(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110998739 A
(43)申请公布日 2020.04.10

(21)申请号 201880051716.0

(22)申请日 2018.08.03

(30)优先权数据
15/671,898 2017.08.08 US

(85)PCT国际申请进入国家阶段日
2020.02.07

(86)PCT国际申请的申请数据
PCT/IB2018/055836 2018.08.03

(87)PCT国际申请的公布数据
WO2019/030627 EN 2019.02.14

(71)申请人 国际商业机器公司
地址 美国纽约阿芒克

(72)发明人 罗衡 张平
A·B·福库伊-恩库特彻 胡建英

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅 彭梦晔

(51)Int.Cl.
G16C 20/30(2019.01)
G16C 20/70(2019.01)

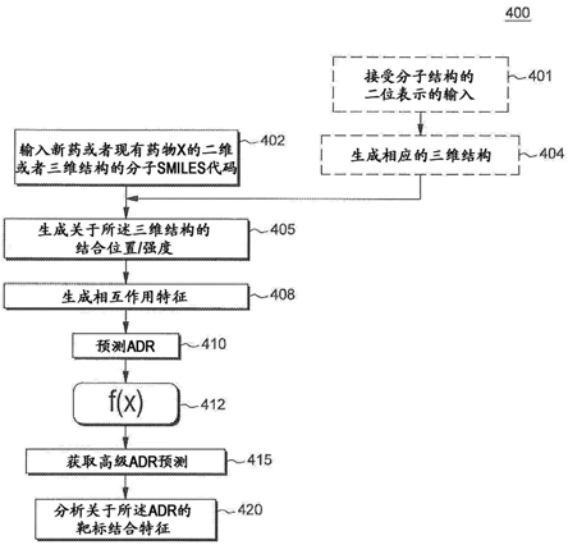
权利要求书4页 说明书13页 附图11页

(54)发明名称

不良药物反应的预测

(57)摘要

用于预测药物不良反应(ADR)的系统框架和方法。为药物小分子和独特的人类蛋白质制备了以三维表示的结构,并使用分子对接产生了它们之间的对接分数。使用分子对接功能开发了机器学习模型来预测ADR。使用所述机器学习模型,它可以基于药物-靶标相互作用特征和已知的药物-ADR关系成功预测药物诱导的ADR。通过进一步分析排名高级或与所述ADR紧密相关的所述结合蛋白,会发现所述ADR机制的可能解释。基于分子对接特征的所述机器学习ADR模型不仅有助于对新药或现有已知药物分子进行ADR预测,而且具有为ADR基本机理提供可能解释或假设的优势。



CN 110998739 A

1.一种自动预测药物的不良药物反应的方法,包括:

在处理器处接收与药物结构相关的数据;

使用所述处理器为所述药物计算多个药物-靶标相互作用特征,每个所述药物-靶标相互作用特征存在在所述药物结构和多个独特的、高分辨率靶蛋白结构中的每个之间;

在所述处理器处运行与对应的一种或多种已知药物不良反应(ADR)相关的一种或多种分类器模型;

使用所述一种或多种分类器模型中的每一种,基于涉及所述药物和所述一种或多种已知ADR的所述药物-靶标相互作用特征预测一种或多种ADR;以及

通过所述处理器生成指示所述预测的一个或多个ADR的输出。

2.根据权利要求1所述的方法,其中,所述多个药物-靶标相互作用特征的计算还包括:

使用所述处理器产生与所述药物结构和所述靶蛋白之间的结合潜力相关的分子对接分数;以及

使用所述处理器根据计算的所述对接分数对所述药物进行所述靶蛋白排名。

3.根据权利要求1或2所述的方法,其中,所接收的所述与药物结构相关的数据是药物分子的二维(2-D)表示,所述方法还包括:

将所述二维药物分子表示形式转换为所述药物分子结构的三维(3D)表示形式,其中每个所述药物-靶标相互作用特征在所述三维药物结构和多个独特的、高分辨率的靶标蛋白质结构的每个的结合受体之间。

4.根据权利要求1、2或3所述的方法,还包括:通过以下步骤确定预测的ADR的根本原因:

所述处理器识别出高级的靶蛋白结构,所述高级的靶蛋白结构涉及细胞表达或细胞分化;以及

确定涉及所述靶蛋白结构的所述细胞表达或细胞分化是否与和所述靶蛋白结构相关的所述预测的ADR相关。

5.根据任一前述权利要求所述的方法,进一步包括:

使用所述处理器训练与所述一个或多个已知ADRs中的每一个相对应的逻辑回归分类器模型,以基于每一个所述药物-靶标相互作用特征和相应的已知药物-ADR关系预测相应的ADR。

6.根据权利要求5所述的方法,其中,所述逻辑回归分类器模型的训练包括:

在所述处理器处接收与多个药物中的每一种的结构有关的数据;

在所述处理器处接收与多个蛋白质靶标中的每一个的结构有关的数据;

在所述处理器处获得包括多个药物中的每一个与多个靶标之间的分子对接分数的多个药物-靶标特征;

在所述处理器处获得包括一个或多个已知ADR的列表和对应的已知ADR-药物关系的数据;以及

在所述处理器处实现一种机器学习技术,以训练所述逻辑回归分类器模型基于所述分子对接分数和所述已知的ADR-药物关系预测ADR。

7.根据权利要求5或6所述的方法,其中,所述训练包括:

使用处理器收集第一特征矩阵,所述第一特征矩阵包含表示作为行的所述药物结构、

作为列的蛋白质和作为特征的所述分子结合得分的数据；

由所述处理器映射每一个所述药物结构与药物不良反应(ADR)之间的关系;以及

使用所述处理器为每个ADR确定所述药物是否与所述ADR相关联；

如果所述药物与所述ADR相关联,则根据第一二进制值将药物-ADR配对进行分类;否则,如果所述药物与所述ADR不相关联,则将所述药物分类为第二二进制值；

使用所述处理器收集二进制标签矩阵,所述矩阵包含作为行的药物和作为列的ADR；

使用所述分子对接得分作为特征,使用所述第一矩阵和所述第二矩阵为每个ADR开发所述逻辑回归分类器模型。

8.根据权利要求5、6或7所述的方法,其中用于特定ADR的每个逻辑回归分类器模型包括用于预测药物结构与所述特定ADR相关联的置信度得分的对应的逻辑回归函数,所述训练进一步包括::

所述处理器为对应的逻辑回归函数生成系数集用于指示与由一个特定的ADR预测所指示的一个或多个蛋白质靶标相关联的多个对应的分子对接得分的权重贡献。

9.根据权利要求8所述的方法,还包括:通过以下步骤确定预测的ADR的根本原因：

对于分类器模型,获得指示权重贡献的逻辑回归函数的每个所述生成系数的绝对值；

确定指示对分类器模型具有最大贡献的靶标蛋白质的最大权重贡献者;和

从对所述分类器模型贡献最大的所述靶蛋白质中识别出与所述特定ADR预测相关的蛋白质机制类型。

10.根据任一前述权利要求所述的方法,进一步包括：

修饰所述药物结构以避免与诱发所述预测的ADR的靶蛋白发生相互作用。

11.一种自动预测药物不良药物反应的系统,包括：

至少一个存储器存储设备;以及

一个或多个硬件处理器可操作地连接到所述至少一个存储器存储设备,所述一个或多个硬件处理器配置为：

接收与药物结构相关的数据；

为所述药物计算多个药物-靶标相互作用特征,每个所述药物-靶标相互作用特征存在在所述药物结构和多个独特的、高分辨率靶蛋白结构中的每个之间；

运行与对应的一种或多种已知药物不良反应(ADR)相关的一种或多种分类器模型；

使用所述一种或多种分类器模型,基于涉及所述药物和所述一种或多种已知ADR的所述药物-靶标相互作用特征预测一种或多种ADR;以及

生成指示所述预测的一个或多个ADR的输出。

12.根据权利要求11所述的系统,其中,为了计算所述多个药物-靶标相互作用特征,所述一个或多个硬件处理器还被配置为：

产生与所述药物结构和所述靶蛋白之间的结合潜力相关的分子对接分数;以及

根据计算的所述对接分数对所述药物进行所述靶蛋白排名。

13.根据权利要求11或12所述的系统,其中,所接收的所述与药物结构相关的数据是药物分子的二维(2-D)表示,所述一个或多个硬件处理器还被配置为：

将所述二维药物分子表示形式转换为所述药物分子结构的三维(3D)表示形式,其中每个所述药物-靶标相互作用特征在所述三维药物结构和多个独特的、高分辨率的靶标蛋白

质结构的每个的结合受体之间。

14.根据权利要求11、12或13所述的系统，其中，所述一个或多个硬件处理器还被配置为通过以下方式确定预测的ADR的根本原因：

识别出高级的靶蛋白结构，所述高级的靶蛋白结构涉及细胞表达或细胞分化；以及

确定涉及所述靶蛋白结构的所述细胞表达或细胞分化是否与和所述靶蛋白结构相关的所述预测的ADR相关。

15.根据权利要求11至14中的任一项所述的系统，其中，所述一个或多个硬件处理器还被配置为：

训练与所述一个或多个已知ADRs中的每一个相对应的逻辑回归分类器模型，以基于每一个所述药物-靶标相互作用特征和相应的已知药物-ADR关系预测相应的ADR。

16.根据权利要求15所述的系统，其中，为了训练所述逻辑回归分类器模型，所述一个或多个硬件处理器还被配置为：

接收与多个药物中的每一种的结构有关的数据；

接收与多个蛋白质靶标中的每一个的结构有关的数据；

获得包括多个药物中的每一个与多个靶标之间的分子对接分数的多个药物-靶标特征；

获得包括一个或多个已知ADR的列表和对应的已知ADR-药物关系的数据；以及

实现一种机器学习技术，以训练所述逻辑回归分类器模型基于所述分子对接分数和所述已知的ADR-药物关系预测ADR。

17.根据权利要求15或16所述的系统，其中，为了训练所述逻辑回归分类器模型，所述一个或多个硬件处理器还被配置为：

收集第一特征矩阵，所述第一特征矩阵包含表示作为行的所述药物结构、作为列的蛋白质和作为特征的所述分子结合得分的数据；

映射每一个所述药物结构与药物不良反应(ADR)之间的关系；以及

为每个ADR确定所述药物是否与所述ADR相关联；

如果所述药物与所述ADR相关联，则根据第一二进制值将药物-ADR配对进行分类；否则，如果所述药物与所述ADR不相关联，则将所述药物分类为第二二进制值；

收集二进制标签矩阵，所述矩阵包含作为行的药物和作为列的ADR；

使用所述分子对接得分作为特征，使用所述第一矩阵和所述第二矩阵为每个ADR开发所述逻辑回归分类器模型。

18.根据权利要求15、16或17所述的系统，其中用于特定ADR的每个逻辑回归分类器模型包括用于预测药物结构与所述特定ADR相关联的置信度得分的对应的逻辑回归函数，其中训练在逻辑回归分类器模型中，一个或多个硬件处理器进一步配置为：

所述处理器为对应的逻辑回归函数生成系数集用于指示与由一个特定的ADR预测所指示的一个或多个蛋白质靶标相关联的多个对应的分子对接得分的权重贡献。

19.根据权利要求18所述的系统，其中，所述一个或多个硬件处理器还被配置为通过以下方式确定预测的ADR的根本原因：

对于分类器模型，获得指示权重贡献的逻辑回归函数的每个所述生成系数的绝对值；

确定指示对分类器模型具有最大贡献的靶标蛋白质的最大权重贡献者；和

从对所述分类器模型贡献最大的所述靶蛋白质中识别出与所述特定ADR预测相关的蛋白质机制类型。

20.根据权利要求11至19中的任一项所述的系统,其中,所述一个或多个硬件处理器还被配置为:

修饰所述药物结构以避免与诱发所述预测的ADR的靶蛋白发生相互作用。

不良药物反应的预测

技术领域

本发明总体上涉及用于预测药物不良反应的系统和方法,尤其涉及用于预测候选药物和未检测到的市售药物的不良反应以及确定相关靶标的潜在药物不良反应(ADR)的框架。其他方面允许使用该框架来评估有关某些ADR的行动机制。

背景技术

已经开发了机器学习模型以预测药物不良反应并提高药物安全性。尽管某些预测方法很有效,但是大多数机器学习模型都无法提供足够的生物学解释(如果有的话)来预测结果,尤其是与靶标绑定有关的信息。

药物不良反应(ADR)很复杂,并且可能因个体而异。确定相关靶标不仅有助于了解ADR的机制,而且有助于集中于潜在的致病方面,例如基因突变,从而有助于改善精密医学。

尽管已经开发出使用多种特征(例如,化学结构、结合测定和表型信息)和模型(例如,逻辑回归、随机森林和支持向量机)来预测药物不良反应的计算方法,但大多数研究集中在特征多样性和模型性能上,而不是机制解释的假设生成。

发明内容

一种通过仅需要药物分子的结构输入来预测新药物或候选药物可能的ADR的系统、方法和计算机程序产品。此外,可以识别/突出显示可能在导致此类ADR中起关键作用的相关结合靶标。

根据一个实施例,提供了一种方法,该方法自动预测新药的药物不良反应或预测当前市售药物的未检测到的药物不良反应。

该方法包括:在处理器处接收关于药物分子结构的数据;使用所述处理器为所述药物计算多个药物-靶标相互作用特征,每个药物-靶标相互作用特征在所述药物分子结构与多个独特的高分辨率靶蛋白结构中的相应一个之间相关;在所述处理器处运行与对应的一个或多个已知药物不良反应(ADR)相关的一个或多个分类器模型;使用所述一个或多个分类器模型中的每一种,基于所述药物-靶标相互作用特征和已知药物的ADR关系预测一个或多个ADR;以及由所述处理器生成指示所述预测的一个或多个ADR的输出。

在另一个实施例中,提供了一种系统,该系统自动预测药物的药物不良反应。该系统包括:至少一个存储器存储设备;和可操作地连接到所述至少一个存储器存储设备的一个或多个硬件处理器,所述一个或多个硬件处理器被配置为:接收关于药物的分子结构的数据;以及为所述药物计算多个药物-靶标相互作用特征,每个药物-靶标相互作用特征存在于所述药物分子结构以及多个独特的高分辨率靶蛋白结构中的每一个;运行与一个或多个已知药物不良反应(ADR)相关的一个或多个分类器模型;使用每个所述分类器模型,根据涉及所述药物和已知药物-ADR关系的所述药物-靶标相互作用特征,预测一个或多个ADR;并生成指示所述预测的一个或多个ADR的输出。

在另一方面,提供了一种用于执行操作的计算机程序产品。该计算机程序产品包括存

储介质,所述存储介质可由处理电路读取并且存储由所述处理电路运行的用于运行方法的指令。所述方法与上面列出的相同。

附图说明

现在将参考附图仅以示例的方式描述本发明的实施例,其中:

图1大体描绘了在一个实施例中实现用于预测关于用于ADR的相关药物靶标和机制的假设的方法的系统框架100;

图2A是这种特征数据矩阵的示例可视化,其包括作为行的所述药物,作为列的所述靶蛋白以及作为特征的所述计算结合得分;

图2B是这种二元标签矩阵的示例可视化,其包括作为行的药物和作为列的ADR标签;

图3概念性地描绘了根据一个实施例的用于一般地预测ADR并确定未知或新药物结构的潜在ADR机制的所述方法;

图4示出了根据一个实施例的用于确定新的或现有的药物分子的靶结合预测和ADR的示例性方法;

图5示出了示例性计算机系统界面显示,其描绘了用于根据本发明的所述方法进行处理的未知或新药物分子的所述输入;

图6A示出了针对特定示例的痤疮样皮炎ADR以其各自的置信度预测的前三(3)种药物的生成列表;

图6B显示了指示针对莫米松(Mometasone)的最主要的预测结合蛋白的表格;

图7示出了进一步的分析步骤700,其可以用于产生关于第一病例研究示例的痤疮样皮炎ADR的原因的假设;

图8描绘了排名高级的蛋白质的一个例子,根据所述已开发的ADR模型,可以从该蛋白质确定糖皮质激素受体是第二大贡献因素;

图9示出了可用于产生关于第二病例研究示例的白内障子囊ADR的原因的假设的进一步分析步骤;

图10显示了针对示例性第一病例研究,药物莫米松与已知蛋白的孤核受体γ(the orphan nuclear receptor gamma)(RORγt)配体结合域之间的预测结合构象;

图11示意性地示出了示例性计算机系统/计算设备,其可用于实现本发明的实施例;和

图12示出了根据本发明的另一示例性系统。

详细说明

一种用于根据药物分子的结构输入来预测药物不良反应(ADR)的系统、方法和计算机程序产品。该系统和方法通过突出可能在引起ADR中起关键作用的相关结合靶标来进一步产生假设。更具体地,提供了一种系统框架,用于实施用于自动生成与所述药物的三维结构相关联的相互作用得分并符合结构库中的此得分的方法。

图1示出了由计算机系统运行的方法100的概述,该方法用于从代表新药化合物结构的数据中预测ADR。最初,计算机系统(例如图11中所示的系统)首先获取代表药物分子的数据和代表多种蛋白质结构的数据,并运行分子对接程序以生成药物-靶标相互作用特征,即分子对接分数。在一个实施方案中,该方法包括从数据库中提取药物分子的二维或三维结构,所述数据库例如可商购的DrugBank Version 5.0数据库资源102(例如,可从

www.drugbank.ca获得)。众所周知,DrugBank资源102将详细的药物(即化学、药理学和药物)数据与全面的药物靶标(即序列、结构和途径)结合起来。一个实施方案中,为了获得药物组或药物库104,所述计算机系统收集用于编码DrugBank 5.0中所有小分子的分子结构的SMILES(简化分子输入线输入系统,Simplified Molecular-Input Line-Entry System)符号。

在另一个实施例中,对于所述药物组104中的药物分子,所述计算机系统可以访问用于基于输入化学方程式或代表二维分子的图形来生成相关三维分子结构的工具,例如通过由Marvin Beans(例如,可从ChemAxon Marvin Beans 6.0.1获得)的程序工具"MolConverter"生成的界面使用"molconvert"命令行。在一个实施例中,所述Marvin Beans是用于化学绘图和可视化的应用程序和API,以及用于在各中二维和三维文件格式(例如分子文件格式、图形格式等)之间转换文件的Molconverter工具。

进一步地,在一个实施例中,对于所述药物组104中的三维药物分子,所述系统可以首先除去没有可旋转键(例如乙酸钙)或太大(具有大于1200的分子量,例如,苯磺酸顺阿曲库铵(cisatracurium besylate))的药物分子。因为它们可能不会产生有意义的对接(docking)得分,例如太大而无法放入蛋白质袋中。

如图1进一步所示,所述计算机系统还获得表示多种蛋白质结构的数据。为了讨论的目的,使用人蛋白质但是本发明可以适用于其他动物蛋白质类型。对于蛋白质集合,系统收集PDBBind数据库资源112(例如,可从www.pdbbind.org.cn获得)或类似蛋白质数据库的一般集合,其是晶体结构的选择来源。选择人蛋白质114,并且针对每种蛋白质仅选择具有最佳分辨率的唯一结构。经由计算机系统的接口,用户可以通过经由例如到PDBBind数据库资源112的接口输入选择特定蛋白质:根据分辨率、PD、唯一选择和PDBBind标准。

在一个实施方案中,从所述PDBBind数据库112提取的是代表独特的人蛋白质靶标的数据。根据选择的标准从PDBBind数据库112中选择靶蛋白:(1)高质量:提取的所有蛋白结构均应具有1.98 ± 0.47Å量级的高分辨率;(2)可靶向的:所述结构具有可用的实验性配体结合数据;(3)独特的人类蛋白质:这些结构代表独特的人类蛋白质,即,对于一种蛋白质,选择可获得的高级分辨率的晶体结构之一;(4)定义明确的结合包:结构具有嵌入的配体以定义结合袋(binding pockets)。

在选择和提取出药物分子集104和独特的靶蛋白集114之后,该方法使用自动对接工具(例如AutoDock Tools 1.5.6)(例如,可从autodock.scripps.edu获得)准备结构文件。在一个实施方案中,使用AutoDock工具的制备脚本将Gasteiger电荷加到所述药物和靶结构两者中。众所周知,AutoDock工具是配置为准备文件的软件程序,所述文件可用来预测小分子(例如基材或候选药物)如何与已知三维(例如靶蛋白)结构的受体结合。在一个实施方案中,蛋白质的结合袋以原始嵌入的配体为中心,固定大小为$25 \times 25 \times 25$Å^3,以减少基于口袋的变化。

继续图1的所述方法100,在107处所述方法包括使用带有固定的随机种子和其他默认参数的AutoDock Vina 1.1.2研究工具(例如,可从vina.scripps.edu获得)将每种药物分子从集104向蛋白质集114的每个蛋白质结构对接。众所周知,AutoDock Vina是用于执行分子对接的软件程序,它提供了高度准确的结合模式预测,即计算分子对接得分107(或分子

结合得分)及它们之间的构象。在一个实施例中,对于其输入和输出,AutoDock Vina使用被AutoDock工具和AutoDock4使用的相同的PDBQT(蛋白质数据库、部分电荷(Q)和原子类型(T)格式)分子结构文件格式。所需要的只是对接的分子结构以及包括结合位点在内的搜索空间的规格。提取最低的对接分数和相应的结合构象,并将其存储为药物-靶标相互作用特征集117。

在导致所述对接得分的产生的图1的方法步骤的基础上,在一个实施例中,收集了特征数据矩阵。图2A是这样的特征数据矩阵150(二维矩阵)的示例可视化,其包括作为行的所述药物104、作为列的所述靶蛋白114以及作为特征的相互作用的药物/靶蛋白的各个计算结合得分107以形成药物-靶标相互作用特征集117。

返回图1,在并行(同步)或后续过程中,所述方法100从SIDER(副作用资源)数据库122执行收集数据,例如包含从药品标签中提取不良药物反应(ADR)信息的SIDER数据库Version4.1,作为一组ADR标签127(可以在http://sideeffects.embl.de上找到)的基本事实。在一个实施方案中,该方法使用DrugBank同义词将药物名称从SIDER数据库映射到DrugBank ID。因此,收集了从SIDER数据库已知的现有药物-ADR关系。

在一个实施例中,基于导致产生ADR标签127的图1的所述方法步骤,收集到代表第二二进制标签矩阵的数据。图2B是这样的二进制标签矩阵160的示例可视化,其包括作为行的药物104和作为列的ADR标签127。对于每个ADR,如果药物已知会引起ADR,则将药物-ADR配对标签128标记为二进制值,例如“1”(阳性),表示所述药物引起ADR;否则,所述药物-ADR配对标签128被标记为“0”(阴性)二进制值,意味着所述药物与所述ADR之间没有关系。

在一个实施例中,该方法可以首先包括过滤步骤,以过滤含有少于预定量的阳性药物(例如五种阳性药物)的ADR,因为它们的阳性样品太少。

返回图1,在后续过程中,计算机实现的方法包括开发和评估机器学习模型130,该模型可用于基于药物-靶标相互作用特征和已知药物-ADR关系预测新药的ADR。也就是说,将第一收集的特征矩阵150和第二收集的二进制标签矩阵160(图2A,2B的)视为训练数据集,所述方法100定义了机器学习问题:Y=f(X)使得特征(Xs):是对接分数,标签(Ys):是否引起ADR。对于每个ADR,开发了一个相应的预测模型,尤其是,使用所述蛋白质结合得分作为特征,为每个ADR开发了一个具有L2正则化的逻辑回归分类器。在一个实施例中,所述分类器可以在具有sklearn Version 0.17.1的Python 2.7.12(例如,Anaconda®4.1.1软件)来实现(Anaconda®是Continuum Analytics公司得克萨斯州奥斯汀78701的注册商标)。

在一个实施例中,为每个ADR生成一个逻辑分类器模型。在一个实施例中,训练ADR模型包括:对于特定的ADR,一次获得一个ADR列,例如图2B中的列118,其具有代表所述标签(Ys)的所述二进制值;并获得整个特征矩阵f(X)例如图2A中所示的所述药物相互作用特征矩阵150。为了建立所述分类器,对于每个ADR,存在与所述一个标签列118(图2B)相对应的输入数据,并且对于每个药物样本108(一个或多个行104的)的每个输入,分别对应于多个特征(分子对接分数),例如图2A中的第114列。第104行有多个药物样本。

在一个实施例中,对于特定的ADR模型,这些输入在一个逻辑回归函数中被接收,例如:

$$f(x)=\frac{1}{1+e^{-(a+b_1x_1+b_2x_2+\cdots+b_{600}x_{600})}}$$

在给定药物x的情况下,针对600种蛋白质的分子对接分数是(x_1,x_2,…,x_{600})的向量。

在模型训练过程中获得了系数

$(b_1, b_2, \ldots, b_{600})$

以及常数值α。该方法包括计算f(x)作为药物x可能引起此特定ADR的预测置信度得分(范围:0%至100%)。

在一个实施方案中,Anaconda®Python的所述sklearn包可以在计算机系统上实现开发逻辑回归模型,并且在一个实施例中,系数通过最小化成本函数(其是预测和实际值之间的汇总差(aggregated difference))来确定。使用L2正则化可以得出具有最佳预测性能的系数。用于Python编程语言的Scikit-learn软件机器学习库也可以用于开发所述ADR模型。

在一个实施例中,使用所述机器学习数学技术在逻辑回归ADR模型构造中计算出的系数依赖于用来了解ADR机制相关的靶标分析。

在一个实施例中,为了选择模型的最佳参数,经过十倍交叉验证,正则化类型(L1和L2)和参数(C=0.001、0.01、0.1、1、10、100和1000)的不同组合可以探索验证,并且可以基于接收机工作特性曲线(AUROC)下的最佳区域来选择最佳参数。为了证明所述分子对接的ADR预测性能,针对训练集中的所述药物生成了七种不同类型的结构指纹以进行特征比较。所述七个结构指纹是E状态、扩展连接指纹(ECFP)-6、功能类指纹(FCFP)-6、FP4、Klekota-Roth方法、MACCS和PubChem结构描述符(被称为E状态、ECFP6、FCFP6、FP4、KR、MACCS和PubChem)。在通过精确调用曲线(AUPR)值下的AUROC和面积上的十倍交叉验证比较了分子对接与这些结构指纹的所述预测性能之后,基于具有所述最佳参数的分子对接特征开发了所述最终模型130。

应该理解,可以开发不同类型的预测模型来预测ADR。例如,虽然如上所述为每个ADR构建了一个单独的模型,但也可能仅开发了一个可以预测所有ADR的模型。对于这种替代方法,需要收集ADR的功能,以便训练集中的每一行都代表一个药物-ADR配对,并且它同时包含药物和ADR特征。该行的标签为阳性(代表已知药物-ADR关联)或阴性(代表未知药物-ADR关联)。

如图1的133进一步所示,开发的模型随后可用于对训练集中尚不存在的药物进行ADR预测。此外,在135处,通过分析与ADR预测相关联的蛋白质结合特征,例如,在排名靠前的对接得分和校正方面,可以确定所述ADR的所述可能机制。

图3在概念上描绘了根据一个实施例的用于总体上预测ADR并确定输入到所述系统的未知或新药物结构301(例如,药物X)的基础ADR机制的方法300。在建立训练集数据之后,包括所述药物相互作用矩阵(例如,如图2A所示)和所述ADR标签矩阵(例如,如图2B所示)的生成,以及在使用上述逻辑回归分类器中开发每个ADR机器学习模型之后,确定新药的ADR的所述方法如图3所示。最初,该方法包括:获得新药/未知药X的分子结构,其中可能包括被测试的新药的物理三维结构301。然后,将新药物结构301输入到所述AutoDock程序或类似的对接工具310,例如AutoDock Vina,其中针对所述多个独特靶蛋白304中的每一个获得所述新药的所述分子结合得分。在对接中,获得每个靶蛋白相互作用的靶分子结合得分(相互作用得分),以产生所述新药x对每个靶蛋白的对接得分的向量315。然后可以通过所述靶标对所述药物X的相互作用分数对所说靶标进行排序,以表明哪种靶蛋白与所述新药的结合效果最好。另外,可以在药物X和靶标库之间获得构象。

然后,交互结果用于通过所述机器学习模型f(x)预测ADR。此外,可以实施功能分析以了解ADR的基本机制。

因此,如图3所示,然后将构建的ADR预测模型f(x)330应用于与每个靶标(可以排序)相关的对接分数的向量315。即,基于所述药物X和所述靶标库之间的每个相互作用得分,所述应用模型基于所述相互作用得分预测药物X的潜在ADR350。

在一个实施例中,通过置信度分数对所述ADR进行排名。例如,可以将所述药物X的高级结合靶标用于研究所述药物-ADR关系的所述潜在机制。参见例如下文的第一病例研究实施例1。

备选地,可以通过基于模型的特征/系数分析来识别所述ADR的最相关靶标,以了解所述ADR的所述机制。参见例如下文的第二病例研究示例2。

图4示出了示例性方法400,其基于所述相互作用得分的结果以及潜在所述ADR的机制确定来确定新的(或现有的)药物分子,例如,训练集中不存在的药物X的靶标结合预测和ADR。

图4中,在402,在第一实施例中,首先接收药物X的三维分子结构的符号数据表示。对于现有或已知的药物结构,可以获得在402处输入到计算机系统的新药物X的分子SMILES代码表示。

在替代实施例中,如图4所示,在401,可以首先接收表示用户生成的新(候选)药物的二维分子或化学式的数据作为输入到所述系统中。一旦被接收到所述系统中,如在404处所示,所述系统调用计算机实现的程序或工具,该程序或工具用于访问分子转换工具以生成所述新的(候选)药物配方的相应的三维分子结构。这样的工具可以包括在Marvin Beans中可用的Molconverter命令行程序工具(例如,可从ChemAxon Marvin Beans 6.0.1获得)。

或者首先通过从预先存在的列表中选择并输入已知药物配方并获得相应的SMILES代码表示(如图4中的402所述),或者首先接收用户生成的一维字符串或药物X的二维结构表示并将其转换为相应的三维分子结构表示,如图4A中的404所示,然后,如图4的405所示,确定所述三维结构内的结合位置和区域。使用分子对接工具,可以在相当大的准确性范围内预测所述新药X的所述三维结构的所述小分子配体在所述靶蛋白结构的所述适当靶结合位点内的所述构象。这可以通过实现诸如AutoDock之类的程序来执行。使用用于所述输入药物配方的数据,所述系统进一步产生针对所述靶蛋白的相互作用特征,即,获得针对每种靶蛋白库的分子结合得分和确认。另外,在405处执行所述药物X-靶标相互作用的排名和可视化。然后,在图4中,在410处,该方法运行所述机器学习的ADR模型412以对所述新药X的ADR进行预测和排名。在该步骤中,可以生成输出置信度得分,其指示所述输入药物(例如,新药X)引起与ADR相关的药物-蛋白质相互作用的可能性。然后在415,进行进一步分析以确定所述高级ADR预测,并在420确定所述新药的可能原因或解释。该系统然后可以生成输出,该输出包括:所述预测的结合靶标,包括药物X的结合得分和构象;药物X的所述预测ADR以及与所述ADR相关的所述目标蛋白。

病例研究示例1

在第一示例案例研究中,确定了药物莫米松诱导痤疮样皮炎ADR。因此,使用图4的示例性方法400,首先以莫米松的分子SMILES代码输入计算机系统。然后,在405,生成与所述提取的靶蛋白库的相互作用特征,即所述分子对接分数。

图5示出了示例性计算机系统界面显示500,其描绘了用于根据本发明方法进行处理的未知药物或新药物的输入。为了说明的目的,第一示例药物502(例如,莫米松)及其从DrugBank获得的相应的SMILES作为输入505。在一个实施例中,可以通过响应于选择通过用户界面“药物列表”标签507而显示的药物列表来选择用于输入的药物。在进一步的实施例中,用户可以在所述系统中输入一维字符串或二维结构表示或与潜在新药相关的新化学式的渲染,并通过调用应用程序接口访问计算机实现的应用程序,该应用程序根据输入的分子结构的一维或二维渲染构建优化的三维分子对象的工具。在任一实施例中,在输入新药物的三维结构(例如,在505处药物莫米松的一维渲染)之后,通过选择“提交”界面按钮510将现有或新药物配方输入到AutoDock Vina程序中。AutoDock Vina程序采用构象搜索算法,并采用产生新药物502与集合中所有靶蛋白的所述相互作用515、所述结合能的定量预测的功能。在一个示例性实施方案中,产生了600个靶蛋白的相互作用得分,并且可以显示每种药物-靶蛋白相互作用得分。列出了具有相应蛋白质标识符(PDBID)515的药物520,以及由AutoDock Vina程序生成的相应药物相互作用分数530。在一实施例中,这些分数根据它们的绑定的分数530进行排名。

然后,如在图4的步骤410所述,该方法运行所述ADR模型412以预测所述新药或现有药物(例如莫米松)的ADR。

在第一说明性示例中,作为针对每种输入药物的相互作用分数530运行每个ADR模型的输出,生成了关于该药物将提供与当前ADR相关的药物-蛋白质相互作用的置信度分数。如图6A的图表600所示,产生了针对痤疮样皮炎ADR的具有各自置信度605的前三(3)种药物的列表。

众所周知,痤疮样皮炎(统一医学语言系统概念ID:C0234708)是痤疮样皮肤丘疹。如图6A所示,运行所述痤疮样皮炎ADR模型的预测结果表明,莫米松(DrugBank ID:DB00764)是导致该ADR的测试集中排名高级的药物,置信度为0.649。据报道,皮肤丘疹是由莫米松使用引起的局部不良反应,这证实了这一预测。

为了理解该ADR的潜在机制,可以进行药物X的靶标结合分析和ADR特异性特征分析。在一个实施方案中,该方法获得新药与所有靶蛋白的对接分数。对于第一个病例研究示例,调用过程来确定莫米松的顶级结合蛋白,并按其对接分数对其进行排名。图6B显示表650,其指示莫米松的最主要预测结合蛋白。如图6B所示,孤核受体γ(RORγt)的配体结合域(蛋白数据库ID,或PDB ID:3BOW)被预测为前3位的对于莫米松的结合靶652,它具有-10.4的结合分数。

图10示出在第一病例研究示例中,莫米松药物1001和所述孤核受体γ之间的预测的结合构象1000的可视化(RORγt)的配体结合域1010(例如,PDB ID:3BOW)。在图10中,在受体1010的三维结构中示出了配体1001的三维结构,其示出了配体对接到受体的结合腔1012中,从而该配体10012中的每一个与所述预测的结合构象相关联的相互作用能的精确预测被确定。所述蛋白质靶标1010的“稀粘”蛋白质残基1007显示在所述蛋白质靶标1010的结合腔1012内,并且与所述配体1001紧密相互作用。

在一个实施方案中,为了避免这种ADR相互作用,可以开发药物修饰或开发新药以最小化或避免与3BOW蛋白的结合。可替代地,可以重新设计或修改现有的药物结构以最小化或避免与3BOW蛋白的结合。这样的修饰包括本领域已知的那些,包括但不限于改变配体的长

度、大小和/或形状、改变空间构型、极性和氢键方面，例如，添加杂原子(氧，氮等)或基团，而氢键可避免与被确定为ADR根本原因的蛋白质发生相互作用。

如以上关于图1所提到的，在进一步的分析步骤135中，可以产生关于ADR的原因的假设。图7示出了进一步的分析步骤700，其可以用于生成关于第一病例研究示例的痤疮样皮炎ADR的原因的假设。在研究中，已经发现在IL-17表达细胞和Th17相关信号存在或诱发痤疮样病变705。在708处，示出的是Th17细胞分化和IL-17的生产需要RORγt。可以在710处假定通过与RORγt结合从而影响Th17/IL-17水平，所述莫米松药物702引起痤疮样皮炎712的发生。

病例研究示例2

在第二病例研究示例中，所述计算机系统执行基于模型的特征分析，即系数分析，包括分析所述ADR模型的所述特征系数并根据所述系数对所述靶标进行排名，以了解与所述ADR相关的机制。

在第二个病例研究示例中，可能确定了一种可以诱导白内障子囊的药物-ADR。因此，根据图1的进一步分析步骤133，将来自600种蛋白质特征(图2A)中的每一个的对接得分向量就针对白内障子囊ADR的标记向量(图2B)进行分析，以评估其各自的性能。

作为分析的结果，所述方法确定与受试者ADR有关的最重要的蛋白质特征，其由相应的ADR模型加权。图3示出了示例性表800，其根据针对该ADR模型的白内障子囊ADR的逻辑回归系数的绝对值来指示与该白内障子囊ADR有关的前三(3)个蛋白质特征。因此，在第二病例研究示例中，获得了所述系数(b_1,b_2,…,b_{600})的绝对值，以指示相应的蛋白质靶蛋白1-600对ADR预测(例如白内障子囊)的贡献权重。绝对值越大，表示对模型的贡献越大。

在图8的表800中所示的分析中，根据所述开发的ADR模型，确定糖皮质激素受体805是第二大贡献因素。

图9示出了可以用于生成第二病例研究示例的白内障子囊ADR912的原因的假设的进一步分析步骤900。为了了解这种ADR的潜在机制，据研究报道，类固醇诱导的继发子囊性白内障仅与具有糖皮质激素活性的类固醇有关，其中糖皮质激素受体激活905及其继发变化(细胞增殖和分化抑制等)908起到关键作用。因此，将确定与糖皮质激素受体结合的药物(例如，新药X)对于白内障子囊的发生可能是重要的。

因此，从基于特征的分析中，有可能找到与ADR相关的蛋白质靶标，从而产生有助于探索和理解ADR机制的假说。

从上述案例研究中，该方法不仅可以预测药物分子的ADR，而且可以通过结合靶标提供可能的机理解释。由于ADR复杂且因人而异，因此这种解释可能会为毒理学研究人员提供线索，从而提出假设并帮助设计有关ADR机制的湿实验室实验，从而改善药物的安全性评估。由于这些方法仅需要药物分子的结构信息来预测ADR，因此在其他类型的候选药物信息受到限制的早期药物开发阶段使用它是可行的。

图11示意性地示出了示例性计算机系统/计算装置，其适用于实现本发明的实施例。

现在参考图11，描绘了一种计算机系统框架200，该计算机系统框架200运行用于预测和产生关于相关药物靶标和药物不良反应机理的假设的方法。在一些方面，系统200可以包括计算设备、移动设备或服务器。在一些方面，计算设备200可以包括例如个人计算机、膝上型计算机、平板电脑、智能设备、智能电话、智能可穿戴设备、智能手表或任何其他类似的计

算设备。

计算系统200包括至少一个处理器252，例如用于存储操作系统和/或程序指令的存储器254、网络接口256、显示设备258、输入设备259以及任何其他公共特征到计算设备。在一些方面，计算系统200可以，例如，是被配置225或基于web或基于云的服务器220以及通过公共或专用通信网络99与数据库230的网站进行通信的任何计算设备。此外，示出系统200的一部分是另一个存储器260，用于临时存储提取的药物-靶标相互作用特征和药物-ADR信息，例如，用于建立ADR模型。例如，在一个实施例中，另外的存储器260可以提供结构库包括已鉴定药物和人类蛋白质靶标的数据库，以及通过分子对接计算出的相互作用谱。

在一个实施例中，如图11所示，设备存储器254存储程序模块，这些程序模块为系统提供了预测和生成有关药物靶标和药物不良反应机理的假设的能力。例如，药物/新药结构处理程序模块265配备有计算机可读指令、数据结构、程序组件和应用程序界面，用于与Drugbank数据库V 5.0网站进行交互，以处理和处理详细的药物(即化学药品，药理和药物数据)。靶蛋白处理模块270具有用于与PDBBind 112相互作用的计算机可读指令、数据结构、程序组件和应用界面。用于选择和加工靶蛋白的数据库网站。对接工具处理器模块275提供有计算机可读指令、数据结构、程序组件和应用程序界面，用于与AutoDock Vina对接程序进行交互以生成药物与所选靶标蛋白之间的分子对接分数。一个ADR-药物提取处理器模块280被用于与用于获得来自特定药品标签中提取的信息ADR的SIDER数据库交互设置有计算机可读指令、数据结构、程序组件和应用程序接口。机器学习工具处理器模块285具有计算机可读指令、数据结构、程序组件和应用程序接口，用于与监督的机器学习程序进行交互以生成逻辑回归ADR模型。另一个程序模块是分析管理器处理程序模块290，该模块具有计算机可读指令、数据结构、程序组件和应用程序接口，用于根据图4的步骤对新药进行ADR预测分析和假设生成。

在图11中，处理器252可包括例如微控制器、现场可编程门阵列(FPGA)或配置为执行各种操作的任何其他处理器。处理器252可以被配置为根据图1和图4的方法执行指令。这些指令可以被存储，例如，在存储器254。

在一个实施例中，计算机系统200是实现多个处理器的机器。由于分子对接过程是最耗时的过程，即每次要处理新药时，它都需要对接600种蛋白质，因此多个控制处理器单元(例如CPU 252A、252B、252C)可以通过并行计算对接过程来加快这一过程。例如，代替一个分子对接600个蛋白质的分子，一台50核的机器可以一次进行50个对接。在一个实施例中，计算机系统200可以是多核机器，由此，核数越多，计算速度就越快。对于ADR模型开发，多核将有助于加快参数测试。例如，如果需要测试10组参数，则一台10核机器可以批量进行。

存储器254可以包括例如易失性存储器形式的非暂时性计算机可读介质，诸如随机存取存储器(RAM)和/或高速缓冲存储器或其他。存储器254可以包括例如其他可移动/不可移动、易失性/非易失性存储介质。仅作为非限制性示例，存储器254可以包括便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪存)、存储器、便携式光盘只读存储器(CD-ROM)、光学存储设备、磁性存储设备或上述的任意合适组合。

网络接口256被配置为例如经由有线或无线连接向数据库网站服务器220发送数据或从数据库网站服务器220接收数据或信息。例如，网络接口256可以利用无线技术和通信协

议，例如，蓝牙、WIFI(例如，802.11a/b/g/n)、蜂窝网络(例如，CDMA、GSM，、M2M和3G/4G/4GLTE)、近场通信系统、卫星通信、通过局域网(LAN)、通过广域网(WAN)或允许计算设备200向服务器发送信息或从服务器接收信息的任何其他形式的通信220，例如，以从各个数据库中选择特定的靶蛋白结构数据或指定小分子药物结构数据。

显示设备258可以包括例如计算机监视器、电视、智能电视、集成到诸如膝上型计算机、智能电话、智能手表、虚拟现实耳机、智能可穿戴设备的个人计算设备中的显示屏、设备或任何其他向用户显示信息的机制。在一些方面，显示器258可以包括液晶显示器(LCD)、电子纸/电子墨水显示器、有机LED(OLED)显示器或其他类似的显示技术。在一些方面，显示器258可以是触敏的，并且还可以用作输入设备。

输入设备259可以包括例如键盘、鼠标、触敏显示器、小键盘、麦克风或其他类似的输入设备，或者可以单独使用或一起使用以提供功能的任何其他输入设备。具有与计算设备200进行交互的能力的用户。

在药物开发的早期阶段，制药公司可以使用该系统框架200来预测候选药物的潜在ADR并确定相关靶标。因此，他们可以选择其他预测更安全或更不可能与危险靶标结合的候选药物，以避免ADR。此外，在上市后阶段，制药公司可以使用该系统框架200来识别有关某些ADR的动作机制。通过根据框架研究相关靶标，他们可能会发现可能改变针对这些靶标的ADR敏感性的遗传突变。因此，他们可以建议具有特定基因突变的患者调整高风险药物(又称精密药物)的使用。

图12示出了根据本发明的示例计算系统。应当理解，所描绘的计算机系统仅仅是合适的处理系统的一个示例，并且无意于暗示对本发明的实施例的使用范围或功能的任何限制。例如，所示的系统可以与许多其他通用或专用计算系统环境或配置一起操作。适用于图12中所示系统的众所周知的计算系统、环境和/或配置的示例可以包括但不限于个人计算机系统、服务器计算机系统、瘦客户端、胖客户端、手持或膝上型设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络PC、小型计算机系统、大型计算机系统以及包括上述任何系统或设备的分布式云计算环境等。

在一些实施例中，可以在由计算机系统执行的，体现为存储在存储器16中的程序模块的计算机系统可执行指令的一般上下文中描述计算机系统。通常，程序模块可以包括根据本发明执行特定任务和/或实现特定输入数据和/或数据类型的例程、程序、对象、组件、逻辑、数据结构等(例如，参见图1)。

计算机系统的组件可以包括，但不限于，一个或多个处理器或处理单元12、存储器16和一个总线14可操作地耦合各种系统组件，包括存储器16到处理器12。在一些实施例中，处理器12可执行从存储器16加载的一个或多个模块10，其中程序模块体现了使处理器执行本发明的一个或多个方法实施例的软件(程序指令)。在一些实施例中，模块10可以被编程到处理器12的集成电路中，该集成电路从存储器16，存储设备18，网络24和/或其组合加载。

总线14可以代表几种类型的总线结构中的任何一种或多种，包括使用各种总线架构中的任何一种的存储器总线或存储器控制器、外围总线、加速图形端口以及处理器或本地总线。作为示例而非限制，此类体系结构包括行业标准体系结构(ISA)总线、微通道体系结构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)本地总线和外围组件互连(PCI)总线。

该计算机系统可以包括各种计算机系统可读介质。这样的介质可以是计算机系统可访问的任何可用介质，并且可以包括易失性和非易失性介质、可移动和不可移动介质。

存储器16(有时称为系统存储器)可以包括易失性存储器形式的计算机可读介质，诸如随机存取存储器(RAM)、高速缓存存储器和/或其他形式。计算机系统可以进一步包括其他可移动/不可移动、易失性/非易失性计算机系统存储介质。仅作为示例，可以提供存储系统18以用于读取和写入不可移动的非易失性磁性介质(例如，“硬盘驱动器”)。尽管未示出，但是用于从可移动非易失性磁盘(例如“软盘”)进行读取和写入的磁盘驱动器，以及用于从可移动非易失性光盘进行读取或写入的光盘驱动器可以提供CD-ROM，DVD-ROM或其他光学介质之类的磁盘。在这种情况下，每个都可以通过一个或多个数据介质接口连接到总线14。

该计算机系统还可以与一个或多个外部设备26通信，例如键盘、指示设备、显示器28等；使用户能够与计算机系统交互的一个或多个设备；和/或使计算机系统能够与一个或多个其他计算设备进行通信的任何设备(例如，网卡，调制解调器等)。这种通信可以通过输入/输出(I/O)接口20发生。

仍然，计算机系统可以经由网络与一个或多个网络24通信，例如局域网(LAN)、通用广域网(WAN)和/或公共网络(例如，因特网)、适配器22。如图所示，网络适配器22通过总线14与计算机系统的其他组件通信。应当理解，尽管未示出，但是其他硬件和/或软件组件也可以与计算机系统结合使用。示例包括但不限于：微代码、设备驱动程序、冗余处理单元、外部磁盘驱动器阵列、RAID系统、磁带驱动器和数据档案存储系统等。

本发明可以是处于任何可能的技术细节集成水平的系统，方法和/或计算机程序产品。该计算机程序产品可以包括其上具有用于使处理器执行本发明的方面的计算机可读程序指令的计算机可读存储介质。

计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如，通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0080] 在此参考根据本发明实施例的方法，装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各方面。将理解，流程图图示和/或框图的每个框以及流程图图示

和/或框图中的框的组合可以由计算机可读程序指令来实现。

用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路配置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等，以及过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)，该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

可以将这些计算机可读程序指令提供给通用计算机，专用计算机或其他可编程数据处理设备的处理器，以产生机器，使得该指令经由计算机或其他处理器执行可编程数据处理设备，创建用于实现流程图和/或框图方框中指定的功能/动作的装置。这些计算机可读程序指令也可以存储在计算机可读存储介质中，该计算机可读存储介质可以指导计算机，可编程数据处理装置和/或其他设备以特定方式起作用，从而使得其中存储有指令的计算机可读存储介质。包括制品的制品，该制品包括用于实现流程图和/或框图方框中指定的功能/动作的方面的指令。

这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它

们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

这里使用的术语仅出于描述特定实施例的目的，并且不旨在限制本发明。如本文所使用的，单数形式的“一”、“一个”和“该”也意图包括复数形式，除非上下文另外明确指出。将进一步理解的是，当在本说明书中使用术语“包括”和/或“包括”时，指定存在所述特征、整数、步骤、操作、元件和/或组件，但是不排除存在或一个或多个其他特征、整数、步骤、操作、元素、组件和/或其组的添加。所附权利要求中的所有元件的相应结构，材料，作用和等同物旨在包括用于与如特别要求保护的其他要求保护的元件组合地执行功能的任何结构、材料或作用。已经出于说明和描述的目的给出了本发明的描述，但并不意图是穷举的或将本发明限制为所公开的形式。在不脱离本发明范围的情况下，许多修改和变化对本领域普通技术人员将是显而易见的。选择和描述实施例是为了最好地解释本发明的原理和实际应用，并使本领域的其他普通技术人员能够理解本发明的各种实施例，这些实施例具有适合于预期的特定用途的各种修改。

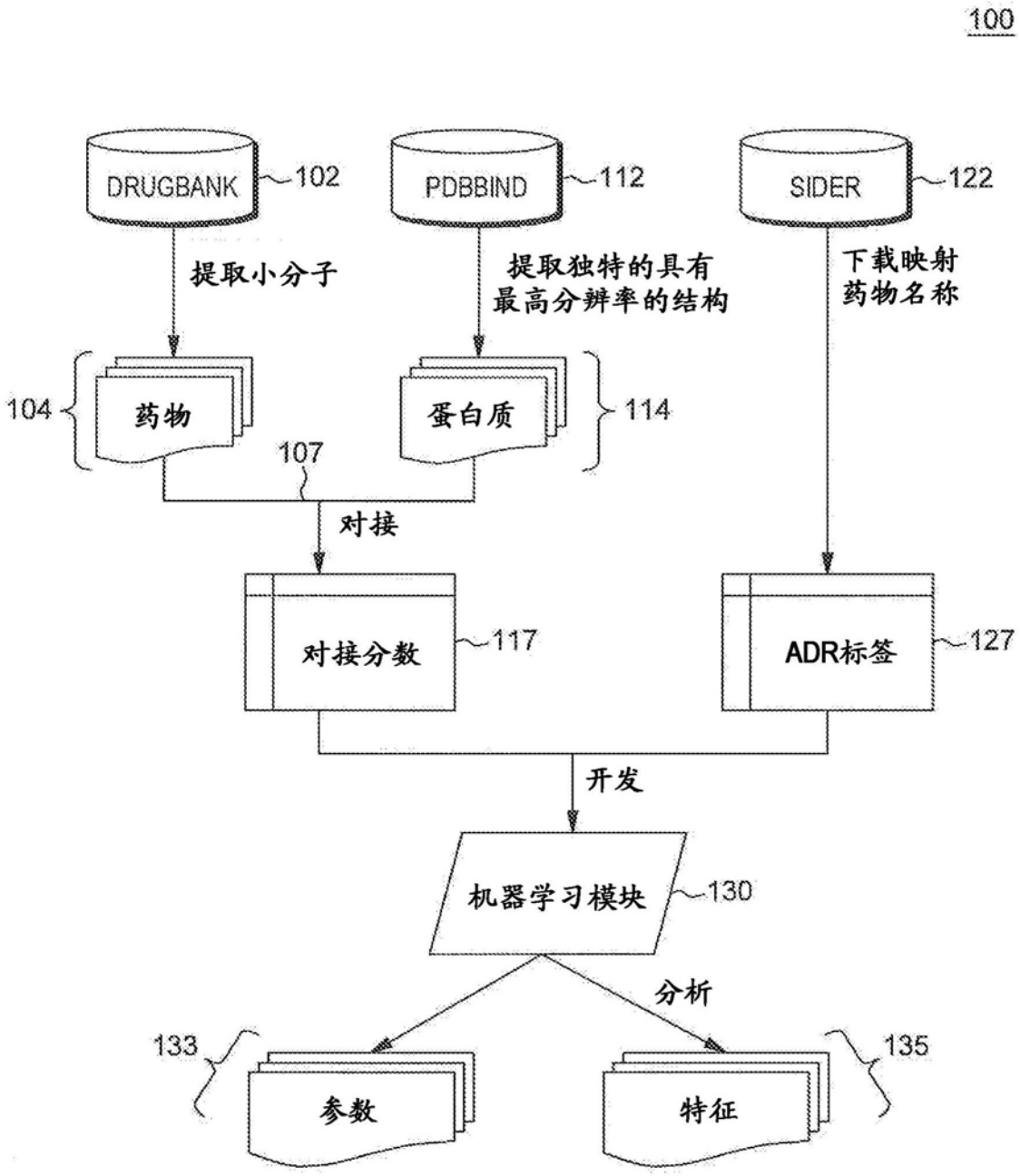

100
DRUGBANK
102
PDBBIND
112
SIDER
122
提取小分子
提取独特的具有
最高分辨率的结构
下载映射
药物名称
104
药物
蛋白质
114
107
对接
对接分数
117
ADR标签
127
开发
机器学习模块
130
分析
133
参数
特征
135

图1

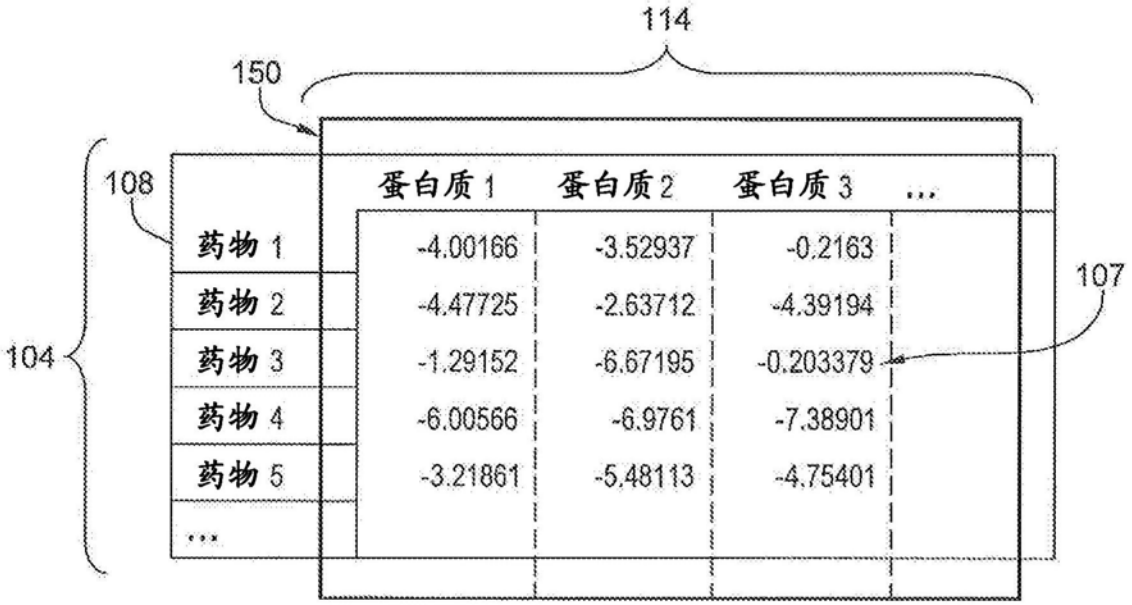

114
150
108
104
107
蛋白质 1
蛋白质 2
蛋白质 3
...
药物 1
药物 2
药物 3
药物 4
药物 5
...
-4.00166
-3.52937
-0.2163
-4.47725
-2.63712
-4.39194
-1.29152
-6.67195
-0.203379
-6.00566
-6.9761
-7.38901
-3.21861
-5.48113
-4.75401

图2A

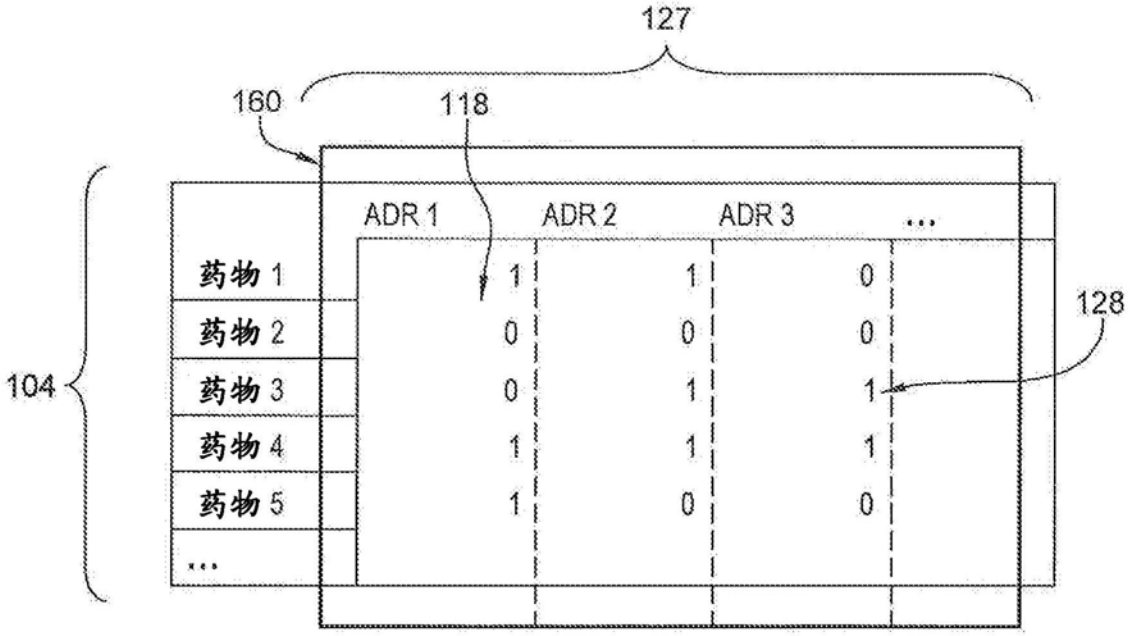

127
160
118
104
128
ADR 1
ADR 2
ADR 3
...
药物 1
药物 2
药物 3
药物 4
药物 5
...
1 1 0
0 0 0
0 1 1
1 1 1
1 0 0

图2B

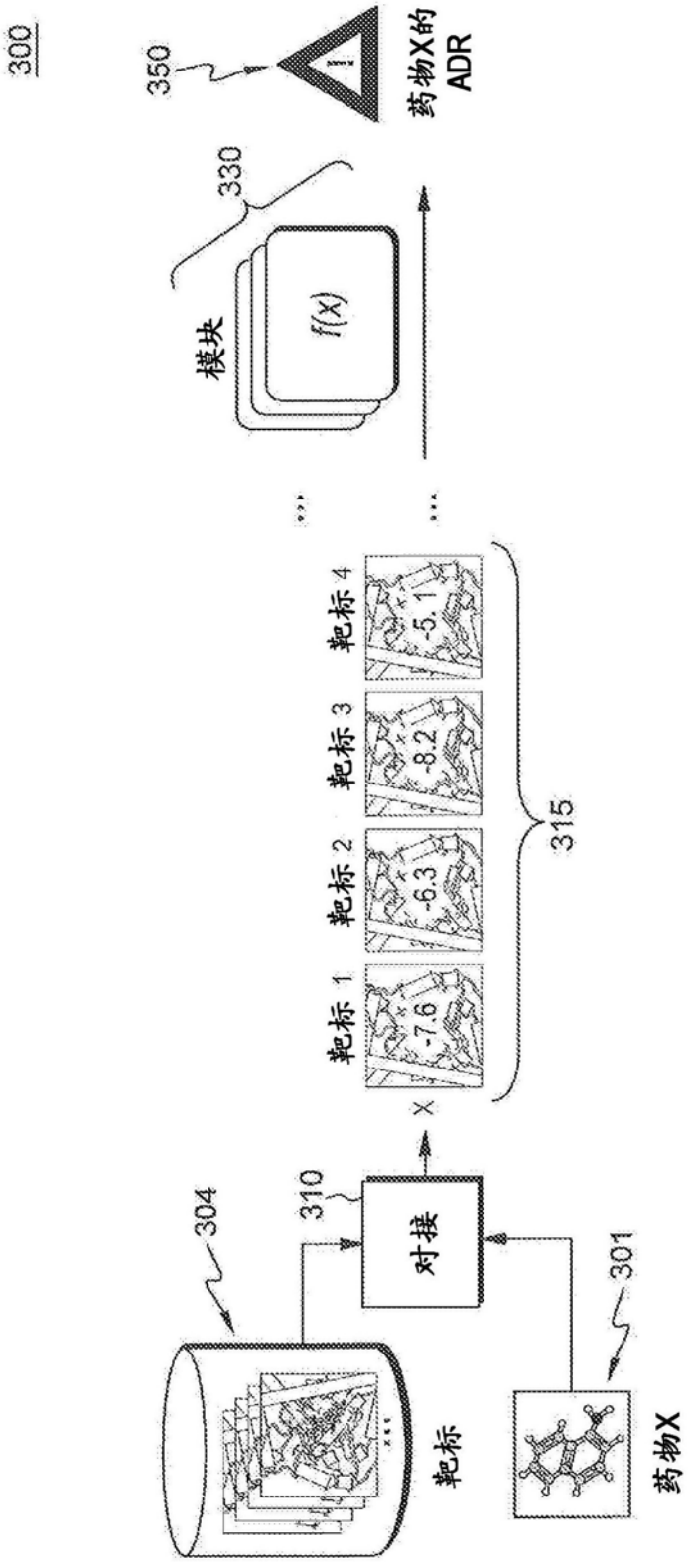

300
350
药物X的ADR
330
模块
f(x)
靶标 1
-7.6
靶标 2
-6.3
靶标 3
-8.2
靶标 4
-5.1
315
X
310
对接
304
靶标
301
药物X

图3

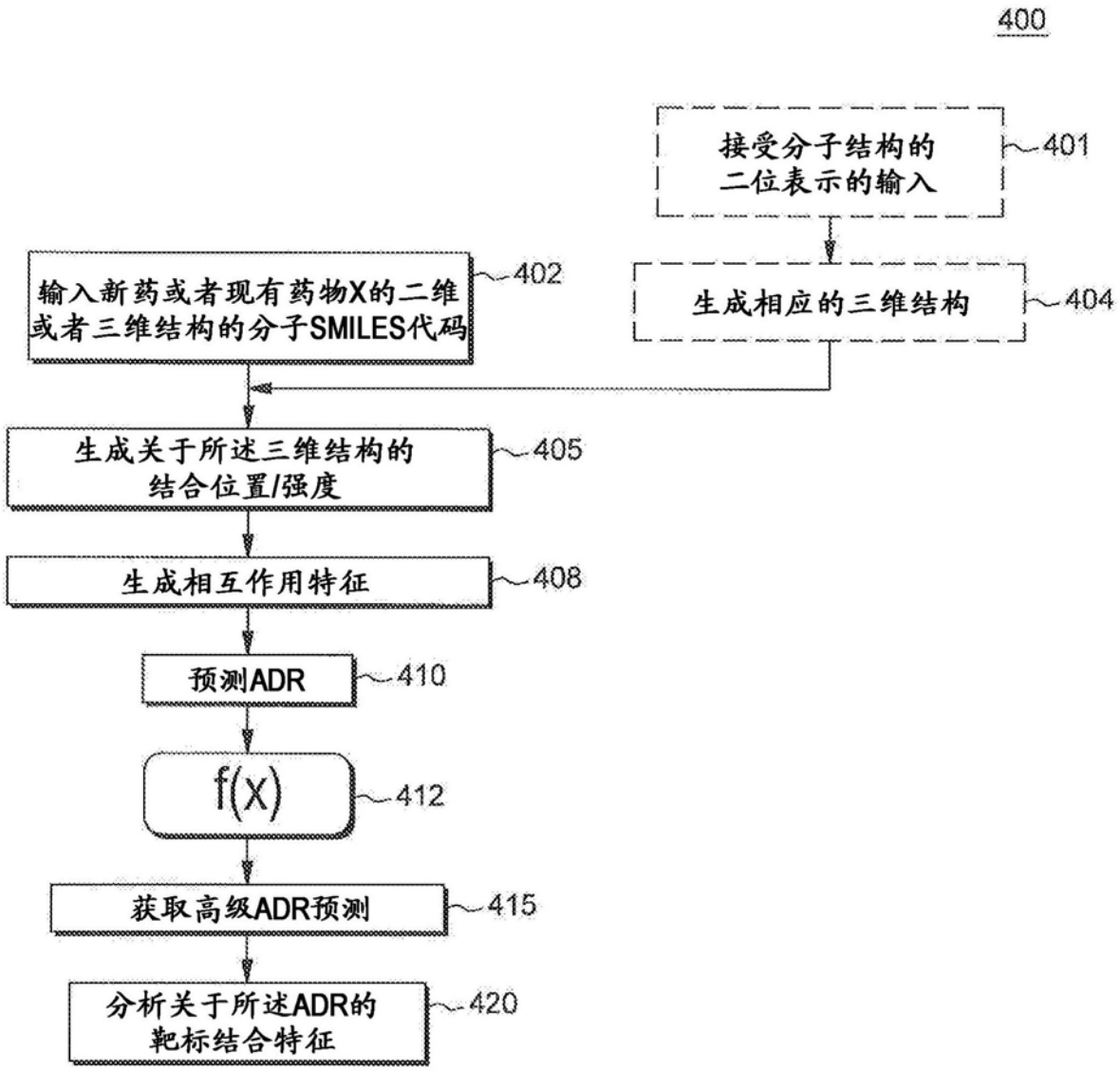

400
接受分子结构的
二位表示的输入
401
输入新药或者现有药物X的二维
或者三维结构的分子SMILES代码
402
生成相应的三维结构
404
生成关于所述三维结构的
结合位置/强度
405
生成相互作用特征
408
预测ADR
410
f(X)
412
获取高级ADR预测
415
分析关于所述ADR的
靶标结合特征
420

图4

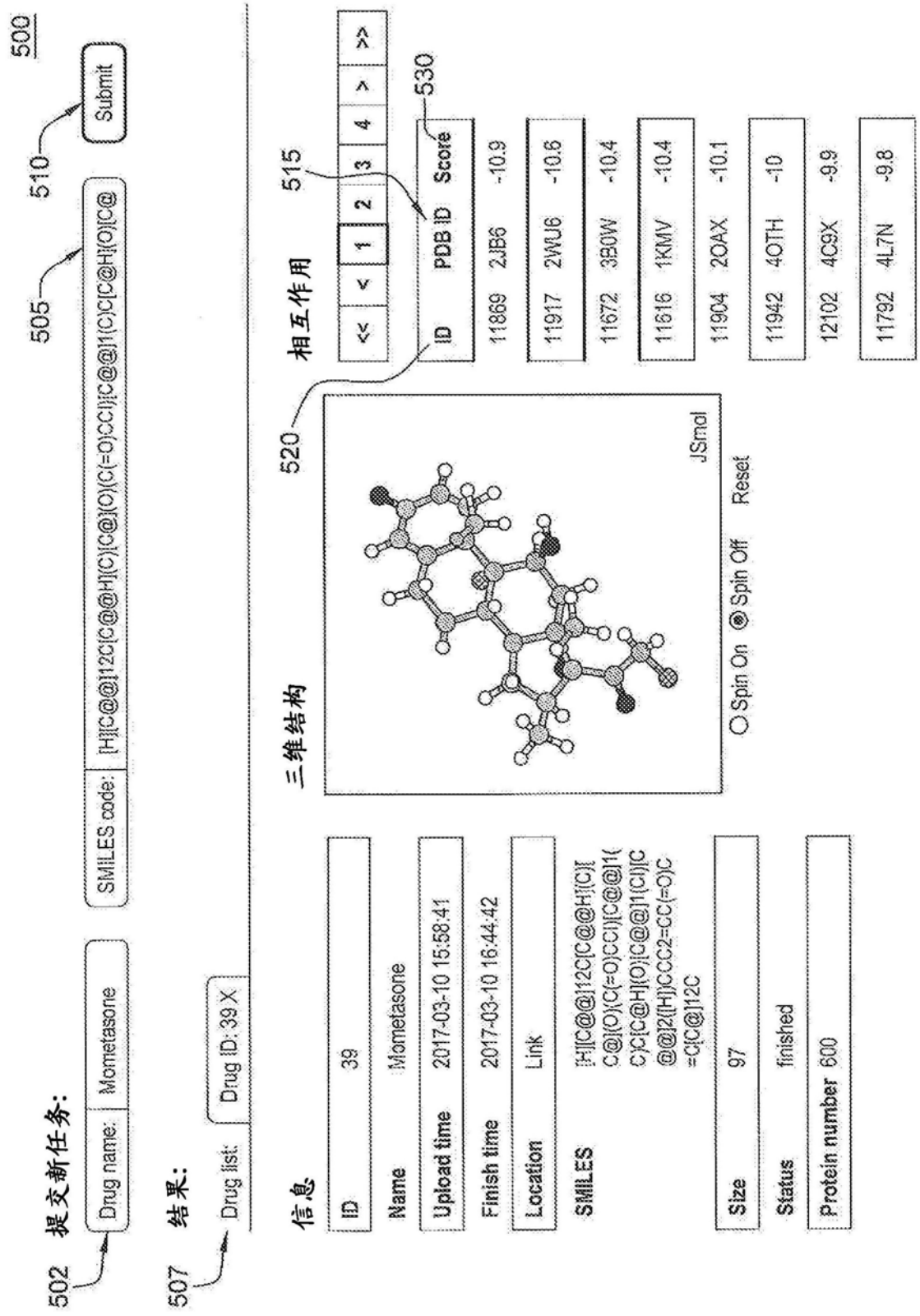

500
提交新任务:
502
Drug name:
Mometasone
SMILES code:
[H][C@@]12C[C@@H](C)[C@](O)(C(=O)CC)[C@@]1(C)C[C@H](O)[C@
505
510
Submit
结果:
507
Drug list:
Drug ID: 39 X
信息
ID 39
Name Mometasone
Upload time 2017-03-10 15:58:41
Finish time 2017-03-10 16:44:42
Location Link
SMILES [H][C@@]12C[C@@H](C)[C@](O)(C(=O)CC)[C@@]1(C)C[C@H](O)[C@@]2([H])CCC2=CC(=O)C=C[C@]12C
Size 97
Status finished
Protein number 600
三维结构
520
JSmol
Spin On
Spin Off
Reset
相互作用
515
<< < 1 2 3 4 > >>
530
ID PDB ID Score
11869 2JB6 -10.9
11917 2WU6 -10.6
11672 3B0W -10.4
11616 1KMV -10.4
11904 2OAX -10.1
11942 4OTH -10
12102 4C9X -9.9
11792 4L7N -9.8

图5

600

药物名称	置信度
Mometasone	0.649
Ixabepilone	0.587
Vinblastine	0.558

605

图6A

650

PDB ID	蛋白质	结合分数
2JB6	Fab Fragment Mor03268	-10.9
2WU6	CDC2-like kinase isoforms 3 (CLK3)	-10.6
3BOW	孤核受体γ（RORγt）的配体结合域	-10.4

652

图6B

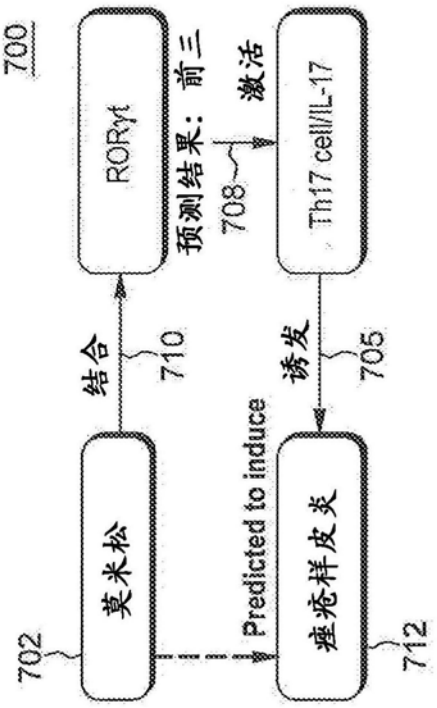

700
ROR γt
预测结果：前三
708
激活
Th17 cell/IL-17
结合
710
诱发
705
Predicted to induce
莫米松
702
痤疮样皮炎
712

图7

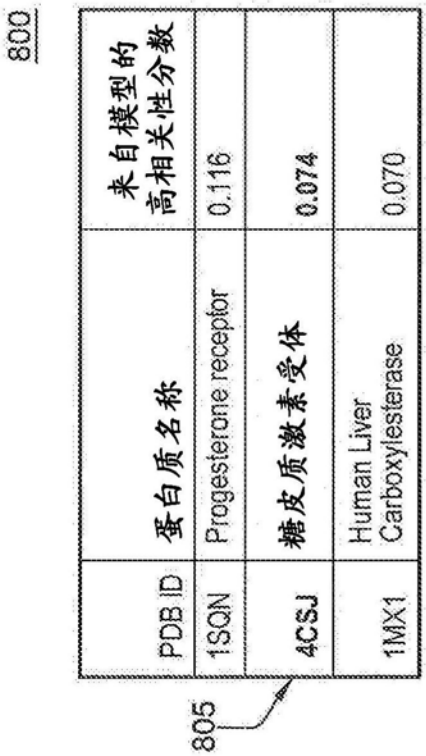

800
PDB ID
蛋白质名称
来自模型的高相关性分数
1SQN
Progesterone receptor
0.116
4CSJ
糖皮质激素受体
0.074
1MX1
Human Liver Carboxylesterase
0.070
805

图8

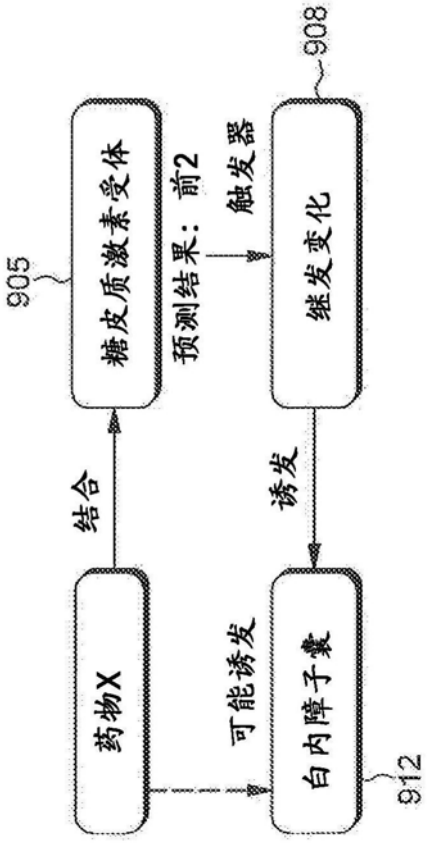


图9

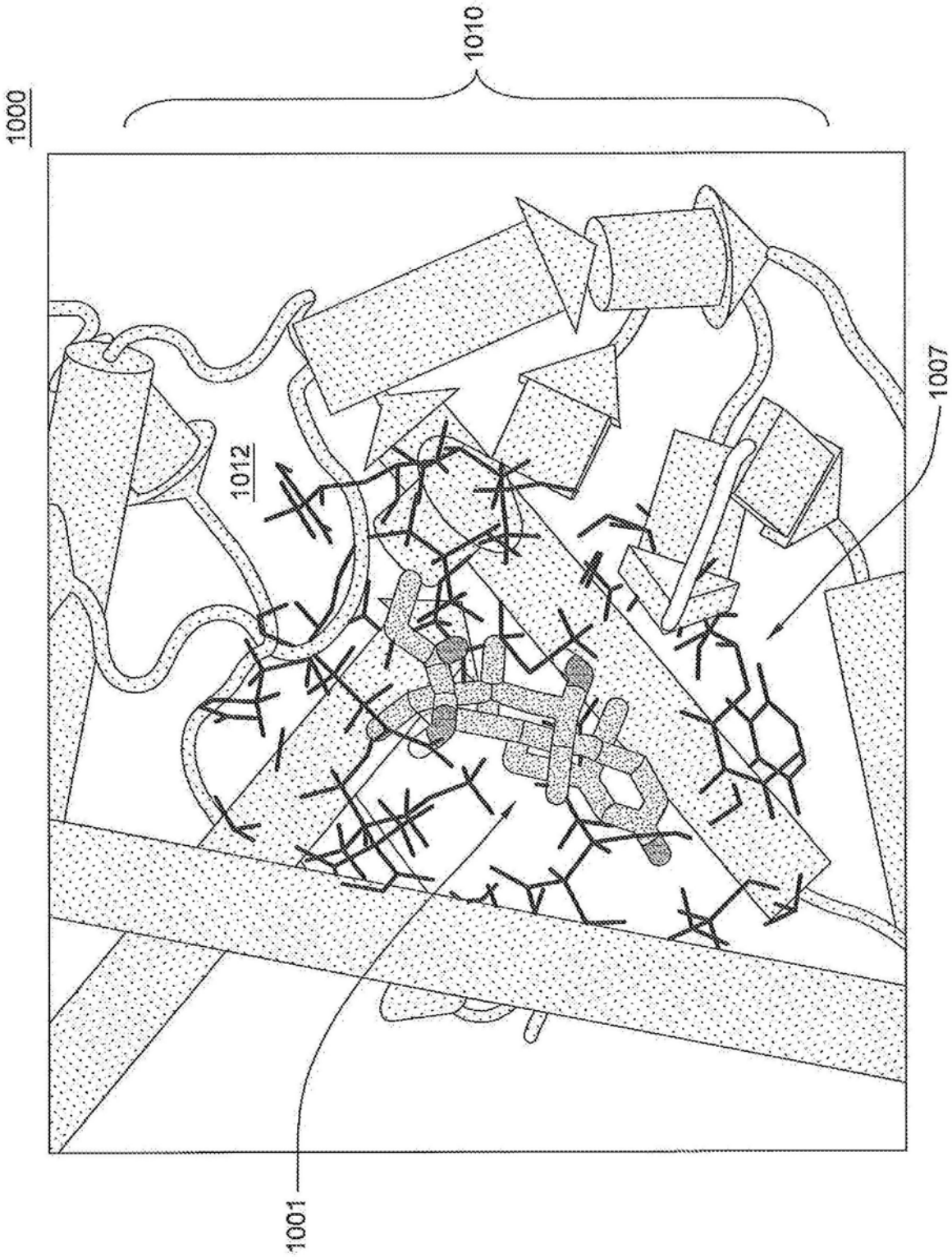


图10

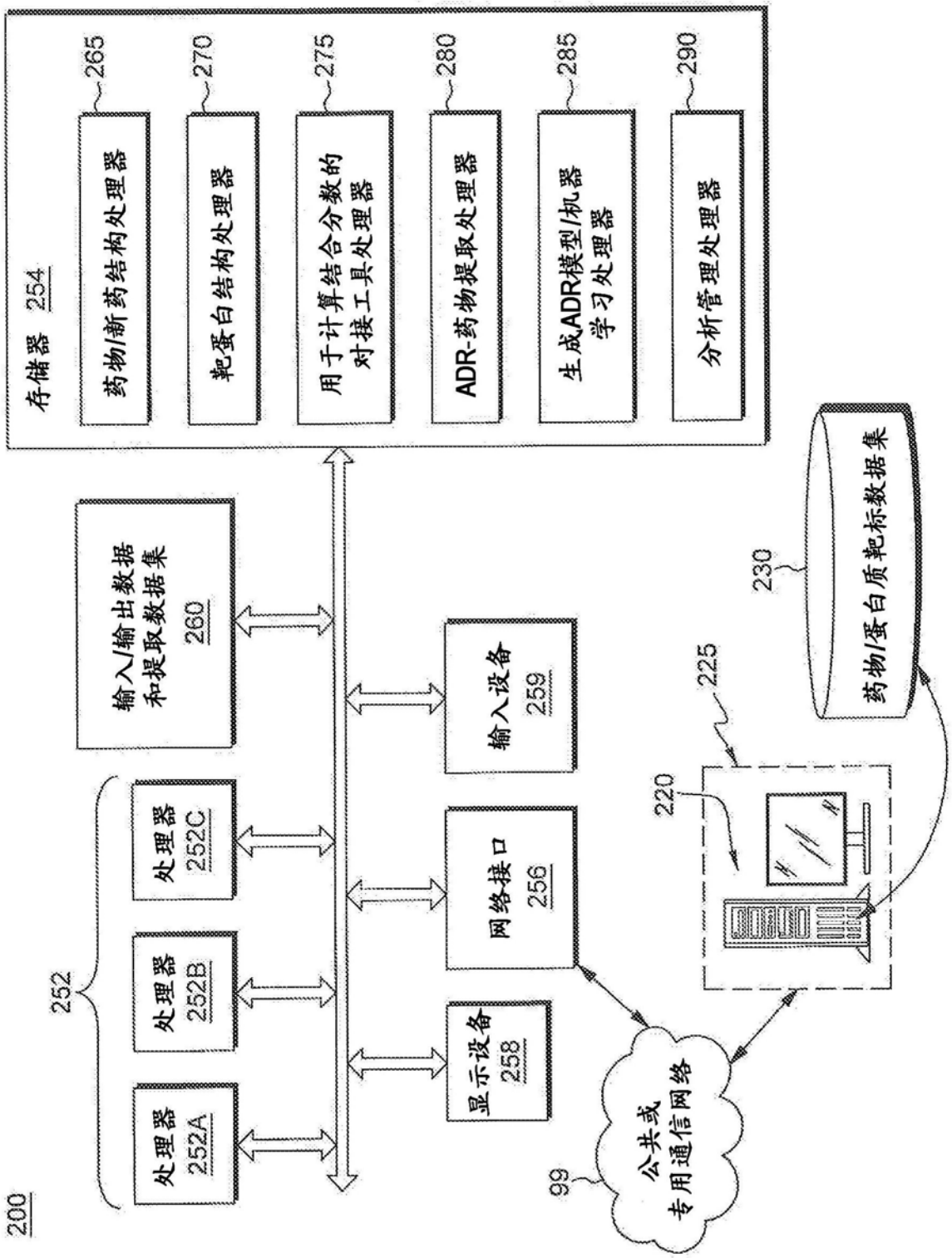

200
252
处理器 252A
处理器 252B
处理器 252C
显示设备 258
网络接口 256
输入设备 259
输入/输出数据和提取数据集 260
存储器 254
药物/新药结构处理器 265
靶蛋白结构处理器 270
用于计算结合分数的对接工具处理器 275
ADR-药物提取处理器 280
生成ADR模型/机器学习处理器 285
分析管理处理器 290
99
公共或专用通信网络
220
225
230
药物/蛋白质靶标数据集

图11

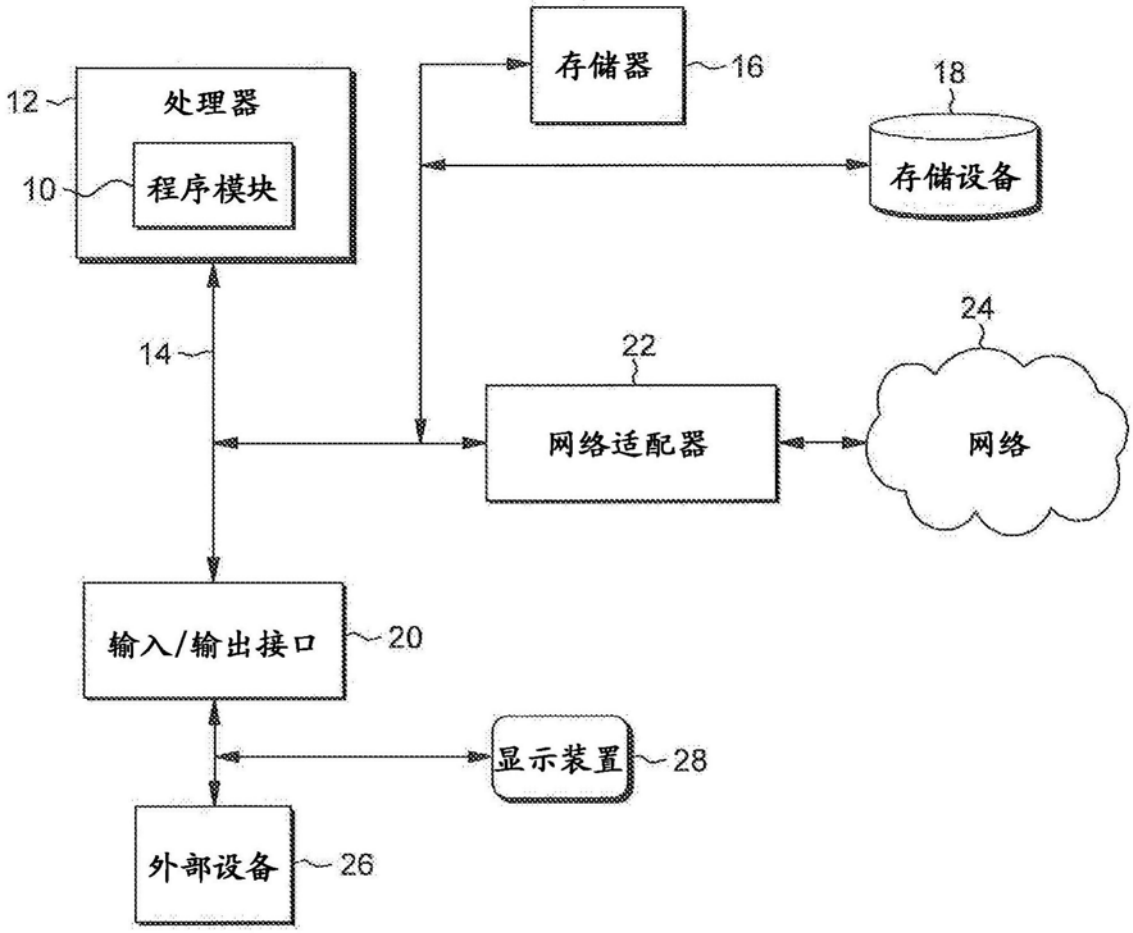

存储器
16
18
12
处理器
存储设备
10
程序模块
24
14
22
网络适配器
网络
输入/输出接口
20
显示装置
28
外部设备
26

图12