**DEEPre: sequence-based enzyme EC number prediction by deep learning**

**DEEPre:用深度学习预测基于序列的酶EC数**

本篇论文来自SCI二区BIOINFORMATICS. 研究内容为对于使用EC数字系统编码的酶的功能的预测，研究领域序列分析。就题目而言，不了解酶的数字编码系统应该很难理解酶EC数是什么，由于我是在找蛋白质序列分析的文章中发现这篇的，对于酶的了解不多，因此建议题目直接说明是预测酶的功能的，对于EC数字系统可以在文章中进行介绍。

**Abstract**

Annotation of enzyme function has a broad range of applications, such as （研究领域及内容）metagenomics, industrial biotechnology, and diagnosis of enzyme deficiency-caused diseases. However,（业界存在的问题） the time and resource required make it prohibitively expensive to experimentally determine the function of every enzyme. Therefore, computational enzyme function prediction has become increasingly important. In this paper, （提出本文的方法）we develop such an approach, determining the enzyme function by predicting the Enzyme Commission number.

（本文的方法的简要介绍）We propose an end-to-end feature selection and classification model training approach, as well as an automatic and robust feature dimensionality uniformization method, DEEPre, in the field of enzyme function prediction. （本文方法的优点）Instead of extracting manually crafted features from enzyme sequences, our model takes the raw sequence encoding as inputs, extracting convolutional and sequential features from the raw encoding based on the classification result to directly improve the prediction performance. （实验结果）The thorough cross-fold validation experiments conducted on two largescale datasets show that DEEPre improves the prediction performance over the previous state-ofthe-art methods. In addition, our server outperforms five other servers in determining the main class of enzymes on a separate low-homology dataset. Two case studies demonstrate DEEPre’s ability to capture the functional difference of enzyme isoforms.

摘要第一句介绍研究领域和研究内容。第二句的However转折和第三句的Therefore进行呼应，引出由传统生物方法的缺点说明利用人工智能算法确定酶的功能很有必要，进而在第四句引出本文的方法。第二段的第一句简要介绍了本文方法具体是什么模式，第二句介绍DEEpre的优点。与传统特征提取方法相比，本文根据分类结果从原始编码中提取卷积特征和序列特征，基于上述优点，第三句抛出实验结果证明DEEpre与之前的先进方法相比提高了预测精度。

小结：逻辑清晰，但是对于方法做出的贡献可以详细列成几点做出说明，而不是跟以前的方法作比较说明。

**1. Introduction**

（引言的第一段对应摘要第一段的第一、二、三句，详细说明了酶功能注释的应用领域和作用，以及传统的鉴定酶功能的方法）

（主题句）Enzymes, an essential kind of proteins in the human body, catalyzing reactions in vivo, play a vital role in regulating biological processes. Annotation of enzyme function has a broad range of applications, such as metagenomics, industrial biotechnology, and diagnosis of enzyme deficiency-caused diseases. （说明酶的重要性）The dysfunction of certain enzymes would cause serious metabolic diseases. （举例说明）For example, …... （转折，说明以前方法的不足）However, conducting experiments requires significant amount of time and expert efforts, which may not cope with the rapid increase in the number of new enzymes. （提出新方法）In this context, computational methods emerged to assist biologists in determining enzyme function and guiding the direction of setting up the validating experiments.

引言第一段介绍了酶的重要性，并且指出传统生物实验方法来确定酶的功能的困难，顺势提出本文的新方法。

（引言第二段对应标题提出的EC数字系统，并详细介绍了这种数字系统）

According to SWISS-PROT (Bairoch and Apweiler, 2000) (released on September 7, 2016), among the 539566 manually annotated proteins, 258733 proteins are enzymes. （点出本文用的分类方案）Such a large number of enzymes are usually classified using the Enzyme Commission (EC) system (Cornish-Bowden, 2014), （过渡句）the most well known numerical enzyme classification scheme, which specifies the function of an enzyme by four digits. （详细介绍这种分类方案）This classification system has a tree structure. After the root of the tree, there are two main nodes, standing for enzyme and non-enzyme proteins, respectively. The enzyme main node extends out six successor nodes, corresponding to the six main enzyme classes:（分类介绍酶的种类） (i) oxidoreductases, (ii) transferases, (iii) hydrolases, (iv) lyases, (v) isomerases and (vi) ligases, represented by the first digit. Each main class node further extends out several subclass nodes, specifying the enzyme’s subclasses, represented by the second digit. With the same logic, the third digit indicates the enzyme’s sub-subclasses and the fourth digit denotes the sub-sub-subclasses. Take Type II restriction enzyme, which is annotated as EC 3.1.21.4,（举例说明） as an example, the ‘3’ denotes that it is an hydrolase; the ‘1’ indicates that it acts on ester bonds; the ‘21’ shows that it is an endodeoxyribonuclease producing 5-phosphomonoesters; and the ‘4’ suggests that it is a Type II site-specific deoxyribonuclease. By predicting the EC numbers precisely, computational methods can annotate the function of enzymes. It should also be noted that a substantial number of enzymes annotated with some reactions in databases such as UniProt or Brenda do not have EC numbers associated, which is out of the scope of this study.

引言第二段呼应题目，为读者解答酶的EC数字系统是什么，并为下文提出的基于这种系统设计的分类器做铺垫。

（引言第三段主要介绍了目前已经开发基于酶的EC值确定酶功能的方法）

A number of computational methods have already been proposed to determine the enzyme function by predicting enzyme EC numbers. （总说主要的研究方向）There have been three main research directions of this problem since (des Jardins et al., 1997), who used machine learning methodologies and sequence information to investigate the problem for the first time. （分开说每个方向的方法）Firstly, because it is commonly believed that structures determine function, some researches….. Second, the common assumption that enzymes with high sequence similarity tend to have similar functionality leads to a number of studies utilizing sequence similarity….. Thirdly, extracting features from the sequence and classifying the enzyme using machine learning algorithms is the most extensively studied direction…...

引言的第三段没有在摘要中有所体现，主要是为了承上启下，首先先指出目前主要的研究方向，承接上段说的基于这种EC值目前存在的方法，引起下段这些目前存在的方法的不足。

（引言第四段介绍目前蛋白质预测所面临的问题）

（主题句，引出问题）In addition to those difficulties, another issue in the protein general function prediction field（特征维度不均匀影响分类精度） is the feature dimensionality nonuniformity problem, which usually lies in the sequence-lengthdependent features, （举例说明）such as PSSM (position-specific scoring matrix). For example, in this paper, the dimensionality of PSSM can range from 50 by 20 to 5000 by 20, according to the corresponding sequence length. （提出解决方法）The feature uniformity requirement of mainstream classifiers has pushed out three strategies to this problem. （分别说明方法）First, avoiding using the sequence-length-dependent features is the most straightforward solution to the problem…... The second solution is to manually derive sequence-length-independent features from the sequence-length-dependent features…... The third solution is to systematically generate sequence-lengthindependent features, such as string kernels (Dai et al., 2017, Leslie et al., 2002, 2004; Ra ¨ tsch et al., 2005; Wang et al., 2014), which, however, do not consider the classification problem when extracting features. （Despite进行转折，引出下文更好的解决方法）Despite the previous success of these three strategies, they still heavily depend on either manually designed or pre-defined features, which are most likely to be suboptimal. To take full advantage of the bursting of data in recent years, a more robust, automatic framework to extract problem-specific sequence-length-independent features from the sequence-length-dependent ones for dealing with the dimensionality problem is desired.

引言的第四段是为了说明第三段目前存在的方法的问题，第一句主题句点出问题，以一个例子说明要解决这些问题是迫在眉睫的，之后再提出可以改进的方法。

（引言的第五段对应引言第二段，详细介绍了本文所提出的方法）

（主题句，通过第四段的问题，提出更佳的解决方法）To conquer the aforementioned limitations, which are homology requirement, feature design and feature dimensionality nonuniformity, here we propose a novel level-by-level prediction approach based on deep learning, by only utilizing the sequence information. The enzyme sequence is represented by two kinds of raw encoding, sequence-length-dependent encoding, such as raw sequence one-hot encoding and PSSM, and sequence-length-independent encoding, such as functional domain (FunD) encoding. （特征编码方法）Those two kinds of raw representations are combined into a deep learning model with a novel architecture to perform dimensionality uniformization, feature selection and classification model training simultaneously. （本文的贡献）This paper makes the following contributions: （第一个创新点，分层预测结构，算法具有鲁棒性）(i) We propose a framework for hierarchical EC number prediction, the idea of which can also be applied to hierarchical classification of protein general function. （第二个创新点：特征编码算法）(ii) To solve the feature dimensionality nonuniformity problem, we propose a robust, automatic framework based on deep learning to extract problem-specific sequence-length-independent features from the sequence-length-dependent ones. （第三个创新点：新的分类器）(iii) We propose a sequence-based enzyme EC number predictor, DEEPre, which is based on the above two frameworks. (iv) Two case studies demonstrate our tool’s ability of performing functionality prediction of different enzyme isoforms caused by alternative splicing. (v) We investigate the importance of local information in determining the functionality of an enzyme.

引言的第五段是由第四段的despite转折引出的，目的是正式提出本文的方法，并对此方法进行一个详细的阐述。

小结：对于现有问题的描述篇幅太多，应该适当减少。但是段与段之间连接紧密，这是比较好的地方。

**2. Related work**

（首先介绍经典的酶功能预测方法）

2.1 EzyPred

EzyPred (Shen and Chou, 2007) is a three-level EC number predictor, （定语从句，说明这个方法的作用）which predicts whether an input protein sequence is an enzyme, and its main class and subclass if it is. （提出EzyPred所使用的特征）It uses two features, pseudo PSSM (Pse-PSSM) and FunD encoding. （分开介绍两种特征）Pse-PSSM is …… FunD encoding captures the local FunD information, which could be referred to Section 3.2.5. With these two features, （EzyPred所使用的分类器）EzyPred uses optimized evidence-theoretic k-nearest neighbor (OET-KNN) as the classifier, which is an improved version of KNN.…… （EzyPred的优点）Although having been developed for 10 years, EzyPred still remains as one of the state-of-the-art methods in predicting enzyme function. Its server is easy-to-use with a user-friendly interface as well.

2.2SVM-prot

SVM-Prot was proposed in 2004 (Cai et al., 2003, 2004) and updated in 2016 (Li et al., 2016). （介绍SVM-prot的功能）It can not only predict enzyme functional families but also non-enzyme functional families. It represents the protein sequence （使用的特征）using 13 properties, ……（所使用的分类器，以及改进之后的分类器）The original version used support vector machines (SVM) as the classifier, while the 2016 update made two more classifiers, KNN and probabilistic neural networks, available.

2.3 COFACTOR

（主题句，介绍COFACTOR）COFACTOR (Roy et al., 2012; Zhang et al., 2017) is a structurebased protein function annotation web-server. （COFACTOR的具体思想）In terms of EC number prediction, for an input structural model, which can be obtained either by experiments or computational modeling, it threads the structure against the template library, whose entries’ annotation has already been validated by experiments, to identify the template enzyme with the most similar folds and functional sites. ……（此方法的附加功能说明）In addition to enzyme function prediction, the server can predict the Gene Ontology (GO) terms and protein-ligand binding interactions as well. （实验表明了算法的优越性）COFACTOR has been proved successful in protein–ligand binding interaction prediction in the CASP9 competition (Moult et al., 2011).

2.4 EFICAz

（主题句，概括EFICAz）EFICAz (Arakaki et al., 2009; Kumar and Skolnick, 2012; Tian et al., 2004) is an EC number prediction server using combined approaches. In addition to using the sequence similarity, it also incorporates the PROSITE and PFAM database information. The original version consists of four components（EFICAz的组成，分开说明） (i) pairwise sequence comparison-based enzyme function inference, (ii) conservation controlled hidden Markov model (HMM) iterative procedure for enzyme family classification-based functionally discriminating residue identification, (iii) multiple PFAM-based functionally discriminating residue recognition and (iv) multiple PROSITE pattern recognition. （分类策略）Those four components work independently, determining the final prediction by voting. （更新之后的版本）In the later updates in 2009 and 2012, two more components, multiple PFAM family-based SVM evaluation and conservation controlled HMM iterative procedure for enzyme family classification-based SVM evaluation, and larger databases were added.……

2.5Deep learning and hierarchical classification

Since (Krizhevsky et al., 2012), deep learning has become an extremely popular machine learning method.（指出深度学习的框架） Its two main architectures, convolutional neural network (CNN) and recurrent neural network (RNN), have made a profound contribution to many bioinformatic problems, （深度学习在生物信息学的应用）such as genetic analysis (Xiong et al., 2015), sequence binding specificity prediction (Alipanahi et al., 2015), and cryo-EM image processing (Wang et al., 2016a). Instead of being a pure classifier that depends on the manually designed features such as SVM,（CNN的功能说明） CNN is ……（RNN的功能说明） RNN has ……（本文结合了CNN和RNN）In our article, we combined the advantages of CNN and RNN, using CNN to conduct feature extraction and dimensionality compression starting from the raw 2D encoding matrices, and using RNN to extract the sequential, long-term interactions within the input sequence.

（主题句，指出酶功能的预测是一个分层分类问题）A classification problem with a tree structure in the label space, such as the enzyme function prediction problem discussed in this article, is often regarded as a hierarchical classification problem. Because this kind of problems can be regarded as multi-label classification and multi-class classification at the same time, （解决这类分类问题的不同角度）the solutions to the problem can be classified into three categories based on different angles to the problem (Silla and Freitas, 2011), namely，flat classification approach, local classifier approach, and global classifier approach. （总结句，指出本文用的方法）According to the property of our problem, we chose the local classifier approach, which constructs one classifier for each internal node, to be the overall strategy.

小结：第二章主要介绍了目前主流的几种预测酶功能的方法，每一种方法占据一个段落进行总结说明，这是值得学习的地方，但是我认为可以把深度学习咋生物信息学的应用去掉或者移动到引言最后一段的部分，这样看起来更加简练。

**3. Materials and methods**

3.1Datasets

（本文使用的数据集）We adopt three datasets in this paper. （介绍第一个数据集）The first dataset is a widely used one from (Shen and Chou, 2007), constructed from the ENZYME database (released on May 1, 2007), with 40% sequence similarity cutoff. More details of that dataset could be referred to (Shen and Chou, 2007). （数据集的名称）This dataset is denoted as the KNN dataset in the rest of the paper. Following the same procedure of constructing the KNN dataset, （本文构建了新的数据集）we constructed a larger dataset using up-to-date databases. The steps of constructing the dataset are as follows:（以下是构建步骤）

（分成五点详细说明构建步骤）

i. The SWISS-PROT (released on September 7, 2016) database was separated into enzymes and non-enzymes based on the annotation.

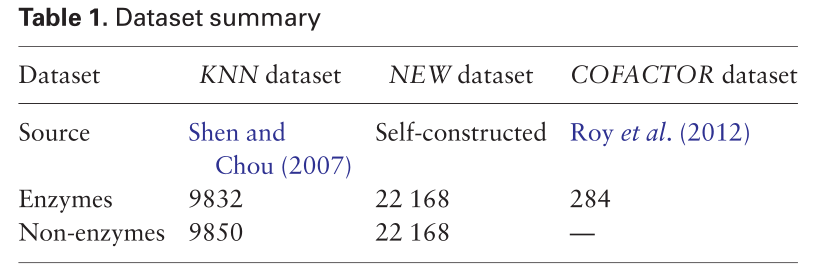
ii. To guarantee the uniqueness and correctness, enzyme sequences with more than one set of EC numbers or incomplete EC number annotation were excluded.

iii. To avoid fragment data, enzyme sequences annotated with ‘fragment’ or with <50 amino acids were excluded. Enzyme sequences with more than 5000 amino acids were also excluded.

iv. To remove redundancy bias, we used CD-HIT (Fu et al., 2012) with 40% similarity threshold to sift upon the raw dataset, resulting in 22 168 low-homology enzyme sequences.

v. To construct the non-enzyme part, 22 168 non-enzyme protein sequences were randomly collected from the SWISS-PROT (released on September 7, 2016) non-enzyme part, which were also subject to the (ii–iv) steps.

（数据集的名称）This larger dataset would be referred to as the NEW dataset in the rest of this article.

Other than KNN and NEW, which will be used as the benchmark to evaluate the proposed method based on cross-fold validation, it is also important to test the generalization power of the proposed method. This can be done by training the model on one dataset, and testing it on an independent and non-overlapping dataset, to avoid being overfitted on a particular dataset. （第三个数据集用于验证）Thus, the third dataset, the benchmark dataset from (Roy et al., 2012), is used for cross-dataset validation. （这个数据集需要满足的要求）This non-homologous dataset was collected from PDB, satisfying two requirements: (i) the pair-wise sequence similarity within the dataset is below 30%, and (ii) there is no selfBLAST hit within the dataset to ensure that there are no enzymes that are homologous to each other in this set (Roy et al., 2012). All enzymes in this dataset have experimentally determined 3D structures. To avoid overlaps between the training and testing datasets, sequences contained in both our training dataset and this dataset were removed, which reduced the size of the dataset from 318 to 284. This benchmark dataset would be referred to as the （数据集的名称）COFACTOR dataset in the following. Table 1 summarizes the three datasets.

本节主要介绍了本文使用的数据集，以及构建新数据集的详细步骤。

3.2 Sequence representation

（总结说明本文使用的特征表示方法）

The deep learning framework explained in Section 3.3 ……. （根据两种性质，分开介绍特征表示方法）Therefore, we use the following raw features, constructed from the input sequence directly, to represent the sequences. Based on their dimensionality, they can be classified into two categories, sequence-length-dependent features and sequence-length independent features. The first four features described below belong to the former while the last one belongs to the latter.

3.2.1 Sequence one-hot encoding

（主题句，说明这种特征表示获取的序列信息是什么）To preserve the original sequence information, we use one-hot encoding as the first raw representation of the input sequence. （特征编码方法的具体介绍）This encoding uses one 1 and nineteen 0 s to represent each amino acid. For example, A is encoded as 1 01... 019 ð Þ, while C is encoded as 011 02... 019 ð Þ. For each input protein sequence, （生成矩阵，维度为L\*20）the one-hot encoding would produce an L by 20 matrix……

3.2.2 Position specific scoring matrix

To provide the evolutional information to the training model, we deploy PSSM as the second sequence representation, （由别的程序获取PSSM）which was obtained through PSI-BLAST (Altschul et al., 1997) from BLASTþ (Camacho et al., 2009) with three iterations, E-value being 0.002, against SWISS-PROT (released on May 11, 2016).

3.2.3 Solvent accessibility

Solvent accessibility describes the openness of a local region. Because such information is unavailable directly from the database, we use DeepCNF (Wang et al., 2016b) to predict it. （溶剂可溶性也是使用程序直接获取）Taking the protein sequence as the input, DeepCNF outputs the possibilities of each amino acid of the sequence being in the state of buried, medium or exposed, respectively. The three states are defined by two solvent accessibility thresholds. Buried is defined as less than 10%; exposed is defined as >40%; and medium is defined within the range of 10 and 40%. （生成L\*3的矩阵）This encoding produces an L by 3 matrix. More details could be referred to (Wang et al., 2016b).

3.2.4 Secondary structure one-hot encoding

（主题句，说明蛋白质的局部折叠信息）An amino acid could be in one of the three main secondary structure states, alpha-helix, beta-sheet and random coil, which indicate the protein’s local folding information. Similar to solvent accessibility, （也是使用现成的程序直接获取）we take advantage of DeepCNF (Wang et al., 2016b) to predict the secondary structure of a given sequence, （生成L\*3的矩阵）whose result is an L by 3 matrix, （矩阵每一行的含义）each row of which shows the possibility of the amino acid folding into alpha-helix, beta-sheet or random coil, respectively. The details could be referred to (Wang et al., 2016 b).

3.2.5 Functional domain

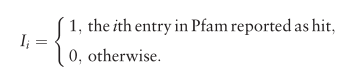
（主题句，指出蛋白质所含有的FunD代表的信息）Usually, a protein sequence contains one or several FunDs, which provide distinct functional and evolutional information. Pfam (Finn et al., 2016) is a collection of such FunDs, each represented by an HMM. Searching against the database and encoding in the following way generates the FunD encoding used in our model.（FunD编码的步骤）

i ……

ii ……

（FunD编码生成16306维向量）

As a result, the FunD encoding of a protein sequence would be:



本节主要介绍了几种本文需要使用的特征表示方法。

3.3 Classification model

（总结段，指出酶功能预测的结构和具体的方法）

The enzyme function prediction problem （酶功能预测是分层分类）has a tree-structured label space, which makes it a typical hierarchical classification problem. （每个节点建立分类器）To solve this kind of problems, we propose a level-by-level prediction framework, building a model for each internal label node. The model contains two main components, namely, the problem-specific feature extractor, which is able to perform dimensionality uniformity and feature extraction, and the classifier. （总结句，指出本文方法的优越性）Such a novel, end-toend model can perform feature selection and classifier training simultaneously in a virtuous circle, making it more likely to achieve high performance.

3.3.1 Level-by-level strategy

（提出问题）As have been discussed in the hierarchical classification part, because of the relative small size (22 168 data points are assigned to 58 classes until the second digit) and, even worse, the extreme imbalance property（举例说明） (e.g. the NEW dataset contains 22 168 sequences belonging to non-enzyme while only 10 sequences belonging to subclass 1.20) of the data,（解决上述问题的方法） we choose the local classifier approach for this problem. （与相关工作的2.5小节对应，详细介绍了分层分类）Particularly, the level-by-level prediction strategy is used.……

3.3.2 Deep neural network model

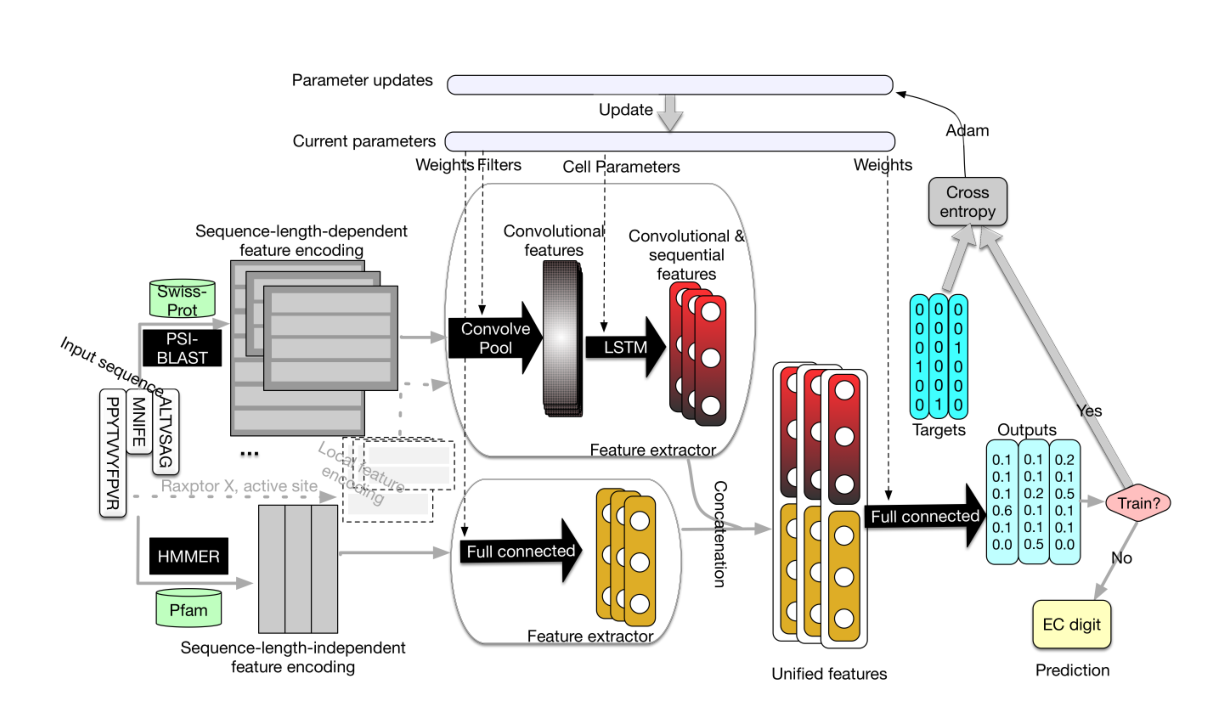
（第一段主要介绍了网络的整体架构）

（主题句，单独介绍每一级如何构造网络）For each level of prediction, we build the end-to-end model based on several deep neural network components. In terms of the sequence-length-dependent features, （针对特征表示，使用CNN和RNN提取特征）such as PSSM, we build a feature extractor exploiting the CNN component to extract convolutional features from the input map and, after that, a RNN component, comprised of long short-term memory (LSTM) cells, to extract sequential features from the output of the previous component. As for the sequence-length-independent feature,（不同于3.1小节介绍的其他特征，由于FunD维数固定，所以使用单独的技术进行特征降维） i.e. the FunD encoding, which is a vector, we use a fully connected component to perform dimensionality reduction and feature extraction. （网络的分类层由softmax构成）We employ a fully connected component to combine those different pieces of information together, followed by a softmax layer for classification. ……（深度网络的特性描述）During training, the training error is back-propagated to each component. ……

（主题句，指出第一段的模型的不足）The high complexity and flexibility of the proposed model bring high risks of overfitting. （解决方法）We adopt several methods to alleviate the problem. （分开说明每种方法）The first method is weight decay, （权重衰减）……The second method is dropout（dropout） (Srivastava et al., 2014). ……. The third method is batch normalization（批量归一化） (Ioffe and Szegedy, 2015). ……（本文提出的解决过拟合的方法）To conquer the issue, in addition to normalizing the data before inputting them in the model, we also normalize the input of each internal layer. In addition to the advantage of mitigating the overfitting problem, this manipulation would also reduce the strong dependency of knowledge-intensive initialization when training the model and allow a larger learning rate when tuning the model.

（指出了方法的优缺点，十分中肯）

（主题句，承接上段提出的问题和解决方法，给出具体的措施）We choose adaptive moment estimation (Adam) as the optimizer (Kingma and Ba, 2014), which is an improved version of stochastic gradient descent, to minimize the weighted cross entropy loss. In this way, （本文方法解决的问题）our method could handle the class imbalance issue by rescaling predictions of each class by its weight. Instead of setting the learning rate as a hyper-parameter manually as in stochastic gradient descent and momentum, this method computes the adaptive learning rate of each individual parameter by estimating the first and （此方法的缺点）second movement of the gradients at the cost of computational time and memory. Essentially, this optimizer combines the advantage of RMSprop (Tieleman and Hinton, 2012), ……

（主题句，解决数据量少的方式是迁移学习）When training the second-digit prediction models, we adopt an idea that is similar to transfer learning. Since the limited number of data is further divided into six parts corresponding to the six main classes, the amount of data belonging to each main class is insufficient to produce a model with the ability to extract features and being generalized well. （具体做法）To solve this problem, we pre-train the CNN component and the RNN component by using all the training data. Then for training each second-digit prediction model, we fix the parameters of those components and only fine tune those fully connected components using the specific subset of the training data.

（本段介绍了文中使用的是TensorFlow和电脑配置要求）In practice, we use TensorFlow (Abadi, 2016) as the framework to construct the deep neural network. With two Pascal Titan X cards, it takes around 4 h to obtain a well-trained model. In Supplementary Section S2, we provide details on setting the model parameters.

小结：第三章主要介绍了蛋白质序列的特征表示方法和预测框架，分别以小节的内容展示，条理清楚，小节中的每段上下文衔接也很紧密，值得学习。

**4. Results and discussion**

4.1Evaluation criteria

（简单介绍了评价指标）

For the enzyme or non-enzyme prediction, since it is a binary classification problem, （二分类问题所用的指标）we use accuracy, Cohne’s Kappa Score (Viera and Garrett, 2005), precision, recall and F1score to evaluate the classifiers’ performance. （多分类问题所用的指标）For other predictions, since they are multi-class classification problems, we use accuracy, Cohen’s Kappa Score, Macro-precision, Macro-recall and Macro-F1score to evaluate the classifiers’ performance, whose definitions are in Supplementary Section S3.

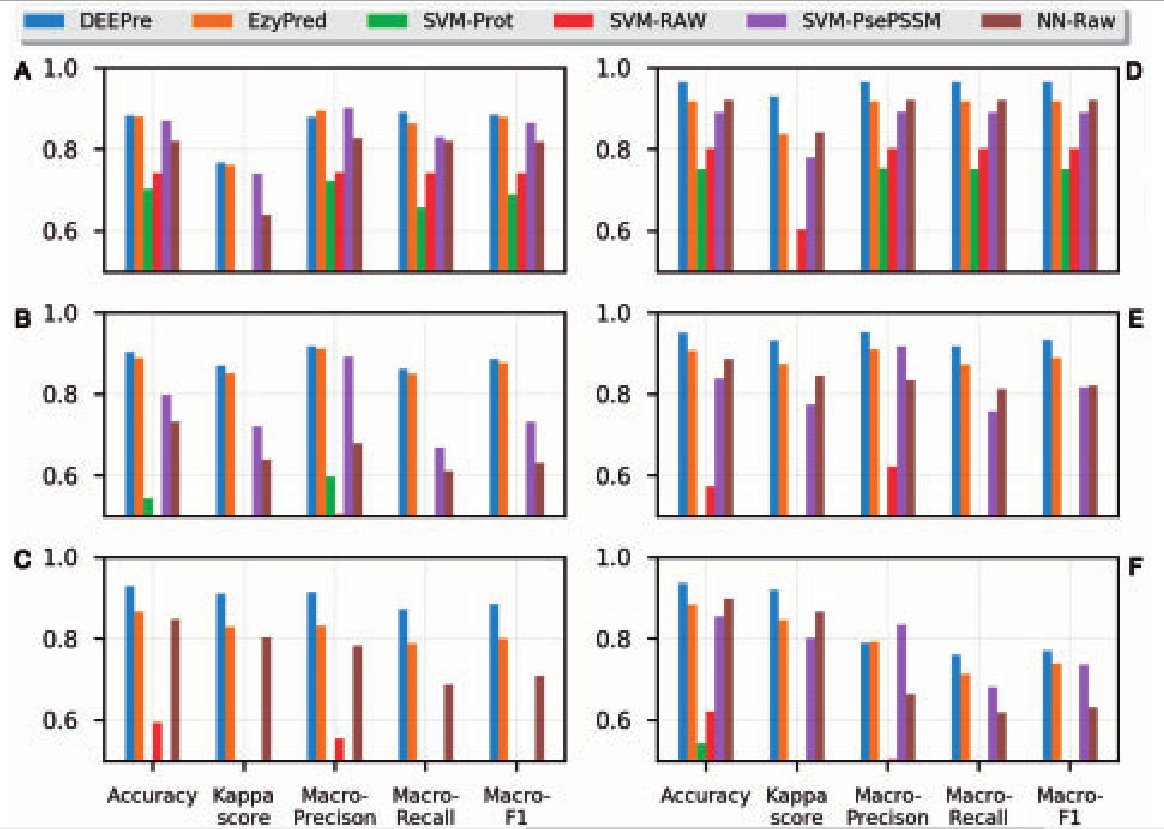
4.2Compared methods

（本节介绍了比较方法）

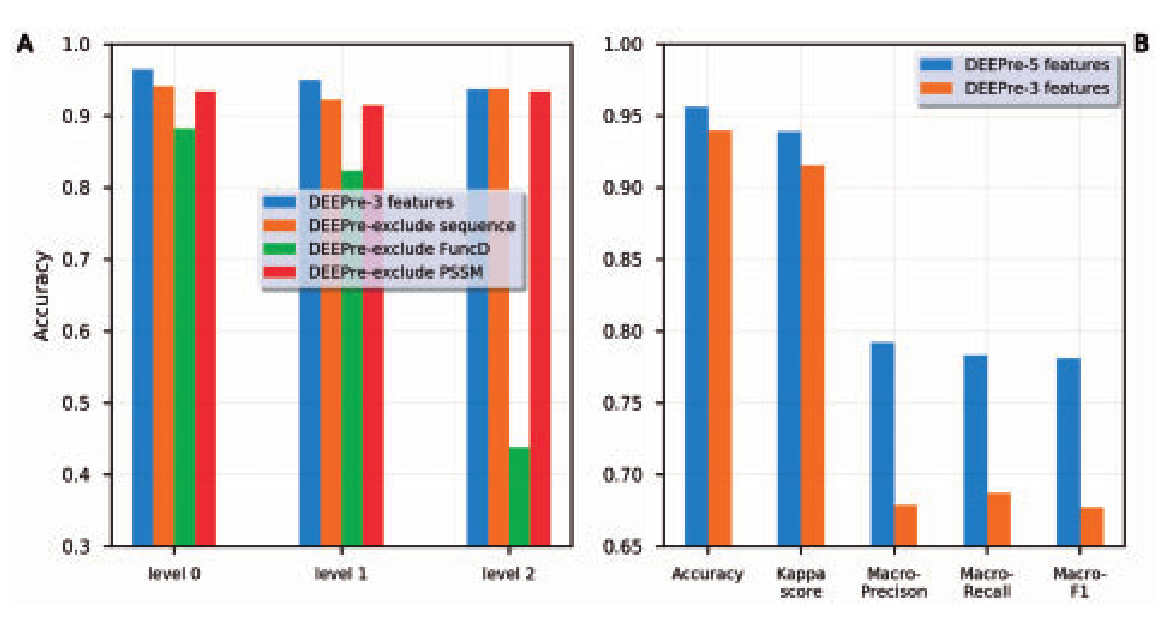
For the cross-fold validation, in which training and testing are based on different parts within the same dataset, （比较对象是五种方法）we compare our method with five other methods, （对五种方法的简单介绍）including two state-of-the-art methods, EzyPred (Shen and Chou, 2007) and SVM-Prot (Li et al., 2016), and three baseline methods. （分别介绍三种基线方法）One of the baseline methods uses SVM with the raw features used in our model; another baseline method uses SVM with Pse-PSSM; and the last baseline method uses the traditional neural network with our raw features. Due to the unchangeable database of EFICAz (Kumar and Skolnick, 2012) and COFACTOR (Zhang et al., 2017), we do not include them in the cross-fold validation comparison. However, we perform crossdataset validation, where the training and testing are performed on different datasets, to compare our method with EzyPred, SVM-Prot, COFACTOR and EFICAz.

4.3Cross-fold validation

（使用五折交叉验证）Here we report the 5-fold cross validation results, （结合图文说明实验结果）which are shown in Figure 2. （先总说本文的方法优于其他比较的方法）Our method almost always outperforms the other methods in both the KNN dataset and the NEW dataset across the five criteria and across the three hierarchical levels of prediction. （再详细说针对新数据集，本文方法在两级预测的效果都比较好）As for the NEW dataset, DEEPre outperforms the other five methods consistently in Levels 0 and 1 prediction across the five criteria. （指出没有改进的地方）As for the Level 2 prediction, the only criterion that DEEPre does not improve over the existing methods is the Macro-Precision, which is an unweighted average of precision of each label. （可能的原因）The appearance of very small classes (e.g. subclass 1.20 only has 10 enzymes) in the second level prediction might be the reason for this result. In terms of the KNN dataset, although the smaller dataset makes the improvement of DEEPre over the other methods in Level 0 prediction less significant, it still significantly outperforms the other methods in Levels 1 and 2 classification.

4.4Feature importance analysis

（总-分，主题句，指出为什么要进行特征重要度分析）It is believed that both global features and local features determine the function of a protein. For detailed function, local information would weigh even more in determining it. （全局特征的来源）The features extracted by the convolutional component and the recurrent component from PSSM and sequence raw encoding could be considered as global features （局部特征的来源）while the FunD encoding would be considered as a local feature. （比较重要度的方法）We remove the three input raw encoding one by one and show the comparison of their performance on the NEW dataset. The comparison is shown in Figure 3A. （结合图文给出结果）It is clear that as the level goes deeper, the importance of FunD is evidently increasing, which demonstrates the well-recognized hypothesis. To further prove it, we design another experiment, in which we input more local feature encodings, including secondary structure and solvent accessibility, into our model. Details of this experiment could be referred to Supplementary Section S4. （使用不同的特征，在2级预测上预测性能的比较）Figure 3B shows the performance comparison of this model and the previous model in Level 2 prediction. It is clear that the additional local features further improve the performance of our model, with accuracy improved by 1.8% while Macro-precision, Macro-recall and Macro-F1 score improved by at least 11%.

4.5Cross-dataset validation

In this experiment, we directly compare the performance of different servers in predicting the first digit and the second digit of an enzyme. （指出数据集）We use the COFACTOR benchmark dataset, which is proved to be a difficult dataset in the enzyme function prediction field (Roy et al., 2012), as the test dataset.（对数据集进行预处理） First, we eliminate the sequences in the COFACTOR benchmark data which overlap with the DEEPre’s training database (NEW) by 40% sequence similarity filtering, reducing the data size from 318 to 284,（实验结果，首先说明本文方法具有比较强的预测能力）for the first-digit prediction, DEEPre outperforms the other servers consistently across the five criteria, improving the accuracy by at least 6% over the other servers, including COFACTOR. This is significant because COFACTOR requires 3D structures of enzymes whereas DEEPre only requires the sequence information. （对于别的预测器的说明）On the other hand, we should admit that we have changed the original COFACTOR dataset ……（通过转折说明本文方法的优点）However, DEEPre is a sequencebased statistical method, ……（鲁棒性）It is worth noting that EC numbers have regular corrections, such as deletions and transfers. We check all the corrections that are related to the test enzymes in the COFACTOR dataset and find that none of them influences the comparison reported here.

4.6Third-digit and fourth-digit prediction

（说明第三位数和第四位数的结果）Using the same framework described above, we are also able to predict the enzyme’s third digit, which represents its sub-subclass, （第三位数的预测的实验结果良好）on the NEW dataset. The accuracy across all the sub-subclasses is 0.9415; the Kappa score is 0.8918; the macro-precision is 0.8942; the macro-recall is 0.8578; and the macro-F1 score is 0.8665. （指出第四位数的预测目前的不可行性）Regarding the fourth-digit prediction, more data are needed to perform normal machine learning training-and-testing procedure.（举例说明） For example, within the sub-subclass 1.1.1 in the NEW dataset, there are 188 classes. Each of those classes has <40 enzyme sequences, with 175 classes having <10 enzyme sequences. Using the current dataset with such distribution would lead to unreliable results.

4.7Case study

（针对某一种具体的酶测试预测器）

（简单介绍谷氨酰胺酶）Glutaminase is a phosphate-activated enzyme, （这种酶的作用）which catalyzes the first step of glutaminolysis, hydrolysing glutamine into glutamate (Curthoys and Watford, 1995). （生物上酶的功能）The alternative splicing of its messager RNA results in its three isoforms, with Isoforms 1 and 3 being capable of catalyzing while Isoform 2 lacking the catalytic activity (Li et al., 2017). To validate our model’s ability to distinguish the different functionality of different isoforms, we obtained the sequences of the three Glutaminase isoforms from the UniProt and put them into our model. （计算机实验上酶的功能）Our model predicted that Isoforms 1 and 3 of Glutaminase were hydrolases acting on carbon-nitrogen bonds, being consistent with the experimental results. （两者结果一致，证明了预测器的良好性能）Our model also recognized Isoform 2 as non-enzyme, which is consistent with the experimental result as well.

（简要说明极光激酶B的作用）Aurora kinases B is a key enzyme regulating chromosomal segregation during mitosis, ……（极光激酶B的形态）Aurora kinases B has five isoforms ……（为了进一步验证我们的模型处理同种型功能预测的能力）To further validate our model’s ability of handling isoforms’ functionality prediction, we collected the sequence of the five isoforms from the database and put them into our model. Our model’s result is consistent with the experimental results. Particularly, （模型结果与实验结果一致）our model predicted the functionality of the Isoform 3 successfully, despite its sequence’s large difference from the ‘canonical’ sequence.

（指出本文方法在两个个例研究上表现出的良好性能）The detailed performance comparison of different servers on these two case studies could be referred to Supplementary Section S6. Among the five compared methods, only our method and EzyPred produced correct predictions for both cases.

小结：第四章介绍了实验结果，作者从不同的角度出发，做了很多对比试验，对于目前存在的缺陷也做了举例说明，而且还额外做了特征重要度分析和个例研究，实验结果很充分，并且效果也很好，每一小节的排版也很用心，段之间很多呈现出总分的结构，值得借鉴。

**5.Conclusion**

（第一段一步步递进总结全文）

In this article, （提出新的想法）we proposed a novel end-to-end feature extraction and classifier training method for enzyme function prediction. The method proposed in this paper would force the model to learn to extract features by itself and adapt the parameters of the classifier simultaneously so that it can improve the performance in a virtuous circle. （在不同数据集上效果都良好）The thorough experiments conducted on two datasets demonstrate the high performance of our method in both a smaller dataset from 10 years ago and a larger dataset constructed half a year ago. The cross-dataset validation experiment proves the performance of our model in handling sequences with no close homologs. Although it is just a starting point,（服务器的作用） the user-friendly server, DEEPre, will provide users a good guess of enzyme function and help them set up downstream experiments. Since DEEPre predicts a score for each candidate value of a certain EC digit, it can be potentially used to detect the enzyme promiscuity (Carbonell and Faulon, 2010; Mellor et al., 2016), which means that some enzymes show multiple activities by either accepting multiple substrates or catalyzing multiple reactions. Our webserver provides the predicted scores for all candidate EC values. （本文的思想对其他问题也具有意义）In addition to providing the server in the enzyme function prediction field, we believe the idea proposed in this paper can be quite helpful in handling the feature length nonuniformity problem and the dataset evolvement in a wide spectrum of computational biology problems.

（对于特征表示的总结）Among the global features, the most important one is the FunD (Fig. 3A)……（本文思想的扩展应用）Furthermore, the robust, automatic framework based on deep learning to extract problem-specific sequence-length-independent features from the sequence-length-dependent features can also be extended to other features in addition to the features mentioned in this article.

（对未来的展望，分成两点介绍未来的工作方向）There are two directions of the future work. First, more robust methods for the fourth-digit prediction are needed. The increasing number of enzymes that have experimentally validated functions, as well as the advance in method development for learning from imbalanced-data and small samples (Maadooliat et al., 2016), provide potential solutions to the problem. Second, instead of predicting the EC numbers for enzymes, it is practically useful to predict enzymatic reactions of the enzymes. The use of reaction fingerprints, for instance, could be one viable solution for this (Segler and Waller, 2017). Another possible solution is through the use of descriptors of the reaction centers as in (Rahman et al., 2014).

小结：总结的语言很简练，第一段介绍了本文所做的贡献，第二对应前面的特征汇重要度分析，再次强调特征对于捕获蛋白质信息的重要性，第三段展望未来，篇幅分配合理。

**References**

（专著）Abadi, M. (2016) Tensorflow: Learning functions at scale. Acm Sigplan.Notices, 51, 1–1.

（专利）Carbonell, P. and Faulon, J.-L. (2010) Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics (Oxford, England), 26, 2012–2019.

（学位论文）Kingma, D. and Ba, J. (2014) Adam: A method for stochastic optimization.arXiv Preprint arXiv, 1412.6980.

总结：总体而言，这篇论文富有逻辑，条理清晰，擅长举例说明一些问题，这点值得学习。但是文章中对于基础知识的介绍太多，并且有些排版也不太好，这是需要注意的地方，本文的思想虽然很值得借鉴的，但是建议将预测框架结合图文描述的更详细一点，对于一些从别的程序中直接提取的特征可以一笔带过，不做详细说明，实验部分做的很完善，但是结论部分可以不用再次拿出单独的一段介绍特征重要度，可以在结论的第一段用一句话强调就好。这篇文章总的来说写的十分不错，篇幅安排也比较合理，不管是方法还是写作手法都很值得借鉴学习。