具有深度蛋白质语言模型的无偏生物体不可知且高度灵敏的信号肽预测器

Junbo Shen, Qinze Yu, Shenyang Chen, Qingxiong Tan, Jingchen Li, and Yu Li
†1,2,3,6,7,8

[1] 中国香港特别行政区香港中文大学计算机科学与工程系
[2] The CUHK Shenzhen Research Institute, Hi-Tech Park, Nanshan, Shenzhen, 518057, China
[3] 上海人工智能实验室，上海，中国
[4] 华盛顿大学计算机科学与工程系，圣路易斯，MO 63130，美国
[5] 佐治亚理工学院，亚特兰大，GA 30332，美国
[6] 麻省理工学院医学工程与科学研究所，美国马萨诸塞州剑桥
[7] 哈佛大学维斯仿生工程研究所，美国马萨诸塞州波士顿
[8] 麻省理工学院和哈佛大学布罗德研究所，美国马萨诸塞州剑桥

抽象的

信号肽（SP）是位于蛋白质N末端的短肽。将跨膜蛋白和分泌蛋白靶向并转移到正确的位置至关重要。与识别信号肽的传统实验方法相比，计算方法更快、更高效，对于分析数千甚至数百万个蛋白质序列，尤其是宏基因组数据更加实用。最近提出了计算工具来对信号肽进行分类并预测切割位点位置。然而，他们中的大多数人忽视了这些任务中极端的数据不平衡问题。此外，几乎所有这些方法都依赖于蛋白质的额外组信息来提高其性能，然而，这些信息可能并不总是可用的。为了解决这些问题，我们提出了无偏生物体不可知信号肽网络（USPNet），这是一种利用蛋白质语言模型的信号肽分类和切割位点预测深度学习方法。我们建议应用标签分布感知的边缘损失来处理数据不平衡问题，并使用蛋白质的进化信息来丰富表示并克服物种信息依赖。大量的实验结果表明，我们提出的方法在多个标准上的分类性能显着优于之前的所有方法 10%。对模拟蛋白质组数据和与生物体无关的实验的其他研究进一步表明，我们的模型是一种更通用和更强大的工具，不依赖于蛋白质的其他组信息。在此基础上，我们设计了一个完整的信号肽发现管道，以从宏基因组数据中探索前所未有的信号肽。 所提出的方法高度敏感，揭示了 347 个预测为候选新 SP，我们的候选肽与训练数据集中最接近的信号肽之间的序列同一性最低，仅为 13%。有趣的是，训练集中候选者和 SP 之间的 TM 分数大多高于 0.8。对经过实验验证的新型 SP 的进一步分析提供了证据，表明尽管 USPNet 不依赖任何结构信息作为输入，但它是根据进化和结构信息而不是序列相似性来检测 SP 的。结果表明，USPNet 仅用原始氨基酸序列和大型蛋白质语言模型就可以学习 SP 结构，从而能够有效地发现远离现有知识的新型 SP。

* 同等第一作者。
†通讯作者。邮箱：liyu@cse.cuhk.edu.hk

1 引言 信号肽（SP）是一段短氨基酸序列，作为特异性靶向信号引导蛋白质并将其转移至分泌途径[1]。它具有三域结构：带正电的 N 区、疏水性的 H 区和不带电的 C 区 [2]。 SP 作为特定片段引导蛋白质到达正确位置，然后被 C 区附近的切割位点切割。因此，信号肽的鉴定对于研究蛋白质的目的地和功能至关重要[3,4,5,6]。

由于 SP 的全面实验鉴定可能非常耗时和资源消耗，因此人们提出了许多计算工具来对信号肽进行分类并预测切割位点。第一次尝试是1983年提出的制定规则[7]。 Von Heijne 首先应用统计方法仅基于 78 种真核蛋白来揭示信号肽切割位点附近的模式 [7]。此外，还提出了生成模型，例如隐马尔可夫模型（HMM），以促进信号肽的识别。这些模型侧重于详细分析这三个功能区（N 区、C 区和 H 区），并通过捕获信号肽不同区域之间的关系来构建[8,9,10,11,12]。与生成模型不同，提出了一些基于同源性的方法[13]。这些方法的预测基于现有知识库中的序列与输入序列之间的相似性。此外，它们可以实现与生成模型类似的预测性能。

最近，监督模型在信号肽识别方面取得了很大进展。查询序列被编码为嵌入向量，然后输入模型以直接计算每种信号肽类型的概率。在这些方法中，基于机器学习的模型在其卓越的性能中发挥着重要作用[14]。 DeepSig将深度卷积神经网络（DCNN）架构应用于信号肽的识别和切割位点位置的预测[15]。此外，SignalP5.0 的出现并对之前提出的所有方法进行了基准测试 [16]，而 SignalP6.0 [17] 能够预测之前模型未能检测到的所有 5 种信号肽。这些方法在任务中取得了先进的性能，但大多数都遭受了极端的类不平衡，因此在小类数据上表现不佳[18,19,20]。此外，这些方法通常在很大程度上依赖于有关生物体群体的附加信息来提高其性能。然而，现实中从宏基因组数据中获取足够的群体信息是不切实际的[21, 22]。一个强大的工具应该只需要氨基酸序列即可产生准确的预测结果。

在信号肽分类中，要解决的关键问题是训练数据的不平衡和对象对群体信息的依赖。拉奥等人。 [23]介绍了一种基于变压器的语言模型ESM-1b，该模型证明仅从大规模蛋白质序列中学习的信息就可以隐式编码功能和结构信息，并有利于各种下游任务，例如二级结构预测和接触预测，表现优于很大程度上是特定数据训练的模型。拉奥等人。[24]还发现，将多重序列比对（MSA）集成到模型中（称为 MSA 转换器）可以带来更优异的性能。这些蛋白质语言模型 [25, 26] 在有限注释数据的问题上得到了改进。受此启发，我们提出了基于 BiLSTM [27] 框架和蛋白质语言模型的无偏生物体不可知信号肽预测器（USPNet），以对信号肽进行分类并预测其切割位点位置。我们利用先进的基于 MSA 的蛋白质语言模型来丰富表示，以帮助编码序列的组信息。此外，我们将类平衡损失与标签分布感知边缘（LDAM）损失[28]结合起来作为USPNet的损失函数以提高泛化能力。我们的模型是端到端的，仅以原始氨基酸作为输入。对所有五种类型的信号肽和非信号肽类型蛋白质进行分类是有效的。我们将我们的模型与重新分类的 SignalP5.0 基准集上的几个与任务相关的深度学习模型进行比较。值得注意的是，与之前最先进的方法相比，USPNet 在多个类别上实现了超过 10% 的 MCC 改进。此外，我们的模型显着优于 SignalP6。切割位点预测的召回率为0。在我们精心设计的域转移独立集上，USPNet 也比其他模型表现更好，这表明我们的方法在信号肽分类上的泛化。为了进一步展示 USPNet 的效力，我们收集了大肠杆菌（K12 菌株）以及其他 7 种生物体的蛋白质组数据。当应用 USPNet 从整个蛋白质组数据中检测信号肽时，它几乎检索到了所有信号肽，这是所有正在测试的模型中最好的之一。此外，我们通过进行与生物体无关的实验来彻底评估组信息依赖性，该实验删除了输入中的组信息。 USPNet 训练有素的编码器具有 MSA 信息以及蛋白质语言模型，可捕获丰富的进化和功能信息，使模型在缺乏群体信息的情况下保持稳健。这种高度灵敏的信号肽预测能力使得能够从大量宏基因组资源中挖掘新的SP。因此，我们建立了从处理宏基因组数据到进行新型信号肽检测的完整流程。我们从多个来源收集猪肠道宏基因组数据进行案例研究，最终从数百万个序列中筛选出347个肽作为与现有SP序列一致性较低且可能是新型信号肽的候选肽。

结果表明，USPNet能够提供SP的进化和结构信息，并有效地发现与现有知识相距甚远的候选信号肽。

我们工作的主要贡献总结如下：

- 我们引入了无偏生物体不可知信号肽网络 (USPNet)，它能够预测所有 5 种已知类型的信号肽。大量实验表明，所提出的方法在信号肽分类方面比其他信号肽预测器实现了最先进的性能。我们将 USPNet 应用于独立组和蛋白质组范围的研究。与之前的方法相比，该模型在多个标准上均取得了 10% 的改进，并且保持了 90% 以上的性能。

- 我们提供两个版本的 USPNet 供使用。一种构建多序列比对 (MSA) 并使用 MSA 转换器生成嵌入来丰富我们的表示。另一种利用进化规模建模（ESM）嵌入[23]，我们将其命名为 USPNet-fast。 USPNet的第一个版本具有更好的预测能力，USPNet-fast可以将推理速度提高20倍。我们方便用户根据应用场景选择工具。

- 我们解决了信号肽预测中的极端不平衡问题。考虑到以前的算法主要基于交叉熵损失来训练模型，我们建议应用标签分布感知损失（LDAM）来提高不太频繁的类的泛化能力[28]。我们通过将类平衡损失与 LDAM 损失相结合，提出了一种改进的损失函数，使 USPNet 从小类中学习有用的信息。

- 我们建立了一个完整的流程来从原始宏基因组数据中检测信号肽。我们揭示了 347 个预测为候选新 SP，候选肽与训练数据集中最接近的信号肽之间的序列同一性最低，仅为 13%。值得注意的是，训练集中候选者和 SP 之间的 TM 分数大多高于 0.8。我们还检索了我们的研究基因组的所有 4 个经过实验验证的信号肽，这些信号肽有文献支持，并且不存在于训练数据集中。结果表明，USPNet 无需从蛋白质折叠模型中额外输入即可学习进化和结构信息，因此可以以理想的速度发现序列同一性低但结构相似性高的肽。

2 结果

2.1 USPNet 是一个从宏基因组数据预测信号肽的管道

如图 1.a 所示，该流程可以从宏基因组数据中预测信号肽，甚至发现新的 SP 候选物。我们方法的基本架构是具有自注意力机制 [30] 的 Bi-LSTM [27, 29]，并且我们利用基于蛋白质语言模型的编码器来丰富表示（图 1.b）。USPNet 以氨基酸序列为输入，同时预测信号肽类型和相应的切割位点。考虑到蛋白质N端信号肽的长度通常在5到30之间，我们将70作为蛋白质长度的截止值，这意味着每个输入序列最多包含70个氨基酸。然后该序列进入特征提取模块，并与嵌入层一起使用。由于常见残基类型的数量为20，因此它将输入序列转换为L×20维矩阵，其中L是序列的长度。特别是，我们在生成的嵌入的头部添加一个 L×4 维向量来存储组信息（真核生物、革兰氏阳性菌、革兰氏阴性菌和古细菌）。然后，生成的嵌入被输入到我们的 BiLSTM 部分。它由具有自注意力的 Bi-LSTM 层和 CNN 组成，可同时提取长距离依赖性的前向和后向以及序列的全局/局部特征。在 Bi-LSTM 模块之后，我们开发了一个基于 MLP 的模块来分别预测切割位点和 SP 类型。为了整合更多信息来分类信号肽，我们引入了 MSA 嵌入。具体来说，我们首先为每个序列生成多序列比对（MSA），然后将 MSA 输入到预先训练的 MSA Transformer 模型 [24] 以从其最后一层获得嵌入。 为了努力解决数据不平衡问题，我们的损失函数是通过将类平衡损失与标签分布感知边缘（LDAM）损失相结合来设计的[28]。详细信息在方法部分介绍。

除了上述方法之外，我们还引入了 USPNet 的另一个版本，称为 USPNet-fast，它用 ESM-1b 嵌入代替 MSA 嵌入[23]，并保持其他模块不变（图 1.c）。 USPNet-fast 能够更快地预测信号肽和切割位点，并且性能不会大幅下降。

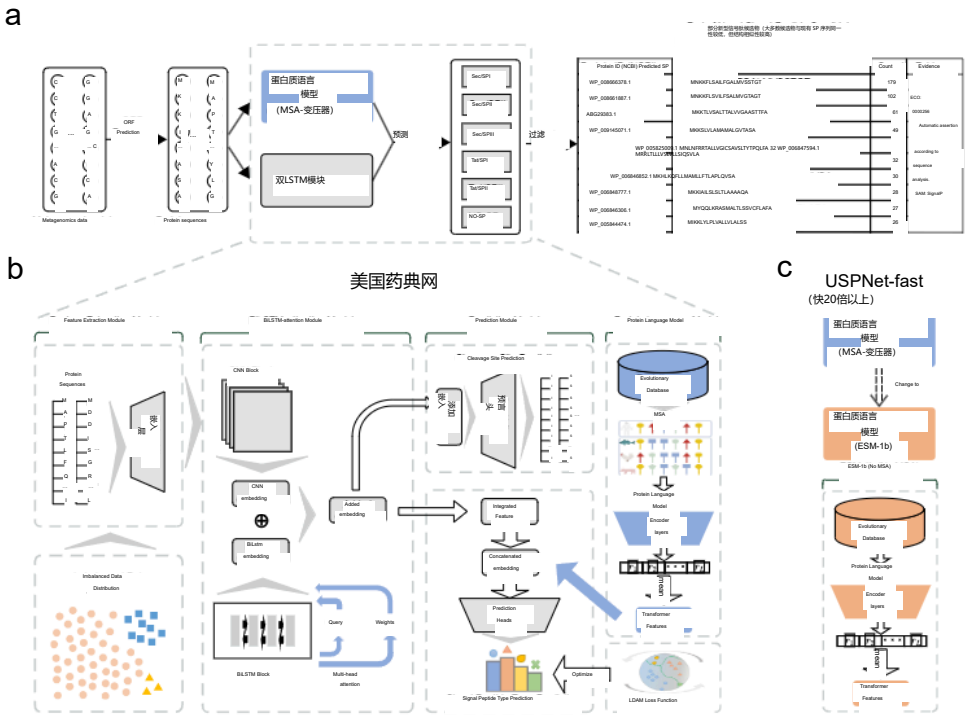

图 1：用于预测信号肽 (SP) 和切割位点的 USPNet 工作流程。 A。 USPNet 从宏基因组数据中发现信号肽的流程。USPNet 以原始氨基酸序列作为输入，并使用 BiLSTM 模块以及蛋白质语言模型来获得级联嵌入作为预测的输入。在这里，我们列出了 UniProt 数据库从 347 个候选肽中自动注释为 SP 的前 10 个最常出现的肽。 b. USPNet 的详细架构。训练数据不平衡。蛋白质序列经过特征提取模块，然后传递到 BiLSTM 模块，该模块包括具有自注意力的 Bi-LSTM 层和用于提取序列的长距离依赖性和特征的 CNN。对于 SP 类型预测，USPNet 结合了由预先训练的 MSA Transformer 模型生成的 MSA 嵌入。随后基于 MLP 的模块可预测切割位点和信号肽类型。训练中采用标签分布感知裕度（LDAM）损失来解决数据不平衡问题。 C。 USPNet-fast 用 ESM-1b 取代了 MSA 转换器，ESM-1b 不需要 MSA，因此推理速度更快。

2.2 USPNet 在基准数据集上优于之前的方法

信号肽类型预测

USPNet能够同时预测信号肽的类型和切割位点。为了公平地分析性能，我们在从 SignalP5.0 和 SignalP6.0 [16, 17] 发布的数据派生的重新分类、扩展和同源性减少的数据集上训练和测试我们的模型。训练和基准数据的组合与同源分区的 SignalP6.0 数据集相同。对于信号肽类型预测，训练数据（排除数据集中的基准数据）包含 13679 个序列，根据六种不同类型的标签可以将其分为六部分。然而，数据极不平衡。如表1所示，带有主要类标签的序列是带有次要类标签的序列的十倍。因此，为了减轻这种偏差并实现对性能的公平评估，我们重点关注属于不太频繁类别的不同序列划分的马修斯相关系数（MCC）。我们使用 MCC 的两个版本作为测量，即 MCC1（将特定信号肽与 TM/Globular (NO-SP) 型蛋白质进行对比）和 MCC2（另外还包括阴性集中的所有剩余序列）。

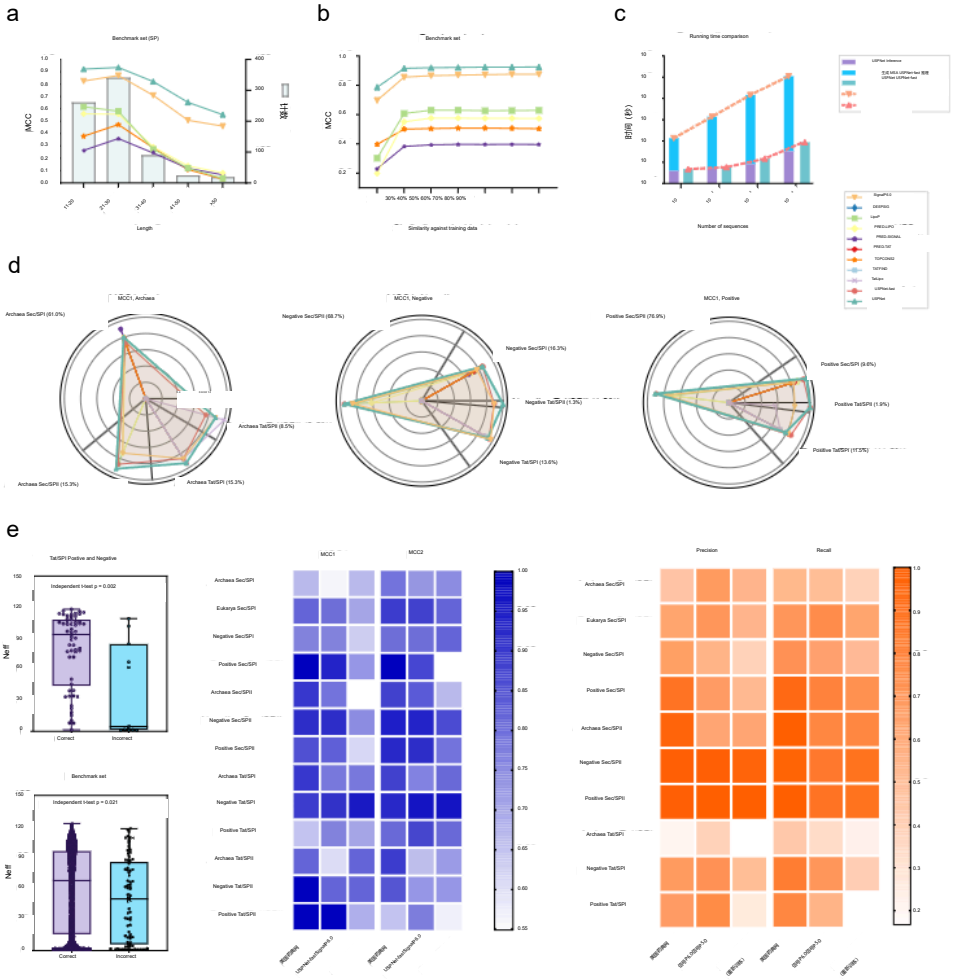

图 2：USPNet 在不同信号肽类型和生物体群体中表现出稳健的性能。 A。基准集中信号肽性能的长度分布。大多数 SP 的长度都低于 50 个 AA，并且当 SP 超过 30 个 AA 时，所有模型的性能都会下降。 b.基准集和训练集在不同序列身份下的性能。 C。 USPNet 和 USPNet-fast 之间的信号肽预测运行时间比较。 MSA 生成步骤非常耗时，约占总处理时间的 95%。 d. USPNet 和其他模型在不同性能角度的雷达图。我们从生物群体的角度比较不同模型的MCC1以及每种SP类型的比例信息。 USPNet 是每个生物群体中最强大的信号肽预测器。 e. USPNet、USPNet-fast、SignalP6.0 和 SignalP5.0 在信号肽类型预测和信号肽切割位点预测方面的比较。以及 Gram+ 和 Gram- 的 Tat/SPI 上的 MSA 的 Neff 分数以及基准集。 （独立t检验：两个独立样本的t检验）。

我们在基准数据集上比较了几种已知信号肽预测模型的不同生物组的信号肽分类性能。总共选取了16种方法进行比较；其中，SignalP6.0和SignalP5.0-retrained是在与我们相同的训练数据集上进行训练的。其他方法的结果是直接从其公开可用的网络服务器获得的，这导致由于缺乏同源划分而导致潜在的性能高估。由于手头的任务涉及多类分类，

表1：本研究采用的完整数据集组成的统计，括号内为基准数据集编号。这里，SP表示Sec/SPI，L表示Sec/SPII，P表示Sec/SPIII，T表示Tat/SPI，TL表示Tat/SPII，N/C表示TM/Globular(NO-SP)。

数据集	生物体	SP	L	P	T	TL	N/C	总计
SP类型预测	真核生物	2040 (146)	-	-	-	-	14356 (5581)	16396 (5727)
	革兰氏阳性	142 (15)	516 (120)	4 (0)	39 (18)	8 (3)	226 (81)	935 (237)
	革兰氏-阴性	356 (61)	1087 (257)	56 (0)	313 (51)	19 (5)	933 (133)	2764 (507)
	古细菌	44 (36)	12 (9)	10 (0)	13 (9)	6 (5)	110 (81)	195 (140)

我们根据信号肽的类型剖析结果。对于带有 Sec/SPI 标签的数据，我们的模型在几乎所有生物体和指标上都优于其他方法。唯一的例外是 PRED-SIGNAL，专门设计用于仅检测 Sec/SPI SP，它在古生菌的 MCC1 上略胜于我们。尽管如此，我们的模型仍然在 MCC2 上表现出卓越的性能（图 2.d 和补充图 3）。在考虑 Sec/SPII SP 分类时，只有 5 种方法展示了其能力，并提供了三个生物组的相应数据。 USPNet 在所有指标中无疑具有最佳性能。对于古细菌的MCC1，我们比其他人至少高出6%。此外，甚至我们的 USPNet-fast 也比所有其他竞争对手做得更好。

从 SP 长度和序列同一性的角度来看性能时，我们发现短 SP 的检测性能良好，并且当 SP 长于 30 个 AA 时，所有方法都会下降（图 2.a）。当序列同一性低于 40% 时，USPNet 的性能下降最小（图 2.b），对远离训练数据的蛋白质表现出更好的泛化能力。很明显，USPNet 在预测性能方面取得了令人印象深刻的提升，尤其是在小类别上。

我们还对 USPNet 和 SignalP6.0 之间的 MCC1 和 MCC2 进行了头对头比较，如图 2.e 所示。 USPNet 在大多数 SP 类型上取得了显着的改进。特别是对于带有 Sec/SPII 标签的数据，我们观察到所有 4 组的数据增长了 10.0%。在数据最少的 Tat/SPII 类上，USPNet 能够检索几乎所有 SP。尽管 SignalP6.0 表现出值得称赞的性能，但与 USPNet 相比仍存在不足。然而，在革兰氏阳性和革兰氏阴性的 Tat/SPI 上，USPNet 落后于 USPNetfast 和 SignalP6.0。这主要归因于错误预测中多重序列比对 (MSA) 的质量不佳。我们计算了我们方法中 MSA 转换器使用的正确预测和错误预测的有效序列数 (Neff)（图 2.e）。具体来说，对于 Tat/SPI，正确预测和错误预测之间存在显着的 Neff 差距：前者的中位数约为 90，而后者仅为 4。在整个基准集上，正确预测的 SP 的 MSA 质量也很差。更好，表明高质量的 MSA 生成可以提高 USPNet 的性能。总体而言，基准集的结果符合我们的假设，即基于边际的损失公式允许进一步扩展稀有类别的分类边界，并在某些方面避免过度拟合。 USPNet 具有更好的泛化能力，是一种无偏的多类 SP 预测器，能够预测所有 5 种 SP。除了基准集之外，我们还对完整的训练集和基准集进行了 5 倍交叉验证（补充表 5-9）。 由于同源划分在交叉验证中不起作用，USPNet 表现出了更令人印象深刻的性能，特别是对于 Tat/SPII 和 Sec/SPIII 信号肽等次要类别。

信号肽切割位点预测

一般来说，USPNet 的重点是生成一个预测器来对各种信号肽类型进行分类，并更好地概括不同类别。然而，在对 SP 进行分类之前，找出蛋白质的精确切割位点也是实际应用中的一个重要步骤。在 USPNet 中，我们利用上下文注意力矩阵的注意力权重来获取与每个氨基酸相关的信息，随后通知切割位点决策。为了进行比较，我们使用精确率和召回率来评估USPNet相对于SignalP6.0和SignalP5.0重新训练的切割位点预测的性能，结果可以从图2.e中观察到。我们注意到 USPNet 切割位点预测的整体性能与 SignalP6.0 相当。特别是，我们的模型在召回方面显着优于其他两个模型，这表明我们更擅长在生物体中找到尽可能多的信号肽，特别是那些属于难以捉摸的小类的信号肽。然而，我们的模型在某些类别中的精度略低于 SignalP6.0，这可能是由于分配给次要类别的权重被放大，导致误报率可能增加。尽管如此，用户可以根据自己的具体需求调整权重，增强USPNet的有效性和对其应用场景的适用性。

最低相似度值 表 2：1715 个蛋白质（新 SP 候选者）中前 10 个最常出现的预测信号肽序列，以及我们从猪肠道宏基因组数据中找出的一些经过实验验证的 SP。 （Seq-sim：重构训练集中最高序列相似度值，TM-score：重构训练集中最高结构相似度（TMscore）值）

新型 SP 候选蛋白 ID (NCBI) 预测 SP 计数证据 WP 008666378.1 MNKKFLSAILFGALMVSSTGT 179

可湿性粉剂 008661887.1 MNKKFLSVILFSALMVGTAGT 102 ABG29383.1 MKKTLVSALTTALVVGAASTTFA 61 可湿性粉剂 009145071.1 MKKSLVLAMAMALGVTASA 49 可湿性粉剂 005825009.1 MNLNFRRTALLVGICSAVSLTYTPQLFA 32 可湿性粉剂 006847594.1 MRRLTLLLVSLVLLSIQSVLA 32 WP 006846852.1 MKHLKQFLLMAMLLFTLAPLQVSA 30 WP 006848777.1 MKKIAILSLSLTLAAAAQA 28 WP 006846306.1 MYQQLKRASMALTLSSVCFLAFA 27 WP 005844474.1 MIKKLYLPLVVALVLAL SS 26 实验验证条目 (UniProtKB/Swiss-Prot ID) 预测 SP证据 Seq-sim TM 分数 P02768 MKWVTFISLLFLFSSAYS [41] 0.222 0.990 P28800 MALLWGLLALILSCLSSLCSAQ [42] 0.35 0.902 Q2UVX4 MKPTSGPSLLLLLLASLPMALG [43] 0.381 0.709 Q3MHN5 MKRILVFLLAVAF VHA [44] 0.125 0.961

生态：0000256
启动师言
根据
序列比对
SAM:信号P

我们的候选肽和特定信号肽之间的结构同源性。我们的方法以原始氨基酸序列作为输入，不依赖于结构信息。然而，所发现的候选者似乎已根据结构相似性被筛选出来。为了进一步检验我们的假设，接下来我们研究 4 个经过验证的 SP。我们原始的训练集包含与这些经过验证的 SP 相似度超过 60% 的 SP。为了确保 USPNet 检测 SP 时不依赖于序列相似性，我们从训练集中排除与 4 个已验证 SP 序列同一性超过 40% 的 SP，并重新训练我们的模型。这次，我们的模型仍然能够检测到所有 4 个 SP，平均而言，它们与训练集具有相当高的结构相似性（表 2）。基于文献的评估和数据库搜索结果表明，我们的 USPNet 可扩展到大量数据。虽然它是基于序列的方法，但它可以发现序列同一性低但结构相似性高的肽，这表明我们的方法已经学会以相对理想的速度在没有任何相关输入的情况下表示蛋白质折叠信息。它可能有助于从宏基因组数据中筛选信号肽，并发现与目前已知的显着不同的 SP。

3 讨论

我们将 USPNet 开发为一种深度学习方法，专注于解决信号肽预测中现有的长尾数据分布问题和组信息依赖问题。为了解决数据不平衡的问题，我们首先引入 LDAM，它与大多数基于深度学习的信号肽预测器正交。由于单一损失函数的效果有限基于并考虑到 LDAM 的特点，我们建议在我们的方法中将类平衡损失和 LDAM 损失结合在一起，我们将其命名为 CB-LDAM 损失，以增强小类和大类边界的泛化类。此外，开发的蛋白质语言模型包含丰富的进化和结构信息，并显示出其改进大量下游任务的卓越能力。因此，我们建议引入 MSA 变压器和 ESM 嵌入来实现端到端的生物体不可知模型。最后，我们提出了一种新的模型架构。基于注意力的 BiLSTM 模型允许最大限度地集成和提取输入序列不同位置之间关系的可能性，并有助于提高性能。

尽管综合实验已经证实，与最先进的方法相比，所有这些技术的整合可以在信号肽预测方面取得更好的性能，但我们在识别古细菌 Sec/SPI 信号肽方面的结果并不那么令人印象深刻。我们认为造成这种现象的原因是我们的损失函数调整了不同类型 SP 的权重，从而使 USPNet 倾向于将肽分为小类。在实际使用过程中，为了使模型更加具有针对性，我们建议用户可以根据自己的需求调整不同类型的权重。

正如前面所强调的，我们开发了一个完整的管道来发现新的信号肽从metage-

经济学数据。我们选择USPNet-fast进行研究的原因是多序列比对可能非常耗时，导致无法在合理的时间内完成大规模筛选。 347 个候选新颖 SP 可作为开放科学资源使用。潜在地，新的信号肽可用于提高异源蛋白质的分泌效率。我们为发现过程提供了一个范例，并且很容易重现。除了我们在研究中使用的猪肠道数据外，人类肠道和心脏组织等其他资源也是信号肽库。我们希望我们的管道能够为相关研究提供新的见解。

除了信号肽预测之外，数据不平衡和对象依赖性在生物预测问题中也很常见。我们的模型开发策略可以转移到更一般的生物信息学领域[45,46,47,48,46,49]。总之，USPNet 代表了一个广泛适用的预测信号肽甚至蛋白质序列的框架。考虑到它可以与管道中的其他工具集成，我们相信 USPNet 将有助于研究广泛的信号肽问题。

4 方法

4.1 数据采集

我们在研究中总共收集了四种数据集，包括基准数据、独立数据集、蛋白质组数据和宏基因组数据。

基准数据

训练和基准测试集与 SignalP6.0 [17] 中引入的相同，它对 SignalP5.0 [16] 发布的数据重新分类了一些 SP 类型。此外，它还添加了一些来自 UniProt20 [50] 和 Prosite21 [51] 的新 SP，以及来自 UniProt 和 TOPDB22 [52] 的新可溶性和跨膜蛋白。然后，使用 G´ıslason 等人引入的同源划分方法删除了新数据集中的部分数据。 [53]。训练数据总共包含 13679 个序列，基准测试数据包含 6611 个序列。原始训练数据集包含来自四个生物群的蛋白质：真核生物、革兰氏阳性菌、革兰氏阴性菌和古细菌。为了验证小类更好泛化的效果，我们执行了五种单独的 SP 类型：Sec/SPI、Sec/SPII、Tat/SPI、Tat/SPII 和 Sec/SPIII SP。其他蛋白质相应地被认为是TM/Globular (NO-SP)类型，如表1所示。具有6个单独标签的数据集具有长尾标签分布，这意味着NO-SP类型在不同标签中在数量上更占优势。值得注意的是，存在三种少数类SP：Tat/SPI包含365个数据点，更糟糕的是，只有33个数据被标记为Tat/SPII信号肽，70个数据被标记为Sec/SPIII信号肽。在训练和基准集中，蛋白质序列的长度可以变化。然而，信号肽的长度通常在 10-50 个 AA 之间（图 2）。因此，我们将截止值设置为 70 个 AA，以突出信号肽并避免长蛋白质序列的影响。

独立测试集

我们策划的独立测试数据集被命名为 SP22。它从SwissProt数据库[50]中收集了经过验证的SP蛋白，该数据库是在2020年11月之后发布的，即SignalP 6.0数据集收集日期。具体来说，它是通过以下步骤生成的：（1）我们去除2020年11月之前的蛋白质，以及由少于30个氨基酸组成的蛋白质；(2)我们选择真核生物、革兰氏阳性菌和革兰氏阴性菌；(3) 我们从 Swiss-Prot 数据库（2022 年 04 月发布）中选择含有信号肽的蛋白质，并具有可信标签 ECO:0000269（实验注释）和 ECO:0000305（手动策划注释）。此外，为了更好地保证SP22数据集的独立性，应用CD-HIT[54]去除与SiganlP6.0数据集中的蛋白质相似性超过40%的冗余蛋白质，并将SP22的内部冗余削减至 80%。 SP22总共有来自39个物种的43条蛋白质序列，这些序列在训练集中很少出现，并且所有序列都含有信号肽。其中，真核生物组31个，革兰氏阴性组6个，革兰氏阳性组6个。

全蛋白质组数据

我们从 UniProt 数据库收集参考蛋白质组数据。对于大肠杆菌（菌株 K12），UniProt 报告了 496 种带有信号肽的蛋白质，其中 141 种经过实验验证，355 种通过各种方式预测。收集经实验验证的 SP 作为信号肽，并带有可信标签 ECO:0000269 和 ECO:0000305

来自 Swiss-Prot 数据库（2022 年 4 月发布）的所有蛋白质组。我们还研究了 UniProt 中提供的其他 7 种参考蛋白质组，包括枯草芽孢杆菌、谷氨酸棒杆菌、耐辐射奇球菌、Haloferax volcanii、Methanocaldococcus jannaschii、激烈火球菌和嗜热栖热菌。

宏基因组数据

我们从五个来源收集猪肠道宏基因组：PRJEB38078 [55]、PRJNA561470 [56]、PRJNA647157 [57]、CNP0000824 [58] 和全球微生物组保护协会 (GMBC) [59]。它们用于预测 sORF 序列。为了确保我们的 sORF 确实得到表达，我们使用了来自 PRIDE 项目 PXD006224 [60] 的元蛋白质组数据集。我们选择与我们预测的信号肽相同的序列。

4.2 USPNet成立

考虑到蛋白质N端信号肽的长度，我们将蛋白质L的输入序列长度设置为70，这意味着每个输入序列最多包含70个氨基酸。考虑到不同残基类型的数量为20，我们采用嵌入层来提取特征，从而将输入序列转换为L×20维矩阵。同时，我们使用one-hot编码，用4个二进制数表达输入序列的4种群体信息（真核生物、革兰氏阳性菌、革兰氏阴性菌和古细菌）。然后，我们复制它们，将输入辅助序列转换为 L×4 维矩阵。如果没有提供组信息，则所有条目都指定为 0。

对于 USPNet，模型架构包含两个关键组件：BiLSTM 注意力模型和用于预测信号肽类型和切割位点的特征提取模块。

BiLSTM模块

我们首先描述 BiLSTM 模块。它获取蛋白质序列的 L×20 维嵌入向量，并将其输入具有 252 个隐藏单元的全连接层。全连接层的输出有两种用途。首先，它被传递到两个 CNN 层以生成 L×512 维表示以供将来使用。 CNN 可以捕获重要的主题并聚合整个输入序列中有用的局部和全局信息。其次，输出与辅助矩阵连接形成L×256维矩阵作为BiLSTM层的输入。它具有 128 个隐藏单元，可提取序列在前向和后向方向上的长距离依赖性。这样，我们通过连接两个方向的输出，在更高级别上集成来自输入序列的信息，以便以后进行预测，可以写为：

$$h=[\overrightarrow{h};\overleftarrow{h}] \quad \square \quad \square$$
$$=\mathrm{BiLSTM}\left(e,\overrightarrow{h},\overleftarrow{\theta}\right), \tag{1}$$

在处理序列时，$\overrightarrow{h}$ 表示时间点 t 在向前和向后方向上的隐藏状态。 e 表示输入序列元素 x 的嵌入表示。 θ表示BiLSTM层的参数。 BiLSTM 层采用嵌入输入 e，即前向隐藏状态 $\overrightarrow{h}$，后向隐藏状态参数 $\overleftarrow{\theta}$ 作为输入并输出串联的隐藏状态 h。 BiLSTM 层输出 256 维向量。

为了更好地聚合来自不同特征子空间的信息，我们在 BiLSTM 模块中开发了一种多头注意力机制，每个头都有缩放的点积注意力：

$$\text{注意力}(Q,K,V)=\mathrm{softmax}(\sqrt{})V, \tag{2}$$

其中 是缩放因子，dis 是查询和键的维度。我们在每个头中执行方程（2）并将多个单头输出连接在一起：

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\text{头},\ldots,\text{头})W, \quad \square$$
$$\text{头}=\text{注意力} \qquad \text{功率,功率,功率} \tag{3}$$

其中投影是参数矩阵 W ∈ R，W ∈ R，W ∈ R； Q 是前一个 BiLSTM 层的输出； K 和 V 与 BiLSTM 层的输入相同。 h代表磁头数。我们采用 h = 2 来形成并行注意力层。对于每个头，我们

对于每个头，使用 d= 256，d= d= d/h = 128。在多头注意力模块中，我们采用残差连接步骤来稳定训练并帮助减轻可能的梯度消失问题。多头注意力模块的输出进一步与其输入连接，然后按元素与先前 CNN 层的表示相加。之后，该向量最终被输入到具有 256 个隐藏单元的第二个 BiLSTM 层。

为了预测切割位点，我们在第二个 BiLSTM 层之上构建了一个三层 MLP。因此，BiLSTM 的输出被输入到三个全连接层以生成 L×11 维矩阵。 11维向量对应于SP区域标签，蛋白质序列中的每个氨基酸有10个SP区域标签，并且对于长度小于70的输入序列的空位有1个注释。输出矩阵用于直接预测所有序列中切割位点的存在。输入蛋白质序列的氨基酸位置。

此外，第二个 BiLSTM 层的输出还用于预测信号肽类型。输出向量被重塑，并首先经过一个全连接层以降低其维度并产生 512 维向量。同时，我们在此处包含 MSA 嵌入以集成更多信息。预训练的 MSA Transformer 模型 [24] 生成 768 维嵌入，我们使用全连接层将有关这些嵌入的高级信息聚合为 64 维向量。处理后的结果与全连接层先前输出的向量连接起来。因此，级联结果具有 512+64 个通道，并由具有 256 个隐藏单元的全连接层进行汇总。最后，输出被转发到归一化线性投影层以进行预测。

针对数据不平衡问题设计的损失函数

大多数与信号肽相关的现有知识库都存在极端的数据不平衡问题。信号肽序列的少数类别的数量通常远小于非信号肽序列的数量。这种情况可能会导致低资源数据类型的泛化能力较差。现有技术，例如成本敏感的重采样和重加权，可以有效应对数据不平衡带来的挑战。然而，对于信号肽预测，这些方法对大小写敏感并且容易过度拟合。还提出了后校正方法来处理数据不平衡，但不能确保泛化能力的提高。受普通经验风险最小化（ERM）算法的启发，我们引入 LDAM 损失并将其与重新加权相结合来解决该问题。

LDAM损失侧重于通过引入margin项△y来修正交叉熵函数，以提高类的泛化能力。它表明，适当的余量应该在主要类别和次要类别的泛化之间实现良好的权衡。多类别的类别相关边距经验证具有以下形式：

$$\triangle y = \frac{C}{n_j^{-}}, \tag{4}$$

其中n是第j类的样本量，C是要调整的超参数。

然而，margin项改变了正确类别的logits幅度，影响了泛化，这增加了表示学习的难度，并使模型对参数设置敏感。在这里，我们通过在分类器的最后一层应用归一化线性层来引入每个类别的代理向量[61]。我们对线性投影层的输入和权重进行归一化，以将特征向量和权重向量的内积限制在 [-1, 1] 中。实验证明，归一化不仅可以提高模型的鲁棒性，而且可以加速训练过程中的收敛。

根据经验，带有缩放因子的softmax交叉熵损失广泛应用于强化学习和相关领域[62]。一方面，如果缩放因子太大，类间隔将变得接近0，从而影响泛化。另一方面，较小的比例因子将导致偏离目标函数。在我们的方法中，使用缩放因子作为超参数来确保最终的 logits 位于合理的范围内：

$$L(x, y) = -\log \frac{\square \text{是}}{\text{是}+ \quad \sum_{i=y} e^{\square}}, \tag{5}$$

在哪里

$$\triangle j = \frac{C}{n_j^{-}} \_ f \text{ 或 } j \in \{1, \ldots, K\}, \tag{6}$$

z表示groundtruth标签的logit分数，z表示其他标签的logit分数，s表示缩放因子。 z 和 z 都被归一化为代理向量。

最终的目标函数联合优化信号肽预测和切割位点预测：

$$L = -\frac{1}{N}\sum_{j=1}^{N}\sum_{y=1}^{K} L(x,y), \tag{7}$$

$$L = -\frac{1}{N*L}\sum_{j=1}^{N*L}\sum_{y=1}^{K} L(x,y), \tag{8}$$

$$Luspnet = L + \tau L, \tag{9}$$

其中 τ = 1。 L 表示信号肽预测的目标函数，其中类别数 K = 6。 L 表示切割位点预测的目标函数，其中区域标签的数量 K= 11。 N 是输入蛋白质的数量序列，L 是输入序列的最大长度，设置为

70.

丰富表征的蛋白质语言模型

与其他蛋白质语言模型相比，MSA Transformer [24]充分利用了强大的模型架构和大型进化数据库。此外，自监督学习擅长在许多不同尺度上编码蛋白质序列的特性，这有助于模型的高性能。经验证，在没有序列以外的生物信号的情况下，模型仍然可以学习氨基酸的结构、蛋白质序列和进化同源性[23]。此外，在多个家庭上训练的模型的泛化效果优于基于单个家庭的效果。因此，我们相信，通过添加 MSA 转换器的表示，可以捕获单个组内序列的相似性，并且 USPNet 将变得更强大，可以区分来自不同组生物的序列。

为了在我们的方法中应用 MSA 转换器，我们首先通过使用 HHblits[64] 搜索更新的 UniClust30[63] 为每个序列生成多序列比对 (MSA)。多样性最大化二次采样策略[24]是一种贪婪策略，从参考开始，将平均汉明距离最高的序列添加到当前序列集，用于将完整 MSA 的比对序列数量减少到 128 个。大小大于 128。然后我们将 MSA 输入到预先训练的 MSA Transformer 模型 [24]，以从模型的最后一层生成 768 维嵌入。

我们模型的另一个版本 USPNet-fast 用 ESM-1b 嵌入替换了 MSA 嵌入 [23]。
ESM-1b 采用单个序列作为输入来生成嵌入，因此它使 USPNet-fast 能够更快地预测信号肽和切割位点，并且不会造成太大的性能退化。

培训详情

在USPNet中，我们应用Adam优化器[65]，初始学习率为2×10，权重衰减为1×10[32]。 epoch 总数设置为 300，采用提前停止策略在训练过程中获得最佳模型。所有实验都在四块具有 32GB 内存的 V100 GPU 卡上运行。
USPNet 不依赖 PSSM 和 HMM 配置文件来丰富嵌入或增强性能。无需基于进化概况的特征即可更直接地进行预测，从而确保更短的处理时间。对于 SignalP6.0 和 SignalP5.0，我们使用它们在代码存储库中提供的超参数来训练模型。

4.3 绩效评估

对于实验中采用的指标，我们将它们总结为两个组成部分：用于评估分类性能和切割位点（CS）预测性能的指标。
对于分类，我们使用马修斯相关系数（MCC）作为测量值。我们通过序列标签来考虑真/假阳性/阴性。 MCC可写为：

$$\text{中冶} = \text{p}\frac{TP \times TN - FP \times FN}{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}, \tag{10}$$

其中TP表示真阳性数，TN表示真阴性数，FP表示假阳性数，FN表示假阴性数。在这里，我们同时包含 MCC1 和 MCC2 来评估性能，因此计算两次。对于MCC1，我们仅将球状和/或跨膜蛋白作为阴性集，并将相关信号肽类型的蛋白作为阳性集。然后，对于MCC2，将所有其余序列添加到负集。

在消融研究中，我们不仅测量 MCC 对不同生物群体的分类性能，还应用整体 MCC、Kappa 和平衡准确率进行评估。这是因为在消融研究部分，一些模型在小类中的表现非常接近，整体表现会更直接直观地显示差异。 Kappa 广泛用于多类分类的一致性测试和测量。并且考虑到数据集的极度不平衡，准确性无法体现模型的实际性能。因此，我们采用平衡精度代替。平衡精度对样本总数的真阳性和真阴性预测进行归一化：

$$\text{平衡精度} = \frac{TPR + TNR}{2}, \tag{11}$$

其中TPR表示真阳性率，TNR表示真阴性率。

在CS预测部分，我们以精度和召回率作为衡量标准。精度表示正确的 CS 预测占预测 CS 总数的比例，召回率表示正确的 CS 预测占 CS 的真实标签数量的比例。值得注意的是，我们不能容忍预测偏差；换句话说，只有准确的站点预测才会被认为是正确的。

为了估计多序列比对的质量，应用有效序列的数量（Neff）。它可以计算为以下函数：

$$N = \sum_{i=1}^{N} \frac{1}{\text{重量}}, \tag{12}$$

其中N是MSA中的序列数，权重是MSA中任意两个同源序列i和j之间的序列同一性。

4.4 3-D 结构预测

考虑到计算资源，我们应用 AlphaFold2 的 ColabFold [66] 版本来执行图 5.d 和 e 中的 3-D 结构预测。具体来说，我们将同步运行模型的数量设置为1，并使用琥珀松弛来细化预测。其他设置保持默认。为了结构的可视化和随后的对齐操作，我们利用 Pymol [67]。而且由于训练集有大量序列，为了节省时间，训练集和 347 个候选序列的结构推断由 ESMFold [68] 进行，TM 分数由 TM-align [69] 计算（随后的3对样本结构比对仍然由ColabFold进行）。

5 数据可用性

我们使用的所有数据集都列在方法部分并且是公开可用的。支持本研究主要发现的所有其他相关数据，例如宏基因组学研究的结果，可在文章和补充信息文件中获得，或根据合理要求从相应作者处获得。本文提供了源数据。

6 代码可用性

USPNet的开源代码可以在https://github.com/ml4bio/USPNet找到，并且产生本文主要结果的实验也存储在这个存储库中。

参考

[1] von Heijne, G. 信号肽的生与死。自然 396, 111–113 (1998)。

[2] Heijne, G. V. 信号肽。膜生物学杂志 115, 195–201 (1990)。

[3] Bradshaw, N., Neher, S. B., Booth, D. S. & Walter, P. 信号序列激活 srp rna 的催化开关。科学 323, 127–130 (2009)。

[4] Craig, L.、Forest, K. T. 和 Maier, B. iv 型菌毛：动力学、生物物理学和功能后果。《自然》评论微生物学 17, 429–440 (2019)。

[5] 段G.-F.等人。信号肽通过与氨基末端结构域结合来抑制 gluk1 表面和突触运输。自然通讯 9, 4879 (2018)。

[6]江F.等。 N 端信号肽有助于将 PVC 复合物工程化为有效的蛋白质递送系统。科学进展 8，eabm2343 (2022)。

[7] Heijne, G. V. 信号序列切割位点附近的氨基酸模式。欧洲生物化学杂志 133 (1983)。

[8] Bendtsen, J. D.、Nielsen, H.、Von Heijne, G. 和 Brunak, S. 改进的信号肽预测：Signalp 3.0。分子生物学杂志 340, 783–795 (2004)。

[9] Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. 使用动态贝叶斯网络进行跨膜拓扑和信号肽预测。 PLoS 计算生物学 4，e1000213 (2008)。

[10] Ehsan, A.、Mahmood, K.、Khan, Y.D.、Khan, S.A. 和 Chou, K.-C。用于信号肽分类的数学生物学新模型。科学报告 8, 1039 (2018)。

[11] Janda, C. Y. 等人。信号识别颗粒对信号肽的识别。自然 465, 507–510 (2010)。

[12] 迈达尼，A. 等人。大型语言模型生成跨不同家族的功能蛋白质序列。

自然生物技术 1-8 (2023)。

[13] Frank, K. & Sippl, M. J. 基于序列比对技术的高性能信号肽预测。生物信息学 24, 2172–2176 (2008)。

[14] Petersen, T. N.、Brunak, S.、Von Heijne, G. 和 Nielsen, H. Signalp 4.0：区分跨膜区域的信号肽。自然方法 8, 785–786 (2011)。

[15] Savojardo, C.、Martelli, P. L.、Fariselli, P. 和 Casadio, R. Deepsig：深度学习改进了蛋白质中的信号肽检测。生物信息学 10 (2017)。

[16] Armenteros，J. J. A. 等人。 Signalp 5.0 使用深度神经网络改进信号肽预测。

自然生物技术 37, 420–423 (2019)。

[17] Teufel, F. 等人。 Signalp 6.0 使用蛋白质语言模型预测所有五种类型的信号肽。自然生物技术 1-3 (2022)。

[18]容克，A.S.等人。革兰氏阴性细菌中脂蛋白信号肽的预测。蛋白质科学 12, 1652–1662 (2003)。

[19] Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. 脂蛋白信号的预测

具有隐马尔可夫模型的革兰氏阳性细菌中的肽。蛋白质组研究杂志 7, 5082–5093 (2008)。

[20] Bendtsen, J. D.、Nielsen, H.、Widdick, D.、Palmer, T. 和 Brunak, S. 双精氨酸信号肽的预测。 BMC 生物信息学 6, 1–9 (2005)。

[21] 帕索利，E. 等人。通过实验中心可访问、精选的宏基因组数据。自然方法 14, 1023–1024 (2017)。

[22] Sczyrba, A. 等人。宏基因组解释的批判性评估——宏基因组软件的基准。自然方法 14, 1063–1071 (2017)。

[23] 里夫斯，A.等人。生物结构和功能是通过将无监督学习扩展到 2.5 亿个蛋白质序列而产生的。美国国家科学院院刊 118 (2021)。

[24] Rao, R. M. 等人。 MSA变压器。在Meila, M. 和Zhang, T.（编辑）第38届国际会议记录中

机器学习国际会议，卷。机器学习研究论文集 139，8844–8856（PMLR，2021）。

[25] Biswas, S.、Khimulya, G.、Alley, E. C.、Esvelt, K. M. 和 Church, G. M. 具有数据高效深度学习的低 n 蛋白质工程。自然方法 18, 389–396 (2021)。

[26] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. 将理性蛋白质工程与基于序列的深度表示学习相结合。自然方法 16, 1315–1322 (2019)。

[27] Thireou, T. & Reczko, M. 用于预测亚细胞的双向长短期记忆网络

真核蛋白质的定位。 IEEE/ACM 计算生物学和生物信息学汇刊 4, 441–446 (2007)。

[28] Cao, K.、Wei, C.、Gaidon, A.、Arechiga, N. 和 Ma, T. 学习具有标签分布感知边缘损失的不平衡数据集。arXiv 预印本 arXiv:1906.07413 (2019)。

[29] Armenteros，J. J. A. 等人。使用深度学习检测靶向肽中的序列信号。生命科学联盟2（2019）。

[30] Mnih, V.、Heess, N.、Graves, A. 等人。视觉注意力的循环模型。神经信息处理系统的进展 27 (2014)。

[31] McInnes, L.、Healy, J. 和 Melville, J. Umap：降维的均匀流形逼近和投影。 arXiv 预印本 arXiv:1802.03426 (2018)。

[32] Lin, T.-Y.、Goyal, P.、Girshick, R.、He, K. 和 Doll´ar, P. 密集物体检测的焦点损失。 IEEE 计算机视觉国际会议论文集，2980-2988（2017）。

[33] Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. 基于有效数量的类平衡损失

样品。 IEEE/CVF 计算机视觉和模式识别会议论文集，9268–9277 (2019)。

[34] 布拉特纳，F. R. 等人。大肠杆菌 k-12 的完整基因组序列。科学 277, 1453–1462 (1997)。

[35] Ma，Y.等。使用深度学习从人类肠道微生物组中鉴定抗菌肽。

自然生物技术 40, 921–931 (2022)。

[36] Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp：一种超快一体化 fastq 预处理器。生物信息学 34，i884–i890 (2018)。

[37] Steinegger, M.、Mirdita, M. 和 Soding, J. 蛋白质水平组装可成倍提高宏基因组样本中蛋白质序列的回收率。自然方法 16, 603–606 (2019)。

[38] Han，S.等。新型信号肽改善大肠杆菌中重组金黄色葡萄球菌α毒素h35l的分泌。安布快车 7, 1–14 (2017)。

[39] Consortium，T. U. UniProt：2023 年通用蛋白质知识库。核酸研究 51，D523–D531 (2022)。

[40] 跳跃者，J. 等人。使用 alphafold 进行高度准确的蛋白质结构预测。自然 596, 583–589 (2021)。

[41] Patterson, J. E. & Geller, D. M. 牛微粒体白蛋白：牛白蛋白原的氨基末端序列。
生物化学和生物物理研究通讯 74, 1220–1226 (1977)。

[42] Christensen, S. 和 Sottrup-Jensen, L. 牛 α2-抗纤溶酶 N 末端和反应位点序列。 FEBS 信件 312, 100–104 (1992)。

[43]巴尔多，A.等人。脂肪素酰化刺激蛋白系统和细胞内甘油三酯合成的调节。临床研究杂志 92, 1543–1547 (1993)。

[44] Nykjaer, A. 等人。内吞途径对于肾脏摄取和激活类固醇 25-(oh) 维生素 d3 至关重要。细胞 96, 507–515 (1999)。

[45] 李，Y.等。 Deepre：通过深度学习进行基于序列的酶 ec 数预测。生物信息学 34, 760–769 (2018)。

[46] Yu, Q., Dong, Z., Fan, X., Zong, L. & Li, Y. Hmd-amp：蛋白质语言驱动的分层多标签深度森林，用于注释抗菌肽。 arXiv 预印本 arXiv:2111.06023 (2021)。

[47] Lam, J. H. 等人。预测 rna 成分在蛋白质表面的结合偏好的深度学习框架。自然通讯 10, 1–13 (2019)。

[48] Wei，J.，Chen，S.，Zong，L.，Gao，X.和Li，Y.利用深度学习预测蛋白质-RNA相互作用：结构很重要。arXiv 预印本 arXiv:2107.12243 (2021)。

[49] 李Y.等。 Hmd-arg：用于注释抗生素抗性基因的分层多任务深度学习。微生物组 9, 1–12 (2021)。

[50] Consortium, U. Uniprot：全球蛋白质知识中心。核酸研究 47，D506–D515 (2019)。

[51] Sigrist, C.J. 等人。 prosite 的新的和持续的开发。核酸研究 41，D344–D347 (2012)。

[52] Dobson, L.、Lango, T.、Rem´enyi, I. 和 Tusn´ady, G. E. 加快 topdb 数据库的拓扑数据收集。核酸研究 43，D283–D289 (2015)。

[53] G´ıslason, M. H.、Nielsen, H.、Armenteros, J. J. A. 和 Johansen, A. R. 用指针神经网络预测 gpi 锚定蛋白。生物技术当前研究 3, 6–13 (2021)。

[54] Li, W. & Godzik, A. Cd-hit：用于聚类和比较大量亲数据的快速程序

tein或核苷酸序列。生物信息学 22，1658 – 1659 (2006)。网址 https://doi.org/10.1093/bioinformatics/btl158。 https://academic.oup.com/bioinformatics/article-pdf/22/13/1658/484588/btl158.pdf。

[55]Youngblut，N.D.等人。大规模宏基因组组装揭示了新的动物相关微生物基因组、生物合成基因簇和其他遗传多样性。 Msystems 5，e01045–20 (2020)。

[56] Looft, T.、Bayles, D.、Alt, D. 和 Stanton, T.coriobacteriaceae 菌株 68-1-3 的完整基因组序列，一种来自猪肠道的新型粘液降解分离株。基因组公告 3，e01143–15 (2015)。

[57] 周S.等。藏猪肠道微生物群中宏基因组组装基因组和碳水化合物降解基因的表征。微生物学前沿 11, 595066 (2020)。

[58]陈，C.等人。普氏菌 COPI 会增加配方奶喂养猪的脂肪积累。微生物组 9, 1–21 (2021)。

[59]格鲁辛，M.等人。工业化人类微生物组中水平基因转移率升高。细胞 184，2053–2067（2021）。

[60]蒂洛卡，B.等人。营养研究中的日粮变化塑造了猪粪便微生物群的结构和功能组成——从几天到几周。微生物组 5, 1–15 (2017)。

[61] Wang, F.、Xiang, X.、Cheng, J. 和 Yuille, A. L. Normface：用于人脸验证的 L2 超球面嵌入。

第 25 届 ACM 国际多媒体会议记录，1041–1049 (2017)。

[62] Liu, J., Krishnamachari, B., Zhou, S. & Niu, Z. Deepnap：通过深度强化学习的数据驱动基站休眠操作。IEEE 物联网杂志 5，4273–4282 (2018)。

[63]米尔迪塔，M.等人。 Uniclust 聚类和深度注释的蛋白质序列和比对数据库。

核酸研究 45，D170–D176 (2017)。

[64] Steinegger, M. 等人。 Hh-suite3 用于快速远程同源性检测和深度蛋白质注释。 BMC 生物信息学 20, 1–15 (2019)。
[65] Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。 arXiv 预印本 arXiv：1412.6980 (2014)。

[66] 米尔迪塔，M.等人。 Colabfold：让所有人都能参与蛋白质折叠。自然方法 19, 679–682 (2022)。

[67] DeLano, W. L. 等人。 Pymol：一种开源分子图形工具。 CCP4 新闻。蛋白质晶体学 40, 82–92 (2002)。
[68] 林，Z.等人。进化尺度上的蛋白质序列语言模型能够实现准确的结构预测。 BioRxiv 2022, 500902 (2022)。
[69]Zhang, Y. & Skolnick, J. Tm-align：基于 tm-score 的蛋白质结构比对算法。核酸研究 33, 2302–2309 (2005)。