



(21) 申请号 202180069629.X

(22) 申请日 2021.11.23

(30) 优先权数据

63/118,921 2020.11.28 US

(85) PCT国际申请进入国家阶段日

2023.04.11

(86) PCT国际申请的申请数据

PCT/EP2021/082698 2021.11.23

(87) PCT国际申请的公布数据

W02022/112257 EN 2022.06.02

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 S·科尔 O·罗内伯格

M·非古尔诺夫 A·普里泽尔

(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 40/20 (2006.01)

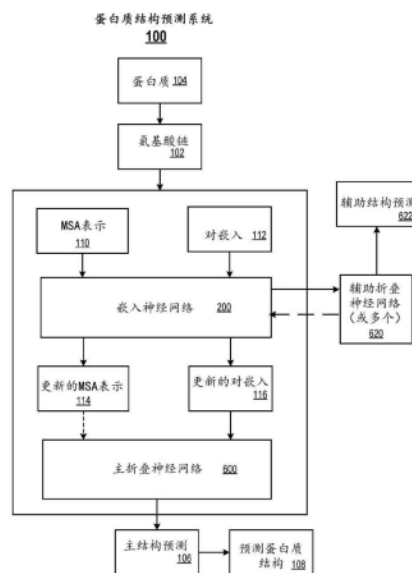
权利要求书4页 说明书28页 附图11页

(54) 发明名称

使用辅助折叠网络预测蛋白质结构

(57) 摘要

用于训练结构预测神经网络的方法、系统和装置,包括在计算机存储介质上编码的计算机程序,结构预测神经网络包括嵌入神经网络和主折叠神经网络。根据一个方面,一种方法包括:获得表征训练蛋白质的训练网络输入;使用嵌入神经网络和主折叠神经网络处理训练网络输入以生成主结构预测;对于一个或多个辅助折叠神经网络的集合中的每个辅助折叠神经网络,至少处理嵌入神经网络的对应中间输出以生成辅助结构预测;确定目标函数的梯度,目标函数包括关于辅助折叠神经网络中的每一个的相应辅助结构损失项;以及基于梯度更新嵌入网络参数和主折叠参数的当前值。



1. 一种训练结构预测神经网络的方法, 其中, 结构预测神经网络包括: (i) 嵌入神经网络, 嵌入神经网络具有多个嵌入参数并且被配置为接收表征蛋白质的网络输入并根据嵌入参数处理网络输入以生成关于网络输入的嵌入输出; 以及 (ii) 主折叠神经网络, 主折叠神经网络具有多个主折叠参数并且被配置为接收所述嵌入输出并根据主折叠参数处理主折叠输入以生成定义蛋白质的预测结构的主结构预测, 所述方法包括:

获得表征训练蛋白质的训练网络输入和指定训练蛋白质的目标蛋白质结构的数据;

使用嵌入神经网络并根据嵌入参数的当前值来处理训练网络输入, 以生成关于训练网络输入的训练嵌入输出;

使用主折叠神经网络并根据主折叠参数的当前值来处理训练嵌入输出, 以生成定义训练蛋白质的主预测结构的主结构预测;

对于每个具有相应多个辅助折叠参数的一个或多个辅助折叠神经网络的集合中的每个辅助折叠神经网络, 使用辅助折叠神经网络并根据辅助折叠神经网络的相应辅助折叠参数的当前值来至少处理嵌入神经网络的对应中间输出, 以生成定义训练蛋白质的辅助预测结构的辅助结构预测;

确定目标函数的梯度, 所述目标函数包括:

主结构损失项, 表征以下之间的相似性: (i) 由主结构预测定义的主预测结构和 (ii) 训练蛋白质的目标蛋白质结构; 和

每个辅助折叠神经网络的相应辅助结构损失项, 其表征以下之间的相似性: (i) 由辅助结构预测神经网络生成的辅助结构预测定义的辅助预测结构和 (ii) 训练蛋白质的目标蛋白质结构; 以及

基于梯度更新嵌入网络参数、主折叠参数和一个或多个辅助折叠神经网络的相应辅助折叠参数的当前值。

2. 根据任一前述权利要求所述的方法, 其中, 每个辅助折叠神经网络具有与主折叠神经网络相同的神经网络架构。

3. 根据权利要求2所述的方法, 其中, 一个或多个辅助折叠神经网络的集合包括多个辅助折叠神经网络, 并且其中, 目标函数约束辅助折叠神经网络以共享参数值。

4. 根据权利要求2所述的方法, 其中, 目标函数约束辅助折叠神经网络和主折叠神经网络以共享参数值。

5. 根据前述权利要求中任一项所述的方法, 其中:

网络输入包括蛋白质中的每个氨基酸对的相应初始对嵌入,

嵌入神经网络包括更新块序列, 其中每个更新块执行操作, 所述操作包括:

接收块输入, 块输入包括蛋白质中的每个氨基酸对的相应当前对嵌入; 和

更新蛋白质中的每个氨基酸对的相应当前对嵌入, 以生成蛋白质中的每个氨基酸对的相应的更新的对嵌入, 并且

嵌入输出至少包括由序列中的最后一个更新块生成的更新的对嵌入。

6. 根据权利要求5所述的方法, 其中, 每个辅助折叠神经网络与序列中的更新块中不是所述序列中的最后一个更新块的不同更新块对应, 并且其中, 每个辅助折叠神经网络被配置为至少接收由对应的更新块生成的更新的对嵌入作为输入。

7. 根据权利要求6所述的方法, 其中:

网络输入还包括初始多序列比对 (MSA) 嵌入, MSA 嵌入表示与蛋白质中的每个链对应的相应多序列比对,

到更新块中的每一个的块输入进一步包括当前 MSA 嵌入, 并且

由每个更新块执行的操作进一步包括:

更新当前 MSA 嵌入以生成更新 MSA 嵌入。

8. 根据权利要求 7 所述的方法, 其中, 每个辅助折叠神经网络的输入还包括由对应的更新块生成的更新 MSA 嵌入。

9. 根据权利要求 7 或 8 中任一项所述的方法, 其中, 每个辅助折叠神经网络被配置为:

从由辅助折叠神经网络生成的辅助结构预测生成经变换的结构预测, 经变换的结构预测具有与所述更新的对嵌入相同的维度;

将经变换的结构预测与更新的对嵌入组合以生成进一步更新的对嵌入; 以及

提供进一步更新的对嵌入作为序列中在当前更新块之后的更新块的输入。

10. 根据权利要求 9 所述的方法, 其中:

辅助折叠预测包括结构参数, 结构参数为每个氨基酸指定蛋白质结构中的该氨基酸中的指定原子的预测 3-D 空间位置; 以及

生成经变换的结构预测包括:

从由结构参数指定的氨基酸的预测的 3-D 空间位置生成距离图, 对于蛋白质中的每个氨基酸对, 距离图表征蛋白质结构中的该氨基酸对之间的相应估计距离; 和

从距离图生成具有与更新的对嵌入相同的维度的经变换的距离图。

11. 根据权利要求 9 所述的方法, 其中:

辅助折叠预测包括指定距离图的结构参数, 对于蛋白质中的每个氨基酸对, 距离图表征蛋白质结构中的该氨基酸对之间的相应估计距离; 并且

生成经变换的结构预测包括:

从由结构参数指定的距离图生成具有与初始对嵌入相同的维度的经变换的距离图。

12. 根据任一前述权利要求所述的方法, 还包括:

在训练之后, 获得表征新蛋白质的新网络输入; 以及

使用经训练的结构预测神经网络处理新的网络输入, 以生成定义新蛋白质的预测结构的新主结构预测。

13. 一种系统, 包括:

一个或多个计算机; 以及

通信地耦合到一个或多个计算机的一个或多个存储设备, 其中, 一个或多个存储设备存储指令, 指令在由一个或多个计算机执行时使一个或多个计算机执行根据权利要求 1-12 中任一项所述的相应方法的操作。

14. 一个或多个非暂时性计算机存储介质, 存储指令, 指令在由一个或多个计算机执行时使一个或多个计算机执行根据权利要求 1-12 中任一项所述的相应方法的操作。

15. 一种获得配体的方法, 其中, 配体是药物或工业酶的配体, 所述方法包括:

执行根据权利要求 12 所述的方法以确定靶蛋白的预测结构;

评估一个或多个候选配体与靶蛋白的预测结构的相互作用; 以及

取决于评估结果选择一个或多个候选配体作为配体。

16. 根据权利要求15所述的方法, 其中, 靶蛋白包括受体或酶, 并且其中, 配体是受体或酶的激动剂或拮抗剂。

17. 根据权利要求15或16所述的方法, 其中, 配体是药物, 并且方法包括:
执行根据权利要求12所述的方法以确定多个靶蛋白中的每一个的预测结构;
评估一个或多个候选配体与每个靶蛋白的预测结构的相互作用; 以及
选择一个或多个候选配体作为配体, 以 i) 获得与靶蛋白中的每一个相互作用的配体, 或 ii) 获得仅与靶蛋白中的一个相互作用的配体。

18. 一种获得多肽配体的方法, 其中, 配体是药物或工业酶的配体, 所述方法包括
对于一个或多个候选多肽配体中的每一个, 执行根据权利要求12所述的方法以确定候选多肽配体的预测结构;

获得靶蛋白的目标蛋白质结构;

评估一个或多个候选多肽配体中的每一个的预测结构与目标蛋白质结构之间的相互作用; 以及

取决于评估结果选择一个或多个候选多肽配体中的一个作为多肽配体。

19. 根据权利要求18所述的方法, 其中, 靶蛋白包括受体或酶, 并且其中, 配体是受体或酶的激动剂或拮抗剂; 或者其中, 多肽配体包含抗体, 并且靶蛋白包含抗体靶标, 特别是病毒或癌细胞蛋白, 并且其中, 抗体与抗体靶标结合以提供治疗效果。

20. 一种获得疾病的诊断性抗体标志物的方法, 所述方法包括:

对于一个或多个候选抗体中的每一个, 执行根据权利要求12所述的方法以确定候选抗体的预测结构;

获得靶蛋白的目标蛋白质结构;

评估一个或多个候选抗体中的每一个的预测结构与目标蛋白质结构之间的相互作用; 以及

取决于评估结果选择一个或多个候选抗体中的一个作为诊断性抗体标志物。

21. 根据权利要求15-20中任一项所述的方法, 其中, 评估候选配体中的一个的相互作用包括确定候选配体的相互作用分数, 其中, 相互作用分数包括候选配体与靶蛋白之间的相互作用的度量。

22. 根据权利要求15-21中任一项所述的方法, 还包括: 合成配体。

23. 根据权利要求22所述的方法, 还包括: 在体外和体内测试配体的生物活性。

24. 一种识别蛋白质错误折叠疾病的存在的方法, 包括:

执行根据权利要求12所述的方法以确定蛋白质的预测结构;

获得从人体或动物体获得的蛋白质版本的结构;

将蛋白质的预测结构与从人体或动物体获得的蛋白质版本的结构进行比较; 以及

取决于比较结果识别蛋白质错误折叠疾病的存在。

25. 一种系统, 包括:

一个或多个计算机; 以及

通信地耦合到一个或多个计算机的一个或多个存储设备, 其中, 一个或多个存储设备存储指令, 指令在由一个或多个计算机执行时使一个或多个计算机实现:

结构预测神经网络, 包括: (i) 嵌入神经网络, 具有多个嵌入参数, 并且被配置为接收表

征蛋白质的网络输入并根据嵌入参数处理网络输入以生成关于网络输入的嵌入输出,以及(ii)主折叠神经网络,具有多个主折叠参数,并且被配置为接收嵌入输出并根据主折叠参数处理主折叠输入以生成定义蛋白质的预测结构的主结构预测;

一个或多个辅助折叠神经网络的集合,其中,每个辅助折叠神经网络具有相应的多个辅助折叠参数,并且被配置为根据辅助折叠神经网络的相应辅助折叠参数的当前值至少处理嵌入神经网络的中间输出,以生成定义蛋白质的辅助预测结构的辅助结构预测;

训练子系统,被配置为执行操作,所述操作包括:

获得表征训练蛋白质的训练网络输入和指定训练蛋白质的目标蛋白质结构的数据;

使用嵌入神经网络并根据嵌入参数的当前值来处理训练网络输入,以生成关于训练网络输入的训练嵌入输出;

使用主折叠神经网络并根据主折叠参数的当前值来处理训练嵌入输出,以生成定义训练蛋白质的主预测结构的主结构预测;

对于所述一个或多个辅助折叠神经网络的集合中的每个辅助折叠神经网络,使用辅助折叠神经网络至少处理嵌入神经网络的对应中间输出,以生成定义训练蛋白质的辅助预测结构的辅助结构预测;

确定目标函数的梯度,目标函数包括:

主结构损失项,表征以下之间的相似性:(i)由主结构预测定义的主预测结构和(ii)训练蛋白质的目标蛋白质结构;和

每个辅助折叠神经网络的相应辅助结构损失项,表征以下之间的相似性:(i)由辅助结构预测神经网络生成的辅助结构预测定义的辅助预测结构和(ii)训练蛋白质的目标蛋白质结构;以及

基于梯度更新嵌入网络参数、主折叠参数和一个或多个辅助折叠神经网络的相应辅助折叠参数的当前值。

使用辅助折叠网络预测蛋白质结构

[0001] 对相关申请的交叉引用

[0002] 本申请要求于2020年11月28日提交的美国临时专利申请序列号63/118,921的申请日的权益,并且其全部内容通过引用并入本文。

技术领域

[0003] 本说明书涉及预测蛋白质结构。

背景技术

[0004] 蛋白质由一个或多个氨基酸序列(“链”)指定。氨基酸是有机化合物,其包括氨基官能团和羧基官能团,以及氨基酸特异性的侧链(即原子团)。蛋白质折叠是指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历蛋白质折叠后蛋白质氨基酸序列中原子的3-D构型。当在通过肽键连接的序列中时,氨基酸可以被称为氨基酸残基。

[0005] 能够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的输入生成输出,例如预测输出。一些机器学习模型是参数模型,并且基于接收的输入和模型的值来生成输出。一些机器学习模型是深度模型,其采用多层模型来针对接收的输入生成输出。例如,深度神经网络是深度机器学习模型,其包括输出层和一个或多个隐藏层,每个隐藏层将非线性变换应用于接收的输入以生成输出。

发明内容

[0006] 本说明书描述了一种神经网络系统,其在一个或多个位置中的一个或多个计算机上实现为计算机程序,用于预测蛋白质结构。

[0007] 如本说明书通篇所用,术语“蛋白质”可以理解为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域,例如蛋白质的氨基酸链的一部分,其能够几乎独立于蛋白质的其余部分进行蛋白质折叠。作为另一个实例,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的蛋白质复合物。

[0008] 蛋白质中的氨基酸链的“多重序列比对”(MSA)指定该氨基酸链与多个另外的氨基酸链——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可以定义该氨基酸链中的位置与多个另外的氨基酸链中的相应位置之间的对应关系。氨基酸链的MSA可以例如通过使用任何适当的计算序列比对技术(例如渐进比对构建)处理氨基酸链的数据库来生成。MSA中的氨基酸链可以理解为具有进化关系,例如,其中MSA中的每个氨基酸链可以共享共同的祖先。氨基酸链的MSA中的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。

[0009] 实体(例如,氨基酸对)的“嵌入”能够指作为数值的有序集合的实体的表示,例如,数值的向量或矩阵。

[0010] 蛋白质的结构能够由一组结构参数定义。定义蛋白质的结构的一组结构参数能够表示为数值的有序集合。下面更详细地描述用于定义蛋白质的结构的可能结构参数的一些示例。

[0011] 在一个示例中,对于蛋白质中的每个氨基酸,定义蛋白质的结构的结构参数包括:(i)位置参数,和(ii)旋转参数。

[0012] 氨基酸的位置参数能够指定蛋白质的结构中的氨基酸中指定原子的预测3-D空间位置。指定原子能够是氨基酸中的阿尔法碳原子,即氨基酸中与氨基官能团、羧基官能团和侧链键合的碳原子。作为另一示例,指定原子能够是氨基酸中的贝塔碳原子。氨基酸的位置参数能够在任何适当的坐标系中表示,例如三维 $[x, y, z]$ 笛卡尔坐标系。

[0013] 氨基酸的旋转参数能够指定氨基酸在蛋白质的结构中的预测“方位”。更具体地,旋转参数能够指定3-D空间旋转操作,如果将其应用于位置参数的坐标系,则使得氨基酸中的三个“主链”原子相对于旋转坐标系采取固定位置。氨基酸中的三个主链原子能够指氨基酸中链接的一系列氮、阿尔法碳和羧基碳原子。氨基酸的旋转参数能够表示为例如行列式等于1的标准正交 3×3 矩阵。

[0014] 通常,氨基酸的位置和旋转参数定义氨基酸的自我中心参考系(egocentric reference frame)。在该参考系中,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,阿尔法碳-贝塔碳键)可以沿着定义的方向。

[0015] 在另一个示例中,定义蛋白质的结构的结构参数能够包括表征蛋白质中的每个氨基酸对之间的相应估计距离(例如,以埃为单位测量)的“距离图”。距离图能够表征氨基酸对之间的估计距离,例如通过氨基酸对之间的一组可能距离的概率分布。

[0016] 在另一个示例中,定义蛋白质的结构的结构参数能够定义蛋白质的结构中的每个氨基酸中的每个原子的三维(3D)空间位置。

[0017] 本文所述的蛋白质结构预测系统能够用于获得配体,诸如药物或工业酶的配体。例如,获得配体的方法可以包括获得靶氨基酸序列,特别是靶蛋白(例如药物靶标)的氨基酸序列,并使用蛋白质结构预测系统基于靶氨基酸序列处理输入以确定靶蛋白的(三级)结构,即预测蛋白质结构。然后,该方法可以包括评估一个或多个候选配体与目标蛋白质结构的相互作用。该方法还可以包括取决于相互作用的评估结果选择一个或多个候选配体作为配体。

[0018] 在一些实施方式中,评估相互作用可以包括评估候选配体与目标蛋白质结构的结合。例如,评估相互作用可以包括识别以足够的亲和力结合生物效应的配体。在一些其他实施方式中,评估相互作用可以包括评估候选配体与目标蛋白质结构的缔合,其对靶蛋白(例如酶)的功能具有影响。评估可以包括评估候选配体与目标蛋白质结构之间的亲和力,或评估相互作用的选择性。可以根据具有最高亲和力的候选配体来选择候选配体。

[0019] 候选配体(或多个)可以衍生自候选配体的数据库,和/或可以通过修改候选配体数据库中的配体来衍生,例如通过修改候选配体的结构或氨基酸序列,和/或可以通过候选配体的逐步或迭代组装/优化来衍生。

[0020] 候选配体与目标蛋白质结构的相互作用的评估可以使用计算机辅助方法执行,其中显示候选配体和目标蛋白质结构的图形模型用于用户操作,和/或评估可以部分或完全自动执行,例如使用标准分子(蛋白质-配体)对接软件。在一些实施方式中,评估可以包括

确定候选配体的相互作用分数,其中相互作用分数包括候选配体与靶蛋白之间的相互作用的测量。相互作用分数可以取决于相互作用的强度和/或特异性,例如,取决于结合自由能的分数的。可以取决于其分数来选择候选配体,例如,当候选配体具有最好相互作用分数时。

[0021] 在一些实施方式中,靶蛋白包括受体或酶,并且配体是受体或酶的激动剂或拮抗剂。在一些实施方式中,该方法可以用于识别细胞表面标志物的结构。然后可以将其用于识别与细胞表面标志物结合的配体,例如抗体或标记物(诸如荧光标记物)。这可以用于识别和/或治疗癌细胞。

[0022] 在一些实施方式中,配体是药物,并且确定多种靶蛋白中的每一种的预测结构,并且评估一个或多个候选配体与靶蛋白中的每一种的预测结构的相互作用。然后可以选择一个或多个候选配体以获得与每个靶蛋白(功能性)相互作用的配体,或获得仅与一种靶蛋白(功能性)相互作用的配体。例如,在一些实施方式中,会期望获得对多个药物靶标有效的药物。另外或替代地,会期望筛选药物的脱靶效应。例如,在农业中,确定设计用于一种植物物种的药物不与另一种不同的植物物种和/或动物物种相互作用能够是有用的。

[0023] 在一些实施方式中,候选配体(或多种)可以包括小分子配体,例如分子量<900道尔顿的有机化合物。在一些其他实施方式中,候选配体(或多种)可以包括多肽配体,即由氨基酸序列定义的多肽配体。

[0024] 在一些情况下,蛋白质结构预测系统能够用于确定候选多肽配体(例如药物或工业酶的配体)的结构。然后可以评估其与目标蛋白质结构的相互作用;已经可以使用结构预测神经网络或使用常规物理探索技术——诸如x射线晶体学和/或磁共振技术和/或低温电子显微镜——确定目标蛋白质结构。

[0025] 在另一方面,提供一种使用蛋白质结构预测系统来获得多肽配体(例如,分子或其序列)的方法。该方法可以包括获得一个或多个候选多肽配体的氨基酸序列。该方法可以还包括使用蛋白质结构预测系统来确定候选多肽配体的(三级)结构。该方法可以还包括通过生物信息和/或通过物理探索获得靶蛋白的目标蛋白质结构,并评估一个或多个候选多肽配体中的每一个的结构与目标蛋白质结构之间的相互作用。该方法可以还包括取决于评估结果选择一个或多个候选多肽配体作为多肽配体。

[0026] 如前所述,评估相互作用可以包括评估候选多肽配体与目标蛋白质结构的结合,例如识别以足够的亲和力结合生物效应的配体,和/或评估候选多肽配体与对靶蛋白(例如酶)的功能有影响的目标蛋白质结构的缔合,和/或评估候选多肽配体与目标蛋白质结构之间的亲和力,或评估相互作用的选择性。在一些实施方式中,多肽配体可以是适体。同样,可以根据具有最高亲和力的多肽候选配体来选择多肽候选配体(或多种)。

[0027] 如前所述,选择的多肽配体可以包括受体或酶,并且配体可以是受体或酶的激动剂或拮抗剂。在一些实施方式中,多肽配体可以包括抗体,并且靶蛋白可以包括抗体靶标,例如病毒,特别是病毒外壳蛋白,或在癌细胞上表达的蛋白质。在这些实施方式中,抗体与抗体靶标结合以提供治疗效果。例如,抗体可以结合靶标并充当特定受体的激动剂;可替代地,抗体可以防止另一种配体与靶标结合,并因此防止相关生物途径的激活。

[0028] 该方法的实施可以还包括合成,即制备小分子或多肽配体。配体可以通过任何常规化学技术合成和/或可以是已经可获得的,例如可以来自化合物库或可以使用组合化学来合成。

[0029] 该方法可以还包括在体外和/或体内测试配体的生物活性。例如,可以测试配体的ADME(吸收、分布、代谢、排泄)和/或毒理学属性,以筛选出不合适的配体。测试可以包括例如使候选小分子或多肽配体与靶蛋白接触并测量蛋白质的表达或活性的改变。

[0030] 在一些实施方式中,候选(多肽)配体可以包括:分离抗体、分离抗体的片段、单可变结构域抗体、双特异性或多特异性抗体、多价抗体、双可变结构域抗体、免疫缀合物、纤连蛋白分子、adnectin、DARPin、avimer、亲和体、抗运载蛋白(anticalin)、affilin、蛋白质表位模拟物或其组合。候选(多肽)配体可以包括具有突变或化学修饰的氨基酸Fc区的抗体,例如,当与野生型Fc区相比时,其阻止或降低ADCC(抗体依赖性细胞毒性)活性和/或增加半衰期。候选(多肽)配体可以包括具有不同CDR(互补决定区)的抗体。

[0031] 本文所述的蛋白质结构预测系统也能够用于获得疾病的诊断性抗体标志物。还提供了一种方法,对于一个或多个候选抗体中的每一个,例如如上所述,所述方法使用蛋白质结构预测系统来确定候选抗体的预测结构。该方法还可以涉及获得靶蛋白的目标蛋白质结构,评估一个或多个候选抗体中的每一个的预测结构与目标蛋白质结构之间的相互作用,并取决于评估结果选择一个或多个候选抗体中的一种作为诊断性抗体标志物,例如选择对目标蛋白质结构具有最高亲和力的一个或多个候选抗体。该方法可以包括制备诊断性抗体标志物。诊断性抗体标志物可以用于通过检测其是否与从患者获得的样本(例如体液样本)中的靶蛋白结合来诊断疾病。如上所述,能够使用对应的技术来获得治疗性抗体(多肽配体)。

[0032] 错误折叠的蛋白质与许多疾病相关联。因此,在另一方面,提供一种使用蛋白质结构预测系统来识别蛋白质错误折叠疾病的存在的方法。该方法可以包括获得蛋白质的氨基酸序列并使用蛋白质结构预测系统来确定蛋白质的结构。该方法可以还包括例如通过常规(物理)方法获得从人体或动物体获得的蛋白质版本(version)的结构。然后,该方法可以包括将蛋白质的结构与从身体获得的版本的结构进行比较,并取决于比较结果识别蛋白质错误折叠疾病的存在。也就是说,可以通过与生物信息确定的结构比较来确定来自身体的蛋白质版本的错误折叠。

[0033] 通常,识别蛋白质错误折叠疾病的存在可以涉及获得蛋白质的氨基酸序列,使用蛋白质的氨基酸序列来确定蛋白质的结构,如本文所述,并将蛋白质的结构与基线版本的蛋白质的结构进行比较,取决于比较的结果识别蛋白质错误折叠疾病的存在。例如,比较的结构可以是突变体和野生型蛋白质的结构。在实施方式中,野生型蛋白质可以用作基线版本,但原则上任一种都可以用作基线版本。

[0034] 在一些其他方面,如上文或本文所述的计算机实施的方法可以用于从其氨基酸序列识别靶蛋白上的活性/结合/阻断位点。

[0035] 根据一个方面,提供了一种训练结构预测神经网络的方法,其中结构预测神经网络包括:(i)嵌入神经网络,具有多个嵌入参数并且被配置为接收表征蛋白质的网络输入并根据嵌入参数处理网络输入以生成关于网络输入的嵌入输出,以及(ii)主折叠神经网络,具有多个主折叠参数并且被配置为接收嵌入输出并根据主折叠参数处理主折叠输入以生成定义蛋白质的预测结构的主结构预测,该方法包括:获得表征训练蛋白质的训练网络输入和指定训练蛋白质的目标蛋白质结构的数据;使用嵌入神经网络并根据嵌入参数的当前值来处理训练网络输入,以生成关于训练网络输入的训练嵌入输出;使用主折叠神经网络

并根据主折叠参数的当前值来处理训练嵌入输出,以生成定义训练蛋白质的主预测结构的主结构预测;对于每个具有相应多个辅助折叠参数的一个或多个辅助折叠神经网络的集合中的每个辅助折叠神经网络,使用辅助折叠神经网络并根据辅助折叠神经网络的相应辅助折叠参数的当前值来至少处理嵌入神经网络的对应中间输出,以生成定义训练蛋白质的辅助预测结构的辅助结构预测;确定目标函数的梯度,所述目标函数包括:主结构损失项,表征(i)由主结构预测定义的主预测结构和(ii)训练蛋白质的目标蛋白质结构之间的相似性;以及每个辅助折叠神经网络的相应辅助结构损失项,表征以下之间的相似性:(i)由辅助结构预测神经网络生成的辅助结构预测定义的辅助预测结构和(ii)训练蛋白质的目标蛋白质结构;以及基于梯度更新嵌入网络参数、主折叠参数和一个或多个辅助折叠神经网络的相应辅助折叠参数的当前值。

[0036] 在一些实施方式中,每个辅助折叠神经网络具有与主折叠神经网络相同的神经网络架构。

[0037] 在一些实施方式中,一个或多个辅助折叠神经网络的集合包括多个辅助折叠神经网络,并且目标函数约束辅助折叠神经网络以共享参数值。

[0038] 在一些实施方式中,目标函数约束辅助折叠神经网络和主折叠神经网络以共享参数值。

[0039] 在一些实施方式中,网络输入包括蛋白质中每个氨基酸对的相应初始对嵌入,嵌入神经网络包括更新块序列,并且每个更新块执行包括以下的操作:接收包括蛋白质中的每个氨基酸对的相应当前对嵌入的块输入;以及更新蛋白质中的每个氨基酸对的相应的当前对嵌入,以生成蛋白质中的每个氨基酸对的相应的更新的块嵌入,并且嵌入输出至少包括由序列中的最后一个更新块生成的更新的块嵌入。

[0040] 在一些实施方式中,每个辅助折叠神经网络与序列中的更新块中不是所述序列中的最后一个更新块的不同更新块对应,并且每个辅助折叠神经网络被配置为至少接收由对应更新块生成的更新的块嵌入作为输入。

[0041] 在一些实施方式中,网络输入还包括初始多序列比对(MSA)嵌入,其表示与蛋白质中的每个链相对应的相应多序列比对,每个更新块的块输入还包括当前MSA嵌入;并且由每个更新块执行的操作还包括:更新当前MSA嵌入以生成更新的MSA嵌入。

[0042] 在一些实施方式中,每个辅助折叠神经网络的输入还包括由对应的更新块生成的更新的MSA嵌入。

[0043] 在一些实施方式中,每个辅助折叠神经网络被配置为:从由辅助折叠神经网络生成的辅助结构预测生成经变换的结构预测,该经变换的结构预测具有与更新后的块嵌入相同的维度;将经变换的结构预测与更新的块嵌入组合以生成进一步的更新的块嵌入;以及将进一步更新的块嵌入作为输入提供给所述序列中在所述当前更新块之后的更新块。

[0044] 在一些实施方式中,辅助折叠预测包括结构参数,所述结构参数对于每个氨基酸指定氨基酸中的指定原子在蛋白质结构中的预测3-D空间位置;并且生成经变换的结构预测包括:从由结构参数指定的氨基酸的预测3-D空间位置生成距离图,对于蛋白质中的每个氨基酸对,距离图表征蛋白质的结构中的氨基酸对之间的相应估计距离;以及从距离图生成具有与更新的块嵌入相同的维度的经变换的距离图。

[0045] 在一些实施方式中,辅助折叠预测包括指定距离图的结构参数,对于蛋白质中的

每个氨基酸对,距离图表征蛋白质的结构中的氨基酸对之间的相应估计距离;以及

[0046] 生成经变换的结构预测包括:从由结构参数指定的距离图生成具有与初始对嵌入相同的维度的经变换的距离图。

[0047] 在一些实施方式中,该方法还包括:在训练之后,获得表征新蛋白质的新网络输入;以及使用经训练的结构预测神经网络处理新的网络输入,以生成定义新蛋白质的预测结构的新主结构预测。

[0048] 可以实现本说明书中描述的主题的特定实施例,以便实现以下优点中的一个或多个。

[0049] 本说明书中描述的系统使用神经网络预测蛋白质的结构,该神经网络至少在训练期间,除了生成预测蛋白质的结构的实际预测输出的主折叠神经网络和生成主折叠神经网络的输入的嵌入神经网络之外,还包括一个或多个辅助折叠神经网络。每个辅助折叠神经网络与嵌入神经网络内的多个更新块中的一个对应并接收来自嵌入神经网络内的多个更新块中的一个的输入。在训练期间包括辅助折叠神经网络作为神经网络的一部分允许嵌入神经网络接收更丰富的训练信号,并且因此学习生成更高质量的嵌入,例如,相对于仅基于由主折叠神经网络进行的预测中的误差以及可选地基于一个或多个辅助任务的输出中的误差的学习,一旦神经网络已经被训练,就导致由主折叠神经网络生成更高质量的预测。

[0050] 在一些情况下,辅助神经网络在训练之后也包括在神经网络中,并且每个辅助神经网络使用神经网络生成的辅助结构预测来修改对应更新块的输出。这可以进一步提高作为输入提供给主折叠神经网络的所生成的嵌入的质量,进一步提高由主折叠神经网络在训练之后生成的预测的质量。

[0051] 本说明书中描述的系统可以通过单次前向传递联合训练的神经网络的集合来预测蛋白质的结构,这可能需要少于一秒。相反,一些常规系统通过经由可能的蛋白质结构的空间的扩展搜索过程来预测蛋白质的结构,以优化标量得分函数,例如使用模拟退火或梯度下降技术。这样的搜索过程可能需要数百万次搜索迭代并且消耗数百个中央处理单元(CPU)小时。通过神经网络集合的单次前向传递来预测蛋白质结构可以使本说明书中描述的系统比通过迭代搜索过程预测蛋白质结构的系统消耗更少的计算资源(例如,存储器和计算能力)。

[0052] 蛋白质的结构决定蛋白质的生物学功能。因此,确定蛋白质结构可以促进理解生命过程(例如,包括许多疾病的机制)和设计蛋白质(例如,作为药物,或作为用于工业过程的酶)。例如,哪些分子(例如,药物)将与蛋白质结合(以及结合将在何处发生)取决于蛋白质的结构。由于药物的有效性能够受到它们与蛋白质(例如,在血液中)结合的程度的影响,因此确定不同蛋白质的结构会是药物开发的重要方面。然而,使用物理实验(例如,通过x射线晶体学)确定蛋白质结构会是耗时且非常昂贵的。因此,本说明书中描述的蛋白质预测系统可以促进涉及蛋白质的生物化学研究和工程领域(例如,药物开发)。

[0053] 在附图和下面的描述中阐述了本说明书的主题的一个或多个实施例的细节。根据说明书、附图和权利要求,主题的其他特征、方面和优点将变得显而易见。

附图说明

[0054] 图1示出示例蛋白质结构预测系统。

- [0055] 图2示出嵌入神经网络的示例架构。
- [0056] 图3示出嵌入神经网络的更新块的示例架构。
- [0057] 图4示出MSA更新块的示例架构。
- [0058] 图5示出对更新块的示例架构。
- [0059] 图6示出折叠神经网络的示例架构。
- [0060] 图7示出氨基酸中的键之间的扭转角。
- [0061] 图8是未折叠蛋白质和折叠蛋白质的图示。
- [0062] 图9是用于训练结构预测神经网络的示例过程的流程图。
- [0063] 图10示出用于生成蛋白质中的氨基酸链的MSA表示的示例过程。
- [0064] 图11示出用于为蛋白质中的每个氨基酸对生成相应的对嵌入的示例过程。
- [0065] 各个附图中相同的附图标记和名称指示相同的元件。

具体实施方式

[0066] 图1示出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中的一个或多个计算机上实现为计算机程序的系统的示例,其中实现了下面描述的系统、组件和技术。

[0067] 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的数据以生成主结构预测106,即定义预测蛋白质结构108——即蛋白质104的结构的预测——的一组结构参数。也就是说,蛋白质104的预测结构108可以由一组结构参数(例如,包括在主结构预测106中)定义,该组结构参数共同定义蛋白质经历蛋白质折叠后蛋白质的预测三维结构。

[0068] 定义预测蛋白质结构108的结构参数可以如前所述。例如,它们可以包括例如蛋白质104中的每个氨基酸的位置参数和旋转参数、表征蛋白质中的每个氨基酸对之间的估计距离的距离图、蛋白质的结构中的每个氨基酸中的每个原子或骨架原子的相应空间位置,或其组合,如上所述。

[0069] 为了生成定义预测蛋白质结构108的主结构预测106,系统100可以生成:(i)蛋白质的多序列比对(MSA)表示110(在一些实施方式中),以及(ii)蛋白质的一组“对”嵌入112,如接下来将更详细描述。

[0070] 蛋白质的MSA表示110包括蛋白质中的每个氨基酸链的MSA的相应表示。蛋白质中的氨基酸链的MSA表示可以被表示为嵌入的 $M \times N$ 阵列(即,具有M行和N列的嵌入的2-D阵列),其中N是氨基酸链中的氨基酸的数量。MSA表示的每一行可以对应于蛋白质中的氨基酸链的相应MSA序列。参考图10描述用于生成蛋白质中的氨基酸链的MSA表示的示例性过程。

[0071] 系统100从蛋白质中的氨基酸链的MSA表示生成蛋白质104的MSA表示110。

[0072] 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识别为蛋白质中的单个氨基酸链的MSA表示。

[0073] 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的氨基酸链的MSA表示组装到嵌入的分块对角2-D阵列来生成蛋白质的MSA表示110,即,其中蛋白质中的氨基酸链的MSA表示形成对角线上的块。系统100可以将对角线上的块外部的2-D阵列中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分配任意排序,并且蛋白质中的氨基酸链的MSA表示可以在分块对角矩阵中被相应地排序。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,以此类推。

[0074] 通常,蛋白质的MSA表示110可以被表示为嵌入的2-D阵列。在整个说明书中,蛋白质的MSA表示的“行”指定蛋白质的MSA表示的嵌入的2-D阵列的行。类似地,蛋白质的MSA表示的“列”指定蛋白质的MSA表示的嵌入的2-D阵列的列。

[0075] 一组对嵌入112包括对应于蛋白质104中的每个氨基酸对的相应对嵌入。通常,对嵌入表示、即编码关于蛋白质中的一对氨基酸之间的关系的信息。一对氨基酸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的氨基酸对的集合由下式给出:

$$[0076] \quad \{(A_i, A_j) : 1 \leq i, j \leq N\} \quad (1)$$

[0077] 其中,N是蛋白质中的氨基酸的数量, $i, j \in \{1, \dots, N\}$ 对蛋白质中的氨基酸编索引, A_i 是由i编索引的蛋白质中的氨基酸,并且 A_j 是由j编索引的蛋白质中的氨基酸。如果蛋白质包括多个氨基酸链,则蛋白质中的氨基酸可以根据蛋白质中的氨基酸链的排序从 $\{1, \dots, N\}$ 顺序编索引。也就是说,来自第一氨基酸链的氨基酸被首先依次索引,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。该组对嵌入112可以被表示为对嵌入的2-DN×N阵列,例如,其中2-D阵列的行由 $i \in \{1, \dots, N\}$ 编索引,2-D阵列的列由 $j \in \{1, \dots, N\}$ 编索引,并且2-D阵列中的位置(i,j)由氨基酸对 (A_i, A_j) 的对嵌入占据。

[0078] 参考图11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对嵌入的示例过程。

[0079] 系统100使用MSA表示110和对嵌入112两者生成定义预测蛋白质结构108的结构参数,因为两者具有互补特性。MSA表示110的结构可以明确地取决于与蛋白质中的每个氨基酸链相对应的MSA中的氨基酸链的数量。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中的氨基酸链的数量没有明确的依赖性。相反,对嵌入112表征蛋白质104中的相应氨基酸对之间的关系,并且在没有明确参考MSA的情况下表达,因此是用于预测蛋白质结构108的方便且有效的数据表示。

[0080] 系统100根据嵌入神经网络200的一组参数的值,使用嵌入神经网络200处理MSA表示110和对嵌入112,以更新MSA表示110和对嵌入112。也就是说,嵌入神经网络200处理MSA表示110和对嵌入112以生成更新的MSA表示114和更新的对嵌入116。

[0081] 嵌入神经网络200通过在MSA表示110和对嵌入112之间共享信息来更新MSA表示110和对嵌入112。更具体地,嵌入神经网络200在基于当前对嵌入112更新当前MSA表示110和基于当前MSA表示110更新当前对嵌入112之间交替。

[0082] 参考图2更详细地描述嵌入神经网络200的示例架构。

[0083] 系统100从更新的对嵌入116、更新的MSA表示114或两者生成用于主折叠神经网络600的网络输入,并使用主折叠神经网络600处理网络输入以生成主结构预测106,即,生成定义预测蛋白质结构的结构参数。

[0084] 在一些实施方式中,主折叠神经网络600处理更新的对嵌入116以生成距离图,对于蛋白质中的每个氨基酸对,距离图包括对于蛋白质结构中的氨基酸对之间的一组可能距离的概率分布。例如,为了生成蛋白质结构中的氨基酸对之间的一组可能距离集合上的概率分布,折叠神经网络可以将一个或多个全连接神经网络层应用于与该氨基酸对相对应的

更新的对嵌入116。

[0085] 在一些实施方式中,主折叠神经网络600通过使用几何注意操作处理从更新的MSA表示114和更新的对嵌入116两者导出的网络输入来生成结构参数,该几何注意操作明确地推出蛋白质结构中的氨基酸的3-D几何形状。参考图6描述实现几何注意机制的主折叠神经网络600的示例架构。

[0086] 训练引擎从端到端训练蛋白质结构预测系统100以优化损失函数,损失函数包括在本文中被称为结构损失的项。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)训练输入,包括蛋白质的初始MSA表示和初始对嵌入;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。用于训练系统100的目标蛋白质结构可以使用实验技术来确定,例如x射线晶体学或冷冻电子显微拷贝。

[0087] 结构损失可以表征以下之间的相似性:(i)由主折叠神经网络600生成的预测蛋白质结构,和(ii)应当由主折叠神经网络600生成的目标蛋白质结构。

[0088] 例如,如果预测的结构参数定义蛋白质中的每个氨基酸的预测的位置参数和预测的旋转参数,则结构损失可以由下式给出:

$$[0089] \quad \mathcal{L}_{structure} = \frac{1}{N^2} \sum_{i,j=1}^N \left(1 - \frac{|t_{ij} - \widetilde{t}_{ij}|}{A} \right)_+ \quad (2)$$

$$[0090] \quad t_{ij} = R_i^{-1}(t_j - t_i) \quad (3)$$

$$[0091] \quad \widetilde{t}_{ij} = \widetilde{R}_i^{-1}(\widetilde{t}_j - \widetilde{t}_i) \quad (4)$$

[0092] 其中,N是蛋白质中的氨基酸的数量,, t_i 表示氨基酸i的预测位置参数, R_i 表示由氨基酸i的预测旋转参数指定的 3×3 旋转矩阵, $\widetilde{t}_i$ 是氨基酸i的靶位置参数, $\widetilde{R}_i$ 表示由氨基酸i的靶旋转参数指定的 3×3 旋转矩阵,A是常数, R_i^{-1} 指由预测旋转参数 R_i 指定的 3×3 旋转矩阵的逆, $\widetilde{R}_i^{-1}$ 指由靶旋转参数 $\widetilde{R}_i$ 指定的 3×3 旋转矩阵的逆,并且 $(\cdot)_+$ 表示线性修正单元(ReLU)运算。

[0093] 参考等式(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 $|t_{ij} - \widetilde{t}_{ij}|$ 进行平均。 t_{ij} 项定义氨基酸j在氨基酸i的预测参考系中的预测空间位置,并且 $\widetilde{t}_{ij}$ 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的预测和实际旋转敏感,因此比仅对氨基酸之间的预测和实际距离敏感的损失项携带更丰富的信息。

[0094] 作为另一个示例,如果预测结构参数定义了蛋白质的每个氨基酸中的每个原子的预测空间位置,则结构损失可以是(i)原子的预测空间位置和(ii)原子的靶(例如,基准真值(ground truth))空间位置之间的平均误差(例如,平方误差)。

[0095] 优化结构损失鼓励系统100生成准确地近似真实蛋白质结构的预测蛋白质结构。

[0096] 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,即,损失函数可以包括与辅助损失中的每一个对应的附加项。辅助损失可以例如基于预

测结构中的氨基酸中的原子之间的键的键角和/或键长,或基于预测结构中的不同氨基酸中的原子的接近度,惩罚具有在天然世界中不太可能发生的特征的预测结构。

[0097] 训练引擎可以例如使用随机梯度下降训练技术,通过多次训练迭代在训练数据上训练结构预测系统100。

[0098] 另外,至少在训练期间,系统100还包括一个或多个辅助折叠神经网络620。每个辅助折叠神经网络620通常具有与主折叠神经网络600相同的架构,并且生成相应的辅助结构预测622,辅助结构预测622是与主结构预测106相同类型的输出,但是基于不同的信息。

[0099] 特别地,如下面将更详细描述,每个辅助折叠神经网络620接收包括嵌入神经网络200的不同中间输出的输入,即,代替更新的对嵌入116(以及可选地更新的MSA表示114),每个辅助折叠神经网络接收对嵌入的不同中间集合(以及可选地对应的中间MSA表示)作为输入。

[0100] 每个辅助折叠神经网络620处理从网络620接收的中间输出生成的输入,以生成辅助结构预测622,其像主结构预测106一样,预测输入蛋白质的结构。

[0101] 为了考虑包括辅助折叠神经网络(或多个)620,训练引擎增加损失函数以包括用于每个辅助折叠神经网络620的相应附加结构损失项。每个神经网络620的结构损失项表征(i)由辅助折叠神经网络620生成的预测蛋白质结构和(ii)应该由系统生成的目标蛋白质结构之间的相似性。每个辅助结构损失项通常与用于评估由主折叠神经网络600生成的主结构预测106的结构损失(即,上述结构损失之一)具有相同的形式。

[0102] 在训练期间,训练引擎将每个辅助结构损失项的梯度反向传播到嵌入神经网络200中,使得除了主结构损失项之外,还基于辅助结构损失项的梯度来更新嵌入神经网络200的参与生成中间输出的那些组件的参数。换句话说,由任何给定的辅助折叠神经网络620生成的辅助结构预测用于更新给定的辅助折叠神经网络620的参数的参数值和嵌入神经网络200的组件的参数值,由给定的折叠神经网络620使用以生成辅助结构预测的中间输出的生成以给定的辅助折叠神经网络620的参数的参数值和嵌入神经网络200的组件的参数值为条件。

[0103] 可选地,辅助结构损失项中的每一个可以在整体损失函数中被分配比主结构损失更低的权重,即,使得作为主结构预测的结果生成的梯度信号被赋予比作为任何给定辅助结构预测的结果生成的梯度信号更多的权重。

[0104] 在一些实施方式中,训练引擎在训练期间对辅助神经网络(或多个)620的参数的值施加约束。作为一个示例,训练引擎可以约束辅助神经网络(或多个)620的参数的值,使得网络620共享参数值,即,约束任何给定参数的值对于所有辅助神经网络620是相同的。作为另一示例,训练引擎可以约束辅助神经网络(或多个)620和主折叠网络600的参数的值,使得网络620和网络600共享参数值,即,约束任何给定参数的值对于所有辅助神经网络620和主折叠网络600是相同的。

[0105] 可选地,训练引擎还可以在训练神经网络时利用附加的辅助损失。特别地,下面图6的描述描述了可以基于由主折叠网络600生成的中间结构参数来计算的附加辅助损失。相同的辅助损失也可以应用于由每个辅助折叠网络620生成的中间结构参数。

[0106] 可选地,辅助神经网络620也可以在训练之后被包括作为系统100的一部分。在这些情况下,每个辅助折叠神经网络620使用由神经网络620生成的预测输出622来更新由辅

助折叠神经网络620作为输入接收的嵌入神经网络200的中间输出。然后,辅助神经网络将更新的中间输出提供给嵌入神经网络以进行进一步处理,即,代替原始中间输出。因此,在这些情况下,由辅助折叠神经网络(或多个)620做出的预测可以用于改善由嵌入神经网络200生成的更新的嵌入的质量,该更新的嵌入作为输入提供给主折叠神经网络600。

[0107] 在一些实施方式中,代替反馈由任何给定辅助折叠神经网络620做出的预测622或除了反馈由任何给定辅助折叠神经网络620做出的预测622之外,神经网络620可以将由网络620生成的特征作为生成预测622的一部分提供给嵌入神经网络。特别地,神经网络620可以选择由网络620生成的任何适当的中间输出,将该输出变换为特征,然后将特征反馈到嵌入神经网络中。

[0108] 图2示出嵌入神经网络200的示例架构,该嵌入神经网络200被配置为处理MSA表示110和对嵌入112以生成更新的MSA表示114和更新的对嵌入116。

[0109] 嵌入神经网络200包括更新块序列202-A-N。在整个说明书中,“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的子网络。

[0110] 嵌入神经网络中的每个更新块被配置为接收包括MSA表示和对嵌入的块输入,并且处理块输入以生成包括更新的MSA表示和更新的对嵌入的块输出。

[0111] 特别地,虽然图2和随后的附图的描述描述了嵌入神经网络200的架构,其中每个块更新对嵌入和MSA表示两者,但是所描述的用于将辅助折叠神经网络并入结构预测系统中的技术可以与包括多个块的嵌入神经网络的任何适当的架构一起使用,其中每个块更新对嵌入。作为下面描述的架构的替代架构的特定示例,嵌入神经网络200可以采用如下架构:其中使用如上所述的MSA表示110来初始化对嵌入112,并且其中每个块仅接收当前对嵌入并通过应用如下所述的自注意来更新那些嵌入,但不依赖于MSA表示。

[0112] 嵌入神经网络200将在嵌入神经网络200的网络输入中包括的MSA表示110和对嵌入112提供给第一更新块(即,在更新块序列中)。第一更新块处理MSA表示110和对嵌入112以生成更新的MSA表示和更新的对嵌入。

[0113] 对于第一更新块之后的每个更新块,嵌入神经网络200向该更新块提供由前一更新块生成的MSA表示和对嵌入,并且对于除了最后一个更新块之外的每个更新块,向下一更新块提供由该更新块生成的更新的MSA表示和更新的对嵌入。

[0114] 嵌入神经网络200通过使用更新块202-A-N的序列重复更新MSA表示110和对嵌入112来逐渐丰富MSA表示110和对嵌入112的信息内容。

[0115] 嵌入神经网络200可以提供由最终更新块(即,在更新块序列中)生成的更新的MSA表示114和更新的对嵌入116作为网络输出。

[0116] 如图2中所示,系统也包含对应于更新块202-M(其中块202-M是更新块序列中的任何更新块;在一些实施方案中,其可以是第一更新块202-A,但其通常不是最后一个更新块202-N)且生成辅助结构预测622的单个辅助折叠神经网络620。辅助折叠神经网络620至少接收由对应的更新块——即,更新块202-M——生成的更新的对嵌入作为输入,并且使用那些更新的对嵌入来生成辅助结构预测622。可选地,如上所述,辅助折叠神经网络620也可以接收由更新块202-M生成的更新的MSA表示。辅助折叠网络620可以以上面和下面参考主折叠网络600描述的任何方式生成预测622。

[0117] 在训练期间,训练引擎使用从辅助结构预测622计算的结构损失来通过反向传播

更新辅助折叠神经网络620的参数值以及更新块202M和更新块序列中更新块202M之前的更新块——即,更新块202A-L——的参数值。包括该信号作为通过系统反向传播的梯度信号的一部分可以向更新块提供更丰富的反馈,并且导致由嵌入神经网络200生成更高质量的最终输出。

[0118] 虽然图2中示出仅单个辅助折叠网络620,但系统可包含多个辅助神经网络620,每个辅助神经网络620对应于更新块中的不同更新块。作为一些示例,每个更新块可以具有对应的辅助神经网络620,序列中的每隔一个更新块可以具有对应的辅助神经网络620,或者仅在序列中的特定位置之后的更新块可以具有对应的辅助神经网络620。

[0119] 因此,更一般地,辅助神经网络620中的每一个对应于更新块中的不同的一个,并且至少接收由对应的更新块生成的更新后的对嵌入作为输入。每个辅助折叠网络620使用所接收的输入来生成辅助结构预测,并且训练引擎使用从辅助结构预测计算的损失通过反向传播来更新辅助神经网络、对应的更新块以及在更新块序列中在对应的更新块之前的任何更新块的参数值。

[0120] 在一些实施方式中,在系统的训练完成之后不使用辅助神经网络(或多个)620。

[0121] 在其他实施方式中,辅助神经网络(或多个)620在训练之后保持系统的一部分。在这些实施方式中,在训练期间和训练之后,即在推断时,每个辅助折叠网络620使用由网络620生成的预测622来进一步更新由网络620从对应的更新块作为输入接收的对嵌入。辅助折叠网络620然后将这些进一步更新的对嵌入作为输入提供到紧接在序列中的对应更新块之后的更新块,即,取代由对应更新块最初产生的更新的嵌入。

[0122] 在图2的示例中,辅助折叠网络620更新从更新块202-M接收的对嵌入,并提供进一步更新的对嵌入作为下一更新块——即,更新块202-N——的输入,代替由更新块202-M产生的对嵌入。

[0123] 辅助折叠网络620能够以各种方式中的任何一种使用辅助结构预测622来更新从更新块202-M接收的对嵌入。

[0124] 然而,通常,网络620将结构预测622变换为具有与更新的对嵌入相同的维度的嵌入,然后经变换的结构预测与更新的对嵌入组合以生成进一步更新的对嵌入。

[0125] 在一些实施方式中,网络620直接组合经变换的结构预测和更新的对嵌入,例如,网络620将经变换的结构预测和更新的对嵌入相加、平均或级联。

[0126] 在一些其他实施方式中,网络620可以通过应用学习的组合——即,与嵌入神经网络和主折叠神经网络的训练联合学习的组合——来组合更新的对嵌入和变换的结构预测。作为特定示例,网络620可以应用学习的交叉注意机制来合并来自经变换的结构预测的信息。与下面描述的自注意机制不同,在交叉注意机制中,查询是从更新的对嵌入生成的,而密钥和值是从经变换的结构预测生成的。

[0127] 特别地,如上所述,在一些实施方式中,辅助结构预测622是表征蛋白质中的每个氨基酸对之间的相应估计距离的距离图。在这些情况下,网络620可以将距离图投影到与更新的对嵌入相同的维度,以生成经变换的结构预测。

[0128] 例如,系统可以通过将每个距离分配给多个仓中的仓来离散距离图中的每个距离,每个仓对应于不同的距离值范围。然后,系统可以例如通过将离散化距离的独热编码表示乘以嵌入矩阵来嵌入每个离散化距离。独热编码表示将离散化距离表示为向量,其在对

应于离散化距离的位置处包括“1”并且在对应于所有其他离散化距离的所有其他位置处包括“0”。

[0129] 作为另一示例,系统可以将内核函数应用于距离图中的每个距离以生成距离的表示,然后嵌入表示。例如,内核函数可以等于函数 $\exp(-\text{distance})$ 或更一般地与函数 $\exp(-\text{distance})$ 成正比。

[0130] 作为另一示例,如上所述,在一些其他实施方式中,辅助结构预测622包括定义蛋白质中的每个氨基酸中的指定原子的位置的结构参数。网络620可以基于由结构参数定义的位置,通过计算蛋白质中的氨基酸中的指定原子对之间的距离来构建距离图。在这些情况下,网络620可以将距离图投影到与更新的对嵌入相同的维度,以生成如上所述的经变换的结构预测。

[0131] 图3示出嵌入神经网络200的更新块300的示例架构,即,如参考图2所描述。

[0132] 更新块300接收包括当前MSA表示302和当前对嵌入304的块输入,并处理块输入以生成更新的MSA表示306和更新的对嵌入308。

[0133] 更新块300包括MSA更新块400和对更新块500。

[0134] MSA更新块400使用当前对嵌入304来更新当前MSA表示302,并且对更新块500使用更新的MSA表示306(即,由MSA更新块400生成的)来更新当前对嵌入304。

[0135] 通常,MSA表示和对嵌入能够编码互补信息。例如,MSA表示能够编码关于一组进化相关的氨基酸链中不同位置的氨基酸的身份之间的相关性的信息,并且对嵌入能够编码关于蛋白质中的氨基酸之间的相互关系的信息。MSA更新块400使用在对嵌入中编码的互补信息来丰富MSA表示的信息内容,并且对更新块500使用在MSA表示中编码的互补信息来丰富对嵌入的信息内容。作为这种丰富的结果,更新的MSA表示和更新的对嵌入编码与预测蛋白质结构更相关的信息。

[0136] 更新块300在本文中被描述为首先使用当前对嵌入304来更新当前MSA表示302,然后使用更新的MSA表示306来更新当前对嵌入304。描述不应理解为将更新块限制为以此顺序执行操作,例如,更新块能够首先使用当前MSA表示来更新当前对嵌入,并且随后使用更新的对嵌入来更新当前MSA表示。

[0137] 更新块300在本文中被描述为包括MSA更新块400(即,其更新当前MSA表示)和对更新块500(即,其更新当前对嵌入)。不应将描述理解为将更新块300限制为仅包括一个MSA更新块或仅包括一对更新块。例如,更新块300能够包括多个MSA更新块,其在MSA表示被提供给对更新块以用于更新当前对嵌入之前多次更新MSA表示。作为另一示例,更新块300能够包括使用MSA表示多次更新对嵌入的多个对更新块。

[0138] MSA更新块400和对更新块500能够具有使它们能够执行其所描述功能的任何适当架构。

[0139] 在一些实施方式中,MSA更新块400、对更新块500或两者包括一个或多个“自注意”块。如贯穿本文档所使用的,自注意块通常是指更新嵌入的集合——即,接收嵌入的集合并输出更新的嵌入——的神经网络块。为了更新给定嵌入,自注意块能够确定给定嵌入与一个或多个选择的嵌入中的每一个之间的相应“注意权重”,然后使用(i)注意权重和(ii)选择的嵌入来更新给定嵌入。为了方便起见,可以说自注意块使用对于选择的嵌入“上”的注意来更新给定的嵌入。

[0140] 例如,自注意块可以接收输入嵌入的集合 $\{x_i\}_{i=1}^N$,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 ,自注意块可以确定注意权重 $[a_{i,j}]_{j=1}^N$,其中 $a_{i,j}$ 表示 x_i 与 x_j 之间的注意权重,为:

$$[0141] \quad [a_{i,j}]_{j=1}^N = \text{softmax}\left(\frac{(W_q x_i) K^T}{c}\right) \quad (5)$$

$$[0142] \quad K^T = [W_k x_j]_{j=1}^N \quad (6)$$

[0143] 其中, W_q 和 W_k 是学习的参数矩阵, $\text{softmax}(\cdot)$ 表示soft-max归一化操作,并且c是常数。使用注意权重,自注意层可以将嵌入 x_i 更新为:

$$[0144] \quad x_i \leftarrow \sum_{j=1..N} a_{i,j} \cdot (W_v x_j) \quad (7)$$

[0145] 其中, W_v 是学习的参数矩阵。 $(W_q x_i)$ 能够被称为用于输入嵌入 x_i 的“查询嵌入”, $W_k x_j$ 能够被称为用于输入嵌入 x_j 的“关键字(key)嵌入”,并且 $W_v x_j$ 能够被称为用于输入嵌入 x_j 的“值嵌入”。

[0146] 参数矩阵 W_q (“查询嵌入矩阵”)、 W_k (“关键字嵌入矩阵”)和 $(W_v$ “值嵌入矩阵”)是自注意块的可训练参数。在MSA更新块400和对更新块500中包括的任何自注意块的参数能够被理解是更新块300的参数,其能够作为参考图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。通常,查询、关键字和值嵌入矩阵的(训练的)参数对于不同的自注意块是不同的,例如,使得在MSA更新块400中包括的自注意块能够具有不同的查询、关键字和值嵌入矩阵,其具有与对更新块500中包括的自注意块不同的参数。

[0147] 在一些实施方式中,MSA更新块400、对更新块500或两者包括以对嵌入为条件的一个或多个自注意块,即,实现以对嵌入为条件的自注意操作的一个或多个自注意操作。为了使自注意操作以对嵌入为条件,自注意块能够处理对嵌入以生成与每个注意权重相对应的相应“注意偏置”。例如,除了根据等式(5) - (6) 确定注意权重 $[a_{i,j}]_{j=1}^N$ 之外,自注意块还能

够生成对应的一组注意偏置 $[b_{i,j}]_{j=1}^N$,其中, $b_{i,j}$ 表示 x_i 与 x_j 之间的注意偏置。自注意块能够通过将学习的参数矩阵应用于嵌入对 $h_{i,j}$,即,对于由(i, j) 编索引的蛋白质中的氨基酸对,来生成注意偏置 $b_{i,j}$ 。

[0148] 自注意块能够例如通过将注意权重和注意偏置求和(或以其他方式组合)来确定一组“偏置的注意权重” $[c_{i,j}]_{j=1}^N$,其中 $c_{i,j}$ 表示 x_i 与 x_j 之间的偏置的注意权重。例如,自注意块能够将嵌入 x_i 与 x_j 之间的偏置的注意权重确定为:

$$[0149] \quad c_{i,j} = a_{i,j} + b_{i,j}$$

[0150] 其中, $a_{i,j}$ 是 x_i 与 x_j 之间的注意权重,并且 $b_{i,j}$ 是 x_i 与 x_j 之间的注意偏置。自注意块能够使用偏置的注意权重来更新每个输入嵌入 x_i ,例如:

$$[0151] \quad x_i \leftarrow \sum_{j=1..N} c_{i,j} \cdot (W_v x_j) \quad (8)$$

[0152] 其中 W_v 是学习的参数矩阵。

[0153] 通常,对嵌入编码表征蛋白质的结构和蛋白质的结构中的氨基酸对之间的关系的的信息。将以对嵌入为条件的自注意操作应用于一组输入嵌入允许以由在对嵌入中编码的蛋白质结构信息通知的方式来更新输入嵌入。嵌入神经网络的更新块能够使用以对嵌入为条件的自注意块来更新和丰富MSA表示和对嵌入本身。

[0154] 可选地,自注意块可以具有多个“头部”,每个头部生成与每个输入嵌入相对应的相应更新的嵌入,即,使得每个输入嵌入与多个更新的嵌入相关联。例如,每个头部可以根据参考等式(5) - (7) 描述的参数矩阵 W_q 、 W_k 和 W_v 的不同值生成更新的嵌入。具有多个头部的自注意块能够实现“选通”操作,以组合由用于输入嵌入的头部生成的更新的嵌入,即,生成与每个输入嵌入相对应的单个更新的嵌入。例如,自注意块能够使用一个或多个神经网络层(例如,全连接神经网络层)来处理输入嵌入,以生成每个头部的相应的选通值。然后,自注意块能够根据选通值组合与输入嵌入相对应的更新的嵌入。例如,自注意块能够将输入嵌入 x_i 的更新的嵌入生成为:

$$[0155] \quad \sum_{k=1}^K \alpha_k \cdot x_i^{next} \quad (9)$$

[0156] 其中, k 对头部编索引, α_k 是头部 k 的选通值, x_i^{next} 是由头部 k 针对输入嵌入 x_i 生成的更新的嵌入。

[0157] 参考图4描述使用以对嵌入为条件的自注意块的MSA更新块400的示例架构。参考图4描述的示例MSA更新块通过使用以当前对嵌入为条件的自注意块处理当前MSA表示的行,基于当前对嵌入来更新当前MSA表示。

[0158] 参考图5描述使用以对嵌入为条件的自注意块的对更新块500的示例架构。参考图5描述的示例对更新块通过计算更新的MSA表示的外积均值,将外积均值的结果加到当前对嵌入,并使用以当前对嵌入为条件的自注意块来处理当前对嵌入,基于更新的MSA表示来更新当前对嵌入。

[0159] 如图3中所示,更新块300具有对应的辅助折叠网络620。辅助折叠网络620接收由对更新块500生成的更新的对嵌入308(以及可选地由MSA更新400生成的更新的MSA表示306)作为输入,并且如上所述,从接收到的输入生成辅助结构预测622。然后,训练引擎使用辅助结构预测622来更新辅助折叠网络600、更新块300的组件以及更新块序列中更新块300之前的任何块的组件的参数值。

[0160] 在其中训练期间也使用辅助折叠网络620的实施方式中,辅助折叠网络620也使用辅助结构预测620以在将对嵌入提供给下一块之前进一步更新更新的对嵌入308,即,使得提供给序列中的下一块的对嵌入是进一步更新的对嵌入,其是如上所述的更新的对嵌入308和经变换的结构预测的组合。

[0161] 图4示出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)基于当前对嵌入来更新当前MSA表示306。

[0162] 为了更新当前MSA表示302,MSA更新块400使用以当前对嵌入为条件的自注意操作(即,“逐行”自注意操作)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嵌入对为条件的“逐行”自注意块402,例如,如参考图3所述,以针对当前MSA表示302的每一行生成更新的嵌入。可选地,MSA更新块能够将逐行自注意块402的输入加到逐行自注意块402的输出。使逐行自注意块402以当前对嵌入为条件使得MSA更新块400能够使用来自当前对嵌入的信息来丰富当前MSA表示302。

[0163] 然后,MSA更新块使用不以当前对嵌入为条件的自注意操作(即,“逐列”自注意操作)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嵌入对为条件的“逐列”自注意块404,以生成当前MSA表示302的每一列的更新的嵌入。作为不以当前对嵌入为条件的结果,逐列自注意块404使用注意权重(例如,如参考等式(5)-(6)所描述的)而不是偏置的注意权重(例如,如参考等式(8)所描述的)针对当前MSA表示的每一列生成更新的嵌入。可选地,MSA更新块能够将逐列自注意块404的输入加到逐列自注意块404的输出。

[0164] 然后,MSA更新块使用例如将一个或多个全连接神经网络层应用于当前MSA表示302的转换块来处理当前MSA表示302。可选地,MSA更新块400能够将转换块406的输入加到转换块406的输出。

[0165] MSA更新块能够输出由逐行自注意块402、逐列自注意块404和转换块406执行的操作产生的更新的MSA表示306。

[0166] 图5示出对更新块500的示例架构。对更新块500被配置为接收当前对嵌入304,并且(至少部分地)基于更新的MSA表示306来更新当前对嵌入304。

[0167] 为了更新当前对嵌入304,对更新块500将外积均值操作502应用于更新的MSA表示306,并将外积均值操作502的结果加到当前对嵌入304。

[0168] 外积均值操作定义了一系列操作,当被应用于表示为嵌入的 $M \times N$ 阵列的MSA表示时,所述一系列操作生成嵌入的 $N \times N$ 阵列,即,其中 N 是蛋白质中的氨基酸的数量。当前对嵌入304也能够被表示为嵌入的 $N \times N$ 阵列,并且将外积均值502的结果加到当前对嵌入304是指对嵌入的两个 $N \times N$ 阵列求和。

[0169] 为了计算外积均值,对更新块生成张量 $A(\bullet)$,例如,由下式给出:

$$\begin{aligned}
 & A(res1, res2, ch1, ch2) \\
 [0170] \quad & = \frac{1}{|rows|} \sum_{rows} LeftAct(row, res1, ch1) \\
 & \quad \cdot RightAct(row, res2, ch2)
 \end{aligned}$$

[0171] 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的数量, $|rows|$ 是MSA表示中的行数, $LeftAct(row, res1, ch1)$ 是应用于位于由“row”索引的行和由“res1”索引的列处的MSA表示的嵌入的通道 $ch1$ 的线性操作(例如,由矩阵乘法定义),并且 $RightAct(row, res2, ch2)$ 是应用于位于由“row”索引的行和由“res2”索引的列处的MSA表示的嵌入的通道 $ch2$ 的线性操作(例如,由矩阵乘法定义)。外积均值的结果通

过对张量A的(ch1, ch2)维度进行平坦化和线性投影来生成。可选地,对更新块能够执行一个或多个层归一化操作(例如,如参考Jimmy Lei Ba等人的“Layer Normalization”arXiv:1607.06450所描述的)作为计算外积均值的一部分。

[0172] 通常,更新的MSA表示306编码关于在一组进化相关的氨基酸链中的不同位置中的氨基酸的身份之间的相关性的信息。在更新的MSA表示306中编码的信息与预测蛋白质的结构相关,并且通过将在更新的MSA表示中编码的信息并入到当前对嵌入中(即,通过外积均值502的方式),对更新块500能够增强当前对嵌入的信息内容。

[0173] 在使用更新的MSA表示更新当前对嵌入304(即,通过外积均值502)之后,对更新块500使用以当前对嵌入为条件的自注意操作(即,“逐行”自注意操作)将当前对嵌入的布置的每一行中的当前对嵌入更新为 $N \times N$ 阵列。更具体地,对更新块500将当前对嵌入的每一行提供给也以当前对嵌入为条件的“逐行”自注意块504,例如,如参考图3所述,以生成用于每一行的更新的对嵌入。可选地,对更新块可以将逐行自注意块504的输入加到逐行自注意块504的输出。

[0174] 然后,对更新块500使用也以当前对嵌入为条件的自注意操作(即,“逐列”自注意操作)来更新当前对嵌入的 $N \times N$ 阵列的每一列中的当前对嵌入。更具体地,对更新块500将当前对嵌入的每个列提供给“逐列”自注意块506,该“逐列”自注意块506也以当前对嵌入为条件,以针对每个列生成更新的对嵌入。可选地,对更新块能够将逐列自注意块506的输入加到逐列自注意块506的输出。

[0175] 然后,对更新块500使用例如将一个或多个全连接神经网络层应用于当前对嵌入的转换块来处理当前对嵌入。可选地,对更新块500能够将转换块508的输入加到转换块508的输出。

[0176] 对更新块能够输出由逐行自注意块504、逐列自注意块506和转换块508执行的操作产生的更新的对嵌入308。

[0177] 图6示出主折叠神经网络600的示例架构,其生成定义预测蛋白质结构108的一组结构参数106。例如,在实施方式中,折叠神经网络600通过处理包括蛋白质中的每个氨基酸对的相应对嵌入116的输入以生成结构参数106的值,基于对嵌入来确定蛋白质的预测结构。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。虽然参考主折叠神经网络600描述图6中所示的架构,但辅助折叠神经网络(或多个)620也可各自具有此相同架构,不同之处在于每一个辅助折叠神经网络620将从嵌入神经网络中的不是更新块序列中的最后一个更新块的对应更新块接收输入。

[0178] 在实施方式中,折叠神经网络600为蛋白质中的每个氨基酸生成结构参数,其可以包括:(i)位置参数,和(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氨基酸中指定原子的预测3-D空间位置。氨基酸的旋转参数可以指定蛋白质结构中氨基酸的预测“取向”。更具体地,旋转参数可以指定3-D空间旋转操作,如果将其应用于位置参数的坐标系,则使得氨基酸中的三个“主链”原子相对于旋转坐标系采取固定位置。

[0179] 在实施方式中,折叠神经网络600对于蛋白质中的每个氨基酸生成结构参数,结构参数能够包括:(i)位置参数,和(ii)旋转参数。如前所述,氨基酸的位置参数可以指定蛋白质的结构中的氨基酸中的指定原子的预测3-D空间位置。氨基酸的旋转参数可以指定蛋白质的结构中的氨基酸的预测“方位”。更具体地,旋转参数可以指定3-D空间旋转操作,如果

将其应用于位置参数的坐标系,则使得氨基酸中的三个“主链”原子相对于旋转坐标系采取固定位置。

[0180] 在实施方式中,折叠神经网络600接收从最终MSA表示、最终对嵌入或两者导出的输入,并生成定义蛋白质的预测结构的结构参数106的最终值。例如,折叠神经网络600可以接收输入,该输入包括:(i)蛋白质中的每个氨基酸对的相应对嵌入116,(ii)蛋白质中的每个氨基酸的“单个”嵌入602的初始值,以及(iii)蛋白质中的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预测结构108的结构参数106的最终值。

[0181] 蛋白质结构预测系统100能够向折叠神经网络600提供作为嵌入神经网络的输出生成的对嵌入,如参考图1所述。

[0182] 蛋白质结构预测系统100能够从MSA表示114生成氨基酸的初始单个嵌入602,即,其被生成为嵌入神经网络的输出,如参考图1所述。例如,如上所述,MSA表示114能够被表示为嵌入的2-D阵列,其列数等于蛋白质中的氨基酸的数量,其中每一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能够通过来自与氨基酸相关联的MSA表示114的列的嵌入进行求和(或以其他方式组合)来生成蛋白质中的每个氨基酸的初始单个嵌入。作为另一示例,蛋白质结构预测系统100能够通过从MSA表示114的行中提取嵌入来生成蛋白质中的氨基酸的初始单个嵌入,所述行对应于其结构正在被估计的蛋白质的氨基酸序列。

[0183] 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。

[0184] 折叠神经网络600能够通过重复更新单个嵌入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个嵌入606(即,生成更新的单个嵌入612)并更新当前结构参数608(即,生成更新的结构参数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,可以与更新块交错的其他神经网络层或块。

[0185] 每个更新块610能够包括:(i)几何注意块616,以及(ii)折叠块618,接下来将更详细地描述其中的每一个。

[0186] 几何注意块616使用“几何”自注意操作来更新当前单个嵌入,该“几何”自注意操作明确地推断蛋白质结构中的氨基酸的3-D几何形状(即,如由结构参数定义的)。更具体地,为了更新给定单个嵌入,几何注意块616确定给定单个嵌入与一个或多个选择的单个嵌入中的每一个之间的相应注意权重,其中注意权重取决于当前单个嵌入、当前结构参数和对嵌入。然后,几何注意块616使用(i)注意权重、(ii)选择的单个嵌入和(iii)当前结构参数来更新给定单个嵌入。

[0187] 为了确定注意权重,几何注意块616处理每个当前单个嵌入以生成对应的“符号查询”嵌入、“符号关键字”嵌入和“符号值”嵌入。例如,几何注意块616可以针对与第*i*氨基酸相对应的单个嵌入 h_i 生成符号查询嵌入 q_i 、符号关键字嵌入 k_i 和符号值嵌入 v_i ,如下:

$$[0188] \quad q_i = \text{Linear}(h_i) \quad (10)$$

$$[0189] \quad k_i = \text{Linear}(h_i) \quad (11)$$

[0190] $v_i = \text{Linear}(h_i)$ (12)

[0191] 其中, $\text{Linear}(\cdot)$ 是指具有独立学习参数值的线性层。

[0192] 几何注意块616附加地处理每个当前单个嵌入以生成对应的“几何查询”嵌入、“几何关键字”嵌入和“几何值”嵌入。每个单个嵌入的几何查询嵌入、几何关键字嵌入和几何值嵌入各自是最初在对应氨基酸的局部参考系中生成的3-D点,然后使用氨基酸的结构参数将其旋转和平移到全局参考系。例如,几何注意块616可以针对与第i氨基酸相对应的单个嵌入 h_i 生成几何查询嵌入 q_i^p 、几何关键字嵌入 k_i^p 和几何值嵌入 v_i^p ,如下:

[0193] $q_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (13)

[0194] $k_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (14)

[0195] $v_i^p = R_i \cdot \text{Linear}_p(h_i) + t_i$ (15)

[0196] 其中, $\text{Linear}_p(\cdot)$ 是指具有独立学习参数值的线性层,其将 h_i 投射到3-D点(上标p表示数量是3-D点), R_i 表示由第i氨基酸的旋转参数指定的旋转矩阵,并且 t_i 表示第i氨基酸的位置参数。

[0197] 为了更新对应于氨基酸i的单个嵌入 h_i ,几何注意块616可以生成注意权重 $[a_j]_{j=1}^N$,其中N是蛋白质中的氨基酸的总数,并且 a_j 是氨基酸i与氨基酸j之间的注意权重,如下:

[0198] $[a_j]_{j=1}^N = \text{softmax} \left(\left[\frac{q_i \cdot k_j}{\sqrt{m}} + \alpha |q_i^p - k_j^p|_2^2 + (b_{i,j} \cdot w) \right]_{j=1}^N \right)$ (16)

[0199] 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的符号关键字嵌入,m表示 q_i 和 k_j 的维度, α 表示学习的参数, 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的几何关键字嵌入, $|\cdot|_2$ 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对嵌入116,并且w是学习的权重向量(或一些其他学习的投影操作)。

[0200] 通常,氨基酸对的嵌入隐含地编码与该对中的氨基酸之间的关系相关的信息,例如,该对中的氨基酸之间的距离。通过部分地基于氨基酸i和j的对嵌入来确定氨基酸i与氨基酸j之间的注意权重,折叠神经网络600利用来自对嵌入的信息来丰富注意权重,从而提高预测折叠结构的准确度。

[0201] 在一些实施方式中,几何注意块616生成多组几何查询嵌入、几何关键字嵌入和几何值嵌入,并且在确定注意权重时使用每组生成的几何嵌入。

[0202] 在生成对应于氨基酸i的单个嵌入 h_i 的注意权重之后,几何注意块616使用注意权重来更新单个嵌入 h_i 。特别地,几何注意块616使用注意权重来生成“符号返回”嵌入和“几何返回”嵌入,然后使用符号返回嵌入和几何返回嵌入来更新单个嵌入。几何注意块124可以将氨基酸i的符号返回嵌入 o_i 生成为例如:

[0203] $o_i = \sum_j a_j v_j$ (17)

[0204] 其中, $[a_j]_{j=1}^N$ 表示注意权重 (例如, 参考等式 (16) 所定义的), 并且每个 v_j 表示氨基酸 j 的符号值嵌入。几何注意块 616 可以将氨基酸 i 的几何返回嵌入 o_i^p 生成为例如:

$$[0205] \quad o_i^p = R_i^{-1} \cdot \left(\sum_j a_j v_j^p - t_i \right) \quad (18)$$

[0206] 其中, 几何返回嵌入 o_i^p 是 3-D 点, $[a_j]_{j=1}^N$ 表示注意权重 (例如, 参考等式 (16) 定义), 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。能够理解, 几何返回嵌入最初在全局参考系中生成, 然后旋转和平移到对应氨基酸的局部参考系。

[0207] 几何注意块 616 可以使用对应的符号返回嵌入 o_i (例如, 根据等式 (17) 生成的) 和几何返回嵌入 o_i^p (例如, 根据等式 (18) 生成的) 将氨基酸的单个嵌入 h_i 更新为例如:

$$[0208] \quad h_i^{next} = \text{LayerNorm} \left(h_i + \text{Linear}(o_i, o_i^p, |o_i^p|) \right) \quad (19)$$

[0209] 其中, h_i^{next} 是氨基酸 i 的更新的单个嵌入, $|\cdot|$ 是范数, 例如, L_2 范数, 并且 $\text{LayerNorm}(\cdot)$ 表示层归一化操作, 例如, 如参考以下所描述的: J.L.Ba, J.R.Kiros, G.E.Hinton, "Layer Normalization" arXiv:1607.06450 (2016)。

[0210] 使用具体的 3-D 几何嵌入来更新氨基酸的单个嵌入 606, 例如, 如参考等式 (13) - (15) 所述, 使得几何注意块 616 能够在更新单个嵌入时推断 3-D 几何形状。此外, 每个更新块以对整个蛋白质结构的旋转和平移不变的方式更新单个嵌入和结构参数。例如, 将相同的全局旋转和平移操作应用于提供给折叠神经网络 600 的初始结构参数将使得折叠神经网络 600 生成预测结构, 该预测结构以相同的方式全局旋转和平移, 但在其他方面是相同的。因此, 应用于初始结构参数的全局旋转和平移操作不会影响由折叠神经网络 600 从初始结构参数开始生成的预测蛋白质结构的准确度。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, 这是因为折叠神经网络 600 自动学习在蛋白质结构的所有旋转和平移上进行泛化。

[0211] 氨基酸的更新的单个嵌入可以在被提供给折叠块 618 之前通过几何注意块 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。

[0212] 在几何注意块 616 更新氨基酸的当前单个嵌入 606 之后, 折叠块 618 使用更新的单个嵌入 612 来更新当前结构参数 608。例如, 折叠块 618 可以将氨基酸 i 的当前位置参数 t_i 更新为:

$$[0213] \quad t_i^{next} = t_i + \text{Linear}(h_i^{next}) \quad (20)$$

[0214] 其中, t_i^{next} 是更新的位置参数, $\text{Linear}(\cdot)$ 表示线性神经网络层, 并且 h_i^{next} 表示氨基酸 i 的更新的单个嵌入。在另一示例中, 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 可以将当前旋转参数 R_i 更新为:

$$[0215] \quad w_i = \text{Linear}(h_i^{\text{next}}) \quad (21)$$

$$[0216] \quad R_i^{\text{next}} = R_i \cdot \text{QuaternionToRotation}(1 + w_i) \quad (22)$$

[0217] 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 的更新的单个嵌入, $1 + w_i$ 表示具有实部 1 和虚部 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。使用等式 (21) - (22) 更新旋转参数确保更新的旋转参数定义有效的旋转矩阵, 例如具有行列式 1 的标准正交矩阵。

[0218] 折叠神经网络 600 可以提供由最终更新块 610 生成的更新的结构参数作为定义预测蛋白质结构 108 的最终结构参数 106。折叠神经网络 600 可以包括任何适当数量个更新块, 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单个的一组参数值。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, 例如通过稳定训练并降低过度拟合的可能性。

[0219] 在训练期间, 训练引擎能够基于评估最终结构参数 106 的准确度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。在一些实施方式中, 训练引擎能够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更新的结构参数的准确度。

[0220] 可选地, 在训练期间, 训练引擎能够应用“停止梯度”操作以防止梯度反向传播通过每个更新块的某些神经网络参数, 例如, 用于计算更新的旋转参数的神经网络参数 (如等式 (21) - (22) 中所述)。应用这些停止梯度操作能够改善在训练期间计算的梯度的数值稳定性。

[0221] 通常, 由折叠神经网络 600 生成的预测蛋白质结构 108 与对应的基准真值蛋白质结构之间的相似性能够例如通过相似性测量来度量, 该相似性测量将相应的准确度分数分配给预测蛋白质结构中的多个原子中的每一个。例如, 相似性测量能够将相应的准确度分数分配给预测蛋白质结构中的每个碳阿尔法原子。预测蛋白质结构中的原子的准确度分数能够表征预测蛋白质结构中的原子的位置与基准真值蛋白质结构中的原子的实际位置的符合程度。能够将预测蛋白质结构与基准真值蛋白质结构进行比较以生成预测蛋白质结构中的原子的准确度分数的相似性测量的示例是参考以下文献描述的 1DDT 相似性测量: V.Mariani 等人, “1DDT: a local superposition-free score for comparing protein structures and models using distance difference tests”, *Bioinformatics*, 2013 年 11 月 1 日; 29(21) 2722-2728。

[0222] 折叠神经网络 600 能够被配置为针对预测蛋白质结构 108 中的一个或多个原子中的每一个生成相应的置信度估计 650。预测蛋白质结构中的原子的置信度估计 650 表征预测蛋白质结构中的原子的预测准确度分数 (例如, 1DDT 准确度分数), 即, 其将由将预测蛋白质结构与 (潜在未知的) 基准真值蛋白质结构进行比较的相似性测量生成。在一个示例中, 预测蛋白质结构中的原子的置信度估计 650 能够定义在一组区间上的离散概率分布, 该组区间形成原子的准确度分数的可能值范围的分区。离散概率分布能够将相应概率与区间中的每一个相关联, 该相应概率定义实际准确度分数被包括在区间中的可能性。例如, 准确度分

数的可能值的范围可以是 $[0, 100]$ ，并且置信度估计650可以定义一组区间 $\{[0, 2), [2, 4), \dots, [98, 100]\}$ 上的概率分布。在另一示例中，预测蛋白质结构中的原子的置信度估计650能够是数值，即，直接预测原子的准确度分数。

[0223] 在一些实施方式中，折叠神经网络600针对蛋白质的每个氨基酸中的指定原子(例如，阿尔法碳原子)生成相应的置信度估计650。折叠神经网络600能够生成蛋白质中的氨基酸中的指定原子的置信度估计650，例如，通过使用一个或多个神经网络层(例如，全连接层)处理由折叠神经网络中的最后一个更新块生成的氨基酸的更新的单个嵌入。

[0224] 结构预测系统能够基于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，结构预测系统能够生成氨基酸的置信度分数，作为氨基酸中的阿尔法碳原子的准确度分数的可能值上的概率分布的预期值。

[0225] 结构预测系统能够生成整个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。

[0226] 在结构预测系统的训练期间，训练引擎能够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，该辅助损失测量以下之间的误差：(i)由折叠神经网络600生成的置信度估计，和(ii)通过将预测蛋白质结构与基准真值蛋白质结构进行比较而生成的准确度分数。误差可以是例如交叉熵误差。

[0227] 由结构预测系统生成的置信度估计能够以各种方式使用。例如，预测蛋白质结构中的原子的置信度估计能够指示结构的哪些部分已经被可靠地估计，因此适合于进一步的下游处理或分析。作为另一个示例，每蛋白质置信度分数能够用于对蛋白质的结构的一组预测进行排序，例如，由相同的结构预测系统通过处理表征相同蛋白质的不同输入而生成的预测，或者由不同的结构预测系统生成的预测。

[0228] 由结构参数106指定的位置和旋转参数能够定义蛋白质的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)。然而，结构参数106不一定定义蛋白质的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。特别地，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、 Φ 角、psi角、chi1角、chi2角、chi3角和chi4角，如参考图7所示。

[0229] 可选地，折叠神经网络600的更新块610中的一个或多个能够生成定义蛋白质的每个氨基酸中的每个原子的相应的预测空间位置的输出。为了生成氨基酸中的原子的预测空间位置，更新块能够使用一个或多个神经网络层来处理氨基酸的更新的单个嵌入，以生成氨基酸中的原子之间的键的扭转角的预测值。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。每个扭转角可以表示为例如2-D向量。

[0230] 更新块能够基于以下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，和(ii)氨基酸的更新的结构参数(例如，位置和旋转参数)。举例来说，更新块能够根据预定义函数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。更新块能够通过根据氨基酸的更新的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，蛋白质中的所有氨基酸共有的参考系)中的空间位置。举例来说，更新块能够通过将由更新的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的空间位置，并且随后将由更新的位置参数定义的平移操作应用于旋转的空间位置来确定原子在全局参考系中的空间位置。

[0231] 在一些实施方式中,替代输出最终结构参数或与输出最终结构参数组合,折叠神经网络600输出由最终更新块生成的蛋白质的氨基酸中的原子的预测空间位置。

[0232] 参考图6描述的折叠神经网络600在本文中被表征为接收基于MSA表示114和由嵌入神经网络生成的对嵌入116的输入,例如,如参考图2所述。然而,通常,能够使用任何适当的技术来生成折叠神经网络的输入(例如,单个嵌入602和对嵌入116)。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)能够由其他折叠神经网络执行,例如,具有接收不同输入的不同架构。

[0233] 图7示出氨基酸中的键之间的扭转角,例如 ω 角、 Φ 角、psi角、chi1角、chi2角、chi3角、chi4角和chi5角。

[0234] 图8是未折叠蛋白质和折叠蛋白质的图示。未折叠蛋白质是氨基酸的无规则卷曲。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,诸如阿尔法螺旋(例如,如802所示)和贝塔折叠。

[0235] 图9是用于训练包括嵌入神经网络和主折叠神经网络的结构预测神经网络的示例过程900的流程图。为了方便起见,过程900将被描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如,根据本说明书适当编程的蛋白质结构预测系统(例如,图1的蛋白质结构预测系统100)可以执行过程900。

[0236] 系统获得包括表征给定蛋白质的数据的训练输入(步骤902)。例如,数据可以包括:(i)初始多序列比对(MSA)表示,其表示对应于蛋白质中的每条链的相应MSA;和(ii)蛋白质中的每个氨基酸对的相应初始对嵌入。

[0237] 系统通过处理训练输入获得指定应当由系统100生成的目标蛋白质结构的数据(步骤904)。

[0238] 系统使用嵌入神经网络和主折叠神经网络处理训练输入,例如,如上所述,以生成主结构预测(步骤906)。

[0239] 对于一个或多个辅助折叠神经网络中的每一个,系统处理包括由嵌入神经网络中的“隐藏”更新块中的对应一个生成的更新的嵌入的输入,以生成辅助结构预测(步骤908)。隐藏的更新块是序列中不是最后一个更新块的更新块。在辅助结构预测被反馈到嵌入神经网络中的实施方式中,系统在执行步骤906的同时执行步骤908,即,因为嵌入神经网络的输出取决于由辅助折叠神经网络(或多个)生成的辅助结构预测(或多个)。在辅助结构预测不被反馈到嵌入神经网络中的实施方式中,系统可以在执行步骤906的同时执行步骤908,或者在执行步骤906之后、即在生成主结构预测之后执行步骤908。

[0240] 系统计算关于损失函数的神经网络的每个参数的梯度(步骤910),损失函数至少包括:(a)表征(i)由主折叠神经网络生成的主结构预测定义的预测蛋白质结构和(ii)应该由系统生成的目标蛋白质结构之间的相似性的主结构损失;以及(b)每个辅助神经网络(或多个)的相应辅助结构损失。给定辅助神经网络的辅助结构损失表征(i)由给定辅助折叠神经网络生成的辅助结构预测定义的预测蛋白质结构与(ii)应该由系统生成的目标蛋白质结构之间的相似性。

[0241] 一旦满足阈值条件,例如,已经计算了整批训练输入的梯度,系统就可以根据用于训练的优化器(例如,Adam、rmsProp或SGD)的更新规则从梯度计算对于嵌入神经网络、主折叠神经网络和辅助折叠神经网络的参数值的更新。然后,系统可以将更新应用于参数的当

前值,即,从当前值中减去更新或将更新添加到当前值。

[0242] 在训练之后,系统可以在不执行步骤904、910的情况下执行过程900,并且在其中来自辅助折叠网络的输出不被反馈到嵌入神经网络中的实施方式中,步骤908,然后提供主结构预测或指定由主结构预测定义的结构的其他信息作为由输入表征的蛋白质的系统的输出。

[0243] 图10示出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能够实现过程1000的操作。

[0244] 为了生成蛋白质中的氨基酸链的MSA表示1010,系统100获得蛋白质的MSA1002,其可以包括例如数千个MSA序列。

[0245] 系统100将该组MSA序列划分为“核心”MSA序列组1004和“额外”MSA序列组1006。核心MSA序列组能够比额外MSA序列组1006小(例如,小一个数量级)。系统100能够将该组MSA序列划分为核心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。

[0246] 对于每个额外MSA序列1006,系统100能够确定额外MSA序列与每个核心MSA序列1004之间的相应相似性测量(例如,基于汉明距离)。然后,系统100能够将每个额外MSA序列1006与额外MSA序列1006最相似(即,根据相似性测量)的对应核心MSA序列1004相关联。与核心MSA序列1004相关联的额外MSA序列组1006能够被称为“MSA序列集群”1008。也就是说,系统100确定与每个核心MSA序列1004对应的相应MSA序列集群1008,其中对应于核心MSA序列1004的MSA序列集群1008包括与核心MSA序列1004最相似的一组额外MSA序列1006。

[0247] 系统100能够基于核心MSA序列和MSA序列集群1008生成蛋白质中的氨基酸链的MSA表示。MSA表示1010能够由嵌入的 $M \times N$ 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表示中的嵌入能够通过 $(i, j) \in \{(i, j) : i=1, \dots, M, j=1, \dots, N\}$ 来索引。

[0248] 为了生成在MSA表示1010中的位置 (i, j) 处的嵌入,系统100能够获得定义在核心MSA序列 i 中的位置 j 处的氨基酸的身份的嵌入(例如,独热嵌入)。系统100还能够基于在与核心MSA序列 i 对应的MSA序列集群1008中的额外MSA序列1006中的位置 j 处的每个可能氨基酸的相对出现频率来确定一组可能氨基酸的概率分布。然后,系统100能够通过组合(例如,级联)以下来确定在MSA表示中的位置 (i, j) 处的嵌入:(i)定义在核心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序列中的位置 j 对应的可能氨基酸的概率分布。

[0249] 在一些情况下,对于一个或多个核心MSA序列,(基准真值)蛋白质结构可以是已知的。特别地,对于一个或多个核心MSA序列,核心MSA序列中的氨基酸中的键之间的扭转角(例如, ω 角、 Φ 角、 ψ 角等)的值可以是已知的。如果核心MSA序列 i 中的氨基酸的扭转角的值是已知的,则系统100能够至少部分地基于核心MSA序列 i 中的氨基酸 j 的扭转角的值来生成在MSA表示中的位置 (i, j) 处的嵌入。例如,系统能够使用一个或多个神经网络层来生成扭转角的值的嵌入,并且将扭转角的值的嵌入级联到在MSA表示中的位置 (i, j) 处的嵌入。

[0250] 图11示出用于针对蛋白质中的每个氨基酸对生成(初始化)相应的对嵌入112的示例过程1100。参考图1描述的蛋白质结构预测系统100能够实现过程1100的操作。

[0251] 系统100能够使用蛋白质的MSA表示1102生成对嵌入112。参考图1和图10更详细地

描述生成蛋白质的MSA表示。系统100能够基于蛋白质中的每条氨基酸链的相应MSA表示生成蛋白质的MSA表示1102。为了生成蛋白质中的氨基酸链的MSA表示,系统100能够使用比用于生成参考图1描述的MSA表示110更多(例如,多一个数量级)的MSA序列。因此,由系统100用于生成对嵌入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)。在一些实施方式中,系统100能够使用参考图10描述的额外MSA序列1006来生成MSA表示1102。

[0252] 在生成MSA表示1102之后,系统100处理MSA表示1102以从MSA表示1102生成对嵌入1104,例如,通过将外积均值操作应用于MSA表示1102,并将对嵌入1104识别为外积均值操作的结果。

[0253] 系统100使用嵌入神经网络1106处理MSA表示1102和对嵌入1104。嵌入神经网络1106能够通过MSA表示1102与对嵌入1104之间共享信息来更新MSA表示1102和对嵌入1104。更具体地,嵌入神经网络1106能够在基于对嵌入1104更新MSA表示1102与基于MSA表示1102更新对嵌入1104之间交替。

[0254] 嵌入神经网络1106能够具有基于参考图2-5描述的嵌入神经网络架构的架构,即,使用逐行和逐列自注意块来更新对嵌入1104和MSA表示1102。

[0255] 在一些实施方式中,嵌入神经网络1106能够使用逐列“全局”自注意操作来更新MSA表示1102的每一列中的嵌入。更具体地,嵌入神经网络1106能够将MSA表示的每一列中的嵌入提供给逐列全局自注意块,以针对当前MSA表示的每一列生成更新的嵌入。为了实现全局逐列自注意,自注意块能够针对列中的每个嵌入生成相应的查询嵌入,然后对查询嵌入求平均以生成列的单个“全局”查询嵌入。然后,逐列自注意块使用单个全局查询嵌入来执行自注意操作,这能够将自注意操作的复杂度从二次(即,每一列的嵌入数量)降低到线性。使用全局自注意操作能够降低逐列自注意操作的计算复杂度,以使得能够对包括大量(例如,数千个)嵌入的MSA表示1102的列执行逐列自注意操作。

[0256] 在使用嵌入神经网络1106更新对嵌入1104和MSA表示1102之后,系统100能够将对嵌入112识别为由嵌入神经网络1106生成的更新的对嵌入。系统100能够丢弃由嵌入神经网络1106生成的更新的MSA表示,或者以任何适当的方式使用它。

[0257] 作为生成对嵌入112的一部分,系统100能够在对应于蛋白质中的每个氨基酸对的相应对嵌入中包括相对位置编码信息。该系统能够通过以下方式包括对应的对嵌入中的相同氨基酸链中包括的氨基酸对的相对位置编码信息:计算表示分离氨基酸链中的氨基酸对的氨基酸数量的有符号差,将结果剪切到预定义区间,使用独热编码向量表示剪切值,对独热编码向量应用线性变换,并将线性变换的结果加到对应的对嵌入。该系统能够通过将默认编码向量加到对应的对嵌入来包括没有包括在对应的对嵌入中的相同氨基酸链中的氨基酸对的相对位置编码信息,该默认编码向量指示该氨基酸对没有包括在相同氨基酸链中。

[0258] 系统100还能够至少部分地基于一组一个或多个模板序列1110来生成对嵌入112。每个模板序列1110是蛋白质中的氨基酸链的MSA序列,其中模板序列1110的折叠结构例如从物理实验中已知。

[0259] 系统100能够生成每个模板序列1110的相应模板表示1112。模板序列1110的模板表示1112包括对应于模板序列中的每个氨基酸对的相应嵌入,例如,使得长度 n (即,具有 n

个氨基酸)的模板序列的模板表示1112能够表示为嵌入的 $n \times n$ 阵列。系统100基于例如以下项来生成在模板序列1110的模板表示1112中的位置 (i, j) 处的嵌入:(i)表示在模板序列的位置 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)由在模板序列中(即模板序列的折叠结构中)的位置 i 和 j 处的氨基酸中的相应碳阿尔法原子的空间位置的差异定义的单位向量,其中单位向量在氨基酸 i 或氨基酸 j 的参考系中计算,和(iii)在模板序列中的位置 i 和 j 处的氨基酸中的相应碳阿尔法原子的空间位置之间的距离的离散/装仓表示。

[0260] 系统100能够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,以生成对应于每个模板序列1110的相应更新的模板表示1116。模板更新块能够包括例如逐行自注意块(例如,其更新模版表示的每一行中的嵌入)、逐列自注意块(例如,其更新模版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版表示中的每个嵌入)。

[0261] 在生成更新的模版表示1116之后,系统100使用更新的模版表示1116来更新对嵌入112。例如,系统100能够使用对在更新的模版表示1116的对应 (i, j) 位置处的嵌入的“交叉注意”来更新在每个位置 (i, j) 处的相应的对嵌入112。在用于更新在位置 (i, j) 处的对嵌入112的交叉注意操作中,从在位置 (i, j) 处的对嵌入生成查询嵌入,并且从在更新的模版表示1116的对应 (i, j) 位置处的嵌入生成关键字和值嵌入。

[0262] 使用模板序列1110更新对嵌入112使得系统100能够利用表征进化相关模板序列1110的蛋白质结构的信息来丰富对嵌入,从而增强对嵌入112的信息内容,并提高使用对嵌入预测的蛋白质结构的准确度。

[0263] 本说明书结合系统和计算机程序组件使用术语“配置”。对于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,意味着系统已经在其上安装了软件、固件、硬件或它们的组合,这些软件、固件、硬件或它们的组合在操作中使得系统执行操作或动作。对于被配置为执行特定操作或动作的一个或多个计算机程序,意味着一个或多个程序包括当由数据处理装置执行时使装置执行操作或动作的指令。

[0264] 本说明书中描述的主题和功能操作的实施例能够在数字电子电路、有形体现的计算机软件或固件、计算机硬件(包括本说明书中公开的结构及其结构等同物)或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能够实现为一个或多个计算机程序,即,在有形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,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质能够是机器可读存储设备、机器可读存储基板、随机或串行存取存储器设备、或它们中的一个或多个的组合。可替代地或另外地,程序指令能够被编码在人工生成的传播信号上,例如,机器生成的电、光或电磁信号,其被生成以编码信息以传输到合适的接收器装置以供数据处理装置执行。

[0265] 术语“数据处理装置”是指数据处理硬件,并且包含用于处理数据的所有种类的装置、设备和机器,例如包括可编程处理器、计算机或多个处理器或计算机。该装置还能够是或还包括专用逻辑电路,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,该装置能够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。

[0266] 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)能够以任何形式的编程语言编写,包括编译或解释语言,或者声明或过程

语言;并且它能够以任何形式部署,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。程序可以但不必对应于文件系统中的文件。程序能够存储在保存其他程序或数据的文件的一部分中,例如,存储在标记语言文档中的一个或多个脚本,存储在专用于所讨论的程序的单个文件中,或者存储在多个协调文件中,例如,存储一个或多个模块、子程序或代码部分的文件。计算机程序能够被部署为在一个计算机上或在位于一个站点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。

[0267] 在本说明书中,术语“引擎”广泛地用于指代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。通常,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,能够在相同的一个或多个计算机上安装和运行多个引擎。

[0268] 本说明书中描述的过程和逻辑流程能够由执行一个或多个计算机程序的一个或多个可编程计算机执行,以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也能够由专用逻辑电路(例如,FPGA或ASIC)执行,或者由专用逻辑电路和一个或多个编程计算机的组合执行。

[0269] 适合于执行计算机程序的计算机能够基于通用或专用微处理器或两者,或任何其他类型的中央处理单元。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执行或运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设备。中央处理单元和存储器能够由专用逻辑电路补充或并入专用逻辑电路中。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设备(例如磁盘、磁光盘或光盘),或者可操作地耦接以从其接收数据或向其传送数据或两者。然而,计算机不需要具有这样的设备。此外,计算机能够嵌入在另一设备中,仅举几例,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设备,例如通用串行总线(USB)闪存驱动器。

[0270] 适用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、介质和存储器设备,包括例如半导体存储器设备,例如EPROM、EEPROM和闪存设备;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。

[0271] 为了提供与用户的交互,本说明书中描述的主题的实施例能够在具有用于向用户显示信息的显示设备(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户能够通过其向计算机提供输入的键盘和定点设备(例如,鼠标或轨迹球)的计算机上实现。其他种类的设备也能够用于提供与用户的交互;例如,提供给用户的反馈能够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;并且能够以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,计算机能够通过向用户使用的设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从用户设备上的网络浏览器接收的请求将网页发送到网络浏览器。此外,计算机能够通过向个人设备(例如,运行消息传送应用的智能电话)发送文本消息或其他形式的消息并且从用户接收作为回应的响应消息来与用户交互。

[0272] 用于实现机器学习模型的数据处理装置还能够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的公共和计算密集部分,即推理、工作负载。

[0273] 能够使用机器学习框架(例如,TensorFlow框架、微软的Cognitive Toolkit框架、

Apache的Singa框架或Apache的MXNet框架)来实现和部署机器学习模型。

[0274] 本说明书中描述的主题的实施例能够在计算系统中实现,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,应用服务器),或者包括前端组件(例如,具有用户能够通过其与本说明书中描述的主题的实现交互的图形用户界面、web浏览器或app的客户端计算机),或者一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件能够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。

[0275] 计算系统能够包括客户端和服务端。客户端和服务端通常彼此远离,并且通常通过通信网络进行交互。客户端和服务端的关系借助于在相应计算机上运行并且彼此具有客户端-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)发送到用户设备,例如,用于向与充当客户端的设备交互的用户显示数据和从其接收用户输入的目的。能够在服务端处从设备接收在用户设备处生成的数据,例如,用户交互的结果。

[0276] 虽然本说明书包含许多具体实施方式细节,但是这些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,而是作为可以特定于特定发明的特定实施例的特征的描述。在本说明书中在分开实施例的上下文中描述的某些特征也能够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能够单独地或以任何合适的子组合在多个实施例中实现。此外,尽管特征可以在上面被描述为以某些组合起作用并且甚至最初如此要求保护,但是来自所要求保护的组合的一个或多个特征在一些情况下能够从组合中删除,并且所要求保护的组合可以针对子组合或子组合的变型。

[0277] 类似地,虽然在附图中描绘了操作并且在权利要求中以特定顺序叙述了操作,但是这不应被理解为要求以所示的特定顺序或按顺序执行这些操作,或者执行所有示出的操作,以实现期望的结果。在某些情况下,多任务和并行处理会是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中都需要这种分离,并且应当理解,所描述的程序组件和系统通常能够一起集成在单个软件产品中或封装到多个软件产品中。

[0278] 已经描述了主题的特定实施例。其他实施例在所附权利要求书的范围内。例如,权利要求中记载的动作能够以不同的顺序执行并且仍然实现期望的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定顺序或相继顺序来实现期望的结果。在一些情况下,多任务和并行处理会是有利的。

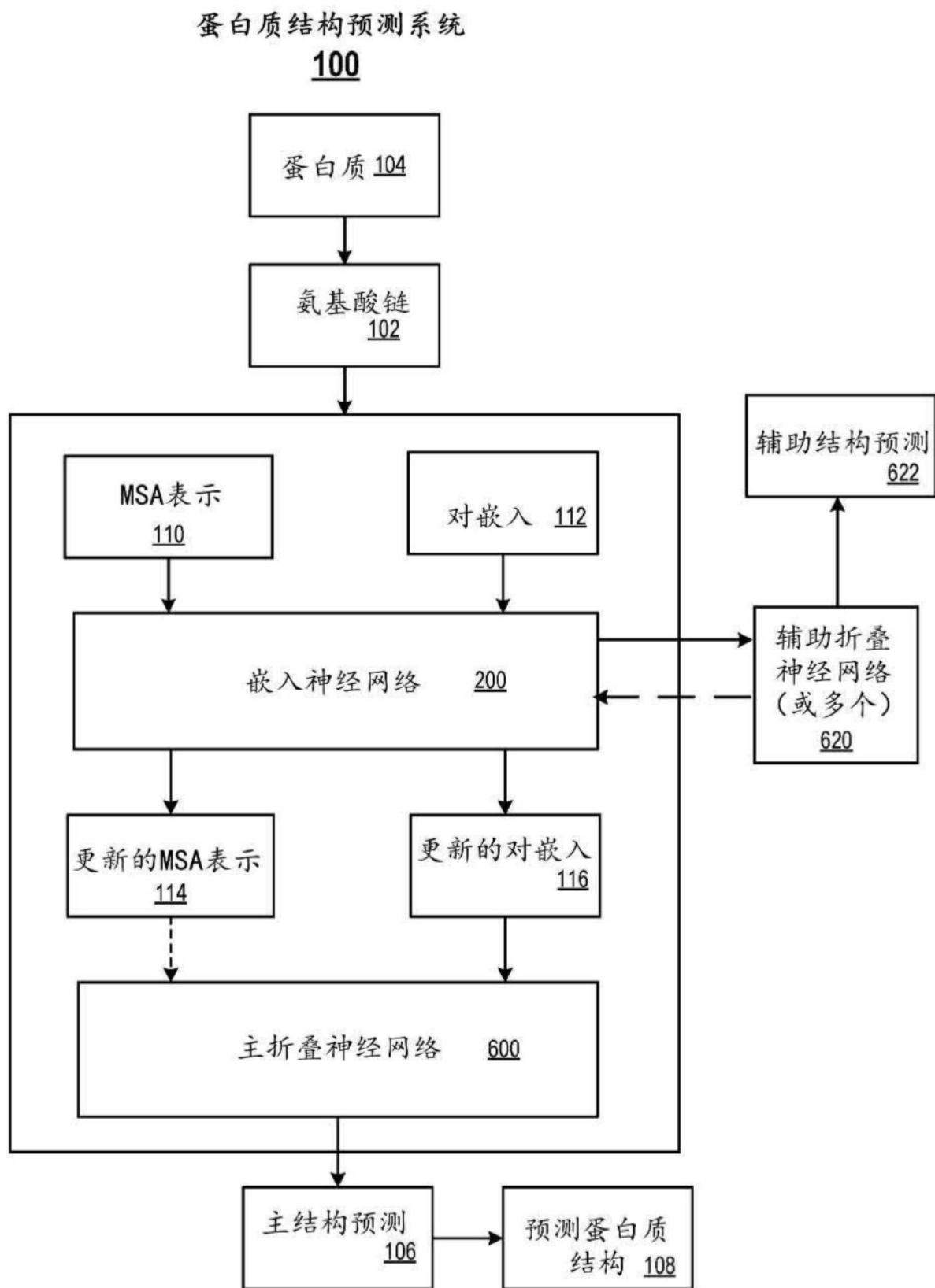


图1

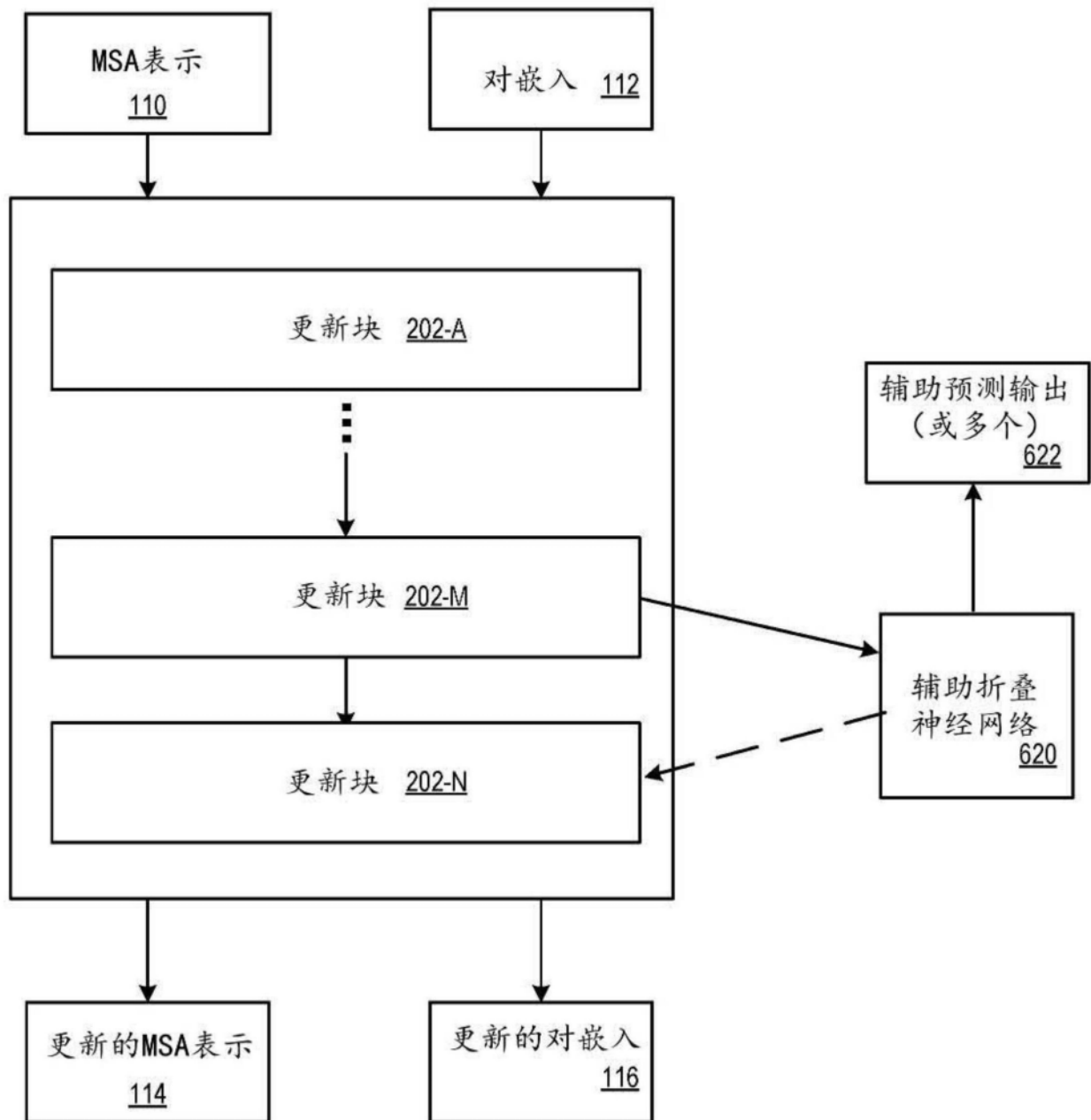
嵌入神经网络 200

图2

嵌入神经网络更新块

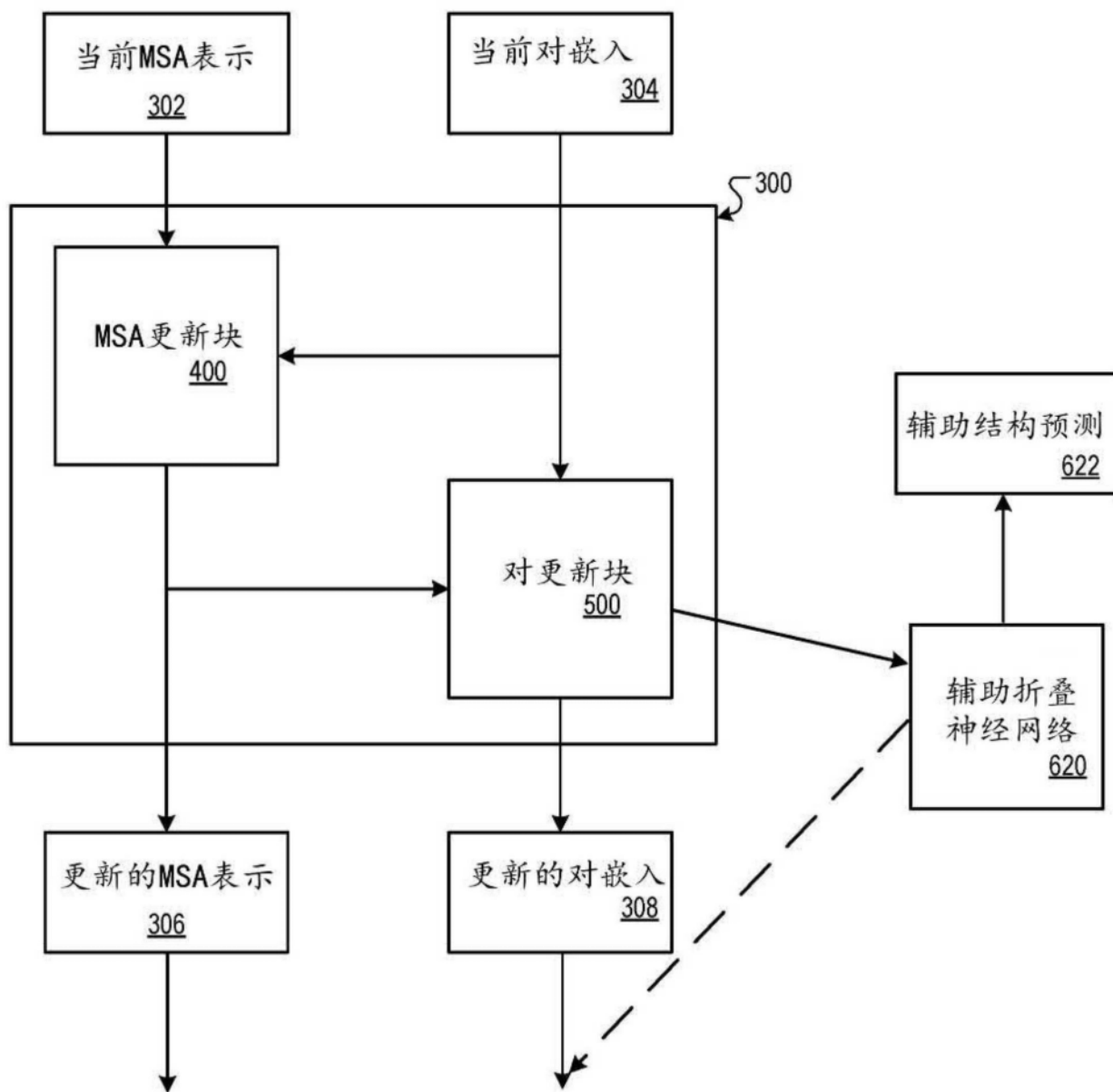


图3

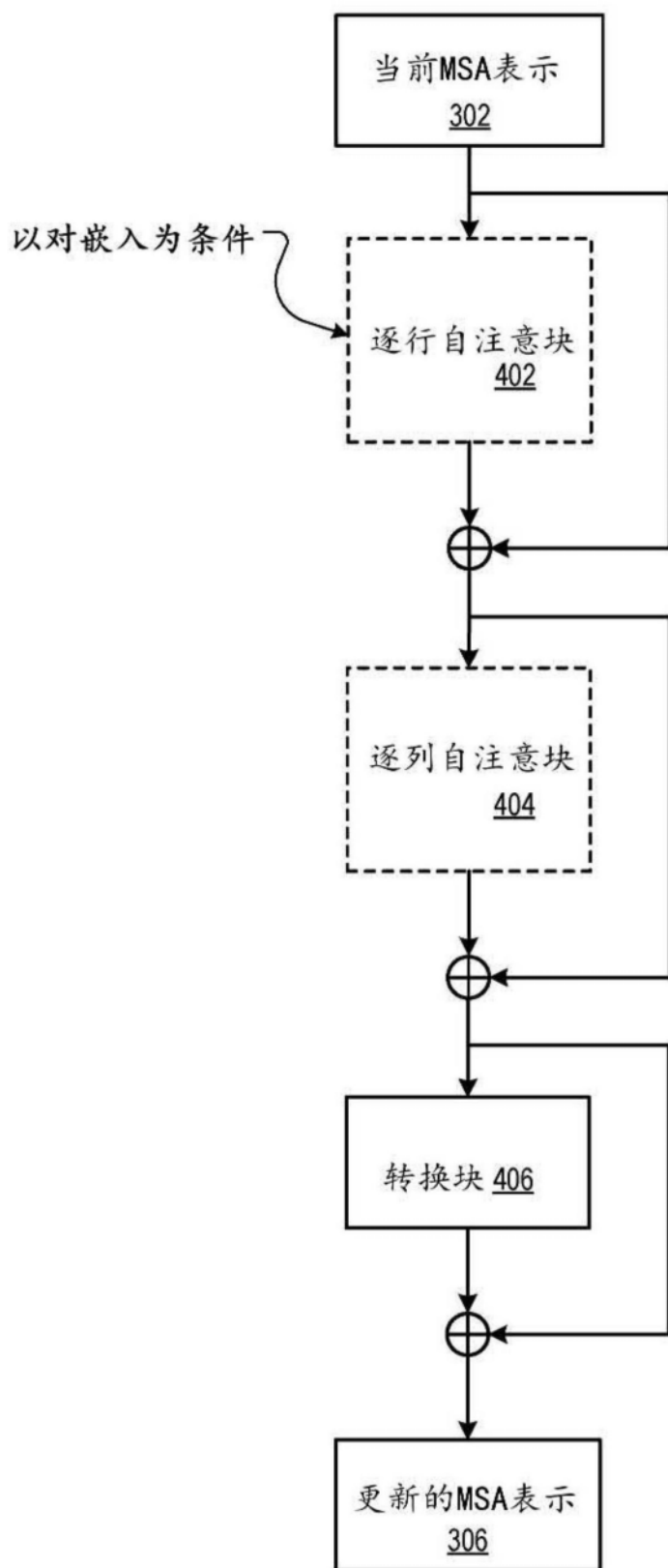
MSA更新块 400

图4

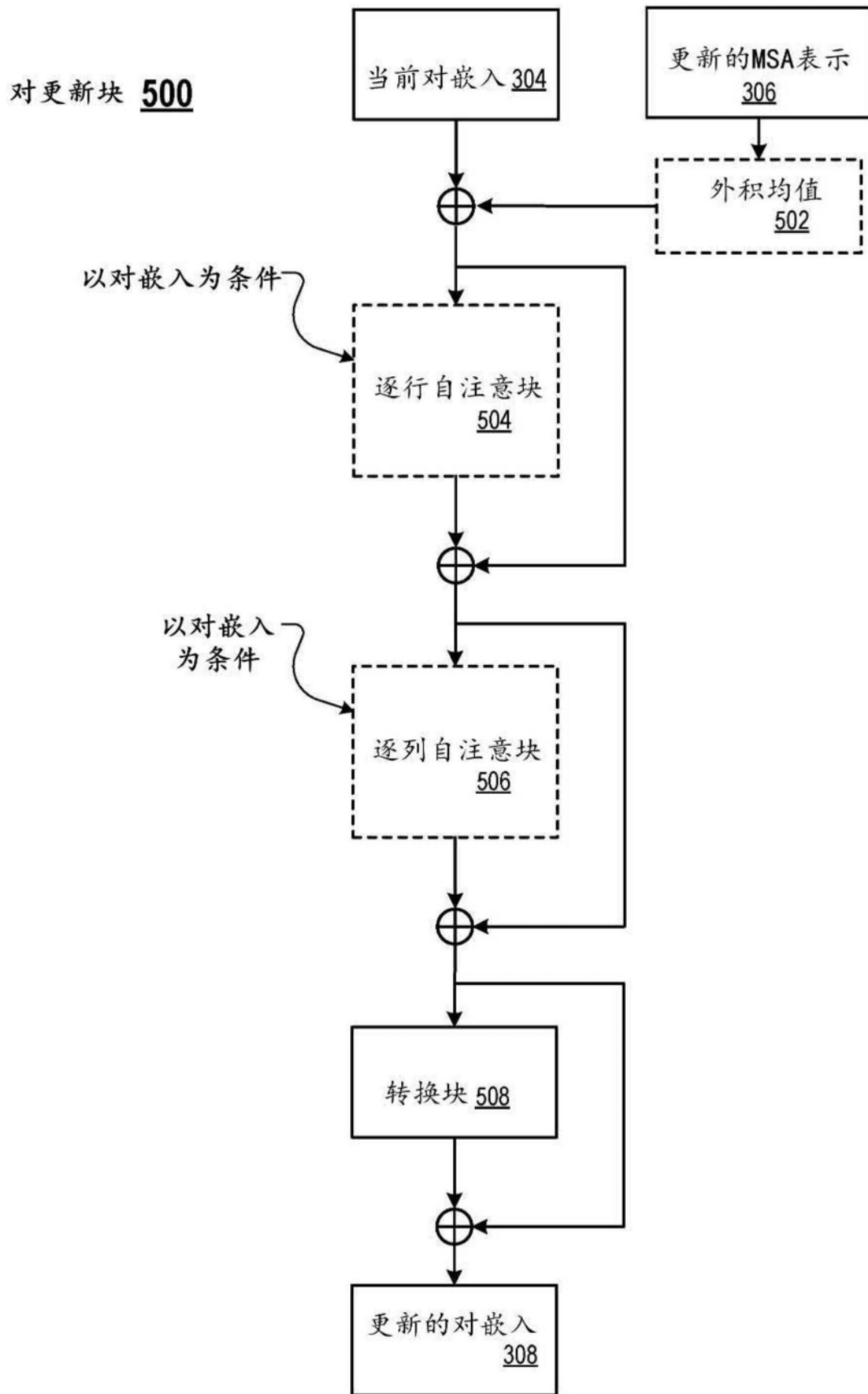


图5

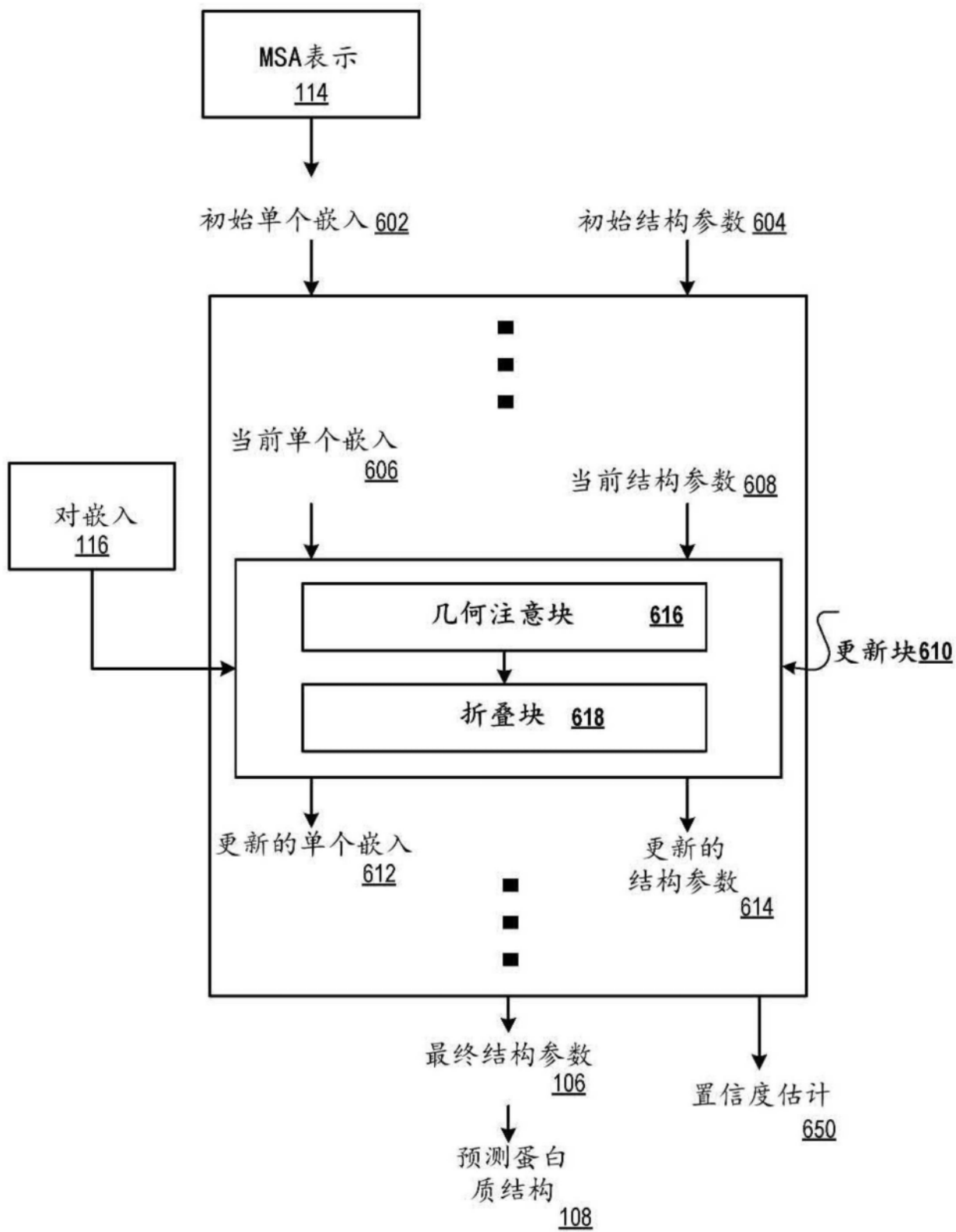
折叠神经网络 600

图6

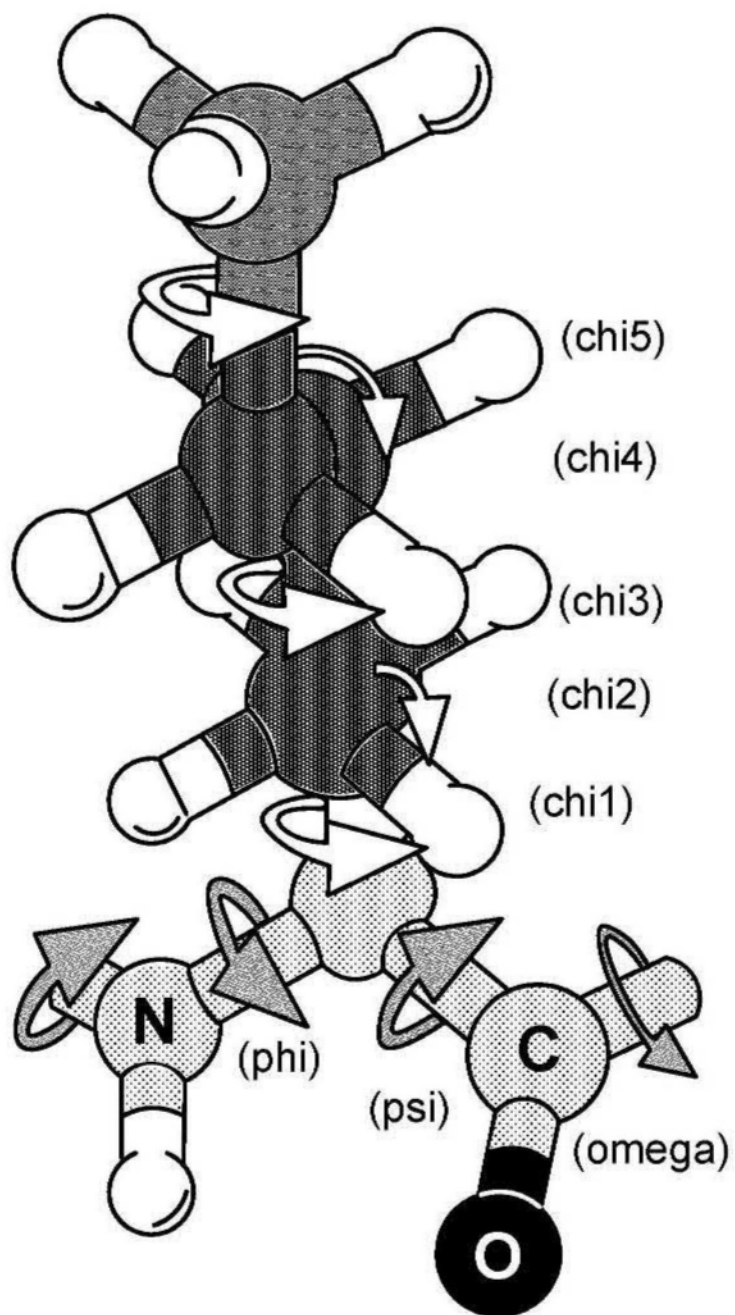


图7

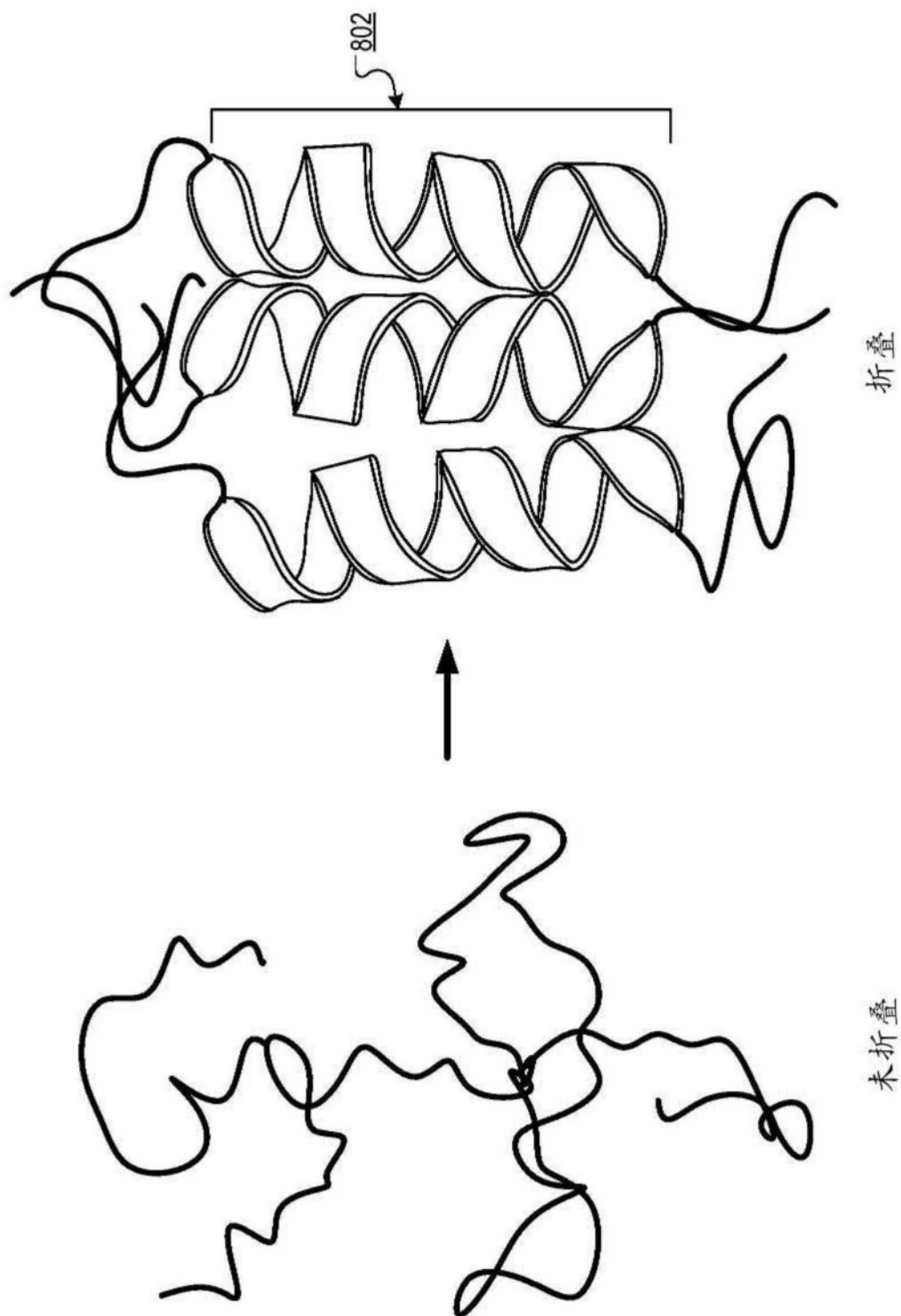


图8

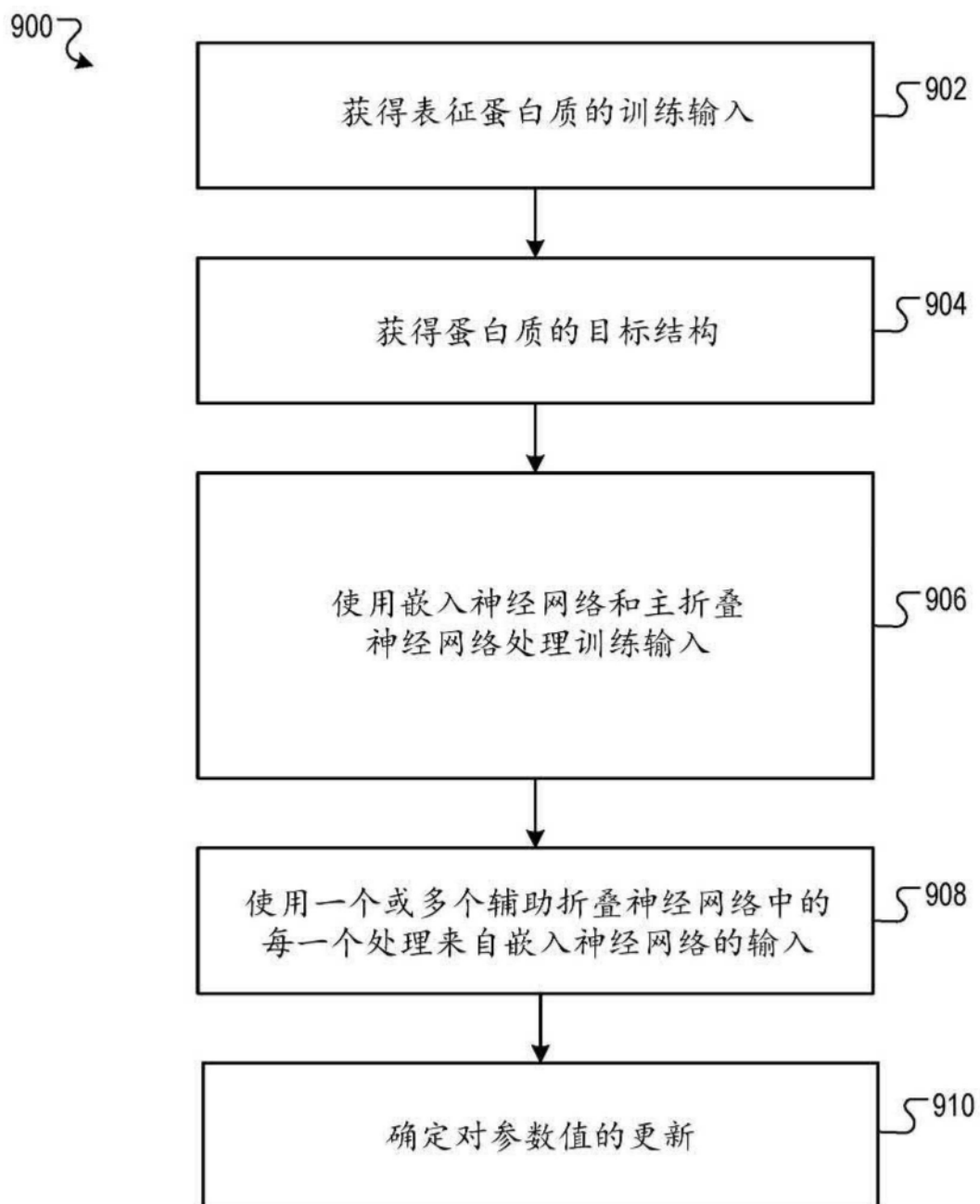


图9

生成蛋白质中的每个氨基酸链的MSA表示

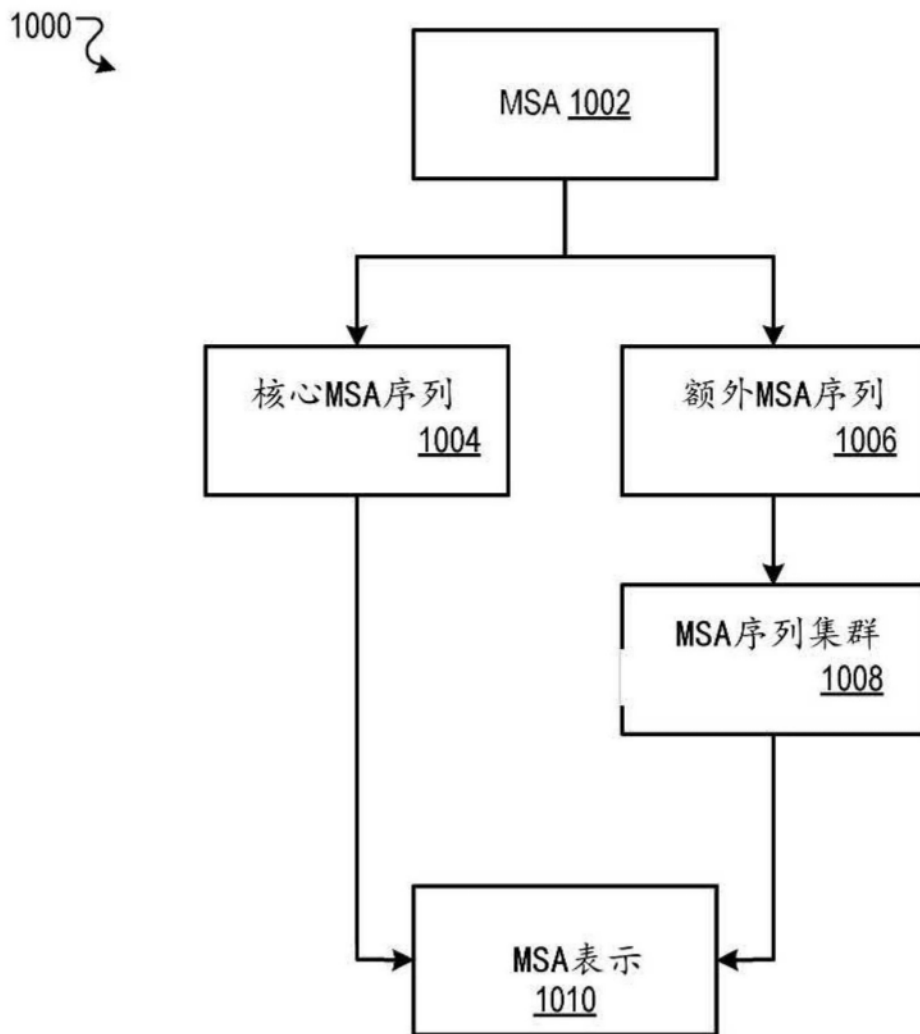


图10

生成初始对嵌入

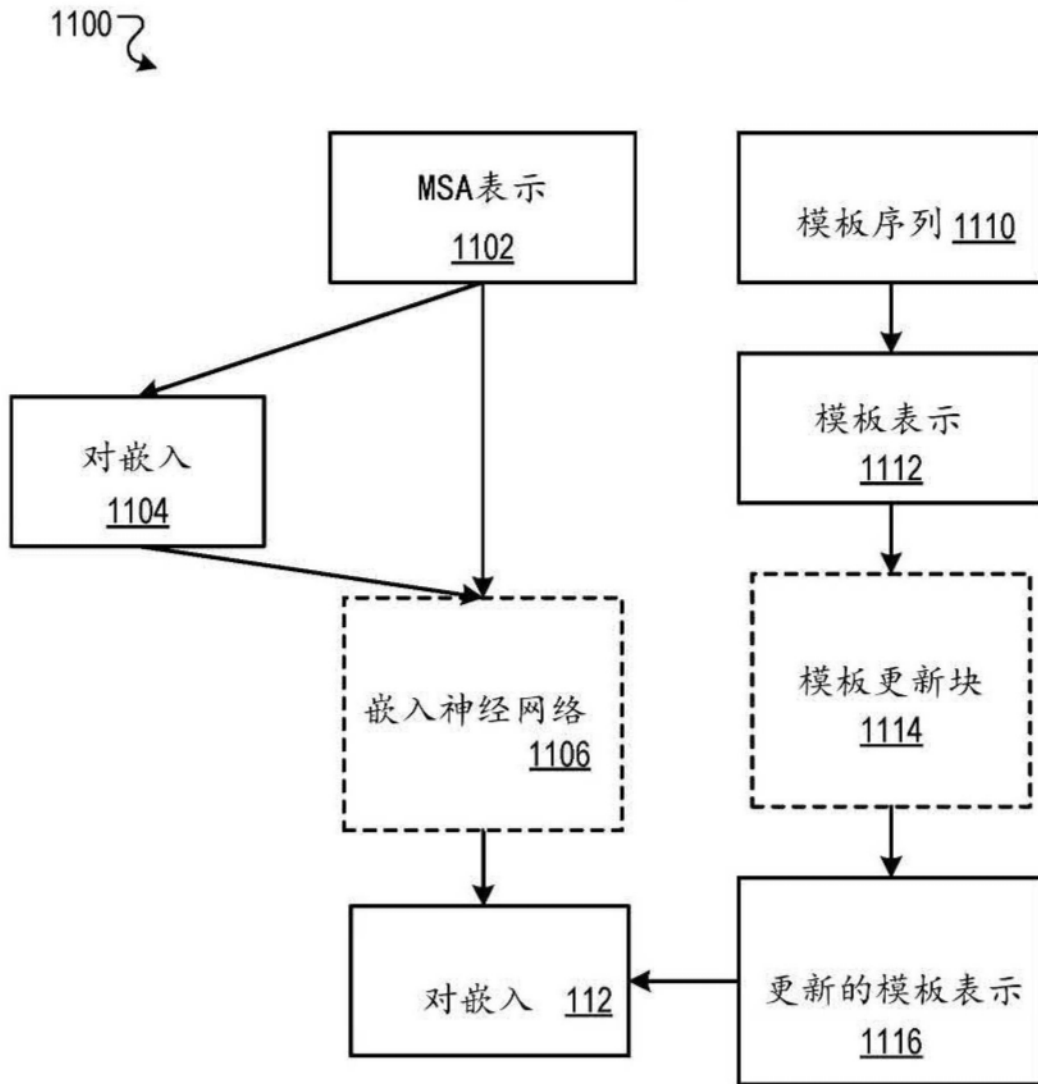


图11