人工智能在新药发现中的应用进展

黄芳[1]，杨红飞[1*]，朱迅[2**]

（1. 杭州费尔斯通科技有限公司，浙江 杭州 310051；2. 吉林大学基础医学院，吉林 长春 130021）

[摘要] 人工智能在新药研发领域中发挥着至关重要的作用。目前，自然语言处理、机器学习、深度学习、知识图谱等人工智能关键技术已广泛应用于新药研发的各个环节，全球多家人工智能企业与制药企业也开启了深度合作模式，为生物医药的发展带来了新的机遇。介绍了机器学习方法和深度学习方法在新药发现领域的应用进展及相关企业，并总结了人工智能应用于新药发现的机遇与挑战，旨在为从事人工智能 + 新药研发工作的科研技术人员提供思路与参考。

[关键词] 人工智能；大数据；机器学习；深度学习；药物研发；靶点发现；药物筛选

[中图分类号] TP18 [文献标志码] A [文章编号] 1001-5094（2021）07-0502-10

Progress in the Application of Artificial Intelligence in New Drug Discovery

HUANG Fang[1], YANG Hongfei[1], ZHU Xun[2]

(1. Hangzhou Firestone Technology Co., Ltd., Hangzhou 310051, China; 2. College of Basic Medical Sciences, Jilin University, Changchun 130021, China)

[Abstract] Artificial intelligence plays an important role in drug research and development. At present, the key artificial intelligence technologies such as natural language processing, machine learning, deep learning and knowledge mapping have been widely used in the whole process of new drug research. A number of enterprises around the world in the field of artificial intelligence have started their deep cooperation with the pharmaceutical industry, creating new opportunities for the development of the biomedical field. This paper introduces the application of machine learning and deep learning methods for drug development in some enterprises, and summarizes the opportunities and challenges artificial intelligence faces in its application in new drug discovery in order to provide some insightful reference for relevant scientific researchers in their adoption of artificial intelligence in the field of new drug research and development.

[Key words] artificial intelligence; big data; machine learning; deep learning; drug research and development; target discovery; drug screening

众所周知，一款新药从研发到上市平均需要花费 10 年以上的时间以及投入高昂的资金，然而仅有 10% 的新药能被批准进入临床研究，最终只有更小比例的药物分子获批上市。曾有投资人将新药“从实验室进入临床试验阶段”形容为“死亡之谷”。

人工智能（artificial intelligence，AI）现在还处于起步阶段。AI 起初被大规模应用于医疗影像，然后逐渐渗透到药物研发领域。近年来，越来越多的 AI 企业投资 AI+ 新药研发赛道，以及海外人才的回归，给中国 AI+ 新药研发注入一股新力量。从医疗领域全景来看，AI 尚未介入很多细分领域，还需要更长的时间、更系统化的解决方案。要实现 AI 在医疗领域的全面落地，需要不断优化升级 AI 系统，提升 AI 的智能化和个性化。虽然 AI 在医疗健康领域处于起步阶段，但普及到各细分领域的潜力巨大。

AI 能够实现在生物医药产业自上游到下游的投入使用，且虚拟筛选、靶点发现等部分应用场景已经能够为企业带来实际收益。新型冠状病毒肺炎（COVID-19）疫情发生后，越来越多的生物医药企业和研究机构通过将其业务与 AI 结合来完成创新突破，在新药开发、生产运营，甚至商业战略中都有所应用。AI 技术在生物医药领域中的应用涉及药物研发、医学影像、辅助治疗、基因治疗等方面，药物研发在全球医疗 AI 市场中的份额最大，占比达到 35%。靶点发现与筛选成为 AI+ 新药发现中最为热门的应用领域，AI 通过深度学习技术快速发现药物

接受日期：2021-06-22

* 通信作者：杨红飞，火石创造创始人兼 CEO；

研究方向：产业大数据；

Tel：0571-86885331；E-mail：yanghf@hsmap.com

** 通信作者：朱迅，教授，博士生导师；

研究方向：免疫学；

Tel：0431-85619476；E-mail：zxunzhux@vip.sohu.com

与疾病，以及疾病与基因间的连接关系，进而缩短靶点发现周期。在化合物合成方面，AI 可通过模拟小分子化合物的药物特性，在较短时间内挑选出最佳模拟化合物进行合成试验，大幅提高化学合成路线设计速度，以降低操作成本。

目前，AI 算法模型被诸多学者提出，随着药物研发数据的高速累积和数字化转型，以及 AI 技术的加速发展，决策树（DT）、随机森林（RF）和支持向量机（SVM）等机器学习模型以及深度神经网络（DNN）、卷积神经网络（CNN）和循环神经网络（FNN）等深度学习算法逐渐被应用于药物发现领域。本综述主要介绍机器学习和深度学习方法在药物发现领域的应用进展以及相关企业。

1 人工智能技术与算法模型简介

新药研发是一个漫长且高投入的过程，高通量筛选、药物基因组学等技术加速了药物开发，引领其步入大数据时代，药物发现大数据可用“十个 V”来描述，即：数量（volume）、速度（velocity）、品种（variety）、准确性（veracity）、有效性（validity）、词汇（vocabulary）、场合（venue）、可视化（visualization）、波动性（volatility）以及价值（value）[1]。基于数据库在药物发现不同阶段的应用和相关性，可将其分为 6 类：1）全面化学分子库，如 Enamine、PubChem 和 ChEMBL；2）药物 / 类药化合物库，如 DrugBank、AICD 和 e-Drug3D；3）收集药物靶标，包括基因组学和蛋白组学数据的数据库，如 BindingDB、Supertarget 和 Ligand Expo；4）存储通过筛选、代谢和功效研究获得的生物学数据的数据库，如 HMDB、TTD、WOMBAT 和 PKPB_DB；5）药物毒性数据库，如 DrugMatrix、SIDER 和 LTKB 基准数据集；6）临床数据库，如 ClinicalTrials.gov、EORTC 和 PharmaGKB[1]。

AI 领域中的自然语言处理、机器学习、深度学习、知识图谱、计算机视觉等相关技术，有助于解决药物研发领域的痛点。这些技术、算法模型在蛋白结构及蛋白-配体相互作用预测、药物靶点发现、活性化合物筛选等新药发现环节均已得到广泛应用[2–6]。各环节常用的 AI 方法详见图 1。

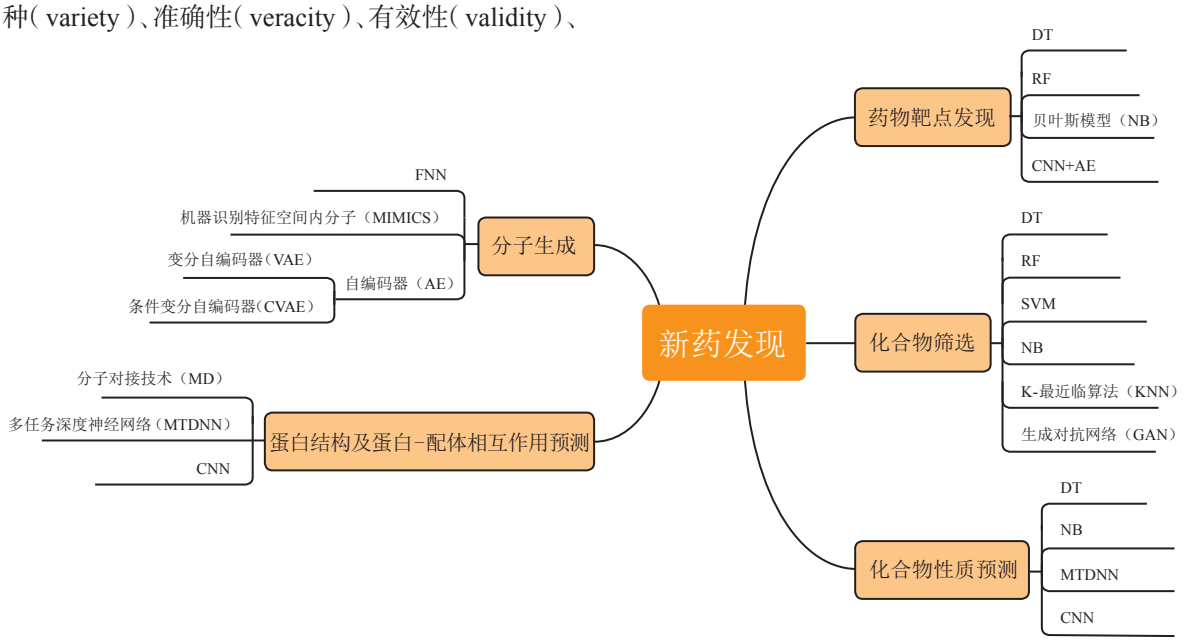


图 1 新药发现各环节常用的人工智能技术

Figure 1 Artificial intelligence techniques used in all aspects of new drug discovery

2 人工智能在药物发现中的应用

2.1 药物靶点识别

靶点是新药研发的基础。当前，药物研究的竞争主要集中体现在药物靶点研究上，早期药物靶点确定对研发项目成功至关重要。

DT 算法是一种常用的机器学习算法，具有条

理清晰、程序严谨、定量与定性分析相结合、方法简单、易于掌握、应用性强、适用范围广等优点。RF 算法是一种基于 Bagging 的集成学习方法，可处理分类、回归等问题，RF 分类器通过将许多 DT 结合来提升分类的正确率。目前，DT、RF 分类器可用于预测药物靶点，Costa 等 [7] 构建了一个基于 DT 的分类器，通过该分类器预测与疾病相关的基因，最后发现了多种转录因子在代谢通路和细胞外定位中的调控作用。Kumari 等 [8] 通过自助法采样提升了 RF 算法的稳定性，成功从潜在靶点中筛选出最有可能获得成功并应用于临床的靶点。Zeng 等 [9] 开发了 deepDTnet 深度学习方法，该系统嵌入了 15 种类型的网络，包括化学、基因组、表型和细胞网络，可以将最大的生物医学网络数据集成在一起，通过异构网络中的深度学习对已知药物进行靶标识别，以加速药物的重新利用、减少药物开发中的障碍。Madhukar 等 [10] 提出 BANDIT（Bayesian ANalysis to determine Drug Interaction Targets）可以准确预测药物与特定靶标的相互作用，不仅可用于识别多种多样的小分子的特定靶标，而且可用于区分同一靶标上的不同作用模式。

机器学习还可以预测肿瘤对药物的反应。Iorio 等 [11] 研究了全基因组基因表达、DNA 甲基化、基因拷贝数和体细胞突变数据对药物反应的影响。该研究组通过 3 种不同的分析框架，即方差分析、逻辑模型和机器学习算法（弹性网络回归和 RF）来定义“癌症功能事件”（cancer functional event，CFE）对药物敏感性预测的贡献。Iorio 等的研究成果可帮助新药研发工作者更好地利用肿瘤细胞系来了解哪些药物将为哪些患者提供最有效的治疗。

2.2 化合物高通量筛选

化合物筛选是指通过规范化的实验手段，从大量化合物中选择对某一特定靶点具有较高活性的化合物的过程，该过程需要较长的时间和成本。AI 可以通过对现有化合物数据库信息的整合和数据提取、机器学习，提取与化合物毒性、有效性相关的关键信息，从而大幅提高筛选的成功率，降低研发成本和工作量。

李瑾 [12] 利用化合物活性分类方法 ENS-VS 构建蛋白质和配体亲和力模型 ComplexNet，用于预测初步筛选出的小分子与靶标蛋白的结合强度，进行精细筛选。筛选过程分 3 步：首先，通过集成 SVM、朴素贝叶斯及 DT 这 3 种分类算法将蛋白质-配体相互作用特征和配体结构进行特征融合，解决活性化合物与非活性化合物样本数量严重不平衡的问题以及提高靶标蛋白的适用性、稳定性；其次，通过 Spark 大数据平台实现 ENS-VS 方法的并行加速，提高活性化合物筛选的执行效率；最后，基于 DUD-E 标准数据库针对靶标已知的活性化合物数量和是否出现新的靶标蛋白特性分别构建蛋白家族特异性模型、靶标特异性模型与通用模型。实验结果表明，ENS-VS 方法能有效提高活性化合物筛选的命中率，并且可与任意分子对接程序联合使用，对提高基于结构的虚拟筛选方法的成功率具有极其重要的意义。Wu 等 [13] 利用生物信息学和结构基因组学的方法系统分析了新型冠状病毒（SARS-CoV-2）基因编码的蛋白，将其作为主要或潜在的药物治疗靶点，并将 SARS-CoV-2 基因序列与 SARS-CoV 和 MARS-CoV 等冠状病毒进行了比对，通过 AI 计算机虚拟筛选方法发现一些具有抗病毒、抗菌和抗炎作用的临床药物和天然产物对上述靶蛋白表现出较高的亲和力，为 COVID-19 的治疗提供了新的可能。SVM 分类模型能够处理小数据集中的高维变量，还可以处理分类和回归问题，其分类效果强于 DT 与 RF 这 2 种机器学习方法。Poorinmohammad 等 [14] 通过建立 SVM 分类模型对人类免疫缺陷病毒（HIV）多肽进行分类，预测准确率达到 96.76%。SVM 用 MATLAB 编写的 svm 源程序可以实现 SVM 分类或提取，用于化合物库的虚拟筛选，有学者通过组合 SVM 和分子对接方法自动筛选化合物库，显著提高了活性化合物的命中率和富集因子，节省了计算资源 [15]。

细胞活力测定、细胞信号通路分析和疾病相关表型分析这 3 种基于细胞表型的方法常被用于筛选先导化合物。结合了 AI 技术的表型筛选更加高效，适用于更为复杂的病理生理过程，且能在细胞水平利用表型改变来筛选新化合物 [16]。SVM、RF 或贝叶斯等机器学习技术已被成功应用于药物发现阶段的化合物筛选环节。Cyclica 开发了名为“Ligand

Express”的云端蛋白质组学筛选平台[17]，该平台使用生物信息学和系统生物学技术将药物与蛋白的互动关系呈现为图像，利用 AI 对小分子化合物进行全面评估，帮助改善药物活性、预防药物副作用，以及发现能与小分子化合物结合的新靶点，制药科学家正在积极利用该平台探索药物发现新领域。SVM 和朴素贝叶斯模型已成功应用于哺乳动物雷帕霉素靶蛋白（mTOR）抑制剂的虚拟筛选。Narain 等[18]通过 AI 贝叶斯神经网络推断方法分析转移性前列腺癌（PC-3）细胞蛋白质组数据，生成每个特定因子的独特概率模型，再根据功能变量子网的 Burt 约束度量排名找到潜在的前列腺癌生物标志物 Filamin-A 和 Filamin-B 等。中国科学院上海生命科学研究院陈洛南教授团队利用 AI 克服了区分疾病样本和正常样本的分子生物标志物覆盖率低和假阳性率高的问题，确定了基于多维数据复杂疾病的网络标志物及动态网络标志物筛选方法[19–20]。

2.3 预测药物的吸收、分布、代谢、排泄和毒性

预测药物的吸收、分布、代谢、排泄和毒性（ADMET）是药物设计和药物筛选中十分重要的方法。过去，药物 ADMET 性质研究以体外研究技术与计算机模拟等方法相结合，研究药物在机体内的动力学表现。目前市场上有数十种计算机模拟软件，包括 ADMET Predicator、MOE、Discovery Studio 和 Shrodinger 等，该类软件现已在国内外的药品监管部门、企业 [如晶泰科技（XtalPi）、Numerate 等] 和科研院所得到了广泛应用。为了进一步提升 ADMET 性质预测的准确度，已有生物科技企业探索通过 DNN 算法有效提取结构特征，加速药物的早期发现和筛选过程。例如晶泰科技通过应用 AI 高效地动态配置药物晶型，完整地预测一个小分子药物所有可能的晶型，大大缩短了晶型开发周期，更有效地挑选出合适的药物晶型，减少了研发成本[21]。普林斯顿大学化学系的 Abigail G. Doyle 教授与默克公司的研究人员合作，利用 RF 算法对氨基化反应条件进行优化，准确预测具有多维变量的 Buchwald-Hartwig 偶联反应收率，结果表明，RF 算法可以利用高通量实验获得的数据来预测多维化学空间中合成反应的性能和化学反应收率，该机器学习算法模型将会在药物发现领域被广泛应用[22]。

严重药物不良反应是新药开发过程中导致失败的关键因素。王昊[23]通过构建贝叶斯网络预测模型进行药物不良反应的预测，结果发现该模型对导致呼吸困难发生频率在 1% 以上药物的预测准确率可以达到 86.76%，机器学习模型能够作为有效工具在药物发现阶段对其进行安全性评估。毒性是新药研发的一项重要指标，在药物发现阶段排除毒性大的化合物对于新药研发相当有利。Goh 等[24]构建了 CNN 毒性评估模型，将其用于预测分子的各种性质如毒性、活性和溶解性等，与多层感知机深度神经网络（MLPDNN）相比，发现 CNN 在活性与溶解度的预测方面表现更优异。

2.4 蛋白结构及蛋白-配体相互作用预测

靶点发现是新药研发的关键，而蛋白质功能分类研究有助于深入理解靶点蛋白特征，是解决药物靶点发现难点的有效途径。随着 AI、大数据等技术的迅速发展，蛋白质功能预测已成为蛋白质功能注释的重要手段，也成为药物靶点发现领域的前沿问题[25]。序列同源性比对、CNN 等多种计算方法被应用于蛋白质功能预测研究，方法论是同源蛋白具有相似功能[26]。

谷歌 DeepMind 团队开发出的 AI 产品 AlphaFold2，可根据氨基酸序列准确预测蛋白质结构，预测结果已接近实验数据的水平，且预测的准确度可与冷冻电子显微镜（cryo-EM）、核磁共振或 X 射线晶体学等实验技术媲美[27]。谷歌 DeepMind 开发的 AlphaFold[28] 深度学习系统可以快速预测 SARS-CoV-2 的蛋白质结构，为 COVID-19 疫苗设计提供有价值的信息，而使用传统的实验方法获得蛋白质结构可能需要数月时间[29]。洪嘉俊[30]通过基于 CNN 的蛋白质二进制编码表示策略构建了蛋白质功能预测模型，结果表明，CNN 预测 GO 家族蛋白的准确率在 66%～98% 之间，显著高于 SVM、概率神经网络（PNN）和 KNN 这 3 种机器学习方法，表明 CNN 模型在真实世界中具有很好的假阳性控制率。由于目前的细菌Ⅳ型分泌系统效应蛋白（T4SE）预测方法存在假阳性率高等缺点，洪嘉俊针对 T4SE 和非 T4SE 数据特征分别建立了

T4SE 的 CNN 预测模型，通过采用与 Bastion4 方法完全相同的建模数据集进行评估，基于蛋白质二级结构特征、位置特异性评分矩阵和序列 One-hot 编码技术这 3 种方式建立的模型预测准确率分别为 95.6%、98.9% 和 96.7%，效果显著高于 Bastion4，表明 CNN 模型可以用于 T4SE 的注释，且可以很好地控制假阳性率。

DNN 在蛋白结构预测、蛋白质-配体相互作用预测方面也有应用。AlphaFold 利用高效训练的 DNN 从主序列中预测蛋白质的性质，通过 DNN 预测氨基酸对之间的距离和相邻肽键之间的 φ - ψ 角，探索蛋白质结构的微观结构，以找到与预测相匹配的结构 [31]。Ragoza 等 [32] 使用 CNN 对蛋白配体复合物构建打分函数，通过打分函数评价蛋白-配体相互作用，该打分函数在蛋白-配体预测和虚拟筛选中的打分表现比 AutoDock Vina 更好，但是也存在实际计算的结果可能会远大于实验观察值的偏差问题，因此 CNN 在该方面的应用还有一定的改进空间。刘桂霞等 [33] 基于 DNN 构建蛋白质相互作用预测框架，预测框架在酿酒酵母蛋白质数据集上的准确率达到 95.67%，精确度达到 96.38%，该预测框架可以解决较高假阳性率和假阴性率的问题，整合蛋白质特征数据；张丽娜 [34] 提出基于多源特征的提取策略，利用集成学习方法构建蛋白质-配体相互作用预测模型，该方法的敏感性和 Youden 指数均优于单分类器预测模型，可以有效解决数据不平衡问题。Cunningham 等 [35] 基于 6 个常见的球形蛋白结合域（PBD）家族构建了 HSM 模型，其能准确预测跨多个蛋白质家族的 PBD-肽相互作用的亲和力，HSM 具有较高的灵活性，适用于在疾病中对突变的 PBD 和肽进行建模，以及基于肽的药物的设计。

2.5 分子生成

AI 可以通过对海量化合物或药物分子的学习获得化合物分子结构和成药性方面的规律，再根据规律生成很多自然界从未存在过的化合物，将其作为候选药物分子，有效构建拥有一定规模且高质量的分子库。高质量的小分子库是药物研发人员一直关注的问题，研究者们利用深度学习技术设计了变分自动编码器（VAE）、生成对抗网络（GAN）、自回归模型（如 PixelRNN 和 PixelCNN）等不同的分子生成模型。

Yang 等 [36] 提出基于分子片段的 AI 分子设计新算法，该算法模型是基于带约束的 Transformer 神经网络架构 SyntaLinker，可以快速自动生成满足特定链接段约束条件的大量新颖的分子结构。神经网络 SyntaLinker 由多个注意力机制（attention）模块构成，SyntaLinker 利用其编码层和解码层对输入的分子片段结构序列进行处理，将分子片段自动连接起来，且结合约束信息，填充链接段，从而生成一个完整的分子。未来这种基于片段连接的分子设计算法能被用于实际的药物开发项目中，为药物化学家提供更多具有启发性的化学结构。曲晋慷 [37] 对新型药物设计方法进行创新，提出通过深度分子生成模型 DGMM、深度迁移分子生成模型 T-DGMM、深度强化分子生成模型 R-DGMM 这 3 种模型生成潜在抗 HIV 活性分子，以扩增潜在抗 HIV 活性分子库。DGMM 基于 MLSTM、SRU、QRNN 这 3 种循环单元进行构造可以生成结构有效、新颖且性质无偏的分子；T-DGMM 通过搭建抗 HIV 活性预测模型 AAPM 可以生成潜在抗 HIV 活性分子，扩增潜在抗 HIV 活性分子库；R-DGMM 采用基于策略梯度的强化学习方法 REINFORCE 搭建模型，生成抗 HIV 药物利匹韦林的相似物，适用于潜在抗 HIV 活性分子库扩增。谭小芹 [38] 基于循环神经网络建立了分子生成模型，进行多靶点 GPCR 分子库的自动设计，再对生成的分子进行活性、可合成性、类药性等多方面评估过滤，最终得到了具有潜在治疗精神疾病活性的候选化合物。同时，基于序列到序列（Seq2Seq）模型建立分子生成模型，该模型可以生成一个基于骨架的虚拟分子库，然后通过激酶谱预测模型对分子库进行虚拟筛选，最终筛选得到可抑制细胞中促炎因子的表达和盘状结构域受体家族成员 1（DDR1）自磷酸化的化合物。

在分子设计领域，生成模型还处于起步阶段，其面临着以下挑战：1）如何提高模型的泛化能力；2）如何提高对真实数据进行推断的能力；3）如何提高生成新分子的能力。此外，分子生成模型的性能难以评估。如何建立基准以便于量化比较模型性能，

而非通过预测分子溶解度或药物相似性等方法进行比较仍充满挑战[39]。

3 全球人工智能新药发现企业及市场规模

伴随 AI 技术的迅猛发展，新药研发工作者希望通过 AI 技术解决医药行业痛点，包括降低药物的研发成本、缩短其研发周期、控制新药研发风险，在此基础上，一批 AI 企业相继出现。

国内外多家 AI 企业与药企开启了深度战略合作模式，利用其自主设计的人工智能技术平台助力制药企业进行新药研发（见表 1）。

基于 AI 技术的药物设计公司 Atomwise 拥有的 AtomNet® 是第一虚拟药物发现平台，其核心技术是 CNN。Atomwise 已与多家制药公司开展约 1 000 个项目，主要包括肿瘤、传染病、神经系统疾病、心血管疾病、免疫性疾病、内分泌系统疾病、COVID-19 等领域的药物研究。

晶泰科技以 AI、量子物理、量子化学及云计算为核心，推动 AI 赋能的数字化药物研发新基建，为创新药研发增效提速。晶泰科技 AI 药物发现平台，在分子生成、虚拟筛选、高精度活性预测等 AI+ 药物发现的关键环节具有独到的技术优势，能实现超大型化学空间的探索，百万级的新分子结构生成及全面、综合的成药性、活性、ADMET 等性质的评估，完成高质量的先导化合物开发和临床前候选化合物开发。

表 1 人工智能企业与制药企业在新药研发领域的战略合作

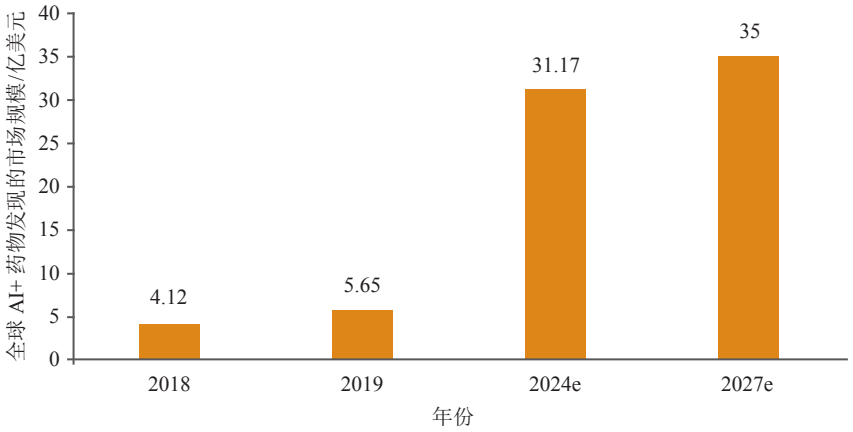
Table 1 Strategic cooperation between artificial intelligence enterprises and drug manufacturers in the field of new drug research and development

AI 企业	合作药企	主要研究方向	人工智能技术/平台	AI+ 药物发现业务
Atomwise	默克、豪森、辉瑞、默沙东、Bridge Biotherapeutics、翰森制药	中枢神经系统（CNS）疾病、肿瘤、感染性疾病和糖尿病等领域药物	深度学习（CNN）/Atom Net®、Pandomics 平台	利用深度学习 AI 技术平台，用于基于结构的小分子药物发现工作，主要应用场景涉及靶点确定、化合物筛选、药物设计
IBM Watson	辉瑞	免疫肿瘤学（Immuno-oncology）中的新药物识别、联合疗法和患者选择策略	Watson for Drug Discovery 的机器学习	用于免疫肿瘤学中的新药物识别，涉及靶点确定、化合物筛选、药物设计、分子生成，帮助生命科学研究者发现新的药物靶点和替代性的药物的适应证
Exscientia	华东医药、GSK、赛诺菲、葛兰素史克、默克、强生	慢性阻塞性肺病治疗药物、抗肿瘤（如卵巢癌和乳腺癌）药物、免疫调节药	深度学习（贝叶斯）/ Centaur Chemist™、Centaur Biologist™、AI 驱动平台	凭借先进的 Centaur Chemist™ AI 平台进行自动化药物研发指导，设计特定靶标的新分子
晶泰科技	华东医药、辉瑞、PhoreMost	肿瘤、内分泌疾病和自身免疫疾病领域药物	量子物理模型和 AI 算法、小分子药物模拟算法平台	利用智能药物研发 ID4 平台寻找新药靶点，发现和设计小分子抑制剂，同时针对发现的靶点，利用机器学习和量子物理算法，探索超大型化学空间，快速开发高质量的先导化合物
Insitro Medicine	药明康德	非酒精性脂肪性肝炎治疗药物，神经退行性疾病疗法	深度学习 [GAN 和强化学习（RL）]/nDexer™	利用机器学习算法加快药物发现和开发的 3 个领域，包括疾病靶标识别、合成生物学（生成生物学）和新型分子（生成化学）数据的生成，以及预测临床试验结果
BenevolentAI、ProteinQure、Berg Health	阿斯利康（AstraZeneca）	治疗慢性肾病和特发性肺纤维化的新药靶点发现，帕金森病等神经系统疾病药物研究	机器学习、深度学习（GAN、CNN）	利用 AI 和机器学习技术，加快发现可治疗慢性肾病和特发性肺纤维化的新药
燧坤智能	维亚生物、保诺科技、信邦制药	苗头和先导化合物的发现及优化，针对恶性肿瘤靶点的小分子药物，治疗神经病痛及老年痴呆等重大疾病的药物	创新药研发综合平台、Silexon® AI4D™ 技术平台与基于结构的药物研发（SBDD）和基于片段的药物研发（FBDD）新药筛选平台、新药研发 CRDMO（合同研究、工艺开发和生产）一站式技术平台	利用 AI 算法发掘疾病靶点、发现已知药物新适应证、提升新药筛选效率、提高大分子产量，主要业务聚焦在药物重定向和虚拟高通量筛选这 2 个方向

注：表格内容由火石创造根据各制药企业及 AI 企业官网信息整理

伴随药物研发数据的高速累积和药企数字化转型，以及 AI 技术的加速发展，AI 在新药发现的应用日益增多，其优势也得到突出体现。互联网数据资讯网（BCC）数据显示，AI 在医疗健康产业所

有应用场景中，新药发现的市场规模与增长速度均占据第一位，预计 2024 年市场规模将达到 31.17 亿美元，年均复合增长率（CAGR）为 40.7%；根据大观研究（Grand View Research）的最新报告，到 2027 年，全球 AI+ 药物发现的市场规模预计将达到 35 亿美元，CAGR 为 28.8%（见图 2）。



注：数据来源于 BCC 与 Grand View Research；“e”表示预测

图 2 人工智能在新药发现领域的市场规模

Figure 2 Market size of artificial intelligence in new drug discovery

火石数据库资料显示，国内从事 AI+ 药物发现的企业有晶泰科技、深度智药、云势软件、望石智慧等，主要分布在北京（7 家）、上海（4 家）、杭州（2 家）和深圳（2 家）等地（见表 2）；但总数较少，不足 20 家。

表 2 国内主要从事 AI+ 药物发现的公司及其业务布局

Table 2 Major domestic companies applying artificial intelligence in drug discovery and their business layout

重点企业	药物发现相关业务		
	靶点发现	化合物筛选	化合物合成
晶泰科技（北京、深圳）		√	√
冰洲石生物（上海、纽约）		√	√
云势软件（北京）	√		
深度智耀（北京、上海、沈阳）		√	√
亿药科技（北京）	√	√	
宇道生物（上海）		√	
意嘉健康（北京）	√		
望石智慧（北京）		√	√
燧坤智能（南京）	√	√	
分迪科技（成都）		√	
费米子（广州）	√	√	√
智药科技（深圳、上海）	√	√	√
元气知药（北京、南京）	√	√	√
METiS（杭州）	√	√	√
成都先导（成都）	√	√	√
赛恪科技（杭州）		√	

注：表格内容由火石创造根据各 AI 企业官网信息整理

2015—2020 年，我国药物发现 CRO 市场 CAGR 达到 28.2%，2020 年市场规模约为 131.5 亿元；预计未来 5 年，创新药研发速度不断加快，我国药物发现 CRO 市场仍将保持快速增长态势，到 2025 年

市场规模将达到 385.2 亿元。

4 人工智能应用于新药发现的机遇与挑战

受 DNN 或递归神经网络（RNN）技术快速发展的影响，AI 技术在药物靶点发现、化合物合成、化合物筛选、晶型预测、药理作用评估、药物重定向、新适应证开发等多个场景中应用广泛，应用优势也愈加凸显。TechEmergence 研究报告显示，AI 可以将新药研发的成功率从 12% 提高到 14%。此外，AI 在化合物合成和筛选方面可节约 40%~50% 的时间，每年为制药行业节约 260 亿美元的化合物筛选成本 [40]。基于此，药物研发领域数字化转型加速，各大制药公司都在迫切寻找能够缩短新药研发周期、有效提高研发成功率、开发有竞争力的创新药物的解决方案。

AI 在新药研发中的应用面临政策瓶颈、人才匮乏、技术壁垒、数据质量不确定等方面的挑战。第一，从政策瓶颈来看，新技术的引进改变原有药物研发模式，而现在尚无针对性的政策指南出台。第二，从人才壁垒来看，高端复合型人才缺失较严重，限制创新发展。未来需要国家出台相关人才政策，培养复合型高端人才。第三，从技术壁垒来看，自然语言、知识图谱以及知识问答、分析决策和语义搜索等需要较大提升。第四，从数据质量挑战性来看，AI 模型基于数据学习，数据学习导致了结果的不确定性，新药研发系统工程加上 AI 双系统的不确定性也会导致新药研发结果的不确定性。近年来，出现了一些来源于临床相关模型的高通量数据，例如用于高通量测试的异质细胞系统及其参数（3D 细胞模型中的细胞间相互作用和渗透性）和患者衍生的测试系统，这些系统产生的数据将来可能会对药物发现产生重大影响；但当前阶段，可用于 AI 挖掘的数据仍相对较少，需要生成足够大量的数据才能真正在上述系统里使用 [41]。

5 结语与展望

尽管在多数情况下化学数据可大规模获得并成功用于配体设计和合成，但这些数据并不能满足 AI 药物发现的需求，且大量可用于模型建立的测定数据（如小分子的各种体外物理化学性质）也并不能很好发挥作用。因此，未来需要更多的高质量化合物数据进行 AI 研究，包括化合物的体外活性 / 毒性指数，以及正确剂量 / 药代动力学数据等。在后期阶段，还需要化合物在动物模型中的药效和毒性数据。此外，我们还需要更有效地进行临床试验，以获得高质量化合物临床数据。

AI 分析药物在体内活性时的数据非常有限，使得计算机不能很好地做出决策，主要影响因素有：第一，没有一个可以比较的基准；第二，可选择的化学结构非常多；第三，在化学领域验证药物的有效性非常难，实验中使用数据往往具有稀疏性和保密性的特性。

值得一提的是，大量描述化学特性的数据能够使计算机生产相应的配体，但配体发现不等于药物发现。在未来，我们需要更多了解药物的生物学特性，了解它们在人体内的一系列反应。此外，临床成功率比时间和成本更重要，我们需要让更多高质量候选化合物进入临床，更好地验证靶点，以及选择合适的患者进行临床试验，提高临床成功率，从而生成有用的数据，从本质上推动 AI+ 药物发现领域的进展。

[参考文献]

[1] Zhao L L, Ciallella H L, Aleksunes L M, *et al*. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling[J]. *Drug Discov Today*, 2020, 25(9): 1624-1638.

[2] Rashid M. Artificial intelligence effecting a paradigm shift in drug development[J]. *SLAS Technol*, 2021, 26(1): 3-15.

[3] Hessler G, Baringhaus K H. Artificial intelligence in drug design[J]. *Molecules*, 2018, 23(10): 2520. DOI:10.3390/molecules23102520.

[4] Krishnaveni C, Arvapalli S, Sharma J V C, *et al*. Artificial intelligence in pharma industry-a review[J]. *Int J Innov Pharm Sci Res*, 2019, 7(10): 37-50.

[5] 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.

[6] Xiong Z P, Wang D Y, Liu X H, *et al*. Pushing the boundaries of

molecular representation for drug discovery with the graph attention mechanism[J]. *J Med Chem*, 2020, 63(16): 8749-8760.

[7] Costa P R, Acencio M L, Lemke N. A machine learning approach forgenome-wide prediction of morbid and druggable human genes based on systems-level data[J]. *BMC Genomics*, 2010, 11(Suppl 5): S9.

[8] Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms[J]. *Comput Biol Med*, 2015, 56: 175-181.

[9] Zeng X, Zhu S, Lu W, *et al*. Target identification among known drugs by deep learning from heterogeneous networks[J]. *Chem Sci*, 2020, 11: 1775-1797.

[10] Madhukar N S, Khade P K, Huang L, *et al*. A Bayesian machine learning approach for drug target identification using diverse data types[J]. *Nat Commun*, 2019, 10(1): 1-14.

[11] Iorio F, Knijnenburg T A, Vis D J, *et al*. A landscape of pharmacogenomic interactions in cancer-ScienceDirect[J]. *Cell*, 2016, 166(3): 740-754.

[12] 李瑾 . 基于机器学习技术的药物虚拟筛选方法研究 [D]. 重庆 : 西南大学 , 2020.

[13] Wu C R, Liu Y, Yang Y Y, *et al*. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods[J/OL]. *Acta Pharmaceutica Sinica B*, 2020, 10(5). [2021-06-10]. https://doi.org/10.1016/j.apsb.2020.02.008.

[14] Poorinmohammad N, Mohabatkar H, Behbahani M, *et al*. Computational prediction of anti HIV-1 peptides and *in vitro* evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides[J]. *J Pept Sci*, 2015, 21(1): 10-16.

[15] Xie Q Q, Zhong L, Pan Y L, *et al*. Combined SVM-based and docking based virtual screening for retrieving novel inhibitors of c-Met[J]. *Eur J Med Chem*, 2011, 46(9): 3675-3680.

[16] Zheng W, Thorne N, McKew J C. Phenotypic screens as a renewed approach for drug discovery[J]. *Drug Discov Today*, 2013, 18(21/22): 1067-1073.

[17] Cyclica. The Ligand Express™ Platform guides drug repurposing study[EB/OL]. (2018-12-11)[2020-04-30]. https://static1.squarespace.com/static/60802f83c72d97003aaa070d/t/608ad5ed3479534fb72bc1f8/1619711470952/Cyclica_case_The+Ligand+Express+platform.pdf.

[18] Narain N R, Diers A R, Lee A, *et al*. Identification of Filamin-A and -B as potential biomarkers for prostate cancer[J]. *Fut Sci OA*, 2017, 3(1) : 524-532.

[19] Wang L, Liu Z P, Zhang X S, *et al*. Prediction of hot spots in protein interfaces using a random forest model with hybrid features[J]. *Protein Eng Des Sel*, 2012, 25(3): 119-126.

[20] Liu Z P, Wang Y, Zhang X S, *et al*. Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains[J]. *BMC Syst Biol*, 2010, 4(Suppl2): S11.

[21] Zhang P Y, Wood G P F, Ma J, *et al*. Harnessing cloud architecture for crystal structure prediction calculations[J]. *Cryst Growth Des*, 2018, 18(11): 6891-6900.

[22] Ahneman D T, Estrada J G, Lin S, *et al*. Predicting reaction performance in C-N cross-coupling using machine learning[J]. *Science*, 2018, 360(6385): 186-190.

[23] 王昊 . 基于机器学习方法的药物不良反应预测 [D]. 厦门 : 厦门大学 , 2012.

[24] Goh G B, Siegel C, Vishnu A, *et al*. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models[EB/OL]. (2017-06-20)[2020-01-01]. https://arxiv.org/abs/1706.06689.

[25] Jiang Y X, Oron T R, Clark W T, *et al*. An expanded evaluation of protein function prediction methods shows an improvement in accuracy[J]. *Genome Biol*, 2016, 17(1): 1-19.

[26] Watson J D, Laskowski R A, Thornton J M. Predicting protein function from sequence and structural data[J]. *Curr Opin Struct Biol*, 2005, 15(3): 275-284.

[27] Callaway E. DeepMind's AI for protein structure is coming to the masses[EB/OL]. (2021-07-15)[2021-07-16]. https://www.nature.com/articles/d41586-021-01968-y.

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

[29] Alimadadi A, Aryal S, Manandhar I, *et al*. Artificial intelligence and machine learning to fight COVID-19[J]. *Physiol Genomics*, 2020,

52(4): 200-202.

[30] 洪嘉俊 . 基于深度学习的蛋白质功能预测及药物靶点发现研究 [D]. 杭州 : 浙江大学 , 2020.

[31] Service R F. Google's DeepMind aces protein folding[EB/OL]. (2018-12-06)[2021-03-22]. https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding.

[32] Ragoza M, Hochuli J, Idrobo E, *et al*. Protein-ligand scoring with convolutional neural networks[J]. *J Chem Inf Model*, 2017, 57(4): 942-957.

[33] 刘桂霞 , 王沫沅 , 苏令涛 . 基于深度神经网络的蛋白质相互作用预测框架 [J]. 吉林大学学报（工学版）: 2019, 49(2): 570-577.

[34] 张丽娜 . 基于机器学习的蛋白质类别及蛋白质-配体相互作用预测研究 [D]. 济南 : 山东大学 , 2017.

[35] Cunningham J M, Koytiger G, Sorger P K, *et al*. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning[J]. *Nat Methods*, 2020, 17(2): 175-183.

[36] Yang Y, Zheng S, Su S, *et al*. SyntaLinker: automatic fragment linking with deep conditional transformer neural networks[J]. *Chem Sci*, 2020, 11(31): 8312-8322.

[37] 曲晋慷 . 基于深度学习的潜在抗 HIV 活性分子生成新方法研究 [D]. 兰州 : 兰州大学 , 2020.

[38] 谭小芹 . 基于虚拟筛选和深度生成模型的药物发现与优化研究 [D]. 上海 : 中国科学院大学 , 2021.

[39] Schwalbe-Koda D, Gómez-Bombarelli R. Generative models for automatic chemical design[EB/OL]. (2019-07-02)[2021-05-22]. https://arxiv.org/abs/1907.01632v1.

[40] Wong C H, Siah K W, Lo A W. Estimation of clinical trial success ratesand related parameters[J]. *Biostatistics*, 2019, 20(2): 273-286.

[41] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet[J]. *Drug Discov Today*, 2020, 26(2): 511-524.

[专家介绍] 杨红飞：产业大数据专家，火石创造创始人兼 CEO。火石产业大脑总设计师，通过基于全球的产业数据自动化采集、机器学习、场景化模型构建和智能分析，将产业数据产品化、服务化，助推战略性新兴产业高质量发展。作为负责人承担过多个科技部、发改委产业专项及省重点研发计划项目；中国第一个生物医药产业发展指数 CBIB 设计者。牵头制定《生物经济产业分类目录》企业标准，连续 3 年参与国家发改委、中国生物工程学会《中国生物产业发展报告》的编写，指导编制首个聚焦产业新基建的白皮书。其个人拥有多项国家发明专利，是 2018 年唯一的“杭州市领军型青年创业团队”核心成员，杭州市科技专家库入选者、杭州市滨江区青年科技英才。同时还担任中国医学装备协会理事、浙江省数字经济学会成员、《药学进展》青年编委。

[专家介绍] 朱迅：医学博士，著名免疫学家，吉林大学教授，博士生导师，原国家新药咨询委员，国家自然科学基金专家组成员，《药学进展》副主编，火石创造战略顾问。先后主持或参与承担国家自然科学基金委员会、卫生部、国家医药管理局、吉林省科委、吉林省计委、日本厚生省等资助的课题 20 多项，共发表论文 200 多篇，参加编写或主编专著及教材 20 多部。获卫生部、吉林省科委等科技成果奖 5 次。1991 年获吉林省第二届青年科技奖；1993 年享受国务院政府特殊津贴；1996 年获“第二届全国中青年医学科技之星”称号；1996 年被国家人事部、国家教委评为“全国优秀留学回国人员”；1997 年获卫生部“笹川医学奖学金优秀归国进修人员奖”；1997 年被评为“吉林省有突出贡献的中青年专业技术人才”；2000 年入选教育部“高等学校骨干教师资助计划”及教育部“跨世纪优秀人才培养计划”。曾任白求恩医科大学副校长，参加国内多家医药产业园区及医药企业的战略咨询，技术及项目论证等数十次，曾担任多家医药公司的顾问。多次参加国务院研究发展中心、国家发改委、国家药品监督管理局、科技部等组织的咨询及论证会；应邀在国际及全国性会议或论坛上做大会报告或主题演讲（讲座）100 余次；并多次组织或主持了全国性及区域性的“生物技术”研讨会或专题报告会。引荐、协助了 20 余名留学生回国创业或加盟国内制药公司，其中多家企业已经成为各自领域的行业龙头。