(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2022 年 1 月 27 日 (27.01.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/017405 A1

(51) 国际专利分类号:
G16B 20/00 (2019.01) *G16B 50/30* (2019.01)
G16B 40/00 (2019.01) *G16C 20/50* (2019.01)

(21) 国际申请号: PCT/CN2021/107509

(22) 国际申请日: 2021 年 7 月 21 日 (21.07.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010704024.0 2020年7月21日 (21.07.2020) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (**TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED**) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: **徐挺洋 (XU, Tingyang)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 **张吉应 (ZHANG, Jiying)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 **叶菲 (YE, Fei)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 **荣钰 (RONG, Yu)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 **黄俊洲 (HUANG, Junzhou)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 北 京 派 特 恩 知 识 产 权 代 理 有 限 公 司 (**CHINA PAT INTELLECTUAL PROPERTY OFFICE**); 中国北京市海淀区海淀南路21号中关村知识产权大厦B座2层, Beijing 100080 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** MEDICINE SCREENING METHOD AND APPARATUS AND ELECTRONIC DEVICE

(54) **发明名称**: 一种药物筛选方法、装置及电子设备

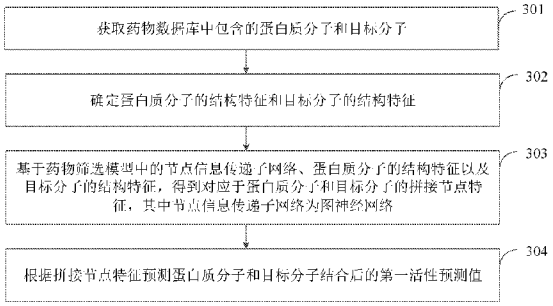


图 3A

301 Obtain protein molecules and target molecules comprised in a medicine database

302 Determine structural characteristics of the protein molecules and structural characteristics of the target molecules

303 On the basis of a node information transfer sub-network in a medicine screening model, the structural characteristics of the protein molecules and the structural characteristics of the target molecules, obtain jointed node characteristics corresponding to the protein molecules and the target molecules, the node information transfer sub-network being a graph neural network

304 According to the jointed node characteristics, predict a first activity predicted value after the protein molecules and the target molecules are bound

(57) **Abstract:** A medicine screening method and apparatus, an electronic device, and a computer readable storage medium. The method comprises: determining structural characteristics of protein molecules and structural characteristics of target molecules (302); on the basis of a node information transfer sub-network in a medicine screening model, the structural characteristics of the protein molecules and the structural characteristics of the target molecules, obtaining jointed node characteristics corresponding to the protein molecules and the target molecules, the node information transfer sub-network being a graph neural network (303); and according to the jointed node characteristics, predicting a first activity predicted value after the protein molecules and the target molecules are bound (304).

(57) **摘要**: 一种药物筛选方法、装置、电子设备及计算机可读存储介质，方法包括：确定蛋白质分子的结构特征和目标分子的结构特征(302)；基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络(303)；根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值(304)。

WO 2022/017405 A1

[见续页]

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) **指定国**(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

一种药物筛选方法、装置及电子设备

相关申请的交叉引用

本申请基于申请号为202010704024.0、申请日为2020年07月21日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

本申请涉及信息处理技术，尤其涉及药物筛选方法、装置、电子设备及计算机可读存储介质。

背景技术

药物研发是现代社会中受到广泛关注的高技术产业，药物研发中的关键环节是药物筛选，即是指对可能作为药物使用的物质（如蛋白质）进行活性或其他属性的评估。

然而，在传统的药物筛选过程中，通常是由研发人员手动进行相关实验，进而进行药物筛选，这导致药物筛选存在成本高、研发周期长及成功率低等问题。

针对于此，相关技术并未提供有效的解决方案。

发明内容

本申请实施例提供了一种药物筛选方法，由电子设备执行，所述方法包括：

获取药物数据库中包含的蛋白质分子和目标分子；

确定所述蛋白质分子的结构特征和所述目标分子的结构特征；

基于药物筛选模型中的节点信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，其中所述节点信息传递子网络为图神经网络；

根据所述拼接节点特征预测所述蛋白质分子和所述目标分子结合后的第一活性预测值。

本申请实施例还提供了一种药物筛选装置，包括：

信息传输模块，配置为获取药物数据库中包含的蛋白质分子和目标分子；

信息处理模块，配置为：

确定所述蛋白质分子的结构特征和所述目标分子的结构特征；

基于药物筛选模型中的节点信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，其中所述节点信息传递子网络为图神经网络；

根据所述拼接节点特征预测所述蛋白质分子和所述目标分子结合后的第一活性预测值。

本申请实施例还提供了一种电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于运行所述存储器存储的可执行指令时，实现本申请实施例提供的药物筛选方法。

本申请实施例还提供了一种计算机可读存储介质，存储有可执行指令，所述可执

行指令被处理器执行时实现本申请实施例提供的药物筛选方法。

附图说明

图 1 为本申请实施例提供的药物筛选方法的使用场景示意图；
图 2 为本申请实施例提供的电子设备的组成结构示意图；
图 3A 为本申请实施例提供的药物筛选方法的流程示意图；
图 3B 为本申请实施例提供的药物筛选方法的流程示意图；
图 4 为本申请实施例提供的确定蛋白质分子的结构特征的示意图；
图 5 为本申请实施例提供的确定目标分子的结构特征的流程示意图；
图 6 为本申请实施例提供的确定目标分子的结构特征的示意图；
图 7 为本申请实施例提供的药物筛选方法的流程示意图；
图 8 为本申请实施例提供的训练药物筛选模型的处理过程示意图。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，所描述的实施例不应视为对本申请的限制，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

在以下的描述中，涉及到“一些实施例”，其描述了所有可能实施例的子集，但是可以理解，“一些实施例”可以是所有可能实施例的相同子集或不同子集，并且可以在不冲突的情况下相互结合。

在以下的描述中，所涉及的术语“第一\第二”仅仅是区别类似的对象，不代表针对对象的特定排序，可以理解地，“第一\第二”在允许的情况下可以互换特定的顺序或先后次序，以使这里描述的本申请实施例能够以除了在这里图示或描述的以外的顺序实施。在以下的描述中，所涉及的术语“多个”是指至少两个。

对本申请实施例进行进一步详细说明之前，对本申请实施例中涉及的名词和术语进行说明，本申请实施例中涉及的名词和术语适用于如下的解释。

1）响应于：用于表示所执行的操作所依赖的条件或者状态，当满足所依赖的条件或状态时，所执行的一个或多个操作可以是实时的，也可以具有设定的延迟；在没有特别说明的情况下，所执行的多个操作不存在执行先后顺序的限制。

2）基于：用于表示所执行的操作所依赖的条件或者状态，当满足所依赖的条件或状态时，所执行的一个或多个操作可以是实时的，也可以具有设定的延迟；在没有特别说明的情况下，所执行的多个操作不存在执行先后顺序的限制。

3）分子：由组成的原子按照一定的键合顺序和空间排列而结合在一起的整体，这种键合顺序和空间排列关系称为分子结构。在本申请实施例中，大分子可以是指相对分子质量在 5000 以上的生物学物质，如蛋白质、核酸、多糖等；小分子可以是指相对分子质量在 1000 以下的生物学物质，如小肽、寡肽、寡糖、寡核苷酸、维生素等。

4）蛋白质分子：由氨基酸以“脱水缩合”的方式组成的多肽链经过盘曲折叠形成的具有一定空间结构的物质。

5）药物筛选：在本申请实施例中，是指将药物筛选的过程在计算机上模拟，对化合物可能的活性作出预测，进而对比较有可能成为药物的化合物进行有针对性的实体筛选。可以表现为应用分子对接技术，筛选需要获知药物作用靶标的分子结构，通过分子模拟手段计算化合物库中的小分子与靶标结合的能力，预测候选化合物的生理活性，建立合理的药效团模型、准确测定或预测靶标蛋白质的分子结构、精确和快速地

计算候选化合物与靶标相互作用的自由能变化是进行药物筛选的关键。

图 1 为本申请实施例提供的药物筛选方法的使用场景示意图，参见图 1，终端包括终端 10-1 和终端 10-2，其中终端 10-1 位于开发人员侧，用以控制药物筛选模型的训练与使用，终端 10-2 位于用户侧，用以请求进行药物筛选；终端通过网络 300 连接服务器 200，网络 300 可以是广域网或者局域网，又或者是二者的组合，使用无线或有线链路实现数据传输 。

终端 10-2 位于用户侧，用于发出药物筛选请求，请求对药物数据库中包含的蛋白质分子和目标分子进行筛选。

在一些实施例中，服务器 200 用于布设药物筛选装置以实现本申请所提供的药物筛选方法，服务器 200 可以布设经过训练的药物筛选模型，以实现在不同的环境中（例如对靶向性药物或者化学药物进行筛选的环境中）对药物进行筛选。在使用药物筛选模型之前，可以对药物筛选模型进行训练，示例过程包括：基于药物数据库中的药物信息参数，确定与药物筛选模型相匹配的训练样本集合，其中，训练样本集合包括至少一组训练样本；通过药物筛选模型提取与训练样本相匹配的特征集合；根据与训练样本相匹配的特征集合对药物筛选模型进行训练，以确定与药物筛选模型相适配的模型参数。当然，本申请所提供的药物筛选装置可以对同一目标分子在不同药物筛选环境中对应的药物筛选模型进行训练，最终在用户界面（UI，User Interface）上呈现出通过药物筛选模型所确定出的蛋白质分子和目标分子的结合产物的活性检测结果（如第一活性预测值和/或第二活性预测值），该活性检测结果还可以供其他应用程序调用。当然，与相应的药物数据库相匹配的药物筛选模型也可以迁移至不同的药物筛选进程（例如靶向性药物筛选进程、化学药物筛选进程或者高分子药物筛选进程）。

对药物筛选模型训练完成之后，服务器 200 就可以通过药物筛选模型进行药物筛选，示例过程如：获取药物数据库中包含的蛋白质分子和目标分子；确定蛋白质分子的结构特征和目标分子的结构特征；基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络；根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值。服务器 200 可以得到的第一活性预测值对药物数据库中的分子进行筛选，即进行药物筛选，在完成对药物的筛选时，可以将筛选结果发送至终端，如发送至终端 10-1 或终端 10-2，其中，还可以结合第二活性预测值进行药物筛选，第二活性预测值的确定过程将在后文进行阐述。

在一些实施例中，终端（如终端 10-1 或终端 10-2）也可以布设药物筛选装置以实现本申请所提供的药物筛选方法，即在终端本地实现药物筛选模型的训练，并根据训练后的药物筛选模型进行药物筛选。

在一些实施例中，服务器 200 可以进行药物筛选模型的训练，并将训练后的药物筛选模型发送至终端，以使终端根据训练后的药物筛选模型在本地实现药物筛选。

下面对本申请实施例的电子设备的结构做详细说明，电子设备可以各种形式来实施，如带有药物筛选装置处理功能的专用终端，也可以为设置有药物筛选装置处理功能的服务器，例如图 1 中的服务器 200。图 2 为本申请实施例提供的电子设备的组成结构示意图，可以理解，图 2 仅仅示出了电子设备的示例性结构而非全部结构，根据需要可以实施图 2 示出的部分结构或全部结构。

本申请实施例提供的电子设备包括：至少一个处理器 201、存储器 202、用户接口 203 和至少一个网络接口 204。电子设备中的各个组件通过总线系统 205 耦合在一起。可以理解，总线系统 205 用于实现这些组件之间的连接通信。总线系统 205 除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，

在图 2 中将各种总线都标为总线系统 205。

其中，用户接口 203 可以包括显示器、键盘、鼠标、轨迹球、点击轮、按键、按钮、触感板或者触摸屏等。

可以理解，存储器 202 可以是易失性存储器或非易失性存储器，也可包括易失性和非易失性存储器两者。本申请实施例中的存储器 202 能够存储数据以支持终端（如图 1 示出的终端 10-1 或终端 10-2）的操作。这些数据的示例包括：用于在终端（如图 1 示出的终端 10-1 或终端 10-2）上操作的任何计算机程序，如操作系统和应用程序。其中，操作系统包含各种系统程序，例如框架层、核心库层、驱动层等，用于实现各种基础业务以及处理基于硬件的任务。应用程序可以包含各种应用程序。

在一些实施例中，本申请实施例提供的药物筛选装置可以采用软硬件结合的方式实现，作为示例，本申请实施例提供的药物筛选装置可以是采用硬件译码处理器形式的处理器，其被编程以执行本申请实施例提供的药物筛选方法。例如，硬件译码处理器形式的处理器可以采用一个或多个应用专用集成电路（ASIC，Application Specific Integrated Circuit）、DSP、可编程逻辑器件（PLD，Programmable Logic Device）、复杂可编程逻辑器件（CPLD，Complex Programmable Logic Device）、现场可编程门阵列（FPGA，Field-Programmable Gate Array）或其他电子元件。

作为本申请实施例提供的药物筛选装置采用软硬件结合实施的示例，本申请实施例所提供的药物筛选装置可以直接体现为由处理器 201 执行的软件模块组合，软件模块可以位于存储介质（计算机可读存储介质）中，存储介质位于存储器 202，处理器 201 读取存储器 202 中软件模块包括的可执行指令，结合必要的硬件（例如，包括处理器 201 以及连接到总线 205 的其他组件）完成本申请实施例提供的药物筛选方法。

作为示例，处理器 201 可以是一种集成电路芯片，具有信号的处理能力，例如通用处理器、数字信号处理器（DSP，Digital Signal Processor），或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等，其中，通用处理器可以是微处理器或者任何常规的处理器等。

作为本申请实施例提供的药物筛选装置采用硬件实施的示例，本申请实施例所提供的装置可以直接采用硬件译码处理器形式的处理器 201 来执行完成，例如，被一个或多个 ASIC、DSP、PLD、CPLD、FPGA 或其他电子元件执行实现本申请实施例提供的药物筛选方法。

本申请实施例中的存储器 202 用于存储各种类型的数据以支持药物筛选装置的操作。这些数据的示例包括：用于在药物筛选装置上操作的任何可执行指令，实现本申请实施例的药物筛选方法的程序可以包含在可执行指令中。

在一些实施例中，本申请实施例提供的药物筛选装置可以采用软件方式实现，图 2 示出了存储在存储器 202 中的药物筛选装置，其可以是程序和插件等形式的软件，并包括一系列的模块，作为存储器 202 中存储的程序的示例，可以包括以下的软件模块：信息传输模块 2081 和信息处理模块 2082。当药物筛选装置中的软件模块被处理器 201 读取到随机存取存储器（Random Access Memory，RAM）中并执行时，将实现本申请实施例提供的药物筛选方法。

在一些实施例中，服务器 200 可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络（CDN，Content Delivery Network）、以及大数据和人工智能平台等基础云计算服务的云服务器，服务器 200 可为实体设备，也可为虚拟化设备。终端（如图 1 示出的终端 10-1 或终端 10-2）可以是智能手机、平板电脑、笔记本电脑、台式计算机等，但并

不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请实施例中不做限制。

在实际应用中，本申请实施例提供的药物筛选模型可以应用于结构生物学以及医学领域，通过药物筛选模型实现药物发现、分子优化、分子合成等。

将结合图 2 示出的药物筛选装置说明本申请实施例提供的药物筛选方法，参见图 3A，图 3A 为本申请实施例提供的药物筛选方法的一个流程示意图，可以理解地，图 3A 所示的步骤可以由运行药物筛选装置的各种电子设备执行，例如可以是带有药物筛选装置的专用终端、药物数据库服务器或者药品提供商的服务器集群。

在图 3A 中，为了克服传统药物筛选方式所造成的药物筛选不准确以及效率低的缺陷，本申请所提供的技术方案使用了人工智能技术，从而大幅度减少相关实验所需的时间和费用，同时增加药物筛选的准确性，提升药物筛选的效率。下面针对图 3A 示出的步骤进行具体说明。

步骤 **301**：获取药物数据库中包含的蛋白质分子和目标分子。

例如，电子设备在接收到药物筛选请求（来自用户或其他电子设备）时，获取药物数据库中包含的蛋白质分子和目标分子。其中，目标分子可以是药物小分子，蛋白质分子可以是能被药物分子（如药物小分子）作用的标靶大分子。

若想要对药物数据库中的化合物可能的活性进行预测，进而对有可能成为临床药物的化合物进行有针对性的实体筛选，可以通过应用分子对接技术来实现。例如，将能被药物分子作用的标靶大分子与药物小分子进行拼接，形成新的化合物，并预测形成的化合物的生理活性。

步骤 **302**：确定蛋白质分子的结构特征和目标分子的结构特征。

在本申请的一些实施例中，确定蛋白质分子的结构特征和目标分子的结构特征，可以通过以下方式实现：

确定蛋白质分子中的不同氨基酸链的空间位置；基于不同氨基酸链的空间位置，确定每对氨基酸之间的距离，并对每对氨基酸之间的距离进行标准化处理，得到标准氨基酸距离；基于标准氨基酸距离及氨基酸距离阈值，确定蛋白质分子对应的氨基酸矩阵图；基于蛋白质分子对应的氨基酸矩阵图，确定蛋白质分子的结构特征；确定目标分子所对应的原子和化学键，并基于目标分子所对应的原子和化学键，确定目标分子的结构特征。

例如，参考图 4，图 4 为本申请实施例中蛋白质分子的结构示意图，其中，在药物筛选的过程中，由于分子的结构无法直接输入一个神经网络进行训练和学习，所以需要投射到一个矢量化空间，即进行特征化处理。而在分子表示方法上，由于分子是不同的原子通过化学键建立连接，因此，可以看成一个由节点和边构成的图。例如，参考图 4，蛋白质具有空间结构，其是通过氨基酸链在空间中折叠形成的，而基于其空间结构就可以计算出每对氨基酸之间的距离，其中标准化氨基酸之间的空间距离参考公式 1 如下：

$$\overline{S}_{ij}=\frac{1}{1+d_{ij}/d'} \quad \text{公式 1}$$

这里 d' 为缩放尺度，用于进行标准化处理，例如 d' 可以取 $3.8\,\mathring{A}$；$\overline{S}_{ij}$ 表示第 i 个氨基酸到第 j 个氨基酸之间的距离。得到了标准氨基酸距离（如 d_{ij}）之后可以结合固定阈值（threshold）d_0（即氨基酸距离阈值），计算出蛋白质图的邻接矩阵（即氨基酸矩阵图），其中，蛋白质图的邻接矩阵计算参考公式 2：

$$A_{ij}=\begin{cases}1, d_{ij}<d_0\\ 0, d_{ij}>d_0\end{cases} \quad \text{公式 2}$$

再以氨基酸为图顶点就可以得到蛋白质图$G_{protein}$。如图4所示，蛋白质分子包括氨基酸A、C、D、E及F，在计算出的标准氨基酸距离中，d_{AC}、d_{CD}、d_{DE}以及d_{DF}均小于氨基酸距离阈值d_0，因此，基于氨基酸A与氨基酸C之间的连接、氨基酸C与氨基酸D之间的连接、氨基酸D与氨基酸E之间的连接、以及氨基酸D与氨基酸F之间的连接建立邻接矩阵，再以涉及到的氨基酸为图顶点就可以得到蛋白质图，该蛋白质图即体现了蛋白质分子的结构特征。

对于目标分子来说，可以根据目标分子的有机结构来确定目标分子的结构特征。作为示例，参考图5，图5为本申请实施例提供的确定目标分子的结构特征的一个流程示意图，图5所示的步骤可以由运行药物筛选装置的各种电子设备执行，例如可以是如带有药物筛选装置的专用终端、药物数据库服务器或者药品提供商的服务器集群。将结合图5示出的步骤进行说明。

步骤501：确定目标分子的有机结构。

步骤502：基于目标分子的有机结构，确定目标分子所对应的原子与化学键。

步骤503：将目标分子所对应的原子作为目标分子的结构特征中的节点。

步骤504：将目标分子所对应的化学键作为目标分子的结构特征中的边线。

步骤505：通过目标分子的结构特征中的节点、以及目标分子的结构特征中的边线，确定目标分子的结构特征。

例如，以目标分子对应的原子为节点，以目标分子对应的化学键为边，确定出目标分子对应的分子图（又称小分子图），该分子图即体现了目标分子的结构特征。为了便于理解，提供了如图6所示的目标分子及确定出的分子图的示意图。

当确定出目标分子的结构特征之后，继续执行后续步骤，即通过药物筛选模型对蛋白质分子与目标分子进行筛选。

继续参考图3A的**步骤303**：基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络。

例如，基于蛋白质分子的结构特征和目标分子的结构特征，确定药物筛选模型中的节点信息传递子网络的输出，该输出即为对应于蛋白质分子和目标分子的拼接节点特征。其中，节点信息传递子网络为图神经网络（Graph Neural Network，GNN），或者也可以是图神经网络的一部分。GNN是一种直接作用于图结构上的神经网络，主要针对非欧几里得空间结构(图结构)的数据进行处理。图神经网络可以由两个模块组成：传播模块(Propagation Module)和输出模块(Output Module)，传播模块用于图中节点之间传递信息并更新状态，输出模块用于基于图的节点和边的向量表示，根据不同的任务定义目标函数。因此，通过确定与目标氨基酸链对应的节点相连接的所有节点，可以将结构种类多样的蛋白质分子中不同氨基酸链的信息嵌入至图神经网络中所不断产生的新节点，实现嵌入表示。

在本申请的一些实施例中，基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，可以通过以下方式实现：

基于蛋白质分子的结构特征，确定目标节点的目标节点特征，目标节点对应于蛋白质分子中的氨基酸；基于蛋白质分子的结构特征，确定一端为目标节点的边线的相连边线特征；基于蛋白质分子的结构特征，确定与目标节点相连的相连节点的相连节点特征；节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，得到对应于蛋白质分子和目标分子的拼接节点特征。

例如，可以基于蛋白质分子的结构特征，确定目标节点的目标节点特征，其中，节点对应于蛋白质分子中的氨基酸，即目标节点可以对应于目标氨基酸。基于蛋白质分子的结构特征，确定一端为目标节点的边线的相连边线特征，其中，可以确定一端为目标节点的部分边线的相连边线特征，也可以确定一端为目标节点的所有边线的相连边线特征，后者能够提升得到的相连边线特征的准确性和全面性。基于蛋白质分子的结构特征，还确定与目标节点相连的相连节点的相连节点特征，同理，这里可以确定与目标节点相连的部分相连节点的相连节点特征，或者与目标节点相连的所有相连节点的相连节点特征。通过节点信息传递子网络对目标节点特征、相连边线特征和相连节点特征进行处理，即可得到对应于蛋白质分子和目标分子的拼接节点特征。

值得说明的是，在确定节点信息传递子网络输出的过程中，可以通过节点信息传递子网络产生对应于目标氨基酸的新节点（例如对应目标氨基酸的目标节点嵌入表示），实现对蛋白质分子中氨基酸链的嵌入。

在一些实施例中，节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，得到对应于蛋白质分子和目标分子的拼接节点特征，可以通过以下方式实现：

节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，产生对应于目标节点的目标节点嵌入表示；基于目标节点嵌入表示，节点信息传递子网络得到对应于蛋白质分子的蛋白质节点嵌入表示向量；基于目标分子的结构特征，节点信息传递子网络得到对应于目标分子的目标分子节点嵌入表示向量；拼接蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接节点特征。

在本申请实施例中，如图 7 所示，节点信息传递子网络（图神经网络或图神经网络的一部分）可以分别针对蛋白质分子和目标分子进行处理。对于蛋白质分子，节点信息传递子网络对目标节点特征、相连边线特征和相连节点特征进行处理，产生对应于目标节点的目标节点嵌入表示，然后，可以将所有节点分别对应的节点嵌入表示（即是将各个节点分别作为目标节点，得到相应的目标节点嵌入表示）进行结合（包括但不限于拼接），得到对应于蛋白质分子的蛋白质节点嵌入表示向量，如此，通过节点信息传递子网络实现对蛋白质分子的嵌入表示。

对于目标分子，同理，节点信息传递子网络可以对目标分子的结构特征进行处理，得到对应于目标分子的目标分子节点嵌入表示向量，如此，实现对目标分子的嵌入表示。

最后，对蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量进行拼接处理，得到拼接节点特征。

在一些实施例中，节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，产生对应于目标节点的目标节点嵌入表示，可以通过以下方式实现：

基于目标节点特征，得到目标节点初始状态特征；基于相连节点的相连节点特征，得到相连节点状态特征；节点信息传递子网络的第一信息汇集函数，对相连节点状态特征和相连边线特征进行结合，得到目标节点信息特征；基于目标节点初始状态特征和目标节点信息特征，节点信息传递子网络的更新函数对目标节点状态特征进行更新；根据更新后的相连节点状态特征，节点信息传递子网络产生目标节点嵌入表示。

例如，可以对目标节点特征进行进一步的特征提取处理，得到目标节点初始状态特征；同理，对相连节点特征进行进一步的特征提取处理，得到相连节点状态特征。通过节点信息传递子网络的第一信息汇集函数对相连节点状态特征和相连边线特征进行结合，得到目标节点信息特征，其中，第一信息汇集函数如拼接函数，即这里的结

合可以是指拼接处理，但并不限于此。然后，通过节点信息传递子网络的更新函数对目标节点初始状态特征和目标节点信息特征进行处理，处理结果即为更新后的目标节点状态特征，其中，更新函数可以用于进行线性处理（如线性变换运算）及偏置处理。根据对目标节点状态特征进行更新的方式，同理可以对每个节点的节点状态特征进行更新。最后，通过节点信息传递子网络对更新后的相连节点状态特征进行处理，得到目标节点的目标节点嵌入表示。

在一些实施例中，根据更新后的相连节点状态特征，节点信息传递子网络产生目标节点嵌入表示，可以通过以下方式实现：

节点信息传递子网络的第二信息汇集函数，对更新后的相连节点状态特征和相连节点特征进行结合，得到目标节点嵌入特征；节点信息传递子网络的激活函数，对目标节点嵌入特征进行处理，得到目标节点嵌入表示。

例如，可以通过节点信息传递子网络的第二信息汇集函数，对更新后的相连节点状态特征和相连节点特征进行结合，得到目标节点嵌入特征，其中，第二信息汇集函数如拼接函数，即这里的结合可以是指拼接处理，但并不限于此。然后，通过节点信息传递子网络的激活函数对目标节点嵌入特征进行激活处理，得到目标节点嵌入表示。

为了便于理解，以节点信息传递子网络是信息传递网络（MPNN，Message Passing Neural Networks）的情况进行举例说明。MPNN 的前向传播包括两个阶段，第一个阶段称为 message passing（信息传递）阶段，第二个阶段称为 readout（读取）阶段。这里给定图结构 G=(V,E)，其中 V 可以代表节点 v 的集合，E 可以代表边线 e 的集合。信息传递阶段会执行多次信息传递过程。例如，对于对应蛋白质图（蛋白质分子的结构特征）中一个特定氨基酸的节点 v，可以参考公式 3 和公式 4 进行第 t 次信息传递，每次信息传递的输入至少部分来自于上一次信息传递的输出：

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{wv}) \quad \text{公式 3}$$

$$h_v^{t+1} = u_t(h_v^t, m_v^{t+1}) \quad \text{公式 4}$$

对于对应目标氨基酸的目标节点 v 来说，这里的节点信息传递子网络（即 MPNN 模型 $F_{\text{node}}^D(\text{G})$）可以通过聚集与目标节点 v 连接的相连节点 w 经过了 t 次信息传递后的相连节点状态特征h_w^t、以及每个节点 w 到目标节点 v 之间边线的相连边线特征e_{wv}，通过信息传递来产生一个新的节点 v。例如参考公式 5、公式 6 以及公式 7，给出了一种实施示例：

$$h_v^0 = \sigma(W_{in} x_v) \text{公式 5}$$

$$m_v^{d+1} = \sum_{k \in N(v)} cat(h_k^d, e_{vk}) \quad \text{公式 6}$$

$$h_v^{d+1} = \sigma(W_\alpha m_v^{d+1} + \mathrm{h}_v^0) \text{公式 7}$$

本领域技术人员会理解，这里的公式 5 描述的是对目标节点 v 的初始节点信息x_v（即目标节点特征）进行特征提取得到目标节点初始状态特征h_v^0。公式 6 和 7 描述的是每次信息传递的过程。这里 N(v)是与节点 v 相连的相连节点 k 的集合，$\sigma(\cdot)$是神经网络的激活函数。显然，这里第一信息汇集函数采用的是拼接函数 $cat(\cdot)$，节点 v 和相连节点 k 之间的相连边线的相连边线特征 e_{vk} 作为 μ_{attached}，以与邻接点 k 经过 d 次信息传递后的相连节点状态特征h_k^d进行拼接，得到 d 次信息传递后的目标节点信息特征m_v^{d+1}。节点更新函数（对应上文的更新函数）采用的是线性变换运算加偏置操作。显然，信息传递后，会更新得到目标节点 v 新的目标节点状态特征h_v^d。在一些实施例中，W_{in}和W_α这两个权重在节点更新过程中共享。

经过 D 次信息传递之后，可以采用一个额外的信息传递步骤计算出目标节点嵌入表示h_v^0（可对应新的节点）作为节点信息传递子网络的输出。在一些实施例中，这个

额外的信息传递步骤可参考公式 8 和公式 9 的形式：

$$m_v^0 = \sum_{k \in N(v)} cat(h_k^d, x_k) \quad \text{公式 8}$$

$$h_v^0 = \sigma(W_0 m_v^0) \quad \text{公式 9}$$

本领域技术人员会理解，在这个实施例中，公式 8 描述的是对相连节点 k 经过 d 次信息传递后的相连节点状态特征h_k^d、以及相连节点 k 的相连节点特征x_k进行拼接处理（信息汇集），得到目标节点嵌入特征m_v^0，最后公式 9 根据m_v^0、输出参数W_0和激活函数$\sigma(\cdot)$，得到节点 v 的目标节点嵌入表示h_v^0。这个目标节点嵌入表示h_v^0对应于目标氨基酸。显然，经过信息传递得到的目标节点嵌入表示h_v^0聚集了节点 v 连接的所有节点 k 的相连节点特征，也聚集了节点 k 与节点 v 之间边线 d 的相连边线特征。在一些实施例中，可以将所有 n 个节点 v 各自对应的目标节点嵌入表示h_v^0一起作为节点信息传递子网络的输出，也即是将下列公式中的蛋白质节点嵌入表示向量H_a作为$F_{\text{node}}^D(G)$的最后输出：

$$\mathrm{H_a} = \left[\mathrm{h_1^0}, \ldots\ldots \mathrm{h_n^0}\right] \quad \text{公式 10}$$

在一些实施例中，拼接蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接节点特征，可以通过以下方式实现：

确定与药物筛选模型相匹配的自注意力读出函数；通过自注意力读出函数、蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量，确定蛋白质分子的结构特征中的第一节点特征向量和目标分子的结构特征中的第二节点特征向量；对第一节点特征向量和第二节点特征向量进行拼接，得到对应于蛋白质分子和目标分子的拼接节点特征。

例如，给定信息传递网络（如节点信息传递子网络）的输出$H \in R^{n*a}$，自注意力权重矩阵 S（自注意力权重矩阵与自注意力读出函数相匹配）可以通过公式 11 表示为：

$$S = solfmax(W_2 \, tanh(W_1 H^T)) \quad \text{公式 11}$$

其中，$\mathrm{W_1} \in \mathrm{R^{h_{attn} \times a}}$，$\mathrm{W_2} \in \mathrm{R^{r \times h_{attn}}}$都是可学习参数。前序公式中，$W_1$是一个线性变换，将 a 维空间的 n 个节点嵌入变换到h_{attn}维空间中，然后经过双曲正切函数$\tanh(\cdot)$进行非线性映射，接着W_2将h_{attn}维空间中的嵌入再线性变换到 r 维空间中，得到的 r 个不同角度的节点重要性分布，值越大代表这个节点的越重要，最后再经过 solfmax(·)函数让每个视角的重要性值的总和为 1，使其符合一个权重分布的特性。

得到了 n 个节点对应的自注意力权重矩阵$\mathrm{S} \in \mathrm{R^{r \times n}}$之后，可以根据自注意力权重矩阵 S 和信息传递网络的输出 H，确定固定大小的包含了节点重要性的图的向量表示：

$$\xi = flatten(SH), \ \xi \in \mathrm{R^{r \times a}} \quad \text{公式 12}$$

其中，flatten(·)表示将矩阵SH展开成一维向量。

进一步地，还可以将蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量拼接起来，即可结合小分子和蛋白质的信息，并基于拼接起来的向量表示预测蛋白质分子和目标分子结合之后的活性。在一些实施例中，可参考公式 13 的形式：

$$pred_a = FCN_a(cat(\xi_a^P, \xi_a^M)) \quad \text{公式 13}$$

其中，$cat(\cdot)$是拼接函数，FCN是全连接神经网络，ξ_a^P是蛋白质图（蛋白质分子的结构特征）经过节点信息传递子网络得到的H_a^P与自注意力读出函数结合后得到的节点特征向量表示（即第一节点特征向量）。类似地，ξ_a^M是小分子图（目标分子的结构特征）经过节点信息传递子网络得到的H_a^M与自注意力读出函数结合后得到的节点特征向量表示（即第二节点特征向量），$pred_a$表示拼接节点特征。

步骤 304：根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值。

根据得到的拼接节点特征，即可预测蛋白质分子和目标分子的结合产物的活性预测值，为了便于区分，将这里得到的活性预测值命名为第一活性预测值。相较于相关技术提供的方案中需要大量实验才能得到第一活性预测值，本申请实施例结合图神经

网络能够准确、快速地确定第一活性预测值，可以大大节省人力成本和时间成本。

在一些实施例中，步骤 304 之后，还包括：基于第一活性预测值，对药物数据库中的分子进行筛选。

一个第一活性预测值对应一个蛋白质分子和一个目标分子，因此，可以根据第一活性预测值对药物数据库中的分子进行筛选。例如，在蛋白质分子固定的情况下，可以确定该蛋白质分子分别与多个目标分子进行结合所得到的第一活性预测值，并根据第一活性预测值对多个目标分子进行筛选，如将最大的若干个第一活性预测值（如最大的一个第一活性预测值）对应的目标分子作为筛选出的目标分子，即筛选结果；又例如，在目标分子固定的情况下，可以确定该目标分子分别与多个蛋白质分子进行结合所得到的第一活性预测值，并根据第一活性预测值对多个蛋白质分子进行筛选，如将最大的若干个第一活性预测值（如最大的一个第一活性预测值）对应的蛋白质分子作为筛选出的蛋白质分子，即筛选结果；又例如，也可以在存在多个蛋白质分子以及多个目标分子的情况下，对蛋白质分子和目标分子进行共同筛选。

参见图 3B，图 3B 为本申请实施例提供的药物筛选方法的一个流程示意图，在图 3A 示出的步骤 302 之后，还可以在**步骤 305** 中，基于边线信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接边线特征。

在本申请实施例中，药物筛选模型还可包括边线信息传递子网络，该边线信息传递子网络同样可以是图神经网络或图神经网络的一部分。

在得到蛋白质分子的结构特征以及目标分子的结构特征后，可以通过边线信息传递子网络对蛋白质分子的结构特征以及目标分子的结构特征进行处理，得到对应于蛋白质分子和目标分子的拼接边线特征。

在本申请的一些实施例中，基于边线信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接边线特征，可以通过以下方式实现：

基于蛋白质分子的结构特征，确定目标边线的目标边线特征，目标边线特征对应于蛋白质分子中相连的两个氨基酸；基于蛋白质分子的结构特征，确定临边的临边边线特征，临边的第一端节点对应于相连的两个氨基酸之一，临边的第二端节点与第一端节点相连；确定与第二端节点对应的相邻节点特征；边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，得到对应于蛋白质分子和目标分子的拼接边线特征。

例如，基于蛋白质分子的结构特征，确定目标边线的目标边线特征，其中，边线对应满足特定条件的两个氨基酸之间的关系，例如两个氨基酸在蛋白质图中相连；目标边线可以是指任意一个边线。基于蛋白质分子的结构特征，确定目标边线的临边的临边边线特征，其中，该临边的第一端节点对应于相连的两个氨基酸之一，该临边的第二端节点与第一端节点相连。这里，可以确定目标边线的部分临边的临边边线特征，也可以确定目标边线的所有临边的临边边线特征。

对于临边来说，还确定与第二端节点对应的相邻节点特征。最后，通过边线信息传递子网络对目标边线特征、临边边线特征和相邻节点特征进行处理，得到对应于蛋白质分子和目标分子的拼接边线特征。

在一些实施例中，边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，得到对应于蛋白质分子和目标分子的拼接边线特征，可以通过以下方式实现：

边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，产生对应于第一端节点的边线嵌入表示；基于边线嵌入表示，边线信息传递子网络得到对应于蛋白质分子的蛋白质边线嵌入表示向量；基于目标分子的结构特征，边线信息传递子网络得到对应于目标分子的目标分子边线嵌入表示向量；拼接蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接边线特征。

例如，通过边线信息传递子网络对目标边线特征、临边边线特征和相邻节点特征进行处理，得到对应于第一端节点的边线嵌入表示。基于边线嵌入表示确定对应于蛋白质分子的蛋白质边线嵌入表示向量，例如可以将所有的边线嵌入表示进行结合（如拼接），得到蛋白质边线嵌入表示向量。同时，通过边线信息传递子网络对目标分子的结构特征进行处理，得到对应于目标分子的目标分子边线嵌入表示向量。

最后，对蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量进行拼接处理，得到对应于蛋白质分子和目标分子的拼接边线特征。

在一些实施例中，边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，产生对应于第一端节点的边线嵌入表示，可以通过以下方式实现：

基于目标边线特征，得到目标边线初始状态特征；基于临边边线特征，得到临边边线状态特征；边线信息传递子网络的第一信息传递函数对临边边线状态特征和相邻节点特征进行结合，得到目标边线信息特征；基于目标边线信息特征和目标边线初始状态特征，边线信息传递子网络的更新函数对目标边线状态特征进行更新；根据更新后的临边边线状态特征，边线信息传递子网络产生边线嵌入表示。

例如，对目标边线特征进行进一步的特征提取处理，得到目标边线初始状态特征；对临边边线特征进行进一步的特征提取处理，得到临边边线状态特征。通过边线信息传递子网络的第一信息传递函数对临边边线状态特征和相邻节点特征进行结合，得到目标边线信息特征。然后，通过边线信息传递子网络的更新函数对目标边线信息特征和目标边线初始状态特征进行处理，处理结果即为更新后的目标边线状态特征，其中，更新函数可以用于进行线性处理（如线性变换运算）及偏置处理。根据对目标边线状态特征进行更新的方式，同理可以对每个边线的边线状态特征（如临边边线状态特征）进行更新。最后，通过边线信息传递子网络对更新后的临边边线状态特征进行处理，得到边线嵌入表示。

在一些实施例中，根据更新后的临边边线状态特征，边线信息传递子网络产生边线嵌入表示，可以通过以下方式来实现：

边线信息传递子网络的第二信息传递函数，对更新后的临边边线状态特征和相邻节点特征进行结合，得到对应于第一端节点的边线嵌入特征；边线信息传递子网络的激活函数，对边线嵌入特征进行处理，得到边线嵌入表示。

例如，可以通过边线信息传递子网络的第二信息传递函数，对更新后的临边边线状态特征和相邻节点特征进行结合，得到对应于第一端节点的边线嵌入特征，其中，第二信息传递函数如拼接函数，即这里的结合可以是指拼接处理，但并不限于此。然后，通过边线信息传递子网络的激活函数对边线嵌入特征进行激活处理，得到边线嵌入表示。

为了便于理解，以边线信息传递子网络是 MPNN 的情况进行举例说明。对于给定的目标边线特征e_{vw}，其对应的目标边线信息特征$\mathrm{m}_{\mathrm{vw}}^{\mathrm{d}}$和目标边线状态特征$h_{vw}^{d}$可以通过公式 14、公式 15 以及公式 16 进行计算：

$$h_{vw}^{0}=\sigma(W_{in}e_{vw}) \quad \text{公式 14}$$

$$m_{vw}^{\mathrm{d+1}}=\sum_{k\in N(v)\backslash w}^{\Sigma}cat(h_{kv}^{d},x_{k}) \quad \text{公式 15}$$

$$h_{vw}^{d+1} = \sigma(W_b m_{vw}^{d+1} + h_{vw}^{0}) \quad \text{公式 16}$$

在公式 14-16 中，本领域技术人员可以理解，目标边线特征e_{vw}对应的临边集合 kv 为一端是节点 v、除了边线 vw 之外的所有边线的集合，也即是$k \in N(v) \setminus w$。公式 14 基于目标边线特征得到目标边线初始状态特征。这里的信息传递函数（见公式 15，即第一信息传递函数）与上述节点信息传递子网络中的信息传递函数（见公式 3）相似，将临边集合中每个边线 vk 对应的经过 d 轮信息传递后的临边边线信息特征、以及每个边线 vk 对应的关联特征$\mu_{attached}$（即临边集合中每个边线 vk 对应的除节点 v 外的端点 k 的节点特征x_k，即相邻节点特征）进行拼接。这里的节点更新函数（见公式 16）也与上述节点信息传递子网络中的节点更新函数（见公式 7）相似，采用的是线性变换运算加偏置的操作，基于目标边线信息特征和目标边线初始状态特征对目标边线状态特征进行更新。

经过循环的 D 步的信息传递之后，同样可以通过采用额外一轮的节点信息聚集，将边的信息转移到两端节点的信息中，产生最后的目标边线嵌入表示h_v^0。在一些实施例中，这额外的一轮信息聚集可以通过公式 17 和公式 18 的形式进行实现:

$$m_v^0 = \sum_{k \in N(v)} cat(h_{kv}^d, x_k) \quad \text{公式 17}$$

$$h_v^0 = \sigma(W_0 m_v^0) \quad \text{公式 18}$$

本领域技术人员会理解，在这个实施例中，公式 17 描述的是将临边 kv 经过 D 次信息传递后的临边边线状态特征h_{kv}^d、以及这些边线 kv 另一端的相邻节点 k 的相邻节点特征x_k进行拼接处理，得到边线嵌入特征m_v^0，最后公式 18 根据m_v^0、输出参数W_0和激活函数$\sigma(\cdot)$，得到节点 v 的边线嵌入表示h_v^0。

在一些实施例中，可以将所有 n 个节点 v 对应的边线嵌入表示h_v^0一起作为边线信息传递子网络的输出，也即是将下列公式中的边线嵌入表示向量H_b作为$F_{\text{edge}}^D(G)$的最后输出:

这里最后的输出结果$F_{\text{edge}}^D(G)$可以表示为:

$$\mathrm{H_b} = \left[\mathrm{h_1^0}, \ldots\ldots \mathrm{h_n^0}\right]。$$

在一些实施例中，边线信息传递子网络也可以对目标分子的结构特征进行处理，输出对应于目标分子的目标分子边线嵌入表示向量。

在一些实施例中，拼接蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接边线特征，可以通过以下方式来实现:

确定与药物筛选模型相匹配的自注意力读出函数；通过自注意力读出函数、蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量，确定蛋白质分子的结构特征中的第一边线特征向量、以及目标分子的结构特征中的第二边线特征向量；对第一边线特征向量和第二边线特征向量进行拼接，得到对应于蛋白质分子和目标分子的拼接边线特征。

例如，可以参照公式 11 来得到自注意力权重矩阵 S。在一些实施例中，为了使节点信息传递子网络和边线信息传递子网络提取的特征信息在训练过程中可以有信息交互，可以使公式 11 中的注意力参数W_1和W_2在这两个网络上进行共享，也即共用一套W_1和W_2。

得到了 n 个节点对应的自注意力权重矩阵 $S \in R^{r \times n}$之后，可以根据自注意力权重矩阵 S 和来自信息传递网络的输入 H 得到固定大小的包含了节点重要性的图的向量表示ξ。

进一步地，还可以将蛋白质表示和目标分子表示拼接起来，即可结合小分子和蛋白质的信息，并基于拼接起来的向量表示预测蛋白质分子和目标分子结合之后的活性。在一些实施例中，可参考公式 19 的形式:

$$pred_b = FCN_b(cat(\xi_b^P, \xi_b^M)) \quad \text{公式 19}$$

其中，$cat(\cdot)$是拼接函数，FCN是全连接神经网络，ξ_b^P是蛋白质图经过边线信息传递子网络得到的H_b^P与自注意力读出函数结合后得到的边线特征向量表示（即第一边线特征向量）。类似地，ξ_b^M是小分子图经过边线信息传递子网络得到的H_b^M与自注意力读出函数结合后得到的边线特征向量表示（即第二边线特征向量），$pred_b$表示拼接边线特征。

在**步骤 306**中，根据拼接边线特征预测蛋白质分子和目标分子结合后的第二活性预测值。

这里，根据拼接边线特征预测蛋白质分子和目标分子结合后的活性预测值，为了便于区分，将这里的活性预测值命名为第二活性预测值。

在一些实施例中，步骤 306 之后，还包括：基于第一活性预测值和第二活性预测值，对药物数据库中的分子进行筛选。

本领域技术人员应理解，最终进行药物筛选的活性预测值可以是$pred_a$和$pred_b$中的至少一个，也可以是两者的均值，或者基于其他方法对两者进行计算得到的最终活性预测值，本申请不做限定。此外，以上涉及的全连接FCN中的权重参数可以根据训练集进行训练得到。

本领域技术人员应理解，在一些实施例中，药物筛选模型可以仅基于节点信息传递子网络预测目标分子与蛋白质分子结合后的活性，例如通过公式 13 进行预测。类似的，在一些实施例中，药物筛选模型也可以仅基于边线信息传递子网络预测目标分子与蛋白质分子结合后的活性，例如通过公式 19 进行预测。

本领域技术人员应理解，本申请实施例中需要进行筛选的可以是蛋白质分子，也可以是目标分子，也可以对蛋白质分子和目标分子进行共同筛选，对此不做限定。

参考图 8，图 8 为本申请实施例提供的训练药物筛选模型的过程示意图。可以理解地，图 8 所示的步骤可以由运行药物筛选装置的各种电子设备执行，例如可以是如带有药物筛选装置的专用终端、药物数据库服务器或者药品提供商的服务器集群。下面将结合图 8 示出的步骤进行说明。

步骤 801：基于药物数据库中的药物信息参数，确定与药物筛选模型相匹配的训练样本集合。

其中，训练样本集合包括至少一组训练样本。本领域技术人员应了解，训练样本通常应该包括目标分子的结构，以及该目标分子与特定蛋白质分子结合后测试记录得到的活性标签（或称活性标签值）。

步骤 802：通过药物筛选模型提取与训练样本相匹配的特征集合。

步骤 803：根据与训练样本相匹配的特征集合对药物筛选模型进行训练，以确定与药物筛选模型相适配的模型参数。

这里，训练后的药物筛选模型可以用于对蛋白质分子和目标分子的结合进行活性预测。

在一些实施例中，基于药物数据库中的药物信息参数，还可以确定与药物筛选模型相匹配的验证样本集合，该验证样本集合用于结合训练样本集合对药物筛选模型进行训练。例如，验证样本集合可以用于验证根据训练样本集合训练后的药物筛选模型是否达到预期训练效果（如设定的精确率、召回率或 F1 分数等），若达到，则确定训练完成；若未达到，则根据训练样本集合继续训练。

在本申请的一些实施例中，药物筛选模型的训练方法还包括：

确定与药物筛选模型相匹配的多维损失函数；基于多维损失函数对药物筛选模型的参数（权重参数）进行调整，其中，调整后的药物筛选模型用于对蛋白质分子和目标分子的结合进行活性预测。在一些实施例中，在对药物筛选模型的训练过程中，可

以使用多个损失函数来对模型进行多监督训练。在一些实施例中，损失函数可以采用双分支均方差值损失函数（MSE，Mean Square Error）的形式。例如，双分支均方差值损失函数可以包括公式 20 以及公式 21 中的至少之一：

$$L_{pred_a} = MSE(pred_a, Target) \quad \text{公式 20}$$

$$L_{pred_b} = MSE(pred_b, Target) \quad \text{公式 21}$$

公式 20 用于计算预测值$pred_a$与训练样本的活性标签之间的均方差以作为损失值，公式 21 用于计算预测值$pred_b$与训练样本的活性标签之间的均方差以作为损失值，如此，基于两个损失值中的至少之一在药物筛选模型中进行逆向传播，以对药物筛选模型中的可训练参数进行更新。

在本申请的一些实施例中，为了使得公式 13 和公式 19 计算得到的两个活性预测值相同，也可以在损失函数中包含例如公式 22 的差异性损失，也即将两个活性预测值的区别纳入损失函数中：

$$L_{dis} = MSE(pred_a, pred_b) \quad \text{公式 22}$$

由此，可以有效的限制某一类的极值分布，从而限制此类的离散程度，有效的提升算法对不平衡数据的鲁棒性，也可以有效地防止药物筛选模型的处理结果过拟合。

同时考虑到实际应用中不但可以通过固定的药物筛选服务器实现本申请的方案，同时由于药物数据库中蛋白质分子和目标分子的数量众多，因此还可以通过药物筛选服务器群组（集群）实现本申请的方案。

下面继续说明本申请实施例提供的药物筛选装置实施为软件模块的示例性结构，在一些实施例中，如图 2 所示，存储在存储器 202 的药物筛选装置中的软件模块可以包括：信息传输模块 2081，配置为获取药物数据库中包含的蛋白质分子和目标分子；信息处理模块 2082，配置为：确定蛋白质分子的结构特征和目标分子的结构特征；基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络；根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值。

在一些实施例中，信息处理模块 2082，还配置为：基于第一活性预测值，对药物数据库中的分子进行筛选。

在一些实施例中，信息处理模块 2082，还配置为：确定蛋白质分子中的不同氨基酸链的空间位置；基于不同氨基酸链的空间位置，确定每对氨基酸之间的距离，并对每对氨基酸之间的距离进行标准化处理，得到标准氨基酸距离；基于标准氨基酸距离及氨基酸距离阈值，确定蛋白质分子对应的氨基酸矩阵图；基于蛋白质分子对应的氨基酸矩阵图，确定蛋白质分子的结构特征；确定目标分子所对应的原子和化学键，并基于目标分子所对应的原子和化学键，确定目标分子的结构特征。

在一些实施例中，信息处理模块 2082，还配置为：基于蛋白质分子的结构特征，确定目标节点的目标节点特征，目标节点对应于蛋白质分子中的氨基酸；基于蛋白质分子的结构特征，确定一端为目标节点的边线的相连边线特征；基于蛋白质分子的结构特征，确定与目标节点相连的相连节点的相连节点特征；节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，得到对应于蛋白质分子和目标分子的拼接节点特征。

在一些实施例中，信息处理模块 2082，还配置为：节点信息传递子网络基于目标节点特征、相连边线特征、和相连节点特征，产生对应于目标节点的目标节点嵌入表示；基于目标节点嵌入表示，节点信息传递子网络得到对应于蛋白质分子的蛋白质节点嵌入表示向量；基于目标分子的结构特征，节点信息传递子网络得到对应于目标分

子的目标分子节点嵌入表示向量；拼接蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接节点特征。

在一些实施例中，信息处理模块2082，还配置为：基于目标节点特征，得到目标节点初始状态特征；基于相连节点的相连节点特征，得到相连节点状态特征；节点信息传递子网络的第一信息汇集函数，对相连节点状态特征和相连边线特征进行结合，得到目标节点信息特征；基于目标节点初始状态特征和目标节点信息特征，节点信息传递子网络的更新函数对目标节点状态特征进行更新；根据更新后的相连节点状态特征，节点信息传递子网络产生目标节点嵌入表示。

在一些实施例中，信息处理模块2082，还配置为：节点信息传递子网络的第二信息汇集函数，对更新后的相连节点状态特征和相连节点特征进行结合，得到目标节点嵌入特征；节点信息传递子网络的激活函数，对目标节点嵌入特征进行处理，得到目标节点嵌入表示。

在一些实施例中，信息处理模块2082，还配置为：确定与药物筛选模型相匹配的自注意力读出函数；通过自注意力读出函数、蛋白质节点嵌入表示向量和目标分子节点嵌入表示向量，确定蛋白质分子的结构特征中的第一节点特征向量和目标分子的结构特征中的第二节点特征向量；对第一节点特征向量和第二节点特征向量进行拼接，得到对应于蛋白质分子和目标分子的拼接节点特征。

在一些实施例中，信息处理模块2082，还配置为：基于边线信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接边线特征；根据拼接边线特征预测蛋白质分子和目标分子结合后的第二活性预测值。

在一些实施例中，信息处理模块2082，还配置为：基于蛋白质分子的结构特征，确定目标边线的目标边线特征，目标边线特征对应于蛋白质分子中相连的两个氨基酸；基于蛋白质分子的结构特征，确定临边的临边边线特征，临边的第一端节点对应于相连的两个氨基酸之一，临边的第二端节点与第一端节点相连；确定与第二端节点对应的相邻节点特征；边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，得到对应于蛋白质分子和目标分子的拼接边线特征。

在一些实施例中，信息处理模块2082，还配置为：边线信息传递子网络基于目标边线特征、临边边线特征和相邻节点特征，产生对应于第一端节点的边线嵌入表示；基于边线嵌入表示，边线信息传递子网络得到对应于蛋白质分子的蛋白质边线嵌入表示向量；基于目标分子的结构特征，边线信息传递子网络得到对应于目标分子的目标分子边线嵌入表示向量；拼接蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量，得到对应于蛋白质分子和目标分子的拼接边线特征。

在一些实施例中，信息处理模块2082，还配置为：基于目标边线特征，得到目标边线初始状态特征；基于临边边线特征，得到临边边线状态特征；边线信息传递子网络的第一信息传递函数对临边边线状态特征和相邻节点特征进行结合，得到目标边线信息特征；基于目标边线信息特征和目标边线初始状态特征，边线信息传递子网络的更新函数对目标边线状态特征进行更新；根据更新后的临边边线状态特征，边线信息传递子网络产生边线嵌入表示。

在一些实施例中，信息处理模块2082，还配置为：边线信息传递子网络的第二信息传递函数，对更新后的临边边线状态特征和相邻节点特征进行结合，得到对应于第一端节点的边线嵌入特征；边线信息传递子网络的激活函数，对边线嵌入特征进行处理，得到边线嵌入表示。

在一些实施例中，信息处理模块2082，还配置为：确定与药物筛选模型相匹配的自注意力读出函数；通过自注意力读出函数、蛋白质边线嵌入表示向量和目标分子边线嵌入表示向量，确定蛋白质分子的结构特征中的第一边线特征向量、以及目标分子的结构特征中的第二边线特征向量；对第一边线特征向量和第二边线特征向量进行拼接，得到对应于蛋白质分子和目标分子的拼接边线特征。

在一些实施例中，信息处理模块2082，还配置为：基于第一活性预测值和第二活性预测值，对药物数据库中的分子进行筛选。

在一些实施例中，药物筛选装置还包括训练模块，配置为：基于药物数据库中的药物信息参数，确定与药物筛选模型相匹配的训练样本集合，其中，训练样本集合包括至少一组训练样本；通过药物筛选模型提取与训练样本相匹配的特征集合；根据与训练样本相匹配的特征集合对药物筛选模型进行训练，以确定与药物筛选模型相适配的模型参数。

在一些实施例中，训练模块还配置为：确定与药物筛选模型相匹配的多维损失函数；基于多维损失函数对药物筛选模型的参数进行调整；其中，调整后的药物筛选模型用于对蛋白质分子和目标分子的结合进行活性预测。

在一些实施例中，药物筛选模型的训练过程中，损失函数包括以下至少之一：第一活性预测值与训练样本的活性标签之间的均方差损失函数；第二活性预测值与活性标签之间的均方差损失函数；第一活性预测值与第二活性预测值之间的均方差损失函数。

本申请实施例提供一种存储有可执行指令的计算机可读存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本申请实施例提供的药物筛选方法。

在一些实施例中，计算机可读存储介质可以是FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言（包括编译或解释语言，或者声明性或过程性语言）来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言（HTML，Hyper Text Markup Language）文档中的一个或多个脚本中，存储在专用于所讨论的程序的单个文件中，或者，存储在多个协同文件（例如，存储一个或多个模块、子程序或代码部分的文件）中。

作为示例，可执行指令可被部署为在一个电子设备上执行，或者在位于一个地点的多个电子设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个电子设备上执行。

综上，相比于传统的药物筛选技术，本申请实施例至少存在以下技术效果：1）通过药物筛选模型，在没有人工干预的情况下快速地给出药物-靶向蛋白可能的相互作用对，从而节省药物研发试验成本，加快了药物新功能的挖掘和发现，节省药物筛选成本，也提升了用户的使用体验；2）不但能够有效地通过药物筛选模型表现蛋白质图的结构特征和小分子图的结构特征，实现准确地将蛋白质分子和目标分子进行结合，还能够高效地对药物数据库中所包含的数量巨大的蛋白质分子和目标分子进行处理，提升药物筛选的效率，节省药物筛选的时间。

以上，仅为本申请的实施例而已，并非用于限定本申请的保护范围，凡在本申请的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本申请的保护范围之内。

权利要求书

1、一种药物筛选方法，由电子设备执行，所述方法包括：

获取药物数据库中包含的蛋白质分子和目标分子；

确定所述蛋白质分子的结构特征和所述目标分子的结构特征；

基于药物筛选模型中的节点信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，其中所述节点信息传递子网络为图神经网络；

根据所述拼接节点特征预测所述蛋白质分子和所述目标分子结合后的第一活性预测值。

2、根据权利要求1所述的方法，其中，所述方法还包括：

基于所述第一活性预测值，对所述药物数据库中的分子进行筛选。

3、根据权利要求1所述的方法，其中，所述确定所述蛋白质分子的结构特征和所述目标分子的结构特征，包括：

确定所述蛋白质分子中的不同氨基酸链的空间位置；

基于所述不同氨基酸链的空间位置，确定每对氨基酸之间的距离，并对所述每对氨基酸之间的距离进行标准化处理，得到标准氨基酸距离；

基于所述标准氨基酸距离及氨基酸距离阈值，确定所述蛋白质分子对应的氨基酸矩阵图；

基于所述蛋白质分子对应的氨基酸矩阵图，确定所述蛋白质分子的结构特征；

确定所述目标分子所对应的原子和化学键，并基于所述目标分子所对应的原子和化学键，确定所述目标分子的结构特征。

4、根据权利要求1所述的方法，其中，所述基于药物筛选模型中的节点信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，包括：

基于所述蛋白质分子的结构特征，确定目标节点的目标节点特征，所述目标节点对应于所述蛋白质分子中的氨基酸；

基于所述蛋白质分子的结构特征，确定一端为所述目标节点的边线的相连边线特征；

基于所述蛋白质分子的结构特征，确定与所述目标节点相连的相连节点的相连节点特征；

所述节点信息传递子网络基于所述目标节点特征、所述相连边线特征、和所述相连节点特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征。

5、根据权利要求4所述的方法，其中，所述节点信息传递子网络基于所述目标节点特征、所述相连边线特征、和所述相连节点特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，包括：

所述节点信息传递子网络基于所述目标节点特征、所述相连边线特征、和所述相连节点特征，产生对应于所述目标节点的目标节点嵌入表示；

基于所述目标节点嵌入表示，所述节点信息传递子网络得到对应于所述蛋白质分子的蛋白质节点嵌入表示向量；

基于所述目标分子的结构特征，所述节点信息传递子网络得到对应于所述目标分子的目标分子节点嵌入表示向量；

拼接所述蛋白质节点嵌入表示向量和所述目标分子节点嵌入表示向量，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征。

6、根据权利要求5所述的方法，其中，所述节点信息传递子网络基于所述目标节点特征、所述相连边线特征、和所述相连节点特征，产生对应于所述目标节点的目标节点嵌入表示，包括：

基于所述目标节点特征，得到目标节点初始状态特征；

基于所述相连节点的相连节点特征，得到相连节点状态特征；

所述节点信息传递子网络的第一信息汇集函数，对所述相连节点状态特征和所述相连边线特征进行结合，得到目标节点信息特征；

基于所述目标节点初始状态特征和所述目标节点信息特征，所述节点信息传递子网络的更新函数对目标节点状态特征进行更新；

根据更新后的所述相连节点状态特征，所述节点信息传递子网络产生所述目标节点嵌入表示。

7、根据权利要求6所述的方法，其中，所述根据更新后的所述相连节点状态特征，所述节点信息传递子网络产生所述目标节点嵌入表示，包括：

所述节点信息传递子网络的第二信息汇集函数，对更新后的所述相连节点状态特征和所述相连节点特征进行结合，得到目标节点嵌入特征；

所述节点信息传递子网络的激活函数，对所述目标节点嵌入特征进行处理，得到所述目标节点嵌入表示。

8、根据权利要求5所述的方法，其中，所述拼接所述蛋白质节点嵌入表示向量和所述目标分子节点嵌入表示向量，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，包括：

确定与所述药物筛选模型相匹配的自注意力读出函数；

通过所述自注意力读出函数、所述蛋白质节点嵌入表示向量和所述目标分子节点嵌入表示向量，确定所述蛋白质分子的结构特征中的第一节点特征向量和所述目标分子的结构特征中的第二节点特征向量；

对所述第一节点特征向量和所述第二节点特征向量进行拼接，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征。

9、根据权利要求1至8任一项所述的方法，其中，所述药物筛选模型还包括边线信息传递子网络，所述方法还包括：

基于所述边线信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征；

根据所述拼接边线特征预测所述蛋白质分子和所述目标分子结合后的第二活性预测值。

10、根据权利要求9所述的方法，其中，所述基于所述边线信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征，包括：

基于所述蛋白质分子的结构特征，确定目标边线的目标边线特征，所述目标边线特征对应于所述蛋白质分子中相连的两个氨基酸；

基于所述蛋白质分子的结构特征，确定临边的临边边线特征，所述临边的第一端节点对应于所述相连的两个氨基酸之一，所述临边的第二端节点与所述第一端节点相连；

确定与所述第二端节点对应的相邻节点特征；

所述边线信息传递子网络基于所述目标边线特征、所述临边边线特征和所述相邻节点特征，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征。

11、根据权利要求 10 所述的方法，其中，所述边线信息传递子网络基于所述目标边线特征、所述临边边线特征和所述相邻节点特征，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征，还包括：

所述边线信息传递子网络基于所述目标边线特征、所述临边边线特征和所述相邻节点特征，产生对应于所述第一端节点的边线嵌入表示；

基于所述边线嵌入表示，所述边线信息传递子网络得到对应于所述蛋白质分子的蛋白质边线嵌入表示向量；

基于所述目标分子的结构特征，所述边线信息传递子网络得到对应于所述目标分子的目标分子边线嵌入表示向量；

拼接所述蛋白质边线嵌入表示向量和所述目标分子边线嵌入表示向量，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征。

12、根据权利要求 11 所述的方法，其中，所述边线信息传递子网络基于所述目标边线特征、所述临边边线特征和所述相邻节点特征，产生对应于所述第一端节点的边线嵌入表示，包括：

基于所述目标边线特征，得到目标边线初始状态特征；

基于所述临边边线特征，得到临边边线状态特征；

所述边线信息传递子网络的第一信息传递函数对所述临边边线状态特征和所述相邻节点特征进行结合，得到目标边线信息特征；

基于所述目标边线信息特征和所述目标边线初始状态特征，所述边线信息传递子网络的更新函数对目标边线状态特征进行更新；

根据更新后的所述临边边线状态特征，所述边线信息传递子网络产生所述边线嵌入表示。

13、根据权利要求 12 所述的方法，其中，所述根据更新后的所述临边边线状态特征，所述边线信息传递子网络产生所述边线嵌入表示，包括：

所述边线信息传递子网络的第二信息传递函数，对所述更新后的临边边线状态特征和所述相邻节点特征进行结合，得到对应于所述第一端节点的边线嵌入特征；

所述边线信息传递子网络的激活函数，对所述边线嵌入特征进行处理，得到所述边线嵌入表示。

14、根据权利要求 11 所述的方法，其中，所述拼接所述蛋白质边线嵌入表示向量和所述目标分子边线嵌入表示向量，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征，包括：

确定与所述药物筛选模型相匹配的自注意力读出函数；

通过所述自注意力读出函数、所述蛋白质边线嵌入表示向量和所述目标分子边线嵌入表示向量，确定所述蛋白质分子的结构特征中的第一边线特征向量、以及所述目标分子的结构特征中的第二边线特征向量；

对所述第一边线特征向量和所述第二边线特征向量进行拼接，得到对应于所述蛋白质分子和所述目标分子的拼接边线特征。

15、根据权利要求 9 所述的方法，其中，所述方法还包括：

基于所述第一活性预测值和所述第二活性预测值，对所述药物数据库中的分子进行筛选。

16、根据权利要求 9 所述的方法，其中，所述药物筛选模型的训练过程中，损失函数包括以下至少之一：

第一活性预测值与训练样本的活性标签之间的均方差损失函数；

第二活性预测值与所述活性标签之间的均方差损失函数；

第一活性预测值与第二活性预测值之间的均方差损失函数。

17、一种药物筛选的装置，所述装置包括：

信息传输模块，配置为获取药物数据库中包含的蛋白质分子和目标分子；

信息处理模块，配置为：

确定所述蛋白质分子的结构特征和所述目标分子的结构特征；

基于药物筛选模型中的节点信息传递子网络、所述蛋白质分子的结构特征以及所述目标分子的结构特征，得到对应于所述蛋白质分子和所述目标分子的拼接节点特征，其中所述节点信息传递子网络为图神经网络；

根据所述拼接节点特征预测所述蛋白质分子和所述目标分子结合后的第一活性预测值。

18、一种电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于运行所述存储器存储的可执行指令时，实现权利要求 1 至 16 任一项所述的药物筛选方法。

19、一种计算机可读存储介质，存储有可执行指令，用于被处理器执行时，实现权利要求 1 至 16 任一项所述的药物筛选方法。

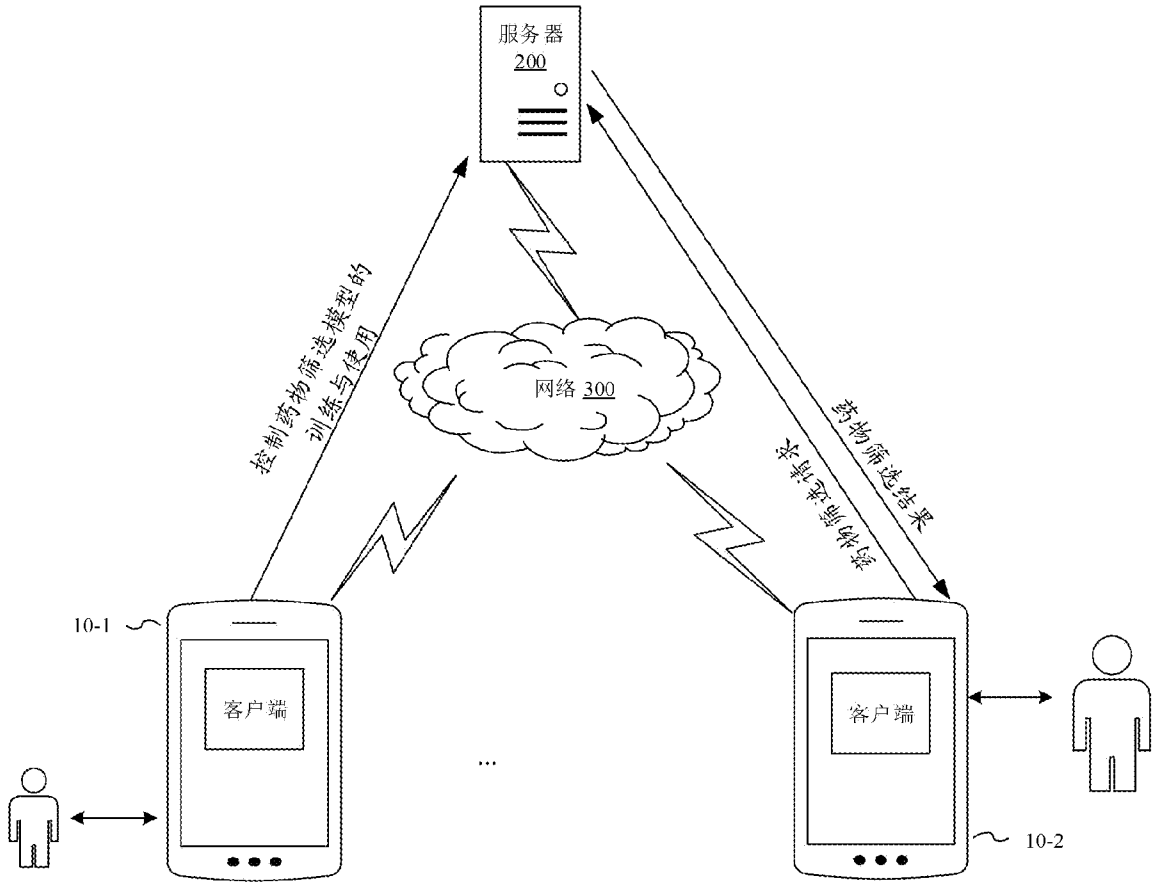

服务器
200
网络 300
控制药物筛选模型的
训练与使用
药物筛选请求
药物筛选结果
10-1
客户端
客户端
10-2
...

图 1

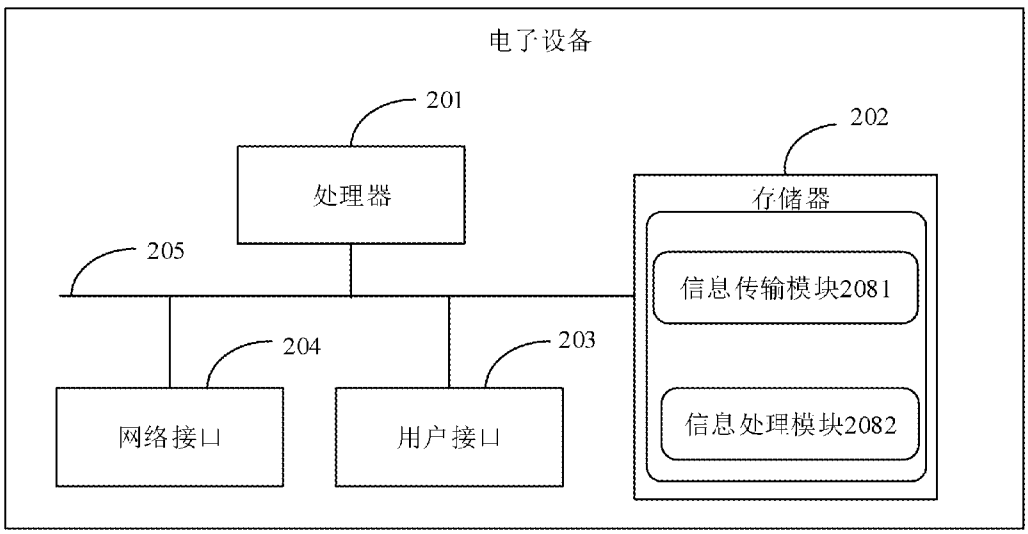

电子设备
201
处理器
202
存储器
205
信息传输模块2081
204
203
网络接口
用户接口
信息处理模块2082

图 2

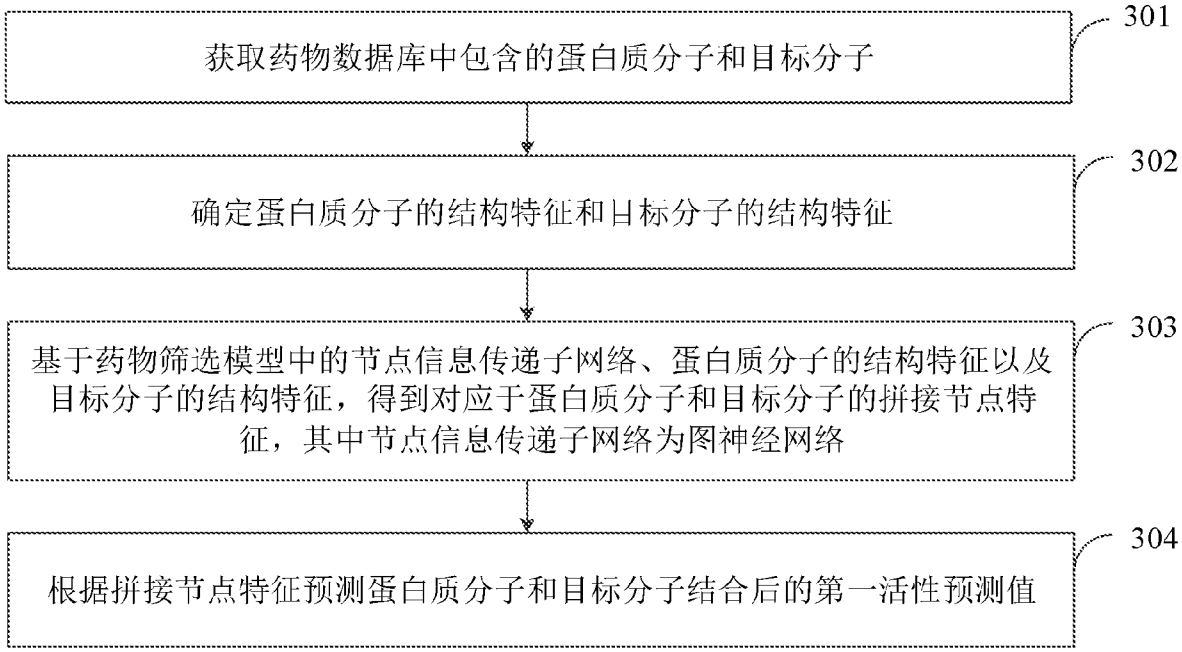

301
获取药物数据库中包含的蛋白质分子和目标分子
302
确定蛋白质分子的结构特征和目标分子的结构特征
303
基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络
304
根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值

图 3A

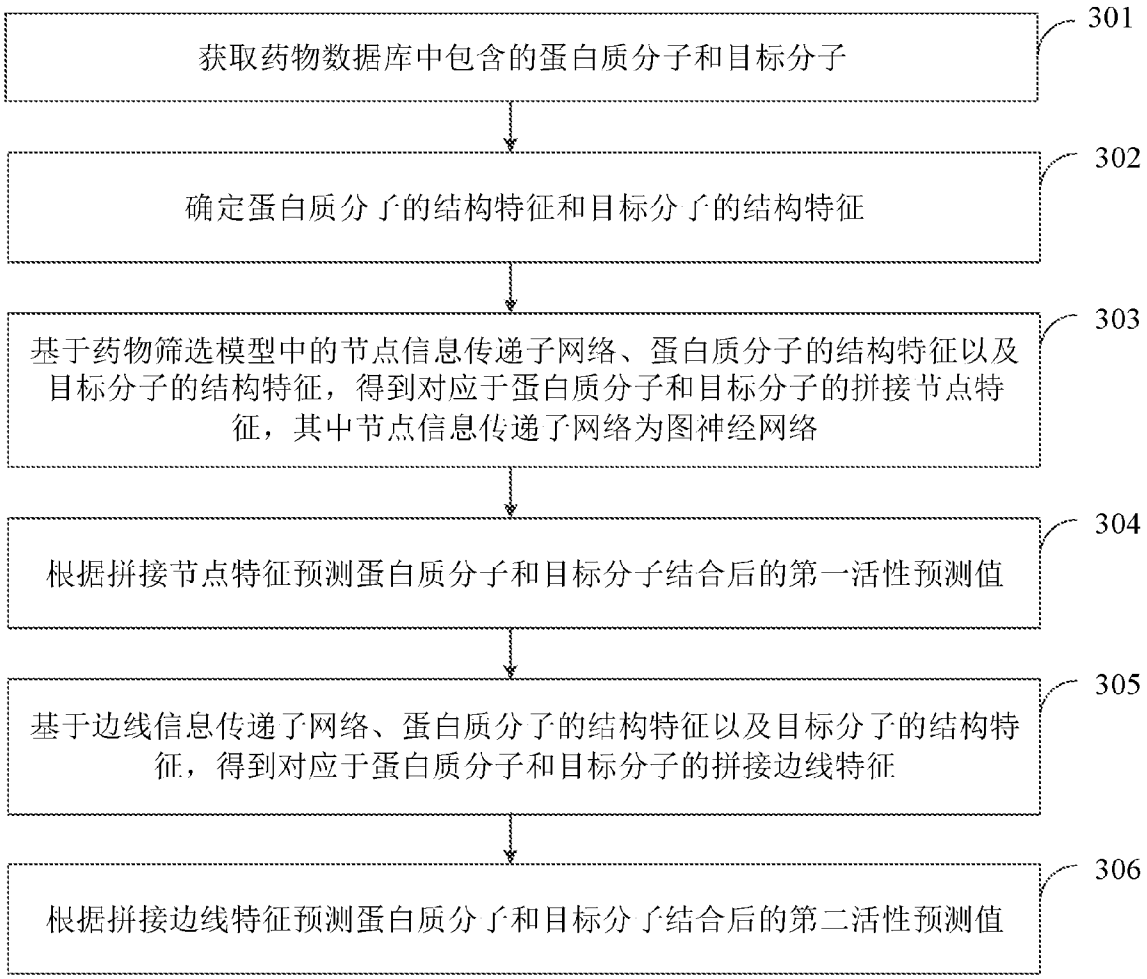

301
获取药物数据库中包含的蛋白质分子和目标分子
302
确定蛋白质分子的结构特征和目标分子的结构特征
303
基于药物筛选模型中的节点信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接节点特征，其中节点信息传递子网络为图神经网络
304
根据拼接节点特征预测蛋白质分子和目标分子结合后的第一活性预测值
305
基于边线信息传递子网络、蛋白质分子的结构特征以及目标分子的结构特征，得到对应于蛋白质分子和目标分子的拼接边线特征
306
根据拼接边线特征预测蛋白质分子和目标分子结合后的第二活性预测值

图 3B

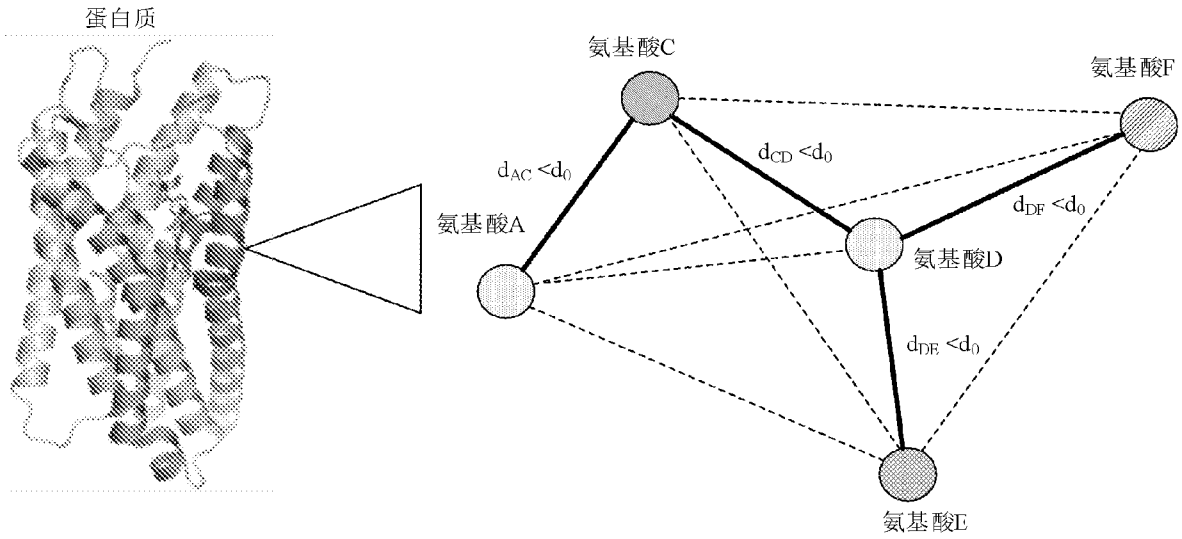

蛋白质
氨基酸C
氨基酸F
$d_{AC}<d_0$
$d_{CD}<d_0$
$d_{DF}<d_0$
氨基酸A
氨基酸D
$d_{DE}<d_0$
氨基酸E

图 4

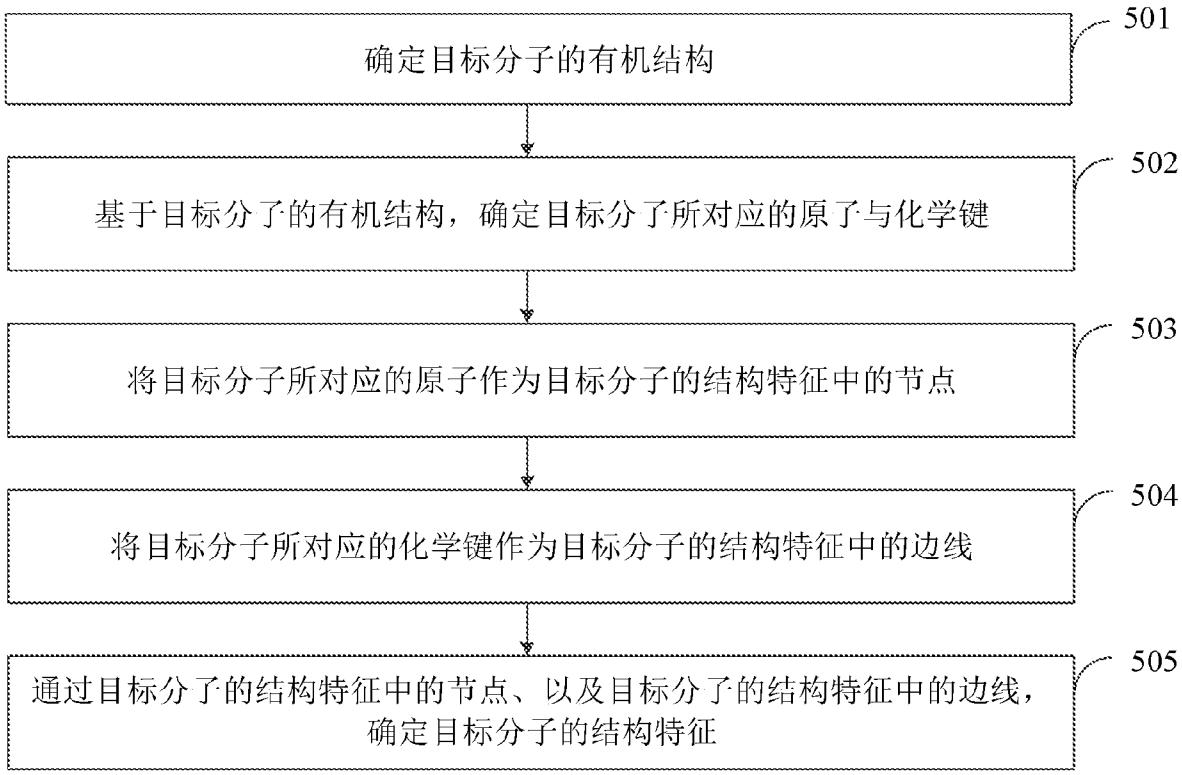

501
确定目标分子的有机结构
502
基于目标分子的有机结构，确定目标分子所对应的原子与化学键
503
将目标分子所对应的原子作为目标分子的结构特征中的节点
504
将目标分子所对应的化学键作为目标分子的结构特征中的边线
505
通过目标分子的结构特征中的节点、以及目标分子的结构特征中的边线，确定目标分子的结构特征

图 5

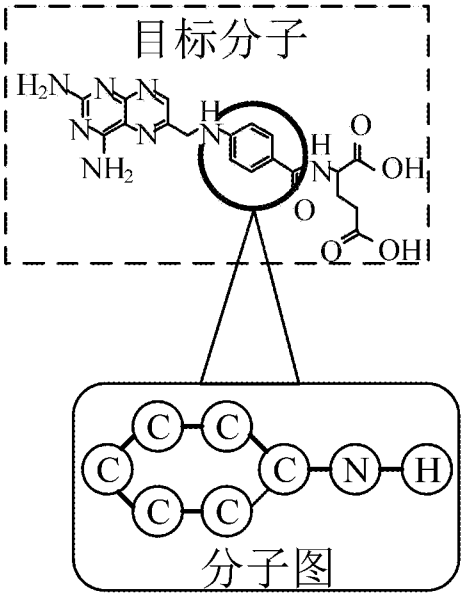

目标分子
分子图

图 6

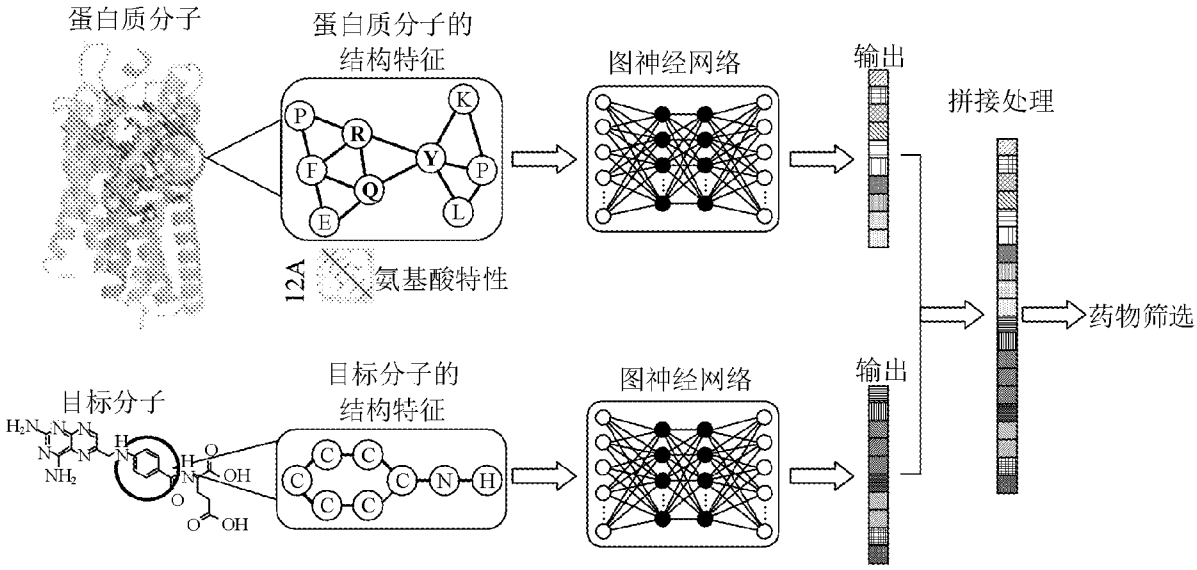

蛋白质分子
蛋白质分子的
结构特征
图神经网络
输出
拼接处理
药物筛选
12A
氨基酸特性
目标分子
目标分子的
结构特征
图神经网络
输出

图 7

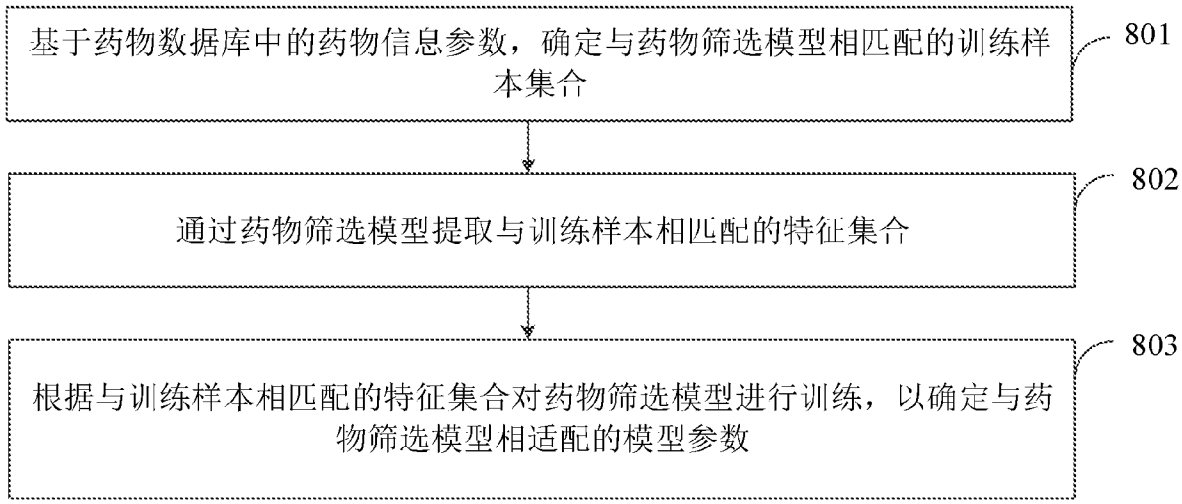

基于药物数据库中的药物信息参数，确定与药物筛选模型相匹配的训练样本集合
801
通过药物筛选模型提取与训练样本相匹配的特征集合
802
根据与训练样本相匹配的特征集合对药物筛选模型进行训练，以确定与药物筛选模型相适配的模型参数
803

图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/107509

A. CLASSIFICATION OF SUBJECT MATTER

G16B 20/00(2019.01)i; G16B 40/00(2019.01)n; G16B 50/30(2019.01)n; G16C 20/50(2019.01)n

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B, G16C

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, EPODOC, WPI, CNKI, ISI, IEEE: 图神经网络, 图网络, 图卷积神经网络, 药物, 选, 蛋白质, 基因, 分子, 节点, 边, 特征, 拼接, 连接, 预测, 活性, GNN, GCN, MPNN, protein, gene, screen, drug, molecular, node, edge, concatnate, feature, predict, activity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PY	CN 111816252 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 23 October 2020 (2020-10-23) claims 1-15, description paragraphs [0074]-[0077], [0093]-[0102], [0145]-[0202]	1-19
PY	CN 112151128 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 29 December 2020 (2020-12-29) description paragraphs [0187]-[0199], [0210]-[0240]	1-19
A	US 2019272468 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 05 September 2019 (2019-09-05) entire document	1-19
A	WO 2018220368 A1 (GTN LTD.) 06 December 2018 (2018-12-06) entire document	1-19
A	CN 104350141 A (OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY et al.) 11 February 2015 (2015-02-11) entire document	1-19

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 October 2021	**27 October 2021**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/107509

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111816252	A	23 October 2020	None			
CN	112151128	A	29 December 2020	None			
US	2019272468	A1	05 September 2019	CN	112533941	A	19 March 2021
				EP	3762405	A1	13 January 2021
				AU	2019231255	A1	01 October 2020
				KR	20200129130	A	17 November 2020
				WO	2019173401	A1	12 September 2019
				CA	3093246	A1	12 September 2019
WO	2018220368	A1	06 December 2018	EP	3646250	A1	06 May 2020
				US	2021081804	A1	18 March 2021
CN	104350141	A	11 February 2015	JP	2013255467	A	26 December 2013
				CA	2876051	A1	19 December 2013
				EP	2862924	A1	22 April 2015
				AU	2013275211	A1	15 January 2015
				SG	11201408017V	A	29 January 2015
				WO	2013187468	A1	19 December 2013
				KR	20150008179	A	21 January 2015
				RU	2014153560	A	10 August 2016
				US	2015169822	A1	18 June 2015

国际检索报告

国际申请号
PCT/CN2021/107509

A. 主题的分类

G16B 20/00(2019.01)i; G16B 40/00(2019.01)n; G16B 50/30(2019.01)n; G16C 20/50(2019.01)n

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G16B, G16C

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, EPODOC, WPI, CNKI, ISI, IEEE: 图神经网络，图网络，图卷积神经网络，药物，选，蛋白质，基因，分子，节点，边，特征，拼接，连接，预测，活性，GNN, GCN, MPNN, protein, gene, screen, drug, molecular, node, edge, concatnate, feature, predict, activity

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PY	CN 111816252 A (腾讯科技深圳有限公司) 2020年 10月 23日 (2020-10-23) 权利要求1-15，说明书第[0074]-[0077]、[0093]-[0102]、[0145]-[0202]段	1-19
PY	CN 112151128 A (腾讯科技深圳有限公司) 2020年 12月 29日 (2020-12-29) 说明书第[0187]-[0199]、[0210]-[0240]段	1-19
A	US 2019272468 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 2019年 9月 5日 (2019-09-05) 全文	1-19
A	WO 2018220368 A1 (GTN LTD) 2018年 12月 6日 (2018-12-06) 全文	1-19
A	CN 104350141 A (学校法人冲绳科学技术大学院大学学园 等) 2015年 2月 11日 (2015-02-11) 全文	1-19

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2021年 10月 11日	2021年 10月 27日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088	欧晓丹
传真号 (86-10)62019451	电话号码 86-(10)-53961416

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/107509

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	111816252	A	2020年 10月 23日	无			
CN	112151128	A	2020年 12月 29日	无			
US	2019272468	A1	2019年 9月 5日	CN	112533941	A	2021年 3月 19日
				EP	3762405	A1	2021年 1月 13日
				AU	2019231255	A1	2020年 10月 1日
				KR	20200129130	A	2020年 11月 17日
				WO	2019173401	A1	2019年 9月 12日
				CA	3093246	A1	2019年 9月 12日
WO	2018220368	A1	2018年 12月 6日	EP	3646250	A1	2020年 5月 6日
				US	2021081804	A1	2021年 3月 18日
CN	104350141	A	2015年 2月 11日	JP	2013255467	A	2013年 12月 26日
				CA	2876051	A1	2013年 12月 19日
				EP	2862924	A1	2015年 4月 22日
				AU	2013275211	A1	2015年 1月 15日
				SG	11201408017V	A	2015年 1月 29日
				WO	2013187468	A1	2013年 12月 19日
				KR	20150008179	A	2015年 1月 21日
				RU	2014153560	A	2016年 8月 10日
				US	2015169822	A1	2015年 6月 18日