**基于多输入的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法**

**技术领域**

本发明涉及合成生物学和代谢工程领域，具体的，涉及一种基于深度学习的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法。

**背景技术**

近年来，合成生物学不断取得突破，代谢工程大放异彩。对于酶促反应的预测和对蛋白质的酶活性的预测在合成生物学和代谢工程中是极其重要而基础的工作。

目前，对于酶促反应的预测的方案有两种类型：类型一，基于专家规则和模板库，利用原子映射找到反应中心，进行官能团替换；类型二，基于序列或图的无模板方法，利用传统机器学习或者seq-2-seq序列模型进行预测。

目前对于酶促反应的预测方法，都需要假设反应一定是能够正向进行的酶促反应。而实际应用中，我们往往不知道反应是否是能够正向进行的酶促反应。目前对于酶促反应的预测只能预测单一产物。而实际应用中，我们的产物往往不止一个。目前对于酶促反应的预测必须对酶的活性加以描述。而实际应用中，我们只知道蛋白质序列，不知道该蛋白质是否是酶，更不知道该蛋白质有什么活性。

目前，对于蛋白质的酶活性预测方法的方案有两种类型：类型一，基于酶活性库，利用序列比对的方式将蛋白质比对到已知活性的酶活性库中；类型二，基于深度学习，对酶的EC码进行预测。

目前对于蛋白质的酶活性预测方法的方法，类型一极大受限于酶活性库的大小，对于库中没有的酶就无法预测，类型二完全不具有可解释性且预测效果差，酶活性的EC码是人为定义的编码，在理化水平上与蛋白质的性质毫无联系。

为此，本发明提出了一种基于多输入的Transformer框架的酶促反应预测和酶活性预测方法，该方法了能够在未知反应是否是酶促反应、未知反应物、未知活性的蛋白质的情况下，对酶促反应进行预测，得到反应的产物，并通过对产物和反应物的比较，推导出蛋白质的酶活性，从而对蛋白质的酶活性进行预测。

**发明内容**

本发明针对上述现有技术中存在的不足，提供了一种基于多输入的Transformer框架的酶促反应预测和酶活性预测方法，可以对于酶促反应和酶活性进行预测，可大大促进合成生物学和代谢工程中对于代谢网络或其它与酶促反应或与酶或与化学反应相关工作的推进。

本发明的技术方案是这样实现的：

基于多输入的Transformer框架的酶促反应预测和酶活性预测方法，其特征在于，包括以下步骤：

（1），搜集酶促反应的数据和非酶蛋白质的数据，包括对反应的描述和对酶的描述和对非酶蛋白质的描述；

（2），对于步骤(1)的描述，将对反应的描述转化为SMILES表达的反应物、生成物和反应式，将对酶的描述和对非酶蛋白质的描述转化成蛋白质序列；

（3），对于步骤(2)的反应物SMILES表达式，利用RDKit进行数据清理得到合理的SMILES表达式；对于步骤（2）的蛋白质序列，进行去重，并只保留合适长度的蛋白质；

（4），对于步骤（3）的反应物SMILES表达式，构建SMILES字典，利用Transformer编码器进行编码；对于步骤(3)的蛋白质序列，利用Transformer编码器进行编码；

（5），对于步骤（4）的SMILES编码和蛋白质编码，二者进行运算后再利用Transformer编码器进行混合编码；

（6），对于步骤（5）的混合编码输出，将步骤（5）的混合编码输出输入到Transformer解码器进行解码；

（7），对于步骤（6）的解码输出，输入到前馈神经网络层运算，得到输出再输入到Softmax层进行运算，得到步骤（4）的SMILES字典的每个token的概率；

（8），对于步骤（7）的概率输出，得到预测的生成物SMILES表达式；

（9），将步骤（8）预测的生成物SMILES与步骤（4）的反应物的SMILES表达式进行比较，推断出蛋白质的活性。

采用了上述技术方案的本发明的设计思想及有益效果是：

本发明的技术方案，提出了一种基于多输入的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法，采用注意力机制和混合注意力机制，对酶促反应的结果进行预测，并通过比较预测结果与反应物推断出蛋白质的酶活性。

另外，根据本发明的上述实施例的基于多输入的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法还可以具有以下附加的技术特征：

在本发明的一个实施例中，多输入的Transformer的训练需要利用已知的反应物-蛋白质-生成物三元组。反应物和生成物可以由SMILES表达式进行描述。反应物和生成物也可以由SMARTS表达式进行描述，因为SMARTS表达式可以表达SMILES表达式所表达的内容。

在本发明的一个实施例中，训练多输入的Transformer模型的数据可以进行数据增强，包括但不限于将化学式的SMILES表达式进行同分异构体变换、将多反应物的反应物的顺序进行调换、将多产物的反应的产物的顺序进行调换。

在本发明的一个实施例中，多输入的Transformer的训练数据分为训练集、测试集和验证集。

**附图说明**

本发明上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明实施例的基于多输入的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法流程图；

图2为根据本发明实施例的计算流程图。

**具体实施方式**

以下详细描述本发明的优选实施例，所述实施例的示例在附图中示出，其中自始至终

相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附

图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

实施例一，参考图1所示，一种基于多输入的Transformer框架的酶促反应预测和蛋白质的酶活性预测方法，其特征在于，包括：

步骤S101：收集得到反应物O=C[O-]与反应物CSCCC(N)C(=O)NC(CC(C)C)C(=O)NC(Cc1ccccc1)C(=O)O与反应物[H+]，蛋白质MAVRRVLKMGEPLLREVAAAVTRFDAELVALVTDMDDTMRAFSGAGIAAPQIGVSARVVIFELRDNPRYPHLAPVPYTVLVNPVLTPLGGEQDEGWEGCLSVPGMRGLVPRLRRLRYQGFDAHGAPIDRTVEGFHARVVQHEIDHLDGILFPQRVRDLRNFGFEDALTGHMTPMPD。这是一个多反应物的反应。进行数据清理和增强后，得到SMILES反应式：

O = C [O-] > C S C C C ( N ) C ( = O ) N C ( C C ( C ) C ) C ( = O ) N C ( C c 1 c c c c c 1 ) C ( = O ) O > [H+]。生成物C S C C [C@H] ( N C = O ) C ( = O ) N [C@@H] ( C C ( C ) C ) C ( = O ) N [C@@H] ( C c 1 c c c c c 1 ) C ( = O ) O > O。 其中反应物与生成物又由符号>分隔为化学分子。经过清理后的蛋白质序列为MAVRRVLKMGEPLLREVAAAVTRFDAELVALVTDMDDTMRAFSGAGIAAPQIGVSARVVIFELRDNPRYPHLAPVPYTVLVNPVLTPLGGEQDEGWEGCLSVPGMRGLVPRLRRLRYQGFDAHGAPIDRTVEGFHARVVQHEIDHLDGILFPQRVRDLRNFGFEDALTGHMTPMPD，清除训练集中与该蛋白质序列过于相似的蛋白质序列；

步骤S102：将步骤S101得到的反应物SMILES序列输入图2所示的反应物的Transformer编码器，得到大小为批量大小乘以窗口大小乘以反应物特征维度的张量。将步骤S101得到的蛋白质序列输入到图2所示的蛋白质的Transformer编码器，得到大小为批量大小乘以窗口大小乘以蛋白质特征维度的张量；

步骤S103：将步骤S102得到的反应物张量和蛋白质张量，输入到图2所示的混合计算模块，计算得到大小为批量大小乘以窗口大小乘以特征维度的张量；将得到的混合张量，输入到图2所示的混合编码器模块，计算得到大小为批量大小乘以窗口大小乘以特征维度的张量；

步骤S104：将步骤S103得到的编码，输入到图2所示的解码器模块，进行解码运算后得到生成物的预测值CSCC[C@H](NC=O)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](Cc1ccccc1)C(=O)O和O；

步骤S105：比较步骤S104得到的生成物预测值与作为输入的反应物，找到官能团的差异，根据已有的化学反应知识，确定蛋白质的酶活性为3.5.1.88。

综上所述，本发明提出了一种基于多输入的Transformer框架的酶促反应预测和酶活性预测方法，该方法了能够在未知反应是否是酶促反应、未知反应物、未知活性的蛋白质的情况下，对酶促反应进行预测，得到反应的产物。并能够通过对产物和反应物的比较，推导出蛋白质的酶活性，从而对蛋白质的酶活性进行预测。

在本说明书的描述中，参考术语“ 一个实施例”、“ 一些实施例”、“ 示例”、“ 具体示例”、或“ 一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结

合和组合。

此外，术语“ 第一”、“ 第二”仅用于描述目的，而不能理解为指示或暗示相对重要性

或者隐含指明所指示的技术特征的数量。由此，限定有“ 第一”、“ 第二”的特征可以明示或者隐含地包括至少一个该特征。在本发明的描述中，“ 多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本发明的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本发明的实施例所属技术领域的技术人员所理解。

在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执

行系统、装置或设备取指令并执行指令的系统)使用，或结合这些指令执行系统、装置或设

备而使用。就本说明书而言，"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装

置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下：具有一个或多个布线的电

连接部(电子装置) ，便携式计算机盘盒(磁装置) ，随机存取存储器(RAM) ，只读存储器

(ROM)，可擦除可编辑只读存储器(EPROM或闪速存储器)，光纤装置，以及便携式光盘只读存储器(CDROM)。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如，如果用硬件来实现和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，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模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

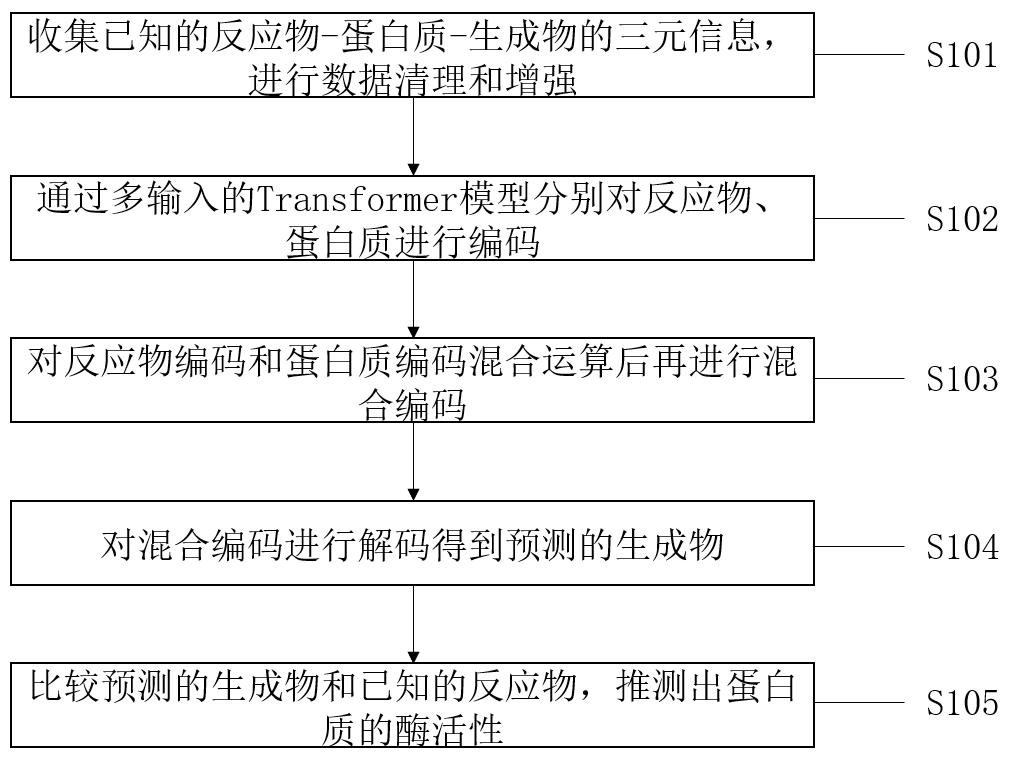


图1

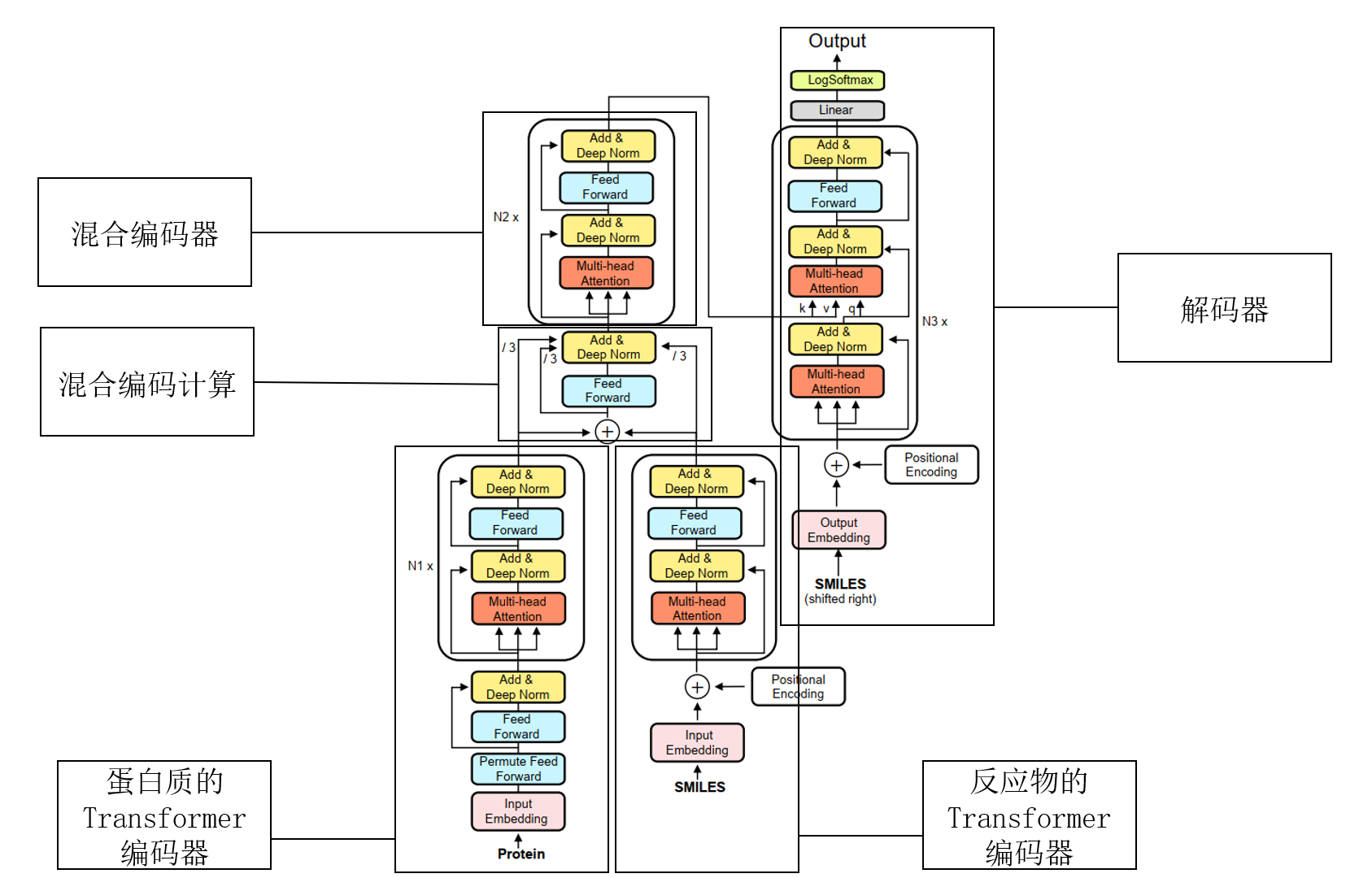


图2