(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2025 年 2 月 20 日 (20.02.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/036438 A1

(51) 国际专利分类号:
G16B 20/50 (2019.01)

(21) 国际申请号: PCT/CN2024/112259

(22) 国际申请日: 2024 年 8 月 15 日 (15.08.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202311028534.0 2023年8月15日 (15.08.2023) CN

(71) 申请人: 上海金斯康生物科技有限公司(**GENSCRIPT (SHANGHAI) BIOTECH CO., LTD.**) [CN/CN]; 中国上海市浦东新区中国(上海)自由贸易试验区荷丹路186号1幢2层, Shanghai 200131 (CN)。南京金斯瑞生物科技有限公司(**NANJING GENSCRIPT BIOTECH CO., LTD.**) [CN/CN]; 中国江苏省南京市江宁科学园雍熙路28号, Jiangsu 211100 (CN)。

(72) 发明人:李根(**LI, Gen**); 中国上海市浦东新区中国(上海)自由贸易试验区荷丹路186号1幢2层, Shanghai 200131 (CN)。樊隆(**FAN, Long**); 中国上海市浦东新区中国(上海)自由贸易试验区荷丹路186号1幢2层, Shanghai 200131 (CN)。张宁(**ZHANG, Ning**); 中国江苏省南京市江宁科学园雍熙路28号, Jiangsu 211100 (CN)。

(74) 代理人:北京华睿卓成知识产权代理事务所(普通合伙)(**CHENG & PENG INTELLECTUAL PROPERTY LAW OFFICE**); 中国北京市东城区东长安街1号东方广场东方经贸城东一办公楼12层12室, Beijing 100738 (CN)。

(54) **Title:** PREDICTION FOR INFLUENCE OF MUTATION ON PROTEASE ACTIVITY

(54) 发明名称: 预测突变对蛋白酶活性的影响

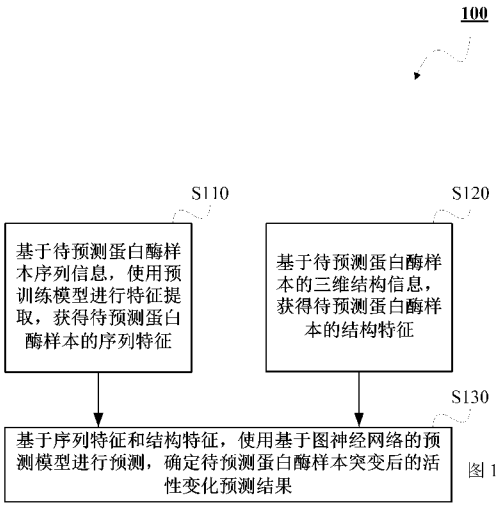


图 1

S110 On the basis of sequence information of a protease sample to be subjected to prediction, perform feature extraction by using a pre-training model, so as to obtain a sequence feature of said protease sample

S120 On the basis of three-dimensional structure information of said protease sample, obtain a structural feature of said protease sample

S130 On the basis of the sequence feature and the structural feature, perform prediction by using a prediction model based on a graph neural network, and determine an activity change prediction result after said protease sample mutates

(57) **Abstract:** The present disclosure relates to a prediction for the influence of a mutation on protease activity. A method for predicting the influence of a mutation on protease activity comprises: on the basis of sequence information of a protease sample to be subjected to prediction, performing feature extraction by using a pre-training model, so as to obtain a sequence feature of said protease sample (S110); on the basis of three-dimensional structure information of said protease sample, obtaining a structural feature of said protease sample (S120); and on the basis of the sequence feature and the structural feature, performing prediction by using a prediction model

WO 2025/036438 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

based on a graph neural network, and determining an activity change prediction result after said protease sample mutates (S130). Further disclosed is a method for constructing a prediction model used for predicting the influence of a mutation on protease activity. By means of the method in the present disclosure, prediction is performed on the basis of an advanced natural language processing model and a graph neural network, in combination with a sequence and a three-dimensional structure, and by means of an artificial intelligence model, showing a performance superior to that of other existing methods, and proving the wide application prospects of the method in protease engineering.

(57) 摘要: 本公开涉及预测突变对蛋白酶活性的影响。在一种用于预测突变对蛋白酶活性影响的方法中，基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得待预测蛋白酶样本的序列特征(S110)；基于待预测蛋白酶样本的三维结构信息，获得待预测蛋白酶样本的结构特征(S120)；基于序列特征和结构特征，使用基于图神经网络的预测模型进行预测，确定待预测蛋白酶样本突变后的活性变化预测结果(S130)。还公开了一种用于预测突变对蛋白酶活性影响的预测模型的构建方法。本公开的方法基于先进的自然语言处理模型和图神经网络，结合序列和三维结构，通过人工智能模型进行预测，表现出优于其他现有方法的性能，证明了其在蛋白酶工程中的广阔的应用前景。

预测突变对蛋白酶活性的影响

相关申请的交叉引用

本申请要求于2023年8月15日提交中国专利局的申请号为202311028534.0、名称为“预测突变对蛋白酶活性的影响”的中国专利申请的优先权，并将其全部内容通过引用结合在本申请中。

技术领域

本公开涉及蛋白酶优化，更具体涉及预测蛋白酶上的氨基酸突变对蛋白酶活性的影响。

背景技术

酶是一种重要的生物催化剂，由于具有高选择性、生物相容性和反应的温和性，在工业生物催化尤其是医药中间体的生产中有广泛的应用前景。绝大多数的酶是蛋白质，而蛋白质的分子结构和功能会受到温度、pH、激活剂等因素影响，因此蛋白酶的催化活性也只有在一定条件下才能表现出来，通常需要对蛋白酶的催化活性进行工程改造才能适应不同的生产环境和功能需求。

定向进化和理性设计是两种常见的蛋白酶优化策略。定向进化策略已经被成功应用于活性、稳定性、底物特异性、立体选择性等蛋白酶性质的改造，美国科学家弗朗西丝·阿诺德（Frances H. Arnold）因此拿下2018年诺贝尔化学奖。然而，定向进化需要构建大规模的突变体文库，建立高通量筛选手段，会耗费大量的人力，物力以及财力。定向进化方法也难以完成对序列空间的全面搜索，存在根本缺陷。理性设计策略依靠对蛋白酶结构与功能关系的认识预测可能的突变型，而后通过定点突变的手段在目的基因中构建突变型。从本质上说，与定向进化相比，理性设计改造蛋白酶的效率更高。另外，理性设计方法具有普适性，一种有效的理性设计策略可以普遍应用于多种蛋白酶的改造。然而，目前理性设计方法的准确率普遍较低，应用范围远没有定向进化广泛。近年来，人工智能辅助的蛋白质工程逐渐发展成为一种高效的蛋白设计新策略，在蛋白质的结构

预测、功能预测、稳定性预测和抗体亲和力预测等多个方面显现出独特的优势，成为继理性设计和定向进化之后的又一次技术浪潮。

近年来，随着计算机技术的飞速发展，各种神经网络模型正在向更准确、更高效的方向快速更新迭代，人工智能（AI）辅助的酶工程逐渐发展成为一种高效的蛋白酶设计新策略。为了应对大规模基因测序带来的蛋白质序列数据库的爆炸式增长，目前最大的蛋白质语言模型的参数量已经达到了 150 亿（参见 Lin, Zeming 等. 2023. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. *Science* 379(6637): 1123 - 30。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）。不断发展的高性能计算和自然语言处理（NLP）的进步使研究人员能够使用大型蛋白质数据库来增强对相对较小的实验数据集的序列属性或注释的预测。

经过数十年的发展，蛋白质的稳定性预测已经有数十种基于 AI 的方法，并且性能已经超越了传统的理性设计方法。与蛋白酶的稳定性预测不同，目前仅有 SCANEER（Sequence Co-Evolutionary Analysis To Control Efficiency Of Enzyme Reactions）方法基于序列协同进化分析寻找进化上存在的可替代氨基酸用以提高蛋白酶活性（参见 Kim, Donghyo 等. 2022. “Enzyme activity engineering based on sequence co-evolution analysis”. *Metabolic Engineering* 74: 49 - 60。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）。SCANEER 评估了多序列比对（MSA）中氨基酸对的共同进化关系，在 MSA 中未观察到或很少观察到的氨基酸对即氨基酸替换则不能预测，最终导致每个蛋白酶可预测的突变数目仅占所有可能突变的 47%，这将极大地限制了该方法的应用范围。如何将现在先进的 AI 方法应用到蛋白酶活性的预测方法还存在空白。

目前将自然语言处理应用到蛋白酶活性预测遇到诸多问题。首先由于人工智能算法严重依赖数据，初始数据的数量和质量决定了训练得到的模型的泛化性能（参见 Usmanova, Dinara R 等. 2018. “Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation”. *Bioinformatics* 34(21): 3653 - 58。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）。数据集的样本数量不足或者质量过低会导致模型出现过拟合或者欠拟合的问题，这可能是限制基于 AI 预测蛋白酶活性的其中一个重要因素。同时由于蛋白酶进化和实验的偏好性导致数据集不平衡，进而影响预测模型的性能。其次，针对不同的底物在不同的温度、

pH 等实验条件下，同一个突变对蛋白酶会产生不同的活性影响，这种多标签数据也给预测工作带来了极大的困难。

发明内容

稳定性和催化活性是蛋白酶的两个最重要的属性，其中稳定性研究较多，已经发展出数十种基于 AI 的方法预测突变对蛋白酶稳定性的影响方法并且性能优秀，而催化活性一直受限于数据集和方法的问题至今停滞不前，仅有 SCANEER 方法（如上已引用）能够预测单点突变对蛋白酶活性的影响，但是该方法由于基于协同进化信息仅能够预测进化中出现过的突变，限制较多，无法广泛应用。基于以上提到的现阶段预测突变对蛋白酶活性影响方法存在的缺陷，本公开提出了一种用于预测突变对蛋白酶活性的影响的方法。它是一种基于新的自然语言处理方法和图结构数据的深度学习方法，采用 T5（即，Transfer Text-to-Text Transformer）预训练模型提取序列特征、图神经网络模型（如图注意力网络，Graph Attention Network（GAT））提取结构特征，成功地将蛋白质语言模型和图注意力网络模型用于预测由单点错义突变引起的蛋白酶活性变化。

根据本公开的第一方面，提供了一种用于预测突变对蛋白酶活性影响的方法。所述方法可以包括如下步骤：基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；基于待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

在根据本公开第一方面的方法中，所述待预测蛋白酶样本序列信息可以包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息。

另一方面，所述待预测蛋白酶样本序列信息可以包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息。由此，根据所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息，可以获得所述待预测蛋白酶样本在突变后的氨基酸序列信息。

优选地，所述指定突变信息包括指定突变位点和/或指定突变后氨基酸类型。

在根据本公开第一方面的方法中，所述的基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征可以进一步包括：将所

述待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息分别使用所述预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息；将所述突变前的序列特征信息和所述突变后的序列特征信息进行拼接，得到所述待预测蛋白酶样本的序列特征。

在根据本公开第一方面的方法中，所述预训练模型可以采用以下模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL；较优选地，所述预训练模型为 ProtT5-XL 或 ProtT5-XXL；更优选地，所述预训练模型为 ProtT5-XL。

在根据本公开第一方面的方法中，所述待预测蛋白酶样本的三维结构信息可以包括通过以下数据库或预测软件中的至少一种获取的待预测蛋白酶样本的三维结构信息：PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold 、Modeller 和 Swiss-model。

在根据本公开第一方面的方法中，所述待预测蛋白酶样本的三维结构信息可以包括相互作用网络、二级结构、氨基酸残基距离或物理环境。优选地，所述待预测蛋白酶样本的三维结构信息包括相互作用网络。

优选地，所述待预测蛋白酶样本的三维结构信息通过以下方式获取：获取待预测蛋白酶样本的三维结构的坐标；根据获取的三维结构的坐标计算蛋白酶三维结构中氨基酸残基对之间的距离；当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个待预测蛋白酶样本的邻接矩阵。

相应地，在根据本公开第一方面的方法中，所述的基于待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征可以包括：以突变位点为中心，从所述整个待预测蛋白酶样本的邻接矩阵中截取指定大小的邻接矩阵，作为所述待预测蛋白酶样本的结构特征。在一些实施方式中，以突变位点为中心，从整个待预测蛋白酶样本的邻接矩阵中截取 N*N（N 为大于或等于 1 的整数）的邻接矩阵。从整个待预测蛋白酶样本的邻接矩阵中截取 N*N 的邻接矩阵可以是基于突变位点左右两侧的序列进行截取，也可以是基于与突变位点之间的距离进行截取。在一个实施方式中，以突变位点为中心，选择整个待预测蛋白酶样本的邻接矩阵中与突变位点距离小于 10 埃的氨基酸，作为待预测蛋白酶样本的结构特征。

在根据本公开第一方面的方法中，优选地，所述的基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果可以进一步包括：将将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，最终得到关于活性变化的分类。

优选地，所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征可以包括：将所述序列特征和所述结构特征输入到图神经网络；通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；将若干层图网络结构分别输出的特征合并；采用池化操作对特征进行压缩；根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。优选地，这里的池化操作可以是平均池化、最大池化或 K-max 池化。这里的活性变化的概率特征可以是蛋白酶样本突变后活性升高的概率值。

在根据本公开第一方面的方法中，所述的基于图神经网络的预测模型可以通过以下步骤构建出来的：获取训练集，所述训练集包括多个训练样本的信息，每个所述训练样本的信息包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息以及突变活性变化标签；基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征；基于样本蛋白酶的三维结构信息，获得所述多个训练样本的结构特征；基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得所述的基于图神经网络的预测模型。

优选地，所述基于图神经网络的预测模型是分类器。

优选地，所述图神经网络是图卷积网络或图注意力网络。

优选地，所述的基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得所述的基于图神经网络的预测模型可以进一步包括：将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。

优选地，所述决策树模型是梯度提升决策树（GBDT）模型。更优选地，使用 LightGBM 框架来实现 GBDT 模型。

优选地，可以获取公开数据库中的实验数据作为训练集。与此同时，可以通过对公开数据库中的实验数据做反转处理将训练集变为原来的两倍大小。

根据本公开的第二方面，提供一种用于预测突变对蛋白酶活性影响的预测模型的构建方法。所述方法可以包括如下步骤：获取训练集，所述训练集包括多个训练样本的信息，每个所述训练样本的信息包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息以及突变活性变化标签；基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征；基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征；基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型。

在根据本公开第二方面的方法中，所述的基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征可以包括：将所述样本蛋白酶突变前和突变后的氨基酸序列信息分别使用所述预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息；将所述突变前的序列特征信息和所述突变后的序列特征信息进行拼接，得到所述多个训练样本的序列特征。

在根据本公开第二方面的方法中，所述预训练模型可以采用以下模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL；较优选地，所述预训练模型为 ProtT5-XL 或 ProtT5-XXL；更优选地，所述预训练模型为 ProtT5-XL。

在根据本公开第二方面的方法中，所述的样本蛋白酶的三维结构信息可以包括通过以下数据库或预测软件中的至少一种获取的样本蛋白酶的三维结构信息：PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold 、Modeller 和 Swiss-model。

在根据本公开第二方面的方法中，所述样本蛋白酶的三维结构信息可以包括相互作用网络、二级结构、氨基酸残基距离或物理环境。优选地，所述样本蛋白酶的三维结构信息包括相互作用网络。

优选地，所述样本蛋白酶的三维结构信息可以通过以下方式获取：获取样本蛋白酶的三维结构的坐标；根据获取的三维结构的坐标计算样本蛋白酶三维结构中氨基酸残基

对之间的距离；当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个样本蛋白酶的邻接矩阵。

相应地，在根据本公开第二方面的方法中，所述的基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征可以包括：以突变位点为中心，从所述整个样本蛋白酶的邻接矩阵中截取指定大小的邻接矩阵，作为所述样本蛋白酶的结构特征。在一些实施方式中，以突变位点为中心，从整个样本蛋白酶的邻接矩阵中截取 N*N（N 为大于或等于 1 的整数）的邻接矩阵。从整个样本蛋白酶的邻接矩阵中截取 N*N 的邻接矩阵可以是基于突变位点左右两侧的序列进行截取，也可以是基于与突变位点之间的距离进行截取。在一个实施方式中，以突变位点为中心，选择整个样本蛋白酶的邻接矩阵中与突变位点距离小于 10 埃的氨基酸，作为样本蛋白酶的结构特征。

在根据本公开第二方面的方法中，所述的基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型可以进一步包括：将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。

优选地，所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征可以包括：将所述序列特征和所述结构特征输入到图神经网络；通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；将若干层图网络结构分别输出的特征合并；采用池化操作对特征进行压缩；根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。优选地，池化操作可以是平均池化、最大池化或 K-max 池化。

优选地，所述预测模型是分类器。

优选地，所述图神经网络是图卷积网络或图注意力网络。

优选地，所述决策树模型是 GBDT 模型。更优选地，使用 LightGBM 框架来实现 GBDT 模型。

优选地，可以获取公开数据库中的实验数据作为训练集。与此同时，可以通过对公开数据库中的实验数据做反转处理将训练集变为原来的两倍大小。

根据本公开的第三方面，提供一种用于预测突变对蛋白酶活性影响的系统。所述系统可以包括：获取模块，用于获取待预测蛋白酶样本序列信息和三维结构信息；处理模块，用于通过输入待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果。所述处理模块进一步包括如下子模块：序列特征子模块，用于基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；结构特征子模块，用于基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；模型预测子模块，用于基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

根据本公开第三方面的用于预测突变对蛋白酶活性影响的系统可以通过计算机实现，优选地，可以通过执行计算机程序以实现以下操作：获取待预测蛋白酶样本序列信息和三维结构信息；通过输入待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果，该操作进一步包括如下子操作：基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

根据本公开的第四方面，提供一种非瞬时性计算机可读存储介质，用于存储计算机程序。所述计算机程序包括指令。所述指令在由电子设备的处理器执行时使所述电子设备实施根据本公开的第一方面的用于预测突变对蛋白酶活性影响的方法或根据本公开的第二方面的用于预测突变对蛋白酶活性影响的预测模型的构建方法。

根据本公开的第五方面，提供一种计算机系统。所述计算机系统包括：处理器、存储器和计算机程序。所述计算机程序存储在所述存储器中并且被配置为由所述处理器执行。所述计算机程序包括用于实施根据本公开的第一方面的用于预测突变对蛋白酶活性影响的方法或根据本公开的第二方面的用于预测突变对蛋白酶活性影响的预测模型的构建方法的指令。

以上方法的特点是通过图神经网络，特别是图注意力网络（Graph Attention Network，GAT），将蛋白酶的一级序列和三维结构建立连接进而预测突变效应，因此

有可能为应对不断增加的序列和结构数据带来的注释挑战提供新的见解。对于实验数据存在质量低和偏好性的问题，本公开从 D3DistalMutation 数据库中提取并筛选出了 5449 个含有实验突变信息的数据，并且将数据集进行了反转（反转的具体操作描述请见下文），这意味着最终用于训练的数据集中包含有 10998 个突变信息，并且造成活性降低的突变和造成活性提高的突变数目相等。最终获得的模型，不仅能预测指定蛋白酶序列中任意位置的突变，且在实验测试集上的性能超出预期。该方法可以与预测蛋白质突变后稳定性的方法合并使用，以解决蛋白酶稳定性和活性之间的这种负相关，即所谓的活性-稳定性权衡机制。不仅能够提升蛋白酶的稳定性，同时解决活性可能下降的问题，应用前景较为广泛。

本公开的方法的有益技术效果包括：

1. 创新性高：基于自行构建的数据集，采用最新的 AI 方法预测突变对蛋白酶活性的影响，填补了本领域的空白。

2. 性能优越：后文中通过实施例中的方法与仅有的 SCANEER 对比结果可以看出，本公开的方法性能更优。

3. 应用前景广泛：该方法可以预测现在仅有的 SCANEER 无法预测的突变对蛋白酶活性的影响。其中不能预测的突变部分更有意义，因为 SCANEER 能够预测的部分与理性设计的方法重合度很高，所筛选出的结果容易被其他专利或者文献所限制。而本公开的方法可以发现进化信息难以发现的有益突变位点。

附图说明

通过以下详细的描述并结合附图将更充分地理解本公开，其中相似的元件以相似的方式编号，其中：

图 1 是根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法的流程图。

图 2 是根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法与用于预测突变对蛋白酶活性影响的预测模型的构建方法的关联示意图。

图 3 是本公开所使用的用于预测突变对蛋白酶活性影响的预测模型的构建方法的流程图。

图 4 是邻接矩阵示意图。

图 5 是形成节点特征矩阵的示意图。

图 6 是基于结构的邻接矩阵构建的示意图。

图 7 是基于序列的节点特征矩阵构建的示意图。

图 8 是预测模型的框架示意图。

图 9 示出了根据本公开的用于预测突变对蛋白酶活性影响的系统的示意框图。

具体实施方式

除非另有说明，本公开所用的技术和科学术语具有与本公开所属领域的普通技术人员通常所理解的含义。

下面通过实施例，并结合附图，对本公开的技术方案作进一步详细的说明。除非另有说明，下文描述的实施例的方法和材料均为可以通过市场购买获得的常规产品。本公开所属领域的技术人员将会理解，下文描述的方法和材料，仅是示例性的，而不应视为限定本公开的范围。

从预测方法的总体角度来看，通过图 1 示出了根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法的流程图。

如图 1 中所述，根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法 100 开始于步骤 S110，在此步骤，基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得待预测蛋白酶样本的序列特征。

这里所述的待预测蛋白酶样本序列信息可以包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息。

另一种情况下，所述待预测蛋白酶样本序列信息也可以包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息。在此情况下，可以根据所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息获得所述待预测蛋白酶样本在突变后的氨基酸序列信息。所述指定突变信息可以包括指定突变位点和/或指定突变后氨基酸类型。例如，所述指定突变信息是指定突变位点的情况下，可以将该位点上的氨基酸进行变化，从而形成所述待预测蛋白酶样本在突变后的氨基酸序列信息。也即，在所述指定突变信息不包括指定突变后氨基酸类型的情况下，突变后的氨基酸序列信息可以是将突变位点的氨基酸替换为其他 19 种氨基酸的 19 条氨基酸序列。在所述指定突变信息还包括指定突变后氨基酸类型的情况下，突变后的氨基酸序列信息可以是一条氨基酸序列。另一种情况是所述指定突变信息只包括指定突变后氨基酸类型不包括突变位点，在此情况下，需要将突变前序列的每个位点依次替换成指定的氨基酸类型，形成 N 条突变后的

序列（N 为序列上位点数量）。此外，另一种极端情况是，所指定的突变位点是多个位点甚至所有位点，或者所指定的突变后氨基酸类型为多种类型或所有其他 19 种类型，这种情况下，则需要遍历所有的突变可能性，以形成最多 19*N 条突变后的序列。这里所述的“氨基酸类型”是指二十种组成生命体中蛋白质主要单元的不同氨基酸，包括指甘氨酸、丙氨酸、缬氨酸、亮氨酸、异亮氨酸、甲硫氨酸（蛋氨酸）、脯氨酸、色氨酸、丝氨酸、酪氨酸、半胱氨酸、苯丙氨酸、天冬酰胺、谷氨酰胺、苏氨酸、天门冬氨酸、谷氨酸、赖氨酸、精氨酸和组氨酸。

本领域技术人员应理解，在后续步骤，需要对突变前和突变后的氨基酸序列分别进行特征提取，最后拼接成关于氨基酸序列的特征，因此，无论最初获取的蛋白酶序列信息是什么，在执行步骤 S110 时，需要形成突变前和突变后的氨基酸序列。

具体地，在步骤 S110，将待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息分别使用预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息。然后，还是在步骤 S110，将突变前的序列特征信息和突变后的序列特征信息进行拼接，得到待预测蛋白酶样本的序列特征。

在自然语言处理研究领域中，随着计算机算力的不断增强，越来越多的通用语言表征的预训练模型（Pre-trained Models，PTMs）逐渐涌现出来。这对下游的预处理任务非常有帮助，能够从海量未标注的数据上学习语言本身的知识，而后在少量带有标签的数据上微调，从而使下游任务能够更好地学习到语言本身的特征和特定任务的知识。通过在蛋白质或者核酸序列语料库上训练，可以获得能够用于生物领域的预训练模型，现有的在蛋白质序列语料库上训练的模型主要有 ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL。尽管同样也可以自行在蛋白质序列语料库上训练获得预训练模型，但通常受限于成本和时间问题，自行训练的模型性能通常无法达到上述模型的水平。

如前所述，预训练模型本质上是自然语言模型。在生物领域一般采用针对生物领域的特点和用途进行改进与优化的预训练模型。该预训练模型已从现有蛋白质语料库中的大量未标注序列数据上学习生物领域相关知识，能够用于从训练样本的蛋白质序列信息中提取与突变相关的特征；再基于提取的突变相关特征进行训练，仅需使用少量带标签的训练样本即可进行模型的参数优化。例如，本公开中所述的预训练模型可以采用以上提及并在以下再次罗列的模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、

ProtElectra、ProtT5-XL 和 ProtT5-XXL。较优选地，预训练模型可以是 ProtT5-XL 或 ProtT5-XXL。更优选地，预训练模型为 ProtT5-XL。即，这里采用的预训练模型优选为基于最新的自然语言处理方法的采用 T5（即，Transfer Text-to-Text Transformer）预训练模型。

步骤 S110 的目的是得到待预测蛋白酶样本的序列特征。而接下来要描述的步骤 S120 的目的则是得到待预测蛋白酶样本的结构特征。

在步骤 S120，基于待预测蛋白酶样本的三维结构信息，获得待预测蛋白酶样本的结构特征。

除了通过预训练模型获得相关特征，还可以结合蛋白酶的三维结构信息获得相关特征用于补充信息。三维结构信息可以通过 PDB 数据库（https://www.rcsb.org/）（参见 Berman, Helen M 等. 2000. “The protein data bank”. *Nucleic acids research* 28(1): 235 - 42。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）获得或者通过预测软件获得，例如以下预测软件（括号中为其参考介绍或来源）：

AlphaFold2（https://alphafold.ebi.ac.uk/）；

I-TASSER（https://zhanggroup.org/I-TASSER/）；

RoseTTAFold（参见 Minkyung Baek et al. “Accurate prediction of protein structures and interactions using a three-track neural network”. *Science* 373, 871-876 (2021). DOI:10.1126/science.abj8754）；

Modeller（https://salilab.org/modeller/）；

Swiss-model（https://swissmodel.expasy.org/）等。

可以获得的蛋白酶的三维结构信息通常包括二级结构、相互作用网络、氨基酸残基距离、物理环境以及通过三维结构获得的能量信息等。其中，相互作用网络因为包含了所有残基对之间的接触信息，是基于结构的预测方法中最常用的特征之一。

根据本公开的优选实施例，所利用的三维结构信息是相互作用网络。更具体地说，正是基于相互作用网络的特征而选择了基于图神经网络（GNN）及其各种变体的预测模型。

本领域技术人员应该理解，本公开并没有限制步骤 S110 与步骤 S120 的先后顺序。也就是说，序列特征和结构特征的获得，可以是依照任何次序进行的，也可以是同时进行的。

最后，在步骤 S130，基于在步骤 S110 获得的序列特征和在步骤 S120 获得的结构特征，使用基于图神经网络的预测模型进行预测，确定待预测蛋白酶样本突变后的活性变化预测结果。

基于图神经网络的预测模型实际上是分类器。

在步骤 S130，将基于序列的特征和基于结构的特征放入分类器中对特定的任务进行预测。分类器一般分为机器学习方法和深度学习方法。机器学习方法一般常用：线性回归（Linear Regression）算法、支持向量机（Support Vector Machine，SVM）算法、最近邻居/k-近邻（K-Nearest Neighbors，KNN）算法、逻辑回归（Logistic Regression，LR）算法、决策树（Decision Tree）算法、k-平均（K-Means）算法、随机森林（Random Forest）算法、朴素贝叶斯（Naive Bayes）算法、梯度增强（Gradient Boosting）算法、集成（Ensemble Learning）方法。深度学习方法一般包括：卷积神经网络（Convolutional Neural Network）、循环神经网络（Recurrent Neural Network）、递归神经网络（Recursive Neural Network）、长短期记忆网络（LSTM）。图神经网络（GNN）是一种基于图结构数据的深度学习模型，用于处理节点和边的结构化数据。图卷积网络（Graph Convolutional Network，GCN）是一种基于卷积神经网络（CNN）的 GNN 模型，通过对邻居节点的信息进行聚合，实现了对节点特征的有效表示和学习。图注意力网络（Graph Attention Network，GAT）是一种基于注意力机制的 GNN 模型，通过对邻居节点的注意力权重进行学习，实现了对节点特征的自适应聚合，可以更好地处理不同节点之间的关系。GraphSAGE（Graph Sampling and Aggregation， SageConv）是一种基于采样的 GNN 模型，通过对邻居节点的采样和聚合，实现了对节点特征的有效学习和表示。GraphSAGE 可以处理大规模图数据，同时保证了模型的效率和精度。图同构性网络（Graph Isomorphism Network，GINConv）是一种基于图同构性的 GNN 模型，通过对邻居节点的特征进行聚合，并将聚合结果与节点自身的特征进行组合，实现了对节点特征的有效学习和表示。扩散卷积神经网络（Diffusion Convolutional Neural Network，DCNN）是一种基于扩散过程的 GNN 模型，通过对节点特征进行扩散和聚合，实现了对节点特征的有效学习和表示。这些方法都有各自的优点和适用范围，可以根据具体的任务和数据特点选择合适的方法。

在本公开的一个优选实施例中，所述图神经网络是图卷积网络或图注意力网络。

此外，在本公开的一个优选实施例中，在步骤 S130，可以将所述序列特征和所述结构特征首先输入到图神经网络，得到关于活性变化的概率特征；然后，将所述活性变

化的概率特征与来自其他预测方法的补充特征（参见后文的详细描述）相结合，输入到决策树模型，最终得到关于活性变化的分类。在一个更优选实施例中，所采用的决策树模型是梯度提升决策树模型。使用 LightGBM 框架来实现该决策树模型。后文对于实施例的描述中将对此给出更加详细的说明。

有关预测模型的构建方法，可参见后文的详细描述。

如前所述，稳定性和催化活性是蛋白酶的两个最重要的属性，其中稳定性研究较多，已经发展出数十种基于 AI 的方法预测突变对蛋白酶稳定性的影响方法并且性能优秀，而催化活性一直受限于数据集和方法的问题至今停滞不前，仅有 SCANEER 方法能够预测单点突变对蛋白酶活性的影响，但是该方法由于基于协同进化信息仅能够预测进化中出现过的突变，限制较多，无法广泛应用。

本公开所提供的技术与上述现有技术的区别在于，本公开的技术将图神经网络，特别是图卷积网络或图注意力网络，引入到对蛋白酶突变后活性变化的预测中，与现有技术中成熟的蛋白质自然语言模型（预训练模型）相结合，共同作用于预测结果的确定。换句话说，使得待预测蛋白酶样本的序列特征（根据一级序列得到）与结构特征（根据三维结构得到）相结合，共同作为基于图神经网络的预测模型的输入，从而输出得到关于蛋白酶突变后活性变化的预测结果。这样确定的预测结果优于通过现有技术的方法确定的预测结果。

图 2 给出了根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法与用于预测突变对蛋白酶活性影响的预测模型的构建方法的关联示意图 200。在图 2 所示的示意图 200 中，虚线左侧示出的是基于图神经网络的预测模型构建方法，而虚线右侧示出的是使用基于图神经网络的预测模型进行预测的方法。

图 2 中虚线右侧的流程其实就是图 1 所示的根据本公开的实施例的用于预测突变对蛋白酶活性影响的方法。图 2 中虚线左侧的流程还会在下文中详细介绍（例如图 3 及其相应文字描述）。从图 2 可以看出，无论是模型构建还是实际预测过程，都需要使用两种特征提取的步骤，即：序列特征和结构特征的提取。对于训练样本来说，其包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息，也包括突变活性变化标签。如后文所介绍的，突变活性变化标签可以为蛋白酶活性变强或变弱（上升或下降）。因此，在模型构建过程中，使用训练样本中的序列信息来提取序列特征。通

过三维结构信息来提取结构特征。通过作为输入数据的序列特征和结构特征、作为输出数据的训练样本中的突变活性变化标签，对基于 GNN 的预测模型即分类器进行充分训练，从而得到最终的基于 GNN 的用于预测突变对蛋白酶活性影响的预测模型。对于构建好的预测模型，就可以投入到蛋白酶突变活性变化预测的实际工作中去，即图 2 中虚线右侧的流程。此外需要说明的是，图 2 中，对于序列信息与序列特征，用空心箭头表示其信号传递过程；而对于三维结构信息和结构特征，则用相区别的实心箭头表示其信号传递过程；此外，其他的信号流、控制流也都使用实心箭头来表示。

下面来详细介绍根据本公开的实施例的用于预测突变对蛋白酶活性影响的预测模型的构建方法的流程。

图 3 是本公开所使用的用于预测突变对蛋白酶活性影响的预测模型的构建方法 300 的流程图。

本领域技术人员应该知道，如图 3 中所示，在步骤 S310，首先应当获取训练集，所述训练集包括多个训练样本的信息。具体地说，每个所述训练样本的信息包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息以及突变活性变化标签。

人工智能（AI）算法严重依赖数据，初始数据的数量和质量决定了训练得到的模型的泛化性能，最直接的体现在模型的预测准确度上。数据集的样本数量不足或者质量过低会导致模型出现过拟合或者欠拟合的问题，为了解决这个问题，最有效的方法是获取更多的实验数据作为训练集。现有获取实验数据的主要来源为公开数据库，对于基于序列的方法预测突变对蛋白酶催化活性影响的任务，可以通过对数据库中的数据做反转处理将数据集变为原来的两倍大小。反转，即在相同的环境条件下，相同位置的相反的突变是相反的活性作用。根据以上定义，将原始数据与反转后的数据进行合并后可以变为原来的两倍数据量。通过反转这种方法，既扩大了原始的数据集，也同时平衡了数据集，使最终预测模型的鲁棒性更强。

需要注意的是，这种通过对原始数据集反转而实现数据集扩大的方式仅仅针对训练集实施。

在下文的具体实施例中，将更加详细描述训练数据集的获取。

突变活性变化标签可以用于反映氨基酸突变对蛋白酶催化活性的影响趋势。例如，突变活性变化标签可以为催化活性变强或变弱（活性上升或下降）。更具体地，可参见

后文具体实施例中关于构建训练数据集的说明。可以理解的是，使用突变活性变化信息作为训练样本的标签，使训练获得的预测模型可以输出关于氨基酸突变对蛋白酶催化活性影响的预测结果。优选地，本申请提到的蛋白酶突变可以是蛋白酶氨基酸序列的单点突变。

在步骤 S320，基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征。

具体地说，在步骤 S320，将所述样本蛋白酶突变前和突变后的氨基酸序列信息分别使用所述预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息。然后，还是在步骤 S320，将突变前的序列特征信息和突变后的序列特征信息进行拼接，得到所述多个训练样本的序列特征。

如前文所述，预训练模型采用以下模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL。较优选地，所述预训练模型为 ProtT5-XL 或 ProtT5-XXL。更优选地，所述预训练模型为 ProtT5-XL。即，这里采用的预训练模型优选为基于最新的自然语言处理方法的采用 T5（即，Transfer Text-to-Text Transformer）预训练模型。

步骤 S320 的目的是得到训练样本的序列特征。而接下来要描述的步骤 S330 的目的则是得到训练样本的结构特征。

在步骤 S330，基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征。

如前所述，可以通过以下数据库或预测软件中的至少一种获取样本蛋白酶的三维结构信息：PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold、Modeller 和 Swiss-model 等。

可以获得的蛋白酶的三维结构信息通常包括二级结构、相互作用网络、氨基酸残基距离、物理环境以及通过三维结构获得的能量信息等。其中，相互作用网络因为包含了所有残基对之间的接触信息，是基于结构的预测方法中最常用的特征之一。

根据本公开的优选实施例，这里所采用的三维结构信息是相互作用网络。更具体地说，正是基于相互作用网络的特征而选择了基于 GNN 的预测模型。

本领域技术人员应该理解，本公开并没有限制步骤 S320 与步骤 S330 的先后顺序。也就是说，序列特征和结构特征的获得，可以是依照任何次序进行的，也可以是同时进行的。

在步骤S340，基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型。

如前文所述，这个预测模型实际上是分类器。

在本公开的一个优选实施例中，所述图神经网络是图卷积网络（GCN）或图注意力网络（GAT）。

有关预测模型的构建方法，可参见后文实施例中的详细描述。

此外，需要注意的是：对于序列特征，通过将已经处理好的蛋白酶活性变化数据集放入预训练模型中可以通过迁移学习把预训练好的模型迁移到新的任务上。在迁移学习中，有两种常用的应用方式：特征提取和微调。在特征提取中，可以在预先训练好网络结构后，修改或添加一个简单的分类器，将原来任务上的预先训练好的网络作为另一个目标任务的特征提取器，只对最后增加的分类器参数进行重新学习，而预先训练好的网络参数不会被修改或冻结。本公开的优选实施例采用了特征提取的迁移学习方式。

在模型训练过程中，使用序列特征、结构特征作为预测模型的输入数据，使用突变活性变化标签作为预测模型的输出目标，对所述预测模型进行充分训练。随着预测模型的参数的不断优化，预测模型的输出数据将与突变活性变化标签所表示的预测结果相吻合。

在本公开的一种优选实施方式中，步骤S340可以进一步包括：将序列特征和结构特征输入到图神经网络，得到关于活性变化的概率特征；然后，将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。

更具体地说，在本公开的一些实施例中，可以基于多个训练样本的序列特征、结构特征和突变活性变化标签，对预测模型（分类器）进行有监督训练，得到训练后的分类器。有监督训练是利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程。在一些实施例中，可以构建损失函数，如交叉熵损失函数，通过损失函数来反映分类器预测的蛋白酶突变后催化活性变化预测结果与训练样本实际的突变活性变化标签的差异，进而根据该差异对分类器的参数进行调整，得到分类器的模型优化参数。当损失函数满足预定条件时，如损失函数收敛，或损失函数值小于预设值，或训练集和测试集损失函数值同时降到最低点或者训练集损失函数值到最低点，分类器训练完成。在一些实施例中，可以采用参数搜索算法调整分类器参数，例如，搜索算法，如网格搜索、

贝叶斯搜索等。在一些实施例中，有监督训练为根据突变活性变化标签训练预测模型的过程。

本领域技术人员应该理解，通过例如图 3 的方法得到的预测模型可用于图 1 中所示的预测方法。

下面将给出具体示例实施例。

实施例方法与材料

1. 训练数据集

为了克服数据质量差所带来的问题，从 2021 年发表的 D3DistalMutation 数据库（参见 Wang, Xiaoyu 等 . 2021. “D3DistalMutation: a database to explore the effect of distal mutations on enzyme activity”. *Journal of Chemical Information and Modeling* 61(5): 2499 - 2508。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）中筛选出 5449 个包含活性变化信息的实验样本， 其中降低活性的突变 5279 个样本，提高活性的突变 220 个样本， 具体的筛选标准如下：

（1）去除所有的多点突变样本，将训练数据集限制为单点突变数据集；

（2）去除中性突变（对蛋白酶活性无影响的突变），并将活性降低或活性丧失的突变统一定义为降低活性的突变；

（3）去除同一个突变具有不同标签的样本。这里需要去除具有不同标签的样本的原因在于：同一个突变具有不同的标签的样本，可能是实验条件不同或实验误差导致的，这就很难给这类数据定义标签， 若不去除这类样本， 可能会给模型带来噪音进而影响准确性；

（4）将数据反转合并。

最终，用于训练模型的数据集包含有 10898 个样本。

2. 测试数据集

测试数据来源于 BRENDA 数据库（https://brenda-enzymes.org/news.php?）（参见 Chang, Antje 等. 2021. “BRENDA, the ELIXIR core data resource in 2021: new developments and updates”. *Nucleic acids research* 49(D1): D498 - 508。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）、文献（例如，参见 Amorosi, Clara J 等. 2021. “Massively parallel characterization of CYP2C9 variant enzyme

activity and abundance". *The American Journal of Human Genetics* 108(9): 1735 - 51、Cheng, Ya-Shan 等. 2015. "Improving the catalytic performance of a GH11 xylanase by rational protein engineering". *Applied microbiology and biotechnology* 99: 9503 - 10、You, Chun, Hanying Yuan, Qiang Huang 和 Hong Lu. 2010. "Substrate molecule enhances the thermostability of a mutant of a family 11 xylanase from Neocallimastix patriciarum". *African Journal of Biotechnology* 9(9)。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）和专利（WO2020234200A1；EP3854872A1；WO2018195850；WO2017109262A1）。筛选标准如下：

（1）去除非单点突变样本，将测试数据集限制为单点突变数据集；

（2）去除中性突变样本，并将活性降低或活性丧失的突变统一定义为降低活性的突变；

（3）去除具有多个标签的样本。多个标签的样本就是同一个突变位点的蛋白酶活性报道的有不同活性的，就需要删除该突变的样本，保证数据的准确性；

（4）去除与训练集重复的样本。

最终，用于测试模型效果的数据集包含有 6103 个样本, 其中降低活性和提高活性的突变样本数分别为 4950 和 1153。

2.1 S3459

SCANEER 由于算法的限制只能预测进化中出现过的突变。在测试集 6103 个数据中，有 3459 个样本是 SCANEER 能够预测的，在这其中有 1797 个样本来自 DMS（Deep Mutational Scanning）。DMS 是在蛋白质的基因中引入随机突变，然后通过压力筛选或高通量测序等方法来评估这些突变对蛋白质功能的影响。剩下的 1662 个样本为非 DMS 数据。

2.2 S2644

所有的测试数据中， SCANEER 不能预测的突变有 2644 个样本。同 2.1，这部分数据也分为 DMS 数据和非 DMS 数据， 样本量分别为 1492 和 1152。

3. 预训练模型

用于本方法的 ProtT5-XL（ProtT5-XL-U50/ ProtT5-XL-UniRef50）模型是 Ahmed Elnaggar 等人在 2021 年发布的，使用 T5 模型首先在 BFD 数据集上训练，再用

UniRef50 数据集对模型做微调优化得到的预训练蛋白模型。该模型含有 24 层注意力结构，每一层由 32 个自注意力头和 1024 个神经元构成。可参见 Elnaggar, Ahmed 等. 2021. "Prottrans: Toward understanding the language of life through self-supervised learning". *IEEE transactions on pattern analysis and machine intelligence* 44(10): 7112 - 27。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分。

4. 图卷积神经网络模型构建

本优选实施例基于图卷积网络（GCN）的变体： 图注意力网络（Graph Attention Network， GAT）（参见 Veličković, Petar 等. 2017. "Graph attention networks". *arXiv preprint arXiv:1710.10903*。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）。

4.1 GAT-1 模型构建

4.1.1 邻接矩阵

通过 PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold 、Modeller 和 Swiss-model 等预测软件获得每条蛋白酶的三维结构，根据获得的蛋白酶三维结构的坐标计算蛋白酶三维结构中的氨基酸残基对之间距离，当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0（自身为 0），即为整个蛋白酶的邻接矩阵。

在本实施例中以突变位点为中心左右各取 10 个氨基酸，即 10 个连续氨基酸作为残基序列，从整个蛋白酶的邻接矩阵中截取 21*21 的邻接矩阵。

图 4 是邻接矩阵示意图。如图所示，长度为 100 个氨基酸的邻接矩阵，其中假设突变位点为残基 12，则方框为该残基的 21*21 的邻接矩阵。使用 McBASCCovariance 算法（参见 McLachlan, Andrew D. 1971. "Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551". *Journal of molecular biology* 61(2): 409 - 24。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）计算邻接矩阵中对应的 21 个氨基酸两两之间的共进化分数，组成 21*21 的共进化分数矩阵， 将其与邻接矩阵相乘得到更新后的 21*21 邻接矩阵。

4.1.2 节点特征矩阵

将一条蛋白酶序列信息输入到预训练模型后，首先模型将蛋白酶序列的每一个氨基酸通过编码器进行编码，每一个氨基酸编码后对应 1024 长度的向量，即一个长度为 N 的蛋白酶序列将会得到 N*1024 长度的向量。该模型拥有 24 层结构，为了能够最大程度的获取有关突变的信息，可以选取最后一层作为输出，并分别提取野生型和突变型对应突变位置的 1024 维的特征向量，将野生型（即突变前）和突变型（即突变后）的特征向量依次拼接为 2048 维。以 21 为滑窗大小，突变位点为中心，前后 10 个氨基酸残基进行拼接，最终得到 21*2048 维特征作为节点特征矩阵，结构如图 5 所示。

4.1.3 模型的构建

将 4.1.1 和 4.1.2 中生成的矩阵分别作为结构特征和序列特征输入图卷积神经网络中。通过矩阵相乘与聚合操作，三层图网络结构更新节点特征分别为 32、64 和 64 维特征向量，每层采用 4 个注意力头，通过残差网络结构将三层图网络输出的节点特征合并为 21*640 维特征向量。然后采用平均池化操作将更新以后的特征矩阵压缩为 1*640 维特征向量表示为 21 个氨基酸残基序列的特征表征。具体的参数为：三层图卷积输出维度分别是 32，64，64；全连接层神经元数量为 64，2；优化器采用 Adam；损失函数为二元交叉熵损失；‘relu’作为激活函数；‘softmax’作为输出层分类函数；l2 正则化项系数为 0.05，dropout 率为 0.3；学习率为 0.001。

这里所说的聚合操作是通过注意力机制来实现的，其中每个节点都有一个注意力权重分布，将特征矩阵和注意力系数进行加权求和， 即可得到节点的聚合特征向量。通过这样的聚合操作，每个节点都能够利用其邻居节点的特征信息进行特征更新和信息传递。

这里所说的残差网络是一种深度卷积神经网络结构，它通过引入残差连接来构建网络层之间的跳跃连接。即每个网络层的输入不仅仅是前一层的输出，还包括一个直接连接的跳跃连接。这个跳跃连接将输入直接添加到输出中，形成了残差。其目的是解决深层网络训练中的梯度消失和网络难以训练的问题。

这里所说的平均池化操作是一种常用的降低卷积层输出的特征维度的技术， 是通过将卷积层输出的特征求平均值实现的。其目的是有效减少网络参数，防止出现过拟合现象。本领域技术人员应理解，这里的平均池化操作也可以替换为最大池化或 K-max 池化等任意一种池化操作。

4.2 GAT-2 模型构建

直接使用 4.1.1 和 4.1.2 中生成的矩阵输入图卷积神经网络中。通过矩阵相乘与聚合操作，三层图网络结构中每一层图网络结构更新节点特征成 128 维特征向量，每层采用 4 个注意力头，通过残差网络结构将三层图网络输出的节点特征合并为 21*1536 维特征向量。然后采用平均池化操作将更新以后的特征矩阵压缩为 1*1536 维特征向量表示为 21 个氨基酸残基序列的特征表征。具体的参数为：三层图卷积输出维度分别是 128，128，128；全连接层神经元数量为 64，2；优化器采用 Adam；损失函数为二元交叉熵损失；‘relu’作为激活函数；‘softmax’作为输出层分类函数；l2 正则化项系数为 0.01，dropout 率为 0.5；学习率为 0.001。

4.3 GAT-3 模型构建

4.3.1 一阶邻接矩阵

首先按照 4.1.1 获得整个蛋白酶的邻接矩阵，然后以突变位点残基为中心取所有空间上与突变位点距离小于 10 埃的氨基酸，计算获得这些氨基酸组成的邻接矩阵。

4.3.2 节点特征矩阵

将一条蛋白酶序列信息输入到预训练模型后，首先模型将蛋白酶序列的每一个氨基酸通过编码器进行编码，每一个氨基酸编码后对应 1024 长度的向量，即一个长度为 N 的蛋白酶序列将会得到 N * 1024 长度的向量。该模型拥有 24 层结构，为了能够最大程度的获取有关突变的信息，本实施例选取了最后一层作为输出，并同时提取野生型和突变型对应突变位置的 1024 维的特征向量，将野生型和突变型的特征向量依次拼接为 2048 维，最终获得 N_f*2048 维特征作为节点特征矩阵，其中 N_f 是从 4.3.1 中获得的对应氨基酸的数量，顺序保持一致，不同蛋白酶获得的氨基酸数量可能不同，采用与最大 N_f 为基准，不足的补 0，这样所有的蛋白酶都能够获得相同大小的矩阵。

4.3.3 模型的构建

将 4.3.1 和 4.3.2 中生成的矩阵分别作为结构特征和序列特征输入图卷积神经网络中。通过矩阵相乘与聚合操作，两层图网络结构分别更新节点特征成 128 和 256 维特征向量，每层采用 4 个注意力头，通过残差网络结构将两层图网络输出的节点特征合并为 N_f*1536 维特征向量。然后采用平均池化操作将更新以后的特征矩阵压缩为 1*1536 维特征向量表示为 N_f 个氨基酸残基序列的特征表征。具体的参数为：两层图卷积输出维度

分别是 128，256；全连接层神经元数量为 128，2；优化器采用 Adam；损失函数为二元交叉熵损失；‘relu’作为激活函数；‘softmax’作为输出层分类函数；l2 正则化项系数为 0.05，dropout 率为 0.3；学习率为 0.001。

5. LightGBM 模型构建

梯度提升决策树（Gradient Boosting Decision Tree，GBDT）（参见 Friedman, Jerome H. 2001. “Greedy function approximation: a gradient boosting machine”. *Annals of statistics*: 1189 - 1232。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）是机器学习中一种常用的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型。而 LightGBM（Light Gradient Boosting Machine，或简称为 LGBM）是一个实现 GBDT 算法的框架（参见 Ke, Guolin 等. 2017. “Lightgbm: A highly efficient gradient boosting decision tree”. *Advances in neural information processing systems* 30。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）。本公开的方法中所使用的 LightGBM 为版本 3.3.3。

5.1 LGBM-1 模型构建

5.1.1 特征提取

从 4.1 构建的 GAT-1 模型获取概率值（代表蛋白酶活性升高的概率）作为一维特征，基于 PremPS（参见 Chen, Yuting 等. 2020. “PremPS: Predicting the impact of missense mutations on protein stability”. *PLoS computational biology* 16(12): e1008543。在此通过援引，将上述文献的全部内容合并到本公开中，使之成为本公开的内容的一部分）和 SCANEER（如前所述）方法描述的特征提取补充特征 11 维，拼接后共获取特征 12 维。其中，由 PremPS 补充 10 维，由 SCANEER 补充 1 维。

5.1.2 模型的构建

将获取的 12 维特征输入到 LightGBM 框架中进行训练，学习速率设置为 0.05，extra_trees 设置为 True， 其他参数使用默认值。

5.2 LGBM-2 模型构建

5.2.1 特征提取

同 5.1.1，从 4.2 和 4.3 构建的 GAT-2 和 GAT-3 模型中获取概率值（代表蛋白酶活性升高的概率）作为两维特征， 基于 PremPS 和 SCANEER 方法描述的特征提取补充特征 11 维，拼接后共得到特征 13 维。

5.2.2 模型的构建

将获取的 13 维特征输入到 LightGBM 框架中进行训练，学习速率设为 0.05，extra_trees 设置为 True， 其他参数使用默认值。

下面进行一些总结。

图 6 示出了基于结构的邻接矩阵构建的示意图。图 7 示出了基于序列的节点特征矩阵构建的示意图。图 8 示出了预测模型的框架示意图。

根据以上的实施例，可以看出，在构建预测模型的过程中，样本蛋白酶的三维结构信息可以通过以下方式获取：获取样本蛋白酶的三维结构的坐标；根据获取的三维结构的坐标计算样本蛋白酶三维结构中氨基酸残基对之间的距离；当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个样本蛋白酶的邻接矩阵。基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征可以包括：以突变位点为中心，从所述整个样本蛋白酶的邻接矩阵中截取指定大小的邻接矩阵，作为所述样本蛋白酶的结构特征。基于序列特征、结构特征以及突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型可以进一步包括：将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征可以进一步包括：将所述序列特征和所述结构特征输入到图神经网络；通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；将若干层图网络结构分别输出的特征合并；采用池化操作对特征进行压缩；根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。这里所说的池化操作可以是平均池化、最大池化或 K-max 池化。

相应地，在将预测模型应用于蛋白酶突变后活性变化预测的过程中，待预测蛋白酶样本的三维结构信息通过以下方式获取：获取待预测蛋白酶样本的三维结构的坐标；根

据获取的三维结构的坐标计算蛋白酶三维结构中氨基酸残基对之间的距离；当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个待预测蛋白酶样本的邻接矩阵。基于待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征包括：以突变位点为中心，从所述整个待预测蛋白酶样本的邻接矩阵中截取指定大小的邻接矩阵，作为所述待预测蛋白酶样本的结构特征。基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果可以进一步包括：将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类。更具体地，所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征包括：将所述序列特征和所述结构特征输入到图神经网络；通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；将若干层图网络结构分别输出的特征合并；采用池化操作对特征进行压缩；根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。同样地，这里所说的池化操作可以是平均池化、最大池化或 K-max 池化。

6. 性能指标

本实施例通过以下指标进行评估，即曲线下面积（AUC）、精确度（precision）、召回率（recall）。具体定义如下：

真阳性(True Positive，TP)

假阳性(False Positive，FP)

真阴性(True Negative，TN)

假阴性(False Negative，FN)

其中，TP (True Positive)、TN (True Negative) 表示分类正确的阳性样本和阴性样本个数，FP (False Positive)、FN (False Negative) 表示分类错误的阳性样本和阴性样本个数，P、N 表示阳性样本和阴性样本数。

AUC 是一种流行的与参数无关的度量，用于描述二元分类器。AUC 即为接收者操作特征曲线（ROC）的曲线下面积，数值越大代表分类器性能越好，最大数值为 1。

精确度（precision），是指在所有预测为阳性样本的结果中，正确预测的结果所占的比例；精确度计算公式为：precision = TP/(TP+FP)，其中，TP（True Positive）为被判定为分类正确的阳性样本，FP（False Positive）为被判定为分类错误的阳性样本。

召回率（recall），是指预测正确的阳性样本数量占总的阳性样本数量的比例；召回率计算公式为：recall = TP/(TP+FN)，其中，FN（False Negative）为被判定为分类错误的阴性样本。

结果评价

1. 模型性能

通过在自行设计的平衡数据集上训练模型，该数据集包含来自 D3DistalMutation 数据库的 1303 个蛋白酶的 10898 个单点突变对酶活性的影响。在五折交叉验证中，模型对测试数据的性能如下表 1 所示。

评价指标	精确度	召回率	AUC
LGBM-1	0.96	0.93	0.97
LGBM-2	0.95	0.96	0.99

表 1：本公开的预测模型在训练集上的五折交叉验证性能指标

2. 与其他方法的比较

为了评估本公开的方法与 SCANEER 方法的差异，将本公开的模型与仅有的预测方法 SCANEER 进行比较。SCANEER 评估了多序列比对（MSA）中氨基酸对的共同进化关系，在 MSA 中未观察到或很少观察到的氨基酸对即氨基酸替换则不能预测，最终导致每个蛋白酶可预测的突变数目仅占所有可能突变的 47%（SCANEER 的作者假设，在 MSA 中未观察到或很少观察到的氨基酸替换，会导致蛋白酶功能丧失，在进化过程中已经被淘汰。所以，作者把 gap 超过 20%或不在排名前 N 的共进化氨基酸对上的位置排除了。即，在 MSA 中，gap 超过 20%或完全保守的位置，以及不在排名前 N 的共进化氨基酸对的位置上的氨基酸无法计算其突变效应（N=蛋白酶长度*2）），这将极大地限制了该方法的应用范围。所以，实验测试集的 6103 个样本中， SCANEER 能预测的样本数目为 3459 个， 其中降低活性和提高活性的突变样本分别为 2855 和 604。3459 个样本中，DMS 和非 DMS 数据样本量分别为 1797 和 1662。比较结果如下表 2 所示。

测试集	方法	评价指标		
		精确度	召回率	AUC
S3459	LGBM-1	0.77	0.12	0.69
	LGBM-2	0.74	0.15	0.68
	SCANEER	0.75	0.10	0.68
S1797	LGBM-1	0.67	0.05	0.73
	LGBM-2	0.70	0.09	0.75
	SCANEER	0.50	0.05	0.73
S1662	LGBM-1	0.78	0.13	0.68
	LGBM-2	0.74	0.16	0.68
	SCANEER	0.78	0.11	0.66

表 2：本公开的模型与 SCANEER 方法在可以比较的实验测试集上的性能比较

从上表中可以看出，在 S3459 和 S1662 上，本公开的方法在精确度和召回率上比 SCANEER 稍优或者基本持平，而在 S1797 上精确度有很大的优势。

S2644 数据集是 SCANEER 不能预测的数据，其中降低活性和提高活性的突变样本分别为 2095 和 549。 这部分数据中，DMS 数据和非 DMS 数据的样本量分别为 1492 和 1152。本公开的方法在这部分数据集上的表现如下表 3 所示。

数据	方法	评价指标		
		精确度	召回率	AUC
S2644	LGBM-1	0.69	0.10	0.71
	LGBM-2	0.73	0.15	0.68
S1492	LGBM-1	0.81	0.13	0.81
	LGBM-2	0.78	0.19	0.78
S1152	LGBM-1	0.62	0.08	0.72
	LGBM-2	0.70	0.14	0.72

表 3：本公开的模型在 SCANEER 无法预测的实验测试数据上的性能

结果显示，在 SCANEER 无法预测的实验测试数据上，本公开的方法仍能得到良好的性能，由此看出本公开的方法展现出的优势。

预测流程

这里以两种情况为例，描述例如预测模型进行蛋白酶活性变化预测的流程。

一种情况是输入的蛋白酶序列信息包括单条蛋白酶序列和突变位点信息。首先，根据输入的单条蛋白酶序列和突变位点信息，进行蛋白酶突变位点预处理，具体地，将输入蛋白酶序列记为原始序列，根据突变位点信息获取突变后的序列。然后，进行氨基酸嵌入层提取，也就是前文所说的序列信息的提取。即，将原始序列输入自然语言处理 T5 预训练模型读取嵌入特征；将突变后的蛋白酶序列输入自然语言处理 T5 预训练模型读取嵌入特征；根据突变位点提取氨基酸嵌入特征，并将突变前后特征进行拼接，作为后续模型的输入。接着是图网络的构建，需要蛋白酶嵌入层特征网络和蛋白酶残基相互作用特征网络。将以上的特征网络应用于基于 GAT 的预测模型，预测氨基酸突变前后蛋白酶活性变化概率。接着，在蛋白酶活性变化概率的基础上，利用 LightGMB 预测氨基酸突变前后蛋白酶活性变化。最后，对预测结果进行输出与展示。对于示例的这种情况而言，可以输出单点突变预测结果。

另一种情况是输入的蛋白酶序列信息仅包括单条蛋白酶序列。首先，根据输入的单条蛋白酶序列，进行蛋白酶突变位点预处理，具体地，将输入蛋白酶序列记为原始序列，依次将蛋白酶序列的每个位点氨基酸突变为其他 19 种氨基酸记为突变后的序列。然后，进行氨基酸嵌入层提取，也就是前文所说的序列信息的提取。即，将原始序列输入自然语言处理 T5 预训练模型读取嵌入特征；将突变后的蛋白酶序列输入自然语言处理 T5 预训练模型读取嵌入特征；根据突变位点提取氨基酸嵌入特征，并将突变前后特征进行拼接，作为后续模型的输入。接着是图神经网络的构建，需要蛋白酶嵌入层特征网络和蛋白酶残基相互作用特征网络。将以上的特征网络应用于基于 GAT 的预测模型，预测氨基酸突变前后蛋白酶活性变化概率。接着，在蛋白酶活性变化概率的基础上，利用 LightGMB 预测氨基酸突变前后蛋白酶活性变化。最后，对预测结果进行输出与展示。对于示例的这种情况而言，可以输出全部提高酶活性的突变概率表格和/或全提高酶活性的突变可能性排序作为最终的预测结果。

结论

更高活性的蛋白酶突变体对于学术界和工业界寻找这些蛋白酶更多的应用是必要的，本公开提出了一种新方法用于发现氨基酸突变对蛋白酶活性的影响。本公开的方法基于先进的自然语言处理模型和图神经网络模型，结合序列和三维结构，可应用于不同的生物领域，并且表现出优于其他方法的性能，证明了本公开的方法在蛋白酶工程中的广阔的应用前景。

下面讨论本公开的方法优越于现有技术的原因。从以上的描述中可以看出：

1. 总的来说，本公开的方法基于蛋白酶的三级结构和氨基酸序列信息，通过 AI 训练构建预测模型。

2. 本公开所构建的预测模型可以应用于蛋白酶突变对其活性的预测。

3. 用于本公开的方法的训练数据集经过反转后，既扩大了原始的数据集，也同时平衡了数据集，使最终预测模型的鲁棒性更强。

本领域技术人员应该理解，基于本公开的用于预测突变对蛋白酶活性影响的方法，可以开发出一种用于预测突变对蛋白酶活性影响的系统。图 9 示出了根据本公开的用于预测突变对蛋白酶活性影响的系统的示意框图。具体地说，用于预测突变对蛋白酶活性影响的系统 900 包括获取模块 910 和处理模块 920。

获取模块 910 用于获取待预测蛋白酶样本序列信息和三维结构信息。

处理模块 920 用于通过输入待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果。如图 9 中所示，处理模块 920 可以进一步包括：序列特征子模块 921，用于基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；结构特征子模块 922，用于基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；模型预测子模块 923，用于基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

本领域技术人员应该理解，以上所述的用于预测突变对蛋白酶活性影响的系统可以通过计算机实现。例如，可以通过执行计算机程序以实现以下操作：获取待预测蛋白酶样本序列信息和三维结构信息；通过输入待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果，该操作进一步包括如下子操作：基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待

预测蛋白酶样本的序列特征；基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

也就是说，尽管在根据本公开的用于预测突变对蛋白酶活性影响的系统中，将对应于预测方法的各个操作功能描述为模块或子模块，但本领域技术人员应该理解，这样的模块或子模块可以是电路元器件或其他实体组件，也可以不是由电路元器件或其他实体组件构成的，而是通过计算机程序来构建的功能模块。

此外，本领域普通技术人员应该认识到，本公开的方法可以实现为计算机程序。如上结合附图所述，通过一个或多个程序执行上述实施例的方法，程序中的指令使得计算机或处理器执行结合附图所述的算法。这些程序可以使用各种类型的非瞬时计算机可读介质存储并提供给计算机或处理器。非瞬时计算机可读介质包括各种类型的有形存贮介质。非瞬时计算机可读介质的示例包括磁性记录介质（诸如软盘、磁带和硬盘驱动器）、磁光记录介质（诸如磁光盘）、CD-ROM（紧凑盘只读存储器）、CD-R、CD-R/W 以及半导体存储器（诸如 ROM、PROM（可编程 ROM）、EPROM（可擦写 PROM）、闪存 ROM 和 RAM（随机存取存储器））。进一步，这些程序可以通过使用各种类型的瞬时计算机可读介质而提供给计算机。瞬时计算机可读介质的示例包括电信号、光信号和电磁波。瞬时计算机可读介质可以用于通过诸如电线和光纤的有线通信路径或无线通信路径提供程序给计算机。

例如，根据本公开的一个实施例，可以提供一种非瞬时性计算机可读存储介质，用于存储计算机程序，所述计算机程序包括指令，所述指令在由电子设备的处理器执行时使所述电子设备实施如上所述的用于预测突变对蛋白酶活性影响的方法。

另外，根据本公开公开的内容，还可以提供一种计算机系统，所述计算机系统包括：处理器；存储器；和计算机程序。计算机程序存储在所述存储器中并且被配置为由所述处理器执行。所述计算机程序包括用于实施以上所述的用于预测突变对蛋白酶活性影响的指令。

另一方面，根据本公开的一个实施例，可以提供一种非瞬时性计算机可读存储介质，用于存储计算机程序，所述计算机程序包括指令，所述指令在由电子设备的处理器执行时使所述电子设备实施如上所述的用于预测突变对蛋白酶活性影响的预测模型的构建方法。

另外，根据本公开公开的内容，还可以提供一种计算机系统，所述计算机系统包括：处理器；存储器；和计算机程序。计算机程序存储在所述存储器中并且被配置为由所述处理器执行。所述计算机程序包括用于实施以上所述的用于预测突变对蛋白酶活性影响的预测模型的构建方法的指令。

本公开的实施方式并不限于上述实施例所述，在不偏离本公开的精神和范围的情况下，本领域普通技术人员可以在形式和细节上对本公开做出各种改变和改进，而这些均被认为落入了本公开的保护范围。

权 利 要 求 书

1. 一种用于预测突变对蛋白酶活性影响的方法，其特征在于，所述方法包括：

基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；

基于待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；

基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

2. 根据权利要求 1 所述的方法，其特征在于，所述待预测蛋白酶样本序列信息包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息。

3. 根据权利要求 1 所述的方法，其特征在于，所述待预测蛋白酶样本序列信息包括所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息。

4. 根据权利要求 3 所述的方法，其特征在于，所述方法进一步包括：根据所述待预测蛋白酶样本在突变前的氨基酸序列信息和指定突变信息获得所述待预测蛋白酶样本在突变后的氨基酸序列信息。

5. 根据权利要求 3 或 4 所述的方法，其特征在于，所述指定突变信息包括指定突变位点和/或指定突变后氨基酸类型。

6. 根据权利要求 2 或 4 所述的方法，其特征在于，所述的基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征进一步包括：

将所述待预测蛋白酶样本在突变前的氨基酸序列信息和在突变后的氨基酸序列信息分别使用所述预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息；

将所述突变前的序列特征信息和所述突变后的序列特征信息进行拼接，得到所述待预测蛋白酶样本的序列特征。

7. 根据权利要求 1 所述的方法，其特征在于，所述预训练模型采用以下模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL；较优选地，所述预训练模型为 ProtT5-XL 或 ProtT5-XXL；更优选地，所述预训练模型为 ProtT5-XL。

8. 根据权利要求 1 所述的方法，其特征在于，所述待预测蛋白酶样本的三维结构信息包括通过以下数据库或预测软件中的至少一种获取的待预测蛋白酶样本的三维结构信息：PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold 、Modeller 和 Swiss-model。

9. 根据权利要求 1 所述的方法，其特征在于，所述待预测蛋白酶样本的三维结构信息包括相互作用网络、二级结构、氨基酸残基距离或物理环境；优选地，所述待预测蛋白酶样本的三维结构信息包括相互作用网络。

10. 根据权利要求 8 或 9 所述的方法，其特征在于，所述待预测蛋白酶样本的三维结构信息进一步通过以下方式获取：

获取待预测蛋白酶样本的三维结构的坐标；

根据获取的三维结构的坐标计算蛋白酶三维结构中氨基酸残基对之间的距离；

当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个待预测蛋白酶样本的邻接矩阵。

11. 根据权利要求 10 所述的方法，其特征在于，所述的基于待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征包括：

以突变位点为中心，从所述整个待预测蛋白酶样本的邻接矩阵中截取指定大小的邻接矩阵，作为所述待预测蛋白酶样本的结构特征。

12. 根据权利要求 1 所述的方法，其特征在于，所述的基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果进一步包括：

将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；

将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类。

13. 根据权利要求 12 所述的方法，其特征在于，所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征包括：

将所述序列特征和所述结构特征输入到图神经网络；

通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；

将若干层图网络结构分别输出的特征合并；

采用池化操作对特征进行压缩；

根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。

14. 根据权利要求 1 所述的方法，其特征在于，所述的基于图神经网络的预测模型是通过以下步骤构建出来的：

获取训练集，所述训练集包括多个训练样本的信息，每个所述训练样本的信息包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息以及突变活性变化标签；

基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征；

基于样本蛋白酶的三维结构信息，获得所述多个训练样本的结构特征；

基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得所述的基于图神经网络的预测模型。

15. 根据权利要求 14 所述的方法，其特征在于，所述基于图神经网络的预测模型是分类器。

16. 根据权利要求 14 所述的方法，其特征在于，所述图神经网络是图卷积网络或图注意力网络。

17. 根据权利要求 14 所述的方法，其特征在于，所述的基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得所述的基于图神经网络的预测模型进一步包括：

将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；

将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；

以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。

18. 根据权利要求 12 或 17 所述的方法，其特征在于，所述决策树模型是梯度提升决策树（GBDT）模型。

19. 根据权利要求 18 所述的方法，其特征在于，使用 LightGBM 框架来实现 GBDT 模型。

20. 根据权利要求 14 所述的方法，其特征在于，所述的获取训练集的步骤包括：

获取公开数据库中的实验数据作为训练集；

通过对公开数据库中的实验数据做反转处理将训练集变为原来的两倍大小。

21. 一种用于预测突变对蛋白酶活性影响的预测模型的构建方法，其特征在于，所述方法包括：

获取训练集，所述训练集包括多个训练样本的信息，每个所述训练样本的信息包括样本蛋白酶的三维结构信息、样本蛋白酶突变前和突变后的氨基酸序列信息以及突变活性变化标签；

基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征；

基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征；

基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型。

22. 根据权利要求 21 所述的方法，其特征在于，所述的基于样本蛋白酶突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得所述多个训练样本的序列特征包括：

将所述样本蛋白酶突变前和突变后的氨基酸序列信息分别使用所述预训练模型进行特征提取，获得突变前的序列特征信息和突变后的序列特征信息；

将所述突变前的序列特征信息和所述突变后的序列特征信息进行拼接，得到所述多个训练样本的序列特征。

23. 根据权利要求 21 或 22 所述的方法，其特征在于，所述预训练模型采用以下模型中的一种或多种的组合来实现：ESM-1b、UniRef、ProteinBert、TAPE、ProtGPT2、ProtTXL、ProtBert、ProtXLNet、ProtAlbert、ProtElectra、ProtT5-XL 和 ProtT5-XXL；较优选地，所述预训练模型为 ProtT5-XL 或 ProtT5-XXL；更优选地，所述预训练模型为 ProtT5-XL。

24. 根据权利要求 21 所述的方法，其特征在于，所述样本蛋白酶的三维结构信息包括通过以下数据库或预测软件中的至少一种获取的样本蛋白酶的三维结构信息：PDB 数据库、AlphaFold2、I-TASSER、RoseTTAFold 、Modeller 和 Swiss-model。

25. 根据权利要求 21 所述的方法，其特征在于，所述样本蛋白酶的三维结构信息包括相互作用网络、二级结构、氨基酸残基距离或物理环境；优选地，所述样本蛋白酶的三维结构信息包括相互作用网络。

26. 根据权利要求 24 或 25 所述的方法，其特征在于，所述样本蛋白酶的三维结构信息进一步通过以下方式获取：

获取样本蛋白酶的三维结构的坐标；

根据获取的三维结构的坐标计算样本蛋白酶三维结构中氨基酸残基对之间的距离；

当两个残基对之间的阿尔法碳原子距离小于 10 埃时，表示节点之间存在边的关系，记为 1，否则为 0，自身记为 0，得到整个样本蛋白酶的邻接矩阵。

27. 根据权利要求 26 所述的方法，其特征在于，所述的基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征包括：

以突变位点为中心，从所述整个样本蛋白酶的邻接矩阵中截取指定大小的邻接矩阵，作为所述样本蛋白酶的结构特征。

28. 根据权利要求 21 所述的方法，其特征在于，所述的基于所述序列特征、所述结构特征以及所述突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型进一步包括：

将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征；

将所述活性变化的概率特征与来自其他预测方法的补充特征相结合，输入到决策树模型，得到关于活性变化的分类；

以所述突变活性变化标签作为输出目标，对所述预测模型的参数进行调节。

29. 根据权利要求 28 所述的方法，其特征在于，所述的将所述序列特征和所述结构特征输入到图神经网络，得到关于活性变化的概率特征包括：

将所述序列特征和所述结构特征输入到图神经网络；

通过矩阵相乘与聚合操作，使用若干层图网络结构对特征进行更新后输出；

将若干层图网络结构分别输出的特征合并；

采用池化操作对特征进行压缩；

根据输出层分类函数对特征进行活性变化分类，输出关于活性变化的概率特征。

30. 根据权利要求 21 所述的方法，其特征在于，所述预测模型是分类器。

31. 根据权利要求 21 所述的方法，其特征在于，所述图神经网络是图卷积网络或图注意力网络。

32. 根据权利要求 28 所述的方法，其特征在于，所述决策树模型是 GBDT 模型。

33. 根据权利要求 32 所述的方法，其特征在于，使用 LightGBM 框架来实现 GBDT 模型。

34. 根据权利要求 21 所述的方法，其特征在于，所述的获取训练集的步骤包括：

获取公开数据库中的实验数据作为训练集；

通过对公开数据库中的实验数据做反转处理将训练集变为原来的两倍大小。

35. 一种用于预测突变对蛋白酶活性影响的系统，其特征在于，所述系统包括：

获取模块，用于获取待预测蛋白酶样本序列信息和三维结构信息；

处理模块，用于通过输入所述待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果，所述处理模块进一步包括如下子模块：

序列特征子模块，用于基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；

结构特征子模块，用于基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；

模型预测子模块，用于基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

36. 一种通过计算机实现的用于预测突变对蛋白酶活性影响的系统，所述计算机实现的系统通过执行计算机程序以实现以下操作：

获取待预测蛋白酶样本序列信息和三维结构信息；

通过输入待预测蛋白酶样本序列信息和三维结构信息，得到所述待预测蛋白酶样本突变后的活性变化预测结果，该操作进一步包括如下子操作：

基于所述待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得所述待预测蛋白酶样本的序列特征；

基于所述待预测蛋白酶样本的三维结构信息，获得所述待预测蛋白酶样本的结构特征；

基于所述序列特征和所述结构特征，使用基于图神经网络的预测模型进行预测，确定所述待预测蛋白酶样本突变后的活性变化预测结果。

37. 一种非瞬时性计算机可读存储介质，用于存储计算机程序，所述计算机程序包括指令，所述指令在由电子设备的处理器执行时使所述电子设备实施如权利要求 1-20 中任一项所述的用于预测突变对蛋白酶活性影响的方法或如权利要求 21-34 中任一项所述的用于预测突变对蛋白酶活性影响的预测模型的构建方法。

38. 一种计算机系统，所述计算机系统包括：

处理器；

存储器；和

计算机程序，其中，所述计算机程序存储在所述存储器中并且被配置为由所述处理器执行，所述计算机程序包括用于实施如权利要求 1-20 中任一项所述的用于预测突变对蛋白酶活性影响的方法或如权利要求 21-34 中任一项所述的用于预测突变对蛋白酶活性影响的预测模型的构建方法的指令。

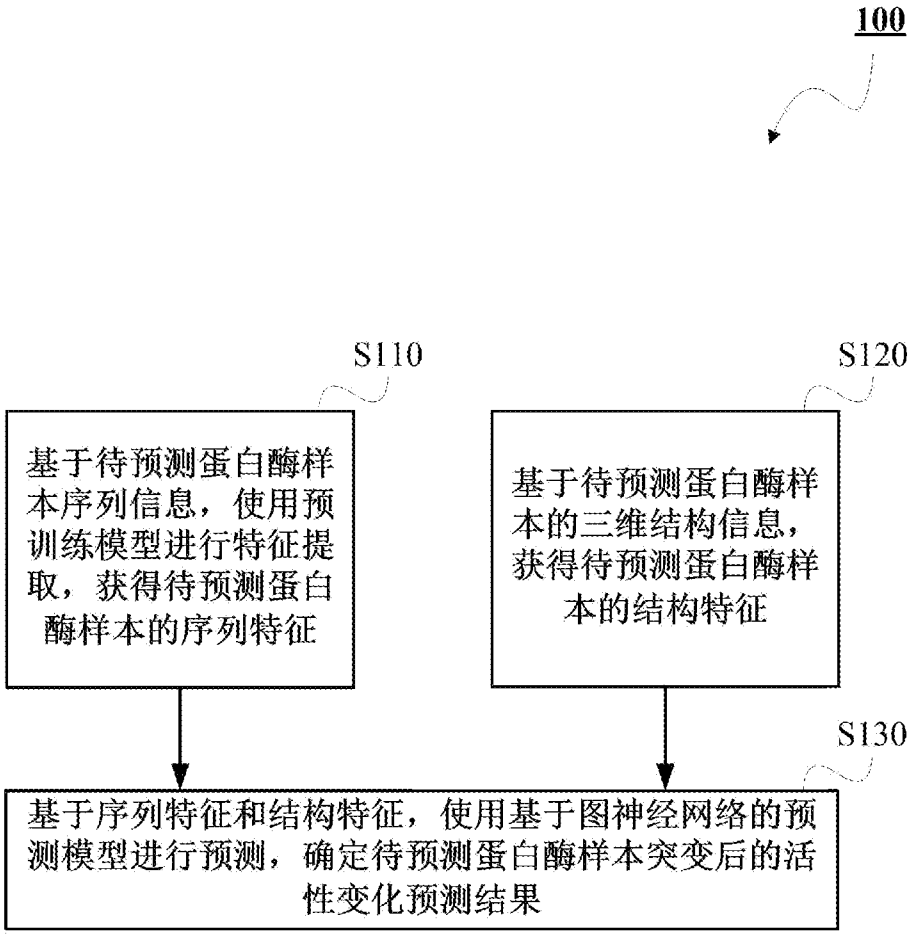

100
S110
基于待预测蛋白酶样本序列信息，使用预训练模型进行特征提取，获得待预测蛋白酶样本的序列特征
S120
基于待预测蛋白酶样本的三维结构信息，获得待预测蛋白酶样本的结构特征
S130
基于序列特征和结构特征，使用基于图神经网络的预测模型进行预测，确定待预测蛋白酶样本突变后的活性变化预测结果

图 1

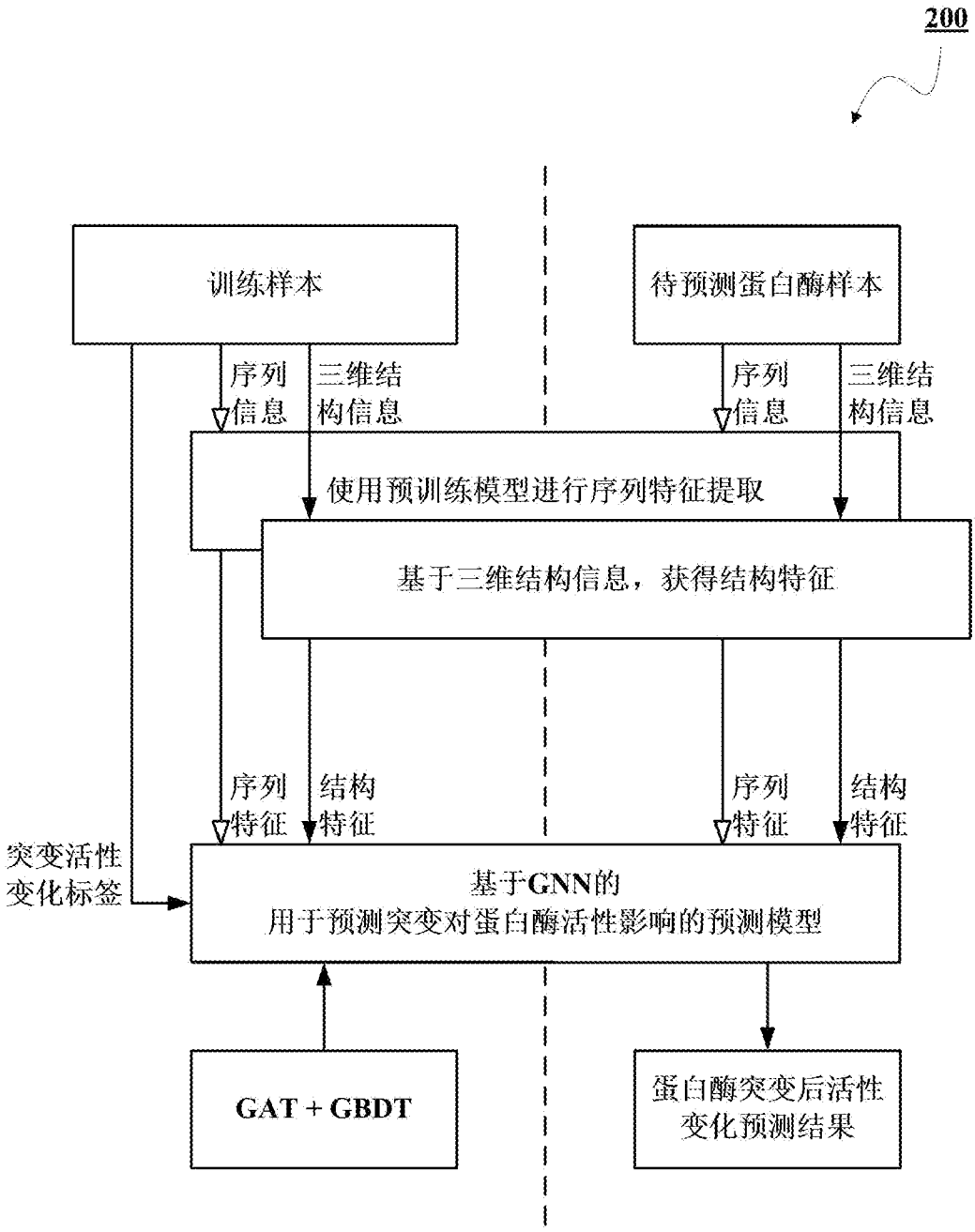

200
训练样本
待预测蛋白酶样本
序列信息
三维结构信息
序列信息
三维结构信息
使用预训练模型进行序列特征提取
基于三维结构信息，获得结构特征
序列特征
结构特征
序列特征
结构特征
突变活性变化标签
基于GNN的
用于预测突变对蛋白酶活性影响的预测模型
GAT + GBDT
蛋白酶突变后活性变化预测结果

图 2

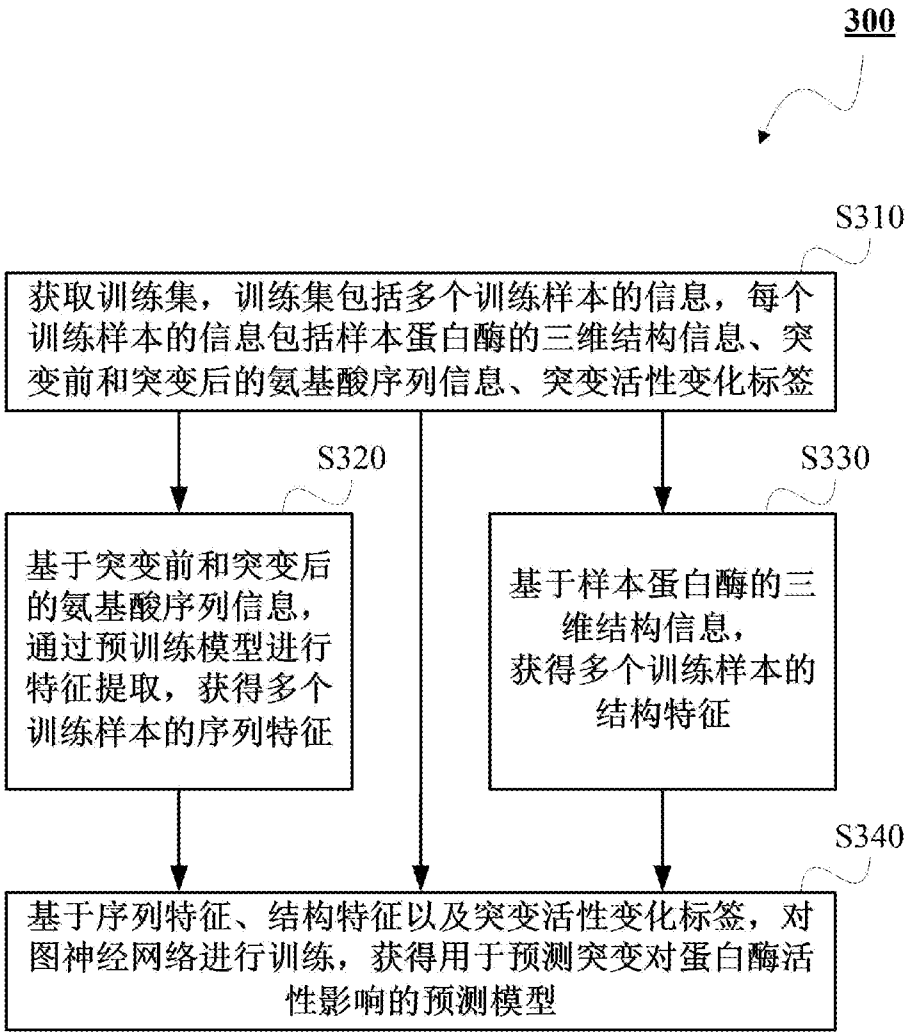
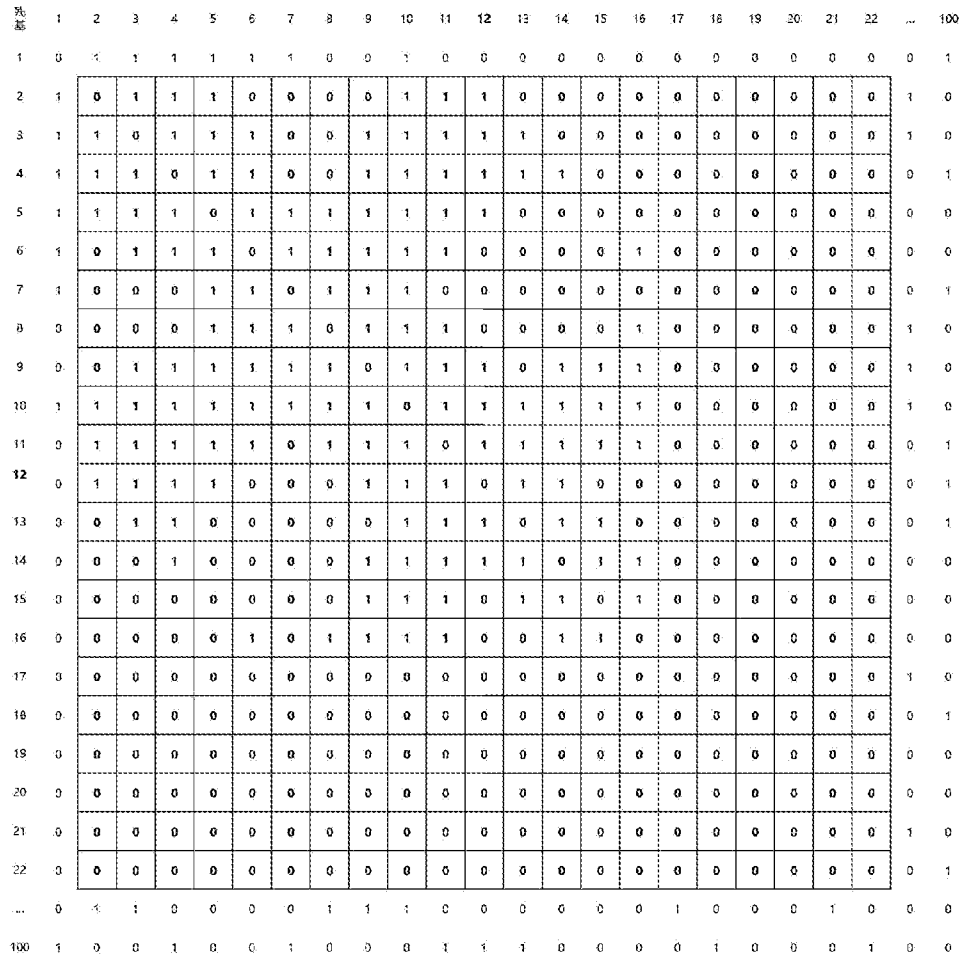

300
S310
获取训练集，训练集包括多个训练样本的信息，每个训练样本的信息包括样本蛋白酶的三维结构信息、突变前和突变后的氨基酸序列信息、突变活性变化标签
S320
基于突变前和突变后的氨基酸序列信息，通过预训练模型进行特征提取，获得多个训练样本的序列特征
S330
基于样本蛋白酶的三维结构信息，获得多个训练样本的结构特征
S340
基于序列特征、结构特征以及突变活性变化标签，对图神经网络进行训练，获得用于预测突变对蛋白酶活性影响的预测模型

图 3



残基	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	...	100
1	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0
3	1	1	0	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0
4	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1
5	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	1	1	1	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	1	1	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0
9	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0	0	1	0
10	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	1	0
11	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1
12	0	1	1	1	1	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	1
13	0	0	1	1	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	1
14	0	0	0	1	0	0	0	0	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0
16	0	0	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
...	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
100	1	0	0	1	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0

图 4

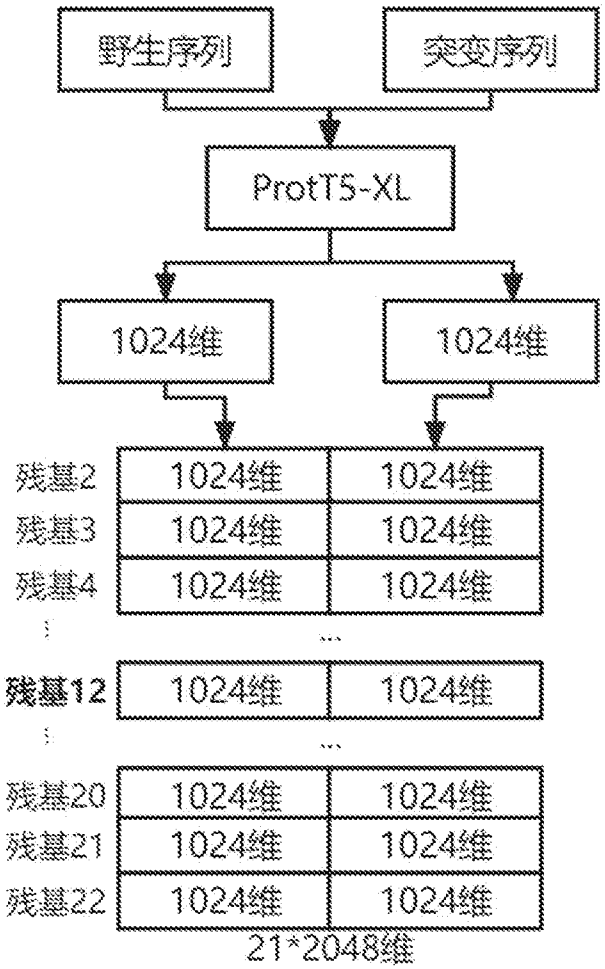

野生序列
突变序列
ProtT5-XL
1024维
1024维
残基2
残基3
残基4
残基12
残基20
残基21
残基22
1024维
21*2048维

图 5

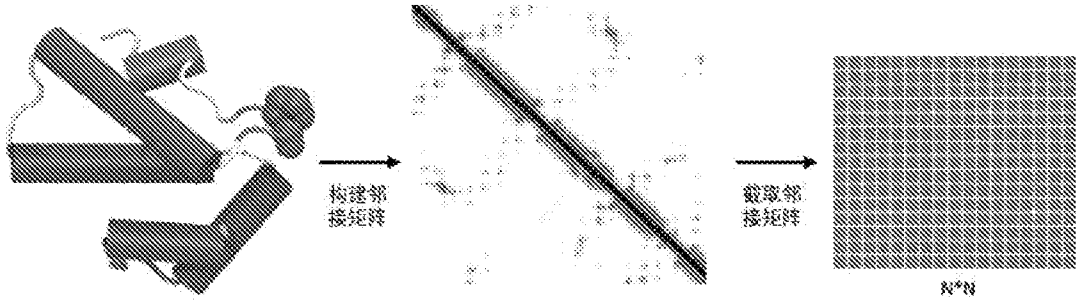

构建邻接矩阵
截取邻接矩阵
N*N

图 6

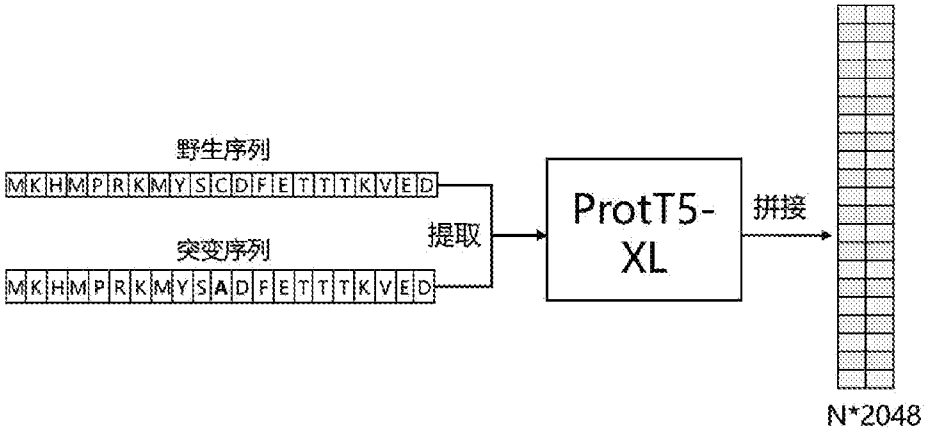

野生序列
M K H M P R K M Y S C D F E T T T K V E D
突变序列
M K H M P R K M Y S A D F E T T T K V E D
提取
ProtT5-XL
拼接
N*2048

图 7

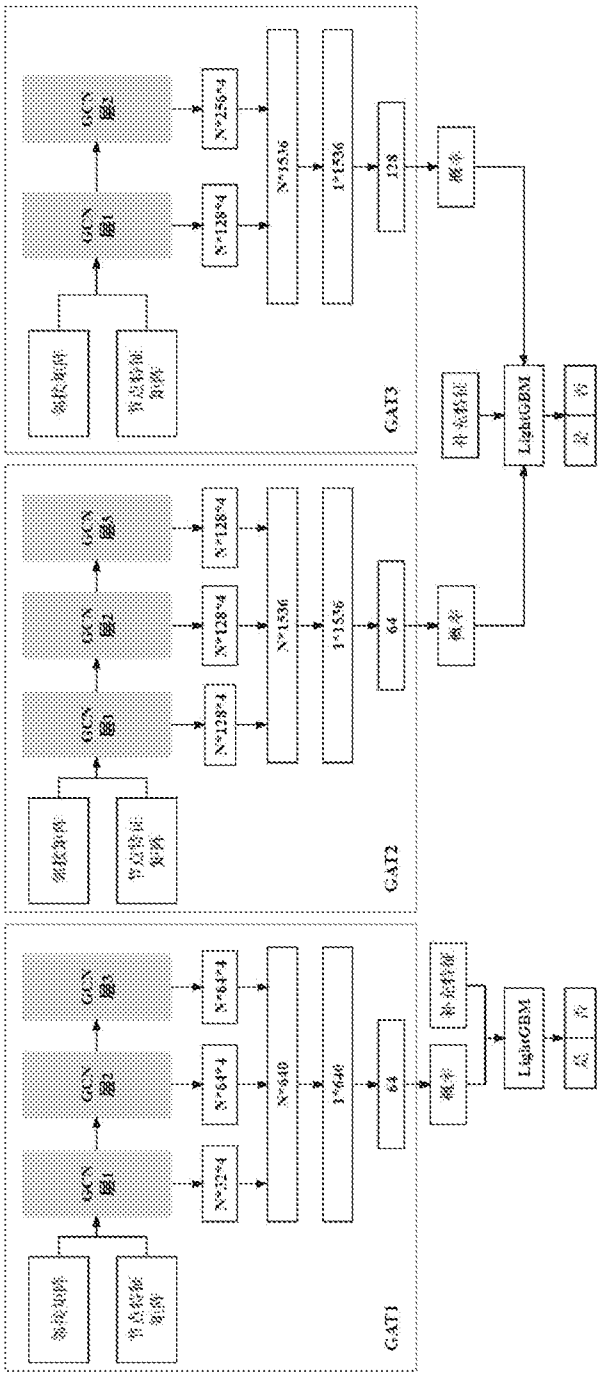


图 8

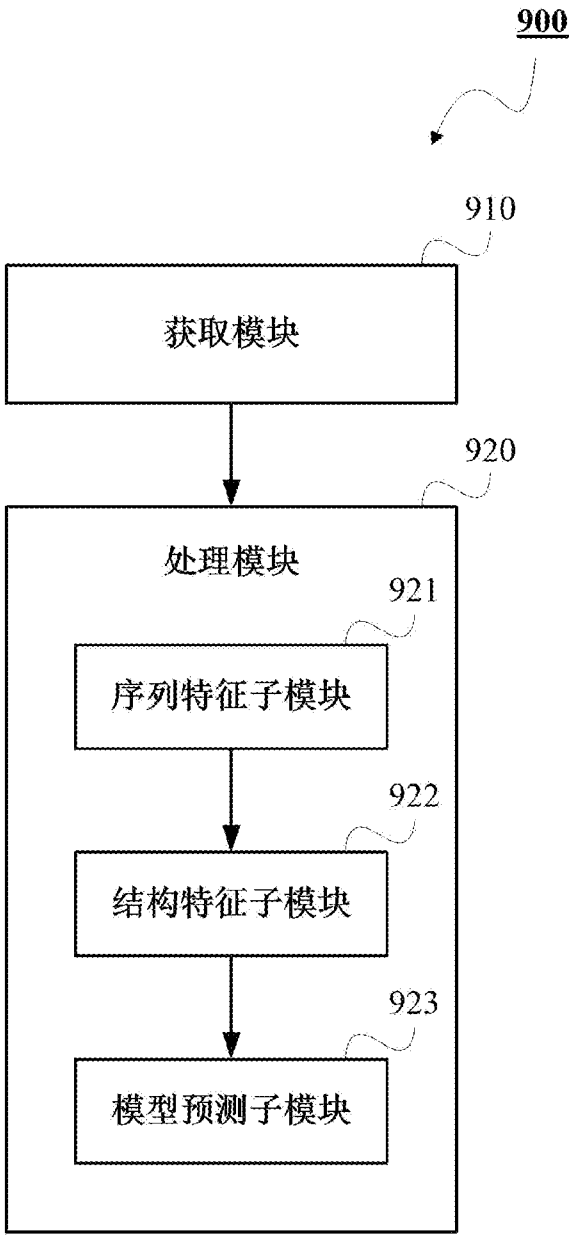


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/112259

A. CLASSIFICATION OF SUBJECT MATTER

G16B20/50(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC:G16B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, DWPI, ENTXTC, CJFD, CNKI, GOOGLE SCHOLAR, BING: 蛋白, 蛋白酶, 活性, 序列, 三维, 结构, 特征, 图神经网络, 预测, protein, protease, activity, sequence, tertiary, 3D, structure, feature, graph neural network, GNN.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110993037 A (ZHEJIANG UNIVERSITY OF TECHNOLOGY) 10 April 2020 (2020-04-10) description, paragraphs 0019-0051	1-38
A	CN 109872771 A (CODEXIS, INC.) 11 June 2019 (2019-06-11) entire document	1-38
A	CN 115197925 A (SHANGHAI YINTAI INFORMATION TECHNOLOGY CO., LTD.) 18 October 2022 (2022-10-18) entire document	1-38
A	CN 115512785 A (OCEAN UNIVERSITY OF CHINA) 23 December 2022 (2022-12-23) entire document	1-38
A	CN 115631786 A (INSTITUTE OF INFORMATION ON TRADITIONAL CHINESE MEDICINE, CACMS) 20 January 2023 (2023-01-20) entire document	1-38
A	KR 102284532 B1 (STANDIGM INC.) 03 August 2021 (2021-08-03) entire document	1-38

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 October 2024	**27 October 2024**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/112259

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2022122692 A1 (FLAGSHIP PIONEERING INNOVATIONS VI LLC) 21 April 2022 (2022-04-21) entire document	1-38

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/112259

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 110993037 A	10 April 2020	None	
CN 109872771 A	11 June 2019	US 2015134315 A1	14 May 2015
		AU 2014324670 A1	24 March 2016
		AU 2014324670 B2	21 November 2019
		IL 244458 A0	21 April 2016
		IL 244458 B	30 April 2020
		DK 3049979 T3	17 February 2020
		BR 112016006284 A2	01 August 2017
		BR 112016006284 B1	26 July 2022
		HUE 048104 T2	28 May 2020
		RU 2016116261 A	01 November 2017
		RU 2016116261 A3	02 March 2018
		RU 2694321 C2	11 July 2019
		ES 2774965 T3	23 July 2020
		KR 20160062079 A	01 June 2016
		KR 102341026 B1	21 December 2021
		WO 2015048573 A1	02 April 2015
		EP 3049979 A1	03 August 2016
		EP 3049979 B1	01 January 2020
		JP 2016537699 A	01 December 2016
		JP 6309086 B2	11 April 2018
		US 2022238179 A1	28 July 2022
		US 2020020415 A1	16 January 2020
		US 11342046 B2	24 May 2022
		CA 2923758 A1	02 April 2015
		CA 2923758 C	30 August 2022
		CN 105814573 A	27 July 2016
		IN 201647013558 A	31 August 2016
		SG 11201601692 A1	28 April 2016
		CN 105814573 A	27 July 2016
		CN 105814573 B	29 March 2019
		SG 11201601692 B	11 August 2017
		NZ 717647 A2	26 June 2020
CN 115197925 A	18 October 2022	None	
CN 115512785 A	23 December 2022	None	
CN 115631786 A	20 January 2023	None	
KR 102284532 B1	03 August 2021	None	
US 2022122692 A1	21 April 2022	IL 285402 A	30 September 2021
		EP 3924971 A1	22 December 2021
		KR 20210125523 A	18 October 2021
		JP 2022521686 A	12 April 2022
		JP 7492524 B2	29 May 2024
		CA 3127965 A1	20 August 2020
		WO 2020167667 A1	20 August 2020
		IN 202147040243 A	10 September 2021
		CN 113412519 A	17 September 2021
		HK 40066804 A0	26 August 2022
		CN 113412519 B	21 May 2024

国际检索报告

国际申请号
PCT/CN2024/112259

A. 主题的分类

G16B20/50(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC:G16B

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT,DWPI,ENTXTC,CJFD,CNKI,GOOGLE SCHOLAR,BING:蛋白,蛋白酶,活性,序列,三维,结构,特征,图神经网络,预测,protein,protease,activity,sequence,tertiary,3D,structure,feature,graph neural network,GNN.

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 110993037 A (浙江工业大学) 2020年4月10日 (2020 - 04 - 10) 说明书第0019-0051段	1-38
A	CN 109872771 A (科德克希思公司) 2019年6月11日 (2019 - 06 - 11) 全文	1-38
A	CN 115197925 A (上海茵肽信息科技有限公司) 2022年10月18日 (2022 - 10 - 18) 全文	1-38
A	CN 115512785 A (中国海洋大学) 2022年12月23日 (2022 - 12 - 23) 全文	1-38
A	CN 115631786 A (中国中医科学院中医药信息研究所) 2023年1月20日 (2023 - 01 - 20) 全文	1-38
A	KR 102284532 B1 (STANDIGM INC) 2021年8月3日 (2021 - 08 - 03) 全文	1-38
A	US 2022122692 A1 (FLAGSHIP PIONEERING INNOVATIONS VI LLC) 2022年4月21日 (2022 - 04 - 21) 全文	1-38

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2024年10月10日	2024年10月27日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 石松婷 电话号码 (+86) 027-59371858

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/112259

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110993037	A	2020年4月10日	无			
CN	109872771	A	2019年6月11日	US	2015134315	A1	2015年5月14日
				AU	2014324670	A1	2016年3月24日
				AU	2014324670	B2	2019年11月21日
				IL	244458	A0	2016年4月21日
				IL	244458	B	2020年4月30日
				DK	3049979	T3	2020年2月17日
				BR	112016006284	A2	2017年8月1日
				BR	112016006284	B1	2022年7月26日
				HUE	048104	T2	2020年5月28日
				RU	2016116261	A	2017年11月1日
				RU	2016116261	A3	2018年3月2日
				RU	2694321	C2	2019年7月11日
				ES	2774965	T3	2020年7月23日
				KR	20160062079	A	2016年6月1日
				KR	102341026	B1	2021年12月21日
				WO	2015048573	A1	2015年4月2日
				EP	3049979	A1	2016年8月3日
				EP	3049979	B1	2020年1月1日
				JP	2016537699	A	2016年12月1日
				JP	6309086	B2	2018年4月11日
				US	2022238179	A1	2022年7月28日
				US	2020020415	A1	2020年1月16日
				US	11342046	B2	2022年5月24日
				CA	2923758	A1	2015年4月2日
				CA	2923758	C	2022年8月30日
				CN	105814573	A	2016年7月27日
				IN	201647013558	A	2016年8月31日
				SG	11201601692	A1	2016年4月28日
				CN	105814573	A	2016年7月27日
				CN	105814573	B	2019年3月29日
				SG	11201601692	B	2017年8月11日
				NZ	717647	A2	2020年6月26日
CN	115197925	A	2022年10月18日	无			
CN	115512785	A	2022年12月23日	无			
CN	115631786	A	2023年1月20日	无			
KR	102284532	B1	2021年8月3日	无			
US	2022122692	A1	2022年4月21日	IL	285402	A	2021年9月30日
				EP	3924971	A1	2021年12月22日
				KR	20210125523	A	2021年10月18日
				JP	2022521686	A	2022年4月12日
				JP	7492524	B2	2024年5月29日
				CA	3127965	A1	2020年8月20日
				WO	2020167667	A1	2020年8月20日
				IN	202147040243	A	2021年9月10日
				CN	113412519	A	2021年9月17日
				HK	40066804	A0	2022年8月26日
				CN	113412519	B	2024年5月21日