(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116508017 A
(43)申请公布日 2023.07.28

(21)申请号 202180074273.9
(22)申请日 2021.10.29
(30)优先权数据
2017177.3 2020.10.29 GB
(85)PCT国际申请进入国家阶段日
2023.04.28
(86)PCT国际申请的申请数据
PCT/GB2021/052813 2021.10.29
(87)PCT国际申请的公布数据
WO2022/096861 EN 2022.05.12
(71)申请人 艾克斯赛安西娅人工智能有限公司
地址 英国邓迪市
(72)发明人 丹尼尔·詹姆斯·克劳瑟
大卫·纳加内斯·卡隆
吉列尔莫·塞拉诺·纳赫拉
(74)专利代理机构 北京高沃律师事务所 11569
专利代理师 赵晓琳
(51)Int.Cl.
G06F 16/35(2006.01)

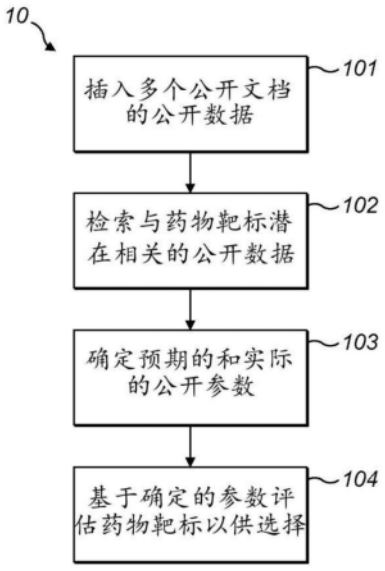
权利要求书5页 说明书21页 附图9页

(54)发明名称

计算药物靶标选择

(57)摘要

本发明提供了一种用于计算药物靶标选择的方法。所述方法包括从至少一个公开数据源获取公开数据，并链接至包括历史公开文档和当前公开文档在内的多个公开文档。所述方法包括搜索公开数据，以为每个公开文档提供指示，说明各公开文档是否与一个或多个药物靶标相关。所述方法包括基于搜索的历史公开文档的公开数据确定一个或多个药物靶标的预期公开参数，并基于搜索的当前公开文档的公开数据确定所述一个或多个药物靶标中每个的实际公开参数；以及，基于相对于预期公开参数的实际公开参数评估一种或多种药物靶标，以供选择。



CN 116508017 A

1.计算药物靶标选择的方法,包括:

从至少一个公开数据源获取公开数据,并链接至包括历史公开文档和当前公开文档的多个公开文档;

检索公开数据,为每个公开文档提供指示,说明公开文档是否分别与一个或多个药物靶标相关;

基于历史公开文档中检索的公开数据确定一个或多个药物靶标的预期公开参数,并基于当前公开文档中检索的公开数据确定每个药物靶标的实际公开参数;和,

基于相对于预期公开参数的实际公开参数,评估所述一个或多个药物靶标中的每一个以进行选择。

2.根据权利要求1所述的方法,包括对每个药物靶标,一个或多个定义的指代所述药物靶标的字符表达式,其中所述检索公开数据包括检索每个药物靶标的一个或多个字符表达式的公开数据。

3.根据权利要求2所述的方法,包括对于每个药物靶标,将所述一个或多个字符表达式中的每一个分类为安全字符表达式或不安全字符表达式,其中所述分类基于公开数据中所述字符表达式的实例指代所述药物靶标的可能性。

4.根据权利要求3所述的方法,其中,如果从其中一个公开文档中检索的公开数据包含安全字符表达式,则确定所述公开文档与所述药物靶标相关。

5.根据权利要求3或权利要求4所述的方法,其中一个或多个字符表达式被用户定义为安全字符表达式。

6.根据权利要求2至5任一项所述的方法,其中一个或多个字符表达式不安全特性被用户定义为指示相应字符表达式不安全,并且检索的公开数据中显示一个或多个字符表达式不安全特性的字符表达式被分类为不安全字符表达式。

7.根据权利要求6所述的方法,其中所述一个或多个用户定义的字符表达式不安全特性包括一个或多个:

与特定自然语言中的单词相对应的字符表达式;

具有小于规定字符数的字符表达式,任选地,其中规定的字符数为3;和,

定义为至少指代两个不同药物靶标的字符表达式。

8.根据权利要求2至7任一项所述的方法,其中一个或多个字符表达式模糊度特性被定义为将赋予一个或多个字符表达式模糊度得分,并且其中基于相应赋予的模糊度得分,每个所述字符表达式被分类为安全字符表达式或不安全字符表达式。

9.根据权利要求8所述的方法,其中一个或多个字符表达式模糊度特性是用户定义的。

10.根据权利要求8或权利要求9所述的方法,如果字符表达式的模糊度得分大于规定的阈值模糊度得分,则所述字符表达式被分类为不安全字符表达式。

11.根据权利要求8至10任一项所述的方法,其中对于每个药物靶标,一个或多个字符表达式模糊度特性包括以下一项或多项:

公开数据中包括所定义的指代所述药物靶标的一个或多个字符表达式的公开文档总数;

相对于公开数据中包括所定义的指代所述药物靶标的一个或多个字符表达式的公开文档总数,公开数据中包括指代所述药物靶标的已定义字符表达式其中一个的公开文档

数；

指代药物靶标的已定义字符表达式其中一个中的字符数；

指代药物靶标的已定义字符表达式其中一个的每个字符在公开数据中出现的频率，任选地，所述一个字符表达式中每个字符所述频率的和，任选地，所述和的对数；

包括所述一个已定义字符表达式的一个或多个药物靶标的已定义字符表达式的数量；

公开数据中包括已定义字符表达式中的一个的公开文档还包括所选字符表达式的概率，所述已定义字符表达式不同于所选字符表达式，所述所选表达式为从指代药物靶标的已定义字符表达式中选择的是安全字符表达式的字符表达式；和

公开数据中包含所选字符表达式的公开文档，还包括不同于所选字符表达式的已定义字符表达式的概率。

12.根据权利要求8至11任一项所述的方法，包括应用机器学习算法，基于所述一个或多个字符表达式的模糊度特性，赋予每个字符表达式的模糊度得分。

13.根据权利要求12所述的方法，其中所述机器学习算法使用所述一个或多个字符表达式不安全特性来赋予所述一个或多个字符表达式中每一个模糊度得分。

14.根据权利要求12或13所述的方法，其中所述机器学习算法包括正样本未标记学习技术。

15.根据权利要求12至14任一项所述的方法，其中所述机器学习算法包括随机森林分类器的应用。

16.根据权利要求12至15任一项所述的方法，其中在所述机器学习算法每次迭代后，用户检查赋予的模糊度得分子集，以确定是否手动更改任何赋予的模糊得分子集。

17.根据权利要求16所述的方法，其中所述子集对应于赋予的模糊度得分最高的字符表达式的规定数量。

18.根据前述任一权利要求所述的方法，其中至少一些公开文档的公开数据包括引用数据，所述引用数据指示一个公开文档对多个公开文档中一个或多个其他公开文档的引用，其中检索所述公开数据包括使用引用数据确定被同一公开文档引用的公开文档对。

19.根据权利要求18所述的方法，包括确定每一对识别的公开文档对的共引用值，所述共引用值表示同时引用所述公开文档对的公开文档数。

20.根据权利要求19所述的方法，包括基于确定的共引用值和引用所述公开文档对的公开文档，将公开文档对分配给多个公开文档群之一。

21.根据权利要求20所述的方法，其中所述将公开文档对分配给多个公开文档群之一包括应用贪婪优化算法。

22.根据权利要求20或权利要求21所述的方法，包括确定多个公开文档群中每一个是否将所述群与一个药物靶标相关联。

23.根据权利要求22所述的方法，当从属于权利要求2时，其中所述确定包括确定指代所述一个药物靶标的已定义字符表达式哪一个存在于所述群的每个公开文档的公开数据中。

24.根据权利要求22或权利要求23所述的方法，当从属于权利要求3时，其中所述确定包括确定在其公开数据中至少包括一个安全字符表达式的公开文档在群中的比例。

25.根据权利要求24所述的方法，如果所述比例大于规定的阈值比例，则确定所述群与

所述药物靶标中的一个相联系。

26.根据权利要求18至25任一项所述的方法，当从属于权利要求2时，其中检索公开文档对包括检索分别包括定义为指代药物靶标之一的至少一个字符表达式的公开文档对。

27.根据权利要求20至26任一项所述的方法，其中至少一些公开文档的公开数据不包括引用数据，并且对于每个公开文档中，所述方法包括基于其公开数据，特别地基于所述公开数据中一个或多个所定义的指代所述药物靶标的字符表达式，确定是否将所述公开文档分配给与药物靶标之一相关的一个群。

28.根据权利要求27所述的方法，当从属于权利要求3时，其中，如果所述公开文档的公开数据至少包含一个安全字符表达式的实例，则决定将上述公开文档分配给所述与一个药物靶标相关的一个群。

29.根据权利要求28所述的方法，其中，如果上述公开文档的公开数据不包括至少一个安全字符表达式的实例，则使用机器学习算法确定是否将上述公开文档分配给与一个药物靶标相关的一个群。

30.根据权利要求29所述的方法，其中所述机器学习算法包括正样本未标记学习技术。

31.根据权利要求29或权利要求30所述的方法，其中所述机器学习算法包括机器学习分类器的应用，任选地至少包括逻辑回归分类器极限树分类器、高斯过程分类器、k近邻分类器、岭回归分类器、随机森林分类器、以及支持向量机分类器之一。

32.根据前述任一权利要求所述的方法，其中，对于每个药物靶标，预期公开参数是与所述药物靶标相关的公开文档的预期数量，实际公开参数是与所述药物靶标相关的公开文档的实际数量。

33.根据前述任一权利要求所述的方法，其中，对于每个药物靶标，

预期公开参数为：与所述药物靶标相关的临床试验的预期数量；与所述药物靶标相关的综述公开文档的预期数量；以及与确定的公司规模相关的公开文档的预期数量之一；并且

实际公开参数为：与所述药物靶标相关的临床试验的实际数量；与所述药物靶标相关的综述公开文档的实际数量；以及与确定的公司规模相关的公开文档的实际数量之一。

34.根据前述任一权利要求所述的方法，其中包括使用历史公开文档中检索的公开数据训练的机器学习算法确定预期公开参数。

35.根据权利要求34所述的方法，其中所述机器学习算法是递归神经网络算法。

36.根据前述任一权利要求所述的方法，其中评估药物靶标以进行选择包括基于各药物靶标实际和预期公开参数的比较对所述药物靶标进行排序。

37.根据权利要求36所述的方法，其中根据各药物靶标实际和预期公开参数之间差异的参数指示对所述药物靶标进行排序。

38.根据前述任一权利要求所述的方法，包括确定药物靶标对之间的靶标-靶标共现参数，所述靶标-靶标共现参数是基于检索的公开数据中的指示确定的，其中公开文档的药物靶标对的两个药物靶标相关联，每个靶标-靶标共现参数指示所述两个药物靶标出现在一个药物靶标对的公开文档数；并且所述方法包括基于确定的靶标-靶标共现参数评估的一个或多个药物靶标以进行选择。

39.根据前述任一权利要求所述的方法，包括检索公开数据，为每个公开文档提供指

示，说明各公开文档是否与一个或多个疾病相关。

40.根据权利要求39所述的方法，包括为每种疾病定义一个或多个与所述疾病相关的字符表达式，其中检索公开数据包括，则检索每种疾病的一个或多个字符表达式的公开数据。

41.根据权利要求39或权利要求40所述的方法，包括确定每个药物靶标和每种疾病之间的靶标-疾病共现参数，基于从检索的公开数据中，公开文档每个药物靶标和每种疾病相关的指示，确定靶标-疾病共现参数，每个靶标-疾病共现参数指示存在一个所述药物靶标和一个所述疾病的公开文档数；并且所述方法包括基于确定的靶标-疾病共现参数评估一个或多个药物靶标以进行选择。

42.根据前述任一权利要求所述的方法，包括将主题建模算法应用于与每个药物靶标相关的公开文档的公开数据，以获得与每个药物靶标相关的一个或多个主题，并基于获得的一个或多个主题评估一个或多个药物靶标以进行选择。

43.根据权利要求42所述的方法，包括对每个药物靶标，基于获得的一个或多个主题确定所述药物靶标与一个或多个公开文档相关性的错误。

44.根据权利要求42或权利要求43所述的方法，其中主题建模算法至少包括潜在狄利克雷分配算法及非负矩阵分解算法之一。

45.根据前述任一权利要求所述的方法，其中与一个或多个公开文档有关的公开数据包括公开文档标题、公开文档摘要、以及与公开文档相关的一个或多个关键词中的一个或多个。

46.根据前述任一权利要求所述的方法，其中所述公开数据包括多个公开文档中每一个的公开日期。

47.根据权利要求46的所述的方法，其中所述公开日期决定了每一份公开文档是历史公开文档还是当前公开文档。

48.根据权利要求46或权利要求47所述的方法，其中公开文档的公开日期早于预期截止日期，则定义为历史公开文档。

49.根据权利要求48所述的方法，其中公开文档的公开日期晚于预期截止日期，则定义为当前公开文档。

50.根据任何权利要求46至49所述的方法，其中出版日期在预期截止日期范围内的公开文档被定义为当前公开文档。

51.根据前述任一权利要求所述的方法，其中至少一个公开数据源包括至少一个在线公开数据源。

52.根据前述任一权利要求所述的方法，其中一个或多个药物靶标包括一个或多个基因，任选地，一个或多个人类基因，任选地由所述一个或多个基因编码的一个或多个蛋白质。

53.根据前述任一权利要求所述的方法，包括使用对一个或多个药物靶标的评估，以为选择在药物发现项目中使用的至少一个药物靶标提供信息。

54.根据权利要求53所述的方法，包括基于所述评估选择药物发现项目中使用的至少一个药物靶标，来设计所述药物发现项目。

55.根据权利要求53或54所述的方法，包使用至少一个选定的药物靶标开展药物发现

项目。

56.根据权利要求55所述的方法，其中进行药物发现项目包括选择和测试针对至少一个选定的药物靶标的化合物。

57.通过计算机处理器执行的非晶体管、计算机可读存储介质存储指令，所述计算机处理器执行前述任一权利要求所述的方法。

58.用于药物靶标选择的计算机设备，所述计算机设备被配置为：

从至少一个公开数据源获取公开数据，并链接至多个公开文档，包括历史公开文档和当前公开文档；

检索所述公开数据，为每个公开文档提供指示，说明各公开文档是否与一个或多个药物靶标相关；

基于历史公开文档中检索的公开数据，确定一个或多个药物靶标中每一个的预期公开参数；并基于当前公开文档的公开数据，确定一个或多个药物靶标中每一个的实际公开参数；以及

基于相对于预期公开参数的实际公开参数，评估一个或多个药物靶标的每一个，以供选择。

计算药物靶标选择

技术领域

[0001] 本发明涉及用于计算选择靶分子或基因(例如药物靶标)的方法和系统,分子(例如药物)将被设计为以最佳方式与靶分子或靶基因相互作用。

背景技术

[0002] 药物发现是鉴定候选化合物以进入药物开发的下一阶段的过程,例如临床前试验。这些候选化合物需要满足进一步开发的某些标准。现代药物发现涉及初步筛选“命中”化合物的鉴定和优化。特别是,需要根据目标标准对此类化合物进行优化,其中可包括对许多不同特性的优化。要优化的特性可以包括,例如:针对目标生物靶标的活性;对非目标生物靶标的选择性;较低概率的毒性;良好的药物代谢和药物代谢动力学(ADME)特性。只有满足特定要求的化合物才能成为能够继续进行药物开发过程的候选化合物。

[0003] 因此,鉴定和选择生物或药物靶标,然后对命中的化合物进行优化是药物发现过程中的一个关键步骤。事实上,靶标鉴定和优先排序是药物发现过程和新药开发的第一个关键步骤。药物靶标是存在于生物体中与药物相互作用的某种物质,例如蛋白质或核酸,药物与之相互作用,例如结合。这种与药物的相互作用导致药物靶标的行为发生变化。有前景的药物靶标可以是与待决的特定疾病相关的靶标,例如药物靶标改变疾病或在疾病的病理生理学中发挥作用。

[0004] 由于存在大量潜在的药物靶标,选择药物靶标的过程变得复杂。例如,对于一个人类疾病,有数以万计的可成为新药的靶标的基因表达蛋白质。此外,由于按医学分类,存在数千种人类疾病,因此有数百万种,特别地,数亿种可能的靶标疾病组合。因此,解决办法的检索空间非常大,以至于对每个组合或假设进行实验测试是不可行的。

[0005] 传统上,药物靶标是由药物化学家根据已发表的科学文献(如学术期刊和公共数据库)逐案鉴定的。也就是说,传统上,大量的靶标鉴定是通过科学家个体利用他们的专业知识解释科学文献进行的。然而,这种方法的一个日益严重的问题是剧增的可供检索的公共数据,如学术论文。生命科学领域有数千万篇已发表的科学论文,成千上万的基因组和成百上千个数据库。事实上,不考虑其他数据来源,如预印本和临床试验报告,每天都有成千上万篇同行评审的文章发表。显然,人类在选择药物靶标时不可能掌握所有可用的数据来源。换言之,不断增长的发表率使得很难维持概观,以鉴定有前景的新的或现有的药物靶标。

[0006] 优化药物靶标的鉴定和选择对于优化整个药物发现过程至关重要。特别是,针对特定药物发现项目的药物靶标的最佳选择可以增加在更短的时间内,即在更短的项目设计周期内鉴定候选化合物的概率。反过来,这减少了特定项目的相关时间和/或相关成本。

[0007] 本发明就是在这种背景下提出的。

发明内容

[0008] 本发明提供了一个鉴定可与特定疾病相关的药物的生物靶标的改进方法,以减少

药物发现过程的总时间和/或相关成本，例如提高作为特定药物发现项目一部分的候选化合物的鉴定效率。此外，本发明提供了用于药物发现的方法。特别地，在包括选择至少一个药物靶标的方法中，所述方法可以包括基于所述至少一个药物靶标进行药物发现项目；以及任选地选择和/或合成和/或测试针对所述选定的至少一个药物靶标的潜在治疗性化合物。

[0009] 根据本发明的一个方面，提供了一个用于计算药物靶标选择的方法。所述方法包括从至少一个公开数据源获取公开数据，并链接至与包括历史公开文档和当前公开文档的多个出版物。所述方法包括检索公开数据以针对每个公开文档提供关于相应的公开文档是否与一个或多个药物靶标相关的指示。所述方法包括基于从历史公开文档中检索到的公开数据来确定一个或多个药物靶标中的每一个的预期公开参数，以及基于从当前公开文档中检索到的公开数据来确定一个或多个药物靶标中每一个的实际公开参数。所述方法包括基于其实际公开参数相对于其预期公开参数来评估一个或多个药物靶标中的每一个以进行选择。

[0010] 所述方法可以包括为每个药物靶标定义一个或多个字符表达式指代药物靶标，并且其中检索公开数据包括在公开数据中检索每个药物靶标的一个或多个个字符表达式。

[0011] 所述方法可以包括针对每个药物靶标将一个或多个字符表达中的每个字符表达分类为安全字符表达或不安全字符表达。所述分类可以基于公开数据中一个字符表达式指代药物靶标的可能性。

[0012] 在一些实施方案中，如果从公开文档之一检索到的公开数据包括安全字符表达，则确定该公开文档与药物靶标相关联。

[0013] 一个或多个字符表达式可以是用户定义的，以分类为安全字符表达式。

[0014] 一个或多个字符表达式不安全特性可以是用户定义的，以指示对应的字符表达式是不安全的。所检索的公开数据中表现出一个或多个字符表达式不安全特性的字符表达式可以被分类为不安全字符表达式。

[0015] 所述一个或多个用户定义的字符表达式不安全特性可以包括以下一个或多个：与特定自然语言中的单词相对应的字符表达式；具有少于规定数量字符的字符表达式，任选地，其中所述规定数量为三；以及被定义为指代至少两个不同的药物靶标的字符表达式。

[0016] 一个或多个字符表达式模糊度特性可以被定义为赋予一个或多个字符表达式模糊度得分。每个字符表达式可以基于相应赋予的模糊度得分被分类为安全字符表达式或不安全字符表达式。

[0017] 一个或多个字符表达式模糊度特性中可以是用户定义的。

[0018] 如果字符表达式的模糊度得分大于规定的阈值模糊度得分，则所述字符表达式可以被分类为不安全字符表达式。

[0019] 对于每个药物靶标，所述一个或多个字符表达式模糊度特性可以包括以下一个或多个：公开数据中包括一个或多个指代药物靶标的已定义字符表达式的公开文档总数；相对于公开数据中包括一个或多个指代药物靶标的已定义字符表达式的公开文档总数，公开数据中包括一个或多个指代药物靶标的已定义字符表达式的公开文档数；指代药物靶标的已定义字符表达式中其中一个的字符数；指代药物靶标的已定义字符表达式中其中一个的每个字符出现在所述公开数据中的频率，任选地是所述一个字符表达式中每个字符的频率

之和,任选地为所述和的对数;包括所述一个已定义表达式的药物靶标已定义字符表达式数;公开数据中包括已定义字符表达式中的一个的公开文档还包括所选字符表达式的概率,所述已定义字符表达式不同于所选字符表达式,所述所选表达式为从指代药物靶标的已定义字符表达式中选择的是安全字符表达式的字符表达式;以及,公开数据中包括所选择字符表达式的公开文档,还包括不同于所选字符表达式的已定义字符表达式的概率。

[0020] 所述方法可以包括应用机器学习算法,以基于一个或多个字符表达式模糊度特性将模糊度得分赋予一个或多个字符表达式中的每一个。

[0021] 所述机器学习算法可以使用一个或多个字符表达式的不安全特性来将模糊度得分赋予一个或多个字符表达式中的每一个。

[0022] 所述机器学习算法可以包括正样本未标记学习(positive-unlabelled learning)技术。

[0023] 所述机器学习算法可以包括随机森林分类器的应用。

[0024] 在一些实施方案中,在机器学习算法的每次迭代后,用户检查先赋(ascribed)模糊度得分的一个子集,以确定是否手动更改先赋模糊度得分的任何一个子集。

[0025] 所述子集可以对应于具有最高先赋模糊度得分的规定数量的字符表达式。

[0026] 至少一些公开文档的公开数据可以包括引用数据,所述引用数据指示由一个公开文档对多个公开文档中的一个或多个其他公开文档所作的引用。检索公开数据可以包括使用引用数据来识别已经被同一公开文档引用的公开文档对。

[0027] 所述方法可以包括,对于每个识别的公开文档对,确定共同引用值,所述共同引用值表示引用这对公开文档的多个公开文档。

[0028] 所述方法可以包括基于所确定的共同引用值和引用公开文档对的公开文档,将公开文档对分配给多个公开文档群中的一个。

[0029] 在一些实施方案中,将公开文档对分配给多个群中的一个包括贪心优化算法的应用。

[0030] 所述方法可以包括,对于多个公开文档群中的每一个,确定是否将所述群与药物靶标之一相关联。

[0031] 所述确定可以包括确定指代一个药物靶标的已定义字符表达式中哪些存在于群中的每个公开文档的公开数据中。

[0032] 所述确定可以包括确定群中在其公开数据中包括至少一个安全字符表达式的公开文档的比例。在一些实施方案中,如果所述比例大于规定的阈值比例,则确定群与一个药物靶标相关联。

[0033] 在一些实施方案中,检索所述公开文档对包括检索包括至少一个指代一个药物靶标的已定义字符表达式的公开文档对。

[0034] 在一些实施方案中,至少一些公开文档的公开数据不包括引用数据。对于每一个公开文档,所述方法可以包括基于其公开数据,特别是基于在其公开数据中一个或多个指代药物靶标的已定义字符表达式,确定是否将公开文档分配给与药物靶标之一相关的群中的一个。

[0035] 在一些实施方案中,如果公开文档的公开数据包括至少一个安全字符表达式,则确定将公开文档分配给与一个药物靶标相关的一个群。

[0036] 在一些实施方案中，如果公开文档的公开数据不包括至少一个安全字符表达式，则使用机器学习算法来确定是否将公开文档分配给与药物靶标之一相关的群中的一个。

[0037] 所述机器学习算法可以包括正样本未标记学习技术。

[0038] 机器学习算法可以包括机器学习分类器的应用，任选地，至少一个：逻辑回归分类器、额外的树分类器、高斯过程分类器、k近邻分类器、岭回归分类器、随机森林分类器、以及支持向量机分类器。

[0039] 对于每个药物靶标，预期公开参数可以是与药物靶标相关的公开文档的预期数量，而实际公开参数可以为与药物靶标相关的公开文档的实际数量。

[0040] 对于每个药物靶标，预期公开参数可以是以下之一：与药物靶标相关的临床试验的预期数量；与药物靶标相关的综述公开文档的预期数量；以及，与所定义的公司规模相关的公开文档的预期数量；并且，实际公开参数可以是以下之一：分别为，与药物靶标相关的临床试验的实际数量；与所述药物靶标相关联的综述公开文档的实际数量；以及，与所定义的公司规模相关的公开文档的实际数量。

[0041] 在一些实施方案中，确定预期公开参数包括使用通过从历史公开文档中检索到的公开数据进行训练的机器学习算法。

[0042] 所述机器学习算法可以是递归神经网络算法。

[0043] 在一些实施方案中，评估药物靶标以供选择包括基于药物靶标各自的实际和预期公开参数的比较来对药物靶标进行排名。

[0044] 药物靶标可以根据指示其各自的实际和预期公布参数之间的差异的参数进行排序。

[0045] 所述方法可以包括确定药物靶标对之间的靶标-靶标共现参数，所述靶标-靶标共存参数是基于来自检索到的公开数据的指示来确定的，所述公开数据是与两个药物靶标相关的公开文档对的公开数据。每个靶标-靶标共现参数可以指示成对的两个药物靶标都出现在公开文档的数量。所述方法可以包括基于所确定的靶标-靶标共现参数来评估一个或多个药物靶标以供选择。

[0046] 所述方法可以包括检索公开数据以针对每个公开文档提供关于相应的公开文档是否与一个或多种疾病相关的指示。

[0047] 所述方法可以包括，对于每种疾病，定义指代该疾病的一个或多个字符表达式。检索公开数据可以包括在公开数据中检索每种疾病的一个或多个字符表达式。

[0048] 所述方法可包括确定每个药物靶标和每种疾病之间的靶标-疾病的共现参数。基于检索的公开数据中，公开文档每个药物靶标和每种疾病相关的指示，确定靶标-疾病共现参数。每个靶疾病共现参数可指示出现药物靶标之一和疾病之一的公开文档的数量。所述方法可包括基于确定的靶标-疾病的共现参数数评估一个或多个药物靶标。

[0049] 所述方法可包括将主题建模算法应用于与每个药物靶标相关的公开文档的公开数据，以获得与每个药物靶标相关的一个或多个主题。所述方法可包括基于获得的一个或多个主题评估一个或多个药物选择靶标。

[0050] 所述方法可包括，对每个药物靶标，基于获得的一个或多个主题，确定所述药物靶标与一个或多个公开文档相关性的错误。

[0051] 主题建模算法可包括以下至少一个：潜在狄利克雷分配算法；和非负矩阵分解算

法。

[0052] 与一个或多个公开文档相关的公开数据可包括一个或多个:公开文档的标题;公开文档摘要;以及与公开文档相关的一个或多个关键词。

[0053] 公开数据可包括多个公开文档的公开日期。

[0054] 公开日期可决定每个公开文档是历史公开文档还是当前公开文档。

[0055] 在一些实施方案中,公开日期在规定截止日期之前的公开文档被定义为历史公开文档。

[0056] 在一些实施方案中,公开日期在规定截止日期之后的公开文档被定义为当前公开文档。

[0057] 在一些实施方案中,公开日期在规定截止日期范围内的公开文档被定义为当前公开文档。

[0058] 至少一个公开数据源可包括至少一个在线公开数据源。

[0059] 一个或多个药物靶标可包括一个或多个基因,任选地,一个或多个人类基因,任选地,由这些基因编码的一个或多个蛋白质。

[0060] 所述方法可以包括使用一个或多个药物靶标的评估来为选择至少一个药物靶标用于药物发现项目提供信息。

[0061] 所述方法可包括设计药物发现项目,基于所述评估选择至少一个用于药物发现项目的药物靶标。

[0062] 所述方法可包括利用至少一个所选的药物靶标开展药物发现项目。

[0063] 在一些实施方案中,进行药物发现项目包括选择,任选地,(计算机模拟、体外和/或体内)合成和测试针对至少一个所选药物靶标的化合物。

[0064] 根据本发明的另一方面,提供了一种用于鉴定对药物靶标/靶标分子具有结合亲和力的药物/化合物的方法,所述方法包括开展药物发现项目(例如,基于一个根据本文所公开的方面和实施方案鉴定药物靶标的方法),以及任选地,选择和/或合成和/或测试针对至少一个所选药物靶标的化合物,以确定对药物靶标具有治疗活性的化合物;其中"治疗活性"可包括但不限于所需的结合特性(如亲和力、选择性);抑制特性;激动剂或拮抗剂特性。

[0065] 应理解,本文公开的任何方面或实施方案的任何特征都可以与本文公开的任何其他方面或实施方案的任何特征相结合,并且所有此类特征组合均被考虑并在此公开,除非此类组合明显不兼容。

[0066] 根据本发明的另一方面,提供了一个非晶体管、计算机可读存储介质存储指令,当由计算机处理器执行时,所述指令使得计算机处理器执行上述方法。

[0067] 根据本发明的另一方面,提供了用于药物靶标选择的计算机设备。所述计算机设备被配置为从至少一个公开数据源获取、接收或下载公开数据,并连接至多个公开文档(包括历史公开文档和当前公开文档)。所述计算机设备被配置为检索公开数据,以为每个公开文档提供指示,即相应的公开文档是否与一个或多个药物靶标相关。所述计算机设备被配置为根据历史公开文档中的检索的公开数据确定一个或多个药物靶标的预期公开参数,并基于当前公开文档中检索的公开数据确定每个药物靶标的实际公开参数。所述计算机设备被配置为基于其相对于预期公开参数的实际公开参数对一个或多个药物靶标中的每一个进行评估。

附图说明

[0068] 现在将参考以下附图对本发明的实施例进行描述，其中：

[0069] 图1总结了根据本发明的计算药物选择方法的步骤；

[0070] 图2示出了一个图形数据库，显示了使用图1的方法确定特定目标基因与相关的公开文档之间的关系；

[0071] 图3示出了使用图1的方法确定的与不同基因相关的预测与真实公开动态的比较；

[0072] 图4示出了使用图1的方法确定的不同疾病组相关的不同基因相对于真实公开的预测动态；

[0073] 图5示出了使用图1的方法确定的显示基因-基因连接和基因疾病连接的共现性网络；和，

[0074] 图6示出了使用图1的方法确定的与不同提取主题相关的不同公开组的不同目标基因的出版物数量的时间线。

具体实施方式

[0075] 分子或药物设计可以被视为一个多维优化问题，利用假设生成和实验周期来推进知识。每个化合物的设计都可以被认为是在实验中被证伪的假设。实验结果表示为构效关系，它构建了一个关于化学结构可能包含所需特征的假设的情形。药物设计过程也是一个优化问题，因为每个项目都需要一个定义的产品特性——即药物靶标功能——用于分析命中化合物的所需的特定属性。

[0076] 药物发现过程通常在称为设计周期的迭代中进行。每次迭代时，合成一组分子或化合物，并测量其生物性质。对其活性进行了分析，并基于先前迭代的经验提出了新的化合物集合。在找到临床候选药物之前重复此过程。与活性一样，测量的生物特性可以包括选择性、毒性、吸收、分布、代谢和排泄中的一个或多个。

[0077] 药物发现过程一般耗时且成本较高。因此，在过程的任何阶段都可以发现效率，这有助于减少与药物发现项目相关的时间和成本。制药企业正在积极寻找降低其流失率、药物研发所需时间以及相关研发成本的方法。

[0078] 选择开发新药的药物靶标是药物发现过程中的第一个决定，也是最重要的单个决策。从历史上看，基于对现有文档的科学解释，靶标鉴定基本上是在个案基础上进行的。然而，除了预印本、专利文件数据和临床试验报告的公开外，每天还有数千篇同行评议文章被发表。在线资源PubMed®允许检索和访问生命科学和生物医学信息领域的公开文档。仅PubMed®就包含了数以百万计的公开文档——特别是3000多万份出版物——以及每九年左右翻一番的科学产出。这创造了“未发现的公共知识”集合，因为人类(如药物化学家)显然不能跟上已公开文献的所有发展。反过来，这使得人类更难根据现有文献就药物靶标的鉴定和选择做出明智决定。

[0079] 潜在药物靶标的巨大检索空间也使人类难以做出最佳选择。对于一个人类疾病，理论上可在特定疾病行为中起作用的成千上万种基因中，存在数百万种基因疾病组合可以进行调查，但在实践中显然不可行。

[0080] 上述问题意味着使用计算方法分析大量可用信息和大量可能的基因疾病组合成为一个有吸引力的提案。特别是，存在对机器学习(ML)、人工智能(AI)和其他计算方法及其

他计算的高需求，这些方法可以利用当前知识，并通过优化药物靶标的识别和选择，促进对这一庞大文档量的概述的维护。

[0081] 本发明认识到，计算方法可用于识别已公开文献中关于潜在药物靶标(例如基因)的趋势，其可用于为选择药物靶标用于特定药物发现项目提供信息。本发明的优点在于，它提供了一个药物靶标选择的计算方法，可以检测与特定基因相关的变化趋势，例如，关于特定基因的科学突破。

[0082] 根据本发明，计算药物靶标选择方法的第一步包括从至少一个公开数据源获取公开数据。例如，公开数据源可以是在线公开数据源，可包括PubMed®等数据库，所述PubMed®可以访问数百万份特定目标领域的学术或期刊文章形式的公开文档。公开数据可额外地或可选地从其他来源获取，如已发表的临床试验报告数据、已发表的专利文件数据和/或已发表的文章预印本。可以理解，公开数据可以从任何合适的公开数据源以及任何数量的此类合适来源获得和获取。

[0083] 获取给定公开文档的公开数据中包含的信息可取决于文件的特定类型或获得公开数据的特定来源。例如，公开数据可仅限于作为给定公开文档的开源数据提供的数据，并且只能在付费墙后访问其他信息。

[0084] 获取的公开数据与多种不同的公开文档有关，即每个公开文档都有与之相关的公开数据。为了检测文献中的长期趋势，获取公开数据的公开文档可被分割为历史公开文档和当前公开文档。这样的分割可以在基于历史文献的可能预期观察到的趋势和基于当前公开的实际观察到的趋势之间进行比较，这将在下文中更详细地讨论。可以理解的是，历史公开数据和当前公开数据的获取可分开进行，也可同时进行。

[0085] 与多数公开文档中至少一些相关的公开数据可包括这些公开文档的出版日期或与之相关。这可被用于确定或定义哪些文档被鉴定为历史公开文档，哪些文档被标识为当前公开文档。出版日期可是相关文件公开的具体日期、月份或年份。纯粹出于示例目的，公开日期在预期截止日期之前的公开文档被定义为历史公开文档，公开日期在预期截止日期之后具的公开文档可被定义为当前公开文档。然而，文档的公开日期可以以任何适当的方式使用，以确定其是历史性的还是当前的文档，例如，公开日期在预期截止日期范围内的公开文档被可以定义为当前公开文档。

[0086] 本发明的下一步包括检索获取的公开数据，以为每个公开文档提供指示，即各自的公开文档是否与一个或多个药物靶标(例如基因)相关。也就是说，检索每个公开文档的公开数据中包含的信息，以识别每个公开文档与一个或多个药物靶标的潜在关联或链接。这样的检索可以以任何合适的方式进行。一个选择是在公开数据中检索目标药物靶标的提及。就此而言，公开数据可能包括一个或多个标题、摘要以及一个或多个与公开文档相关的关键词等数据。这些信息通常可以从存储期刊论文的在线公开数据库中获得，例如，这些信息可以很容易地作为与不同公开文档相关的出版物数据的一部分。

[0087] 在一个示例中，定义了(例如，由用户定义)一个或多个目标药物靶标的名称，例如目标基因，并自动检索公开数据的内容以查找所定义的药物靶标名称。例如，如果在与特定公开文档相关的公开数据中找到所定义药物靶标名称之一，则该药物靶标可被视为与特定公开文档相关或链接。例如，根据公认的命名法，所定义的药物靶标的名称可以是被批准命名，例如被批准的基因命名。在下文中，"基因命名"是指人类基因命名委员会所接受的

19084个人类蛋白质编码基因中的任何一个基因的批准命名;然而,人们会明白,这纯粹是为了说明目的,是非限定性的。

[0088] 通过计算方法自动分析生物医学文献的一个重要障碍是使用来自不同竞争来源的非冗余替代基因(药物靶标)同义词、命名和首字母缩略词,它们在其他研究领域可具有其他含义。也就是说,在文献中,单个药物靶标,如基因,可以以许多不同的方式指代,并被该领域接受。也可以是特定药物靶标的一个或多个同义词在不同的环境中与一个完全不同的概念或具有完全不同的含义的术语或表达重合,或者是其中的一部分。这些因素使得通过自动(计算)分析很难明确地确定哪些包含与药物靶点名称重合的参考的公开文档实际上是指该药物靶点。

[0089] 为了在公开数据中识别多个以不同方式、多个不同名称或不同语言提及的目标药物靶标的参考文献,所述方法可包括为每个药物靶标定义一个或多个指代所述药物靶标的字符表达式或同义词。这些字符表达式可以由用户定义,并且可以包括可以计算检索的任何合适字符。例如,合适的字符可包括一个或多种自然语言中使用的字母或其他类型的符号。然后,对于每个药物靶标,检索公开数据可包括检索每个药物靶标的一个或多个所定义的字符表达式或同义词的公开数据。

[0090] 在这种情况下,如果药物靶标是基因,“基因同义词”可以指科学界指代或已经指代的任何可能的基因名称变体。批准的基因名称——如上所述——也包括在基因同义词中。举例来说,“EGFR”是批准的基因名称,其中“EGFR”、“表皮生长因子受体”、“ERBB1”、“ErbB1”、“c-erbB1”、“HER1”和“ERBB”是基因同义词。

[0091] 定义为潜在地指代特定药物靶标的不同字符表达式可从不同来源获得。例如,在目标人类基因的情况下,不同人类基因的各种字符表达式,即同义词,可以从不同来源收集,以采样提到人类基因名称的潜在公开文档。

[0092] 如上所述,生物医学文献自动分析的另一个问题是,公开数据中药物靶标的公认同义词的实例,即所定义的字符表达式的实例,可能实际指代或不指代特定文档中目标药物靶标。事实上,在不同的环境中,一个特定字符表达式用于指代两个不同基因并不常见。因此,必须在科学文献中消除生物医学实体的歧义,以便准确分析不同药物靶标的公开动态。

[0093] 对药物靶标定义的每个字符表达式,例如基因的每个不同同义词,可被视为具有与药物靶标相关的不同程度的模糊度。即,文献中一些同义词具有特定目标基因不同的含义,例如这些同义词可以被视为有与药物靶标相关的更大程度的模糊度,因为很可能文献中这些同义词的实例指代与目标基因不同物质。另一方面,当一个特定的字符表达式或同义词没有指代目标特定基因之外(共同)的物质的含义时,则这个同义词可能被视为具有较低程度的模糊度,因为文献中的同义词很可能指代目标基因。

[0094] 由于不同的原因,可会出现与基因同义词(药物靶标字符表达式)相关的模糊度水平。例如,一个被称为“杂乱基因名(同音异义词)”的基因名可以被视为任何与一个以上基因同义的基因名。这可包括以前的官方基因命名(根据公认的命名法),因为这些基因命名尚未被文献公开。作为一个说明性实例,“CDH3”和“钙黏蛋白3”对于基因命名“CHD15”和“CHD3”是混杂的。此外,“ARP1”是基因命名“NR2F2”、“ACTR1A”、“ACTR1B”、“ANGPTL1”、“APOBEC2”、“ARFRP1”和“PITX2”的基因同义词。作为另一个实例,一个被称为“嵌套基因同

义词”的基因同义词可以被视为另一个基因同义词的一部分。例如,“胰岛素”是“胰岛素受体”的嵌套基因同义词。此外,“TNF”是“TNF受体超家族成员1A”(基因命名“TNFRSF1A”)和“TNF受体关联因子2”(基因命名“TRAF2”)的嵌套基因同义词。

[0095] 为了更准确地分析检索的公开数据,本发明的方法因此可以包括对公开数据中发现的每个药物靶标(例如基因)的一个或多个所定义的字符表达式(例如基因同义词)进行分类,以成为安全的字符表达式或不安全的字符表达式。所述分类基于公开数据中字符表达式指代药物靶标的可能性,即与字符表达式相关的模糊度水平。特别地,安全的字符表达式可能具有与药物靶标相关的相对较低的模糊度水平,而不安全的字符表达式可能具有与药物靶标相关的相对较高的模糊度水平。例如,当字符表达式为基因同义词时,″不安全的基因同义词″可能包括在其他研究领域或在不同环境中具有不同含义的基因同义词,例如出现在英语字典中的词。例如,“STAR”基因命名可能被认为是不安全的字符表达式,而其基因同义词“类固醇合成急性调节蛋白”则与之相反。作为另一个实例,“CCP4”可能被认为是不安全的,因为它既是基因同义词,也是晶体学软件的名称。

[0096] 如果确定从一个公开文档中检索的公开数据包含安全的字符表达式(针对特定药物靶标),则可确定该公开文档与该药物靶标相关。换言之,该公开文档被视为与待决的特定药物靶标(如目的基因)相关。对于特定的药物靶标,至少一些所定义的潜在指代药物靶标的字符表达式,即文献中出现的字符表达式,可被视为绝对安全的字符表达式。特别地,一个或多个字符表达式可能被用户定义为安全的字符表达式。也就是说,有一些字符表达式没有模糊度——或者模糊度水平很低——因此,先验地知道,公开数据中这些字符表达式的实例实际上指代相关药物靶标,而不考虑这些实例出现在公开数据中的环境。因此,当在公开数据中找到此类字符表达式时,可以自动将其分类为安全的字符表达式。这意味着,当在某个公开文档的公开数据中找到此类字符表达式时,可以自动确定所述公开文档与字符表达式为所定义的同义词的药物靶标相关或链接。

[0097] 字符表达式的一个或多个特征可以被定义,例如由用户定义,以指示显示此类特征的字符表达式具有高模糊度,因此它们不安全。特别地,检索公开数据中显示此类“字符表达式不安全特性”的字符表达式可会自动被分类为不安全字符表达式。用户定义的字符表达式不安全特性的示例可以是与特定自然语言中的单词对应的任何字符表达式,例如英语词典中的单词(见上文的实例“STAR”)。用户定义的字符表达式不安全特性的另一个实例可是字符表达式小于规定的字符数量。例如,规定的数字可以是三个或任何其他适当定义的数字。用户定义的字符表达式不安全特性的另一个例子是被定义为指代至少两个不同药物靶标的字符表达式(见上文提到的“杂乱基因名称”)。这样,至少包含一个所定义的不安全特性的字符表达式被视为绝对不安全。可以理解,字符表达式的任何合适特性都可以被定义为表示字符表达式高模糊度,因此不安全。

[0098] 根据上述定义,可能存在大量已定义字符表达式不被视为绝对安全或绝对不安全。在这种情况下,可以确定或计算与剩余字符表达式相关的模糊度水平,以便将这些字符表达式分类为安全或不安全。确定哪些字符表达式或同义词具有潜在的高模糊度(因此被视为不安全)的一个选项是执行特征工程以获取表征不安全同义词的变量,然后基于获得的变量对每个同义词赋予模糊度水平。例如,更长的基因名模棱两可的可能性小。更常见地,字符表达式模糊度特性可以被定义为任何合适的方式,以赋予一个或多个字符表达式

模糊度得分。其可以被用户定义，例如基于特征工程或其他。基于相应赋予的模糊度得分，这些字符表达式中的每一个都可被分类为安全字符表达式或不安全字符表达式。例如，如果字符表达式的模糊度得分大于规定的模糊度得分阈值，则可以将其归类为不安全字符表达式。

[0099] 基于获得的字符表达式模糊度特性(例如通过特征工程)，机器学习算法可被应用于对每个字符表达式或同义词进行模糊度评分。特别地，机器学习算法可以使用字符表达式不安全特性来赋予每个未分类为安全或不安全的字符表达式模糊度得分。也就是说，模糊度得分被赋予未标记的同义词集合，即不是先前被标记为安全的同义词集合或先前被标记为不安全的同义词集合。模糊度评分用于将尚未标记的同义词标记为安全或不安全。为了实现这一点，机器学习算法可以包括应用正样本未标记学习技术，例如正样本未标记装袋策略(positive-unlabelled bagging strategy)，以及分类方案，例如随机森林分类器。

[0100] 机器学习算法可以作为迭代过程运行。算法每次迭代后，用户可检查上赋予的糊度得分的子集，以确定是否手动改变其中任何一个，即纠正算法的分类，以训练算法并提高后续迭代的准确性。例如，所述子集可对应于的规定数量的同义词或字符表达式，其被赋予了最高的模糊度得分(因此被算法认为是最不安全的)。

[0101] 字符表达式模糊度特性—由特征工程获得，举例来说—可包括公开数据中包括所定义的指代特定药物靶标的字符的公开文档的总数。模糊度特性可包括，相对于公开数据中包括指代特定药物靶标的已定义字符表达式的公开文档的总数，公开数据中包括指代所述药物靶标的已定义字符表达式其中一个的公开文档数。模糊度特性可包括其中一个指代药物靶标的已定义字符表达式的字符数。例如，较短的表达式通常可被视为比较长的表达式更不明确。模糊度特性可包括公开数据中其中一个指代药物靶标的已定义字符表达式中每个字符的出现频率。更具体地，为特定字符表达式中每个字符的频率之和，即整个表达式的频率分数。可以使用任何适用于总频率得分的指标，例如该总得分的对数。例如，包含较少共同字符的同义词或字符表达式可比完全由公开数据中常见字符组成的同义词(或通常被认为常见)的更明确。特定字符表达式或同义词另外的模糊度特性可基于包括特定字符表达式的药物靶标的已定义字符表达式数。换言之，模糊度特性可基于与特定同义词相关的嵌套同义词(如上所定义)的数量，即包含待决的特定基因同义词的其他基因同义词的数量。另一个模糊度特性可能是公开数据中包括已定义字符表达式中的一个的公开文档还包括所选字符表达式的概率，所述已定义字符表达式不同于所选字符表达式，所述所选表达式为从指代药物靶标的已定义字符表达式中选择的是安全字符表达式的字符表达式。表达不同的是，模糊度特性可以是在给定的特定公开文档的公开数据中发现目标基因同义词的条件概率，因为文本中出现了一个相同基因命名(如上所定义)的(其他)基因同义词。另一个模糊度特性本质上可以是上述的“反向概率”，即公开数据中包含选定字符表达式(即待决的同义词)的公开文档也包含该药物靶标的另一个所定义的字符表达式的概率。表达不同的是，模糊度特性可以是在给定的特定公开文档的公开数据中发现同一基因命名的一个(其他)基因同义词的条件概率，因为文本中出现了目标基因同义词。作为最后的实例，模糊度特性可以基于待决的字符表达式是否为特定药物靶标的可接受字符表达式，例如，待决的基因同义词是否为基因命名。

[0102] 本发明的方法可以使用标记的字符表达式—即标记的安全或不安全—取决于表

达的相关模糊度，来明确地将每个药物靶标(如人类基因)与每种药物靶标(例如人类基因)关联或链接至获取了公开数据的公开文档的子集。为此，可以使用基于共同引用网的方法。也就是说，对公开文档的引用可被用于更准确地确定在公开数据中对特定药物靶标的字符表达式的提及实际上是否意味着所述特定药物靶标与公开文档相关(或所述字符表达式是否在不同的语境中提及，使得所述字符表达式实际上不指代药物靶标)。具体地，以下更详细描述的共同引用方法可用于减少或消除检索公开数据中(即其公开数据提及了已定义字符表达式(基因同义词)的公开文档)的“假正样本”—这表明公开文档可能与药物靶标(基因)相关的字符表达式相关—但实际上与药物靶标不相关或相链接。这种方法可以被认为是基于一个假设，即包含″假正样本″的公开文档将倾向归属于与包含″真正样本″的公开文档不同的涉及不同研究领域的出版物群，即公开文档包含文本中所定义的实际上指代目标药物靶标的字符表达式。这样，公开文档的识别的群可能被确定为(作为一个整体)与目标基因相关联或不相关联。

[0103] 为了分析不同公开文档的具体引用，至少获取的一些公开文档的公开数据可包括一个公开文档对多个公开文档中一个或多个其他公开文档所作引用的引用数据。所述方法可涉及在公开数据中识别所谓的“共同引用”。共同引用可被视为出现第三个文档同时引用两个公开文档。也就是说，如果“出版物A”和“出版物B”在“出版物C”的参考列表中，则“出版物A”与“出版物B”间存在共引用。检索公开数据的步骤可包括使用所获取的引用数据识别被同一(第三)公开文档引用的(第一和第二)公开文档对。特别地，为了获得指代每个特定药物靶标的出版物群，检索公开文档对的这一步骤包括检索分别包括定义为其中一个药物靶标(基因)的至少一个一个字符表达式(基因同义词)的公开文档对。

[0104] 可以使用已识别的共同引用(即公开文档对)获得共同引用网。对于每一对已识别的公开文档，可确定代表同时引用引用公开文档对的(不同)公开文档数量的共同引用值。也就是说，可以获得一个加权的共同引用图，其中边缘的权重表示第三个出版物同时引用(共同引用)两个出版物的频率。当两份出版物被反复共引用时，假定这是两份出版物都属于同一研究领域的强烈表示。反过来，这意味着假定共同引用对中的两份出版物要么是“真正样本”，要么是“假正样本”。

[0105] 一旦建立了共同引用网，将向不同的公开文档群分配公开文档对。每个群都包括公开文档，所述公开文档包括所定义的特定药物靶标的字符表达式实例；然而，并非所有群都包括与特定药物靶标相关的公开文档，即某些群可能由字符表达实例处于与特定药物靶标不同环境的文档组成。

[0106] 因此，所述方法可包括根据其确定的共同引用值和引用这些公开文档对的公开文档，将公开文档对分配给多个公开文档群之一。这可以使用适当的群检测技术自动执行。例如，将公开文档对分配给一个群可包括(快速)贪婪优化算法的应用。

[0107] 一旦获得了一定数量的公开文档群，就需要将已识别的群与其他群区分开来。特别地，所述方法可包括，对于多个公开文档群的每一个，确定是否将该群与一个药物靶标相关联。为此，可使用特定群公开文件中出现的字符表达式(如上所述进行确定)的相对“安全性”。这可涉及确定或识别特定群中每个公开文档的公开数据中存在哪些与所定义的指代特定药物靶标的字符表达式。确定群中有多少安全字符表达式可用于确定该群是否与相关药物靶标相关。例如，可以确定群中待决公开文档的公开数据中包含至少一个安全字符表

达式的比例。如果确定的比例大于规定的阈值比例，则可决定将该群与目标药物靶标相关联。可选地，一个或多个安全字符表达式比例最高的群可被视为与相关药物靶标相关。

[0108] 上述共同引用方法可产生的一个问题是，一些公开文档的公开数据可能不包括引用数据，即特定公开文档的引用详情。这可是一个特别的问题，在这种情况下，需要获取开放获取出版物的公开数据，因为引用数据通常无法从此类来源获得。

[0109] 因此，对于公开数据不包括引用数据的每个公开文档，所述方法可包括基于其公开数据，特别是基于其公开数据中的一个或多个指代药物靶标的已定义字符表达式，确定是否将公开文档分配给与一个药物靶标相关联的一个群。例如，如果公开文档的公开数据包括至少一个安全字符表达式的实例，则可确定将公开文档分配给与相关药物靶标相关的一个群。另一方面，如果公开文档的公开数据不包括安全字符表达式，则可以使用机器学习算法，例如正向未标记学习技术，确定是否将公开文档分配给与相关药物靶标相关的一个群。所述机器学习算法可以应用机器学习分类器，如逻辑回归分类器、极限树分类器、高斯过程分类器、K近邻分类器、岭回归分类器、随机森林分类器和支持向量机分类器。也就是说，一个正向未标记装袋方法可用于训练多个分类器，使用公开数据中包含的单词/表达式(如标题、摘要等)将未连接的出版物(无引文数据)与先前的计算机共同引用网组件相关联。

[0110] 上述对获取的公开数据进行检索的步骤提供了文献中哪些公开文档与特定药物靶标(如基因)相关的准确指示。反过来，这允许对一个或多个药物靶标随时间的公开动态进行更准确和可靠的分析，例如，与给定基因相关的随时间的公开率。

[0111] 根据本发明，所述方法的下一步包括基于检索的公开数据确定每个目标药物靶标的预期公开参数和实际公开参数。特别地，预期公开参数是基于历史公开文档中的公开数据确定的。具体而言，例如，使用历史公开文档计算特定基因的历史公开动态，然后使用这些历史公开动态来确定或预测预期公开参数，例如通过外推法。作为说明性实例，可以使用历史公开数据计算给定基因连续几年的公开动态，例如使用公开数据中与历史公开文档相关的出版日期，并且可以使用这些计算的(历史)公开动态来预测该给定基因的当前公开动态。可以使用经过训练的机器学习算法，使用历史公开文档中检索到的公开数据，例如递归神经网络算法，确定预期公开参数。实际公开参数基于当前公开文档中的公开数据确定。

[0112] 预期和实际公开参数可以是与特定药物靶标相关的公开动态的任何一个或多个方面的度量或指示。例如，预期和实际公开参数可分别是预期和实际的公开文档数，例如给定年份的公开文档数量。任选地，或另外的，预期和实际公开参数可包括与待决的特定药物靶标相关的预期和实际临床试验的数量、与特定药物靶标相关的预期与实际数量的综述公开文档，以及与特定公司规模相关的预期与公开文档的实际数量。在每种情况下，需要能够在获取的公开数据中获得相关信息，以确定相关参数。例如，某些公开文档的公开数据可表明该公开文档是否与大型或中型制药公司有关。例如，如果作者隶属于大型制药公司的手稿引用了其他出版物，则这些引用可归类为“大型制药公司”引用。相反，引用该手稿的作者隶属于大型制药公司的出版物不一定被归类为"大制药公司"引用。

[0113] 根据本发明，为了检测文献中的新兴趋势，所述方法随后包括根据目标药物靶标的相对于其预期公开参数的实际公开参数进行评估，以进行选择。例如，评估药物靶标可包括根据各药物靶标的实际和预期公开参数的比较，对列表中的药物靶标(优先)排序。特别

地，如果各药物靶标的实际和预期公开参数之间存在(显著)差异，则药物靶标可被视为潜在目标以供选择，因为这可能意味着相对于根据历史公开数据可能预期的，目标药物靶点发生了阶跃性变化。

[0114] 一般来说，可以发现所述方法可以产生对公开动态的准确预测，即实际公开参数通常与预期公开参数一致。然而，对于一小部分药物靶标子集，如基因，实际或真实的出版物或引用数量可远远高于预期。当实际的出版物或引用数量超出预测时，这可被解释为所述出版物动态发生了实质性变化，无法简单地通过目标基因的公开历史来解释，例如，暗示最近可能在该领域发生了有意义的发现。术语“趋势性(trendiness)”可被定义为给定基因的预测和实际出版物和引用数量之间发生倍数变化的概率。该指标可被用于识别学术界(使用所有出版物)或制药行业(使用制药公司的出版物)的“最具趋势性(trendiest)”基因。

[0115] 所述方法可包括使用对药物靶标的评估以为选择至少一个药物靶标用于药物发现项目提供信息，例如根据趋势性基因排序表进行。特别地，所述方法可包括通过基于所述评估选择药物发现项目中使用的至少一个药物靶标，从而设计药物发现项目。所述方法可包括，至少部分基于上述评估，使用选定的一个药物靶标开展药物发现项目。这样药物发现项目可以包括选择化合物并针对至少一个选定的药物靶标进行测试，例如，鉴定对疾病靶标具有治疗活性的化合物。本公开的方法可涉及合成至少一个可对所选药物靶标具有结合活性的化合物。

[0116] 在分析特定药物靶标以进行选择时，考虑两个不同药物靶标之间的关系也是有用的。特别地，可能会发现目标基因可在关联网中聚类。对文献中一个基因与其他基因之间的关系进行考察是有用的，因为这可能意味着对一个基因的鉴定，所述基因的公开动态导致了一个或多个基因的倍数变化，所述一个或多个基因的公开动态的变化方式意味着它们是目标基因。

[0117] 在这方面，对药物靶标公开动态的分析可包括确定药物靶标对之间的靶标-靶标共现参数。该参数可以基于检索的公开数据的指示来确定，其中公开文档中的药物靶标对是相关的，即公开文献与两个不同的药物靶点相关。每个靶标-靶标共现参数可指示出现药物靶标对的发行文档的数量。可基于确定的靶标-靶标共现参数评估药物靶标以供选择。

[0118] 制药行业的潜在目标药物靶标可以是与特定疾病相关的药物靶标。所述检索与出版物相关的药物靶标的公开数据的方法也可适用于与出版物相关特定疾病。在这方面，所述方法可包括检索获取的公开数据，以为每个公开文档提供指示，说明公开文档是否分别与一个或多种疾病相关。与上述药物靶标的方法类似，这可涉及为每种疾病定义一个或多个指代上述疾病的字符表达式，检索每种疾病字符表达式的公开数据。

[0119] 作为一个非限定性的实例，疾病名称及其同义词可以从Bioportal的医学主题词表(Medical Subject Headings，MeSH)本体中获得。MeSH本体包含不同的疾病节点的4818个不同层次的本体。例如，每种疾病的字典可以用首选名称和别称创建。然后在公开数据(如标题、摘要等)中使用上述针对基因的相应技术对疾病进行明确。

[0120] 所述方法可包括确定每个药物靶标和每种疾病之间的靶标-疾病共现参数。基于从检索的公开数据中，公开文档每个药物靶标和每种疾病相关的指示，确定靶标-疾病共现参数，每个靶标-疾病共现参数指示出现所述药物靶标之一和疾病之一的公开文档的数量。

可基于确定的靶标-疾病共现参数评估药物靶标以供选择。

[0121] 为了评估药物靶点的选择，可能需要更深入地了解为什么特定药物靶点的出版物发生了变化，这也许与某种疾病有关。这样，可以对提到目标基因的出版物组进行分析。例如，所述方法可能包括对与目标药物靶标相关的公开文档的公开数据应用主题建模算法，以获得与所述药物靶标相关的一个或多个主题，然后基于获得的主题对药物靶标进行评估，以供选择。主题可以看作是一组相似词的集合，具体到一组文档。非负矩阵分解可被用于为每个查询生成潜在主题集。特别地，主题建模算法可包括潜在狄利克雷分配算法和/或非负矩阵分解算法。主题检测还可被用于根据基于获得的一个或多个主题的检索公开数据，确定对药物靶标而言，一个或多个公开文档与所述药物靶标相关性的错误，从而进一步帮助文献中药物靶标关联的准确性。

[0122] 图1总结了根据本发明的计算药物靶标选择方法10的步骤。在步骤101中，从至少一个公开数据源接收、获取或下载公开数据，例如存储公开文档(如文章、期刊论文等)的在线数据库。公开文档包括历史公开文档和当前公开文档。公开数据可以包括与公开文档相关的公开日期、作者、标题、摘要、关键词、引用等。

[0123] 在步骤102，检索接收到的公开数据，以为每个公开文档提供指示，即公开文档是否分别与一个或多个潜在药物靶标(例如基因)相关。特别地，这可涉及检索公开数据，以查找一个或多个所定义的指代每个药物靶标的字符表达式的提及或实例。在一个公开文档的公开数据中提及这些字符表达式，表明所述公开文档可能与特定潜在药物靶标有关。可以采取进一步步骤，以确定公开文档是否与潜在药物靶标相关。例如，可以建立公开数据中字符表达式的相对“安全性”(如上所述)，以表明公开数据中的字符表达式确实指代目标的潜在药物靶标。可以采取进一步步骤，根据检索的公开数据将公开文档聚类为群，以确定公开文档聚类是否实际上与目标药物靶标相关。一般来说，对公开数据进行检索被用于建立一组与一个或多个潜在药物靶标相关的公开文档。

[0124] 在步骤103中，基于历史公开文档中(即与历史公开文档相关)检索的公开数据确定每个潜在药物靶标的预期公开参数。此外，基于当前公开文档中检索的公开数据，确定了每个潜在药物靶标的实际或真实公开参数。公开参数可以是描述每个潜在药物靶标的公开动态(随时间变化)的任何合适参数。例如，公开参数可以指示与特定药物靶标相关的每个日历年度的公开文档数量。预期或预测的公开参数可以通过基于历史公开数据确定药物靶标的历史公开动态并外推这些动态来预测当前或未来的公开动态来确定。

[0125] 在步骤104，可以基于每个潜在药物靶标相对于其预期公开参数的实际公开参数对其进行评估以进行选择。特别地，给定潜在药物靶标的预期和实际参数之间的差异可能表明药物靶标的假定变化，并可能表明有兴趣进一步研究选择作为药物靶点。所述评估可能包括基于上述分析(即预测与实际值之间的差异，以及基于预测信心的差异)创建潜在药物靶标的靶标清单，以便优先选择与任何疾病或生物选择机制相关的潜在药物靶标。该评估可以为各种应用(例如设计和执行特定的药物发现项目)的药物靶点选择提供信息。

[0126] 本发明的方法可以在任何合适的计算设备上实现，例如在一个或多个计算机处理器上实现的一个或多个功能单元或模块。这些功能单元可以由使用传统或客户端处理器和存储器在任何合适的计算基板上运行的合适软件提供。一个或多个功能单元可以使用公共计算基板(例如，它们可以在同一服务器上运行)或单独的基板，或者一个或两个基板可以

分布在多个计算设备之间。计算机存储器可以存储执行所述方法的指令，所述处理器可以执行存储指令以执行所述方法。

[0127] 在不超出所附权利要求范围的情况下，可以对上述实例进行许多修改。

[0128] 在下文中，描述了上述概述的计算药物靶标选择方法的具体非限制性实施例

[0129] 实施例1：

[0130] PubMed®快照(baseline)于2019年12月公开，包含超过3000万份出版物、来自开源数据的约1.7亿条引用、近900万名作者和近3亿条MeSH注释。PubMed®可以使用图形数据库平台Neo4J高效地查询关系，如作者、参考文献或注释，将其转换为图形数据库。结果数据库包含五种不同的节点类型：出版物、作者、人类蛋白质编码基因、人类疾病以及医学主题词表术语。出版物节点具有从PubMed®快照中提取的多个属性：PubMed标识码、标题、摘要、关键词、作者、从属关系、出版日期、杂志以及文章类型(例如，文章、综述或临床试验)。还包括一个属性聚合从属关系数据，以了解制药公司是否参与了出版物的作者身份。有五种类型的关系(边)：引用(从出版物到出版物)、发表(从作者到出版物)、MeSH注释(从MeSH术语到出版物)、基因注释(从基因到出版物)及疾病注释(从疾病到出版物)。在该数据库的准备过程中，实施了一条消除歧义的方法，将人类蛋白质编码基因命名和人类疾病明确链接到单个出版物。

[0131] 从不同来源(Ensembl、UniProt、HGCN、Entrez和OpenTarget)收集了人类基因同义词，以对提及人类基因名称的潜在出版物进行采样。值得注意的是，每个人类基因平均大约有10个同义词，其中许多同义词不明确(当脱离环境考虑时)。超过30％的基因命名至少有一个混杂的同义词，大约10％的基因命名在不同的环境中有另一个含义，在英语词典中至少有一个基因同义词，并且近50％的基因命名具有嵌套同义词。综合这些问题，在19082个基因命名中，几乎60％的基因符号至少有一个类型的不明确。为了确定哪些同义词潜在不清楚，进行特征工程以获得表征不安全同义词(例如，较长的基因名称更不容易产生歧义)的变量。接下来，使用带有工程化特征的随机森林分类器的正样本未标记装袋(PU)策略来计算基因同义词"不安全"的概率。

[0132] 更具体地，使用了人类基因命名委员会(HGNC)注释的19082种蛋白质编码人类基因。删除了与医学主题词表(MeSH)数据库中包含的疾病名称相同的基因同义词。这主要发生在基因以与之相关的疾病命名时，例如“李法美尼症候群”，作为基因TP53的基因同义词，或“马凡综合征”作为“FBN1”的基因同义词。

[0133] 使用自举汇聚法(bootstrap-aggregating)的正样本未标记(PU)学习修改版本将基因同义词分类为类别“安全”或“不安全”。PU学习是一个半监督学习形式，在先验未标记数据中反复发现正样本实例。为了构建一个二进制分类器，将未标记类(U)区分为不安全(P，正样本)和安全(N，负样本)类，工程化了一系列特征，例如基因同义词中字符的组合频率(例如:“ZNF”比“EDA”更安全，因为字符“Z”和“F”在PubMed ®语料库中出现的频率低于“E”、“D”和“A”)，或对于给定基因同义词，文本中出现其他基因同义词的概率(将“STAR”给定“类固醇合成急性调节蛋白”的概率高，但将“类固醇合成急性调节蛋白”给定“STAR”的概率低，因为“STAR”更不明确)。

[0134] PU学习使用随机森林分类器进行了五次迭代。纯正样本类(不安全)是结合英语词

典中的基因同义词、少于三个字符的基因同义词和混杂的基因同义词构建的。以主动学习的方式，每次迭代后，如果分类错误，则手动重标记1000个不安全概率最高的实例。例如，真正样本不安全同义词，如基因家族(如“G蛋白偶联受体”)、表型(如“威廉姆斯综合征”)和其他生物实体(如“细胞表面抗原”)包含于下一次迭代的真正样本集中。假正样本，如“胸腺生成素”或“微管蛋白α-1C链”，被纳入剩余迭代的新真负样本类。

[0135] 五次迭代后，基因同义词被认为不安全，如果：(i)它被包含在英语词典中；(ii)它是少于三个字符的单词；(iii)随机森林分类器的预测得分高于0.5；(iv)它是一个杂乱的基因同义词。

[0136] 为了将每个人类基因与出版物的子集相联系，实施了基于共引用网络和机器学习的消除歧义的方法。使用ElasticSearch的regex字段收集了与任何同义词匹配的出版物的标题、摘要和关键词。特别地，ElasticSearch API检索引擎用于检索出版物的PubMed®标识码，所述出版物在标题、摘要或关键词中包含基因或疾病同义词。这些PubMed®标识码后来用于通过其Python驱动程序使用Cypher语言从Neo4J检索出版物的属性。正则表达式用于避免嵌套名称模糊度，使用查找和模糊匹配来解释大小写、标点符号和字母大小写的变体(例如，“ErbB-1”、“erbB1”、“ERBB1”、“ErbB1”)。

[0137] 为了检测出版物群，使用了共引用网，即加权图，其中边的权重表示第三出版物同时引用(共同引用)两个出版物的频率。来自iGraph的快速贪婪调制算法用于通过检测标题和摘要中“安全基因同义词”的存在来确定共引用网中的群和关注目标基因的出版物的不同群。如果提及至少一个安全同义词的出版物相对仅提及不安全同义词的出版物的比例高大于0.1％，则群中的每个出版物都标记有目标基因命名。

[0138] 最后，由于仅使用了PubMed Central(PMC)中包含的开放获取出版物的引用，在PubMed®中46％的出版物未连接至共引用图。提及安全同义词的未连接出版物自动链接至目标基因命名。基于链接至目标基因的群和废弃群的文本语料库(关键词、标题和摘要)中的单词，使用二进制逻辑回归分类器的PU方法装袋策略，将其余的未连接出版物链接至目标基因。使用了Scikit Learn中的所有可用机器分类器，但由于速度和准确率之间的比值，选择了逻辑回归。

[0139] 每个语料库都经过以下预处理：(i)删除非字母数字字符；(ii)通过空白符(whitespace)进行分词(tokenisation)或拆分(split)；(iii)从NLTK(自然语言工具包)中删除停止词；(iv)小写转换；(v)删除长度小于三个字符的词语(token)；(vi)删除表示整数的词语；和(vii)词干提取(例如，“消除歧义(disambiguated)”、“消除歧义(disambiguations)”、“消除歧义(disambiguating)转换为“消除歧义(disambiguat)”。使用TF IDF(词频-逆文档频率)对整个语料库中至少2个计数且频率低于0.6的词语(单字母词(gram)、双字母词、三字母词、四字母词)列表进行矢量化。当在训练集中存在少于1000份目的基因的未标记出版物时，创建辅助负样本类，以增加培训数据中的负样本实例。这种辅助负样本类包括1000份提到了与目标基因不同的基因的出版物的随机样本。

[0140] 为了测试消除歧义方法的性能，将消除歧义结果与GeneRif(手动整理的注释)、DISEASES(计算注释)和UniProt(计算和手动整理的注释)的基因出版物注释进行比较。平均而言，消除歧义方法恢复了这些数据库中85％以上的出版物。GeneRif和Uniprot注释不

一定在标题或摘要中包含基因同义词,因此这些出版物超出了所述方法。用UniProt消除歧义的结果平均准确率为70％,UniProt是唯一一个规模相似的消除歧义出版物的集合。最后,图数据库中包含了消除歧义的基因出版物注释。

[0141] 这些向量由Python库sklearn中的所有可用机器学习分类器提供:极限树分类器、高斯过程分类器、k近邻分类器、逻辑回归分类器、岭回归分类器、随机森林分类器、以及支持向量机分类器。所有分类器都经过了超参数调试和3倍交叉验证的训练,以避免在50PU-装袋迭代中出现过拟合。丢失的函数被修改以考虑类的不平衡性。考虑到消歧方法的准确性-速度平衡,选择了逻辑回归(LOG)分类器。

[0142] 用于基因实体识别的相同程序被用于检测疾病实体、共引用网和机器学习。医学主题词表(MeSH)本体是通过查询BioOntology提供的Rest-AP下载的。每种疾病都是本体中的一个节点。疾病同义词是从本体中的“概念列表术语”中获得的,以收集疾病的首选和替代解释方法。疾病的另一个同义词是通过将含逗号的同义词的顺序颠倒:“尿崩症(Insipidus,Diabetes)”到“尿崩症(Diabetes Insipidus)”。

[0143] 使用在消除歧义、并通过呈现这些标签的出版物总数进行归一化后的出版物中基因/疾病标签的同时发生,计算了基因和疾病的共现。还计算了基因-基因和基因-疾病关联的互信息度量。

[0144] 每个疾病MeSH项都与其在节点“疾病”下的MeSH本体中最低祖先相关。计算基因-疾病共现后,每个基因链接至最常见的祖先疾病项。

[0145] 为了检测文献中的未来趋势,从消除歧义的图表数据库中收集了给定人类基因的公开动态。这些时序数据包括大中型制药公司的出版物、临床试验、综述和出版物数量,以及每个日历年度上述类别的出版物的引用。

[0146] 对于大多数基因,该模型产生了对公开动态的准确预测,但对于少量基因子集,出版物或引用的实际数量远远高于预期。基因的趋势可以被视为观察给定基因的预测和实际出版物数量之间倍数变化幅度的概率。预测中的错误不可避免地更高,因为基因只与少量出版物有关。为了更正这个,基于出版物的初始数量(20％、40％、60％、80％和100％)生成了五个仓室(Bin)。使用Scikit Learn提供的高斯核密度估计器(带宽＝0.1,剩余参数为默认值)计算了五个仓室中每个的预测和观察到的实际倍数变化分布。获得概率密度函数下的面积等于1。趋势性是被观察到的倍数变化被左侧的概率密度函数右尾部限制的面积。这提供了对特定仓室中该基因倍数变化的极端程度的估计。

[0147] 使用具有编码器-解码器架构的递归神经网络模型,用1980年至2013年的时序数据预测2014年至2019年期间每个类别的每个基因公开动态,所述架构前面有一个注意层,其中编码器和解码器均由门控循环单元(GRU)的五个隐藏层组成。该模型使用Tensorflow-GPU后端在Keras中实现。最小-最大归一化用于恢复训练前的时序数据。优化器为RMSprop,损失计算为测录误差。30％的时序数据被保留用于训练期间的验证。

[0148] 输入数据采用两个形式:累积和差分。使用了多重归一化(“无”、“极小极大(Minmax)”、“对数”、“标准”及其组合)。使用不同的归一化获得了类似的结果,最终选择了极小极大。使用了编码器-解码器的形式的多个递归神经网络(RNN)架构(GRU、LSTM),具有不同数量的神经元(1、5、10、20、50)。模型与平均精度缩放误差(MASE)进行了比较,这是一个无偏性方法,通过比较每个模型在多大程度上优于重复最后值的幼稚模型来比较时序数

据预测模型。选择5个神经的GRU是因为它是最小MASE的简约模型。

[0149] 为了确定药物目标趋势基因，计算了出版物标题和摘要中基因和疾病的归一化互信息。许多趋势基因簇形成了获得基因-基因和基因-疾病关联网的趋势通路。丰富的生物过程基因本体(GO)条目被用于揭示前100种趋势性基因的共同通路。学术界和制药界最丰富的GO条目是T细胞共刺激、坏死性凋亡和凋亡的执行阶段。这些生物过程富含趋势基因，这可反映出这些研究领域在当前生物医学研究中产生了最大的创新和期望。

[0150] 在检测到基因趋势之后，下一步是了解为什么这些基因有趋势，并整理消除歧义中可能的错误。带着这一目标，实行了主题检测方法，作为自动、快速的发现工具，用于研究提及目标基因的出版物组。在这种情况下，使用了主题建模算法。主题是特定于一组文档的类似单词的集合。使用了两个不同的主题检测算法：潜在狄利克雷分配(LDA)算法和非负矩阵分解(NMF)算法。两个算法都将大小为NxM的非负矩阵“A”，其中N是出版物的数量，M是获得的用于命名实体识别的TF IDF向量的维数，分解为大小为NxK的非负因子矩阵W和大小为KxM的矩阵H，其中WxH是矩阵A的近似值。矩阵W包含给定出版物属于潜在主题的关联强度，而H包含潜在主题和给定n-字母词之间关联强度。两个算法的Scikit Learn实现均用于生成用户使用默认参数定义的主题的“k”数，直到收敛(容忍度1e-12)。主题时间线是通过计算每个日历年度所有提及目标基因的出版物的主题概率的平均值和标准偏差获得的。

[0151] 还可以设计一个综述推荐系统，以加快对网中包含大部分信息的出版物的检索。平均有2.9个综述引用了任何至少提及一个基因名称的出版物。目标是使阅读时间最小化，基因子网内信息最大化。算法汇总了主题和网络信息，以获得最以查询为中心(query-centric)的综述，所述主题和网络信息来自提及目标基因的出版物的引用子图。主题信息来自从主题检测算法获得的潜在主题。使用了出版物的主题概率和引用网的聚合页面排位(PageRank)得分。网络信息由子图的页面排位分数捕获。用户可以选择其愿意阅读的区间数的综述(R)：介于2-3或3-50之间。然后，为每组出版物定义了三个矩阵：(i)大小为NxR的二进制低秩矩阵，具有包括引用邻接网的N个出版物和R个综述；(ii)包含页面排位得分的Nx1加权矩阵；以及(iii)具有N个出版物和K个用户定义主题的主题概率的NxK矩阵。每个综述的得分被定义为其参考文献的页面排位得分之和，而综述组合的分数则定义为索引NxR矩阵的行的和乘以Nx1页面排位向量获得的向量之和。结果后来被总最大得分归一化，定义为引用所有基因出版物的假定综述。对于这种方法，最好的综述是引用具有最高页面排位得分的出版物的综述。最后，为了最小化综述的数量，发现了同时最大化累积页面排位得分并最小化其组合引用的重叠的组合。通过这种方式，可以获得涵盖所述领域主要主题和出版物的少量综述集。该推荐系统可被用于选择最佳的综述子集，以评估为什么基因可能有趋势性。

[0152] 总的来说，每个基因的出版物数量是非常可预测的。然而，有时基因呈现明显多于预期的出版物，这意味着最近发生了无法从公开动态中解释的突破。“趋势性”指标可以从基因组尺度快速分析的文献中识别新出现的靶标。趋势性与基因-疾病关联性相结合，以优先考虑潜在药物靶标：与疾病相关但尚未被纳入药物出版物的新兴基因值得作为潜在靶标进行研究。可以观察到，趋势基因通常聚类为相同的生物通路。

[0153] 总之，所述的实例方法包括从PubMed®快照下载公开数据并使用获取的信息创建图形数据库。获得人类编码基因名称和同义词的综合集合，所述方法涉及对潜在不明

确(不安全)基因名称的自动确定。所述图数据库通过结合共引用网本体和二进制分类器用明确的基因命名进行注释。所述方法涉及使用递归神经网络预测每个基因的出版趋势。当基因的出版物或引用明显多于模型的预期时,它被认为是“趋势性的”。所述方法任选地涉及出版物集合的自动主题检测,该算法被用于量化趋势基因出版物中主题随时间的演变。任选地,可以实施综述推荐系统,利用引用网和主题检测中的信息,以推荐最有效的综述集,从而探索可以被实施的文献。

[0154] 图2说明了当使用不同的上述技术或步骤时,为特定基因创建的图表数据库的实例。特别地,图2(a)说明了PubMed®中提及任何基因命名LRWD1,包括ORCA的基因同义词的公开文档自己的引用网。节点表示公开文档,节点的大小表示引用的数量。边指示文件之间的引用,包括引用方向。图2(b)说明了与图2(a)相同的公开文档子集的共同引用网。边的粗细表示共同引用文件对的次数。图2(c)说明了使用iGraph的快速贪婪算法获得的发行文档的不同群,如上所述。每个群都与不同的主题相关。例如,存在有一个被称为“虎鲸”的群201、“虎鲸植物”群或群202、“果蝇中的LRWD1”群203和“异染色质中的LRWD1”群204。图2(d)说明了同一共引用网中每个公开文档标题或摘要中的安全同义词数量。图2(e)说明了添加了综述文档的引用网,以显示综述文档对任何公开文档的引用。图2(f)说明了由综述推荐系统定义的综述信息,范围从0到1。

[0155] 图3说明了不同基因趋势和基因-基因-疾病共现的检测。特别地,图3(a)显示了2019年不同基因的预测出版物数量与实际出版物数量的对数散点图。类似地,图3(b)、3(c)和3(d)分别显示了2019年不同基因的综述文档、引用、来自“大型”制药公司的引用的预测数量与综述文档、引用、来自“大型”制药公司的引用的实际数量。例如,实际发表数量远大于预测值的基因(即,其节点位于指示对数线性关系的线上)可被视为有趋势性的和目标潜在药物靶标。

[0156] 图4说明了与不同疾病组相关的不同基因的趋势性(根据MeSH亲本类别),命名为Log_2(预测/实际)。特别地,图4(a)说明了所有(一般)公开文档的出版物、综述、引用和来自综述的引用的平均趋势,图4(b)说明了大型和中型制药公司的出版物、综述、引用和综述的引用的平均趋势。

[0157] 图5说明了CD274第一个邻居的基因-基因-疾病共现性网络。疾病和基因节点用其定义的名称标记,根据定义指标,基因节点的大小表示其“趋势性”。边表明基因-疾病和基因-基因关联,边的宽度反映了每种情况的共现次数。

[0158] 图6说明了提及不同目标基因的出版物数量的主题时间表,即探索了与一些趋势性基因相关的主题的演变。特别地,图6(a)、6(b)和6(c)分别显示了提及免疫检查点抑制剂、编程性死亡和炎性凋亡通路的任何基因的出版物的主题时间线。在每种情况下,显示四个主题时间线。潜在的四个主题是使用消除歧义后用基因注释的所有出版物的非负因子分解得出的。所有时间线都显示了2013年后主题的增加,表明了这些基因变成“趋势性的”的原因。

[0159] 参考图6(a),对于免疫检查点抑制剂(CD274、PDCD1、TGIT和CTLA4),主题时间线表明,自2010年以来,讨论这些免疫检查点抑制剂的生物学作用的主题时间线标记为601所指示的出版物的可能性迅速下降,这与标记为602的主题显著增加相吻合,这些主题讨论了癌症治疗和靶向这四种不同跨膜免疫球蛋白的单克隆抗体。通过这种方式,主题检测方法可

以捕捉从生物学描述到临床应用的研究的演变。

[0160] 参考图6(b),编程性死亡通路成员(RIPK1、RIPK3和MLKL)的主题时间线表明,在过去十年中,如主题时间线标记611所示,在细胞凋亡的背景下讨论这些基因的出版物的可能性有所下降,以支持关于新发现的细胞死亡形式(坏死性凋亡通路),及这一通路的转化医学观点的出版物,如小鼠、治疗和活性或癌症等词语所表明的(由主题时间线标记612指示)。

[0161] 参考图6(c),炎性凋亡通路成员(CGAS、TMEM173、GSDMA和GSDMAD)的主题时间线显示,自2013年以来,讨论TMEM173激动剂在癌症免疫治疗中的治疗机会的出版物迅速增加(由主题时间线标记621指示),而剩余主题似乎包含有关基因生物化学和生物学作用的信息。

[0162] 下面描述了说明上述示例方法的一些示例案例研究。

[0163] 免疫检查点抑制剂:CTLA4、CD274、PDCD1、TIGIT

[0164] CTLA4、PDCD1(PD-1)、CD274(PD-L1)和TIGIT是2019年学术界和制药界趋势性的基因。CTLA4、PDCD1、CD274和TIGIT基因编码四种不同的跨膜免疫球蛋白,作为共抑制受体:检查点或"断裂"防止T细胞发挥功能的适应性免疫反应。CTLA4与其类似物CD28竞争CD80和CD86,以防止T细胞的过早激活。PDCD1-CD274相互作用对抗可能已经激活T效应细胞的阳性信号。TIGIT与CD155相互作用下调自然杀伤细胞和T淋巴细胞。癌细胞试图破坏这些检查点,目前有7个FDA批准的单克隆抗体靶向3个蛋白(CTLA4:伊匹单抗;PDCD1:纳武单抗、派姆单抗、西米普利单抗;CD274:阿替利珠单抗、阿维单抗)和多个候选靶向TIGIT(BGB-A1217、OMP-313M32、MTIG7192A、AB154)。

[0165] 神经退行性变:TREM2和C9orf72

[0166] 最近的发现正在彻底改变对神经退行性疾病的理解。C9orf72编码参与体内运输和自噬的鸟嘌呤核苷酸交换因子。C9orf72启动子或内含子区域的六核苷酸重复扩增是肌萎缩侧索硬化和额颞叶痴呆的散发性和家族形式的主要原因之一。反义寡核苷酸被用于抑制C9orf72或CRISPR-Cas9系统的转录,以靶向DNA或RNA中的GGGGCC重复。

[0167] TREM2基因编码在巨噬细胞、破骨细胞、树突状细胞和脑小胶质细胞中表达的跨膜免疫球蛋白受体。TREM2变体与多囊性脂膜样骨发育不良并硬化性白质脑病、晚发性阿尔茨海默病、额颞叶痴呆、肌萎缩侧索硬化症和帕金森病相关。TREM2通过TYROBP/DAP12激活通路,促进促进炎症反应,促进细胞废物、凋亡细胞残余物和病原体的吞噬。目前,两个独立的组产生了抗TREM2抗体,刺激小胶质细胞去除淀粉样斑块。此外,其中一个Alenco与Abbvie合作的小组开发的单克隆抗体已进入I期临床试验。

[0168] cGAS-STING的DNA传感:cGAS、TMEM173、GSDMD、GSDMA

[0169] 细胞内核酸受体信号通路导致细胞炎性凋亡,这是一种溶解性促炎型细胞死亡,参与抗病毒、抗菌和抗癌反应。cGAS是一种核苷酸转移酶,在识别双链DNA时催化产生环鸟苷酸-腺苷酸(GAMP)。TMEM173(STING)与cGAMP结合,促进TBK1和IRF3的激活,增加编码I型干扰素基因的转录。GSDMA和GSDMD是质膜上的成孔效应蛋白,可释放IL-1β和IL-18等促炎性白细胞介素。cGAS-STING通路与非酒精性脂肪肝、系统性红斑狼疮、血管和肺综合征、黄斑变性、布卢姆综合征、Aicardi-Goutières综合征、癌症、DNA损伤、神经退行性变等多种自身免疫性和慢性炎症性疾病相关。目前,TMEM173和GSDMD的临床试验正在进行中,而GSDMA

和cGAS的临床试验尚未见报道。

[0170] 坏死性凋亡:RIPK1、RIPK3和MLKL

[0171] RIPK1、RIPK3和MLKL形成了肿瘤坏死因子诱导坏死性凋亡通路的一部分。所述通路与多种疾病相关:全身炎症反应综合征、溃疡性结肠炎、银屑病、类风湿性关节炎、神经退行性疾病甚至癌症。TNFR1、FasL、TRAIL和TLR都可以激活RIPK1决定细胞的命运:炎症、凋亡或坏死。如果半胱天冬蛋白酶-8被抑制,RIPK1和RIPK3形成坏死小体,随后磷酸化MLKL。MLKL形成同源三聚体,迁移到质膜,与高度磷酸化的磷酸肌醇结合,在膜上形成孔道,破坏细胞完整性。RIPK1的发现可以追溯到1995年。此后,4个抑制剂方案通过人体Ⅱ期安全性试验取得进展。第一个提及MLKL的出版物是最近的,尽管缺乏激酶活性,但制药公司自2013年以来引用其出版物的次数增加了60倍。虽然目前还没有临床试验,但已知的化学抑制剂至少有三种。

[0172] 力学生物学:YAP1/WWTR1、PIEZO1和PIEZO2

[0173] 细胞利用环境中的机械线索来指导增殖和迁移等行为。力作为信号传递到控制基因表达的细胞核。机械力是器官和组织稳态、形态发生和再生的关键调节因子,是癌症、转移、纤维化和心肌肥大等疾病的重要方面。YAP1/WWTR1(TAZ)是转录共激活剂和机械转接器。YAP/TAZ在癌症中过度激活,其抑制减少动脉粥样硬化和纤维化,触发肺动脉高压,并且是肠道上皮再生所必需的。PIEZO1和PIEZO2是两个机械敏感性阳离子通道,在细胞数量调节和迁移、听觉、神经和血管发育、体感功能、本体感觉等方面发挥关键作用。最近,Piezo通道被关联到多种疾病上,如关节弯曲、呼吸暂停、先天性淋巴管发育不良、痛觉过敏、疟疾、胰腺炎、干瘪细胞增多症、戈登综合征、马-沃氏综合征和5型远端关节挛缩综合征。在过去几年中,机械转导信号通路的发现受到了显著关注,并可为治疗这些疾病的新治疗策略打开大门。

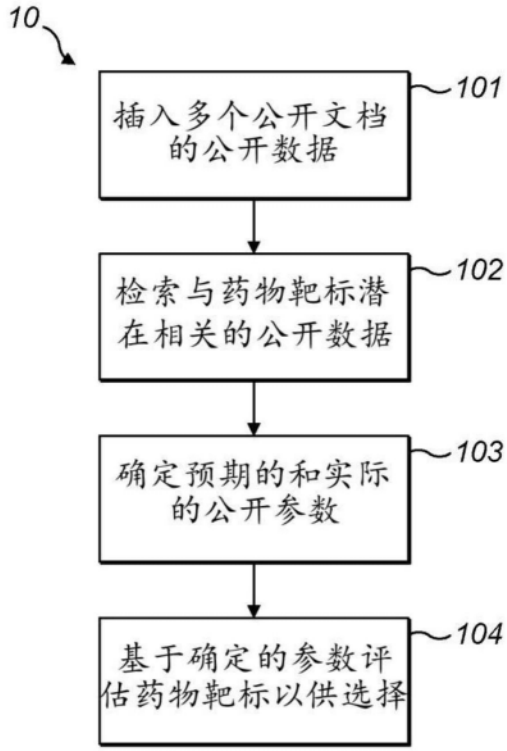

10
插入多个公开文档的公开数据
101
检索与药物靶标潜在相关的公开数据
102
确定预期的和实际的公开参数
103
基于确定的参数评估药物靶标以供选择
104

图1

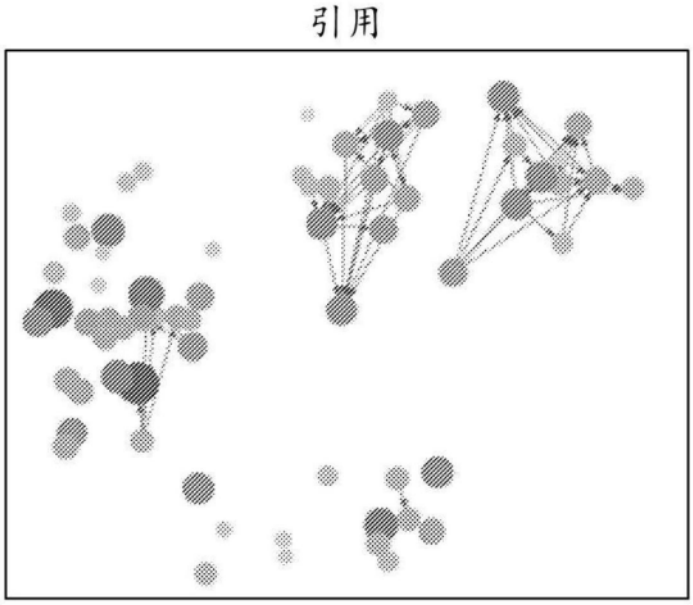

引用

图2(a)

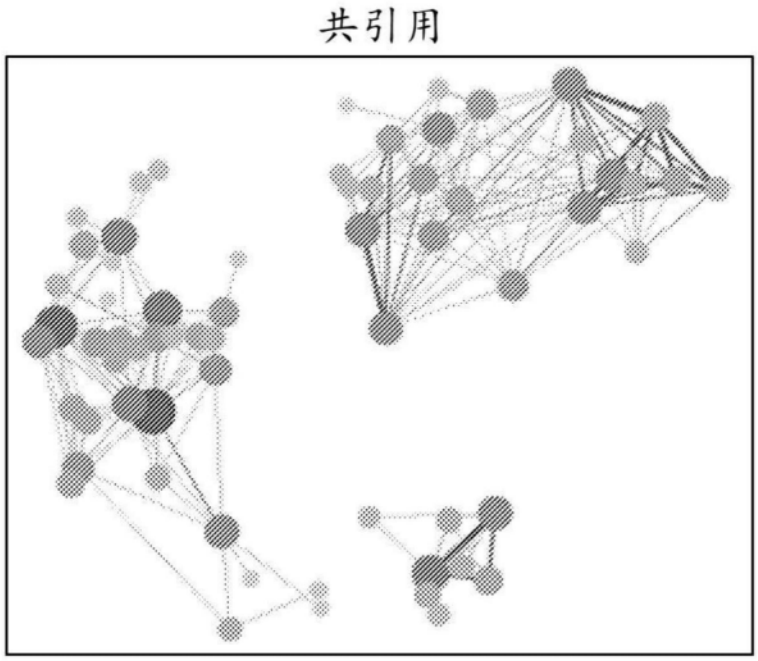


图2(b)

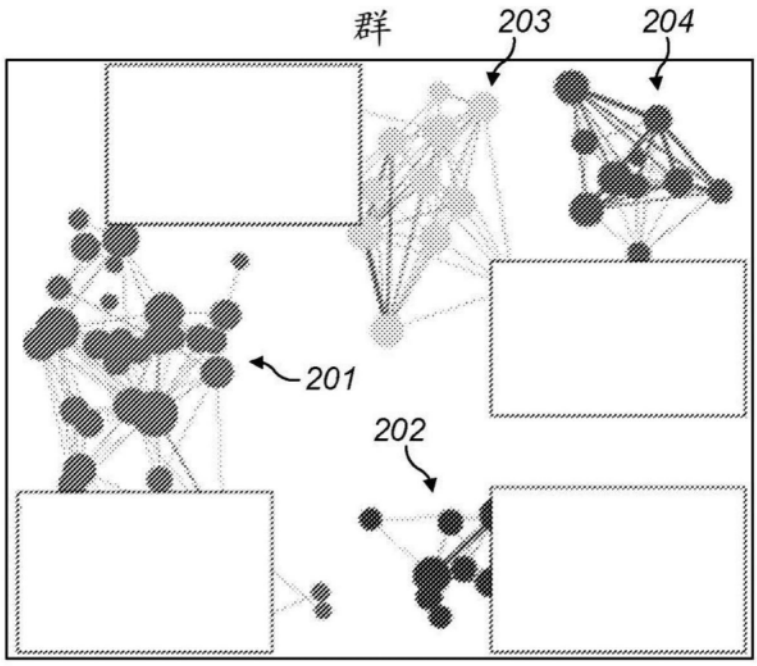


图2(c)

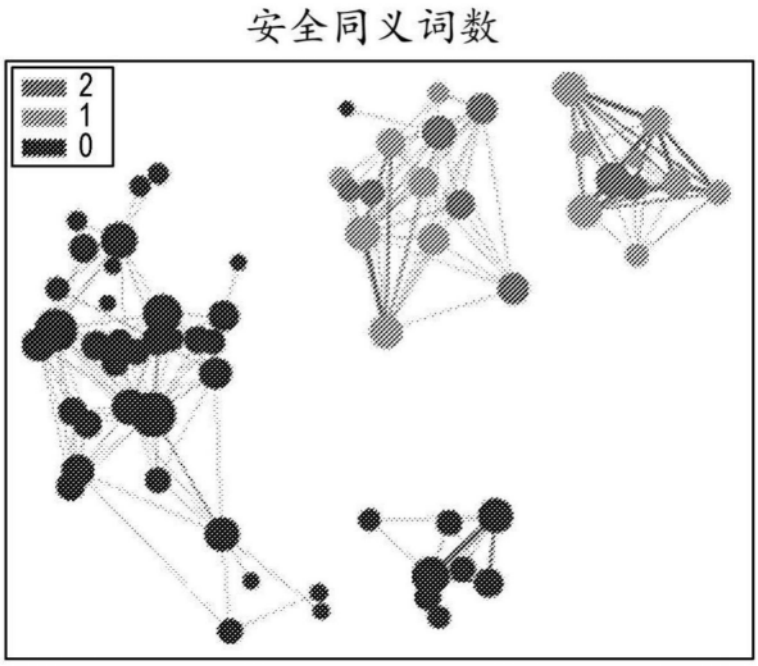


图2(d)

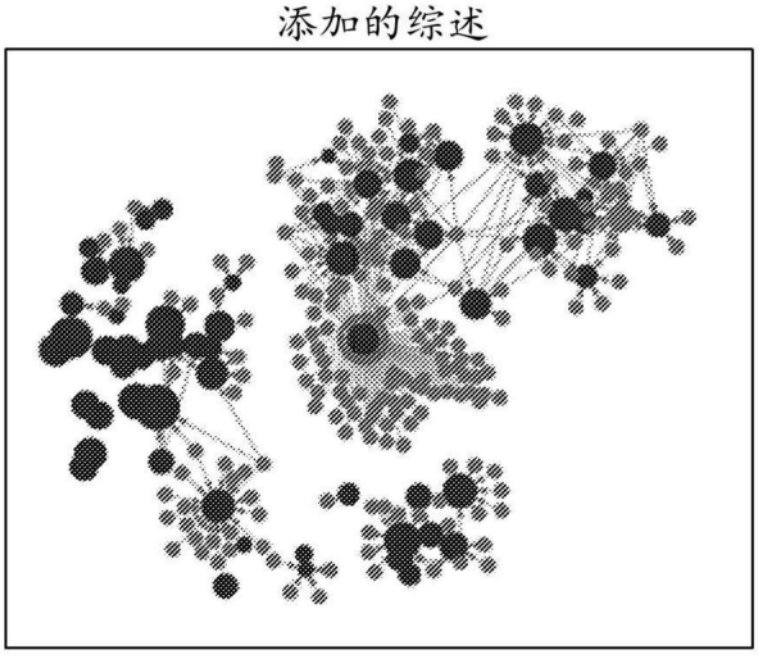


图2(e)

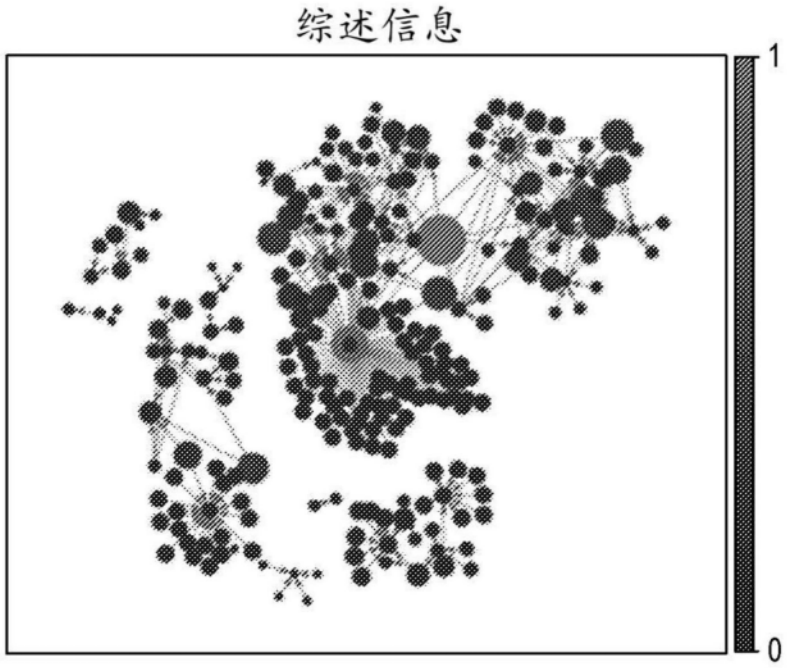

综述信息
1
0

图2(f)

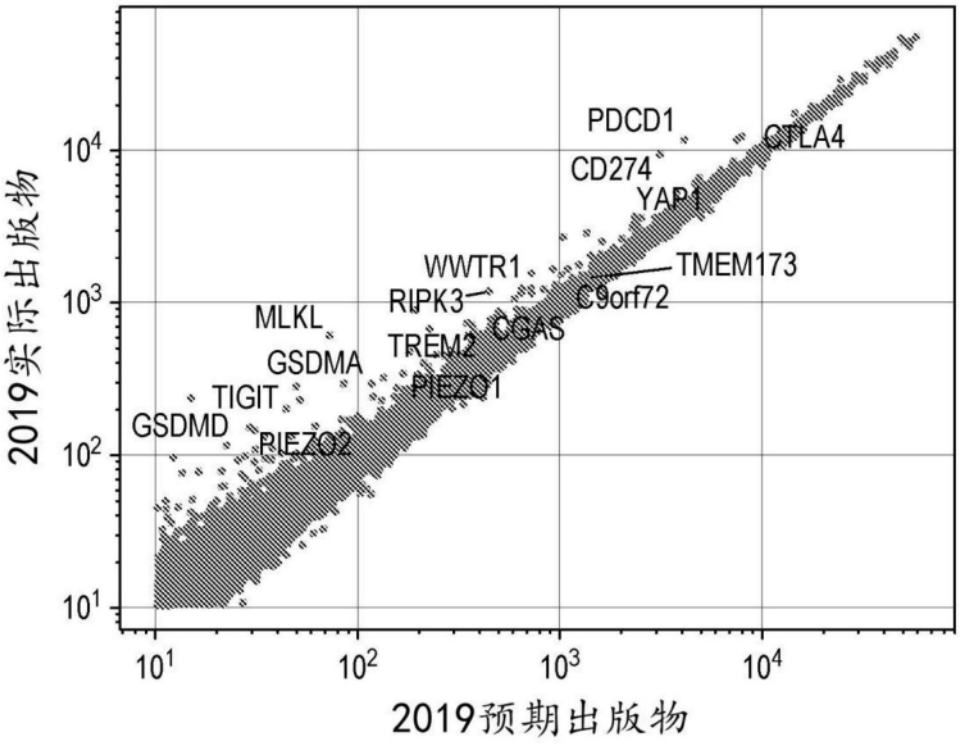

2019实际出版物
2019预期出版物
PDCD1
CTLA4
CD274
YAP1
WWTR1
TMEM173
RIPK3
C9orf72
MLKL
CGAS
TREM2
GSDMA
PIEZO1
TIGIT
GSDMD
PIEZO2

图3(a)

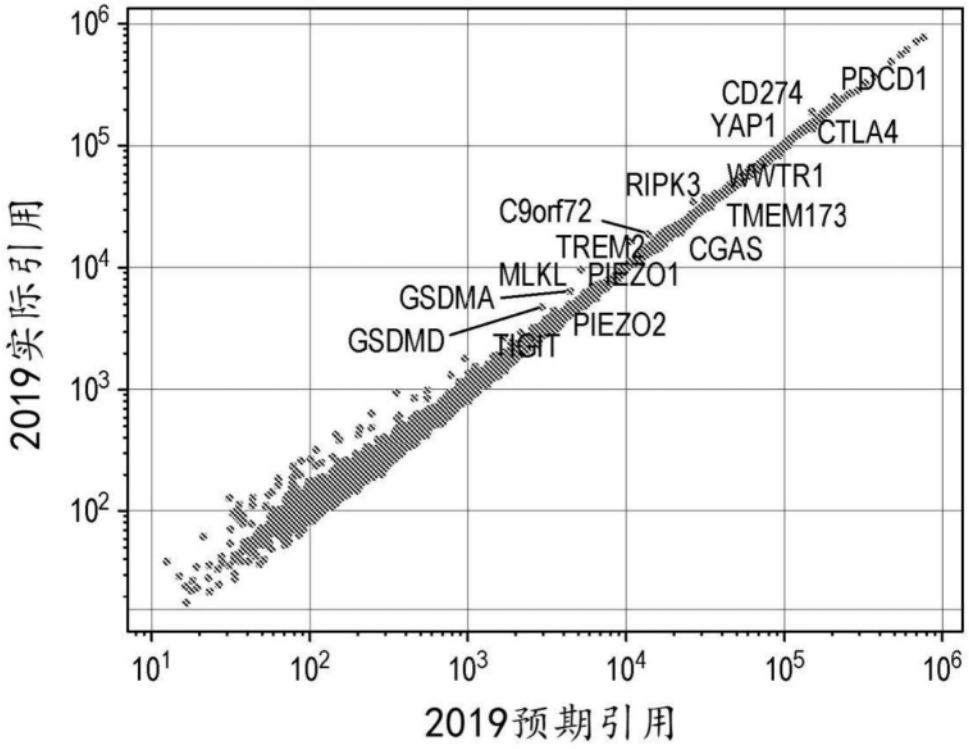

2019实际引用
2019预期引用
PDCD1
CD274
YAP1
CTLA4
RIPK3
WWTR1
C9orf72
TMEM173
TREM2
CGAS
MLKL
PIEZO1
GSDMA
PIEZO2
GSDMD
TIGIT

图3(b)

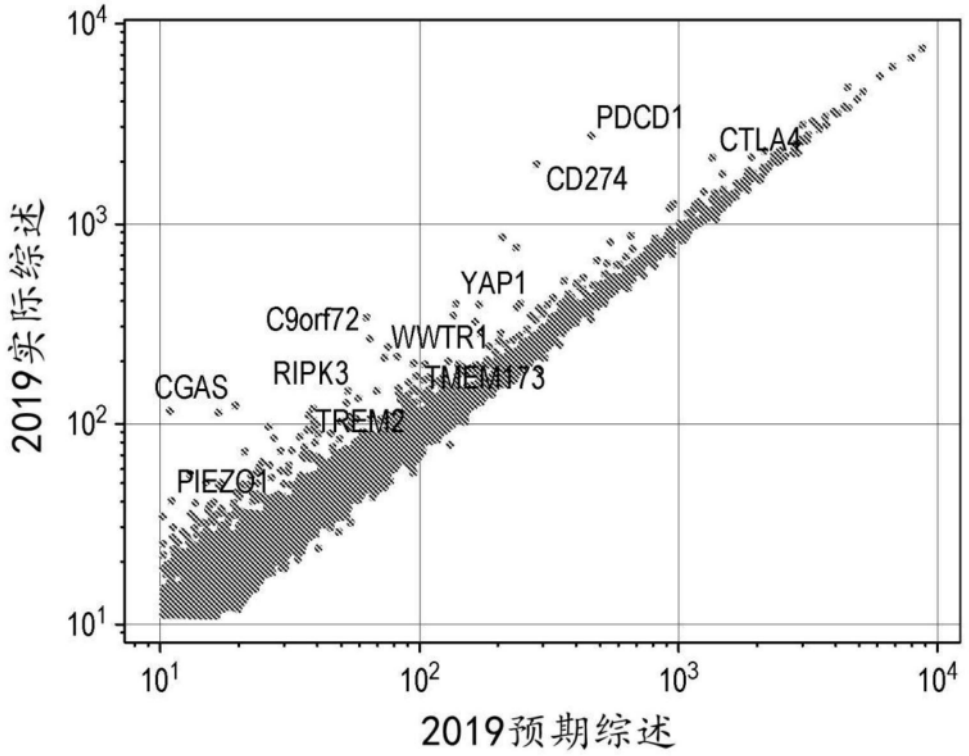

2019实际综述
2019预期综述
PDCD1
CTLA4
CD274
YAP1
C9orf72
WWTR1
RIPK3
TMEM173
CGAS
TREM2
PIEZO1

图3(c)

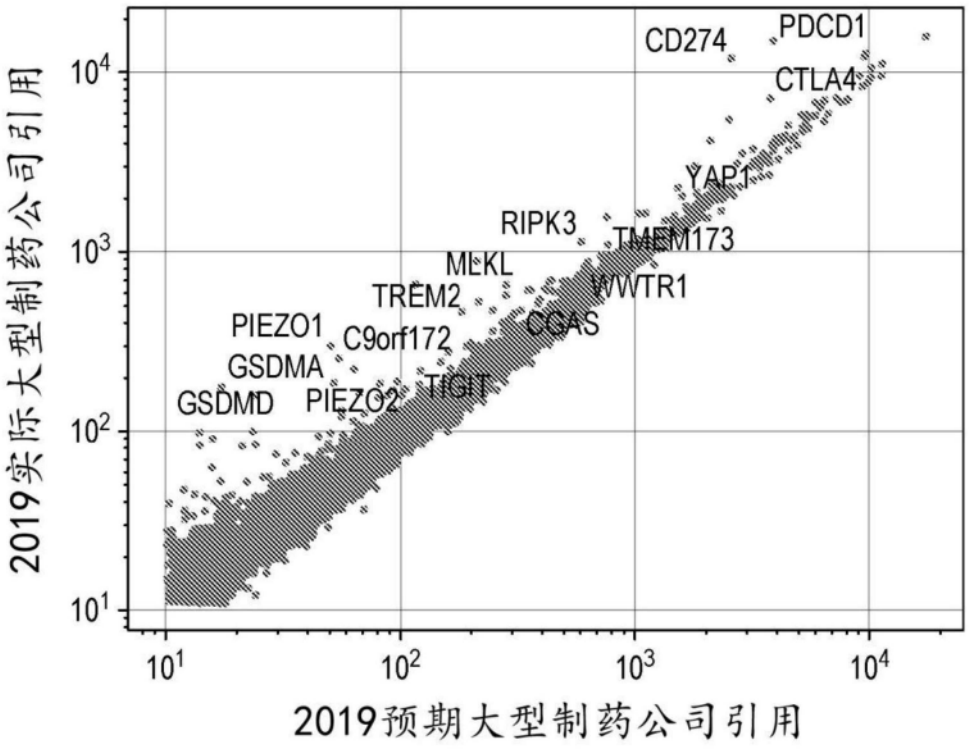

CD274
PDCD1
CTLA4
YAP1
RIPK3
TMEM173
MLKL
TREM2
WWTR1
CGAS
PIEZO1
C9orf172
GSDMA
TIGIT
GSDMD
PIEZO2
10^4
10^3
10^2
10^1
2019实际大型制药公司引用
2019预期大型制药公司引用

图3(d)

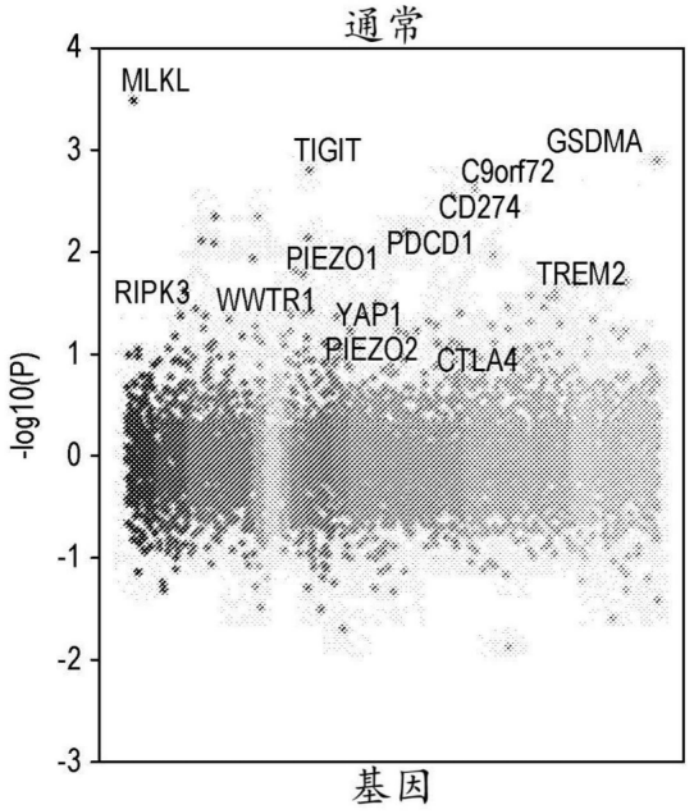

通常
MLKL
TIGIT
GSDMA
C9orf72
CD274
PDCD1
PIEZO1
TREM2
RIPK3
WWTR1
YAP1
PIEZO2
CTLA4
4
3
2
1
0
-1
-2
-3
-log10(P)
基因

图4(a)

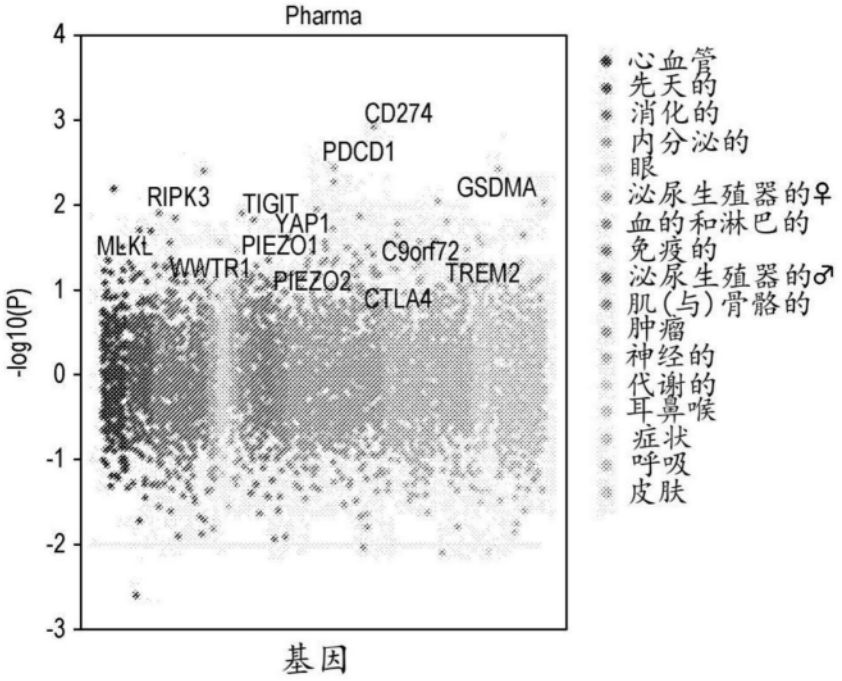

Pharma
-log10(P)
基因
CD274
PDCD1
GSDMA
RIPK3
TIGIT
YAP1
MLKL
PIEZO1
C9orf72
WWTR1
PIEZO2
TREM2
CTLA4
心血管
先天的
消化的
内分泌的
眼
泌尿生殖器的♀
血的和淋巴的
免疫的
泌尿生殖器的♂
肌(与)骨骼的
肿瘤
神经的
代谢的
耳鼻喉
症状
呼吸
皮肤

图4(b)

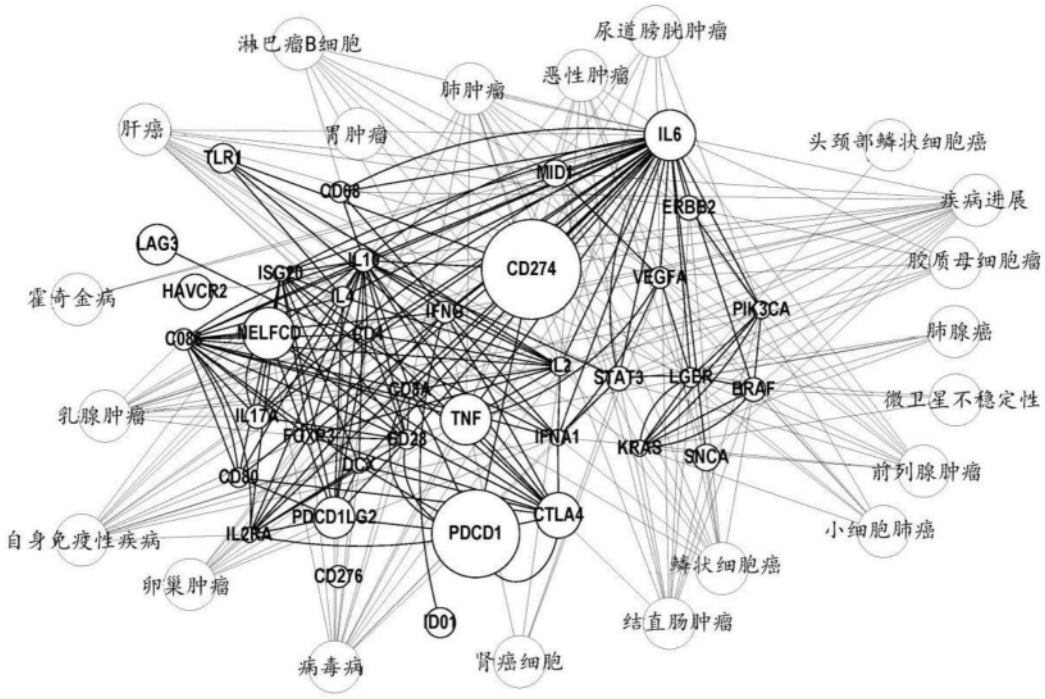

淋巴瘤B细胞
尿道膀胱肿瘤
肺肿瘤
恶性肿瘤
肝癌
胃肿瘤
IL6
头颈部鳞状细胞癌
TLR1
MIDI
CD96
ERBB2
疾病进展
LAG3
CD274
胶质母细胞瘤
霍奇金病
HAVCR2
ISG20
IL10
VEGFA
IL4
IFNG
PIK3CA
肺腺癌
CD8A
NELFCD
CD4
IL2
STAT3
LGER
BRAF
微卫星不稳定性
乳腺肿瘤
IL17A
CD8A
TNF
IFNA1
KRAS
SNCA
CD80
DCK
前列腺肿瘤
自身免疫性疾病
PDCD1LG2
PDCD1
CTLA4
小细胞肺癌
IL2RA
鳞状细胞癌
卵巢肿瘤
CD276
结直肠肿瘤
IDO1
病毒病
肾癌细胞

图5

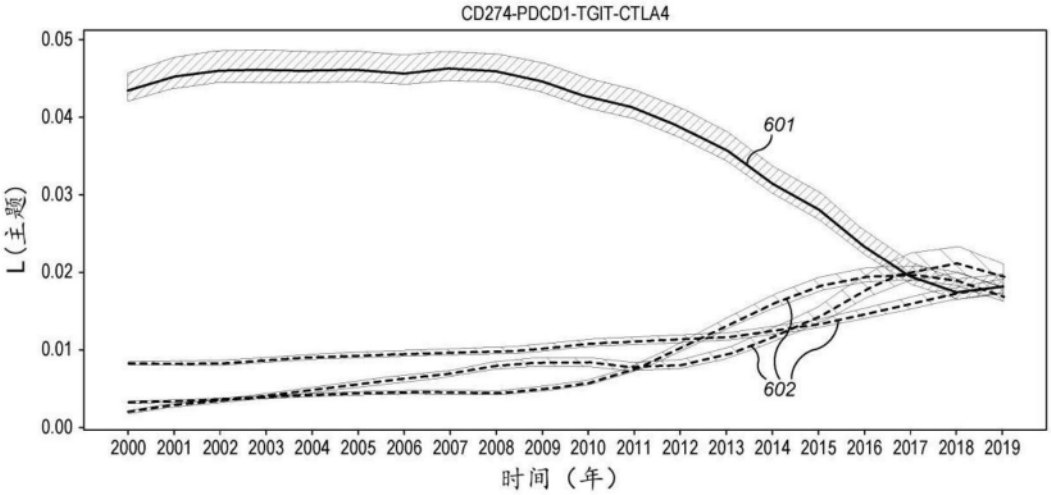

CD274-PDCD1-TGIT-CTLA4
L(主题)
0.05
0.04
0.03
0.02
0.01
0.00
601
602
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
时间（年）

图6(a)

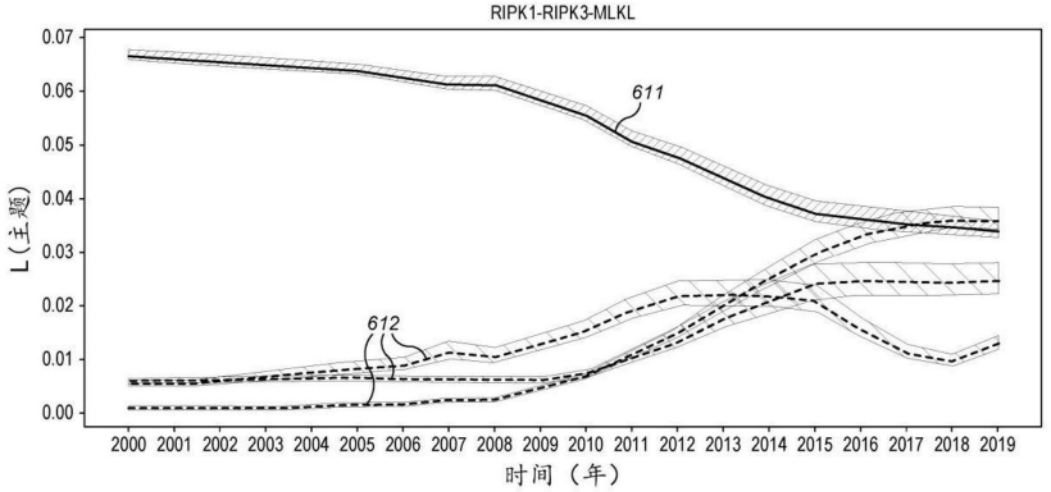

RIPK1-RIPK3-MLKL
L(主题)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
611
612
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
时间（年）

图6(b)

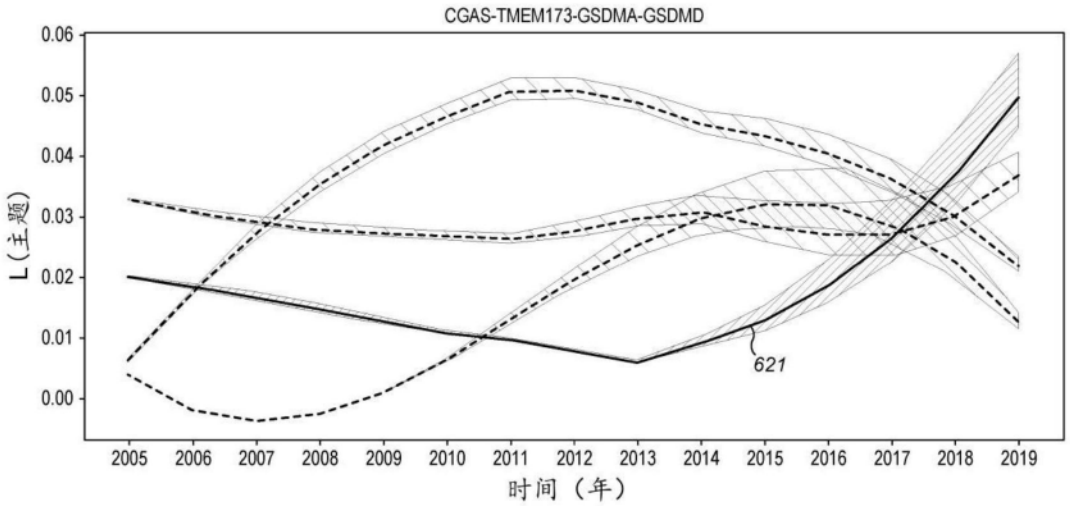

CGAS-TMEM173-GSDMA-GSDMD
0.06
0.05
0.04
0.03
0.02
0.01
0.00
L(主题)
621
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
时间（年）

图6(c)