DOI: 10.13376/j.cbls/20250120
文章编号：1004-0374(2025)10-1251-12

· 情报研究 ·

基于专利分析蛋白质结构预测领域发展态势

梁亚茹[1#]，王平洋[1#]，鲁璟哲[1#]，邵旭倩[1#]，王　帆[1#]，姜　鑫[1]，李怡心[1]，李丽媛[2*]
(1 国家知识产权局专利局医药生物发明审查部，北京 100088；2 首都医科大学附属北京世纪坛医院，北京 100038)

摘　要：随着人工智能技术的不断发展，蛋白质结构预测领域研究取得了重大突破，在药物研发、生物能源、生物材料等众多领域展现出巨大的应用潜力。为了更好地了解全球蛋白质结构预测领域的研究现状，科学推动蛋白质结构预测产业高质量发展，本文基于专利数据，对全球范围内蛋白质结构预测领域的申请态势、区域布局、主要申请人、技术演进、产业发展机遇与挑战等方面进行分析，并重点比较了 DeepMind、腾讯等全球主要研究机构的专利布局特点。结果表明，我国蛋白质结构预测领域的专利数量排在世界前列，但核心底层算法、高质量数据和跨学科高端人才与美国相比仍存在差距；新药研发是蛋白质结构预测技术布局应用研发的重点和热点，在合成生物学等其他领域仍有大量空白待填。针对这些问题，提出我国蛋白质结构预测领域的产业发展策略。

关键词：蛋白质结构预测；人工智能；专利分析
中图分类号：Q5-3；Q51　　**文献标志码：**A

Research progress of protein structure prediction based on patent analysis

LIANG Ya-Ru[1#], WANG Ping-Yang[1#], LU Jing-Zhe[1#], SHAO Xu-Qian[1#],
WANG Fan[1#], JIANG Xin[1], LI Yi-Xin[1], LI Li-Yuan[2*]
(1 Department of Pharmaceutical and Biological Invention Examination, Patent Office of the National Intellectual Property Administration of China, Beijing 100088, China; 2 Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, China)

Abstract: With the continuous development of artificial intelligence technology, significant breakthroughs have been made in the research of protein structure prediction, demonstrating great application potential in many fields such as drug development, bioenergy, and biomaterials. In order to better understand the current research status of protein structure prediction globally and scientifically promote the high-quality development of the protein structure prediction industry, this paper, based on patent data, analyzes the application trends, regional distribution, major applicants, technological evolution, and opportunities and challenges in the field of protein structure prediction worldwide. It also focuses on comparing the patent layout characteristics of major global research institutions such as DeepMind and Tencent. The results show that the number of patents in the field of protein structure prediction in China ranks among the top in the world. However, there are still gaps compared with the United States in terms of core underlying algorithms, high-quality data, and interdisciplinary high-end talents. New drug development is the focus and hotspot of the application and R&D layout of protein structure prediction technology, and there are still a large number of gaps to be filled in other fields such as synthetic biology. In light of these findings, this paper proposes industrial development strategies for protein structure prediction in China.

Key words: protein structure prediction; artificial intelligence; patent analysis

收稿日期：2025-02-23；**修回日期：**2025-04-24
基金项目：国家自然科学基金面上项目(81974503)；北京市医院管理中心2023年度(第八批)青苗计划(QML20230702)
#共同第一作者
*通信作者：李丽媛，E-mail: liliyuan89@126.com

随着医疗改革的持续深化以及生物技术和信息技术的融合创新，我国生物医药步入大发展阶段，国家相继出台多项政策推动人工智能与生物医药深度融合，以加速产业高质量发展。2021 年 3 月发布的《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》首次将“人工智能 + 生物医药”纳入国家战略科技力量布局；同年 12 月，九部委联合印发的《医药工业发展规划》则聚焦技术落地，明确要求在药物发现、临床前研究等关键环节实现 AI 技术深度整合。在国家重点研发计划支持下，国内已建成 20 余个 AI 药物研发共性技术平台，带动蛋白质结构预测领域专利申请量大幅攀升，逐步形成从基础算法到产业应用的完整布局。

蛋白质结构预测是指通过计算机模拟和算法分析预测蛋白质的三维结构，其利用海量生物数据，挖掘蛋白质序列与结构间的内在规律，开发出一系列创新型算法与模型，极大提升了结构预测的精度与速度 [1-4]。其中，谷歌 DeepMind 推出的 AlphaFold 系列模型能够以极高的精度预测众多未知蛋白质及其复合物的三维结构，不仅大大加速了重组蛋白质结构解析以及原位结构解析的过程，甚至在一定程度上，通过提供高精度的三维结构模型，使得蛋白质功能研究不再过度依赖耗时耗力的实验结构解析 [5-8]。相关科学家 David Baker、Demis Hassabis 和 John Jumper 也因对蛋白质设计和蛋白质结构预测方面的突出贡献，荣获 2024 年诺贝尔化学奖。目前主流的蛋白质结构预测方法主要分为基于模板的结构预测方法、无模板的结构预测方法和基于深度学习的结构预测方法三大类。尽管不同的结构预测方法所涉及的具体预测过程不同，但其基本步骤存在共性，往往包括构象初始化 (含模板识别和从头建模)、构象搜索、结构筛选、全原子结构重建和结构优化等 [9, 10]。

鉴于蛋白质结构预测技术蕴含的巨大发展潜力，本文利用全球专利数据，基于专利计量分析方法 [11-13]，对蛋白质结构预测相关专利的申请态势、技术热点和技术演进展开深入分析，全面探讨该领域发展现状，以期为我国未来蛋白质结构预测领域研发攻关及宏观决策提供支撑和启示。

1 研究态势分析

1.1 数据与方法

采用 incoPat 专利数据库作为检索工具，以蛋白质结构预测为主题，检索策略为：IPC = (G16B15 OR G16B40 OR G06N3 OR G06N20 OR G16B30 OR G16H50 OR G16B20 OR G06F19 OR G16B25 OR G16B35 OR G16C20) AND ((BACKGROUND-ART= (蛋白 (2W) 结构) AND 构象 AND 预测) OR (TIABC = (蛋白 (4W) 结构) AND (预测 OR 构象 OR 计算) AND (序列 OR 残基) AND (图神经网络 OR 拓扑结构 OR 欧氏距离 OR 矩阵 OR TM-align OR 损失函数 OR MSA)) OR TIABC = (protein structure OR protein fold) AND (neural network OR MSA OR distance map) OR TI = (protein AND structure* AND predict*))。检索截止时间为 2024 年 5 月 18 日，经人工阅读降噪后获得相关专利 1 164 项。

1.2 专利申请趋势

蛋白质结构预测领域专利申请量变化大致可划分为三个阶段：(1) 技术萌芽期 (1991—1999 年)，全球年专利申请量仅个位数；(2) 缓慢发展期 (2000—2012 年)，随着机器学习技术的快速发展，蛋白质结构预测技术迎来一段相对平稳的发展阶段，2005 年 I-TASSER 首次发布并被广泛应用，自此更多研究团队基于该算法陆续开发出多种蛋白质结构预测方法；(3) 快速发展期 (2013 至今)，以 ResNet、Transformer 为代表的深度学习算法为蛋白质研究带来革命性变革，领域关注度不断提高。2018 年，DeepMind 的 AlphaFold 问世；2019 年，蛋白质结构预测专利全球年申请量首次突破 100 件并迅速增加；2020 年，AlphaFold2 凭借深度学习架构大幅提升蛋白质结构预测精度 [5]，推动 AI 赋能相关专利申请的突破性增长 (图 1)。从当前趋势来看，该领域专利申请量尚未达到平台期，预计未来会有更多创新主体加入，专利申请或将迎来下一轮爆发式增长。

1.3 主要技术来源地和专利布局地分析

优先权号是申请人首次在某个国家或地区提交专利申请时获得的申请号，同一基础申请对应唯一优先权号。利用优先权号进行国别统计，分析全球范围内蛋白质结构预测专利申请的主要技术来源地。结果显示中国和美国是蛋白质结构预测的主要贡献地，其中以中国为首，贡献了 787 项专利申请，其次是美国、欧洲、韩国、日本、加拿大、以色列和印度 (图 2A)。公开号在申请公开阶段获得，同一申请在不同国家或地区可能有多个公开号。利用公开号进行国别统计，分析全球范围内蛋白质结构预测技术的专利布局，发现中国、美国和世界知识

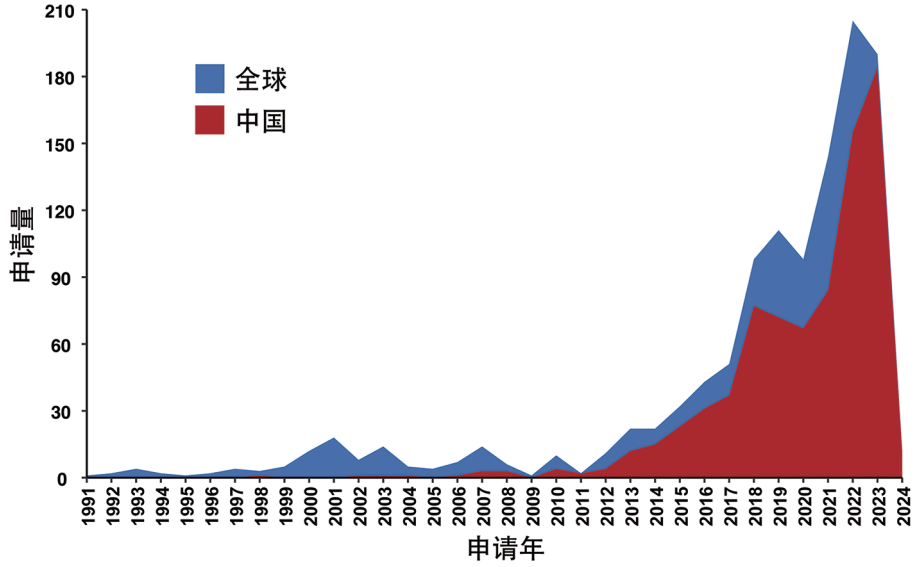
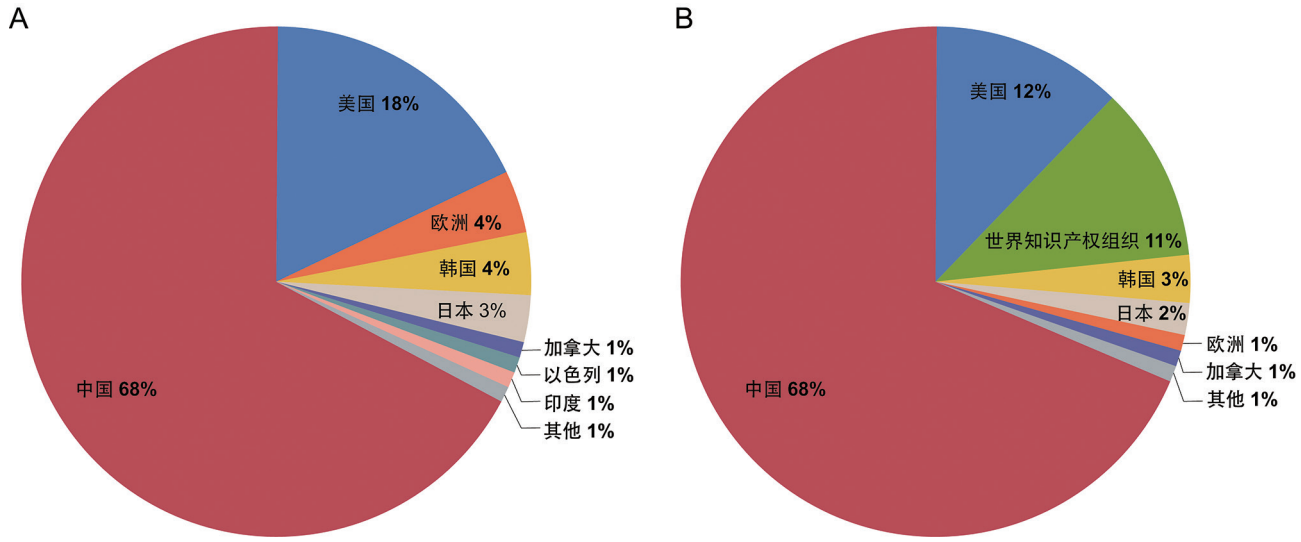


图1　蛋白质结构预测领域专利申请量年度分布



A：技术来源地分布；B：专利布局地分布

图2　全球蛋白质结构预测技术分布情况

产权组织是蛋白质结构预测领域专利公开数量最多的国家和地区。其中，在世界知识产权组织公开的专利申请多为 PCT 申请在国际阶段的公开，申请人后续可以选择进入不同的目标国，体现了蛋白质结构预测领域申请人重视国际市场的专利布局。从在国家阶段的公开数量来看，位居前四的依次为中国、美国、韩国和日本，在中国公开的专利为 790 件，远超在其他国家的公开总和，部分原因是这些申请大部分是中国申请，这也反映出我国是该领域主要的技术来源国 (图 2B)。

1.4　主要申请人

对全球各申请人的申请量进行分析可知，蛋白质结构预测领域全球排名前十的申请人中，浙江工业大学以 177 件专利申请遥遥领先，上海交通大学、DeepMind 和腾讯分别位居第二、第三和第四，中国石油大学、南京理工大学和 Illumina 紧随其后。国外申请量排名靠前的专利申请人多为公司，包括 DeepMind、Illumina、日本电气、PHARMCADD、Flagship Poineering 等；而我国申请量排名靠前的申请人中，大多是高校和科研院所，包括浙江工业大

学、上海交通大学、中国石油大学、南京理工大学、中南大学、之江实验室、中国海洋大学、北京航空航天大学等 (图 3)。可以看出，国内外创新主体在该领域研发模式和研发路径存在差异，国外诸如 DeepMind 等科技公司创新能力极强，凭借其算法、算力、人才、资金优势研发出一系列突破性技术。相比之下，国内虽然也有腾讯这样的大型科技公司投入蛋白质结构预测领域的技术研发，但其在基础研究和科技突破方面相较于国外科技巨头仍有一定距离；而高校和科研院所作为国内基础研究的主力军和重大科技突破的策源地，为该领域的创新发展做出主要的贡献，上述差异也一定程度上反映出我国蛋白质结构预测技术更多处于科研孵化阶段，离以公司为主导的产业化应用还有一定距离。

1.5 技术构成

1.5.1 输入端数据

输入端数据是指构建预测模型中需要输入的数据，包括：序列信息，二级结构、二面角等非完整蛋白结构 (多作为辅助特征用于功能预测模型)，同源蛋白已知结构信息，化学小分子结构等非蛋白质结构信息，理化性质信息，相互作用，三维照片和其他数据类型。蛋白质结构预测主要是单独依赖序列信息，或者序列信息结合序列信息或同源蛋白已知结构信息 (如 PDB 等结构数据库中获得的)，依赖性质、功能或图像的预测研究还比较少 (图 4)。当前，大部分结构数据通过如 X 射线晶体学、冷冻

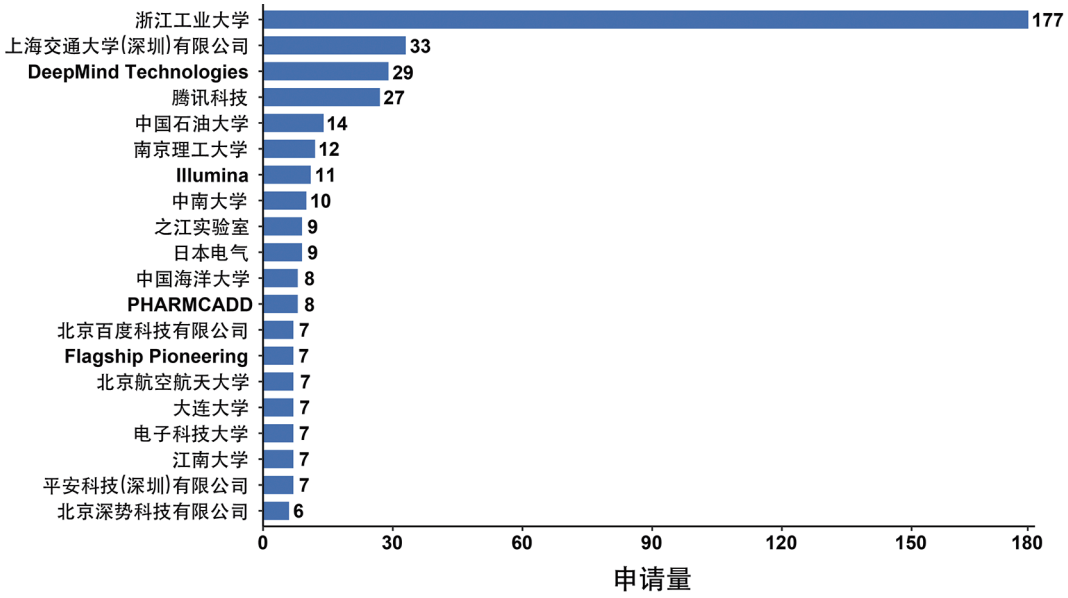


图3 全球蛋白质结构预测技术重要申请人排名

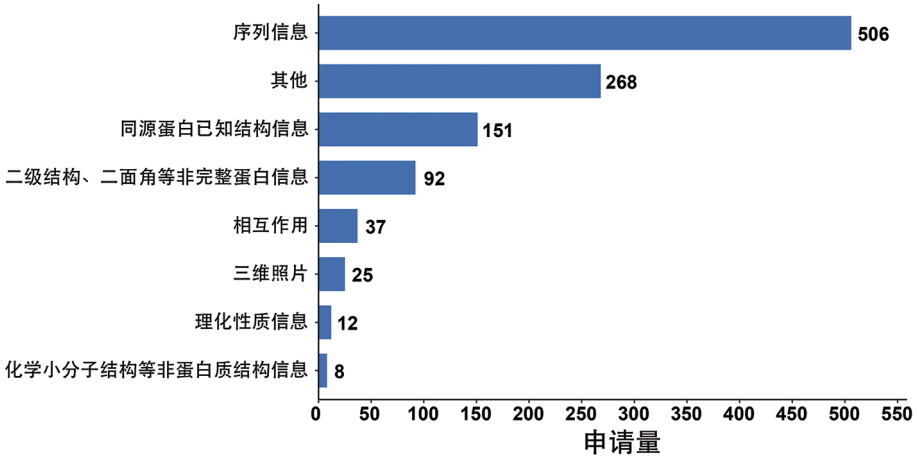


图4 输入端数据类型分析

电镜等实验手段获取，虽精度很高但耗时耗力；PDB 等三维结构数据集中样本规模不足，现存数据量难以满足对结构预测准确性进行有效评估的需求，成为制约蛋白质结构预测技术发展的关键掣肘之一；相比而言，序列数据的获取相对容易，导致蛋白质序列数据远远超过已知的结构数据，但是当无法获得同一性较高的同源蛋白时，基于多序列比对 (MSA) 的相关算法经常得不到合理的模型 [14]。随着多模态算法的发展，可以预期蛋白质序列数据、结构数据、分子动力学、蛋白质组学研究结果、小角散射数据以及一些其他实验相关的数据都有望作为有效信息加入预测模型。

1.5.2　预测方法

对蛋白质结构预测专利申请涉及的预测方法统计发现，基于深度学习的结构预测方法占比最高，达到 58%，其次是基于模板的结构预测方法，占比 23%，最后是无模板的结构预测方法，占比 19%。从蛋白质结构预测方法类型 - 时间变化趋势可以看出，随着人工智能领域的发展，从 2019 年开始，基于深度学习的结构预测方法快速增长并占据主导，远超其他两种方式 (图 5)。截至目前，基于深度学习的结构预测方法专利占总量的一半以上，其主要应用卷积神经网络、循环神经网络 (包括长短期记忆网络) 等多种神经网络技术和 Transformer 架构；基于模板的结构预测方法中使用 Swiss-Model 较多；无模板的结构预测方法中使用 Rosetta 较多，AlphaFold 其次。

1.5.3　应用方式

目前蛋白质结构预测领域的技术主要应用于通用的蛋白质结构预测，这种通用的预测不涉及特定的需求或目标，重点在于实现对蛋白质结构的认识和解析。除了上述通用性应用，该技术还被应用于蛋白质功能预测、药物设计与开发、个性化医学等特定需求的领域。自 20 世纪 90 年代起，有关蛋白质结构预测的研究便持续进行，而蛋白质性质、功能、设计的研究则开始于 21 世纪初 (图 6A)。在蛋白质结构预测中，三级结构预测占比 61%，二级结构预测占比 17%，四级 (多亚基) 结构预测占比 6% (图 6B)。蛋白质性质预测中亲和力预测略多，但差异不明显。蛋白质功能预测以蛋白质相互作用预测为主。在个性化医学中，预测突变蛋白的致病性为主要应用方向。

2　技术分支及演进

蛋白质结构预测技术主要包括基于模板的结构预测、无模板的结构预测和基于深度学习的结构预测方法。

2.1　蛋白质结构预测技术发展路线

2.1.1　基于模板的结构预测方法

基于模板的结构预测方法包括同源建模法和折叠识别法 (穿线法)。同源建模法是基于查询序列和已知结构模板蛋白的序列相似性，假定二者结构相似，以此预测查询序列结构，具体步骤为：运用 BLAST 等比对算法找出相似模板蛋白，优化模板蛋白结构以消除构象错误，再通过残基将查询序列映射到模板蛋白结构上。20 世纪 90 年代初，瑞士生物信息研究所的 Swiss-Model[15] 与加州大学 Sali Lab 的 Modeller[16] 推动了同源建模技术发展。为了解决 Modeller 需 30%~40% 以上序列同源性才能获得准确三维结构的问题，WO0204685A1 公开了一种改进方法，借助多重参考序列配对信息与实验结构，在低于 30% 同源性时也能提升配对准确性。

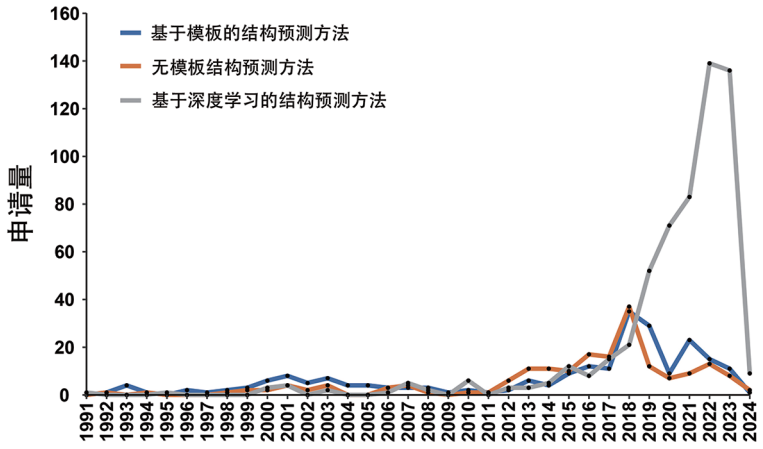
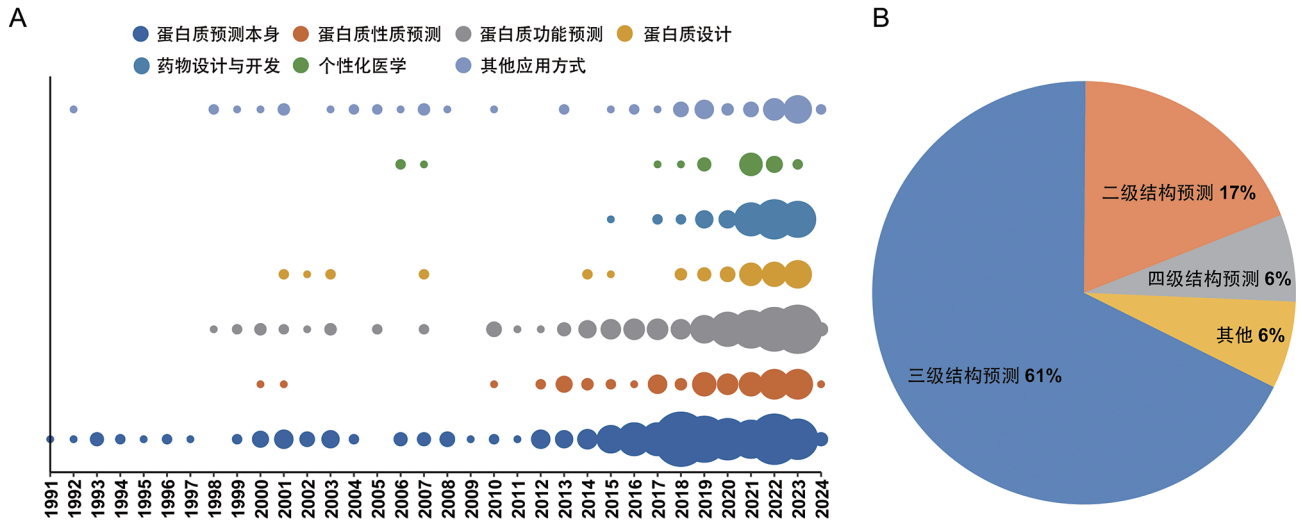


图5　预测方法类型分析



A：应用方式-时间气泡图，横轴为时间(年)，纵轴为不同的应用方式，气泡大小反映年申请量；B：结构预测类型

图6 应用方式分析

2003 年，许锦波团队开发了 Raptor[17]，基于纳入成对相互作用偏好的比对模型，运用整数编程实现全局最优且有效的穿线算法，开创整数规划和线性规划用于蛋白质穿线的先河 (US2004219601A1)。2010 年左右，张阳实验室的 I-TASSER[18] 基于线程技术，经迭代逐步精细化蛋白质三维结构，多次在 CASP 中获得优异成绩，后续还衍生出 C-I-TASSER、D-I-TASSER 等工具。

2.1.2 无模板的结构预测方法

无模板的结构预测方法依据蛋白质物理化学性质，如氨基酸间相互作用 (氢键、疏水作用、静电作用等) 和能量最小化原理，直接从氨基酸序列预测三维结构，不依赖已知模板。1998 年，华盛顿大学 David Baker 团队开发 Rosetta[19] 工具，采用“蒙特卡洛模拟”技术，借助蛋白质片段库随机折叠片段并计算能量，搜索可能结构，其高度可定制与灵活，成为生物学和药物设计领域极其重要的工具。2001 年，David T. Jones 团队开发 FragFold[20]，用模拟退火算法组装高分辨率结构提取的超次级结构片段，但准确性预测方面尚需完善。Xncor 公司公开了蛋白质设计自动化用于产生计算预筛选的蛋白质二级文库的用途、使用该文库的方法和组合物 (WO03014325A2)。其他代表性方法还有张阳实验室的 QUARK、C-QUARK 和 D-QUARK，David Baker 实验室的 RosettaCM、trRosetta，许锦波实验室的 RaptorX、RaptorX-Contact，程建林实验室的 CONFOLD、CONFOLD2 等 [21]，但大多并未申请专利。浙江工业大学在此期间布局了诸多相关专利，包括 CN103714265A、CN103984878A 和 CN106778059A，分别公开了一种基于蒙特卡洛局部抖动和片段组装的蛋白质三维结构预测方法、基于树搜索和片段组装的蛋白质结构预测方法和基于 Rosetta 局部增强的群体蛋白质结构预测方法。

2.1.3 基于深度学习的结构预测方法

基于深度学习的结构预测方法是通过深度学习模型直接从蛋白质的氨基酸序列预测其三维结构 [22, 23]。模型经过大量序列与对应三维结构数据训练，学习二者复杂的映射关系，从而能够预测新的氨基酸序列的结构。2018 年，DeepMind 开发的 AlphaFold 利用深度学习，尤其是卷积神经网络，从蛋白质序列直接预测其空间结构 (WO2020058177A1)，该模型在 CASP 竞赛中打破了 I-TASSER 多年的连冠纪录，标志着人工智能在该领域取得突破。2020 年，AlphaFold2[5] 引入 Transformer 架构提升蛋白质间接触点预测精度，进而提升整体结构预测精度，且能更高效处理蛋白质家族相关结构。AlphaFold2 的推出伴随着一个重要的里程碑 —AlphaFold 蛋白质结构数据库的建立，该数据库由 DeepMind 与 EBI 合作创建，为全球科研人员提供开放且易于访问的蛋白质结构数据，包括由 AlphaFold2 预测的大规模蛋白质结构数据集。DeepMind 基于 AlphaFold2 布局了系列专利，包括 US2021398606A1、US20211-66779A1、WO2022112248A1、WO2022112257A1、WO2022112260A1、WO2022089805A1、WO2022-167325A1 和 WO2023- 094335A1 等。除了 AlphaFold 系列模型外，其他代表性的方法包括 David Baker 实

验室的 RoseTTAFold、Meta AI 的 ESMFold、David T. Tones 实验室的 DMP- fold2、百图生科的 HelixFoldSingle (CN115458042A) 和 HelixFold-Multimer (CN116052758A) 等 (图 7)。

2.2　重要申请人的专利布局

2.2.1　DeepMind

DeepMind 在 2018 年申请的专利 WO2020058177A1 涉及 AlphaFold 的算法框架，它将复杂的蛋白质结构预测问题转化为蛋白质折叠性质预测约束求解，用改进梯度下降法求最佳结构，展现出对深度学习和生物学的深刻理解。AlphaFold2 的整体框架包括三部分：一是特征提取模块，根据氨基酸序列生成 MSA 和成对表示，表征共进化和结构约束信息；二是编码模块，由 Evoformer 构成，利用特征信息推理蛋白质的空间和进化关系，其模块通过自注意力机制和几何变换更新 MSA 及成对表示，专利 WO2022112248A1 阐释了其架构和实现；三是结构解码模块，使用不变点注意力更新单一表示，提高预测准确性和稳定性。AlphaFold2 有诸多算法改进。2018 年和 2019 年申请的专利 US2021398606A1、US2021166779A1 体现其引入注意力机制的亮点，增强了预测能力。它利用自蒸馏进行模型训练，专利 WO2022089805A1 体现此设计思路，可学习更多结构特征，预测更复杂多样的结构。AlphaFold2 还采用端到端框架减少信息损失和噪声，通过循环机制更新优化输出，WO2022112257A1 涉及中间损失函数的辅助损失函数以优化性能，WO2022112260A1 涉及循环机制提升性能。AlphaFold 系列算法不仅能预测单链蛋白质三维结构，还可预测多蛋白质复合物结构，为蛋白质功能研究、设计和药物研发提供了有力的工具。2021 年的 WO2022167325A1 和 WO2023094335A1 涉及蛋白质设计领域应用，如受体、酶的激动剂或抑制剂及抗体设计；2022 年的 WO2024079204A1 应用于个性化医学领域，可预测蛋白突变体致病性，有利于生物标志物挖掘和致病机制研究 (图 8)。

2.2.2　腾讯

腾讯 AI Lab 的 tFold 算法具有以下三个特点：一是“多数据来源融合”技术，可挖掘多组多序列比对的共进化信息；二是“深度交叉注意力残差网络”，显著提升蛋白质二维结构信息 (如残基对距离矩阵) 预测精度；三是“模板辅助自由建模”方法，整合自由和模板建模的三维构型信息，增强最终三维建模准确性。2019—2022 年申请的专利 CN110-706738A、CN114613429A、CN114974397A 和 CN115132277A 均涉及蛋白质结构预测算法的设计与研发，包括序列特征扩增模型、构象筛选、知识蒸馏、结构质量评估 [24, 25] 等方面。“云深智药”平台基于 6 大模块实现两大功能：小分子药物和大分子药物 (主要是抗体药物) 的发现。其蛋白质结构预测模块基于 tFold 提升建模精度；抗体模块在 tFold 基础上开发了相关功能，涵盖抗体三维结构、复合物结构预测和亲和力优化；虚拟筛选模块利用人工智能学习小分子结构与生物活性关系。腾讯 2021—2023 年在蛋白质结构预测的下游应用领域申请大量专利，在药物设计与开发领域，CN115472239A 涉及药物活性预测、CN114283878A 涉及蛋白质药物设计、CN117012270A 涉及抗体结构预测、CN11-6959558A 涉及抗原抗体复合物结构预测、CN11-6959589A 涉及抗体结构索引的构建；在蛋白质功能或性质预测领域，也申请了一系列涵盖结合位点预测、相互作用预测以及亲和力预测的专利，包括 CN11433980A、CN117037936A 和 CN115116538A 等 (图 9)。

3　产业发展机遇与挑战

3.1　中美占据技术主导，美国占据先机，中国乘势追赶

在“百年未有之大变局”下，中美在生物医药关键技术创新领域的战略竞争态势凸显 [26]。以人工智能为代表的新一代信息技术快速发展，为生物医药创新带来历史机遇，使其成为中美战略竞争的重点领域，其中，蛋白质结构预测作为人工智能深度参与并加速生物医药创新的典型细分领域，正成为两国科技博弈的重要战场。专利分析结果显示，中国在该领域的专利申请总量已超越美国，但质量维度仍存在一定差距。在技术创新层面，中国在蛋白质结构预测领域专利申请虽起步晚于美国，但近年来却逐渐超越美国成为最大的技术来源国与目标国。中国的高校和科研院所在该领域的研发创新实力不容小觑，多所高校专利申请量位居前列，贡献了大量的新思路与技术。相比之下，美国则呈现“企业主导”态势，DeepMind、英伟达等企业凭借强大的算力资源与算法开发人才，掌握了如 AlphaFold2 这样的全球领先的核心算法、平台和技术。随着两国科创竞争加剧，预计在蛋白质结构预测这一高潜力领域的研发投入将持续增加。尽管中国高校和科研院所专利申请量可观，但美国大型公司凭借资金、

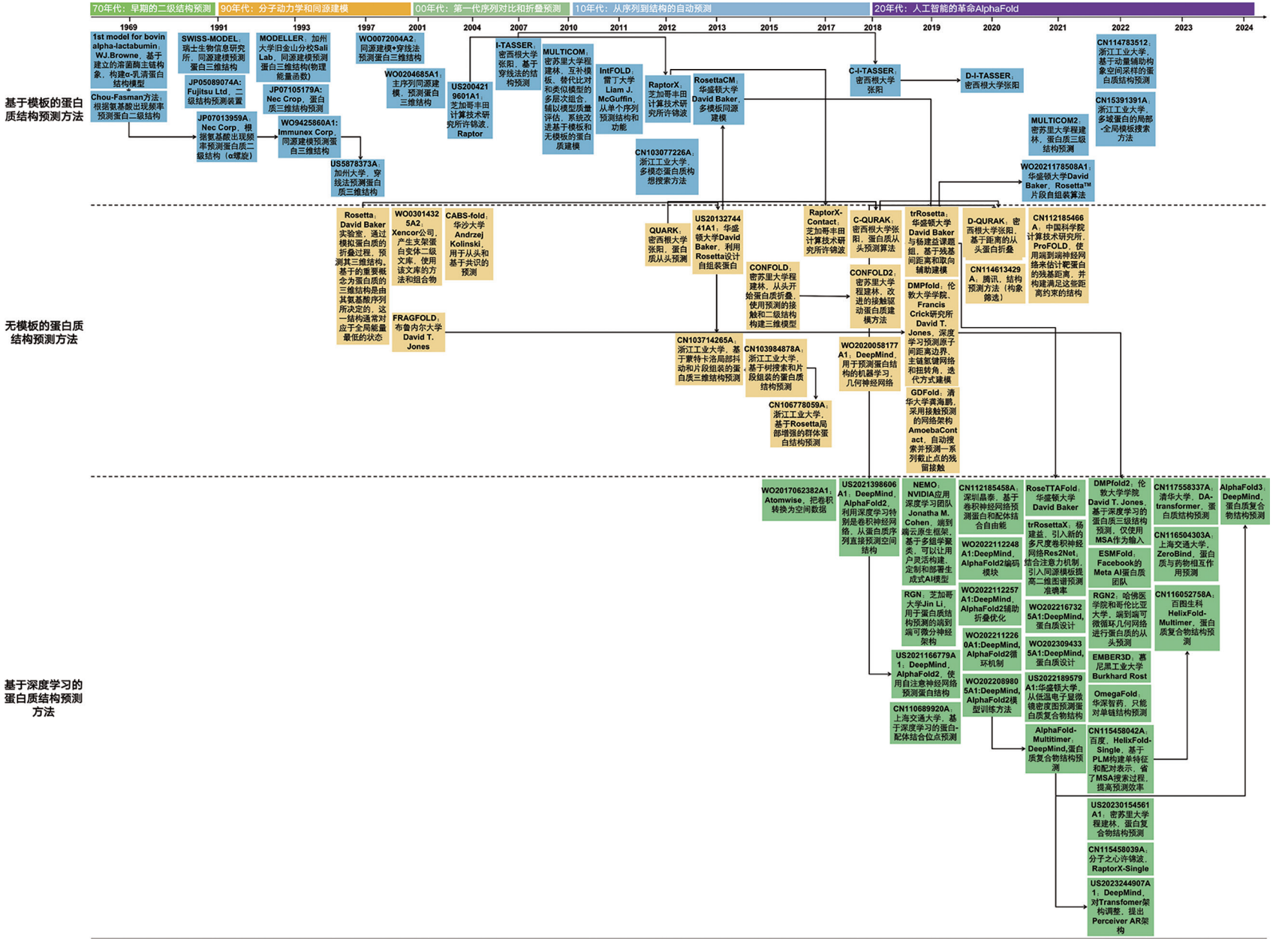


图7 蛋白质结构预测领域全球技术路线

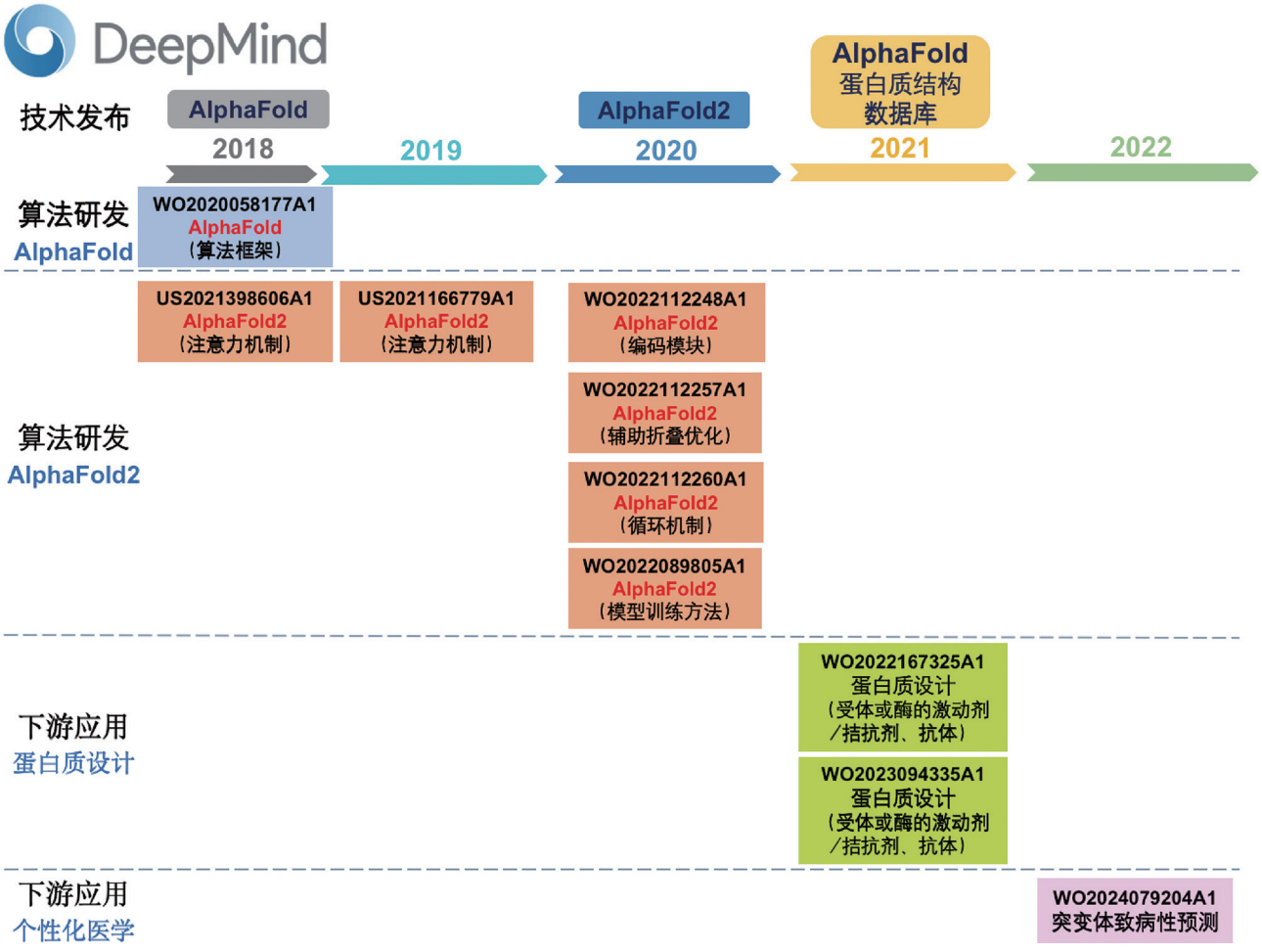


图8 DeepMind公司专利布局

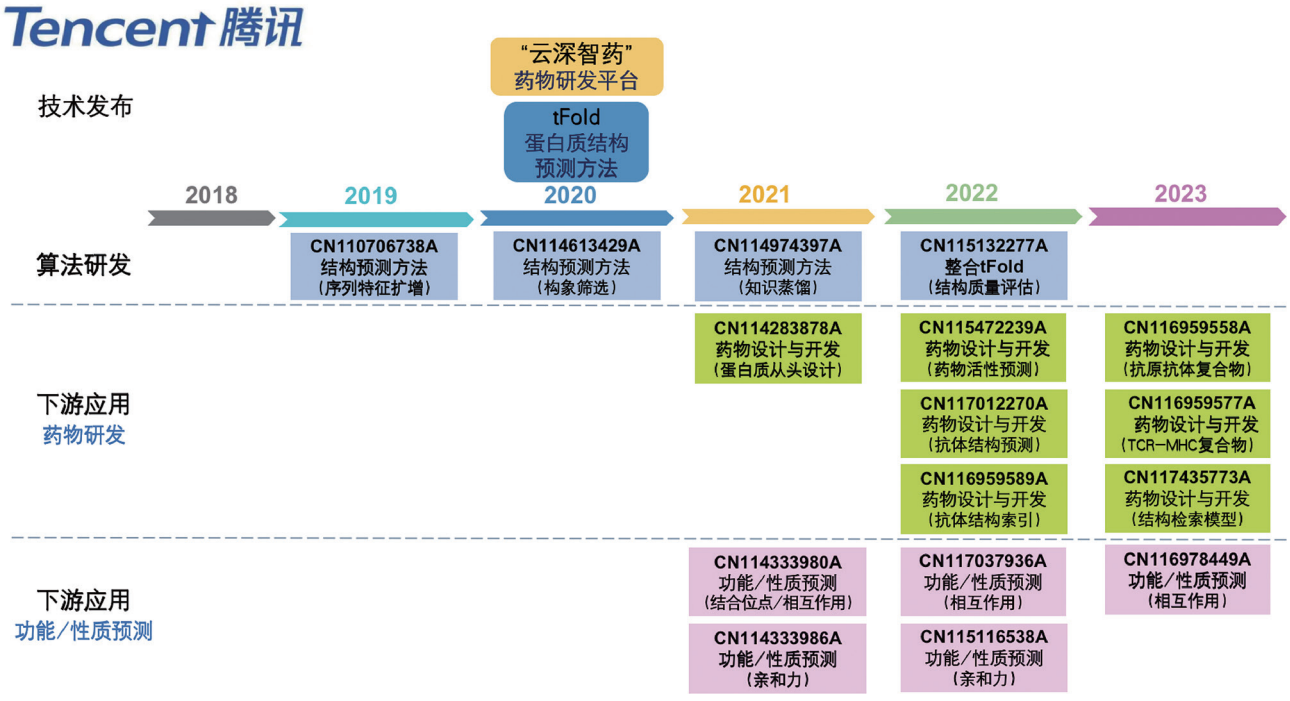


图9 腾讯公司专利布局

算力及人才优势，掌握核心底层算法技术，将会进一步巩固其研发优势，抢占技术先机。在市场应用方面，中美两国企业积极布局蛋白质结构预测技术在新药研发产业链的应用，但产业转化维度呈现显著梯度差。美国已形成“基础算法 - 计算平台 - 药物管线”的垂直生态，例如，英特尔结合自身优势为 AlphaFold2 提供算力优化，DeepMind 公司开发的 AlphaFold 系列算法高效解析蛋白质结构，Generate Biomedicines 开发蛋白质生成模型 Chroma，通过人工智能技术理解蛋白质序列、结构与其功能之间的关系，并应用于肿瘤、传染病药物管线。反观中国，虽涌现出百图生科、云深智药、医疗智能体等 AI 新药研发机构，角逐药物研发、精准诊断等市场，但关键环节仍受制于算力基础设施缺口和复合型人才不足 [27]。

3.2 全球蛋白质结构预测技术方兴未艾，应用端热点集中、空白待填

蛋白质结构预测已全面融入新药开发全流程，在靶点发现、化合物虚拟筛选等环节广泛应用，有效提升研发效率与成功率，并显著降低成本。专利数分析表明，新药研发是蛋白质结构预测技术当下最具潜力与价值的应用领域，也是相关专利涉及的重点下游应用。同时，蛋白质结构预测技术在合成生物学领域也有涉足。例如，通过开发蛋白质设计算法分析酶活性以指导新型催化剂设计，利用蛋白质自组装特性开发纳米材料等。但相较于新药研发，该技术在合成生物学应用端的研究较少，存在诸多待填补的空白。合成生物学涵盖医药、能源、材料等多个领域，创新发展可从根本上改变经济发展模式，创造社会财富，促进社会稳定和谐。因此，我国应重视蛋白质结构预测技术在合成生物学中的创新应用，提前布局，抢占研发先机，推动生物制造产业高质量发展。

3.3 数字资源平台滞后，数据质量制约发展

专利分析显示大量专利仅涉及单域蛋白质结构预测，这反映出蛋白质结构预测模型训练对实验测定的蛋白质结构数据的依赖，PDB 数据库中单域蛋白结构信息丰富，为模型训练提供更多数据，侧面体现数据对研发的制约。对比国内外数据资源平台建设，根据市场研究机构 Synergy Research Group 的数据显示，截至 2024 年底，全球运营中的超大规模数据中心数量已达 1 136 个，其中，54% 位于美国，中国和欧洲占比分别为 16% 和 15%[28]。美国高度重视数据库平台建设，在生物医学领域拥有 PubMed、GenBank 等重要数据库，还开发针对机器学习的大型数据集，并将拓展至更多前沿领域。我国 2010 年后才开启生物医药重大科学基础设施建设，虽已建成部分中心，但在数据平台建设、数据资源管理与共享等方面经验和能力不足，高质量数据集匮乏，严重制约蛋白质结构预测技术研发。我国需加强数据资源平台建设与维护，构建自主数据库，完善数据管理规则，营造良好数据环境。

3.4 跨学科人才短缺，加强交流迫在眉睫

专利分析表明，我国高校和科研院所在蛋白质结构预测领域专利申请多，研发实力较强。这源于该领域研发需多学科知识与思维，高校和科研院所汇聚了生物信息学、计算化学、深度学习等多学科人才，其协同创新效率在基础研究阶段表现突出。然而，基于 2024 年美国保尔森基金会下属的智库公布的“全球人工智能人才追踪”调查可知，美国在顶级 AI 人才方面有着明显的领先优势，其拥有全球 60% 的顶级 AI 研究机构，是全球 57% 的精英 AI 人才的首选就业地；中国仅次于美国，但也只有 12% 的精英 AI 人才首选在中国就业 [29]。可见美国跨学科高端研发人才资源丰富，这是其掌握核心算法技术的重要因素。因此，扩充与建设跨学科高端人才队伍是我国发展蛋白质结构预测领域、缩小与美国技术差距的关键。一方面，要发挥高校和科研院所人才优势，推进产学研融合，加强人才交流与合作，畅通科技成果转化渠道；另一方面，加大生物医药与人工智能交叉学科教育投入，优化人才培养体系，加强国际交流合作，引进海外高端人才。

4 结语

本文以蛋白质结构预测技术为切入点，分析了全球和中国相关专利态势与关键技术演进路线。在此基础上，对我国该行业发展提出以下建议。

一是坚持创新驱动发展。蛋白质结构预测技术的核心在于算法创新，随着人工智能技术的深度融合，蛋白质结构预测在技术研发和算法开发方面有望取得更多突破，包括：预测效率和准确度的提升，预测范围从单结构域蛋白拓展至多结构域蛋白、从单一蛋白发展至蛋白质与其他生物分子的复合物、从静态结构预测升级为动态结构预测等。在国内外产业政策的积极支持与鼓励下，资金和人才不断涌入蛋白质结构预测技术创新领域。该领域的算法创新，将成为驱动技术进步和产业发展的“加速器”，人工智能算法框架的创新，以及生物数据获取技术

的革新，都将有力助推蛋白质结构预测领域实现革命性进步。

二是拓宽下游应用场景。蛋白质结构预测技术的应用正迅速扩展到多个领域。在生物医药领域，除药物研发外，该技术还可以应用于精准医疗，例如，通过蛋白质结构预测算法，依据基因突变后的序列预测其结构，从蛋白质结构与功能角度深入探究基因突变位点引发疾病的机制，助力生物标志物挖掘。在生物制造领域，应用场景更为多元，例如，将蛋白质结构预测与蛋白质设计相结合以改造酶类蛋白质，提升相关行业的生产效率；赋能生物材料领域，辅助设计与优化生物材料；推动作物蛋白质改良，提升食品产量和质量；助力研发针对性强、高效的生物降解剂，解决污染物处理问题，改善生态环境。蛋白质结构预测技术在各领域的深度渗透，将催生出更为多元的应用场景，推动产业转型升级。

三是加强数据平台建设。受限于已知的蛋白质结构数据的有限性，人工智能在构建高阶蛋白质复杂构象预测模型及蛋白质动态结构预测模型时，面临数据不足的困境。因此，需持续加强蛋白质结构数据的积累，推进蛋白质结构数据库的开发与完善，为基于人工智能的蛋白质结构预测模型的构建与训练提供丰富的数据来源。在数据平台建设与数据质量提升方面，应进一步完善平台建设的顶层设计与规划，明确建设目标与需求，制定数据标准体系，强化对数据源的统一管理与调度，建立数据源间的关联关系，完善数据质量监控与反馈体系。此外，在基于大数据训练及构建模型的领域，随着技术在产业中的深入渗透，数据安全与隐私、数据开放与共享等问题备受关注。该领域在技术发展的同时，需重视保障数据安全与隐私，制定契合技术及产业发展的合理规范和政策；同时，也应积极促进数据的开放与共享，助力技术研发的交流与合作，共同推进技术革新。

四是重视人才交流培养。在教育体系方面，应进一步突破传统学科与专业的固有壁垒，推动跨学科课程设置。例如，加强蛋白质结构预测技术所需的生物信息学学科建设；重视通识教育与专业教育的融合，强化计算机技术相关的通识教育，提升人才的综合素质及跨界能力。在人才交流方面，需进一步加强高校和企业的人才交流，让蛋白质结构预测领域的高校发明人更深入地了解企业的算法研发需求及技术应用场景，促使高校发明人在进行技术创新时充分考虑产业需求，研发出更易落地的技术及方法。同时，使该领域的企业发明人更好地跟进生物、化学等领域的基础研究进展，加深对蛋白质结构预测技术背后生物学知识的理解，并将其转化为算法研发的助力。

[参 考 文 献]

[1] 王芳, 李洪进, 李虎阳. 蛋白质结构预测的方法探究. 现代信息科技, 2022, 6: 122-5

[2] 曹卫, 潘宪明. 蛋白质结构预测进展. 生物化学与生物物理进展, 2023, 50: 1190-4

[3] Peng CX, Liang F, Xia YH, et al. Recent advances and challenges in protein structure prediction. J Chem Inf Model, 2024, 64: 76-95

[4] 张贵军, 侯铭桦, 彭春祥, 等. 多结构域蛋白质结构预测方法综述. 电子科技大学学报, 2022, 51: 820-9

[5] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583-9

[6] 张弘, 王慧洁, 鲁睿捷, 等. 蛋白质结构预测模型AlphaFold2的应用进展. 生物工程学报, 2024, 40: 1406-20

[7] 陈志航, 季梦麟, 戚逸飞. 人工智能蛋白质结构设计算法研究进展. 合成生物学, 2023, 4: 464-87

[8] Jussupow A, Kaila VRI. Effective molecular dynamics from neural network-based structure prediction models. J Chem Theory Comput, 2023, 19: 1965-75

[9] 王超, 朱建伟, 张海仓, 等. 蛋白质三级结构预测算法综述. 计算机学报, 2018, 41: 760-79

[10] 邓海游, 贾亚, 张阳. 蛋白质结构预测. 物理学报, 2016, 65: 178701

[11] 吕璐成, 郑丽丽, 赵亚娟. 克力芝药物全球专利布局与研发态势分析. 科学观察, 2020, 15: 45-51

[12] 上官晨虹, 全毅恒, 陈琛. 基于专利计量的天麻药品研发态势分析. 中草药, 2022, 53: 4915-24

[13] 吕璐成, 郑丽丽, 赵亚娟. 法匹拉韦药物全球专利布局与研发态势分析. 科学观察, 2020, 15: 1-10

[14] 黄鹤, 吴桐, 王闻达, 等. 蛋白质复合物结构预测: 方法与进展. 合成生物学, 2023, 4: 507-23

[15] Schwede T, Kopp J, Guex N, et al. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res, 2003, 31: 3381-5

[16] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol, 1993, 234: 779-815

[17] Xu J, Li M, Kim D, et al. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol, 2003, 1: 95-117

[18] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc, 2010, 5: 725-38

[19] Bonneau R, Tsai J, Ruczinski I, et al. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins, 2001, Suppl 5: 119-26

[20] Jones DT, Bryson K, Coleman A, et al. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins, 2005, 61 Suppl 7: 143-51

[21] Kinch LN, Pei J, Kryshtafovych A, et al. Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (CASP14). Proteins, 2021, 89: 1673-86
[22] 杨璐. 基于深度学习的蛋白质二级结构预测[D]. 无锡: 江南大学, 2023
[23] 包晨. 深度学习算法在蛋白质结构预测中的应用[D]. 无锡: 江南大学, 2020
[24] Liu D, Zhang B, Liu J, et al. Assessing protein model quality based on deep graph coupled networks using protein language model. Brief Bioinform, 2024, 25: bbad420
[25] 刘栋, 崔新月, 王浩东, 等. 蛋白质结构模型质量评估方法综述. 物理学报, 2023, 72: 248702
[26] 王楠, 王国强. 新竞争格局下中美生物医药创新对比研究. 中国软科学, 2023, (01): 22-31
[27] 刘晓凡, 孙翔宇, 朱迅. 人工智能在新药研发中的应用现状与挑战. 药学进展, 2021, 45: 494-501
[28] Hyperscale Data Center Count Hits 1,136; Average Size Increases; US Accounts for 54% of Total Capacity. https://www.srgresearch.com/articles/hyperscale-data-center-count-hits-1136-average-size-increases-us-accounts-for-54-of-total-capacity#:~:text=New%20data%20from%20Synergy%20Research%20Group%20shows%20that,2024%2C%20having%20doubled%20over%20the%20last%20five%20years[EB/OL]. [2025-09-16]
[29] The global AI talent tracker 2.0. MacroPolo. (2024, March 6). The Global AI Talent Tracker 2.0. https://archive-macropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/[EB/OL]. [2025-09-16]