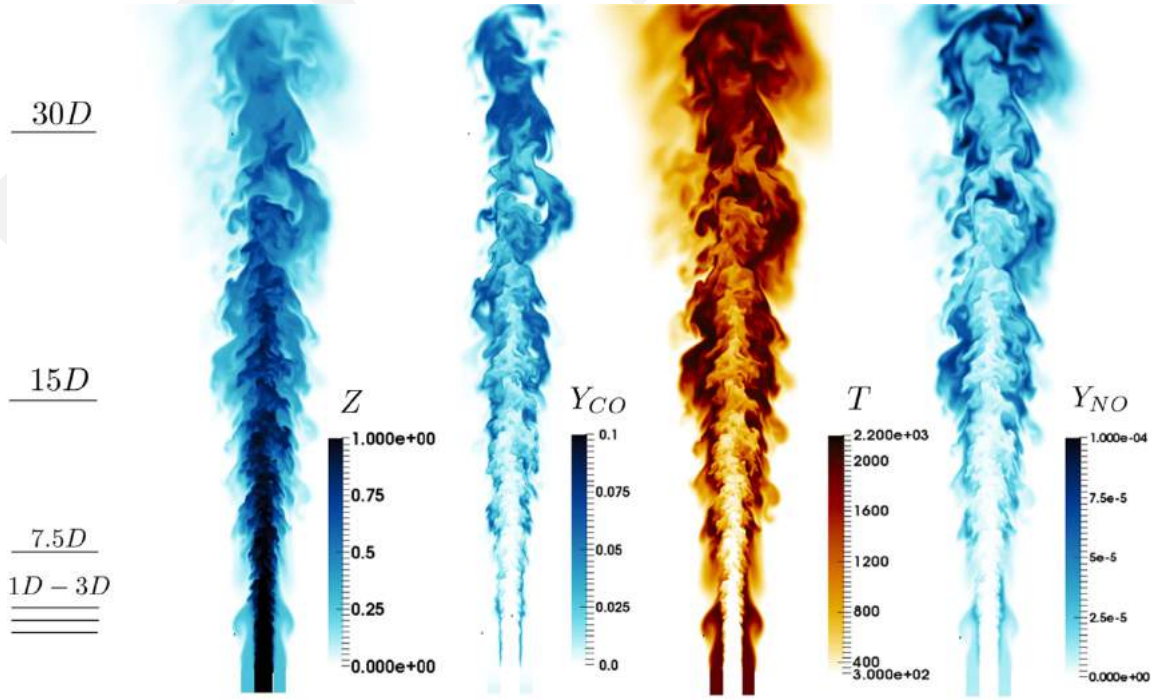


图. 基于第一性的流体-反应动力学耦合计算的燃烧仿真, 来自 Perrine Pepiot et al. [21]

II. AI4S+燃烧反应过程

燃烧化学机理体现了燃烧过程中的物质生成与转化以及能量释放过程，是一个典型的非线性、高维度的体系。近二十年来，化学领域针对燃烧化学反应设计的反应机理从单步反应发展成为组分数量上千、反应数量近万的庞大体系。传统计算模拟方法在刻画燃烧反应及相应的产物方面存在精度和效率的瓶颈问题。

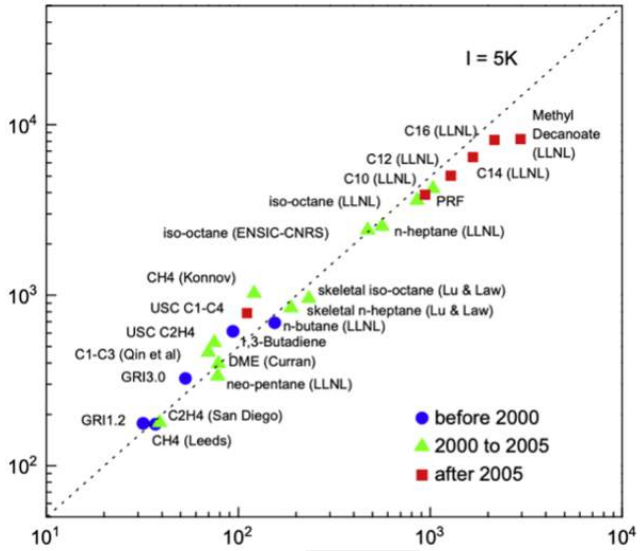


图. 燃烧反应过程异常复杂, x=number of species, y=number of reactions[8]

大量燃烧反应数据积累促成燃烧大数据时代的到来。伴随着各个尺度下物理模型的快速发展，具有强大自学习和拟合能力的 AI 手段为燃烧研究带来曙光。

如麻省理工的 William H. Green 教授专注于利用计算模拟预测时间演变下化学反应过程及反应产物；清华大学杨斌教授专注于燃烧反应动力学的研究，包括模型构建、反应路径调控等；大连理工大学贾明教授专注于利用数值模拟实现实用燃料的化学反应机理研究……科学家们开始着力于 AI 在燃烧机理的探索和研究，同时也取得了不少喜人的成果。AI4S 方法在燃烧领域的发展，有望克服传统方法“算不准，算不动”的核心困难，并深化科学家对燃烧反应机制的理解，优化燃料利用方案，指导高性能发动机的设计制造，拓展化石燃料的应用研究边界。

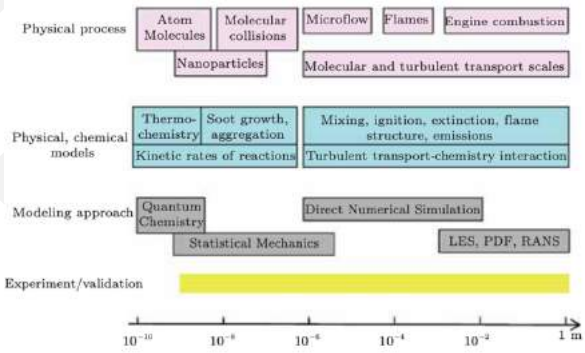


图. 燃烧过程中涉及的空间尺度[9]

III. AI4S+燃烧污染机理研究

除了助力燃烧过程本身的研究以外，AI4S 为燃烧产物的研究提供了新的路径。每年，我们都能从新闻中听到北美遭遇各种不同程度的火灾。2023 年，故事依旧继续，北美再次面临"前所未有"的大火灾难。从加利福尼亚到华盛顿，从美国西海岸到东海岸，以及加拿大，无数生命和财产在这场毁灭性的大火面前备受威胁。火情导致至少 36 人丧生，数千座房屋被毁。这场大火引发的烟雾已经扩散到美国大部分地区，甚至在东海岸的纽约和华盛顿，也能发现烟雾迹象。此外，加拿大也未能幸免，全国已有超过 2 万人被迫撤离。这场灾难性的大火不仅影响了居民的生活，空气质量的严重恶化也让人们对公共卫生问题感到担忧。

地球"生病了"，在全球极端天气愈发频繁的今天，人们更加期盼科学能对"病理"做出诊断，并给出有效的应对。在山火造成的污染中包含煤烟颗粒（Soot），由黑碳和其他物质构成，可以长距离传播，影响空气质量，干扰航行，同时也会加剧全球气候变化。更糟糕的是，这些颗粒通常含有多环芳烃（PAHs），这是公认的致癌物，对人体健康构成附加风险。而减少碳烟排放取决于对从燃料到火焰中碳烟颗粒的物理和化学途径的认识。[10]

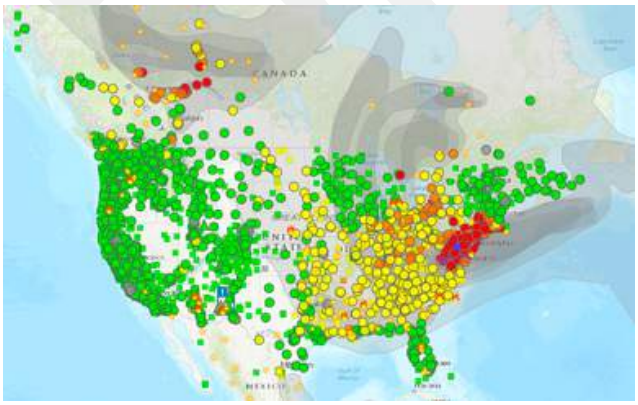


图. 2023 年北美大火期间的空气质量数据：美国环保局数据
EPA Fire and Smoke Map on AQI for June 6, 2023

PAHs 更为人所知的是其也会在高温烹调中出现，也是专家建议少使用烧烤类食物的主要原因之一。根据香港食品安全署的报告，当脂肪抵受摄氏 200 度以上的热力时会分解， 产生 PAHs。摄氏 500 至 900 度的高温， 尤其是摄氏 700 度以上， 最有利 PAHs 形成。烹煮食物的温度越高， 会产生越多的 PAHs 。其他有机物质(例如蛋白质和碳水化合物) 受热分解时也会产生 PAHs ， 但脂肪受热分解时所产生的 PAHs 最多。肉类直接接触火焰时， 所含的脂肪会分解而产生 PAHs 。此外， 食物的脂肪溶化后滴在热源上， 也会产生 PAHs ，然后随着升起来的烟又落到肉面上。

科学家们一直在研究燃烧过程中的煤烟形成和 PAHs 排放。传统的实验室方法虽然提供了宝贵的数据，但其局限性也相当明显。首先，燃烧和野火等极端环境下的原位实验非常具有挑战性。其次，从理论方面模拟煤烟形成和 PAHs 排放也充满了难题。这些过程非常复杂，涉及到极端环境下物理扩散与化学反应的复杂耦合，这些模拟需要巨大的计算资源和时间，成本十分昂贵。此外，模型中的不准确性可能导致大的误差，限制了其预测的可靠性。

人工智能（AI）的崛起为这个问题提供了新的解决方案。使用 AI 解决科学问题，即 AI for Science，在今年受到各界关注，其中的一个代表性技术，是 DeePMD 深度势能分子动力学 —— 这种基于深度学习的模型为我们理解和模拟燃烧过程中的煤烟形成和 PAHs 排放提供了独特的方法。DeePMD 就像一个高清摄影机，能捕捉到煤炭和多环芳烃在火灾中形成的每一个微妙的步骤，为科学家提供了前所未有的细节。DeePMD 利用深度学习的力量，有效

地捕捉了这些过程的复杂性，解决了实验方法和传统模拟面临的困难。

目前对于碳烟形成过程还存在着许多未知，许多研究正在进行中。然而碳烟形成过程中化学和物理反应的复杂性使其在数值和实验上都具有挑战性，实验手段难以揭示该过程中的原子级细节，AI 的快速发展为从理论角度理解碳烟的形成提供全新的机会。[11]为探索煤碳烟中多环芳烃(PAHs)的生长过程，上海华东师范大学和上海纽约大学朱通课题组训练开发描述煤燃烧反应过程的第一性原理精度

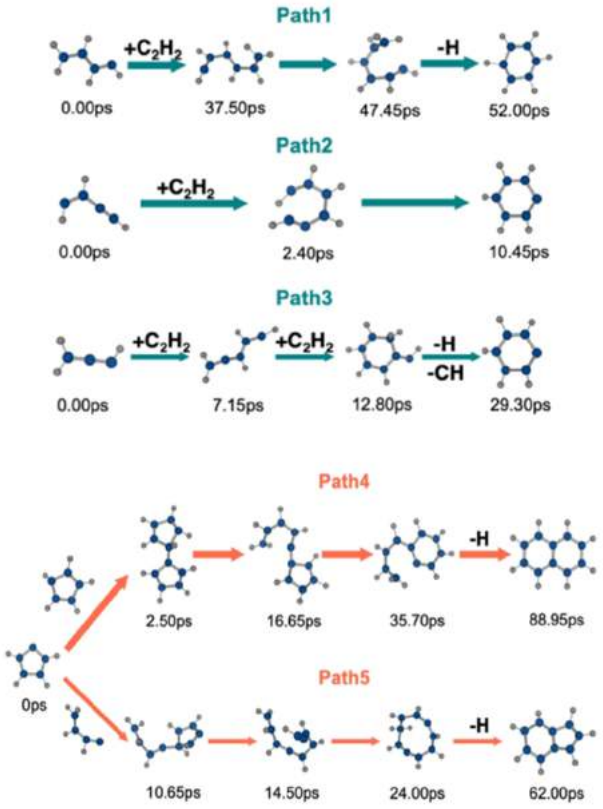


图. PAHs 部分反应路径示意图

AI4S 模型，以 0.1 飞秒的分辨率对 4000 多个原子的煤燃烧反应系统模拟小烃类反应物产生 PAHs 和碳烟的过程。该模拟展现了数十种形成多环芳烃的反应途径，揭示了导致大的多环芳烃形成和增长的关键步骤，提供了碳烟形成初期的详细信息。[11]

另外中国国家爆炸科学与技术重点实验室陈东平团队针对煤碳烟形成前体 PAHs 进行研究，探索 PAHs 引发碳烟成核的机制。团队基于第一性原理数据集训练开发描述 PAHs 反应的 AI4S 模型，并从原子角度研究数十纳米范围内碳烟粒子的演化过程。基于该模型进行分子动力学模拟，团队分析 PAHs 与煤中不同大小的乙炔反应得到的产物，明确物理相互作用增强了化学成核作用，为探索碳烟粒子演化路径提供基础。[12] 除上述提到的 2 篇工作外，相关的研究成果不在少数。AI4S 也真正帮从理论角度为碳烟的形成机制提供丰富的解析，为化石能源的清洁使用提供指引。

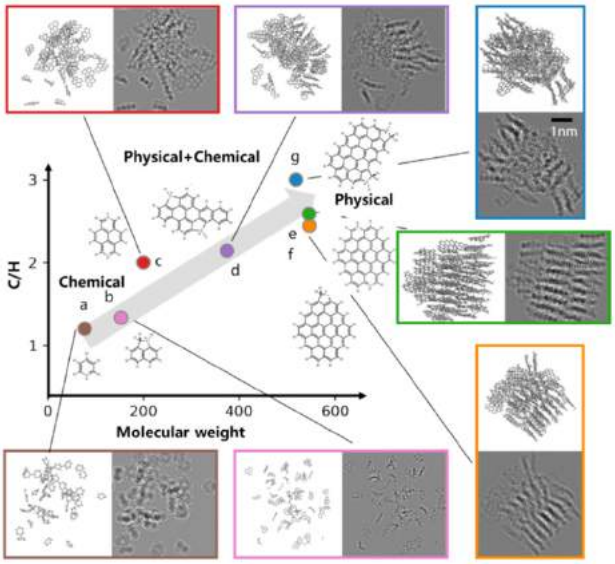


图. PAHs 反应过程产物

IV. AI4S+高能材料

除了研究日常应用场景中的燃烧反应，AI4S 还能帮助探索剧烈的燃烧反应过程，如含能材料、高能炸药等。这类高能量密度材料的反应过程都极为剧烈和复杂，相关的化学反应网络包含了数千甚至上万个基本反应，同时这些反应过程都发生在较大范围的时间尺度上。实验上难以复现并详细剖析具体的化学反应过程，深刻理解其详细演化过程及机制成为巨大的困难。科学家希望在原子水平上详细探究高能量密度材料的反应机理，深入研究其性质，为新型高能量密度材料的设计和发展提供支持。[13]

相关材料体系中，含能材料六硝基六氮杂异伍兹烷(CL-20)常用作推动剂，然而其对环境刺激的高度敏感大大降低了安全性，严重限制其应用。

华东师范大学和上海纽约大学朱通课题组系统地探究了 β-CL-20 和 CL-20/ TNT（2,4,6-三硝基甲苯）共晶体两种含能材料体系的反应过程，基于 AI4S 模型进行分子动力学模拟发现了许多重要的中间组分及其在分解过程中的相关反应路径，明确在 CL-20/TNT 共晶中加入 TNT 分子，使体系中 CL-20 与 TNT 分子之间形成分子间氢键，提高了共晶的热稳定性。

这为 CL-20 与其他含能材料共晶结合后稳定性提高的机理提供丰富的理论指导，从原子角度提出改善 CL-20 安全性的方案，拓展 CL-20 的应用边界。

另外，北京理工大学陈东平团队结合深度势能模型与虚拟现实分子动力学方法研究新型的含能材料—

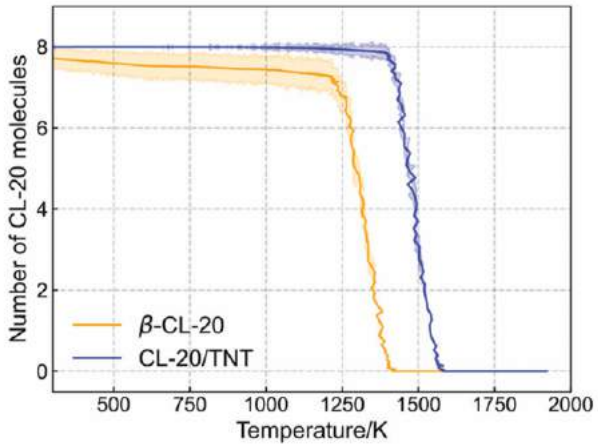


图. 在 β-CL-20 体系，CL-20 分子在温度升高到大约 1200 K 时开始分解，而在 CL-20/TNT 共晶体系中 CL-20 分子在温度升高到 1400 K 左右时才开始分解。CL-20/TNT 共晶体系的热稳定性有明显的提升。[14]

—ICM-102。ICM-102 是近年来国产自研的一类新型含能材料，兼具低感度与高能量密度的优点。然而，目前关于其反应机理与能量释放机制较少，限制了该材料在实际装备中的应用。

团队通过虚拟现实采样以及第一性原理建模，分析 ICM-102 含能分子的分解过程，实现 ICM- 102 分解复杂反应网络的全景可视化，同时提出了 ICM-102 分解生成水的主要机理。

该成果突出了基于神经网络的分子动力学模拟在第一性原理精度下探索极端条件的复杂反应机制，提高实验的真实性和保真度，为发展新型的高能量密度材料提供新方法。[16]

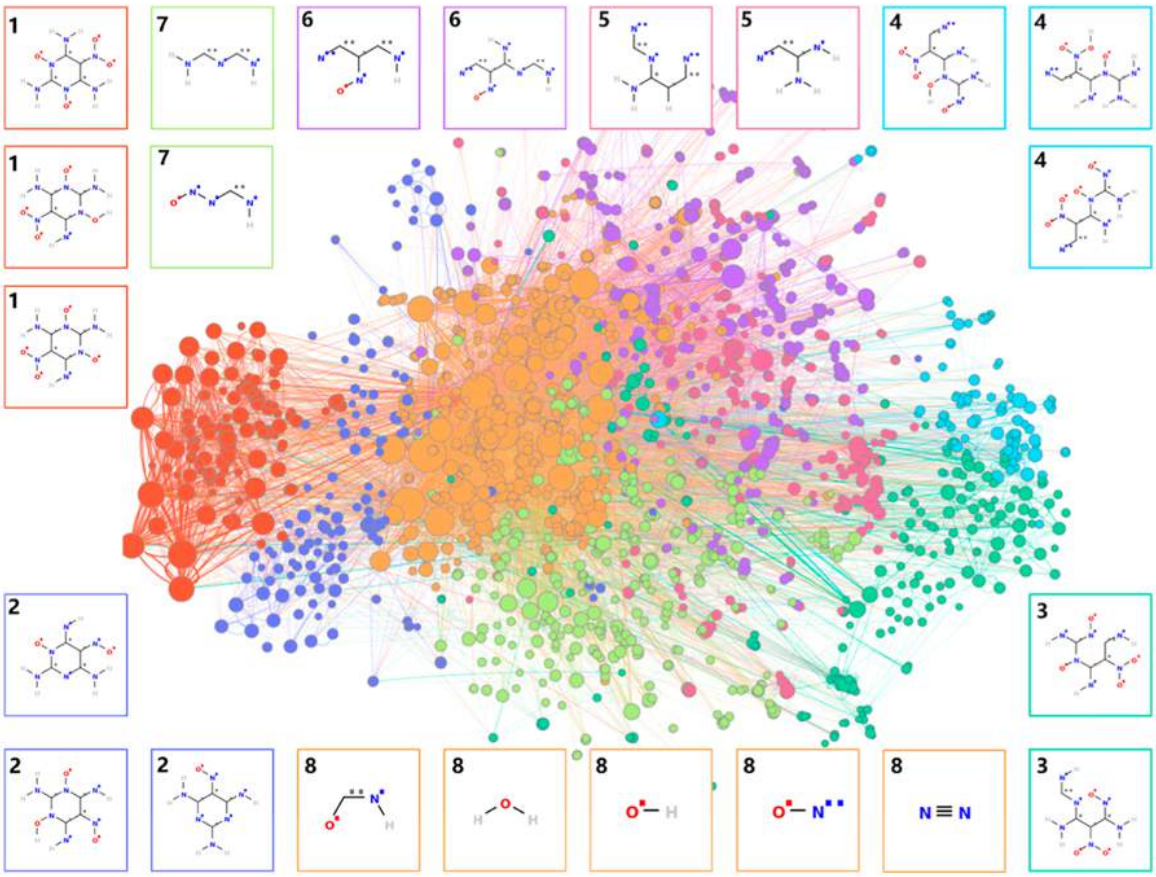


图. ICM-102 分解的全景可视化。每个点和每条线分别代表一个物种和物种之间的反应。网点大小与网络中物种的观测频率成正比。只有频率为>5 的物种被显示出来。插图显示了每个基团的代表性分子结构。[16]

表 5. 燃烧中的科学问题与 AI4S 实践

研究方向	技术难点 & 科学问题	AI4S 应用
反应路径及机制	实验和传统手段难以描述 1 fs 下的化学反应过程	实现超高的时间分辨率的化学反应过程描述，对化学反应产物进行预测[17]
纳米燃料	纳米颗粒在燃料中难以均匀分散，同时伴随着沉积聚集，影响发动机性能	预测添加燃料纳米颗粒对柴油发动机排放特性和性能的影响[18]
燃烧相位	缺少高保真燃烧模型描述	构建数据驱动的线性参数变化模型来预测和控制燃烧相位[19]
燃油喷射	数值模拟优化耗费大量的时间	优化压缩点火发动机的喷射时间和燃料喷射压力
爆震	限制火花点火发动机热效率的提高	依据每次循环预测爆震开始，深入了解随机爆震燃烧
发动机运行条件	传统数值模拟方法耗时长，且不能保证发动机的最优能效	优化重型汽油压燃式发动机的运行条件和活塞顶凹腔设计
发动机性能	发动机整体性能的预测	利用仿真模拟加速发动机设计 [19]
发动机排放	排放产物对环境以及发动机热效率的影响	预测发动机排放产物并进行控制

参考资料

[1] BP 世界能源统计年鉴 2021 (第 70 版)

[2] 西南石油大学 https://www.swpu.edu.cn/info/1015/12279.htm

[3] Yin F, Rao A G. A review of gas turbine engine with inter-stage turbine burner[J]. Progress in Aerospace Sciences, 2020, 121: 100695.

[4] Graphcore https://mp.weixin.qq.com/s/bWPUNd-VqkTFeCI9fVjm3g

[5] Lapeyre C J, Misdariis A, Cazard N, et al. Training convolutional neural networks to estimate turbulent sub-grid scale reaction rates[J]. Combustion and Flame, 2019, 203: 255-264.

[6] 《力学者说》第 2 期 基于机器学习的湍流模型研究进展 https://www.bilibili.com/video/BV1BM4y1A7Ra/?vd_source=4102f6f3c9bce85328a8321286c2b8cc

[7] Nvidia, Siemens Energy Taps NVIDIA to Develop Industrial Digital Twin of Power Plant in Omniverse, https://blogs.nvidia.com/blog/2021/11/15/siemens-energy-nvidia-industrial-digital-twin-power-plant-omniverse/

[8] Lu T, Law C K. Toward accommodating realistic fuel chemistry in large-scale computations[J]. Progress in Energy and Combustion Science, 2009, 35(2): 192-215.

[9] Yiguang J U. Recent progress and challenges in fundamental combustion research[J]. 力学进展, 2014, 44(1): 201402.

[10] Xi J, Yang G, Cai J, et al. A review of recent research results on soot: The formation of a kind of carbon-based material in flames[J]. Frontiers in Materials, 2021, 8: 695485.

[11] Wang B, Zeng J, Cao L, et al. Growth of Polycyclic Aromatic Hydrocarbon and Soot Inception by in silico Simulation[J]. 2022.

[12] Chu Q, Chen D. Toward full ab initio modeling of soot formation in a nanoreactor[J]. Dongping, Toward Full Ab Initio Modeling of Soot Formation in a Nanoreactor, 2022.

[13] Deep Modeling

[14] Cao L, Zeng J, Wang B, et al. Ab initio neural network MD simulation of thermal decomposition of a high energy material CL-20/TNT[J]. Physical Chemistry Chemical Physics, 2022, 24(19): 11801-11811.

[15] L. Zhou et al. , Machine learning for combustion, Energy and AI 7 (2022) 100128

[16] Chu Q, Luo K H, Chen D. Exploring complex reaction networks using neural network-based molecular dynamics simulation[J]. The Journal of Physical Chemistry Letters, 2022, 13: 4052-4057.

[17] Zeng J, Cao L, Xu M, et al. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation[J]. Nature communications, 2020, 11(1): 1-9.

[18] Ağbulut Ü, Gürel A E, Sarıdemir S. Experimental investigation and prediction of performance and emission responses of a CI engine fuelled with different metal-oxide based nanoparticles–diesel blends using different machine learning algorithms[J]. Energy, 2021, 215: 119076.

[19] Irdmousa B K, Rizvi S Z, Veini J M, et al. Data-driven modeling and predictive control of combustion phasing for RCCI engines[C]//2019 American Control Conference (ACC). IEEE, 2019: 1617-1622.

[20] Jay Gould, https://developer.nvidia.com/blog/nvidia-announces-modulus-a-framework-for-developing-physics-ml-models-for-digital-twins/

[21] Prediction of flame structure and pollutant formation of Sandia flame D using Large Eddy Simulation with direct integration of chemical kinetics, Combustion and Flame, Volume 188, 2018, Pages 180-198, ISSN 0010-2180, https://doi.org/10.1016/j.combustflame.2017.08.028.

AI4S 科研实践（13）：DeepFlame —— "AI4S 原生"的燃烧流体仿真

燃烧，特别是多相与湍流中的燃烧，涉及一系列复杂多尺度问题的交叉融合，是长期以来宏观尺度科学计算的痛点领域。基于宏观连续介质 Navier-Stokes 方程数值求解的计算流体力学 CFD 已经在航空航天、气象预测等领域发挥着不可替代的作用。计算机求解连续偏微分方程需要进行数值离散，基于空间网格化的有限体积方法作为应用最为广泛的方法之一，已经在各类国外商业 CFD 软件中实现并工程化，例如 Fluent、StarCCM+等。

另一方面，宏观层面的化学反应动力学计算普遍使用 Arrhenius 公式求解基元反应速率。燃烧系统通常包含成百上千的组分以及更多的基元反应个数，因此需要在时间维度上求解一个大型的常微分方程组。对于此类方程组的求解，国外也已经有类似的商业软件如 ChemkinPro、CosiLab 等。

然而，带有燃烧反应的流体力学计算需要对以上两个方面进行交叉耦合，面临更大挑战。虽然各类商业软件进行了不少功能上的开发与尝试，目前的计算精度和效率都比较低，还难以达到工业应用场景的要求。

基于 AI for Science 研究范式，DeepFlame 项目整理 OpenFOAM、Cantera、Torch 等开源平台，结合异构并行与 AI 加速器等新一代算力基础设施，旨在建设高精度、高效率、简单易用、覆盖面广的燃烧反应流的数值模拟程序，由北京大学陈帜发起和主导开发，南方科技大学张天汉和上海交通大学许志钦提供 AI 建模支持，依托北京科学智能研究院（AISI）的科研合作平台，由 DeepModeling 社区提供代码托管及相关技术支持。

在融合 AI 功能方面，项目组完成了一个示范性的验证。对于限制计算速度的化学反应速率刚性问题，DeepFlame 既可以调用 Cantera 集成的 SUNDIALS CVODE 求解器[3]，也能够兼容 pyTorch 原生格式的神经网络替代模型（相关算法可参考[4]）。

Source: Runze Mao, etc, DeepFlame: A deep learning empowered open-source platform for reacting flow simulations, Computer Physics Communications, Volume 291, 2023, 108842

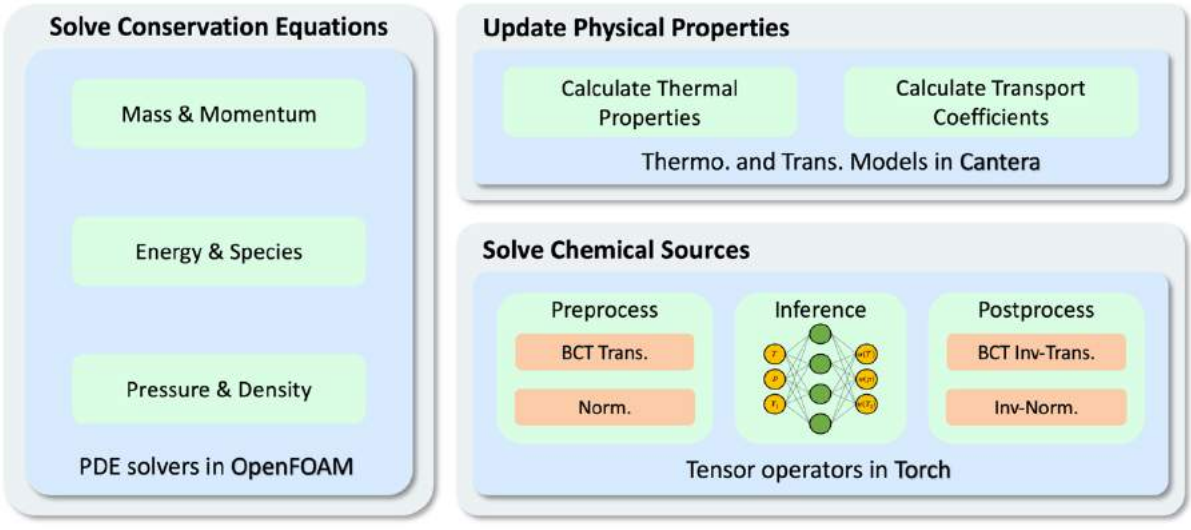


图. DeepFlame 融合 AI 与 PDE 科学计算的示意图

AI4S 科研实践（14）：《自然•通讯》收录华东师范大学团队航空发动机燃烧反应路径研究

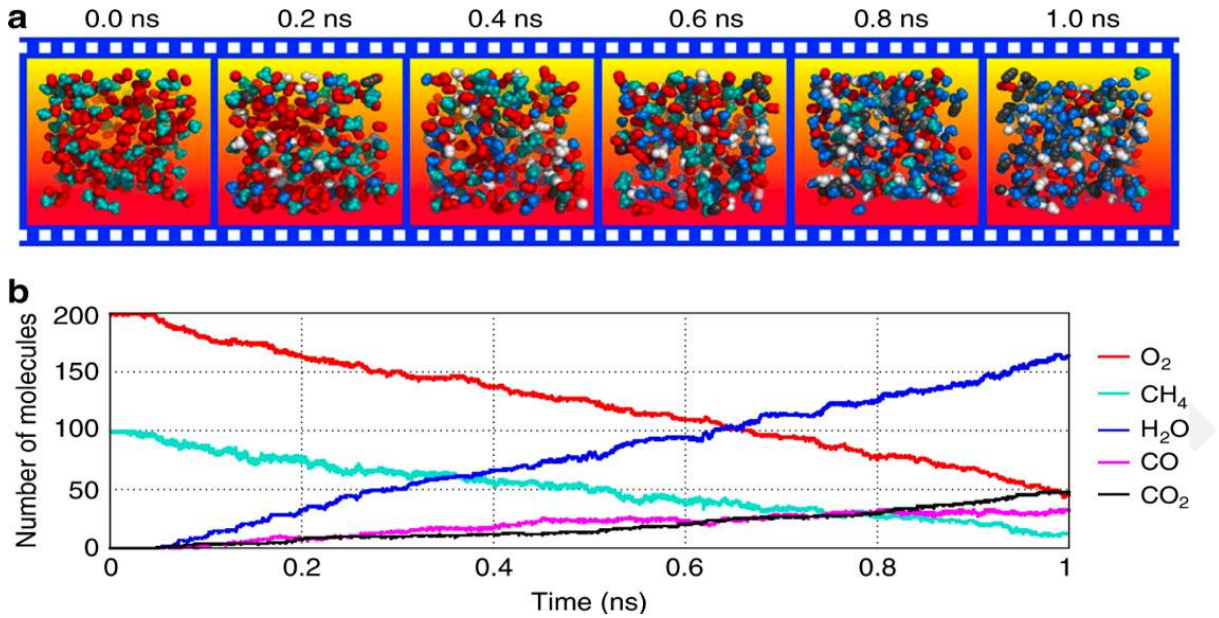


图. 燃烧反应产物随时间变化[1]

航空发动机是国防、交通等领域的核心装备，其水平代表了一个国家的科技和工业实力。研究发动机往往离不开研究发动机内部的燃烧反应。发现和解析燃料燃烧的反应机理，对发动机设计和性能优化具有重要的意义。然而实际情况下，发动机都通常是在高温高压的严苛工况下进行运作，这种条件实验手段难以复现，同时传统计算模拟方法无法承担实际工况的完全刻画，其全面研究更是巨大的困难。

随着 AI 在科学计算领域的蓬勃发展，人们看到了 AI4S 方法对大规模准确研究燃烧反应机理的全新可能。2020 年，华东师范大学朱通老师课题组利用 AI4S 模型成功探索烷烃燃烧反应路径。通过构建甲烷 AI4S 模型，描述燃烧中的各种复杂反应，复现多年来积累的甲烷燃烧骨架反应机理，还创新性地发现了数百个实验数据库中没有的新反应，进一波揭示甲烷燃烧的完整反应网络。

另外，他们还研究了拥有 35,496 个结构的正十二烷热解机理。正十二烷主要作为有机溶剂，蒸馏剂和柴油进行使用。通过构建了正十二烷的 AI4S 模型，模拟正十二烷的热解过程，打开正十二烷的应用边界。相关工作分别在《Nature Communications》[1] 和《Energy & Fuels》[2] 上进行发表。

Source:

[1] Zeng, J., Cao, L., Xu, M. et al. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. Nat Commun 11, 5713 (2020). https://doi.org/10.1038/s41467-020-19497-z

[2] Exploring the Chemical Space of Linear Alkane Pyrolysis via Deep Potential GENerator, Jinzhe Zeng, Linfeng Zhang, Han Wang, and Tong Zhu, Energy & Fuels 2021 35 (1), 762-769, DOI: 10.1021/acs.energyfuels.0c03211

4.3 电池与 AI4S

4.3.1 电池研发的特点：多场景，多尺度，多技术栈

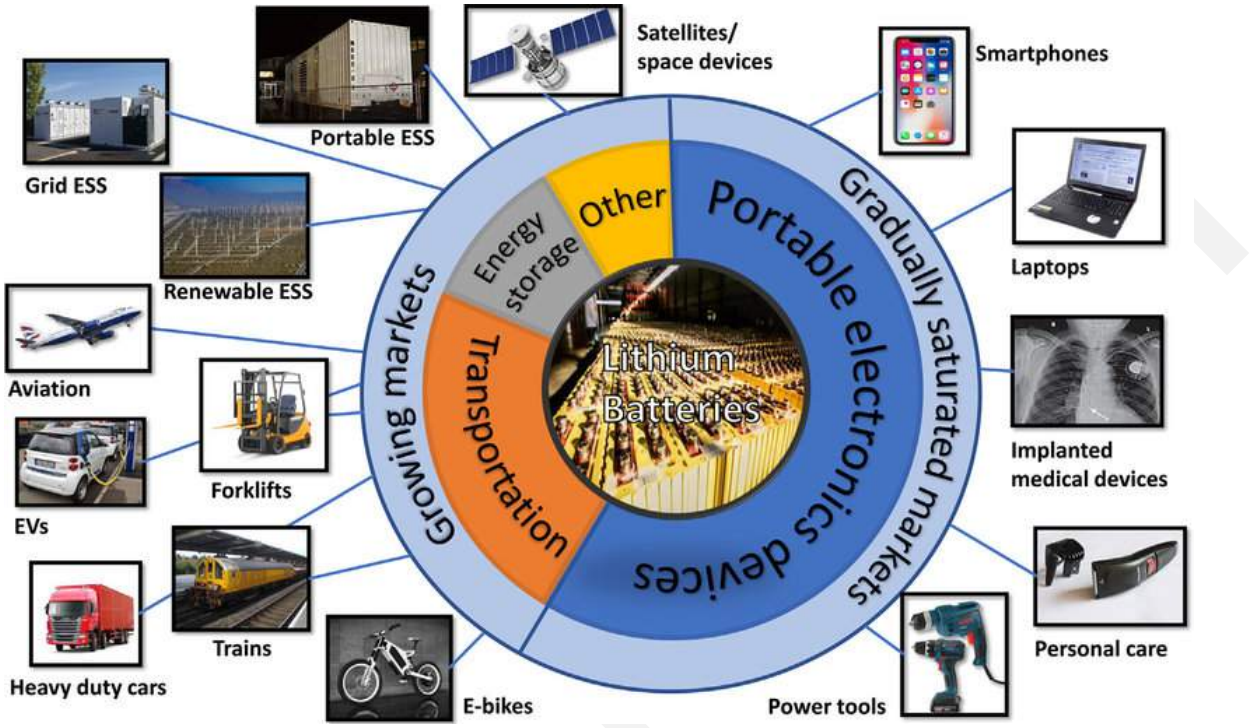


图. 电池的应用场景十分广泛 [1]

电能的大规模使用是人类文明进步的重要标志。2019 年诺贝尔化学奖认可了锂电池对人类文明发展的重要价值和贡献，这离不开数十年来许多科研人员的付出。

（一）电池的应用场景非常多样：随着消费电子，新能源汽车，储能等需求的飞速增长，产业对电池性能的要求愈发多样。新型电池研发的需求正快速膨胀。

（二）电池是典型的跨尺度科学研究体系：除了研究电极材料本身的不同原子间的相互作用外，还需要研究正负极和电解质之间的相互作用，充放电过程中电池材料本身的组分变化，以及组装成整个电芯之后，周围环境对电芯的影响等，这些问题都需要借助跨尺度的手段来解决。过去实验方法研发投入大、周期长，难以实现快速的试错验证迭代过程。而传统计算模拟方法中，基于量子力学的计算模拟受制于计算量，计算规模局限在数百原子，时间尺度和统计能力严重不足；基于分子动力学的计算模拟又面临着准确率低、调参难度大等问题。AI4S 研发范式通过生成具有第一性原理精度的跨尺度物理模型，高效求解出不同规模体系下材料的结构和反应动态过程，贴近真实应用场景，指导工艺生产。

电池研发的关键是在低成本条件下提高安全性和能量密度。安全性是电池质量的"底线"，缺乏安全性

也就没有任何应用场景可言。而随着越来越多应用场景关注到电池的便携性及使用过程中不产生污染物排放的优势和产业价值，对电池的需求开始变得越来越迫切，高能量密度的电池设计是基本要求。能量密度是电池容量的标志，容量越大意味着电池续航时间和里程更长，实现在较少充电次数下的长时间运行。除此之外，同其他产品商业化一样，电池成本的考量是不可避免。低成本是电池大规模商业化的重要前提，这意味着在较低的研发投入下，研究的电池体系能实现在短时间内快速充电以及较高的循环寿命。

过去几十年，数以十亿计的资金投入到电池研发中，使得能量密度和安全性得到大幅提高，进而推动了手机、笔记本电脑以及混合动力和电动汽车（EV）等变革性技术。然而要面向更多的应用场景，大规模实现完全电动化则需要优化电池设计，在降低研发投入的同时进一步提高能量密度、安全性和循环寿命。

（三）电池的技术路线也非常多样：在锂电池体系下正极负极电解质均有不同的技术路线各有侧重；此外，也有钠电池，燃料电池等新型路线正被探索

Source：

[1] Huizhe Niu et al. Recent Advances in Application of Ionic Liquids in Electrolyte of Lithium Ion Batteries, Journal of Energy Storage,

https://doi.org/10.1016/j.est.2021.102659.

[2] J. M. Tarascon, M. Armand; Nature, 2001

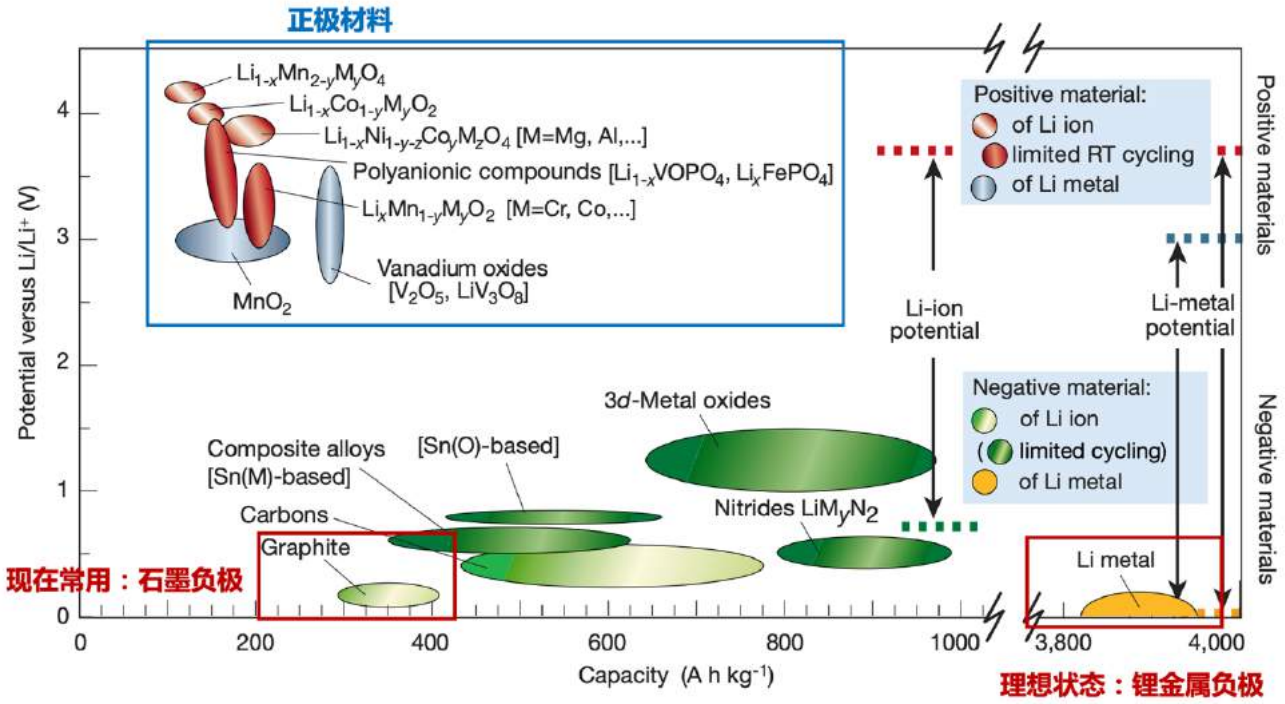


图. 电池正负极材料的技术路线和理论性能 [2]

4.3.2 AI4S 解决电池研发的"多尺度"与"干湿结合"难题，加快能源新材料开发应用效率

Source: DP.Technology

电池由正极、负极、电解质、隔膜以及其他非活性物质构成。作为电化学储能器件，电池工作的核心目标是实现化学能和电能的转换：充电时锂离子从正极材料中脱出并迁移到负极，同时电子在外电路从正极迁移到负极，靠近负极的锂离子通过获得外电路的电子而嵌入到负极材料中。放电过程则与之相反[1]。 新型高能量密度可充放锂电池的发展，需要开发高性能新材料体系和优化电池设计。无论是根据已有的知识经验进行实验验证的研发模式，亦或是基于计算模拟手段进行材料设计的研发模式，他们都需要对锂电池中的基础科学问题进行深入地研究[2]。

针对不同的性能指标和设计目标，业界不同的企业也开始分别发力：斯坦福大学崔屹教授成立的 Amprius 公司，主要是生产以纳米硅负极材料的锂离子电池，通过将硅负极纳米化以降低膨胀带来的影响。清陶能源、QuantumScape、Solid Power 等企业则致力于研发锂电池的固态电解质材料，提高锂电池的能量密度，增长循环寿命。除此之外，

还有不少企业致力于高能量密度的正极材料研究，如国内有名的容百科技、当升科技等。

在全球面临能源危机和环境挑战的当下，能源材料的研究显得尤为重要。这类材料在电池、太阳能电池和各种能源材料中发挥着核心作用，对于实现可持续能源解决方案具有决定性影响。然而，传统范式下能源材料研究速度相对缓慢，从理论概念到实际产品的转化往往需要长达几十年的时间。例如，可充电锂电池的概念在 1970 年代首次提出，但直到 1991 年，索尼才成功地将其商业化，期间历经了大约二十年的时间。太阳能电池的发展过程也是如此。光伏效应最早在 1839 年由法国物理学家埃德蒙·贝克勒尔首次发现。然而，将阳光直接转化为电能的第一个实用光伏电池，却是在 1954 年由贝尔实验室研制出来的。从基础科学的发现到实际应用的开发，经历了一个多世纪的时间。

能源材料研究的挑战主要源于其复杂性和多样性。这些材料的微观结构与性能表现之间的关系错综复杂，人工经验和传统实验手段难以精确揭示其内在联系。此外，能源材料的合成和优化往往需要大量的实验和时间成本，限制了研究的进展。

锂电池的主要部分是正极、负极，电解液和隔膜。其中，正、负极材料决定电池能量密度，电解液和隔膜使锂离子自由穿梭从而完成电路的“闭环”，其对于电池安全起核心作用。目前，电动汽车主要使用的是石墨负极的锂离子电池。这种电池体系的能量密度和充电速度都受到其材料性能的限制。石墨负极的锂离子电池已经逐渐接近其理论极限，这就意味着仅仅依靠优化现有的电池设计和制造工艺来进一步提高电池性能最多只能实现“微优化”，而无法达到“大突破”。

要想实现显著的性能提升，我们只能求助于开发新的电池材料体系。例如，硅负极和锂金属负极的电池体系具有比现有的石墨负极的锂离子电池更高（3~10 倍）的理论能量密度，因此具有巨大的潜力。然而，这些新材料体系的商业化应用还面临着重重挑战。比如，硅在锂离子插入/脱出过程中体积膨胀的问题，锂金属在充电过程中形成的锂枝晶的问题，都对电池的稳定性和安全性构成了威胁。所以，要解决电动汽车的里程焦虑问题，最根本的是需要在电池材料科学方面取得重大突破。

值得注意的是，当我们想到材料研发时，可能第一个进入我们大脑的画面是实验室和显微镜。然而由于电池的“动态”，显微镜并不足以支持电池的研发。事实上，领先的电池厂商大都在实验室配备了先进电镜等表征设备。这些设备可以在材料工程师提出设计并制备后观测验证其微观的原子排布结构，进而佐证其材料设计和工艺设计的合理性。

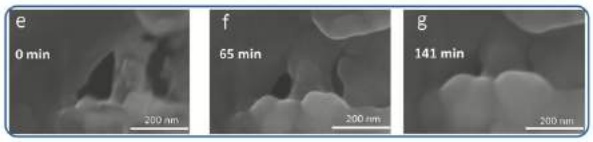
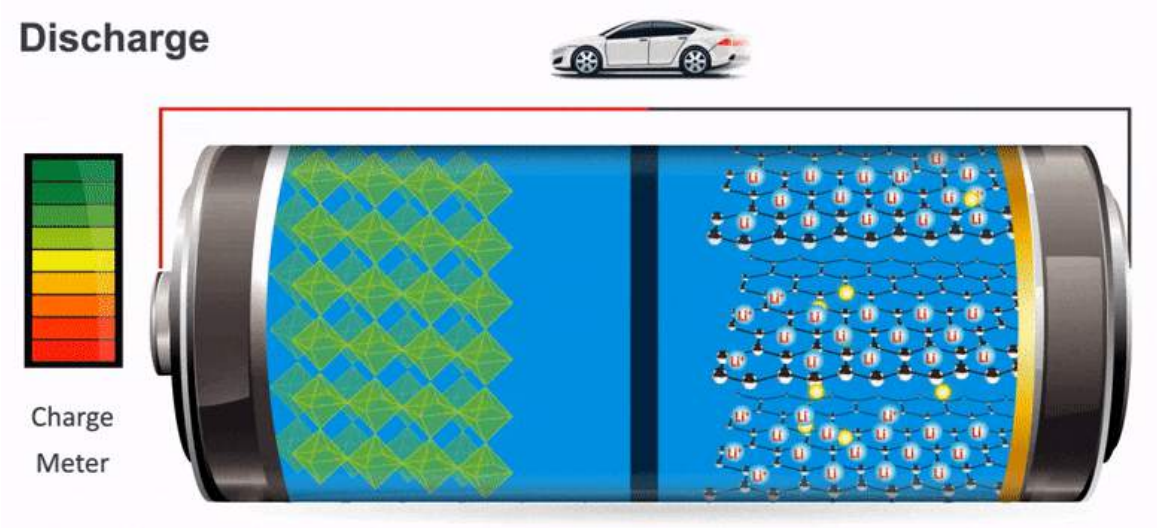


图. 锂金属负极材料缺陷的扫描电镜图像， Adv. Sci. 202105574

然而，电镜所拍摄的是“照片”，而可靠的预测电池性能所需要的是“影片”。电池和桌子不一样，不是一个静态的物件，电池之所以能发挥功能，就是因为它在工作中能发生化学变化，因此，设计电池光依靠照片是不够的，而是要理解影片，以及影片背后的机理。比如，在 EV 驾驶时，电池处于放电模式，此时锂离子从正极“游泳”到负极并堆积在石墨层中；而充电时锂离子从负极游回正极。当我们要求电池以更快的速度充电时，我们实质上是在微观尺度迫使锂离子更加“匆忙”的嵌入正极材料。举个

不恰当的例子，当我们以低速充电时，锂离子是整齐有序的进入正极材料，而快充时，就如中午食堂开饭，小学生们一齐涌入，这会导致混乱和危险。在电池快充时，有部分锂离子会"撞"入奇怪的位置，"卡住出不来了"，在下次放电的时候也无法响应，这部分锂离子实质上失去了活性，而当越来越多的锂离子失去活性时，电池的容量就会明显的受损。这就是为什么 EV 厂商都不建议日常高频率高强度的使用快充超充。

在最理想的情况下，材料工程师希望能在材料设计时就能对电池进行建模仿真，观察其在充放电过程中的实际变化，考察其性能、安全性等多目标方向上的表现，对已有设计进行改善和迭代。这种先进的研发范式，就是理性设计（de novo design）。



电池充放电示意图，USDoE

表：主要电池性能指标和对应的科学问题

电池性能指标	部分核心问题列举
高能量密度	现有正极材料的实际克容量距离其理论克容量还有很大差距； 硅材料及锂金属负极产业化面临的挑战； 新化学体系在实际电芯中的性能验证。
高安全性	电池材料及其界面的热安全风险和电芯内部物理化学短路的电安全风险的管控； 不同化学体系的单体电芯充放电过程的热安全、电安全、结构安全的管理策略； 面向真实使用场景的电池器件系统级的市场安全的管理，以电动汽车为例，汽车碰撞、汽车充电、汽车进水、汽车结露、电池包设计、电芯制造等全场景、全周期、全方位、全系统安全。
长循环寿命	物理化学力学等过程导致的电池材料老化及其界面恶化（正极材料的溶解、电极材料的相变化、电解液分解、界面膜的形成）、极片的结构破坏和功能失效（例如集流体本身的腐蚀、集流体和活性材料的接触不良、反复膨胀收缩导致的极片开裂）、电芯的结构破坏和功能失效（例如电芯内部短路、电解液大量消耗、产气大量积累等）
高倍率	真实电芯的充放电微观动力学过程中的限速步骤的识别和优化； 结构安全、热安全和电安全等外部约束下的充电策略优化。

Source:
[1]《中国物理》
[2] 储能科学与技术

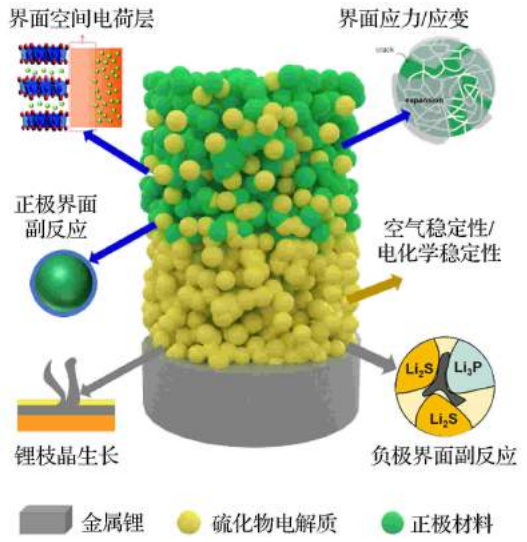
表 6：主要电池材料体系的理论优势、技术难点和 AI4S 的实践示例

	材料体系	优缺点	技术难点/科学问题	AI4S 应用
负极	石墨	成本低、结构稳定，克容量低	嵌锂的微观和介观机制；天然石墨存储性能差；人造石墨高耗能	反应路径/扩散机制/相变机制/枝晶生长机制探索，揭示反应机理
	纯硅	克容量高、嵌锂电位较低、脱嵌锂过程体积变化大	低首效，循环过程中体积反复膨胀收缩导致的颗粒粉化；相应电解液质的开发	
	氧化亚硅	克容量较高、嵌锂电位较低、脱嵌锂过程体积变化较大	低首效	
	硅碳复合材料	较高的克容量和中等的体积膨胀	硅碳混合比例及颗粒形貌对体积膨胀的影响；提高硅含量面临挑战；硅材料和碳材料脱嵌锂过程的交互关系	
	金属锂	克容量高、嵌锂电位低	锂枝晶生长的微观机制和调控策略；锂枝晶生长和 SEI 膜生长的竞争协同机制；相应电解质的开发	
正极	三元材料	理论克容量较高、结构稳定、平均电压较高、电子电导较差、压实密度较高	锂镍混排的综合影响、各种过渡金属元素的作用和组成调控、颗粒内和颗粒间的裂纹生长机制	搜索新的正极材料体系和结构，为新一代正极材料的快速发现提供理论和计算支撑 解析材料失效过程的具体微观机制，如相变、重构和反应是如何发生的，明确各种元素在这些过程的角色
	磷酸铁锂	理论克容量低、结构稳定、热稳定性好、平均电压低、电子电导差、压实密度低	充放电过程相分离的微观机制和调控策略；碳包覆和纳米化的作用机制和替代方案。	
	钴酸锂	理论克容量较高、结构稳定、电子电导优异、平均电压高、压实密度高、成本高	高电压下的过渡金属溶出、释氧、表面相变和 CEI 生长、脱锂末端 O3/O1 滑移相变的微观机制和抑制策略	

	尖晶石锰酸锂	理论克容量低、结构稳定、成本低	锰溶出问题； Jahn-Teller 效应的调控	
	富锂锰基材料	理论克容量高、结构不稳定、平均电压高、成本低	释氧导致的结构不稳定、循环过程较快的电压降和较高的容量损失、阴离子氧化还原的微观机制	
	硫正极	能量密度高	硫在转化成 Li_2S_x 过程中会产生多硫化物，溶解在电解液中	
电解质	液态	成本低，电化学稳定	高电压下电解液的氧化分解； 和锂金属体系、硅负极体系的匹配性	解析材料的电子结构和几何结构，确定原子成键信息和排布情况，明确扩散机制
	高分子/聚合物	容易加工、可制备大容量电芯	锂离子传输速率和途径以及配位环境的研究	
	固态陶瓷	耐高电压、安全性好	不同掺杂元素对陶瓷材料的结构变化，锂离子传输速率和途径的影响	
	固态硫化物	高离子电导率、耐高压	低成本下空气稳定的保持	
电芯	方形	技术难度相对较低，可定制化程度高	保证散热，维护电池的安全性	电芯尺度设计
	软包	能量密度高	及时的热扩散保障电池的安全性	
	刀片	散热性、安全性更高	刀片电池由磷酸铁锂材料构成，低温性能差；热失控过程的避免	
	圆柱	技术最成熟	能量密度的提升，对 BMS 的要求	
Pack	串并联组合方式	产线自动化成本低	结构设计、热设计、轻量化对电池最大能量效益发挥以及安全性的影响	自动化生产
BMS	电池生命周期监测	实现电池的实时监控	模型优化从而对电池实施良好的监控	实时生命周期监控

AI4S 实践（15）：AI4S 帮助中科院物理所、字节跳动等开发新型硫化物固态电解质

近年来,随着对电池安全性要求的提高,全固态电池成为研究热点。与液态电解质相比,固态电解质可有效改善安全性。其中,硫化物固态电解质因具有超高离子导电率和易冷压制备的优点,被认为是实现全固态锂电池的最有希望材料之一。但目前硫化物固态电解质在界面稳定性和空气稳定性方面仍存在一定问题。因此设计新型的硫化物固态电解质以获得更优异的综合性能,是目前的研究热点和发展方向。



基于硫化物电解质的固态锂电池界面面临的主要挑战 [2]

2023 年 7 月，《自然•通讯》报道了中科院物理所、字节跳动等机构关于硫化物固态电解质的研究[1]。该论文的主要技术思路和贡献如下:

通过调控 Si 在 Li6AsS5I 中的掺杂量,优化了电解质的离子传导性,获得了室温下 10.4 mS/cm 的高导电率。这是利用 Si 替换 As 引入额外的 Li+并激活新的迁移通道所致。与磷基硫化物电解质相比,Li6+xSixAs1-xS5I 系列电解质在湿空气中稳定性更好,结构和性能不易降解。这是 As 与 S 键合更紧密所致。

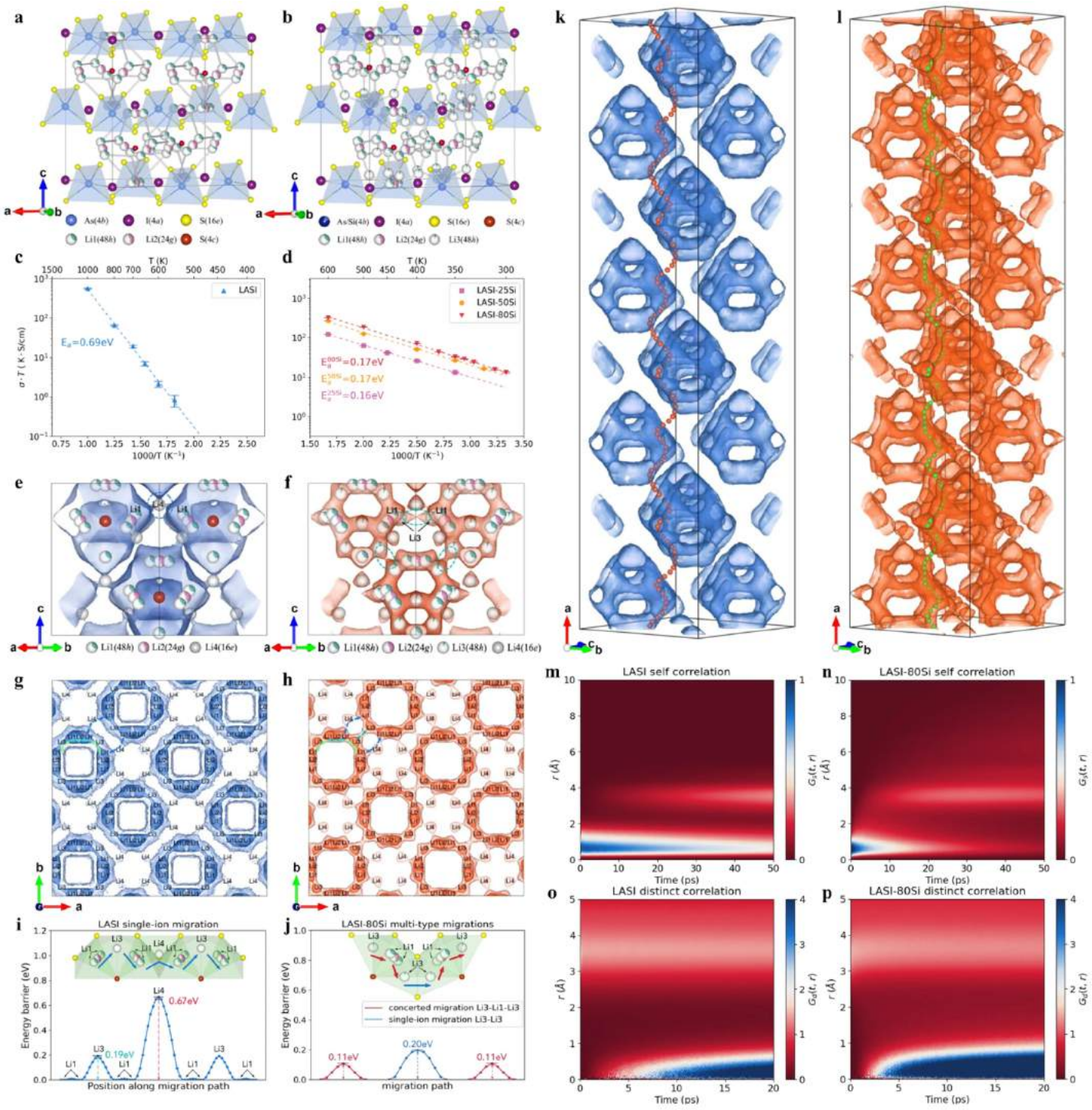
将该电解质与 Li-In 负极和 TiS2 正极配合,制备出全固态电池。获得了高达 6.25 万次的优异循环稳定性。电解质中的 Li2S 和 LiI 功能相起助催作用,提供额外的 Li+补偿,提高容量、循环性能和倍率能力。获得了 9.26 mAh/cm2 的高面容量和 24.45 mA/cm2 的大电流密度。这显示该电解质用于全固态电池有巨大应用潜力。该电解质兼具多功能,如高离子传导性、界面钝化效应、Li+补偿等,是一种非常有前景的新型硫化物固体电解质。

该工作中采用了基于深度学习的分子动力学方法 deepmd 来模拟电解质的结构优化和离子迁移。deepmd 能比传统的量子力学方法扩展模拟系统的大小和时间尺度,同时保持计算准确度。相比传统的首原理分子动力学模拟,deepmd 方法使得模拟时间从几十皮秒级别扩展到几十纳秒量级,模拟系统中的原子数也从几百个提高到几千个。这更好地揭示了 Li6+xSixAs1-xS5I 体系中 Li+的迁移机制,获得了更准确可靠的传导率数据。关于这部分的示意图，请见下页。

这项研究为设计新型功能电解质材料提供了思路和指导,有利于发展高安全、高容量的全固态锂电池。

Source: [1] Lu, P., Xia, Y., Sun, G. et al. Realizing long-cycling all-solid-state Li-In||TiS2 batteries using Li6+xMxAs1-xS5I (M=Si, Sn) sulfide solid electrolytes. Nat Commun 14, 4077 (2023). https://doi.org/10.1038/s41467-023-39686-w

[2] Zhang Qiao-Bao, Gong Zheng-Liang, Yang Yong. Advance in interface and characterizations of sulfide solid electrolyte materials. Acta Phys. Sin., 2020, 69(22): 228803. doi: 10.7498/aps.69.20201581



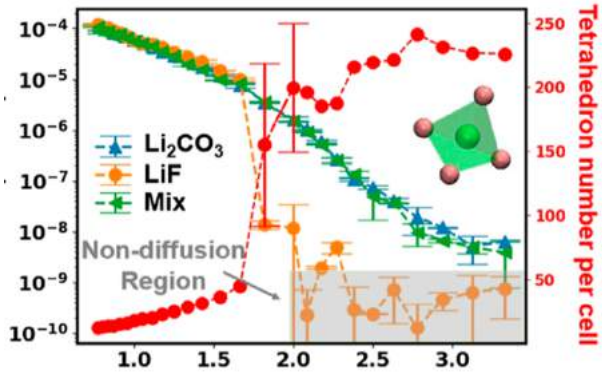
其中 a、b 为经过第一性原理 DFT 和分子动力学模拟优化后的 LASI 和 LASI-80Si 的晶体结构图,模拟温度为 400K。c、d 分别为 LASI 和 LASI-25Si、LASI-50Si 以及 LASI-80Si 离子导电率随温度的阿伦尼乌斯关系图。e、f 展示了在单元胞中选取的三个笼子结构内,LASI 在 1000K 和 LASI-80Si 在 400K 下的三维锂离子概率密度分布,其中不同笼子结构之间的特殊迁移通道用虚线圆圈标注。g、h 沿[001]方向观察 LASI 在 1000K 和 LASI-80Si 在 400K 的二维锂离子概率密度分布,其中双跳迁移(Li1(48h)-Li2(24g)-Li1(48h))、同笼内跳迁(Li1(48h)-Li3(48h)-Li1(48h))、不同笼子间跳迁(Li1(48h)-Li4(16e)-Li1(48h))以及新的不同笼子间跳迁(Li3(48h)-Li3(48h))通道用不同颜色的双向箭头表示。i、j 分别给出了 LASI 的单离子跳迁模式和 LASI-80Si 的多类型跳迁模式下的迁移能垒。插入图示出了对应锂离子迁移通道中多个锂位点的连接。k、l 展示了相同等值面下,LASI 和 LASI-80Si 在 4×1×1 超胞中的三维锂离子概率密度分布,其中锂离子沿 a 轴方向的轨迹用于直观显示。m-p 为 LASI 和 LASI-80Si 在 600K 下分子动力学模拟过程中自相关和区分相关凡夫函数。

AI4S 实践（16）：AI4S 先进科研平台助力北京大学许审镇组在顶刊《JACS》发表复杂固态电解质界面 SEI 机理研究

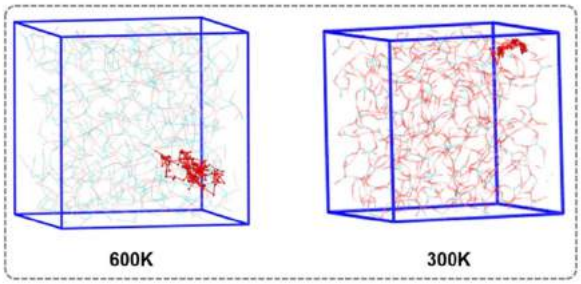
SEI(固体电解质界面)的形成和稳定性在电池研究中有非常重要的地位。在锂电池的第一次充放电过程中，电极与电解液会发生反应形成所谓的固态电解质膜（SEI）。合理的 SEI 可以抑制后续电解质分解,保护负极,从而确保电池循环稳定性。但是 SEI 的过度生长会消耗电解质和锂离子,降低 Coulombic 效率。不同电解质和负极对 SEI 的形成机理和性能影响巨大。开发稳定的 SEI 一直是实现高性能电池的关键目标之一。

研究 SEI 的关系可以提供设计优化电池的重要指导。一个重要的争议是"引入无定型特征的 SEI 膜是否有利于提高锂离子电导"，如果是，那么设计具有无定型特征的人造 SEI 膜对于提高锂电池的倍率性能就可以起到至关重要的作用。因此，研究无定型 SEI 组分中锂离子的扩散机理尤为重要。

2023 年，国际顶刊 JACS 发表北京大学许审镇组工作[1]。研究者借助深度势能生成器 DP-GEN 训练了覆盖从室温到高温（300-1300 K）的无定型 Li2CO3、LiF 以及二者 1:1 混合的 DP 势函数，通过 DeePMD 模拟研究了锂离子在其中的扩散机理，重点解释了局域环境对锂离子扩散行为的影响。



在混合体系中,锂离子的扩散率接近 Li2CO3 的扩散率。在低温(500K 以下)将 Li2CO3 引入无定形 LiF 中,可以有效抑制 Li-F 正四面体结构的形成,从而提高混合体系中的锂离子导电率



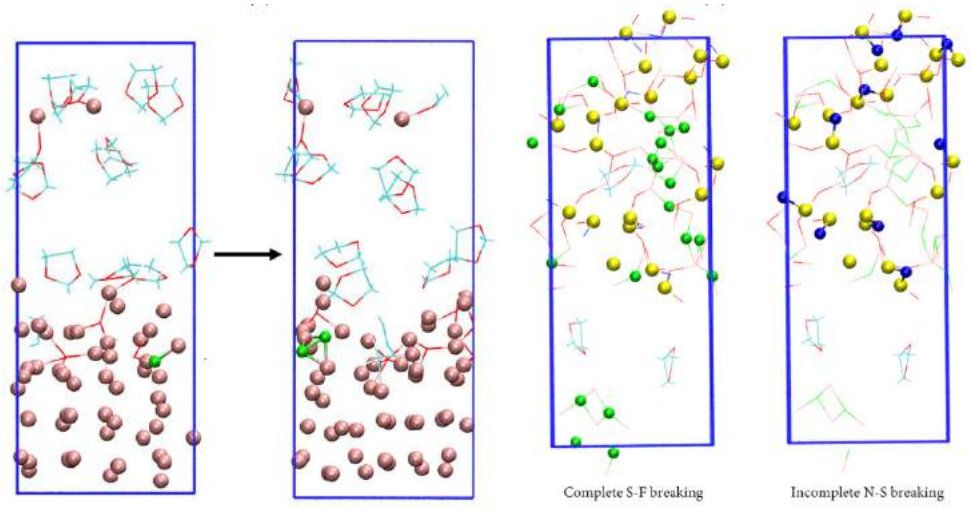
Li 离子在 SEI 成分 Li2O3 中的扩散

研究表明，LiF 中 Li 离子扩散的 non-Arrhenius 行为是由于局域环境明显改变所导致的 non-diffusion 特征决定的。这说明对于 LiF 而言，设计具有无定型特征的 SEI 膜对 Li 离子的扩散没有帮助，因为 Li-F 规整四面体的自发形成会严重阻碍 Li 离子的输运。相反，在 Li2CO3 中，发现不论是在高温还是低温下，均保持无定型的特征，因此相似的局域环境决定了 Li2CO3 中 Li 离子扩散基本符合 Arrhenius 行为。进一步对 1:1 混合的 Li2CO3:LiF 模拟发现，Li 离子的扩散系数与 Li2CO3 接近。从局域结构来看，Li2CO3 的引入明显抑制了 Li-F 四面体的形成，因而促进了 Li 离子的扩散。从 SEI 膜设计层面来讲，尽管实验上已经对于 LiF 的作用有了大量研究，但是本工作表明设计富含 LiF 的 SEI 膜也可能会导致一些潜在的问题，因为 LiF 中 Li-F 规整四面体的存在导致了 Li 离子扩散能力的下降，因此要避免大块体相 LiF 的单独析出。

本工作揭示了局域环境对锂离子扩散机理的影响，并对实验上 SEI 膜的设计提出了理性指导。研究在深势科技推出的先进科研平台 Bohrium 上完成。Bohrium 提供了 AI4S"四梁" 中基础算法算力的深度整合，为科研工作者提供了生产力新工具。

Source: J. Am. Chem. Soc. 2023, 145, 2, 1327–1333

AI4S 实践（17）：三星研究院 (SRC-B) 使用 AI4S 实现 SEI 形成过程超长仿真



(左) LiF 发生离子聚集形成 Li2F2。(右) 在 10 M 界面的系统中,S-F 键发生完全断裂,而 N-S 键发生不完全断裂。为了清楚起见,我们在 VMD 中只使用 CPK 绘图样式显示相关原子。

电池内部电极和电解质之间的界面结构,以及如何设计这种界面,是现代电池化学研究的核心。固体电解质界面(SEI)的组成和结构会影响电池的安全性、循环寿命、能量密度和成本等多个关键指标。通过调整电解质的浓度、组成结构等方面来优化固体电解质界面,是开发先进电池的重要策略之一。例如,研究表明电解质溶液的浓度会显著影响锂金属电池中固体电解质界面的组成和结构。通过调节锂离子与氧/氟的连接网络,以及改变取电子基团的数量和位置,研究者能够合理设计电解质溶液中的溶剂分子,使电池在溶剂化和库仑效率等多个方面达到优异的性能。简而言之,优化电极和电解质之间的界面结构,是设计高性能电池的关键所在。

三星是世界最大的电池研发与生产厂商之一。2023 年，三星研究院 SRC-B 预发表文章，提出了一种称为 SIS-MLPMD 的方法，使用深度势能、增强采样等方法，实现锂金属电池的超长模拟。DeepModeling 社区对研究中深度势能的使用提供了讨论与帮助。

文章的技术思路是：1）利用不稳定的物理信息(温度)作为采样指示,无需事先了解反应网络或训练多个 MLP 进行不确定性评估。2）通过比较不同 MLP 和性质之间的基础不稳定一致性,证明仅使用温度不稳定信息本身就足以指导采样。3）在 1M 和 10M Li|LiFSI/DOL 界面系统中实现了超过 1ns 的稳定 MLPMD 模拟。

固体电解质界面(SEI)的形成对电池安全、循环寿命、能量密度等方面具有关键影响,需要较长时间尺度的动力学模拟才能足够理解。目前公认的短时间量子力学模拟难以捕捉界面反应的全部机理,需要一种低缩放的方法将模拟时间延长到纳秒量级。机器学习势能使得延长模拟时间成为可能,但需要有效的采样策略获得足够的训练数据。

本研究将基于 AIMD 的 MLPMD 模拟时间延长到 1-10 ns,这对研究 SEI 形成机理具有重要意义。超长时间尺度的模拟可以揭示一些短时间模拟难以观察到的重要结构和反应，如 LiF 的离子聚集，这对未来电池的设计与仿真提供了技术思路。

Source: Data Efficient and Stability Indicated Sampling for Developing Reactive Machine Learning Potential to Achieve Ultra-long Simulation in Lithium Metal Batteries , 10.26434/chemrxiv-2023-4x3gr

4.3.3 "Beyond Lithium" -- AI4S 赋能钠电池的基础理论建设

长期来看，面临着日益增长的电池需求，锂离子电池的进一步发展将受锂资源短缺限制，而资源储备丰富、具备成本和安全性优势的钠离子电池路线崭露头角，未来或将成为电池行业的重要备选路线。[1]

钠离子电池的应用前景非常广阔，其中在储能场景中具有巨大的应用潜力。然而现有的钠电池材料中，在离子电导率方面的表现还不尽人意，这也是限制钠电池发展的主要因素之一。钠电池材料的优化以及电导率的提高需要对钠离子的扩散过程有更清楚的基础微观认识。

图. 钠电池示意图 [2]

Source:

[1] 中国锂电行业发展 德勤观察 2.0"电池风云

[2] Chem. Soc. Rev., 2017,46, 3529-3614

AI4S 实践（18）：《Energy Environ. Sci》等期刊收录 AI4S 钠电池固态电解质研究，发现提高核心参数电导率的新理论思路

杜克大学 Mayanak K. Gupta 团队近日发表在《Energy & Environmental Science》的工作利用 AI4S 方法考虑 Na 空位和温度引起晶格振动的声子行为对 Na_3PS_4 固态电解质中 Na 离子扩散作用的影响。

Na_3PS_4 是最常见的硫化物固态电解质，具有两种晶体结构，分别是四方晶相和立方晶相，前者为低温稳定相，后者为高温稳定相。而四方相 Na_3PS_4 在 50 ℃ 时，离子电导率不能满足实际应用的要求。[1] 团队通过结合高分辨率的中子非弹性散射(inelastic neutron scattering, INS)实验和准弹性中子散射(quasi-elasticneutron scattering, QENS) 实验验证在结构相变附近的软晶体结构中非谐波低能模式能加速离子传导，这为从微观上揭示 Na 固态电解质中离子传导的动力学机制，进一步寻找和设计高电导率的 Na 固态电解质提供新的思路。[2]

北京大学蒋鸿课题组发表在《Inorganic Chemistry Frontiers》中的工作研究了富钠反钙钛矿 Na_3OBr 电解质中的离子迁移率。Na_3OBr 电解质具有较好的电化学性能，然而在室温下电导率低，团队基于第一性原理密度泛函理论(DFT)计算建立了 Na_3OBr 的深度势能模型，直接计算了 Na^+在不同温度下的扩散系数，确定了扩散系数与空位浓度的比例关系，并发现迁移势垒对空位浓度相对不敏感，这促进对反钙钛矿型材料结构和性质的理解，帮助进一步探究钠离子固态电解质的应用前景。[3]

除此之外，厦门大学化学化工学院的程俊教授和杨勇教授通过理论和实验方面的手段，利用深度势能方法进行动态核磁共振波谱（Nuclear Magnetic Resonance spectroscopy, NMR）的计算，降低碱

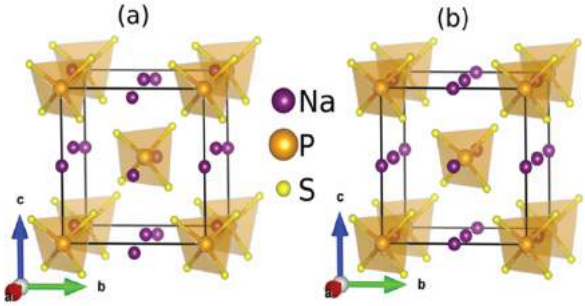


图. Na3PS4 的(a)低温四方相和(b)高温超离子立方相的晶体结构。

金属离子快速扩散过程中 NMR 谱峰的指认难度，该方法应用于顺磁性 P2 型钠离子电池正极材料并通过实验验证。相关工作发表在《Chemical Science》《Angew. Chem. Int. Ed.》中[4,5]。

利用 AI4S 范式研究钠电池的工作不在少数，篇幅所限就不一一列举了。

Source:

[1] 高永晟, 陈光海, 王欣然,等. 钠离子电池电解质安全性:改善策略与研究进展[J]. 储能科学与技术, 2020, 9(5):9.

[2] Gupta M K, Ding J, Osti N C, et al. Fast Na diffusion and anharmonic phonon dynamics in superionic Na 3 PS 4[J]. Energy & Environmental Science, 2021, 14(12): 6554-6563.

[3] Li H X, Zhou X Y, Wang Y C, et al. Theoretical study of Na+ transport in the solid-state electrolyte Na 3 OBr based on deep potential molecular dynamics[J]. Inorganic Chemistry Frontiers, 2021, 8(2): 425-432.

[4] M. Lin, X. Liu, Y. Xiang, F. Wang, Y. Liu, R. Fu, J. Cheng, Y. Yang, Unravelling the Fast Alkali-Ion Dynamics in Paramagnetic Battery Materials Combined with NMR and Deep-Potential Molecular Dynamics Simulation, Angew. Chemie Int. Ed. 60 (2021) 12547–12553.

[5] M. Lin, J. Xiong, M. Su, F. Wang, X. Liu, Y. Hou, R. Fu, Y. Yang, J. Cheng, A machine learning protocol for revealing ion transport mechanisms from dynamic NMR shifts in paramagnetic battery materials, Chem. Sci. (2022).

4.4 太阳能与 AI4S

太阳能是人类"取之不尽用之不竭"的清洁能源。在太阳能的有效利用中，最常见的是以光伏效应为基本原理的光电转换方式。太阳能电池在近几十年来快速发展，得益于它无污染、可再生的优点，相关研究也越来越受瞩目。

制作太阳能电主要是以半导体材料为基础，其工作原理是利用半导体材料吸收光能后进行光电子转移反应。根据所用的材料不同，太阳能电池可以分为：①硅太阳能电池；②以无机盐如砷化镓、铜铟镓硒等多元化合物为材料的电池；③有机太阳能电池；④钙钛矿太阳能电池。然而无论以哪种材料制作太阳能电池，都离不开以下必要条件：①材料的带隙不能太宽；②具有较高的光电转换效率；③材料本身对环境不造成污染；④材料便于工业化生产且性能稳定。[1]

针对"光电转换效率的提高"和"成本低"两大太阳能电池制备的核心因素，科研人员也在针对不同的材料体系面临的核心困难进行发力。受限于传统研发方式耗时长的，微观结构表征困难的缺点，学科数据的积累和物理化学机制的理解为 AI4S 方法加速太阳能电池研发创造更多机会。以下表格将针对四种常见的光伏材料体系进行分析。

光伏体系	理论优势	技术难点	AI4S 赋能
晶硅（Si）	量产技术成熟，光电转化效率高	光电效率已经到达天花板，硅的生产存在一定的环境污染问题	微观缺陷调控（本征缺陷、杂质缺陷、复合缺陷、晶界）优化太阳能电池的光电效率
砷化镓（GaAs）	在可见光区域具有良好的光吸收性能[2]	严重的光腐蚀问题破坏材料的结构	从微观角度解释材料的光腐蚀机理
铜铟镓硒（CIGS）	具有最高的光吸收系数；可制造为柔性电池	不同的铟/镓比例及柔性衬底对光电效率的影响[3]	不同掺杂元素的影响
钙钛矿材料	光电转化效率高	材料的结构不稳定从而导致使用寿命短	相转变过程/结构稳定性预测；器件设计

钙钛矿太阳能电池是目前发展最快的太阳能电池技术之一。从第一次利用钙钛矿进行太阳能发电的 2009 年开始到现在，钙钛矿的光电转化效率从 3.8%快速发展到 25%，直逼同期单晶硅电池的理论效率。除了具有优异的光电转化效率外，钙钛矿还展现出较低生产成本的潜力，其巨大的产业价值吸引了越来越多的科研团队以及企业投身其中，目前国内已有布局的企业超过 15 家，如宁德时代、协鑫纳米、纤纳光电、国家能源集团等。学术界的研究中，领域顶刊中与钙钛矿相关的工作同样不在少数。

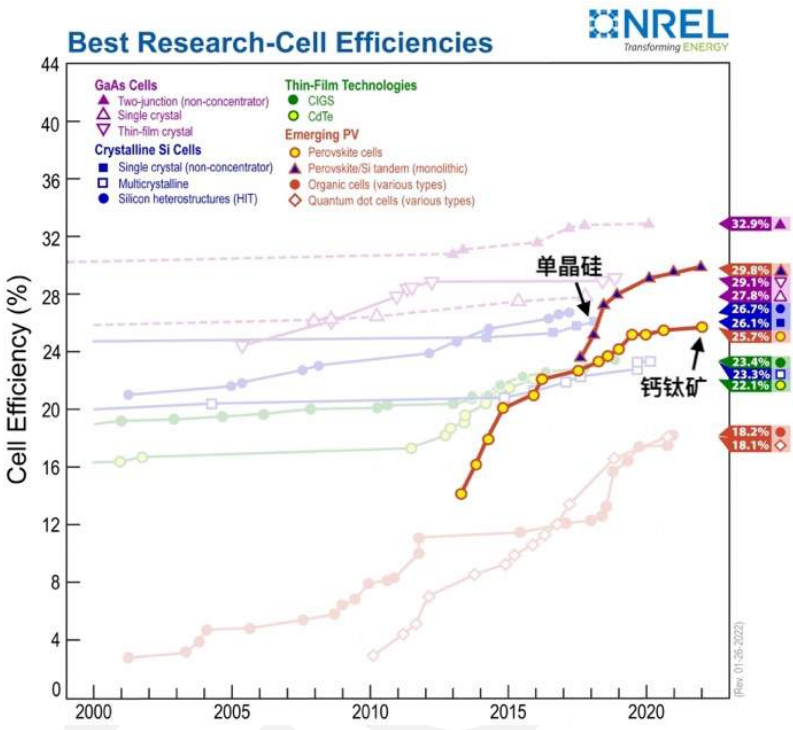


图. 各光伏技术路线发展历史 [5]

钙钛矿太阳能电池中的材料通常是指具有 ABX_3 结构的有机或无机钙钛矿，其中 A 位置为铯(Cs)、甲基铵(MA)或甲脒(FA)；B 位置为 Pb 或 Sn；X 位置为 Cl, Br 或 I。一般来说，钙钛矿太阳能电池的光伏性能归因于以下几个方面：吸收系数、带隙、载流子扩散长度、载流子传输能力等。尤其需要关注的是，有机-无机杂化钙钛矿材料是钙钛矿太阳能体系中的佼佼者：低成本制造，具有更高的介电常数，高载流子扩散长度和速度，出色的光吸收能力，表现出独特的光学和电学性能。[3]

即使钙钛矿太阳能电池的研究已经表现出相对较好的光电转化优势，但它依然面临着长期运行不稳定这一重要挑战。钙钛矿太阳能电池的不稳定性主要来源于两个方面：一个是材料本身的组分以及结构缺陷；另一个是外部服役环境，如光照时长、温度、湿度等，上述都会加速钙钛矿太阳能电池的光电转换效率衰减，降低使用寿命。目前已有研究的钙钛矿电池平均使用寿命大约在 1000 个小时，而晶硅太阳能电池的使用寿命超过 20 年。尽管钙钛矿太阳能电池在制造成本上远胜于晶硅电池，但其寿命短、稳定性差、天然不稳定，导致效率衰减过快，是产业化落地的最大掣肘。

如何在保证光电转化效率的前提下提高材料的结构稳定性，是现阶段钙钛矿材料研究的重点。通过合适的掺杂、包覆、以及理解材料内部光子到电子转化过程的机理等方式，帮助设计优化其结构。

为了提高钙钛矿太阳能电池的性能，全球各个研究课题组针对钙钛矿电池的不同层面展开深入的研究，包括带隙，缺陷演化，相变过程，界面相互作用以及外部环境如热处理、光照、湿度等等因素的影响。相应研究硕果累累，不少发布在学术顶刊上，如近期南京大学谭海仁团队发表在《Nature》上利用实验表征和 DFT 计算，研究钝化分子表面吸附对钙钛矿太阳能电池转换效率的影响，通过钝化窄带隙钙钛矿晶粒表面缺陷来提升薄膜的载流子

扩散长度从而提高电池转换效率的工作；美国加州大学洛杉矶分校 Yang Yang 团队结合实验和 DFT 计算手段深入研究表面钝化对钙钛矿太阳能电池的影响，指导最优的表面钝化策略，相应成果也发布在《Nature》上……相应的工作还有非常的多。相关的钙钛矿工作除了有实验研究之外，近年来计算模拟手段在这当中也越来越发挥作用。从微观角度帮助深入理解光电转化过程中的物理化学机制，为实验提供深刻的理论指导。然而随着科学家以及相应的科学问题研究对高精度、高效率的计算模拟方法要求越来越高，计算手段也面临着挑战。用最典型的 CsPbI3 钙钛矿举例。其较低的生产成本，较高的光电转换效率，较好的工业加工性使得过去几年成为钙钛矿光伏材料的新星。然而其稳定性依旧是一大挑战。

2022 年，华北电力大学 Zhengyang Gao 团队通过构建 AI4S 模型，研究离子缺陷对 CsPbI3 钙钛矿材料的结构稳定性的影响。过去已有不少实验和理论研究都表明离子缺陷显著抑制了钙钛矿材料的结构稳定性，影响材料在太阳能电池性能如光电转换效率，发光寿命等的发挥，然而由于传统方法的局限性，在控制缺陷的浓度和类型对性能分析上仍然缺少一定的说服力。而传统的 DFT 计算和分子动力学(MD)模拟虽然可以明确地控制缺陷的浓度和类型从而在原子水平上研究这一问题，然而高昂的成本限制了高精度 DFT 对原子和时间尺度的评估，经验参数的不足约束了经典 MD 模拟对钙钛矿材料模拟的准确性。随着近些年来 AI 的快速发展，AI4S 的新范式在计算精度和速度上做了很好的融合和平衡，为解决上述挑战提供了全新的机会。Zhengyang Gao 团队基于 DFT 计算数据建立了描述 CsPbI3 体系的 AI4S 模型，该模型的计算结果和 AIMD 的结果相比非常接近（右图），同时计算效率较 AIMD 相比加速了上万倍，比经验方法的计算速度更快。借助该模型他们对五万个原子的体系在不同温度下进行 1 纳秒时间尺度的分子动力学模拟，获得了 CsPbI3 材料结构演化过程，进一步分析表面和体内离子缺陷对 CsPbI3 钙钛矿材料结构稳定性影响，为实验设计和制备高稳定钙钛矿材料奠定了基础。[4]

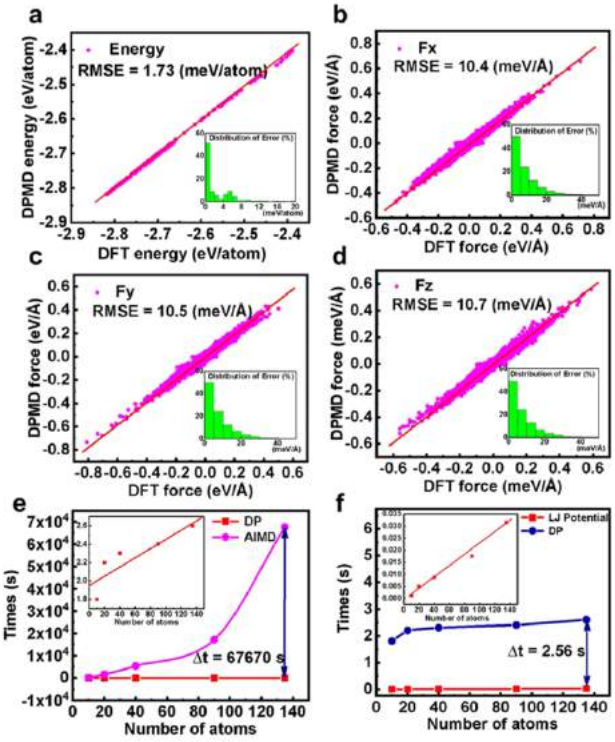


图. AI4S 模型精度接近 AIMD 模型，同时计算成本远低于 AIMD，与经验模型相当

参考资料：

[1] 梁宗存, 沈辉, 李戬洪. 太阳能电池研究进展[J]. 能源工程, 2000(4):4.

[2] 材料导报 35.5(2021):6.

[3] Kumar N S, Naidu K C B. A review on perovskite solar cells (PSCs), materials and applications[J]. Journal of Materiomics, 2021, 7(5): 940-956.

[4] Yang W, Li J, Chen X, et al. Exploring the Effects of Ionic Defects on the Stability of CsPbI3 with a Deep Learning Potential[J]. ChemPhysChem, 2022, 23(7): e202100841.

[5] NERL www.nrel.gov/pv/cell-efficiency.html

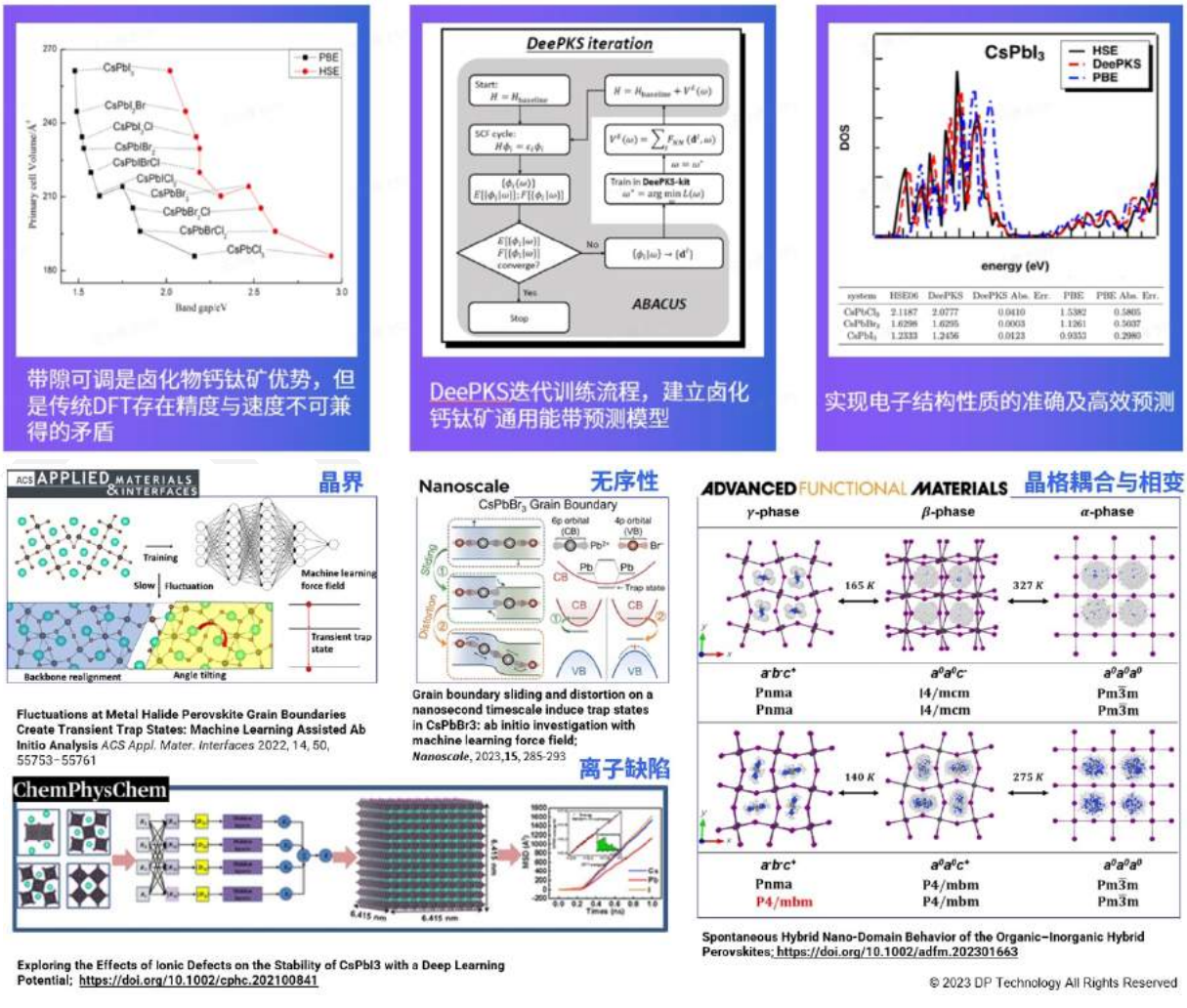
AI4S 实践（19）：DeePKS 基于钙钛矿带隙预测的高通量筛选技术路线

钙钛矿系列光伏材料中，卤化物钙钛矿因高效目带隙可调，收到业界关注，而精确预测其电子结构性质对设计高效稳定的光伏器件非常重要。在常用密度泛函中 HSE06 准确但慢；PBE 快但低估带隙。

2023 年，北京科学智能研究院团队在 arXiv 发布文章[1]，提出了一个通用的 DeePKS 模型,可以高效准确地计算卤化钙钛矿类材料的电子结构性质。其技术思路如下：使用 HSE06 混合泛函作为高精度目标方法生成标签,使用 PBE 泛函作为低成本基准方法；构建 DeePKS 模型,使用神经网络拟合 HSE06 和 PBE 的能量差,以 PBE 的效率获得接近 HSE06 的精度；通过对 7 种卤化钙钛矿的 460 个配置的迭代训练,获得一个通用的 DeePKS 模型，模型不仅可以准确预测能量和力,还可以复现 HSE06 的带隙和态密度，显著提升 PBE 的精度。即使自旋-轨道耦合效应没有考虑在训练中，DeePKS + SOC 也可以给出与 HSE06+SOC 一致的结果。

基于高效精确的计算能力，研究者可以进一步根据带隙来实现对于配方的高通量筛选。相关团队同时对钙钛矿建立了深度势能的分子动力学模型，可进一步研究其稳定性、相变、分解等问题，在多个尺度建立了钙钛矿的高精度数学模型，为相关材料器件体系的理性设计提供了技术路线的参考。

Source: [1] arXiv:2306.14486



4.5 核能与 AI4S

在全球能源紧缺的背景下，核能作为一种清洁、可持续、能大规模应用的能源，在世界能源体系中扮演着越来越关键的角色。目前核能的主要使用方式是发电，据英国石油公司 BP 统计，2020 年全球核能发电量占全球发电总量的 10%。常见的核能释放过程主要有核裂变和核聚变两种形式。核裂变是利用中子去撞击一个质量较重的原子核，使其分裂为两个或两个以上的中等质量的原子核。裂变反应过程是持续进行的，是非常典型的链式反应。跟核裂变的原理不同，核聚变是帮助 2 个原子克服电磁力和强核力，聚合形成新的原子并释放能量。聚变反应过程中释放的能量非常大，其能量密度远超于现有能源的能量密度。[8]

当前全球核能正在向安全性更高、经济性更好、多用途、更加灵活、能够满足更广泛用户需求发展。[2]这当中除了研究新能核能系统和先进核能技术外，对核反应的基本原理研究也必不可少。然而核反应过程涉及多个学科，当中的物理化学现象非常复杂，除此之外核反应网络还包含了大量的基本反应，实验上难以对实际工况进行实践从而完成详细反应路径的探索，另外理论上又缺乏相应的物理模型进行指导，这使得核反应的基础研究一直停留在起步阶段。而 AI4S 强大的处理复杂问题的能力为核反应的基础研究提供新的路径。在过去的一段时间里，国内外科研人员针对核反应过程的安全性、可持续性开展了大量的研发工作。

相关科研实践中，北京大学薛建明课题组利用描述核反应堆中抗辐照材料二硫化钼(MoS2)的 AI4S 模型，研究其在辐照过程中复杂微观物理机制，包括碰撞、局部熔融、重结晶、缺陷的聚集和迁移等多种复杂过程，帮助筛选和设计高性能的抗辐照损伤材料。[3]北京大学陈默涵课题组用分子模拟技术研究核聚变过程中处于高温高压状态下温稠密等离子体的性质[4]，打破大规模模拟困难的局面，推动实际应用规模下核反应过程的研究和发展。另外，华东理工大学路贵民课题组利用 AI4S 方法研究熔融氯化镧(LaCl3)，理解利用液态金属和熔盐进行化学高温处理辐射核燃料过程，从而解释该过程当中与理想状态偏离的部分，推动可持续的核燃料循环。[5]除上述提到的几项工作之外，核反应相关的研究已经有不少积累，这里就不详细展开。

反应类型	技术难点/科学问题	AI4S 展望 [6,7]
核裂变	核结构和材料在高温、辐照、腐蚀环境下的力学性能及其老化过程对核电装置和构件的性能和服役寿命的影响[1]；辐射核燃料的处理等	跨尺度建模对沸腾、流动、核反应等复杂物理化学现象进行解释，设计新型核能系统
核聚变	点火条件；托卡马克装置中的磁场控制；耐中子和电磁辐射的第一壁材料选择；温稠密物质的反应过程研究等	有效融合实验与模拟数据的 AI4S 方法，最大限度地提高对聚变等离子体和燃烧等离子体状态的预测性理解，实现对托卡马克装置的实时反馈，弥补聚变等离子体约束和稳定性预测方面的差距，有效指导实验。

Source:

1. 中国科学: 物理学 力学 天文学 49(11), 114601 (2019)

2. 《大力推进先进核能技术研发应用，创新引领核电积极安全有序发展》, 规划编制核能专业组

3. Wang H, Guo X, Zhang L, et al. Deep learning inter-atomic potential model for accurate irradiation damage simulations[J]. Applied Physics Letters, 2019, 114(24): 244101.

4. Zhang Y, Gao C, Liu Q, et al. Warm dense matter simulation via electron temperature dependent deep potential molecular dynamics[J]. Physics of Plasmas, 2020, 27(12): 122704.

5. Feng T, Zhao J, Liang W, et al. Molecular dynamics simulations of lanthanum chloride by deep learning potential[J]. Computational Materials Science, 2022, 210: 111014.

6. Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science, 2020

7. 红义 杨. An Overview of the Application of Neural Network Algorithm in the Nuclear Field of China[J]. Nuclear Science and Technology, 2020, 08(1):19-34.

8. Wikipedia, https://en.wikipedia.org/wiki/Nuclear_power

AI4S 实践（20）：DeepMind 更新其核聚变物理仿真能力，等离子体控制精度高达 65%

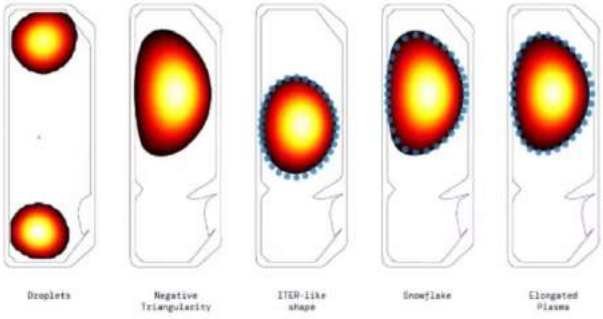
核聚变，在过程中质量耗损而释放出巨大的能量，是目前人类已知能源理论中的圣杯。数十年来，科学家和工程师们不断研究托卡马克装置，约束等离子体，从而达成可控核聚变的目的。如何有效控制等离子体，是通往核聚变的关键。

2022 年 DeepMind 与士洛桑联邦理工学院等离子体中心的物理学家 Magnetic control of tokamak plasmas through deep reinforcement learning 论文登上《Nature》。研究中， DeepMind 训练 AI 学习精准控制托卡马克内包含等离子体的磁场。[1]

为托卡马克装置开发控制软件绝非易事。整个托克马克装置涉及多个物理尺度，且其最底层的物理理论并不完全完备。传统的经验参数（例如需要读取哪些传感器输入、如何响应某些特定变化之类）无法达到实践所需的精度。因此，核心挑战是 “如何建立一个严谨的测量与建模迭代过程，由工程师对控制流程做出持续调整，以近实时方式确保反应堆始终拥有理想的释能水平”。

DeepMind 团队从瑞士等离子体中心的托卡马克模拟器硬件起步；研究人员在 AI 编程中设置了物理限制，要求模型直接丢弃那些无法在模拟器中产生准确结果的配置，通过 AI+Science 融合的方式，提高 AI 的准确率。经过反复迭代，DeepMind 训练出一款深度强化学习程序，足以控制模拟器完成种种复杂的等离子体配置操作。

研究团队用一个大型神经网络每秒对 90 种等离子体的形状和位置完成一万次训练，从而不断对磁场变化如何塑造等离子体进行长程预测，并相应地调整 19 块磁铁的电压。然后用这个神经网络来训练一个小的系统，学习执行第一个网络所推荐的最佳



DeepMind AI 生成的等离子体形态

决策。“这是迄今为止强化学习在现实世界中最具挑战性的应用之一，”DeepMind 研究科学家 Martin Riedmiller 表示 [2, 3]

2023 年，DeepMind 对其算法进行进一步更新。通过精心设计代理的奖励函数,等离子体形状控制精度提高了达 65%。提供积分误差信号也显著减少了等离子体电流控制中的稳态偏差。团队引入 “片段化 episode”可以通过将长 episode 拆分为多个短片段并行学习来加速训练。这比在完整 episode 上学习结果更平滑,训练时间缩短 3 倍。传递学习也被证明可用于快速适应新操作条件。

从结果而言，模拟性能的提高转化为更好的实际结果。TCV 托卡马克的实验验证了 X 点误差减少 59.7%,并且采用片段化 episode 平稳运行。 这项研究使 RL 在实际托卡马克控制中更加可行。它也指出了模拟到实际的传递和训练加速方面的进一步工作。[4]

Source:
[1] Degrave, J., Felici, F., Buchli, J. et al. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature 602, 414–419 (2022).
[2] 澎湃新闻, https://j.021east.com/p/1645099491034568
[3] InfoQ, https://www.infoq.cn/article/lqgwkaxwfbirzo1xehds
[4] https://arxiv.org/pdf/2307.11546.pdf

4.6 氢能源与 AI4S

氢（H）是宇宙中最常见的元素，氢气也是重要的工业原料和能源载体。氢能具有可再生、清洁无污染、储运方便、利用率高、热值高（同质量化石燃料的 3 ~ 4 倍）等特点。氢能的产业链可以分为：制备、存储输运及应用（氢燃料电池）。在各环节中，氢能都面临着一系列基础科学挑战：电子输运、转移等电子层面问题，分子扩散、吸附、氢键等分子层面动力学问题，相变等分子层面热力学问题等。AI4S 借助微尺度模拟和跨尺度建模等方法，可实现微观和介观物理的模拟和分析，完善氢能源基础研究，加速氢能源产业落地。

二维碳材料被认为是储氢的优秀技术路线。2021 年，《Journal of Physics Chemistry C》收录了 University of Pittsburg 团队对石墨烷（Graphane）的研究，研究人员对饱和储氢状态的石墨烷进行了基于深度势能的 AI4S 第一性原理建模，在声子态密度（DoS），热力学性质，等性质研究都达到了 DFT 精度，远超传统的经验多体势能建模方法，并观察到了良好的泛化性。

过程	技术路线	核心挑战	AI4S 展望
制备	化石能源；工业副产气制氢	制备过程产生污染; 产品需要提纯	通过模拟二氧化碳的产生及释放路径，找到捕集二氧化碳的方法或筛选合适的物质
	电解水制氢	贵金属催化剂成本高，同时催化活性、稳定性有待提升	微观模拟催化反应实现对催化流程的建模，从而构建高比表面积、高稳定性、高活性催化剂结构。
	生物质制氢; 太阳能光催化	仍处于实验室阶段	理解微观反应过程，发现反应机理
存储及输运	气态储运	高压气态储氢	需要加压设备，成本高
	液态储运	液氢存储密度低，输运设备要求高; 脱氢性能差	对低温下具有良好脱氢性能的储氢材料进行筛选
	固体储运	储氢，储氢性能有待提高	设计优化固态储氢材料，提高材料储氢性能
燃料电池	催化剂	催化剂依赖贵金属 Pt，成本高昂; 催化剂稳定性有待提升	利用金属掺杂的方式降低重金属催化剂的用量 载体效应（抗氧化载体）提高催化剂稳定性
	质子交换膜	质子交换膜的传导率、力学性能、稳定性等性质的优化	扩散/动态迁移过程模拟，指导新型质子交换膜的分子设计；利用材料复合方式改性质子交换膜
	电极/器件	流道、输运、热交换设计	器件仿真，设计优化

4.7 热电技术与 AI4S

热能被普遍认为是最清洁、却最难发电的能源类型之一。目前人类对热能发电的利用主要集中在地热发电，即利用地热驱动发电机发电，其对地域本身的环境禀赋要求极高，限制了其在全球范围的大规模应用。然而，随着对新兴热电材料的研究不断深入，其灵活，小型，低维护成本等特点正越来越收到重视。其潜在用途包括特种电源、绿色能源、环境能量收集与工业余热发电等。其中，最令人兴奋的是其在物联网（IOT）领域成为关键基础设施的可能性。在物流、制造、医疗等 IOT 常见场景中，如何在缺乏外部干预的条件下，持续为各种 IOT 设备供电，是 IOT 能否大规模应用的关键。可以说，电源问题将直接影响物联网整个行业的兴衰。热电材料凭借其温差发电，是一类具有无机械运动、无噪音、免维护、无污染、高可靠、长寿命的电源，也是物联网的"理论最优解"之一。

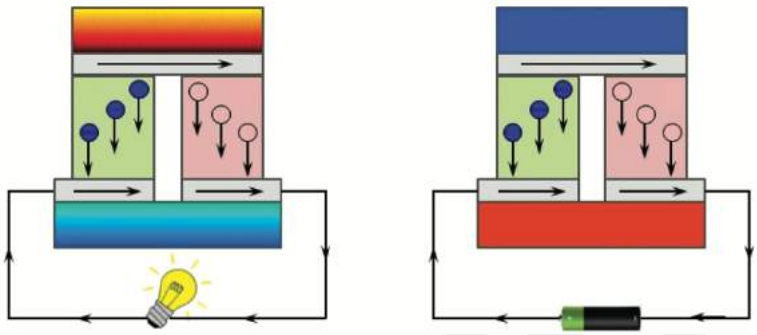


图. 热电发电原理示意图 [2]

热电材料的基本原理是由于温差的存在，一方面热端载流子具有比冷端附近载流子更高的动能，另一方面半导体材料中热端附近受热激发进入导带或价带的载流子数量也将高于冷端附近，从而引起材料内部载流子从热端到冷端的扩散。这样，冷端附近由于载流子的聚集会形成一个自建电场从而阻碍从热端向冷端输运的载流子。而当温差逐渐消失的时候，这一过程就会趋于平衡。[2]

硼亚磷化物（$B_{12}P_2$）是一种高温热电材料，因为它具有良好的热稳定性和化学惰性。然而，到目前为止，$B_{12}P_2$ 的热性能尚未得到很好的揭示，这限制了其应用。2022 年《Energy and AI》收录武汉大学和内华达大学研究成果。其中，研究人员利用深度势能对 $B_{12}P_2$ 进行了分子动力学建模。用其对晶体 $B_{12}P_2$ 的各向同性晶格导热系数进行了计算预测，得到 39.70±4.38 W/m·K，与实验数据较为一致。同时模拟了更大温度范围（300 ~ 1500K）导热系数和温度的关系（与 1/T 成比例），并发现其原因是声子散射导致的振动密度和声子参与率下降。该模型也成功计算了 $B_{12}P_2$ 的弹性性能（如剪切模量等）。该模型为以后探索 $B_{12}P_2$ 高温热电材料增加了理论基础，为这种技术的最终工业化提供助力。[1]

Source：

[1] Xiaona Huang, Yidi Shen, Qi An, Nanotwinning induced decreased lattice thermal conductivity of high temperature thermoelectric boron subphosphide (B12P2) from deep learning potential simulations, Energy and AI, Volume 8, 2022, 100135, ISSN 2666-5468, https://doi.org/10.1016/j.egyai.2022.100135.

[2] 热电材料的基本原理、关键问题及研究进展，Kai Guo，Jingtai Zhao，Article · January 2015，DOI: 10.3969/j.issn.0253-9608.2015.03.003)

4.8 储能技术与 AI4S

储能一般通过介质或设备将能量存储起来，在需要时再释放。[1] 常见的可再生能源如风能、太阳能等发电存在间隙性和不稳定性，这也间接造成了电网负荷的波动，储能的出现可以很好的弥补上述的弊端，除此之外，像电动车等移动设备的供能也离不开储能系统发挥作用。可以说，储能技术是可再生能源接入、智能电网以及电动汽车发展不可或缺的环节。

储能技术的分类多种多样，按照能量储存方式，最常见的方式有以下四种：机械储能，电化学储能，热储能以及电磁储能。其中机械储能主要包括抽水蓄能、压缩空气储能等，电化学储能主要包括锂离子电池、钠硫电池等，热储能主要包括熔融盐，电磁储能主要包括超级电容器储能等。上述各储能手段主要是将机械能、化学能以及热能等其他形式的能量最终转换为电能进行使用。

目前应用规模最大，技术最成熟的储能方式为抽水储能，同时它也是现在储能技术的主流。除此之外，热储能和电化学储能是当前发展聚焦的技术路线。相比抽水蓄能，电化学储能受地理条件影响较小，建设周期短，可灵活运用于电力系统各环节及其他各类场景中，是目前广受关注的储能方式。[4]

不同的储能方式主要是适应不同情况的储能要求。目前，成本高是影响储能大规模发展的一个主要因素，其关键材料、制造工艺和能量转化效率也是各种技术面临的共同挑战和瓶颈。目前以提升寿命、低成本、高安全为突出特征的储能电池是主要研究方向。[2]

技术名称	技术难点，科学问题	AI4S 优势
锂离子电池	安全运维难度高：随着使用时长增加，锂离子电池的理化特性会发生非线性变化，宏观表现为电池一致性发散、产热增加、安全边界演化、微小故障多发等特征[5]。	组分搜索、结构搜索提供最佳材料配方，为工艺生产提供有效指导； 高精度跨尺度建模揭示从电极材料到宏观储能系统尺度的性能演化规律，帮助电池储能安全的有效管控。[5]
钠硫电池	离子电导率低，难满足实际电池应用； 高温下容易发生热失控，安全性差[3]。	
熔盐	对熔盐结构和物理化学性质（熔点、热导、热容、黏度等）随组分和温度的变化规律缺乏清晰理解。	组分搜索优化，高精度建模，预测熔盐结构和物理化学性质，建立结构和性质间的构效关系，指导熔盐储能材料设计。

I. 化学储能

除了蓄水储能之外，化学储能（电池）是目前最普遍的储能手段。其中主流商用方案仍以锂电池为主。锂电池之外，钠电池凭借其自身的特殊性质，在储能领域也同样发挥着重要作用。然而钠电池的使用面临着较大的安全隐患：如，钠电池在运行过程中固态电解质的裂缝导致电池整体的温度升高，从而造成电池的热失控现象等。对于化学储能系统的设计，除了要保证固态电解质具有基本的离子导电性能外，同时还要保证足够的力学性能以及减小在充放电过程中可能出现的的表面局部电流密度集中，减缓固态电解质的电化学退化速度。*关于 AI4S 在电池的应用，请见 "本章 4.3 小节"

II. 热储能

与电化学储能一样，热储能作为一种能量高密度化、转换高效化、应用成本化的大容量规模化储能方式具有相当大的发展潜力，如 KCl-$CaCl_2$ 熔盐是未来聚光太阳能发电中传热和蓄热的潜在材料。华东理工大学路贵民团队基于第一性原理模拟结果训练的深度神经网络生成描述 KCl-$CaCl_2$ 熔盐的 AI4S 模型，进行了深度势分子动力学模拟，以预测 1100 K 下 KCl-$CaCl_2$ 熔盐的局部结构和热物理性质。在这项工作中，团队发现随着 $CaCl_2$ 的加入，随着网络结构的形成和增长，空间位阻效应越来越强烈，但 $CaCl_2$ 的增加使 Ca^{2+} 的配位壳层更加动态和活跃，两种反应不断竞争。然后，通过深度势分子动力学模拟计算剪切粘度、热容和热导率等热物性。可见，基于 AI4S 模型可以平衡复杂原子相互作用的正确描述，克服经典分子模拟中潜在参数缺失的挑战，比第一原理分子动力学模拟更有效，这将为下一个聚光太阳能发电提供一种探索更多熔融共晶盐的新方法。[6]

Source:
1. Wikipedia
2. 国际能源研究中心
3. 胡英瑛, 吴相伟, 温兆银. 储能钠硫电池的工程化研究进展与展望[J]. *储能科学与技术*, 2021, 10(3): 781-799
4. 《锂电储能行业深度报告：应用场景多点开花，万亿市场即将开启》
5. 中国科学报 朱汉斌 《锂离子电池储能系统热-安全管控技术获突破》
6. Bu M, Liang W, Lu G, et al. Local structure elucidation and properties prediction on KCl–CaCl2 molten salt: a deep potential molecular dynamics study[J]. Solar Energy Materials and Solar Cells, 2021, 232: 111346.
7. Gupta M K, Ding J, Osti N C, et al. Fast Na diffusion and anharmonic phonon dynamics in superionic Na 3 PS 4[J]. Energy & Environmental Science, 2021, 14(12): 6554-6563.
8. Pegolo P, Baroni S, Grasselli F. Temperature-and vacancy-concentration-dependence of heat transport in Li3ClO from multi-method numerical simulations[J]. npj Computational Materials, 2022, 8(1): 1-9.

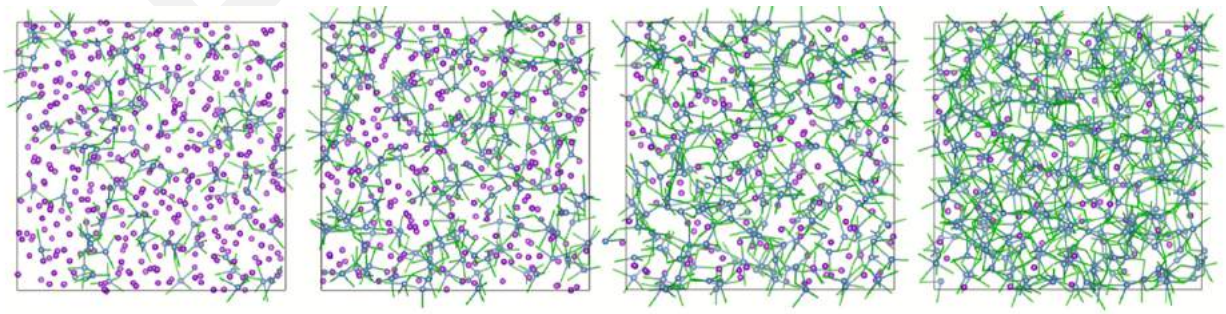


图. AI4S 模拟 KCl-CaCl2 熔盐中 CaCl2 分别为 20 mol%、40 mol%、60 mol%和 80 mol%的不同样品

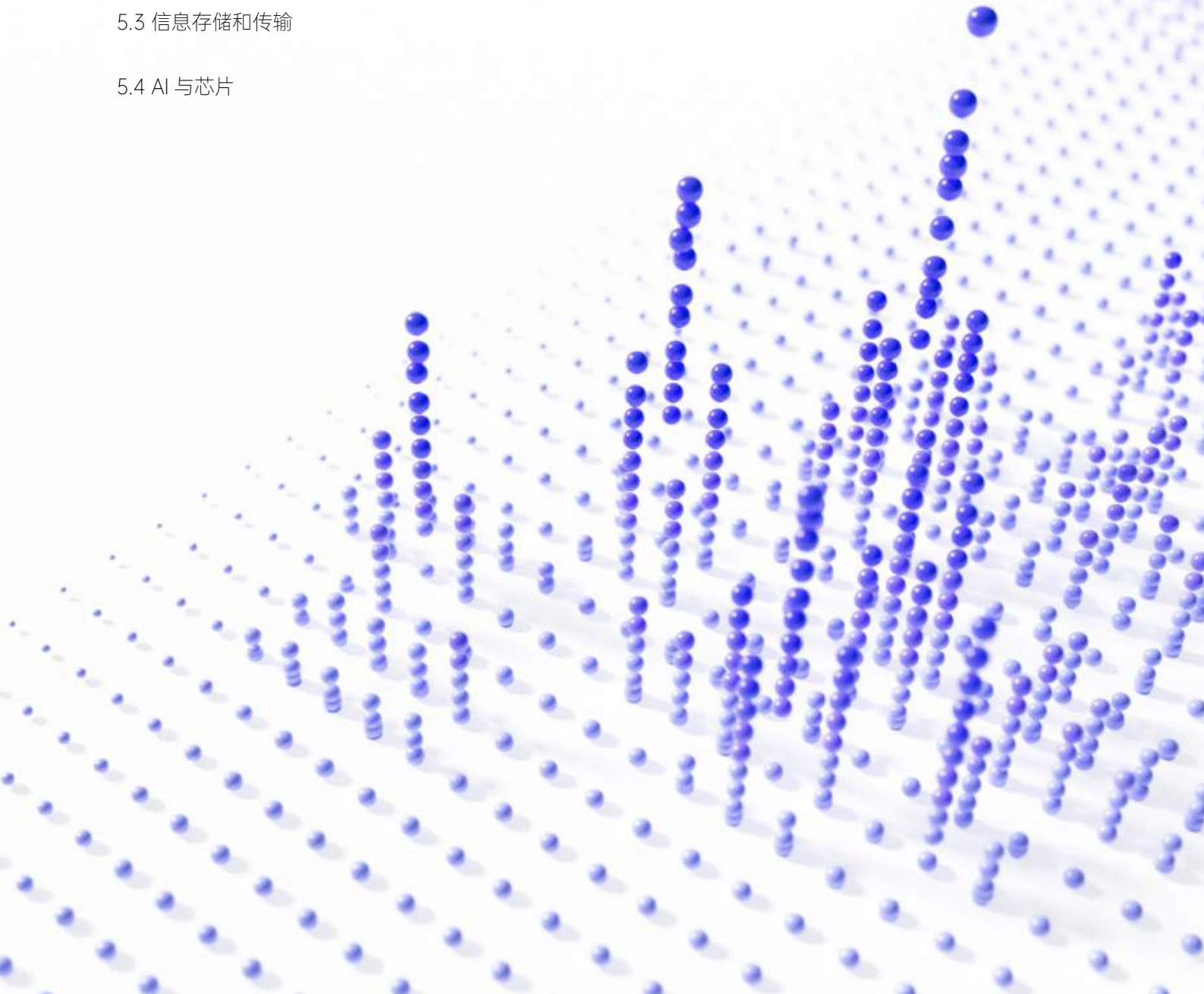
第五章：AI for Electronic Engineering & Computer Science 原理与实践

5.1 半导体材料和工艺: More Moore & More than Moore

5.2 显示材料

5.3 信息存储和传输

5.4 AI 与芯片



5.1 半导体材料与工艺

半导体技术是信息时代的基础。随着摩尔定律的衰退，关于半导体未来的发展，形成了两个不同的路线："more Moore" 和 "more than Moore"。前者是为了解决随着器件尺度不断减小，新制程的开发愈发困难的现状；而后者则是开发硅基半导体之外的新材料体系和工艺路线。

在传统硅基半导体技术中，三星近日官宣 3nm 芯片量产工艺。三星和台积电等头部芯片企业在研发先进制程时已经遇到微观尺度的建模瓶颈，靠传统 DFT 软件在效率和精度上难以做到顾全。随着器件尺度的不断减小，两大挑战成为硅基半导体材料发展的瓶颈。

第一个是由于尺寸缩小，栅氧化层的厚度仅仅只有几个原子那么厚，量子隧穿效应此时凸显，加剧了晶体管的漏电效应，因此如何减小漏电同时又能保证高性能计算和移动计算，成为该挑战的破局点，如 high-κ 栅极电介质、栅极全方位的新架构等。AI4S 有望通过对微尺度的高效高精度建模，助力相关新材料和器件体系的开发和应用。

第二，由于制程进步，对工具、工艺等环节要求越来越高，其中原子层沉积、光刻、蚀刻等工艺环节尤为严重。AI4S 有望通过工艺仿真→预测→搜索优化流程为工艺的改进提供助力。[1,2,3]

传统硅基之外，如何寻找新一代半导体材料也是业界关注的焦点[7]。氮化镓（GaN）、碳化硅（SiC）、砷化镓（AsGa）氧化镓（Ga_2O_3）等半导体材料凭借其高击穿电子场、高饱和电子漂移速度、高电子密度、高迁移率、高热导率、高抗辐射能力、可承受大功率等特点，在高温、高频、抗辐射及大功率器件方面有较为广泛的应用。[4] 由于目前如碳化硅等材料长晶速度慢、加工难度高（切片、薄化、抛光等）、缺陷密度较难去除等问题，使得该类材料仍需要进一步研究开发。

AI4S 借助高性能计算、物理建模和机器学习相结合的方式，一方面可以通过仿真模拟等手段，对半导体材料的工艺进行模拟，加速工艺的商业化应用，另一方面又可以对缺陷的生长过程进行计算建模，实现对缺陷的预测，此外在化学抛光等方面 AI4S 又可以结合高通量实验发掘最有效的抛光液材料。[6]

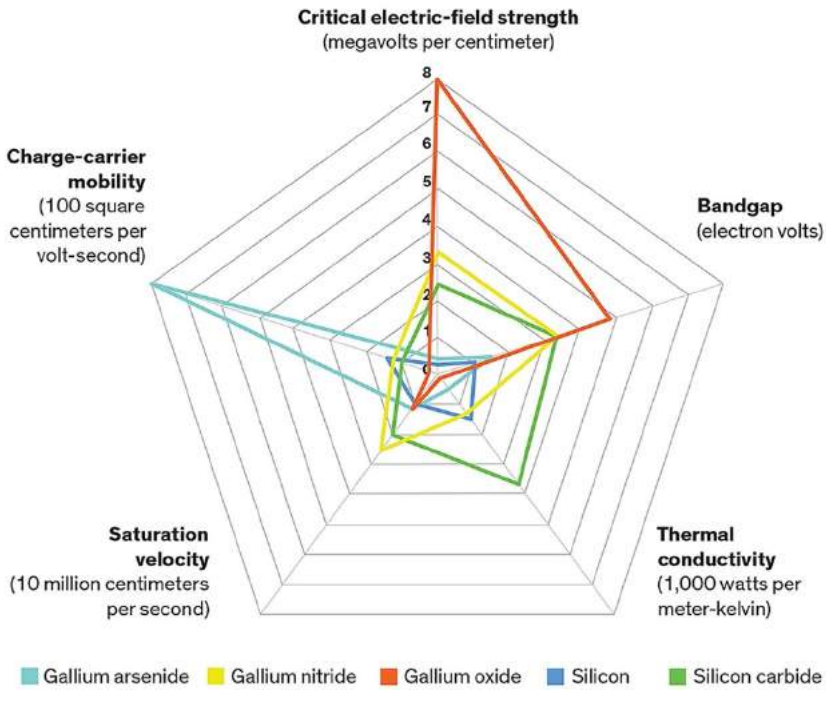
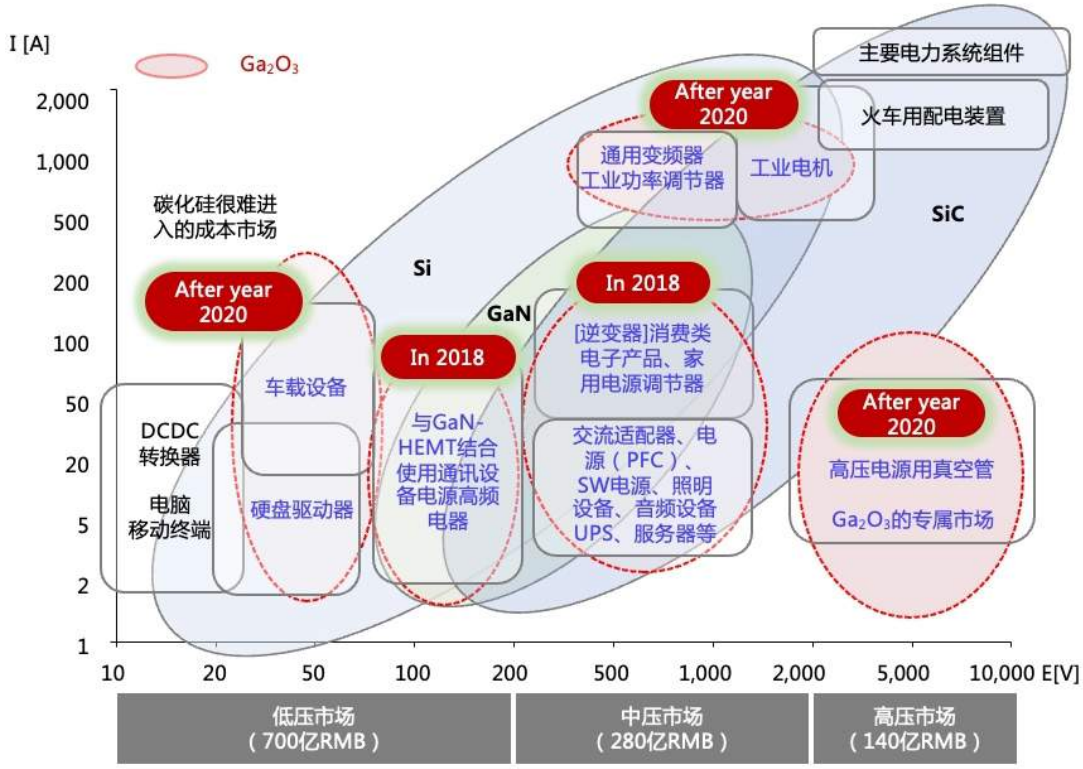


图. 不同半导体技术路线的理论性能 [1]



图：在电流和电压需求方面 Si，SiC，GaN 和 Ga_2O_3 功率电子器件的应用[5]

Source:

[1]. IRDS 2021 update more moore;

[2]. https://zhuanlan.zhihu.com/p/356942072

[3]. https://zhuanlan.zhihu.com/p/21262704

[4]. 2022-5-26 海通国际证券 比亚迪：中国"智"造

[5]. https://www.toutiao.com/article/6974930997062189599/?&source=m_redirect&wid=1657696174333

[6]. 2022-5-19 华安证券-国防军工深度报告：第三代半导体，能源转换链"绿芯"材料

[7] IEEE https://spectrum.ieee.org/gallium-oxide-the-supercharged-semiconductor

表 7：半导体设计与工艺中的 AI4S

		当前科研挑战	AI4S 应用
半导体材料	硅	传统 TCAD 考虑量子效应时计算成本高；而经验参数精度不足	高精度建模，从材料到器件（如 FinFET）模拟微观结构，预测性质
	GaN	深入理解 GaN 的微观结构和性质	
	SiC	通过研究微观结构来优化 SiC 的导热性、缺陷的理解、长晶等动力学、热力学过程	
芯片设计	EDA	消化巨量计算资源；且需要大量人工参与	AI 辅助电路设计
先进制程与工艺	晶圆加工	展新外延工艺；激光切割精确控制；研磨液、研磨垫选择	对反应过程进行原子尺度模拟，并与更高尺度模型（如 kMC）耦合，对工艺过程进行数字孪生；理性工艺参数设置，智能调控，提高经济效益
	氧化	氧化过程模拟	
	沉积	金属等材料的选取：耐腐蚀性、高导电率、高纯度、低接触电阻、稳定性；沉积过程模拟	
	光刻	光敏材料选取等；光化学反应模拟等	
	刻蚀	化学刻蚀的腐蚀液选取、反应模拟等；等离子体刻蚀的物质选取、激发态的游离基及化学性质活泼的中心原子团的分析等；工作气体等选取、刻蚀工艺模拟	
	互联	引入满足导线导电性要求、降低介电常数和满足可靠性要求的新材料	
	封装	封装材料选取；封装工艺模拟	
	测试	模拟测试过程、发掘测试问题	

5.1.1 “More Moore” -- AI4S 为硅半导体先进制程开发提供新工具

半导体工艺和器件仿真软件，是建立在半导体物理基础之上的数值仿真工具，可以对不同工艺条件进行仿真，取代或部分取代昂贵、费时的工艺实验；也可以对不同器件结构进行优化，以获得理想的特性。据国际半导体技术路线图（ITRS）报告，通过减少实验批量、缩短研发周期，半导体工艺和器件仿真软件可将全球半导体芯片全行业的研发总成本降低 40%。

随着集成电路技术的发展，行业龙头的先进制程进入 3nm 以下的技术节点，这对芯片所涉及的工艺、材料、器件结构都造成了巨大的挑战。 相比于旧的制程，新节点器件具有完全不同的器件结构和物理效应。传统的半导体工艺和器件仿真软件方法面临着巨大的挑战，其主要困难来自于：

1. 当器件到达深纳米尺度甚至原子尺度时，量子效应将起重要作用，而传统的模型无法精确、完整的处理量子效应的内容。其中集成电路中各类短沟道效应的出现即为最具代表性的表现。同时随着集成电路特征尺寸的减小，因为传统半导体工艺和器件仿真软件无法解决量子效应的问题，故单个器件的漏电也变得难以调控。

2. 传统的半导体工艺和器件仿真软件设计方法和软件基于大量的实验获得核心参数。然而当器件达到纳米尺度后，借助实验去获得可靠、有效的数据变得愈发困难，常常陷入时间长、成本高却产出少的困境。

3. 对于新的半导体材料和器件，目前成熟的 TCAD 方法还没有积累足够的数据，以进行精确仿真。

由此，基于密度泛函理论（DFT）的 Atomistic TCAD 逐渐成为产研各界的发力方向。

全球范围，EDA 龙头 Synopsys 收购 Quantumwise A/S 公司，通过结合 NEGF-DFT（非平衡格林函数-密度泛函）技术和传统 TCAD 基于的漂移扩散模型来实现技术的闭环；另一龙头 Silvaco 与 NEMO5 合作推进 Atomistic TCAD。NEMO5 聚焦在在原子和亚原子尺度，预测半导体和金属纳米线中的电子结构、声子结构和量子输运性质。国内企业鸿之微也开发了基于 NEGF-DFT 技术的原子尺度 TCAD，可以用于不同材料类型、不同器件结构的纳米器件的模拟。[1]

Atomistic TCAD 模拟半导体工艺过程的主要目标之一是预测半导体材料中掺杂原子的分布，这将对半导体器件的特性产生决定性影响。轻推弹性带（NEB）[2,3]方法通过在已知的初始状态和最终状态之间使用少量图像进行线性插值，据虚弹簧和实势能面给出的力优化每幅图像中的原子坐标，可以找到平衡位置之间的最小能量路径（MEP），因此许多研究人员使用它来研究各种材料中杂质的扩散机制。然而，由于 Atomistic TCAD 是基于第一性原理进行计算，其昂贵的计算成本限制了其在更接近真实环境的大规模原子系统中的应用。

同前文所述，AI4S 范式已在多个领域证明了其对大规模体系高效高精度的建模能力。具体到半导体领域，学界已有团队成功使用 AI4S 方法进行原子尺度的高效建模，在 DFT 精度下对半导体掺杂等问题进行有效研究。

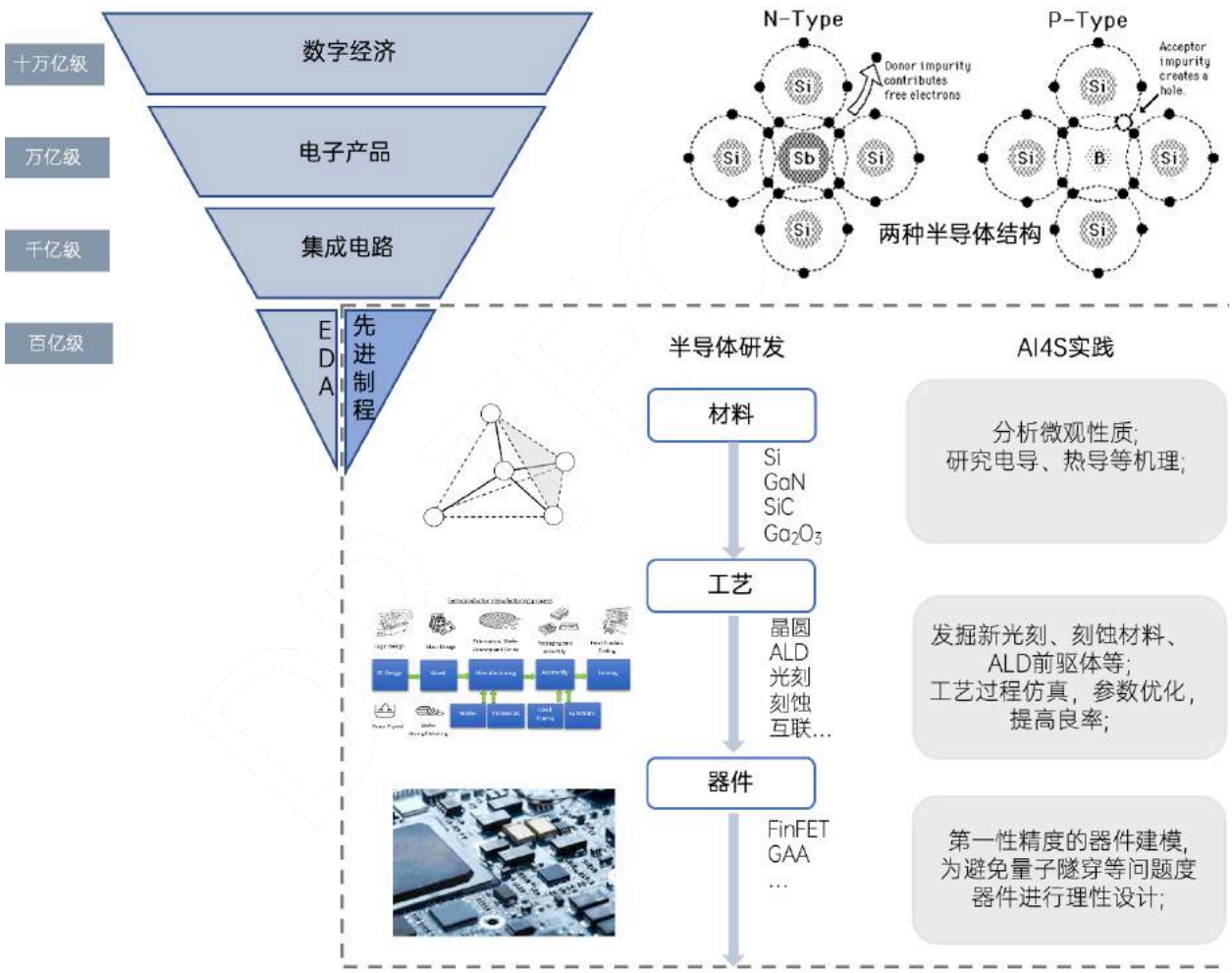


图. 半导体产业格局和 AI4S 实践 (Figure credit: DP Technology)

Source:

[1] https://zhuanlan.zhihu.com/p/358750347

[2] https://blog.csdn.net/yu1581274988/article/details/104931792

[3] https://commons.wikimedia.org/wiki/File:Silicon_Crystal_structure.svg

AI4S 实践（21）：湖南大学利用 AI4S 方法将半导体掺杂工艺仿真速度提高数万倍

一定数量的杂质掺入到半导体材料的工艺，是为了改变半导体材料的电学特性，从而得到所需的电学参数。根据掺杂原子情况可以分为两类：P 型掺杂（如硼等元素）：提供额外的空位实现空穴导电，及 N 型掺杂（如磷）：提供额外的电子实现电子导电。由于半导体很多电子特性都是和掺杂及其浓度有关，而掺杂的量较少，且掺杂会产生复杂的原子、电子相互作用，因此半导体掺杂是下一代半导体工艺中的核心难题。

湖南大学刘杰博士和团队利用深度势能建立了一种研究半导体中掺杂原子迁移模型的高效、准确方法。通过基于主动学习的采样方案，减少训练集冗余，同时将最小能量路径(MEP)附近的结构加入训练，进一步优化模型，以实现更精确的 NEB 计算。用提出来的建模方法对三个掺杂系统（硅掺硼、硅掺锂和氮化镓掺碳）进行轻推弹性带（NEB）模拟，结果表明，基于 AI4S 的方法比基于密度泛函理论（DFT）的原子掺杂迁移建模方法快 10^4~10^5倍，同时保持了 DFT 水平的高精度。

与基于 DFT 的 NEB 计算结果相比，本研究中的计算方法预测最小能量路径(MEP)的迁移能垒仅偏离约为 10-2eV（相对误差在 0.4%-3%之间），预测鞍点处的掺杂原子位置仅偏离约为 10-2Å（平均和最大差异在 0.004-0.02 Å 和 0.01-0.06 Å 之间）研究人员还测试了 AI4S 范式在研究掺杂原子迁移中的作用，使用 $Si_{64}B$、$Si_{64}Li$ 和 $Ga_{48}N_{48}C$ 作为模型系统，三者都是半导体应用中典型的掺杂体系。对于含有三种元素的 $Ga_{48}N_{48}C$ 系统，虽然需要更多的样品才能获得准确的电位，但通过不超过 3000 个样品的 AI 训练，也可以准确描述掺杂剂的迁移路径。这表明该方法适用于研究掺杂剂在各种半导体材料中的迁移行为。

整个建模过程完全基于第一性原理计算，没有任何经验参数。因此，对于经验参数较少的先进制程和新型半导体材料中的扩散和激活过程（物理机制尚不清楚，工艺实验成本很高），此项研究中的原子建模方法很有参考价值。

Source:

Xi Ding, Ming Tao, Junhua Li, Mingyuan Li, Mengchao Shi, Jiashu Chen, Zhen Tang, Francis Benistant, Jie Liu, Efficient and accurate atomistic modeling of dopant migration using deep neural network, Materials Science in Semiconductor Processing, Volume 143, 2022, 106513, ISSN 1369-8001, https://doi.org/10.1016/j.mssp.2022.106513.

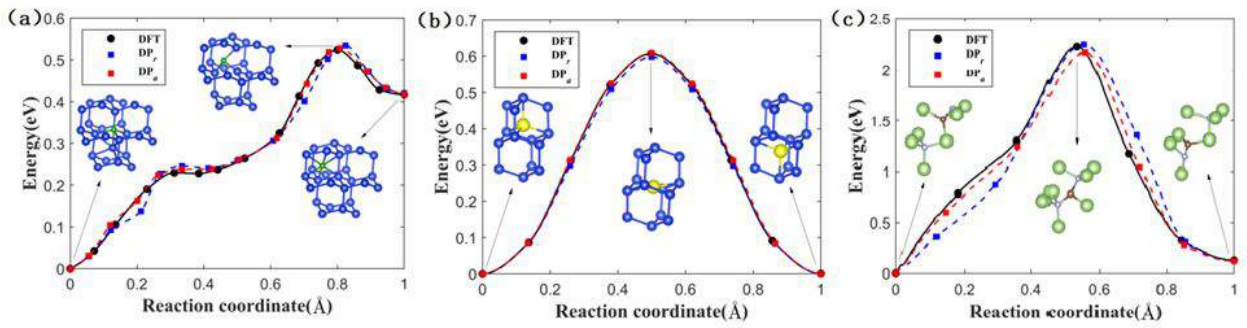


图. 各系统基于 DFT 和深度势能的 NEB 计算掺杂原子迁移路径结果

AI4S 实践（22）：《ACS Appl. Mater. Interfaces》报道原子层沉积(ALD)的化学反应动力学模拟，推动半导体工艺仿真的数字孪生

原子层沉积（ALD）是一种通过自限表面反应在不同衬底上沉积薄膜材料的气相技术，是半导体、显示材料、能源材料等领域的核心工艺步骤。与传统物理气相沉积（PVD）和化学气相沉积（CVD）相比，ALD 可以精确地控制薄膜的厚度、成分和结晶度，实现薄膜在衬底上均匀、共形和无针孔地生长，除此之外，ALD 还可以应用于更加广泛的温度范围。因此对于半导体先进制程，ALD 至关重要。

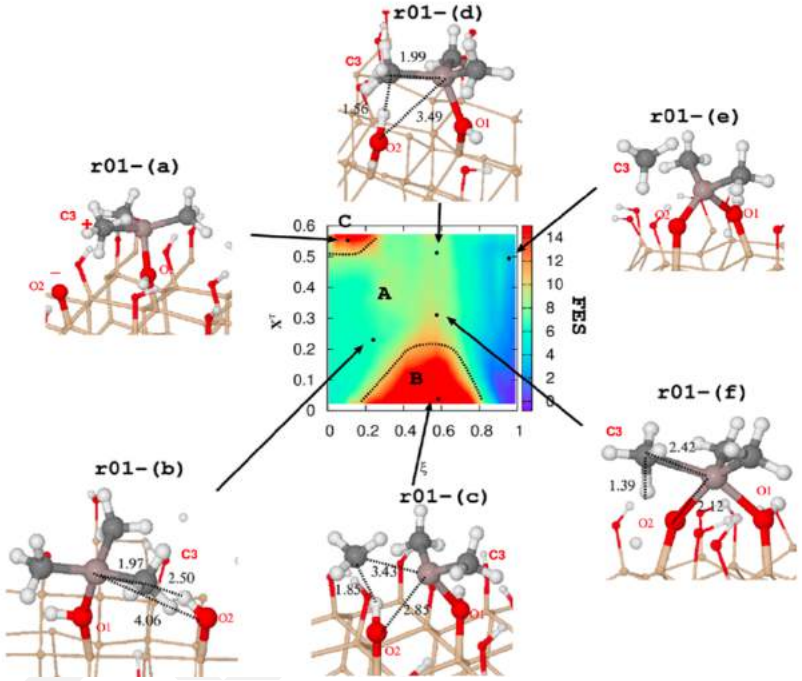


图. $Al(CH_3)_3$和 H_2O 在 OH/Si(111)表面 ALD 反应自由能面 [1]

ALD 的工艺反应过程十分复杂，传统的实验手段和建模分析方法存在明显的局限，如难以在化学空间内进行合适的前驱体、氧化物筛选；传统方法无法对多尺度界面问题进行有效建模；反应过程跨越的时空尺度大，无法实现准确的描述……上述科学问题也正是工业上难以进行大规模应用的原因。

ALD 过程包括初始吸附和随后的表面化学反应。近日，韩国庆北大学 Hiroya Nakata 等人利用基于深度势能的 AI4S 方法研究了 $Al(CH_3)_3$和水在 OH/Si(111)表面的 ALD 化学反应。

研究者将 $Al(CH_3)_3$和 H_2O 在 OH/Si(111)表面的 ALD 反应归纳为 6 条路径, 使用短时间的第一性计算和二维伞形采样技术(TDUS) 对 6 条反应路径(r01-r06)进行构型采样，以获得训练 AI4S 模型的初始数据集。本项目由 Samsung 资助实施，利用大规模集群共进行了大约 639,400 次第一性原理计算以产生训练神经网络需要的数据集；研究者表示，相比纯使用第一性原理计算， AI4S 方法减少了 100 ~ 1000 倍计算量需要。与此同时，深度势能方法仍达到了 DFT 精度。

模型进行吸附动力学模拟，以其中一条反应路径 r01 为例做详细介绍。上图为由模型得到的在 300K 下沿反应 r01 的二维自由能面(FES)，以及一些重要构型。可以看出 B 区和 C 区的自由能较高，因此该模型描述了反应应通过相对平坦的 A 区进行，例如 r01-(d)和 r01-(f)结构生成最终产物 r01-(e)。

基于上述的研究思路，研究人员可以探索通入 $Al(CH_3)_3$的最佳接触速度以及温度效应，为 ALD 工艺的理性优化方案提供准确、具体的科学依据。[1]

Source: [1] Nakata H, Filatov M, Choi C H. Accelerated Deep Learning Dynamics for Atomic Layer Deposition of Al (Me) 3 and Water on OH/Si (111)[J]. ACS Applied Materials & Interfaces, 2022.

AI4S 实践（23）： AI4S 模拟仿真硅基半导体在太空等极端工况条件下辐照损伤

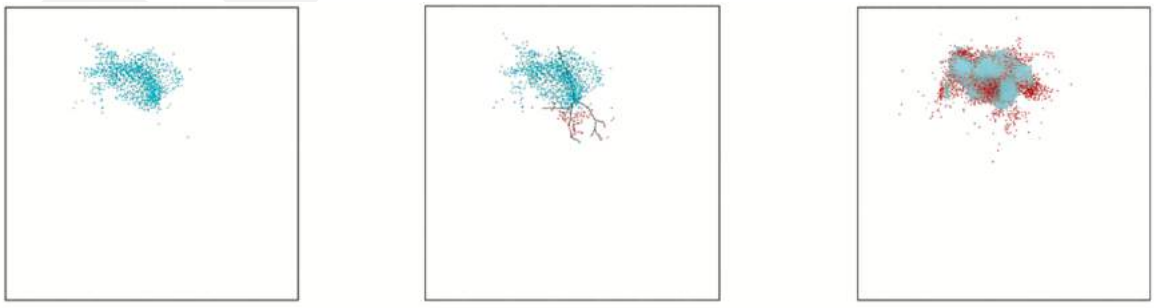
太空探测任务中,航天器上的电子设备会受到来自太阳和宇宙的强烈辐射影响,如高能粒子辐射、伽马射线等。这些强辐射会对半导体材料中的原子结构产生破坏和位移,导致半导体器件的参数发生改变,甚至失效。为了保障航天器的正常工作,必须研发出抗辐射性能优异的半导体器件。

2023 年，哈尔滨工业大学牛宏伟等在《Computational Materials Science》发表论文[1]。团队采用了先进的机器学习势方法,可以高效准确地模拟辐射条件下半导体材料的微观结构变化和缺陷产生过程。分子模拟提供了一种独特的方法来研究辐射诱导缺陷的动态产生、聚集和演变过程,以指导新型半导体组件的设计和增强。本文开发了一个更高效的机器学习势方法,可以达到 DFT 精度,用来研究硅材料的辐射损伤。该势方法的精确性通过比较静态属性、缺陷形成能和位移阈值能与实验和 DFT 数据进行了验证。通过模拟单个原子冲击,发现与经验势相比,机器学习势的原子冲击影响空间范围更大。最后通过顺序模拟多个原子冲击,发现前几个激发的原子产生了无定形区域,后续原子冲击穿过无定形区域时能量损耗更多,新的缺陷生成减少了 36%。

基于该模型,研究者系统地研究了不同能量的原子撞击下硅材料的辐射损伤过程,揭示了辐射条件下硅材料缺陷的产生、聚集和演变规律。这为设计防辐射措施,提高半导体器件的抗辐射能力提供了理论借鉴。

该研究展示了机器学习势在材料科学研究中的应用潜力。这种新方法不仅可以推动半导体抗辐射性能的提升,也可以广泛运用于材料设计的各个方面,实现材料性能的突破。相信机器学习势能为太空探测提供更可靠的电子设备支持,推动人类空间科技事业的发展。

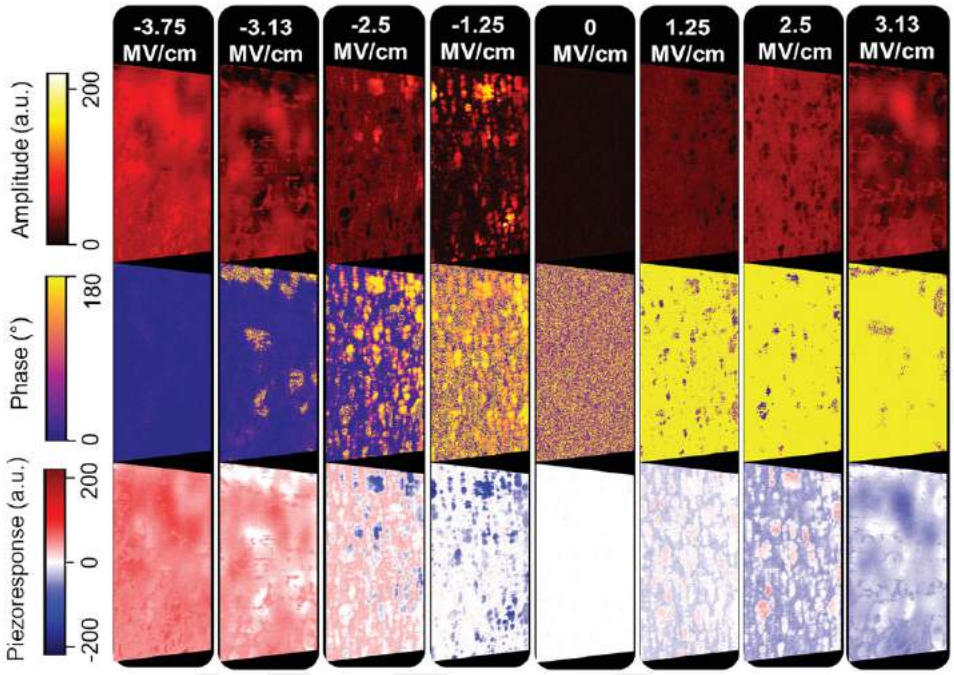
Source: [1] A machine-learning interatomic potential to understand primary radiation damage of silicon, Computational Materials Science 218 (2023) 111970



(a) 第一个 4 个原子弹击原子(PKA)产生的点缺陷(蓝色)组成的球形区域。(b) 第 5 个原子弹击原子(PKA)穿过了这个球形缺陷区域(蓝色),并在更深处产生了新的缺陷(红色)。黑线表示的是第 5 个原子弹击原子激活过程中能量原子的迁移路径。(c) 后续 8 个原子弹击原子穿过了这个球形区域(蓝色)。新产生的缺陷为红色。

AI4S 实践（24）：《AFM》 报道高 k 材料 ZrO2 反铁电效应在工况中工作与失效机理

随着 CMOS 器件的不断缩小,传统 SiO2 栅介质的等效氧化层厚度已经很难再进行缩薄。High-k 材料的使用保证了 MOS 器件可以继续按照摩尔定律发展到更小的工艺节点,是保持器件性能提升的必要选择。它已成为 CMOS 器件特别是微处理器的关键材料与必要组成部分,对延续半导体行业的发展起着举足轻重的作用。当前主流 CMOS 工艺中,栅介质几乎都采用了 high-k 材料。它也推动了新材料、新工艺的发展与突破。



BE-PFM 图像显示,在施加 350nm x 350nm 范围内不同的直流电场时,ZrO2 材料电场驱动的相变的快照式演变。在没有施加直流电场时,ZrO2 不具压电性。

HfO2 是目前高 k 材料的主流，在新型高 k 材料中，ZrO2 拥有更强大的理论性能，备受关注。通常 HfO2 的介电常数在 20-25 之间,而掺杂 Hf 的 ZrO2 可以达到 40-60。ZrO2 的性能来自其特殊的反铁电效应：反铁电材料具有极大的介电响应,这源自其非中心对称结构和可逆的电场驱动相变行为。ZrO2 的反铁电效应来源于其四方相和正交相之间的可逆相变。这种相变使其介电常数出现突变,从正常的 20-30 剧增到 40-60。

2023 年，领域顶刊 Advanced Functional Materials 报道了德国 TU Dresden 大学研究组的成果 [1] 。团队称这是人类首次在纳米尺度直接观察到了 ZrO2 基反铁电材料中的电场驱动的相变,证实了可逆的电场诱导相变是 ZrO2 反铁特有的非线性介电行为的起源。这一发现对于深入理解和利用反铁电性质具有重要意义。通过电场偏置的频率调制压电响应力显微成像技术(BE-PFM)直接区分了钉扎态的 La 掺杂 HfO2 铁电体和 ZrO2 反铁电体,明确 ferro 电畴不参与形成 ZrO2 的双复合循环电化学循环。。同时，团队发现 Hf 掺杂可提高 ZrO2 的压电响应,降低相变电场,这为优化 ZrO2 反铁电性能提供了思路。

团队使用基于深度学习的分子动力学，建立了电场作用下 ZrO2 从非极性四方相转变到极性正交相的模型。模拟了该过程中材料的压电响应、体积变化等,获得了与实验结果高度吻合的仿真结果。该研究可以视为实现了对 ZrO2 反铁电材料工作机理的物理模拟。模拟条件代表了实际工作中的电场作用,结果揭示了 ZrO2 的失效机制与可逆相变的关系。

Source: [1] https://doi.org/10.1002/adfm.202303636

5.1.2 "More than Moore" -- AI4S 探索第三代半导体技术路线

由于硅（Si）本身的微观性质和工艺所限，研究人员不断探索新的技术路线，对不同半导体的应用场景进行探索。其中，在功率半导体和射频等领域，均已实现了对硅的部分替代。以氮化镓（GaN）和碳化硅（SiC）为代表的新一代功率半导体材料具有耐高温、耐高压、高频率、大功率及抗辐射等特性，被达摩院评为 2021 年十大科技趋势。但由于材料工艺和成本因素限制，仍需进一步研究优化。传统的"炒菜试错"的材料研究手段需要大量的合成实验、验证实验，不仅耗时而且低效。对于功率半导体的研究，主要聚焦于其微观结构、性质及工艺。

碳化硅（SiC）是第三代半导体和核电站组件的基本材料。其导带间隙大，是传统 Si 的三倍，因此对能量、热量的提取比 Si 更容易。传统 Si 的可承受温度的极限是 175°C，而 SiC 可以达到 300°C，若封装方式正确的话甚至可以达到 900°C。且 SiC 的极限电子雪崩效应（雪崩效应是引起 pn 结击穿的一种机制）是 Si 的 8-10 倍。因此，SiC 是一种很好的功率半导体材料。但其应用很大程度上取决于它的导热性，而碳化硅的导热性又与微观结构高度相关，因此分子模拟（MD）是解决纳米器件或微观结构中热传递机制的常用方法。但传统的分子模拟对原子间电势的描述缺乏准确性，限制了传统模拟手段的应用。

中科院的研究团队利用 AI4S 方法建模，基于第一性原理的计算，对有 8000 个原子的立方碳化硅（3C-SiC）在 300K 的温度下进行分子动力学模拟，并将模拟得到的红外共振频率和声子线宽等结构数据与 Vashishta（VA）电位和实验得到的值进行比较，模型的结果与实验测量非常吻合。该结果表明可以对大型 SiC 系统进行分子动力学模拟计算得到介电特性，而且此结果的准确性与基于第一性原理计算的准确性一致。该研究为研究介电光谱提供了一种新方法，且可以推广到其他介电材料。

氧化镓的别名是三氧化二镓，氧化镓（Ga_2O_3）是一种宽禁带半导体，Eg=4.9eV，同时其生产成本较低，具有较广阔的技术前景。Ga_2O_3 是一种透明的氧化物半导体材料，在光电子器件方面有广阔的应用前景 ，被用作于 Ga 基半导体材料的绝缘层，以及紫外线滤光片。它还可以用作 O2 化学探测器。[2]

但由于氧化镓导热性能差，相比于 SiC，其导热性只有前者的十分之一，此外氧化镓制造 p 型半导体的难度高，一定程度上限制了氧化镓的应用。

来自美国 University of Notre Dame 的研究团队对 b- Ga_2O_3 的热导率开发了深度势能模型，基于第一性原理计算，通过分子动力学模拟（MD）方法对该物质的热导率、声子传输特性进行了建模分析，计算的在三个晶体方向的晶格热导率值与实验结果十分吻合，且发现了 b- Ga_2O_3 室温下热导率的各向异性。[3] 之后，研究人员结合 Green-Kubo 模态量化了不同声子模式对热传输的贡献，发现光学声子模式（phonon modes）在热传输中发挥了关键作用。

该研究为解决 Ga_2O_3 低热导率问题提出了新方向，同时也为探索复杂半导体材料热传输物理做出了贡献。此外，该模型的成功为建立其他特性（如机械特性等）做了铺垫，有利于加速人们对 Ga_2O_3 的进一步理解，推动第三代半导体材料的商业化落地及应用。

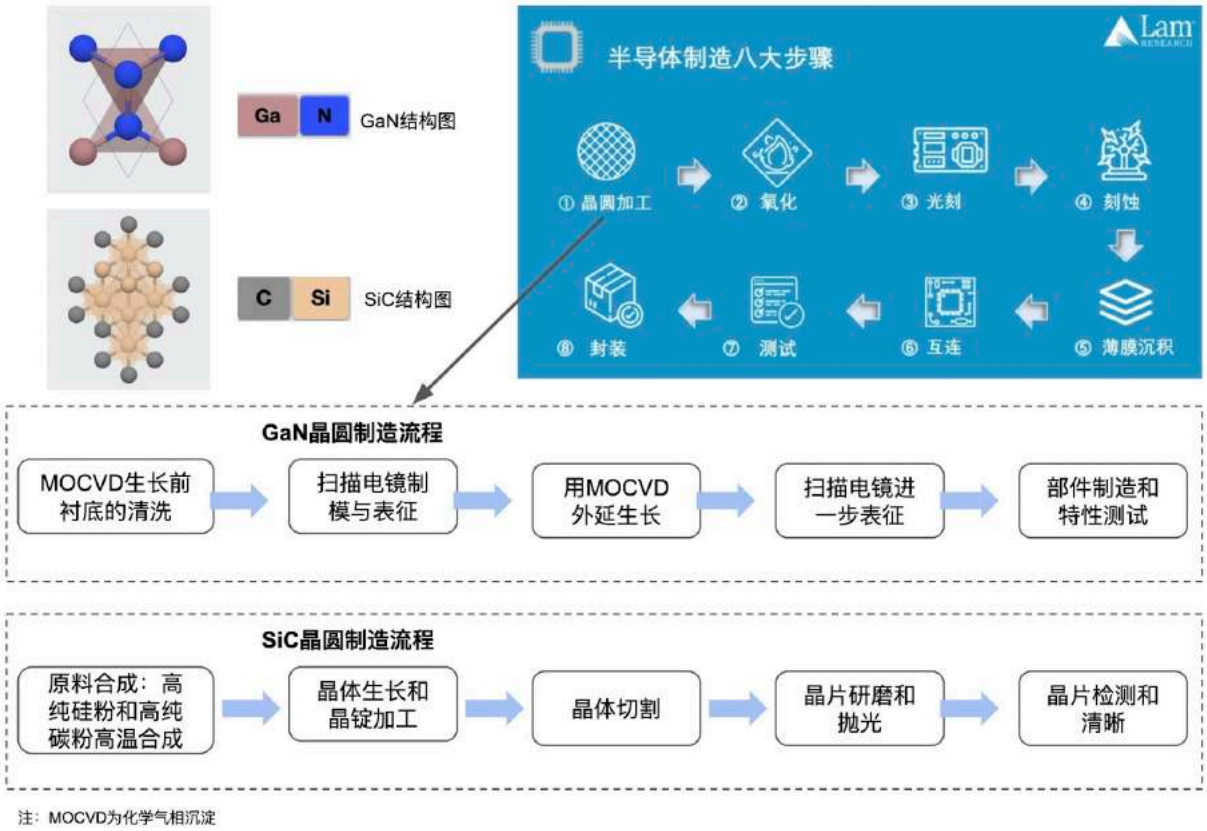


图. GaN、SiC 材料制备流程是半导体材料的核心（来源：Lam Research [1]）

Source：

[1] Lam Research, retrieved at https://www.siscmag.com/news/show-4355.html

[2] A deep neural network interatomic potential for studying thermal conductivity of β-Ga2O3 Appl. Phys. Lett. 117, 152102 (2020); https://doi.org/10.1063/5.0025051

AI4S 实践（25）：从量子力学到有限元，多尺度研究 GaN-BAs 高性能功率半导体器件

逐渐缩小的电子器件对半导体设备的散热性要求逐渐变高。因此寻找高导热性的半导体材料成为关键。BAs-GaN 是一种高效的散热材料，但由于缺乏对其原子间势能的进一步理解，使得对该类材料的应用受限。在一项研究中，湖南大学与 TU Darmstadt 等研究机构利用机器学习构建了砷化硼(BAs)、氮化镓(GaN)及其异质结的准确和高效的智能势，并进行了从量子力学到有限元的多尺度模拟研究 BAs-GaN 异质结的热传输特性。结果表明 BAs-GaN 界面的界面导热系数高达 265 MW m-2K-1,归因于 BAs 和 GaN 的晶格振动匹配良好。进一步的有限元分析揭示了粒径和界面的竞争关系,不同尺度下的有效热导率呈现不同变化趋势。

但是 BAs-GaN 异质结的界面热阻是一个关键因素,会影响整体的散热性能。需要对界面热传导机制进行理论分析和计算,为实际的器件集成提供指导。

本文利用 AI for Science 多尺度模拟的方法,深入预测和揭示了 BAs-GaN 异质结的界面热传导特性,可以为进一步优化这种散热方案提供理论支持。在纳米尺度的分子动力学模拟的基础上,利用有限元方法建立了微米级的 BAs-GaN 异质结模型。

具体而言，团队使用机器学习方法,通过 DFT 计算训练构建了 BAs 和 GaN 的深度势能来描述原子间力，从而进行大规模分子动力学模拟,计算 BAs 和 GaN 的热导率,以及 BAs-GaN 的界面热传导系数。此后，基于 MD 结果,建立 BAs-GaN 微米级有限元模型,计算不同粒径大小下的有效热导率。最后使用 FEM 模型桥接了纳米尺度的热传导机理与实际器件的热设计需求。这样,整个多尺度模型从量子级构建经典力场出发,再到连续介质计算,不同方法在各自适用尺度发挥优势。

这项技术有望解决目前微纳电子器件面临的热管理难题,推动相关器件的持续 miniaturization 和高性能化。值得各界研究者关注。

Source: [1] https://arxiv.org/pdf/2201.00516.pdf

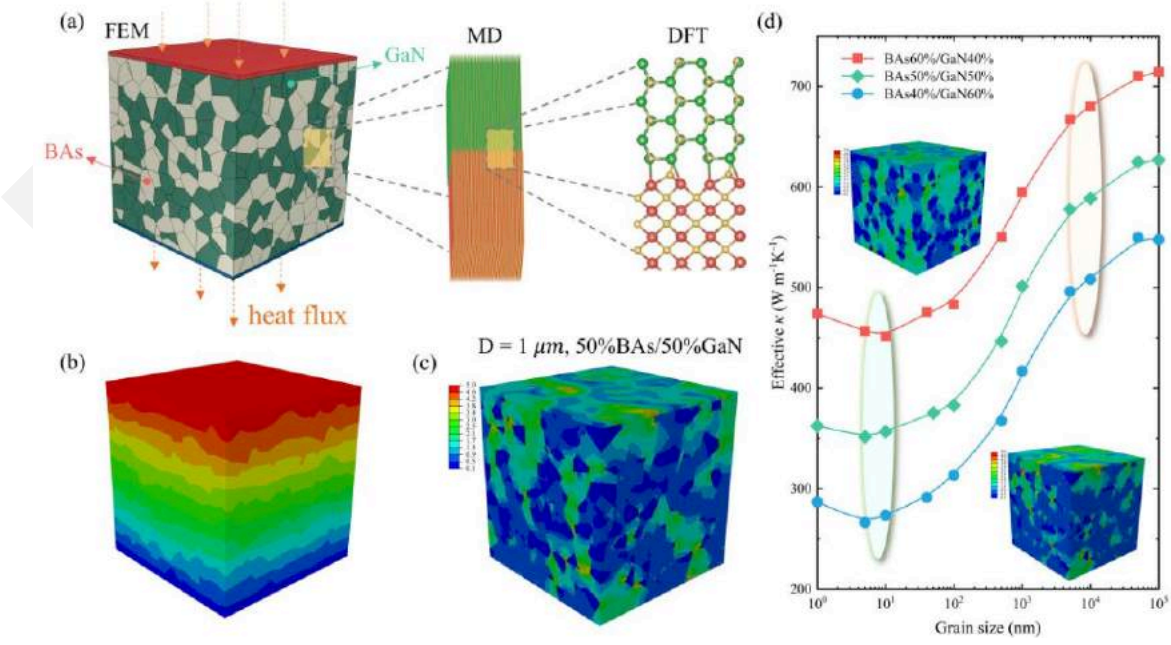


图. 不同百分比 BAs 和 GaN 的有限元模型及其有效热导率。(a)含 50%BAs 和 50%GaN 的有限元模型,以及多尺度示意图。(b-c)在典型 1 微米尺度下的温度分布和热通量分布。(d)不同比例 BAs 和 GaN 模型中随粒径变化的有效热导率。插入图显示了 40%BAs - 60%GaN 和 60%BAs - 40%GaN 的热通量分布。彩色椭圆标出了随粒径增大有效热导率的转折点。

5.2 显示材料

显示产业是涉及光电子材料与器件、光学工程、微电子与电子学等领域的综合性产业，新一代技术变革往往由材料和器件的突破所带动。近年来，随着第五代移动通信、大数据、人工智能等新一代信息技术的出现，新型显示与超高清视频、柔性、传感、印刷电子等技术交叉融合，丰富了新型显示终端产品体系。

印刷显示、Mini/Micro-LED 显示、激光显示、电子纸显示等新型显示技术和产品基本具备了产业化条件，引导了显示材料、器件、装备到制造技术在内的整个显示产业链的一次全面技术革新。但同时消费者更苛刻的需求正在推动制造商开发更高色彩饱和度、更明亮、更高空间分辨率且适应性更强的屏幕。但目前主要的显示材料如高温 OLED 材料、激光显示仍处于科研阶段，距离商业化仍有一段道路。[12,13]

借助 AI4S 方法，尤其是 AI 强化的跨尺度建模技术，有机会实现高效地筛选和设计发光材料、OLED 材料（如荧光材料、磷光材料）、激光显示材料（如镓铟磷，氮化镓）、有机半导体材料、QD 材料（如硒化镉），并且模拟如发光、空穴/电子传输和迁移等微观性质，同时仿真材料、设备以及终端产品中的应用，辅助高通量筛选，同时探索外界条件对显示材料的影响，缩短实验与产业化距离，从而加速高温 OLED 等显示材料商业化。新的研发范式也有机会帮助国内企业追赶国际先进水平，突破壁垒。

LCD vs. OLED （图片来自 Corning）

AI4S 实践（26）：《Advanced Optical Materials》报道基于自然科学大模型的高通量 OLED 材料配方筛选工作流

OLED 技术自 1987 年首次示范以来,因其卓越的色彩表现和柔性制造能力,逐渐成为消费电子主流显示技术。尽管 OLED 商业化多年,但提高效率和扩大色域等方面仍面临挑战,需要发现新型发光材料。传统材料设计依赖专家经验和大量实验验证,但试错成本高。OLED Ir(III)体系是一种基于有机发光二极管（OLED）技术的发光材料体系，其中使用了含铱（Ir）的磷光材料。这种体系具有高效、低功率消耗、高亮度和高稳定性等优点，因此在电子显示领域有广泛的应用。其中，Ir(III)配合物材料具有较高的光效率和发光寿命，可以用于制备高效的红、绿、蓝光发射器件。OLED Ir(III)体系在智能手机、平板电脑、电视、汽车仪表盘等领域都有着广泛的应用前景。

高通量虚拟筛选是突破性新工具,可以全面探索化学空间,预测最有前途的候选材料。但筛选标准的生成需要大量实验数据或量子力学计算。实验数据有限,精确量子力学计算又非常耗时,这限制了高通量虚拟筛选的应用。

分子表征学习通过预训练学习大量无标记数据,再精调得到物性预测模型,可有效解决上述问题。2022 年提出的 Uni-Mol 直接以 3D 结构为输入、基于 Transformer 骨架的自然科学领域预训练大模型,是预测材料结构相关性质的有竞争力工具。它利用大量目标系统的 3D 构型信息,在预测有机和医药分子能量特性上媲美最优方法。

2023 年北京科学智能研究院（AISI）团队在国际权威期刊 Advanced Optical Materials 发表论文,将 Uni-Mol 应用于有机发光材料的高通量虚拟筛选[1]。发光特性与 3D 构型高度相关。结合有限量子

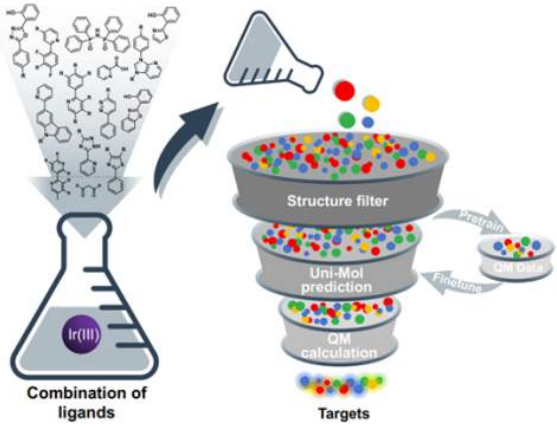


图. 高通量筛选 OLED 材料流程示意图

力学基准计算和百万级自动构建结构的 Uni-Mol 训练,可以高效预测 Ir(III)配合物等典型磷光材料的光物性,筛选出潜在优良候选。该计算协议可以轻松转移到其他有机材料如 TADF 和有机光伏材料,使各领域材料设计更有效。通过利用 Uni-Mol 强大的预测能力，团队极大地降低额外的计算成本，同时高通量的筛选迭代也能够进一步提高模型的预测效果（如下图左图所示）。这种大模型训练和 QM 小规模计算相互迭代的思路也将成为材料研发的一种新型范式。

其技术思路技术思路如下:

1）构建 Ir(III)配合物候选池

- 收集文献报道的 278 种双齿配体,随机组合生成数百万个 Ir(III)配合物候选分子。
- 使用半经验量子化学方法(GFN2-xTB)对所有候选分子进行优化,得到 3D 结构。
- 设置结构合理性过滤器,过滤掉非物理构型。得到约 160 万个合理候选分子。

2）对 Uni-Mol 进行改造

- 使用 GFN2-xTB 计算的轨道能量作为预训练标签,在全候选池上对 Uni-Mol 进行预训练。

- 随机选择约 1500 个分子,使用高精度量子化学方法计算 HOMO、LUMO、T1 激发能等光物性作为 fine-tune 的数据集。
- 通过预训练+微调,得到可以预测 Ir(III)配合物发光特性的 Uni-Mol 模型。

3）候选分子筛选

- 根据 Uni-Mol 预测的发光色坐标,筛选出红、黄、绿、蓝发光体。
- 根据预测的量子产额等筛选高效发光体。
- 根据配体种类数、分子量等指标排除难以合成的复杂分子。
- 得到用于显示和照明的潜在优良发光体。预测已知优良发光体性能,验证模型准确性。

这项研究首次将分子表征学习 Uni-Mol 应用于有机发光材料虚拟筛选，使用大规模 OLED 专项数据进行预训练,替代原始的一般分子数据，并提出端到端的自动化计算流程,实现大规模高通量材料筛选。这项研究为 OLED 技术的发展提供了一种新的高效材料筛选方案。其机器学习方法也可扩展到其他光电子材料的设计,值得业界关注。它开辟了利用先进计算工具辅助材料发现的新思路。

Source: [1] Automatic Screen-out of Ir(III) Complex Emitters by Combined Machine Learning and Computational Analysis, https://doi.org/10.1002/adom.202301093

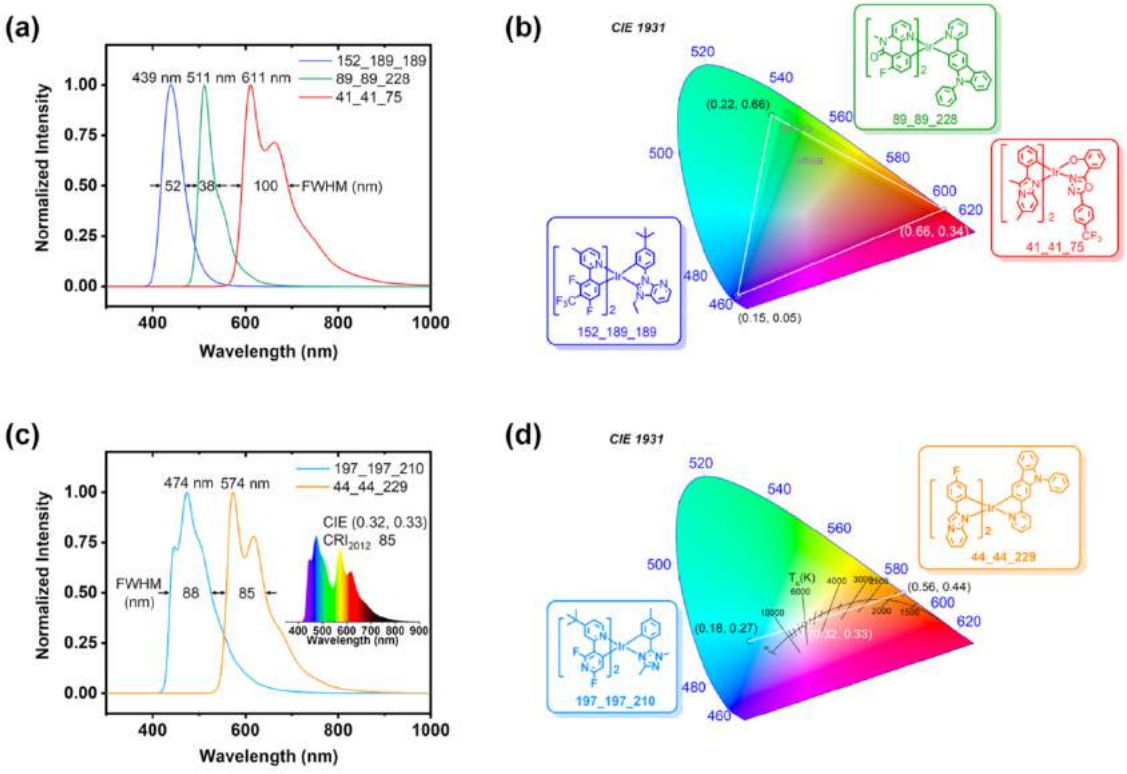


图. 用于显示[(a)和(b)]和照明[(c)和(d)]的筛选出候选物的预测发射谱和 CIE 坐标。(a) 展示了用于显示的红色(41 41 75)、绿色(89 89 228)和蓝色(152 189 189)发光体的预测发射谱。其中绿色和蓝色发光体具有非常窄的半高宽度(分别为 38 nm 和 52 nm),可以实现更广阔的色域。(b) 展示了这三种颜色发光体可以覆盖的色域接近 DCI-P3 标准,明显宽于 sRGB 标准。(c) 展示了用于照明的天蓝色(197 197 210)和黄橙色(44 44 229)发光体的预测发射谱。两者主发射峰分别位于 474 nm 和 574 nm,半高宽度均大于 85 nm。(d) 通过调节两种颜色的混合比例,可以获得 1931 CIE (0.32,0.33)的理想白光。对应的 CRI 为 85,适合日常照明。

5.3 信息存储和传输

存储信息科学行业的重要环节，而存储材料则是这一环节的核心基础。通过各类存储材料，人类不断扩充存储设备的存储容量、提升存储速度、降低单位价格等，助力信息科学迅速发展。 在信息爆炸的时代，随着人工智能、云计算等技术的推动，再加上人们处理的科学、工业等问题复杂性增加，全球数据量正呈现出爆炸式扩展和增加，国际数据公司（IDC）预测，全球 90%的数据量将在这几年内产生，2025 全球数据量将达到 163ZB。因此如何存储这些数据、如何发展更好的存储材料，成为当下制约信息科学进一步发展的难题。[1]

在信息传输中，光纤，即光导纤维，是一种由玻璃或塑料制成的纤维，是信息领域通信中必要的材料。随着信息科学的不断发展以及对电信通讯需求的增长，其限制性也日益凸显出来，如质地脆、机械强度低、维修成本高、对高强度光纤表现不好等，成为限制光通信发展的关键因素。随着 5G 和未来更高性能的网络技术的发展以及人们对高清视频和高质量数据的需求，光纤技术正向着更长波段、更高强度的光线发展，因此研究长波长光纤材料及耐高强度激光材料，需要突破传统的二氧化硅纤维材料。

AI4S 可以加速相关新材料的选择和应用。首先是对传统石英纤维的深入探索，AI4S 可以通过微观模拟二氧化硅材料及掺杂其他元素实现短时间内获得性能的优化，扩展现有材料的使用边界。AI4S 还能助力新型传输材料的发现及应用。如硫化物、聚合物、氟化物（如氟化锆）等面对长波段、高强度光有较好的表现，AI4S 可以从多尺度对其进行建模分析，从微观研究其构效关系，并以此为基础通过高通量筛选找到其它潜在的新型光纤材料。在实际生产过程中可以充分利用已有物质、协调、优化光纤制造流程和用料，降低成本。AI4S 还可以在研究阶段模拟光纤的实际运行情况，发现潜在问题，即使调整，加速产业化落地。AI4S 可以通过优化光纤的外壁材料，如环氧树脂、聚氨酯等材料，提高光纤整体的耐腐蚀性等性能，拓宽应用场景，降低后期维修成本。除此之外，AI4S 亦可以对量子通信等前沿技术中涉及的材料等进行筛选和优化，助力信息传输领域的"渐进式"和"跨越式"发展。

AI4S 通过新范式筛选加速新材料优化、发现和筛选，助力信息存储发展。除了前文提到的 AI4S 对半导体材料的促进作用之外，其还可以模拟、分析材料磁性结构等，解决关键构效问题，同时依靠高通量筛选手段，低成本、高效率筛选新的可用磁性材料，加速磁性存储材料发展。亦可以筛选新型材料用于存储技术，如铁电材料、纳米材料、新型氧化物、相变存储材料等，构建、合成低能耗、高存储量的材料，扩大存储材料空间；同时可以模拟新存储材料器件的性质，加快新材料产业化的验证过程。[2-3]

Source:

[1]. IDC
https://www.idc.com/getdoc.jsp?containerId=prUS47560321

[2]. https://doi.org/10.1103/physrevlett.120.145301

[3]. https://doi.org/10.1038/s41524-022-00712-y

AI4S 实践（27）：AI4S 构建二维铁电材料精确力场，为 FeRAM 的发展增加理论储备

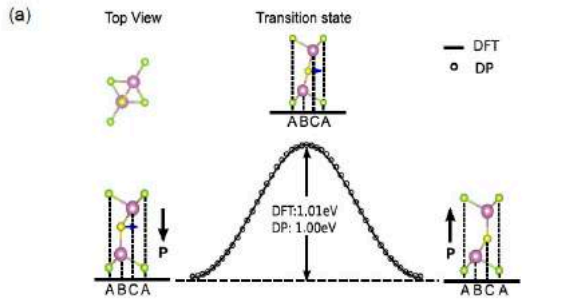
二维铁电材料（二维范德瓦尔斯 vdW 铁电体）是一类具有原子层厚度的层状材料，在半导体（多态存储器、光电二极管等）以及能源（钙钛矿太阳能电池、超级电容器等）领域具有广阔的应用前景。铁电的功能很大程度上取决于外部电/应力场作用下的极化开关的动态，而这种开关则是信息储存写入与擦除的关键。而在这其中，二维极化开关的机制和动力学在很大程度上仍未被研究。由于铁电的磁畴切换需要较高的空间/时间分辨率，传统的模拟和计算收到有较大的瓶颈。

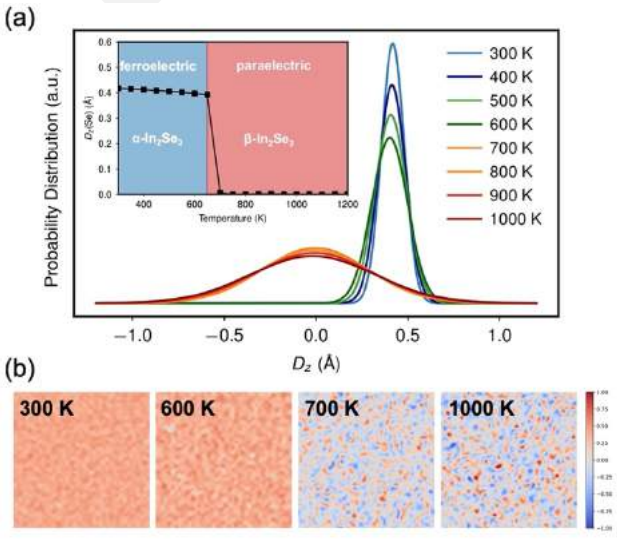
来着复旦大学、西湖大学等高校的研究团队利用 AI4S 方法开发了一个基于深度神经网络（DNN）的力场，适用于铁电单层 α-In_2Se_3。该模型对 In_2Se_3 种原子能量和原子力的预测误差仅为 5.72meV/atom。

基于此，模型模拟了 DFT 对单层 In_2Se_3 的各种动力学特性的结果。一些极化转换过程的能垒与 DFT 值相当吻合（右上图）。尤其是重现了单层 α-In_2Se_3 的 180°的畴壁运动（电流驱动），与 DFT 计算结果一致，对于理解铁电开关发挥重要作用。

该模型精确描述了 α-In_2Se_3 的相变过程及再现了 DFT 对单层 In_2Se_3 的各种相的热力学特性的结果。特别是，单层 α-In_2Se_3 的温度驱动的铁电-准电相变被 AI4S 模型有效模拟捕捉到，其理论 TC ≈ 650 K 与最近的实验值（≈ 700 K）相匹配（右下图）。

该工作为进一步研究单层 α-In_2Se_3 和相关系统中二维铁电性的动力学铺平了道路，是研究其动态过程的起点。研究者相信该 AI4S 模型未来能够得到系统的改进和扩展，从而能够实现大规模模拟层间滑动铁电性和多层的 Moiré铁电性（解释铁电性），为日后二维铁电材料的研究增加基础。[1]





Source:

[1] Wu, Jing and Bai, Liyi and Huang, Jiawei and Ma, Liyang and Liu, Jian and Liu, Shi, Accurate force field of two-dimensional ferroelectrics from deep learning, PhysRevB.104.174107, 10.1103/PhysRevB.104.174107

5.4 “AI 设计芯片” 与 “AI 专用芯片”

传统芯片设计需要行业专家的多年积累的经验以及复杂的仪器、软件等，使得芯片设计门槛高。AI 技术充分利用已有的数据和设计工具，将工作方式变成可持续积累模式。AI 驱动的芯片设计涉及在工具流程中使用机器学习等人工智能技术来设计、验证和测试半导体器件。例如,找到功耗、性能和面积(PPA)之间最佳平衡的解决方案空间非常大。存在大量的可变输入参数,会导致不同的结果。从本质上讲,人类不可能在给定的时间框架内探索所有这些组合来找到最佳结果,这会牺牲掉一些性能。

今天的 AI 芯片设计解决方案通常使用强化学习来探索解决方案空间并识别优化目标。作为一种决策科学,强化学习通过与环境的互动和观察环境的响应方式来学习最佳行为,以获得最大回报。这个过程涉及一边学习一边探索,有点像试错法。因此,强化学习会随着时间的推移产生更好的结果。

强化学习非常适合基于电子设计自动化(EDA)工作负载,因为它可以全面分析复杂问题,以人类单独无法实现的速度求解。强化学习算法可以快速适应和响应环境变化,并以持续的动态方式进行学习。

半导体行业也正在开始探索基于生成对抗网络的 AI 在芯片开发中的应用。对于 EDA,芯片设计相关的数据在很大程度上是专有的,生成对抗网络具有支持更定制化平台或可能增强内部流程以提高生产力的潜力。通过将更大的智能和更快的速度这两者强大的结合应用于否则重复的任务,AI 驱动的芯片设计可以产生更好的硅芯片结果和大大提高的工程生产力。AI 芯片设计具有各种优势,包括:

- **提高 PPA**。在每次芯片设计中,都有提高其目标应用的 PPA 的机会。但是,在一个近乎无限的设计选择和大规模设计空间中,人类不可能在项目时间框架内找到正确的选择。AI 可以通过探索这些大规模设计空间来识别优化领域,以提高 PPA。
- **提高生产力**。工程师一直报告在资源缩减和人才短缺的环境下工作量繁重。通过处理迭代任务,AI 使工程师能够专注于芯片设计的差异化和质量,同时满足上市时间目标。由于一个项目的学习可以保留并应用于下一个项目,AI 可进一步提高芯片开发过程的效率。

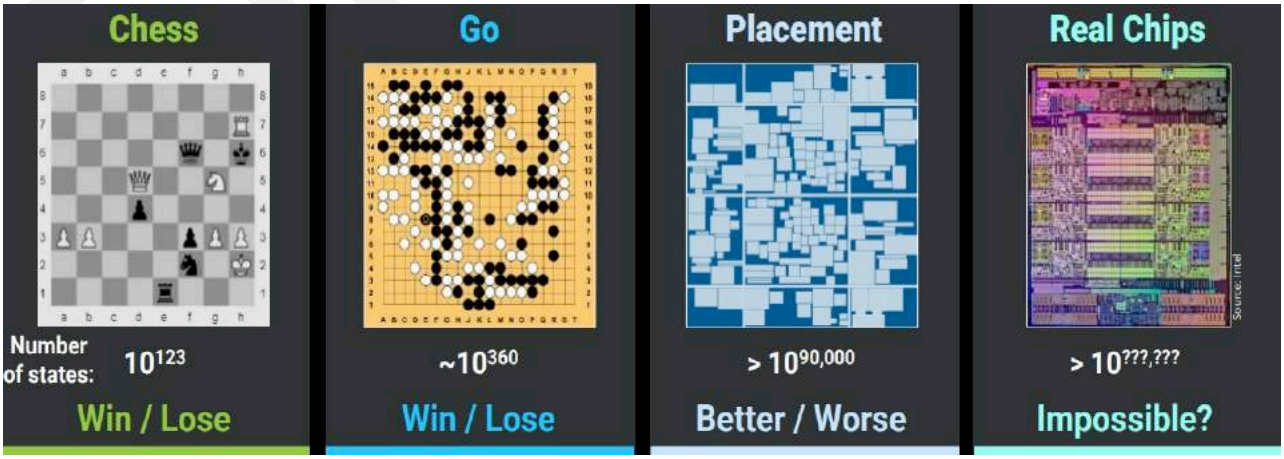


图.芯片设计的计算复杂度远高于围棋等领域 [4]

- **更快的设计迁移**。在 AI 的支持下,芯片设计团队可以更快地将其设计从一个工艺节点迁移到另一个节点。

AI 驱动的芯片设计也存在一些独特的挑战。作为一个相对较新的努力领域,能够将 AI 技术集成到不同的芯片设计解决方案中,需要对此有深入的理解。随着影响半导体行业的人才短缺,该行业将需要找到那些有优化 EDA 流程与 AI 技术专长的人员,以及增强 EDA 算法的计算平台的人员。

用于 AI 训练的数据集也非常有限。该行业所做的大部分工作都是专有的。怀疑论也是一个挑战,因为有些工程师怀疑机器如何可能会得出比他们更好的结果。

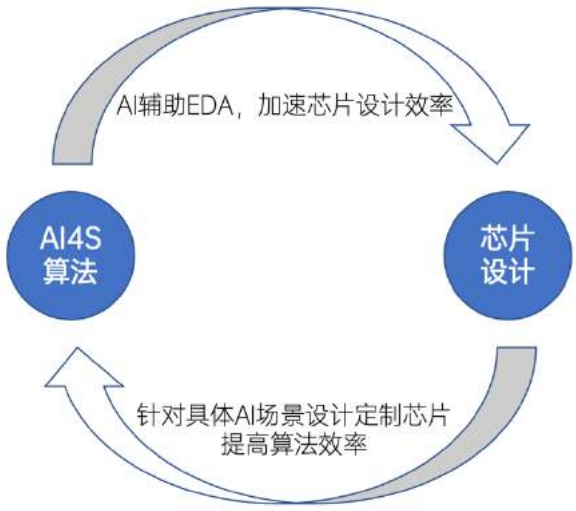
AI 技术正在成为 EDA 流程中日益普遍的一部分,它增强了从单片 SoC 到多粒子系统的各种内容的开发。它们将继续帮助交付更高质量的硅芯片,并缩短周转时间。芯片开发过程中还有许多其他步骤可以通过 AI 进行增强。虽然这个领域存在挑战,但挑战中也存在机遇。通过提高生产力和结果,AI 可以帮助填补人才短缺以及资深工程师离职时产生的知识差距留下的空白。此外,探索 AI 可以以其他方式增强芯片设计的机会也存在,包括 AI 芯片。

在具体实践中，Google 团队采用 AI 算法设计芯片，可以在 6 小时以内完成人工所需数月时间完成的芯片设计工作。此外，AI4S 已用于现有芯片的检测工序，通过深度学习、大量数据和机器视觉结合，实现高精度、高速度排除缺陷器件，缩短生产周期，提高器件良率，实现生产效率提高 20%以上。美国国防先进局 DARPA 在其雄心勃勃的 15 亿美元 "电子文艺复兴计划" 中也特别列出了 AI for Chip Design 作为核心课题 [1]

此外，芯片设计软件（EDA）全球龙头 Synopsys 也于 2020 年发布了其芯片设计的自主人工智能应用程序——DSO.ai。其客户英特尔通过将 DSO.ai 并入区块布局布线（PnR）流程中，有助于缩短芯片设计周期并实现最佳 PPA，减少了人工/ECO 收敛工作中的搅动，并及时向 RTL 所有者提供反馈以修复严重违反时间路径的问题 [2,3]

AI 算法的发展和芯片技术发展相互依存

AI 的算法可以帮助工程师更快的设计出 PPA 更好的芯片；反过来，利用 FPGA 等技术，工程师也可以设计针对特定场景的 AI 加速芯片，提高 AI 算法的效率。近日，芯片巨头 AMD 斥巨资收购 FPGA 龙头 Xilinx，引起业界震动，也显示了这一业界领袖对这一技术路线的重视。



Source:
[1] https://spectrum.ieee.org/darpas-15billion-remake-of-us-electronics-progress-report
[2] Synopsys, DSO.ai 两周年小考成绩揭晓：设计效率+5 倍，功耗 -25%, https://mp.weixin.qq.com/s/pfbBrNI-he2jhCbioBvyrg
[3] Synopsys https://www.synopsys.com/ai/what-is-ai-chip-design.html
[4] www.deccanherald.com/business/technology/could-ai-designed-chips-herald-a-revolution-in-compute-1022986.html

AI4S 实践（28）：《npj Computational Materials》收录湖南大学利用非冯架构加速 AI 分子动力学模拟的工作

自 1946 年发明至今，冯·诺依曼架构一直占据统治地位，是手机、台式机、笔记本、计算服务器、超级计算中心的底层基础架构。目前，需要运行 MD 计算（包括 DeePMD）时，使用冯·诺依曼架构计算机是几乎所有研究人员的唯一选择，这已成为一种固有范式。遗憾的是，冯·诺依曼架构中，计算单元（例如 CPU/GPU）和存储单元（例如内存）是互相独立的（即"存算分离"），导致计算总耗时和总功耗的绝大部分（>90%）消耗于存储单元、计算单元之间的频繁数据搬运，俗称"存储墙（memory wall）"和"功耗墙（power wall）"瓶颈。这严重制约了 MD 计算性能的提升。

近日，湖南大学的研究人员提出了基于新型非冯·诺依曼架构的分子动力学（non von-Neumann molecular dynamics, NVNMD），同时实现了 AIMD 级别的高精度、CMD 经典分子动力学级别的高速度。NVNMD 的核心计算模块中，存储单元和计算单元紧密融为一体（即"存算一体"），避免了频繁的数据搬运，极大缓解了 MD 计算中的"存储墙"和"功耗墙"瓶颈，将计算速度提升了约 1-2 个数量级；将计算功耗降低了 2-3 个数量级。[1]

在算法层面，为实现从传统 CPU/GPU 向新型非冯·诺依曼芯片架构的"范式转移"，湖南大学研究团队对 DeePMD 进行了若干修改和创新：（1）用整数等离散数据类型，取代了浮点数等连续数据类型；（2）用移位等运算，取代了乘法等运算；（3）用离散神经网络，取代了连续神经网络；（4）用逼近近似，取代了三角函数求解；等。这些修改和创新旨在使用新型非冯·诺依曼架构下有限的硬件资源，实现尽可能高的片内并行度，以实现高速 MD 计算。

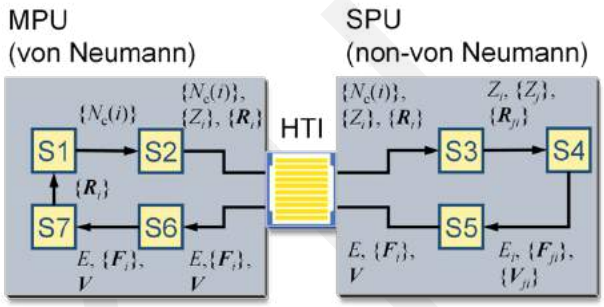


图. NVNMD 算法架构图

如下图 a 所示，使用 NVNMD 运行 GeTe 的分子动力学，复现了 GeTe 的整个相变过程（晶体相→液相→无定形相→晶体相）。 如下图 b 所示，使用 NVNMD 运行 $Li_{10}Ge_2SP_{12}$的分子动力学，得到的 500K 下的 Li 原子的均方位移与文献值基本一致。

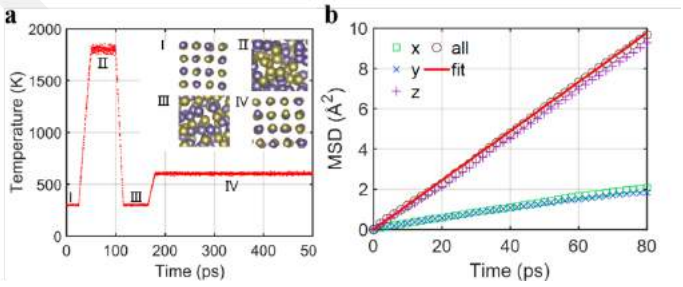


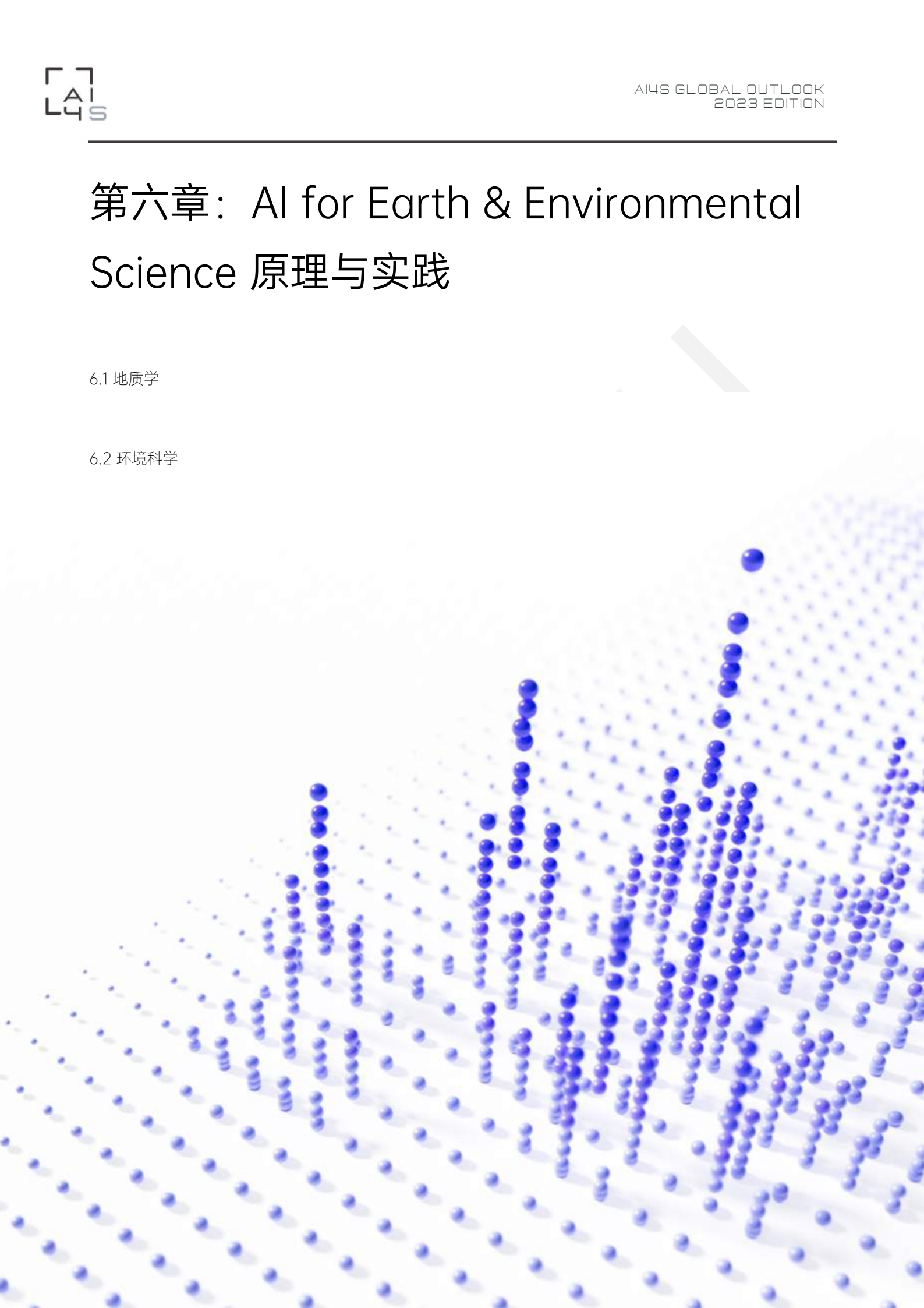
图.(a) GeTe 相变过程 (b) $Li_{10}Ge_2PS_{12}$中 Li 原子的均方位移

Source：
[1] Mo, P., Li, C., Zhao, D. et al. Accurate and efficient molecular dynamics based on machine learning and non von Neumann architecture. npj Comput Mater 8, 107 (2022). https://doi.org/10.1038/s41524-022-00773-

第六章：AI for Earth & Environmental Science 原理与实践

6.1 地质学

6.2 环境科学



6.1 地质学

地球科学，是行星科学的专门分支。通常会以物理、地理、地质、气象、数学、化学、生物的角度研究。它包括各种生产所需的矿产资源；建筑所需的砂、石、水泥原料；海洋物产；气候；天体的运行，时时刻刻影响着我们的生产生活。地球科学的范围很广，涵盖地质学、海洋学、气象学和天文学等领域[1]。

其中，地质学是对地球的起源探讨压力与时间、历史与结构进行研究的学科。主要研究地球的物质组成、内部构造、外部特征、各圈层间的相互作用和演变历史[2]。由于目前实验手段的限制，研究人员无法自由的对地球内部形态和性质进行有效的观测，因此计算建模手段成为了重要的研究手段。

在地球科学的基础研究中，AI 与地质学的结合更是广泛且深入。例如，通过分析和模拟地质年代学数据，AI 可以帮助我们重建地球的地质历史和演变过程；通过学习和理解岩石圈、大气层和水圈等地球系统的相互作用，AI 可以预测和模拟地球的未来气候和环境变化；通过研究地壳的构造、组成和动力学过程，AI 可以帮助我们预测和防止地质灾害，如滑坡、地裂缝等。

在地震预测中，AI 不仅可以通过分析海量的地震数据进行预警，更重要的是，它可以借助于地震学的基础理论，例如弹性波传播、地壳应力积累与释放、板块构造运动等，模拟地震的生成和传播过程，进而更准确地预测地震的可能性和规模。

在矿产勘探方面，AI 可以通过分析和学习地质、地球物理、遥感等数据，预测矿产的分布和丰度。这不仅可以提高矿产勘探的效率和精度，也可以减少环境的破坏和影响。而且，随着近年来深海和太空矿产勘探的兴起，AI 在这些极端环境中的应用将更加重要和必要。此外，AI 也能够结合地质学的基本理论，例如岩石学、矿物学、地球物理学等，以理解矿产的形成、分布和探测方法。AI 不仅可以通过学习和分析各类地质、地球物理和遥感数据预测矿产的可能分布，而且能根据矿物的地质-地球物理特性，识别出有矿化潜力的地层和地质构造，进一步提高矿产勘查的准确性和效率。美国 KoBold Metals 公司利用人工智能指导矿产勘探，在最近一轮融资中筹集了 2 亿美元，公司估值超过 10 亿美元。KoBold 将清洁能源和人工智能这两大热门投资领域结合，改进了对铜、锂等关键电池金属的勘探效率。它利用机器学习分析地质数据，识别出高潜力的勘探区域。如今 BHP 等矿产巨头也与 KoBold 合作，认识到剩余矿石储量越来越深，越难以用传统技术发现。KoBold 旨在用科学、技术驱动的方法转变矿产勘探。

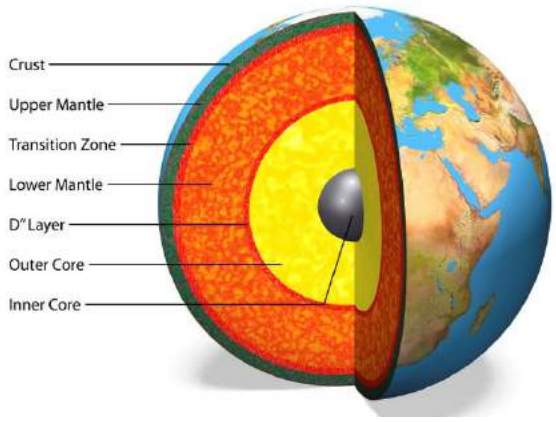
此外，AI 还可以与地下水学、火山学、土壤学等地质学分支学科深度结合，从而实现更多的可能性。例如，AI 可以通过学习和理解地下水的流动规律和地层过滤特性，预测和管理地下水资源；AI 可以通过分析和模拟火山岩浆的运动和地壳应力状态，预警火山喷发；AI 可以根据土壤的物理化学性质和环境因素，预测土壤侵蚀和沙尘暴等地质灾害。

AI 在地质学领域的应用和可能性是广阔且多元的。通过深度学习、数据和基本原理的融合，AI for Science 不仅可以帮助我们更好地理解和保护地球，也可以帮助我们有效地开发和利用地球的资源。而这些只是冰山一角，随着 AI 技术的进步，其在地质学领域的应用将更加深入和广泛。

6.1.1 地球物理学（Geophysics）

以地球热演化为例，硅酸盐熔体控制了早期岩浆海洋的动态，随后影响了岩石学和地球动力学过程，如晶体沉降和地核-地幔分异。在今天的地球上，硅酸盐熔体暂时存在于地壳和上地幔中，并被认为存在于像核幔边界(CMB)那样的深度。因此研究硅酸盐熔体（及液体）会让人们对地球核心及地球演化过程有更全面、深刻的了解。其中，热导率、粘度以及电导率是硅酸盐熔体核心关注点。第一性原理是强大的辅助研究方法，但由于准确性、计算体系及成本的限制，未能广泛应用。

路易斯安那州立大学的研究人员利用 AI4S 加速分子动力学模拟，模拟了压力增大过程中 $MgSiO_3$ 熔体的粘度变化，其发现低压下，熔体的粘度随着压力的增加（最高到 ~ 6GPa）而降低，之后才开始随着进一步压缩而增加。同时该组研究人员也对硅酸盐熔体的导电率进行了模拟，发现随着压力的增加，导电率先是增加，而后减少，该现象有助于解释从遥测中得到的地球浅层电导率曲线的陡峭上升。[3]



Picture credit: Simone Brandt / Getty Images

来自加州大学洛杉矶分校的研究人员则利用深度势能工具对硅酸盐熔体（$MgSiO_3$）的热导率进行了模拟。发现热导率随着压力增大而增加，在等温加热的时候几乎保持不变。基于此，研究人员发现热导率对压力的依赖性来自于压缩时体积模量的增加，而微弱的温度依赖性则是来自于结构无序导致的声子平均自由路径的饱和。得到的结论与经典的 Debye 模型有很好的契合，对于硅酸盐溶液热导率的研究及模拟对深入理解地幔有着重要意义。[4]

展望未来，AI4S 有机会实现在第一性原理精度下，进行整个地幔的压力-温度模拟，将地球物理学带入新时代。

Source:

[1] Wikipedia, https://zh.wikipedia.org/wiki/%E5%9C%B0%E7%90%83%E7%A7%91%E5%AD%A6

[2] Wikipedia, https://zh.wikipedia.org/wiki/%E5%9C%B0%E8%B4%A8%E5%AD%A6

[3]. Haiyang Luo, Bijaya B. Karki, Dipta B. Ghosh, Huiming Bao, Anomalous Behavior of Viscosity and Electrical Conductivity of MgSiO3 Melt at Mantle Conditions, Volume 48, Issue 13, 16 July 2021

[4]. Jie Deng and Lars Stixrude, Thermal Conductivity of Silicate Liquid Determined by Machine Learning Potentials, Volume 48, Issue 17, 8 September 2021

AI4S 实践（29）：《自然•通讯》报道 AI 赋能基础科学研究地球内核对地震的影响

在地震研究中，地球内部的物质运动和其性质对于解释地震波的传播特性至关重要。在地震波通过地球内部时，它们会受到各种物理过程的影响，包括物质的密度、弹性性质、温度和化学成分等。其中，离子扩散是一种重要的物理过程，影响地球内部的电导率和声速，进而影响地震波的传播。

2023 年，一篇发表于《自然•通讯》的论文揭示了地球内核的复杂结构以及这些结构是如何影响地震等自然现象的[1]。地球的内核被认为是异质性和各向异性的，这种特性在过去一直难以解释。然而，这项研究通过探索内核中的超离子态铁-氢合金的特性，提供了一种新的理解角度。在这种状态下，氢像液体一样在铁的晶格中高速扩散，这种特性可能解释了为何地球内核的性质会影响地震的产生和行进。这篇文章为我们理解地球内部最深处的工作机制提供了新的视角，为地震学研究带来新的启示。

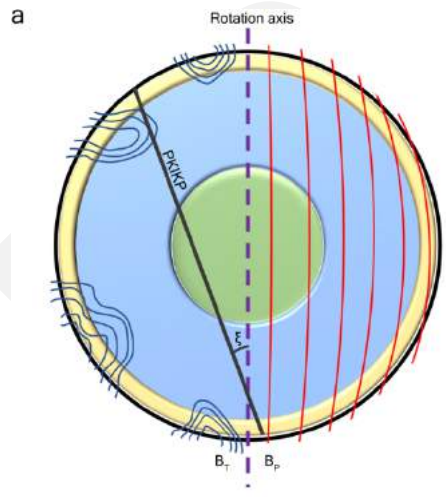


图. 地幔中的内核地磁场和深度相关各向异性纹理变化示意图

文章由上海高压科学与技术先进研究中心毛河光院士领衔，作者研究了超离子态下的铁-氢合金。在这种状态下，氢离子在六方密堆积（hcp）铁晶格中的扩散是各向异性的，沿 c 轴的扩散障碍能最低。当存在外部电场时，c 轴指向电场方向的铁-氢晶格的排列是能量最优的。由于这种效应，由北-南向地磁场扩散到内核的铁-氢合金的 c 轴并行于赤道平面。这种由地磁场驱动的排列结构产生了明显的地震各向异性，这解释了内核中的地震速度各向异性，表明内核结构和地磁场之间存在强耦合。

首先，论文通过密度泛函理论(DFT)计算了弹性属性和氢离子传输属性。接着，采用 ab initio 分子动力学(AIMD)模拟在固定的温度、压力和体积下，计算了 hcp-Fe-H 合金的单元参数。AIMD 模拟后，随机添加了氢原子以构建 Fe64H16 结构。然后，通过解决应力—应变关系，计算了弹性常数。之后，计算了体积模量和剪切模量，并使用 Voigt 平均方案计算了压缩波速度、剪切波速度和体声速度。

之后，研究人员使用了神经网络势能（NNP）的方法，使用 DeePMD-kit 软件包进行训练，模拟了在 360 GPa 和 6000 K 下的超离子 FeH0.25 合金的行为。通过神经网络模型，我们可以更深入地研究复杂的物理过程，如氢离子在超离子 FeH0.25 合金中的扩散行为。

然后，研究人员进行了非平衡分子动力学模拟（NEMD），使用经过机器学习 NNP 训练的模型，进一步研究了 FeH0.25 在外部电场下的各向异性扩散行为。由于 AI for Science 先进方法的存在，我们得以在相当复杂的条件下（如高温、高压和外部电场的存在）模拟和理解物质的行为。

在另一篇研究中，意大利 The Abdus Salam International Centre for Theoretical Physics 研究

中心的研究员同样使用 AI for Science 方法对地球的内核进行研究，特别是内核中六方最密堆积(hcp)铁的性质 [2]。这个研究采用了深度学习驱动的分子动力学模拟，以解释地震观测数据并理解地球内核的物理属性。

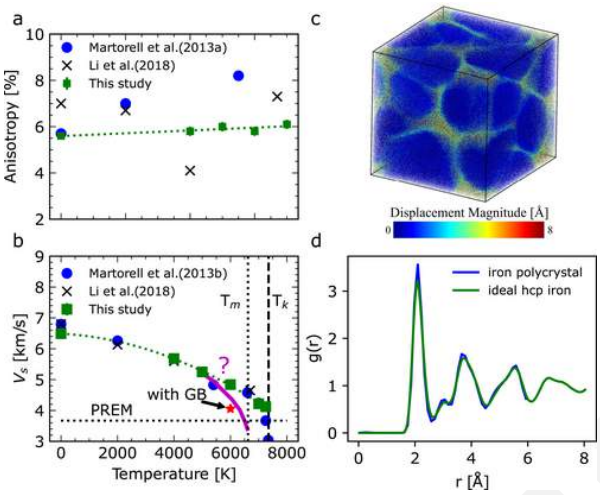


图. 压缩波各向异性（a）和地球内核条件下 hcp 铁的剪切波速度（b），与之前的研究（Li 等人，2018; Martorell，Brodholt 等人，2013; Martorell，Vočadlo 等人，2013）以及初步参考地球模型（PREM）（Dziewonski 和 Anderson，1981）进行比较。红色星号表示粘性晶界（GB）对剪切波速度的影响。品红色线表示在 Martorell，Vočadlo 等人（2013）中观察到的预熔效应。在 a 和 b 中，绿色虚线是通过线性拟合和二阶多项式拟合得到的眼观引导线。c 显示了用剪切波速度计算的多晶铁样品。立方体边长约为 200 Å，模拟盒中包含 1,164,419 个原子。我们根据它们的运动性对每个原子进行着色。高移动性的原子总是位于晶界。即使存在这些扩展缺陷，与理想 hcp 铁的配对分布函数进行比较，hcp 铁晶体的结构仍然是六方最密堆积，如（d）所示。

具体来说，这个研究工作的几个关键步骤包括：

建立深度学习势能模型：这篇文章使用了迭代策略来生成训练数据集，并采用第一原理计算方法来确定这些配置的电子结构，以获取用于训练深度学习势能模型的力量、能量和压力数据。

计算 hcp 铁的粘度：该论文主要关注位错蠕变对 hcp 铁塑性的贡献。研究者使用由 MD 模拟直接估计的空位扩散率来描述应变率。

评估压缩波各向异性：通过使用深度学习的原子势，研究者在 MD 模拟中更有效地增强了时间和长度的规模，从而得出了 hcp 铁弹性各向异性的更可靠估计。

分析剪切波速度：研究者使用 Voigt 平均方案来估计多晶体骨架的剪切模量和剪切波速度，结果显示，如果不考虑光元素的存在，该值高于地球物理观测值。此外，作者们发现，具有黏性晶界的 hcp 铁的多晶体可能会导致剪切速度降低。

计算粘度：研究者利用获得的自扩散率和剪切体积模量来评估 hcp 铁在 360 GPa 下的粘度。

这篇论文主要的贡献在于，它采用了深度学习驱动的分子动力学模拟，使得研究者可以在更大的时间和长度尺度上研究 hcp 铁的物理性质。这种方法在解释地震观测数据、理解地球内核的性质，以及预测地震行为等方面可能具有很大的价值。

Source:

[1] Sun, S., He, Y., Yang, J. *et al.* Superionic effect and anisotropic texture in Earth's inner core driven by geomagnetic field. *Nat Commun* **14**, 1656 (2023).

[2] Li, Z., & Scandolo, S. (2022). Elasticity and viscosity of hcp iron at Earth's inner core conditions from machine learning-based large-scale atomistic simulations. *Geophysical Research Letters*, 49, e2022GL101161. https://doi.org/10.1029/2022GL101161

6.1.2 同位素地球化学（Isotope geochemistry）

对地球科学的研究，另一个重要的方向是对早期地球历史和物质演化的研究，尤其是围绕惰性气体同位素测年原理的相关课题。惰性气体在化学是惰性的，这意味着在地球历史上，它们的成分没有被化学或生物过程改变，只有扩散、吸附或离子植入等物理过程才能导致显著的元素和同位素分馏。因此对这部分稀有气体及其同位素研究对于了解地球早期历史、地幔-地壳的物质演化、地球的脱气和大气圈的形成、自然界核类的和转化过程及同位素年龄测定都有重要意义。[1]

为研究这一课题，路易斯安那州立大学的研究人员将视线聚焦于 He 在钠长石和玄武岩（典型地壳岩石）中的同位素问题。其中，He 的流失主要来自岩浆脱气过程。由于这一过程的极端反应条件（高温低压），传统的实验和计算手段均无法满足精度要求。因此研究人员利用 AI4S 研究地幔钠长石中扩散 He 同位素分馏，并在零压力附近模拟玄武岩在 3000K、2200K 和 1700K 的熔体，得到的数据与经典 MD 计算结果及实验数据较为一致。他们发现扩散因子值可用于定量估计岩浆脱气过程中的 He 损失，并追溯岩浆源区域的 He 浓度和同位素组成。[2]

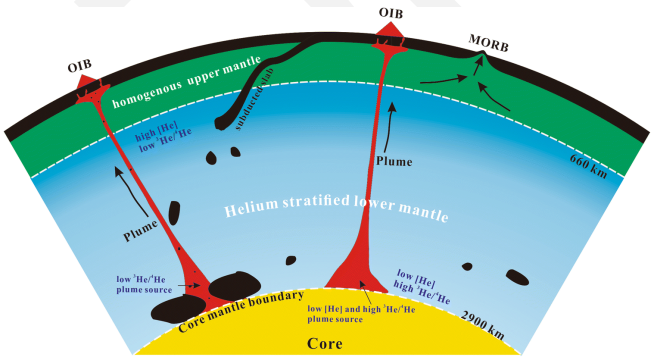


图. He 同位素在地幔/岩浆脱气过程中的示意图[5]

除了惰性气体，科学家也关注锂同位素相关的问题。

锂(Li)是自然界最轻的金属元素, 有 2 个天然稳定同位素 ^{6}Li 和 ^{7}Li, 丰度分别为 7.5%和 92.5%。高达 ~ 15%的相对质量差使得 Li 在地质过程中容易产生较大的同位素分馏。在部分熔融过程中, Li 的中等不相容的地球化学特征导致其在壳-幔体系中发生分异。同时, Li 也是强烈的流体活动性元素, 且优先在液体相中富集。已有研究表明地球上不同储库的 δ^{7}Li 值差异可高达 80‰, 这些特性使得 Li 同位素体系被广泛应用于地壳物质再循环和地幔交代作用的示踪。[3]但目前对 Li 的研究主要聚焦于矿物中，对熔融态或者玻璃态研究较少。尤其是目前对理解扩散状况和机制的重要因素仍然不清楚。

路易斯安那州大学的研究人员借助 AI4S 模拟了钠长石中 Li 及其同位素的扩散，得到的扩散率与实验结果有了很好的匹配，以此为基础，模拟了玄武岩熔体中 Li 的扩散，来评估扩散因子的作用。该结果表明，硅酸盐熔体中的锂同位素扩散在很大程度上取决于熔体成分，扩散因子在玄武岩熔体中的温度依赖性比钠长石中的温度依赖性强。温度和成分对可以从离子孔隙度以及 Li 扩散与硅酸盐熔体

网络流动性之间的耦合关系来定性解释，而水的存在往往会削弱扩散因子的温度依赖性。[4]

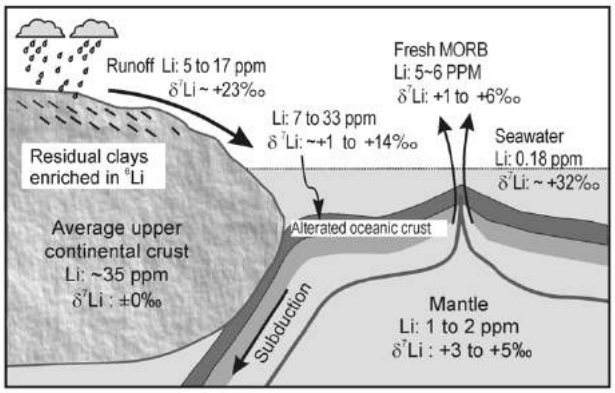


图. Li 同位素在地球活动中的"代谢"过程 [6]

Source：

[1]. 《地幔脱气作用和大气圈惰性气体形成与演变》，朱岳年，许多，张友学，1998

[2]. Luo, H., Karki, B.B., Ghosh, D.B., Bao, H. (2021) Diffusional fractionation of helium isotopes in silicate melts. Geochem. Persp. Let. 19, 19–22.

[3]. SU Ben-xun. 2017. Applications of Li Isotopes in Mantle Geochemistry. Kuangwu Yanshi Diqiuhuaxue Tongbao, 36(1): 6-13

[4]. Haiyang Luo, Bijaya B. Karki, Dipta B. Ghosh, Huiming Bao, Deep neural network potentials for diffusional lithium isotope fractionation in silicate melts, Geochimica et Cosmochimica Acta, Volume 303, 2021, Pages 38-50, ISSN 0016-7037, https://doi.org/10.1016/j.gca.2021.03.031.

[5] Liu, Y. Chemical structure of the Earth's mantle defined by fast diffusion elements like helium. Acta Geochim 39, 1–3 (2020). https://doi.org/10.1007/s11631-019-00381-7

[6] Tang, Yan-Jie, Hong-fu Zhang and Ji-Feng Ying. "Review of the Lithium Isotope System as a Geochemical Tracer." International Geology Review 49 (2007): 374 - 388.

6.2 环境科学

6.2.1 天气预测

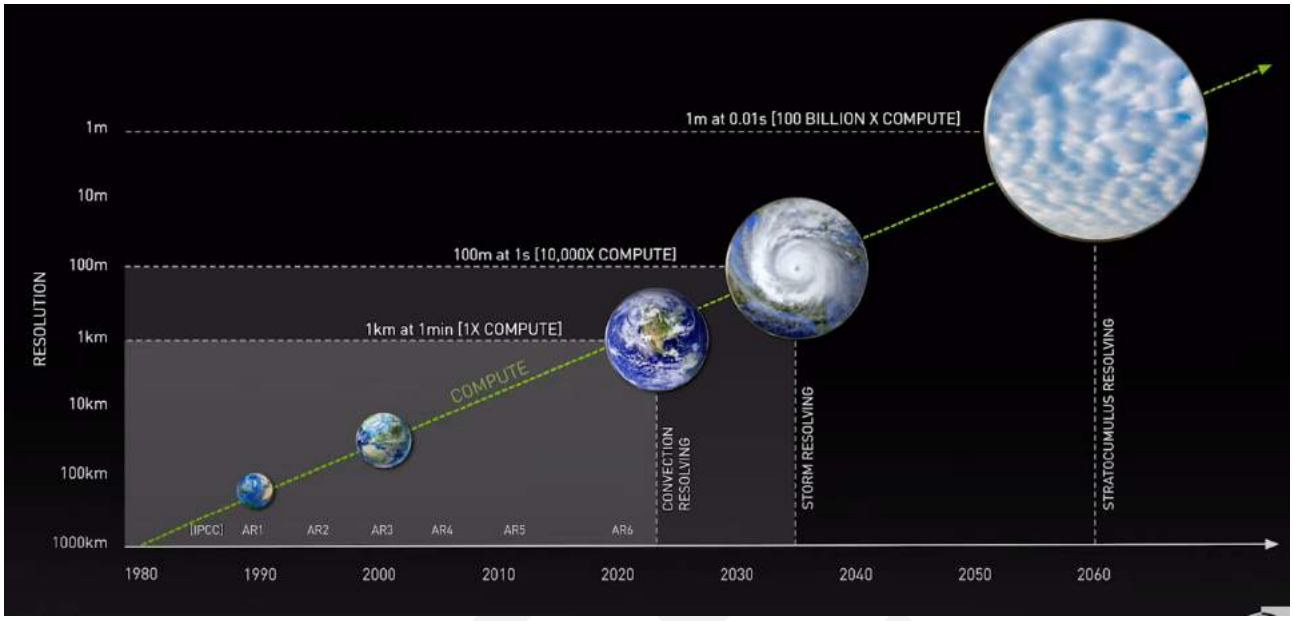


图. 人类对高精度气象模拟的需求不断升级，对新算力和新算法的要求不断提高 (Source: Nvidia [4])

现代天气预报来自于各地的超级计算机的预测，通过高数百万次计算解出对温度、风、降雨等天气事件的预测的方程式。预测准确性和速度两者兼顾是十分重要的，这对传统计算方式提出了挑战。在传统算法体系下，即使世界最强大的计算机也无法很好的预测天气。

纵观历史，天气预测对国家和个人均具有重要的价值。“降水临近预报”（precipitation nowcasting）对交通、农业、体育等众多依赖天气动态作决策的场景有着至关重要的意义。为了保障生产生活的有序进行，降水临近预报必须尽可能及时且准确地预测多维度的降雨信息，包括降雨时间、降雨量、降雨位置、降雨强度等等。

此前的预测手段主要采用数值天气预报 (NWP) 系统。该系统主要是通过模拟大气的耦合物理方程，估计降雨概率，从而生成多个真实的降水预报结果。然而，在提前 0 至 2 小时的降水临近预报，NWP 却无法胜任，因为 NWP 的模型基于第一性物理原理，计算缓慢，难以保证计算结果的时效性。

后来，人们又使用基于雷达进行观测的降雨预测方法。雷达概率预测系统与 NWP 的拟合方式相近，但新增了根据对流方程和雷达源项模拟降水的方法，通过光流估计，能得到较准确的概率预测。

深度学习崛起后，研究人员开始用深度学习模型开发降雨预测系统，希望摆脱对对流方程的依赖，在兼顾物理准确性的情况下将预测效率提高至“实时”级别。深度学习主要通过在大型雷达观测资料库上训练模型，而不是依赖内在的物理设置，可以更好地模拟传统方法难以预测的非线性降水现象，比如对流启动与强降水。[3]

AI4S 实践（30）：从 DeepMind 到华为、AI for Science 不断突破气象预测

2021 年，《Nature》收录 DeepMind 新成果。研究人员用深度生成模型取代了大气物理方程，实现了 200 万平方公里的大气层的物理仿真。研究者称这是目前最先进的可操作性降雨预测技术，使用了深度学习加雷达对未来的降雨率进行直接预测，

图. DeepMind 基于雷达的 AI4S 模型，通过过去 20 分钟的数据预测未来 90 分钟的天气

优于基于雷达的风估计来确定大致降雨范围的传统方法，且不会受到太多物理条件的约束。[1]

"降水临近预报"对农业生产、灾备管理、空中交通管制与海事服务等场景有巨大的商业和社会价值。为了确保这些行业的正确运营，降水临近预报必须尽可能准确地预测多维度的降雨信息，包括降雨时间、降雨量、降雨位置、降雨强度等等。

此前，降雨预测主要采用数值天气预报 (NWP) 系统。由于 NWP 计算量巨大，启动时间都要超过 2 小时，且难以规避非高斯数据的同化；因此 NWP 难以满足"降水临近预报"的要求。

DeepMind 用深度生成模型（DGM）开发了一种基于观察的临近预报概率方法。从本质上讲，DGM 是一种统计模型，可以学习数据的概率分布，并允许从学习到的分布中生成样本。由于生成模型基本上是概率性的，因此它们能够在给定的历史雷达上从未来雷达的条件分布中模拟大量样本，生成与集合方法类似的预测集合。

与传统方法相比，DeepMind 的方法在统计上有显著改善。他们联合英国国家气象局（Met Office）的 50 多位气象学家进行了认知任务评估，表明了：与目前广泛使用的其他降雨临近预报方法相比，这些气象专家在 89% 的案例中会首先选择 DGM。

相关研究中，华盛顿大学和微软合作利用 AI4S 进行天气预测，为天气预测提出新路径。该团队借助机器模式识别功能，对过去 40 年的天气数据进行

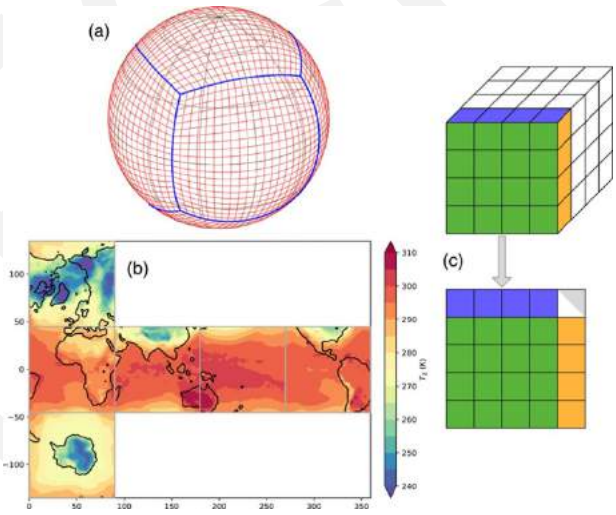


图. 研究人员将地球表面划分成一个立方体的网格（左上），然后将六面平展成一个二维形状（左下）[2]

学习，进而产生新的天气预测模型，即使相比全球最顶尖的传统预测模型仍有不准确的地方，但其面对与传统预测模型进行相同数量的预测时，可以节省 7000 多倍的算力，实现更快获取结果。同时，这种高效率也扩大了可预测天气的范围，实现对一个足够大系统的预测。随着对模型的不断训练和完善，未来对天气的预测有望变得更加准确、快速。AI4S 为天气预带来了颠覆性改变。[2]

此外，伯克利劳伦斯国家实验室，Caltech 和英伟达的研究员基于 Nvidia 的 AI 物理引擎

（Modulus）对气象数据进行研究，以气象台观察数据作为训练集，经过 4h，128 张 A100 卡的模型训练就能实现对 30km 级分辨率的极端气候预测（如飓风），比传统计算方法加速了 100,000 倍 [4]

2023 年，华为气象模型"盘古"登上《Nature》[5] 华为云盘古大模型研发团队发现，AI 气象预报模型的精度不足主要有两个原因：第一，原有的 AI 气象预报模型都是基于 2D 神经网络，无法很好地处理不均匀的 3D 气象数据；第二，AI 方法缺少数学物理机理约束，因此在迭代的过程中会不断积累迭代误差。为此，团队创造性地提出了适应地球坐标系统的三维神经网络（3D Earth-Specific Transformer）来处理复杂的不均匀 3D 气象数据，并且使用层次化时域聚合策略来减少预报迭代次数，从而减少迭代误差。通过在 43 年的全球天气数据上训练深度神经网络，盘古气象大模型在精度和速度方面超越传统数值预测方法。

同在 2023 年，一项名为"高技巧极端降水临近预报大模型"（Skilful Nowcasting of Extreme Precipitation with NowcastNet）为题发表在《自然》（Nature）上，同时被《自然·新闻和观点》以"The Outlook for AI Weather Prediction"为题做了报道。目前，该大模型已经在国家气象中心短临预报业务平台（SWAN 3.0）部署上线，将为全国极端降水天气短临预报业务提供支撑。

该项目由清华大学软件学院王建民、龙明盛团队与国家气象中心、国家气象信息中心 完成。项目提出了临近预报大模型 NowcastNet，将数据驱动与物理驱动两大科学范式紧密结合，显著提高了公里尺度下 0 ~ 3 小时极端降水的预报能力，在全国 62 位气象预报专家的过程检验中大幅领先国际上的同类方法。

中尺度演变网络是一种通过对平流等物理性质显著的中尺度演变过程进行建模的技术。这个模型是基于质量守恒定律（物质连续性方程）设计的，使用可微的神经演变算子对降水场的未来演变进行模拟。神经演变算子在当前降水场的基础上，预测未来的中尺度降水场，结合了运动场的输送和强度残差场的加成，生成中尺度雷达场的预测。这种方法易于结合基于梯度的优化方法，因为它具有良好的可微性。中尺度演变网络还可以根据过去的雷达场直接预测未来的运动场和强度残差，以实现十公里尺度降水过程的端到端模拟。预测过程中的累积误差可以通过反向传播进行优化。

在中尺度演变网络的基础上，研究团队开发了一种多尺度条件生成机制，能够在中尺度预测结果上补充对流尺度的演变细节，以模拟对流、生消等混沌效应更显著的公里尺度降水过程。这种机制考虑了湍流效应引起的对流尺度混沌，并通过对抗学习和空间自适应归一化技术，输入随机噪声模拟系统混沌，并保证新的预测在中尺度上与原预测保持一致，以保持物理机理的约束。

NowcastNet 是通过这两阶段网络的设计实现的，能在预测中同时考虑物理性质显著的平流运动和混沌效应更强的对流演变。这种设计实现了降水预测的位置、强度和细节的同步提升，解决了以往模型预测位置和强度不准，或预报模糊、弥散的问题 (案例见下图)。

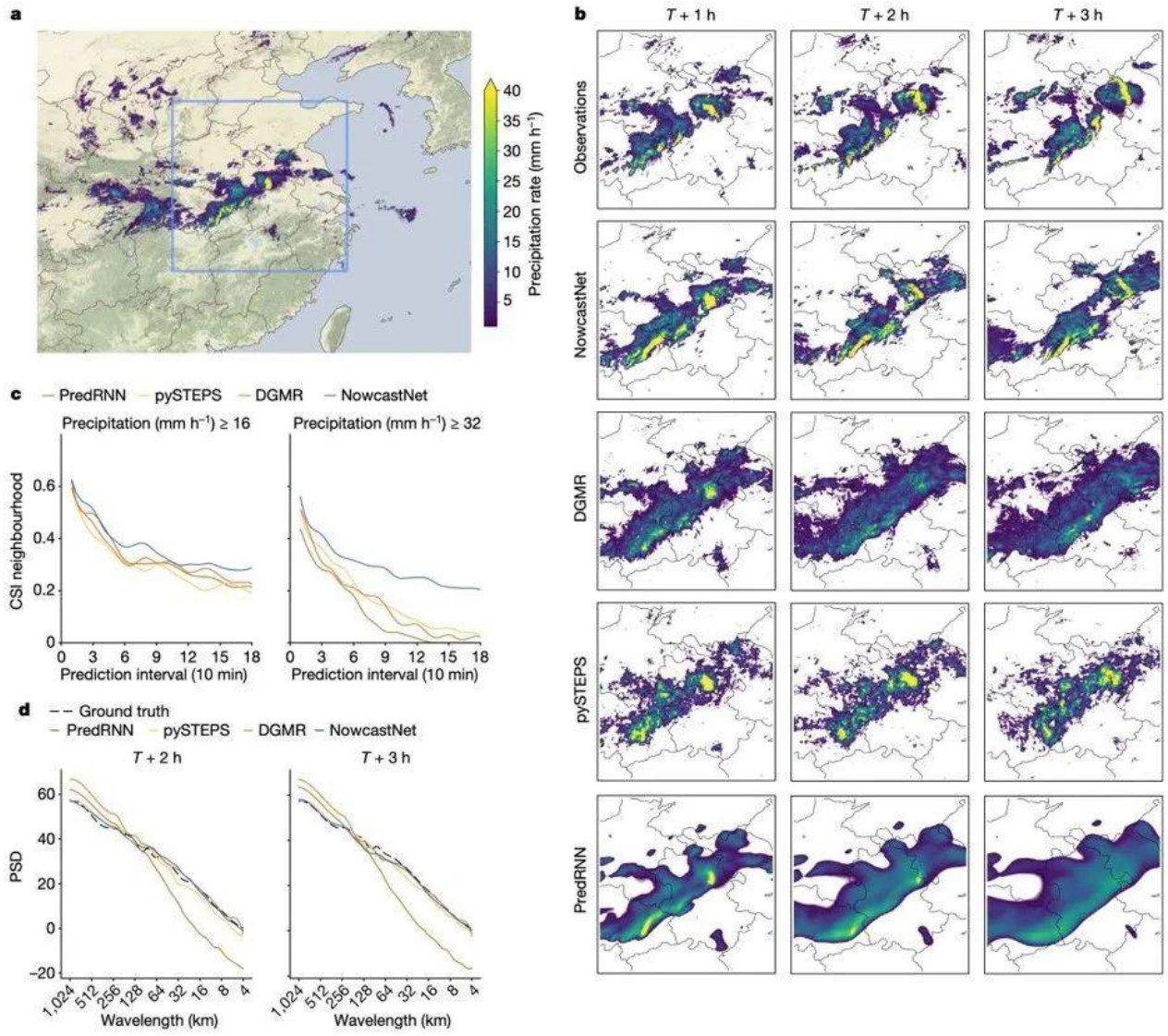


图. 以中美两国的典型极端天气过程为例：2021 年 5 月 14 日 23 时 40 分，中国江淮地区出现强降水过程，湖北、安徽等多个地区发布了暴雨红色预警，NowcastNet 可以准确预测出三个强降水超级单体的变化过程。

source：

[1] Ravuri, S., Lenc, K., Willson, M. et al. Skilful precipitation nowcasting using deep generative models of radar. Nature 597, 672–677 (2021).

[2] Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. Journal of Advances in Modeling Earth Systems, 12, e2020MS002109.

[3] AI 科技评论, cloud.tencent.com/developer/article/1887058

[4] Nvidia GTC 2021, youtube.com/watch?v=heshd3L6Kdk

[5] Bi, K., Xie, L., Zhang, H. et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538 (2023).

[6] Zhang, Y., Long, M., Chen, K. et al. Skilful nowcasting of extreme precipitation with NowcastNet. Nature 619, 526–532 (2023). https://doi.org/10.1038/s41586-023-06184-4

AI4S 实践（31）：《自然•通讯》报道 AI4S 助力宏观气象现象的微观机理研究

大气气溶胶核化是一个令科学家着迷并长期致力于研究的主题。然而这一研究却仍然充满挑战与未解之谜。这一过程涉及到大气中微小颗粒物质的形成，其结果是形成了大约占全球一半的云凝结核。这些微小的颗粒，在影响全球气候变化、天气模式、甚至人类健康方面发挥了重要作用。然而，详细的核化机制依然不为人知。

在一篇原创研究中，中国科学院安徽光学精密机械研究所大气物理化学实验室的科学家们尝试深入了解这个复杂过程。他们利用深度神经网络、反应力场，开发了一个全新的模拟流程，可以以接近完全基于第一原理（ab initio）的精度模拟气溶胶核化过程[1]。这种新的技术将对我们理解全球气候系统产生深远影响，特别是在我们面临日益严重的气候变化挑战的当下。此外，该研究还发现，过去对酸碱形成率的报告可能被大大低估，特别是在污染环境中。这意味着，我们需要重新审视在多种环境中观察到的酸碱核化，以便我们能够更准确地理解和预测大气污染的影响。

具体来说，工作流程的第一步是通过元动力学（Metadynamics）与主动学习技术（Active Learning Techniques）准备全面的数据集。元动力学是一种增强采样方法，能够有效地探索和采样自由能面，而主动学习则是一种迭代过程，通过反复选择信息丰富的数据来进行模型训练，这样可以加速数据收集过程，使其更为高效。然后，基于这个数据集，训练一个深度神经网络的力场（DNN-FF），使得可以进行强大的核化分子动力学（Nucleation MD）模拟。

接下来，通过使用泊松分布，可以从这些模拟中推导出碰撞率常数（Collision Rate Constants）。泊松分布是一种统计分布，它可以描述在给定的时间或空间区间内发生的事件的概率。最后，将静态量子化学热力学基础的蒸发率（Evaporation Rates）与基于 DNN-FF 的 MD 导出的碰撞率常数相结合，导入一个簇动力学模型（Cluster Dynamics Model），以提供基于第一原理的动力学，用于模拟大气气溶胶核化。

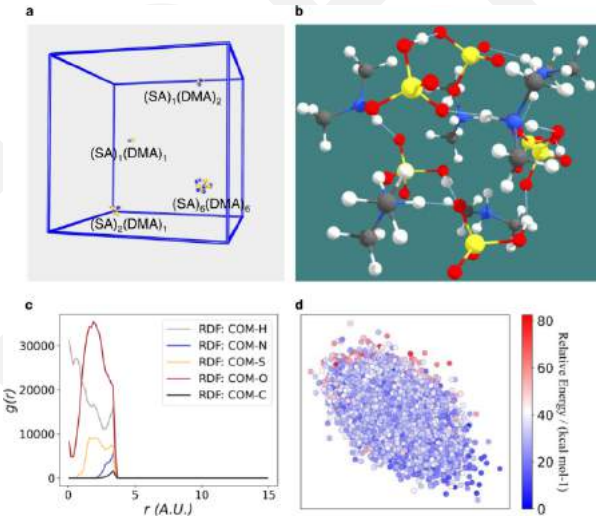


图. a 展示了在 1 纳秒处的分子动力学快照。青色、白色、红色、蓝色和黄色的圆圈分别代表 C（碳）、H（氢）、O（氧）、N（氮）和 S（硫）原子。为了清晰起见，N 和 S 原子的半径被放大了三倍。b 是轨迹中最稳定的异构体。c 是簇质心（COM）与五种元素之间的径向分布函数（RDF）。d 是利用全局原子位置平滑重叠（SOAP）核的异构体的核主成分分析（KPCA）映射。源数据作为源数据文件提供。

这种方法允许科学家以前所未有的精度来模拟和理解大气气溶胶核化过程，这是一个影响全球气候和天气模式的关键过程。

Source: [1] Jiang, S., Liu, YR., Huang, T. et al. Towards fully ab initio simulation of atmospheric aerosol nucleation. Nat Commun 13, 6067 (2022).

6.2.2 污染治理与碳中和

从污染和碳排放的源头出发，利用 AI4S 方法对相关反应机理进行深入研究并针对其环境影响进行定向优化的可能性，前文已于 4.2 进行了详述，此处不再重复。值得注意的是，对污染的治理很大程度与催化相关。

比如汽车尾气的核心治理技术即"三相催化剂"。使用 AI for Science 方法，科学家可以在原子尺度研究制作工艺、纳米尺度设计等环节如何影响催化效果，并基于第一性认知对下一代催化工业进行引导和助力。

2022 年底，国际顶级学术期刊 JACS (美国化学会期刊) 报道了清华大学 Li Jun 等关于贵金属纳米颗粒在 Ce Si 氧化物基底(尾气催化剂常用设置) 烧结过程中活性的变化[5]。研究者们用原子级别的透射电镜以及计算模型来观察金（Au）纳米颗粒在不同载体上的动态相互作用。研究发现，与在非晶态硅石（amorphous silica）上的金纳米颗粒相比，附着在氧化铈（ceria）上的金纳米颗粒的接触角度较小，且明显的不易移动，特别是在表面步骤处。

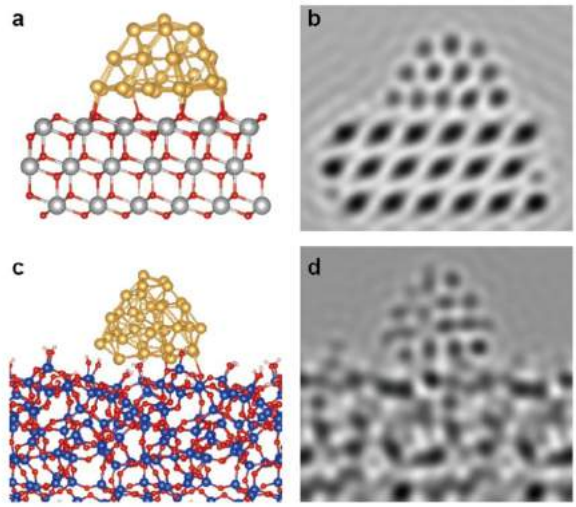


图. DeePMD 对微观形貌的仿真与 TEM 电镜观测高度一致[5]

这种移动性的差异被归因于"金属亲和性"。金属亲和性决定了金属与载体材料之间的交互，类似于水分子对于特定表面的亲水性。通过第一性原理分子动力学（AIMD）和基于机器学习的深度势能分子动力学模拟，研究者们直接捕捉到了金在硅石表面的聚结过程，以及金在氧化铈上的强烈定位。

这篇文章的结论是，催化剂载体的金属亲和性是决定其烧结和失活关键因素，无论载体是活性的还是惰性的。

2023 年，华东理工研究员等使用相同思路对 Pd 贵金属催化一氧化碳(煤气)至无毒的二氧化碳的微观机理进行探究。通过模拟，研究人员确定了催化剂在动力学稳态下可能的活性相，并揭示了它们的共同特性。他们发现，在 CO/O2 反应混合物存在时，Pd(111)表面形成了一个动态稳定的部分氧化非化学计量的钯氧化物(PdO0.44 层)，其中氧原子插入到 Pd(111)亚层中，驱动表面氧化物的形成。有趣的是，第一性原理微观动力学分析表明，这种自发演变的 PdO0.44 表面在稳态下比 Pd(111)和过氧化 PdO 催化剂具有更高的 CO 氧化催化活性。这一发现有助于解释在 CO 氧化过程中观察到的"PdOx"活性相的长期难题。

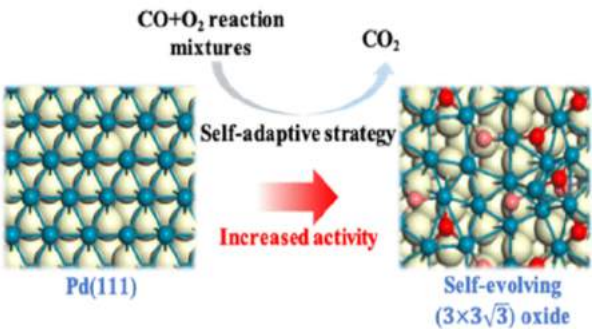


图. AI4S 仿真模拟 Pd 纳米颗粒的形貌在催化过程动态变化

此外，许多传统能源形式，如煤炭和石油，需要经过转化才能用于日常生活和工业生产。这些转化过程通常涉及到将原始能源转化为电力、热力或化学燃料等更便于使用的形式。

这种转化过程的效率对我们利用能源的效率有重大影响。如果转化过程的效率低，那么我们就需要更多的原始能源来产生相同量的电力或热力。这不仅增加了能源成本，还可能增加环境污染，因为大部分能源转化过程都会产生一定量的排放物。

此外，转化过程的设计和优化也可以影响能源的环境影响。例如，通过设计更环保的转化过程，我们可以将原始能源转化为更低污染的形式，或者减少在转化过程中产生的有害排放物。这对于实现可持续发展和减少气候变化的影响具有重要意义。

因此，改进能源转化过程的效率和环保性，如通过开发新的催化剂和优化化学反应过程，是当前能源科学研究的重要目标。

2023 年国际权威期刊 ACS Catalysis 报道了英国女王大学研究人员的相关工作。论文主要探索了氧空位（OVs）在氧化物催化剂中的作用，特别是在二氧化锌（ZnO）表面对一氧化碳（CO）和氢（H2）活化的影响[6]。论文的技术思路如下：

利用机器学习势和基于遗传算法的全局优化，以及密度泛函理论验证，研究者们探索了不同浓度的氧空位在 ZnO 表面的分布。研究者们发现，当存在氧空位时，ZnO 表面会从纤锌矿结构转变为体心四方结构。这些空位形成了一个 Zn3 簇位点，使得氢的自裂解和 C-O 键的断裂变得可能。他们还发现，内在位点（Zn3c 和 O3c 位点）的活性几乎不变，而由氧空位产生的位点的活性则强烈依赖于氧空位的浓度。此外，表面上氧空位的分布会明显影响反应过程，例如当氧空位沿[12$\bar{1}$0]方向排列时，C-O 键解离的能垒显著降低。

研究者们揭示了氧空位在催化反应中的重要作用，特别是它们在 CO 和 H2 活化中的作用。在某些氧化物催化剂中，氧空位的存在、浓度和分布决定了催化剂的活性。研究者们提出，通过调节氧空位的浓度和分布，可以优化催化剂的设计，从而对氢和一氧化碳的活化产生重大影响。这项研究不仅为理解氧空位在催化剂活性中的作用提供了新的视角，也为通过调节氧空位的浓度和分布来优化催化剂设计提供了新的思路。

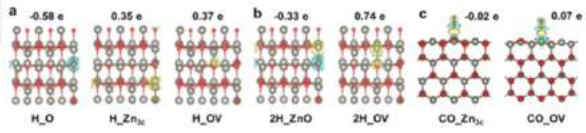


图. ZnO(10$\bar{1}$0)表面上 0.67 ML（单层分子）氧空位（OVs）情况下 H（a）、2H（b）和 CO（c）的可能吸附构型，以及在等值面值为 5 × 10^−3 eÅ^−3 时对应的电荷密度差异。黄色和蓝色泡泡分别表示电荷的积累和耗竭。

除了降低碳排放，人类也需要开发"减碳技术"，如碳捕捉。现有技术仍聚焦于借助自然力量对二氧化碳进行改造，使得其面临高额基础设施投资成本，从而无法实现低价、高效捕碳，这也导致碳捕捉技术无法大规模应用。前文 3.4.4 中叙述了 IBM 研究院使用 AI for Science 对 MOF 材料二氧化碳"捕捉"的研究，此处不再复述。

AI4S 有望通过探索新的碳捕捉路径，从研发端出发找到捕碳低成本路径，如将二氧化碳高效转化为淀粉、燃料等物质，实现高效"再利用"；也可以借助 AI4S 优化和管理系统的作用，模拟捕捉过程，降低实际生产中的浪费等，拉近实验与产业距离，推进碳捕捉技术未来的产业化应用。[1,2,3,4]

在污染治理领域，AI4S 有望通过优化能源使用全流程、揭秘空气中污染物的反应路径等方式，帮助深入研究酸雨、光化学烟雾、雾霾等环境污染问题，助力发展绿色转型。

环境污染问题是各国面临的共同问题，如何解决该问题也是提升人类生活质量的关键。近年来我国多次强调保护环境的重要性，十四五也明确提出"推动绿色发展，促进人与自然和谐共生"。面对更高要求的绿色发展，AI4S 新范式有机会为环境科学开辟一条发展的"快车道"。

source：
[1] 碳中和：https://www.sohu.com/a/477943681_313745 、
[2] https://www.cdmfund.org/30049.html 、
[3] 碳中和白皮书：
https://www.sohu.com/a/523994746_120189950 ；
[4] 碳中和技术：
https://www.sohu.com/a/470047224_120056227
[5] J. Am. Chem. Soc. 2022, 144, 45, 20601–20609
[6] ACS Catal. 2023, 13, 8, 5104–5113

AI4S 实践（32）：《Science》收录加州大学伯克利团队成果：揭示并模拟影响空气质量和气候的关键原理过程，为解决酸雨等问题提出新理论

人类日常的汽车、工业为期中含有较多的氮氧化合物，这些物质排放到空气中会形成各类污染现象，如酸雨、光化学烟雾等。其中酸雨是目前中国华南地区常遭遇的问题，极大的影响了日常人们出行和生活，同时也对自然界中动植物产生不可逆的影响。因此如何降低空气中的氮氧化合物成为环境科学研究的热点问题之一。N_2O_5是造成酸雨等污染的重要元凶，其在空气中的反应过程是决定污染程度的关键。

水汽凝胶对 N_2O_5 的活性吸收是清除对流层中氮氧化物的主要途径。然而该反应过程非常快速，很难通过实验手段观测并了解其微观机理。因此，分子动力学模拟是了解这类复杂化学反应体系微观过程的有效工具。然而，现有的力场模型不能准确模拟 N_2O_5 活化反应过程，必须依赖于第一性原理精度的从头算分子动力学模拟（AIMD）。AIMD 计算非常昂贵，极大的限制了其在更大的空间以及时间尺度上进行采样。

加州大学伯克利分校 David T. Limmer 组通过使用深度势能训练了第一性原理精度的 AI4S 模型。该模型可以经济且准确的复现水和 N_2O_5 的势能面，为系统研究水汽凝胶对 N_2O_5 活性吸收的微观过程提供了强力的工具。[1]

这篇工作发现：由于液-气界面溶液进行水解和竞争性的再蒸发，水汽凝胶对 N_2O_5 活性吸收过程不是水相介导，而是主要在液-气界面上完成的。基于此发现，作者还提出了 N_2O_5 界面吸收模型，该模型可以很好的解释实验观测结果。AI4S 以极低的成本实现从微观角度理解整个反应热力学和动力学过程，为解决酸雨问题提出了新的理论思路。

Source：[1] MIRZA GALIB, DAVID T. LIMMER, uptake of N2O5 by atmospheric aerosol is dominated by interfacial processes, Science 371.6532 (2021): 921-925.

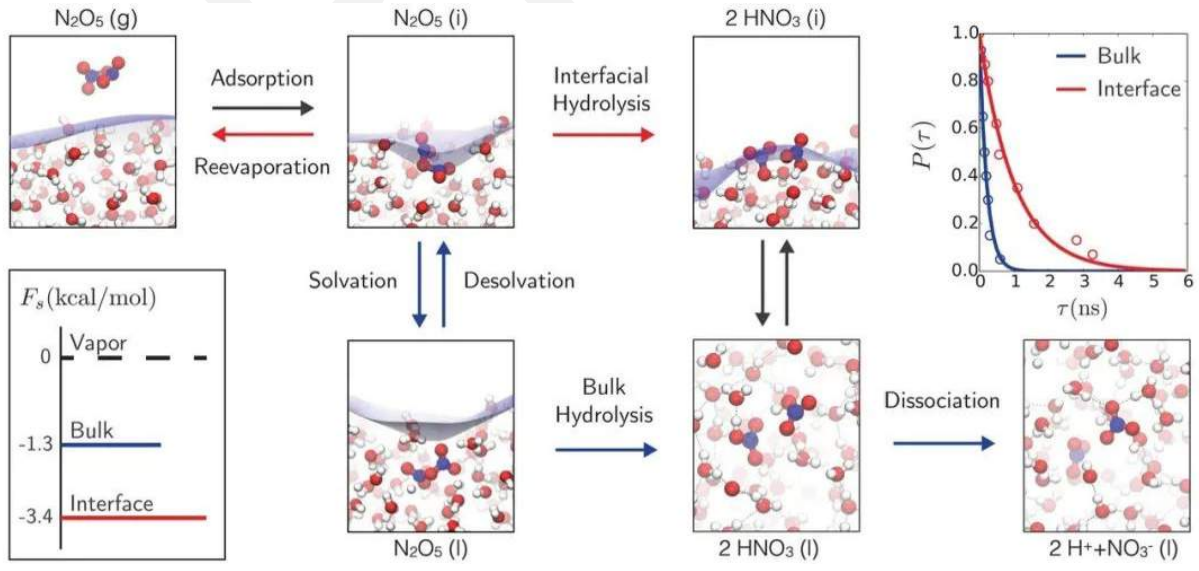


图. N_2O_5在纯水中吸收和反应涉及的物理和化学步骤（界面模型）[1]

6.2.3 海水淡化

水是生命的源泉。中国及众多国家都面临着严重的水资源短缺问题，地表水中的海水约占整个水圈的97.5%。

海水淡化是人类将海水转化成淡水的方法。除了能耗、成本较高的"烧水"方法外，反渗透是全球应用最广的技术路线。[5]

海水淡化膜利用反渗透原理，使水依靠外界的力或能量，从海水侧透过半透膜进入淡水侧，实现海水的淡化。目前海水淡化膜仍存在许多关键问题未被解决。

第一点是淡化海水后剩余的高盐度、高碱性的"废海水"对海洋生态等会产生一定危害，如何处理这部分废水成为一个关键问题；

第二点由于淡化膜利用反渗透原理，现有膜材料有能耗高、产出低的特点；

第三点则是由于海水中物质复杂，淡化膜适用性有限，使得其使用寿命短；此外微生物、颗粒物堵塞膜通道、制备膜材料有限、成本高等问题均或多或少限制了海水淡化膜的发展。

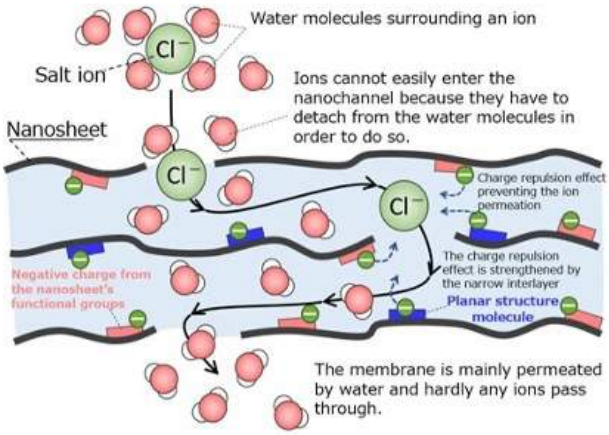


图. 海水淡化膜淡化海水机理

AI4S 有望为解决上述问题带来新思路。AI4S 可以模拟渗透过程，对膜材料进行筛选，发现更有潜力、更适合生产的新淡化膜材料；此外 AI4S 可以结合海水数据和海水中分子、微生物的模拟轨迹，进行优化、调节膜表面、膜参数等性质，增加膜的使用寿命和适用性，以期最终提高相关产业的经济效益。

2022 年 Temple 大学研究员在《自然•通讯》发表文章，增加我们对于离子在水中如何影响水的结构和性质的理解，这是水淡化技术所需要的基础研究[6]。水淡化是一种去除水中盐分和其他溶解物质的过程，以便把海水或其他盐水转化为淡水。在这个过程中，我们需要理解离子如何分布在水中，以及它们是如何影响水分子间的相互作用的。

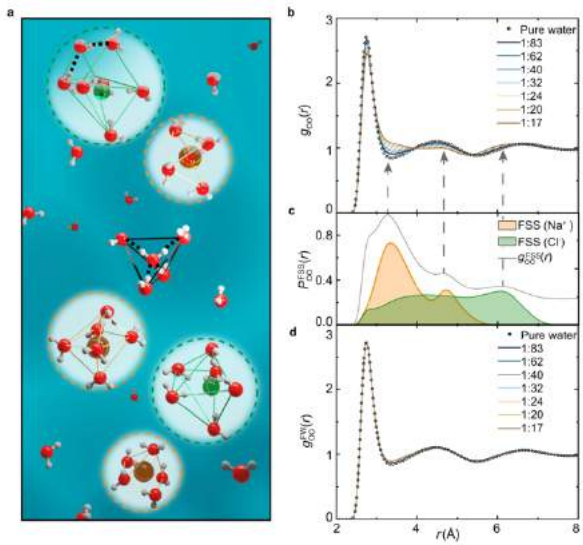


图. a 是一张钠氯化物溶液微观结构的示意图。红色、白色、黄色和绿色球分别代表氧、氢、Na+离子和 Cl-离子/原子。虚线的黄色/绿色圆圈显示了 Na+/Cl-的第一溶剂壳(FSSs)的半径，其定义为 O-Na/O-Cl 径向分布函数的第一个极小值。虚线的黑线表示水分子之间的氢键。Cl-的 FSS 中的氢键链表明水的四面体结构部分恢复。b 显示了纯水和不同浓度下的 NaCl 溶液中 O-O 径向分布函数($g$oo(r))。c 显示了 Na+或 Cl-离子的 FSS 中 O-O 距离概率分布(PFSSoo(r))。灰线gFSSoo(r)表示所有 FSSs 对$g$oo(r)的贡献。gFSSoo(r)的绝对值被缩放。d 将 NaCl 溶液中的 O-O 径向分布函数(gFWoo(r))与纯水的$g$oo(r)进行了比较。

在这篇文章中，作者们使用基于密度泛函理论的机器学习分子动力学模型来研究不同浓度的氯化钠（NaCl）、氯化钾（KCl）和溴化钠（NaBr）溶液。他们发现，溶解在水中的离子并不会像纯水受到高压影响时那样扭曲水的结构。相反，计算得出的结构变化仅限于侵入氢键网络的离子第一溶剂壳，超过这个范围，氧的径向分布函数相对于纯水没有发生主要变化。

这些发现对于淡水化过程有一定的间接影响，因为它们揭示了盐分对水结构的影响程度和方式，这可能会影响淡水化过程的效率和效果。例如，如果我们更好地理解了水中离子的分布和水分子间的相互作用，我们可能能设计出更有效的淡水化方法，例如改进膜过滤技术或开发新的淡水化技术。

在各种膜材料体系中，石墨烯被认为是理论性能最优的终极方案，然而学界对其微观性质和机理的研究仍处于初期阶段。《Carbon》2022 年收录了复旦大学研究组的成果。[1] 研究者使用深度势能对单层石墨烯进行高精度高效建模，并对其复杂的空位缺陷（vacancy defect）进行了详尽的探索，为其在相关场景用的应用提供了前期的理论探索。

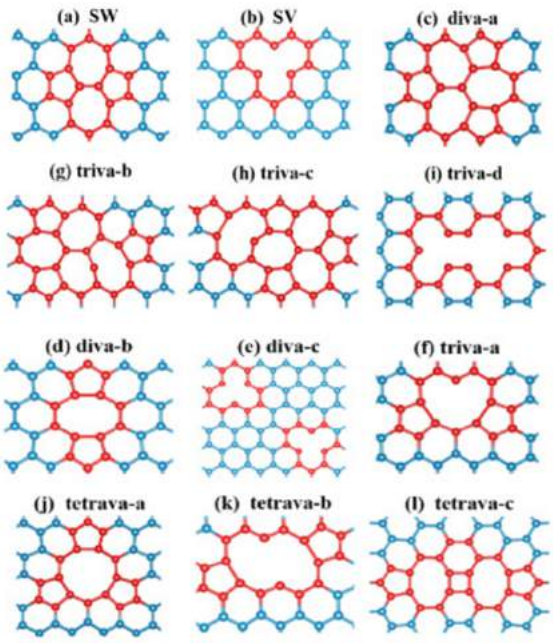


图. 石墨烯的各类缺陷 [1]

除了传统的反渗透原理，全球科研人员也在探索基于其他原理的海水淡化方法。

其中，东北大学研究组提出基于 TiN 吸收太阳能对海水进行加速蒸腾后冷凝回收的系统 [2]；滑铁卢大学研究组提出基于三维有序层级多孔的 TiN 对离子进行吸附脱盐的系统。[3]

业界对 TiN 的微观结构和机理的研究仍处于早期积累阶段。日本中央大学和伊藤忠的研究人员利用深度势能方法对 TiN 微观结构进行了建模和研究，使模型达到了 AIMD 精度，为相关研究的进一步开展提供了前期验证。该成果发表于《Computational Materials Science》[4]

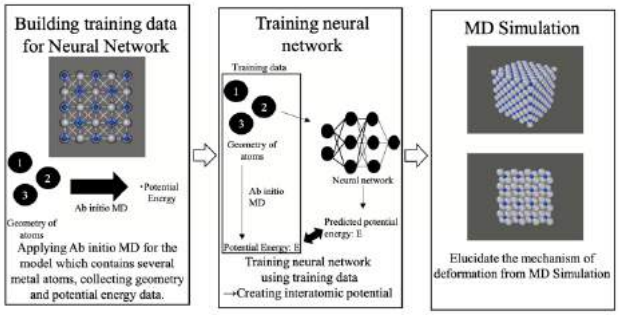


图. Chuo Univ. 基于深度势能的 TiN 模型 [4]

整体而言，海水淡化技术中,反渗透(RO)技术仍是主流和核心技术。RO 技术的关键是膜材料,目前广泛使用的高分子膜材料,其分子结构复杂,进行全原子模拟存在困难,开发准确的分子力场也不容易。

使用机器学习势能或大规模分子表征模型对这类高分子材料进行模拟，理论上可以有效描述分子间相互作用力,并有机会用于指导膜材料的优化设计。

这对推动膜材料的发展以提高海水淡化技术效率,降低成本,具有重要意义。这方面的工作仍在探索阶段，我们会对此持续关注

淡水资源的日益稀缺,已经成为限制人类可持续发展的重大威胁。依靠科技进步才能实现海水淡化等新型淡水资源的开发利用。我们需要系统性地研究和开发应用 AI 等前沿技术,来推动海水淡化技术的进一步发展,而不能停留在局限的小试或碰运气的层面。充分利用 AI 的预测建模能力,可以显著提高海水淡化系统的效率;利用机器学习等方法开发新型膜材料,可以降低成本。

海水淡化技术的系统性研究和开发,既需要政府的支持与规划,也需要企业的投入与实践。同时,也需要科研人员保持开放和协作的心态,在国际上进行广泛合作。只有这样,我们才能取得实质性的进展,以科技进步真正解决淡水短缺问题,造福人类。我们应该为我们的子孙后代负责任,不断推进这项事业。

Source:

[1] Jinjin Wang, Hong Shen, Riyi Yang, Kun Xie, Chao Zhang, Liangyao Chen, Kai-Ming Ho, Cai-Zhuang Wang, Songyou Wang, A deep learning interatomic potential developed for atomistic simulation of carbon materials, Carbon, Volume 186, 2022, Pages 1-8, ISSN 0008-6223, https://doi.org/10.1016/j.carbon.2021.09.062.

[2] Yanpei Tian, Xiaojie Liu, Shilin Xu, Jiansheng Li, Andrew Caratenuto, Ying Mu, Ziqi Wang, Fangqi Chen, Ruizhe Yang, Jun Liu, Marilyn L. Minus, and Yi Zheng, "Recyclable and efficient ocean biomass-derived hydrogel photothermal evaporator for thermally-localized solar desalination", Desalination, 10.1016/j.desal.2021.115449。

[3] Wu, Yuchen; Jiang, Gaopeng; Liu, Guihua; Lui, Gregory; et al.， 3D ordered hierarchically porous non-carbon electrode for highly effective and efficient capacitive deionization, Journal of materials chemistry A, ISSN : 2050-7496

[4] Takeru Miyagawa, Kazuki Mori, Nobuhiko Kato, Akio Yonezu, Development of neural network potential for MD simulation and its application to TiN, Computational Materials Science, Volume 206, 2022, 111303, ISSN 0927-0256,

[5] Huaon, https://www.huaon.com/channel/other/676428.html

[6] Zhang, C., Yue, S., Panagiotopoulos, A.Z. et al. Dissolving salt is not equivalent to applying a pressure on water. Nat Commun 13, 822 (2022). https://doi.org/10.1038/s41467-022-28538-8

第七章：浅谈 AI for 工业仿真的机遇

工业仿真是利用计算机技术模拟工业系统的过程，目的是在虚拟环境中测试不同的方案,以优化生产流程和设备使用。工业仿真在汽车、航空航天和电子等行业有着广泛的应用，例如在汽车行业,可以通过仿真软件来模拟车辆的气动学性能、碰撞安全性、动力系统等,评估设计方案,优化产品性能。在航空航天领域,可以测试飞机的气动布局、发动机性能、控制系统等。

人工智能（AI）在过去的十年里对工业设计和仿真领域产生了显著影响。传统的设计过程，主要依赖人类的直觉和经验，现在正在被 AI 驱动的方法补充，甚至在某些情况下，被 AI 方法取代。特别是在有限元分析（FEA）、计算流体动力学（CFD）求解以及基于自动化工业几何构造（AIGC）的工业设计生成和参数优化方面等，AI 已经显现出其巨大的潜力和价值。AI 有望在工业设计的不同阶段做出贡献。

7.1 生成式设计

AI 在工业设计中的重要角色是生成式设计。生成式设计算法利用机器学习技术，在给定一组约束（如材料、成本、性能等）和目标的前提下，生成众多设计选项。这种方法使工程师能够探索比传统方法更广阔的设计空间，从而产生创新且有时与直觉相反的设计建议。这些设计往往有更好的性能，减少了材料的使用，更优化地适应了如 3D 打印这样的制造技术。

从最直观的角度，工业设计的第一步 —— 概念设计，已经是 AI 所触及的领域。诸如 stable diffusion/midjourney 等生成式 AI 工具可以帮助设计师从抽象的概念/词汇出发，生成高质量的渲染图，为汽车、纺织等等领域提供源源不断的灵感

Prompt: Barbie's pink sportscar, generated by DALL•E

其次，AI 可以通过"点云"等方式处理工程级别的设计需求。点云是一种在三维空间中表示物体形状的方法，每个点都包含了该点的位置信息。在工业设计中，可以通过 3D 扫描设备获取现实世界物体的点云数据，然后再用来生成 3D 模型，进行进一步的设计和优化。在生成式设计中，人工智能可以用来处理这些点云数据，生成新的或改进的设计。2022 年 12 月，OpenAI 发布 Point•E 即是此类应用。具体而言，深度学习算法可以接收一个初始的点云作为输入，然后通过迭代的过程，改变这些点的位置，使得新生成的设计能够满足特定的性能目标。这种方法可以用来优化设计，比如减少飞机翼的阻力，或者提高建筑的结构稳定性。

此外，GAN 和 VAE 技术也在工业设计领域有应用前景。生成对抗网络 (GANs) 是一种深度学习模型，它包括两个子模型：生成模型和判别模型。生成模型试图创建与真实数据集相似的新数据，而判别模型试图区分生成的数据和真实数据。通过生成模型和判别模型的相互对抗和协同学习，GAN 可

以生成非常接近真实数据的新数据。在工业设计中，GAN 可以应用于生成设计概念，为设计师提供新的设计灵感。例如，在产品设计中，通过对大量现有产品图像的学习，GAN 能够生成具有创新性的新产品设计草图。同时，判别模型也可以作为一种质量控制工具，帮助筛选生成的设计草图，保证它们在一定程度上与现有的优秀设计相似。

变分自编码器 (VAEs) 是一种生成模型，它可以学习输入数据的潜在表示，并从这个潜在表示生成新的数据。VAEs 通过优化重构损失和潜在空间的分布，使得我们可以对潜在空间进行平滑且连续的插值，从而生成多样性的新数据。在工业设计中，VAE 可以用于探索设计的可能性，即通过学习和理解现有设计的潜在特性，VAE 可以生成新的设计草案。例如，在汽车工业设计中，我们可以使用 VAE 来学习汽车的外形设计，然后在潜在空间中进行插值或随机采样，生成新的汽车外形设计。此外，VAE 的潜在空间还能反映出设计的各个因素之间的关系，比如我们可以通过操纵潜在空间中的某些维度，来看看它们如何影响最终的设计。

Point•E, by OpenAI

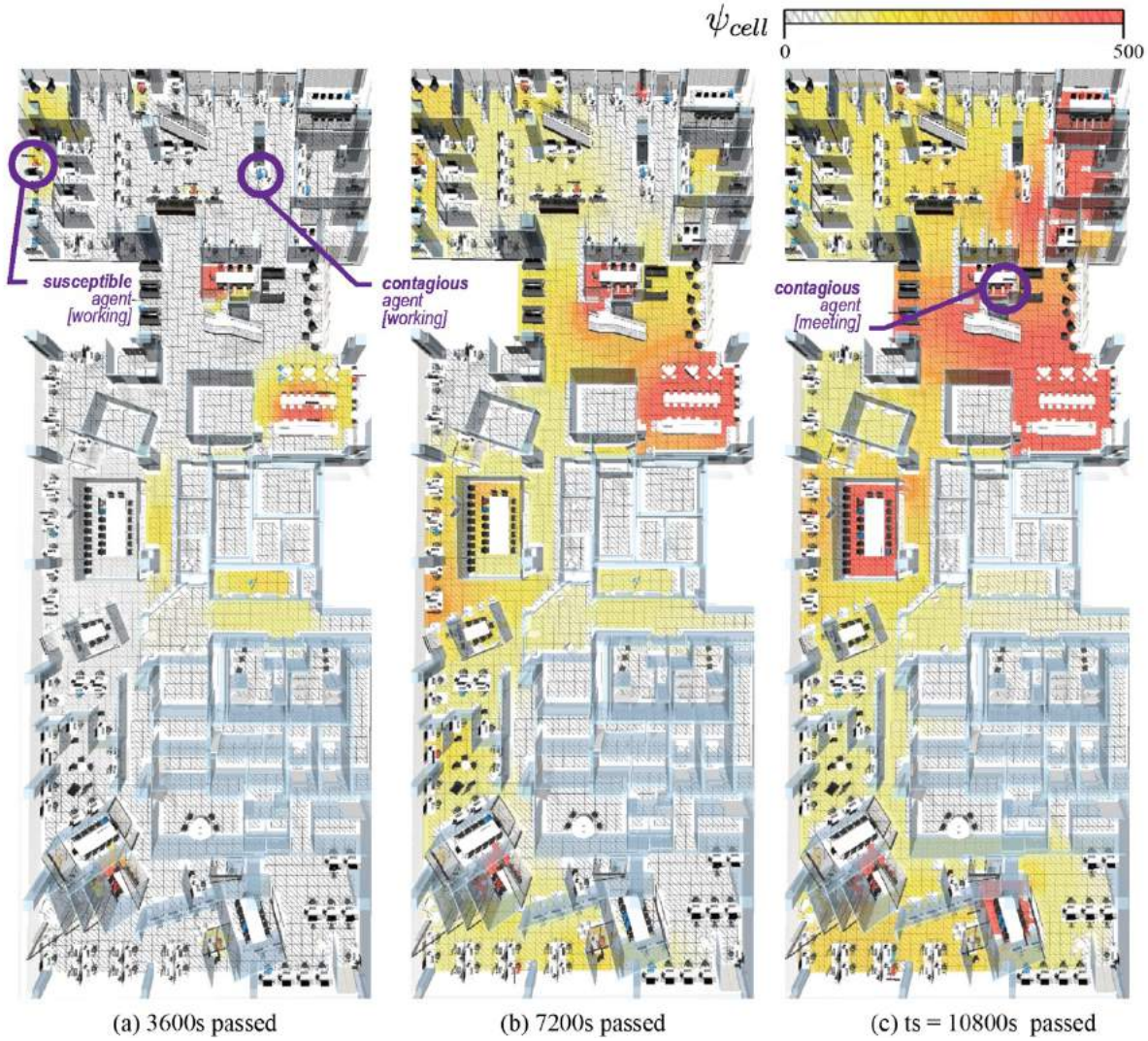
AI4S 实践（33）：Autodesk Research 使用 AIGC 将公共卫生需求融入房屋设计

2023 年，Autodesk 研究部门开发了一种生成式设计方法,通过全面的多智能体仿真来捕捉建筑内人员的行为,从而处理病毒传播的复杂性。该仿真同时跟踪病毒是如何从带原者传播到空气和表面,以及最终造成的感染情况,从而产生更安全的设计。其论文发表于期刊《Sustainable Cities and Society》[1]。

文章还介绍了具体的工作流程:首先生成大量不同的建筑布局,然后对每个布局进行病毒传播风险的仿真测试,系统通过学习哪些设计元素可以降低风险来产生新的低风险设计。最终的设计方案比基准设计降低了 20%的病毒空气传播风险。

这种方法的创新之处在于将复杂的人员行为模拟融入到生成式设计中,从而比仅仅使用简单距离指标等要全面和准确。这有助于我们在后疫情时代设计出更安全的建筑。

Source: [1] Bokyung Lee, et al, Generative design for COVID-19 and future pathogens using stochastic multi-agent simulation, Sustainable Cities and Society, 2023, https://doi.org/10.1016/j.scs.2023.104661.



病毒粒子随时间在空气和表面积聚的动画。需要注意的是,高风险区域主要集中在会议室以及如厨房和休息室等人员流动频繁的空间。

7.2 逆设计 / 逆问题

反向设计是 AI 可以产生重大影响的另一个领域。反向设计在给定一组特定的期望输出或性能指标的前提下，利用优化算法逆向工作，确定实现这些输出的最优设计参数。这种方法可以与生成式设计结合，以得到更好的结果。强化学习和贝叶斯优化的最新进展正在为更高效的反向设计过程铺平道路，使得我们可以在庞大的设计空间中更快地找到最优设计。

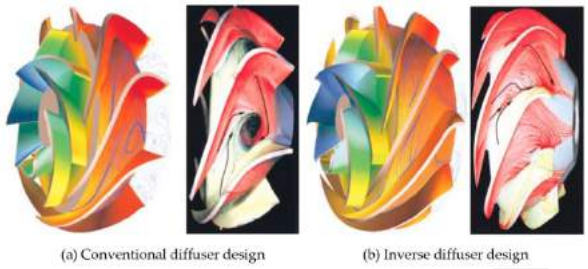


图. 水利机械 Diffuser 逆设计 [1]

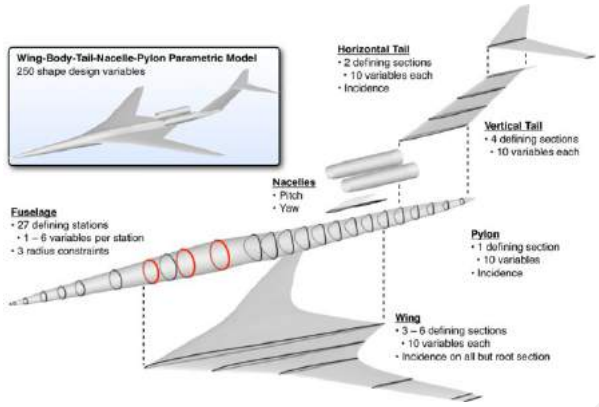
逆问题与逆设计：逆问题指的是一些或所有的初始条件、边界条件或 PDE 的系数未知的情况，目标是从观察到的数据中确定或恢复这些未知数。逆问题通常假设观察到的数据在物理上是合理的，并代表了 PDE 的解。例如，在流体力学中，观察到的数据可能是涡度场，而只有初始条件是未知的。那么，逆问题就是确定会产生这样的涡度场的初始条件 u0。另一方面，逆设计更具体地指的是一种设计或优化方法，其中给定了预定义的目标，目标是根据目标优化系统配置。例如，给定一个可以模拟前向流体动力学的代理模型 u，目标可能是设计一个可以将流体引导到期望位置的表面。对于逆设计，可能并不存在精确的解，但是，我们仍然可能希望优化所提出的解以尽可能地满足目标。从某种意义上说，逆设计也可以被认为是一种特殊类型的逆问题，其中的目标不仅是确定未知的参数或系数，而且是设计一个按照特定方式运作的系统。[1]

在逆问题领域，AI 有可能创造新的机会，我们举例如下

- 流体动力学基础：学习流体动力学的替代模型通常需要使用昂贵的经典求解器来获取训练数据。另一种方法是考虑一个反问题，任务是仅基于 3D 动态流体场景的多视图视频推断出底层动力学。
- 系统识别：传统上，估计物体的物理属性需要进行许多物理实验和使用特别设计的算法。这里有一个有希望的反问题是直接从视觉观察中推断物理属性。
- 地球物理全波形反演：在地球物理学中，可以从地面上的地震波测量中推断地下的性质，如密度或波速，这个问题被称为全波形反演。这些地下属性对于能源勘探或地震预警等应用非常重要，但由于问题规模大，这些属性通常很难测量。
- 流体同化和历史匹配：流体同化的目标是从时空领域的稀疏观测中恢复整个流体场。流体同化可以应用于模拟地下流动。地质模型被调整，以使预测与历史观测相匹配，这是一个被称为历史匹配的任务。
- 医学成像的断层扫描：断层扫描的目标是仅使用表面测量来恢复物体的内部结构。例如，在医学成像中，电阻抗断层扫描(EIT) 可以通过测量皮肤上的电压分布来推断内部器官的状态，当电流被注入时，这避免了侵入性的测量或辐射暴露。

在逆设计中，我们确定了一些 AI 可以辅助并且有大量机会的应用：

- 飞机形状设计：在气动学中，一个重要的挑战是设计飞机的形状以减小阻力。这涉及到模拟空气流体动力学和它与飞机边界形状的交互。



Parametric wing-body-tail model used for inverse design [3]

- 离子推进器设计：在航天工程中，有效推进器的设计非常重要。例如，霍尔效应推进器(HET)是最吸引人的电推进(EP)技术之一，因为它具有高比冲和高推力密度。一个关键的问题是如何设计推进器的形状和材料排列，考虑到其复杂的等离子体动力学。
- 控制核聚变：解决控制核聚变可以为无限的清洁和廉价能源铺平道路。在磁约束托卡马克中，控制核聚变的两种主要方法之一，一个关键的挑战是优化外部磁场和壁设计，以便将等离子体塑造成具有良好稳定性、约束性和能量排放的配置。
- 芯片制造：芯片制造中的许多过程都涉及反向设计。一个重要的应用是等离子体沉积。具体来说，问题是如何设计电介质单元的形状，使得等离子体沉积到基板上尽可能平滑。
- 水下机器人形状设计：在水下机器人中，一个重要的问题是设计机器人的形状以达到多个目标，包括减小阻力、提高能效、提高操控性以及改善某些声学属性。
- 应对气候变化：在许多应对气候变化的方法中，反向设计可以发挥重要作用，包括改善建筑材料、优化碳捕获、太阳能地质工程以及设计碳信用和政策。
- 纳米光子学：纳米光子学专注于设计与电磁波波长接近的结构。为与光交互的微尺度结构、纳米尺度结构或拓扑模式设计原则方法在激光发生、数据存储、芯片设计和太阳能电池设计等应用中具有重要意义。
- 电池设计：深度学习启用的反向设计在电池设计中有巨大的潜力。例如，它可以用于电池界面的反向设计，这对于开发高性能可充电电池非常重要。除了电池本身，机器学习中的超参数搜索技术可以用于加速锂离子电池的高循环寿命充电协议的实验探索，这对于电动汽车至关重要。

7.3 设计验证（正向模拟仿真）

一旦一个设计被生成，AI 也可以在验证阶段提供帮助。正向建模是一个过程，在这个过程中，最终设计被模拟以预测其在真实环境下的性能。这通常涉及解决复杂的方程和模型，可能需要花费大量的时间和计算资源。机器学习模型可以被训练用来预测新的模拟结果，大大加速了验证过程。

以 FEA 为例，FEA 是一种强大的工程工具，被广泛用于机械、土木、航空航天等领域的结构分析，以验证工程师的设计能在现实中成立。然而，FEA 模型的构建和求解过程需要大量的计算，耗时耗能。这就是 AI 技术发挥作用的地方：通过深度学习，AI 可以从大量的 FEA 模型和结果中学习和提取有用的信息，实现更快速、更精确的求解。此外，AI 也可以用于自动化的网格划分和优化，大大提高了 FEA 模型的构建效率。

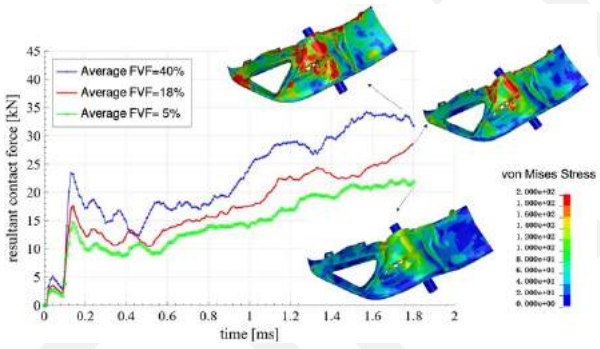


图. LsDyna+机器学习对复合材料的多尺度非线性建模 [5]

在航空等领域，CFD 被用于验证设计的气动性（如飞机能否真的飞的起来）。CFD 求解则涉及到复杂的流体动力学问题，包括流体的流动、热量传递以及物质输送等。传统的 CFD 模型求解通常需要复杂的数值计算，并且对计算资源的需求非常大。然而，通过引入 AI 技术，我们可以利用神经网络和机器学习算法来预测流场的行为，以实现更快、更高效的 CFD 求解。

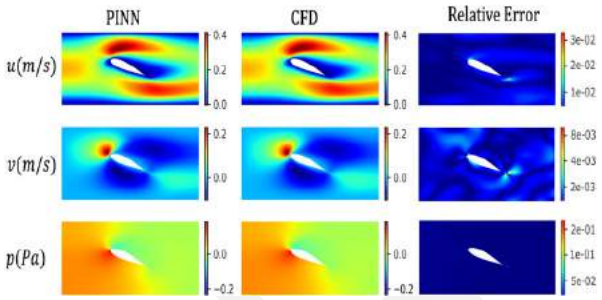


图. 在一项机翼气动仿真中 PINN 表现与 CFD 高度一致 [6]

在各工业行业中，制造可行性都是从设计落地到现实的关键评估。即使最好的设计，如果不能有效和经济地制造，也是无用的。AI 可以通过分析诸如制造能力、成本、时间等因素，来帮助评估设计的制造可行性，这对设计的成功实施至关重要。

在制造之后，AI 可以帮助预测产品的寿命并据此进行维护计划。预测性维护模型可以使用实时数据在故障发生前预测故障，从而提高产品的可靠性，减少停机时间。

工业设计是长流程的系统性工作，如何兑现 AI 在工业仿真中的潜力，需要产业上下游和研发链条上各方的通盘考虑。成功的使用 AI，有机会给企业带来额外的经济效益，并形成竞争优势 [2]

提升研发效率

如今用于工程和科学计算的计算机模拟正在从 AI 获得越来越多的帮助（在某些情况下，甚至被 AI 替代），这大大降低了成本，并帮助工程师更快地找到最佳答案。

运行模拟通常需要大量的费用，需要超级计算机对海量数据集进行计算和执行高度复杂的计算。但

是，如果你可以建立一个关于物理运作方式的机器学习（ML）模型，你就不需要每次都运行模拟，因为你的 ML 推理模型可以从数据中推导出答案。

这意味着你可以更快、更便宜地获取到关于某种设计如何表现的答案。从长期来看，机器学习、物理信息神经网络和其他基于 AI 的工具将成为所有工程师和科学家在 R&D 开发中提高生产力的标准工具。

企业知识存留

那些善于创建精良的 AI-物理模型的组织将获得坚实的竞争优势。这样的能力将帮助他们建立基础物理理解，了解不同产品设计在现实生活中的表现。这对于组织在研究和产品开发中保留知识具有巨大的影响。

AI-物理模型可以捕获关于物理对象行为的最佳实践知识——这些信息通常是由专家（如科学家、工程师或设计师）头脑中的信息。例如，一名飞机工程师积累了关于机翼形状最佳设计方法的知识，这些知识将指导他们关于使用数字模拟探索哪些设计选项的选择。然而，这种信息现在可以通过 AI 捕获，然后提供一份工程师可以进一步用数字模拟探索的设计建议的简短列表。最关键的是，假设一家公司可以创建一个关于机翼设计最佳实践的 ML 模型（通过训练 AI 工具关于飞机机翼的知识），那么无论工程师是否离开公司，它都可以保留那种专业知识。这也给组织带来了更大的灵活性。如果一家公司想要构建一种在强风中更稳定的新型飞机，一个 ML 应用可以快速产生机翼形状的最佳选项，帮助组织快速建立新的原型以进入新的市场。

机遇和挑战

拥抱 AI 的组织将加速工程和科学发现，同时开发创新的新解决方案，这些解决方案使用传统方法在计算上将无法实现。

尽管 AI 充满了希望，但各行业的组织都需要建立工程和研究的最佳实践，以帮助确保他们安全地转型，最大化收益而无需无谓的风险。企业仍需要做大量的工作，以提供让 AI 工具在正确方向上变得聪明的必要数据。此外，AI 的法律问题依然很模糊。组织必须仔细审查 AI 的输出以确保准确性，同时警惕任何道德红旗。安全也是另一个需要考虑的重要因素，以确保 AI 实践不会无意中暴露知识产权或专有信息。

当然，在我们探索这项新技术的初期，为组织使用 AI 设置防护栏是必要的。但是，有了一些深思熟虑的措施，AI 可以安全地开启全新的研发可能性，帮助组织行动更快，变得更加灵活，并发现创造未来的更好方式。

总的来说，AI 有潜力革新整个工业设计和仿真的周期，从初步设计阶段，到验证，再到制造和维护。公司通过采用 AI，可以增强创新能力，缩短开发时间，提高产品性能，最终获得竞争优势。然而，AI 在工业设计中的整合也带来了诸如数据隐私、算法透明度、和员工重新技能培训等挑战，这些都需要得到解决，以便成功实施。我们也应认识到，AI 并非万能的，它只是工程师们的一个工具。在使用 AI 时，我们还需要充分考虑到实际的工程背景和需求，以确保得到的仿真和设计结果能够真正满足工业生产的需要。

虽然工业仿真并非是本报告所重点关注的领域，但其对生产生活的影响广泛且深渊。回望过去 20 年并展望未来 10 年，工业仿真和科学计算领域日新

月异（见下页图），本章挂一漏万，仅做摘要性讨论。

Source:

[1] Figure9, Three-Dimensional Inverse Design Method for Hydraulic Machinery, Energies 2019, 12(17), 3210; https://doi.org/10.3390/en12173210

[2] Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems, arXiv:2307.08423

[3] Wintzer, Mathias & Kroo, I & Aftosmis, Michael & Nemec, M. (2023). Conceptual Design of Low Sonic Boom Aircraft Using Adjoint-Based CFD.

[4] How AI physics has the potential to revolutionise product design, https://www.weforum.org/agenda/2023/06/how-ai-physics-potential-to-revolutionise-product-design/

[5] LS-DYNA Machine Learning–Based Multiscale Method for Nonlinear Modeling of Short Fiber–Reinforced Composites, Journal of Engineering Mechanics Volume 149, Issue 3 https://doi.org/10.1061/JENMDT.EMENG-6945

[6] Yubiao Sun, Ushnish Sengupta, Matthew Juniper,Physics-informed deep learning for simultaneous surrogate modeling and PDE-constrained optimization of an airfoil geometry,Computer Methods in Applied Mechanics and Engineering,Volume 411,2023,116042,ISSN 0045-7825,https://doi.org/10.1016/j.cma.2023.116042.

DP Technology
深势科技制图

云、GPU、AI：科学计算与工业仿真的三条主线

	Cloud	GPU	AI
CAE、有限元 ...	★	☆	☆
流体力学、燃烧仿真	★	★	★
量子力学，分子动力学	★	★	★
生命科学，药物发现	☆	★	★
EDA	☆	☆	★
气象预测	☆	☆	★

★ 方案成熟
☆ 探索中

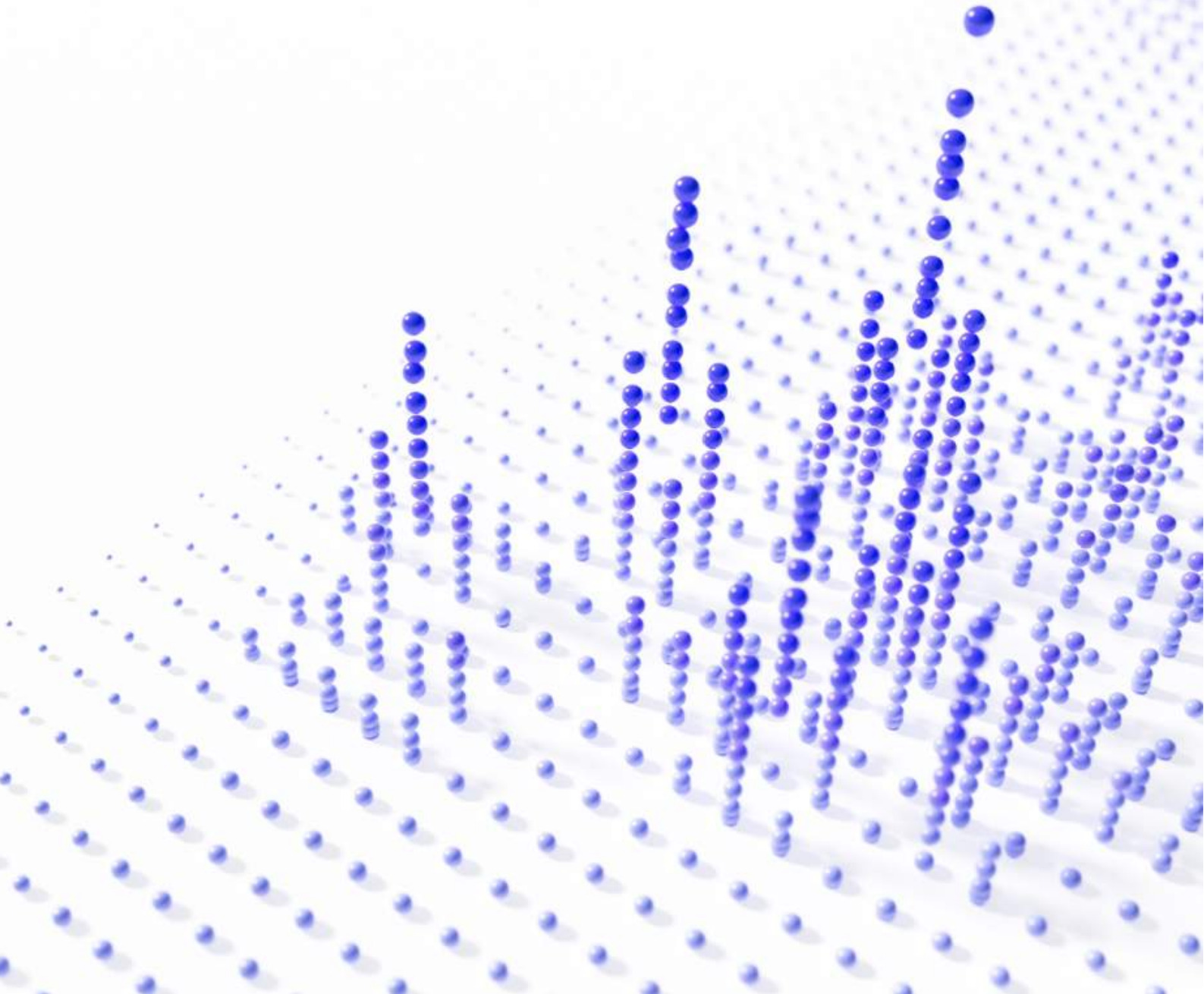
云计算已经深入科学计算的各个场景，头部玩家已全面布局，整体生态愈发完整

GPU 在众多场景中均取得了明显的性能加速；但是 GPU 在相关场景的商业应用受限于成本及适配

新兴 AI for Science 技术正在更多场景中涌现，加速原始创新，提高科研效率

	Cloud	GPU	AI
Pre-2010	AWS, Azure 等云厂商起步	CUDA, OpenCL	/
2010s	超过 1000 家仿真软件商上云	VASP, Gromacs, Ansys mechanical 等软件加入 GPU 支持	AI 加速仿真技术涌现，如 AlphaFold, DeePMD, PINN 等
2020s	云原生仿真软件和云原生科学计算平台涌现	GPU 原生仿真软件涌现	AI原生的计算手段不断完善；生成式 AI 设计，智能自动化实验设计(DoE)，AI 工作流等涌现

PART III:
AI for Science 应用案例和产业观点



宁德时代：拥抱 AI4S 攻坚电池、光伏能源新材料

在电动汽车（EV）领域，"里程焦虑"已经成为阻碍广泛采用电动车的重要心理障碍。一方面，电动汽车的车主希望 EV 车辆的行驶里程能够与他们的燃油驱动的竞品相媲美，甚至超过。另一方面，人们希望未来的电动汽车可以像加注汽油般快速充电，从而实现"充电五分钟，狂飙一千里"。目前，电动汽车主要使用的是石墨负极的锂离子电池。这种电池体系的能量密度和充电速度都受到其材料性能的限制。要想实现显著的性能提升，我们只能求助于开发新的电池材料体系。例如，硅负极和锂金属负极的电池体系具有比现有的石墨负极的锂离子电池更高（3~10 倍）的理论能量密度，因此具有巨大的潜力。然而，这些新材料体系的商业化应用还面临着重重挑战。比如，硅在锂离子插入/脱出过程中体积膨胀的问题，锂金属在充电过程中形成的锂枝晶的问题，都对电池的稳定性和安全性构成了威胁。

所以，要解决电动汽车的里程焦虑问题，最根本的是需要在电池材料科学方面取得重大突破。这需要我们不断深化对材料在原子和分子层面的理解，以及构建更精细、更精确的电池性能模型和预测工具。这是一个复杂而艰巨的任务，但也是一个充满机遇和可能性的领域，值得我们全力以赴。

AI for Science 为能源材料研究开辟了新的途径。比如，全球能源巨头 —— 宁德时代，就利用了人工智能来研究锂金属负极和钙钛矿相变等问题。他们使用了深度势能与先进表征结合来研究这些材料的分子细节，从而能够在分子层面理解材料性能的变化，为材料设计和优化提供了理论指导。同时，这种研究方法也大大提高了研究效率，为能源材料的快速发展和应用提供了可能。

1）AI 加速硅负极、锂金属负极等新电池材料安全性与性能研究

产业一般认为，硅负极技术距离产品化较近，处于"临门一脚"的状态，因此具有极高的商业价值。然而，硅负极的实际应用却因其在锂化/脱锂(放电/充电)过程中的巨大体积膨胀/收缩而受到严重阻碍。为了更好地开发硅阳极，有必要了解脱/嵌锂过程的反应机制以及相应的微结构演变。但是，硅阳极的锂化/去锂化涉及晶相和非晶相之间的复杂转变，且传统的 MD 模拟受限于准确性和效率的困境，导致该问题仍具有挑战性。

2023 年，深势科技联合宁德时代 21C 创新实验室与北京大学数学科学学院和材料科学与工程学院、北京科学智能研究院，采用深度势能方法[1]研究硅基负极在脱嵌锂过程的相变和结构演化。相关研究成果发表在领域顶刊 《AFM》上[1].。

研究人员借助深度势能训练了覆盖整个成分空间的高精度 Li-Si 势函数模型，进一步通过 DeePMD 再现了 c-Si/a-Si 锂化和 c-Li3.75Si/a-Li4.5Si 脱锂过程的电化学性质，捕捉了锂化/脱锂过程的结构转变特征，如晶体与非晶体之间的电压平台差、c-Li15-δSi4 到 a-Li15-δSi4 的相变引起的电压滞后等，揭示了锂化和脱锂反应路径的差异及原子尺度机理，为硅阳极的电化学性能和相变反应提供了重要的见解（见下页图）。

对于产业界，这项研究揭示了硅阳极中相变和结构演变的热力学洞见，也为我们提供了改进硅阳极稳定性的新途径。这些洞见和优化策略在实际的产业应用中可能带来益处，推动下一代锂离子电池技术的进步。

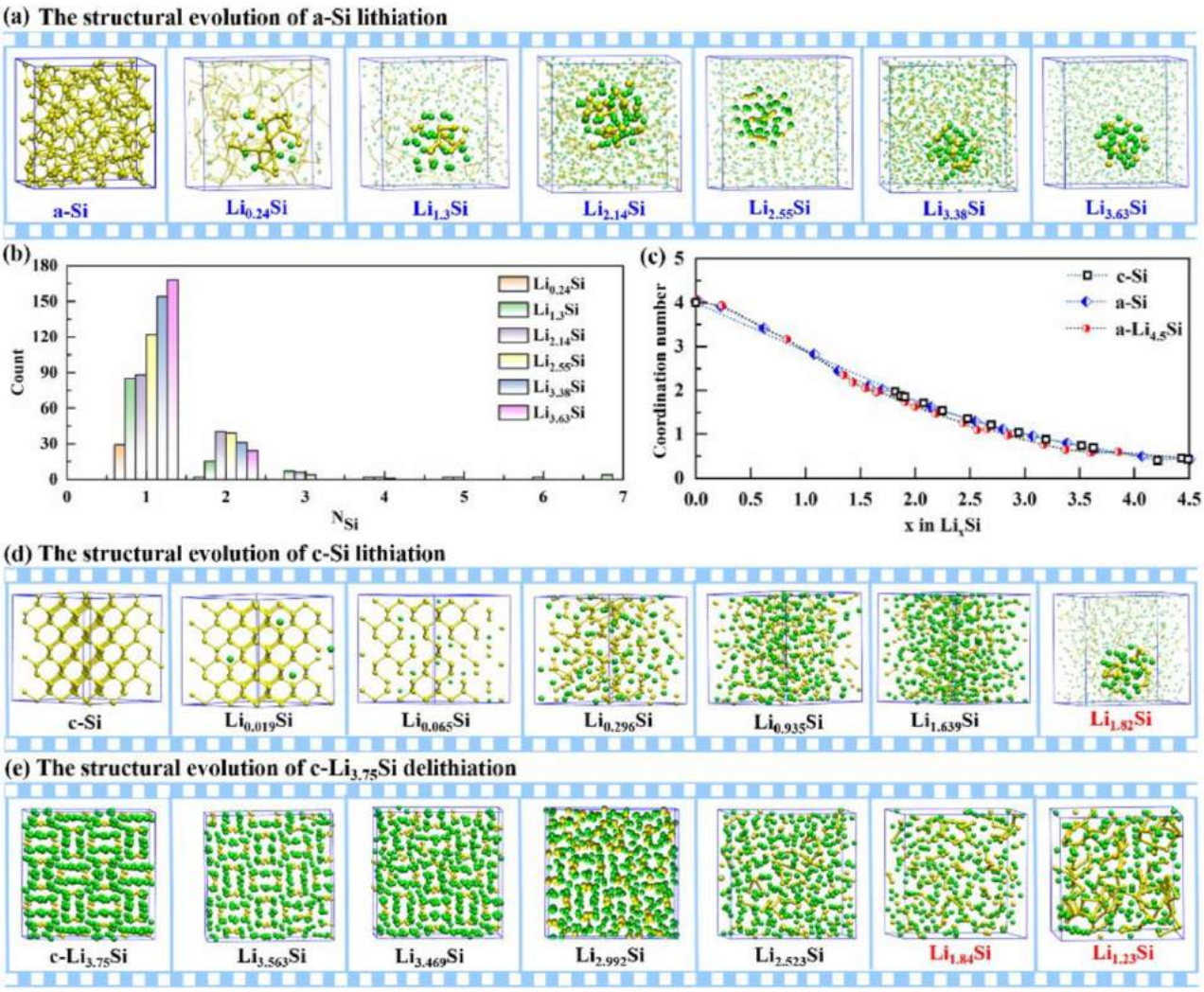


图. (a) 非晶态硅锂化过程从非晶态硅到 Li3.63Si 的典型快照。突出显示了各种非晶态 aLixSi 的一些 Si 原子,以显示 Si 的键合结构(r≤2.40 Å)。(b) 比较了不同 Si 原子(NSi,r≤2.40 Å)的 Si 簇数量在非晶态锂化硅的各种 a-LixSi 结构中。(c) 在切除半径 2.98 Å 的第一个配位壳层中,随着 Li 含量的增加,非晶态/晶态 Si 和非晶态 Li4.5Si 系统中的平均 Si-Si 配位数。(d) 晶态锂化过程从晶态 Si 到 Li1.82Si 的典型快照。(e) 非晶态脱锂过程从晶态 Li3.75Si 到 Li1.23Si 的典型快照。绿色和黄色球代表 Li 和 Si 原子。[1]

在锂金属电池体系中，充放电过程中最大的问题在于形成金属锂的沉积和溶解，导致锂枝晶的产生，材料活性降低。锂枝晶生长存在安全隐患，破坏电极结构，严重制约了锂二次电池的商业应用。目前，人们对锂的沉积（生长）机制在原子尺度上还了解甚少，同时对锂枝晶的成因也复杂多样。不同的研究结果之间甚至看上去互相矛盾。例如，一些研究认为大电流密度会加剧锂枝晶的生长，但一些研究则认为高电流密度可以缓解枝晶生长。这些问题与矛盾的解决需要人们在原子尺度更加详细地研究金属锂的行为与生长机制。

2022 年，宁德时代、北京大学、中科院物理所等机构联合课题组，构建描述金属锂沉积（生长）机制的 AI4S 模型，该模型具有第一性原理计算精

度，是在原子尺度上进行金属锂沉积（生长）模拟的关键。

团队利用该模型发现了金属锂的两种自修复机制，即表面自修复和体相自修复。表面自修复现象是指在金属锂的均匀沉积中，无论初始表面是平整还是有缺陷，在经过一定时间的沉积之后其表面都会趋于平整。基于 AI4S 模型模拟发现由于金属锂枝晶尖端的锂原子势能较高，凹槽处锂原子势能较低，势能差使得锂表面的原子更多的从尖端向凹槽处移动。该现象也通过原子力显微镜的观察被证实。

而当金属锂表面发生非均匀沉积时，两个区域的锂枝晶相互接触从而使得两个枝晶融合在一起，中间的孔洞消失，完成金属锂的体相自修复。这种现象也同样被实验所证实。除上述关于金属锂的自修复机理外，该研究还利用 AI4S 模型进一步阐述了不同温度与沉积速率下锂枝晶的形貌，指导通过利用金属锂的自修复效应来避免金属锂枝晶的生成。研究得出结论，自愈合在纳米尺度上迅速发生，因此，使用若干综合方法最小化锂颗粒之间的空隙可以有效促进无枝晶锂的形成。

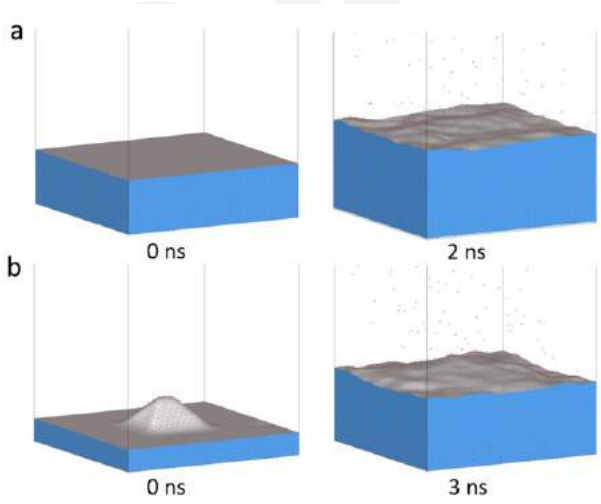


图. 锂沉积的模拟仿真

该工作发表于国际权威期刊《Advanced Science》[1] 这项研究也揭示了锂电池中的自愈合过程，这可能会引导产业界寻找新的防止锂枝晶形成的方法。这项研究也为我们提供了新的视角来理解和改进锂金属负极的性能，例如提高充电温度和改进电流集电器的表面均匀性，这些都可能在实际的产业应用中带来重大的益处。

值得注意的是，电池是典型的多尺度复杂系统，电池材料决定了电池性能的上限，而其他部分的设计则觉得了电池的实际性能可以多接近起理论上限。比如电池中的锂金属负极并不直接与导线相连，而是需要通过集电器。这是桥接负极材料和外部电路的必不可少的组件，极大地影响了锂离子电池的容量，倍率能力和长期稳定性。铜箔具有导电性好，形成的氧化保护膜，质地柔软，粘结性好，制造工艺成熟，价格相对较低的优点。因此，它们被选作锂电池集电器的主要材料。

宁德时代联合研究组使用深度势能对锂金属负极-铜箔的复杂界面问题进行研究，发表于《Advanced Materials Interfaces》[2]。论文提供了一种以原子精度研究这种界面性质的方法，从而为改善电池性能提供了新的工具。其次，这项研究揭示了锂在铜表面的吸附行为，这将有助于设计更好的集电器材料，从而提高电池的效率和可靠性。

基于这项研究，宁德时代与北京大学项目组进一步对无负极体系中锂在铜表面沉积行为进行仿真，发表于《Small》[3]。研究者使用深度神经网络势函数 LiCu-NNIP 描述 Li-Cu 体系的相互作用,保证计算精度。进而利用表面相似性分析(SSA)方法定量分析铜基底表面上沉积的锂层的结构。下一步，研究者通过计算势能分布、动能学特征等分析锂原子在不同 Cu 表面上的运动规律，从而模拟多晶 Cu 基

底的均匀沉积和不均匀沉积,比较不同 Cu 表面对锂沉积的影响。研究认为 Cu 基底的米勒指数对 Li 沉积形成的结构有显著影响。减少商用 Cu 箔中的 Cu(110)相分数,可有效改善无电极 Li 金属电池中的 Li 沉积可逆性和稳定性。

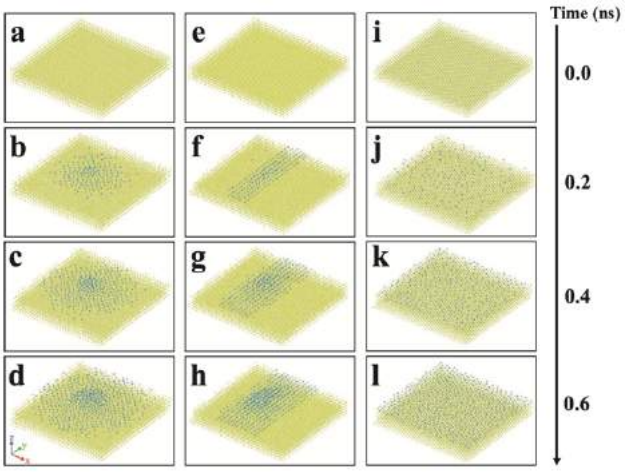


图. 锂在不同铜表面的沉积过程仿真 100, 110, 111 (左到右)

2）AI 探究钙钛矿光伏新材料的寿命机理

钙钛矿材料在太阳能电池方面的应用，有其独特的优势：钙钛矿的光电转化效率发展迅速，已经直逼当下主流的单晶硅体系；钙钛矿的生产成本更低：其生产所需的原材料成本低，且对缺陷的容忍度较高（也就是对工艺精度的容忍度高）；钙钛矿有较高的光吸收系数，因此可以做成薄膜，适应更多的使用场景。由于其独特的性质，钙钛矿是目前发展最快的太阳能技术路线之一。其巨大的产业价值吸引了越来越多的科研团队以及企业投身其中。目前国内已有布局的企业超过 15 家，如宁德时代、隆基光伏、协鑫纳米、纤纳光电、国家能源集团等。学术界的研究中，领域顶刊中与钙钛矿相关的工作同样不在少数。

一般来说，钙钛矿太阳能电池的优劣归因于多个方面：吸收系数、载流子传输能力等，这些关键指标影响着最终成品的效率、寿命、成本等，进而影响着相关企业的竞争力和市场地位。尤其需要关注的是，有机-无机杂化钙钛矿材料是钙钛矿太阳能体系中的佼佼者：这种材料制造成本低，具有更高的介电常数，高载流子扩散长度和速度，出色的光吸收能力。

即使钙钛矿太阳能电池已经表现出相对较好的光电转化优势，但它依然面临着长期运行不稳定这一重要挑战。钙钛矿太阳能电池的不稳定性主要来源于两个方面：一个是材料本身的组分以及结构缺陷；另一个是外部服役环境，如光照时长、温度、湿度等，上述都会加速钙钛矿太阳能电池的光电转换效率衰减，降低使用寿命。目前已有研究的钙钛矿电池平均使用寿命大约在 1000 个小时，而晶硅太阳能电池的使用寿命超过 20 年。尽管钙钛矿太阳能电池在制造成本上胜于晶硅电池，但其寿命短、稳定性差、天然不稳定、铅泄露风险，是产业化落地的最大掣肘。如何在保证光电转化效率的前提下提高材料的结构稳定性，是现阶段研究的重点。

在当前的钙钛矿研究中，有机-无机复合体系（FA/MA/Cs）PbI3 是最受关注的材料体系。2022 年，深势科技、北京科学智能研究院联合宁德时代 21C 创新实验室、宁波材料所采用深度势能方法在研究有机无机复合钙钛矿的微观结构方面取得重大发现，研究成果在 《AFM》 上发表[4]。

由于有机-无机复合体系的结构复杂性，晶体学表征方法受噪声影响非常大，对其结构的理解仍然非常不足。比如实验发现，随着温度降低，在相变点，FAPbI3 钙钛矿晶体学衍射谱上的噪音突然增多，谱峰难以标定。在常规体系中，随着温度降低，一方面，体系的对称性趋于降低，衍射峰会发生劈裂；另一方面，由于热扰动诱发的非谐震动减

少，衍射峰倾向于变窄。而高温立方相的一个衍射峰，随着温度降低至四方相，劈裂成的两个衍射峰，而在低温正交相再次"fuse"成一个模糊的宽峰。这一非常规现象被多篇文献报道，导致 FAPbI3 钙钛矿的低温真实结构至今未知。

深度势能方法兼具第一性原理精度和近经验势的速度。非常适合需要长时间分子动力学的低温物相问题。而对于有机无机杂化体系，由于分子旋转弛豫时间长达数十到数百皮秒，要想获得有机无机杂化体系的低温平衡结构，所需的分子动力学模拟时长又比多数无机体系多出数十到数百倍。该工作借助深度势能生成器 DP-GEN 训练了有机无机杂化钙钛矿 FAPbI3 和 MAPbI3 的 DP 势，通过分子动力学模拟研究了两种材料的低温构型。模拟发现，在低温下，有机-无机复合钙钛矿形成一种耦合畴结构，即在单畴的晶格中嵌套着分子取向空间分布不均匀形成的多畴。根据 Scherrer 公式，这种纳米尺寸的畴结构导致晶面间距非均匀分布，衍射峰变宽，甚至于原本劈裂的衍射峰"fuse"。

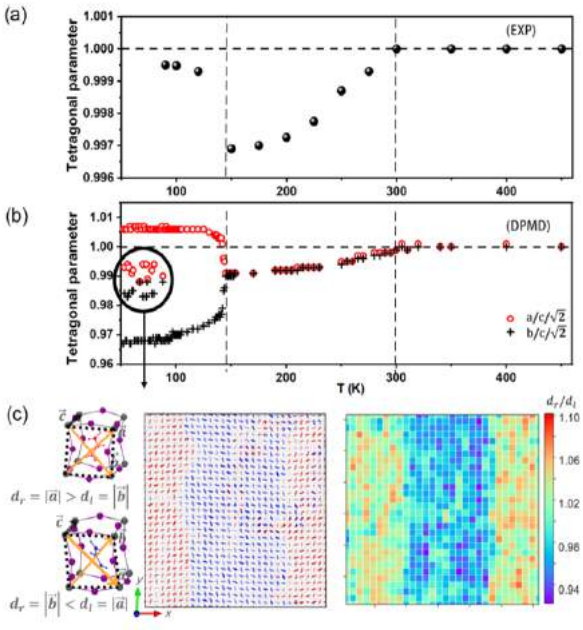


图. DeePMD 对钙钛矿相变的仿真预测与 x 光衍射实验高度一致

钙钛矿在工作环境中失效报废的过程，就是其微观组织结构的改变过程。

该工作首次探究清楚了低温下 FAPbI3 钙钛矿非常规的衍射现象与钙钛矿微观精细结构的对应关系，为解决钙钛矿光伏材料的稳定性问题提供了理论基础。

小结

在人工智能的推动下，我们可以预见，未来的能源材料研究将更加深入、准确，研发速度也将大大提高。这将对能源行业的发展和全球可持续能源的实现产生深远影响。

创新不是无源之水。目前中国在新能源领域已走到世界前列，除了正面攻坚科学难题，别无捷径。业界学界已有共识，新能源的瓶颈在于能源材料，而吃透材料的机理是实现高效理性设计的第一步，也是从"原始创新"走向"商业成功"的必经之路。

Source:

[1] Unraveling the Atomic-scale Mechanism of Phase Transformations and Structural Evolutions during (de)Lithiation in Si Anodes, doi.org/10.1002/adfm.202303936

[2] Self-Healing Mechanism of Lithium in Lithium Metal, doi.org/10.1002/advs.202105574

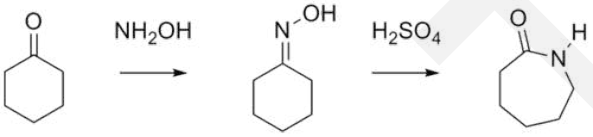
[3] A Deep Neural Network Interface Potential for Li-Cu Systems, doi.org/10.1002/admi.202201346

[4] The mechanism of Li deposition on the Cu substrates in the anode-free Li metal batteries, doi.org/10.1002/smll.202205416

[5] Spontaneous Hybrid Nano-Domain Behavior of the Organic–Inorganic Hybrid Perovskites , doi.org/10.1002/adfm.202301663

中国石化石油化工科学研究院：结合 AI4S 与化工催化场景

尼龙 6 是当今最重要的合成纤维之一,广泛用于纺织、汽车、电子等行业。尼龙 6 的单体 ε-己内酰胺是一种大规模的基础化学品,全球年产量达到 900 万吨。目前,95%以上的己内酰胺都是通过环己酮肟的液相 Beckmann 重排反应生产的。这个方法使用浓硫酸作为催化剂,反应条件极为恶劣,不仅腐蚀设备,还产生大量的废氨基硫酸盐,严重污染环境。气相 Beckmann 重排反应可以 100%利用原子转换成产品,不产生任何副产物,是绿色生产己内酰胺的理想途径。但是这个反应在工业中很难实现,因为反应条件非常苛刻,目前的固体酸催化剂活性不高,容易发生多种副反应, 导致己内酰胺选择性低。研究人员一直在寻找新的催化剂和反应条件,以实现这个反应的工业化。



Source: Wikipedia

中石化石油化工科学研究院有限公司的科研人员最近在《Nano Research》和《ACS Applied Materials & Interface》期刊上报道了一个重要进展。他们设计合成了一系列纯硅分子筛催化剂,系统地研究了催化剂组成、硅醇类型以及溶剂效应对反应的影响。研究发现,MFI 结构的 S-1 纯硅分子筛,具有大量硅醇巢式活性位点, not 与环己酮肟的相互作用最强,催化活性也最高。这一发现与他们使用密度泛函理论进行的计算机模拟结果一致。

为进一步验证这一点,研究人员又改变了溶剂的种类。计算结果显示,乙醇分子与硅醇巢式活性位点的相互作用能最低,实验也证明在乙醇溶剂中,S-1 分子筛催化剂的己内酰胺选择性可达 97%。这表明溶剂可以微调催化剂活性位点的化学环境,从而优化反应结果。这项工作首次系统地研究了纯硅分子筛催化剂在气相 Beckmann 重排反应中的效果,不仅揭示了硅醇巢式活性位点的优势,也展示了计算机模拟在预测和指导催化剂设计中的强大能力。这为工业上生产绿色己内酰胺提供了可行的新方法。该研究团队下一步的计划是继续优化催化剂,提高其稳定性,以实现这个反应的大规模商业化应用。

这项研究为提高其他类似的固体酸催化气相反应的选择性提供了参考。无论是在应用还是基础方面,都有重要的意义。它展示了计算机模拟在催化剂设计中的必要性,可以事先预测最佳的催化剂和反应条件,避免大量的试错,极大地提高研发的效率。这对于将基础研究的成果快速转化为应用,实现产业化具有重要意义。这项创新不仅可望推动绿色尼龙 6 单体的生产,也将对纤维、塑料等行业产生深远的影响。它为投资者提供了一个具有巨大市场价值的新技术方向。

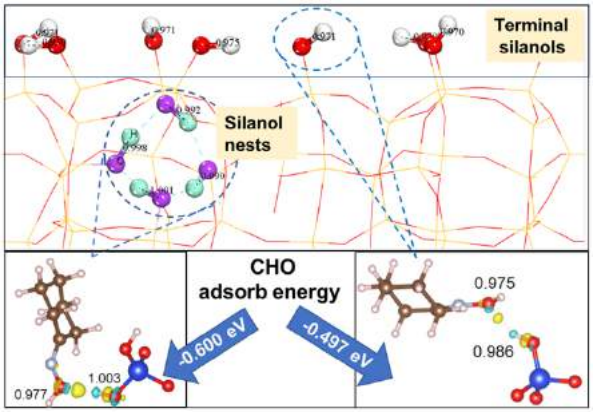


图. 终端硅醇和硅醇巢的理论预测 O-H 键长,以及 CHO 基团与终端硅醇和硅醇巢的相互作用

在此项工作中，研究团队使用了 DeepModeling 社区项目 -- 国产量子物理软件 ABACUS 来进行复杂的密度泛函计算，传统上这类工作一般由国外软件完成。ABACUS 在本项目中的成功应用为更大范围的自主可控提供了经验。同时，ABACUS 软件内置了 DeePKS (深度学习辅助泛函计算) 模组，该模组同为 DeepModeling 社区项目，由张林峰博士等领衔，代表着世界领先水平，也是 AI for Science 在亚原子尺度电子结构问题的优秀应用。

据悉，石科院团队也在探索自然科学大模型在该领域的应用前景。该团队应用 Uni-Mol (分子表征大模型）对新型材料 / 催化剂进行高通量筛选。Uni-Mol 发表于 AI 顶会 ICLR 2023，是全球首个基于 3D 分子构型的大规模预训练模型，可直接应用于大量下游任务，如分子性质预测、结构生成、分子对接等。最近，Uni-Mol 的模型框架了大幅升级，并在国际权威学术竞赛 OGB-LSC 的量子化学性质预测上获得了榜首。

团队表示：使用 Uni-Mol 等自然科学大模型，可以极大降低重复训练特定场景 AI 模型的时间和成本；同时，预训练大模型也保持了较高的精度，可以有效的对亿级以上的设计空间进行富集和初筛，对于催化剂等领域是非常令人兴奋的新工具。

中石化石油化工科学研究院有限公司夏长久博士表示："这是一项令人振奋的研究。新的 AI 技术与大大模型有机会帮助解决了人们对传统工业领域高温高压下复杂反应过程和机理认识的盲区，是未来工业技术研发的新范式和发展方向。我们与深势科技的合作，帮助我们对工业催化过程的关键基础科学问题展开攻关，为长期持续的工艺优化和催化剂创新提供理论基础。"

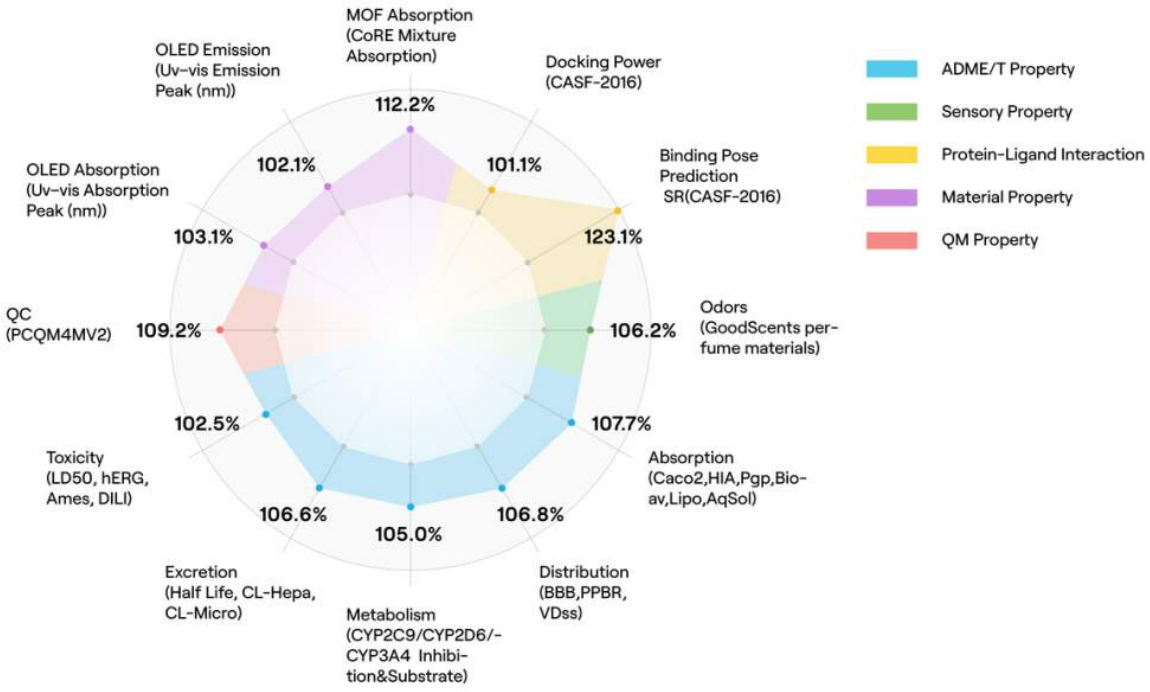
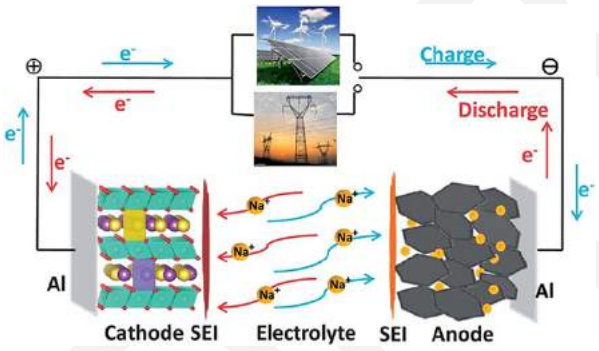


图. Uni-Mol 在各类任务上都表现优异，超越之前的最好方法。图中内部灰色区域为之前的最好方法的效果，外部多种颜色区域描述的是是 Uni-Mol 在多种任务上超出之前最好方法的百分比。

多氟多：AI4S 干湿结合，形成纳电掺杂问题科研生产力

钠在地球上的储量丰富，比锂的储量大得多。这意味着如果我们能成功研发出可靠的钠电池，那么电池的供应将会更加稳定，价格也可能更低。钠电池的理论能量密度和锂电池 (磷酸铁锂) 有重叠，这意味着它们有潜力提供类似的电池性能。然而，钠电池也面临着许多技术挑战。钠的离子半径比锂大，这使得钠在电极材料中的扩散速度较慢，而且能量也较低。（上面的一条不是原因，锂的电位本身就比钠低）这就意味着钠电池需要在更高的电压下运行才能达到与锂电池相当的能量输出，但这又会带来电池结构稳定性的问题。高压下，电极材料的结构可能会发生变化，导致电池性能下降（脱钠或锂过程中，结构都会变，但不是性能下降的原因）。同时，钠也有可能与电解液反应，形成"死区"，这也会损害电池的性能。



钠电池体系示意图 Source: fchem.2020.00635

为了克服钠电池的技术挑战，科研人员正在专注于电极材料的"包覆"和"掺杂"技术，同时使用高通量筛选来加快材料研发的速度。

1. 包覆（Coating）：在电极材料表面上加一层保护膜是一种常用的策略来改善电池性能。这层包覆材料可以防止电解液直接接触活性材料，从而避免不良化学反应，如电极材料的溶解或锂离子的不可逆沉积。这种保护膜可以有效提高电池的循环稳定性和安全性。针对钠电池，科研人员也在研究使用包覆技术来提高电池性能。

2. 掺杂（Doping）：掺杂是一种改变材料性能的方法，通过将少量的其他元素添加到材料中，可以改变材料的电化学性能。例如，掺杂可以提高离子的扩散速度，提高电极的电导率，或者改善材料的结构稳定性。在钠电池的研究中，掺杂也是一种重要的策略来优化电极材料。

3. 高通量筛选（High Throughput Screening）：在新型电池材料的研发过程中，科研人员需要测试大量的材料组合以找到最优的配方。高通量筛选是一种使用自动化设备和数据分析技术，快速测试大量样品的方法。这种方法可以大大加快材料研发的速度，并帮助科研人员找到最优的材料和工艺参数。

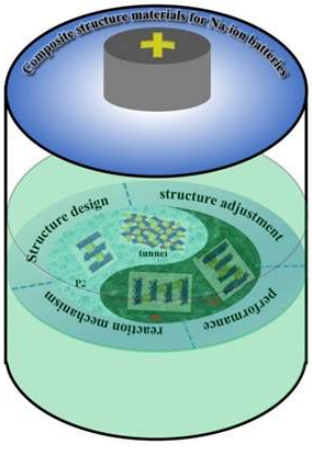
一般来说，高通量筛选分为 "干"、"湿"两种模式。"湿"模式即高通量实验，通过自动化机器人对大量配方进行合成、测试，并收集数据。这种模式成本高，精度好。"干"模式即虚拟筛选，通过数字化建模，对不同配方进行仿真计算，预测其性能。这种模式的效果取决于数字化建模和仿真的 "效率" 和 "精度"。如果精度低，则筛选过程约等于胡乱猜测，无法达到"富集"的目的；如果"效率低"，则筛选过程难以规模化，同样面临成本和时间的挑战。多氟多认为，AI for Science 能提供强大的虚拟筛选能力，有效兼顾效率和精度；进一步将 AI for Science 与实验进行干湿结合，则可在新钠电材料与工艺开发中实现生产力的提升。

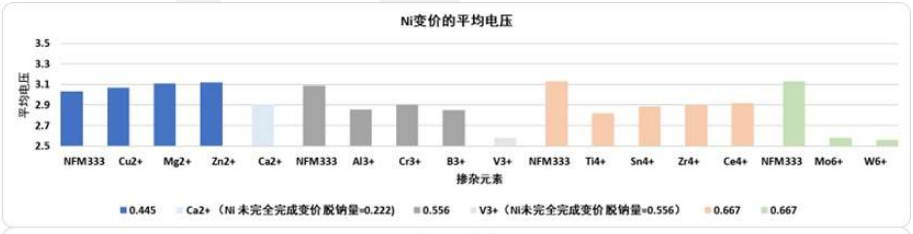
多氟多企业代表表示：电池作为当下热门领域，先进算法和行业实际应用的结合也同样重要。多氟多

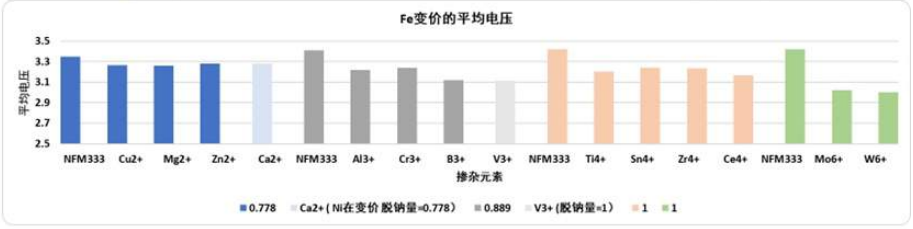
与 AI for Science 引领者深势科技合作，利用深势科技的电池自动化设计平台 —— Piloteye，结合创新算法和行业 know-how，打破过去电池研发中计算与实验交流的鸿沟，除此之外，操作便捷及用户友好的 Piloteye 也能很好地降低研发人员使用传统计算模拟类软件的门槛，让研发更聚焦于行业的关键问题，形成企业研发所需的体系性生产力。

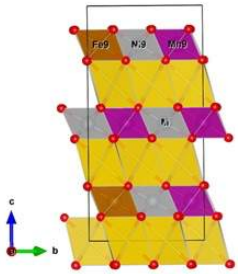
深势科技高级研究员、材料研发总监王晓旭博士表示：当前电池研发领域面临着"研发周期长"、"成本高"、"创新难"的瓶颈。通过"实验试错"手段开展的电池研发，需要经过多个环节，从新材料开发到化学体系整合到电芯样品设计制造测试到电芯产品大规模量产，整个研发周期可能需要数年时间，需要大量的资金投入，包括设备、人力、原材料等。虽然计算手段在电池研发领域应用日益增多，但是"传统计算模拟"手段仍面临着许多挑战，最明显的则是"计算效率和计算精度难以兼顾"，这成为计算方法在模拟电池工作真实场景、反映电池工作真实问题的瓶颈。随着市场对电池设计的要求越来越多样化，而电池研发的痛点难以快速被解决，电池行业内"卷"的生态实在无可避免。

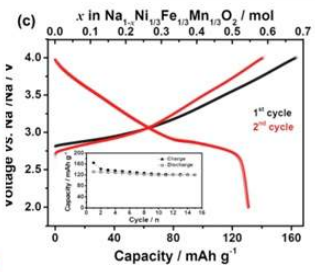
要解决"卷"的生态，则需要通过创新技术和加强合作来实现。王晓旭博士表示，在 AI for Science 新范式的发展下，利用人工智能等新技术来突破电池研发的难点，开发更高效、更可靠的电池设计自动化平台成为电池研发行业的重要趋势。深势科技 Piloteye（电池设计自动化）平台基于多尺度模拟算法的突破、算法工程化的实践和产品开发能力，可以更快速、精准地完成电池的设计和研发，持续提升电池研发的创新效能。











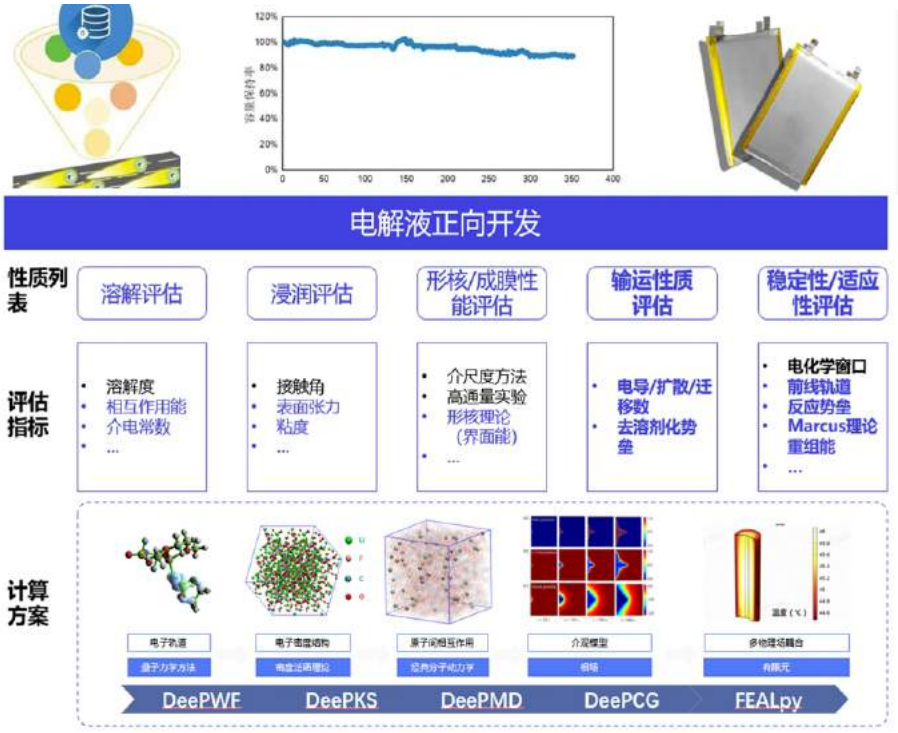
Step	电子转移	价态分布
NFM333		9Ni2+/9Fe3+/9Mn4+
NFM333脱1/3Na	9Ni2+→9Ni3+	9Ni3+/9Fe3+/9Mn4+
NFM333脱2/3Na	9Ni3+→9Ni4+	9Ni4+/9Fe3+/9Mn4+
NFM333脱3/3Na	9Fe3→9Fe4+	9Ni4+/9Fe4+/9Mn4+

多氟多+深势科技：高通量计算研究参杂和相变

金羽新能：AI4S 驱动高通量筛选工作流

锂金属电池是使用锂金属作为负极的新型电池，具有能量密度高的特点，有望解决新能源汽车的"里程焦虑"。但由于在循环过程中存在锂枝晶生长、体积膨胀、SEI 膜不稳定等困难和问题，目前其商业化、产品化进程受到阻碍。

近日，国内电池制造新势力金羽新能宣布，其 440Wh/Kg 高能量密度电池下线。据悉该电池标称容量 10Ah，产品循环寿命超过 300 圈。作为深势科技的战略合作伙伴，此次金羽新能节点性电池的下线，标志着 AI for Science 在工业应用领域的又一次进阶。



金羽新能+深势科技：电解液正向开发工作流

金羽新能表示：公司研究团队始终聚焦在新型电池前沿技术，全力把技术转化为新能源产品。其中，锂金属电池就是一个重要的探索方向。金羽新能联合深势科技，面向锂金属电池电解液/电解质优化搭建的工业设计平台，将电解液/电解质的实验和计算充分配合，形成"计算设计与实验验证循环优化"的先进模式。

值得注意的是，此次金羽新能 440Wh/Kg 高能量密度电池的成功下线，正是依托深势科技工业级电解质配方设计与优化的多尺度建模、机器学习、高性能计算和高通量筛选的核心技术，突破计算模拟到实验需求的瓶颈，促进电解质配方的靶向设计。这极大地提升了锂金属电池电解液/电解质的研发效率。基于此，新型电解液/电解质快速迭代，使锂金属电池的性能有了有效提升，商业化、产品化进程得到大幅推进。而作为《锂金属蓄电池及电池组总规范》的主要起草单位，金羽新能前期开展了大量相关基础探索工作。

基于 AI for Science 在电解液研发的工作流，金羽新能也申请了相关专利。该方法包括：对电解液体系所含的成分进行结构优化，获取成分初始构象在能量最低时的总能量及分子结构信息;获取电解液体系的密度; 对每个原子指派与初始构象匹配的力场形式;构建电解液体系模型并进行初始结构优化;随后进行分子动力学计算，对能量稳定后的体系中各种粒子的构象进行统计，得到粒子的径向函数值等数据;选择电中性-局部溶剂化结构分别进行 DFT 计算，得到相应的 HOMO/LUMO 能级;获取 Li 粒子均方位移以及电解液扩散系数;分析预测电解液性能。该方法能够使模拟结果具有较高的准确性，且模拟过程所需时间较短，成本较低。

英矽智能：端到端 AI4S 实现 "First-in-class" 药物的高效研发

特发性肺纤维化（IPF）是一种导致肺功能进行性、不可逆转下降的慢性肺部疾病，在全球范围内存在大量的未尽医疗需求，随着病情进展和肺部损伤的加重，患者呼吸功能不断恶化，严重者最后可致死亡。目前已有的药物吡非尼酮和尼达尼布在临床上存在着毒副作用大等众多问题。英矽智能基于生成对抗网络、深度强化学习、预训练模型及其他机器学习技术，构建高效的人工智能药物研发台 Pharma.AI，其中包括靶点识别和发现引擎 PandaOmics™、分子设计和生成引擎 Chemistry42™和临床实验结果预测和优化平台 Inclinico。利用人工智能的端到端平台，英矽智能从 IPF 适应症出发，利用 PandaOmics™发现了 Target X，进而利用分子设计和生成引擎 Chemistry42™针对该全新的靶点进行化合物生成，仅仅在合成了 78 个化合物后，成功提名了 ISM001-055 为临床前候选化合物（PCC）。2021 年 11 月，ISM001-055 在澳大利亚进入临床 0 期试验阶段，并完成首例受试者的微剂量给药。ISM001-055 是全球首个由 AI 发现具有全新靶点和全新分子结构的候选药物。历时 18+9 月从靶点发现进入到澳大利亚临床试验中，2022 年 02 月，新西兰启动了 I 期临床试验，目前已顺利完成单次给药的爬坡试验，并取得了良好的安全性和耐受性结果。2022 年 05 月，ISM001-055 已获得中国国家药品监督管理局药品审评中心（CDE）1 类新药临床试验默示许可，获批在中国进入 I 期临床试验阶段。2022 年 07 月，ISM001-055，已完成在中国 I 期临床试验中的首次健康受试者给药。临床前数据显示，该候选药物显著改善了肌成纤维细胞的形成，而控制肌成纤维细胞的形成有助于减缓纤维化的发展。该项目是全球范围内首个由人工智能发现的全新靶点全新小分子的临床阶段小分子药物，也是该靶点目前唯一一个临床阶段的在研管线，是真正意义上的"first in class"药物。

晶泰科技：AI 药物发现+自动化实验

人工智能正在与生物医药行业进行深度融合，随着 AI 技术的迭代，不管是靶点发现，药物分子发现，乃至临床研究，越来越多公司正在致力于让药物研发更加数字化和智能化，越来越多人工智能参与的管线也将进入临床或相继面世造福患者。

作为以智能化、自动化驱动药物研发的引领者，晶泰科技开创了智能算法、自动化实验与专家经验相结合的智能化自动化药物研发平台（晶泰智药），包括小分子药物发现 ID4Inno™和抗体药物发现 XupremAb™两大特色技术平台，以量子物理、人工智能、云计算及大规模实验机器人集群等前沿技术与能力，助力药物研发走向"计算密集型""自动化密集型"，让药物研发更智能，让生物创新触手可及。

近期，晶泰科技与礼来就某个未披露的创新靶点展开 AI 小分子新药发现合作，利用晶泰智药小分子药物发现平台 ID4Inno™ 研发原创新药，填补未满足的临床用药需求，该合作预付款及里程碑总收益可达 2.5 亿美元。

但 AI 与生物医药的融合并不是一蹴而就的，这是一个系统性建设过程。AI 的核心要素有三个，即算力、算法与数据。其中，数据需要长期积累，也是重要的基础设施，目前 AI 制药还存在"数据鸿沟"，按照传统模式内生的数据，还无法满足机器学习的需求。因此建设创新的、能够生产数据的基础设施，让有价值的数据能够更有效、更便利地被获取，药物研发才能真正走向智能化。

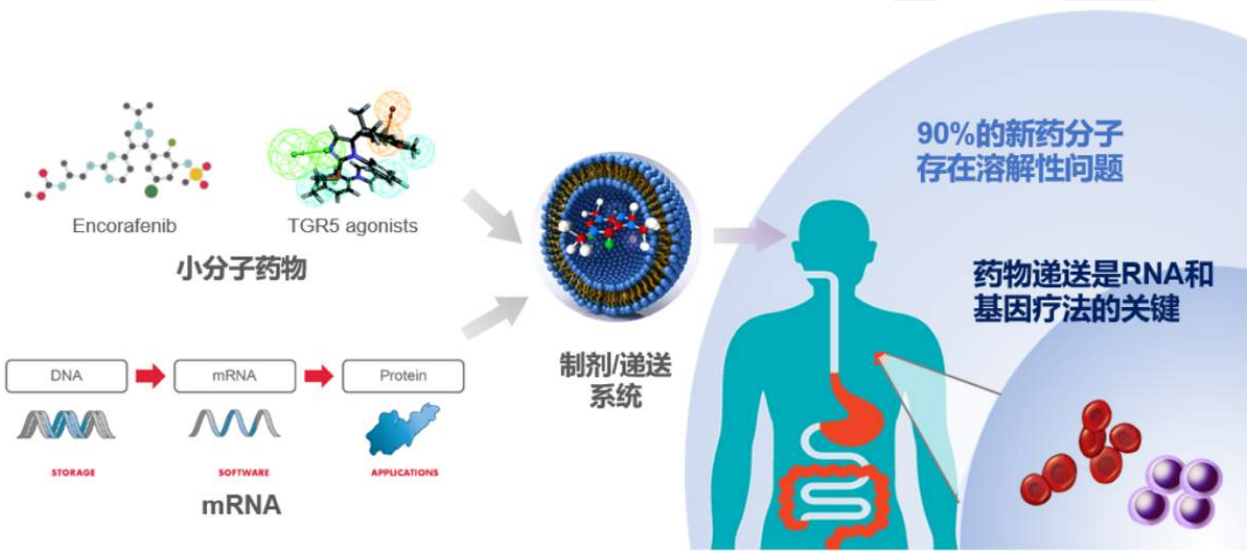
晶泰科技早在 2019 年便开始布局自动化实验室，致力于形成 AI 及自动化的数据闭环，应用于小分子药物发现及抗体药物发现领域，提升化学合成，固态化学研究等领域的研发效率及标准化程度，同时反哺 AI，提速人工智能的发展进程，助力科学创新。

作为晶泰科技旗下专注以 AI 与机器人技术助力科学创新的专属品牌，晶泰智造期望通过构建高效智能研发体系，建设可提高数据生产通量的工具，赋能自动化实验室新基建，以更高效、精确、智能的方式为科学家提供稳定可靠的实验支持。现有独立自动化工作站、桌面型智能仪器、定制自动化解决方案、自动化实验室集群平台等。通用型自动化产品：XmartChem™ 智能合成工作站、XtalComplete™ 智能结晶工作站、ChemPlus™ 桌面型固体加样仪；按需定制自动化解决方案：功能材料合成表征自动化系统、食品农残检测全流程自动化系统、药品质量控制自动化系统等。

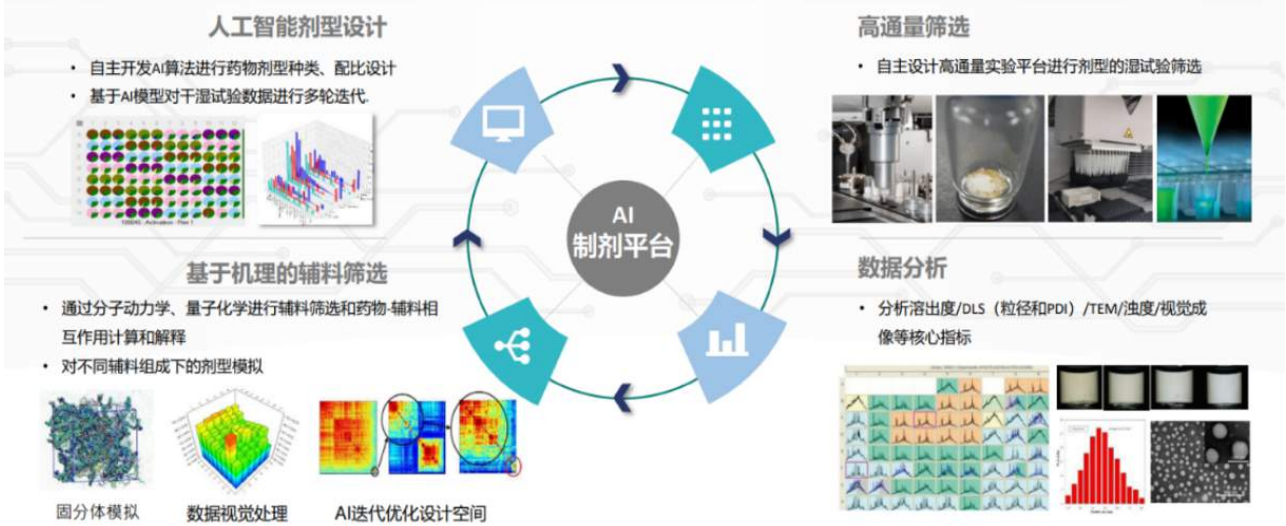
AI 和自动化的相互促进与迭代，将有利于实现 AI4S 数据闭环，打通临床前药物研发的各个环节，不仅仅为制药，也为化工、检测、功能材料、生物基材料等行业打好地基，促进科研创新。

剂泰医药：AI+药物递送

剂泰医药是以人工智能（AI）驱动、利用精准靶向的药物递送和药物发现技术，为患者开发更有效的治疗药物的创新型生物技术公司。目前，剂泰医药已同时布局 RNA 药物和小分子新药，致力于以"AI+药物递送"推动全新或同类最佳药物的研发，以及对现有疗法进行改良，更高效、更精准地为全球患者开发更有效的治疗药物，满足其未被满足的医疗需求。 剂泰医药聚集了 AI 数据驱动算法、机制驱动的量子力学和分子动力学模拟以及高通量实验平台等底层技术，最大化运用各项交叉学科先进技术，实现高效、安全的多器官靶向药物递送，以此驱动规模化药物发现。

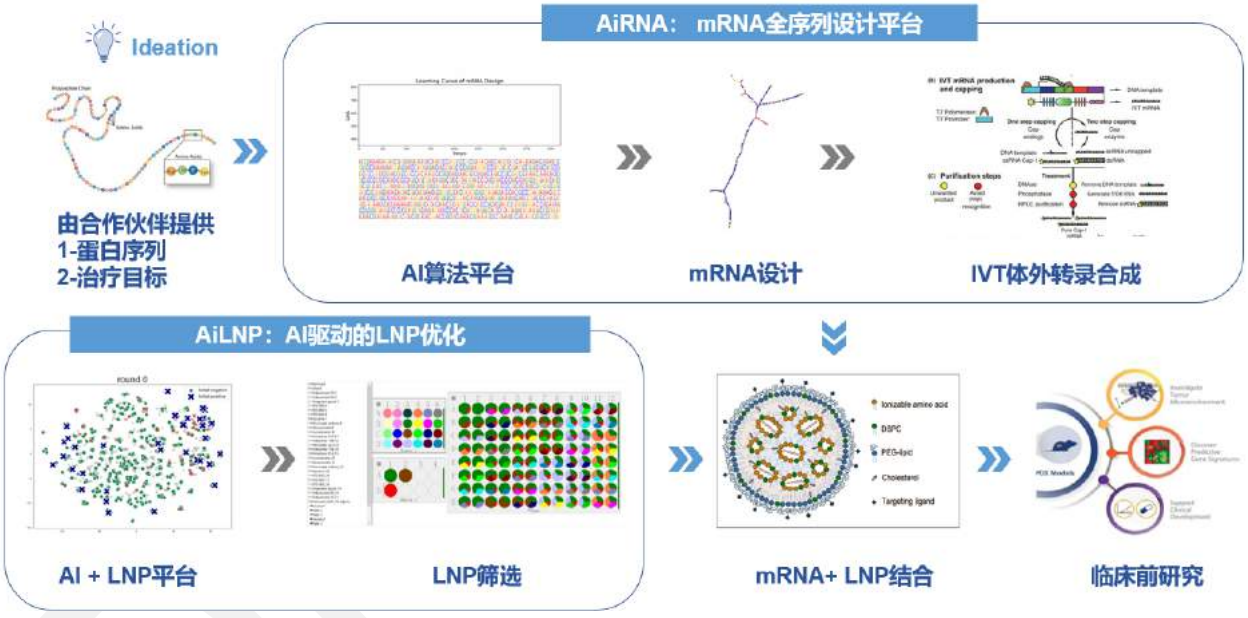


基于上述底层技术，公司搭建了自有的 AiLNP（AI 驱动核酸递送系统设计平台）、AiRNA（AI 驱动 mRNA 序列设计平台）、AiTEM（AI 驱动小分子制剂设计平台）核心技术平台，以预测在特定微环境下的小分子和核酸药的理化和生物特性，以 AI 驱动干、湿实验迭代，实现了更有效的创新递送材料设计、核酸序列设计与优化，并设计更具优势的制剂配方和核酸递送系统。



在小分子新药研发中，90%新药为难溶药物，药物递送技术可以实现对现有药物的制剂改良，从而更快、更高效地研发出毒性更低、疗效更佳的小分子新药，提高患者依从性。剂泰医药通过 AiTEM 平台，以自主开发的 AI 算法进行药物剂型种类、配比设计，并基于 AI 模型对干湿试验数据进行多轮迭代，以实现药物研发效率的大幅提升。

对于核酸（RNA）药物研发领域，因其成药靶点数量更多、药物设计更快、靶向特异性更高，有望成为继小分子药物和抗体药物后的第三大药物，引发新一代制药行业科技浪潮。剂泰医药建立了世界领先的端到端 mRNA 药物发现平台——AiLNP 和 AiRNA 平台，可将药物递送系统效率进行指数级优化。AiRNA 以 AI 算法驱动 mRNA 密码子优化、二级结构优化及蛋白折叠效率优化，并打造专利 UTR 库, 增强蛋白表达。通过标准化的工作流程，该平台确保了高质量的 mRNA 合成。AiLNP 平台则构建了具备自主知识产权的 mRNA 序列设计、LNP 递送系统设计一体化的 RNA 药物开发能力，实现了业内较难实现的肺、脾、肌肉、神经等多器官、组织的靶向递送。



通过持续、深入地探究 AI 在药物递送和药物发现领域的应用，剂泰医药正在将全新的 AI 技术与传统药物开发相结合，更高效、更精准地推动创新药物的研发。同时，运用核心技术平台赋能医药生态圈，加速革命性药物的研发进程，为未来带来更多的技术变革和更新，实现医药行业的持续进步。

未知君：LLM+微生物基因组

从 2007 年人类微生物组计划开始，我们对于肠道微生物跟人体的健康和疾病的关联了解越来越多，也越来越了解到肠道微生物的复杂性，人体肠道微生物包含了 10 倍于人体的细胞数量和 100 倍于人体的基因数量。在这样的一个数量和复杂度上，越来越多地要依赖于大数据和人工智能来进行研究。目前 AI 在肠道菌群和微生物组的研究中的应用主要在三个方面，第一，是疾病风险预测建模和诊断的生物标志物发现；第二，是分析疾病治疗和干预的靶点以及有相关治疗功能的微生物和微生物代谢产物的发现；第三，是微生物基因和代谢通路功能的学习和预测，以及通过生成式人工智能进行功能基因及其产物和代谢通路的设计和优化。

通过分析疾病病人和正常人的肠道菌群的宏基因组和其他组学数据，可以分析疾病相关的标志物（包括菌种或者菌株信息，宏基因组信息，宏代谢物信息）从而开发相关的疾病诊断工具和疾病风险预测模型，近年来已经有很多的文章报道，通过 AI 和跨队列数据的方法，建立起数十种疾病的风险预测模型，甚至还包括对于一些药物跟肠道微生物作用之后的疗效预测等等。在未知君，我们也通过不断搜集大量的公共数据和自建数据集，进行健康人肠道菌群基线的研究，并建立起很多疾病的风险预测模型。值得一提的是，我们目前也开始采用大模型和多模态的人工智能算法，把更多维度的肠道微生物组学数据应用到模型中，从而更好地提高算法的精度，随着该领域数据的不断增加，模型的改进速率也会加快。

同时，通过深入分析疾病和正常人肠道菌群的区别，我们可以从菌株和菌种层面预测某些菌株或者菌种对人体健康和疾病的作用。这一方法用到了未知君的药物研发的菌株提名工作中，通过综合数据分析，以及文献挖掘建立起来的知识图谱，可以更快和更准确地找到关键菌株，从而极大提到活菌药物的开发效率和极大减少实验验证数量。

目前大模型和语言模型发展非常迅猛，微生物的基因组其实也是一个标准的字符串，通过语言模型的引入，我们可以通过 AI 更好地学习基因组里面一些底层的信息，从而更好地预测和挖掘基因的功能信息，代谢通路信息，代谢产物信息等等。在未知君，我们通过语言模型对微生物基因组进行功能预测，比传统的仅仅通过序列相似度预测有更高的精确度。随着我们对于基因功能和代谢通路的模型更加精确，以及结合蛋白质设计等等新的 AI 工具的使用，下一步可以根据相应的功能和功能产物去设计相关的序列，从而更进一步打开利用微生物功能造福人类健康和疾病的大门。

德睿智药：AI 加速药物发现

德睿智药是一家专注于使用人工智能驱动新药研发的公司，希望通过推动人工智能和新药研发领域多种前沿技术渗透融合，持续输出兼具差异化和高临床价值的候选药物，从而让更多生命重获健康。基于药物化学、计算化学、生物学等多学科知识，德睿智药打造了 AI4S 工程化产品，包含 Molecule Pro, Molecule Dance、PharmKG 三大 AI 药物研发平台。在加快药物发现速度、降低药物研发成本之外，德睿智药致力于利用新一代机器学习的方式，去解决一些难以解决的问题，例如难成药靶点和全新靶点的药物研发。

目前，AI 平台的核心机器学习模块表现已大幅优于国际领先软件以及领域内当前最优（State-of-the-Art）模型，并在数十个新药研发项目中得到充分验证，可为小分子药物临床前研发节省约 70%的研发时间及成本，相较传统方式极大提高新药研发效率。利用一站式 AI 药物研发平台，德睿智药首个自研 GLP-1RA 小分子口服药物 MDR-001，从项目启动到获得 IND 批件仅用时 19 个月，其中在 8 个月内仅合成验证不到 100 个分子即获得具有 Best-in-Class 潜力的临床前候选化合物（PCC）。目前，MDR-001 已实现中美双获批临床，针对肥胖适应症和二型糖尿病，并于 2023 年 6 月成功完成了临床 I 期首例受试者给药。

目前，许多 AI 技术已被应用于生物制药领域，这像是制药业的一项工业革命，而现阶段正是这场革命的开始。德睿智药已在多个 *Nature*，*Cell* 系列子刊及 ICML 等国际顶刊及会议上发表数十篇论文，涵盖对以上 AI 技术的落地应用与改进。比如公司团队发表在 *Nature Biomedical Engineering* 上的研究建立了一种基于 AI 融合多维分子信息的虚拟筛选算法，为快速筛选药物先导化合物提供了全新策略。发表在 *Advanced Science* 上的论文提出了首个基于蛋白质动态信息的预训练模型 ProtMD，该模型首创引入蛋白质动态空间信息，在药物-蛋白亲和力预测任务等多个下游任务的表现上大幅超越目前最优模型。发表在 Nature 系列期刊 *Communications Chemistry* 上的论文则介绍了一种多视角异构图神经网络模型 PharmHGT，通过首创引入官能团信息，该模型在下游分子性质预测任务上达到目前最优表现，也为该领域的研究提供了一个新的 Benchmark。

如今在 AI 技术加持下，很多 AI 制药公司在与传统药企的合作项目中交付了令人满意的成果，大规模的合作以及复购正在被 AI 制药企业实现。德睿智药正与某跨国药企、多家生物制药公司深化 AI 技术合作，已收获总额过亿美元的药物管线合作订单，并成功完成了对多个上市药企的 PCC 交付。

AI 辅助药物研发是一个循序渐进的技术进化和选代过程，类似于 AI 大模型的技术在未来还有非常大的潜力，比如 AI 有望更高效地将海量的机器语言转换为研发人员可以理解的自然语言，从而有望更加高效地辅助靶点发现等下游任务，协助攻克一些业界难题，比如能否通过知识图谱更好地理解复杂疾病的生物学机理、发现更优更通用的生物标志物和靶点通路。德睿智药希望基于专家经验和自有数据，加速 AI 平台的大模型迭代升级，从而把一些更好的药物更快速地研发出来，同时可以在人体上面有更好的疗效和更高的临床成功率。

青云瑞晶：结构解析

苏州青云瑞晶生物科技有限公司是一家专业的 CRO（研发服务提供商），公司在结构生物学技术服务，药物固态研究和新材料结构表征，三个业务板块，为客户提供最先进的解决方案。公司自主开发了国际领先的 MicroED 结构解析平台，结合冷冻电镜单颗粒（CryoEM-SPA）及同步辐射单晶 X 射线衍射（XRD），为客户解决难点问题。 现代新药研究与开发至关重要的第一步是寻找、确定和制备药物筛选靶标—分子药靶。得到靶标蛋白结合位点的形状和电荷特征信息，测定并解析蛋白质-配体复合物的晶体结构，从而获得二者相互作用的信息。这些信息可以在保持先导化合物框架的同时，对药物分子进行结构优化，从而提高药物的活性和选择性。而在这一过程之中，将会运用到一个十分重要而基础的学科——结构生物学。

根据蛋白/蛋白-配体复合物能否结晶为初步的划分依据，不同的蛋白可以使用不同的结构解析手段进行研究。X 射线衍射技术作为最传统的结构解析技术，方法成熟且成本低。但对晶体的尺寸具有一定的要求，10 微米是能够获得衍射数据的极限尺寸。若晶体小于 X 射线要求尺寸，可以尝试 MicroED 的方法。其入射光束为高能电子，与晶体作用更强。因此，只需要少量微纳米级的晶粒就可以快速、高效地获得高分辨率的衍射数据，大幅降低了对样品形状、纯度和尺寸的要求。MicroED 被认为是化学、合成、生物等领域底层的颠覆性工具，入选《Science》2018 年度全球十大科技突破。「青云瑞晶」创始团队来自瑞典斯德哥尔摩大学 MicroED 技术发明课题组，在 MicroED 技术的应用上具有丰富的经验。若蛋白无法结晶，则可使用 Cryo-EM SPA（冷冻电镜单颗粒法）。这是一种电子显微技术，用透射电镜进行图像采集，最后对收集的成千上万张图片进行三维重构，从而得到目标蛋白的近原子分辨率结构数据。

2022 年，布宜诺斯艾利斯大学的 Claudio N. Cavasotto 发表文章，对比了 22 个靶点蛋白使用 AlphaFold 预测的结构和实际的结构，所有的蛋白预测结构几乎都与实际结构存在偏差。而在另一项使用结构信息进行虚拟对接的实验中，即使一些预测的蛋白结构的主链或侧链与实际相比直发生一些十分细小的变化，反映在虚拟对接的评分结果中，都变成了巨大的分数差异。当前的药物设计策略，需要高质量的数据结合先进的算法。通过实验验证的药物靶点结构能够保证数据库的真实和准确性。真实的数据配合优越的算法才能设计出真正有效的新药。青云瑞晶在结构生物学服务上拥有全面的技术手段，丰富的实验经验和完善的一站式服务体系，可以满足药物研发阶段对靶点结构研究的所有需求。

中国人民大学高瓴人工智能学院

孙浩、长聘副教授、博导 | 智能科学计算：知识嵌入与知识发现

面向复杂动力系统（例如流体、大气系统、电磁系统等），开展高效且精准的科学计算（例如建模、仿真、数据同化、定律探索等），已经成为现代科学研究、工程设计和优化的重要工具。经典数值方法须已知完备的控制方程和初始/边界条件，虽然时效性长，但在解决复杂高维系统科学计算问题时，特别是数据同化、参数化建模与模型不确定性量化、求解反问题等任务，面临计算资源消耗大、效率低的瓶颈问题，同时无法克服因模型假设带来的误差。常见的人工智能方法则需要依赖大量训练数据，还存在可解释性差、泛化性弱、误差不可控等基础科学问题，亟需解决。

在科学计算领域，复杂动力/动态系统的状态演化通常可由一组线性或非线性的微分方程描述（如 ODE、PDE），先验知识的完备程度归为完备、半完备、未知三大类（如上图所示）。因此，可将面向复杂动力/动态系统智能科学计算方法分为"知识嵌入"和"知识发现"两个维度。需要解决的关键科学问题包括：1）在"知识嵌入"维度，如何将机制/机理/知识有效嵌入机器学习，实现可解释、可通用、可扩展、可泛化、低数据依赖性的科学计算模型，同时兼顾不确定性量化，实现对高维复杂动力/动态系统的高效准确仿真和数据驱动建模；2）在"知识发现"维度，如何利用有限数学算子作为基础符号单元，建立简明、易计算、表示能力强的数学方程，构建高效的自动化搜索策略，实现高效率、自动化符号回归，从数据中自动发现用于描述复杂动力/动态系统状态演化的物理定律或控制方程，揭示复杂动态系统所蕴含的变量逻辑关系、潜在动力机制、非线性规律和时空演化行为本质。

因此，将数据驱动的机器学习与机制/知识驱动的符号计算融合，深入开展新一代可通用、可解释、可扩展、可泛化、低数据依赖性的智能科学计算研究，实现"面向科学计算的机制融合机器学习"基础理论、创新方法和先进技术的突破，势在必行。通过该交叉研究，推动"面向数据驱动科学索的可解释人工智能"领域知识融通发展为完整的交叉知识体系，为跨学科领域应用（如实时短龄天气预报、多物理场模拟、湍流仿真等）提供新的技术支撑。

北邮网络与交换技术全国重点实验室

北京邮电大学"网络与交换技术全国重点实验室"（简称"实验室"）是我国信息科技领域的国家战略科技力量，也是我国信息科技人才的重要培养基地。在网络通信领域新理论研究、关键技术突破、重大应用示范、专业人才储备等方面处于全国领先水平。

AI for Science（AI4S）代表了一种新兴的研究范式，该范式旨在通过更强大的人工智能方法来加速或优化人类对物理世界的模拟、洞察、新发现或控制。现有的 AI4S 研究主要集中在计算层面，然而在感知、网络以及信息系统层面仍然存在许多挑战。实验室利用自身的专业积累和特色优势，组织对 AI4S 关键问题进行前瞻性研究并攻克核心技术，包括构建基于智简网络架构和通信支持的 Science for AI（S4AI）系统，例如"感知-通信"一体化网络，信息网络安全与隐私计算的系统，从而更好地建立人工智能的"世界模型"。此外，通过更强大的信息智能基础设施，团队开展 AI4S 在通信网络的前沿交叉研究工作。实验室在张平院士带领下，提出并引领了"语义通信"相关工作，致力于挖掘通信和网络引入人工智能所带来的增益，达到熵减（信息熵和系统熵）意义下的系统性能最优。通过探索信息在更高维度层面的生成、传输和变换机理，在网络内部实现语义承载模型的分发、更新和演化；提出了智简网络的云端、边缘和终端三层协同架构，并支持模型更新和演化后的自动分发。工作可用于解决极端环境下通信能力受限导致的难题等。

在生物医学方面，AI4S 的应用更是广泛且深入。以实验室近年来成立的"信息智能与数字健康研究所"为例，团队开展了生物启发计算的基础理论（Science for AI，S4AI）研究，提出认知科学启发的智能语义计算方法，探索多模态感知的语义空间结构，揭示与人类专家认知的关联映射。此外，基于更强大的智能信息处理方法，构建多模态基础大模型等来加速生物医学的新发现，在病毒进化预测、AI 抗体药设计以及个性化用药方面取得了引领性成果。例如，王光宇研究员团队提出了基于大规模预训练模型的蛋白质功能分析新框架 UniBind，将蛋白质表示为残基和原子级别的图结构，通过 BindFormer 模块来提取几何和能量信息，并采用多任务学习来融合海量、异构的生物实验数据，实现蛋白质间相互作用（PPI）的准确预测，并用于预测病毒可能的变异和演化路径（Nature Medicine 发表）。

总之，AI4S 正在引领人类进入一个全新的科学研究时代，无论是在通信网络领域还是生物医学领域，AI 都提供了加速我们理解世界、解决问题的新工具。

浙江大学材料学院

洪子健，浙江大学百人计划研究员、博士生导师

大语言模型和材料科学的第一个结合点是基于大语言模型的材料类文献综述和论文编辑校正。相对于传统的文献综述，大语言模型框架下的文献综述因其具有较大的文献数量和强的总结能力，但同时也存在对相关术语不够熟悉和缺乏高质量数据集等问题导致训练效果不理想。另一个风险就是现在学界对于基于大模型的论文撰写和校正存在一定的争议，而相关的业界规范和立法约束缺失，相关的学术伦理风险需要格外注意。

第二个方向是利用大语言模型辅助计算材料学。大语言模型具有强大的编程和作图能力，而这种能力也是计算材料学不可或缺的。在大语言模型训练集中也不可避免的采用了大量材料计算软件的应用手册以及文献中关于材料计算的具体参数设置，为大语言模型在计算材料学中的应用提供了方便。例如，浙江大学洪子健研究员探究了 ChatGPT 在计算材料学中的应用和前景展望，提出了 ChatGPT 可以在三个层面上辅助计算材料学，包括：原子结构构建，编写计算程序和绘图代码等（Z. Hong, ChatGPT for Computational Materials Science: A Perspective, Energy Materials Advances, 4, 0026, 2023）。他以锂金属表面模型的构建，硅和二氧化硅能带计算代码的编写和三维箭头图的 Matlab 程序编写为例展示了 ChatGPT 的能力。在 GPT-4 开发出来前，生成的内容尚不能够满足实际的需求。例如锂金属表面模型构建时，只能获得 16 个随机分布的锂原子团。能带代码也是错漏百出，编造出了许多软件本身没有的指令集。而且只会将硅生搬硬套到二氧化硅。在 GPT-4 开发出来后，他发现 ChatGPT 在经过人类"教育"后，可疑获得正确的锂金属模型，同时能带计算的代码错误明显减少，而且还知道在代码中明确硅和二氧化硅的具体区别。他还总结了使用 ChatGPT 这些大语言模型工具的优势：(1) 工具已用，对计算材料领域的初学者友好；(2) 在领域专家的指导下，它能够从交流中学习到领域知识并纠正错误，从而正确地执行任务，并且其迭代速度很快；(3) 即使其目前尚不知道确切的答案，也可以给我们一些提示或建议。当然，文章也提到了这些工具的弊端，例如：(1) 其输出结果取决于版本，无法保证结果的一致性。(2)如果没有经过合适的训练，它可能会在代码中犯下非常简单的错误。因此，现阶段人为干预和仔细检查依然是必须的。(3) 应该关注科学伦理问题。在撰写论文时，一些大学和杂志社对使用 ChatGPT 等大语言模型有特定的政策/禁令。

第三个方向就是大模型与语音识别和高通量实验的结合。一个全自动的开发过程将包括语音命令和识别，大语言模型结合高性能计算平台的计算，再结合高通量自动实验机器人将模拟结果进行实验验证和反馈，最后对结果进行反复迭代和优化。当然，最大的一个障碍在于理论和实验的无缝衔接上。高通量理论计算的结果通常精度有限，可靠性相对较低，而高通量实验也存在实验步骤繁琐，各个实验步骤的衔接较为困难。将两者相结合，所能开展的材料体系目前相对较为局限。

大语言模型的发展对于材料科学来说是一个很好机遇，也是未来 AI4S 的一个很重要的前沿方向。作为材料学科的专业人士，不仅要应用这些大语言模型，还要试着主动去参与这些大语言模型的开发，特别是国产大语言模型的开发。目前来看，国内的大语言模型们虽然功能弱一些，但其发展势头，迭代速度也是不容小视。

厦门大学信息材料与工业智能实验室

洪文晶教授、博士生导师

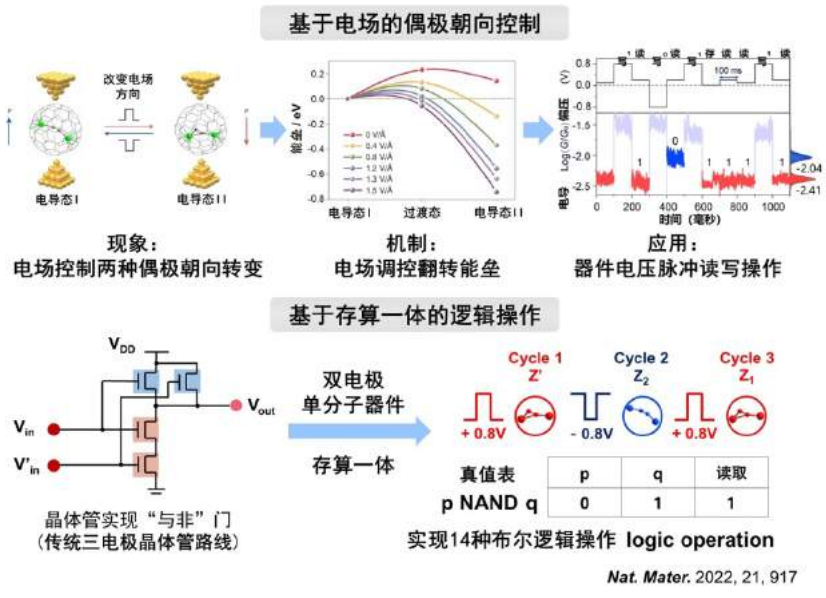
AI4S 在未来将不是一种可能的选择路径，而很大可能将成为科研人员探索未知世界和研发尖端科技的必要工具。在现阶段，我认为对于交叉学科的科研人员而言，AI4S 的关键在于寻找适合这个阶段的 AI4S 能力的重要问题来加以解决，比如面向世界科技前沿的下一代计算芯片和相关电子化学材料研究、比如面向国家需求的新能源相关材料和储能系统研究、比如面向人民生命健康的生物医药研究，还比如面向经济主战场能够显著提升产业竞争力的自主智能仪器和研发系统开发，这些工作都是厦门大学 π-lab 团队正在探索的方向。

在最近的三五年，我坚信 AI4S 将很快突破 demo 应用场景的阶段，而在基础研究和技术创新都将扮演不可或缺的角色，而这需要 AI 理论方法与不同学科领域的研究人员相向而行。

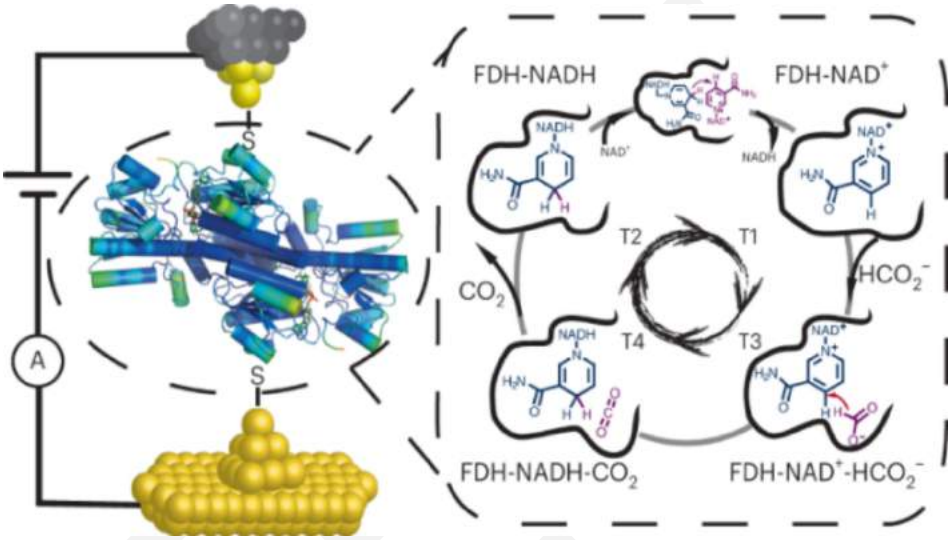
实践 1：Beyond Moore 厦门大学利用 AI4S 方法发展下一代分子集成电路

单分子电子学旨在探索单个分子作为逻辑器件的潜力，探索基于分子的 Beyond Moore 下一代计算芯片解决方案；同时提供了从分子水平获取复杂化学和生物体系组装和反应过程信息的独特研究工具。为了满足高通量的单分子材料研发需要，厦门大学洪文晶教授团队正在致力于研发一套能够完全无人智能化研究的单分子表征仪器和研究系统：该系统能够根据预设研究目标，通过自然语言处理技术精准高效获取文献数据，为体系筛选提供决策支持；基于决策产生的分子结构开展无人智能化的分子合成，并通过智能化的单分子电学表征科学仪器实现目标分子体系的高通量表征；结合开源发布的 XMe Code（github.com/Pilab-XMU）进行智能化的数据分析，基于产生的多模态数据进行数据库 XMe Database 的构建，并结合机器学习算法实现对海量数据的智能分析，真正让 AI 赋能单分子电子学基础研究中合成、表征、分析、决策的全链条环节。

基于单个分子实现逻辑运算长期以来都是单分子电子学这个学科领域的终极目标，但在分子电子学这个设想在 1974 年提出之后的近 50 年间却迟迟未能得到解决，其主要技术挑战在于单分子器件的精准电学表征和原子级制备集成。基于团队开发的智能科学仪器和数据分析算法，团队通过与厦门大学谢素原院士合作，基于内嵌金属富勒烯(Sc2C2@Cs(hept)-C88)提出了单分子尺度的存算一体器件方案，首次在实验上实现了单分子器件 14 种逻



辑运算的原理性验证（上图）。并与英国兰卡斯特大学 Colin Lambert 院士合作开展的密度泛函理论计算表明，非易失性存储行为来自富勒烯笼中[Sc2C2]基团的偶极重定向。该成果揭示了未来单分子电子学器件可能具有显著有别于传统三电极晶体管器件的工作机制和集成架构，为未来高密度集成、低能耗操作的存算一体乃至 AI 计算芯片的发展提供了信息材料和器件基础，(Nature Materials, 2022, 21, 917-923)。美国凝聚态物理学家 Douglas Natelson 教授在《自然-材料》上评价该研究将逻辑和存储器件推进至极致尺度。



同时，单分子电子学的智能科学仪器和数据分析方法还可以用于酶催化等复杂反应过程中信息的提取和分析。团队通过与厦门大学方柏山教授合作，在近期将科学仪器应用于酶催化研究，建立了基于单分子电学表征的酶催化过程研究平台，通过谱聚类的智能算法对大量测试数据进行挖掘分析后，发现甲酸脱氢酶其分别结合还原型或氧化型辅酶 I 后呈现不同的电导值。这种现象与教科书中广泛接受的 Theorell-Chance 机制（1955 年诺贝尔生理学或医学奖）不同，其催化结束时不经历脱辅酶状态。结合多尺度模拟，团队提出反应结束后结合在甲酸脱氢酶中的还原型辅酶通过氢负离子原位转移直接转化为氧化型辅酶，并直接开启新的催化循环（上图），而与王斌举教授合作开展的计算表明这一新机制在能量上是更为有利的。(Nature Catalysis, 2023, 6, 266-275)。面向单分子电子学这一基础前沿交叉研究领域数据获取和分析存在具有挑战的特点，AI4S 的研究范式充分体现其在特征数据提取和大空间搜索领域的独特优势，揭示了 AI4S 在基础研究领域的重要潜力。

参考文献：

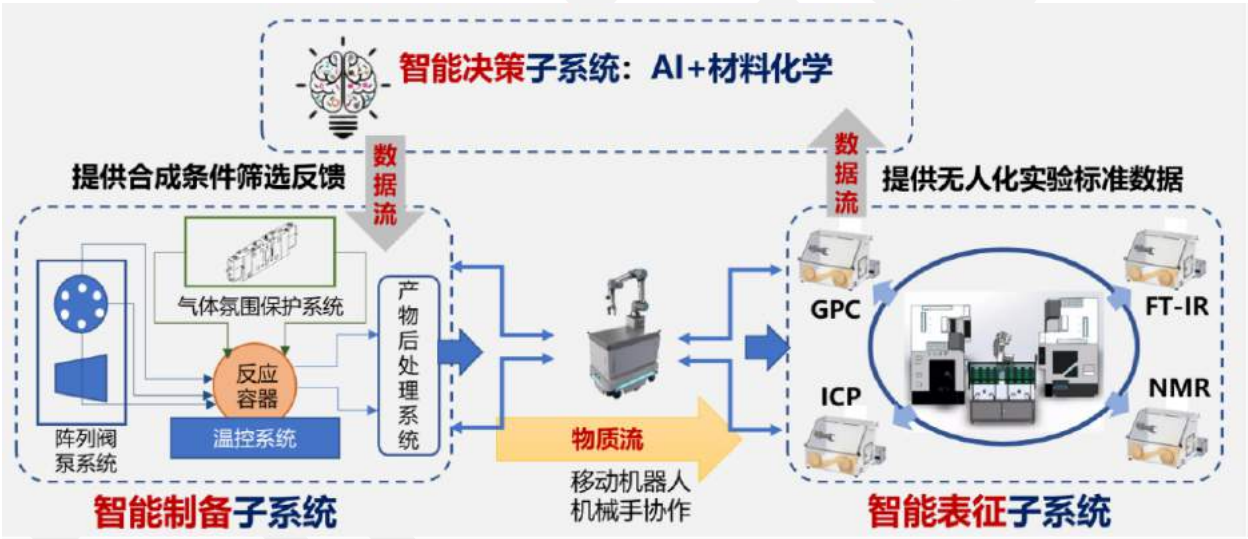
Li, Jing#; Hou, Songjun#; Yao, Yang-Rong#; Zhang, Chengyang#; Wu, Qingqing; Wang, Hai-Chuan; Zhang, Hewei; Liu, Xinyuan; Tang, Chun; Wei, Mengxi; Xu, Wei; Wang, Yaping; Zheng, Jueting; Pan, Zhichao; Kang, Lixing; Liu, Junyang; Shi, Jia; Yang, Yang; Lambert, Colin J.*; Xie, Su-Yuan*; Hong, Wenjing*. Room-temperature logic-in-memory operations in single-metallofullerene devices. *Nature Materials*, 2022, 21, 917-923

Zhang, Aihui#; Zhuang, Xiaoyan#; Liu, Jia#; Huang, Jiacheng#; Lin, Luchun; Tang, Yongxiang; Zhao, Shiqiang; Li, Ruihao; Wang, Binju*; Fang, Baishan*; Hong, Wenjing*. Catalytic cycle of formate dehydrogenase captured by single-molecule conductance. *Nature Catalysis*, 2023, 6, 266-275

实践 2：基于自主智能研发系统的智慧材料研发实验室与 AI4S

厦门大学洪文晶教授团队正在嘉庚创新实验室的支持下，通过与宁德时代等产业龙头企业的密切合作，建立智慧材料研发实验室系统，以此通过建立 AI 驱动的智能材料研发新范式，突破长期以来以人为主的试错研发模式导致的材料研发周期较长且研发成本较高的瓶颈。该系统包括基于智能机器人的精确位移控制系统，自动化高通量的材料器件高效合成系统，对所制备材料和器件进行表征的材料综合表征系统，以及能够对上述系统所获取的数据进行学习并进行反应参数自主优化的智能材料设计决策系统。

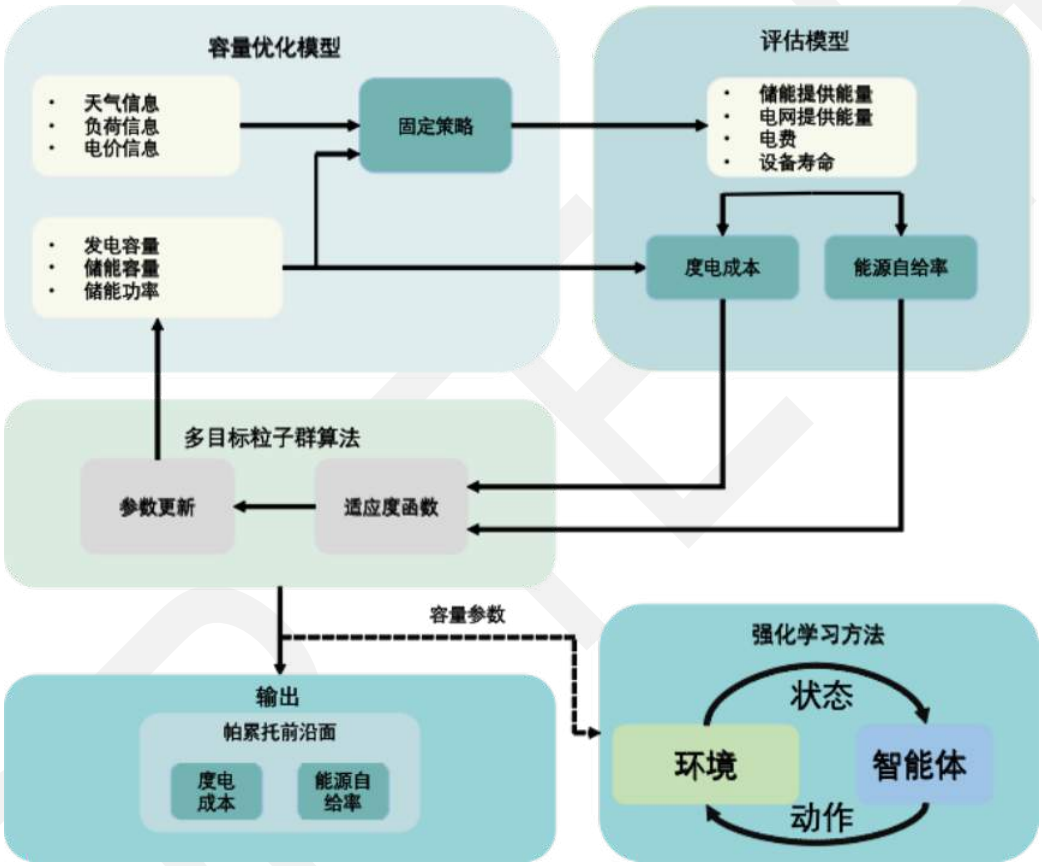
该实验室通过与多个国产科学仪器厂商的密切合作实现了对核磁、拉曼、红外等一系列表征仪器的无人智能操作改造，以此通过基于所开发的智能实验室平台获取实验数据，通过机器学习算法建立可用于材料性能的优化预测模型。针对特定材料性能需求，通过海量实验数据进行自主训练得到针对不同材料的最佳制备条件，并结合智能材料设计系统进行优化迭代，以此实现新材料的设计、合成和性能测试。目前该系统已经在面向锂电池、燃料电池催化剂和复合材料三个方向进行了初步探索，初步实现一周时间的无人连续运转和智能研发，在电池研发领域迭代效率显著优于传统研究范式。通过这一智能化研发平台的建设，我们希望探索基于 AI4S 的材料研发新范式和新系统。



实践 3：基于电氢联储的新型储能策略与 AI4S

储能技术是构筑基于新能源的能源体系中不可或缺的关键环节，是应对可再生能源波动性和随机性的有效解决方案。将电池储能和氢储能相结合，形成电-氢耦合储能系统，可以实现快速响应、高效率和储能容量与储能功率解耦等特点，有望提供长时、高效、大容量、低成本的储能系统解决方案。为了优化利用现有资源并应对潜在挑战，电-氢耦合微电网系统需要智能可靠的规划和调度系统。人工智能技术可提供弹性、高效、可靠和可扩展的解决方案，解决电网耦合微电网规划和调度中的问题。

厦门大学洪文晶教授团队在嘉庚创新实验室的支持下，开发了一种基于多目标粒子群-强化学习算法的氢电联储系统规划和调度框架（下图），该框架使用多目标粒子群算法快速生成经济可行的容量配置方案，并将可行的配置方案作为参数传入到强化学习算法中进行系统运行优化。我们的微电网两阶段规划与运行方法旨在研究在不同的电力市场、负荷水平和设备成本下电-氢耦合微电网系统的技术经济性。该调度算法能够有效应对电价和负荷波动，降低系统建设成本，设计的微电网系统平均电能成本显著低于从电力市场购买电力的价格，能够有效降低 10% 的系统运行成本。我们希望 AI4S 的氢电联储系统研究，能够为未来储能系统的大规模应用提供基于人工智能的解决方案。



西湖大学人工智能与科学仿真发现实验室

吴泰霖 | AI 加速大规模地下流体仿真

西湖大学人工智能与科学仿真发现实验室由吴泰霖博士创建。吴泰霖博士课题组长期开展 AI + Science 学科交叉的核心、普适问题。主要研究方向包括：（1）开发机器学习算法（基于图神经网络、扩散模型）用于大规模、多尺度科学仿真（用于流体、材料、等离子体）和科学设计（蛋白质设计、材料设计、机械设计）；（2）开发机器学习算法发现科学系统普适规律和内部结构（用于生命科学和物理）。吴泰霖博士已在以上 AI 用于科学仿真（流体、等离子体、机械）和科学发现（发现物理方程）领域做出多项基础、开创性工作。本文将简述吴泰霖博士在 AI 加速地下流体仿真方面做出的开创性应用工作。

地下流体仿真是能源和环境领域的核心任务之一。从油田的多相流体模拟到地下水管理，从碳捕获到碳储存，都需要对地下流体（例如水、石油、天然气、二氧化碳等）在岩石中的动力学进行准确而快速的仿真，从而更精确地控制地下流体，提高经济效益，更好地实现碳中和的目标。传统基于第一性原理（偏微分方程）的数值方法（例如有限差分、有限元等）尽管精确，但其核心缺点是仿真速度非常慢，往往需要上千 CPU 的集群通过数小时才能对一个中等大小（约百万网格）的系统进行仿真。近些年也有多个通过数据驱动的代理模型工作加速对系统的仿真，但其模型表达性不足，也最多只应用在了二维的比较简单的系统，离实际的三维、大规模复杂地下流体系统还有很大距离。

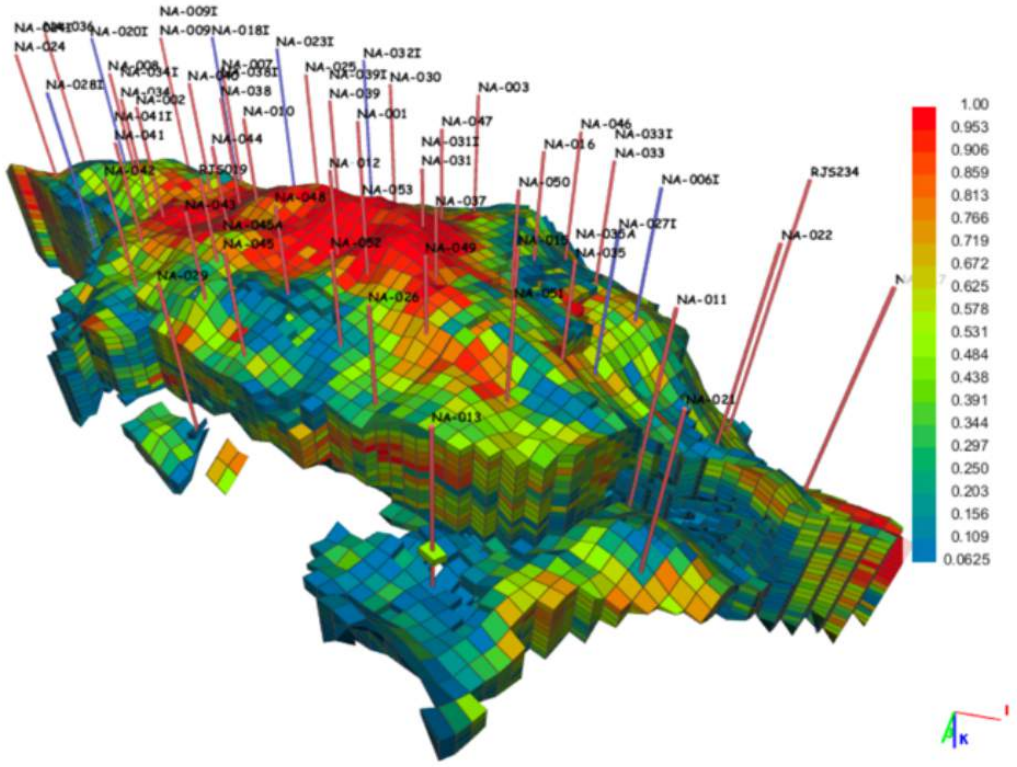


图 | 油田中的地下流体仿真。通过将高压水打入注入井（蓝色垂直线），推动地下流体（石油、天然气等）从生产井（红色垂直线）中流出。为更好设计井的位置，需要对整个地下流体进行准确、快速的仿真

吴泰霖博士与沙特阿美（世界最大石油公司）合作，开发出了首个能够模拟三维、大规模地下流体的图神经网络代理模型：混合图神经网络仿真（HGNS）模型 [1]（见下图），并已在沙特阿美公司生产线的仿真系统部署。HGNS 模型将地下区域划分成一个三维规则网格，网格的每个单元格可以是正常的岩石，或者是生产井、注入井、岩石裂缝等。HGNS 把每个单元格作为图的一个节点，与周围相邻的六个单元格连边，整个网格构成了一个多种节点类型的异构图。每个节点的状态包括单元格内的平均压强和水、石油等的饱和度，它们随时间变化。每个节点还有渗透性、多孔性、空隙量、高度等十多个静态参数（不随时间变化，但随空间变化）。HGNS 模型把以上静态参数和在时间 t 的动态状态作为输入，通过一个图神经网络预测地下流体饱和度在时间 t+1 的状态，通过一个 3D-U-Net 预测压强在时间 t+1 的状态。在预测时，可以将以上过程迭代，从而预测系统长时间（几十年）之后的未来状态。

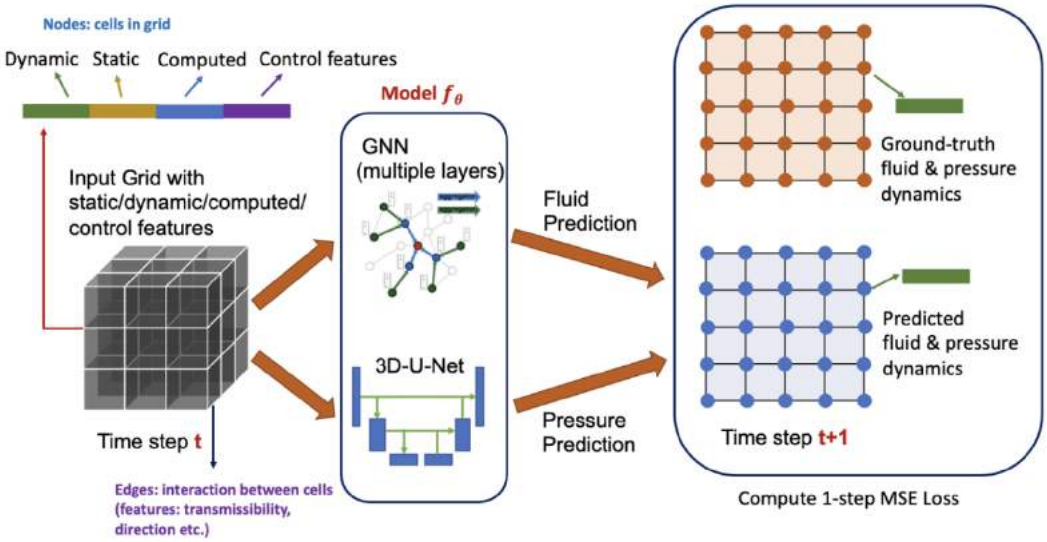
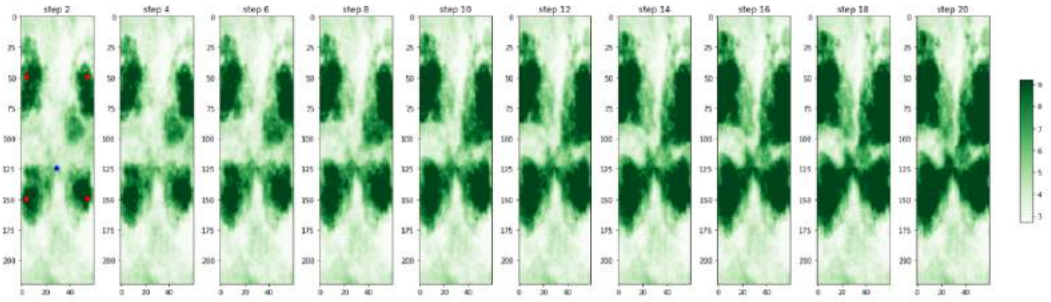


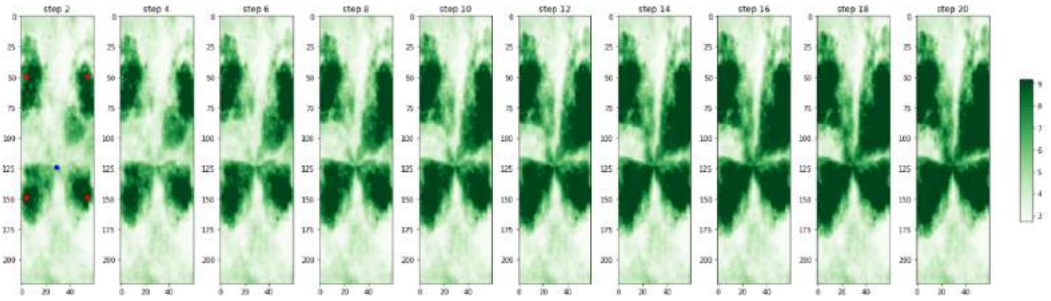
图 | HGNS 模型架构

本工作的核心创新点包括：1）**降低多步误差累积**：针对数据驱动的代理模型长时预测的误差累积问题，提出了训练中降低多步误差的损失函数方法。相比起通常的训练中降低单步误差的损失函数，极大减小了误差累积，使模型能更准确预测长时间后的未来。2）**首次大规模系统仿真**：首次将图神经网络应用在三维、上百万网格的系统仿真中，相比起此前工作，其系统的大小提高了两个数量级。针对这样大规模系统仿真中遇到的无法在单个 GPU 中放下整个图的问题，提出了基于子图的训练和推理方法，使得在训练时能基于 10^4 大小的子图进行训练，而推理时可以应用在 10^6 到 10^7 大小的网格中。

本工作在一个困难的、含一百二十万网格的系统的数据集上对 HGNS 进行了评估，并于标准的深度学习代理模型（CNN、3D-U-Net、基本的图神经网络）进行了比较。评估发现，HGNS 模型能够更准确模拟系统 20 个月的未来，并与求解器的仿真结果有很高的相符度（下图），其误差远小于其他基线模型。与标准的求解器相比，其仿真速度快了 18 倍。本模型已在沙特阿美公司生产线的仿真系统部署，其在单个 GPU 上对千万网格系统的仿真速度快于其求解器在上千个 CPU 上的仿真速度。这是图神经网络首次在真实的应用场景进行的大规模部署。



(a) HGNS rollout of water volume (barrel) for 20 steps (months)



(b) Ground-truth of water volume (barrel) for 20 steps

此工作只是 AI4S 在地下流体的初步应用。展望未来，AI4S 将在地下流体的仿真中扮演越来越重要、甚至是核心的角色。尤其是，对于复杂的地质条件的反向估计、多尺度多分辨率的仿真、更加精确的设计和控制方面，传统方法面临巨大挑战，AI 方法对于解决以上问题有很大的发展空间和应用前景，能够极大提升仿真的速度和准确性。展望未来，由于能源和环境领域有万亿计的产值，碳中和将成为未来的重要趋势，更加准确和高效 AI 地下流体仿真方法将能帮助决策者和生产者更好地规划未来，产生巨大的经济效益和社会价值。

参考文献：

Tailin Wu, Qinchen Wang, Yinan Zhang, Rex Ying, Kaidi Cao, Rok Sosic, Ridwan Jalali, Hassan Hamam, Marko Maucec, Jure Leskovec. "Learning Large-scale Subsurface Simulations with a Hybrid Graph Network Simulator." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD), 2022.

清流资本：投资像 AI4S 这样的前沿科学领域是一种"双赢"策略

作为曾深耕互联网产业十数年的投资机构，清流资本的数智化投资板块一直坚持着两条主线：以云和 AI 为代表的互联网技术向下渗透带来的平台级新基础设施机遇，以及各个垂直行业的数字化、智能化带来的效率升级机遇。其中，AI 技术本身的发展和 AI 在各垂直应用场景的解决方案，一直是清流资本密切关注的方向。

现代人类文明最重要的几大行业：生命健康、新能源、新材料、新技术，过去面临的主要瓶颈都在微观层面的工业设计和计算上，由于微观粒子之间多体作用的数学问题太复杂，一直没有好的求解方法，"维数灾难"问题严重制约了人类科学的进步与发展。

站在这个历史拐点上，AI for Science 出现并产生了突破，AI 成为了一个对高维函数极为有效的拟合和求解工具，能系统化解决科学研究中的高维问题。深势科技作为该领域的领导者，提出并引领了 AI for Science（AI4S）理念，这种理念能将精确研发推向新的层次，实现微观世界的分子模拟，为科学研究带来了跨时代的飞跃。深势科技的计算平台已经应用于药物和材料研发领域，极大地提高了新材料的发现效率，推动了相关产业的快速发展。AI4S 不仅加速了人类科学研究的进展，为人类探索世界的新边界提供了有效的工具，而且为生命健康、新能源、新材料等领域提供了高效的实用工具，搭建起了科学和产业发展之间的桥梁。

AI+分子模拟被誉为人工智能技术继计算机视觉、语音和自然语言处理等赛道后的下一个颠覆性应用场景。其所带来的更深层次的、有系统方法支持的框架成为了研究微观世界的新型基础设施。这种体系将涵盖不同尺度的扩展空间，有可能催生跨尺度的研究协同机制，带来全新的研究方法的机会。

从清流资本的角度来看，我们坚信 AI4S 将对未来的基础科学，甚至下游的科技产品和智能制造产生深远影响。这种影响可能是间接的，但当技术原理发生变化时，下游产品的形态和发展路径必然会受到深远影响。这是因为，科学的进步和技术的创新直接塑造了产品的特性和功能，进而改变了其在市场中的竞争地位。在长期的视角下，技术优势往往会转化为市场优势，为投资者带来稳定且可观的回报。

投资像 AI4S 这样的前沿科学领域，可以视为一种"双赢"策略。从社会角度来看，这些投资为人类解决重大科学问题、推动科技进步，提供了强大的动力。从经济角度来看，新技术和新科学常常带来范式的转变，创新了产品和服务，颠覆了既有的市场，为投资者带来了巨大的商业价值和金融回报。以 AI4S 为例，其将有可能催生出全新的药物研发方式，新材料的发现与应用，以及其他未知但前景广阔的应用领域，为社会带来福利的同时，也将产生巨大的经济价值。

清流资本始终坚持与时俱进，以开放的视野和前瞻的视角，寻找并投资这些能够引领未来，具有创新力和潜力的深度科技项目。这既是我们为社会做出的贡献，也是我们追求长期稳健回报的战略选择。立足于高远的愿景，投资机构的目标不仅是商业上的盈利和成功，更重要的是发挥资本对先进科技和实体经济的支持作用。清流资本始终坚持"取之于民，用之于民"的理念，我们希望通过支持像深势科技这样在探索、技术、创新和贡献上具有突破性的企业，为中国乃至世界的科学技术发展贡献力量。这也符合我们对 AI for Science 的深度理解和愿景，我们期待它能为更多行业和领域带来创新，让世界因为科技而变得更美好。

元璟资本：AI 能更大幅度的推动人类社会的发展

在基础科学领域分为两种流派，第一性原理驱动和数据驱动：第一性原理驱动更加本源导向，对问题的理解更深刻，解决更优雅，但对复杂问题的效率比较低；数据驱动效率更高，结果导向，但难以理解问题的本质

今天的机器学习是在用大规模的算力去解决大规模的数据问题，本质是一种数据驱动的解决方法。而纯粹基于数据的机器学习在 science 领域最大的问题是，无法做到可解释和精准，虽然可以做到在一些领域解决问题，但要推动基础学科和科研的进步，还是需要回归到第一性原理，去理解问题的原理。

因此，只有让机器学习能够理解基础学科的运行原理，才能这种做到让机器学习做到跨领域和多尺度的去解决复杂科学问题（拥有在 science 领域内的泛化能力）

如何将科学的第一性原理和机器学习结合是 AI for Science 最重要的问题：让机器学习的大规模计算能力，结合上第一性原理，去解决复杂和大规模的问题，帮助科学研究扩展到更高维的空间，理解人类依靠人脑很难去理解和抽象的高维问题，实现科研的突破。今天人类在真实世界遇到的大部分问题都是复杂的系统化问题，只有让机器学习能够理解真实世界的运行原理，才能更好的解决真实世界遇到的问题

AI for Science 可以将原来依赖人的创意和灵感来推动的科学研发变成一个系统化和体系化的新范式，这是一个类似农耕时代迈入工业革命的突破，而且发生在人类智慧的顶尖—科研领域，无论是对产业还是对人类未来大发展都有巨大的意义。

原来科研和产业界总是有着明确的距离，一个重要的原因是科研本身是一个小规模的突破，而产业界需要的是大规模的效率，而 AI for Science 通过将科研的过程和目标都变成大规模，能解决更复杂的实际产业问题，直接拉进了科研和产业的距离，不管在药物研发还是在新材料领域，这样的促进都已经出现。

今天，算力，机器学习，人工智能领域的快速发展给了 AI for Science 最好的基础，科学计算，基础学科今天可以和这些新兴领域彻底结合，这是一个时代的机遇。在这样跨学科结合的新领域里，有大量没有解决的问题，同时也就有更多的机遇；而在这个新的领域，全世界都在同一个起跑线，无论是 Google 微软还是国内的学术界和产业界，都在类似的时间开始往这个方向探索，中国的团队有机会再这样的新范式的变革中，做到全球领先。

如果我们认为未来 AI 能更大幅度的推动人类社会的发展，AI for Science 就是最重要的方向，能够真正的帮主人类去理解我们身处的世界，小到原子，大到宇宙。

九合创投：AI for Science 有望推动更多技术平台的诞生

九合创投自 2011 年成立以来，始终聚焦技术变革带来的结构性投资机会，管理着 5 支人民币基金和 1 支美元基金，被投企业共 200 余家，是青云科技（688316.SH）、鹰瞳科技（02251.HK）、探迹、众合云科、Momenta、态创生物等上市公司和独角兽企业的早期投资人。

如何理解 AI4S？AI for Science 就是让 AI 帮助科学家从海量复杂的数据中，快速提炼特征和规律，然后科学家通过建立模型，筛选出真正有意义的规律，解决实际的科研问题。AI 能够辅助科学家在不同的假设条件下，进行大量的验证和试错，显著提升科研效率，改变了科学研究的范式。

AlphaFold2 是 AI for Science 的成功案例。蛋白质结构预测是个非常适合通过 AI 解决的科研问题，biotech 领域积累了大规模的蛋白质结构数据，AlphaFold2 的模型架构则充分利用了数据驱动的深度学习模型，具备模型上的优越性，为科学家提供了高效的计算工具。

AIGC 浪潮来临后，AI for Science 出现更多的可能性。不同背景的科学家们打破学科界限，共同研究计算、生物、化学等交叉领域的新课题，与此同时也给产业应用带来新一波的机会。

我们投资的域新说是一家专注于全新蛋白设计的公司，利用 AIGC 算法设计全新的生物元件，将生物元件应用于基因电路构建，使得菌株及催化酶可以进行自然进化，最终生成全新功能性蛋白。蛋白质生成和设计对生物技术、医学等具有重要意义，能广泛应用在酶的改造与设计、新药研发、基因治疗与基因编辑等领域。

九合投资的青云瑞晶、态创生物等也是 AI for Science 的典型代表。青云瑞晶是基于 MicroED 微晶电子衍射的小分子药物发现平台，是"AI+化学"及"AI+生物"的探索，能结合人工智能算法，缩小需搜索的化学空间，降低所需计算的变量，从而解决计算精确度与计算量的矛盾，克服目前难以成药靶点，研发出有效的药物，并极大地缩减研发周期和成本。

在新材料研发领域，AI for science 也是重要的趋势，但在 AI 应用场景和后续科研成果的商业化方面有待突破。目前其落地商业化路径不太像 AI 在医药中那般成闭环，场景上还需要尝试突破，新材料大部分研发还是属于实验科学，AI 的加速还是需要建立在理论基础之上，包括材料基因组大数据等基础设施建设完善之后，可能在此基础之上才能更好的促进 AI 在材料研发中的应用。

本质上来讲，AI for Science 是辅助科研创新的工具，发挥作用基于数据和计算两大要素，带来了模型与数据驱动的科研新范式。如果学科有丰富的数据，并且科研问题和大规模计算相关，需要强大的计算能力，AI for Science 将更有可能取得突破。

未来，AI for Science 有望推动更多技术平台的诞生，解决生物技术、材料、能源等领域的产业问题。

创世伙伴：加速跨学科"合作共赢"的规模化成果

创世伙伴 CCV 合伙人梁宇：这一波的人工智能新浪潮，点燃了炙热的面向公众的科幻造梦之火。可见的表象领域，文字、图像、视频、音频和 3D 如火如荼的进入 AIGC 建设的大工地之喧嚣随处可见。但在 AI4s, 远离探照灯的幕布背后，对 AI 新范式在科研领域的破坏性创新才真正开始萌芽。AI 对社会分工的破坏和重建已经开始数年。这背后的真正意义，是生产力的巨大解放和随之人类福祉的极大提升。这是来自 AI 的馈赠：

加速预测蛋白质结构，生命科学将迎来前所未见的加速发展和创新。在可见的将来癌症的被攻克，衰老机制的揭示，病毒的解构和应对跨物种的研究，将有可能重构我们对著名哲学问题的回答："我是谁，我来自哪里"。而这一切，将由 AI4S 的子领域贡献者，添砖加瓦。

韦伯接过哈勃的班，源源不断传回的天量数据让我们有机会看到 135 亿年前的深空信号。新 AI 基础设施的建设，有机会很快向我们揭示新知。也许有机会，我们将彻底打破 100 年来基础物理界的沉闷，为"大爆炸理论"的新辩论提供崭新的正反面证据，并目睹新理论的诞生。AI4S 在这个领域的渗透，将非常有力地重构我们的宇宙观。兴许新 "哥白尼时刻"正在急速奔向我们而来。

人类的巨大进步和进化力量来自协作。新 AI 将有机会通过新算力，新数据，新模型把各自然学科已经碎片化的各远端的独立分支重新收拢在一起，并系统化地提供崭新的视角和维度。跨学科的研究，由于知识结构积累所需的时间和训练已经变得原来越困难。自然规律，在人们具有巨大创造力的黄金年龄 30 岁往下。教育，知识积累和融会贯通占据了大量的时间。肉眼可见，战后大量科研领域的创新往往来自跨学科。AI4S 的介入，可以系统化地将跨学科组织方法论变为非常接近的现实，一扫过去孤岛化，随机化的结果产出。把各巨量数据间关联关系的发掘和推演交给 AI, 将判断和修正交给科学家：这样的组合将爆发巨大的"研究生产力"，日新月异地为科研阵地提供源源不断的炮弹。

现在我们谈及的教育，是 200 年工业文明的产物。AI 新范式带来的冲击，将一定程度影响教育体系，教育是否能如期待般源源不断的向业界输送工程，研究，产业人才值得期待。知识的积累和记忆能力，AI 已经显著超越碳基生物，我们的教育是否应该主动走向科研的产业链条上游：判断，决策，和创造?

这一切，都有赖于 AI 新范式下的基础设施建设和投入产出比的不断优化，即科研成本的断崖式降低。可见的通过自然语言将代码自生成和可视化，算力的云化，和天量数据的存储快速检索的工程技术普及，将有机会将科学研究 "原子化"和 "部落化"的加速趋势延缓，将科研要素重新组织，重新定义"分工合作"，加速跨学科的"合作共赢"的规模化成果浪潮的到来。

新范式下，新工具集，新模型，新计算平台，新组织和沟通方式将由各位来协同建设。这里孕育的巨大商业机会,潜在的伟大公司和伟大科研成果的诞生，有赖于行业中具备洞察力和真正好奇心驱动，兴趣驱动的个人和组织，在 AI 范式新灯塔的指引下，砥砺前行并享受其中。

结语：理性之光再次照亮科学大地

从生命的基本组成（蛋白质）到世界工业的基本要素（材料）到各个科学技术领域，AI4S 不只是解决具体问题的有力工具，更是重新定义科学问题的系统性思路。

前计算机时代，我们定义问题的方式是"此问题是否能观测并总结规律"；计算机时代，我们定义问题的方式是"此问题是否能被多项式有效近似并通过计算机模拟"；而 AI4S，让我们可以重新提问"此问题是否有原理和数据？"哪怕只满足一条，我们也有机会将问题推进到前人无法触及的程度。

当我们用 AI4S 的视角重新审视以前被学界定义为"无法攻克"的难题时，我们发现了全新的可能性，并在不少场景中取得了史无前例的突破。如果我们能更广泛的应用这个思路，系统性的重新审视过去的科学问题和科学成果，或可将理性的光照入科学未解之领域。

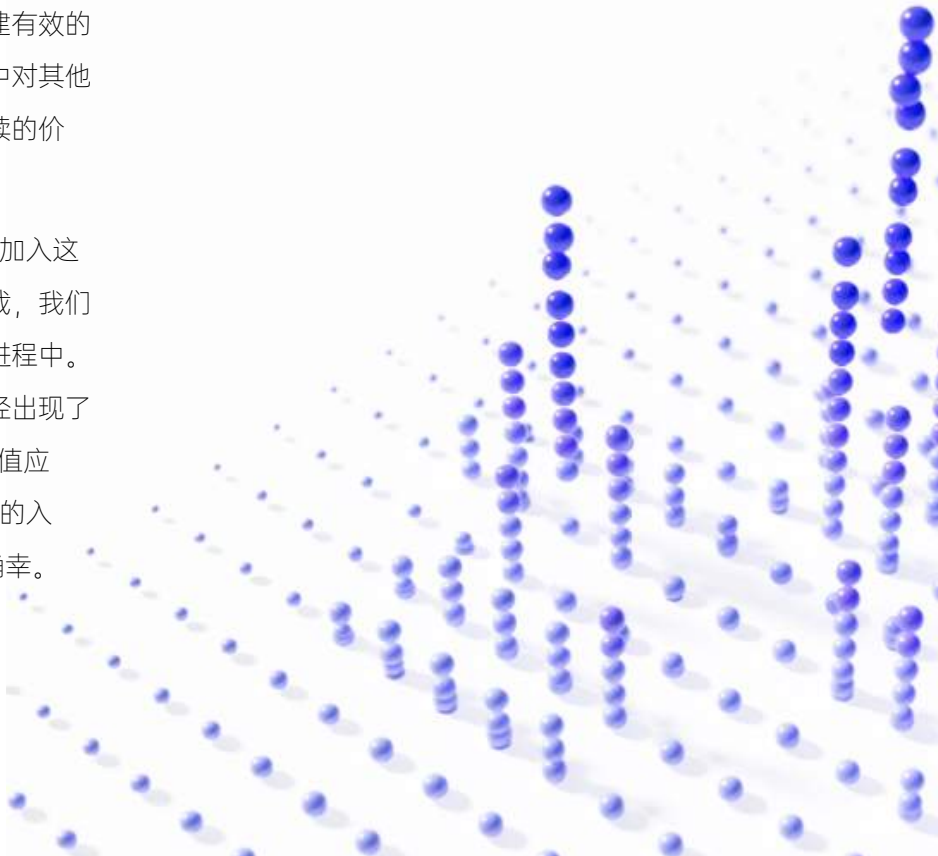
不仅如此，AI4S 还能长期反哺科学：AI4S 在不断的实践中积累高质量数据，在小范围内构建有效的原理，然后和数据反复迭代，在整个过程中对其他科技路线的发展（比如量子计算）提供持续的价值。

面对 AI4S 的巨大机会，我们很兴奋有机会加入这样一项事业中来，就如一步伟大作品的连载，我们有机会见证甚至亲自参与到这样一段历史进程中。然而，我们也必须时刻提醒自己，虽然已经出现了 AlphaFold2、DeePMD、PINN 这样的高价值应用，但我们仍旧只是停留在 AI4S 这片富矿的入口，万万不可沉迷于摘得"低垂之果"的小确幸。

同时，在兴奋之余仍需保持冷静和理性。任何一项新兴技术的发展都有周期波动，我们也都经历过许多技术领域过热后跌落神坛。泡沫并非全然有害，也会吸引更多优秀人才和资源注入我们的领域。古今中外伟大的机构和组织也无一不是穿越了周期，在泡沫破碎后成长起来的。身为行业的从业者和开拓者，如果致力于推动技术边界的不断扩张，我们就需要保持定力：聚焦真需求而非概念的炒作；关注长期价值而非短期利益；团结一切可以团结的力量而不是领域不大却过早陷入内卷。

我们期待以这部《展望》，抛砖引玉，让更多人了解、熟悉 AI for Science，甚至加入到这样的事业中来。

受限于团队的视野和时代的局限，本文所述的很可能只是 AI4S 的冰山一角。科学的大地浩瀚无垠，愿"理性之光"再次将其照亮！



附录 1：学术及产业各界声音*

US Department of Energy	"From chemistry to materials sciences to biology, the use of ML and deep learning (DL) techniques opens the potential to move beyond today's heuristics-based experimental design and discovery to AI-enhanced strategies of the future." [1]
Nobel Laureate in Chemistry 81', **Roald Hoffmann**	The wave of AI represented by machine learning and artificial neural network techniques has broken over us. Let's stop fighting, and start swimming. . .We will see in detail that most everything we do anyway comes from an intertwining of the computational, several kinds of simulation, and the building of theories [2]
Roberto Car, National Academy of Sciences, USA	By learning the potential energy surface from quantum mechanics, Machine Learning methods make possible simulations of AIMD quality at FF cost [3]
Demis Hassabis, co-founder and CEO of DeepMind	"I believe that AI will become a kind of meta-solution for scientists to deploy, enhancing our daily lives and allowing us all to work more quickly and effectively. If we can deploy these tools broadly and fairly, fostering an environment in which everyone can participate in and benefit from them, we have the opportunity to enrich and advance humanity as a whole."[4]
Yann LeCun, Chief AI Scientist of Facebook	"I believe we need to find new concepts that would allow machines to: - learn how the world works by observing like babies. - learn to predict how one can influence the world through taking actions" [5]
Jensen Huang, CEO of Nvidia	"The super computer (AI) of the future will enable a complete reinvention of physics, chemistry, fluid dynamics, and every aspect that is today modeled based on first order principles. This will enable multi-physics approaches that will help us understand the world around us at a far deeper level." [6]
Aart de Geus, Chairman and co-CEO of Synopsys	"The ability to use AI to design AI chips, that's the ultimate cool" [7]

Kevin Scott, EVP and CTO, Microsoft	AI4Science is an effort deeply rooted in Microsoft's mission, applying the full breadth of our AI capabilities to develop new tools for scientific discovery so that we and others in the scientific community can confront some of humanity's most important challenges [8]
Iya Khalil, Global Head of the AI Innovation Lab at Novartis	Not only can AI learn from our past experiments, but, with each new iteration of designing and testing in the lab, the machine learning algorithms can identify new patterns and help guide the early drug discovery and development process. Hopefully in doing this we can augment our human scientists' expertise so they can design better molecules faster. [8]
Cathie Wood, CEO of Ark Invest	"The ground is shifting underneath us … with five major innovation platforms evolving at the pace of early 1900s. They are DNA sequencing, Robotics, Energy storage, Artificial Intelligence, Blockchain … and they are converging with one another in massively profound ways. [9]
Joris Poort, CEO of Rescale	"Simulations can not only predict how a single human-designed product might perform, but it can also predict the performance of a full range of potential designs … applying Artificial Intelligence techniques to biology, and accelerating new drug discoveries by analyzing cells 20 times faster using machine learning on supercomputers." [10]

中国科学院院士、复旦大学光电研究院院长褚君浩	“我们基于物质科学新发现和微纳米器件制造新技术，基于脑科学和认知科学发展促进了高度智能化，信息技术在不断地发展，数字、物理和认知网络将实现高度融合，人类将处于无所不在的网络之中。” [11]
图灵奖获得者、中国科学院院士姚期智	“AI 研究的特色，就在于它能够巧妙结合若干元素，通过学科交叉产生拥有‘大能量’的新核心技术。AI 在投入具体场景应用的具体过程中，会生成各种新的问题。如何用各种学科工具去解决实际的问题，对研究者而言，重点是“要推进 AI 的基础研究，以激发原始性的创新突破。” [12]
中国科学院院士、清华大学人工智能研究院名誉院长张钹	“发展第三代人工智能，必须把第一代人工智能知识驱动和第二代人工智能数据驱动结合起来，必须充分利用知识、数据、算法、算力四个要素来解决人工智能现在存在的不安全性问题。” [13]
施一公，中科院院士，西湖大学校长	“AI 一定会影响我们的研究，不管你是否愿意被它影响。AI 的到来，让我们更清楚的知道，科学应该是一种探索，而不是谋生。” [14]
高文院士，中国工程院院士，北京大学教授	“现在计算机往前走，肯定走不了太远，会面临各种各样的限制，要想变得更强大，只有两条出路：一条是采用类脑模型做的加速器，来处理人工智能当前无法解决的问题；另一条是结合量子学，是采用全新的计算方式——算力上去了，一些大的、难的问题就能解决。而这两条出路，都需要人工智能与其他学科的交叉研究。” [15]
鄂维南，中科院院士，AISI 院长	从数学的意义上来说，深度学习提供了一个逼近高维函数的工具。由此产生的影响是巨大的，因为我们在很多场景下都会遇到高维函数……传统科学领域是 AI 更大的发展空间所在……AI for Science 不光有助于大量科学问题的解决，也是推动制造业转型和实体经济发展的重要一环。[16]
陈十一，中科院院士，南方科技大学校长（原）	而真正对智慧城市,数字孪生有用的,是将数字经济与工业软件的可预测性结合起来,这就需要 AI+CAE。能用 AI 的就用 AI,我们最近发现,一些简单的流体力学问题,利用 AI 解决可以节省 10 的 3 次方级的计算空间。所以我觉得下一个版本的数字经济应该是 AI 和 CAE 的结合。比方说,福岛海啸来临时,面对核辐射威胁的前提下,整个城市应该怎么样规划疏散方案?这时,我们就可以通过深度学习和

	CAE 的仿真马上得出疏散方案。为了解决像这样的紧急公共安全事件,数字孪生城市需要有预测性,也就需要 AI+CAE [17]
复旦大学教授，计算物质科学教育部重点实验室主任，龚新高	随着研究体系越来越复杂，研究精度要求越来越高，第一性原理方法在材料发现和物性研究方面的代价也越来越昂贵，无论在空间尺度还是在时间尺度上已经遇到了难以克服的瓶颈。如何解决这些瓶颈问题，是目前计算物质科学面临的最严重挑战。为克服这些挑战，我们将人工智能算法与传统计算物质科学研究方法相结合，探索计算物质科学研究新方法 [18]
中科院院士，北京大学副校长，张平文	目前，（人工智能）这方面中国仅次于美国，世界排名第二。 我国优势主要体现在两点，一是数据，AI 非常依赖数据，我国人口多、数据多，占有极大优势；二是应用场景丰富，国家重视并在政策上给予了最大力度支持。 [19]
清华大学教授、智能产业研究院（AIR）院长张亚勤	从 AlphaGO、AlphaZero 在围棋领域战胜人类，到 AlphaFold2 在医疗领域高精度预测蛋白质结构等，深度学习算法和人工智能技术正在逐步改变我们的物理世界和数字世界。未来五至十年，深度学习还会是人工智能最重要的算法，但未来无疑我们需要知识+数据驱动的融合算法，需要结合符号逻辑、知识型推理和第一性原理的新范式 [20]
中国工程院院士，中国科学院计算技术研究所学术所长，中国科学院大学计算机科学与技术学院院长孙凝晖	人工智能=A+B+C+D+E …… 第一层，C 是指计算，D 是指数据，我们的芯片、高性能计算机、大数据处理技术都是 AI 的基础设施。第二层，AI 的算法层，就是 A 和 B，A 是算法，B 是知识，我们的智能信息处理方向、泛在计算、生物信息处理、知识库，都在这一层。最后一层，E 是生态与应用 [21]

* 注：引用顺序不分先后

Source:

[1] Stevens, Rick, Taylor, Valerie, Nichols, Jeff, Maccabe, Arthur Barney, Yelick, Katherine, and Brown, David. AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science. United States: N. p., 2020. Web. doi:10.2172/1604756.

[2] R. Hoffmann, J.-P. Malrieu, Angew. Chem. Int. Ed. 2020, 59, 12590.

[3] Roberto Car, Presentation at Supercomputing Frontiers Europe 2021, retrieved at https://www.youtube.com/watch?v=LZAdD-kv6aY

[4] Demis Hassabis, The mind in the machine: Demis Hassabis on artificial intelligence, Financial Times, retrieved on 7.28.22 https://www.ft.com/content/048f418c-2487-11e7-a34a-538b4cb30025

[5] Yann LeCun, LinkedIn, retrieved on 7.28.22, https://www.linkedin.com/posts/yann-lecun_ai-activity-6932436820454502400-hV39?utm_source=linkedin_share&utm_medium=member_desktop_web

[6] Pearls Of Wisdom From Jensen Huang, Forbes, retrieved on 7.28.22 https://www.forbes.com/sites/karlfreund/2022/03/24/pearls-of-wisdom-from-jensen-huang/?sh=4a904edb4bab

[7] AI Design Innovation Is Transforming the World Around Us, Synopsys, retrieved on 7.28.22 https://www.synopsys.com/implementation-and-signoff/ml-ai-design.html

[8] Christopher Bishop, AI4Science to empower the fifth paradigm of scientific discovery, retrieved on 7.28.22 https://www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/

[9] Cathie Wood, BIS2020, retrieved at https://www.youtube.com/watch?v=cKeZBWYVcDk

[10] Joris Poort, How Cloud-Based Supercomputing Is Changing R&D, retrieved on 7.28.22 https://hbr.org/2021/11/how-cloud-based-supercomputing-is-changing-rd

[11] 中科院院士褚君浩：第四次工业革命以智能化为特征 低碳技术与人工智能是关键，新浪财经（每日经济新闻），检索于 https://finance.sina.cn/china/gncj/2022-03-01/detail-imcwipih5943701.d.html?vt=4&cid=76729&node_id=76729

[12] AI 离不开数学，但 AI 面临的问题不全是数学问题，文汇报，检索于 http://dzb.whb.cn/2020-08-12/7/detail-695226.html

[13] 张钹院士：加快发展第三代人工智能，推动应用更加安全和公平，战略前沿技术，检索于 https://mp.weixin.qq.com/s/ND_RBMe0q2vg0jF6QBHHDw

[14] 施一公与南开师生交流科研心得，中国科学报，检索于 https://news.sciencenet.cn/htmlnews/2022/4/477823.shtm

[15] 2019 年末，10 位院士对 AI 的深度把脉（上），AI 科技评论，检索于 https://mp.weixin.qq.com/s/mLYBrlP7ilpo8z5j1nxWBQ

[16] 专访 | 鄂维南院士提出应用数学"新的曙光"：AI for Science 将人工智能与基础科学深度融合，DeepTech 深科技，检索于 https://mp.weixin.qq.com/s/glB3tBGxsZmnQCSGmySmBg

[17] 自主工业软件的创新与发展，China Daily，检索于 https://tech.chinadaily.com.cn/a/202207/04/WS62c283a4a3101c3ee7addadc.html

[18] 龚新高，基于人工智能的计算物质科学, retrieved at https://www.sqz.ac.cn/physical-device-15

[19] 张平文，北京智源大会，retrieved at https://www.math.pku.edu.cn/pzhang/zh/post/20200625_rgzn/

[20] 张亚勤，智能科学：无尽的前沿，retrieved at https://air.tsinghua.edu.cn/info/1007/1385.htm

[21] 计算所技术发展处整理，孙凝晖所长在计算所 2017 年度技术创新工作会议上的讲话, retrieved at http://www.ict.ac.cn/zjgd/202007/t20200723_5644801.html

附录 2：AI4S 相关论文索引*

Journal	Title	Author	Year	DOI
Science	Reactive uptake of N2O5 by atmospheric aerosol is dominated by interfacial processes	Mirza Galib, David T. Limmer	2021	10.1126/science.abd7716
Science	Geometric deep learning of RNA structure.	Townshend R J L, Eismann S, Watkins A M, et al.	2021	DOI:10.1126/science.abe5650
Nature	Highly accurate protein structure prediction with AlphaFold	Jumper, J., Evans, R., Pritzel, A. et al.	2021	doi.org/10.1038/s41586-021-03819-2
Nature	A graph placement methodology for fast chip design	Mirhoseini, A., Goldie, A., Yazgan, M. et al.	2021	doi.org/10.1038/s41586-021-03544-w
Nature	Skilful precipitation nowcasting using deep generative models of radar	Ravuri, S., Lenc, K., Willson, M. et al.	2021	doi.org/10.1038/s41586-021-03854-z
Nature	Magnetic control of tokamak plasmas through deep reinforcement learning	Degrave, J., Felici, F., Buchli, J. et al.	2022	doi.org/10.1038/s41586-021-04301-9
Nature	Highly accurate protein structure prediction for the human proteome.	Tunyasuvunakool, K., Adler, J., Wu, Z. et al.	2021	doi.org/10.1038/s41586-021-03819-2
Nature methods	Effective gene expression prediction from sequence by integrating long-range interactions.	Avsec, Ž., Agarwal, V., Visentin, D., et al.	2021	doi.org/10.1038/s41592-021-01252-x
Nature methods	Predicting 3D genome folding from DNA sequence with Akita.	Fudenberg, G., Kelley, D. R., & Pollard, K. S.	2020	doi.org/10.1038/s41592-020-0958-x
Nature biotechnology	Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.	Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J.	2015	doi.org/10.1038/nbt.3300
Proceedings of the National Academy of Sciences	Signatures of a liquid–liquid transition in an ab initio deep neural network model for water	Gartner, Thomas E. Zhang, Linfeng et al.	2020	doi/abs/10.1073/pnas.2015440117
Energy & Environmental Science	Fast Na diffusion and anharmonic phonon dynamics in superionic Na3PS4	Gupta M K, Ding J, Osti N C, et al.	2021	doi.org/10.1039/D1EE01509E

Nature Communication	Ab initio phase diagram and nucleation of gallium	Niu, H., Bonati, L., Piaggi, P.M. et al	2020	doi.org/10.1038/s41467-020-16372-9
Nature Communication	Transforming solid-state precipitates via excess vacancies	Bourgeois, L., Zhang, Y., Zhang, Z. et al	2020	https://doi.org/10.1038/s41467-020-15087-1
Nature Communication	Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation	Zeng J, Cao L, Xu M, et al.	2020	doi.org/10.1038/s41467-020-19497-z
Nature Communications	RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning.	J. Singh, J. Hanson, K. Paliwal, and Y. Zhou,	2019	doi.org/10.1038/s41467-019-13395-9
Nature Communications	Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement.	Xiong P, Wu R, Zhan J, et al.	2021	doi.org/10.1038/s41467-021-23100-4
Chemical Science	The Importance of the Compact Disordered State in the Fuzzy Interactions between Intrinsically Disordered Proteins	Wang D, Wu S, Wang D, et al.	2022	doi.org/10.1039/D1SC06825C
Chemical Science	A machine learning protocol for revealing iontransport mechanisms from dynamic NMR shifts inparamagnetic battery materials	Lin M, Xiong J, Su M, et al.	2022	10.1039/D2SC01306A
ACS Nano	2D Heterostructure of Amorphous CoFeB Coating Black Phosphorus Nanosheets with Optimal Oxygen Intermediate Absorption for Improved Electrocatalytic Water Oxidation	Huayu Chen, Junxiang Chen, et al	2021	Doi.org/10.1021/acsnano.1c04715
Physical Review Letters	Phase Diagram of a Deep Potential Water Model	Zhang L , Wang H , Car R	2021	doi.org/10.1103/PhysRevLett.126.236001
Association for Computing Machinery	Extending the limit of molecular dynamics with ab initio accuracy to 10 billion atoms	Zhuoqiang Guo, and Weile Jia et al.	2022	doi.org/10.1145/3503221.3508425
Carbon	A deep learning interatomic potential developed for atomistic simulation of carbon materials	Jinjin Wang, Songyou Wang et al.	2022	doi.org/10.1016/j.carbon.2021.09.062
Cement and Concrete Research	A deep learning potential applied in tobermorite phases and extended to calcium silicate hydrates	Yang Zhou, Haojie Zheng, Weihuan Li, Tao Mac, Changwen Miao	2022	https://doi.org/10.1016/j.cemconres.2021.106685

Science Bulletin	Neural network representation of electronic structure from ab initio molecular dynamics	Qiangqiang Gu, Linfeng Zhang, Ji Feng	2022	https://doi.org/10.1016/j.scib.2021.09.010.
NPJ Computational Materials	Accurate and efficient molecular dynamics based on machine learning and non von Neumann architecture	Mo, P., Li, C., Zhao, D. et al.	2022	https://doi.org/10.1038/s41524-022-00773-
npj Computational Materials	Temperature-and vacancy-concentration-dependence of heat transport in Li3ClO from multi-method numerical simulations	Pegolo P, Baroni S, Grasselli F.	2022	doi.org/10.1038/s41524-021-00693-4
Advanced Science	Self-Healing Mechanism of Lithium in Lithium Metal	Jiao J, Lai G, Zhao L, et al.	2022	doi.org/10.1002/advs.202105574
Angewandte Chemie	Unravelling the fast alkali-ion dynamics in paramagnetic battery materials combined with NMR and deep-potential molecular dynamics simulation	Lin M, Liu X, Xiang Y, et al.	2021	doi.org/10.1002/ange.202102740
Angewandte Chemie	Simulation vs. Understanding: A Tension, in Quantum Chemistry and Beyond. Part A. Stage Setting	R. Hoffmann, J.-P.	2020	https://doi.org/10.1002/anie.201902527
Computational and Theoretical Chemistry	Crystal Structure Prediction of Binary Alloys via Deep Potential	Haidi Wang, Yuzhi Zhang, Linfeng Zhang, Han Wang	2020	doi.org/10.3389/fchem.2020.589795
Soft Matter	Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved	Shu Wang, Zhan Ma, Wenxiao Pan	2020	https://doi.org/10.1039/D0SM01019G
Journal of Materials Science & Technology	Theoretical prediction on thermal and mechanical properties of high entropy (Zr0.2Hf0.2Ti0.2Nb0.2Ta0.2)C by deep learning potential	Fu-Zhi Dai, Bo Wen, Yinjie Sun, Huimin Xiang, Yanchun Zhou	2020	doi.org/10.1016/j.jmst.2020.01.005.
MRS Communications	Carbyne: from the elusive allotrope to stable carbon atom wires	C.S Casari, A.Milani	2018	doi.org/10.1557/mrc.2018.48
Journal of Chemical Theory and Computation	Combined Deep Learning and Classical Potential Approach for Modeling Diffusion in UiO-66	Siddarth K. Achar, Jacob J. Wardzala, Leonardo Bernasconi, Linfeng Zhang and .Karl Johnson	2022	https://doi.org/10.1021/acs.jctc.2c00010

Materials Science in Semiconductor Processing	Efficient and accurate atomistic modeling of dopant migration using deep neural network	Xi Ding, , Jie Liu et al.	2022	https://doi.org/10.1016/j.mssp.2022.106513
ACS Applied Materials & Interfaces	Accelerated Deep Learning Dynamics for Atomic Layer Deposition of Al (Me) 3 and Water on OH/Si (111)	Nakata H, Filatov M, Choi C H.	2022	DOI: 10.1021/acsami.2c01768
Geophysical Research Letters	Anomalous Behavior of Viscosity and Electrical Conductivity of MgSiO3 Melt at Mantle Conditions	Haiyang Luo, Bijaya B. Karki, Dipta B. Ghosh, Huiming Bao	2021	10.1029/2021GL093573
Geophysical Research Letters	Thermal Conductivity of Silicate Liquid Determined by Machine Learning Potentials	Jie Deng and Lars Stixrude	2021	10.1029/2021GL093806
Geochemical Perspectives Letters	Diffusional fractionation of helium isotopes in silicate melts	Luo, H., Karki, B.B., Ghosh, D.B., Bao, H	2021	doi: 10.7185/geochemlet.2128
Geochimica et Cosmochimica Acta	Deep neural network potentials for diffusional lithium isotope fractionation in silicate melts	Haiyang Luo, Bijaya B. Karki, Dipta B. Ghosh, Huiming Bao	2021	doi.org/10.1016/j.gca.2021.03.031.
Computational Materials Science	Development of neural network potential for MD simulation and its application to TiN	Takeru Miyagawa, Kazuki Mori, Nobuhiko Kato, Akio Yonezu	2022	https://doi.org/10.1016/j.commatsci.2022.111303
Solar Energy Materials and Solar Cells	Local structure elucidation and properties prediction on KCl-CaCl2 molten salt: A deep potential molecular dynamics study	Bu M, Liang W, Lu G, et al.	2021	doi.org/10.1016/j.solmat.2021.111346
The Journal of Physical Chemistry Letters	Exploring complex reaction networks using neural network-based molecular dynamics simulation	Chu Q, Luo K H, Chen D.	2022	doi.org/10.1021/acs.jpclett.2c00647
Inorganic Chemistry Frontiers	Theoretical study of Na+ transport in the solid-state electrolyte Na 3OBr based on deep potential molecular dynamics	Li H X, Zhou X Y, Wang Y C, et al.	2021	doi.org/10.1039/D0QI00921K
Combustion and Flame	Training convolutional neural networks to estimate turbulent sub-grid scale reaction rates	Lapeyre C J, Misdariis A, Cazard N, et al.	2019	doi.org/10.1016/j.combustflame.2019.02.019
Applied Physics Letters	Deep learning inter-atomic potential model for accurate irradiation damage simulations	Wang H, Guo X, Zhang L, et al.	2019	doi.org/10.1063/1.5098061

Energy & Fuels	Exploring the Chemical Space of Linear Alkane Pyrolysis via Deep Potential GENerator	Zeng J, Zhang L, Wang H, et al.	2020	doi.org/10.1021/acs.energyfuels.0c03211
Physical Chemistry Chemical Physics	b initio neural network MD simulation of thermal decomposition of a high energy material CL-20/TNT	Cao L, Zeng J, Wang B, et al.	2022	doi.org/10.1039/D2CP00710J
Computational Materials Science	Molecular dynamics simulations of lanthanum chloride by deep learning potential[	Feng T, Zhao J, Liang W, et al.	2022	doi.org/10.1016/j.commatsci.2021.111014
ChemPhysChem	Exploring the Effects of Ionic Defects on the Stability of CsPbI3 with a Deep Learning Potential	Yang W, Li J, Chen X, et al.	2022	doi.org/10.1002/cphc.202100841
Physics of Plasmas	Warm dense matter simulation via electron temperature dependent deep potential molecular dynamics	Zhang Y, Gao C, Liu Q, et al.	2020	doi.org/10.1063/5.0023265
Theoretical and Computational Chemistry	Growth of Polycyclic Aromatic Hydrocarbon and Soot Inception by in silico Simulation	Wang B, Zeng J, Cao L, et al.	2022	10.26434/chemrxiv-2022-qp8fc
Energy and AI	Machine learning for combustion	Zhou L, Song Y, Ji W, et al.	2022	doi.org/10.1016/j.egyai.2021.100128
Energy and AI	Nanotwinning induced decreased lattice thermal conductivity of high temperature thermoelectric boron subphosphide (B12P2) from deep learning potential simulations	Xiaona Huang, Yidi Shen, Qi An	2022	doi.org/10.1016/j.egyai.2022.100135.
Physics Review b	Accurate force field of two-dimensional ferroelectrics from deep learning	Wu, Jing Liu, Shi et al.	2021	10.1103/PhysRevB.104.174107
Proteomics	DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions.	Zhang, F., Song, H., Zeng, M., Li, Y., Kurgan, L., & Li, M.	2019	doi.org/10.1002/pmic.201900019
ChemRxiv	Uni-Mol: A Universal 3D Molecular Representation Learning Framework	Zhou G, Gao Z, Ding Q, et al.	2022	DOI:10.26434/chemrxiv-2022-jjm0j
Journal of Controlled Release	Computational pharmaceutics-A new paradigm of drug delivery	Wang W, Ye Z, Gao H, et al.	2021	DOI: 10.1016/j.jconrel.2021.08.030
Communications of the ACM	Artificial intelligence for synthetic biology.	Mohammed Eslami, Aaron Adler, Rajmonda S. et al.	2022	doi.org/10.1145/3500922

Proc Natl Acad Sci USA	Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence	Washburn JD, Mejia-Guerra MK, Ramstein G., et al.	2019	doi.org/10.1073/pnas.1814551116
Nature Computational Science	Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics	Wang D, Wang Y, Chang J, et al.	2022	doi.org/10.1038/s43588-021-00173-1

* 列表仅收录了报告中所提及的部分文章，排序不分先后