药物-靶点相互作用预测中的方法、数据及软件的发展现状

牛步盈[1,2]，孙菁菁[1,2]，石江珊[1,2]，郑明月[1,2*]，李叙潼[1,2*](1.中国科学院上海药物研究所，上海 201203；2.中国科学院大学，北京 100049)

摘要：识别药物-靶点相互作用一直是新药发现进程中的重要步骤。由于体外实验的高成本性，基于计算机的药物-靶点相互作用预测方法将大大缩小化学空间的搜索范围，加快药物发现的进程。本文将介绍与药物-靶点预测密切相关的数据库软件和用于药物-靶点相互作用预测的主要方法，进而对药物-靶点相互作用预测问题进行总结和讨论。

关键词：药物-靶点相互作用；新药发现；计算模型；人工智能

中图分类号：R966　　**文献标志码：**A　　**文章编号：**1007-7693(2022)21-2809-10

DOI: 10.13748/j.cnki.issn1007-7693.2022.21.017

引用本文：牛步盈, 孙菁菁, 石江珊, 等. 药物-靶点相互作用预测中的方法、数据及软件的发展现状[J]. 中国现代应用药学, 2022, 39(21): 2809-2818.

Development of Methods, Data and Software for Drug Target Interaction Prediction

NIU Buying[1,2], SUN Jingjing[1,2], SHI Jiangshan[1,2], ZHENG Mingyue[1,2*], LI Xutong[1,2*](*1.Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; 2.University of Chinese Academy of Sciences, Beijing 100049, China*)

ABSTRACT: Identifying drug-target interaction is an important step in the discovery of new drugs. Because of the high cost of the experiment *in vitro*, the computer based predictive method of drug-target interaction will greatly reduce the scope of the chemical space and accelerate the process of drug discovery. This paper introduces the data and software that are closely related to the drug-target prediction and main methods of its in-silico prediction. Then, the problems of drug target interaction prediction are summarized and discussed.

KEYWORDS: drug-target interaction; drug discovery; in-silico model; artificial intelligence

药物发现是一个高代价、高风险、低回报的过程，相比较药物靶点的搜索范围，药物分子可探索的化学空间显得十分庞大。目前，大量已知化合物与蛋白质的相互作用谱尚不清楚，PubChem数据集所包含的大部分化合物大多均没有公认的与靶点相互作用的概况[1]，这大大限制了化学分子的药用。在湿实验中尝试详尽揭示潜在的化学分子与靶点的相互作用是十分耗时耗力的。因此，研究者倾向于通过构建计算模型来发现药物-靶点相互作用关系，从而减少要通过湿实验验证的药物靶点候选物的搜索空间，最大限度地减少工作量和成本[2]。

药物-靶点相互作用指药物与相应靶点结合而引起靶点蛋白行为或功能的变化。药物通常指与靶点结合能够引起生理变化的化学分子；靶点通常指生物体内与药物分子结合后，其结构和行为发生改变的生物大分子(核酸或蛋白质)。多数药物靶点属于4个蛋白质家族：核受体、离子通道、G蛋白偶联受体和酶[1]。

药物-靶点相互作用的预测可以促进药物发现[3]、药物重新定位[4]和药物不良反应[5]预测等研究。药物重定位旨在使用已获批的药物来治疗不同疾病，由于已获批药物的化学性质、不良反应、安全性等均是已知的，“老药新用”展现出广阔的前景，通过药物-靶点相互作用预测进行药物重定位能够极大地节省时间和精力。多重药理学表明[1]：一种药物可能对多个靶点具有活化或者抑制作用，从而导致其有多种药理作用。一方面，对非治疗靶点的作用可能产生不良反应，若在临床试验早期通过药物-靶点相互作用预测发现特异性作用于靶点的药物，有助于减少不良反应带来的研发失败；另一方面，设计对多个生物途径有效

基金项目：临港实验室(LG202102-01)；上海市科技重大专项

作者简介：牛步盈，女，博士生　E-mail: niubuying@simm.ac.cn　***通信作者：**郑明月，男，博士，研究员　E-mail: myzheng@simm.ac.cn　李叙潼，女，博士　E-mail: lixutong@simm.ac.cn

的分子，就有可能创造出高效且低毒的新药，并且不容易产生耐药性，将有助于改善费力又昂贵的药物测试过程。

药物-靶点相互作用预测方法通常分为 3 类：基于配体的方法、基于对接的方法和化学基因组学方法[2]。基于配体的方法假设相似的配体倾向于结合相似的靶点，根据分子与已知配体的相似度来预测新的药物-靶点相互作用；基于对接的方法须已知配体和靶点蛋白的三维结构数据，并根据两者相结合的位置姿势、亲和力和能量等给出两者具有相互作用的预测概率；基于化学基因组学的方法主要通过靶点蛋白和药物分子的序列结构信息以及已知的药物-靶点相互作用关系数据来训练算法，从而探索新的药物-靶点相互作用。当配体分子的结构不能详尽表示或靶点蛋白的三维结构特征未知时，前 2 种方法将大大受限，化学基因组学预测药物-靶点相互作用则有助于突破这种局限性。

本文主要针对药物-靶点相互作用介绍常用数据库以及预测工具软件，从基于相似性的方法、基于特征的方法和基于深度学习等方面介绍药物-靶点预测的主要方法。最后，对当前药物-靶点相互作用预测存在的问题进行说明，并对药物-靶点相互作用预测的未来进行展望。

1 数据库和软件

1.1 数据库

药物-靶点相互作用预测基于大量化学分子以及蛋白靶点的信息，一系列的数据库提供了药物、靶点以及其相互作用的信息可用于药物-靶点相互作用的有效预测。

1.1.1 DrugBank DrugBank[6]数据库包含超过 50 万个关于 FDA 批准的药物以及正在通过 FDA 批准程序的实验药物，其机制、相互作用和靶点的全面分子信息。丰富且高质量的数据资源使其在药物基因组学、药物蛋白质组学、药物转录组学、药物代谢组学、药动学、药剂学和药物发现的研究中均有着广泛的应用。

1.1.2 ZINC ZINC[7]是一个免费的用于虚拟筛选化合物的商业数据库，它提供了超过 2.3 亿个分子的 3D 结构，库中的分子均被赋予了生物学上相关的质子化态，并以分子量、LogP 和可旋转键的数量等性质信息加以注释。此外，提供多个对接程序接口、用户自定义分子操作以及基于 Web 数据库搜索和浏览功能。

1.1.3 PubChem PubChem[8]是由美国国立卫生研究院建立的小分子生物特性公共资源库，提供了生物筛选结果数据的快速检索、整合、比较，探索结构-活性分析和目标选择性检查等功能。它包含了约 1.12 亿种化合物、2.96 亿种生物活性物和 18 万种蛋白。

在最新的版本[9]中，添加了元素周期表和元素页面、路径页面和知识面板。此外，为应对 2019 年新型冠状病毒肺炎(COVID-19)疫情，PubChem 创建了一个特殊的数据收集，其中包含与 COVID-19 和严重急性呼吸综合征冠状病毒 2 (SARS-CoV-2)相关的数据。

1.1.4 ChEMBL ChEMBL 数据库[10]包含大量药物类生物活性化合物的功能和 ADMET(即体内吸收、分布、代谢、排泄和毒性特性的评估)信息。最新版本[11]添加了实验和靶点注释信息，包括临床候选靶点和适应证以及药物代谢途径和计算结构警报信息。

1.1.5 化学品相互作用的搜索工具(search tool for interactions of chemicals，STITCH) STITCH 数据库[12]来自实验、数据库和文献化学品整合成一个单一的、易于使用的资源。最新的版本中新增了提供化学物质结合亲和力的网络视图以及过滤与特定组织无关的蛋白质和化学物质等功能。

1.1.6 BindingDB BindingDB 数据库[13]中约有 250 万个蛋白质-小分子相互作用数据，涉及约 108 万个小分子和 8 千个蛋白质。BindingDB 提供了可交叉搜索的多种查询类型，包括文本、化学结构、蛋白质序列和数字亲和性等以及通过最大化学相似度、支持向量机等方法进行虚拟化合物筛选的工具。

1.1.7 治疗靶点数据库[14](therapeutic target database，TTD) TTD 提供已知治疗蛋白和核酸靶点、靶向疾病、通路信息以及针对每个靶点的相应药物-配体的信息。最新的版本[15]中包含 38 760 种药物以及 3 578 种靶点，提供了靶点的 PDB 形式与 AlphaFold 形式交互连接以及对多入口靶点序列或药物结构高级检索的功能。

1.1.8 SuperTarget SuperTarget 数据库[16]包括医学适应证、药物不良反应、药物代谢、通路和靶点蛋白的基因本体术语。目前数据库包含超过 6 000 个靶点蛋白以及 19.6 万个化合物，标注了超

过 330 000 个药物-靶点相互作用关联。用户界面将药物筛选和靶点相似性纳入查询的选项，查询界面支持通过构建复杂的查询来查找特定靶点或药物的功能。

1.1.9 UniPort UniPort 知识库[17]是专为蛋白质开发的，包含约 2.2 亿个蛋白质序列及蛋白生物功能的相关信息。它从文献中提取注释信息并添加到已查的条目中，并在未查的条目中使用自动化系统(如基于关联规则的注释器)提供的注释对其进行补充。此外，它还建立了为 UniProt 提供新的条目和新的注释公开提交界面用来丰富数据库中的数据。

1.1.10 京都基因和基因组百科全书[18] (Kyoto Encyclopedia of Genes and Genomes，KEGG) KEGG 是一个系统分析基因功能的知识库，将基因组信息与高阶功能信息联系起来。由 3 个子数据库组成，分别为基因子数据库、通路子数据库和配体子数据库。此外，还提供了浏览基因组图、比较 2 组基因组图和操作表达图的 Java 图形工具，以及序列比较、图比较和路径计算的计算工具。

1.1.11 Pfam Pfam 数据库[19]是一个包含蛋白注释信息且可进行多序列比对的蛋白家族和结构域的数据库，广泛用于分析新基因组、宏基因组及指导特定蛋白质和系统的实验工作。此外，为促进 COVID-19 的研究，此库修订了包含 SARS-CoV-2 蛋白质组的 Pfam 条目，并建立了新的 Pfam 未覆盖区域的条目。

1.2 软件

药物和靶点特征的提取是预测药物-靶点相互作用的基础，有各种各样的在线网站工具以及软件包来提取药物的化学描述符和靶点蛋白的不同性质特征。

1.2.1 PROFEAT PROFEAT[20]是一个从氨基酸序列中计算蛋白质和多肽的常用特征的网络服务器。其计算的特征包括蛋白质和多肽的 11 个特征描述符组、400 多个小分子描述符组以及蛋白质-蛋白质和蛋白质-小分子相互作用的衍生特征。在预测特定结构或功能类别的蛋白质、蛋白质间的相互作用、特定功能的多肽和小分子的定量结构活性关系等方面有着广泛的应用。

1.2.2 PYDPI PYDPI[21]是一个从氨基酸序列和化学结构分别计算蛋白质和药物特征的综合平台。可从氨基酸序列计算蛋白质和多肽常用的结构和物理化学特征，从药物分子的拓扑结构计算分子描述符，以及蛋白质-蛋白质相互作用和蛋白质-配体相互作用描述符。它提供了 42 种描述符类型，包括 9 890 个蛋白质描述符，13 种描述符类型，包括 615 个药物描述符。此外，该平台还提供了 7 种药物分子指纹系统，包括原子对指纹、拓扑指纹、拓扑扭转指纹、电拓扑状态指纹、摩根/环形指纹、MACCS 密钥、FP4 密钥。

1.2.3 RDKit RDKit[22]是一个用于为化学分子生成各种描述符(包括 SMILES、2D 和 3D 分子描述符等)的化学信息学工具包，且支持子结构检索、分子序列化、相似性比较、多样性分析等功能。

1.2.4 ChemDes ChemDes[23]是一个免费的基于网络的平台。该平台集成了 Pybel、CDK、RDKit 等多个软件包用于计算分子描述符和分子指纹。它支持指纹相似度计算、MOPAC 优化和格式转换等功能。目前，其具有生成 3 679 个分子描述符和 59 种分子指纹的计算能力。

1.2.5 jCompoundMapper jCompoundMapper[24]是一个用于化学指纹的开源 Java 库和命令行工具，它为化学图谱提供了主流的分子指纹算法，比如深度优先搜索指纹、最短路径指纹、自相关指纹等等，为数据挖掘提供了标准的指纹，也为新指纹编码的开发提供了基础。

1.2.6 Protr Protr[25]是一个基于氨基酸序列生成蛋白质和多肽的各种数值表示的 R 包。它支持用户生成自定义属性的蛋白描述符，蛋白序列相似性计算，基于位置特定评分矩阵计算蛋白质图谱特征。

1.2.7 Pse-in-One Pse-in-One[26]是一种特征提取工具，支持根据蛋白质、DNA、RNA 序列特征以及自定义规则产生所需要的向量。其中包含了 8 种基于序列或伪序列组成的蛋白质特征提取方法，覆盖了 547 种氨基酸理化性质。

1.2.8 ChemoPy ChemoPy[27]是依赖于 Pybel、RDKit 等其他软件包，用于计算化学分子常用结构和物理化学特征的开源软件包。它计算了由 19 个描述符组成的 16 个药物特征组，其中包括 1 135 个描述符值。它支持 7 种类型的分子指纹，包括拓扑指纹、电拓扑状态指纹、拓扑扭转指纹、MACCS 键、FP4 键、原子对指纹和摩根指纹。

1.2.9 SIMCOMP SIMCOMP[28]用于计算 2 个化合物之间的化学结构相似性，通过构建以原子为

节点，共价键为边的二维图，根据图中最大的子结构来判断化合物之间的相似性，并给出全局相似性得分。

1.2.10 Open Babel Open Babel[29]旨在描述多种化学数据，对分子建模、化学性质、生物化学等相关领域的数据进行存储、搜索、转换、分析。它可以解析 110 多种化学文件格式，并提供可扩展的分子指纹和分子力学的相关功能。

2 方法

2.1 基于相似性的方法

该方法通常基于以下假设[30]：若已知药物分子 d 与靶蛋白 p 具有相互作用，则认为与药物分子 d 相似性高的分子与靶蛋白 p 也具有相互用；与靶蛋白 p 相似性高的蛋白与药物分子 d 也具有相互作用；同时与药物分子 d 和靶蛋白 p 相似性高的药物分子也具有相互作用。

选择恰当的评估分子相似性的方法对相似度的衡量显得尤为重要。仅考虑 2D 的化学分子结构及靶点蛋白的氨基酸序列来评估分子相似性，对药物-靶点间的相互作用进行预测是不准确的。例如：药物 D00316 和 D01132 共享许多靶点，但其 2D 化学结构相似性仅为 0.275，在与 D01132 最相似的 25 种药物中，D00316 仅排第 10[31]，见图 1。因为仅基于简单的序列结构信息不足以充分表征分子的相关特征，捕捉到影响药物-靶点相互作用的独特点。

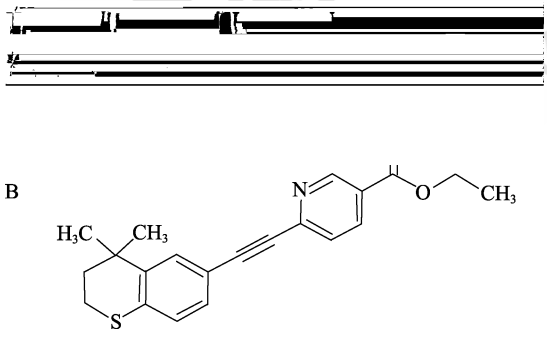

图 1 D00316(A)和 D01132(B)的二维结构

Fig. 1 2D chemical structure of D00316(A) and D01132 (B)

在药物 2D 化学结构之上添加非结构的信息，例如：理化性质、药理治疗特性、作用的器官或系统的分类信息等；对于靶点除蛋白氨基酸序列信息外添加了功能类别信息，即根据目标催化的化学反应分类或蛋白质编码基因的注释功能来衡量靶点相似度[31]。由于分子官能团通常代表化合物的特征以及与其他分子的反应机制，并且常见官能团的数量非常少，因此可以使用官能团的组合来唯一地表示一种药物；使用蛋白质的生化以及理化特征来编码靶点蛋白[32]，把它们编码后做嵌入形成向量，所有向量构成了一个向量空间，每个向量都可以看作是空间中的一个点。基于欧氏距离、曼哈顿距离、马氏距离等的最近邻算法来比较向量之间的相似度。

2.1.1 基于二分图的方法 网络或图 G 是由 1 组顶点(节点)V 和 1 组边(线)E 组成，这些边将各个顶点连接起来，由此构成图[33]。二分图是图的一种特殊模型，其中顶点 V 可分为 2 个互不相交的子集，且图中每条边连接分属于 2 个不同子集的顶点。

药物-靶点相互作用的形式与二分图十分类似，因此，研究人员通过二分图构建药物-靶点关系并对其进行相关预测。Bleakley 等[34]将药物和靶点设为节点，药物-靶点之间的相互作用设为边，通过引入局部模型，将边的预测转换为分类问题，即转而判断节点之间是否具有相互作用的方式(图 2)，来训练构成的二分图局部模型，从而预测 4 种重要的药物靶点作用，包括酶、离子通道、GPCRs 和核受体。最终通过文献检索，在 10 个预测结果最高的药物-靶点相互作用中至少得到 4 个经过验证的结果。

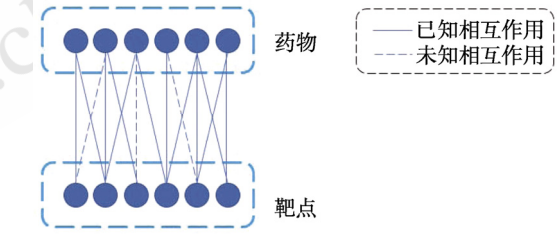

图 2 药物-靶点相互作用二分图

Fig. 2 Bipartite graph of drug-target interaction

由于传统的二分图局部模型面对新的候选药物或靶点时，无法做出有效预测，Mei 等[35]在二分图局部模型的基础上引入了基于邻居的相互作用特征推断的方式，将与新候选药物(或靶点)相似性较大的已知药物(或靶点)作为新候选药物(或靶点)的邻居，它们的相互作用作为训练标签来预测新的候选药物(或靶点)。为减小二分图中错误枢纽(指某些节点作为 k 邻近出现在其他节点中的次数过高)的影响，Buza 等[36]开发了一种 Hubness-aware 回归模型，其中引入误差修正的 k 邻近算法并在

多模态相似性空间中对药物-靶点进行增强表示后，通过映射进行集成。经验证，该模型产生了更多的药物-靶点相互作用的有效信息。

2.1.2 基于网络的方法 通过同质性网络图构建，将药物化学结构数据和蛋白质序列数据与已知的药物-靶点相互作用数据相结合，以预测潜在的药物-靶点相互作用。然而简单的同质性网络图结构没有考虑更多类型的生物节点，因此多个异质性相似网络组合应运而生。

异质性相互作用网络，例如：药物-不良反应关联网络、蛋白质-疾病关联网络、化学相似性等网络组合使用能够多维度描述药物和靶点相关信息。An等[37]提出多路复用网络中的嵌入框架，基于 15 个异质性信息网络构建节点相似网络，应用随机游走提取网络中节点的拓扑信息来构建梯度决策提升树模型完成药物-靶点相互作用分类任务。Thafar等[38]基于异质性信息网络图(包括药物-药物相似度相互作用图、靶点-靶点相似度相互作用图以及药物-靶点相互作用图)，使用多图嵌入、图挖掘结合基于特征和基于相似性的方法，将新型药物-靶点相互作用的识别建模为异构网络中的链接预测问题。

尽管异质图可以在单个网络中集成多种类型的实体和相互作用，但聚合节点或边的属性来获得图表示方法，仍具有一定的挑战性。Wan 等[39]通过将异质图转变为具有定向和加权边的同质图，将所需预测的化合物节点和蛋白质节点间的距离问题转化为是否具有相互作用的概率问题。其中，将化合物和蛋白质节点表示为 1 024 维向量，边通过 Dice 相似性系数进行表示后，以端到端的方式进行特征提取、降维以及判别分析。与SEA[40-41]、DeepDTA[42]、DTINet[43]等基于异构网络的方法相比，该方法在 AUROC 和 AUPRC 上取得最优结果。此外，在预测新的潜在相互作用中，有 3 对得到了文献验证。

基于相似性的药物-靶点预测方法实现简单有效，但是该方法基于相似的药物和靶点结构具有一致的相互作用的假设，该类模型往往倾向于发现不同药物-靶点对的共同点，难以捕捉单个药物-靶点对的独特点。此外，由于相似度矩阵计算的高复杂度，每种基于相似性方法在扩展到大型数据集时耗时较长，具有一定的局限性。

2.2 基于特征的方法

基于特征的预测方法选取药物和靶点的特征描述符，并将其分别表示为带有二进制标签且具有一定长度的特征向量。通过降维等方法降低特征维度，降低数据稀疏度，放大有效特征来提高模型的性能和效率。通过特征拼接或向量计算等方法来将药物-靶点对的特征整合，输入至支持向量机、随机森林和核方法等分类器中进行模型的训练，对药物-靶点对进行预测，将其分为正负相互作用的 2 类。

笔者将从以下 3 个方面介绍基于特征的方法，分为特征向量化、特征提取以及模型方法。

2.2.1 特征向量化

2.2.1.1 基本特征向量化 根据分子的特征、分子指纹以及各种分子描述符构建分子配体向量，如通过化学结构以及质谱数据[44]，将分子编码为代表某些功能基团或片段存在的指纹特征向量[45]；或根据拓扑描述符，功能描述符，分子性质描述等构建配体分子的特征向量。

基于蛋白质氨基酸序列、短肽序列[46]及蛋白质物化性质指纹信息[47]构建靶点向量，使用径向基函数用于支持向量机分类。或将蛋白质序列编码为包含生物进化信息的位置特异性评分矩阵[3,45]，从而构建靶点向量。

2.2.1.2 基于相似性的向量化 通过相似性来表征药物和靶点的分子特征从而构建特征向量，基于不同的核函数分别计算化学分子以及靶点的相似性来构建向量，例如：基于化学分子图使用 Tanimoto 核函数[48]来计算分子相似度从而构建配体分子向量；使用狄拉克核函数[49]、多任务核函数[50]等来计算蛋白质相似性从而构建靶点分子。

2.2.1.3 基于数据库的向量化 药物和靶点向量的编码分别对应数据库中的相关信息：药物分子的化学结构对应 PubChem 数据库中定义的 881 个唯一化学子结构[51](去掉了化学性质相同、结构相同重复的药物)；靶蛋白的基因组信息中除去从 UniProt 数据库中获得的信息之外，相关蛋白结构域对应 Pfam 数据库中 876 个结构域[52]。每个配体分子和靶蛋白分别对每维结构的存在或不存在进行 1 或 0 编码，从而构建特征向量。

2.2.2 特征提取 将相应的靶点与配体表示成向量后，通常需要对向量进行拼接，常用的拼接方法分别有直接拼接法和张量相乘法。此后通过适当的核函数将其投射到分类器可分的空间中，进行分类或回归任务预测。

由于高维特征空间表示张量的稀疏性，许多降维的方法(主成分分析法、SVD[1]、PLS[1]、Laplacian Eigenmaps[1]、最小哈希散列[53]等)可减少上述问题，提升模型的性能。

此外，受益于 Word2vec 等无监督表示学习方法的启发，SPVec[45](图 3)通过组合 SMILES2Vec[54] 和 ProtVec[55]构建了新的 SPVec 向量来表示特定的药物-靶点相互作用，通过负采样方法实现的 Skip-gram 模型来训练将原始数据(如 SMILES 字符串和蛋白质序列)自动表示为连续、信息丰富和低维向量的新方法，以避免人工提取特征的数据稀疏性和比特位碰撞[45]。

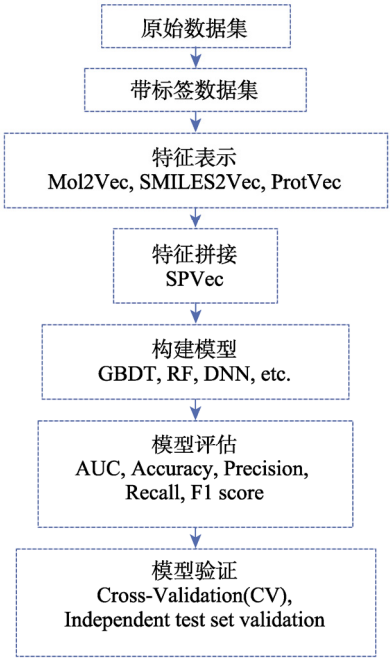

图 3 SPVec 模型
Fig. 3 SPVec model

2.2.3 模型方法

2.2.3.1 随机旋转森林 最初由 Rodríguez 等[56]提出的旋转森林分类器[3,57-58]是一种通过多个差分分类器集成来提升模型分类性能的集成分类器，已被广泛应用于药物-靶点相互作用的预测中。它首先随机划分样本集，并通过不同的变换方法对样本子集进行转换以增加样本集的差异性。分别输入至不同的决策树分类器中进行预测，最后模型通过集成各子分类器的结果给出最终的预测结果。

Li 等[57]使用位置特异性得分矩阵将蛋白质序列转换为包含生物进化信息的数值描述符，然后使用离散余弦变换算法提取隐藏特征并与化学子结构描述子集成。使用旋转森林分类器预测药物与靶蛋白之间是否存在相互作用。该模型在酶、离子通道、GPCRs 和核受体基准数据集上的平均准确率分别达到 0.914 0，0.891 9，0.872 4 和 0.811 1。

2.2.3.2 LightGBM LightGBM 是一种基于决策树算法的快速、分布式、高性能的梯度增强框架，用于排序、分类和许多其他机器学习任务，相较于 XGBoost 可在所需内存更少的情况下处理更多的数据。

Mahmud 等[2]通过伪位置特异性评分矩阵，二肽组成和伪氨基酸组成提取蛋白质序列的特征向量；并使用 MACCS 子结构提取药物特征。使用 FastUS 算法处理类不平衡问题和 MoIFS 算法来去除不相关和冗余的特征后，输入至 LightGBM 分类器中来识别药物-靶点相互作用。该模型显示出优越的性能，并且可用于发现未知疾病或感染的新药。

2.2.3.3 核方法 通过核方法可以整合多种信息进行预测，Van Laarhoven 等[59]引入高斯相互作用轮廓核，通过正则化最小二乘法来预测药物-靶点相互作用。同时引入 RLS-Kron 算法，使用 Kronecker 结合了药物核函数及靶点核函数，进一步优化了性能。Hao 等[60]将非线性核融合技术与 RLS 相结合，模型在 AUC 和 AUPR 上都呈现出更好的结果。

为了进一步优化多核学习的问题，Nascimento 等[61]提出了 KronRLS-MKL 算法来自动选择和组合药物-靶点相互作用中的核方法。通过 L2 正则化的方法产生一个非稀疏基本内核的组合，该方法可以处理药物-蛋白质的相互作用矩阵，并且不需要对药物-靶点网络进行再抽样，这种内核的非稀疏组合增加了模型的泛化能力。

2.2.3.4 矩阵分解法 矩阵分解法将药物靶点的关联矩阵分解为表征药物和靶点的低阶矩阵，从而得到药物和靶点的特征空间。Gönen 等[62]将药物-靶点相互作用预测看作二分类问题，提出了双核贝叶斯矩阵因式分解的模型。它结合了降维、矩阵分解和二进制分类，仅利用化合物之间的化学相似性和靶点蛋白之间的基因组相似性来预测药物-靶点相互作用网络。

为实现更好的泛化能力，Xia 等[63]提出了自定进度学习的协作矩阵分解方法来识别未知药物-靶

点相互作用对的模型。其将药物和靶点映射到一个低阶特征空间后，将药物的化学结构信息、蛋白质序列信息和已知的药物-靶点相互作用信息整合到正则化的最小二乘中。自定进度学习可以规避不良的局部极小值并增加鲁棒性，尤其针对数据存在严重噪声和丢失的情况。

相较于基于相似性的方法，基于特征的方法能够更加详尽地捕捉药物和靶点的几何及化学特征，该方法在大型数据集上更加有效，使得模型预测更加稳健。

2.3 基于深度学习的方法

2.3.1 卷积神经网络 Shim 等[64]基于相似性的模型，将二维卷积神经网络应用于药物和靶点的 2 个相似性矩阵的列向量之间外积的计算，以预测药物-靶点结合的亲和力。为进一步捕捉原子和氨基酸之间的复杂相互作用，Zhao 等[65]提出了 1 种基于卷积神经网络和注意力机制的端到端的仿生模型(图 4)来预测药物与靶点之间的相互作用。

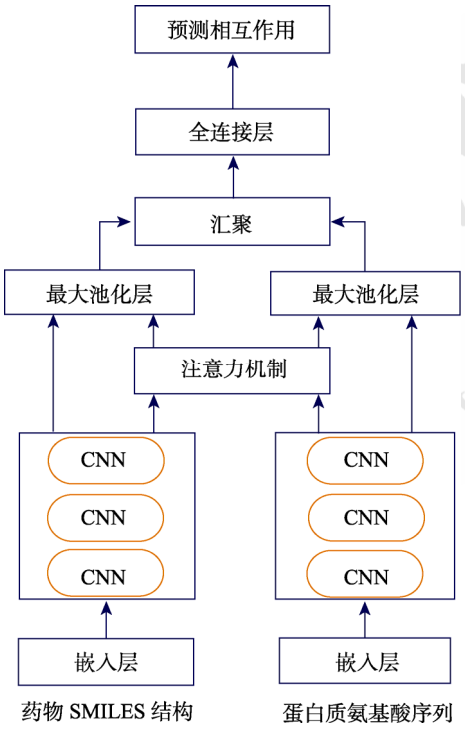

图 4 仿生模型结构

Fig. 4 Structure of biomimetic model

该模型的 2 个嵌入层将输入的 SMILES 字符串和氨基酸序列转换为相应的嵌入向量，得到相应的嵌入矩阵。将 2 个独立的卷积神经网络分别作用于药物和蛋白质，从而有效地提取序列的语义信息。同时，引入特殊的注意力机制——HyperAttention，在药物、蛋白质子序列间的维度上进行语义相互依赖性的建模，相比现有的最先进的基准模型，其 AUC 和 AUPR 均得到了显著提高。

2.3.2 Transformer Hu 等[66]提出了基于 Transformer 来评估蛋白质和药物之间结合力的模型，该模型结合了蛋白质序列、蛋白质三维结构图和药物 SMILES 字符串的相关特征信息。在大规模蛋白质序列语料库上进行了预训练后，在 PDB 数据集上对模型进行了评估，相比基于三维结构的经典深度学习模型有更好的结果。

Huang 等[67]基于 Transformer 设计了 Moltrans 模型，该模型由 3 部分组成，分别为频率连续子序列(Frequent Consecutive Sub-sequence，FCS)挖掘模块、Transformer 编码模块和相互作用预测模块。首先通过 FCS 挖掘模块将输入的药物和蛋白质数据分解成一组显式的子结构序列，再使用 Transformer 编码器获得每个子结构增强的上下文信息并嵌入到模型中。在相互作用预测模块中，应用卷积神经网络捕捉药物分子和蛋白质分子子结构之间的相互作用。最后通过 Transformer 解码器得到药物和靶点之间相互作用可能性的概率值。该模型具有良好的解释性的同时比最佳的基准模型的效果提高了约 25%。

由于传统的 Transformer 模型在小规模数据集上易出现过拟合的现象，Chen 等[68]提出了 TransformerCPI 模型，使用一维卷积的门控卷积网络和门控线性单元代替了传统 Transformer 编码器中的自注意力机制层，提升了模型的性能。

2.3.3 知识图谱 知识图谱是一种将关系信息表示为图的数据表示模型，其中图的节点表示实体，边表示实体之间的关系，通常以头实体、关系和尾实体的三元组的形式来进行表示。

KGE_NFM 方法[69]通过将知识图谱和推荐系统相结合，构建了一个系统化的药物-靶点相互作用预测框架。此框架首先通过知识图谱学习图谱中各种药物相关概念实体的低维表示，然后通过集成知识图谱中所学到的多组学信息和药物与蛋白的结构表征信息实现药物-靶点相互作用的预测。在基准数据集上，KGE_NFM 模型相比于其他传统方法的预测准确度提高了 15%~30%。

为了建立元素之间的微观联系以及各元素的基本领域知识，Fang 等[70]基于化学元素周期表，构建了化学元素知识图谱。化学元素知识图谱中

包含了元素之间的关系及其基本的化学性质，例如：周期性、金属性等。基于此，本文提出一种知识增强的分子图对比学习框架，利用化学元素知识图谱指导原始分子图的增强过程，并针对分子增强图设计了知识感知的消息传递网络，通过最大化正样本对之间的一致性和难负样本对之间的差异性构建对比损失以优化模型。实验结果表明，其在涵盖不同分子属性的 8 个测试数据集上取得了最佳的性能表现。

采用深度学习的方法能够更加有效地预测药物和靶点之间的潜在相互作用，但该模型可解释性差，且在小数据集和数据不平衡的情况下，模型难以取得良好的预测效果。

3 总结

在过去十年中，在药物发现的早期阶段采用深度学习模型预测药物与靶点的相互作用大大降低了药物发现的成本。

药物-靶点预测中出现了一系列的问题：对于深度学习模型，某些药物靶点间数据量缺乏，Popova 等[71]尝试将强化学习策略应用于低数据量的药物-靶点相互作用预测模型以提高模型的泛化能力。在分类任务中，只有阳性相互作用的靶点-配体对存在于数据库中，研究者常根据未标记的药物靶点对中随机选择阴性样本，因此引入了大量的“假阳性”样本。许多减少假阳性的负样本提取方法，例如：使用聚类方法[72]对未知的相互作用中的未标记互作和负互作进行分类；以及基于正非标记学习的反向样本提取方法[4]提取“阴性药物-靶点相互作用样本”，用于筛选强反向药物-靶点相互作用样本等。为避免随机组合和基于相似性方法产生负样本而带来的配体偏差，Chen 等[68]构建了 2 个“标签反转数据集”，使在训练集中属于正样本(或负样本)的配体，在测试集中进行标签反转(图 5)，从而降低配体偏置在模型训练过程中产生的影响。

基于相似性的方法研究原理完善，方法简单明了；基于特征的方法能够更好地表征药物和靶点，获得更有效的预测结果；深度学习方法更适用于大规模数据，能够获得更加精确的结果。受益于集成模型的启发，药物-靶点相互作用预测方法间并没有明确的壁垒，越来越多的药物-靶点相互作用预测方法呈现出多模式结合的趋势。

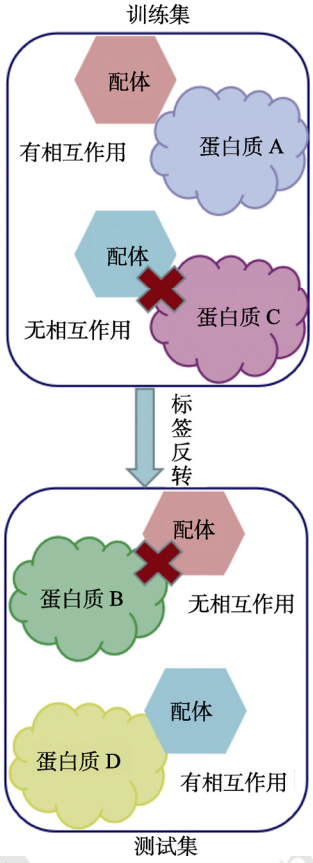

图 5 标签反转
Fig. 5 Label reversal

随着可探索化学空间的扩大，药物-靶点相互作用预测方法不断地拓展，计算机辅助药物-靶点预测模型准确性、鲁棒性等的提升，药物-靶点相互作用预测模型将进一步通过自动化药物发现的早期阶段助力药物发现过程。

REFERENCES

[1] SACHDEV K, GUPTA M K. A comprehensive review of feature based methods for drug target interaction prediction[J]. J Biomed Inform, 2019(93): 103159.

[2] MAHMUD S M H, CHEN W Y, LIU Y S, et al. PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques[J]. Brief Bioinform, 2021, 22(5): bbab046.

[3] LI Y, LIU X Z, YOU Z H, et al. A computational approach for predicting drug-target interactions from protein sequence and drug substructure fingerprint information[J]. Int J Intell Syst, 2021, 36(1): 593-609.

[4] PENG L H, ZHU W, LIAO B, et al. Screening drug-target interactions with positive-unlabeled learning[J]. Sci Rep, 2017, 7(1): 8087.

[5] BAN T, OHUE M, AKIYAMA Y. Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction[C]//2017 IEEE 7th International

Conference on Computational Advances in Bio and Medical Sciences. Orlando, FL, USA. IEEE, 2017: 1-6.
[6] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Res, 2018, 46(D1): D1074-D1082.
[7] IRWIN J J, SHOICHET B K. ZINC: a free database of commercially available compounds for virtual screening[J]. J Chem Inf Model, 2005, 45(1): 177-182.
[8] WANG Y L, XIAO J, SUZEK T O, et al. PubChem: a public information system for analyzing bioactivities of small molecules[J]. Nucleic Acids Res, 2009, 37(Web Server issue): W623-W633.
[9] KIM S, CHEN J, CHENG T J, et al. PubChem in 2021: New data content and improved web interfaces[J]. Nucleic Acids Res, 2021, 49(D1): D1388-D1395.
[10] GAULTON A, HERSEY A, NOWOTKA M, et al. The ChEMBL database in 2017[J]. Nucleic Acids Res, 2017, 45(D1): D945-D954.
[11] GAULTON A, BELLIS L J, BENTO A P, et al. ChEMBL: a large-scale bioactivity database for drug discovery[J]. Nucleic Acids Res, 2012, 40(Database issue): D1100-D1107.
[12] SZKLARCZYK D, SANTOS A, VON MERING C, et al. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data[J]. Nucleic Acids Res, 2016, 44(D1): D380-D384.
[13] GILSON M K, LIU T Q, BAITALUK M, et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology[J]. Nucleic Acids Res, 2016, 44(D1): D1045-D1053.
[14] CHEN X, JI Z L, CHEN Y Z. TTD: therapeutic target database[J]. Nucleic Acids Res, 2002, 30(1): 412-415.
[15] ZHOU Y, ZHANG Y T, LIAN X C, et al. Therapeutic target database update 2022: Facilitating drug discovery with enriched comparative data of targeted agents[J]. Nucleic Acids Res, 2022, 50(D1): D1398-D1407.
[16] HECKER N, AHMED J, VON EICHBORN J, et al. SuperTarget Goes quantitative: Update on drug-target interactions[J]. Nucleic Acids Res, 2012, 40(Database issue): D1113-D1117.
[17] CONSORTIUM U. UniProt: the universal protein knowledgebase in 2021[J]. Nucleic Acids Res, 2021, 49(D1): D480-D489.
[18] KANEHISA M, GOTO S. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Res, 2000, 28(1): 27-30.
[19] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: The protein families database in 2021[J]. Nucleic Acids Res, 2021, 49(D1): D412-D419.
[20] RAO H B, ZHU F, YANG G B, et al. Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence[J]. Nucleic Acids Res, 2011, 39(Web Server issue): W385-W390.
[21] CAO D S, LIANG Y Z, YAN J, et al. PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies[J]. J Chem Inf Model, 2013, 53(11): 3086-3096.
[22] LANDRUM G. RDKit: Open-source cheminformatic software[EB/OL]. 2006. http://www.rdkit.org.
[23] DONG J, CAO D S, MIAO H Y, et al. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation[J]. J Cheminform, 2015(7): 60.
[24] HINSELMANN G, ROSENBAUM L, JAHN A, et al. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints[J]. J Cheminform, 2011, 3(1): 3.
[25] XIAO N, CAO D S, ZHU M F, et al. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences[J]. Bioinformatics, 2015, 31(11): 1857-1859.
[26] LIU B, LIU F L, WANG X L, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences[J]. Nucleic Acids Res, 2015, 43(W1): W65-W71.
[27] CAO D S, XU Q S, HU Q N, et al. ChemoPy: freely available python package for computational biology and chemoinformatics[J]. Bioinformatics, 2013, 29(8): 1092-1094.
[28] HATTORI M, TANAKA N, KANEHISA M, et al. SIMCOMP/SUBCOMP: Chemical structure search servers for network analyses[J]. Nucleic Acids Res, 2010, 38(Web Server issue): W652-W656.
[29] O'BOYLE N M, BANCK M, JAMES C A, et al. Open Babel: An open chemical toolbox[J]. J Cheminform, 2011(3): 33.
[30] BAGHERIAN M, SABETI E, WANG K, et al. Machine learning approaches and databases for prediction of drug-target interaction: A survey paper[J]. Brief Bioinform, 2021, 22(1): 247-269.
[31] SHI J Y, YIU S M, LI Y M, et al. Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering[J]. Methods, 2015(83): 98-104.
[32] HE Z S, ZHANG J, SHI X H, et al. Predicting drug-target interaction networks based on functional groups and biological features[J]. PLoS One, 2010, 5(3): e9603.
[33] BLEAKLEY K, BIAU G, VERT J P. Supervised reconstruction of biological networks with local models[J]. Bioinformatics, 2007, 23(13): i57-i65.
[34] BLEAKLEY K, YAMANISHI Y. Supervised prediction of drug-target interactions using bipartite local models[J]. Bioinformatics, 2009, 25(18): 2397-2403.
[35] MEI J P, KWOH C K, YANG P, et al. Drug-target interaction prediction by learning from local information and neighbors[J]. Bioinformatics, 2013, 29(2): 238-245.
[36] BUZA K, PEŠKA L. Drug-target interaction prediction with Bipartite Local Models and hubness-aware regression[J]. Neurocomputing, 2017(260): 284-293.
[37] AN Q, YU L. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions[J]. Brief Bioinform, 2021, 22(6): bbab275.
[38] THAFAR M A, OLAYAN R S, ASHOOR H, et al. DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques[J]. J Cheminform, 2020, 12(1): 44.
[39] WAN X Z, WU X L, WANG D Y, et al. An inductive graph neural network model for compound-protein interaction prediction based on a homogeneous graph[J]. Brief Bioinform, 2022, 23(3): bbac073.
[40] KEISER M J, SETOLA V, IRWIN J J, et al. Predicting new molecular targets for known drugs[J]. Nature, 2009, 462(7270):

175-181.
[41] KEISER M J, ROTH B L, ARMBRUSTER B N, et al. Relating protein pharmacology by ligand chemistry[J]. Nat Biotechnol, 2007, 25(2): 197-206.
[42] ÖZTÜRK H, ÖZGÜR A, OZKIRIMLI E. DeepDTA: deep drug-target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.
[43] LUO Y N, ZHAO X B, ZHOU J T, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information [J]. Nat Commun, 2017, 8(1): 573.
[44] NAGAMINE N, SAKAKIBARA Y. Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data[J]. Bioinformatics, 2007, 23(15): 2004-2012.
[45] ZHANG Y F, WANG X G, KAUSHIK A C, et al. SPVec: A Word2Vec-inspired feature representation method for drug-target interaction prediction[J]. Front Chem, 2020(7): 895.
[46] LESLIE C, ESKIN E, WESTON J, et al. Mismatch string kernels for SVM protein classification[J]. Adv Neur Informat Process Syst, 2003: 1441-1448.
[47] VENKATARAJAN M S, BRAUN W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties[J]. Mol Modeling Annu, 2001, 7(12): 445-453.
[48] RALAIVOLA L, SWAMIDASS S J, SAIGO H, et al. Graph kernels for chemical informatics[J]. Neural Netw, 2005, 18(8): 1093-1110.
[49] BORGWARDT K M, ONG C S, SCHÖNAUER S, et al. Protein function prediction via graph kernels[J]. Bioinformatics, 2005, 21(Suppl 1): i47-i56.
[50] EVGENIOU T, MICCHELLI C A, PONTIL M. Learning multiple tasks with kernel methods[J]. J Mach Learn Res, 2005(6): 615-637.
[51] SHI J Y, ZHANG A Q, ZHANG S W, et al. A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization[J]. BMC Syst Biol, 2018, 12(Suppl 9): 136.
[52] TABEI Y S, PAUWELS E, STOVEN V, et al. Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers[J]. Bioinformatics, 2012, 28(18): i487-i494.
[53] BRODER A Z, CHARIKAR M, FRIEZE A M, et al. Min-wise independent permutations[J]. J Comput Syst Sci, 2000, 60(3): 630-659.
[54] GOH G B, HODAS N O, SIEGEL C, et al. SMILES2Vec: an interpretable general-purpose deep neural network for predicting chemical properties[EB/OL]. 2017: arXiv: 1712.02034[stat.ML]. https://arxiv.org/abs/1712.02034.
[55] ASGARI E, MOFRAD M R K. Continuous distributed representation of biological sequences for deep proteomics and genomics[J]. PLoS One, 2015, 10(11): e0141287.
[56] RODRÍGUEZ J J, KUNCHEVA L I, ALONSO C J. Rotation forest: A new classifier ensemble method[J]. IEEE Trans Pattern Anal Mach Intell, 2006, 28(10): 1619-1630.
[57] LI Y, HUANG Y, YOU Z H, et al. Drug-target interaction prediction based on drug fingerprint information and protein sequence[J]. Molecules, 2019, 24(16): 2999.
[58] WANG L, YOU Z H, LI L P, et al. Incorporating chemical sub-structures and protein evolutionary information for inferring drug-target interactions[J]. Sci Rep, 2020, 10(1): 6641.
[59] VAN LAARHOVEN T, NABUURS S B, MARCHIORI E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. Bioinformatics, 2011, 27(21): 3036-3043.
[60] HAO M, WANG Y L, BRYANT S H. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique[J]. Anal Chim Acta, 2016(909): 41-50.
[61] NASCIMENTO A C A, PRUDÊNCIO R B C, COSTA I G. A multiple kernel learning algorithm for drug-target interaction prediction[J]. BMC Bioinformatics, 2016(17): 46.
[62] GÖNEN M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization[J]. Bioinformatics, 2012, 28(18): 2304-2310.
[63] XIA L Y, YANG Z Y, ZHANG H, et al. Improved prediction of drug-target interactions using self-paced learning with collaborative matrix factorization[J]. J Chem Inf Model, 2019, 59(7): 3340-3351.
[64] SHIM J, HONG Z Y, SOHN I, et al. Prediction of drug-target binding affinity using similarity-based convolutional neural network[J]. Sci Rep, 2021, 11(1): 4416.
[65] ZHAO Q, ZHAO H, ZHENG K, et al. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism[J]. Bioinformatics, 2021: btab715.
[66] HU F, HU Y, ZHANG J, et al. Structure Enhanced Protein-Drug Interaction Prediction using Transformer and Graph Embedding[C]//2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020.
[67] HUANG K X, XIAO C, GLASS L M, et al. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction[J]. Bioinformatics, 2021, 37(6): 830-836.
[68] CHEN L F, TAN X Q, WANG D Y, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments[J]. Bioinformatics, 2020, 36(16): 4406-4414.
[69] YE Q, HSIEH C Y, YANG Z Y, et al. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system[J]. Nat Commun, 2021, 12(1): 6775.
[70] FANG Y, ZHANG Q, YANG H, et al. Molecular Contrastive Learning with Chemical Element Knowledge Graph[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022.
[71] POPOVA M, ISAYEV O, TROPSHA A. Deep reinforcement learning for de novo drug design[J]. Sci Adv, 2018, 4(7): eaap7885.
[72] KEUM J, NAM H. SELF-BLM: Prediction of drug-target interactions via self-training SVM[J]. PLoS One, 2017, 12(2): e0171839.

收稿日期：2022-08-25
(本文责编：沈倩)