

(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 12 月 19 日 (19.12.2024)



(10) 国际公布号  
**WO 2024/254739 A1**

- (51) 国际专利分类号:  
**G16B 15/20** (2019.01)
- (21) 国际申请号: PCT/CN2023/099729
- (22) 国际申请日: 2023 年 6 月 12 日 (12.06.2023)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 深圳华大生命科学研究院 (BGI SHENZHEN) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。
- (72) 发明人: 苏子栋 (SU, Zidong); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。赵子裕 (ZHAO, Ziyu); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。郭斐 (GUO, Fei); 中国广东省深圳市盐田区

北山工业区综合楼, Guangdong 518083 (CN)。黎宇翔 (LI, Yuxiang); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。董宇亮 (DONG, Yuliang); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

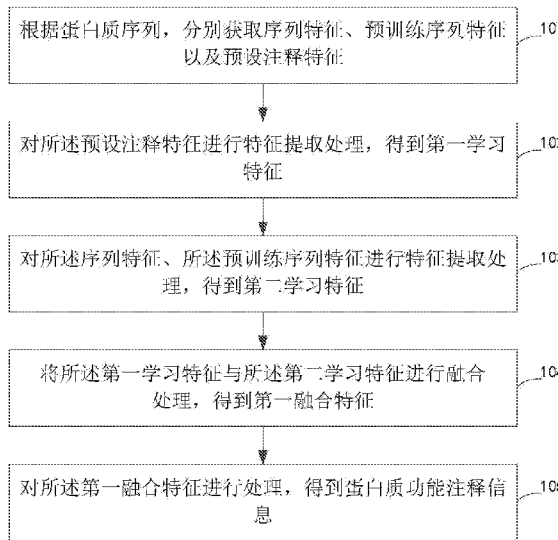
(74) 代理人: 北京法胜知识产权代理有限公司 (FASHENG INTELLECTUAL PROPERTY COMPANY, LTD.); 中国北京市海淀区悦秀路99号1单元4层401, Beijing 100085 (CN)。

(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,

(54) Title: PROTEIN FUNCTIONAL ANNOTATION METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 蛋白质功能注释的方法及装置、电子设备和存储介质

[图 1]



- 101 Separately acquire a sequence feature, a pre-trained sequence feature, and a preset annotation feature on the basis of a protein sequence
- 102 Perform feature extraction on the preset annotation feature to obtain a first learning feature
- 103 Perform feature extraction on the sequence feature and the pre-trained sequence feature to obtain a second learning feature
- 104 Fuse the first learning feature and the second learning feature to obtain a first fused feature
- 105 Process the first fused feature to obtain protein functional annotation information

(57) Abstract: The present disclosure provides a protein functional annotation method and apparatus, an electronic device, and a storage medium. The method comprises: separately acquiring a sequence feature, a pre-trained sequence feature, and a preset annotation feature on the basis of a protein sequence; performing feature extraction on the preset annotation feature to obtain a first learning feature; performing feature extraction on the sequence feature and the pre-trained sequence feature to obtain a second learning feature; fusing the first learning feature and the second learning feature to obtain a first fused feature; and processing the first fused feature to obtain protein functional annotation information. Compared with the related art, the present disclosure relates to acquiring a sequence feature,

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))。

a pre-trained sequence feature, and a preset annotation feature on the basis of a protein sequence, and fusing the three different aspects of features, such that protein functional annotation information obtained is more accurate.

(57) 摘要: 本公开提供了一种蛋白质功能注释的方法及装置、电子设备和存储介质, 根据蛋白质序列, 分别获取序列特征、预训练序列特征以及预设注释特征; 对预设注释特征进行特征提取处理, 得到第一学习特征; 对序列特征、预训练序列特征进行特征提取处理, 得到第二学习特征; 将第一学习特征与第二学习特征进行融合处理, 得到第一融合特征; 对第一融合特征进行处理, 得到蛋白质功能注释信息。与相关技术相比, 本公开基于蛋白质序列获取序列特征、预训练序列特征以及预设注释特征, 通过对前述三种不同角度的特征的融合处理, 使得获得的蛋白质功能注释信息更加准确。

# 说明书

## 发明名称: 蛋白质功能注释的方法及装置、电子设备和存储介质 技术领域

[0001] 本公开涉及数据处理技术领域，尤其涉及一种蛋白质功能注释的方法及装置、电子设备和存储介质。

### 背景技术

[0002] 蛋白质中氨基酸的顺序决定了蛋白质的空间结构，蛋白质的结构又决定了蛋白质的在生物学上的功能。蛋白质功能注释的目的是用算法对蛋白质可能具有的功能进行预测。当前蛋白质功能注释算法主要分类三类，包括仅基于序列信息的算法，基于序列和结构信息的算法以及基于序列和交互网络的算法。

[0003] 由于蛋白质的结构信息和交互网络信息需要通过复杂实验获得，相对难以获取，而蛋白质的氨基酸序列信息是最易获得的蛋白质的信息，因此仅基于序列信息算法应用最为广泛。但是相对于其他使用了难以从序列中获得的额外信息的算法，仅基于序列信息的蛋白质功能注释算法的预测准确率较低。

[0004] 发明内容

[0005] 本公开提供了一种蛋白质功能注释的方法及装置、电子设备和存储介质。其主要目的在于解决仅基于序列信息的蛋白质功能注释算法的预测准确率较低的问题。

[0006] 根据本公开的第一方面，提供了一种蛋白质功能注释的方法，其中，包括：

[0007] 根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；

[0008] 对所述预设注释特征进行特征提取处理，得到第一学习特征；

[0009] 对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；

[0010] 将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；

[0011] 对所述第一融合特征进行处理，得到蛋白质功能注释信息。

[0012] 可选的，在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之后，所述方法还包括：

- [0013] 基于第一预设层数的全连接层，对所述预设注释特征进行特征提取处理，得到第三学习特征；
- [0014] 基于第二预设层数的卷积层，分别对所述序列特征、所述预训练序列特征进行特征提取处理，得到第四学习特征与第五学习特征；
- [0015] 将所述第三学习特征、所述第四学习特征以及所述第五学习特征进行融合处理，得到第二融合特征；其中，所述融合处理为加和处理；
- [0016] 基于第三预设层数的全连接层，对所述第二融合特征进行分类计算，得到蛋白质功能注释信息。
- [0017] 可选的，所述根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征包括：
- [0018] 基于预设编码算法，对所述蛋白质序列进行编码得到所述序列特征；
- [0019] 基于预设蛋白质语言模型，对所述序列特征进行预训练处理，得到所述预训练序列特征；
- [0020] 在预设蛋白质数据库中，查找与所述蛋白质序列对应的所述预设注释特征。
- [0021] 可选的，所述对所述预设注释特征进行特征提取处理，得到第一学习特征包括如下步骤：
- [0022] 将所述预设注释特征输入第一预设数量的残差全连接模块；
- [0023] 将第一个残差全连接模块的输出特征作为第二个残差全连接模块的输入特征进行特征提取处理；
- [0024] 按照如上步骤循环执行第一预定次数，以获取所述第一学习特征，其中，所述第一预定次数等于所述第一预设数量。
- [0025] 可选的，所述将所述预设注释特征输入所述残差全连接模块中的第一个残差全连接模块包括：
- [0026] 将所述预设注释特征输入归一化层进行归一化处理，得到归一化处理后的所述预设注释特征；
- [0027] 将所述归一化处理后的预设注释特征输入非线性层进行非线性映射，得到非线性映射后的预设注释特征；

- [0028] 将所述非线性映射后的预设注释特征输入全连接层进行特征提取，得到第一预设注释特征；
- [0029] 将所述第一预设注释特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二预设注释特征；
- [0030] 将所述预设注释特征与所述第二预设注释特征进行加和处理，得到第一输出特征。
- [0031] 可选的，基于第四预设层数的卷积层，对所述序列特征进行处理，得到第六学习特征、对所述预训练序列特征进行处理，得到第七学习特征；
- [0032] 将所述第六学习特征与所述第七学习特征进行加和处理，得到第三融合特征；
- [0033] 所述对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征包括如下步骤：
- [0034] 将所述第三融合特征输入第二预设数量的残差卷积模块；
- [0035] 将第一个所述残差卷积模块的输出特征作为第二个所述残差卷积模块的输入特征进行特征提取处理；
- [0036] 按照如上步骤循环执行第二预定次数，得到所述第二学习特征，其中，所述第二预定次数等于所述第二预设数量。
- [0037] 可选的，所述将所述第三融合特征输入所述残差卷积模块中的第一个残差卷积模块包括：
- [0038] 将所述第三融合特征输入归一化层进行归一化处理，得到归一化处理后的第三融合特征；
- [0039] 将所述归一化处理后的所述第三融合特征输入非线性层进行非线性映射，得到非线性映射后的第三融合特征；
- [0040] 将所述非线性映射后的第三融合特征输入卷积层进行特征学习，得到第二输出特征；
- [0041] 将所述第二输出特征依次输入所述归一化层、所述非线性层以及所述卷积层进行处理，得到第三输出特征；
- [0042] 将所述第三融合特征与所述第三输出特征进行加和处理，得到第四输出特征。

- [0043] 可选的，所述将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征包括：
- [0044] 将所述第一学习特征与所述第二学习特征进行拼接处理，得到第一融合特征。
- [0045] 可选的，所述对所述第一融合特征进行处理，得到蛋白质功能注释信息包括：
- [0046] 将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征；
- [0047] 基于第五预设层数的全连接层，对所述蛋白质融合特征进行分类处理，得到所述蛋白质功能注释信息。
- [0048] 可选的，所述将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征包括：
- [0049] 将所述第一融合特征输入归一化层进行归一化处理，得到归一化处理后的第一融合特征；
- [0050] 将所述归一化处理后的第一融合特征输入非线性层进行非线性映射，得到非线性映射后的第一融合特征；
- [0051] 将所述非线性映射后的第一融合特征输入全连接层进行特征提取，得到第一蛋白质融合特征；
- [0052] 将所述第一蛋白质融合特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二蛋白质融合特征；
- [0053] 将所述第一融合特征与所述第二蛋白质融合特征进行加和处理，得到所述蛋白质融合特征。
- [0054] 可选的，在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之前，所述方法还包括：
- [0055] 基于训练用蛋白质序列，对蛋白质功能注释模型进行训练。
- [0056] 可选的，所述将训练用蛋白质序列输入蛋白质功能注释模型进行训练包括：
- [0057] 根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征；
- [0058] 对所述训练用预设注释特征进行特征提取处理，得到第一训练用学习特征；
- [0059] 对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第二训练用学习特征；

- [0060] 将所述第一训练用学习特征与所述第二训练用学习特征进行融合处理，得到第一训练用融合特征；
- [0061] 对所述第一训练用融合特征进行处理，得到训练用蛋白质功能注释信息。
- [0062] 可选的，在根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征之后，所述方法还包括：
- [0063] 基于第一预设层数的训练用全连接层，对所述训练用预设注释特征进行特征提取处理，得到第三训练用学习特征；
- [0064] 基于第二预设层数的训练用卷积层，分别对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第四训练用学习特征与第五训练用学习特征；
- [0065] 将所述第三训练用学习特征、所述第四训练用学习特征以及所述第五训练用学习特征进行融合处理，得到第二训练用融合特征；其中，所述融合处理为加和处理；
- [0066] 基于第三预设层数的训练用全连接层，对所述第二训练用融合特征进行分类计算，得到训练用蛋白质功能注释信息。
- [0067] 根据本公开的第二方面，提供了一种蛋白质功能注释的装置，包括：
- [0068] 获取单元，用于根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；
- [0069] 第一处理单元，用于对所述预设注释特征进行特征提取处理，得到第一学习特征；
- [0070] 第二处理单元，用于对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；
- [0071] 第一融合单元，用于将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；
- [0072] 第三处理单元，用于对所述第一融合特征进行处理，得到蛋白质功能注释信息。
- [0073] 可选的，所述装置还包括：

- [0074] 第四处理单元，用于在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之后，基于第一预设层数的全连接层，对所述预设注释特征进行特征提取处理，得到第三学习特征；
- [0075] 第五处理单元，用于基于第二预设层数的卷积层，分别对所述序列特征、所述预训练序列特征进行特征提取处理，得到第四学习特征与第五学习特征；
- [0076] 第二融合单元，用于将所述第三学习特征、所述第四学习特征以及所述第五学习特征进行融合处理，得到第二融合特征；其中，所述融合处理为加和处理；
- [0077] 第六处理单元，用于基于第三预设层数的全连接层，对所述第二融合特征进行分类计算，得到蛋白质功能注释信息。
- [0078] 可选的，所述获取单元包括：
- [0079] 编码模块，用于基于预设编码算法，对所述蛋白质序列进行编码得到所述序列特征；
- [0080] 预训练模块，用于基于预设蛋白质语言模型，对所述序列特征进行预训练处理，得到所述预训练序列特征；
- [0081] 查找模块，用于在预设蛋白质数据库中，查找与所述蛋白质序列对应的所述预设注释特征。
- [0082] 可选的，所述第一处理单元包括：
- [0083] 第一输入模块，用于将所述预设注释特征输入第一预设数量的残差全连接模块；
- [0084] 第一处理模块，用于将第一个残差全连接模块的输出特征作为第二个残差全连接模块的输入特征进行特征提取处理；
- [0085] 第一循环模块，用于按照第一处理模块的处理方法循环执行第一预定次数，以获取所述第一学习特征，其中，所述第一预定次数等于所述第一预设数量。
- [0086] 可选的，所述第一输入模块还用于：
- [0087] 将所述预设注释特征输入归一化层进行归一化处理，得到归一化处理后的所述预设注释特征；
- [0088] 将所述归一化处理后的预设注释特征输入非线性层进行非线性映射，得到非线性映射后的预设注释特征；

- [0089] 将所述非线性映射后的预设注释特征输入全连接层进行特征提取，得到第一预设注释特征；
- [0090] 将所述第一预设注释特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二预设注释特征；
- [0091] 将所述预设注释特征与所述第二预设注释特征进行加和处理，得到第一输出特征。
- [0092] 可选的，所述第二处理单元包括：
- [0093] 第二处理模块，用于基于第四预设层数的卷积层，对所述序列特征进行处理，得到第六学习特征、对所述预训练序列特征进行处理，得到第七学习特征；
- [0094] 第一加和模块，用于将所述第六学习特征与所述第七学习特征进行加和处理，得到第三融合特征；
- [0095] 第二输入模块，用于将所述第三融合特征输入第二预设数量的残差卷积模块；
- [0096] 第一提取模块，用于将第一个所述残差卷积模块的输出特征作为第二个所述残差卷积模块的输入特征进行特征提取处理；
- [0097] 第二循环模块，用于按照第一提取模块的处理方法循环执行第二预定次数，得到所述第二学习特征，其中，所述第二预定次数等于所述第二预设数量。
- [0098] 可选的，所述第二输入模块还用于：
- [0099] 将所述第三融合特征输入归一化层进行归一化处理，得到归一化处理后的第三融合特征；
- [0100] 将所述归一化处理后的所述第三融合特征输入非线性层进行非线性映射，得到非线性映射后的第三融合特征；
- [0101] 将所述非线性映射后的第三融合特征输入卷积层进行特征学习，得到第二输出特征；
- [0102] 将所述第二输出特征依次输入所述归一化层、所述非线性层以及所述卷积层进行处理，得到第三输出特征；
- [0103] 将所述第三融合特征与所述第三输出特征进行加和处理，得到第四输出特征。
- [0104] 可选的，所述第一融合单元还用于：
- [0105] 将所述第一学习特征与所述第二学习特征进行拼接处理，得到第一融合特征。

- [0106] 可选的，所述第三处理单元包括：
- [0107] 整合模块，用于将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征；
- [0108] 分类模块，用于基于第五预设层数的全连接层，对所述蛋白质融合特征进行分类处理，得到所述蛋白质功能注释信息。
- [0109] 可选的，所述整合模块还用于：
- [0110] 将所述第一融合特征输入归一化层进行归一化处理，得到归一化处理后的第一融合特征；
- [0111] 将所述归一化处理后的第一融合特征输入非线性层进行非线性映射，得到非线性映射后的第一融合特征；
- [0112] 将所述非线性映射后的第一融合特征输入全连接层进行特征提取，得到第一蛋白质融合特征；
- [0113] 将所述第一蛋白质融合特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二蛋白质融合特征；
- [0114] 将所述第一融合特征与所述第二蛋白质融合特征进行加和处理，得到所述蛋白质融合特征。
- [0115] 可选的，所述装置还包括：
- [0116] 训练单元，用于在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之前，基于训练用蛋白质序列，对蛋白质功能注释模型进行训练。
- [0117] 可选的，所述训练单元包括：
- [0118] 训练用获取模块，用于根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征；
- [0119] 训练用第一处理模块，用于对所述训练用预设注释特征进行特征提取处理，得到第一训练用学习特征；
- [0120] 训练用第二处理模块，用于对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第二训练用学习特征；

- [0121] 训练用第一融合模块，用于将所述第一训练用学习特征与所述第二训练用学习特征进行融合处理，得到第一训练用融合特征；
- [0122] 训练用第三处理模块，用于对所述第一训练用融合特征进行处理，得到训练用蛋白质功能注释信息。
- [0123] 可选的，所述训练单元还包括：
- [0124] 训练用第四处理模块，用于在根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征之后，基于第一预设层数的训练用全连接层，对所述训练用预设注释特征进行特征提取处理，得到第三训练用学习特征；
- [0125] 训练用第五处理模块，用于基于第二预设层数的训练用卷积层，分别对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第四训练用学习特征与第五训练用学习特征；
- [0126] 训练用第二融合模块，用于将所述第三训练用学习特征、所述第四训练用学习特征以及所述第五训练用学习特征进行融合处理，得到第二训练用融合特征；其中，所述融合处理为加和处理；
- [0127] 训练用第六处理模块，用于基于第三预设层数的训练用全连接层，对所述第二训练用融合特征进行分类计算，得到训练用蛋白质功能注释信息。
- [0128] 根据本公开的第三方面，提供了一种电子设备，包括：
- [0129] 至少一个处理器；以及
- [0130] 与所述至少一个处理器通信连接的存储器；其中，
- [0131] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行前述第一方面所述的方法。
- [0132] 根据本公开的第四方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中，所述计算机指令用于使所述计算机执行前述第一方面所述的方法。
- [0133] 根据本公开的第五方面，提供了一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现如前述第一方面所述的方法。

[0134] 本公开提供了一种蛋白质功能注释的方法及装置、电子设备和存储介质，根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；基于第一预设数量的残差全连接模块，对所述预设注释特征进行特征提取处理，得到第一学习特征；基于第二预设数量的残差卷积模块，对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；对所述第一融合特征进行处理，得到蛋白质功能注释信息。与相关技术相比，本公开基于蛋白质序列获取序列特征、预训练序列特征以及预设注释特征，通过对前述三种不同角度的特征的融合处理，使得获得的蛋白质功能注释信息更加准确；有效提升了蛋白质功能注释的预测准确率。

[0135] 应当理解，本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征，也不用于限制本申请的范围。本申请的其它特征将通过以下的说明书而变得容易理解。

## 附图说明

- [0136] 附图用于更好地理解本方案，不构成对本公开的限定。其中：
- [0137] 图1为本公开实施例提供的一种蛋白质功能注释的方法的流程示意图；
- [0138] 图2为本公开实施例提供的另一种蛋白质功能注释的方法的流程示意图；
- [0139] 图3为一种蛋白质功能注释模型MergeNet算法网络架构示意图；
- [0140] 图4为本公开提供的一种蛋白质功能注释信息预测结果示意图；
- [0141] 图5为本公开实施例提供的另一种蛋白质功能注释的方法的流程示意图；
- [0142] 图6为一种蛋白质序列特征的示意图；
- [0143] 图7为一种蛋白质功能注释模型ResMergeNet算法网络架构示意图；
- [0144] 图8为本公开实施例提供的一种蛋白质功能注释模型训练的流程示意图；
- [0145] 图9为本公开实施例提供的另一种蛋白质功能注释模型训练的流程示意图；
- [0146] 图10为本公开实施例提供的一种蛋白质功能注释的装置的结构示意图；
- [0147] 图11为本公开实施例提供的另一种蛋白质功能注释的装置的结构示意图；
- [0148] 图12为本公开实施例提供的示例电子设备的示意性框图。

## 具体实施方式

- [0149] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。
- [0150] 随着蛋白质计算工具的发展，目前可以从蛋白质序列信息中计算出不同角度的序列特征。然而，当前仅基于序列信息的蛋白质功能注释算法往往只使用了其中一种序列特征，无法很好地融合多种蛋白质序列特征是影响当前仅基于序列信息的蛋白质功能注释算法预测性能的主要原因。
- [0151] 下面参考附图描述本公开实施例的蛋白质功能注释的方法及装置、电子设备和存储介质。
- [0152] 图1为本公开实施例所提供的一种蛋白质功能注释的方法的流程示意图。
- [0153] 如图1所示，该方法包含以下步骤：
- [0154] 步骤101，根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征。
- [0155] 在本公开实施例中，蛋白质序列为需要进行蛋白质功能注释的蛋白质序列。所述序列特征为根据蛋白质序列生成的编码特征(One-Hot特征)，所述预训练序列特征为蛋白质语言模型ProtBert根据蛋白质的One-Hot特征计算出预训练特征，所述预设注释特征为基于蛋白质的序列在InterPro数据库中检索的InterPro特征。需要说明的是，本公开实施例中所述序列特征以One-Hot编码得到的One-Hot特征为例进行说明，所述预训练序列特征以用蛋白质语言模型根据序列特征预先进行计算出的特征为例进行说明，所述预设注释特征以用InterProScan工具根据序列在InterPro数据库中检索出的InterPro特征为例进行说明，但这并不构成对所述序列特征、所述预训练序列特征以及所述预设注释特征的限定。
- [0156] 根据需要进行蛋白质功能注释的蛋白质序列，通过One-Hot编码技术蛋白质序列编码成蛋白质序列的One-Hot特征矩阵；蛋白质语言模型ProtBert根据One-Hot特征可以挖掘蛋白质序列的内部模式信息，并将蛋白质序列的内部模式信

息显式地表达出来；InterPro特征整合了多种与蛋白质功能强相关的特征，利用InterPro特征可以以较高的精度对蛋白质的功能进行分类。

- [0157] DeepGOPlus算法使用蛋白质序列的One-Hot特征作为算法唯一的输入特征；但是序列One-Hot特征较为底层，算法模型很难高效的从中学到功能类别信息，导致模型预测准确率不高。虽然使用InterPro特征作为DeepGOZero算法唯一的输入特征，预测准确率相对较高，但是并非所有蛋白质都具有InterPro特征，仅利用InterPro特征的模型无法对所有蛋白质进行分类。本公开实施例通过使用前述三种蛋白质序列特征进行学习，可以有效地提高蛋白质功能注释的准确率。
- [0158] 步骤102，对所述预设注释特征进行特征提取处理，得到第一学习特征。
- [0159] 在本公开实施例中，预训练特征基于One-Hot特征计算得到，与One-Hot特征差异较小；而InterPro特征与One-Hot特征、预训练特征的差异性很大。在蛋白质功能注释模型的网络浅层简单的将三种特征相加，会破坏每个特征的各自的特性，进而影响蛋白质功能注释模型的预测准确率。但是简单地加深蛋白质功能注释模型的网络层数会造成梯度消失的问题，影响信息在网络内部的传递，进而也会影响影响蛋白质功能注释模型的预测准确率。本公开实施例通过将预设注释特征(InterPro特征)输入第一预设数量的残差全连接模块进行特征提取处理，可以减缓梯度消失的问题，使得蛋白质功能注释模型可以加深网络层，并在网络的高层很好地融合不同的蛋白质序列特征。
- [0160] 在蛋白质功能注释模型中，使用数量更多的残差全连接模块会出现过拟合的特征，影响预测的准确率，而数量较少的残差全连接模块也会影响预测的准确率；具体的，所述第一预设数量设置在4个左右更为合适；需要说明的是，本公开实施例以4个残差全连接模块进行说明，但这并不构成对所述第一预设数量的限定。
- [0161] 步骤103，对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征。

- [0162] 在本公开实施例中，请参考步骤102的说明，为了缓解蛋白质功能注释模型中梯度消失的问题，将所述序列特征、所述预训练序列特征输入第二预设数量的残差卷积模块进行特征提取处理。
- [0163] 在蛋白质功能注释模型中，使用数量更多的残差卷积模块会出现过拟合的特征，影响预测的准确率，而数量较少的残差卷积模块也会影响预测的准确率；具体的，所述第二预设数量设置在8个左右更为合适；需要说明的是，本公开实施例以8个残差卷积模块进行说明，但这并不构成对所述第二预设数量的限定。
- [0164] 步骤104，将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征。
- [0165] 在本公开实施例中，由于InterPro特征与One-Hot特征、预训练特征的差异性很大，因此将One-Hot特征、预训练特征以及InterPro特征分别输入第二预设数量的残差卷积模块、第一预设数量的残差全连接模块进行特征提取处理；并在蛋白质功能注释模型的网络高层将提取到的第一学习特征和第二学习特征进行融合处理，得到第一融合特征。将这些来自不同特征进行融合，可以弥补不同特征之间的不足，从而提高整体特征的质量，提高蛋白质功能注释模型的预测准确率。
- [0166] 步骤105，对所述第一融合特征进行处理，得到蛋白质功能注释信息。
- [0167] 在本公开实施例中，将所述第一学习特征和所述的第二学习特征简单的融合后，并不能够直接的得到蛋白质的功能注释信息。还需要对第一融合特征再进行一次特征提取，充分挖掘One-Hot特征、预训练特征以及InterPro特征之间的优势，实现特征之间的信息互补。通过对第一融合特征进行特征提取，生成蛋白质功能注释信息。
- [0168] 本公开提供了一种蛋白质功能注释的方法，根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；基于第一预设数量的残差全连接模块，对所述预设注释特征进行特征提取处理，得到第一学习特征；基于第二预设数量的残差卷积模块，对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；对所述第一融合特征进行处理，得到蛋白质功能注释

信息。与相关技术相比，本公开基于蛋白质序列获取序列特征、预训练序列特征以及预设注释特征，通过对前述三种不同角度的特征的融合处理，使得获得的蛋白质功能注释信息具有更多的蛋白质功能信息；有效提升了蛋白质功能注释的预测准确率。

[0169] 为了清楚说明本公开实施例，本公开实施例提供了另一种蛋白质功能注释的方法的流程示意图。

[0170] 如图2所示，该方法包含以下步骤：

[0171] 步骤201，根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征。

[0172] 有关步骤201可参阅上述实施例中的详细说明，故在此不再进行一一赘述。

[0173] 步骤202，基于第一预设层数的全连接层，对所述预设注释特征进行特征提取处理，得到第三学习特征。

[0174] 具体在本公开实施例中，提供了一种在网络浅层将多种序列特征融合的蛋白质功能注释模型(MergeNet算法),弥补了使用单一特征的模型算法的缺陷，提高了仅使用One-Hot特征的算法预测准确率，增加了仅使用InterPro特征的算法应用普适性。

[0175] 具体的，请参阅图3；MergeNet用一层全连接层(MLP)对InterPro特征进行初步的学习，得到学习后的InterPro特征 $I \in \mathbb{R}^d$ ；其中d为隐藏层的特征维度，在本实施例中将d设置成1024，并且所有隐藏层的特征维度d均保持一致。

[0176] 步骤203，基于第二预设层数的卷积层，分别对所述序列特征、所述预训练序列特征进行特征提取处理，得到第四学习特征与第五学习特征。

[0177] 具体在本公开实施例中，请继续参阅图3；将根据蛋白质序列得到的One-Hot特征H和预训练特征E输入一层一维卷积层(CNN)进行特征的初步学习，得到 $H \in \mathbb{R}^{N \times d}$ ， $E \in \mathbb{R}^{N \times d}$ ，其中d为卷积层的特征维度，然后将H和E在残基维度保留最大的特征值以进行最大池化操作，得到 $H \in \mathbb{R}^d$ 和 $E \in \mathbb{R}^d$ 。

[0178] 步骤204，将所述第三学习特征、所述第四学习特征以及所述第五学习特征进行融合处理，得到第二融合特征；其中，所述融合处理为加和处理。

[0179] 具体在本公开实施例中，MergeNet在网络的浅层(即三种特征均经过一个特征学习层后)将三种特征通过加和处理的方式进行融合，得到蛋白质特征 $P = H + E + I \in \mathbb{R}^d$ 。示例性的，加和处理的方式为

$$H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} H + E = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}。$$

[0180] 步骤205，基于第三预设层数的全连接层，对所述第二融合特征进行分类计算，得到蛋白质功能注释信息。

[0181] 具体在本公开实施例中，将蛋白质特征P送入全连接层计算出各个功能类别的逻辑概率 $C \in \mathbb{R}^{class}$  (蛋白质功能注释信息),其中class为蛋白质功能的总数，在本实施例中class为5901。预测结果示意图如图4所示，图中选取了算法预测出的十个蛋白质的三个功能，其中第一列为预测的蛋白质名称，后面三列为对每个蛋白质所预测出的功能，其中用“|”前面表示算法所预测的基因本体论中功能的编号，“|”后面表示蛋白质具有该功能的概率值。

[0182] 为了清楚说明本公开实施例，本公开实施例提供了另一种蛋白质功能注释的方法的流程示意图。

[0183] 如图5所示，该方法包含以下步骤：

[0184] 步骤301，基于预设编码算法，对所述蛋白质序列进行编码得到所述序列特征。

[0185] 步骤302，基于预设蛋白质语言模型，对所述序列特征进行预训练处理，得到所述预训练序列特征。

[0186] 具体在本公开实施例中，所述预设编码算法为One-Hot编码算法，利用One-Hot编码算法对待注释的蛋白质序列进行编码生成One-Hot特征。需要说明的是，本公开实施例并不对所述预设编码算法进行限定。利用蛋白质语言模型ProtBert对One-Hot特征进行预训练处理。蛋白质语言模型可以挖掘蛋白质序列的内部模式信息，并将蛋白质序列的内部模式信息显式地表达出来。

[0187] 示例性的，请参阅图6，利用One-Hot编码基于序列计算出蛋白质序列的One-Hot特征 $H \in \mathbb{R}^{N \times 20}$ ,其中N为蛋白质中氨基酸的数量，20为每个残基的特征维度，也是氨基酸的类别总数。如果该氨基酸对应的是20种氨基酸类别中的第j个，

则 $H_{ij}=1$ ， $H_i$ 其他19个位置上的值均为0；将序列的OneHot特征送入预训练模型ProtBert得到预训练特征 $E \in \mathbb{R}^{N \times 30}$ ，其中30是预训练特征中残基的特征维度。与One-Hot特征不一样的是，预训练特征E不是由0和1组成的二进制特征矩阵，而是由连续的浮点数组成的特征矩阵，以SwissProt中由八个残基组成的蛋白质SKSP4\_ASCTR为例，其序列为：CHYIFNTC，序列的One-Hot特征和预训练特征如图6所示，其序列OneHot特征的维度为 $H \in \mathbb{R}^{8 \times 20}$ ，其序列预训练特征的维度为 $E \in \mathbb{R}^{8 \times 30}$ 。

- [0188] 步骤303，在预设蛋白质数据库中，查找与所述蛋白质序列对应的所述预设注释特征。
- [0189] 具体在本公开实施例中，请参阅图7，图7为ResMergeNet模型算法网络架构示意图。使用InterProScan工具从InterPro数据库检索出与功能类别强相关的蛋白质InterPro特征 $I \in \mathbb{R}^M$ ，其中M为蛋白质InterPro特征的维度，即InterPro特征的类别总数。
- [0190] 步骤304，将所述预设注释特征输入第一预设数量的所述残差全连接模块中进行特征提取处理；得到所述第一学习特征。
- [0191] 所述步骤具体包括：将所述预设注释特征输入所述残差全连接模块中的第一个残差全连接模块；将所述第一个残差全连接模块的输出特征作为第二个残差全连接模块的输入特征进行特征提取处理；按照如上步骤循环执行第一预定次数，以获取所述第一学习特征，其中，所述第一预定次数等于所述第一预设数量。
- [0192] 所述将所述预设注释特征输入所述残差全连接模块中的第一个残差全连接模块包括：将所述预设注释特征输入归一化层进行归一化处理，得到归一化处理后的所述预设注释特征；将所述归一化处理后的预设注释特征输入非线性层进行非线性映射，得到非线性映射后的预设注释特征；将所述非线性映射后的预设注释特征输入全连接层进行特征提取，得到第一预设注释特征；将所述第一预设注释特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二预设注释特征；将所述预设注释特征与所述第二预设注释特征进行加和处理，得到第一输出特征。

- [0193] 具体在本公开实施例中，请继续参阅图7，将InterPro特征 $I \in \mathbb{R}^M$  (M为InterPro特征的类别总数)送入4个顺序连接的残差全连接模块(ResMLP)中进行高层特征的提取。在每个残差全连接模块中，输入特征首先通过归一化层(Norm)进行特征的归一化，然后通过非线性层(ReLU)进行非线性映射，接着通过全连接层(MLP)进行特征的学习，再接着顺序通过归一化层、非线性层和全连接层进行特征的学习，此时将该残差全连接模块的初始输入特征与经过学习的特征进行残差加和，得到高层InterPro特征 $I \in \mathbb{R}^d$  (第一学习特征)。
- [0194] 步骤305，基于第四预设层数的卷积层，对所述序列特征进行处理，得到第六学习特征、对所述预训练序列特征进行处理，得到第七学习特征。
- [0195] 具体在本公开实施例中，请继续参阅图7，将One-Hot特征与预训练特征分别输入一层一维卷积层进行特征的初步提取，得到 $H \in \mathbb{R}^{N \times d}$ 和 $E \in \mathbb{R}^{N \times d}$ ，其中N为蛋白质中氨基酸的数量，d为卷积层的特征维度。需要说明的是，本公开实施例以一层一维卷积层为例进行说明，这并不构成对第四预设层数的限定。
- [0196] 步骤306，将所述第六学习特征与所述第七学习特征进行加和处理，得到第三融合特征。
- [0197] 具体在本公开实施例中，在经过一层一维卷积层的学习后，模型将这两个特征进行加和处理，加和处理后的特征为第三融合特征 $P_s$ ，即 $P_s = H + E \in \mathbb{R}^{N \times d}$ 。
- [0198] 步骤307，将所述第三融合特征输入第二预设数量的所述残差卷积模块中进行特征提取处理，得到所述第二学习特征。
- [0199] 所述步骤具体包括：将所述第三融合特征输入所述残差卷积模块中的第一个残差卷积模块；将所述第一个所述残差卷积模块的输出特征作为第二个所述残差卷积模块的输入特征进行特征提取处理；按照如上步骤循环执行第二预定次数，得到所述第二学习特征，其中，所述第二预定次数等于所述第二预设数量。
- [0200] 所述将所述第三融合特征输入所述残差卷积模块中的第一个残差卷积模块包括：将所述第三融合特征输入归一化层进行归一化处理，得到归一化处理后的第三融合特征；将所述归一化处理后的所述第三融合特征输入非线性层进行非线性映射，得到非线性映射后的第三融合特征；将所述非线性映射后的第三融合特征输入卷积层进行特征学习，得到第二输出特征；将所述第二输出特征依次

输入所述归一化层、所述非线性层以及所述卷积层进行处理，得到第三输出特征；将所述第三融合特征与所述第三输出特征进行加和处理，得到第四输出特征。

- [0201] 具体在本公开实施例中，请继续参阅图7，将第三融合特征 $P_s$ 送入8个顺序连接的残差卷积模块中进行高层特征的提取。在每个残差卷积模块中，输入特征首先通过归一化层(Norm)进行特征的归一化，然后通过非线性层(ReLU)进行非线性映射，接着通过卷积层(CNN)进行特征的学习，再接着顺序通过归一化层、非线性层和卷积层进行特征的学习。在经过所有残差卷积模块的学习后，将第四输出特征在残基维度保留最大的特征值以进行最大池化操作，得到高层序列特征 $P_s \in \mathbb{R}^d$  (第二学习特征)。
- [0202] 步骤308，将所述第一学习特征与所述第二学习特征进行拼接处理，得到第一融合特征。
- [0203] 具体在本公开实施例中，高层序列特征 $P_s$  (第二学习特征)和高层InterPro特征 $I$  (第一学习特征)变得更为抽象，并且特征分布变得相似，均与蛋白质的功能高度相关。ResMergeNet将高层序列特征 $P_s$ 和高层InterPro特征 $I$ 进行拼接，得到的特征称为高层蛋白质特征(第一融合特征)，即 $P = \text{concatenate}(P_s, I) \in \mathbb{R}^{2d}$ ，其中 $\text{concatenate}(\cdot)$ 表示将两个向量按照特征维度进行拼接。
- [0204] 步骤309，将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征。
- [0205] 所述步骤具体包括：将所述第一融合特征输入归一化层进行归一化处理，得到归一化处理后的第一融合特征；将所述归一化处理后的第一融合特征输入非线性层进行非线性映射，得到非线性映射后的第一融合特征；将所述非线性映射后的第一融合特征输入全连接层进行特征提取，得到第一蛋白质融合特征；将所述第一蛋白质融合特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二蛋白质融合特征；将所述第一融合特征与所述第二蛋白质融合特征进行加和处理，得到所述蛋白质融合特征。
- [0206] 具体在本公开实施例中，请继续参阅图7，将高层蛋白质特征(第一融合特征)， $P = \text{concatenate}(P_s, I) \in \mathbb{R}^{2d}$ 输入残差整合模块(Integrate)中进行蛋白质特征的进一步

提取。在残差整合模块中，输入特征首先通过归一化层(Norm)进行特征的归一化，然后通过非线性层(ReLU)进行非线性映射，接着通过全连接层(MLP)进行特征的学习，再接着顺序通过归一化层、非线性层和全连接层进行特征的学习，此时将该残基全连接模块的初始输入特征与经过学习的特征进行残差加和，得到 $P \in \mathbb{R}^d$  (蛋白质融合特征)。

[0207] 步骤310，基于第五预设层数的全连接层，对所述蛋白质融合特征进行分类处理，得到所述蛋白质功能注释信息。

[0208] 具体在本公开实施例中，请继续参阅图7，将得到的蛋白质融合特征P送入全连接层(MLP)中进行蛋白质功能的分类，得到蛋白质的功能类别的逻辑概率 $C \in \mathbb{R}^{class}$  (蛋白质功能注释信息)。

[0209] 作为本公开实施例一种可能的实现方式中，在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之前，所述方法还包括：

[0210] 基于训练用蛋白质序列，对蛋白质功能注释模型进行训练。

[0211] 具体在本公开实施例中，在使用蛋白质功能注释模型进行蛋白质功能注释之前，将训练用的蛋白质序列输入蛋白质功能注释模型进行训练，以提高模型的预测准确率。

[0212] 为了清楚说明本公开实施例，本公开实施例提供了一种蛋白质功能注释模型地训练的方法的流程示意图。

[0213] 如图8所示，该方法包含以下步骤：

[0214] 步骤401，根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征。

[0215] 步骤402，对所述训练用预设注释特征进行特征提取处理，得到第一训练用学习特征。

[0216] 步骤403，对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第二训练用学习特征。

[0217] 步骤404，将所述第一训练用学习特征与所述第二训练用学习特征进行融合处理，得到第一训练用融合特征。

[0218] 步骤405，对所述第一训练用融合特征进行处理，得到训练用蛋白质功能注释信息。

[0219] 有关步骤101至步骤105可参阅上述实施例中的详细说明，故在此不再进行一一赘述。

[0220] 进一步地，在本实施例一种可能的实现方式中，在根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征之后，本公开实施例提供了另一种蛋白质功能注释模型地训练的方法的流程示意图。

[0221] 如图9所示，该方法包含以下步骤：

[0222] 步骤501，根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征。

[0223] 步骤502，基于第一预设层数的训练用全连接层，对所述训练用预设注释特征进行特征提取处理，得到第三训练用学习特征。

[0224] 步骤503，基于第二预设层数的训练用卷积层，分别对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第四训练用学习特征与第五训练用学习特征。

[0225] 步骤504，将所述第三训练用学习特征、所述第四训练用学习特征以及所述第五训练用学习特征进行融合处理，得到第二训练用融合特征；其中，所述融合处理为加和处理。

[0226] 步骤505，基于第三预设层数的训练用全连接层，对所述第二训练用融合特征进行分类计算，得到训练用蛋白质功能注释信息。

[0227] 有关步骤201至步骤205可参阅上述实施例中的详细说明，故在此不再进行一一赘述。

[0228] 与上述的蛋白质功能注释的方法相对应，本发明还提出一种蛋白质功能注释的装置。由于本发明的装置实施例与上述的方法实施例相对应，对于装置实施例中未披露的细节可参照上述的方法实施例，本发明中不再进行赘述。

[0229] 图10为本公开实施例提供的一种蛋白质功能注释的装置的结构示意图，如图10所示，包括：获取单元61、第一处理单元62、第二处理单元63、第一融合单元64和第三处理单元65。

- [0230] 获取单元61，用于根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；
- [0231] 第一处理单元62，用于对所述预设注释特征进行特征提取处理，得到第一学习特征；
- [0232] 第二处理单元63，用于对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；
- [0233] 第一融合单元64，用于将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；
- [0234] 第三处理单元65，用于对所述第一融合特征进行处理，得到蛋白质功能注释信息。
- [0235] 本公开提供了一种蛋白质功能注释的装置，根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；对所述预设注释特征进行特征提取处理，得到第一学习特征；对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；对所述第一融合特征进行处理，得到蛋白质功能注释信息。与相关技术相比，本公开基于蛋白质序列获取序列特征、预训练序列特征以及预设注释特征，通过对前述三种不同角度的特征的融合处理，使得获得的蛋白质功能注释信息更加准确；有效提升了蛋白质功能注释的预测准确率。
- [0236] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述装置还包括：
- [0237] 第四处理单元66，用于在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之后，基于第一预设层数的全连接层，对所述预设注释特征进行特征提取处理，得到第三学习特征；
- [0238] 第五处理单元67，用于基于第二预设层数的卷积层，分别对所述序列特征、所述预训练序列特征进行特征提取处理，得到第四学习特征与第五学习特征；
- [0239] 第二融合单元68，用于将所述第三学习特征、所述第四学习特征以及所述第五学习特征进行融合处理，得到第二融合特征；其中，所述融合处理为加和处理；

- [0240] 第六处理单元69，用于基于第三预设层数的全连接层，对所述第二融合特征进行分类计算，得到蛋白质功能注释信息。
- [0241] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述获取单元61包括：
- [0242] 编码模块611，用于基于预设编码算法，对所述蛋白质序列进行编码得到所述序列特征；
- [0243] 预训练模块612，用于基于预设蛋白质语言模型，对所述序列特征进行预训练处理，得到所述预训练序列特征；
- [0244] 查找模块613，用于在预设蛋白质数据库中，查找与所述蛋白质序列对应的所述预设注释特征。
- [0245] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述第一处理单元62包括：
- [0246] 第一输入模块621，用于将所述预设注释特征输入第一预设数量的残差全连接模块；
- [0247] 第一处理模块622，用于将所述第一个残差全连接模块的输出特征作为第二个残差全连接模块的输入特征进行特征提取处理；
- [0248] 第一循环模块623，用于按照第一处理模块的处理方法循环执行第一预定次数，以获取所述第一学习特征，其中，所述第一预定次数等于所述第一预设数量。
- [0249] 进一步地，在本实施例一种可能的实现方式中，所述第一输入模块621还用于：
- [0250] 将所述预设注释特征输入归一化层进行归一化处理，得到归一化处理后的所述预设注释特征；
- [0251] 将所述归一化处理后的预设注释特征输入非线性层进行非线性映射，得到非线性映射后的预设注释特征；
- [0252] 将所述非线性映射后的预设注释特征输入全连接层进行特征提取，得到第一预设注释特征；

- [0253] 将所述第一预设注释特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二预设注释特征；
- [0254] 将所述预设注释特征与所述第二预设注释特征进行加和处理，得到第一输出特征。
- [0255] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述第二处理单元63包括：
- [0256] 第二处理模块631，用于基于第四预设层数的卷积层，对所述序列特征进行处理，得到第六学习特征、对所述预训练序列特征进行处理，得到第七学习特征；
- [0257] 第一加和模块632，用于将所述第六学习特征与所述第七学习特征进行加和处理，得到第三融合特征；
- [0258] 第二输入模块633，用于将所述第三融合特征输入第二预设数量的残差卷积模块；
- [0259] 第一提取模块634，用于将所述第一个所述残差卷积模块的输出特征作为第二个所述残差卷积模块的输入特征进行特征提取处理；
- [0260] 第二循环模块635，用于按照第一提取模块的处理方法循环执行第二预定次数，得到所述第二学习特征，其中，所述第二预定次数等于所述第二预设数量。
- [0261] 进一步地，在本实施例一种可能的实现方式中，所述第二输入模块633还用于：
- [0262] 将所述第三融合特征输入归一化层进行归一化处理，得到归一化处理后的第三融合特征；
- [0263] 将所述归一化处理后的所述第三融合特征输入非线性层进行非线性映射，得到非线性映射后的第三融合特征；
- [0264] 将所述非线性映射后的第三融合特征输入卷积层进行特征学习，得到第二输出特征；
- [0265] 将所述第二输出特征依次输入所述归一化层、所述非线性层以及所述卷积层进行处理，得到第三输出特征；

- [0266] 将所述第三融合特征与所述第三输出特征进行加和处理，得到第四输出特征。
- [0267] 进一步地，在本实施例一种可能的实现方式中，所述第一融合单元64还用于：
- [0268] 将所述第一学习特征与所述第二学习特征进行拼接处理，得到第一融合特征。
- [0269] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述第三处理单元65包括：
- [0270] 整合模块651，用于将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征；
- [0271] 分类模块652，用于基于第五预设层数的全连接层，对所述蛋白质融合特征进行分类处理，得到所述蛋白质功能注释信息。
- [0272] 进一步地，在本实施例一种可能的实现方式中，所述整合模块651还用于：
- [0273] 将所述第一融合特征输入归一化层进行归一化处理，得到归一化处理后的第一融合特征；
- [0274] 将所述归一化处理后的第一融合特征输入非线性层进行非线性映射，得到非线性映射后的第一融合特征；
- [0275] 将所述非线性映射后的第一融合特征输入全连接层进行特征提取，得到第一蛋白质融合特征；
- [0276] 将所述第一蛋白质融合特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二蛋白质融合特征；
- [0277] 将所述第一融合特征与所述第二蛋白质融合特征进行加和处理，得到所述蛋白质融合特征。
- [0278] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述装置还包括：
- [0279] 训练单元610，用于在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之前，基于训练用蛋白质序列，对蛋白质功能注释模型进行训练。
- [0280] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述训练单元610包括：

- [0281] 训练用获取模块6101，用于根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征；
- [0282] 训练用第一处理模块6102，用于对所述训练用预设注释特征进行特征提取处理，得到第一训练用学习特征；
- [0283] 训练用第二处理模块6103，用于对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第二训练用学习特征；
- [0284] 训练用第一融合模块6104，用于将所述第一训练用学习特征与所述第二训练用学习特征进行融合处理，得到第一训练用融合特征；
- [0285] 训练用第三处理模块6105，用于对所述第一训练用融合特征进行处理，得到训练用蛋白质功能注释信息。
- [0286] 进一步地，在本实施例一种可能的实现方式中，如图11所示，所述训练单元610还包括：
- [0287] 训练用第四处理模块6106，用于在根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征之后，基于第一预设层数的训练用全连接层，对所述训练用预设注释特征进行特征提取处理，得到第三训练用学习特征；
- [0288] 训练用第五处理模块6107，用于基于第二预设层数的训练用卷积层，分别对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第四训练用学习特征与第五训练用学习特征；
- [0289] 训练用第二融合模块6108，用于将所述第三训练用学习特征、所述第四训练用学习特征以及所述第五训练用学习特征进行融合处理，得到第二训练用融合特征；其中，所述融合处理为加和处理；
- [0290] 训练用第六处理模块6109，用于基于第三预设层数的训练用全连接层，对所述第二训练用融合特征进行分类计算，得到训练用蛋白质功能注释信息。
- [0291] 需要说明的是，前述对方法实施例的解释说明，也适用于本实施例的装置，原理相同，本实施例中不再限定。
- [0292] 根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0293] 图12示出了可以用来实施本公开的实施例的示例电子设备700的示意性框图。

电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0294] 如图12所示，设备700包括计算单元701，其可以根据存储在ROM(Read-Only Memory，只读存储器)702中的计算机程序或者从存储单元708加载到RAM(Random Access Memory，随机访问/存取存储器)703中的计算机程序，来执行各种适当的动作和处理。在RAM 703中，还可存储设备700操作所需的各种程序和数据。计算单元701、ROM 702以及RAM 703通过总线704彼此相连。I/O(Input/Output，输入/输出)接口705也连接至总线704。

[0295] 设备700中的多个部件连接至I/O接口705，包括：输入单元706，例如键盘、鼠标等；输出单元707，例如各种类型的显示器、扬声器等；存储单元708，例如磁盘、光盘等；以及通信单元709，例如网卡、调制解调器、无线通信收发机等。通信单元709允许设备700通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0296] 计算单元701可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元701的一些示例包括但不限于CPU(Central Processing Unit，中央处理单元)、GPU(Graphic Processing Units，图形处理单元)、各种专用的AI(Artificial Intelligence，人工智能)计算芯片、各种运行机器学习模型算法的计算单元、DSP(Digital Signal Processor，数字信号处理器)、以及任何适当的处理器、控制器、微控制器等。计算单元701执行上文所描述的各个方法和处理，例如蛋白质功能注释的方法。例如，在一些实施例中，蛋白质功能注释的方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元708。在一些实施例中，计算机程序的部分或者全部可以经由ROM 702和/或通信单元709而被载入和/或安装到设备700上。当计算机程序加载到RAM 703并由计算单元

701执行时，可以执行上文描述的方法的一个或多个步骤。备选地，在其他实施例中，计算单元701可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行前述蛋白质功能注释的方法。

[0297] 本文中以上描述的系统和技术和各种实施方式可以在数字电子电路系统、集成电路系统、FPGA(Field Programmable Gate Array，现场可编程门阵列)、ASIC(Application-Specific Integrated Circuit，专用集成电路)、ASSP(Application Specific Standard Product，专用标准产品)、SOC(System On Chip，芯片上系统的系统)、CPLD(Complex Programmable Logic Device，复杂可编程逻辑设备)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0298] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0299] 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、RAM、ROM、EPROM(Electrically Programmable Read-Only-Memory，可擦除可编程只读存储器)或快闪存储器、光纤、CD-ROM(Compact Disc Read-

Only Memory, 便捷式紧凑盘只读存储器)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0300] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(Cathode-Ray Tube, 阴极射线管)或者LCD(LiquidCrystal Display, 液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0301] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:LAN(Local Area Network, 局域网)、WAN(Wide Area Network, 广域网)、互联网和区块链网络。

[0302] 计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务端的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与VPS服务("Virtual Private Server", 或简称"VPS")中,存在的管理难度大,业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0303] 其中,需要说明的是,人工智能是研究使计算机来模拟人的某些思维过程和智能行为(如学习、推理、思考、规划等)的学科,既有硬件层面的技术也有软件层面的技术。人工智能硬件技术一般包括如传感器、专用人工智能芯片、云计

算、分布式存储、大数据处理等技术；人工智能软件技术主要包括计算机视觉技术、语音识别技术、自然语言处理技术以及机器学习/深度学习、大数据处理技术、知识图谱技术等几大方向。

[0304] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

[0305] 上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

## 权利要求书

[权利要求 1]

一种蛋白质功能注释的方法，其特征在于，包括：  
根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；  
对所述预设注释特征进行特征提取处理，得到第一学习特征；  
对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；  
将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；  
对所述第一融合特征进行处理，得到蛋白质功能注释信息。

[权利要求 2]

根据权利要求1所述的方法，其特征在于，在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之后，所述方法还包括：  
基于第一预设层数的全连接层，对所述预设注释特征进行特征提取处理，得到第三学习特征；  
基于第二预设层数的卷积层，分别对所述序列特征、所述预训练序列特征进行特征提取处理，得到第四学习特征与第五学习特征；  
将所述第三学习特征、所述第四学习特征以及所述第五学习特征进行融合处理，得到第二融合特征；其中，所述融合处理为加和处理；  
基于第三预设层数的全连接层，对所述第二融合特征进行分类计算，得到蛋白质功能注释信息。

[权利要求 3]

根据权利要求1所述的方法，其特征在于，所述根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征包括：  
基于预设编码算法，对所述蛋白质序列进行编码得到所述序列特征；  
基于预设蛋白质语言模型，对所述序列特征进行预训练处理，得到所述预训练序列特征；

在预设蛋白质数据库中，查找与所述蛋白质序列对应的所述预设注释特征。

[权利要求 4]

根据权利要求1所述的方法，其特征在于，所述对所述预设注释特征进行特征提取处理，得到第一学习特征包括如下步骤：

将所述预设注释特征输入第一预设数量的残差全连接模块；

将第一个残差全连接模块的输出特征作为第二个残差全连接模块的输入特征进行特征提取处理；

按照如上步骤循环执行第一预定次数，以获取所述第一学习特征，其中，所述第一预定次数等于所述第一预设数量。

[权利要求 5]

根据权利要求4所述的方法，其特征在于，所述将所述预设注释特征输入所述残差全连接模块中的第一个残差全连接模块包括：

将所述预设注释特征输入归一化层进行归一化处理，得到归一化处理后的所述预设注释特征；

将所述归一化处理后的预设注释特征输入非线性层进行非线性映射，得到非线性映射后的预设注释特征；

将所述非线性映射后的预设注释特征输入全连接层进行特征提取，得到第一预设注释特征；

将所述第一预设注释特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二预设注释特征；

将所述预设注释特征与所述第二预设注释特征进行加和处理，得到第一输出特征。

[权利要求 6]

根据权利要求3所述的方法，其特征在于，基于第四预设层数的卷积层，对所述序列特征进行处理，得到第六学习特征、对所述预训练序列特征进行处理，得到第七学习特征；

将所述第六学习特征与所述第七学习特征进行加和处理，得到第三融合特征；

所述对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征包括如下步骤：

将所述第三融合特征输入第二预设数量的残差卷积模块；

将第一个所述残差卷积模块的输出特征作为第二个所述残差卷积模块的输入特征进行特征提取处理；

按照如上步骤循环执行第二预定次数，得到所述第二学习特征，其中，所述第二预定次数等于所述第二预设数量。

[权利要求 7]

根据权利要求6所述的方法，其特征在于，所述将所述第三融合特征输入所述残差卷积模块中的第一个残差卷积模块包括：

将所述第三融合特征输入归一化层进行归一化处理，得到归一化处理后的第三融合特征；

将所述归一化处理后的所述第三融合特征输入非线性层进行非线性映射，得到非线性映射后的第三融合特征；

将所述非线性映射后的第三融合特征输入卷积层进行特征学习，得到第二输出特征；

将所述第二输出特征依次输入所述归一化层、所述非线性层以及所述卷积层进行处理，得到第三输出特征；

将所述第三融合特征与所述第三输出特征进行加和处理，得到第四输出特征。

[权利要求 8]

根据权利要求1所述的方法，其特征在于，所述将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征包括：

将所述第一学习特征与所述第二学习特征进行拼接处理，得到第一融合特征。

[权利要求 9]

根据权利要求1所述的方法，其特征在于，所述对所述第一融合特征进行处理，得到蛋白质功能注释信息包括：

将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征；

基于第五预设层数的全连接层，对所述蛋白质融合特征进行分类处理，得到所述蛋白质功能注释信息。

- [权利要求 10] 根据权利要求9所述的方法，其特征在于，所述将所述第一融合特征输入整合模块进行整合处理，得到蛋白质融合特征包括：  
将所述第一融合特征输入归一化层进行归一化处理，得到归一化处理后的第一融合特征；  
将所述归一化处理后的第一融合特征输入非线性层进行非线性映射，得到非线性映射后的第一融合特征；  
将所述非线性映射后的第一融合特征输入全连接层进行特征提取，得到第一蛋白质融合特征；  
将所述第一蛋白质融合特征依次输入所述归一化层、所述非线性层以及所述全连接层进行处理，得到第二蛋白质融合特征；  
将所述第一融合特征与所述第二蛋白质融合特征进行加和处理，得到所述蛋白质融合特征。
- [权利要求 11] 根据权利要求1所述的方法，其特征在于，在根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征之前，所述方法还包括：  
基于训练用蛋白质序列，对蛋白质功能注释模型进行训练。
- [权利要求 12] 根据权利要求11所述方法，其特征在于，所述将训练用蛋白质序列输入蛋白质功能注释模型进行训练包括：  
根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征；  
对所述训练用预设注释特征进行特征提取处理，得到第一训练用学习特征；  
对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第二训练用学习特征；  
将所述第一训练用学习特征与所述第二训练用学习特征进行融合处理，得到第一训练用融合特征；  
对所述第一训练用融合特征进行处理，得到训练用蛋白质功能注释信息。

[权利要求 13]

根据权利要求12所述的方法，其特征在于，在根据训练用蛋白质序列，分别获取训练用序列特征、训练用第一序列特征以及训练用预设注释特征之后，所述方法还包括：

基于第一预设层数的训练用全连接层，对所述训练用预设注释特征进行特征提取处理，得到第三训练用学习特征；

基于第二预设层数的训练用卷积层，分别对所述训练用序列特征、所述训练用第一序列特征进行特征提取处理，得到第四训练用学习特征与第五训练用学习特征；

将所述第三训练用学习特征、所述第四训练用学习特征以及所述第五训练用学习特征进行融合处理，得到第二训练用融合特征；其中，所述融合处理为加和处理；

基于第三预设层数的训练用全连接层，对所述第二训练用融合特征进行分类计算，得到训练用蛋白质功能注释信息。

[权利要求 14]

一种蛋白质功能注释的装置，其特征在于，包括：

获取单元，用于根据蛋白质序列，分别获取序列特征、预训练序列特征以及预设注释特征；

第一处理单元，用于对所述预设注释特征进行特征提取处理，得到第一学习特征；

第二处理单元，用于对所述序列特征、所述预训练序列特征进行特征提取处理，得到第二学习特征；

第一融合单元，用于将所述第一学习特征与所述第二学习特征进行融合处理，得到第一融合特征；

第三处理单元，用于对所述第一融合特征进行处理，得到蛋白质功能注释信息。

[权利要求 15]

一种电子设备，其特征在于，包括：

至少一个处理器；以及

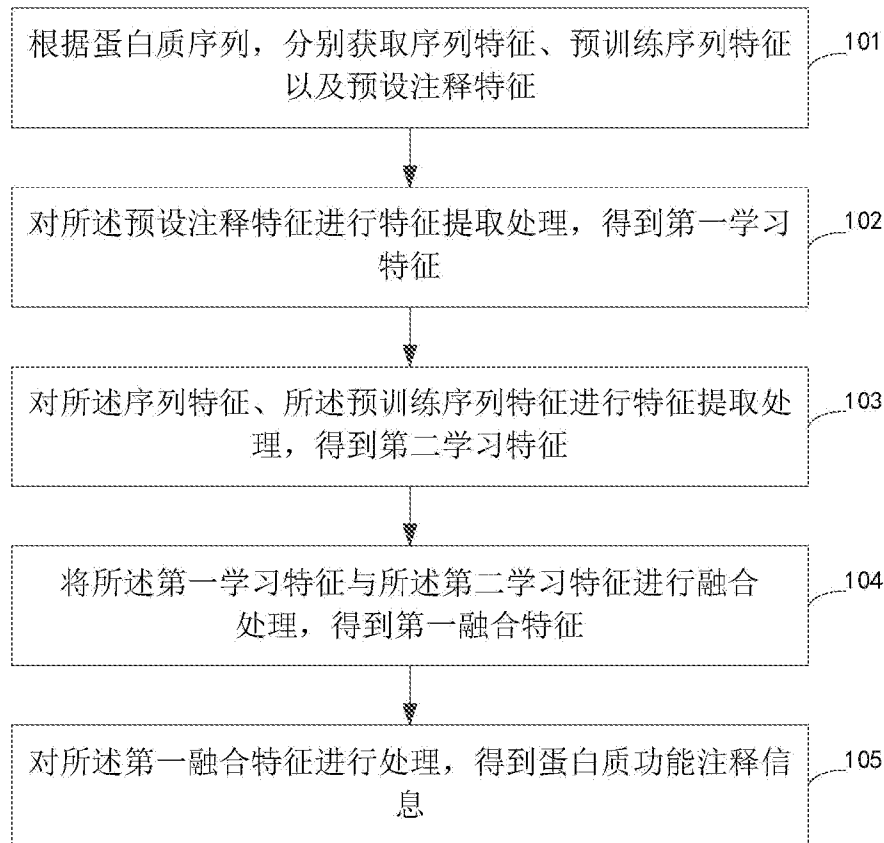
与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-13中任一项所述的方法。

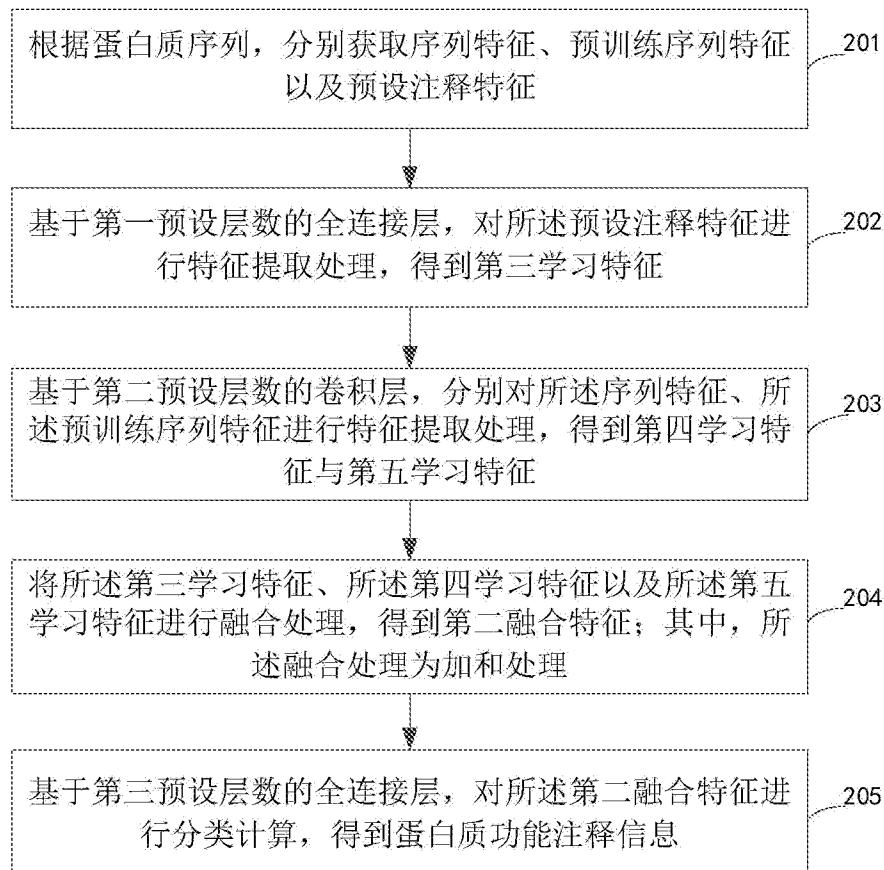
[权利要求 16] 一种存储有计算机指令的非瞬时计算机可读存储介质，其特征在于，所述计算机指令用于使所述计算机执行根据权利要求1-13中任一项所述的方法。

[权利要求 17] 一种计算机程序产品，其特征在于，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求1-13中任一项所述的方法。

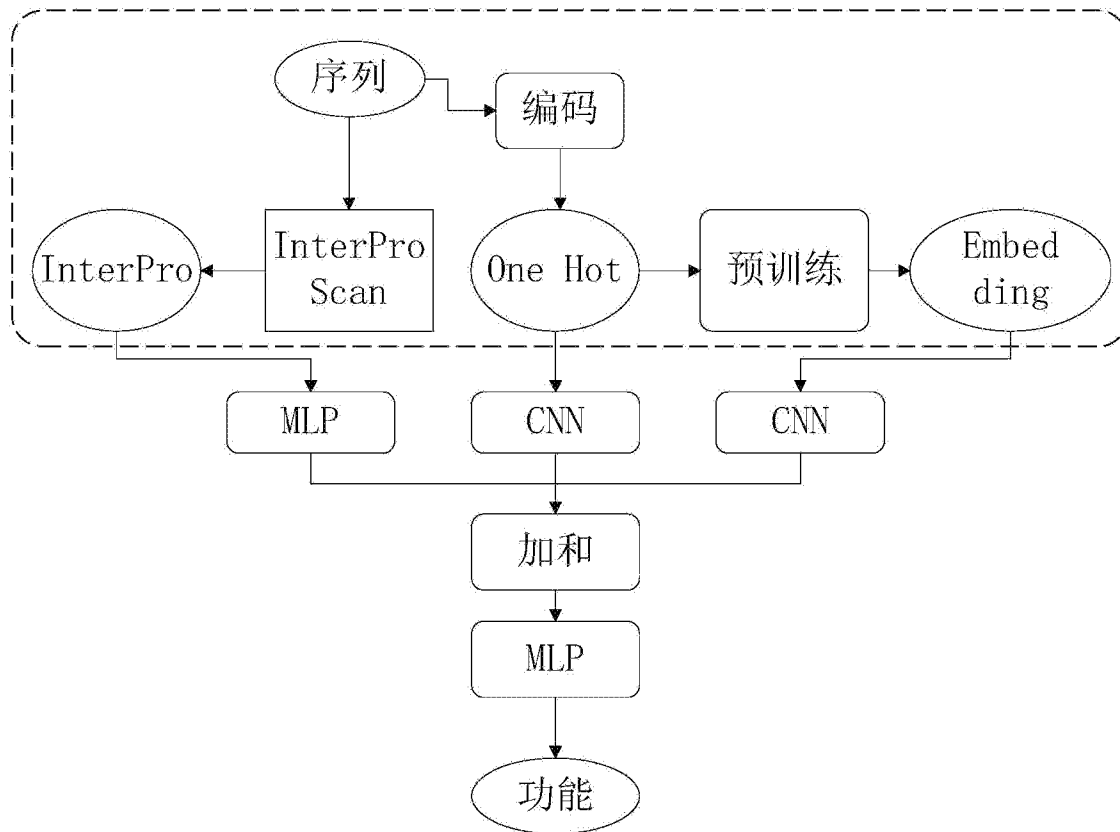
[图 1]



[图 2]



[图 3]

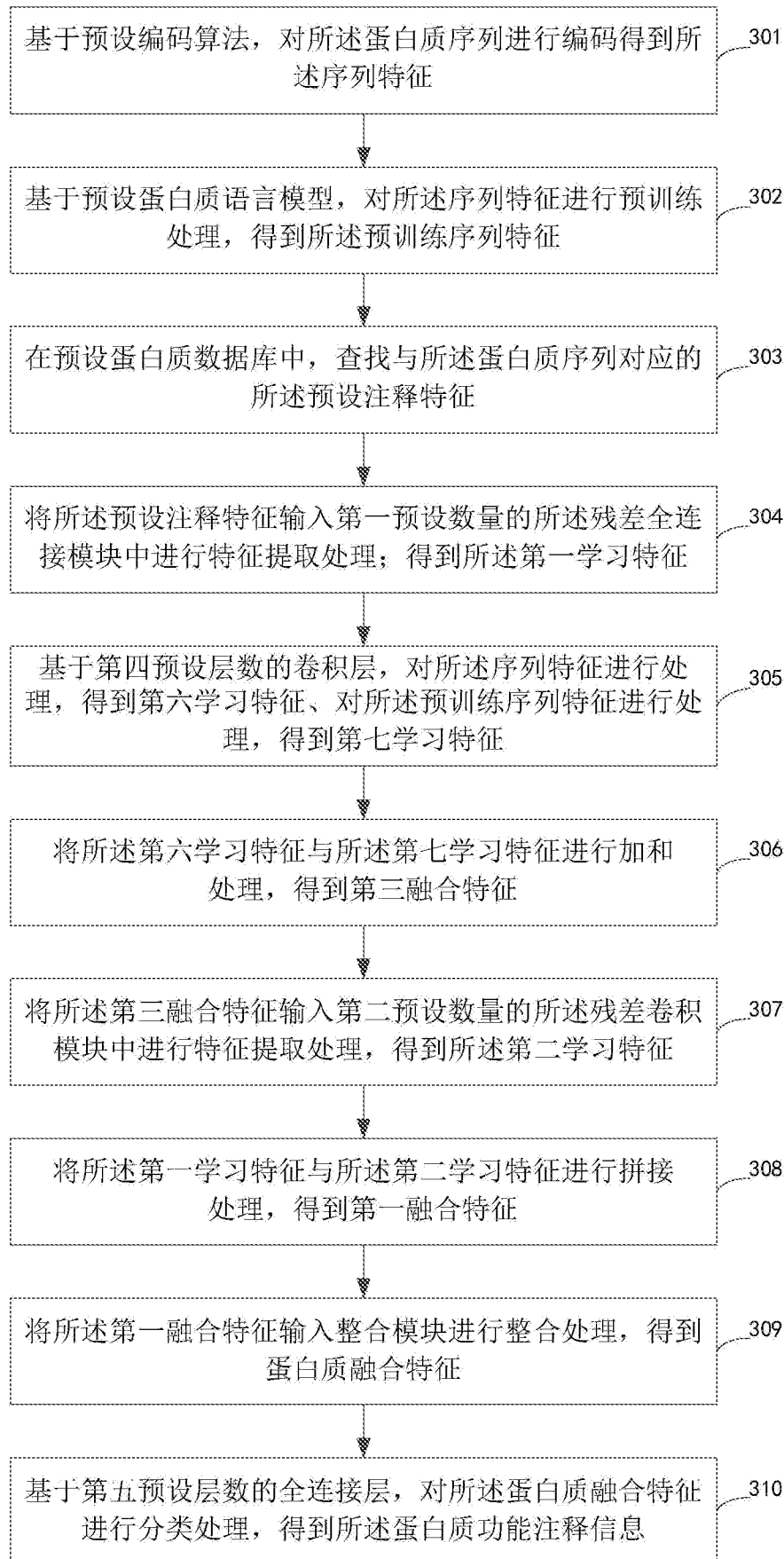
细则 26,  
30.06.2023

[图 4]

|            |                  |                  |                  |
|------------|------------------|------------------|------------------|
| protein_1  | GO:0008150 0.449 | GO:0005575 0.442 | GO:0110165 0.442 |
| protein_2  | GO:0005575 0.547 | GO:0110165 0.547 | GO:0016020 0.324 |
| protein_3  | GO:0008150 0.677 | GO:0003674 0.647 | GO:0003824 0.647 |
| protein_4  | GO:0008150 0.603 | GO:0003674 0.545 | GO:0003824 0.432 |
| protein_5  | GO:0005575 0.767 | GO:0110165 0.767 | GO:0016020 0.699 |
| protein_6  | GO:0008150 0.732 | GO:0003674 0.549 | GO:0005488 0.503 |
| protein_7  | GO:0005575 0.887 | GO:0110165 0.887 | GO:0008150 0.626 |
| protein_8  | GO:0003674 0.915 | GO:0003824 0.915 | GO:0008150 0.727 |
| protein_9  | GO:0008150 0.662 | GO:0009987 0.627 | GO:0003674 0.619 |
| protein_10 | GO:0008150 0.762 | GO:0003674 0.634 | GO:0005488 0.634 |

细则 26,  
30.06.2023

[图 5]



[图 6]

细则 26,  
30.06.2023

```

[[0 0 0 0 0 0 1 0 0 0 0 0 0 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 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 1 0 0 0 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 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 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 0 0 0 0 0 0 0 0 0 0 0 0 0]]

```

OneHot 特征 维度[8\*20]

```

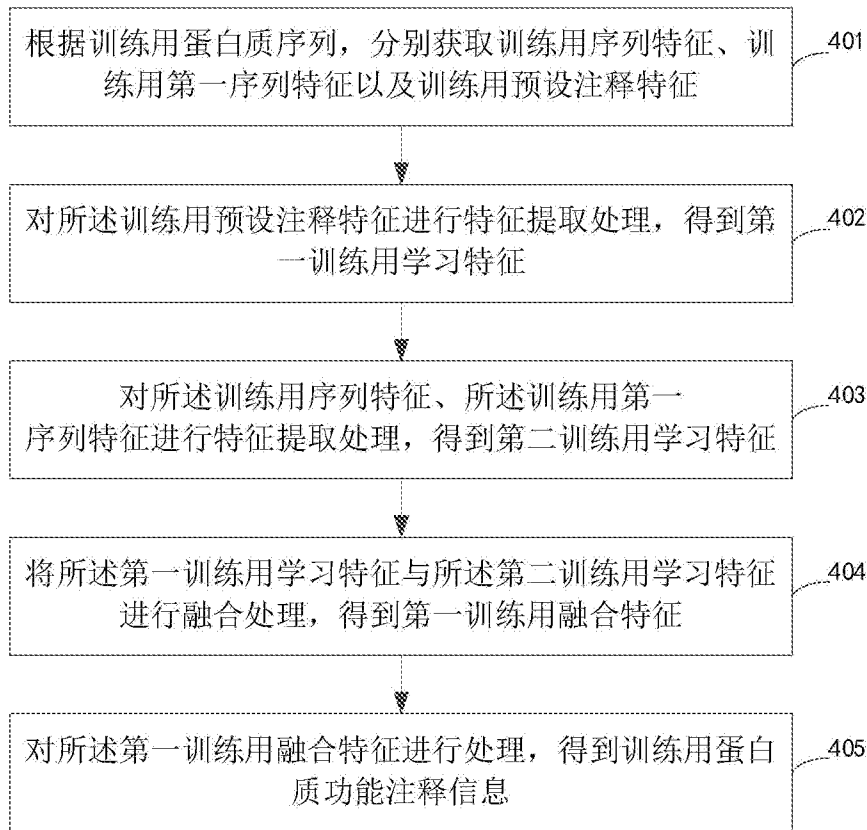
[[-17.58 -17.76 -17.46 -20.68 -19.58 -0.19 3.74 -0.64
 -0.35 0.2 -0.24 -0.75 0.21 -0.63 -0.78 -0.22
 -0.47 -1.14 -0.87 -1.57 -1.98 0.39 -2.15 -2.15
 -2.48 -4.74 -17.25 -17.28 -17.44 -17.78]
 [-19.49 -19.15 -19.15 -20.21 -21.18 0.35 0.76 -0.34
 0.11 0.65 -0.28 -0.32 0.24 -0.54 -0.24 3.11
 -0.69 -1.04 -0.76 -1.16 -1.6 -1.32 -1.99 -2.24
 -2.18 -5.87 -19.22 -19.46 -19.64 -20.03]
 [-19.72 -19.6 -19.83 -19.77 -20.86 0.6 1.12 0.24
 0.42 1.52 0.28 0.01 0.88 -0.16 0.35 0.47
 0.47 -0.43 -0.05 -1.02 -1.36 -0.95 -1.59 -1.93
 -2.02 -5.43 -19.56 -19.55 -19.41 -19.72]
 [-21.2 -21.15 -21.73 -19.27 -22.52 0.17 0.82 -0.34
 -0.16 4. -0.19 -0.48 0.26 -0.63 -0.32 0.07
 -0.31 -0.73 -0.55 -1.22 -1.78 -1.31 -1.84 -2.17
 -2.38 -5.59 -21.37 -20.91 -20.92 -21.44]
 [-17.92 -17.71 -17.52 -21.3 -19.55 0.37 3.78 -0.12
 -0.05 0.27 -0.42 -0.39 0.25 -0.53 -0.69 -0.18
 -0.57 -0.77 -0.7 -1.19 -1.69 -1.28 -1.96 -2.26
 -1.87 -6.04 -17.69 -17.71 -17.59 -17.79]
 [-19.43 -19.12 -19.45 -18.74 -20.83 0.27 0.94 -0.08
 0.16 0.55 -0.02 -0.36 4.05 -0.17 -0.42 0.15
 -0.05 -0.82 -0.41 -1.24 -1.76 -1.27 -1.88 -2.01
 -2.13 -4.99 -19.3 -19.33 -19.24 -19.38]
 [-21.18 -21.1 -21.58 -19.28 -22.47 0.31 0.56 -0.18
 -0.02 3.54 -0.19 -0.42 0.29 -0.13 -0.59 -0.13
 -0.31 -0.67 -0.45 -1.09 -1.45 -1.36 -1.46 -1.95
 -2.03 -5.48 -21.4 -20.93 -20.87 -21.26]
 [-20.21 -18.82 -19.26 -20.94 -19.23 3.26 0.88 -0.19
 -0.17 0.66 -0.27 -0.66 0.68 -0.51 -0.5 -0.09
 -0.65 -0.52 -0.56 -1.66 -1.9 -1.37 -1.76 -2.4
 -2.38 -5.82 -18.34 -20.12 -18.89 -19.22]]

```

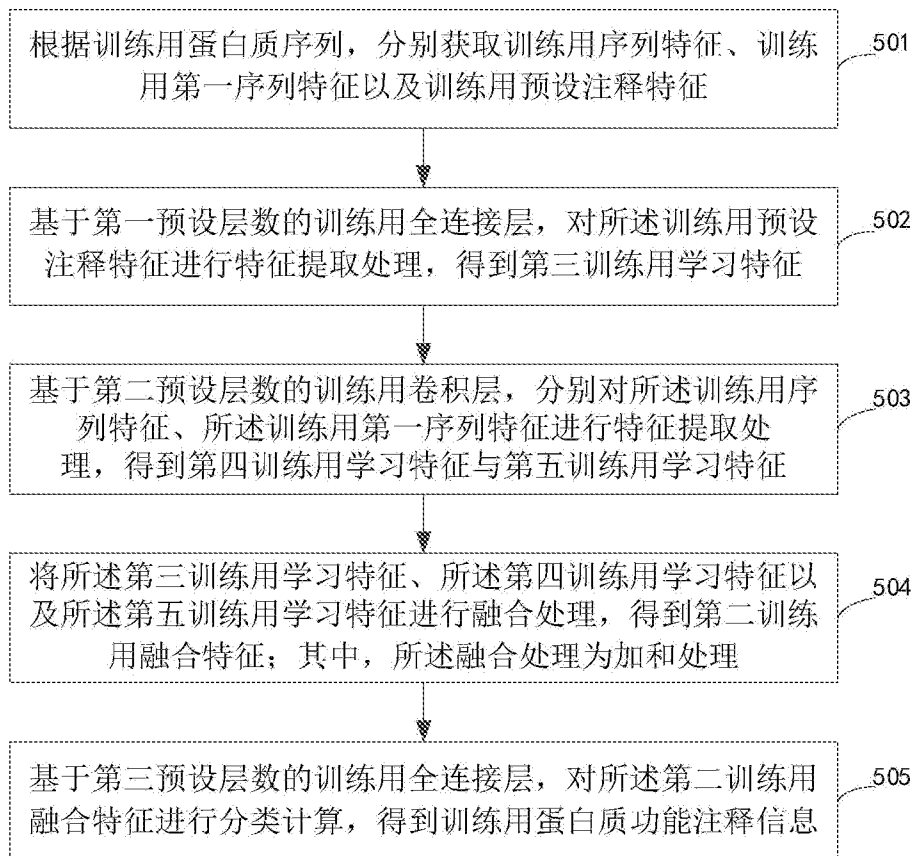
预训练特征 维度[8\*30]



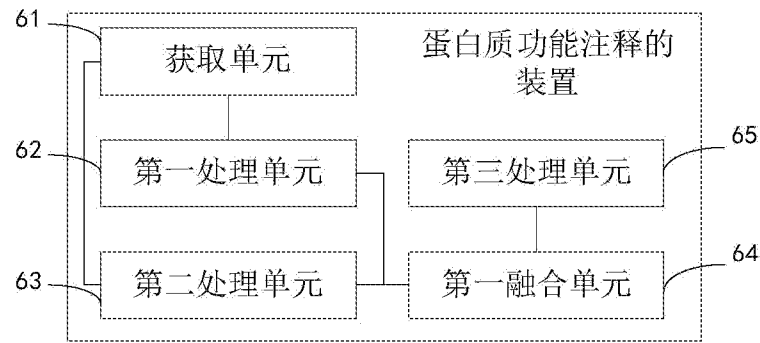
[图 8]



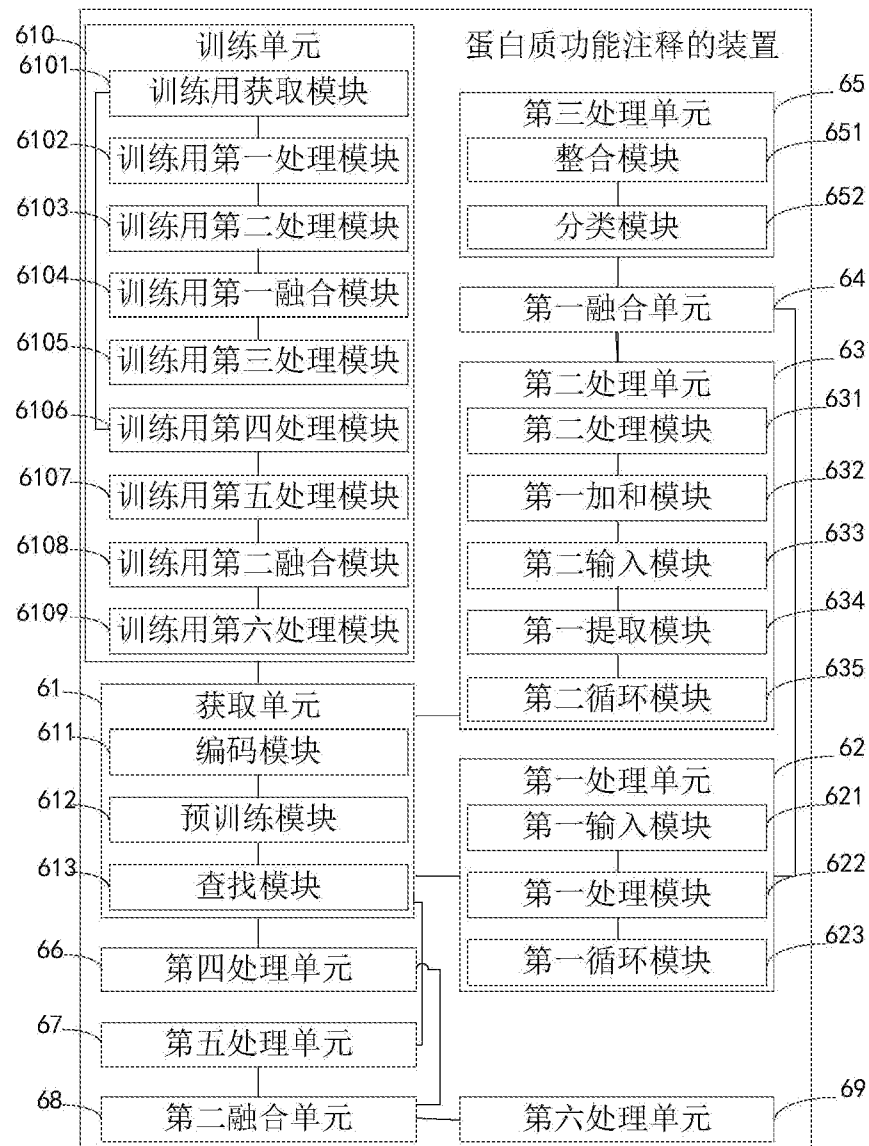
[图 9]



[图 10]



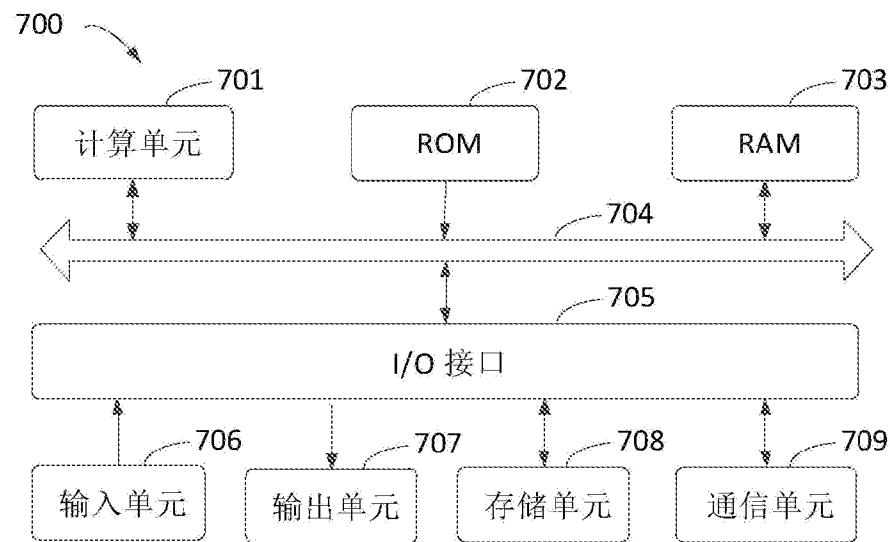
[图 11]



细则 26,  
30.06.2023

细则 26,  
30.06.2023

[图 12]



## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/099729

**A. CLASSIFICATION OF SUBJECT MATTER**

G16B 15/20(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)

CJFD, CNTXT, ENTXTC, VEN, CNKI, IEEE: 残差, 蛋白质, 酶, 性质, 属性, 功能, 点位, 独热, 分类, 谷歌, 特征, 融合, 序列, 预训练, 注释, residual, protein, function+, one hot, bert, pre 1w train+, CNN, DNN, feature, fus+, google, interpro, PTM, merg+

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                         | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Y         | CN 110070909 A (CENTRAL SOUTH UNIVERSITY) 30 July 2019 (2019-07-30)<br>description, paragraphs [0033]-[0051], and figure 1 | 1-4, 6, 8-9, 11-17    |
| Y         | CN 114898811 A (TSINGHUA UNIVERSITY) 12 August 2022 (2022-08-12)<br>description, paragraphs [0050]-[0057]                  | 1-4, 6, 8-9, 11-17    |
| A         | CN 116092577 A (OCEAN UNIVERSITY OF CHINA) 09 May 2023 (2023-05-09)<br>entire document                                     | 1-17                  |
| A         | CN 112562790 A (CHINA UNIVERSITY OF PETROLEUM) 26 March 2021 (2021-03-26)<br>entire document                               | 1-17                  |
| A         | US 2022172055 A1 (GOOGLE LLC) 02 June 2022 (2022-06-02)<br>entire document                                                 | 1-17                  |
| A         | WO 2022185179 A1 (GLAXOSMITHKLINE BIOLOGICALS SA) 09 September 2022<br>(2022-09-09)<br>entire document                     | 1-17                  |

☐ 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

31 January 2024

Date of mailing of the international search report

20 February 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**  
**Information on patent family members**

International application No.

**PCT/CN2023/099729**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) |            |    | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|------------|----|--------------------------------------|
| CN                                        | 110070909  | A  | 30 July 2019                         | None                    |            |    |                                      |
| CN                                        | 114898811  | A  | 12 August 2022                       | None                    |            |    |                                      |
| CN                                        | 116092577  | A  | 09 May 2023                          | None                    |            |    |                                      |
| CN                                        | 112562790  | A  | 26 March 2021                        | None                    |            |    |                                      |
| US                                        | 2022172055 | A1 | 02 June 2022                         | EP                      | 3953939    | A1 | 16 February 2022                     |
|                                           |            |    |                                      | WO                      | 2020210591 | A1 | 15 October 2020                      |
| WO                                        | 2022185179 | A1 | 09 September 2022                    | None                    |            |    |                                      |

A. 主题的分类

G16B 15/20(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))

CJFD、CNTXT、ENTXTC、VEN、CNKI、IEEE: 残差, 蛋白质, 酶, 性质, 属性, 功能, 点位,独热, 分类, 谷歌, 特征, 融合, 序列, 预训练, 注释, residual, protein, function+, one hot, bert, pre 1w train+, CNN, DNN, feature, fus+, google, interpro, PTM, merg+

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         |
|-----|------------------------------------------------------------------------------------|-----------------|
| Y   | CN 110070909 A (中南大学) 2019年7月30日 (2019 - 07 - 30)<br>说明书第[0033]-[0051]段及附图1        | 1-4、6、8-9、11-17 |
| Y   | CN 114898811 A (清华大学) 2022年8月12日 (2022 - 08 - 12)<br>说明书第[0050]-[0057]段            | 1-4、6、8-9、11-17 |
| A   | CN 116092577 A (中国海洋大学) 2023年5月9日 (2023 - 05 - 09)<br>全文                           | 1-17            |
| A   | CN 112562790 A (中国石油大学(华东)) 2021年3月26日 (2021 - 03 - 26)<br>全文                      | 1-17            |
| A   | US 2022172055 A1 (GOOGLE LLC) 2022年6月2日 (2022 - 06 - 02)<br>全文                     | 1-17            |
| A   | WO 2022185179 A1 (GLAXOSMITHKLINE BIOLOGICALS SA) 2022年9月9日 (2022 - 09 - 09)<br>全文 | 1-17            |

☐ 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1月31日

国际检索报告邮寄日期

2024年2月20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

王欣玥

电话号码 (+86) 010-53961379

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告  
关于同族专利的信息

国际申请号  
PCT/CN2023/099729

| 检索报告引用的专利文件 |            |    | 公布日<br>(年/月/日) | 同族专利             | 公布日<br>(年/月/日) |
|-------------|------------|----|----------------|------------------|----------------|
| CN          | 110070909  | A  | 2019年7月30日     | 无                |                |
| CN          | 114898811  | A  | 2022年8月12日     | 无                |                |
| CN          | 116092577  | A  | 2023年5月9日      | 无                |                |
| CN          | 112562790  | A  | 2021年3月26日     | 无                |                |
| US          | 2022172055 | A1 | 2022年6月2日      | EP 3953939 A1    | 2022年2月16日     |
|             |            |    |                | WO 2020210591 A1 | 2020年10月15日    |
| WO          | 2022185179 | A1 | 2022年9月9日      | 无                |                |