

AI+蛋白质 行业研究报告

2024年4月

序言

在AlphaFold等人工智能技术的推动下,人类在蛋白质的研究中取得了历史性突破,AI作为一种革命性工具在生命科学领域展现出巨大潜力。

蛋白质是生命活动的主要承担者,与人类的健康和疾病密切相关。传统的蛋白质研究方法存在着实验周期长、成本高昂等问题。人工智能技术,特别是深度学习技术,已经被广泛应用于蛋白质的结构预测。通过训练深度学习模型,使其能够从氨基酸序列预测蛋白质的三维结构,加速蛋白质设计和改造,最终赋能药物研发与生物制造。

AI在蛋白质领域的突破,反映出人工智能正在成为支撑经济转型不可或缺的基础设施和核心能力,契合我国培育发展新质生产力的时代要求。作为“人工智能+”的关键领域,AI+蛋白质无疑将推动生命科学领域的产业发展,解决更多医药创新难题。

尽管人工智能在蛋白质研究领域展现出巨大的潜力,特别是在预测蛋白质结构和功能方面取得了显著成就,但它仍然面临着一些技术和实践上的挑战。例如,数据质量问题是AI应用中的一个关键限制因素。由于蛋白质结构的复杂性,需要大量高质量的数据来训练和验证AI模型。目前,公共数据库如PDB(蛋白质数据银行)提供了大量的结构信息,但这些数据往往存在偏差和不完整性,这可能影响AI模型的准确性和泛化能力。此外,算法的可解释性也是一个重要问题。AI模型,尤其是深度学习模型,通常被视为“黑箱”,其运行过程缺乏透明度,这在需要精确理解分子机制的生物学研究中尤为突出。

然而,随着AI技术的不断进步和蛋白质数据的日益丰富,这些问题有望得到解决。例如,随着高通量测序技术的发展,蛋白质序列数据能够以前所未有的速度和规模生成。此外,算力的增强和算法的优化,AI模型的学习能力和处理速度也在不断提升。这些进步预示着AI在蛋白质行业的应用将迎来更广阔的发展空间。

总之,尽管存在挑战,但随着技术的不断进步和数据资源的丰富,AI在蛋白质研究和应用方面的前景仍然充满希望。未来,我们可以期待AI在推动生命科学发展和改善人类健康方面发挥更加重要的作用。

目录

1	AI+蛋白质行业概述	01
2	AI+蛋白质技术分析	08
3	AI+蛋白质应用场景	22
4	AI+蛋白质投融资及发展预测	35



AI+蛋白质行业概述

1.1 发展阶段

20世纪60年代，科学家们主要依赖X射线、NMR等常规手段来解析蛋白质的结构；

2012年以来，以ResNet、Transformer为代表的深度学习算法被提出，为蛋白质研究带来革命性方法；2020年以来，AI与蛋白质研究相结合的产业应用实现了突破性增长；

2021年以来，AlphaFold2、RoseTTAFold等前沿算法的开源，加速了产业应用落地；

2022年以来，随着ProteinMPNN等工具的出现，标志着蛋白质设计进入新的发展阶段。

人工智能技术的不断进步，其在解析蛋白质三维结构的能力上已获得全球科学界的高度认可。特别是自2020年以来，AI与蛋白质研究相结合的产业应用实现了突破性增长，这一趋势不仅揭示了人工智能在生物医药领域的巨大应用潜力，也预示着其在未来可能为疾病治疗、药物设计等关键领域带来革命性的变革。

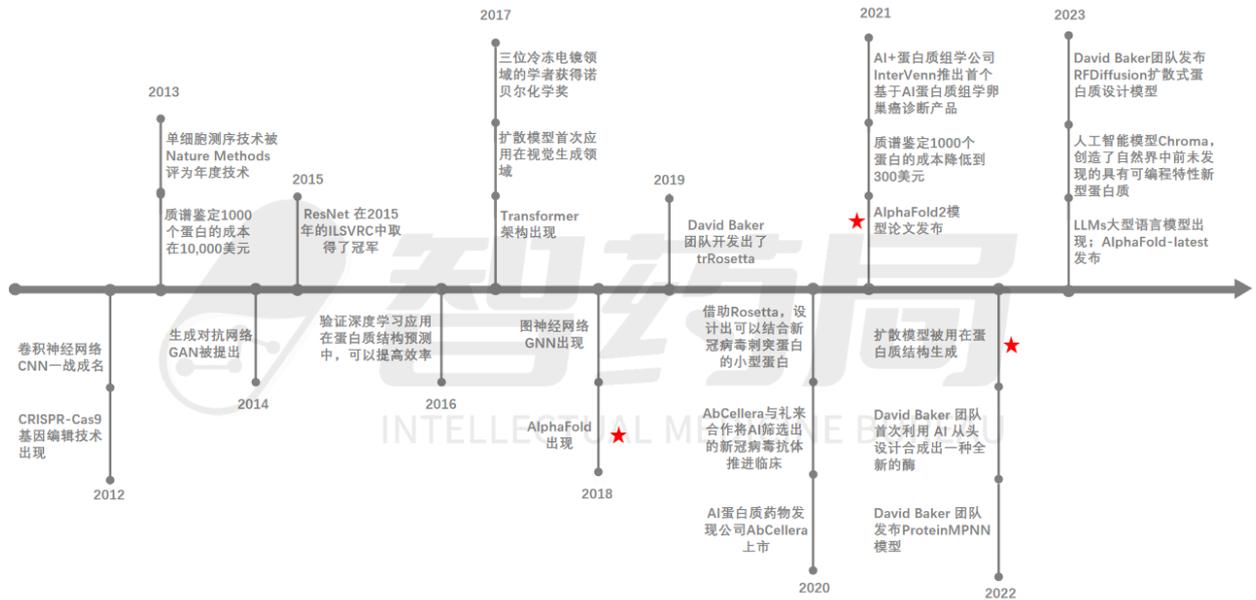
人类对蛋白质的研究可追溯至20世纪60年代，当时科学家们主要依赖X射线、NMR（核磁共振）等常规手段来解析蛋白质的结构。随着对蛋白质生物化学机制的深入理解和计算技术的不断突破，科学家们开始尝试利用计算方法进行蛋白质结构的预测。

2021年7月，DeepMind开源了AlphaFold 2.0，并推出了一个包含350,000种三维蛋白质结构的数据库。几个月后又公布了另一个包含2亿种蛋白质结构的数据库，几乎覆盖了所有科学上已知的蛋白质。这一进展几乎颠覆了传统蛋白质研究的方式，将解析一个蛋白质结构的时间缩短了十万倍。

蛋白质的研究不应仅限于对其结构的解析，而应更深入地探索其生物学功能及其与其他分子的相互作用。自AlphaFold取得突破后，机器学习和深度学习等人工智能技术被广泛应用于蛋白质功能注释、蛋白质-蛋白质相互作用以及蛋白质-小分子对接等方面，同时质谱和单细胞测序技术的进步，使得获取数据的成本不断降低。技术的迭代和成本的下降，像两台发动机加速下一代蛋白质组学的到来。

目前，基于 AI 的蛋白质结构、功能等的预测、蛋白质设计以及蛋白质组学研究已经广泛应用于新药研发、合成生物、疾病诊断等多个跨学科领域，这些应用不仅加速了新药从理论到临床的转化过程，还促进了对复杂生物系统的深入理解，为应对全球性的健康挑战和环境问题提供了创新的解决策略。

图 1-1 AI + 蛋白质行业的发展大事件



来源：智药局

自 2012 年以来，深度学习取得了突破性进展，在蛋白质研究领域展现出了显著的效果。尤以 AlphaFold 2.0 的出现为分水岭，开启了蛋白质结构预测的新纪元，在行业发展过程中，技术创新始终是第一推动力，从 CNN 到 Transformer，从基于结构到基于序列，AI 的变革不断带来新的可能性和变革性机会。

1.2 产业链分析

上游——算力、算法和数据构建产业基石，生命科学技术和设备的迭代是关键变量；

中游——以 AlphaFold 为首的蛋白质结构预测、以 David Baker 为代表的蛋白质设计以及人工智能与蛋白质组学相结合；

下游——分为药物研发和生物制造两大领域。

人工智能技术与生物技术的融合，包括对大规模蛋白质数据库的深度挖掘以及相关算法模型的层出不穷，已经使得研究人员能够更加精确地预测、优化和设计具有特定功能的蛋白质。这种技术的进步为疾病治疗、生物催化和材料科学等领域带来了新的研究工具和方法。

例如，David Baker 教授和他的团队利用这些先进技术，成功设计了一系列具有新颖结构和功能的蛋白质，这类工作不仅推动了基础科学的发展，也为实际应用开辟了新的可能性，如开发新型疫苗、生物催化剂和治疗性蛋白质。这些成果展示了人工智能在推动生物科学前沿研究方面的巨大潜力。

目前，人工智能在蛋白质产业链中的应用已经取得了显著的成果和进展。它不仅推动了生物医药领域的发展，还影响了化工、食品、能源等非医药领域，为产业链上下游企业带来了商机和挑战。

图 1-2 AI 蛋白质行业产业链



仅为部分企业展示, 企业排序不分先后

来源: 智药局

1.2.1 产业链上游

产业链上游包括支撑人工智能运行的算力、算法和数据，生物实验的硬件设备，以及基因编辑技术。

其中，英伟达、AMD 等产业龙头为 AI 提供了坚实的算力基础；谷歌云、亚马逊云和阿里云等成熟云服务供应商则为 AI 应用提供了稳定、高效的软件支持。在蛋白质相关的人工智能算法领域，国内大部分公司在 AI 研发方面选择采用开源的算法底层框架，如 TensorFlow、PyTorch 等。常用的包含蛋白质信息的数据库，比如，蛋白质结构信息 PDB 数据库、蛋白质序列和注释信息 UniProt 数据库、蛋白质功能信息 Gene Ontology 数据库、蛋白质-蛋白质相互作用信息 STRING 数据库等，为产业链的深入研究和发​​展提供了丰富的数据资源。

其次，生物制造所需的原材料、设施和先进技术也是上游环节的重要组成部分，例如，质谱仪、蛋白质试剂、冷冻电镜、基因编辑技术等都是这一环节的关键要素。

目前，全球质谱仪市场主要被赛默飞、安捷伦、WATERS、岛津等国际行业巨头垄断；国内以谱育科技、禾信仪器为代表的企业，在产品性能上已经达到国际水平。试剂供应商包括赛默飞、CST 等，为产业开展蛋白质组学活动提供基础试剂。冷冻电镜主要有三大供应商：赛默飞、日本电子和日立，是实验获得蛋白质结构的有力工具。

基因编辑技术的发展经历了从 ZFNs 到 TALENs，再到 CRISPR/Cas9 的演变过程，2012年发明第三代基因编辑技术，不仅实现了对人类细胞的基因编辑，与前两代相比，基因编辑成本下降超过90%，靶向精度逐渐达到临床要求。

在 AI 蛋白质产业链中，上游环节承担着关键的角色，构成了整个产业链的基石，而随着产业中下游的快速发展，尤其是大模型的兴起，对上游信息产出提出了更高的要求：不仅在数量上呈现出大规模的趋势，更在质量上追求更高的精确度和功能性。这种需求的增长将催动上游产业技术进一步发展。

1.2.2 产业链中游

技术创新是推动蛋白质结构预测领域发展的主旋律，目前，DeepMind、Profluent Bio、Genesis Therapeutics 等均以蛋白质预测为主要研究方向。国内百图生科、分子之心、华深智药、深势科技等公司也在深入探索该领域。

蛋白质从头设计作为一种前沿的生物技术，通过计算生物学和人工智能的深度融合，研究人员能够设计出具有特定结构和功能的蛋白质，从而颠覆传统的生物制造模式，为生命科学产业带来根本性的变革。代表企业

有 Arzeda、Generate:Biomedicines、分子之心、途深智合等。与此同时，蛋白质优化技术则更加注重实际应用场景的需求，这一领域的技术已经逐渐成熟，并开始进入商业化阶段。代表企业有 Cradle Bio、RevolKa、天鹭科技等。

AI 蛋白质组学公司利用人工智能技术对蛋白质组进行深入研究，以推动生物标记物发现、药物发现、体外诊断等应用场景的发展。国外以 Olink、Somalogic、Matchpoint 等为代表性企业，国内目前形成了以西湖欧米、珞米科技等为代表的专注于蛋白质组学研究的企业。

在 AI 蛋白质产业链中，中游环节作为连接上下游的关键纽带，其创新密度之高尤为引人注目，近年来该领域涌现出一批充满活力的初创公司。这类企业通常以差异化的技术和产品为基础，并积极探索商业模式和市场策略的创新，最终加快了生物医药、能源、食品等下游应用领域的发展进程。

1.2.3 产业链下游

人工智能+蛋白质的下游应用主要可以分为药物研发和生物制造两大类，两者都蕴含着巨大的潜力和广泛的应用前景。

AI 蛋白质预测和蛋白质设计可以大大加速新药研发进程。通过预测蛋白质的结构和功能，研究人员可以更准确地确定药物与蛋白质之间的相互作用，从而设计出更有效的小分子药物和生物药。还可以通过预测蛋白质与特定分子的相互作用来检测疾病标志物，这对在疾病的早期发现、诊断和预后评估方面具有重要价值。

在临床中，AI 蛋白质组学可以对各种临床样本的蛋白质组与蛋白质修饰组分析，揭示某一疾病的生理、病理过程，分析导致某种生理现象或病理状态的机制，对患者进行精准分子分型和合理用药指导。还可以通过寻找异常表达蛋白或异常的蛋白质修饰作为潜在药物靶点，以及进行后续药物开发。

AI 蛋白质下游中还有一部分企业处于非医药领域，而是集中在食品、生物燃料、化工领域。蛋白质设计中很重要的一部分为酶工程，在食品、轻工业、医药等多个领域得到了重要应用。使用AI进行蛋白质设计或优化，可以获得性能更好的工业酶，2023年，合成生物公司 Arzeda 成功利用人工智能平台研发出一种酶，能够提高将甜菊提取物转化为高纯度且更昂贵的甜味剂的效率。

总体来看，AI+蛋白质将给产业链下游带来巨大价值，且呈现两种实现路径：一种是深度赋能生物医药管线研发，通过解决药物研发的重要问题来获取高附加值；另一种则以生物制造为典型，相关公司通过提升蛋白质工程化水平，以场景为出发点，以打造通用平台为最终目标。

AI+蛋白质技术分析



2023年以来，人工智能在蛋白质领域的研究取得了全方位的发展。AlphaFold 嵌入蛋白质开发流程，David Baker 团队重构蛋白质设计思路，大模型切入蛋白质组学，传统方法与人工智能结合以提高效率。

DeepMind 团队在预测模型中占据领先地位，2023年，AlphaFold 发布最新版本模型 AlphaFold-latest，将蛋白质预测的准确率提升到一个新的高度，AlphaFold 成为蛋白质开发进程中有力的验证工具。而 David Baker 更偏向应用，其团队的研究方向横跨制药、材料、酶催化等多领域，从蛋白质序列出发，重构蛋白质设计和药物开发的底层逻辑。技术进步引发蛋白质组学数据的爆发式增长，让大模型有了用武之地，面对复杂繁多的蛋白质组数据，AI 蛋白质组学在靶点发现和疾病诊断等方面的应用有巨大的潜力。还有很多公司使用基于物理和机器学习的方法进行蛋白质的预测和设计。

图 2-1 AI算法发展历程



来源：智药局

人工智能在蛋白质研究和应用领域正逐渐展现出其强大的潜力。在这一过程中，已有多家创新型企业崭露头角，值得一提的是，Generate:Biomedicines 的 Chroma 模型成功创造了自然界中前所未有的新型蛋白质，为蛋白质研究带来了突破性的进展。而百图生科的 HelixFold-Single 模型则能够在不依赖多序列对比的情况下，仅凭初级结构（氨基酸序列）预测原子的三维坐标，从而实现了对蛋白质结构的精准预测。

2.1 蛋白质结构和功能预测

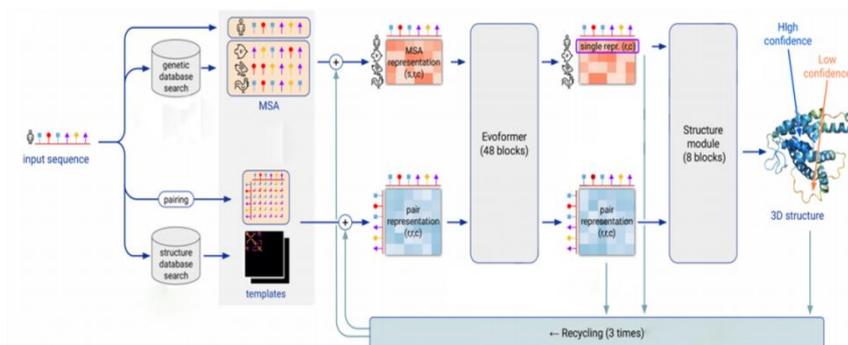
蛋白质的三维结构和功能在其生物学角色中具有决定性作用，深入理解氨基酸序列与蛋白质结构和功能之间的关系，对于推动生物学的理解和医学的应用具有重大意义。

2.1.1 人工智能在蛋白质结构预测方面的应用

蛋白质功能由结构主导，目前有效的蛋白质结构解析工具包括：X射线、核磁共振和冷冻电镜，但是这类方法往往伴随着高成本和低效率。当面对某个全新的、复杂的蛋白质时，人类可利用的工具是有限的。

人工智能打破了这种局面，2016年，许锦波等开创性地将深度残差网络（ResNet）架构成功地应用到结构预测领域中，显著提升了蛋白质残基接触预测，在这个工作基础上有大量结合共进化和深度学习的算法出现，其中代表性的工作如 AlphaFold（主要引入残基距离预测）和山东大学杨建益与 David Baker 团队开发的 trRosetta（主要引入了二面角信息等），均采用了 ResNet。随后，2020年，AlphaFold 2.0 在蛋白质结构预测竞赛 CASP14 中惊艳亮相，获得了 98.5% 的蛋白质预测率，AlphaFold 2.0 的结构框架更加复杂，核心模块是基于 Transformer。这三种方法都是使用多序列对比（MSA）作为模型的输入，这类方法虽然出现得早，但仍是目前主流方法之一。AlphaFold 2.0 之后，结构预测主要是基于 AlphaFold 2.0 的基础上做的改进。如图 2-2 所示，为 AlphaFold 2.0 模型的架构。

图 2-2 AlphaFold2.0 模型架构



来源：Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

2021年，David Baker团队在Science上公布了其开源蛋白质预测工具RoseTTAFold，其使用自然语言处理（NLP）的方法来直接从MSA中学习共进化信息，这种模型的结构预测精度接近CASP14中AlphaFold 2.0的精度。自此，基于蛋白质序列的预训练模型又称为蛋白质语言模型（PLM）的方法也被使用在蛋白质结构预测中。2022年底，科技公司Meta首次推出ESM-2和ESMFold，这是迄今为止发布的规模最大、最复杂的蛋白质语言模型之一。2024年，David Baker团队发布的新结构预测方法RoseTTAFold All-Atom（RFAA），可以生成一个生物单元的所有原子的三维坐标，包括蛋白质、核酸、小分子、金属和化学修饰。

近期，结合单序列蛋白质语言模型和结构预测模块的算法，在一些孤儿蛋白或者人工设计蛋白上展示了不错的结构预测性能，如HelixFold-single、OmegaFold等。这类方法的共同之处是使用了蛋白质语言模型的表征信息替换掉原始MSA的输入信息，并采用了类似AlphaFold 2.0的Evoformer模块和结构模块，在MSA质量较低的蛋白质上展示了更好的预测性能。

图 2-3 人工智能+蛋白质结构预测算法

序号	时间	模型	团队	架构	应用范围	缺点
1	2018年	AlphaFold	DeepMind	输入：多序列对比信息； 算法：使用深度神经网络，预测氨基酸链的性质，比如两两之间的距离分布、夹角分布等；	蛋白质结构预测	预测精度不够，无法预测孤儿蛋白结构
2	2019年	trRosetta	杨建益课题组与David Baker课题组	输入：多序列对比信息； 算法：使用深度残差神经网络，预测接触矩阵或距离矩阵，在得到该约束条件后再进行能量优化；	蛋白质结构预测	主要在残基间方向预测中精度更高，但处于初步阶段
3	2020年	AlphaFold2	DeepMind	输入：多序列对比信息； 算法：采用基于attention的Transformer架构，从网络中学习到特征，预测三维结构坐标；	蛋白质结构预测	在识别蛋白质-配体相互作用方面的预测能力较弱
4	2021年	RoseTTAFold	David Baker团队	输入：多序列对比信息； 算法：使用NLP的方法来直接从MSA中学习共进化信息，建立蛋白质三维结构模型；	蛋白质结构预测、蛋白质-蛋白质复合物	模型规模相比于AlphaFold2较小，缺少侧链坐标的推断
5	2023年	HelixFold-single	百图生科宋乐团队	输入：蛋白质一级序列； 算法：采用PLM作为MSA的替代，通过结合预先训练的蛋白质语言模型和AlphaFold2的基本模块得到最终端到端的可微模型；	孤儿蛋白或者人工设计蛋白结构预测	更适合预测MSA质量较低的蛋白质结构
6	2024年	RoseTTAFold All-Atom	David Baker团队	输入：蛋白质序列、小分子指纹信息等； 算法：通过对扩散去噪任务进行微调，直接在小分子和其他非蛋白质分子周围构建蛋白质结构，来生成结合口袋；	包括蛋白质、核酸、小分子、金属和化学修饰等单个生物单元的结构预测	在训练数据有限的领域可能遇到困难

来源：智药局

AlphaFold 2.0 端到端三维蛋白质结构预测技术的成功，无疑是该领域的一大里程碑。众多公司现已采纳并应用AlphaFold 2.0或同等精度的模型，如RoseTTAFold，进行结构验证工作。使用人工智能辅助的蛋白质结构预测已经成为蛋白质研究的基础工具之一。目前，研究焦点已转向孤儿蛋白的性能评估等领域（如OmegaFold和OpenFold），并尝试采用类似GPT-4的大型语言模型（例如ESMFold）的不同架构，以提升模型的泛化能力和预测效率。

展望未来，蛋白质结构预测领域的算法创新，尤其是底层技术的突破，可能会成为人工智能在蛋白质研究领域中的关键技术壁垒。这些创新将有助于推动蛋白质科学的进步，为疾病治疗、药物开发和生物技术等领域带来深远的影响。

2.1.2 人工智能在蛋白质功能预测方面的应用

准确标注蛋白质功能对于理解生物活动的本质、诊断疾病的病因以及加速新药的开发至关重要。尽管研究者在这一领域取得了显著进展，但仍有大量蛋白质的功能未知。目前，UniProt 数据库（收录蛋白质序列）中不到 1% 的蛋白质有功能注释，而这些注释大多通过昂贵且耗时的生物实验获得，人工智能蛋白质功能预测技术能够消除已知序列的蛋白质数量与已知功能的蛋白质数量之间的部分差距。

人工智能蛋白质功能预测是基于氨基酸序列、二级和三级结构信息、蛋白质-蛋白质相互作用（PPI）网络等不同类型的生物数据以及结合文献信息的深度学习方法。

直接从氨基酸序列推断蛋白质功能是长久以来一直在研究的方向，早期，基于序列相似性的 BlastKNN 是一个广泛使用的基准方法。近年来，深度学习技术在从各种数据类型中提取特征方面显示出了强大的能力，DeepGO-CNN 是第一个使用深度学习的序列基模型，它通过堆叠的 CNN 层来提取特征并预测蛋白质的功能。大型语言模型也显示出显著潜力，DeepGO-SE 方法通过使用预训练的大型语言模型从蛋白质序列预测 GO（Gene Ontology）功能。

AlphaFold 2.0 等先进的算法的进展为基于结构信息预测蛋白质功能提供了可能性。例如 DeepFRI，这是一个基于图卷积网络（GCN）的模型，结合了蛋白质序列和结构信息来预测蛋白质功能；还有 GAT-GO，是一种基于结构的图注意力网络（GAT）方法来改善蛋白质功能预测效果。目前如何检测结构中的重要基序和相对应的特征，仍是提高蛋白质功能预测性能的关键。

在 PPI 网络中，节点代表蛋白质，而边代表它们之间的相互作用，反映了蛋白质参与的复杂生物过程，因此，PPI 信息在蛋白质功能预测中的位置同样重要。DeepGO 是首个将蛋白质序列和 PPI 网络集成的基于深度学习的预测模型。受 DeepGO 启发，DeepGOA 更高效、稳健地提取序列特征。图神经网络（GNNs）的成功为蛋白质功能预测提供了更多策略。

DeepGraphGO 是一个端到端模型，利用 GNNs 从 PPI 网络中提取信息以预测蛋白质功能。PPIs 与功能紧密相关，但这些方法也面临一些限制，如何将已知的 PPI 信息应用于新测序的生物体，是一个待解决的难题。

另外，在蛋白质功能预测领域中，基于序列和文献的预测方法是一种新兴的研究方向。这些方法试图通过结合蛋白质的序列信息和已发布的生物医学文献来预测蛋白质功能，克服了单一数据源可能遇到的限制。这种

方法面临的挑战包括处理文献描述的复杂性和从大规模数据中收集相关信息的困难。

图 2-4 人工智能+蛋白质功能预测算法

序号	时间	模型	发表期刊	架构	缺点
1	1980s	BlastKNN	/	输入：蛋白质序列； 算法：通过比较目标蛋白与已知功能的蛋白质之间的相似性，然后采用一种相似性加权算法来预测目标蛋白的功能；	早期机器学习算法，效率低
2	2018年	DeepGO	Bioinformatics	输入：蛋白质序列和PPI网络； 算法：基于卷积神经网络（CNN）的深度学习分类模型，使用3-mer编码蛋白质序列，取序列特征。对于PPI网络，采用DeepWalk生成每个蛋白质的256维网络拓扑特征；	层次化分类网络需要巨大的内存资源，难以应用于大规模标签
3	2020年	DeepGOCNN	Bioinformatics	输入：蛋白质序列； 算法：从蛋白质序列中提取特征以预测功能，通过堆叠的CNN层来提取特征并预测蛋白质的功能	最大预测蛋白质长度为2000
4	2020年	DeepGOA	IEEE/ACM Transactions on Computational Biology and Bioinformatics	输入：蛋白质序列和PPI网络； 算法：使用Word2vec生成序列的残基级嵌入，然后输入到Bi-LSTM和多尺度CNN层中以提取全局和局部特征；	具有特殊特征的蛋白质功能无法预测，蛋白质序列最长为1000
5	2021年	DeepGraphGO	Bioinformatics	输入：蛋白质序列和PPI网络； 算法：端到端模型，利用GNNs从PPI网络中提取信息以预测蛋白质功能；	缺乏PPIs信息的新测序生物体难以预测
6	2021年	DeepFRI	Nature Communications	输入：蛋白质序列和结构； 算法：融合了自监督语言模型和图卷积网络，从蛋白质序列的自监督模型中提炼出的序列特征及蛋白质结构来预测其功能；	蛋白质序列训练集局限于PDB
7	2022年	GAT-GO	Briefings in Bioinformatics	输入：蛋白质序列和结构； 算法：基于图注意力网络（GAT）的方法，它利用预训练的蛋白质序列语言模型。GAT-GO输入蛋白质序列，提取序列特征、残基级特征和结构特征，从而预测功能；	对于长序列蛋白质，对蛋白质结构数据精度要求很高
8	2024年	DeepGO-SE	Nature Machine Intelligence	输入：蛋白质序列； 算法：使用预训练的大型语言模型从蛋白质序列预测GO功能，通过生成多个近似GO模型，并用神经网络预测蛋白质功能的真值；	当序列与PPIs结合时性能最佳，许多新蛋白质没有已知的相互作用限制了组合模型的应用

来源：智药局

人工智能在进行蛋白质功能预测时，依赖于现有的数据集和已知的生物学信息，这种依赖性也暴露了AI在功能预测上的局限性。未来，蛋白质预测领域的重大突破可能源于人工智能算法框架的创新或生物数据获取技术的革新。这些突破有望为蛋白质功能预测带来革命性的进步。

2.2 蛋白质与其他生物分子作用预测

蛋白质与其他生物分子的相互作用是生命活动的核心，而传统的实验方法在预测这些相互作用时存在局限。人工智能算法，如基于搜索的分子对接算法和基于几何神经网络的深度学习算法，为预测这些相互作用提供了新途径。

蛋白质与其他生物分子相互作用的过程是生命活动的基本组成部分，精确预测蛋白质分子表面可能的结合位点的位置对于很多科学及应用问题都是很有帮助的，比如，药物靶标注释、药物设计、药物副作用预测等等。传统的实验方法虽能提供关于这些相互作用的数据，但通常耗时费力且易产生假阳性结果。

计算方法在预测这些相互作用方面的作用日益凸显，蛋白质与其他生物分子相互作用预测的算法框架，包括基于能量的蛋白质分子对接算法，以及基于几何神经网络等的人工智能算法。

人工智能分子对接算法可以分为基于搜索的分子对接算法，代表性的算法包括 Glide、Autodock 等；以及基于回归的深度学习分子对接算法，主要包括 EquiBind、TANKBind、DIFFDOCK 等。基于几何神经网络等的人工智能算法，包括从氨基酸序列、蛋白质结构以及蛋白质-蛋白质相互作用（PPI）网络中蕴含的信息中抓取特征，例如，HIGH-PPI 模型使用两个GNN, 从两个视角分别进行学习以预测 PPI。

预测复合物的结构是理解蛋白质如何作用的另一种方式，经典的蛋白质复合物结构预测算法多采用蛋白质分子对接法。蛋白质分子对接方法是基于已给定的结构来预测复合物结构。例如，EquiDock 模型，预测蛋白质-蛋白质复合物结构；以及整体思路与 DiffDock 接近，使用机器学习方法预测蛋白质-蛋白质复合物结构。另一方面，David Baker团队运用共进化分析，并结合 AlphaFold 2.0 和 RosettaFold，开发了 RoseTTAFold2NA 预测蛋白质复合物结构。这类方法比较依赖数据库以及数据的精确度。

随着单体蛋白质结构预测算法的进步，端到端的蛋白质复合物结构预测算法亦表现出卓越的性能。Deepmind 团队在 AlphaFold 2.0 的基础上，开源了 AlphaFold-Multimer，专门用于端到端的复合物结构预测。同时，深势科技也开发出了可训练版本的 UniFold-Multimer，其性能与 AlphaFold-Multimer 相当。其中，最具代表性的就是 DeepMind 团队 AlphaFold-latest，它不仅成功地预测了许多蛋白质的结构，在蛋白质与蛋白质、核酸、小分子等复合物的结构预测中的准确度最高。

图 2-5 人工智能+蛋白质与其他生物分子相互作用的算法

序号	时间	模型	团队	架构	应用范围
1	2018年	DNN-PPI	天津大学团队	DNN-PPI基于深度神经网络，包含卷积神经网络（CNN）和长短期记忆（LSTM）两个独立的顺序层，从蛋白质序列中学习的特征以自动预测PPI。	预测蛋白质-蛋白质相互作用
3	2021年	AlphaFold-Multimer	DeepMind	基于深度学习，先建立复合物的多序列比对，以推断进化关系，然后用与AlphaFold2基本相同的深度学习方法来预测三级结构。	预测蛋白质复合物三维结构
2	2021年	EquiDock	麻省理工学院	EquiDock模型，假设蛋白质内的构象在结合过程中没有发生构象变化的情况下，在单个未结合的结构中计算预测蛋白质-蛋白质复合物的3D结构的模型。	预测蛋白质-蛋白质复合物结构
6	2022年	EquiBind	麻省理工学院	以EquiDock为基础，依赖SE(3)-等变图神经网络，以配体分子图的随机三维构象和受体结构作为输入，预测结合蛋白质配体构象。	预测蛋白质-小分子复合物结构
4	2022年	DeepTrio	浙江大学团队	DeepTrio是一种使用掩码多个并行卷积神经网络进行蛋白质-蛋白质相互作用的一个深度学习框架，它允许模型学习相对性质并研究每个残基对预测结果的贡献，并通过热图可视化蛋白质的重要性图。	预测蛋白质-蛋白质相互作用
5	2022年	UniFold-Multimer	深势公司团队	端对端的蛋白质复合物结构预测，输入蛋白质多聚体的一级结构，预测蛋白质的三级结构，同时给出预测结果的置信度。	预测蛋白质复合物结构
7	2022年	TANKBind	星药科技联合复旦大学、中山大学共同研发	基于图神经网络模型，用三维结合构象表示分子之间的几何关系，从而提高预测的准确性。	预测蛋白质-小分子配体结构和亲和力
8	2022年	DIFFDOCK	麻省理工学院	基于深度学习的分析对接模型，以配体和靶标的结构信息作为输入，之后对配体进行了一定的构象转换（平移，旋转，扭转）来生成新的配体构象，最后则是对这些生成的配体进行一个合理性的评分以及排名。	预测蛋白质-小分子结合结构
9	2023年	DIFFDOCK-PP	麻省理工学院	算法整体思路与DiffDock接近，使用机器学习方法学习将未结合的蛋白质结构翻译和旋转为其结合构象，对评分模型生成的不同姿势，进行排序并选择最佳姿势。	预测蛋白质-蛋白质结合结构
10	2023年	HIGH-PPI	腾讯 AI Lab 联合香港科技大学、中国科学院大学相关团队	基于层次图学习技术的深度学习，描述蛋白质之间的相互作用，每种蛋白质就是一个节点，蛋白质之间的相互作用就是图的边，关键氨基酸或残基组合就是图的节点，物理位置相邻的残基以边相连。	预测蛋白质-蛋白质相互作用
11	2023年	RoseTTAFold2NA	David Baker团队	端到端的深度学习方法，使用与RoseTTAFold相同的数据集进行训练，并增加了所有RNA、蛋白质-RNA和蛋白质-DNA复合物的结构信息。	预测核酸结构和蛋白质-核酸复合物结构
12	2023年	AlphaFold-latest	DeepMind	最新版本的AlphaFold。	对包括蛋白质、核酸、小分子、离子和修饰残基在内的复合物进行联合结构预测

来源：智药局

在生物分子相互作用的研究领域，端对端预测方法的引入标志着一场技术革新，它极大地扩展了人类对于蛋白质与其他生物分子相互作用方式的认识。这一技术革新在实际应用中展现出巨大的潜力，将对药物发现方式产生颠覆性的影响。2024年，由DeepMind拆分出的公司Isomorphic Labs与诺华和礼来的巨额合作订单，充分验证了人工智能技术在生物制药领域的应用价值。

2.3 蛋白质设计

人工智能技术在蛋白质设计领域的发展，特别是通过优化和从头设计方法，能够帮助科学家们创造出从前无法制造的功能性蛋白质、疫苗和其他类型的分子。

对蛋白质的精准预测，可以更好地理解蛋白的结构和功能，实现蛋白质设计，从而应用于药物研发、合成生物等领域。

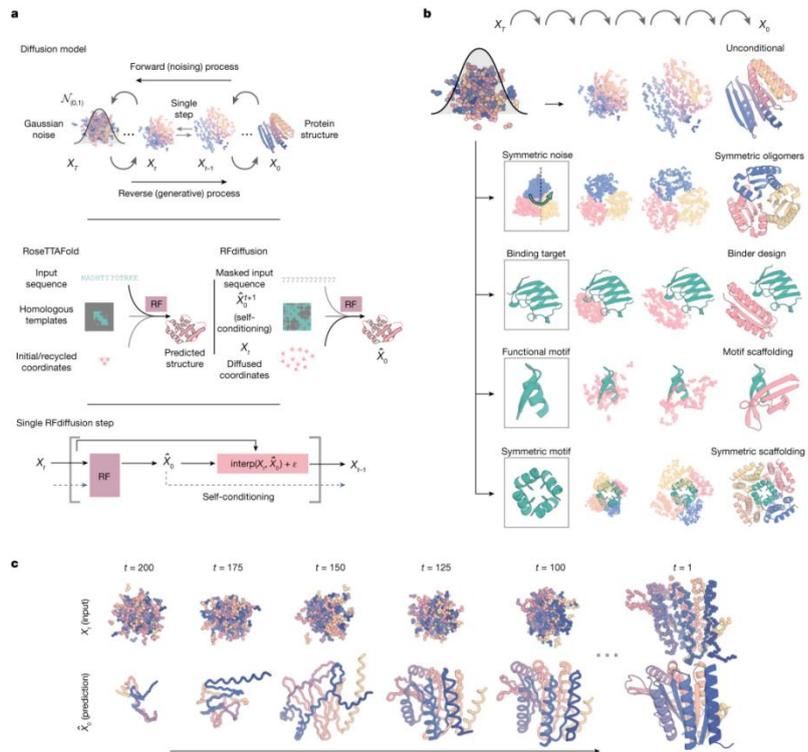
早期的蛋白质设计使用 PCR 技术，依托研究人员的经验在基因特定位点引入突变，从而改变蛋白质对应位置的氨基酸残基种类，进行蛋白质改造；之后的计算设计是基于结构模拟与能量计算来进行蛋白质设计，可以构建自然界不存在的蛋白质；近年来，随着结构生物学、计算生物学及人工智能技术的迅猛发展，人工智能技术成为指导蛋白质设计的新思路。

人工智能蛋白质设计可以分为两大方向：蛋白质优化与蛋白质从头设计。蛋白质优化是通过修改一个自然序列（定向进化），提升其特定功能，如亲和力的增强、催化活性的提高以及稳定性的增强等；蛋白质从头设计，包括从结构设计序列、从功能设计序列和从功能设计结构，在这一领域，David Baker 教授及其团队的研究成果尤为突出。

从结构生成序列的过程又被称为蛋白质逆折叠，2019年，David Baker 团队发布了 trRosetta，能够寻找自然折叠和新结构的稳定序列，从而实现蛋白质设计。“逆折叠”问题的另一个解决方法是训练模型可以明确地预测给定一个结构的序列，例如，ProteinMPNN 是一个基于图的神经网络。

从功能到结构的蛋白质设计，以蛋白质功能为起点，目前GANs、扩散模型和蛋白质语言模型为主流，例如，RFdiffusion 将扩散模型与预先训练的蛋白质结构预测模型（RoseTTAfold）的结合，能够用于各种蛋白质设计应用，并得到了广泛的实验验证。如图所示，为 RFDiffusion 蛋白质设计实例流程图。

图 2-6 RFDiffusion 蛋白质设计实例流程



来源: Watson, J.L., Juergens, D., Bennett, N.R. et al. De novo design of protein structure and function with RFDiffusion. Nature (2023)

从功能到序列的蛋白质设计，只需要识别可以执行该功能的氨基酸序列。蛋白质语言模型是有效探索蛋白质序列空间的良好候选模型。例如，ProGen 给定一系列输入的蛋白质属性，能够从头开始迭代地生成蛋白质序列。

图 2-7 人工智能+蛋白质设计算法

序号	时间	模型	团队	架构	应用范围
1	2019年	trRosetta	David Baker 团队	基于结构神经网络的蛋白质序列设计方法。	蛋白质逆折叠
2	2022年	ProteinMPNN	David Baker 团队	基于图神经网络的蛋白质序列设计方法。	用于设计单体蛋白质、对称重复结构、纳米颗粒和目标蛋白质结合物。
3	2023年	RFDiffusion	David Baker 团队	基于扩散模型，结合预先训练的蛋白质结构预测模型RoseTTAfold的蛋白质设计方法。	用于稳定单体设计、对称低聚物设计、功能基序或酶活性位点的支架设计，以及蛋白质结合物设计。
4	2023年	ProGen	Salesforce Research、Tierra Biosciences和加州大学的研究团	以无监督学习的方式在一个大型多样的蛋白质序列数据库中进行训练，学习通用的蛋白质表示。训练后，ProGen 可以根据提示从头生成蛋白质序列	从功能生成蛋白质序列

来源：智药局

当前，AI 算法在从头生成蛋白质中取得了显著进展，特别是在抗体和多肽的设计与合成方面。基于深度学习的扩散模型和变分自动编码器等技术，以及类似 GPT-4 这样的蛋白质语言模型，已经在处理蛋白质序列等生物数据方面展现出卓越的性能。这些工具不仅提高了蛋白质设计的精确性和速度，而且通过自动化和智能化的方法，正在逐步替代传统的计算和实验方法。

这种转变不仅缩短了蛋白质从概念到实验室原型的时间，还提高了设计的精确性和可预测性，使得定制化蛋白质解决方案成为可能，为生物制药、生物材料和合成生物学等领域带来了根本性的变革。AI 技术在蛋白质科学领域的应用，正在开启一个创新和应用的新纪元。

2.4 蛋白质组学

人工智能与蛋白质组学的结合，为靶点发现、生物标志物识别和精准医疗提供了更加高效、精准并具有创新性的数据分析手段。

高通量质谱技术的出现推动了蛋白质组学的发展，带来了海量的数据。人工智能技术的进步和高性能计算能力的提升使得大规模数据的处理和分析已经成为可能。蛋白质组学与人工智能的结合，在靶点发现、生物标志物的发现以及精准医疗领域的应用有很大的潜力。

在靶点发现方面，人工智能通过对多组学数据的深度挖掘和分析，能够迅速识别出与疾病发生发展密切相关的关键靶点。例如，Insilico Medicine（英矽智能）用于靶点发现的 PandaOmics 平台，通过 AI 对患者基因组学和蛋白质组学等医学多组学数据的挖掘分析来实现新靶点的发现。

在生物标志物的发现方面，通过对蛋白质表达谱的精准分析，结合机器学习和模式识别技术，研究人员能够筛选出与特定疾病高度相关的生物标志物。例如，2024年，复旦大学研究团队采用大规模蛋白质组学数据和人工智能算法，发现了4种可预测未来痴呆风险的血浆蛋白质，可提前15年预测痴呆发病风险。

在精准医疗领域，通过对患者个体的蛋白质组学数据进行全面分析，结合人工智能的预测模型，医生能够为患者量身定制出最佳的治疗方案。Matthias Mann 团队发布的 AlphaPept 软件，可用于 DDA 蛋白质质谱数据分析，在大规模蛋白质组学研究中具有优越性能和灵活性。另外，人工智能技术中的 NLP 还能够理解和分析生物医学文献中的大量文本信息，提取出与蛋白质组学相关的关键信息，为科研人员提供有价值的参考，这种跨学科的融合将不断拓展应用场景。

图 2-8 人工智能+多组学算法

序号	时间	应用	团队	架构	应用范围
1	2020年	PandaOmics平台	Insilico Medicine (英矽智能)	通过AI对患者基因组学和蛋白质组学等医学多组学数据的挖掘分析,对正常组织和疾病组织比较分析,结合数据来整合靶点信息。	靶点发现
2	2022年	AI赋能的蛋白质组学平台	西湖大学郭天南团队及西湖欧米团队	基于人工智能神经网络结合蛋白质组特征构建了一个可以用于甲状腺结节良恶性评估的深度学习模型	疾病诊断及精准医疗
3	2024年	人工智能+生物标志物	复旦大学研究团队	使用一种人工智能算法模型,结合痴呆症相关生物标志物与其他数据,成功预测了包括阿尔茨海默病在内的三种常见痴呆症的发病,准确率约为90%	生物标志物的发现及疾病预防

来源:智药局

人工智能与蛋白质组学的结合,可以类比于多模态大模型在处理和复杂数据集方面的强大能力。人工智能在蛋白质组学中的应用将能够整合基因组学、转录组学、蛋白质组学和代谢组学等多个层面的数据。这种跨层次的数据整合将有助于构建更为全面的生物系统模型,还能通过分析个体的蛋白质组数据,为精准医疗和个性化治疗提供支持。

3

AI+蛋白质应用场景

AI 对蛋白质的开发目前集中在蛋白质本身结构、功能的预测和蛋白质与其他生物分子相互作用的预测，以及蛋白质设计和蛋白质组学分析。在产业应用中，人工智能+蛋白质设计的占比最高，预测技术更多地嵌入进现有的药物开发和蛋白质工程流程中，而蛋白质组学分析的应用处于初期阶段。

人工智能+蛋白质的应用场景非常广泛，已成为药物研发、精准医疗、合成生物等领域中均有应用。以 David Baker 团队为例，其参与孵化的 AI 蛋白质公司，基于深度学习开发的一系列工具，在药物研发、多肽设计、小分子结合蛋白设计、新型材料开发等重要生物问题取得了出色的表现，体现出人工智能在蛋白质中的应用价值。

3.1 新药研发

人工智能技术通过精确预测蛋白质结构，模拟生物分子互动，辅助药物筛选和设计，结合蛋白质组学数据分析，加速药物发现和靶点库扩充。

蛋白质既是部分药物的构成成分，也是多数药物发挥效力的核心靶点。利用人工智能技术对蛋白质结构进行精确预测，有助于深化对药物作用机制的理解，从而指导候选药物分子的筛选工作。通过模拟蛋白质与其他生物分子的互动过程，可以精准识别药物与靶点之间的结合部位，为药物研发提供有力支持。此外，基于人工智能的端到端蛋白质设计技术，可以实现从零开始设计具有特定功能的蛋白质类药物。同时，将人工智能与蛋白质组学数据相结合，能够系统地分析人体内蛋白质与疾病、药物之间的关联，补充靶点库，加快药物发现的进程。

3.1.1 蛋白质药物研发

蛋白质类药物包括抗体药物、多肽药物、环肽药物和迷你蛋白等。人工智能+蛋白质药物研发的企业比例占所有AI蛋白质企业的比例高达44%，包括AbCellera、BigHat Biosciences、Generate:Biomedicines、华深智药、呈元科技等企业。

人工智能在从头设计抗体药物方面表现出色，2023年，Nature 期刊发表了一篇利用蛋白质表面特征指纹图谱的机器学习方法，从头设计新的蛋白质，这些蛋白质与癌症免疫治疗靶标（PD-1、PD-L1、CTLA-4）或新冠病毒靶标（S蛋白）的结合亲和力与自然产生的抗体相当。针对没有固定构象的多肽和无序蛋白，深度学习工具也取得了出色的结果，比如 David Baker 实验室设计了能够结合与阿尔茨海默病有关的淀粉样蛋白形成肽的结合蛋白，实现了对淀粉样蛋白Abeta形成的完全抑制。

除了抗体药物和多肽药物，还有一些不常见但很有潜力的蛋白质类药物，比如环肽药物和迷你蛋白。环状多肽在合成时并不受自然界氨基酸的限制，可以使用非天然氨基酸，加了非天然氨基酸之后，可以实现原来无法实现的功能。2024年1月，默克与 Unnatural Products (UNP) 签署了价值2.2亿美元的 Biobucks 协议，进一步进军大环肽领域。迷你蛋白具有性质稳定、方便运输等特性，并且分子量小，可以实现精准地调控靶点蛋白的生理功能的效果。2022年11月，利用 AI 从头设计合成迷你蛋白质的生物技术公司 AI Proteins 宣布完成1820万美元的种子轮融资。

从商业模式来看，AI蛋白质公司与生物制药公司通过或合作或被收购的方式建立蛋白质药物发现的AI能力。例如：BigHat（专注于抗体发现）与安进、AbCellera（专注于抗体发现）与艾伯维以及Generate:Biomedicines（专注于蛋白质工程）与安进的合作。以及2021年，基因泰克（Genentech）收购了 Prescient Design（这是一家AI驱动的生物技术公司，将机器学习应用于抗体发现）。

图 3-1 部分蛋白质药物公司情况

序号	成立时间	企业名称	平台	应用场景	最新融资轮次	商业模式	合作企业	管线情况/成果
1	2011年	ABSci	Integrated Drug Creation™ 平台	抗体药物研发——从头生成、先导化合物优化、细胞因子生物	2021年7月上市	自研管线/合作管线	诺华、梯瓦、维森特、IGM 公司	数条药物管线
2	2012年	AbCellera	抗体药物开发平台	抗体药物研发	2020年纳斯达克上市	自研管线	阿斯利康、Almirall、艾伯维	数余条管线，ABCL575和ABCL635两条目前正在IND研究
3	2012年	LabGenius	基于机器学习的蛋白质药物发现&EVA™ 蛋白质工程平台	抗体药物研发（癌症、炎症）——优化抗体	2020年A轮	合作	赛诺菲	LabGenius 与赛诺菲合作优化 NANOBODY® 蛋白，获得积极成果
4	2013年	瀚科迈博	AI-计算机辅助抗体设计平台	抗体药物研发（肿瘤靶向治疗和肿瘤免疫治疗）	2023年A+轮	自研/合作管线	安科生物	8条管线，未进入临床
5	2014年	BenchSci	知识图谱连接的 ASCEND 平台	SaaS——重组蛋白、RNAi、动物模型、基因编辑技术等试剂的选择	2023年D轮	合作	赛诺菲、诺和诺德、艾伯维、武田制药、GSK、诺华、强森、百济神州、默克	/
6	2017年	Juvena Therapeutics	JuvNET平台和专有蛋白质库	分泌蛋白设计	2022年A轮	自研管线	/	数余条管线，涉及多种肌病和代谢类疾病以及未披露的管线
7	2018年	Generate Biomedicines	Chroma生成式人工智能模型	蛋白质从头生成	2023年C轮	自研管线/合作管线	安进、美国安德森癌症中心	十余条管线，包括免疫学、传染性和肿瘤学等领域
8	2018年	Nabla Bio	蛋白质语言建模的人工智能蛋白质设计平台 AbMap™ 平台	抗体药物设计-从头生成	2021年种子轮	自研管线	/	未披露
9	2019年	BigHat Biosciences	抗体药物开发平台	抗体药物研发——优化抗体、从头生成	2022年B轮	自研管线/合作管线	艾伯维、安进、默克	5条管线以及与安进和默克合作的未披露的管线
10	2021年	Seismic Therapeutic	IMPACT™ 平台	蛋白质从头设计和优化	2022年A轮	自研管线	/	针对泛IgG蛋白酶和PD-1的2条管线，均处于IND授权开发阶段
11	2021年	华深智药	Helixon Design抗体药物发现平台；OmegaFold蛋白质预测	抗体药物研发	2022年A轮	自研管线	/	多条用于肿瘤和自身免疫性疾病的大分子药物管线
12	2022年	赛得康生物	AbZed蛋白模块技术平台	多特异性重组蛋白	2023年种子轮	自研/合作管线	晶泰科技	3条管线，未进入临床
13	2022年	呈元科技	生成式AI设计平台 SynCore	针对PPI靶点的环肽药物设计	2023年PreA轮	自研管线	罗氏	10条管线，未进入临床

来源：智药局

人工智能在蛋白质药物研发领域的应用正迅速扩展，且已逐步进入临床试验阶段。目前人工智能开发的蛋白质类药物中，最快的已经进入了临床2期，为 ZielBio 公司开发的用于实体瘤的抗凝集素单克隆抗体。此外，还有2款药物处于临床1期，包括 Peptilogics 公司开发的用于治疗假体周围感染的肽类抗生素，以及 SparX 公司开发的用于治疗胃癌的靶向 Claudin 18.2 的单克隆抗体。

2024年3月，David Baker 团队取得了新的突破，他们首次利用生成式人工智能来设计并制造全新的抗体。这一成果被 Nature 誉为“具有里程碑意义的时刻”，充分证明了 AI 工具在药物发现环节的应用价值。

然而，AI 工具的大规模应用并非一片坦途，目前仍面临一系列挑战，部分原因在于，药物研发需要满足极高的安全性标准，任何一款新药都必须经过严格的测试和长期的临床试验，而这一验证周期通常耗时数年，涉及大量的资源投入和复杂的监管审批流程。在这种背景下，分散风险和缩短交付周期成为实现AI工具大规模应

用的关键策略。

3.1.2 靶点发现

小分子药物用于治疗疾病已有数百年，但仍有许多与蛋白质靶点的交互未知。靶点是药物发挥作用的关键，因此靶点发现和确认是药物开发的重要步骤。AI可以通过分析大量的生物信息学数据来识别和验证新的靶点，这对于发现新的治疗途径和开发创新药物至关重要。

首先，多组学分析的药物靶点预测是一种基于多组学数据训练机器学习/深度学习模型进行靶点发现。通过整合基因组学、转录组学、蛋白质组学等多维度信息，对药物靶点进行预测。其次，大语言模型下的药物-靶点关系获取则是一种基于自然语言处理技术的方法。这种方法利用大语言模型对生物医药领域的知识进行建模，通过语义分析和信息抽取等技术，从各种蛋白质数据库、药品数据库等生物医药信息中获取药物-靶点关系，还可以对未知的药物-靶点关系进行预测和推断。

靶点发现作为药物开发流程的重要环节，通常并非孤立存在。目前，人工智能技术在癌症、神经性疾病以及感染性疾病等领域的新靶点发现方面展现出显著成果，部分基于人工智能发现的靶点已顺利进入临床试验阶段。2024年1月，PIKfyve 抑制剂 VRG50635 正式启动临床1b期概念验证研究，PIKfyve 是基于 Verge 公司的 AI 平台在患病人体组织中发现的 ALS（渐冻症）治疗靶点。这一研究为 ALS 的治疗提供了新的可能。

图 3-2 部分进行靶点发现的平台情况

序号	发布时间	技术/平台	来源	技术框架	应用场景	成果
1	2021年	Advanced Covalent Exploration (ACE) 平台	Matchpoint Therapeutics	基于机器学习算法，指导靶点的优先排序和特有骨架预测；识别致病蛋白的新型共价结合物并提供整个蛋白质组的选择性评估；	组学数据理解靶点和疾病关系，以及可成药蛋白质分子发现	研发管线最初将聚焦于免疫学领域，在赛诺菲和Vertex的支持下成立的，旨在开发创新共价小分子药物；
2	2022年	PandaOmics3.0	Insilico Medicine (英矽智能)	使用AI分析患者基因组学、转录组学和蛋白质组学等多组学数据；	潜在靶点识别和老药新用策略的开发	与剑桥大学合作发现阿尔茨海默病等多种疾病治疗新靶点；使用转录组数据，提名51个与衰老相关的潜在靶点；发现MYT1作为乳腺癌等癌症治疗靶点的潜力；等等
3	2023年	PandaOmics4.0—ChatPandaGPT	Insilico Medicine (英矽智能)	在PandaOmics中增加了基于Transformer的知识图谱功能；	使用大型语言模型实现问答功能，发现最佳潜在药物靶点	
4	2023年	ConPlex	麻省理工学院	基于大语言模型的AI算法，将目标蛋白质与潜在药物分子相匹配；	可应用于小分子药物和治疗性抗体的筛选	扩大靶点库；新药物发现
5	2023年	大模型 xTrimo、AI 蛋白质生成平台 AIGP	百图生科	将大规模蛋白质语言模型 (PLM) 与 AlphaFold2 相结合	新靶点以及新药物的发现	与启愈生物合作发现新靶点和新药物；与启德医药合作推进ADC 药物开发；与索智生物合作开展IBD协同双靶点研究；

来源：智药局

人工智能蛋白质组学研究在靶点发现中的应用正展现出强大的发展势头，AI 的介入能够揭示新的生物标记和治疗途径。然而，这一领域的发展同样面临着挑战，包括对高质量、大规模生物样本数据的需求，以及算法的准确性和可解释性问题。因此，未来的研究应更加注重数据质量和算法的改进，以提高 AI 在蛋白质辅助靶点发现中的准确性和可靠性。

3.1.3 难成药靶点

人类疾病相关的蛋白靶点中约 85% 为难成药的靶点，这是因为一些蛋白难以解析和观察到，即使使用传统的电镜、核磁共振等方法也只能观察到它们的静态结构。AI 技术为解决难成药靶点的药物研发提供了新的工具和方法，有望改变传统药物研发的模式，提高药物发现的效率和成功率。

AI 技术，尤其是深度学习算法，可以用于预测蛋白质的三维结构，这对于理解蛋白质的功能和寻找药物作用的潜在位点至关重要。例如，AI 技术还可以用于开发药物设计平台，如 ENSEM 的 Kinetic Ensemble[®] 平台，该平台结合计算和人工智能深度学习方法与先进的实验技术，发现难成药靶点的隐匿性结合位点，从而加速基于结构的药物设计。

除了对靶点的识别和改造，AI 还在新治疗方法的研发中发挥着重要作用。例如，针对非底物结合位点的变构药物研发，2022年12月，Nimbus Therapeutics 通过 AI 和计算辅助药物发现，其中一款针对口服选择性变构酪氨酸激酶2（TYK2）的变构小分子抑制剂以60亿美元授权给武田制药，其中首付款部分高达40亿美元，变构机制药物已展示出巨大的市场价值。

目前，多数专注于难成药靶点的公司聚焦于单个或多个靶点的结构研究。基于此，他们开发出包括小分子药物、变构药物、分子胶、PROTAC 药物等在内的创新药物。这些药物的市场竞争相对较小，但由于缺乏同类型成功药物的参照，因此需承担创新的风险。

图3-3 部分专注于难成药靶点的公司情况

序号	成立时间	企业名称	平台	技术/应用场景	最新融资轮次	商业模式	合作企业	管线情况/成果
1	2013年	宇道生物	变构药物研发平台ALLOSTAR™	难成药靶点-变构药物	2022年A+轮	自研/合作管线	青煜医药	5条管线, 未进入临床
2	2017年	Redesign Science	综合物理、化学、实验数据以及深度学习及主动学习的NUVO™平台	难成药靶点——蛋白质相互作用预测、蛋白质-DNA和RNA结合预测、变构生物学	2021年未披露	自研管线	/	8条管线, 主要针对肿瘤学和免疫学
3	2018年	DeepCure	PocketExpander™平台(口袋扩展器)、MolGen™平台(分子设计)	难成药靶点——蛋白质-蛋白质相互作用预测	2021年A轮	自研管线	BioSero(合成实验室)	8条管线, 主要针对炎症和免疫学
4	2019年	Genesis Therapeutics	人工智能平台GEMS	难成药靶点——蛋白质-配体结构和相互作用预测	2023年B轮	合作CRO	礼来、基因泰克、斯坦福大学	未披露
5	2019年	Ventus Therapeutics	ReSOLVE™平台	难成药靶点	2020年A轮	自研管线/合作管线	诺和诺德	十余条自研与合作管线, 针对CGAS、NLRP3以及其他未披露靶点
6	2020年	德睿智药	一站式AI药物设计平台Molecule Pro™; 蛋白质动态模拟、结构预测平台Molecule Dance™; 基于知识图谱的数据挖掘平台PharmKG™	难成药靶点——蛋白质动态模拟、结构预测	2023年A+轮	自研/合作管线、CRO	德琪医药、皓元医药、复宏汉森、美迪西	10余条管线, 其中2条进入临床阶段
7	2021年	CHARM Therapeutics	DragonFold平台	难成药靶点——蛋白质-配体共折叠预测, 绕过亲和力预测	2023年未披露	合作	/	主要针对肿瘤学
8	2021年	Congruence Therapeutics	Revenir™平台、蛋白生物物理特性的专有数据库	针对蛋白质折叠错误的靶点设计小分子药物	2023年A轮	自研管线	/	5条管线, 包括癌症、帕金森、肥胖等疾病
9	2021年	Gandeeva Therapeutics	SPOTLIGHT™平台(靶标选择)、HYPERFOCUS™平台(冷冻电镜)、CRYO-CADD™平台(药物发现)	难成药靶点——蛋白质相互作用预测、蛋白质-药物预测	2022年A轮	自研管线	Moderna(冷冻电镜研究蛋白质表面结构合作)	4条管线, 主要针对肿瘤学, 1条进入IND申请
10	2021年	ENSEM	Kinetic Ensemble®平台、EnsemNET™平台、EnsemGEN™平台	难成药靶点——蛋白质口袋预测、蛋白质-配体相互作用预测	2022年A轮	自研管线/合作管线	百济神州	3条管线, 其中1条授权给百济神州, 主要针对肿瘤学
11	2021年	Isomorphic Labs	新一代AlphaFold与人工智能药物开发平台	难成药靶点-蛋白质-蛋白质相互作用、蛋白质-小分子相互作用等研究	/	CRO	诺华、礼来	2024年与礼来合作获4500万美元的预付款; 与诺华合作获3750万美元的预付款
12	2021年	予路乾行	多尺度分子动力学+人工智能药物研发平台Divamics	难成药靶点——蛋白质三维结构、蛋白质-配体复合物结构、PPI结合界面等从头建模和分析优化	2023年未知轮	CRO	立迪生物、甫康药业、星浩澎博	已与多家国内外生物医药企业合作, 达成近20个项目里程碑

来源: 智药局

3.2 精准医疗

人体细胞内的蛋白质是生命活动的关键，异常蛋白质可能由基因变异或RNA不稳定引起。蛋白质组学揭示细胞生理活动，结合AI技术，有助于理解细胞机制和推动新药研发、精准医学。

人体细胞中包含着数千种蛋白质，这些蛋白质主导着人体的运行。想要了解某个细胞中是否存在异常蛋白质，可以通过基因测序检测致病蛋白是否因为基因变异产生。但这只是部分可能，更多的变异蛋白质是由转录过程中RNA的不稳定导致的，所以只依赖基因组学信息是不够的，直接检测细胞中蛋白质组成可以帮助我们了解细胞内的各种重要生理活动。

与基因组相比，蛋白质组学数据是复杂和动态的，在不同的健康状态下，人体内的蛋白质每时每刻都在发生变化。通过将AI技术与蛋白质组数据相结合，不仅能够深入理解细胞功能的复杂机制，更能够加速新药研发、精准医学的实现。

3.2.1 生物标志物

在医学领域，生物标志物通常用于疾病诊断、病情监测和预后评估。下一代测序（NGS）揭示了大量与癌症相关的基因位点，同时，得益于高通量蛋白质组学技术所产生的庞大数据集，人工智能技术得以在蛋白质组学数据分析领域发挥巨大作用，从而发现与疾病或生理状态相关的生物标志物。此外，在精准医疗领域，将蛋白质生物标志物与基因型数据相结合，可以帮助深入了解疾病过程，进行癌前筛查和癌症诊断，为患者提供个性化的治疗建议。

目前，全球前20家大型药企都在使用蛋白质组学发现新生物标记物。美国早筛巨头 Exact Sciences 收购了蛋白质组学公司 OmicEra；Illumina 也宣布进军蛋白质组学；新晋诺贝尔化学奖得主 Carolyn Bertozzi 创立的 InterVenn，2021年已经推出全球首个基于AI和蛋白质组学的卵巢癌诊断产品。基于人工智能蛋白质组学的研究处于初期阶段，2022年，西湖大学推出人工智能算法与高通量蛋白质组学结合的用于甲状腺结节良恶性评估模型，对甲状腺结节的分类准确率分别高达89%和85%；2024年，Nature 子刊发表使用多组学数据开发和验证衰老生物标志物。

图 3-4 部分人工智能生物标志物发现的公司情况

序号	成立时间	企业名称	技术	应用场景	平台	成果
1	2018年	InterVenn	使用机器学习模型，破译糖蛋白组学数据中错综复杂的模式和关系	癌症相关生物标志物的发现和识别	GlycoVision™平台	AI质谱蛋白组技术，实现卵巢癌和直肠癌的诊断
2	2020年	西湖欧米	AI 深度学习与蛋白质组学数据分析	临床样本蛋白质组研究、癌症相关生物标志物的发现和识别	AI赋能的蛋白质组学平台；西湖欧米AIDD平台 iTalks Series	基于万例血清蛋白组预测代谢综合征风险、人工智能结合蛋白质组学辅助甲状腺结节良恶性判别等论文
3	2021年	珞米生命	基于AI和纳米技术，实现高通量、大规模、和非靶向地富集人类血浆样本中的低丰度蛋白质	生物标记物的发现和筛选、液体活检新方法的建立、诊断试剂盒的开发等	Proteonano™人类血浆蛋白质组学NGP平台；Kepler Pro 干湿结合平台	运用Proteonano™，在有限的样本的情况下已可以发现61个差异表达蛋白；

来源：智药局

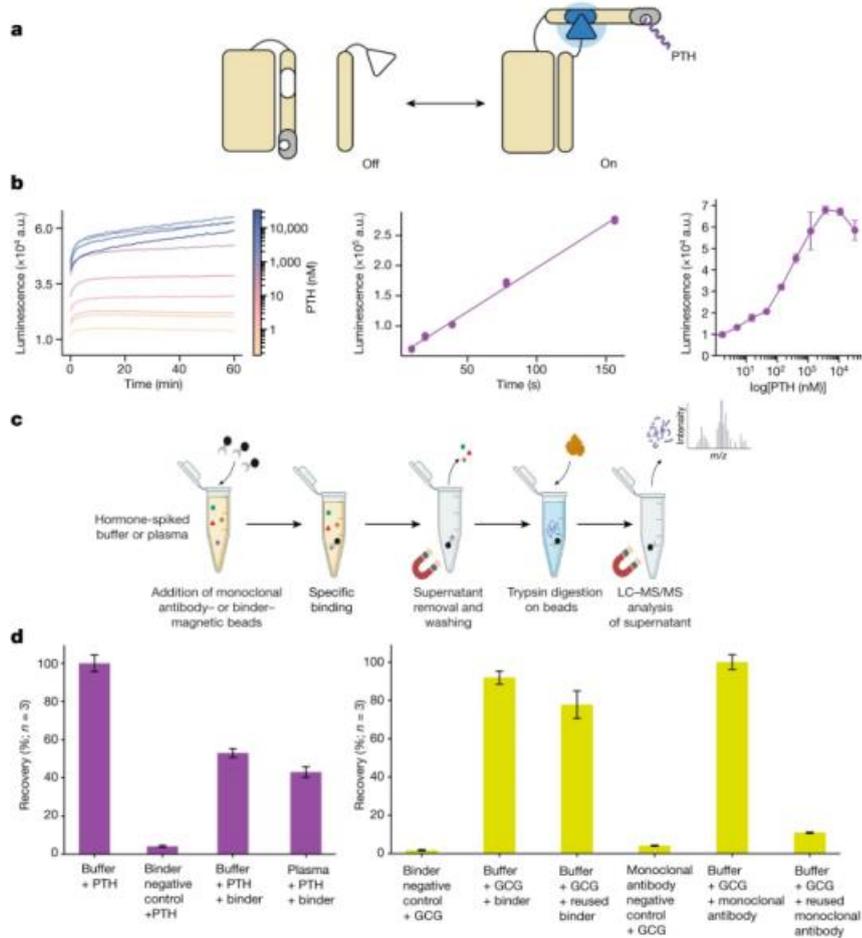
AI 蛋白质组学研究正逐渐成为精准医疗和体外诊断行业的创新引擎。企业在这一领域的商业模式正从单一的技术或设备提供商，转向综合性健康解决方案供应商，涵盖了从数据分析平台到个性化医疗咨询的全方位服务。

3.2.2 生物传感器

生物传感器是一种利用生物活性物质作为催化剂的传感器，它们被广泛应用于体外诊断中，用于检测各种生物标志物。AI 蛋白质可以辅助优化生物传感器中的生物活性材料，发现更合适的生物识别元件，提高体外诊断的效率和准确度。

2023年末，David Baker 在 Nature 期刊发表的论文，从头设计出更高结合力的结合蛋白，该研究从头设计的高亲和力和特异性的蛋白质，使得甲状旁腺激素和胰高血糖素的富集和后续检测可以通过质谱进行，并构建了基于生物发光的蛋白质生物传感器。这些从头设计的结合蛋白，可以与各种具有挑战性的生物标志物结合，还能够用于开发更便宜的抗体替代品，对于药物开发、疾病检测和环境监测等领域具有重要意义。

图 3-5 从头设计的结合蛋白在传感和检测中的应用



来源：Vázquez Torres, S., Leung, P.J.Y., Venkatesh, P. et al. De novo design of high-affinity binders of bioactive helical peptides. *Nature* 626, 435–442 (2024)

在生物科学、信息科学和材料科学的共同推动下，生物传感器技术飞速发展，其在临床诊断、食品安全、环境保护、运动健康等多领域展现出了广阔的应用前景，而这项技术与人工智能算法的结合催生了“智能生物传感”这一新理念，不仅能够识别和量化特定的生物分子或生物活动，还能够通过集成的智能系统进行数据处理、存储和决策支持，具备生物计算、记忆存储、自动校准、模块化和标准化等特点。

3.3 合成生物

合成生物学通过改造、优化和设计蛋白质，下游应用涵盖医药、农业、食品、化工等多个领域，在具体商业化上，赛道内玩家们呈现出两种不同的路线。

在合成生物领域，通过精确控制和改造蛋白质的结构和功能，为各行各业提供了创新的解决方案，具有巨大的应用潜力和经济价值。合成生物学领域中蛋白质的应用非常广泛，涉及多个重要领域，包括但不限于以下几个方面：

(1) AI 技术可设计和改造酶类蛋白质，提高农业、食品、美容等行业的生产效率。合成生物公司 Arzeda 在酶改造领域取得显著成果，他们利用先进的酶改造平台，研发出高效转化甜味剂的酶，降低生产成本。此外，Arzeda 还为英国石油公司研发出了新型酶，用于提高石油开采和生产过程中的效率。值得一提的是，2023年，David Baker 团队相继发表了两篇关于人工智能设计功能酶的文章，这些研究不仅展示了该方法的强大和通用性，还为未来酶改造领域的发展提供了有力支持。

(2) 酶和原料药在药物研发中具有重要价值。人工智能技术通过设计和优化微生物代谢途径和酶活性，高效合成复杂有机分子，作为药物合成的中间原料，提高药物生产效率和降低成本。还可改造细胞因子的结构和功能，创造新型细胞因子药物，具有更强活性、稳定性和低副作用，适用于肿瘤、炎症和自身免疫性疾病等领域。

(3) 蛋白质也是许多新型材料的灵感来源。自然界已经存在的一些生物材料已经展现出了比传统材料更好的性质，比如蜘蛛丝的单位抗拉强度比钢都要强。首个获批药物 COVID-19 疫苗就使用了 David Baker 团队的通过计算设计的蛋白质纳米粒子技术。除了医疗领域，AI 可以帮助设计和优化生物材料和能源材料，例如，生物可降解塑料、生物膜，以及燃料电池和太阳能电池的催化剂和吸附剂。

(4) 在食品领域，AI 可帮助开发更高营养和更低成本的优质蛋白质。在农业领域，AI 可改进作物蛋白质，提高产量和质量，减少环境污染。2021年，NotCo 公司推出了人工智能平台 Giuseppe，使用 AI 设计和预测植物基蛋白质的结构，以模仿动物性食品中的营养成分和口感，为植物基食品的开发提供优化建议。2023年，中国农科院蔬菜花卉研究研究员团队联合其他学者基于人工智能蛋白设计策略改造马铃薯块茎，显著增加了马铃薯面团的粘度和营养，提高其在食品加工领域中的应用。

(5) 通过对蛋白质的结构和功能进行预测和优化，可以研发出更具针对性和高效性的生物降解剂，用于解决环境中的有害物质和废弃物处理问题。利用人工智能蛋白质可以开发能够降解塑料和其他污染物的酶类蛋白质，有助于解决环境污染问题，还可以开发出能够检测环境污染物的生物传感器。此外，人工智能技术还可以应用于生物修复等领域，通过设计具有特定功能的蛋白质，以恢复和改善受损的生态系统。

图 3-6 部分AI+合成生物公司情况

序号	成立时间	企业名称	平台	应用场景	最新融资轮次	商业模式	合作企业	管线情况/成果
1	2008年	Arzeda	Intelligent Protein Design Technology™ 平台	农业、能源	2022年B轮	合作	Amyris、联合利华、Zymergen	提高将甜菊提取物转化为甜味剂效率的酶，预计在2023年四季度正式上市
2	2020年	Enzymit	深度学习酶设计平台	酶的从头设计和优化——生长因子、重组蛋白等	2022年种子轮	合作	Aleph Farms	与Aleph Farms合作开发生长因子替代品，已用于细胞培养肉获批上市；
3	2020年	引加生物	核心蛋白技术平台	核心蛋白原料的开发——细胞治疗、mRNA疫苗、新药研发、分子诊断和肿瘤新药伴随诊	2023年A轮	合作	上海交大免疫学研究所、上药康希诺、智峪生科、泉心泉意	/
4	2021年	Cradle Bio	Cradle生成式人工智能模型	蛋白质从头设计——疫苗、多肽和抗体、工业酶等	2023年A轮	合作	强生、grifols、novozymes、Arkeon、Enzymit	与超过12个研究和开发项目合作
5	2021年	Monod Bio	LucCage生物传感器平台和LuxSit从头设计光素酶平台	生物传感器——医疗诊断、食品污染、环境检测	2022年种子轮	合作	/	使用AI算法——Family-wide Hallucination创造出全新的酶，可以有效催化荧光素底物发光
6	2021年	智峪生科	ZCloud 全生态平台	医疗健康、生物制药、食品保健、美容化妆品和农业畜牧业等	2023年A轮	合作	维亚生物、引加生物、海正药业	药物中间体、抗菌药物、功能食品添加剂、抗氧化剂等数条生产线
7	2021年	力文所	Lésign®蛋白质设计平台	工业环境、制药、食品添加剂、消费品	2023年天使轮	合作	恺休生物	/
8	2022年	Profluent Bio	ProGen模型	蛋白质从头设计——制药、工业酶等	2023年种子轮	合作	/	/
9	2022年	粒影生物	人工智能辅助蛋白质设计平台	抗体、重组蛋白等原料蛋白；酶、功能蛋白和IVD原料的改造	2023年Pre A轮	合作	英国癌症研究所、弗朗西斯·克里克研究所	成为弗朗西斯·克里克研究所供应商
10	2022年	分子之心	AI大分子优化与设计平台MoleculeOS	抗体药物、多肽药物、小蛋白药物、新材料、新能源等	2023年战略	合作	凯赛生物	/

来源：智药局

在合成生物学领域，人工智能的应用主要体现在两种不同的商业模式上。

第一种是平台型公司，例如 Generate:Biomedicines，它们致力于打造一个能够根据客户需求定制蛋白质的生物制造平台。这类公司的优势在于能够迅速响应市场需求，人工智能蛋白质设计在合成生物中的应用，一定程度上绕过了生物制药在伦理和安全性方面的复杂挑战，使得人工智能的应用更容易被公众接受。这些公司通常在技术创新和成本效益方面保持领先，充当着行业的推动者和验证者，不断优化和更新其人工智能算法。

第二种是专注于特定应用领域的公司，它们通过在特定垂直市场中实现闭环验证，确立了自身的竞争优势。这些公司通过积累丰富的数据资源、客户基础和实践经验，建立了强大的行业壁垒。虽然在初期可能需要较大的投资，但这些公司通过深耕特定领域，能够建立起客户的信任，并逐渐形成稳定的市场地位。

这两类公司都在以各自的方式推动合成生物学的进步，通过人工智能技术的创新应用，为未来的生物科技发展提供了多样化的解决方案和坚实的基础。



AI+蛋白质投融资及发展预测

4.1 投融资情况

技术进步和应用落地的加速推进，

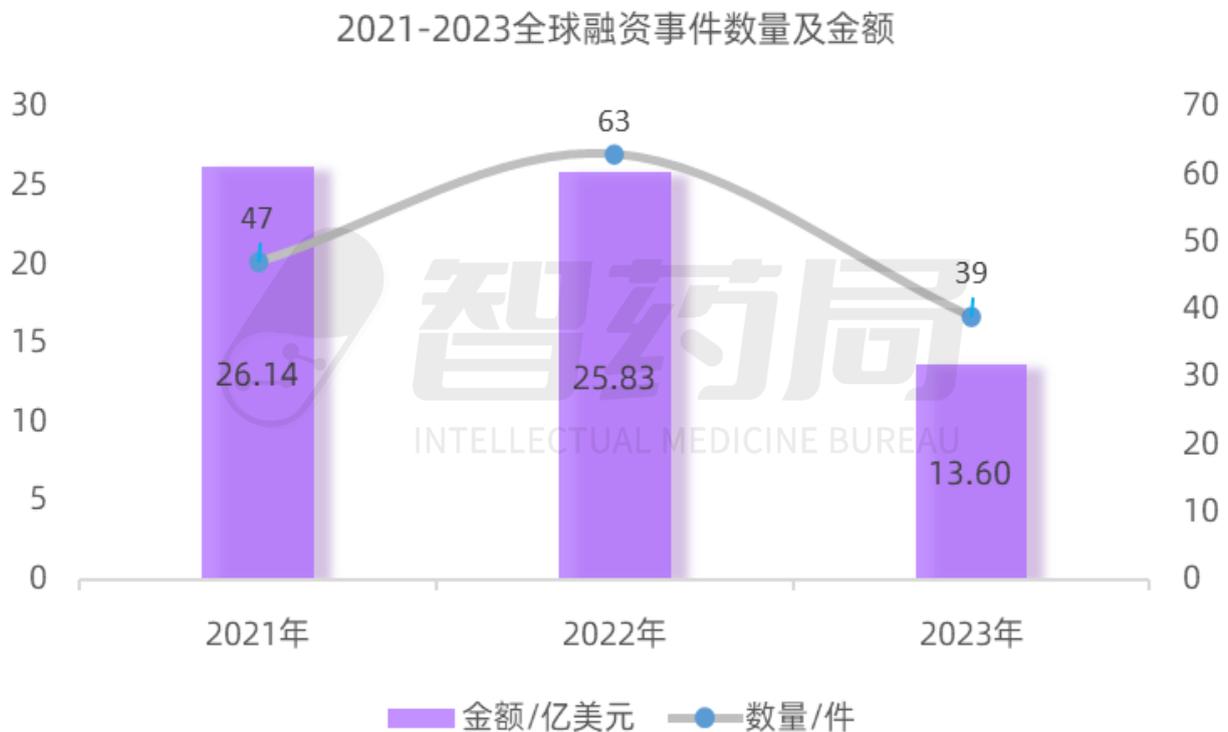
AI 蛋白质行业成为近年风口。

以 AlphaFold 为代表的科学突破，让市场看到了人工智能在生命科学领域应用的未来商业价值。

近三年（2021-2023）全球融资活动共149起，主要分布在美国、中国、加拿大、以色列、英国、德国、日本和瑞士。美国的金融市场和科技创新环境相对成熟，为企业提供了稳定的融资渠道和商业支持；而中国在生物技术和人工智能等领域的发展取得显著进展，出台多项利好政策，成为投资者重点关注的市场。

在全球视角下，近三年完成融资的企业创立时间集中在2015-2022年间，正好契合底层技术迭代的里程碑时点。

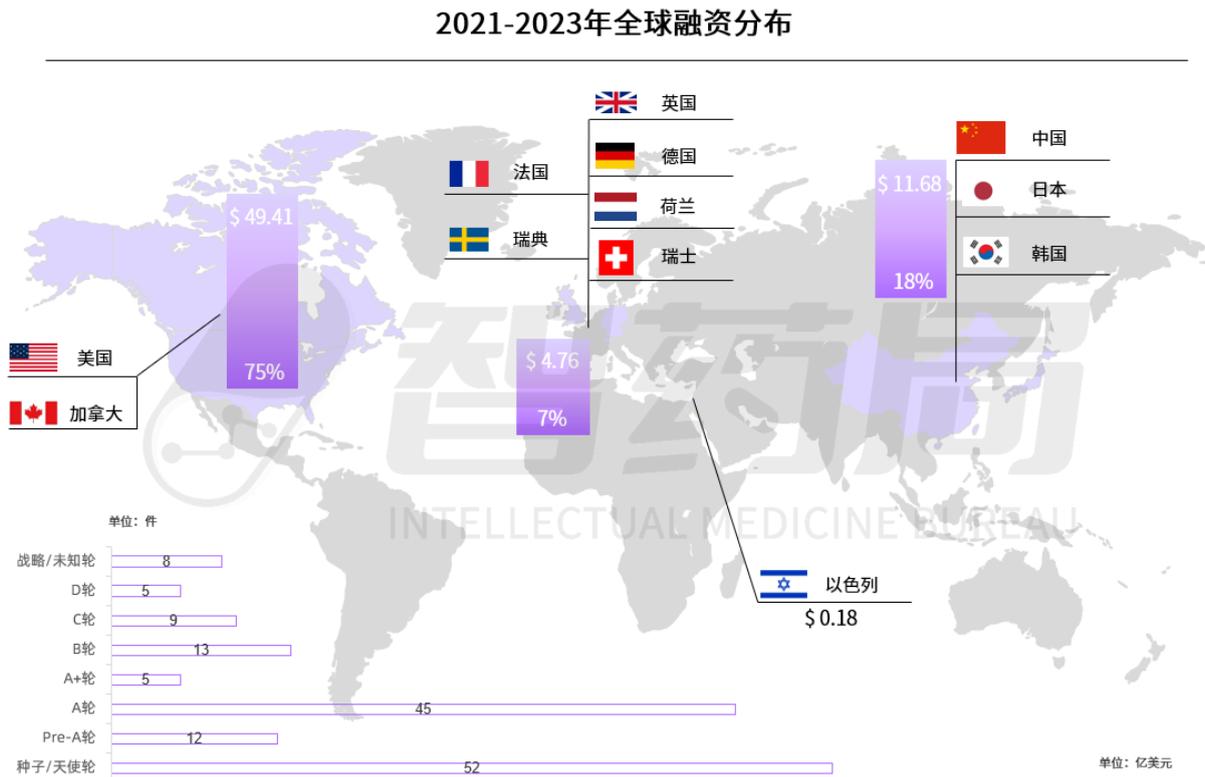
»» 图 4-1 2021-2023年全球融资情况



来源：智药局

全球 AI 蛋白质领域的融资活动在2021年达到 47 起，披露的融资总额为 26.14 亿美元。这一时期，超过 90% 的企业处于 A 轮或更早的融资阶段，行业整体进入快速创新阶段。2022年全球投融资活动数量迎来峰值，全年融资事件增至 63 起，披露金额为 25.83 亿美元。2023年，市场投资节奏有所放缓，投资者的态度变得更加审慎，全球融资事件减少至 39 起，披露金额同比下降 47.35%，仅为 13.6 亿美元。

图 4-2 近三年全球融资分布图



来源: 智药局

另一方面，全球融资金额分布同样呈现出显著的地域差异。根据2021至2023年的数据，北美地区在AI蛋白质融资总额达到了49.41亿美元，占据了全球融资的主导地位；其次是亚洲地区，其融资总额为11.68亿美元，以极高的活跃度位列第二；欧洲地区在这一领域的融资总额为4.76亿美元；中东地区以以色列为主。

融资轮次方面，在过去的三年中，早期融资活动（包括种子轮、天使轮以及A轮阶段）共发生了114起，占据了总融资事件的绝大多数，比例高达76.5%。我们认为，市场仍处于行业早期，未来将不断有新的资金涌入。

图 4-3 近三年过亿美元融资事件汇总

近三年过亿美元融资事件					
序号	公司	成立时间	地区	轮次	披露金额
1	Generate:Biomedicines	2018	美国	B轮	3.7亿美元
2	Generate:Biomedicines	2018	美国	C轮	2.73 亿美元
3	英矽智能	2014	香港	C轮	2.55亿美元
4	Exscientia	2012	英国	D轮	2.25亿美元
5	Immunai	2018	美国	B轮	2.15亿美元
6	InterVenn Biosciences	2017	美国	C轮	2.01亿美元
7	FogPharma	2015	美国	D轮	1.78亿美元
8	Ventus	2019	美国	C轮	1.4亿美元
9	ABSci	2011	美国	未披露	1.25亿美元
10	Cellarity	2017	美国	B轮	1.23亿美元
11	Cellarity	2017	美国	C轮	1.21亿美元
12	Seismic	2020	美国	B轮	1.21亿美元
13	Seismic	2020	美国	A轮	1.01 亿美元
14	百图生科	2020	北京	A轮	上亿美元
15	Exscientia	2012	英国	C轮	1亿美元

来源：智药局

近三年，全球共发生了15起过亿美元融资事件，美国地区以超过70%的占比，拥有领先地位，披露总融资金额约15.17亿美元。在这一全球科技浪潮中，中国区域的科技创新企业同样具有竞争力，英矽智能和百图生科以技术优势成功获得了大额资金支持。

目前，投资者开始更加关注企业的商业模式可行性、市场前景以及技术的商业应用潜力，这种转变导致融资事件数量和融资金额有所减少。在这样的背景下，科技创新和成功的商业化落地依然是企业和投资者关注的焦点。

4.2 市场规模

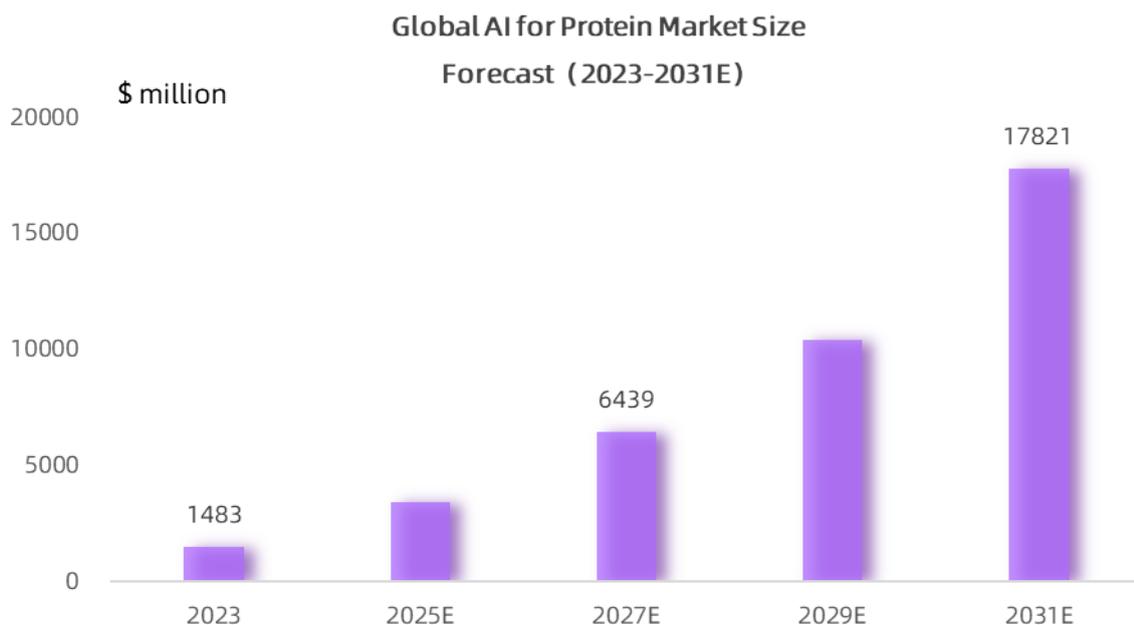
预计到2031年，市场规模将从2023年的14.83亿美元增长至178亿美元，年复合增长率约36.5%。

人工智能与蛋白质结合的应用正迎来快速发展，特别是在药物发现、精准医疗和合成生物等领域。根据 MedMarket Insights 的测算，2023年AI蛋白质市场规模已达14.83亿美元，得益于 AI 大模型与生命科学的高适配性，预计到2031年市场规模将增长至178亿美元，年复合增长率约为36.5%。

其中，合成生物学或将成为最具潜力的下游领域，预计到2031年市场规模将达到90亿-110亿美元。具体应用中，食品、农业以及快消品行业将占据超过三分之一的市场份额；医药领域，特别是生物药与药物中间体的生产，将占据约四分之一的比重。化工品、生物能源、降解材料拥有低成本+可持续的巨大优势，有望在2031年之后实现大规模合成，进而占据超过三分之一的市场份额。未来，AI蛋白质技术有望以酶工程作为切入口，拓展行业内更广泛的应用场景。

其次是药物研发领域，2031年市场规模预计在35亿-45亿美元之间，AI辅助的蛋白质药物开发部分项目已进入临床2期，预计2027年左右将迎来发展高峰。此外，精准医疗市场规模约为20亿-30亿美元，得益于AI大模型处理多维复杂数据的能力，算力的提升将进一步推动人工智能与蛋白质结合在应用层面实现指数级增长。

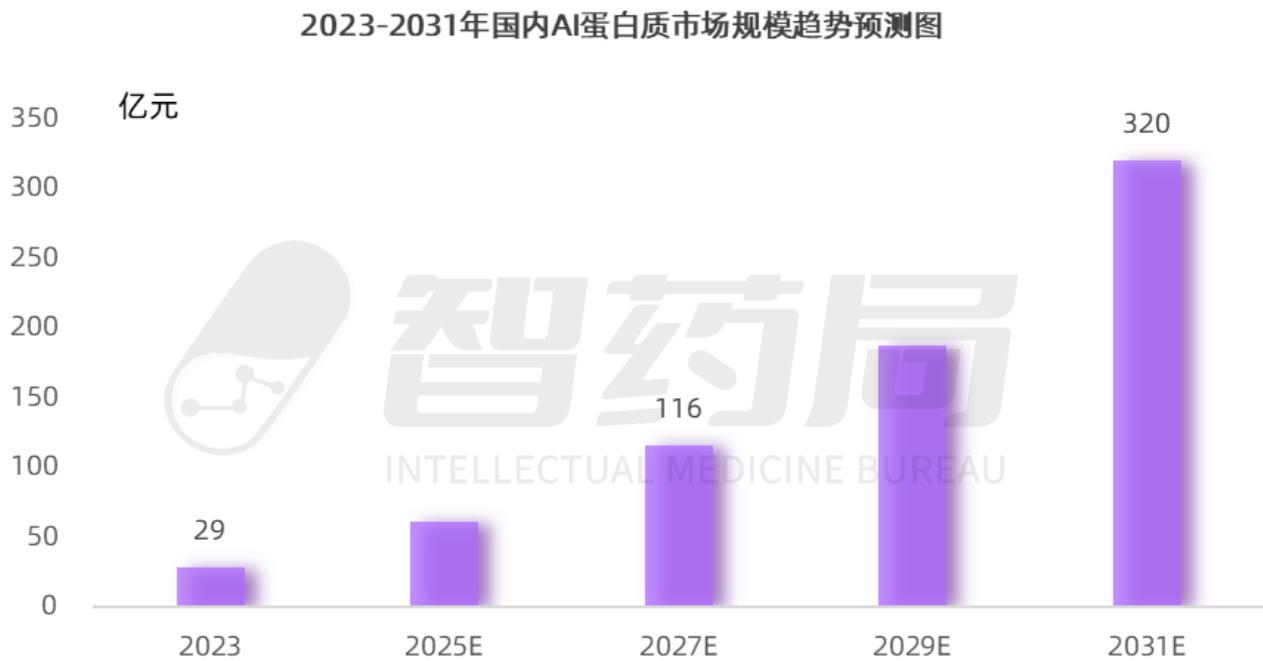
图 4-4 2023—2031年全球AI蛋白质市场规模预测



Source: MedMarket Insights, 2023

中国在AI蛋白质领域发展迅速，市场规模也在持续扩大。据智药局预测，2023年，中国AI蛋白质行业的市场规模将达到29亿元，并预计到2031年将增长至320亿元。这一增长趋势得益于中国成熟的工业制造体系和活跃的投融资市场。同时，AI蛋白质技术正逐步改变传统药物研发的模式，中国创新药领域有望在全球药物研发市场占据更加重要的地位。

图 4-5 2023—2031年中国AI蛋白质市场规模预测

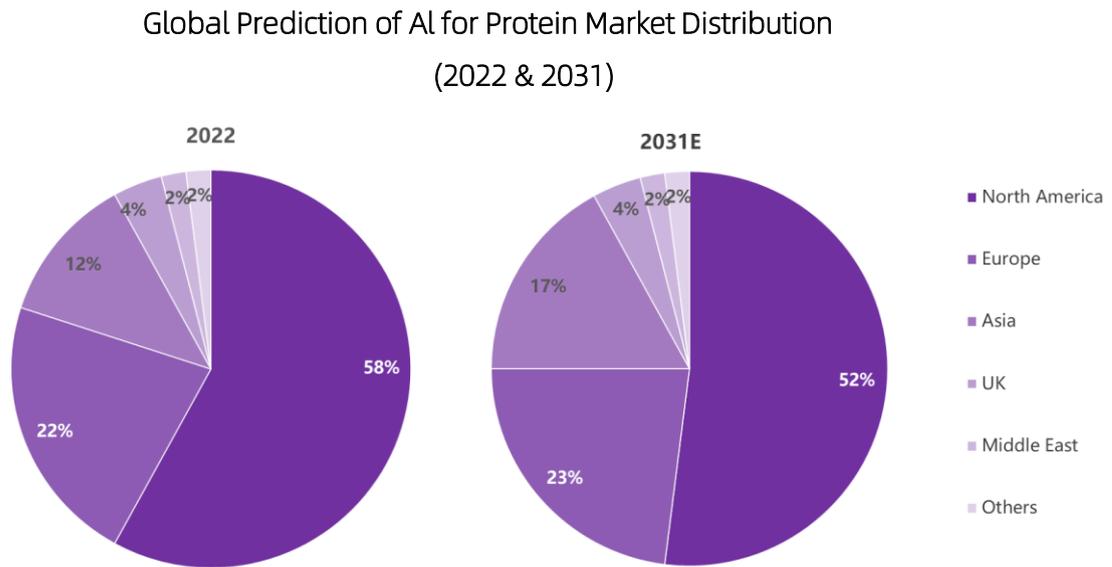


来源：智药局

美国目前在 AI 蛋白质研究和应用方面处于全球领先地位，占据约58%的市场份额。欧洲占22%，以 Cradle Bio 等合成生物企业为代表，与大型制药企业合作密切。亚洲（主要为中国和部分日本企业）市场份额约12%，中国新兴的蛋白质研究企业增长迅速，但与美国仍有差距。英国和中东地区分别占4%和2%，其他地区占2%。

AI 蛋白质市场正经历全球性的竞争格局变化，尽管美国在 AI 蛋白质领域具有先发优势，但亚洲尤其是中国的发展势头强劲，预计到2031年，将在市场中占据更大份额。

图 4-6 2023—2031年中国AI蛋白质市场规模预测



Source: MedMarket Insights, 2023

4.3 未来趋势展望

人工智能引领生命科学进入研究新范式

人工智能技术的突破性进展正在重塑生命科学的研究范式。特别是在蛋白质研究领域，深度学习和机器学习算法的应用使得科学家能够以空前的速度和准确度预测蛋白质的三维结构，并进行功能解析。这些技术能够处理庞大的生物信息数据集，揭示蛋白质间的相互作用及其作为生物标志物的角色，为深入理解疾病机制提供了新的洞见。

2024年3月，David Baker 领导的团队利用生成式人工智能技术，成功设计出全新的抗体，标志着人工智能在药物设计领域的重大进展。随着算法的持续改进和计算能力的增强，人工智能将推动个性化医疗和精准治疗的前进步伐，预示着研究新时代的到来。

AI+蛋白质产业化进入从1到100的发展阶段

AI蛋白质技术的应用已经度过初步探索阶段，在生物医药、能源化工、食品科学等多个领域证明了其在降低成本、提高效率以及实现创新突破方面的重要价值。

随着生物数据的不断积累和人工智能技术的快速进步，AI在蛋白质研究领域的应用成果将日益丰富，相关的政策支持、资金投入和人才培养也将得到加强，推动产业链的完善和生态系统的构建。

可以预见的是，未来AI蛋白质技术将加速落地，推动产业的转型和升级。

生物医药是最具挑战性的下游领域

人工智能蛋白质技术正引领生物医药行业进入一个新时代，这一过程漫长但可期。

一方面，生物医药研发过程的复杂性、临床试验的多期验证、监管审批的严格性以及后期数据监测收集等因素导致一项新技术的应用验证需要较为漫长的周期，AI蛋白质要在该领域产生真正价值需要大量的时间和资源投入，整个过程也将面临众多可预期和不可预期的风险。

另一方面，AI蛋白质无疑将推进人类对新疗法的探索，推动传统小分子药物和抗体、疫苗为代表的生物制剂的开发，此外，AI在蛋白质组学大数据中的应用，使得个性化医疗成为可能，通过分析患者的个体信息，为患者提供定制化的治疗方案。

未来，生物医药行业正逐步从传统的实验驱动模式转变为数据和智能驱动的新模式，预示着未来在疾病治疗 and 健康管理方面将实现更高效、更精准的革新，更多元化的创新应用。

合成生物学将诞生未来十年最具潜力的应用场景

麦肯锡预计到2025年，合成生物学与生物制造的经济影响将达到1000亿美元；2030-2040年，合成生物学每年带来的经济影响将达到1.8万亿-3.6万亿美元。

但眼下，绝大部分人工生物系统的构建与优化仍然依赖于反复试错，缺乏理性设计的能力，难以实现定量可控，这极大制约了合成生物学的发展。

在此背景下，人工智能已展现出设计具有更高催化效率和特定底物特异性的蛋白质产品的潜力，为客户提供合成新物质的整体设计方案、利用人工智能与生物计算技术进行虚拟筛选并进行规模化量产，加快 DBTL（设计—构建—测试—学习）循环，从而显著提升行业迭代能力。此外，人工智能还可以帮助研究人员处理和分析大规模的蛋白质组学数据，用以预测生物系统的行为和反应。

在数据积累的基础上，结合近年来不断发展的机器学习及深度学习等技术，合成生物学有望逐步从“试错”转向“设计”，通过建立定量理论模型和人工智能数据集，实现对生命过程的“真理解”，并在此基础上进行理性设计和功能预测。

蛋白质设计领域将出现超级平台型公司

随着人工智能技术的发展，蛋白质设计领域将出现一批超级平台型公司。这类平台型公司凭借其强大的 AI 技术实力和丰富的生物信息学资源，将成为行业的领导者。

人工智能在生物领域体现出越来越强的通用性，突出表现为其在跨学科的应用能力，随着下游产业应用落地的进一步拓展，有望打通多领域不同诉求，构建统一的底层框架，这将催生超级平台型公司的诞生。

平台型公司提供的一站式服务平台，从蛋白质序列设计到结构预测，再到功能验证，极大地促进了生物技术的研发进程。它们的出现不仅改变了传统的研发模式，还推动了行业内的合作与创新。随着这些超级平台型公司的不断壮大，预计它们将在全球范围内引领蛋白质设计和生物医药行业的未来发展，重塑行业的商业模式，推动生物技术领域的创新和合作。

AI+蛋白质赋能多元化应用场景

人工智能蛋白质技术的应用正在迅速扩展到多个领域，包括生物医药、农业科技、环境监测等。在生物医药领域，AI技术的应用已经从药物发现延伸到精准医疗和基因编辑。在农业科技领域，AI 技术正在帮助科学家设计更耐旱、抗病的作物品种。在环境监测领域，AI 技术通过分析大量数据，有助于更好地理解 and 预测环境变化。

在未来几年，将涌现一批聚焦于细分领域业务的企业，结合自身禀赋，以推出具体方案来解决行业痛点。

AI+蛋白质技术整合到特定业务流程的速度加快，产品和服务的市场契合度有望持续提升。

未来十年：人工智能蛋白质技术催生生命科学大模型时代

展望未来十年，人工智能蛋白质技术预计将催生一系列生命科学大模型的诞生。这些模型将能够处理和析海量的生物数据，为疾病机理研究、药物筛选和基因编辑等领域带来重大突破。目前，产业界也在积极探索生物大模型可行的技术路径，希望在生命科学领域复现出大语言模型的“辉煌”。

人工智能蛋白质技术的发展和应用，预示着生命科学将进入一个全新的数据驱动时代。在这个时代中，数据和算法将成为科学研究的核心，推动我们对生命现象的理解达到前所未有的深度。这不仅将为医学和健康领域带来革命性的变化，还将为人类的生活质量和福祉提供更强有力的支持。

随着科技的持续发展和应用领域的不断拓宽，生命科学领域将在未来十年迎来充满机遇的全新发展阶段。

声明

本报告由智药局制作。报告中文字、数据受中国知识产权法等法律保护。除报告中引用的第三方数据及其他公开信息，报告所有权归深圳智药信息咨询有限公司所有。

智药局对报告中所引用的第三方数据及其他公开信息不承担任何责任或义务。在任何情况下，本报告仅供读者参考。

如果您对报告中的内容存在异议，可通过zhiyaoju3@163.com联系我们。
未经允许，不得对报告进行加工和改造。欢迎转载或引用。如有转载或引用请及时与我们联系，并注明出处【智药局】。

智药咨询旗下

【智药局】作为国内第一家聚焦AI+药物研发的内容平台，实时追踪AI+药物研发技术的发展及应用、行业机会、市场需求等，未来将携手各方讲好新药创智的故事，助力前沿创新科技在医药产业中的应用与落地。



【智药投】作为医药投融资领域的Top3内容平台，聚焦上市和Pre-IPO药企在A股、科创板、港股等资本市场的信息传递，致力于挖掘优质公司的内在价值，在一、二级市场具有广泛认知度和深厚影响力。



