· 药咖论坛 ·

生物数据驱动的药物设计研究进展

杨皓[1,2]，白芳[1,2*]

（1. 上海科技大学免疫化学研究所，上海 201210；2. 上海科技大学生命科学与技术学院，上海 201210）

[摘要] 近年来，人工智能（artificial intelligence，AI）与药物科学的深度融合，为突破传统药物研发“高投入、低效率”的瓶颈提供了新契机。生物数据因覆盖范围广、信息量大等优势，已在基于AI的药物发现各环节中得到广泛应用。综述主要生物、药学数据的类型及规模，具体包括多组学数据、蛋白质互作网络、化学分子-生物活性关联数据及三维结构数据等；讨论各类数据在药物研发关键阶段——靶标发现、先导化合物发现与优化及成药性评估中的应用进展，以期为构建标准化、多模态融合且临床转化高效的生物数据驱动药物设计体系提供参考。

[关键词] 药物设计；数据驱动；人工智能；生物数据库

[中图分类号] R914.2；R965 [文献标志码] A [文章编号] 1001-5094（2025）12-1056-10

DOI: 10.20053/j.issn1001-5094.202509080692

Research Progress of Biodata-Driven Drug Design

YANG Hao[1,2], BAI Fang[1,2]

(1. Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai 201210, China; 2. School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China)

[Abstract] In recent years, the deep integration of artificial intelligence (AI) with pharmaceutical sciences has offered new opportunities to overcome the traditional bottlenecks of drug development characterized by high costs and low efficiency. Owing to their wide coverage and rich information content, biological data have been widely applied across various stages of AI-driven drug discovery. This paper summarizes the main types and scales of biological and pharmaceutical data, including multi-omics data, protein-protein interaction networks, chemical structure-bioactivity correlation data, and three-dimensional structural information, and discusses recent advances in the application of these data to key stages of drug development, such as target identification, lead compound discovery and optimization, and druggability evaluation, aiming to provide some reference for constructing a standardized, multi-modally integrated and clinically translatable biodata-driven drug design system.

[Key words] drug design; data-driven; artificial intelligence; biological database

传统药物研发普遍面临周期漫长、成本高昂及失败率高等严峻挑战。据统计，一款新药的平均研发周期长达10～15年[1]；而DiMasi等[2]的研究显示，将一种新药推向市场的平均成本估计高达25.58亿美元。与此同时，研发失败率亦居高不下，约90%的候选药物在进入临床试验阶段后宣告失败[1]。上述数据凸显了传统药物研发“高投入、低效率”的固有瓶颈，其关键挑战在于靶标确认难度大、化合物筛选与优化效率偏低。

在此背景下，寻求并构建新的研发模式已成为医药领域亟待解决的重要问题。自20世纪60年代分子对接概念提出以来，计算机辅助药物设计（computer-aided drug design，CADD）技术显著缩短了先导化合物发现周期[3]；进入人工智能（artificial intelligence，AI）时代后，深度学习等算法进一步革新了药物研发全流程。以蛋白质结构预测工具AlphaFold[4]为代表，其取得的里程碑式进展，有效缓解了长期制约CADD的靶点三维结构数据稀缺问题——通过基于氨基酸序列精准预测蛋白质结构，极大加速了靶点验证与基于结构的药物发现（structure-based drug discovery，SBDD）进程。

支撑这一研发模式升级的关键在于高质量生物数据：多组学数据、三维结构信息、药物-靶点活性数据及药代动力学性质[吸收、分布、代谢、排泄与毒性（absorption，distribution，metabolism，excretion，and toxicity，ADMET）]标签等多模态信息，为AI模型提供了丰富的特征输入。这些数据使深度网络能够捕捉复杂的生物医学规律，在靶标发现、先导化合物发现与优化以及临床前研究（见图1）等环节全面提升决策效率，为降低研发成本、缩短研发周期及提高研发成功率奠定了重要基础[5]。

接受日期：2025-11-18

基金项目：上海市计算生物学专项项目（No. 24JS2850200）

* 通信作者：白芳，研究员

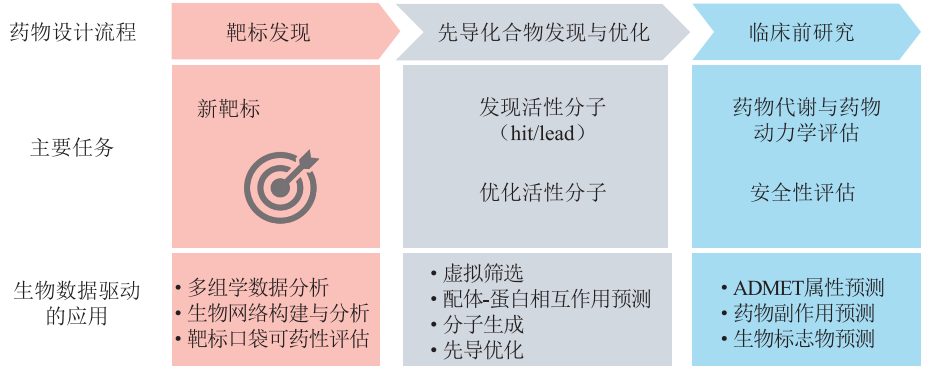
研究方向：药物设计，人工智能

E-mail：baifang@shanghaitech.edu.cn

1 生物数据：AI药物设计的关键驱动

高质量、多模态的生物数据正将药物研发从“经验驱动”全面推向“数据驱动”。一方面，跨层级

整合的多组学信息能够将基因突变、转录异常与代谢失衡映射至同一特征空间，为解析疾病因果网络、挖掘可药（druggable）靶点提供系统性依据。另一方面，蛋白质三维结构数据的大规模快速积累使可药位点的精准刻画成为现实，进而显著压缩“靶点-配体”搜索空间。与此同时，药物-靶点活性数据与 ADMET 标签的持续丰富，为生成式分子设计和成药性（developability）性评估提供了闭环反馈机制。经严格质量控制且完整保留生物学背景的多模态数据，既为深度模型提供了可信的训练基准，也支撑了模型解释能力与跨场景应用性能的提升，为 AI 药物设计的加速落地筑牢基础。



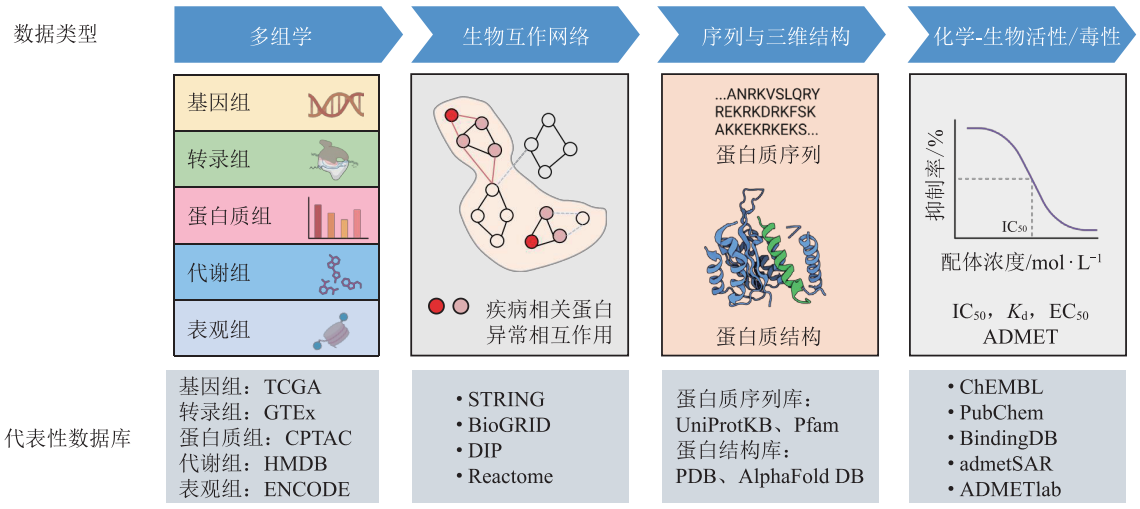
hit：苗头化合物；lead：先导化合物；ADMET：absorption, distribution, metabolism, excretion, and toxicity（吸收、分布、代谢、排泄与毒性）

图 1 生物数据在药物设计各阶段的应用示意图

Figure 1 Schematic diagram of biological data application in different stages of drug design

2 生物数据类型与资源

随着高通量测序、结构生物学方法等实验技术的快速发展，各类生物数据资源在药物发现与设计过程中的作用愈发突出。如图 2 所示，常见生物数据类型可大致分为 4 个主要层面：其一，多组学数据涵盖基因组、转录组、蛋白质组、代谢组和表观组等，为从分子到细胞层面解析疾病发生机制提供了丰富依据；其二，生物互作网络 [如蛋白-蛋白相互作用（protein-protein interactions，PPIs）图谱] 能够反映生物体内复杂的功能调控模块，为潜在药物靶点识别提供支撑；其三，序列与结构数据包含蛋白质一维氨基酸序列和三维空间构象，是 SBDD 与构效关系（structure-activity relationship，SAR 研究的重要基础；其四，化学-生物活性与毒性数据聚焦于小分子与靶标的结合常数、活性指标及 ADMET 性质，为候选化合物筛选与优化提供关键评价标准。对上述 4 类多模态数据的交叉整合，可为深度学习及传统算法提供多维特征输入，进而提升靶标发现、先导化合物筛选与优化的效率及成功率。



IC_{50}：半数抑制浓度；K_d：平衡解离常数；EC_{50}：半数有效浓度；ADMET：absorption, distribution, metabolism, excretion, and toxicity（吸收、分布、代谢、排泄与毒性）

图 2 常见的生物数据类型及代表性数据库

Figure 2 Common types of biological data and their representative databases

2.1 多组学数据

常见的组学类型包括基因组学、转录组学、蛋白质组学、代谢组学及表观组学，其对应的组学数据广泛应用于药物研发领域。伴随大规模公共计划的推进，多组学（multi-omics）数据量呈指数级增长：癌症基因组图谱（The Cancer Genome Atlas，TCGA）系统记录 30 余种癌症的基因变异与临床信息[6]；基因型 - 组织表达项目（Genotype-Tissue Expression Project，GTEx）V10 版本在 V8 基础上新增 12% 的 RNA 测序（Ribonucleic Acid Sequencing，RNA-seq）样本，并扩展表达数量性状位点（expression Quantitative Trait Loci，eQTL）分析覆盖度；蛋白质层面的临床蛋白质组肿瘤分析联盟（Clinical Proteomic Tumor Analysis Consortium，CPTAC）[7] 与蛋白质组学鉴定数据存储库（PRoteomics IDEntifications database，PRIDE）[8] 提供高分辨率磷酸化数据和质谱原始数据；人类代谢组数据库（Human Metabolome Database，HMDB）[9] 与代谢组学数据共享库（MetaboLights）[10] 汇集代谢指纹信息；DNA 元件百科全书（Encyclopedia of DNA Elements，ENCODE）/ 表观基因组学路线图计划（Roadmap Epigenomics Project）[11] 则清晰描摹调控元件与组蛋白修饰特征。人类细胞图谱（Human Cell Atlas，HCA）[12] 目前已累计公开逾 6 500 万条单细胞多组学记录，并正向“10 亿级细胞图谱”目标迈进。

多组学整合旨在将基因组、转录组、蛋白质组、代谢组与表观组等层级信息投射至统一特征空间，从而同步捕捉突变、表达和代谢失衡在疾病网络中的协同效应。这一全景式视角已被证实可显著提升病理机制解析深度，为精准医学和 AI 驱动药物设计提供系统性证据[13]。

多组学特征可用于筛选网络中心性高且功能关键的潜在靶点，提示耐药相关旁路通路，驱动生成式模型朝特定生物通路或表型定向优化分子结构；同时辅助将患者细分为代谢驱动型、免疫驱动型等亚群，为临床试验设计和个性化医疗实施提供支持。

尽管多组学数据库的规模为研究提供了前所未有的机遇，但实际应用中仍面临诸多挑战。首先，由非生物学因素（如实验时间、操作人员、试剂批次差异）引起的技术异质性（即批次效应）可能显著以至于掩盖了真实生物学差异。虽然有 ComBat-seq[14]、Harmony[15] 等统计学校正算法用于减轻该问题的影响，但此类方法仍存在局限——可能在去除批次效应的同时错误移除部分真实生物学信号（即“过度校正”）。其次，组学数据普遍具有“高维低样本量”（high-dimensional，low-sample-sizes，HDLSS）特征[16]，即特征数量远大于样本数量，这种特性极易导致 AI 模型训练出现过拟合。虽可通过特征选择、正则化等降维技术应对，但这些策略并非完美：特征选择存在丢失关键信息的风险，降维后的组合特征往往缺乏清晰生物学可解释性。此外，数据异质性与标准化缺失也是主要障碍。多组学数据本质上结构存在差异（如稀疏数据与连续数据），目前仍缺乏广泛认可的标准化流程。尽管学界正推动可查找、可访问、可互操作、可重用（findable，accessible，interoperable，reusable，FAIR）原则[17] 等数据标准，但这些框架的普适性和实施统一性仍有待提升。

2.2 生物大分子互作网络

生物大分子互作网络为连接基因组层线索与分子层干预手段提供了桥梁，在靶标筛选、作用机制阐释中发挥重要作用。其中，PPIs 网络记录了细胞内蛋白质间的物理或功能联系，构成复杂的互作调控框架。大量临床与实验证据表明，癌症、神经退行性疾病等病症往往伴随关键互作的丢失或异常增强[18-19]。

为系统地研究 PPIs，搜索工具用于检索交互基因/蛋白质数据库（Search Tool for the Retrieval of Interacting Genes/Proteins，STRING）、生物互作数据集综合数据库（Biological General Repository for Interaction Datasets，BioGRID）、蛋白质相互作用数据库（Database of Interacting Protein，DIP）等多个数据库相继建立，提供了丰富的蛋白质互作信息。其中，STRING v12.0（2025 年 3 月更新）整合实验、预测及文本挖掘等多源证据，已覆盖 5 930 万余种蛋白序列对应的 12 535 个物种，汇总约 200 亿条相互作用数据，并按置信度分层供用户下载[20]。BioGRID 4.4.245（2025 年 5 月发布）以人工文献标引为核心优势，收录约 220 万条非冗余物理互作数据，实行每月滚动更新[21]；创建时间较早的 DIP 侧重高质量实验验证互作数据，目前仍维护约 1.1 万条唯一互作记录[22]，常被用于小规模方法验证。这些资源为 PPIs 网络的构建与分析提供了坚实基础，辅助研究者精准识别关键网络节点及潜在药物靶点。

2.3 生物分子序列与结构信息

蛋白质的一级序列决定其三维折叠构象，而三维结构又直接限定其生物学功能，因此序列与结构数据共同构成分子层药物设计不可或缺的基础框架。

在序列数据方面，通用蛋白质知识库（Universal Protein Knowledgebase，UniProtKB）[23] 每 4 周更新

一次，截至 2024 年 6 月的 2024_06 版本已收录超 2.5 亿条蛋白序列及功能注释，为同源序列搜索、遗传变异功能注释及高深度多序列比对（multiple-sequence alignment，MSA）构建提供数据支撑。蛋白质家族数据库（Protein families database，Pfam）37.0[24] 进一步将 21 979 个蛋白家族的隐藏马尔可夫模型与标准 MSA 打包发布，便于快速提取进化共变信号。已有研究证实，MSA 的深度与多样性是 AI 结构预测精度的决定性因素——这一点在 AlphaFold 及 DeepMSA2 的技术突破中得到充分验证[4, 25]。

在结构数据方面，蛋白质结构数据库（Protein Data Bank，PDB）[26] 是全球最大的生物大分子结构数据库，截至 2024 年 9 月，已收录超过 22.4 万条实验解析的结构数据。而蛋白质-配体结合数据库（PDBbind）2024 版本[27] 为其中 2.7 万余个蛋白质-配体复合物配套了高质量亲和力数据，成为虚拟筛选和打分函数验证的黄金基准。序列库的持续扩充与结构资源的爆发式增长形成双向驱动：丰富的序列多样性为结构预测模型提供更全面的训练数据，高精度的预测结构又反过来指导实验解析和蛋白设计工作。

2.4 化学-生物活性与 ADMET 数据

化学-生物活性数据详细记录了小分子在多种体外或体内体系中的结合解离常数（K_D）、抑制常数（K_i）及半数抑制浓度（IC_{50}），为先导化合物筛选、SAR/定量构效关系（quantitative structure-activity relationship，QSAR）建模和候选分子优化提供量化依据。以欧洲分子生物学实验室化学数据库（ChEMBL Database，ChEMBL）35[28] 为例，其主库联合 SureChEMBL、生物化学实体数据库（Chemical Entities of Biological Interest，ChEBI）等子库已整理约 2 112 万条标准化活性记录，覆盖 1.6 万个靶点条目，是目前应用最广泛的开放式靶点-配体知识库。结合数据库（Binding Database，BindingDB）[29] 截至 2025 年 4 月已收集超过 300 万条来自文献和专利的亲和力数据，覆盖约 130 万个化合物及 9 500 个靶点。高通量筛选结果主要集中在公共化学生物活性测定数据库（PubChem BioAssay Database，PubChem BioAssay）[30]，其公开的 260 万余份实验方案涵盖近 460 万个化合物，为深度学习模型提供了数十亿级规模的活性标签[31]。

与活性数据互为补充的 ADMET 信息直接决定候选分子的药代动力学特性及安全窗口。ADMET 属性预测平台 ADMETlab3.0[32] 是综合性 ADMET 研究工具，整合了来自 ChEMBL、PubChem 和在线化学建模平台（Online Chemical Modeling Environment，OCHEM）等开放资源的超 40 万条数据，提供统一检索接口和应用程序编程接口功能。admetSAR3.0[33] 则整合超过 37 万条高质量实验 ADMET 数据，覆盖 104 652 个独特化合物；该平台采用基于对比学习的多任务图神经网络（graph neural network，GNN）框架，支持 119 个 ADMET 终点的预测，且具备用户友好的操作界面，兼容多种输入方式和输出格式。

活性与 ADMET 数据结合前述序列和结构信息，共同构成“生成—评价”闭环的关键环节：化学空间探索模型首先依据大规模活性标签完成虚拟筛选，随后借助并行属性预测实现多目标优化。基于深度学习的多任务框架目前已能同时评估多个 ADMET 端点，显著压缩苗头化合物到先导化合物（hit-to-lead）的研发周期；与此同时，持续扩充且经过严格质控的公共数据也反过来推动了自监督化学表示、主动学习及不确定性估计等方法的发展。

3 基于生物数据的靶标发现与验证

3.1 差异组学驱动的候选靶标筛选

差异组学分析已成为药物靶标发现的重要策略，其关键思路是比较疾病与正常状态下各类分子（转录本、蛋白质、代谢物等）的表达差异，从而识别与病理过程密切相关的关键基因或蛋白。

传统研究通常聚焦单一层面的显著性检验（如 DESeq2[34]、*limma*[35]），但单一组学仅能提供片段化信息，难以全面揭示复杂疾病的分子调控网络。近年来，多组学整合方法通过将突变负荷、差异表达、蛋白翻译后修饰等多源特征投射到共享潜在空间，使模型可同时捕获基因-基因、基因-蛋白及代谢通路间的高阶关联。这类方法的实现形式多样，包括 GNN、协同矩阵分解、张量分解、多视角自编码器及变分贝叶斯框架等，均已被用于学习跨组学的联合特征表示[36–37]。Li 等[38] 提出的 CGMega，将多组学特征与基于图形注意力的深度学习框架相结合，在乳腺癌和急性髓系白血病数据集中显著提升了候选靶点预测准确率，并解析出与疾病表型匹配的癌症基因模块。Lan 等[39] 开发的 MULGONET，基于网络富集的统计策略及可解释性设计，将基因本体（gene ontology，GO）信号映射到通路图、PPIs 网络和转录调控网络，用于预测癌症复发并挖掘生物标志物，该模型在胃癌、胰腺癌和膀胱癌等多个数据集上验证后，展现出优异的预测性能。

随着公共组学数据的持续积累及批次效应校正流程的日益标准化，差异组学驱动的靶标发现正从

“平面候选列表”向多模态系统建模演进，为后续结构可药性（druggablity）评估与候选化合物优化提供了更精确、可操作的研究起点。

3.2 PPIs 网络关键节点挖掘

PPIs 网络为理解疾病通路提供了系统性视角。研究表明，位于网络枢纽或关键簇的节点在病理状态下更易产生“放大效应”，因此被视为优先干预的潜在药物靶点[40]。传统关键节点挖掘主要分为 2 个步骤：先在全局网络中通过度中心性、介数中心性等指标度量节点拓扑重要性，再叠加疾病特异性信号（如突变负荷、差异表达等），评估其与疾病的功能关联度。

近年的研究方法不仅在网络中编码高阶邻域信息，还明确融合突变负荷、差异表达和表观修饰等多组学证据，以提升对疾病驱动基因的识别能力。例如，Ren 等[41]提出的 MONet 框架结合图卷积网络和图注意力网络对 PPIs 网络进行特征编码，将获得的特征向量输入多层感知机开展半监督学习，实现癌症驱动基因识别。在乳腺癌和肺腺癌数据集中，MONet 将靶点排序的受试者操作特征曲线下面积（area under the receiver operating characteristic curve，AUROC）从 0.89 提升至 0.93，成功识别出 *APOBEC2*、*GDNF* 等未被报道的驱动基因。Xu 等[42]提出的 SSCI 方法，通过掩码-重构预训练策略优化网络结构表征，在外部验证数据集中实现了超过 10% 的召回率提升，凸显无标签预训练在稀疏互作图分析中的优势。

总体而言，PPIs 网络关键节点挖掘已从基于单一拓扑指标的“中心性分析阶段”迈入多模态深度整合阶段，旨在从系统尺度精准锁定高影响力节点，为后续分子界面设计、分子胶靶点选择及组合疗法优化提供高置信度起点。

3.3 结构数据评估靶标口袋可药性

在 SBDD 中，蛋白质表面结合口袋的可药性评估是靶点遴选的首要环节。随着高分辨率晶体学、冷冻电子显微镜及 AlphaFold[43] 预测结构的激增，可药性评估形成 3 条互补技术路径：其一，几何特征分析（如 Fpocket[44]、SiteMap[45]）通过 α-球或栅格填充算法量化口袋体积、深度与封闭性，可快速捕获经典静态口袋；其二，理化属性量化（如 PockDrug-Server[46]），在疏水-亲水分布、电荷极性及氢键供体-受体等方面构建多元指标，并结合片段对接能量区分浅表缝隙；其三，基于大规模蛋白质-配体复合物的深度学习预测（如 PocketVec[47]、P2Rank[48] 等），利用三维卷积或图嵌入技术获取口袋向量表征，在独立验证集中将可药位点预测的 AUROC 提升至 0.90 以上，但仍存在可解释性不足的问题。

对于传统上被认为“难可药性”的 PPIs 界面，片段对接与直接耦合分析（fragment docking and direct coupling analysis，Fd-DCA）方法[49]可捕获瞬时凹陷及共进化热区，定位瞬时别构口袋，为蛋白降解靶向嵌合体、分子胶或界面竞争抑制剂设计提供结构依据。

随着动态结构数据库与深度生成模型的发展，可药性评估正从静态几何阈值判断向动态系统化量化演进，为下游分子对接、虚拟筛选与分子生成提供更多样、准确的口袋预测结果。

3.4 靶标发现方法的评价局限性与挑战

评估基于生物数据的靶标发现方法有效性时，需正视其在关键评价指标上的局限性。首先，高假阳性率是当前计算方法面临的普遍问题，该挑战贯穿计算药物设计各环节：例如“正向”虚拟筛选中，有研究指出其命中率仅约 12%[50]；而“反向”靶标垂钓过程中，现行打分函数的局限性——尤其对特定口袋性质（如疏水性、尺寸）的“打分偏见”，会产生大量假阳性靶点，干扰真实靶点识别。其次，这引发实际应用中敏感性与特异性的平衡问题：靶标发现早期探索阶段，计算模型需从海量数据中富集潜在候选集，策略上通常优先保障高敏感性（降低假阴性率）以避免遗漏有价值靶点，这必然导致特异性偏低，产生的大量假阳性结果需依赖后续成本更高、通量更低的生物学实验验证。

此外，不同方法（如前文提及的 MONet[41] 和 SSCI[42]）性能比较的统计学意义问题，目前领域内难以开展严格对比，主要原因在于缺乏统一、广泛认可的“金标准”验证集。一项针对新型冠状病毒 [严重急性呼吸综合征冠状病毒 2（severe acute respiratory syndrome coronavirus 2，SARS-CoV-2）] 主蛋白酶虚拟筛选研究的批判性综述[51]显示，61 篇相关论文中，超 67% 未报告任何验证方法，其余论文的验证手段也各不相同。数据集、验证流程及“阳性”定义的高度异质性，导致不同算法性能比较常缺乏统计学意义，结果难以跨研究直接对比。

4 基于生物数据的先导化合物发现与优化

4.1 结构数据驱动的分子对接与虚拟筛选

分子对接与虚拟筛选已成为发现先导化合物的常用计算手段，其技术核心是基于已知或预测的蛋白质三维结构，快速模拟小分子进入结合口袋的空间构象并定量评估结合亲和力。DOCK[52]、AutoDock4[53]、和 Glide[54] 等经典工具通过构象采样

与打分函数结合，实现受体-配体“预处理—对接计算—结果排序”全流程自动化，可在百万级化合物库高效筛选高亲和力候选分子。相较于传统实验筛选，虚拟筛选显著缩短了筛选周期并降低研发成本；而随着蛋白质结构资源的持续扩充及口袋预测方法的迭代升级，其应用范围正持续扩展。

盐野义制药（Shionogi）开发的恩司特韦（ensitrelvir，代号：S-217622）是结构数据驱动虚拟筛选的成功案例[55]。作为一种口服非肽类非共价SARS-CoV-2 3CL蛋白酶抑制剂，其研发全程采用SBDD策略：通过虚拟筛选识别初始命中化合物，随后进行生物学活性验证与结构优化，最终获得药代动力学性质良好的候选药物，为新型冠状病毒感染提供了重要治疗选择。

深度学习技术的融入，推动结构数据驱动的分子对接与虚拟筛选从“构象采样+经验打分”模式，向“端到端几何生成/评分”体系演进。此类模型通过学习蛋白质口袋与配体的联合表征，直接捕获二者间非线性相互作用，同步优化构象搜索精度与能量评估可靠性。早期的DeepDock采用混合高斯密度网络[56]，将口袋几何-理化特征与配体节点距离分布进行联合建模，在CASF-2016基准测试中，其排名相关系数提升至0.83，同时保持优异的筛选效能。后续开发的DiffDock[57]、TankBind[58]、DockFormer[59]等“生成—对齐”框架，进一步通过旋转-平移等协变操作减少对初始构象的依赖，提升复杂体系对接稳定性。Cao等[60]提出的SurfDock在上述技术路线基础上，创新性融入蛋白质表面形貌特征，利用几何扩散网络在蛋白曲面上同步优化配体平移、旋转及扭转角，在PoseBusters和Astex Diverse数据集上的对接成功率分别达78%和93%，显著优于其他基线方法。

随着深度学习框架的持续迭代与复合物结构库的不断积累，基于结构的虚拟筛选技术已能够在千万级化合物库内快速、可靠地筛选高亲和力候选分子，为先导化合物发现和后续优化提供了更高通量且更准确的计算支撑。

4.2 结构信息引导的分子生成

传统分子对接与虚拟筛选依赖现有小分子化学库，难以突破已知化学空间的局限，无法高效生成兼具新颖性与类药性的分子实体。生成式AI技术的发展为探索更加广阔的化学空间提供了全新技术路径。

早期分子生成方法多聚焦分子自身结构特征，未能有效整合靶标蛋白质的三维结构信息，导致生成分子与靶点的结合特异性不足。结构信息引导的分子生成作为近年来药物设计领域的研究热点，通过将靶标结合口袋的三维结构特征嵌入分子生成过程，在优化配体-靶点结合构象的同时兼顾分子类药性与可合成性，实现“结构匹配—功能优化”的协同调控。Dorna等[61]提出的TAGMol在扩散生成过程中引入“目标感知梯度”机制，同步考量分子与靶点的亲和力及药理属性，在多个靶基准测试中将Vina得分提升约22 %，显著优化了生成分子的结合活性。Lin等[62]开发的DiffBP进一步实现全原子层面的条件扩散生成，直接输出配体三维坐标，相较于传统顺序自回归方法，有效减少分子内物理冲突，同时提升生成多样性与结合能预测一致性。Teng等[63]提出的DTMol将预训练分子表示与扩散Transformer相结合，在百万级口袋-配体复合物数据上实现端到端对接与生成一体化，鉴定出2种新的实验验证有效的Janus激酶2抑制剂，证实了技术路线的实用性。

上述研究表明，随着多模态结构数据的积累与深度生成架构的创新，结构引导的分子生成正向预测精度提升、可解释性增强及化学空间覆盖拓展的方向发展，为先导化合物发现提供了可规模化应用的计算工具。

4.3 基于深度学习的药物-靶标亲和力预测

药物-靶标亲和力（drug-target affinity，DTA）预测旨在量化小分子与蛋白质的结合强度，是计算药物发现中的关键技术环节。传统DTA预测方法主要分为两类：基于结构的分子对接与基于特征相似性的机器学习模型（如KronRLS-MKL[64]和SimBoost[65]），但这类方法在处理大规模异质性数据及捕获复杂分子间相互作用时存在明显局限。

Öztürk等[66]提出的DeepDTA模型，首次利用深度学习方法，仅基于小分子的简化分子线性输入规范表征和蛋白质一级序列信息，通过卷积神经网络提取特征，实现了亲和力的回归预测。在KIBA数据集上，该模型将一致性指数提升至0.86，显著优于传统特征工程依赖性方法。基于DeepDTA的核心框架，研究者后续开发出WideDTA[67]、GraphDTA[68]、TEFDTA[69]和DMFF-DTA[70]等改进模型，通过引入GNN、注意力机制等先进技术，更高效地捕获药物分子结构特征与靶点序列信息的关联模式，进一步提升预测性能并增强模型可解释性。

综观从DeepDTA到DMFF-DTA的技术演进，DTA预测正由早期序列-化学单模态分析，逐步迈向口袋几何、图嵌入和自监督预训练融合的多模态研究阶段。在预测精度提升的同时，模型的可解释性（如显式特征对齐、注意力热图）、泛化能力及

跨研究可比性（依托治疗学数据共享基准平台）均得到显著改善，为后续生成式分子设计、虚拟筛选和多目标优化提供了可靠的活性预测依据。

4.4 基于生物数据的临床前成药性评估

通过 4.3 节所述方法初步确认分子的靶标亲和力后，临床前成药性评估成为决定候选分子研发价值的关键环节。该评估体系涵盖基础理化属性、可合成性及复杂体内 ADMET 特性等多维度指标。

在基础属性评估方面，研究者常用两类计算指标：一是基于物理化学特性的类药性评分 [药物相似性定量评估（quantitative estimation of drug-likeness，QED），取值范围 0～1]；二是预测合成难易程度的 SA-Score。这些定量评估在相关模型工作中均有体现，例如 TAGMol[61] 生成分子的平均 QED 为 0.55、平均 SA-Score 为 0.56，DiffBP[62] 生成的分子具有“良好类药性特征”。需强调的是，QED 和 SA-Score 本质上是“代理评估模型”，其预测结果与真实实验室合成难度及体内药效间存在一定偏差。目前 AI 从头（*de novo*）生成分子中成功进入临床试验的案例仍较为罕见，这一现象凸显了计算预测与实际药物开发间的转化鸿沟。

仅依赖 QED、SA-Score 等基础指标，难以全面规避药物研发后期失败风险。因此，精准预测候选分子的复杂体内 ADMET 特性，已成为成药性评估的关键。近年来，整合多任务学习、GNN 等技术的 ADMET 预测模型，凭借其多端点高效准确预测能力，逐步取代传统串联评估流程，实现研发早期对候选分子药代动力学性质与安全性的系统性评估，显著加速优化迭代进程并降低临床前开发失败率。

目前已有多个高性能 ADMET 预测平台投入实际应用。ADMETlab 3.0[32] 整合多数据库资源，构建覆盖 119 个预测端点的模型体系，采用有向信息传递神经网络及其与快速化学开发工具包生成描述符结合的变体架构，可同时支持回归与分类任务，在多个端点预测中表现优异。admetSAR 3.0[33] 以 37 万条实验数据为训练基础，结合对比学习的 GNN 模型，有效增强了预测泛化能力与实际应用价值。工业界层面，拜耳、勃林格-英格翰等公司 [71] 基于内部数据训练多任务 GNN 模型，其泛化性和外部适应性均优于单任务模型，为企业内部药物研发提供了重要支撑。

总体而言，临床前成药性评估正从传统单任务预测向多任务协同学习方向演进。面对标签稀缺、预测端点异构等实际挑战，新一代模型展现出更强的适应性和稳定性。随着高质量实验数据的持续积累和深度学习模型架构的不断革新，成药性预测工具正从辅助分析手段逐步升级为研发决策支撑引擎，在虚拟筛选、药物优化和安全性评估等关键环节中发挥日益重要的作用。

5 结语与展望

本文系统梳理了多组学数据、蛋白质互作网络、序列-结构信息以及化学-生物活性与 ADMET 数据在现代药物设计中的应用进展，完整覆盖从靶标发现与验证，到先导化合物发现与优化，再到临床前成药性评估的全流程关键技术环节。现有研究进展表明，以深度学习为代表的 AI 技术，正凭借强大的数据整合能力，将多模态、高维度的零散生物学证据转化为可量化的研发决策依据，推动药物研发从传统“经验试错”模式，加速向高效、精准的“数据驱动、闭环迭代”模式转型。

然而，要实现这一研发模式的全面落地，未来仍需在多个关键挑战上取得实质性突破。其中，以下几个方面尤为关键。

首先是数据层面的标准化与共享难题。如 2.1 节所述，多组学数据具有高度异质性且常分散于“数据孤岛”。未来行业发展的方向之一是推动 FAIR 数据原则 [17] 的全行业严格采纳，确保数据具备机器可读和跨平台可互操作特性，为 AI 模型的自动化数据整合提供基础支撑。同时，数据隐私保护需求日益凸显，联邦学习 [72] 作为核心技术解决方案，允许模型在数据不出本地的前提下实现协同训练，为敏感生物医学数据的安全利用提供技术路径。

其次是模型可解释性的提升。深度学习固有的“黑盒”特性，使其预测结果难以获得临床研究人员和监管机构的完全信任。因此，模型发展需从单纯追求预测精度转向构建机制驱动型 AI 体系。具体而言，需将已知的生物学知识作为先验嵌入模型，使其预测逻辑可追溯至明确的生物学依据，能启发新的生物学假设。

第 3 个关键挑战是跨模态验证鸿沟的弥合。如 3.4 节所讨论，计算（*in silico*）预测普遍存在的高假阳性率严重拖累了研发效率。未来的解决方案之一是构建高保真生物数字孪生系统 [73–74]，这类系统并非简单的分子结构模型，而是动态整合多组学数据、结构生物学信息及临床真实世界数据的多尺度、可计算虚拟患者模型。通过该系统，AI 可在 *in silico* 环境中高效模拟药物体内反应过程，在投入昂贵湿实验前完成大量无效候选分子的初步验证与淘汰，形成“计算生成—虚拟验证”的闭环。

随着 FAIR 数据生态的逐步成熟、可解释 AI 算

法的落地应用以及生物数字孪生技术的持续进步，结构信息与表型功能之间的实时互馈将成为可能，从而有望大幅缩短药物设计从计算生成到临床验证的迭代周期，最终加速精准化、个体化创新疗法的研发与临床转化进程，为医药健康领域的发展提供重要支撑。

[参考文献]

[1] Sun D, Gao W, Hu H, *et al*. Why 90% of clinical drug development fails and how to improve it?[J]. *Acta Pharm Sin B*, 2022, 12(7): 3049–3062.

[2] DiMasi J A, Grabowski H G, Hansen R W. Innovation in the pharmaceutical industry: new estimates of R&D costs[J]. *J Health Econ*, 2016, 47: 20–33.

[3] Sadybekov A V, Katritch V. Computational approaches streamlining drug discovery[J]. *Nature*, 2023, 616(7958): 673–685.

[4] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, 596(7873): 583–589.

[5] Zhang K, Yang X, Wang Y, *et al*. Artificial intelligence in drug development[J]. *Nat Med*, 2025, 31(1): 45–59.

[6] Weinstein J N, Collisson E A, Mills G B, *et al*. The cancer genome atlas pan-cancer analysis project[J]. *Nat Genet*, 2013, 45(10): 1113–1120.

[7] Mertins P, Mani D R, Ruggles K V, *et al*. Proteogenomics connects somatic mutations to signalling in breast cancer[J]. *Nature*, 2016, 534(7605): 55–62.

[8] Perez-Riverol Y, Csordas A, Bai J, *et al*. The PRIDE database and related tools and resources in 2019: improving support for quantification data[J]. *Nucleic Acids Res*, 2019, 47(D1): D442–D450.

[9] Wishart D S, Guo A, Oler E, *et al*. HMDB 5.0: the human metabolome database for 2022[J]. *Nucleic Acids Res*, 2022, 50(D1): D622–D631.

[10] Haug K, Cochrane K, Nainala V C, *et al*. MetaboLights: a resource evolving in response to the needs of its scientific community[J]. *Nucleic Acids Res*, 2020, 48(D1): D440–D444.

[11] Project Consortium E N C O E E, Moore J E, Purcaro M J, *et al*. Expanded encyclopaedias of DNA elements in the human and mouse genomes[J]. *Nature*, 2020, 583(7818): 699–710.

[12] Regev A, Teichmann S A, Lander E S, *et al*. The human cell atlas[J]. *eLife*, 2017, 6: e27041.

[13] Wu Y, Xie L. AI-driven multi-omics integration for multi-scale predictive modeling of genotype-environment-phenotype relationships[J]. *Comput Struct Biotechnol J*, 2025, 27: 265–277.

[14] Zhang Y, Parmigiani G, Johnson W E. *ComBat-seq*: batch effect adjustment for RNA-seq count data[J]. *NAR Genom Bioinform*, 2020, 2(3): lqaa078.

[15] Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with Harmony[J]. *Nat Methods*, 2019, 16(12): 1289–1296.

[16] Baião A R, Cai Z, Poulos R C, *et al*. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches[J]. *Brief Bioinform*, 2025, 26(4): bbaf355.

[17] Wise J, de Barron A G, Splendiani A, *et al*. Implementation and relevance of FAIR data principles in biopharmaceutical R&D[J]. *Drug Discov Today*, 2019, 24(4): 933–938.

[18] Gao M, Skolnick J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected[J]. *Proc Natl Acad Sci USA*, 2010, 107(52): 22517–22522.

[19] Tracy T E, Madero-Pérez J, Swaney D L, *et al*. Tau interactome maps synaptic and mitochondrial processes associated with neurodegeneration[J]. *Cell*, 2022, 185(4): 712–728.e14.

[20] Szklarczyk D, Nastou K, Koutrouli M, *et al*. The STRING database in 2025: protein networks with directionality of regulation[J]. *Nucleic Acids Res*, 2025, 53(D1): D730–D737.

[21] Stark C, Breitkreutz B J, Reguly T, *et al*. BioGRID: a general repository for interaction datasets[J]. *Nucleic Acids Res*, 2006, 34(Database issue): D535–D539.

[22] Xenarios I, Salwínski L, Duan X J, *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions[J]. *Nucleic Acids Res*, 2002, 30(1): 303–305.

[23] Consortium U. UniProt: the universal protein knowledgebase in 2025[J]. *Nucleic Acids Res*, 2025, 53(D1): D609–D617.

[24] Paysan-Lafosse T, Andreeva A, Blum M, *et al*. The Pfam protein families database: embracing AI/ML[J]. *Nucleic Acids Res*, 2025, 53(D1): D523–D534.

[25] Zheng W, Wuyun Q, Li Y, *et al*. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data[J]. *Nat Methods*, 2024, 21(2): 279–289.

[26] Berman H M, Westbrook J, Feng Z, *et al*. The protein data bank[J]. *Nucleic Acids Res*, 2000, 28(1): 235–242.

[27] Wang R, Fang X, Lu Y, *et al*. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures[J]. *J Med Chem*, 2004, 47(12): 2977–2980.

[28] Zdrazil B. Fifteen years of ChEMBL and its role in cheminformatics and drug discovery[J]. *J Cheminform*, 2025, 17(1): 32.

[29] Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities[J]. *Nucleic Acids Res*, 2007, 35(Database issue): D198–D201.

[30] Wang Y, Xiao J, Suzek T O, *et al*. PubChem's BioAssay database[J]. *Nucleic Acids Res*, 2012, 40(database issue): D400–D412.

[31] An S, Lee Y, Gong J, *et al*. InertDB as a generative AI-expanded resource of biologically inactive small molecules from PubChem[J]. *J Cheminform*, 2025, 17(1): 49.

[32] Fu L, Shi S, Yi J, *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 Res*, 2024, 52(W1): W422−W431.

[33] Gu Y, Yu Z, Wang Y, *et al*. admetSAR3.0: a comprehensive platform for exploration, prediction and optimization of chemical ADMET properties[J]. *Nucleic Acids Res*, 2024, 52(W1): W432−W438.

[34] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. *Genome Biol*, 2014, 15(12): 550.

[35] Ritchie M E, Phipson B, Wu D, *et al*. *limma* powers differential expression analyses for RNA-sequencing and microarray studies[J]. *Nucleic Acids Res*, 2015, 43(7): e47.

[36] Valous N A, Popp F, Zörnig I, *et al*. Graph machine learning for integrated multi-omics analysis[J]. *Br J Cancer*, 2024, 131(2): 205−211.

[37] Ballard J L, Wang Z, Li W, *et al*. Deep learning-based approaches for multi-omics data integration and analysis[J]. *BioData Min*, 2024, 17(1): 38.

[38] Li H, Han Z, Sun Y, *et al*. CGMega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection[J]. *Nat Commun*, 2024, 15: 5997.

[39] Lan W, Tang Z, Liao H, *et al*. MULGONET: an interpretable neural network framework to integrate multi-omics data for cancer recurrence prediction and biomarker discovery[J/OL]. *Fundam Res*, 2025[2025-11-18].https://doi.org/10.1016/j.fmre.2025.01.004.

[40] Jeong H, Mason S P, Barabási A L, *et al*. Lethality and centrality in protein networks[J]. *Nature*, 2001, 411(6833): 41−42.

[41] Ren Y, Zhang T, Liu J, *et al*. MONet: cancer driver gene identification algorithm based on integrated analysis of multi-omics data and network models[J]. *Exp Biol Med (Maywood)*, 2025, 250: 10399.

[42] Xu J, Hao J, Liao X, *et al*. SSCI self-supervised deep learning improves network structure for cancer driver gene identification[J]. *Int J Mol Sci*, 2024, 25(19): 10351.

[43] Varadi M, Anyango S, Deshpande M, *et al*. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models[J]. *Nucleic Acids Res*, 2022, 50(D1): D439−D444.

[44] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection[J]. *BMC Bioinformatics*, 2009, 10: 168.

[45] Halgren T A. Identifying and characterizing binding sites and assessing druggability[J]. *J Chem Inf Model*, 2009, 49(2): 377−389.

[46] Hussein H A, Borrel A, Geneix C, *et al*. PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins[J]. *Nucleic Acids Res*, 2015, 43(W1): W436−W442.

[47] Comajuncosa-Creus A, Jorba G, Barril X, *et al*. Comprehensive detection and characterization of human druggable pockets through binding site descriptors[J]. *Nat Commun*, 2024, 15(1): 7917.

[48] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure[J]. *J Cheminform*, 2018, 10(1): 39.

[49] Bai F, Morcos F, Cheng R R, *et al*. Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis[J]. *Proc Natl Acad Sci USA*, 2016, 113(50): E8051−E8058.

[50] Adeshina Y O, Deeds E J, Karanicolas J. Machine learning classification can reduce false positives in structure-based virtual screening[J]. *Proc Natl Acad Sci USA*, 2020, 117(31): 18477−18488.

[51] Macip G, Garcia-Segura P, Mestres-Truyol J, *et al*. Haste makes waste: a critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition[J]. *Med Res Rev*, 2022, 42(2): 744−769.

[52] Meng E C, Shoichet B K , Kuntz I D. Automated docking with grid-based energy evaluation[J]. *J Comput Chem*, 1992, 13(4): 505−524.

[53] Morris G M, Huey R, Lindstrom W, *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility[J]. *J Comput Chem*, 2009, 30(16): 2785−2791.

[54] Friesner R A, Banks J L, Murphy R B, *et al*. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy[J]. *J Med Chem*, 2004, 47(7): 1739−1749.

[55] Unoh Y, Uehara S, Nakahara K, *et al*. Discovery of S-217622, a noncovalent oral SARS-CoV-2 3CL protease inhibitor clinical candidate for treating COVID-19[J]. *J Med Chem*, 2022, 65(9): 6499−6512.

[56] Méndez-Lucio O, Ahmad M, del Rio-Chanona E A, *et al*. A geometric deep learning approach to predict binding conformations of bioactive molecules[J]. *Nat Mach Intell*, 2021, 3(12): 1033−1039.

[57] Corso G, Stärk H, Jing B, *et al*. DiffDock: diffusion steps, twists, and turns for molecular docking [PP/OL]. arXiv (2022-10-04) [2025-08-28]. https://arxiv.org/abs/2210.01776.

[58] Lu W, Wu Q, Zhang J, *et al*. TankBind: trigonometry-aware neural networks for drug-protein binding structure prediction[PP/OL]. bioRxiv (2022-06-06) [2025-08-28]. https://doi.org/10.1101/2022.06.06.495043.

[59] Yang Z, Ji J, He S, *et al*. DockFormer: a transformer-based molecular docking paradigm for large-scale virtual screening[PP/OL]. arXiv (2024-11-11) [2025-08-28]. https://doi.org/10.48550/arXiv.2411.06740.

[60] Cao D, Chen M, Zhang R, *et al*. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein-ligand complex prediction[J]. *Nat Methods*, 2025, 22(2): 310−322.

[61] Dorna V, Subhalingam D, Kolluru K, *et al*. TAGMol: target-aware gradient-guided molecule generation[PP/OL]. arXiv(2024-06-03) [2025-08-28]. https://doi.org/10.48550/arXiv.2406.01650.

[62] Lin H, Huang Y, Zhang O, *et al*. DiffBP: generative diffusion of 3D molecules for target protein binding[PP/OL]. arXiv (2024-07-14) [2025-08-28]. https://doi.org/10.48550/arXiv.2211.11214.

[63] Teng H, Wang R, Shen Y, *et al*. DTMol: pocket-based molecular docking using diffusion transformers[PP/OL]. bioRxiv(2025-04-18) [2025-08-28]. https://doi.org/10.1101/2025.04.13.648103.

[64] Nascimento A C A, Prudêncio R B C, Costa I G. A multiple kernel learning algorithm for drug-target interaction prediction[J]. *BMC Bioinformatics*, 2016, 17: 46.

[65] He T, Heidemeyer M, Ban F, *et al*. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines[J]. *J Cheminform*, 2017, 9(1): 24.

[66] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction[J]. *Bioinformatics*, 2018, 34(17): i821–i829.

[67] Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drug-target binding affinity[PP/OL]. arXiv (2019-02-04) [2025-08-28]. https://doi.org/10.48550/arXiv.1902.04166.

[68] Nguyen T, Le H, Quinn T P, *et al*. GraphDTA: predicting drug-target binding affinity with graph neural networks[J]. *Bioinformatics*, 2021, 37(8): 1140–1147.

[69] Li Z, Ren P, 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(1): btad778.

[70] He H, Chen G, Tang Z, *et al*. Dual modality feature fused neural network integrating binding site information for drug target affinity prediction[J]. *NPJ Digit Med*, 2025, 8(1): 67.

[71] Walter M, Borghardt J M, Humbeck L, *et al*. Multi-Task ADME/PK prediction at industrial scale: leveraging large and diverse experimental datasets[J]. *Mol Inform*, 2024, 43(10): e202400079.

[72] Eden R, Chukwudi I, Bain C, *et al*. A scoping review of the governance of federated learning in healthcare[J]. *NPJ Digit Med*, 2025, 8(1): 427.

[73] Vallée A. Digital twins for personalized medicine require epidemiological data and mathematical modeling: viewpoint[J]. *J Med Internet Res*, 2025, 27: e72411.

[74] Shmatko A, Jung A W, Gaurav K, *et al*. Learning the natural history of human disease with generative transformers [J]. *Nature*, 2025, 647(8088): 248–256.

[专家介绍]白芳：上海科技大学免疫化学研究所研究员、博士生导师、课题组组长，国家海外高层次青年人才计划入选者。大连理工大学/中国科学院上海药物研究所联合培养博士，美国莱斯大学博士后，曾任美国德克萨斯大学休斯顿健康科学中心助理教授。研究方向以药物设计新计算方法的研发为重点，聚焦新药设计与药物作用机制解析等应用研究，尤其在抗感染药物、抗肿瘤药物研发领域形成特色优势，致力于通过计算生物学与药物化学的交叉融合解决临床药物研发关键问题。已在 *PNAS*、*Nat Commun*、*Adv Sci* 等国际权威期刊上发表论文 80 余篇，连年入选斯坦福大学发布的“全球前 2% 顶尖科学家榜单”。

（责任编辑：邢爱敏）