

# 大健康与新医疗

BIG DATA Health  
and New Medical

2023 年  
第 23

上海科学技术情报研究所  
上海市前沿技术发展研究中心  
技术与创新支持中心(TISC)

# 人工智能在蛋白质设计中的发展与应用

## 编者按

蛋白质是各类生命活动不可缺少的承担者，其序列决定了折叠后的三维结构和功能。这些具有特定功能的蛋白质在生物医学等多个领域具有重要的应用价值。计算蛋白质设计可以根据所需的蛋白功能和结构设计氨基酸序列，生成自然界中不存在的蛋白质。传统计算蛋白质设计通常采用能量函数和特定的搜索优化算法获得设计的序列。近年来，随着先进算法的发展、大数据的积累和计算机硬件算力的增长，人工智能技术得到了蓬勃发展，并逐渐应用于蛋白质设计领域。

# 目 录

目 录	2
基础研究	3
➤ David Baker 团队设计出响应刺激的双态铰链蛋白	3
➤ AI 制药公司推出生成式 AI 蛋白质设计模型	4
➤ 生成式 AI 设计出非天然蛋白质	6
资本投入	7
➤ 分子之心宣布获得超亿元战略投资	7
➤ AI 蛋白质设计公司力文所完成数千万元天使轮融资	7
➤ 前 Meta 研究人员成立 EvolutionaryScale，完成 4000 万美元种子轮融资	8
应用案例	10
➤ AI 生成蛋白平台 AIGP	10

## 基础研究

### David Baker 团队设计出响应刺激的双态铰链蛋白

通过人工智能（AI）辅助设计，开发出一种铰链样蛋白质，这种铰链蛋白同时具有两种明确的构象，在与目标蛋白结合时显示出稳定的构象变化，因此可以根据这种特异性定制出“蛋白质开关”。这项工作为产生响应生物刺激的蛋白质开关提供了研究基础，为蛋白设计领域带来全新变革！

在自然界中，许多天然蛋白质可以在两种构象之间转换以响应环境刺激，例如目标分子的结合、翻译后修饰或 pH 的变化。这个过程在结构上类似于晶体管控制计算设备中的电子信息流，只不过蛋白质传递的是生化信息。

然而，天然蛋白质都是经过漫长的进化才衍化得来，人工设计的蛋白往往很难做到在两种折叠状态之间进行这样的构象转换。因为人工蛋白质设计通常旨在优化一个单一的、非常稳定的构象，使其拥有折叠能量景观的全局最小值。

因此，如何设计出同时具有两种不同构象且结构完整的人工蛋白质是蛋白质设计领域的一个重大挑战。其中，最关键的一个要求是，所设计的蛋白质需要同时具有两个不同最小值的能量景观，就像一座山峰之中同时有两个山坳，

一片沙漠之中同时有两个绿洲。

在这项研究中，David Baker 团队描述了一种铰链蛋白的设计概念，这种蛋白具有两个明确定义和结构的构象状态：在没有配体的情况下表现为 X 设计状态，而在配体存在的情况下通过构象变化转换为 Y 设计状态。

研究小组通过 X 射线晶体学、电子显微镜、双电子-电子共振光谱等技术对设计的铰链蛋白的蛋白质结构、结合动力学和构象平衡进行了全面表征。研究结果表明，尽管存在显著的结构差异，但这两种构象状态的设计具有原子水平的精度，并且构象平衡和结合平衡是紧密耦合的。

因此，这种铰链蛋白可以广泛适用于蛋白质开关的设计，就像是电子电路中的晶体管一样，研究人员可以将蛋白质开关与外部输出和输入耦合，其状态转换通过与天然蛋白（例如胰高血糖素、神经肽、分泌素）结合而不是人工设计肽。

基于这种设计的铰链蛋白可以用于创建生物传感设备，并将它们合并到更大的蛋白质系统中，以解决各种突出的设计挑战：铰链蛋白可以作为一种模块，在其内部安装特定的酶活位点，从而实现两种明确构象的切换——当底物结合

时有利于一种状态，当产物释放时有利于另一种状态。

这对于之前的 LOCKR 开关而言是不可能的，因为 LOCKR 开关只有一个明确的构象，另一个状态往往是无序的。缺乏定义的第二状态使得它不适合于构建蛋白质开关或基于离散状态的计算系统中的机械耦合。

总的来说，这项新研究开发出了同时具有两种明确状态的铰链蛋白，可以应用于“蛋白质开关”的设计。这种双态开关的研究能使蛋白质设计超越静态结构，转向更复杂的多态组装和生物传感器设计，为蛋白设计提供了更广阔的可能性。

**资料来源：**

[1]Science 重磅：AI 设计蛋白新突破，David Baker 团队设计出响应刺激的双态铰链蛋白！  
[EB/OL].(2023.08.22).[2024.01.08].<https://news.bioon.com/article/f50be87899c8.html>

[2]Praetorius F, Leung PJY, Tessmer MH, Broerman A, Demakis C, Dishman AF, Pillai A, Idris A, Juergens D, Dauparas J, Li X, Levine PM, Lamb M, Ballard RK, Gerben SR, Nguyen H, Kang A, Sankaran B, Bera AK, Volkman BF, Nivala J, Stoll S, Baker D. Design of stimulus-responsive two-state hinge proteins. *Science*. 2023 Aug 18;381(6659):754-760

## AI 制药公司推出生成式 AI 蛋白质设计模型

蛋白质是生命活动的执行者，但创造它们却是一项复杂的任务，需要数十亿年的进化。基于计算的蛋白质设计，旨在通过可编程的方式自动设计功能蛋



白来缩短这一漫长的进化过程。该领域在过去几十年里取得了相当大的进展，但大多数从头设计尚未接近自然界中天然蛋白质的复杂性和多样性。

地球诞生生命以来的 30 亿年时间里，进化产生了巨大的蛋白质多样性，然而，这在蛋白质的全部潜力面前微不足道，这一巨大潜力为我们从头设计蛋白质带来了无限可能，但确定如何有效地探索可设计蛋白质结构的空间，是当下面临的一个巨大挑战。

2023 年 11 月 15 日，生成式 AI 制药公司 Generate:Biomedicines 的研究人员在国际顶尖学术期刊 Nature 上发表了题为：Illuminating protein space with a programmable generative model（用可编程的生成式模型照亮蛋白质空间）的研究论文。

该研究开发了一种名为 Chroma 的生成式人工智能模型，该模型建立在扩散模型（Diffusion Models）和图神经网络（Graph Neural Networks）的框架上，能够从头生成高质量、多样化和创新的蛋白质结构。

研究团队使用 Chroma 生成了 310 个自然界中不存在的蛋白质，并通过实验验证了这些蛋白质可以表达、折叠，并具有良好的生物物理特性。



Generate:Biomedicines 成立于 2018 年，旨在通过人工智能（AI）技术来理解蛋白质序列、结构与其功能之间的关系，从而从头设计前所未见的全新蛋白质，定制蛋白质疗法，以改善肿瘤学、免疫学和传染病等领域的药物开发。

Generate:Biomedicines 目前已累计融资近 7 亿美元，2020 年 9 月，获得 Flagship 的 5000 万美元融资，2021 年 11 月，该公司完成了 3.7 亿美元 B 轮融资，还与安进公司达成了 19 亿美元的合作。2023 年 9 月，该公司完成了 2.73 亿美元 C 轮融资，值得一提的是，C 轮投资方还包括制药巨头安进（Amgen）、人工智能计算领导者英伟达（NVIDIA）。

在人工智能革命之前，蛋白质设计方法仅限于基于自然界已有的蛋白质生成设计，起局限性显而易见，因为自然界中的蛋白质只是可能的蛋白质景观的一小部分。相比之下，生成式人工智能方法强调从头设计全新的蛋白质，超越自然界所能达到的范围。

该研究开发的生成式人工智能模型 Chroma，能够在外部约束条件下从头设计蛋白质，这些约束条件涉及对称性、亚结构、形状，甚至自然语言提示。





研究团队对 310 个由 Chroma 生成的蛋白质进行了实验表征，结果显示，这些生成的、自然界不存在的蛋白质可以表达、折叠，并具有良好的生物物理特性。

研究团队还解析了其中 2 个生成的蛋白质（UNC\_079 和 UNC\_239）的 X 射线晶体结构，结果显示，观察到的结构与预期设计高度匹配（均方根误差分别为 1.1Å 和 1.0Å），这表明了用 Chroma 生成蛋白质结构是可行的。

探索蛋白质结构空间以产生物理上合理和可设计的构象，一直是蛋白质设计领域长期存在的挑战。现有的蛋白质设计工具往往把大量时间花费在了寻找合理的蛋白质骨架上，而忽略了设计的蛋白质的实际功能。

研究团队表示，Chroma 有潜力解决这个问题，使蛋白质设计的重点从生成可行的结构转向特定任务——即设计这个蛋白质要实现的目的。通过对 30 亿年中进化产生的蛋白质的学习，找到组装稳定蛋白质的新方法，像 Chroma 这样的生成式人工智能模型已经准备好推动生物分子多样性进入新时代。

**资料来源：**

[1]这家融资 7 亿美元的 AI 制药公司发表 Nature 论文，推出生成式 AI 蛋白质设计模型，已

免费开源[EB/OL].(2023.11.22).[2024.01.08].<https://news.bioon.com/article/389d80181263.html>

[2]Ingraham JB, Baranov M, Costello Z, et.al. Illuminating protein space with a programmable generative model. Nature. 2023 Nov 15

## 生成式 AI 设计出非天然蛋白质

据报道，加拿大多伦多大学研究人员开发了一种人工智能系统，可以使用生成扩散来创建自然界中不存在的蛋白质。该系统有望使治疗蛋白的设计和测试更加高效和灵活，从而加速人类药物开发。研究发表在最新一期《自然·计算科学》杂志上。

蛋白质由氨基酸链组成，氨基酸链折叠成的三维形状反过来又决定了蛋白质的功能。这些折叠的三维形状经过数十亿年的发展，多种多样且复杂，但数量是有限的。因此，研究人员开始尝试设计非自然界产生的折叠模式。

这一研究的主要难题是对折叠的“想象”，因为很难预测哪种折叠是真实的，并在蛋白质结构中起作用。通过将基于生物物理学的蛋白质结构表示与图像生成空间的扩散方法相结合，科学家找到了解决这个问题的途径，创建了被称为 ProteinSGM 的新系统。

该模型从图像表示（图像信息在计算机中的表示和存储方式）中学习，

并以非常高的速度生成全新的蛋白质。研究人员表示，除了优化图像生成过程存在挑战外，对系统产生的蛋白质进行验证也很困难，因为该系统产生的许多结构与自然界中发现的任何结构都不同。

根据指标，几乎所有产生的结构看起来都合理，但研究人员需要进一步的证据。他们转向求助于人工智能“欧米伽折叠”（深度思维公司“阿尔法折叠 2”的改进版本），测试后确认，几乎所有的新序列都折叠成了所需的新蛋白质结构。再辅以实验室的物理测试，研究人员最终确信这些都是正确的蛋白质折叠。

**资料来源：**

生成式 AI 设计出非天然蛋白质[J]. 电子产品可靠性与环境试验，2023，41（3）：70

## 资本投入

### 分子之心宣布获得超亿元战略投资

2月20日，AI蛋白质设计平台公司分子之心宣布获得超亿元战略投资，由合成生物学龙头凯赛生物领投，联想创投跟投，天使轮领投资方红杉中国本轮继续追加投资。融资将用于AI蛋白质优化与设计平台 MoleculeOS 进一步开发，以及在生物制药、合成生物学等产业领域的应用探索。成立仅一年，分子



之心已经快速完成 2 轮融资。

分子之心表示，获得凯赛生物战略投资后，双方还将在业务层面展开深度合作，借助分子之心自研的 AI 蛋白质优化与设计平台 MoleculeOS，融合凯赛生物在合成生物领域的 20 余年产业经验，联合推动合成生物学产线升级和新品研发。

分子之心由全球知名计算生物学家、“AI 蛋白质折叠奠基人”许锦波教授创立。基于团队多年 AI 蛋白质结构预测与设计的经验，分子之心自主研发了国内首个功能完整的人工智能驱动的蛋白质预测和设计平台“MoleculeOS”，运用数据驱动的 AI 方法，快速识别、改造甚至从头设计最合适的蛋白质，从而颠覆大分子药物设计、合成生物学、环境保护等领域研发范式。当前，分子之心已经基于 MoleculeOS 平台构建全场景的 AI 蛋白质发现、优化与设计能力，其中蛋白质从头设计、蛋白质优化、蛋白质以及复合物结构预测、蛋白-蛋白对接、蛋白质侧链预测、蛋白质功能预测等十余项关键 AI 算法计算结果领先全球。

资料来源：

智药邦 2023.02.20 新闻



## AI 蛋白质设计公司力文所完成数千万元天使轮融资

近日，国内创新的 AI 蛋白质设计公司杭州力文所生物科技有限公司（简称“力文所”）宣布完成数千万元天使轮融资。本轮融资由凯泰资本领投，磐霖资本、红什资本跟投，种子轮领投方真格基金追加投资。融资资金将主要用于 AI 蛋白质设计平台的优化和开发，及推动平台孵化建设多条产品管线。力文所成立于 2021 年，致力于 AI 蛋白质设计研究。创始人王浩博是共进化和蛋白质设计专家，其在哈佛大学博士后期间师从 Sergey Ovchinnikov，学习蛋白质共进化和人工智能。Sergey 出身于蛋白质设计领域国际领军实验室 Baker Lab，其在 Alphafold2 时代，探索了基于 AF2 的蛋白质设计和扩散模型，是 AI 蛋白质设计领域的先驱。公司核心研发团队来自哈佛大学、波士顿大学、北京大学、中科院等国内外知名大学和研究机构。Lésign® 是力文所的蛋白质设计平台，该平台能将生物系统发育以及物理势能信息嵌入 AI 共进化分析模型，可有效利用自然界进化试错以及物理模型的校正，引入大量突变，产生多样蛋白序列；还可基于通用的扩散式生成模型，对蛋白结构进行重新设计生成。本



轮融资由凯泰资本领投，磐霖资本、红什资本跟投，种子轮领投资方真格基金追加投资。

资料来源：

AI 蛋白质设计公司「力文所」完成数千万元天使轮融资[EB/OL].(2023.07.27).[2024.01.08].  
<https://finance.sina.com.cn/tech/roll/2023-07-27/doc-imzeaura6527711.shtml>

## 前 Meta 研究人员成立 EvolutionaryScale，完成 4000 万美元种子轮融资

2023 年 6 月，来自著名的 Meta 公司的前研究人员成立的一家开创性的人工智能生物技术公司 EvolutionaryScale 成功获得了 4000 万美元种子轮融资，以通过大幅扩大其人工智能模型的规模来推进其研究工作。

此次融资由 Lux Capital 领投，著名人工智能投资者 Nat Friedman 和 Daniel Gross 跟投。完成后 EvolutionaryScale 的估值为 2 亿美元。

这家新公司由 Alexander Rives 领导，其曾是 Meta AI 蛋白质折叠团队的负责人，该部门今年 4 月被解散。

这支由八人组成的创新团队，全部来自同一公司。他们从 GPT-4 和 Bard 等项目中汲取灵感，巧妙地设计出一种基于 Transformer 的蛋白质预测模型--



ESMFold。他们的创造有望通过全面的数据分析彻底改变人们对蛋白质分子的理解，从而预测迄今为止未知蛋白质的结构。该人工智能模型利用一个以蛋白质分子为重点的庞大数据集进行了细致的训练，最终形成了一个拥有 7 亿个潜在三维结构的强大数据库。这些结构是打开解决方案宝库的钥匙，从治疗各种疾病的突破性药物开发，到解决污染和工业化学品生产的生态友好型方法，不一而足。

由氨基酸链编织而成的蛋白质作为一系列生物实体的基本构件，存在于细菌和人类细胞中。蛋白质的独特功能与其独特的形状密切相关，在与生物框架内的各种化学物质或其他蛋白质相互作用时，蛋白质的形状会发生改变。这些形状变化特征在靶向药物的开发中发挥着关键作用，可针对蛋白质的特定区域来解决疾病问题。然而，预测这些蛋白质结构的任务是一个错综复杂的难题。这种预测能力赋予科学家解码蛋白质功能的能力，从而引导药物设计的方向，并将重点放在三维构型上。

2020 年，谷歌的子公司 DeepMind 推出了 AlphaFold，具有预测蛋白质

结构的卓越能力。诺贝尔奖获得者 Venki Ramakrishnan 对这一里程碑式的成就大加赞赏，称其为"惊人的进步"，有望彻底改变生物研究。这一突破无疑推动了这一领域的发展，但蛋白质与潜在靶向药物之间错综复杂的相互作用仍然是一项艰巨的挑战，因为阐明这些相互作用的复杂性是一个巨大的障碍。

在去年 11 月，该团队在《Science》杂志上发表了一篇引人注目的论文，强调了他们的模型比 AlphaFold 快了 60 倍这一了不起的成就。值得注意的是，这些预测在速度和准确性之间进行了微妙的权衡，不过人工智能与药物开发的结合主要是逐步提高了效率。另一方面生物研究中类似于文本研究的变革性突飞猛进的分水岭时刻尚未到来。因此传统制药企业对人工智能和生物学最终的融合仍持怀疑态度。

但值得注意的是，基于人工智能研究领域有了许多巨额投资。如 Inflection AI 在 6 月份获得了 13 亿美元的巨额融资，Cohere 在 5 月份宣布获得 2.7 亿美元的巨额融资，掀起了轩然大波。Adept 最近在 3 月份获得了 3.5 亿美元的注资。这股热潮延伸到了人工智能的基础架构，备受关注的



Hugging Face 披露的 2.35 亿美元巨额投资就凸显了这一点，该公司的估值达到了惊人的 45 亿美元。

Rives 和他富有远见的团队以务实的态度认识到了他们工作的"登月"性质。EvolutionaryScale 预计其第一年将花费 3800 万美元，其中 1600 万美元用于开发计算能力。随着计划的展开，成本也会激增，第二年将达到 1.61 亿美元，第三年将达到 2.78 亿美元。在这个复杂的开发计算能力的过程中，计算费用分别达到 1 亿美元和 2 亿美元。在整个战略蓝图中，始终回荡着一个主题：生物人工智能模型产生实际产品和疗法的潜力是一项需要时间的工作，可能会跨越漫长的十年。

EvolutionaryScale 的开创性方法和巨额资金标志着人工智能和生物技术的融合迈出了关键一步。虽然人工智能在生物学领域的变革潜力尚未完全商业化，但像 EvolutionaryScale 这样的初创企业正在塑造未来创新的格局，强调了持续投资和合作的必要性，以实现承诺的突破。

**资料来源：**

[1]智药邦 2023.09.06 新闻

[2]Former Meta researchers established AI biotech startup EvolutionaryScale, securing \$40 million funding from Lux Capital[EB/OL],[2024.01.08].



<https://multiplatform.ai/former-meta-researchers-established-ai-biotech-startup-evolutionaryscale-securing-40-million-funding-from-lux-capital/>

## 应用案例

### AI 生成蛋白平台 AIGP

百度创始人李彦宏牵头创立的生命科学平台公司百图生科，正式对外发布其“生命科学版 ChatGPT”——由生命科学大模型驱动的 AI 生成蛋白平台 AIGP。

恰如人类自然语言可以拆解成 26 个字母、词、句子、段落的嵌套结构，生命语言可拆解成 20 个氨基酸字母、蛋白质、细胞、生命体。如此多的相似性，使得生成式 AI 不仅能写诗编程，也能被用于解决困扰生命科学科研人员已久的难题。

AIGP 平台由百图生科在过去 2 年多时间内、近 300 位 AI+BioTech 专家组成的跨国技术团队共同打造。根据不同模块的输入和要求，AIGP 平台可在较短时间内设计和生成具有特定性质的蛋白质。据介绍，这将大幅简化蛋白质生产流程，用几十分钟、几小时来设计蛋白质，几天就能把蛋白质制备出来。



其背后的核心引擎，是百图生科成立两年多以来一直致力于打造的生命科学大模型 xTrimo。该模型拥有千亿参数，从跨物种、跨模态的生命信息中学习蛋白质如何构成和实现功能、如何相互作用、如何组合和调控细胞功能的关键规律。

有近 20 家合作伙伴已与百图生科开展 AIGP 联合研发合作，方向涉及高性能弹头设计、新功能蛋白质设计、靶点挖掘和调控蛋白设计等领域，其中多个项目取得了阶段性的发现成果

### **AIGP 平台：3 大功能模块，12 项核心能力，6 月上线**

目前，百图生科 AIGP 平台设置了 3 大功能模块。

(1) Function to Protein Design (F2P)：从功能到蛋白设计，根据结构、功能、可开发性等指标设计/优化蛋白质，生成形状和理化性质的需求，生成一系列满足用户需求的蛋白设计，包括新功能、新结构、新酶，并提供蛋白质参数优化能力。百图生科企业发展副总裁、产业基金董事总经理瞿佳润 (Vicky) 分享说，这类功能更多是跟合成生物学公司、酶类公司以及一些保健、医美行业有相关性。



(2) Protein to Protein Design (P2P) : 从蛋白到蛋白设计，给定抗原等目标蛋白，设计与之以特定方式结合的抗体等蛋白，比如针对新冠、渐冻症等生成蛋白，根据石油等分子生成酶。该部分提供四个功能点：高亲和力、高精度表位、高特异性、高序列差异化的设计。在这个方向，百图生科的主要合作方是 Biotech 和 Pharma，可以设定到一些大分子药物。

(3) Cell to Protein Design (C2P) : 从细胞到蛋白设计，给定细胞，发现调控细胞功能的靶点蛋白并设计相应的调控蛋白。比如输入胃癌病人的疾病信息，AIGP 的平台能够分析胃癌病人多组学数据，找到病人的胃癌靶点，根据靶点生成一个抗体甚至一系列有多样性的蛋白，最后一键式生成的蛋白回到自动实验室里面合成蛋白。其功能点能够做到细胞分类、细胞调控靶点、组织特异性靶点、药物组合效应的预测。

由于制药需要高度专业性。AIGP 平台暂时不会面向大众开放。百图生科计划于 2023 年 6 月起将部分功能模块进一步开放，让专业用户可以直接自主使用，在更多的研究场景调用 AI 的蛋白质生成能力，激发更多的生命科学探索。

**生物学大模型 xTrimo : 预测蛋白质结构的速度、准确度超过**



## AlphaFold

AIGP 背后的千亿大模型体系，也有一个四层嵌套形式，最底层是对蛋白质的数据进行预训练，往上一层是蛋白质相互作用的预训练，还有一层是对细胞体系的预训练模型，最上层有一系列针对蛋白质设计特别关键的参数的预测模型。

要充分吸收生命科学里的数据，需将模型、训练和工程上的很多创新做有机融合。

百图生科 CTO 兼首席 AI 科学家宋乐说，ChatGPT 依赖于人类反馈来输出舒适的、对人没有恶意的交流，AI 生成蛋白质也是如此，需要经过专家信息及高通量实验室数据的输入，才能将生命科学大模型调整到能生成有用蛋白的程度。


生命科学大模型 xTrimo 的炼成过程是一个闭环，涉及大模型体系、高通量验证、数据平台三大部分，数据平台既有大量公开数据，也有百图生科自驾实验室产出的数据

**资料来源：**

生命科学迎“ChatGPT 时刻”！AIGP 平台三大功能加速蛋白质生成，6 月起向专业公众开放



[EB/OL].(2023.03.30).[2024.01.08].<https://zhuanlan.zhihu.com/p/618343344>



地址：上海市永福路 265 号  
邮编：200031  
编辑：李春霞  
责编：姚恒美  
编审：林鹤