◎热点与综述◎

药物-靶点相互作用预测的计算方法综述

张 然[1,2],王学志[1],汪嘉葭[1],孟 珍[1]
1.中国科学院 计算机网络信息中心 大数据技术与应用发展部,北京 100083
2.中国科学院大学 计算机科学与技术学院,北京 100049

摘 要:药物-靶点相互作用预测旨在发现可作用于特定蛋白质的潜在药物,在药物重定位、药物副作用预测、多重药理学和耐药性的研究中都发挥着重要作用。随着计算机处理能力的进步和计算算法的不断更新,药物-靶点相互作用预测的计算方法展现出时间短、成本低、精度高、范围广的优势,受到了广泛的关注,并取得了显著的进展。为了梳理其研究发展历程,探讨未来的研究方向,就药物-靶点相互作用预测的背景和意义进行简要概述;将方法分为基于分子对接、基于药物结构、基于文本挖掘和基于化学基因组四类进行综述,并对每类方法进行对比分析,详细阐述每类方法的数据需求及应用场景;对现有研究存在的局限性和面临的挑战进行讨论,展望未来的研究方向,为后续研究提供参考和借鉴。

关键词:药物-靶点相互作用预测;药物发现;数据挖掘;生物信息
文献标志码:A **中图分类号:**TP391.72 doi:10.3778/j.issn.1002-8331.2210-0108

Survey on Computational Approaches for Drug-Target Interaction Prediction

ZHANG Ran[1,2], WANG Xuezhi[1], WANG Jiajia[1], MENG Zhen[1]
1.Department of Big Data Technology and Application Development, Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China
2.School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Drug-target interaction prediction aims to discover potential drugs acting on specific proteins, and plays an important role in drug repositioning, drug side effect prediction, polypharmacology and drug resistance research. With the advancement of computer processing and the continuous updating of computing algorithms, the computational drug-target interaction prediction has shown the advantages of short time, low cost, high precision and wide range, which has received extensive attention and made remarkable progress. In order to sort out the development history and explore the future research direction, the background and significance of drug-target interaction prediction are firstly introduced in brief. Secondly, the methods are classified into four types: molecular docking-based, drug structure-based, text mining-based and chemogenomic-based methods. A comparative analysis of each method is carried out, and the data requirements and application scenarios for each type of methods are described in detail. Finally, the limitations and challenges of the existing research are discussed, and the future research directions are prospected to provide references for follow-up research.

Key words: drug-target interaction prediction; drug discovery; data mining; bioinformatics

药物发现可以分为药物早期发现和药物开发两大阶段。药物早期发现过程包含药物作用靶点的选择、先导化合物的确定、活性化合物的筛选和候选药物的选定。候选药物确定后,新药研发就进入开发阶段,开发阶段包括临床前研究和临床研究[1]。

药物-靶点相互作用(drug-target interaction,DTI)在药物发现中起着至关重要的作用。DTI预测的主要目标是发现可作用于特定蛋白质的潜在药物,它在药物重

基金项目:中国科学院战略性先导科技专项(XDA16021400,XDB31000000,XDB38030300)。
作者简介:张然(1998—),女,硕士研究生,研究方向为生物信息数据挖掘;王学志(1979—),通信作者,男,博士,研究员,研究方向为生物大数据分析技术,E-mail:wxz@cnic.cn;汪嘉葭(1987—),女,博士,副研究员,研究方向为生物信息学;孟珍(1982—),女,硕士,高级工程师,研究方向为生物多源异构数据的融合管理与关联技术。
收稿日期:2022-10-09 **修回日期:**2023-01-03 **文章编号:**1002-8331(2023)12-0001-13

定位、药物副作用预测、多重药理学和耐药性的研究中都发挥着重要作用[2-3]。

绝大多数药物的靶点是蛋白质,蛋白质在生物过程中起着核心作用,疾病的发生往往与蛋白质的结构功能变化相关。药物分子通常通过与靶蛋白表面的特定位点结合来发挥作用,以达到缓解疾病的目的。在体外或体内实验方法中,测量药物与靶点的结合亲和力是昂贵耗时的,给研制新药的经济成本和时间成本都带来了挑战。据研究统计,开发一种新的分子实体药物大约需要18亿美元,新药申请的批准则需要9至12年[4-5]。从2019年底开始,具有高传染性和高隐匿性的COVID-19引发的疫情已在全球范围内迅速传播,对人类健康构成严重威胁[6]。治疗COVID-19的药物开发显然不宜按照传统的新药研发流程[7],因此药物重定位研究成为研究热点。药物重定位发现经过严格测试验证的药物的临床新用途,大大缩短了药物研发进程,在应对突发性疾病和治疗罕见病中展现了突出优势。药物除了作用于主要的治疗靶点外,还可能与其他蛋白质发生作用,了解药物靶点信息有助于预估脱靶毒性和次优疗效,也为药物副作用的分子机制研究提供了新的视角。

近年来,计算机处理能力的进步和计算算法的不断更新使得DTI计算预测方法成为早期药物发现的流行工具,在疾病相关的miRNA预测[8-10]、疾病基因预测[11]、蛋白质-蛋白质相互作用预测[12]和蛋白质亚细胞定位预测[13]等生物信息领域也取得了一定的成果。DTI的计算预测方法缩小了实验验证DTI的搜索范围,具有巨大的研究潜力和广泛的应用前景,学术界和工业界对于技术的发展有着持续迫切的需求。

1　DTI预测的计算方法综述

基于计算的方法在探索潜在DTI中展现了时间短、成本低、精度高、范围广的优势,因而受到了广泛的关注[14]。如图1所示,基于不同的数据类型和应用场景,DTI预测的计算方法可分为四大类:基于分子对接的方法、基于药物结构的方法、基于文本挖掘的方法与基于化学基因组的方法。

1.1　基于分子对接的方法

基于分子对接的方法对蛋白质的三维(3-Dimensional,3D)结构进行建模,模拟药物与蛋白质的对接过程,依赖于分析生物分子3D结构的能力。药物分子能否发挥特定生物效应取决于其能否与蛋白质上特定位点结合产生相互作用,掌握感兴趣靶点的结构是实施基于分子对接方法的先决条件[15]。

1.1.1　蛋白质结构建模

随着基因组学和蛋白质组学的发展进步,大量候选药物靶点被发现,基于结构的计算机辅助药物设计已成为一种常用的药物发现技术[16]。随着X射线晶体学和核磁共振波谱学等生物物理技术的飞速发展与广泛应用,大量蛋白质的3D结构变得清晰可见。基于分子对接的方法采用仿真建模和可视化技术快速筛选大型化合物库并确定潜在的药物,大幅增加了新药成功开发的概率。

蛋白质的结构建模方法一般有三种:原子表示、表面表示和网格表示[17]。蛋白质的排序和打分均基于势能函数时,通常采用蛋白质的原子表示[18]。表面表示则使用几何特征描述分子形貌,表示为光滑的凸面、凹面和鞍形表面[19],然后通过配体分子和蛋白质结合位点表面的互补排列来引导对接。网格表示[20]将蛋白质分子的表面与内部分开,在三个维度上扫描分子,基于傅里叶变换计算的相关函数来确定分子之间的重叠程度。网格表示方法通过将能量势存储在网格点上来表示蛋白质的物理化学性质。

1.1.2　药物与蛋白质的对接模拟

在获得蛋白质的结构表示之后,需要对药物与蛋白

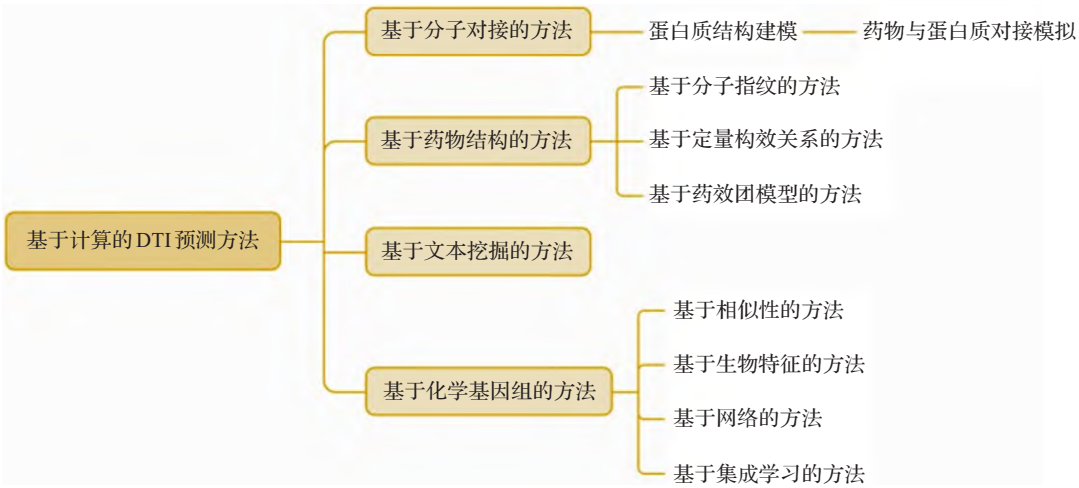
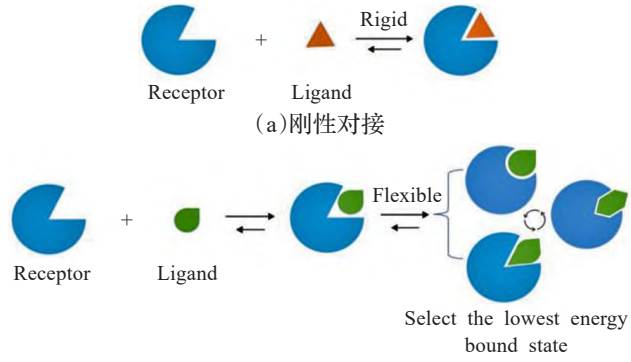


图1　基于计算的DTI预测方法分类

Fig.1　Taxonomy of computational DTI prediction methods

质进行对接模拟。分子对接[21]过程将药物分子逐一放在靶点分子的活性位点处,通过不断优化化合物的位置、构象等,寻找药物小分子与靶点大分子结合的最佳构象,通过打分函数选出与真实构象最为接近且与靶点大分子结合亲和力最佳的药物小分子。

对接方法根据对接过程中药物和蛋白质分子构象变化的灵活程度可分为刚性对接与柔性对接[22-23],如图2所示。刚体对接方法基于锁钥原理将分子视为刚性结构,仅考虑药物与靶点之间静态的物理化学互补性,参与对接的分子不改变其构象,仅可改变空间位置与姿态。刚性对接方法的简化程度高,计算量相对较小,适用于时间紧迫的情况。柔性对接方法也被称为诱导契合方法,在对接过程中允许分子构像自由变化,由于变量随着体系的原子数呈几何级数增长,计算量非常大,但适合精确模拟考察分子对接情况。此外,还有一种半柔性对接方法,半柔性对接方法基于启发式的构象空间探索方法,允许药物小分子构象在一定的范围内变化,而蛋白质大分子是刚性的,适合处理小分子与大分子间的对接。



(a)刚性对接

(b)柔性对接

图2　分子对接方法

Fig.2　Molecular docking models

分子对接实验可能会产生数十万个药物-靶点复合物构象,需要快速准确地评估药物与蛋白质相互作用的能量。打分函数对复合物构象进行排序,选择出有效结合模式,并预测药物-靶点结合亲和力,实现从靶点到先导化合物再到药物的优化。打分函数一般分为四种类型:基于力场的打分函数、基于经验的打分函数、基于知识的打分函数和基于共识的打分函数。

基于力场的打分函数基于经典的分子力学,使用从实验数据和量子力学计算得出的参数和拟合分子间相互作用得到的变量[24]。基于经验的打分函数则将参数与实验数据相匹配,通过回归分析实验数据获得不同参数的权重,目前已有几种基于经验的打分函数应用在商用对接产品中[25-27]。基于知识的打分函数使用通过实验确定的复杂结构中包含的信息,如原子间距比平均距离更适用于表示分子间的接触,较低频率发生的相互作用倾向于较低的亲和力,目前也有一些基于知识的打分函数在实际生产中得以应用[28-31]。基于共识的打分函数则采用集成学习的思想,使用不同的打分函数多次重新评估预测构象,综合多个结果以提升准确性[32],研究结果表明结合具有互补优势的打分函数具有重要的意义。

基于分子对接的方法是预测药物与蛋白质相互作用的常用方法之一,然而当蛋白质的3D结构无法获得时,此类方法将不再适用。同时基于分子对接的方法需要模拟药物与蛋白质之间复杂的结合过程,庞大的搜索空间和巨大的计算规模给计算时间和计算成本带来挑战,使得基于分子对接的方法普遍效率低下。

1.2　基于药物结构的方法

基于药物结构的方法不需要掌握靶点的3D结构,仅利用已有的药理学相关知识,分析与感兴趣靶点发生相互作用的药物结构,保留与相互作用相关的理化性质,舍弃其余的无关信息,揭示并预测药物化学结构与药理活性的关系[33]。

基于药物结构方法的核心是获取药物分子的描述符。药物分子描述符通常通过基于知识的图论方法、分子力学或量子力学工具得到,可以在多个复杂层次上描述药物分子的结构信息和理化性质。

常见的基于药物结构的方法有三种:一种是基于分子指纹的方法,基于已知药物的化学相似性,构建具有生物学意义的集群;一种是基于定量构效关系(quantitative structure activity relationship,QSAR)的方法,从化学结构预测生物活性,对影响感兴趣靶点的化学结构特征进行加权;另一种则是基于药效团模型的方法,重点关注具有相同药理作用的药效团,通过生成针对感兴趣靶点的药效团以实现DTI预测。

1.2.1　基于分子指纹的方法

基于分子指纹的方法识别药物分子的结构,根据结构相似性对药物分子进行聚类,计算成本较低,完全依赖于药物结构。基于分子指纹的方法关注整个药物分子,而不仅是具有生物活性的部分,因此有效避免了过拟合现象,但同时也使得模型容易受到非必要特征和有限评价空间的影响[34]。这类方法由于计算成本低同时兼具有效性,在当时是虚拟筛选的重要方法之一[35-36]。相似性集成方法[37]基于药物的化学相似性来比较药物靶点,使用统计模型排序以避免由于偶然性导致的化学相似性。

1.2.2　基于定量构效关系的方法

基于QSAR的DTI预测方法首先生成药物分子有关理化性质和结构特性的描述符,构建模型来识别描述符与靶点相互作用的关系,最后预测使用相同描述符编码的测试药物分子的活性[38]。基于QSAR方法的结果不仅取决于初始药物的质量,还取决于所选择的最优描述符集。过多的特征会给模型增加噪声,为DTI预测带来负担,因此选用特征选择技术去除不必要的特征以最

小化模型自由度数量是十分必要的。目前一些基于QSAR的方法通过计算信息增益[39]和F-分数[40]来选择特征,使用正交前向选择和后向消除算法来选择特征子集[41],采用遗传算法[42]、群优化[43]和输入敏感性分析[44]等算法来进行分析研究。基于QSAR方法不需要生产测试,就可有效识别新化合物的特征。

1.2.3 基于药效团模型的方法

药效团是指药物分子中对药物活性作出重要贡献的特征元素及其空间排列形式。有效的药效团通常包含与靶点相互作用的官能团、非共价相互作用类型以及原子间距等信息。基于药效团模型的方法通过叠加多个具有生物活性的药物分子生成针对感兴趣靶点的药效团以实现DTI预测[45],涉及药物分子的2D或3D结构表示,并需要结合分子活性来确定重叠位点。

药物分子通常采用基于点或基于属性的技术来进行对齐。基于点的对齐技术通过最小化欧氏距离来叠加成对的原子特征,并使用均方根距离(root mean squared distance,RMSD)来最大化重叠[46]。基于属性的对齐技术则利用分子描述符来计算高斯函数表示的相互作用能分布,将高斯函数间的重叠程度作为打分函数来最大化重叠[47]。

目前已有一些比较成熟的药效团生成软件包,例如Phase[48]、Catalyst[49]、DISCO[50]、GASP[51]等,分别使用不同方法来进行分子排列和药效团特征提取。

与基于分子对接的方法不同,基于药物结构的方法并不需要了解蛋白质的结构,在早期的虚拟筛选中取得了很好的效果,然而当与目标靶点产生相互作用的药物数量不足时,基于药物结构的方法则表现不佳。

1.3 基于文本挖掘的方法

科学文献为研究人员提供了大量丰富的信息,它是评估特定领域最新技术的起点,也是构建研究假设的基石。随着生物医学研究产生的实验数据量和发表论文数量的指数级增长,手工从文献数据库中检索信息并与实验数据相结合是十分困难的。文本挖掘在药物发现领域已经取得了广泛的应用,基于文本挖掘的方法可以从科学文献中自动挖掘药物与蛋白质的关联,发现可以用于干预治疗疾病的候选药物靶点[52]。

Campillos等人对药品说明书的副作用描述进行分类,将具有相似副作用的药物与其已知的靶点组合在一起,推断作用在相同靶点的药物组[53];Zaravinos等人创建血管生成成分的文献网络,研究其在尿路上皮细胞癌中的作用[54];Wu等人构建了基于网络的冠心病研究平台的知识库,采用文本挖掘技术结合人工确认从文献摘要中提取冠心病相关基因,并将基因映射到生物学通路,便于分子机制剖析和新药物的发现[55];PharmGKB[56]、Chem2bio2rdf[57]等方法也是基于语义相似性来衡量药物与靶点之间的关联。通过寻找文献中生物实体的共现关系同样是发现药物与靶点关联的一种简单流行的技术。DTI受基因调控,基因影响着药物与靶点的结合能力,Zhu等人提出一种从文献的共现信息中挖掘药物与基因关系的概率模型[58]。发现药物的候选靶点需要全面丰富的信息,将文献挖掘出的结果与靶点的成药性、相关组织中的表达以及已知的副作用等信息联系起来,以便对特定蛋白质作为候选药物靶点的适用性作出综合判断[59]。

基于文本挖掘的方法依赖于海量的生物信息领域文献中蕴含的有效信息,但其通常基于关键字搜索,难以对领域名词进行实体对齐和消歧,也难以解决文献中化合物、基因名称等存在的冗余问题,同时也无法检测出新的生物学发现。

1.4 基于化学基因组的方法

为了解决传统方法成本高、受限多的问题,基于化学基因组的方法[60]在药物发现和药物重定位领域中应运而生,逐渐展现出优势。DTI预测中经常涉及到四种类型的生物实体,即药物、蛋白质、疾病和副作用。基于化学基因组的方法通常将药物的化学空间和靶蛋白的基因组空间整合到统一的药理学空间。这类方法可以利用丰富的生物学药物学数据,但其主要挑战在于缺乏大量真实的DTI样本和经实验验证的DTI负样本。

基于化学基因组的方法如图3所示,又可根据实现方法的差异细化为不同的类别:基于相似性的方法、基于生物特征的方法、基于网络的方法以及基于集成学习的方法。

1.4.1 基于相似性的方法

基于相似性的方法基于“联想有罪推定”假设,即相似的药物倾向于结合相似的靶点,相似的靶点也倾向于与相似的药物结合[61-62]。基于相似性的方法采用的相似性度量策略也各有不同。

最近邻方法根据邻居的信息来进行预测[63-64],逐渐从邻居推荐方法发展为基于邻域的方法。邻居推荐方法一般使用邻居的加权平均信息进行预测。基于邻域的方法通过集成邻域以实现更稳健的预测,比如同时使用Jaccard相似度、Cosine相似度和Pearson相关相似度来计算相似度得分。由于最近邻方法一般采用比较简单的相似函数,研究往往将最近邻方法与其他方法结合以提升模型的性能。加权最近邻算法[65]使用数据集中已知药物分子的化学和相互作用信息构建新药物的相互作用得分谱,可用作预测未知DTI。基于相似等级的预测器[66]计算趋势指数和逆趋势指数两项指标,分别衡量每个药物-靶点对倾向于具有相互作用和不具有相互作用的可能性。

二分局部模型(bipartite local model,BLM)[67]分别训练基于药物结构相似性的药物局部模型和基于蛋白质序列相似性的靶点局部模型,从药物端和靶点端生成

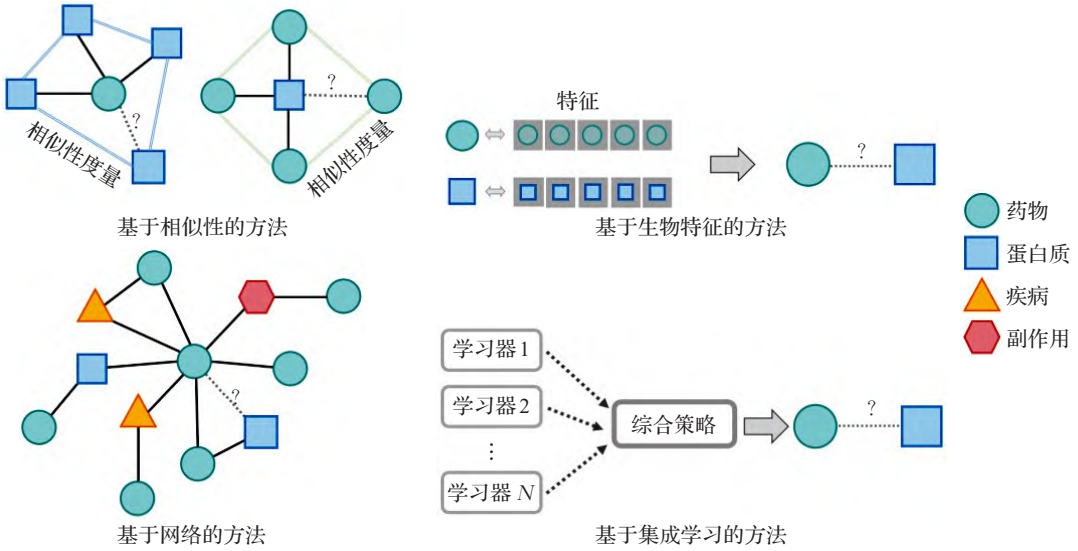


图3　基于化学基因组的DTI预测方法分类

Fig.3　Taxonomy of chemogenomic-based DTI prediction methods

两个独立的预测结果,通过聚合两端结果获得最终的预测结果。由于BLM方法对于未知的候选药物和靶点相互作用的预测具有一定的局限性,BLM-NII[68]将邻居的交互配置文件推断作为标签集成到现有的BLM方法中,有助于发现潜在的DTI。ECkNN[69]则扩展了BLM,使其集成了中心感知回归技术,采用丰富的特征来表征相似性空间中的药物和靶点。

矩阵分解方法则将DTI预测视为矩阵补全问题。DTI矩阵被分解为两个潜在的特征矩阵,两矩阵相乘可以逼近原始矩阵。KBMF2K[70]基于药物分子之间的化学相似性和靶蛋白之间的基因组相似性定义了两个核矩阵,结合了贝叶斯概率公式,提出了主动学习策略以及概率矩阵分解方法;MSCMF[71]扩展了协同过滤的加权低秩近似,集成了化学结构相似性、基因组序列相似性、ATC相似性、GO相似性和PPI网络相似性多个相似性矩阵,将药物和靶点投影到一个共同的低秩特征空间中来预测潜在的DTI;Ezzat等人提出了两种矩阵分解方法GRMF和WGRMF[72],通过图正则化的方式隐式地执行流形学习,并将原始DTI矩阵中的0转换为相互作用似然值;DNILMF[73]则通过融合药物和靶点相似性矩阵,并添加约束来计算关联。

最近邻方法通常基于一阶相似度构建邻域,使得相似度无法传递。局部模型复杂度比全局模型低得多,但一般难以处理药物与靶点均未参与训练的情况。矩阵分解方法通常具有更好的性能,但同时伴随高计算复杂度。

1.4.2　基于生物特征的方法

基于生物特征的方法的核心思想是提取药物和靶点的生物学特征,将药物和靶点信息转化为特征向量,从而推断潜在的DTI。

研究人员从不同角度来描述药物和靶点的信息。Tabei等人运用张量积方法提取与DTI相关的药物化学亚结构和蛋白质结构域之间的潜在关联特征[74];Xiao等人对药物分子指纹进行离散傅里叶变换以获得药物特征,并根据假氨基酸组成提取G蛋白偶联受体特征[75];PDTPS[76]则整合了蛋白质序列和药物化学结构。

研究人员使用不同的特征提取方法来捕获特征。DTINet[77]应用带重启的随机游走和奇异值分解,寻找从药物空间到蛋白质空间的最佳投影,使得药物的投影特征向量在几何上接近已知相互作用的靶点特征向量;Hirohara等人提出将药物特征使用分子线性输入规范序列表示,之后转换为二维矩阵,使用神经网络提取药物特征[78];Wang等人设计堆叠自动编码器模型从药物分子结构和蛋白质序列中提取具有高代表性的特征[79];Shi等人利用伪位置特异性评分矩阵和FP2分子指纹技术提取药物和靶点的特征,并采用LASSO对提取的特征信息进行降维处理[80];Ru等人采用基于距离的top-n-gram算法提取蛋白质的特征,并使用通用描述符表示药物的结构信息[81]。

1.4.3　基于网络的方法

基于网络的方法构建了拓扑网络,采用网络来建模多实体类型之间复杂的关联,之后基于现有的数据资源进行计算和推理,进行DTI预测。根据构建的网络中拓扑结构的不同,可以细分为基于二分网络的方法和基于复杂网络的方法。

基于二分网络的方法大多数仅集成了药物结构、蛋白质序列和DTI信息,仅包含药物-靶点这一种网络拓扑结构。Cheng等人开发了DTI推理方法,其中基于网络的推理仅使用药物-靶点二分网络的拓扑相似性来推断已知药物的新靶点[82]。基于药物-靶点二分网络的推理方法为基于分子多重药理学空间发现潜在DTI提供了有效手段。

复杂网络在数据建模方面具有很强的灵活性,包含药物、蛋白质、疾病、副作用等多种生物实体和多种生物实体间关联的网络可以抽象为复杂网络,DTI预测任务也可相应定义为网络表示学习下游的链接预测任务。基于复杂网络的方法能够利用多样化信息和多视图视角,更敏锐地捕捉生物实体之间的关系。NRWRH[83]将药物结构相似性网络、蛋白质序列相似性网络和已知DTI网络集成到复杂网络中,使用带重启的随机游走方法推断潜在的DTI。DASPfind[84]同样结合了药物相似性网络、蛋白质相似性网络和DTI二分网络形成了复杂网络,但采用指数衰减函数来融合连接药物-靶点对的所有通路,通过遍历所有通路来预测DTI。

深度学习是当前人工智能领域的研究热点,在语音识别,图像识别和自然语言处理等众多任务中展现出强大的性能优势。近年来,将深度学习方法应用于药物发现的研究也在不断增加[85-86]。基于深度学习的方法使用神经网络自动学习多类型实体的特征以及它们之间复杂的关联关系。目前较为先进的DTI预测方法采用基于深度学习的网络表示学习方法,卷积神经网络(convolutional neural networks,CNN),图卷积网络(graph convolutional networks,GCN),图注意力网络(graph attention networks,GAT)等深度学习方法均在DTI预测领域得以广泛应用,并取得了较好的实验结果,表1展示了近年来具有代表性的基于网络结合深度学习的DTI预测方法。

深度学习方法具有适应能力强、应用范围广、可移植性强等优点,但同时也不可避免地存在可解释性偏弱的问题。可解释性在某种程度上反映了人类对模型决策或预测结果的理解程度。DTI预测任务属于与人类生命安全息息相关的生物医药领域,模型的可解释性也与用户对模型的信任程度紧密相关。强可解释性的模型更容易获得领域研究人员的接纳与认可,便于进一步地推广应用。大多数深度学习模型都是由数据驱动的“黑盒”模型,高度依赖于训练数据,可解释性普遍较差。用户难以理解每个参数和最终结果之间的关联,进而难以对其进行调试和优化。

1.4.4 基于集成学习的方法

除了前面提到的单一学习方法外,DTI预测任务还可以采用bagging、boosting、stacking等主流集成学习框架,对多种学习方法进行偏差权衡,提升预测算法的性能。

Zhang等人提出一种stacking集成框架,采用支持向量机作为元学习器以获得更好的DTI预测结果[96];Ezzat等人提出了利用特征降维和集成学习的DTI预测框架,采用三种降维方法,训练决策树和岭回归两个基础分类器,构建EnsemDT和EnsemKRR两个集成模型变体[97];Sharma等人提出了一种bagging的集成框架BE-DTI'[98],采用降维和主动学习方法来处理数据不平衡的DTI预测任务;Yang等人提出基于stacking的集成框架NegStacking[99],使用逻辑回归作为元学习器自适应地组合弱学习器,同时应用特征子间距和超参数扰动来增强集成多样性;Pliakos等人提出树集成学习和输出空间重建方法[100],在重建网络上学习多输出双聚类树的集成;Xuan等人提出基于非负矩阵分解和梯度提升决策树(gradient boosting decision tree,GBDT)的方法NGDTP[101],采用基于矩阵分解的网络表示学习方法来学习药物和蛋白质的低维向量表示,并设计了基于GBDT的预测模型,通过构建多个决策树来获得药物与蛋白质的关联分数;Thafar等人提出基于图神经网络和boosting集成方法的模型DTi2Vec[102],采用AdaBoost和XGBoost两个集成分类器来实现DTI预测。

集成学习方法通过集成不同的信息集来提升预测算法的性能,并且充分结合了不同方法来解决DTI中未

表1 基于网络结合深度学习的DTI预测方法

Table 1 Network-based DTI prediction methods combined with deep learning

方法	特点	数据集
NeoDTI[87]	采用神经网络进行邻域信息聚合和拓扑消息传递	Luo数据集
DeepDTnet[88]	采用堆叠式去噪自动编码器学习药物与蛋白质的低维向量表示	基于DrugBank、TTD、PharmGKB等数据库构建数据集
GCN-DTI[89]	构建药物-蛋白质对(drug-protein pair,DPP)网络,利用GCN将药物和蛋白质特征与DPP网络的结构信息相结合	Yamanashi数据集、DrugBank5.0.3数据集
EEG-DTI[90]	堆叠三层GCN以聚合多类型的邻居节点的特征,将每个节点的表示连接在不同层中,有效避免了梯度消失	Luo数据集、Yamanashi数据集
DTI-GAT[91]	利用GAT的自注意力机制捕获网络的拓扑结构信息	Yamanashi数据集、DrugBank_approved数据集
MultiDTI[92]	构建了异构网络的联合学习框架,采用词级CNN将药物和靶点的一维序列信息转化为固定长度的序列特征	Luo数据集
DTI-MGNN[93]	使用GAT来学习DPP网络的拓扑图和特征图,利用GCN有效结合拓扑结构和语义特征	基于DrugBank、HPRD、SIDER等数据库构建数据集
HGDTI[94]	利用双向长短时记忆网络来提取药物与蛋白质特征,采用注意力机制聚合异构邻居	Luo数据集
DeepMGT-DTI[95]	利用包含多层图信息的transformer网络来捕获药物分子特征,采用CNN捕捉靶点序列中的局部残差信息	基于KEGG、PubChem、DrugBank等数据库构建数据集

知相互作用的问题。集成学习方法性能较强的原因在于优化了特征提取过程,缓解了正负样本严重失衡的问题,捕获到了复杂的药物与靶点的隐藏特征。然而,集成学习方法由于集成多个信息集,使得模型十分庞大,训练过程较为复杂,效率普遍不高。

1.4.5　方法比较

基于化学基因组的方法将药物与蛋白质映射到同一空间来预测DTI,是近年来发展较快且较有前景的DTI预测方法。表2对比分析了基于化学基因组的四类方法,分别阐述其优势与局限性,便于研究人员在现有基于化学基因组方法的基础上开展深入研究。

基于相似性的方法利用了已有一定研究基础的相似性打分函数,但严重依赖于可见数据,其泛化能力较差;基于生物特征的方法建立了药物与蛋白质之间的间接连接,但人工提取特征过程使得结果依赖于专家的先验知识和直观感受;基于网络的方法利用网络结构来建模数据,同时与深度学习方法结合也是目前研究的一大热点趋势,然而深度学习方法的"黑盒"属性使得模型面向领域研究人员的解释推广困难重重;基于集成学习的方法集成多个信息集以提升模型性能,但同时也扩大了计算规模,增大了计算成本。

2　DTI预测的计算方法对比分析

根据对文献的梳理和总结,基于分子对接的方法对药物和蛋白质的结合构象进行模拟;基于药物结构的方法分析药物结构预测其药理活性;基于文本挖掘的方法从文本数据中挖掘DTI;基于化学基因组的方法则将药物的化学空间和靶蛋白的基因组空间整合到统一的药理学空间。每类方法的研究重点不同,对原始数据的需求不同,应用场景也有所不同。本章将对四类方法的数据类型、方法优势局限性展开分析,并讨论不同方法的应用场景。

2.1　数据类型

不同DTI预测方法对DTI预测任务的定义不同,对于原始数据的需求也各不相同。一般来说,所利用的原始数据越接近分子相互作用底层原理层面,方法的可解释性越强,方法的复杂度也越高。

基于分子对接的方法模拟药物与靶点的结合过程,基于丰富的药理学知识,依赖于蛋白质结构来计算相互作用能。蛋白质结构获取的主要来源是PDB蛋白质数据库[103],PDB中包含上万种蛋白质结构,其中大部分是通过X射线晶体学确定的,少部分是通过核磁共振波谱法确定的。

基于药物结构的方法收集分析与感兴趣靶点相互作用的化合物结构,从而发现对其生物活性最重要的结构性质,依赖于通过分子力学或量子力学等工具生成的药物分子描述符,包含分子量、可旋转键、原子间距、电负性、极化率、对称性等丰富的特征信息。药物分子的相关数据可从DrugBank[104]、PubChem[105]、ChemDB[106]等数据库中获取。

基于文本挖掘的方法从科学文献等文本数据中发现药物与靶点的关系,主要依赖于PubMed文献数据库[107]、药物说明书数据以及现有的一些知识库。

基于化学基因组的方法则利用已经观测到的DTI正样本和未被观测到的DTI负样本,DTI数据可以从DrugBank[104]等数据库获得或通过实验测定,同时可以引入疾病、副作用等实体类型以及药物结构与蛋白质序列等特征,增加信息源以辅助DTI预测。

2.2　优势局限性分析

正如计算机科学中的"没有免费的午餐定理",没有一种学习算法可以在所有情形下总是产生最准确的模型,这意味着每种DTI预测方法都有自身独特的优势,同时不可避免地伴随着某些方面的局限性。表3列出了基于分子对接的方法、基于药物结构的方法、基于文本挖掘的方法和基于化学基因组的方法各自的优势和局限性。

基于分子对接的方法从分子相互作用机制视角出发,能够模拟药物小分子与靶点大分子的结合构象,结果的可解释性很强。然而此类方法并不适用于蛋白质结构过于复杂或无法获得蛋白质结构的情况,同时需要对所有可能的构象进行采样,使得效率普遍较低。

基于药物结构的方法则充分利用了药物的化学结构和理化性质,并不需要掌握蛋白质的3D结构。由于其严重依赖于药物结构与靶点的关联关系,当已知与感

表2　基于化学基因组的DTI预测方法比较

Table 2　Comparison of chemogenomic-based DTI prediction methods

算法	优势	局限性
基于相似性的方法	(1)不需要特征选择和特征提取 (2)相似性打分函数有一定的研究基础	(1)往往具有高计算复杂度,难以扩展到大规模数据集上 (2)严重依赖于可见数据,泛化能力较差
基于生物特征的方法	(1)药物与蛋白质投影到同一特征空间 (2)建立了药物与蛋白质的长间接连接	(1)大多方法需要人工提取特征 (2)结果极大依赖于专家的先验知识及直观感受
基于网络的方法	(1)网络在数据建模方面具有很强的灵活性 (2)与深度学习方法结合展现出性能优势	(1)尚未完全捕获网络中的结构与语义信息 (2)深度学习方法可解释性较差
基于集成学习的方法	(1)优化了特征提取过程 (2)有效缓解了类不平衡的问题	集成模型计算规模庞大,训练过程复杂,效率普遍不高

表3 DTI预测的计算方法比较

Table 3 Comparison of computational DTI prediction methods

算法	优势	局限性
基于分子对接的方法	预测结果的可解释性强	(1)不适用于结构过于复杂或无法获得结构的蛋白质 (2)搜索空间与计算规模巨大,效率普遍较低
基于药物结构的方法	不需要了解蛋白质的结构	与目标靶点关联的药物数量很少时,表现不佳
基于文本挖掘的方法	(1)利用了文献中蕴含的有价值信息 (2)文本挖掘方法已有一定研究基础	(1)难以对领域名词进行实体对齐和实体消歧 (2)极大受限于语义表达的多样性和冲突性
基于化学基因组的方法	抽象为具有研究基础的机器学习任务	(1)缺乏大量DTI正样本和经实验验证的DTI负样本 (2)可解释性普遍较差

兴趣靶点产生相互作用的药物数量很少时,基于药物结构的方法则表现不佳。

基于文本挖掘的方法自动挖掘文献等文本数据中蕴含的有价值信息,文本挖掘方法也有一定研究基础。然而文本数据中存在大量药物与蛋白质名称不统一、冗余现象,同时方法极大受限于语义表达的差异,表达的多样性和冲突性严重限制其性能。

基于化学基因组的方法将药物和蛋白质映射到化学基因组空间,抽象为分类、排序与回归等任务,可以借助已有研究基础的机器学习方法进行DTI预测。但目前缺乏大量真实的DTI正样本和经实验验证的DTI负样本数据,同时部分机器学习方法也被喻为“黑盒”,其可解释性普遍较差。

2.3 应用场景

虽然没有一种方法可以被称为最优模型,但在具体研究场景下某些方法独特的优势对于任务起着关键作用,以至于可以忽略其不足与缺陷,成为基于当前应用场景的最优方法。研究人员在具体应用场景中如何选择合适的方法以解决当前任务是十分重要的。

基于分子对接的方法模拟药物与蛋白质的结合过程,能够预测药物小分子与靶点结合位点的结合构象。由于药物发现与生命安全紧密相关,模型的可解释性则凸显得更加重要,药物的开发必须具有严格的原理论证和完备的药理学知识支持。基于分子对接的方法可以从药物与靶点分子结合机制层面进行解释,适合用于药物靶点初筛之后的复核与可行性分析。

基于药物结构的方法利用药物分子的理化性质和结构特性预测其药理活性,尝试发现药物分子中对其活性作出重要贡献的结构以推理DTI关系,其预测结果具有一定的解释性,适合用于药效团的推理发现以辅助药物作用机制的研究。同时有些蛋白质的结构十分复杂难以分析,如离子通道和G蛋白偶联受体,甚至还有些蛋白质结构无法获得,在这种情况下基于药物结构的方法提供了一种间接的研究思路,仅利用药物结构尝试解释药物与靶点相互作用机制,对后续药物发现提供很大帮助。

基于文本挖掘的方法利用生物信息文献和数据库中的文本数据自动挖掘DTI关系,是自然语言处理方法在DTI预测领域的应用。与前两种方法相比,其计算成本显著下降,并严重依赖于文献数据库中已有的DTI关系,并且不能检测出新的药物学发现,适用于整合分析现有的DTI知识,进行DTI预测的前期调研和实验数据集的构建。

基于化学基因组的方法将药物的化学空间和蛋白质的基因组空间整合到药理学空间,它完全脱离了药物与蛋白质分子相互作用机制,仅从药物与蛋白质是否会发生相互作用的结果反向学习药物与蛋白质的特征表示。虽然部分基于化学基因组的方法捕获的特征难以解释,但其在一定程度上是高效准确的,适合用于药物靶点的初筛,缩小药物蛋白质结构分析与对接模拟的范围,并用于发现潜在的DTI关系。

3 挑战与展望

目前国内外已有许多研究团队致力于基于计算的DTI预测方法的探索,DTI预测技术不断更新迭代,性能也在不断攀升,然而还存在一些研究缺口与方法局限,这同时也为DTI预测的计算方法指明了未来的研究方向。

(1)问题定义

基于分子对接的方法和基于药物结构的方法从分子相互作用机制角度出发,其余大部分基于机器学习的DTI预测模型通常是在过于简化的设置下构建和评估的,这可能会导致结果过于乐观并偏离真实情况。DTI预测可以归纳为分类任务、排序任务或回归任务。常见的任务抽象是忽略分子浓度和定量亲和力等其他重要因素,将DTI预测视为一个简单的二分类任务。考虑到药物-靶点的结合亲和力具有剂量依赖性,DTI预测问题更适合归纳为排序任务或者回归任务。归纳为排序任务获得与感兴趣靶点可能发生相互作用的药物列表。归纳为回归任务可以获得药物与靶点结合强度值,使得模型更适合推广至真实的药物开发任务中。

(2)数据集构建

DTI预测方法依赖于相似药物和相似靶点的确定性,现有数据集中缺乏统一定义,难以判定其一致性。如何对从不同数据源获得的异构数据中的药物和靶点

进行统一的标识是一个挑战。同时在生物医学大数据时代,数据稀疏并伴有部分丢失是十分常见的,DTI预测数据集普遍存在噪声、数据缺失和数据不一致问题。如何基于大量稀疏具有噪声的DTI数据进行数据填补也是未来一个重要的研究课题。

DTI预测训练过程需要大量经实验验证的负样本数据集,如何建立无偏负样本DTI数据集是关键的一步。现有方法大多使用未知DTI构建负样本,然而数据集中未知DTI并不与药物和靶点真正无相互作用等价,同时少量已知DTI与大量未知DTI会导致严重的数据失衡。目前已有一些集成学习方法有效缓解了类不平衡问题[98],同时也有一些半监督方法在不使用负样本情况下进行了尝试[108-110]。未来研究重点可以从有监督学习向半监督学习倾斜,半监督学习仅使用少量有标签数据和大量无标签数据,在数据集不平衡情况下可能产生比有监督学习更可靠的预测。

(3)方法设计

截至目前,基于网络结合深度学习方法进行DTI预测仍是较为高效精确的方法,然而现有方法尚未完全捕获复杂网络中大量丰富的信息。

药物结构和蛋白质序列中蕴含着对于DTI预测任务十分重要的药理学知识。Transformer具有超强的序列建模能力与全局信息感知能力,目前在众多自然语言处理和计算机视觉任务上已经刷新了最优性能[111]。利用transformer对序列特征进行编码可能会获得更具药理学意义的药物与蛋白质低维向量表示。此外,网络中的拓扑结构也包含着关键的上下文信息。基于无标签数据的自监督学习方法在DTI预测领域颇具研究前景,预训练技术[112]和图对比学习[113]等自监督图表示学习方法在DTI预测任务上也具有一定的研究潜力,可能会捕获到更加丰富的拓扑特征和语义信息,进而提升DTI预测性能。

基于分子对接的方法与基于药物结构的方法均基于药理学知识从分子视角出发,而结合深度学习的DTI预测方法却很难从生物学和病理学的角度来理解潜在的药物作用机制。由于药物开发与人类的生命安全息息相关,不具备可解释性的DTI预测方法使得模型的推广应用困难重重。寻找DTI预测方法中高准确性和强可解释性之间的平衡是目前学术界和工业界共同的愿景。

(4)结果评估

DTI预测任务目前还没有统一专用的评估指标。基于分子对接方法模拟药物与靶点结合的复合物构象,可靠的评估指标需要将真实复合物构象排在首位,常用的评估方式为计算模拟构象与实验确定的真实构象之间的均方根偏差,针对具有对称官能团或全分子对称的药物分子,计算其最小对称校正均方根偏差。采用结合亲和力预测的方法通常计算预测亲和力值与实验测试得到的亲和力值之间的Pearson相关系数。考虑其相关性不一定是线性的,也可选用Spearman相关系数。采用机器学习的方法则常常使用灵敏度、特异性、平均百分比排名(mean percentage ranking,MPR)、精确召回曲线下面积(area under the precision-recall curve,AUPR)、ROC曲线下面积(area under the ROC curve,AUROC)等评估指标。由于DTI预测任务自身的独特性,同时存在严重的正负样本不均衡问题,设计一个充分考虑代价成本、简洁性、可信性、准确性和因果相关性的评估指标是十分必要的,引领着DTI预测方法进一步改进优化的方向。

4　结语

药物开发属于知识密集、技术密集、资金密集型的技术领域,同时具有高成本、长周期、高风险等特点,是衡量一个国家医药行业技术发展的重要标志。经过数十年的药物研发进程,我国与国际先进水平之间的差距逐渐缩小,且取得了阶段性的科研成果,然而我国在药物研发最为关键的靶点选择上,原始创新的新药却还是很少,靶点选择仍是药物开发过程中的一大挑战。DTI计算预测为药物开发提供了重要的初筛依据,推动着药物开发的进程,引领着我国药物研发领域的革新,具有深刻的科研意义和应用价值。

本文将现有DTI预测的主流计算方法根据不同的数据类型和应用场景分为基于分子对接的方法、基于药物结构的方法、基于文本挖掘的方法与基于化学基因组的方法四大类,对每一类方法进行详细介绍说明,并对它们需要的数据类型、优势局限性以及应用场景展开了对比分析,之后讨论了现有的研究缺口与方法局限,对未来研究方向进行了展望,为未来研究人员提供参考和借鉴。

参考文献:

[1] DREWS J.Drug discovery:a historical perspective[J].Science,2000,287(5460):1960-1964.

[2] MASOUDI-NEJAD A,MOUSAVIAN Z,BOZORGMEHR J H.Drug-target and disease networks:polypharmacology in the post-genomic era[J].In Silico Pharmacology,2013,1(1):1-4.

[3] MASOUDI-SOBHANZADEH Y,OMIDI Y,AMANLOU M,et al.Drug databases and their contributions to drug repurposing[J].Genomics,2020,112(2):1087-1095.

[4] PAUL S M,MYTELKA D S,DUNWIDDIE C T,et al.How to improve R&D productivity:the pharmaceutical industry's grand challenge[J].Nature Reviews Drug Discovery,2010,9(3):203-214.

[5] DICKSON M,GAGNON J P.Key factors in the rising cost

of new drug discovery and development[J].Nature Reviews Drug Discovery,2004,3(5):417-429.
[6] CHENG L,HAN X,ZHU Z,et al.Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2[J].Briefings in Bioinformatics,2021,22(2):1442-1450.
[7] CHONG C R,SULLIVAN D J.New uses for old drugs[J].Nature,2007,448(7154):645-646.
[8] ZENG X,LIU L,LÜ L,et al.Prediction of potential disease-associated microRNAs using structural perturbation method[J].Bioinformatics,2018,34(14):2425-2432.
[9] ZHANG X,ZOU Q,RODRIGUEZ-PATON A,et al.Meta-path methods for prioritizing candidate disease miRNAs[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2017,16(1):283-291.
[10] SHI H,WU Y,ZHENG Z Q,et al.A discussion of micrornas in cancers[J].Current Bioinformatics,2014,9(5):453-462.
[11] ZENG X,LIAO Y,LIU Y,et al.Prediction and validation of disease genes using HeteSim scores[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2016,14(3):687-695.
[12] ZENG J,LI D,WU Y,et al.An empirical study of features fusion techniques for protein-protein interaction prediction[J].Current Bioinformatics,2016,11(1):4-12.
[13] WANG Z,ZOU Q,JIANG Y,et al.Review of protein subcellular localization prediction[J].Current Bioinformatics,2014,9(3):331-342.
[14] EZZAT A,WU M,LI X L,et al.Computational prediction of drug-target interactions using chemogenomic approaches:an empirical survey[J].Briefings in Bioinformatics,2019,20(4):1337-1357.
[15] WANG T,WU M B,ZHANG R H,et al.Advances in computational structure-based drug design and application in drug discovery[J].Current Topics in Medicinal Chemistry,2016,16(9):901-916.
[16] BAMBINI S,RAPPUOLI R.The use of genomics in microbial vaccine development[J].Drug Discovery Today,2009,14(5/6):252-260.
[17] KITCHEN D B,DECORNEZ H,FURR J R,et al.Docking and scoring in virtual screening for drug discovery:methods and applications[J].Nature Reviews Drug Discovery,2004,3(11):935-949.
[18] TAYLOR J S,BURNETT R M.DARWIN:a program for docking flexible molecules[J].Proteins:Structure,Function,and Bioinformatics,2000,41(2):173-191.
[19] CONNOLLY M L.Analytical molecular surface calculation[J].Journal of Applied Crystallography,1983,16(5):548-558.
[20] KATCHALSKI-KATZIR E,SHARIV I,EISENSTEIN M,et al.Molecular surface recognition:determination of geometric fit between proteins and their ligands by correlation techniques[J].Proceedings of the National Academy of Sciences,1992,89(6):2195-2199.
[21] YANG C,CHEN E A,ZHANG Y.Protein-ligand docking in the machine-learning era[J].Molecules,2022,27(14):4568.
[22] HALPERIN I,MA B,WOLFSON H,et al.Principles of docking:an overview of search algorithms and a guide to scoring functions[J].Proteins:Structure,Function,and Bioinformatics,2002,47(4):409-443.
[23] DIAS R,DE AZEVEDO J,WALTER F.Molecular docking algorithms[J].Current Drug Targets,2008,9(12):1040-1047.
[24] HALGREN T A.Merck molecular force field.I.Basis,form,scope,parameterization,and performance of MMFF94[J].Journal of Computational Chemistry,1996,17(5/6):490-519.
[25] BÖHM H J.The computer program LUDI:a new method for the de novo design of enzyme inhibitors[J].Journal of Computer-Aided Molecular Design,1992,6(1):61-78.
[26] RAREY M,KRAMER B,LENGAUER T,et al.A fast flexible docking method using an incremental construction algorithm[J].Journal of Molecular Biology,1996,261(3):470-489.
[27] JAIN A N.Surflex:fully automatic flexible molecular docking using a molecular similarity-based search engine[J].Journal of Medicinal Chemistry,2003,46(4):499-511.
[28] DEWITTE R S,SHAKHNOVICH E I.SMoG:de novo design method based on simple,fast,and accurate free energy estimates.1.Methodology and supporting evidence[J].Journal of the American Chemical Society,1996,118(47):11733-11744.
[29] MITCHELL J B O,LASKOWSKI R A,ALEX A,et al.BLEEP—potential of mean force describing protein-ligand interactions:I.Generating potential[J].Journal of Computational Chemistry,1999,20(11):1165-1176.
[30] SHIMADA J,ISHCHENKO A V,SHAKHNOVICH E I.Analysis of knowledge-based protein-ligand potentials using a self-consistent method[J].Protein Science,2000,9(4):765-775.
[31] VELEC H F G,GOHLKE H,KLEBE G.DrugScoreCSD knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction[J].Journal of Medicinal Chemistry,2005,48(20):6296-6303.
[32] FEHER M.Consensus scoring for protein-ligand interactions[J].Drug Discovery Today,2006,11(9/10):421-428.
[33] ACHARYA C,COOP A E,POLLI J,et al.Recent advances in ligand-based drug design:relevance and utility of the conformationally sampled pharmacophore approach[J].Current Computer-Aided Drug Design,2011,7(1):10-22.
[34] AUER J,BAJORATH J.Molecular similarity concepts and search calculations[M]//Bioinformatics.[S.l.]:Humana Press,

2008:327-347.
[35] HUTTER M C.Graph-based similarity concepts in virtual screening[J].Future Medicinal Chemistry,2011,3(4):485-501.
[36] WILLETT P.Similarity-based virtual screening using 2D fingerprints[J].Drug Discovery Today,2006,11(23/24):1046-1053.
[37] KEISER M J,ROTH B L,ARMBRUSTER B N,et al.Relating protein pharmacology by ligand chemistry[J].Nature Biotechnology,2007,25(2):197-206.
[38] TROPSHA A.Best practices for QSAR model development,validation,and exploitation[J].Molecular Informatics,2010,29(6/7):476-488.
[39] KENT J T.Information gain and a general measure of correlation[J].Biometrika,1983,70(1):163-173.
[40] CHEN Y W,LIN C J.Combining SVMs with various feature selection strategies[M]//Feature extraction.Berlin,Heidelberg:Springer,2006:315-324.
[41] MAO K Z.Orthogonal forward selection and backward elimination algorithms for feature subset selection[J].IEEE Transactions on Systems,Man and Cybernetics,Part B (Cybernetics),2004,34(1):629-634.
[42] WHITLEY D,SUTTON A M.Genetic algorithms-a survey of models and methods[M]//Handbook of natural computing.Berlin,Heidelberg:Springer,2012:637-671.
[43] GOODARZI M,FREITAS M P,JENSEN R.Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3 inhibitory activities[J].Journal of Chemical Information and Modeling,2009,49(4):824-832.
[44] MUELLER R,RODRIGUEZ A L,DAWSON E S,et al.Identification of metabotropic glutamate receptor subtype 5 potentiators using virtual high-throughput screening[J].ACS Chemical Neuroscience,2010,1(4):288-305.
[45] WOLBER G,SEIDEL T,BENDIX F,et al.Molecule-pharmacophore superpositioning and pattern matching in computational drug design[J].Drug Discovery Today,2008,13(1/2):23-29.
[46] POPTODOROV K,LUU T,LANGER T,et al.Pharmacophores and pharmacophores searches[M]//Methods and principles in medicinal chemistry.Weinheim,Germany:Wiley-VCH,2006:17-47.
[47] GOODFORD P J.A computational procedure for determining energetically favorable binding sites on biologically important macromolecules[J].Journal of Medicinal Chemistry,1985,28(7):849-857.
[48] DIXON S L,SMONDYREV A M,KNOLL E H,et al.PHASE:a new engine for pharmacophore perception,3D QSAR model development,and 3D database screening:1.Methodology and preliminary results[J].Journal of Computer-Aided Molecular Design,2006,20(10):647-671.
[49] KUROGI Y,GUNER O F.Pharmacophore modeling and three-dimensional database searching for drug design using catalyst[J].Current Medicinal Chemistry,2001,8(9):1035-1055.
[50] MARTIN Y C,BURES M G,DANAHER E A,et al.A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists[J].Journal of Computer-Aided Molecular Design,1993,7(1):83-102.
[51] JONES G,WILLETT P,GLEN R C.A genetic algorithm for flexible molecular overlay and pharmacophore elucidation[J].Journal of Computer-Aided Molecular Design,1995,9(6):532-549.
[52] FLEUREN W W M,ALKEMA W.Application of text mining in the biomedical domain[J].Methods,2015,74:97-106.
[53] CAMPILLOS M,KUHN M,GAVIN A C,et al.Drug target identification using side-effect similarity[J].Science,2008,321(5886):263-266.
[54] ZARAVINOS A,VOLANIS D,LAMBROU G I,et al.Role of the angiogenic components,VEGFA,FGF2,OPN and RHOC,in urothelial cell carcinoma of the urinary bladder[J].Oncology Reports,2012,28(4):1159-1166.
[55] WU L,LI X,YANG J,et al.CHD@ ZJU:a knowledge-base providing network-based research platform on coronary heart disease[J].Database,2013.
[56] HEWETT M,OLIVER D E,RUBIN D L,et al.PharmGKB:the pharmacogenetics knowledge base[J].Nucleic Acids Research,2002,30(1):163-165.
[57] CHEN B D,JIAO X,WANG D,et al. Chem2Bio2RDF:a semantic framework for linking and data mining chemogenomic and systems chemical biology data[J].BMC Bioinformatics,2010,11:255.
[58] ZHU S,OKUNO Y,TSUJIMOTO G,et al.A probabilistic model for mining implicit 'chemical compound-gene' relations from literature[J].Bioinformatics,2005,21:245-251.
[59] HOPKINS A L,GROOM C R.The druggable genome[J].Nature Reviews Drug Discovery,2002,1(9):727-730.
[60] YAMANISHI Y.Chemogenomic approaches to infer drug-target interaction networks[J].Data Mining for Systems Biology,2013,939:97-113.
[61] KLABUNDE T.Chemogenomic approaches to drug discovery:similar receptors bind similar ligands[J].British Journal of Pharmacology,2007,152(1):5-7.
[62] SCHUFFENHAUER A,FLOERSHEIM P,ACKLIN P,et al.Similarity metrics for ligands reflecting the similarity of the target proteins[J].Journal of Chemical Information and Computer Sciences,2003,43(2):391-405.
[63] ZHANG W,ZOU H,LUO L,et al.Predicting potential side effects of drugs by recommender methods and ensem-

ble learning[J].Neurocomputing,2016,173:979-987.
[64] ZHANG W,CHEN Y,LIU F,et al.Predicting potential drug-drug interactions by integrating chemical,biological, phenotypic and network data[J].BMC Bioinformatics,2017, 18(1):1-12.
[65] VAN LAARHOVEN T,MARCHIORI E.Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile[J].PloS One,2013,8(6): e66952.
[66] SHI J Y,YIU S M.SRP:A concise non-parametric similarity-rank-based model for predicting drug-target interactions[C]// 2015 IEEE International Conference on Bioinformatics and Biomedicine(BIBM),2015:1636-1641.
[67] BLEAKLEY K,YAMANISHI Y.Supervised prediction of drug-target interactions using bipartite local models[J]. Bioinformatics,2009,25(18):2397-2403.
[68] 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.
[69] BUZA K,PEŠKA L.Drug-target interaction prediction with bipartite local models and hubness-aware regression[J]. Neurocomputing,2017,260:284-293.
[70] GÖNEN M.Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization[J].Bioinformatics,2012,28(18):2304-2310.
[71] ZHENG X,DING H,MAMITSUKA H,et al.Collaborative matrix factorization with multiple similarities for predicting drug-target interactions[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2013:1025-1033.
[72] EZZAT A,ZHAO P,WU M,et al.Drug-target interaction prediction with graph regularized matrix factorization[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2016,14(3):646-656.
[73] HAO M,BRYANT S H,WANG Y.Predicting drug-target interactions by dual-network integrated logistic matrix factorization[J].Scientific Reports,2017,7(1):1-11.
[74] TABEI Y,PAUWELS E,STOVEN V,et al.Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers[J].Bioinformatics,2012,28(18):487-494.
[75] XIAO X,MIN J L,WANG P,et al.iGPCR-drug:a web server for predicting interaction between GPCRs and drugs in cellular networking[J].PloS One,2013,8(8):e72234.
[76] MENG F R,YOU Z H,CHEN X,et al.Prediction of drug-target interaction networks from the integration of protein sequences and drug chemical structures[J].Molecules, 2017,22(7):1119.
[77] LUO Y,ZHAO X,ZHOU J,et al.A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information[J].Nature Communications,2017,8(1):1-13.
[78] HIROHARA M,SAITO Y,KODA Y,et al.Convolutional neural network based on SMILES representation of compounds for detecting chemical motif[J].BMC Bioinformatics,2018,19(19):83-94.
[79] WANG L,YOU Z H,CHEN X,et al.A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network[J].Journal of Computational Biology,2018,25(3):361-373.
[80] SHI H,LIU S,CHEN J,et al.Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure[J].Genomics, 2019,111(6):1839-1852.
[81] RU X,WANG L,LI L,et al.Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm[J].Computers in Biology and Medicine, 2020,119:103660.
[82] CHENG F,LIU C,JIANG J,et al.Prediction of drug-target interactions and drug repositioning via network-based inference[J].PLoS Computational Biology,2012,8 (5):e1002503.
[83] CHEN X,LIU M X,YAN G Y.Drug-target interaction prediction by random walk on the heterogeneous network[J].Molecular BioSystems,2012,8(7):1970-1978.
[84] BA-ALAWI W,SOUFAN O,ESSACK M,et al.DASP-find:new efficient method to predict drug-target interactions[J].Journal of Cheminformatics,2016,8(1):1-9.
[85] GAWEHN E,HISS J A,SCHNEIDER G.Deep learning in drug discovery[J].Molecular Informatics,2016,35(1): 3-14.
[86] EKINS S.The next era:deep learning in pharmaceutical research[J].Pharmaceutical Research,2016,33(11): 2594-2603.
[87] WAN F,HONG L,XIAO A,et al.NeoDTI:neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions[J]. Bioinformatics,2019,35(1):104-111.
[88] ZENG X,ZHU S,LU W,et al.Target identification among known drugs by deep learning from heterogeneous networks[J].Chemical Science,2020,11(7):1775-1797.
[89] ZHAO T,HU Y,VALSDOTTIR L R,et al.Identifying drug-target interactions based on graph convolutional network and deep neural network[J].Briefings in Bioinformatics,2021,22(2):2141-2150.
[90] PENG J,WANG Y,GUAN J,et al.An end-to-end heterogeneous graph representation learning-based framework for drug-target interaction prediction[J].Briefings in Bioinformatics,2021,22(5):430.
[91] WANG H,ZHOU G,LIU S,et al.Drug-target interac-

tion prediction with graph attention networks[J].arXiv:2107.06099,2021.

[92] ZHOU D,XU Z,LI W T,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[J].Bioinformatics,2021,37(23):4485-4492.

[93] LI Y,QIAO G,WANG K,et al.Drug-target interaction predication via multi-channel graph neural networks[J].Briefings in Bioinformatics,2022,23(1):346.

[94] YU L,QIU W,LIN W,et al.HGDTI:predicting drug-target interaction by using information aggregation based on heterogeneous graph neural network[J].BMC Bioinformatics,2022,23(1):1-18.

[95] ZHANG P,WEI Z,CHE C,et al.DeepMGT-DTI:Transformer network incorporating multilayer graph information for drug-target interaction prediction[J].Computers in Biology and Medicine,2022,142:105214.

[96] ZHANG R.An ensemble learning approach for improving drug-target interactions prediction[C]//Proceedings of the 4th International Conference on Computer Engineering and Networks.Cham:Springer,2015:433-442.

[97] EZZAT A,WU M,LI X L,et al.Drug-target interaction prediction using ensemble learning and dimensionality reduction[J].Methods,2017,129:81-88.

[98] SHARMA A,RANI R.BE-DTI':ensemble framework for drug target interaction prediction using dimensionality reduction and active learning[J].Computer Methods and Programs in Biomedicine,2018,165:151-162.

[99] YANG J,HE S,ZHANG Z,et al.NegStacking:drug-target interaction prediction based on ensemble learning and logistic regression[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2021,18(6):2624-2634.

[100] PLIAKOS K,VENS C.Drug-target interaction prediction with tree-ensemble learning and output space reconstruction[J].BMC Bioinformatics,2020,21(1):1-11.

[101] XUAN P,CHEN B,ZHANG T.Prediction of drug-target interactions based on network representation learning and ensemble learning[J].IEEE/ACM Transactions on Computational Biology and Bioinformatics,2021,18(6):2671-2681.

[102] THAFAR M A,OLAYAN R S,ALBARADEI S,et al.DTi2Vec:drug-target interaction prediction using network embedding and ensemble learning[J].Journal of Cheminformatics,2021,13(1):1-18.

[103] BERMAN H M,WESTBROOK J,FENG Z,et al.The protein data bank[J].Nucleic Acids Research,2000,28(1):235-242.

[104] 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 Research,2018,46(1):1074-1082.

[105] KIM S,CHEN J,CHENG T,et al.PubChem 2019 update:improved access to chemical data[J].Nucleic Acids Research,2019,47(1):1102-1109.

[106] CHEN J,SWAMIDASS S J,DOU Y,et al.ChemDB:a public database of small molecules and related chemoinformatics resources[J].Bioinformatics,2005,21(22):4133-4139.

[107] CANESE K,WEIS S.PubMed:the bibliographic database[J].The NCBI Handbook,2013,2(1).

[108] XIA Z,WU L Y,ZHOU X,et al.Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces[C]//International Symposium on Optimization and Systems Biology,2010:1-16.

[109] MA T,XIAO C,ZHOU J,et al.Drug similarity integration through attentive multi-view graph auto-encoders[J].arXiv:1804.10850,2018.

[110] HUANG K,XIAO C,GLASS L M,et al.MolTrans:molecular interaction transformer for drug-target interaction prediction[J].Bioinformatics,2021,37(6):830-836.

[111] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[J].Advances in Neural Information Processing Systems,2017,30.

[112] HAO B,ZHANG J,YIN H,et al.Pre-training graph neural networks for cold-start users and items representation[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining,2021:265-273.

[113] LIN Z,TIAN C,HOU Y,et al.Improving graph collaborative filtering with neighborhood-enriched contrastive learning[C]//Proceedings of the ACM Web Conference,2022:2320-2329.