DOI: 10.13345/j.cjb.230679
Jul. 25, 2024, 40(7): 2070-2086

·综 述·

人工智能在蛋白质-配体结合亲和力预测中的研究进展

云轶楠[1,2,3#]，刘诗梦[3#]，代琦[1]，张瑾[2*]，王筱[3]

1 浙江理工大学 生命科学与医药学院，浙江 杭州 310018
2 嘉兴学院 生物与化学工程学院，浙江 嘉兴 314001
3 嘉兴欣贝莱生物科技有限公司，浙江 嘉兴 314001

云轶楠, 刘诗梦, 代琦, 张瑾, 王筱. 人工智能在蛋白质-配体结合亲和力预测中的研究进展[J]. 生物工程学报, 2024, 40(7): 2070-2086.
YUN Yinan, LIU Shimeng, DAI Qi, ZHANG Jin, WANG Xiao. Advances in using artificial intelligence for predicting protein-ligand binding affinity[J]. Chinese Journal of Biotechnology, 2024, 40(7): 2070-2086.

摘 要：蛋白质与配体的结合在生命过程中发挥重要作用，计算蛋白质-配体结合亲和力(protein-ligand binding affinity, PLBA)有助于解析蛋白质功能、筛选与蛋白靶点结合的药物以及进行酶的改造等。近年来，人工智能(artificial intelligence, AI)发展迅速，因其特征提取能力强、算法准确度高、计算速度快等优势，已广泛应用于 PLBA 预测。本文介绍了 AI 预测的建立过程、相关资源、应用场景以及面临的挑战和潜在解决办法，为相关研究提供借鉴。
关键词：人工智能；蛋白质-配体结合亲和力；药物研发；酶工程

Advances in using artificial intelligence for predicting protein-ligand binding affinity

YUN Yinan[1,2,3#], LIU Shimeng[3#], DAI Qi[1], ZHANG Jin[2*], WANG Xiao[3]

1 School of Life Sciences and Medicine, Zhejiang Sci-Tech University, Hangzhou 310018, Zhejiang, China
2 School of Biological and Chemical Engineering, Jiaxing University, Jianxing 314001, Zhejiang, China
3 Jiaxing Synbiolab Biotech Co. Ltd., Jiaxing 314001, Zhejiang, China

Abstract: The binding of proteins and ligands is a crucial aspect of life processes. The calculation of the protein-ligand binding affinity (PLBA) offers valuable insights into protein

资助项目：国家自然科学基金(32172708)
This work was supported by the National Natural Science Foundation of China (32172708).
#These authors contributed equally to this work.
*Corresponding author. E-mail: zhangjin7688@163.com
Received: 2023-10-08; Accepted: 2024-01-10; Published online: 2014-01-17

function, drug screening targets protein receptors, and enzyme modifications. In recent years, artificial intelligence (AI) has experienced rapid advancements, becoming widely used in PLBA prediction. This is attributed to its robust feature extraction ability, superior algorithm accuracy, and speedy calculations. Our paper aims to provide a comprehensive overview of AI predication process, associated resources, application scenarios, challenges, and potential solutions, serving as a valuable reference for the relevant research endeavors.

Keywords: artificial intelligence; protein-ligand binding affinity; drug development; enzyme engineering

蛋白质作为生命过程中最重要的生物大分子，往往通过与配体结合发挥作用[1-2]，其中配体通常为小分子化合物[3-5]。蛋白质-配体结合亲和力(protein-ligand binding affinity, PLBA)是描述蛋白质与配体相互作用的定量指标[6]，计算PLBA对于理解蛋白质功能、辅助药物发现和酶改造工作意义重大。然而，直接测量PLBA的实验方法存在操作繁琐、耗时长、成本高等问题[7]。

人工智能(artificial intelligence, AI)的快速发展为PLBA的研究提供了新的策略。通过AI算法预测PLBA的大致流程如下：研究人员收集含有蛋白质和配体信息及PLBA的已知数据集，利用AI算法建立输入(蛋白质配体序列或结构)和输出(PLBA)的关系函数，在多次训练后得到精度较高的预测模型，最终用于预测未知的PLBA[4,8]。本文将重点介绍AI预测模型的建立流程及相关资源，并结合案例介绍不同应用场景下PLBA研究的进展，最后探讨现存挑战和可能的解决思路。

1 基于AI的PLBA预测模型的建立流程

构建AI预测模型一般分为数据准备(data)、模型构建(model)和性能评估(evaluation)三个阶段(图1)，研究人员从数据库(database)中收集蛋白质、配体及PLBA信息，将蛋白质和配体信息转化成计算机能识别的数据表示(data representation)[4]，然后构建包含数据表示和PLBA的训练集(train set)和测试集(test set)，基于AI算法(algorithm)的PLBA预测模型在训练集中学习特征以建立蛋白质、配体与PLBA的映射关系，最后模型预测测试集中的PLBA，通过比较预测结果与真实值的差异来评估模型的性能。

1.1 数据准备

通常情况下构建基于AI的预测模型仅有20%的时间用于编写算法，而80%的时间则主要用于准备数据[9]，可见高效获取数据是加快模型构建的关键一步。在数据准备的过程中，丰富的数据有利于AI模型更加充分地学习蛋白质和配体的特征，而准确性高、代表性强且格式统一的数据有利于AI模型更加准确地捕捉蛋白质、配体结构和PLBA的关系，简而言之，数据的数量和质量是提高AI模型泛化性能的基础[10-11]。接下来将逐一介绍基于AI的PLBA预测模型构建之初需要做的数据准备，包括数据资源的收集和数据表示的构建。

1.1.1 数据资源

本文总结了包含蛋白质、配体信息及PLBA的数据库，其中PDBbind[12]数据库包含大量由实验确定的蛋白质-配体复合物的结构信息和PLBA数据，研究人员可以利用PDBbind中的数据开发和验证基于AI的PLBA预测模型；DUD-E[13]包含配体药物和非配体药物，CASF-2016[14]包含真阳性和假阳性的PLBA数据，这两种数据库可以帮助研究人员评估模型筛选能力和预测能力(表1)。

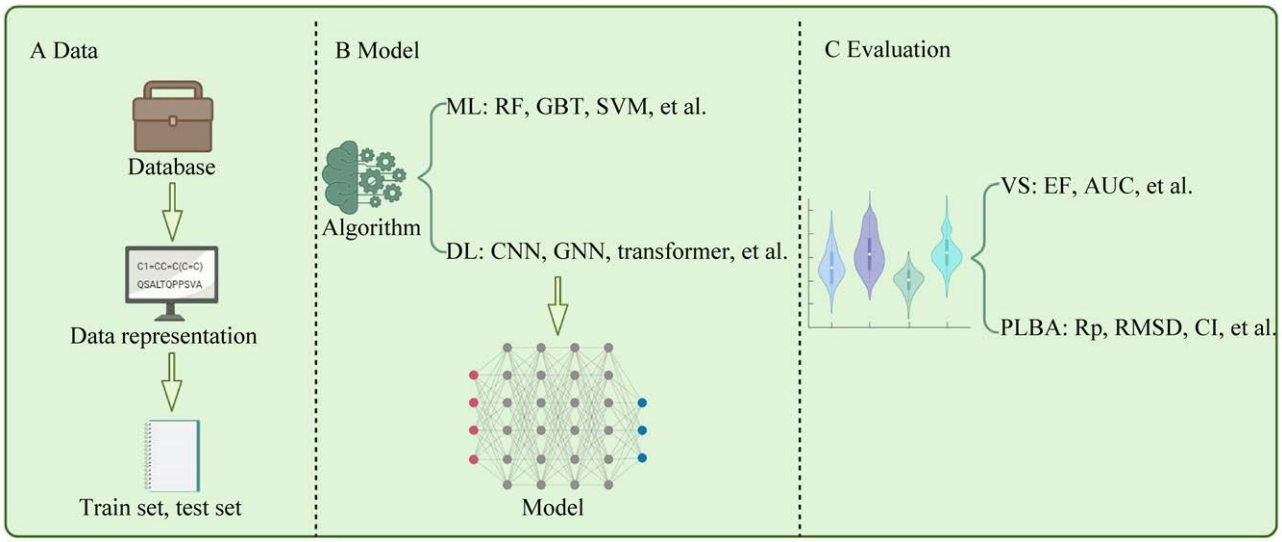

图 1　基于 AI 的 PLBA 预测模型建立流程　图中 ML 表示机器学习(machine learning)，RF 表示随机森林(random forest)，GBT 表示梯度增强树(gradient boosting tree)，SVM 表示支持向量机(support vector machine)，DL 表示深度学习(deep learning)，CNN 表示卷积神经网络(convolutional neutral network)，GNN 表示图神经网络(graph neural network)，transformer 表示转换器，VS 表示虚拟筛选(virtual screening)，EF 表示富集因子(enrichment factor)，AUC 表示曲线下面积(area under the curve)，Rp 表示 Pearson 相关系数(Pearson's correlation coefficient)，RMSD 表示均方根误差(root mean square deviation)，CI 表示一致性指数(concordance index)

Figure 1　The process of building a model for predicting PLBA based on AI algorithm. ML stands for machine learning, RF stands for random forest, GBT stands for gradient boosting tree, and RF stands for random forest, SVM stands for support vector machine, DL stands for deep learning, CNN stands for convolutional neutral network, GNN stands for graph neural network, VS stands for Virtual Screening, EF stands for enrich factor, AUC stands for area under the curve, Rp stands for Pearson's correlation coefficient, RMSD stands for root mean square deviation , and CI stands for concordance index.

1.1.2　数据表示

由于计算机系统只能识别数字向量，所以需要将原始数据转换成计算机能够识别的数据表示[21]。蛋白质和配体相应的数据表示方法包括简化分子输入行系统(simplified molecular input line entry system, SMILES)[22]、指纹(fingerprint, FP)[23]和图(graph)[24]等(表 2)。在构建 AI 模型时，研究人员需要根据蛋白质、配体的结构大小以及复杂程度选择合适的数据表示。

1.2　模型构建

AI 算法是一种基于已有数据、知识或经验自动识别数据特征，并在相似环境中做出预测或决策的技术，主要包括机器学习(machine learning, ML)及 ML 的分支——深度学习(deep learning, DL)。2010−2023 年间研究人员相继开发了很多基于 AI 的 PLBA 预测模型，该类模型的开发有效缩小相关实验范围，极大地推动了生物学研究进程。

ML 算法包括随机森林(random forest, RF)[4]算法、梯度增强树(gradient boosting tree, GBT)[4]算法、极端梯度提升(extreme gradient boosting, XGB)算法[42]和支持向量机(support vector machine,

表 1　蛋白质和配体信息数据库

Table 1　Protein and ligand information databases

Database	Description	Link
PDBbind[12]	A database focused on curating and sharing binding affinity information for protein-ligand interactions	http://www.pdbbind.org.cn/
DUD-E[13]	A database focused on benchmarking virtual screening methods	https://dude.docking.org/
CASF-2016[14]	A database specifically designed for the evaluation of docking and scoring methods for PLBA	http://www.pdbbind-cn.org/casf.asp/
PubChem[15]	An open chemistry database mostly contains small molecules (also large molecules) with their properties and active	https://pubchem.ncbi.nlm.nih.gov/
DrugBank[16]	A comprehensive pharmaceutical database	https://www.drugbank.com/
UniProt[17]	A database providing comprehensive information on protein sequences and functions across various species	https://www.uniprot.org
PDB[18]	A database providing 3D structural data of biological molecules	https://www.rcsb.org
Brenda[19]	A database providing detailed information on enzyme function and properties	https://www.brenda-enzymes.org/
BindingDB[20]	A database specializing in molecular binding data	https://www.bindingdb.org/rwd/bind/index.jsp

表 2　蛋白质和配体的常用数据表示

Table 2　Common data representations of protein and ligand

Data representation	Data structure	Description	Characteristic
SMILES[22]	A string of ASCII characters	SMILES conveys information about the atoms, connectivity, and stereochemistry of molecules	SMILES is a molecular representation that is highly efficient for computer storage and transmission. However, it can exhibit significant variation in notation for structurally similar small molecules, which may hinder machine learning models' ability to capture fine molecular features[22]
FP[23]	A string of binary characters	FP encompasses a broader spectrum of molecular details, including topological structure, charge state, ring systems, and substructures	FP, a molecular representation method, offers simplicity in its mathematical structure, minimal storage demands, and swift AI model training and testing[23]. However, it mainly comprises binary data, leading to the loss of finer molecular details
Graph[24]	Includes edge vector, node vector and node connection relation matrix	In addition to the aforementioned molecular information, graph also encapsulates details such as atomic electronegativity, formal charges, radii, as well as bond characteristics and lengths	The graph representation method for molecules is a cutting-edge approach, encoding richer molecular information for an intuitive and flexible structure representation[24]. However, it's worth noting that with large or complex molecules, this method's use demands more computational resources and time for processing

SVM)[4]算法等。2010 年，Ballester 等[25]提出了基于 RF 的 RF-Score 模型，该模型首次将 ML 算法用于 PLBA 预测。RF-Score 以 12 Å 距离内蛋白质和配体原子对的出现次数作为特征，其

中包括 4 种蛋白质原子类型 C、N、O 和 S，9 种配体原子类型 C、N、O、F、P、S、Cl、Br 和 I，共计 36 个特征向量，在含 1 300 条数据的 PDBbind v2007[26]一般集(general set)上训练，在含 195 条数据的 PDBbind v2007[26]核心集(core set)上表现出比传统评分函数更优越的预测能力，代表将 ML 算法应用于评分函数(scoring funciton, SF)设计的时代正式到来。

DL 是一种依靠多层神经网络实现复杂特征的学习和特征表示的技术，包括卷积神经网络(convolutional neural network, CNN)[4]算法、图神经网络(graph neural network, GNN)[4]算法以及转换器(transformer)[42]算法等。2017 年，Ragoza 等[31]首次使用基于 CNN 的评分函数预测 PLBA，并通过测试证明了用 CNN 设计评分函数的潜力与可行性。该模型采用了一个以结合位点为中心、边长为 24 Å 的网格，将不同类型的蛋白质和配体原子之间的距离作为特征，包括 16 种蛋白质原子和 18 种配体原子。在结合位点预测任务中，该模型在 CSAR[52]数据集上进行训练，包含 745 个正样本和 3 251 个负样本，并在包含 195 条数据的 PDBbind v2013[12]核心集上测试，在虚拟筛选(virtual screening, VS)任务中，该模型在含 22 645 条正样本和 1 407 145 条负样本的 DUD-E[13]数据集上训练，并在 ChEMBL[32](含 11 406 个正样本和 10 000 个负样本)和 MUV[33](含 1 913 个正样本和 1 177 989 个负样本)数据集上测试。结果显示，该模型在结合位点预测和虚拟筛选任务中均优于 AutoDock Vina[53]评分函数。

深圳先进院 Li 等[29]、Yan[30]等和 Wang 等[39]以 OnionNet 为核心成功开发了一系列用于预测 PLBA 的模型，在 PDBbind v2016 基准测试中取得了显著的精度优势。2019 年该团队还开发了基于 CNN 的预测模型 OnionNet[29]，在 30.5 Å 距离内，该模型以配体原子为中心定义了 60 个间隔为 0.5 Å 的同心圆边界，将相邻边界之间的蛋白质原子与配体原子的接触数作为特征，蛋白质和配体原子类型均为 C、H、O、N、P、S、卤素以及其他元素(8 种)，共 3 840 (8×8×60)个特征向量，在含有 11 906 条数据的 PDBbind v2016[35]精炼集(refined set)上训练，在含有 398 条数据的 PDBbind v2013[12]和 v2016[54]核心集上测试性能，其预测值与实验值的相关性可达 73%，精度 R 为 0.816。2021 年团队进一步优化 OnionNet 并开发了 OnionNet-2[39]，将特征改为相邻边界之间的氨基酸残基与配体原子的接触数，相邻边界之间的特征量从 8 种蛋白质原子扩大至 21 种氨基酸残基，此次特征选取使模型更加准确地学习到蛋白质配体相互作用，精度 R 提升至 0.864。此外，团队还设计了评分函数修正项 OnionNet-SFCT[30]以提高传统评分函数的对接和筛选能力，在 CASF-2016 基准测试中，OnionNet-SFCT 与 AutoDock Vina 联用实现了富集能力最强的筛选效果，性能是单独使用 AutoDock Vina 的 2 倍。

2021 年，浙江大学侯廷军教授团队和浙江大学吴健教授团队、中南大学曹东升团队以及腾讯量子实验室[35]共同开发基于 GNN 的预测模型 IGN，该模型在 PLBA 预测、VS 以及小分子结合构象预测任务中展现了优越的泛化性能。IGN 以配体图、蛋白质图和蛋白质-配体图作为特征，其特征图的节点表示相关原子，两种不同类型的边对应共价连接和非共价连接，同时嵌入了距离、角度、面积等信息以全面表示蛋白质和配体的三维结构，使用两个独立的图卷积模块学习分子内和分子间的相互作用，当蛋白质原子和配体原子的距离在阈值 8 Å 以内则视为可能存在相互作用。IGN 在含有 8 298 条数据的 PDBbind v2016[36]一般集上训练，在含有

357 条数据的 PDBbind v2016[36]核心集上测试，其精度 R 可达 0.837，在虚拟筛选和小分子结合构象预测任务中表现出更好或相似的性能。

2023 年，之江实验室、百度大数据以及香港科技大学[46]共同开发了基于 GNN 的 CurvAGN，该模型将曲率、同源性等信息纳入特征，规避了传统 GNN 只考虑蛋白质和配体距离、忽略角度的问题，更加全面地提取了蛋白质和配体的结构特征。CurvAGN 以蛋白质-配体相互作用图作为特征，其中包括 36 维向量用于编码原子，26 维向量用于编码原子距离，使用曲率块(curvature block)编码蛋白质和配体的多尺度曲率信息，使用自适应注意机制(adaptive graph attention mechanism)将几何结构，包括角度、距离和多尺度曲率以及远距离分子相互作用和图的异源性(heterophily)，纳入蛋白质-配体复合物的表示中，在含有 11 993 条数据的 PDBbind v2016[35]一般集上训练，在含有 290 条数据的 PDBbind v2016[35]核心集上测试得到较为优越的预测性能(精度 R 为 0.831)，证明了其特征选取具有一定的可靠性和有效性。

此外，国内外相关课题组如中南大学 Wang 等[48]开发了基于 GNN 的预测模型 GraphScoreDTA、中国矿业大学 Zhang 等[43]开发了基于 transformer 的 MRBDTA、河北师范大学 Zhang 等[44]开发了基于 GNN 的 SS-GNN、复旦大学药学院 Zhang 等[49]开发了基于 GNN 的 PLANET、中山大学 He 等[51]开发了基于 GNN 的 NHGNN-DTA、美国东北大学 Chatterjee 等[55]开发了基于 CNN 的 AI-Bind 等，均为 PLBA 预测工作做出有效贡献(表 3)。

与 ML 相比，具有多层网络结构的 DL 擅长处理大规模数据(数据集大于 10 000)并能够自动提取数据特征，不仅弥补了 ML 难以捕捉数据隐藏信息的不足，且其性能通常比 ML 更加优越[56-57]。DL 的网络结构复杂且训练集庞大，导致其需要更强大的算力支持和更长的计算时间[58]，因此在预测 PLBA 的任务中需要充分考虑预测精度、运行速度和算法难度的因素来选择最适合的算法。

1.3 模型评估

评估模型的准确性和精确度需要建立合适的性能指标，虚拟筛选任务需要评估模型区分结合和非结合配体的能力[4]，常用指标有富集因子(enrichment factor, EF)[4]、曲线下面积(area under the curve, AUC)[4]等，PLBA 预测任务需要衡量模型的预测值与实验数据的相关性及差距[4]，常用指标有 Pearson 相关系数(Pearson's correlation coefficient, R_p)[4]、均方根误差(root mean square deviation, RMSD)[4]和一致性指数(concordance index, CI)[41]等 (表 4)。

1.4 研究蛋白质-配体相互作用的预测工具

在蛋白质配体的相互作用的实际研究中，Taba[59]和 CSatDTA[60]是 PLBA 预测工具，相比之下，CSatDTA 有着简洁的页面供用户提交数据，PLBA 预测值可在页面自动呈现，Taba 需要下载至本地，用户可以根据需求选择合适的算法来预测 PLBA。MANORAA[61]通过分析结合口袋、原子距离和活性位点等信息在线研究 PLBA，并为药物设计提供定制的生物医学数据及其注释。gRINN[62]是蛋白质分子动力学研究工具，可计算成对氨基酸的相互作用能，同时提供相互作用能和能量网络的可视化界面，需要下载到本地使用。DeepSite[63]、COACH[64]和 DoGSiteScorer[65]都是蛋白质结合位点预测工具，在提交数据后，DeepSite 会在页面中央呈现相应的蛋白质 3D 结构图(橙色部分表示预测的结合口袋)，右侧呈现结合口袋的坐标及打分，COACH 会在页面上给出来自不同算法的

预测结果，页面左侧呈现蛋白质 3D 结构图(绿色部分表示预测的结合口袋)，右侧给出结合口袋的排名、打分、大小、可能的结合配体以及结合位点的残基等信息，用户可以将详细结果下载至本地查看，DoGSiteScorer 会在页面左侧显示蛋白质 3D 结构图(结合口袋用不同颜色表示，其中红色表示可药性高，绿色表示可药性低)，页面右侧以表格形式显示结合口袋的形状尺寸、官能团、口袋原子以及氨基酸分布等信息。综合而言，DeepSite 提供简洁的结果，COACH 提供多种结合位点预测算法供用户选择并有效预测潜在的结合配体以及蛋白质-配体结合模板，DoGSiteScorer 更关注结合口袋和成药性的具体信息，用户可以根据其研究需求选择合适的蛋白质结合位点预测工具。这些预测工具界面友好、操作便捷，使用者不需要编程，计算过程也不需要庞大的算力，极大提升了 PLBA 的研究效率(表 5)。

表 3　基于 AI 的 PLBA 预测模型及其性能

Table 3　Models for predicting PLBA based on AI algorithm and their performances

Year	Model	Algorithm	Performance	Test set
2010	RF-Score[25]	RF[4]	R=0.776	PDBbind v2007[26]
2011	CScore[27]	CMAC[28]	R=0.801	PDBbind v2007[26]
2015	RF-Score v3[29]	RF[4]	R=0.803	PDBbind v2007[26]
2017	PLEIC-SVM[30]	SVM[4]	AUC=0.930	DUD[13]
2017	CNN Scoring[31]	CNN[4]	AUC=0.779 AUC=0.522	ChEMBL[32], MUV[33]
2018	PotentialNet[34]	DNN[34]	R=0.822	PDBbindv2007[26]
2021	IGN[35]	GNN[4]	R=0.837	PDBbind v2016[36]
2019	OnionNet[37]	CNN[4]	R=0.816	PDBbind v2016[36]
2021	ECIF-GBT[38]	GBT[4]	R=0.866	CSAF-2016[14]
2021	OnionNet-2[39]	CNN[4]	R=0.864	PDBbind v2016[36]
2021	OPRC-GBT[40]	GBT[4]	R=0.835	PDBbind v2016[36]
2022	PLA-MoRe[41]	GIN[41] Transformer[4]	CI=0.886 CI=0.874	Davis[41], KIBA[41]
2022	3D-RISM-AI[42]	XGB[42]	R=0.800	PDBbind v2016[36]
2022	MRBDTA[43]	Transformer[42]	CI=0.901 CI=0.892	Davis[41], KIBA[41]
2023	SS-GNN[44]	GNN[4]	R=0.832	PDBbind v2016[36]
2023	PPS-ML[45]	GBT[4]	R=0.843	PDBbind v2016[36]
2023	CurvAGN[46]	GNN[4]	R=0.831	PDBbind v2016[36]
2023	GPCNDTA[47]	GNN[4]	CI=0.903 CI=0.907	Davis[41], KIBA[41]
2023	GraphScoreDTA[48]	GNN[4]	R=0.831	CASF-2016[14]
2023	PLANET[49]	GNN[4]	R=0.811	CASF-2016[14]
2023	3DProtDTA[50]	GNN[4]	CI=0.909 CI=0.858	Davis[41], KIBA[41]
2023	NHGNN-DTA[51]	GNN[4]	CI=0.914 CI=0.907	Davis[41], KIBA[41]

Note: CI stands for concordance index.

表 4　基于 AI 的 PLBA 预测模型的评估指标

Table 4　Evaluation indicators of models used to predict PLBA based on AI algorithm

Task	Evaluation indicator	Introduction
VS	EF[4]	EF primarily focuses on the degree of enrichment of active compounds at specified proportions, aiding in the assessment of the model's efficiency in identifying potential active compounds across the entire compound library. Computed by comparing the ratio of active compounds identified at a threshold to the overall distribution, EF values range from 0 to positive infinity. A higher EF value indicates that the model more effectively enriches active compounds.
	AUC[4]	AUC is a key metric that primarily assesses the overall performance of a model in distinguishing between positive and negative instances across the entire range of classification thresholds. Computed by integrating the receiver operating characteristic (ROC) curve, AUC values range from 0 to 1. A higher AUC value signifies superior screening performance, with 1 indicating perfect discrimination, where all true positives are ranked higher than all true negatives, and 0.5 representing random chance.
Predicting PLBA	R_p[4]	R_p plays a fundamental role in assessing the linear relationship between predicted and actual PLBA, elucidating both the strength and direction of the correlation. Computed by dividing the covariance of predicted and actual binding affinities by the product of their standard deviations, R_p values range from −1 to 1. An R_p value of 1 indicates a perfect positive linear correlation, signifying that as the predicted PLBA rises, the actual PLBA consistently increases. Conversely, an R_p value of −1 suggests a perfect negative linear correlation, denoting a consistent decrease in predicted PLBA as the actual PLBA increases. As R_p approaches 1, it reflects a progressively stronger positive correlation, while an Rp nearing −1 signifies a more robust negative correlation. In contrast, an R_p value close to 0 implies a weak or no linear correlation between the predicted and actual PLBA.
	RMSD[4]	RMSD serves as a critical metric when evaluating a model's predictive accuracy for PLBA. Computed by determining the square root of the average squared differences between predicted and actual PLBA, RMSD values range from 0 to positive infinity. Lower RMSD values indicate a more precise alignment between predicted and actual binding strengths. This implies that the model is effectively capturing the intricate details of protein-ligand interactions. Conversely, higher RMSD values suggest a greater discrepancy, highlighting potential limitations in the model's ability to accurately predict binding affinities.
	CI[41]	CI provides insights into the consistency, strength, and direction of the correlation between predicted and actual PLBA. Computed by dividing the covariance of predicted and actual binding affinities by the product of their standard deviations, CI values fall within the range of −1 to 1. A CI value of 1 indicates a perfect positive linear correlation, signifying that as the predicted PLBA increases, the actual PLBA consistently increases. Conversely, a CI value of −1 suggests a perfect negative linear correlation, denoting a consistent decrease in predicted PLBA as the actual PLBA increases. As CI approaches 1, it reflects a progressively stronger positive correlation, while a CI nearing −1 signifies a more robust negative correlation. In contrast, a CI value close to 0 implies a weak or no linear correlation between the predicted and actual PLBA.

2　基于 AI 的 PLBA 预测模型的应用场景

2.1　基于 AI 的 PLBA 预测模型助力药物发现

药物研发从立项到开始销售平均需要 12 年，成本在 28 亿美元左右[72-73]，其中药物发现需要花费 5−6 年时间。为了缩短研发周期、降低研发成本、提升发现药物候选化合物的效率，AI 算法已广泛应用于多种类型的药物研发项目[74-75](图 2)，其中，研究人员开发了诸多基于 AI 的 PLBA 预测模型以探索药物与靶点之间的相互作用(drug-target interaction)，如表 3 提到的

MRBDTA、GPCNDTA、GraphScoreDTA 以及 3DPRotDTA 等，计算药物与靶点的结合亲和力可以有效排除脱靶可能性大的配体，低成本、高效率地筛选出与靶点结合亲和力高的先导化合物，提升药物研发效率[76]。

近年来，从疾病机制到临床研究等生物医学数据呈现出爆炸式增长，以疾病种类为切入点统计了 2010−2023 年美国食品和药物管理局(Food and Drug Administration, FDA)[77]批准的作用于已知靶点的新分子实体(new molecular entity, NME)药物数量(图 3)，可见针对肿瘤学(Oncology)、感染(infection)、神经学(Central Nervous System, CNS)方面的药物研究数量最多。在这些药物研究过程中，基于 AI 的 PLBA 预测模型不仅能在海量化合物库中筛选潜在抑制剂，还能通过神经网络捕获关键残基，进而推断蛋白靶点与配体药物的作用机理。Chen 等[78]用来自 Atomwise[79]公司的 AtomNet 从 400 万种化合物中确定了一种非小细胞肺癌的潜在抑制剂，还利用该模型识别出分子的氢键、芳香度和单碳键等重要化学基团。Krishnan 等[80]开发了基于深度学习的从头药物设计方法，该方法不仅

表 5　研究蛋白质-配体相互作用的预测工具

Table 5　Predictive tools to assist in the study of protein-ligand interactions

Tool	Description	Case	Link
Taba[59]	Taba is a supervised machine learning-based scoring function for protein-ligand binding affinity	Taba assists in studying the interaction between cyclin-dependent kinase 2 (CDK2) and related compounds[66]	http://www.taba.bio.br/
CSatDTA[60]	CSatDTA is an online tool based on CNN for predicting the binding affinity of protein-ligand interactions	−	http://nsclbio.jbnu.ac.kr/tools/CSatDTA/
MANORAA[61]	MANORAA is a machine learning platform to guide protein-ligand design and affinity calculation by anchors and influential distances	MANORAA assists in studying the interaction between cystic fibrosis transmembrane conductance regulator (CFTR) and genistein[67]	http://manoraa.org
gRINN[62]	gRINN is a software tool for residue interaction-energy based investigation of protein molecular dynamics simulations	gRINN assists in studying the interaction between ebolavirus interferon antagonist VP35 and ubiquitin (Ub)[68]	http://grinn.readthedocs.io
DeepSite[63]	DeepSite is a protein binding site prediction tool based on deep convolutional neural networks (CNN), designed to capture distinctive features of binding sites	DeepSite assists in studying the interaction between acetylcholinesterase (AChE) and its potential inhibitors[69]	https://www.playmolecule.com/deepsite/
COACH[64]	COACH is an online tool for predicting protein binding sites that combines the binding-specific substructure algorithm TM-SITE with the sequence-profile alignment algorithm S-SITE	COACH assists in studying the structure and function of alkaline phosphatase (ALP)[70]	https://seq2fun.dcmb.med.umich.edu//COACH/
DoGSiteScorer[65]	DoGSiteScorer is an online tool for predicting potential binding pockets and sub-pockets of proteins. This tool also analyzes the geometric and physicochemical properties of these pockets and estimates the druggability of drugs	DoGSiteScorer assists in studying the interaction between receptor of glycation end product (RAGE) and aldose reductase (ALR2)[71]	https://proteins.plus/

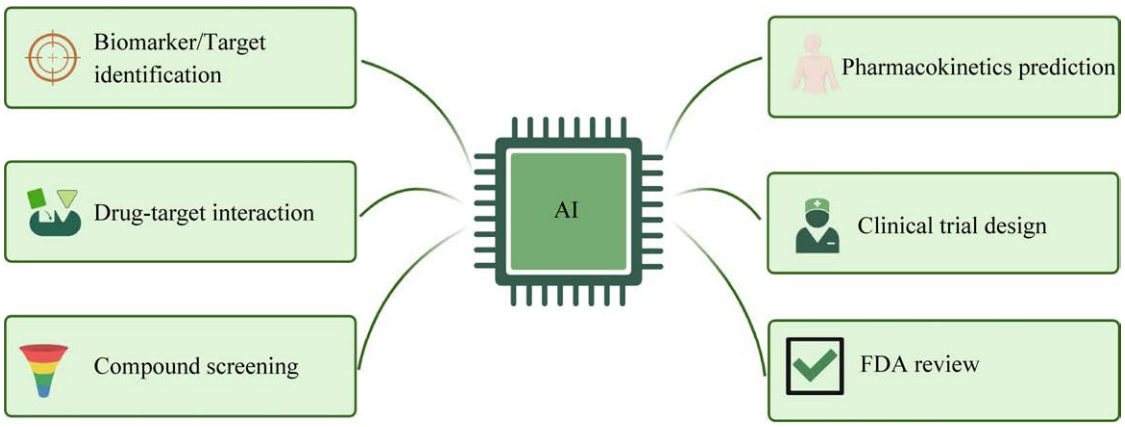

图 2　AI 算法应用到药物研发多个领域

Figure 2　AI algorithm is used in many areas of drug development.

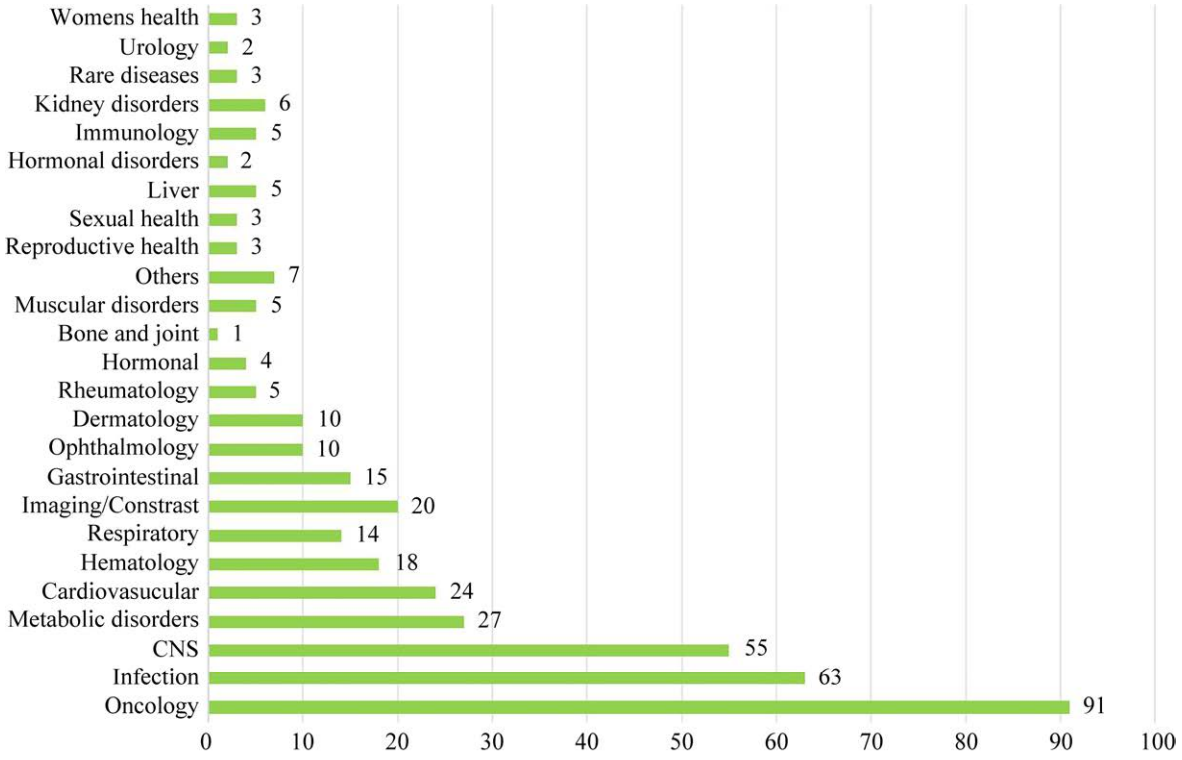

图 3　2010–2023 年 FDA 批准的 NME 药物分类汇总　数据来自 FDA 官网(https://www.fda.gov/)

Figure 3　Summary of NME drug classifications approved by FDA from 2010 to 2023. The data is sourced from FDA official website (https://www.fda.gov/).

可以根据靶点蛋白生成候选分子，还可以根据 ML 算法预测的 PLBA 对候选分子进行排名和过滤，该方法对蛋白靶点 JAK2 和 DRD2 的 10 000 个小分子抑制剂进行采样，最终分别筛选到 30 个和 80 个满足要求的新分子，这种根据靶点蛋白设计候选分子的方法进一

步缩短了候选药物分子的筛选时间，推动了药物研发进程(表 6)。此外，德睿智药与西湖大学、厦门大学科研团队[89]共同开发基于 DL 的预测模型 ProTMD，该模型引入蛋白质动态时空信息，在药物-蛋白质结合亲和力预测任务中的性能优于其他模型，是辅助药物化学专家筛选出高活性小分子的有力工具。

2019 年新型冠状病毒疫情暴发，AI 模型被应用于 SARS-CoV-2 病毒蛋白靶点与配体药物相互作用的研究，如 Choi 等[90]利用 MT-DTI[91]筛选出 20 种潜在抑制剂，Anwaar 等[92]利用 DeepDTA[93]和分子对接[94]技术筛选出 49 种针对不同 SARS-CoV-2 病毒蛋白的 FDA 批准药物，促进了相关药物的再利用，Jablonský 等[95]利用基于 XGB[95]的模型和分子对接技术筛选出 4 种蛋白靶点 3CLpro 的潜在抑制剂，

表 6 利用基于 AI 的 PLBA 预测模型研究靶点案例

Table 6 Cases studies on targets using models for predicting PLBA based on AI algorithm

Disease	Target	Research progress
Oncology	Cyclin-dependent kinase2 (CDK2)	The model identified key residues involved in the interaction between CDK2 and inhibitors[66]
Oncology	Ovarian tumor protease 7B (OTUD7B)	The model is capable of sifting through 4 million of compound libraries to identify small molecule 7B with significant inhibitory activity against the protein target OTUD7B[78-79]
Oncology	Janus kinase2 (JAK2)	The model screened and identified 30 potential inhibitors from a compound library containing thousands of compounds[80]
CNS	Dopamine receptor D2 (DRD2)	The model screened and identified 80 potential inhibitors from a compound library containing thousands of compounds[80]
Infection	3-dehydroquinic acid dehydratase (DHQD)	The model identified key residues and electrostatic interactions involved in the interaction between DHQD and inhibitors[81]
Infection	HIV-1 protease	The model excels in predicting the high-precision affinity between HIV protease and inhibitors, concurrently identifying differences in affinity between wild-type and mutant proteases and inhibitors[82]
Oncology	Epidermal growth factor receptor (EGFR)	Utilizing a combination of methods, the model pinpointed 360 potential inhibitors for the L858R/T790M/C797S mutant EGFR from a selection of 1.76 million compounds[83]
Oncology	NIMA kinase 7 (NEK7)	Incorporating molecular docking and other techniques, the model sieved through over 1 200 compounds, identifying potential inhibitors with higher affinity than the FDA-approved drug dabrafenib[84]
Oncology	Microtubule affinity-regulating kinase 3 (MARK3)	The model captured the essential residues of the protein target MARK3 and the inhibitor nilotinib, enabling the visualization of critical interactions[85]
CNS	Cannabinoid receptor 2 (CB2R)	The model enables precise design of potential CB2R inhibitors, serving as a guide for drug development targeting CB2R. Additionally, it offers an online platform[86]
Oncology	Phosphatidylinositol 5-phosphate 4-kinase (PI5P4Ks)	The model is capable of discerning the crucial residues and functional groups implicated in the interaction between the protein target PI5P4Ks and inhibitors[87]
Oncology	Leukocyte tyrosine kinase (LTK)	The model, in conjunction with experiments, facilitated the screening of 1 highly inhibitory compound from a pool of 1.77 million molecular compounds. Furthermore, it revealed the indispensable hydrogen bonding interaction of LTK with the inhibitor[88]

Williams 等[96]将 OnionNet-SFCT 与基于多层感知器(multilayer perceptron, MLP)的预测模型结合，发现突变 使药物 Ly-CoV555 与抗体 B.1.1.529 的结合亲和力降低了 99.5%，该数据已得到临床证实。上述案例均印证了基于 AI 的 PLBA 预测模型辅助筛选抑制剂的可行性，相信在 AI 的推动下，这些先导化合物将以更短的时间和更低的成本得以上市。

2.2 AI 预测方法推动酶工程

米氏常数 K_m 描述了酶对特定底物的结合亲和力，是酶动力学和细胞生物学研究的重要参数之一[97]，具体而言，K_m 值表示在酶-底物系统中酶活性达到峰值一半时的底物浓度，K_m 值越小，表示酶对底物的结合亲和力越高，而 K_m 值越大，表示结合亲和力越低。2006 年 Borger 等[98]结合 AI 算法和交叉验证(cross validation)技术构建 K_m 预测模型，该模型主要针对同一种酶对不同底物的反应情况。为设计出通用的 K_m 预测模型，2021 年 Kroll 等[97]将酶的 UniRep[99]向量与其底物的不同分子指纹相结合，使用基于 DL 的预测模型为 47 个基因组规模的代谢模型进行了完整的 K_m 预测，包括人、小鼠、酿酒酵母和大肠杆菌，该模型从基因组角度分析酶动力学参数，加强了代谢物浓度与细胞生物学的联系。2022 年 Maeda 等[100]结合 ML 算法和参数优化策略构建了 MLAGO 预测模型，该模型不仅在较小的计算量下准确地预测了酶的 K_m 值，还为研究人员提供了 web 应用程序，可访问 https://sites. google.com/view/kazuhiro-maeda/software-tools-web-apps。此外，Yan 等[101]开发了基于纤维二糖和 β-葡萄糖苷酶的 K_m 预测模型，该模型预测精度较高，可以帮助研究人员更深入地理解酶与纤维二糖的亲和性和催化特性，研究 β-葡萄糖苷酶及其底物相互作用对于相关生物能源和生物燃料的生产具有重要意义。

此外，Goldman 等[102]通过基于结构的池化策略来构建酶与底物相互作用的预测模型，该模型可以准确预测酶的底物特异性，其性能在发现激酶抑制剂的任务中优于基线模型。Mou 等[103]利用基于 ML 的预测模型捕获蛋白质与配体相互作用的关键残基以及配体和活性位点的亲疏水性，最终准确预测出脂肪族、芳香族和芳基脂肪族腈酶的底物范围。AI 模型预测酶的催化性质、底物特异性以及底物范围有助于深入理解酶催化机理、充分挖掘酶的催化潜能，大大提高了成功获取高效催化剂的可能性，为以酶催化生产为主的生物制造和社会经济可持续发展增添新动力。

3 结论

本文综述了 AI 在 PLBA 预测中的应用，介绍了基于 AI 的 PLBA 预测模型的构建流程以及辅助研究蛋白质配体相互作用的线上平台，讨论了 AI 预测模型的应用场景。近年来，AI 融合药物研发、酶设计等领域取得了令人振奋的成果，但仍然面临着蛋白质和配体的特征考虑不全面、数据质量不高以及模型可解释性差等挑战。在后续的研究工作中上述问题有望得到改善，在数据方面，研究人员可以尝试选择偏差较小的 LIT-PCBA[104]数据集或 TocoDDB[105]数据集训练网络；在模型方面，研究人员可以尝试在模型中引入注意力(attention)[106]机制使模型更加关注蛋白质和配体中与亲和力相关的重要区域，或采用梯度加权类激活映射(gradient-weighted class activation mapping, Grad-CAM)[107]技术可视化神经网络捕获到的蛋白质配体相互作用关键区域，进而提高 AI 预测模型的解释性。

相信随着 ML 和 DL 算法的改进、并行计算能力的提升和数据的丰富，AI 方法将进一步融合生物医学、生物化学以及生物学等基础科学，推动药物研发和酶工程的发展，为人类健康作出更大贡献。

REFERENCES

[1] JONES DS, SILVERMAN AP, COCHRAN JR. Developing therapeutic proteins by engineering ligand-receptor interactions[J]. Trends in Biotechnology, 2008, 26(9): 498-505.

[2] LUCHINAT E, BARBIERI L, CREMONINI M, PENNESTRI M, NOCENTINI A, SUPURAN CT, BANCI L. Determination of intracellular protein-ligand binding affinity by competition binding in-cell NMR[J]. Acta Crystallographica Section D Structural Biology, 2021, 77(10): 1270-1281.

[3] BOYLES F, DEANE CM, MORRIS GM. Learning from the ligand: using ligand-based features to improve binding affinity prediction[J]. Bioinformatics, 2020, 36(3): 758-764.

[4] KEVIN C, ALEXIS G, MYRTILLE D, STÉPHANIE B, LUIZ AS. Machine-learning methods for ligand-protein molecular docking[J]. Drug Discovery Today, 2021, 27(1): 151-164.

[5] VEERAMALAI M, GILBERT D. A novel method for comparing topological models of protein structures enhanced with ligand information[J]. Bioinformatics, 2008, 24(23): 2698-2705.

[6] CHEN P, SHEN HM, ZHANG YZ, WANG B, GU PY. SGNet: sequence-based convolution and ligand graph network for protein binding affinity prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(5): 3257-3266.

[7] KAIRYS V, BARANAUSKIENE L, KAZLAUSKIENE M, MATULIS D, KAZLAUSKAS E. Binding affinity in drug design: experimental and computational techniques[J]. Expert Opinion on Drug Discovery, 2019, 14(8): 755-768.

[8] 卞佳豪, 杨广宇. 人工智能辅助的蛋白质工程[J]. 合成生物学, 2022, 3(3): 429-444.
BIAN JH, YANG GY. Artificial intelligence-assisted protein engineering[J]. Synthetic Biology Journal, 2022, 3(3): 429-444 (in Chinese).

[9] VAMATHEVAN J, CLARK D, CZODROWSKI P, DUNHAM I, FERRAN E, LEE G, LI B, MADABHUSHI A, SHAH P, SPITZER M, ZHAO SR. Applications of machine learning in drug discovery and development[J]. Nature Reviews Drug Discovery, 2019, 18(6): 463-477.

[10] YANG Y, UROLAGIN S, NIROULA A, DING XS, SHEN BR, VIHINEN M. PON-tstab: protein variant stability predictor. importance of training data quality[J]. International Journal of Molecular Sciences, 2018, 19(4): 1009.

[11] YU LA, WANG SY, LAI KK. An integrated data preparation scheme for neural network data analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 217-230.

[12] WANG RX, FANG XL, LU YP, YANG CY, WANG SM. The PDBbind database: methodologies and updates[J]. Journal of Medicinal Chemistry, 2005, 48(12): 4111-4119.

[13] CHEESERIGHT TJ, MACKEY MD, MELVILLE JL, VINTER JG. FieldScreen: virtual screening using molecular fields. application to the DUD data set[J]. Journal of Chemical Information and Modeling, 2008, 48(11): 2108-2117.

[14] SU MY, YANG QF, DU Y, FENG GQ, LIU ZH, LI Y, WANG RX. Comparative assessment of scoring functions: the CASF-2016 update[J]. Journal of Chemical Information and Modeling, 2019, 59(2): 895-913.

[15] KIM S, CHEN J, CHENG TJ, GINDULYTE A, HE J, HE SQ, LI QL, SHOEMAKER BA, THIESSEN PA, YU B, ZASLAVSKY L, ZHANG J, BOLTON EE. PubChem in 2021: new data content and improved web interfaces[J]. Nucleic Acids Research, 2021, 49(D1): D1388-D1395.

[16] WISHART DS, KNOX C, GUO AC, CHENG DA, SHRIVASTAVA S, TZUR D, GAUTAM B, HASSANALI M. DrugBank: a knowledgebase for drugs, drug actions and drug targets[J]. Nucleic Acids Research, 2008, 36(database issue): D901-D906.

[17] APWEILER R. UniProt: the universal protein knowledgebase[J]. Nucleic Acids Research, 2004, 32(90001): 115D-119.

[18] BURLEY SK, BERMAN HM, KLEYWEGT GJ, MARKLEY JL, NAKAMURA H, VELANKAR S. Protein data bank (PDB): the single global macromolecular structure archive[J]. Methods in Molecular Biology (Clifton, N J), 2017, 1607: 627-641.

[19] SCHOMBURG I, JESKE L, ULBRICH M, PLACZEK S, CHANG A, SCHOMBURG D. The BRENDA enzyme information system-from a database to an expert system[J]. Journal of Biotechnology, 2017, 261: 194-206.

[20] LIU TQ, LIN Y, WEN X, JORISSEN RN, GILSON MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities[J]. Nucleic Acids Research, 2007, 35(suppl_1): D198-D201.

[21] YANG KK, WU Z, ARNOLD FH. Machine-learning-guided directed evolution for protein engineering[J]. Nature Methods, 2019, 16(8): 687-694.

[22] CHENG Y, GONG YS, LIU YS, SONG BS, ZOU Q. Molecular design in drug discovery: a comprehensive review of deep generative models[J]. Briefings in Bioinformatics, 2021, 22(6): bbab344.

[23] YANGJB, CAI YY, ZHAO KR, XIE HB, CHEN XJ.

Concepts and applications of chemical fingerprint for hit and lead screening[J]. Drug Discovery Today, 2022, 27(11): 103356.
[24] AN X, CHEN X, YI DY, LI HY, GUAN YF. Representation of molecules for drug response prediction[J]. Briefings in Bioinformatics, 2022, 23(1): bbab393.
[25] BALLESTER PJ, MITCHELL JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking[J]. Bioinformatics, 2010, 26(9): 1169-1175.
[26] CHENG TJ, LI X, LI Y, LIU ZH, WANG RX. Comparative assessment of scoring functions on a diverse test set[J]. Journal of Chemical Information and Modeling, 2009, 49(4): 1079-1093.
[27] OUYANG XC, DANIEL HANDOKO S, KWOH CK. Cscore: a simple yet effective scoring function for protein-ligand binding affinity prediction using modified cmac learning architecture[J]. Journal of Bioinformatics and Computational Biology, 2011, 9(supp01): 1-14.
[28] ALBUS JS. A new approach to manipulator control: the cerebellar model articulation controller (CMAC)[J]. Journal of Dynamic Systems, Measurement, and Control, 1975, 97(3): 220-227.
[29] LI HJ, LEUNG KS, WONG MH, BALLESTER PJ. Improving AutoDock vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets[J]. Molecular Informatics, 2015, 34(2/3): 115-126.
[30] YAN YN, WANG WJ, SUN ZX, ZHANG JZH, JI CG. Protein-ligand empirical interaction components for virtual screening[J]. Journal of Chemical Information and Modeling, 2017, 57(8): 1793-1806.
[31] RAGOZA M, HOCHULI J, IDROBO E, SUNSERI J, KOES DR. Protein-ligand scoring with convolutional neural networks[J]. Journal of Chemical Information and Modeling, 2017, 57(4): 942-957.
[32] GAULTON A, BELLIS LJ, BENTO AP, CHAMBERS J, DAVIES M, HERSEY A, LIGHT Y, McGLINCHEY S, MICHALOVICH D, AL-LAZIKANI B, OVERINGTON JP. ChEMBL: a large-scale bioactivity database for drug discovery[J]. Nucleic Acids Research, 2012, 40(D1): D1100-D1107.
[33] ROHRER SG, BAUMANN K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data[J]. Journal of Chemical Information and Modeling, 2009, 49(2): 169-184.
[34] FEINBERG EN, SUR D, WU ZQ, HUSIC BE, MAI HH, LI Y, SUN SS, YANG JY, RAMSUNDAR B, PANDE VS. PotentialNet for molecular property prediction[J]. ACS Central Science, 2018, 4(11): 1520-1530.
[35] JIANG DJ, HSIEH CY, WU ZX, KANG Y, WANG JK, WANG EC, LIAO B, SHEN C, XU L, WU J, CAO DS, HOU TJ. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions[J]. Journal of Medicinal Chemistry, 2021, 64(24): 18209-18232.
[36] LIU ZH, SU MY, HAN L, LIU J, YANG QF, LI Y, WANG RX. Forging the basis for developing protein-ligand interaction scoring functions[J]. Accounts of Chemical Research, 2017, 50(2): 302-309.
[37] ZHENG LZ, FAN JR, MU YG. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction[J]. ACS Omega, 2019, 4(14): 15956-15965.
[38] SÁNCHEZ-CRUZ N, MEDINA-FRANCO JL, MESTRES J, BARRIL X. Extended connectivity interaction features: improving binding affinity prediction through chemical description[J]. Bioinformatics, 2021, 37(10): 1376-1382.
[39] WANG ZC, ZHENG LZ, LIU Y, QU YY, LI YQ, ZHAO MW, MU YG, LI WF. OnionNet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells[J]. Frontiers in Chemistry, 2021, 9: 753002.
[40] WEE J, XIAKL. Ollivier persistent ricci curvature-based machine learning for the protein-ligand binding affinity prediction[J]. Journal of Chemical Information and Modeling, 2021, 61(4): 1617-1626.
[41] LI QY, ZHANG XC, WU LL, BO XC, HE S, WANG SQ. PLA-MoRe: a protein-ligand binding affinity prediction model *via* comprehensive molecular representations[J]. Journal of Chemical Information and Modeling, 2022, 62(18): 4380-4390.
[42] OSAKI K, EKIMOTO T, YAMANE T, IKEGUCHI M. 3D-RISM-AI: a machine learning approach to predict protein-ligand binding affinity using 3D-RISM[J]. The Journal of Physical Chemistry B, 2022, 126(33): 6148-6158.
[43] ZHANG L, WANG CC, CHEN X. Predicting drug-target binding affinity through molecule representation block based on multi-head attention and skip connection[J]. Briefings in Bioinformatics, 2022, 23(6): bbac468.
[44] ZHANG SK, JIN YZ, LIU TM, WANG Q, ZHANG ZH, ZHAO SL, SHAN B. SS-GNN: a simple-structured graph neural network for affinity prediction[J]. ACS Omega, 2023, 8(25): 22496-22507.
[45] LIU R, LIU X, WU J. Persistent path-spectral (PPS) based machine learning for protein-ligand binding

affinity prediction[J]. Journal of Chemical Information and Modeling, 2023, 63(3): 1066-1075.
[46] WU JQ, CHEN HY, CHENG MH, XIONG HY. CurvAGN: curvature-based adaptive graph neural networks for predicting protein-ligand binding affinity[J]. BMC Bioinformatics, 2023, 24(1): 378.
[47] ZHANG L, WANG CC, ZHANG Y, CHEN X. GPCNDTA: prediction of drug-target binding affinity through cross-attention networks augmented with graph features and pharmacophores[J]. Computers in Biology and Medicine, 2023, 166:107512.
[48] WANG KL, ZHOU RY, TANG J, LI M. GraphscoreDTA: optimized graph neural network for protein-ligand binding affinity prediction[J]. Bioinformatics, 2023, 39(6): btad340.
[49] ZHANG XY, GAO HT, WANG HJ, CHEN ZH, ZHANG Z, CHEN XC, LI Y, QI YF, WANG RX. PLANET: a multi-objective graph neural network model for protein-ligand binding affinity prediction[J]. Journal of Chemical Information and Modeling, 2023: online ahead of print.
[50] VOITSITSKYI T, STRATIICHUK R, KOLEIEV I, POPRYHO L, OSTROVSKY Z, HENITSOI P, KHROPACHOV I, VOZNIAK V, ZHYTAR R, NECHEPURENKO D, YESYLEVSKYY S, NAFIIEV A, STAROSYLA S. 3DProtDTA: a deep learning model for drug-target affinity prediction based on residue-level protein graphs[J]. RSC Advances, 2023, 13(15): 10261-10272.
[51] HE HH, CHEN GX, CHEN CYC. NHGNN-DTA: a node-adaptive hybrid graph neural network for interpretable drug-target binding affinity prediction[J]. Bioinformatics, 2023, 39(6): btad355.
[52] DUNBAR JB Jr, SMITH RD, YANG CY, UNG PMU, LEXA KW, KHAZANOV NA, STUCKEY JA, WANG SM, CARLSON HA. CSAR benchmark exercise of 2010: selection of the protein-ligand complexes[J]. Journal of Chemical Information and Modeling, 2011, 51(9): 2036-2046.
[53] TROTT O, OLSON AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading[J]. Journal of Computational Chemistry, 2010, 31(2): 455-461.
[54] LI Y, SU MY, LIU ZH, LI J, LIU J, HAN L, WANG RX. Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark[J]. Nature Protocols, 2018, 13(4): 666-680.
[55] CHATTERJEE A, WALTERS R, SHAFI Z, AHMED OS, SEBEK M, GYSI D, YU R, ELIASSI-RAD T, BARABÁSI AL, MENICHETTI G. Improving the generalizability of protein-ligand binding predictions with AI-Bind[J]. Nature Communications, 2023, 14: 1989.
[56] LeCUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[57] 曲戈, 朱彤, 蒋迎迎, 吴边, 孙周通. 蛋白质工程: 从定向进化到计算设计[J]. 生物工程学报, 2019, 35(10): 1843-1856.
QU G, ZHU T, JIANG YY, WU B, SUN ZT. Protein engineering: from directed evolution to computational design[J]. Chinese Journal of Biotechnology, 2019, 35(10): 1843-1856 (in Chinese).
[58] XU YT, VERMA D, SHERIDAN RP, LIAW A, MA JS, MARSHALL NM, McINTOSH J, SHERER EC, SVETNIK V, JOHNSTON JM. Deep dive into machine learning models for protein engineering[J]. Journal of Chemical Information and Modeling, 2020, 60(6): 2773-2790.
[59] Da SILVA AD, BITENCOURT-FERREIRA G, de AZEVEDO WF Jr. Taba: a tool to analyze the binding affinity[J]. Journal of Computational Chemistry, 2020, 41(1): 69-73.
[60] GHIMIRE A, TAYARA H, XUANZY, CHONG KT. CSatDTA: prediction of drug-target binding affinity using convolution model with self-attention[J]. International Journal of Molecular Sciences, 2022, 23(15): 8453.
[61] TANRAMLUK D, PAKOTIPRAPHA D, PHOOCHAIJAROEN S, CHANTRAVISUT P, THAMPRADID S, VANICHTANANKUL J, NARUPIYAKUL L, AKAVIPAT R, YUVANIYAMA J. MANORAA: a machine learning platform to guide protein-ligand design by anchors and influential distances[J]. Structure, 2022, 30(1): 181-189.e5.
[62] SERÇINOĞLU O, OZBEK P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations[J]. Nucleic Acids Research, 2018, 46(W1): W554-W562.
[63] JIMÉNEZ J, DOERR S, MARTÍNEZ-ROSELL G, ROSE AS, de FABRITIIS G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks[J]. Bioinformatics, 2017, 33(19): 3036-3042.
[64] YANG JY, ROY A, ZHANG Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment[J]. Bioinformatics, 2013, 29(20): 2588-2595.
[65] VOLKAMER A, KUHN D, RIPPMANN F, RAREY M. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment[J]. Bioinformatics, 2012, 28(15): 2074-2075.
[66] BITENCOURT-FERREIRA G, DUARTE Da SILVA A, FILGUEIRA de AZEVEDO W Jr. Application of machine learning techniques to predict binding affinity

for drug targets: a study of cyclin-dependent kinase 2[J]. Current Medicinal Chemistry, 2020, 28(2): 253-265.

[67] RODRAT M, WONGDEE K, TEERAPORNPUNTAKIT J, THONGBUNCHOO J, TANRAMLUK D, AEIMLAPA R, THAMMAYON N, THONAPAN N, WATTANO P, CHAROENPHANDHU N. Vasoactive intestinal peptide and cystic fibrosis transmembrane conductance regulator contribute to the transepithelial calcium transport across intestinal epithelium-like Caco-2 monolayer[J]. PLoS One, 2022, 17(11): e0277096.

[68] RODRÍGUEZ-SALAZAR CA, van TOL S, MAILHOT O, GALDINO G, TERUEL N, ZHANG LH, WARREN AN, GONZÁLEZ-OROZCO M, FREIBERG AN, NAJMANOVICH RJ, GIRALDO MI, RAJSBAUM R. Ebola virus VP35 interacts non-covalently with ubiquitin chains to promote viral replication creating new therapeutic opportunities[J]. BioRxiv: the Preprint Server for Biology, 2023: 2023.07.14.549057.

[69] ALMALKI SG, ALQURASHI YE, ALTURAIKI W, ALMAWASH S, KHAN A, AHMAD P, IQBAL D. Antioxidant, LC-MS analysis, and cholinesterase inhibitory potentials of phoenix dactylifera cultivar khudari: an *in vitro* enzyme kinetics and *in silico* study[J]. Biomolecules, 2023, 13(10): 1474.

[70] SONG X, QIN YG, ZHANG YH, ZHOU YB, CHEN D, XIE DH, LI ZX. Functional characterization of alkaline phosphatases involved alarm pheromone in the vetch aphid Megoura viciae[J]. iScience, 2023, 26(11): 108115.

[71] KAZMI S. Molecular docking analysis of calcium channel blockers with ALR2 and RAGE[J]. Bioinformation, 2023, 19(1): 28-31.

[72] ÁLVAREZ-MACHANCOSESÓ, FERNÁNDEZ-MARTÍNEZ JL. Using artificial intelligence methods to speed up drug discovery[J]. Expert Opinion on Drug Discovery, 2019, 14(8): 769-777.

[73] DANA D, GADHIYA SV, ST SURIN LG, LI D, NAAZ F, ALI Q, PAKA L, YAMIN MA, NARAYAN M, GOLDBERG ID, NARAYAN P. Deep learning in drug discovery and medicine; scratching the surface[J]. Molecules (Basel, Switzerland), 2018, 23(9): 2384.

[74] PUN FW, OZEROV IV, ZHAVORONKOV A. AI-powered therapeutic target discovery[J]. Trends in Pharmacological Sciences, 2023, 44(9): 561-572.

[75] QURESHI R, IRFAN M, GONDAL TM, KHAN S, WU J, HADI MU, HEYMACH J, LE XN, YAN H, ALAM T. AI in drug discovery and its clinical relevance[J]. Heliyon, 2023, 9(7): e17575.

[76] PEI JP, WANG G, FENG L, ZHANG JF, JIANG TT, SUN Q, OUYANG L. Targeting lysosomal degradation pathways: new strategies and techniques for drug discovery[J]. Journal of Medicinal Chemistry, 2021, 64(7): 3493-3507.

[77] HORTON LR. Food and drug administration[J]. BMJ, 2007, 334(7584): 55-56.

[78] CHEN JF, BOLHUIS DL, LAGGNER C, KONG DY, YU L, WANG XD, EMANUELE MJ, BROWN NG, LIU PD. AtomNet-aided OTUD7B inhibitor discovery and validation[J]. Cancers, 2023, 15(2): 517.

[79] WALLACH I, DZAMBA M, HEIFETS A. AtomNet: a deep convolutional neural network for bioactivity prediction in "structure-based drug discovery"[EB/OL]. 2015: arXiv: 1510.02855. http://arxiv.org/abs/1510.02855.pdf"

[80] KRISHNAN SR, BUNG N, VANGALA SR, SRINIVASAN R, BULUSU G, ROY A. *De novo* structure-based drug design using deep learning[J]. Journal of Chemical Information and Modeling, 2022, 62(21): 5100-5109.

[81] de ÁVILA MB, de AZEVEDO WF Jr. Development of machine learning models to predict inhibition of 3-dehydroquinate dehydratase[J]. Chemical Biology & Drug Design, 2018, 92(2): 1468-1474.

[82] LEIDNER F, KURT YILMAZ N, SCHIFFER CA. Target-specific prediction of ligand affinity with structure-based interaction fingerprints[J]. Journal of Chemical Information and Modeling, 2019, 59(9): 3679-3691.

[83] CHOI G, KIM D, OH J. AI-based drug discovery of TKIs targeting L858R/T790M/C797S-mutant EGFR in non-small cell lung cancer[J]. Frontiers in Pharmacology, 2021, 12: 660313.

[84] AZIZ M, EJAZ SA, ZARGAR S, AKHTAR N, ABORODE AT, WANI TA, BATIHA GES, SIDDIQUE F, ALQARNI M, AKINTOLA AA. Deep learning and structure-based virtual screening for drug discovery against NEK7: a novel target for the treatment of cancer[J]. Molecules (Basel, Switzerland), 2022, 27(13): 4098.

[85] YUAN WN, CHEN GX, CHEN CYC. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction[J]. Briefings in Bioinformatics, 2022, 23(1): bbab506.

[86] DELRE P, CONTINO M, ALBERGA D, SAVIANO M, CORRIERO N, MANGIATORDI GF. ALPACA: a machine learning platform for affinity and selectivity profiling of CAnnabinoids receptors modulators[J]. Computers in Biology and Medicine, 2023, 164: 107314.

[87] GIM M, CHOE J, BAEK S, PARK J, LEE C, JU M, LEE SM, KANG J. ArkDTA: attention regularization guided by non-covalent interactions for explainable drug-target binding affinity prediction[J]. Bioinformatics, 2023, 39(supplement_1): i448-i457.

[88] ZHANG XJ, ZHANG O, SHEN C, QU WL, CHEN SC, CAO HQ, KANG Y, WANG Z, WANG EC, ZHANG JT, DENG YF, LIU FR, WANG TY, DU HY, WANG LC, PAN PC, CHEN GY, HSIEH CY, HOU TJ. Efficient and accurate large library ligand docking with KarmaDock[J]. Nature Computational Science, 2023, 3(9): 789-804.

[89] WU F, JIN ST, JIANG YH, JIN XR, TANG BW, NIU ZM, LIU XR, ZHANG Q, ZENG XX, LI SZ. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding[J]. Advanced Science (Weinheim, Baden-Wurttemberg, Germany), 2022, 9(33): e2203796.

[90] CHOI Y, SHIN B, KANG K, PARK S, BECK BR. Target-centered drug repurposing predictions of human angiotensin-converting enzyme 2 (ACE2) and transmembrane protease serine subtype 2 (TMPRSS2) interacting approved drugs for coronavirus disease 2019 (COVID-19) treatment through a drug-target interaction deep learning model[J]. Viruses, 2020, 12(11): 1325.

[91] SHIN B, PARK S, KANG K, HO JC. "Self-attention based molecule representation for predicting drug-target interaction"[EB/OL]. 2019: arXiv: 1908.06760. http://arxiv.org/abs/1908.06760.pdf.

[92] ANWAAR MU, ADNAN F, ABRO A, KHAN RA, REHMAN AU, OSAMA M, RAINVILLE C, KUMAR S, STERNER DE, JAVED S, JAMAL SB, BAIG A, SHABBIR MR, AHSAN W, BUTT TR, ASSIR MZ. Combined deep learning and molecular docking simulations approach identifies potentially effective FDA approved drugs for repurposing against SARS-CoV-2[J]. Computers in Biology and Medicine, 2022, 141: 105049.

[93] ÖZTÜRK H, ÖZGÜR A, OZKIRIMLI E. DeepDTA: deep drug-target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.

[94] PINZI LC, RASTELLI G. Molecular docking: shifting paradigms in drug discovery[J]. International Journal of Molecular Sciences, 2019, 20(18): 4331.

[95] JABLONSKÝ M, ŠTEKLÁČ M, MAJOVÁ V, GALL M, MATÚŠKA J, PITOŇÁK M, BUČINSKÝ L. Molecular docking and machine learning affinity prediction of compounds identified upon softwood bark extraction to the main protease of the SARS-CoV-2 virus[J]. Biophysical Chemistry, 2022, 288: 106854.

[96] WILLIAMS AH, ZHAN CG. Fast prediction of binding affinities of SARS-CoV-2 spike protein and its mutants with antibodies through intermolecular interaction modeling-based machine learning[J]. The Journal of Physical Chemistry B, 2022, 126(28): 5194-5206.

[97] KROLL A, ENGQVIST MKM, HECKMANN D, LERCHER MJ. Deep learning allows genome-scale prediction of Michaelis constants from structural features[J]. PLoS Biology, 2021, 19(10): e3001402.

[98] BORGER S, LIEBERMEISTER W, KLIPP E. Prediction of enzyme kinetic parameters based on statistical learning[J]. Genome Informatics International Conference on Genome Informatics, 2006, 17(1): 80-87.

[99] ALLEY EC, KHIMULYA G, BISWAS S, AlQURAISHI M, CHURCH GM. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16(12): 1315-1322.

[100] MAEDA K, HATAE A, SAKAI Y, BOOGERD FC, KURATA H. MLAGO: machine learning-aided global optimization for Michaelis constant estimation of kinetic modeling[J]. BMC Bioinformatics, 2022, 23(1): 455.

[101] YAN SM, SHI DQ, NONG H, WU G. Predicting *K*m values of beta-glucosidases using cellobiose as substrate[J]. Interdisciplinary Sciences: Computational Life Sciences, 2012, 4(1): 46-53.

[102] GOLDMAN S, DAS R, YANG KK, COLEY CW. Machine learning modeling of family wide enzyme-substrate specificity screens[J]. PLoS Computational Biology, 2022, 18(2): e1009853.

[103] MOU ZY, EAKES J, COOPER CJ, FOSTER CM, STANDAERT RF, PODAR M, DOKTYCZ MJ, PARKS JM. Machine learning-based prediction of enzyme substrate scope: application to bacterial nitrilases[J]. Proteins, 2021, 89(3): 336-347.

[104] TRAN-NGUYEN VK, JACQUEMARD C, ROGNAN D. LIT-PCBA: an unbiased data set for machine learning and virtual screening[J]. Journal of Chemical Information and Modeling, 2020, 60(9): 4263-4273.

[105] ZHANG XJ, SHEN C, WANG TY, KANG Y, LI D, PAN PC, WANG JK, WANG GA, DENG YF, XU L, CAO DS, HOU TJ, WANG Z. Topology-based and conformation-based decoys database: an unbiased online database for training and benchmarking machine-learning scoring functions[J]. Journal of Medicinal Chemistry, 2023, 66(13): 9174-9183.

[106] LI M, LU ZL, WU YF, LI YH. BACPI: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction[J]. Bioinformatics, 2022, 38(7): 1995-2002.

[107] WANGSH, ZHANG YD. Grad-CAM: understanding AI models[J]. Computers, Materials & Continua, 2023, 76(2): 1321-1324.

(本文责编 陈宏宇)