(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116343910 A
(43)申请公布日 2023.06.27

(21)申请号 202310327119.9

(22)申请日 2023.03.30

(71)申请人 南京师范大学
地址 210024 江苏省南京市鼓楼区宁海路122号

(72)发明人 顾彦慧 陈晓健 张先锋 郝渊苏 文羽昕 廖楚悦

(74)专利代理机构 南京苏高专利商标事务所(普通合伙) 32204
专利代理师 柏尚春

(51)Int.Cl.
G16B 15/30(2019.01)
G16B 40/00(2019.01)
G06N 3/0475(2023.01)
G06N 3/094(2023.01)

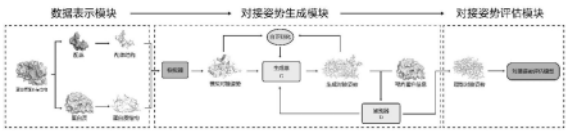
权利要求书2页 说明书5页 附图1页

(54)发明名称

基于图神经网络的蛋白质与配体之间对接姿势的预测方法

(57)摘要

本发明公开了一种基于图神经网络的蛋白质与配体之间对接姿势的预测方法，首先，获取蛋白质-配体复合物的生物信息样本集，样本集包括样本数据和样本标注数据；其次，构建基于图神经网络的对接姿势生成模型和基于多视角的对接姿势评估模型，进一步调节模型的参数，通过训练得到的结构生成模型对样本数据进行处理，获得蛋白质配体的姿势对接实际输出；最后，利用主流的姿势对接结构评价指标对输出结果进行稳定性评估。本发明直接利用配体蛋白质的生物结构信息生成最优的对接姿势结构，并通过多角度的综合评估模型对生成结果进行评估，从而提高对配体-蛋白质姿势结构对接预测的准确性，以及提高对配体-蛋白质姿势结构对接预测结果评估的有效性。



CN 116343910 A

1.一种基于图神经网络的蛋白质与配体之间对接姿势的预测方法,其特征在于,包括以下步骤:

(1)获取蛋白质-配体复合物的生物信息样本集,并对其进行预处理;样本集包括样本数据和样本数据的样本标注,对样本数据编码,得到特征向量;

(2)构建基于图神经网络的对接姿势生成模型,并使用生成对抗网络对其进行训练;固定靶向蛋白,对配体分子利用多步姿态预测模拟原子的对接姿势结构,由基于多视角的对接姿势判别器评估对接姿势,计算损失差值,调整对接姿势中原子的空间位置;反复迭代,直到判别结果满足阈值要求,对于所有原子的输出将被认为是最终的预测姿态;

(3)构建基于图神经网络的对接姿势评估模型,评估生成的理想对接姿势;对蛋白质配体的理想对接姿势实际输出,依据现有的主流评价指标,对预测得到的对接姿势结果进行评估。

2.根据权利要求1所述的基于图神经网络的蛋白质与配体之间对接姿势的预测方法,其特征在于,步骤(1)所述的对生物信息样本集进行预处理过程如下:

去除可旋转键少于两个的蛋白质-配体复合物以及具有一个以上配体的蛋白质,去除缺失残基或重复残基的蛋白质,形成包含N个蛋白质-配体复合物及其天然的结合亲和力标注;

通过根据化学键和原子排列等结构性质,将三维生物分子数据的分子图转换为二维邻接矩阵,所得的邻接矩阵符合图神经网络的输入格式;

从而得到蛋白质分子结构邻接矩阵、配体分子结构邻接矩阵、和天然的结合亲和力样本标注。

3.根据权利要求1所述的基于图神经网络的蛋白质与配体之间对接姿势的预测方法,其特征在于,步骤(2)中,一种基于生成式对抗网络的对接姿势结构生成的训练方法,具体包括如下步骤:

(21)获取蛋白质配体初始模拟对接姿势结构;利用AutoDock对接姿势结构模拟模型,固定蛋白质分子,改变配体分子的空间位置分布,进行随机对接姿势结构模拟;

(22)构建基于图神经网络的对接姿势生成模型,生成候选的对接姿势;根据模拟的对接姿势结构,初始化神经网络模型生成器;通过训练样本集中抽取的样本数据以及生成器利用定义的噪声分布,对配体分子利用多步姿态预测模拟原子的对接姿势结构,计算每个配体原子的运动并输出移动向量,获得生成器的对接姿势结构的实际输出;

(23)构建基于多视角的对接姿势有效性判别模型,对生成模型的实际输出进行评估;以原始数据集PDBbind2016中天然的结合亲和力标注确定为基准结果,将生成的配体的对接姿势结果与靶向蛋白进行多视角的亲和力预测,判断生成的对接姿势结果的有效性;将计算对接姿势的结合亲和力与基准结果的偏差,反馈给生成模型进行参数的调节优化;

(24)对生成对接姿势进行重复训练迭代,使得对接姿势在亲和力判别指标上达到理想预期,获得训练完成的模型并输出理想的对接姿势结果。

4.根据权利要求1所述的基于图神经网络的蛋白质与配体之间对接姿势的预测方法,其特征在于,所述步骤(3)实现过程如下:

利用评估函数对均方根偏差RMSD指标进行衡量,定义如下:

$$RMSD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2}$$

其中,δ表示某一帧的原子的位置减去参考系中它的位置,即位置偏移量,T表示时间,x表示原子某时刻的位置;RMSD值表示各原子运动幅度的大小,该值越大,说明该原子的运动的空间范围越大,原子的空间位阻也就越小;在评估模型中,RMSD越小,代表生成的对接姿势结构越准确有效。

5.根据权利要求3所述的基于图神经网络的蛋白质与配体之间对接姿势的预测方法,其特征在于,步骤(23)所述判断生成的对接姿势结果的有效性,具体包括如下步骤:

(231)数据特征表示;获取输入数据,即配体分子图结构,蛋白质分子图结构,整段蛋白质序列和生成的蛋白质配体对接姿势,其中蛋白质配体对接姿势利用二部图进行表示;

(232)数据编码;利用图神经网络的AttentiveFP模型、BIGGNN模型和ProBert模型分别对输入数据编码,转换成特征向量;

(233)亲和力预测;将编码得到的特征向量利用多层感知机预测亲和力结果,得到的结合亲和力值与天然的结合亲和力标注进行比较,判别生成姿势结果的有效性;

(234)计算预测结果与实际姿势的偏差;使用了四个指标来进行衡量,平均绝对误差MAE,均方根误差RMSE,均方根误差标准差SD,皮尔逊相关系数R,这四个指标的定义如下:

$$MAE=\frac{1}{|D|}\sum_{i=1}^{|D|}\left(|y_i-\hat{y}_i|\right)$$

$$RMSE=\sqrt{\frac{1}{D}\sum_{i=1}^{D}\left(|y_i-\hat{y}_i|\right)^2}$$

$$SD=\sqrt{\frac{1}{D-1}\sum_{i=1}^{D}\left[y_i-\left(a+b\hat{y}_i\right)\right]^2}$$

$$R=\frac{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)\left(y_i-\bar{y}_i\right)}{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)^2\left(y_i-\bar{y}_i\right)^2}$$

其中,D是数据集中的样本数,y和$\hat{y}$分别是实验确定的和模型预测的蛋白质-配体结合亲和力的值,a和b分别为回归线的截距和斜率;利用这四项指标计算的值确定损失函数,并反馈给对接姿势生成模型,进行参数的调节和优化。

基于图神经网络的蛋白质与配体之间对接姿势的预测方法

技术领域

[0001] 本发明属于计算机辅助药物设计领域，具体涉及一种基于图神经网络对蛋白质与配体之间对接姿势的预测方法。

背景技术

[0002] 在计算机辅助药物设计的过程中，筛选出特定蛋白质与配体之间结合亲和力高的对接姿势一直是一项难题。传统的方法考虑结合现有方法，生成无限多个对接姿势的组合，并在此基础上筛选得到最满足条件的一组姿势，忽略了药物分子与靶点蛋白质之间的全局信息，此外现有的对接姿势信息是比较少的，结合亲和力的评估也需要通过大量的实验，良好的筛选结果需要对数据集进行大量的标注。而标注量不足往往会造成对配体分子与靶点蛋白质结合亲和力预测的不准确，且此类方法中实验需要的资金成本与标注所需的人工成本是大多数科研项目难以承受的。而新兴的利用三维分子结构对对接姿势预测的方法，在预测蛋白质与配体之间的结合亲和力方面十分具有前景，提高了计算机辅助药物设计的效率与准确性。

[0003] 在现有的利用分子三维结构对对接姿势与亲和力预测的方法中，为了提高预测的效率，采用了各种抽样方法，其中例如Glide对姿势的全局信息采用了蒙特卡洛采样法，从而提高了在面对大量药物分子配体时对对接姿势预测的准确性与速度。然而类似于Glide的各种方法往往忽略了对对接姿势的生物学优化，而没有经过优化的姿势信息，会造成预测对接姿势与亲和力时的不准确，在药物生产过程中迫切需要更为智能且准确的预测模型。此外，在这类方法中往往只考虑了分子内部作用力与分子间作用力的其中一种，没有对两者进行综合讨论，在对亲和力的预测过程中，丢失了一些重要信息，造成了预测的不准确。

[0004] 总体而言，目前对蛋白质与配体之间对接姿势的预测还有很多局限性，例如：预测时耗费的时间长，现有的对接姿势信息少，对分子空间结构信息的利用不充分，对结合亲和力的预测不够准确。这主要是由于候选药物分子具有数据量大、空间结构复杂的特点。本发明针对现有蛋白质配体对接姿势预测方法的局限性，提出了一种基于图神经网络和生成对抗网络结合的预测模型，以解决配体对接姿势结构预测效率低并提高对配体蛋白质结合亲和力评估的有效性。

发明内容

[0005] 发明目的：本发明提供一种基于图神经网络对蛋白质与配体之间对接姿势的预测方法，实现了对蛋白质与配体分子之间的全局信息的充分利用，避免了在传统预测方法中丢失分子间重要信息的缺陷，并大大提高了对靶点蛋白质与配体分子之间对接姿势进行预测的准确性与效率。

[0006] 技术方案：本发明旨在一种基于图神经网络对蛋白质与配体之间对接姿势的预测方法，具体包括以下步骤：

[0007] (1)获取蛋白质-配体复合物的生物信息样本集,并对其进行预处理;样本集包括样本数据和样本数据的样本标注,对样本数据编码,得到特征向量;

[0008] (2)构建基于图神经网络的对接姿势生成模型,并使用生成对抗网络对其进行训练;固定靶向蛋白,对配体分子利用多步姿态预测模拟原子的对接姿势结构,由基于多视角的对接姿势判别器评估对接姿势,计算损失差值,调整对接姿势中原子的空间位置;反复迭代,直到判别结果满足阈值要求,对于所有原子的输出将被认为是最终的预测姿态;

[0009] (3)构建基于图神经网络的对接姿势评估模型,评估生成的理想对接姿势;对蛋白质配体的理想对接姿势实际输出,依据现有的主流评价指标,对预测得到的对接姿势结果进行评估。

[0010] 进一步地,步骤(1)所述的对生物信息样本集进行预处理过程如下:

[0011] 去除可旋转键少于两个的蛋白质-配体复合物以及具有一个以上配体的蛋白质,去除缺失残基或重复残基的蛋白质,形成包含N个蛋白质-配体复合物及其天然的结合亲和力标注;

[0012] 通过根据化学键和原子排列等结构性质,将三维生物分子数据的分子图转换为二维邻接矩阵,所得的邻接矩阵符合图神经网络的输入格式;

[0013] 从而得到蛋白质分子结构邻接矩阵、配体分子结构邻接矩阵、和天然的结合亲和力样本标注。

[0014] 进一步地,步骤(2)中,一种基于生成式对抗网络的对接姿势结构生成的训练方法,具体包括如下步骤:

[0015] (21)获取蛋白质配体初始模拟对接姿势结构;利用AutoDock对接姿势结构模拟模型,固定蛋白质分子,改变配体分子的空间位置分布,进行随机对接姿势结构模拟;

[0016] (22)构建基于图神经网络的对接姿势生成模型,生成候选的对接姿势;根据模拟的对接姿势结构,初始化神经网络模型生成器;通过训练样本集中抽取的样本数据以及生成器利用定义的噪声分布,对配体分子利用多步姿态预测模拟原子的对接姿势结构,计算每个配体原子的运动并输出移动向量,获得生成器的对接姿势结构的实际输出;

[0017] (23)构建基于多视角的对接姿势有效性判别模型,对生成模型的实际输出进行评估;以原始数据集PDBbind2016中天然的结合亲和力标注确定为基准结果,将生成的配体的对接姿势结果与靶向蛋白进行多视角的亲和力预测,判断生成的对接姿势结果的有效性;将计算对接姿势的结合亲和力与基准结果的偏差,反馈给生成模型进行参数的调节优化;

[0018] (24)对生成对接姿势进行重复训练迭代,使得对接姿势在亲和力判别指标上达到理想预期,获得训练完成的模型并输出理想的对接姿势结果。

[0019] 进一步地,所述步骤(3)实现过程如下:

[0020] 利用评估函数对均方根偏差RMSD指标进行衡量,定义如下:

[0021] $$RMSD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2}$$

[0022] 其中,δ表示某一帧的原子的位置减去参考系中它的位置,即位置偏移量,T表示时间,x表示原子某时刻的位置;RMSD值表示各原子运动幅度的大小,该值越大,说明该原子的运动的空间范围越大,原子的空间位阻也就越小;在评估模型中,RMSD越小,代表生成的对接姿势结构越准确有效。

[0023] 进一步地，步骤(23)所述判断生成的对接姿势结果的有效性，具体包括如下步骤：

[0024] (231)数据特征表示；获取输入数据，即配体分子图结构，蛋白质分子图结构，整段蛋白质序列和生成的蛋白质配体对接姿势，其中蛋白质配体对接姿势利用二部图进行表示；

[0025] (232)数据编码；利用图神经网络的Attentive FP模型、BIGGNN模型和ProBert模型分别对输入数据编码，转换成特征向量；

[0026] (233)亲和力预测；将编码得到的特征向量利用多层感知机预测亲和力结果，得到的结合亲和力值与天然的结合亲和力标注进行比较，判别生成姿势结果的有效性；

[0027] (234)计算预测结果与实际姿势的偏差；使用了四个指标来进行衡量，平均绝对误差MAE，均方根误差RMSE，均方根误差标准差SD，皮尔逊相关系数R，这四个指标的定义如下：

[0028] $$MAE=\frac{1}{|D|}\sum_{i=1}^{|D|}\left(\left|y_i-\hat{y}_i\right|\right)$$

[0029] $$RMSE=\sqrt{\frac{1}{D}\sum_{i=1}^{D}\left(\left|y_i-\hat{y}_i\right|\right)^2}$$

[0030] $$SD=\sqrt{\frac{1}{D-1}\sum_{i=1}^{D}\left[y_i-\left(a+b\hat{y}_i\right)\right]^2}$$

[0031] $$R=\frac{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)\left(y_i-\bar{y}_i\right)}{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)^2\left(y_i-\bar{y}_i\right)^2}$$

[0032] 其中，D是数据集中的样本数，y和$\hat{y}$分别是实验确定的和模型预测的蛋白质-配体结合亲和力的值，a和b分别为回归线的截距和斜率；利用这四项指标计算的值确定损失函数，并反馈给对接姿势生成模型，进行参数的调节和优化。

[0033] 有益效果：与现有技术相比，本发明的有益效果：本发明采用图神经网络，实现了对蛋白质与配体分子之间的全局信息的充分利用，避免了在传统预测方法中丢失分子间重要信息的缺陷，并大大提高了对靶点蛋白质与配体分子之间对接姿势进行预测的准确性与效率；本发明采用基于生成对抗网络的对接姿势结构生成模型和基于图神经网络的对接姿势评估模型两种优化模型，充分利用生物分子的全局信息，根据所得信息生成对蛋白质-配体对接姿势的模型并进行预测评估。

附图说明

[0034] 图1为本发明的流程图；

[0035] 图2为本发明蛋白质-配体复合物生物信息数据预处理流程示意图；

[0036] 图3为本发明蛋白质-配体生成姿势结构评估方法流程示意图。

具体实施方式

[0037] 下面结合附图对本发明作进一步详细说明。

[0038] 如图1所示，本发明提出一种基于图神经网络的蛋白质与配体之间对接姿势的预测方法，具体包括如下步骤：

[0039] 步骤1：获取蛋白质-配体复合物的生物信息样本集，并对其进行预处理；样本集包括样本数据和样本数据的样本标注，对样本数据编码，得到特征向量。

[0040] 提取原始PDBbind2016数据集中的数据，以获得其中的配体与蛋白质的生物分子结构信息，去除可旋转键少于两个的蛋白质-配体复合物以及具有一个以上配体的蛋白质，去除缺失残基或重复残基的蛋白质，形成包含N个蛋白质-配体复合物及其天然的结合亲和力标注。

[0041] 然后根据化学键和原子排列等结构性质，将三维生物分子数据转换为二维邻接矩阵(该邻接矩阵符合Pytorch_Geometric图神经网络框架的输入格式)。再将得到的蛋白质-配体组合多维信息用于对接姿势结构生成，以随机获取大量初始对接姿势结构，目前对接姿势结构生成已有成熟的模型，我们采用的是AutoDock对接姿势结构生成模型。最后的训练样本集与测试样本集，就是根据对蛋白质-配体对接姿式结构进行的亲和力预测与标注，进而随机划分得到的。其中，每组样本数据包括蛋白质分子结构邻接矩阵、配体分子结构邻接矩阵、蛋白质配体对接姿式结构和相应的结合亲和力样本标注。

[0042] 步骤2：构建基于图神经网络的对接姿势生成模型，并使用生成对抗网络对其进行训练；固定靶向蛋白，对配体分子利用多步姿态预测模拟原子的对接姿势结构，由基于多视角的对接姿势判别器评估对接姿势，计算损失差值，调整对接姿势中原子的空间位置；反复迭代，直到判别结果满足阈值要求，对于所有原子的输出将被认为是最终的预测姿态。如图2所示，具体包括以下步骤：

[0043] (2.1)获取蛋白质配体初始模拟对接姿势结构；利用对接姿势结构生成模型AutoDock，固定蛋白质分子，改变配体分子的空间位置分布，进行随机对接姿势结构生成；

[0044] (2.2)构建基于图神经网络的对接姿势生成模型，生成候选的对接姿势；初始化神经网络模型生成器，通过训练样本集中抽取的样本数据以及生成器利用定义的噪声分布，对配体分子利用多步姿态预测模拟原子的对接姿势结构，计算每个配体原子的运动并输出移动向量，获得生成器的对接姿势结构的实际输出；

[0045] (2.3)构建基于多视角的对接姿势有效性判别模型，对生成模型的实际输出进行评估；以原始数据集PDBbind2016中天然的结合亲和力标注确定为基准结果，将生成的配体的对接姿势结果与靶向蛋白进行多视角的亲和力预测，判断生成的对接姿势结果的有效性；将计算对接姿势的结合亲和力与基准结果的偏差，反馈给生成模型进行参数的调节优化。

[0046] 判断生成的对接姿势结果的有效性，如图3所示，具体过程如下：

[0047] 1)数据特征表示；获取输入数据，即配体分子图结构，蛋白质分子图结构，整段蛋白质序列和生成的蛋白质配体对接姿势，其中蛋白质配体对接姿势利用二部图进行表示。

[0048] 2)数据编码；利用图神经网络的Attentive FP模型、BIGGNN模型和ProBert模型分别对输入数据编码，转换成特征向量。

[0049] 3)亲和力预测；将编码得到的特征向量利用多层感知机预测亲和力结果，得到的结合亲和力值与天然的结合亲和力标注进行比较，判别生成姿势结果的有效性。

[0050] 4)计算预测结果与实际姿势的偏差；使用了四个指标来进行衡量，平均绝对误差

MAE,均方根误差RMSE,均方根误差标准差SD,皮尔逊相关系数R,这四个指标的定义如下:

[0051] $$MAE=\frac{1}{|D|}\sum_{i=1}^{|D|}\left(|y_i-\hat{y}_i|\right)$$

[0052] $$RMSE=\sqrt{\frac{1}{D}\sum_{i=1}^{D}\left(|y_i-\hat{y}_i|\right)^2}$$

[0053] $$SD=\sqrt{\frac{1}{D-1}\sum_{i=1}^{D}\left[y_i-\left(a+b\hat{y}_i\right)\right]^2}$$

[0054] $$R=\frac{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)\left(y_i-\bar{y}_i\right)}{\sum_{i=1}^{D}\left(\hat{y}_i-\bar{\hat{y}}_i\right)^2\left(y_i-\bar{y}_i\right)^2}$$

[0055] 其中,D是数据集中的样本数,y和$\hat{y}$分别是实验确定的和模型预测的蛋白质-配体结合亲和力的值,a和b分别为回归线的截距和斜率。利用这四项指标计算的值确定损失函数,并反馈给对接姿势生成模型,进行参数的调节和优化。

[0056] (2.4)对生成对接姿势进行重复训练迭代,使得对接姿势在亲和力判别指标上达到理想预期,获得训练完成的模型并输出理想的对接姿势结果。

[0057] 步骤3:构建基于图神经网络的对接姿势评估模型,评估生成的理想对接姿势;对蛋白质配体的理想对接姿势实际输出,依据现有的主流评价指标,对预测得到的对接姿势结果进行评估。

[0058] 利用传统的评估函数对均方根偏差RMSD指标进行衡量,其定义如下:

[0059] $$RMSD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\delta_i\right)^2}$$

[0060] 其中,δ表示某一帧的原子的位置减去参考系中它的位置,即位置偏移量,T表示时间,x表示原子某时刻的位置;RMSD值表示各原子运动幅度的大小,该值越大,说明该原子的运动的空间范围越大,原子的空间位阻也就越小。在评估模型中,RMSD越小,代表生成的对接姿势结构越准确有效。

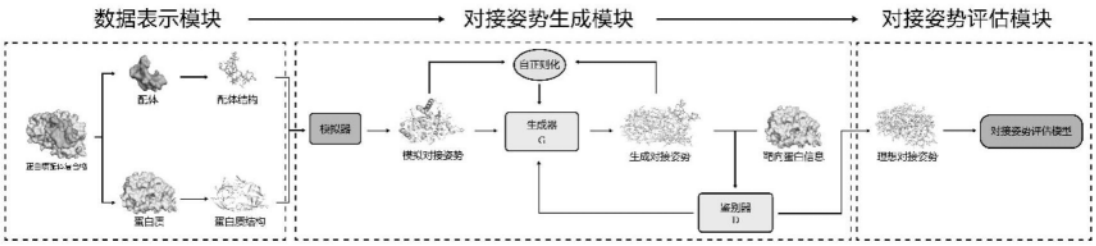

数据表示模块
对接姿势生成模块
对接姿势评估模块
配体
配体结构
蛋白质
蛋白质结构
模拟器
模拟对接姿势
自环到化
生成器
G
生成对接姿势
鉴别器
D
对接姿势评估模型

图1

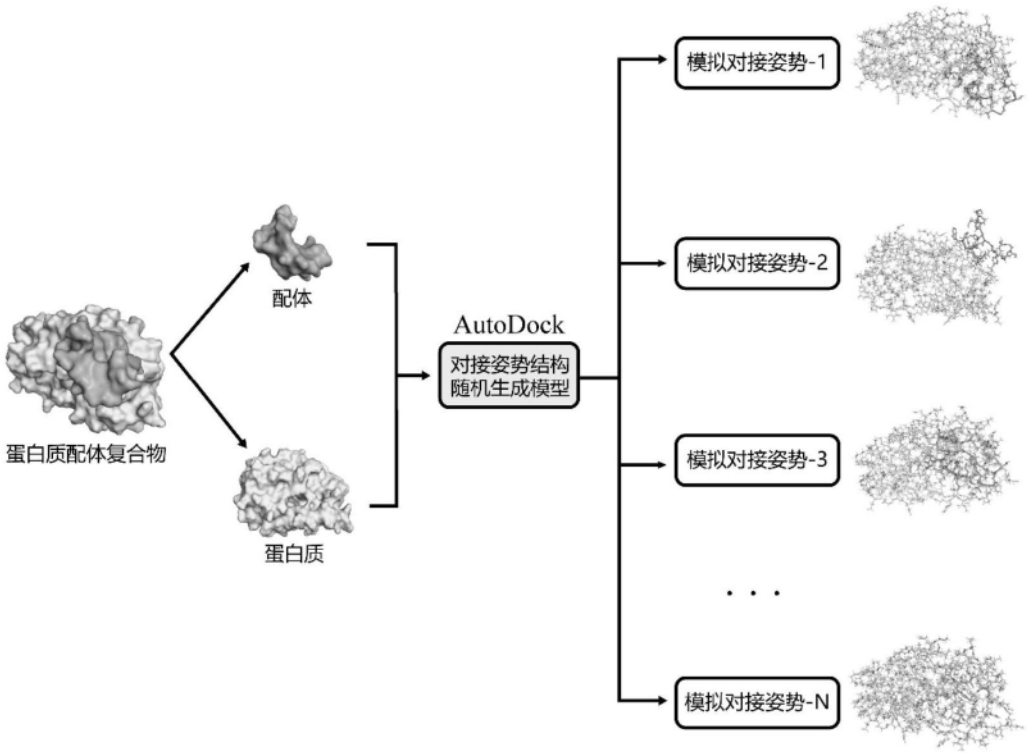

蛋白质配体复合物
配体
蛋白质
AutoDock
对接姿势结构
随机生成模型
模拟对接姿势-1
模拟对接姿势-2
模拟对接姿势-3
模拟对接姿势-N

图2

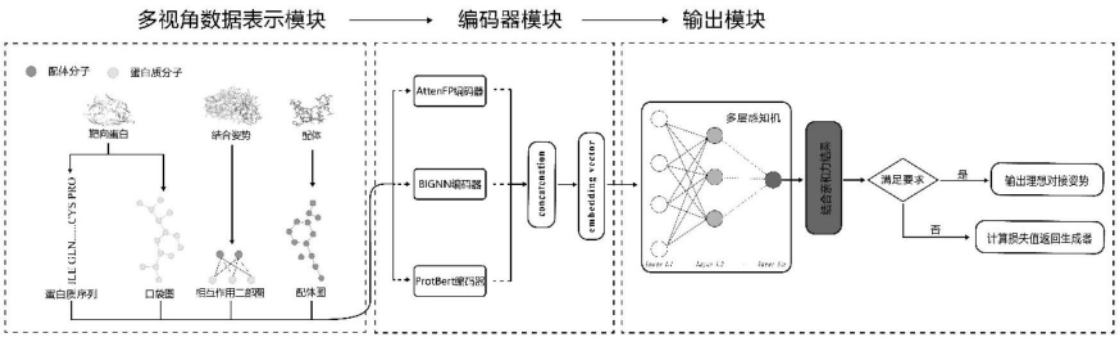

多视角数据表示模块
编码器模块
输出模块
配体分子
蛋白质分子
靶向蛋白
结合姿势
配体
蛋白质序列
口袋图
相互作用二部图
配体图
AttenFP编码器
BIGNN编码器
ProtBert编码器
concatnation
embedding vector
多层感知机
满足要求
是
否
输出理想对接姿势
计算损失值返回生成器

图3