



中国人工智能系列白皮书

——人工智能与药物发现

中国人工智能学会

二〇二二年九月

《中国人工智能系列白皮书》编委会

主 任：戴琼海

执行主任：王国胤

副 主 任：陈 杰 刘成林 刘 宏 孙富春 王恩东 王文博
赵春江 周志华

委 员：班晓娟 曹 鹏 陈 纯 陈松灿 邓伟文 董振江
杜军平 付宜利 古天龙 桂卫华 何 清 胡国平
黄河燕 季向阳 贾英民 焦李成 李 斌 刘 民
刘庆峰 刘增良 鲁华祥 马华东 苗夺谦 潘 纲
朴松昊 钱 锋 乔俊飞 孙长银 孙茂松 陶建华
王卫宁 王熙照 王 轩 王蕴红 吾守尔·斯拉木
吴晓蓓 杨放春 于 剑 岳 东 张小川 张学工
张 毅 章 毅 周国栋 周鸿祎 周建设 周 杰
祝烈煌 庄越挺

《中国人工智能系列白皮书——人工智能与药物发现》编委会

主 任:

张学工

副 主 任:

高 琳 沈红斌 汪小我 汪增福 赵兴明

秘 书 长:

王 颖

常 务 委 员:

蔡宏民 杜朴风 高 琳 古 槿 蒋庆华 姜 伟

雷秀娟 李 敏 刘治平 沈红斌 宋晓峰 汪小我

王 颖 汪增福 魏彦杰 鱼 亮 张 法 张绍武

张世华 张学工 张治华 赵兴明 章 乐 章 文

邹 权

本书编写组

何 松 李 昊 刘 琦 刘世超 宋 弢 施建宇

涂仕奎 魏乐义 辛弘毅 曾湘祥 章 文

全书统稿: 章 文

前言

《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》中提出了“瞄准人工智能、生命健康等前沿领域”、“聚焦人工智能关键算法等关键领域，加快推进基础理论、基础算法、装备材料等研发突破与迭代应用”等规划。发展新一代人工智能是我国在科技革命与产业变革等战略问题上的重要抓手。药物是诊断、缓解、治疗或预防疾病的物质，对于改善人类健康和保证生活质量具有非常重要的作用。塔夫茨大学药物开发研究中心的研究表明，在过去30年，研发一种新药需要近26亿美元的资金投入与近14年的时间投入，这个数字依然在不断提升。在高昂研发成本的驱使下，制药公司目前正在寻找可以提高研发效率和转化率的高新技术。

计算机辅助药物设计自20世纪60年代被提出，以计算化学、计算机科学和生物学等学科为基础，对靶标蛋白质与配体药物的结合过程进行计算模拟、预测，评估药物分子结构与其生物活性、毒性和代谢等性质的相互关系，进行药物分子的发现与优化。高通量技术的发展和应用产生了丰富的药物、疾病、基因和蛋白质等数据，使得开展人工智能药物发现成为可能。近年来，以Google公司AlphaFold为代表的的人工智能系统在生命科学领域取得了重要突破，推动了人工智能等关键领域在药物研发上的应用。深度学习(Deep Learning, DL)、自然语言处理(Natural Language Processing, NLP)和知识图谱(Knowledge Graph, KG)等人工智能关键技术已广泛应用于药物发现的各个环节，如肿瘤靶点识别、苗头化合物筛选、药物从头设计、药物重定位、药物属性预测、药物相互作用预测、药物发现中的可解释性模型和大规模预训练模型等。人工智能辅助药物发现深刻改变了药物发现的方法和途径，极大提高了药物发现效率、缩短开发进程，加速了生物技术的创新变革，加深人类对生命科学中的分子机制的认

知。开展基于人工智能技术的药物发现研究，符合科技革命和国家发展规划的需求，是落实“面向世界科技前沿、面向国家重大需求、面向人民生命健康”战略的重要举措。

本白皮书收集了目前国内外人工智能与药物发现交叉领域的最新理论研究成果，并介绍了人工智能技术在药物发现领域中的应用。编写过程中的贡献者包括：辛弘毅（第一章）、施建宇（第二章）、涂仕奎（第三章）、曾湘祥（第四章）、宋弢（第四章）、魏乐义（第五章）、刘世超（第六章）、刘琦（第七章）、李昊（第八章）、何松（第八章）、章文（统稿与第六章）及其博士生李梦露、刘旋和王紫嫣（整理与校稿），在此一并表示感谢。

目 录

前言.....	1
第 1 章 人工智能与肿瘤靶点识别	1
1.1 人工智能与肿瘤靶点识别概述	1
1.2 人工智能与肿瘤建模	2
1.2.1 人工智能与肿瘤转录组模型	2
1.2.2 人工智能与单细胞表观肿瘤模型	5
1.2.3 人工智能与多模态肿瘤模型	12
1.3 人工智能与靶点识别	15
1.3.1 人工智能与基于单细胞 RNA 的靶点发现.....	15
1.3.2 人工智能与基于表观的靶点发现	17
1.3.3 人工智能与基于多组学测序技术的药物靶点发现	18
1.4 人工智能在肿瘤靶点识别中的发展前景	21
1.5 本章小节	22
第 2 章 人工智能与苗头化合物筛选	23
2.1 人工智能与苗头化合物筛选概述	23
2.2 基于深度学习的苗头化合物筛选	25
2.2.1 CPI 数据库	25
2.2.2 蛋白质和化合物典型特征表示	26
2.2.3 基于深度学习的 CPI 预测模型	27
2.3 深度学习在苗头化合物筛选中的发展前景	34

2.3.1 趋势与挑战	34
2.3.2 实际应用	35
2.4 本章小节	36
第 3 章 人工智能与药物从头设计	38
3.1 基于人工智能的药物从头设计概述	38
3.2 深度生成模型与小分子药物从头设计	39
3.2.1 小分子药物合理结构的生成模型	39
3.2.2 满足生化性质要求的小分子药物生成模型	40
3.2.3 基于靶点蛋白结构的小分子药物生成模型	43
3.3 深度生成模型与大分子药物从头设计	46
3.3.1 基于深度学习的核酸类药物设计	47
3.3.2 基于深度学习的蛋白和多肽设计	48
3.4 本章小节	50
第 4 章 人工智能与药物重定位	52
4.1 药物重定位概述	52
4.2 药物重定位数据库	52
4.3 表示学习	53
4.3.1 基于序列的表示	53
4.3.2 基于网络/图的表示学习	56
4.4 药物重定位的深度学习模型	57
4.4.1 以靶点为中心的模型	57

4.4.2	以疾病为中心的模型	60
4.4.3	模型评估	61
4.5	药物重定位的应用	62
4.6	本章小节	65
第 5 章	人工智能与药物属性预测	67
5.1	人工智能与药物属性预测概述	67
5.2	多肽药物属性预测	69
5.2.1	多肽属性预测方法	70
5.2.2	研究难点	73
5.3	药物属性预测最新研究进展	74
5.3.1	基于元学习的多肽药物生物活性预测	74
5.3.2	基于图神经网络的多肽毒性预测	75
5.4	本章小节	78
第 6 章	人工智能与药物相互作用预测	79
6.1	人工智能与药物相互作用预测概述	79
6.2	人工智能与药物互作用预测方法	80
6.2.1	基于文献数据的提取方法	80
6.2.2	基于药物关联数据的预测方法	83
6.3	人工智能在药物相互作用预测中的发展前景	89
6.3.1	构建标准数据集	89
6.3.2	药物事件预测	90

6.3.3 预测高阶药物相互作用	91
6.3.4 整合多源数据分析	92
6.4 本章小节	92
第 7 章 药物发现中的大规模预训练模型	93
7.1 分子表征	93
7.2 预训练.....	95
7.3 分子预训练	97
7.3.1 基于 Mask Language Model 的分子预训练	98
7.3.2 基于生成式模型的分子预训练	99
7.3.3 基于对比学习的分子预训练	100
7.3.4 基于几何特征的分子预训练	101
7.3.5 基于领域知识的分子预训练	102
7.4 分子预训练范例	103
7.4.1 确定预训练任务与模型结构	103
7.4.2 构建运算平台	104
7.4.3 设计微调策略	105
7.4.4 模型微调与评估	106
7.5 本章小节	107
第 8 章 药物发现中的可解释人工智能模型	108
8.1 药物发现中的可解释人工智能模型概述	108
8.2 可解释人工智能技术 (XAI)	109

8.2.1 可解释机器学习	109
8.2.2 图结构的可解释技术	110
8.2.3 建模后的可解释技术	112
8.2.4 知识嵌入的可解释技术	114
8.2.5 针对注意力机制能否提供可解释的辨析	115
8.3 可解释人工智能在药物设计中的应用	116
8.3.1 XAI 与定量构效关系 (QSAR)	116
8.3.2 XAI 与联合用药	118
8.3.3 XAI 与分子属性预测	119
8.3.4 XAI 与药靶互作	120
8.3.5 XAI 与药物不良反应预测	121
8.3.6 XAI 与新药设计	122
8.4 可解释人工智能在药物发现中的前景展望	122
8.5 本章小节	124
参考文献.....	125

第 1 章 人工智能与肿瘤靶点识别

1.1 人工智能与肿瘤靶点识别概述

肿瘤药物研发是人工智能（Artificial intelligence, AI）的重要应用场景。靶点识别是肿瘤药物研发的关键抓手。近年来，在肿瘤多组学大数据的驱动下，人工智能逐渐成为肿瘤靶点研究中必不可少的研究手段。早期的肿瘤靶点研究模式较为简单，以检测肿瘤高突变率基因为主。目前已经获批进入临床的肿瘤靶向药大部分就是靶向这些高突变率的基因编码的致癌蛋白^[1]。然而经过临床的长时间测试，人们发现，这样的靶向方案能覆盖的肿瘤患者群体过于有限，即使是能满足靶向治疗条件的患者，也很容易出现耐药甚至转移复发的情况^[2]。近年来，生物分子测量技术的不断突破，使得人们能够从不同分子层面建立全面的肿瘤异常模型，为肿瘤靶点研究创造了新的契机。肿瘤靶点的研究从传统的关注高突变基因的单一思路，逐渐发展为多层面、多角度的研究思路^[3]。随着技术的普及和成本的下降，无论是反映肿瘤病人个体间差异的批量组学数据，还是反映肿瘤细胞间差异的单细胞组学数据都在快速产生和累积。爆发式增长的肿瘤组学大数据，为人工智能在肿瘤研究上的应用提供了数据基础。同时，组学数据具有维度高、噪声大、数据类型多样等特点，分析难度较大，也确实需要量身定制的分析方法来进行去噪和模式抽提。

日益丰富的组学测量技术为发现新的肿瘤靶点提供了契机。组学通常指生物学中对各类研究对象（一般为生物分子）的集合所进行的系统性研究，如基因组学、蛋白质组学、转录组学、代谢组学等。传统的批量（bulk）组学技术是以个体为研究对象，将待测生物样本中所有细胞混合在一起进行分子测定，只能反应肿瘤个体间的差异。新兴的单细胞组学技术能对肿瘤样本中的每个细胞进行分子测量，全面刻画肿瘤细胞间及肿瘤免疫微环境的异质性，为破解肿瘤耐药性产生

机制、研发新的肿瘤靶点提供了强大工具^[4,5]。近年来，组学测量技术不断融入主流的临床肿瘤学，科学研究表明可改善临床结果的多种分子靶向药也逐渐获批进入临床，加速了肿瘤治疗范式的改变，例如：曲妥珠单抗或威罗非尼等靶向药已成为表达 HER2 靶点的乳腺癌患者和有 BRAF 靶点突变的黑色素瘤患者的临床治疗标准，以免疫细胞为靶向目标的免疫检查点抑制剂也获批可用于治疗微卫星不稳定性特点的肿瘤患者^[6]。

人工智能为基于组学数据的肿瘤靶点识别提供了强大的计算工具。人工智能，更具体地说是机器学习 (Machine learning, ML) 分支，可以处理大规模异构数据集，并识别出数据中的潜藏模式。而随着技术的普及和成本的下降，肿瘤样本的批量组学和单细胞组学数据都在快速产生和累积，为人工智能在肿瘤研究上的应用提供了重要的数据基础。此外，组学数据具有维度高、噪声大、数据类型多样等特点，分析难度较大，需要量身定制的分析方法来进行去噪和模式抽提。目前，决策树、支持向量机等众多人工智能模型均已广泛应用到了组学数据建模和肿瘤靶点识别中^[7,8]。

1.2 人工智能与肿瘤建模

1.2.1 人工智能与肿瘤转录组模型

1.2.1.1 肿瘤转录组异质性

癌症的一大普遍特点是转录失调^[9]。在细胞内部，调节网络由一组连接的途径组成，其中途径是细胞中发生的化学反应链，通路是基因的集合，这些基因相互作用可以实现特定的细胞功能，调节细胞的状态，它们共同构成了细胞调节网络。为了使细胞正常运作，通路基因的表达水平需要得到很好的控制。然而，正常细胞和癌细胞存在许多差异表达的基因，癌细胞中的异常表达可能通过抑制或刺激途径使途径失调，这可能会影响细胞的适应性（即增殖能力），这种转录组

上的差异即为癌症中的转录组异质性。转录组指的是细胞内所有转录产物的集合,包括信使 RNA、核糖体 RNA、转运 RNA 及非编码 RNA,细胞的转录组可以随外部环境条件转变^[10]。转录组异质性在癌细胞中会急剧增加,这来源于 DNA 拷贝数异常,细胞所处环境的刺激,基因之间的相互作用混乱等。从转录层面来看,癌症是一种细胞调节网络混乱的疾病,因而进行转录组上的研究可为我们提供癌细胞更全面更独特的信息。对于基因组相同的细胞,也可能因其所处环境不同而表现出不同的转录状态。基于转录组的研究统称为转录组学,能够研究统计单个细胞或特定类型的细胞、组织、器官或发育阶段的细胞群内生产的各类 RNA 分子的类型和数量。在肿瘤细胞中,基因突变及环境的改变都会导致转录组的异质性,从而使癌细胞获得不同的功能特点,包括增殖、DNA 修复、侵袭、血管生成、衰老和细胞凋亡等等,这些仅从基因组角度是无法观测到的,而单细胞 RNA 测序可绘制出细胞的转录图谱,从而清晰地展示细胞的转录特点。

随着单细胞 RNA 测序技术的发展,近几年单细胞核糖核酸测序(scRNA-seq)已在世界范围内得到广泛应用^[11]。单细胞测序技术可谓是科技发展史上的一大创举,可以精细区分不同细胞类型,使得在单细胞水平研究分子机制成为可能。2009 年, Tang 等人^[12]提出了首个 scRNA-seq 方法,开辟了单细胞水平 RNA 测序的新领域。随后又有多种改进的技术,如 Drop-seq、Seq-Well、DroNC-seq 和 SPLiT-seq 等,值得注意的是,基于 droplet-based 的技术(Drop-seq^[13]、InDrop^[14]和 Chromium^[15])通常可以提供更大的细胞通量,而且与全转录 scRNA-seq 相比,每个细胞的测序成本更低,因而被广泛应用于肿瘤单细胞研究。目前,商业化的单细胞测序技术以 10x Genomics 为主,下文的数据分析也将以此为基础。在此基础上,2017 年美国安德森癌症中心的研究人员在 Cell 上发表了“地形”单细胞测序技术(Topographic single cell sequencing, TSCS)^[16],该方法提供了细胞位

置的空间信息，能更准确地从空间上获得单个肿瘤细胞的具体特征，能够在早期癌症研究方面提供有力的支持。

1.2.1.2 人工智能与单细胞转录组数据分析

单细胞数据处理和人工智能算法结合极为紧密，目前已有多种算法可以从繁杂的 RNA 测序序列中提取出用于生物学分析的转录组信息。转录组学数据的预处理主要包括质控、批次矫正、插补、降维和特征提取等步骤，下面将简略介绍这些过程的作用及现有算法。

由于转录本覆盖的偏差、低捕获效率和低测序覆盖度等因素，scRNA-seq 数据的技术噪声水平比较高，破损、死亡或与多个细胞混合的细胞中会生成部分低质量的数据，这些低质量的细胞将阻碍下游的分析，并可能导致数据的误读，因此需要对测序数据进行质控（Quality control, QC）。目前质控方法主要根据基因的数量、唯一比对率、表达基因/转录的数量比对率和线粒体 RNA 的质量等。测序过程中的操作差异、平台差异、测序方法差异等会引入系统错误、技术混淆和生物变异，导致一个批次的基因表达谱与另一个批次的基因表达谱存在系统差异，这种差异有可能会掩盖真实的生物学差异，导致分析结果错误^[17]。因而需要对测序数据进行批次矫正。现广泛使用的去批次矫正有 Harmony, LIGER 和 Seurat 3。2020 年 ASTAR 团队对 15 种批次矫正方法从多批次、多技术、模拟数据情况下识别细胞类型等多个角度进行了比对分析，得出 Harmony 是综合运行速度和结果准确性的最优批次整合方法。

单细胞 RNA-seq 数据通常包含许多由于原始 RNA 扩增失败而导致的缺失（dropouts），最近针对这些缺失开发了一些新的插补算法，比如 SAVER^[18]、MAGIC^[19]、ScImpute^[20]、DrImpute^[21]和 AutoImpute^[22]等。其中 SAVER 利用基于 UMI 的 scRNA-seq 数据来恢复所有基因的真实表达水平；MAGIC 通过构建基于马尔可夫亲和度的基因表达

图来进行基因表达的计算；ScImpute 可以利用其他类似细胞中不太可能受 dropout 影响的相同基因的信息，在不引入新的偏差的情况下计算 dropout 值；DrImpute 则基于集群将 dropout 中的零从真正的零中分离出来；AutoImpute 基于自编码通过学习 scRNA-seq 数据的固有分布来寻找缺失的值。

由于单细胞 RNA 数据是超高维的，数据降维可降低实验误差与数据噪声的影响，并挖掘数据内部的本质结构特征，便于后续计算以及数据可视化。主流的降维和特征提取算法可以分为基于矩阵分解的、基于图的和基于神经网络的降维算法三大类，其中主流的为主成分分析、t-随机邻域嵌入、均匀流形逼近和投影。主成分分析（Principal components analysis, PCA）是最常用的线性降维方法。t-随机邻域嵌入（t-distributed stochastic neighbor embedding, t-SNE）是一种非线性降维方法，能够根据在邻域图上随机游走的概率分布在数据中找到其结构关系。均匀流形逼近和投影（Uniform Manifold Approximation and Projection, UMAP）是基于 k-近邻理论使用随机梯度下降优化结果。

1.2.2 人工智能与单细胞表观肿瘤模型

1.2.2.1 肿瘤中的表观遗传模型

染色质结构定义了 DNA 形式的遗传信息在细胞内的组织状态，其中基因组这种精确紧凑结构的组织极大地影响了基因被激活或沉默的能力。表观遗传学最初被 C.H.Waddington^[23]定义为“基因及其产物之间的因果相互作用，从而导致表型的形成”，涉及到理解染色质结构及其对基因功能的影响。Waddington 的定义最初是指表观遗传学在胚胎发育中的作用。然而，表观遗传学的定义随着时间的推移已经演变，因为它涉及到各种各样的生物过程。目前对表观遗传学的定义是“研究独立于原始 DNA 序列变化而发生的基因表达的可遗传变化”。这些可遗传的变化大部分是在分化过程中建立的，并在细胞分

裂的多个周期中稳定地维持，使细胞在包含相同遗传信息的同时具有不同的身份。这种基因表达模式的遗传性是由表观遗传修饰决定的，包括 DNA 中胞嘧啶碱基的甲基化、组蛋白的翻译后修饰以及核小体沿 DNA 的定位。这些修饰的补充，统称为表观基因组。如果不能正确维护可遗传的表观遗传标记，可能会导致各种信号通路的不适当激活或抑制，并导致癌症等疾病状态^[24]。

表观遗传学领域表明除了大量的基因改变外，人类癌细胞还存在全局表观遗传学异常^[25]，这些基因遗传和表观遗传的改变在癌症发展的各个阶段相互作用^[26]。目前癌症的基因起源已被广泛接受，表观遗传改变可能是某些形式癌症的关键初始事件^[27]。这些发现促使研究人员开始探索表观遗传学在癌症的起始和传播中的作用^[28]。与基因突变不同的是，表观遗传畸变具有潜在的可逆性，因此可以通过找到表观遗传层面肿瘤的靶点信息，治疗患者，使癌细胞恢复为正常状态。表观遗传畸变具有的可逆性使此类举措具有广阔的前景^[29]。

第二代测序技术的进步，推动了表观遗传学的发展。例如，ChIP-Seq^[30]方法可以绘制组蛋白修饰和转录因子结合的全基因组图谱；利用 ATAC-Seq^[31]可以测定染色质可及性；使用 Hi-C^[32]等方法确定高阶染色质结构；DNA 甲基化则可使用 RRBS^[33]、WGBS^[34]或基于阵列的技术来确定。此外，表观基因组也被用于癌症诊断与辅助治疗。为了实现这些目标，我们需要实现自动化决策系统，应用于癌症的临床预防、诊断和治疗中^[35]。当前生物表观遗传辅助诊断仍面临许多挑战，尤其是临床环境数据分布广泛、模态多且高度复杂，使得单独调查单个实验-对照数据的传统方法效果有限。机器学习技术能够集成大型和复杂的数据集，推动临床诊断的发展^[36]，并帮助医生进一步解析临床表观遗传数据^[35]（图 1-1）。

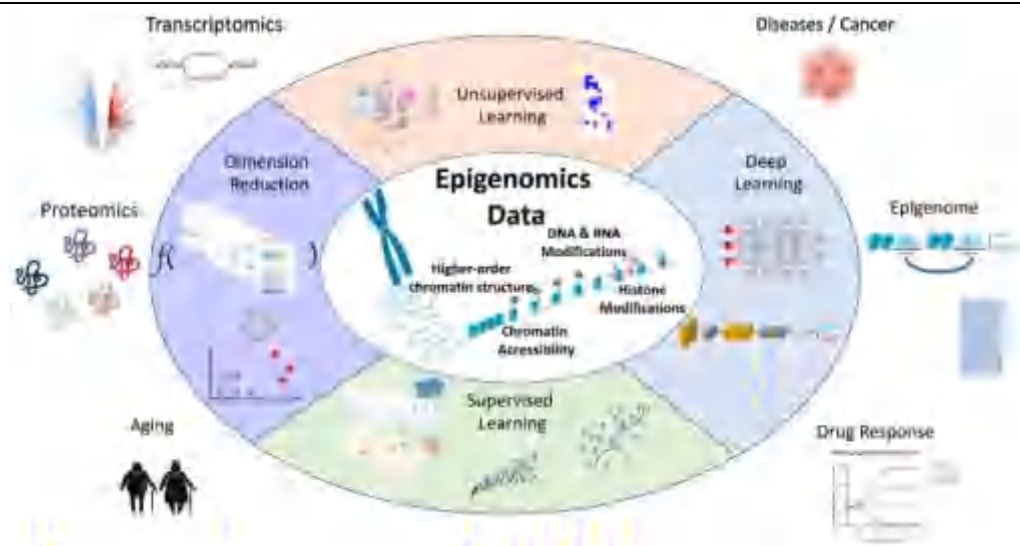


图 1-1 机器学习在表观遗传的应用^[35]

1.2.2.2 人工智能与基于甲基化测序技术的肿瘤表观遗传模型

基因甲基化是表观遗传的主要形式之一，同癌症的发生与发展有重要联系^[37]。DNA 甲基化代表基因组的直接修饰，并调控基因表达。目前几乎所有肿瘤都已发现特异性的基因甲基化标记物。相比正常细胞，癌细胞的基因甲基化水平有显著下降，是癌细胞基因调控失稳的重要原因之一。在正常组织中，细胞通过基因甲基化实现表达调控，当相关基因调控区的 CpG 岛大量发生甲基化时，便可阻止该基因的表达，实现基因沉默^[38]。DNA 甲基化导致基因沉默的已知机制大致可分为三种：1) DNA 甲基化干扰转录因子对 DNA 元件的识别与结合^[39]；2) 序列特异性的甲基化 DNA 结合蛋白与启动子区甲基化 CpG 岛结合，募集组蛋白去乙酰化酶 (HDAC)，形成转录抑制复合物，进而阻止转录因子与启动子区靶序列的结合，最终阻止基因转录表达^[40]；3) DNA 甲基化通过改变染色质结构，使染色质结构更加紧密，影响转录因子与 DNA 结合，进而使转录失活^[41]。由此可见，癌细胞通过降低自身的甲基化水平，可大量转录本应处于静默态的基因，降低细胞的表达调控稳定性，进而实现快速增殖、耐药重编程等特性^[42]。

甲基化与癌症发生的因果关系存在两类主要学说：1) 细胞在癌

变后启动了重编程通路，随后对 DNA 进行了甲基化改写，即癌变导致甲基化重编程；2) 癌症由正常细胞的甲基化紊乱发展而来，即甲基化错误导致癌症。两大学说均有若干证明。众多研究发现低甲基化在多种恶性肿瘤乳腺癌、子宫颈癌、脑癌中可见，并且在免疫缺陷的许多癌症患者中，染色体 1 和 16 上的中心周围染色质区域严重低甲基化，这些都证实了甲基化紊乱与癌症的关联^[24]。

基因甲基化在临床中被大量用于癌症早筛和诊断。例如，mSEPT9 基因甲基化是结直肠癌的重要生物标记物，其编码的 SEPT9 蛋白，在细胞代谢中发挥重要作用，并被 FDA 批准用于结直肠癌诊断中。该蛋白可阻止细胞过快分裂或以不受控制的方式增殖，从而达到抑癌基因的效果。当 SEPT9 启动子区域甲基化时，SEPT9 蛋白停止表达，最终导致上皮细胞癌变，最终发展为结肠癌^[43,44]。再如，RASSF1A 基因甲基化是肺癌的关键生物标志物。RASSF1A 基因参与细胞周期调节、诱导细胞凋亡和稳定微管等多种细胞生理功能。RASSF1A 基因甲基化会导致 RASSF1A 基因表达静默，进而干扰细胞在出现基因损伤后经由细胞周期检查点机能诱导细胞凋亡，进而促进了癌细胞的存活和生长^[45]。

基因甲基化靶点的发现主要依靠人工智能算法，其关键在于从大量高噪声数据中识别出同癌症发生相关度高的基因甲基化信号，主要包括以下难点：1) 基因甲基化数据多为组织级测序数据。该数据混合了各癌症克隆亚型以及各类正常细胞型的甲基化信号，数据采样率低、随机性大。2) 基因甲基化导致基因模式较多、差异度大，CpG 岛在 DNA 中分布广泛。单一基因的表达静默或开启存在多种不同的甲基化模式，细胞癌变往往是多基因共同作用的结果。3) 维度高、数据量相对有限。

机器学习在表观遗传领域的研究多集中于分类问题。问题核心是如何建立一个模型，能够准确预测出给定样本的类别信息，例如从阵

列数据中区分正常与癌症样本。主要的方法包括支持向量机 (Support vector machine, SVM)、决策树 (Decision tree, DT)、随机森林 (Random forest, RF) 和朴素贝叶斯 (Naive bayes, NB) 等。SVM 依赖于对数据进行高维拓扑, 并在拓扑空间中找到分类超平面实现分类。Wayne Xu 等研究者在识别黑色素瘤和软组织肉瘤的问题上, 正确地分类了 76 个样本中的 75 个^[46]。决策树对数据进行逐层分类分割, 逐步细化分类结果, 可适应癌症 DNA 甲基化数据的高度异质性。Atsushi Kaneda 等研究者使用 DT 在结直肠癌样本的测试集上达到了 95% 的准确率, 并识别出三种结直肠癌亚型表观层面的生物标志物^[47]。NB 是另一种广泛使用的监督学习方法, 可以整合数据中存在的不确定性, 并且易于解释。其理论核心为贝叶斯定理中的条件概率模型^[48]。深度学习也广泛应用在 DNA 甲基化数据上。如电子科技大学 Shicai Fan 团队提出 MRCNN 使用卷积神经网络 (Convolutional neural networks, CNN) 根据附近的 DNA 序列预测全基因组甲基化水平^[49], 该方法以 93.2% 的准确度预测甲基化与非甲基化区域。哈尔滨工业大学王亚东团队在 2019 年使用变分自动编码器 (Variational auto-encoders, VAE) 和 t-SNE 来压缩 450K 甲基化数据以进行逻辑回归分类^[50], 体现了 VAE 编码对解释复杂的高维非线性数据的价值^[35]。

针对数据维度过高的挑战, 已有工作主要从特征排序、特征选择和特征融合三个方向着手降低维度。特征排序通过假设检验检测特征同标签关联, 并对特征的重要度进行排名。例如 T 检验计算 P 值来衡量零假设^[51], 即潜在假设是患者样本和对照组样本都符合正态分布。Wilcoxon 检验 (Wtest) 评估两个分布之间的差异, 其作为 T 检验的替代^[52]。卡方检验 (chi-squared test, Chi2) 则评估两个互斥类中的一个特征是否具有统计学显著性差异^[53]。Li Zhou 等人使用假设检验的数据挖掘方法, 评估 hsa-mir-3923 (MicroRNA 的一种) 表达与临床相关及病理调控变量的关系, 发现胃癌中 66 个基因与 hsa-mir-3923

可能存在密切关系^[53]。特征选择主要依靠在机器学习模型中融入特征筛选压力。其中递归特征消除（Recursive feature elimination, RFE）是一种常用的具有特征系数的分类模型特征选择框架。特征将递归评估其模型系数，系数最小的特征将被移除。例如 Alhasan Alkuhlani 等研究者使用 SVM-RFE 算法，分别为乳腺癌、结肠癌和肺癌数据集选择了 24、13 和 27 个最佳 CpG 位点的子集，这些最佳 CpG 位点子集的分类准确率分别为 100%、100%和 97.67%。Stefan M. Pfister 等研究者基于随机森林模型建立了 100 种已知的中枢神经系统肿瘤诊断系统，该方法可能对诊断精度有实质性的影响^[54]。特征融合通过特征之间的关系将多特征融合为单特征，以降低特征数目。吉林大学周丰丰团队提出 ReGear，使用线性回归将原始的甲基化位点特征拟合成基因特征，以大幅降低特征维度，在乳腺癌和胃癌的病例中获得了更好的分类预测准确率^[55]。

1.2.2.3 人工智能与基于染色质可及性的肿瘤表观遗传模型

随着单细胞染色质可及性测序技术（ATAC-seq）的出现，染色质可及性已成为癌症研究的重要问题之一。单细胞染色质可及性测序技术的本质是在单细胞水平检测基因所在染色体是否处于开放状态：处于染色质开放状态的基因可被转录并表达。染色质处于闭合状态的基因则被静默。染色质可及性是表观遗传调控的一种表征，通常与 DNA 甲基化相关。相比基因甲基化测序，染色质开合提供了更为直接和确定的观测——处于打开状态的基因处于非静默状态，而处于闭合状态的基因则一定处于静默状态。目前发现染色质开合同癌症的发生、发展、产生耐药性有相关性，可用于癌症诊断和预后预测等临床问题中。

单细胞染色质可及性测序技术的核心原理与单细胞转录组测序技术较为相似，均基于微液滴微流控测序技术。然而，相比 scRNA-seq 数据，scATAC-seq 数据分析更具挑战性。其核心难点有三：1)

scATAC-seq 数据高度稀疏。当前 scATAC-seq 测序技术仅能覆盖>1% 的基因组。此外，测序中存在的高度随机性，导致单个细胞中，scATAC-seq 数据信号极为微弱、信噪比低，仅能测量极少数染色质的打开基因。2) 缺乏领域知识与标准。目前对于各类型细胞的染色质可及性研究较为有限，领域缺乏对各类细胞染色质可及性的全面、深入的认知，相关数据库并不完善。3) 数据维度极高。由于 scRNA-seq 测序技术侧重于度量蛋白编码基因的 RNA 表达量，数据维度同基因数量相同——约为 3 万。相比之下，scATAC-seq 技术不仅仅局限于测量蛋白编码基因区间的染色质开合状态，其测量范围被扩大到全基因组。

目前已经开发了各种分析工具来使用 scATAC-seq 数据研究单细胞表观基因组，可以分为四大类。第一类是无监督学习算法，包括聚类和降维。chromVAR 利用开放染色质区域中出现的转录因子 (Transcription factor, TF) 基序，使用流型学习中的 t-SNE 算法将单个细胞的偏差校正向量投影到二维上。该算法的优势在于它可用于计算与染色质可及性显著相关的 TF 结合谱，能够精确地聚类 scATAC-seq 剖面，并表征与染色质可及性变异相关的已知和新的序列基序^[56]；另一种 scABC 算法则仅依赖于基因组区域内的读取计数模式，通过使用无监督的 k-medoids 聚类来聚类细胞，并证明了细胞类型特异性开放启动子可以更好地识别细胞类型特异性表达^[57]；SCRAT 则是一种较为成熟的方便用户使用的软件，用于根据不同特征（例如基因集、转录因子结合基序位点等）方便地总结调控活动。利用这些特征，用户可以识别异质生物样本中的细胞亚群，推断每个亚群的细胞身份，并发现显示亚群之间不同活动的基因集和转录因子等显著特征^[58]。第二类是将染色质可及性当作序列模型，用自然语言处理的方法进行分析。例如 Cusanovich 等人对构建小鼠器官单细胞图谱的大规模研究，使用潜在语义分析 (Latent semantic analysis, LSA) 来识别细胞簇，确

定了数百种具有复杂性状的细胞类型，这些数据定义了单细胞角度下常见哺乳动物细胞关于调控基因组的体内景观^[59]。Carmen Bravo González-Blas 提出了概率框架 cisTopic，对增强子和稳定细胞状态联合建模，用于分析造血细胞、大脑和转录因子扰动的单细胞 ATAC-seq 数据集。实验表明该算法模型可以有效识别细胞类型、识别增强子、相关转录因子，挖掘细胞异质性信息^[60]。第三类是使用图或网络的模型。例如 Cicero 等提出了基于图 Lasso 的预测 DNA 的顺式调控元件方法，通过使用相似细胞组的采样和聚合来量化假定的调控元件之间的相关性，发现这些预测的相互作用与其他染色质 3D 结构数据兼容^[61]。第四类为综合分析工具，如 Scasat^[62]和 SnapATAC^[63]等。

1.2.3 人工智能与多模态肿瘤模型

1.2.3.1 基于多组学的肿瘤研究模式

多组学 (multi-omics) 分析，是指同时在多种不同的生物分子层面对研究对象进行的定量分析，通过将基因组、表观组、转录组、蛋白组和代谢组等不同模态的分子数据进行整合，来揭示潜藏在数据后的生物规律。目前，肿瘤多组学数据日益丰富和强大。传统的肿瘤多组学分析以个体为研究对象，已有一众大型国际合作项目和众多小型研究积累了大量宝贵的临床样本，The Cancer Genome Atlas (TCGA, <https://cancergenome.nih.gov>)、International Cancer Genome Consortium (ICGC, <https://icgc.org>) 等国际项目所收录的肿瘤样本均超万例，每个样本均测量了基因组、转录组等多种组学特征，为描述肿瘤个体间差异提供了群体信息和宝贵的临床信息。新兴的单细胞多模态技术将肿瘤多组学分析的分辨率从个体水平提升至单个细胞水平，为深入解析肿瘤内部细胞在不同分子模态上的差异提供了有力工具。图 1-2 简要总结了目前已实现的单细胞多组学技术^[64]，单细胞多组学技术为描述肿瘤内部异质性提供了强大的工具，被 Nature Method 评为 2019 年

年度方法。



图 1-2 已有的单细胞多模态技术^[64]

1.2.3.2 人工智能与多组学数据建模

癌症多组学模型从信息整合方式上来讲可分为早期整合、中期整合和后期整合三类（图 1-3）^[7,8]。早期整合采用简单的矩阵拼接的方式将不同组学的数据拼接为一个矩阵，该方法虽简单直接，但会面临“维度诅咒”问题。后期整合是指在单一组学的建模、分析完成后，对每个组学层面得到的分析结果进行整合，该方法虽较为稳定，但一般需要大量的人工干预对每种组学层面的结果进行解读，并手工融合不同组学得到的结果。此外，以上两种方法均未充分考虑多组学数据的内在异质性及不同组学之间的潜在联系。人工智能为多组学数据建模提供了第三种方案，即通过建立可兼容不同数据特性的机器学习模型，对癌症多组学数据进行系统性建模，在兼容不同数据类型的同时，实现对不同分子层面间的关联的模拟。

目前，人工智能已成为多组学数据建模的重要手段，已广泛应用于肿瘤分子分型、药物响应预测、靶点发现、生存期预测等诸多方面。相比于以个体为研究对象的 **bulk** 多组学技术，单细胞多组学技术通

量高，单次即可产生成千上万个样本（细胞）的测量结果，更加适合人工智能这种需要大样本才能充分发挥效力的模型，同时单细胞技术噪声大、信噪比低，更需要人工智能模型进行去噪。

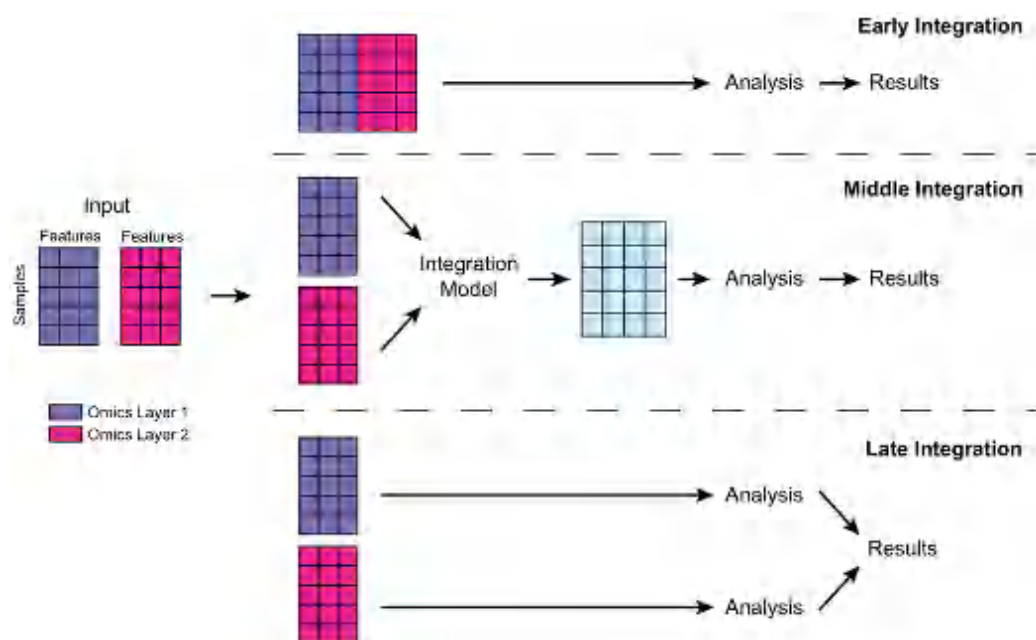


图 1-3 多组学数据整合的三种主要思路^[8]

人工智能在单细胞多组学领域的应用可分为两种场景。第一种场景是不同组学的测量对象（细胞）不匹配。由于技术复杂度和测量成本等因素，会发生不同批次的细胞分别测了一种单细胞组学的场景。通常，研究人员会假设每种组学测的细胞来自同一个分布，即不同组学测的细胞不同，但每种组学测的细胞群体在细胞构成、细胞状态上不应该有显著差异。目前，最近邻^[65]、典型相关分析^[66]、非负矩阵分解^[67,68]、流形对比^[69,70]、统计模型^[71]、变分自编码器^[72]等多种机器学习方法均已应用到不匹配场景下的单细胞多组学分析中。第二种场景是不同组学的测量对象是同一批细胞，即真正意义上的单细胞多模态。这类场景中常见的人工智能模型可分为三类^[73]。第一类是基于矩阵分解的方法，将每种组学数据描述为一个特征矩阵和一个系数矩阵的乘

积，并假设细胞在不同组学层面共享相同的系数矩阵，这类方法中代表性的工作有 MOFA+^[66]等；第二类是基于神经网络的方法，利用神经网络的高度灵活性和强大的特征提取功能，让神经网络学习到可以同时编码多种组学特征的低维向量，从而实现对单细胞多组学数据的整合，这类方法中代表性的工作有 scMVAE^[74]、totalVI^[75]等；第三类是基于网络表示的方法，先用单个组学特征构建细胞的相似性网络，然后基于网络融合的方式得到单细胞多组学网络表示，这类方法中代表性工作有 Seurat v4^[76]。

1.3 人工智能与靶点识别

1.3.1 人工智能与基于单细胞 RNA 的靶点发现

随着近年来关于 RNA 失调的深入研究，以 RNA 失调作为肿瘤抗原 (Tumor antigen, TA) 的来源，寻找新的免疫治疗靶点成为肿瘤研究的一大热点^[77]，对不同 RNA 亚型及参与 RNA 加工的蛋白质促成癌症的机制发现，为治疗干预提供了新的机遇^[78]。如 circRNA 在癌症中过度表达也展示了其作为疾病生物标志物的潜力^[79]。此外，2017 年 Balzeau J. 等人发现 let-7 miRNA 变体通过靶向癌基因 (包括 KARS 和 MYC) 抑制肿瘤的发展^[80]，因此癌症中最常见的 miRNA 的减少，即 let-7 miRNA 变体被认为是一个潜在的治疗靶点。化学修饰寡核苷酸整合或重新引入 miRNA 可能成为未来一种新的治疗方式，但在 miRNA 传递方面仍存在挑战。目前已开始了部分 miRNA 模拟物和 miRNA 抑制剂治疗肿瘤的药物试验^[81]。此外，miRNA 治疗间皮瘤的 I 期试验已获得初步成功^[82]，RNA 靶向药物的临床应用指日可待。

最近有研究表明，选择性剪接的异常转录本有可能作为免疫检查点抑制物 (Immune checkpoint inhibitors, ICI) 治疗的新分子标志物。选择性剪接广泛存在于癌症转录组中，有助于形成“癌症标志”，这是区别癌细胞与正常细胞的关键表型特征^[83, 84]。对于癌症相关的选择性

剪接可以通过多种机制调节癌症的进展，例如通过产生促进细胞增殖、抑制细胞死亡、避免抗肿瘤免疫或促进侵袭和转移的蛋白质异构体^[84]。ICI 治疗中，以非同义突变所产生的癌症特异性新抗原为靶点，然而在一些癌症中该方法并不奏效，近年的研究表明，含有移码突变和异常剪接模式的转录本也会产生抗原肽^[85-88]，异常转录物作为 ICI 的生物标志物可能具有重要潜力。如 2021 年 Yutaka Suzuki 团队利用第三代转录组测序检测到可作为非小细胞肺癌潜在新抗原转录本的异常剪接异构体，这为非小细胞肺癌的治疗提供了新的靶点^[88]。

转录组可用于肿瘤及免疫细胞的亚群分析，为肿瘤的发展、免疫逃逸和耐药性等研究提供更全面的信息^[89, 90]。2017 年张泽民团队通过对肝癌患者的外周血、肿瘤和邻近正常组织中的 T 细胞的转录组学分析，鉴定了 11 个 T 细胞亚群，描绘了其发育轨迹及每个亚群的特征基因，证明了肿瘤浸润淋巴细胞在免疫疗法开发和预测中的关键作用^[91]。次年，该团队对于非小细胞肺癌的研究表明，肿瘤浸润淋巴细胞的组成，状态及异质性与肺癌预后高度相关，转录组学有潜力用于癌症预后预测^[92]。转录组层面的亚群分析可提供亚群独特的蛋白等标志和可针对特定肿瘤亚群、免疫抑制性细胞亚群制定靶点药物。

从转录组出发的肿瘤细胞之间及肿瘤细胞与基质的相互作用研究也有望为肿瘤治疗提供有潜力的靶点。当前细胞间通信的识别方法有两种：(1) 依赖于一种细胞类型中受体基因与另一种细胞类型中相应配体基因的表达水平的比较^[9]。CellPhoneDB 方法首先计算一种类型中受体基因的平均表达和另一种细胞类型中配体基因的平均表达^[93]，然后通过基于图形的方法生成零分布，以评估统计显著性^[94]，并在随机排列所有细胞的类型标签后重新计算均值，最后观察到的均值与零分布进行比较来评估其统计显著性。(2) 通过计算一种细胞类型的受体基因表达与另一种细胞类型中相应配体基因表达在所有 scRNA-seq 数据集中的相关性来识别特定的通信。2019 年，Browaeys

R.等通过将基因表达数据与细胞内信号传导和基因调控网络的先验知识相结合，开发了 NicheNet 算法^[95]。NicheNet 通过将基因的表达数据与配体-靶点链接的先验知识模型相结合，推断相互作用细胞之间的活性配体-靶点链接。

1.3.2 人工智能与基于表观的靶点发现

表观遗传信号是最早发现的癌症治疗靶点之一，相关药物开发可以追溯到 1970 年代的分化剂（作用于 DNA 甲基化）^[96]。表观调控异常是肿瘤细胞维持恶性和侵入性的原因之一，是重要的潜在药物靶点。从表观调控异常信号中搜索癌症的生物标记物具有广阔的临床应用前景，可用于开发能逆转肿瘤表观遗传异常的药物，抑制癌细胞增殖或延缓癌细胞恶性发展进程^[97]。

基于表观组学分析的人工智能方法可应用于癌症亚型分类，协助指定个性化临床治疗方案。目前甲基化分析已被用于预测胆管癌^[98]、非典型畸胎样/横纹肌样肿瘤的生存、复发风险或治疗结果^[99]、脑肿瘤^[100]或肺肿瘤^[101]。基于蛋白修饰微阵列数据分析可预测癌症复发，例如前列腺癌和膀胱癌^[102, 103]。Jurmeister P 等人对原发性肿瘤进行了 DNA 甲基化分析，开发了基于神经网络的分类模型，在 279 名 HNSC 和 LUSC 患者以及正常肺对照的验证队列中正确分类了 96.4% 的病例，为后续选择临床治疗方案提供支持^[104]。基于 DNA 甲基化的癌症分类器也被用于确定未知原发性癌症的原发部位，以辅助治疗决策和改善预后。Moran 等人描述了一种基于微阵列 DNA 甲基化特征的临床诊断方案，从大约 3000 个肿瘤样本中训练机器学习模型，并在测试集中取得近 100% 的肿瘤分类准确度^[105]。Rong Xu 等人使用来自癌症基因组图谱 TCGA 的 18 种不同癌症起源的 7,339 名患者的 DNA 甲基化数据开发了基于深度神经网络（Deep neural network, DNN）的癌症起源分类器。与现有的基于病理学和基于基因表达微阵列数据的

模型相比更准确，并具有在临床环境中易于实施的独特优势^[106]。

人工智能方法在表观组学分析中的另一用途是肿瘤-健康细胞差异分析和靶点识别^[107]。近年，以 Vorinostat 为代表的表观遗传靶向药物已陆续获 FDA 批准进入市场^[108]。同时，新靶标不断被发现，如 DNMT1^[109]、PRMT^[110]等。Chip-seq 等染色质免疫沉淀测序技术提供了分析蛋白质与 DNA 交互作用的新手段，从染色质-蛋白结合的角度解析细胞的表观遗传机制与作用^[111]，由此产生的各类大量表观组学数据为人工智能方法开发和靶点发现提供了大量机遇^[112]。例如，清华大学谭春燕团队使用 SVM 预测了基于由 VEGFR-2、Abl-1 和 ERK-2 介导的经过充分研究的抗癌信号网络，并进一步开发出了同时靶向这三种蛋白质的烟酰胺类化合物 NEPT^[113]。

1.3.3 人工智能与基于多组学测序技术的药物靶点发现

肿瘤的发生发展是自身异常基因突变积累等内因与免疫系统失稳等外因共同作用的结果。肿瘤靶点识别也有两种主要思路，一种是针对肿瘤细胞的异常特征识别能够直接作用于肿瘤细胞的靶点，另一种是根据肿瘤免疫微环境的特点，识别作用于免疫细胞的靶点，从而提升其对肿瘤的抑制效果。

1.3.3.1 人工智能结合多组学发现肿瘤靶点

在肿瘤靶点识别上，人工智能结合多组学数据已在乳腺癌、肝癌、卵巢癌和胰腺癌等多种癌症中发现了有临床价值的靶点。复旦大学邵志敏团队利用网络融合方法 SNF (Similarity network fusion)^[114]对数百例中国三阴性乳腺癌 (Triple-negative breast cancer, TNBC) 临床样本的基因突变、拷贝数变异和转录组进行建模，发现东亚 TNBC 患者的分子特征显著异于白种人患者，并提出东亚 TNBC 患者可分为四种亚型，且通过肿瘤多组学模型识别出了部分亚型的潜在靶点^[115]。

苏黎世理工的 Christos D.等研究者对 mTOR 驱动的肝细胞癌小鼠进行了转录组、蛋白组、微小 RNA 等多组学测量，并利用基于先验网络的网络融合方法 NetICS^[116]进行了分析，检测出 74 个候选靶点，并在体外实验中发现靶向其中两个基因 YAP1 和 GRB2，对抑制 mTOR 信号通路激活型肝细胞癌有一定效果^[117]。广东省中医院的梁雪芳团队通过构建自编码器模型实现对卵巢癌的 mRNA、微小 RNA、CNV 三种模态数据进行联合建模和特征抽提，然后基于瓶颈层输出得到的三种模态的共同表征进行无监督聚类，识别出两种亚型，通过进一步的差异基因分析和共表达分析，识别出 34 个与卵巢癌有关的靶点基因^[118]。凯特琳癌症研究中心的 Ronglai Shen 等人提出联合隐变量模型 iCluster 来整合不同的组学特征并进行聚类^[119]。后有研究者将 iCluster 应用到肝细胞癌的研究中对 SNV、CNV、甲基化、转录组等模态进行整合，将肝细胞癌分成了两个子类型，并识别出 17 个可能的靶基因^[120]。北京大学汤富酬团队基于相关性分析及 k 近邻等模型，对胰腺导管腺癌样本的单细胞多模态（甲基化、染色质开放程度、转录组）数据进行了建模和分析，成功在单细胞水平识别出新的表观靶点^[121]。斯坦福大学 Jeffrey Granja 等研究者在对混合表型急性白血病（Mixed phenotype acute leukemia, MPAL）的研究中，将自然语言处理中的潜在语义分析模型（Latent semantic indexing, LSI）应用到单细胞多模态（表面蛋白、转录组、染色质开放）数据上实现特征抽提和不同模态的共同投影，识别出不同表型的 MPAL 共有的恶性特征和调控白血病相关基因的转录因子^[122]。

1.3.3.2 人工智能结合多组学发现免疫靶点

2018 年获得诺贝尔医学奖的免疫疗法是利用患者自身免疫系统来预防、控制和消除癌症的治疗方法^[123]。目前在临床上已有数种方案可选，如免疫检查点抑制剂（Immune checkpoint blockade, ICB）^[124]、

过继性细胞疗法 (adoptive cellular therapy) [125] 等, 但目前均面临不具备普适性、整体响应率较低的问题。免疫系统本身具备识别并消灭肿瘤细胞的功能, 然而在肿瘤不断演化的过程中, 肿瘤细胞能够逐渐习得躲避免疫系统杀伤的能力, 从而产生“免疫逃逸”, 甚至利用免疫细胞的功能机制, 加速自身的增殖甚至转移。

充分解构肿瘤免疫微环境, 是发现免疫靶点的重要基础和前提。研究表明, 肿瘤微环境会将 T 细胞长期暴露于抗原的持续慢性的刺激之下, 这使得 T 细胞逐渐丧失效应功能, 从而无法识别、消除肿瘤细胞。T 细胞耗竭导致的细胞功能和状态改变, 在表观组、转录谱和代谢等方面均有异常体现 [126]。此外研究还发现, 肿瘤会通过招募免疫抑制细胞 Treg 来达到抑制杀伤性 T 细胞的免疫效能, 从而使肿瘤免于被攻击 [127], 肿瘤还可通过重编程自身的代谢模式, 将整个肿瘤免疫微环境变成营养贫乏、乳酸富集、缺氧的状态, 使其非常不利于效应 T 细胞的生存和功能发挥 [128]。目前, 人们对肿瘤免疫微环境的认知还非常有限, 充分了解肿瘤免疫微环境的细胞构成, 认识每种细胞型在肿瘤发生发展过程中所发挥的作用, 有望能提升已有免疫疗法的响应率, 为开发新的免疫疗法提供可能。

基于多组学的人工智能模型在解构肿瘤免疫微环境中具有巨大潜力。哥伦比亚大学的 Benjamin Izar 及美国博德研究所 (Broad Institute) 的 Aviv Regev 等人结合单细胞蛋白-转录组双模态技术与 CRISPR 技术研究了黑色素瘤对免疫检查点抑制剂产生耐药性的机制, 研究者采用 Elastic-net 进行特征选择, 并构建正则化线性模型 MIMOSCA [129], 分析 CRISPR 带来的扰动对基因表达量的影响, 最终在复现已知耐药性相关因素之余, 也发现了 CD58 表达缺失这一新的免疫检查点抑制剂耐药机制 [130]。北京大学张泽民团队基于密度聚类及非线性可视化方法 tSNE 对非小细胞肺癌免疫微环境中的 T 细胞群体进行了单细胞多模态 (TCR 库、转录组、蛋白) 数据分析, 发现除肿瘤相关 CD8 T

细胞呈现耗竭状态之外，还有两种子类细胞呈现出耗竭前的状态。此外，还发现一种与恶性预后高度相关的激活态 Treg 亚型，并识别出其基因标志^[92]。吉林大学刘子玲团队采用基于非负矩阵分解和 Lasso 回归的模型对非小细胞肺癌的多模态（突变、CNV、甲基化、基因表达）数据进行了建模和分析，识别出四种具有不同免疫特性的亚型，通过对比不同的亚型，发现拷贝数异常对免疫检查点相关基因的表达量有重要关联，然后基于调控网络分析，识别出 7 个可能与免疫表型有关的关键基因^[131]。深圳大学吴松团队基于改进版一致聚类算法 CrossICC^[132]综合分析了膀胱癌的单细胞转录组、bulk 转录组、突变、CNV 等数据，识别出四类癌症子型，并发现同时有抑制性免疫微环境和免疫耗竭特征的亚型更可能对免疫治疗没有响应，而在特定亚型中 TGFβ 的表达水平可更准确地预测免疫治疗响应率^[133]。

1.4 人工智能在肿瘤靶点识别中的发展前景

目前，组学技术还在不断革新，人们观测肿瘤状态的手段愈发丰富多样。空间转录组技术可以实现在准确记录细胞空间位置的同时，测量细胞的转录组信息，且分辨率已可达到亚细胞尺度，这将为研究肿瘤免疫微环境的构成、提升免疫治疗响应率提供强大的推动力^[134, 135]。时序单细胞技术可以实现对同一个细胞在两个时间点的表达谱进行测量^[136]，使直接观测肿瘤发展过程及肿瘤耐药性产生过程的表达谱变化成为可能。此外，组学技术结合 CRISPR 基因编辑技术拥有强大的效力^[129]，可人为设定基因异常事件并测量每个细胞对此的响应，促进未知肿瘤驱动基因的发现，也为传统研究中难以靶向的功能丢失型突变提供了寻找具有合成致死效应的靶基因^[2]的新途径。组学测量技术和 CRISPR 等干预技术的共同进步，为破解肿瘤机制和发现新的肿瘤靶点提供了强有力的技术支持，同时也积累了不同条件下的海量肿瘤组学数据，为人工智能的应用提供了数据土壤。

基于人工智能的组学大数据研究，不仅在肿瘤药物及靶点研发中大放异彩，在广义的药物研发上同样举足轻重。单细胞组学可细致刻画各种药物临床实验中的实验组和对照组在不同分子层面的差异，辅助判定药物有效性，并揭示药物作用机制。但组学数据的高异质性、低信噪比、高维度、批次效应等数据特点，也对人工智能模型提出了区别于其他应用领域的独特要求。

1.5 本章小节

近年来测序技术的一系列突破为肿瘤靶点识别带来了新的契机。如何利用人工智能方法，高效地处理海量生物数据，从中寻找生物规律，是近年来的研究热点。目前，人工智能技术已被广泛应用于基因组、转录组、蛋白组和表观组等分析中，并引导发现了多个肿瘤靶点以及癌症发生发展的机制。随着人工智能技术在模型稳定性、可解释性、可迁移性以及与生物领域知识融合等方向的发展，可以期待人工智能技术在肿瘤靶点识别中得到更广泛的应用，进一步推动肿瘤靶点识别和肿瘤生物信息学的发展。

第 2 章 人工智能与苗头化合物筛选

2.1 人工智能与苗头化合物筛选概述

从第一种化合物药物诞生至今，科研人员一直在药物研发领域投入大量精力来对抗各种疾病，以提高人们的医疗保健水平^[137]。新型小分子药物的开发通常从生物学家确定疾病靶标开始，然后通过筛选技术在数以万计乃至数以百万计的化合物中挖掘出一组能够抑制或激活特定疾病靶标的活性分子。之后，再进行一系列的药代动力学、药效学、毒性测试以及结构修改来获得若干候选药物的苗头/先导化合物（图 2-1）。接着，经过大量动物测试以及多阶段临床试验之后筛选出最佳候选药物。最后，经过药监局审核批准之后，候选药物成为上市药物，进而可以被患者服用^[138]。然而，由于新药研发需要进行大量的实验且具有极高的失败率，通常完成一个新药的研发要花费 10-20 年以及 5-26 亿美元^[139-141]。

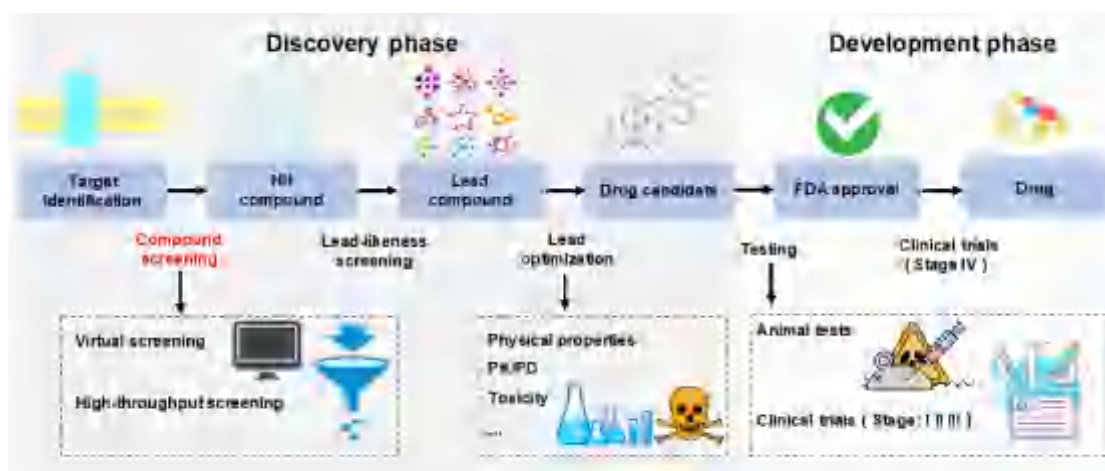


图 2-1 药物发现与发展过程^[142]

药物研发过程中，高通量筛选（High throughput screening, HTS）^[143]和虚拟筛选（Virtual screening, VS）^[144]是获得苗头/先导化合物的两种传统技术。然而，HTS 难以构建涵盖大量化合物的筛选库，VS

则需要数量众多的高质量三维结构^[145, 146]。这极大地限制了药物研发的速度。为解决这一局限性，工业界和学术界寻求利用人工智能技术来加速苗头/先导化合物筛选的进程。

近年来，深度学习作为人工智能中最重要的领域之一^[147]，不仅在自然语言处理^[148]和计算机视觉^[149]等多个领域得到广泛的应用并取得了巨大的进展，而且因其强大的学习和表征能力，同样也广泛应用于药物研发领域，比如靶标识别^[150]、化合物-蛋白质相互作用（Compound-protein interaction, CPI）预测^[151]、候选药物的理化性质预测^[152]、ADME/T（吸收、分布、代谢、排泄和毒性）预测^[153]和化学合成预测^[154]等。综上，深度学习技术逐渐成为降低药物研发成本、缩短研发时间、加速药物研发进程的希望之星^[155]。

其中，CPI 预测作为活性化合物筛选及寻找苗头化合物的关键步骤，不仅能够降低药物研发成本、缩短新药研发时间，而且可以提高新药研发的成功率。深度学习能够加速 CPI 预测主要基于以下两个方面。（1）现有大量的 CPI 数据可用。目前各种数据库中小分子和蛋白质之间的相互作用已经收集了数十亿条，深度学习可以通过自动挖掘化合物、蛋白质及其相互作用之间的隐空间关联进行高效、快速的 CPI 筛选^[156-159]。（2）各种形式的生物和化学数据都可以通过特定的深度学习模型实现自动提取特征。深度学习一般可以处理四种类型的数据：序列（如语音）、网格（如图像）、图（如网络）和决策流（如 GO 游戏）。对于 CPI 预测而言，化合物可以表示为序列^[160]或直接表示为分子图，蛋白质可以表示为序列^[161]或三维网格^[162]，CPI 可以被视为一个网络，包含两类节点：化合物和蛋白质，节点之间的边是它们的相互作用。

下文首先总结了 CPI 预测中常见的数据库；其次介绍了化合物、蛋白质的典型特征表示方法；之后，从设计范式的角度介绍了最先进的 29 种基于深度学习的预测模型，包括 11 种经典 Y 型框架模型、9

种基于注意力机制的模型以及 9 种基于复合物的模型；最后，总结了当前 CPI 预测的挑战和发展趋势，并简要介绍了若干典型应用案例。

2.2 基于深度学习的苗头化合物筛选

2.2.1 CPI 数据库

目前，生物实验已经积累了许多 CPI 数据。这些数据不仅包括小分子与蛋白质的相互作用，还包括它们之间的由 IC₅₀、K_i、K_d 和 EC₅₀ 等指标进行度量的结合亲和力。

STITCH 是目前最大的 CPI 数据库，包含通过实验测定和预测的 CPI^[163]，该数据库包含约 16 亿对相互作用，900 万种蛋白质和 43 万种化合物之间的结合亲和力数据。BindingDB 是第二大 CPI 数据库，它收集了 100 多万个小分子化合物和 8,000 多个潜在靶蛋白之间的 200 万个结合亲和力数据^[164]。与 STITCH 和 BindingDB 相比，PDBbind 是一个源自 Protein Data Bank (PDB) 的中型 CPI 数据库，它提供了超过 17,000 个实验确定的化合物-蛋白质复合物结构和亲和力数据，并额外提供了结合位点数据^[165]。与 PDBbind 类似，Binding MOAD 是 PDB 的另一个子集，它收集了超过 38,000 个具有高质量配体信息的蛋白质晶体结构，并使用从文献中提取的实验测定亲和力数据对其进行注释^[166]。此外，KIBA^[167]、Davis^[168]和 DUD-E^[169]也是研究中普遍使用的三个小型数据库。

药物靶点蛋白相关的综合数据库主要有：KEGG、DrugBank 和 TTD，包含已批准的药物、未批准的化合物、实验验证的靶点、蛋白质、途径、疾病和其他生物对象。其中，KEGG 整合了基因组、化学和系统功能信息^[170]；DrugBank 包含有关药物和药物靶点的详细信息^[171]；TTD 提供了靶点、靶向疾病状况、代谢通路信息，以及相应药物和配体^[172]。这些数据库中提供的已批准、已验证的靶标可以作为 CPI 预测模型的结果验证。

PubChem 和 ChEMBL 是两个综合性化合物数据库。PubChem 包含各种类型的化合物信息，包括 2D 和 3D 分子结构、化学和物理特性、生物活性数据、药理学、毒理学、药物靶点、代谢、安全性、相关专利和科学论文等^[173]。除了二维分子结构之外，ChEMBL 包含了 log P、分子量和 Lipinski 参数等计算预测的属性以及结合常数、药理学和 ADME/T 等科学文献中提取的生物活性数据^[174]。

2.2.2 蛋白质和化合物典型特征表示

应用机器学习的首要步骤是特征表示。传统的特征工程通常将化合物和蛋白质编码为高维特征向量，其中每个维度都反映了化合物和蛋白质的特定属性。根据化学结构的维度，化合物的特征描述符包括基于结构表示的 1 维、2 维及 3 维等。分子指纹 (fingerprint) 是经典的化合物特征提取方法。化合物的分子指纹可分为：基于子结构、基于路径、圆形、基于药效团以及复合分子指纹 (图 2-2)。

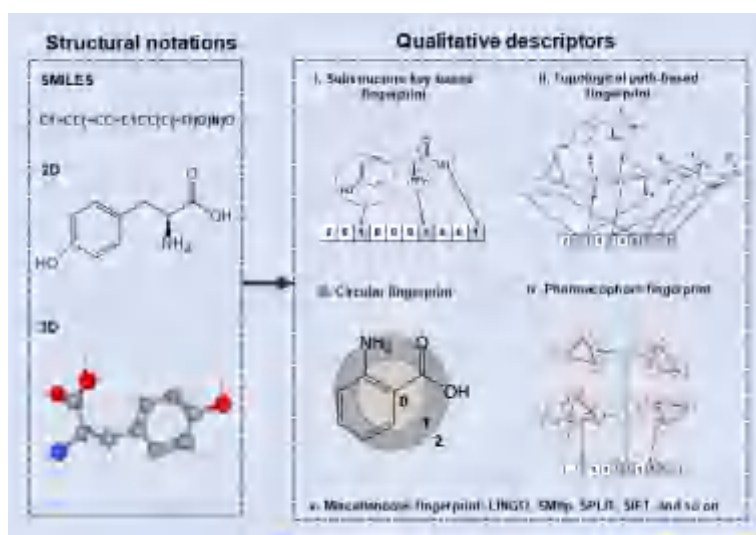


图 2-2 化合物结构表示及其定量描述符^[142]

从特征工程中衍生出的蛋白质描述符主要包括基于序列和基于结构的描述符。(1) 基于序列的描述符大致可分为基于 k-order 氨基

酸组成和基于物理化学性质的描述符。基于 **k-order** 氨基酸组成的描述符反映了蛋白质序列中氨基酸 **k-mers** (**k** 个氨基酸组成的短肽) 的出现频率^[175]。基于物理化学性质的描述符则利用每个氨基酸的物理和化学性质 (例如疏水性、范德华力和极性等) 将氨基酸序列映射为实值序列进行特征提取; (2) 基于结构的描述符可以大致分为基于拓扑结构、基于几何和基于距离图的描述符。基于拓扑结构的描述符根据从分子图生成的原子连接指数来描述氨基酸^[176,177]。基于几何的描述符则反映了与形状、大小、空间中的原子位置等相关的蛋白质结构特征^[178-180]。基于距离图的描述符首先通过计算 **C α** 原子之间或靠近残基之间的成对距离来获得蛋白质的距离图, 然后利用矩阵分解、网络或图像处理技术生成描述符值^[181]。

2.2.3 基于深度学习的 CPI 预测模型

本节将介绍基于深度学习的 **CPI** 预测方法。首先介绍了经典 **Y** 型架构模型。**Y** 型架构的两个分支分别编码化合物和蛋白质, 以获得相应的嵌入表示 (图 2-3 a)。然后, 描述了基于注意力机制的可解释模型。该模型通常利用额外的注意力层来表示形成相互作用的关键化合物-蛋白质特征。之后, 概述了基于绑定复合物的模型, 这些模型捕获了形成相互作用的细节因素。最后, 评估了这些模型在二元预测任务 (分类) 和结合亲和力预测任务 (回归) 中的性能。

2.2.3.1 经典 Y 型架构模型

作为最早的基于深度学习的 **CPI** 预测方法之一, **DeepDTA**^[182] 提供了一个 **Y** 型框架, 其中一个分支使用简化分子线性输入规范 (Simplified molecular input line entry system, **SMILES**) 编码化合物, 另一个使用一维序列作为原始蛋白质表示来编码蛋白质, 然后再由两个独立的卷积神经网络模块分别编码为相应的嵌入向量。接着将化合

物和蛋白质的嵌入表示拼接后输入到一个或多个全连接层，最后输出结合亲和力的预测结果。此外，化合物和蛋白质更多的特征表示也可以被整合到这个 Y 型框架中。作为 DeepDTA 的扩展，WideDTA 使用配体最大共同结构作为额外的原始化合物表示，使用蛋白质域和功能注释作为额外的原始蛋白质表示^[183]。类似地，DeepConv-DTI^[184]用 CNN 提取蛋白质序列的特征表示，并使用摩根指纹表示化合物。MDeePred 模型构建了多种类型的蛋白质特征(氨基酸的序列、结构、进化和理化特性)，同时使用圆形分子指纹表示化合物，然后再分别使用 CNN 和前馈神经网络将蛋白质和化合物编码为对应的嵌入向量，最后进行拼接并输入前向 DNN 中进行预测^[185]。

由于存在大量活性未知的化合物和不完全注释的蛋白质，近期的工作试图利用大量未标记的化合物字符串和未标记的蛋白质序列来改善序列数据的表示。DeepCPI^[186]借鉴自然语言处理 (Natural language processing, NLP) 技术，利用潜在语义分析对化合物进行编码，利用 Word2Vec 以无监督的方式对蛋白质序列进行编码，并将生成的化合物和蛋白质的表征共同输入多模态 DNN。GANsDTA 采用两个生成对抗网络 (Generative adversarial networks, GANs) 分别作为化合物 SMILES 字符串和蛋白质序列的无监督特征提取器，并进一步将提取的特征输入一维 CNN 进行结合亲和力预测^[187]。除了基于 CNN 对 SMILES 字符串和氨基酸序列进行特征编码外，MultiDTI^[188]建立了一个额外的异构网络，利用网络中化合物、蛋白质、副作用和疾病之间的关联作为约束条件来生成化合物和蛋白质的最终表示。

由于化合物结构可以直接表示为分子图，图神经网络在小分子特征表示方面大放异彩。例如，GraphDTA 使用四种类型的图神经网络获得基于化合物的图表示，包括图卷积网络 (Graph convolutional network, GCN)、图注意力网络 (Graph attention network, GAT)、图同构网络 (Graph isomorphism network, GIN) 和 GAT-GCN 组合，并采

用多层一维 CNN 来获得基于序列的蛋白质^[189]表示。类似地, MONN 利用 GCN 获得分子图表示^[190], 对原子和化学键使用独热编码, 同时利用一维 CNN 对经过 BLOSUM62 数值化处理的蛋白质序列进行表示。另外, 蛋白质也可以通过 distance map^[191]或 contact map^[192]来进行表示。例如, DGraphDTA 首先通过 PconsC4^[193]从每个序列生成蛋白质的 contact map^[194], 然后将该图构建为蛋白质图, 其中节点为氨基酸, 边表示其相邻关系, 最后在分子图和蛋白质图上采用图神经网络, 分别获得化合物表示和蛋白质表示^[195]。

2.2.3.2 基于注意力机制的模型

尽管上述方法实现了高精度的 CPI 预测, 但它们不能明确指出哪些因素对相互作用有贡献, 以及相应的贡献程度。由于注意力机制在揭开“黑箱”方面具有重要的能力, 将注意力层整合到经典 Y 型架构模型中有利于解释化合物与蛋白质形成相互作用的原因。近期的 CPI 预测工作通过将注意力层整合到经典模型中, 在寻找成对的关键蛋白质子序列(如残基或 n-gram 氨基酸)和化合物关键子结构方面起到了重要作用。它们的研究结果表明, 注意力机制有利于解释化合物与蛋白质相互作用的原因。

大多数基于注意力机制的模型都是分别针对化合物和蛋白质设计注意力模块(图 2-3 b)。Gao 等人在 CNN 以及 LSTM (Long short-term memory) 之后分别使用了两个注意力模块^[196], 揭示了对结合有关键作用的蛋白质残基和化合物原子。同样, Abbasi 等人提出了 DeepCDA^[197], 将化合物 SMILES 字符串、蛋白质序列分别传入一个 LSTM 块和一个 CNN 块, 然后通过注意力机制来表明化合物子结构和蛋白质残基之间交互的强度。Zheng 等人设计双向 LSTM 结合多头注意力模块来解释对相互作用形成具有重要作用的关键蛋白质残基和化合物原子^[198]。

此外，一些研究为化合物和蛋白质设计了联合注意力模块。**AttentionDTA** 沿袭传统的 Y 型框架，在用数字编码化合物/蛋白质序列后，使用两个一维 CNN 来提取化合物和蛋白质各自的表征，然后应用联合注意力模块来捕获化合物子序列和蛋白质子序列，从而帮助寻找结合位点^[199]。**Tsubaki** 等人应用联合注意力机制来捕获化合物的子结构和 3-gram 氨基酸对于形成 CPI 的贡献^[200]。**Chen** 等人提出的 **TransformerCPI** 模型，首先利用 **Word2Vec** 获得蛋白质的预训练嵌入向量，然后将其传入 **Transformer** 编码器，之后通过 **GCN** 获得化合物的嵌入向量，最后利用 **Transformer** 解码器中的多头注意力层来表明化合物原子和 3-gram 氨基酸对于形成 CPI 的贡献程度^[201]。**MATT_DTI** 使用了一个额外的关系感知自注意力模块来加强药物化合物的信息，然后采用联合多头注意力模块来模拟化合物表征和蛋白质表征之间的桥梁^[202]。

2.2.3.3 基于复合物的模型

通常大量的蛋白质结构是很难获取的，但当有化合物-蛋白质复合物时，设计基于复合物的模型有助于 CPI 的预测。在早期阶段，**AtomNet** 直接采用三维 CNN，将化合物-蛋白质复合物离散成三维网格后，获得活性复合物和非活性复合物的表征^[162]。一些类似的工作^[157, 203, 204]采用不同的三维 CNN 结构来编码复合物。由于三维网格的计算复杂性，近期一些工作关注化合物-蛋白质结合口袋的特征表示而不是整个复合物的特征表示法来加快复合物嵌入表示^[205-207]。例如，**Lim** 等人提出了一个以距离感知的图聚焦算法来获得三维口袋嵌入，该算法利用两个相邻矩阵上的一个共享门增强图注意力层来编码原子，并在嵌入空间刻画非共价相互作用的复合物和其单个分离结构的差异^[205]。**Cang** 等人通过采用新的代数拓扑描述符（**Element-specific persistent homology, ESPH**）^[206]将蛋白质-化合物复合物的特征转化为

多通道一维图像表示，从而利用多通道 CNN 进一步编码复合物^[207]。

此外，一些工作也在 Y 型框架下分别表示蛋白质结构和化合物结构。Gonczarek 等人通过对每个分子分别应用可学习的原子卷积和 softmax 操作，生成固定大小的蛋白质和小分子的指纹图谱^[208]。Gomes 等人^[209]通过扩展 Y 型框架，提出了三个并行的 ACNN，这是带有径向池化层的三维 CNN 变体，通过刻画蛋白质与化合物的联合体与复合体之间的差异来表示结合口袋中的非共价相互作用。Torng 等人建立了一个预先训练好的图自动编码器来提取蛋白质口袋（而不是整个蛋白质）的通用特征，并把预先训练好的架构作为口袋图卷积层来表示蛋白质^[210]（图 2-3 c）。

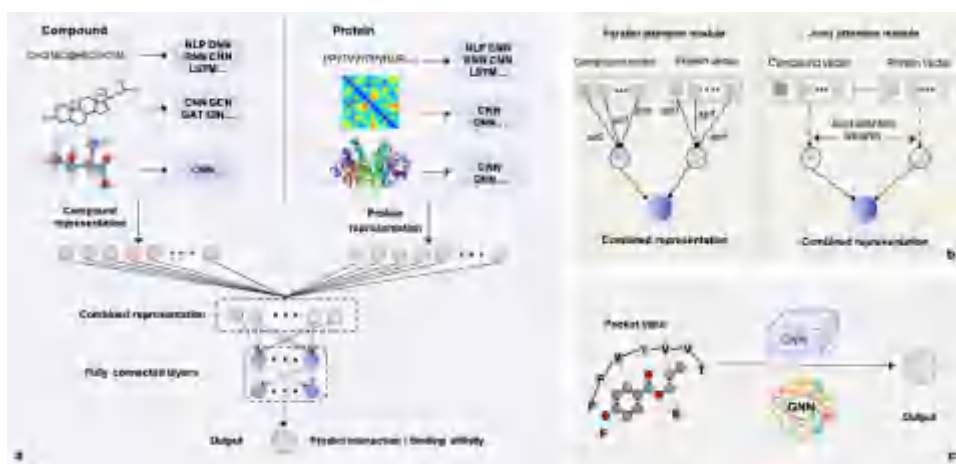


图 2-3 基于深度学习的 CPI 预测模型框架^[142]

2.2.3.4 CPI 预测模型性能评估

在本节中，我们将在二元预测任务和亲和力预测任务下比较最先进的基于深度学习的 CPI 预测模型。前者是一项分类任务，需要区分化合物是否与蛋白质结合；后者是一项回归任务，要求推断化合物与蛋白质结合的强度。总共对比了 29 个基于深度学习的预测模型和一个集成预测模型 DeepPourse^[211]，分别在三个数据集上评估了 13 个模型的二元分类任务（表 2-1）和 12 个回归预测任务（表 2-1）。

为了在二元预测任务中进行公平比较，我们选择了大多数论文中经常使用的三个数据集作为基准数据集：DUD-E、Davis 和 Human。这些模型的性能通常使用受试者工作特征曲线下面积（Receiver operating characteristic-area under curve, AUC）来进行衡量，AUC 越大表明预测效果越好^[161]。（1）在 DUD-E 数据集上对比了 8 个模型，其中 MONN（AUC=0.974）^[161]，DrugVQA（AUC=0.972）^[198]和 Lim 等人（AUC=0.968）^[205]这三个最近发布的模型显著优于其他 5 个模型。（2）在 Davis 数据集上对比了 4 个模型，其中 MolTrans(AUC=0.907) 优于其他模型。（3）在 Human 数据集上调查了 4 个模型，其中 TransformerCPI^[212]是最好的（AUC=0.973）^[213]。此外，DrugVQA 在 DUD-E 和 Human 上均表现出令人满意的性能。总的来说，二元预测任务中的优秀模型包括 MONN^[161]、DrugVQA^[198]、Lim^[205]、TransformerCPI^[212]和 MolTrans^[214]。

表 2-1 CPI 分类任务预测性能评估^[142]

数据库	方法	年份	AUC	数据划分
DUD-E	Ragoza et al.	2017	0.868	3-fold cross-validation
DUD-E	Tornig et al.	2019	0.886	4-fold cross-validation
DUD-E	DrugVQA	2020	0.972	3-fold cross-validation
DUD-E	AtomNet	2015	0.895	Train (72targets) Test (30targets)
DUD-E	Gonczarek et al.	2018	0.904	Train (72targets) Test (30targets)
DUD-E	Tsubaki et al.	2019	0.940	Train (72targets) Test (30targets)
DUD-E	Lim et al.	2019	0.968	Train (72targets) Test (25targets)
DUD-E	MONN	2020	0.974	Train (72 targets) Test (30 targets)
Davis	DeepDTA	2018	0.880	5-fold cross-validation
Davis	Tsubaki et al.	2019	0.840	5-fold cross-validation

数据库	方法	年份	AUC	数据划分
Davis	DeepConv-DTI	2019	0.884	5-fold cross-validation
Davis	MolTrans	2021	0.907	5-fold cross-validation
Human	Tsubaki et al.	2019	0.97	5-fold cross-validation
Human	GraphDTA	2020	0.96	5-fold cross-validation
Human	DrugVQA	2020	0.964	5-fold cross-validation
Human	TransformerCPI	2020	0.973	5-fold cross-validation

在结合亲和力预测任务中，我们选择了广泛使用的 Davis 数据集作为基准数据集，并将 12 个基于深度学习的回归模型进行比较（表 2-2）。亲和力预测的性能使用一致性指数（Consistency index, CI）和均方误差（Mean square error, MSE）来衡量。CI 越大，MSE 越低，预测效果越好。调查结果表明，DGraphDTA^[195]在 CI 和 MSE 方面表现出最佳性能。此外，我们发现尽管基于注意力机制的模型（例如 AttentionDTA、DeepCDA、MATT_DTI）有更好的解释性，但它们并未优于经典模型 DGraphDTA。因此，良好的特征表示对于 CPI 预测至关重要。另外，与二元任务相比，亲和力预测任务比二元预测任务更难。

表 2-2 CPI 亲和力预测任务性能评估^[142]

方法	年份	蛋白质表示	化合物表示	CI	MSE
DeepDTA	2018	1D+CNN	1D+CNN	0.878	0.261
DeepCPI	2019	1D+NLP	1D+NLP	0.867	0.293
WideDTA	2019	1D+CNN	1D+CNN	0.886	0.262
AttentionDTA	2019	1D+CNN	1D+CNN	0.893	0.216
GANsDTA	2020	1D+GAN	1D+GAN	0.881	0.276
DeepGS	2020	1D+CNN	1D+CNN&2D+GAT	0.882	0.252
MDeePred	2020	2D+CNN	1D+DNN	0.886	0.254

方法	年份	蛋白质表示	化合物表示	CI	MSE
DeepCDA	2020	1D+CNN+LSTM	1D+CNN+LSTM	0.891	0.248
DeepPourse	2020	1D+ACC	2D+MPNN	0.881	0.242
GraphDTA	2020	1D+CNN	2D+GIN	0.893	0.229
DGraphDTA	2020	2D+GNN	2D+GNN	0.904	0.202
MATT_DTI	2021	1D+CNN	1D+CNN	0.891	0.227

2.3 深度学习在苗头化合物筛选中的发展前景

2.3.1 趋势与挑战

尽管当前基于深度学习的模型展示了良好的 CPI 预测性能，但仍然存在如下趋势和挑战：

(1) 如何利用大量未标记的化合物和蛋白质。无监督学习（如 DeepCPI^[186]）、半监督学习（如 GANsDTA^[187]）和预训练策略（如 DeepAffinity^[160]）的成功应用表明，利用丰富的未标记数据（序列）可以增强化合物和蛋白质表示，从而实现更好的 CPI 预测。如何利用更多的深度学习技术（例如对比学习）学习更好的化合物和蛋白质表示成为了未来发展趋势。

(2) 由于具有标记的化合物表示数据较为稀缺，因此可以通过自监督学习对未标记数据的模型进行预训练，然后将学习到的模型转移到下游任务。最近的一些研究工作利用自监督学习方法来表示基于化合物 SMILES 字符串^[215]和分子图的表示^[216]。然而，这些方法主要侧重于学习节点级表示，不能显式地学习全局图级表示，导致图级任务的收益有限^[217]。因此，开发有效的图级自监督方法非常重要。

(3) Y 型架构趋向于演变为叉状框架，以适应多种化合物或蛋白质来源（例如 WideDTA^[183]、DeepGS^[218]）。虽然利用 3D 结构可增强化合物和蛋白质的表示。然而，这些结构不仅需要高额算力，而且

基于复合物的模型与利用化合物和蛋白质序列以及分子图的模型相比并没有得到显著的性能提升。一种可能的解决方案是获得更复杂的复合物数据,而另一种是开发基于迁移学习的模型来将化合物和靶标表示从序列域迁移到 3D 复杂域,后者在未来几年可能更实用。

(4) 现有基于深度学习的预测模型^[161, 205, 209]的分析表明,在表征化合物-蛋白质结合对时,非共价相互作用的表示是至关重要的。结合位点有助于理解非共价相互作用。目前,基于深度学习的方法已经逐渐应用在结合位点预测中,例如 DeepSite^[219]和 DeepSurf^[220]。这些方法可以整合到现有模型中来增强 CPI 预测。

(5) 注意力机制已在一定程度上展示了其对于 CPI 形成机制的可解释性^[160, 196, 221]。但是,当前的研究方法仅通过一个或几个示例进行评估,在大规模数据集中可能出现不一致的解释性结果^[161]。此外,目前还缺乏统一的标准来系统地评估各种基于注意力模型的可解释性。因此,应该形成统一的评估指标,以便挖掘化合物和蛋白质的结合规则。相比之下,与基于深度学习的“黑盒”模型相比,“白盒”模型可能是捕捉结合机制的新尝试^[222]。总之,提高化合物-蛋白质结合机制的可解释性是未来发展趋势,不仅可以筛选潜在 CPI,还可以指导后续的先导化合物优化。

2.3.2 实际应用

CPI 预测只是寻找苗头/先导化合物的第一步。为了加快苗头化合物的发现过程,一些网络服务器根据预测的 CPI 来执行化合物虚拟筛选^[223]。例如,基于 DNN 的 DeepScreening 可根据结合亲和力针对特定靶标进行大规模的化合物筛选^[224]。BindScope 是一个基于 CNN 的交互式网络应用程序,提供了大规模的活性/非活性化合物分类。

除了 CPI 预测中的通用模型和 Web 服务器外,研究人员还开发了针对特定疾病的方法。例如,Zhang 等针对 2019-nCov (SAR-COV-

2) 主要治疗靶点 3C 样蛋白酶, 开发了基于深度神经网络的药物筛选管线, 用于快速筛选其候选配体和多肽药物^[225]。TranScreen 通过在多任务化合物数据库 (MoleculeNet^[226]) 上预训练一组 GCN, 将学习到的 GCN 迁移到人类癌症主要突变来源之一 p53 基因的活性化合物筛选过程中^[227]。

近年来, 随着深度学习的快速发展, 基于深度学习的小分子药物研发也取得重大突破且部分候选药物已进入临床试验。例如, Insilico Medicine 和药明康德在 2019 年通过深度学习模型, 仅在 23 天就产生了 6 个先导化合物, 并在 46 天内从中筛选出 1 个具备良好药代动力学行为的候选药物, 最后通过实验验证了其对于 DDR1 激酶的高效抑制作用^[151]。2021 年 4 月 9 日, 由 Exscientia 设计的首个基于 AI 技术的肿瘤免疫分子成功进入临床 (<https://www.exscientia.ai/>)。2021 年 4 月 30 日, 药物牧场 (DRUG FARM) 基于 AI 技术发现了治疗乙肝的候选药物 DF-006 并获批国际中心 (新西兰) 一期临床, 是中国首个从靶点发现进入临床的全球首创新药 (<https://www.drugfarm.com/home>)。Tan 等人通过开发 RNN 和 MTDNN (一种多任务 DNN), 成功筛选出了具有所需靶标活性的新型抗精神病分子^[228]。Liu 等人通过建立预训练的自注意力消息传递神经网络 (P-SAMPNN), 鉴定了五种抗骨质疏松症生物活性天然产物^[229]。这些案例都证明了深度学习技术赋予新药研发的广阔场景。

深度学习技术能够快速发现新的活性化合物并产生新的苗头/先导化合物, 可推动新型药物发现范式的发展。随着我国医药产业创新发展的加速推进, 人工智能技术尤其是深度学习技术或将成为制药行业的一个宝贵工具。

2.4 本章小节

CPI 筛选是活性化合物筛选及寻找苗头化合物过程中的重要环

节。基于深度学习的 **CPI** 预测模型利用化学和生物大数据加快了筛选过程。本章对 **CPI** 预测方法进行了全面的调查。首先，简要地回顾了小分子化合物、蛋白质、两者复合物的常见数据库。其次，介绍了化合物、蛋白质的典型特征表示方法，包括由传统特征工程产生的各种分子指纹和不同的描述符以及基于序列和基于结构的描述符的蛋白质特征。然后，从设计范式的角度简要介绍了最先进的深度学习 **CPI** 预测模型，包括经典两分支特征表示的 **Y** 型模型、利用注意力机制挖掘关键成对“蛋白质子序列-化合物子结构”的 **Y** 型模型、聚焦化合物-蛋白质的复合物或其结合袋表示的模型。接着，在经典数据集上研究对比了各类预测模型的性能。最后，总结了当前基于深度学习的 **CPI** 预测方法存在的趋势、挑战以及实际应用。

第 3 章 人工智能与药物从头设计

3.1 基于人工智能的药物从头设计概述

从头药物设计是根据靶点结构直接构造出形状和性质互补的全新配体分子，因其能提出结构全新的具有启发性的先导化合物，在药物研发过程中具有重要的原创性意义。

计算机科学的不断进步，以及量子化学、分子力学和分子动力学等计算方法对药物科学研究的渗透，催生了计算机辅助药物设计（Computer aided drug design, CADD）技术。该技术使得计算机科学中的大数据运算、数据库和图形学方法能够广泛应用于药物小分子和生物大分子的化学结构研究。CADD 的出现为构象分析、药物相互作用模式认定、机制推测和结构-药效研究提供了先进的技术手段，可以缩短药物发现周期、减少研究投入与风险，因而得到了学术界和产业界的重视和关注。生命科学、药学、化学等医学前沿技术的不断进步，产生了丰富的药物分子及其对应的药理活性、基因组学、蛋白质组学和结构以及药靶结合结构等数据，为 CADD 技术的发展提供了数据基础。例如，第二章中所提到的公共化合物数据库 PubChem^[230]、ChEMBL^[231]、DrugBank^[232]以及 PDBBind 数据库^[233]，（此外还有 DrugMatrix 数据库^[234]和 PharmGKB 数据库^[235]。前者是药物安全和毒性数据库，包含 600 种药物的毒理学数据，还包含药物治疗下的大规模大鼠基因表达数据。后者覆盖药物分子临床信息的药物基因组学知识资源。）

如此浩繁复杂、多源多样、异质非结构化的化学和生物数据，既为药物设计技术的进步提供了前所未有的可能性，也对具体的存储管理、准确的建模分析和高效的应用设计提出了考验和挑战。高校和学术研究机构的新药研发人员大多数仍在使用上一代 CADD 技术，靶向先导化合物筛选命中率为 0.1%~1%，计算模拟误差大、实验验证成

本高，因此，国际上能进行自主新药研发的高校和科研机构较少。近年来，以深度学习为代表的人工智能日渐成为各种场景下的应用、数据分析和建模的利器，其蓬勃发展为 CADD 技术的发展提供了新的机遇。深度神经网络具有强大的数据逼近和拟合能力，通过优化损失函数 (loss function) 能够进行参数优化和模型学习，可以自动地进行数据潜在特征的抽取和挖掘，在面对维度高、分布复杂、难以刻画的数据时具有巨大优势。因此，深度学习模型可以高效地从药靶生化数据中挖掘和提取相关的复杂模式和特征，并可以作为启发式函数为搜索优化算法提供有效的指导信息，进而更好地完成如预估药物的理化性质、预测药靶结合的强弱程度、从头生成满足指定性质的新分子等药物设计的相关任务。

3.2 深度生成模型与小分子药物从头设计

近年来，为满足药物分子的结构合理性、生化性质、靶向亲和力等不同方面的设计需求，研究人员使用基于深度神经网络的数据样本生成模型即深度生成模型，对药物分子从头生成方法进行了广泛的探索。

3.2.1 小分子药物合理结构的生成模型

有机物分子的化学结构可能性可达到 10^{60} ，但具有结构合理性的可能化合物空间相对较小。因此，在具有结构合理性的化学空间中探索有机物结构的自然分布（即生成具有结构合理性的新颖分子）就成为了一个重要的问题。有研究者^[236]使用一个药物分子数据库训练和测试深度学习模型中结构较为简单的循环神经网络 (Recurrent neural network, RNN) 模型，仅使用了数据库 0.1% 的样本就能重建恢复出 68.9% 的分子样本的结构，可见深度学习模型对于分子结构分布具有很好的拟合能力。Mahmood 等人^[237]使用了掩码图模型进行分子图结

构的随机生成。掩码图模型借鉴了自然语言处理中掩码语言模型的思想，对生成的小分子药物的图结构进行随机遮罩，再训练消息传递神经网络（Message passing neural network, MPNN）模型以补全图结构中被遮罩的部分。而后采用吉布斯采样算法，以既有的小分子药物或者按照数据集中分子的分布采样得到的初始样本为基础，每次随机遮罩部分原子和化学键，使用训练得到的消息传递神经网络模型补全图结构。重复多次后，就可以得到结构合理且具有高新颖性的小分子。

3.2.2 满足生化性质要求的小分子药物生成模型

药物分子进入人体组织器官发挥作用的前提条件是需要满足一定的物理化学性质和生物化学性质，如脂水分配系数、类药性和自然产物类似性等。因而，符合相关生化性质要求的小分子结构生成是一个重点研究方向。

基于药物分子的一维描述形式是 SMILES 字符串，许多研究者的通用思路是使用神经网络的序列模型来学习分子生成策略的概率分布，而后使用各种不同的优化算法对分子生成策略作进一步提升。ChemTS 模型^[238]就将分子生成任务建模为马尔可夫序列决策问题，通过 RNN 学习分子分布，得到分子生成的策略 $\Pr(X_i|X_{1,2,\dots,i-1})$ ，此过程也同时包含了 SMILES 语法规则的学习（如括号匹配，环的闭合规则），然后通过蒙特卡洛树搜索（Monte carlo tree search, MCTS）针对特定的优化目标（如 logP）进行分子化学空间搜索。具体来说，树中每个节点代表 SMILES 的一个符号，每一个分子的 SMILES 序列对应一条从根节点到叶节点的路径。树搜索的过程可看做树不断扩展生长的过程，每一次的搜索都从根节点出发，通过 UCT 公式选择性质好的节点，直到选到叶节点，然后通过 RNN 学到的策略概率对叶节点进行扩展，通过 rollout 的奖励值进行反馈，进而指导下一次树搜索朝着更理想的方向进行。通过大量搜索，该模型找到的分子化学性质

会越来越符合既定的优化目标。Popova 等人^[239]提出了一种基于强化学习的针对生化性质的小分子药物生成方法。该方法以小分子药物的随机生成模型和对于给定生化性质的预测模型为基础，采用强化学习中的策略梯度算法，将随机生成模型和生化性质预测模型结合，把性质预测模型的输出作为强化学习算法当中的奖励值，通过策略梯度更新随机生成模型的参数，使得模型生成的小分子药物具有所需的生化性质。这种方法还通过引入栈循环神经网络作为随机生成模型，解决了传统上使用普通循环神经网络的生成模型不具备计数能力、生成的小分子药物结构正确性即“合法性”等问题。实验结果表明，经过策略梯度算法调整后的生成模型所生成分子的生化性质与原本的随机生成模型得到的分子生化性质有显著的偏移，证明该模型能够生成出具有所需生化性质的小分子药物。类似的，Olivecrona 等人^[240]通过一种基于策略的强化学习方法来优化调整预先训练的 RNN 模型，以生成具有用户定义属性的分子。在生成具有多巴胺 2 型受体活性的小分子的测试示例中，该模型生成的结构中 95% 以上的分子具有活性。此外还有一些关于编解码器的应用研究，例如借助变分自编码器将分子编码为一维隐空间向量，再利用多层感知机（Multi-layer perceptron, MLP）对其进行性质预测与改进，最后利用解码器进行解码以生成优化后的分子^[241]。Kusner 等人^[242]借助上下文无关的语法规则（context free grammar）对分子的编码解码进行限制，给予变分自编码器（VAE）关于如何产生合法分子字符串的显式知识，从而使得 VAE 模型能够生成更多满足给定性质的合理分子，同时也学习到了更平滑的向量表示空间。

药物的二维分子图是基于分子的原子-化学键型的表示形式，研究人员普遍利用图神经网络（Graph neural networks, GNN）相关模型进行特征表示学习和分子生成。例如 GraphVAE 模型^[243]同样使用 VAE，不同于前文提及的方法^[241]，该方法的输入输出均为图结构，利用图

匹配去训练 VAE 网络。Jin 等人^[244]将分子图进行了符合化学语义的压缩，即将分子中的环和一些官能团表示为节点，使得分子的生成更具化学语义合理性和结构合法性——VAE 编码器将分子表示成若干亚结构（构建单元）和亚结构之间连接方式，通过训练集学习构建单元的特性、出现频率和连接规则，最后通过解码器，将新拼装出的亚结构组合进行解码，得到全新的分子拓扑图结构。尽管用于堆砌连接构建分子的片段种类有限，但事实上深度神经网络产生出的结构差异巨大，这使得最终产生的分子仍具有较高创新性。为了提高 VAE 生成的分子合法性，受启发于约束优化问题，研究者采用拉格朗日乘数法对 VAE 中的损失函数添加约束，使得修改后的损失函数与原损失函数保持相同解^[245]。Li 等人^[246]使用了顺序图生成模型，模型中可以并入条件标记以生成分子特性接近指定目标分数的分子。为更直接地提高分子的给定性质要求，研究者提出了如 GCPN、MolGAN 的结合强化学习（Reinforcement learning, RL）算法的图生成模型。GCPN^[247]将分子的生成看成序列决策问题，将图生成建模为添加节点和边的马尔可夫决策过程，同时使用策略梯度优化对抗性损失和特定领域的奖励。MolGAN^[248]整合了 convGNN（Convolutional graph neural network）、对抗性生成网络和强化学习目标，以生成具有所需特性的分子。MolGAN 由生成器和鉴别器组成，它们相互竞争以提高生成器的真实性。在 MolGAN 中，生成器试图提出一个伪图及其特征矩阵，而鉴别器旨在将伪样本从经验数据中区分开来。此外，MolGAN 引入了与鉴别器并行的奖励网络，以鼓励生成的分子图具有指定属性。与 GCPN 通过一系列动作来生成图不同，MolGAN 可以直接生成完整的图，这对小分子生成很有效。另一类重要的深度生成模型是自回归流（Autoregressive Flow）。GraphAF^[249]是基于自回归流模型的分子图生成模型的代表之一，它具有强大的数据分布建模能力和进行数据密度估计的模型灵活性，在训练时还可以有效地进行并行计算。针对优质

分子的多目标优化任务，Xie 等人^[250]提出了 MARS 模型，其主要思路是将寻找优质分子的问题转化为关于分子综合评价的采样问题，模型包含三个重要组成部分：(1) MCMC 分子采样框架；(2) 分子图修改模型；(3) 模型的自适应训练。具体而言，在 MCMC 的采样过程中，MARS 利用一个建议分布 (Proposal distribution) 来进行分子状态的转移，即从任意初始分子状态 x_0 开始，在每一个时间点 t ，MARS 都根据当前分子状态 x_{t-1} 和建议分布 $q(x'|x_{t-1})$ 来生成一个新的候选分子 x' ，再根据接受概率 $A(x_{t-1}, x')$ 接受或拒绝候选的 x' ，如此迭代重复，便能生成一系列的分子^[243]。对于建议分布 $q(x'|x_{t-1})$ (即原本分子到候选分子的转移)，MARS 为基于结构片段的分子图优化显式地建模，并用 MPNN 网络对其进行参数化。关于分子图结构的具体修改，MARS 考虑分子结构片段的添加与删除两种操作，即从分子结构片段库中选取一个片段并将其拼接到某个特定的原子上，以及切断某条特定的化学键并移除与其相连的片段，这使得 MARS 的修改方式能够尽量覆盖可以合法改造的结构。以上方法从不同的深度学习模型设计的角度，对于一维分子序列和二维分子图的生成和给定性质优化进行了探索。

3.2.3 基于靶点蛋白结构的小分子药物生成模型

目前基于靶点蛋白结构的药物设计发展十分迅速，但仍然面临很大挑战。相关设计方法一般包括虚拟筛选和基于结构的全新药物设计，前者是在已知的化合物分子库中通过计算手段直接搜索筛选得到针对给定靶点的高活性化合物，而后者是根据分子和靶点结合部位从头生成符合要求的化合物，这可以突破已知化合物的数据限制，从而发现新颖的候选活性药物分子。我国有不少课题组在这方面利用计算化学知识和传统优化方法进行研究。例如，北京大学来鲁华教授团队长期开发和维护药物从头设计程序系统 LigBuilder，经过四代发展的

LigBuilder 系统已拓展到针对多靶标的药物设计和共价化合物设计 [251, 252]。浙江大学侯廷军课题组 [253] 通过对计算模拟方法以及体内/体外生物活性评价实验的整合, 针对多种重要药物靶点进行了先导化合物的筛选与优化。

近年来, 深度学习方法逐渐在基于靶点蛋白结构的药物设计方面发挥出巨大潜力。**Grechishnikova** 等人 [254] 尝试利用 **Transformer** 模型将基于蛋白的分子设计转变为一个“翻译任务”, 即将蛋白序列看作“蛋白语言”, **SMILES** 看作“分子语言”, 每一个 **SMILES** 序列会对应一个蛋白序列, 通过将成对的蛋白序列和分子序列输入 **Transformer**, 学习分子蛋白的联合分布, 得到有条件的分子生成策略。**Transformer** 采用注意力机制, 能够捕获长序列之间 (如蛋白序列) 的依赖关系。该方法在推理生成时也采用自回归的方式: 即给定一个靶蛋白序列, 借助束搜索策略搜索分子空间, 从头生成分子。**Zhavoronkov** 等人 [255] 利用一种叫做 **GENTRL** (**Generative tensorial reinforcement learning**) 的 **GAN** 模型, 通过分子结构空间与连续隐空间的互相映射, 实现分子结构的生成和演化, 进而实现了酪氨酸激酶盘状结构域受体 (**DDR1**) 靶点活性分子的快速 **AI** 设计。该工作在短短 21 天内, 就以酪氨酸激酶 **DDR1** 为靶点, 设计了 40 个潜在抑制剂, 其中选择合成的 6 个化合物中, 有 4 个具有 **DDR1** 抑制能力。

空间结构是决定药物分子性质和理解其在真实物理世界中进行靶向作用原理的最关键因素之一, 因此捕捉受体和配体分子的三维空间结构特征对于分子生成至关重要。**Fabritiis** 等人 [256] 基于 **BicycleGAN**, 设计出 **LIGANN**。其网络结构主要分为两部分: 第一部分以蛋白质口袋为输入, 通过 **BicycleGAN** 输出其对应配体的形状, 其输入输出均用三维图像的像素格式来表示; 第二部分将配体的形状通过一个捕获网络最终生成 **SMILES** 分子。**Xu** 等人 [257] 设计出一种包含蛋白关键残基信息的库伦矩阵 (**coulomb matrix**), 计算矩阵对应的

特征值向量来表示蛋白口袋的三维结构信息，该结构作为条件 RNN 的输入。由此训练出来的模型可以生成与该蛋白口袋结合更优的配体分子。Luo 等人^[258]提出了一种基于三维坐标的方法，来生成针对指定靶点蛋白质口袋的小分子药物。该方法利用了靶点蛋白质口袋中原子的三维坐标信息和已知小分子药物配体的三维坐标信息，通过 k 近邻构建原子图结构，使用类似图神经网络的模型预测目标三维空间中的位置是否存在原子以及存在何种原子的概率。然后通过口袋空间中进行网格状的采样近似地得到不同种类原子在口袋空间中分布的条件概率，通过束搜索采样得到小分子药物的构成原子和对应的三维坐标。最后通过基于化学规则、原子的种类和三维坐标信息构建出生成的小分子药物。该方法通过图神经网络进行编码，并利用了靶点蛋白质口袋的三维坐标信息，使得所生成的小分子药物的原子也具有三维坐标的信息，解决了传统基于序列的模型在这一方面可解释性不足的问题。为了直接根据靶标蛋白结合位点特征生成三维分子结构，来鲁华、裴剑锋教授团队设计了 DeepLigBuilder 模型^[259]，该模型将配体神经网络 (Ligand neural network, L-Net) 和 MCTS 相互结合，完成基于结构的药物从头设计任务。网络结构上，L-Net 的状态编码器由图卷积网络结构构建，结合了图聚合和旋转协方差等特征，增加了网络的接受域的大小，并且在训练过程中加入 3D 误差，进而使其更有效地训练和采样。状态编码器分析现有的结构并将信息编码成连续表示后，策略网络使用该表示来决定分子应该如何编辑，即添加多少原子到分子中、每个新原子和键的类型以及新原子的 3D 位置。与其它三维分子生成模型相比，L-Net 的主要优势是能够端到端地生成三维分子结构。这一特性使得 L-Net 的使用和扩展更加方便，能以即插即用的方式与许多技术结合，实现目标导向的分子设计。在这项工作中，研究人员将其与 MCTS 结合，并在蛋白结合袋内直接生成具有高预测亲和力的配体，这也是第一次将 3D 生成模型与 MCTS 结合来解决

与基于结构的药物发现相关的问题。MCTS 通过迭代构建搜索树，树中的每个节点代表分子生成过程中的一个中间状态。在每次迭代中，模型首先从搜索树中选择一个有希望的状态 (**selection**)，枚举该状态的可能操作 (**expansion**)，并执行 **rollout** 以生成其余的分子结构 (**simulation**)。通过收集奖励信息并反向传播，更新每个节点的 Q 值。其中，研究人员使用了 **smina** 软件提供的对接得分作为奖励函数。MCTS 负责寻找高结合亲和力的分子，而 L-Net 用于生成新颖、类药性好且易于合成的分子。以 SARS-CoV-2 主要蛋白酶 Mpro 为例，作者使用 DeepLigBuilder 完成了针对 Mpro 已知共价抑制剂的结构优化和非共价抑制剂的从头设计任务，展示了 DeepLigBuilder 发现新分子与靶蛋白新相互作用的能力。此外，借鉴计算机视觉中广泛应用的卷积神经网络模型的思想，曾湘祥教授团队^[260]设计了 GEOM-CVAE 模型。模型主要分为三部分，第一部分是基于 3D 空间结构的分子可视化表征，即将分子空间结构编码为图像的表征方式，将分子空间坐标转换为图像的 RGB 属性后使用 CNN 进行特征提取，而后送入 VAE 模型。第二部分是基于蛋白质表面的几何特征的图表征，即将蛋白质的结构以 3D 网格的形式表示，并通过基于曲率的二次误差度量算法简化网格结构至原来的 1/5。每一次蛋白质结构的简化都伴随着切比雪夫 (Chebyshev) 图卷积的操作，每一次网格的简化也可以看作是图采样和图神经网络中的图池化。几何图卷积网络中最后两层输出的向量表征信息，作为限制条件输入到第三部分 GEOM-CVAE 解码器，从而生成具有靶点结合特性的分子 SMILES 字符串。以上方法通过不同方式引入和利用分子三维信息，提出了基于靶点结构的小分子结构生成的多种设计策略，为设计策略的发展提供了丰富的参考思路。

3.3 深度生成模型与大分子药物从头设计

随着科研人员在代谢通路、病理机制、大分子的结构和作用等分

子生物学和结构生物学的研究中取得了越来越多的进展,大分子正日渐成为攻克复杂疾病的利器。相比于小分子半衰期短、毒性较大、特异性差、专利易被突破的不足之处,大分子具有特异性强、功效高、安全性高、半衰期长、仿制壁垒高等优势,且在复杂系统疾病治疗中具有不可替代性。因而相比于成熟的小分子药物研发,大分子药物研发也正呈现出崛起之势。在 2021 年度全球十大畅销药物中,生物大分子如疫苗和单克隆抗体已占据绝大多数^[261]。学术界和产业界日渐积累的大分子数据信息为深度学习的应用提供了数据的基础。

3.3.1 基于深度学习的核酸类药物设计

随着新冠疫情的全球蔓延,mRNA 疫苗等核酸类药物因具有免疫原性强、核酸序列设计和改造的速度快等优点而日益受到关注。在 mRNA 核酸序列的各个功能模块中,5'和 3'端 UTR 序列可以影响整个 mRNA 的翻译效率和稳定性,因而成为了设计研发的重点之一。5'-UTR 序列的平均长度为 200 个碱基左右,如果随机探索所有可能的序列就会产生组合爆炸式的复杂度,加之湿实验成本高且效率低,这严重阻碍了 mRNA 疫苗的研发速度。深度学习模型因能有效捕捉和提取序列中隐含的特征,从而可以为湿实验提供预测和指导,使 mRNA 疫苗研发过程降本增效。Seelig 团队^[262]基于大规模平行报告基因检测(Massively parallel reporter assays, MPRA)技术,结合湿实验和深度学习干实验对 mRNA 的 5'-UTR 序列进行了优化探索。首先,MPRA 测得大批量序列信息及其关联特征,如 mRNA 序列表达量以及序列上核糖体的载量等,构建深度学习模型的训练数据集;其次,训练一维 CNN 模型,捕获序列的高层次特征,预测平均核糖体载量;然后利用卷积网络模型进行下游任务的设计和优化,例如,将训练好的预测模型作为适应度函数,与遗传算法突变、反传梯度及深度生成模型等方法结合来生成序列。采用了上述类似的思路,

Vaishnav 等人^[263]基于巨量平行报告基因检测 (Gigantic parallel reporter assays, GPRA) 进行顺式调控元件设计, 通过 GPRA 得到 mRNA 序列信息及其对应的表达量, 然后构建 Transformer 模型, 预测输入序列的表达量, 而后将 Transformer 模型作为适应度函数配合遗传算法进行基因表达的工程优化。但是, 预测模型结合遗传算法的序列设计模式具有一定的局限性, 例如需要实验测得海量数据来训练模型; 预测模型因有过拟合的风险而存在一定偏差; 遗传算法具有收敛速度慢、参数多、调参难、优化结果受预测模型影响大和不具备自主生成样本能力等局限。直接从生成模型入手, Linder 等人^[264]设计了深度探索网络, 具体而言是从均匀分布中随机采样出两个随机变量来生成两个位置权重矩阵 (Position weight matrix, PWM), 然后再分别根据两个 PWM 采样生成出两条 mRNA 序列, 进而通过神经网络, 如 VAE 网络, 来最小化两条序列的相似度和预测表型如核糖体载量等。Hu 等人^[265]使用强化学习模型来预测核糖体密度, 模型中的策略网络用来解释该核糖体密度分布的密码子特征, 而模型中的预测器用于预测密码子对应的核糖体密度, 因而模型可用于设计高产能的 mRNA 序列。

3.3.2 基于深度学习的蛋白和多肽设计

蛋白质和多肽设计是生成与具有所需功能的蛋白质和多肽相对应的新氨基酸序列的过程, 其主要策略是定向进化, 使用多轮随机诱变和高通量筛选来选择最有前途的序列。计算方法已作为随机诱变的替代方法, 用于提高设计序列的质量。早期的计算方法依赖于进化曲线和半经验能量函数来指导序列空间的探索, 深度学习方法因具有可以利用蛋白质和多肽的序列结构大型数据集中的信息、构建更准确地捕获蛋白质和多肽序列功能的模型等特点而备受关注。

在使用自回归模型方面, Ingraham 等人^[266]开发了 Structured

Transformer 模型，模型的编码器将表示成主链扭转角以及残基对之间的距离、相对平移和旋转值的蛋白质结构作为输入，模型的解码器迭代生成氨基酸残基。**Structured Transformer** 生成天然氨基酸的概率更高，并且它能够以比 **Rosetta**^[267] 更高的准确度恢复 NMR 蛋白质结构中的正确氨基酸。**Strokach** 等人^[268] 开发了一种图神经网络 **ProteinSolver**，其中输入节点属性和边属性分别表示氨基酸对之间的身份和距离，并且该网络通过重建恢复出被遮罩的氨基酸种类的方式进行训练。**ProteinSolver** 生成了具有所需拓扑结构的稳定蛋白质序列，正如一系列计算验证技术和表达纯化蛋白质的圆二色谱所证实的那样，相比于不利用结构信息的 **Transformer** 模型，该方法能更准确地预测蛋白质稳定性和亲和力的变化。在使用深度生成模型方面，**Eguchi** 等人^[269] 将距离矩阵作为输入，训练 VAE 生成与输入距离矩阵和相应蛋白质结构的扭转角相匹配的 3D 坐标，从而生成具有预定拓扑结构的蛋白质主链。在对约 11,000 个免疫球蛋白结构进行 VAE 训练后，生成的模型能够生成与预期键长、键角和扭转角相匹配的新型免疫球蛋白主链，并学习有意义的潜在表示，可以找到具有所需形状和特征的主链。除了 VAE，深度生成模型中著名的 GAN 也已被用于生成和改进距离矩阵，并生成具有特定折叠和功能的新型蛋白质序列。**Anand** 和 **Huang**^[270] 训练了一个 GAN 模型，使用 2D 卷积、池化和上采样层来生成与新蛋白质折叠相对应的距离矩阵，进而可以通过解凸优化目标函数或使用经过训练的模型将距离矩阵映射到坐标的方式，从距离矩阵中重建蛋白质骨架。**Repecka** 等人^[271] 在苹果酸脱氢酶（Malate dehydrogenases, MDH）序列的数据集上训练了一个采用卷积和注意力层的 GAN 模型。由模型生成的蛋白序列经过湿实验验证，约 24% 具有酶活性，具有不错的应用潜力。在结合预测和搜索优化模型方面，魏冬青教授团队^[272] 面向免疫原性肽筛选任务开发了 **TransMut** 框架。**TransMut** 框架由用于肽-人类白细胞抗原（Human

leukocyte antigen, HLA) 复合物 (pHLA) 结合预测的 TransPHLA 和一个自动优化的突变肽 (Automatic optimization of mutant peptides, AOMP) 程序组成。TransPHLA 模型的核心思想是将自注意力应用于肽、HLA 和 pHLA 对以获得结合分数, 模型由四个主要子模块组成:

(1) 嵌入块 (除了序列中氨基酸的种类编码, 还增加了位置编码来描述序列的位置信息); (2) 编码器块 (应用多个自注意力模块, 专注于序列的不同分量, 并屏蔽序列的填充位置, 防止误导模型); (3) 特征优化块 (先上升后下降的带有陀螺通道的全连接层用于处理之前的自注意力块得到的特征, 以达到更好的特征表示); (4) 投影块 (多个全连接层用于预测最终的 pHLA 结合得分)。将所提出的 TransPHLA 模型与之前的 14 种 pHLA 结合预测方法进行了比较, 包括最先进的方法、免疫表位数据库 (IEDB) 推荐的方法、九种 IEDB 基线方法和三种最近的基于注意力的方法。TransPHLA 不仅以更高的效率实现了更好的性能, 而且解决了许多 HLA 等位基因和可变长度多肽方法的局限性。当用户提供包含源肽和目标 HLA 等位基因对时, AOMP 程序可以搜索对目标 HLA 等位基因具有更高亲和力且不超过四个突变位置的突变肽。TransPHLA 和 AOMP 程序共同组成 TransMut 框架, 将 Transformer 应用于生物分子结合和突变领域。该框架可应用于任何生物分子突变任务, 例如表位优化或药物设计, 尤其适用于疫苗开发。

综上所述, 大分子药物设计的需求多种多样, 而针对其特定需求的深度学习设计方法方兴未艾, 相比于较为成熟的小分子设计拥有更大的未知探索空间。

3.4 本章小节

药物设计是极具挑战性的生物医学问题, 也是需要长产业链、高投资的研发过程。在小分子药物、核酸类药物和蛋白多肽类药物等设

计场景中，针对药物分子的结构合理性、生化性质、特定靶点亲和力等性质要求，研究者越来越多地关注、使用和迭代更新各种深度生成模型来进行从头药物设计，以期高效地发现具有启发性的小分子和大分子结构。

尽管在基于深度生成模型的从头药物设计领域已有不少相关研究工作，但许多挑战仍尚待解决：在算法模型方面，如何使模型具有泛化性，具有更好的训练收敛速度，在不同数据分布下具有可迁移性，在物理化学层面具有可解释性，与已有成熟的 CADD 理论技术的融合性；在设计要求方面，如何更好地利用三维结构信息，挖掘蛋白质三维结构以及配体-受体结合模式，更精准地进行基于靶点结构的药物设计；在效果评价方面，在计算如分子的合理性、新颖性、成药性、可合成性、理化性质分值、靶点对接打分和结构分布参数等评价分数时，如何设计更通用、更公允、更高效、物理解释性更强的基准和指标计算方式等。

总而言之，基于 AI 的从头药物设计技术正处于蓬勃发展阶段，许多算法模型、计算和实验手段的验证方法都需要更进一步的设计和提升。可以相信，基于 AI 的从头药物设计技术很快将更好地辅助药物化学研究人员设计和优化新颖的药物分子，加速药物发现的研发循环。

第 4 章 人工智能与药物重定位

4.1 药物重定位概述

新药物的研发具有投资巨大、周期漫长、风险程度高和回报效益高等特点。从已获批准或成熟的临床药物中有效识别新的适应症在药物发现中起着至关重要的作用，可以绕过开发一个治疗性药物所必须的多项批准前测试^[273]，这个过程也被称为药物重定位或再定位。

在药物重定位流程中，传统的机器学习方法可以利用人工构建的描述符更好地预测各种下游任务（例如，分子特性、药物-靶点结合亲和力和化合物-蛋白质相互作用等），为后续的临床试验确定候选药物^[274]。然而，这些方法只能处理固定大小的输入，并且大多数机器学习方法严重依赖于特征工程^[275]和领域知识。与传统的机器学习技术不同，深度学习具有映射海量数据中输入特征和输出决策的之间关系的特点，进而从输入数据中自动学习多级表示，而且不需要额外的用户输入。

本章对药物重定位方法进行总结归纳，重点介绍表示方法和深度学习模型的最新进展。首先，总结与药物重定位相关的广泛使用的数据库。其次，分别对基于序列和基于图的表示方法进行概述，重点研究了两种药物重定位的深度学习模型，即基于靶点和基于疾病的模型。最后，全面介绍药物重定位技术的几个应用。

4.2 药物重定位数据库

计算性药物重定位方法随着基因组学数据、表型数据和全息数据的急剧增长，增加了发现新候选药物的机会^[276]。现有的数据库存储了来自不同系列化合物的潜在细胞靶点。例如，KEGG 数据库^[277]，它包含了来自基因、蛋白质、生物途径和人类疾病的大规模分子数据集。DrugBank^[171]将详细的药物信息和相应的药物靶点相结合，共计

13,791 个药物条目,其中包括了 2,653 个由食品和药物监管局(FDA)批准的小分子药物。

药物重定位涉及化学、生物学分子、药物-靶点的相互作用和疾病等数据。首先,这些数据大多可以直接下载或者通过 API(应用程序接口)获得,便于整合到计算方法中。其次,可以对各种数据源或跨数据库比较分析,选择所需要的输入。例如,DrugBank 可以通过阅读文档或者检查数据统计来获得药物-靶点相互作用数据、临床药物分类、化学结构、路径和药物组合信息。此外,整合多组学数据可以为计算性药物重定位提供新的视角。

4.3 表示学习

深度学习作为机器学习的一个分支,将人工神经网络与多层非线性处理单元相结合,从原始输入中逐步提取高级特征^[147]。事实上,深度学习方法的性能在很大程度上体现在有效的数据表示上。这意味着可以让一个系统使用一套技术自动从原始数据中提取特征或发现分类所需的表示,此过程被称为表示学习,是端到端深度学习的基本步骤之一^[278]。因此,研究者们致力于将深度学习方法整合到输入数据的特征表示设计中,使其更容易提取到有效信息。目前,用于药物重定位的表示学习主要可以分为基于序列的方法和基于图的方法。

4.3.1 基于序列的表示

基于序列的表示方法可以克服部分现有的蛋白质/靶点结构数据的局限性和昂贵的分子对接模拟的需求,而现有的蛋白质和化合物序列的生物信息可以快速推进药物重定位。对于分子化合物,一个关键的一维表示是 SMILES^[279],它是一种基于化学键规则的拓扑信息的文本符号(图 4-1a)。此外,化学指纹,如圆形指纹^[280],是分子的二维表示。它通过反复搜索每个原子周围的部分结构,然后使用哈希函

数将分子转换为二进制向量（图 4-1b）。然而，这种表示方法生成的向量不仅是高维和稀疏的，还可能因为使用了哈希函数从而导致“比特碰撞”。此外，受自然语言处理（NLP）中预训练语言模型的启发，Mol2vec^[281]被提出并被认为是最具有代表性的方法，它将化合物分子的子结构视为“单词”，将化合物视为“句子”，并且使用 Word2Vec 来生成化合物的嵌入表示^[282]。虽然这些方法取得了很好的效果，但一维或二维表示的显著缺陷是缺失了关于键长和三维构象的信息，而这些信息可能对药物靶点的结合细节很重要。因此，三维表示法将在未来引起更多的关注（图 4-1c）。

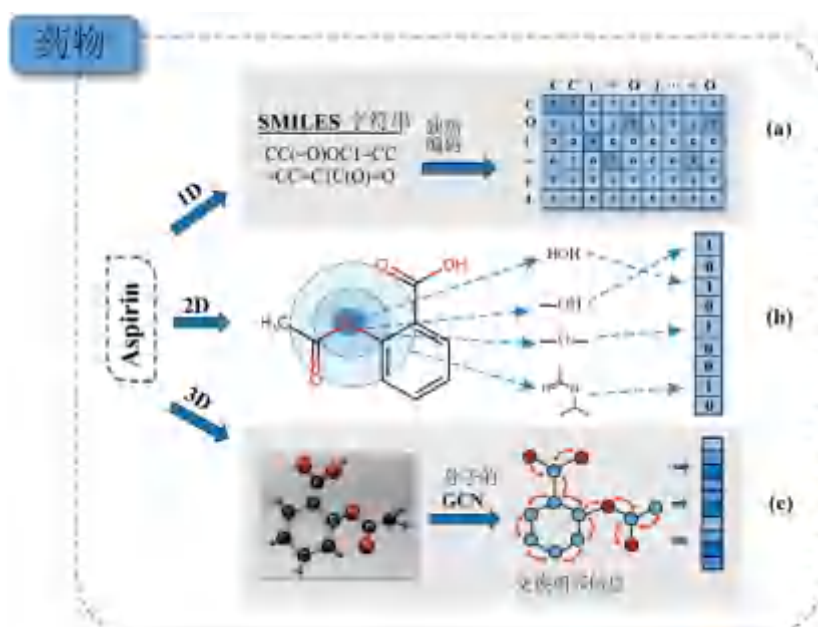


图 4-1 药物表示法

同样，蛋白质序列一般由 20 种标准氨基酸构成，其中每个氨基酸都可以通过独热编码（图 4-2a）进行简单的编码。此外，蛋白质可以用一个二维（2D）距离图（图 4-2b）来表示，它计算出三维蛋白质结构中所有可能的氨基酸残基对之间的距离。受 NLP 的嵌入技术启发，ProtVec^[283]和 doc2vec^[284]被开发出来，用于生成蛋白质序列中不

重叠的 3-gram 子序列，并通过使用 word2vec 技术对其分布式表示进行基于 skip-gram 模型的预训练。然而，这些方法通常只专注于学习与上下文无关的表示。因此，Ethan 等人设计了一个统一的表示学习方法，将 RNN 应用于从未标记的氨基酸序列中，进而学习蛋白质的统计表示。这些表示含有丰富的语义信息，同时在结构、进化和生物物理上也更合理。Strodthoff 等人^[285]提出了一个通用的深度序列模型，首先使用未标记的蛋白质序列进行预训练，然后在下游分类任务中进行微调。然而，在上述的蛋白质表示只用到了序列信息，忽略了蛋白质的物理、化学和生物特征。因此，Rifaioglu 等人提出了一种新的特征化方法，根据蛋白质的物理、化学和生物特征，将蛋白质序列表示为数字矩阵^[185]。与药物相似，基于序列的表示方法没有考虑到蛋白质的三维结构信息。谷歌 DeepMind 开发的 AlphaFold 深度学习系统^[286]已经发布了基于基因序列的蛋白质三维结构预测，其将训练时间压缩到了几天内，而传统的实验方法可能需要数月的时间。最近，DeepMind 发布了 AlphaFold2 的源代码，并且免费公开了他们预测的人类蛋白质的三维结构^[287]。

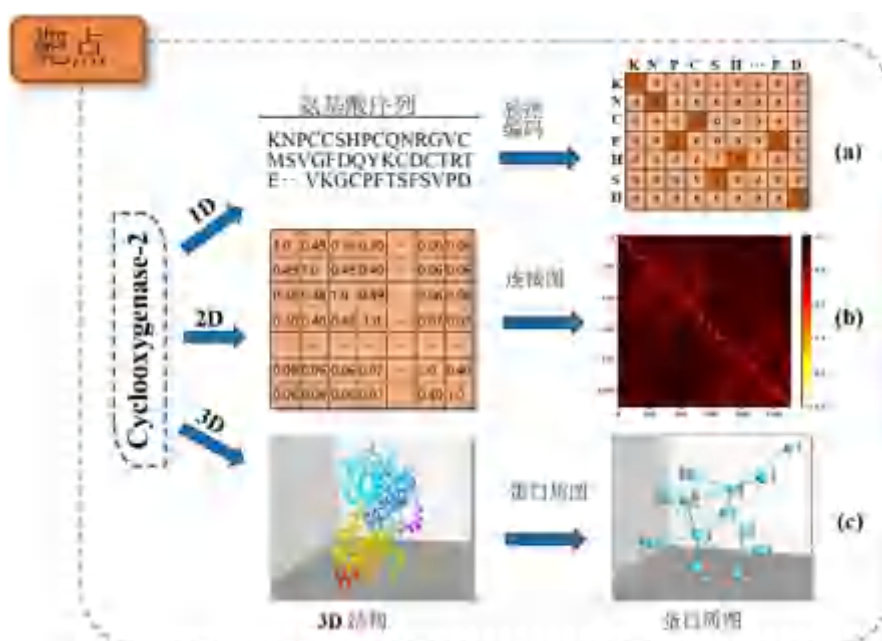


图 4-2 靶点表示法

4.3.2 基于网络/图的表示学习

最新的多组学技术和系统生物学方法生成了大规模的异质生物网络，为开发基于图或网络的药物重定位方法提供了机会^[150]。由于化合物或蛋白质(包括两者之间的化学关联)可以被编码成图或网络，基于图的表示方法逐渐成为提高药物重定位性能的一种新型解决方案。

最近，图神经网络(GNNs)已经发展为图相关任务的最先进方法，并成功应用于节点级别和图级别的分类任务中^[288, 289]。图神经网络的优势在于通过考虑相邻结点的结构和聚合各层之间的消息来自动提取特征。对于化合物而言，SMILES字符串可以通过RDKit转换为分子图，并将分子图中的原子和键分别表示为图中的顶点和边(图4-2c)。对于蛋白质而言，可以将蛋白质的各种非氢原子表示为蛋白质图的顶点，并且这种表示方法在构造上具备旋转不变性。ProteinGCN^[290]有效地利用了原子间的方向和距离，并通过图卷积的方式捕捉了局部结构信息(图4-2c)。与那些保持一阶或二阶相似性的GNNs相比，网络嵌入技术可以学习到全局特征。它通常将节点和边分别映射为向量表示，该向量最大限度地保留了图的全局属性(例如，结构信息)^[291]。一旦获得节点表示，深度学习模型就可以应用到基于网络的任务上，包括节点分类^[292]、节点聚类^[293]和链接预测^[294]。概率图也是一个重要的基于图的深度学习方法，其结合了神经生成模型、基于梯度的优化和神经推理技术。此外，在生物序列上训练的变分自动编码器(VAE)已经被证实可以学习对各种下游任务有益且具有生物学意义的表示。VAE作为自动编码器的一个变种，其在输入空间和潜在空间之间提供了一个随机映射，这个映射在训练过程中被规范化，确保其潜在空间有能力生成一些新数据。将VAE应用于蛋白质建模领域的一个例子就是学习细菌荧光素酶^[295]的表示。由此产生的连续实值表示可以用来生成*luxA*细菌荧光素酶的新型功能变体。

4.4 药物重定位的深度学习模型

药物重定位工具通常归类为“以靶点为中心”和“以疾病为中心”两类方法，用于预测未知的药物-靶点和药物-疾病的相互作用。靶向捕获策略^[296]通过编码药物的化学结构来筛选靶向蛋白质，从而提供详细的药理学解释。然而，仅预测靶点不能完全描述疾病的特征，因此，有效地识别药物和疾病之间的关联对于理解潜在的生物学机制至关重要。本节重点介绍用于药物重定位的基于靶点和基于疾病的深度学习学习方法。

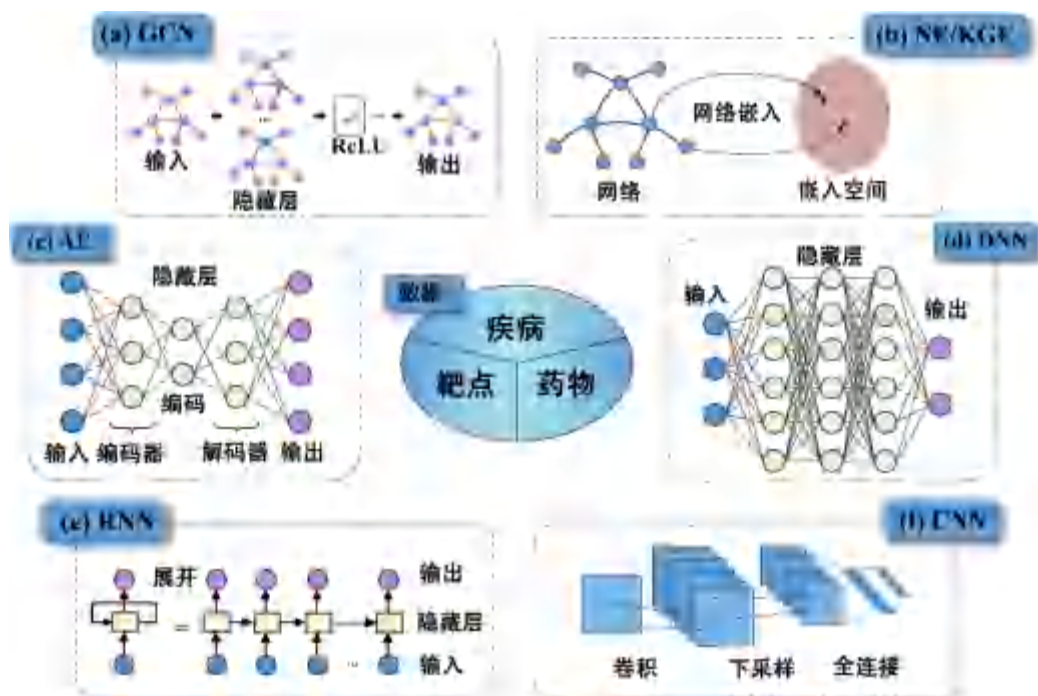


图 4-3 深度学习模型的架构: (a) 图卷积网络 (GCN); (b) 网络嵌入 (NE) 或知识图谱嵌入模型 (KGE)^[297]; (c) 自动编码器 (AE); (d) 全连接深度神经网络 (DNN); (e) 递归神经网络 (RNN); (f) 卷积神经网络 (CNN)。

4.4.1 以靶点为中心的模型

目前，许多深度学习方法被应用于发现基于分子结构的药物-靶点相互作用。卷积运算可用于处理不同长度的氨基酸序列，并捕获在

药物-靶点相互作用 (Drug target interaction, DTI) 预测^[184]中起关键作用的广义蛋白质类的局部残基模式。为了充分利用化合物-蛋白质相互作用 (CPI) 数据, 一种基于多通道 PCM 的 DNN (图 4-3d) 框架 MCPINN^[298]被用于预测 DTI。该框架主要有三个模块, 分别是特征抽取器、端到端学习器和分类器, 它将 SMILE、ECFP 和 Mol2vec^[281]嵌入的向量以及 ProtVec^[283]嵌入的氨基酸序列结合作为输入。

最近, 许多重定位方法采用分子图的形式来表示化合物。Tsubaki 等人将 GNN 和 CNN (图 4-3f) 相结合, 开发了一种端到端的 CPI 预测方法, 该方法能够将分子图和蛋白质序列转换为低维实值向量。类似地, Gao^[299]等人利用 LSTM 和 GCN (图 4-3a) 分别将蛋白质和药物结构映射到密集的嵌入空间中, 并使用双向注意机制^[300]计算药物-蛋白质的相互作用, 从而实现可解释性。然而, 基于序列的 CPI 模型^[212]仍然存在一些局限性, 例如分割方法以及隐藏的配体偏差, 会导致模型的预测性能偏高。为了解决这些局限性, 出现了一个名为 TransformerCPI 的具有自注意力机制的 transformer 架构^[301], 它使用 GCN 学习每个原子的表示, 并通过嵌入^[212]将蛋白质转换为序列的向量表示。与之前的方法相比, TransformerCPI 在更严格的标签反转实验中取得了最佳性能。

基于网络的方法也被应用于已知药物的靶点识别, 以应对药物副作用并加速药物的重新利用。例如, Luo 等人^[302]首先采用无监督的方法从异构网络中学习药物和靶点的低维向量表示, 然后采用归纳矩阵完成法预测新的 DTI。然而, 将特征学习与预测任务分离可能不是最佳解决方案。随后, 进一步提出了一种基于神经网络的 DTI 预测方法, 称为 NeoDTI^[303]。NeoDTI 集成了异构网络中节点的邻域信息并自动学习了药物和靶点的表示。然而这个方法倾向于只保留局部邻近性, 因此又引入了深度自动编码器用于从异构网络中自动学习高质量的特征, Zeng 等人^[150]使用正无标记矩阵来预测新的 DTI, 称为

deepDTnet，它集成了大型生物医学网络数据集用于靶点识别，并最大限度减小了药物开发中的转化差距。对比实验表明，**deepDTnet** 在识别已知药物的新分子靶点方面实现了高达 **0.963** 的 **AUC** 指标。此外，基于深度学习的框架 **AOPEDF** 也被应用于预测 **DTI**，该框架是一种任意顺序的近邻嵌入深度森林^[304]。它构建了 **9** 种药物网络，考虑不同网络的互补序邻近信息，并且用较少的超参数获得了更高的性能。此外，案例研究表明，**AOPEDF** 预测出的多个分子靶点与几种上市药物（如阿立哌唑、利培酮和氟哌啶醇）的药物滥用障碍的作用机制有关，并得到了实验分析的证实。消融实验进一步分析了其潜在优势，将其替换为 **LINE**^[305]（即 **LINE1st** 和 **LINE2nd**）进行特征提取，并将所设计的深度森林分类器与具有相同特征的传统方法（包括支持向量机、随机森林和深度神经网络）进行了比较，结果表明，**AROPE** 保留的高阶近邻可以为分类提供更有效的信息，而深度森林分类器的性能最好。

上述大多数研究都集中在二元分类任务上，其目的是确定药物-靶点对是否存在相互作用关系。然而事实上，药物和靶点之间存在结合亲和力值，相对二元分类任务而言，预测结合亲和力值的药物重定位回归任务更具挑战性。例如，**Karimi** 等人提出了一种名为 **DeepAffinity** 的半监督深度学习模型^[306]。**DeepAffinity** 结合了 **RNN**（图 4-3e）和 **CNN** 来编码分子表示，并使用未标记和标记数据预测亲和力。此外，**DeepAffinity** 引入了注意力机制^[300]，通过计算不同分子片段的注意力权重并选择权重最大的片段来解释预测，这可以进一步应用于预测结合位点和结合特异性来源。**GraphDTA**^[189] 也被用于预测 **DTA**（药物靶点结合亲和力），但不同的是，它使用 **GNN** 而不是 **CNN** 来学习化合物的表示。然而，在上述方法中，蛋白质的物理、化学和生物学特性通常被忽略。因此，**Rifaioğlu** 等人^[185] 提出了一种新的蛋白质特征化方法，该方法集成了多种类型的蛋白质特征，将序列、

结构、进化和理化性质等蛋白质特征转化为二维向量，并在 CPA（复合蛋白质亲和力）预测性能方面取得了显著提升。

4.4.2 以疾病为中心的模型

识别药物-疾病对之间的相互作用对于以疾病为中心的药物重定位至关重要。目前，现有的方法大致可以分为基于相似性的方法和基于网络的方法。

已有很多方法被用于计算药物和疾病之间的相似性。这些方法通过将药物或疾病特征与已知的药物-疾病关联相结合，在药物重定位方面取得了一定的成功。例如，一个名为 SNF-CVAE^[307]的方法被用于预测新的药物-疾病相互作用。它集成了相似性度量、相似性选择、相似性网络融合(SNF)和集体变分自动编码器(Collective VAE, CVAE)^[308]进行非线性分析，提高了药物-疾病相互作用预测的准确性。同时，两个案例研究显示，SNF-CVAE 预测的候选药物有可能用于治疗阿尔茨海默病和幼年类风湿性关节炎，临床试验和已发表的研究验证了这一点。此外，Xuan 等人^[309]提出了一种基于 CNN 和双向 LSTM 的药物重定位新方法，其中基于 CNN 的模块用于从药物-疾病对的相似性和关联性中学习药物-疾病对的原始表示，而基于 BiLSTM 的模块被用于学习药物-疾病的路径表征，通过注意力机制平衡不同路径的贡献。

基于网络的方法通过结合不同生物网络之间的图形信息来进行药物重定位。例如，Su 等人^[291]总结了网络嵌入（图 4-3b）方法在生物医学数据中的应用，并讨论了它们的可扩展性和局限性。此外，一种基于网络的深度学习方法 deepDR^[310]，也被应用于药物重定位。该算法首先通过多模态深度自编码器从 10 个网络中学习药物的高级特征，然后结合临床报告的药物-疾病对，将学习到的药物表示进行编码，最后通过变分自动编码器进行解码（图 4-3c），以推断出候选药

物。然而，**deepDR** 只考虑了药物的信息，而没有考虑疾病的相互作用。**Wang** 等人^[311]从大规模数据库中收集了蛋白质、药物和疾病的相互作用信息，为利用蛋白质相互作用（**PPI**）改进药物重定位评估提供了经验。他们设计了一种基于双 **GCN** 的方法来合并域间信息，通过使用异构的多关系网络（即知识图谱）来建模。在另一项研究中，**Mohamed** 等人^[312]研究了知识图谱嵌入（**Knowledge graph embedding, KGE**）模型，重点关注在各种生物任务中可扩展性和准确性表现最好的模型，并进一步讨论了使用 **KGE** 建模生物系统的机遇和挑战。

4.4.3 模型评估

药物重定位任务分为分类和回归两大类。对于回归任务，选取均方根误差（**Root mean squared error, RMSE**）、平均绝对误差（**Mean absolute error, MAE**）和一致性指数（**Consistency index, CI**）来评估模型性能。**MAE**^[189]表示预测和实际值之间的平方差的平均值，**RMSE**^[185]通过取预测和实际值之间的平方差的平均值的平方根来衡量误差的平均大小，而 **CI**^[185]衡量随机选择的两个结合亲和力值不同的化合物-靶蛋白对处于正确顺序的概率。在分类任务中，准确率（**Accuracy**）、**AUC**、精确召回曲线下面积（**Area under the precision-recall curve, AUPR**）和 **F1-score** 这些常用指标，用来评估分类器性能。准确率将总体准确性定义为预测值与真实值之间一致的概率。**AUC**^[150]是一个衡量二元分类器区分正类和负类能力的指标。然而对于高度不平衡的数据，**AUC** 在评估预测算法的性能方面可能过于乐观，而 **AUPR** 在这种情况下可以提供更好的评估。**PR** 曲线^[150]显示了在不同的决策阈值下准确度和召回率之间的权衡。**F1** 分数^[307]是测试准确性的衡量标准，它由测试的精确率和召回率计算而来，其中精确率表示被分为正例的样本中实际为正例的比例，而召回率表示的是在样本中的正例中有多少被预测正确了。最近，第一个统一框架 **Therapeutics Data**

Commons (TDC) [313]发布了，它能够系统地评估整个治疗领域的机器学习模型的性能。

除了通过计算指标来评估模型性能，还要对置信度排名靠前的候选药物名单进行实验验证或临床验证。常用的实验验证方法包括体外实验和体内动物实验。Zeng 等人通过实验验证了 deepDTnet 预测的拓扑替康（一种已批准的拓扑异构酶抑制剂）是一种新的、直接的人类维甲酸受体相关孤儿受体- γ t (ROR- γ t) 的抑制剂。随后，该团队发现拓扑替康通过特异性靶向 ROR- γ t 在多发性硬化症小鼠模型中显示出治疗潜力。一个经典的临床验证方法是使用来自健康保险索赔或电子健康记录的电子患者数据进行病例对照观察性研究，通过对 720 万个病例对照观察，研究小组发现氟替卡松（一种已批准的糖皮质激素受体激动剂）的使用与阿尔茨海默病的发病率降低显著相关[314]。通过使用超过 2.2 亿患者的大型医疗保健数据库和最先进的药物流行病学分析，Cheng 等人发现羟氯喹（一种已批准的免疫抑制药物）与降低冠状动脉疾病 (Coronary artery disease, CAD) 风险有关。此外，体外实验表明，羟氯喹可减弱人类主动脉内皮细胞中促炎细胞因子介导的激活[315]。总之，将计算预测和实验或临床验证相结合，将提供可行的策略来确定可重用的候选药物，以直接在患者身上测试。

4.5 药物重定位的应用

药物重定位已经被证明是一种有前景的药物发现和开发策略，可用于应对各种人类疾病，如罕见疾病[316]、神经退行性疾病[317]、癌症[278]和传染病。本节以 COVID-19 为例，说明药物重定位策略在对抗 COVID-19 大流行中起到的加速治疗发展的作用（图 4-4）。

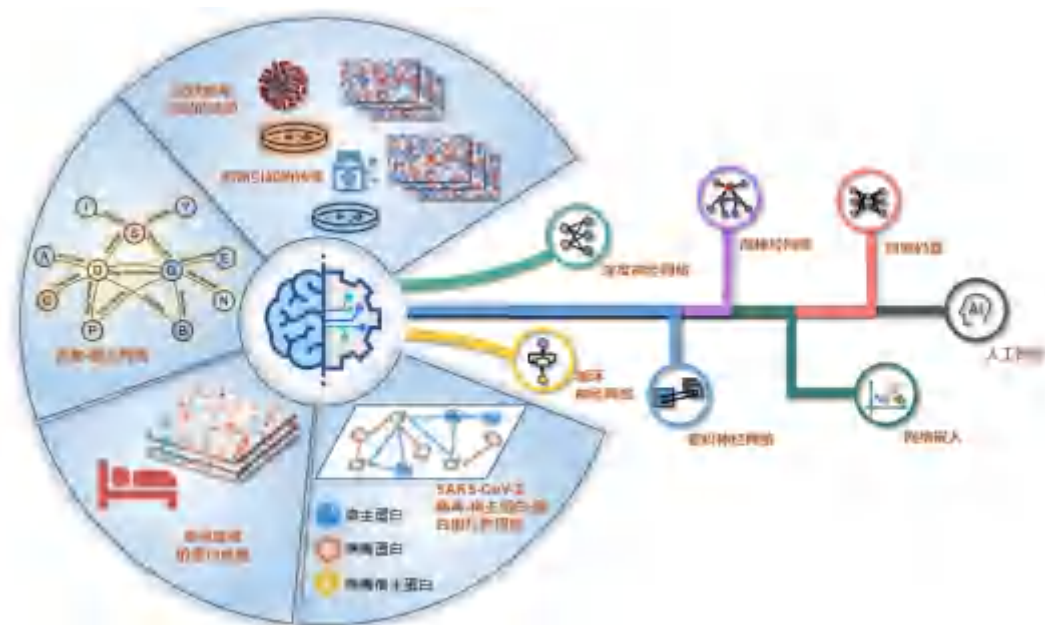


图 4-4 基于深度学习的药物重定位架构示意图，用于开发新的宿主靶向治疗以应对 COVID-19 和未来的大流行。从 SARS-CoV-2 病毒-宿主蛋白-蛋白相互作用组的角度来看，经批准的针对特定人类蛋白质/靶点的药物可能为 COVID-19 提供潜在的宿主靶向治疗，因为 COVID-19 可能与人体细胞和组织共享生理特性。

截至 2021 年 9 月 13 日，正在进行的全球 COVID-19 大流行已经导致超过 2.24 亿个确诊病例和约 400 万的死亡病例。尽管有可用的疫苗，但目前尚无有效的 COVID-19 治疗方法。针对 COVID-19 大流行，目前的紧急任务是制定有效的预防和治疗策略。为了解决这个困境，Belyaeva 等人提出了一个基于自动编码器的平台，该平台系统地集成了可用的转录组学、蛋白质组学和结构数据^[318]。作者强调了将丝氨酸/苏氨酸和酪氨酸激酶作为潜在靶点的重要性，并通过计算确定了几种候选药物（即多沙普兰、达沙替尼和利巴韦林）对患有 COVID-19 的老年人的优先次序。除了针对 COVID-19 的宿主靶向疗法，专门针对 SARS-CoV-2 病毒蛋白的抗病毒药物重定位也是一种有效的方法。例如，SARS-CoV-2 的主蛋白酶（Mpro）是最有利的药物靶点之一。一项研究整合了数学和深度学习（MathDL），提供了 137

个 Mpro 的晶体结构中候选抑制剂的结合亲和性的可靠排名^[319]。该团队使用 MathDL 确定了 71 个候选的 SARS-CoV-2 MPro 共价键抑制剂。张平团队提出了一种基于神经网络的方法 DeepCE，该方法利用图神经网络和多头注意机制^[300]来预测受化合物干扰的化学亚结构-基因和基因-基因关联。DeepCE 使用了一种数据增强方法^[320]，从 L1000 数据集的不可靠实验（即平均皮尔逊相关性得分 <0.7 ）中提取有用的信息，并且与其他先进的方法相比，DeepCE 具有更好的性能^[321]。该团队进一步将 DeepCE 应用于 COVID-19 药物重定位，并通过计算确定了一组符合 COVID-19 临床证据的候选药物^[321]。

尽管这些研究表明，深度学习在识别可用于 COVID-19 的可再用药物（包括靶向宿主疗法和抗病毒治疗）方面具有潜力，但最近的一些研究也取得了与深度学习方法相当的性能，且采用了更简单的策略。例如，用于 SARS-CoV-2 病毒-宿主相互作用组探索和药物-靶点识别的网络平台 CoVex^[322]，实现了基于网络的候选药物预测的系统医学算法，并挖掘整合了病毒-宿主-药物相互作用组，以寻找药物靶点和药物重定位的候选物。一种多模式集合预测方法^[323]将人工智能、网络传播和网络邻近性相结合，并在人类细胞中对排名靠前的药物进行了实验筛选，最终确定了四种可重用药物（地高辛、氟伐他汀、氮卓斯汀和金诺芬），这些药物对治疗 COVID-19 具有潜力。然而，这些预测都没有经过临床前和临床随机对照临床试验的验证。因此，所有预测的候选药物都必须通过实验分析和随机临床试验进行验证，才能推荐用于 COVID-19 患者。

另外，Zeng 等人开发一种基于网络的综合性深度学习方法 Cov-KGE^[324]。该团队在 PubMed 数据库的基础上，利用 CoV-KGE 构建了一个全面知识图谱，涵盖了 39 种类型 1500 万条的连接药物、疾病、蛋白质/基因、通道和表达的边，使用 COVID-19 试验数据作为验证集。CoV-KGE 已被证明在识别可用于 COVID-19

的可重用药物方面具有很高的性能，它确定了 41 种用于 COVID-19 的高置信度可再用药物（包括地塞米松^[325]和褪黑素），并通过富集分析 SARS-CoV-2 感染的人类细胞基因表达和蛋白质组数据进行了验证。随后，该团队从大型 COVID-19 注册数据库发现，褪黑激素的使用与 SARS-CoV-2 实验室检测结果呈阳性的可能性降低 28% 相关^[326]。clinicaltrials.gov 数据库中，至少有 8 项临床试验即将或正在进行中，以测试褪黑激素在 COVID-19 治疗中的临床效果。将计算策略和真实的患者数据验证相结合，将挖掘出更多有潜力的可重用候选药物^[327]。在 BenevolentAI 的知识图谱^[328]中，baricitinib 被筛选为可能治疗 COVID-19 的候选药物。针对 COVID-19 患者的几个 II 期随机双盲试验（单独使用 baricitinib 进行治疗或 baricitinib 与现有抗病毒药物联合治疗的试验）也正在进行。最近，在一项 3 期、双盲、随机和安慰剂对照试验中，baricitinib 与降低 COVID-19 住院成人患者的死亡率呈相关关系^[329]，这是深度学习方法用于 COVID-19 药物重定位开发的首个成功案例。

随着 COVID-19 患者涌入世界各地的医院，医生们正在努力寻找有效的抗病毒疗法来拯救生命。深度学习方法为快速开发有效的 COVID-19 大流行治疗干预措施提供了希望^[329]。深度学习方法可以最大限度地缩小临床前检测结果和临床结果之间的转化差距，这是快速制定针对 COVID-19 大流行的有效治疗策略的一个关键问题。从转化的角度看，深度学习工具如果得到广泛应用，也可能有助于制定其他复杂人类疾病的有效治疗策略，包括进一步的大流行和其他新出现的传染病。

4.6 本章小节

对于即将到来的大数据驱动的药物研究和药物发现，尤其是在药

物重定位方面，深度学习是一股充满希望的浪潮。与依赖显式物理方程的物理模型不同，深度学习通过设计模式识别算法来映射小分子经验观测值之间的数学关系，能更有效地处理大规模数据集。深度学习利用深度和专用的架构从原始数据中学习有用的特征。与传统的依靠领域知识手工构造分子描述符的机器学习方法相比，深度学习可以从简单的输入中自动学习并提取化学结构的特定任务表示。然而，深度学习方法的局限性在于需要大规模、高质量的数据集来进行模型训练，以及揭示预测背后的生物学意义的可解释性。虽然传统的机器学习方法可以在某些领域很好地解决特定任务，但随着数据的爆炸性增长和 **AlphaFold** 的成功落地，有理由相信深度学习将在不久的将来为药物重定位带来里程碑式的发展。

第 5 章 人工智能与药物属性预测

5.1 人工智能与药物属性预测概述

传统的药物研发是基于化学家或药剂师的经验，依靠模拟和实验完成的，但这样会产生高达 26 亿美元的研发成本与长达 14 年的上市时间。不仅如此，在通过传统技术发现的药物分子中，90% 以上都未能在人体临床实验中取得成功^[330]。随着药物属性数据的指数级增长以及人工智能的快速发展，许多制药公司开始使用人工智能算法对候选药物进行筛选与测试，这可以大大降低研发成本。将人工智能技术应用于药物研发领域，可以较为准确地预测特定药物对哪些症状具有较好的疗效与较高的安全性，从而减少研发时间、缩减投入成本、提升研发效率、还可以充分利用现有医疗资源^[331]。

在与药物相关的应用场景中，有两个问题具有实际应用价值与现实意义：第一个问题是预测药物分子的性质，对于给定药物分子，通过分析其性质，如水溶性、似药性或者与特殊蛋白的亲合性，可以极大降低相关测定的投入；第二个问题是基于特定属性的药物分子优化，在药物设计中，对于指定属性的化合物的筛选仅局限于已知的数据集中，通过条件生成模型生成潜在的具有指定属性的药物分子，可以加速药物研发过程。

准确预测药物属性不仅可以帮助确定药物的功能，还可以应用于药物设计中的药物属性定向优化中。药物属性预测模型包含几个重要的组成部分：高质量的数据集、适当的分子表示、强大的学习算法和严格的性能评估指标。人工智能的优点在于不需要依赖专家定义的药物特征集，只需要提供可以改善特定任务的学习表征^[332]。将这些表征输入到深度神经网络中捕获化学结构和生物活性之间的相关性，这样得到的模型优于传统的定量构效关系（Quantitative structure-activity relationship, QSAR）方法。

在药物属性预测模型中，虚拟筛选目前已广泛应用于预测生物活性、生物分布以及药物的物理特性^[310]。如果虚拟筛选能够成功应用于药物研发过程中，那么这些虚拟筛选方法就可以通过减少实验筛选的时间和费用并扩大可探索化学空间的方式来加速药物发现过程。目前的实验筛选方法只能获得数百万个分子，而虚拟筛选能够在短时间内对数十亿个分子进行评估。在虚拟筛选的过程中，人工智能算法用于学习特定的分子亚结构和目标特性之间的关联，这类似于药物化学家分析分子的方法。研究者们使用带有属性的药物分子训练时，这些算法可以系统地捕捉到人类难以掌握的复杂模式。这类基于数据的人工智能计算方法不仅可以准确预测药物属性，还可以提升开发和发现药物的速度。

针对药物属性预测在不同的场景具有不同的含义，需要预测的属性也各不相同，比如量子力学属性、物理化学属性、生物物理学属性和生物效应类属性等。量子力学属性包括原子坐标、能量以及电荷等属性；物理化学属性包括水溶性、极性表面积、生物利用度、辛醇溶解度、代谢稳定性、沸点和熔点、疏水性、溶剂化自由能、被动膜通透性和血脑通透性这 10 个属性，其中水溶性是常用属性；生物物理学属性包括亲和力、功效和活性；生物效应类属性包括副作用、毒性和药物代谢动力学 (Absorption、Distribution、Metabolism、Excretion、Toxicity, ADMET)。在上述属性预测任务中，有的属于分类任务，即 0 与 1 的二分类，如毒性预测任务；有的属于回归任务，即预测属性的数值，如定量评估类药性任务。

多肽类药物作为近年来新兴的一种药物，具有较低的免疫原性、较高的生物活性、较好的安全性和组织渗透性、较小的药用剂量、较难蓄积在组织中、较方便合成和改造等优点，在治疗心血管和免疫等方面疾病、对肿瘤和细菌的抑制等方面都具有显著的疗效^[333]。值得注意的是，由于多肽类药物可以诱导肿瘤细胞出现凋亡、抑制肿瘤细

胞生长、增强细胞免疫力和抗肿瘤效应，在未来该类药物可能作为一种高效低毒类药物造福人类^[334,335]。因此，预测多肽类药物的属性具有重要的医学研究价值，也逐渐成为领域内广泛研究的课题。

5.2 多肽药物属性预测

多肽通常是一段肽链（肽链长度一般不超过 50），其大小通常大于小分子化合物（分子量在 500 以内），但小于蛋白药物（分子量在 5000 以上），它作为信号分子广泛存在于生物体细胞的各个角落，并参与机体内众多生理反应。多肽的获取方式一般包含三种：从动物、植物或微生物体内获取；酶解蛋白质；人工合成。动物、植物或微生物体内的天然多肽可以直接提取，比如，在海洋环境中，各种各样的生物体蕴含丰富的潜在成药物物质来源。海洋环境中各种多肽的研究以及临床试验结果表明，一部分海洋衍生多肽有助于人类癌症的治疗^[336-339]。

通过特定的酶水解蛋白质，可得到各种各样的多肽物质，其中一部分被确定对细菌的生长具有抑制作用，还可以帮助治疗肿瘤。比如，有研究表明鱼副产物中的蛋白质通过此类方法获得的多肽对于氧化、细菌和肿瘤均具有抑制作用^[340,341]。除了上述两种多肽获取方式以外，随着科学技术的发展、研究人员对肽类结构的深入挖掘，人工合成技术作为一类新的多肽获取方法得到了广泛关注^[342]。

多肽在人类的健康中起着至关重要的作用，它可以作为生长因子、激素、神经递质和抗感染剂等^[343]，不同大小和功能来源的多肽已被广泛用作治疗药物。例如，抗癌肽由于其抗癌活性而被用于癌症治疗^[344]；抗炎肽最近被用作抗炎剂治疗阿尔茨海默氏病和类风湿性关节炎等^[345]各种炎性疾病；细胞穿膜肽被证明是将药物递送到细胞中的转载体^[346]。与传统的基于蛋白质的生物药物相比，多肽类药物具有较低的生产复杂性以及较低的合成成本。因此，发掘多肽潜在的免

疫特性对于发现新颖且有效的治疗肽具有重要意义。

然而，研发多肽药物及发现多肽先导化合物是一项风险高、耗资大的工程。发现和设计活性多肽是多肽药物研发的关键步骤。传统的方法是利用湿实验处理多肽或改造已有多肽的结构，进而观测多肽的药理活性以获得活性多肽^[334, 347]。这类实验方法过程复杂、耗时耗力、无法预测候选多肽的相关理化属性、且无法提取潜在的多肽与活性之间的关联关系。不同于上述传统的实验方法，研究人员基于多肽数据，设计计算方法从数据中提取有效信息以构建多肽数据与活性之间的关系预测模型，找出传统实验方法中难以发现的、潜在的多肽与活性之间关系的规律，深层次地分析多肽及其抗菌活性^[339, 347, 348]。

5.2.1 多肽属性预测方法

测序技术的发展产生了海量的蛋白质数据，但是大部分蛋白质序列的功能还未被实验测定。这意味着在如此大规模的蛋白质序列中，可能蕴含有大量具有治疗属性的功能多肽未被发现。该问题的关键是如何在海量的数据中快速、准确地对这些序列进行精准识别与属性分析。针对这一关键问题，目前的解决方法可以归纳为以下四类：基于序列比对的方法、基于模糊逻辑模型的方法、基于语言生成模型的方法以及基于机器学习的方法^[349]。

5.2.1.1 基于序列比对的方法

作为一种用于确定待测序列和数据库中已知序列的相似性的序列分析方法，序列比对首先按照一定的规律将待测序列和已知序列进行排列，然后比对待测序列和已知序列得到分析结果^[350]。通过待测序列和已知序列之间的序列比对，可以确定序列之间的相似性，基于序列之间的相似性还可以进行序列之间的同源性分析。比如，Wang 等人^[351]将序列比对方法与特征选择方法相结合，设计了一种新的抗菌

肽预测方法。在序列比对部分，该方法首先假设有一个查询肽和训练集（一组抗菌肽），利用 BLASTP（Protein basic local alignment search tool）工具^[352]计算查询肽和训练集中每个肽之间的高比值片段对（High-scoring segment pairs, HSP）得分，若训练集某一个肽与查询肽之间的得分最高，则该训练集的肽与查询肽类别一致。

Ng 等人^[353]采用与上述序列比对类似的方法建立了一种预测对免疫系统至关重要的抗菌肽的新方法。该方法利用 BLASTP 建立序列比对方法，计算测试序列和所有训练序列之间的 HSP 评分。最终测试序列的分类取决于训练序列中 HSP 得分最高的类别。为了挖掘出鸡中潜在的新型抗菌肽，Xiao 等人^[354]利用多序列比对算法 ClustalW，计算具有和不具有最后一个外显子序列的所有已知 cathelicidin 前体之间的氨基酸差异比例，然后采用邻接法^[344]构建系统进化树。最终该方法成功鉴别出三种鸡类 cathelicidin 抗菌肽，一系列功能分析也表明，这些抗菌肽是迄今为止发现的最有效的抗菌肽之一，具有强大的抗菌性和脂多糖中和活性。

基于序列比对的预测方法虽然容易供研究人员使用，但是还没有一个好的方法能够确定使用的相似性度量是否合理，而且预测与已知某类特性的多肽之间存在较大差异的待测序列的类别，基于序列比对的方法精度还有待提升。此外，基于序列比对的预测方法无法预测与训练序列匹配度极低的待测序列。

5.2.1.2 基于模糊逻辑模型的方法

通过定义模糊集合和规则库，模糊逻辑模型根据需要将因变量作为独立变量的一个函数，从而对因变量进行预测^[349]。比如，Mikut 和 Hilpert 开发了一种基于模糊的技术用于抗菌肽数据集的可视化和基于规则的描述。它提供了一种利用分子描述符中的模糊项来定量描述氨基酸功能特征的系统方法。最终该方法应用于来自高通量筛选实验

的 1,609 个肽的数据集中识别抗菌肽^[355]。基于多肽的一些物理化学属性与其免疫性之间存在着模糊模式，Fernandes 等人提出了序列相似性和物理化学搜索方法，然后通过模糊推理系统寻找最适合每个结构域的抗菌肽^[356]。

基于模糊逻辑模型的方法中模糊规则的设计和定义较为简单，无需建立复杂精准的数学模型。但大多数情况下，研究人员需要凭借自身经验建立隶属度函数，共性函数的总结较为艰难，往往需要针对不同类型的多肽样本构建不同的模糊逻辑模型，因此泛化性较弱^[349]。

5.2.1.3 基于语言生成模型的方法

在分子生物学中，组成多肽的每一种氨基酸分别用不同的字母表示。因此，可以用一个按一定规律排列的字母字符串来表示一条多肽序列。Loose 等人^[357]将天然抗菌肽的氨基酸作为一种正式的语言，使用 TEIRESIAS 算法^[358]从一组天然抗菌肽序列中建立规则语法，并利用这套语法创造了新的、非天然的抗菌肽序列。此外，Loose 等人在对抗微生物肽数据库 APD^[359]中的数据进行收集分析过程中，发现了 684 个与抗菌肽活性有关的语法规则，对这些语法规则进行整合、建模有助于发现新的抗菌肽。Zuo 等人^[360]为了开发一种预测抗菌肽防御素蛋白的计算方法，首次引入了由蛋白质区块 (Protein Blocks) 方法^[361]得到的简化氨基酸字母表 (Reduced amino acid alphabet, RAAA) 的新方案来描述不同防御素家族和亚家族的长片段，进而利用多样性增量 (Increment of diversity, ID) 结合还原氨基酸字母表 (ID_RAAA) 的 n 肽组成来预测抗菌肽的四个防御素科 (脊椎动物、植物、昆虫以及其他) 和三个主要的脊椎动物亚科 (α -防御素、 β -防御素和 θ -防御素)。

基于语言模型的方法将抗菌肽序列看作由不同字符组成的句子，从中学习并建立有效的语法规则，进而用于下游预测任务。由于基于

语言模型的方法对训练数据中的现有语义模式非常依赖，导致在与训练数据中语义模式不一致的新样本上预测能力较弱，难以发现和识别出包含新的语义模式的抗菌肽。

5.2.1.4 基于机器学习的方法

机器学习方法采用推理、归纳以及模型拟合等方法在已知数据上进行知识学习，找出数据中的潜在规律，并运用、扩展到未知数据。针对多肽活性与生物化学属性之间的线性关系和非线性关系的挖掘，机器学习方法均适用，且机器学习方法还可以解决复杂且缺乏一般性理论的多肽识别和预测问题。目前，大多数研究将多肽的预测看作分类问题，采用的是有监督机器学习模型。基于已知样本数据中的有效信息的挖掘，该类机器学习模型构建了分类器预测多肽类型，推断待测样本属于每种类别的可能性，从而实现多肽的类别预测。

目前，主要采用基于序列水平的预测算法来解决计算性多肽识别问题。Torrent 等人^[362]设计了一个基于抗菌肽理化属性的人工神经网络方法，该方法不仅能够识别活性肽，还可以评估其抗菌能力。结果表明该方法能够将序列衍生的抗菌肽理化属性与抗菌活性联系起来。Holton 等人^[363]使用 N-to-1 神经网络构建了一个细胞穿透肽预测模型，该模型输入为可变长度 N 的肽序列（N 在 5 到 30 之间），输出为输入肽序列被细胞穿透的可能性（即是否属于细胞穿透肽）。

5.2.2 研究难点

尽管目前已有的工作已经取得了一些可喜的进展，但仍然存在一些挑战。首先，与蛋白质不同，多肽序列通常很短。对于这样的多肽，可以使用的背景信息不多。因此，难以捕获有效且具有判别性的多肽特征去区分不同的多肽。尽管目前已经有一些特征表示方法从不同角度捕捉多肽的特异性，例如二级结构信息、初级序列信息、谱图信息

和理化信息等，但如何将不同类型的信息用于特征表示是一大难点。其次，目前通过实验标注的多肽样本数量还比较有限，特别是一些多肽类型，例如抗病毒肽、抗癌肽以及细胞穿膜肽等。因此如何更加高效的对多肽样本进行标注，以及如何对小样本多肽建立鲁棒的计算模型是另一大难点。

5.3 药物属性预测最新研究进展

5.3.1 基于元学习的多肽药物生物活性预测

识别各种生物活性肽的工作取得了很大进展，但仍然存在以下不足：（1）许多基于机器学习的方法都受到样本数较少的影响，低容量标记样本（经实验验证的生物活性肽）无法使使用监督学习训练的模型得到较高的鲁棒性，这会导致模型存在过拟合和泛化能力差的问题。

（2）大多数现有的利用工程特性的方法都是针对特定的功能肽设计的，没有通用的计算方法可以同时准确预测不同肽的生物活性。更重要的是，它们不能挖掘新的肽生物活性，这也是监督学习的一个限制。

（3）大多数基于机器学习的多肽预测器仍然使用人工设计的统计特征进行构建，这在很大程度上依赖于研究人员的先验知识。此外，人工设计的特征无法捕获不同肽功能的高潜在非线性信息，且缺乏对不同肽功能预测任务的适应性。也就是说，他们可能在一项特定任务上表现良好，但在其他任务上表现不佳。（4）特征工程通常会产生数百维的特征向量，这将导致维数灾难。

针对上述问题，He 等人提出了 MIMML 方法。MIMML^[364]是一个通过联合优化最大化互信息与最小化交叉熵以改进现有的元学习算法原型网络（ProtoNet）^[365]的模型（图 5-1）。它的优点如下：（1）MIMML 专门为生物活性肽的挖掘和预测而设计，能够统一预测多种生物活性肽；（2）它基于嵌入技术而不基于特征工程和人工设计的特征，能够通过微调或推断等方式来预测多肽是否具有某种特定的功能

活性；(3) 嵌入使用的主干是 TextCNN^[366]，即在进行元学习前先在所有基类上进行监督预训练；(4) 使用来自各种功能肽的少量样本，通过元学习获取了各种功能之间的区别信息并表征了功能差异，能够在下游功能肽预测任务中表现良好，尤其是在小样本情况下，与传统方法相比性能有显著提升。MIMML 为生物序列分析中的少数样本学习问题提供了同类解决方案，有利于新功能肽的发现。

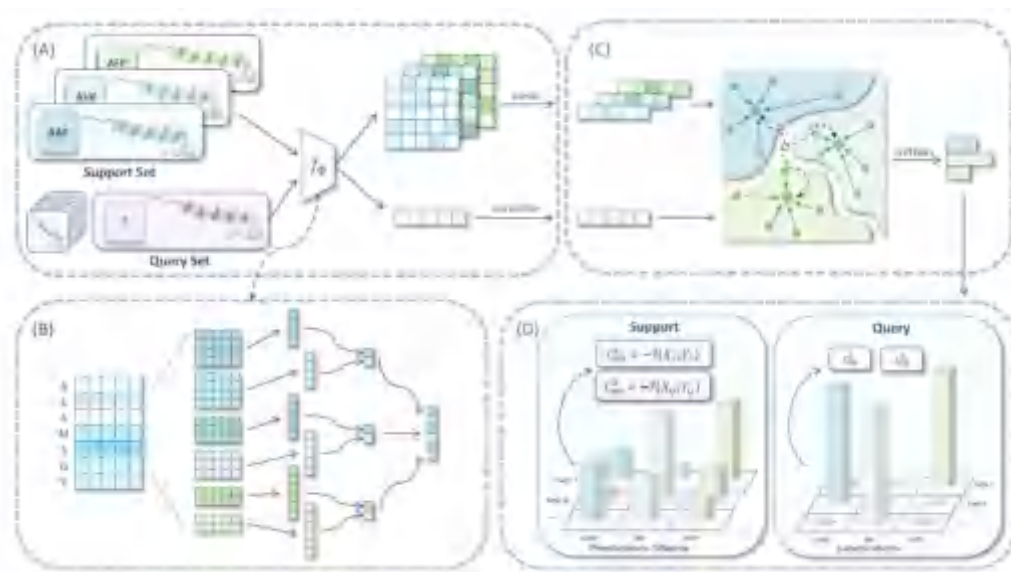


图 5-1 MIMML 的框架图。(A) 嵌入模块，(B) 基于多通道卷积的 TextCNN 作为特征提取器，(C) 原型模块，对支持集的嵌入进行平均，(D) 优化模块，根据交叉熵损失和互信息损失进行优化。

5.3.2 基于图神经网络的多肽毒性预测

相比于小分子药物，多肽大多是通过水解或肾过滤来被人体清除，氨基酸为其水解产物，因此多肽类药物的代谢产物毒性很低，且多肽类药物通常是特异性设计，其模板为内源性多肽，一般具有较高的靶标亲和力。与大分子蛋白药物相比，化学合成多肽的技术发展已经趋于稳定，合成过程中多肽与杂质或副产物容易分离，在质量、纯度、产量和成本等方面均具有优势。因此，这些特性使得多肽成为一种很

有前途的治疗药物。自胰岛素问世以来，已经有 80 多种多肽药物进入市场用于各种各样的疾病的治疗，比如糖尿病、骨质疏松症、慢性疼痛以及多发性硬化症等^[367]。

大部分多肽对人体是没有危害的，但自然界中也存在许多有毒性的多肽，例如多数蛇毒和河豚毒素。在合成用于肿瘤免疫治疗的多肽疫苗时，由于制剂也属于多肽，因此在输入人体之前，必须判断合成得到的多肽是否具有毒性，以免引起严重的医疗事故。因此多肽的毒性研究是多肽药物开发中的一个重要问题。鉴定多肽毒性最直接有效的方法是在实验室中进行生物实验，但这是一种昂贵、劳动强度大、耗时长的方法。随着潜在的治疗性多肽的数量大量增长，通过实验筛选毒性的成本和时间花费越来越高昂，如何快速准确地鉴定多肽毒性成为了一个巨大的挑战。此外，从动物试验中获得的结果对人类毒性反应几乎没有指导作用^[368]。

为了解决该问题，相关的计算方法被开发用来识别可能的毒性肽。多肽毒性预测是一个二分类问题，已知的有毒肽作为正样本，其他多肽作为负样本。现有的多肽毒性预测方法基于模型构建的主要思想可以分为两类：基于相似性的多肽毒性预测方法和基于机器学习的多肽毒性预测方法。基于相似性的方法使用比对搜索工具来衡量序列对的局部和全局序列相似性，例如 BLAST^[369]，或者从同源序列中推断出序列毒性^[370]。然而，这类方法存在一些局限性：（1）潜在的制药多肽需要有同源的毒性肽；（2）在处理大量数据时，这类方法的性能会大幅下降；（3）使用者需要提前设置 e 值的下限，并且计算序列相似性的算法选取也是任意的，没有固定的标准，这可能会影响预测性能。

与基于相似性的方法不同，基于机器学习的方法侧重于使用正样本和负样本来捕获与毒性相关的信息，以此来预测多肽的毒性。例如，ClanTox 使用从原始序列获得的 545 维特征作为输入，训练基于树分类器的预测器来分析动物毒素^[371]。ToxinPred 利用支持向量机和提取

的多肽序列的各种统计特征来区分有毒和无毒的多肽^[372]。这类方法需要研究者在特征工程中花费大量的时间，还需要专业知识作为支撑，因此存在一定的局限性。

与传统机器学习方法相比，深度学习方法能够自动提取多肽的序列特征而不需要先验知识。比如，Wei 等人基于多肽的结构信息和进化信息，使用图神经网络以及注意力机制来预测肽的毒性^[373]，其框架图如图 5-2 所示。Pan 等人利用序列信息和蛋白质领域的知识，有效地对具有不同长度的有毒和无毒蛋白质进行分类。然而，这种方法不是专门为肽设计的，需要搜索给定蛋白质的蛋白质域的嵌入^[374]。Wei 等人提出一种既能预测多肽毒性也能预测蛋白质毒性的方法，该方法利用信息瓶颈理论以及迁移学习方法，将毒性蛋白质中有效信息迁移到多肽中，即解决了毒性肽样本小的问题，也提高了预测性能^[375]。

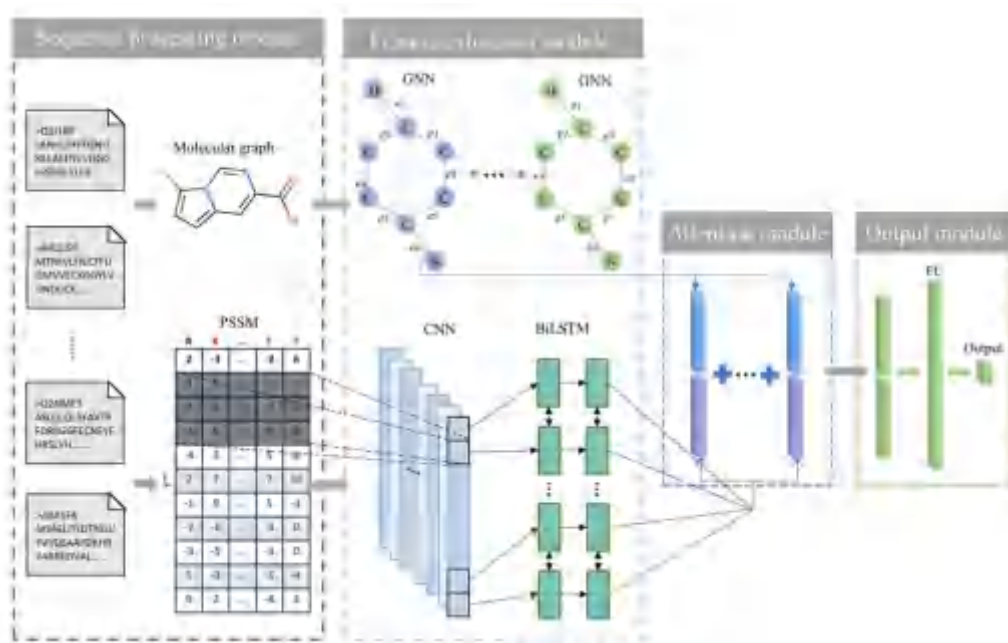


图 5-2 ATSE 的流程图。ATSE 方法包括四个模块：(1) 序列处理模块，可以从初级肽序列生成分子图和 PSSM。(2) 特征提取模块，分别从分子图和 PSSM 中提取特征。(3) 注意力模块，将特征提取模块的输出作为输入。(4) 输出模块，将得到的特征表示输入以生成预测的毒性概率。

5.4 本章小节

自 20 世纪 60 年代人们将 QSAR 模型引入新药研发以来，以药物发现为目的的计算预测模型便得到了广泛的应用。如今常见的药物发现操作都依赖于药物属性预测模型来指导药物分子的搜索和属性优化。人工智能作为一项强大而高效的新兴技术，已经广泛应用于药物属性预测中。许多先进的人工智能技术，比如图神经网络、多视图学习、对比学习和元学习等，都被第一时间用在了药物属性预测领域。与此同时，多肽类药物这种近年来展现出巨大潜力的药物，由于具有易合成、免疫原性低和安全性高等多种传统药物所不具备的优点，开始逐渐被人们重视起来。因此，对于多肽药物的属性预测模型有着良好的应用前景。相信在不远的将来，使用药物属性预测模型辅助发现的廉价且有效的多肽药物能够切实促进人类健康事业的发展。

第 6 章 人工智能与药物相互作用预测

6.1 人工智能与药物相互作用预测概述

对人类来说，药物是预防、治疗及诊断疾病的必不可少的物质。随着现代药学、药理学和药物治疗学研究的飞速发展，人们找到了治疗各种各样疾病的多种药物。药物相互作用（DDI）一直以来被描述为一种药物的药效因另一种药物的存在而发生的变化。随着获得批准的药物数量迅速增长，通过使用联合药物处方来治疗患者疾病已成为一种普遍现象。然而，多种药物的共同使用会引起重大的药物相互作用潜在风险，增加了治疗管理的复杂性。根据已有研究显示^[376]，DDI 可以引起药理学（Pharmacy, PC）、药代动力学（Pharmacokinetics, PK）和药效学（Pharmacodynamics, PD）的不同反应类型：PC DDI 的发生是由物理或化学的不相容性导致的^[377,378]；药物与药物的相互影响在于吸收、分布、代谢和排泄（ADME）过程，即会发生 PK DDI^[379]；当一种药物因另一种同浓度的药物而产生药效学反应时，会出现 PD DDI^[380]。因药物相互作用的不良影响的存在，一些上市的新药相继被召回，至今已有多个先例（如因 DDI 引起横纹肌溶解的降脂药“西立伐他汀”、引起尖端扭转型室速的胃药“西沙比利”等）。因此，药物相互作用的发现极大地提高了现代医疗的质量、并促进了安全有效的联合用药策略。

近年来，人工智能（AI）已经在药物相互作用的检测和分析中得到了广泛的应用，通过设计算法对药物相互作用数据进行建模、分析和解释，以构建自动化的智能药物开发系统。大量与药物相关的信息和数据（例如，生物医学文本、电子健康记录、事后系统和公共数据库）的可用性为人工智能计算模型的发展提供了坚实基础。机器学习方法作为一种数据驱动的人工智能技术，其相关应用在药物开发领域已经展现出广阔的前景。相较于依赖复杂的物理和化学专业知识的传

统方法，机器学习方法聚焦于海量药物数据，通过对数据进行学习，从而识别药物数据中的潜在模式和丰富知识。随着计算机算力的迅速提升，以人工神经网络为代表的深度学习模型因其强大的特征提取能力，在药物相互作用检测领域取得了巨大成功。与依赖于人工提取特征的传统机器学习方法不同，深度学习方法能够通过多层非线性特征变换获得数据中更高层次的抽象特征。此外，相较于传统的机器学习方法，深度学习方法表现出了更优越的性能，在药物相互作用检测等领域取得了更令人满意的结果，该领域运用的深度学习方法主要有深度神经网络（DNN）、卷积神经网络（CNN）、循环神经网络（RNN）、生成对抗网络（GAN）和图神经网络（GNN）等。本章将详细阐述现有的基于人工智能的药物相互作用检测方法及其应用前景。

6.2 人工智能与药物相互作用预测方法

近年来，人工智能方法在药物相互作用检测领域得到了广泛的应用，而丰富的药物数据是人工智能检测方法的基石。根据药物数据源的不同，现有基于人工智能的药物相互作用检测方法主要分为基于文献数据的提取方法和基于药物关联数据的预测方法。

6.2.1 基于文献数据的提取方法

随着呈指数级增长的生物学文献的出现，大量关于药物的知识隐藏在这些非结构化文献中，例如已发表的文章、科学期刊、书籍和技术报告。生物学文献中包含大量有价值的关于药物相互作用的描述信息，但是计算机很难高效处理上述非结构化的文本信息。自然语言处理（NLP）作为一种重要的技术，将人类语言转换为计算机易于操作的表示形式。因此，基于文献数据的提取方法主要使用自然语言处理技术从生物学文献中提取药物-药物相互作用。该类型方法旨在从文本信息中检测 DDI，其目标是通过分析药物实体对提及的文档

来识别实体对之间的特定关系。2011 年和 2013 年 DDIExtraction 挑战^[381,382]引起了广泛关注，他们为基于文献数据的 DDI 提取任务提供了带标注的语料库。因此，很多研究者们参与了这一挑战任务，大多数所研究的方法是从文本信息中提取有用的特征来检测药物相互作用并进行相关类别的分类。

一般来说，根据使用技术的不同，我们大致将这些方法分为两种类型：基于深度学习的方法和传统分类器的方法。传统分类器的方法通常设计一个特征提取器来将原始数据转换为合适的表示或特征向量，大多使用线性核或非线性核的支持向量机 (SVM) 来进行分类。基于深度学习的方法则是通过神经网络模型自动学习数据的特征表示并完成 DDI 的分类。基于深度学习的方法通常采用 NLP 技术，使用卷积神经网络、循环神经网络等方法来检测 DDI。

6.2.1.1 数据来源

基于文献的提取方法中使用的数据来源主要是公开的文献语料库。例如，DDIExtraction 2013 语料库是整合了 DrugBank 和 MEDLINE 数据库的数据集，由 792 个 DrugBank 文档和 233 个 MEDLINE 摘要组成。DrugBank 提供了广泛的与生化和药理学信息相关的文档。MEDLINE 是一个国际综合性的生物医学信息书目数据库，全部数据由美国国家医学图书馆制作。目前，很多方法都是基于 DDIExtraction 2013 语料库来提取和分类 DDI。该语料库将 DDI 分类为四种不同类型：**mechanism**（药代动力学机制）、**effect**（药效学机制）、**advice**（建议不同时服用这两种药物）和 **int**（无进一步信息）。

6.2.1.2 基于传统分类器的 DDI 提取方法

此类方法通常包括两个阶段：DDI 的识别和分类。在 DDI 识别阶段，他们从信息丰富的文献句子中提取有用的特征，以识别具有相

相互作用的药物对。在分类阶段，这些已识别的 DDI 被分为四种类型（即 mechanism、effect、advice、int）。

基于传统分类器的方法通常使用 SVM 对 DDI 进行分类。这些基于 SVM 的模型依靠精心设计的特征或内核，利用两个向量之间的相似函数的替换点积^[383]，并且可以根据内核的不同分为：基于非线性内核和基于线性内核的方法。Chowdhury 等人^[384]利用具有非线性内核的 SVM 在 DDIExtraction 2013 挑战赛中实现了最佳性能。Kim 等人^[385]使用词汇和句法特征推广线性内核模型，线性内核提供了强大的基线性能，适用于大规模数据集。

6.2.1.3 基于深度学习的 DDI 提取方法

不同于传统特征工程和冗余的特征处理方法，深度学习方法能够从数据中学习特征，此类方法主要使用卷积神经网络 CNN 和循环神经网络 RNN 及其变体模型来进行 DDI 提取。

卷积神经网络使用卷积核从局部相关数据中提取合适的特征，并通过多个堆叠层之间的非线性映射为 DDI 分类器生成特征集^[386]。基于 CNN 的 DDI 提取方法通常包含五个步骤：预处理、嵌入、卷积、池化和分类。Liu 等人^[164]提出了一种基于 CNN 的 DDI 提取方法，给定预处理后的生物学句子，该模型生成词嵌入并编码词与两种药物在句子中的相对距离以生成两个位置嵌入。然后将词嵌入与位置嵌入输入 CNN 来进行预测相互作用关系。Dewi 等人^[387]提出的 DeepCNN 模型是对 10 层 CNN 的 DCNN^[388]模型的扩展，相对于 DCNN，其改进的部分是多通道词嵌入，可以扩大词汇量并减少未知词的数量，能够提升模型的性能。

循环神经网络是一种输入为序列数据，并在序列的演进方向进行递归，且将所有节点（循环单元）按链式连接的递归神经网络^[389]。Kavuluru 等人^[390]提出了一种字符级循环神经网络（char-RNN）来完

成 DDI 提取任务。常规的词级 RNN 的输入是预训练的词向量和位置向量生成的词嵌入,而字符级 RNN 的输入是在字符嵌入上使用 LSTM 生成的词向量。通常,并非所有单词对特定句子的贡献都相同,注意力机制允许模型学习不同模态之间的特点,这在机器翻译^[391]、对话系统^[392]和关系分类^[393,394]等 NLP 任务中引起了广泛关注。在 DDI 提取任务中,单词与候选药物对的相关性可以反映单词对句子中药物对之间相互作用的贡献。Zheng 等人^[395]提出了一种具有注意力机制的双向 RNN 模型,以捕获全局语义表示。

CNN 可以使用具有不同大小的卷积核来学习特征并取得了令人满意的结果,而 RNN 在学习句子序列特征方面表现良好^[396]。因此,一些方法考虑将 CNN 和 RNN 结合起来,利用它们的优势来提高性能。Shen 等人^[397]提出 Drug2vec 来学习表示,该方法使用 CNN 来捕获药理学特征和药物类别特征,同时使用 Bi-LSTM 学习文本描述特征的表示。此外,Wu 等人^[398]提出了一种基于 GRU-CNN 的模型,该模型将带有注意力机制的堆叠 GRU 单元与 CNN 相结合,使用词和距离嵌入作为特征,堆叠的 GRU 网络用于学习单词特征,CNN 用于学习位置特征。总的来说,CNN、RNN 及其变体的性能优于使用 SVM 的传统基于分类器的方法^[384,385,399]。在深度学习方法中,具有多层的架构可以通过使用分层表示生成更合适和完整的特征来增强性能,例如,DCNN^[388]和 DeepCNN^[387]。此外,包括 Transformer 模型、注意力机制和其他特殊嵌入或特征在内的一些创新可以提高模型的性能。深度学习技术在 DDI 检测中显示出巨大的潜力和广阔的应用前景,为研究人员提供了一个极具前景的方向。

6.2.2 基于药物关联数据的预测方法

基于药物关联数据的预测方法利用数据库中已知的药物-药物相互作用关联和药物特征建立模型来预测潜在的药物相互作用。随着众

多公共药物数据库的建立,基于化学和生物学知识的预测模型在 DDI 预测中表现出巨大的潜力。数据库中多种类型的生物医学实体通过不同类型的关系进行相互关联,蕴涵了丰富的图(网络)结构信息。面对复杂、异构和多模态的不同数据源,学者们提出了许多基于机器学习的方法来揭示药物实体之间的隐藏关系。这类方法通常将 DDI 预测作为链接预测任务,以检测药物对之间是否存在相互作用。基于药物关联数据的预测方法可大致分为两类:基于传统分类器的预测方法和基于深度学习的预测方法。

6.2.2.1 数据来源

与基于文献的提取方法不同,基于药物关联数据的预测方法倾向于利用包含各种药物信息的多个数据源。目前广泛使用的药物特定数据库中,包括 DrugBank、FAERS、SIDER、TWO-SIDES 和 OFFSIDES。DrugBank 提供了诸多综合信息,包括 FDA 批准的小分子和生物技术药物,是一个整合了生物信息学和化学信息学资源的药物知识库。FAERS 是一个包含 FDA 记录的不良事件信息的数据库。SIDER 提供了有关已上市药物及其记录的药物不良反应的信息。TWO SIDES 是由 Tatonetti^[400]通过挖掘来自 FAERS 的 DDI 引起的副作用而开发的数据集,包含 645 种药物和由 63,473 种不同药物组合引起的副作用。OFFSIDES 数据库包含 438,802 种药物副作用。多样化的数据源为构建预测模型提供了异构和多模态数据,这些数据集也被广泛应用于 DDI 预测任务并被用于评估 DDI 预测模型性能。

6.2.2.2 基于传统分类器的预测方法

基于传统分类器的预测方法通常利用药物之间的相似性和相异性来构建特征,然后应用传统分类器预测潜在的 DDI。相似性和相异性在模式分类和聚类中起着关键作用^[401-404]。因此,大多数传统的基

于分类器方法的基本概念是：如果药物 A 和药物 B 之间存在相互作用，并且药物 C 与药物 A 相似，那么药物 B 和药物 C 之间可能存在关联。这些方法通常利用各种相似性度量并将它们集成以构建分类特征，然后根据某种分类规则和不同的分类技术，生成潜在 DDI 的概率。例如，Cheng 等人^[405]提出了一种异构网络辅助推理 (HNAI) 框架，使用药物对的多种相似性作为每个药物对的特征，并采用五种预测模型，包括逻辑回归 (Logistic regression, LR)、朴素贝叶斯 (NB)、决策树 (DT)、支持向量机 (SVM) 和 K 最近邻 (k-nearest neighbor, KNN) 来构建预测模型。Qian 等人^[406]使用特征相似性和特征选择方法^[407]构建基于梯度提升的分类器，以加快训练过程并实现稳健的预测性能。

6.2.2.3 基于深度学习的预测方法

基于深度学习的预测方法主要是应用深度学习技术提取深层的特征来预测潜在的 DDI。目前，广泛使用的技术包括深度神经网络 (DNN)、图嵌入 (Graph embedding)、图神经网络 (GNN) 及其变体。

深度神经网络是具有多个全连接层的人工神经网络。基于 DNN 的方法通常利用各种药物数据并使用深度神经网络模型来构建预测框架，将从 DNN 中学习到的更具表现力的数据特征表示用于 DDI 预测。Ryu 等人^[408]基于药物的化学结构信息计算每个药物的特征表示，并将药物对的特征输入深度神经网络模型来预测 DDI。在该工作基础上，Lee 等人^[409]融合多种药物特征来提升模型性能，Deng 等人^[410]进一步构建多通道深度神经网络来学习药物的多模态表示，进而提升模型的预测 DDI 的准确性。

大部分医学实体间的关系可以抽象为图结构，例如异构交互网络和分子图，这种数据我们称之为图数据。为了分析图数据，研究人员

开发了许多基于图的嵌入方法来解决生物学问题^[411-413]。图嵌入方法旨在将图转换为保留结构图信息的低维空间表示，然后将学习到的低维表示作为特征进行预测。近些年，得益于深度学习在图结构数据应用上的重大进展，图神经网络技术开始被应用于药物相互作用预测领域。药物的药效与其性质和功能息息相关，这种性质和功能基本由分子结构所决定。基于图嵌入的方法能够对药物的分子结构进行图建模（即构建分子图），将原子看作图中的节点，将原子间的化学键看作图中的边，然后在分子图上应用图神经网络技术来学习每个原子的特征表示，进而得到药物的特征表示。

基于图嵌入的方法不仅可以对分子进行建模，也能够对药物关联网络以及药物与其他生物实体关系进行建模。该方法将药物看作关联网络中的节点，将药物间的关系看作关联网络中的边，并在关联网络上应用图神经网络技术来学习每个药物的特征表示。**Ma** 等人^[414]基于多种药物特征构建不同的药物相似性网络，并使用多视角图自动编码器来对每个药物相似性网络进行建模，同时使用注意力机制来决定每个视图的权重。**Zitnik** 等人^[415]构建了由蛋白质-蛋白质相互作用、药物-蛋白质靶点相互作用和药物-药物相互作用组成的异构网络，其中每条药物-药物相互作用表示不同类型的药物副作用关系，并应用图卷积网络在异构网络上进行多关系预测。**Lin** 等人^[416]提出了知识图神经网络（KGNN）来对包含药物和其他实体（如药物和基因）的关系的知识图谱进行建模以编码丰富的语义关系。**Yu** 等人^[417]运用子图提取技术精准定位知识图谱中的高质量关系数据，并运用图神经网络技术在子图上学习节点的代表。Lyu 等人^[418]同时对知识图谱和其它异构的药物特征进行建模以学习药物的多模态表征。深度学习模型是解决 DDI 预测问题的一个常用方法，并且在一些研究中取得了令人满意的结果。与传统的分类器相比，深度学习方法通常会产生更具表现力的表示并获得更好的结果^[409,410]。此外，研究表明整合异质药物特

征对于 DDI 预测具有积极影响^[405, 419-421]，这也是当前一个重要的研究趋势。

6.2.2.4 基于矩阵分解的预测方法

矩阵分解作为协同过滤算法之一，近些年在生物信息学任务中取得了令人满意的结果^[422-424]。DDI 预测任务可以表述为矩阵补全任务，旨在对未观察到的药物相互作用进行预测。基于矩阵分解的方法通常对药物-药物相互作用的邻接矩阵进行运算操作。典型的矩阵分解方法包括非负矩阵分解、奇异值分解 (Singular value decomposition, SVD) 等。Rohani 等人^[422]提出了一种称为 ISCMF 的方法，该方法在 DDI 矩阵上采用相似约束矩阵分解。该方法计算子结构、目标、副作用等 8 个相似度，构建一个综合相似性矩阵。Shtar 等人^[425]提出了一种仅使用已知 DDI 作为输入来预测潜在 DDI 的方法，即邻接矩阵分解 (AMF)。AMF 对药物-药物相互作用的邻接矩阵进行矩阵分解，并共享行和列的潜在因素，利用 Adam 优化元素乘法的权重和偏差。

一些方法基于流形学习算法、人工神经网络等开发了新的矩阵分解模型。流形学习是将药物特征纳入经典矩阵分解方法的替代方法，其将高维数据投影到低维空间中，并学习更多潜在信息以重建原始特征。Zhang 等人^[426]将基于药物特征的流形正则化引入矩阵分解以进行 DDI 预测。具体来说，作者计算药物相似性并将它们用作特征空间中的流形，然后应用流形正则化以将流形近似地保持在低维空间中。Liu 等人^[427]提出了一种名为 CLML 的协作线性流形学习模型，该模型使用嵌入在目标网络 (即药物-药物相互作用网络) 和辅助网络之间的流形协同优化节点相似性的一致性，构建了两个流形来测量数据的相关性，这两个流形都是通过迭代的协作学习策略共同重建的，学习到的目标网络揭示了收敛后的预测结果。

6.2.2.5 基于网络传播的预测方法

基于网络传播的方法通常是收集生物医学网络中的属性和结构信息以进行 DDI 预测。在医学和药理学领域，生物医学网络包含各种实体（例如药物、蛋白质）之间的复杂关系^[428-430]。节点的高阶接近度、网络属性、共享属性、局部邻域的相似性和节点之间的关系强度是需要研究的重要信息。基于网络扩散的主流预测方法包括标签传播、随机游走、概率软逻辑（Probabilistic soft logic, PSL）模型、图遍历算法等。

Zhang 等人^[431]开发了一个综合标签传播框架，考虑到高阶相似性和特征集成。标签传播算法通过迭代过程估计其他未标记药物与参考药物发生 DDI 的概率。Park 等人^[432]提出了另一种基于传播的方法，在蛋白质-蛋白质相互作用（Protein-protein interaction, PPI）网络上采用带有重启算法的随机游走（Random walk with restart, RWR）来模拟信号传播，进而计算每个药物对的概率得分。PSL 是一种统计关系学习（Statistical relational learning, SRL）框架，用于关系域中的集体概率推理^[433]。Sridhar 等人^[434]应用 PSL 从多个基于药物的相似性和已知相互作用的网络中推断潜在的 DDI。PSL 模型假设相似的药物可能与相同的药物发生相互作用，该方法将有效的最大后验推断用于 DDI 预测。

此外，还有一些方法考虑定义度量函数来衡量网络关系的强度或实体之间的距离。基于测量关系强度的方法通常构建一个生物医学网络，包括蛋白质、通路等附加的生物医学元素。然后定义一个度量函数来衡量网络关系的强度或候选药物之间的距离，以进行 DDI 预测。例如，Lee 等人^[435]构建具有各种生物实体的异构生物信息网络。该模型的基本概念是：如果两种药物在网络中共享其它的生物实体，则它们往往会发生相互作用。因此作者开发了衡量两种药物关系强度的度量函数来预测潜在的 DDI 并采用图遍历算法统计药物之间的路径数

量，通过计算加权和得到不同类型路径的最优组合。

6.2.2.6 基于集成学习的预测方法

作为机器学习和模式识别的研究热点之一，集成学习算法通过构建多个分类器，然后对其预测结果进行投票来对新数据进行分类^[436]。由于其小样本数据处理、复杂数据结构和高维方面的独特优势，在计算生物学中的应用越来越广泛^[437]。集成学习方法能够结合多个机器学习模型以实现高精度的预测，并可以减少过度拟合训练数据的现象^[400]。

Zhang 等人^[420]采用三种具有代表性的方法来构建基于各种药物数据的预测模型用于集成学习，包括邻居推荐、随机游走和扰动矩阵方法。Deepika 等人^[438]提出了一种半监督学习框架，使用 node2vec 将构建的特征网络中的药物表示为低维特征向量，采用基于正向未标记学习的算法将特征向量分别送入 bagging SVM。作者训练了一个 bagging SVM 作为元分类器，使用基于分类器的输出来生成最终的预测结果。

6.3 人工智能在药物相互作用预测中的发展前景

6.3.1 构建标准数据集

在实际应用中，数据噪声、数据量不足和样本不均衡等挑战给模型的预测精度带来了障碍。DDI 预测的主要困难是缺乏足够的已知药物的相互作用数据和经过实验验证的负样本数据。因此，只有少数样本被标记标签，而大多数样本没有标记的标签，也缺乏非相互作用的药物对的黄金标准数据库。面对原始 DDI 数据集中未标记的样本，大多数方法将其视为负样本，而忽略了未标记的样本可能包含潜在的正样本这一事实，这会对模型的性能产生不利影响。正样本无标签学习（PU learning）是解决该问题的技术之一。例如，Zhang 等人^[421]通

过 LPU 算法^[439]在 PU 设置中构建一个专门的随机森林分类器，以预测新的 DDI。Deepika 等人^[438]应用 bagging SVM^[440]分类器预测 DDI，并产生了较为满意的结果。

6.3.2 药物事件预测

尽管许多基于机器学习的药物相互作用预测方法已经取得了巨大成功，但是它们中的大多数都聚焦于预测药物之间是否存在相互作用。然而，药物互作用可能会引起多种不同的后果或后续事件。

以图 6-1 为例，药物 Itraconazole 和药物 Dabrafenib 发生作用会引起机体的血清浓度下降，但其和药物 Abemaciclib 发生作用后会增加不良反应程度，甚至造成一定的风险。因此，相较于仅预测药物间是否存在相互作用，对于探究联合用药，药物事件预测对揭露其背后的隐藏机制是十分有用的。

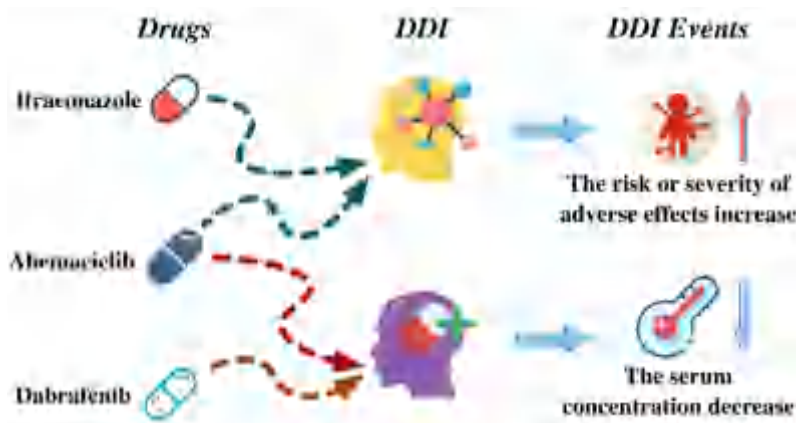


图 6-1 药物相互作用事件^[418]

药物事件预测是有意义的也是有挑战性的，越来越多的研究者开始关注该问题，并构建和整合了很多高质量的数据集。Ryu 等人^[408]将从 DrugBank 收集的药物事件数据分类为 86 种类型。Deng 等人^[410]定义了一个标准协议来分析 DrugBank 中的药物事件数据，并选择了 65

种主要事件用于分析。表 6-1 概括了多种药物事件数据集的统计资料。

表 6-1 药物事件数据集

数据集构建者	药物数目	药物互作用关系数目	药物事件类型数目
Tatonetti 等人 ^[400]	645	63,473	1,318
Ryu 等人 ^[408]	1,704	191,400	86
Deng 等人 ^[410]	572	74,528	65

受益于这些高质量的数据集，许多基于深度学习的药物事件预测方法已经被提出。然而，在药物事件预测方面仍有很大可以改进的空间。比如，药物事件数据集中存在有标签的训练样本数量不足的挑战。因此，建立更大、更完整的药物事件数据集具有重要的意义。其次，药物事件数据集中的类别不平衡问题给分类器的性能带来了严峻挑战，其中少数类样本往往被错误地分类为多数类。采样、成本敏感学习、少样本学习和其他类不平衡学习方法可以用于缓解上述问题。

6.3.3 预测高阶药物相互作用

药物组合是提高治疗效果和降低毒性的有效方法^[441]，这已成为治愈疾病的主要治疗策略^[442]。在过去的几年中，大多数方法都集中在两种药物同时被使用时的 DDI 分析。尽管如此，高阶药物相互作用也值得关注。高阶药物相互作用表示 DDI 存在三种或更多药物的组合服用^[443]。预测高阶药物相互作用具有指导性，并有助于估计多药同时服用对药物不良反应影响的巨大潜力。

Du 等人^[444]通过病历数据库挖掘高阶药物相互作用对肌病的定向影响。除了估计将药物添加到已知组合中的每种风险外，作者还设计了一种可视化方法来设计定向的 DDI 模式。Zhang 等人^[445]通过经验贝叶斯估计方法确定高阶药物相互作用对过度肌病风险的影响。作

者通过采用不同数量的药物组合，从二元到六元，来研究高阶药物相互作用。实验结果表明，肌病风险随着药物相互作用顺序的增加而增加。

6.3.4 整合多源数据分析

生物医学文献的数据包含丰富的语义和句法信息，对相互作用关系提取有很大的贡献价值。而数据库中记录了已知的 DDI 和药物特征，不同类型的生物医学实体是相互关联的，包含了复杂的关系和隐藏的结构信息。来自文献的文本数据和来自各种数据库的药物特征提供了及时和充分的药物信息。因此，整合这两个数据源可以补充药物的知识和特征，从而提高模型的预测性能。

6.4 本章小节

近年来，人工智能技术在药物相互作用的预测和分析中得到了广泛应用，通过设计智能分析算法对药物相互作用数据进行了建模、分析和解释。本章系统性地总结了基于人工智能的药物相互作用预测方法，根据药物数据源的不同对人工智能预测方法进行分类，并详细介绍了这些预测方法的特点和策略。最后，本章讨论和展望了人工智能方法在药物相互作用预测中的应用前景。

第 7 章 药物发现中的大规模预训练模型

7.1 分子表征

开展基于人工智能的药物小分子研究，第一步是通过人工智能模型可以理解的方式对分子进行表征。分子表征过程是将真实存在于物理世界的分子使用数学的方式表示出来，并驱动模型通过这种表示方式理解分子真实的情况，进而达到对分子建模的目标。在早期的 QSAR^[446, 447] 研究中，传统机器学习模型的弱学习能力只能对目标关系进行线性和简单的非线性建模。为了能够使模型学到分子的化学结构与分子的物理、化学、生物性质之间的复杂关系，在对分子进行表征时，往往需要专家对分子进行手动描述。这种描述的核心之处是设计特定的分子描述符来对分子进行表示^[448-450]。在分子中，通过对任务有重要影响的特征进行数字上的标注，降低模型学习难度。描述符的设计对于最终建模结果的影响巨大，直接决定了最终结果的好坏，因此基于分子描述符的分子研究严重依赖于专家的背景知识。针对 AI 分子或 AI 药物的研究中，具体任务数量繁多，这些不同的任务往往关注分子不同方面的特性，而针对每个任务单独设计特定描述符的代价过高，这在一定程度上限制了基于计算的分子/药物研究。

幸运的是，人工智能技术的兴起与发展为相关研究领域带来了巨大的改变，人工智能模型所具有的强大学习能力可以学习到深入的非线性的关系^[147]、自动从原始数据中学习任务相关的特征。该特性将研究者们从描述符的设计中解放了出来，因为人工智能模型可以直接接收分子的原始表示作为输入，并在训练过程中自动对分子进行理解。早期的工作中，研究人员一般使用 SMILES^[279] 来对分子进行表示。SMILES 是一种线性的分子表示方式，最早由 Arthur Weininger 与 David Weininger 于上个世纪八十年代提出，最终由日光公司完善并规范化。SMILES 设计之初的目的是为了能够以一种十分简易的方式来

储存分子，因此 **SMILES** 的优点是数据类型简单，可以以极小的储存代价在没有歧义的情况下完整地表示一个分子。另一方面，由于近些年自然语言处理 **NLP** 领域的飞速发展，**SMILES** 作为一种化学语言天然地与自然语言有着极高的相似性，可以直接被应用在成熟的 **NLP** 方法中。依靠着计算方法的成熟与储存上的高效性，**SMILES** 逐渐成为了基于 **AI** 的分子研究中最先被广泛使用的分子表示方式。

然而 **SMILES** 的简易性与高效性是通过牺牲对于分子结构的直观表示而得到的，这使得 **SMILES** 对于分子的表示较为抽象，限制了其进一步的应用。分子原本的结构是空间中的三维结构，其中的三维信息是由键与键之间的角度、二面角与键长决定。由于许多化学键存在着可旋转的性质，分子的三维构象并不固定，因此一般情况下，针对分子信息的记录可以将空间中的三维信息简化为二维的拓扑信息。为了能够在线性空间中表示二维空间的信息，**SMILES** 首先将分子中的环状结构打断，并使用成对的数字来记录开环处两端的原子。当分子中出现分支结构时，会选择其中一条链作为主链，其他的链被当作支链放在小括号内进行记录，当支链记录完成后，对于主链的记录继续进行，直至分子被完全记录。最终分子被压缩成为一条线性的表示，尽管这种线性表示结合 **SMILES** 的规则可以完整复原出分子原本的二维结构，但实际上对于 **SMILES** 规则的理解却成为了众多模型的痛点。在使用 **SMILES** 作为输入分子相关预测任务中，模型需要通过输入的 **SMILES** 重建出分子原本的结构，并在此基础上学习分子结构与性质的内在关系，这无疑增加了模型学习的难度。

另一方面，**SMILES** 中支链的优先记录机制导致原本在分子中某一主链上相连的两个原子，在 **SMILES** 中可能会被一长段支链所隔开，这就带来了另一个问题。在分子生成任务，对于这一 **AI** 分子研究中最重要任务，**SMILES** 的特性使得模型难以对正处于生成过程中的分子进行全面的的有效性检验。生成分子的有效性是分子生成任

务中的一项基础挑战，模型需要生成满足化学规则的分子。上述问题使得基于 SMILES 的分子生成工作需要长期面对生成的分子不满足 SMILES 语法规则、不符合化学规则的情况。

面对以上两点问题，研究人员迫切地需要一些解决方法来解决基于 SMILES 的研究工作中模型对于分子的理解问题。在 SMILES 之后，针对于分子图 (Molecule graph) 的研究逐渐兴起，分子图由于其对分子结构的直观展示而被广泛认为是一种更好的分子表示方式。尽管基于分子图的研究可以避免上述提到的 SMILES 面对的种种问题，但模型能否真正理解分子仍然是研究人员们所担忧的点。此外，受限于药物设计领域极低的有标注数据量，面对具体任务时，人工智能模型无法得到充分训练以及过拟合等问题同样困扰着相关研究人员。通过特殊的任务形式，利用起药物设计领域“有标注数据少、无标注数据多”的特性，使人工智能模型能够对分子进行良好的理解、降低人工智能模型在不同任务中的学习难度、提升人工智能模型在分子任务上的具体表现，成为了 AI 药物研发中的一项重要内容。

7.2 预训练

预训练的理念源自于自然语言处理领域。自然语言处理与 AI 药物分子研究最相似之处在于两者所面对的数据都是离散的，即自然语言与分子都由离散的非数值型元素组成，例如句子由不同的字与词组成、分子由不同的原子组成，并且这些基础元素所表达的含义都难以直接通过数值的形式表示出来。与之形成对比的，计算视觉 (Computer vision, CV) 研究中所面对的图像 (Image) 则完全不同，图像中的每个像素点都由一组连续的数字组成，并且数字的大小与其所表达的含义直接关联，因此针对于图像的研究并不需要进行表征操作，图像本身便是一类已经表征好的数据。

人工智能模型的本质是对数字的运算，因此进行自然语言研究的

第一步便是将离散的字、词转为恰当的数值表示。在最早期的研究中，为了在数值上表现出词与词之间的离散性，并避免数值大小对于词含义的影响，研究人员们通常使用 **one-hot** 方法来对词进行编码，转为数值表示。在 **one-hot** 编码中，每个词都由一条 n 维的向量表示， n 为所有词的种类，在这个词向量中，仅有词对应维度的数值为 1，其余维度都用 0 来表示。**one-hot** 的编码规则可以有效避免数值大小对于词含义的影响，但是也会使得所有词在编码空间中的距离相同。然而，在实际的自然语言中，词与词之间的距离有着显著的差异，例如同义词在语义空间中的距离远小于反义词间的距离，而反义词间的距离又远小于不相关词之间的距离。

为了能在词的表征上表现出词与词之间的距离与关联，研究人员们将对于词的编码与表征部分从各个特定任务中独立出来，作为一类全新的研究任务。**word2vec**^[282, 451]是这一时期最具代表性的工作类型之一，通过特殊的任务将离散的词在连续的空间中表示出来，最终得到 **embedding** 作为词的连续表示。**word2vec** 利用句子中相近的词之间具有一定关联性、通过跳字模型（**Skip-gram**）与连续的词袋模型（**Continuous bag of words, CBOW**）来对词的语义进行学习。使用 **word2vec** 得到 **embedding** 代替原始的词并开展具体的下游任务往往会取得更好的效果。在一定程度上可以认为 **word2vec** 的工作是一种初步的“预训练工作”，一个重要的思想此时已经形成：“将一个特定的任务分成共性学习与特性学习的两个部分，并使用大规模的非该任务数据来对共性部分进行学习”。在 **NLP** 研究中，共性部分是对于词在不同任务中的所表示的意思固定。

在 **word2vec** 之后，研究人员们再次发现了新的不足之处。自然语言中存在着大量的多义词，并且那些只具有一种语义的词语在不同的语境下也通常会表现出不同的含义，而 **word2vec** 只能固定地表达每个词的同一种语义。尽管在下游任务的训练中可以对 **embedding** 进

行微调，但 **word2vec** 的不足依然限制了语言模型的对于自然语言的学习与理解。随后提出了用于解决上述问题的预训练思想。从 **word2vec** 到预训练，不变的是通过特定的任务来学习词的表示，核心的变化仅发生在 **embedding** 的形式上。这一时间点上的代表性工作有 **ELMo**^[452]、**BERT**^[453] 与 **GPT**^[454]。

与 **word2vec** 将词表示为一条连续的向量并作为一种显式的 **embedding** 不同，预训练工作则将对于词的理解存储于模型内部，是一种隐式的 **embedding**。出于上述隐式 **embedding** 的特性，提取预训练过程学到的语义信息通常需要使用原始模型在特定的任务上进行微调，而非拿到特定的分子表示后再使用其他模型对特定任务进行学习。因此，预训练工作通常由两部分组成：(1) 设计特定的预训练任务以充分学习语义信息；(2) 针对不同下游任务设计特定的微调策略。预训练工作的预训练与微调过程共享模型，因此预训练模型的设计也会对其在下游任务中的应用范围产生影响。

7.3 分子预训练

随着大规模预训练研究模式在 **NLP** 领域的开展与成熟，利用大规模无标签数据和自监督任务使模型学习数据共性信息的思想，也被 **AI** 药物小分子设计领域所关注。从早期使用 **word2vec** 思想进行分子表征（相关工作例如 **Mol2Vec**^[281]），到中期沿用 **NLP** 的预训练思路开展分子预训练，到最新的针对分子特性开发预训练方法，已经有大量的分子预训练工作被推出。

在预训练工作中，预训练任务的定义与学习直接关乎模型在下游任务中的表现，针对不同的分子表示方式与目标学习成果设计特定的自监督任务是预训练工作的核心。这里，我们按照分子预训练工作中预训练任务的种类对主流的一些分子预训练工作进行介绍。

7.3.1 基于 Mask Language Model 的分子预训练

由 BERT 首次提出的 Mask Language Model (MLM) 是 NLP 领域最为成熟的预训练模型, 通过将样本中的任意部分随机遮盖(mask), 再使用 transformer 编码器基于样本中未被遮盖的部分来复原出被遮盖的部分, 模型可以有效挖掘出样本上下文之间的支撑关系, 进而学习样本的语义信息。由于 MLM 在 NLP 领域取得辉煌的表现, 并且相关的参数设定也基本被确定, 因此大多数分子预训练工作也都沿用了 MLM 的思想来对分子这一数据进行学习。

SMILES-BERT^[455]首先将 MLM 应用在 SMILES 这一分子表示方式上, 以解决模型对于 SMILES 的理解问题。SMILES-BERT 对输入 SMILES 中的全部 token, 按照 15% 的概率进行随机 mask, 并使用一个 transformer encoder 对被遮盖的 token 进行复原。随后的 MolBERT^[456]与 ChemBERTa^[457]则在 SMILES-BERT 基础上进行了改进。MolBERT 将两条 SMILES 同时输入进模型之中, 并在 MLM 之外加入了对两条 SMILES 进行是否为同一分子的判断以及包含了 200 条分子特性的 PhysChemPred 任务。ChemBERTa 虽然沿用了 SMILES-BERT 的单 SMILES 输入形式, 但是引入了动态 mask, 即每条 SMILES 每次被输入到模型前都会被重新 mask, 而非 SMILES-BERT 中每一条 SMILES 只有一种 mask 情况。结合了优化后的训练方法, ChemBERTa 相较于 SMILES-BERT 展现出了显著的性能优势。ChemFormer^[458]则是在模型方面进行了改动, 使用完整 transformer 的 encoder+decoder 结构对被 mask 的 SMILES 进行复原。与此前的 encoder-based 工作不同, ChemFormer 将 MLM 中分子信息提取与 masked token 复原两个部分分别由 encoder 与 decoder 完成, 基于自回归模式的 decoder 在复原时将整个分子复原出来, 而不再局限于被 mask 的部分。

除了 SMILES, 也有基于分子图的预训练工作采用了 MLM 这一

任务类型，代表性的工作有 MG-BERT^[459]、Grover^[460]和 MPG^[216]。MG-BERT 在原子层面对分子图进行了 MLM 的学习，同时通过在分子图中添加氢原子来隐式地添加化学键信息（在 AI 分子研究中，氢原子通常被忽略）。Grover 对分子图进行了两个层面上的 mask，分别是基于原子的 mask 与基于化学键的 mask。在基于原子的 mask 中，分子中的被选定 mask 的原子与其直接相连的原子，以及其中化学键均被 mask；在基于化学键的 mask 中，分子中的被选定 mask 的化学键与其两端原子，以及这些原子相连的化学键均被 mask。此外，Grover 也在 motif 层面对分子图内的不同部分进行了预测。MPG 只进行了原子层面的 MLM，但额外加入了成对的半分子图区分（Pairwise half-graph discrimination, PHD）任务。PHD 首先对分子图进行对半拆分以及随机重组，再通过区分重组后的两个子图是否源于同一原图使模型在分子维度上对分子的表征进行学习。

此外，也有工作同时使用多种分子表示方式，基于 MLM 进行分子预训练。相关工作如 Dual-view Molecule Pre-training (DMP)^[461]，同时使用 SMILES 与分子图两种分子表示方式。DMP 在 SMILES 中随机对 token 进行 mask，而在分子图中随机对原子进行 mask，同时保证这些 mask 在两种分子表示方式中的一致性，并分别使用两个模型进行 MLM 任务。在两个单独的 MLM 任务之外，DMP 使用了一个正则项将两个模型关联，使其在训练过程中互相促进。

7.3.2 基于生成式模型的分子预训练

在基于 MLM 的预训练工作中，由于模型需要根据被 mask 部分的“上下文”来对 mask 部分进行复原，因此模型内部通常是双向的运算机制，即每个 token 都能观察到它前面与后面的 token。MLM-based 模型中的双向运算机制使得这类模型在性质预测相关的任务中往往有着良好的表现，因为双向运算可以更好地提取样本中的特征，但同

时也使模型在生成式任务中表现不尽如人意。在生成式任务中，由于新生成的部分依赖于已生成的部分，并且无法观测到未来会生成的部分，因此执行的是单向运算机制。这与 MLM-based 模型的学习过程不同，会导致模型表现不佳（生成式任务分为自回归生成与非自回归生成，非自回归生成不存在上述问题，但由于非自回归生成的表现通常远弱于自回归，仅在运算速度上有优势，因此主流的生成式任务是基于自回归模式的生成）。

针对上述问题，基于生成式模型的预训练工作被提出来。相关的工作有 X-MOL^[462]、PanGu^[463]等。X-MOL 利用了 SMILES 的不唯一性，通过分子的一条 SMILES 生成分子的另外一条 SMILES 来达到理解 SMILES 和表征分子的目的。PanGu 则是在此基础上将数据类型拓展到了分子图，通过分子图来生成分子的 SMILES。除此之外，还有一些研究使用分子的 SMILES 来生成分子指纹。需要注意的是，由于分子图的生成难度较大、耗时较长，基于分子图的生成式预训练工作目前还较为少见。

7.3.3 基于对比学习的分子预训练

对比学习是目前一类较为流行的自监督任务，其理念为对数据样本进行扩增、并使模型学习源自同一样本的扩增样本之间的相似之处与源自不同样本间的差异之处，进而学习样本的共性信息。AI 分子研究同样可以借助这样思想进行预训练工作，例如基于 SMILES 的 MM-Deacon^[464]与基于分子图的 MolCLR^[465]、MoCL^[466]等。

对比学习的核心问题在于如何对样本进行扩增，在分子领域，不同的分子表示通常对应着不同的扩增方式。MM-Deacon 首先将分子转为 SMILES 与 IUPAC（International union of pure and applied chemistry）两种表示方式，并使用两个独立的模型分别对 SMILES 与 IUPAC 进行理解与表征。在训练时，MM-Deacon 中的两个模型被要

求最小化同一分子间的表征差距，并最大化不同分子间的表征差距。

MolCLR 对分子图进行了三种随机修改，包括原子遮盖、化学键删除以及子图删除，针对每个分子图，每次修改便得到一个新的扩增分子图。随后 MolCLR 模型被用来对这些扩增后的分子进行表征，同时源自同一分子图的扩增分子图需有相近的表征，而对不同分子扩增得到的分子图表征则需相对较远。MoCL 与 MolCLR 相似，但 MoCL 认为对分子图上的原子、化学键以及子图进行删除会大幅度影响分子本身的性质，进而导致同一分子使用不同删除方式得到的扩增分子图之间可能存在较大的性质差异。MoCL 基于领域知识设计了 230 条替换规则，并将分子图扩展的方式改为了基于这些规则进行 motif 维度上的替换，以保证替换前后分子间性质的相似性。

7.3.4 基于几何特征的分子预训练

随着 AI 分子研究的不断发展，研究人员们希望通过加入空间几何信息来对小分子进行更进一步的建模^[467]，而此前的工作都只考虑了分子的拓扑结构信息。新提出来了一系列涉及到分子几何特征的预训练工作，代表工作如 GeomGCL^[468]、GEM^[469]及 Uni-Mol^[470]等。

GeomGCL 同样也是基于对比学习而构建，但 GeomGCL 认为对分子的任何改变都会影响分子的性质，因此引入了分子的几何来进行分子的数据扩增。首先每个分子由两种方式进行表示——分子图与分子几何表征，随后两个编码模型分别对分子图与分子几何表征进行编码，源自于同一分子的两种表示方式得到的编码需尽可能相近，而源自不同的分子的表征则需距离较远。

GEM 则是完全针对于分子几何性质的预训练工作。GEM 接收分子图作为输入，并对分子内化学键的键长、键角进行预测，在局部结构层面上对分子的几何特征进行学习。此外，GEM 预训练任务也包含了原子间距离的预测，分子中两个原子在空间中的距离受到分子整

体的几何结构影响，因此这一任务是在分子的全局结构层面对其几何特征进行学习。

Uni-Mol 在分子表示方式、任务设计与模型构建上都带来的巨大创新。在分子表示方式上，除了原子维度的表示，Uni-Mol 创新性地额外使用原子对来对分子的几何性质进行表示。同时，Uni-Mol 在模型结构上将原子维度的表示与原子对维度的表示联系在一起，使两者共同参与分子的表征计算。而在任务设计方面，尽管 Uni-Mol 沿用了 MLM 的思想，分别在原子和原子对维度进行 mask 并训练模型复原，但 Uni-Mol 额外引入了蛋白质口袋模块，并在蛋白质口袋模块上也进行了同样的预训练任务。额外引入的蛋白质口袋模块既能够在预训练阶段对蛋白质进行学习，又可以避免完整蛋白质过大而引起的建模困难问题。最终 Uni-Mol 在多项下游任务中取得了 SOTA (state of the art) 的效果，尤其是蛋白质-配体结合预测中，Uni-Mol 更是超越了一众基于理论计算的方法。

7.3.5 基于领域知识的分子预训练

上述预训练任务类型都围绕着分子的结构展开，还有一些工作则利用领域知识来构建预训练任务，代表工作有 MoLR^[471]，以及上文提到 MoCL。MoLR 利用了化学反应信息，使用对比学习的方式来进行分子预训练。MoLR 认为化学反应的反应物与生成物在一定程度具有等效关系，并基于这种关系设计了数据扩增方法。在 MoLR 的训练中，来自同一反应的反应物与生成物为正向样本，在经过编码后应相距较近，来自不同反应的反应物与生成物为负向样本，在经过编码后距离较远。MoLR 提出的基于化学反应的对比学习避免了传统对比学习中数据扩增方法对于样本本身的破坏。

需要注意的是，通常分子的性质都与另外一些分子相关联（包括小分子与大分子），而基于分子结构的预训练只能学习到与分子自身

相关的性质。基于领域知识的分子预训练可以有效避免上述问题，但关于领域知识的数据同时也是有标签数据，这意味着基于领域知识的分子预训练无法利用分子研究领域的大规模无标签数据。

7.4 分子预训练范例

本节选取目前可以被微调至下游任务种类最多的 X-MOL 作为范例，明确分子预训练工作中所包含的工作内容、工作流程以及需要注意的点。

7.4.1 确定预训练任务与模型结构

7.4.1.1 预训练任务设计

通常，针对预训练任务设计的第一步是明确预训练的目标，并将这个目标抽象化为一个可学习的人工智能任务。X-MOL 的目标是通过预训练的方式去使模型学到 SMILES 的语法规则。X-MOL 提出以下假设：“模型学会了 SMILES 的语法规则”与“模型可以将一条 SMILES 复原为一个分子结构”以及“模型可以将一个分子结构转换为一条 SMILES”互为对方的充分必要条件。当把分子结构视作模型内部对于分子的一种隐式表示时，整个任务就变成了“模型接收一条 SMILES 作为输入并理解其所代表的分子结构，随之返回属于这个分子的另一条 SMILES”。以上是 X-MOL 设计时的思路以及最终采用的预训练任务。

7.4.1.2 模型结构设计

在明确了预训练任务后，需要针对任务本身的特性进行模型结构的设计，以保证模型能够在预训练任务中有效学习预训练目标。此外，针对模型结构的设计也需要考虑模型未来将会被应用的下游任务，设计出的模型结构需要对这些任务兼容。在这一部分，我们将介绍 X-

MOL 如何设计其模型结构。

为尽可能地保证输入端与输出端对 SMILES 理解方式的对齐, X-MOL 采用共享参数的方式强制使 Encoder 与 Decoder 保持相同的 SMILES 理解方式。并且通过 attention mask 的方式在一个 Encoder 模型内实现了 Encoder 与 Decoder 内部不同的特征抽取方式、完成了逻辑上的 Encoder-Decoder 结构, 避免 Encoder 与 Decoder 的时序问题, 并大大降低了对于显存的占用。X-MOL 的模型结构如图 7-1 所示。

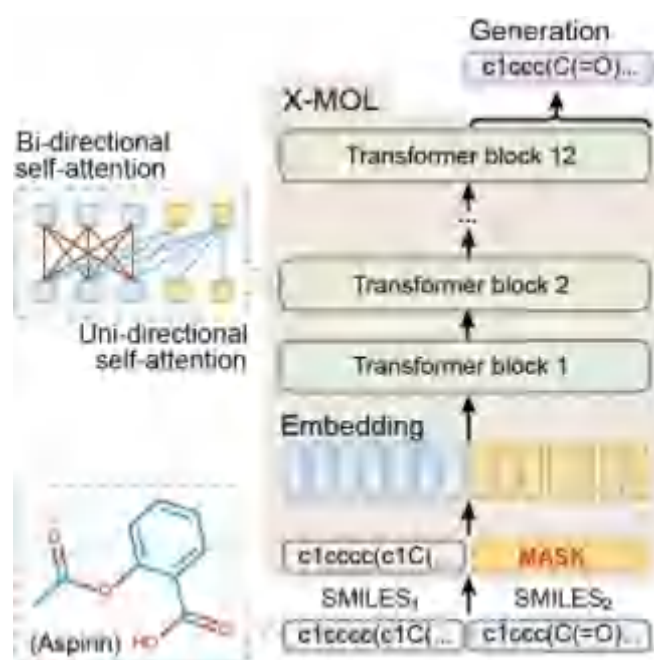


图 7-1 X-MOL 模型架构图

7.4.2 构建运算平台

预训练工作通常使用具有强大学习能力的超大模型, 在大规模的无标签数据上进行学习, 这对整个计算平台的计算力与数据收集都带来了巨大的考验。因此除了预训练任务的设计问题, 能使预训练工作顺利运行的工程问题也十分重要。

在训练数据方面, X-MOL 选择了 ZINC15^[472]数据库中的全部数据, 总计有超过 11 亿个小分子被应用在 X-MOL 的预训练过程中。

借助 Hadoop 技术与百度的云计算平台，X-MOL 在时间可控的情况下对如此大规模的数据进行预处理。在数据预处理的过程中，超过 1000 颗 CPU 核心被调用来同时对这 11 亿个小分子进行处理。对于 SMILES 的标准化和随机化等操作，则是通过化学信息学包 RDKit^[473]来完成。

在模型训练方面，X-MOL 中所有模型均使用百度的 PaddlePaddle 计算框架搭建构。包括预训练与微调在内的所有模型训练均依托于百度公司的 PaddleCloud 云计算平台，所有预训练任务的单次训练由 8 或 16 张 Tesla P40 GPU（24GB 显存）同时完成，在微调时则使用 4-8 张 Tesla P40 GPU 同时完成。在此计算条件下，X-MOL 的单次预训练大约持续 4 天时间，而微调至下游任务时的训练时间根据任务的不同而有所不同。

7.4.3 设计微调策略

预训练模型需要被微调至具体任务中才能体现出其预训练阶段所需到的共性信息，因此设计合理、能有效利用共性信息并能够完成具体下游任务的微调策略将决定预训练工作是否具有实际的应用价值。我们以 X-MOL 为例详细介绍如何将经过预训练的模型微调至多种下游任务当中。

7.4.3.1 微调至预测任务

将 X-MOL 微调至所有预测任务的策略如图 7-2a 所示。预测任务的核心是从输入数据中抽取特征，并使用特定的 readout 操作来将这些特征汇总成为预测目标。常见的 readout 操作的种类有求平均方法与虚拟点方法。大部分基于分子图的工作使用求平均法，将分子图模型输出的所有原子进行加和/求平均，用于对整个分子进行表示，进而完成性质预测。虚拟点方法常见于基于序列的分子模型中，通过引入

一个虚拟的、可以观察到分子全貌的点（在不同工作中称呼不同，例如[CLS]、[BOS]等等）来对整个分子进行表征。此外，如果预训练模型并不完全适用于预测任务，则需对预训练模型本身结构进行修改。例如 X-MOL 在被微调至所有预测任务时，“Decoder”部分被完全删除，从而使 X-MOL 的 Encoder-Decoder 结构简化为单纯的 Encoder 结构以提取分子中的特征。

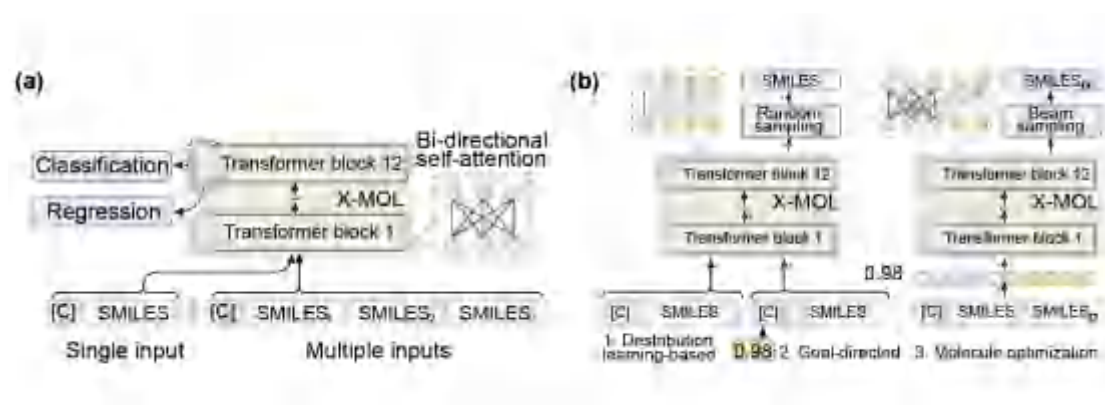


图 7-2 X-MOL 的微调方法示意图

7.4.3.2 微调至生成任务

由于分子生成任务与分子性质预测任务在内部计算的方向上有本质区别，因此除了基于生成式的预训练任务等使用单向计算机制的工作外，绝大多数工作均无法被微调至分子生成任务。需要注意的是，这里所说的无法被微调并不指工程上无法将模型应用于分子生成任务，而是由于计算机制的改变，使得预训练的效果失效、无法对下游任务起到足够的支持作用。而基于生成式预训练任务的工作，例如 X-MOL 等，在被微调至分子生成任务时，沿用预训练阶段的任务运行方式即可，详细的策略如图 7-2b 所示。

7.4.4 模型微调与评估

预训练工作的最后部分是将模型微调至各个具体的下游任务中，

根据不同的模型在不同任务中的实验结果对预训练工作进行全面的评估。在对预训练模型的下游任务中的评估中，通常在不同任务中使用预训练模型与专注于该任务的先进工作进行对比，以展示出预训练模型的性能。同时，评估过程还需要加入与自身冷启动模型的对比(冷启动意为保证预训练模型结构不变，但不适用预训练的到的参数，只使用随机初始化的参数)，用于排除预训练模型结构带来的影响。X-MOL 经过微调后在众多的分子任务上都取得了令人振奋的表现^[462]，表明了大规模预训练与分子表征学习对于整个 AI 分子领域的促进作用。

7.5 本章小节

本章重点关注于小分子药物的预训练模型构建和应用。首先，对于小分子的常见表征方式进行了详细介绍；其次，对于几种常见小分子的预训练模型进行介绍，具体包括：基于 Mask Language Model 的分子预训练、基于生成式模型的分子预训练、基于对比学习的分子预训练、基于几何特征的分子预训练和基于领域知识的分子预训练五类。最后，以 X-MOL 模型为例，对分子预训练的模型构建过程和下游应用进行介绍。总结来说，小分子预训练模型可以有效地利用海量的无标签数据，在众多的下游分子任务上均取得了令人振奋的表现，表明了大规模预训练与分子表征学习对于整个 AI 药物设计领域的重要促进作用。

第 8 章 药物发现中的可解释人工智能模型

8.1 药物发现中的可解释人工智能模型概述

随着 AI 技术在自动驾驶、医疗诊断和金融保险等应用领域的不断发展，AI 系统做出决策或者提出建议的同时，向用户、开发人员和监管机构提供解释是相当重要的。尤其在生物医学领域，决策和建议需要更高的透明度和可问责性。2021 年，“透明性和可解释性”包括在联合国教育、科学及文化组织（United nations educational, scientific and cultural organization, UNESCO）发布的首个全球性 AI 伦理协议《人工智能伦理建议书》中提出的十大 AI 原则。可解释人工智能技术也是美国国防部高级研究计划局（Defense advanced research projects agency, DAPRA）启动的重要研究计划，旨在实现“第三代 AI 系统”。中国国务院在 2017 年印发的《新一代人工智能发展规划》中提出“实现具备高可解释性、强泛化能力的人工智能”，得到了产业界和学术界的广泛认可和积极响应。

生命科学领域的飞速发展，使得疾病存在更多可使用的治疗方案。理论上，几乎所有生物学过程都可以被药物靶向。然而针对特定疾病的一大挑战是，如何寻找出具有合适药理学、毒理学和药代动力学等特性的类药小分子。对于当前海量的生物医药数据，虽然计算机研究人员能够设计出有效的预测模型，但这些预测模型的结果难以被生物学家、药物学家、化学家等理解，且这种不同领域之间的理解难度正在逐渐扩大，这也使得生物医药研究人员很难信服于人工智能预测模型的结果，进而使得人工智能技术在药物研发中的应用受到限制^[193]。

AI 药物发现研究的重点之一就是探索如何解释研究人员构建出的人工智能模型的预测结果，进而开发出更符合化学背景、更易于生物医药研究人员理解的 AI 模型^[474]，给出药物作用机理解释，提升药物安全性、优化有机合成设计，促进化学信息学家、药物化学家和数

据科学家之间的合作。可解释人工智能技术在药物发现研究中的主要内容具体体现在以下几个方面：(1) 实现模型透明化，阐述 AI 系统给出药物设计的计算流程；(2) 明确决策来由，证明 AI 系统提供的预测结果是可以接受的；(3) 提供新知识，发掘药物设计和分子发现中关键标志物；(4) 评估可靠性，量化预测结果的不确定性^[475, 476]。

目前，药物发现研究中的可解释人工智能技术仍处于起步阶段，但正快速向前发展。相信在不久的将来，人工智能终将补齐在感知、记忆、推理等方面的功能“短板”，“黑箱”模型将变得更加透明。在药物研发领域，通过机器智能可以帮助药物学家快速、高效地处理海量数据，做出合理决策。

8.2 可解释人工智能技术 (XAI)

8.2.1 可解释机器学习

在许多实际场景中，内在可解释的机器学习算法有着广泛的应用，比如，树模型、广义加性模型^[477-480]。简单明了的统计意义特征和可视化方法的使用，对人们发现数据中潜在的基本规律并完成最终决断具有辅助作用。同时，因为本身具有良好的可理解性，可解释的机器学习算法广泛应用于金融保险、医疗健康等众多风控领域。

线性回归是经典的内在可解释机器学习模型，回归系数是自变量与因变量相关性的直接体现，代表了输入特征的重要程度，系数越大表示对应自变量起的作用越大，特征越重要。线性回归模型透明度高，解释起来也非常简单。作为一种广义线性模型，逻辑回归是最常用的适用于二分类的模型，是线性回归模型的拓展，其可解释性特征与线性回归模型类似。

但是，当特征与结果之间的关系是非线性的或者特征之间存在联合作用时，线性回归和逻辑回归将失效。此时，决策树模型可以发挥重要作用。决策树模型能够从一组有特征和标签的数据中总结出决策

规则并用树状图的结构来呈现，可以很好的解释模型结果。在决策树模型中，根据特征被选中的次数和信息量来衡量特征的重要性，次数越多、信息量越大，该特征越重要。在回归问题中，评估均方误差(MSE)的变化量被用于评估特征重要性。此外，通过可视化决策树模型可以清晰地看到模型的决策路径，易于人们理解和使用。通过将树模型中的规则融入广义加性模型，RuleFit 模型^[481]自动将特征交互添加到线性模型中，使得模型的精度和可解释性都得到了很大的提升。此外，还有朴素贝叶斯模型、K 近邻方法等多种内在可解释机器学习模型。这些算法在建模阶段能够帮助开发人员理解模型，进行模型的对比选择并在必要时调整优化；在投产阶段能够通过模型的运算机制，对模型的结果进行解释^[482]。然而这些算法难以胜任复杂的分析任务，应用场景存在很大局限性。

8.2.2 图结构的可解释技术

在现实世界中，许多复杂数据的本质都是图，例如社交网络、生化结构、药靶关系等。基于图结构的算法相比于序列算法能够更自然地表达这类数据的关联关系，因而出现许多基于图的神经网络技术，例如图卷积神经网络^[483]、图注意力网络^[484]、变分图编码器^[485]以及异质图注意力网络(Heterogeneous graph attention network, HAN)^[486]等。这种灵活的建模方式使得图网络技术本身具备一定的内在可解释性，例如知识图谱、语义图等本身由人类可理解的抽象概念组成，但这种建模方式同时也提升了解释难度。

解释图模型的挑战在于：图是非网格数据，每个节点都有不同数目的邻居并且处于不同的局部拓扑结构中。因此，好的解释方法需要融合图的结构。近期有综述对这类解释方法进行了总结^[487]，根据其解释对象的不同将这些方法划分为实例层面的解释方法和模型层面的解释方法。在此重点介绍一下实例层面的解释方法，实例层面的方

法与特征工程的思想有些相似，旨在探究影响模型预测的重要特征，即回答哪些输入特征更加重要，哪些图的模式更有利于正确决策。根据获取特征重要性得分的具体方式的不同，实例层面的解释方法可以被进一步划分为：基于梯度或特征的方法、基于扰动的方法、基于分解的方法和基于代理的方法。下面具体介绍这几种方法的主要思想。

8.2.2.1 基于梯度或特征的方法

基于梯度或特征是一种直观的解释思路，其思想是将梯度或隐空间的特征图近似于输入特征的重要性。通常，梯度值或者隐含特征图的值越高，意味着输入的特征越重要。该类的代表方法有 SA^[488]、GBP^[488]、CAM^[489]和 Grad-CAM^[489]等。

8.2.2.2 基于扰动的方法

基于扰动的方法主要是在具有不同扰动的输入的情况下，监测输出结果的变化情况。当预测结果相比于原始预测发生了较大变化，则表明输入中重要部分被扰动；当预测结果与原始预测变化微小，则说明重要输入未被扰动，即被扰动的是无关紧要的信息。扰动的常见方式是对节点、节点的特征、边以及边的特征添加掩码。该类的代表方法有 GNNExplainer^[490]、PGExplainer^[491]和 GraphMask^[492]。

8.2.2.3 基于分解的方法

基于分解的方法将原始模型的预测分解为若干项，将预测出的打分分配到输入空间，通过反向传播的方式逐层分配预测得分，直到输入层。最后通过组合，可以表示边的重要性、节点的重要性以及游走路径的重要性。该类方法的主要代表有 LRP^[488]、Excitation BP^[489]和 GNN-LRP^[493]。

8.2.2.4 基于代理的方法

基于代理的方法基本思想是简化输入与输出之间的非线性关系，即虽然难以解释原始的深度神经网络，但是可以采用简单且自洽的代理模型来近似代替。为了获得对给定输入数据预测结果的解释，该类方法需要先对输入数据进行抽样，以获得目标数据周围的关系表示，通常使用简单的机器学习模型而不是复杂的深度神经网络作为代理模型。该类方法的主要代表有 GraphLime^[494]、RelEx^[495]和 PGM-Explainer^[496]。

相比于上述实例层面的解释方法，模型层面的解释提供了更高层次的见解，但目前该方向的研究工作还比较少。模型层面的解释方法基于生成思想，不考虑任何具体的输入实例，而是对图神经网络工作的一般性原理进行解释，特定的输入图模式会在图神经网络上产生特定行为。XGNN^[497]作为唯一的模型级解释的典型方法，它通过一个可训练的图生成器，生成对目标任务可以产生最优预测的图，并将生成的图作为目标预测的解释。

8.2.3 建模后的可解释技术

在可解释技术作用的阶段，可以将其分为内在（或原生）的可解释技术和建模后的可解释技术（post-hoc）。内在可解释技术作用于建模过程中，强调模型自身具备可解释性，除了可以使用“透明度”高的机器学习算法或者基于规则构建模型以外，该类技术的典型场景还有视觉问答（Visual question answering, VQA）^[498]和原型网络设计^[499]。VQA 的一般解释方式是同时训练一个模型和一个语言的解释器。原型网络的解释方式则是参照人类加工和处理抽象信息的过程赋予模型解释的步骤或者环节。内在可解释技术的研究应用还比较少，因此不作为该节的主要内容。

建模后的可解释技术的主要思路仍然是将深度模型视为黑盒，不

做显式的拆解，而是通过假设和检验去观察模型，据此解释模型的实际工作方式。此类方法的解释系统一般与原 AI 系统之间是解耦的，可以做到解释不依赖于模型，即与模型无关，是可解释人工智能的研究重点。建模后的可解释技术根据其解释目标的不同，可以分为局部可解释和全局可解释。下面重点介绍一些经典的建模后可解释技术。

8.2.3.1 基于局部的解释方法

基于局部的解释方法以特征归因法为主要代表。特征归因是一种分析模型决策对于特征依赖程度的度量方法。该类解释方法的思想是：具有越高重要性的特征，模型对其依赖程度越高。该类方法的典型代表有 LIME^[500]、LEMNA^[501]、DeepLift^[502]和 SHAP^[503]等。

LIME 方法使用线性模型对解释黑盒模型进行局部代理，例如在文本或图像任务下，LIME 给出模型决策的依据主要源于句子中的哪些单词或图像中的哪些超像素块。在 LIME 基础上，LEMNA 方法优化了 LIME 的局部线性代理，同时在模型中整合了特征之间的依赖性信息。上述两种方法本质上是从前馈计算的角度进行特征归因，与之不同的是从反向传播角度进行特征归因的方法，例如 DeepLift，这种方法可以避免前馈计算的效率问题和饱和问题。以图像任务为例，其最终归因单元可以到像素级别。SHAP 方法则从博弈论利益分配的角度对参与决策的不同特征进行打分，该方法相比于前述的方法有更严格的数学理论证明。

8.2.3.2 基于全局的解释方法

基于全局的解释方法聚焦于提供针对整个模型的可解释性。该类方法往往借助于简单的可理解的代理模型来模拟复杂模型的行为，其代表工作有 Feature selector^[504]、MAME^[505]、ACE^[506]和 Global model on CEM^[507]等。

8.2.4 知识嵌入的可解释技术

基于知识或者规则对网络结构进行有约束的设计，本身就体现了一种直观的可解释思想，属于上述提到的内在可解释模型。在可解释人工智能的理论研究中，目前该类工作仍然较少，但是在很多应用领域如生物医学、物理等，这种结合具体领域知识的做法已经显示出优势。下文主要简介该方向的几个经典工作，以深度学习在生物领域的应用为主。

结合生物系统知识的神经网络具有独特优势，因为生物系统的层次性和复杂关系非常适合作为领域先验知识指导深度模型的设计和模型训练策略的制定。从单细胞内存储遗传物质的基因、功能各异的调控元件，到具有生物功能的蛋白质和中间代谢产物，再到细胞层面甚至生物个体层面的表型，它们共同构成很多基本模块进而形成复杂高级系统，例如基因调控网络、蛋白质互作网络和细胞通讯网络等。

已有部分工作尝试将这些知识融入特定的深度学习任务中，这些知识为模型的层次、信息流、甚至神经元赋予了具体的生物含义，使得神经网络由浅层到深层的计算与生命中由底层到表层的生命过程对应起来，因而从可解释角度使得模型具备了“不言自明”的性质。

例如，DCell^[508]主要研究基因互作与细胞生长速率的关系，将 Gene Ontology 数据库^[223, 509] (<http://geneontology.org/>) 中从基因到蛋白质到细胞器再到整个细胞的层级结构，作为构建深度神经网络的约束条件，从而实现对网络的知识嵌入。DSPN^[510]将基因调控网络嵌入到一个可解释的深度玻尔兹曼机，从而将全基因组关联研究的变异与基因联系起来，避免了从基因型直接预测表型，将从基因到表型的多种中间产物分别作为网络的中间层，实现了更准确的疾病预测，并且辅助发现了精神疾病中的关键基因和通路。P-NET^[511]通过在神经网络里逐层嵌入病人的病理数据以及基因、通路和生物过程的层级约束，实现了对前列腺癌患者治疗耐药状态的分层和评估，且该网络的可解

释性允许对关键基因预测能力进行打分，从而帮助研究人员发现了新的候选致癌基因，并得到了体外实验验证。

8.2.5 针对注意力机制能否提供可解释的辨析

在可解释领域，注意力机制能否为模型提供解释已经成为近年来关注的焦点。注意力机制能够根据任务需求选择最合适的输入，最早由 Bahdanau 作为一种软对齐的方式引入机器翻译任务^[512]，之后随着谷歌提出的完全基于注意力机制的 Transformer^[513]以及 Bert^[514]模型的广泛应用，注意力机制目前已经成为很多人工智能研究的标配。目前，注意力机制也被用于分析模型的可解释性，因为其功能和狭义的可解释（确认哪些输入内容对模型性能更重要）是重合的，并且注意力机制非常适合可视化，因此很多研究工作都以可视化注意力权重的方式反应模型透明度的提升。然而，这一做法往往会被批评没有对可解释性进行定义，学术界针对注意力机制能否提供可解释性展开了一场持续至今的辩论。

部分学者认为注意力机制不能用于提供解释，例如：Serrano 等人^[515]使用擦除中间表示的方式，发现注意力权重不能识别出最终输出所对应的输入部分，说明其解释力度偏弱；Jain 等人^[516]通过引入对抗注意力权重来测试模型输出的变化，发现存在完全不同的注意力权重使模型具有相同输出的情况。随之，也有学者对上述观点进行了反驳，认为注意力机制可以用于解释，例如：Wiegrefe 等人^[517]反驳 Jain 的实验设定，主要提出两点理由：（1）注意力机制需要与模型联合分析才具有意义，因为权重分布不是独立存在的，由于经历了前向和反向的计算过程，注意力权重是无法同整个模型割裂开的，因此，单纯使用对抗分布来否定注意力权重对于解释的贡献是没有意义的；（2）注意力权重提供解释的存在性并不蕴含排他性，也就是说，注意力权重只是提供一个解释而不保证提供唯一的解释，在计算方面这一点是

易于理解的，尤其对于中间表示的特征向量维度较高，而最终输出的类别个数较少的情况，适用于中间用于减少维度的映射函数可以有很大的灵活性。

到目前为止，关于注意力机制能否为模型提供解释仍是一个开放问题。大部分学者认同的观点是首先要对模型任务是否需要使用注意力机制进行分辨，如果任务本身过于简单，使用注意力机制无法带来显著性能提升，此时注意力权重的分布往往是不可预测的，对可解释的贡献也微乎其微；然而当注意力机制可以有效提升模型性能时，注意力本身能够为很多解释方法，如基于梯度、基于传播和基于遮挡的方法，提供很好的方向性参考。未来注意力机制和可解释技术能否突破现有概念壁垒，产生进一步的碰撞，值得我们瞩目以待。

8.3 可解释人工智能在药物设计中的应用

8.3.1 XAI 与定量构效关系 (QSAR)

作为一个应用统计数学方法，定量构效关系 (Quantitative structure-activity relationship, QSAR) 是对药物分子的化学结构与生物活性、毒性间的关系进行定量分析的模型。近几十年来，该方向已经积累了大量研究工作^[518-520]，相关的 QSAR 数学模型已发表很多，按照分子结构的维度不同，可以分为二维 QSAR 和三维 QSAR，分别基于分子的二维结构和三维空间结构。机器学习算法可以拟合出更为精准的 QSAR 模型，但不能明确给出回归方程的物理意义以及药物-受体间的作用模式。

目前在 QSAR 方面的可解释性可以分为以下几个方面：第一，基于先验知识的可解释策略例如 Hongming 等人和 Irene 等人分别提出基于 R 基团分解的 SVM 模型和基于 Rivality 指标的分类模型，试图将特定基团与活性关联起来^[521, 522]。第二，基于集成学习的可解释策略。例如 Chia-Hsiu 等人提出的基于集成学习的 QSPR 模型，兼顾了

高准确性和可解释性^[523]。其集成学习的基分类器也是随机森林、AdaBoost 等具有解释性的模型。该集成模型分别从几个基分类器的决策树重要特征排序方面进行解释。第三，基于注意力机制的可解释策略。例如 Pavel 等人提出的 Transformer-CNN 模型，使用内置于 Transformer 的自注意力机制，辨识出对于活性重要程度不同的子结构，从而实现可解释^[524]。同样基于注意力的另一项工作，注意到化学家对化合物的理解和语言使用者对单词的理解存在相似之处，因而将反应预测任务转换为语言体系里的翻译问题，目标是将表示反应物的文本序列映射到表示产物的文本序列^[525]。此外，QSAR 的可解释性一般会以对子结构着色等可视化方式展示^[526]。

另外，涉及不确定性估计的方法通过量化预测误差，也可以提供模型的解释。其中，已有一些将不确定性估计与 QSAR 相结合的方法。这些方法中有一部分是基于距离的，如 Sheridan 等人使用注意力和门控机制增强后的图神经网络，通过比较未知分子与训练集中已知各分子的距离来估计对该分子预测的不确定性^[527]；Liu 等人根据输入特征提出了一种基于分子相似性的领域适用性度量，该度量可以应用于包括深度神经网络在内的很多机器学习方法^[528]；Janet 等人将这种基于分子相似性的计算扩展到模型内部的隐层表示中，得到的度量具有优越的校准性能，并且适用于无机和有机化学^[529]。还有一些方法通过内在模型或事后的方式来处理不确定性，比如 Obrezanova 等人提出应用高斯过程来预测分子在吸收、代谢和排泄等方面的特性，该方法不需要限制模型先验参数且适用于大量分子描述符^[530]；Schroeter 等人使用多种机器学习方法预测包括 600 多种药物在内的大约 4,000 种化合物的水溶性^[531]。存在于化学结构空间之外的预测一般被认为是不可靠的，因此，不断有用于评估预测结果的可信度的新技术被提出，Bosc 等人聚焦保形预测（Conformal prediction），提出依赖于校准集先验知识的评估方法^[532]。

此外，有学者针对基于深度学习的 QSAR 类方法是否还需要考虑领域适用性的问题进行了专门的探讨，认为化学结构等约束与模型准确性的评估息息相关，因此，这些约束本身应该作为可解释模型的一部分^[533]。采用图的数据结构来表征这些分子描述符的约束是一种自然不过的做法，例如 Nembri 等人通过图卷积神经网络，将 QSAR 思想应用于筛选与细胞色素 P450 相关的药物分子等^[534]。

8.3.2 XAI 与联合用药

随着人们对癌症的了解以及认知加深，研究人员越来越多的关注于抗癌药物的发现和设计。然而，由于长期服药导致出现耐药性，单一药物治疗特定疾病的效果存在局限性。因此，越来越多的人采用多类药物混合治疗的方法，即联合用药^[535-537]。联合用药已广泛用于一些疾病的治疗并取得了成功，如艾滋病、真菌或细菌感染^[538-540]。随着药物数量的迅速增加又产生了新的问题：可能的联合药物组合变得极多，对所有可能的药物组合进行试验是不现实的。深度学习技术的应用极大地提高了联合药物筛选的效率^[541-545]。但是，由于生物知识无法完全融入深度学习模型，许多联合药物筛选的计算模型缺乏可解释性、透明性，极大限制了它们的临床应用^[546, 547]。可解释人工智能技术为应对这一挑战带来新的机遇。一方面，研究表明，基因互作关系、基因必需性以及药靶互作关系是影响联合药物效果的重要因素。TranSynergy 模型^[548]采用自注意力机制的增强深度学习技术对药靶关系、基因互作关系和基因必需性信息进行建模，并通过沙普利加性基因富集分析方法，挖掘了与联合药物作用相关的新基因，提高了联合药物预测的性能和可解释性。此外，药物组合产生的副作用也是联合用药面临的重要挑战，药物互作关系具有典型的图结构特征，Decagon 方法^[415]采用图卷积神经网络模型对多模态药物互作关系进行建模，准确预测了药物组合的副作用。另一方面，反向蛋白质组学

技术 (Reverse phase protein array, RPPA) 作为一种有别于常规蛋白组学的检测方法, 可以实现上千例样本的几百种靶点平行比对, 能够为联合用药研究提供重要解决方案。通过在 RPPA 靶向蛋白质组学数据上建立机器学习计算框架, CellBox 模型^[549]有效实现了靶向药物组合效果和功能预测。该模型不但具有精准度高、抗噪性能优异、拓展性强等优势, 还具有两方面的可解释性: 透明性和可追溯性。一方面, CellBox 采用定义好的具有生物可解释的微分方程 (Ordinary differential equation, ODE) 数学模型, 在 ODE 模型中, 每个参数代表细胞成分或表型数量之间的直接且定量的互作关系。另一方面, 通过在细胞中给定扰动, ODE 模型能够给出该扰动如何在有向网络中进行传播, 进而提出细胞反应的机制假设。CellBox 模型是可解释人工智能在联合用药研究中的重要探索。

8.3.3 XAI 与分子属性预测

药物发现领域中, 分子属性预测是一项基本任务, 深度学习技术的发展大大加快了寻找候选药物的速度、减少了候选药物挖掘的成本^[550]。然而, 由于现有的深度学习计算模型存在一大挑战, 即难以保证高精度的同时, 模型具有可解释性。如果无法解释分子属性预测模型的分析结果, 药物学家很难相信某个预测算法给出的“武断”决策, 进而投入巨大的资金进行药物的后期研发^[551]。将分子表示成图, 利用图神经网络进行分子属性预测能够一定程度上兼具预测准确性和可解释性。研究人员因此开发出 Attentive FP 模型^[474]用于药物发现, 该模型首先用图表示一个分子, 然后模型中使用注意力机制以有效提取图的局部和非局部特征, 以及远距离节点相互作用, 实现了针对特定领域的分子结构的非局部特性学习, 具有一定的可解释性。这一注意力机制的添加, 有助于药物学家或化学家直接从各种属性数据中挖掘出分子结构更深层的知识, 超越经验和直觉。基于官能团的图自监督

学习方法 MGSSL^[552]同时考虑了原子层级和官能团层级的自监督任务,使得预训练得到的图神经网络可以有效捕捉分子图中官能团的结构和语义信息,提升下游分子属性预测的性能,并能够生成官能团 motif 树,这些官能团信息是神经网络学习的重要信息。

模型的不确定性评估是模型可解释性研究的重要内容^[475],在很多高风险应用中具有重要意义。对不确定性的定量同时也是分子属性预测问题的重要研究内容之一^[553]。结合贝叶斯策略和半监督图卷积神经网络能够实现分子属性的不确定性校准预测^[554],例如熔点和水溶性。通过选择不确定性最大的分子,形成一套主动学习方法,并能够有效应用于不同化合物的研究。除此之外,还有一系列诸如随机森林、映射指纹和机器学习集成等与神经网络兼容的不确定性定量技术,这些技术在分子属性预测领域具有越来越大的应用价值。然而大量实验表明,尚没有一种不确定性定量技术能够在所有性能指标和数据集中具有显著优势^[553],所以需要更多研究成果推动该领域的发展。

8.3.4 XAI 与药靶互作

药物-靶标的相互作用关系预测,在药物发现过程中至关重要。在新药研发和药物重定位中,基于疾病相关的靶标筛选新的候选药物,以及发现已知药物-已知靶标的新关联关系,均属于这一类别。随着人工智能技术的发展,基于药物结构、细胞反应等智能模型使药物-靶标相互作用预测的准确性逐渐提升^[555,556]。药物-靶标相互作用预测通常包含药物-靶标关联关系预测和药物-靶标亲和力预测两个方面,分别被建模为二分类和回归问题^[557]。但大部分研究均基于不可解释的深度学习黑盒模型,即使有较高的准确率,仍较难取信于药物研发者。

近期,针对药物-靶标相互作用这一预测任务,可解释性预测模型受到了研究人员的广泛关注,它们可以同时提升模型精度和可解释性。这类模型的解释性主要是通过赋权药物和靶标作用的重要基团体现

的。Gao 等人基于注意力池化网络构建药物-靶标关联关系二分类预测方法，在药物和靶蛋白两个通道的表征学习器中加入了注意力机制，实现了一定程度的可解释性^[558]。相较于关联关系二分类预测，可解释人工智能在药物-靶标亲和力预测方面的研究更显深入。DeepAffinity 通过单独和联合的注意力机制混合模型，提高亲和力预测的可解释性，该方法也成为亲和力预测领域中的一项极具参考价值的研究^[559]。ML-DTI 通过基于多头注意力和位置感知注意力机制的互学习模型，既实现药物和靶标表征之间的交互学习，也实现了学习过程中的可解释^[560]。Brighter 等人提出的基于多视图自注意力机制的药物-靶标亲和力预测模型，在不同视图上利用自注意力机制实现了可解释性^[561]。

8.3.5 XAI 与药物不良反应预测

药物不良反应是药物研发失败的主要原因之一，已经成为一个重要的公共卫生问题。随着人工智能技术的发展，一系列基于机器学习和深度学习的精准预测药物反应方法被提出^[562-564]，为这一领域研究提供了新的技术支撑。近年来，研究人员在提升药物反应预测算法性能的同时，也聚焦于将可解释人工智能技术应用于药物反应预测研究，尝试挖掘用于表示药物分子的关键特征集合，并解释这些关键特征如何影响药物不良反应。例如 IBM 健康计算中心研究团队开发的可解释深度学习计算框架^[565]不但能够准确预测药物不良反应，同时通过在神经网络模型中部署注意力机制的方式，不良反应具有重要意义。药物不良反应是药物化学结合和患者生物系统之间复杂相互作用的结果。受这一思想启发，基于结构学习的因果分析模型 CASTLE^[563]通过融合药物化学信息和生物属性，将因果特征筛选问题转化为贝叶斯网络结构学习中的等价父子发现这一经典问题，实现药物不良反应的关键因子识别，并在 12 种器官特异的药物不良反应中取得了优异

性能。药物毒性反应将导致严重功能紊乱和器质损害，是严重的药物不良反应。分子图编码的卷积神经网络框架 MGE-CNN^[566]建立了 3 种高质量的急性口服毒性预测模型，包括回归模型、多分类模型和多任务模型。通过对分子指纹进行正向、反向探索，该模型相比于传统模型更加支持浅层机器学习方法，提升了模型的可解释性。此外，通过自动特征学习，MGE-CNN 框架能够将相应的激活值映射到片段空间，进而挖掘出与急性口服毒性相关的化学结构。

8.3.6 XAI 与新药设计

目前，可解释人工智能方法在新药设计方面并没有相关研究。我们认为，XAI 应用于上述几个领域后，将会在以下两个方面推动可解释的新药理性设计：

第一，通过基于 XAI 的分子属性预测研究，揭示高活性基团与特定属性（如低肝毒性等）强关联的子结构，并将其作为先验知识或者约束条件，整合进新药设计模型中；

第二，通过基于 XAI 的药物-靶标相互作用预测和药物反应预测等研究，揭示高活性基团或子结构与靶蛋白结合位点、细胞反应和表型等关联关系，也可将其作为先验知识或者约束条件，整合进新药设计模型中。

8.4 可解释人工智能在药物发现中的前景展望

追求人工智能的全面可解释是一项难度很大的挑战，而具体到药物设计中的可解释人工智能更是任重道远^[567]。如何设计对照试验、论证智能驱动下的种种假设^[568]对于提升人工智能的可靠性具有关键意义。现阶段虽然已经发展出多种解释手段，但其中大多数都是针对特定任务量身定制的，这种针对性解决方案在提升性能的同时，也提高了可解释技术的通用门槛。因此，开发基于统一解释框架、兼容药

物设计需求的可解释人工智能技术仍然是一项巨大挑战。

理想情况下，药物设计所采用的可解释人工智能应该具备足够高的透明度、充分的决策公平、丰富的信息量和恰当的不确定性估计等特性，但是目前仍然缺乏同时兼顾这些特征的良好方法^[475]，而且大部分基于深度学习的药物发现工作都没有考虑领域适用性的限制^[569, 570]，也就是模型学习所搜索的假设空间应该满足现有知识框架下的化学限制，而这些限制本身应当被视为可解释本身的组成部分^[533]。所以，短期内的解决策略仍是以智囊团或专家知识与人工智能应用个例相结合的方式为主。深入了解药物设计领域的系统知识将大有裨益，人工智能的决策依据需要提供足够的信息量，而这些领域知识往往可以决定哪些模型决策需要进一步解释，哪些解释对于用户而言可能成为创新的来源，哪些解释是意料之中甚至是不必要的^[571]。目前，这样的解释方案的实现还需要不同领域的专家们的共同努力。借助交叉融合，贯通领域知识是必经之路。

总体而言，基于人工智能的药物设计模型现已形成了一套用于描述模型决策空间的语言体系，深入理解该语言体系有利于进一步发掘其内涵并且了解其局限性。建立具有明确化学含义且适用于机器学习的简单分子表征，如氨基酸序列^[572, 573]、三维空间表示^[574, 575]等，已经被证明是一种可行的方式。已有一部分工作依赖于先进的分子表征，如哈希二进制指纹^[576, 577]、拓扑化学几何描述^[578]等，这些表征可以把药物设计需要遵循的结构特征作为先验知识。广泛运用易嵌化学知识的分子表征是实现可解释人工智能的一种直观思路。因此，发展适应人工智能的可解释分子表示将会是未来药物设计的重要研究内容，包括如何克服现有分子表征中信息量与可理解度之间的矛盾。从长期来看，这种将领域知识与先进模型相结合的策略，也将为药物设计中的生化模拟带来巨大优势。

8.5 本章小节

现阶段，药物发现中的可解释人工智能已经取得了初步进展，基于不同原理的可解释方法都在为“以人为本的可解释”目标提供可行的探索方向。该领域的可解释人工智能技术具有试错成本高、迭代周期长等现实挑战，因此，需要结合药物设计中大量积累的专家知识。然而，具体到可解释方法设计中，如何对先验知识进行提取、抽象和运用，决定了先验知识能否有效辅助发现新知识。当前，可解释人工智能所学习到的信息已有部分可以解析为化学家和生物学家可理解的知识，但从整体而言，这种可解释程度距离人类认知还相去甚远。由于药物发现的风险敏感性，人类难以向不确定性做出妥协，因此，这条认知鸿沟注定要以机器走向人类的方式得以解决，这也意味着科学家需要以超越经验和直觉的方式，站在人工智能的角度，反向回溯药物发现的知识需求。可以预料，如此的需求，将会吸引跨领域专家打破领域知识的结构壁垒，发挥交叉融合的最大优势。在日趋平权化的这一领域，可解释人工智能将加速降低药物发现的技术门槛和推动新一轮知识涌现。

参考文献

- [1] Sharaf R, Montesion M, Hopkins JF et al. A pan-cancer landscape of telomeric content shows that RAD21 and HGF alterations are associated with longer telomeres, *Genome Medicine* 2021,184(1):792-809.e723.
- [2] Huang A, Garraway LA, Ashworth A et al. Synthetic lethality as an engine for cancer drug target discovery, *Nature Reviews Drug Discovery* 2020,19(1):23-38.
- [3] Hahn WC, Bader JS, Braun TP et al. An expanded universe of cancer targets, *Cell* 2021,184(5):1142-1155.
- [4] Vucic EA, Thu KL, Robison K et al. Translating cancer ‘omics’ to improved outcomes, *Genome Research* 2012,22(2):188-195.
- [5] Yu K-H, Snyder M. Omics Profiling in Precision Oncology*, *Molecular & Cellular Proteomics* 2016,15(8):2525-2536.
- [6] Ballester PJ, Carmona J. Artificial intelligence for the next generation of precision oncology, *npj Precision Oncology* 2021,5(1):1-3.
- [7] Computational strategies for single-cell multi-omics integration, *Computational and Structural Biotechnology Journal* 2021,19:2588-2596.
- [8] Cai Z, Poulos RC, Liu J et al. Machine learning for multi-omics data integration in cancer, *iScience* 2022,25(2):103798.
- [9] Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities, *Experimental & Molecular Medicine* 2020,52(9):1452-1465.
- [10] Kashima Y, Sakamoto Y, Kaneko K et al. Single-cell sequencing techniques from individual to multiomics analyses, *Experimental & Molecular Medicine* 2020,52(9):1419-1427.
- [11] Suzuki Y. *Single Molecule and Single Cell Sequencing*. Springer, 2019.
- [12] Tang F, Barbacioru C, Wang Y et al. mRNA-Seq whole-transcriptome analysis of a single cell, *Nature methods* 2009,6(5):377-382.
- [13] Macosko E, Goldman M. Drop-Seq Laboratory Protocol 2015:19.
- [14] Klein AM, Mazutis L, Akartuna I et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells, *Cell* 2015,161(5):1187-1201.
- [15] Zheng F, Zhang W, Chu X et al. Genome sequencing of strain *Cellulosimicrobium* sp. TH-20 with ginseng biotransformation ability, *3 Biotech* 2017,7(4):237.
- [16] Casasent AK, Schalck A, Gao R et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing, *Cell* 2018,172(1-2):205-217. e212.
- [17] Tran HTN, Ang KS, Chevrier M et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data, *Genome Biology* 2020,21(1):12.
- [18] Huang M, Wang J, Torre E et al. SAVER: gene expression recovery for single-cell RNA sequencing, *Nat Methods* 2018,15(7):539-542.
- [19] Dijk Dv, Nainys J, Sharma R et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data, *BioRxiv* 2017:111591.
- [20] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data, *Nat Commun* 2018,9(1):997.
- [21] Gong W, Kwak I-Y, Pota P et al. DrImpute: imputing dropout events in single cell RNA sequencing data, *BMC Bioinformatics* 2018,19(1).
- [22] Talwar D, Mongia A, Sengupta D et al. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data, *Sci Rep* 2018,8(1):16329.
- [23] Waddington CH. The Epigenotype, *International Journal of Epidemiology* 2012,41(1):10-13.
- [24] Akhavan-Niaki H, Samadani AA. DNA Methylation and Cancer Development: Molecular Mechanism, *Cell Biochemistry and Biophysics* 2013,67(2):501-513.
- [25] Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer, *Nature Reviews Genetics* 2002,3(6):415-428.
- [26] Jones PA, Laird PW. Cancer-epigenetics comes of age, *Nature Genetics* 1999,21(2):163-167.
- [27] Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer, *Nature Reviews Genetics* 2006,7(1):21-33.
- [28] Jones PA, Martienssen R. A Blueprint for a Human Epigenome Project: The AACR Human Epigenome Workshop, *Cancer Research* 2005,65(24):11241-11246.

- [29] Yoo CB, Jones PA. Epigenetic therapy of cancer: past, present and future, *Nature Reviews Drug Discovery* 2006,5(1):37-50.
- [30] Landt SG, Marinov GK, Kundaje A et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, *Genome Research* 2012,22(9):1813-1831.
- [31] Buenrostro JD, Wu B, Chang HY et al. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide, *Current Protocols in Molecular Biology* 2015,109(1):21.29.21-21.29.29.
- [32] Berkum NLv, Lieberman-Aiden E, Williams L et al. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes., *JoVE (Journal of Visualized Experiments)* 2010(39):e1869.
- [33] Meissner A, Gnirke A, Bell GW et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis, *Nucleic Acids Research* 2005,33(18):5868-5877.
- [34] Cokus SJ, Feng S, Zhang X et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, *Nature* 2008,452(7184):215-219.
- [35] Arslan E, Schulz J, Rai K. Machine Learning in Epigenomics: Insights into Cancer Biology and Medicine, *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 2021,1876(2):188588.
- [36] Holzinger A, Jurisica I. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In: Holzinger A., Jurisica I. eds). *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges*. Berlin, Heidelberg: Springer, 2014, 1-18.
- [37] Merkel A, Esteller M. Experimental and Bioinformatic Approaches to Studying DNA Methylation in Cancer, *Cancers* 2022,14(2):349.
- [38] Aslibekyan S, Claas SA, Arnett DK. Clinical applications of epigenetics in cardiovascular disease: the long road ahead, *Translational Research* 2015,165(1):143-153.
- [39] Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation, *Nature Reviews Genetics* 2016,17(9):551-565.
- [40] Fuks F. DNA methylation and histone modifications: teaming up to silence genes, *Current Opinion in Genetics & Development* 2005,15(5):490-495.
- [41] Jones PA, Takai D. The Role of DNA Methylation in Mammalian Epigenetics, *Science* 2001,293(5532):1068-1070.
- [42] Wilting SM, van Boerdonk RAA, Henken FE et al. Methylation-mediated silencing and tumour suppressive function of hsa-miR-124 in cervical cancer, *Molecular Cancer* 2010,9(1):167.
- [43] Zhao G, Liu X, Liu Y et al. Aberrant DNA Methylation of SEPT9 and SDC2 in Stool Specimens as an Integrated Biomarker for Colorectal Cancer Early Detection, *Frontiers in Genetics* 2020,11.
- [44] Xiufang WU, Qiong NaN, Xiaohong Z et al. Diagnostic Value of Plasma SEPT9 Methylation Test for Colorectal Cancer, *Chinese General Practice* 2021,24(15):1915-1919.
- [45] Pfeifer GP, Yoon J-H, Liu L et al. Methylation of the RASSF1A Gene in Human Cancers 2002,383(6):907-914.
- [46] Huang S, Cai N, Pacheco PP et al. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics, *Cancer Genomics & Proteomics* 2018,15(1):41-51.
- [47] Yagi K, Akagi K, Hayashi H et al. Three DNA Methylation Epigenotypes in Human Colorectal Cancer, *Clinical Cancer Research* 2010,16(1):21-33.
- [48] Feng P, Chen W, Lin H. Prediction of CpG island methylation status by integrating DNA physicochemical properties, *Genomics* 2014,104(4):229-233.
- [49] Tian Q, Zou J, Tang J et al. MRCNN: a deep learning model for regression of genome-wide DNA methylation, *BMC Genomics* 2019,20(2):192.
- [50] Wang Z, Wang Y. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders, *BMC Bioinformatics* 2019,20(18):568.
- [51] Paul JK, Iype T, R D et al. Characterization of fibromyalgia using sleep EEG signals with nonlinear dynamical features, *Computers in Biology and Medicine* 2019,111:103331.
- [52] Seravalli V, Miller JL, Blitzer MG et al. A comparison of first trimester blood pressures obtained at the time of first trimester pre-eclampsia screening and those obtained during prenatal care visits, *European Journal of Obstetrics & Gynecology and Reproductive Biology* 2020,248:77-80.
- [53] Yang X, Zhang Z, Zhang L et al. MicroRNA hsa-mir-3923 serves as a diagnostic and prognostic

- biomarker for gastric carcinoma, *Scientific Reports* 2020,10(1):4672.
- [54] Capper D, Jones DTW, Sill M et al. DNA methylation-based classification of central nervous system tumours, *Nature* 2018,555(7697):469-474.
- [55] Cai J, Xu Y, Zhang W et al. A comprehensive comparison of residue-level methylation levels with the regression-based gene-level methylation estimations by ReGear, *Briefings in Bioinformatics* 2021,22(4):bbaa253.
- [56] Schep AN, Wu B, Buenrostro JD et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data, *Nature Methods* 2017,14(10):975-978.
- [57] Zamanighomi M, Lin Z, Daley T et al. Unsupervised clustering and epigenetic classification of single cells, *Nature Communications* 2018,9(1):2410.
- [58] Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT, *Bioinformatics* 2017,33(18):2930-2932.
- [59] Cusanovich DA, Hill AJ, Aghamirzaie D et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility, *Cell* 2018,174(5):1309-1324.e1318.
- [60] Bravo González-Blas C, Minnoye L, Papasokrati D et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data, *Nature Methods* 2019,16(5):397-400.
- [61] Pliner HA, Packer JS, McFaline-Figueroa JL et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data, *Molecular Cell* 2018,71(5):858-871.e858.
- [62] Baker SM, Rogerson C, Hayes A et al. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool, *Nucleic Acids Research* 2019,47(2):e10.
- [63] Fang R, Preissl S, Li Y et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC, *Nature Communications* 2021,12(1):1337.
- [64] Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods, *Experimental & Molecular Medicine* 2020,52(9):1428-1442.
- [65] John CR, Watson D, Barnes MR et al. Spectrum: fast density-aware spectral clustering for single and multi-omic data, *Bioinformatics* 2020,36(4):1159-1166.
- [66] Argelaguet R, Arnol D, Bredikhin D et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data, *Genome Biology* 2020,21(1):111.
- [67] Duren Z, Chen X, Zamanighomi M et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations, *Proceedings of the National Academy of Sciences* 2018,115(30):7723-7728.
- [68] Welch JD, Kozareva V, Ferreira A et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity, *Cell* 2019,177(7):1873-1887.e1817.
- [69] Cao K, Bai X, Hong Y et al. Unsupervised topological alignment for single-cell multi-omics integration, *Bioinformatics* 2020,36(Supplement_1):i48-i56.
- [70] Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona, *Bioinformatics* 2022,38(1):211-219.
- [71] Campbell KR, Steif A, Laks E et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers, *Genome Biology* 2019,20(1):54.
- [72] Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding, *Nature Biotechnology* 2022:1-9.
- [73] Stanojevic S, Li Y, Garmire LX. Computational Methods for Single-Cell Multi-Omics Integration and Alignment:26.
- [74] Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data, *Briefings in Bioinformatics* 2021,22(4):bbaa287.
- [75] Gayoso A, Steier Z, Lopez R et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI, *Nature Methods* 2021,18(3):272-282.
- [76] Hao Y, Hao S, Andersen-Nissen E et al. Integrated analysis of multimodal single-cell data, *Cell* 2021,184(13):3573-3587.e3529.
- [77] Pan Y, Kadash-Edmondson KE, Wang R et al. RNA Dysregulation: An Expanding Source of Cancer Immunotherapy Targets, *Trends in Pharmacological Sciences* 2021,42(4):268-282.
- [78] Agrawal AA, Yu L, Smith PG et al. Targeting splicing abnormalities in cancer, *Current Opinion in Genetics & Development* 2018,48:67-74.
- [79] Vo JN, Cieslik M, Zhang Y et al. The Landscape of Circular RNA in Cancer, *Cell* 2019,176(4):869-881.e813.

- [80] Li Y, Choi PS, Casey SC et al. MYC through miR-17-92 Suppresses Specific Target Genes to Maintain Survival, Autonomous Proliferation, and a Neoplastic State, *Cancer Cell* 2014,26(2):262-272.
- [81] Hanna J, Hossain GS, Kocerha J. The Potential for microRNA Therapeutics and Clinical Research, *Frontiers in Genetics* 2019,10.
- [82] van Zandwijk N, Pavlakis N, Kao SC et al. Safety and activity of microRNA-loaded minicells in patients with recurrent malignant pleural mesothelioma: a first-in-man, phase 1, open-label, dose-escalation study, *The Lancet Oncology* 2017,18(10):1386-1396.
- [83] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation, *Cell* 2011,144(5):646-674.
- [84] Oltean S, Bates DO. Hallmarks of alternative splicing in cancer, *Oncogene* 2014,33(46):5311-5318.
- [85] Turajlic S, Litchfield K, Xu H et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis, *The Lancet Oncology* 2017,18(8):1009-1021.
- [86] Smart AC, Margolis CA, Pimentel H et al. Intron retention is a source of neoepitopes in cancer, *Nature Biotechnology* 2018,36(11):1056-1058.
- [87] Shen L, Zhang J, Lee H et al. RNA Transcription and Splicing Errors as a Source of Cancer Frameshift Neoantigens for Vaccines, *Scientific Reports* 2019,9(1):14184.
- [88] Oka M, Xu L, Suzuki T et al. Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer, *Genome Biology* 2021,22(1):9.
- [89] Gohil SH, Iorgulescu JB, Braun DA et al. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy, *Nature Reviews Clinical Oncology* 2021,18(4):244-256.
- [90] Ren L, Li J, Wang C et al. Single cell RNA sequencing for breast cancer: present and future, *Cell Death Discovery* 2021,7(1):1-11.
- [91] Zheng C, Zheng L, Yoo J-K et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing, *Cell* 2017,169(7):1342-1356.e1316.
- [92] Guo X, Zhang Y, Zheng L et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing, *Nature Medicine* 2018,24(7):978-985.
- [93] Vento-Tormo R, Efremova M, Botting RA et al. Single-cell reconstruction of the early maternal-fetal interface in humans, *Nature* 2018,563(7731):347-353.
- [94] Smillie CS, Biton M, Ordovas-Montanes J et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis, *Cell* 2019,178(3):714-730.e722.
- [95] Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes, *Nature Methods* 2020,17(2):159-162.
- [96] Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation, *Cell* 1980,20(1):85-93.
- [97] Bates SE. Epigenetic Therapies for Cancer, *New England Journal of Medicine* 2020,383(7):650-663.
- [98] Goepfert B, Toth R, Singer S et al. Integrative Analysis Defines Distinct Prognostic Subgroups of Intrahepatic Cholangiocarcinoma, *Hepatology* 2019,69(5):2091-2106.
- [99] Johann PD, Erkek S, Zapatka M et al. Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes, *Cancer Cell* 2016,29(3):379-393.
- [100] Sahm F, Schrimpf D, Stichel D et al. DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis, *The Lancet Oncology* 2017,18(5):682-694.
- [101] Duruisseaux M, Martínez-Cardús A, Calleja-Cervantes ME et al. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: a multicentre, retrospective analysis, *The Lancet Respiratory Medicine* 2018,6(10):771-781.
- [102] Seligson DB, Horvath S, Shi T et al. Global histone modification patterns predict risk of prostate cancer recurrence, *Nature* 2005,435(7046):1262-1266.
- [103] Ellinger J, Schneider A-C, Bachmann A et al. Evaluation of Global Histone Acetylation Levels in Bladder Cancer Patients, *Anticancer Research* 2016,36(8):3961-3964.
- [104] Jurmeister P, Bockmayr M, Seegerer P et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck

- metastases, *Science Translational Medicine* 2019,11(509):eaaw8513.
- [105] Moran S, Martínez-Cardús A, Sayols S et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis, *The Lancet Oncology* 2016,17(10):1386-1395.
- [106] Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model, *PLOS ONE* 2020,15(5):e0226461.
- [107] Paluszczak J, Baer-Dubowska W. [Epigenome and cancer: new possibilities of cancer prevention and therapy?], *Postepy biochemii* 2005,51(3):244-250.
- [108] Xu WS, Parmigiani RB, Marks PA. Histone deacetylase inhibitors: molecular mechanisms of action, *Oncogene* 2007,26(37):5541-5552.
- [109] Chen S, Wang Y, Zhou W et al. Identifying Novel Selective Non-Nucleoside DNA Methyltransferase 1 Inhibitors through Docking-Based Virtual Screening, *Journal of Medicinal Chemistry* 2014,57(21):9028-9041.
- [110] Mao R, Shao J, Zhu K et al. Potent, Selective, and Cell Active Protein Arginine Methyltransferase 5 (PRMT5) Inhibitor Developed by Structure-Based Virtual Screening and Hit Optimization, *Journal of Medicinal Chemistry* 2017,60(14):6289-6304.
- [111] Johnson DS, Mortazavi A, Myers RM et al. Genome-Wide Mapping of in Vivo Protein-DNA Interactions, *Science* 2007,316(5830):1497-1502.
- [112] Li G-B, Yang L-L, Yuan Y et al. Virtual screening in small molecule discovery for epigenetic targets, *Methods* 2015,71:158-166.
- [113] Jin F, Gao D, Wu Q et al. Exploration of N-(2-aminoethyl)piperidine-4-carboxamide as a potential scaffold for development of VEGFR-2, ERK-2 and Abl-1 multikinase inhibitor, *Bioorganic & Medicinal Chemistry* 2013,21(18):5694-5706.
- [114] Wang B, Mezlini AM, Demir F et al. Similarity network fusion for aggregating data types on a genomic scale, *Nature Methods* 2014,11(3):333-337.
- [115] Jiang Y-Z, Ma D, Suo C et al. Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies, *Cancer Cell* 2019,35(3):428-440.e425.
- [116] Dimitrakopoulos C, Hindupur SK, Häfliger L et al. Network-based integration of multi-omics data for prioritizing cancer genes, *Bioinformatics* 2018,34(14):2441-2448.
- [117] Dimitrakopoulos C, Hindupur SK, Colombi M et al. Multi-omics data integration reveals novel drug targets in hepatocellular carcinoma, *BMC Genomics* 2021,22(1):592.
- [118] Guo L-Y, Wu A-H, Wang Y-x et al. Deep learning-based ovarian cancer subtypes identification using multi-omics data, *BioData Mining* 2020,13(1):10.
- [119] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, *Bioinformatics* 2009,25(22):2906-2912.
- [120] Shen Y, Xiong W, Gu Q et al. Multi-Omics Integrative Analysis Uncovers Molecular Subtypes and mRNAs as Therapeutic Targets for Liver Cancer, *Frontiers in Medicine* 2021,8.
- [121] Fan X, Lu P, Wang H et al. Integrated single-cell multiomics analysis reveals novel candidate markers for prognosis in human pancreatic ductal adenocarcinoma, *Cell Discovery* 2022,8(1):1-16.
- [122] Granja JM, Klemm S, McGinnis LM et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia, *Nature Biotechnology* 2019,37(12):1458-1465.
- [123] Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice, *Nature Reviews Immunology* 2020,20(11):651-668.
- [124] Robert C. A decade of immune-checkpoint inhibitors in cancer therapy, *Nature Communications* 2020,11(1):3801.
- [125] Mikkilineni L, Kochenderfer JN. CAR T cell therapies for patients with multiple myeloma, *Nature Reviews Clinical Oncology* 2021,18(2):71-84.
- [126] Yi JS, Cox MA, Zajac AJ. T-cell exhaustion: characteristics, causes and conversion, *Immunology* 2010,129(4):474-481.
- [127] Cinier J, Hubert M, Besson L et al. Recruitment and Expansion of Tregs Cells in the Tumor Environment—How to Target Them?, *Cancers* 2021,13(8):1850.
- [128] Koyama S, Nishikawa H. Mechanisms of regulatory T cell infiltration in tumors: implications for innovative immune precision therapies, *Journal for ImmunoTherapy of Cancer*

2021,9(7):e002591.

[129] Dixit A, Parnas O, Li B et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens, *Cell* 2016,167(7):1853-1866.e1817.

[130] Frangieh CJ, Melms JC, Thakore PI et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion, *Nature Genetics* 2021,53(3):332-341.

[131] Zheng Y, Tang L, Liu Z. Multi-omics analysis of an immune-based prognostic predictor in non-small cell lung cancer, *BMC Cancer* 2021,21(1):1322.

[132] Zhao Q, Sun Y, Liu Z et al. CrossICC: iterative consensus clustering of cross-platform gene expression data without adjusting batch effect, *Briefings in Bioinformatics* 2020,21(5):1818-1824.

[133] Hu R, Tao T, Yu L et al. Multi-Omics Characterization of Tumor Microenvironment Heterogeneity and Immunotherapy Resistance Through Cell States-Based Subtyping in Bladder Cancer, *Frontiers in Cell and Developmental Biology* 2022,9.

[134] Ståhl PL, Salmén F, Vickovic S et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics, *Science* 2016,353(6294):78-82.

[135] Chen A, Liao S, Cheng M et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays, *Cell* 2022,0(0).

[136] Qiu Q, Hu P, Qiu X et al. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq, *Nature Methods* 2020,17(10):991-1001.

[137] Sneader W. *Drug discovery: a history*. John Wiley & Sons, 2005.

[138] Rifaioglu AS, Atas H, Martin MJ et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, *Briefings in bioinformatics* 2019,20(5):1878-1912.

[139] Sun M, Zhao S, Gilvary C et al. Graph convolutional networks for computational drug development and discovery, *Briefings in bioinformatics* 2020,21(3):919-935.

[140] Paul SM, Mytelka DS, Dunwiddie CT et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nature reviews Drug discovery* 2010,9(3):203-214.

[141] Avorn J. The \$2.6 billion pill—methodologic and policy considerations, *N Engl J Med* 2015,372(20):1877-1879.

[142] Du B-X, Qin Y, Jiang Y-F et al. Compound–protein interaction prediction by deep learning: Databases, descriptors and models, *Drug Discovery Today* 2022,27(5):1350-1366.

[143] Macarron R, Banks MN, Bojanic D et al. Impact of high-throughput screening in biomedical research, *Nature reviews Drug discovery* 2011,10(3):188-195.

[144] Schneider G. Virtual screening: an endless staircase?, *Nature reviews Drug discovery* 2010,9(4):273-276.

[145] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *Journal of computational chemistry* 2010,31(2):455-461.

[146] Jr FRS. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery, *Current Opinion in Pharmacology* 2010,10(6):738-744.

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

[148] Lee RS. Natural language processing. *Artificial Intelligence in Daily Life*. Springer, 2020, 157-192.

[149] Feng X, Jiang Y, Yang X et al. Computer vision algorithms and hardware implementations: A survey, *Integration* 2019,69:309-320.

[150] Zeng X, Zhu S, Lu W et al. Target identification among known drugs by deep learning from heterogeneous networks, *Chemical Science* 2020,11(7):1775-1797.

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

[152] Heng T, Yang D, Wang R et al. Progress in Research on Artificial Intelligence Applied to Polymorphism and Cocrystal Prediction, *ACS Omega* 2021,6(24):15543-15550.

[153] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery, *Drug Discovery Today* 2019,24(5):1157-1165.

[154] Paul D, Sanap G, Shenoy S et al. Artificial intelligence in drug discovery and development, *Drug Discovery Today* 2021,26(1):80.

- [155] 刘晓凡, 孙翔宇, 朱迅. 人工智能在新药研发中的应用现状与挑战, *药学进展* 2021,45(7):8.
- [156] Tian K, Shao M, Wang Y et al. Boosting compound-protein interaction prediction by deep learning, *Methods* 2016,110:64-72.
- [157] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction, *Bioinformatics* 2018,34(21):3666-3674.
- [158] Zhang L, Tan J, Han D et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery, *Drug Discovery Today* 2017,22(11):1680-1685.
- [159] Carpenter KA, Cohen DS, Jarrell JT et al. Deep learning and virtual drug screening, *Future medicinal chemistry* 2018,10(21):2557-2567.
- [160] Karimi M, Wu D, Wang ZY et al. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks, *Bioinformatics* 2019,35(18):3329-3338.
- [161] Li S, Wan F, Shu H et al. MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities, *Cell Systems* 2020,10(4):308-322. e311.
- [162] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery, *arXiv preprint arXiv:1510.02855* 2015.
- [163] Szklarczyk D, Santos A, von Mering C et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data, *Nucleic acids research* 2016,44(D1):D380-D384.
- [164] Gilson MK, Liu T, Baitaluk M et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic acids research* 2016,44(D1):D1045-D1053.
- [165] Su M, Yang Q, Du Y et al. Comparative assessment of scoring functions: the CASF-2016 update, *Journal of chemical information and modeling* 2018,59(2):895-913.
- [166] Smith RD, Clark JJ, Ahmed A et al. Updates to binding MOAD (mother of all databases): polypharmacology tools and their utility in drug repurposing, *Journal of molecular biology* 2019,431(13):2423-2433.
- [167] Tang J, Szwajda A, Shakyawar S et al. Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis, *Journal of Chemical Information and Modeling* 2014,54(3):735-743.
- [168] Davis MI, Hunt JP, Herrgard S et al. Comprehensive analysis of kinase inhibitor selectivity, *Nature biotechnology* 2011,29(11):1046-1051.
- [169] Mysinger MM, Carchia M, Irwin JJ et al. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, *Journal of medicinal chemistry* 2012,55(14):6582-6594.
- [170] Kanehisa M, Furumichi M, Sato Y et al. KEGG: integrating viruses and cellular organisms, *Nucleic acids research* 2021,49(D1):D545-D551.
- [171] Wishart DS, Feunang YD, Guo AC et al. DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic Acids Res* 2018,46(D1):D1074-D1082.
- [172] Wang Y, Zhang S, Li F et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics, *Nucleic acids research* 2020,48(D1):D1031-D1041.
- [173] Kim S, Chen J, Cheng T et al. PubChem 2019 update: improved access to chemical data, *Nucleic acids research* 2019,47(D1):D1102-D1109.
- [174] Mendez D, Gaulton A, Bento AP et al. ChEMBL: towards direct deposition of bioassay data, *Nucleic acids research* 2019,47(D1):D930-D940.
- [175] Rao H, Zhu F, Yang G et al. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Research* 2011,39(suppl_2):W385-W390.
- [176] Yang L, Shu M, Ma K et al. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues, *Amino acids* 2010,38(3):805-816.
- [177] Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides, *Journal of Molecular Structure* 2007,830(1-3):106-115.
- [178] Berkholz DS, Krenesky PB, Davidson JR et al. Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry, *Nucleic*

- Acids Research 2010,38(suppl_1):D320-D325.
- [179] Kurgan L, Miri Disfani F. Structural protein descriptors in 1-dimension and their sequence-based predictions, *Current Protein and Peptide Science* 2011,12(6):470-489.
- [180] Hvidsten TR, Kryshafovich A, Komorowski J et al. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins, *Bioinformatics* 2003,19(suppl_2):ii81-ii91.
- [181] Shi J, Yiu S, Zhang Y et al. Effective moment feature vectors for protein domain structures, *PloS one* 2013,8(12):e83788.
- [182] Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction, *Bioinformatics* 2018,34(17):821-829.
- [183] Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drug-target binding affinity, *arXiv preprint arXiv:1902.04166* 2019.
- [184] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences, *PLoS computational biology* 2019,15(6):e1007129.
- [185] Rifaioglu AS, Cetin Atalay R, Cansen Kahraman D et al. MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery, *Bioinformatics* 2021,37(5):693-704.
- [186] Wan FP, Zhu Y, Hu HL et al. DeepCPI: A Deep Learning-based Framework for Large-scale in silico Drug Screening, *Genomics Proteomics & Bioinformatics* 2019,17(5):478-495.
- [187] Zhao L, Wang J, Pang L et al. GANsDTA: predicting drug-target binding affinity using GANs, *Frontiers in genetics* 2020,10:1243.
- [188] Zhou D, Xu Z, Li W et al. MultiDTI: drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network, *Bioinformatics* 2021,37(23):4485-4492.
- [189] Nguyen T, Le H, Quinn TP et al. GraphDTA: predicting drug-target binding affinity with graph neural networks, *Bioinformatics* 2021,37(8):1140-1147.
- [190] Ishiguro K, Maeda S-i, Koyama M. Graph Warp Module: an Auxiliary Module for Boosting the Power of Graph Neural Networks in Molecular Graph Analysis, *arXiv preprint arXiv:1902.01020* 2019.
- [191] Chen S, Sun Z, Lin L et al. To Improve Protein Sequence Profile Prediction through Image Captioning on Pairwise Residue Distance Map, *Journal of chemical information and modeling* 2019,60(1):391-399.
- [192] Wang S, Sun S, Li Z et al. Accurate de novo prediction of protein contact map by ultra-deep learning model, *PLOS Computational Biology* 2017,13(1):e1005324.
- [193] Michel M, Menéndez Hurtado D, Elofsson A. PconsC4: fast, accurate and hassle-free contact predictions, *Bioinformatics* 2019,35(15):2677-2679.
- [194] Wu Q, Peng Z, Anishchenko I et al. Protein contact prediction using metagenome sequence data and residual neural networks, *Bioinformatics* 2020,36(1):41-48.
- [195] Jiang M, Li Z, Zhang S et al. Drug-target affinity prediction using graph neural network and contact maps, *RSC Advances* 2020,10(35):20701-20712.
- [196] Gao KY, Fokoue A, Luo H et al. Interpretable drug target prediction using deep neural representation. In: *IJCAI*. 2018, p. 3371-3377.
- [197] Abbasi K, Razzaghi P, Poso A et al. DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks, *Bioinformatics* 2020,36(17):4633-4642.
- [198] Zheng S, Li Y, Chen S et al. Predicting drug-protein interaction using quasi-visual question answering system, *Nature Machine Intelligence* 2020,2(2):134-140.
- [199] Zhao Q, Xiao F, Yang M et al. AttentionDTA: prediction of drug-target binding affinity using attention model. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, p. 64-69. IEEE.
- [200] Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, *Bioinformatics* 2019,35(2):309-318.
- [201] Chen L, Tan X, Wang D et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, *Bioinformatics (Oxford, England)* 2020,36(16):4406-4414.
- [202] Zeng Y, Chen X, Luo Y et al. Deep drug-target binding affinity prediction with multiple

- attention blocks, *Briefings in bioinformatics* 2021,22(5):bbab117.
- [203] Ragoza M, Hochuli J, Idrobo E et al. Protein–ligand scoring with convolutional neural networks, *Journal of chemical information and modeling* 2017,57(4):942-957.
- [204] Jiménez J, Skalic M, Martínez-Rosell G et al. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks, *Journal of chemical information and modeling* 2018,58(2):287-296.
- [205] Lim J, Ryu S, Park K et al. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation, *Journal of chemical information and modeling* 2019,59(9):3981-3988.
- [206] Cang Z, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction, *International journal for numerical methods in biomedical engineering* 2018,34(2):e2914.
- [207] Cang Z, Wei G-W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, *PLOS Computational Biology* 2017,13(7):e1005690.
- [208] Gonczarek A, Tomczak JM, Zaręba S et al. Interaction prediction in structure-based virtual screening using deep learning, *Computers in biology and medicine* 2018,100:253-258.
- [209] Gomes J, Ramsundar B, Feinberg EN et al. Atomic convolutional networks for predicting protein-ligand binding affinity, *arXiv preprint arXiv:1703.10603* 2017.
- [210] Torng W, Altman RB. Graph convolutional neural networks for predicting drug-target interactions, *Journal of chemical information and modeling* 2019,59(10):4131-4149.
- [211] Huang K, Fu T, Glass LM et al. DeepPurpose: a deep learning library for drug–target interaction prediction, *Bioinformatics* 2020,36(22-23):5545-5547.
- [212] Chen L, Tan X, Wang D et al. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, *Bioinformatics* 2020,36(16):4406-4414.
- [213] Liu H, Sun J, Guan J et al. Improving compound–protein interaction prediction by building up highly credible negative samples, *Bioinformatics* 2015,31(12):i221-i229.
- [214] Huang K, Xiao C, Glass LM et al. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction, *Bioinformatics* 2021,37(6):830-836.
- [215] Pesciullesi G, Schwaller P, Laino T et al. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates, *Nature communications* 2020,11(1):1-8.
- [216] Li P, Wang J, Qiao Y et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery, *Briefings in Bioinformatics* 2021,22(6):bbab109.
- [217] Hu W, Liu B, Gomes J et al. Strategies for pre-training graph neural networks, *arXiv preprint arXiv:1905.12265* 2019.
- [218] Lin X, Zhao K, Xiao T et al. DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction, *arXiv preprint arXiv:2003.13902* 2020.
- [219] Jiménez J, Doerr S, Martínez-Rosell G et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks, *Bioinformatics* 2017,33(19):3036-3042.
- [220] Mylonas SK, Axenopoulos A, Daras P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins, *Bioinformatics* 2021,37(12):1681-1690.
- [221] Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, *Bioinformatics* 2019,35(2):309-318.
- [222] Yang JH, Wright SN, Hamblin M et al. A white-box machine learning approach for revealing antibiotic mechanisms of action, *Cell* 2019,177(6):1649-1661. e1649.
- [223] Singh N, Chaput L, Villoutreix BO. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace, *Briefings in bioinformatics* 2021,22(2):1790-1818.
- [224] Liu Z, Du J, Fang J et al. DeepScreening: a deep learning-based screening web server for accelerating drug discovery, *Database* 2019,2019.
- [225] Zhang H, Saravanan KM, Yang Y et al. Deep learning based drug screening for novel coronavirus 2019-nCov, *Interdisciplinary Sciences, Computational Life Sciences* 2020:1.
- [226] Wu Z, Ramsundar B, Feinberg EN et al. MoleculeNet: a benchmark for molecular machine learning, *Chemical Science* 2018,9(2):513-530.

- [227] Salem M, Khormali A, Arshadi AK et al. TranScreen: Transfer Learning on Graph-Based Anti-Cancer Virtual Screening Model, *Big Data and Cognitive Computing* 2020,4(3):16.
- [228] Tan X, Jiang X, He Y et al. Automated design and optimization of multitarget schizophrenia drug candidates by deep learning, *European Journal of Medicinal Chemistry* 2020,204:112572.
- [229] Liu Z, Huang D, Zheng S et al. Deep learning enables discovery of highly potent anti-osteoporosis natural products, *European Journal of Medicinal Chemistry* 2021,210:112982.
- [230] Wang Y, Xiao J, Suzek TO et al. PubChem: a public information system for analyzing bioactivities of small molecules, *Nucleic Acids Res* 2009,37(Web Server issue):W623-633.
- [231] Gaulton A, Hersey A, Nowotka M et al. The ChEMBL database in 2017, *Nucleic Acids Research* 2017,45(D1):D945-D954.
- [232] Wishart DS, Knox C, Guo AC et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Research* 2008,36:D901-D906.
- [233] Liu ZH, Su MY, Han L et al. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions, *Accounts of Chemical Research* 2017,50(2):302-309.
- [234] Yao D, Zhang L, Zheng M et al. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease, *Sci Rep* 2018,8(1):11018.
- [235] Whirl-Carrillo M, McDonagh EM, Hebert JM et al. Pharmacogenomics Knowledge for Personalized Medicine, *Clinical Pharmacology & Therapeutics* 2012,92(4):414-417.
- [236] Arus-Pous J, Blaschke T, Ulander S et al. Exploring the GDB-13 chemical space using deep generative models, *Journal of Cheminformatics* 2019,11.
- [237] Mahmood O, Mansimov E, Bonneau R et al. Masked graph modeling for molecule generation, *Nature Communications* 2021,12(1).
- [238] Yang XF, Zhang JZ, Yoshizoe K et al. ChemTS: an efficient python library for de novo molecular generation, *Science and Technology of Advanced Materials* 2017,18(1):972-976.
- [239] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design, *Science Advances* 2018,4(7).
- [240] Olivecrona M, Blaschke T, Engkvist O et al. Molecular de-novo design through deep reinforcement learning, *Journal of Cheminformatics* 2017,9.
- [241] Gomez-Bombarelli R, Wei JN, Duvenaud D et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *Acs Central Science* 2018,4(2):268-276.
- [242] Kusner MJ, Paige B, Hernandez-Lobato JM. Grammar Variational Autoencoder, *International Conference on Machine Learning*, Vol 70 2017,70.
- [243] Simonovsky M, Komodakis N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, *Artificial Neural Networks and Machine Learning - Iccann 2018*, Pt I 2018,11139:412-422.
- [244] Jin WG, Barzilay R, Jaakkola T. Junction Tree Variational Autoencoder for Molecular Graph Generation, *International Conference on Machine Learning*, Vol 80 2018,80.
- [245] Ma TF, Chen J, Xiao C. Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders, *Advances in Neural Information Processing Systems 31 (Nips 2018)* 2018,31.
- [246] Li YB, Zhang LR, Liu ZM. Multi-objective de novo drug design with conditional graph generative model, *Journal of Cheminformatics* 2018,10.
- [247] You JX, Liu BW, Ying R et al. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation, *Advances in Neural Information Processing Systems 31 (Nips 2018)* 2018,31.
- [248] De Cao N, Kipf T. MolGAN: An implicit generative model for small molecular graphs, *arXiv preprint arXiv:1805.11973* 2018.
- [249] Shi C, Xu M, Zhu Z et al. Graphaf: a flow-based autoregressive model for molecular graph generation, *arXiv preprint arXiv:2001.09382* 2020.
- [250] Xie Y, Shi C, Zhou H et al. Mars: Markov molecular sampling for multi-objective drug discovery, *arXiv preprint arXiv:2103.10432* 2021.
- [251] Yuan YX, Pei JF, Lai LH. LigBuilder 2: A Practical de Novo Drug Design Approach, *Journal of Chemical Information and Modeling* 2011,51(5):1083-1091.
- [252] Yuan YX, Pei JF, Lai LH. LigBuilder V3: A Multi-Target de novo Drug Design Approach, *Frontiers in Chemistry* 2020,8.

- [253] Chong CM, Kou MT, Pan PC et al. Discovery of a novel ROCK2 inhibitor with anti-migration effects via docking and high-content drug screening, *Molecular Biosystems* 2016,12(9):2713-2721.
- [254] Grechishnikova D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem, *Scientific Reports* 2021,11(1).
- [255] Zhavoronkov A, Ivanenkov YA, Aliper A et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors, *Nature Biotechnology* 2019,37(9):1038-+.
- [256] Skalic M, Sabbadin D, Sattarov B et al. From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design, *Molecular Pharmaceutics* 2019,16(10):4282-4291.
- [257] Xu MY, Ran T, Chen HM. De Novo Molecule Design Through the Molecular Generative Model Conditioned by 3D Information of Protein Binding Sites, *Journal of Chemical Information and Modeling* 2021,61(7):3240-3254.
- [258] Luo S, Guan J, Ma J et al. A 3D Generative Model for Structure-Based Drug Design, *Advances in Neural Information Processing Systems* 2021,34.
- [259] Li YB, Pei JF, Lai LH. Structure-based de novo drug design using 3D deep generative models, *Chemical Science* 2021,12(41):13664-13675.
- [260] Li CY, Yao JF, Wei W et al. Geometry-Based Molecular Generation With Deep Constrained Variational Autoencoder, *Ieee Transactions on Neural Networks and Learning Systems* 2022.
- [261] Urquhart L. Top companies and drugs by sales in 2021, *Nature Reviews Drug Discovery* 2022,21(4):251-251.
- [262] Castillo-Hair SM, Seelig G. Machine Learning for Designing Next-Generation mRNA Therapeutics, *Accounts of Chemical Research* 2022,55(1):24-34.
- [263] Vaishnav ED, de Boer CG, Molinet J et al. The evolution, evolvability and engineering of gene regulatory DNA, *Nature* 2022,603(7901):455-+.
- [264] Linder J, Bogard N, Rosenberg AB et al. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences, *Cell Systems* 2020,11(1):49-+.
- [265] Hu HL, Liu XG, Xiao A et al. Riboexp: an interpretable reinforcement learning framework for ribosome density modeling, *Briefings in Bioinformatics* 2021,22(5).
- [266] Ingraham J, Garg VK, Barzilay R et al. Generative models for graph-based protein design, *Advances in Neural Information Processing Systems* 32 (Nips 2019) 2019,32.
- [267] Rohl CA, Strauss CE, Misura KM et al. Protein structure prediction using Rosetta. *Methods in enzymology*. Elsevier, 2004, 66-93.
- [268] Strokach A, Becerra D, Corbi-Verge C et al. Fast and Flexible Protein Design Using Deep Graph Neural Networks, *Cell Systems* 2020,11(4):402-+.
- [269] Eguchi RR, Anand N, Choe CA et al. IG-VAE: generative modeling of immunoglobulin proteins by direct 3D coordinate generation, *Biorxiv* 2020.
- [270] Anand N, Huang PS. Generative Modeling for Protein Structures, *Advances in Neural Information Processing Systems* 31 (Nips 2018) 2018,31.
- [271] Repecka D, Jauniskis V, Karpus L et al. Expanding functional protein sequence spaces using generative adversarial networks, *Nature Machine Intelligence* 2021,3(4):324-333.
- [272] Chu YY, Zhang Y, Wang QK et al. A transformer-based model to predict peptide-HLA class I binding and optimize mutated peptides for vaccine design, *Nature Machine Intelligence* 2022,4(3):300-+.
- [273] Zhou Y, Hou Y, Shen J et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2, *Cell discovery* 2020,6(1):1-18.
- [274] Vamathevan J, Clark D, Czodrowski P et al. Applications of machine learning in drug discovery and development, *Nature Reviews Drug Discovery* 2019,18(6):463-477.
- [275] Dong G, Liu H. Feature engineering for machine learning and data analytics. CRC Press, 2018.
- [276] Sharifi-Noghabi H, Zolotareva O, Collins CC et al. MOLI: multi-omics late integration with deep neural networks for drug response prediction, *Bioinformatics* 2019,35(14):i501-i509.
- [277] Kanehisa M, Furumichi M, Tanabe M et al. KEGG: new perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Res* 2017,45(D1):D353-D361.
- [278] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives, *IEEE transactions on pattern analysis and machine intelligence* 2013,35(8):1798-1828.

- [279] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of chemical information and computer sciences* 1988,28(1):31-36.
- [280] Glen RC, Bender A, Arnby CH et al. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME, *IDrugs* 2006,9(3):199.
- [281] Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition, *Journal of Chemical Information and Modeling* 2018,58(1):27-35.
- [282] Mikolov T, Chen K, Corrado G et al. Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* 2013.
- [283] Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics, *PloS one* 2015,10(11):e0141287.
- [284] Yang KK, Wu Z, Bedbrook CN et al. Learned protein embeddings for machine learning, *Bioinformatics* 2018,34(15):2642-2648.
- [285] Strodthoff N, Wagner P, Wenzel M et al. UDSMProt: universal deep sequence models for protein classification, *Bioinformatics* 2020,36(8):2401-2409.
- [286] Senior AW, Evans R, Jumper J et al. Improved protein structure prediction using potentials from deep learning, *Nature* 2020,577(7792):706-710.
- [287] Tunyasuvunakool K, Adler J, Wu Z et al. Highly accurate protein structure prediction for the human proteome, *Nature* 2021,596(7873):590-596.
- [288] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907* 2016.
- [289] Veličković P, Cucurull G, Casanova A et al. Graph attention networks, *arXiv preprint arXiv:1710.10903* 2017.
- [290] Sanyal S, Anishchenko I, Dagar A et al. ProteinGCN: Protein model quality assessment using graph convolutional networks, *BioRxiv* 2020.
- [291] Su C, Tong J, Zhu Y et al. Network embedding in biomedical data science, *Brief Bioinform* 2020,21(1):182-197.
- [292] Loetsch J, Ultsch A. A machine-learned computational functional genomics-based approach to drug classification, *European journal of clinical pharmacology* 2016,72(12):1449-1461.
- [293] Karim MR, Beyan O, Zappa A et al. Deep learning-based clustering approaches for bioinformatics, *Brief Bioinform* 2021,22(1):393-415.
- [294] Mohamed SK, Novacek V, Nounu A. Discovering protein drug targets using knowledge graph embeddings, *Bioinformatics* 2020,36(2):603-610.
- [295] Hawkins-Hooker A, Depardieu F, Baur S et al. Generating functional protein variants with variational autoencoders, *PLoS computational biology* 2021,17(2):e1008736.
- [296] Liu X, Xu Y, Li S et al. In Silicotarget fishing: addressing a “Big Data” problem by ligand-based similarity rankings with data fusion, *J Cheminform* 2014,6(1):1-14.
- [297] Ji S, Pan S, Cambria E et al. A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems* 2021.
- [298] Lee M, Kim H, Joe H et al. Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery, *J Cheminform* 2019,11(1):1-16.
- [299] Gao Y, Fokoue A, Luo H et al. Interpretable Drug Target Prediction Using Deep Neural Representation. In: *international joint conference on artificial intelligence*. 2018, p. 3371-3377.
- [300] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* 2014.
- [301] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017, p. 5998-6008.
- [302] Luo Y, Zhao X, Zhou J et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, *Nature communications* 2017,8(1):1-13.
- [303] Wan F, Hong L, Xiao A et al. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions, *Bioinformatics* 2019,35(1):104-111.
- [304] Zeng X, Zhu S, Hou Y et al. Network-based prediction of drug-target interactions using an arbitrary-order proximity embedded deep forest, *Bioinformatics* 2020,36(9):2805-2812.
- [305] Tang J, Qu M, Wang M et al. Line: Large-scale information network embedding. In:

- Proceedings of the 24th international conference on world wide web. 2015, p. 1067-1077.
- [306] Karimi M, Wu D, Wang Z et al. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, *Bioinformatics* 2019,35(18):3329-3338.
- [307] Jarada TN, Rokne JG, Alhadj R. SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder, *Knowledge-Based Systems* 2021,212:106585.
- [308] Chen Y, de Rijke M. A collective variational autoencoder for top-n recommendation with side information. In: *Proceedings of the 3rd workshop on deep learning for recommender systems*. 2018, p. 3-9.
- [309] Xuan P, Ye Y, Zhang T et al. Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations, *Cells* 2019,8(7):705.
- [310] Zeng X, Zhu S, Liu X et al. deepDR: a network-based deep learning approach to in silico drug repositioning, *Bioinformatics* 2019,35(24):5191-5198.
- [311] Wang Z, Zhou M, Arnold C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing, *Bioinformatics* 2020,36(Supplement_1):i525-i533.
- [312] Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models, *Brief Bioinform* 2021,22(2):1679-1693.
- [313] Huang K, Fu T, Gao W et al. Therapeutics data Commons: machine learning datasets and tasks for therapeutics, *arXiv e-prints* 2021:arXiv: 2102.09548.
- [314] Xu J, Zhang P, Huang Y et al. Multimodal single-cell/nucleus RNA sequencing data analysis uncovers molecular networks between disease-associated microglia and astrocytes with implications for drug repurposing in Alzheimer's disease, *Genome research* 2021,31(10):1900-1912.
- [315] Cheng F, Desai RJ, Handy DE et al. Network-based approach to prediction and population-based validation of in silico drug repurposing, *Nature communications* 2018,9(1):1-12.
- [316] Chang Y, Park H, Yang H-J et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature, *Sci Rep* 2018,8(1):1-11.
- [317] Fang J, Zhang P, Wang Q et al. Artificial intelligence framework identifies candidate targets for drug repurposing in Alzheimer's disease, *Alzheimer's research & therapy* 2022,14(1):1-23.
- [318] Belyaeva A, Cammarata L, Radhakrishnan A et al. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing, *Nature communications* 2021,12(1):1-13.
- [319] Nguyen DD, Gao K, Chen J et al. Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning, *Chemical Science* 2020,11(44):12036-12046.
- [320] Van Dyk DA, Meng X-L. The art of data augmentation, *Journal of Computational and Graphical Statistics* 2001,10(1):1-50.
- [321] Pham T-H, Qiu Y, Zeng J et al. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing, *Nature Machine Intelligence* 2021,3(3):247-257.
- [322] Sadegh S, Matschinske J, Blumenthal DB et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing, *Nature communications* 2020,11(1):1-9.
- [323] Gysi DM, Do Valle Í, Zitnik M et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19, *Proceedings of the National Academy of Sciences* 2021,118(19).
- [324] Zeng X, Song X, Ma T et al. Repurpose open data to discover therapeutics for COVID-19 using deep learning, *Journal of proteome research* 2020,19(11):4624-4636.
- [325] Group RC. Dexamethasone in hospitalized patients with Covid-19, *New England Journal of Medicine* 2021,384(8):693-704.
- [326] Zhou Y, Hou Y, Shen J et al. A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19, *PLoS biology* 2020,18(11):e3000970.
- [327] Zhou Y, Wang F, Tang J et al. Artificial intelligence in COVID-19 drug repurposing, *The*

- Lancet Digital Health 2020,2(12):e667-e676.
- [328] Segler MH, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI, *Nature* 2018,555(7698):604-610.
- [329] Marconi VC, Ramanan AV, de Bono S et al. Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial, *The Lancet Respiratory Medicine* 2021,9(12):1407-1418.
- [330] Hecht D, Fogel GB. Computational intelligence methods for ADMET prediction, *Front Drug Des Discov* 2009,4(27):351-377.
- [331] Freedman DH. Hunting for new drugs with AI, *Nature* 2019,576(7787):S49-S53.
- [332] Yang K, Swanson K, Jin W et al. Analyzing learned molecular representations for property prediction, *Journal of Chemical Information Modeling* 2019,59(8):3370-3388.
- [333] 王永亮, 陈敬蕊, 孔令聪 et al. 蛋白多肽类药物长效化技术研究进展, *经济动物学报* 2021:1-8.
- [334] 王锐. 临床应用的多肽药物及其研发进展. In: 2018 年中国药学会大会. 中国四川成都, 2018, p. 42.
- [335] 杨树青, 窦丽鑫, 王宏远. 生物抗菌多肽在口腔疾病治疗中的应用, *口腔颌面修复学杂志* 2019,20(03):188-192.
- [336] Sarfaraj HM, Sheeba F, Saba A et al. Marine natural products: A lead for Anti-cancer, *Indian Journal of Geo-Marine Sciences* 2012.
- [337] Adrian TE. Novel marine-derived anti-cancer agents, *Current pharmaceutical design* 2007,13(33):3417-3426.
- [338] 李明明, 赵欣然, 戴建芳 et al. 多肽药物及组装体在癌症免疫治疗中的应用(英文), *Science China Materials* 2019,62(11):1759-1781.
- [339] 范亲, 代奇轩, 张萌 et al. 自组装多肽-药物结合物的研究进展, *中国医院药学杂志* 2018,38(03):338-343.
- [340] Howard A, Udenigwe CC. Mechanisms and prospects of food protein hydrolysates and peptide-induced hypolipidaemia, *Food Function* 2013,4(1):40-51.
- [341] Kim S-K, Wijesekara I. Marine-derived Peptides: Development and Health Prospects. 2013, 1-3.
- [342] 彭博, 郭中敏, 陆家海. 人工合成抗菌肽的常用方法及应用前景, *中国抗生素杂志* 2012,37(03):176-183.
- [343] Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions, *Drug discovery today* 2015,20(1):122-128.
- [344] Borghouts C, Kunz C, Groner B. Current strategies for the development of peptide-based anti-cancer therapeutics, *Journal of peptide science: an official publication of the European Peptide Society* 2005,11(11):713-726.
- [345] Gupta S, Sharma AK, Shastri V et al. Prediction of anti-inflammatory proteins/peptides: an insilico approach, *Journal of translational medicine* 2017,15(1):1-11.
- [346] Wei L, Xing P, Su R et al. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency, *Journal of Proteome Research* 2017,16(5):2044-2053.
- [347] 唐川, 刘俊成, 周兴智 et al. 蛋白多肽类药物载体应用研究进展, *沈阳药科大学学报* 2020,37(01):51-56.
- [348] 李琬琼, 高艳锋. 抗肿瘤多肽药物研究进展, *药学进展* 2019,43(10):759-766.
- [349] 曹隽喆, 顾宏. 基于计算方法的抗菌肽预测, *计算机学报* 2017,40(12):2777-2796.
- [350] Frank K, Sippl MJ. High-performance signal peptide prediction based on sequence alignment techniques, *Bioinformatics* 2008,24(19):2172-2176.
- [351] Wang P, Hu L, Liu G et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods, *PloS one* 2011,6(4):e18476.
- [352] Altschul SF, Gish W, Miller W et al. Basic local alignment search tool, *Journal of molecular biology* 1990,215(3):403-410.
- [353] Ng XY, Rosdi BA, Shahrudin S. Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity, *Biomed Res Int* 2015,2015:212715.
- [354] Xiao Y, Cai Y, Bommineni YR et al. Identification and functional characterization of three

chicken cathelicidins with potent antimicrobial activity, *Journal of Biological Chemistry* 2006,281(5):2858-2867.

[355] Mikut R, Hilpert K, Therapeutics. Interpretable features for the activity prediction of short antimicrobial peptides using fuzzy logic, *International Journal of Peptide Research* 2009,15(2):129-137.

[356] Fernandes FC, Porto WF, Franco OL. A wide antimicrobial peptides search method using fuzzy modeling. In: *Brazilian Symposium on Bioinformatics*. 2009, p. 147-150. Springer.

[357] Loose C, Jensen K, Rigoutsos I et al. A linguistic model for the rational design of antimicrobial peptides, *Nature* 2006,443(7113):867-869.

[358] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm, *Bioinformatics* 1998,14(1):55-67.

[359] Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education, *Nucleic acids research* 2016,44(D1):D1087-D1093.

[360] Zuo Y-C, Li Q-Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet, *Peptides* 2009,30(10):1788-1793.

[361] de Brevern AG. New assessment of a structural alphabet, *In silico biology* 2005,5(3):283-289.

[362] Torrent M, Andreu D, Nogués VM et al. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model, *Plos One* 2011,6(2):e16968.

[363] Holton TA, Pollastri G, Shields DC et al. CPPpred: prediction of cell penetrating peptides, *Bioinformatics* 2013,29(23):3094-3096.

[364] He W, Jiang Y, Jin J et al. Accelerating bioactive peptide discovery via mutual information-based meta-learning, *Briefings in Bioinformatics* 2022,23(1):bbab499.

[365] Sasson O, Vaaknin A, Fleischer H et al. ProtoNet: hierarchical classification of the protein space, *Nucleic acids research* 2003,31(1):348-352.

[366] Chan WKB, Zhang H, Yang J et al. GLASS: a comprehensive database for experimentally validated GPCR-ligand associations, *Bioinformatics* 2015,31(18):3035-3042.

[367] Muttenthaler M, King GF, Adams DJ et al. Trends in peptide drug discovery, *Nature Reviews Drug Discovery* 2021,20(4):309-325.

[368] Mumtaz MM, Pohl HR. Interspecies uncertainty in molecular responses and toxicity of mixtures. *Molecular, Clinical and Environmental Toxicology*. Springer, 2012, 361-379.

[369] Altschul SF, Madden TL, Schäffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research* 1997,25(17):3389-3402.

[370] Negi SS, Schein CH, Ladics GS et al. Functional classification of protein toxins as a basis for bioinformatic screening, *Scientific Reports* 2017,7(1):1-11.

[371] Naamati G, Askenazi M, Linial M. ClanTox: a classifier of short animal toxins, *Nucleic acids research* 2009,37(suppl_2):W363-W368.

[372] Gupta S, Kapoor P, Chaudhary K et al. In silico approach for predicting toxicity of peptides and proteins, *Plos One* 2013,8(9):e73957.

[373] Wei L, Ye X, Xue Y et al. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism, *Briefings in Bioinformatics* 2021,22(5):bbab041.

[374] Pan X, Zuallaert J, Wang X et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity, *Bioinformatics* 2021,36(21):5159-5168.

[375] Wei L, Ye X, Sakurai T et al. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning, *Bioinformatics* 2022,38(6):1514-1524.

[376] Beijnen JH, Schellens JH. Drug interactions in oncology, *The lancet oncology* 2004,5(8):489-496.

[377] Qato DM, Wilder J, Schumm LP et al. Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the United States, 2005 vs 2011, *JAMA internal medicine* 2016,176(4):473-482.

[378] Huang J, Niu C, Green CD et al. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network, *PLoS Comput Biol* 2013,9(3):e1002998.

[379] Percha B, Altman RB. Informatics confronts drug-drug interactions, *Trends in pharmacological sciences* 2013,34(3):178-184.

[380] Rowland M. Introducing pharmacokinetic and pharmacodynamic concepts, *DRUGS*

- AND THE PHARMACEUTICAL SCIENCES 2008,179:1.
- [381] Segura Bedmar I, Martínez P, Sánchez Cisneros D. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts 2011.
- [382] Segura Bedmar I, Martínez P, Herrero Zazo M. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). 2013. Association for Computational Linguistics.
- [383] Vilar S, Friedman C, Hripcsak G. Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media, *Briefings in bioinformatics* 2018,19(5):863-877.
- [384] Chowdhury MFM, Lavelli A. FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013, p. 351-355.
- [385] Kim S, Liu H, Yeganova L et al. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach, *Journal of biomedical informatics* 2015,55:23-30.
- [386] Szmids E, Kacprzyk J. A similarity measure for intuitionistic fuzzy sets and its application in supporting medical diagnostic reasoning. In: *International conference on artificial intelligence and soft computing*. 2004, p. 388-393. Springer.
- [387] Dewi IN, Dong S, Hu J. Drug-drug interaction relation extraction with deep convolutional neural networks. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017, p. 1795-1802. IEEE.
- [388] Sun X, Ma L, Du X et al. Deep convolution neural networks for drug-drug interaction extraction. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018, p. 1662-1668. IEEE
- [389] Witten IH, Frank E. *Data mining: practical machine learning tools and techniques with Java implementations*, *Acm Sigmod Record* 2002,31(1):76-77.
- [390] Kavuluru R, Rios A, Tran T. Extracting drug-drug interactions with word and character-level recurrent neural networks. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017, p. 5-12. IEEE.
- [391] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. In: *Advances in neural information processing systems*. 2017, p. 5998-6008.
- [392] Zhou X, Li L, Dong D et al. Multi-turn response selection for chatbots with deep attention matching network. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, p. 1118-1127.
- [393] Zhou P, Shi W, Tian J et al. Attention-based bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, p. 207-212.
- [394] Wang L, Cao Z, De Melo G et al. Relation classification via multi-level attention cnns. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, p. 1298-1307.
- [395] Zheng W, Lin H, Luo L et al. An attention-based effective neural model for drug-drug interactions extraction, *BMC bioinformatics* 2017,18(1):445.
- [396] Zhang Y, Zheng W, Lin H et al. Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths, *Bioinformatics* 2018,34(5):828-835.
- [397] Shen Y, Yuan K, Li Y et al. Drug2vec: Knowledge-aware feature-driven method for drug representation learning. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018, p. 757-800. IEEE.
- [398] Wu H, Xing Y, Ge W et al. Drug-drug interaction extraction via hybrid neural networks on biomedical literature, *Journal of biomedical informatics* 2020,106:103432.
- [399] Bokharaeian B, Díaz A. NIL_UCM: Extracting Drug-Drug interactions from text through combination of sequence and tree kernels. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013, p. 644-650.
- [400] Tatonetti NP, Patrick PY, Daneshjou R et al. Data-driven prediction of drug effects and interactions, *Science translational medicine* 2012,4(125):125ra131-125ra131.
- [401] Bicego M, Murino V, Figueiredo MA. Similarity-based classification of sequences using

- hidden Markov models, *Pattern Recognition* 2004,37(12):2281-2291.
- [402] Chen Y, Garcia EK, Gupta MR et al. Similarity-based classification: Concepts and algorithms, *Journal of Machine Learning Research* 2009,10(3).
- [403] Yang M-S, Wu K-L. A similarity-based robust clustering method, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004,26(4):434-448.
- [404] Bicego M, Murino V, Figueiredo MA. Similarity-based clustering of sequences using hidden Markov models. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. 2003, p. 86-95. Springer.
- [405] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, *Journal of the American Medical Informatics Association* 2014,21(e2):e278-e286.
- [406] Qian S, Liang S, Yu H. Leveraging genetic interactions for adverse drug-drug interaction prediction, *PLOS Computational Biology* 2019,15(5):e1007068.
- [407] Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, *The annals of applied statistics* 2011,5(1):232.
- [408] Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions, *Proceedings of the National Academy of Sciences* 2018,115(18):E4304-E4311.
- [409] Lee G, Park C, Ahn J. Novel deep learning model for more accurate prediction of drug-drug interaction effects, *BMC bioinformatics* 2019,20(1):415.
- [410] Deng Y, Xu X, Qiu Y et al. A multimodal deep learning framework for predicting drug-drug interaction events, *Bioinformatics* 2020.
- [411] Yue X, Wang Z, Huang J et al. Graph embedding on biomedical networks: methods, applications and evaluations, *Bioinformatics* 2020,36(4):1241-1251.
- [412] Zhao C, Qiu Y, Zhou S et al. Graph embedding ensemble methods based on the heterogeneous network for lncRNA-miRNA interaction prediction, *BMC genomics* 2020,21(13):1-12.
- [413] Yu Z, Huang F, Zhao X et al. Predicting drug–disease associations through layer attention graph convolutional network, *Briefings in bioinformatics* 2020.
- [414] Ma T, Xiao C, Zhou J et al. Drug similarity integration through attentive multi-view graph auto-encoders. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: AAAI Press, 2018, 3477–3483.
- [415] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics* 2018,34(13):i457-i466.
- [416] Lin X, Quan Z, Wang Z-J et al. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. 2020. *IJCAI*.
- [417] Yu Y, Huang K, Zhang C et al. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization, *Bioinformatics* 2021,37(18):2988-2995.
- [418] Lyu T, Gao J, Tian L et al. MDNN: A Multimodal Deep Neural Network for Predicting Drug-Drug Interaction Events. In: *IJCAI*. 2021.
- [419] Liu S, Huang Z, Qiu Y et al. Structural Network Embedding using Multi-modal Deep Auto-encoders for Predicting Drug-drug Interactions. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, p. 445-450. IEEE
- [420] Zhang W, Chen Y, Liu F et al. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, *BMC bioinformatics* 2017,18(1):18.
- [421] Zhang Y, Qiu Y, Cui Y et al. Predicting Drug-drug Interactions using Multi-modal Deep Auto-encoders based Network Embedding and Positive-unlabeled Learning, *Methods* 2020.
- [422] Rohani N, Eslahchi C, Katanforoush A. Iscmf: Integrated similarity-constrained matrix factorization for drug–drug interaction prediction, *Network Modeling Analysis in Health Informatics and Bioinformatics* 2020,9(1):1-8.
- [423] Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization, *Bioinformatics* 2012,28(18):2304-2310.
- [424] Zhang W, Yue X, Lin W et al. Predicting drug-disease associations by using similarity constrained matrix factorization, *BMC bioinformatics* 2018,19(1):1-12.
- [425] Shtar G, Rokach L, Shapira B. Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures, *PloS one* 2019,14(8):e0219796.
- [426] Zhang W, Chen Y, Li D et al. Manifold regularized matrix factorization for drug-drug

- interaction prediction, *Journal of biomedical informatics* 2018,88:90-97.
- [427] Liu J, Jin X, Hong Y et al. Collaborative linear manifold learning for link prediction in heterogeneous networks, *Information Sciences* 2020,511:297-308.
- [428] Cami A, Manzi S, Arnold A et al. Pharmacointeraction network models predict unknown drug-drug interactions, *PloS one* 2013,8(4):e61468.
- [429] Hopkins AL. Network pharmacology: the next paradigm in drug discovery, *Nature chemical biology* 2008,4(11):682-690.
- [430] Chang RL, Xie L, Xie L et al. Drug off-target effects predicted using structural analysis in the context of a metabolic network model, *PLoS Comput Biol* 2010,6(9):e1000938.
- [431] Zhang P, Wang F, Hu J et al. Label propagation prediction of drug-drug interactions based on clinical side effects, *Scientific reports* 2015,5(1):1-10.
- [432] Park K, Kim D, Ha S et al. Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks, *PloS one* 2015,10(10):e0140816.
- [433] Kimmig A, Bach S, Broecheler M et al. A short introduction to probabilistic soft logic. In: *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*. 2012, p. 1-4.
- [434] Sridhar D, Fakhraei S, Getoor L. A probabilistic approach for collective similarity-based drug-drug interaction prediction, *Bioinformatics* 2016,32(20):3175-3182.
- [435] Lee K, Lee S, Jeon M et al. Drug-drug interaction analysis using heterogeneous biological information network. In: *2012 IEEE International Conference on Bioinformatics and Biomedicine*. 2012, p. 1-5. IEEE.
- [436] Dietterich TG. Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*. 2000, p. 1-15. Springer.
- [437] Yang P, Hwa Yang Y, B Zhou B et al. A review of ensemble methods in bioinformatics, *Current Bioinformatics* 2010,5(4):296-308.
- [438] Deepika S, Geetha T. A meta-learning framework using representation learning to predict drug-drug interaction, *Journal of biomedical informatics* 2018,84:136-147.
- [439] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, p. 213-220.
- [440] Mordelet F, Vert J-P. A bagging SVM to learn from positive and unlabeled examples, *Pattern Recognition Letters* 2014,37:201-209.
- [441] Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations, *Nature communications* 2019,10(1):1-11.
- [442] Dancey JE, Chen HX. Strategies for optimizing combinations of molecularly targeted anticancer agents, *Nature reviews Drug discovery* 2006,5(8):649-659.
- [443] Chiang CW, Zhang P, Wang X et al. Translational high-dimensional drug interaction discovery and validation using health record databases and pharmacokinetics models, *Clinical Pharmacology & Therapeutics* 2018,103(2):287-295.
- [444] Du L, Chakraborty A, Chiang CW et al. Graphic mining of high-order drug interactions and their directional effects on myopathy using electronic medical records, *CPT: pharmacometrics & systems pharmacology* 2015,4(8):481-488.
- [445] Zhang P, Du L, Wang L et al. A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases, *CPT: pharmacometrics & systems pharmacology* 2015,4(8):474-480.
- [446] Cherkasov A, Muratov EN, Fourches D et al. QSAR modeling: where have you been? Where are you going to?, *Journal of medicinal chemistry* 2014,57(12):4977-5010.
- [447] Tropsha A. Best practices for QSAR model development, validation, and exploitation, *Molecular informatics* 2010,29(6-7):476-488.
- [448] Balaban AT, Devillers J. *Topological indices and related descriptors in QSAR and QSPAR*. CRC Press, 2014.
- [449] Karelson M. *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience, 2000.
- [450] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies, *Chemical reviews* 1996,96(3):1027-1044.
- [451] Mikolov T, Sutskever I, Chen K et al. Distributed representations of words and phrases

- and their compositionality, *Advances in neural information processing systems* 2013,26.
- [452] Peters ME, Neumann M, Iyyer M et al. Deep contextualized word representations. In: *Proceedings of NAACL-HLT*. 2018, p. 2227-2237.
- [453] Devlin J, Chang M-W, Lee K et al. Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* 2018.
- [454] Radford A, Narasimhan K, Salimans T et al. Improving language understanding by generative pre-training 2018.
- [455] Wang S, Guo Y, Wang Y et al. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019, p. 429-436.
- [456] Fabian B, Edlich T, Gaspar H et al. Molecular representation learning with language models and domain-relevant auxiliary tasks, *arXiv preprint arXiv:2011.13230* 2020.
- [457] Chithrananda S, Grand G, Ramsundar B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, *arXiv preprint arXiv:2010.09885* 2020.
- [458] Irwin R, Dimitriadis S, He J et al. Chemformer: a pre-trained transformer for computational chemistry, *Machine Learning: Science and Technology* 2022,3(1):015022.
- [459] Zhang X-C, Wu C-K, Yang Z-J et al. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction, *Briefings in Bioinformatics* 2021,22(6):bbab152.
- [460] Rong Y, Bian Y, Xu T et al. Self-supervised graph transformer on large-scale molecular data, *Advances in neural information processing systems* 2020,33:12559-12571.
- [461] Zhu J, Xia Y, Qin T et al. Dual-view molecule pre-training, *arXiv preprint arXiv:2106.10234* 2021.
- [462] Xue D, Zhang H, Xiao D et al. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis, *bioRxiv* 2021:2020.2012. 2023.424259.
- [463] Lin X, Xu C, Xiong Z et al. PanGu Drug Model: Learn a Molecule Like a Human, *bioRxiv* 2022.
- [464] Guo Z, Sharma PK, Du L et al. MM-Deacon: Multimodal molecular domain embedding analysis via contrastive learning, *arXiv preprint arXiv:2109.08830* 2021.
- [465] Wang Y, Wang J, Cao Z et al. MolCLR: molecular contrastive learning of representations via graph neural networks, *arXiv preprint arXiv:2102.10056* 2021.
- [466] Sun M, Xing J, Wang H et al. MoCL: Contrastive Learning on Molecular Graphs with Multi-level Domain Knowledge, *arXiv preprint arXiv:2106.04509* 2021.
- [467] Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations, *Nature Machine Intelligence* 2021:1-10.
- [468] Li S, Zhou J, Xu T et al. GeomGCL: Geometric Graph Contrastive Learning for Molecular Property Prediction, *arXiv preprint arXiv:2109.11730* 2021.
- [469] Fang X, Liu L, Lei J et al. Geometry-enhanced molecular representation learning for property prediction, *Nature Machine Intelligence* 2022,4(2):127-134.
- [470] Zhou G, Gao Z, Ding Q et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework, *ChemRxiv* 2022.
- [471] Wang H, Li W, Jin X et al. Chemical-Reaction-Aware Molecule Representation Learning, *arXiv preprint arXiv:2109.09888* 2021.
- [472] Sterling T, Irwin JJ. ZINC 15—ligand discovery for everyone, *Journal of chemical information and modeling* 2015,55(11):2324-2337.
- [473] Tabak HF, Van der Horst, G., Smit, J., Winter, A. J., Mul, Y., Koerkamp, G. M. Discrimination between RNA circles, interlocked RNA circles and lariats using two-dimensional polyacrylamide gel electrophoresis, *Nucleic Acids Res* 1988,16(14):6597-6605.
- [474] Xiong Z, Wang D, Liu X et al. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism, *Journal of Medicinal Chemistry* 2020,63(16):8749-8760.
- [475] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence, *Nature Machine Intelligence* 2020,2(10):573-584.
- [476] Lipton ZC. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, *Queue* 2018,16(3):31–57.
- [477] Zhang X, Tan S, Koch P et al. Axiomatic Interpretability for Multiclass Additive Models,

- Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019.
- [478] Zhuang H, Wang X, Bendersky M et al. Interpretable Ranking with Generalized Additive Models, Proceedings of the 14th ACM International Conference on Web Search and Data Mining 2021.
- [479] Vasić M, Petrović A, Wang K et al. MoET: Mixture of Expert Trees and its application to verifiable reinforcement learning, Neural Networks 2022,151:34-47.
- [480] Jones MG, Khodaverdian A, Quinn JJ et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia, Genome Biology 2020,21(1):92.
- [481] Friedman JH, Popescu BE. Predictive learning via rule ensembles, The Annals of Applied Statistics 2008,2(3):916-954, 939.
- [482] 伟 刘, 王赛涵, 辛益博 et al. 深度态势感知与智能化战争, 国防科技 2021,42(3):9-17.
- [483] Kipf T, Welling M. Semi-Supervised Classification with Graph Convolutional Networks, ArXiv 2017,abs/1609.02907.
- [484] Velickovic P, Cucurull G, Casanova A et al. Graph Attention Networks, ArXiv 2018,abs/1710.10903.
- [485] Kipf T, Welling M. Variational Graph Auto-Encoders, ArXiv 2016,abs/1611.07308.
- [486] Wang X, Ji H, Shi C et al. Heterogeneous Graph Attention Network, The World Wide Web Conference 2019.
- [487] Yuan H, Yu H, Gui S et al. Explainability in Graph Neural Networks: A Taxonomic Survey 2020,arXiv preprint arXiv:2012.15445.
- [488] Baldassarre F, Azizpour H. Explainability Techniques for Graph Convolutional Networks 2019,arXiv preprint arXiv:1905.13686.
- [489] Pope PE, Kolouri S, Rostami M et al. Explainability Methods for Graph Convolutional Neural Networks, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019:10764-10773.
- [490] Ying R, Bourgeois D, You J et al. GNNExplainer: Generating Explanations for Graph Neural Networks, Advances in neural information processing systems 2019,32:9240-9251.
- [491] Luo D, Cheng W, Xu D et al. Parameterized Explainer for Graph Neural Network. Advances in Neural Information Processing Systems. 2020, 19620-19631.
- [492] Schlichtkrull MS, Cao ND, Titov I. Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking 2020,arXiv preprint arXiv:2010.00577.
- [493] Schnake T, Eberle O, Lederer J et al. Higher-Order Explanations of Graph Neural Networks via Relevant Walks, IEEE Transactions on Pattern Analysis and Machine Intelligence 2021:1-1.
- [494] Huang Q, Yamada M, Tian Y et al. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks 2020,arXiv preprint arXiv:2001.06216.
- [495] Zhang Y, Defazio D, Ramesh A. RelEx: A Model-Agnostic Relational Model Explainer. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, 2021, 1042–1049.
- [496] Vu MN, Thai MT. PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks, ArXiv 2020.
- [497] Yuan H, Tang J, Hu X et al. XGNN: Towards Model-Level Explanations of Graph Neural Networks, Association for Computing Machinery 2020.
- [498] Wu J, Mooney R. Faithful Multimodal Explanation for Visual Question Answering. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2019.
- [499] Chen C, Li O, Barnett A et al. This looks like that: deep learning for interpretable image recognition, Advances in Neural Information Processing Systems 32 2018.
- [500] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, the 22nd ACM SIGKDD International Conference 2016.
- [501] Guo W, Mu D, Xu J et al. LEMNA: Explaining Deep Learning based Security Applications. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada: Association for Computing Machinery, 2018, 364–379.
- [502] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating

- activation differences. Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, NSW, Australia: JMLR.org, 2017, 3145–3153.
- [503] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777.
- [504] Chen J, Song L, Wainwright MJ et al. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, ArXiv 2018.
- [505] Ramamurthy KN, Vinzamuri B, Zhang Y et al. Model Agnostic Multilevel Explanations 2020.
- [506] Ghorbani A, Wexler J, Zou J et al. Towards Automatic Concept-based Explanations, 33rd Conference on Neural Information Processing Systems 2019.
- [507] Pedapati T, Balakrishnan A, Shanmugam K et al. Learning Global Transparent Models Consistent with Local Contrastive Explanations, arXiv: Learning 2020.
- [508] Ma J, Yu MK, Fong S et al. Using deep learning to model the hierarchical structure and function of a cell, Nature Methods 2018,15(4):290-298.
- [509] Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature Genetics 2000,25(1):25-29.
- [510] Wang D, Liu S, Warrell J et al. Comprehensive functional genomic resource and integrative model for the human brain, Science 2018,362(6420).
- [511] Elmarakeby HA, Hwang J, Arafeh R et al. Biologically informed deep neural network for prostate cancer discovery, Nature 2021,598(7880):348-352.
- [512] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate, Computer Science 2014.
- [513] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010.
- [514] Devlin J, Chang MW, Lee K et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2018, Volume 1 (Long and Short Papers), pages 4171–4186.
- [515] Serrano S, Smith NA. Is Attention Interpretable, meeting of the association for computational linguistics 2019.
- [516] Jain S, Wallace BC. Attention is not Explanation, arXiv: Computation and Language 2019.
- [517] Wiegrefe S, Pinter Y. Attention is not not Explanation, arXiv: Computation and Language 2019.
- [518] Abramenko N, Kustov L, Metelytsia L et al. A review of recent advances towards the development of QSAR models for toxicity assessment of ionic liquids, Journal of Hazardous Materials 2020,384:121429.
- [519] Ghasemi F, Mehridehnavi A, Pérez-Garrido A et al. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, Drug Discov Today 2018,23(10):1784-1790.
- [520] Simões RS, Maltarollo VG, Oliveira PR et al. Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges, Frontiers in Pharmacology 2018,9:74.
- [521] Chen H, Carlsson L, Eriksson M et al. Beyond the scope of Free-Wilson analysis: building interpretable QSAR models with machine learning algorithms, J Chem Inf Model 2013,53(6):1324-1336.
- [522] Luque Ruiz I, Gómez-Nieto M. Building of Robust and Interpretable QSAR Classification Models by Means of the Rivality Index, J Chem Inf Model 2019,59(6):2785-2804.
- [523] Chen CH, Tanaka K, Kotera M et al. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications, J Cheminform 2020,12(1):19.
- [524] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation, J Cheminform 2020,12(1):17.
- [525] Schwaller P, Gaudin T, Lányi D et al. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models, Chem Sci 2018,9(28):6091-6098.

- [526] Marcou G, Horvath D, Solov'ev V et al. Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions, *Mol Inform* 2012,31(9):639-642.
- [527] Sheridan RP, Feuston BP, Maiorov VN et al. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR, *J Chem Inf Comput Sci* 2004,44(6):1912-1928.
- [528] Liu R, Wallqvist A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds, *J Chem Inf Model* 2019,59(1):181-189.
- [529] Janet JP, Duan C, Yang T et al. A quantitative uncertainty metric controls error in neural network-driven chemical discovery, *Chem Sci* 2019,10(34):7913-7922.
- [530] Obrezanova O, Csanyi G, Gola JM et al. Gaussian processes: a method for automatic QSAR modeling of ADME properties, *J Chem Inf Model* 2007,47(5):1847-1857.
- [531] Schroeter TS, Schwaighofer A, Mika S et al. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules, *J Comput Aided Mol Des* 2007,21(12):651-664.
- [532] Bosc N, Atkinson F, Felix E et al. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery, *J Cheminform* 2019,11(1):4.
- [533] Liu R, Wang H, Glover KP et al. Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure-Activity Relationship Models Based on Deep Neural Networks?, *Journal of Chemical Information and Modeling* 2019,59 1:117-126.
- [534] Nembri S, Grisoni F, Consonni V et al. In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9, *Int J Mol Sci* 2016,17(6).
- [535] Raghavendra NM, Pingili D, Kadasi S et al. Dual or multi-targeting inhibitors: The next generation anticancer agents, *European Journal of Medicinal Chemistry* 2018,143:1277-1300.
- [536] Lehar J, Krueger AS, Avery W et al. Synergistic drug combinations tend to improve therapeutically relevant selectivity, *Nature Biotechnology* 2009,27(7):659-666.
- [537] Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era, *Nature Biotechnology* 2012,30(7):679-692.
- [538] Murphy EM, Jimenez HR, Smith SM. Current clinical treatments of AIDS, *Advances in Pharmacology* 2008,56:27-73.
- [539] Groll AH, Tragiannidis A. Recent advances in antifungal prevention and treatment, *Seminars in Hematology* 2009,46(3):212-229.
- [540] Tamma PD, Cosgrove SE, Maragakis LL. Combination therapy for treatment of infections with gram-negative bacteria, *Clinical Microbiology Reviews* 2012,25(3):450-470.
- [541] Chen X, Ren B, Chen M et al. NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning, *PLOS Computational Biology* 2016,12(7):e1004975.
- [542] Julkunen H, Cichonska A, Gautam P et al. Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects, *Nature Communications* 2020,11(1):6136.
- [543] Li H, Li T, Quang D et al. Network Propagation Predicts Drug Synergy in Cancers, *Cancer Research* 2018,78(18):5446-5457.
- [544] Gayvert KM, Aly O, Platt J et al. A Computational Approach for Identifying Synergistic Drug Combinations, *PLOS Computational Biology* 2017,13(1):e1005308.
- [545] Wildenhain J, Spitzer M, Dolma S et al. Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning, *Cell Systems* 2015,1(6):383-395.
- [546] Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations, *Nature Communications* 2019,10(1):1197.
- [547] Menden MP, Wang D, Mason MJ et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen, *Nature Communications* 2019,10(1):2674.
- [548] Liu Q, Xie L. TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations, *PLOS Computational Biology* 2021,17(2):e1008653.
- [549] Yuan B, Shen C, Luna A et al. CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy, *Cell Syst* 2021,12(2):128-140.e124.

- [550] Wieder O, Kohlbacher S, Kuenemann M et al. A compact review of molecular property prediction with graph neural networks, *Drug Discovery Today: Technologie* 2020,37:1-12.
- [551] Schneider G. Mind and machine in drug design, *Nature Machine Intelligence* 2019,1(3):128-130.
- [552] Zhang Z, Liu Q, Wang H et al. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction, *ArXiv* 2021,abs/2110.00987.
- [553] Hirschfeld L, Swanson K, Yang K et al. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction, *Journal of Chemical Information and Modeling* 2020,60(8):3770-3780.
- [554] Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning, *Chemical Science* 2019,10(35):8154-8163.
- [555] Xie L, He S, Song X et al. Deep learning-based transcriptome data classification for drug-target interaction prediction, *BMC Genomics* 2018,19(Suppl 7):667.
- [556] Wen M, Zhang Z, Niu S et al. Deep-Learning-Based Drug-Target Interaction Prediction, *J Proteome Res* 2017,16(4):1401-1409.
- [557] Bagherian M, Sabeti E, Wang K et al. Machine learning approaches and databases for prediction of drug-target interaction: a survey paper, *Brief Bioinform* 2021,22(1):247-269.
- [558] Gao KY, Fokoue A, Luo H et al. Interpretable drug target prediction using deep neural representation. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: AAAI Press, 2018, 3371–3377.
- [559] Karimi M, Wu D, Wang Z et al. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks, *Bioinformatics* 2019,35(18):3329-3338.
- [560] Yang Z, Zhong W, Zhao L et al. ML-DTI: Mutual Learning Mechanism for Interpretable Drug-Target Interaction Prediction, *J Phys Chem Lett* 2021,12(17):4247-4261.
- [561] Agyemang B, Wu WP, Kpiebaareh MY et al. Multi-view self-attention for interpretable drug-target interaction prediction, *J Biomed Inform* 2020,110:103547.
- [562] Jia P, Hu R, Pei G et al. Deep generative neural network for accurate drug response imputation, *Nat Commun* 2021,12(1):1740.
- [563] Liu M, Cai R, Hu Y et al. Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning, *Journal of the American Medical Informatics Association* 2013,21(2):245-251.
- [564] Wang F, Zhang P, Cao N et al. Exploring the associations between drug side-effects and therapeutic indications, *Journal of Biomedical Informatics* 2014,51:15-23.
- [565] Dey S, Luo H, Fokoue A et al. Predicting adverse drug reactions through interpretable deep learning framework, *BMC Bioinformatics* 2018,19(21):476.
- [566] Xu Y, Pei J, Lai L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction, *J Chem Inf Model* 2017,57(11):2672-2685.
- [567] Schneider P, Schneider G. De Novo Design at the Edge of Chaos, *Journal of Medicinal Chemistry* 2016,59 9:4077-4086.
- [568] Sheridan RP. Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It?, *Journal of Chemical Information and Modeling* 2019,59 4:1324-1337.
- [569] Sahigara F, Mansouri K, Ballabio D et al. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models, *Molecules* 2012,17:4791 - 4810.
- [570] Mathea M, Klingspohn W, Baumann K. Chemoinformatic Classification Methods and their Applicability Domain, *Molecular Informatics* 2016,35.
- [571] Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation", *AI Magazine* 2017,38:50-57.
- [572] Nagarajan D, Nagarajan T, Roy N et al. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria, *The Journal of Biological Chemistry* 2017,293:3492 - 3509.
- [573] Müller AT, Hiss JA, Schneider G. Recurrent Neural Network Model for Constructive Peptide Design, *Journal of Chemical Information and Modeling* 2018,58 2:472-479.
- [574] Jiménez-Luna J, Cuzzolin A, Bolcato G et al. A Deep-Learning Approach toward Rational

Molecular Docking Protocol Selection, *Molecules* 2020,25.

[575] Jiménez J, Kalj M, Martínez-Rosell G et al. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *Journal of Chemical Information and Modeling* 2018,58 2:287-296.

[576] Rogers D, Hahn M. Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling* 2010,50 5:742-754.

[577] Awale M, Reymond JL. Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of ZINC and GDB-17, *Journal of Chemical Information and Modeling* 2014,54 7:1892-1907.

[578] Todeschini R, Ballabio D, Consonni V. Novel Molecular Descriptors Based on Functions of New Vertex Degrees. 2010, 73-100.