



(10) 申请公布号 CN 120226084 A

(43) 申请公布日 2025. 06. 27

(21) 申请号 202380081905.3

(22) 申请日 2023.10.11

(30) 优先权数据

63/415,117 2022.10.11 US

63/479,653 2023.01.12 US

(85) PCT国际申请进入国家阶段日

2025.05.27

(86) PCT国际申请的申请数据

PCT/EP2023/078227 2023.10.11

(87) PCT国际申请的公布数据

W02024/079204 EN 2024.04.18

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 Z·埃维塞克 G·诺瓦蒂 J·程

(74) 专利代理机构 北京市柳沈律师事务所

11105

专利代理师 金玉洁

(51) Int.Cl.

G16B 15/20 (2006.01)

G16B 20/50 (2006.01)

G16B 30/10 (2006.01)

G16B 40/20 (2006.01)

权利要求书4页 说明书33页 附图18页

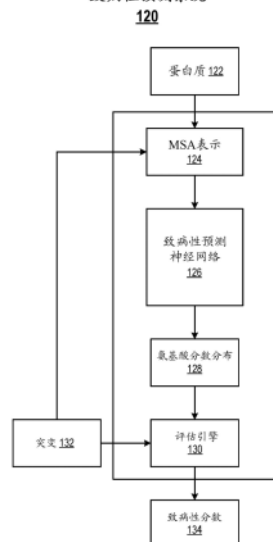
(54) 发明名称

使用氨基酸分数分布进行蛋白质突变的致病性预测

(57) 摘要

公开了用于生成致病性分数的方法、系统和装置,其包括在计算机存储介质上编码的计算机程序,该致病性分数表征蛋白质突变是致病性突变的可能性,其中该突变通过在蛋白质的氨基酸序列中的突变位置处用替代氨基酸替换原始氨基酸来修改蛋白质的氨基酸序列。在一个方面,一种方法包括:生成致病性预测神经网络的网络输入,其中该网络输入包括表示蛋白质的多序列比对(MSA)的MSA表示;使用致病性预测神经网络处理网络输入以生成氨基酸集上的分数分布;以及使用该氨基酸集上的分数分布来生成致病性分数。

致病性预测系统



1. 一种由一个或多个计算机执行的方法,所述方法包括:
生成致病性分数,所述致病性分数表征蛋白质突变是致病性突变的可能性,
其中所述突变通过在所述蛋白质的氨基酸序列中的突变位置处用替代氨基酸替换原始氨基酸来修改所述蛋白质的所述氨基酸序列,
其中生成所述致病性分数包括:
生成致病性预测神经网络的网络输入,其中所述网络输入包括表示所述蛋白质的多序列比对MSA的MSA表示;
使用所述致病性预测神经网络处理所述网络输入以生成氨基酸集上的分数分布;以及
基于以下两项之间的差值生成所述致病性分数:(i)在所述分数分布下所述原始氨基酸的分数,以及(ii)在所述分数分布下所述替代氨基酸的分数。
2. 根据权利要求1所述的方法,其中所述MSA表示包括与所述MSA中每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入;并且
其中生成所述MSA表示包括:
掩蔽所述MSA表示中与所述蛋白质的所述氨基酸序列中的所述突变位置相对应的嵌入。根据权利要求2所述的方法,其中使用所述致病性预测神经网络处理所述网络输入以生成所述氨基酸集上的所述分数分布包括:
使用所述致病性预测神经网络的嵌入子网络处理所述MSA表示以生成经更新MSA表示,其中所述经更新MSA表示包括与所述MSA中每个蛋白质的所述氨基酸序列中的每个位置相对应的相应经更新嵌入;并且
使用所述致病性预测神经网络的投影子网络处理与所述蛋白质的所述氨基酸序列中的所述突变位置相对应的所述经更新嵌入,以生成所述氨基酸集上的所述分数分布。
3. 根据权利要求2所述的方法,其中所述致病性预测神经网络的所述嵌入子网络包括一个或多个自注意力神经网络层。
4. 根据权利要求3所述的方法,其中所述致病性预测神经网络的所述嵌入子网络包括一个或多个按行或按列的自注意力神经网络层。
5. 根据任一项前述权利要求所述的方法,其中所述致病性预测神经网络已经通过包括以下各项的操作进行训练:
训练所述致病性预测神经网络以执行蛋白质去掩蔽任务;以及
训练所述致病性预测神经网络以执行致病性预测任务。
6. 根据权利要求5所述的方法,其中训练所述致病性预测神经网络以执行所述蛋白质去掩蔽任务包括,针对多个训练蛋白质中的每个蛋白质:
基于所述训练蛋白质生成所述致病性预测神经网络的网络输入,其中:
所述网络输入包括表示所述训练蛋白质的MSA的MSA表示;并且
所述MSA表示包括与所述训练蛋白质的氨基酸序列中的一个或多个经掩蔽位置中的每个经掩蔽位置相对应的相应经掩蔽嵌入;
处理所述致病性预测神经网络的所述网络输入以生成网络输出,所述网络输出针对所述训练蛋白质的所述氨基酸序列中的每个经掩蔽位置定义针对位于所述训练蛋白质的所述氨基酸序列中的所述经掩蔽位置处的氨基酸身份的相应预测;以及
通过所述致病性预测神经网络反向传播掩蔽损失的梯度,其中所述掩蔽损失衡量由所

述致病性预测神经网络针对每个经掩蔽位置生成的相应预测的准确性。

7. 根据权利要求5至6中任一项所述的方法,进一步包括:

训练所述致病性预测神经网络以执行蛋白质结构预测任务。

8. 根据权利要求7所述的方法,其中训练所述致病性预测神经网络以执行所述蛋白质结构预测任务包括,针对多个训练蛋白质中的每个训练蛋白质:

基于所述训练蛋白质生成所述致病性预测神经网络的网络输入,其中所述网络输入包括表示所述训练蛋白质的MSA的MSA表示;

使用所述致病性预测神经网络处理所述网络输入以生成所述致病性预测神经网络的中间输出;

使用折叠神经网络处理所述致病性预测神经网络的所述中间输出以生成定义所述训练蛋白质的经预测结构的结构参数集;以及

将结构损失的梯度通过所述折叠神经网络反向传播到所述致病性预测神经网络中,其中所述结构损失衡量所述训练蛋白质的所述经预测结构的误差。

9. 根据权利要求5至8中任一项所述的方法,其中所述致病性预测神经网络在被训练以执行所述致病性预测任务之前被预训练以执行所述蛋白质去掩蔽任务和所述蛋白质结构预测任务。

10. 根据任一项前述权利要求所述的方法,其中所述致病性预测神经网络已经在包括第一训练阶段和第二训练阶段的至少两个训练阶段上进行训练;

其中所述第一训练阶段包括:

在致病性训练示例集上训练所述致病性预测神经网络,其中每个致病性训练示例定义:(i) 训练蛋白质的氨基酸序列,(ii) 所述训练蛋白质的所述氨基酸序列的一个或多个突变,以及(iii) 所述一个或多个突变中的每个突变的相应目标致病性分数;

确定每个训练示例的相应过滤分数;以及

使用所述过滤分数对所述致病性训练示例集进行过滤。

11. 根据权利要求10所述的方法,其中针对每个训练示例,确定所述训练示例的所述过滤分数包括:

针对所述训练示例中包括的每个突变,使用所述致病性预测神经网络确定所述突变的经预测致病性分数;以及

针对所述训练示例中包括的每个突变,基于以下两项之间的误差确定所述突变的误差分数:(i) 所述突变的所述经预测致病性分数,以及(ii) 所述突变的所述目标致病性分数;以及

至少部分地基于所述训练示例中所包括的所述突变的所述误差分数来确定所述训练示例的所述过滤分数。

12. 根据权利要求10或11中任一项所述的方法,其中使用所述过滤分数对所述致病性训练示例集进行过滤包括:

基于所述过滤分数确定所述致病性训练示例集上的概率分布;

根据所述概率分布从所述致病性训练示例集中采样所述致病性训练示例的子集;以及从所述致病性训练示例集中去除所述致病性训练示例集的所采样子集。

13. 根据权利要求10至12中任一项所述的方法,其中所述第二训练阶段包括:

在经过滤的致病性训练示例集上再训练所述致病性预测神经网络。

14. 根据任一项前述权利要求所述的方法,进一步包括:

确定所述突变的所述致病性分数是否满足阈值;以及
响应于确定所述突变的所述致病性分数满足所述阈值:
将所述突变分类为致病性突变。

15. 根据权利要求14所述的方法,其中确定所述突变的所述致病性分数满足所述阈值包括:

确定所述突变的所述致病性分数超过所述阈值。

16. 根据任一项前述权利要求所述的方法,进一步包括使用所述致病性分数来识别受试者的疾病原因。

17. 根据权利要求1至16中任一项所述的方法,所述方法被用于获得用于有害生物防治的经突变蛋白质,所述方法进一步包括:

确定参考蛋白质的多个突变的所述致病性分数,每个突变定义相应的经突变蛋白质;
以及

基于所确定的致病性分数,选择所述突变中的一个突变来定义用于有害生物防治的所述经突变蛋白质。

18. 根据权利要求1至16中任一项所述的方法,所述方法被用于筛选一个或多个活生物体中是否存在具有致病性突变的蛋白质,所述方法进一步包括:

针对所述生物体中的每个生物体获得所述蛋白质的所述氨基酸序列;

针对所获得的氨基酸序列中包括突变的每个氨基酸序列,确定所述突变的所述致病性分数;以及

处理每个所确定的致病性分数以确定所述突变是否具有致病性。

19. 根据权利要求1至16中任一项所述的方法,所述方法被用于确定细菌或病毒对活生物体的致病程度,所述方法进一步包括:

获得人造蛋白质的氨基酸序列,其中所述人造蛋白质是由于所述活生物体被所述细菌或病毒感染而在所述活生物体内部制成的,并且其中所述人造蛋白质是所述生物体中的天然存在的蛋白质的突变;

确定所述天然存在的蛋白质的所述突变的所述致病性分数;

根据所述致病性分数确定所述细菌或病毒的致病程度。

20. 根据任一项前述权利要求所述的方法,进一步包括提供所述致病性分数以显示在用户界面上。

21. 一种系统,包括:

一个或多个计算机;以及

一个或多个存储装置,所述一个或多个存储装置通信地耦合到所述一个或多个计算机,其中所述一个或多个存储装置存储指令,所述指令在由所述一个或多个计算机执行时,致使所述一个或多个计算机执行如权利要求1至20中任一项所述的相应方法的操作。

22. 一个或多个非暂时性计算机存储介质,所述计算机存储介质存储指令,所述指令在由一个或多个计算机执行时致使所述一个或多个计算机执行如权利要求1至20中任一项所述的相应方法的操作。

23. 一种方法, 包括:

维护突变-致病性数据库, 所述突变-致病性数据库包括针对多个蛋白质中的每个蛋白质定义所述蛋白质的一个或多个突变中的每个突变的相应致病性分数的数据, 其中所述致病性分数已经使用权利要求1至20中任一项所述的方法生成;

从受试者获得遗传物质;

基于来自所述受试者的所述遗传物质, 识别所述受试者的所述遗传物质中编码的各自具有一个或多个突变的一个或多个蛋白质;

基于以下各项确定针对所述受试者的诊断: (i) 所述突变-致病性数据库, 以及 (ii) 从所述受试者的所述遗传物质中识别的所述蛋白质突变。

24. 根据权利要求23所述的方法, 其中所述突变-致病性数据库包括表征人类蛋白质组中的多个蛋白质的数据。

25. 根据权利要求24所述的方法, 其中所述突变-致病性数据库包括表征所述人类蛋白质组中的所有蛋白质的数据。

26. 根据权利要求23至25中任一项所述的方法, 其中所述受试者是人类受试者。

27. 根据权利要求23至26中任一项所述的方法, 其中确定针对所述受试者的所述诊断包括:

确定从所述受试者的所述遗传物质中识别的特定蛋白质突变在所述突变-致病性数据库中满足阈值的致病性分数相关联; 以及

作为响应, 诊断所述受试者患有与所述特定蛋白质突变相关联的医学病症。

使用氨基酸分数分布进行蛋白质突变的致病性预测

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年10月11日提交的美国临时申请号63/415,117和于2023年1月12日提交的美国临时申请号63/479,653的优先权。先前申请的公开内容被视为本申请公开内容的一部分并且通过引用并入本申请公开内容中。

背景技术

[0003] 本说明书涉及预测突变蛋白质的致病性。

[0004] 蛋白质由一个或多个氨基酸序列(“链”)指定。氨基酸是一种有机化合物,其包括氨基官能团和羧基官能团以及该氨基酸特有的侧链(即,原子团)。蛋白质折叠是指用来将一个或多个氨基酸序列折叠成三维(3-D)构型的物理过程。蛋白质的结构定义了蛋白质经过蛋白质折叠之后蛋白质氨基酸序列中的原子的3-D构型。当氨基酸以肽键链接成序列时,它们可以被称为氨基酸残基。

[0005] 可以使用机器学习模型进行预测。机器学习模型接收输入并基于接收到的输入生成输出,例如经预测的输出。一些机器学习模型是参数模型,并且基于接收到的输入和模型参数的值生成输出。一些机器学习模型是深度模型,深度模型采用多层模型来为接收到的输入生成输出。例如,深度神经网络是一种深度机器学习模型,该深度机器学习模型包括输出层和一个或多个隐藏层,一个或多个隐藏层各自将非线性变换应用于接收到的输入以生成输出。

发明内容

[0006] 本说明书描述了一种在一个或多个位置的一个或多个计算机上实现为计算机程序的神经网络系统,该神经网络系统用于预测突变蛋白质的致病性。

[0007] 如整个本说明书中所使用的,术语“蛋白质”可以被理解是指由一个或多个氨基酸序列(或“链”)指定的任何生物分子。例如,术语“蛋白质”可以指蛋白质结构域,例如,蛋白质的氨基酸链的可以几乎独立于蛋白质的其余部分进行蛋白质折叠的一部分。作为另一个示例,术语“蛋白质”可以指蛋白质复合物,即,包括共同折叠成蛋白质结构的多个氨基酸链。

[0008] “突变”蛋白质可以指由基因的经突变版本(例如脱氧核糖核酸(DNA)核苷酸序列的经突变版本)编码的蛋白质。突变蛋白质可以具有与“原始”蛋白质(即,由基因的非突变版本编码的蛋白质)不同的氨基酸序列。原始蛋白质可以是任何参考蛋白质,即,被选择作为参考的蛋白质;例如它可以但不必由该蛋白质的野生型版本定义。

[0009] 基因可能以多种可能的方式中的任一种方式发生突变。例如,基因可以包括“错义突变”。错义突变可以指点突变,其中单个DNA核苷酸的改变产生编码不同氨基酸的密码子。

[0010] 由于突变蛋白质具有与原始(未突变的)蛋白质不同的氨基酸序列,因此突变蛋白质可以具有与原始蛋白质不同的特性。例如,突变蛋白质可以具有与原始蛋白质不同的稳定性、不同的结构或不同的结合界面。

[0011] 在一些情况下,活生物体(例如人类)中的蛋白质的一种或多种生物功能依赖于蛋白质的特性,例如蛋白质的稳定性、结构或结合界面。蛋白质的突变可能影响蛋白质的特性,并且进而可能影响蛋白质在生物体中发挥其生物功能的能力。具体地,某些蛋白质突变可能具有“致病性”,例如导致更高的患病几率,例如遗传性疾病,例如镰状细胞性贫血、囊性纤维化、癌症等。

[0012] 除非另有规定,“氨基酸集”可以指标准的二十(20)种氨基酸,例如:丙氨酸、精氨酸、天冬酰胺、天冬氨酸、半胱氨酸、谷氨酰胺、谷氨酸、甘氨酸、组氨酸、异亮氨酸、亮氨酸、赖氨酸、蛋氨酸、苯丙氨酸、脯氨酸、丝氨酸、苏氨酸、色氨酸、酪氨酸、缬氨酸,或标准二十(20)种氨基酸的任何适当的子集或超集,例如蛋白氨基酸。

[0013] 氨基酸集上的“分数分布”或“氨基酸分数分布”可以指定义该氨基酸集中的每种氨基酸的相应分数的数据。

[0014] 蛋白质中的某个氨基酸链的“多序列比对”(MSA)指定该氨基酸链与多个附加氨基酸链(例如来自其他蛋白质,例如同源蛋白质)的序列比对。更具体地,MSA可以定义氨基酸链中的位置与多个附加氨基酸链中的对应位置之间的对应关系。可以例如通过使用任何适当的计算序列比对技术(例如渐进式比对构建)处理氨基酸链数据库来生成氨基酸链的MSA。MSA中的氨基酸链可以被理解为具有进化关系,例如,MSA中的每个氨基酸链可以具有共同的祖先。在氨基酸链的MSA中,氨基酸链中的氨基酸之间的相关性可以编码与预测氨基酸链的结构相关的信息。

[0015] 实体(例如氨基酸对)的“嵌入(embedding)”可以指将实体表示为数值的有序集合,例如数值的向量或矩阵。

[0016] 蛋白质的结构可以通过结构参数集来定义。定义蛋白质结构的结构参数集可以被表示为数值的有序集合。接下来将更详细地描述用于定义蛋白质结构的可能结构参数的几个示例。

[0017] 在一个示例中,针对蛋白质中的每个氨基酸,定义蛋白质结构的结构参数包括:(i)位置参数,以及(ii)旋转参数。

[0018] 氨基酸的位置参数可以指定蛋白质结构中的氨基酸中的指定原子的经预测3-D空间位置。该指定原子可以是氨基酸中的 α 碳原子(alpha carbon atom),即,氨基酸中与氨基官能团、羧基官能团以及侧链键合的碳原子。氨基酸的位置参数可以用任何适当的坐标系(例如三维 $[x, y, z]$ 笛卡尔坐标系)来表示。

[0019] 氨基酸的旋转参数可以指定蛋白质结构中的氨基酸的经预测“取向”。更具体地,旋转参数可以指定3-D空间旋转操作,如果将该空间旋转操作应用于位置参数的坐标系,则导致氨基酸中的三个“主链”原子相对于旋转后的坐标系处于固定位置。氨基酸中的这三个主链原子可以指氨基酸中所链接的氮原子、 α 碳原子和羰基碳原子的序列。氨基酸的旋转参数可以被表示为例如行列式等于1的正交 3×3 矩阵。

[0020] 一般来说,氨基酸的位置参数和旋转参数定义了该氨基酸的以自我为中心的参考系。在这个参考系中,每个氨基酸的侧链可以从原点开始,并且沿着侧链的第一个键(即, α 碳- β 碳键)可以沿着所定义的方向。

[0021] 在另一个示例中,定义蛋白质结构的结构参数可以包括“距离图”,该距离图表征蛋白质中的每个氨基酸对之间的相应估计距离(例如以埃为单位来测量)。距离图可以例如

通过氨基酸对之间的可能距离集上的概率分布来表征氨基酸对之间的估计距离。

[0022] 在另一个示例中,定义蛋白质结构的结构参数可以定义蛋白质结构中的每个氨基酸中的每个原子的三维(3D)空间位置。

[0023] 根据第一方面,提供了一种由一个或多个计算机执行的方法,该方法包括:生成致病性分数,该致病性分数表征蛋白质突变是致病性突变的可能性,其中该突变通过在该蛋白质的氨基酸序列中的突变位置处用替代氨基酸替换原始氨基酸来修改该蛋白质的该氨基酸序列,其中生成该致病性分数包括:生成致病性预测神经网络的网络输入,其中该网络输入包括表示该蛋白质的多序列比对(MSA)的MSA表示;使用该致病性预测神经网络处理该网络输入以生成氨基酸集上的分数分布;以及基于以下两项之间的差值生成该致病性分数:(i)在该分数分布下该原始氨基酸的分数,以及(ii)在该分数分布下该替代氨基酸的分数。

[0024] 在一些实现方式中,该MSA表示包括与该MSA中每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入;并且生成该MSA表示包括:掩蔽该MSA表示中与该蛋白质的该氨基酸序列中的该突变位置相对应的嵌入。

[0025] 在一些实现方式中,使用该致病性预测神经网络处理该网络输入以生成该氨基酸集上的该分数分布包括:使用该致病性预测神经网络的嵌入子网络处理该MSA表示以生成经更新MSA表示,其中该经更新MSA表示包括与该MSA中每个蛋白质的该氨基酸序列中的每个位置相对应的相应经更新嵌入;并且使用该致病性预测神经网络的投影子网络处理与该蛋白质的该氨基酸序列中的该突变位置相对应的该经更新嵌入,以生成该氨基酸集上的该分数分布。

[0026] 在一些实现方式中,该致病性预测神经网络的该嵌入子网络包括一个或多个自注意力神经网络层。

[0027] 在一些实现方式中,该致病性预测神经网络的该嵌入子网络包括一个或多个按行(row-wise)或按列(column-wise)自注意力神经网络层。

[0028] 在一些实现方式中,该致病性预测神经网络已经通过包括以下各项的操作进行了训练:训练该致病性预测神经网络以执行蛋白质去掩蔽(unmasking)任务;以及训练该致病性预测神经网络以执行致病性预测任务。

[0029] 在一些实现方式中,训练该致病性预测神经网络以执行该蛋白质去掩蔽任务包括,针对多个训练蛋白质中的每个蛋白质:基于该训练蛋白质生成该致病性预测神经网络的网络输入,其中:该网络输入包括表示该训练蛋白质的MSA的MSA表示;并且该MSA表示包括与该训练蛋白质的氨基酸序列中的一个或多个经掩蔽位置中的每个经掩蔽位置相对应的相应经掩蔽嵌入;处理该致病性预测神经网络的网络输入以生成网络输出,该网络输出针对该训练蛋白质的该氨基酸序列中的每个经掩蔽位置定义针对位于该训练蛋白质的该氨基酸序列中的该经掩蔽位置处的氨基酸身份的相应预测;以及通过该致病性预测神经网络反向传播掩蔽损失的梯度,其中该掩蔽损失衡量由该致病性预测神经网络针对每个经掩蔽位置生成的相应预测的准确性。

[0030] 在一些实现方式中,该方法进一步包括训练该致病性预测神经网络以执行蛋白质结构预测任务。

[0031] 在一些实现方式中,训练该致病性预测神经网络以执行该蛋白质结构预测任务包

括,针对多个训练蛋白质中的每个训练蛋白质:基于该训练蛋白质生成该致病性预测神经网络的网络输入,其中该网络输入包括表示该训练蛋白质的MSA的MSA表示;使用该致病性预测神经网络处理该网络输入以生成该致病性预测神经网络的中间输出;使用折叠神经网络处理该致病性预测神经网络的该中间输出以生成定义该训练蛋白质的经预测结构的结构参数集;以及将结构损失的梯度通过该折叠神经网络反向传播到该致病性预测神经网络中,其中该结构损失衡量该训练蛋白质的该经预测结构的误差。

[0032] 在一些实现方式中,该致病性预测神经网络在被训练以执行该致病性预测任务之前被预训练以执行该蛋白质去掩蔽任务和该蛋白质结构预测任务。

[0033] 在一些实现方式中,该致病性预测神经网络已经在包括第一训练阶段和第二训练阶段的至少两个训练阶段上进行训练;其中该第一训练阶段包括:在致病性训练示例集上训练该致病性预测神经网络,其中每个致病性训练示例定义:(i) 训练蛋白质的氨基酸序列,(ii) 该训练蛋白质的该氨基酸序列的一个或多个突变,以及(iii) 该一个或多个突变中的每个突变的相应目标致病性分数;确定每个训练示例的相应过滤分数;以及使用这些过滤分数对该致病性训练示例集进行过滤。

[0034] 在一些实现方式中,针对每个训练示例,确定该训练示例的该过滤分数包括:针对该训练示例中包括的每个突变,使用该致病性预测神经网络确定该突变的经预测致病性分数;以及针对该训练示例中包括的每个突变,基于以下两项之间的误差确定该突变的误差分数:(i) 该突变的该经预测致病性分数,以及(ii) 该突变的该目标致病性分数;以及至少部分地基于该训练示例中所包括的这些突变的这些误差分数来确定该训练示例的该过滤分数。

[0035] 在一些实现方式中,使用这些过滤分数对该致病性训练示例集进行过滤包括:基于这些过滤分数确定该致病性训练示例集上的概率分布;根据该概率分布从该致病性训练示例集中采样这些致病性训练示例的子集;以及从该致病性训练示例集中去除致病性训练示例的所采样子集。

[0036] 在一些实现方式中,该第二训练阶段包括:在经过滤的致病性训练示例集上再训练该致病性预测神经网络。

[0037] 在一些实现方式中,该方法进一步包括:确定该突变的该致病性分数是否满足阈值;以及响应于确定该突变的该致病性分数满足该阈值:将该突变分类为致病性突变。

[0038] 在一些实现方式中,确定该突变的该致病性分数满足该阈值包括:确定该突变的该致病性分数超过该阈值。

[0039] 在一些实现方式中,该方法进一步包括使用该致病性分数来识别受试者的疾病原因。

[0040] 在一些实现方式中,本文描述的方法用于获得用于有害生物防治的经突变蛋白质,例如包括:确定参考蛋白质的多个突变的该致病性分数,每个突变定义相应的经突变蛋白质;以及基于所确定的致病性分数,选择这些突变中的一个突变来定义用于有害生物防治的该经突变蛋白质。

[0041] 在一些实现方式中,本文描述的方法用于筛选一个或多个活生物体中是否存在具有致病性突变的蛋白质,例如包括:针对这些生物体中的每个生物体获得该蛋白质的该氨基酸序列;针对所获得的氨基酸序列中包括突变的每个氨基酸序列,确定该突变的该致病

性分数;以及处理每个所确定的致病性分数以确定该突变是否具有致病性。

[0042] 在一些实现方式中,本文描述的方法用于确定细菌或病毒对活生物体的致病程度,例如包括:获得人造蛋白质的氨基酸序列,其中该人造蛋白质是由于该活生物体被该细菌或病毒感染而在该活生物体内部制成的,并且其中该人造蛋白质是该生物体中的天然存在的蛋白质的突变;确定该天然存在的蛋白质的该突变的该致病性分数;根据该致病性分数确定该细菌或病毒的致病程度。

[0043] 在一些实现方式中,该方法进一步包括提供该致病性分数以显示在用户界面上。

[0044] 根据另一个方面,提供了一种系统,其包括:一个或多个计算机;以及通信地耦合到该一个或多个计算机的一个或多个存储装置,其中该一个或多个存储装置存储指令,这些指令当由该一个或多个计算机执行时,致使该一个或多个计算机执行本文描述的方法的操作。

[0045] 根据另一个方面,提供了一个或多个非暂时性计算机存储介质,该计算机存储介质存储指令,这些指令在由一个或多个计算机执行时致使该一个或多个计算机执行本文描述的方法的操作。

[0046] 根据另一个方面,提供了一种方法,包括:维护突变-致病性数据库,该突变-致病性数据库包括针对多个蛋白质中的每个蛋白质定义该蛋白质的一个或多个突变中的每个突变的相应致病性分数的数据,其中该致病性分数已经使用本文描述的方法生成;从受试者获得遗传物质;基于来自该受试者的该遗传物质,识别该受试者的该遗传物质中编码的各自具有一个或多个突变的一个或多个蛋白质;以及基于以下各项确定针对该受试者的诊断:(i) 该突变-致病性数据库,以及(ii) 从该受试者的该遗传物质中识别的该蛋白质突变。

[0047] 从受试者获得遗传物质可以包括例如从受试者获得生物样本,例如,血液样本、或唾液样本、或口腔拭子、或组织活检。然后可以从生物样本中提取诸如脱氧核糖核酸(DNA)等遗传物质。可以执行基因测试,例如诸如DNA测序、DNA微阵列分析、聚合酶链式反应(PCR)分析或下一代测序(NGS),以识别在受试者的遗传物质中编码的例如与参考基因组相比各有一个或多个突变的一个或多个蛋白质。

[0048] 在一些情况下,该突变-致病性数据库包括表征人类蛋白质组中的多个蛋白质的数据。

[0049] 在一些情况下,该突变-致病性数据库包括表征该人类蛋白质组中的所有蛋白质的数据。

[0050] 在一些情况下,该受试者是人类受试者。

[0051] 在一些情况下,确定针对该受试者的该诊断包括:确定从该受试者的该遗传物质中识别的特定蛋白质突变在该突变-致病性数据库中满足阈值的致病性分数相关联;以及作为响应,诊断该受试者患有与该特定蛋白质突变相关联的医学病症。该阈值可以是任何适当的阈值,例如0.5、0.8、0.9或0.99。

[0052] 可以实现本说明书中描述的主题的特定实施例,以便实现以下优势中的一个或多个优势。

[0053] 本说明书描述了一种致病性预测神经网络,该致病性预测神经网络可以生成致病性分数,这些致病性分数预测蛋白质突变是否具有致病性,例如是否可能导致疾病。为了确定蛋白质的氨基酸序列中的某个位置处的突变是否具有致病性,该致病性预测神经网络可

以处理该蛋白质的MSA以生成氨基酸集上的分数分布。然后可以基于以下两项之间的差值生成突变的致病性分数：(i) 在分数分布下该位置处的原始氨基酸的分数，以及(ii) 在分数分布下该位置处的替代氨基酸的分数。处理MSA使得致病性预测神经网络能够利用编码与突变致病性相关的信息的蛋白质进化历史(如本说明书稍后将更详细描述)。此外，将致病性分数计算为被指派给原始氨基酸和替代氨基酸的氨基酸分数之间的差值编码了一种归纳偏置，即，评估致病性涉及将原始氨基酸与替代氨基酸进行比较。

[0054] 致病性预测神经网络可以被预训练(或联合训练)以执行蛋白质结构预测的辅助任务。蛋白质的结构在确定蛋白质突变是否将具有致病性方面起着重要作用，并且训练致病性预测神经网络以执行辅助蛋白质结构预测任务可以使得致病性预测神经网络对蛋白质结构进行隐式推理。因此，训练致病性预测神经网络以执行蛋白质结构预测的辅助任务可以允许致病性预测神经网络在更少的训练数据上进行训练，或者在更少的训练迭代中进行训练，或者两者兼有，从而减少计算资源的消耗。

[0055] 致病性预测神经网络可以被预训练(或联合训练)以执行蛋白质去掩蔽的辅助任务。去掩蔽任务需要致病性预测神经网络基于氨基酸序列的剩余未掩蔽部分中的上下文信息来解码蛋白质氨基酸序列中的经掩蔽位置处的氨基酸的身份。学习有效地执行去掩蔽任务可以使得致病性预测神经网络对蛋白质生物化学进行隐式推理，并且从而可以允许致病性预测神经网络在更少的训练数据上进行训练，或者在更少的训练迭代中进行训练，或者两者兼有，从而减少计算资源的消耗。

[0056] 本说明书的主题的一个或多个实施例的细节在附图和以下描述中进行阐述。根据描述、附图和权利要求，该主题的其他特征、方面和优点将变得显而易见。

附图说明

[0057] 图1示出了示例致病性预测系统。

[0058] 图2示出了致病性预测神经网络的示例架构。

[0059] 图3示出了用于训练致病性预测神经网络的示例训练系统。

[0060] 图4是用于生成致病性分数的示例过程的流程图，该分数表征蛋白质突变是致病性突变的可能性。

[0061] 图5是用于训练致病性预测神经网络的多阶段过程的示例过程的流程图。

[0062] 图6A示出了蛋白质的MSA的MSA表示的示例。

[0063] 图6B示出了由致病性预测神经网络生成的氨基酸集上的分数分布的示例。

[0064] 图7示出了示例蛋白质结构预测系统。

[0065] 图8示出了嵌入神经网络的示例架构。

[0066] 图9示出了嵌入神经网络的更新块的示例架构。

[0067] 图10示出了MSA更新块的示例架构。

[0068] 图11示出了成对更新块的示例架构。

[0069] 图12示出了折叠神经网络的示例架构。

[0070] 图13示出了氨基酸中的键之间的扭转角。

[0071] 图14是未折叠蛋白质和经折叠蛋白质的图示。

[0072] 图15是用于预测蛋白质结构的示例过程的流程图。

- [0073] 图16示出了用于生成蛋白质中的氨基酸链的MSA表示的示例过程。
- [0074] 图17示出了为蛋白质中的每个氨基酸对生成相应成对嵌入的示例过程。
- [0075] 图18A至图18B示出了致病性预测系统的性能(例如分类准确性)。
- [0076] 在各个附图中,相同的附图标记和名称指示相同的元素。

具体实施方式

[0077] 图1示出了示例致病性预测系统120。致病性预测系统120是在一个或多个位置的一个或多个计算机上实现为计算机程序的系统的示例,其中实现下文描述的系统、组件和技术。

[0078] 致病性预测系统120被配置为处理定义以下各项的数据以生成致病性分数134:(i) 蛋白质122的氨基酸序列,以及(ii) 蛋白质122的氨基酸序列的突变132。致病性分数134表征突变132是致病性突变(例如,将导致生物体有更高患病可能性的突变)的可能性。蛋白质的氨基酸序列可以是任何参考序列,例如野生型序列。生物体可以是动物(例如人类)或植物或微生物(诸如细菌)。

[0079] 突变132可以指定:(i) 蛋白质122的氨基酸序列中的突变位置,以及(ii) 氨基酸序列中的突变位置的替代氨基酸(即,在突变下,其将替换氨基酸序列中的突变位置处的原始氨基酸)。突变132可以是例如编码蛋白质122的基因发生错义突变的结果。

[0080] 致病性分数134可以被表示为例如标量数值,例如,其中致病性分数的值越高可以表示蛋白质突变122致病的可能性越高。

[0081] 由致病性预测系统120针对蛋白质突变生成的致病性分数134可以被用于多种可能的应用中的任一种应用中。例如,致病性分数134可以被用于识别导致遗传疾病的特定基因。更具体地,可以对疑似患有遗传疾病的受试者进行基因测序,并且可以处理该受试者的基因数据以识别该受试者基因组中的经突变蛋白质编码基因集,例如通过将蛋白质的基因序列与参考基因序列进行比较。本说明书中描述的致病性预测系统120可以被用于针对由经突变蛋白质编码基因中的每个经突变蛋白质编码基因引起的蛋白质突变生成相应的致病性分数。与高致病性分数(例如,分数高于阈值或相对高于其他蛋白质的分数)相关联的经突变蛋白质编码基因然后可以被指定为受试者遗传疾病的候选原因。该系统的一些进一步应用将在稍后描述。

[0082] 致病性预测系统120包括致病性预测神经网络126和评估引擎130,接下来将分别对它们进行详细描述。

[0083] 为了生成突变132的致病性分数134,致病性预测系统120获得定义蛋白质122的多序列比对(MSA)的数据,并生成表示蛋白质122的MSA的MSA表示124。(在整个本文档中,蛋白质的MSA应理解为包括蛋白质本身;也就是说,蛋白质的MSA包括:(i) 蛋白质(蛋白质的参考版本),以及(ii) 与(参考)蛋白质同源的其他蛋白质集)。

[0084] MSA表示124包括嵌入的集合,该集合包括与MSA中包括的每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入。下文参考图7更详细地描述用于生成蛋白质122的MSA表示124的示例技术。

[0085] 可选地,致病性预测系统120可以“掩蔽(mask)”MSA表示124中与(参考)蛋白质的氨基酸序列中的突变位置(即,由突变132指定)相对应的嵌入。经掩蔽嵌入是指将嵌入替换

为默认(经掩蔽)嵌入(例如预定义嵌入(例如,在每个条目中具有预定义值,例如零))或随机嵌入(例如,嵌入中每个条目的值都是从概率分布(例如高斯分布)中采样的)。一般来说,掩蔽嵌入是指改变嵌入。

[0086] 蛋白质的MSA编码与预测蛋白质突变的致病性相关的信息。例如,对于蛋白质的氨基酸序列中的每个位置,MSA编码指示在蛋白质进化历史过程中该位置处的氨基酸已经改变为不同氨基酸的频率的信息。如果蛋白质氨基酸序列中某个位置处的氨基酸在该蛋白质的进化史上已经频繁发生突变,那么该位置处的突变可能不太容易致病。反之,如果蛋白质氨基酸序列中某个位置处的氨基酸在该蛋白质的进化史上很少发生突变,那么该位置处的突变可能更容易致病。直观地说,如果蛋白质氨基酸序列中某个位置处的突变导致致病性,那么进化压力可能导致该位置发生突变的氨基酸序列在蛋白质的进化史中出现的频率降低。

[0087] 致病性预测系统120使用致病性预测神经网络126处理包括MSA表示124的网络输入,以生成氨基酸集上的分数分布128。下文将参考图2更详细地描述致病性预测神经网络126的示例架构。

[0088] 评估引擎130基于由致病性预测神经网络126生成的氨基酸分数分布128来生成突变位置处的突变132的致病性分数134。例如,评估引擎130可以基于以下两项之间的差值(例如通过确定该差值的度量)来生成突变132的致病性分数134: (i) 在分数分布128下(即,由其给出)突变位置处的原始氨基酸的分数,以及(ii) 在分数分布128下突变位置处的替代氨基酸的分数。突变位置处的原始氨基酸是指蛋白质122的原始(即,未突变)氨基酸序列中突变位置处的氨基酸。也就是说,评估引擎130可以处理在突变位置具有经掩蔽嵌入的MSA表示(如前所述),以确定分数分布128,并且然后可以根据分数分布中原始氨基酸和替代氨基酸的分数(例如,根据这些分数之间的差值)生成突变132的致病性分数134。

[0089] 在一些实现方式中,评估引擎130可以使用致病性分数134将突变132分类为“致病的”或“良性的”。例如,如果突变132的致病性分数134满足(例如,超过)阈值,则评估引擎130可以将突变132分类为致病的。如果突变的致病性分数134不满足(例如,不超过)阈值,则评估引擎130可以将突变132分类为良性的。一般来说,该系统可以被配置为用相对较大的分数或相对较小的分数来指示致病性突变。

[0090] 致病性预测系统120可以提供致病性分数134、突变分类132或两者,例如,用于存储在存储器中、或用于通过用户界面呈现给用户、或用于下游系统使用、或用于任何其他适当的目的。

[0091] 图2示出了致病性预测神经网络126的示例架构。致病性预测神经网络126可以被包括在致病性预测系统中,例如参考图1描述的致病性预测系统120。

[0092] 致病性预测神经网络126被配置为接收包括蛋白质的MSA的MSA表示124的网络输入。MSA表示124包括嵌入的集合,该集合包括与MSA中包括的每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入。在实现方式中,与蛋白质氨基酸序列中的“突变”位置相对应的嵌入是经掩蔽嵌入。

[0093] 致病性预测神经网络126处理网络输入以生成氨基酸分数分布128。致病性预测系统120可以使用氨基酸分数分布128来生成致病性分数,该致病性分数表征蛋白质氨基酸序列中的突变位置的突变的致病可能性,如上文参考图1所述。

[0094] 致病性预测神经网络126包括嵌入神经网络136 (也称为嵌入子网络)和投影神经网络140 (也称为投影子网络),接下来将分别对它们进行详细描述。

[0095] 嵌入神经网络136被配置为处理致病性预测神经网络126的网络输入(包括MSA表示124),以生成经更新MSA表示138。经更新MSA表示138包括原始MSA表示124中包括的每个嵌入的经更新版本,包括与蛋白质氨基酸序列中的突变位置相对应的嵌入的经更新版本。

[0096] 嵌入神经网络136可以具有任何适当的神经网络架构,使得嵌入神经网络136能够执行其所描述的功能,例如,处理MSA表示124以生成经更新MSA表示138。具体地,嵌入神经网络136可以包括任意适当数量(例如,5层、10层或15层)并且以任何适当的配置(例如,作为层的线性序列)连接的任何适当类型的神经网络层(例如,自注意力层、交叉注意力层、全连接层、卷积层等)。

[0097] 下文将参考图8描述嵌入神经网络的示例实现方式。参考图8描述的示例嵌入神经网络被配置为接收另外包括“成对嵌入”集的输入(即,除了MSA表示124之外),并且被配置为生成另外包括经更新成对嵌入集的输出(即,除了经更新MSA表示138之外),如下文将参考图8更详细地描述的。一般来说,成对嵌入可以是来自氨基酸序列中的氨基酸对得到的嵌入。

[0098] 投影神经网络140被配置为处理来自经更新MSA表示138的与蛋白质的氨基酸序列中的突变位置相对应的经更新嵌入,以生成该氨基酸集上的分数分布。

[0099] 投影神经网络140可以具有任何适当的神经网络架构,使得投影神经网络140能够执行其所描述的功能,例如,处理来自经更新MSA表示的嵌入以生成氨基酸分数分布。具体地,投影神经网络140可以包括任意适当数量(例如,1层、5层、10层或15层)并且以任何适当的配置(例如,作为层的线性序列)连接的任何适当类型的神经网络层(例如,线性层、自注意力层、交叉注意力层、全连接层、卷积层等)。

[0100] 图3示出了示例训练系统142。训练系统142是在一个或多个位置的一个或多个计算机上实现为计算机程序的系统的示例,其中实现下文描述的系统、组件和技术。

[0101] 训练系统142被配置为在训练示例集144上训练致病性预测神经网络126。一般来说,致病性预测神经网络被训练以执行致病性预测任务;它还可以被训练以执行蛋白质去掩蔽任务(去除掩蔽嵌入的影响)和/或蛋白质结构预测任务(预测训练蛋白质的结构)。

[0102] 每个训练示例可以是例如:致病性训练示例、经掩蔽训练示例或结构训练示例。包括经掩蔽训练示例和结构训练示例可以很有用,但不是必要的。

[0103] 致病性训练示例定义:(i)训练蛋白质的氨基酸序列,(ii)训练蛋白质的一个或多个突变,以及(iii)一个或多个突变中的每个突变的相应目标致病性分数。突变中的每个突变可以指定:(i)训练蛋白质的氨基酸序列中的突变位置,以及(ii)训练蛋白质的氨基酸序列中的突变位置的替代氨基酸。稍后将描述生成目标致病性分数的一些方式。训练蛋白质可以包括生成致病性分数所针对的生物体的蛋白质,但不必局限于这样的蛋白质。

[0104] 经掩蔽训练示例定义:(i)训练蛋白质的氨基酸序列,以及(ii)训练蛋白质氨基酸序列中被指定为“经掩蔽”位置的一个或多个位置。

[0105] 结构训练示例定义:(i)训练蛋白质的氨基酸序列,以及(ii)训练蛋白质的结构(例如,由结构参数集表示)。

[0106] 训练系统142可以在以下各项上训练致病性预测神经网络126:(i)致病性训练示

例以优化致病性损失154, (ii) 经掩蔽训练示例以优化掩蔽损失152, 以及 (iii) 结构训练示例以优化结构损失150, 如接下来将更详细地描述的。

[0107] 为了在致病性训练示例上训练致病性预测神经网络126, 训练系统142可以向致病性预测神经网络126提供表征对应训练蛋白质的网络输入。网络输入可以包括训练蛋白质的MSA的MSA表示, 其中MSA表示包括与MSA中每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入。对于由训练示例指定的每个突变, 训练系统142都掩蔽与训练蛋白质的氨基酸序列中该突变的突变位置相对应的嵌入。

[0108] 训练系统142使用致病性预测神经网络126处理训练蛋白质的网络输入, 以生成由训练示例指定的每个突变的突变位置的相应氨基酸分数分布。对于由训练示例指定的每个突变, 训练系统然后基于由致病性预测神经网络126针对突变位置生成的氨基酸分数分布来生成相应的经预测致病性分数134。(上文参考图1描述了从氨基酸分数分布生成突变的致病性分数的示例技术)。

[0109] 训练系统142评估由致病性训练示例指定的每个突变的致病性损失154, 并且然后将致病性训练示例的整体致病性损失确定为每个突变的致病性损失的组合(例如, 平均值或总和)。训练系统142可以基于以下两项之间的误差(例如, 铰链损失、交叉熵损失或任何其他适当的损失)来评估突变的致病性损失: (i) 突变的经预测致病性分数, 以及(ii) 突变的目标致病性分数。突变的目标致病性分数可以按照稍后参考图5描述的那样来获得。作为一个示例, 目标致病性分数可以具有指示生物体群体或生物体群组中突变的流行率的值。然后, 该生物体或生物体群组可以包括这样的生物体: 致病性预测系统120(在推断时)被用于为其生成致病性分数——但由于不同生物体之间可能存在相当大的相似性, 因此致病性预测系统120不必在来自其在推断时被用于的同一生物体的数据上进行训练。

[0110] 训练系统142可以例如使用反向传播来确定致病性损失154相对于致病性预测神经网络126的参数的梯度。训练系统142然后可以使用梯度(例如, 使用适当的梯度下降优化技术)调整致病性预测神经网络126的参数的当前值。

[0111] 为了在经掩蔽训练示例上训练致病性预测神经网络126, 训练系统可以向致病性预测神经网络126提供表征对应训练蛋白质的网络输入。网络输入可以包括表示训练蛋白质的MSA的MSA表示, 其中MSA表示包括与MSA中每个蛋白质的氨基酸序列中的每个位置相对应的相应嵌入。训练系统142可以掩蔽MSA表示124中与训练蛋白质的氨基酸序列中已经被经掩蔽训练示例指定为经掩蔽位置的位置相对应的每个嵌入。

[0112] 训练系统142然后使用致病性预测神经网络126处理训练蛋白质的网络输入, 以生成与训练蛋白质的氨基酸序列中的每个经掩蔽位置相对应的相应氨基酸分数分布。训练系统142评估每个经掩蔽位置的掩蔽损失, 并且然后将训练示例的整体掩蔽损失确定为每个经掩蔽位置的掩蔽损失的组合(例如, 平均值或总和)。训练系统142可以基于在经掩蔽位置的分数分布下训练蛋白质中经掩蔽位置处的氨基酸的分数(例如基于由于掩蔽导致的分数变化)来评估经掩蔽位置的掩蔽损失。掩蔽损失可以衡量由致病性预测神经网络针对每个经掩蔽位置生成的相应预测的准确性。

[0113] 训练系统142可以例如使用反向传播来确定掩蔽损失152相对于致病性预测神经网络126的参数的梯度。训练系统142然后可以使用梯度(例如, 使用适当的梯度下降优化技术)调整致病性预测神经网络126的参数的当前值。

[0114] 通过针对掩蔽损失152来训练致病性预测神经网络126,训练系统142训练致病性预测神经网络126以执行去掩蔽任务。去掩蔽任务需要致病性预测神经网络基于氨基酸序列的剩余未掩蔽部分中的上下文信息来解码蛋白质氨基酸序列中的经掩蔽位置处的氨基酸的身份。学习有效地执行去掩蔽任务可以将对蛋白质生物化学的理解隐式地编码在致病性预测神经网络的参数值中,并且从而可以提高致病性预测神经网络126在致病性预测任务上的性能。

[0115] 为了在结构训练示例上训练致病性预测神经网络126,训练系统142可以向致病性预测神经网络126提供表征对应训练蛋白质的网络输入。网络输入可以包括表示训练蛋白质的MSA的MSA表示。训练系统142使用致病性预测神经网络126处理训练蛋白质的网络输入,并将致病性预测神经网络的中间输出提供给折叠神经网络146。(神经网络的中间输出是指由神经网络的一个或多个隐藏层产生的输出;可以使用任意层)。折叠神经网络146被配置为处理致病性预测神经网络126的中间输出以生成定义训练蛋白质的经预测结构148的数据。

[0116] 折叠神经网络146可以被配置为接收致病性预测神经网络126的任何适当的中间输出。例如,致病性预测神经网络126可以包括嵌入神经网络,如参考图8所述,并且折叠神经网络146可以被配置为接收嵌入神经网络的一些或全部输出。对于具有下文参考图8描述的架构的嵌入神经网络,折叠神经网络可以接收由嵌入神经网络生成的经更新MSA表示和/或经更新成对嵌入。

[0117] 折叠神经网络可以具有任何适当的神经网络架构,使得折叠神经网络能够执行其所描述的功能,特别是处理致病性预测神经网络的中间输出以生成经预测蛋白质结构。具体地,折叠神经网络可以包括任意适当数量(例如,5层、10层或15层)并且以任何适当的配置(例如,作为层的线性序列)连接的任何适当类型的神经网络层(例如,自注意力层、交叉注意力层、全连接层、卷积层等)。下文参考图12描述折叠神经网络的示例实现方式。

[0118] 结构训练示例的结构损失150衡量以下两项之间的误差:(i)由折叠神经网络146生成的经预测蛋白质结构148,以及(ii)由结构训练示例指定的目标蛋白质结构。下文例如参考方程(2)-(4)描述了结构损失示例。训练系统142例如使用反向传播来确定结构损失150相对于致病性预测神经网络126和折叠神经网络146的参数的梯度。训练系统142然后可以使用梯度(例如,通过适当的梯度下降优化技术)调整致病性预测神经网络126和折叠神经网络146的参数的当前值。也就是说,训练系统142将结构损失150的梯度通过折叠神经网络146反向传播到致病性预测神经网络126中。

[0119] 在训练致病性预测神经网络126之后,不需要折叠神经网络146。因此,在训练致病性预测神经网络126之后,可以丢弃折叠神经网络146。

[0120] 通过针对结构损失150来训练致病性预测神经网络126和折叠神经网络146,训练系统142训练致病性预测神经网络126和折叠神经网络146以共同执行蛋白质结构预测的任务。学习有效地执行蛋白质结构预测任务可以将对蛋白质生物化学的理解隐式地编码在致病性预测神经网络的参数值中,并且从而可以提高致病性预测神经网络在致病性预测任务上的性能。

[0121] 更具体地,蛋白质的结构编码了有关蛋白质突变的可能致病性的信息。例如,蛋白质可以折叠以将疏水性氨基酸嵌套在蛋白质内部,同时将亲水性氨基酸暴露在蛋白质外

部。会将疏水性氨基酸置于蛋白质外部的蛋白质突变可能显著改变蛋白质的特性,例如导致蛋白质变得不稳定或者改变蛋白质的形状,并且从而有可能具有致病性。针对结构损失150来训练致病性预测神经网络126可以使致病性预测神经网络126能够推理蛋白质结构,并且从而更准确地执行致病性预测。

[0122] 在一些实现方式中,训练系统142首先训练致病性预测神经网络126,以优化掩蔽损失152和结构损失150。然后,训练系统142禁用掩蔽损失152和结构损失150,并仅针对致病性损失154来训练致病性预测神经网络126。也就是说,训练系统142可以针对掩蔽损失152和结构损失150来预训练致病性预测神经网络126,并且然后针对致病性损失对致病性预测神经网络126进行微调。在一些情况下,在微调期间,训练系统142可以针对致病性损失和结构损失150两者(即,而不是禁用结构损失150)来训练致病性预测神经网络。

[0123] 在针对致病性损失154对致病性预测神经网络进行微调之前针对掩蔽损失152和结构损失150来预训练致病性预测神经网络可以提高致病性预测神经网络在致病性预测任务上的性能。具体地,预训练可以为致病性预测神经网络的参数值生成有效的初始化,并且可以使得致病性预测神经网络使用很少的训练数据和较少的训练迭代就能在致病性预测任务上实现可接受的性能。

[0124] 在一些情况下,用于训练致病性预测神经网络的致病性训练示例可能是“有噪的”,例如,可能通过产生不可忽略数量的具有不准确目标致病性分数的致病性训练示例的过程生成。为了解决这个问题,训练系统142可以在训练过程中从训练数据中过滤(即去除)有噪的致病性训练示例,并且从而正则化和改进对致病性预测神经网络的训练。下文参考图5更详细地描述用于在从训练数据中过滤有噪致病性训练示例的同时训练致病性预测神经网络126的多阶段训练过程的示例。

[0125] 图4是用于生成致病性分数的示例过程158的流程图,该分数表征蛋白质突变是致病性突变的可能性。突变通过在蛋白质的氨基酸序列的突变位置处用替代氨基酸替换原始氨基酸来修改蛋白质的氨基酸序列。在实现方式中,该过程涉及获得表示突变的数据。为了方便起见,过程158将被描述为由位于一个或多个位置中的一个或多个计算机的系统来执行。例如,致病性预测系统(例如,图1的致病性预测系统120)在根据本说明书适当编程后可以执行过程158。

[0126] 该系统生成致病性预测神经网络的网络输入(160)。网络输入包括表示蛋白质的多序列比对(MSA)的MSA表示。在实现方式中,掩蔽MSA表示中与蛋白质氨基酸序列中的突变位置相对应的嵌入。

[0127] 该系统使用(经训练的)致病性预测神经网络处理网络输入,以生成氨基酸集上的分数分布(162)。在实现方式中,致病性预测神经网络已被训练以优化(最小化)致病性损失,该致病性损失表征蛋白质突变的致病性与致病性预测系统对蛋白质突变致病性的预测(即,源自氨基酸集上的分数分布,更具体地源自蛋白质突变的致病性分数)之间的差值。

[0128] 该系统基于以下两项之间的差值生成致病性分数:(i)在分数分布下原始氨基酸的分数,以及(ii)在分数分布下替代氨基酸的分数(164)。

[0129] 图5是用于实现用于训练致病性预测神经网络的多阶段过程的示例过程166的流程图。为了方便起见,过程166将被描述为由位于一个或多个位置中的一个或多个计算机的系统来执行。例如,训练系统(例如,图3的训练系统142)在根据本说明书适当编程后可以执

行过程166。

[0130] 该系统生成致病性训练示例集(168)。每个致病性训练示例定义:(i) 训练蛋白质的氨基酸序列,(ii) 训练蛋白质的一个或多个突变,以及(iii) 一个或多个突变中的每个突变的相应目标致病性分数。

[0131] 在一些实现方式中,该系统可以基于表征生物体群体(例如人类群体、或灵长类动物群体、或哺乳动物或其他动物的群体、或植物群体、或微生物群体)中蛋白质突变的发生频率的数据集来生成或获得致病性训练示例集。此外或替代地,该系统可以使用已知的生物数据(例如来自适当的数据库)来获得突变致病性的度量。

[0132] 具体地,该系统可以将群体中至少阈值百分比的生物体中发生的每个蛋白质突变识别为“良性”突变,即,非致病性突变。直观地说,具有严重致病性的蛋白质突变不太可能在群体中频繁出现,例如,因为具有严重致病性突变的生物体将更有可能从群体中消失,并且不太可能繁殖。因此,在群体中至少以阈值频率发生的蛋白质突变不太可能具有严重致病性,并且可以被识别为良性突变以用于生成用于训练致病性预测神经网络的训练示例。

[0133] 如果蛋白质突变尚未在生物体群体中发生,则该系统可以将该蛋白质突变指定为致病性突变。因此,为了识别致病性突变,该系统可以随机生成蛋白质突变集,去除生物体群体中已经发生的任何突变,并且然后将剩余的突变指定为致病性突变。直观地说,蛋白质突变可能不会在群体中发生是因为:(i) 群体(有效群体大小)不够大,不足以让突变反映在群体中的至少一个生物体中,或(ii) 蛋白质突变具有严重致病性,以至于具有该蛋白质突变的生物体极不可能繁殖或存活。

[0134] 在如上所述识别出良性突变集和致病性突变集之后,该系统可以继续生成或获得致病性训练示例。具体地,为了生成致病性训练示例,该系统可以识别训练蛋白质和与训练蛋白质相关联的一个或多个蛋白质突变。该系统可以将第一致病性分数(例如零(0))与已被指定为良性的每个训练蛋白质突变相关联,并且将第二致病性分数(例如一(1))与已被指定为致病的每个训练蛋白质突变相关联。

[0135] 应当理解,上述识别良性和致病性蛋白质突变的过程可能导致生成“有噪的”致病性训练示例。例如,致病性蛋白质突变发生在群体中,并且因此蛋白质突变在群体中出现至少阈值次数并不一定意味着该蛋白质突变是良性的。反之,蛋白质突变可能未发生在群体中,因为群体(有效群体大小)不够大,而不是因为蛋白质突变具有致病性。为了解决这个问题,该系统可以过滤致病性训练示例集以去除可能不准确的致病性训练示例,如下文参考步骤(172)描述的。

[0136] 该系统在训练示例集(170)上训练致病性预测神经网络。该训练示例集包括在步骤(168)处生成的致病性训练示例、以及经掩蔽训练示例和结构训练示例。参考图3更详细地描述了在包括致病性训练示例、经掩蔽训练示例和结构训练示例的训练示例集上训练致病性预测神经网络的示例技术。

[0137] 在一些实现方式中,该系统过滤致病性训练示例集(172)。具体地,对于每个蛋白质突变,该系统可以使用致病性预测神经网络生成该蛋白质突变的经预测致病性分数。然后,该系统基于蛋白质突变的(i) 目标致病性分数与(ii) 经预测致病性分数之间的误差生成蛋白质突变的误差分数。在此和其他地方,蛋白质突变的目标致病性分数可以是例如指示蛋白质突变被指定为良性还是致病性的二元0/1分数,例如在步骤(168)处。直观地说,蛋

白质突变的误差分数较大指示经训练的致病性预测神经网络生成与蛋白质突变是良性还是致病性的最初指定相反的(自信的)预测。蛋白质突变的较大误差分数可以指示该蛋白质突变被错误标记为良性或致病性。

[0138] 该系统可以至少部分地基于对应训练蛋白质的氨基酸序列中包括的突变的误差分数来生成每个致病性训练示例的过滤分数。例如,该系统可以至少部分地基于对应训练蛋白质的氨基酸序列中包括的突变的误差分数的组合(例如乘积)来生成致病性训练示例的过滤分数。该系统可以采用多种方式使用过滤分数来过滤致病性训练示例。作为一个示例,该系统可以使用过滤分数(例如通过softmax)生成致病性训练示例集上的概率分布,根据概率分布采样多个致病性训练示例(例如,使得具有较高过滤分数的示例更有可能被选择),并且从致病性训练示例集中过滤(去除)所采样的致病性训练示例。

[0139] 致病性训练示例的过滤分数可以另外基于对应训练蛋白质的其他方面,即,除了蛋白质突变的误差分数之外。例如,如果对应训练蛋白质中包括的(致病性或良性)突变数量与例如通常在蛋白质中观察到的预期的(致病性或良性)突变数量有显著偏差(例如,偏离至少阈值量),则致病性训练示例的过滤分数可以具有更高的值。致病性预测神经网络可以很容易将具有比预期(致病性或良性)突变数量明显更高或更低的(致病性或良性)突变数量的训练蛋白质与天然蛋白质区分开,从而可能影响训练的有效性。

[0140] 基于过滤分数对致病性训练示例集进行过滤可以提高致病性训练示例集的质量,例如,通过去除具有错误标记的蛋白质突变的训练示例以及通过去除具有非自然人为因素(unnatural artifacts)的训练示例。

[0141] 该系统可以在包括经过滤的致病性训练示例集的训练示例集上再训练致病性预测神经网络(174)。也就是说,在致病性预测神经网络的再训练期间,不使用被过滤(被去除)的致病性训练示例。因此,例如与步骤(170)中训练的原始致病性预测神经网络相比,经再训练的致病性预测神经网络可以被训练以在更少的训练迭代中实现更高的预测准确性。

[0142] 图6A示出了蛋白质的MSA的MSA表示的示例。MSA表示的每一行表示MSA中的相应MSA蛋白质。每一行包括相应的嵌入集,该行所表示的MSA蛋白质的氨基酸序列中的每个位置包括一个嵌入。MSA表示的第二行表示蛋白质,并且嵌入176表示蛋白质氨基酸序列中的突变位置并且在示例MSA表示中被掩蔽。致病性预测神经网络可以处理MSA表示以生成致病性分数,该致病性分数表征蛋白质氨基酸序列中的突变位置的突变的致病可能性。

[0143] 图6B示出了由致病性预测神经网络生成的氨基酸集上的分数分布的示例。在图6B中,氨基酸使用其单字母缩写来表示。氨基酸L是蛋白质氨基酸序列中突变位置处的原始氨基酸。致病性预测系统可以基于在分数分布下氨基酸L的分数(0.55)与氨基酸N的分数0.44之间的差值来评估在突变位置处从氨基酸L到氨基酸N的突变的致病可能性。致病性预测系统可以基于在分数分布下氨基酸L的分数(0.55)与氨基酸R的分数0.01之间的差值来评估在突变位置处从氨基酸L突变到氨基酸R的可能性。分数分布表明氨基酸R的突变比氨基酸N的突变更有可能致病($L - N$ 差值小于 $L - R$ 差值)。

[0144] 图7示出了示例蛋白质结构预测系统100。蛋白质结构预测系统100是在一个或多个位置的一个或多个计算机上实现为计算机程序的系统的示例,其中实现下文描述的系统、组件和技术。

[0145] 系统100被配置为处理定义蛋白质104的一个或多个氨基酸链102的数据,以生成

定义经预测蛋白质结构108 (即,对蛋白质104的结构的预测)的结构参数集106。也就是说,蛋白质104的经预测结构108可以由共同定义在蛋白质经历蛋白质折叠之后蛋白质的经预测三维结构的结构参数集106定义。

[0146] 定义经预测蛋白质结构108的结构参数106可以如前所述。例如,它们可以包括例如蛋白质104中每个氨基酸的位置参数和旋转参数、表征蛋白质中每个氨基酸对之间的估计距离的距离图、蛋白质结构中每个氨基酸中每个原子或骨架 (backbone) 原子的相应空间位置、或者它们的组合,如上所述。

[0147] 为了生成定义经预测蛋白质结构108的结构参数106,系统100生成:(i) 蛋白质的多序列比对 (MSA) 表示110,以及 (ii) 蛋白质的“成对”嵌入112的集,如下文将更详细地描述的。

[0148] 蛋白质的MSA表示110包括蛋白质中每个氨基酸链的MSA的相应表示。蛋白质中氨基酸链的MSA表示可以被表示为 $M \times N$ 嵌入阵列 (即,具有M行和N列的2-D嵌入阵列),其中N是氨基酸链中的氨基酸数量。MSA表示的每一行可以对应于蛋白质中氨基酸链的相应MSA序列。参考图16描述了用于生成蛋白质中氨基酸链的MSA表示的示例过程。

[0149] 系统100根据蛋白质中氨基酸链的MSA表示生成蛋白质104的MSA表示110。

[0150] 如果蛋白质仅包括单个氨基酸链,则系统100可以将蛋白质104的MSA表示110识别为蛋白质中单个氨基酸链的MSA表示。

[0151] 如果蛋白质包括多个氨基酸链,则系统100可以通过将蛋白质中这些氨基酸链的MSA表示组装成块对角线2-D嵌入阵列来生成蛋白质的MSA表示110,即,蛋白质中氨基酸链的MSA表示形成对角线上的块。系统100可以将2-D阵列中在对角线上的块之外的每个位置处的嵌入初始化为默认嵌入,例如零向量。蛋白质中的氨基酸链可以被指派任意顺序,并且蛋白质中氨基酸链的MSA表示可以在块对角矩阵中进行相应地排序。例如,第一氨基酸链的MSA表示 (即,根据排序) 可以是对角线上的第一块,第二氨基酸链的MSA表示可以是对角线上的第二块,依此类推。

[0152] 一般来说,蛋白质的MSA表示110可以被表示为2-D嵌入阵列。在整个本说明书中,蛋白质的MSA表示的“行”是指定义蛋白质的MSA表示的2-D嵌入阵列的行。类似地,蛋白质的MSA表示的“列”是指定义蛋白质的MSA表示的2-D嵌入阵列的列。

[0153] 成对嵌入112的集包括与蛋白质104中的每个氨基酸对相对应的相应成对嵌入。一般来说,成对嵌入表示即编码有关蛋白质中的氨基酸对之间的关系的信息。氨基酸对是指包括蛋白质中的第一氨基酸和第二氨基酸的有序元组,即,使得蛋白质中的可能氨基酸对的集如下给出:

$$[0154] \quad \{(A_i, A_j): 1 \leq i, j \leq N\} \quad (1)$$

[0155] 其中N是蛋白质中的氨基酸的数量, $i, j \in \{1, \dots, N\}$ 索引蛋白质中的氨基酸, A_i 是蛋白质中由i索引的氨基酸,并且 A_j 是蛋白质中由j索引的氨基酸。如果蛋白质包括多个氨基酸链,则蛋白质中的氨基酸可以按照蛋白质中氨基酸链的顺序从 $\{1, \dots, N\}$ 依次索引。也就是说,首先按顺序索引来自第一氨基酸链的氨基酸,然后是来自第二氨基酸链的氨基酸,然后是来自第三氨基酸链的氨基酸,依此类推。成对嵌入112的集可以被表示为2-D $N \times N$ 成对嵌入阵列,例如,其中2-D阵列的行由 $i \in \{1, \dots, N\}$ 索引,2-D阵列的列由 $j \in \{1, \dots, N\}$ 索

引,并且2-D阵列中的位置 (i, j) 被该氨基酸对 (A_i, A_j) 的成对嵌入所占据。

[0156] 参考图17描述了用于生成(初始化)与蛋白质中的每个氨基酸对相对应的相应成对嵌入的示例过程。

[0157] 系统100使用MSA表示110和成对嵌入112两者来生成定义经预测蛋白质结构108的结构参数106,因为这两者具有互补特性。MSA表示110的结构可以明确地取决于与蛋白质中每个氨基酸链相对应的MSA中氨基酸链的数量。因此,MSA表示110可能不适合用于直接预测蛋白质结构,因为蛋白质结构108对MSA中氨基酸链的数量没有明确的依赖关系。相比之下,成对嵌入112表征蛋白质104中的相应氨基酸对之间的关系,并且在表达时不明确参考MSA,并且因此是用于预测蛋白质结构108的方便且有效的数据表示。

[0158] 系统100使用嵌入神经网络200根据嵌入神经网络200的参数集的值来处理MSA表示110和成对嵌入112,以更新MSA表示110和成对嵌入112。也就是说,嵌入神经网络200处理MSA表示110和成对嵌入112以生成经更新MSA表示114和经更新成对嵌入116。

[0159] 嵌入神经网络200通过在MSA表示110与成对嵌入112之间共享信息来更新MSA表示110和成对嵌入112。更具体地,嵌入神经网络200在基于当前成对嵌入112更新当前MSA表示110与基于当前MSA表示110更新当前成对嵌入112之间交替进行。

[0160] 参考图8更详细地描述嵌入神经网络200的示例架构。

[0161] 系统100从经更新成对嵌入116、经更新MSA表示114或两者生成折叠神经网络600的网络输入,并使用折叠神经网络600处理网络输入以生成定义经预测蛋白质结构的结构参数106。

[0162] 在一些实现方式中,折叠神经网络600处理经更新成对嵌入116以生成距离图,该距离图针对蛋白质中的每个氨基酸对包括在蛋白质结构中该氨基酸对之间的可能距离集上的概率分布。例如,为了生成蛋白质结构中氨基酸对之间的可能距离集上的概率分布,折叠神经网络可以将一个或多个全连接的神经网络层应用于与该氨基酸对相对应的经更新成对嵌入116。

[0163] 在一些实现方式中,折叠神经网络600通过使用明确推理蛋白质结构中氨基酸的3-D几何形状的几何注意力操作处理从经更新MSA表示114和经更新成对嵌入116两者得到的网络输入来生成结构参数106。参考图12描述了实现几何注意力机制的折叠神经网络600的示例架构。

[0164] 训练引擎可以端到端地训练蛋白质结构预测系统100,以优化本文被称为结构损失的目标函数。训练引擎可以在包括多个训练示例的训练数据集上训练系统100。每个训练示例可以指定:(i) 包括蛋白质的初始MSA表示和初始成对嵌入的训练输入,以及(ii)应当由系统100通过处理训练输入生成的目标蛋白质结构。可以使用实验技术(例如X射线晶体学)来确定用于训练系统100的目标蛋白质结构。

[0165] 结构损失可以表征以下两项之间的相似性:(i) 由系统100生成的经预测蛋白质结构,与(ii) 该系统应当已经生成的目标蛋白质结构。

[0166] 例如,如果经预测结构参数定义了蛋白质中每个氨基酸的经预测位置参数和经预测旋转参数,则结构损失 $\mathcal{L}_{\text{结构}}$ 可以由下式给出:

$$\mathcal{L}_{\text{结构}} = \frac{1}{N^2} \sum_{i,j=1}^N \left(1 - \frac{|t_{ij} - \tilde{t}_{ij}|}{A} \right)_+ \quad (2)$$

[0167]

$$t_{ij} = R_i^{-1}(t_j - t_i) \quad (3)$$

$$\tilde{t}_{ij} = \tilde{R}_i^{-1}(\tilde{t}_j - \tilde{t}_i) \quad (4)$$

[0168] 其中N是蛋白质中的氨基酸的数量, t_i 表示氨基酸i的经预测位置参数, R_i 表示由氨基酸i的经预测旋转参数指定的 3×3 旋转矩阵, $\tilde{t}_i$ 是氨基酸i的目标位置参数, $\tilde{R}_i$ 表示由氨基酸i的目标旋转参数指定的 3×3 旋转矩阵, A是常数, R_i^{-1} 是指由经预测旋转参数 R_i 指定的 3×3 旋转矩阵的逆, $\tilde{R}_i^{-1}$ 是指由目标旋转参数 $\tilde{R}_i$ 指定的 3×3 旋转矩阵的逆, 并且 $(\cdot)_+$ 表示修正线性单元(ReLU)操作。

[0169] 参考方程(2) - (4)定义的结构损失可以被理解为对蛋白质中的每个氨基酸对的损失 $|t_{ij} - \tilde{t}_{ij}|$ 进行平均。项 t_{ij} 定义氨基酸j在氨基酸i的预测参考系中的经预测空间位置, 并且 $\tilde{t}_{ij}$ 定义氨基酸j在氨基酸i的实际参考系中的实际空间位置。这些项对氨基酸i和j的经预测旋转和实际旋转很敏感, 并且因此与仅对氨基酸之间的经预测距离和实际距离敏感的损失项相比携带更丰富的信息。

[0170] 作为另一个示例, 如果经预测结构参数定义了蛋白质的每个氨基酸中每个原子的经预测空间位置, 那么结构损失可以是以下两项之间的平均误差(例如, 平方误差): (i) 原子的经预测空间位置, 以及(ii) 原子的目标(例如, 基准真实)空间位置。

[0171] 优化结构损失促使系统100生成准确近似真实蛋白质结构的经预测蛋白质结构。

[0172] 除了优化结构损失之外, 训练引擎还可以训练系统100以优化一个或多个辅助损失。辅助损失可以惩罚具有自然界中不太可能出现的特性的经预测结构, 例如基于经预测结构中氨基酸中的原子之间的键的键角和/或键长, 或基于经预测结构中不同氨基酸中的原子的接近度。

[0173] 训练引擎可以通过多次训练迭代在训练数据上训练结构预测系统100, 例如, 使用随机梯度下降训练技术。

[0174] 图8示出了嵌入神经网络200的示例架构, 该嵌入神经网络被配置为处理MSA表示110和成对嵌入112以生成经更新MSA表示114和经更新成对嵌入116。

[0175] 嵌入神经网络200包括更新块序列202-A至202-N。在整个本说明书中, “块”是指神经网络的一部分, 例如, 包括一个或多个神经网络层的神经网络子网络。

[0176] 嵌入神经网络中的每个更新块被配置为接收包括MSA表示和成对嵌入的块输入, 并处理块输入以生成包括经更新MSA表示和经更新成对嵌入的块输出。

[0177] 嵌入神经网络200将嵌入神经网络200的网络输入中包括的MSA表示110和成对嵌入112提供给(即, 更新块序列中的)第一更新块。第一更新块处理MSA表示110和成对嵌入112以生成经更新MSA表示和经更新成对嵌入。

[0178] 对于第一更新块之后的每个更新块, 嵌入神经网络200向更新块提供由前一个更新块生成的MSA表示和成对嵌入, 并将由该更新块生成的经更新MSA表示和经更新成对嵌入提供给下一个更新块。

[0179] 嵌入神经网络200通过使用更新块序列202-A至202-N重复更新MSA表示110和成对嵌入112来逐渐丰富它们的信息内容。

[0180] 嵌入神经网络200可以提供由(即,更新块序列中的)最终更新块生成的经更新MSA表示114和经更新成对嵌入116作为网络输出。

[0181] 图9示出了嵌入神经网络200(即,如参考图8所述)的更新块300的示例架构。

[0182] 更新块300接收包括当前MSA表示302和当前成对嵌入304的块输入,并处理该块输入以生成经更新MSA表示306和经更新成对嵌入308。

[0183] 更新块300包括MSA更新块400和成对更新块500。

[0184] MSA更新块400使用当前成对嵌入304来更新当前MSA表示302,并且成对更新块500使用(即,由MSA更新块400生成的)经更新MSA表示306来更新当前成对嵌入304。

[0185] 一般来说,MSA表示和成对嵌入可以编码互补信息。例如,MSA表示可以编码有关进化相关的氨基酸链集中不同位置的氨基酸身份之间的相关性的信息,而成对嵌入可以编码有关蛋白质中的氨基酸之间的相互关系的信息。MSA更新块400使用成对嵌入中编码的互补信息来丰富MSA表示的信息内容,而成对更新块500使用MSA表示中编码的互补信息来丰富成对嵌入的信息内容。由于这种丰富,经更新MSA表示和经更新成对嵌入编码了与预测蛋白质结构更相关的信息。

[0186] 本文将更新块300描述为首先使用当前成对嵌入304来更新当前MSA表示302,并且然后使用经更新MSA表示306来更新当前成对嵌入304。该描述不应理解为将更新块限制为按此顺序执行操作,例如,更新块可以首先使用当前MSA表示来更新当前成对嵌入,并且然后使用经更新成对嵌入来更新当前MSA表示。

[0187] 更新块300在本文被描述为包括MSA更新块400(即,更新当前MSA表示)和成对更新块500(即,更新当前成对嵌入)。该描述不应理解为将更新块300限制为包括仅一个MSA更新块或仅一个成对更新块。例如,更新块300可以包括多个MSA更新块,这些更新块在将MSA表示提供给成对更新块以用于更新当前成对嵌入之前多次更新MSA表示。作为另一个示例,更新块300可以包括多个成对更新块,这些成对更新块使用MSA表示来多次更新成对嵌入。

[0188] MSA更新块400和成对更新块500可以具有任何适当的架构,以使它们能够执行其所描述的功能。

[0189] 在一些实现方式中,MSA更新块400、成对更新块500或两者包括一个或多个“自注意力”块。如整个本文档中所使用的,自注意力块通常是指更新嵌入集合(即,接收嵌入集合并输出经更新嵌入)的神经网络块。为了更新给定嵌入,自注意力块可以确定给定嵌入与一个或多个所选嵌入中的每个所选嵌入之间的相应“注意力权重”,并且然后使用以下各项更新给定嵌入:(i)注意力权重,以及(ii)所选嵌入。为了方便起见,自注意力块可以被说成使用对所选嵌入“施加”的注意力来更新给定的嵌入。

[0190] 例如,自注意力块可以接收输入嵌入 $\{x_i\}_{i=1}^N$ 的集合,其中N是蛋白质中的氨基酸的数量,并且为了更新嵌入 x_i ,自注意力块可以确定注意力权重 $[a_{i,j}]_{j=1}^N$,其中 $a_{i,j}$ 表示 x_i 与 x_j 之间的注意力权重,如下所示:

$$[0191] \quad [a_{i,j}]_{j=1}^N = \text{softmax}\left(\frac{(W_q x_i) K^T}{c}\right) \quad (5)$$

$$K^T = [W_k x_j]_{j=1}^N \quad (6)$$

[0192] 其中 W_q 和 W_k 是学习到的参数矩阵, $\text{softmax}(\cdot)$ 表示soft-max归一化运算, 并且 c 是常数。使用注意力权重, 自注意力层可以更新嵌入 x_i , 如下所示:

$$[0193] \quad x_i \leftarrow \sum_{j=1..N} a_{i,j} \cdot (W_v x_j) \quad (7)$$

[0194] 其中 W_v 是学习到的参数矩阵。 $(W_q x_i)$ 可以被为输入嵌入 x_i 的“查询嵌入”, $W_k x_j$ 可以被为输入嵌入 x_j 的“键嵌入”, 并且 $W_v x_j$ 可以被为输入嵌入 x_j 的“值嵌入”。

[0195] 参数矩阵 W_q (“查询嵌入矩阵”)、 W_k (“键嵌入矩阵”) 以及 W_v (“值嵌入矩阵”) 是自注意力块的可训练参数。MSA更新块400和成对更新块500中包括的任何自注意力块的参数可以被理解为更新块300的参数, 这些参数可以作为对参考图1描述的蛋白质结构预测系统100的端到端训练的一部分进行训练。一般来说, 查询嵌入矩阵、键嵌入矩阵和值嵌入矩阵的(经训练)参数对于不同的自注意力块是不同的, 例如, 使得MSA更新块400中包括的自注意力块可以具有带有与成对更新块500中包括的自注意力块不同的参数的不同的查询嵌入矩阵、键嵌入矩阵和值嵌入矩阵。

[0196] 在一些实现方式中, MSA更新块400、成对更新块500或两者包括一个或多个以成对嵌入为条件的自注意力块, 即, 实现以成对嵌入为条件的自注意力操作。为了使自注意力操作以成对嵌入为条件, 自注意力块可以处理成对嵌入以生成与每个注意力权重相对应的相应“注意力偏置”。例如, 除了根据方程(5) - (6) 确定注意力权重 $[a_{i,j}]_{j=1}^N$ 之外, 自注意力块还可以生成对应的注意力偏置集 $[b_{i,j}]_{j=1}^N$, 其中 $b_{i,j}$ 表示 x_i 与 x_j 之间的注意力偏置。自注意力块可以通过将学习到的参数矩阵应用于(即, 蛋白质中由 (i, j) 索引的氨基酸对的)成对嵌入 $h_{i,j}$ 来生成注意力偏置 $b_{i,j}$ 。

[0197] 自注意力块可以例如通过对注意力权重和注意力偏置求和(或以其他方式组合)来确定“有偏(biased)注意力权重”集 $[c_{i,j}]_{j=1}^N$, 其中 $c_{i,j}$ 表示 x_i 与 x_j 之间的有偏注意力权重。例如, 自注意力块可以确定嵌入 x_i 与 x_j 之间的有偏注意力权重 $c_{i,j}$, 如下所示:

$$[0198] \quad c_{i,j} = a_{i,j} + b_{i,j}$$

[0199] 其中 $a_{i,j}$ 是 x_i 与 x_j 之间的注意力权重, 并且 $b_{i,j}$ 是 x_i 与 x_j 之间的注意力偏置。自注意力块可以使用有偏注意力权重来更新每个输入嵌入 x_i , 例如:

$$[0200] \quad x_i \leftarrow \sum_{j=1..N} c_{i,j} \cdot (W_v x_j) \quad (8)$$

[0201] 其中 W_v 是学习到的参数矩阵。

[0202] 一般来说, 成对嵌入编码了表征蛋白质结构以及蛋白质结构中的氨基酸对之间的

关系的信息。对输入嵌入集应用以成对嵌入为条件的自注意力操作允许输入嵌入以受成对嵌入中编码的蛋白质结构信息影响的方式来更新。嵌入神经网络的更新块可以使用以成对嵌入为条件的自注意力块来更新和丰富MSA表示和成对嵌入本身。

[0203] 可选地,自注意力块可以具有多个“头”,每个头生成与每个输入嵌入相对应的相应经更新嵌入,即,使得每个输入嵌入与多个经更新嵌入相关联。例如,每个头可以根据参考方程 (5) - (7) 描述的参数矩阵 W_q 、 W_k 和 W_v 的不同值生成经更新嵌入。具有多个头的自注意力块可以实现“门控(gating)”操作,以将由这些头针对输入嵌入生成的经更新嵌入组合起来,即,生成与每个输入嵌入都对应的单个经更新嵌入。例如,自注意力块可以使用一个或多个神经网络层(例如,全连接的神经网络层)处理输入嵌入,以为每个头生成相应的门控值。自注意力块然后可以根据门控值组合与输入嵌入相对应的经更新嵌入。例如,自注意力块可以针对输入嵌入 x_i 生成经更新嵌入,如下所示:

$$[0204] \quad \sum_{k=1}^K \alpha_k \cdot x_i^{next} \quad (9)$$

[0205] 其中 k 索引头, α_k 是头的门控值 k ,并且 x_i^{next} 是头针对输入嵌入 x_i 生成的经更新嵌入 k 。

[0206] 参考图10描述了使用以成对嵌入为条件的自注意力块的MSA更新块400的示例架构。参考图10描述的示例MSA更新块基于当前成对嵌入,通过使用以当前成对嵌入为条件的自注意力块处理当前MSA表示的行来更新当前MSA表示。

[0207] 参考图11描述了使用以成对嵌入为条件的自注意力块的成对更新块500的示例架构。参考图11描述的示例成对更新块基于经更新MSA表示,通过计算经更新MSA表示的外积均值、将外积均值的结果添加到当前成对嵌入、并使用以当前成对嵌入为条件的自注意力块处理当前成对嵌入来更新当前成对嵌入。

[0208] 图10示出了MSA更新块400的示例架构。MSA更新块400被配置为接收当前MSA表示302,以(至少部分地)基于当前成对嵌入来更新当前MSA表示306。

[0209] 为了更新当前MSA表示302,MSA更新块400使用以当前成对嵌入为条件的自注意力操作(即,仅对特定行中的嵌入进行操作的“按行”自注意力操作)来更新当前MSA表示的每一行中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每一行中的嵌入提供给以当前成对嵌入为条件的“按行”自注意力块402(例如,如参考图9所述),以生成当前MSA表示302的每一行的经更新嵌入。可选地,MSA更新块可以将按行自注意力块402的输入添加到按行自注意力块402的输出。使按行自注意力块402以当前成对嵌入为条件使得MSA更新块400能够使用来自当前成对嵌入的信息来丰富当前MSA表示302。

[0210] 然后,MSA更新块使用不以当前成对嵌入为条件的自注意力操作(即,仅对特定列中的嵌入进行操作的“按列”自注意力操作)来更新当前MSA表示的每一列中的嵌入。更具体地,MSA更新块400将当前MSA表示302的每一列中的嵌入提供给不以当前成对嵌入为条件的“按列”自注意力块404,以生成当前MSA表示302的每一列的经更新嵌入。由于不以当前成对嵌入为条件,按列自注意力块404使用注意力权重(例如,参考方程(5) - (6)所述)而不是有偏注意力权重(例如,参考方程(8)所述)来生成当前MSA表示的每一列的经更新嵌入。可选

地,MSA更新块可以将按列自注意力块404的输入添加到按列自注意力块404的输出。

[0211] 然后,MSA更新块使用转换块(transition block)处理当前MSA表示302,该转换块例如将一个或多个全连接的神经网络层应用于当前MSA表示302。可选地,MSA更新块400可以将转换块406的输入添加到转换块406的输出。

[0212] MSA更新块可以输出由按行自注意力块402、按列自注意力块404和转换块406执行的操作产生的经更新MSA表示306。

[0213] 图11示出了成对更新块500的示例架构。成对更新块500被配置为接收当前成对嵌入304,并(至少部分地)基于经更新MSA表示306来更新当前成对嵌入304。

[0214] 为了更新当前成对嵌入304,成对更新块500将外积均值运算502应用于经更新MSA表示306,并将外积均值运算502的结果添加到当前成对嵌入304。

[0215] 外积均值运算定义了运算序列,该运算序列当被应用于表示为 $M \times N$ 嵌入阵列的MSA表示时,生成 $N \times N$ 嵌入阵列,即,其中N是蛋白质中的氨基酸的数量。当前成对嵌入304也可以被表示为 $N \times N$ 嵌入阵列,并且将外积均值502的结果添加到当前成对嵌入304是指将这两个 $N \times N$ 嵌入阵列相加。

[0216] 为了计算外积均值,成对更新块生成张量 $A(\cdot)$,例如,由下式给出:

$$A(res1, res2, ch1, ch2) = \frac{1}{|rows|} \sum_{rows} LeftAct(row, res1, ch1) \cdot RightAct(row, res2, ch2)$$

[0218] 其中 $res1, res2 \in \{1, \dots, N\}$, $ch1, ch2 \in \{1, \dots, C\}$,其中C是MSA表示的每个嵌入中的通道数, $|rows|$ 是MSA表示中的行数, $LeftAct(row, res1, ch1)$ 是被应用于MSA表示中位于由“row”索引的行和由“res1”索引的列处的嵌入的通道 $ch1$ 的线性运算(例如,由矩阵乘法定义),并且 $RightAct(row, res2, ch2)$ 是被应用于MSA表示中位于由“row”索引的行和由“res2”索引的列处的嵌入的通道 $ch2$ 的线性运算(例如,由矩阵乘法定义)。外积均值的结果是通过对张量A的 $(ch1, ch2)$ 维度进行扁平化和线性投影来生成的。可选地,作为计算外积均值的一部分,成对更新块可以执行一个或多个层归一化操作(例如,参考以下文献描述的:Jimmy Lei Ba等人的“Layer Normalization [层归一化]”,arXiv:1607.06450)。

[0219] 一般来说,经更新MSA表示306编码了有关进化相关的氨基酸链集中不同位置的氨基酸身份之间的相关性的信息。经更新MSA表示306中编码的信息与预测蛋白质的结构相关,并且通过将经更新MSA表示中编码的信息合并到当前成对嵌入中(即,通过外积均值502),成对更新块500可以增强当前成对嵌入的信息内容。

[0220] 在使用经更新MSA表示(即,通过外积均值502)来更新当前成对嵌入304之后,成对更新块500使用以当前成对嵌入为条件的自注意力操作(即,“按行”自注意力操作)将当前成对嵌入的排列的每一行中的当前成对嵌入更新为 $N \times N$ 阵列。更具体地,成对更新块500将当前成对嵌入的每一行提供给也以当前成对嵌入为条件的“按行”自注意力块504(例如,如参考图9所述),以生成每一行的经更新成对嵌入。可选地,成对更新块可以将按行自注意力块504的输入添加到按行自注意力块504的输出。

[0221] 然后,成对更新块500使用也以当前成对嵌入为条件的自注意力操作(即,“按列”自注意力操作)来更新当前成对嵌入的 $N \times N$ 阵列的每一列中的当前成对嵌入。更具体地,

成对更新块500将当前成对嵌入的每一列提供给也以当前成对嵌入为条件的“按列”自注意力块506,以生成每一行的经更新成对嵌入。可选地,成对更新块500可以将按列自注意力块506的输入添加到按列自注意力块506的输出。

[0222] 然后,成对更新块500使用转换块处理当前成对嵌入,该转换块例如将一个或多个全连接的神经网络层应用于当前成对嵌入。可选地,成对更新块500可以将转换块508的输入添加到转换块508的输出。

[0223] 成对更新块500可以输出由按行自注意力块504、按列自注意力块506和转换块508执行的操作产生的经更新成对嵌入308。

[0224] 图12示出了折叠神经网络600的示例架构,该折叠神经网络生成定义经预测蛋白质结构108的结构参数集106。折叠神经网络600可以被包括在参考图1描述的蛋白质结构预测系统100中。

[0225] 折叠神经网络600为蛋白质中的每个氨基酸生成结构参数,这些结构参数可以包括:(i)位置参数以及(ii)旋转参数。如前所述,氨基酸的位置参数可以指定蛋白质结构中的氨基酸中的指定原子的经预测3-D空间位置。氨基酸的旋转参数可以指定蛋白质结构中的氨基酸的经预测“取向”。更具体地,旋转参数可以指定3-D空间旋转操作,如果将该空间旋转操作应用于位置参数的坐标系,则导致氨基酸中的三个“主链”原子相对于旋转后的坐标系处于固定位置。

[0226] 在实现方式中,折叠神经网络600接收从最终MSA表示、最终成对嵌入或两者得到的输入,并生成定义蛋白质经预测结构的结构参数106的最终值。例如,折叠神经网络600可以接收以下输入:(i)蛋白质中每个氨基酸对的相应成对嵌入116,(ii)蛋白质中每个氨基酸的“单”嵌入602的初始值,以及(iii)蛋白质中每个氨基酸的结构参数604的初始值。折叠神经网络600处理输入以生成共同表征蛋白质的经预测结构108的结构参数106的最终值。

[0227] 蛋白质结构预测系统100可以向折叠神经网络600提供如参考图1所述作为嵌入神经网络的输出而生成的成对嵌入。

[0228] 蛋白质结构预测系统100可以根据MSA表示114生成氨基酸的初始单嵌入602,即,如参考图1所述作为嵌入神经网络的输出生成的嵌入。例如,如上所述,MSA表示114可以被表示为具有等于蛋白质中的氨基酸数的2-D嵌入阵列,其中每列与蛋白质中的相应氨基酸相关联。蛋白质结构预测系统100可以针对蛋白质中的每个氨基酸,通过对来自MSA表示114中与该氨基酸相关联的列的嵌入进行求和(或以其他方式组合)来生成初始单嵌入。作为另一个示例,蛋白质结构预测系统100可以通过从MSA表示114中与其结构正被估计的蛋白质的氨基酸序列相对应的行中提取嵌入来生成蛋白质中氨基酸的初始单嵌入。

[0229] 蛋白质结构预测系统100可以生成具有默认值的初始结构参数604,例如,其中每个氨基酸的位置参数被初始化为原点(例如,笛卡尔坐标系中的 $[0,0,0]$),并且每个氨基酸的旋转参数被初始化为 3×3 单位矩阵。

[0230] 折叠神经网络600可以通过重复更新单嵌入606和结构参数608的当前值(即,从它们的初始值开始)来生成最终的结构参数106。更具体地,折叠神经网络600包括更新神经网络块610的序列,其中每个更新块610被配置为更新当前单嵌入606(即,生成经更新单嵌入612)并更新当前结构参数608(即,生成经更新结构参数614)。除了更新块之外,折叠神经网络600还可以包括其他神经网络层或块,例如可以与更新块交错。

[0231] 每个更新块610可以包括：(i) 几何注意力块616, 以及(ii) 折叠块618, 接下来将更详细地描述这些块中的每个块。

[0232] 几何注意力块616使用明确推理蛋白质结构中氨基酸的3-D几何形状(即, 如由结构参数定义的几何形状)的“几何”自注意力操作来更新当前的单嵌入。更具体地, 为了更新给定的单嵌入, 几何注意力块616确定给定的单嵌入与一个或多个所选单嵌入中的每个单嵌入之间的相应注意力权重, 其中注意力权重取决于当前单嵌入、当前结构参数和成对嵌入。然后, 几何注意力块616使用以下各项来更新给定的单嵌入：(i) 注意力权重、(ii) 所选单嵌入、以及(iii) 当前结构参数。

[0233] 为了确定注意力权重, 几何注意力块616处理每个当前单嵌入以生成对应的“符号查询”嵌入、“符号键”嵌入和“符号值”嵌入。例如, 针对与第*i*个氨基酸相对应的单嵌入 h_i , 几何注意力块616可以生成符号查询嵌入 q_i 、符号键嵌入 k_i 和符号值嵌入 v_i , 如下所示：

$$q_i = \text{Linear}(h_i) \quad (10)$$

$$[0234] \quad k_i = \text{Linear}(h_i) \quad (11)$$

$$v_i = \text{Linear}(h_i) \quad (12)$$

[0235] 其中 $\text{Linear}(\cdot)$ 是指具有独立学习到的参数值的线性层。

[0236] 几何注意力块616另外处理每个当前单嵌入以生成对应的“几何查询”嵌入、“几何键”嵌入和“几何值”嵌入。每个单嵌入的几何查询嵌入、几何键嵌入和几何值嵌入都是最初在对应氨基酸的局部参考系中生成并且然后使用该氨基酸的结构参数被旋转和平移到全局参考系的3-D点。例如, 针对与第*i*个氨基酸相对应的单嵌入 h_i , 几何注意力块616可以生成几何查询嵌入 q_i^p 、几何键嵌入 k_i^p 和几何值嵌入 v_i^p , 如下所示：

$$q_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (13)$$

$$[0237] \quad k_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (14)$$

$$v_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i \quad (15)$$

[0238] 其中 $\text{Linear}_p(\cdot)$ 是指具有独立学习到的参数值的线性层, 这些线性层将 h_i 投影到3-D点(上标*p*指示该量是3-D点), R_i 表示由第*i*个氨基酸的旋转参数指定的旋转矩阵, 并且 t_i 表示第*i*个氨基酸的位置参数。

[0239] 为了更新与氨基酸*i*相对应的单嵌入 h_i , 几何注意力块616可以生成注意力权重 $[a_j]_{j=1}^N$, 其中*N*是蛋白质中氨基酸的总数, 并且 a_j 是氨基酸*i*与氨基酸*j*之间的注意力权重, 如下所示：

$$[0240] \quad [a_j]_{j=1}^N = \text{softmax} \left(\left[\frac{q_i \cdot k_j}{\sqrt{m}} + \alpha |q_i^p - k_j^p|_2^2 + (b_{i,j} \cdot w) \right]_{j=1}^N \right) \quad (16)$$

[0241] 其中 q_i 表示氨基酸*i*的符号查询嵌入, k_j 表示氨基酸*j*的符号键嵌入, m 表示 q_i 和 k_j 的维数, α 表示学习到的参数, q_i^p 表示氨基酸*i*的几何查询嵌入, k_j^p 表示氨基酸*j*的几何键嵌入, $|\cdot|_2$ 是 L_2 范数, 并且 $b_{i,j}$ 是与包括氨基酸*i*和氨基酸*j*的氨基酸对相对应的成对嵌入116, 并且 w 是学习到的权重向量(或某个其他学习到的投影操作)。

[0242] 一般来说,氨基酸对的成对嵌入隐式编码了有关该对中的氨基酸之间的关系(例如,该对中的氨基酸之间的距离)的信息。通过部分地基于氨基酸*i*和*j*的成对嵌入来确定氨基酸*i*与氨基酸*j*之间的注意力权重,折叠神经网络600利用来自成对嵌入的信息丰富了注意力权重,并且从而提高经预测折叠结构的准确性。

[0243] 在一些实现方式中,几何注意力块616生成多组几何查询嵌入、几何键嵌入和几何值嵌入,并使用每个生成的几何嵌入集来确定注意力权重。

[0244] 在生成与氨基酸*i*相对应的单嵌入 h_i 的注意力权重之后,几何注意力块616使用注意力权重来更新单嵌入 h_i 。具体地,几何注意力块616使用注意力权重来生成“符号返回”嵌入和“几何返回”嵌入,并且然后使用符号返回嵌入和几何返回嵌入来更新单嵌入。几何注意力块124可以生成氨基酸*i*的符号返回嵌入 o_i ,例如,如下所示:

$$[0245] \quad o_i = \sum_j a_j v_j \quad (17)$$

[0246] 其中 $[a_j]_{j=1}^N$ 表示(例如,参考方程(16)定义的)注意力权重,并且每个 v_j 表示氨基酸*j*的符号值嵌入。几何注意力块616可以生成氨基酸*i*的几何返回嵌入 o_i^p ,例如,如下所示:

$$[0247] \quad o_i^p = R_i^{-1} \cdot \left(\sum_j a_j v_j^p - t_i \right) \quad (18)$$

[0248] 其中几何返回嵌入 o_i^p 是3-D点, $[a_j]_{j=1}^N$ 表示(例如,参考方程(16)定义的)注意力权重, R_i^{-1} 是由氨基酸*i*的旋转参数指定的旋转矩阵的逆,并且 t_i 是氨基酸*i*的位置参数。应当理解,几何返回嵌入最初在全局参考系中生成,并且然后被旋转和平移到对应氨基酸的局部参考系中。

[0249] 几何注意力块616可以使用(例如,根据方程(17)生成的)对应的符号返回嵌入 o_i 和(例如,根据方程(18)生成的)几何返回嵌入 o_i^p 来更新氨基酸*i*的单嵌入 h_i ,例如,如下所示:

$$[0250] \quad h_i^{next} = \text{LayerNorm} \left(h_i + \text{Linear}(o_i, o_i^p, |o_i^p|) \right) \quad (19)$$

[0251] 其中 h_i^{next} 是氨基酸*i*的经更新单嵌入, $|\cdot|$ 是范数,例如 L_2 范数,并且 $\text{LayerNorm}(\cdot)$ 表示层归一化操作,例如,参考以下文献描述的:J.L. Ba、J.R. Kiros、G.E. Hinton的“Layer Normalization”,arXiv:1607.06450 (2016)。

[0252] 使用具体的3-D几何嵌入来更新氨基酸的单嵌入606(例如,如参考方程(13)-(15)所述)使得几何注意力块616能够在更新单嵌入时推理3-D几何。此外,每个更新块以对整体蛋白质结构的旋转和平移保持不变的方式来更新单嵌入和结构参数。例如,对被提供给折叠神经网络600的初始结构参数应用相同的全局旋转和平移操作将致使折叠神经网络600生成以相同方式全局旋转和平移但在其他方面保持不变的经预测结构。因此,对初始结构参数应用全局旋转和平移操作不会影响折叠神经网络600从初始结构参数开始生成的经预测蛋白质结构的准确性。由折叠神经网络600生成的表示的旋转和平移不变性有助于训

练,例如,因为折叠神经网络600自动学习以在蛋白质结构的所有旋转和平移上进行泛化。

[0253] 氨基酸的经更新单嵌入可以由几何注意力块616中的一个或多个附加神经网络层(例如,线性神经网络层)进一步变换,然后被提供给折叠块618。

[0254] 在几何注意力块616更新氨基酸的当前单嵌入606之后,折叠块618使用经更新单嵌入612来更新当前结构参数608。例如,折叠块618可以更新氨基酸i的当前位置参数 t_i ,如下所示:

$$[0255] \quad t_i^{next} = t_i + \text{Linear}(h_i^{next}) \quad (20)$$

[0256] 其中 t_i^{next} 是经更新位置参数, $\text{Linear}(\cdot)$ 表示线性神经网络层,并且 h_i^{next} 表示氨基酸i的经更新单嵌入。在另一个示例中,氨基酸i的旋转参数 R_i 可以指定旋转矩阵,并且折叠块618可以更新当前旋转参数 R_i ,如下所示:

$$w_i = \text{Linear}(h_i^{next}) \quad (21)$$

$$[0257] \quad R_i^{next} = R_i \cdot \text{QuaternionToRotation}(1 + w_i) \quad (22)$$

[0258] 其中 w_i 是三维向量, $\text{Linear}(\cdot)$ 是线性神经网络层, h_i^{next} 是氨基酸i的经更新单嵌入, $1 + w_i$ 表示实部为1、虚部为 w_i 的四元数,并且 $\text{QuaternionToRotation}(\cdot)$ 表示将四元数变换为等效的 3×3 旋转矩阵的操作。使用方程(21) - (22)来更新旋转参数确保了经更新旋转参数定义有效的旋转矩阵,例如行列式为一的正交矩阵。

[0259] 折叠神经网络600可以提供由最终更新块610生成的经更新结构参数,作为定义经预测蛋白质结构108的最终结构参数106。折叠神经网络600可以包括任意适当数量的更新块,例如5个更新块、25个更新块或125个更新块。可选地,折叠神经网络的更新块中的每个更新块可以共享在折叠神经网络的训练期间联合更新的一个参数值集。在更新块610之间共享参数值减少了折叠神经网络的可训练参数的数量,并且因此可以促进折叠神经网络的有效训练,例如,通过使训练稳定并降低过拟合的可能性。

[0260] 在训练期间,训练引擎可以基于评估最终结构参数106的准确性的结构损失来训练结构预测系统的参数,包括折叠神经网络600的参数,如上所述。在一些实现方式中,训练引擎可以进一步评估在最终更新块(即,产生最终结构参数)之前的更新块610中的一个或多个更新块的辅助结构损失。更新块的辅助结构损失评估由更新块生成的经更新结构参数的准确性。

[0261] 可选地,在训练期间,训练引擎可以应用“停止梯度(stop gradient)”操作,以防止梯度通过每个更新块的某些神经网络参数(例如,用于计算经更新旋转参数的神经网络参数(如方程(21) - (22)所述))反向传播。应用这些停止梯度操作可以提高在训练期间计算的梯度的数值稳定性。

[0262] 一般来说,由折叠神经网络600生成的经预测蛋白质结构108与对应的基准真实蛋白质结构之间的相似性可以例如通过为经预测蛋白质结构中的多个原子中的每个原子指派相应的准确性分数的相似性度量来衡量。例如,相似性度量可以为经预测蛋白质结构中的每个 α 碳原子指派相应的准确性分数。经预测蛋白质结构中的原子的准确性分数可以表征经预测蛋白质结构中的原子的位置与基准真实蛋白质结构中的原子的实际位置的吻合程度。可以将经预测蛋白质结构与基准真实蛋白质结构进行比较以生成经预测蛋白质结构

中的原子的准确性分数的相似性度量的示例是参考以下文献描述的1DDT相似性度量:V. Mariani等人“1DDT: a local superposition-free score for comparing protein structures and models using distance difference tests [1DDT:使用距离差值测试比较蛋白质结构和模型的局部无叠加分数]”,生物信息学,2013年11月1日;29(21) 2722-2728。

[0263] 折叠神经网络600可以被配置为针对经预测蛋白质结构108中的一个或多个原子中的每个原子生成相应的置信度估计650。经预测蛋白质结构中的原子的置信度估计650表征经预测蛋白质结构中的原子的经预测准确性分数(例如,1DDT准确性分数),即,该经预测准确性分数将通过将经预测蛋白质结构与(可能未知的)基准真实蛋白质结构进行比较的相似性度量生成。在一个示例中,经预测蛋白质结构中的原子的置信度估计650可以定义区间集上的离散概率分布,这些区间形成原子准确性分数的可能值范围的分区。离散概率分布可以将相应的概率与区间中的每个区间相关联,这定义了实际准确性分数被包括在该区间中的可能性。例如,准确性分数的可能值范围可以是[0, 100],并且置信度估计650可以定义区间{[0, 2), [2, 4), ..., [98, 100]}集上的概率分布。在另一个示例中,经预测蛋白质结构中的原子的置信度估计650可以是数值,即,直接预测原子的准确性分数。

[0264] 在一些实现方式中,折叠神经网络600为蛋白质的每个氨基酸中的指定原子(例如, α 碳原子)生成相应的置信度估计650。折叠神经网络600可以例如通过使用一个或多个神经网络层(例如,全连接层)对由折叠神经网络中的最后一个更新块生成的氨基酸的经更新单嵌入进行处理来生成蛋白质中的氨基酸中的指定原子的置信度估计650。

[0265] 结构预测系统可以基于经预测蛋白质结构中的原子的置信度估计650来生成与蛋白质中的每个氨基酸相对应的相应置信度分数。例如,结构预测系统可以生成氨基酸的置信度分数,作为氨基酸中 α 碳原子的准确性分数的可能值上的概率分布的预期值。

[0266] 结构预测系统可以生成整个经预测结构的置信度分数,例如,作为蛋白质中的氨基酸的置信度分数的平均值。

[0267] 在结构预测系统的训练期间,训练引擎可以通过反向传播辅助损失的梯度来调整结构预测系统的参数值,该辅助损失衡量以下两项之间的误差:(i)由折叠神经网络600生成的置信度估计值,以及(ii)通过将经预测蛋白质结构与基准真实蛋白质结构进行比较生成的准确性分数。该误差可以是例如交叉熵误差。

[0268] 由结构预测系统生成的置信度估计可以以多种方式使用。例如,经预测蛋白质结构中的原子的置信度估计可以指示结构的哪些部分已经得到可靠估计,因此适合进一步的下游处理或分析。作为另一个示例,每个蛋白质的置信度分数可以被用于对蛋白质结构的预测集(例如,由同一结构预测系统通过处理表征同一蛋白质的不同输入生成的预测,或由不同的结构预测系统生成的预测)进行排名。

[0269] 由结构参数106指定的位置参数和旋转参数可以定义蛋白质氨基酸中的主链原子的空间位置(例如,以 $[x, y, z]$ 笛卡尔坐标表示)。然而,结构参数106不一定定义蛋白质氨基酸中的剩余原子(例如氨基酸的侧链中的原子)的空间位置。具体地,氨基酸中的剩余原子的空间位置取决于氨基酸中的键之间的扭转角的值,例如, ω 角(omega-angle)、 ϕ 角(phi-angle)、 ψ 角(psi-angle)、 χ_1 角(chi1-angle)、 χ_2 角(chi2-angle)、 χ_3 角(chi3-angle)和 χ_4 角(chi4 angle),如参考图13所示。

[0270] 可选地,折叠神经网络600的更新块610中的一个或多个更新块可以生成定义蛋白质的每个氨基酸中每个原子的相应经预测空间位置的输出。为了生成氨基酸中原子的经预测空间位置,更新块可以使用一个或多个神经网络层处理氨基酸的经更新单嵌入,以生成氨基酸中的原子之间的键的扭转角的经预测值。神经网络层可以是例如嵌入有残差连接的全连接神经网络层。每个扭转角都可以被表示为例如2-D向量。

[0271] 更新块可以基于以下各项来确定氨基酸中的原子的空间位置:(i)氨基酸的扭转角的值,以及(ii)氨基酸的经更新结构参数(例如,位置参数和旋转参数)。例如,更新块可以根据预定义函数来处理扭转角,以生成氨基酸中的原子在氨基酸的局部参考系中的空间位置。更新块可以通过根据氨基酸的经更新结构参数旋转和平移原子的空间位置来生成氨基酸中的原子在全局参考系(即,对蛋白质中的所有氨基酸而言都是共同的)中的空间位置。例如,更新块可以通过以下方式来确定原子在全局参考系中的空间位置:将由经更新旋转参数定义的旋转操作应用于原子在局部参考系中的空间位置以生成旋转后的空间位置,并且然后将由经更新位置参数定义的平移操作应用于旋转后的空间位置。

[0272] 在一些实现方式中,作为输出最终结构参数的替代或组合,折叠神经网络600输出由最终更新块生成的蛋白质氨基酸中的原子的经预测空间位置。

[0273] 参考图12描述的折叠神经网络600在此被表征为接收基于(例如,如参考图8所述)由嵌入神经网络生成的MSA表示114和成对嵌入116的输入。然而,一般来说,折叠神经网络的输入(例如,单嵌入602和成对嵌入116)可以使用任何适当的技术来生成。此外,由折叠神经网络执行的操作的各个方面(例如,预测蛋白质的每个氨基酸中的原子的空间位置)可以由其他折叠神经网络(例如具有接收不同输入的不同架构)执行。

[0274] 图13示出了氨基酸中的键之间的扭转角,例如 ω 角、 ϕ 角、 ψ 角、 χ_1 角、 χ_2 角、 χ_3 角、 χ_4 角和 χ_5 角。

[0275] 图14是未折叠蛋白质和经折叠蛋白质的图示。未折叠蛋白质是氨基酸的随机卷曲结构。未折叠蛋白质经历蛋白质折叠并折叠成3D构型。蛋白质结构通常包括稳定的局部折叠模式,诸如 α 螺旋(例如802所示)和 β 折叠片。

[0276] 图15是用于预测包括一条或多条链的蛋白质结构的示例过程900的流程图,其中每条链指定一个氨基酸序列。为了方便起见,过程900将被描述为由位于一个或多个位置中的一个或多个计算机的系统来执行。例如,蛋白质结构预测系统(例如,图1的蛋白质结构预测系统100)在根据本说明书适当编程后可以执行过程900。

[0277] 该系统获得表示与蛋白质中的每条链相对应的相应多序列比对(MSA)的初始MSA表示(902)。

[0278] 该系统获得蛋白质中的每个氨基酸对的相应初始成对嵌入(904)。

[0279] 该系统使用嵌入神经网络处理包括初始MSA表示和初始成对嵌入的输入,以生成包括最终MSA表示和蛋白质中的每个氨基酸对的相应最终成对嵌入的输出。

[0280] 嵌入神经网络包括更新块序列。每个更新块具有相应的更新块参数集,并且被配置为接收当前MSA表示和蛋白质中的每个氨基酸对的相应当前成对嵌入。每个更新块:(i)基于当前成对嵌入来更新当前MSA表示,并(ii)基于经更新MSA表示来更新当前成对嵌入(906)。

[0281] 该系统基于使用最终MSA表示、最终成对嵌入或两者来确定蛋白质的经预测结构

(908)。

[0282] 图16示出了用于生成蛋白质中的氨基酸链的MSA表示1010的示例过程1000。参考图1描述的蛋白质结构预测系统100可以实现过程1000的操作。

[0283] 为了生成蛋白质中的氨基酸链的MSA表示100,系统100获得蛋白质的MSA 1002,其可以包括例如数千个MSA序列。

[0284] 系统100将MSA序列集分为“核心”MSA序列集1004和“额外”MSA序列集1006。核心MSA序列集可以小于额外MSA序列集1006 (例如,小一个数量级)。系统100可以将MSA序列集划分为核心MSA序列1004和额外MSA序列1006,例如,通过随机选择预定数量的MSA序列作为核心MSA序列,并将剩余的MSA序列识别为额外MSA序列1006。

[0285] 对于每个额外MSA序列1006,系统100可以确定额外MSA序列与每个核心MSA序列1004之间的相应相似性度量(例如,基于汉明距离)。然后,系统100可以将每个额外MSA序列1006与跟该额外MSA序列1006最相似的对核心MSA序列1004相关联(即,根据相似性度量)。与核心MSA序列1004相关联的额外MSA序列集1006可以被称为“MSA序列簇”1008。也就是说,系统100确定与每个核心MSA序列1004相对应的相应MSA序列簇1008,其中与核心MSA序列1004相对应的MSA序列簇1008包括与核心MSA序列1004最相似的额外MSA序列集1006。

[0286] 系统100可以基于核心MSA序列和MSA序列簇1008生成蛋白质中的氨基酸链的MSA表示。MSA表示1010可以由 $M \times N$ 嵌入阵列表示,其中M是核心MSA序列的数量(即,使得每个核心MSA序列与MSA表示的相应行相关联),并且N是氨基酸链中的氨基酸的数量。MSA表示中的嵌入可以通过 $(i, j) \in \{(i, j): i = 1, \dots, M, j = 1, \dots, N\}$ 来索引。

[0287] 为了生成MSA表示1010中的位置 (i, j) 处的嵌入,系统100可以获得定义核心MSA序列i中位置j处的氨基酸的身份的嵌入(例如,独热嵌入)。系统100还可以基于与核心MSA序列i相对应的MSA序列簇1008中的额外MSA序列1006中的位置j处每种可能氨基酸出现的相对频率,确定可能氨基酸集上的概率分布。然后,系统100可以通过组合(例如,连结)以下各项来确定MSA表示中的位置 (i, j) 处的嵌入:(i) 定义核心MSA序列i中的位置j处的氨基酸的身份的嵌入,以及(ii) 与核心MSA序列i中的位置j相对应的可能氨基酸上的概率分布。

[0288] 在一些情况下,核心MSA序列中的一个或多个核心MSA序列的(基准真实)蛋白质结构可能是已知的。具体地,对于核心MSA序列中的一个或多个,核心MSA序列中的氨基酸中的键之间的扭转角值(例如, ω 角、 ϕ 角、 ψ 角等)可以是已知的。如果核心MSA序列i中的氨基酸的扭转角值是已知的,那么系统100可以至少部分基于核心MSA序列i中的氨基酸j的扭转角值来生成MSA表示中的位置 (i, j) 处的嵌入。例如,该系统可以使用一个或多个神经网络层生成扭转角值的嵌入,并将扭转角值的嵌入连结到MSA表示中的位置 (i, j) 处的嵌入。

[0289] 图17示出了用于为蛋白质中的每个氨基酸对生成(初始化)相应的成对嵌入112的示例过程1100。参考图1描述的蛋白质结构预测系统100可以实现过程1100的操作。

[0290] 系统100可以使用蛋白质的MSA表示1102来生成成对嵌入112。参考图7和图16更详细地描述了生成蛋白质的MSA表示。系统100可以基于蛋白质中每个氨基酸链的相应MSA表示来生成该蛋白质的MSA表示1102。为了生成蛋白质中氨基酸链的MSA表示,系统100可以使用比被用于生成参考图1描述的MSA表示110的MSA序列更多的MSA序列(例如,多一个数量级)。因此,由系统100用于生成成对嵌入112的MSA表示1102可以具有比参考图1描述的MSA

表示110更多的行(例如,多一个数量级)。在一些实现方式中,系统100可以使用参考图16描述的额外MSA序列1006来生成MSA表示1102。

[0291] 在生成MSA表示1102之后,系统100处理MSA表示1102以根据MSA表示1102生成成对嵌入1104,例如,通过对MSA表示1102应用外积均值运算,并将成对嵌入1104识别为外积均值运算的结果。

[0292] 系统100使用嵌入神经网络1106来处理MSA表示1102和成对嵌入1104。嵌入神经网络1106可以通过在MSA表示1102与成对嵌入1104之间共享信息来更新MSA表示1102和成对嵌入1104。更具体地,嵌入神经网络1106可以在基于成对嵌入1104更新MSA表示1102与基于MSA表示1102更新成对嵌入1104之间交替进行。

[0293] 嵌入神经网络1106可以具有基于参考图2至图5描述的嵌入神经网络架构的架构,即,使用按行自注意力块和按列自注意力块来更新成对嵌入1104和MSA表示1102。

[0294] 在一些实现方式中,嵌入神经网络1106可以使用按列“全局”自注意力操作来更新MSA表示1102的每一列中的嵌入。更具体地,嵌入神经网络1106可以将MSA表示的每一列中的嵌入提供给按列全局自注意力块,以生成当前MSA表示的每一列的经更新嵌入。为了实现全局按列自注意力,自注意力块可以为列中的每个嵌入生成相应的查询嵌入,并且然后对查询嵌入进行平均以生成该列的单个“全局”查询嵌入。然后,按列自注意力块使用单个全局查询嵌入来执行自注意力操作,这可以将自注意力操作的复杂度从二次方(即,每列的嵌入数量)降低到线性。使用全局自注意力操作可以降低按列自注意力操作的计算复杂度,从而能够在MSA表示1102的包括大量(例如,数千个)嵌入的列上执行按列自注意力操作。

[0295] 在使用嵌入神经网络1106更新成对嵌入1104和MSA表示1102之后,系统100可以将成对嵌入112识别为由嵌入神经网络1106生成的经更新成对嵌入。系统100可以丢弃由嵌入神经网络1106生成的经更新MSA表示,或者以任何适当的方式使用它。

[0296] 作为生成成对嵌入112的一部分,系统100可以包括与蛋白质中的每个氨基酸对相对应的相应成对嵌入中的相对位置编码信息。该系统可以通过以下方式将包括在同一氨基酸链中的氨基酸对的相对位置编码信息包括在对应的成对嵌入中:计算表示氨基酸链中分开这对氨基酸的之间的氨基酸数量的有符号差,将结果截取到预定义区间,使用独热编码向量来表示经截取的值,对独热编码向量应用线性变换,并将线性变换的结果添加到对应的成对嵌入中。该系统可以通过以下方式将未包括在同一氨基酸链中的氨基酸对的相对位置编码信息包括在对应的成对嵌入中:将默认编码向量添加到对应的成对嵌入中,以指示这对氨基酸未包括在同一氨基酸链中。

[0297] 系统100还可以至少部分地基于一个或多个模板序列1110的集来生成成对嵌入112。每个模板序列1110都是蛋白质中的氨基酸链的MSA序列,其中模板序列1110的经折叠结构是例如通过物理实验已知的。

[0298] 系统100可以生成每个模板序列1110的相应模板表示1112。模板序列1110的模板表示1112包括与模板序列中的每个氨基酸对相对应的相应嵌入,例如,使得长度为 n (即,具有 n 个氨基酸)的模板序列的模板表示1112可以被表示为 $n \times n$ 嵌入阵列。系统100基于例如以下各项来生成模板序列1110的模板表示1112中的位置 (i, j) 处的嵌入:(i)表示模板序列中的位置 i 和位置 j 处的氨基酸的身份的相应嵌入(例如,独热嵌入),(ii)由模板序列中(即,在模板序列的经折叠结构中)的位置 i 和 j 处的氨基酸中相应 α 碳原子的空间位置差

值定义的单位向量,其中单位向量是在氨基酸*i*或氨基酸*j*的参考系中计算的,以及(iii)模板序列中的位置*i*和*j*处的氨基酸中相应 α 碳原子的空间位置之间的距离的离散化/分仓(binned)表示。

[0299] 系统100可以使用一个或多个模板更新块1114的序列来处理每个模板表示,以生成与每个模板序列1110相对应的相应经更新模板表示1116。模板更新块可以包括例如按行自注意力块(例如,其更新模板表示的每一行中的嵌入)、按列自注意力块(例如,其更新模板表示的每一列中的嵌入)和转换块(例如,其将一个或多个神经网络层应用于模板表示中的嵌入中的每个嵌入)。

[0300] 在生成经更新模板表示1116之后,系统100使用经更新模板表示1116来更新成对嵌入112。例如,系统100可以对经更新模板表示1116的对应(*i, j*)位置处的嵌入使用“交叉注意力”来更新每个位置(*i, j*)处的相应成对嵌入112。在用于更新位置(*i, j*)处的成对嵌入112的交叉注意力操作中,查询嵌入由位置(*i, j*)处的成对嵌入生成,并且键嵌入和值嵌入由经更新模板表示1116的对应(*i, j*)位置处的嵌入生成。

[0301] 使用模板序列1110更新成对嵌入112使得系统100能够使用表征进化相关模板序列1110的蛋白质结构的信息来丰富成对嵌入,从而增强成对嵌入112的信息内容,并提高使用成对嵌入预测的蛋白质结构的准确性。

[0302] 如前所述,本说明书中描述的致病性预测系统120具有多种可能的应用,包括使用致病性分数来识别动物或植物受试者的疾病原因。

[0303] 在另一个示例应用中,该系统被用于获得用于有害生物防治的经突变蛋白质。这可以包括确定多个蛋白质突变(例如相对于参考氨基酸序列)的致病性分数,每个突变定义相应的经突变蛋白质。然后,可以基于这些突变的致病性分数选择这些突变中的一个突变来定义用于有害生物防治的经突变蛋白质。例如,可以对致病性分数进行排名,并可以根据分数来选择致病性最强的突变,例如,具有最高分数的突变。作为另一个示例,可以选择具有中等致病性分数的突变,例如以便促进突变在有害生物群体中传播。为了使用经突变蛋白质来进行有害生物防治,可以培育有害生物以产生所选的经突变蛋白质。

[0304] 在另一个示例应用中,该系统被用于筛选一种或多种活生物体,例如动物(包括人类)或植物,以确定是否存在具有致病性突变的蛋白质。这可以包括获得这些生物体中的每个生物体的蛋白质的氨基酸序列,例如通过先前描述的基因测序。一般来说,这些序列中的一些序列将对应蛋白质的经突变版本。对于所获得的包括突变(例如与参考蛋白质的参考氨基酸序列相比)的氨基酸序列中的每个氨基酸序列,该系统确定该突变的致病性分数。然后可以处理每个所确定的致病性分数以确定突变是否具有致病性,例如通过将致病性分数分类为良性或致病性。致病性分数或分类可以例如为关于是否或如何提供治疗的临床决策提供输入。

[0305] 在另一个示例应用中,该系统被用于确定细菌或病毒对活生物体的致病程度。这可以包括获得人造蛋白质的氨基酸序列,其中人造蛋白质是在活生物体内制造的,例如由活生物体或细菌制造,作为活生物体被细菌或病毒感染的结果。人造蛋白质可以是由细菌生成的毒素或由活生物体制造的蛋白质,例如病毒蛋白质,并且可以是生物体中天然存在的蛋白质的突变。可以确定天然存在的蛋白质的突变的致病性分数,并用该致病性分数来

确定细菌或病毒的致病程度。例如,致病性分数可以被分为良性或致病性,或者可以比较细菌或病毒的不同菌株的相对致病性。这可以用于例如确定群体中致病菌株的比例是增加还是减少,例如以触发警报;或根据所确定的致病程度来选择细菌或病毒以用于制造疫苗,例如以识别具有相对较低致病性的细菌或病毒菌株来制造疫苗。

[0306] 图18A和图18B示出了通过接受者操作者曲线下面积 (auROC) 评估的本说明书中描述的系统在临床策划的分类基准上的性能 (例如,分类准确性)。误差条示出了1000次自举重采样的95%置信区间。图表1802比较了当前系统 (“AlphaMissense”) 与其他预测器在留出的82,868个ClinVar错义变异上的性能。图表1804比较了当前系统 (“AlphaMissense”) 和其他预测器在来自202个驱动基因的热点中的868个癌症突变与1734个随机选择的负变异上的性能。图表1806比较了当前系统 (“AlphaMissense”) 和其他预测器在区分来自解密发育障碍队列 (DDD) 患者和健康对照的新发 (de novo) 变异方面的性能;考虑了215个DDD相关基因的353个患者变异和57个对照变异。图表1808比较了当前系统 (“AlphaMissense”) 和其他预测器在对进化受限程度较高的区域上的ClinVar变异 (3430个致病性变异和1185个良性变异) 进行分类方面的性能。

[0307] 本说明书在涉及系统和计算机程序组件时使用术语“配置”。对于将被配置为执行特定操作或动作的一个或多个计算机的系统,意味着该系统已经在其上安装了在操作中使该系统执行该操作或动作的软件、固件、硬件或它们的组合。对于将被配置为执行特定操作或动作的一个或多个计算机程序,意指一个或多个程序包括在由数据处理设备执行时致使设备执行操作或动作的指令。

[0308] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路系统中、在有形地体现的计算机软件或固件中、在计算机硬件 (包括本说明书中公开的结构及其结构等效物) 或在它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序,即,在有形的非暂时性存储介质上编码的计算机程序指令的一个或多个模块,以供数据处理设备执行或控制数据处理设备的操作。计算机存储介质可以是机器可读存储装置、机器可读存储基板、随机或串行存取存储器装置或它们中的一个或多个的组合。替代地或另外,可以在人工生成的传播信号 (例如,机器生成的电、光或电磁信号) 上对程序指令进行编码,该人工生成的传播信号被生成以对信息进行编码,以用于传输到合适的接收器设备以供数据处理设备执行。

[0309] 术语“数据处理设备”是指数据处理硬件,并且包括用于处理数据的所有种类的设备、装置和机器,例如包括可编程处理器、计算机或多个处理器或计算机。该设备还可以是或进一步包括专用逻辑电路系统,例如FPGA (现场可编程门阵列) 或ASIC (专用集成电路)。除了硬件之外,该设备还可以可选地包括为计算机程序创建执行环境的代码,例如,构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组的代码。

[0310] 计算机程序,其也可以被称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码,可以用任何形式的编程语言 (包括编译型或解释型语言或者声明式或过程式语言) 编写;并且它可以以任何形式部署,包括作为独立程序或作为模块、组件、子例程或适合在计算环境中使用的其他单元。程序可以但不必与文件系统中的文件相对应。程序可以存储在保存其他程序或数据 (例如存储在标记语言文档中的一个或多个脚本) 的文件的一部分中、存储在专用于所讨论的程序的单个文件中,或存储在多个协调文件 (例如存储一

个或多个模块、子程序或代码部分的文件)中。计算机程序可以被部署为在一个计算机上或在位于一个站点处或分布在多个站点上并通过数据通信网络互连的多个计算机上执行。

[0311] 在本说明书中,术语“引擎”广泛地用于指被编程为执行一个或多个特定功能的基于软件的系统、子系统或过程。一般地,引擎将被实现为安装在一个或多个位置中的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下,一个或多个计算机将专用于特定引擎;在其他情况下,多个引擎可以安装并运行在相同的一个或多个计算机上。

[0312] 本说明书中所描述的过程和逻辑流可以由一个或多个可编程计算机执行,该一个或多个可编程计算机执行一个或多个计算机程序以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路系统(例如,FPGA或ASIC)或由专用逻辑电路系统和一个或多个编程计算机的组合来执行。

[0313] 适合于执行计算机程序的计算机可以基于通用或专用微处理器或两者,或任何其他种类的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的基本元件是用于履行或执行指令的中央处理单元和用于存储指令和数据的一个或多个存储器装置。中央处理单元和存储器可以由专用逻辑电路系统补充或并入专用逻辑电路系统中。通常,计算机还将包括用于存储数据的一个或多个大容量存储装置,例如磁盘、磁光盘或光盘,或者操作性地耦合以从其接收数据或将数据传输到其或两者。然而,计算机不必具有此类装置。此外,计算机可以嵌入另一装置中,例如,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储装置,例如通用串行总线(USB)闪存驱动器,仅举几例。

[0314] 适于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器装置,包括例如半导体存储器装置(例如EPROM、EEPROM和闪存存储器装置)、磁盘(例如内部硬盘或可移动磁盘)、磁光盘以及CD-ROM和DVD-ROM盘。

[0315] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在具有用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户可以通过其向计算机提供输入的键盘和指向装置(例如,鼠标或轨迹球)的计算机上实现。也可以使用其他种类的装置来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感觉反馈,例如,视觉反馈、听觉反馈或触觉反馈;并且可以按任何形式接收来自用户的输入,包括声音、语音或触觉输入。另外,计算机可通过向用户使用的装置发送文档以及从用户使用的装置接收文档来与用户交互;例如,通过响应于从用户的装置上的web浏览器接收到的请求而向web浏览器发送网页。此外,计算机可以通过向个人装置(例如,运行消息传递应用的智能电话)发送文本消息或其他形式的消息,并且作为回应从用户接收响应消息来与用户交互。

[0316] 用于实现机器学习模型的数据处理设备还可以包括例如专用硬件加速器单元,用于处理机器学习训练或生产(即,推断、工作负载)的通用和计算密集部分。

[0317] 可以使用机器学习框架(例如,TensorFlow框架、Microsoft Cognitive Toolkit框架,Apache Singa框架或Apache MXNet框架)来实现和部署机器学习模型。

[0318] 本说明书中描述的主题的实施例可以在计算系统中实现,该计算系统包括后端组件(例如,作为数据服务器),或者包括中间件组件(例如,应用程序服务器),或者包括前端组件(例如,具有用户可以通过其与本说明书中描述的主题的实现方式交互的图形用户界

面、web浏览器或应用的客户端计算机),或者一个或多个此类后端、中间件或前端组件的任何组合。系统的组件可以通过任何形式或介质的数字数据通信(例如,通信网络)互连。通信网络的示例包括局域网(LAN)和广域网(WAN),例如互联网。

[0319] 计算系统可包括客户端和服务端。客户端和服务端一般彼此远离并且通常通过通信网络进行交互。客户端和服务端的关系凭借在相应的计算机上运行并使彼此具有客户端-服务端关系的计算机程序而产生。在一些实施例中,服务端将数据(例如,HTML页面)发送到用户装置,例如,用于向与充当客户端的装置进行交互的用户显示数据并从中接收用户输入的目的。可以在服务端处从装置接收在用户装置处生成的数据,例如,用户交互的结果。

[0320] 虽然本说明书包含许多特定的实现方式细节,但是这些细节不应被解释为对任何发明的范围或可能要求保护的范围的限制,而应被解释为对可能是特定发明的特定实施例所特有的特征的描述。在本说明书中在单独实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反,在单个实施例的上下文中描述的各种特征也可以在多个实施例中单独地或以任何合适的子组合来实现。此外,尽管特征可以在上面描述为以某些组合起作用并且甚至最初如此要求保护,但是在一些情况下可以从组合中删除来自所要求保护的组合的一个或多个特征,并且所要求保护的组合可以涉及子组合或子组合的变体。

[0321] 相似地,虽然在附图中描绘了操作并且在权利要求中以特定的次序叙述了操作,但是这不应理解为要求以所示的特定次序或以顺序次序来执行此类操作,或者要求执行所有示出的操作,来实现期望的结果。在某些情况中,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应理解为在所有实施例中都需要此类分离,并且应理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或封装到多个软件产品中。

[0322] 已经描述了本主题的特定实施例。其他实施例在所附权利要求的范围内。例如,权利要求中陈述的动作可以按不同的次序执行并仍实现期望结果。作为一个示例,附图描绘的过程不一定要求所示出的特定次序或者顺序次序来实现期望结果。在一些情况下,多任务处理和并行处理可以是有利的。

[0323] 本说明书主题的某些新颖方面在所附权利要求中进行了阐述,并在附录A中附有进一步的描述。

致病性预测系统

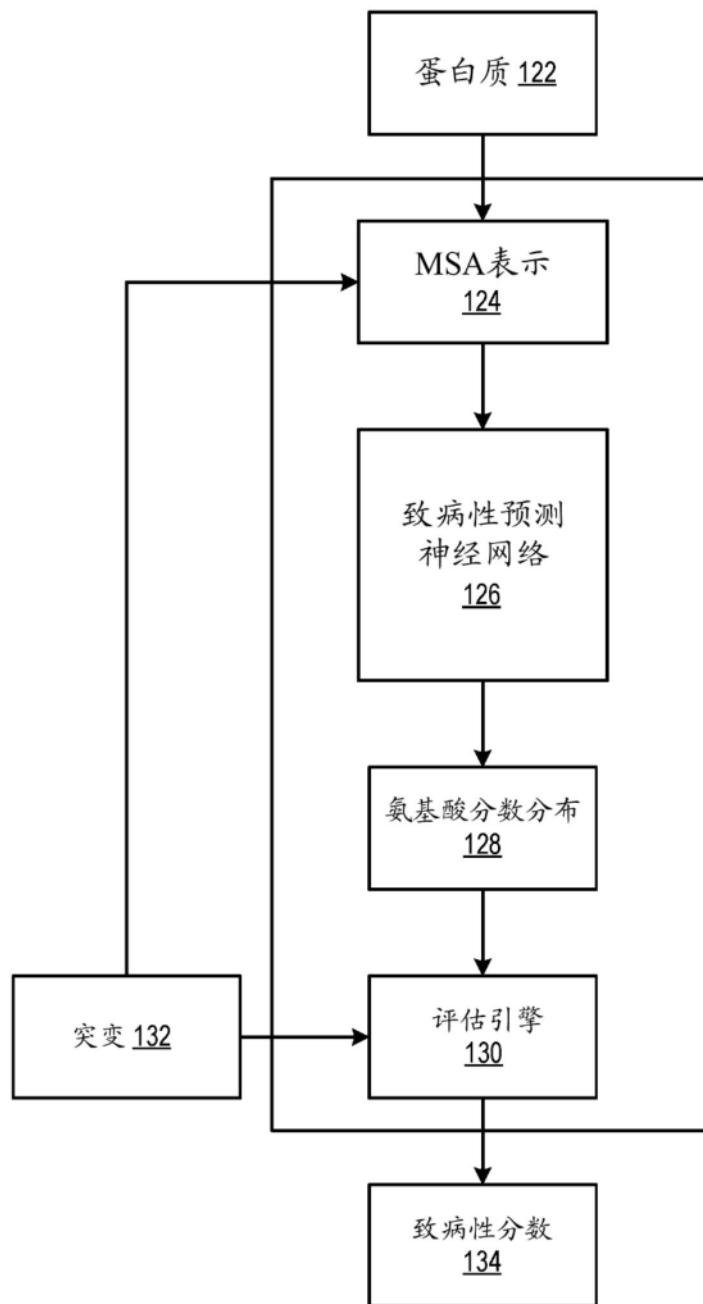
120

图1

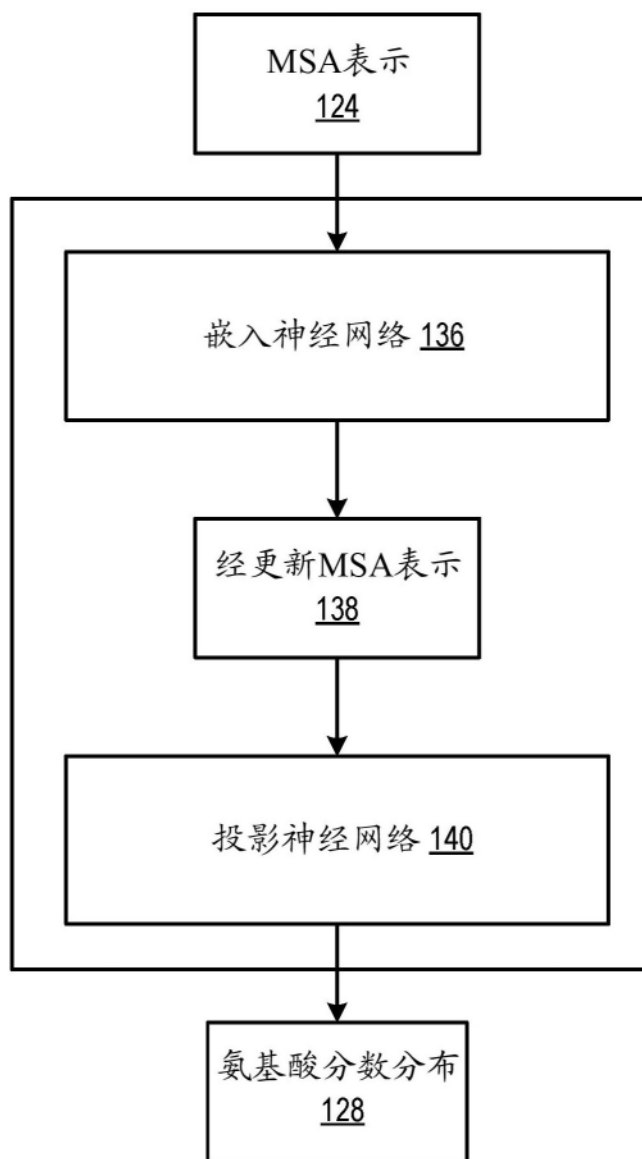
致病性预测神经网络 126

图2

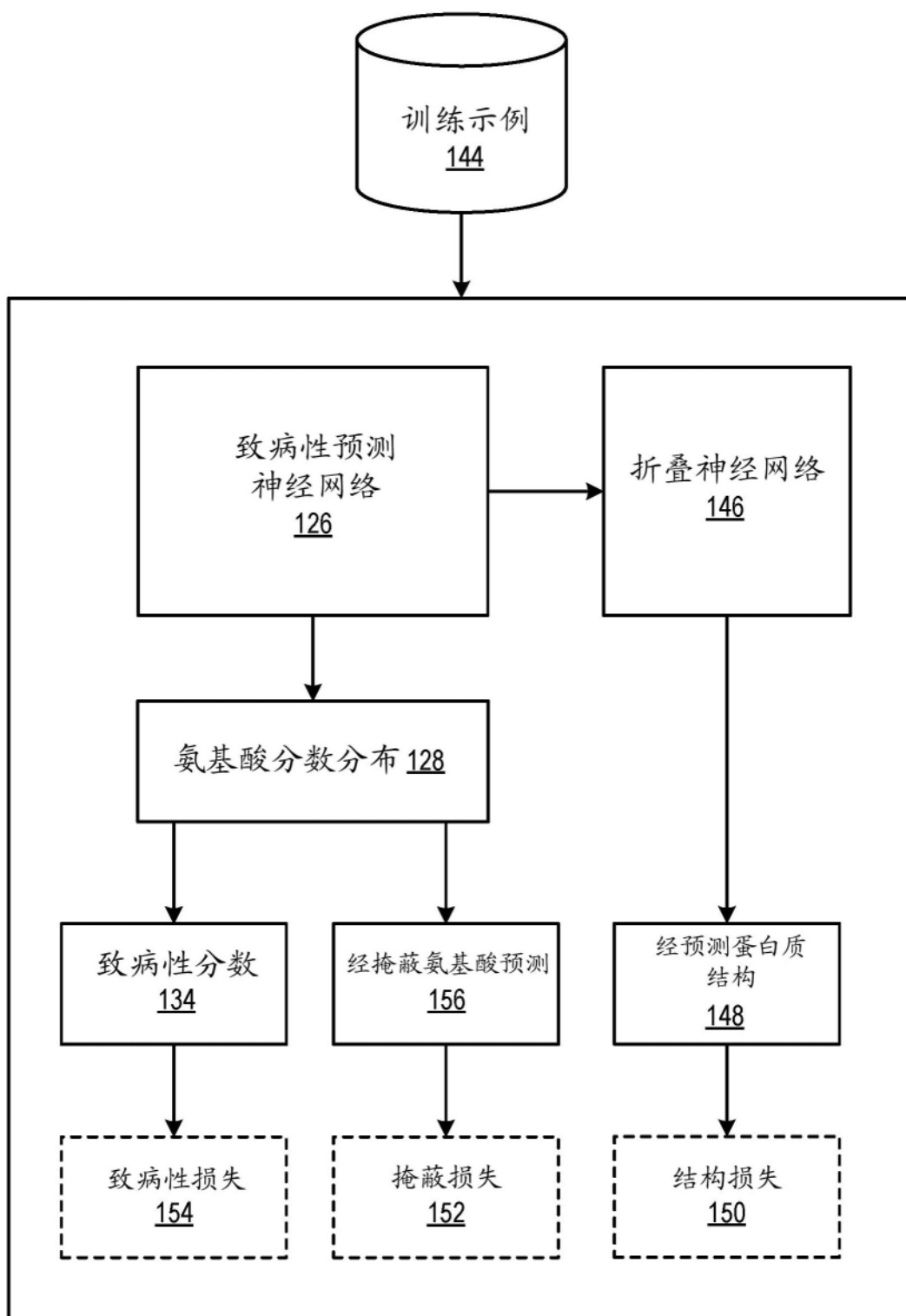
训练系统 142

图3

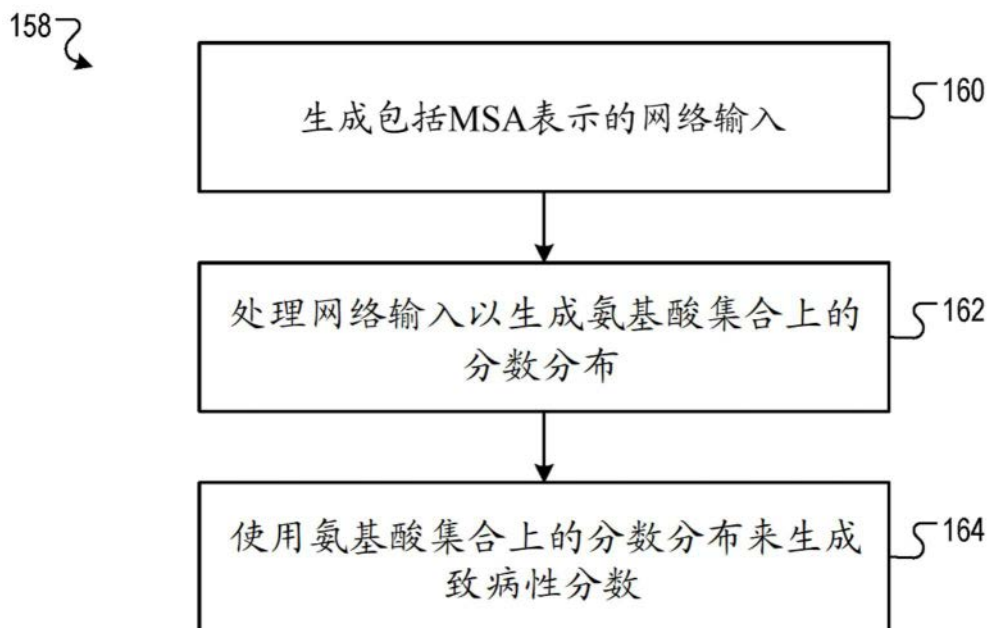


图4

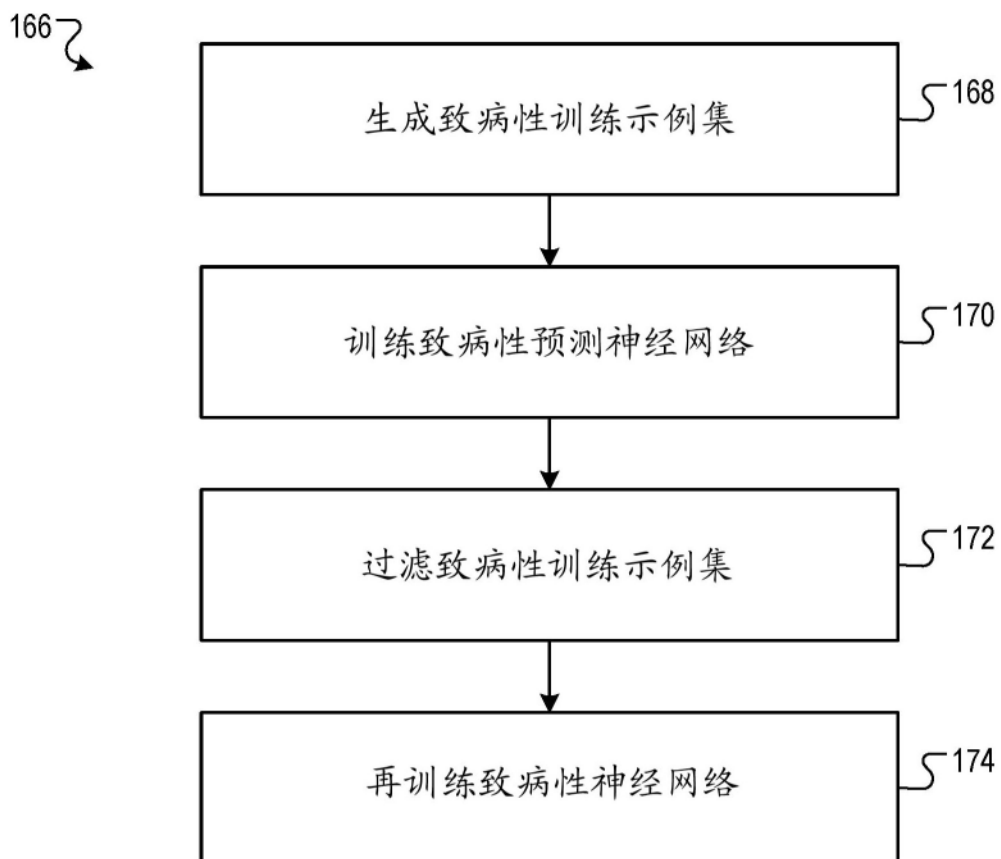


图5

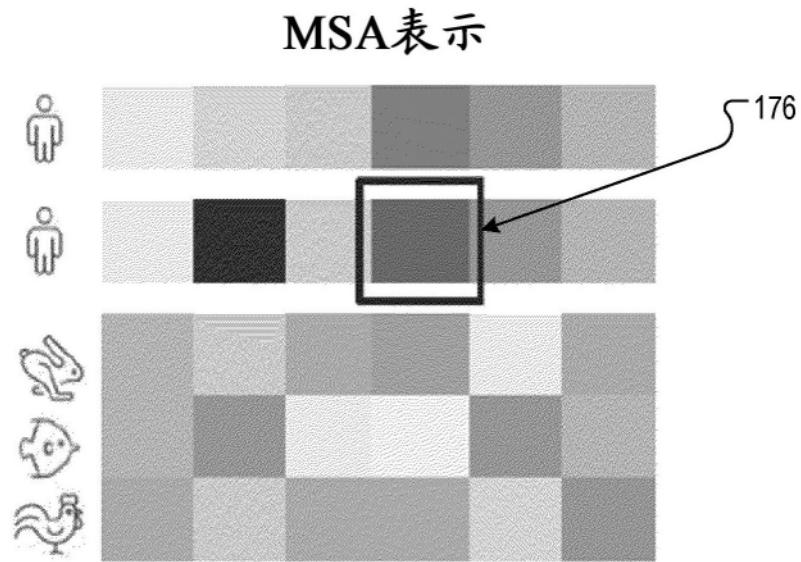


图6A



图6B

蛋白质结构预测系统

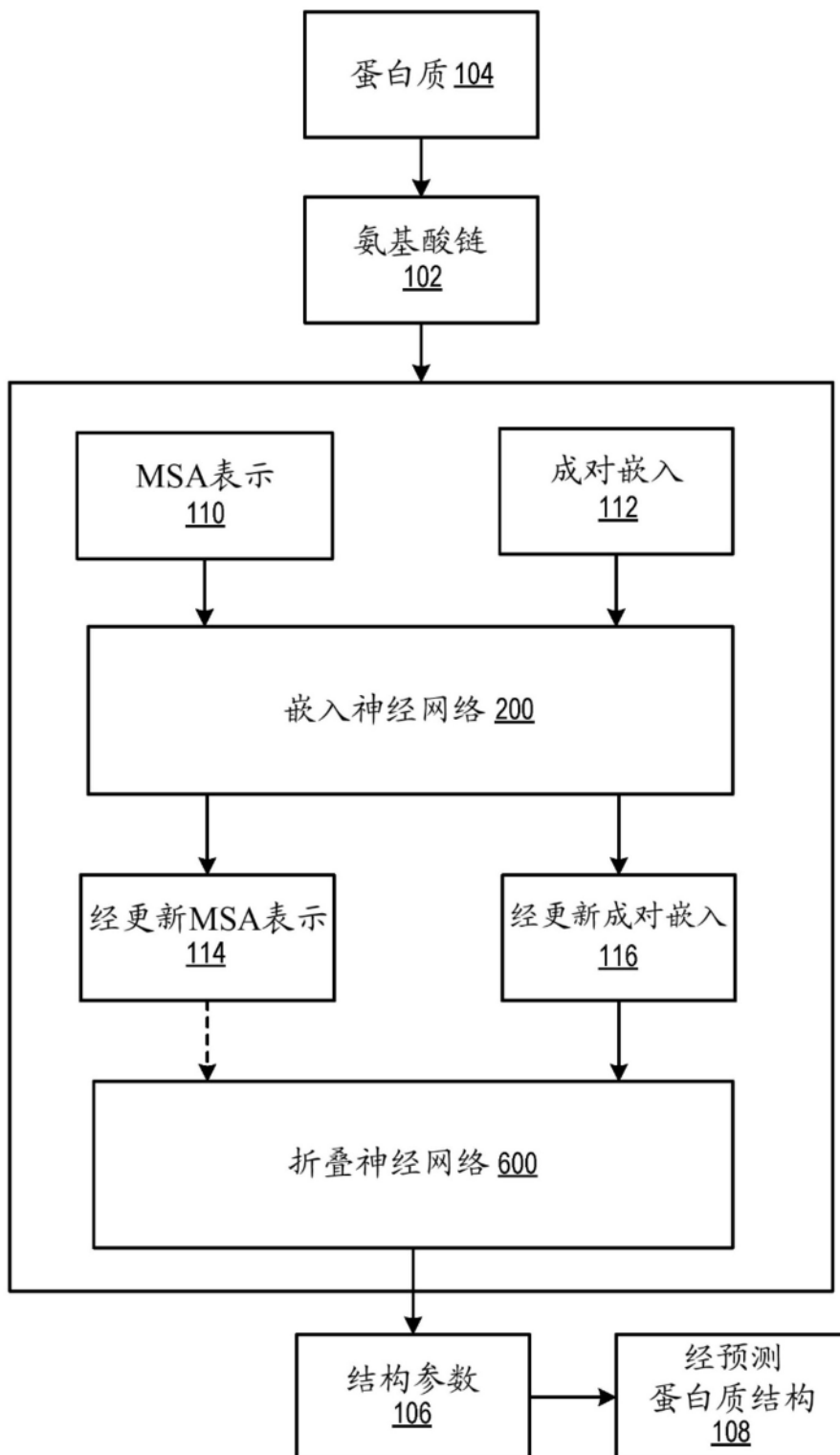
100

图7

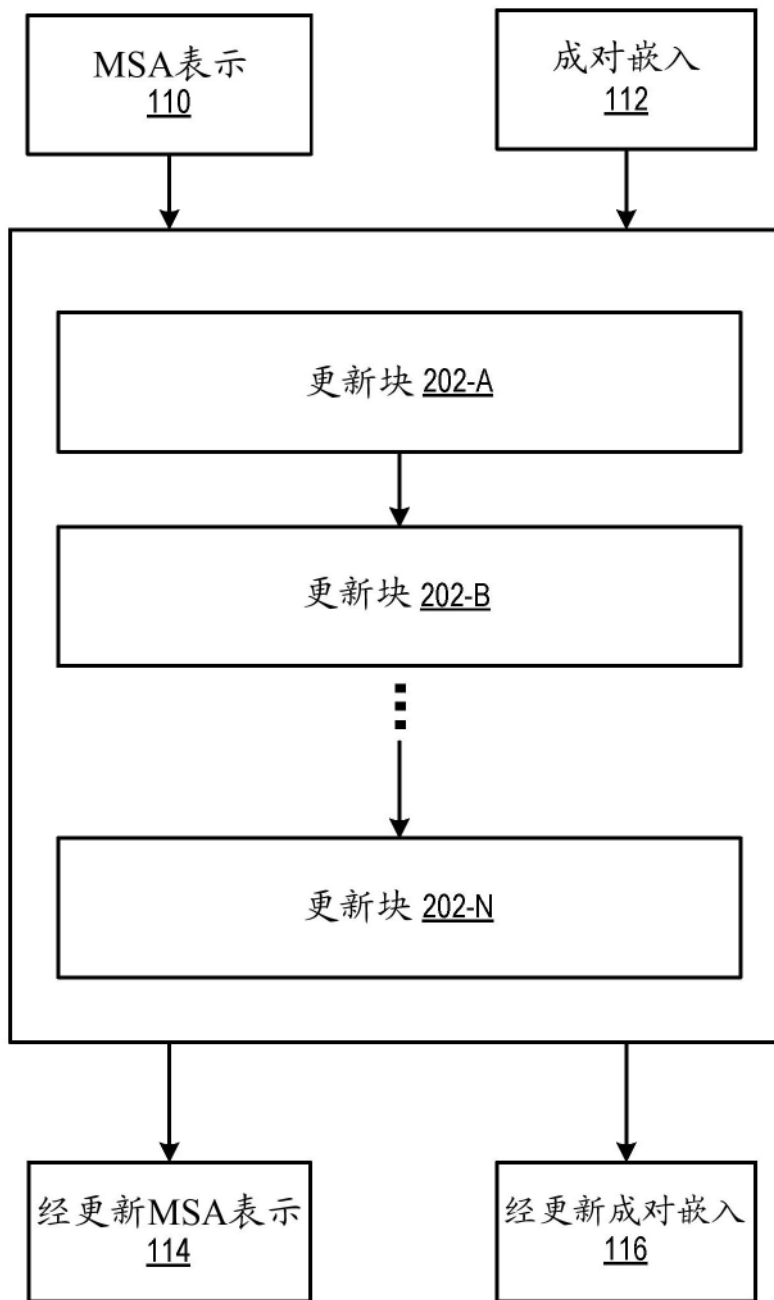
嵌入神经网络 200

图8

嵌入神经网络更新块

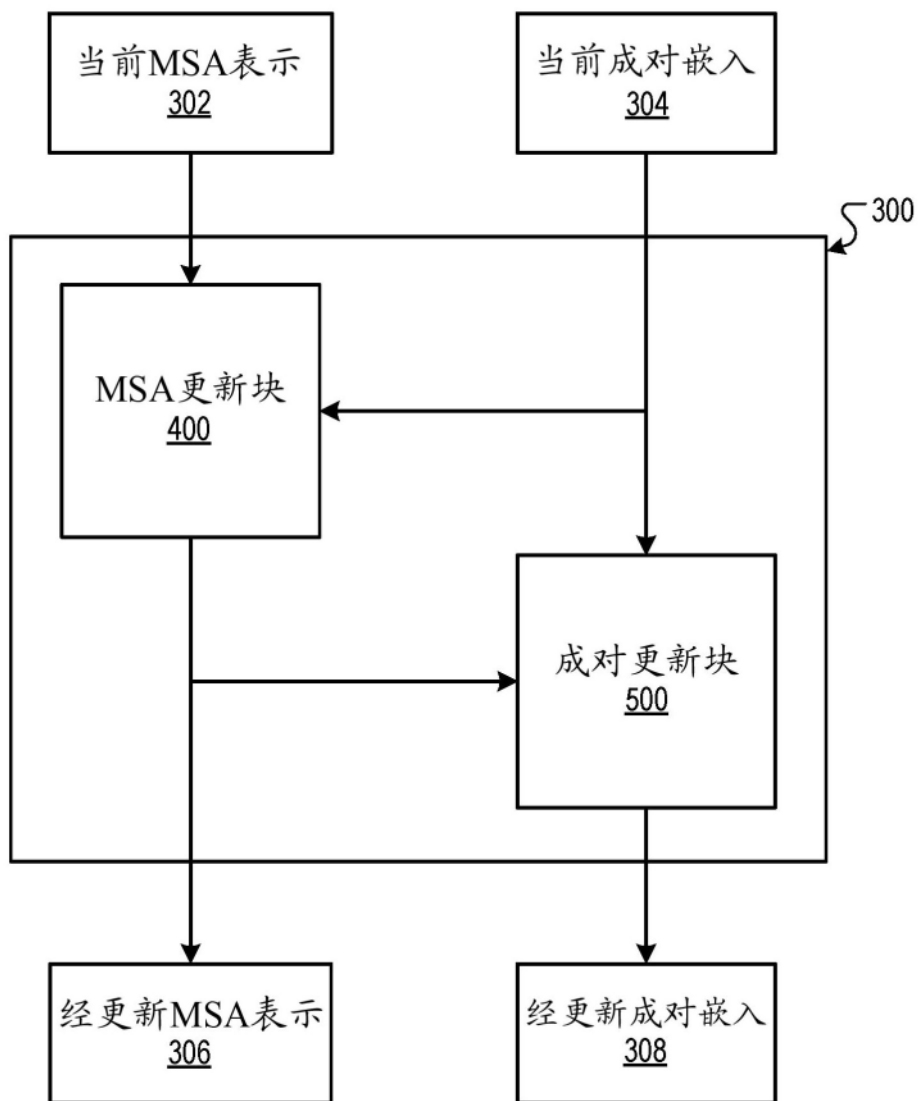


图9

MSA更新块 400

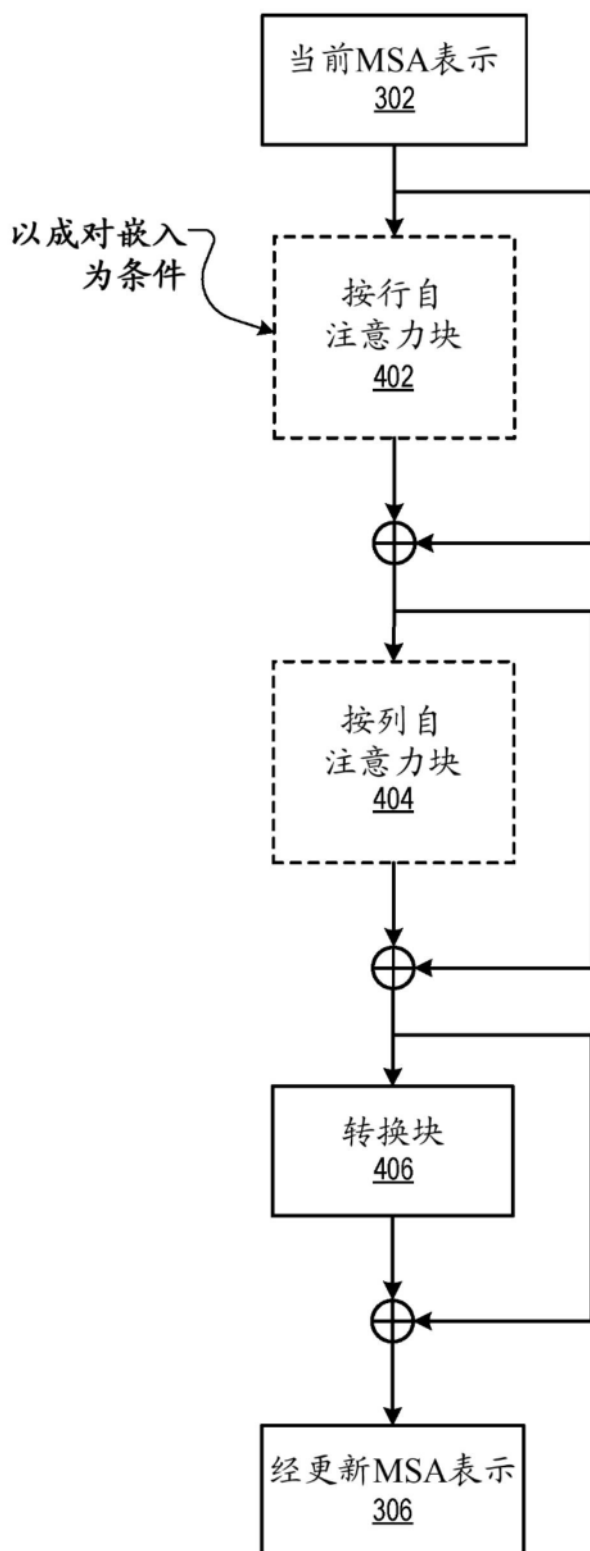


图10

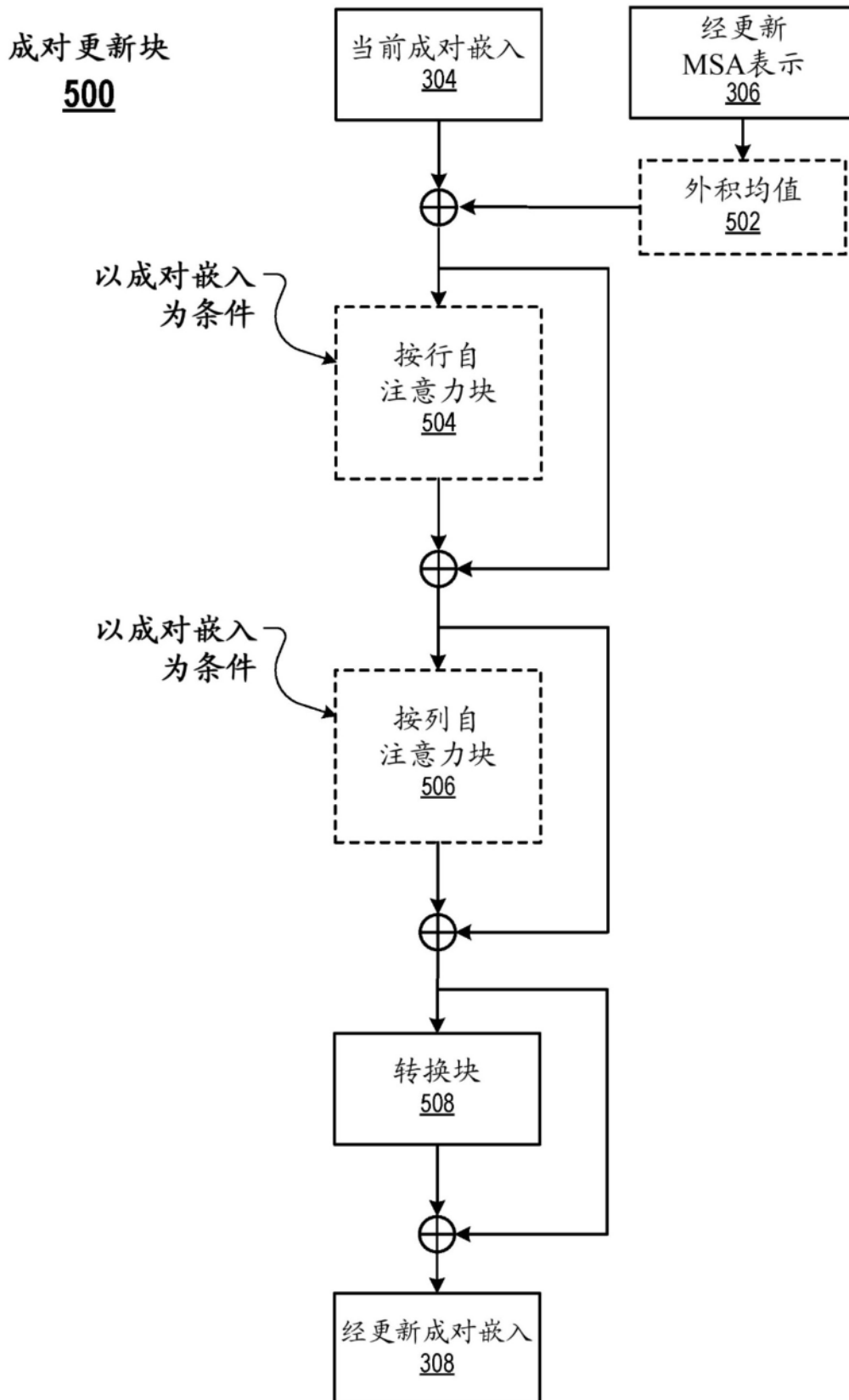


图11

折叠神经网络 600

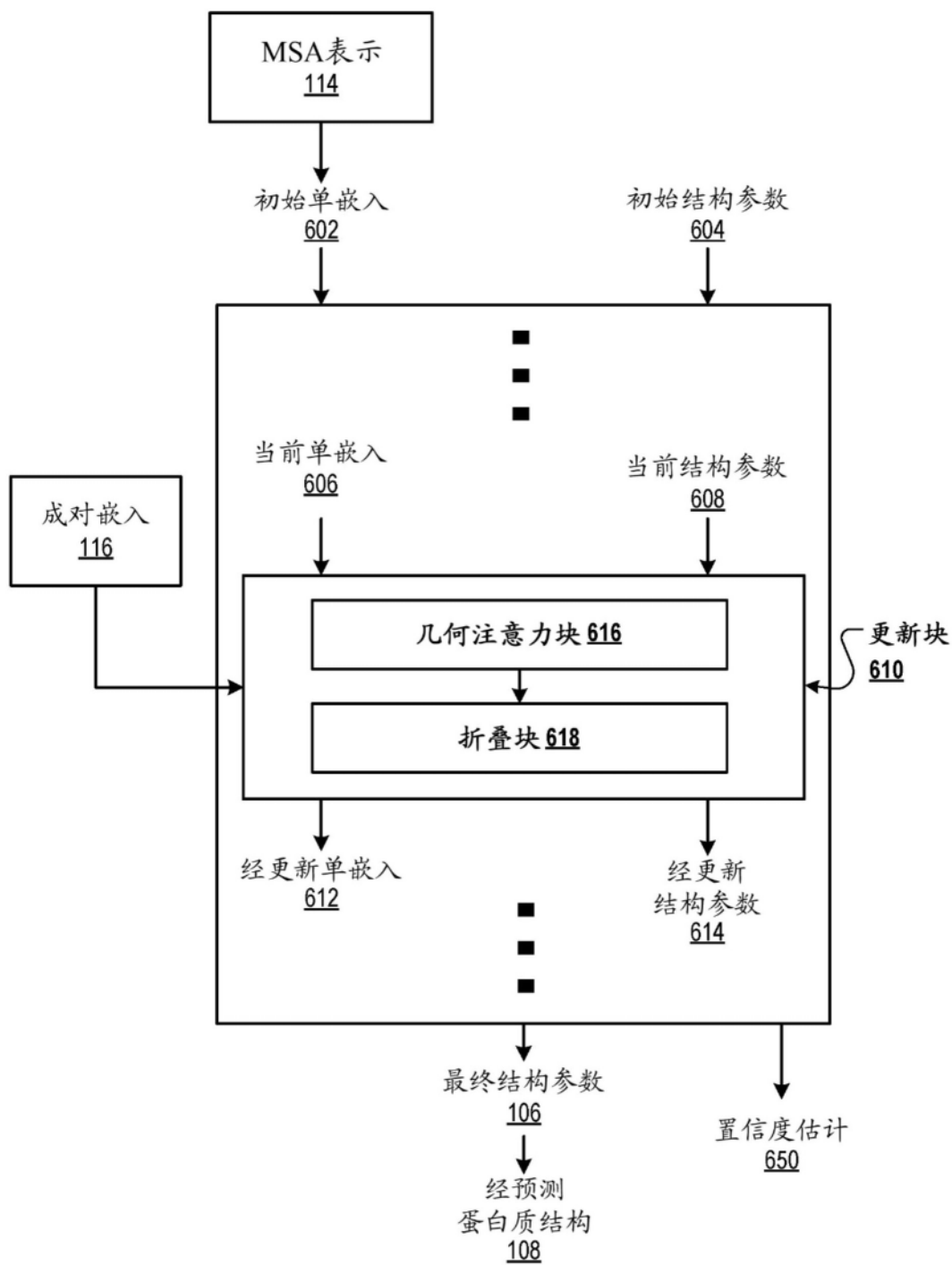


图12

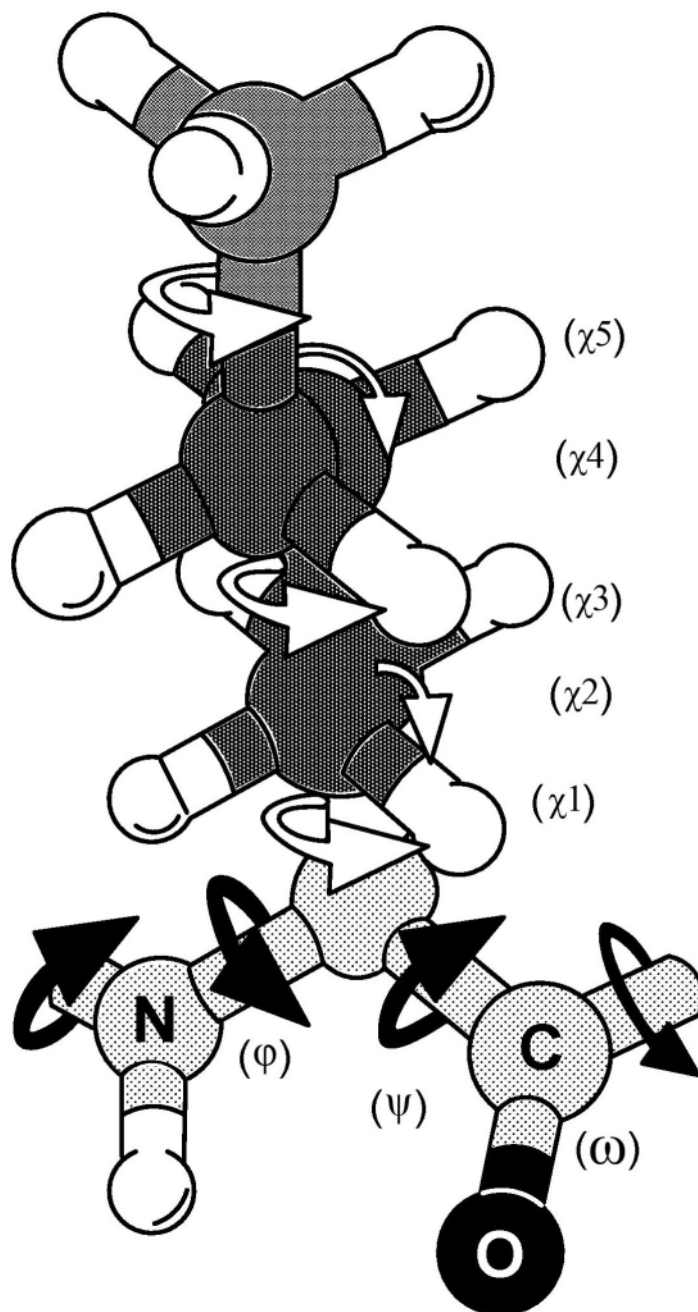


图13

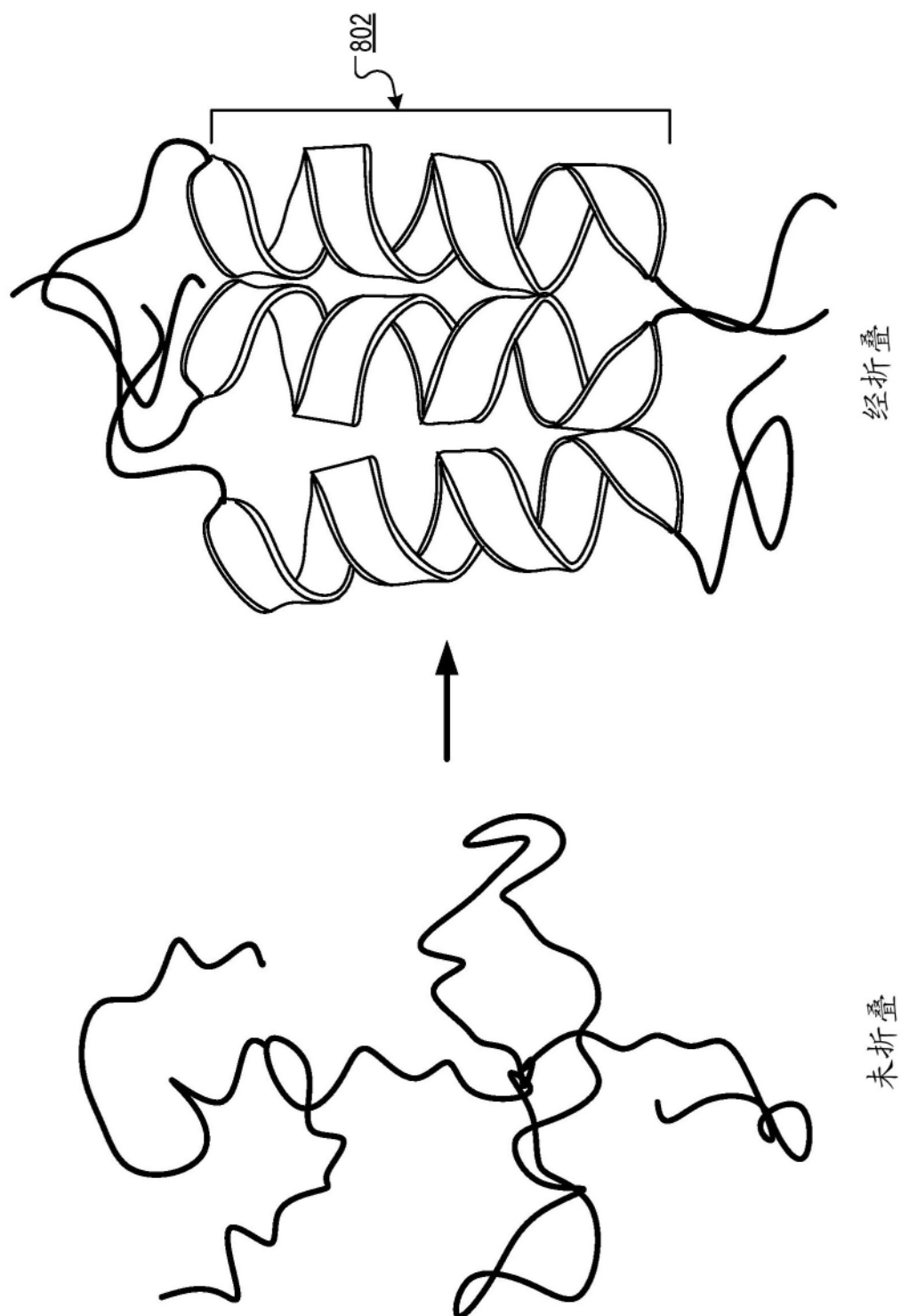


图14

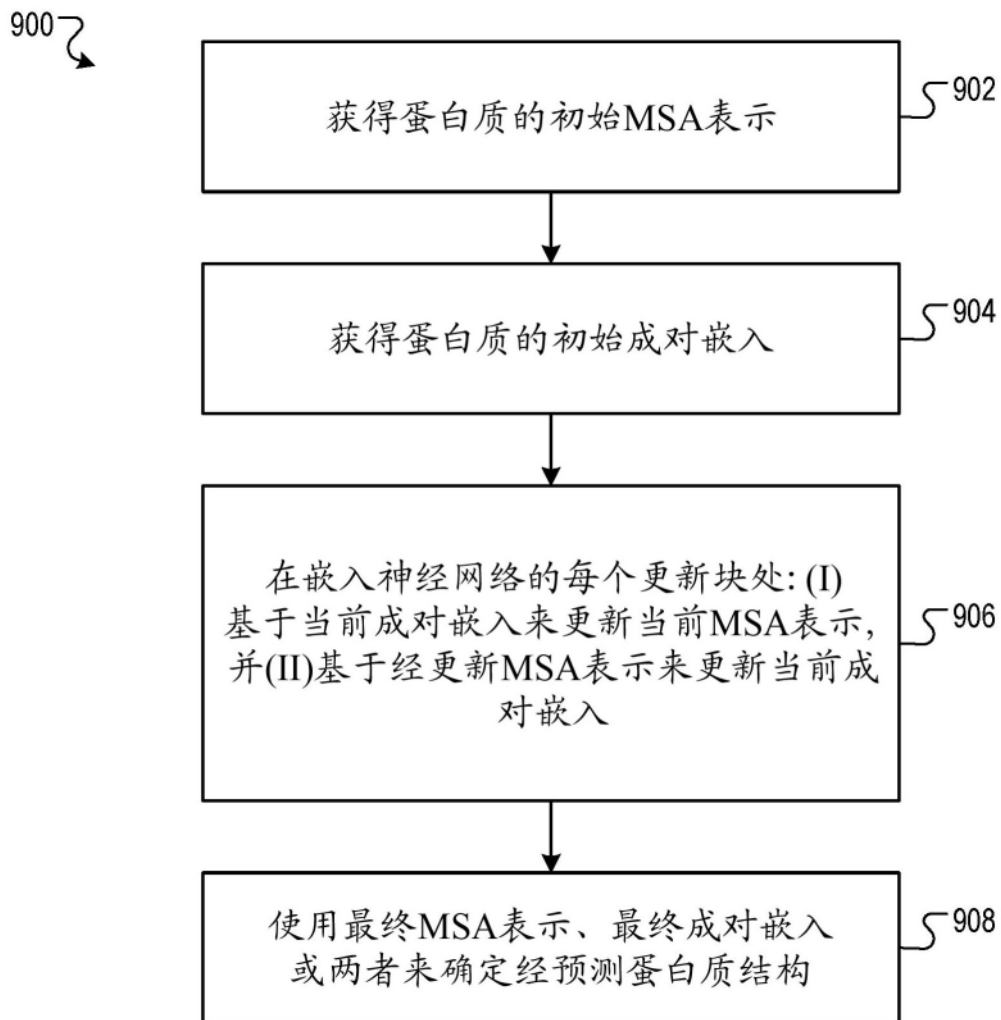


图15

生成蛋白质中的氨基酸链的MSA表示

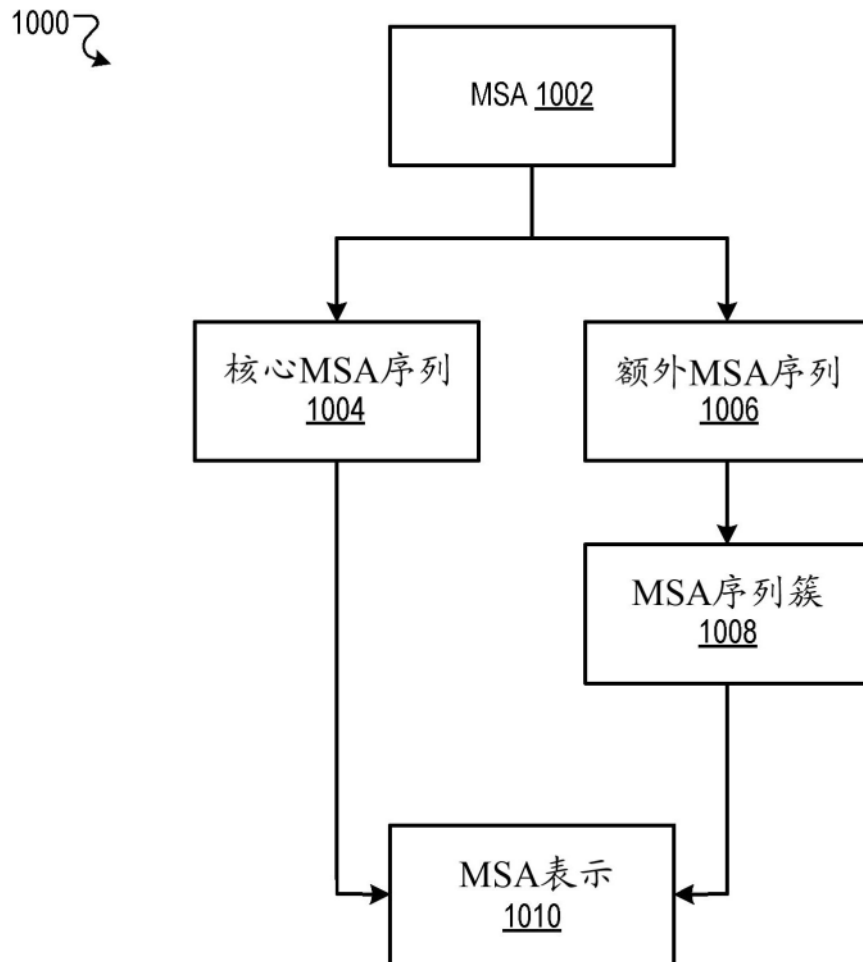


图16

生成初始成对嵌入

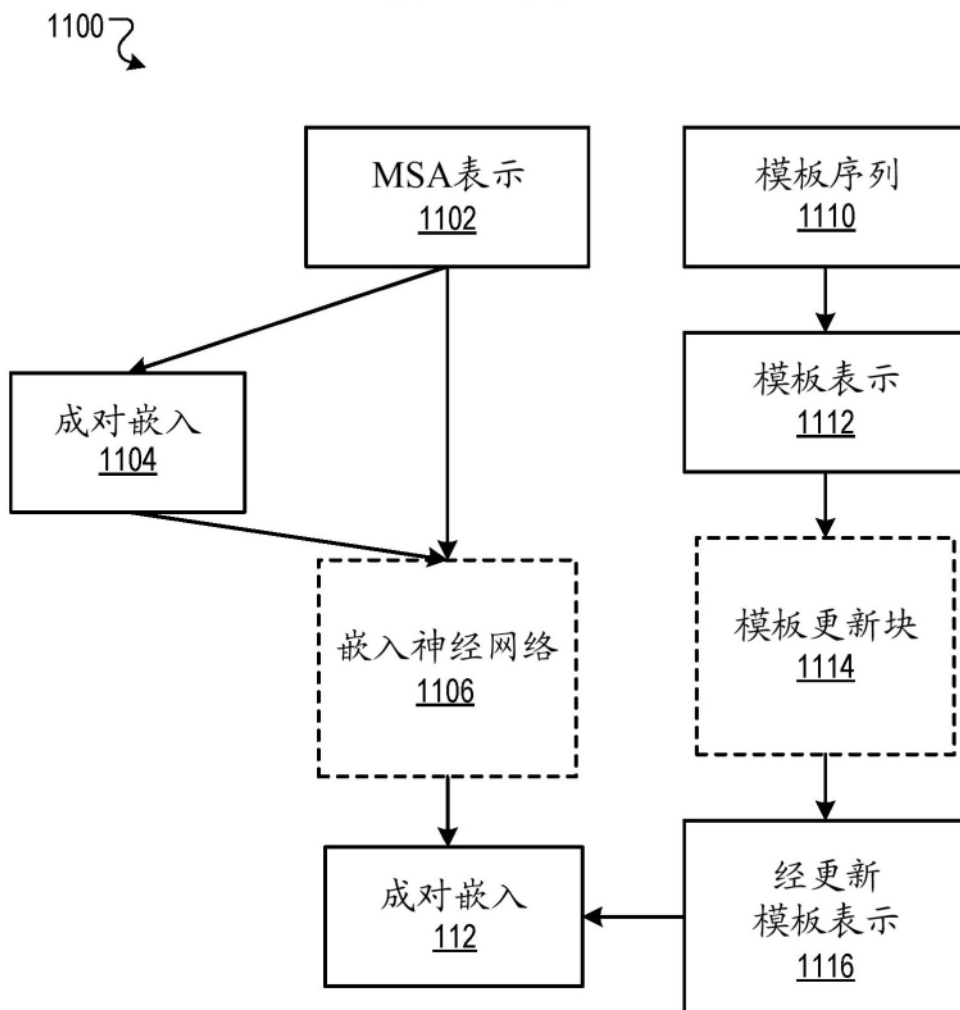


图17

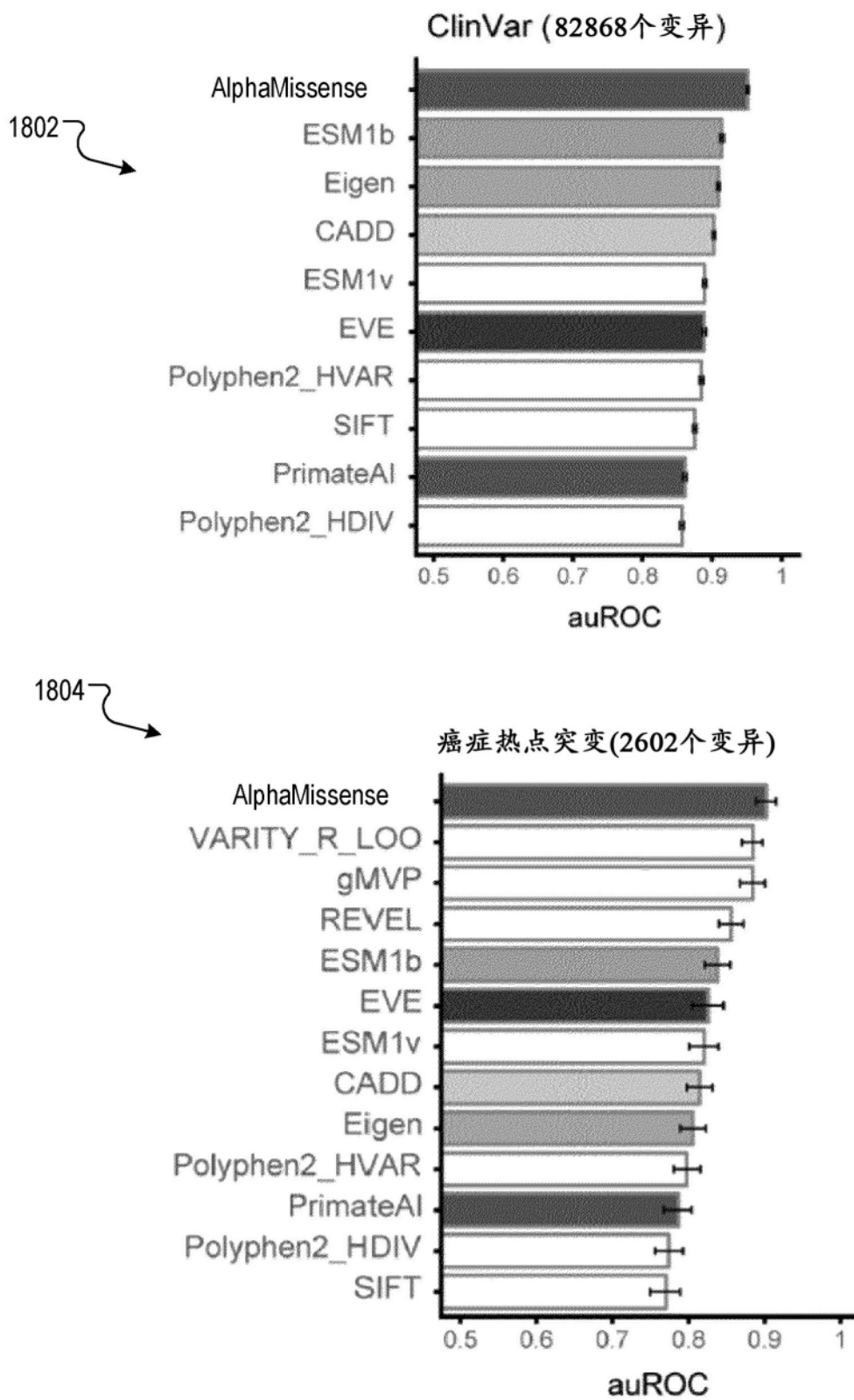


图18A

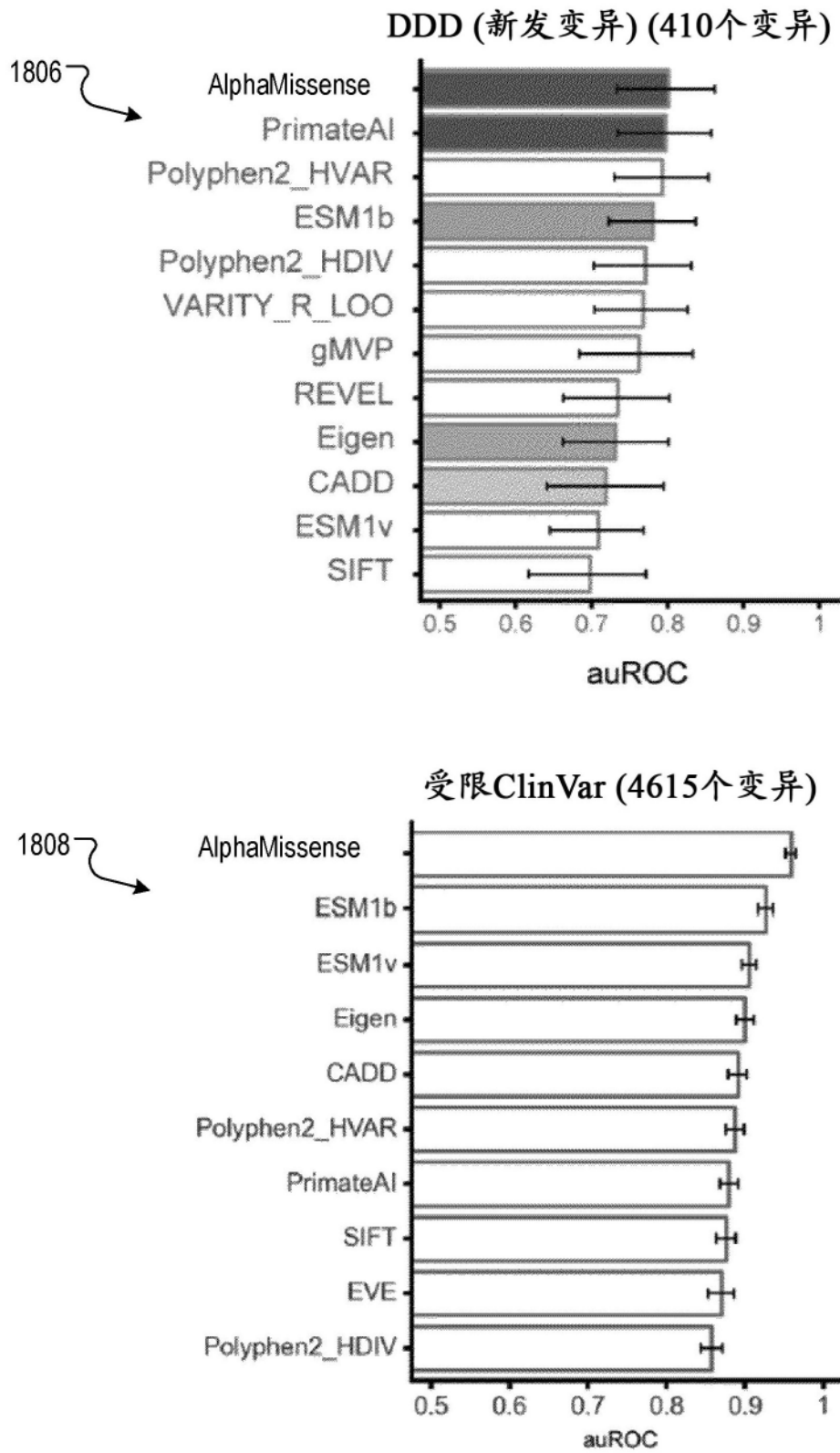


图18B