(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117616505 A
(43)申请公布日 2024.02.27

(21)申请号 202280048777.8
(22)申请日 2022.06.15
(30)优先权数据
63/210,930 2021.06.15 US
(85)PCT国际申请进入国家阶段日
2024.01.09
(86)PCT国际申请的申请数据
PCT/US2022/033685 2022.06.15
(87)PCT国际申请的公布数据
WO2022/266259 EN 2022.12.22
(71)申请人 旗舰先锋创新VI有限责任公司
地址 美国马萨诸塞州
(72)发明人 F·A·沃尔夫 R·哈达德 N·M·普拉吉斯
(74)专利代理机构 深圳市百瑞专利商标事务所(普通合伙) 44240
专利代理师 金辉
(51)Int.Cl.
G16B 15/30(2006.01)
G16B 40/20(2006.01)
G16B 25/10(2006.01)
G16B 5/00(2006.01)

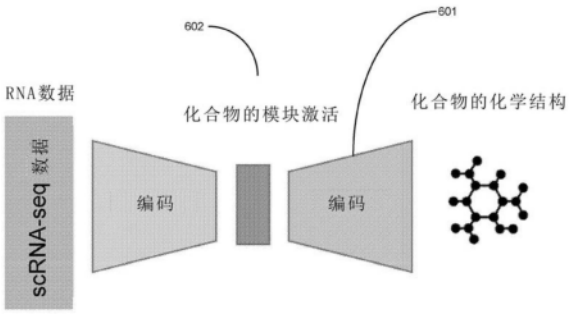
权利要求书18页 说明书83页 附图26页

(54)发明名称

用于使用指纹分析将化合物与生理状况相关联的系统和方法

(57)摘要

本发明提供了用于将化合物与生理状况相关联的系统和方法。获得化合物化学结构的指纹并将其输入到输出一个或多个计算出的激活评分的模型。每个激活评分表示模块的集合中的细胞组分模块,其中每个模块包括细胞组分的子集,并且所述模块的集合中的第一模块与所述生理状况相关联。当针对所述第一模块的所述激活评分满足阈值标准时,所述化合物被识别为与所述生理状况相关联。在一些方面,每个激活评分表示与所述生理状况相关联的扰动特征,并且当针对第一扰动特征的所述激活评分满足阈值标准时,识别出所述化合物。本发明还提供了用于训练将化合物与生理状况相关联的模型的系统和方法。



CN 117616505 A

1.一种将测试化学化合物与目的生理状况相关联的方法,所述方法包括:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取细胞组分模块的集合,其中

所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的相应的独立子集,

针对所述多种细胞组分的每个相应的独立子集的对应的多个基于细胞的测定丰度值跨与所述生理状况相关联的多种不同状态而单独相关,并且

所述细胞组分模块的集合中的第一细胞组分模块与所述目的生理状况相关联;

(C)响应于将所述化学结构的所述指纹输入到模型中,其中所述模型包括100个或更多个参数,检索针对所述细胞组分模块的集合中的每个细胞组分模块的相应的激活评分,作为来自所述模型的输出;以及

(D)当针对所述第一细胞组分模块的所述激活评分满足第一阈值标准时,将所述测试化学化合物与所述目的生理状况相关联。

2.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为器官的细胞的测定丰度值。

3.根据权利要求2所述的方法,其中所述器官为心脏、肝脏、肺部、肌肉、脑、胰腺、脾脏、肾脏、小肠、子宫或膀胱。

4.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为组织的细胞的测定丰度值。

5.根据权利要求4所述的方法,其中所述组织为骨骼、软骨、关节、气管、脊髓、角膜、眼睛、皮肤或血管。

6.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为多个干细胞中的细胞的测定丰度值。

7.根据权利要求6所述的方法,其中所述多个干细胞为多个胚胎干细胞、多个成体干细胞或多个诱导性多能干细胞(iPSC)。

8.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为多个原代人细胞中的细胞的测定丰度值。

9.根据权利要求8所述的方法,其中所述多个原代人细胞为多个CD34+细胞、多个CD34+造血干细胞、多个祖细胞(HSPC)、多个T细胞、多个间充质干细胞(MSC)、多个气道基底干细胞或多个诱导性多能干细胞。

10.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为脐带血中、外周血中或骨髓中的细胞的测定丰度值。

11.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为实体组织中的细胞的测定丰度值。

12.根据权利要求11所述的方法,其中所述实体组织为胎盘、肝脏、心脏、脑、肾脏或胃肠道。

13.根据权利要求1所述的方法,其中所述基于细胞的测定丰度值为多个分化细胞的测定丰度值。

14.根据权利要求13所述的方法,其中所述多个分化细胞为多个巨核细胞、多个成骨细

胞、多个软骨细胞、多个脂肪细胞、多个肝细胞、多个肝间皮细胞、多个胆管上皮细胞、多个肝星状细胞、多个肝窦内皮细胞、多个库普弗细胞、多个隐窝细胞、多个血管内皮细胞、多个胰管上皮细胞、多个胰管细胞、多个腺腔中心细胞、多个腺泡细胞、多个朗格尔汉斯小岛、多个心肌细胞、多个纤维母细胞、多个角质形成细胞、多个平滑肌细胞、多个I型肺泡上皮细胞、多个II型肺泡上皮细胞、多个克拉拉细胞、多个纤毛上皮细胞、多个基底细胞、多个杯状细胞、多个神经内分泌细胞、多个库尔契茨基细胞、多个肾小管上皮细胞、多个尿路上皮细胞、多个柱状上皮细胞、多个肾小球上皮细胞、多个肾小球内皮细胞、多个足细胞、多个血管系膜细胞、多个神经细胞、多个星形胶质细胞、多个小胶质细胞或多个少突胶质细胞。

15.根据权利要求1至14中任一项所述的方法,其中所述对应的多个基于细胞的测定丰度值为多个细胞的单细胞核糖核酸(RNA)测序(scRNA-seq)数据。

16.根据权利要求15所述的方法,其中通过将细胞的不同等分试样暴露于一种或多种已知影响所述生理状况的参考化合物而得到与所述生理状况相关联的所述多种不同状态,此外还得到对照状态,在所述对照状态下,细胞的等分试样并未免于暴露于已知影响所述生理状况的化合物。

17.根据权利要求1至14中任一项所述的方法,其中所述对应的多个基于细胞的测定丰度值来自大量RNA序列。

18.根据权利要求1至14中任一项所述的方法,其中所述对应的多个基于细胞的测定丰度值来自单细胞RNA测序。

19.根据权利要求1至18中任一项所述的方法,其中所述细胞组分模块的集合由所述第一细胞组分模块组成。

20.根据权利要求1至18中任一项所述的方法,其中所述细胞组分模块的集合包括多个细胞组分模块并且所述模型为包括多个成分模型的集成模型,并且其中所述多个成分模型中的每个成分模型响应于将所述化学结构的所述指纹输入到所述多个成分模型中的每个成分模型中而提供针对不同细胞组分模块的激活评分。

21.根据权利要求1至20中任一项所述的方法,所述方法进一步包括根据所述测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算所述指纹。

22.根据权利要求20或21所述的方法,其中所述多个成分模型中的每个成分模型为对应的神经网络。

23.根据权利要求22所述的方法,其中所述对应的神经网络为全连接神经网络、消息传递神经网络或其组合。

24.根据权利要求20或21所述的方法,其中所述多个成分模型中的成分模型为逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

25.根据权利要求22所述的方法,其中

所述对应的神经网络为对应的全连接神经网络和对应的消息传递神经网络的组合,

响应于将所述化学结构的所述指纹输入到所述对应的全连接神经网络和所述对应的消息传递神经网络中,将所述对应的全连接神经网络的第一输出和所述对应的消息传递神经网络的第二输出组合,以确定一个或多个计算出的激活评分中的针对所述细胞组分模块的集合中的对应的细胞组分模块的激活评分。

26.根据权利要求1所述的方法,其中

所述细胞组分模块的集合为多个细胞组分模块,

包括所述第一细胞组分模块的所述多个细胞组分模块的第一子集与所述目的生理状况相关联,

所述多个细胞组分模块的第二子集与所述目的生理状况不关联,并且

当针对所述第一细胞组分模块的相应的计算出的激活评分满足所述第一阈值标准并且针对所述多个细胞组分模块的所述第二子集中的细胞组分模块的相应的计算出的激活评分满足所述第一阈值标准之外的第二阈值标准时,所述测试化学化合物与所述目的生理状况关连。

27.根据权利要求1至26中任一项所述的方法,所述方法进一步包括通过包括以下的过程来识别所述第一细胞组分模块:

以电子形式获得一个或多个第一数据集,所述一个或多个第一数据集包括或共同包括:

对于第一多个细胞中的每个相应的细胞,其中所述第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态:

对于所述多种细胞组分中的每种相应的细胞组分,其中所述多种细胞组分包括10种或更多种细胞组分:

所述相应的细胞组分在所述相应的细胞中的对应的丰度,

由此获取或形成多个向量,所述多个向量中的每个相应的向量(i)对应于所述多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素,所述对应的多个元素中的每个相应的元素具有对应的计数,所述对应的计数表示所述相应的细胞组分在所述第一多个细胞中的所述相应的细胞中的所述对应的丰度;

使用所述多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块,所述多个候选细胞组分模块中的每个候选细胞组分模块包括所述多种细胞组分的子集,其中所述多个细胞组分模块布置在由(i)所述多个候选细胞组分模块和(ii)所述多种细胞组分或其表示来确定维度的潜在表示中,并且其中所述多个细胞组分模块包括多于十个细胞组分模块;

以电子形式获得一个或多个第二数据集,所述一个或多个第二数据集包括或共同包括:

对于第二多个细胞中的每个相应的细胞,其中所述第二多个细胞包括二十个或更多个细胞并且共同表示提供所述目的生理状况的信息的多个协变量:

对于所述多种细胞组分中的每种相应的细胞组分:

所述相应的细胞组分在所述相应的细胞中的对应的丰度,

由此获得由(i)所述第二多个细胞和(ii)所述多种细胞组分或其所述表示来确定维度的细胞组分计数数据结构;

通过以下来形成激活数据结构:使用所述多种细胞组分或其所述表示作为公共维度来组合所述细胞组分计数数据结构和所述潜在表示,其中所述激活数据结构包括:对于所述多个细胞组分模块中的每个细胞组分模块,

对于所述第二多个细胞中的每个细胞,相应的激活权重;以及

针对所述多个协变量中的每个相应的协变量,使用以下两者之间的差异来训练候选细胞组分模型:(i)在将所述协变量的指纹输入到所述候选细胞组分模型中时针对由所述候选细胞组分模型表示的每个细胞组分模块的计算出的激活,以及(ii)针对由所述候选细胞组分模型表示的每个细胞组分模块的实际激活,其中所述训练响应于所述差异来调整与所述候选细胞组分模型相关联的多个协变量参数。

28.根据权利要求27所述的方法,其中所述多个协变量参数包括:

对于所述多个细胞组分模块中的每个相应的细胞组分模块:

对于每个相应的协变量:

对应的参数,所述对应的参数指示所述相应的协变量是否跨所述第二多个细胞与所述相应的细胞组分模块相关;并且所述方法进一步包括:

在训练所述候选细胞组分模型时使用所述多个协变量参数来识别所述多个候选细胞组分模块中的所述第一细胞组分模块。

29.根据权利要求27或28所述的方法,其中所述多种经注释的细胞状态中的经注释的细胞状态为所述第一多个细胞中的细胞在暴露条件下暴露于化合物。

30.根据权利要求29所述的方法,其中所述暴露条件为暴露的持续时间、所述化合物的浓度或暴露的持续时间与所述化合物的浓度的组合。

31.根据权利要求1至30中任一项所述的方法,其中所述多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。

32.根据权利要求27至30中任一项所述的方法,其中

所述多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合,并且

所述相应的细胞组分在所述第一或第二多个细胞中的所述相应的细胞中的所述对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。

33.根据权利要求11所述的方法,其中

所述多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合,并且

所述相应的细胞组分在所述第一或第二多个细胞中的所述相应的细胞中的所述对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq或其任何组合。

34.根据权利要求1至30或32至33中任一项所述的方法,其中使用所述多个向量来识别所述多个候选细胞组分模块中的每个候选细胞组分模块包括使用所述多个向量中的每个向量的每组对应的多个元素来将相关模型应用于所述多个向量。

35.根据权利要求34所述的方法,其中所述相关模型包括图聚类。

36.根据权利要求34所述的方法,其中所述图聚类为基于皮尔逊相关的距离度量上的莱顿聚类。

37.根据权利要求34所述的方法,其中所述图聚类为鲁汶聚类。

38.根据权利要求27至37中任一项所述的方法,其中所述多个细胞组分模块由介于10

个与2000个之间的细胞组分模块组成。

39.根据权利要求27至37中任一项所述的方法,其中所述多种细胞组分由介于100种与8,000种之间的细胞组分组成。

40.根据权利要求27至37中任一项所述的方法,其中多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

41.根据权利要求1至40中任一项所述的方法,其中所述目的生理状况为疾病。

42.根据权利要求27所述的方法,其中所述目的生理状况为疾病,并且所述第一多个细胞包括表示所述疾病的细胞和不表示所述疾病的细胞,如由所述多种经注释的细胞状态所指示。

43.根据权利要求27所述的方法,其中所述多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态、暴露于化学化合物或其任何组合。

44.根据权利要求27所述的方法,其中所述训练所述候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中所述多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且所述多个成本函数中的每个相应的成本函数具有公共的权重因子。

45.根据权利要求1至44中任一项所述的方法,其中所述测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。

46.根据权利要求45所述的方法,其中所述测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。

47.根据权利要求45所述的方法,其中所述测试化学化合物为满足里宾斯基五规则标准中的至少三个标准的有机化合物。

48.根据权利要求1至19中任一项所述的方法,其中所述模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

49.根据权利要求1至48中任一项所述的方法,所述方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从所述测试化学化合物的化学结构生成所述指纹。

50.根据权利要求1至18或20至49中任一项所述的方法,其中所述细胞组分模块的集合包括五个或更多个细胞组分模块。

51.根据权利要求1至18或20至50中任一项所述的方法,其中所述细胞组分模块的集合包括十个或更多个细胞组分模块。

52.根据权利要求1至18或20至50中任一项所述的方法,其中所述细胞组分模块的集合包括100个或更多个细胞组分模块。

53.根据权利要求1至52中任一项所述的方法,其中所述相应的细胞组分模块中的所述多种细胞组分的所述独立子集包括五种或更多种细胞组分。

54.根据权利要求1至52中任一项所述的方法,其中所述相应的细胞组分模块中的所述多种细胞组分的所述独立子集由与所述目的生理状况相关联的分子途径中的介于两种与20种之间的细胞组分组成。

55.根据权利要求1至54中任一项所述的方法,其中所述第一阈值标准为以下要求:所

述第一细胞组分模块具有阈值激活评分。

56.一种计算机系统,其包括一个或多个处理器以及存储器,所述存储器存储用于进行用于将测试化学化合物与目的生理状况相关联的方法的指令,所述方法包括:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取细胞组分模块的集合,其中

所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的相应的独立子集,

针对所述多种细胞组分的每个相应的独立子集的对应的多个基于细胞的测定丰度值跨与所述生理状况相关联的多种不同状态而单独相关,并且

所述细胞组分模块的集合中的第一细胞组分模块与所述目的生理状况相关联;

(C)响应于将所述化学结构的所述指纹输入到模型中,其中所述模型包括100个或更多个参数,检索针对所述细胞组分模块的集合中的每个细胞组分模块的相应的激活评分,作为来自所述模型的输出;以及

(D)当针对所述第一细胞组分模块的所述激活评分满足第一阈值标准时,将所述测试化学化合物与所述目的生理状况相关联。

57.一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将测试化学化合物与目的生理状况相关联,所述计算机包括一个或多个处理器以及存储器,所述一个或多个计算机程序共同编码进行包括以下的方法的计算机可执行指令:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取细胞组分模块的集合,其中

所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的相应的独立子集,

针对所述多种细胞组分的每个相应的独立子集的对应的多个基于细胞的测定丰度值跨与所述生理状况相关联的多种不同状态而单独相关,并且

所述细胞组分模块的集合中的第一细胞组分模块与所述目的生理状况相关联;

(C)响应于将所述化学结构的所述指纹输入到模型中,其中所述模型包括100个或更多个参数,检索针对所述细胞组分模块的集合中的每个细胞组分模块的相应的激活评分,作为来自所述模型的输出;以及

(D)当针对所述第一细胞组分模块的所述激活评分满足第一阈值标准时,将所述测试化学化合物与所述目的生理状况相关联。

58.一种将测试化学化合物与目的生理状况相关联的方法,所述方法包括:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取扰动特征的集合,其中

所述扰动特征的集合中的每个相应的扰动特征包括多种细胞组分的相应的独立子集,

所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞

状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态;

(C)将所述指纹输入到模型中,其中

所述模型包括100个或更多个参数,

所述模型响应于将所述指纹输入到所述模型中而输出一个或多个计算出的激活评分,

所述一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示所述扰动特征的集合中的对应的扰动特征;以及

(D)当针对所述扰动特征的集合中的第一扰动特征的所述相应的计算出的激活评分满足第一阈值标准时,将所述化学化合物与所述目的生理状况相关联。

59.根据权利要求58所述的方法,所述方法进一步包括根据所述测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算所述指纹。

60.根据权利要求58或59所述的方法,其中所述模型包括神经网络。

61.根据权利要求60所述的方法,其中所述神经网络为全连接神经网络、消息传递神经网络或其组合。

62.根据权利要求58至61中任一项所述的方法,其中所述模型为包括多个成分模型的集成模型,并且其中所述多个成分模型中的每个成分模型响应于将所述化学结构的所述指纹输入到多个成分模型的集合中的每个成分模型中而提供针对所述扰动特征的集合中的不同的扰动特征的激活评分。

63.根据权利要求62所述的方法,其中所述多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

64.根据权利要求62或63所述的方法,其中所述多个成分模型中的每个成分模型为对应的神经网络。

65.根据权利要求64所述的方法,其中所述对应的神经网络为全连接神经网络、消息传递神经网络或其组合。

66.根据权利要求63或64所述的方法,其中所述多个成分模型中的成分模型为逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

67.根据权利要求65所述的方法,其中

所述对应的神经网络为全连接神经网络和消息传递神经网络的组合,并且

响应于将所述化学结构的所述指纹输入到所述全连接神经网络和所述消息传递神经网络中,将第一神经网络的第一输出和第二神经网络的第二输出组合,以确定所述一个或多个计算出的激活评分中的针对所述扰动特征的集合中的第一扰动特征的激活评分。

68.根据权利要求58所述的方法,其中

所述扰动特征的集合为多个扰动特征,

包括所述第一扰动特征的所述多个扰动特征的第一子集与所述目的生理状况相关联,

所述多个扰动特征的第二子集与所述目的生理状况不关联,并且

当针对所述第一扰动特征的所述相应的计算出的激活评分满足所述第一阈值标准并且针对所述多个扰动特征的所述第二子集中的扰动特征的所述相应的计算出的激活评分满足所述第一阈值标准之外的第二阈值标准时,所述测试化学化合物与所述目的生理状况

关连。

69.根据权利要求58至68中任一项所述的方法,其中所述目的生理状况为疾病。

70.根据权利要求58所述的方法,其中所述测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。

71.根据权利要求70所述的方法,其中所述测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。

72.根据权利要求70所述的方法,其中所述测试化学化合物为满足里宾斯基五规则标准中的至少三个标准的有机化合物。

73.根据权利要求58所述的方法,其中所述模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

74.根据权利要求58至73中任一项所述的方法,所述方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从所述测试化学化合物的化学结构生成所述指纹。

75.根据权利要求58至74中任一项所述的方法,其中所述扰动特征的集合由所述第一扰动特征组成。

76.根据权利要求58至74中任一项所述的方法,其中所述扰动特征的集合包括五个或更多个扰动特征。

77.根据权利要求58至74中任一项所述的方法,其中所述扰动特征的集合包括十个或更多个扰动特征。

78.根据权利要求58至74中任一项所述的方法,其中所述扰动特征的集合包括100个或更多个扰动特征。

79.根据权利要求58至74中任一项所述的方法,其中所述第一阈值标准为以下要求:所述第一扰动特征具有阈值激活评分。

80.一种计算机系统,其包括一个或多个处理器以及存储器,所述存储器存储用于进行用于将测试化学化合物与目的生理状况相关联的方法的指令,所述方法包括:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取扰动特征的集合,其中

所述扰动特征的集合中的每个相应的扰动特征包括多种细胞组分的相应的独立子集,

所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态;

(C)将所述指纹输入到模型中,其中

所述模型包括100个或更多个参数,

所述模型响应于将所述指纹输入到所述模型中而输出一个或多个计算出的激活评分,

所述一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示所述扰动

特征的集合中的对应的扰动特征;以及

(D)当针对所述扰动特征的集合中的第一扰动特征的所述相应的计算出的激活评分满足第一阈值标准时,将所述化学化合物与所述目的生理状况相关联。

81.一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将测试化学化合物与目的生理状况相关联,所述计算机包括一个或多个处理器以及存储器,所述一个或多个计算机程序共同编码进行包括以下的方法的计算机可执行指令:

(A)获得所述测试化学化合物的化学结构的指纹;

(B)获取扰动特征的集合,其中

所述扰动特征的集合中的每个相应的扰动特征包括多种细胞组分的相应的独立子集,

所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态;

(C)将所述指纹输入到模型中,其中

所述模型包括100个或更多个参数,

所述模型响应于将所述指纹输入到所述模型中而输出一个或多个计算出的激活评分,

所述一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示所述扰动特征的集合中的对应的扰动特征;以及

(D)当针对所述扰动特征的集合中的第一扰动特征的所述相应的计算出的激活评分满足第一阈值标准时,将所述化学化合物与所述目的生理状况相关联。

82.一种将化学化合物与目的生理状况相关联的方法,所述方法包括:

在包括存储器和一个或多个处理器的计算机系统处:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对所述多种化合物中的每种化合物的相应的数值激活评分,其中所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构,

对于所述细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的细胞组分模块的相应的计算出的激活评分,以及(ii)所述细胞组分模块的集合中的所述相应的细胞组分模块针对所述相应的化合物的所述相应的数值激活评分,其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数,并且其中所述多个参数包括100个或更多个参数,从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

83.根据权利要求82所述的方法,其中所述细胞组分模块的集合由单个细胞组分模块组成。

84.根据权利要求82所述的方法,其中所述细胞组分模块的集合包括多个细胞组分模块。

85.根据权利要求82所述的方法,其中所述细胞组分模块的集合由介于二百个与五百个之间的细胞组分模块组成。

86.根据权利要求82所述的方法,其中所述多种化合物由介于10种与1×10^{6}种之间的化合物组成。

87.根据权利要求82所述的方法,其中所述多种化合物由介于100种与100,000种之间的化合物组成。

88.根据权利要求82所述的方法,其中所述多种化合物由介于1000种与100,000种之间的化合物组成。

89.根据权利要求82至88中任一项所述的方法,其中所述训练(C)根据回归算法响应于与每种相应的化合物相关联的针对所述细胞组分模块的集合中的每个相应的细胞组分模块的每个差异而调整与所述未经训练的模型相关联的所述多个参数。

90.根据权利要求89所述的方法,其中所述回归算法优化与每种相应的化合物相关联的针对所述细胞组分模块的集合中的每个相应的细胞组分模块的每个差异的最小二乘误差。

91.根据权利要求82至90中任一项所述的方法,其中所述经训练的模型包括神经网络。

92.根据权利要求91所述的方法,其中所述神经网络为全连接神经网络、消息传递神经网络或其组合。

93.根据权利要求82至90中任一项所述的方法,其中所述经训练的模型为多个成分模型的集成模型,并且所述多个成分模型中的每个相应的成分模型输出针对所述多个细胞组分模块中的不同细胞组分模块的计算出的激活评分。

94.根据权利要求93所述的方法,其中所述多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

95.根据权利要求93所述的方法,其中所述多个成分模型中的每个成分模型为对应的神经网络。

96.根据权利要求95所述的方法,其中所述对应的神经网络为全连接神经网络、消息传递神经网络或其组合。

97.根据权利要求82至96中任一项所述的方法,其中

所述细胞组分模块的集合为多个细胞组分模块,

所述多个细胞组分模块的第一子集与所述目的生理状况相关联,并且

所述多个细胞组分模块的第二子集与所述目的生理状况不关联。

98.根据权利要求82至97中任一项所述的方法,所述方法进一步包括通过包括以下的过程来识别所述多个细胞组分模块中的细胞组分模块:

以电子形式获得一个或多个第一数据集,所述一个或多个第一数据集包括或共同包括:

对于第一多个细胞中的每个相应的细胞,其中所述第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态:

对于所述多种细胞组分中的每种相应的细胞组分,其中所述多种细胞组分包括10种或更多种细胞组分:

所述相应的细胞组分在所述相应的细胞中的对应的丰度,

由此获取或形成多个向量,所述多个向量中的每个相应的向量(i)对应于所述多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素,所述对应的多个元素中的每个相应的元素具有对应的计数,所述对应的计数表示所述相应的细胞组分在所述第一多个细胞中的所述相应的细胞中的所述对应的丰度;

使用所述多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块,所述多个候选细胞组分模块中的每个候选细胞组分模块包括所述多种细胞组分的子集,其中所述多个细胞组分模块布置在由(i)所述多个候选细胞组分模块和(ii)所述多种细胞组分或其表示来确定维度的潜在表示中,并且其中所述多个细胞组分模块包括多于十个细胞组分模块;

以电子形式获得一个或多个第二数据集,所述一个或多个第二数据集包括或共同包括:

对于第二多个细胞中的每个相应的细胞,其中所述第二多个细胞包括二十个或更多个细胞并且共同表示提供所述目的生理状况的信息的多个协变量:

对于所述多种细胞组分中的每种相应的细胞组分:

所述相应的细胞组分在所述相应的细胞中的对应的丰度,

由此获得由(i)所述第二多个细胞和(ii)所述多种细胞组分或其所述表示来确定维度的细胞组分计数数据结构;

通过以下来形成激活数据结构:使用所述多种细胞组分或其所述表示作为公共维度来组合所述细胞组分计数数据结构和所述潜在表示,其中所述激活数据结构包括:对于所述多个细胞组分模块中的每个细胞组分模块,

对于所述第二多个细胞中的每个细胞,相应的激活权重;

使用以下两者之间的差异来训练候选细胞组分模型:(i)在将所述激活数据结构输入到候选模型中时对所述多个协变量中的每个协变量在表示于所述激活数据结构中的每个细胞组分模块中的不存在或存在的预测,以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在,其中所述训练响应于所述差异而调整与所述候选细胞成分模型相关联的多个协变量参数。

99.根据权利要求98所述的方法,其中所述多个协变量参数包括:

对于所述多个细胞组分模块中的每个相应的细胞组分模块:

对于每个相应的协变量:

对应的参数,所述对应的参数指示所述相应的协变量是否跨所述第二多个细胞与所述相应的细胞组分模块相关;以及

在训练所述候选细胞组分模型时使用所述多个协变量参数来识别所述多个候选细胞组分模块中的细胞组分模块。

100.根据权利要求99所述的方法,其中所述多种经注释的细胞状态中的经注释的细胞

状态为所述第一多个细胞中的细胞在暴露条件下暴露于化合物。

101.根据权利要求99所述的方法,其中所述暴露条件为暴露的持续时间、所述化合物的浓度或暴露的持续时间与所述化合物的浓度的组合。

102.根据权利要求82至101中任一项所述的方法,其中所述多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。

103.根据权利要求98所述的方法,其中所述相应的细胞组分在所述第一或第二多个细胞中的所述相应的细胞中的所述对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。

104.根据权利要求98所述的方法,其中所述相应的细胞组分在所述第一或第二多个细胞中的所述相应的细胞中的所述对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq或其任何组合。

105.根据权利要求98所述的方法,其中使用所述多个向量来识别所述多个候选细胞组分模块中的每个候选细胞组分模块包括使用所述多个向量中的每个向量的每组对应的多个元素来将相关模型应用于所述多个向量。

106.根据权利要求105所述的方法,其中所述相关模型包括图聚类。

107.根据权利要求106所述的方法,其中图聚类方法为基于皮尔逊相关的距离度量上的莱顿聚类,或者为鲁汶聚类。

108.根据权利要求82至107中任一项所述的方法,其中所述多种细胞组分由介于100种与8,000种之间的细胞组分组成。

109.根据权利要求98所述的方法,其中多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

110.根据权利要求82至109中任一项所述的方法,其中所述目的生理状况为疾病。

111.根据权利要求98中任一项所述的方法,其中所述生理状况为疾病,并且所述第一多个细胞包括表示所述疾病的细胞和不表示所述疾病的细胞,如由所述多种经注释的细胞状态所指示。

112.根据权利要求98所述的方法,其中所述多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。

113.根据权利要求98所述的方法,其中所述训练所述候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中所述多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且所述多个成本函数中的每个相应的成本函数具有公共的权重因子。

114.根据权利要求82至113中任一项所述的方法,其中多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。

115.根据权利要求82至113中任一项所述的方法,其中多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。

116.根据权利要求82至113中任一项所述的方法,其中多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。

117.根据权利要求82至116中任一项所述的方法,其中所述经训练的模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

118.根据权利要求82至117中任一项所述的方法,所述方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从所述对应的化学结构生成每个相应的指纹。

119.根据权利要求82所述的方法,其中所述细胞组分模块的集合包括五个或更多个细胞组分模块。

120.根据权利要求82所述的方法,其中所述细胞组分模块的集合包括十个或更多个细胞组分模块。

121.根据权利要求82所述的方法,其中所述细胞组分模块的集合包括100个或更多个细胞组分模块。

122.一种计算机系统,其包括一个或多个处理器以及存储器,所述存储器存储用于进行用于将化学化合物与目的生理状况相关联的方法的指令,所述方法包括:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对所述多种化合物中的每种化合物的相应的数值激活评分,其中所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构,

对于所述细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的细胞组分模块的相应的计算出的激活评分,以及(ii)所述细胞组分模块的集合中的所述相应的细胞组分模块针对所述相应的化合物的所述相应的数值激活评分,其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数,并且其中所述多个参数包括100个或更多个参数,从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

123.一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将化学化合物与目的生理状况相关联,所述计算机包括一个或多个处理器以及存储器,所述一个或多个计算机程序共同编码进行包括以下的方法的计算机可执行指令:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对所述多种化合物中的每种化合物的相应的数值激活评分,其中所述细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构,

对于所述细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的细胞组分模块的相应的计算出的激活评分,以及(ii)所述细胞组分模块的集合中的所述相应的细胞组分模块针对所述相应的化合物的所述相应的数值激活评分,其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数,并且其中所述多个参数包括100个或更多个参数,从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

124.一种将化学化合物与目的生理状况相关联的方法,所述方法包括:

在包括存储器和一个或多个处理器的计算机系统处:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得扰动特征的集合中的每个相应的扰动特征针对所述多种化合物中的每种对应的化合物的相应的数值激活评分,其中所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于所述对应的化合物引起的相应的受扰动的细胞状态;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构,

对于所述扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的扰动特征的相应的计算出的激活评分,以及(ii)所述扰动特征的集合中的所述相应的扰动特征针对所述对应的化合物的所述相应的数值激活评分,其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数,并且其中所述多个参数包括100个或更多个参数,从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

125.根据权利要求124所述的方法,其中所述扰动特征的集合由单个扰动特征组成。

126.根据权利要求124所述的方法,其中所述扰动特征的集合由介于二百个与五百个之间的扰动特征组成。

127.根据权利要求124至126中任一项所述的方法,其中所述多种化合物由介于10种与1×10^{6}种之间的化合物组成。

128.根据权利要求124至126中任一项所述的方法,其中所述多种化合物由介于100种与100,000种之间的化合物组成。

129.根据权利要求124至126中任一项所述的方法,其中所述多种化合物由介于1000种与100,000种之间的化合物组成。

130.根据权利要求124至129中任一项所述的方法,其中所述训练(C)根据回归算法响应于与每种对应的化合物相关联的针对所述扰动特征的集合中的每个相应的扰动特征的每个差异而调整与所述未经训练的模型相关联的所述多个参数。

131.根据权利要求130所述的方法,其中所述回归算法优化与每种对应的化合物相关联的针对所述扰动特征的集合中的每个相应的扰动特征的每个差异的最小二乘误差。

132.根据权利要求124至131中任一项所述的方法,其中所述经训练的模型包括神经网络。

133.根据权利要求132所述的方法,其中所述神经网络为全连接神经网络、消息传递神经网络或其组合。

134.根据权利要求124所述的方法,其中所述经训练的模型为多个成分模型的集成模型,并且所述多个成分模型中的每个相应的成分模型响应于将相应的化学结构的指纹输入到所述多个成分模型的集合中的每个成分模型中而输出针对多个扰动特征的集合中的不同的扰动特征的集合的计算出的激活评分。

135.根据权利要求134所述的方法,其中所述多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

136.根据权利要求134所述的方法,其中所述多个成分模型中的每个成分模型为对应的神经网络。

137.根据权利要求136所述的方法,其中所述对应的神经网络为全连接神经网络、消息传递神经网络或其组合。

138.根据权利要求124至137中任一项所述的方法,其中

所述扰动特征的集合包括多个扰动特征,

所述多个扰动特征的第一子集与所述目的生理状况关联,并且

所述多个扰动特征的第二子集与所述目的生理状况不关联。

139.根据权利要求124至138中任一项所述的方法,其中所述目的生理状况为疾病。

140.根据权利要求124至139中任一项所述的方法,其中多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。

141.根据权利要求124至140中任一项所述的方法,其中多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。

142.根据权利要求124至140中任一项所述的方法,其中多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。

143.根据权利要求124所述的方法,其中所述经训练的模型包括逻辑回归模型、神经网络模型、支持向量机、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

144.根据权利要求124至143中任一项所述的方法,所述方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从所述对应的化学结构生成每个相应的指纹。

145.根据权利要求124所述的方法,其中所述扰动特征的集合包括五个或更多个扰动特征。

146.根据权利要求124所述的方法,其中所述扰动特征的集合包括十个或更多个扰动特征。

147.根据权利要求124所述的方法,其中所述扰动特征的集合包括100个或更多个扰动特征。

148.根据权利要求124所述的方法,所述方法进一步包括通过包括以下的程序来获得所述扰动特征的集合中的相应的扰动特征的相应的数值激活评分:

以电子形式获取单细胞转变特征,所述单细胞转变特征表示未改变的细胞状态与改变的细胞状态之间的差异细胞组分丰度的测度,其中

所述改变的细胞状态通过从所述未改变的细胞状态到所述改变的细胞状态的细胞转变而出现,

(i)所述未改变的细胞状态、(ii)所述改变的细胞状态以及(iii)从所述未改变的细胞状态到所述改变的细胞状态的所述转变中的至少一者与所述目的生理状况相关联,并且

所述单细胞转变特征包括参考多种细胞组分的标识以及针对多种参考细胞组分中的每种相应的细胞组分的对应的第一显著性评分,所述对应的第一显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及所述未改变的细胞状态与所述改变的细胞状态之间的细胞状态变化;以及

比较所述单细胞转变特征和所述相应的扰动特征,由此确定所述相应的扰动特征的所述相应的数值激活评分。

149.根据权利要求148所述的方法,其中比较所述单细胞转变特征和所述扰动特征以确定所述相应的扰动特征的所述相应的数值激活评分包括:针对有所述单细胞转变特征的所述参考多种细胞组分中的每种相应的细胞组分,

将所述相应的细胞组分的所述第一显著性评分与对应的细胞组分在所述相应的扰动特征中的所述对应的显著性评分进行比较。

150.根据权利要求148或149所述的方法,其中所述相应的扰动特征的所述激活评分为所述相应的扰动特征相对于所述扰动特征的集合中的其他扰动特征与所述单细胞转变特征的相关性的相对排名。

151.根据权利要求150所述的方法,其中所述相对排名通过Wilcoxon秩和检验、t检验、逻辑回归或广义线性模型来确定。

152.根据权利要求148至151中任一项所述的方法,其中所述单细胞转变特征的所述未改变的细胞状态与所述相应的扰动特征的所述第一细胞状态或所述第二细胞状态相同。

153.根据权利要求148至151中任一项所述的方法,其中所述单细胞转变特征的所述未改变的细胞状态与所述相应的扰动特征的所述第一细胞状态和所述第二细胞状态两者均不同。

154.根据权利要求148至153中任一项所述的方法,所述方法进一步包括:

修剪有所述单细胞转变特征的所述参考多种细胞组分以及有所述相应的扰动特征的所述相应的多种细胞组分以限制与转录因子的比较。

155.根据权利要求124至154中任一项所述的方法,其中所述多个扰动特征中的相应的扰动特征的所述受扰动的细胞状态由尚未暴露于所述多种化合物中的化合物的对照细胞表示。

156.根据权利要求124至154中任一项所述的方法,其中所述多个扰动特征中的相应的扰动特征的所述受扰动的细胞状态由跨已经暴露于所述多种化学化合物中除了与所述相应的扰动特征相关联的化合物之外的化学化合物的不相关的受扰动的细胞的平均数表示。

157.一种计算机系统,其包括一个或多个处理器以及存储器,所述存储器存储用于将化学化合物与目的生理状况相关联的指令,方法包括:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得扰动特征的集合中的每个相应的扰动特征针对所述多种化合物中的每种对应的化合物的相应的数值激活评分,其中所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于所述对应的化合物引起的相应的受扰动的细胞状态;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构,

对于所述扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的扰动特征的相应的计算出的激活评分,以及(ii)所述扰动特征的集合中的所述相应的扰动特征针对所述对应的化合物的所述相应的数值激活评分,其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数,并且其中所述多个参数包括100个或更多个参数,从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

158.一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将化学化合物与目的生理状况相关联,所述计算机包括一个或多个处理器以及存储器,所述一个或多个计算机程序共同编码进行包括以下的方法的计算机可执行指令:

(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,从而获得多个指纹;

(B)以电子形式获得扰动特征的集合中的每个相应的扰动特征针对所述多种化合物中的每种对应的化合物的相应的数值激活评分,其中所述扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对所述相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,所述对应的显著性评分量化以下两者之间的关联:所述相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中所述相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且所述相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于所述对应的化合物引起的相应的受扰动的细胞状态;以及

(C)训练未经训练的模型

对于所述多种化合物中的每种相应的化合物的每个相应的化学结构，

对于所述扰动特征的集合中的每个相应的扰动特征，使用以下两者之间的相应的差异来进行：

(i)在将所述相应的化合物的所述化学结构的所述指纹输入到所述未经训练的模型中时针对所述相应的扰动特征的相应的计算出的激活评分，以及(ii)所述扰动特征的集合中的所述相应的扰动特征针对所述对应的化合物的所述相应的数值激活评分，其中所述训练(C)响应于所述差异而调整与所述未经训练的模型相关联的多个参数，并且其中所述多个参数包括100个或更多个参数，从而获得将化学化合物与所述目的生理状况相关联的经训练的模型。

159.根据前述权利要求中任一项所述的方法，其中所述模型为回归器。

用于使用指纹分析将化合物与生理状况相关联的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求2021年6月15日提交的名称为“SYSTEMS AND METHODS FOR ASSOCIATING COMPOUNDS WITH PHYSIOLOGICAL CONDITIONS USING FINGERPRINT ANALYSIS”的美国临时专利申请号63/210,930以及2021年6月15日提交的名称为“COMPUTATIONAL MODELING PLATFORM”的美国临时专利申请号63/210,679的优先权,两件申请中的每一件特此通过引用以其整体并入。

技术领域

[0003] 本发明总体涉及用于将化合物与生理状况相关联的系统和方法。

背景技术

[0004] 对细胞机制的研究对于理解疾病来说很重要。

[0005] 生物组织是动态且高度网络化的多细胞系统。特定细胞中的亚细胞网络的功能障碍会改变细胞行为的整体形势并导致疾病状态。现有的药物发现努力试图表征导致细胞从健康状态转变为疾病状态的分子机制,并鉴定逆转或抑制这些转变的药理学方法。过去的努力还寻求鉴定表征这些转变的分子签名,以及鉴定逆转这些签名的药理学方法。

[0006] 关于通过表面标志物富集的组织或细胞中的大量细胞集合的分子数据会掩盖群体中的各个细胞的表型和分子多样性。这些大量细胞集合中的细胞的异质性致使当前旨在阐明疾病驱使机制的努力的结果具有误导性,或者甚至完全不正确。新方法诸如单细胞RNA测序可以在分子水平上表征各个细胞。这些数据为以较高分辨力理解不同的细胞状态提供了基础,并揭示了细胞所拥有的丰富且显着的状态多样性。

[0007] 当解释单细胞数据时存在重大挑战,即这些数据的稀疏性(忽略存在于细胞中的分子的存在)和噪声,这些分子测量的准确度具有不确定性。因此,需要新方法来得到对用于控制单种细胞状态的药理学方法的深刻理解,并对应地解决疾病。

[0008] 此外,通常不能将复杂的疾病分解为单个或几个分子靶点。尽管针对体外疾病模型的高通量成像技术和高通量筛选最近取得进展,但将从基于体外的筛选方法产生的候选靶点转化为有功效的药物是艰巨的任务,该任务通常涉及回到相对缓慢和低效的基于分子靶点的药物发现方法。

[0009] 鉴于上述背景,本领域所需要的是用于识别用于药物发现的候选化合物的系统和方法。

发明内容

[0010] 本公开解决了上述缺点。本公开用与目的生理状况(例如,目的表型、疾病、细胞状态和/或细胞过程)相对应的细胞组分数据(例如,基因的丰度和/或扰动特征)来至少部分地解决这些缺点,并使用潜在表示和机器学习来确定细胞组分的模块(例如,子集)与目的生理状况之间的关联(例如,权重和/或相关)。特别地,本公开提供了用于阐明各种生理状

况(诸如疾病)背后的分子机制的系统和方法。

[0011] 本公开的一方面提供了一种将测试化学化合物与目的生理状况相关联的方法。该方法包括(A)获得测试化学化合物的化学结构的指纹。

[0012] 该方法进一步包括(B)获取细胞组分模块的集合。细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的相应的独立子集。针对多种细胞组分的每个相应的独立子集的对应的多个基于细胞的测定丰度值跨与生理状况相关联的多种不同状态而单独相关。细胞组分模块的集合中的第一细胞组分模块与目的生理状况相关联。

[0013] 该方法进一步包括(C)响应于将化学结构的指纹输入到模型中,检索针对细胞组分模块的集合中的每个细胞组分模块的相应的激活评分,作为来自该模型的输出。在一些实施例中,模型包括50个或更多个参数、100个或更多个参数、1000个或更多个参数,或10,000个或更多个参数。

[0014] 该方法进一步包括(D)当针对第一细胞组分模块的激活评分满足第一阈值标准时,将测试化学化合物与目的生理状况相关联。

[0015] 在一些实施例中,基于细胞的测定丰度值为器官的细胞的测定丰度值。在一些此类实施例中,器官为心脏、肝脏、肺部、肌肉、脑、胰腺、脾脏、肾脏、小肠、子宫或膀胱。

[0016] 在一些实施例中,基于细胞的测定丰度值为组织的细胞的测定丰度值。在一些实施例中,组织为骨骼、软骨、关节、气管、脊髓、角膜、眼睛、皮肤或血管。

[0017] 在一些实施例中,基于细胞的测定丰度值为多个干细胞中的细胞的测定丰度值。在一些实施例中,多个干细胞为多个胚胎干细胞、多个成体干细胞或多个诱导性多能干细胞(iPSC)。

[0018] 在一些实施例中,基于细胞的测定丰度值为多个原代人细胞中的细胞的测定丰度值。在一些此类实施例中,多个原代人细胞为多个CD34+细胞、多个CD34+造血干细胞、多个祖细胞(HSPC)、多个T细胞、多个间充质干细胞(MSC)、多个气道基底干细胞或多个诱导性多能干细胞。

[0019] 在一些实施例中,基于细胞的测定丰度值为脐带血中、外周血中或骨髓中的细胞的测定丰度值。

[0020] 在一些实施例中,基于细胞的测定丰度值为实体组织中的细胞的测定丰度值。在一些此类实施例中,实体组织为胎盘、肝脏、心脏、脑、肾脏或胃肠道。

[0021] 在一些实施例中,基于细胞的测定丰度值为多个分化细胞的测定丰度值。在一些此类实施例中,多个分化细胞为多个巨核细胞、多个成骨细胞、多个软骨细胞、多个脂肪细胞、多个肝细胞、多个肝间皮细胞、多个胆管上皮细胞、多个肝星状细胞、多个肝窦内皮细胞、多个库普弗细胞、多个隐窝细胞、多个血管内皮细胞、多个胰管上皮细胞、多个胰管细胞、多个腺腔中心细胞、多个腺泡细胞、多个朗格尔汉斯小岛)、多个心肌细胞、多个纤维母细胞、多个角质形成细胞、多个平滑肌细胞、多个I型肺泡上皮细胞、多个II型肺泡上皮细胞、多个克拉拉细胞、多个纤毛上皮细胞、多个基底细胞、多个杯状细胞、多个神经内分泌细胞、多个库尔契茨基细胞、多个肾小管上皮细胞、多个尿路上皮细胞、多个柱状上皮细胞、多个肾小球上皮细胞、多个肾小球内皮细胞、多个足细胞、多个血管系膜细胞、多个神经细胞、多个星形胶质细胞、多个小胶质细胞或多个少突胶质细胞。

[0022] 在一些实施例中,对应的多个基于细胞的测定丰度值为多个细胞的单细胞核糖核

酸(RNA)测序(scRNA-seq)数据。在一些此类实施例中,通过将细胞的不同等分试样暴露于一种或多种已知影响生理状况的参考化合物而得到与生理状况相关联的多种不同状态,此外还得到对照状态,在该对照状态下,细胞的等分试样并未免于暴露于已知影响生理状况的化合物。

[0023] 在一些实施例中,对应的多个基于细胞的测定丰度值来自大量RNA序列。

[0024] 在一些实施例中,对应的多个基于细胞的测定丰度值来自单细胞RNA测序。

[0025] 在一些实施例中,细胞组分模块的集合由第一细胞组分模块组成。

[0026] 在一些实施例中,细胞组分模块的集合包括多个细胞组分模块,并且模型为包括多个成分模型的集成模型。响应于将化学结构的指纹输入到多个成分模型中的每个成分模型中,多个成分模型中的每个成分模型提供针对细胞组分模块的集合中的不同细胞组分模块的激活评分。

[0027] 在一些实施例中,该方法进一步包括根据测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算指纹。

[0028] 在一些此类实施例中,多个成分模型中的每个成分模型为对应的神经网络(例如,全连接神经网络、消息传递神经网络或其组合)。在一些实施例中,对应的神经网络为对应的全连接神经网络和对应的消息传递神经网络的组合,响应于将化学结构的指纹输入到对应的全连接神经网络和对应的消息传递神经网络中,将对应的全连接神经网络的第一输出和对应的消息传递神经网络的第二输出组合,以确定一个或多个计算出的激活评分中的针对细胞组分模块的集合中的对应的细胞组分模块的激活评分。

[0029] 在一些此类实施例中,多个成分模型中的成分模型为逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0030] 在一些实施例中,细胞组分模块的集合为多个细胞组分模块,多个细胞组分模块的包括第一细胞组分模块的第一子集与目的生理状况相关联,多个细胞组分模块的第二子集与目的生理状况不关联,并且当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准且针对多个细胞组分模块的第二子集中的细胞组分模块的相应的计算出的激活评分满足第一阈值标准之外的第二阈值标准时,测试化学化合物与目的生理状况关连。

[0031] 在一些实施例中,该方法进一步包括通过包括以下的过程来识别第一细胞组分模块:以电子形式获得一个或多个第一数据集,该一个或多个第一数据集包括或共同包括:对于第一多个细胞中的每个相应的细胞,其中第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态:对于多种细胞组分(例如,至少10、20、30、100或1000种或更多种细胞组分中的每种相应的细胞组分:相应的细胞组分在相应的细胞中的对应的丰度,由此获取或形成多个向量,多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分并且(ii)包括对应的多个元素,对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示第一多个细胞中的相应的细胞中的相应的细胞组分的对应的丰度。该方法进一步包括使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块,多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集,其中多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确

定维度的潜在表示中,并且其中多个细胞组分模块包括多于十个细胞组分模块。该方法进一步包括以电子形式获得一个或多个第二数据集,该一个或多个第二数据集包括或共同包括:对于第二多个细胞(其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量)中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度,由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示来确定维度的细胞组分计数数据结构。该方法进一步包括:通过使用多种细胞组分或其表示作为公共维度对细胞组分计数数据结构和潜在表示进行组合来形成激活数据结构,其中该激活数据结构包括:对于多个细胞组分模块中的每个细胞组分模块:对于第二多个细胞中的每个细胞,相应的激活权重;以及针对多个协变量中的每个相应的协变量,使用以下两者之间的差异来训练候选细胞组分模型:(i)在将协变量的指纹输入到候选细胞组分模型中时针对由候选细胞组分模型表示的每个细胞组分模块的计算出的激活,以及(ii)针对由候选细胞组分模型表示的每个细胞组分模块的实际激活,其中该训练响应于差异来调整与候选细胞组分模型相关联的多个协变量参数。在一些此类实施例中,多个协变量参数包括:对于多个细胞组分模块中的每个相应的细胞组分模块:对于每个相应的协变量:对应的参数,该对应的参数指示相应的协变量是否跨第二多个细胞与相应的细胞组分模块相关;并且该方法进一步包括:在训练候选细胞组分模型时使用多个协变量参数来识别多个候选细胞组分模块中的第一细胞组分模块。在一些此类实施例中,该方法进一步包括:多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件下对化合物的暴露(例如,暴露的持续时间、化合物的浓度或暴露的持续时间与化合物的浓度的组合)。

[0032] 在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。

[0033] 在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合,并且相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过比色测量、荧光测量、发光测量或共振能量转移(FRET)测量来确定。

[0034] 在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合,并且相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq或其任何组合。

[0035] 在一些实施例中,使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。在一些此类实施例中,相关模型包括图聚类(例如,基于皮尔逊相关的距离度量上的莱顿聚类、鲁汶聚类等)。

[0036] 在一些实施例中,多个细胞组分模块由介于10个与2000个之间的细胞组分模块或介于100种与8,000种之间的细胞组分组成。在一些实施例中,多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

[0037] 在一些实施例中,目的生理状况为疾病。

[0038] 在一些实施例中,目的生理状况为疾病,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态所指示。

[0039] 在一些实施例中,多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态、暴露于化学化合物或其任何组合。

[0040] 在一些实施例中,该训练该候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0041] 在一些实施例中,测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。在一些此类实施例中,测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的至少三个标准的有机化合物。在一些实施例中,模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0042] 在一些实施例中,该方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从测试化学化合物的化学结构生成指纹。

[0043] 在一些实施例中,细胞组分模块的集合包括五个或更多个细胞组分模块、十个或更多个细胞组分模块,或者100个或更多个细胞组分模块。

[0044] 在一些实施例中,相应的细胞组分模块中的多种细胞组分的独立子集包括五种或更多种细胞组分。

[0045] 在一些实施例中,相应的细胞组分模块中的多种细胞组分的独立子集由与目的生理状况相关联的分子途径中的介于两种与20种之间的细胞组分组成。

[0046] 在一些实施例中,第一阈值标准为以下要求:第一细胞组分模块具有阈值激活评分。

[0047] 本公开的另一方面提供了一种将测试化学化合物与目的生理状况相关联的方法。

[0048] 该方法包括(A)获得测试化学化合物的化学结构的指纹。

[0049] 该方法进一步包括(B)获取扰动特征的集合,其中该扰动特征的集合中的每个相应的扰动特征包括多种细胞组分的相应的独立子集,扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。

[0050] 该方法进一步包括(C)将指纹输入到模型中,其中该模型包括50、100、500、1000或10,000个或更多个参数,该模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分,一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征。

[0051] 该方法进一步包括(D)当针对扰动特征的集合中的第一扰动特征的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况相关联。

[0052] 在一些实施例中,该方法进一步包括根据测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算指纹。

[0053] 在一些实施例中,模型包括神经网络。在一些此类实施例中,神经网络为全连接神经网络、消息传递神经网络或其组合。

[0054] 在一些实施例中,模型为包括多个成分模型的集成模型,并且多个成分模型中的每个成分模型响应于将化学结构的指纹输入到多个成分模型的集合中的每个成分模型中而提供针对扰动特征的集合中的不同扰动特征的激活评分。

[0055] 在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0056] 在一些实施例中,多个成分模型中的每个成分模型为对应的神经网络(例如,对应的神经网络为全连接神经网络、消息传递神经网络或其组合)。

[0057] 在一些实施例中,多个成分模型中的成分模型为逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0058] 在一些实施例中,对应的神经网络为全连接神经网络和消息传递神经网络的组合,并且响应于将化学结构的指纹输入到全连接神经网络和消息传递神经网络中,将第一神经网络的第一输出和第二神经网络的第二输出组合,以确定一个或多个计算出的激活评分中的针对扰动特征的集合中的第一扰动特征的激活评分。

[0059] 在一些实施例中,扰动特征的集合为多个扰动特征,多个扰动特征的包括第一扰动特征的第一子集与目的生理状况相关联,多个扰动特征的第二子集与目的生理状况不关联,并且当针对第一扰动特征的相应的计算出的激活评分满足第一阈值标准且针对多个扰动特征的第二子集中的扰动特征的相应的计算出的激活评分满足第一阈值标准之外的第二阈值标准时,测试化学化合物与目的生理状况关连。

[0060] 在一些实施例中,目的生理状况为疾病。

[0061] 在一些实施例中,测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。

[0062] 在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。在一些此类实施例中,测试化学化合物为满足里宾斯基五规则标准中的至少三个标准的有机化合物。

[0063] 在一些实施例中,模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0064] 在一些实施例中,该方法进一步包括使用使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从测试化学化合物的化学结构生成指纹。

[0065] 在一些实施例中,扰动特征的集合由第一扰动特征组成。

[0066] 在一些实施例中,扰动特征的集合包括五个或更多个扰动特征、十个或更多个扰动特征,或者100个或更多个扰动特征。

[0067] 在一些实施例中,第一阈值标准为以下要求:第一扰动特征具有阈值激活评分。

[0068] 本公开的另一方面提供了一种将化学化合物与目的生理状况相关联的方法。

[0069] 该方法包括在包括存储器和一个或多个处理器的计算机系统处:(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,由此获得多个指纹。

[0070] 该方法进一步包括(B)以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对多种化合物中的每种化合物的相应的数值激活评分,其中细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。

[0071] 该方法进一步包括(C)训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分,以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分,其中该训练(C)响应于该差异而调整与未经训练的模型相关联的多个参数(并且其中该多个参数包括50、100、200、500、1000或10,000个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0072] 在一些实施例中,细胞组分模块的集合由单个细胞组分模块组成。

[0073] 在一些实施例中,细胞组分模块的集合包括多个细胞组分模块。

[0074] 在一些实施例中,细胞组分模块的集合由介于二百个与五百个之间的细胞组分模块组成。

[0075] 在一些实施例中,多种化合物由介于10种与1×10^{6}种之间的化合物组成。

[0076] 在一些实施例中,多种化合物由介于100种与100,000种之间的化合物组成。

[0077] 在一些实施例中,多种化合物由介于1000种与100,000种之间的化合物组成。

[0078] 在一些实施例中,该训练(C)根据回归算法响应于与每种相应的化合物相关联的针对细胞组分模块的集合中的每个相应的细胞组分模块的每个差异而调整与未经训练的模型相关联的多个参数。在一些此类实施例中,回归算法优化与每种相应的化合物相关联的针对细胞组分模块的集合中的每个相应的细胞组分模块的每个差异的最小二乘误差。

[0079] 在一些实施例中,经训练的模型包括神经网络(例如,全连接神经网络、消息传递神经网络或其组合)。

[0080] 在一些实施例中,经训练的模型为多个成分模型的集成模型,并且多个成分模型中的每个相应的成分模型输出针对多个细胞组分模块中的不同细胞组分模块的计算出的激活评分。在一些此类实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0081] 在一些实施例中,多个成分模型中的每个成分模型为对应的神经网络。在一些此类实施例中,对应的神经网络为全连接神经网络、消息传递神经网络或其组合。

[0082] 在一些实施例中,细胞组分模块的集合为多个细胞组分模块,多个细胞组分模块的第一子集与目的生理状况相关联,并且多个细胞组分模块的第二子集与目的生理状况不关联。

[0083] 在一些实施例中,该方法进一步包括通过包括以下的过程来识别多个细胞组分模

块中的细胞组分模块:以电子形式获得一个或多个第一数据集,该一个或多个第一数据集包括或共同包括:对于第一多个细胞中的每个相应的细胞,其中第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态:对于多种细胞组分中的每种相应的细胞组分,其中多种细胞组分包括5、10、15、20、25、50或100种或更多种细胞组分:相应的细胞组分在相应的细胞中的对应的丰度,由此获取或形成多个向量。多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素。对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。使用多个向量以识别多个候选细胞组分模块中的每个候选细胞组分模块,多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集。多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中,并且多个细胞组分模块包括多于3、5、10、15、20或100个细胞组分模块。一个或多个第二数据集是以电子形式获得的,该一个或多个第二数据集包括或共同包括:对于第二多个细胞(其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量)中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度,由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示来确定维度的细胞组分计数数据结构。激活数据结构是通过以下来形成的:使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示,其中激活数据结构包括:对于多个细胞组分模块中的每个细胞组分模块:对于第二多个细胞中的每个细胞,相应的激活权重。候选细胞组分模型是使用以下两者之间的差异来训练的:(i)在将激活数据结构输入到候选模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测,以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在。该训练响应于差异而调整与候选细胞组分模型相关联的多个协变量参数。

[0084] 在一些实施例中,多个协变量参数包括:对于多个细胞组分模块中的每个相应的细胞组分模块:对于每个相应的协变量:对应的参数,该对应的参数指示相应的协变量是否跨第二多个细胞与相应的细胞组分模块相关;并且在训练候选细胞组分模型时使用多个协变量参数来识别多个候选细胞组分模块中的细胞组分模块。

[0085] 在一些实施例中,多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件下暴露于化合物。

[0086] 在一些实施例中,暴露条件为暴露的持续时间、化合物的浓度或暴露的持续时间与化合物的浓度的组合。

[0087] 在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。

[0088] 在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。

[0089] 在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq或其任何组合。

[0090] 在一些实施例中,使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。在一些此类实施例中,相关模型包括图聚类(例如,基于皮尔逊相关的距离度量上的莱顿聚类,或者为鲁汶聚类)。

[0091] 在一些实施例中,多种细胞组分由介于100种与8,000种之间的细胞组分组成。

[0092] 在一些实施例中,多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

[0093] 在一些实施例中,目的生理状况为疾病。

[0094] 在一些实施例中,生理状况为疾病,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态所指示。

[0095] 在一些实施例中,多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。

[0096] 在一些实施例中,该训练该候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0097] 在一些实施例中,多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。

[0098] 在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。在一些此类实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。

[0099] 在一些实施例中,经训练的模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0100] 在一些实施例中,该方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从对应的化学结构生成每个相应的指纹。

[0101] 在一些实施例中,细胞组分模块的集合包括五个或更多个细胞组分模块、十个或更多个细胞组分模块,或者100个或更多个细胞组分模块。

[0102] 本公开的另一方面提供了一种将化学化合物与目的生理状况相关联的方法。该方法例如可以在包括存储器和一个或多个处理器的计算机系统处进行。

[0103] 该方法包括(A)以电子形式获得多种化合物中的每种相应的化合物的对应的化学结构的相应的指纹,由此获得多个指纹。

[0104] 该方法进一步包括(B)以电子形式获得扰动特征的集合中的每个相应的扰动特征针对多种化合物中的每种对应的化合物的相应的数值激活评分,其中扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化。相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。

[0105] 该方法进一步包括(C)训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的扰动特征的相应的计算出的激活评分,以及(ii)扰动特征的集合中的相应的扰动特征针对对应的化合物的相应的数值激活评分。训练(C)响应于差异而调整与未经训练的模型相关联的多个参数,由此获得将化学化合物与目的生理状况相关联的经训练的模型。在一些实施例中,多个参数包括50、100、200、500、1000、10,000或1×10^6个或更多个参数。

[0106] 在一些实施例中,扰动特征的集合由单个扰动特征组成。

[0107] 在一些实施例中,扰动特征的集合由介于二百个与五百个之间的扰动特征组成。

[0108] 在一些实施例中,多种化合物由介于10种与1×10^6种之间的化合物组成。在一些实施例中,多种化合物由介于100种与100,000种之间的化合物组成。在一些实施例中,多种化合物由介于1000种与100,000种之间的化合物组成。

[0109] 在一些实施例中,该训练(C)根据回归算法响应于与每种对应的化合物相关联的针对扰动特征的集合中的每个相应的扰动特征的每个差异而调整与未经训练的模型相关联的多个参数。在一些此类实施例中,回归算法优化与每种对应的化合物相关联的针对扰动特征的集合中的每个相应的扰动特征的每个差异的最小二乘误差。

[0110] 在一些实施例中,经训练的模型包括神经网络(例如,全连接神经网络、消息传递神经网络或其组合)。

[0111] 在一些实施例中,经训练的模型为多个成分模型的集成模型,并且多个成分模型中的每个相应的成分模型响应于将相应的化学结构的指纹输入到多个成分模型的集合中的每个成分模型中而输出针对多个扰动特征的集合中的不同的扰动特征的集合的计算出的激活评分。在一些此类实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0112] 在一些实施例中,多个成分模型中的每个成分模型为对应的神经网络(例如,全连接神经网络、消息传递神经网络或其组合)。

[0113] 在一些实施例中,扰动特征的集合包括多个扰动特征,多个扰动特征的第一子集与目的生理状况相关联,并且多个扰动特征的第二子集与目的生理状况不关联。

[0114] 在一些实施例中,目的生理状况为疾病。

[0115] 在一些实施例中,多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。

[0116] 在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。

[0117] 在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。

[0118] 在一些实施例中,经训练的模型包括逻辑回归模型、神经网络模型、支持向量机、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0119] 在一些实施例中,该方法进一步包括使用Daylight、BCI、ECFP4、EcFC、MDL、APFP、TTFP、UNITY 2D指纹、RNNS2S或GraphConv从对应的化学结构生成每个相应的指纹。

[0120] 在一些实施例中,扰动特征的集合包括五个或更多个扰动特征、十个或更多个扰动特征,或者100个或更多个扰动特征。

[0121] 在一些实施例中,该方法进一步包括通过包括以下的程序来获得扰动特征的集合中的相应的扰动特征的相应的数值激活评分:以电子形式获取单细胞转变特征,该单细胞转变特征表示未改变的细胞状态与改变的细胞状态之间的差异细胞组分丰度的测度,其中改变的细胞状态通过从未改变的细胞状态到改变的细胞状态的细胞转变而出现,(i)未改变的细胞状态、(ii)改变的细胞状态以及(iii)从未改变的细胞状态到改变的细胞状态的转变中的至少一者与目的生理状况相关联,并且单细胞转变特征包括参考多种细胞组分的标识以及对应的第一显著性评分(针对多种参考细胞组分中的每种相应的细胞组分),该对应的第一显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及未改变的细胞状态与改变的细胞状态之间的细胞状态变化。此外,比较单细胞转变特征和相应的扰动特征,从而确定相应的扰动特征的相应的数值激活评分。

[0122] 在一些实施例中,比较单细胞转变特征和扰动特征以确定相应的扰动特征的相应的数值激活评分包括:针对有单细胞转变特征的参考多种细胞组分中的每种相应的细胞组分,将相应的细胞组分的第一显著性评分与对应的细胞组分在相应的扰动特征中的对应的显著性评分进行比较。

[0123] 在一些实施例中,相应的扰动特征的激活评分为相应的扰动特征(相对于扰动特征的集合中的其他扰动特征)与单细胞转变特征的相关性的相对排名。

[0124] 在一些实施例中,相对排名通过Wilcoxon秩和检验、t检验、逻辑回归或广义线性模型来确定。

[0125] 在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态或第二细胞状态相同。

[0126] 在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态和第二细胞状态两者均不同。

[0127] 在一些实施例中,该方法进一步包括:修剪有单细胞转变特征的参考多种细胞组分以及有相应的扰动特征的相应的多种细胞组分以限制与转录因子的比较。

[0128] 在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的细胞状态由尚未暴露于多种化合物中的化合物的对照细胞表示。

[0129] 在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的细胞状态由跨已经暴露于多种化学化合物中的化学化合物(除了与相应的扰动特征相关联的化合物之外)的不相关的受扰动的细胞的平均数表示。

[0130] 在所公开的实施例中的一些实施例中,模型为回归器。

[0131] 本公开的另一方面提供了一种计算机系统,其具有一个或多个处理器和存储器,该存储器存储用于供一个或多个处理器执行的一个或多个程序,该一个或多个程序包括用于进行本文所公开的方法和/或实施例中的任一者的指令。

[0132] 本公开的另一方面提供了一种非暂时性计算机可读存储介质,其存储配置成用于供计算机执行的一个或多个程序,该一个或多个程序包括用于执行本文所公开的方法和/

或实施例中的任一者的指令。

[0133] 对于本领域技术人员而言,根据以下详细说明,本公开的另外的方面和优点将变得显而易见,其中仅示出和描述了本公开的说明性实施例。如将认识到的,本公开能够具有其他且不同的实施例,并且其若干细节能够在各种明显的方面进行修改,而所有这些都不脱离本公开。因此,附图和说明书将在本质上被视为是说明性的而非限制性的。

附图说明

[0134] 在附图中,通过实例而非限制的方式绘示了本文所公开的实施例。贯穿附图,相似的附图标记指代对应的部分。

[0135] 图1展示了根据本公开的实施例的示例性系统和计算装置的框图。

[0136] 图2A和2B共同提供了根据本公开的各个实施例的用于将多种细胞组分与目的生理状况相关联的示例方法的过程和特征的流程图。

[0137] 图3A、3B、3C、3D和3E提供了根据本公开的各个实施例的用于将测试化学化合物与目的生理状况相关联的示例方法的过程和特征的流程图,其中虚线框表示任选元素。

[0138] 图4展示了根据本公开的一些实施例的细胞组分的多个向量的示例和细胞组分模块的潜在表示的示例。

[0139] 图5展示了根据本公开的一些实施例的细胞组分计数数据结构的示例和示例激活数据结构。

[0140] 图6展示了根据本公开的一些实施例的训练模型来调整多个化合物权重的方法的示例。

[0141] 图7提供了根据本公开的一些实施例的用于将测试化学化合物与目的生理状况相关联的示例方法的过程和特征的流程图,其中虚线框表示任选元素。

[0142] 图8提供了根据本公开的实施例的用于将化学化合物与目的生理状况相关联的示例方法的过程和特征的流程图,其中虚线框表示任选元素。

[0143] 图9提供了根据本公开的实施例的用于将化学化合物与目的生理状况相关联的示例方法的过程和特征的流程图,其中虚线框表示任选元素。

[0144] 图10A、10B、10C、10D和10E展示了根据本公开的实施例的用于预测用于激活脂肪酸相关细胞程序的化学结构的示例方法的性能和4折验证。图10A展示了用于预测化学结构的模型架构的示意图。图10B展示了1,200种随机选择的化合物的测试集的性能。图10C展示了1,200种化合物的测试集(相比训练集具有不同的骨架)的性能。图10D展示了在体外前脂肪细胞测定中基于转录激活的对米色化(beigeing)相关模块的验证。图10E展示了针对靶模块的对从5百万种化合物的数据库中提取的预测化合物的优化。

[0145] 图11展示了根据本公开的实施例的用于预测用于激活与胎儿红细胞生成和T细胞耗竭相关的细胞行为的化学结构的示例方法的验证。

[0146] 图12展示了根据本公开的实施例的用于使用单细胞RNA测序(scRNA-seq)来评估已知含哌啶化合物(“KPCC”)和六种新合成的苗头化合物(hit)“合成苗头化合物”)对人前脂肪细胞基因模块激活的影响的示例方法的示意图。

[0147] 图13展示了根据本公开的实施例的KPCC和六种合成苗头化合物对期望转录变化的激活的影响。

[0148] 图14A、14B、14C和14D提供了识别细胞组分模块的流程图,其中任选元素由虚线框指示。

具体实施方式

[0149] 简介。

[0150] 鉴于上述背景,本公开描述了一种以对疾病至关重要的细胞过程和程序为目标的药物发现方法。在一些方面,这种方法是通过以下来实现的:使用生理状况(例如,细胞程序、细胞过程和/或细胞状态)的经计算工程化的表示和化合物的化学结构来预测化学结构相关模态及其性质。然后可以将经编码的化学结构映射到细胞程序和/或细胞状态的表示上,从而将化合物与生理状况相关联。

[0151] 举例来说,在一些方面,本公开提供了用于获得分子图谱(例如,基因模块)及目的生物过程(例如,细胞程序和/或细胞状态)与化合物的化学结构之间的关联的系统和方法。这些关联可以用于预测新的化学结构(诸如具有相似功能或结构性质的那些化学结构),以进行药物发现。

[0152] 在一些实施例中,具有预测能力的计算建模架构用于通过跨一个或多个结构域和/或数据类型生成生理相关化学结构的潜在表示来发现这些关联。关联可以衍生自例如提供细胞行为图谱的扰动数据,诸如响应于细胞暴露于一种或多种化合物的差异基因表达或细胞状态转变。在一些实现方式中,该方法使用潜在表示和机器学习来组合并确定多种结构域(例如,分子的、细胞的、临床的、体内的、体外的、基于知识的结构域等)和/或多种数据类型(转录、遗传、表观遗传、协变量等)之间的相关性以预测生理相关化学结构。

[0153] 在示例实施例中,本公开提供了一种使用针对化合物的潜在表示的建模方法。对于多种化合物中的每种相应的化合物,该方法包括生成潜在表示,该潜在表示储存表示相应的化合物诱导多种生理状况中的每种生理状况的可能性的向量。生理状况可以包括与特定表型、细胞过程和/或疾病相关联的细胞状态转变和/或细胞组分模块(例如,基因模块)。因此,该方法生成充当模型的多任务训练标记的矩阵表示,该矩阵表示由化合物和生理状况(例如,细胞状态和/或基因模块)确定维度,表示为例如n种_化合物x n种_细胞_状态或n种_化合物x n个_基因_模块。

[0154] 针对用于将化合物与生理状况相关联的机器学习模型的输入包括每种化合物的规范异构SMILES表示和/或基于图形的表示,其对化合物的化学结构进行编码,并进一步用于训练该模型。提供训练标记作为将每种化合物与每种生理状况相关联的数值激活评分。举例来说,针对每种化合物的向量可以包括多个相关联的权重,其中每个权重指示该化合物诱导相应的生理状况(诸如相应的细胞状态、细胞状态转变、扰动特征和/或相应的基因模块的激活)的可能性。

[0155] 在接收矩阵表示作为输入后,训练该模型以通过解决回归问题来从化学结构中学习细胞状态(例如,扰动特征)和/或基因模块激活。使用两个示例模型架构来解决回归问题。第一模型在SMILES字符串的标准指纹上使用全连接网络,其中网络架构为具有ReLU激活的3层网络。第二模型包括缺乏DGL库的MPNN网络。这些模型中的每一个均是通过优化回归预测的最小二乘误差来相互独立地训练的。在测试时,对这些模型的预测求平均,从而形成包括第一模型和第二模型的集成模型。然后,集成模型可以用于确定化合物和生理状况

之间的关联,该关联可以进一步应用于获得对来自化学结构的可能的生理激活的预测和/或对可能诱导特定生理状况的化学结构的预测。

[0156] 有利地,本文所公开的系统和方法通过提供用于药物发现的系统的可扩展的方法来解决上述缺点。举例来说,与药物发现相关的常规机器学习方法利用采用3D蛋白质和化学结构表示的在计算机上运行的靶点筛选能力,结合深度学习方法和高性能计算来计算候选化合物对靶点库的作用方式。然而,这些方法属于以靶点为中心的筛选范式,其不会充分解决生物过程背后的动态且高度网络化的多细胞系统的复杂性。其他用于药物发现的常规方法使用机器学习方法以基于转录组数据或成像数据来模拟单细胞和细胞系如何响应于扰动。在此类方法中,使用高通量数据集来学习疾病的表型表示和细胞体外系统的化合物扰动。这些用于预测会诱导或抵消表型疾病响应的化合物。然而,传统的高通量数据建模方法仍然由于缺乏管理和识别大量候选靶点的潜力而处于劣势。对从高通量筛选中获得的每个潜在候选的验证是费力的过程,通常需要进行基于分子靶点的优化或合成数百甚至数千种化合物来进行体外筛选。

[0157] 与这些方法相比,本公开有利地提供了用于获得表示化学结构数据(例如,对化合物处理的细胞响应)的系统和方法,然后跨与生物过程相关联的细胞状态、扰动特征和/或细胞组分(例如,目的生理状况中涉及的基因模块或扰动特征)的表示对该数据进行映射。然而,这种与靶点无关的方法允许对候选靶点进行系统管理和优化,从而弥合靶点发现与跨系统预测性转译之间的巨大差距。

[0158] 举例来说,如下面的实例中所说明,使用本文所公开的系统和方法的实施例来识别参与脂肪酸代谢的候选药效团。如实例4中所进一步说明,基于候选药效团的预测转译产生6种新的化学实体,发现所有新的化学实体在人脂肪细胞上经测试时均激活参与脂肪酸相关细胞过程的基因模块。候选药效团的识别和6种新的化学实体的设计的进行无需针对蛋白质靶点的高通量筛选、识别或优化,也无需数百或数千种新化合物的合成。因此,与常规的基于分子靶点或基于表型的方法相比,本文提供的系统和方法从靶点发现到预测性转译和验证提高药物发现和开发过程的简易性和效率。

[0159] 有利地,本公开进一步提供了通过改善用于有针对性地确定化合物与生理状况之间的关联(例如,权重和/或相关)的模型的训练和使用来改善化合物与生理状况的关联的各种系统和方法。机器学习模型的复杂度包括时间复杂度(运行时间,或针对给定输入大小n的算法的速度的测度)、空间复杂度(空间要求,或执行针对给定输入大小n的算法所需的计算能力或存储器的量)或两者。复杂度(以及随后的计算负担)适用于对给定模型的训练和由其进行的预测。

[0160] 在一些情况下,计算复杂度受到附加算法或交叉验证方法的实现、合并以及/或者一个或多个参数(例如,权重和/或超参数)的影响。在一些情况下,计算复杂度表示为输入大小n的函数,其中输入数据为实例的数量(例如,训练样品的数量)、维度p(例如,特征的数量)、树的数量$n_{树}$(例如,对于基于树的方法)、支持向量的数量n_{sv}(例如,对于基于支持向量的方法)、邻居的数量k(例如,对于k最近邻模型)、类的数量c和/或第i层的神经元的数量n_i(例如,对于神经网络)。那么,就输入大小n来说,计算复杂度的近似(例如,以大0符号)表示运行时间和/或空间需求如何随着输入大小的增加而增加。相对于输入大小的增加,函数的复杂度可能以更慢或更快的速率增加。计算复杂度的各种近似包括但不限于常数(例如,0

(1))、对数(例如,0(log n))、线性(例如,0(n))、对数线性(例如,0(n log n))、二次(例如,0(n^2))、多项式(例如,0(n^c))、指数(例如,0(c^n))和/或阶乘(例如,0(n!))。在一些情况下,随着输入大小的增加,较简单的函数伴随较低级别的计算复杂度,如常数函数的情况,而较复杂的函数(诸如阶乘函数)可以响应于输入大小的轻微增加而表现出复杂度的大幅增加。

[0161] 机器学习模型的计算复杂度可以类似地通过函数来表示(例如,以大0符号),并且复杂度可能根据模型的类型、一个或多个输入或维度的大小、用途(例如,训练和/或预测)和/或是否正在评定时间或空间复杂度而变化。举例来说,决策树模型的复杂度近似为针对训练的0(n^2p)以及针对预测的0(p),而线性回归模型的复杂度近似为针对训练的0(p^2n+p^3)以及针对预测的0(p)。对于随机森林模型,训练复杂度近似为0($n^2pn_{树}$),并且预测复杂度近似为0($pn_{树}$)。对于梯度提升模型,复杂度近似为针对训练的0($npn_{树}$)以及针对预测的0($pn_{树}$)。对于内核支持向量机,复杂度近似为针对训练的0(n^2p+n^3)以及针对预测的0($n_{sv}p$)。对于朴素贝叶斯模型,复杂度表示为针对训练的0(np)以及针对预测的0(p),并且对于神经网络,复杂度近似为针对预测的0($pn_1+n_1n_2+\cdots$)。K最近邻模型的复杂度近似为针对时间的0(knp)以及针对空间的0(np)。对于逻辑回归模型,复杂度近似为针对时间的0(np)以及针对空间的0(p)。对于逻辑回归模型,复杂度近似为针对时间的0(np)以及针对空间的0(p)。

[0162] 如上面所描述,对于机器学习模型,计算复杂度决定可扩展性,并且因此决定模型(例如,回归器)对于增加输入、特征和/或类大小以及对于模型架构的变化的整体有效性和可用性。在大规模数据集的背景下,如在基因表达数据集包括针对至少10、至少100、至少1000或更多个细胞获得的至少10、至少100、至少1000或更多个基因的丰度的情况下,在如此大的数据集上进行的函数的计算复杂度可能对许多现有系统的能力造成压力。此外,随着输入特征的数量(例如,细胞组分(例如,基因)的数量和/或化合物的数量)和/或实例的数量(例如,细胞、细胞状态注释、扰动特征、模块和/或协变量的数量)的增加,加上技术进步、注释的渐增的可用性以及日益扩展的下游应用和可能性,任何给定分类模型的计算复杂度都可能很快压倒由相应的系统的规格提供的时间容量和空间容量。

[0163] 因此,使用具有最小输入大小(例如,至少10、至少100、至少1000或更多种化合物;对于相应的细胞组分模块,至少10、至少50、至少100或更多种细胞组分;至少5、至少10、至少100或更多个扰动特征;和/或至少5、至少10、至少100或更多个细胞组分模块)和/或对应最小数量的参数(例如,至少50、至少100或至少1000个参数以及/或者对应于输入到机器学习模型的所有特征的每个可能配对的参数)的机器学习模型来将化合物与生理状况相关联,就这一点来说,计算复杂度成比例地增加,使得其无法在头脑中进行,而该方法解决了计算问题。举例来说,在本公开的实施例中,获得由多个至少10个细胞组分模块和多种至少50种化合物来确定维度的激活评分矩阵包括获得至少500个参数(例如,权重)。在本公开的另一实施例中,针对多种至少50种化合物中的每种化合物、针对多个至少10个扰动特征中的每个扰动特征获得相应的激活权重包括获得至少500个激活权重。对附加的输入特征和/或实例(包括但不限于细胞状态转变、细胞组分、细胞、化合物、协变量、样品、时间点、重复和/或批次的数量)施加类似的最小值将类似地影响该方法的计算复杂度。

[0164] 有关机器学习模型的计算复杂度的更多细节提供在以下中:可于thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms在线获取的2018年4月16日发布的“机器学习算法的计算复杂度”;Hastie,

2001,The Elements of Statistical Learning,Springer,New York;以及Arora和Barak,2009,Computational Complexity:A Modern Approach,Cambridge University Press,New York;其中的每一者均特此通过引用以其整体并入本文。

[0165] 现将详细参考实施例,其实例示出于附图中。在以下详细描述中,阐述众多特定细节以便提供对本公开之透彻理解。然而,将显而易见的是,本领域普通技术人员可在没有这些特定细节的情况下实践本公开。在其他情况下,未详细描述熟知的方法、程序、组件、电路及网路,以免不必要地混淆实施例之范畴。

[0166] 对于本文中描述为单个例项之组件、操作或结构,可提供复数个例项。最后,各种组件、操作和数据储存之间的边界在某种程度上是任意的,并且在特定说明性组态的上下文中说明特定操作。设想到其他形式之功能性,且其可属于实施方案之范围内。一般而言,实例组态中呈现为单独组件的结构及功能性可实施为组合结构或组件。类似地,呈现为单个组件之结构及功能性可实施为单独组件。这些和其他变化、修改、添加及改良落入实施方案的范围内。

[0167] 还应当理解,虽然术语“第一”、“第二”等在本文中可用于描述各种元件,但这些元件不应受这些术语限制。这些术语仅用于将一个元件与另一元件区分开。例如,第一数据集可以被称为第二数据集,并且类似地,第二数据集可以被称为第一数据集,而不脱离本发明的范围。第一数据集和第二数据集都是数据集,但它们不是相同的数据集。

[0168] 本文中所使用之术语仅出于描述特定实施方案之目的,且并不意欲限制申请专利范围。如实施方案及随附申请专利范围之描述中所使用,除非上下文另外清楚地指示,否则单数形式“一(a/an)”及“该(the)”意欲亦包括复数形式。还应理解,如本文所使用的,术语“和/或”是指并涵盖相关联的所列项目中的一者或多者的任何及所有可能组合。应进一步理解,在用于本说明书中时,术语“包含(comprises及/或comprising)”指定所陈述特征、整体、步骤、操作、元素及/或组分的存在,但并不排除一个或多个其他特征、整体、步骤、操作、元素、组分及/或其群组的存在或添加。

[0169] 如本文中所使用,取决于上下文,术语“若”可解释为意谓“当”或“一旦”或“响应于确定”或“根据确定”或“响应于侦测到”所陈述的先决条件为真。类似地,取决于上下文,片语“若判定(所陈述的先决条件为真)”或“若(所陈述的先决条件为真)”或“当(所陈述的先决条件为真)时”可解释为意谓“一旦判定”或“回应于判定”或“根据判定”或“一旦侦测到”或“回应于侦测到”所陈述的先决条件为真。

[0170] 此外,当参考编号被给予“第i”指示时,参考编号是指通用成分、集合或实施例。举例来说,称为“第i细胞成分”的细胞成分是指多个细胞成分中的第i细胞成分。

[0171] 前述描述包括实现说明性实施方案的实例系统、方法、技术、指令序列和计算机程序产品。出于解释之目的,阐述众多特定细节以便提供对本发明主题之各种实施方案的理解。然而,熟习此项技术者将显而易见的,可在无此等特定细节之情况下实践本发明主题之实施方案。通常,没有详细示出公知的指令实例、协议、结构和技术。

[0172] 出于解释之目的,已参考特定实施方案描述了前述描述。然而,以下说明性论述并不意欲为详尽的或将实施方案限于所揭示之精确形式。鉴于以上教导,许多修改和变化是可能的。选择并描述该等实施方案係为了最佳地解释原理及其实际应用,藉此使熟习此项技术者能够最佳利用该等实施方案以及具有适合所涵盖之特定用途之各种修改的各种实

施方案。

[0173] 为了清晰起见,并未展示及描述本文中所描述之实施方案的所有常规特征。应了解,在开发任何此类实际实施方案时,作出众多实施方案特定决策以便实现设计者的特定目标,诸如遵守用例及业务相关约束,且此等特定目标将在一个实施方案至另一实施方案之间以及在一个设计者至另一设计者之间变化。此外,应了解,此类设计工作可能为复杂且耗时的,但对于受益于本发明的熟习此项技术者而言,为工程之常规任务。

[0174] 本说明书的某些部分根据对信息的操作的算法和符号表示来描述本发明的实施例。这些算法描述和表示通常被数据处理领域的技术人员用来有效地向本领域的其它技术人员传达其工作的实质。虽然在功能上、计算上或逻辑上描述了这些操作,但是这些操作被理解为由计算机程序或等效电路、微代码等来实现。

[0175] 说明书中使用的语言主要是出于可读性和指导性的目的而选择的,并且可能没有被选择来描绘或限制本发明的主题。因此,本发明的范围不限于该详细描述,而是由基于本申请提出的任何权利要求来限定。因此,本发明实施例的公开旨在说明而非限制本发明的范围。

[0176] 一般而言,权利要求和说明书中使用的术语旨在被解释为具有本领域普通技术人员所理解的普通含义。下文定义某些术语以提供额外的清楚性。在普通含义与所提供的定义之间冲突的情况下,将使用所提供的定义。

[0177] 本文未直接定义的任何术语应理解为具有如本发明领域内理解的通常与其相关联的含义。本文讨论了某些术语以向从业者提供描述本发明各方面的组合物、装置、方法等以及如何制备或使用它们的额外指导。应当理解,可以以多于一种的方式来说相同的事情。因此,可替代的语言和同义词可以用于本文讨论的任何一个或多个术语。无论术语是否在本文中被详细阐述或讨论,都没有意义。提供了一些同义词或可替换的方法、材料等。描述一个或几个同义词或等同物不排除使用其它同义词或等同物,除非明确说明。实例的使用,包括术语的实例,仅用于说明的目的,并不限制本发明的各方面的范围和含义。

[0178] 定义。

[0179] 如本文所用,术语“约”或“大约”意指在由普通技术人员中的一者所确定的特定值的可接受误差范围内,这部分地取决于如何测量或确定该值,例如,测量系统的限制。举例来说,在一些实施例中,根据本领域中的实践,“约”意指在1或大于1个标准偏差内。在一些实施例中,“约”意指给定值的±20%、±10%、±5%或±1%的范围。在一些实施例中,术语“约”或“大约”意指在值的数量级内、5倍内或2倍内。当在本申请和权利要求书中描述特定值时,除非另外指出,否则可以假设术语“约”意指所述特定值在可接受的误差范围内。本文的详细描述内的所有数值均由“约”所指示的值修饰,并且考虑将由所属领域的技术人员预期的实验误差和变化形式。术语“约”可具有一般熟习此项技术者通常所理解之含义。一些实施例中,术语“约”是指±10%。一些实施例中,术语“约”是指±5%。

[0180] 如本文所用,术语“丰度”、“丰度水平”或“表达水平”是指存在于一个或多个细胞中的细胞组分(例如,基因产物诸如RNA物种(例如mRNA或miRNA),或蛋白质分子)的量,或跨多个细胞存在的细胞组分的平均量。当提及mRNA或蛋白质表达时,该术语通常是指与特定基因组座位(例如,特定基因)相对应的任何RNA或蛋白质物种的量。然而,在一些实施例中,丰度可以指与产生多种mRNA或蛋白质同种型的特定基因相对应的mRNA或蛋白质的特定同

种型的量。可以使用基因名称、染色体位置或任何其他基因作图度量来标识基因组座位。

[0181] 如本文可互换使用的,“细胞状态”或“生物学状态”是指细胞或细胞群体的状态或表型。例如,细胞状态可以是健康的或患病的。细胞状态可以为多种疾病中的一种。细胞状态可以是对化合物处理和/或分化的细胞谱系的响应。细胞状态可以通过一种或多种细胞组分(包括但不限于一个或多个基因、一种或多种蛋白质和/或一种或多种生物学途径)的测度(例如,激活、表达和/或丰度测度)来表征。

[0182] 如本文所用,“细胞状态转变”或“细胞转变”是指细胞的状态从第一细胞状态到第二细胞状态的转变。在一些实施例中,第二细胞状态为改变的细胞状态(例如,健康的细胞状态到患病的细胞状态)。在一些实施例中,相应的第一细胞状态和第二细胞状态中的一个是未扰动状态,并且相应的第一细胞状态和第二细胞状态中的另一者是由细胞暴露于条件引起的扰动状态。受扰动的状态可能是由于细胞暴露于化合物而引起的。细胞状态转变可以通过细胞中细胞组分丰度的变化来标记,并因此通过由细胞产生的细胞组分(例如,mRNA、转录因子)的身份和量(例如,扰动特征)来标记。

[0183] 如本文所用,涉及针对一个细胞或多个细胞的细胞组分丰度测量的术语“数据集”在一些上下文中可以指从单细胞收集的高维数据集(例如,单细胞细胞组分丰度数据集)。在其他上下文中,术语“数据集”可以指从单细胞收集的多个高维数据集(例如,多个单细胞细胞组分丰度数据集),多个数据集中的每个数据集均是从多个细胞中的一个细胞收集的。

[0184] 如本文所用,术语“差异丰度”或“差异表达”是指存在于第一实体(例如,第一细胞、第一多个细胞和/或第一样品)(与第二实体(例如,第二细胞、第二多个细胞和/或第二样品)相比)中的细胞组分的量和/或频率的差异。在一些实施例中,第一实体为以第一细胞状态(例如,患病的表型)为特征的样品,并且第二实体为以第二细胞状态(例如,正常或健康的表型)为特征的样品。举例来说,细胞组分可以为多核苷酸(例如,mRNA转录本),与以第二细胞状态为特征的实体相比,以第一细胞状态为特征的实体中该多核苷酸以升高的水平或以降低的水平存在。在一些实施例中,细胞组分可以为多核苷酸,与以第二细胞状态为特征的实体相比,以第一细胞状态为特征的实体中该多核苷酸以较高频率或以较低频率被检测到。细胞组分在量、频率或两者方面的可能具有差异化丰度。在一些情况下,如果一个实体中的细胞组分的量与另一个实体中的细胞组分的量在统计学上显著不同,则在两个实体之间细胞组分具有差异化丰度。举例来说,如果细胞组分在一个实体中相比在另一个实体中的存在量大至少约120%、至少约130%、至少约150%、至少约180%、至少约200%、至少约300%、至少约500%、至少约700%、至少约900%或至少约1000%,或者如果细胞组分在一个实体中是可检测到的而在另一个中不是可检测到的,则该细胞组分在两个实体中具有差异化丰度。在一些情况下,如果在实体的第一子集(例如,表示经注释的细胞状态的第一子集的细胞)中与在实体的第二子集(例如,表示经注释的细胞状态的第二子集的细胞)中相比检测到细胞组分的频率在统计学上显著较高或较低,则该细胞组分在两个实体集合中被差异化表达。举例来说,如果在一个实体集合中与另一个实体集合相比观察到细胞组分的频率高或低至少约120%、至少约130%、至少约150%、至少约180%、至少约200%、至少约300%、至少约500%、至少约700%、至少约900%或至少约1000%,则该细胞组分在两个实体集合中被差异化表达。

[0185] 如本文所用,术语“健康的”是指以健康的状态为特征的样品(例如,从拥有良好健

康的受试者获得)。健康的受试者可以证明不存在任何恶性或非恶性疾病。“健康的”个体可能患有与所测定的病症无关的其他疾病或病症,该其他疾病或病症通常不能被视为“健康的”。

[0186] 如本文所用,涉及细胞的术语“扰动”(例如,细胞的扰动或细胞扰动)是指细胞对一种或多种条件(诸如通过一种或多种化合物进行处理)的任何暴露。这些化合物可以称为“扰动原”。在一些实施例中,扰动原可以包括,例如,小分子、生物制品、治疗剂、蛋白质、与小分子组合的蛋白质、ADC、如siRNA或干扰RNA等的核酸、过表达野生型和/或突变型shRNA的cDNA、过表达野生型和/或突变型向导RNA的cDNA(例如,Cas9系统或其他基因编辑系统),或任何前述的任何组合。扰动可以诱导或者特征可以为细胞的表型的变化和/或细胞中的一种或多种细胞组分的表达或丰度水平的变化(例如,扰动特征)。举例来说,扰动的特征可以是细胞的转录图谱的变化。

[0187] 如本文所用,术语“样品”、“生物学样品”或“患者样品”是指取自受试者的任何样品,其可以反映与受试者相关联的生物学状态。样品的示例包括但不限于受试者的血液、全血、血浆、血清、尿液、脑脊液、粪便、唾液、汗液、眼泪、胸膜液、心包液或腹膜液。样品可以包括从活的或死亡的受试者得到的任何组织或物质。样品可以为无细胞样品。样品可以包含一种或多种细胞组分。举例来说,样品可以包含核酸(例如,DNA或RNA)或其片段,或蛋白质。术语“核酸”可以指脱氧核糖核酸(DNA)、核糖核酸(RNA)或其任何杂合体或片段。样品中的核酸可以为无细胞核酸。样品可以为液体样品或固体样品(例如,细胞或组织样品)。样品可以为体液。样品可以为大便样品。可以处理样品以对组织或细胞结构进行物理破坏(例如,离心和/或细胞裂解),从而将细胞内成分释放到溶液中,该溶液可以进一步含有可以用于准备样品以供分析的酶、缓冲液、盐、去污剂等。

[0188] 如本文所用,如化合物的指纹方面的术语“指纹”是该化合物的数字摘要。此类数字摘要的非限制性示例包括日光(Daylight)指纹、BCI指纹、ECFC4指纹、ECFP4指纹、EcFC指纹、MDL指纹、原子对指纹(APFP指纹)、拓扑扭转指纹(TTFP)指纹、UNITY 2D指纹、RNNS2S指纹或GraphConv指纹。参见Franco,2014,“The Use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation,”J.Cheminform 6,第5页,以及Rensi和Altman,2017,“Flexible Analog Search with Kernel PCA Embedded Molecule Vectors,”Computational and Structural Biotechnology Journal,doi:10.1016/j.csbj.2017.03.003,其中的每一者均特此通过引用并入。另参见Raymond和Willett,2002,“Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases,”Journal of Computer-Aided Molecular Design 16,第59-71页,以及Franco等人,2014,“The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation”Journal of chemoinformatics 6(5),其中的每一者均特此通过引用并入。

[0189] 如本文所用,术语“分类”可以指与实体(例如,细胞、样品、细胞组分、细胞组分模块等)的特定性质(例如,细胞过程、协变量、细胞状态注释等)相关联的任何一个或多个编号或其他一个或多个字符。举例来说,符号“+”(或者字“正”)可以表明,实体被分类为对于特定性质是正的(例如,细胞组分模块与目的细胞过程正关联)。在另一示例中,术语“分类”

可以指对实体与特定性质之间的相关(例如,相应的协变量与相应的细胞组分模块之间的相关)的确定。在一些实施例中,分类为相关系数和/或权重。分类可以是二元的(例如,正的或负的)或者具有更多分类级别(例如,从1到10或0到1的数值范围)。术语“截止”和“阈值”可以指操作中使用的预定数字。举例来说,截止值可以指高于其的实体被排除的值。阈值可以为高于或低于其的特定分类适用的值。这些术语中的任一个均可以在任一这些上下文中使用。

[0190] 如本文中可互换使用的,术语“分类器”、“模型”、算法、“回归器”和/“或分类器”是指机器学习模型或算法。在一些实施例中,模型为无监督学习算法。无监督学习算法的一个示例为聚类分析。

[0191] 在一些实施例中,模型为有监督机器学习。有监督学习算法的非限制性示例包括但不限于逻辑回归、神经网络、支持向量机、朴素贝叶斯算法、最近邻算法、随机森林算法、决策树算法、提升树算法、多项逻辑回归算法、线性模型、线性回归、梯度提升、混合模型、隐马尔可夫模型、高斯NB算法、线性判别分析或其任何组合。在一些实施例中,模型为多项分类器算法。在一些实施例中,模型为2阶段随机梯度下降(SGD)模型。在一些实施例中,模型为深度神经网络(例如,深度-宽度样本级模型)。在一些实施例中,本公开的分类器或模型具有25或更多、100或更多、1000或更多、10,000或更多、100,000或更多或者1×10^6或更多个参数,并且因此对模型的计算无法在头脑中进行。

[0192] 此外,如本文中所用,术语“参数”是指算法、模型、回归器和/或分类器中可能影响(例如,修改、定制和/或调整)算法、模型、回归器和/或分类器中的一个或多个输入、输出和/或函数的任何系数或类似地,内部或外部元素(例如,权重和/或超参数)的任何值。举例而言,在一些实施例中,参数係指可用以控制、修改、定制和/或调整演算法、模型、回归器和/或分类器之行为、学习和/或效能的任何係数、权重和/或超参数。在一些情况下,参数用以增加或减少输入(例如,特征)对算法、模型、回归器和/或分类器的影响。作为非限制性示例,在一些实施例中,参数用于增加或减少(例如,神经网络的)节点的影响,其中节点包括一个或多个激活函数。将参数指派至特定输入、输出和/或函数不限于用于给定演算法、模型、回归器和/或分类器的任一范式,但可用于任何合适的演算法、模型、回归器和/或分类器架构中以实现所需效能。在一些实施例中,参数具有固定值。在一些实施例中,可手动地及/或自动地调整参数之值。在一些实施例中,通过算法、模型、回归器和/或分类器的验证和/或训练过程(例如,通过误差最小化和/或反向传播方法)来修改参数的值。在一些实施例中,本发明之演算法、模型、回归器和/或分类器包括多个参数。在一些实施例中,多个参数为n个参数,其中:$n\geq2$;$n\geq5$;$n\geq10$;$n\geq25$;$n\geq40$;$n\geq50$;$n\geq75$;$n\geq100$;$n\geq125$;$n\geq150$;$n\geq200$;$n\geq225$;$n\geq250$;$n\geq350$;$n\geq500$;$n\geq600$;$n\geq750$;$n\geq1,000$;$n\geq2,000$;$n\geq4,000$;$n\geq5,000$;$n\geq7,500$;$n\geq10,000$;$n\geq20,000$;$n\geq40,000$;$n\geq75,000$;$n\geq100,000$;$n\geq200,000$;$n\geq500,000$、$n\geq1\times10^6$、$n\geq5\times10^6$,或$n\geq1\times10^7$。因此,本公开的算法、模型、回归器和/或分类器无法在头脑中进行。在一些实施例中,n介于10,000与1×10^7之间、介于100,000与5×10^6之间,或介于500,000与1×10^6之间。在一些实施例中,本公开的算法、模型、回归器和/或分类器在k维空间中操作,其中k为5或更大的正整数(例如,5、6、7、8、9、10等)。因此,本公开的算法、模型、回归器和/或分类器无法在头脑中进行。

[0193] 神经网络。在一些实施例中,模型为神经网络(例如,卷积神经网络和/或残差神经

网络)。神经网络模型,也称为人工神经网络(ANN),包括卷积和/或残差神经网络模型(深度学习模型)。神经网络可以为机器学习模型,其可以训练成将输入数据集映射到输出数据集,其中该神经网络包括组织成多层节点的互连节点组。举例来说,神经网络架构可以至少包括输入层、一个或多个隐藏层以及输出层。神经网络可以包括任意总数的层和任意数量的隐藏层,其中隐藏层用作可训练特征提取器,该可训练特征提取器允许将输入数据集映射到一个输出值或输出值的集合。如本文所用,深度学习模型(DNN)可以为包括多个隐藏层(例如,两个或更多个隐藏层)的神经网络。神经网络的每个层均可以包括多个节点(或"神经元")。节点可以接收直接来自输入数据或者来自前一层中的节点的输出的输入,并进行特定操作,例如,求和操作。在一些实施例中,从输入到节点的连接与参数(例如,权重和/或加权因子)相关联。在一些实施例中,节点可以对输入的所有对x_i及其相关联参数的乘积求和。在一些实施例中,加权总和以偏差b偏移。在一些实施例中,可以使用阈值或激活函数f来对节点或神经元的输出进行门控,该阈值或激活函数可以为线性或非线性函数。激活函数可以为例如修正线性单元(ReLU)激活函数、泄露型ReLU激活函数或其他函数,诸如饱和双曲正切函数、恒等函数、二进制阶跃函数、逻辑函数、arcTan函数、softsign函数、参数修正线性单元函数、指数线性单元函数、softPlus函数、弯曲恒等函数、softExponential函数、Sinusoid函数、Sine函数、高斯函数或sigmoid函数,或其任何组合。

[0194] 可以在训练阶段使用一个或多个训练数据集来"教导"或"学习"加权因子、偏差值和阈值,或神经网络的其他计算参数。举例来说,可以使用来自训练数据集的输入数据和梯度下降或反向传播方法来训练参数,使得ANN所计算的一个或多个输出值与包括在训练数据集中的示例一致。参数可以从反向传播神经网络训练过程获得。

[0195] 多种神经网络中的任一种均可以适合用于分析受试者的图像。示例可以包括但不限于前馈神经网络、径向基函数网络、递归神经网络、残差神经网络、卷积神经网络、残差卷积神经网络等,或其任何组合。在一些实施例中,机器学习采用经预训练的和/或经转移学习的ANN或深度学习架构。根据本公开,卷积和/或残差神经网络可以用于分析受试者的图像。

[0196] 举例来说,深度神经网络模型包括输入层、多个单独参数化的(例如,加权的)卷积层和输出评分器。卷积层中的每一层以及输入层的参数(例如,权重)促成与深度神经网络模型相关联的多个参数(例如,权重)。在一些实施例中,至少100个参数、至少1000个参数、至少2000个参数或至少5000个参数与深度神经网络模型相关联。照此,深度神经网络模型需要使用计算机,因为它们无法在头脑中解决。换句话说,在给定对模型的输入的情况下,模型输出在此类实施例中需要使用计算机而非在头脑中确定。参见,例如,Krizhevsky等人,2012,"Imagenet classification with deep convolutional neural networks,"刊于Advances in Neural Information Processing Systems 2,Pereira、Burges、Bottou、Weinberger编著,第1097-1105页,Curran Associates,Inc.中;Zeiler,2012"ADADELTA:an adaptive learning rate method,"CoRR,卷abs/1212.5701;以及Rumelhart等人,1988,"Neurocomputing:Foundations of research,"ch.Learning Representations by Back-propagating Errors,第696-699页,Cambridge,MA,USA:MIT Press,其中的每一者均特此通过引用并入。

[0197] 适合用作模型的神经网络模型,包括卷积神经网络模型,公开于例如以下中:

Vincent等人,2010,“Stacked denoising autoencoders:Learning useful representations in a deep network with a local denoising criterion,”J Mach Learn Res 11,第3371-3408页;Larochelle等人,2009,“Exploring strategies for training deep neural networks,”J Mach Learn Res 10,第1-40页;以及Hassoun,1995,Fundamentals of Artificial Neural Networks,Massachusetts Institute of Technology,其中的每一者均特此通过引用并入。适合用作模型的附加示例神经网络公开于以下中:Duda等人,2001,Pattern Classification,第二版,John Wiley&Sons,Inc.,New York;以及Hastie等人,2001,The Elements of Statistical Learning,Springer-Verlag,New York,其中的每一者均特此通过引用以其整体并入。适合用作模型的附加示例神经网络也描述于以下中:Draghici,2003,Data Analysis Tools for DNA Microarrays,Chapman&Hall/CRC;以及Mount,2001,Bioinformatics:sequence and genome analysis,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,New York,其中的每一者均特此通过引用以其整体并入。

[0198] 支持向量机。在一些实施例中,模型为支持向量机(SVM)。适合用作模型的SVM模型描述于例如以下中:Cristianini和Shawe-Taylor,2000,“An Introduction to Support Vector Machines,”Cambridge University Press,Cambridge;Boser等人,1992,“A training algorithm for optimal margin models,”刊于Proceedings of the5th Annual ACM Workshop on Computational Learning Theory,ACM Press,Pittsburgh,Pa.,第142-152页中;Vapnik,1998,Statistical Learning Theory,Wiley,New York;Mount,2001,Bioinformatics:sequence and genome analysis,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.;Duda,Pattern Classification,第二版,2001,John Wiley&Sons,Inc.,第259、262-265页;和Hastie,2001,The Elements of Statistical Learning,Springer,New York;以及Furey等人,2000,Bioinformatics 16,第906-914页,其中的每一者均特此通过以其整体并入。当用于分类时,SVM将给定的二进制标记的数据集与最大程度地远离标记的数据的超平面分离。对于没有线性分离的情况,SVM可以结合自动实现对特征空间的非线性映射的内核技术运行。由SVM在特征空间中发现的超平面可以对应于输入空间中的非线性决策边界。在一些实施例中,与SVM相关联的多个参数(例如,权重)定义超平面。在一些实施例中,超平面由至少10、至少20、至少50或至少100个参数定义,并且SVM模型需要计算机来进行计算,因为它无法在头脑中解决。

[0199] 朴素贝叶斯模型。在一些实施例中,模型为朴素贝叶斯模型。适合用作模型的朴素贝叶斯模型公开于例如以下中:Ng等人,2002,“On discriminative vs.generative classifiers:A comparison of logistic regression and naive Bayes,”Advances in Neural Information Processing Systems,14,其特此通过引用并入。朴素贝叶斯分类器为基于应用贝叶斯定理和特征之间的强(朴素)独立假设的一系列“概率分类器”中的任何分类器。在一些实施例中,它们与核密度估计相结合。参见,例如,Hastie等人,2001,The elements of statistical learning:data mining,inference,and prediction,Tibshirani和Friedman编著,Springer,New York,其特此通过引用并入。

[0200] 最近邻模型。在一些实施例中,模型为最近邻模型。最近邻模型可以是基于存储器的并且不包括要拟合的模型。对于最近邻居,给定查询点x_0(测试受试者),k个训练点$x_{(r)}$,

r,...,识别距离x_0最近的k(此处为训练受试者),并且然后使用k个最近邻居对点x_0进行分类。这里,距这些邻居的距离为判别基因集的丰度值的函数。在一些实施例中,特征空间中的欧几里德距离用于确定距离为$d_{(i)}=||x_{(i)}-x_{(o)}||$。通常,当使用最近邻模型时,用于计算线性判别式的丰度数据被标准化成具有均值零和方差1。可以改进最近邻规则来解决不相等类先验、差异错误分类成本和特征选择的问题。这些改进中的许多改进涉及对邻居进行某种形式的加权投票。对于有关最近邻分析的更多信息,参见Duda,Pattern Classification,第二版,2001,John Wiley&Sons,Inc;以及2001,The Elements of Statistical Learning,Springer,New York,其中的每一者均特此通过引用并入。

[0201] K最近邻模型为非参数机器学习方法,其中输入由特征空间中的k个最近训练样品组成。输出为类隶属。对象通过其邻居的多数投票来分类,其中该对象被指派到在其k个最近邻居中最常见的类(k为正整数,通常很小)。如果k=1,则将该对象仅仅指派给该单个最近邻居的类。参见,Duda等人,2001,Pattern Classification,第二版,John Wiley&Sons,其特此通过引用并入。在一些实施例中,求解k最近邻模型所需的距离计算的数量使得使用计算机来求解针对给定输入的模型,因为该模型无法在头脑中进行。

[0202] 随机森林、决策树和提升树模型。在一些实施例中,模型为决策树。适合用为模型的决策树总体由以下描述:Duda,2001,Pattern Classification,John Wiley&Sons,Inc.,New York,第395-396页,其特此通过引用并入。基于树的方法将特征空间分区成矩形的集合并且然后在每个矩形中拟合模型(如常数)。在一些实施例中,决策树为随机森林回归。可以使用的一个特定模型为分类和回归树(CART)。其他特定决策树模型包括但不限于ID3、C4.5、MART和随机森林。CART、ID3和C4.5描述于以下中:Duda,2001,Pattern Classification,John Wiley&Sons,Inc.,New York,第396-408页和第411-412页,其特此通过引用并入。CART、MART和C4.5描述于以下中:Hastie等人,2001,The Elements of Statistical Learning,Springer-Verlag,New York,第9章,其特此通过引用以其整体并入。随机森林描述于以下中:Breiman,1999,"Random Forests--Random Features," Technical Report 567,Statistics Department,U.C.Berkeley,1999年9月,其特此通过引用以其整体并入。在一些实施例中,决策树模型包括至少10、至少20、至少50或至少100个参数(例如,权重和/或决策)并且需要计算机来进行计算,因为它无法在头脑中解决。

[0203] 回归。在一些实施例中,模型采用回归。回归算法可以为任何类型的回归。举例来说,在一些实施例中,回归为逻辑回归。在一些实施例中,回归为具有套索、L2或弹性网络正则化的逻辑回归。在一些实施例中,从考虑中(移除)修剪具有未能满足阈值的对应的回归系数的那些经提取的特征。在一些实施例中,对处置多范畴响应的逻辑回归模型的泛化被用作模型。逻辑回归公开于以下中:Agresti,An Introduction to Categorical Data Analysis,1996,第5章,第103-144页,John Wiley&Son,New York,其特此通过引用并入。在一些实施例中,模型采用公开于以下中的回归模型:Hastie等人,2001,The Elements of Statistical Learning,Springer-Verlag,New York。在一些实施例中,逻辑回归模型包括至少10、至少20、至少50、至少100或至少1000个参数(例如,权重)并且需要计算机来进行计算,因为它无法在头脑中解决。

[0204] 线性判别分析。线性判别分析(LDA)、正态判别分析(NDA)或判别函数分析可以是Fisher线性判别的泛化,该Fisher线性判别为用于统计学、模式标识和机器学习中以查找

表征或分离两类或更多类的对象或事件的特征的线性组合的方法。所得组合可以用作本公开的一些实施例中的模型(线性模型)。

[0205] 混合模型和隐马尔可夫模型。在一些实施例中,模型为混合模型,诸如描述于以下中的:McLachlan等人,Bioinformatics 18(3):413-422,2002。在一些实施例中,特别地,在包括时间分量的那些实施例中,模型为隐马尔可夫模型,诸如描述于以下中的:Schliep等人,2003,Bioinformatics 19(1):i255-i263。

[0206] 聚类。在一些实施例中,模型为无监督聚类模型。在一些实施例中,模型为有监督聚类模型。适合用作模型的聚类描述于例如以下中:Duda和Hart,Pattern Classification and Scene Analysis,1973,John Wiley&Sons,Inc.,New York(下文中为“Duda 1973”)的第211-256页,该文献特此通过引用以其整体并入。聚类问题可以描述为在数据集中寻找自然分组问题中的一个。为识别自然分组,可以解决两个问题。首先,可以确定测量两个样品之间的相似性(或相异性)的方式。该度量(例如,相似性度量)可以用于确保一个聚类中的样品比其他聚类中的样品彼此更相似。其次,可以确定用于使用相似性量度将数据分区成聚类的机制。开始聚类研究的一种方式可以是定义距离函数并计算训练集中所有样品对之间的距离的矩阵。如果距离为相似性的良好度量,则同一聚类中的参考实体之间的距离可以显著小于不同聚类中的参考实体之间的距离。然而,聚类可能不采用距离度量。例如,可以使用非度量相似性函数s(x,x')来比较两个向量x和x'。s(x,x')可以为对称函数,其值在x和x'以某种方式“相似”时较大。一旦选择了用于测量数据集中的点之间的“相似性”或“相异性”的方法,则聚类可以使用测量数据的任何分区的聚类质量的标准函数。可以使用使标准函数极值化的数据集的分区来对数据进行聚类。可以用于本公开的特定示例性聚类技术可以包括但不限于分级聚类(使用最近邻域算法、最远邻域算法、平均连接算法、质心算法或平方和算法的聚集聚类)、k均值聚类、模糊k均值聚类和帕特里克贾维斯(Jarvis-Patrick)聚类。在一些实施例中,聚类包括无监督聚类(例如,没有预先设想数量的聚类和/或没有预先确定聚类指派)。

[0207] 模型和提升的集成。在一些实施例中,使用(两个或更多个)模型的集成。在一些实施例中,提升技术诸如AdaBoost与许多其他类型的学习算法结合使用以改善模型的性能。在该方法中,本文所公开的任何模型的输出或其等效物组合成表示经提升的模型的最终输出的加权总和。在一些实施例中,使用本领域已知的任何集中趋势测度,包括但不限于均值、中值、众数、加权均值、加权中值、加权众数等来组合来自模型的多个输出。在一些实施例中,使用投票方法来组合多个输出。在一些实施例中,模型的集成中的相应的模型是加权的或未加权的。

[0208] 如本文所用,术语“未经训练的模型”(例如,“未经训练的回归器”和/或“未经训练的分类器”)是指尚未在训练数据集上训练的机器学习模型,诸如回归器或分类器。如本文所用,术语“训练模型”是指训练未经训练或经部分训练的模型的过程。举例来说,在一些实施例中,训练模型包括获得布置在潜在表示中的多个细胞组分模块和下面所讨论的细胞组分计数数据结构。布置在潜在表示中的多个细胞组分模块以及细胞组分计数数据结构被组合以形成激活数据结构,该激活数据结构结合针对激活数据结构中的多个细胞组分模块的每个细胞组分模块的的多个协变量中的每个协变量的实际不存在或存在而被应用为针对未经训练或经部分训练的模型的集体输入(下文中为“主要训练数据集”),以在协变量模块

相关性上训练未经训练或经部分训练的模型,由此获得经训练的模型。此外,应当认识到,术语“未经训练的模型”不排除在未经训练的模型的此类训练中使用转移学习技术的可能性。举例来说,以下提供了此类转移学习的非限制性示例:Fernandes等人,2017,“Transfer Learning with Partial Observability Applied to Cervical Cancer Screening,”Pattern Recognition and Image Analysis:Iberian Conference Proceedings第8版,第243-250页,其特此通过引用并入。在使用转移学习的情况下,除了主要训练数据集的数据之外,还向上面所描述的未经训练的模型提供附加数据。也就是说,在转移学习实施例的非限制性示例中,未经训练的模型接收(i)主要训练数据集和(ii)附加数据。通常,该附加数据呈从另一辅助训练数据集学习的系数(例如,回归系数)的形式。此外,虽然已经公开了对单个辅助训练数据集的描述,但是应当认识到,在本公开中在训练未经训练的模型时可以用于补充主要训练数据集的辅助训练数据集的数量不存在限制。举例来说,在一些实施例中,使用两个或更多个辅助训练数据集、三个或更多个辅助训练数据集、四个或更多个辅助训练数据集或者五个或更多个辅助训练数据集来通过转移学习来补充主要训练数据集,其中每个此类辅助训练数据集与主要训练数据集不同。在此类实施例中可以使用任何方式的转移学习。举例来说,考虑除了主要训练数据集之外还存在第一辅助训练数据集和第二辅助训练数据集的情况。可以使用转移学习技术(例如,二维矩阵乘法)将(通过对第一辅助训练数据集应用模型诸如回归)从第一辅助训练数据集学习到的系数应用于第二辅助训练数据集,这继而可以产生经训练的中间模型,然后将其系数应用于主要训练数据集,并将该应用与主要训练数据集本身结合应用于未经训练的模型。替代性地,可以(例如,通过单独的独立矩阵乘法)将(通过对第一辅助训练数据集应用模型诸如回归)从第一辅助训练数据集学习到的第一系数集和(通过对第二辅助训练数据集应用模型诸如回归)从第二辅助训练数据集学习到的第二系数集各自单独地应用于主要训练数据集的单独实例,并且然后可以将系数到主要训练数据集的单独实例的这两种应用与主要训练数据集本身(或某种简化形式的主要训练数据集,诸如从主要训练集学习到的主成分或回归系数)结合应用于未经训练的模型,以便训练该未经训练的模型。在任一示例中,使用从第一和第二辅助训练数据集得出的关于协变量模块相关性的知识(例如,附加细胞状态注释、附加协变量和/或其细胞组分丰度等),以结合经协变量标记的主要训练数据集来训练该未经训练的模型。

[0209] 如本文中可互换使用的,术语“神经元”、“节点”、“单元”、“隐藏神经元”、“隐藏单元”等是指神经网络的经由激活函数和一个或多个参数(例如,系数和/或权重)来接受输入并提供输出的单元。举例来说,隐藏神经元可以接受来自在先层的一个或多个输入,并提供用作后续层的输入的输出。在一些实施例中,神经网络包括仅一个输出神经元。在一些实施例中,神经网络包括多个输出神经元。一般来说,输出为预测值,诸如目的状况(诸如协变量、细胞状态注释或目的细胞过程)的概率或可能性、二元确定(例如,存在或不存在、正的或负的结果)和/或标记(例如,分类和/或相关系数)。对于单类分类模型,输出可以为输入特征(例如,一个或多个细胞组分模块)具有状况(例如,协变量、细胞状态注释和/或目的细胞过程)的可能性(例如,相关系数和/或权重)。对于多类分类模型,可以生成多个预测值,其中每个预测值针对每个目的状况来指示输入特征的可能性。

[0210] 如本文所用,术语“参数”是指模型、分类器、算法中可能影响(例如,修改、定制及/或调整)模型、分类器、算法中的一个或多个输入、输出和/或函数的任何系数或类似地,内

部或外部元素(例如,权重及/或超参数)的任何值。在一些实施例中,参数为调节模型中的一个或多个输入、输出或函数的系数(例如,权重)。举例来说,参数的值可以用于升高或降低针对模型的输入(例如,特征)的影响的权重。特征可以与诸如在逻辑回归、SVM或朴素贝叶斯模型中的参数相关联。参数的值可以替代性地或另外地用于升高或降低神经网络(例如,其中节点包括定义输入到输出的变换的一个或多个激活函数)中的节点、类或(例如,多个细胞中的一个细胞的)实例的影响的权重。将参数指派到特定输入、输出、函数或特征不限于用于给定模型的任一范式,但可以用于任何合适的模型架构中以实现最优性能。在一些情况下,对与模型的输入、输出、函数或特征相关联的参数(例如,系数)的引用可以类似地用作其数量、性能或优化的指示符,诸如在机器学习模型的计算复杂度的上下文中。在一些实施例中,参数具有固定值。在一些实施例中,(例如,使用超参数优化方法)可手动地和/或自动地调整参数的值。在一些实施例中,参数的值通过模型验证和/或训练过程(例如,通过误差最小化和/或反向传播方法,如本文别处所描述)来修改。

[0211] 如本文所用,术语"向量"是元素的枚举列表,诸如元素数组,其中每个元素具有所指派的含义。如此,如本公开中所使用的术语"向量"与术语"张量"可互换。作为示例,如果向量包括多个细胞中针对相应的细胞组分的丰度计数,则向量中存在针对多个细胞中的每一个的预定元素。为了便于呈现,在一些情况下,向量可以被描述为是一维的。然而,本公开不限于此。任何维度的向量均可以在本公开中使用,只要定义了对向量中的每个元素表示何物的描述(例如,元素1表示多个细胞中的细胞1的丰度计数等等)。

[0212] I.示例性系统实施例

[0213] 由于已经提供了本公开的一些方面的概述和本公开中使用的一些定义,所以结合图1描述了示范性系统的细节。

[0214] 图1提供了展示根据本公开的一些实施例的系统100的框图。系统100提供对与目的细胞过程相关联的多个细胞组分模块中的一个或多种细胞组分模块的确定。在图1中,系统100被绘示为计算装置。计算机系统100的其他拓扑是可能的。举例来说,在一些实施例中,系统100实际上可以构成在网络中链接在一起的数个计算机系统,或者为云计算环境中的虚拟机或容器。如此,图1所示的示范性拓扑仅用于以本领域技术人员将容易理解的方式描述本公开的实施例的特征。

[0215] 参考图1,在一些实施例中,计算机系统100(例如,计算装置)包括网络接口104。在一些实施例中,网络接口104通过一个或多个通信网络(例如,通过网络通信模块158)将系统内的系统100计算装置以及任选的外部系统和装置彼此互连。在一些实施例中,网络接口104任选地通过网络通信模块158经由因特网、一个或多个局域网(LAN)、一个或多个广域网(WAN)、其他类型的网络或此类网络的组合来提供通信。

[0216] 网络的实例包括万维网(WWW)、内联网和/或无线网络,如蜂窝电话网络、无线局域网(LAN)和/或城域网(MAN),以及通过无线通信的其它装置。无线通信任选地使用多个通信标准、协议及技术中的任一者,包括:全球移动通信系统(GSM)、增强型数据GSM环境(EDGE)、高速下行链路分组接入(HSDPA)、高速上行链路分组接入(HSUPA)、仅数据演进(EV-DO)、HSPA、HSPA+、双小区HSPA(DC-HSPDA)、长期演进(LTE)、近场通信(NFC)、宽带码分多路访问(W-CDMA)、码分多路访问(CDMA)、时分多路访问(TDMA)、蓝牙、无线保真(Wi-Fi)(例如,IEEE 802.11a、IEEE 802.11ac、IEEE 802.11ax、IEEE 802.11b、IEEE 802.11g和/或IEEE

802.11n)、互联网协议语音(VoIP)、Wi-MAX、用于电子邮件的协议(例如,互联网消息访问协议(IMAP)和/或邮局协议(POP))、即时消息处理(例如,可扩展消息处理和存在协议(XMPP)、即时消息处理和存在利用扩展的会话发起协议(SIMPLE)、即时消息处理和存在服务(IMPS)),以及/或者短消息服务(SMS),或任何其他合适的通信歇息,包括截至本文件的申请日期尚未开发的通信协议。

[0217] 在一些实施例中,系统100包括一个或多个处理单元(CPU)102(例如,处理器、处理核心等)、一个或多个网络接口104、包括(任选地)显示器108和输入系统105(例如,输入/输出接口、键盘、鼠标等)以供使用者使用的用户接口106、存储器(例如,非持久性存储器107、持久性存储器109),以及用于互连前述部件的一个或多个通信总线103。一个或多个通信总线103可选地包括互连并控制系统组件之间的通信的电路(有时称为芯片组)。非持久性存储器107通常包括高速随机存取存储器,诸如DRAM、SRAM、DDR RAM、ROM、EEPROM、闪存,而持久性存储器109通常包括CD—ROM、数字多功能盘(DVD)或其它光存储、磁带盒、磁带、磁盘存储或其它磁存储装置、磁盘存储装置、光盘存储装置、闪存装置或其它非易失性固态存储装置。持久性存储器109可选地包括远离CPU 102的一个或多个存储装置。持久性存储器109和非持久性存储器109内的非暂时性存储装置包括非暂时性计算机可读存储介质。在一些实施例中,非持久性存储器107或替代地非暂时性计算机可读存储介质有时结合持久性存储器109存储以下程序、模块和数据结构或其子集:

[0218] • 任选的操作系统156(例如,ANDROID、iOS、DARWIN、RTXC、LINUX、UNIX、OS X、WINDOWS,或者嵌入式操作系统诸如VxWorks),其包括用于处理各种基本系统服务和用于进行硬件相关任务的程序;

[0219] • 任选的网络通信模块(或指令)158,用于将系统100与其它装置和/或通信网络104连接;

[0220] • 化合物结构数据存储区120,其包括多种化合物中的每种化合物的相应的化学结构122(例如,122-1、…122-R)或其表示(例如,化学结构的指纹);

[0221] • 细胞组分模块数据存储区130,其包括细胞组分模块132(例如,132-1、…132-K)的集合,细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分134(例如,134-1-1、…134-1-Z)的子集;

[0222] • 扰动数据存储区140,其包括扰动特征142(例如,142-1、…142-P)的集合,扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分144(例如,144-1-1、…144-1-Q)(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化;

[0223] • 激活数据结构150,其包括:对于多种化合物中的每种相应的化合物的每个相应的化学结构152(例如,152-1、…152-R),

[0224] o任选地对于细胞组分模块的集合中的每个相应的细胞组分模块,相应的数值激活评分154(例如,154-1-1、…154-1-K),以及/或者

[0225] o任选地对于扰动特征的集合中的每个相应的扰动特征,相应的数值激活评分156(例如,156-1-1、…156-1-P);以及

[0226] • 模型,其包括多个参数(例如,100个或更多个参数),其中响应于针对相应的化

学结构的计算出的激活评分与数值激活评分之间的差异而调整该多个参数。

[0227] 在各种实施例中,以上标识的元件中的一个或多个存储在前述存储器装置中的一个或多个中,并且对应于用于进行上述功能的指令的集合。以上标识的模块、数据或程序(例如,指令的集合)不需要被实现为单独的软件程序、过程、数据集或模块,并且因此这些模块和数据的各种子集可以在各种实现方式中被组合或以其他方式重新布置。在一些实现方式中,非持久性存储器107可选地存储以上鉴定的模块和数据结构的子集。此外,在一些实施例中,存储器存储以上未描述的附加模块和数据结构。在一些实施例中,以上标识的元素中的一个或多个存储在计算机系统中,而不是存储在系统100的计算机系统中,该计算机系统可由系统100寻址,使得系统100可在需要时检索全部或一部分此类数据。

[0228] 尽管图1描绘了“系统100”,但是该图更多地旨在作为可能存在于计算机系统中的各种特征的功能描述,而不是作为本文所描述的实施方案的结构示意图。在实践中,并且如本领域普通技术人员所认识到的,单独示出的项目可以被组合并且一些项目可以被分离。此外,尽管图1描绘了非持久性存储器107中的某些数据和模块,但是这些数据和模块中的一些或全部可以存储在持久性存储器109中或一个以上的存储器中。举例来说,在一些实施例中,至少化合物结构数据存储区120和激活数据结构150存储在远程存储装置中,该远程存储装置可以为基于云的基础设施的一部分。在一些实施例中,至少化合物结构数据存储区120和激活数据结构150存储在基于云的基础设施上。在一些实施例中,化合物结构数据存储区120和激活数据结构150也可以存储在一个或多个远程存储装置中。

[0229] 虽然已经参考图1公开了根据本公开的系统,但是现在参考图2、3、7、8、9和14详细描述根据本公开的方法200、300、700、800、900和1500。

[0230] II.将测试化学化合物与目的生理状况相关联的方法

[0231] 生理状况。

[0232] 参考图3A至3E,本公开的一个方面提供了一种将测试化学化合物与目的生理状况相关联的方法300。

[0233] 在一些实施例中,目的生理状况为疾病。

[0234] 在一些实施例中,该疾病选自由以下组成的组:传染病或寄生虫病;肿瘤;血液或造血器官的疾病;免疫系统的疾病;内分泌、营养或代谢疾病;精神、行为或神经发育障碍;睡眠-觉醒障碍;神经系统的疾病;视觉系统的疾病;耳朵或乳突的疾病;循环系统的疾病;呼吸系统的疾病;消化系统的疾病;皮肤的疾病;肌肉骨骼系统或结缔组织的疾病;泌尿生殖系统的疾病;与性健康有关的病症;与怀孕、分娩或产褥期有关的疾病;某些源自围产期的病症;以及发育异常。在一些实施例中,疾病为ICD-11MMS或国际疾病分类的一个或多个条目。ICD提供了一种对疾病、伤害和死亡原因进行分类的方法。世界卫生组织(WHO)发布了ICD,以使记录和追踪确诊疾病的实例的方法标准化。

[0235] 在一些实施例中,目的生理状况为疾病刺激物,诸如疾病先决条件或共病。

[0236] 在一些实施例中,目的生理状况出现在细胞系统中或者是在细胞系统的背景下测量的。在一些实施例中,目的生理状况出现在一个或多个细胞中或者是在一个或多个细胞的背景下测量的,其中该一个或多个细胞包括单细胞、细胞系、活检样品细胞和/或经培养的原代细胞。在一些实施例中,目的生理状况为人细胞中出现的生理状况。在一些实施例中,目的生理状况为样品(诸如本文所描述的任何样品(参见,例如,定义:样品))中出现的

生理状况。在一些实施例中,目的生理状况为受试者(诸如人或动物)中出现的生理状况。

[0237] 在一些实施例中,目的生理状况为目的细胞过程或者与其相关。

[0238] 在一些实施例中,目的细胞过程为畸变细胞过程。在一些实施例中,目的细胞过程为与疾病相关的细胞过程。举例来说,如上面所描述,在一些实施例中,该方法提供对疾病至关重要的细胞过程和程序的靶向和阐明。在一些实施例中,目的细胞过程指示疾病的任何特性背后的机制(包括但不限于疾病的发作、进展、症状、严重性和/或消退)或者与其相关。在一些实施例中,目的细胞过程为功能途径。在一些实施例中,目的细胞过程为信号转导途径。在一些实施例中,目的细胞过程为(例如,化合物、小分子和/或治疗剂)作用机制。在一些实施例中,目的细胞过程由转录网络(例如,基因调控网络)表征和/或调节。在一些实施例中,目的细胞过程为在第一细胞状态与第二细胞状态之间的转变期间出现的细胞过程。

[0239] 在一些实施例中,目的细胞过程为注释,诸如基因集富集测定(GSEA)注释、基因本体注释、功能和/或信号转导途径注释和/或细胞特征注释。注释可以获自任何公共知识数据库,包括但不限于NIH基因表达汇编(Gene Expression Omnibus)(GEO)、EBI ArrayExpress、NCBI、BLAST、EMBL-EBI、GenBank、Ensembl、KEGG途径数据库、基于网络的集成细胞特征库(Library of Integrated Network-based Cellular Signatures)(LINCS)L1000数据集、Reactome途径数据库、基因本体(Gene Ontology)项目和/或任何疾病特定数据库。

[0240] 因此,在一些实施例中,目的生理状况为如本文所描述的任何相应的疾病、功能途径、信号转导途径、作用机制、转录网络、差别和/或细胞或生物学过程。

[0241] 在一些实施例中,目的生理状况为表型。举例来说,在一些实施例中,目的生理状况为化合物、小分子和/或治疗剂的生理表现,诸如毒性和/或疾病的消退。在一些实施例中,生理状况为使用实验数据测量的表型,该实验数据包括但不限于流式细胞术读数、成像和显微镜注释(例如,H&E载玻片、IHC载玻片、放射学图像和/或其他医学成像),以及/或者细胞组分数据。

[0242] 在一些实施例中,目的生理状况为毒性的测度。在一些实施例中,生理状况为核受体的抑制或激活,和/或核受体的抑制量或激活量。在一些实施例中,生理状况为生物学途径(例如,应激反应途径)的抑制或激活和/或抑制量或激活量。描述了大约10,000种化合物的可用于本公开的示例核受体和示例应激反应途径以及这些核受体和示例应激反应途径的抑制或激活数据,如以下中所描述:Huang等人,2016,“Modelling the Tox21 10K chemical profiles for in vivo toxicity prediction and mechanism characterization,”Nat Commun.7,第10425页,其特此通过引用并入。

[0243] 在一些实施例中,目的生理状况的特征在于细胞组分(例如,细胞组分模块)的集合的激活和/或扰动特征(例如,多种分析物响应于扰动的差异表达图谱)。

[0244] 举例来说,在一些实施例中,目的生理状况为包括细胞组分的集合的细胞组分模块。设想任何类型的分析物(例如,基因、转录本、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合)均用于相应的细胞组分模块中的细胞组分的集合中。在一些实施例中,细胞组分模块与本领域已知的任何细胞或生物学过程及其任何畸变相关,如对于本领域技术人员来说显而易见的。适合与当前公开的系统和方法一起使用的细胞组分模块进一步描述

在下面标题为“细胞组分和细胞组分模块”的章节中。

[0245] 在一些实施例中,目的生理状况为特征在于第一细胞状态与第二细胞状态之间的差别的扰动特征(例如,细胞状态转变特征)。

[0246] 在一些此类实施例中,目的生理状况通过患病的状态(例如,从患病的受试者和/或患病的组织获得的细胞)与健康的状态(例如,从健康的或对照受试者和/或组织获得的细胞)之间的差别来识别。举例来说,在一些实施例中,患病的状态通过细胞的功能的丧失、细胞的功能的获得、细胞的进展(例如,细胞转变为分化的状态)、细胞的停滞(例如,细胞不能转变为分化的状态)、细胞的侵入(例如,细胞出现在异常位置)、细胞的消失(例如,细胞在细胞正常存在的位置不存在)、细胞的紊乱(例如,细胞内和/或细胞周围的结构、形态和/或空间变化)、细胞的网络的丧失(例如,细胞的消除后代细胞或该细胞下游的细胞中的正常效应的变化)、细胞的网络的获得(例如,细胞的触发该细胞下游的细胞中的后代细胞中的新下游效应的变化)、细胞的过剩(例如,细胞的超丰)、细胞的不足(例如,细胞的低于临界阈值的密度)、细胞中细胞组分比率和/或量的差异、细胞中转变速率的差异,或其任何组合。

[0247] 适合与当前公开的系统和方法一起使用的扰动特征进一步描述于下面标题为“扰动特征”的章节中。

[0248] 在一些实施例中,目的生理状况包括多种生理状况(例如,细胞过程、细胞组分模块和/或扰动特征)。在一些实施例中,目的生理状况包括至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少15、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90或至少100种生理状况。在一些实施例中,目的生理状况包括不超过200、不超过100、不超过90、不超过80、不超过70、不超过60、不超过50、不超过20或不超过10种生理状况。在一些实施例中,目的生理状况包括1至5、5至10、2至20、10至50或20至100种生理状况。在一些实施例中,目的生理状况包括落入以不低于3种生理状况开始并以不高于200种生理状况结束的另一范围内的多种生理状况。

[0249] 在一些实施例中,本公开的化合物为满足里宾斯基五规则标准的化学化合物。在一些实施例中,本公开的化合物为满足以下里宾斯基五规则中的两条或更多条规则、三条或更多条规则或所有四条规则的有机化合物:(i)不超过五个氢键供体(例如,OH和NH基团),(ii)不超过十个氢键受体(例如,N和O),(iii)分子量低于500道尔顿,以及(iv)LogP低于5。之所以称为“五倍率法则”,是因为四项标准中的三项都涉及数字五。参见,Lipinski,1997,Adv.Drug Del.Rev.23,3,所述文献特此通过引用以其整体并入本文。在一些实施例中,除了里宾斯基五规则外,本公开的化合物还满足一个或多个标准。举例来说,在一些实施例中,本公开的化合物具有五个或更少个芳香族环、四个或更少个芳香族环、三个或更少个芳香族环,或两个或更少个芳香族环。

[0250] 参考框302,方法300包括获得测试化学化合物的化学结构的指纹。

[0251] 举例来说,在一些实现方式中,将测试化学化合物应用于机器学习方法包括将分子数据(例如,化合物的化学结构)变换为可以由机器学习模型读取和操纵的格式。

[0252] 参考图3A的框304,将化学结构变换为机器学习可读格式的一种方法包括使用简化分子输入行输入系统(SMILES)(其将分子表示为文本字符串)来确定化学结构的“指纹”。因此,在一些实施例中,该方法进一步包括根据测试化学化合物的简化分子输入行输入系

统(SMILES)字符串表示来计算指纹。使用SMILES字符串的分子指纹分析进一步描述于例如以下中:Honda等人,2019,“SMILES Transformer:Pre-trained Molecular Fingerprint for Low Data Drug Discovery,”arXiv:1911.04738,其特此通过引用以其整体并入本文。

[0253] 将化学结构变换为机器学习可读格式的另一方法包括确定基于图的分子指纹。在基于图的分子指纹分析中,初始分子结构由图表示,在该图中节点代表各个原子,并且边代表原子之间的键。基于图的方法提供数个优点,包括能够以较低的大小要求高效地编码多个子结构并因此降低计算负担,以及能够编码对指纹之间的结构相似性的指示。基于图的指纹分析进一步描述于例如以下中:Duvenaud等人,2015,“Convolutional networks on graphs for learning molecular fingerprints,”NeurIPS,第2224-2232页,其特此通过引用以其整体并入本文。在一些实施例中,指纹是从图卷积网络生成的。在一些实施例中,指纹是从空间图卷积网络,诸如图注意力网络(GAT)、图同构网络(GIN)或基于图子结构索引的近似图(SAGA)生成的。在一些实施例中,指纹是从谱图卷积网络,诸如采用切比雪夫多项式滤波(Chebyshev polynomial filtering)的谱图卷积生成的。

[0254] 参考图3A的框306,在一些实施例中,指纹是使用SMILES变换器、ECFP4、RNNS2S和/或GraphConv从化学结构生成的。

[0255] 模型架构。

[0256] 参照图3B的框308,该方法包括将指纹输入到模型中。在一些实施例中,模型包括多个(例如,100、200、300、500、1000、10,000或更多个)参数。

[0257] 在一些实施例中,模型包括多个参数(例如,权重和/或超参数)。在一些实施例中,用于模型的多个参数包括至少10、至少50、至少100、至少500、至少1000、至少2000、至少5000、至少10,000、至少20,000、至少50,000、至少100,000、至少200,000、至少500,000、至少1百万、至少2百万、至少3百万、至少4百万或至少5百万个参数。在一些实施例中,用于模型的多个参数包括不超过8百万、不超过5百万、不超过4、不超过1百万、不超过500,000、不超过100,000、不超过50,000、不超过10,000、不超过5000、不超过1000或不超过500个参数。在一些实施例中,用于模型的多个参数包括10至5000、500至10,000、10,000至500,000、20,000至1百万或1百万至5百万个参数。在一些实施例中,用于模型的多个参数落入从不低于10个参数开始到不高于8百万个参数结束的另一范围内。

[0258] 在一些实施例中,模型的训练进一步由一个或多个超参数(例如,可以在训练期间调适的一个或多个值)来表征。在一些实施例中,超参数值是在训练期间调适(例如,调整)的。在一些实施例中,超参数值是基于训练数据集的特定元素和/或一个或多个输入(例如,细胞、细胞组分模块、协变量等)来确定的。在一些实施例中,超参数值是使用实验优化来确定的。在一些实施例中,超参数值是使用超参数扫描来确定的。在一些实施例中,超参数值是基于现有模板或默认值来指派的。

[0259] 在一些实施例中,一个或多个超参数中的相应的超参数包括学习率。在一些实施例中,学习率为至少0.0001、至少0.0005、至少0.001、至少0.005、至少0.01、至少0.05、至少0.1、至少0.2、至少0.3、至少0.4、至少0.5、至少0.6、至少0.7、至少0.8、至少0.9或至少1。在一些实施例中,学习率不超过1、不超过0.9、不超过0.8、不超过0.7、不超过0.6、不超过0.5、不超过0.4、不超过0.3、不超过0.2、不超过0.1、不超过0.05或不超过0.01,或更小。在一些实施例中,学习率为0.0001至0.01、0.001至0.5、0.001至0.01、0.005至0.8或0.005至1。在

一些实施例中,学习率落入从不低于0.0001开始到不高于1结束的另一范围内。在一些实施例中,一个或多个超参数进一步包括正则化强度(例如,L2权重惩罚、丢弃率等)。举例来说,在一些实施例中,模型(例如,神经网络)使用对多个隐藏神经元中的每个隐藏神经元的对应的参数(例如,权重)的正则化来训练。在一些实施例中,正则化包括L1或L2惩罚。

[0260] 在一些实施例中,一个或多个超参数中的相应的超参数为损失函数。在一些实施例中,损失函数为均方误差、扁平化均方误差、二次损失、平均绝对误差、平均偏差误差、折页(hinge)、多类支持向量机和/或交叉熵。在一些实施例中,损失函数为梯度下降算法和/或最小化函数。

[0261] 在一些实施例中,模型与一个或多个激活函数相关联。在一些实施例中,一个或多个激活函数中的激活函数为双曲正切函数(tanh)、sigmoid函数、softmax函数、高斯函数、玻尔兹曼加权平均函数、绝对值函数、线性函数、修正线性单元(ReLU)函数、有界修正线性函数、软修正线性函数、参数化修正线性函数、平均值函数、max函数、min函数、符号函数、平方函数、平方根函数、多重二次函数、反二次函数、反多重二次函数、多调和样条函数、swish函数、mish函数、高斯误差线性单元(GeLU)函数和/或薄板样条函数。模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分。

[0262] 参考图3B的框310,在一些实施例中,模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。在一些实施例中,模型为回归器。在一些实施例中,模型为本文所公开的任何模型(参见,例如,定义:模型)。

[0263] 参考图3B的框312,在一些实施例中,模型包括神经网络。

[0264] 在一些实施例中,神经网络为具有ReLU激活的全连接神经网络。举例来说,在一些实施例中,模型为神经网络,该神经网络包括:对应的一个或多个输入,其中对应的一个或多个输入中的每个输入针对测试化学化合物的化学结构,对应的第一隐藏层包括对应的多个隐藏神经元,其中对应的多个隐藏神经元中的每个隐藏神经元(i)与多个输入中的每个输入全连接,(ii)与第一激活函数类型相关联,并且(iii)与用于神经网络的多个参数中的对应的参数(例如,权重);以及一个或多个对应的神经网络输出,其中对应的一个或多个神经网络输出中的每个相应的神经网络输出(i)直接或间接接收对应的多个隐藏神经元中的每个隐藏神经元的输出作为输入,并且(ii)与第二激活函数类型相关联。在一些此类实施例中,神经网络为全连接网络。

[0265] 在一些实施例中,神经网络包括多个隐藏层。如上面所描述,隐藏层位于输入层与输出层之间(例如,以捕获额外的复杂度)。在存在多个隐藏层的一些实施例中,每个隐藏层可以具有相同或不同相应数量的神经元。

[0266] 在一些实施例中,每个隐藏神经元(例如,在神经网络中的相应的隐藏层中)与对输入数据进行函数(例如,线性或非线性函数)的激活函数相关联。一般来说,激活函数的目的是将非线性引入到数据中,以便神经网络在初始数据的表示上进行训练并且随后可以“拟合”或生成新的(例如,以前未见过的)数据的附加表示。对激活函数(例如,第一和/或第二激活函数)的选择取决于神经网络的用例,因为某些激活函数(例如,双曲正切函数和/或sigmoid函数)可能导致数据集的极端处的饱和。举例来说,在一些实施例中,激活函数(例如,第一激活函数和/或第二激活函数)选自本领域已知的任何合适的激活函数,包括但不

限于本文所公开的任何激活函数。

[0267] 在一些实施例中,每个隐藏神经元进一步与基于激活函数确定的促成神经网络的输出的参数(例如,权重和/或偏差值)相关联。在一些实施例中,隐藏神经元用任意参数(例如,随机化的权重)来初始化。在一些替代实施例中,隐藏神经元用预定的参数集来初始化。

[0268] 在一些实施例中,神经网络中(例如,跨一个或多个隐藏层)的多个隐藏神经元为至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少11、至少12、至少13、至少14、至少15、至少16、至少17、至少18、至少19、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400或至少500个神经元。在一些实施例中,多个隐藏神经元为至少100、至少500、至少800、至少1000、至少2000、至少3000、至少4000、至少5000、至少6000、至少7000、至少8000、至少9000、至少10,000、至少15,000、至少20,000或至少30,000个神经元。在一些实施例中,多个隐藏神经元不超过30,000、不超过20,000、不超过15,000、不超过10,000、不超过9000、不超过8000、不超过7000、不超过6000、不超过5000、不超过4000、不超过3000、不超过2000、不超过1000、不超过900、不超过800、不超过700、不超过600、不超过500、不超过400、不超过300、不超过200、不超过100或不超过50个神经元。在一些实施例中,多个隐藏神经元为2至20、2至200、2至1000、10至50、10至200、20至500、100至800、50至1000、500至2000、1000至5000、5000至10,000、10,000至15,000、15,000至20,000或20,000至30,000个神经元。在一些实施例中,多个隐藏神经元落入从不低于2个神经元开始到不高于30,000个神经元结束的另一范围内。

[0269] 在一些实施例中,神经网络包括1至50个隐藏层。在一些实施例中,神经网络包括1至20个隐藏层。在一些实施例中,神经网络包括至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少11、至少12、至少13、至少14、至少15、至少16、至少17、至少18、至少19、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90或至少100个隐藏层。在一些实施例中,神经网络包括不超过100、不超过90、不超过80、不超过70、不超过60、不超过50、不超过40、不超过30、不超过20、不超过10、不超过9、不超过8、不超过7、不超过6或不超过5个隐藏层。在一些实施例中,神经网络包括1至5、1至10、1至20、10至50、2至80、5至100、10至100、50至100或3至30个隐藏层。在一些实施例中,神经网络包括落入从不低于1个层开始到不高于100个层结束的另一范围内的多个隐藏层。

[0270] 在一些实施例中,神经网络包括浅层神经网络。浅层神经网络是指具有少量隐藏层的神经网络。在一些实施例中,此类神经网络架构提高神经网络训练的效率并且由于训练中涉及减少数量的层而节省计算能力。在一些实施例中,神经网络包括一个隐藏层。在一些实施例中,神经网络包括两个、三个、四个或五个隐藏层。

[0271] 在一些实施例中,神经网络为消息传递神经网络。消息传递神经网络是指用于图上的有监督学习的框架(例如,化学结构的基于图的表示),在该图中,节点表示原子,并且边代表原子之间的键。一般来说,消息传递神经网络在正向传递中包括两个阶段:消息传递阶段和读出阶段。消息传递阶段运行T个间隔的周期,并且包括根据消息函数M_t和顶点更新函数U_t来更新图中的每个节点处的隐藏状态。读出阶段使用读出函数R来计算图的特征向量。在一些实施例中,消息传递神经网络包括卷积网络(例如,空间图卷积网络和/或谱图卷积网络)、门控图神经网络(GG-NN)、交互网络、分子图卷积、深度张量神经网络和/或基于拉普拉斯(Laplacian)的方法。参见,例如,Gilmer等人,2017,"Neural Message Passing for

Quantum Chemistry,"arXiv:1704.01212v2,其特此通过引用以其整体并入本文。

[0272] 参照图3B的框314,在一些实施例中,模型为多个成分模型的集成模型。举例来说,参考框316,在一些实施例中,一个或多个计算出的激活评分中的每个计算出的激活评分为多个成分模型中的每个成分模型的输出的集中趋势测度。

[0273] 参考图3B的框318,在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型、线性回归模型或多个神经网络。

[0274] 在一些实施例中,集成模型包括至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400或至少500个成分模型。在一些实施例中,集成模型包括不超过500、不超过400、不超过300、不超过200或不超过100个成分模型。在一些实施例中,集成模型包括不超过100、不超过50、不超过40、不超过30或不超过20个成分模型。在一些实施例中,集成模型包括介于1与50之间、介于2与20之间、介于5与50之间、介于10与80之间、介于5与15之间、介于3与30之间、介于10与500之间、介于2与100之间或介于50与100之间个的成分模型。在一些实施例中,集成模型包括另一范围的组件模型,以不低于2个成分模型开始并且以不高于500个成分模型结束。

[0275] 在一些实施例中,集成模型通过组合从多个成分模型获得的多个输出(例如,激活评分)来形成。在一些实施例中,来自分类器的多个输出(例如,激活评分)使用以下来进行组合:本领域已知的任何集中趋势测度,包括但不限于均值、中值、众数、加权均值、加权中值、加权众数、算术均值、中列数、中轴数、三均值和/或缩尾(Winsorized)均值。举例来说,来自集成模型的最终确定可以基于跨集成模型中所有成分模型的输出的平均值来获得。

[0276] 在一些实施例中,多个输出使用投票方法来组合。举例来说,在一些实施例中,多个输出通过以下来组合:统计来自集成模型中的每个成分模型的输出(例如,激活评分)的数量,该输出指示相应的化学结构和相应的目的生理状况之间的关联。在一些实施例中,来自成分模型的多个输出(例如,激活评分)是使用多数投票来组合的。在一些此类实施例中,来自成分模型的多个输出通过以下来组合:当对指示关联的输出的统计(例如,对超过阈值标准的激活评分的统计)大于投票阈值时,确定相应的化学结构与相应的目的生理状况之间的关联。在一些实施例中,投票阈值为来自集成模型中的多个成分模型的总投票的至少50%。在一些实施例中,投票阈值为来自集成模型中的多个成分模型的总投票的至少20%、至少30%、至少40%、至少50%、至少60%、至少70%、至少80%、至少90%或至少95%。

[0277] 在一些实施例中,集成模型中的每个成分模型是未加权的(例如,每个成分模型在集成模型中具有一票)。在一些实施例中,集成模型中的一个或多个成分模型被进一步加权(例如,在集成模型中具有大于1票)。

[0278] 在一些实施例中,该方法包括获得单个集成模型或多个集成模型。设想本领域已知的任何架构均用于集成模型。举例来说,在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。在一些实施例中,多个成分模型包括多个神经网络。

[0279] 参照图3B的框320,在一些实施例中,模型为多个神经网络的集成模型。参考图3B

的框322,在一些实施例中,模型为包括多个神经网络的集成模型,其中多个神经网络中的第一神经网络为具有ReLU激活的全连接神经网络,并且多个神经网络中的第二神经网络为消息传递神经网络。在一些此类实施例中,第一神经网络为全连接3层神经网络,其接受化学结构的呈SMILES表示形式的分子指纹作为输入。在一些实施例中,第二神经网络为消息传递神经网络(MPNN),其接受化学结构的呈基于图的表示形式的分子指纹作为输入。

[0280] 细胞组分和细胞组分模块。

[0281] 如上面所描述,再次参考框308,响应于将化学结构的指纹输入到模型中,模型输出针对细胞组分模块的集合的一个或多个计算出的激活评分。参考图3C的框326,一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块。

[0282] 参考图3C的框328,细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。

[0283] 在一些实施例中,细胞组分为基因、基因产物(例如,mRNA和/或蛋白质)、碳水化合物、脂质、表观遗传特征、代谢物和/或其组合。在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。在一些实施例中,多种细胞组分包括:核酸,包括DNA、修饰的(例如,甲基化的)DNA、RNA(包括编码(例如,mRNA)或非编码RNA(例如,sncRNA);蛋白质,包括转录后修饰的蛋白质(例如,磷酸化的、糖基化的、肉豆蔻酰化的蛋白质等);脂质;碳水化合物;核苷酸(例如,三磷酸腺苷(ATP)、二磷酸腺苷(ADP)和单磷酸腺苷(AMP)),包括环状核苷酸,诸如环状单磷酸腺苷(cAMP)和环状单磷酸鸟苷(cGMP);其他小分子细胞组分,诸如氧化和还原形式的烟酰胺腺嘌呤二核苷酸(NADP/NADPH);及其任何组合。

[0284] 在一些实施例中,多种细胞组分包括至少5、至少10、至少15、至少20、至少25、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少2000、至少3000、至少4000、至少5000、至少6000、至少7000、至少8000、至少9000、至少10,000、至少20,000、至少30,000、至少50,000,或超过50,000种细胞组分。在一些实施例中,多种细胞组分包括不超过70,000、不超过50,000、不超过30,000、不超过10,000、不超过5000、不超过1000、不超过500、不超过200、不超过100、不超过90、不超过80、不超过70、不超过60、不超过50或不超过40种细胞组分。在一些实施例中,多种细胞组分由介于二十种与10,000种之间的细胞组分组成。在一些实施例中,多种细胞组分由介于100种与8,000种之间的细胞组分组成。在一些实施例中,多种细胞组分包括5至20、20至50、50至100、100至200、200至500、500至1000、1000至5000、5000至10,000或10,000至50,000种细胞组分。在一些实施例中,多种细胞组分落入从不低于5种细胞组分开始到不高于70,000种细胞组分结束的另一范围内。

[0285] 作为示例,在一些实施例中,多种细胞组分包括任选地在RNA水平上测定的多个基因。在一些实施例中,多个基因包括至少5、至少10、至少15、至少20、至少25、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900或至少1000个基因。在一些实施例中,多个基因包含至少1000、至少2000、至少3000、至少4000、至少5000、至少10,000、至少30,000、至少50,000或多于50,000个基因。在一些实施例中,多个基因包括5至20、20至50、50至100、100至

200、200至500、500至1000、1000至5000、5000至10,000或10,000至50,000个基因。

[0286] 作为另一示例,在一些实施例中,多种细胞组分包括多种蛋白质。在一些实施例中,多种蛋白质包括至少5、至少10、至少15、至少20、至少25、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900或至少1000种蛋白质。在一些实施例中,多种蛋白质包含至少1000、至少2000、至少3000、至少4000、至少5000、至少10,000、至少30,000、至少50,000或多于50,000种蛋白质。在一些实施例中,多种蛋白质包括5至20、20至50、50至100、100至200、200至500、500至1000、1000至5000、5000至10,000或10,000至50,000种蛋白质。

[0287] 不要求细胞组分模块中的每种细胞组分均是唯一的。举例来说,考虑细胞组分模块A含有细胞组分1、3和10的情况。细胞组分模块的集合中的其他细胞组分模块也可以含有这些细胞组分。这里,术语"独立"意指特定细胞组分模块中的多种细胞组分的子集作为整体是唯一的。因此,考虑上面的示例细胞组分模块A,细胞组分模块的集合中的另一细胞组分模块可以含有细胞组分1、3和10,只要它进一步含有细胞组分模块A不含有的其他细胞组分。进一步考虑上面的示例细胞组分模块A,细胞组分模块的集合中的另一细胞组分模块可以限于细胞组分1、3和10的子集,但不要求它进一步含有细胞组分模块A不含有的其他细胞组分(但是,它也可能含有此类附加细胞组分)。

[0288] 在一些实施例中,细胞组分模块的集合中的每个细胞组分模块在多种细胞组分的相应独立子集中包括相同或不同数量的细胞组分。在一些实施例中,与每个相应的细胞组分模块相对应的细胞组分的每个相应独立子集是细胞组分的唯一子集(例如,非重叠的,其中多种细胞组分中的每种细胞成分被分组为不超过一个模块)。在一些实施例中,第一细胞组分模块具有细胞组分的第一子集,其与对应于第二细胞组分模块的细胞组分的第二子集重叠(例如,重叠的,其中多种细胞组分中的至少一种细胞组分是两个或更多个不同的模块所共同的)。

[0289] 参考图3C的框330,在一些实施例中,相应的细胞组分模块中的多种细胞组分的独立子集包括五种或更多种细胞组分。在一些实施例中,多个细胞组分模块中的相应的细胞组分模块中的多种细胞组分的独立子集包括至少2、至少5、至少10、至少15、至少20、至少25、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少2000或至少3000种细胞组分。在一些实施例中,多种细胞组分的独立子集包括不超过5000、不超过3000、不超过1000、不超过500、不超过200、不超过100、不超过90、不超过80、不超过70、不超过60或不超过50种细胞组分。在一些实施例中,多种细胞组分的独立子集包括5至100、2至300、20至500、200至1000或1000至5000种细胞组分。在一些实施例中,多种细胞组分的独立子集落入从不低于2种细胞组分开始到不高于5000种细胞组分结束的另一范围内。

[0290] 在一些实施例中,相应的细胞组分模块中的多种细胞组分的独立子集由与目的生理状况相关联的细胞过程(例如,分子途径)中的细胞组分组成。举例来说,参考图3C的框332,在一些实施例中,相应的细胞组分模块中的多种细胞组分的独立子集由与目的生理状况相关联的分子途径中的介于两种与20种之间的细胞组分组成。

[0291] 参考图3D的框334,细胞组分模块的集合中的至少第一细胞组分模块与目的生理状况相关联。事实上,细胞组分模块中的许多细胞组分模块可以与目的生理状况相关联。

[0292] 参考图3D的框336,一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块。

[0293] 参考图3D的框338,细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。

[0294] 参考图3D的框340,在一些实施例中,细胞组分模块的集合为多个细胞组分模块。包括第一细胞组分模块的多个细胞组分模块的第一子集与目的生理状况相关联。也就是说,此类细胞组分模块表示在目的生理状况中涉及的细胞组分。举例来说,相对于某种基线、野生型状态下的细胞,在表示目的生理状况的细胞中可以下调或上调此类细胞组分模块的此类细胞组分。此外,多个细胞组分模块的第二子集与目的生理状况不关联。也就是说,此类细胞组分模块的细胞组分表示在目的生理状况中未涉及的细胞组分。举例来说,相对于某种基线、野生型状态下的细胞,在表示目的生理状况的细胞中没有下调或上调此类细胞组分。在此类实施例中,当针对第一细胞组分模块(其在细胞组分模块的第一子集中)的相应的计算出的激活评分满足第一阈值标准并且针对多个细胞组分模块的第二子集中的细胞组分模块的相应的计算出的激活评分满足第二阈值标准时,化学化合物与目的生理状况关连。下面就图3E的框348来讨论示例第一阈值标准。一般来说,所寻求的是与细胞组分模块的第一子集中的细胞组分模块关连(如通过计算出的激活评分满足第一阈值所证明)但与细胞组分模块的第二子集中的细胞组分模块不关连(如通过计算出的激活评分满足第二阈值所证明)的化学化合物。举例来说,在一些实施例中,满足第一阈值需要激活评分高于第一预定的数值,而满足第二阈值需要激活评分低于第二预定的数值,其中确切的第一和第二预定的数值取决于应用。

[0295] 如上面所指示,在一些实现方式中,该方法包括使用一种或多种类型的分子数据(例如,细胞组分)来表征目的生理状况(例如,细胞过程)。此类分子数据可以包括具有可测量属性(例如,丰度和/或表达水平)的任何分析物,诸如组学图谱(例如,转录组学、蛋白质组学、代谢组学等)。

[0296] 一般来说,当与细胞过程相关联时,细胞组分(例如,基因)的细胞组分模块可以被认为是由一系列切换事件产生的,其中在相似时间切换的细胞组分(例如,基因)一起形成模块。因此,举例来说,在一些实施例中,相应的细胞组分模块包括多种细胞组分的相应子集,其中细胞组分的子集是基于与相应的目的生理状况(例如,目的细胞过程)相关联的行为的相似性来分组的。在示例中,与相应的目的生理状况相关联的细胞组分模块可以包括跨具有相应的生理状况的多种细胞类型行为相似(例如,表现出相似的表达图谱)的基因的子集。

[0297] 参考图3D的框342,在一些实施例中,细胞组分模块的集合由第一细胞组分模块组成。

[0298] 参考图3D的框344,在一些实施例中,细胞组分模块的集合包括五个或更多个细胞组分模块。在一些实施例中,细胞组分模块的集合包括20或更多、30或更多、40或更多、50或更多、60或更多、70或更多、80或更多、90或更多,或者100或更多个细胞组分模块。

[0299] 在一些实施例中,细胞组分模块的集合包括至少5、至少10、至少15、至少20、至少25、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少2000、至少3000、至

少4000或至少5000个细胞组分模块。在一些实施例中,细胞组分模块的集合包括不超过10,000、不超过5000、不超过2000、不超过1000、不超过500、不超过300、不超过200、不超过100、不超过90、不超过80、不超过70、不超过60或不超过50个细胞组分模块。在一些实施例中,细胞组分模块的集合由介于10个与2000个之间的细胞组分模块组成。在一些实施例中,细胞组分模块的集合由介于50个与500个之间的细胞组分模块组成。在一些实施例中,细胞组分模块的集合包括5至20、20至50、50至100、100至200、200至500、500至1000、1000至5000或5000至10,000个细胞组分模块。在一些实施例中,细胞组分模块的集合落入从不低于5个细胞组分模块开始到不高于10,000个细胞组分模块结束的另一范围内。

[0300] 在一些实施例中,该方法进一步包括识别与目的生理状况相关联的细胞组分模块。下面结合图14A至14D在标题为识别细胞组分模块的章节中讨论此类方法。

[0301] 激活评分。

[0302] 如图3B的框308中所描述,模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分。一般来说,经训练的模型(框308的模型)的输出是通过以下过程来定义的:对包括标记(例如,数值激活评分)的训练数据集进行学习并调整多个参数直至经训练的模型的输出满足(诸如通过验证步骤)最低性能水平。下面在标题为“模型训练”的章节中进一步公开对模型进行训练。

[0303] 在一些实施例中,一个或多个计算出的激活评分中的激活评分为针对与相应的化合物相对应的相应的细胞组分模块的相应的激活权重。举例来说,在一些实施例中,激活评分为如在下面标题为“识别细胞组分模块”的章节中参考图2A至2B和14A至14D所描述获得的并且在图5中的激活数据结构中展示的激活权重,其中激活评分指示相应的(例如,第一)细胞组分模块的与用相应的化合物进行处理相关和/或响应于用相应的化合物进行处理的激活(例如,诱导和/或差异表达)。

[0304] 因此,在一些此类实施例中,经训练的模型提供计算出的激活评分作为输出,该计算出的激活评分指示测试化学化合物与目的生理状况的关联(例如,第一细胞组分模块与目的生理状况相关联)。然后参考图3E的框348,该方法包括当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连(例如,确定其关联)。

[0305] 参考图3E的框350,在一些实施例中,第一阈值标准为以下要求:第一细胞组分模块具有阈值激活评分。一般来说,所寻求的是与目的生理状况关连(如通过计算出的激活评分满足第一阈值所证明)的化学化合物。举例来说,在一些实施例中,满足第一阈值需要激活评分高于第一预定的数值。

[0306] 举例来说,在一些实施例中,激活评分被表示为“0”与“1”之间(或者某个其他范围“A”至“B”,其中A和B为两个不同的数字)的归一化的连续值,其中较接近于“1”的值(例如,.89、.90、.91、.92等)指示细胞组分模块(以及细胞组分模块所表示的化学化合物)与目的生理状况之间有强关联。较接近于“0”的值(例如,0.01、0.02、0.03、0.04等)指示细胞组分模块(以及细胞组分所表示的化学化合物)与目的生理状况之间没有关联。在此类情况下,在“0”与“1”之间(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)选择第一阈值,并且当激活评分高于第一阈值时,细胞组分模块(以及其所表示的化学结构)被认为与目的生理状况相关联,而当激活评分低于第一阈值时,细胞组分模块(以及其所表示的化学结

构)被认为与目的生理状况不关联。在一些此类实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)内的归一化的值,并且第一阈值为0与1之间、0.10与0.90之间、0.20与0.80之间、0.30与0.70之间、0.50与0.99之间、0.60与0.99之间、0.70与0.99之间、0.80与0.99之间或0.90与0.99之间的值。

[0307] 作为另一示例,在一些实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或者某个其他范围“A”至“B”,其中A和B为两个不同的数字)内的归一化的值,其中较接近于“1”的值(例如,.89、.90、.91、.92等)指示细胞组分模块(以及细胞组分模块所表示的化学化合物)与目的生理状况之间没有关联。较接近于“0”的值(例如,0.01、0.02、0.03、0.04等)指示细胞组分模块(以及细胞组分所表示的化学化合物)与目的生理状况之间有关联。在此类情况下,在“0”与“1”之间(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)选择第一阈值,并且当激活评分低于第一阈值时,细胞组分模块(以及其所表示的化学结构)被认为与目的生理状况相关联,而当激活评分高于第一阈值时,细胞组分模块(以及其所表示的化学结构)被认为与目的生理状况不关联。在一些此类实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)内的归一化的值,并且第一阈值为0与1之间、0.10与0.90之间、0.20与0.80之间、0.30与0.70之间、0.50与0.99之间、0.60与0.99之间、0.70与0.99之间、0.80与0.99之间或0.90与0.99之间的值。

[0308] 参考图3E的框352,在一些实施例中,细胞组分模块的集合为多个细胞组分模块(例如,介于两个与1000个之间、介于10个与100个之间、介于2个与100个之间、介于4个与50个之间的细胞组分模块),并且框348的关连要求,细胞组分模块的集合中的每个细胞组分模块的相应的计算出的激活评分满足第一阈值标准。举例来说,考虑细胞组分模块的集合由两个细胞组分模块A和B组成的情况。图3E的框352要求,细胞组分模块A和B的激活评分各自满足第一阈值条件。举例来说,考虑这样的情况,其中细胞组分模块A具有计算出的激活评分0.25,细胞组分模块B具有计算出的激活评分0.75,满足第一阈值条件要求每个激活评分大于0.4。在这种情况下,细胞组分模块的集合不满足图3E的框352的要求,因为每个激活评分不大于要求阈值0.4。

[0309] 参考图3E的框354,在一些实施例中,细胞组分模块的集合为多个细胞组分模块(例如,介于两个与1000个之间、介于10个与100个之间、介于2个与100个之间、介于4个与50个之间的细胞组分模块),并且框348的关连要求,跨细胞组分模块的集合中的每个细胞组分模块的相应的计算出的激活评分的集中趋势测度满足第一阈值标准。举例来说,考虑细胞组分模块的集合由两个细胞组分模块A和B组成的情况。图3E的框354要求,细胞组分模块A和B的激活评分的某个集中趋势测度满足第一阈值条件。举例来说,考虑这样的情况,其中集中趋势测度为求平均,细胞组分模块A具有计算出的激活评分0.25,细胞组分模块B具有计算出的激活评分0.75,并且满足第一阈值条件要求平均激活评分大于0.4。在这种情况下,细胞组分模块的集合满足图3E的框354的要求,因为它们具有0.25+0.75/2或0.5的平均激活评分,其大于要求阈值0.4。在一些实施例中,集中趋势测度为细胞组分模块的集合中的每个细胞组分模块的每个相应的计算出的激活评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0310] 化合物。

[0311] 在一些实施例中,测试化学化合物为小分子、生物制品、蛋白质、与小分子组合的蛋白质、ADC、诸如siRNA或干扰RNA等的核酸、过表达野生型和/或突变型shRNA的cDNA、过表达野生型和/或突变型向导RNA的cDNA(例如,Cas9系统或其他细胞成分编辑系统),和/或任何前述的任何组合。

[0312] 在一些实施例中,测试化学化合物是无机或有机的。

[0313] 举例来说,参考图3E的框356,在一些实施例中,测试化学化合物为具有小于2000道尔顿(Da)的分子量的有机化合物。在一些实施例中,测试化学化合物具有至少10Da、至少20Da、至少50Da、至少100Da、至少200Da、至少500Da、至少1kDa、至少2kDa、至少3kDa、至少5kDa、至少10kDa、至少20kDa、至少30kDa、至少50kDa、至少100kDa或至少500kDa的分子量。在一些实施例中,测试化学化合物具有不超过1000kDa、不超过500kDa、不超过100kDa、不超过50kDa、不超过10kDa、不超过5kDa、不超过2kDa、不超过1kDa、不超过500Da、不超过300Da、不超过100Da或不超过50Da的分子量。在一些实施例中,测试化学化合物具有10Da至900Da、50Da至1000Da、100Da至2000Da、1kDa至10kDa、5kDa至500kDa或100kDa至1000kDa的分子量。在一些实施例中,测试化学化合物具有落在从不低于10道尔顿开始到不高于1000kDa结束的另一范围内的分子量。

[0314] 参考图3E的框358,在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。里宾斯基五规则(例如,RO5)标准为指南集,其用于评估药物相似性,诸如以确定具有相应的药理或生物活性的相应的化合物是否具有适合人类中施用的对应的化学或物理性质。里宾斯基五规则包括以下用于确定化合物的药物相似性的标准:(i)分子量小于500Da,(ii)不超过5个氢键供体,(iii)不超过10个氢键受体,以及(iv)辛醇-水分配系数log P不大于5。

[0315] 参考图3E的框360,在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的至少两、三或四个标准的有机化合物。在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的零、一、二、三或所有四个标准的有机化合物。

[0316] 在一些实施例中,测试化学化合物选自数据库。提供来自药物筛选的结果、注释和/或一般信息(诸如化合物靶点和化合物的化学性质)的合适化合物数据库的示例包括但不限于药物癌症敏感性基因组学、癌症治疗响应门户、联系图(Connectivity Map)、PharmacoDB、生物同位可交换替换物库(Base of Bioisosterically Exchangeable Replacements)(BoBER)和/或药物库(DrugBank)。在一些实施例中,测试化学化合物选自提供关于基因和基因产物、扰动诱导的细胞组分特征和/或途径注释的信息的数据库。注释可以获自任何公共知识数据库,包括但不限于NIH基因表达汇编(GEO)、EBI ArrayExpress、NCBI、BLAST、EMBL-EBI、GenBank、Ensembl、KEGG途径数据库、基于网络的集成细胞特征库(LINCS)L1000数据集、Reactome途径数据库和/或基因本体项目。

[0317] 在实际应用中使用方法300的结果。

[0318] 在一些实施例中,上面结合图3描述的方法300用于针对目的生理状况评估多种测试化学化合物。在此类实施例中,使多种测试化合物中的每种测试化合物均经过图3的方法300。因此,如果存在100种测试化合物和一种目的生理状况,则在此类实施例中,方法300运行100次,其中100次中的每个实例用于测试化合物中不同的一种。

[0319] 此外,在一些实施例中,上面结合图3描述的方法300用于评估针对多种目的生理

状况的多种化合物。在此类实施例中,对于每种目的生理状况,使多种测试化合物中的每种相应的每种测试化合物均经过图3的方法300。因此,如果存在100种测试化合物和两种目的生理状况,则在此类实施例中,方法300运行200次,其中200次中的每个实例用于测试化合物中针对第一或者第二目的生理状况的不同的一种。

[0320] 在一些实施例中,多种测试化合物包括至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少100、至少200、至少300、至少400、至少500、至少800、至少1000、至少2000、至少3000、至少4000、至少5000、至少8000、至少10,000、至少20,000、至少30,000、至少50,000、至少80,000、至少100,000、至少200,000、至少500,000、至少800,000、至少1百万或至少2百万种测试化合物,并且存在单种目的生理状况。在一些此类实施例中,方法300运行至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少100、至少200、至少300、至少400、至少500、至少800、至少1000、至少2000、至少3000、至少4000、至少5000、至少8000、至少10,000、至少20,000、至少30,000、至少50,000、至少80,000、至少100,000、至少200,000、至少500,000、至少800,000、至少1百万或至少2百万次,以产生至少5、至少10、至少15、至少20,至少30、至少40、至少50、至少100、至少200、至少300、至少400、至少500、至少800、至少1000、至少2000、至少3000、至少4000、至少5000、至少8000、至少10,000、至少20,000、至少30,000、至少50,000、至少80,000、至少100,000、至少200,000、至少500,000、至少800,000、至少1百万或至少2百万个激活评分,每种测试化合物一个激活评分。

[0321] 在一些实施例中,多种化合物包括不超过1千万、不超过5百万、不超过1百万、不超过500,000、不超过100,000、不超过50,000、不超过10,000、不超过8000、不超过5000、不超过2000、不超过1000、不超过800、不超过500、不超过200或不超过100种测试化合物。在一些实施例中,多种化合物由10至500、100至10,000、5000至200,000或10,000至1百万种测试化合物组成。

[0322] 在一些实施例中,多种测试化合物介于10种与1×10^6种测试化合物之间。在一些实施例中,多种测试化合物介于100种与100,000种测试化合物之间。在一些实施例中,多种测试化合物介于1000种与100,000种测试化合物之间。

[0323] 因此,方法300可以用于获得针对大量测试化合物的激活评分。针对这些激活评分应用第一阈值可以用于从经过测试的许多测试化合物中识别与目的生理状况相关联的测试化合物。在典型的实施例中,选定数量的测试化合物具有指示它们与目的生理状况相关联的激活评分,而其他测试化合物则不具有。对选定数量的测试化合物的分析可以用于确定测试化合物的导致与目的生理状况的关联的分子性质。举例来说,可以对选定数量的测试化合物(其具有指示其与目的生理状况相关联的激活评分)的化学结构进行针对其结构的相似性的目视检查,该相似性将该选定数量的测试化合物与不与目的生理状况相关联的测试化合物区分开来。然后可以将此类分子性质并入新测试分子中,该新测试分子未包括在由模型601评估的初始测试分子中,并且不用于训练模型601。

[0324] 此外,可以使用更正式的方法来分析测试化合物(满足以及不满足由方法300施加的第一阈值的测试化合物)。举例来说,可以使用子结构挖掘来识别测试化合物内的子结构,该子结构使此类化合物与目的生理状况相关联。子结构挖掘的示例包括但不限于MOSS(Borgetl和Meinl,2006“Full Perfect Extension Pruning for Frequent Graph Mining,”Proc.Workshop on Mining Complex Data(MCD 2006at ICDM 2006,Hong Kong,

China,IEEE Press,Piscataway,NJ,USA,其特此通过引用并入,以及MOFA(Meinl和Worlein,2006“Mining Molecular Datasets on Symmetric Processor Systems,”International conference on Systems,man and Cybernetics 2,第1269-1274页,其特此通过引用并入)。

[0325] 另外,可以使用最大公共子结构(MCS)分析来识别测试化合物内的子结构,该子结构使此类化合物与目的生理状况相关联。MCS分析的示例包括但不限于LIBMCS(Chemaxon,Library MCS,2008)、MCSS(OEChem TK版本2.0.0、OpenEye Scientific Software,Santa Fe,NM.http://www.eyesopen.com)和CncMCS(http://www.chemnavigator.

[0326] com/cnc/products/downloads.asp)。

[0327] 另外,可以使用SMARTS来识别测试化合物内的子结构,该子结构使此类化合物与目的生理状况相关联。SMART分析的示例为CDK Descriptor GUI。

[0328] 另外,可以使用频繁子图挖掘(Frequent Subgraph Mining)来识别测试化合物内的子结构,该子结构使此类化合物与目的生理状况相关联。频繁子图挖掘的示例为ParMol(埃尔朗根-纽伦堡大学(Uni Erlangen))。

[0329] 另外,可以使用图和化学挖掘来识别测试化合物内的子结构,该子结构使此类化合物与目的生理状况相关联。图和化学挖掘的示例为PAFI/AFGen(明尼苏达大学Karypis实验室)。

[0330] 扰动特征。

[0331] 如上面所描述,在一些实施例中,目的生理状况为扰动特征(例如,特征在于第一细胞状态与第二细胞状态之间的响应于扰动的差别)。因此,本公开的另一方面提供了一种将测试化学化合物与目的生理状况相关联的方法700。在一些实施例中,目的生理状况为疾病。

[0332] 参考框702,该方法包括获得测试化学化合物的化学结构的指纹。设想了如上面标题为“生理状况”和“化合物”的章节中所公开的生理状况、化合物、指纹和/或获得指纹的方法的任何合适的实施例,包括其任何替换、修改、添加、删除和/或组合,如本领域技术人员显而易见的。

[0333] 举例来说,在一些实施例中,测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。在一些实施例中,测试化学化合物为满足里宾斯基五规则标准中的至少三个标准的有机化合物。在一些实施例中,该方法进一步包括根据测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算指纹。在一些实施例中,指纹是使用SMILES变换器、ECFP4、RNNS2S或GraphConv从化学结构生成的。

[0334] 参考框704,该方法进一步包括将指纹输入到模型中,其中该模型包括100个或更多个参数,模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分,并且一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征。

[0335] 设想了模型的任何合适的实施例,诸如上面标题为“模型架构”的章节中公开的那些实施例,以及其任何替换、修改、添加、删除和/或组合,如对本领域技术人员显而易见的。举例来说,在一些实施例中,模型包括神经网络。在一些此类实施例中,神经网络为具有

ReLU激活的全连接神经网络。在一些实施例中,神经网络为消息传递神经网络。

[0336] 在一些实施例中,模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0337] 在一些实施例中,模型为多个成分模型的集成模型,一个或多个计算出的激活评分中的每个计算出的激活评分为多个成分模型中的每个成分模型的输出的集中趋势测度。

[0338] 在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0339] 在一些实施例中,多个成分模型包括多个神经网络。在一些此类实施例中,多个神经网络中的第一神经网络为具有ReLU激活的全连接神经网络,并且多个神经网络中的第二神经网络为消息传递神经网络。

[0340] 如上面所定义,扰动是指细胞对一种或多种状况(诸如通过一种或多种化合物进行处理)的任何暴露。在一些实施例中,扰动特征为细胞中的一种或多种细胞组分的由扰动诱导的表达或丰度水平的变化。

[0341] 示例扰动包括但不限于基因敲低、针对刺激的细胞响应、组织生长及再生,和/或用化合物进行处理或暴露于化合物。示例扰动原包括但不限于小分子、生物制品、治疗剂、蛋白质、与小分子组合的蛋白质、ADC、核酸(诸如siRNA或干扰RNA、过表达野生型和/或突变型shRNA的cDNA、过表达野生型和/或突变型向导RNA的cDNA(例如,Cas9系统或其他基因编辑系统)),或任何前述的任何组合。

[0342] 在一些实施例中,扰动以系统水平(例如,结合或对接活性)和/或相对于下游效应和器官级表型来表征。在一些实施例中,扰动被表征为在分子、细胞和/或组织水平上对扰动原的响应的驱使或潜在机制的函数(例如,通过在扰动之前或之后识别或测量生物标志物、细胞活力和/或药物-蛋白质相互作用)。举例来说,扰动的测量结果可以包括表型测量结果(例如,IC50值)和/或细胞组分特征(例如,组学图谱)。

[0343] 在一些实施例中,相应的扰动和/或对应的扰动特征获自公开可用的数据库,诸如药物癌症敏感性基因组学、癌症治疗反应门户、联系图、PharmacoDB、生物同位可交换替换物库(BoBER)、药物库、人细胞图谱、分子特征数据库(MSigDB)和/或Enrichr。可以从其获得扰动数据的其他合适的数据库包括NIH基因表达汇编(GEO)、EBI ArrayExpress、NCBI、BLAST、EMBL-EBI、GenBank、Ensembl、KEGG途径数据库、基于网络的集成细胞特征库(LINCS)L1000数据集、Reactome途径数据库和/或基因本体项目。

[0344] 获得扰动数据的方法包括使用例如以下来测量细胞组分数据:perturb-seq、CRISP-seq、CROP-seq、CRISPRi、TAP-seq、CRISPRa、perturb-CITE-seq、sci-Plex、多重、MIX-seq、CyTOF和/或scRNA-seq。获得扰动数据的方法进一步包括获得组学数据的任何方法,包括质谱(例如,LCMS、GCMS)、流式细胞术、定量聚合酶链式反应(qPCR)、凝胶电泳、基因芯片分析、微阵列、细胞荧光分析、荧光显微术、共焦激光扫描显微术、激光扫描细胞术、亲和层析、手动批量模式分离、电场悬浮、测序和/或其任何组合。在一些实施例中,设想本文所公开的用于获得细胞组分丰度值的任何方法用于获得扰动数据(例如,针对扰动特征)。

[0345] 在一些实施例中,扰动特征的集合由第一扰动特征组成。在一些实施例中,扰动特

征的集合包括五个或更多个扰动特征。在一些实施例中,扰动特征的集合包括十个或更多个扰动特征。在一些实施例中,扰动特征的集合包括100个或更多个扰动特征。

[0346] 在一些实施例中,扰动特征的集合包括至少2、至少3、至少4、至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少800、至少1000、至少2000或至少5000个扰动特征。在一些实施例中,扰动特征的集合包括不超过10,000、不超过5000、不超过1000、不超过800、不超过500、不超过200、不超过100、不超过50或不超过20个扰动特征。在一些实施例中,扰动特征的集合包括5至50、2至100、20至500、10至1000、800至5000或50至2000个扰动特征。在一些实施例中,扰动特征的集合落入从不低于2个扰动特征开始到不高于10,000个扰动特征结束的另一范围内。

[0347] 参考框706,扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。

[0348] 在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由尚未暴露于多种化合物中的化合物的对照细胞表示。在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由跨已经暴露于多种化学化合物中的化学化合物(除了与相应的扰动特征相关联的化合物之外)的不相关的受扰动的细胞的平均数表示。

[0349] 在一些实施例中,细胞状态的变化是指未改变的细胞状态与改变的细胞状态之间的变化,其中改变的细胞状态通过从未改变的细胞状态到改变的细胞状态的细胞转变而出现。此外,(i)未改变的细胞状态、(ii)改变的细胞状态以及(iii)从未改变的细胞状态到改变的细胞状态的转变中的至少一者与目的生理状况相关联。

[0350] 在一些实施例中,作为非限制性示例,可以使用在以下中公开的任何方法来确定扰动特征的集合中的相应的扰动特征:2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691,其特此通过引用并入。

[0351] 在某些实施例中,可以存在扰动(例如,细胞暴露于特定化学组合物)的协变量。举例来说,化学组合物的协变量可以包括:化学组合物的特定剂量、测量暴露于化学组合物的细胞以量化细胞组分的时间和/或暴露于化学组合物的细胞的身份(例如,细胞系)。在一些实施例中,仅当扰动的协变量的阈值量也被预测影响特定细胞转变时,才预测扰动(例如,细胞暴露于特定化学组合物)影响特定细胞转变。换句话说,在一些实施例中,特定扰动特征的计算出的激活评分至少部分地通过以下来确定:有特定扰动特征的化学组合物的协变量是否也被预测影响与目的生理状况相关联的特定细胞转变。

[0352] 一般来说,如上面所描述,经训练的模型的输出是通过以下过程来定义的:对包括标记(例如,数值激活评分)的训练数据集进行学习并调整多个参数直至经训练的模型的输出满足(诸如通过验证步骤)最低性能水平。下面在标题为“模型训练”的章节中进一步公开对模型进行训练。因此,在一些此类实施例中,经训练的模型提供针对第一扰动特征的计算出的激活评分作为输出,该计算出的激活评分指示测试化学化合物与目的生理状况的关联

(例如,其中第一扰动特征与关联于目的生理状况的细胞状态转变相关联)。

[0353] 然后参考框708,该方法包括当针对扰动特征的集合中的第一扰动特征的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连。

[0354] 设想如上面标题为“激活评分”的章节中所公开的激活评分的任何合适的实施例用于获得一个或多个计算出的激活评分,其中每个激活评分表示扰动特征的集合中的对应的扰动特征,包括其任何替换、修改、添加、删除和/或组合,如对本领域技术人员显而易见的。

[0355] 一般来说,所寻求的是与目的生理状况关连(如通过计算出的激活评分满足第一阈值标准所证明)的化学化合物。举例来说,在一些实施例中,满足第一阈值需要激活评分高于第一预定的数值。

[0356] 举例来说,在一些实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或者某个其他范围“A”至“B”,其中A和B为两个不同的数字)的归一化的值,其中较接近于“1”的值(例如,.89、.90、.91、.92等)指示扰动特征(以及扰动特征所表示的化学化合物)与目的生理状况之间有强关联。较接近于“0”的值(例如,0.01、0.02、0.03、0.04等)指示扰动特征(以及扰动特征所表示的化学化合物)与目的生理状况之间没有关联。在此类情况下,在“0”与“1”之间(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)选择第一阈值,并且当激活评分高于第一阈值时,扰动特征(以及其所表示的化学结构)被认为与目的生理状况相关联,而当激活评分低于第一阈值时,扰动特征(以及其所表示的化学结构)被认为与目的生理状况不关联。在一些此类实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)内的归一化的值,并且第一阈值为0与1之间、0.10与0.90之间、0.20与0.80之间、0.30与0.70之间、0.50与0.99之间、0.60与0.99之间、0.70与0.99之间、0.80与0.99之间或0.90与0.99之间的值。

[0357] 作为另一示例,在一些实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或者某个其他范围“A”至“B”,其中A和B为两个不同的数字)的归一化的值,其中较接近于“1”的值(例如,0.89、0.90、0.91、0.92等)指示扰动特征(以及扰动特征所表示的化学化合物)与目的生理状况之间没有关联。较接近于“0”的值(例如,0.01、0.02、0.03、0.04等)指示扰动特征(以及扰动特征所表示的化学化合物)与目的生理状况之间有关联。在此类情况下,在“0”与“1”之间(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)选择第一阈值,并且当激活评分低于第一阈值时,扰动特征(以及其所表示的化学结构)被认为与目的生理状况相关联,而当激活评分高于第一阈值时,扰动特征(以及其所表示的化学结构)被认为与目的生理状况不关联。在一些此类实施例中,激活评分被表示为“0”与“1”之间的连续数值范围(或某个其他范围“A”至“B”,其中A和B为两个不同的数字)内的归一化的值,并且第一阈值为0与1之间、0.10与0.90之间、0.20与0.80之间、0.30与0.70之间、0.50与0.99之间、0.60与0.99之间、0.70与0.99之间、0.80与0.99之间或0.90与0.99之间的值。

[0358] 在一些实施例中,第一阈值标准为以下要求:第一扰动特征具有阈值激活评分。

[0359] 在一些实施例中,第一阈值标准为以下要求:第一扰动特征在扰动特征的集合中至少具有阈值排名,其中扰动特征的集合基于扰动特征的集合中的每个扰动特征与参考特征(例如,单细胞转变特征)的比较来排名。适合用于将化学化合物与生理状况相关联的将扰动特征与参考特征(例如,单细胞转变特征)进行比较的方法进一步详细描述于下面标题

为“针对扰动特征的数值激活评分”的章节中。

[0360] 在一些实施例中,关连要求,扰动特征的集合中的每个扰动特征的相应的计算出的激活评分满足阈值标准。在一些实施例中,关连要求,跨扰动特征的集合中的每个扰动特征的相应的计算出的激活评分的集中趋势测度满足阈值标准。在一些实施例中,集中趋势测度为扰动特征的集合中的每个扰动特征的每个相应的计算出的激活评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0361] 在一些实施例中,扰动特征的集合介于两个与100个扰动特征之间,并且关连要求,扰动特征的集合中的每个扰动特征的相应的计算出的激活评分满足阈值标准。在一些实施例中,扰动特征的集合介于两个与100个扰动特征之间,并且关连要求,跨扰动特征的集合中的每个扰动特征的相应的计算出的激活评分的集中趋势测度满足阈值标准。在一些实施例中,集中趋势测度为扰动特征的集合中的每个扰动特征的每个相应的计算出的激活评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0362] 在一些实施例中,扰动特征的集合为多个扰动特征,多个扰动特征的包括第一扰动特征的第一子集与目的生理状况相关联,多个扰动特征的第二子集与目的生理状况不关联,并且当针对第一扰动特征的相应的计算出的激活评分满足第一阈值标准且针对多个扰动特征的第二子集中的扰动特征的相应的计算出的激活评分满足第二阈值标准时,测试化学化合物与目的生理状况关连。

[0363] 在一些实施例中,第二阈值标准为以下要求:针对多个扰动特征的第二子集中的扰动特征的相应的计算出的激活评分具有阈值激活评分。

[0364] 在一些实施例中,第二阈值标准为以下要求:针对多个扰动特征的第二子集中的扰动特征的相应的计算出的激活评分在扰动特征的集合中至少具有阈值排名,其中扰动特征的集合基于扰动特征的集合中的每个扰动特征与参考特征(例如,单细胞转变特征)的比较来排名。

[0365] 在一些实施例中,关连要求,扰动特征的第二子集中的每个扰动特征的相应的计算出的激活评分满足第二阈值标准。在一些实施例中,关连要求,跨扰动特征的第二子集中的每个扰动特征的相应的计算出的激活评分的集中趋势测度满足第二阈值标准。在一些实施例中,集中趋势测度为扰动特征的集合中的每个扰动特征的每个相应的计算出的激活评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0366] III.将化学化合物与目的生理状况相关联的方法

[0367] 模型训练。

[0368] 本公开的另一方面提供了一种将化学化合物与目的生理状况相关联的方法800。在一些实施例中,目的生理状况为疾病。

[0369] 参考框802,该方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。设想了如上面标题为“生理状况”和“化合物”的章节中所公开的生理状况、化合物、指纹和/或获得指纹的方法的任何合适的实施例,包括其任何替换、修改、添加、删除和/或组合,如本领域技术人员显而易见的。

[0370] 举例来说,在一些实施例中,多种化合物介于10种与1×10^{6}种化合物之间。在一些实施例中,多种化合物介于100种与100,000种化合物之间。在一些实施例中,多种化合物介于1000种与100,000种化合物之间。

[0371] 在一些实施例中,多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。在一些实施例中,每个相应的指纹是使用SMILES变换器、ECFP4、RNNS2S或GraphConv从化学结构生成的。

[0372] 参考框804,该方法包括以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对多种化合物中的每种化合物的相应的数值激活评分,其中细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。设想了如上面标题为"细胞组分和细胞组分模块"和下面标题为"识别细胞组分模块"的章节中所公开的细胞组分、细胞组分模块和/或识别细胞组分模块的方法的任何合适的实施例,包括其任何替换、修改、添加、删除和/或组合,如对本领域技术人员显而易见的。

[0373] 举例来说,在一些实施例中,细胞组分模块的集合为单个细胞组分模块。在一些实施例中,细胞组分模块的集合为多个细胞组分模块。在一些实施例中,细胞组分模块的集合介于二百个与五百个细胞组分模块之间。在一些实施例中,细胞组分模块的集合由单个细胞组分模块组成。在一些实施例中,细胞组分模块的集合包括五个或更多个细胞组分模块。在一些实施例中,细胞组分模块的集合包括十个或更多个细胞组分模块。在一些实施例中,细胞组分模块的集合包括100个或更多个细胞组分模块。在一些实施例中,细胞组分模块的集合为多个细胞组分模块,多个细胞组分模块的第一子集与目的生理状况相关联,并且多个细胞组分模块的第二子集与目的生理状况不关联。

[0374] 在一些实施例中,如图2A至2B中的示例工作流程所展示,该方法进一步包括通过包括以下的过程来识别多个细胞组分模块中的细胞组分模块:以电子形式获得一个或多个第一数据集,该一个或多个第一数据集包括或共同包括:对于第一多个细胞(其中第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态)中的每个相应的细胞,对于多种细胞组分(其中多种细胞组分包括10种或更多种细胞组分)中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度。因此,该方法获取或形成多个向量,多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素,对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。使用多个向量以识别多个候选细胞组分模块中的每个候选细胞组分模块。多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集,其中多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中,并且其中多个细胞组分模块包括多于十个细胞组分模块。

[0375] 一个或多个第二数据集是以电子形式获得的,该一个或多个第二数据集包括或共同包括:对于第二多个细胞(其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量)中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度。因此,获得由(i)第二多个细胞和(ii)多种细胞组分或其表示来确定维度的细胞组分计数数据结构。激活数据结构是通过以下来形成的:使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示,其中激活数据结构包括:对于多个细胞组分模块中的每个细胞组分模块,

对于第二多个细胞中的每个细胞,相应的激活权重。

[0376] 候选细胞组分模型是使用以下两者之间的差异来训练的:(i)在将激活数据结构输入到候选模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测,以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在,其中该训练响应于差异而调整与候选细胞组分模型相关联的多个协变量权重,并且其中多个协变量权重包括:对于多个细胞组分模块中的每个相应的细胞组分模块,对于每个相应的协变量,对应的权重,该对应的权重指示相应的协变量是否跨激活数据结构与相应的细胞组分模块相关。在训练候选细胞组分模型时,使用多个协变量权重来识别多个候选细胞组分模块中的细胞组分模块(例如,与目的生理状况相关联的细胞组分模块)。

[0377] 在一些实施例中,目的生理状况为疾病,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态所记载。在一些实施例中,多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件下暴露于化合物。在一些实施例中,暴露条件为暴露的持续时间、化合物的浓度或暴露的持续时间与化合物的浓度的组合。

[0378] 在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq及其任何组合。在一些实施例中,多种细胞组分由介于100种与8,000种之间的细胞组分组成。

[0379] 在一些实施例中,使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。在一些实施例中,相关模型包括图聚类。在一些实施例中,图聚类方法为基于皮尔逊相关的距离度量上的莱顿聚类,或者为鲁汶聚类。

[0380] 在一些实施例中,多个细胞组分模块由介于10个与2000个之间的细胞组分模块组成。在一些实施例中,多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

[0381] 在一些实施例中,多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。

[0382] 在一些实施例中,该训练该候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0383] 参考框806,该方法进一步包括训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分,以及(ii)细

胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分。

[0384] 在一些实施例中,一个或多个计算出的激活评分中的激活评分为针对与相应的化合物相对应的相应的细胞组分模块的相应的激活权重。举例来说,在一些实施例中,激活评分为如在图2A至2B中所描述获得的并且在图5中的激活数据结构中展示的激活权重,其中激活评分指示相应的(例如,第一)细胞组分模块的与用相应的化合物进行处理相关和/或响应于用相应的化合物进行处理的激活(例如,诱导和/或差异表达)。

[0385] 设想了模型的任何合适的实施例,诸如上面标题为"模型架构"的章节中公开的那些实施例,以及其任何替换、修改、添加、删除和/或组合,如对本领域技术人员显而易见的。举例来说,在一些实施例中,经训练的模型包括神经网络。在一些实施例中,神经网络为具有ReLU激活的全连接神经网络。在一些实施例中,神经网络为消息传递神经网络。在一些实施例中,经训练的模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0386] 在一些实施例中,经训练的模型为多个成分模型的集成模型,并且相应的计算出的激活评分为多个成分模型中的每个成分模型的输出的集中趋势测度。在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。在一些实施例中,多个成分模型包括多个神经网络。在一些实施例中,多个神经网络中的第一神经网络为具有ReLU激活的全连接神经网络,并且多个神经网络中的第二神经网络为消息传递神经网络。

[0387] 参考框808,该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0388] 在一些实施例中,对模型的输入包括多个激活评分,对于多种化合物中的每种化合物来说,每个相应的激活评分对应于多个细胞组分模块中的相应的细胞组分模块。针对每种相应的化合物的与每个相应的细胞组分模块相对应的激活评分用作标记(例如,指示模块与化合物之间实际存在或不存在关联的数值激活评分),用于训练多任务模型以识别模块与化合物之间的关联(例如,权重和/或相关)。举例来说,如上面所描述,在一些实施例中,多个细胞组分模块的第一子集与目的生理状况相关联,并且多个细胞组分模块的第二子集与目的生理状况不关联。因此,在一些此类实施例中,可以使用多个细胞组分模块的第一子集作为标记将关联的实际存在包括在训练数据集中,并且可以使用多个细胞组分模块的第二子集作为标记将关联的实际不存在包括在训练数据集中。

[0389] 在一些实施例中,多种化合物包括至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少100、至少200、至少300、至少400、至少500、至少800、至少1000、至少2000、至少3000、至少4000、至少5000、至少8000、至少10,000、至少20,000、至少30,000、至少50,000、至少80,000、至少100,000、至少200,000、至少500,000、至少800,000、至少1百万或至少2百万种化合物,其中对于多种化合物中的每种化合物,对模型的输入包括针对多个细胞组分模块中的每个相应的细胞组分模块的相应的激活评分。

[0390] 在一些实施例中,多种化合物包括不超过1000万、不超过5百万、不超过1百万、不

超过500,000、不超过100,000、不超过50,000、不超过10,000、不超过8000、不超过5000、不超过2000、不超过1000、不超过800、不超过500、不超过200或不超过100种化合物,其中对于多种化合物中的每种化合物,对模型的输入包括针对多个细胞组分模块中的每个相应的细胞组分模块的相应的激活评分。在一些实施例中,多种化合物由10至500、100至10,000、5000至200,000或10,000至1百万种化合物组成,其中对于多种化合物中的每种化合物,对模型的输入包括针对多个细胞组分模块中的每个相应的细胞组分模块的相应的激活评分。

[0391] 在一些实施例中,如上面所描述,对于多种化合物中的每种化合物,多个数值激活评分中的相应的数值激活评分为针对多个细胞组分模块中的每个相应的细胞组分模块的激活权重(例如在图5中的激活数据结构中所展示)。

[0392] 如上面所描述,在一些实施例中,模型的输出包括一个或多个计算出的激活评分,其指示多种化合物中的相应的化合物(例如,测试化学化合物)是否与多个细胞组分模块中的相应的一个或多个细胞组分模块相关。

[0393] 一般来说,训练模型(例如,神经网络)包括通过反向传播(例如,梯度下降)来更新针对相应的模型的多个参数(例如,权重)。首先,进行正向传播,其中将输入数据(例如,对于多种化合物中的每种相应的化合物,针对多个模块中的每个相应的细胞组分模块的多个激活评分)接受到神经网络中,并基于所选择的激活函数和参数(例如,权重和/或超参数)的初始集来计算输出。在一些实施例中,针对未经训练或经部分训练的模型随机指派(例如,初始化)参数(例如,权重和/或超参数)。在一些实施例中,(例如,通过转移学习)从先前保存的多个参数或从经预训练的模型转移参数。

[0394] 然后通过以下来进行反向传递:计算针对每个相应的参数的与每个层中的每个相应的单元相对应的误差梯度,其中针对每个参数的误差是通过以下来确定的:基于网络输出(例如,相应的化合物与相应的细胞组分模块之间的关联的预测不存在或存在,作为计算出的激活评分)和输入数据(例如,预期值或真实标记;相应的化合物与相应的细胞组分模块之间的关联的实际不存在或存在,作为数值激活评分)来计算损失(例如,误差)。然后通过基于计算出的损失对值进行调整来更新参数(例如,权重),从而训练该模型。

[0395] 举例来说,在机器学习的一些一般实施例中,反向传播为训练具有包括多个权重(例如,嵌入)的隐藏层的网络的方法。首先使用任意选择的初始权重的集来生成未经训练的模型的输出(例如,关联的预测不存在或存在,作为计算出的激活评分)。然后通过以下来将输出与初始输入(例如,关联的实际不存在或存在,作为数值激活评分)进行比较:评估误差函数以计算误差(例如,使用损失函数)。然后更新权重,使得误差最小化(例如,根据损失函数)。在一些实施例中,使用多种反向传播算法和/或方法中的任一者来更新多个权重,如本领域技术人员将显而易见的。

[0396] 在一些实施例中,损失函数为均方误差、二次损失、平均绝对误差、平均偏差误差、折页、多类支持向量机和/或交叉熵。在一些实施例中,训练未经训练或经部分训练的模型包括根据梯度下降算法和/或最小化函数来计算误差。在一些实施例中,训练未经训练或经部分训练的模型包括使用多个损失函数来计算多个误差。在一些实施例中,多个损失函数中的每个损失函数接收相同或不同的加权因子。

[0397] 图6展示了根据本公开的一些实施例的训练模型的方法的示例。激活数据结构(顶部图画)提供对模型的输入,包括指示多个(K个)细胞组分模块中的每个相应的细胞组分模

块与多个(G个)细胞中的每个单元之间的关联的多个激活评分,其中每个细胞表示多种化合物中的相应的化合物。对于多个细胞组分模块中的每个相应的细胞组分模块(中部图画),在训练之前,针对由多个细胞(例如,W种化合物)共同表示的多种化合物中的每种相应的化合物,将对应的权重初始(例如,至随机权重)。因此,多个化合物权重构成化合物权重矩阵(中部图画)。使用以下两者之间的差异来进行对多个化合物权重的调整:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分(例如,预测的),以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分(例如,实际的)(底部图画)。在一些实施例中,例如使用参考图2A至2B和14A至14D在下面标题为“识别细胞组分模块”的章节中描述的用于识别细胞组分模块的方法来获得实际激活,其中多个协变量包括多种化合物。然后可以进行训练(例如,对化合物权重的调整),直至(例如,通过完成最小数量的调整和/或满足最小性能阈值)形成经训练的模型。

[0398] 在一些实施例中,使用误差函数以通过将一个或多个参数的值调整与计算出的损失成比例的量来更新模型(例如,神经网络)中的一个或多个参数(例如,权重),从而训练该模型。在一些实施例中,参数被调整的量通过学习率超参数来计量,该学习率超参数指示参数被更新的程度或剧烈度(例如,较小或较大的调整)。因此,在一些实施例中,该训练基于学习率来更新多个参数的全部或子集。在一些实施例中,学习率为差异学习率。

[0399] 在一些实施例中,训练模型(例如,神经网络)进一步采用对应的多个隐藏神经元中的每个隐藏神经元的对应参数的正则化。举例来说,在一些实施例中,通过向损失函数添加惩罚来进行正则化,其中惩罚与神经网络中的参数的值成比例。一般来说,正则化通过以下来减小模型的复杂度:向一个或多个参数添加惩罚以降低与这些参数相关联的相应的隐藏神经元的重要性。此类实践可以产生更通用的模型并减少数据的过度拟合。在一些实施例中,正则化包括L1或L2惩罚。举例来说,在一些优选实施例中,正则化包括对下参数和上参数的L2惩罚。在一些实施例中,正则化包括空间正则化(例如,基于先验和/或实验知识而确定)或丢弃正则化。在一些实施例中,正则化包括独立优化的惩罚。

[0400] 在一些实施例中,针对多个训练实例中的每个训练实例重复训练过程,该训练过程包括(例如,响应于预测标记与实际标记之间的差异)调整与模型相关联的多个化合物权重。

[0401] 在一些实施例中,多个训练实例包括至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少50、至少100、至少500、至少1000、至少2000、至少3000、至少4000、至少5000或至少7500个训练实例。在一些实施例中,多个训练实例包括不超过10,000、不超过5000、不超过1000、不超过500、不超过100或不超过50个训练实例。在一些实施例中,多个训练实例包括3至10、5至100、100至5000、或1000至10,000个训练实例。在一些实施例中,多个训练实例落入从不低于3个训练实例开始到不高于10,000个训练实例结束的另一范围内。

[0402] 在一些此类实施例中,该训练包括历经多个训练实例重复对模型的参数的调整(例如,经由反向传播),因此增加模型在指示相应的化合物是否与相应的细胞组分模块相关方面的准确性。

[0403] 在一些实施例中,该训练包括转移学习。转移学习进一步描述于例如定义章节中(参见,上面的“未经训练的模型”)。

[0404] 在一些实施例中,在对误差函数的第一评估之后,训练未经训练或经部分训练的模型形成经训练的模型。在一些此类实施例中,在基于对误差函数的第一评估来对一个或多个参数进行第一更新之后,形成经训练的模型。在一些实施例中,在以下之后形成经训练的模型:对误差函数的至少1、至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少20、至少30、至少40、至少50、至少100、至少500、至少1000、至少10,000、至少50,000、至少100,000、至少200,000、至少500,000或至少1百万次评估。在一些此类实施例中,在以下之后形成经训练的模型:基于对误差函数的至少1、至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少20、至少30、至少40、至少50、至少100、至少500、至少1000、至少10,000、至少50,000、至少100,000、至少200,000、至少500,000或至少1百万次评估,来对一个或多个参数进行至少1、至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少20、至少30、至少40、至少50、至少100、至少500、至少1000、至少10,000、至少50,000、至少100,000个、至少200,000、至少500,000或至少1百万次更新。

[0405] 在一些实施例中,当模型满足最低性能要求时,形成经训练的模型。举例来说,在一些实施例中,当在评估误差函数(例如,每种化合物与每个细胞组分模块之间的预测关联与实际关联之间的差异)之后针对经训练的模型计算的误差满足误差阈值时,形成经训练的模型。在一些实施例中,当误差小于百分之20、小于百分之18、小于百分之15、小于百分之10、小于百分之5或小于百分之3时,通过误差函数计算的误差满足误差阈值。

[0406] 在示例实施例中,该训练该模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0407] 在一些实施例中,该训练根据回归模型响应于与每种相应的化合物相关联的针对细胞组分模块的集合中的每个相应的细胞组分模块的每个差异而调整与未经训练的模型相关联的多个参数。在一些实施例中,回归模型优化与每种相应的化合物相关联的针对细胞组分模块的集合中的每个相应的细胞组分模块的每个差异的最小二乘误差。

[0408] 虽然对模型训练的前述讨论描述了获得和使用指示化合物与细胞组分模块之间的关联的激活评分,但实际上,设想指示化合物与任何其他目的生理状况或其任何细胞过程之间的关联的激活评分用于训练并使用模型以将化合物与生理状况相关联。举例来说,如将在以下章节中描述的,本公开的另一方面包括使用扰动特征来训练模型。具体地,在一些实施例中,模型使用针对扰动特征的数值激活评分作为训练标记来训练。然后,如方法700中所描述的,使用经训练的模型以响应于将化学结构指纹输入到模型中而获得一个或多个计算出的激活评分作为输出,其中一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征。

[0409] 获得针对扰动特征的数值激活评分。

[0410] 因此,本公开的另一方面提供了一种用于将化学化合物与目的生理状况相关联的方法900。在一些实施例中,目的生理状况为疾病。

[0411] 参考框902,该方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。设想了如上面标题为“生理状况”和“化合物”的章节中所公开的生理状况、化合物、指纹和/或获得指纹的方法的任何合适的实施例,包括其任何替换、修改、添加、删除和/或组合,如本领域技术人员显而易见的。

[0412] 举例来说,在一些实施例中,多种化合物介于10种与1×10^6种化合物之间。在一些实施例中,多种化合物介于100种与100,000种化合物之间。在一些实施例中,多种化合物介于1000种与100,000种化合物之间。在一些实施例中,多种化学化合物中的每种化学化合物为具有小于2000道尔顿的分子量的有机化合物。在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的每一个。在一些实施例中,多种化学化合物中的每种化学化合物满足里宾斯基五规则标准中的至少三个标准。在一些实施例中,每个相应的指纹是使用SMILES变换器、ECFP4、RNNS2S或GraphConv从化学结构生成的。

[0413] 参考框904,该方法包括以电子形式获得扰动特征的集合中的每个相应的扰动特征针对多种化合物中的每种对应的化合物的相应的数值激活评分。如对本领域技术人员将显而易见的,设想了如上面标题为“扰动特征”的章节中所公开的扰动特征的任何合适的实施例,包括其任何替换、修改、添加、删除和/或组合。

[0414] 举例来说,在一些实施例中,扰动特征的集合为单个扰动特征。在一些实施例中,扰动特征的集合为多个扰动特征。在一些实施例中,扰动特征的集合为介于二百个与五百个扰动特征之间。在一些实施例中,扰动特征的集合包括五个或更多个扰动特征。在一些实施例中,扰动特征的集合包括十个或更多个扰动特征。在一些实施例中,扰动特征的集合包括100个或更多个扰动特征。在一些实施例中,扰动特征的集合为多个扰动特征,多个扰动特征的第一子集与目的生理状况相关联,并且多个扰动特征的第二子集与目的生理状况不关联。

[0415] 参考框906,扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。

[0416] 在一些实施例中,扰动特征的集合中的相应的扰动特征的相应的数值激活评分通过包括以下的程序获得:以电子形式获取单细胞转变特征,该单细胞转变特征表示未改变的细胞状态与改变的细胞状态之间的差异细胞组分丰度的测度。改变的细胞状态通过从未改变的细胞状态到改变的细胞状态的细胞转变而出现,其中(i)未改变的细胞状态、(ii)改变的细胞状态以及(iii)从未改变的细胞状态到改变的细胞状态的转变中的至少一者与目的生理状况相关联。单细胞转变特征包括参考多种细胞组分的标识以及对应的第一显著性评分(针对多种参考细胞组分中的每种相应的细胞组分),该对应的第一显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及未改变的细胞状态与改变的细胞状态之间的细胞状态变化。比较单细胞转变特征和相应的扰动特征以便确定相应的扰动特征的相应的数值激活评分。

[0417] 在一些实施例中,比较单细胞转变特征和扰动特征以确定相应的扰动特征的相应的数值激活评分包括:针对有单细胞转变特征的参考多种细胞组分中的每种相应的细胞组分,将相应的细胞组分在单细胞转变特征中的第一显著性评分与对应的细胞组分在相应的扰动特征中的对应的显著性评分进行比较。

[0418] 在一些实施例中,比较单细胞转变特征和扰动特征以确定相应的扰动特征的相应

的数值激活评分包括:针对有单细胞转变特征的参考多种细胞组分中的每种相应的细胞组分,将多种参考细胞组分中的相应的细胞组分在单细胞转变特征中的第一显著性评分与多种细胞组分中的每种对应的细胞组分在相应的扰动特征中的对应的显著性评分进行比较。

[0419] 在一些实施例中,相应的扰动特征的激活评分为相应的扰动特征(相对于扰动特征的集合中的其他扰动特征)与单细胞转变特征的相关性的相对排名。在一些实施例中,相对排名通过Wilcoxon秩和检验、t检验、逻辑回归或广义线性模型来确定。在一些实施例中,相应的扰动特征的激活评分不基于排名。

[0420] 在一些实施例中,相应的扰动特征的激活评分是针对相应的扰动特征的、针对相应的多种细胞组分中的每种相应的细胞组分的相应的显著性评分的集中趋势测度。在一些实施例中,集中趋势测度为针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0421] 在一些实施例中,相应的扰动特征的激活评分为以下两者之间的差异:(i)针对相应的扰动特征的、针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分的集中趋势测度,以及(ii)针对单细胞转变特征的、针对多种参考细胞组分中的每种相应的细胞组分的对应的第一显著性评分的集中趋势测度。

[0422] 在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态或第二细胞状态相同。在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态和第二细胞状态两者均不同。

[0423] 在一些实施例中,该方法进一步包括修剪有单细胞转变特征的参考多种细胞组分以及有相应的扰动特征的相应的多种细胞组分以限制与转录因子的比较。在一些实施例中,该方法进一步包括修剪有单细胞转变特征的参考多种细胞组分以及有相应的扰动特征的相应的多种细胞组分以限制与另一细胞组分类型(例如,基因、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质和/或其组合)的比较。在一些实施例中,不修剪参考多种细胞组分和相应的多种细胞组分。

[0424] 在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由尚未暴露于多种化合物中的化合物的对照细胞表示。在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由跨已经暴露于多种化学化合物中的化学化合物(除了与相应的扰动特征相关联的化合物之外)的不相关的受扰动的细胞的平均数表示。

[0425] 如上面所讨论,在一些实施例中,作为非限制性示例,可以使用在以下中公开的任何方法来确定扰动特征的集合中的相应的扰动特征:2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691,其特此通过引用并入。

[0426] 相应的扰动特征包括相应的多种细胞组分的标识以及对应的显著性评分(针对相应的多种细胞组分中的每种相应的细胞组分),该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化。相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,而另一者为由细胞暴露于与相应的扰动特征相对应的化合物而引起的相应的受扰动的细胞状态。此外,如上面所讨论,相应的扰动特征包括数值激活评分。在一些实施例中,针对相应的扰动特征的数值激活评分为连续数值范围内的绝对值。在一些实施例中,针对相应的扰动特征的数值激活评分为相对排名,如下面更详细地讨论的。

[0427] 在一些实施例中,扰动特征的集合中的相应的扰动特征的相应的数值激活评分通过包括以下的程序来获得:以电子形式获取单细胞转变特征,该单细胞转变特征表示未改变的细胞状态与改变的细胞状态之间的差异细胞组分丰度的测度。这里,改变的细胞状态通过从未改变的细胞状态到改变的细胞状态的细胞转变而出现。此外,(i)未改变的细胞状态、(ii)改变的细胞状态以及(iii)从未改变的细胞状态到改变的细胞状态的转变中的至少一者与目的生理状况相关联。

[0428] 单细胞转变特征包括参考多种细胞组分的标识以及对应的第一显著性评分(针对多种参考细胞组分中的每种相应的细胞组分),该对应的第一显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及未改变的细胞状态与改变的细胞状态之间的细胞状态变化。在一些实施例中,单细胞转变特征使用在以下中公开的任何方法来确定:2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691,其特此通过引用并入。

[0429] 一旦获得,就将单细胞转变特征与相应的扰动特征进行比较,从而确定相应的扰动特征的相应的数值激活评分。在一些实施例中,可以使用在2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691中公开的用于将单细胞转变特征与相应的扰动特征进行比较以确定相应的扰动特征相对于多个扰动特征中的其他扰动特征的相对排名的任何方法,其中,例如,此类相对排名将随后被视为相应的扰动特征的相应的数值激活评分。

[0430] 在一些实施例中,比较单细胞转变特征和扰动特征以确定相应的扰动特征的相应的数值激活评分包括:针对有单细胞转变特征的参考多种细胞组分中的每种相应的细胞组分,将相应的细胞组分的第一显著性评分与对应的细胞组分在相应的扰动特征中的对应的显著性评分进行比较。在一些此类实施例中,相应的扰动特征的激活评分为相应的扰动特征(相对于扰动特征的集合中的其他扰动特征)与单细胞转变特征的相关性的相对排名。在一些此类实施例中,相对排名通过Wilcoxon秩和检验、t检验、逻辑回归或广义线性模型来确定。在一些实施例中,相应的扰动特征的激活评分不是相应的扰动特征的相关性的相对排名,而是独立于其他扰动特征与单细胞转变特征的排名而确定。

[0431] 在一些实施例中,相应的扰动特征的激活评分不基于排名。举例来说,在一些实施例中,相应的扰动特征的激活评分为多个显著性评分,包括针对相应的扰动特征的、针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分。

[0432] 在一些实施例中,相应的扰动特征的激活评分是针对相应的扰动特征的、针对相应的多种细胞组分中的每种相应的细胞组分的相应的显著性评分的集中趋势测度。在一些实施例中,集中趋势测度为针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分的算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数。

[0433] 在一些实施例中,相应的扰动特征的激活评分为以下两者之间的差异:(i)针对相应的扰动特征的、针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分的集中趋势测度,以及(ii)针对单细胞转变特征的、针对多种参考细胞组分中的每种相应的细胞组分的对应的第一显著性评分的集中趋势测度。

[0434] 在一个实施例中,为进行单细胞转变特征与相应的扰动特征之间的比较,有扰动特征的细胞组分被表示为矩阵。矩阵的每一行与单个扰动(例如,多种化合物中的单种化合

物)相关联。矩阵上的每一列与在相应的状态之间表现出差异丰度的细胞组分中的一种细胞组分相关联。矩阵中的每个条目包括针对特定扰动特征识别的细胞组分的显著性评分(例如,p值、t评分)。过滤该矩阵以仅包括处于单细胞转变特征的细胞组分。该过滤可以使用阈值p值、使用阈值数量的细胞成分等来完成。

[0435] 矩阵中的每个显著性评分均被替换为离散匹配评分。为将每个显著性评分替换为离散匹配评分,识别针对细胞转变的显著上调的细胞组分和针对细胞转变的显着下调的细胞组分。对于通过单细胞转变特征识别的显著上调的细胞组分中的每一种,如果对于该扰动(例如,化学组合物),该细胞组分针对扰动特征也是显著上调的,则矩阵中的针对该细胞组分/扰动组合的显著性评分被替代为离散匹配评分“1”。如果相对于单细胞转变特征,该细胞组分针对扰动特征是显著下调的,则矩阵中的针对该细胞组分/扰动组合的显著性评分被替代为离散匹配评分“-2”。如果细胞组分针对扰动特征没有显著上调或下调,则矩阵中的针对细胞组分/扰动组合的显著性评分被替代为离散匹配评分“0”。

[0436] 相反,对于在单细胞转变特征中识别的显著下调的细胞组分中的每一种,如果对于扰动,该细胞组分也是显着下调的,则矩阵中的针对该细胞组分/扰动组合的显著性评分被替代为离散匹配评分“-1”。如果对于扰动,该细胞组分是显著上调,则矩阵中的针对该细胞组分/扰动组合的显著性评分被替代为离散匹配评分“2”。如果对于扰动,细胞组分没有显著上调或下调,则矩阵中的针对该细胞组分/扰动组合的显著性评分被替代为离散匹配评分“0”。本领域技术人员将认识到,在一些实施例中,这些特定评分替代可以用其他数值来替换。此外,不是进行上调或下调,而是可以采用针对每种细胞组分使用阈值丰度值,其中随后在将上述类标记(例如,“-1”、“2”、“0”等)指派给矩阵的每个元素时考虑给定的细胞组分是高于还是低于阈值丰度值。

[0437] 结果为矩阵,其中行数由扰动的数量(多种化学组合物中的化学组合物的数量以及因此产生的多个扰动特征中的扰动特征的数量)给定,并且列数由来自单细胞转变的差异细胞组分给定,其中矩阵元素条目表示上面所描述的匹配评分。

[0438] 在将矩阵中的显著性评分替换为如上面所描述的离散匹配评分之后,对矩阵的每一行中的离散匹配评分求和以针对每一行生成总和匹配评分。然后,按照递减的总和匹配评分的次序对矩阵的行(每一行对应一扰动特征)进行排名。排名靠前的行与最可能关联于单细胞转变特征的所识别的细胞转变的扰动特征相关联。此外,各行中的每一行的排名可以用作与各列中的每一列相对应的针对扰动特征的激活评分。

[0439] 在一些实施例中,对于矩阵中的每一行的总和匹配评分,对错误细胞成分发现率进行估计,如以下中所讨论:2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691,其特此通过引用并入。

[0440] 在某些实施例中,可以存在扰动(例如,细胞暴露于特定化学组合物)的协变量。举例来说,化学组合物的协变量可以包括:化学组合物的特定剂量、测量暴露于化学组合物的细胞以量化细胞组分的时间和/或暴露于化学组合物的细胞的身份(例如,细胞系)。在一些实施例中,仅当扰动的协变量的阈值量也被预测影响特定细胞转变时,才预测扰动(例如,细胞暴露于特定化学组合物)影响特定细胞转变。换句话说,在一些实施例中,特定扰动特征的数值激活评分至少部分地通过以下来确定:有特定扰动特征的化学组合物的协变量是否也被预测影响与单细胞转变特征评分相关联的特定细胞转变。

[0441] 可以使用将相应的扰动特征与单细胞转变特征进行比较的替代方法来确定相应的扰动特征的数值激活评分。举例来说,可以使用web接口将细胞组分与数据库进行匹配(例如,诸如L1000CDS2.超快速LINCS L1000特性方向标签搜索引擎(An ultra-fast LINCS L1000 Characteristic Direction Signature Search Engine),位于万维网amp.pharm.mssm.edu/L1000CDS2/#/index)。

[0442] 在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态或第二细胞状态相同。在一些实施例中,单细胞转变特征的未改变的细胞状态与相应的扰动特征的第一细胞状态和第二细胞状态两者均不同。

[0443] 在一些实施例中,该方法进一步包括修剪有单细胞转变特征的参考多种细胞组分以及有相应的扰动特征的相应的多种细胞组分以限制与转录因子的比较。在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由尚未暴露于多种化合物中的化合物的对照细胞表示。

[0444] 在一些实施例中,多个扰动特征中的相应的扰动特征的受扰动的状态由跨已经暴露于多种化学化合物中的化学化合物(除了与相应的扰动特征相关联的化合物之外)的不相关的受扰动的细胞的平均数表示。

[0445] 参考框908,该方法进一步包括:训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的扰动特征的相应的计算出的激活评分,以及(ii)扰动特征的集合中的相应的扰动特征针对对应的化合物的相应的数值激活评分。

[0446] 设想了模型的任何合适的实施例,诸如上面标题为“模型架构”的章节中公开的那些实施例,以及其任何替换、修改、添加、删除和/或组合,如对本领域技术人员显而易见的。举例来说,在一些实施例中,经训练的模型包括神经网络。在一些实施例中,神经网络为具有ReLU激活的全连接神经网络。在一些实施例中,神经网络为消息传递神经网络。在一些实施例中,经训练的模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。

[0447] 在一些实施例中,经训练的模型为多个成分模型的集成模型,并且相应的计算出的激活评分为多个成分模型中的每个成分模型的输出的集中趋势测度。在一些实施例中,多个成分模型包括逻辑回归模型、神经网络模型、支持向量机模型、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型或线性回归模型。在一些实施例中,多个成分模型包括多个神经网络。在一些实施例中,多个神经网络中的第一神经网络为具有ReLU激活的全连接神经网络,并且多个神经网络中的第二神经网络为消息传递神经网络。

[0448] 参考框910,该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0449] 设想了用于训练未经训练或经部分训练的模型的任何合适的方法和实施例,诸如上面标题为“模型训练”的章节中公开的那些,包括其任何替换、修改、添加、删除和/或组

合,如对本领域技术人员显而易见的。

[0450] 对于一些实施例,对模型的输入包括多个激活评分,对于多种化合物中的每种化合物,每个相应的激活评分对应扰动特征的集合中的相应的扰动特征。针对每种相应的化合物的与每个相应的扰动特征相对应的激活评分用作标记(例如,指示扰动特征与化合物之间实际存在或不存在关联的数值激活评分),用于训练多任务模型以识别扰动特征与化合物之间的关联(例如,权重和/或相关)。举例来说,如上面所描述,在一些实施例中,多个扰动特征的第一子集与目的生理状况相关联,并且多个扰动特征的第二子集与目的生理状况不关联。因此,在一些此类实施例中,可以使用多个扰动特征的第一子集作为标记将关联的实际存在包括在训练数据集中,并且可以使用多个扰动特征的第二子集作为标记将关联的实际不存在包括在训练数据集中。

[0451] 在一些实施例中,该训练根据回归模型响应于与每种对应的化合物相关联的针对扰动特征的集合中的每个相应的扰动特征的每个差异而调整与未经训练的模型相关联的多个参数。在一些实施例中,回归模型优化与每种对应的化合物相关联的针对扰动特征的集合中的每个相应的扰动特征的每个差异的最小二乘误差。

[0452] 在一些实施例中,模型被训练和/或用于基于针对细胞组分模块的激活评分、扰动特征或两者来将化合物与目的生理状况相关联。在一些实施例中,模型被训练和/或用于基于针对多个结构域(例如,标记类型,诸如模块和/或扰动特征)和/或数据类型(例如,分析物和/或细胞组分,诸如基因表达图谱、代谢组学、蛋白质组学、表观遗传学等)的激活评分来将化合物与目的生理状况相关联。在一些实施例中,模型被训练和/或用于基于针对任何一种或多种目的生理状况(例如,化合物的毒性、疾病状态的消退等)的激活评分来将化合物与目的生理状况相关联。在一些实施例中,跨多个系统来训练该模型,其中系统是指本文所公开的任何一种或多种生理状况、任何一个或多个结构域和/或任何一种或多种数据类型,或对本领域技术人员将显而易见的任何替换、修改、添加、删除和/或组合。举例来说,在一些实施例中,模型被联合训练以共同确定测试化学化合物、具有毒性特性的基因模块的激活和指示疾病消退的扰动特征之间的关联。

[0453] 附加实施例。

[0454] 本公开的另一方面提供了一种计算机系统,其包括一个或多个处理器以及存储器,该存储器存储用于进行用于将测试化学化合物与目的生理状况相关联的方法的指令。该方法包括获得测试化学化合物的化学结构的指纹并将指纹输入到模型中,其中该模型包括100个或更多个参数,该模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分,一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块,细胞组分模块的集合中的每个相应的细胞组分模块包括多个细胞组分模块的独立子集,并且细胞组分模块的集合中的第一细胞组分模块与目的生理状况相关联。该方法进一步包括当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连。

[0455] 本公开的另一方面提供了一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将测试化学化合物与目的生理状况相关联,该计算机包括一个或多个处理器以及存储器,该一个或多个计算机程序共同编码进行方法的计算机可执行指令。该方法包括获得测试化学化合物的化学结构的指纹并将指纹输入到模型中,其

中该模型包括100个或更多个参数,该模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分,一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块,细胞组分模块的集合中的每个相应的细胞组分模块包括多个细胞组分模块的独立子集,并且细胞组分模块的集合中的第一细胞组分模块与目的生理状况相关联。该方法进一步包括当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连。

[0456] 本公开的又一方面提供了一种计算机系统,其包括一个或多个处理器以及存储器,该存储器存储用于进行用于将测试化学化合物与目的生理状况相关联的方法的指令。该方法包括获得测试化学化合物的化学结构的指纹并将该指纹输入到模型中,其中该模型包括100个或更多个参数。模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分。一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征。扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。该方法进一步包括当针对扰动特征的集合中的第一扰动特征的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连。

[0457] 本公开的另一方面提供了一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将测试化学化合物与目的生理状况相关联,该计算机包括一个或多个处理器以及存储器,该一个或多个计算机程序共同编码进行方法的计算机可执行指令。该方法包括获得测试化学化合物的化学结构的指纹并将该指纹输入到模型中,其中该模型包括100个或更多个参数。模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分。一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征。扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。该方法进一步包括当针对扰动特征的集合中的第一扰动特征的相应的计算出的激活评分满足第一阈值标准时,将化学化合物与目的生理状况关连。

[0458] 本公开的又一方面提供了一种计算机系统,其包括一个或多个处理器以及存储器,该存储器存储用于进行用于将化学化合物与目的生理状况相关联的方法的指令。该方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。该方法包括以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对多种化合物中的每种化合物的相应的数值激活评分,其中细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。该方法进一步包括训练未经训练的模型,对

于多种化合物中的每种相应的化合物的每个相应的化学结构,对于细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分,以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分。该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0459] 本公开的另一方面提供了一种存储由计算机可执行的一个或多个计算机程序的非暂时性计算机可读介质,其用于将化学化合物与目的生理状况相关联,该计算机包括一个或多个处理器以及存储器,该一个或多个计算机程序共同编码进行方法的计算机可执行指令。该方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。该方法进一步包括以电子形式获得细胞组分模块的集合中的每个细胞组分模块针对多种化合物中的每种化合物的相应的数值激活评分,其中细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。该方法进一步包括训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分,以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分。该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0460] 本公开的又一方面提供了一种计算机系统,其包括一个或多个处理器以及存储器,该存储器存储用于将化学化合物与目的生理状况相关联的指令,方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。该方法进一步包括以电子形式获得扰动特征的集合中的每个相应的扰动特征针对多种化合物中的每种对应的化合物的相应的数值激活评分。扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。该方法包括训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的扰动特征的相应的计算出的激活评分,以及(ii)扰动特征的集合中的相应的扰动特征针对对应的化合物的相应的数值激活评分。该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0461] 本公开的另一方面提供了一种存储由计算机可执行的一个或多个计算机程序的

非暂时性计算机可读介质,其用于将化学化合物与目的生理状况相关联,该计算机包括一个或多个处理器以及存储器,该一个或多个计算机程序共同编码进行方法的计算机可执行指令,该方法包括以电子形式获得多种化合物中的每种化合物的化学结构的相应的指纹,由此获得多个指纹。该方法进一步包括以电子形式获得扰动特征的集合中的每个相应的扰动特征针对多种化合物中的每种对应的化合物的相应的数值激活评分。扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分,该对应的显著性评分量化以下两者之间的关联:相应的细胞组分的丰度的变化,以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化,其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态,并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。该方法进一步包括训练未经训练的模型,对于多种化合物中的每种相应的化合物的每个相应的化学结构,对于扰动特征的集合中的每个相应的扰动特征,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的扰动特征的相应的计算出的激活评分,以及(ii)扰动特征的集合中的相应的扰动特征针对对应的化合物的相应的数值激活评分。该训练响应于差异而调整与未经训练的模型相关联的多个参数(其中该多个参数包括100个或更多个参数),由此获得将化学化合物与目的生理状况相关联的经训练的模型。

[0462] 本公开的又一方面提供了一种计算机系统,其具有一个或多个处理器和存储器,该存储器存储用于供一个或多个处理器执行的一个或多个程序,该一个或多个程序包括用于进行本文所公开的方法和/或实施例中的任一者的指令。在一些实施例中,当前所公开的方法和/或实施例中的任一者是在具有一个或多个处理器以及存储器的计算机系统处进行的,该存储器存储一个或多个程序以供一个或多个处理器执行。

[0463] 本公开的另一方面提供了一种非暂时性计算机可读存储介质,其存储配置成用于供计算机执行的一个或多个程序,该一个或多个程序包括用于执行本文所公开的方法中的任一种方法的指令。

[0464] IV.识别细胞组分模块

[0465] 在一些实施例中,识别与目的生理状况相关联的细胞组分模块132。这里结合图2和14来讨论此类方法。特别地,参考图14A的框1500,在一些实施例中,该方法进一步包括识别与目的生理状况相关联的第一细胞组分模块132。

[0466] 参考图2A至2B提供了根据本公开的一些实施例的用于将细胞组分与目的生理状况相关联的方法200的示例工作流程。

[0467] 参考图2A的框202和图14A的框1502,该方法包括以电子形式获得一个或多个第一数据集。参考图14B的框1504,一个或多个第一数据集包括或共同包括:对于第一多个细胞中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度。以此方式获得多个向量。

[0468] 在一些实施例中,目的生理状况为疾病,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态所记载。

[0469] 在一些实施例中,图3A的框300的目的生理状况为与疾病相关联的畸变细胞过程,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态

所记载。

[0470] 在一些实施例中,图3A的框300的目的生理状况为与疾病相关联的畸变细胞过程,并且第一多个细胞包括表示疾病状态的细胞和表示健康的或对照状态的细胞,如由多种经注释的细胞状态所记载。

[0471] 在一些实施例中,图3A的框300的目的生理状况为与多种疾病相关联的畸变细胞过程,并且第一多个细胞包括多个细胞子集,每个相应的细胞子集表示多种疾病中的相应的疾病,如由多种经注释的细胞状态所记载。

[0472] 参考图14B的框1506,在一些实施例中,第一多个细胞包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、30、40、50、60、70、80、100、200或1000个或更多个细胞并且共同表示多种(例如,2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、30、40、50、60、70、80、100、200或1000种)经注释的细胞状态。

[0473] 参考图14B的框1508,在一些实施例中,多种细胞组分包括2、3、4、5、6、7、8、9、10、15、20、25、30、35、50、100、500、1000、5000、10,000或更多种细胞组分。在一些实施例中,多种细胞组分由介于2种与10,000种之间的或细胞组分组成。在一些实施例中,多种细胞组分由介于100种与10,000种之间的或细胞组分组成。

[0474] 参考图2A的框204,该方法包括获取或形成多个向量。参考图14A的框1510,多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素。参考图14A的框1512,对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。

[0475] 参考框206,使用多个向量以识别多个候选细胞组分模块中的每个候选细胞组分模块。多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集。多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中,其中多个细胞组分模块包括多于十个细胞组分模块。

[0476] 参考图14B的框1514,在一些实施例中,多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件(例如,暴露的持续时间、化合物的浓度,或暴露的持续时间与化合物的浓度的组合)下暴露于化合物。

[0477] 参考图14B的框1518,在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。

[0478] 参考图14B的框1520,在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。

[0479] 参考图14B的框1522,在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq或其任何组合。

[0480] 参考图14B的框1524,在一些实施例中,目的生理状况为疾病,并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞,如由多种经注释的细胞状态所记载。

[0481] 参考图14B的框1526,使用多个向量以识别多个候选细胞组分模块中的每个候选细胞组分模块,多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子

集。多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中,并且其中多个细胞组分模块包括多于十个细胞组分模块。

[0482] 参考图14C的框1528,在一些实施例中,使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。在一些实施例中,相关模型包括图聚类算法(例如,图聚类方法为基于皮尔逊相关的距离度量上的莱顿聚类,图聚类方法为鲁汶聚类等等)。

[0483] 参考图14C的框1532,在一些实施例中,多个细胞组分模块由介于10个与2000个之间、介于100个与10000个之间、介于20个与5000个之间、介于2个与15,000个之间、介于80个与5000个之间、介于100个与500个之间的细胞组分模块组成。在一些实施例中,多个细胞组分模块介于2个与500个细胞组分模块之间。

[0484] 参考图14C的框1534,在一些实施例中,多种细胞组分由介于10种与2000种之间、介于100种与10000种之间、介于20种与5000种之间、介于2种与15,000种之间、介于80种与5000种之间、介于100种与500种之间的细胞组分组成。在一些实施例中,多种细胞组分介于2种与500种细胞组分之间。

[0485] 参考图14C的框1536,在一些实施例中,多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

[0486] 参考图2A的框208和图14C的框1538,该方法包括以电子形式获得一个或多个第二数据集。一个或多个第二数据集包括或共同包括:对于第二多个细胞(其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量)中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度。因此,获得细胞组分计数数据结构,其中细胞组分计数数据结构由(i)第二多个细胞和(ii)多种细胞组分或其表示确定维度。

[0487] 参考图14C的框1540,在一些实施例中,多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。

[0488] 参考图2B的框210和图14D的框1542,通过以下来形成激活数据结构:使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示。激活数据结构包括:对于多个细胞组分模块中的每个细胞组分模块,对于第二多个细胞中的每个细胞,相应的激活权重。

[0489] 参考图2B的框212和图14D的框1544,该方法进一步包括:使用以下两者之间的差异来训练候选细胞组分模型:(i)在将激活数据结构输入到候选模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测,以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在。该训练响应于差异而调整与候选细胞组分模型相关联的多个协变量权重,其中多个协变量权重包括:对于多个细胞组分模块中的每个相应的细胞组分模块,对于每个相应的协变量,对应的权重,该对应的权重指示相应的协变量是否跨激活数据结构与相应的细胞组分模块相关。

[0490] 参考图14D的框1546,该训练该候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0491] 因此,参考图2C的框214和图14D的框1548,在训练候选细胞组分模型时,使用多个

协变量权重来识别多个候选细胞组分模块中的第一细胞组分模块,其中多个候选细胞组分模块中的第一细胞组分模块与目的生理状况相关联。

[0492] 在一些实施例中,第一和/或第二多个细胞包括至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少100、至少200、至少300、至少400、至少500、至少1000、至少2000、至少3000、至少4000、至少5000、至少10,000、至少20,000、至少30,000、至少50,000、至少80,000、至少100,000、至少500,000或至少1百万个细胞。在一些实施例中,第一和/或第二多个细胞包括不超过5百万、不超过1百万、不超过500,000、不超过100,000、不超过50,000、不超过10,000、不超过5000、不超过1000、不超过500、不超过200、不超过100或不超过50个细胞。在一些实施例中,第第一和/或第二多个细胞包括5至100、10至50、20至500、200至10,000、1000至100,000、50,000至500,000或10,000至1百万个细胞。在一些实施例中,第一和/或第二多个细胞落入从不低于5个细胞开始到不高于5百万个细胞结束的另一范围内。

[0493] 在一些实施例中,第二多个细胞不包括第一多个细胞中所包括的任何细胞。在一些实施例中,第二多个细胞包括第一多个细胞中所包括的细胞的一些或全部。

[0494] 在一些实施例中,多种经注释的细胞状态包括细胞表型、细胞行为、疾病状态、基因突变、基因或基因产物的扰动(例如,敲低、沉默、过表达等)和/或暴露于化合物中的一者或多者。在一些实施例中,多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件下暴露于化合物。举例来说,细胞的暴露包括用一种或多种化合物对细胞进行任何处理。在一些实施例中,一种或多种化合物包括例如小分子、生物制品、治疗剂、蛋白质、与小分子组合的蛋白质、ADC、核酸(例如,siRNA、干扰RNA、过表达野生型和/或突变型shRNA的cDNA、过表达野生型和/或突变型向导RNA的cDNA(例如,Cas9系统或其他细胞成分编辑系统)等),和/或任何前述的任何组合。在一些实施例中,暴露条件为暴露的持续时间、化合物的浓度或暴露的持续时间与化合物的浓度的组合。在一些实施例中,化合物为本文所描述的任何实施例,诸如上面标题为“化合物”的章节中的。

[0495] 在一些实施例中,多种经注释的细胞状态包括对细胞批次、细胞供体、细胞类型、细胞系、疾病状态、时间点、重复和/或相关元数据的一个或多个指示。在一些实施例中,多种经注释的细胞状态包括实验数据(例如,流式细胞术读数、成像和显微镜注释、细胞组分数据等)。在一些实施例中,多种经注释的细胞状态包括一种或多种遗传标志物(例如,拷贝数变异、单核苷酸变体、多核苷酸多态性、插入、缺失、基因融合、微卫星不稳定状态、扩增和/或同种型)。在一些实施例中,多种经注释的细胞状态包括本文所公开的任何协变量和/或本文所公开的任何目的生理状况,诸如在上面标题为“生理状况”的章节中的。

[0496] 设想本文所公开的任何细胞组分和/或任何细胞组分模块,以及其任何实施例、替换、修改、添加、删除和/或组合均用于细胞组分模块的识别,如上面标题为“细胞组分和细胞组分模块”的章节中所描述。举例来说,在一些实施例中,多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。在一些实施例中,多种细胞组分由介于100种与8,000种之间的细胞组分组成。在一些实施例中,多个细胞组分模块由介于10个与2000个之间的细胞组分模块组成。在一些实施例中,多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。

[0497] 在一些实施例中,相应的细胞组分的对应的丰度包括上面所公开的任何细胞组分

的丰度。

[0498] 多种丰度计数技术(例如,细胞组分测量技术)中的任何一种可以用于获得每种相应的细胞组分在每个相应的细胞中的对应的丰度。举例来说,表1列出了根据本公开的一些实施例的用于单细胞细胞组分测量的非限制性技术。

[0499] 在一些实施例中,相应的细胞组分的对应的丰度是使用包括以下的一种或多种方法来测定的:经由荧光、化学发光的微阵列分析、电信号检测、聚合酶链式反应(PCR)、逆转录酶聚合酶链式反应(RT-PCR)、数字微滴PCR(ddPCR)、固态纳米孔检测、RNA开关激活、北方印迹和/或基因表达系列分析(SAGE)。在一些实施例中,相应的细胞组分在第一和/或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。

[0500] 在一些实施例中,第一和/或第二多个细胞中的相应的细胞中的基因表达可以通过以下来测量:对细胞进行测序并且然后对在测序期间识别的每种基因转录本的量进行计数。在一些实施例中,经测序和量化的基因转录本包括RNA,诸如mRNA。在一些实施例中,经测序和量化的基因转录本包括mRNA的下游产物,诸如蛋白质(例如,转录因子)。一般来说,如本文所用,术语“基因转录本”可以用于表示基因转录或翻译的任何下游产物,包括翻译后修饰,并且“基因表达”可以用于一般地指基因转录本的任何测度。

[0501] 在一些实施例中,相应的细胞组分的对应的丰度为RNA丰度(例如,基因表达),并且相应的细胞组分的丰度通过以下来确定:测量对应于相应的基因的一种或多种核酸分子的多核苷酸水平。相应的基因的转录本水平可以根据存在于第一和/或第二多个细胞中的相应的细胞中的mRNA或从其衍生的多核苷酸的量来确定。多核苷酸可以通过多种方法来检测和定量,多种方法包括但不限于微阵列分析、聚合酶链式反应(PCR)、逆转录酶聚合酶链式反应(RT-PCR)、北方印迹、基因表达系列分析(SAGE)、RNA开关、RNA指纹分析、连接酶链式反应、Qβ复制酶、等温扩增方法、链置换扩增、基于转录的扩增系统、核酸酶保护测定(Si核酸酶或核糖核酸酶保护测定)和/或固态纳米孔检测。参见,例如,Draghici,Data Analysis Tools for DNA Microarrays,Chapman and Hall/CRC,2003;Simon等人,Design and Analysis of DNA Microarray Investigations,Springer,2004;Real-Time PCR:Current Technology and Applications,Logan、Edwards和Saunders编著,Caister Academic Press,2009;Bustin A-Z of Quantitative PCR(IUL Biotechnology,No.5),International University Line,2004;Velculescu等人,(1995)Science 270:第484-487页;Matsumura等人,(2005)Cell.Microbiol.7:11-18;Serial Analysis of Gene Expression(SAGE):Methods and Protocols(Methods in Molecular Biology),Humana Press,2008;其中的每一者均特此通过引用以其整体并入本文。

[0502] 在一些实施例中,相应的细胞组分的对应的丰度获自来自第一和/或第二多个细胞中的相应的细胞的经表达的RNA或从其衍生的核酸(例如,cDNA或从cDNA衍生的合并RNA聚合酶启动子的经扩增的RNA),包括天然存在的核酸分子以及合成的核酸分子。因此,在一些实施例中,相应的细胞组分的对应丰度获自非限制性源,诸如总细胞RNA、聚(A)+信使RNA(mRNA)或其小部分、细胞质mRNA或从cDNA转录的RNA(例如,cRNA)。用于制备总RNA和聚(A)+RNA的方法是本领域公知的,并且一般性地描述于以下中:例如,Sambrook等人,Molecular Cloning:A Laboratory Manual(第3版,2001)。可以经以下从目的细胞提取RNA:使用硫氰

酸胍溶解接着CsCl离心(参见,例如,Chirgwin等人,1979,Biochemistry18:5294-5299)、基于硅胶的柱(例如,RNeasy(Qiagen,Valencia,Calif.)或StrataPrep(Stratagene,La Jolla,Calif.)),或使用苯酚和氯仿,如描述于Ausubel等人编著,1989,Current Protocols In Molecular Biology,第III卷,Green Publishing Associates,Inc.,John Wiley&Sons,Inc.,New York,第13.12.1-13.12.5页)。可以例如通过用寡dT纤维素进行选择,或者替代性地通过寡dT引发的总细胞RNA的逆转录来选择聚(A)+RNA。RNA可以通过本领域已知的方法,例如,通过用ZnCl2进行温育来片段化,以产生RNA的片段。

[0503] 在一些实施例中,相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过测序来确定。在一些实施例中,相应的细胞组分在第一和/或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定:单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq及其任何组合。

[0504] 细胞组分丰度测量技术可以基于待测量的期望细胞组分来选择。举例来说,可以使用scRNA-seq、scTag-seq和miRNA-seq来测量RNA表达。具体地,scRNA-seq测量RNA转录本的表达,scTag-seq允许检测稀有mRNA物种,并且miRNA-seq测量微RNA的表达。可以使用CyTOF/SCoP和E-MS/Abseq来测量细胞中的蛋白质表达。CITE-seq同时测量细胞中的基因表达和蛋白质表达两者,并且scATAC-seq测量细胞中的染色质构象。下面的表1提供了用于进行上面所描述的每种细胞组分丰度测量技术的示例方案。

[0505] 表1-示例测量方案

[0506]

技术	方案
RNA-seq	Olsen 等人，(2018)，“单细胞 RNA 测序简介(Introduction to Single-Cell RNA Sequencing)”，Current protocols in molecular biology 122(1)，第 57 页。
Tag-seq	Rozenberg 等人，(2016)，“通过样品多重检测和 PCR 重复检测进行数字基因表达分析：简单方案(Digital gene expression analysis with sample multiplexing and PCR duplicate detection: A straightforward protocol)”，BioTechniques, 61(1)，第 26 页。
ATAC-seq	Buenrostro 等人，(2015)，“ATAC-seq：在全基因组范围内测定染色可及性的方法(ATAC-seq: a method for assaying chromatic accessibility genome-wide)”，Current protcols in molecular biology, 109(1)，第 21 页。
miRNA-seq	Faridani 等人，(2016)，“小 RNA 转录组的单细胞测序(Single-cell sequencing of the small-RNA transcriptome)”，Nature biotechnology, 34(12)，第 1264 页。
CyTOF/SCoPE-MS/Abseq	Bandura 等人，(2009)，“质谱流式细胞术：基于电感耦合等离子体飞行时间质谱的实时单细胞多靶点免疫测定(Mass cytometry: technique for real time single cell multitarget immunoassay based on inductivitely coupled plasma time-of-flight mass spectrometry)”，Analystic chemistry, 81(16)，第 6813 页。 Budnik 等人，(2018)，“SCoPE-ME：单哺乳动物细胞的质谱量化细胞分化期间的蛋白质组异质性(SCoPE-ME: mass scpectrometry of single mammalian cells quantifies proteome heterogenity

[0507]

技术	方案
	during cell differentiation)”，Genome biology, 19(1)，第 161 页。 Shahi 等人，(2017)，“Abseq：利用 Droplep 微流控条形码进行超高通量输出单细胞蛋白质图谱分析(Abseq: Ultrahigh-throughoutput single cell protein profiling with droplep microfluidic barcoding)”，Scientific reports, 7，第 44447 页。
CITE-seq	Stoeckius 等人，(2017)，“单细胞中表位和转录组的同时测量(Simultaneous epitope and transcritome measurement in single cells)”，Nature Methods, 14(9)，第 856 页。

[0508] 在一些实施例中,在单个时间点测量多种细胞组分。在一些实施例中,在多个时间点测量多种细胞组分。举例来说,在一些实施例中,在整个细胞状态转变(例如,分化过程、对暴露于化合物的响应、发育过程等)中在多个时间点测量多种细胞组分。

[0509] 应当理解,这是说明性的而非限制性的,因为本公开涵盖使用对从细胞(例如,单细胞)获得的其他细胞组分的测量结果的类似方法。应当进一步理解,本公开涵盖使用直接从实践本公开中描述的方法的个人或组织进行的实验工作获得的测量结果的方法,以及使用间接例如从由其他人进行并通过任何手段或机制使其可用的实验工作的结果的报告获得的测量结果(包括第三方出版物中报告的数据、数据库、承包商进行的测定或可用于实践所公开的方法的合适输入数据的其他来源)的方法。

[0510] 在一些实施例中,对多种细胞组分在第一和/或第二多个细胞中的对应的丰度(例如,一个或多个第一数据集和/或一个或多个第二数据集)进行预处理。在一些实施例中,预处理包括过滤、归一化、映射(例如,到参考序列)、量化、缩放、反卷积、清理、降维、变换、统计分析和/或聚合中的一者或多者。

[0511] 举例来说,在一些实施例中,基于以下来过滤多种细胞组分:期望的质量,例如,核酸序列的大小和/或质量,或相应的细胞组分的最小和/或最大丰度值。在一些实施例中,过滤部分或其整体通过各种软件工具诸如Skewer来进行。参见,Jiang,H.等人,BMC Bioinformatics 15(182):1-12(2014)。在一些实施例中,使用例如以下针对质量控制来过滤多种细胞组分质:测序数据QC软件诸如AfterQC、Kraken、RNA-SeQC、FastQC,或另一类似软件程序。在一些实施例中,将多种细胞组分归一化,例如,以考虑下拉、扩增和/或测序偏差(例如,可映射性、GC偏差等)。参见,例如,Schwartz等人,PLoS ONE 6(1):e16685(2011)以及Benjamini和Speed,Nucleic Acids Research 40(10):e72(2012),两者的内容特此通过引用以其整体并入,用于所有目的。在一些实施例中,预处理从多种细胞组分中去除细胞组分的子集。在一些实施例中,预处理多种细胞组分的对应的丰度改善(例如,降低)高信噪比。

[0512] 在一些实施例中,预处理包括进行相应的细胞组分在相应的细胞中的对应的丰度与参考丰度的比较。在一些实施例中,参考丰度获自,例如,正常样品、匹配的样品、包括参考丰度值的参考数据集、参考细胞组分诸如管家基因,和/或参考标准品。在一些实施例中,使用以下来进行细胞组分丰度的该比较:任何差异表达检验,包括但不限于均值差异检验、Wilcoxon秩和检验(Mann Whitney U检验)、t检验、逻辑回归和广义线性模型。本领域技术人员将认识到,其他度量也可以用于细胞组分丰度的比较和/或归一化。

[0513] 因此,在一些实施例中,一个或多个第一数据集和/或一个或多个第二数据集中相应的细胞组分在相应的细胞中的对应的丰度包括多种形式中的任一种,包括但不限于原始丰度值、绝对丰度值(例如,转录本编号)、相对丰度值(例如,相对荧光单位、转录组分析和/或基因集表达分析(GSEA))、复合或聚合丰度值、经变换的丰度值(例如,经log2和/或log10变换的)、相对于参考(例如,正常样品、匹配的样品、参考数据集、管家基因和/或参考标准品)的变化(例如,倍数或对数变化)、标准化丰度值、集中趋势测度(例如,均值、中值、众数、加权均值、加权中值和/或加权众数)、离散测度(例如,方差、标准偏差和/或标准误差)、经调整的丰度值(例如,经归一化、缩放和/或误差校正的)、经降维的丰度值(例如,主成分向量和/或潜在成分)及/或其组合。用于使用降维技术来获得细胞组分丰度的方法是本领域已知的并且在下面进一步详述,该方法包括但不限于主成分分析、因子分析、线性判别分析、多维缩放、等距特征映射、局部线性嵌入、hessian特征映射、谱嵌入、t分布随机邻居嵌入及/或其对本领域技术人员显而易见的任何替换、添加、删除、修改和/或组合。参见,例如,Sumithra等人,2015,“A Review of Various Linear and Non Linear Dimensionality Reduction Techniques,”Int J Comp Sci and Inf Tech,6(3),第2354-2360页,其特此通过引用以其整体并入本文。

[0514] 在一些实施例中,使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。

[0515] 在一些实施例中,相关模型包括聚类方法(例如,聚类模型)。在一些实施例中,相关模型包括图聚类方法(例如,模型)和/或非图聚类方法。在一些实施例中,图聚类方法为基于皮尔逊相关的距离度量上的莱顿聚类。在一些实施例中,图聚类方法为鲁汶聚类。

[0516] 举例来说,在一些实现方式中,该方法包括应用基于相关的成本函数。优化基于相关的成本函数包括计算定义细胞组分(例如,基因)之间的邻近关系的最近邻居图,由通过存储每个细胞中每种细胞组分的丰度计数(例如,表达值)而形成的向量表示该细胞组分,并计算细胞组分之间的相关性。将彼此之间具有高相关性的细胞组分确定为最近邻居,并将其用于通过使用图聚类方法(例如,莱顿和/或鲁汶)对图进行聚类来形成细胞组分模块。

[0517] 可以使用多种聚类技术中的任一种,其示例包括但不限于分层聚类、k均值聚类和基于密度的聚类。在实施例中,使用基于分层密度的聚类(称为HDBSCAN,参见,例如,Campello等人,(2015).Hierarchical density estimates for data clustering,visualization,and outlier detection.ACM Trans Knowl Disc Data,10(1),5)。在另一实施例中,使用基于社区检测的聚类,诸如鲁汶聚类(参见,例如,Blondel等人,(2008).Fast unfolding of communities in large networks.J stat mech:theor exp,2008(10),P10008)。在又一实施例中,使用莱顿聚类。通过以下来进行莱顿算法:在社区之间移

动各个节点来确定分区、细化分区并基于经细化的分区来创建聚合网络。基于在该过程的较早步骤中确定的未经细化的分区来进一步分区该聚合网络,并通过移动每个聚合网络内的各个节点来细化新的分区。参见,例如,Traag等人,(2019),"From Louvain to Leiden: guaranteeing well-connected communities,"Sci Rep 9:5233,doi:10.1038/s41598-019-41695-z。在再一实施例中,使用扩散路径算法。

[0518] 一般来说,如鲁汶聚类和/或莱顿聚类等聚类采用硬分区技术,其中每个元素(例如,每种细胞组分)被唯一地指派给单个聚类而不重叠。然而,并且不受任一特定理论的束缚,细胞过程(例如,与目的生理状况相关联的)可以通过细胞内的细胞组分网络之间复杂且动态的相互作用来表征,其中举例来说,单个基因可以与在任意数量的相同或不同的过程和途径中类似地起作用的任意数量的其他基因组合而在细胞内的两个、三个、四个或更多个细胞过程中发挥作用。因此,与细胞内活动的复杂度相似,将细胞组分聚类到第一模块中不一定排除任何其他模块。因此,在一些实施例中,细胞组分模块的识别包括获得具有细胞组分的重叠子集的模块。

[0519] 作为采用使用基于相关的模型的硬划分技术的替代或补充,在一些实施例中,使用多个向量来识别多个细胞组分模块中的每个细胞组分模块包括字典学习模型,该字典学习模型产生作为多种降维组分的多种细胞组分的表示。在一些实施例中,字典学习模型为L0正则化的自动编码器。这些模型的优点为它们不强制模块与细胞组分之间的1:1对应,而是允许细胞组分同时出现在数个模块中。

[0520] 举例来说,在一些实现方式中,该方法包括应用备用自动编码器成本函数。在一些此类情况下,优化稀疏自动编码器成本函数包括:使用如在pytorch或tensorflow中实现的标准训练,通过其权重的L0正则化和重建损失来训练一层自动编码器。

[0521] 重叠分区算法的其他方法是可能的,包括但不限于模糊K均值、重叠K均值(OKM)、加权OKM(WOKM)、重叠分区聚类(OPC)和多聚类重叠K均值扩展(MCOKE),及/或其任何变型或组合。

[0522] 在一些实施例中,可以使用统计技术来将高维数据(例如,对于共同表示多种经注释的细胞状态的第一多个细胞中的每个细胞,多种细胞组分跨多个细胞组分模块的丰度)压缩到较低维空间,同时保留编码在一个或多个第一数据集中的潜在信息的形状。举例来说,如图4的顶部图画所展示,计数矩阵包括:对于第一多个细胞中的每个细胞,对于多种细胞组分中的每种细胞组分,对应的计数(例如,丰度)。计数矩阵可以变换为在图4的底部图画中展示的潜在表示,其中数据被降低到较低维空间,表示基于以下来跨第一多个细胞对细胞组分进行聚类:细胞组分在不同的经注释的细胞状态的条件(例如,细胞类型、暴露条件、疾病等)下的对应丰度的相似性。因此,经聚类的细胞组分被表示为细胞组分模块,其在潜在表示中编码跨多种细胞状态的行为相似性。

[0523] 再次参考在图4中展示的潜在表示,每个行列分组处的条目中的值是通过基于初始输入数据集的降维来确定的。举例来说,每个条目可以包括:对于由相应的列表示的每种相应的细胞组分,对以下的指示:在被包括在由相应的行(例如,权重$_{1\text{-}1}$、权重$_{1\text{-}2}$等)表示的相应的细胞组分模块中的多种细胞组分的子集中的隶属。具体地,在一些实施例中,每个条目为指示相应的细胞组分是否被包括在相应的模块中的权重。在一些实现方式中,权重为对隶属的二元指示(例如,在相应的模块中的存在或不存在分别由1或0指示)。在一些实现

方式中,权重被缩放以指示细胞组分对相应的模块的相对重要性(例如,隶属和/或相关的概率)。

[0524] 在一些实施例中,潜在表示中的相应的维度对应于相应的细胞组分的表示。细胞组分的表示可以例如由细胞组分的非线性表示产生,诸如在潜在表示矩阵中的相应的条目(例如,权重)对应于多种细胞组分的情况下。包括细胞组分的表示的其他实施例包括使用主成分分析获得的潜在表示,在该主成分分析中,每个主成分表示与多种细胞组分相对应的数据的方差和/或其他变换。

[0525] 在一些实施例中,降维技术产生数据的某种有损压缩。然而,所得潜在表示(例如,潜在表示118)的计算存储大小较小,并且因此需要较少的计算处理能力以结合其他下游技术(诸如模型训练)进行分析。因此,多个细胞组分模块布置在潜在表示中增加使用当前时代的计算装置的当前公开的方法的计算可行性。

[0526] 可以使用多种降维技术。在一些实施例中,降维为主成分(PCA)、随机投影、独立成分分析、特征选择、因子分析、Sammon映射、曲线成分分析、随机邻居嵌入(SNE)、Isomap、最大方差展开、局部线性嵌入、t-SNE、非负矩阵因子分解、核主成分分析、基于图的核主成分分析、线性判别分析(LDA)、广义判别分析、统一流形近似和投影(UMAP)、LargeVis、拉普拉斯特征图、扩散图、网络(例如,神经网络)技术和/或Fisher线性判别分析。参见,例如,Fodor,2002,“A survey of dimension reduction techniques,”Center for Applied Scientific Computing,Lawrence Livermore National,Technical Report UCRL-ID-148494;Cunningham,2007,“Dimension Reduction,”University College Dublin,Technical Report UCD-CSI-2007-7;Zahorian等人,2011,“Nonlinear Dimensionality Reduction Methods for Use with Automatic Speech Recognition,”Speech Technologies.doi:10.5772/16863.ISBN 978-953-307-996-7;以及Lakshmi等人,2016,“2016IEEE 6th International Conference on Advanced Computing(IACC),”第31-34页,doi:10.1109/IACC.2016.16,ISBN 978-1-4673-8286-1,其中的每一者均特此通过引用并入。因此,在一些实施例中,降维为主成分分析(PCA),并且每个相应的提取的降维成分包括通过PCA得出的相应的主成分。在此类实施例中,多个主成分中的主成分的数量可以限制为通过PCA计算的主成分的阈值数量。主成分的阈值数量可以为例如至少5、至少10、至少20、至少50、至少100、至少1000、至少1500或任何其他数量。在一些实施例中,通过PCA计算的每个主成分通过PCA被指派特征值,并且第一多个提取的特征的对应子集被限制为阈值数量的被指派最高特征值的主成分。对于多个细胞组分向量中的每个相应的细胞组分向量,将多个降维成分应用于相应的细胞组分向量以形成对应的降维向量,该对应的降维向量包括针对多个降维成分中的每个相应的降维成分的降维成分值。这从多个细胞组分向量形成对应的多个降维向量,由此形成布置在潜在表示中的多个细胞组分模块。

[0527] 在一些实施例中,该方法进一步包括使用布置在潜在表示中的多个细胞组分模块来进行流形学习。一般来说,流形学习用于通过确定数据集中的最大变异来描述高维数据的低维结构。示例包括但不限于力导向布局(Fruchterman,T.M.,&Reingold,E.M.(1991).Graph drawing by force-directed placement.Software:Practice and experience,21(11),1129-1164)(例如,Force Atlas 2)、t分布随机邻居嵌入(t-SNE)、局部线性嵌入(Roweis,S.T.,&Saul,L.K.(2000).Nonlinear dimensionality reduction by locally

linear embedding.Science,290(5500),2323-2326)、局部线性等距映射(ISOMAP,Tenenbaum,J.B.、De Silva,V.和Langford,J.C.(2000).A global geometric framework for nonlinear dimensionality reduction.Science,290(5500),第2319-2323页)、核PCA、基于图形的核PCA、基于亲和力的轨迹嵌入的热扩散势(PHATE)、广义判别分析(GDA)、统一流形近似和投影(UMAP),或核判别分析。特别地在预先已知关于每个细胞的特定细胞类型的一些信息的情况下,可以使用判别分析。力导向布局可用于各种特定实施例,因为它们能够识别新的、较低的维度,该维度对由底层细胞过程产生的底层数据的非线性方面进行编码。力导向布局使用基于物理的模型作为用于确定最佳表示数据的降低的维度的机制。作为示例,力导向布局使用某种物理模拟形式,其中,在该实施例中,一个或多个第一数据集中的每个细胞被指派“排斥”力并且存在全局“重力”,当对针对第一多个细胞计算该全局“重力”时,其识别数据的在这些竞争“力”下“扩散”在一起的扇区。力导向布局对数据的结构做出很少的假设,并且不强制采用去噪方法。

[0528] 流形学习进一步描述于例如以下中:Wang等人,2004,“Adaptive Manifold Learning,”Advances in Neural Information Processing Systems 17,其特此通过引用以其整体并入本文。

[0529] 在一些实施例中,多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。在一些实施例中,多个协变量包括对与第二多个细胞中的一个或多个细胞有关的时间点、重复和/或相关元数据的一个或多个指示。在一些实施例中,多个协变量包括实验数据(例如,流式细胞术读数、成像和显微镜注释、细胞组分数据等)。在一些实施例中,多个协变量包括具有第二多个细胞中的一个或多个细胞的特性的一种或多种遗传标志物(例如,拷贝数变异、单核苷酸变体、多核苷酸多态性、插入、缺失、基因融合、微卫星不稳定状态、扩增和/或同种型)。在一些实施例中,多个协变量包括针对第二多个细胞中的一个或多个细胞的以下中的一者或多者:细胞表型、细胞行为、疾病状态、基因突变、基因或基因产物的扰动(例如,敲低、沉默、过表达等)和/或暴露条件。

[0530] 举例来说,在一些实施例中,协变量为第二多个细胞中的细胞在暴露条件下暴露于化合物或对该暴露的响应。在一些实施例中,细胞的暴露包括用一种或多种化合物对细胞进行任何处理。在一些实施例中,一种或多种化合物包括例如小分子、生物制品、治疗剂、蛋白质、与小分子组合的蛋白质、ADC、核酸(例如,siRNA、干扰RNA、过表达野生型和/或突变型shRNA的cDNA、过表达野生型和/或突变型向导RNA的cDNA(例如,Cas9系统或其他细胞成分编辑系统)等),和/或任何前述的任何组合。在一些实施例中,暴露条件为暴露的持续时间、化合物的浓度或暴露的持续时间与化合物的浓度的组合。

[0531] 在一些实施例中,协变量为施加于一个或多个细胞的在一个或多个细胞中诱导细胞状态转变和/或扰动特征的化合物(例如,扰动原)。

[0532] 在一些实施例中,协变量为与多种细胞组分中的细胞组分或与第二多个细胞中的细胞相关联的知识项(例如,注释)。举例来说,在一些实施例中,协变量为全基因组关联研究(GWAS)注释、基因集富集测定(GSEA)注释、基因本体注释、功能和/或信号转导途径注释和/或细胞特征注释。在一些实施例中,协变量获自本领域已知的任何公共知识数据库,包括但不限于NIH基因表达汇编(GEO)、EBI ArrayExpress、NCBI、BLAST、EMBL-EBI、GenBank、Ensembl、KEGG途径数据库和/或任何疾病特定数据库。在一些实施例中,协变量获自提供扰

动(例如,小分子)诱导的基因表达特征的数据库,诸如基于网络的集成细胞特征库(LINCS)L1000数据集。参见,例如,Duan,2016,“L1000CDS^2:An ultra-fast LINCS L1000 Characteristic Direction Signature Search Engine,”Systems Biology and Applications 2,第16015条,其特此通过引用以其整体并入本文。

[0533] 在一些实施例中,多个协变量包括至少3、至少5、至少10、至少15、至少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少2000或至少3000个协变量。在一些实施例中,多个协变量包括不超过5000、不超过1000、不超过500、不超过200、不超过100、不超过50或不超过20个协变量。在一些实施例中,多个协变量包含3至10、10至50、20至500、200至1000或1000至5000个协变量。在一些实施例中,多个协变量落入从不低于3个协变量开始到不高于5000个协变量结束的另一范围内。

[0534] 在一些实施例中,多个协变量中的每个协变量为施加于一个或多个细胞的诱导细胞状态转变和/或扰动特征的化合物,并且多个协变量为多种化合物。在一些实施例中,多个协变量由多种化合物组成,如上面标题为“化合物”的章节中公开的。

[0535] 图5展示了通过以下形成的示例激活数据结构:使用多种细胞组分或其表示作为共同维度来组合细胞组分计数数据结构(例如,使用共同表示提供目的生理状况的信息的多个协变量的第二多个细胞获得的)和潜在表示。为实现这一点,在一些实施例中,将针对第二多个细胞的计数矩阵(例如,在结构上类似于图4所展示的针对第一多个细胞的计数矩阵)和潜在表示相乘,以便将潜在表示矩阵的权重乘以计数矩阵的归一化的计数。一般来说,可以将两个矩阵乘以公共维度(例如,第一矩阵的x轴和第二矩阵的y轴)。第一和第二矩阵与其公共维度的矩阵乘法产生辅助数据的第三矩阵,其可以作为第一矩阵和/或第二矩阵的替代或补充而应用于未经训练或经部分训练的模型。

[0536] 因此,在一些此类实施例中,计数矩阵具有维度n个_细胞x n个_基因,并且潜在表示具有维度n个_基因x n个_模块,其中n个_细胞为第二多个细胞中的细胞的数量,n个基因为细胞组分(例如,基因)的数量或其表示,并且n个_模块为多个细胞组分模块中的模块的数量。这将计数矩阵中的细胞组分的丰度映射到空间中,在该空间中,每个细胞(例如,对应于一个或多个目的协变量)通过其模块激活来表征,并且在该空间中,所得矩阵表示(例如,激活数据结构)具有维度为n个_细胞x n个_模块(例如,乘以公共维度n个_基因之后)。

[0537] 使用例如矩阵乘法的潜在表示和细胞组分计数数据结构的组合,以及呈矩阵形式的所得激活数据结构,共同展示在图5中。潜在表示(如图5的顶部图画所展示)具有维度Z x K,其中Z为细胞组分的数量或其表示,并且K为细胞组分模块的数量。细胞组分计数数据结构(如左下图画所展示)具有维度G x Z,其中G为第二多个细胞中的细胞的数量,如对于潜在表示来说,Z为细胞组分的数量或其表示。使用Z(细胞组分的数量或其表示)作为公共维度,通过矩阵乘法进行组合,产生具有维度G x K的所得激活数据结构。每个相应的行中针对每个相应的列的每个条目为激活权重,其指示每个相应的细胞组分模块在与相应的列相对应的第二多个细胞中的相应的细胞中的激活。因此,如图5所展示,与模块1相对应的计数包括与细胞1相对应的激活权重$_{1\text{-}1}$、与细胞G相对应的激活权重$_{1\text{-}G}$,依此类推。

[0538] 在一些实施例中,激活数据结构中的多个激活权重包括差异模块激活。在一些实施例中,差异模块激活(例如,激活数据结构中的第二多个细胞中的细胞之间的相应的模块

的差异激活权重)通过以下获得:使用函数(mu_1-mu_2)/(var_1+var_2)$^{-0.5}$来计算v评分,其中mu_i表示具有相应的条件i(例如,协变量i)的跨细胞的模块激活均值,并且var_i表示条件i下的模块激活方差。V评分可以描述为未按分母中的细胞数量归一化的t评分。

[0539] 在一些实施例中,激活数据结构中的第二多个细胞中的每个相应的细胞表示相应的协变量。在一些实施例中,激活数据结构中的第二多个细胞中的每个相应的细胞表示施加于一个或多个细胞的诱导细胞状态转变和/或扰动特征的相应的化合物。

[0540] 因此,在一些实施例中,激活数据结构指示相应的细胞组分模块对应于(例如,关联于和/或响应于)暴露于由第二多个细胞表示的多种化合物中的每种化合物的激活(例如,激活水平或程度)。举例来说,在其中第二多个细胞中的每个相应的细胞表示相应的扰动原(例如,一个或多个细胞所暴露于的化合物和/或诱导细胞状态转变和/或扰动特征的化合物)的一些实施例中,激活数据结构包括:对于多个细胞组分模块中的每个相应的细胞组分模块,相应的激活权重,其指示相应的细胞组分模块的与用相应的化合物进行处理相关和/或响应于用相应的化合物进行处理的激活(例如,诱导和/或差异表达)。

[0541] 在一些实施例中,候选细胞组分模型包括本文所公开的任何模型架构,如上面标题为“模型架构”的章节中描述的。

[0542] 在一些实施例中,候选细胞组分模型为自动编码器、稀疏自动编码器,和/或稀疏多读出、知识耦合的自动编码器。在一些实施例中,候选细胞组分模型是半监督模型。在一些实施例中,候选细胞组分模型为一层神经网络(例如,SoftMax和/或逻辑回归模型)。在一些实施例中,候选细胞组分模型为一维Huber异常值回归器模型。

[0543] 在一些实施例中,候选细胞组分模型为包括多个层的稀疏多读出、知识耦合的自动编码器,其中第一层用于获得潜在表示并且第二层用于获得细胞组分模块知识构建体(例如,协变量权重矩阵)。

[0544] 在一些实施例中,该训练该候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的,其中多个协变量中的每个协变量对应于多个成本函数中的成本函数,并且多个成本函数中的每个相应的成本函数具有公共的权重因子。

[0545] 在一些实施例中,训练候选细胞组分模型为对该模型进行训练以识别细胞组分模块的集合中的与目的生理状况相关联的第一细胞组分模块。本文进一步详细地描述了用于训练模型的方法。设想了本文所公开的任何方法和/或实施例用于训练候选细胞组分模型,如上面标题为“模型训练”的章节中描述的。

[0546] V.实例

[0547] 本文提供了用于将化合物与生理状况相关联的模型的示例性能测量和治疗应用。

[0548] 实例1.预测用于激活脂肪酸相关细胞过程的化学结构。

[0549] 在该实例中,首先定义了细胞组分模块。这是通过获得细胞的表达数据来完成的,其中细胞表示与目的生理状况相关联的不同状态。这一点跟随如最初提交的权利要求27。从细胞中的每一个测量细胞组分丰度值,并且将该数据用于对细胞组分进行聚类。其表达值跨由细胞表示的各种状态而彼此相关的那些细胞组分被分组为细胞组分模块。这产生数个细胞组分模块,其中的每一个均包括细胞组分样品的不同子集。在一些实施例中,虽然每个细胞组分模块具有细胞组分的不同子集,但是一个细胞组分模块中的蜂窝组分与另一细胞组分模块中的细胞组分之间可能存在重叠。

[0550] 此外,在该实例中,以第二训练集的形式获得附加训练数据。第二训练集还包括细胞组分的单细胞丰度数据。然而,在该第二训练集中,每个细胞已暴露于多种训练化学化合物中的不同化学化合物。在该训练集中,已知量为相应的不同化学化合物的指纹,以及暴露于此类化合物的细胞的所得细胞组分丰度数据。第二数据集的数据可以布置为计数矩阵502(如图5所展示),其中第一轴用于细胞组分身份,并且第二轴用于细胞身份。因此,计数矩阵502中的每个元素为给定细胞组分在给定细胞中的丰度。此外,计数矩阵502中的每个相应的列(其对应于特定细胞)用特定细胞所暴露于的特定化合物来标记。因此,计数矩阵502的每一列均用特定化合物(例如,训练化合物)来标记,而每个元素为对应的细胞组分(Y轴)针对对应的细胞(X轴)的计数。

[0551] 如图5所展示,来自第一数据集(潜在表示404)和第二数据集(计数矩阵502)的数据被组合以形成激活数据结构(例如,如图5所展示的激活数据结构504)。举例来说,实现这一点的一种方式是将细胞组分模块布置为潜在表示404中的行,使得第一轴表示细胞组分模块并且第二轴表示细胞组分中的每一种。这样,为产生激活数据结构504,经由矩阵乘法将潜在表示404和计数矩阵502乘以它们的公共轴(细胞组分数量),以得到激活数据结构504。激活数据结构504保留来自计数矩阵502的细胞身份轴和来自潜在表示504的细胞组分模块轴。可以针对不同的细胞类型形成不同的激活结构。也就是说,用于形成计数矩阵502的细胞可以表示特定目的疾病状态。因此,可以针对不同的疾病状态或其他目的表型形成不同的激活数据结构504。

[0552] 转向图6,在一些情况下,激活数据结构504的每一行(来自图5并且现在位于图6的顶部)用作用于不同模型601的训练数据。举例来说,考虑这样的情况,其中模型601包括行604-1的权重(权重$_{1\text{-}1}$至权重$_{1\text{-}W}$)以表示化合物1至W分别激活细胞组分模块1的程度。该模型601在激活数据结构504的行640的元素上进行训练,该元素提供训练化合物1、…、G中的每一者激活细胞组分模块1的程度。在此训练中,首先将细胞1所暴露于的化合物的指纹表示输入到模型601中。响应于该输入,模型601针对细胞组分模块1输出激活值,在图6的命名中称为Pred值$_1$。将该输出激活值与实际激活值进行比较,该实际激活值为激活数据结构504的Act$_{1\text{-}1}$。接下来,将细胞2所暴露于的化合物的指纹表示输入到模型601中。响应于此输入,模型输出激活值(Pred值$_2$)。将该输出激活值与化合物2的实际激活值进行比较,该实际激活值为激活数据结构504的Act$_{1\text{-}2}$。该过程进行至细胞G。将细胞G所暴露于的化合物的指纹表示输入到模型601中。响应于此,模型将输出激活值(Pred值$_G$)。将该输出激活值与细胞G的实际激活值进行比较,该实际激活值为激活数据结构504的Act$_{1\text{-}G}$。在此实例中,W和G具有相同的值。这样,得到化合物的训练集中的每种化合物的所得预测(Pred值),该化合物的训练集用于得出针对细胞组分模块1的如图5所绘示的激活数据结构。将上述计算出的(激活值的)预测与这些化合物中的每一种的上述实际激活值进行比较,并且使用预测激活值与实际激活值之间的差异以使用反向传播和相关模型细化技术来进一步训练模型601。

[0553] 因此,结果为一系列经训练的模型601,每个细胞组分模块一个模型。测试化合物的指纹可以输入到经训练的模型中的每一个中,并且每个相应的经训练的模型601输出预测激活值,其大小指示对应于相应的经训练的模型的细胞组分模块是否被测试化合物激活。鉴于已经描述了该过程的概述,结合本实例中使用的实验数据来描述步骤中的每一个。

[0554] 通过以下过程识别第一细胞组分模块(图1、图4 132-1)。以电子形式获得一个或

多个第一数据集。一个或多个第一数据集包括共同表示多种经注释的(例如,经标记的或已知的)细胞状态的第一多个细胞(例如,二十个或更多个细胞)的数据。第一数据集包括:对于第一多个细胞中的每个相应的细胞,对于多种细胞组分(例如,10种或更多种细胞组分)中的每种相应的细胞组分,相应的细胞组分在相应的细胞中对应的丰度。举例来说,针对每个细胞的转录数据。这样,获取或形成多个向量。多个向量中的每个相应的向量对应于多种组分中的相应的细胞组分,并且包括对应的多个元素。向量的对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。因此,在一些此类实施例中,获得针对多个细胞状态中的每个细胞状态的转录数据。

[0555] 举例来说,形成图4所展示的形式的计数矩阵402。就此示例来说,使用已知诱导前脂肪细胞的代谢活性过程的小分子扰动原。将前脂肪细胞系的等分试样暴露于扰动原24小时,并针对扰动条件下该细胞系的经暴露的等分试样获得scRNA-seq读数。还针对该细胞系的未暴露于扰动原的等分试样获得scRNA-seq读数,并且这些读数表示对照条件。这样,根据图14A的框1504,获得第一数据集,其包括:对于第一多个细胞中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度,由此获取或形成多个向量。也就是说,每种细胞组分(例如,基因)的在暴露于扰动原的细胞和未暴露于扰动原的细胞(对照细胞)两者中测量的表达值形成图4所展示的计数矩阵402的元素。如图4中所展示且图14A的框1510中所指出,计数矩阵402包括针对每种细胞组分的向量,并且因此存在多个向量。多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分,并且(ii)包括对应的多个元素。

[0556] 举例来说,对于细胞组分1(例如,基因1),计数1-1、…、计数1-N为基因1在细胞1至N中的表达的测量结果,其中N个细胞中,一些已暴露于扰动原而一些未暴露,这些计数构成针对细胞组分1的向量的元素。也就是说,根据图14A的框1512,针对细胞组分1的向量的对应的多个元素中的每个相应的元素具有对应的计数,该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。虽然该示例包括两种状态(暴露于扰动原或未暴露于扰动原),但原则上可以涵盖任何数量的状态,诸如不同的扰动原浓度、暴露时间等。

[0557] 根据图14A的框1514,在该实例1中存在两种经注释的状态:对照(不暴露于扰动原)和暴露于扰动原。也就是说,多种经注释的细胞状态中的一种经注释的细胞状态为第一多个细胞中的细胞在暴露条件(例如,暴露持续时间,这里为24小时)下暴露于化合物(这里为扰动原)。虽然该实例由两种状态(暴露于扰动原或未暴露于扰动原)组成,但原则上可以涵盖任何数量的状态,诸如不同的扰动原浓度、暴露时间等。

[0558] 通过过滤和归一化步骤对计数矩阵402进行预处理,产生含有具有高信噪比的数个基因的经预处理的计数矩阵。

[0559] 使用多个向量以识别多个候选细胞组分模块中的每个候选细胞组分模块。多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集。多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中,并且其中多个细胞组分模块包括多于十个细胞组分模块。

[0560] 在一些实施例中,每个候选细胞组分模块为候选转录指纹。

[0561] 在该示例中,使用计数矩阵402以识别细胞组分模块132。这是根据图14B的框1526完成的:使用多个向量(图4的计数矩阵402的每一行)以识别多个候选细胞组分模块中的每个候选细胞组分模块,多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集。

[0562] 这产生由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示,其中多个细胞组分模块包括多于十个细胞组分模块。该潜在表示的示例为图4的潜在表示404,其中,对于每个相应的候选细胞组分模块132,存在对哪些细胞组分在相应的候选细胞组分模块中的指示。

[0563] 根据图14C的框1528形成潜在表示404:使用多个向量(计数矩阵402的细胞组分向量)以识别(潜在表示404的)多个候选细胞组分模块中的每个候选细胞组分模块,方式为使用多个向量中的每个向量的每组对应的多个元素将相关模型应用到多个向量。特别地,优化基于相关的成本函数,这相当于计算定义细胞组分向量之间的邻近关系的最近邻居图,并计算计数矩阵402的细胞组分向量之间的相关性。跨多个细胞彼此具有高度相关性的细胞组分(此处为基因)最终成为最近邻居,并且通过使用莱顿或任何其他图聚类方法对图进行聚类而在潜在表示402内形成细胞组分模块。优化稀疏自动编码器成本函数相当于使用如在pytorch或tensorflow中实现的标准训练,通过其权重的L0正则化和重建损失来训练一层自动编码器)。在此示例中,这使得在训练期间108个细胞组分模块被学习。也就是说,图4的潜在表示404具有108个细胞组分模块132,每个细胞组分模块具有其表达数据在计数矩阵402中可用的细胞组分的独立子集。

[0564] 在108个细胞模块中,当对针对每种细胞成分的跨受扰动的样品和对照样品计算的t评分求平均时,称为“模块78”的细胞组分模块132显示出最强的激活。换句话说,使用计数矩阵数据中的表达数据以通过以下来验证细胞组分:对于潜在表示404中的每个相应的细胞组分模块,在已暴露于扰动原的细胞与未暴露于扰动原的细胞之间,对相应的细胞组分模块中的每种细胞组分的差异表达进行t评分。此外,模块78富集有在与脂肪酸和脂质相关联的生物过程中涉及的细胞组分。总之,模块78由28个基因组成,包括FABP3(代谢活性标志物)。

[0565] 除了细胞组分模块之外,还需要在细胞暴露于训练化合物时的基于细胞的细胞组分响应数据。

[0566] 因此,以电子形式获得一个或多个第二数据集。一个或多个第二数据集包括来自第二多个细胞的数据。第二多个细胞包括二十个或更多个细胞。第二多个单元共同表示提供目的生理状况的信息的多个协变量。举例来说,在一些情况下,多个协变量为训练化合物。然后,对于第二多个细胞中的每个细胞,对于多种细胞组分中的每种相应的细胞组分,获取相应的细胞组分在相应的细胞中的对应的丰度,由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示确定维度的细胞组分计数数据结构。

[0567] 这一点是依据图14C的框1538,其说明,以电子形式获得第二数据集,该第二数据集包括:对于第二多个细胞(其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量(这里为多种不同化学化合物))中的每个相应的细胞,对于多种细胞组分中的每种相应的细胞组分,相应的细胞组分在相应的细胞中的对应的丰度,由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示确定维度的细胞组分

计数数据结构。

[0568] 该计数矩阵的形式的图示为图5的计数矩阵502。如图5的计数矩阵502所展示,对于每种相应的细胞组分(例如,基因),存在针对第二多个细胞中的每个细胞的表达数据。举例来说,跨第二多个细胞测量多个基因中的每一个的转录活性。细胞中的每一个均已暴露于协变量,这里为训练化学化合物。

[0569] 通过以下来形成激活数据结构:使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示,其中激活数据结构包括:对于多个细胞组分模块中的每个细胞组分模块,对于第二多个细胞中的每个细胞,相应的激活权重。

[0570] 对计数矩阵502与潜在表示404进行矩阵乘法以获得图5所展示的激活数据结构504。对于每个相应的细胞组分模块,对于第二多个细胞中的每个细胞,激活数据结构504具有激活值Act_{K-G},其值由潜在表示404与计数矩阵502的对应的矩阵乘法确定。

[0571] 使用以下两者之间的差异来训练候选细胞组分模型:(i)在将激活数据结构输入到候选模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测,以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在,其中该训练响应于差异而调整与候选细胞成分模型相关联的多个协变量权重。

[0572] 激活数据结构502用作用于图6的模型601的训练数据(标记数据),其本身是维度为N个化合物×M个细胞组分模块的潜在表示602。在此示例中,考虑了8000种不同的化合物和108个细胞组分模块。因此,在图5的命名中,Z为108,并且G为8000。激活数据结构以两种方式分割为训练集和测试集。首先,选择“随机分割”,其将1200种化合物分组为测试集,并将其余6800种化合物分组为训练集。另外,使用开源软件包RDKit中的函数定义了“跨骨架分割”,这确保测试集含有相比训练集具有不同的骨架的化合物。

[0573] 如图6所展示,激活数据结构504的每个相应的行是表示哪些化合物可能诱导由相应的行表示的对应的细胞组分模块的细胞组分的向量。模型601的每个实例均在激活数据结构504的行上进行训练。使用6800种训练化合物形成激活数据结构504。对于给定模型601,将特定化学化合物的指纹输入到模型601中,并响应于该输入,计算对应的细胞组分模块的预测激活值。可以将该预测激活值与激活数据结构504中的对应的元素中的实际激活值进行直接比较。因此,这样,可以计算以下两者之间的差异:(i)在将激活数据结构504输入到模型601中时对训练化合物中的每种化合物针对表示于激活数据结构504中的每个细胞组分模块的不存在或存在的预测,以及(ii)每种化合物针对每个细胞组分模块的实际不存在或存在,并将其用于通过响应于差异而调整与候选细胞组分模型相关联的多个协变量权重604来训练模型601。如图6所展示,多个协变量权重包括:对于多个细胞组分模块中的每个相应的细胞组分模块:对于每个相应的协变量:对应的权重,该对应的权重指示相应的协变量是否跨激活数据结构与相应的细胞组分模块相关。在一些实施例中,针对每个细胞组分模块存在不同的模型601。换句话说,参考图6,在一些实施例中,每一行604处于不同的模型601中。因此,在此类实施例中,使用激活数据结构中的对应行(例如,对应于与相应的模型601相同的细胞组分模块的行)来训练每个此类模型601。

[0574] 如图6所展示,经训练的模型601(或多个模型)提供针对每个协变量(这里为训练化学组合物)的权重。也就是说,模型601的潜在表示602提供了描述每个协变量(化学组合物)与细胞组分模块的激活相关联的程度的权重(例如,图6的权重$_{1-1}$或行604-1)。此类权重

被认为是细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分。在针对每个细胞组分模块形成不同的模型601的实施例中,潜在表示602为每个模型601的聚合潜在表示。在一些实施例中,该表示中的每个权重是分类的(例如,化合物影响细胞组分模块“0”或则化合物不影响细胞组分模块“1”。在其他实施例中,每个权重处于连续数值范围内,其中数值范围的一端指示训练化合物极大地影响细胞组分模块,并且数值范围的另一端指示训练化合物不影响细胞组分模块。如本文所用,术语“影响”取决于应用,但通常意味着化合物的不存在或存在改变细胞组分模块中的细胞组分的丰度。

[0575] 为训练模型601,在该示例中,将表示于图6的激活数据结构504中的化合物的SMILES表示变换为ECFP4指纹表示,并且另外变换为图表示。随后训练两个模型。也就是说,在该示例中,模型601为两个不同模型的集成:A)使用全连接神经网络架构以在ECFP4表示上进行训练,B)适应消息传递神经网络(MPNN)以在图表示上进行训练。使用开源软件包pytorch和DGL来进行此训练。训练未经训练的模型601,对于训练集中的每种相应的化合物的每个相应的化学结构,对于细胞组分模块的集合中的每个相应的细胞组分模块,使用以下两者之间的相应的差异来进行:(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分,以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分(获自激活数据结构504),其中该训练响应于差异而调整与未经训练的模型601相关联的多个参数,并且其中多个参数包括100个或更多个参数,由此获得经训练的模型。

[0576] 如上面所指出,在此示例中,模型601为以下两者的集成:(i)SMILES字符串的标准指纹上的全连接网络,其中网络架构为具有ReLU激活的3层网络,以及(ii)缺乏DGL库的MPNN网络。在输入化学结构信息后,模型601提供该模型在其上进行训练的每个细胞组分模块132的激活评分。

[0577] 事实上,在一些实施例中,在该示例中,针对每个细胞组分模块存在单独的集成模型601。换句话说,模型601为多任务编码器,其在输入化学结构时为多个细胞组分模块中的每一个提供单独的激活评分。更进一步地,在一些实施例中,如上面所讨论,针对每个相应的细胞组分模块存在单独的模型601。在此类实施例中,每个此类相应的模型601包括每种化合物相对于对应的细胞组分模型的激活权重。

[0578] 现在经训练的每个相应的模型601提供其对应的细胞组分模块针对任何化合物的激活评分,无论是否是训练集的一部分。也就是说,每个模型601能够报告其对应的细胞组分模块是否与测试化合物相关联。如果是,则模型输出指示其对应的细胞组分模块与测试化合物相关联的评分。在一些实施例中,该评分是分类的(例如,如果对应的细胞组分模块与测试化合物相关联则为“1”,并且如果不关联则为“0”)。在一些实施例中,该评分为概率或可能性(例如,在0至1的数值范围内),其中较接近1的数字(例如,0.85)指示对应的细胞组分模块与测试化合物相关联的可能性。在一些实施例中,该评分在介于“A”与“B”之间的连续数值范围内,其中A和B为两个不同的数字。由于存在数个模型601,每个模型对应于不同细胞组分模块,因此针对数个不同的模型601运行测试化合物以确定哪些细胞组分模块被该化合物激活(与其相关联)。在每种情况下,如上面所讨论,化学结构转换为指纹,并且正是该指纹被应用于每个模型。注意,从生物学角度来看,可以预期,给定的测试化合物可以激活任意数量的不同细胞组分模块(例如,1、2、3、4、5或更多个)。此外,本公开中描述的

方法可以通过模型601尚未在其上进行训练但是对于其已知的是,哪些细胞组分模块应该被测试化合物激活的测试化合物来验证。在此示例中,这是按如下所述完成的。特别地,在该示例中,对用于将化合物与生理状况相关联的经训练的模型601进行了4重验证。此测试跟随如最初提交的权利要求1。

[0579] 首先,针对脂肪酸产生相关细胞组分模块的由以下诱导的激活来获得来自模型601的模型预测:来自高通量筛选的上述1200种随机选择的未见过的化合物,以及上述1200种骨架与6800种化合物的训练集不重叠的化合物。针对随机选择的化合物获得的相应的模型601预测(预测的细胞组分激活评分)展示于图10B中。也就是说,图10B示出了来自两种不同的模型601的结果,一种针对细胞组分模块78(“模块78”),且另一种针对细胞组分模块“90”。模块78表示对细胞代谢重要的脂肪酸相关细胞过程,并且其对应的经训练的模型601表现出高决定系数($R^2=0.28$)。相反,针对与细胞代谢无关的细胞组分“模块90”(模块90中的细胞组分不涉及脂肪酸相关过程)的从同一scRNA-seq数据集学习的训练模型601具有低的决定系数($R^2=0.08$)。所有基准均产生高度显著的相关性(皮尔逊相关系数p_s分别等于~0.5和~0.2)。

[0580] 用最初提交的权利要求1的语言风格来说,该第一验证方法提供了将测试化学化合物(来自高通量筛选的所描述的1200种随机选择的未见过的化合物以及上述1200种骨架与6800种化合物的训练集不重叠的化合物中的一种)与目的生理状况(在此示例中,这里为对细胞代谢重要的脂肪酸相关联细胞过程)的方法。该方法包括在包括存储器和一个或多个处理器的计算机系统处,获得测试化学化合物的化学结构的指纹。因此,在此示例中,获得测试化学化合物的化学结构的指纹,并将其输入到图1的每个模型601中。在最初提交的权利要求1的上下文中,该模型称为模型。此模型涵盖集成模型,其中该集成模型中的每个成分模型包括针对图6的模型601列出的参数的单个行,该行是针对与成分模型相关联的给定的细胞组分模块的权重的参数。将认识到,虽然在图6中此类权重被表示为单个行,但是不要求它们在集成模型的成分模型中呈行格式,其任何等同形式均在本公开的范围内。此外,虽然图6的模型601包括针对每种化合物(该模型针对其进行了训练)的单个权重,该单个权重适合于基于回归的模型601,但在一些实施例中,模型601中的权重的数量与化合物(模型针对其进行了训练)的数量之间没有明确的关系。在一些实施例中,模型601包括100或更多、1000或更多、10,000或更多,或者100,000或更多个参数。

[0581] 根据最初提交的权利要求1,将测试化合物的指纹输入到模型中。如最初提交的权利要求1所指出,该模型包括100个或更多个参数。换句话说,在输入测试化合物的指纹后,对模型输出的计算不能在头脑中进行。模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分。一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块。在该示例中,该模型为模型601的集成,每个模型表示不同细胞组分模块,并且因此集成中的每个模型601输出一个或多个计算出的激活评分中的计算的激活评分,其表示细胞组分模块的集合中的单个对应的细胞组分模块。在这方面,并且如上面所指出,细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。此外,细胞组分模块的集合中的至少第一细胞组分模块与目的生理状况相关联。在该示例中,模块78与目的生理状况相关联。如图10B所指示,(例如,通过针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准)识别出正确激活模

块78并因此与模块78的目的生理状况(对细胞代谢重要的脂肪酸相关的细胞过程)相关联的那些化合物。

[0582] 作为所要求保护的方法的第二验证,随后将针对模块78和90的相应的经训练的模型应用于暴露于前脂肪细胞的某些小分子“合成苗头”(尚未在训练期间引入图6的模型601的另一测试集)的scRNA-seq表征。图10D展示了由针对模块78的经训练的模型601通过合成苗头化合物指示的对激活的高相关性的和忠实的预测,相比之下,由针对模块90的经训练的模型601通过合成苗头指示的激活很少或没有激活。

[0583] 第三,使用针对模块78的经训练的模型601来预测细胞组分模块78(模块78)的针对从公共数据库中的五百万种化合物采样的200,000种化合物的随机子集的细胞组分激活评分。由此,选择预测高度激活细胞组分模块78的前50种化合物,并将其与数据库中的化合物的集合进行比较,包括来自LINCS L1000数据集的化合物和衍生自已知化合物(本文称为已知含哌啶化合物(“KPCC”))的化学结构的合成苗头类似物。该比较的分布展示在图10E中。在分布的尾端,针对细胞组分模块78的经训练的模型601获得的预测识别出显著超过LINCS和合成苗头中的所有化合物的化合物。此方法突出了用于针对特定期望细胞过程来优化化学结构的方法。

[0584] 第四,对在前50个预测中识别的化学结构进行目视检查,并发现其含有表示已知的脂肪组织靶向药效团)且从而正确激活与模块78相关联的细胞组分模块的明显化学结构。

[0585] 此第一实例也跟随最初提交的权利要求58。权利要求1与权利要求58之间的差异在于扰动特征与细胞组分模块中的一者。通过比较已经受到扰动的细胞与没有受到扰动的细胞的表达来获得扰动特征。因此,可以使用已知诱导前脂肪细胞的代谢活性过程的小分子扰动原。将前脂肪细胞系暴露于扰动源24小时,并且针对受扰动的条件和对照条件可以获得scRNA-seq读数。由此,可以获得扰动特征。替代性地,可以通过比较已暴露于用于第二数据集的化学协变量中的任一个的细胞的细胞表达来获得单独的扰动特征。事实上,对于用于第二数据集的化学协变量中的每一个,可以按此方式获得单独的扰动特征。每个此类扰动特征具有潜在表示404中的行的形式,不同的是每个此类权重现在处于连续数值范围,而非二元数值范围内。举例来说,在一些实施例中,每个权重为处于介于0与1之间的连续数值范围(或某个其他范围“A”至“B”,其中A和B为两个不同的数字,诸如-100和100)内的值。在此之后,训练的过程与上面关于对潜在表示404、计数矩阵502、激活数据结构的使用和对成分模型601的训练所讨论的过程相同,其中每个此类模型现在表示扰动特征的集合中的不同的扰动特征。

[0586] 实例2.预测用于激活胎儿红细胞生成程序并阻止T细胞耗竭的化学结构。

[0587] 在两个另外的示例中,在对与胎儿红细胞生成以及与T细胞耗竭相关的两个scRNA-seq数据集上训练两个模型。

[0588] 对于胎儿红细胞生成,用工具化合物CLT-AAA-12来处理CD34造血干细胞,先前对于该工具化合物已确立诱导出胎儿红细胞生成的终点标志物,特别是测定中F细胞的数量,如用流式细胞术所读出。

[0589] 对于T细胞耗竭,用耗竭诱导培养基处理幼稚T细胞。

[0590] 两种细胞系统均以scRNA-seq进行表征。随后,通过以下将药物反射器模型(参见,

2019年7月15日提交的名称为“Methods of Analyzing Cells”的美国专利申请号16/511,691,其特此通过引用并入)应用于scRNA-seq数据集:输入由其相应的样品中的受扰动细胞与对照细胞定义的细胞状态转变。药物反射器为药物反射器潜在表示中的8000种化合物中的每一种指派细胞状态激活评分。这产生具有针对两种转变(胎儿血红蛋白和T细胞耗竭)的细胞状态激活评分的两个向量。这两个向量充当用于模型601的训练数据。

[0591] 使用该模型来预测激活造血干细胞的胎儿红细胞生成和T细胞耗竭的化合物。造血干细胞的胎儿红细胞生成为近年来带来了针对镰状细胞病的突破性CRISPR疗法的细胞过程,而T细胞耗竭是阻碍癌症的检查点抑制剂疗法的更广泛成功的关键机制。

[0592] 使用从公共数据库中的5百万种化合物中采样的2,000种化合物的子集来进行预测,其中随机分割或者在骨架上分割子集。图11的顶部图画示出了本示例的模型在2,000种在骨架上且随机地分割的化合物的测试集上的性能,示出了所采样的化合物的显著的R^2和相关系数p_s,其中苗头化合物CLT-AAA-12的扰动特征与造血干细胞的胎儿红细胞生成相关。图11的底部图画示出了2,000种在骨架上且随机地分割的化合物的测试集的性能,示出了所采样的化合物的显著的R^2和相关系数p_s,其中细胞转变特征与T细胞耗竭相关。因此,图11表明,模型601能够预测诱导与目的扰动特征和/或细胞转变特征相同的细胞行为效应的新的骨架。

[0593] 实例3.基于疾病关键细胞行为的特征归因:预测用于设计新分子的药效团。

[0594] 如实例1中所描述,根据本文所公开的系统和方法预测的化学结构可以用于识别与目的生理状况(例如,脂肪组织靶向的)潜在相关的分子特征,诸如药效团。与实例1中一样,这些药效团可以通过已知的化学结构来验证,或者可以呈现新的结构以供进一步验证。举例来说,基于药效团的算法的示例用例包括利用具有先前在文献中描述的功能含义的药效团的数据库,包括生物同位可交换替换物库(BoBER)数据库。用例的另一示例包括(诸如由药物化学家)应用专业知识来获得有关所识别的药效团在系统对扰动的复杂响应中的作用的直觉知识。

[0595] 进行用于预测用于设计新分子的药效团的模型,其中该模型包括基于使用Teversky相似性的评分选择的来自干预库的小分子的特征化,以得到表示药效团是否包含在化学结构中的表示。将该表示(化学指纹)输入到实例1的针对模块78的模型601中。使用在实例1中识别的脂肪靶向药效团,将实例1的针对模块78的模型用于确定已知含哌啶化合物(“KPCC”)的脂肪靶向药效团的关联以,且与激活评分范围为.04064至0.04633的脂肪酸模块的观察到的转录激活隔离。

[0596] 实例4.基于潜在细胞行为产生合成苗头化合物。

[0597] 作为测试例,基于以下来设计六种新合成的小分子苗头(本文称为“六种合成苗头”):经体外和体内验证的脂肪细胞米色化化合物及其潜在空间表示。六种合成苗头中的每一种均引发人前脂肪细胞的期望细胞行为变化。首先,识别KPCC聚类的药效团。然后通过在该聚类中进行分子的药效团富集以及掺入新型生物等排体来设计分子,产生六种合成苗头的最终设计。这六种结构多样的合成苗头的目标是诱导与包括KPCC在内的已知化学实体(KCE)相同的细胞行为效应。如图13中的示意图所展示,通过以下来确定细胞行为效应:用1 μM KPCC和六种合成苗头来处理人前脂肪细胞24小时,使用scRNA-seq来测量基因表达,并评估由上面实例1中描述的细胞代谢基因模块(模块78)的变化表示的细胞响应。举例来说,

脂肪代谢模块中的基因包括FABP3、FDPS和LPIN1等。

[0598] 对这些化合物对前脂肪细胞的影响的评估揭示,每种合成苗头激活与KPCC相同的脂肪代谢基因模块(图13;框1302中突出显示的模块78)。也就是说,突出显示的框1302示出了由实例1的针对模块78的模型在将化合物的指纹输入到模型中时输出的在图13中的图表的Y轴上列出的激活评分。这些结果为基于模型平台产生可预测地靶向期望细胞行为的合成苗头的能力提供高置信度。特别地,本公开的模型601(例如,实例1的针对模块78的模型601)可以用于预测靶向与生理状况相关联的基因模块的合成苗头,而不需要进行高通量筛选、基于分子靶点的识别或优化,或合成用于验证的数百或数千种新化合物。

[0599] 引用的参考文献及替代实施例

[0600] 本文引用的所有参考文献通过引用整体并入本文并且用于所有目的,其程度如同每个单独的出版物或专利或专利申请被具体地和单独地指出通过引用整体并入用于所有目的。

[0601] 本发明可被实现为计算机程序产品,其包括嵌入于非暂时性计算机可读存储介质中的计算机程序机制。举例来说,计算机程序产品可以包含图1至3和7至9的任意组合中所示的程序模块。这些程序模块可以存储在CD-ROM、DVD、磁碟存储产品或任何其他非暂时性计算机可读数据或程序存储产品上。

[0602] 如熟习此项技术者将显而易见,在不背离本发明之精神及范围的情况下可对其作出许多修改及变化。本文中所描述之特定实施例仅作为实例提供。选择并描述该等实施例以便最佳地解释本发明之原理及其实际应用,以藉此使其他熟习此项技术者能够最佳地利用本发明及具有适合于所涵盖之特定用途之各种修改的各种实施例。本发明仅受随附申请专利范围之条款以及此申请专利范围有权享有之等效物的全部范围的限制。

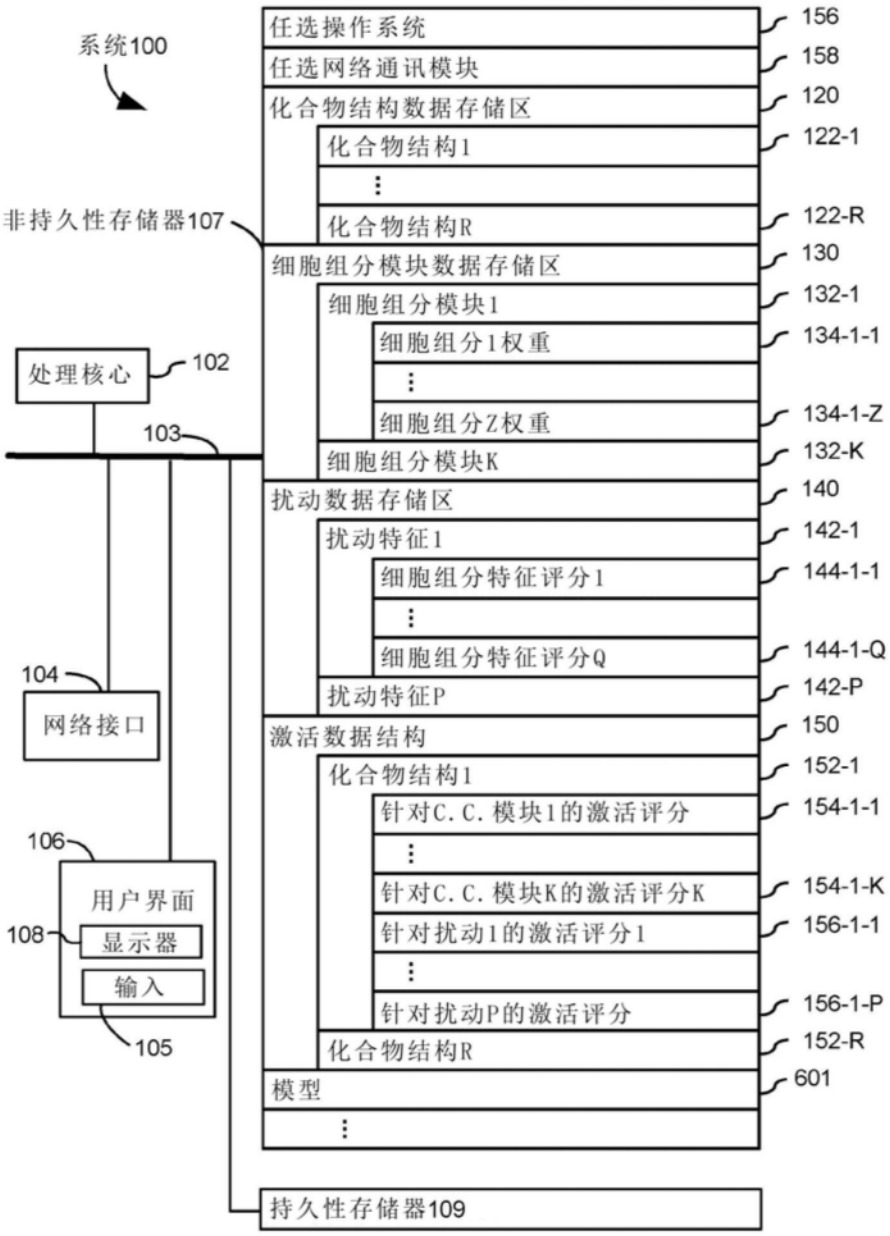

系统100
非持久性存储器107
处理核心
102
103
104
网络接口
106
用户界面
108
显示器
输入
105
任选操作系统
156
任选网络通讯模块
158
化合物结构数据存储区
120
化合物结构1
122-1
化合物结构R
122-R
细胞组分模块数据存储区
130
细胞组分模块1
132-1
细胞组分1权重
134-1-1
细胞组分Z权重
134-1-Z
细胞组分模块K
132-K
扰动数据存储区
140
扰动特征1
142-1
细胞组分特征评分1
144-1-1
细胞组分特征评分Q
144-1-Q
扰动特征P
142-P
激活数据结构
150
化合物结构1
152-1
针对C.C.模块1的激活评分
154-1-1
针对C.C.模块K的激活评分K
154-1-K
针对扰动1的激活评分1
156-1-1
针对扰动P的激活评分
156-1-P
化合物结构R
152-R
模型
601
持久性存储器109

图1

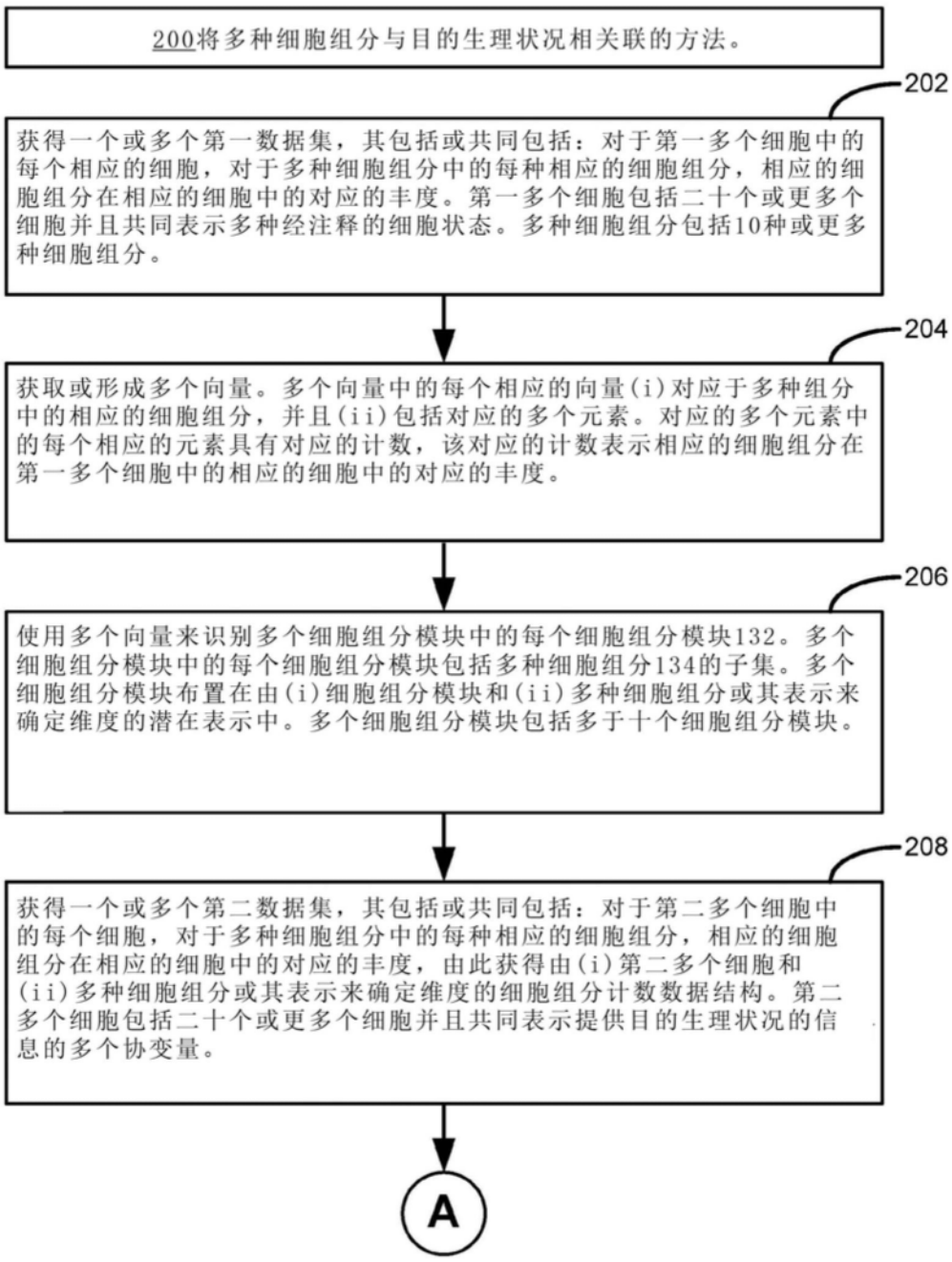

200将多种细胞组分与目的生理状况相关联的方法。
202
获得一个或多个第一数据集，其包括或共同包括：对于第一多个细胞中的每个相应的细胞，对于多种细胞组分中的每种相应的细胞组分，相应的细胞组分在相应的细胞中的对应的丰度。第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态。多种细胞组分包括10种或更多种细胞组分。
204
获取或形成多个向量。多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分，并且(ii)包括对应的多个元素。对应的多个元素中的每个相应的元素具有对应的计数，该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。
206
使用多个向量来识别多个细胞组分模块中的每个细胞组分模块132。多个细胞组分模块中的每个细胞组分模块包括多种细胞组分134的子集。多个细胞组分模块布置在由(i)细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中。多个细胞组分模块包括多于十个细胞组分模块。
208
获得一个或多个第二数据集，其包括或共同包括：对于第二多个细胞中的每个细胞，对于多种细胞组分中的每种相应的细胞组分，相应的细胞组分在相应的细胞中的对应的丰度，由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示来确定维度的细胞组分计数数据结构。第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量。
A

图2A

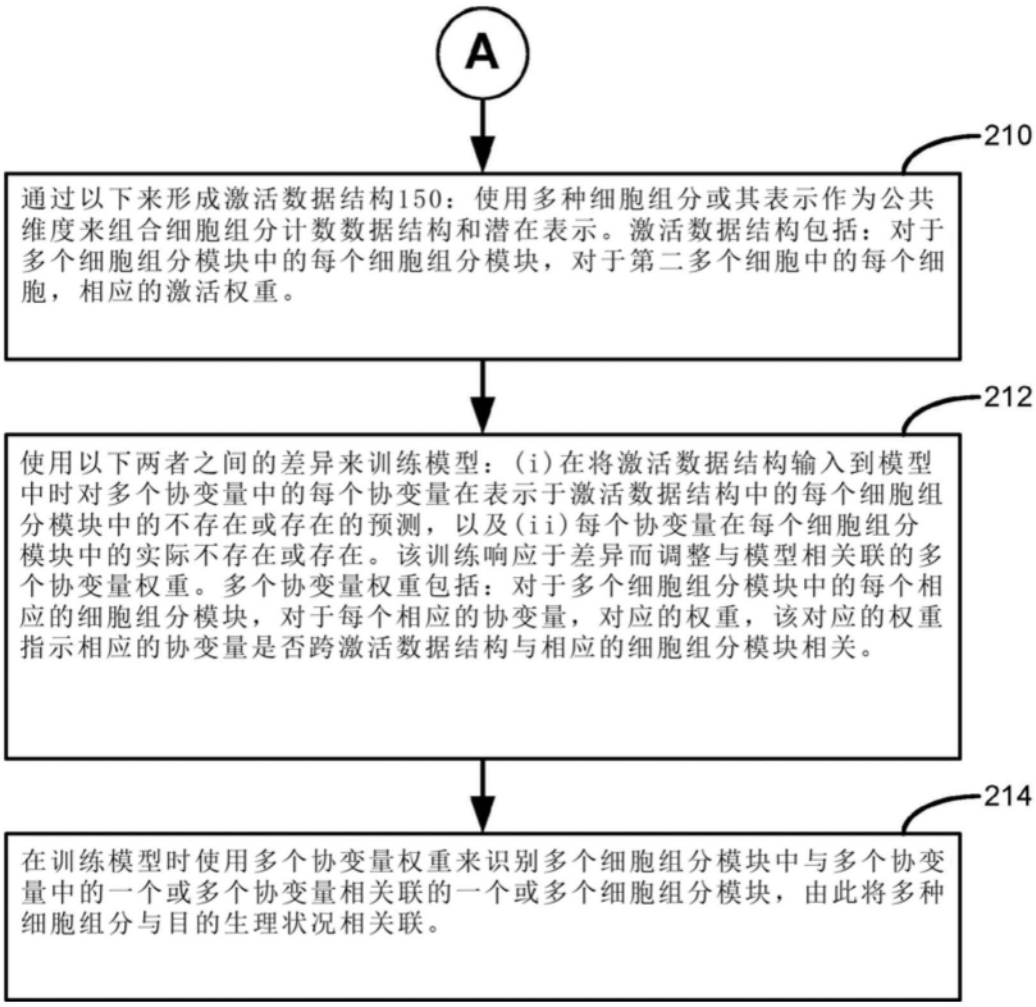

A
210
通过以下来形成激活数据结构150：使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示。激活数据结构包括：对于多个细胞组分模块中的每个细胞组分模块，对于第二多个细胞中的每个细胞，相应的激活权重。
212
使用以下两者之间的差异来训练模型：(i)在将激活数据结构输入到模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测，以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在。该训练响应于差异而调整与模型相关联的多个协变量权重。多个协变量权重包括：对于多个细胞组分模块中的每个相应的细胞组分模块，对于每个相应的协变量，对应的权重，该对应的权重指示相应的协变量是否跨激活数据结构与相应的细胞组分模块相关。
214
在训练模型时使用多个协变量权重来识别多个细胞组分模块中与多个协变量中的一个或多个协变量相关联的一个或多个细胞组分模块，由此将多种细胞组分与目的生理状况相关联。

图2B

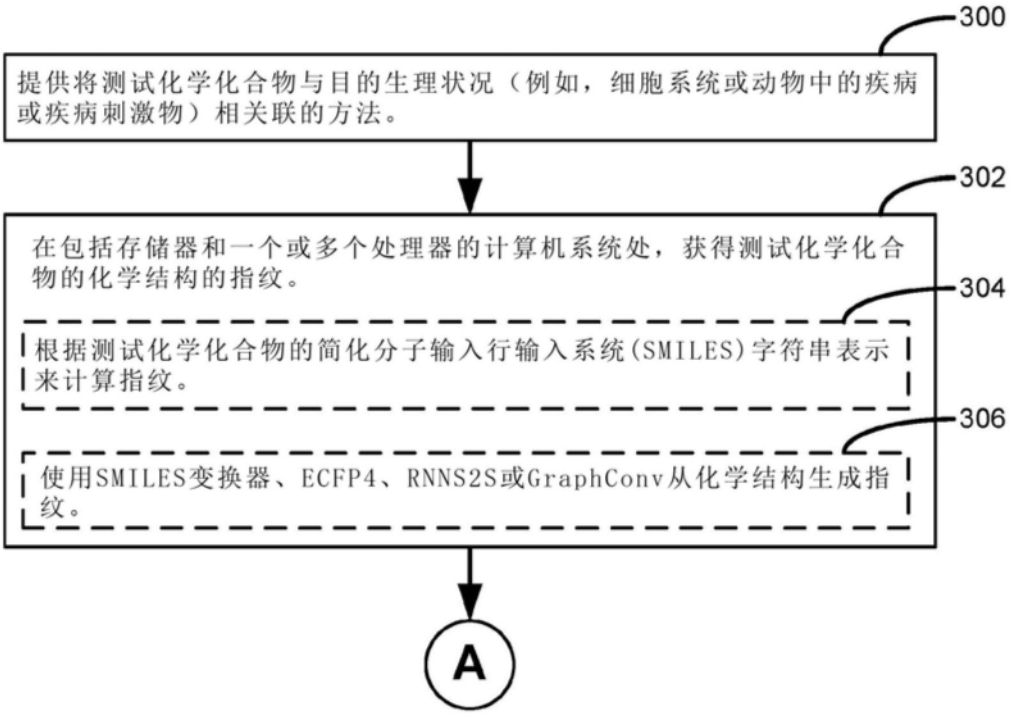

300
提供将测试化学化合物与目的生理状况（例如，细胞系统或动物中的疾病或疾病刺激物）相关联的方法。
302
在包括存储器和一个或多个处理器的计算机系统处，获得测试化学化合物的化学结构的指纹。
304
根据测试化学化合物的简化分子输入行输入系统(SMILES)字符串表示来计算指纹。
306
使用SMILES变换器、ECFP4、RNNS2S或GraphConv从化学结构生成指纹。
A

图3A

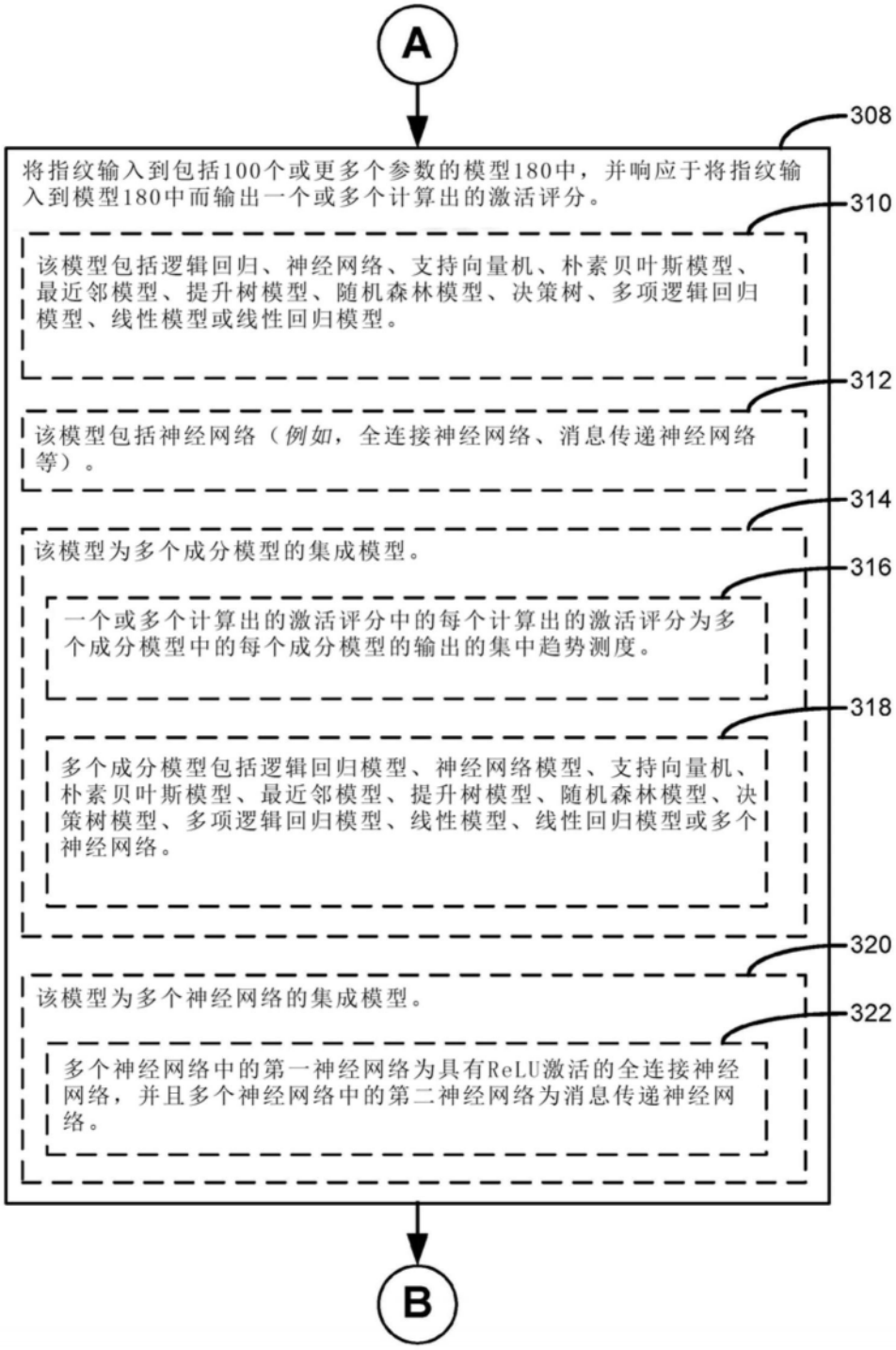

A
308
将指纹输入到包括100个或更多个参数的模型180中，并响应于将指纹输入到模型180中而输出一个或多个计算出的激活评分。
310
该模型包括逻辑回归、神经网络、支持向量机、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树、多项逻辑回归模型、线性模型或线性回归模型。
312
该模型包括神经网络（例如，全连接神经网络、消息传递神经网络等）。
314
该模型为多个成分模型的集成模型。
316
一个或多个计算出的激活评分中的每个计算出的激活评分为多个成分模型中的每个成分模型的输出的集中趋势测度。
318
多个成分模型包括逻辑回归模型、神经网络模型、支持向量机、朴素贝叶斯模型、最近邻模型、提升树模型、随机森林模型、决策树模型、多项逻辑回归模型、线性模型、线性回归模型或多个神经网络。
320
该模型为多个神经网络的集成模型。
322
多个神经网络中的第一神经网络为具有ReLU激活的全连接神经网络，并且多个神经网络中的第二神经网络为消息传递神经网络。
B

图3B

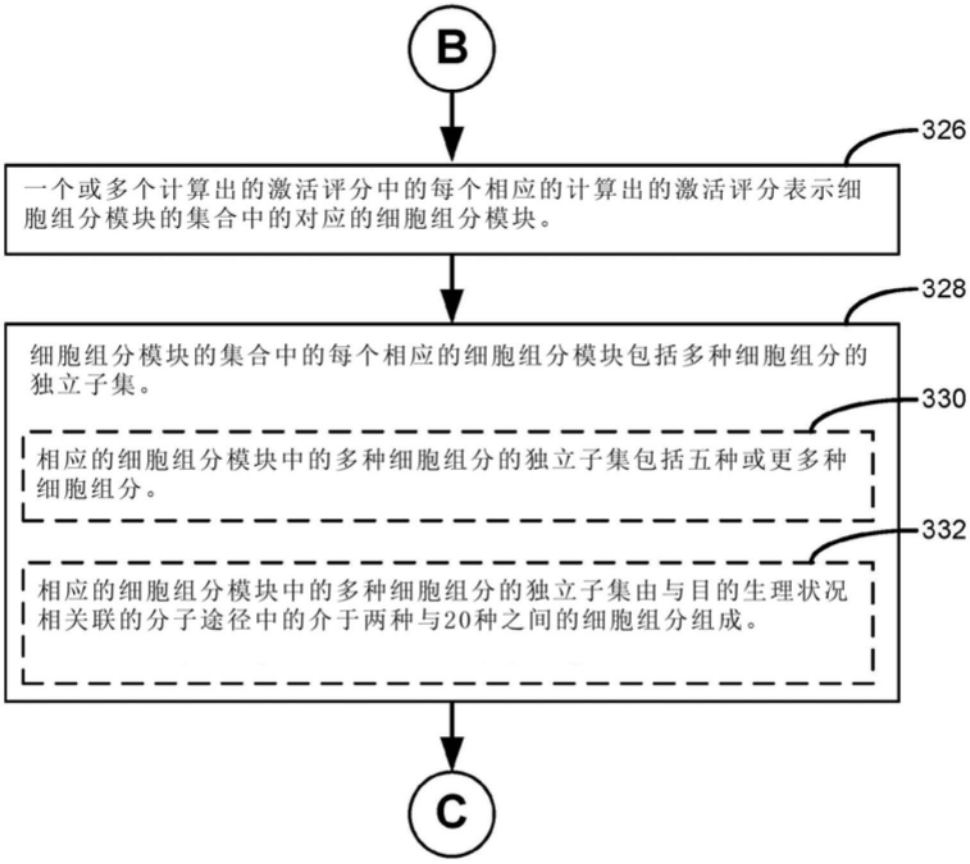

B
326
一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块。
328
细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。
330
相应的细胞组分模块中的多种细胞组分的独立子集包括五种或更多种细胞组分。
332
相应的细胞组分模块中的多种细胞组分的独立子集由与目的生理状况相关联的分子途径中的介于两种与20种之间的细胞组分组成。
C

图3C

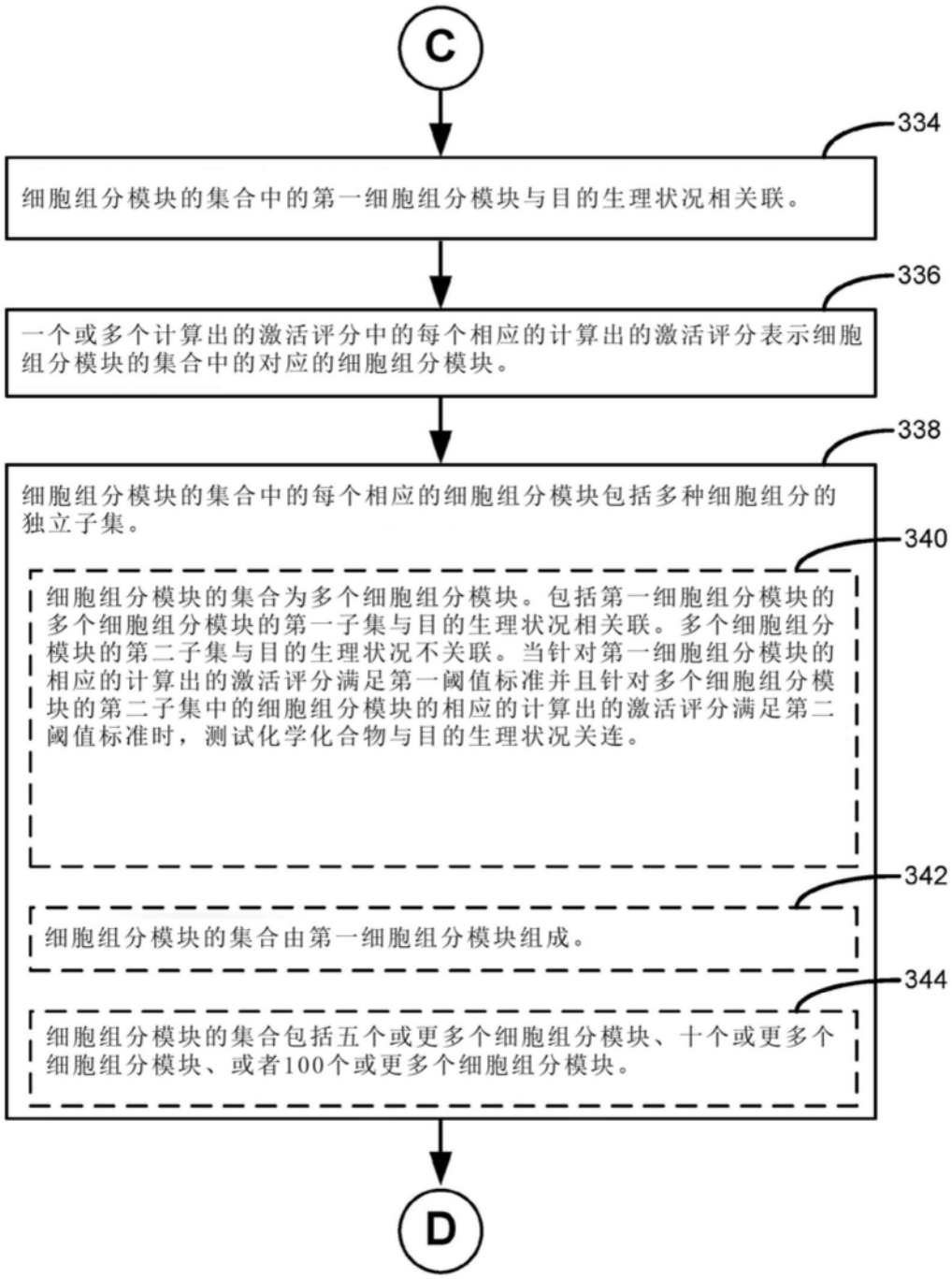

C
334
细胞组分模块的集合中的第一细胞组分模块与目的生理状况相关联。
336
一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示细胞组分模块的集合中的对应的细胞组分模块。
338
细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分的独立子集。
340
细胞组分模块的集合为多个细胞组分模块。包括第一细胞组分模块的多个细胞组分模块的第一子集与目的生理状况相关联。多个细胞组分模块的第二子集与目的生理状况不关联。当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准并且针对多个细胞组分模块的第二子集中的细胞组分模块的相应的计算出的激活评分满足第二阈值标准时，测试化学化合物与目的生理状况关连。
342
细胞组分模块的集合由第一细胞组分模块组成。
344
细胞组分模块的集合包括五个或更多个细胞组分模块、十个或更多个细胞组分模块、或者100个或更多个细胞组分模块。
D

图3D

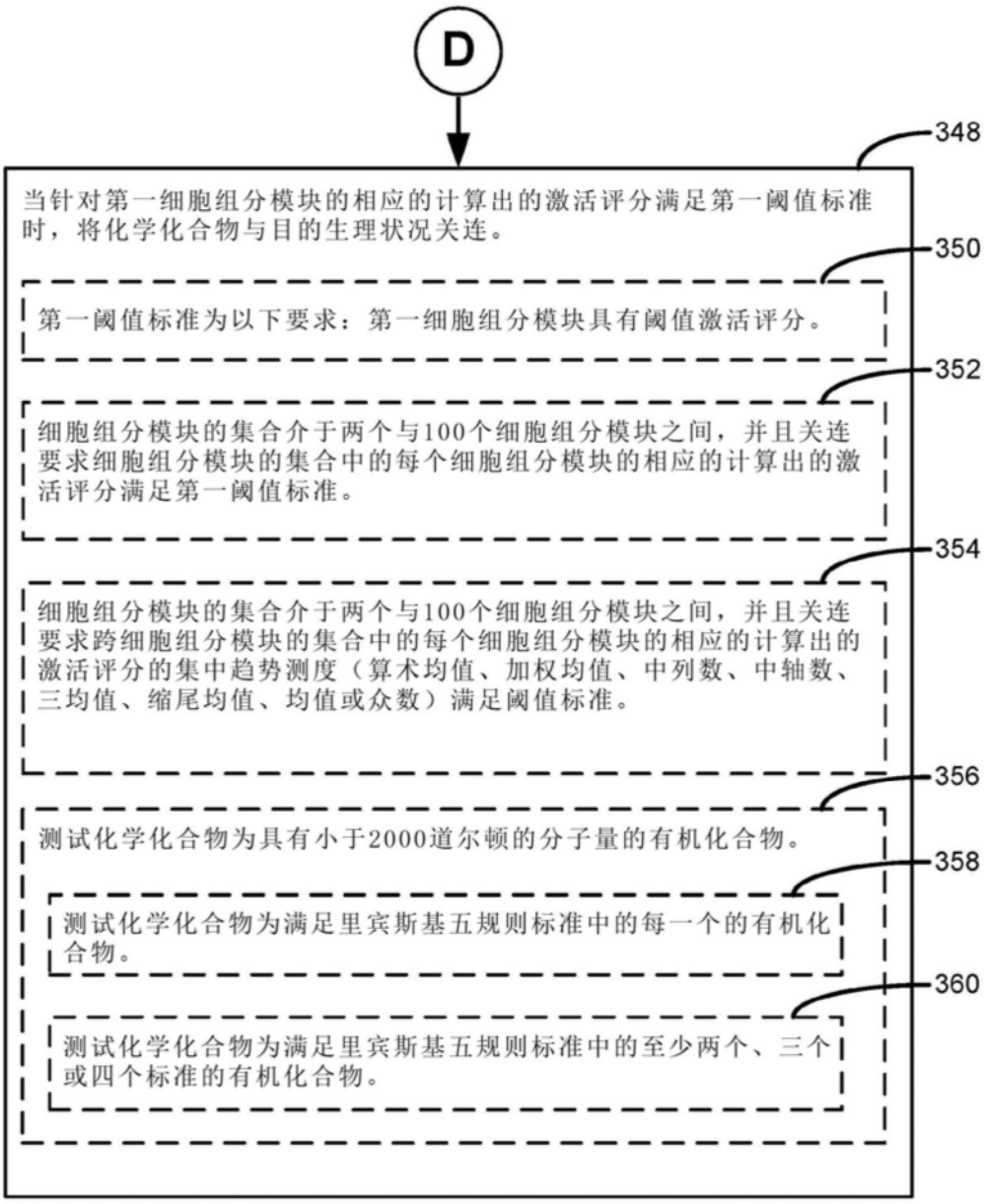

D
348
当针对第一细胞组分模块的相应的计算出的激活评分满足第一阈值标准时，将化学化合物与目的生理状况关连。
350
第一阈值标准为以下要求：第一细胞组分模块具有阈值激活评分。
352
细胞组分模块的集合介于两个与100个细胞组分模块之间，并且关连要求细胞组分模块的集合中的每个细胞组分模块的相应的计算出的激活评分满足第一阈值标准。
354
细胞组分模块的集合介于两个与100个细胞组分模块之间，并且关连要求跨细胞组分模块的集合中的每个细胞组分模块的相应的计算出的激活评分的集中趋势测度（算术均值、加权均值、中列数、中轴数、三均值、缩尾均值、均值或众数）满足阈值标准。
356
测试化学化合物为具有小于2000道尔顿的分子量的有机化合物。
358
测试化学化合物为满足里宾斯基五规则标准中的每一个的有机化合物。
360
测试化学化合物为满足里宾斯基五规则标准中的至少两个、三个或四个标准的有机化合物。

图3E

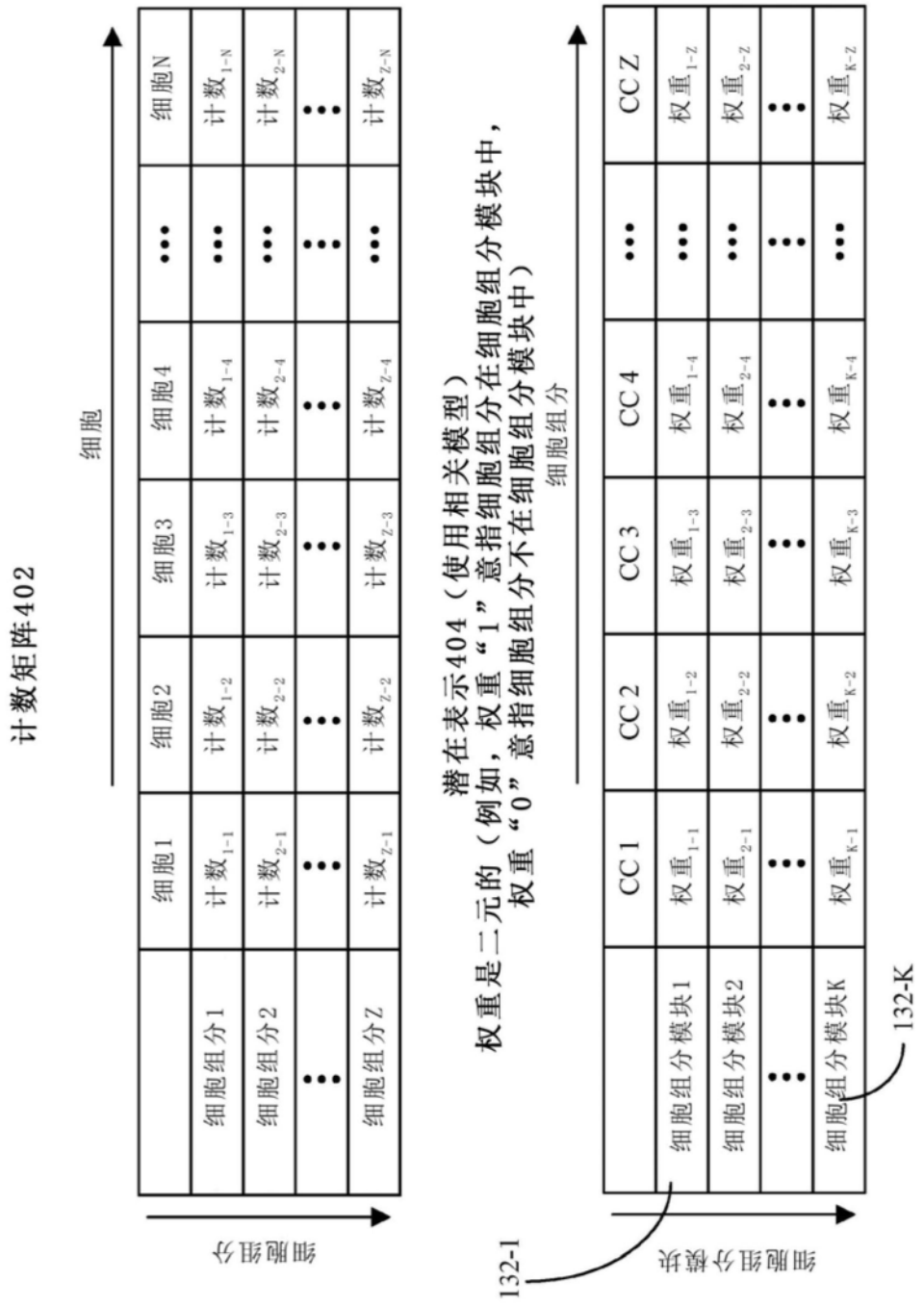

计数矩阵402
细胞
细胞1
细胞2
细胞3
细胞4
细胞N
细胞组分1
细胞组分2
细胞组分Z
计数$_{1-1}$
计数$_{1-2}$
计数$_{1-3}$
计数$_{1-4}$
计数$_{1-N}$
计数$_{2-1}$
计数$_{2-2}$
计数$_{2-3}$
计数$_{2-4}$
计数$_{2-N}$
计数$_{Z-1}$
计数$_{Z-2}$
计数$_{Z-3}$
计数$_{Z-4}$
计数$_{Z-N}$
细胞组分
潜在表示404（使用相关模型）
权重是二元的（例如，权重“1”意指细胞组分在细胞组分模块中，
权重“0”意指细胞组分不在细胞组分模块中）
细胞组分
CC 1
CC 2
CC 3
CC 4
CC Z
细胞组分模块1
细胞组分模块2
细胞组分模块K
权重$_{1-1}$
权重$_{1-2}$
权重$_{1-3}$
权重$_{1-4}$
权重$_{1-Z}$
权重$_{2-1}$
权重$_{2-2}$
权重$_{2-3}$
权重$_{2-4}$
权重$_{2-Z}$
权重$_{K-1}$
权重$_{K-2}$
权重$_{K-3}$
权重$_{K-4}$
权重$_{K-Z}$
132-1
132-K
细胞组分模块

图4

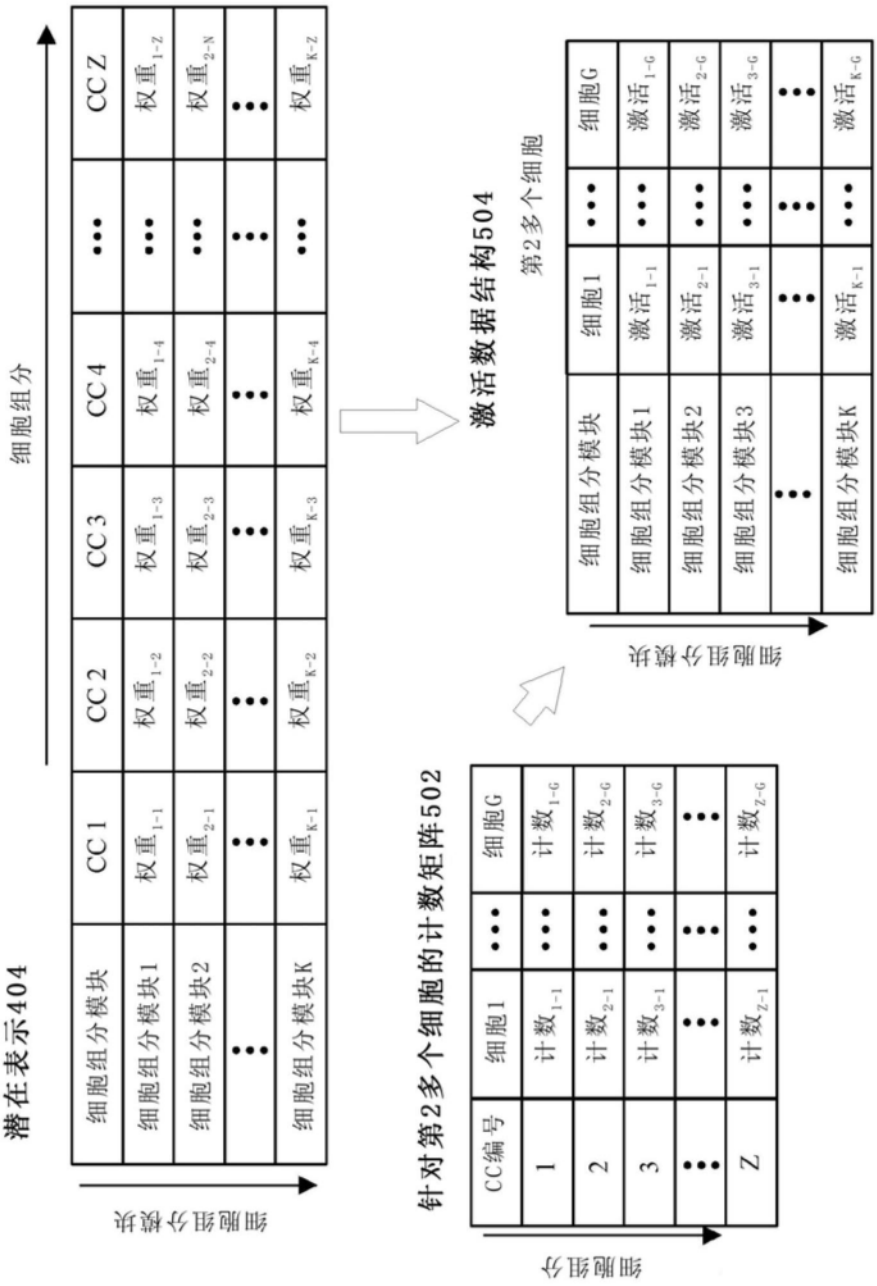

潜在表示404
细胞组分
细胞组分模块
CC 1
CC 2
CC 3
CC 4
CC Z
细胞组分模块1
细胞组分模块2
细胞组分模块K
权重1-1
权重1-2
权重1-3
权重1-4
权重1-Z
权重2-1
权重2-2
权重2-3
权重2-4
权重2-N
权重K-1
权重K-2
权重K-3
权重K-4
权重K-Z
针对第2多个细胞的计数矩阵502
CC编号
细胞1
细胞G
计数1-1
计数1-G
计数2-1
计数2-G
计数3-1
计数3-G
计数Z-1
计数Z-G
激活数据结构504
第2多个细胞
激活1-1
激活1-G
激活2-1
激活2-G
激活3-1
激活3-G
激活K-1
激活K-G

图5

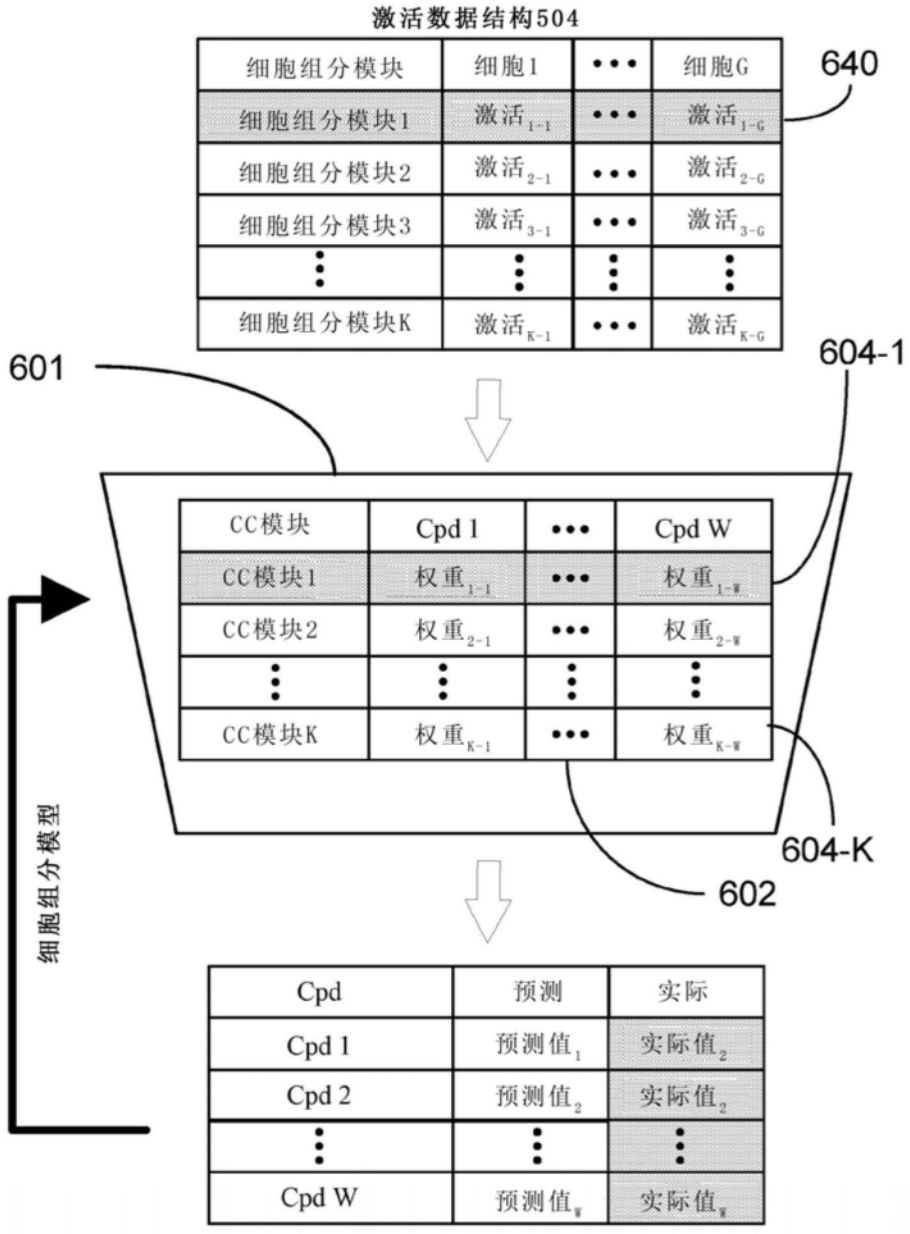

激活数据结构504
细胞组分模块
细胞1
细胞G
640
细胞组分模块1
激活1-1
激活1-G
细胞组分模块2
激活2-1
激活2-G
细胞组分模块3
激活3-1
激活3-G
细胞组分模块K
激活K-1
激活K-G
601
604-1
CC模块
Cpd 1
Cpd W
CC模块1
权重1-1
权重1-W
CC模块2
权重2-1
权重2-W
CC模块K
权重K-1
权重K-W
604-K
602
细胞组分模型
Cpd
预测
实际
Cpd 1
预测值1
实际值2
Cpd 2
预测值2
实际值2
Cpd W
预测值W
实际值W

图6

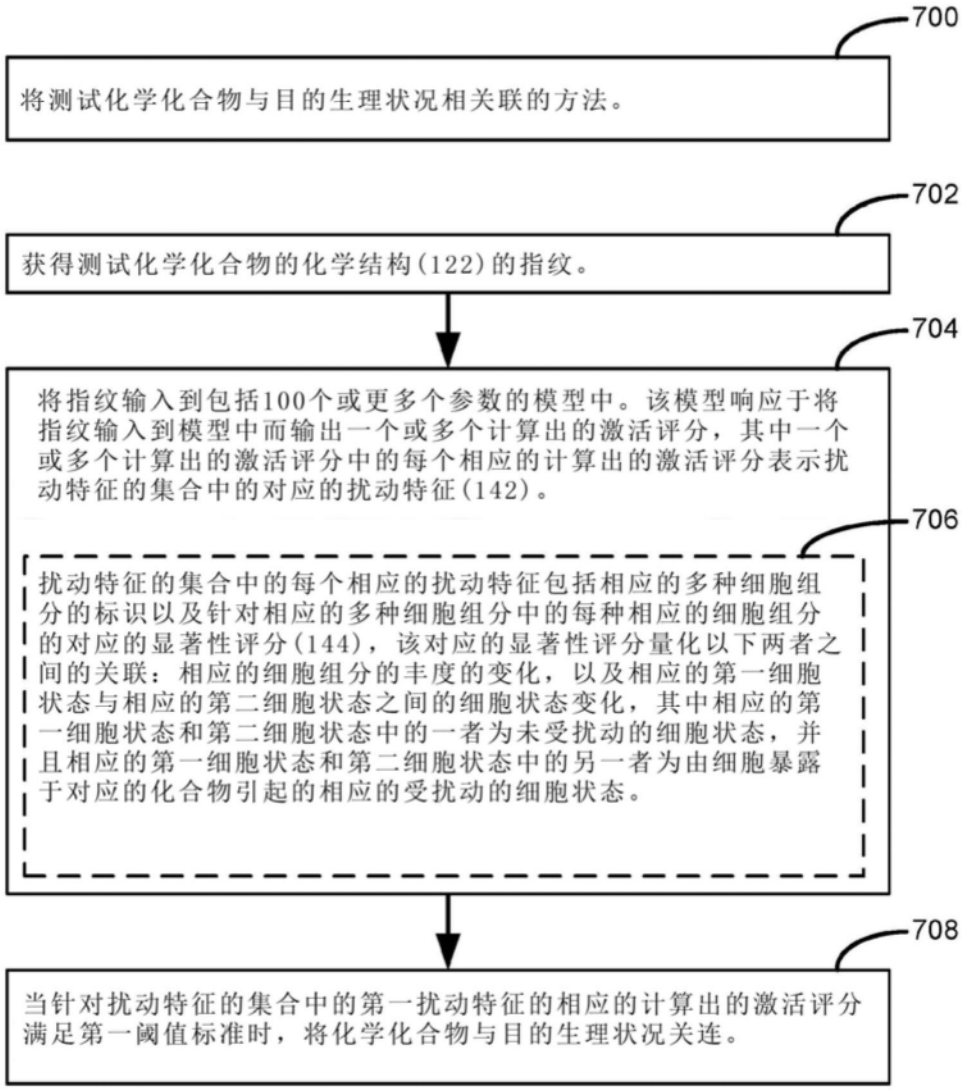

700
将测试化学化合物与目的生理状况相关联的方法。
702
获得测试化学化合物的化学结构(122)的指纹。
704
将指纹输入到包括100个或更多个参数的模型中。该模型响应于将指纹输入到模型中而输出一个或多个计算出的激活评分，其中一个或多个计算出的激活评分中的每个相应的计算出的激活评分表示扰动特征的集合中的对应的扰动特征(142)。
706
扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分(144)，该对应的显著性评分量化以下两者之间的关联：相应的细胞组分的丰度的变化，以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化，其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态，并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。
708
当针对扰动特征的集合中的第一扰动特征的相应的计算出的激活评分满足第一阈值标准时，将化学化合物与目的生理状况关连。

图7

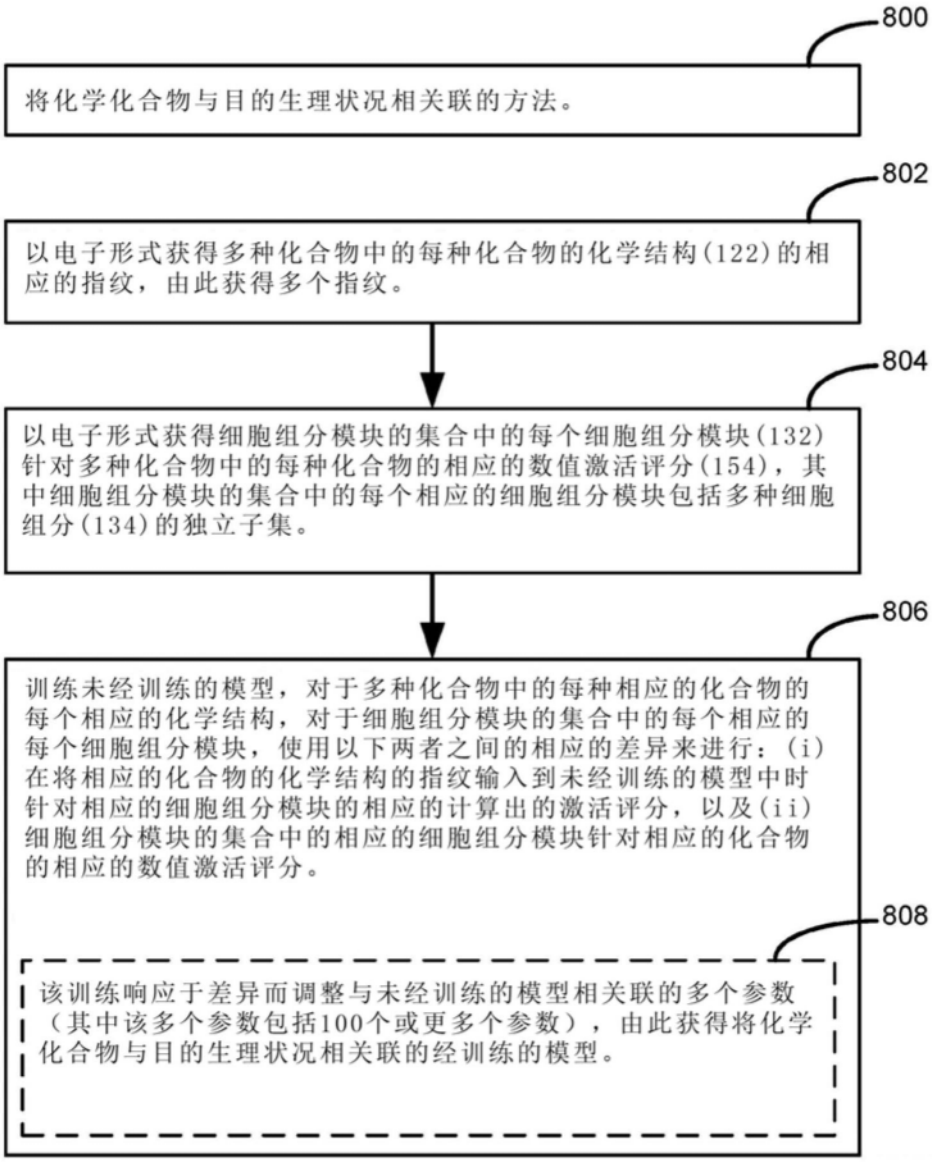

800
将化学化合物与目的生理状况相关联的方法。
802
以电子形式获得多种化合物中的每种化合物的化学结构(122)的相应的指纹，由此获得多个指纹。
804
以电子形式获得细胞组分模块的集合中的每个细胞组分模块(132)针对多种化合物中的每种化合物的相应的数值激活评分(154)，其中细胞组分模块的集合中的每个相应的细胞组分模块包括多种细胞组分(134)的独立子集。
806
训练未经训练的模型，对于多种化合物中的每种相应的化合物的每个相应的化学结构，对于细胞组分模块的集合中的每个相应的每个细胞组分模块，使用以下两者之间的相应的差异来进行：(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的细胞组分模块的相应的计算出的激活评分，以及(ii)细胞组分模块的集合中的相应的细胞组分模块针对相应的化合物的相应的数值激活评分。
808
该训练响应于差异而调整与未经训练的模型相关联的多个参数（其中该多个参数包括100个或更多个参数），由此获得将化学化合物与目的生理状况相关联的经训练的模型。

图8

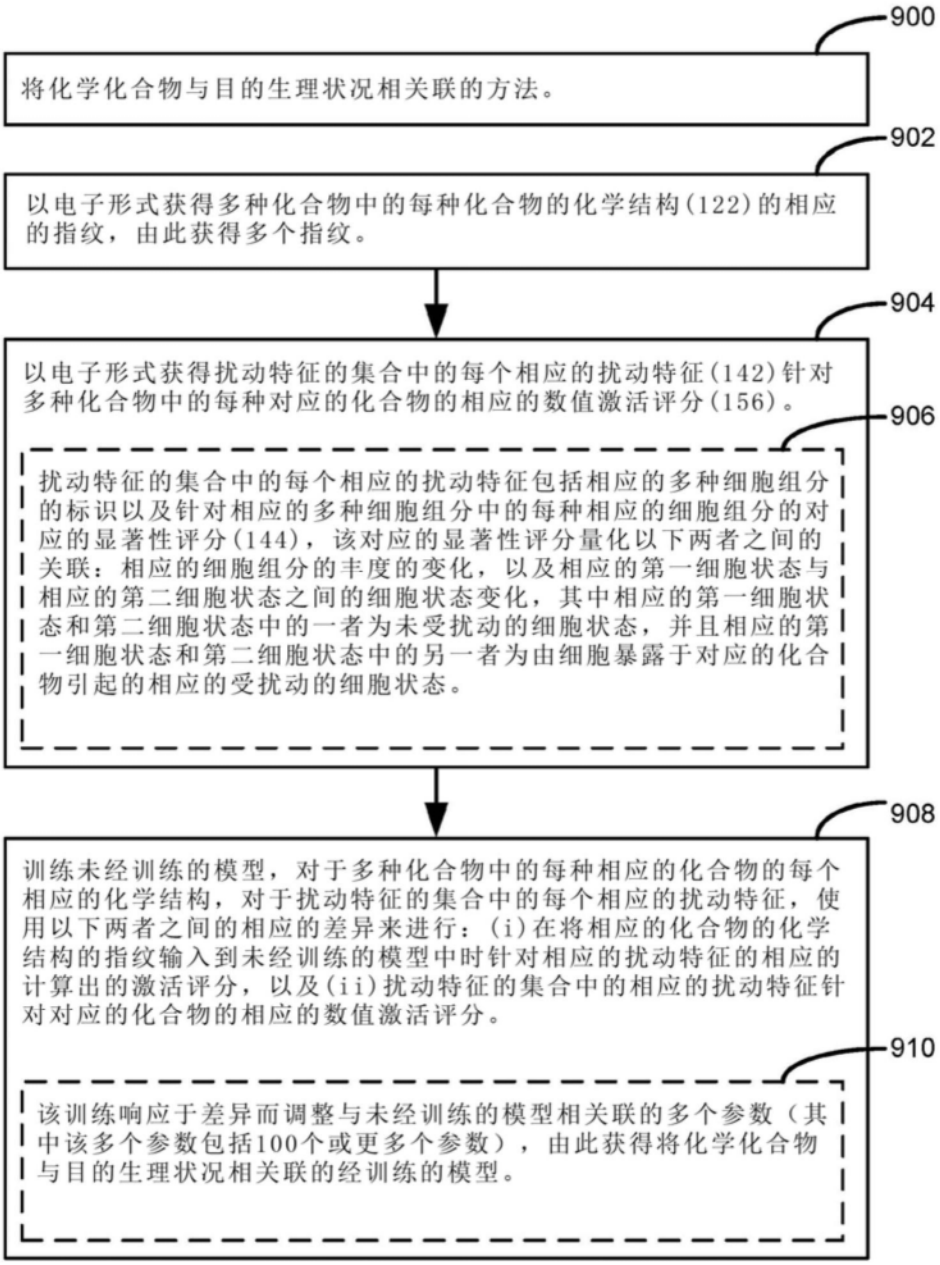

900
将化学化合物与目的生理状况相关联的方法。
902
以电子形式获得多种化合物中的每种化合物的化学结构(122)的相应的指纹，由此获得多个指纹。
904
以电子形式获得扰动特征的集合中的每个相应的扰动特征(142)针对多种化合物中的每种对应的化合物的相应的数值激活评分(156)。
906
扰动特征的集合中的每个相应的扰动特征包括相应的多种细胞组分的标识以及针对相应的多种细胞组分中的每种相应的细胞组分的对应的显著性评分(144)，该对应的显著性评分量化以下两者之间的关联：相应的细胞组分的丰度的变化，以及相应的第一细胞状态与相应的第二细胞状态之间的细胞状态变化，其中相应的第一细胞状态和第二细胞状态中的一者为未受扰动的细胞状态，并且相应的第一细胞状态和第二细胞状态中的另一者为由细胞暴露于对应的化合物引起的相应的受扰动的细胞状态。
908
训练未经训练的模型，对于多种化合物中的每种相应的化合物的每个相应的化学结构，对于扰动特征的集合中的每个相应的扰动特征，使用以下两者之间的相应的差异来进行：(i)在将相应的化合物的化学结构的指纹输入到未经训练的模型中时针对相应的扰动特征的相应的计算出的激活评分，以及(ii)扰动特征的集合中的相应的扰动特征针对对应的化合物的相应的数值激活评分。
910
该训练响应于差异而调整与未经训练的模型相关联的多个参数（其中该多个参数包括100个或更多个参数），由此获得将化学化合物与目的生理状况相关联的经训练的模型。

图9

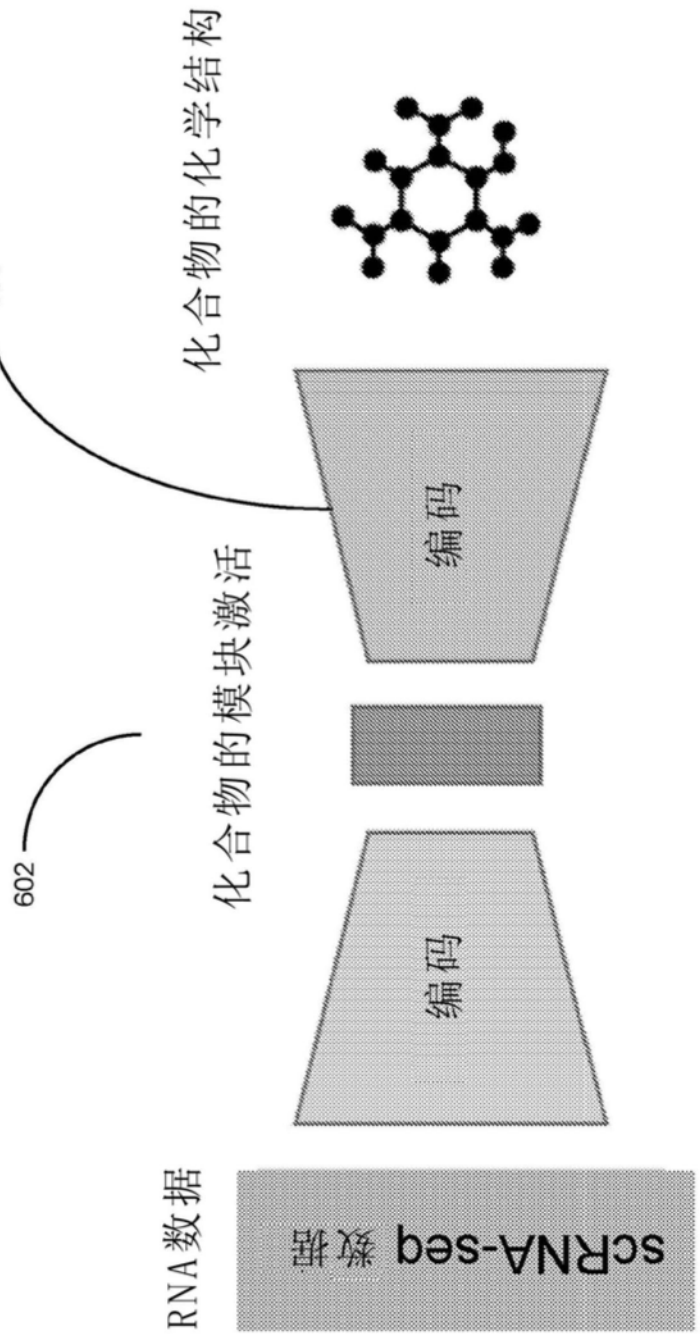


图10A

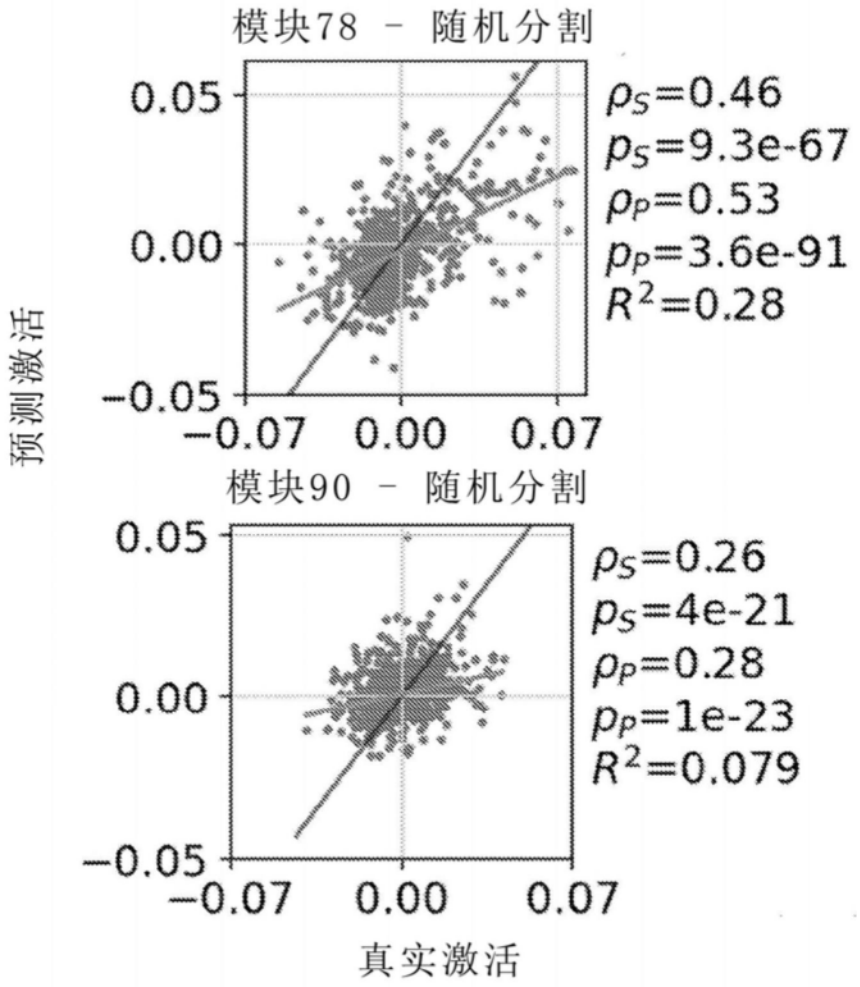

模块78 - 随机分割
0.05
0.00
−0.05
−0.07
0.00
0.07
ρ_S=0.46
p_S=9.3e-67
ρ_P=0.53
p_P=3.6e-91
R^2=0.28
模块90 - 随机分割
0.05
0.00
−0.05
−0.07
0.00
0.07
ρ_S=0.26
p_S=4e-21
ρ_P=0.28
p_P=1e-23
R^2=0.079
预测激活
真实激活

图10B

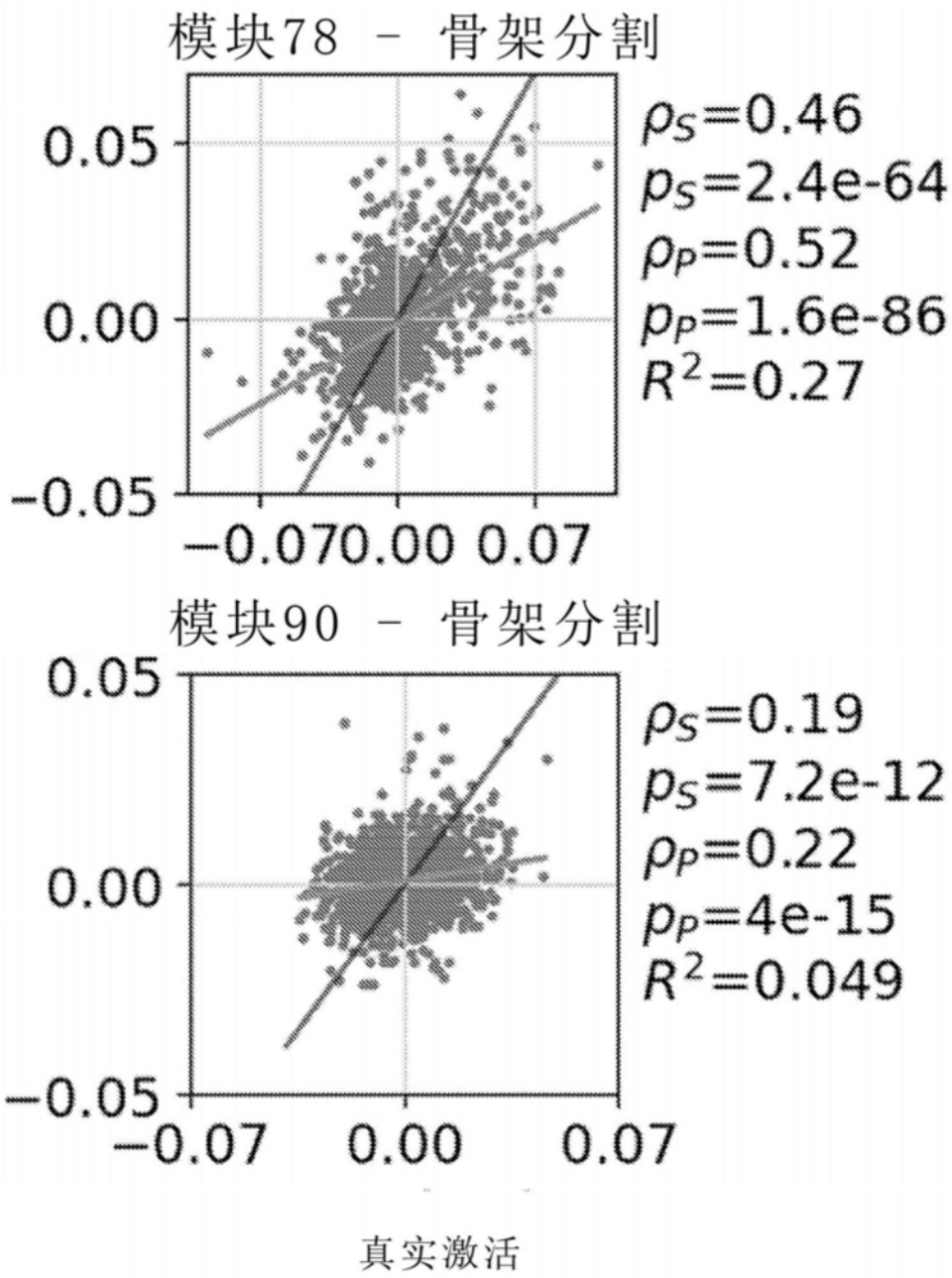

模块78 - 骨架分割
0.05
0.00
−0.05
−0.070.00 0.07
ρ_S=0.46
p_S=2.4e-64
ρ_P=0.52
p_P=1.6e-86
R^2=0.27
模块90 - 骨架分割
0.05
0.00
−0.05
−0.07 0.00 0.07
ρ_S=0.19
p_S=7.2e-12
ρ_P=0.22
p_P=4e-15
R^2=0.049
真实激活

图10C

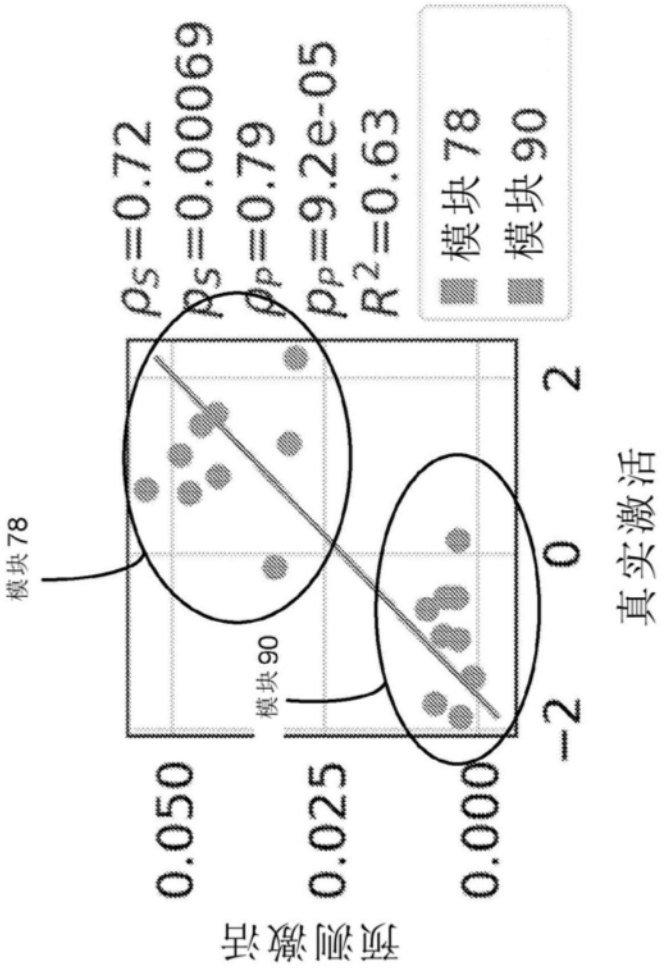

模块 78
模块 90
ρ_S=0.72
ρ_S=0.00069
ρ_P=0.79
ρ_P=9.2e-05
R^2=0.63
模块 78
模块 90
0.050
0.025
0.000
预测激活
2
0
−2
真实激活

图10D

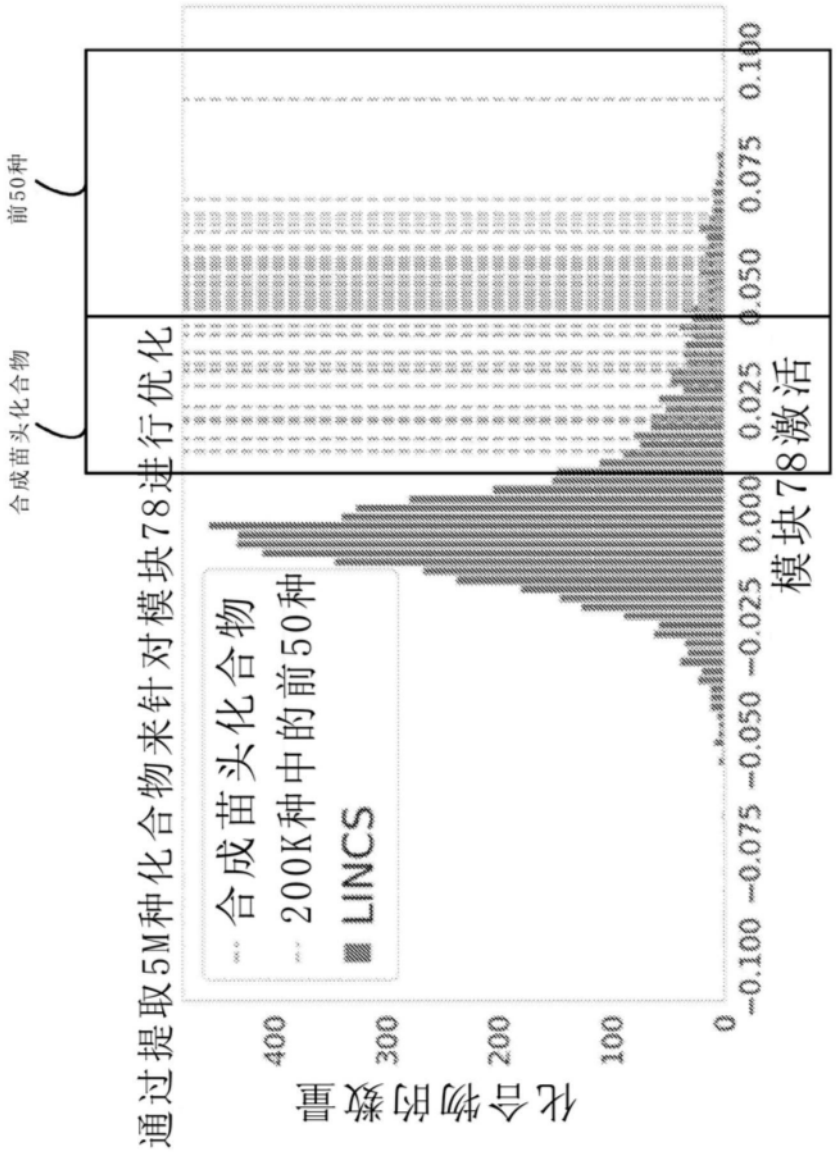

通过提取5M种化合物来针对模块78进行优化
合成苗头化合物
前50种
化合物的数量
模块78激活
合成苗头化合物
200K种中的前50种
LINCS
0
100
200
300
400
-0.100
-0.075
-0.050
-0.025
0.000
0.025
0.050
0.075
0.100

图10E

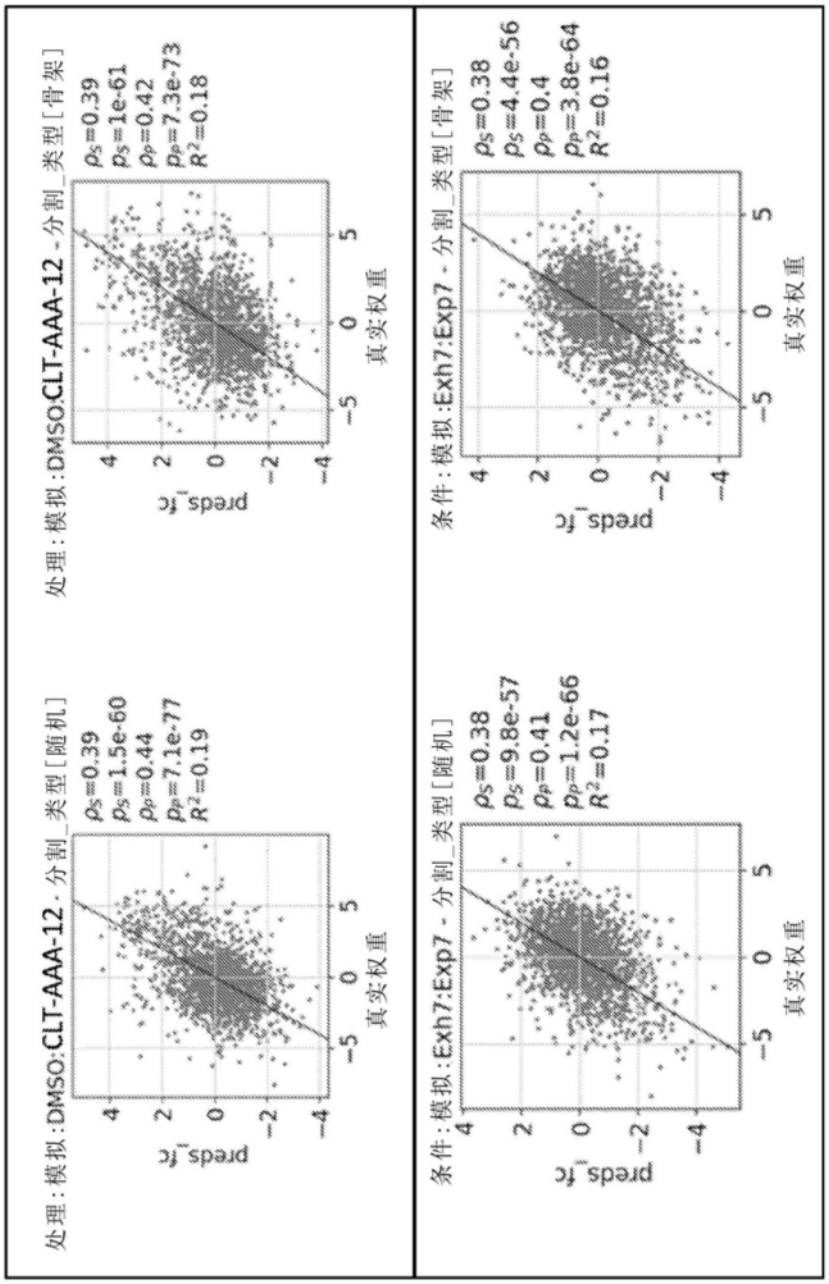

处理：模拟：DMSO:CLT-AAA-12 - 分割_类型[随机]
ρ_S=0.39
p_S=1.5e-60
ρ_P=0.44
p_P=7.1e-77
R^2=0.19
处理：模拟：DMSO:CLT-AAA-12 - 分割_类型[骨架]
ρ_S=0.39
p_S=1e-61
ρ_P=0.42
p_P=7.3e-73
R^2=0.18
条件：模拟：Exh7:Exp7 - 分割_类型[随机]
ρ_S=0.38
p_S=9.8e-57
ρ_P=0.41
p_P=1.2e-66
R^2=0.17
条件：模拟：Exh7:Exp7 - 分割_类型[骨架]
ρ_S=0.38
p_S=4.4e-56
ρ_P=0.4
p_P=3.8e-64
R^2=0.16
preds_i
真实权重

图11

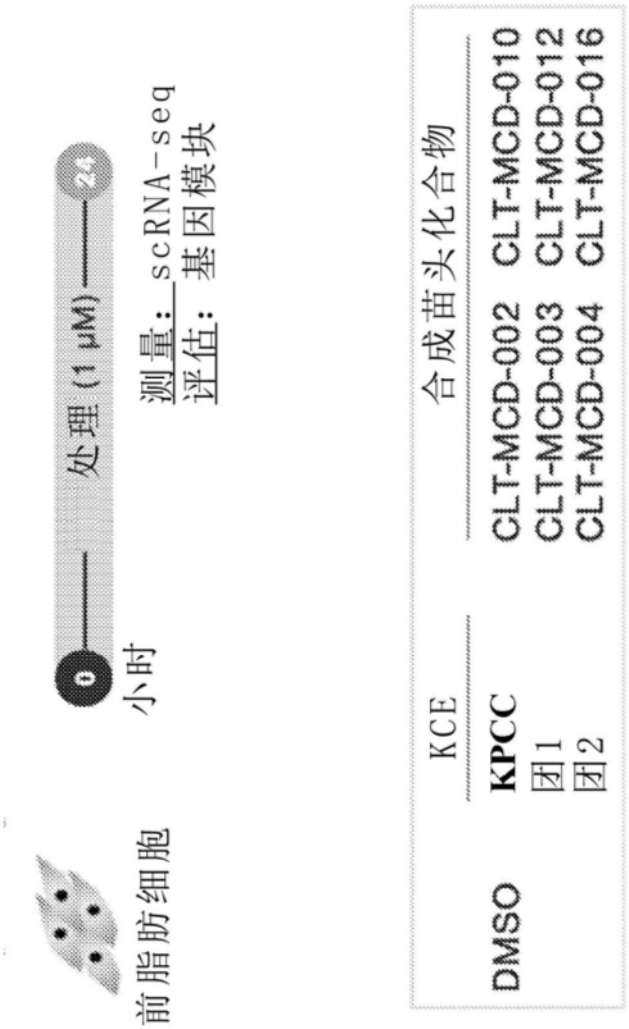

前脂肪细胞
0
处理 (1 μM)
24
小时
测量：scRNA-seq
评估：基因模块
合成苗头化合物
KCE
DMSO
KPCC
团1
团2
CLT-MCD-002
CLT-MCD-003
CLT-MCD-004
CLT-MCD-010
CLT-MCD-012
CLT-MCD-016

图12

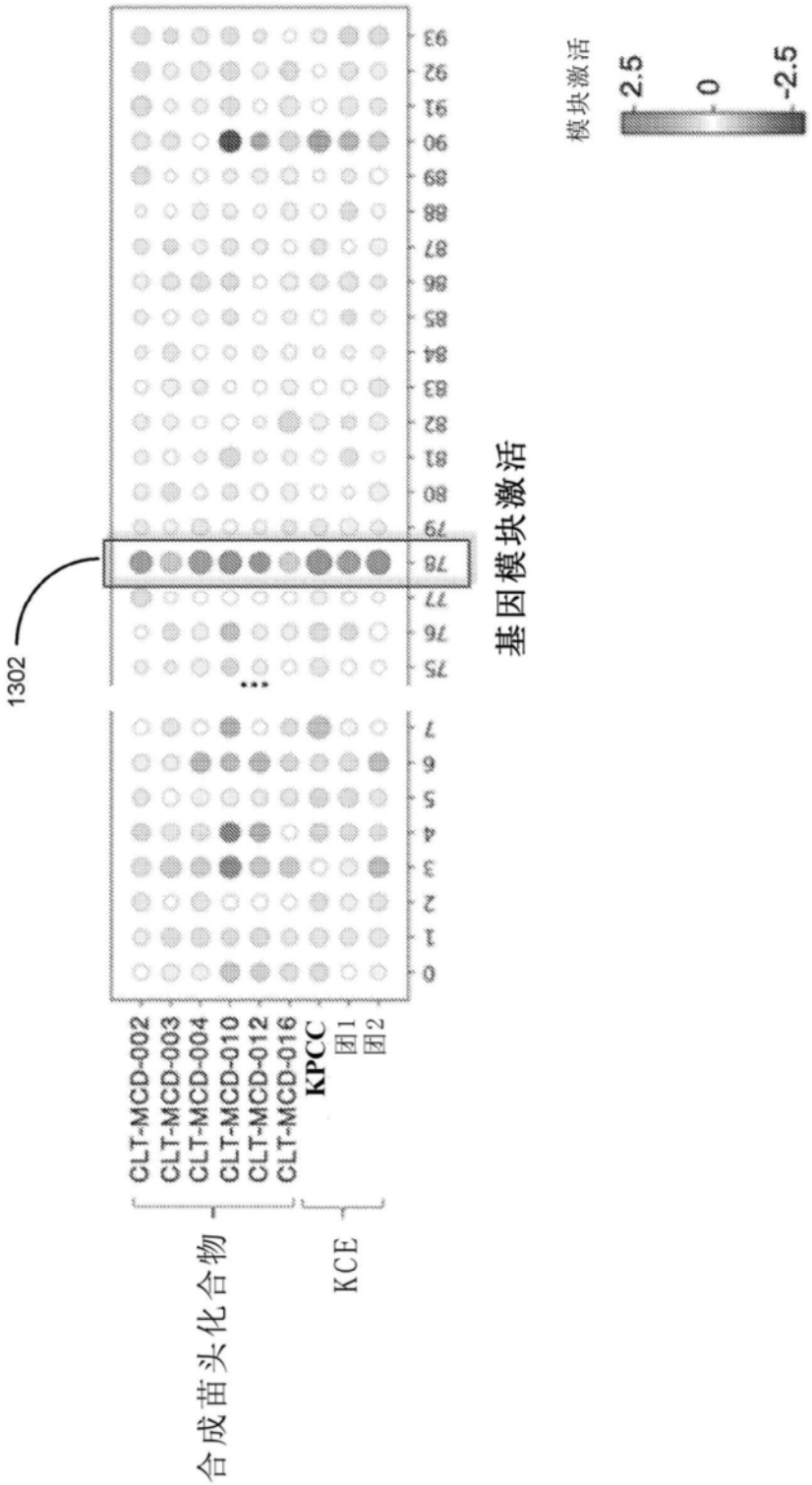

模块激活
2.5
0
-2.5
基因模块激活
1302
CLT-MCD-002
CLT-MCD-003
CLT-MCD-004
CLT-MCD-010
CLT-MCD-012
CLT-MCD-016
KPCC
团1
团2
合成苗头化合物
KCE

图13

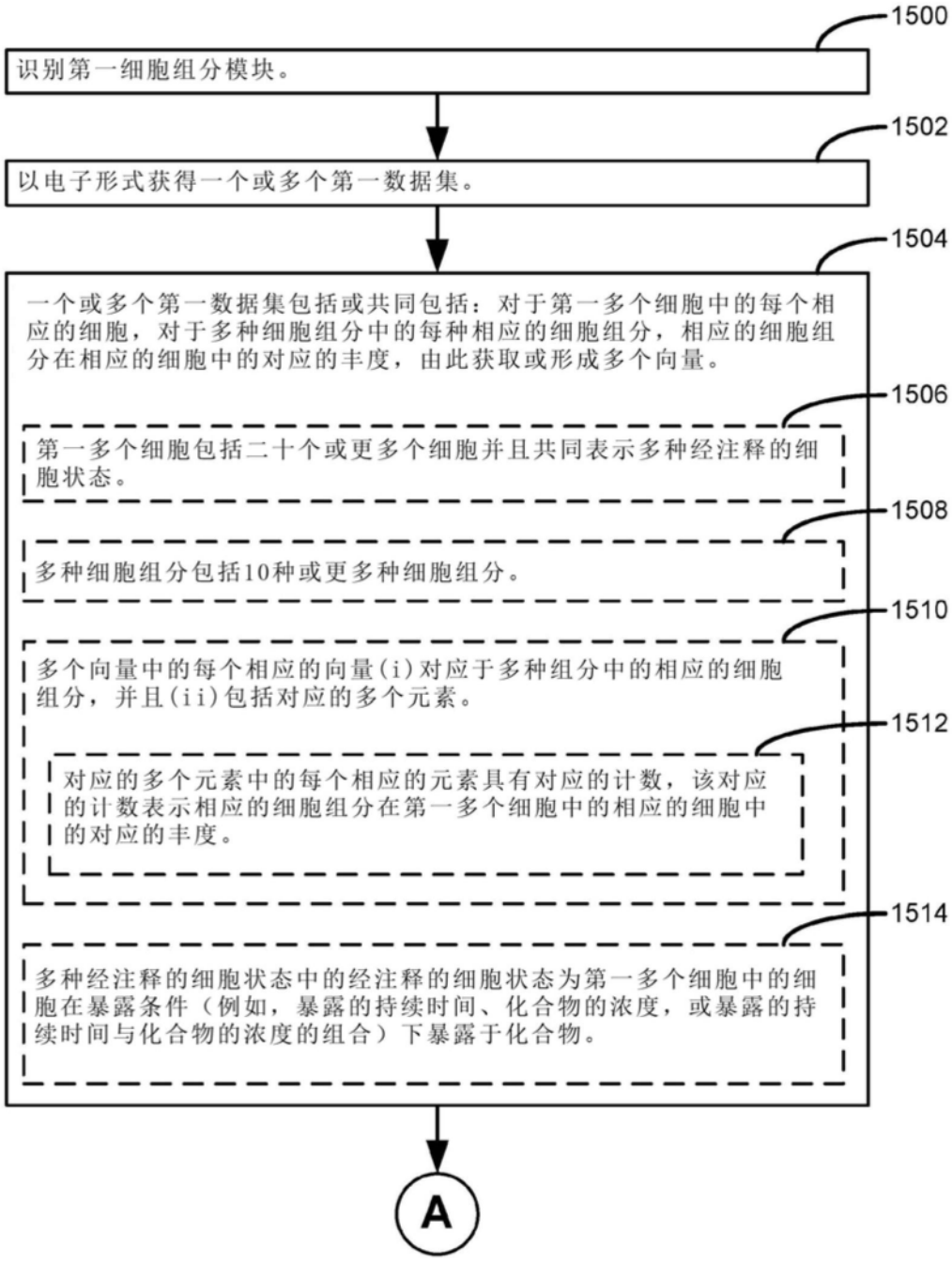

1500
识别第一细胞组分模块。
1502
以电子形式获得一个或多个第一数据集。
1504
一个或多个第一数据集包括或共同包括：对于第一多个细胞中的每个相应的细胞，对于多种细胞组分中的每种相应的细胞组分，相应的细胞组分在相应的细胞中的对应的丰度，由此获取或形成多个向量。
1506
第一多个细胞包括二十个或更多个细胞并且共同表示多种经注释的细胞状态。
1508
多种细胞组分包括10种或更多种细胞组分。
1510
多个向量中的每个相应的向量(i)对应于多种组分中的相应的细胞组分，并且(ii)包括对应的多个元素。
1512
对应的多个元素中的每个相应的元素具有对应的计数，该对应的计数表示相应的细胞组分在第一多个细胞中的相应的细胞中的对应的丰度。
1514
多种经注释的细胞状态中的经注释的细胞状态为第一多个细胞中的细胞在暴露条件（例如，暴露的持续时间、化合物的浓度，或暴露的持续时间与化合物的浓度的组合）下暴露于化合物。
A

图14A

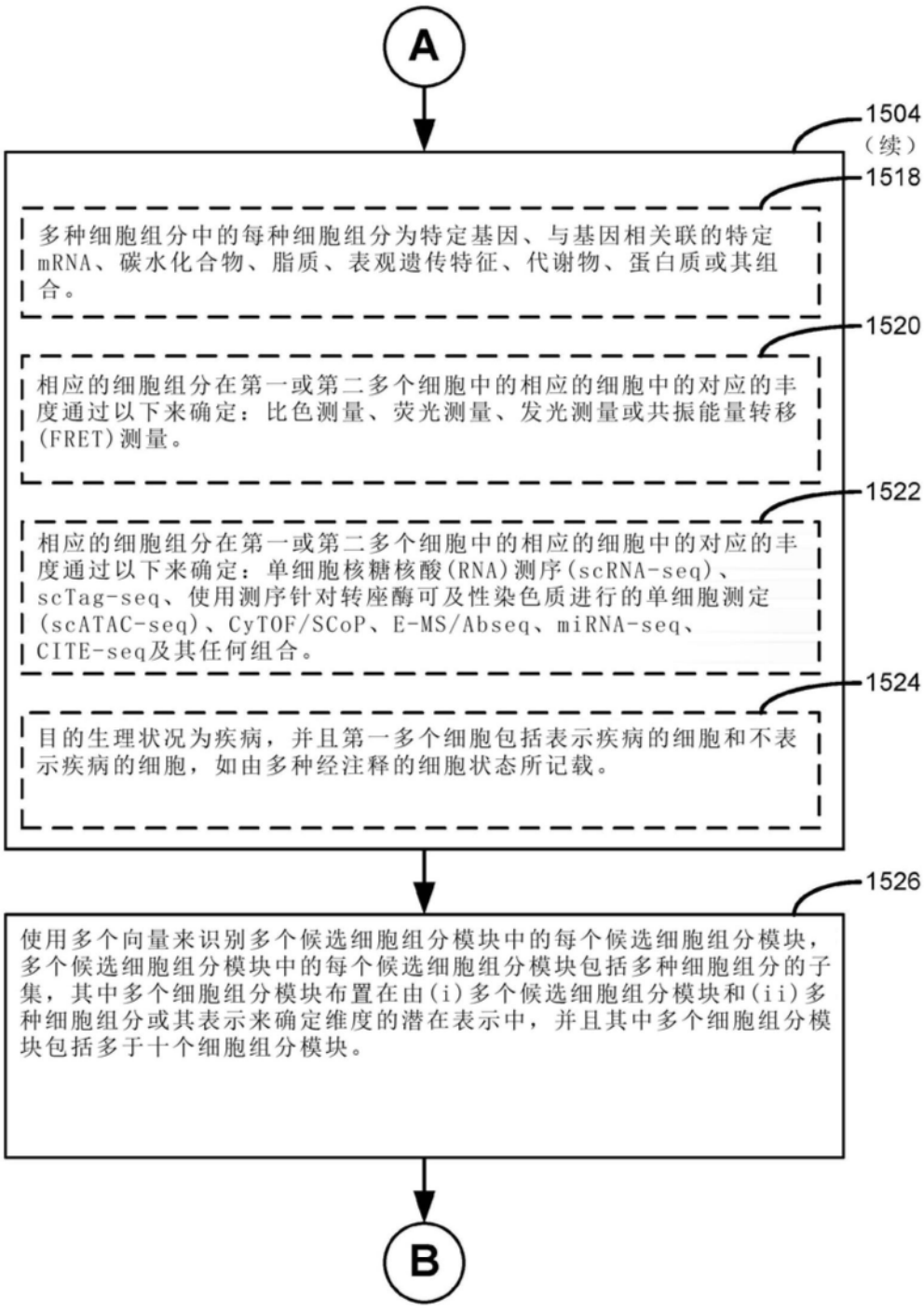

A
1504
（续）
1518
多种细胞组分中的每种细胞组分为特定基因、与基因相关联的特定mRNA、碳水化合物、脂质、表观遗传特征、代谢物、蛋白质或其组合。
1520
相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定：比色测量、荧光测量、发光测量或共振能量转移(FRET)测量。
1522
相应的细胞组分在第一或第二多个细胞中的相应的细胞中的对应的丰度通过以下来确定：单细胞核糖核酸(RNA)测序(scRNA-seq)、scTag-seq、使用测序针对转座酶可及性染色质进行的单细胞测定(scATAC-seq)、CyTOF/SCoP、E-MS/Abseq、miRNA-seq、CITE-seq及其任何组合。
1524
目的生理状况为疾病，并且第一多个细胞包括表示疾病的细胞和不表示疾病的细胞，如由多种经注释的细胞状态所记载。
1526
使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块，多个候选细胞组分模块中的每个候选细胞组分模块包括多种细胞组分的子集，其中多个细胞组分模块布置在由(i)多个候选细胞组分模块和(ii)多种细胞组分或其表示来确定维度的潜在表示中，并且其中多个细胞组分模块包括多于十个细胞组分模块。
B

图14B

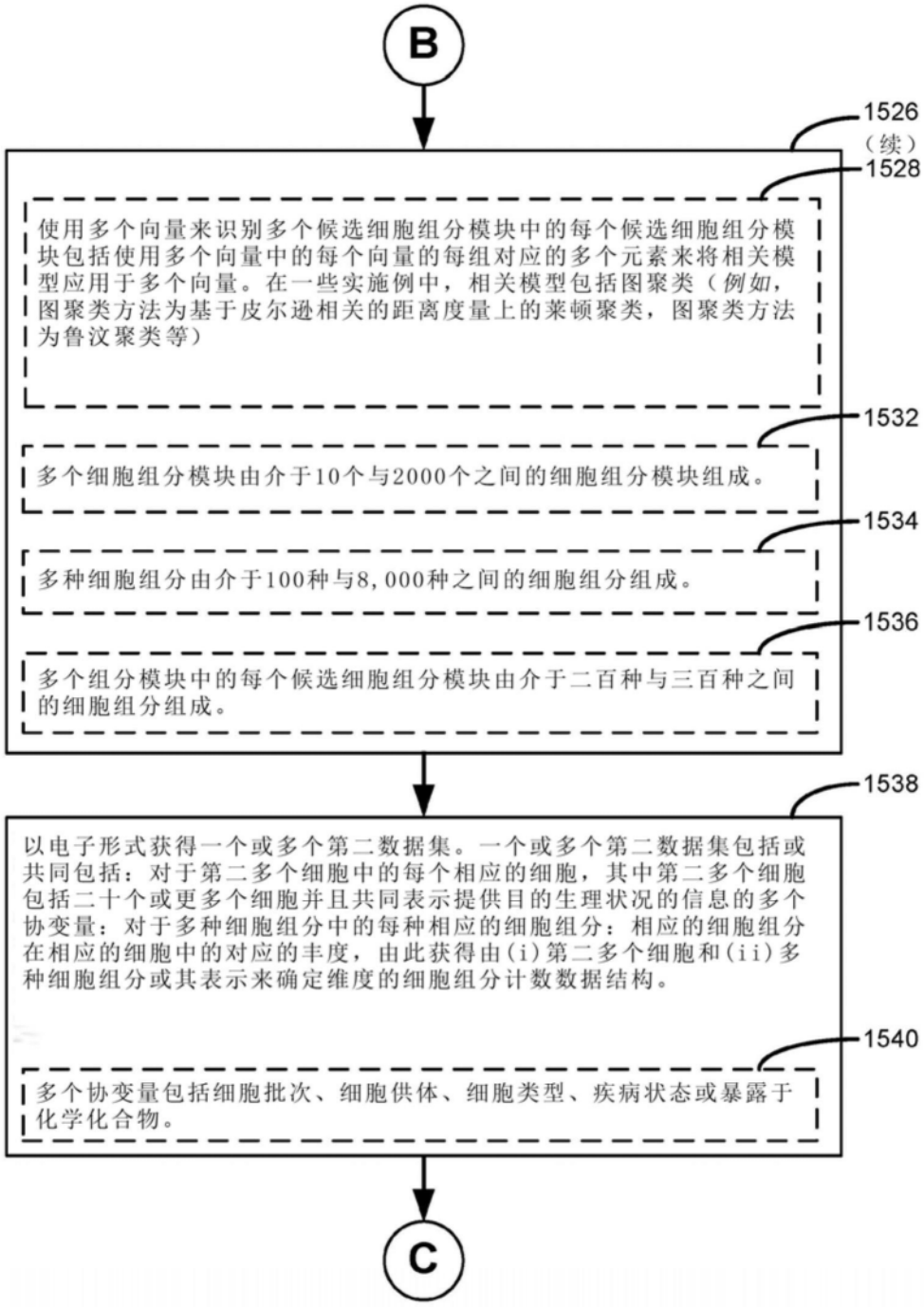

B
1526
（续）
1528
使用多个向量来识别多个候选细胞组分模块中的每个候选细胞组分模块包括使用多个向量中的每个向量的每组对应的多个元素来将相关模型应用于多个向量。在一些实施例中，相关模型包括图聚类（例如，图聚类方法为基于皮尔逊相关的距离度量上的莱顿聚类，图聚类方法为鲁汶聚类等）
1532
多个细胞组分模块由介于10个与2000个之间的细胞组分模块组成。
1534
多种细胞组分由介于100种与8,000种之间的细胞组分组成。
1536
多个组分模块中的每个候选细胞组分模块由介于二百种与三百种之间的细胞组分组成。
1538
以电子形式获得一个或多个第二数据集。一个或多个第二数据集包括或共同包括：对于第二多个细胞中的每个相应的细胞，其中第二多个细胞包括二十个或更多个细胞并且共同表示提供目的生理状况的信息的多个协变量：对于多种细胞组分中的每种相应的细胞组分：相应的细胞组分在相应的细胞中的对应的丰度，由此获得由(i)第二多个细胞和(ii)多种细胞组分或其表示来确定维度的细胞组分计数数据结构。
1540
多个协变量包括细胞批次、细胞供体、细胞类型、疾病状态或暴露于化学化合物。
C

图14C

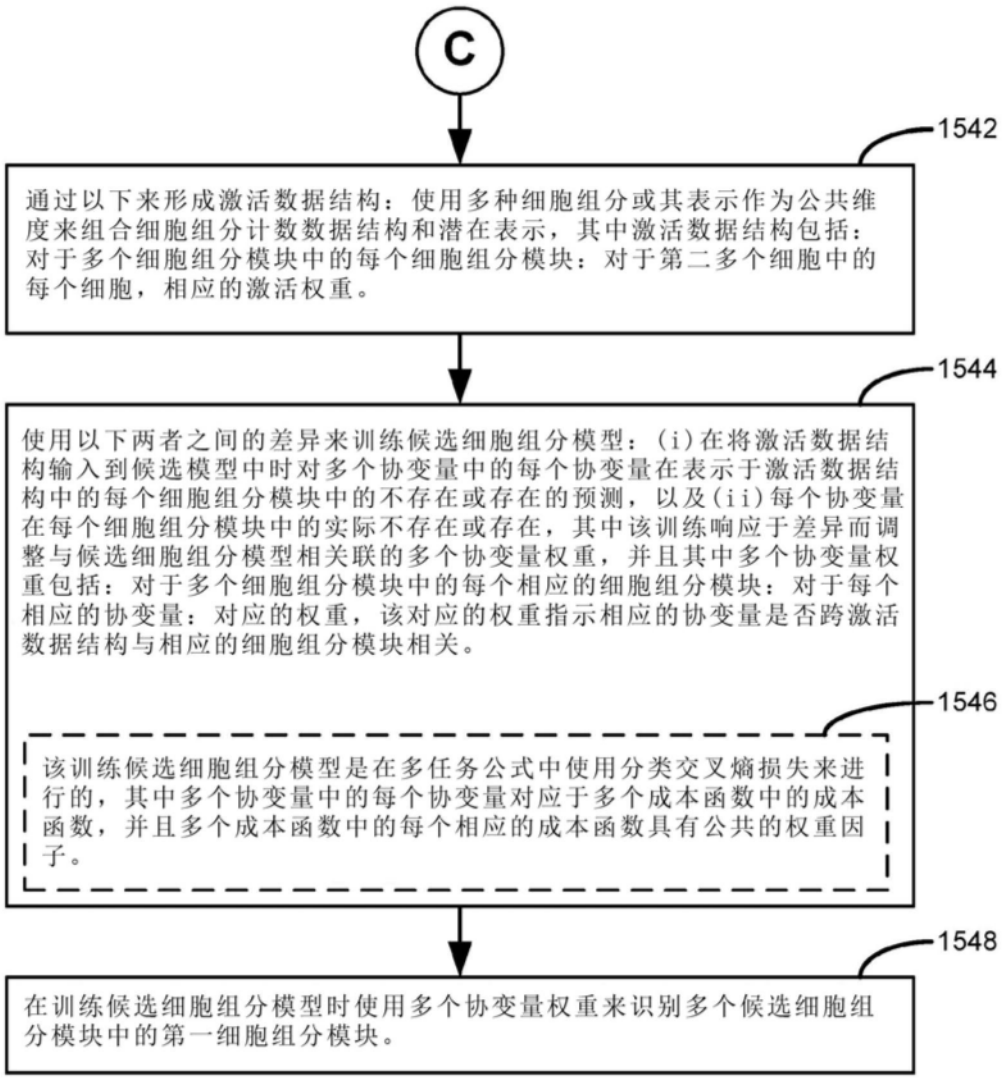

C
1542
通过以下来形成激活数据结构：使用多种细胞组分或其表示作为公共维度来组合细胞组分计数数据结构和潜在表示，其中激活数据结构包括：对于多个细胞组分模块中的每个细胞组分模块：对于第二多个细胞中的每个细胞，相应的激活权重。
1544
使用以下两者之间的差异来训练候选细胞组分模型：(i)在将激活数据结构输入到候选模型中时对多个协变量中的每个协变量在表示于激活数据结构中的每个细胞组分模块中的不存在或存在的预测，以及(ii)每个协变量在每个细胞组分模块中的实际不存在或存在，其中该训练响应于差异而调整与候选细胞组分模型相关联的多个协变量权重，并且其中多个协变量权重包括：对于多个细胞组分模块中的每个相应的细胞组分模块：对于每个相应的协变量：对应的权重，该对应的权重指示相应的协变量是否跨激活数据结构与相应的细胞组分模块相关。
1546
该训练候选细胞组分模型是在多任务公式中使用分类交叉熵损失来进行的，其中多个协变量中的每个协变量对应于多个成本函数中的成本函数，并且多个成本函数中的每个相应的成本函数具有公共的权重因子。
1548
在训练候选细胞组分模型时使用多个协变量权重来识别多个候选细胞组分模块中的第一细胞组分模块。

图14D