# 5 PROTEIN LARGE LANGUAGE MODELS

# 5 蛋白质大语言模型

In the past years, large language models have become increasingly influential in protein research, offering novel insights and capabilities in understanding and manipulating proteins. In this section, we present a comprehensive review of LLMs for proteins (named Prot-LLMs), encompassing detailed discussions on their model architectures, utilized datasets, various capabilities, and corresponding evaluation criteria. The overview of this section is shown in Figure 8.

在过去的几年中，大语言模型在蛋白质研究中变得越来越有影响力，为理解和操作蛋白质提供了新的见解和能力。在本节中，我们对蛋白质大语言模型(Prot-LLMs)进行了全面回顾，包括对其模型架构、使用的数据集、各种能力以及相应评估标准的详细讨论。本节的概述如图8所示。

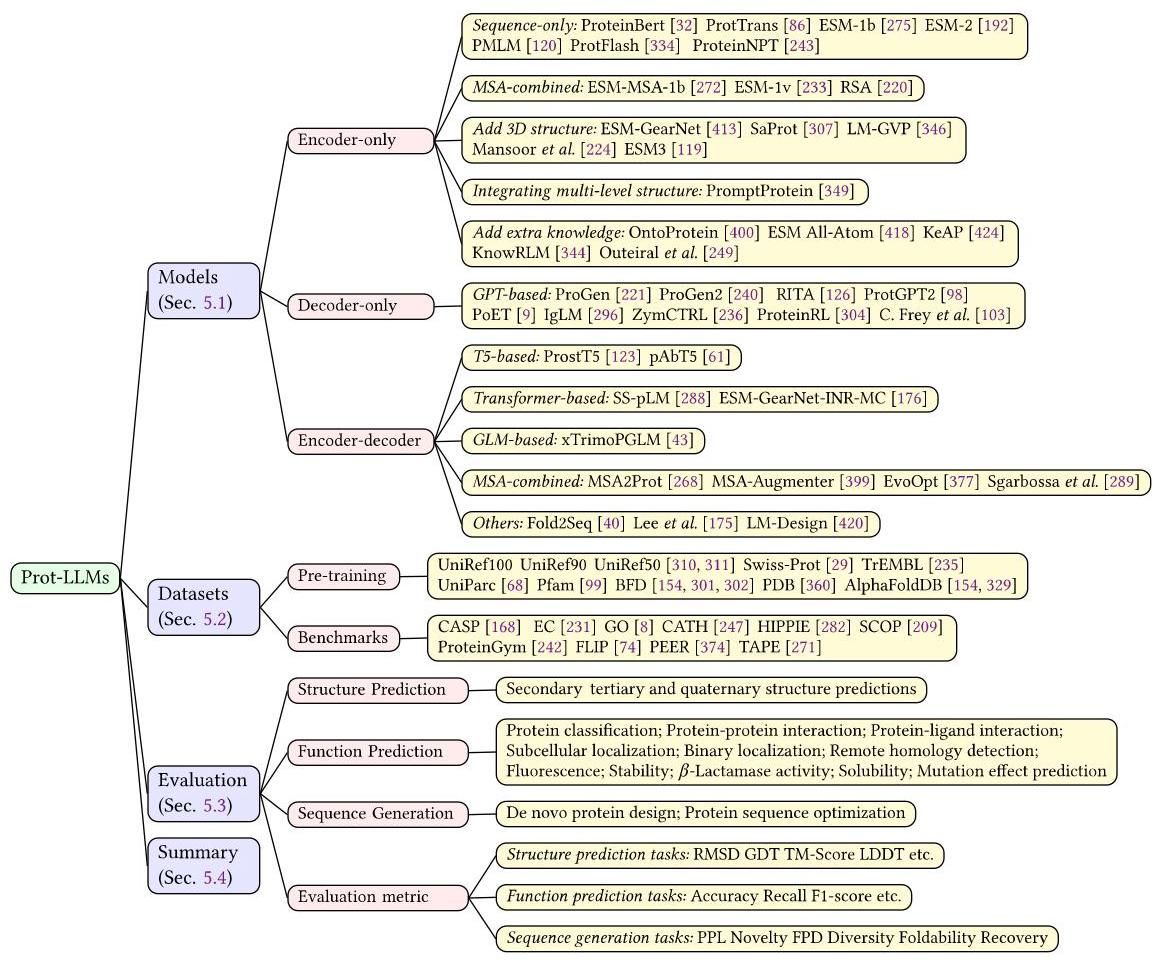


Fig. 8. Chapter overview of Prot-LLMs.

图8. Prot-LLMs章节概述。

# 5.1 Models

# 5.1 模型

In this section, we also classify Prot-LLMs into three main types based on their specific architectures: encoder-only, decoder-only, and encoder-decoder models. These distinct architectures are well-suited for diverse protein research applications. For example, the encoder-only models primarily serve protein function or property prediction purposes, while the decoder-only models are predominantly employed for protein generation tasks. Table 5 provides a summary of Prot-LLMs.

在本节中，我们还根据Prot-LLMs的具体架构将其分为三大类:仅编码器模型、仅解码器模型和编码器-解码器模型。这些不同的架构非常适合各种蛋白质研究应用。例如，仅编码器模型主要用于蛋白质功能或性质预测，而仅解码器模型则主要用于蛋白质生成任务。表5总结了Prot-LLMs。

# 5.1.1 Encoder-only Models.

# 5.1.1 仅编码器模型

Most of the encoder-only Prot-LLMs are built upon the encoder of Transformer, which enables the encoding of protein sequences or structures into fixed-length vector representations. These representations facilitate the identification of patterns and features within proteins, thereby enhancing subsequent analysis and prediction tasks. Self-supervised learning is employed to acquire protein representation, such as masked language modeling (MLM) tasks that reconstruct corrupted tokens based on the surrounding sequence. Prominent pre-trained protein sequence encoders include ESM- 1b [275], ESM-1v [233], and ESM-2 [192], ProteinBert [32], ProtTrans [86]. The ESM series [192, 233, 275] mainly employs a Transformer’s encoder (i.e., BERT [77] and RoBERTa [202]) to predict protein structure and function, which utilizes extensive sequence information from protein databases without relying on manual annotations of the sequences. ProteinBert [32] enhances the classical BERT architecture by incorporating a novel pretraining task for predicting protein functionality. It separates local (character-level) and global (sequence-level) representations, enabling multitask processing in a principled way. ProtTrans [86] trained several auto-encoder models (includes BERT [77], Albert [172], Electra [62]) on a vast of sequence data. Several works have attempted to improve the encoder architecture and the training method. PMLM [120] introduced Pairwise Masked Language Model (PMLM), to directly pretrain the encoder with a focus on capturing co-evolutionary information reflected in residue co-variation within the sequence. This approach considers dependencies between masked tokens. ProtFlash [334] suggested a mixed-chunk attention mechanism that combines multiple positional encodings and a hybrid block attention mechanism incorporating both local and global attention, effectively reducing model complexity. ProteinNPT [243] developed a non-parametric Transformer specifically designed for protein sequences, making it particularly suitable for scenarios involving sparse labels and multi-task learning.

大多数仅编码器的Prot-LLMs都是基于Transformer的编码器构建的，这使得蛋白质序列或结构能够被编码为固定长度的向量表示。这些表示有助于识别蛋白质中的模式和特征，从而增强后续的分析和预测任务。自监督学习被用于获取蛋白质表示，例如基于周围序列重建损坏标记的掩码语言建模(MLM)任务。著名的预训练蛋白质序列编码器包括ESM-1b [275]、ESM-1v [233]和ESM-2 [192]、ProteinBert [32]、ProtTrans [86]。ESM系列[192, 233, 275]主要使用Transformer的编码器(即BERT [77]和RoBERTa [202])来预测蛋白质结构和功能，它利用了蛋白质数据库中的大量序列信息，而不依赖于序列的手动注释。ProteinBert [32]通过引入一种新的预训练任务来预测蛋白质功能，增强了经典的BERT架构。它将局部(字符级)和全局(序列级)表示分开，从而以原则性的方式实现多任务处理。ProtTrans [86]在大量序列数据上训练了多个自动编码器模型(包括BERT [77]、Albert [172]、Electra [62])。一些工作尝试改进编码器架构和训练方法。PMLM [120]引入了成对掩码语言模型(PMLM)，直接预训练编码器，重点捕捉序列中残基共变异反映的共进化信息。这种方法考虑了掩码标记之间的依赖关系。ProtFlash [334]提出了一种混合块注意力机制，结合了多种位置编码和混合块注意力机制，有效地降低了模型复杂性。ProteinNPT [243]开发了一种专门为蛋白质序列设计的非参数Transformer，使其特别适合涉及稀疏标签和多任务学习的场景。

The Multiple Sequence Alignment (MSA) is a computational method that can reveal common features and variation patterns among sequences. By aligning multiple sequences, MSA can unveil shared evolutionary relationships between them and facilitate the identification of functional regions and structural domains. Therefore, MSA has been widely used in protein models. For example, ESM-MSA-1b (MSA Transformer) [272] extends the self-attention mechanism to the MSA setting, which interleaves self-attention across rows and columns to capture dependencies between amino acids and between sequences. Notably, MSA Transformer has become a crucial component of AlphaFold2 [90] and AlphaMissence [54]. However, the utilization of MSA incurs significant computational overheads, and cannot handle orphan proteins. An important alternative method, called Retrieved Sequence Augmentation (RSA) [220], eliminates the need for additional alignment or pre-processing steps. RSA effectively links query protein sequences to a set of structurally or functionally similar sequences in the database and aggregates these sequences for subsequent function prediction tasks.

多序列比对(MSA)是一种计算方法，可以揭示序列之间的共同特征和变异模式。通过比对多个序列，MSA可以揭示它们之间的共同进化关系，并有助于识别功能区域和结构域。因此，MSA已被广泛应用于蛋白质模型中。例如，ESM-MSA-1b(MSA Transformer)[272]将自注意力机制扩展到MSA设置中，交替进行行和列的自注意力，以捕捉氨基酸之间和序列之间的依赖关系。值得注意的是，MSA Transformer已成为AlphaFold2 [90]和AlphaMissence [54]的关键组成部分。然而，使用MSA会带来显著的计算开销，并且无法处理孤儿蛋白质。一种重要的替代方法称为检索序列增强(RSA)[220]，它消除了额外的比对或预处理步骤。RSA有效地将查询蛋白质序列与数据库中结构或功能相似的序列集联系起来，并聚合这些序列用于后续的功能预测任务。

Given that the function of a protein is intricately linked to its structure, another reasonable approach for protein representation would involve the integration of its conformation. There are some recent efforts in this direction. ESM-GearNet [413] combines ESM-1b and GearNet [364] (the state-of-the-art protein structure encoder) to enhance model performance. SaProt [307] introduces a structure-aware vocabulary that integrates residue tokens with structure tokens. LM-GVP [346] compose Transformer blocks and a graph network derived from the protein 3D structure. Mansoor et al. [224] encoded protein sequence and structure through joint training in a semi-supervised manner. PromptPro-tein [349] utilizes prompt-guided multi-task pretraining to fuse different levels of protein structure. ESM3 [119] is an advanced multimodal generative language model that effectively integrates protein sequence, structure, and function analysis. It exhibits exceptional capability in comprehending intricate prompts by leveraging its evolutionary simulation capabilities, making it highly adaptable for biological alignment tasks. It is worth noting that numerous 3D structure encoders (like GearNet [364] and GBPNet [12]) have been proposed to model protein structural information; however, they are beyond the scope of this survey which primarily focuses on LLMs.

鉴于蛋白质的功能与其结构密切相关，另一种合理的蛋白质表示方法将涉及整合其 构象。最近在这一方向上已有一些努力。ESM-GearNet [413]结合了ESM-1b和GearNet [364](最先进的蛋白质结构编码器)以增强模型性能。SaProt [307]引入了一种结构感知词汇表，将残基标记与结构标记整合在一起。LM-GVP [346]由Transformer块和源自蛋白质三维结构的图网络组成。Mansoor等人[224]通过半监督联合训练编码了蛋白质序列和结构。PromptProtein [349]利用提示引导的多任务预训练来融合不同层次的蛋白质结构。ESM3 [119]是一种先进的多模态生成语言模型，有效整合了蛋白质序列、结构和功能分析。它通过利用其进化模拟能力，在理解复杂提示方面表现出卓越的能力，使其在生物对齐任务中具有高度适应性。值得注意的是，许多3D结构编码器(如GearNet [364]和GBPNet [12])已被提出用于建模蛋白质结构信息；然而，它们超出了本调查的范围，本调查主要关注LLMs。

Additionally, some approaches propose to introduce extra knowledge such as Gene Ontology (GO) to enhance the protein representation. OntoProtein [400] considered GO as a factual knowledge graph, which was used to enhance protein representations. ESM All-Atom [418] achieves atom-scale and residue-scale unified molecular modeling by pretraining on multi-scale protein sequences and utilizing a multi-scale position encoding to capture relationships among residues and atoms. KeAP [424] proposed a knowledge-injected auto-encoder that performs token-level knowledge graph exploration for protein representation learning. KnowRLM [344] stands out for its ability to utilize contextual amino acid information from knowledge graphs, thus attaining advantages from both statistical patterns of protein sequences and biochemical properties of amino acids. Some studies improve protein learning representation by enriching pre-trained biological data. For example, Outeiral et al. [249] show that large language models trained on codons, instead of amino acid sequences, provide high-quality representations that outperform comparable state-of-the-art models across a variety of tasks.

此外，一些方法提出引入额外知识，如基因本体(GO)，以增强蛋白质表示。OntoProtein [400]将GO视为事实知识图谱，用于增强蛋白质表示。ESM All-Atom [418]通过在多尺度蛋白质序列上进行预训练并利用多尺度位置编码来捕捉残基和原子之间的关系，实现了原子尺度和残基尺度的统一分子建模。KeAP [424]提出了一种知识注入的自动编码器，用于蛋白质表示学习的标记级知识图谱探索。KnowRLM [344]因其能够利用知识图谱中的上下文氨基酸信息而脱颖而出，从而在蛋白质序列的统计模式和氨基酸的生化特性方面获得优势。一些研究通过丰富预训练的生物数据来改进蛋白质学习表示。例如，Outeiral等人[249]表明，在密码子而非氨基酸序列上训练的大型语言模型提供了高质量的表示，在各种任务中优于可比的最先进模型。

# 5.1.2 Decoder-only Models.

# 5.1.2 仅解码器模型。

Decoder-based Prot-LLMs play a predominant role in the generation of novel proteins, serving as a crucial tool in protein engineering and drug design. GPT [266], a well-established decoder-only architecture, has gained extensive utilization in Prot-LLMs. A representative model is ProGen [221], which utilizes GPT for controllable protein generation. It has been trained on a dataset comprising protein sequences, accompanied by conditioning tags that encode diverse annotations encompassing taxonomic, functional, and locational information. ProGen2 [240] extends the model to 6.4B parameters and is trained on diverse sequence datasets extracted from over one billion proteins from genomic, metagenomic, and immune repertoire databases. ProtGPT2 [98] is another GPT-based autoregressive model that effectively generates protein sequences exhibiting amino acid compositions and disorder propensities comparable to those observed in natural proteins. RITA [126] presents a suite of autoregressive generative models, which is the first systematic investigation of how capabilities evolve with model size for these protein models.

基于解码器的Prot-LLMs在新型蛋白质生成中占据主导地位，是蛋白质工程和药物设计中的关键工具。GPT [266]是一种成熟的仅解码器架构，已在Prot-LLMs中得到广泛应用。代表性模型是ProGen [221]，它利用GPT进行可控蛋白质生成。它已在包含 蛋白质序列的数据集上进行了训练，并附有编码各种注释的条件标签，包括分类学、功能和位置信息。ProGen2 [240]将模型扩展到64亿参数，并在从基因组、宏基因组和免疫库数据库中提取的超过10亿个蛋白质的多样化序列数据集上进行了训练。ProtGPT2 [98]是另一种基于GPT的自回归模型，有效生成具有与天然蛋白质相当的氨基酸组成和无序倾向的蛋白质序列。RITA [126]提出了一套自回归生成模型，这是首次系统研究这些蛋白质模型的能力如何随模型大小演变。

The decoder-based protein language models have been widely applied in specific protein designs. Timothy et al. introduced the protein evolutionary Transformer (PoET), an autoregressive generative model of the distribution over protein families, which can generate sets of related proteins. Zheng et al. proposed the LM-Design [420] that uses language models for structure-based protein design (i.e., inverse folding). ZymCTRL [236] trained a conditional language model on the BRENDA enzyme database [41], aimed at designing customized artificial enzymes by generating specific enzyme classes based on user prompts. IgLM [296] employed autoregressive sequence generation for antibody design. C. Frey et al. [103] evaluate the robustness of our approach on generative modeling of antibody proteins and introduce the distributional conformity score to benchmark protein generative models. ProteinRL [304] is a reinforcement learning framework for fine-tuning generative PLMs for proteins optimized for specific sequence and/or structural properties. ProteinRL fine-tuning guides the PLM towards generating sequences optimized for the defined properties, extending to values rarely or never seen in natural sequences or sequences generated without ProteinRL fine-tuning.

基于解码器的蛋白质语言模型已广泛应用于特定蛋白质设计中。Timothy等人引入了蛋白质进化Transformer(PoET)，这是一种蛋白质家族分布的自回归生成模型，可以生成相关蛋白质集合。Zheng等人提出了LM-Design [420]，它使用语言模型进行基于结构的蛋白质设计(即逆折叠)。ZymCTRL [236]在BRENDA酶数据库[41]上训练了一个条件语言模型，旨在通过基于用户提示生成特定酶类来设计定制的人工酶。IgLM [296]采用自回归序列生成进行抗体设计。C. Frey等人[103]评估了我们在抗体蛋白质生成建模方法上的鲁棒性，并引入了分布一致性评分来基准测试蛋白质生成模型。ProteinRL [304]是一个强化学习框架，用于微调生成性PLMs以优化特定序列和/或结构特性的蛋白质。ProteinRL微调引导PLM生成优化定义特性的序列，扩展到在自然序列或未经过ProteinRL微调生成的序列中很少或从未见过的值。

# 5.1.3 Encoder-decoder Models.

# 5.1.3 编码器-解码器模型。

The encoder-decoder architecture is commonly used for sequence-to-sequence tasks. In the field of protein research, ProstT5 [123], a protein language model based on the T5 architecture [267], is an exemplary encoder-decoder model that facilitates translation between protein sequences and structures. The conversion of protein structure from 3D to

编码器-解码器架构通常用于序列到序列任务。在蛋白质研究领域，ProstT5 [123] 是一种基于T5架构 [267] 的蛋白质语言模型，是一个典型的编码器-解码器模型，它促进了蛋白质序列与结构之间的转换。蛋白质结构从3D到

Table 5. Summary of Prot-LLMs

表5. Prot-LLMs总结

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model | Time | #Parameters | Base model | Pretraining Dataset | Capability | Open- source |
| Encoder-only | ESM-1b [275] | 2020.02 | 650M | RoBERTa | UniRef50 | Secondary struct. pred., Contact pred., etc. | ✓ |
| ESM-MSA-1b [272] | 2021.02 | 100M | ESM-1b | UniRef50 | Secondary struct. pred., Contact pred., etc. | ✓ |
| ESM-1v [233] | 2021.02 | 650M | ESM-1b | UniRef90 | Mutation effect pred. | ✓ |
| ProtTrans [86] | 2021.07 | - | BERT, Albert, Electra | UniRef, BFD | Secondary struct. pred., Func. pred., etc | ✓ |
| PMLM [120] | 2021.07 | 87M - 731M | Trans. enc. | Uniref50/Pfam | Contact pred. | ✘ |
| Mansoor et al. [224] | 2021.09 | 100M | ESM-1b | - | Mutation effect pred. | ✘ |
| ProteinBERT [32] | 2022.02 | 16M | BERT | UniRef90 | Func. pred. | ✓ |
| LM-GVP [346] | 2022.04 | - | Trans. enc | - | Func. pred. | ✓ |
| RSA [220] | 2022.05 | - | ESM-1b | - | Func. pred. | ✓ |
| OntoProtein [400] | 2022.06 | - | BERT | ProteinKG25 | Func. pred. | ✓ |
| ESM-2 [192] | 2022.07 | 8M - 15B | RoBERTa | UniRef50 | Func. pred., Struct. pred. | ✓ |
| PromptProtein [349] | 2023.02 | 650M | RoBERTa | UniRef50, PDB | Func. pred. | ✓ |
| KeAP [424] | 2023.02 | - | RoBERTa | ProteinKG25 | Func. pred. | ✓ |
| ProtFlash [334] | 2023.10 | 79M/174M | Trans. enc | UniRef50 | Func. pred. | ✓ |
| ESM-GearNet [413] | 2023.10 | - | ESM-1b, GearNet | - | Func. pred. | ✓ |
| SaProt [307] | 2023.10 | 650M | BERT | - | Mutation effect pred. | ✓ |
| ProteinNPT [243] | 2023.12 | - | Trans. enc. | - | Fitness pred., Redesign | ✘ |
| Outeiral et al. [249] | 2024.02 |  | Trans. enc. | European Nucleotide Archive | Protein represent learning | ✓ |
| ESM All-Atom [418] | 2024.06 | 35M | RoBERTa | AlphaFold DB | Unified Molecular Modeling | ✘ |
| KnowRLM [344] | 2024.06 | - | Trans. enc. | - | Protein Directed Evolution | ✘ |
| ESM3 [119] | 2024.06 | 98B | RoBERTa | PDB | Seq. pred., Func. pred., Struct. pred. | ✓ |
| Decoder-only | ProGen [221] | 2020.03 | 1.2B | GPT | Uniparc SWISS-Prot | Functional prot. gen. | ✓ |
| ProtGPT2 [98] | 2021.01 | 738M | GPT | Uniref50 | De novo protein design and engineering | ✓ |
| ZymCTRL [236] | 2022.01 | 738M | GPT | BRENDA | Functional enzymes gen. | ✓ |
| RITA [126] | 2022.05 | 1.2B | GPT | UniRef100 | Functional prot. gen. | ✘ |
| IgLM [296] | 2022.12 | 13M | GPT | - | Antibody design | ✓ |
| ProGen2 [240] | 2023.10 |  | GPT | Uniref90, BFD30, PDB | Functional prot. gen. | ✓ |
| ProteinRL [304] | 2023.10 | 764M | GPT | - | Prot. design | ✘ |
| PoET [9] | 2023.11 | 201M | GPT | - | Prot. family. gen. | ✘ |
| C. Frey et al. [103] | 2024.03 | 9.87M/1.03M | GPT | hu4D5 antibody mutant | Functional prot. gen. | ✘ |
| Encoder-Decoder | Fold2Seq [40] | 2021.01 | - | Transformer | - | Prot. design | ✓ |
| MSA2Prot [268] | 2022.04 | - | Transformer | - | Prot. gen., Variant func. pred. | ✘ |
| Sgarbossa et al. [289] | 2023.02 | - | MSA Transformer | - | Prot. gen. | ✓ |
| Lee et al. [175] | 2023.02 | 150M | Transformer | - | Prot. design | ✘ |
| LM-Design [420] | 2023.02 | 664M | Transformer | - | Prot. design | ✓ |
| MSA-Augmenter [399] | 2023.06 | 260M | Transformer | Uniref50 | MSA gen. | ✓ |
| ProstT5 [123] | 2023.07 | 3B | T5 | PDB | Seq.-struct. translation | ✓ |
| xTrimoPGLM [43] | 2023.07 | 100B | GLM | Uniref90. ColdFoldDB | Prot. gen., Func. pred. | ✘ |
| SS-pLM [288] | 2023.08 | 14.8M | Transformer | Uniref50 | Prot. gen. | ✘ |
| pAbT5 [61] | 2023.10 | - | T5 | - | Prot. design | ✘ |
| ESM-GearNet- INR-MC [176] | 2024.04 | - | Transformer | Swiss-Prot. AlphaFoldDB | Prot. gen | ✘ |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 模型 | 时间 | #参数 | 基础模型 | 预训练数据集 | 能力 | 开源 |
| 仅编码器 | ESM-1b [275] | 2020.02 | 650M | RoBERTa | UniRef50 | 二级结构预测，接触预测等 | ✓ |
| ESM-MSA-1b [272] | 2021.02 | 100M | ESM-1b | UniRef50 | 二级结构预测，接触预测等 | ✓ |
| ESM-1v [233] | 2021.02 | 650M | ESM-1b | UniRef90 | 突变效应预测 | ✓ |
| ProtTrans [86] | 2021.07 | - | BERT, Albert, Electra | UniRef, BFD | 二级结构预测，功能预测等 | ✓ |
| PMLM [120] | 2021.07 | 87M - 731M | 编码器 | Uniref50/Pfam | 接触预测 | ✘ |
| Mansoor 等 [224] | 2021.09 | 100M | ESM-1b | - | 突变效应预测 | ✘ |
| ProteinBERT [32] | 2022.02 | 16M | BERT | UniRef90 | 功能预测 | ✓ |
| LM-GVP [346] | 2022.04 | - | 编码器 | - | 功能预测 | ✓ |
| RSA [220] | 2022.05 | - | ESM-1b | - | 功能预测 | ✓ |
| OntoProtein [400] | 2022.06 | - | BERT | ProteinKG25 | 功能预测 | ✓ |
| ESM-2 [192] | 2022.07 | 8M - 15B | RoBERTa | UniRef50 | 功能预测，结构预测 | ✓ |
| PromptProtein [349] | 2023.02 | 650M | RoBERTa | UniRef50, PDB | 功能预测 | ✓ |
| KeAP [424] | 2023.02 | - | RoBERTa | ProteinKG25 | 功能预测 | ✓ |
| ProtFlash [334] | 2023.10 | 79M/174M | 编码器 | UniRef50 | 功能预测 | ✓ |
| ESM-GearNet [413] | 2023.10 | - | ESM-1b, GearNet | - | 功能预测 | ✓ |
| SaProt [307] | 2023.10 | 650M | BERT | - | 突变效应预测 | ✓ |
| ProteinNPT [243] | 2023.12 | - | 编码器 | - | 适应性预测，重新设计 | ✘ |
| Outeiral 等 [249] | 2024.02 |  | 编码器 | 欧洲核苷酸档案库 | 蛋白质表示学习 | ✓ |
| ESM 全原子 [418] | 2024.06 | 35M | RoBERTa | AlphaFold 数据库 | 统一分子建模 | ✘ |
| KnowRLM [344] | 2024.06 | - | 编码器 | - | 蛋白质定向进化 | ✘ |
| ESM3 [119] | 2024.06 | 98B | RoBERTa | PDB | 序列预测，功能预测，结构预测 | ✓ |
| 仅解码器 | ProGen [221] | 2020.03 | 1.2B | GPT | Uniparc SWISS-Prot | 功能蛋白质生成 | ✓ |
| ProtGPT2 [98] | 2021.01 | 738M | GPT | Uniref50 | 从头蛋白质设计与工程 | ✓ |
| ZymCTRL [236] | 2022.01 | 738M | GPT | BRENDA | 功能酶生成 | ✓ |
| RITA [126] | 2022.05 | 1.2B | GPT | UniRef100 | 功能蛋白质生成 | ✘ |
| IgLM [296] | 2022.12 | 13M | GPT | - | 抗体设计 | ✓ |
| ProGen2 [240] | 2023.10 |  | GPT | Uniref90, BFD30, PDB | 功能蛋白质生成 | ✓ |
| ProteinRL [304] | 2023.10 | 764M | GPT | - | 蛋白质设计 | ✘ |
| PoET [9] | 2023.11 | 201M | GPT | - | 蛋白质家族生成 | ✘ |
| C. Frey 等人 [103] | 2024.03 | 9.87M/1.03M | GPT | hu4D5 抗体突变体 | 功能蛋白质生成 | ✘ |
| 编码器-解码器 | Fold2Seq [40] | 2021.01 | - | Transformer | - | 蛋白质设计 | ✓ |
| MSA2Prot [268] | 2022.04 | - | Transformer | - | 蛋白质生成，变体功能预测 | ✘ |
| Sgarbossa 等人 [289] | 2023.02 | - | MSA Transformer | - | 蛋白质生成 | ✓ |
| Lee 等人 [175] | 2023.02 | 150M | Transformer | - | 蛋白质设计 | ✘ |
| LM-Design [420] | 2023.02 | 664M | Transformer | - | 蛋白质设计 | ✓ |
| MSA-Augmenter [399] | 2023.06 | 260M | Transformer | Uniref50 | MSA 生成 | ✓ |
| ProstT5 [123] | 2023.07 | 3B | T5 | PDB | 序列-结构翻译 | ✓ |
| xTrimoPGLM [43] | 2023.07 | 100B | GLM | Uniref90. ColdFoldDB | 蛋白质生成，功能预测 | ✘ |
| SS-pLM [288] | 2023.08 | 14.8M | Transformer | Uniref50 | 蛋白质生成 | ✘ |
| pAbT5 [61] | 2023.10 | - | T5 | - | 蛋白质设计 | ✘ |
| ESM-GearNet- INR-MC [176] | 2024.04 | - | Transformer | Swiss-Prot. AlphaFoldDB | 蛋白质生成 | ✘ |

1D is achieved through the utilization of 3Di-tokens introduced by Foldseek [328]. pAbT5 [61] is another T5-based model that takes into consideration the constraints imposed by protein-protein interactions on generation and design. It can generate complementary heavy or light chains from their pairing partners. xTrimoPGLM [43], based on General Language Model (GLM) architecture [81], is a unified protein language model for understanding and generating protein sequences. It has an unprecedented scale of 100B parameters and consumes 1 trillion training tokens. Considering that the training of protein LLMs involves millions to billions of sequences and billions of parameters, Serrano et al. [288] introduced the Small-Scale protein Language Model (SS-pLM), which significantly reduces the computational burden, democratizing the use of foundational models in protein generation.

一维(1D)是通过利用Foldseek [328]引入的3Di标记实现的。pAbT5 [61]是另一个基于T5的模型，它考虑了蛋白质-蛋白质相互作用对生成和设计的约束。它可以从配对的伙伴中生成互补的重链或轻链。xTrimoPGLM [43]基于通用语言模型(GLM)架构 [81]，是一个用于理解和生成蛋白质序列的统一蛋白质语言模型。它具有前所未有的1000亿参数规模，并消耗了1万亿训练标记。考虑到蛋白质大语言模型(LLMs)的训练涉及数百万到数十亿的序列和数十亿的参数，Serrano等人 [288] 引入了小规模蛋白质语言模型(SS-pLM)，显著减少了计算负担，使基础模型在蛋白质生成中的应用更加普及。

Building upon the Transformer, several models have incorporated MSA modules. MSA2Prot [268] is an MSA-to-protein Transformer, developed axial and cross attentions for encoder and decoder to model sequence probabilities autoregressively. MSA-Augmenter [399] utilizes protein-specific attention mechanisms and large-scale MSAs to generate novel protein sequences that are not present in existing databases. Sgarbossa et al. [289] proposed an iterative method that directly utilizes the masked language modeling to generate sequences within the MSA Transformer framework.

在Transformer的基础上，多个模型已经整合了多序列比对(MSA)模块。MSA2Prot [268] 是一个MSA到蛋白质的Transformer，为编码器和解码器开发了轴向和交叉注意力，以自回归方式建模序列概率。MSA-Augmenter [399] 利用蛋白质特定的注意力机制和大规模MSA生成现有数据库中不存在的新蛋白质序列。Sgarbossa等人 [289] 提出了一种迭代方法，直接在MSA Transformer框架内利用掩码语言建模生成序列。

In the field of structure-based protein design, Cao et al. [40] proposed a Fold2Seq framework, a generative approach based on the Transformer architecture, which aims at designing protein sequences conditioned on specific target folds. To capture intricate sequence-structure relationships, Fold2Seq jointly learns embeddings for both sequences and folding structures, which are derived from the density of secondary structure elements in 3D voxels.

在基于结构的蛋白质设计领域，Cao等人 [40] 提出了Fold2Seq框架，这是一种基于Transformer架构的生成方法，旨在设计特定目标折叠条件下的蛋白质序列。为了捕捉复杂的序列-结构关系，Fold2Seq联合学习序列和折叠结构的嵌入，这些嵌入源自3D体素中二级结构元素的密度。

While recent studies have successfully employed sequence- or structure-based representations to address multiple tasks in protein science, there has been significant oversight in incorporating protein surface information, a critical factor for protein function. Therefore, ESM-GearNet-INR-MC [176] presents a pre-training strategy that incorporates information from protein sequences, 3D structures, and surfaces to improve protein representation learning.

尽管最近的研究已成功采用基于序列或结构的表示来解决蛋白质科学中的多个任务，但在整合蛋白质表面信息方面存在显著疏忽，而蛋白质表面信息是蛋白质功能的关键因素。因此，ESM-GearNet-INR-MC [176] 提出了一种预训练策略，整合了蛋白质序列、3D结构和表面信息，以改进蛋白质表示学习。

Additionally, some approaches utilize reinforcement learning to improve the generation quality. Lee et al. [175] efficiently navigates the latent representation space rather than the protein sequence space by formulating the sequence design task as a Markov decision process and employing model-based reinforcement learning. This approach enables them to generate proteins with enhanced functionality and cellular fitness.

此外，一些方法利用强化学习来提高生成质量。Lee等人 [175] 通过将序列设计任务表述为马尔可夫决策过程并采用基于模型的强化学习，有效地导航潜在表示空间而不是蛋白质序列空间。这种方法使他们能够生成具有增强功能和细胞适应性的蛋白质。

Table 6. Summary of datasets for Prot-LLMs

表6. Prot-LLMs数据集总结

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dataset | Last updated | Scale | Keywords |
| Pretraining | UniRef100 [310, 311] | 2023.11 | 314M | Complete collection of protein sequences from UniProtKB |
| UniRef90 [310, 311] | 150M | Cluster UniRef100 sequences at sequence identity level |
| UniRef50 [310, 311] | 53M | Cluster UniRef100 sequences at sequence identity level |
| UniProtKB/Swiss-Prot [29] | 2023.11 | 570K | High-quality, manually curated protein sequence database |
| UniProtKB/TrEMBL [235] | 251M | Computationally annotated protein sequence database |
| UniParc [68] | 2023.11 | 632M | Comprehensive and non-redundant protein sequence database |
| Pfam [99] | 2023.09 | 47M | Protein family database |
| BFD | 2021.07 | 2.5B | Protein sequences from multiple databases and resources |
| PDB [360] | 2023.12 | 214K | Experimentally determined accurate protein structures |
| PIR [357] | 2024.03 | 513M | Information on functionally annotated protein sequences |
| AlphaFoldDB [154, 329] | 2021.11 | 200M | Protein structures predicted by AlphaFold |
| Benchmark | CASP [168] | 2022.01 | - | Structure prediction competition |
| EC [231] | 2023.11 | 2.6 M | Enzymes classification database |
| GO [8] | 2023.11 | 1.5M | Gene Ontology knowledgebase |
| CATH [247] | 2023.02 | 151M | Classification of protein structures |
| HIPPIE [282] | 2022.04 | 39K | Protein-protein interaction networks |
| SCOP [209] | 2023.01 | 914K | Protein structure classification |
| ProteinGym [242] | 2022.12 |  | Predict the effects of protein mutations |
| FLIP [74] | 2022.01 |  | Fitness landscape prediction (AAV, Thermostability, GB1) |
| PEER [374] | 2022.11 |  | Protein function, Localization, Structure prediction, Protein-protein interaction, Protein-ligand interaction |
| TAPE [271] | 2021.09 |  | Remote homology detection, Secondary structure, Contact, Fluorescence, Stability prediction |
| Reactome [142] | 2023.12 |  | Biological interactions and pathways |
| STRING [312] | 2022.11 | 59.3M | Protein-Protein interaction networks |
| BioGRID [248] | 2023.12 | 271k | Genetic and protein interactions |
| InterPro [253] | 2024.01 |  | Classification of protein families |
| DIP [361] | 2020.02 | 81K | Protein-protein interaction information |
| PROSITE [297] | 2024.03 | 103K | Collection of signatures that identify patterns or profiles in proteins |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 数据集 | 最后更新 | 规模 | 关键词 |
| 预训练 | UniRef100 [310, 311] | 2023.11 | 314M | 来自UniProtKB的完整蛋白质序列集合 |
| UniRef90 [310, 311] | 150M | 在 序列同一性水平上聚类UniRef100序列 |
| UniRef50 [310, 311] | 53M | 在 序列同一性水平上聚类UniRef100序列 |
| UniProtKB/Swiss-Prot [29] | 2023.11 | 570K | 高质量、手动整理的蛋白质序列数据库 |
| UniProtKB/TrEMBL [235] | 251M | 计算注释的蛋白质序列数据库 |
| UniParc [68] | 2023.11 | 632M | 全面且非冗余的蛋白质序列数据库 |
| Pfam [99] | 2023.09 | 47M | 蛋白质家族数据库 |
| BFD | 2021.07 | 2.5B | 来自多个数据库和资源的蛋白质序列 |
| PDB [360] | 2023.12 | 214K | 实验确定的精确蛋白质结构 |
| PIR [357] | 2024.03 | 513M | 功能注释蛋白质序列的信息 |
| AlphaFoldDB [154, 329] | 2021.11 | 200M | 由AlphaFold预测的蛋白质结构 |
| 基准 | CASP [168] | 2022.01 | - | 结构预测竞赛 |
| EC [231] | 2023.11 | 2.6 M | 酶分类数据库 |
| GO [8] | 2023.11 | 1.5M | 基因本体知识库 |
| CATH [247] | 2023.02 | 151M | 蛋白质结构分类 |
| HIPPIE [282] | 2022.04 | 39K | 蛋白质-蛋白质相互作用网络 |
| SCOP [209] | 2023.01 | 914K | 蛋白质结构分类 |
| ProteinGym [242] | 2022.12 |  | 预测蛋白质突变的影响 |
| FLIP [74] | 2022.01 |  | 适应性景观预测(AAV、热稳定性、GB1) |
| PEER [374] | 2022.11 |  | 蛋白质功能、定位、结构预测、蛋白质-蛋白质相互作用、蛋白质-配体相互作用 |
| TAPE [271] | 2021.09 |  | 远程同源性检测、二级结构、接触、荧光、稳定性预测 |
| Reactome [142] | 2023.12 |  | 生物相互作用和通路 |
| STRING [312] | 2022.11 | 59.3M | 蛋白质-蛋白质相互作用网络 |
| BioGRID [248] | 2023.12 | 271k | 遗传和蛋白质相互作用 |
| InterPro [253] | 2024.01 |  | 蛋白质家族分类 |
| DIP [361] | 2020.02 | 81K | 蛋白质-蛋白质相互作用信息 |
| PROSITE [297] | 2024.03 | 103K | 识别蛋白质中模式或特征的签名集合 |

# 5.2 Datasets

# 5.2 数据集

Protein datasets can be classified into two categories based on the availability of annotations: pre-training datasets and benchmarks. The former, which lacks labels, is commonly employed for self-supervised pre-training, whereas the latter, with labeled data, is utilized for supervised fine-tuning or model evaluation. The following present popular pre-training datasets and benchmarks for Prot-LLMs, as summarized in Table 6.

蛋白质数据集根据注释的可用性可以分为两类:预训练数据集和基准数据集。前者缺乏标签，通常用于自监督预训练，而后者包含标注数据，用于监督微调或模型评估。以下介绍了Prot-LLMs常用的预训练数据集和基准数据集，如表6所示。

# Pre-training Datasets:

# 预训练数据集:

* UniProtKB (Swiss-Prot and TrEMBL) [29, 235]. It is a central hub for the collection of functional information on proteins, with extensive and accurately annotated protein sequence data. It comprises Swiss-Prot and TrEMBL, where the former is known for its high level of annotations, accuracy, and reliability, because they are manually annotated and reviewed by experts. TrEMBL contains protein sequences that are automatically annotated and have not yet been reviewed by human curators.
* UniProtKB(Swiss-Prot和TrEMBL)[29, 235]。它是收集蛋白质功能信息的中心枢纽，拥有广泛且准确注释的蛋白质序列数据。它由Swiss-Prot和TrEMBL组成，前者以其高水平的注释、准确性和可靠性著称，因为它们是由专家手动注释和审查的。TrEMBL包含自动注释且尚未经过人工审查的蛋白质序列。
* UniRef100, UniRef90, UniRef50 [310, 311]. UniRef (UniProt Reference Clusters) is a system of databases that provide clustered sets of protein sequences from UniProtKB and selected UniProt Archive records. The primary purpose of UniRef is to provide a comprehensive and non-redundant dataset that improves the efficiency of sequence similarity searches. The databases are organized into three main clusters: UniRef100, UniRef90, and UniRef50, where the latter two are created by clustering UniRef100 sequences that have at least 90% and 50% sequence identity to each other, respectively.
* UniRef100、UniRef90、UniRef50 [310, 311]。UniRef(UniProt参考聚类)是一个数据库系统，提供来自UniProtKB和部分UniProt Archive记录的蛋白质序列聚类集。UniRef的主要目的是提供一个全面且非冗余的数据集，以提高序列相似性搜索的效率。这些数据库分为三个主要聚类:UniRef100、UniRef90和UniRef50，后两者是通过将UniRef100序列分别聚类为至少90%和50%序列相似性生成的。
* UniParc [68]. UniParc (UniProt Archive) is an important component of the UniProt database family, serving as a resource for storing non-redundant protein sequences. Unlike other members of the UniProt database family, UniParc does not provide protein annotation and functional information; it focuses solely on storing protein sequences themselves.
* UniParc [68]。UniParc(UniProt Archive)是UniProt数据库家族的重要组成部分，作为存储非冗余蛋白质序列的资源。与UniProt数据库家族的其他成员不同，UniParc不提供蛋白质注释和功能信息；它仅专注于存储蛋白质序列本身。
* Pfam [99]. Pfam is a widely used protein family database that is used for the classification and annotation of protein sequences. It plays a crucial role in understanding protein function, structure, and evolution. It is built based on the shared conserved domains and functional modules found in protein sequences, providing researchers with classification, annotation, and functional prediction of protein sequences.
* Pfam [99]。Pfam是一个广泛使用的蛋白质家族数据库，用于蛋白质序列的分类和注释。它在理解蛋白质功能、结构和进化方面起着关键作用。它基于蛋白质序列中共享的保守域和功能模块构建，为研究人员提供蛋白质序列的分类、注释和功能预测。
* BFD [154, 301, 302]. The Big Fantastic Database (BFD) is a comprehensive collection of protein sequences gathered from various sources. It includes a vast array of sequence data (2.5B), which is crucial for training machine learning models in tasks like protein structure prediction.
* BFD [154, 301, 302]。Big Fantastic Database(BFD)是一个从各种来源收集的蛋白质序列的综合集合。它包含大量的序列数据(2.5B)，这对于训练机器学习模型进行蛋白质结构预测等任务至关重要。
* PDB [360].The Protein Data Bank (PDB) is a global protein structure database used for storing and sharing resolved 3D protein structure data. It archives detailed information about the 3D structures of proteins, nucleic acids, and complex assemblies. These structures are determined using experimental methods like X-ray crystallography, NMR spectroscopy, and, more recently, cryo-electron microscopy.
* PDB [360]。Protein Data Bank(PDB)是一个全球蛋白质结构数据库，用于存储和共享解析的3D蛋白质结构数据。它存档了蛋白质、核酸和复杂组装的3D结构的详细信息。这些结构是通过X射线晶体学、核磁共振光谱以及最近的冷冻电镜等实验方法确定的。
* PIR [357].The Protein Information Resource(PIR) is a popular protein sequence database that provides information on functionally annotated protein sequences. PIR maintains three databases, the Protein Sequence Database (PSD), the Non-redundant Reference (NREF) sequence database, and the integrated Protein Classification (iProClass) database, which contains annotated protein sequences, classification information, and protein family, function, and structure information.
* PIR [357]。Protein Information Resource(PIR)是一个流行的蛋白质序列数据库，提供功能注释的蛋白质序列信息。PIR维护三个数据库:蛋白质序列数据库(PSD)、非冗余参考(NREF)序列数据库和集成蛋白质分类(iProClass)数据库，其中包含注释的蛋白质序列、分类信息以及蛋白质家族、功能和结构信息。
* AlphaFoldDB [154, 329] AlphaFold Database comprises protein structure predictions generated by DeepMind’s AlphaFold system, offering accurate predictions of protein 3D structures. While these structures are not experimentally determined, they offer high-confidence models that can be incredibly useful, especially for proteins whose structures have not yet been solved experimentally.
* AlphaFoldDB [154, 329]。AlphaFold数据库包含由DeepMind的AlphaFold系统生成的蛋白质结构预测，提供准确的蛋白质3D结构预测。虽然这些结构不是通过实验确定的，但它们提供了高置信度的模型，特别适用于尚未通过实验解析结构的蛋白质。

# Benchmarks:

# 基准数据集:

* CASP [168]. The Critical Assessment of Structure Prediction (CASP) is an influential initiative in the field of protein structure prediction. The goal of CASP is to advance the methods used in predicting protein structures from their amino acid sequences. CASP is essentially a series of competitions, which provides a unified benchmark for evaluating the performance of different methods in protein structure prediction.
* CASP [168]。Critical Assessment of Structure Prediction(CASP)是蛋白质结构预测领域的一个重要倡议。CASP的目标是推进从氨基酸序列预测蛋白质结构的方法。CASP本质上是一系列竞赛，为评估不同方法在蛋白质结构预测中的性能提供了一个统一的基准。
* EC [231]. The Enzyme Commission (EC) dataset is a standardized dataset used to describe the classification and functions of enzymes. Enzymes are protein molecules that are widely present in living organisms and play a crucial role in catalyzing chemical reactions. The EC dataset contains a large number of enzyme records, each with a unique EC number and corresponding description. These descriptions include the enzyme’s name, reaction type, catalytic substrates, and products.
* EC [231]。Enzyme Commission(EC)数据集是一个标准化的数据集，用于描述酶的分类和功能。酶是广泛存在于生物体中的蛋白质分子，在催化化学反应中起着关键作用。EC数据集包含大量酶记录，每个记录都有一个唯一的EC编号和相应的描述。这些描述包括酶的名称、反应类型、催化底物和产物。
* GO [8]. The Gene Ontology (GO) dataset is a widely used gene annotation and functional classification system for describing the biological functions of genes and proteins. It provides a standardized framework for representing gene and gene product attributes across species and databases, helping researchers understand their functions in terms of cellular processes, molecular functions, and cellular components.
* GO [8]。Gene Ontology(GO)数据集是一个广泛使用的基因注释和功能分类系统，用于描述基因和蛋白质的生物学功能。它提供了一个标准化的框架，用于跨物种和数据库表示基因和基因产物的属性，帮助研究人员从细胞过程、分子功能和细胞组分的角度理解它们的功能。
* CATH [247]. The acronym CATH stands for the four main levels at which it classifies protein domains: Class, Architecture, Topology, and Homologous superfamily. CATH is widely used for protein structure classification and annotation. It offers a comprehensive framework for systematically classifying protein structures while providing crucial insights into their function and evolutionary aspects.
* CATH [247]。CATH 是蛋白质结构分类的四个主要级别的缩写:类别(Class)、架构(Architecture)、拓扑(Topology)和同源超家族(Homologous superfamily)。CATH 广泛用于蛋白质结构分类和注释。它提供了一个全面的框架，用于系统地分类蛋白质结构，同时提供对其功能和进化方面的重要见解。
* HIPPIE [282]. The HIPPIE database is a comprehensive database that encompasses information on protein-protein interactions in humans. The data originates from diverse sources, including experimental techniques such as yeast two-hybrid assays, mass spectrometry analysis, and documented interaction relationships reported in the scientific literature. These datasets are meticulously harmonized and standardized to provide precise and dependable insights into protein-protein interactions. This invaluable resource finds extensive applications in functional annotation, systems biology research, drug target discovery, and beyond.
* HIPPIE [282]。HIPPIE 数据库是一个包含人类蛋白质-蛋白质相互作用信息的综合数据库。数据来源于多种来源，包括酵母双杂交实验、质谱分析等实验技术，以及科学文献中报道的相互作用关系。这些数据集经过精心协调和标准化，以提供精确可靠的蛋白质-蛋白质相互作用见解。这一宝贵资源在功能注释、系统生物学研究、药物靶点发现等领域有广泛应用。
* SCOP [209]. SCOP (Structural Classification of Proteins) is a database used for protein structure classification. It helps researchers understand the structure, function, and evolutionary relationships of proteins by assigning protein structures to different levels and categories. The SCOP database employs a hierarchical classification system consisting of four levels: Class, Fold, Superfamily, and Family. Each level has specific definitions and criteria to group protein structures into their respective categories.
* SCOP [209]。SCOP(蛋白质结构分类数据库)是一个用于蛋白质结构分类的数据库。它通过将蛋白质结构分配到不同的级别和类别，帮助研究人员理解蛋白质的结构、功能和进化关系。SCOP 数据库采用了一个由四个级别组成的层次分类系统:类别(Class)、折叠(Fold)、超家族(Superfamily)和家族(Family)。每个级别都有特定的定义和标准，用于将蛋白质结构分组到各自的类别中。
* ProteinGym [241]. The ProteinGym platform comprises a comprehensive collection of Deep Mutational Scanning (DMS) assays meticulously curated to evaluate mutation effect predictors in determining the fitness of mutated proteins. It encompasses two distinct benchmarks: a substitution benchmark encompassing experimental characterization of 1.5 million missense variants across 87 DMS assays, and an indel benchmark comprising analysis of 300,000 mutants across 7 DMS assays.
* ProteinGym [241]。ProteinGym 平台包含一个精心策划的深度突变扫描(DMS)实验集合，用于评估突变效应预测器在确定突变蛋白质适应性方面的表现。它包括两个不同的基准:一个替换基准，涵盖 87 个 DMS 实验中的 150 万个错义变体的实验表征；另一个插入缺失基准，涵盖 7 个 DMS 实验中的 30 万个突变体的分析。
* FLIP [74]. FLIP (Functional Learning for Protein Engineering) is a benchmark designed to encourage rapid scoring of representation learning for protein engineering. Currently, FLIP covers experimental data on adenovirus stability for gene therapy, stability of protein domain B1, immunoglobulin binding, and thermal stability of multiple protein families.
* FLIP [74]。FLIP(蛋白质工程的功能学习)是一个旨在鼓励快速评分蛋白质工程表示学习的基准。目前，FLIP 涵盖了基因治疗中腺病毒稳定性的实验数据、蛋白质域 B1 的稳定性、免疫球蛋白结合以及多个蛋白质家族的热稳定性。
* PEER [374]. The PEER benchmark is a comprehensive and multi-task benchmark for protein sequence understanding, encompassing 17 tasks across five categories: protein function, localization, structure predictions, protein-protein interaction prediction and protein-ligand interaction prediction. This benchmark evaluates diverse sequence-based methodologies for each task, including conventional feature engineering approaches, various sequence encoding methods, as well as large-scale pre-trained protein language models under both single-task learning and multi-task learning settings.
* PEER [374]。PEER 基准是一个全面的多任务基准，用于蛋白质序列理解，涵盖五个类别的 17 个任务:蛋白质功能、定位、结构预测、蛋白质-蛋白质相互作用预测和蛋白质-配体相互作用预测。该基准评估了每个任务中基于序列的多种方法，包括传统的特征工程方法、各种序列编码方法，以及在单任务学习和多任务学习设置下的大规模预训练蛋白质语言模型。
* TAPE [271]. The Tasks Assessing Protein Embeddings (TAPE) dataset is a collection of five biologically relevant semi-supervised learning tasks spanning different domains of protein biology, including secondary structure prediction, contact prediction, remote homology prediction, stability prediction, and fluorescence prediction. TAPE divides the tasks into specific training, validation, and testing sets to ensure the evaluation of each task’s biological relevance and generalizability to real-life scenarios. TAPE serves as a benchmark for a range of semi-supervised protein representation learning methods, encompassing recent advances as well as established sequence-based learning techniques.
* TAPE [271]。TAPE(蛋白质嵌入评估任务)数据集是一个包含五个生物学相关的半监督学习任务的集合，涵盖蛋白质生物学的不同领域，包括二级结构预测、接触预测、远程同源预测、稳定性预测和荧光预测。TAPE 将任务划分为特定的训练、验证和测试集，以确保评估每个任务的生物学相关性和对现实场景的泛化能力。TAPE 作为一系列半监督蛋白质表示学习方法的基准，涵盖了最新进展以及已建立的基于序列的学习技术。
* Reactome [142]. Reactome is a database focused on human biomolecular reactions and pathways. It provides detailed networks of biochemical reactions, including proteins, small molecules, gene expression events, and their associated biological processes. Reactome can be used to validate the model’s predictions about the roles of proteins in biological processes.
* Reactome [142]。Reactome 是一个专注于人类生物分子反应和通路的数据库。它提供了详细的生化反应网络，包括蛋白质、小分子、基因表达事件及其相关的生物过程。Reactome 可用于验证模型对蛋白质在生物过程中作用的预测。
* STRING [312]. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a comprehensive protein-protein interaction network database. It contains a wealth of information about known and predicted direct (physical) and indirect (functional) interactions. Through STRING, researchers can explore whether the proteins are in known interaction networks, or if they participate in unknown yet biologically plausible new interactions, which is crucial for understanding protein functions and complex biological pathways.
* STRING [312]。STRING(搜索相互作用基因/蛋白质的工具)是一个全面的蛋白质-蛋白质相互作用网络数据库。它包含了大量关于已知和预测的直接(物理)和间接(功能)相互作用的信息。通过 STRING，研究人员可以探索蛋白质是否在已知的相互作用网络中，或者它们是否参与了未知但生物学上合理的新相互作用，这对于理解蛋白质功能和复杂生物通路至关重要。
* BioGRID [248]. BioGRID (Biological General Repository for Interaction Datasets) is an online database that specializes in recording and curating protein-protein interactions, genetic interactions, chemical associations, and post-translational modifications from scientific literature. BioGRID serves as a critical tool for validating the predictive capabilities of these models regarding protein interactions and modifications. By leveraging the extensive interaction data in BioGRID, researchers can assess the accuracy of the model’s predictions about how proteins interact with each other and their roles in cellular processes.
* BioGRID [248]。BioGRID(生物相互作用数据集通用存储库)是一个在线数据库，专门记录和整理科学文献中的蛋白质-蛋白质相互作用、遗传相互作用、化学关联和翻译后修饰。BioGRID 是验证这些模型在蛋白质相互作用和修饰方面预测能力的关键工具。通过利用 BioGRID 中广泛的相互作用数据，研究人员可以评估模型对蛋白质如何相互作用及其在细胞过程中作用的预测的准确性。
* InterPro [253]. InterPro is a comprehensive resource that offers functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. It integrates diverse information about protein families, domains, and functional sites from multiple reference databases. LLMs can utilize InterPro to annotate predicted protein sequences with potential functional domains, sites, and family memberships. Furthermore, InterPro’s comprehensive collection of protein signatures aids in validating and refining the predictions made by protein sequence models, ensuring a higher level of accuracy and reliability in protein function prediction and analysis.
* InterPro [253]。InterPro 是一个综合资源，通过对蛋白质序列进行分类并预测域和重要位点的存在，提供蛋白质序列的功能分析。它整合了来自多个参考数据库的关于蛋白质家族、域和功能位点的多样化信息。LLMs 可以利用 InterPro 对预测的蛋白质序列进行潜在功能域、位点和家族成员的注释。此外，InterPro 的全面蛋白质签名集合有助于验证和优化蛋白质序列模型的预测，确保蛋白质功能预测和分析的更高准确性和可靠性。
* DIP [361]. The Database of Interacting Proteins(DIP) is a database that contains protein-protein interaction information that has been compiled through both manual curations and computational methods. It is useful for understanding protein functions, and their relationships with other proteins. It can also be used to study the properties of networks of interacting proteins, evaluate predictions of protein-protein interactions, and explore the evolution of these interactions.
* DIP [361]。相互作用蛋白质数据库(DIP)是一个包含蛋白质-蛋白质相互作用信息的数据库，这些信息通过手动整理和计算方法进行编译。它有助于理解蛋白质功能及其与其他蛋白质的关系。它还可用于研究相互作用蛋白质网络的性质，评估蛋白质-相互作用的预测，并探索这些相互作用的进化。
* PROSITE [297]. The PROSITE is a collection of signatures that identify patterns or profiles in proteins, which can provide information on their biological functions. The signatures in the database are linked to annotation documents that provide information on the protein family or domain detected, including its name, function, 3D structure, and references.
* PROSITE [297]。PROSITE 是一个识别蛋白质中模式或特征的签名集合，可以提供有关其生物学功能的信息。数据库中的签名与注释文档相关联，这些文档提供了有关检测到的蛋白质家族或结构域的信息，包括其名称、功能、三维结构和参考文献。

# 5.3 Evaluation

# 5.3 评估

In the realm of protein computational models, Prot-LLMs are primarily evaluated based on their capabilities in three critical areas: protein function prediction, protein sequence generation, and protein structure prediction. Protein

在蛋白质计算模型领域，Prot-LLMs 主要基于其在三个关键领域的能力进行评估:蛋白质功能预测、蛋白质序列生成和蛋白质结构预测。蛋白质

Structure Prediction. It is a crucial area within the fields of bioinformatics and computational biology. It involves the prediction of the three-dimensional structure of proteins given an input sequence, which includes determining the atomic coordinates and the topological relationships between atoms. Predicting protein structures is of significant importance for understanding protein function, drug design, and biomedical research. This task includes tertiary and quaternary structure predictions. The former is critical for understanding the functional mechanisms of a protein, and the latter plays a vital role in understanding the functionality of multi-subunit complexes like hemoglobin. The models for protein structure prediction often rely on encoder-based Prot-LLMs (i.e., ESM series [192, 233, 275]) to extract the sequence information, followed by structure prediction modules.

结构预测。这是生物信息学和计算生物学领域中的一个关键领域。它涉及根据输入序列预测蛋白质的三维结构，包括确定原子坐标和原子之间的拓扑关系。预测蛋白质结构对于理解蛋白质功能、药物设计和生物医学研究具有重要意义。该任务包括三级和四级结构预测。前者对于理解蛋白质的功能机制至关重要，后者在理解多亚基复合物(如血红蛋白)的功能中起着至关重要的作用。蛋白质结构预测模型通常依赖于基于编码器的 Prot-LLMs(即 ESM 系列 [192, 233, 275])来提取序列信息，然后通过结构预测模块进行预测。

Protein Function Prediction. This task aims to determine the biological function of proteins, including how they operate within an organism and their interactions with other biomolecules. The prediction of protein function encompasses a multitude of subtasks, enumerated below.

蛋白质功能预测。该任务旨在确定蛋白质的生物学功能，包括它们在生物体内的运作方式以及它们与其他生物分子的相互作用。蛋白质功能预测涵盖了许多子任务，如下所述。

* Protein Classification. This refers to the process of categorizing and grouping proteins based on their structure, function, sequence similarity, or other features.
* 蛋白质分类。这是指根据蛋白质的结构、功能、序列相似性或其他特征对蛋白质进行分类和分组的过程。
* Protein-Protein Interaction Prediction. This refers to the focus on identifying and predicting interactions between different proteins. Proteins often interact with each other to perform various biological functions such as signal transduction, enzyme catalysis, and assembly of protein complexes.
* 蛋白质-蛋白质相互作用预测。这是指专注于识别和预测不同蛋白质之间的相互作用。蛋白质通常相互作用以执行各种生物学功能，如信号转导、酶催化和蛋白质复合物的组装。
* Protein-Ligand Interaction Prediction. This refers to the process of predicting the binding interactions between a protein and a small molecule ligand.
* 蛋白质-配体相互作用预测。这是指预测蛋白质与小分子配体之间结合相互作用的过程。
* Subcellular Localization Prediction. This refers to the prediction of the subcellular location of a protein within a cell. The subcellular localization of a protein is crucial for its function and interactions.
* 亚细胞定位预测。这是指预测蛋白质在细胞内的亚细胞位置。蛋白质的亚细胞定位对其功能和相互作用至关重要。
* Binary Localization Prediction. This refers to the process of predicting the subcellular localization of a protein into one of two possible categories.
* 二元定位预测。这是指将蛋白质的亚细胞定位预测为两个可能类别之一的过程。
* Remote Homology Detection. This refers to the process of identifying distant relationships between protein sequences. In the protein world, homology refers to the shared ancestry or evolutionary relationship between two or more proteins.
* 远程同源性检测。这是指识别蛋白质序列之间远缘关系的过程。在蛋白质世界中，同源性指的是两个或多个蛋白质之间的共同祖先或进化关系。
* Fluorescence Landscape Prediction. This refers to the prediction of protein sequences to evaluate their fluorescence spectral characteristics under different environmental conditions.
* 荧光景观预测。这是指预测蛋白质序列以评估其在不同环境条件下的荧光光谱特性。
* Stability Prediction. This refers to the process of predicting the stability of a protein to assess its stability under specific conditions. Protein stability represents its ability to maintain the correct folding state within a biological organism or specific environment.
* 稳定性预测。这是指预测蛋白质的稳定性以评估其在特定条件下的稳定性的过程。蛋白质稳定性代表其在生物体或特定环境中维持正确折叠状态的能力。
* -Lactamase Activity Prediction. This refers to the process of predicting the activity of a protein as a - lactamase enzyme, which is capable of hydrolyzing -lactam bonds and is commonly involved in antibiotic degradation and the development of antibiotic resistance.
* -内酰胺酶活性预测。这是指预测蛋白质作为 -内酰胺酶的活性的过程，该酶能够水解 -内酰胺键，通常参与抗生素降解和抗生素耐药性的发展。
* Solubility Prediction. This refers to the process of predicting the solubility of a compound in a solution. Solubility represents the maximum amount of a compound that can dissolve in a solvent at a specific temperature and pressure.
* 溶解度预测。这是指预测化合物在溶液中的溶解度的过程。溶解度代表在特定温度和压力下，化合物在溶剂中能够溶解的最大量。
* Mutation Effect Prediction. It focuses on understanding how genetic mutations can affect protein function, a critical aspect due to their potential to induce diseases or modify biological processes. Prot-LLMs aim to identify potential mutation sites and predict the functional consequences of these mutations.
* 突变效应预测。它专注于理解基因突变如何影响蛋白质功能，这是一个关键方面，因为它们有可能引发疾病或改变生物过程。Prot-LLMs旨在识别潜在的突变位点并预测这些突变的功能后果。

Protein Sequence Generation. It refers to the computational process of creating novel amino acid sequences that could potentially form functional proteins, which can be used for various applications such as drug design, enzyme engineering, and fundamental biological research. The task of protein sequence generation can be roughly classified into two categories:

蛋白质序列生成。它指的是通过计算过程创建可能形成功能蛋白质的新氨基酸序列，这些序列可用于药物设计、酶工程和基础生物学研究等多种应用。蛋白质序列生成任务大致可分为两类:

* De Novo Protein Design. It involves creating entirely new protein sequences that are not based on existing proteins. This process relies heavily on computational methods to design proteins that can perform desired functions, such as binding to a particular target. The autoregressive generative models (e.g., ProGen series) are often employed for the tasks of protein sequence generation.
* 从头蛋白质设计(De Novo Protein Design)。它涉及创建完全不基于现有蛋白质的全新蛋白质序列。这一过程主要依赖于计算方法来设计能够执行特定功能(如与特定靶标结合)的蛋白质。自回归生成模型(例如ProGen系列)通常用于蛋白质序列生成任务。
* Protein Sequence Optimization. It focuses on modifying existing protein sequences to enhance or alter their functions, employing techniques such as directed evolution. Directed evolution is particularly useful for improving the properties of existing proteins, such as increasing their stability, altering their substrate specificity, or enhancing their catalytic efficiency.
* 蛋白质序列优化。它专注于修改现有蛋白质序列以增强或改变其功能，采用定向进化等技术。定向进化对于改善现有蛋白质的特性特别有用，例如增加其稳定性、改变其底物特异性或提高其催化效率。

Evaluation metric. The evaluation of protein structure prediction models typically includes Root Mean Square Deviation (RMSD), Global Distance Test (GDT) [410], Template Modeling Score (TM-Score) [410], and Local Distance Difference Test (LDDT) [226]. For more details on evaluating the predicted protein structure, readers can refer to [168]. For protein function prediction, the standard metrics in machine learning (e.g., accuracy for classification tasks and correlation coefficient for regression tasks) can be employed to quantitatively assess the performance of these Zhang and Ding, et al. Prot-LLMs. To evaluate the performance of protein sequence generation, one can use the following metrics: the model perplexity, novelty, similarity, and diversity of generated protein sequences, as well as the condition consistency if the generation is conditioned on certain inputs (e.g., protein backbone). We list them as follows.

评估指标。蛋白质结构预测模型的评估通常包括均方根偏差(RMSD)、全局距离测试(GDT)[410]、模板建模评分(TM-Score)[410]和局部距离差异测试(LDDT)[226]。有关评估预测蛋白质结构的更多详细信息，读者可以参考[168]。对于蛋白质功能预测，可以使用机器学习中的标准指标(例如，分类任务的准确性和回归任务的相关系数)来定量评估这些Zhang和Ding等人的Prot-LLMs的性能。为了评估蛋白质序列生成的性能，可以使用以下指标:模型困惑度、生成蛋白质序列的新颖性、相似性和多样性，以及如果生成是基于某些输入(例如蛋白质骨架)的条件一致性。我们将其列出如下。

* Perplexity (PPL). In information theory, perplexity is a measure of uncertainty in the value of a sample from a discrete probability distribution. The larger the perplexity, the less likely it is that a language model can generate the protein sequence. The perplexity of a heldout test set reflects the capability of model to capture the distribution of natural sequences. The formula for PPL is defined as follows:
* 困惑度(Perplexity, PPL)。在信息论中，困惑度是衡量从离散概率分布中抽取样本值的不确定性的指标。困惑度越大，语言模型生成蛋白质序列的可能性就越小。保留测试集的困惑度反映了模型捕捉自然序列分布的能力。PPL的公式定义如下:

where is the entropy of the distribution , and ranges over the protein sequence.

其中 是分布 的熵， 在蛋白质序列上变化。

* Novelty. Novelty refers to the uniqueness or novelty of a protein sequence in comparison to known proteins, which can be reflected by the highest identity between each generated sequence and natural proteins. The calculation involves aligning the amino acid sequences and counting the number of positions in which the amino
* 新颖性。新颖性指的是蛋白质序列与已知蛋白质相比的独特性或新颖性，这可以通过每个生成序列与天然蛋白质之间的最高相似性来反映。计算过程包括比对氨基酸序列并统计氨基酸位置的数量。

acids are identical, defined as

酸是相同的，定义为

where represents the number of identical positions and represents the total number of aligned positions. A protein with a low percentage identity within all natural proteins implies a high degree of novelty.

其中 表示相同位置的数量， 表示对齐位置的总数。在所有天然蛋白质中，具有低百分比同一性的蛋白质意味着高度的新颖性。

* Fréchet Protein Distance (FPD) [144]. This metric quantifies the similarity between a set of generated proteins(G)and a reference set(R). Previous research has demonstrated that protein language model embeddings inherently encapsulate information related to both structure and function, suggesting that the FPD metric provides a comprehensive measure of coverage across the distribution of structural and functional properties. The FPD is mathematically defined as follows:
* 弗雷歇蛋白质距离(Fréchet Protein Distance, FPD)[144]。该指标量化了一组生成蛋白质(G)与参考集(R)之间的相似性。先前的研究表明，蛋白质语言模型嵌入本质上封装了与结构和功能相关的信息，这表明FPD指标提供了对结构和功能属性分布覆盖的全面衡量。FPD的数学定义如下:

where, given the embedding space feature vectors for the reference and generated distributions, represents the feature-wise mean for each set of sequences, denotes the representative covariance matrix, and Tr refers to the trace linear algebra operation, defined as the sum of elements along the main diagonal of a square matrix.

其中，给定参考分布和生成分布的嵌入空间特征向量， 表示每组序列的特征均值， 表示代表性协方差矩阵，Tr指代迹线性代数运算，定义为方阵主对角线上元素的总和。

* Diversity. Sequence diversity assessment involves analyzing the variety of the generated proteins. Tools like BLAST can be used to compare these sequences against known protein databases to identify novel sequences. Metrics such as sequence similarity, percentage of unique sequences, and alignment scores provide a quantitative measure of diversity.
* 多样性。序列多样性评估涉及分析生成蛋白质的多样性。可以使用BLAST等工具将这些序列与已知蛋白质数据库进行比较，以识别新序列。序列相似性、独特序列百分比和比对得分等指标提供了多样性的定量测量。
* Foldability. This metric measures the foldability of a generated protein sequence by assessing the average per-residue confidence score, denoted as pLDDT, across the entire protein sequence. pLDDT serves as an indicator of the structural prediction model’s confidence in its predictions for individual residues. Lower pLDDT scores are frequently associated with intrinsically disordered regions (IDRs) within proteins, which lack a well-defined three-dimensional structure and thus exhibit inherent unfoldability.
* 可折叠性。该指标通过评估整个蛋白质序列中每个残基的平均置信度得分(称为pLDDT)来衡量生成蛋白质序列的可折叠性。pLDDT作为结构预测模型对其单个残基预测置信度的指标。较低的pLDDT得分通常与蛋白质内的内在无序区域(IDRs)相关，这些区域缺乏明确的三维结构，因此表现出固有的不可折叠性。
* Recovery. This metric refers to the success or accuracy in predicting the correct amino acid sequence that corresponds to a given 3D structure (i.e., Structure-based conditional generation). A high recovery rate indicates that the designed sequences are likely to fold the desired structures.
* 恢复率。该指标指的是预测与给定三维结构对应的正确氨基酸序列的成功率或准确性(即基于结构的条件生成)。高恢复率表明设计的序列很可能折叠成所需的结构。

# 5.4 Summary

# 5.4 总结

In this section, we have thoroughly examined the landscape of Sci-LLMs specifically designed for protein languages, namely Prot-LLMs. Our investigation encompasses a comprehensive analysis of their model architectures and capabilities, with a focus on how these models interpret and process protein languages. Additionally, we have curated an extensive collection of datasets, vital for training and benchmarking Prot-LLMs. These datasets are instrumental in providing a rich source of protein-related information, encompassing various aspects like sequence homology, structural configurations, and functional annotations. Finally, we present the evaluation method of Prot-LLMs, which encompasses several downstream tasks that these models are expected to perform, such as functional classification, interaction prediction, and protein sequence design. These tasks are crucial in assessing the models’ practical applicability in real-world biological and pharmaceutical scenarios.

在本节中，我们深入探讨了专为蛋白质语言设计的Sci-LLMs(科学语言模型)，即Prot-LLMs(蛋白质语言模型)。我们的研究涵盖了对其模型架构和能力的全面分析，重点关注这些模型如何解释和处理蛋白质语言。此外，我们整理了大量数据集，这些数据集对于训练和基准测试Prot-LLMs至关重要。这些数据集为提供丰富的蛋白质相关信息提供了重要来源，涵盖了序列同源性、结构配置和功能注释等多个方面。最后，我们介绍了Prot-LLMs的评估方法，包括这些模型预期执行的多个下游任务，如功能分类、相互作用预测和蛋白质序列设计。这些任务对于评估模型在现实世界生物和制药场景中的实际应用性至关重要。