人工智能在药物靶点的筛选及验证方面的应用进展

王 超,肖 辅,李妙竹,潘 颖,丁 晓,任 峰,Zhavoronkov Alex,王亚洲*

(英矽智能科技(上海)有限公司,上海 201203)

摘 要 近年来人工智能发展迅速,随着算力的提升、算法的迭代,人工智能极大方便了生物信息、化学信息和临床数据的收集及处理,为新药研发注入了新的活力。本综述对人工智能在制药领域的发展历程及其主要算法进行了简要介绍,随后结合具体实例对人工智能在药物靶点筛选及验证方面的不同阶段进行了详细描述,包括药物靶点发现、蛋白结构预测以及苗头化合物生成与优化等。最后对人工智能平台"端到端"的一次高效应用过程进行了具体讨论。

关键词 人工智能;靶点发现;蛋白结构预测;苗头化合物生成;端到端应用

中图分类号 TP18;R914.2 **文献标志码** A **文章编号** 1000-5048(2023)03-0269-13

doi:10.11665/j.issn.1000-5048.2023041102

引用本文 王超,肖辅,李妙竹,等.人工智能在药物靶点的筛选及验证方面的应用进展[J].中国药科大学学报,2023,54(3):269-281.

Cite this article as: WANG Chao, XIAO Fu, LI Miaozhu, *et al*. Application progress of artificial intelligence in the screening and identification of drug targets[J]. *J China Pharm Univ*, 2023, 54(3): 269-281.

Application progress of artificial intelligence in the screening and identification of drug targets

WANG Chao, XIAO Fu, LI Miaozhu, PAN Ying, DING Xiao, REN Feng, ZHAVORONKOV Alex, WANG Yazhou*

Insilico Medicine Shanghai Ltd., Shanghai 201203, China

Abstract In recent years, artificial intelligence (AI) has developed rapidly, with improved computing power and algorithms, which has greatly facilitated the collection and processing of biological, chemical information and clinical data, injecting new vitality into the research and development of new drugs. In this review, we began with a brief overview of the development and the main algorithms of AI in drug discovery. Then we elaborated through several specific cases on the various scenarios of AI application, including target identification, protein structure prediction, hit generation and optimization etc. Finally, we focused on a recent example to discuss the high efficiency of "end-to-end" application of AI.

Key words artificial intelligence; target identification; protein structure prediction; hit generation; end-to-end application

1956年达特茅斯会议上,计算机科学家们提出建立具有完全人类智能的新型计算机,从而诞生了人工智能(artificial intelligence,AI)的概念,这是一门用于模拟、延伸和扩展人的智能的新技术科学[1]。AI技术已经在自然语言翻译、语音识别、自动驾驶、医疗、金融等各个领域发挥着越来越重要的作用。而在药物研发领域,AI同样有着极其深入的研究和应用。本文主要对AI在制药领域的发展历程及其算法进行简要介绍,同时对其在药物早期发现阶段的不同应用举例加以说明,希望可以引起科研人员对制药领域AI应用的更多关注和思考。

1 人工智能在制药领域的发展历程

药物研发长久以来一直面临周期长、成本高、成功率低的问题[2]。根据估算,每上市1个药物大

收稿日期 2023-04-11 *通信作者 Tel:13584071909 E-mail:yazhou.wang@insilico.ai

约需要10年时间,研发总成本大约为18亿美元。AI作为新药研发的强力引擎,在全球各大药厂和生物科技公司都受到越来越多的重视。据报道AI技术在化合物合成和筛选方面比传统手段可节约40%～50%的时间,每年为药企节约260亿美元的化合物筛选成本;在临床研究阶段,可节约50%～60%的时间,每年可节约280亿美元的临床试验成本[3]。

自2018年以来,AI在制药领域的发展更是实现了从"0"到"1"的跨越。2018年Heal-X公司通过AI驱动老药新用,发现脆性X综合征候选药物,并在18个月内将项目推进到Ⅱa期临床试验[4]。2019年Deep genomics公司利用AI驱动平台,在18个月内完成肝豆状核变性的全新靶点发现和寡核苷酸候选化合物筛选[5]。同年,Insilico Medicine公司应用基于生成对抗网络神经技术的GENTRL在21 d内完成AI药物发现挑战,生成设计了高活性DDR1激酶抑制剂[6]。2020年,DeepMind公司的AlphaFold完成50年来生物学重大挑战,根据蛋白质的氨基酸序列预测蛋白质的3D结构[7]。2022年2月,Insilico Medicine公司完全基于AI生成的针对特发性肺纤维化的药物进入临床Ⅰ期,这也是仅用时18个月投入260万研发费用发现的候选化合物。该药于2023年2月开始Ⅱ期临床试验,同时获得美国FDA授予的孤儿药资格认定。这是全球首个达到这一研发里程碑的、由AI赋能靶点发现和分子设计的first-in-class在研药物。这一系列里程碑事例均表明AI赋能创新药物研发的强大能力。

2 AI算法简介

数据、算法和算力被认为是AI的三大支柱,持续推动AI领域的发展[8]。机器学习(machine learning, ML)是AI的一种类型,计算机可以自己学习,识别模式然后建立模型,并根据这些模型进行预测;深度学习(deep learning, DL)则是机器学习的一种进阶类型。AI算法可以按照不同的分类标准进行类型划分,例如按照模型训练方式的差异可以分为监督学习(supervised learning, SL)、无监督学习(unsupervised learning, UL),以及强化学习(reinforcement learning, RL),按照模型预测任务的不同可分为分类算法(包括二分类和多分类)、回归算法、聚类算法、降维算法、异常检测算法等(图1)。

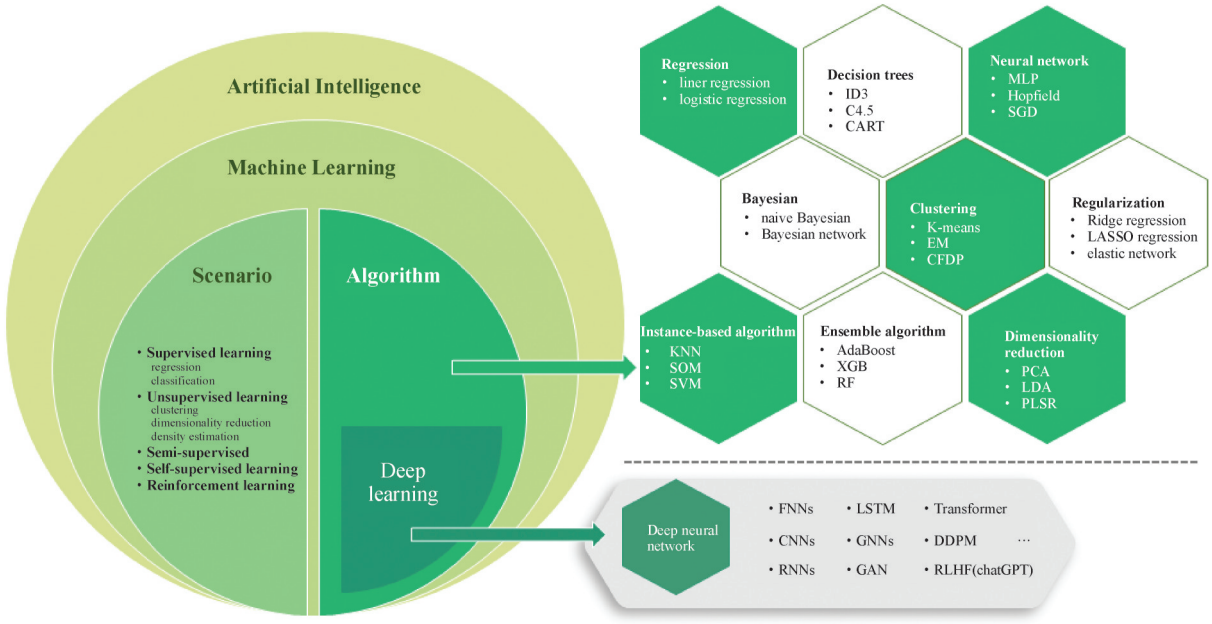

Figure 1 Artificial intelligence, machine learning and deep learning

2.1 监督学习

监督学习是指在带有标签的数据上训练得到数学模型,然后给定新的输入,模型会预测相应的输出值[9]。典型算法包括支持向量机(support vec-

tor machine，SVM）、决策树（decision trees，DT），随机森林（random forest，RF）、朴素贝叶斯（naive bayes，NB）、K-近邻（k-nearest neighbors，KNN）、梯度提升（gradient boosting，GB）、多层感知器（multi-layer perceptron，MLP）、人工神经网络（artificial neural network，ANN）等。

2.2　无监督学习

无监督学习则与监督学习过程相反，在没有标签的数据上训练得到模型[10]。常用的无监督学习算法包括聚类和降维。代表算法包括 K 均值（K-means）、期望最大化（expectation maximization，EM）、主成分分析（principal component analysis，PCA）、线性判别分析（linear discriminant analysis，LDA）、高斯混合模型（gaussian mixture models，GMM）、奇异值分解（singular value decomposition，SVD）、自编码器（autoencoders，AE）等。

2.3　强化学习

与上述学习方式不同，强化学习是另一种机器学习范式，不需要外部提供大量带标签的数据进行训练。强化学习中有两个可交互的对象，智能体（agent）与环境，智能体利用已有动作（action）和经验的反馈不断地与环境进行交互，以实现特定目标或取得最大化的预期利益[11]。

2.4　深度学习

深度学习是机器学习的一个子集，其源于对 ANN 的研究[12]。ANN 是由一系列人工神经元互连构成的网络系统，用来模拟人类大脑神经系统的结构与功能[13]。至今已发展出多种类型的网络架构，包括卷积神经网络（conventional neural networks，CNN）、循环神经网络（recurrent neural networks，RNN）、图卷积神经网络（graph CNN，GCNN）、AE 等。

2022 年 AI 技术驱动的自然语言处理工具 ChatGPT 将大型语言模型（large language model，LLMs）带入了公众视野，这是一种基于大量文本数据的深度学习模型[14]。研究显示 LLMs 处理简化分子线性输入规范（simplified molecular input line entry system，SMILES）字符串库，可产生用于 QSAR 和生成模型的化学语言模型。LLMs 在一些有缺陷的基准上表现出与更广泛使用的技术相当的性能，它们在分子特性预测中的应用还处于起步阶段。

3　AI 在药物早期发现阶段的应用

基于近年来算法的不断发展，AI 也在制药领域发挥着越来越重要的作用，尤其集中于药物早期发现阶段（图 2）。其应用主要包括了药物靶点发现与确证、蛋白结构预测、苗头化合物生成与优化，以及将不同阶段同时进行应用的“端到端”平台。此外，针对逆合成路线设计以及 ADMET 性质预测等也已有很多研究报道。

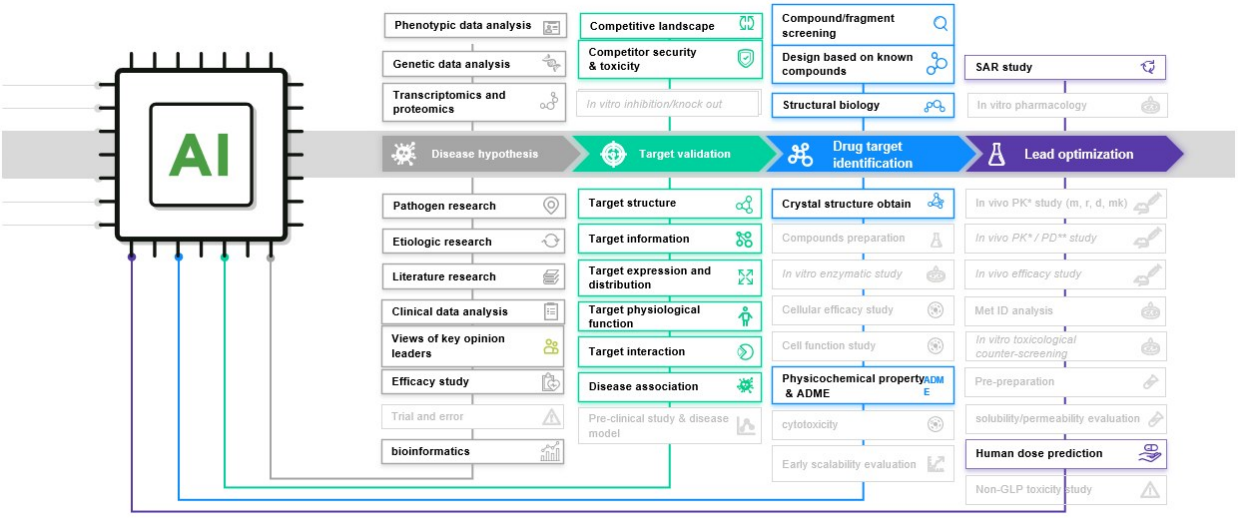

Figure 2　Artificial intelligence empowerment in drug discovery

3.1　药物靶点发现与确证

基于靶点的药物发现是药物研发的主流手段，到 2021 年 9 月，FDA 批准的 1 619 个药物中，共涉及靶点 893 个，其中 667 个为人体靶点（其余为

病原体靶点),药物靶点对整个新药研发项目起到决定性的作用[15]。将系统生物学和AI算法相结合,挖掘多组学数据和患者临床健康信息的关联,联合自然语言处理技术,可以找出潜在的通路、蛋白和机制等与疾病的相关性,以发现新机制和新靶点。

3.1.1 *系统生物学方法* 系统生物学通过研究各个生物系统内部所有组分成分间相互关系,期望最终能够建立整个系统的可理解模型,为有机体绘制完整图谱。AI生物技术公司BERG开发的Interrogative Biology是一个基于AI的系统生物学平台,可生成数据驱动的无偏网络,用于识别靶点和疾病的生物标志物[16]。

将知识图谱技术与系统生物学结合构建生物医药知识图谱已开始在医学实践和研究中发挥关键作用。BenevolentAI公司推出的判断加强认知系统JACS(judgment augmented cognition system)通过发现疾病、药物、试验数据等大量非结构化数据间的新联系,实现药物重定位。2020年,研究人员利用这一AI平台发现经典JAK激酶抑制剂巴瑞替尼(baricitinib)或可用于治疗新型冠状病毒感染[17]。MindRankAI公司参与构建的PharmKG利用异构图注意力神经网络构建药物与疾病之间联系的多关系属性生物医药知识图谱,包含了29种关系种类以及超过8 000种歧义实体[18]。Insilico Medicine公司于自主研发的靶点发现平台内推出基于Transformer的知识图谱功能,从期刊文献中提取信息,将基因、疾病、化合物和生物通路联系起来,并将其与基于大型语言模型的问答功能相结合,以快速识别疾病发展的遗传基础和分子机制,促进药物靶点和生物标志物的识别。

信号通路激活分析则是一种从大规模转录组学和蛋白质组学数据中提取生物学相关特征的强大方法,Insilico Medicine公司的科学家基于信号通路激活分析提出了"通路激活网络分解分析"iPANDA(*in silico* pathway activation network decomposition analysis)方法[19]。iPANDA以疾病患者样本中基因表达水平与正常组样本的平均表达水平之间的倍数变化作为输入,并引入基因重要性因子来表征基因对通路激活的影响程度,进而进行标志信号通路识别(图3)。

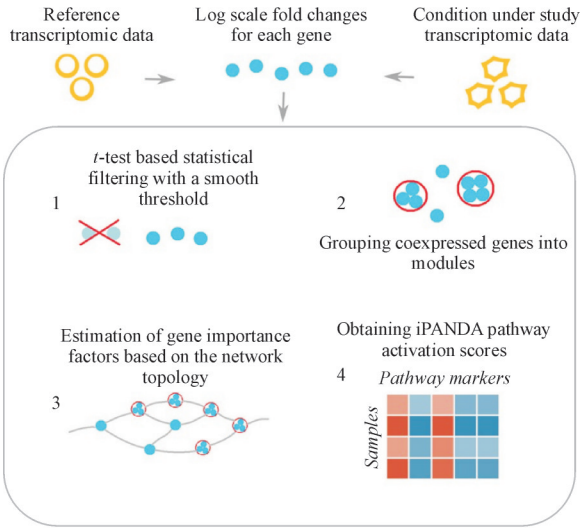

Figure 3 General scheme of iPANDA calculation pipeline[19]

3.1.2 *基于靶点结构的方法* 基于结构的靶点发现计算方法可以作为补充实验方法的策略,例如反向对接、药效团、结合位点相似性和基于指纹交互的方法。其中,反向对接(reverse docking)已经成为确定给定化合物潜在靶点的有效工具之一,不仅用于靶点确证,而且还能预测毒性和不良反应,也可用于发现药物或天然化合物的未知新颖靶点[20]。2008年发布的大规模靶蛋白数据库(potential drug target database, PDTD)涵盖了约1 100个具有3D结构的蛋白质条目,其数据是从文献和几个在线数据库(如TDD、DrugBank和Thomson Pharma)中提取的,包括830个已知或潜在药物靶点的信息,使用该数据库发现了茶多酚和人参皂苷的潜在靶蛋白[21]。2019年杨光富教授团队开发了基于一致性对接方法的药物重定位(老药新用)平台ACID(auto *in silico* consensus inverse docking),用来评估每个蛋白质和给定小分子之间的亲和力,预测准确度提高了10%[22]。

利用药效团模型进行反向找靶也是进行靶标预测的重要方式。李洪林教授团队开发了药效团匹配与潜在识别靶标平台PharmMapper,该平台可通过将所查询化合物的药效团与内部药效团模型数据库进行匹配来执行预测。该课题组在2017年对PharmMapper平台进行了更新,将其药效团数据库进行了6倍规模的扩增。随后通过对这些药效团进行匹配,对抗菌剂卡那霉素(kanamycin)的作用靶点进行了预测[23]。对排名前1%的候选药效团结合实验评估进行筛选,最终选定氨基糖苷磷酸转移酶以及核苷酸转移酶为其作用靶点。

3.2　蛋白结构预测及应用

蛋白质是构成人体细胞、组织的主要物质，参与了从细胞复制到凋亡的绝大多数分子过程，了解复杂蛋白质的功能对于理解生物过程至关重要。蛋白质执行多样性的分子功能是由其精细的三维结构所决定，因此，获取蛋白质精确的3D结构对于洞悉蛋白质发挥功能的具体机制和对其结构与功能的改造极其关键。

半个多世纪以来，研究人员已发展了诸如核磁共振（NMR）、X射线晶体学、冷冻电子显微镜（CryoEM）等多种实验技术来解析蛋白质结构，但基于实验的手段解析蛋白结构耗费大量时间和财力，其测定速度远未满足深层次理解生命现象过程以及复杂药物研发的需求。截至2022年，Uniport数据库中拥有超过2.3亿条蛋白质序列，然而PDB数据库中仅包含约6万个蛋白质的约20万个三维结构，覆盖不到蛋白质的0.1%[24]。

3.2.1　AI预测蛋白结构的算法相关进展　到目前为止，针对蛋白结构预测已有多种算法报道，大致可以分为3类：同源建模、从头建模和基于机器学习的建模[25]。其中同源建模依赖于已知的蛋白质结构；从头建模则仅基于既定的物理定律（量子力学）生成目标蛋白质的3D结构，但受限于自由能的精确计算以及蛋白质的构象空间；基于机器学习的建模，尤其基于深度学习的建模方法是一种数据驱动方法，是最新的新兴方法，其中AlphaFold[7, 26]、RoseTTAFold[27]、ESMFold[28]和Chowdhury[29]等的语言模型最为著名。

为了更好地预测和破解蛋白质三维结构，1994年以来每两年都会在全球范围内举行国际蛋白质结构预测大赛（critical assessment of protein structure prediction，CASP），为参赛者提供测试预测方法的平台。在2018年举办的第13届CASP比赛中，由谷歌DeepMind开发的蛋白结构预测模型AlphaFold大显身手，其预测的43种蛋白质中有25种蛋白质的结构最准确，一举拿下该届比赛冠军[7]。其工作原理主要分两步，第一步是多序列比对以找出最相似的氨基酸序列，并进一步预测每个氨基酸之间的距离矩阵和扭转角等，第二步则会基于氨基酸序列，创造出一个符合物理规则的随机三维结构，然后用深度学习中常用的梯度下降法迭代优化第一步中的预测。

2020年举办的第14届CASP比赛中，升级版AlphaFold2对所有目标蛋白质结构预测的平均GDT得分达到92.4分，其准确性可以与使用冷冻电镜等实验技术解析的3D结构相媲美[26]。作为最先进的蛋白结构预测方法，AlphaFold2采用了端到端（end-to-end）的深度神经网络框架，主要包括输入模块、Evoformer模块和结构模块（图4）。其注意力机制和利用自蒸馏（self-distillation）理念的训练方法，以及端到端的架构、回收机制（recycling approach）、丰富的氨基酸序列和蛋白质结构数据都是促进AlphaFold2成功的重要因素[30]。

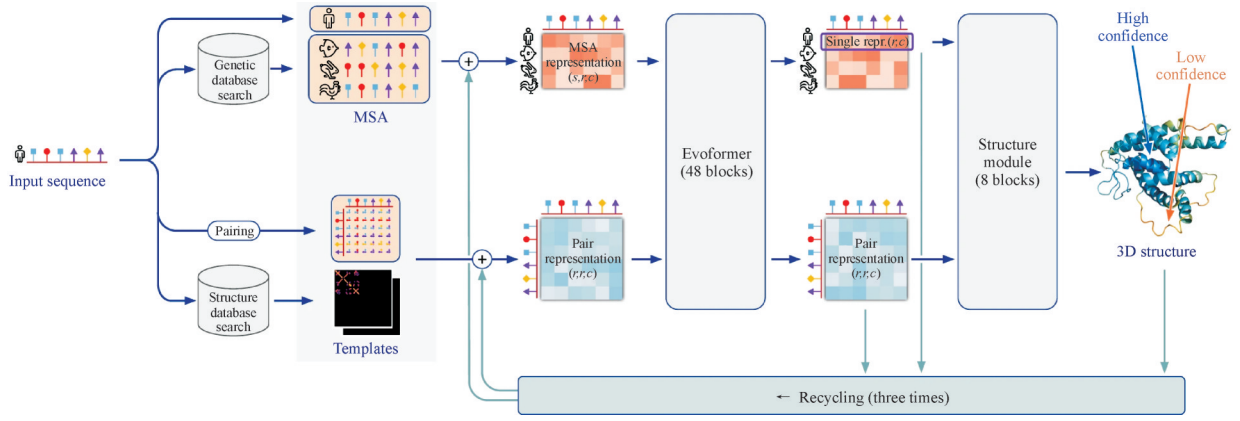

Figure 4　Model architecture of AlphaFold2[26]

尽管AlphaFold的出现成为蛋白质结构预测领域的一个突破，但也应该意识到存在的诸多不足。AlphaFold2对蛋白质无序区域的结构、具有点突变的结构、孤儿蛋白及人工设计蛋白结构、复合物结构、翻译后修饰蛋白结构等面临预测性能低或无法预测的问题，此外模型的弱可解释性和预

测速度较慢也是其局限性之一[31]。

3.2.2 AlphaFold预测蛋白结构的应用——GPR84拮抗剂优化 G-蛋白偶联受体84(G-protein-coupled receptor 84,GPR84)是一种促炎型GPCR受体,可在炎性刺激下高表达而放大机体的免疫反应,被认为与多种炎症或纤维化疾病的发生发展相关。目前已有部分合成的类脂质分子具有一定的GPR84激动活性,但是针对GPR84的拮抗剂仍研究较少(图5)[32]。

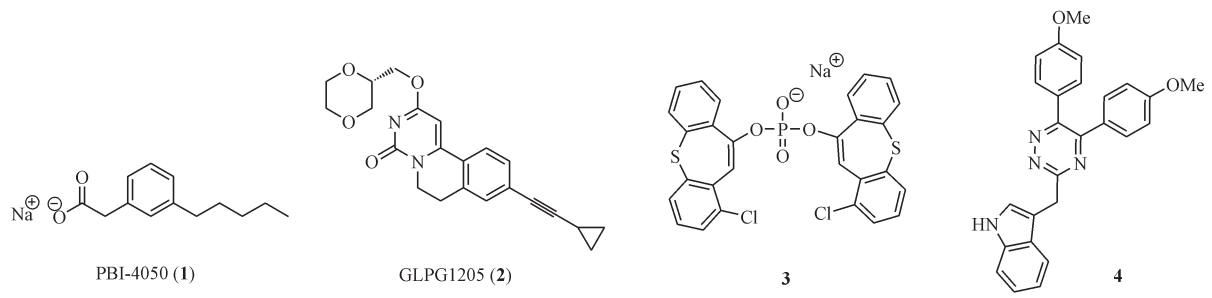

Figure 5 Chemical structures of selected GPR84 antagonists

2022年,Andrew G. Jamieson课题组利用AlphaFold预测的GPR84蛋白结构,针对GPR84拮抗剂进行了一系列构效关系研究[33]。在先前的研究工作中,作者发现苗头化合物**4**是具有高活性及高选择性的GPR84受体拮抗剂。在该预测模型中,由于两个二硫桥键Cys11-Cys166和Cys93-Cys168的形成,限制了细胞外环2(extracellular loop 2,ECL2)区域的构象(图6)。其中三嗪骨架6位的芳基指向细胞外的开放区域,可以接受体积更大的取代基,而5位的芳基取代基则受到ECL2区域构象的限制。

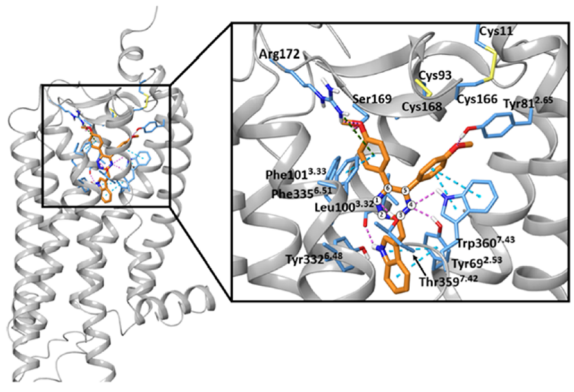

Figure 6 Proposed binding mode of GPR84 antagonist with hGPR84 receptor AlphaFold (AF-Q9NQS5-F1)[33]

基于该结合模式,科研人员重点针对三嗪骨架的5,6位芳基取代基进行了构效探索(图7)。经过一系列化合物筛选后,发现引入吗啉结构的化合物**7**和引入极性酰胺侧链的化合物**8**具有良好的活性和LLE,同时均表现出良好的受体选择性和体外ADME性质。该工作基于AlphaFold预测GPR84蛋白的结合模式进行了详细的构效关系研究,也为后续GRP84拮抗剂的研究提供了基础。

3.3 苗头化合物的生成与优化

由于深度学习的架构非常适合在复杂、非线性的数据集中自动识别相关模式而无须人工特征工程,计算化学家越来越多地采用生成模型来获得新分子并预测他们的特征[34]。基于深度学习方法的优点在于其可以通过模式识别在分子结构数据中生成隐式化学知识,而无须像传统标准从头设计方法依赖于以合成规则或基本物理模型的形式所积累的明确化学知识[35]。化学空间中大约包含1×10^{60}~1×10^{100}种可能的小分子,湿实验研究只能探索其中的极小部分,传统计算机药物设计方法往往仅包含有限数量的片段或采用复杂搜索策略从预定义的化学空间区域采样命中化合物,这限制了在广阔化合物空间中对新分子的有效探索。而基于深度学习的生成模型可以从大量药物数据中学习化学分子结构与其生物和药理性质之间的非线性概率分布,然后对具有所需性质的新生成分子进行设计[36-37],因此可以帮助科学家更有效的探索整个药物类化学空间。

对于生成式建模,目前常用的是基于SMILES[38]的字符串编码生成结构,以及基于图生成建模的图卷积策略网络[39]或深度强化学习[40]来生成分子结构。代表性的研究包括基于SMILES的循环神经网络(recurrent neural networks,RNN)[41-43]、变分自编码器(variational autoencoder,VAE)[44-46],基于

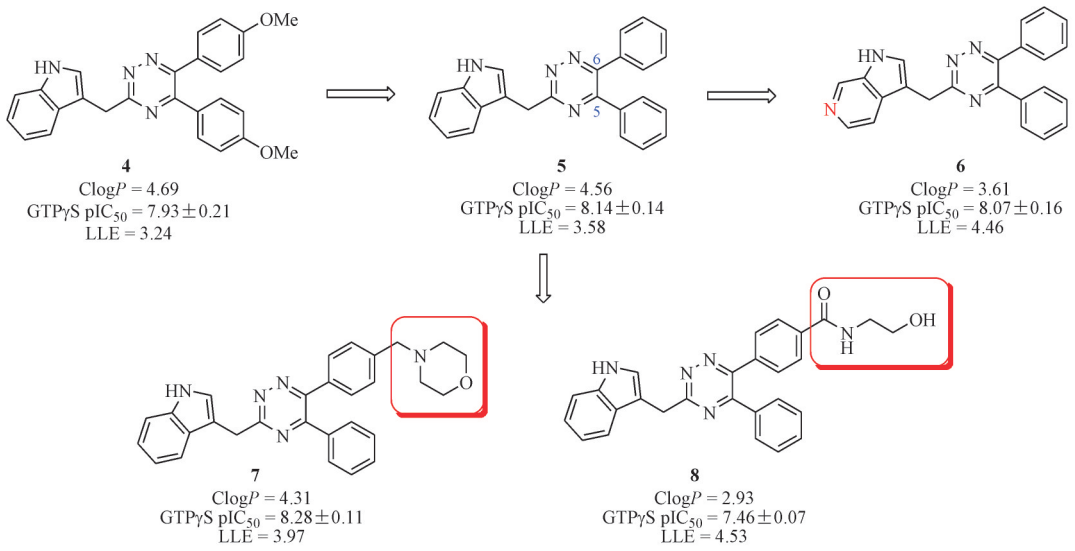

Figure 7　Structure-activity research based on the binding mode

分子图的生成对抗网络（generative adversarial networks，GAN）[47]、目标增强生成对抗网络（objective-reinforced generative adversarial network for inverse-design chemistry，ORGANIC）[48]等。

在针对 VAE 和 GAN 等的改进方案中，又诞生了对抗式自编码器（adversarial autoencoders，AAE）[49]、纠缠条件对抗自编码器（entangled conditional adversarial autoencoder，ECAAE）[50]、双向对抗自动编码器（bidirectional adversarial autoencoder，BiAAE）[51]、高斯混合模型变分自编码器（graph-based variational autoencoder with gaussian mixture hidden space，Graph-GMVAE）[52]、药物生成对抗网络（drugGAN）[53]以及强化对抗神经计算机（reinforced adversarial neural computer，RANC）[54]和生成张量强化学习模型（generative tensorial reinforcement learning，GENTRL）[6]等优化方法。接下来将从几个实例中列举深度学习在苗头化合物生成及优化中的应用。

3.3.1　DDR1 激酶抑制剂的发现——GENTRL 模型的应用　盘状结构域受体 1（discoidin domain receptor 1，DDR1）是一种酪氨酸激酶受体，通过基因敲除以及反义寡核苷酸可以证明 DDR1 在肾纤维化中发挥重要作用。罗氏制药研发团队曾利用 DNA 编码库（DNA encoded library，DEL）筛选出一类具有吲唑结构的螺环化合物，并在针对奥尔波特综合征（Alport syndrome）的 *Col4a3*$^{-/-}$ 小鼠中验证其对肾纤维化具有一定的治疗作用[55]。

2019 年，Insilico Medicine 研发团队利用 GENTRL 模型对 DDR1 受体进行了研究[6]。GENTRL 是一种将强化学习、变分推理和张量分解组合得到的生成式两步机器学习算法，作者首先在 ZINC 数据集上进行预训练，随后在 DDR1 激酶抑制剂和普通激酶抑制剂数据集上进行微调，接着在非激酶抑制剂数据集上进行强化学习。通过警戒结构、反应基团、聚类和多样性排序进行过滤，再使用特异激酶的自组织映射、基于受体结构的药效团模型和 sammon 映射进行评估，得到了具有潜在活性的分子集（图 8）。从该分子集中，根据化学空间和 RMSD 均匀地进行随机挑选，得到 40 个分子，其中 39 个都在专利覆盖范围之外。

在短短 21 d 内针对给定靶点设计出苗头化合物后，作者根据合成难易程度挑选出 6 个化合物进行了合成（图 9）。通过活性测试发现，化合物 **9** 和 **10** 对 DDR1 在酶学和细胞学均表现出良好的抑制活性。在接下来的体外代谢稳定性及体内 PK 测试中，化合物 **9** 均表现出良好的成药性。从利用 GENTRL 进行苗头化合物的生成，到分子的筛选合成并完成初步生物活性测试，整个过程仅耗时 46 d，极大程度地提高了苗头化合物发现的效率。该工作也进一步验证了深度生成模型对于快速发现具有良好生物活性、合成可行性以及成药性分子的作用，有望成为发现药物候选分子的强有力工具。

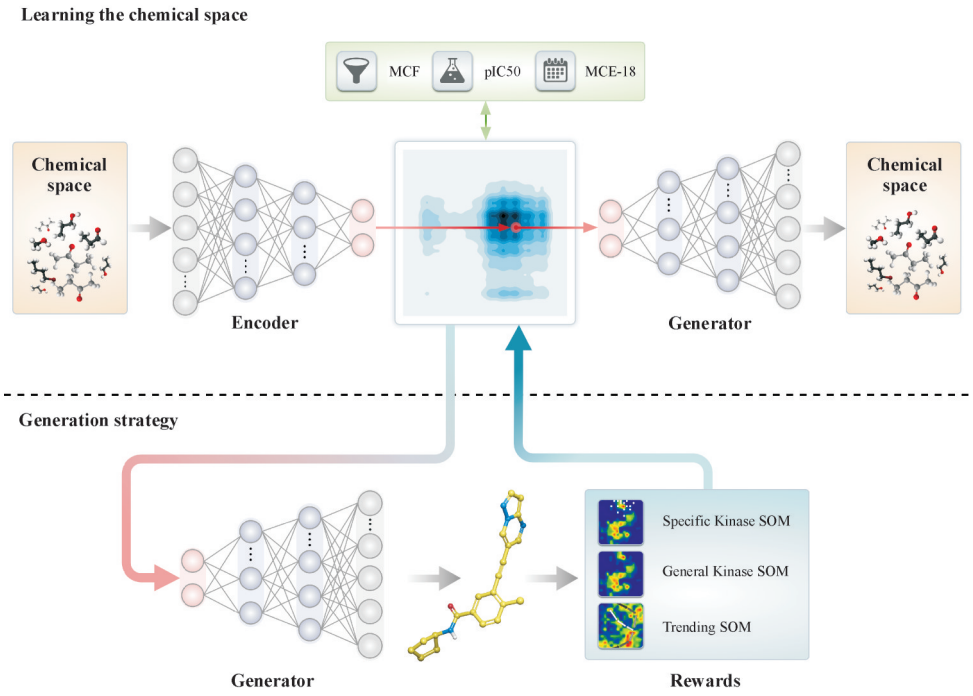

Figure 8 Generative tensorial reinforcement learning model [6]

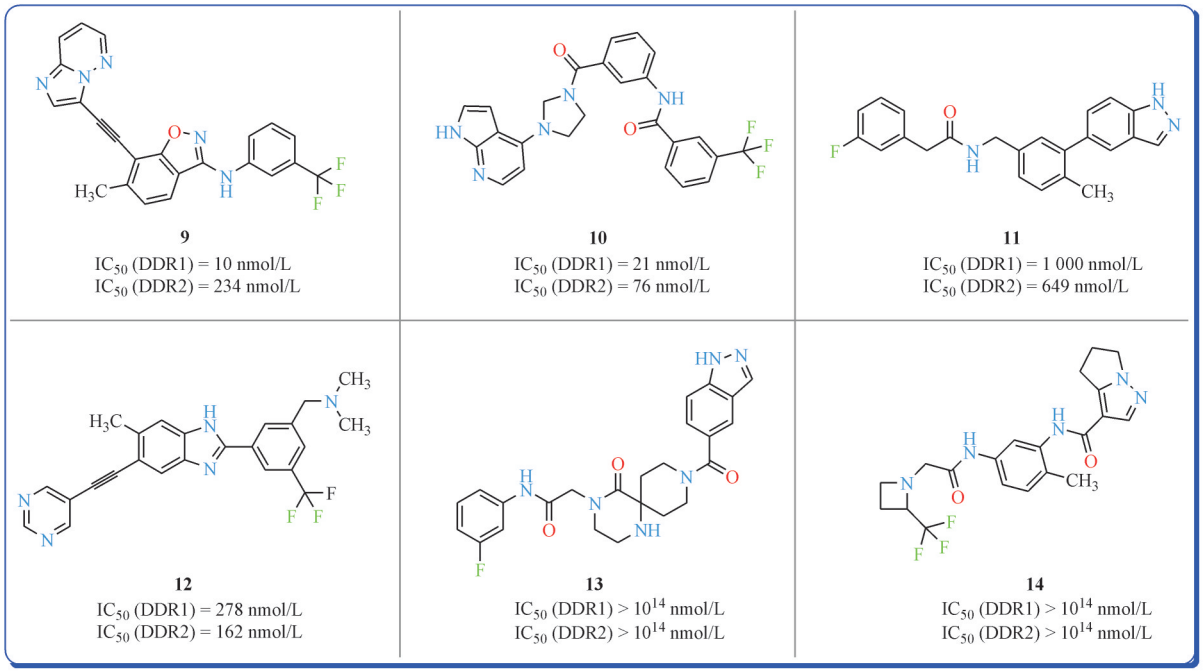

Figure 9 Generated compounds with the highest inhibition activity against human DDR1 kinase

3.3.2 DAO抑制剂的优化——hDAO FEP+的应用 D-氨基酸氧化酶(D-amino acid oxidase,DAO)是一类广泛存在的氧化还原酶,能够催化体内D-氨基酸进行氧化脱氨基,生成相应的α-酮酸。其中D-丝氨酸作为*N*-甲基-D-天冬氨酸受体的协同激动剂,被报道在精神分裂症患者的血清和脑脊液中具有显著降低[56]。目前进入临床阶段的DAO抑制剂仅有SyneuRx公司的苯甲酸钠和Takeda公司的TAK-831,其临床改善认知功能表现仍有待提升。

2022 年来自 Schrödinger 的研发团队利用自由能微扰（free energy perturbation，FEP+）模型以及 AutoDesigner 算法平台，对 DAO 抑制剂进行了优化[57]。AutoDesigner 算法平台首先使用匹配分子对转换、基于反应的枚举、递归结构修剪以及 R 基团修饰等探索化学空间，生成阶段的输出会被漏入中间过滤级联中，在过滤掉剩余分子后，会被作为下一个生成阶段的输入[58]。经组合、去重后得到的最终输出结果会进一步经过 FEP+预测结合活性[59]。

根据已有的抑制剂与 DAO 结合蛋白结构，作者利用 hDAO FEP+模型获得了苗头化合物 **15**，该化合物具有新颖的非酸性的二氢吡嗪二酮结构（图 10）。在进行多个位点的构效关系研究后发现，引入硫醚结构的化合物 **16**、引入并环结构固定分子构象后的化合物 **17** 活性均得到进一步提高。同时针对尾部区域进行延伸位置的子口袋（图 11），以并环化合物 **17** 作为模板，通过 AutoDesigner 进行枚举和筛选，经合成及活性验证后发现化合物 **18** 对 DAO 的 IC_{50} 可达 8 nmol/L。该工作不仅利用自身 AI 平台进行了详尽的构效关系研究，也探索了前人未曾研究过的结合位点，为后续进一步分子优化提供了新的方向。

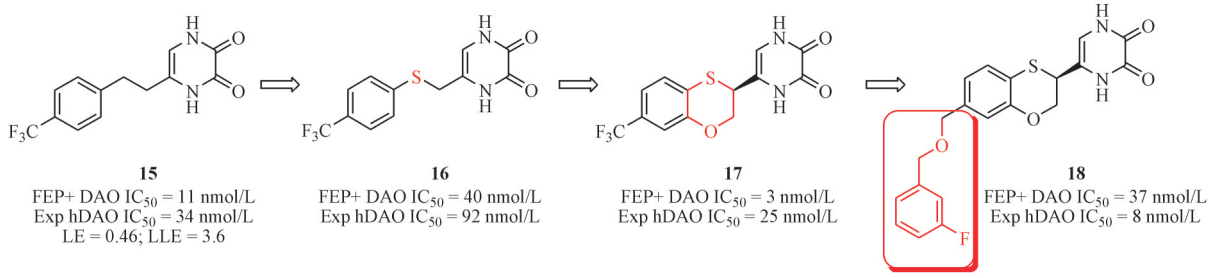

Figure 10　Novel DAO inhibitors identified by hDAO FEP+ model and AutoDesigner algorithm

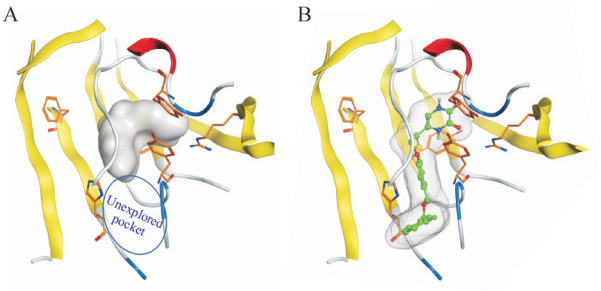

Figure 11　Binding mode of DAO inhibitor **18** (A) Overlay of complex crystal with surface representation (gray) (B) Binding model of new DAO inhibitor with the novel sub-pocket

3.4　AI 平台“端到端”应用——新型 CDK20 抑制剂的发现

在药物早期发现阶段，AI 在某一个环节的应用已经可以极大地提高研发效率，而将多个 AI 平台进行“端到端”应用则可以将 AI 的加持能力发挥到最大水平。2023 年，Insilico Medicine 研发团队瞄准临床急需的肝癌药物，利用自身的生物计算平台 PandaOmics 找到新靶点，借助 AlphaFold 预测蛋白结构，再利用生成化学平台 Chemistry42 寻找苗头化合物结构，结合湿实验验证，开发出了潜在的全球首创 CDK20 抑制剂（图 12）[60]。

Pharma. AI 药物研发平台中，PandaOmics 是一个由 AI 驱动的生物靶点发现引擎，采用先进的多模态深度学习方法，集成 20 多种算法模型、60 多种计算规则，囊括超 1 000 万份组学数据样本、超 34.2 万项临床试验数据、43.3 万个分子的相互作用机制，有能力根据输入数据生成动态结果[61-62]。Chemistry42 平台则是以 GENTRL 模型为核心，集成多种前沿算法模型，同时持续采用奖励机制和 3D 物理结构模块对生成的分子结构进行评估，并在生成算法辅助下进行多维度评分和优化，涵盖药效、代谢稳定性、合成难度等[63]。另外还有一个临床试验结果预测工具 InClinico。该工作以肝细胞癌作为出发点，为了寻找治疗靶点，作者首先获取了包含 1 133 个患者样本和 674 个健康样本的肝细胞癌数据集。利用这些数据，PandaOmics 筛选出 20 个有潜力的靶点，基于靶点与疾病的关联度给出评分，确定了得分最高的靶点——细胞周期蛋白依赖性激酶 20（cyclin-dependent kinase 20，CDK20）[60]。

CDK20 也被称为细胞周期相关激酶（cell cycle-related kinase，CCRK），是 CDK 家族最新鉴定蛋白[64]。尽管针对 CDK 家族其他成员的药物研发已有诸多报道，针对 CDK20 抑制剂的研究工作相

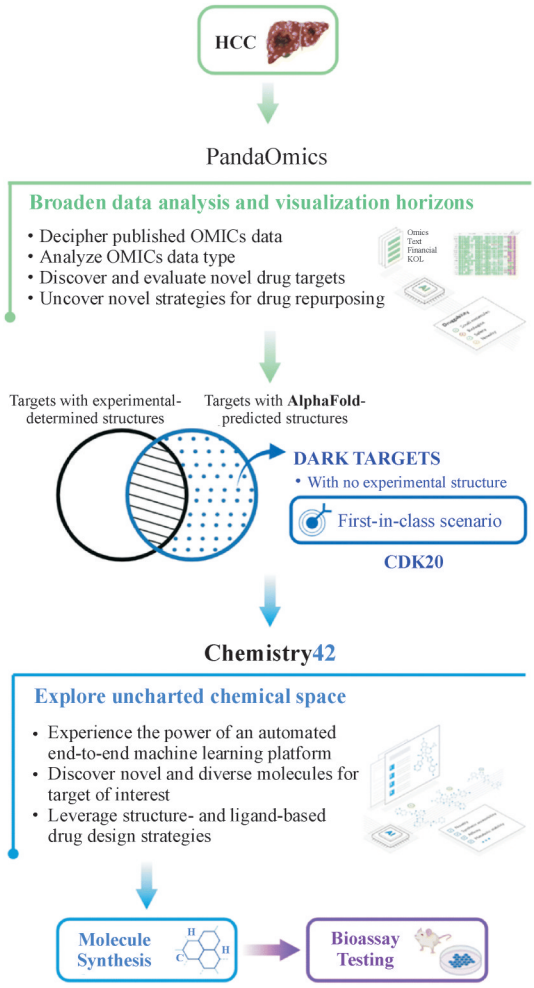

Figure 12 Pipeline to combine AlphaFold with Insilico Medicine end-to end AI-powered PandaOmics and Chemistry42[60]

对较少,一个重要的原因便在于CDK20没有可用的蛋白结构信息。Chemistry42从AlphaFold预测的蛋白结构中发现,CDK20有一个较浅的ATP结合口袋(图13)。把铰链区的Met84作为必须的结合位点,结合这个口袋的结构特性,Chemistry42设计并生成了8 918种分子结构,药化科学家挑选出7个最具潜力的化合物进行合成。

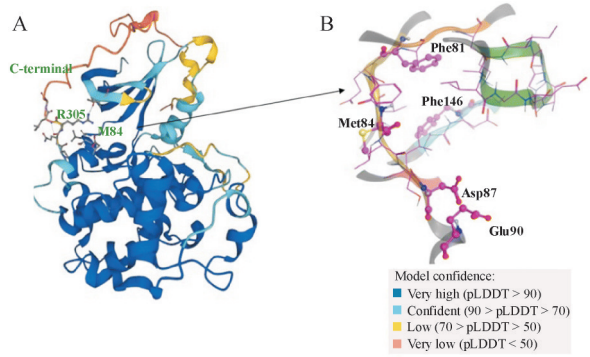

Figure 13 Protein structure and ATP pocket of CDK20

A: AlphaFold predicted structure of CKD20 (AF-Q8IZL9-FL-model_v1); B:ATP pocket of CDK20 with a DFG-in conformation[60]

首轮合成的7个分子中,化合物ISM042-2-001与CDK20表现出良好的结合活性(K_d = 9.2 ± 0.5 μmol/L),仅用时30 d便发现了苗头化合物(图14)。随后通过该化合物与预测蛋白结构的结合模式,作者用Chemisry42进行了第2轮的化合物生成,重点针对苯并嘧啶骨架引入官能团占据"门控开关"区域疏水空间。其中化合物ISM042-2-048的结合活性增强了15倍,在CDK20激酶活性测试中IC_{50}可达33.4 nmol/L。以该化合物进行结合模式验证可以发现,新增加的吡唑环与Lys33形成额外的氢键相互作用,也进一步解释了其活性增强的原因。

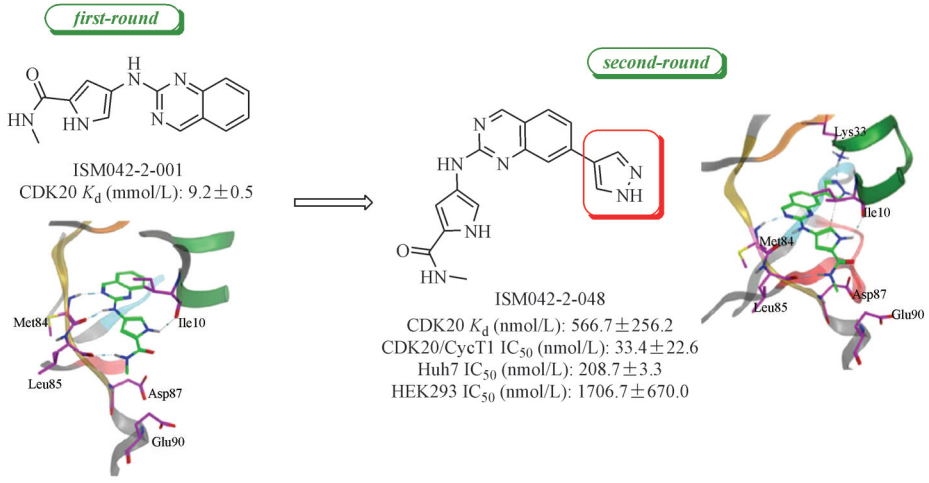

Figure 14 Chemical structures for the selected molecules from Chemistry42 generation

在结构新颖性方面，将化合物 ISM042-2-048 与已报道的 CDK20 抑制剂进行结构比对，该分子与其他化合物均具有较低的相似性。这也是全球首个利用 AlphaFold 预测蛋白结构，针对全新靶点生成全新苗头化合物的实例，突显了 AI“端到端”平台在药物分子发现过程中的速度、效率和准确性。利用该平台针对其他靶点如 GPCR、E3 连接酶的研究工作也在进行中，相信在不久的将来，AI“端到端”平台应用将会为药物研发工作带来更多的奇迹。

4　展　望

新药的研发成本和研发周期不断提高，而成功率却不断下降，很多疾病仍然得不到有效治疗，缺乏真正合适的靶点。AI 技术的发展无疑为新药研发的困境带来强劲动力[65]。一方面 AI 具有通量大、效率高的特点，能够处理海量数据，可以更快缩小搜索范围，包括生物信息，临床数据，化学信息等；另一方面，AI 无偏见的学习和算法的快速迭代优化，能够提高研究人员的“认识带宽”，突破固有的定性思维，减少寻找新靶点并加以验证的时间和成本，尽量避免盲目试错。然而，AI 在该领域的应用当前仍面临如下一些问题。

首先，AI 模型的表现是以数据为基石的，系统全面的高质量数据对于 AI 模型的发展至关重要。尽管目前已进入大数据时代，但对制药领域而言，依旧面临数据量少、数据体系不完整、数据标准不统一、数据共享机制不完善等问题，数据来源的差异、格式的差异等均会带来数据本身的准确性问题。同时，医药研发的数据作为药物研发企业的核心资产，涉及到知识产权，不会轻易对外公开，公开的数据中也会埋没大量的阴性数据。随着未来 AI 在药物研发过程中优质数据的不断积累以及联邦算法等 AI 模型的推广，相信数据的问题会得到进一步改善。

此外，针对药物研发的 AI 算法模型大多只纳入化学指标，生物学指标尚不完整，因此这些算法的落地也是集中于解决药物早期发现阶段的问题，尤其针对苗头化合物的生成、优化以及活性预测等方面。对于研发成本更加巨大的后期开发，包括候选化合物成药性优化，药理、毒理、安全性、有效性等临床前试验综合评价以及临床试验的评价，由于存在种属差异和个体差异，后期数据公开不多且质量不一，评价参数多且复杂等问题，导致某些结果预测困难，AI 仍存在较大的进步空间。

相信随着 AI 技术以及生物医药技术的不断记录和完善，为解决临床需求而诞生的创新药物会越来越多。AI 药物研发的未来也会不断随着数据、算法以及人才的发展，从已经跨越的“0”到“1”，进而发展到“1”到“N”，真正为解决人类健康问题提供切实可行的高效手段。

References

[1] Yang X, Wang YF, Byrne R, *et al*. Concepts of artificial intelligence for computer-assisted drug discovery[J]. *Chem Rev*, 2019, **119**(18): 10520-10594.

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

[3] Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects[J]. *Drug Discov Today*, 2019, **24**(3): 773-780.

[4] Pushpakom S, Iorio F, Eyers PA, *et al*. Drug repurposing: progress, challenges and recommendations[J]. *Nat Rev Drug Discov*, 2019, **18**(1): 41-58.

[5] Merico D, Spickett C, O′Hara M, *et al*. *ATP7B* variant c.1934T > G p.Met645Arg causes Wilson disease by promoting exon 6 skipping[J]. *NPJ Genom Med*, 2020, **5**: 16.

[6] Zhavoronkov A, Ivanenkov YA, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors [J]. *Nat Biotechnol*, 2019, **37**(9): 1038-1040.

[7] Senior AW, Evans R, Jumper J, *et al*. Improved protein structure prediction using potentials from deep learning[J]. *Nature*, 2020, **577**(7792): 706-710.

[8] Huawei T Technologies Co Ltd. *A general introduction to artificial intelligence*[M]//Artificial Intelligence Technology. Singapore: Springer Nature Singapore, 2022: 1-41.

[9] Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development[J]. *Nat Rev Drug Discov*, 2019, **18**(6): 463-477.

[10] Vijayan RSK, Kihlberg J, Cross JB, *et al*. Enhancing preclinical drug discovery with artificial intelligence[J]. *Drug Discov Today*, 2022, **27**(4): 967-984.

[11] Nag S, Baidya ATK, Mandal A, *et al*. Deep learning tools for advancing drug discovery and development[J]. *3 Biotech*, 2022, **12**(5): 110.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015, **521**(7553): 436-444.

[13] McCulloch WS, Pitts W. A logical calculus of the ideas imma-

nent in nervous activity[J]. *Bull Math Biol*, 1990, **52**(1/2): 99-115.

[14] Zhao WX, Zhou K, Li j, *et al*. A survey of large language models [J].*arXiv*,2023:2303.18223.

[15] Santos R, Ursu O, Gaulton A, *et al*. A comprehensive map of molecular drug targets[J]. *Nat Rev Drug Discov*, 2017, **16**(1): 19-34.

[16] Narain N, Kiebish M, Vishnudas V, *et al*. CSAO-1. Interrogative Biology: Unraveling insights into causal disease drivers by use of a dynamic systems biology and Bayesian AI to identify the intersect of disease and healthy signatures[J]. *Neuro Oncol Adv*, 2021, **3**(Supplement_2): ii1.

[17] Richardson P, Griffin I, Tucker C, *et al*. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease[J]. *Lancet*, 2020, **395**(10223): e30-e31.

[18] Zheng SJ, Rao JH, Song Y, *et al*. PharmKG: a dedicated knowledge graph benchmark for biomedical data mining[J]. *Brief Bioinform*, 2021, **22**(4): bbaa344.

[19] Ozerov IV, Lezhnina KV, Izumchenko E, *et al*. In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development[J]. *Nat Commun*, 2016, **7**(1): 1-11.

[20] Lee A, Lee K, Kim D. Using reverse docking for target identification and its applications for drug discovery[J]. *Expert Opin Drug Discov*, 2016, **11**(7): 707-715.

[21] Gao ZT, Li HL, Zhang HL, *et al*. PDTD: a web-accessible protein database for drug target identification[J]. *BMC Bioinformatics*, 2008, **9**: 104.

[22] Wang F, Wu FX, Li CZ, *et al*. ACID: a free tool for drug repurposing using consensus inverse docking strategy[J]. *J Cheminform*, 2019, **11**(1): 73.

[23] Wang X, Shen YH, Wang SW, *et al*. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database[J]. *Nucleic Acids Res*, 2017, **45**(W1): W356-W360.

[24] Dana JM, Gutmanas A, Tyagi N, *et al*. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins[J]. *Nucleic Acids Res*, 2019, **47**(D1): D482-D489.

[25] AlQuraishi M. Machine learning in protein structure prediction [J]. *Curr Opin Chem Biol*, 2021, **65**: 1-8.

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

[27] Baek M, DiMaio F, Anishchenko I, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network[J]. *Science*, 2021, **373**(6557): 871-876.

[28] Lin ZM, 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.

[29] Chowdhury R, Bouatta N, Biswas S, *et al*. Single-sequence protein structure prediction using a language model and deep learning[J]. *Nat Biotechnol*, 2022, **40**(11): 1617-1623.

[30] Xie QZ, Luong MT, Hovy E, *et al*. Self-training with noisy student improves ImageNet classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10684-10695.

[31] Yang ZY, Zeng XX, Zhao Y, *et al*. AlphaFold2 and its applications in the fields of biology and medicine[J]. *Signal Transduct Target Ther*, 2023, **8**(1): 115.

[32] Marsango S, Barki N, Jenkins L, *et al*. Therapeutic validation of an orphan G protein-coupled receptor: the case of GPR84[J]. *Br J Pharmacol*, 2022, **179**(14): 3529-3541.

[33] Mahindra A, Jenkins L, Marsango S, *et al*. Investigating the structure-activity relationship of 1, 2, 4-triazine G-protein-coupled receptor 84 (GPR84) antagonists[J]. *J Med Chem*, 2022, **65**(16): 11270-11290.

[34] Schneider P, Schneider G. *De novo design at the edge of chaos [J]. J Med Chem*, 2016, **59**(9): 4077-4086.

[35] Vanhaelen Q, Lin YC, Zhavoronkov A. The advent of generative chemistry[J]. *ACS Med Chem Lett*, 2020, **11**(8): 1496-1505.

[36] Zhavoronkov A. Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry[J]. *Mol Pharmaceutics*, 2018, **15**(10): 4311-4313.

[37] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering[J]. *Science*, 2018, **361**(6400): 360-365.

[38] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. *J Chem Inf Comput Sci*, 1988, **28**(1): 31-36.

[39] You JX, Liu BW, Ying R, *et al*. Graph convolutional policy network for goal-directed molecular graph generation[J]. *arXiv*, 2018: 1806.02473.

[40] Zhou ZP, Kearnes S, Li L, *et al*. Optimization of molecules via deep reinforcement learning[J]. *Sci Rep*, 2019, **9**(1): 10752.

[41] Gupta A, Müller AT, Huisman BJH, *et al*. Generative recurrent networks for *de novo* drug design[J]. *Mol Inform*, 2018, **37**(1/2): 1700111.

[42] Segler MHS, Kogej T, Tyrchan C, *et al*. Generating focused molecule libraries for drug discovery with recurrent neural networks [J]. *ACS Cent Sci*, 2018, **4**(1): 120-131.

[43] Olivecrona M, Blaschke T, Engkvist O, *et al*. Molecular de-novo design through deep reinforcement learning[J]. *J Cheminform*, 2017, **9**(1): 48.

[44] Kingma DP, Welling M. Auto-encoding variational Bayes[J]. *arXiv*,2013:1312.6114.

[45] Gómez-Bombarelli R, Wei JN, Duvenaud D, *et al*. Automatic chemical design using a data-driven continuous representation of molecules[J]. *ACS Cent Sci*, 2018, **4**(2): 268-276.

[46] Blaschke T, Olivecrona M, Engkvist O, *et al*. Application of generative autoencoder in *de novo* molecular design[J]. *Mol Inform*, 2018, **37**(1/2): 1700123.

[47] De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs[J]. *arXiv*,2018: 1805.11973.

[48] Sanchez-Lengeling B, Outeiral C, Guimaraes GL, *et al*. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC)[J]. *ChemRxiv*, 2017:5309668.v3.

[49] Kadurin A, Aliper A, Kazennov A, *et al*. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology[J]. *Oncotarget*, 2017, **8**(7): 10883-10890.

[50] Polykovskiy D, Zhebrak A, Vetrov D, *et al*. Entangled conditional adversarial autoencoder for *de novo* drug discovery[J]. *Mol Pharmaceutics*, 2018, **15**(10): 4398-4405.

[51] Shayakhmetov R, Kuznetsov M, Zhebrak A, *et al*. Molecular generation for desired transcriptome changes with adversarial autoencoders[J]. *Front Pharmacol*, 2020, **11**: 269.

[52] Yu Y, Xu TY, Li JW, *et al*. A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors[J]. *ACS Omega*, 2021, **6**(35): 22945-22954.

[53] Kadurin A, Nikolenko S, Khrabrov K, *et al*. druGAN: an advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties in silico[J]. *Mol Pharmaceutics*, 2017, **14**(9): 3098-3104.

[54] Putin E, Asadulaev A, Ivanenkov Y, *et al*. Reinforced adversarial neural computer for *de novo* molecular design[J]. *J Chem Inf Model*, 2018, **58**(6): 1194-1204.

[55] Richter H, Satz AL, Bedoucha M, *et al*. DNA-encoded library-derived DDR1 inhibitor prevents fibrosis and renal function loss in a genetic mouse model of alport syndrome[J]. *ACS Chem Biol*, 2019, **14**(1): 37-49.

[56] Verrall L, Burnet PWJ, Betts JF, *et al*. The neurobiology of D-amino acid oxidase and its involvement in schizophrenia[J]. *Mol Psychiatry*, 2010, **15**(2): 122-137.

[57] Tang HF, Jensen K, Houang E, *et al*. Discovery of a novel class of d-amino acid oxidase inhibitors using the Schrödinger computational platform[J]. *J Med Chem*, 2022, **65**(9): 6775-6802.

[58] Bos PH, Houang EM, Ranalli F, *et al*. AutoDesigner, a *De novo* design algorithm for rapidly exploring large chemical space for lead optimization: application to the design and synthesis of d-amino acid oxidase inhibitors[J]. *J Chem Inf Model*, 2022, **62**(8): 1905-1915.

[59] Beuming T, Martín H, Díaz-Rovira AM, *et al*. Are deep learning structural models sufficiently accurate for free-energy calculations? application of FEP+ to AlphaFold2-predicted structures [J]. *J Chem Inf Model*, 2022, **62**(18): 4351-4360.

[60] Ren F, Ding X, Zheng M, *et al*. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor[J]. *Chem Sci*, 2023, **14**(6): 1443-1452.

[61] Pun FW, Liu BHM, Long X, *et al*. Identification of therapeutic targets for amyotrophic lateral sclerosis using PandaOmics-an AI-enabled biological target discovery platform[J]. *Front Aging Neurosci*, 2022, **14**: 914017.

[62] Pun FW, Leung GHD, Leung HW, *et al*. Hallmarks of aging-based dual-purpose disease and age-associated targets predicted using PandaOmics AI-powered discovery engine[J]. *Aging*, 2022, **14**(6): 2475-2506.

[63] Ivanenkov YA, Polykovskiy D, Bezrukov D, *et al*. Chemistry42: an AI-driven platform for molecular design and optimization[J]. *J Chem Inf Model*, 2023, **63**(3): 695-701.

[64] Mok MT, Zhou JY, Tang WS, *et al*. CCRK is a novel signalling hub exploitable in cancer immunotherapy[J]. *Pharmacol Ther*, 2018, **186**: 138-151.

[65] Chen W, Liu XS, Zhang SY, *et al*. Artificial intelligence for drug discovery: resources, methods, and applications[J]. *Mol Ther Nucleic Acids*, 2023, **31**: 691-702.

[**专家介绍**] 王亚洲,博士,高级工程师,英矽智能科技(上海)有限公司药物化学副总监。此前在南京大学、江苏先声药业、南京圣和药业从事新药发现研究工作,历任项目经理、药物化学部部长、小分子新药发现项目总监。主持江苏省博士后科研基金,国家自然科学基金委青年基金及企业重点课题,在抗肿瘤、抗感染及自身免疫疾病领域已获授权发明专利20余件,在*J Med Chem*,*Eur J Med Chem* 和 *Org Lett* 等期刊发表论文20余篇,领导团队获得4个候选化合物,其中2个进入临床开发阶段。2011年博士毕业于中国科学院化学研究所,曾获江苏省"企业博士集聚计划"人才项目资助。