

特色专题

人工智能在药物研发中的进展与展望

李洪林^{1,2}

摘要 药物研发具有周期长、成本高、临床成功率低等特点,传统研发模式已难以满足复杂疾病和创新药物开发的需求。近年来,人工智能(artificial intelligence, AI)凭借其在多源数据融合、复杂模式识别和智能决策等方面的优势,正推动药物研发由经验驱动向数据驱动和知识驱动的新范式转变。本文系统综述了国内外 AI 赋能药物研发的最新研究进展,重点分析了 AI 在药物研发全流程中的关键应用及其技术发展趋势,并结合代表性研究总结了 AI 在提高研发效率、优化决策过程和促进研发模式变革方面的重要作用。综合现有研究表明, AI 已成为连接生物医学数据、计算模型与实验验证的重要桥梁,不仅显著提升了药物研发早期的靶标发现、分子设计和药物评价效率,而且推动了自动化实验、自驱动实验室和虚拟细胞等新型研发模式的发展,为构建智能化药物研发体系提供了新的技术路径。然而, AI 在药物研发中的广泛应用仍受到高质量数据不足、模型可解释性和泛化能力有限、多模态知识融合困难,以及实验验证和标准化评价体系不完善等因素的制约,其临床转化价值仍需要进一步验证。未来,应加强高质量数据资源和共享体系建设,发展融合生物机制与物理规律的可解释 AI 模型,推动生成式 AI、多智能体系统与自驱动实验室深度融合,完善模型评价标准与监管体系,促进人工智能与实验科学协同创新,加快 AI 赋能药物研发向智能化、自动化和闭环化方向发展。

关键词 人工智能; 药物研发; 智能药物设计; 生成式人工智能; 自驱动实验室; 虚拟细胞

药物研发是一个周期长、成本高且风险极高的复杂过程,通常需要经历靶标识别、先导化合物发现与优化、临床前研究以及临床试验等多个阶段。从新药发现到最终获批上市通常需要 10~15 年,研发投入超过 20 亿美元,而进入临床试验的候选药物最终获批上市的比例不足 10%^[1]。大量候选药物在临床开发阶段因疗效不足或安全性问题而失败,尤其在 II 期临床试验中因无法验证治疗有效性导致的失败最为常见,反映出传统研发模式在早期靶标选择、候选分子评价及成药性预测方面仍存在明显局限^[2]。传统药物发现

主要依赖串行实验流程和高通量筛选策略,面对约 10^{60} ~ 10^{100} 数量级的潜在类药化合物空间以及复杂疾病生物学机制时,难以实现高效探索和精准预测,进一步增加了药物研发的不确定性,导致研发周期延长、资源投入增加和创新效率受阻^[3-4]。因此,开发能够提高药物发现效率和成功率的新型技术体系已成为生物医药领域的重要需求。

随着人工智能(artificial intelligence, AI)算法、计算能力及生物医药大数据的快速发展, AI 已成为推动药物研发模式转型的重要技术力量。依托机器学习、深度学习、生成式人工智能及大语言模型(large language models, LLMs)等技术, AI 能够整合多组学、化学结构及临床数据,实现复杂生物医学信息的深度

1. 华东师范大学药学院, 上海 200062

2. 人工智能新药创智中心, 上海 200062

收稿日期: 2026-06-05; 修回日期: 2026-07-19

基金项目: 国家杰出青年科学基金延续项目(82425104)

作者简介: 李洪林, 教授, 研究方向为药物靶标识别和药物发现的计算方法及其应用, 电子信箱: hlli@hsc.ecnu.edu.cn

引用格式: 李洪林. 人工智能在药物研发中的进展与展望[J]. 科技导报, 2026, 44(14): 104-119; doi:10.3981/j.issn.1000-7857.2026.06.00021

挖掘,并在靶标发现、分子生成与优化、药物-靶标相互作用预测、药效及吸收、分布、代谢、排泄、毒性(ADMET)评价、药物重定位和临床试验优化等关键

环节提供数据驱动的解决方案(图1),推动药物研发由传统经验驱动模式向智能化、系统化和预测驱动模式转变。

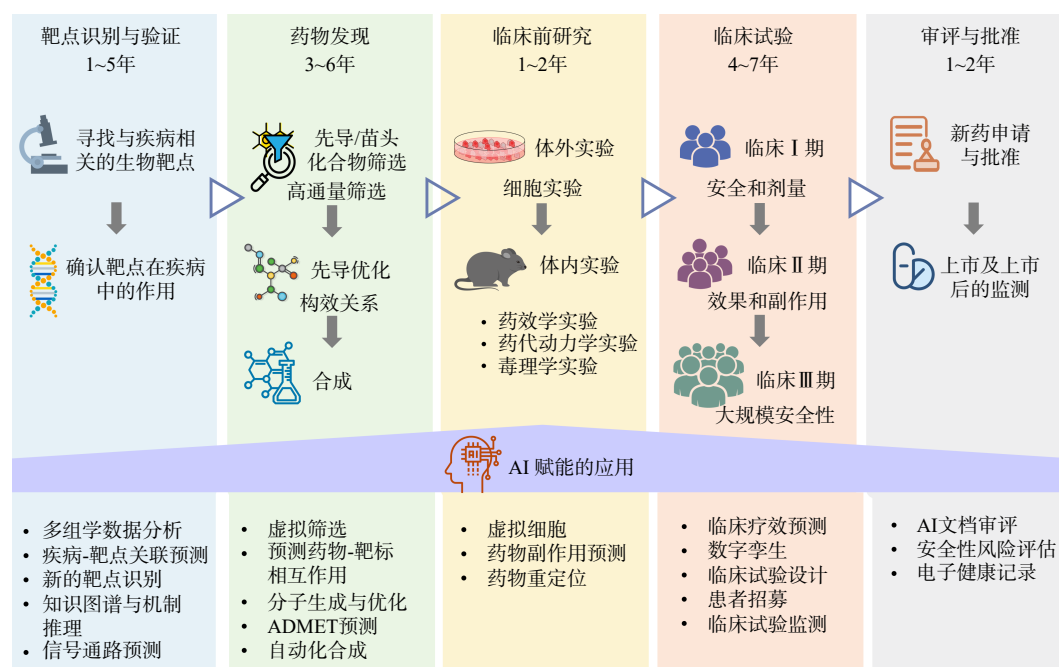


图1 AI 赋能的药物研发全流程

与此同时, AI 驱动的药物研发平台已显示出初步转化应用潜力,多种 AI 设计的候选药物已进入临床研究阶段。例如, Insilico Medicine 基于 Pharma.AI 平台完成特发性肺纤维化(idiopathic pulmonary fibrosis, IPF)相关新靶标发现,并在约 18 个月内推进候选药物进入临床前研究阶段,而传统类似流程通常需要 4~6 年; Exscientia 联合 Sumitomo Dainippon Pharma 利用 AI 辅助设计平台开发 5-HT_{1A} 受体激动剂 DSP-1181,在不到 12 个月内完成候选药物发现并进入 I 期临床研究^[5]。这些进展表明, AI 正在推动药物研发由传统经验驱动模式向数据驱动和智能化模式转变。随着 AI 技术应用范围不断扩大,全球主要制药企业和生物技术公司正在加快 AI 研发平台建设,通过自主研发、技术合作及战略投资等方式推动 AI 与药物研发流程深度融合。AI 正逐步由辅助分析工具发展为贯穿药物研发全流程的重要技术支撑,为缩短研发周期、降低研发成本及提高药物研发成功率提供新的技术路径。

1 靶标识别: AI 引领的单点预测到系统化智能推理的技术演进

靶标发现是药物研发的起始环节,其准确性和可靠性直接影响后续分子设计效率、临床转化潜力以及研发成功率。传统方法主要依赖遗传关联分析、基因微扰、化学筛选以及结构生物学研究等实验策略,但面对复杂疾病网络、多因素调控机制以及大量未知的分子相互作用和生物调控关系时,传统方法在研究通量、多源证据整合及复杂疾病机制解析方面逐渐面临挑战。随着多组学、单细胞、空间组学及临床数据持续积累, AI 逐渐由基于单一特征的关联预测发展为融合多源证据的系统化推理框架。通过表示学习、网络建模和知识推理, AI 能够整合疾病相关性、因果证据、可成药性与安全性信息,推动靶标研究由经验驱动和单因素分析模式向系统化证据评估与机制推理方向发展。

1.1 从单一关联预测到系统化证据推理

靶标发现包括疾病相关候选生成、因果证据评估、可成药性分析、优先级排序及实验验证等环节。

传统方法主要依赖遗传关联、基因微扰、化学筛选和结构生物学实验,容易受到样本规模、实验体系及先验假设的限制,且差异表达或网络相关性并不能直接证明疾病因果关系^[6-7]。人类遗传学证据可提高靶标选择的转化可靠性。大规模数据分析显示,具有遗传学证据支持的药物作用机制获得临床批准的概率约提高 2.6 倍^[8];缺少遗传学支持的项目更易因疗效不足终止,而靶标的组织广泛表达和功能约束与安全性风险相关^[9]。早期人工智能方法主要利用机器学习模型对基因表达、蛋白结构特征以及药物化学性质进行建模,实现疾病相关分子或潜在作用靶标的预测。然而,由于生物系统具有高度网络化特征,仅依赖单一特征的信息往往难以全面反映疾病机制。

AI 的发展使靶标研究由单一分子特征分析扩展至异构网络和多源证据建模。Luo 等^[10]提出 DTINet 模型通过整合药物、蛋白质、疾病及不良反应等网络,并结合网络嵌入方法学习不同实体之间的潜在关联,实现了药物-靶标相互作用预测和药物重定位任务,并对部分预测结果进行了实验验证。该模型在已知药物-靶标关联数据集上的交叉验证中表现出较高的预测性能,其 AUC 和 AUPR 指标均优于传统矩阵分解和随机游走方法。同时,DTINet 能够从复杂异构网络中挖掘潜在药物-疾病关联,为发现新的治疗靶标和药物适应症提供了计算依据。该方法突破了传统模型依赖单一分子特征的限制,为从复杂生物网络中发现潜在治疗靶标提供了新的计算框架。

在此基础上,知识图谱(knowledge graph, KG)的引入推动 AI 靶标发现向知识增强推理方向发展。知识图谱能够整合药物、疾病、基因、蛋白质以及生物通路之间的多维关系,使 AI 模型不仅能够学习数据中的统计关联,还能够结合已有生物医学知识进行推理。例如,Zhang 等^[11]开发的 KGE-UNIT 框架将知识图谱嵌入与卷积神经网络相结合,在药物-靶标相互作用及药物-药物相互作用预测任务中取得优于传统方法的性能。针对药物研发中普遍存在的数据稀疏和冷启动问题,KGE-NFM 进一步结合知识图谱表示学习与神经推荐模型,通过利用药物、靶标及疾病之间的潜在关系,提高了对缺乏已知配体的新靶标或新药物的预测能力^[12]。近期,知识图谱已开始从关联预测进一步向实际药物发现应用拓展。TxGNN 利用大

规模医学知识图谱学习疾病-药物关系,通过零样本学习预测缺乏治疗药物的新疾病适应症,并在多个疾病领域验证了其药物重定位能力^[13];Sudhakar 等^[14]进一步探索基于知识图谱生成具有机制解释性的治疗假设,并通过实验验证其可靠性。这些研究表明,知识图谱正在推动 AI 靶标发现由“预测潜在关联”向“提出可验证生物学假设”转变。

需要指出,药物与靶标相互作用预测及药物重定位主要用于补充潜在关联,不能替代疾病因果靶标的确认。AI 辅助靶标发现应形成“候选生成、证据融合、优先排序、实验验证”的连续流程,综合遗传学、多组学、结构、临床和文献证据,并通过基因编辑、细胞模型、类器官或动物模型验证候选靶标,从而提高靶标决策的可解释性与转化价值。

1.2 多模态数据与算法协同驱动的靶标优选

基因组、转录组、蛋白质组、单细胞组学、空间组学及真实世界数据的积累,为疾病机制解析和靶标优选提供了多层次证据。针对不同数据模态在尺度、噪声和语义上的差异,深度学习可通过表示学习、注意力机制和跨模态融合建立分子、细胞、组织与疾病表型之间的关联。scGPT 和 scFoundation 等基础模型通过大规模单细胞转录组预训练,为细胞状态表征、基因功能分析和扰动效应预测提供了通用框架^[15-16]。PINNACLE 进一步整合单细胞转录组、蛋白质相互作用网络和组织层级信息,生成具有细胞类型特异性的蛋白质表征,并用于类风湿关节炎和炎症性肠病的候选靶标排序^[17]。GET 则从染色质可及性和序列信息中学习细胞类型特异性转录调控规律,为识别上游调控因子提供了新的技术路径^[18]。

现有 AI 靶标发现算法可概括为 4 类互补路径。第 1 类通过多组学和单细胞数据识别疾病相关基因、异常细胞状态及调控网络。第 2 类利用图神经网络和知识图谱整合疾病、基因、蛋白质、药物及通路关系,开展链路预测和候选靶标排序^[13-14]。第 3 类利用蛋白质结构和序列模型评估靶标功能、相互作用及结构可干预性。AlphaFold2 提高了蛋白质结构预测的准确性,AlphaFold3 进一步扩展至蛋白质、核酸、小分子和离子复合物的结构预测^[19-20];ESMFold 则证明蛋白质语言模型能够直接从序列学习结构与进化信息^[21]。第 4 类利用生物医学语言模型进行文献检索、

实体关系抽取和研究假设生成,从而辅助构建可追溯的靶标证据链。

1.3 多智能体协同与实验闭环验证

多智能体系统通过将文献检索、数据库查询、计算分析、候选排序和结果审查分配给不同智能体,提高复杂靶标研究任务的组织与证据整合效率。Drug-Agent 整合机器学习预测、知识图谱查询和文献检索智能体,可对药物-靶标相互作用的异构证据进行整合与冲突总结;其输出仍需结合原始证据和专家审核,不能替代实验验证^[22]。相关研究表明,多智能体的主要优势在于分解复杂任务和整合异构证据,其输出应被视为待验证的候选假设。在靶标发现方面, OriGene 整合疾病生物学、组学分析、药理学和文献工具,对候选靶标进行检索、推理与排序,并提出肝癌相关靶标 GPR160 和结直肠癌相关靶标 ARG2,随后在患者来源类器官和肿瘤组织模型中进行了初步验证^[23]。Co-Scientist 采用生成、反思、排序和演化等智能体迭代优化研究假设,并通过人肝类器官实验验证了其提出的肝纤维化候选靶标^[24]。Robin 则连接文献驱动的假设生成与实验数据分析,在干性年龄相关性黄斑变性研究中筛选候选治疗策略,并通过细胞实验和转录组分析更新后续假设^[25]。这些研究显示,多智能体系统正在推动靶标研究由独立计算预测转向“证

据整合、假设生成、实验验证、结果反馈”的迭代流程。然而,现有系统仍存在文献和数据库偏倚、模型幻觉、智能体误差累积及实验验证规模有限等问题。未来应加强证据来源追踪、智能体交叉核验、不确定性量化和专家监督,使多智能体系统成为可审计的靶标决策支持工具。

1.4 走向因果推理与系统干预的新趋势

AI 辅助靶标发现正在由统计关联预测转向多层次因果证据整合。未来模型将进一步融合人类遗传学、纵向多组学、单细胞扰动和临床数据,以识别疾病进程中的驱动因子,并综合评价靶标的疗效潜力和安全风险^[8-9]。研究对象也将由单一靶标扩展至通路、信号轴和靶标组合,以适应肿瘤及免疫性疾病等复杂病理网络。结构层面的研究将更多考虑蛋白质构象动态,例如, PocketMiner 可利用图神经网络从单个蛋白质结构预测可能开放的隐蔽口袋^[26]。虚拟细胞和扰动响应模型则有望用于候选靶标的计算预筛选和优先级排序,但现有模型的跨数据集泛化和扰动预测能力仍需严格评估^[27]。因此,未来发展的重点是建立可追溯的“因果证据整合、计算预测、实验验证、模型更新”闭环,提高靶标优选的可靠性与转化价值。上述方法在药物研发不同阶段的主要应用如表 1 所示。

表 1 AI 在药物研发不同阶段中的主要应用

药物研发阶段	主要任务	AI方法	代表模型	主要作用	验证方式
靶标发现	疾病关联分析	图神经网络、知识图谱、多组学模型	DTINet ^[10] 、TxGNN ^[13] 、PINNACLE ^[17]	整合疾病-基因-药物关系,发现潜在治疗靶标	文献验证、细胞实验、类器官验证
蛋白质设计	新蛋白生成、结构优化	生成模型、扩散模型	ProteinMPNN ^[28] 、RFdiffusion ^[29]	从序列和结构空间设计功能蛋白	结构预测、实验表达和功能检测
肽药设计	结合肽、抗菌肽设计	扩散模型、强化学习	DiffPepBuilder ^[30] 、AMP-Diffusion ^[31]	探索超大肽序列空间	抗菌实验、结合实验
先导物设计	分子生成、虚拟筛选	分子生成模型、图神经网络	EDM ^[32] 、SurfGen ^[33] 、DiffGui ^[34]	生成具有潜在活性的候选分子	对接、活性实验、药理评价
化学合成	路线规划、条件优化	Transformer、强化学习、智能体	ASKCOS ^[35] 、RXN ^[36] 、ChemOS ^[37]	自动规划和优化实验流程	实验产率验证
药效评价	活性预测、机制解析	多模态学习、表型模型	PhenoModel ^[38] 、Boltz-2 ^[39]	预测结合能力和细胞响应	细胞实验、动物实验
成药性评价	ADMET预测	深度学习、QSAR	ADMETlab3.0 ^[40] 、admetSAR3.0 ^[41]	提前筛选失败候选物	临床前实验
虚拟细胞	系统模拟、药物响应预测	基础模型、多智能体	scGPT ^[15] 、CellForge ^[42]	模拟细胞状态变化	实验数据一致性验证

2 先导物的发现:AI 驱动下的蛋白质设计到小分子生成新策略

先导物是药物研发流程中的关键起点,指具有明确生物活性和可优化潜力的分子骨架。其发现与优化直接决定着新药研发的效率与成功率。然而,传统的药物发现过程往往依赖经验驱动的筛选,研发周期长、成本高、成功率低,且受限于有限的化学空间和结构知识。随着人工智能、计算生物学与结构组学的快速发展,蛋白质设计、肽药设计及小分子生成等新技术正在重塑先导物的发现模式。这些方法通过数据驱动与算法创新,扩展了传统筛选策略的探索范围,使药物设计从被动探索向主动生成转变,为药物创新提供了新的策略和发展方向。

2.1 蛋白质设计在大分子药物设计中的进展

随着人工智能和计算生物学的发展,蛋白质设计为大分子药物的发现提供了全新的思路 and 工具。这一系列工作的起点主要得益于蛋白质结构预测准确性大幅度的提高,如 DeepMind 提出的 AlphaFold 系列模型^[19-20,43],通过深度学习预测蛋白质的三维结构,使研究者能够在短时间内获得数十万种蛋白质的预测结构,一方面可用于靶标结构的预测,另一方面有助于设计蛋白质的结构预测和辅助评价。AlphaFold3 已扩展到对蛋白-蛋白、蛋白-核酸(DNA/RNA)、蛋白-小分子复合物结构与相互作用的预测,为蛋白质设计与药物研发提供了更强的结构基础。同时,来自 FAIR 的研究团队则以语言模型为核心切入蛋白质生成领域,推出了 ESM 系列模型^[21],通过在数亿条蛋白序列上进行预训练,学习序列中的进化规律与隐含结构信息,并进一步发展出可直接预测结构的 ESM-Fold^[21]。这一系列工作首次证明了纯序列语言模型即可捕捉蛋白的结构与功能特征,为蛋白质生成和理解提供了新的研究框架。华盛顿大学 David Baker 团队在蛋白质设计领域做出了奠基性贡献。其开发的 Rosetta 平台^[44]自 1998 年问世以来一直是从头蛋白质设计的重要工具。近期推出的 ProteinMPNN^[28]与 RFdiffusion^[29]模型结合了深度学习与扩散生成策略,使得蛋白质设计的自动化与可扩展性取得长足进步,推动了结构生物学与药物发现的融合。当前,基于流的蛋白质生成模型也逐渐进入大众的视野,来自苹果

公司的研究团队提出首个基于流匹配的蛋白质折叠模型 SimpleFold^[45],大幅简化了架构并降低了对计算复杂网络模块的依赖。在中国,蛋白质设计也正逐渐成为一个活跃的方向,北京大学来鲁华团队设计的 GeoSeqBuilder^[46]在生成序列的同时,也给出了高精度的侧链构象,可以更直接地给出原子之间的相互作用,不需要进行单序列结构预测;西湖大学卢培龙团队通过整合深度学习和基于能量的方法,首次实现了对跨膜孔蛋白的精确从头设计^[47];中国科学技术大学刘淇团队设计的 PocketGen^[48]用于快速高效地生成高亲和力的蛋白质结合位点,极大地提升了蛋白质生成的效率和成功率,成为蛋白质结合口袋生成领域具有代表性的高效计算方法之一。

总体而言,随着 AI 驱动的蛋白质设计技术不断成熟,从蛋白质结构预测到功能导向的理性设计正在形成闭环。这一趋势不仅改善了蛋白质设计的效率,也为探索靶向“难成药”靶标的蛋白质药物设计开辟了新的可能。

2.2 肽类药物的设计和发现

肽类药物因具有较高靶标选择性、良好的生物活性以及较低的脱靶毒性,在肿瘤、感染性疾病和自身免疫疾病等领域展现出重要的应用潜力。与传统的小分子相比,肽类药物能够更好地模拟天然配体或信号通路中的关键片段,从而实现更精准的靶标识别与调控。然而,天然肽普遍存在稳定性较差、易被蛋白酶降解、生物利用度低以及构象优化困难等问题。传统的肽设计主要通过结构优化以提升肽药物的稳定性和药代动力学特性,不仅周期长、成本高,而且难以高效探索庞大的肽序列空间。目前,AI 驱动的肽设计已经逐渐成为新趋势,为不同类型肽药物的研发提供了新的技术路径。

在结合肽设计方面,北京大学来鲁华团队开发的 DiffPepBuilder^[30]采用扩散生成模型(diffusion model)结合蛋白-肽复合物结构约束,能够在已知作用界面的情况下同步生成肽骨架构象和氨基酸序列,并进一步通过二硫键设计提高多肽构象稳定性。实验结果表明,该方法设计的多肽能够保持稳定的结合构象,并在多个测试体系中表现出良好的结合能力,为基于结构的结合肽设计提供了新的解决方案。

针对环肽设计中构象搜索空间巨大和环化效率

低等问题,该团队进一步提出了基于强化学习的环肽设计框架 CYC_BUILDER^[49],该方法利用强化学习优化片段组装过程,并实现头尾酰胺键或二硫键环化,从而高效生成具有稳定构象的候选环肽。此外,团队开发 CycDockAssem^[50]结合分子动力学模拟、片段对接和结构组装策略,为环肽药物发现提供了新的策略。这两种方法在 TNF α 环肽抑制剂设计中均获得了具有较高结合活性的候选分子,验证了 AI 辅助环肽设计的有效性。

抗菌肽作为近年肽类药物研究的热点之一,也成为 AI 驱动分子设计的重要方向。宾夕法尼亚大学 César de la Fuente-Núñez 团队是当前国际上最活跃、最具代表性的人工智能驱动的抗菌肽设计与发现的团队之一。该团队开发的 AMP-Diffusion^[31]基于扩散模型学习天然抗菌肽序列分布,通过系统地探索序列空间来快速发现抗菌肽候选物,实验验证表明,多个生成肽不仅具有针对多重耐药菌株的广谱抗菌活性,而且保持较低的细胞毒性。随后,该团队联合复旦大学 Luis Pedro Coelho 团队建立了全球微生物组抗菌肽资源平台 AMPSphere^[51],利用机器学习预测全球微生物组中的潜在抗菌肽,并对 100 条合成肽进行了 11 种临床相关致病菌株和人类肠道共生菌的作用实验验证,其中 79 条表现出针对临床病原菌或肠道共生菌的显著抗菌活性,显著提高了天然抗菌肽资源的发掘效率。浙江大学侯廷军团队开发的 AMP-Designer^[52]通过巧妙融合大语言模型、提示学习与知识蒸馏等技术,仅耗时 11 d 便设计出 18 种具有广谱抗菌活性的候选肽,并在实验中验证了其抑菌效果,展示了生成式 AI 在抗菌肽快速发现与优化中的应用前景。

除直接生成功能肽外,AI 还广泛应用于免疫相关肽预测。华东师范大学李洪林团队开发的 PISTE^[53]基于 Transformer 架构,融合物理和生物学的先验知识,引入滑动窗口式注意力机制与位置感知编码,捕捉 T 细胞受体-抗原-人类白细胞抗原分子之间的跨尺度相互作用特征,以筛选具有免疫原性的新抗原。该模型在多个肿瘤新抗原数据集上均表现出优于现有方法的预测性能,并成功筛选出具有较高免疫原性的候选新抗原,为个体化肿瘤免疫治疗及免疫肽药物设计提供了新的计算工具。

AI 驱动的肽药设计已逐步形成“结构预测—序列生成—活性预测—实验验证”的完整技术链。扩散模型增强了结合肽和抗菌肽的生成能力,强化学习推动了环肽设计自动化,大语言模型进一步增强了少样本条件下的肽序列优化能力,而融合结构建模与生物学先验知识的方法则提高了免疫相关肽筛选的准确性。随着 AlphaFold 等蛋白质结构预测技术、生成式 AI、多模态学习以及高通量实验平台的不断发展,未来有望实现肽药物从序列设计、结构优化到功能验证的闭环研发流程,为复杂疾病靶标的精准干预和新型肽类药物开发提供更加高效的技术支撑。

2.3 小分子先导化合物的生成

药物分子的化学空间规模巨大且高度多样化,已知的药物空间仅覆盖其中有限范围,这使得当前的药物研发很大程度上仍局限于已知化合物空间的局部探索,难以逃脱经验主导的筛选思维,从而限制了结构与功能层面的创新。在过去几十年中,小分子先导化合物的发现主要依赖高通量筛选和虚拟筛选等方法,这些方法仍受限于预设的化学结构模板,难以产生真正新颖的分子结构。随着人工智能的兴起,以算法驱动的方式能够在广阔的化学空间中高效地搜索潜在的先导化合物,为小分子药物发现提供了新的技术路径。

浙江大学侯廷军团队尝试从加速虚拟筛选速率入手,构建了 KarmaDock、CarsiDock、RTMScore 等分子对接和打分算法,有效提升了虚拟筛选的准确率与效率^[54-57]。同时,该团队还基于物理与深度学习相结合的理念,开发了一系列小分子生成模型如 MCMG、SurfGen、ResGen、Delete 等^[33,58-60],并提出结合强化学习与提示词机制的 Prompt-MolOpt 框架^[61],实现了条件可控的分子优化。阿姆斯特丹大学 Max Welling 团队提出的等变扩散模型 EDM^[32]在三维分子生成中引入 E(3)等变约束,以提高几何表示的一致性;生成结构的化学合理性和稳定性仍需结合能量计算及实验进行验证。中国科学院上海药物研究所郑明月团队提出的 SurfDock^[62]将蛋白质表面信息与扩散生成框架相结合,能够在三维空间中直接生成符合口袋几何约束的配体结合构象,实现了蛋白-配体复合物的精准预测与生成,为生成式药物设计提供了新的范式。华东师范大学李洪林团队近期提出的 DiffGui^[34],

通过原子-化学键协同扩散和药物属性引导的双重机制,大幅提高生成分子的结构合理性与类药性。四川大学杨胜勇团队设计的 PocketFlow^[63]能够在靶蛋白活性口袋内直接生成全新结构配体分子,为快速发现药物先导化合物提供了重要工具。除生成式策略外,基于配体的药物发现仍是先导化合物筛选的重要方向。上海交通大学张健团队发展了一种结合电子密度信息的深度学习方法 ED2Mol,用于提升先导分子的从头生成与优化,并且面向 4 个重大疾病靶标设计并优化出多类活性分子,不仅在常规正构位点上表现出色,在一直难以突破的变构位点分子生成上也获得突破,发现了多类抑制剂与激动剂^[64]。上海科技大学白芳团队提出的 GeminiMol^[65]通过比较分子间的相似性,实现了对与参考配体性质接近分子的快速筛

选;其后续的 PhenoModel^[38]则进一步引入表型信息,以提升筛选的生物学相关性与预测精度。需要指出的是,该团队还将研究拓展至新型降解类药物 PROTAC,提出的 DiffPROTACs^[66]能够基于两端配体生成全新的 linker,并结合降解率预测模型 Deep-PROTACs^[67]对生成分子进行筛选,从而发现具有优异降解效率的候选化合物。

总体而言,AI 驱动的小分子生成与筛选正推动药物设计从经验驱动向数据驱动、从线性探索向迭代优化转变。随着生成模型与物理建模、分子动力学及自动合成平台的结合,小分子先导化合物的发现正迈向端到端的自动化与智能化新阶段,涉及的代表性模型、主要应用及局限性汇总如表 2 所示。

表 2 AI 药物研发流程中的典型模型及应用场景

代表模型	模型类型	主要任务	代表优势	主要局限
DTINet ^[10]	异构网络学习	DTI预测	多源网络融合	偏关联预测
TxGNN ^[13]	KG+GNN	药物重定位	Zero-shot预测	依赖知识图谱
PINNACLE ^[17]	图神经网络	靶标优选	细胞类型特异性表示	依赖PPI质量
scGPT ^[15]	单细胞基础模型	细胞状态建模	大规模预训练	跨平台校正需求
GET ^[18]	Transformer	转录调控预测	多组学整合	依赖ATAC数据
AlphaFold3 ^[20]	深度学习	多组分结构预测	支持蛋白-核酸-配体复合物	动态构象不足
RFdiffusion ^[29]	扩散模型	从头蛋白设计	新蛋白生成能力强	实验验证需求
ProteinMPNN ^[28]	深度学习	蛋白序列设计	快速稳定设计	功能优化有限
DiffPepBuilder ^[30]	扩散模型	结合肽设计	同时设计序列和构象	需已知结合界面
AMP-Diffusion ^[31]	扩散模型	抗菌肽生成	快速探索序列空间	活性需验证
PocketFlow ^[63]	扩散模型	Pocket内生成	靶标导向设计	合成性有限
EDM ^[32]	等变扩散	三维分子生成	保持三维对称性	计算成本较高
Prompt-MolOpt ^[61]	LLM+RL	分子优化	条件可控优化	对Prompt敏感
RXN for Chemistry ^[36]	Transformer	反应预测	高精度预测	依赖历史数据
ASKCOS ^[35]	AI规划	合成路线设计	自动逆合成	路线仍需审核
ChemCrow ^[68]	LLM智能体	自动实验	多工具协同	幻觉风险
Boltz-2 ^[39]	基础模型	结合亲和力预测	接近FEP精度	需独立和前瞻性验证
PhenoModel ^[38]	多模态模型	表型药效预测	融合细胞形态表型信息	依赖表型数据
ADMETlab3.0 ^[40]	AI平台	ADMET预测	多端点预测	部分端点精度有限
CellForge ^[42]	多智能体	虚拟细胞	扰动模拟与数字药筛	数据需求高

3 AI 驱动的化学合成:从自动化实验到自主科研的范式跃迁

化学合成作为创造新物质、发现新药和开发新材料的核心手段,正在经历一场由 AI 驱动的深刻变

革^[69]。在过去 10 年中,自动化技术已在化学实验室中得到广泛应用,使化学家能够通过机械化系统完成物料输送、温控及分析等操作。然而,这类系统主要解决的是执行自动化问题,即让设备按预设程序运

行。要实现真正意义上的智能化合成,则需使系统具备自主决策与自我优化能力,能够在未知化学空间中主动探索最优实验条件、提出新假设,甚至发现新知识,从而形成“AI 规划实验—机器人执行—分析反馈—AI 再决策”的闭环体系。这一转变标志着化学实验正从自动化执行迈向智能化科研阶段^[70-71]。

3.1 AI 在化学合成中的分层应用与系统智能发展

当前,AI 在化学合成中的应用已形成清晰的层次结构,主要集中在:反应预测与路线规划、条件优化与实验设计,以及自动化实验系统的智能控制 3 大核心模块,其价值不仅体现在单一环节的性能提升,更在于多层次协同融合后所展现的系统智能。

在反应预测与路线规划方面,深度学习模型,尤其是基于 Transformer 架构与 GNN 的算法,能够在大规模反应数据上学习化学转化规律,从而实现从产物出发预测可行的合成路径或高精度预测复杂反应的产物。基于 AI 的合成预测软件已被广泛开发与应用。例如,IBM 的 RXN for Chemistry 平台基于神经机器翻译和 Molecular Transformer 开展反应产物预测^[36,72];MIT 开发的 ASKCOS 软件套件则可将复杂目标分子逐步拆解为可购买的基础原料,为化学家提供多条可行的合成路线参考。此外,该系统还可正向预测反应产物,并根据设定反应预测最适宜的试剂、催化剂、溶剂及温度等条件^[35]。部分平台已实现商业化,如 ChemAI、SYNTHIA、Reaxys 数据库的 Predictive Retrosynthesis 等。这一方向的研究通过突破人类基于已知反应类型的思维定式和知识的局限,能在复杂反应网络中挖掘潜在的结构—反应模式关系,进一步拓宽了可探索的反应空间,提高了复杂反应路线设计能力,为有机合成提供新的思路。

在反应优化与实验设计层面,AI 正告别传统依赖经验或穷举筛选的实验优化模式。通过贝叶斯优化、强化学习等算法,研究人员能够在有限实验次数内探索多维参数空间(如温度、溶剂、催化剂、摩尔比等),逐步逼近最优反应结果。以 ChemOS、Deep Reaction Optimizer(DRO)等系统为代表的闭环优化实验实现了“模型预测—实验验证—模型更新”的迭代循环,有效提高实验效率,并在较少实验次数下获得更优反应结果^[37,73-75]。这种“实验—模型共进化”模式,正推动反应优化从经验驱动向数据驱动转变,可同时权衡产

率、成本、步骤经济性、安全性与环境影响,使合成方案更符合实际工业和应用需求。

在自动化实验系统的智能控制层面,AI 驱动的合成系统逐步与自动化平台深度耦合,形成从实验规划到物理执行的闭环体系。典型案例包括 Coscientist、LLM-RDF 与 ChemCrow 等平台^[68,76-79]。这些系统可支持实验方案设计和流程规划,其中部分系统能够连接自动化硬件,并根据传感器反馈调整实验流程。这类自主化实验系统被视为通向未来自驱动实验室(self-driving laboratory, SDL)的重要雏形^[80]。

3.2 AI 驱动化学合成的挑战与未来发展趋势

尽管 AI 驱动的合成技术发展迅速,但其全面落地与自主化仍面临多重挑战:数据层面上,现有反应数据库普遍存在数据稀疏、不平衡以及实验条件记录不完整等问题,严重影响模型的泛化能力与预测可靠性;在模型层面上,当前深度学习方法缺乏对反应机理与物理约束的嵌入,导致模型在化学逻辑上的可解释性不足,限制了其在高风险科研场景中的应用;系统层面上,硬件接口与控制协议尚未实现统一标准,不同厂商的自动化实验平台之间兼容性有限,从而阻碍了基于 AI 合成系统在更大范围内的通用部署^[81]。此外,AI 驱动的合成研究流程仍离不开化学家的知识监督与目标设定,尚未实现真正意义上的科研自驱化。

展望未来,AI 驱动的合成发展将呈现出多维融合与自主演化的趋势。其一,大模型与化学知识图谱的结合将增强系统的知识理解与推理能力,实现从数据驱动向知识驱动的范式转变;其二,多智能体协同体系将成为关键架构,不同智能体分别承担规划、执行、分析与优化等任务,通过通信与反馈形成自组织科研网络;其三,SDL 建设将成为化学研究的重要方向,AI 驱动系统有望根据科研目标自动生成假设、设计实验并验证结果,形成闭环科学发现流程,转变为新化学知识的共同创造者。从宏观层面看,AI 驱动的化学合成不仅重塑了化学研究的技术体系,也将对科研组织模式与科技创新生态产生深远影响。它代表着科研自动化从工具化向智能化的转变,预示着以 AI 为核心的认知科研基础设施的形成。这一变革不仅关乎研究效率的提升,更关乎在未来科技竞争中能否占据源头创新的制高点。

4 候选物评价: 基于 AI 的药效推断与成药性评估方法

先导化合物药效评价旨在把早期具有理想活性的先导化合物向候选药物推进,其关键在于能否在分子层面命中并明确靶标与作用机制、能否在细胞或更高生物体系中产生期望表型,以及在体内是否具有可接受的 ADMET 特性。

4.1 基于靶标的药效预测

基于靶标的药效预测旨在评估化合物能否与预期靶标发生特异、稳定的相互作用并触发可解释的分子机制。常以分子对接作为高通量初筛手段,并通过自由能打分来评估结合的稳定性与关键相互作用残基。这依赖于具有明确的靶标结构,而 AlphaFold2^[19]、ESMFold^[21]等蛋白结构预测方法推动了许多靶标结构的可获得性,从而扩大了结构基础筛选的适用范围,但对配体结合精细构象的预测仍需与实验验证联合使用。

此外,也可通过深度学习模型进行亲和力预测来进行药效评价。中国科学院上海药物研究所郑明月研究员团队开发了 TransformerCPI 系列^[82]、PertKGE^[83]等模型,用于化合物-蛋白质相互作用预测。中南大学联合赫尔辛基大学开发的基于多任务学习模型 DeepDTAGen 可用于药物-靶标亲和力预测并探索靶标感知分子生成^[84]。上海科技大学白芳教授团队开发的 TEFDTA 模型^[85]通过将 Transformer 编码器和指纹表征相结合,用于预测键合和非键合药物-靶标相互作用的结合亲和力。由 MIT Jameel Clinic 与 Recursion 开发的 Boltz-2 在结合亲和力预测方面实现了最先进的性能,其预测的结合亲和力接近自由能扰动方法性能^[39]。

值得注意的是,仅有 9.4% 获批的小分子药物是通过基于靶标发现的^[86],并且即使这些靶向药物,其治疗效果也可能并非完全由预期靶标介导。因此,仅通过基于靶标的药效预测只能作为初步筛选,计算预测得到的靶点结合信息并不一定能作为发挥药效的直接证据。

4.2 基于表型的药效预测

不同于靶标导向的药效预测,基于表型的药效预测不预设有关分子靶标的先验知识,而是直接评估化

合物在细胞或更复杂模型中是否能引发期望的生物学表型,例如,细胞形态变化、疾病状态逆转或特定信号通路调节。这一策略尤其适用于多靶标药物、复杂疾病(如肿瘤、神经退行性疾病)及未知作用机制的新化学分子的药效筛选。中国科学院上海药物研究所郑明月研究员团队开发了 TranSiGen 分子表征模型,通过将分子结构及转录组信息联系起来用于化合物反应预测^[87]。

Cell Painting 是最具代表性的表型筛选技术之一。该方法通过多通道荧光染料标记亚细胞结构,并结合高内涵成像与自动化特征提取,生成细胞形态特征向量。通过与已知机制或药理特征的形态学特征对比,能推测未知化合物的作用机制、潜在毒性类型及药效相关性^[88]。国际上, JUMP-Cell Painting Consortium 发布了多个 Cell Painting 数据集,包含大量的细胞显微镜图像和组学分析,涵盖数十万种化合物以及基因敲除和过表达的扰动,为探索强大的深度学习框架以开发基于表型的药物筛选甚至生成方法提供了丰富的训练样本。上海科技大学白芳团队通过设计双空间对比学习框架,将分子结构与表型信息联系起来,开发了多模态表征模型 PhenoModel^[38],成功鉴定几种针对骨肉瘤和横纹肌肉瘤细胞系的表型生物活性化合物^[38]。中山大学杨跃东团队开发的 MIGA 模型将细胞表型纳入模型学习中,在预测生物活性和作用机制方面表现十分出色^[89]。

除了真实细胞的高内涵成像外,虚拟细胞模型正成为新兴方向。虚拟细胞模型是一个学习型的模拟器,可以模拟细胞和细胞系统在不同条件和变化环境下的运作,例如分化状态、扰动、疾病状态及环境变化等条件下的响应^[90]。CellForge^[42]是一种面向单细胞数据集和扰动预测任务的多智能体建模框架,可协同设计神经网络架构并生成可执行实现;其药效筛选和临床价值仍需验证。

4.3 ADMET 与成药性评价

先导化合物的药效评价不仅要关注其靶标活性与表型效应,更需评估其在体内的 ADMET。在先导化合物进入动物或人体之前,ADMET 特性是衡量其是否具备良好成药性的关键门槛,是从药理学活性到临床可用性之间的关键桥梁,决定了先导化合物能否成为具有开发潜力的候选药物。尽管一个化合物在

分子或表型层面表现优异,但若其 ADMET 性能欠佳,则极可能在临床前或临床阶段失败。

国内华东理工大学唐赟教授团队发布了 admet SAR 系列,从 2012 年就推出了 admetSAR,到 admet SAR2.0,再到 2024 年正式发布的 admetSAR3.0^[41],目前收录超过 10 万个化合物、37 万余条实验数据,支持 ADMET 性质搜索、预测和优化 3 大功能。最近,该团队开发了一个新的综合成药性预测模型 CLaSP,通过整合 ADMET 性质、理化性质与可合成性评价为先导化合物筛选及优化提供可量化、可解释的决策参考^[91];中南大学曹东升教授团队开发了 ADMETlab 系列,其最新发布的 ADMETlab3.0^[40],可预测 119 个 ADMET 端点,覆盖超过 40 万个条目。在国际上,昆士兰大学 David B Ascher 团队开发了 Deep-PK^[92]用于分子的药代动力学和毒性预测;柏林夏利特医学院 Robert Preissner 团队专注于化合物的毒性预测,其开发的 ProTox 毒性预测系列平台(ProTox 3.0)能预测 61 个毒性终点,如急性毒性、器官毒性、临床毒性等^[93];瑞士洛桑大学 Vincent Zoete 团队开发 Swiss ADME 模型^[94]用于小分子药代动力学预测。因此,应尽早并行开展关键 ADMET 预测,从而在早期筛选中淘汰不可开发的化合物,降低后期失败风险。

综上,国内外在基于 AI 对药物的药效与 ADMET 预测方面取得了显著进展。AI 方法已从早期的 QSAR 模型发展为深度学习、图神经网络、自动化机器学习等复杂技术体系,以提升对化合物在体内行为与安全性的预测能力。国内研究也逐渐加速,更多聚焦于多端点联合预测、结构-活性-毒性一体化建模。展望未来,未来趋势包括:(1)更多多模态数据(如组学、生物标志物、临床真实世界数据)与化学结构融合;(2)提升模型可解释性与透明度,便于监管使用;(3)加强针对新化学结构(化学空间外)模型的泛化能力;(4)推动 AI 预测工具与实验/临床验证体系更紧密结合,从而从预测迈向设计与创造。

5 AI 虚拟细胞:引领药学研究迈向智能化的新纪元

作为生命的微观引擎,细胞内部持续进行着复杂的物质交换、能量流动与信号传递,这些过程共同维

系着个体的稳态平衡,并塑造了从生长发育到疾病衰老的生命全景^[95]。然而,要全面理解并准确预测细胞在不同内外扰动下的动态响应,仍是现代生物学期长期未解的核心难题。传统研究主要依赖实验观察与基于规则的数学建模,前者受限于通量与成本,后者则难以刻画细胞系统的高维复杂性与涌现行为。随着生物医学数据的爆发式增长与人工智能技术的快速演进,药学研究正迎来范式性的变革。人工智能与多组学数据的深度融合,正在催生一种面向系统层次与机制推理的新型研究框架——人工智能虚拟细胞(artificial intelligence virtual cell, AIVC)^[90,96],它以数据驱动的建模与智能化推演为核心,旨在实现对细胞行为的可解释预测与可验证模拟,为未来的精准医学与药物研发提供新的研究路径。

5.1 从机理建模到数据驱动的虚拟细胞技术演进

虚拟细胞概念可追溯到十余年前。早期模型主要尝试将已有的生物学知识提炼为一系列数学方程与算法规则,以重构特定细胞(如生殖支原体)的生命周期过程^[97]。这类基于显式知识的模型为理解细胞功能机制提供了重要启示,但其高度依赖先验假设,难以扩展至复杂、多层次的细胞体系。新一代的 AIVC 则实现了根本性的转变。它不再是人类知识的简单编码,而是一个基于多尺度、多模态大型神经网络的数据驱动模型。其核心能力在于直接从海量的组学数据(如基因组、转录组、蛋白质组等)中学习细胞的内在规律。通过整合来自不同物种、细胞类型乃至组织层面的数据, AIVC 致力于构建一个能够跨越不同生物学背景的通用细胞状态表示。这使得模型不仅能理解已知的细胞状态,更能泛化和预测在全新扰动(如药物处理或基因编辑)下,细胞可能进入的前所未见的新状态。近期涌现的 scGPT^[15]、Geneformer^[98]等基础模型,以及能够自动化设计建模流程的 Cell-Forge^[42]多智能体系统,正是这一技术路径上的标志性进展。

5.2 虚拟细胞的应用潜力及其对新药研发的影响

AIVC 的崛起将对多个科研领域产生深远影响,尤其是在新药研发、疾病机理研究以及个性化医疗等方面。

在药物研发领域, AIVC 有望彻底改变广撒网式的传统药物筛选模式。传统研发流程中,大量化合物

需要在细胞和动物模型中进行反复测试,周期长、成本高,且动物模型与人体的差异导致临床转化成功率低。现有模型可在特定数据集上预测转录或细胞状态响应^[99],但对疗效、毒性和临床结局的外推仍需独立验证。例如,通过模拟肿瘤细胞对不同药物的响应,研究者可快速发现潜在的耐药机制或联合用药方案。

在疾病机理研究中,AIVC 为探索复杂的生命过程提供了前所未有的计算分析工具。已有研究构建了类风湿关节炎滑膜关节的模块化多细胞虚拟孪生模型^[100]。通过在虚拟细胞中引入特定基因突变或模拟病理微环境,研究人员可分析细胞状态的动态变化并提出可供实验检验的机制假设,从而辅助研究癌症、神经退行性疾病等复杂疾病的发生发展过程。

最后,个性化医疗将是 AIVC 最具变革性的应用场景之一。未来的临床实践中,医生有望为每位患者构建专属的“数字孪生细胞”^[101]。通过将患者的个人组学数据输入模型,可在虚拟环境中比较不同治疗方案的模拟结果,为个体化治疗研究提供参考^[102]。美国食品药品监督管理局(FDA)于 2025 年提出在单克隆抗体及部分其他药物的特定开发场景中减少、优化或在适用情况下替代部分动物试验^[103],这为新方法学评价提供了监管背景,但不意味着 AIVC 已获得临床监管采信。

5.3 实现虚拟细胞的技术瓶颈与治理挑战

尽管 AIVC 的前景令人振奋,但要真正实现其全部潜力,仍需克服一系列重大的科学技术与伦理挑战。

在数据方面,AIVC 的构建高度依赖于大规模、高质量、标准化的多模态生物数据^[90]。然而,目前全球范围内的生物数据仍存在格式不一、分布零散的问题,亟需构建开放共享的数据基座。此外,为提高模型预测的准确性,还需积累更多反映细胞在不同扰动条件下动态响应的干预性数据,而这类数据正是当前最为稀缺的资源。在模型可解释性与验证方面,深度学习常被称为“黑箱模型”,其决策过程缺乏透明度,这在对可靠性和安全性要求极高的生物医学领域尤为突出^[104]。如何确保 AIVC 的预测结果不仅准确,而且其背后的生物学逻辑可被合理解释,是当前亟待解决的问题。同时,建立统一的模型评估与验证标准,使虚拟预测结果能够与湿实验数据高度一致,是推动

该领域走向成熟的必由之路;在伦理与监管方面,海量生物医学数据的使用涉及患者隐私和数据安全等敏感问题^[105]。如何在推动科研共享的同时,建立健全的数据管理和伦理规范,是所有从业者必须审慎对待的议题。

展望未来,AIVC 未来需要建立与真实实验数据相比较的统一评价基准,“虚拟细胞图灵测试”可作为一种研究设想。长远来看,单个虚拟细胞的成功构建将为虚拟组织、虚拟器官乃至虚拟人体的实现奠定基础。这不仅标志着药学研究范式的根本性变革,也为理解复杂生命系统提供长期研究方向。

6 AI 药物研发方法的局限性与未来挑战

尽管人工智能提升了药物靶标发现、分子设计、药物重定位等多个环节的研发效率,但目前 AI 在药物研发中仍面临诸多技术瓶颈。首先,高质量数据不足仍是限制 AI 性能的核心因素。AI 模型高度依赖大规模、高质量训练数据,而药物研发数据普遍存在样本量有限、实验条件不统一、数据缺失、负样本不足以及数据共享受限等问题,尤其在罕见疾病和新靶标研究中更加突出。同时,目前多数深度学习模型仍属于“黑箱模型”,虽然预测性能较高,但难以解释其决策依据,限制了研究人员对预测结果的生物学理解,也影响了模型在临床和监管中的可信度。可解释人工智能(explainable AI, XAI)及知识图谱推理等方法能够一定程度提升模型透明度,但尚未形成统一的解释评价标准。此外,复杂生物系统的多尺度特性使现有 AI 模型仍难以全面刻画疾病发生发展的动态过程。目前多数模型主要基于单一数据模态开展学习,对基因组学、转录组学、蛋白质组学、临床数据及医学知识等多源异构信息的融合能力仍有限。近年来生成式 AI 能够快速设计大量候选化合物,但生成结果往往更多关注结合活性,而对药物可合成性、ADMET 性质及成药性等多目标优化考虑不足;同时,模型在超出训练数据分布的新疾病、新靶标上的泛化能力仍有限,大多数预测结果仍需依赖体外实验和动物实验进一步验证。目前 AI 更多发挥辅助决策作用,距离实现完全自主的新药研发仍存在较大差距。

与此同时, AI 药物研发还面临数据隐私保护、知

识产权、算法偏倚、模型安全及监管规范等问题。随着大语言模型和智能体(agent AI)逐渐应用于药物研发,模型幻觉、错误推理及自动化实验决策等潜在风险也受到广泛关注。因此,未来的发展不仅需要进一步提升模型预测性能,更应加强多模态数据融合、知识图谱与物理机制结合、可解释人工智能、联邦学习、人机协同及标准化评价体系建设,并建立更加完善的实验验证与监管框架,以提高 AI 药物研发的可靠性、可解释性及临床转化能力。

7 结论

人工智能正以前所未有的深度与广度重塑药物研发,从靶标发现、先导物设计、化学合成到药效与成药性评价,再到虚拟细胞构建,形成贯穿全流程的范式转变。它不仅能够整合海量、多模态生物数据,从局部预测走向系统化、因果化推理,还推动药物研发各阶段从经验探索向自进化、自学习、自推理的智能药物创新体系转型。然而,这一转型也伴随挑战,包括数据标准化与共享、模型可解释性、伦理与监管等问题。未来的发展趋势将是多智能体系统、知识驱动建模、虚拟实验与物理实验平台的深度融合,实现从单一工具到自主发现的全流程闭环操作平台。这种数字-物理协同模式不仅提升药物研发效率和成功率,也为复杂疾病干预和精准医疗提供了可量化、可迭代的智能化路径,标志着药物研发正迈向真正的自主化与系统化新时代。

参考文献 (References)

- [1] Albani F G, Alghamdi S S, Almutairi M M, et al. Artificial intelligence-driven innovations in oncology drug discovery: Transforming traditional pipelines and enhancing drug design[J]. *Drug Design, Development and Therapy*, 2025, 19: 5685–5707.
- [2] Zemnou C T, Kamsu G T, Ngakam R, et al. The application of AI in drug discovery and early development: Impact and challenges[J/OL]. Preprints, 2026: 2026012255. <https://www.preprints.org/manuscript/202601.2255/v1>.
- [3] Zhang K, Yang X, Wang Y F, et al. Artificial intelligence in drug development[J]. *Nature Medicine*, 2025, 31(1): 45–59.
- [4] Ferreira F J N, Carneiro A S. AI-driven drug discovery: A comprehensive review[J]. *ACS Omega*, 2025, 10(23): 23889–23903.
- [5] Dermawan D, Alotaib N. From lab to clinic: How artificial intelligence (AI) is reshaping drug discovery timelines and industry outcomes[J]. *Pharmaceutics*, 2025, 18(7): 981.
- [6] Trajanoska K, Bh  rer C, Taliun D, et al. From target discovery to clinical drug development with human genetics[J]. *Nature*, 2023, 620(7975): 737–745.
- [7] Ochoa D, Hercules A, Carmona M, et al. The next-generation Open Targets Platform: Reimagined, redesigned, rebuilt [J]. *Nucleic Acids Research*, 2023, 51(D1): D1353–D1359.
- [8] Minikel E V, Painter J L, Dong C C, et al. Refining the impact of genetic evidence on clinical success[J]. *Nature*, 2024, 629(8012): 624–629.
- [9] Razuvayevskaya O, Lopez I, Dunham I, et al. Genetic factors associated with reasons for clinical trial stoppage[J]. *Nature Genetics*, 2024, 56(9): 1862–1867.
- [10] Luo Y N, Zhao X B, Zhou J T, et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information[J]. *Nature Communications*, 2017, 8: 573.
- [11] Zhang C C, Zang T Y, Zhao T Y. KGE-UNIT: Toward the unification of molecular interactions prediction based on knowledge graph and multi-task learning on drug discovery[J]. *Briefings in Bioinformatics*, 2024, 25(2): bbae043.
- [12] Ye Q, Hsieh C Y, Yang Z Y, et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system[J]. *Nature Communications*, 2021, 12: 6775.
- [13] Huang K X, Chandak P, Wang Q W, et al. A foundation model for clinician-centered drug repurposing[J]. *Nature Medicine*, 2024, 30(12): 3601–3613.
- [14] Sudhakar S, Ozer B, Chang J K, et al. An experimentally validated approach to automated biological evidence generation in drug discovery using knowledge graphs[J]. *Nature Communications*, 2024, 15: 5703.
- [15] Cui H T, Wang C, Maan H, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI[J]. *Nature Methods*, 2024, 21(8): 1470–1480.
- [16] Hao M S, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics[J]. *Nature Methods*, 2024, 21(8): 1481–1491.
- [17] Li M M, Huang Y P, Sumathipala M, et al. Contextual AI models for single-cell protein biology[J]. *Nature Methods*, 2024, 21(8): 1546–1557.
- [18] Fu X, Mo S T, Buendia A, et al. A foundation model of transcription across human cell types[J]. *Nature*, 2025, 637(8047): 965–973.
- [19] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021,

- 596(7873): 583–589.
- [20] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. *Nature*, 2024, 630(8016): 493–500.
- [21] Lin Z M, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. *Science*, 2023, 379(6637): 1123–1130.
- [22] Inoue Y, Song T C, Wang X L, et al. DrugAgent: Reliable multi-agent integration of conflicting biomedical evidence for drug–target interaction assessment[J/OL]. arXiv, 2024: 2408.13378. <https://arxiv.org/abs/2408.13378>.
- [23] Zhang Z, Qiu Z, Wu Y, et al. OriGene: A self-evolving virtual disease biologist automating therapeutic target discovery[J/OL]. bioRxiv, 2025: 2025.06. 03.657658. <https://www.biorxiv.org/content/10.1101/2025.06.03.657658v1>.
- [24] Gottweis J, Weng W H, Daryin A, et al. Accelerating scientific discovery with co-scientist[J]. *Nature*, 2026, 655(8122): 487–496.
- [25] Ghareeb A E, Chang B, Mitchener L, et al. A multi-agent system for automating scientific discovery[J]. *Nature*, 2026, 655(8122): 497–505.
- [26] Meller A, Ward M, Borowsky J, et al. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network[J]. *Nature Communications*, 2023, 14: 1177.
- [27] Ahlmann-Eltze C, Huber W, Anders S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines[J]. *Nature Methods*, 2025, 22(8): 1657–1661.
- [28] Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. *Science*, 2022, 378(6615): 49–56.
- [29] Watson J L, Juergens D, Bennett N R, et al. *De novo* design of protein structure and function with RFdiffusion[J]. *Nature*, 2023, 620(7976): 1089–1100.
- [30] Wang F H, Wang Y Z, Feng L Y, et al. Target-specific *de novo* peptide binder design with DiffPepBuilder[J/OL]. arXiv, 2024: 2405.00128. <https://arxiv.org/abs/2405.00128>.
- [31] Torres M D T, Chen L T, Wan F P, et al. Generative latent diffusion language modeling yields anti-infective synthetic peptides[J]. *Cell Biomaterials*, 2025, 1(9): 100183.
- [32] Hooeboom E, Satorras V G, Vignac C, et al. Equivariant diffusion for molecule generation in 3D[C/OL]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 8867–8887.
- [33] Zhang O, Wang T Y, Weng G Q, et al. Learning on topological surface and geometric structure for 3D molecular generation[J]. *Nature Computational Science*, 2023, 3(10): 849–859.
- [34] Hu Q Y, Sun C Z, He H, et al. Target-aware 3D molecular generation based on guided equivariant diffusion[J]. *Nature Communications*, 2025, 16: 7928.
- [35] Tu Z K, Choure S J, Fong M H, et al. ASKCOS: Open-source, data-driven synthesis planning[J]. *Accounts of Chemical Research*, 2025, 58(11): 1764–1775.
- [36] Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction[J]. *ACS Central Science*, 2019, 5(9): 1572–1583.
- [37] Sim M, Vakili M G, Strieth-Kalthoff F, et al. ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories[J]. *Matter*, 2024, 7(9): 2959–2977.
- [38] Wang S H, Han Q L, Qin W C, et al. PhenoModel: A multi-modal phenotypic drug design foundation model for discovering novel potential inhibitors of multiple cancer cells[J]. *Acta Pharmaceutica Sinica B*, 2026, 16(3): 1219–1232.
- [39] Passaro S, Corso G, Wohlwend J, et al. Boltz-2: Towards accurate and efficient binding affinity prediction[J/OL]. bioRxiv, 2025: 2025.06. 14.659707. <https://www.biorxiv.org/content/10.1101/2025.06.14.659707v1>.
- [40] Fu L, Shi S H, Yi J C, et al. ADMETlab 3.0: An updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support[J]. *Nucleic Acids Research*, 2024, 52(W1): W422–W431.
- [41] Gu Y X, Yu Z H, Wang Y M, et al. admetSAR3.0: A comprehensive platform for exploration, prediction and optimization of chemical ADMET properties[J]. *Nucleic Acids Research*, 2024, 52(W1): W432–W438.
- [42] Tang X R, Yu Z Y, Chen J P, et al. CellForge: Agentic design of virtual cell models[J/OL]. arXiv, 2025: 2508.02276. <https://arxiv.org/abs/2508.02276>.
- [43] Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning[J]. *Nature*, 2020, 577(7792): 706–710.
- [44] Simons K T, Kooperberg C, Huang E, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions[J]. *Journal of Molecular Biology*, 1997, 268(1): 209–225.
- [45] Wang Y Y, Lu J R, Jaitly N, et al. SimpleFold: Folding proteins is simpler than you think[J/OL]. arXiv, 2025: 2509.18480. <https://arxiv.org/abs/2509.18480>.
- [46] Liu J L, Guo Z, You H T, et al. All-atom protein sequence design based on geometric deep learning[J]. *Angewandte Chemie International Edition*, 2024, 63(50): e202411461.
- [47] Zhu J Y, Liang M F, Sun K, et al. *De novo* design of transmembrane fluorescence-activating proteins[J]. *Nature*, 2025,

- 640(8057): 249–257.
- [48] Zhang Z X, Shen W X, Liu Q, et al. Efficient generation of protein pockets with PocketGen[J]. *Nature Machine Intelligence*, 2024, 6(11): 1382–1395.
- [49] Wang F H, Zhang T T, Zhu J T, et al. Reinforcement Learning-based target-specific *de novo* design of cyclic peptide binders[J]. *Journal of Medicinal Chemistry*, 2025, 68(16): 17287–17302.
- [50] Zhang C S, University P, Wang F H, et al. *De novo* design of cyclic peptide binders based on fragment docking and assembling[J]. *Journal of Chemical Information and Modeling*, 2025, 65(8): 4206–4218.
- [51] Santos-Júnior C D, Torres M D T, Duan Y Q, et al. Discovery of antimicrobial peptides in the global microbiome with machine learning[J]. *Cell*, 2024, 187(14): 3761–3778. e16.
- [52] Wang J K, Feng J W, Kang Y, et al. Discovery of antimicrobial peptides with notable antibacterial potency by an LLM-based foundation model[J]. *Science Advances*, 2025, 11(10): eads8932.
- [53] Feng Z Y, Chen J Y, Hai Y L, et al. Sliding-attention transformer neural architecture for predicting T cell receptor-antigen-human leucocyte antigen binding[J]. *Nature Machine Intelligence*, 2024, 6(10): 1216–1230.
- [54] Zhang X J, Zhang O, Shen C, et al. Efficient and accurate large library ligand docking with KarmaDock[J]. *Nature Computational Science*, 2023, 3(9): 789–804.
- [55] Zhang X J, Shen C, Zhang H T, et al. Advancing ligand docking through deep learning: Challenges and prospects in virtual screening[J]. *Accounts of Chemical Research*, 2024, 57(10): 1500–1509.
- [56] Shen C, Zhang X J, Deng Y F, et al. Boosting protein-ligand binding pose prediction and virtual screening based on residue-atom distance likelihood potential and graph transformer[J]. *Journal of Medicinal Chemistry*, 2022, 65(15): 10691–10706.
- [57] Gu S K, Shen C, Zhang X J, et al. Benchmarking AI-powered docking methods from the perspective of virtual screening[J]. *Nature Machine Intelligence*, 2025, 7(3): 509–520.
- [58] Chen S C, Zhang O, Jiang C R, et al. Deep lead optimization enveloped in protein pocket and its application in designing potent and selective ligands targeting LTK protein[J]. *Nature Machine Intelligence*, 2025, 7(3): 448–458.
- [59] Zhang O, Zhang J T, Jin J Y, et al. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling[J]. *Nature Machine Intelligence*, 2023, 5(9): 1020–1030.
- [60] Wang J K, Hsieh C Y, Wang M Y, et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning[J]. *Nature Machine Intelligence*, 2021, 3(10): 914–922.
- [61] Wu Z X, Zhang O, Wang X R, et al. Leveraging language model for advanced multiproperty molecular optimization via prompt engineering[J]. *Nature Machine Intelligence*, 2024, 6(11): 1359–1369.
- [62] Cao D H, Chen M G, Zhang R Z, et al. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein-ligand complex prediction[J]. *Nature Methods*, 2025, 22(2): 310–322.
- [63] Jiang Y Y, Zhang G, You J, et al. PocketFlow is a data-and-knowledge-driven structure-based molecular generative model[J]. *Nature Machine Intelligence*, 2024, 6(3): 326–337.
- [64] Li M Y, Song K, He J X, et al. Electron-density-informed effective and reliable *de novo* molecular design and optimization with ED2Mol[J]. *Nature Machine Intelligence*, 2025, 7(8): 1355–1368.
- [65] Wang L, Wang S, Yang H, et al. Conformational space profile enhances generic molecular representation learning [J/OL]. *bioRxiv*, 2023: 2023.12. 14.571629. <https://www.biorxiv.org/content/10.1101/2023.12.14.571629v1>.
- [66] Li F L, Hu Q Y, Zhou Y Q, et al. DiffPROTACs is a deep learning-based generator for proteolysis targeting chimeras[J]. *Briefings in Bioinformatics*, 2024, 25(5): bbae358.
- [67] Li F L, Hu Q Y, Zhang X L, et al. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs[J]. *Nature Communications*, 2022, 13: 7133.
- [68] Bran M A, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools[J]. *Nature Machine Intelligence*, 2024, 6(5): 525–535.
- [69] Tom G, Schmid S P, Baird S G, et al. Self-driving laboratories for chemistry and materials science[J]. *Chemical Reviews*, 2024, 124(16): 9633–9732.
- [70] Seifrid M, Pollice R, Aguilar-Granda A, et al. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab[J]. *Accounts of Chemical Research*, 2022, 55(17): 2454–2466.
- [71] Christensen M, Yunker L P E, Shiri P, et al. Automation isn't automatic[J]. *Chemical Science*, 2021, 12(47): 15473–15490.
- [72] Schwaller P, Gaudin T, Lányi D, et al. "Found in Translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models[J]. *Chemical Science*, 2018, 9(28): 6091–6098.
- [73] Jiang Z Y, Liu Y B, Ke Z F. Recent progress in autonomous laboratories for chemical synthesis[J]. *Communications in Computational Chemistry*, 2025, 7(3): 226–242.
- [74] Torres J A G, Lau S H, Anchuri P, et al. A multi-objective

- active learning platform and web app for reaction optimization[J]. *Journal of the American Chemical Society*, 2022, 144(43): 19999–20007.
- [75] Christensen M, Yunker L P E, Adedeji F, et al. Data-science driven autonomous process optimization[J]. *Communications Chemistry*, 2021, 4: 112.
- [76] Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models[J]. *Nature*, 2023, 624(7992): 570–578.
- [77] Ruan Y X, Lu C Y, Xu N, et al. An automatic end-to-end chemical synthesis development platform powered by large language models[J]. *Nature Communications*, 2024, 15: 10160.
- [78] Bayley O, Savino E, Slattery A, et al. Autonomous chemistry: Navigating self-driving labs in chemical and material sciences[J]. *Matter*, 2024, 7(7): 2382–2398.
- [79] Ha T, Lee D, Kwon Y, et al. AI-driven robotic chemist for autonomous synthesis of organic molecules[J]. *Science Advances*, 2023, 9(44): eadj0461.
- [80] Hysmith H, Foadian E, Padhy S P, et al. The future of self-driving laboratories: From human in the loop interactive AI to gamification[J]. *Digital Discovery*, 2024, 3(4): 621–636.
- [81] Maffettone P M, Friederich P, Baird S G, et al. What is missing in autonomous discovery: Open challenges for the community[J]. *Digital Discovery*, 2023, 2(6): 1644–1659.
- [82] Chen L F, Tan X Q, Wang D Y, et al. TransformerCPI: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments[J]. *Bioinformatics*, 2020, 36(16): 4406–4414.
- [83] Ni S K, Kong X T, Zhang Y Y, et al. Identifying compound-protein interactions with knowledge graph embedding of perturbation transcriptomics[J]. *Cell Genomics*, 2024, 4(10): 100655.
- [84] Shah P M, Zhu H M, Lu Z L, et al. DeepDTAGen: A multi-task deep learning framework for drug-target affinity prediction and target-aware drugs generation[J]. *Nature Communications*, 2025, 16: 5021.
- [85] Li Z Q, Ren P X, Yang H, et al. TEFDTA: A transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug-target affinities[J]. *Bioinformatics*, 2024, 40: btad778.
- [86] Sadri A. Is target-based drug discovery efficient? discovery and "off-target" mechanisms of all drugs[J]. *Journal of Medicinal Chemistry*, 2023, 66(18): 12651–12677.
- [87] Tong X C, Qu N, Kong X T, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery[J]. *Nature Communications*, 2024, 15: 5378.
- [88] Seal S, Trapotsi M A, Spjuth O, et al. Cell Painting: A decade of discovery and innovation in cellular imaging[J]. *Nature Methods*, 2025, 22(2): 254–268.
- [89] Zheng S J, Rao J H, Zhang J X, et al. Cross-modal graph contrastive learning with cellular images[J]. *Advanced Science*, 2024, 11(32): 2404845.
- [90] Bunne C, Roohani Y, Rosen Y, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities[J]. *Cell*, 2024, 187(25): 7045–7063.
- [91] Li X M, Chen L, et al. CLaSP: A contrastive learning-guided latent scoring platform for comprehensive drug-likeness evaluation[J]. *Journal of Medicinal Chemistry*, 2025, 68(15): 16750–16765.
- [92] Myung Y, de Sá A G C, Ascher D B. Deep-PK: Deep learning for small molecule pharmacokinetic and toxicity prediction[J]. *Nucleic Acids Research*, 2024, 52(W1): W469–W475.
- [93] Banerjee P, Kemmler E, Dunkel M, et al. ProTox 3.0: A webserver for the prediction of toxicity of chemicals[J]. *Nucleic Acids Research*, 2024, 52(W1): W513–W520.
- [94] Daina A, Michielin O, Zoete V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules[J]. *Scientific Reports*, 2017, 7: 42717.
- [95] Zeng H K. What is a cell type and how to define it?[J]. *Cell*, 2022, 185(15): 2739–2755.
- [96] Yang T, Wang Y Y, Ma F, et al. Build the virtual cell with artificial intelligence: A perspective for cancer research[J]. *Military Medical Research*, 2025, 12(1): 4.
- [97] Karr J R, Sanghvi J C, Macklin D N, et al. A whole-cell computational model predicts phenotype from genotype[J]. *Cell*, 2012, 150(2): 389–401.
- [98] Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology[J]. *Nature*, 2023, 618(7965): 616–624.
- [99] Qi X N, Zhao L H, Tian C Y, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery[J]. *Nature Communications*, 2024, 15: 9256.
- [100] Zerrouk N, Augé F, Niarakis A. Building a modular and multi-cellular virtual twin of the synovial joint in Rheumatoid Arthritis[J]. *npj Digital Medicine*, 2024, 7: 379.
- [101] Laubenbacher R, Mehrad B, Shmulevich I, et al. Digital twins in medicine[J]. *Nature Computational Science*, 2024, 4(3): 184–191.
- [102] Wang H W, Arulraj T, Ippolito A, et al. From virtual patients to digital twins in immuno-oncology: Lessons learned from mechanistic quantitative systems pharmacol-

- ogy modeling[J]. *npj Digital Medicine*, 2024, 7: 189.
- [103] FDA announces plan to phase out animal testing requirement for monoclonal antibodies and other drugs[EB/OL]. (2025-04-10)[2026-07-29]. <https://www.fda.gov/news-events/press-announcements/fda-announces-plan-phase-out-animal-testing-requirement-monoclonal-antibodies-and-other-drugs>.
- [104] Quinn T P, Jacobs S, Senadeera M, et al. The three ghosts of medical AI: Can the black-box present deliver?[J]. *Artificial Intelligence in Medicine*, 2022, 124: 102158.
- [105] Wang Y T, Su Z, Guo S L, et al. A survey on digital twins: Architecture, enabling technologies, security and privacy, and future prospects[J]. *IEEE Internet of Things Journal*, 2023, 10(17): 14965–14987.

Artificial intelligence-enabled drug discovery and development: Recent advances and future perspectives

LI Honglin^{1,2}

1. School of Pharmacy, East China Normal University, Shanghai 200062, China
2. Center for AI-Driven Drug Discovery and Innovation, Shanghai 200062, China

Abstract Drug research and development (R&D) is characterized by long cycles, high costs, and low clinical success rates, and traditional R&D paradigms have struggled to meet the demands of complex diseases and innovative drug development. In recent years, artificial intelligence (AI), leveraging its strengths in multi-source data fusion, complex pattern recognition, and intelligent decision-making, has been driving a paradigm shift in drug R&D from experience-driven to data-driven and knowledge-driven approaches. This article systematically reviews the latest advances in AI-empowered drug R&D worldwide, focusing on key applications across the full drug R&D pipeline and technological development trends, and synthesizes representative studies to summarize AI's vital roles in improving R&D efficiency, optimizing decision-making, and fostering paradigm changes. The review shows that AI has become an essential bridge linking biomedical data, computational models, and experimental validation. It has not only significantly enhanced the efficiency of early-stage target discovery, molecular design, and drug evaluation, but has also spurred the development of novel R&D modalities such as automated experimentation, self-driving laboratories, and virtual cells, thereby providing new technical pathways toward an intelligent drug R&D system. However, the broad application of AI in drug R&D remains constrained by insufficient high-quality data, limited model interpretability and generalizability, difficulties in multimodal knowledge fusion, and the lack of robust experimental validation and standardized evaluation frameworks; its clinical translational value thus requires further verification. Looking ahead, efforts should be directed toward strengthening high-quality data resources and sharing systems, developing interpretable AI models that integrate biological mechanisms and physical laws, promoting deep integration of generative AI, multi-agent systems, and self-driving laboratories, refining model evaluation standards and regulatory frameworks, and fostering collaborative innovation between AI and experimental sciences, so as to accelerate the evolution of AI-empowered drug R&D toward intelligent, automated, and closed-loop paradigms.

Keywords artificial intelligence; drug R&D; intelligent drug design; generative artificial intelligence; self-driving laboratories; virtual cells ●



(责任编辑 王微)