·数据论文·

AcidBasePred: 基于深度学习的蛋白酸碱耐受性预测平台

黄蓉[1,2#]，张鹤渐[1,2#]，吴敏[1,2]，门志月[1]，初环宇[2]，白杰[2]，常宏[2]，程健[2]，廖小平[2]，刘玉万[2]，宋亚囝[1]，江会锋[2*]

1 天津科技大学 生物工程学院，天津 300457
2 中国科学院天津工业生物技术研究所，天津 300308

黄蓉, 张鹤渐, 吴敏, 门志月, 初环宇, 白杰, 常宏, 程健, 廖小平, 刘玉万, 宋亚囝, 江会锋. AcidBasePred: 基于深度学习的蛋白酸碱耐受性预测平台[J]. 生物工程学报, 2024, 40(12): 4670-4681.
HUANG Rong, ZHANG Hejian, WU Min, MEN Zhiyue, CHU Huanyu, BAI Jie, CHANG Hong, CHENG Jian, LIAO Xiaoping, LIU Yuwan, SONG Yajian, JIANG Huifeng. AcidBasePred: a protein acid-base tolerance prediction platform based on deep learning[J]. Chinese Journal of Biotechnology, 2024, 40(12): 4670-4681.

摘　要：酶的结构和活性受环境 pH 值的影响。了解酶对极端 pH 值的适应机制并进行区分，对于阐明酶的分子机制和工业应用具有重要意义。本研究利用 ESM-2 蛋白质语言模型对最适 pH 值大于等于 9 和/或小于等于 5 的微生物的分泌蛋白进行编码，分别获得了 47 725 条和 66 079 条数据。在此基础上，本研究构建了一个基于氨基酸序列判别蛋白酸碱耐受性的深度学习模型。该模型准确率显著超过其他方法，在测试集上的整体准确率为 94.8%，精确率为 91.8%、召回率为 93.4%。同时搭建了一个 web 预测平台(https://enzymepred.biodesign.ac.cn)，用户可以直接提交酶的蛋白质序列，预测其酸碱耐受性。本研究加速了酶在生物技术、制药和化工等多个领域的应用进程，为工业酶的快速筛选与优化提供了强有力的工具。

关键词：酶；蛋白质序列；酸碱耐受性；深度学习；预测平台

资助项目：国家重点研发计划(2021YFC2103500)
This work was supported by the National Key Research and Development Program of China (2021YFC2103500).
#These authors contributed equally to this study.
*Corresponding author. E-mail: jiang_hf@tib.cas.cn
Received: 2024-03-24; Accepted: 2024-05-15; Published online: 2024-05-21

AcidBasePred: a protein acid-base tolerance prediction platform based on deep learning

HUANG Rong[1,2#], ZHANG Hejian[1,2#], WU Min[1,2], MEN Zhiyue[1], CHU Huanyu[2], BAI Jie[2], CHANG Hong[2], CHENG Jian[2], LIAO Xiaoping[2], LIU Yuwan[2], SONG Yajian[1], JIANG Huifeng[2*]

1 School of Biological Engineering, Tianjin University of Science & Technology, Tianjin 300457, China
2 Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

Abstract: The structures and activities of enzymes are influenced by pH of the environment. Understanding and distinguishing the adaptation mechanisms of enzymes to extreme pH values is of great significance for elucidating the molecular mechanisms and promoting the industrial applications of enzymes. In this study, the ESM-2 protein language model was used to encode the secreted microbial proteins with the optimal performance above pH 9 and below pH 5, which yielded 47 725 high-pH protein sequences and 66 079 low-pH protein sequences, respectively. A deep learning model was constructed to identify protein acid-base tolerance based on amino acid sequences. The model showcased significantly higher accuracy than other methods, with the overall accuracy of 94.8%, precision of 91.8%, and a recall rate of 93.4% on the test set. Furthermore, we built a website (https://enzymepred.biodesign.ac.cn), which enabled users to predict the acid-base tolerance by submitting the protein sequences of enzymes. This study has accelerated the application of enzymes in various fields, including biotechnology, pharmaceuticals, and chemicals. It provides a powerful tool for the rapid screening and optimization of industrial enzymes.

Keywords: enzyme; protein sequence; acid-base tolerance; deep learning; prediction platform

酶作为重要的生物催化剂，在医药、食品、环保和能源等多个领域均有着广泛的应用[1]。随着生物技术的不断发展，对蛋白酶性能的要求也日益提高，特别是在极酸或极碱环境下的稳定性[2]，成为评价蛋白酶质量的重要指标。然而，当前对于如何准确预测和评估蛋白酶的酸碱耐受性仍面临诸多挑战，传统的实验方法(比如酸碱滴定曲线法)不仅耗时耗力，而且成本高昂，难以满足大规模筛选和优化的需求。而利用大数据和机器学习[3]的计算方法，可以建立基于蛋白酶序列和结构信息的分类模型，从而实现高效、准确地预测蛋白酶在酸性或碱性条件下的性能表现[4]。这种方法可以加速蛋白酶的设计和优化过程，为相关领域的应用提供更好的催化剂选择。因此，开发一种高效、准确的计算预测模型显得尤为重要。

近年来，随着生物信息学和计算生物学的快速发展，基于机器学习和深度学习[5]的预测模型为这一领域带来了新的机遇。例如 Zhang 等[6]利用随机森林模型建立了一种预测酸性和碱性酶的生物信息学方法。此外，Lin 等[7]利用支持向量机的方法通过最优二肽组成来区分酸性酶和碱性酶。这些模型能够从大量的数据中提取关键特征，建立序列与蛋白质特性之间

的映射关系，从而实现对蛋白酶酸碱耐受性的准确预测。然而，由于实验数据有限并且获取成本昂贵，所以难以全面且深入地学习蛋白质序列内部间的复杂规律。

嗜酸菌的胞外酶由于可以在酸性环境下正常生长而被认为是酸性酶，嗜碱菌的胞外酶可以正常生长在碱性环境下而被认为是碱性酶[8]。本研究收集嗜酸菌和嗜碱菌的分泌蛋白作为预测模型的训练数据集，以期通过深度学习算法对酸性和碱性标签数据进行训练学习，从而得到一个能够准确预测蛋白酶酸碱耐受性的分类模型。研究还通过置信学习的方法对数据集进行清洗，构建独立验证集对模型进行验证并取得较好的预测效果。这一研究有望为工业生产和科研实践提供一种快速、准确的蛋白酶酸碱耐受性预测工具，推动生物领域的技术进步和应用发展。

1 材料与方法

1.1 微生物收集

考虑到在高 pH 环境中存活的生物体的胞外蛋白大概率是碱性蛋白，低 pH 环境中生物体的胞外蛋白应该是酸性蛋白，因此，为了获得足够的标签数据用于模型训练，本研究用生物体的最适生长 pH 值(optimal growth pH, OGpH)来作为天然分泌蛋白酸碱耐受性的标签。从 4 个来源收集不同最适生长 pH 值的生物体(主要是微生物)：(1) 在数据库 BacDive (https://bacdive.dsmz.de)官网的“Advanced search”模块中搜索并下载含有最适生长 pH 值信息的微生物名称。BacDive 是一个包含已经分离纯化得到的细菌菌株各种信息的数据库，内容涉及细菌和古菌的分类学、形态学、生理学、培养条件、来源等[9]。(2) 第 2 个数据来源是维基百科网络搜索。首先从 NCBI 中获取了所有基因组测序微生物的名称，然后使用 Python 语言写的脚本遍历这些测序微生物在维基百科[10]中的定义，如果介绍内容中包含“Acidophilic”“Alkaliphilic”等关键词，则打开该微生物的维基百科网页，人工确认是否为嗜酸或嗜碱生物，随后加入数据集。(3) 第 3 个来源是数据库 ThermoBase (http://togodb.org/db/thermobase)。它目前包含 603 种嗜热或超嗜热生物体的全面描述，除文献资源外，ThermoBase 还报告分类学、代谢、环境和生理信息，参数包括化学渗透的离子、最佳 pH 值和范围、最佳温度和范围、最佳压力和最佳盐度等[11]。类似于 BacDive，本研究在 ThermoBase 首页下载了所有微生物数据并保存具有 OGpH 信息的微生物。(4) 最后一个来源是 AciDB (https://acidb.cl)。它是一个包含约 600 种嗜酸生物的数据库，其中包括每个已测序生物的分类信息和基因组信息，例如基因组大小、G+C 含量、每种嗜酸生物体的最适生长 pH 值和最适温度等[12]。在 AciDB 网页的“Select Columns”模块中选择下拉菜单的“pH optimum”选项，即可显示所有具有最适生长 pH 值信息的微生物名称并下载。

1.2 基因收集

为了提高嗜碱生物的蛋白是碱性蛋白、嗜酸生物的蛋白是酸性蛋白的可能性，本研究收集了上述具有 OGpH 信息微生物的分泌蛋白作为模型的训练数据。首先，利用 NCBI 数据库微生物信息表获取所收集微生物的 GenBank ID 和分类信息，用于下载蛋白组序列(使用命令 ncbi-genome-download--assembly-accessions genbank_id groups--section genbank--formats protein-fasta)。接着，本研究使用 SignalP 6.0[13]检测每条蛋白质序列是否包含信号肽结构(共 5 种：Sec/SPI、Sec/SPII、Tat/SPI、Tat/SPII 和 Sec/SPIII)。SignalP 6.0 可以用来预测来自古细菌、革兰氏阳性细菌、革兰氏阴性细菌和真核

生物的蛋白质中是否存在信号肽。随后，使用TMHMM 2.0[14]工具预测蛋白质中是否包含跨膜结构。如果蛋白质中含有信号肽结构而不包含跨膜结构或者只包含 1 个跨膜结构并且位于蛋白质序列前 60 个氨基酸位点中，则认为该蛋白是分泌蛋白。最终，本研究只保留分泌蛋白的序列信息及其微生物的 OGpH 信息作为模型的训练数据集。

1.3 模型构建

本研究的目标是判别蛋白酶的酸碱耐受性，因此从上述分泌蛋白中按照其微生物的 OGpH 信息提取具有酸性标签和碱性标签的两种数据集，即提取 OGpH 小于等于 5 的微生物分泌蛋白作为低 pH 蛋白数据(low pH protein, LpHP)，OGpH大于等于9的微生物分泌蛋白作为高 pH 蛋白数据(high pH protein, HpHP)，将这些数据以及对应的 2 种标签作为初步的数据集。在构建模型之前，对数据集进行简单的预处理。鉴于实际应用中使用大多数酶的长度，保留长度大于等于 100 并小于等于 1 000 个氨基酸的蛋白质序列。为提高数据质量并保证蛋白质序列能够被正确编码，剔除包含非 20 种常见氨基酸字符或连续 3 个未知氨基酸(比如“XXX”)的序列。本研究在2种标签的数据集中分别随机选择 80%的蛋白质序列作为训练集，剩下 20%作为测试集以供模型进行训练。

ESM-2[15-16]是当前领先的蛋白质语言模型，它基于 Transformer 架构在蛋白质序列上进行训练。虽然 Transformer 架构原本广泛应用于自然语言处理(natural language processing, NLP)任务，但由于氨基酸序列可被视作一种特定语言，Transformer 架构同样被成功地应用于解决与蛋白质相关的生物学问题。该架构中的注意力机制能够捕捉蛋白质的折叠结构、结合位点以及复杂的生物物理特性。ESM-2 经过训练后，可将蛋白质序列编码为具有特定长度的向量特征表示，其中隐含了蛋白质的二三级结构、功能及同源性等信息。受自然语言处理技术的启发，本研究采用迁移学习策略，利用预训练的蛋白质模型 ESM-2 的最后一层平均嵌入向量来获取蛋白质序列的向量特征表示。这些特征表示随后作为输入传递给二分类模型，以进行蛋白质酸碱耐受性分类。本研究构建了一个包含三层全连接层的神经网络：1 280 个输入特征映射至 64 个神经元、64 个神经元映射至 16 个神经元、16 个神经元映射至 2 个输出节点，用于执行最终的分类任务，即预测输入蛋白质序列的酸碱耐受情况。

在每个全连接层之后，都引入了批量归一化层和 ReLU 激活函数，以增强模型的稳定性和非线性表达能力。此外，为了预防过拟合，在第一和第二个隐藏层后分别添加了 Dropout 层。在构建蛋白质酸碱耐受性分类模型时，采用了交叉熵损失函数，以量化模型预测的概率分布与真实标签之间的差异。模型参数的优化则选择了 Adam 优化器，训练的初始学习率设定为 1×10^{-4}，并采用 StepLR 学习率调度策略逐步降低到 1×10^{-6}。这种逐步降低学习率的策略有助于模型在训练初期快速收敛，并在训练后期通过更精细的参数调整来避免过拟合，从而提升模型的泛化能力。为了确保训练过程的稳定性和模型的充分训练，模型在整个数据集上进行了 250 次迭代训练。整个模型在 Pytorch 框架中完成并实现训练过程。

1.4 模型性能评价

本研究计算每个模型的准确率、精确率、召回率和 F_1 值，采用这些指标来评价模型的预测性能。其中准确率是评估模型预测正确与否的基本指标，它计算的是所有样本中被正确预测的样本比例，计算公式(1)如下：

$$\text{Accuracy}=\frac{T_{HpHP}+T_{LpHP}}{T_{HpHP}+T_{LpHP}+F_{HpHP}+F_{LpHP}} \quad (1)$$

其中，T_{HpHP}是HpHP中被预测为HpHP的数量，T_{LpHP}是LpHP中被预测为LpHP的数量，F_{HpHP}是LpHP中被预测为HpHP的数量，F_{LpHP}是TpHP中被预测为LpHP的数量。然而，当面对各类别的数据量不平衡时，准确率可能无法真实反映模型的性能，因此需要结合其他指标进行综合评估。精确率关注的是模型预测为正样本的实例中，真正为正样本的比例，计算公式(2)如下：

$$\text{Precision}=\frac{T_{HpHP}}{T_{HpHP}+F_{HpHP}} \quad (2)$$

召回率则衡量的是所有真正为正样本的实例中，被模型正确预测为正样本的比例，按公式(3)计算：

$$\text{Recall}=\frac{T_{HpHP}}{T_{HpHP}+F_{LpHP}} \quad (3)$$

F_1分数则是一个融合精确率和召回率的综合指标，它能够在两者之间进行权衡，按公式(4)给出更全面的评价：

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (4)$$

最终本研究保留F_1分数最高的模型以供进一步使用。

1.5 独立验证集测试

为了评估模型的预测性能，本研究从数据库BRENDA (https://www.brenda-enzymes.org/)中提取所有最适pH值小于等于5或者大于等于9的蛋白酶。BRENDA是一个基于原始文献的酶功能和分子信息的综合关系数据库，保存了来自6 900多种不同生物体的至少40 000种不同酶的数据[17]。对于具有明确酸性和碱性标签的蛋白酶，剔除可以催化不同反应即具有多种最适pH值的酶，然后在UniProt中下载它们的蛋白质序列。按照训练数据集的处理方式，研究保留长度在100到1 000个氨基酸之间的序列，并删除包含非20种常见氨基酸字符或连续3个未知氨基酸的序列。为了减少同源偏差和冗余，使用MMseqs2[18]进行聚类，使独立验证集之间的序列相似性不超过20%。最重要的是，研究删除了在训练集中存在的序列，这使得对模型的验证更加可信。最后，本研究通过该独立验证集对上述保存的最优模型以及现有的其他酸碱预测工具进行评估和测试。

2 结果与分析

2.1 数据收集

本研究从上述4个途径中共收集到7 078种微生物(图1A)：其中BacDive中包含5 847种具有最适生长pH值信息的微生物；从Wikipedia中得到51个嗜酸菌和57个嗜碱菌；在ThermoBase数据库的统计表中得到523种微生物及其最适生长pH值；最后是AciDB数据库中的600种嗜酸生物。为了训练酸碱耐受性分类模型，研究只保留HpHP和LpHP数据作为模型的训练数据集，其中OGpH大于等于9的微生物有273种，OGpH小于等于5的微生物有835种(图1B)。本研究从NCBI基因组数据库中提取相应的蛋白质序列，其中基因组测序的微生物只有659种，共检索到约200万条相应的蛋白质序列。研究使用SignalP和TMHHM工具预测这些蛋白的信号肽结构和跨膜结构(图1C)，得到共153 278条分泌蛋白，其中OGpH小于等于5的分泌蛋白有99 530条，OGpH大于等于9的分泌蛋白有53 748条(图1D)。然后根据蛋白质序列长度和内容的限制条件对其进行过滤，最终得到47 725个HpHP和66 079个LpHP作为训练模型的数据集。

2.2 数据分析

为了研究模型对数据集特征的提取和识别

情况，本研究分别从两个标签数据集中随机抽取 10%的序列，利用平均嵌入向量的主成分分析(principal components analysis, PCA)[19]检测 ESM-2 嵌入对酸性和碱性蛋白酶分类的潜在适用性(图 2A)。此外，研究利用t-分布邻域嵌入算法 (t-distributed stochastic neighbor embedding, t-SNE)[20]进行特征可视化(图 2B)。从整体来看，不同类别对应的点明显分离，这表明模型在数据集上很好地提取到了易于区分类别的特征，证明了模型在分类任务中的有效性。

2.3 模型训练

将最终得到的数据集按照 8:2 的比例分别作为训练集和测试集结果如图 3A 所示，蛋白质序列通过 ESM-2 编码后得到 1 280 维向量数据并将其输入到多层感知机神经网络模型中，研究使用交叉熵损失函数[21]优化模型。经过 250 轮训练后，损失值变得稳定，模型结束训练。结果显示，模型在测试集上的总体准确率为 93.7%，表明测试集中的大多数蛋白质被正确分类为 HpHP 或 LpHP。考虑到训练数据

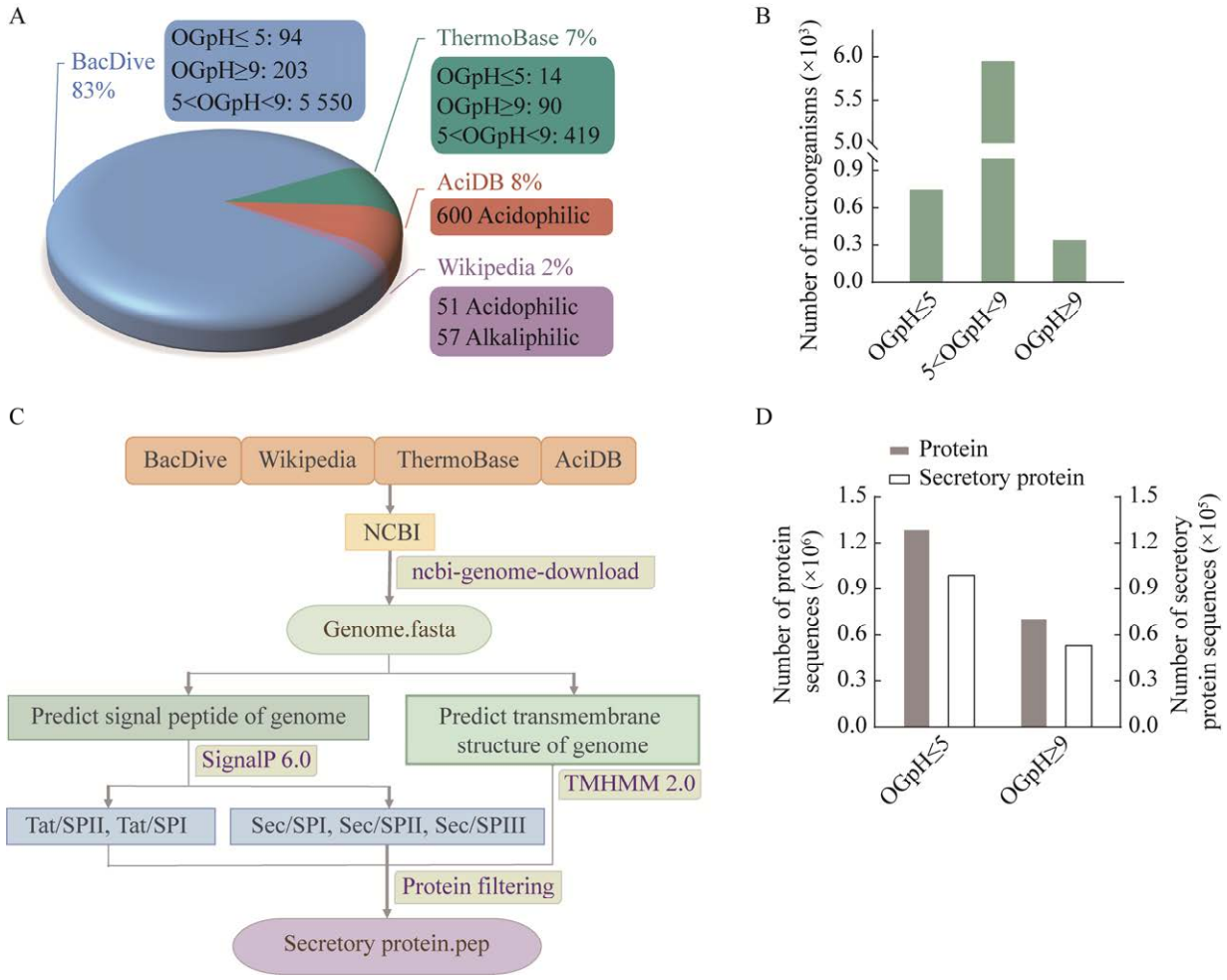

图 1 数据收集 A：微生物的来源. B：微生物的收集. C：提取分泌蛋白的流程. D：蛋白和分泌蛋白的收集

Figure 1 Data Collection. A: Source of microorganisms. B: Collection of microorganisms. C: The process of extracting secretory proteins. D: Collection of proteins and secretory proteins.

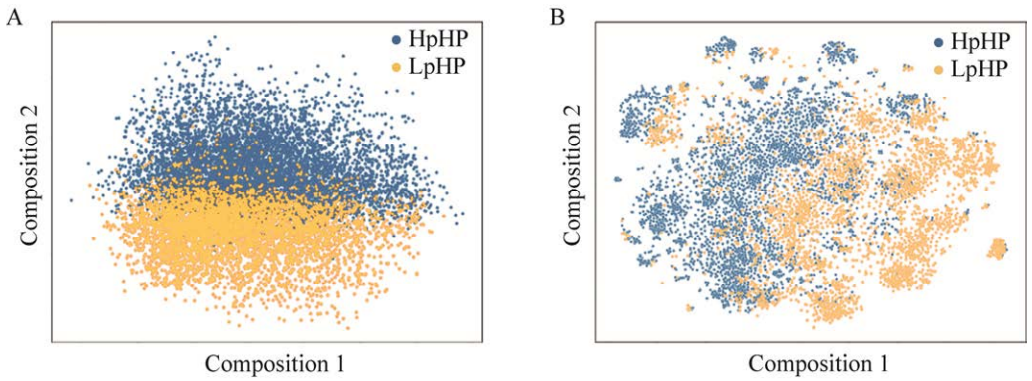

图 2　数据分析　A：利用 PCA 进行数据降维分析. B：利用 t-SNE 进行数据特征可视化.

Figure 2　Data analysis. A: Dimensionality reduction analysis was performed using PCA. B: Data feature visualization was conducted using t-SNE.

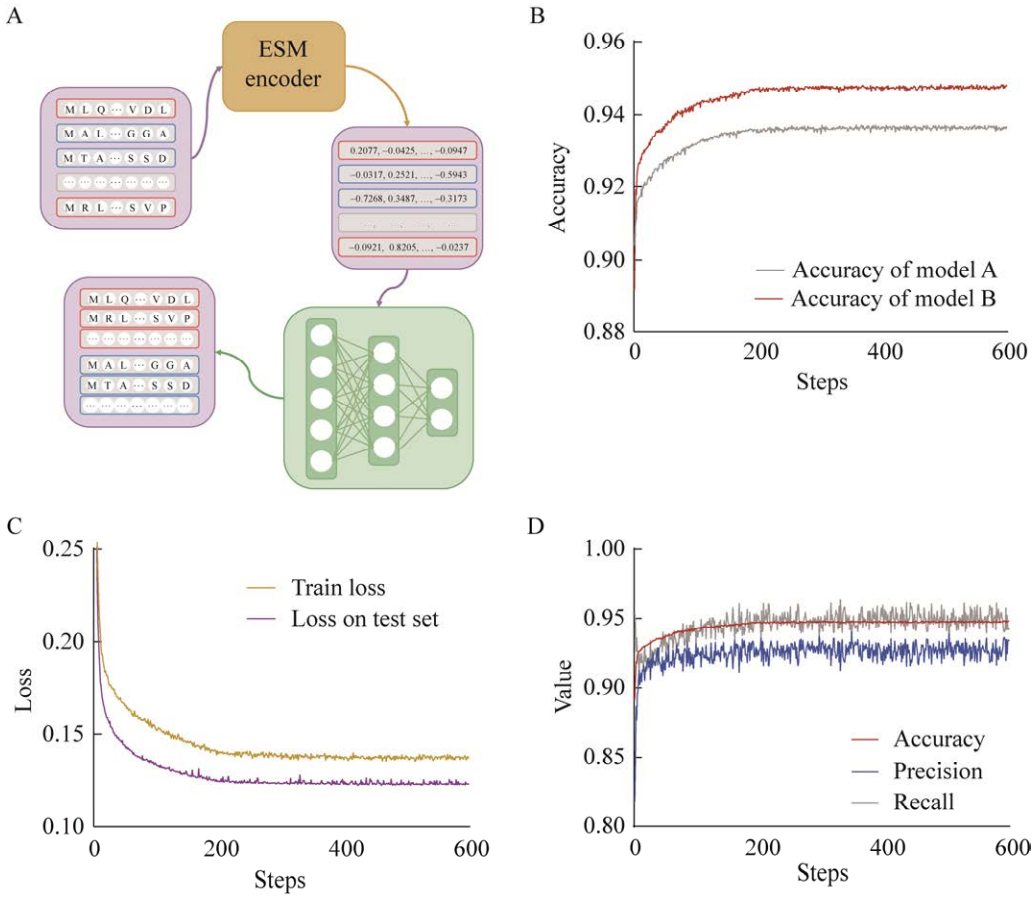

图 3　模型的构建和训练　A：模型的框架图. B：数据清洗对模型准确率的影响. C：模型训练过程中的损失和在测试集上的损失. D：模型训练过程中在测试集上的准确率、精确率和召回率

Figure 3　Model construction and training. A: Frame diagram of the model. B: The impact of data cleaning on the accuracy of the model, where A is the model before data cleaning, and B is the model after data cleaning. C: The train loss and loss on test set during the training of the model. D: Accuracy, precision and recall of test set during the training of the model.

集的不平衡性(58.0%属于 LpHP)，计算了精确率和召回率，以进一步评估模型预测蛋白酶酸碱性的性能。结果显示，所有被预测为 HpHP 的蛋白质序列中有 91.8% (精确率)确实是 HpHP，而模型预测为 LpHP 的所有序列中有 93.4% (召回率)实际上也是 LpHP (表 1)。这些结果表明，本研究构建的酸碱耐受性模型可以有效识别 HpHP 和 LpHP 数据。

2.4 数据清洗

数据的标签错误会导致模型预测效果存在潜在的错误问题。本研究使用工具 Cleanlab[22] 对训练数据集进行了清洗，根据模型预测标签和实际标签的情况识别可能存在问题的数据，最终发现训练集中至少存在 684 个错误标签，测试集中至少存在 1 536 个错误标签，错误率达到2.0%。随后，删除这些标注错误的数据，使用上述相同的参数重新训练判别模型，并保存 F_1 分数最高的模型，再通过独立验证集对最优模型进行评价。相比清洗数据之前的模型，最优模型在测试集上的准确率有明显的提升，从 93.7%上升到 94.8% (图 3B)，loss 值由 0.162 降低到 0.123。图 3C 是清洗数据后模型在训练过程中的损失和在测试集上的损失，经 250 轮后皆趋于稳定，模型训练结束。图 3D 是模型在测试集上的评估指标，包括准确率、精确率和召回率。这些指标在训练过程中逐步上升，最后趋于稳定，且数值均达到 90.0%以上，说明该模型在训练数据集上表现良好。

表 1 蛋白酶酸碱耐受性判别模型总结

Table 1 Summary of proteinase acid-base tolerance discrimination model

		Training set	Testing set
Data	HpHPs	38 180	9 545
	LpHPs	52 863	13 216
Metrics	Accuracy (%)	93.0	93.7
	Precision (%)	91.2	91.8
	Recall (%)	92.4	93.4

Model parameter description: The batch size is set to 64, and the model tends to stabilize after 250 rounds of training.

2.5 独立验证集测试

从 BRENDA 数据库中共收集到 1 062 个已知酸碱性标签的蛋白酶，剔除在训练集中出现过的蛋白质序列后按 40.0%的相似度聚类，得到 286 个酸性蛋白酶和 354 个碱性蛋白酶。为了平衡两种标签的数据，随机删除 68 个碱性蛋白酶使其与酸性蛋白酶的数量相等。使用 ESM-2 将其编码成 1 280 维向量数据后输入到最优模型中预测它们的酸碱耐受性，并统计预测的结果。结果显示，模型可以正确识别 198 条酸性蛋白酶和 133 条碱性蛋白酶，即该模型在独立验证集上的准确率达到 66.7%，而模型预测为碱性的蛋白酶中有 53.6%确实是碱性，预测为酸性的蛋白酶中有 79.8%实际上是酸性。

2.6 AcidBasePred 预测平台

为方便用户使用，本研究还搭建了一个开源的酸碱耐受性 web 预测平台 AcidBasePred，其网址为 https://enzymepred.biodesign.ac.cn。如图 4A 所示，用户可以以蛋白质序列和文件的形式提交批量查询序列，后台将查询序列经 ESM-2 编码为向量表示文件后输入构建好的酸碱耐受性预测模型并判断蛋白的酸碱耐受性。结果页面中可以查看所有用户提交的任务及其内容(图 4B)，出于对数据隐私和安全性的考虑，用户可以选择将预测结果保留在本地后删除网络上的任务记录。由于资源限制，提交查询序列后需要等待几分钟，查询序列越多等待时间相应增加。如果有多个用户提交请求，后台会按照提交时间排队进行处理。结果如图 4C 所示，显示每条查询序列的酸碱性预测情况。

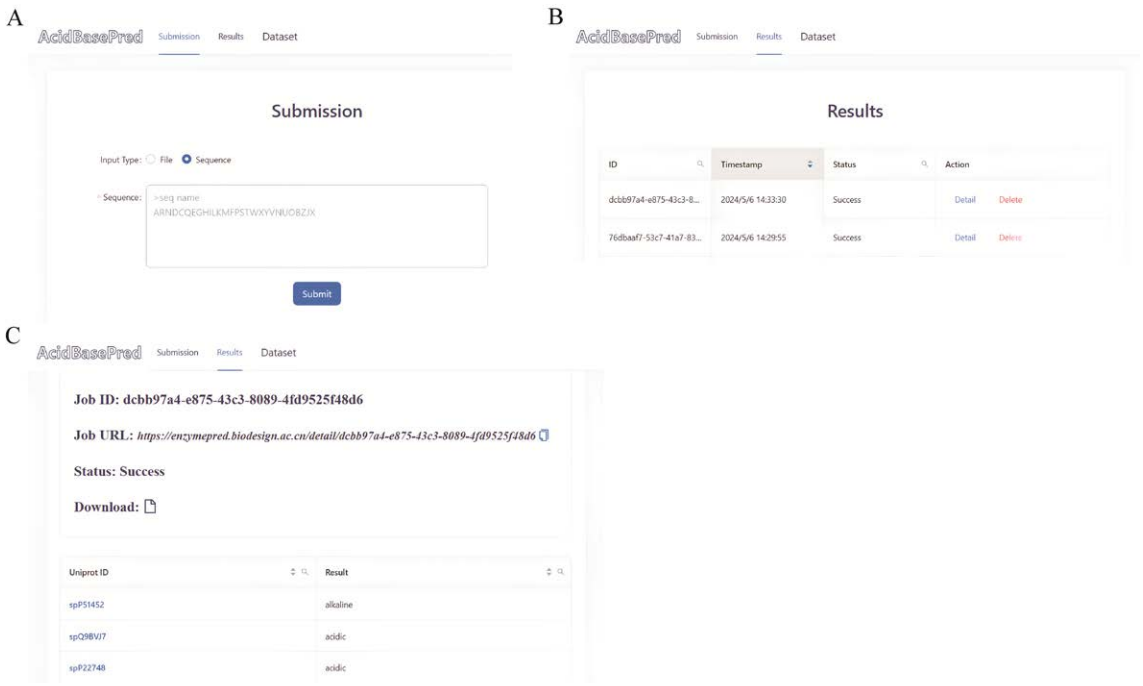

图 4　蛋白酶酸碱耐受性判别模型线上预测平台　A: 首页. B: 结果页面. C: 预测结果

Figure 4　An online platform for the discrimination model of protease acid-base tolerance. A: Homepage. B: Result page. C: Prediction results.

3　讨论与结论

本研究利用微生物最适生长 pH 值的数据集构建了一个基于预训练蛋白质语言模型 ESM-2 的多层感知机模型，以此来判别输入蛋白质序列的酸碱耐受性。为提高模型的可靠性，研究选择分泌蛋白作为模型训练的数据集。本研究采用准确率、精确率等指标对训练的模型进行性能评估，并利用置信学习的方法对潜在的标签错误数据进行清洗，以达到优化模型的目的。最终，优化后的模型在测试集上展现出了高达 94.8%的准确率。为验证模型的泛化能力，还使用了独立验证集进行测试，结果显示模型在该验证集上达到了 66.7%的准确率。

Zhang 等首次利用计算模型区分酸性酶和碱性酶[6]，他们提出了一种基于随机森林的方法，通过对样本进行二级结构特征编码，将准确率提高到 90.7%。然而，准确率仍然不够理想。此外，如果不能正确预测蛋白质的二级结构，就会为进一步的酸碱性酶描述提供错误的信息。2013 年，Fan 等[23]将 AAC、GO、PSSM、ACS、RAAA 等向量组合，构建了基于支持向量机的预测模型；该模型中，准确率、精确率和灵敏度分别达到 93.1%、94.6%和 91.4%。然而，该方法的预测器需要蛋白质在 GO 数据库存在注释信息，并且有研究表明 UniProt 中只有不到 50.0%的蛋白质含有 GO 信息[7]。此外，如果不能在搜索数据集中找到蛋白质的同源序列，则 PSSM 信息也存在不足，从而导致错误的预测。为了提高数据的有效性，Lin 等[7]将数据集的序列一致性降低到 25.0%，提出了一种基于支持向量机的模型，因为蛋白质二级结构中与氢键相关的残基之间具有深度相关性，所以他们通过使用 g-gap 二

肽组成对样本进行编码，获得 96.0%的准确率。2015 年，Khan 等[24]构建了一个 PNN 模型，在第一层计算训练样本和测试样本之间的距离，第二层使用 RBF 函数计算测试样本的概率值，第三层是不同类别的概率值之和，最后一层选择概率值最高的类别作为被测样本的预测类别；该模型输入特征是 SAAC 和 PseAAC 的组合，这两类特征提供了蛋白质的序列、位置和理化性质，对数据集产生了 96.3%的准确率；然而，由于该方法所使用的数据集仅包含 217 条蛋白质序列，远远不足以覆盖蛋白质的复杂空间，因此模型的泛化能力较弱。而本研究收集了数十万条蛋白质序列作为模型的训练数据集，并利用预训练的蛋白质语言模型对数据集进行编码，有助于更好地理解和学习高维蛋白质序列空间中的复杂特征，提高预测模型的准确率和泛化能力。其次，研究使用蛋白质的一级序列信息作为模型的输入，避免了预测二级结构潜在的错误以及同源序列不存在的问题。

为了证明本研究中蛋白酶酸碱耐受性判别模型的可靠性，将其与已发表的方法进行比较。由于只能获得 Lin 等[7]开发的 AcalPred 预测工具的训练数据集且只有在它的网络服务器可以正常使用，所以通过相同独立验证集将它和本研究方法进行比较。为了提高可信度，剔除了独立验证集中存在于 AcalPred 数据集或本研究模型数据集的蛋白质序列。同样，将酸性蛋白酶和碱性蛋白酶控制在相同数量，各 212 个。本研究构建的模型在独立验证集上测试得到的准确率为 65.1%，而 AcalPred 的方法在相同数据集上的准确率是 60.6%。因此，本研究方法的准确率在独立测试集上要比 AcalPred 高 4.5%。

由以上结果可见，本研究模型的优异表现可以归结为 2 个方面的原因。首先，本研究拥有庞大的数据量。AcalPred 预测模型的数据来源于 BRENDA 数据库的蛋白质序列，数据来源单一，且数据量十分有限。而本研究模型收集到的数据量多达 10 万余条蛋白。这可以帮助模型更好地捕捉数据中潜在的分布和特征，也可以提高模型的泛化能力。此外，研究利用多种工具提取极端环境中微生物的胞外分泌蛋白，提高数据质量，降低本研究预测模型的错误率。其次，其他模型是通过侧重于某一种或某几种蛋白质特性进行编码学习，而本研究则使用了大型蛋白质语言模型编码的表示向量作为初始输入来训练酸碱耐受性预测模型。该表示向量隐含了蛋白质的二三级结构、功能以及同源性等信息，已经被证明在下游预测任务的潜在适用性。基于以上原因，本研究提出的方法在预测效果上优于先前的方法。

虽然本研究模型在识别 HpHP 和 LpHP 的工作上具有较好的效果，但还是存在一定的局限性。首先，数据集可能存在噪声或错误标注的情况，导致模型学习到错误的模式或规律，从而产生误差。其次，数据的收集以微生物为单位，样本分布可能不均匀，并且数据集中酸性和碱性样本的数量存在不平衡，这会影响模型在整体数据上的表现以及限制模型对蛋白质特征的泛化能力。在实际应用中，蛋白质的结构和功能复杂多样，该研究模型在处理多样性蛋白上可能会面临一定的挑战。此外，如何选择合适的模型超参数以达到最好的训练效果也是一个挑战(如学习率、批量大小、优化器、激活函数等)。展望未来，通过扩大数据来源范围，有望获得更加全面和系统的认识，这将有助于深入揭示不同生物体内蛋白酶适应环境的机制。运用生物信息学和统计学的分析方法，对搜集到的原始数据进行审查、纠错与标注，为模型的训练提供更可靠、有效的数据支持。另

外，可以将深度强化学习应用于超参数优化，通过代理模型来动态调整以最大化目标性能指标，从而提高模型效果。另一方面，模型在测试集和独立验证集上的准确率差异较大，一个可能的原因是独立验证集中序列的物种来源有一部分并没有包含在训练数据集中，导致模型对于这些没有见过的数据预测效果较差。还有一部分原因是测试集中仅包含分泌蛋白，而独立验证集中分泌蛋白的比例只占 24.0% (137 条)，模型在这些蛋白上的准确率提升为 72.1%。

目前，由于缺乏对该模型的生物学湿实验验证结果以及模型的可解释性不足，导致用户对预测结果的理解和信任度受到限制。这在一定程度上阻碍了模型的广泛应用和进一步的发展。未来需要将本研究的方法应用于生物学实验，将计算方法与湿实验验证相结合，进一步确保该预测平台的可靠性。此外，可以通过强化数据加密和隐私保护机制，实施访问控制和身份验证机制，定期进行安全审计和漏洞修复来增加用户信任度。还要不断改进优化 web 界面的美观程度和功能模块等，以提升用户的使用体验。

总之，本研究为深度学习在预测蛋白酶酸碱耐受性方面的应用拓展了新的可能，为相关领域的科学研究和应用发展提供了有益的启示。随着技术的不断进步和研究的深入，相信本研究训练得到的酸碱耐受性预测模型将有助于推动生物信息学和蛋白质工程领域的进一步发展。

REFERENCES

[1] TRENDELENBURG U. The interaction of transport mechanisms and intracellular enzymes in metabolizing systems[J]. Journal of Neural Transmission. Supplementum, 1990, 32: 3-18.

[2] YAMAGATA Y, MAEDA H, NAKAJIMA T, ICHISHIMA E. The molecular surface of proteolytic enzymes has an important role in stability of the enzymatic activity in extraordinary environments[J]. European Journal of Biochemistry, 2002, 269(18): 4577-4585.

[3] GREENER JG, KANDATHIL SM, MOFFAT L, JONES DT. A guide to machine learning for biologists[J]. Nature Reviews Molecular Cell Biology, 2022, 23: 40-55.

[4] LEE BJ, SHIN MS, OH YJ, OH HS, RYU KH. Identification of protein functions using a machine-learning approach based on sequence-derived properties[J]. Proteome Science, 2009, 7: 27.

[5] DING WZ, NAKAI KT, GONG HP. Protein design *via* deep learning[J]. Briefings in Bioinformatics, 2022, 23(3): bbac102.

[6] ZHANG GY, LI HC, FANG BS. Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition[J]. Process Biochemistry, 2009, 44(6): 654-660.

[7] LIN H, CHEN W, DING H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes[J]. PLoS One, 2013, 8(10): e75726.

[8] HOUGH DW, DANSON MJ. Extremozymes[J]. Current Opinion in Chemical Biology, 1999, 3(1): 39-46.

[9] REIMER LC, SARDÀ CARBASSE J, KOBLITZ J, EBELING C, PODSTAWKA A, OVERMANN J. BacDive in 2022: the knowledge base for standardized bacterial and archaeal data[J]. Nucleic Acids Research, 2022, 50(D1): D741-D746.

[10] ZIMMER C, LEUBA SI, YAESOUBI R, COHEN T. Use of daily Internet search query data improves real-time projections of influenza epidemics[J]. Journal of the Royal Society, Interface, 2018, 15(147): 20180220.

[11] DiGIACOMO J, McKAY C, DAVILA A. ThermoBase: a database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms[J]. PLoS One, 2022, 17(5): e0268253.

[12] NEIRA G, CORTEZ D, JIL J, HOLMES DS. AciDB 1.0: a database of acidophilic organisms, their genomic

information and associated metadata[J]. Bioinformatics, 2020, 36(19): 4970-4971.

[13] TEUFEL F, ALMAGRO ARMENTEROS JJ, JOHANSEN AR, GÍSLASON MH, PIHL SI, TSIRIGOS KD, WINTHER O, BRUNAK S, von HEIJNE G, NIELSEN H. SignalP 6.0 predicts all five types of signal peptides using protein language models[J]. Nature Biotechnology, 2022, 40(7): 1023-1025.

[14] KROGH A, LARSSON B, von HEIJNE G, SONNHAMMER EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes[J]. Journal of Molecular Biology, 2001, 305(3): 567-580.

[15] RIVES A, MEIER J, SERCU T, GOYAL S, LIN ZM, LIU J, GUO DM, OTT M, ZITNICK CL, MA J, FERGUS R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[16] LIN ZM, AKIN H, RAO R, HIE B, ZHU ZK, LU WT, SMETANIN N, VERKUIL R, KABELI O, SHMUELI Y, dos SANTOS COSTA A, FAZEL-ZARANDI M, SERCU T, CANDIDO S, RIVES A. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[17] CHANG A, JESKE L, ULBRICH S, HOFMANN J, KOBLITZ J, SCHOMBURG I, NEUMANN-SCHAAL M, JAHN D, SCHOMBURG D. BRENDA. The ELIXIR core data resource in 2021: new developments and updates[J]. Nucleic Acids Research, 2021, 49(D1): D498-D508.

[18] STEINEGGER M, SÖDING J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets[J]. Nature Biotechnology, 2017, 35(11): 1026-1028.

[19] DAVID CC, JACOBS DJ. Principal component analysis: a method for determining the essential dynamics of proteins[J]. Methods in Molecular Biology, 2014, 1084: 193-226.

[20] DIMITRIADIS G, NETO JP, KAMPFF AR. T-SNE visualization of large-scale neural recordings[J]. Neural Computation, 2018, 30(7): 1750-1774.

[21] GUO C, CHEN XN, CHEN YH, YU CY. Multi-stage attentive network for motion deblurring *via* binary cross-entropy loss[J]. Entropy, 2022, 24(10): 1414.

[22] NORTHCUTT C, JIANG L, CHUANG I. Confident learning: estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-1411.

[23] FAN GL, LI QZ, ZUO YC. Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC[J]. Process Biochemistry, 2013, 48(7): 1048-1053.

[24] KHAN ZU, HAYAT M, KHAN MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model[J]. Journal of Theoretical Biology, 2015, 365: 197-203.

(本文责编　陈宏宇)