TSINGHUA PBCSF 清华五道口 | NIFR 国家金融研究院

研究报告

（2022 年第 7 期 总第 116 期） 2022 年 8 月 15 日

人工智能在新药研发中的应用[1]

资本市场与公司金融研究中心

朱雅姝 安砾

【摘要】行业图谱研究是本中心科技成果转化研究的一项子课题，目标定位于清晰理解前沿科技成果的技术核心、科创企业的技术竞争力及科研工作者的研究进度，从而助力科技成果转化效率的提升。行业图谱研究将以系列形式展开，选取国家战略重点科技领域的商业应用场景逐一进行，时效性较强。

本报告属于行业图谱的第一个系列——生物医药领域。药物发现和研发是制药企业和化学科学家的重要研究领域。人工智能和机器学习技术使制药领域实现了现代化。机器学习和深度学习算法已被应用于多肽合成、虚拟筛选、毒性预测、药物监测和释放、药效团建模、

[1] 感谢资本市场与公司金融研究中心的实习生许喜远同学对本报告的助研工作。许喜远同学是清华大学医学院 2019 级硕士。

定量构效关系、药物重定位、多药理和生理活性等药物发现过程。此外，新的数据挖掘和管理技术为最近开发的建模算法提供了支持。其中蛋白质折叠和蛋白质相互作用的人工智能预测技术是现阶段药物研发中的通用选择，并成为技术实力的关键竞争环节，也是国际企业间专利争夺之处。来自美英的两家企业 Recursion 和 Benevolent AI 并称为全球两大新药物研发引领者，核心技术主要来自专利授权。其中 Benevolent AI 通过知识网络图谱分析与识别，预测了巴瑞替尼（Baricitinib）[2]或可用于治疗新冠肺炎。我国药物研发企业的创立和研发紧跟国际步伐，以自主研发的技术平台占主导，其中生物医药公司英矽智能（Insilico Medicine）已发现一款靶向主蛋白酶（3CL）的全新临床前候选药物，用于治疗新型冠状病毒引起的肺炎。学术领域国内主要研究者的科研进展也集中在蛋白质折叠和蛋白质相互作用的预测开发上，研究成果普遍进入了自创企业或企业合作的转化模式。

[2] 巴瑞替尼（Baricitinib）是一种是 Janus(JAK1/JAK2) 激酶抑制药，由美国制药公司 Incyte 和 Eli Lilly 共同开发，主要用于治疗成人中重类风湿关节炎。巴瑞替尼最早于 2017 年在欧洲获批上市，商品名 Olumiant，中文译名：艾乐明。

目 录

图表目录

新药研发存在周期长、费用高和成功率低等特点，人工智能作为药物研发领域的一个热点方向，已被应用到药物研发的各个阶段，医药领域对人工智能技术越来越重视，目前人工智能技术在药物研发中的应用主要表现为七个场景，分别是：靶点药物研发、候选药物挖掘、化合物筛选、预测 ADMET 性质、药物晶型预测、辅助病理生物学研究和发掘药物新适应症，人工智能可以直接为新药的研发做出贡献，AI＋药物研发与传统模式相比，时间和成本优势明显。AI＋药物研发的结合必然是未来制药行业的发展趋势，对医药领域进行一场巨大的革命，让医药行业迎来新时，随着新冠疫情的爆发，国内外医药从业人员纷纷涉足 AI 人工智能，国内多个科研院所高校企业单位更是创立多个人工智能药物研究所，投入巨额资金。

那么，什么是人工智能？为什么要将人工智能技术和药物研发结合？中国在人工智能药物方面取得了什么进展？在当前全球形势和开放新格局下，国际顶级医疗公司如何利用人工智能技术生成更有效的药物？本文试图对这些问题提供一些解答。

一、引言

（一）人工智能的演进：从机器学习到深度学习

人工智能（Artificial Intelligence，AI），也被称为机器智能，指的是计算机系统从输入或过去的数据中学习的能力，术语“人工智能”通常用于机器在学习和解决问题过程中模仿与人脑相关的认知行为。

机器学习（Machine Learning，ML）可描述为“AI 的应用”，同时也可称为“AI 的子集”。ML 将数据与朴素贝叶斯、决策树

（Decision Tree，DT）、隐马尔可夫模型（Hidden Markov Model，HMM）等算法一起输入机器，使其在不显式编程的情况下进行学习。大约在 20 世纪中旬，Igor Aizenberg 和他的同事们在谈论人工神经网络（Artificial Neural Network，ANN）时，首次提出了“深度学习（Deep Learning，DL）”这一术语。

图 1-1 人工智能的分类

根据《人工智能：现代方法》中的讨论，人工智能有七种分类（图 1-1），分别是推理和问题解决、知识表示、规划和社会智能、感知、机器学习、机器人：运动和操纵，以及自然语言处理。机器学习进一步分为三个重要子集：监督学习、无监督学习和深度学习；而自然语言处理被分为五个主要子集，包括分类、机器翻译、问答、文本生成和内容提取。

（二）药物研发中的革命性过程：大数据和人工智能的作用

大数据可以定义为过于庞大和错综复杂的数据集，无法使用传统的数据分析软件、工具和技术进行分析。大数据的三个主要特征是体积、速度和多样性，其中体积代表产生的大量数据，速度代表这些数据被再现的速率，多样性代表数据集中存在的异质性。随着微阵列、转录组测序技术（Ribonucleic Acid Sequencing，RNA-seq）和高通量测序（High-Throughput Sequencing，HTS）技术的出现，每天都会产生过多的生物医学数据，当代药物发现也因此进入了大数据时代。

如图 1-2 所示，在新药研发中，第一步也是最重要的一步是确定与疾病病理生理学有关的适当靶点（如基因、蛋白质），然后找到可以干扰这些靶点的药物或类药物分子。如今我们可以搜索一系列生物医学数据库来实现，如国家生物技术信息中心（National Center for Biotechnology Information，NCBI）的基因表达综合（Gene Expression Omnibus，GEO）数据集、癌症基因组图谱（The Cancer Genome Atlas，TCGA）和 Arrayexpress 等等。另外，通过出版的文献也可以用于识别靶点，如 PubMed 是各种已出版生物医学文献的数据库，对其进行数据挖掘可以帮助识别不同疾病的靶点。此外，人工智能的发展使得大数据分析变得容易得多，因为现在有无数的 ML 技术可用，这些技术可以帮助提取这些大型生物医学数据集中存在的有用特征、模式和结构。如 Han 等人[1]利用大数据和人工智能在 2019 年开发了 DriverML，这是一个基于 ML 监督学习的工具，可以指出与癌症相关的驱动基因。

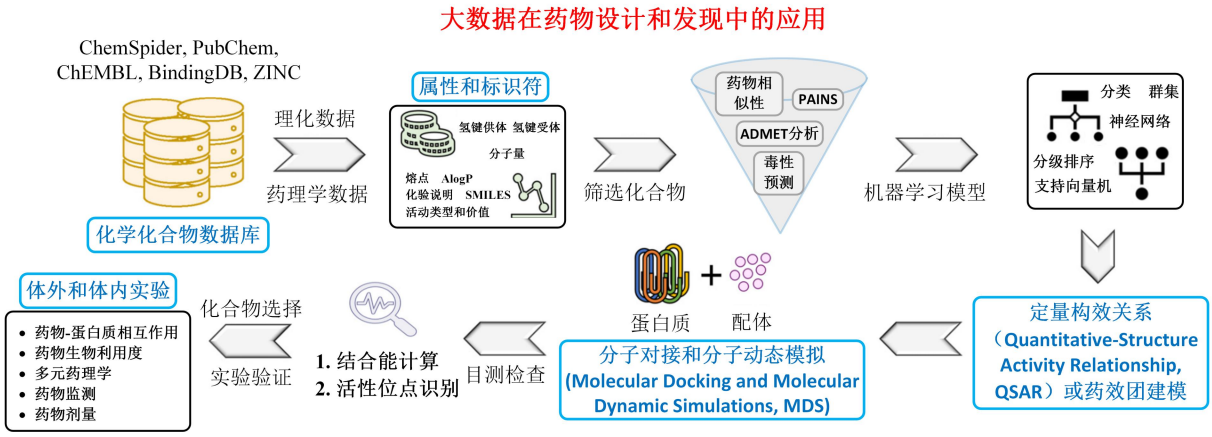

图 1-2 大数据在药物设计和发现中的应用

在确定和验证了合适的靶点之后，下一步是寻找合适的药物或类药物分子，这些分子可以与靶点相互作用并引起所需的反应。在大数据时代，通过支配海量的大型化学数据库，协助寻找针对特定靶点的完美药物。比如 PubChem 是一个免费的化学数据库，其中包含各种化学结构的数据，包括它们的生物、物理、化学和毒性特性；ChEMBL 包含许多具有类似药物特性的生物活性化合物的数据，还包含有关这些化合物的吸收、分布、新陈代谢和排泄（Absorption, Distribution, Metabolism and Excretion，ADME）、毒性特性，甚至它们的靶相互作用的信息；其他的化学数据库还包括 DrugBank、LINCS L1000 和 PDB 等。

（三）本章小结

随着技术的进步和高性能计算机的发展，在计算机辅助药物设计（Computer Aided Drug Design，CADD）中补充了从 ML 到 DL 的一系列人工智能算法。在过去的二十年里，发展了许多用于计算新药研发、定量结构活性关系（Quantitative Structure Activity Relationship，QSAR）和自由能最小化技术的工具。传统的面向化学

的药物发现与人工智能药物设计相结合，提供了一个很好的研究平台。此外，世界各地的系统生物学和化学科学家与计算科学家经过多方面合作，开发现代 ML 算法和原理，可以促进药物的发现和开发。

二、人工智能在药物发现与研发中的应用技术和方法流程

在新药研发过程中，常遇到的瓶颈问题有：①寻找合适的、具有生物活性的药物分子；②药物分子难以通过第二阶段临床试验和其他监管批准。利用基于人工智能的工具和技术，提升药物研发的效率，解决上述所面临的药物研发问题。为此，下面将详细介绍人工智能在药物发现与研发中的应用技术和方法流程，如图 2-1 所示。

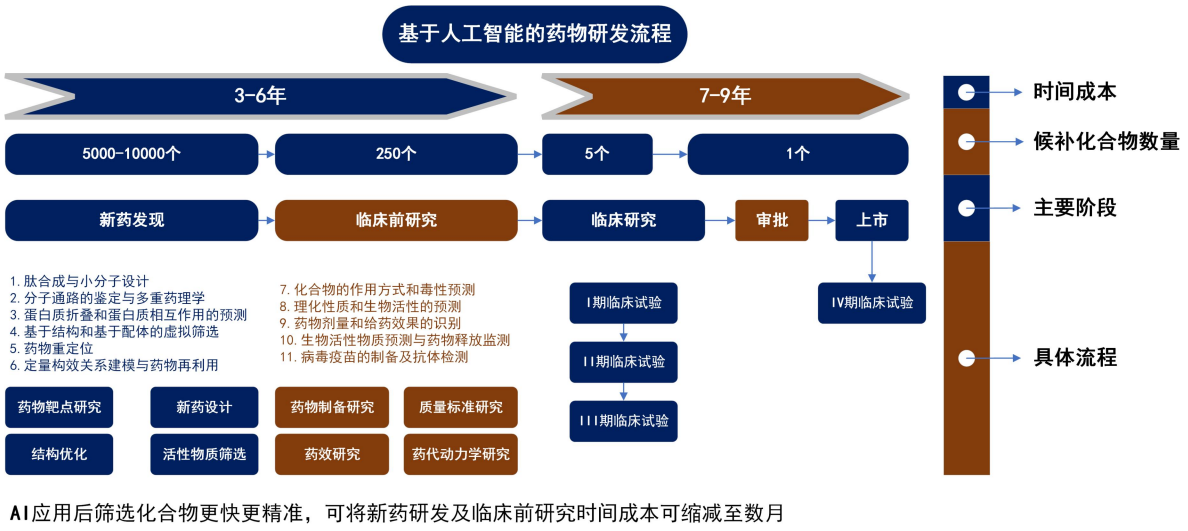

图 2-1 人工智能在药物发现与研发中的应用

（一）肽合成与小分子设计

多肽是一种由大约 2 至 50 个氨基酸组成的生物活性小链，由于它们具有跨越细胞屏障的能力并可以到达所需的靶点，因此越来越多地被用于治疗。深度学习于肽合成与小分子设计的应用概念图如

图 2-2 所示。近年来，研究人员利用人工智能的优势发现了新肽。例如，Yan 等人[2]在 2020 开年发了基于 DL 的短抗菌肽（Antimicrobial peptides，AMPs）鉴定平台 Deep-AmPEP30。AmPEP30 是一种卷积神经网络（Convolutional Neural Network, CNN）驱动的工具，可以根据脱氧核糖核酸（DeoxyriboNucleic Acid，DNA）序列数据预测短 AMP。通过利用该平台，研究人员从一种存在于胃肠道的真菌病原体—光滑梭菌的基因组序列中鉴定出新的 AMPs。另外，小分子是分子量非常低的分子，就像肽一样，利用人工智能也可以用来探索小分子的治疗作用，如 Zhavoronkov 等人[3]设计了一种基于强化学习的小分子从头设计工具 GENTRL（https://github.com/insilicomedicine/GENTRL），并利用它发现了一种新的酶抑制剂，DDR1 激酶。McCloskey 等人[4]将 DNA 编码的小分子库（DNA-Encoded small molecule Libraries，DEL）数据与图卷积神经网络和随机森林等 ML 模型相结合，以发现新的类似药物的小分子。另外，Xing 等人[5]集成了 XGBoost、支持向量机和深度神经网络寻找与类风湿性关节炎有关的靶点的小分子。

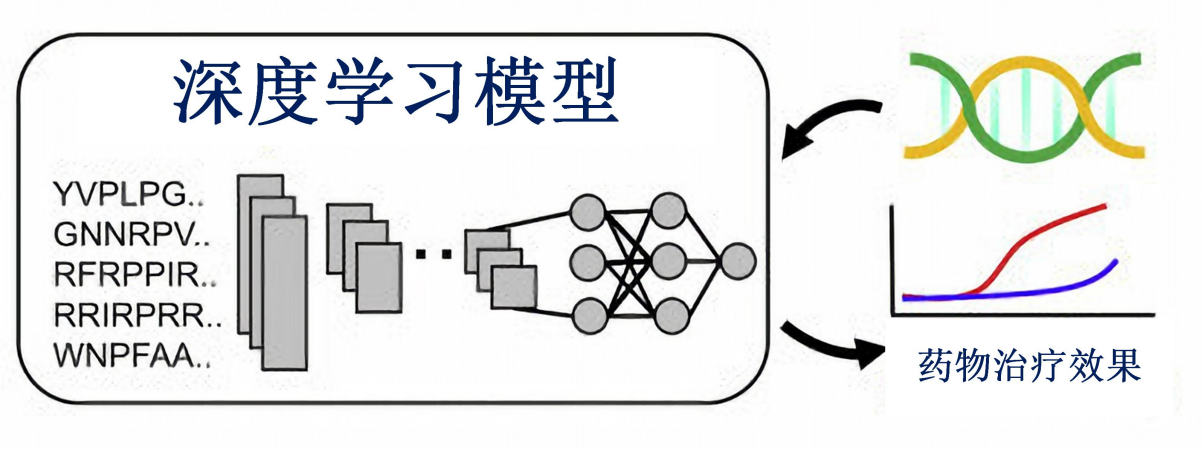

图 2-2 人工智能在肽合成与小分子设计的应用

（二）分子通路的鉴定与多重药理学

人工智能和最大似然算法在药物发现和开发中的重要成果之一是预测和估计疾病网络、药物—药物相互作用和药物—靶点关系的总体拓扑和动力学。如图 2-3 所示，数据库如 DisGeNET、STRTCH、STRING 分别被用于确定基因—疾病关联、药物—靶标关联和分子途径。例如，Gu 等人[6]在 2020 年使用相似性集成方法确定了 197 种最常用中草药的靶点，然后使用 DisGeNET 数据库将这些靶标与不同的疾病联系起来，从而将草药与可用于治疗的疾病联系起来。

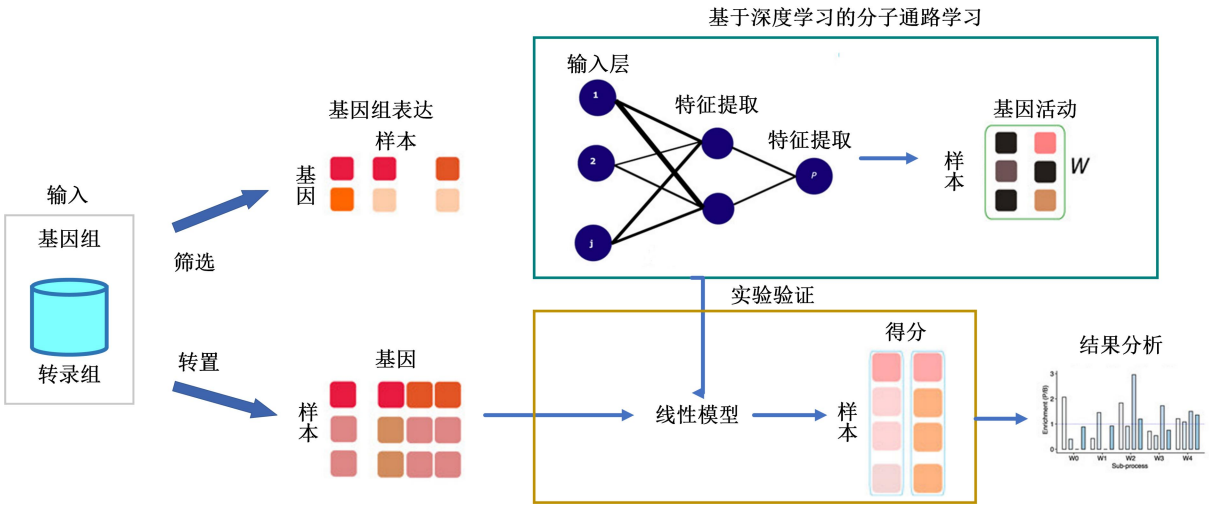

图 2-3 人工智能在分子通路的鉴定与多重药理学的应用

在药物化学中，多重药理学是指在与疾病相关的药物靶标生物网络中设计能够与多个靶点相互作用的单一药物分子。它适合于为复杂疾病，如癌症、神经退行性疾病（Neurodegenerative Diseases，NDDS）、糖尿病和心力衰竭等设计治疗剂。由于强大的挖掘能力和数据分析能力，基于 ML 的方法具有分析牵连分子网络的潜力，大大增加发现多靶配体的概率。此外，ML 模型有助于识别具有不同结合口袋的多靶配体。

（三）蛋白质折叠和蛋白质相互作用的预测

分析蛋白质—蛋白质相互作用（Protein-Protein Interaction, PPI）对于药物开发和发现至关重要，如图 2-4 所示。比如使用贝叶斯网络（Bayesian Network，BN）预测 PPI，其本质是利用基因共表达、基因本体（Gene Ontology，GO）和其他生物过程相似性，集成数据集产生精确的 PPI 网络。已有研究小组使用 BN 结合酵母菌的数据集研究出一种新的层次模型 PCA 集成极限学习机，该工具可以仅使用蛋白质序列信息来预测蛋白质—蛋白质相互作用，提供准确且快速的输出[7]。

机器和统计学习方法，如 K—最近邻、朴素贝叶斯、支持向量机、人工神经网络、决策树和随机森林，用于预测 PPI 中的障碍。通过使用基因共表达、基因本体（Gene Ontology，GO）和其他生物过程相似性数据训练贝叶斯网络，并获取精确和准确的 PPI 预测结果。文献[7]提出使用蛋白质序列信息来预测蛋白质—蛋白质相互作用的新型分层模型主成分分析集成极限学习机。

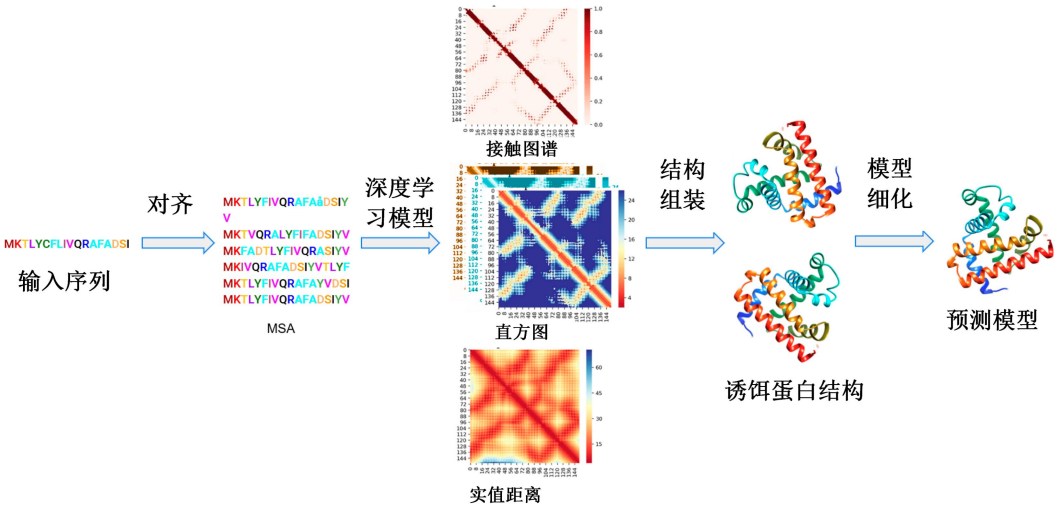

图 2-4 人工智能在蛋白质折叠和蛋白质相互作用的预测

（四）基于结构和基于配体的虚拟筛选

在药物设计和药物发现中，虚拟筛选（Virtual Screening，VS）是 CADD 的重要方法之一，是从化合物库中筛选出有前景的治疗化合物的有效方法。作为高通量筛选的重要工具，它也带来了成本高、准确率低的问题。要将 ML 用于 VS，应该有一个由已知的活性和非活性化合物组成的过滤训练集。这些训练数据用于使用监督学习技术训练模型。然后对训练的模型进行验证，如果它足够精确，则将该模型用于新的数据集，以针对目标筛选具有所需活性的化合物。ML 能够加快 VS 的速度，使其更完善，甚至可以减少 VS 中的误报。

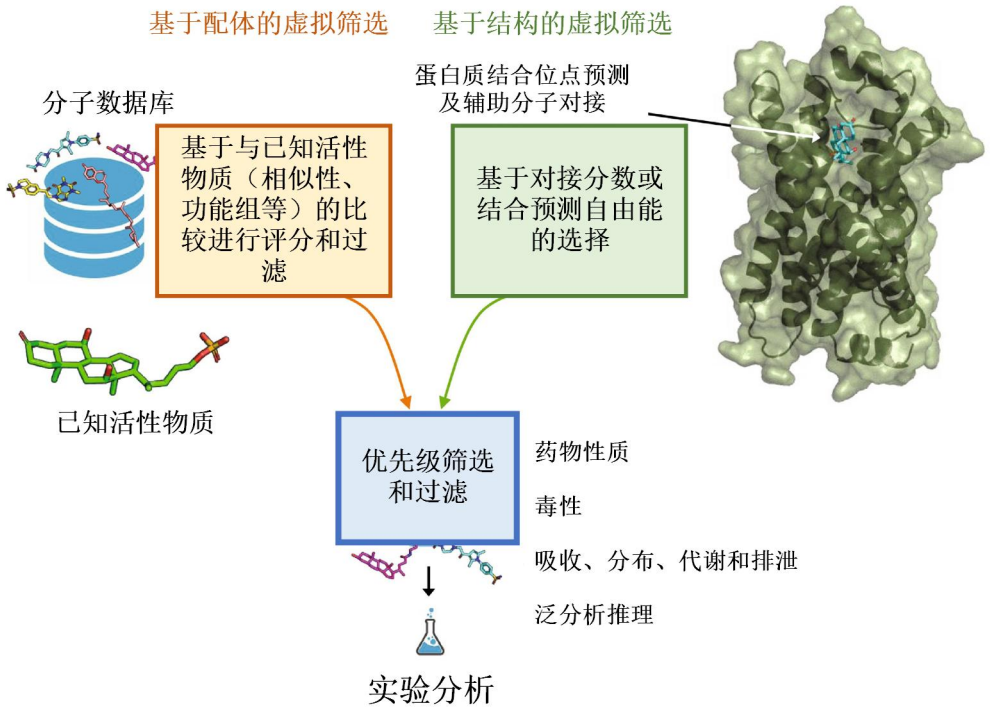

图 2-5 人工智能在药物虚拟筛选的应用

一般来说 VS 分两种（如图 2-5 所示），基于结构的 VS（Structural Based Virtual Screening，SBVS）和基于配体的 VS（Ligand Based Virtual Screening，LBVS）。其中，分子对接是

SBVS 中应用的主要原则，已经开发了几种基于 AI 和 ML 的评分算法[8]，如 NNScore、CScore、SVR-SCORE 和 ID-SCORE；也有算法被开发用于 SBVS 中的分子动态模拟分析以及预测 SBVS 中蛋白质—配体的亲和力，如支持向量机、卷积神经网络和浅层神经网络。类似的， LBVS 也开发了不同的算法和工具，如 SwissSimilarity[9]、METADOCK[10]、HybridSim-V[11]S 和 AutoDock Bias[12]等等。

（五）药物重定位

在新药研发中，先导化合物的筛选是至关重要的，人工智能在识别新的和潜在的先导化合物方面发挥着巨大的作用。在化学空间中有大约 1.06 亿个化学结构，他们来自不同的研究，如基因组研究、临床和临床前研究、体内分析和微阵列分析。利用机器学习模型，如强化模型、Logistic 模型、回归模型和生成模型，根据活性位点、结构和靶结合能力可以筛选出这些化学结构。

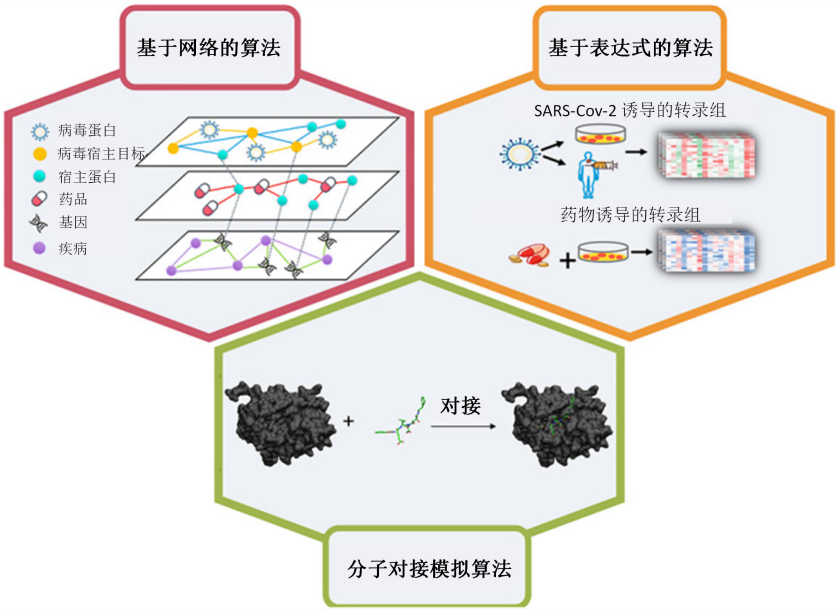

图 2-6 基于 AI 和 ML 的药物重定位方法，包括基于网络的算法、基于表达的算法和分子对接模拟算法

新型冠状病毒肺炎（COVID-19）是由严重急性呼吸综合征冠状病毒 2 型（SARS-CoV-2）引起的急性呼吸道传染病，已成为前所未有的公共卫生危机，对人类生命安全构成严重威胁。自 COVID-19 疫情爆发以来，众多研究人员在使用计算方法来鉴定用于治疗 COVID-19 的候选药物方面做出了重大努力。在本节中，总结了基于 AI 或 ML 的 COVID-19 治疗药物重定位方法的一般类别，包括基于网络的算法，基于表达的算法和分子对接模拟算法（图 2-6 和表 2-1）。

表 2-1 基于 AI 或 ML 的药物重定位研究

AI 或 ML 方法	项目详细信息	代码
正则化拉普拉斯算子[13]	用于识别 SARS-CoV-2 相互作用者的网络标签传播	https://github.com/Murali-group/SARS-CoV-2-network-analysis
密集的全卷积神经网络[14]	通过虚拟药物筛选识别和排序蛋白质-配体相互作用	无
自然语言处理[15]	MT-DTI 筛选潜在的抗病毒药物	无
卷积神经网络[16]	具有结合亲和力的药物-靶标相互作用的识别和排序	无
集成深度学习方法[17]	通过知识图网络发现候选药物	https://github.com/ChengF-Lab/CoV-KGE
图卷积网络[18]	多理变分图自编码器对药物的识别和排序	https://github.com/yejinjkim/drug-repurposing-graph
具有注意机制的图卷积网络[17]	通过医学知识图谱发现候选药物	https://github.com/FangpingWan/NeoDTI
图卷积网络[19]	病毒相关知识图谱的构建	https://github.com/FangpingWan/CoV-DTI
图卷积网络、网络扩散和网络邻近度[20]	通过药物功效筛选鉴定和排序病毒-宿主相互作用	https://github.com/Barabasi-Lab/COVID-19
基于人工智能的平台-InfinityPhenotype[21]	转录组数据分析	无

人工神经网络[22]	转录组、蛋白质组、结构数据和老化特征分析	https://github.com/uhlerlab/covid19_repurposing
具有多头注意力机制的图卷积网络[23]	受新化学物质干扰的基因表达谱分析	https://github.com/pth1993/DeepCE
卷积神经网络[24]	序列同一性和结构相似性分析，分子对接	无
随机森林[14]	对接模拟分数的预测	无
端到端的深度神经网络[25]	蛋白质配体相互作用概率预测，药物对接算法验证	https://github.com/ekraka/SSnet
朴素贝叶斯[26]	基于各种结合能函数的排序，通过对接方法进行验证	无

①基于网络的算法：药物重定位的经典方法是将基于网络的算法[27,28]应用于包含不同类型医学实体（如疾病、药物和蛋白质）之间关系的知识图谱，以识别相关的宿主蛋白靶标或宿主相互作用组的区域。病毒—宿主网络基于这样的假设，即属于复合物或信号通路的人类蛋白质最接近病毒蛋白质的人类相互作用物，并且是潜在的良好抑制目标。Law 等人[13]利用一种特殊的网络标签传播方法与基于正则化拉普拉斯算子 (Regularized Laplacian，RL) 的半监督学习方法相结合，对病毒—宿主相互作用组进行识别，以识别额外的 SARS-CoV-2 相互作用者。基于 5 折交叉验证，RL 在接受者—操作者（Receiver-Operator）特征曲线下的面积达到 0.76。此外，所提出的方法发现了内质网应激、HSPA5 和抗凝血剂之间的联系。上述方法表明，应用于 COVID-19 药物重定位研究的基于网络的策略可以通过整合多种类型的模块，包括病毒—宿主相互作用、蛋白质—蛋白质相互作用组网络和药物靶点来识别有效的可用药物和药物目标网络。

②基于表达的算法：药物重定位 FDA 批准的抗 SARS-CoV-2 药物的一种方法是观察疾病状态下防御基因表达的变化，其可用作有效的疾病描述符或定量表型，同时可用于驱动相反方向的基因表达[29]。Zhu 等人[21]将转录组数据输入基于 AI 的平台 InfinityPhenotype，以揭示天然产物或 FDA 批准药物的功效。实验结果表明，甘草素（Liquiritin）通过模仿 I 型干扰素发挥抗病毒作用，可作为治疗 COVID-19 的竞争候选药物。

③分子对接模拟算法：最近，分子对接模拟的显着改进[30]以及 AI 和 ML 技术的进步已被用于彻底改变药物开发过程。Nguyen 等人[24] 充分利用了数学姿势（MathPose）和卷积神经网络（MathDL）来预测 SARS-CoV-2 3CL 蛋白酶的空间结构和蛋白质—配体结合亲和力。MathPose 对接选定的已知复合物，生成的诱饵复合物被送入 MathDL 进行药物性质评估。根据预测的结合亲和力，作者向 COVID-19 报告了前 15 种潜在的强效药物，这为进一步的药物再利用提供了关键步骤。

在过去的一年中，AI 和 ML 算法[31–33]广泛促进了新型冠状病毒治疗药物研发过程。通常，计算式药物研发系统由三个单元组成：① 靶点发现；② 小分子药物发现；③ 预测临床试验结果（图 2-7）。所有研究总结在表 2-2 中。通过利用目标蛋白的晶体结构和同源模型的知识[34]，Zhavoronkov 等人[35]提出了一种生成化学管道来设计 COVID-19 的新型药物样抑制剂，并通过利用深度学习中的生成自动编码器、生成对抗网络、遗传算法和语言模型来研究药物的化学性

质，产生了几种新的药物化合物以供进一步开发。通过结合蒙特卡洛树搜索算法和多任务神经网络，Srinivasan 等人[36]提出了一种计算方法来发现治疗 COVID-19 的新治疗剂。研究表明，迭代探索设计空间并同时提高替代分子模型准确性的搜索和优化策略显著加快了候选药物的发现。

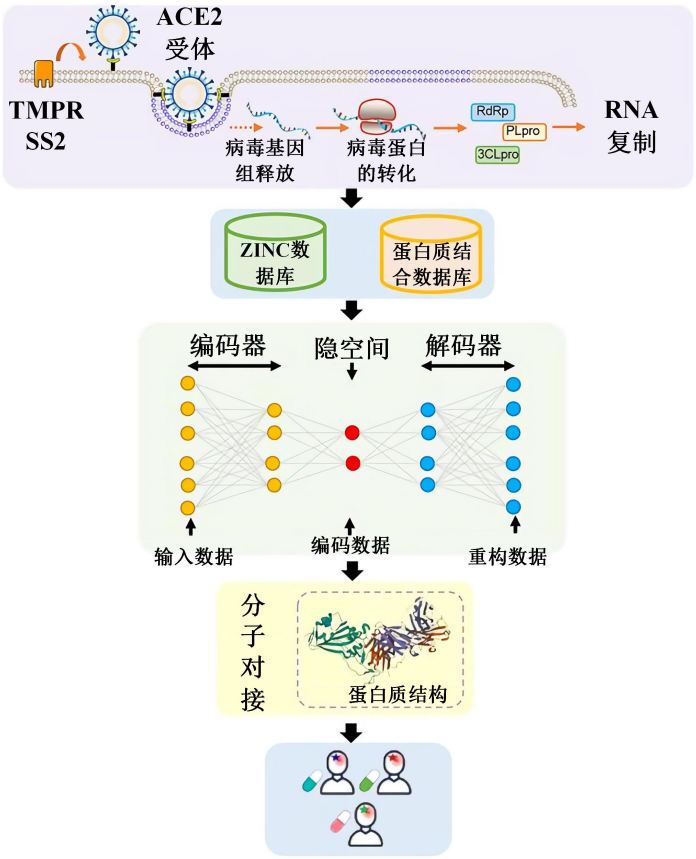

图 2-7 基于 AI 和 ML 方法的药物研发过程

表 2-2 基于 AI 或 ML 的新型冠状病毒药物研发

AI 或 ML 方法	项目详细信息	代码
生成化学管道[35]	利用目标蛋白的晶体和同源模型知识设计新药	无
深度 Q 学习网络[37]	抗病毒药物采集、化学规则拆分结构、片段采集、具有医学化学知识的片段库	https://github.com/tbwxmu/2019-nCov
深度对接模型[38]	通过基于结构的虚拟筛选识别潜	无

	在药物	
深度生成模型[39]	通过分子 SMILES 变分自编码器和有效的多属性控制采样方案生成新药	无
深度生成模型[40]	通过迁移和强化学习生成小分子	无
定向消息传递神经网络结合迁移学习[41]	通过虚拟筛选识别新药	https://github.com/pkuwangsw/COVIDVS
蒙特卡洛树搜索算法和多任务神经网络[36]	通过迭代搜索和再训练策略发现新的候选配体	无

（六）定量构效关系建模与药物再利用

在药物设计和开发中，研究化学结构和理化性质与生物活性之间的关系是至关重要的。定量构效关系（Quantitative Structure-Activity Relationship，QSAR）建模是一种计算方法，通过它可以在化学结构和生物活性之间建立定量的数学模型。传统 QSAR 模型大致分为两类，回归模型（如高斯过程（Gaussian Process，GPs））和分类模型。目前已经开发了多种基于网络的工具和算法，如 Vega 平台 (https://www.vega-qsar.eu/)、QSAR-Co[42]、Transformer-CNN[43]和 Chemception (https://github.com/Abdulk084/Chemception) 等，为 QSAR 建模提供了一条新的途径。在药物设计和发现中，药物重新定位是指对已经针对一种疾病情况开发的药物进行调查，并针对其他疾病情况进行重新定位（如图 2-8 所示）。近年来，基于人工智能的工具和算法的出现为该领域研究提供了平台，如 DrugNet，DRIMC （https://github.com/linwang1982/DRIMC）和 DRRS[44]等。特别是最近，新冠肺炎成为一种全球性的流行病，世界各地的研究

人员开始寻找有前途的治疗剂。在这方面，Hooshmand 等人[45]基于神经网络进行药物重新定位，确定了 16 种潜在的抗 HCoV 可再利用药物，并为新冠肺炎确定了 12 个有前景的药物靶点。

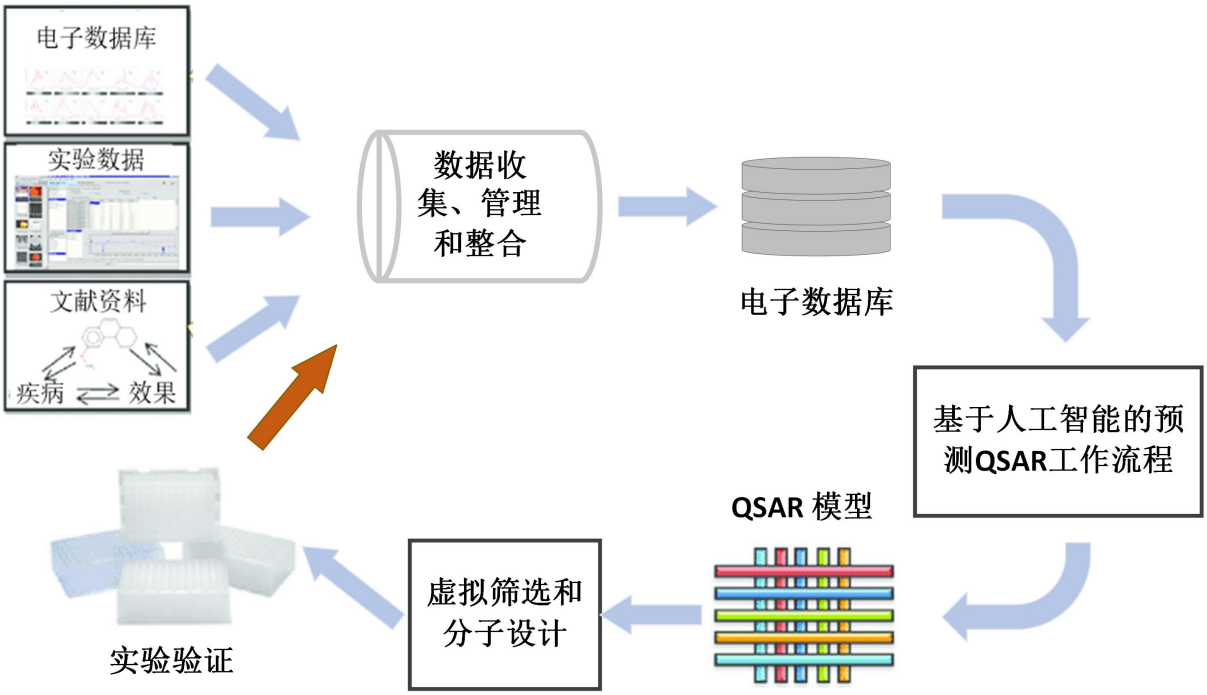

图 2-8 人工智能在定量构效关系建模与药物再利用的应用

（七）化合物的作用方式和毒性预测

药物毒性是指化学分子由于化合物的作用方式或新陈代谢方式而对生物体产生的不利影响。如图 2-9 所示，人工智能可以预测药物分子与靶点结合和未结合时的效应，以及体内安全性分析。已经开发了不同的基于 Web 的工具，如 LimTox、pkCSM、admetSAR 和 Toxtree。DeepTox[46]和 PrOCTOR[47]，用于预测新的毒性化合物和临床试验中毒性概率的预测和分类。例如，Robledo-Cadena 等人[48]使用 PrOCTOR 预测了非甾体抗炎药对顺铂、紫杉醇和阿霉素对宫颈

癌细胞的疗效的影响，并确定了 2,576 种小分子的新适应症，结合了 16 种不同的药物特征，用于治疗 帕金森和 2 型糖尿病。

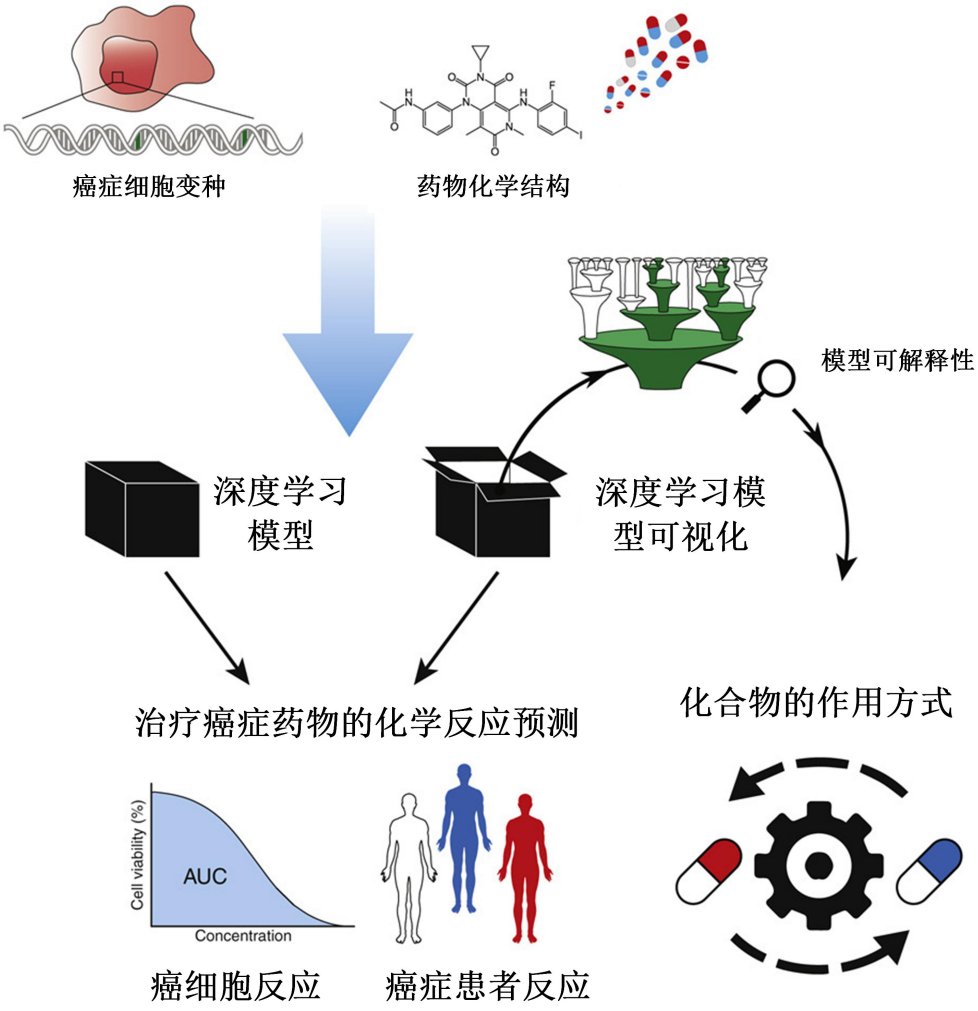

图 2-9 人工智能在化合物的作用方式和毒性预测的应用

（八）理化性质和生物活性的预测

众所周知，每一种化合物都与溶解度、分配系数、电离度、渗透系数等物理化学性质有关，这可能会阻碍化合物的药代动力学特性和药物靶向结合效率。因此，在设计新的药物分子时，必须考虑化合物的物理化学性质。为此，已经开发了不同的基于人工智能的工具来预测这些性质（如图 2-10 所示），包括分子指纹、SMILES

格式、库仑矩阵（Coulomb matrices）和势能测量，这些都用于深度神经网络（Deep Neural Networks，DNN）训练阶段。

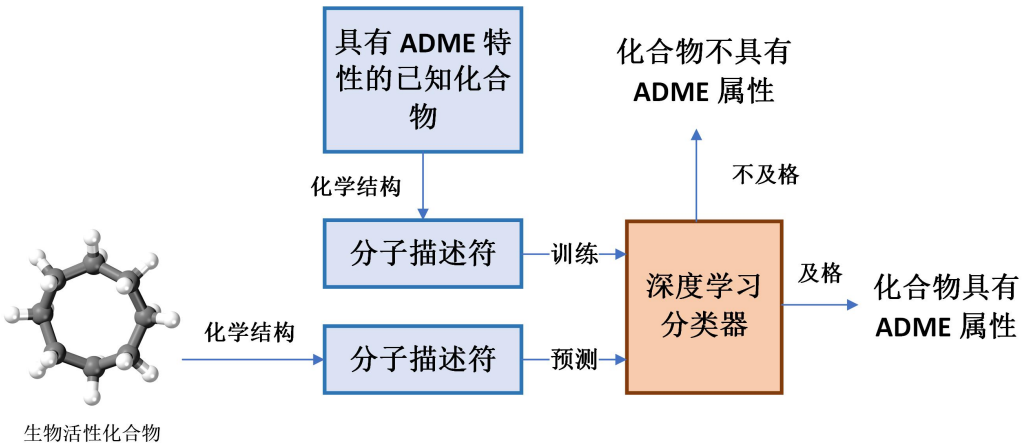

图 2-10 人工智能在理化性质和生物活性的预测的应用

此外，药物分子的治疗活性取决于其与受体或靶点的结合效率，因此，预测化学分子与治疗靶点的结合亲和力对于药物的发现和开发至关重要。人工智能算法的最新进展增强了该过程，使用相似性特征已经开发了几个基于网络的工具，如 ChemMapper 和相似集合方法。此外，还构建了基于 ML 和 DL 的药物靶标亲和力识别模型[49]，如 KronRLS、SimBoost、DeepDTA 和 Padme 等。

（九）药物剂量和给药效果的识别

给病人任何不适当剂量的药物都可能导致不良和致命的副作用，多年来，确定能够以最小毒副作用达到预期效果的药物的最佳剂量一直是一个挑战。随着人工智能的出现，许多研究人员正在借助 ML 和 DL 算法来确定合适的药物剂量，如图 2-11 所示。例如，Shen 等人[50]开发了一个基于人工智能的平台，称为 AI-PRS，用于确定通过抗逆转录病毒疗法治疗艾滋病毒的最佳剂量和药物组合。AI-PRS 是

一种神经网络驱动的方法，它通过抛物线响应曲线（Parabolic Response Curve，PRC）将药物组合和剂量与疗效联系起来。在他们的研究中，10 名 HIV 患者联合使用替诺福韦、法韦伦和拉米夫定，AI-PRS 分析表明替诺福韦的剂量可以减少起始剂量的 33%，而不会导致病毒复发。Tang 等人[51]使用人工神经网络、贝叶斯加性回归树、增强回归树、多元自适应回归样条等 ML 技术来确定免疫抑制药物他克莫司的最佳剂量。此外，Hu 等人[52]使用分类和回归树、多层感知器网络、K—最近邻等技术进行 ML 分析，以找出心脏药物地高辛（Digoxin）的安全初始剂量。此外，Imai 等人[53]开发了一个决策树模型来寻找抗生素药物万古霉素（Vancomycin）的安全起始剂量。

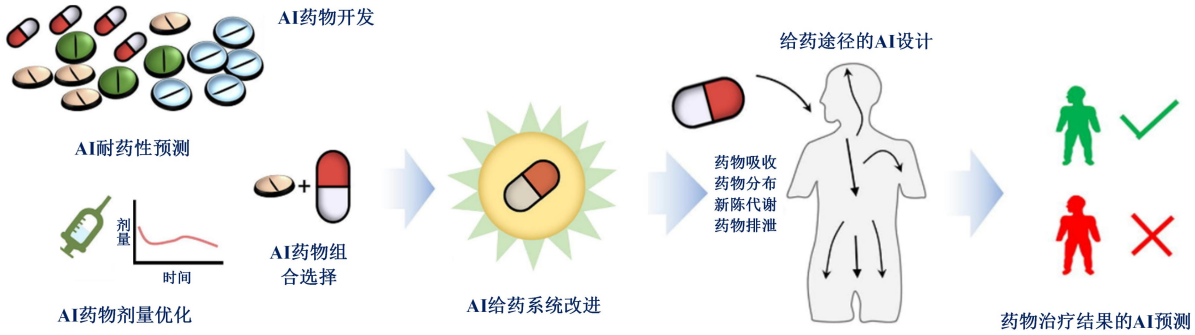

图 2-11 人工智能在药物剂量和给药效果的识别应用

（十）生物活性物质预测与药物释放监测

最近研究已经开发了多种在线工具来分析药物释放，以及选定的生物活性化合物作为载体的可行性，其概念图如图 2-12 所示。最常用的是基于化学特征的药效团评价。为了研究基于配体的化学性质，已经使用 CATALYST 程序建立了各种成功的实验。此外，利用人工智能研究人员可以确定用于与疾病相关的特定靶点的生物活性化合物。

例如，Wu 等人利用集成 DL 和 RF 的方法设计了 WDL-RF (https://zhanglab.ccmb.med.umich.edu/WDL-RF/) 用于测定靶向配体的 G 蛋白偶联受体（G Protein-Coupled Receptors，GPCRs）的生物活性。同样，Cichonska 等人[54]开发了用于确定化合物的生物活性的多核学习方法 pairwiseMKL（https://github.com/aalto-ics-kepaco）[55]，并用于预测化合物的抗癌效力。

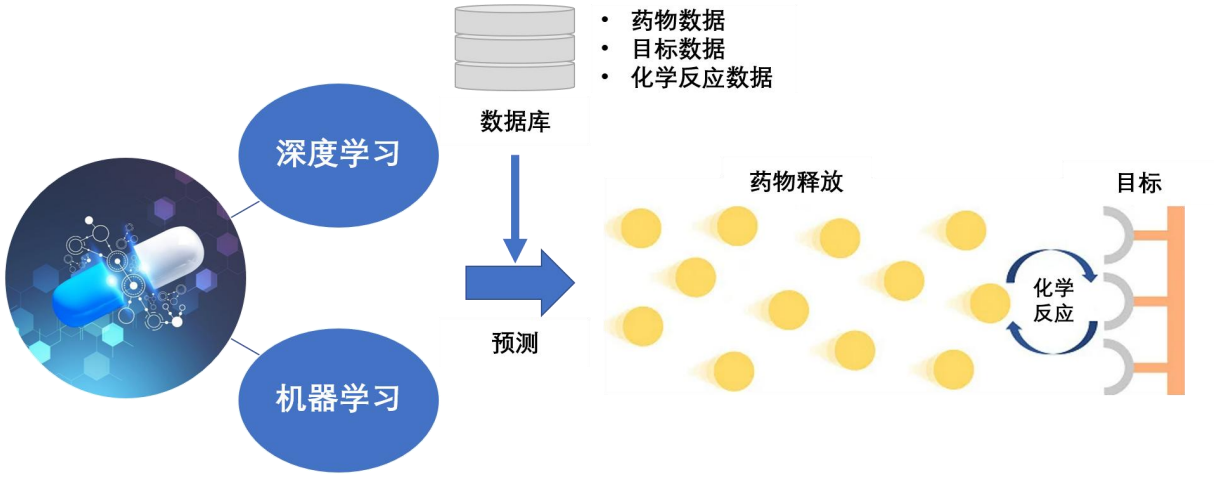

图 2-12 人工智能在生物活性物质预测与药物释放监测的应用

（十一）病毒疫苗的制备及抗体检测

由于数据量巨大，并且需要自动抽象特征学习，人工智能在疫苗发现领域有着重大贡献（图 2-13 和表 2-3）。COVID-19 冠状病毒疾病疫苗的深度学习和机器学习模型主要集中在人工神经网络、梯度提升决策树和深度神经网络等预测算法模型中[56]。Fast 等人[57]使用两种名为 MARIA 和 NetMHCPan4 的人工神经网络算法来识别 SARS-CoV-2 的 T 细胞和 B 细胞表位。该方法识别了 405 个 T 细胞表位（在 MHC-I 类分子和 MHC-II 类分子的结构中都具有很强的分

子代表性分数）以及 S 蛋白上的两个潜在中和 B 细胞表位。这一发现将促进针对 COVID-19 的强效疫苗和中和抗体的开发。

表 2-3 基于 AI 或 ML 的新型冠状病毒疫苗制备研发

AI 或 ML 方法	项目详细信息	代码
人工神经网络[57]	基于病毒蛋白抗原呈递和抗体结合特性鉴定 SARS-CoV-2 T 细胞和 B 细胞表位	无
XGBoost[58]	从非结构蛋白预测候选疫苗	无
前馈神经网络[59]	通过结合稳定性预测来自 SARS-CoV-2 病毒的 HLA 结合肽	无
深度神经网络[60]	可应对病毒变异的多表位疫苗的预测与设计	https://github.com/zikunyang/DCVST

图 2-13 人工智能在病毒疫苗的制备及抗体检测的应用

此外，Prachar 等人[59]提出通过前馈神经网络，确定了 174 个 SARS-CoV-2 表位，这些表位可以稳定地结合 11 个 HLA 同种异体，具有较高的预测结合分数。重要的是，作者评估了目前识别 SARS-CoV-2 相关表位的肽 HLA 预测工具。结果表明，几种算法表现出低稳定性，因此预测的肽很可能引发针对 SARS-CoV-2 的免疫反应。基于此，该研究中经验证的结合或非结合肽被用于进一步开发 COVID-19 的疫苗和治疗。Yang 等人[60]还提出了一种基于深度神经

网络的方法，名为 DeepVacPred，用于预测和设计多表位疫苗。DeepVacPred 构建了一种 694aa 多表位疫苗，包含 16 个 B 细胞表位、82 个 CTL 表位和 89 个 HTL 表位。此外，还对 SARS-CoV-2 的 RNA 突变进行了跟踪，以确保设计的疫苗能够应对病毒的突变。

（十二）本章小结

在过去的几年里，制药行业的数据数字化有了很大的增长。然而，数字化带来的挑战是如何应用这些数据来解决复杂的临床问题。这激发了人工智能的使用，因为它可以通过增强的自动化处理大量数据。人工智能是一个以技术为基础的系统，包括各种先进的工具和网络，可以模仿人类的智能。人工智能利用能够解释和学习输入数据的系统和软件，为实现特定的目标做出独立的决定。人工智能在医药领域的应用正在不断扩大。

三、人工智能在制药行业的应用现状

（一）人工智能在药物发现领域的市场情况

药物发现是一个众所周知的漫长、复杂和昂贵的过程，需要世界上最聪明的人共同努力。随着每一篇新的研究论文的发表和每一种新化合物的测试，理解人类生理和分子机制的复杂性不断增加。随着世界在试图适应和抵御冠状病毒方面面临新的挑战，人工智能提供了新的希望，一种治疗方法可能比以往任何时候都更快发展。

最近对药物开发人工智能的大量投资意味着这些初创公司拥有开发技术的人力和资源。与医疗成像领域的人工智能相比，总投资已经增长了四倍多，尽管两个行业的初创企业数量相当。图 3-1 显

示了基于人工智能的新药研发所需要的条件及关键性技术竞争点。这使得人工智能在药物开发中的平均交易规模比医学成像大 3.5 倍。由于这些人工智能初创企业的员工总数目前在全球接近 1 万人，这笔资金已用于大幅扩大和建设产能。

（二）人工智能新药研发所需要的条件及关键性技术竞争点

人工智能成为国际竞争的新焦点。人工智能是引领未来的战略性技术，世界主要发达国家把发展人工智能作为提升国家竞争力、维护国家安全的重大战略，加紧出台规划和政策，围绕核心技术、顶尖人才、标准规范等强化部署，力图在新一轮国际科技竞争中掌握主导权。针对基于人工智能的新药研发，企业必须把人工智能发展放在企业战略层面系统布局，打造竞争新优势、开拓市场新空间。

图 3-1 基于人工智能的新药研发所需要的条件及关键性技术竞争点

①数据质量：构建有效且可解释的药物发现模型的方法之一是使用与药物相关的实体来构建神经网络模型。然而，这样的算法具有固有的局限性。例如，为了定义和获得更大范围的药物—靶点相互作用数据，设置了较低的亲和力阈值，这可能会导致预测性能出

现偏差。此外，大多数研究都没有评估药物—靶点相互作用可能是功能关联而不是物理结合的可能性，这也影响了数据的质量。

②算法设计：大多数研究中使用的低水平穷举对接算法在寻找具有局部最小值的受体—配体相互作用时遇到困难，导致高度的亲和力可变性。此外，对接算法的高计算负担限制了其在大型化合物库中的应用。此外，对接算法在许多情况下使用不同的评估评分标准生成不同的候选集。因此，对接算法的选择、算法结构的设计和评价标准的设置都需要系统地改进。

③临床试验：药物或疫苗开发与临床应用的计算工作之间的转化差距是计算生物学和医学领域的一个主要且被广泛认可的瓶颈。许多预测的药物和疫苗尚未进入临床试验。尽管如此，由于难以确定临床终点和招募患者队列，因此缺乏对临床益处的优化。

④硬件设施：在人工智能领域进行高额投资，但仅靠人工智能技术是远远不够的，还需要相配套的硬件来支撑高性能算法，同时不占用太多空间和消耗过多电力，以适应临床实际应用。在人工智能核心硬件方面，根据应用领域不同，可以将智能芯片分为云端人工智能芯片、边缘人工智能芯片、新型人工智能芯片 3 类。在云端，通用图形处理器（Graphic Processing Unit，GPU）被广泛应用于神经网络训练和推理；张量处理单元（Tensor Processing Unit，TPU）等定制人工智能芯片使用专用架构实现了比同期中央处理器（Central Processing Unit，CPU）和 GPU 更高的效率；现场可编辑逻辑门阵列（Field Programmable Gate Array，FPGA）在云端推理应

用中也占有一席之地，具有支持大规模并行、推理延时低、可变精度等特点。在边缘计算领域，智能手机是目前应用最为广泛的边缘计算设备，药物研发是未来边缘人工智能计算的最重要应用之一，为此，推理计算能力、功耗和成本是应用于边缘设备人工智能芯片关注的主要因素。目前，云端和边缘设备在各种人工智能应用中通常是配合工作的，随着边缘设备能力不断增强，越来越多的计算工作负载将在边缘设备上执行。新型人工智能芯片主要包括神经形态芯片、近内存计算芯片、存内计算芯片等，目前仍处于探索研发阶段。

总括而言，本报告预计未来用于药物和疫苗开发的先进 AI/ML 方法将在输出生成、多维数据集成、算法结构部署和工作机制解释方面得到改进。

（三）国际顶级医疗公司在人工智能新药研发的最新研究成果

在全球范围内，人工智能制药领域自 2014 年以来兴起。生成对抗网络（Generative Adversarial Networks，GANs）的出现促使该行业利用该技术探索化学分子的生成；同时，图像处理、语音识别等技术也被用于小分子识别和目标发现。2014 至 2015 年是 AI 医药领域发展的起步阶段。第一批成长中的 AI 药企（包括 Exscientia、Insilico Medicine、晶泰科技等）大多诞生于这段时间，并陆续完成早期融资。彼时，生态圈对人工智能药物的认可度并不高。发展初期，AI 制药企业资金和新药研发实力不足，同时也需要进行大量的

前期技术积累工作。而此阶段的 AI 制药初创企业，商业模式几乎都以提供技术服务为主。

接下来的两年（2016 至 2017 年）是 AI 药业相对缓慢的两年。整个行业经历了缓慢发展的过渡期。虽然人工智能在新药研发的某些环节可以加速和提高效率，但其整体优势并不明显。在这阶段，一些 AI 医药初创企业开始尝试纵向拉长技术服务链条，不仅是为了提高新药研发某个点或环节的效率，而是追求更端到端的解决方案（如直接提供分子化合物）。

2018 年，AI 医药领域的发展终于有了小突破和爆发。首批成立的 AI 药企，包括 Exscientia、Atomwise、Recursion、Insilico Medicine 等，开始获得临床候选分子的验证结果。越来越多的人开始相信借助 AI 做药的可能性，也有越来越多的新成员加入到 AI 制药赛道，尤其是在中国。全球已上市的制药企业见下表 3-1，重点关注美国、德国、英国和中国。

表 3-1 全球已上市或进入临床的人工智能制药企业

药物研发公司	国别	研发技术	技术来源	研发合作情况
Benevolent AI	英国	药物重定位	自研	主要以人工智能技术，借助深度学习和知识图谱，提取出能够推动药物研发的知识和新的可以被验证的假说，从而加速药物研发的过程。并通过知识网络图谱分析与识别，筛选出巴瑞替尼（Baricitinib）对治疗新冠肺炎的潜在疗效。
Exscientia	英国	蛋白质折叠和蛋白质相互作用的预测	自研	Exscientia 是一家英国药物研发 AI 技术服务提供商，主要业务是利用已开发的人工智能平台进行自动化药物的研发指导，利用大数据和机器学习方法，自动设计出上百万种与

				特定靶标相关的小分子化合物，缩短新药研发进程。
Atomwise	美国	化合物的作用方式和毒性预测	自研	建立了第一个基于结构的药物设计的深层神经网络，帮助研究人员应对慢性疾病、癌症、突发性硬化、糖尿病、埃博拉、疟疾、耐抗生素细菌等问题。其卷积神经网络具有自主学习空间，能够通过化学特性预测那些药物分子式可能是有效的，并避免毒性问题。
Recursion	美国	理化性质和生物活性的预测	自研	构建其机器学习药物发现平台，同时提供旨在大幅加速 NME 化学和预测安全药理学的新能力。此外，该公司将继续推进临床前和临床资产管线，包括脑海绵状血管畸形和神经纤维瘤的临床计划。2019 年，Recursion 与武田制药（Takeda）合作对武田制药 60 多种临床前和临床化合物的独特适应症的评估，并在 6 种以上的疾病中发现了新的候选治疗药物。
Insilico Medicine 英矽智能	中国	病毒疫苗的制备及抗体检测	自研	在其自主研发的人工智能平台的支持下，英矽智能已在前沿领域布局了快速增长的疗法组合。其内部在研项目中有 7 个项目进入 IND-enabling 阶段，包括一款用于 COVID-19 治疗的新型 3CL 蛋白酶抑制剂临床前候选药物，及两款分别靶向 MAT2A 和 USP1 的“合成致死”抗肿瘤疗法。此外，英矽智能内部在研项目中进展最快的抗纤维化项目也已成功完成了 0 期微剂量组试验，目前正在健康志愿者中进行 1 期临床试验。
齐鲁锐格医药	中国	药物重定位	合作开发	与英伟达密切合作，将高性能计算技术整合到锐格自主研发的 rCARD 平台，利用计算领域领先的 GPU 产品和技术，包括 NVIDIA Clara Discovery Pipeline 中的 GROMACS，Autodock-GPU 等套件，加速药物发现的进程，降低研发成本，提升药物开发的效率，助力锐格医药研发出更多、更好的潜在首创和最佳的创新药。
信华生物	中	理化性质	自研	聚焦于 AI 在肿瘤免疫治疗等面临巨大的未满

药业	国	和生物活性的预测		足医疗需求的领域，致力于打造 AI 驱动智能大分子药物开发平台，提升大分子药物研发的质量与效率。信华生物目前聚焦肿瘤免疫治疗等领域内未满足的医疗需求，多个项目已获得积极的生物验证数据。
复星医药	中国	蛋白质折叠和蛋白质相互作用的预测	合作开发	将英矽智能端到端人工智能驱动的药物发现平台，与复星医药强大的临床开发和商业推广能力相结合，发现和开发创新药物和疗法的组合，并获得英矽智能人工智能平台 PandaOmics 和 Chemistry42 的使用权，以推进公司内部人工智能驱动的药物发现和开发工作。
海和药物研究	中国	蛋白质折叠和蛋白质相互作用的预测	自研	中国工程院院士丁健领衔，建立了以“生物标志物指导下的”精准医疗平台；专注于抗肿瘤创新药物发现、开发、生产及商业化的中国领先的自主创新生物技术公司。
望石智慧	中国	分子通路的鉴定与多重药理学	自研	利用 AI 快速识别处理繁多结构化的研发数据，结合医药研发专家的领域知识，构建了亿级别的超高通量分子筛选系统、多维度分子生成系统和基于映射数据库的分子优化系统，体系化赋能更快、更好的新药发现。
晶泰科技	中国	蛋白质折叠和蛋白质相互作用的预测	合作开发	与辉瑞制药签订战略研发合作，融合量子物理与人工智能，建立小分子药物模拟算法平台，提高算法的精确度和适用广泛度，驱动小分子药物的创新。

国内 AI 制药相关技术企业基本都起步于 2019 年，疫情及国内的创新支持环境为这些科创企业带来了非常有利的机遇。从创业团队的背景来看，几乎都与国际三大 AI 制药企业引领者（Benevolent AI、Atomwise 和 Recursion）有着千丝万缕的联系，其核心技术大多来自创业团队的自主研发以及与院校合作研发。表 3-2 展示了国内外重点 AI 制药企业的概况以及相应的融资信息。

表 3-2 国内外重点 AI 制药企业的概况以及相应的融资信息

公司	融资信息
Benevolent AI	2019 年 9 月 17 日：私募股权轮 - BenevolentAI：9000 万美元 2018 年 4 月 19 日：一轮融资 - BenevolentAI：1.15 亿美元 2015 年 8 月 26 日：风险投资轮 - BenevolentAI：8700 万美元
Exscientia	Exscientia 在 9 轮融资中共筹集了 4.744 亿美元的资金。 他们的最新资金是在 2022 年 1 月 10 日通过 IPO 后股权融资筹集的。 2022 年 1 月 10 日：IPO 后股权 - Exscientia：1 亿美元 2021 年 7 月 8 日：Grant - Exscientia：150 万美元 2021 年 4 月 20 日：D 系列 - Exscientia：2.25 亿美元 * 更多的融资信息请查询网址： https://www.crunchbase.com/organization/exscientia/company_financials
Atomwise	2020 年 09 月 17 日：D 轮 - 9000 万美元 - MSD Partners、Octave Group、Shumway Capital、TpTfOCV Partners、OMX Ventures、Avenir Growth 2019 年 08 月 22 日：C 轮 5300 万美元 – Shumway Capital、TPTF、OCV Partners 2018 年 03 月 1 日：B 轮 3600 万美元 - Shumway Capital、Morgan Noble、Avenir Growth Capital
Recursion	2019 年 7 月 15 日，生物技术公司 Recursion Pharmaceuticals 宣布完成 1.21 亿美元 C 轮融资，本轮融资由 Baillie Gifford 的旗舰投资信托公司 Scottish Mortgage Investment Trust 领投，参与此轮融资的还有新投资者 Intermountain Ventures、明尼苏达大学（University of Minnesota）、德克萨斯理工大学（Texas Tech University System）的董事以及部分天使投资者。
Insilico Medicine 英矽智能	Insilico Medicine 在 10 轮融资中共筹集了 3.063 亿美元。 他们的最新资金是在 2022 年 1 月 11 日从 Venture - Series Unknown 轮中筹集的。 2022 年 1 月 11 日：风险投资 - Insilico Medicine 2021 年 6 月 22 日：C 系列- Insilico Medicine：2.55 亿美元 2020 年 4 月 20 日：B 系列-Insilico Medicine * 更多的融资信息请查询网址： https://www.crunchbase.com/organization/insilico-medicine/company_financials
齐鲁锐格医药	上海医药进行了 2 次投资。 他们最近的一次投资是在 2021 年 2 月 19 日，当时 SciClone Pharmaceuticals 筹集了 1.34 亿美元。 2021 年 2 月 19 日：风险投资轮 - SciClone Pharmaceuticals：1.34 亿美

	元 2016 年 12 月 21 日：天使轮-Healenvoy：人民币 300 万
信华生物药业	该公司无融资信息
复星医药	上海复星医药已在 1 轮融资中筹集了总计 2000 万美元的资金。 这是 2020 年 7 月 9 日进行的企业回合。 2020 年 7 月 9 日： 复星凯特生物企业融资：2000 万美元
海和药物研究	该公司无融资信息
望石智慧	2018 年 7 月 3 日，望石智慧获得数百万美元的天使轮融资，投资方为 SIG 海纳亚洲。 2020 年 3 月 16 日，望石智慧获得 1000 万美元的 A 轮融资，投资方为长岭资本、线性资本。
晶泰科技	2021 年 08 月，晶泰科技已完成近 4 亿美元 D 轮融资，由五源资本、奥博资本、厚朴资本领投，和暄资本、中国生物制药集团作为新股东跟投，腾讯投资、红杉资本、IMO Ventures 等早期股东继续追加投资。其他投资方还包含 Neumann Capital、Artisan Partners 等机构。值得关注的是，Artisan Partners 为美国知名的大长线基金。据了解，晶泰科技已秘密递交招股书，启动赴美上市。 2020 年 9 月，晶泰科技完成 3.188 亿美元 C 轮融资，由晨兴资本、软银愿景基金 2 期、人保资本联合领投，中金资本、招银国际招银电信基金、Mirae Asset（未来资产）、中证投资、中信资本、海松资本、顺为资本、方圆资本、IMO Ventures、Parkway 基金等多家来自全球的投资机构跟投，腾讯、红杉中国、国寿股权投资、SIG 海纳亚洲等早期股东继续追加投资 。

国外相关技术早在 2000 年左右即开始进行院校的科技成果转化，如上海海和药物研究开发股份有限公司等的创立，并以企业研发持续开展。相对而言，我国该领域的研发跟随国际，在新冠疫情之后研究团队开始增加，并注重成果转化的同步进行。基于中国国内院校学者的调研及文献报道，国内学者的研究重点同样集中在递送系统上，如表 3-3 所示：

表 3-3 AI 制药中国学者定位

单位及姓名	技术优势	设立企业	合作企业
中科院上海药物研究所 蒋华良院士	【肽合成与小分子设计】通过计算信息科学等多学科交叉，利用深度神经网络开展大数据和 AI 的药物研发，在多靶点药物分子设计以及药性拓展空间数据库（DrugSpaceX）有杰出贡献，同时还联合华为云发布了基于 ModelArts 平台的药物联邦学习服务。	上海君实生物医药科技股份有限公司 创始人	华为技术有限公司
中科院微生物研究所 王军教授	【理化性质和生物活性的预测】长期专注微生物组的生物信息学方法和应用研究。现阶段在机器学习、人工智能与微生物组结合的领域有多个前沿项目。通过利用深度学习算法，构建了构建了从人类肠道微生物组挖掘潜在新型抗菌肽的模型，预测准确率超 80%。	无	无
中科院深圳先进技术研究院 袁曙光研究员	【定量构效关系建模与药物再利用】主要研究方向为 CPCR 分子机理，计算机辅助药物设计，基于人工智能的药物筛选平台，疾病模型的建立，原生态细胞膜环境下 CPCR 与离子通道的结构与功能，计算机理性设计酶的功能与改造。	深圳阿尔法分子科技有限责任公司	无
北京大学 高毅勤教授	【蛋白质折叠和蛋白质相互作用的预测】发展针对生物和化学体系的理论与计算方法，特别是基于统计力学和机器学习的计算方法与软件的开发，染色质三维结、生物大分子和材料自组装、DNA 别构效应等领域的研究。2021 年 10 月，高毅勤课题组又与华为昇思 MindSpore 团队推出了基于 AlphaFold2 算法的蛋白质结构预测推理工具，效率同比提升 2 至 3 倍。	无	无
中国药科大学 陈亚东教授	【肽合成与小分子设计】主要方向为基于重大疾病的原创小分子药物发现，以及人工智能的药物分子设计及应用研究。团队自主研发的抗肿瘤化合物以 1.5 亿合同价格转让给上海复星医药产业有限公司，	无	北京深度智耀科技有限

	2019 年 11 月该药物获美国 FDA 孤儿药资格认定。		公司
清华大学 彭健教授	【蛋白质折叠和蛋白质相互作用的预测】主要研究领域为生物信息学、化学信息学和机器学习，其合作开发的算法在多项科学挑战赛中获得佳绩，包括蛋白质结构预测技术的关键测试，及转化医学和药物基因组学的挑战。	北京望石智慧科技有限公司	无

（四）本章小结

人工智能技术融合生物医药产业正迎来新的发展阶段，机遇和挑战并存。传统生物医药行业利用高通量筛选方式进行目标化合物的筛选，人力、时间和试错成本均很高。大数据时代下，人工智能平台开发的虚拟计算能基于基因和疾病的关联机制建立数据模型，高效寻找新药靶点，设计和合成有生物活性的分子，预测药物发现和开发过程中的药效、药代、毒性、稳定性等，大幅度降低研发成本，提升创新药研发整体效率和成功率，从而造福全球病患。

尽管取得了巨大的成功，但仍然存在许多挑战，其中有两个最重要的问题：首先，标记不能是二元的，因为药物在生物系统中的作用是复杂的；其次，虽然数据库拥有海量信息，但药物发现中可用的高质量数据并不多。因此，需要一个不仅能提供数据数量而且能提供质量的平台。尽管将人工智能工具融入药物发现过程中存在一些不可避免的障碍，还有大量的工作要做，但毫无疑问，在不久的将来，人工智能将给药物发现和开发过程带来革命性的变化。

四、专业术语解析

人工智能（Artificial Intelligence，AI）

是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

人工神经网络（Artificial Neural Network，ANN）

是由大量处理单元互联组成的非线性、自适应信息处理系统。它是在现代神经科学研究成果的基础上提出的，试图通过模拟大脑神经网络处理、记忆信息的方式进行信息处理。

计算机辅助药物设计（Computer Aided Drug Design，CADD）

是应用量子力学、分子动力学、构效关系等基础理论数据研究药物对酶、受体等的作用的药效模型,从而达到药物设计之目的。

卷积神经网络（Convolutional Neural Network，CNN）

是一类包含卷积计算且具有深度结构的前馈神经网络，是深度学习的代表算法之一。

深度学习（Deep Learning，DL）

是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。

机器学习（Machine Learning，ML）

是专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

蛋白质—蛋白质相互作用（Protein-Protein Interaction, PPI）

是在两个或多个蛋白质分子之间建立的高特异性物理接触，这是由包括静电力、氢键和疏水效应在内的相互作用引导的生化事件的结果。

定量构效关系（Quantitative Structure-Activity Relationship，QSAR）

是一种借助分子的理化性质参数或结构参数，以数学和统计学手段定量研究有机小分子与生物大分子相互作用、有机小分子在生物体内吸收、分布、代谢、排泄等生理相关性质的方法。

参考文献

[1] Han Y , Yang J , Qian X , et al. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies[J]. Nuclc Acids Research, 2019(8):e45-e45.

[2] Yan J , Bhadra P , Li A , et al. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning[J]. Molecular Therapy - Nucleic Acids, 2020.

[3] Zagribeln Yy B A , Zhavoronkov A , Aliper A , et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors[J]. Nature Biotechnology, 2019, 37(9):1038-1040.

[4] Mccloskey K , Sigel E A , Kearnes S , et al. Machine learning on DNA-encoded libraries: A new paradigm for hit-finding[J]. arXiv, 2020.

[5] Xing G , Liang L , Deng C , et al. Activity Prediction of Small Molecule Inhibitors for Antirheumatoid Arthritis Targets Based on Artificial Intelligence[J]. ACS Combinatorial Science, 2020, 22(12): 873–886.

[6] Gu S , Lai L H . Associating 197 Chinese herbal medicine with drug targets and diseases using the similarity ensemble approach[J]. Acta Pharmacologica Sinica, 2019, 41(3): 7.

[7] Zhu-Hong, You, Ying-Ke, et al. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis[J]. BMC Bioinformatics, 2013, 14(10): S10.

[8] Serafim, Mateus Sa Magalhaes, et al. The application of machine learning techniques to innovative antibacterial discovery and development. [J]. Expert opinion on drug discovery, England: 2020, 15(10): 1165 – 1180.

[9] Zoete V , Daina A , Bovigny C , et al. SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening[J]. Journal of Chemical Information & Modeling, 2016:1399 – 1404.

[10] Imbernon B , JM Cecilia, Perez-Sanchez H , et al. METADOCK: A parallel metaheuristic schema for virtual screening methods[J]. Experimental Mechanics, 2018, 32(6):789-803.

[11] Hongjian L , Kwong-S. L , Man-H. W , et al. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques[J]. Nucleic Acids Research, 2016(W1):W436-W441.

[12] Pablo A J , Modenutti C P , Demian A , et al. AutoDock Bias: improving binding mode prediction and virtual screening using known protein-ligand interactions[J]. Bioinformatics, 2019, 35(19): 3836 – 3838.

[13] Law J N , N Tasnina, Kshirsagar M , et al. Identifying human interactors of SARS-CoV-2 proteins and drug targets for COVID-19 using network-based label propagation[J]. arXiv preprint arXiv:2006.01968, 2020.

[14] Zhang, Haiping, et al. Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov.[J]. Interdisciplinary sciences, computational life sciences, 2020, 12(3): 368 − 376.

[15] Bo R B , Shin B , Choi Y , et al. Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model[J]. Computational and Structural Biotechnology Journal, 2020, 18:784-790.

[16] Majumdar S , Nandi S K , Ghosal S , et al. Deep Learning-Based Potential Ligand Prediction Framework for COVID-19 with Drug–Target Interaction Model[J]. Cognitive Computation, 2021(4): 1 − 13.

[17] Zeng X , Song X , Ma T , et al. Repurpose Open Data to Discover Therapeutics for COVID-19 using Deep Learning[J]. Journal of Proteome Research, 2020, 19(11): 4624 − 4636.

[18] Hsieh K L , Wang Y , Chen L , et al. Drug Repurposing for COVID-19 using Graph Neural Network with Genetic, Mechanistic, and Epidemiological Validation[J]. 2020.

[19] Ge Y , Tian T , Huang S , et al. An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19[J]. Signal Transduction and Targeted Therapy, 2021, 6(1): 165.

[20] Gysi D M , Valle T D , Zitnik M , et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19[J]. Proceedings of the National Academy of Sciences of the United States of America, 118(19):e2025581118.

[21] Zhu J , Deng Y Q , Wang X , et al. An artificial intelligence system reveals liquiritin inhibits SARS-CoV-2 by mimicking type I interferon. 2020.

[22] Belyaeva A , Cammarata L , Radhakrishnan A , et al. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing[J]. Nature Communications, 2021, 12(1): 1024.

[23] Pham, Thai-Hoang, et al. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing[J]. Nature machine intelligence, 2021, 3(3): 247 – 257.

[24] Nguyen D D , Gao K , Chen J , et al. Potentially highly potent drugs for 2019-nCoV[J]. bioRxiv, Cold Spring Harbor Laboratory, 2020.

[25] Karki N K , Verma N , Trozzi F , et al. Predicting Potential SARS-COV-2 Drugs - In Depth Drug Database Screening Using Deep Neural Network Framework SSnet, Classical Virtual Screening and Docking. 2020, 22(4).

[26] Mohapatra S , Nath P , Chatterjee M , et al. Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking[J]. PLoS ONE, 2020, 15(11):e0241543.

[27] Liu L , Zhang L R , Dao F Y , et al. A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation.[J]. Molecular therapy. Nucleic acids, 2021, 23: 347 – 354.

[28] Liu L , Li Q Z , Jin W , et al. Revealing Gene Function and Transcription Relationship by Reconstructing Gene-Level Chromatin Interaction.[J]. Computational and structural biotechnology journal, 2019, 17: 195 – 205.

[29] Killick R, Ballard C, Doherty P, et al. Transcription-based drug repurposing for COVID-19.[J]. Virus research, 2020, 290: 198176.

[30] Zulfiqar H , Masoud M S , Yang H , et al. Screening of Prospective Plant Compounds as H1R and CL1R Inhibitors and Its Antiallergic Efficacy through Molecular Docking Approach[J]. Computational and Mathematical Methods in Medicine, 2021, 2021(1):1-9.

[31] Hasan, Md Mehedi, et al. NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning.[J]. Briefings in bioinformatics, England: 2021, 22(6).

[32] Charoenkwan P , Nantasenamat C , Hasan M M , et al. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides[J]. Bioinformatics, 2021.

[33] Phasit C , Wararat C , Chanin N , et al. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides[J]. Briefings in Bioinformatics, 2021(6):6.

[34] Xu B , Liu D , Wang Z , et al. Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family[J]. Cellular and molecular life sciences : CMLS, 78(1):129-141.

[35] Zhavoronkov A , Aladinskiy V A , Zhebrak A , et al. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches[J]. 2020.

[36] Srinivasan S , Batra R , Chan H , et al. Artificial Intelligence-Guided De Novo Molecular Design Targeting COVID-19[J]. ACS Omega, 2021, 6(19):12557-12566.

[37] Tang B , He F , Liu D , et al. AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2[J]. BioRxiv, Cold Spring Harbor Laboratory, 2020.

[38] Ton A T , Gentile F , Hsing M , et al. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds.[J]. Molecular informatics, 2020, 39(8): e2000028.

[39] Chenthamarakshan V , Das P , Hoffman S C , et al. Cogmol: target-specific and selective drug design for COVID-19 using deep generative models[J]. arXiv preprint arXiv:2004.01215, 2020.

[40] Bung N , Krishnan S R , Bulusu G , et al. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence[J]. Future Medicinal Chemistry, 2021(2): 575 − 585.

[41] Wang S , Sun Q , Xu Y , et al. A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2[J]. Briefings in bioinformatics, Oxford University Press, 2021.

[42] Ambure P , Halder A K , H González-Díaz, et al. QSAR-Co: An Open Source Software for Developing Robust Multi-tasking or Multi-target Classification-Based QSAR Models[J]. Journal of Chemical Information and Modeling, 2019, 59(6): 2538 − 2544.

[43] Karpov P , Godin G , Tetko I V . Transformer-CNN: Swiss knife for QSAR modeling and interpretation[J]. Journal of Cheminformatics, 2020, 12(1): 17.

[44] Luo H , Li M , Wang S , et al. Computational Drug Repositioning using Low-Rank Matrix Approximation and Randomized Algorithms[J]. Bioinformatics(11): 1904 − 1912.

[45] Hooshmand S A , · Mohadeseh, Ghobadi Z , et al. A multimodal deep learning-based drug repurposing approach for treatment of COVID-19[J]. Molecular Diversity, 2020:3: 1717 − 1730.

[46] Mayr, Andreas, et al. DeepTox: toxicity prediction using deep learning[J]. Frontiers in Environmental Science, Frontiers, 2016, 3: 80.

[47] Kaitlyn M, Gayvert, Neel S, et al. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials[J]. Cell Chemical Biology, 2016, 23(10): 1294–1301.

[48] Gilvary C , Elkhader J , Madhukar N , et al. A machine learning and network framework to discover new indications for small molecules[J]. PLoS Computational Biology, 2020, 16(8):e1008098.

[49] Zelik R , H Ztürk, A Zgür, et al. ChemBoost: A Chemical Language Based Approach for Protein – Ligand Binding Affinity Prediction[J]. Molecular Informatics, 2020, 40(5): 2000212.

[50] Harnessing Artificial Intelligence to Optimize Long‐Term Maintenance Dosing for Antiretroviral‐Naive Adults with HIV‐1 Infection[J]. Advanced Therapeutics, 2020, 3(4): 1900114.

[51] Tang J , Liu R , Zhang Y L , et al. Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients[J]. Scientific Reports, 2017, 7:42192.

[52] Hu YH, Tai CT, Tsai CF, Huang MW. Improvement of Adequate Digoxin Dosage: An Application of Machine Learning Approach.[J]. Journal of healthcare engineering, 2018, 2018: 3948245.

[53] Imai S , Takekuma Y , Miyai T , et al. A New Algorithm Optimized for Initial Dose Settings of Vancomycin Using Machine Learning[J]. Biological & Pharmaceutical Bulletin, 2020, 43(1):188-193.

[54] Jiansheng W, Qiuming Z , Weijian W , et al. WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest[J]. Bioinformatics, 2018, 2(13):13.

[55] Cichonska, Anna, et al. Learning with multiple pairwise kernels for drug bioactivity prediction. [J]. Bioinformatics, 2018, 34(13): i509 – i518.

[56] Wang J , Wang H , Wang X , et al. Predicting Drug-target Interactions via FM-DNN Learning[J]. Current Bioinformatics, 2020, 15(1): 68 – 76.

[57] Fast E, Chen B. Potential T-cell and B-cell epitopes of 2019-nCoV[J]. BioRxiv, Cold Spring Harbor Laboratory, 2020.

[58] Ong E , Mei U W , Huffman A , et al. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. 2020, 11: 1581.

[59] Prachar M , Justesen S , Steen-Jensen D B , et al. Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools[J]. Scientific Reports, 2020, 10(1): 20465.

[60] Yang Z , Bogdan P , Nazarian S . An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study[J]. Scientific Reports, 2021, 11(1): 3238.

（作者：朱雅姝为清华大学国家金融研究院资本市场与公司金融研究中心高级研究专员。安硕为清华大学五道口金融学院副教授、博士生导师，资本市场与公司金融研究中心副主任。）

联系人：朱雅姝　　邮箱：zhuysh@pbcsf.tsinghua.edu.cn

行业图谱研究项目

一、项目目标和定位

行业图谱是资本市场与公司金融研究中心基于科技成果转化研究的一项子课题，聚焦于科技成果这一核心要素，从技术链视角切入展开的研究项目。科技成果的转化需要对科技成果有清晰、准确、深刻的认识和理解，能够解析科技成果所包含的学术价值、社会价值、经济价值和人文价值等，从而探索科技成果的未来应用场景，以跨越从0到1的商业性转化，通过不断理解优化实现社会产业化，并最终成为科技推动社会发展的历史进程。

然而，由于科技天然具有强大的认知壁垒，其先进性、创新性的特点，使得科技成果面临非专业人士看不懂、不敢判断的知识窘境。在成果转化的操作路径中，执行者可分类为三方：成果供给方、成果接收方及连接双方的中介服务机构。除了成果供给方之外，成果接收方和中介服务机构都面临着知识窘境。成果供给方是科技成果的发明人、创造者，对科技成果的学术价值拥有深度认知，但缺乏商业经验和分析社会需求的能力，很难独立实现成果的成功转化；成果接收方是进行成果商业化、产业化的企业，对社会需求敏感，善于进行商业价值的探索，但由于不具备深厚的科研基础，不能对科技成果进行技术层面的准确分析和判断，影响执行效率；中介服务机构虽然具备政策分析、法律服务等领域的专业能力，但同样面临看不懂技术的知识窘境，导致出现无效推介、不合理的专利布局、未来的专利纠纷等潜在危机。这一需求的断层也间接性地影响经济学称之为成果转化“死亡之谷”时期的存在。因此，如何准确认识科技成果，正确判断科技成果的技术领先度，理解科技成果所处的行业地位和产业链发展格局，对于提高科技成果转化具有极其重要的价值。

本研究以国家十四五规划为导向，重点关注与国家战略需求发展相关的重大创新领域。集中在人工智能、量子信息、集成电路、生命科学、生物育种、空天科技、深地深海、现代能源等前沿领域。对基础科研方向进行应用场景的细分，将相关可转化/转化中的科技成果进行技术链条的梳理，通过专业性的技术解构和解析，形成高逻

辑性、易理解性的技术图谱；并在此基础上，对科技成果产业化应用现状进行行业研究和分析，以全球视野定位领先梯队中的科创企业和学术团队的技术实力。通过行业图谱的研究，不仅可以清晰定位高新科技企业的技术竞争力，而且能够对我国相关行业现状和未来方向有更准确的认识。既为科技成果转化提供了专业性知识体系支撑，也有助于指导城镇产业化发展布局、推动产业链融通创新、引导创业投资基金对“硬科技”的积极性及鼓励金融支持创新体系的建设。

二、研究方法

方法学上，行业图谱研究将进行学科领域分级细化，再对技术在应用场景方向上进行详细分级和解构：

（一）一级分类：从应用产业所属学科的角度，以国家十四五规划为导向，重点关注影响国家安全和发展全局的基础核心领域，包括人工智能、量子信息、集成电路、生命科学、生物育种、空天科技、深地深海、现代能源等。

（二）二级分类：对技术对象进行分类。比如生命科学中包括疫苗、新生物材料、细胞治疗、人工智能、基因技术等技术对象，择一进行技术应用方向分析和流程解析。

1、应用方向的技术流程全景

即对某一技术对象在一个应用方向上的技术流程全景图，从研发到生产、上市的全流程。如新药的研发生产及上市的整体概况图。

2、应用方向的技术产品细分类

对技术对象在此应用方向上所形成的产品种类进行细分，并提炼属性/功能的特点。比如机器学习和深度学习算法在多肽合成、虚拟筛选、毒性预测、药物监测和释放、药效团建模、定量构效关系、药物重定位、多药理和生理活性等药物发现过程。

3、应用方向上某一细分产品的技术开发流程

从上一级分类产品中选定一个细分产品，一般是现阶段技术发展最先进的产品，针对其所应用的场景相关技术开发/生产全流程进行解析和描述。比如：新药研发中蛋白质折叠和蛋白质相互作用的人工智能预测技术，其生产技术流程及其中核心竞争技术环节。

4、领先级国际科创企业及学者团队定位

将国际国内最领先的科创企业进行技术平台和产品性能的比较分析，并将其所具备的技术优势定位于上述图谱中。将国内外学者团队的领先性研究成果/转化状态进行分析，并定位于上述图谱中。比如：国际已上市或进入临床的人工智能制药企业如Exscientia、Atomwise、Recursion、Insilico Medicine的优势技术平台。

三、研究报告形式

行业图谱以结构化脑图为基本形式，辅以文字报告进行解释说明。文字报告的内容框架包括：

概览：概述图谱传递的信息内容、解答的技术问题和目的。

科学背景简述：描述图谱行业背景、技术流程、关键技术平台和竞争点的细节、技术应用的例证及国内外行业发展现状，对图谱做详细内容的补充说明。

专业术语解析：针对重点专业术语进行概念解释。

参考文献。

免责声明

本报告由清华大学五道口金融学院国家金融研究院，资本市场与公司金融研究中心（以下统称“研究中心”）编写。本报告仅供研究使用，并非为提供咨询意见而编写。本报告中的信息均来源于本研究中心认为可靠的已公开资料，但研究中心及其关联机构对信息的准确性及完整性不做任何保证。