

(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
2024 年 11 月 28 日 (28.11.2024)



(10) 国际公布号
WO 2024/239598 A1

- (51) 国际专利分类号:
G16B 30/00 (2019.01) **G16B 40/20** (2019.01)
- (21) 国际申请号: PCT/CN2023/137165
- (22) 国际申请日: 2023 年 12 月 7 日 (07.12.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310588503.4 2023年5月22日 (22.05.2023) CN
- (71) 申请人: 深圳先进技术研究院 (SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY) [CN/CN]; 中国广东省深圳市南山区西丽大学城学苑大道1068号, Guangdong 518055 (CN)。
- (72) 发明人: 余函 (YU, Han); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 罗小舟 (LUO, Xiaozhou); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。
- (74) 代理人: 北京中巡通大知识产权代理有限公司 (BEIJING ZHONG XUN TONG DA

INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市西城区西直门外大街 1 号院 2 号楼 7 层 7C1, Beijing 100044 (CN)。

- (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,

(54) Title: PROTEIN IDENTIFICATION METHOD, ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) 发明名称: 蛋白质鉴定方法、电子设备及存储介质

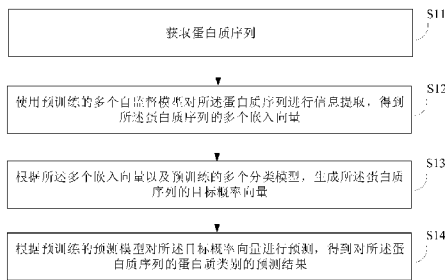


图 2

S11 ACQUIRE A PROTEIN SEQUENCE
S12 USE A PLURALITY OF PRE-TRAINED SELF-SUPERVISED MODELS TO PERFORM INFORMATION EXTRACTION ON THE PROTEIN SEQUENCE, SO AS TO OBTAIN A PLURALITY OF EMBEDDED VECTORS OF THE PROTEIN SEQUENCE
S13 ACCORDING TO THE PLURALITY OF EMBEDDED VECTORS AND A PLURALITY OF PRE-TRAINED CLASSIFICATION MODELS, GENERATE A TARGET PROBABILITY VECTOR OF THE PROTEIN SEQUENCE
S14 ACCORDING TO A PRE-TRAINED PREDICTION MODEL, PREDICT THE TARGET PROBABILITY VECTOR TO OBTAIN A PREDICTION RESULT OF A PROTEIN CATEGORY OF THE PROTEIN SEQUENCE

(57) Abstract: Provided are a protein identification method, an electronic device and a storage medium. The method comprises: acquiring a protein sequence; using a plurality of pre-trained self-supervised models to perform information extraction on the protein sequence, so as to obtain a plurality of embedded vectors of the protein sequence; according to the plurality of embedded vectors and a plurality of pre-trained classification models, generating a target probability vector of the protein sequence; and, according to a pre-trained prediction model, predicting the target probability vector to obtain a prediction result of a protein category of the protein sequence. Using the described method can improve the efficiency of identifying protein sequences.

(57) 摘要: 提供一种蛋白质鉴定方法、电子设备及存储介质。所述方法包括: 获取蛋白质序列; 使用预训练的多个自监督模型对所述蛋白质序列进行信息提取, 得到所述蛋白质序列的多个嵌入向量; 根据所述多个嵌入向量以及预训练的多个分类模型, 生成所述蛋白质序列的目标概率向量; 根据预训练的预测模型对所述目标概率向量进行预测, 得到对所述蛋白质序列的蛋白类别的预测结果。利用上述方法, 能够提高蛋白质序列的鉴定效率。

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))。

蛋白质鉴定方法、电子设备及存储介质

技术领域

本申请涉及生物领域，尤其涉及一种蛋白质鉴定方法、电子设备及存储介质。

背景技术

目前的蛋白质鉴定方案中，依赖于人工设计的复杂特征提取蛋白质序列的关键信息，然后根据提取的关键信息实现蛋白质序列的鉴定。但是人工设计的特征往往表示能力有限，导致提取到的关键信息难以充分地反映蕴含在蛋白质序列内部的信息，造成对蛋白质序列的鉴定效率较低。

发明内容

鉴于以上内容，有必要提供一种蛋白质鉴定方法、电子设备及存储介质，解决了由于人工设计的特征在表达能力上的局限性导致蛋白质序列的鉴定效率较低的技术问题。

一方面，本申请提供一种蛋白质鉴定方法，所述方法包括：获取蛋白质序列，使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量，根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量，根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。

由上述技术方案可知，本申请采用多个自监督模型对蛋白质序列进行信息提取，由于所述多个蛋白质模型为多个不同类型网络架构的自监督模型，因此，能够确保提取到的多个嵌入式向量为所述蛋白质序列中不同级别的信息。此外，本申请依靠预训练的多个自监督模型从蛋白质序列中提取信息，并不依赖于人工设计的特征从蛋白质序列中提取信息，因此能够解决的人工提取的特征的表达能力有限的技术问题，还能够提高信息提取的速度。由于本申请是将每个嵌入向量输入至每个分类模型中，因此可以得到每个嵌入向量在每个分类模型上的输出结果，将所述多个嵌入向量在所述多个分类模型上的输出结果进行拼接，得到所述目标概率向量，所述目标概率向量中融合了所述多个自监督模型以及所述多个分类模型对所述蛋白质序列进行提取的信息，因此所述目标概率向量能够充分地反映蕴含在所述蛋白质序列内部的信息，对所述目标概率向量进行预测，能够提高对蛋白质序列的鉴定效率。

在一些实施例中，所述方法还包括：获取数据集，对所述数据集进行划分，得到第一训练集以及第二训练集，其中，所述第一训练集包括多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签，所述第二训练集包括多个第二蛋白质序列以及每个第二蛋白质序列对应的第二类别标签，所述第一类别

标签以及所述第二类别标签均包括预设蛋白质类别，获取所述多个分类模型对应的多个初始分类器以及所述预测模型对应的预测学习器，根据所述多个自监督模型中的任意一个模型、所述多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签对每个初始分类器进行训练，得到每个初始分类器对应的分类模型，将任一自监督模型以及任一分类模型进行组合，得到多对组合模型，根据每对组合模型对每个第二蛋白质序列进行信息提取，得到每个第二蛋白质序列对应的训练概率向量，使用所述预测学习器对每个训练概率向量进行预测，得到每个第二蛋白质序列的训练预测结果，根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型。

在一些实施例中，所述根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型包括：根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值，根据所述训练损失值对所述预测学习器进行参数调整，直至所述训练损失值满足预设条件，得到所述预测模型。

在一些实施例中，所述根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值包括：根据每个训练预测结果对应的第二类别标签，确定每个训练预测结果是否正确，根据正确的训练预测结果以及所述第二类别标签的标签数量，计算所述训练损失值。

在一些实施例中，所述根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量包括：将每个嵌入向量输入至所述多个分类模型中，得到每个嵌入向量的多个初始概率向量，将所述多个嵌入向量的所有初始概率向量进行拼接，得到所述目标概率向量。

在一些实施例中，所述蛋白质类别包括预设蛋白质类别和非预设蛋白质类别，所述根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果，包括：对所述预设蛋白质类别和所述非预设蛋白质类别分别进行编码，得到类别向量，基于所述类别向量对所述目标概率向量进行分类，确定所述蛋白质序列属于所述预设蛋白质类别的预测概率，根据所述预测概率与预设阈值的比较结果确定所述预测结果。

在一些实施例中，所述根据所述预测概率与预设阈值的比较结果确定所述预测结果包括：若所述预测概率大于或者等于预设阈值，确定所述预测结果为所述蛋白质序列属于所述预设蛋白质类别，或者，若所述预测概率小于所述预设阈值，确定所述预测结果为所述蛋白质序列不属于所述预设蛋白质类别。

在一些实施例中，所述预设蛋白质类别包括嗜热蛋白和嗜冷蛋白。

另一方面，本申请提供一种蛋白质鉴定装置，运行于电子设备，所述装置包括：获取单元，用于获取蛋白质序列，提取单元，用于使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量，生成单元，用于根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量，预测单元，用于根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。

另一方面，本申请提供一种电子设备，所述电子设备包括：存储器，存储至少一个指令；及处理器，执行所述至少一个指令以实现所述的蛋白质鉴定方法。

另一方面，本申请提供一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个指令，所述至少一个指令被电子设备中的处理器执行以实现所述的蛋白质鉴定方法。

附图说明

图 1 是本申请一实施例提供的电子设备的结构图。

图 2 是本申请一实施例提供的蛋白质鉴定方法的流程图。

图 3 是本申请一实施例提供的预测结果的生成方法的流程图。

图 4 是本申请一实施例提供的多个自监督模型、多个分类模型以及预测模型的生成方法的流程图。

图 5 是本申请一实施例提供的蛋白质鉴定装置的功能模块图。

具体实施方式

需要说明的是，本申请中“至少一个”是指一个或者多个，“多个”是指两个或多于两个。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B 可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 的情况，其中 A，B 可以是单数或者复数。本申请的说明书和权利要求书及附图中的术语“第一”、“第二”、“第三”、“第四”等（如果存在）是用于区别类似的对象，而不是用于描述特定的顺序或先后次序。

在本申请实施例中，“示例性的”或者“例如”等词用于表示作例子、例证或说明。本申请实施例中被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念。

嗜热蛋白，通常也被称为耐高温的蛋白，是一类在高温条件下仍然能够保持其结构和功能的蛋白质。嗜热蛋白质通常来自于嗜热菌，特别是古菌。由于嗜热蛋白在增强蛋白稳定性、增强酶活以及适用于高温工业过程的可行性，它们已经在生物技术、食品加工和制药等各个领域实现了广泛的应用。例如，来自嗜热菌的热稳定纤维素酶已被用于将木质纤维素生物质高效转化为生物燃料。然而，鉴定和筛选嗜热蛋白是一项费力且繁琐的过程。为了加快相关领域的发展和更好的理解嗜热蛋白的机制，开发高通量的筛选方法实现快速的嗜热蛋白的鉴定就至关重要。

很多相关计算方法被开发出来用于鉴定嗜热蛋白，然而，相关计算方法往往依赖于人工设计的复杂特征以提取蛋白质序列的关键信息，然后根据提取的关键信息实现对蛋白质序列的鉴定。但是人工设计的特征往往表示能力有限，导致提取到的关键信息难以充分地反映蕴含在蛋白质序列内部的信息，造成对

蛋白质序列的鉴定效率较低。

为了解决上述问题，本申请提供一种蛋白质鉴定方法、电子设备及存储介质。为了使本申请的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本申请进行详细描述。

本申请实施例提供的蛋白质鉴定方法可应用于一个或者多个电子设备中。

如图 1 所示，是本申请一实施例提供的电子设备的结构图。在图 1 中，电子设备 1 包括，但不限于，存储器 12、处理器 13，以及存储在存储器 12 中并可在处理器 13 上运行的计算机程序，例如蛋白质鉴定程序。

所述电子设备 1 是一种能够按照事先设定或存储的指令，自动进行参数值计算和/或 信息处理的设备，其硬件包括，但不限于：微处理器、专用集成电路 (Application Specific Integrated Circuit, ASIC)、 可编程门阵列 (Field-Programmable Gate Array, FPGA)、 数字信号处理器 (Digital Signal Processor, DSP)、 嵌入式设备等。

所述电子设备 1 可以是任何一种可与用户进行人机交互的电子产品，例如，个人计算机、服务器、平板电脑、智能手机等。

所述电子设备 1 还可以包括网络设备和/或用户设备。其中，所述网络设备包括，但不限于单个网络服务器、多个网络服务器组成的服务器组或基于云计算 (Cloud Computing) 的由大量主机或网络服务器构成的云。图 1 仅仅是电子设备 1 的示例，并不构成对电子设备 1 的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如电子设备 1 还可以包括输入输出设备、网络接入设备、总线等。

所述电子设备 1 所处的网络包括但不限于：互联网、广域网、城域网、局域网、虚拟专用网络 (Virtual Private Network, VPN) 等。

如图 2 所示，是本申请一实施例提供的蛋白质鉴定方法的流程图。根据不同的需求，该流程图中各个步骤的顺序可以根据实际要求进行调整，某些步骤可以省略。所述方法的执行主体为电子设备，例如图 1 所示的电子设备 1。

S11，获取蛋白质序列。

在一实施例中，所述蛋白质序列是指需要鉴定类别的蛋白质序列。所述电子设备可以从数据集中获取所述蛋白质序列。例如，所述数据集可以为 Uniprot 数据库、BRENDA 数据库以及常见的包含蛋白质序列的数据库中的部分数据或者所有数据。

在一实施例中，所述蛋白质序列可以为需要鉴定类别的数据集中的序列。

在另一实施例中，所述蛋白质序列可以通过其它方式获取，本申请对所述蛋白质序列不作限制。

S12，使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量。

在一实施例中，所述多个自监督模型包括，但不限于：SeqVec 模型、ProtCNN 模型、ProtTrans 模型和 CPCProt 模型。其中，所述多个自监督模型为预训练至收敛的模型。

在另一实施例中，所述多个自监督模型还可以为其它模型，本申请对所述多个自监督模型不作限制。

在一些实施例中，所述电子设备使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量包括：所述电子设备将每个嵌入向量输入至每个自监督模型中，得到每个嵌入向量在每个自监督模型上对应输出的嵌入向量。

其中，每个蛋白质模型对所述蛋白质序列的信息提取过程可参考相关技术中该模型对蛋白质序列进行表示学习，得到蛋白质嵌入式表示的过程。

在本实施例中，每个嵌入向量的维度以及信息量等信息与所述嵌入向量对应的自监督模型相关。例如，若所述多个蛋白质模型为所述 SeqVec 模型、所述 ProtCNN 模型、所述 ProtTrans 模型和所述 CPCProt 模型，所述多个嵌入向量为 4 个向量，所述 SeqVec 模型提取到的嵌入向量的维度为 1024 维、所述 ProtCNN 模型提取到的嵌入向量的维度为 1100 维、所述 ProtTrans 模型提取到的嵌入向量的维度为 1024 维以及所述 CPCProt 模型提取到的嵌入向量的维度为 1536 维。

在本实施例中，使用多个自监督模型对所述蛋白质序列进行特征提取，而并不依赖人工提取的特征从蛋白质序列中提取信息，因此能够解决的人工提取的特征的表达能力有限的技术问题，还能够提高信息的提取速度。由于每个自监督模型的架构不相同，因此能够提取到所述蛋白质序列中不同级别以及不同维度的嵌入向量。

在本申请的其它实施例中，为了使得所述蛋白质序列具有更多的嵌入表示，所述多个自监督模型的数量可以更多。

S13，根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量。

在一些实施例中，所述多个分类模型包括，但不限于：随机森林模型（Random Forest，RF）、自适应增强模型（AdaBoost，AB）、引导聚集模型或者装袋模型（Bagging，BA）、梯度提升决策树模型 XGBoost（eXtremeGradient Boosting，XGB）和（LightGBM，LGB）等。

在一实施例中，所述电子设备根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量包括：所述电子设备将每个嵌入向量输入至所述多个分类模型中，得到每个嵌入向量的多个初始概率向量，然后所述电子设备将所述多个嵌入向量的所有初始概率向量进行拼接，得到所述目标概率向量。

其中，由于所述多个分类模型为不同的模型，因此每个分类模型根据输入的每个嵌入向量生成每个嵌入向量的过程可参考相关技术中每个分类模型的分方法或者分类公式。

例如，承接上述实施例，若所述多个嵌入向量为所述 SeqVec 模型、所述 ProtCNN 模型、所述 ProtTrans 模型和所述 CPCProt 模型输出的 4 个嵌入向量，所述多个分类模型为所述 AdaBoost 模型、所述 Bagging 模型、所述 XGBoost 模型和所述 LightGBM 模型，将所述 4 个嵌入向量分别输入至 5 个分类模型中，

得到 20 个初始概率向量，若每个初始概率向量的维度均为 1 维，将 20 个 1 维的初始概率向量进行拼接，得到维度为 20 维的目标概率向量。

在本申请的另一实施例中，所述多个分类模型可以为其它模型，本申请对所述不作限制。上述对所述多个分类模型的举例仅为示例，而并不构成对所述多个分类模型的限制。

在本实施例中，可以对所述多个分类模型进行级联，通过由所述多个分类模型集成的模型对每个嵌入向量进行分类，得到所述目标概率向量，能够减少避免单一模型的过拟合问题。此外，由于所述目标概率向量是由多个所述初始概率向量拼接得到，所述目标概率向量融合了所述多个初始概率向量中的信息，使得所述目标概率向量能够充分地反映蕴含在所述蛋白质序列内部的信息。

S14，根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。

在一实施例中，所述蛋白质类别包括预设蛋白质类别和非预设蛋白质类别。其中，所述预设蛋白质类别包括，但不限于：嗜热蛋白和嗜冷蛋白。所述非预设蛋白质类别为除了所述预设蛋白质类别之外的其它蛋白质类别。例如，若所述预设蛋白质类别为嗜热蛋白，则所述非预设蛋白质类别可以为非嗜热蛋白，或者，若所述预设蛋白质类别为嗜冷蛋白，则所述非预设蛋白质类别可以为非嗜冷蛋白。

在一些实施例中，所述预测结果包括所述蛋白质序列属于所述预设蛋白质类别和所述蛋白质序列不属于所述预设蛋白质类别。

在一些实施例中，所述预测模型可以为机器学习模型。例如，所述预测模型可以为支持向量机模型（Support Vector Machines, SVM）。

在本实施例中，由于所述目标概率向量能够充分地反映蕴含在所述蛋白质序列内部的信息，因此对所述目标概率向量进行预测，能够提高对蛋白质序列的鉴定准确性。

由上述技术方案可知，本申请采用多个自监督模型对蛋白质序列进行信息提取，由于所述多个蛋白质模型为多个不同类型网络架构的自监督模型，因此，能够确保提取到的多个嵌入式向量为所述蛋白质序列中不同级别的信息。此外，本申请依靠预训练的多个自监督模型从蛋白质序列中提取信息，并不依赖于人工设计的特征从蛋白质序列中提取信息，因此能够解决的人工提取的特征的表达能力有限的技术问题，还能够提高信息提取的速度。由于本申请是将每个嵌入向量输入至每个分类模型中，因此可以得到每个嵌入向量在每个分类模型上的输出结果，将所述多个嵌入向量在所述多个分类模型上的输出结果进行拼接，得到所述目标概率向量，所述目标概率向量中融合了所述多个自监督模型以及所述多个分类模型对所述蛋白质序列进行提取的信息，因此所述目标概率向量能够充分地反映蕴含在所述蛋白质序列内部的信息，对所述目标概率向量进行预测，能够提高对蛋白质序列的鉴定效率。

在本申请的一些实施例中，所述电子设备将所述目标概率向量输入至所述预测模型中，得到对所述蛋白质序列的蛋白质类别的预测结果。如图 3 所示，

是本申请一实施例提供的预测结果的生成方法的流程图，包括以下步骤：

S141，所述电子设备对所述预设蛋白质类别和所述非预设蛋白质类别分别进行编码，得到类别向量。

在一些实施例中，所述预设蛋白质类别包括，但不限于：嗜热蛋白和嗜冷蛋白。所述非预设蛋白质类别为除了所述预设蛋白质类别之外的其它类别。例如，若所述预设蛋白质类别为嗜热蛋白，则所述非预设蛋白质类别可以为非嗜热蛋白，或者，若所述预设蛋白质类别为嗜冷蛋白，则所述非预设蛋白质类别可以为非嗜冷蛋白。

在一些实施例中，所述电子设备可以对所述预设蛋白质类别和所述非预设蛋白质类别分别进行独热编码（one-hot），得到所述类别向量。例如，所述电子设备可以将所述预设蛋白质类别编码为 1，并将所述非预设蛋白质类别编码为 0，得到由 1 和 0 组成的类别向量。

在本申请的其它实施例中，所述电子设备可以通过其它方式对所述预设蛋白质类别和所述非预设蛋白质类别进行编码，本申请对编码的方式不作限制。

S142，所述电子设备基于所述类别向量对所述目标概率向量进行分类，确定所述蛋白质序列属于所述预设蛋白质类别的预测概率。

在一些实施例中，所述电子设备基于所述类别向量对所述目标概率向量进行分类，确定所述蛋白质序列属于所述预设蛋白质类别的预测概率的方法可参考相关技术中 sigmoid 函数或者 softmax 函数对应的分类公式。

S143，所述电子设备根据所述预测概率与预设阈值的比较结果确定所述预测结果。

在一些实施例中，所述电子设备根据所述预测概率与预设阈值的比较结果确定所述预测结果，包括：若所述预测概率大于或者等于预设阈值，所述电子设备确定所述预测结果为所述蛋白质序列属于所述预设蛋白质类别，或者，若所述预测概率小于所述预设阈值，所述电子设备确定所述预测结果为所述蛋白质序列不属于所述预设蛋白质类别，也可以是所述蛋白质序列属于所述非预设蛋白质类别。

在本申请的一实施例中，若所述电子设备确定所述预测结果为所述蛋白质序列属于所述预设蛋白质类别，所述电子设备输出第一预设数据以表示所述蛋白质序列属于所述预设蛋白质类别，或者，若所述电子设备确定所述预测结果为所述蛋白质序列不属于所述预设蛋白质类别（所述电子设备确定所述预测结果为所述蛋白质序列属于所述非预设蛋白质类别），所述电子设备输出第二预设数据以表示所述蛋白质序列不属于所述预设蛋白质类别。

其中，所述第一预设数据与所述第二预设数据可以为数值、字母或者数值和字母的组合等形式。所述第一预设数据以及所述第二预设数据可以自行设置，本申请对此不作限制。例如，在所述第一预设数据与所述第二预设数据均为数值时，所述第一预设数值可以为 1，所述第二预设数值可以为 0。

在本申请的一些实施例中，在使用所述多个自监督模型、所述多个分类模型以及所述预测模型之前，需要生成所述多个自监督模型、所述多个分类模型

以及所述预测模型。如图 4 所示，是本申请一实施例提供的多个自监督模型、多个分类模型以及预测模型的生成方法的流程图，包括以下步骤：

S21，所述电子设备获取数据集，对所述数据集进行划分，得到第一训练集以及第二训练集。

在一些实施例中，所述数据集包括多个蛋白质序列，所述数据集中的每个蛋白质序列具有对应的类别标签，所述数据集可以包括，但不限于：Uniprot 数据库、BRENDA 数据库以及常见的包含蛋白质序列的数据库中的部分数据或者所有数据。其中，所述第一训练集包括多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签，所述第二训练集包括多个第二蛋白质序列以及每个第二蛋白质序列对应的第二类别标签，所述第一类别标签以及所述第二类别标签均包括预设蛋白质类别。

在一些实施例中，每个第一蛋白质序列具有对应的第一类别标签，每个第二蛋白质具有对应的第二类别标签，所述第一类别标签用于指示每个第一蛋白质序列的类别，所述第二类别标签用于指示每个第二蛋白质序列的类别。

在本实施例中，所述电子设备对所述数据集进行划分的方式可以自行设置，本申请对此不作限制。例如，所述电子设备可以将所述数据集中 80% 的数据作为所述第一训练集，将所述 80% 的数据的类别标签作为所述第一类别标签，然后所述电子设备将所述数据集中 20% 的数据作为所述第二训练集，将所述 20% 的数据的类别标签作为所述第二类别标签。上述举例仅为对所述数据集的一种划分方式的举例，而不构成对所述数据集的划分方式的限制。

S22，所述电子设备获取所述多个分类模型对应的多个初始分类器以及所述预测模型对应的预测学习器。

在一些实施例中，所述多个初始分类器可以包括所述随机森林模型 Random Forest 对应的分类器、所述自适应增强模型 AdaBoost 对应的分类器、所述装袋模型 Bagging 对应的分类器、所述梯度提升决策树模型 XGBoost 对应的分类器和所述梯度提升决策树模型 LightGBM 对应的分类器，所述预测学习器可以为支持向量机模型 SVM 的学习器。

S23，所述电子设备根据所述多个自监督模型中的任意一个模型、所述多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签对每个初始分类器进行训练，得到每个初始分类器对应的分类模型。

在一些实施例中，所述电子设备将每个第一蛋白质序列输入至每个自监督模型中，得到每个第一蛋白质序列在每个自监督模型上对应输出的初始嵌入向量，其中，每个第一蛋白质序列对应的多个初始嵌入向量与该第一蛋白质序列的第一类别标签相对应，然后所述电子设备将每个第一蛋白质序列对应的多个初始嵌入向量分别输入至所述多个初始分类器中，得到每个第一蛋白质序列的每个初始嵌入向量在每个初始分类器上对应输出的分类向量，将每个第一蛋白质序列对应的多个分类向量进行拼接，得到每个第一蛋白质序列对应的拼接概率向量，所述电子设备对每个第一蛋白质序列对应的拼接概率向量进行预测，得到每个第一蛋白质序列的预测类别，根据所述多个第一蛋白质的预测类别与

第一类别标签计算初始损失值，并根据所述初始损失值调整所述多个初始分类器的参数，直至所述初始损失值满足配置条件，得到所述初始分类器对应的分类模型。

在本实施例中，所述多个初始嵌入向量与上文中的多个嵌入向量的生成方式基本相同，所述拼接概率向量与上文中的目标概率向量的生成方式基本相同，故本申请不再重复描述。此外，所述电子设备对每个拼接概率向量进行预测可参考下文中对目标概率向量的预测过程，所述初始损失值的计算方式可参考下文中的训练损失值的计算方式，所述配置条件与下文中的预设条件基本相同。

S24，所述电子设备将任一自监督模型以及任一分类模型进行组合，得到多对组合模型。

例如，若所述多个自监督模型分别以 A、B、C 和 D 进行表示，所述多个分类模型分别以 a、b、c、d 和 e 进行表示。所述电子设备将任一自监督模型以及任一分类模型进行组合，得到的多对组合模型如表 1 所示。

表 1 多对组合模型示例

	自监督模型 A	自监督模型 B	自监督模型 C	自监督模型 D
分类模型 a	第 1 对组合模型 A+a	第 2 对组合模型 B+a	第 3 对组合模型 C+a	第 4 对组合模型 D+a
分类模型 b	第 5 对组合模型 A+b	第 6 对组合模型 B+b	第 7 对组合模型 C+b	第 8 对组合模型 D+b
分类模型 c	第 9 对组合模型 A+c	第 10 对组合模型 B+c	第 11 对组合模型 C+c	第 12 对组合模型 D+c
分类模型 d	第 13 对组合模型 A+d	第 14 对组合模型 B+d	第 15 对组合模型 C+d	第 16 对组合模型 D+d
分类模型 e	第 17 对组合模型 A+e	第 18 对组合模型 B+e	第 19 对组合模型 C+e	第 20 对组合模型 D+e

S25，所述电子设备根据每对组合模型对每个第二蛋白质序列进行信息提取，得到每个第二蛋白质序列对应的训练概率向量。

在一些实施例中，每个训练概率向量的生成方式与所述目标概率向量的生成方式基本相同，故本申请不再重复描述。

S26，所述电子设备使用所述预测学习器对每个训练概率向量进行预测，得到每个第二蛋白质序列的训练预测结果。

在一些实施例中，所述训练预测结果与所述蛋白质序列的预测结果的生成方式基本相同，故本申请在此不再重复描述。

S27，所述电子设备根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型。

在一些实施例中，所述电子设备根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型包括：所述电子设备根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值，根据所述训练损失值对所述预测学习器进行参数调整，直至所述训练损失值满足预设条件，得到所述预测模型。

其中，所述电子设备可以对所述多个自监督模型、所述多个初始分类器和

所述预测学习器的权重和偏置等参数进行调整。所述预设条件可以自行设置，本申请对此不作限制。例如，所述预设条件与所述训练预测结果相对应。例如，所述预设条件可以为所述训练损失值下降至一预设区间、所述训练损失值下降至最低或者所述训练损失值不再变化，其中，所述预设区间可以自行设置，本申请对此不作限制。

在一些实施例中，所述电子设备根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值包括：所述电子设备根据每个训练预测结果对应的第二类别标签，确定每个训练预测结果是否正确，然后所述电子设备根据正确的训练预测结果以及所述第二类别标签的标签数量，计算所述训练损失值。

其中，所述电子设备可以将每个训练预测结果与所述对应的第二类别标签进行比对，若每个训练预测结果指示的蛋白质与所述对应的第二类别标签相同，则所述电子设备确定每个训练预测结果正确，或者，若每个训练预测结果与所述对应的第二类别标签不相同，则所述电子设备确定每个训练预测结果错误。

例如，在任一第二蛋白质序列的训练预测结果为所述任一第二蛋白质序列属于所述预设蛋白质类别时，若所述任一第二蛋白质序列的第二类别标签为所述预设蛋白质类别，则所述电子设备确定任一第二蛋白质序列的训练预测结果正确。或者，在任一第二蛋白质序列的训练预测结果为所述任一第二蛋白质序列属于所述预设蛋白质类别时，若所述任一第二蛋白质序列的第二类别标签不是所述预设蛋白质类别，则所述电子设备确定任一第二蛋白质序列的训练预测结果错误。

在一实施例中，所述电子设备使用集成算法（例如 bagging 算法或者 boosting 算法）对所述多个初始分类器进行集成训练，得到由所述多个分类模型级联而成的模型。

在本实施例中，所述电子设备计算正确的训练预测结果的正确数量，并将所述正确数量与所述标签数量之间的比值确定为所述训练损失值。在所述训练损失值不再变化时，所述电子设备停止调整，得到所述多个自监督模型、所述多个分类模型以及所述预测模型。

在本申请的另外一些实施例中，所述训练损失值还可以通过其它方式进行计算。例如，所述电子设备可以将错误的训练预测结果的错误数量与所述标签数量之间的比值确定为所述训练损失值，在所述训练损失值不再下降或者下降至一预设区间，所述电子设备停止调整，得到所述多个自监督模型、所述多个分类模型以及所述预测模型。

如图 5 所示，是本申请一实施例提供的蛋白质鉴定装置的功能模块图。所述蛋白质鉴定装置 11 包括获取单元 110、提取单元 111、生成单元 112 以及预测单元 113。本申请所称的模块/单元是指一种能够被图 1 中的处理器 13 所获取，并且能够完成固定功能的一系列计算机可读指令段，其存储在图 1 中的存储器 12 中。在本实施例中，关于各模块/单元的功能将在后续的实施例中详述。

在一些实施例中，所述获取单元 110 用于获取蛋白质序列。

所述提取单元 111，用于使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量。

所述生成单元 112，用于根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量。

在一些实施例中，所述生成单元 112，还用于将每个嵌入向量输入至所述多个分类模型中，得到每个嵌入向量的多个初始概率向量，将所述多个嵌入向量的所有初始概率向量进行拼接，得到所述目标概率向量。

所述预测单元 113，用于根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。所述预设蛋白质类别包括嗜热蛋白和嗜冷蛋白。

在一些实施例中，所述蛋白质类别包括预设蛋白质类别和非预设蛋白质类别，所述预测单元 113，还用于根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果，包括：对所述预设蛋白质类别和所述非预设蛋白质类别分别进行编码，得到类别向量，基于所述类别向量对所述目标概率向量进行分类，确定所述蛋白质序列属于所述预设蛋白质类别的预测概率，根据所述预测概率与预设阈值的比较结果确定所述预测结果。

在一些实施例中，所述预测单元 113，还用于根据所述预测概率与预设阈值的比较结果确定所述预测结果包括：若所述预测概率大于或者等于预设阈值，确定所述预测结果为所述蛋白质序列属于所述预设蛋白质类别，或者，若所述预测概率小于所述预设阈值，确定所述预测结果为所述蛋白质序列不属于所述预设蛋白质类别。

在一些实施例中，所述生成单元 112，还用于获取数据集，对所述数据集进行划分，得到第一训练集以及第二训练集，其中，所述第一训练集包括多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签，所述第二训练集包括多个第二蛋白质序列以及每个第二蛋白质序列对应的第二类别标签，所述第一类别标签以及所述第二类别标签均包括预设蛋白质类别，获取所述多个分类模型对应的多个初始分类器以及所述预测模型对应的预测学习器，根据所述多个自监督模型中的任意一个模型、所述多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签对每个初始分类器进行训练，得到每个初始分类器对应的分类模型，将任一自监督模型以及任一分类模型进行组合，得到多对组合模型，根据每对组合模型对每个第二蛋白质序列进行信息提取，得到每个第二蛋白质序列对应的训练概率向量，使用所述预测学习器对每个训练概率向量进行预测，得到每个第二蛋白质序列的训练预测结果，根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型。

在一些实施例中，所述生成单元 112，还用于根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型包括：根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值，根据所述训练损失值对所述预测学习器进行参数调整，直至所述训练损失值满足预

设条件，得到所述预测模型。

在一些实施例中，所述生成单元 112，还用于根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值包括：根据每个训练预测结果对应的第二类别标签，确定每个训练预测结果是否正确，根据正确的训练预测结果以及所述第二类别标签的标签数量，计算所述训练损失值。

由上述技术方案可知，本申请采用多个自监督模型对蛋白质序列进行信息提取，由于所述多个蛋白质模型为多个不同类型网络架构的自监督模型，因此，能够确保提取到的多个嵌入式向量为所述蛋白质序列中不同级别的信息。此外，本申请依靠预训练的多个自监督模型从蛋白质序列中提取信息，并不依赖于人工设计的特征从蛋白质序列中提取信息，因此能够解决的人工提取的特征的表达能力有限的技术问题，还能够提高信息提取的速度。由于本申请是将每个嵌入向量输入至每个分类模型中，因此可以得到每个嵌入向量在每个分类模型上的输出结果，将所述多个嵌入向量在所述多个分类模型上的输出结果进行拼接，得到所述目标概率向量，所述目标概率向量中融合了所述多个自监督模型以及所述多个分类模型对所述蛋白质序列进行提取的信息，因此所述目标概率向量能够充分地反映蕴含在所述蛋白质序列内部的信息，对所述目标概率向量进行预测，能够提高对蛋白质序列的鉴定效率。

在一实施例中，接续前文对图 1 中的电子设备的描述，处理器 13 可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现场可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等，处理器 13 是电子设备 1 的运算核心和控制中心，利用各种接口和线路连接整个电子设备 1 的各个部分，及获取电子设备 1 的操作系统以及安装的各类应用程序、程序代码等。

处理器 13 获取电子设备 1 的操作系统以及安装的各类应用程序。处理器 13 获取应用程序以实现上述各个蛋白质鉴定方法实施例中的步骤，例如图 2、图 3 及图 4 所示的步骤。

示例性的，计算机程序可以被分割成一个或多个模块/单元，例如加速单元，一个或者多个模块/单元被存储在存储器 12 中，并由处理器 13 获取，以完成本申请。一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述计算机程序在电子设备 1 中的获取过程。

存储器 12 可用于存储计算机程序和/或模块，处理器 13 通过运行或获取存储在存储器 12 内的计算机程序和/或模块，以及调用存储在存储器 12 内的数据，实现电子设备 1 的各种功能。存储器 12 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据服务器的使用所创建的数据等。此外，存储器 12 可以包括非易失性存储器，例如硬盘、内存、插接

式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）、至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。

存储器 12 可以是电子设备 1 的外部存储器和/或内部存储器。进一步地，存储器 12 可以是具有实物形式的存储器，如内存条、TF 卡（Trans-flash Card）等等。

电子设备 1 集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机程序来指令相关的硬件来完成，的计算机程序可存储于一计算机可读存储介质中，该计算机程序在被处理器获取时，可实现上述各个方法实施例的步骤。

其中，计算机程序包括计算机程序代码，计算机程序代码可以为源代码形式、对象代码形式、可获取文件或某些中间形式等。计算机可读介质可以包括：能够携带计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（Read-Only Memory, ROM）。

结合图 2，电子设备 1 中的存储器 12 存储多个指令以实现一种蛋白质鉴定方法，处理器 13 可获取多个指令从而实现：获取蛋白质序列；使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量；根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量；根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。

具体地，处理器 13 对上述指令的具体实现方法可参考图 2 对应实施例中相关步骤的描述，在此不赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以处于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选取其中的部分或者全部模块来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本申请的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本申请内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。本申请中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一、第二等词语用来表示名称，而并不表示任何特定的顺序。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。

权 利 要 求 书

1. 一种蛋白质鉴定方法，其特征在于，所述方法包括：

获取蛋白质序列；

使用预训练的多个自监督模型对所述蛋白质序列进行信息提取，得到所述蛋白质序列的多个嵌入向量；

根据所述多个嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量；

根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果。

2. 如权利要求 1 所述的蛋白质鉴定方法，其特征在于，所述方法还包括：

获取数据集，对所述数据集进行划分，得到第一训练集以及第二训练集，其中，所述第一训练集包括多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签，所述第二训练集包括多个第二蛋白质序列以及每个第二蛋白质序列对应的第二类别标签，所述第一类别标签以及所述第二类别标签均包括预设蛋白质类别；

获取所述多个分类模型对应的多个初始分类器以及所述预测模型对应的预测学习器；

根据所述多个自监督模型中的任意一个模型、所述多个第一蛋白质序列以及每个第一蛋白质序列对应的第一类别标签对每个初始分类器进行训练，得到每个初始分类器对应的分类模型；

将任一自监督模型以及任一分类模型进行组合，得到多对组合模型；

根据每对组合模型对每个第二蛋白质序列进行信息提取，得到每个第二蛋白质序列对应的训练概率向量；

使用所述预测学习器对每个训练概率向量进行预测，得到每个第二蛋白质序列的训练预测结果；

根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型。

3. 如权利要求 2 所述的蛋白质鉴定方法，其特征在于，所述根据所述训练预测结果、所述第二类别标签和所述预测学习器，生成所述预测模型包括：

根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值；

根据所述训练损失值对所述预测学习器进行参数调整，直至所述训练损失值满足预设条件，得到所述预测模型。

4. 如权利要求 2 所述的蛋白质鉴定方法，其特征在于，所述根据多个所述训练预测结果以及每个训练预测结果对应的第二类别标签计算训练损失值包括：

根据每个训练预测结果对应的第二类别标签，确定每个训练预测结果是否正确；

根据正确的训练预测结果以及所述第二类别标签的标签数量，计算所述训练损失值。

5. 如权利要求 1 所述的蛋白质鉴定方法，其特征在于，所述根据所述多个

嵌入向量以及预训练的多个分类模型，生成所述蛋白质序列的目标概率向量包括：

将每个嵌入向量输入至所述多个分类模型中，得到每个嵌入向量的多个初始概率向量；

将所述多个嵌入向量的所有初始概率向量进行拼接，得到所述目标概率向量。

6. 如权利要求 1 所述的蛋白质鉴定方法，其特征在于，所述蛋白质类别包括预设蛋白质类别和非预设蛋白质类别，所述根据预训练的预测模型对所述目标概率向量进行预测，得到对所述蛋白质序列的蛋白质类别的预测结果，包括：

对所述预设蛋白质类别和所述非预设蛋白质类别分别进行编码，得到类别向量；

基于所述类别向量对所述目标概率向量进行分类，确定所述蛋白质序列属于所述预设蛋白质类别的预测概率；

根据所述预测概率与预设阈值的比较结果确定所述预测结果。

7. 如权利要求 6 所述的蛋白质鉴定方法，其特征在于，所述根据所述预测概率与预设阈值的比较结果确定所述预测结果包括：

若所述预测概率大于或者等于预设阈值，确定所述预测结果为所述蛋白质序列属于所述预设蛋白质类别；或者

若所述预测概率小于所述预设阈值，确定所述预测结果为所述蛋白质序列不属于所述预设蛋白质类别。

8. 如权利要求 6 所述的蛋白质鉴定方法，其特征在于，所述预设蛋白质类别包括嗜热蛋白和嗜冷蛋白。

9. 一种电子设备，其特征在于，所述电子设备包括：

存储器，存储至少一个指令；及

处理器，执行所述至少一个指令以实现如权利要求 1 至 8 中任意一项所述的蛋白质鉴定方法。

10. 一种计算机可读存储介质，其特征在于：所述计算机可读存储介质中存储有至少一个指令，所述至少一个指令被电子设备中的处理器执行以实现如权利要求 1 至 8 中任意一项所述的蛋白质鉴定方法。

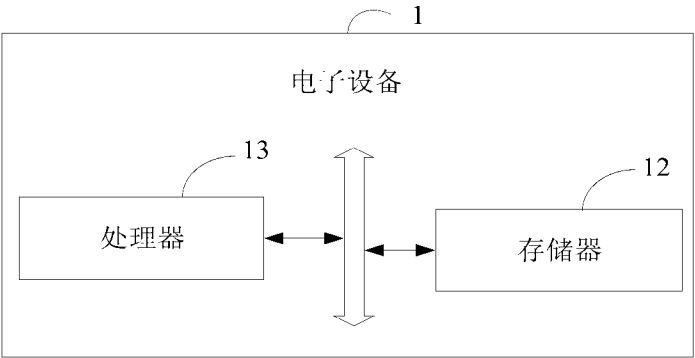


图 1

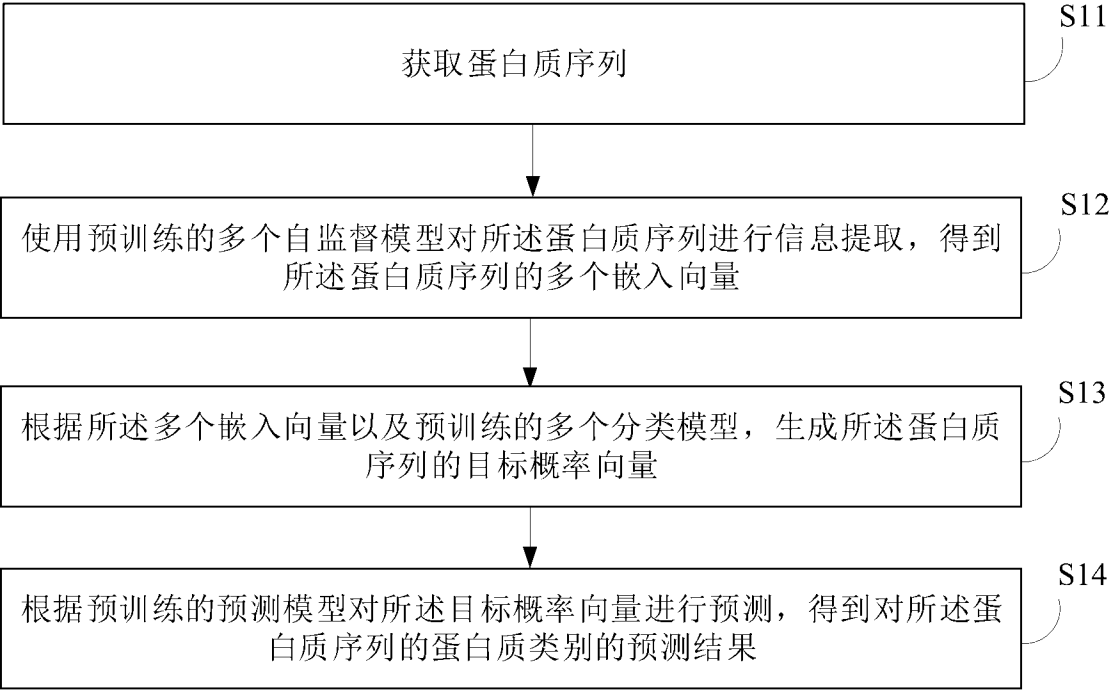


图 2

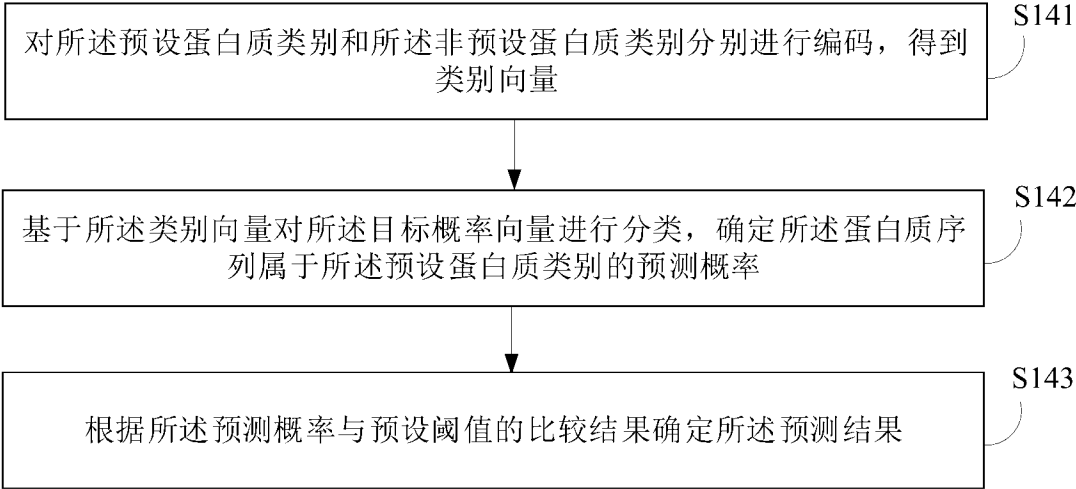


图 3

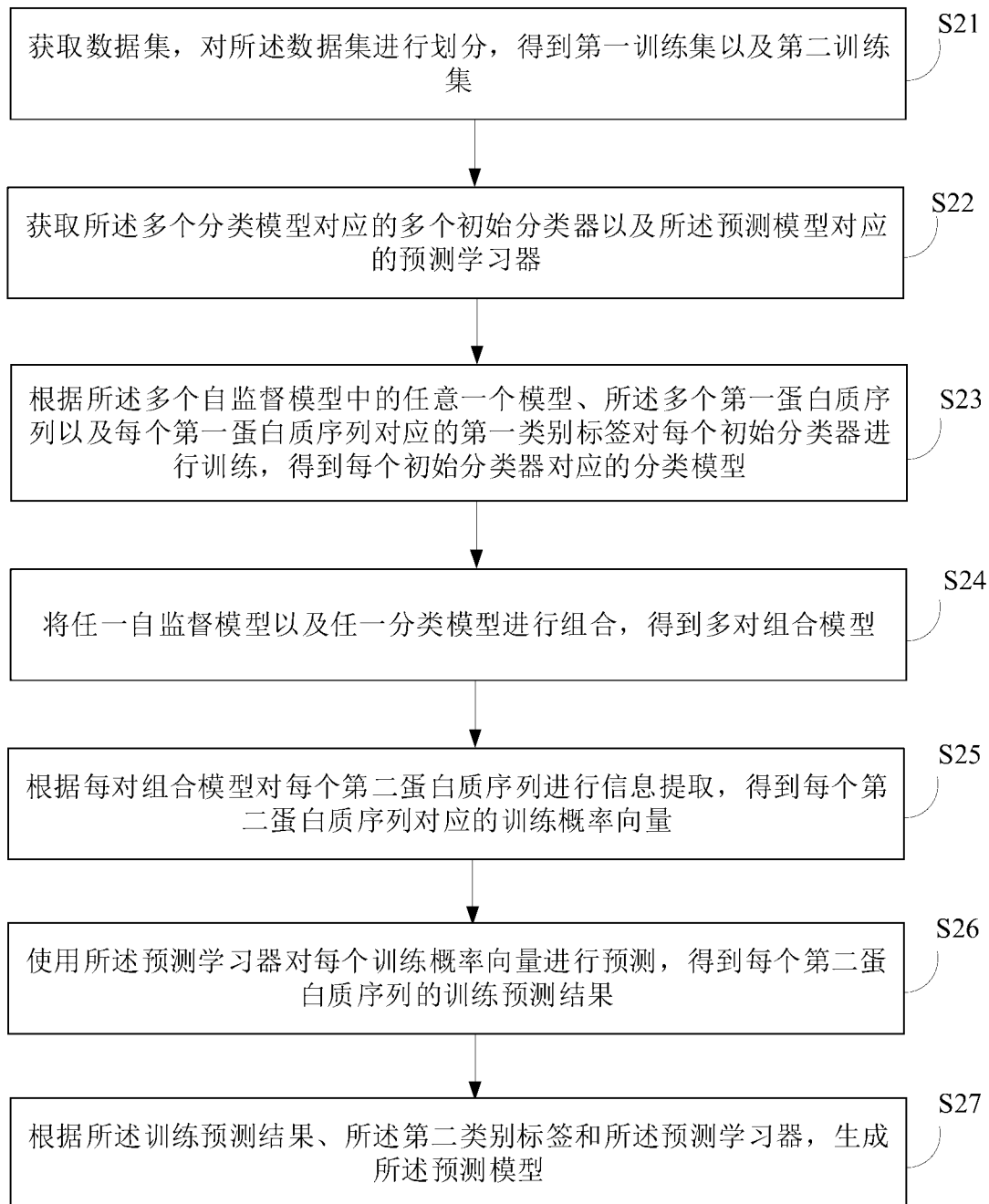


图 4

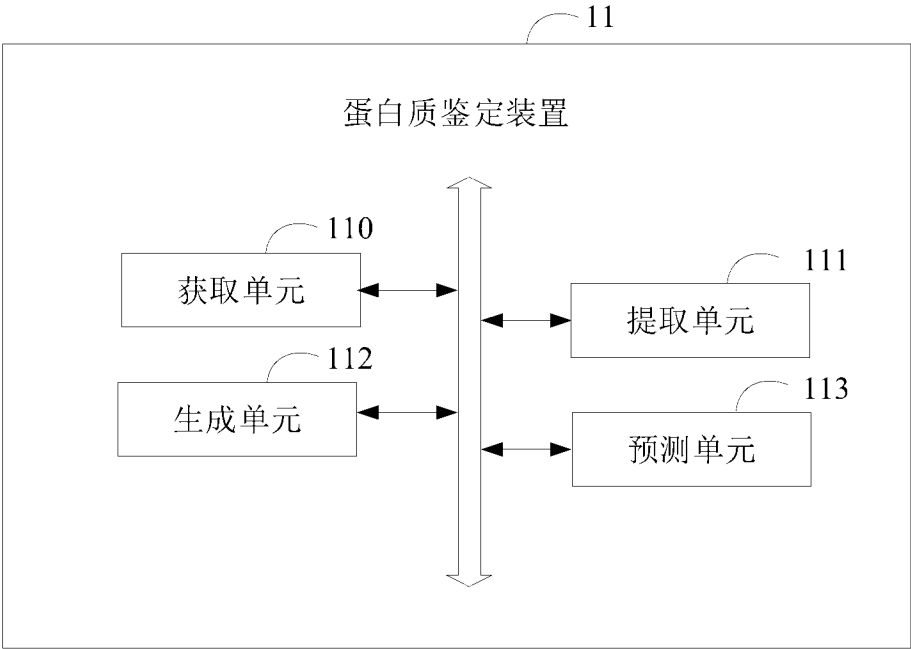


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/137165

A. CLASSIFICATION OF SUBJECT MATTER

G16B 30/00(2019.01)i; G16B 40/20(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)

CNXTX, WPABSC, WPABS, ENTXTC, ENTXT, 超星读秀学术, DUXIU ACADEMIC, CNKI, 万方, WANFANG, ISI web of knowledge, Elsevier Science Direct, Springerlink, pubmed, 百度, BAIDU, 百度学术, BAIDU SCHOLAR, patentics: 蛋白质, 多肽, 鉴定, 自监督, 嵌入向量, 分类, 预测, 概率向量, SeqVec模型, ProtCNN模型, ProtTrans模型, CPCProt模型, 随机森林模型, Random Forest, 自适应增强模型, AdaBoost, 引导聚集模型, 装袋模型, Bagging, 梯度提升决策树模型, XGBoost, XGB, LGB, 机器学习模型, 支持向量机模型, Support Vector Machines, SVM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116646009 A (SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY) 25 August 2023 (2023-08-25) claims 1-10	1-10
X	US 2022165356 A1 (NE47 BIO, INC.) 26 May 2022 (2022-05-26) claims 1-20	1-10
A	CN 112420123 A (SHANGHAI SENSETIME INTELLIGENT TECHNOLOGY CO., LTD.) 26 February 2021 (2021-02-26) claims 1-11	1-10
A	WO 2022185179 A1 (GLAXOSMITHKLINE BIOLOGICALS S.A.) 09 September 2022 (2022-09-09) claims 1-31	1-10
A	CN 116013428 A (CENTRAL SOUTH UNIVERSITY) 25 April 2023 (2023-04-25) claims 1-10	1-10

☒ 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

29 March 2024

Date of mailing of the international search report

09 April 2024

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/137165

C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 114333982 A (BEIJING BAIDU NETCOM SCIENCE & TECHNOLOGY CO., LTD.) 12 April 2022 (2022-04-12) claims 1-17	1-10
A	吴疆等 (WU, Jiang et al.). "半监督学习算法拉普拉斯支持向量机应用于蛋白质结构类预测 (Protein Structural Classes Prediction by Using Laplace Support Vector Machine and Based on Semi-supervised Method)" <i>微型电脑应用 (Microcomputer Applications)</i> , Vol. 36, No. 8, 31 December 2020 (2020-12-31), pages 5-8	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/137165

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	116646009	A	25 August 2023	None			
US	2022165356	A1	26 May 2022	US	2023123770	A1	20 April 2023
				EP	4248446	A1	27 September 2023
				US	11532378	B2	20 December 2022
				WO	2022109462	A1	27 May 2022
CN	112420123	A	26 February 2021	None			
WO	2022185179	A1	09 September 2022	EP	4302300	A1	10 January 2024
CN	116013428	A	25 April 2023	None			
CN	114333982	A	12 April 2022	None			

A. 主题的分类

G16B 30/00(2019.01)i; G16B 40/20(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))

CNTEXT, WPABSC, WPABS, ENTXTC, ENTXT, 超星读秀学术, CNKI, 万方, ISI web of knowledge, Elsevier Science Direct, Springerlink, pubmed, 百度, 百度学术, patentics: 蛋白质、多肽、鉴定、自监督、嵌入向量、分类、预测、概率向量、SeqVec模型、ProtCNN模型、ProtTrans模型、CPCProt模型、随机森林模型、Random Forest、自适应增强模型、AdaBoost、引导聚集模型、装袋模型、Bagging、梯度提升决策树模型、XGBoost、XGB、LGB、机器学习模型、支持向量机模型、Support Vector Machines、SVM

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
PX	CN 116646009 A (深圳先进技术研究院) 2023年8月25日 (2023 - 08 - 25) 权利要求1-10	1-10
X	US 2022165356 A1 (NE47 BIO. INC.) 2022年5月26日 (2022 - 05 - 26) 权利要求1-20	1-10
A	CN 112420123 A (上海商汤智能科技有限公司) 2021年2月26日 (2021 - 02 - 26) 权利要求1-11	1-10
A	WO 2022185179 A1 (GLAXOSMITHKLINE BIOLOGICALS SA) 2022年9月9日 (2022 - 09 - 09) 权利要求1-31	1-10
A	CN 116013428 A (中南大学) 2023年4月25日 (2023 - 04 - 25) 权利要求1-10	1-10
A	CN 114333982 A (北京百度网讯科技有限公司) 2022年4月12日 (2022 - 04 - 12) 权利要求1-17	1-10

☒ 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

2024年3月29日

国际检索报告邮寄日期

2024年4月9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

授权官员

王慧

电话号码 (+86) 010-53962103

PCT/ISA/210 表(第2页) (2022年7月)

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	吴疆等. "半监督学习算法拉普拉斯支持向量机应用于蛋白质结构类预测" 微型电脑应用, 第36卷，第8期，2020年12月31日 (2020 - 12 - 31)， 第5-8页	1-10

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/137165

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	116646009	A	2023年8月25日	无	
US	2022165356	A1	2022年5月26日	US	2023123770 A1 2023年4月20日
				EP	4248446 A1 2023年9月27日
				US	11532378 B2 2022年12月20日
				WO	2022109462 A1 2022年5月27日
CN	112420123	A	2021年2月26日	无	
WO	2022185179	A1	2022年9月9日	EP	4302300 A1 2024年1月10日
CN	116013428	A	2023年4月25日	无	
CN	114333982	A	2022年4月12日	无	