我们必须打破世界

Markus J. Buehler
@ProfBuehlerMIT · Feb 4

如果你在一个普通的 LLM 上训练牛顿所有的著作，然后问它当粒子穿过两个狭缝时会发生什么，它会告诉你粒子会落在两堆上。它永远不会预测到实际出现的干涉模式——那个表明粒子以某种方式同时穿过两个狭缝的模式！这个答案隐藏在经典力学中；它被经典力学所禁止，概率模型也不会揭示它。你无法通过插值实现范式转变！根据定义，发现在数据中创造了不连续性。如果我们训练系统最小化惊讶并最大化可能性，我们实际上是在训练它们抑制科学革命所在的异常现象。在封闭的理论系统内学习不会产生该系统所排除的假设。学习者必须能够重写系统本身。

我的论点很简单：发现需要(1)组合式世界模型构建、(2)对抗性证伪和(3)物理基础的机制。本文的其余部分将介绍我们如何组装这些机制——智能体、图引擎、逆向设计和与模拟器和制造系统相结合的 AI 集群。

在封闭的理论系统上进行统计学习无法推导出该系统所禁止的现象。双缝实验的衍射图案（模拟：https://javalab.org/en/youngs_double_slit_en/）。

当我们在科学领域部署 AI——从材料工程到合成生物学——我们面临一个关键转折点。我们可以构建能够高效检索已知信息的 AI"助手"，或者构建真正具有创造力、发现能力和合成能力的 AI。要实现后者，我们必须设计不仅能够阅读文明集体著作，还能积极撰写新章节、构建物质并改变历史进程的系统。

前向模型的局限性

在我职业生涯的前二十年里，我在麻省理工学院的实验室专注于第一性原理模拟。我们逐个原子地模拟了蜘蛛丝的断裂力学和胶原蛋白的展开过程。这就是"正向问题"：给定一个结构，预测其性质。虽然这种方法很强大，但它不具备可扩展性。当你试图捕捉生物涌现的深层复杂性时，无法通过模拟来穷尽化学空间的组合无限性。此外，我们开始渴望能够创造更深层次洞察的算法——不仅仅是单一模拟的结果，而是一套更深刻、更基础的洞见。

模拟结构（左）以预测性质（右）。虽然准确，但当面对化学空间的无限性时，这种方法会遇到计算瓶颈。

我们转向范畴论，寻求不同领域之间的数学同构——将蜘蛛网的层次结构与音乐的组合结构进行映射。这个见解简单而深刻：如果我理解蜘蛛丝的工作原理，也理解音乐，那么我就有了一个*原理*在起作用。我可以问：这个原理还能在哪些地方应用？子混音、复合材料、完全不同的材料系统。

这是一项很棒的工作，但它只是纸笔上的工作，而且进展缓慢。整个过程完全由人类智能驱动——我们会与这些图表和图形一起坐上数月，寻找能够将一个领域的解决方案映射到另一个领域的函子。我们需要一种方法来自动发现这些关系。

利用范畴论寻求不同领域间的数学同构 - 将蜘蛛网的层次结构与音乐的组合结构进行映射。

所有有趣的事物都发生在事物之间

几年前发生的变化是我们意识到可以自动化这个推理过程。我们早期在 Transformer 方面的工作就派上了用场——不仅是作为文本生成器，更是作为图引擎，用于生成类别、形式推理和对世界进行组合式重构。范畴论是保持映射结构的语言；Transformer 为我们提供了可扩展的基础，用于在约束条件下发现候选对齐，这些对齐表现为学习到的关系图。

思考注意力机制中发生的事情。第一步，模型形成一个软关系图（一个学习到的邻接关系）——计算标记之间的关系，决定什么连接到什么。第二步，它在前馈层中对该图进行计算。这与预定义的图神经网络有着根本的不同。Transformer 发现自身的结构关系，然后基于此进行推理。

通过分层推理引擎进行情境化世界构建。

这一洞见是构建 AI 群体的"缺失环节"。如果我们想要建模物理，就需要建模关系。偏微分方程是跨越时空的场之间的约束。量子力学是振幅和可观测量上的约束系统。如果我们的智能体能"思考"图——将知识表示为符号结构——那么它们就能用我们过去用范畴理论所做的事情：跨领域搜索同构模式，组合出任何单一目的模型都无法找到的解决方案。

重新连接知识的图表示，用于全新假设和组合推理。参见：SciAgents (https://doi.org/10.1002/adma.202413523

发现的架构

我们正在超越简单的代理模型，转向展现涌现智能的系统。关键转变是从正向问题转向逆向问题：

> **旧方法：** "这是一个蛋白质序列；告诉我它的力-延伸曲线。"
>
> **新方法：** "我需要一种具有这种特定非线性力学响应的材料。设计能产生这种响应的蛋白质序列。"

这听起来可能像是一个渐进式的变化，但它代表了我们进行科学研究方式的根本性转变。让我给你一个贯穿我整个职业生涯的例子。

在2000年代初，研究人员发表了里程碑式的工作，他们拉伸单个蛋白质分子——固定一端，拉伸另一端，逐个原子地测量其力-延伸行为。当时，我的实验室通过仔细的模拟和分析，为这些行为创建了标度律。但这始终是正向问题：给定一个蛋白质，它的力学特性是什么？

ForceGen - 一种逆向模型，用于设计蛋白质以实现非线性力学响应。ForceGen，详见：https://www.science.org/doi/10.1126/sciadv.adl4000.

借助我们的新工具，我们现在可以指定所需的力-延伸曲线——任何任意的非线性力学响应——然后系统会设计出能产生该响应的蛋白质序列。它给我提供氨基酸序列。我可以用 AlphaFold 折叠它，并使用分子动力学模拟其物理特性。我可以模拟它并观察它的工作过程。

如果五年前有人告诉我这是可行的，我会相当怀疑。那时我们甚至无法可靠地折叠蛋白质！现在，这不仅成为可能，而且机器可以创建整个代码库来开发此类模型、训练它们、验证它们、应用它们并进行实验。这就是指数级起飞的样子——涌现、协同作用、如今已成为现实的丰富可能性。

VibeGen：为定制运动设计蛋白质。

从专家模型到推理系统

这些逆向设计能力最初来自专门的代理模型。但真正的突破在于我们将它们组合成一个专家系统。

在早期阶段，这催生了 X-LoRA，我将其视为一个原始群体。对于每个输出标记，我们生成一个无声的"思考标记"。这个无声标记（不是自然语言推理，而是内部路由/控制潜在变量）预测了神经网络本身的组织方式：对于这个特定的推理步骤，模块应该如何连接在一起？

X-LoRA 受物质组学原理启发，通过重组结构产生新功能，而不引入新的构建模块。

你可以观察系统在处理查询时如何逐层重新配置自己的架构。专家们用他们嵌入的母语相互交流，无需解码为自然语言。这种方法效果显著——但它仍然是一个单一模型"与自己对话"。为了真正扩展发现能力，我们需要将模型拆分。拆分模型意味着将对抗性动态分布到自主智能体中。

原型集群，X-LoRA。

证明自己是错的技巧

我们探索智能体（如我们的 SPARKS 系统）的核心组成部分是对抗性协作。我们不直接指导 AI 寻找答案，而是要求它找到一个*原理*。该系统有两个基本角色：解释智能体试图将观察到的数据压缩为通用原理或标度律，即能够解释所有已知现象的最简单解释，同时最小化柯尔莫哥洛夫复杂度（一个好的原理能用少量比特编码大量观测数据）。破坏智能体主动生成假设或寻找能够*违反*该原理的数据，试图打破当前的世界模型。

这种递归循环本身反映了科学方法。系统在综合理论和证伪理论之间振荡，推动智能体脱离训练分布，进入新颖领域。它们共同构成了任何单一智能体都无法实现的东西：一个能够观察自身思考的模型——这是打破世界的关键智能体组件。

奖励函数是对科学进步的形式化：最大化解释现象的覆盖范围，同时对未解决的数据进行惩罚，并倾向于更简单的理论。当发现解释者无法处理的新数据点时，解释者必须更新其对世界的模型。

可视化组合的无限性：绿色簇是进化赋予我们的蛋白质。广阔的黄色区域是我们必须打破世界才能探索的"禁区"。迄今为止，我们只看到了可能性的一个小而熟悉的片段。

输出结果令人瞩目。我们的 SPARKS 系统不仅提供一次性解决方案，而是产生标度律、数学表达式和新颖的假设。十年前可能需要一篇博士论文才能完成的工作，现在通过数十或数百次对抗性迭代涌现出来，这些智能体群体协同工作（或相互对抗），即时构建新的世界模型。

使用 mistral.rs 进行推理，针对高吞吐量智能体推理进行了优化。

群体涌现

这里有一个引人深思的问题：由通用 GPT 类前沿模型组成的群体，能否比专门构建的蛋白质扩散模型更好地设计蛋白质？

令人惊讶的是，答案是肯定的。

我称之为"蚂蚁桥梁"现象。单个蚂蚁的智力有限，它们不理解桥梁或建筑学。但集体而言，它们构建的结构在功能和结构上都比自身大得多。这就是涌现：小智力的结合产生了远超其总和的东西。

生物学中的涌现现象，这里是一个蚂蚁桥。

我们在 AI 代理中也发现了相同的涌现行为。我们的 SWARMS 系统从完全相同的代理开始——都是相同的基础模型，没有预定的角色。我们给它们一个设计问题和奖励信号。接下来发生的事情令人瞩目。

这些智能体开始分化。无需任何特殊训练，它们发展出不同的专业方向。一些成为批评者，一些成为规划者，一些成为验证者，一些成为推动边界的创新力量。智能体本身是标准的前沿模型。打破瓶颈的关键在于更多的智能体，它们对抗性地组织起来，通过大规模奖励进行学习。这就是通用性与多样性范式在发挥作用：相同的构建模块，经过倍增和重组，产生涌现功能。群体自我组织成功能性的劳动分工，共同解决了那些能够击败单一用途模型的复杂蛋白质设计问题。

智能群体利用情境学习和涌现智能设计新蛋白质。

最重要的是：它们设计的蛋白质与天然蛋白质相比 *远* 超出分布范围。我们已经跨越了从检索、插值到真正的组合合成的门槛。智能体在推理过程中 *学会了蛋白质设计的艺术* ——通过协作、失败、相互批评和迭代。

这让我们更接近从一开始就设想的真正自适应系统。

自然无法被欺骗

人们合理地担心，生成式人工智能将用"垃圾"淹没科学界——我们将得到听起来合理但纯属幻觉的胡言乱语，这些内容要么缺乏新颖性，无法在实验室中制造，或者明显违反物理定律。费曼在挑战者号事故后的警告在此适用：自然无法被欺骗。防御措施是通过第一性原理推理进行物理基础验证。

组合推理为预测提供了可验证的方法。

AI 科学家不能仅仅存在于概率世界中。推理循环必须包含现实检验。这意味着将 AI 群体直接与物理模拟器（DFT、有限元分析）以及最终的自动化制造相结合。在我们系统中，一个"想法"只是一个假设，直到它经受住现实的严酷考验。

这与当前模型的工作方式有着根本不同。如果你询问一个 LLM 其预测是否正确，它无法知晓——它被训练来预测文本最可能的延续，而非验证物理事实。我们需要能够从外部角度审视自身预测、比较不同模型版本、并根据证据更新信念的系统。

增材制造旨在与 AI 代理（如 BEAVER）原生集成。

我们已经构建了 AI 原生的 3D 打印机，旨在使代理系统能够轻松原型化想法并进行物理测试。计算推理、模拟和物理制造之间的三角关系至关重要。你可以完成惊人的计算工作，但归根结底，我们生活在一个物理世界中，材料必须真正发挥作用。

扩展发现

我经常被问到的一个问题是：学术研究仍然依赖于这些能力出现之前的工作流程。是什么阻碍了其采用？

答案是，我们不能通过培训每个人来使用 AI 来扩展规模。这就像回到 1980 年代，告诉每个人他们需要学习 C++和终端脚本。我们并没有通过这种方式扩展计算能力。我们通过创建操作系统、图形界面和使这些功能易于使用的产品来扩展。

AI 科学也是如此。我们需要足够易于使用的系统，让基因测序领域的生物学专家——他们没有兴趣成为 AI 专家——能够有效利用这些能力。这是一个技术开发挑战，但更重要的是，这是一个产品开发挑战。

我们部署的 AI 科学家越多，采用率就越高，他们产生的数据就越多，他们之间的联系就越多。这形成了一个网络效应。一个研究人员的代理可以验证另一个人的假设。系统可以相互通信，就像 Claude Code 实例在编写软件时已经相互通信一样。

SPARKS 产生新的科学发现，例如标度定律和理论。

想象一下，拥有数百名人工智能博士生同时运行实验，每个人都能够设计材料、模拟材料、在自动化实验室中制作原型，并报告结果——而你作为整个操作的监督者。我们还没有达到那个阶段，但组件已经存在。20 年前的纯理论博士生——编写 DFT 输入文件、运行模拟、分析结果——已经可以自动化。实验方面更具挑战性，但正在逐步实现。

构建世界构建机器

让我印象深刻的是，我们正在从仪器的手动操作者转变为智能的架构师。我们正在成为进步的导演，而不仅仅是工具。

愿景是一台不仅能预测新事物，还能*实现*它的机器。它成为一个塑造世界的系统：构建存在于世界中的材料，创造改变我们对可能认知的知识。这些不仅仅是预测，它们是机器想象之前并不存在的物理现实！

通过制造将想法变为现实。

我们正在构建具有自我指涉能力的系统，能够知道何时预测违反物理定律，并有能力更新自己的信念。能够以图形思考、组合推理、证伪自身理论，并在推理过程中分化为专门智能的系统。这超越了构建预测世界现状的 AI 模型。我们正在训练它们想象世界可能的样子——然后赋予它们构建世界的工具。

随之而来的是非凡的责任。如果这些系统成功运行，它们将代表人类能力的根本性增强。我们必须明智地使用它们。

检索与发现。我们可以在已知真理的死水潭中垂钓，或者建造能够塑造全新材料、思想和发现图景的机器。

要实现这一目标，我们必须首先愿意打破我们所认知的一切：从从已知真理中垂钓，转变为建造能够塑造新思想、洞察和原则景观的机器。我们必须打破当前的世界，才能构建下一个世界。

参考文献与注释

SciAgents: https://doi.org/10.1002/adma.202413523

ForceGen: https://www.science.org/doi/10.1126/sciadv.adl4000

mistral.rs: https://github.com/EricLBuehler/mistral.rs

X-LoRA: https://doi.org/10.1063/5.0203126

生物材料智能: https://link.springer.com/article/10.1557/s43577-025-00987-8

材料组学与范畴理论: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023911

VibeGen: https://arxiv.org/abs/2502.10173