# Language models of protein sequences at the scale of evolution enable accurate structure prediction

# 进化尺度上的蛋白质序列语言模型实现准确的结构预测

Zeming Lin , Halil Akin , Roshan Rao , Brian Hie , Zhongkai Zhu , Wenting Lu , Allan dos Santos Costa , Maryam Fazel-Zarandi , Tom Sercu , Sal Candido , Alexander Rives

林泽铭 、哈利勒·阿金 、罗尚·拉奥 、布莱恩·希 、朱忠凯 、卢文婷 、艾伦·多斯·桑托斯·科斯塔 、玛丽阿姆·法泽尔 - 扎兰迪 、汤姆·塞尔库 、萨尔·坎迪多 、亚历山大·里夫斯

Meta AI, FAIR Team

Meta人工智能公司，FAIR团队

New York University. Work performed as a visiting researcher at Meta AI.

纽约大学。作为Meta人工智能公司的访问研究员开展的工作。

Stanford University. Work performed as a visiting researcher at Meta AI.

斯坦福大学。作为Meta人工智能公司的访问研究员开展的工作。

Massachusetts Institute of Technology. Work performed during internship at Meta AI.

麻省理工学院。在Meta人工智能公司实习期间开展的工作。

\* Equal contribution

\* 同等贡献

† Research and engineering leadership

† 研究与工程领导

Corresponding author, arives@fb.com

通讯作者，arives@fb.com

# Abstract

# 摘要

Large language models have recently been shown to develop emergent capabilities with scale, going beyond simple pattern matching to perform higher level reasoning and generate lifelike images and text. While language models trained on protein sequences have been studied at a smaller scale, little is known about what they learn about biology as they are scaled up. In this work we train models up to 15 billion parameters, the largest language models of proteins to be evaluated to date. We find that as models are scaled they learn information enabling the prediction of the three-dimensional structure of a protein at the resolution of individual atoms. We present ESMFold for high accuracy end-to-end atomic level structure prediction directly from the individual sequence of a protein. ESMFold has similar accuracy to AlphaFold2 and RoseTTAFold for sequences with low perplexity that are well understood by the language model. ESMFold inference is an order of magnitude faster than AlphaFold2, enabling exploration of the structural space of metagenomic proteins in practical timescales.

最近研究表明，大型语言模型会随着规模的增大而发展出涌现能力，超越简单的模式匹配，进行更高级别的推理，并生成逼真的图像和文本。虽然在较小规模上对基于蛋白质序列训练的语言模型进行过研究，但对于随着模型规模扩大它们对生物学的学习情况却知之甚少。在这项工作中，我们训练了参数多达150亿的模型，这是迄今为止评估过的最大的蛋白质语言模型。我们发现，随着模型规模的增大，它们会学习到能够在单个原子分辨率水平上预测蛋白质三维结构的信息。我们推出了ESMFold，用于直接从蛋白质的单个序列进行高精度的端到端原子水平结构预测。对于语言模型能够很好理解的低困惑度序列，ESMFold的准确性与AlphaFold2和RoseTTAFold相近。ESMFold的推理速度比AlphaFold2快一个数量级，能够在实际时间尺度内探索宏基因组蛋白质的结构空间。

# Introduction

# 引言

In linguistics, the distributional hypothesis proposes that meaning can be inferred from text by the way it constrains the patterns of words (1). An analogous idea has been critical for inference from sequences in biology. Because the structure and function of a protein constrains the mutations to its sequence that are selected through evolution (2-4), it should also be possible to infer biological structure and function from sequence patterns (5-9), which would provide insight into some of the most foundational problems in biology (10). However, learning sufficient information from sequence alone to model the complexity and diversity of biological structures and functions remains a considerable challenge.

在语言学中，分布假设认为可以通过文本对单词模式的约束方式从文本中推断出其含义(1)。类似的观点在生物学中从序列进行推断时也至关重要。由于蛋白质的结构和功能会限制其序列在进化过程中所发生的突变(2 - 4)，因此也应该可以从序列模式中推断出生物学结构和功能(5 - 9)，这将有助于深入理解生物学中一些最基础的问题(10)。然而，仅从序列中学习到足够的信息来对生物结构和功能的复杂性与多样性进行建模仍然是一项相当大的挑战。

In natural language processing and artificial intelligence, general purpose language models have shown that their performance on complex tasks improves as compute, data, and model size increases. At certain scales, language models exhibit useful capabilities which emerge as a result of scaling a simple training process to large corpuses of data, e.g. few-shot language translation, commonsense reasoning, and mathematical reasoning (11-14). To this end, we study protein structure as learned by language models bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license. trained purely on protein sequence data with a simple language modeling objective (15-22). Previous work has shown that protein language models can capture some functional (23) and structural properties of proteins, including secondary structure, tertiary contacts, backbone structure, and antibody structure (15,19,24,25).

在自然语言处理和人工智能领域，通用语言模型表明，随着计算资源、数据量和模型规模的增加，它们在复杂任务上的表现会有所提升。在一定规模下，语言模型会展现出一些有用的能力，这些能力是将简单的训练过程应用于大规模数据集的结果，例如少样本语言翻译、常识推理和数学推理(11 - 14)。为此，我们研究了仅使用简单的语言建模目标对蛋白质序列数据进行训练的语言模型所学习到的蛋白质结构(15 - 22)。先前的研究表明，蛋白质语言模型能够捕捉到蛋白质的一些功能(23)和结构特性，包括二级结构、三级接触、主链结构和抗体结构(15、19、24、25)。

Here, we report that large protein language models learn sufficient information to enable accurate, atomic-level predictions of protein structure. First, we introduce ESM-2, in variants up to 15 billion parameters, the largest language model of protein sequences to date. Next, we introduce ESMFold, which uses the information and representations learned by ESM-2 to perform end-to-end 3D structure prediction using only a single sequence as input, allowing us to quantify the emergence of protein structure as the language model is scaled from millions to billions of parameters. Notably, we find that as the size of the language model increases, we also observe consistent improvements in structure prediction accuracy.

在这里，我们报告称大型蛋白质语言模型能够学习到足够的信息，从而实现对蛋白质结构的准确原子水平预测。首先，我们推出了ESM - 2，其变体的参数多达150亿，是迄今为止最大的蛋白质序列语言模型。接下来，我们推出了ESMFold，它利用ESM - 2学习到的信息和表示，仅以单个序列作为输入进行端到端的三维结构预测，使我们能够量化随着语言模型参数从数百万增加到数十亿时蛋白质结构信息的涌现情况。值得注意的是，我们发现随着语言模型规模的增大，结构预测的准确性也会持续提高。

While recent models, AlphaFold2 (26) and RoseTTAFold (27), have achieved breakthrough success in the problem of atomic-resolution structure prediction, they also rely on the use of multiple sequence alignments (MSAs) and templates of similar protein structures to achieve optimal performance. In contrast, by leveraging the internal representations of the language model, ESMFold generates structure predictions using only a single sequence as input, resulting in considerably faster structure prediction. ESMFold also produces more accurate atomic-level predictions than AlphaFold2 or RoseTTAFold when they are artificially given a single sequence as input, and obtains competitive performance to RoseTTAFold given full MSAs as input. Moreover, we find that ESMFold produces comparable predictions to state-of-the-art models for low perplexity sequences, and more generally that structure prediction accuracy correlates with language-model perplexity, indicating that when a language model better understands the sequence, it also better understands the structure.

尽管近期的模型，如AlphaFold2(26)和RoseTTAFold(27)，在原子分辨率结构预测问题上取得了突破性成功，但它们也依赖于多序列比对(MSAs)和相似蛋白质结构模板的使用，以实现最佳性能。相比之下，ESMFold通过利用语言模型的内部表征，仅使用单一序列作为输入来生成结构预测，从而显著加快了结构预测速度。当人为地仅给AlphaFold2或RoseTTAFold输入单一序列时，ESMFold产生的原子级预测比它们更准确；当输入完整的多序列比对时，ESMFold的性能与RoseTTAFold相当。此外，我们发现，对于低困惑度的序列，ESMFold的预测结果与最先进的模型相当，并且更普遍地，结构预测的准确性与语言模型的困惑度相关，这表明当语言模型更好地理解序列时，它也能更好地理解结构。

Since ESMFold’s prediction speed is an order of magnitude faster than existing atomic resolution structure predictors, ESMFold can help address the gap between the rapid growth of protein sequence databases, which increasingly contain billions of sequences (28-30), and the much slower growth of databases of protein structures and functions. Using ESMFold, we rapidly compute 1 million predicted structures representing a diverse subset of metagenomic sequence space, most of which have no annotated structure or function (29). A large fraction of ESMFold’s high-confidence predictions has low similarity to any known experimental structures, which suggests the structural novelty of many metagenomic proteins. Notably, many high-confidence structures also have low sequence similarity to any entry in UniRef90, indicating generalization of the model’s predictions beyond its training dataset and enabling structure-based insight into protein function that would be difficult to obtain from sequence information alone. By leveraging the unprecedented view into the language of protein sequences provided by ESM-2, ESMFold promises to augment our understanding of vast databases of poorly understood protein sequences.

由于ESMFold的预测速度比现有的原子分辨率结构预测器快一个数量级，它有助于弥合蛋白质序列数据库快速增长(其序列数量日益达到数十亿条，参考文献28 - 30)与蛋白质结构和功能数据库增长缓慢之间的差距。利用ESMFold，我们快速计算了100万个预测结构，这些结构代表了宏基因组序列空间的一个多样化子集，其中大多数没有注释的结构或功能(29)。ESMFold的高置信度预测结果中有很大一部分与任何已知的实验结构相似度较低，这表明许多宏基因组蛋白质具有结构新颖性。值得注意的是，许多高置信度结构与UniRef90中的任何条目序列相似度也较低，这表明该模型的预测能够推广到其训练数据集之外，并有助于从结构角度深入了解蛋白质功能，而仅通过序列信息则难以实现这一点。通过利用ESMFold - 2对蛋白质序列语言提供的前所未有的见解，ESMFold有望加深我们对大量难以理解的蛋白质序列数据库的理解。

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license. bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。本预印本的版权所有者(未经同行评审认证)是作者/资助者，他已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议提供。生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。本预印本的版权所有者(未经同行评审认证)是作者/资助者，他已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议提供。

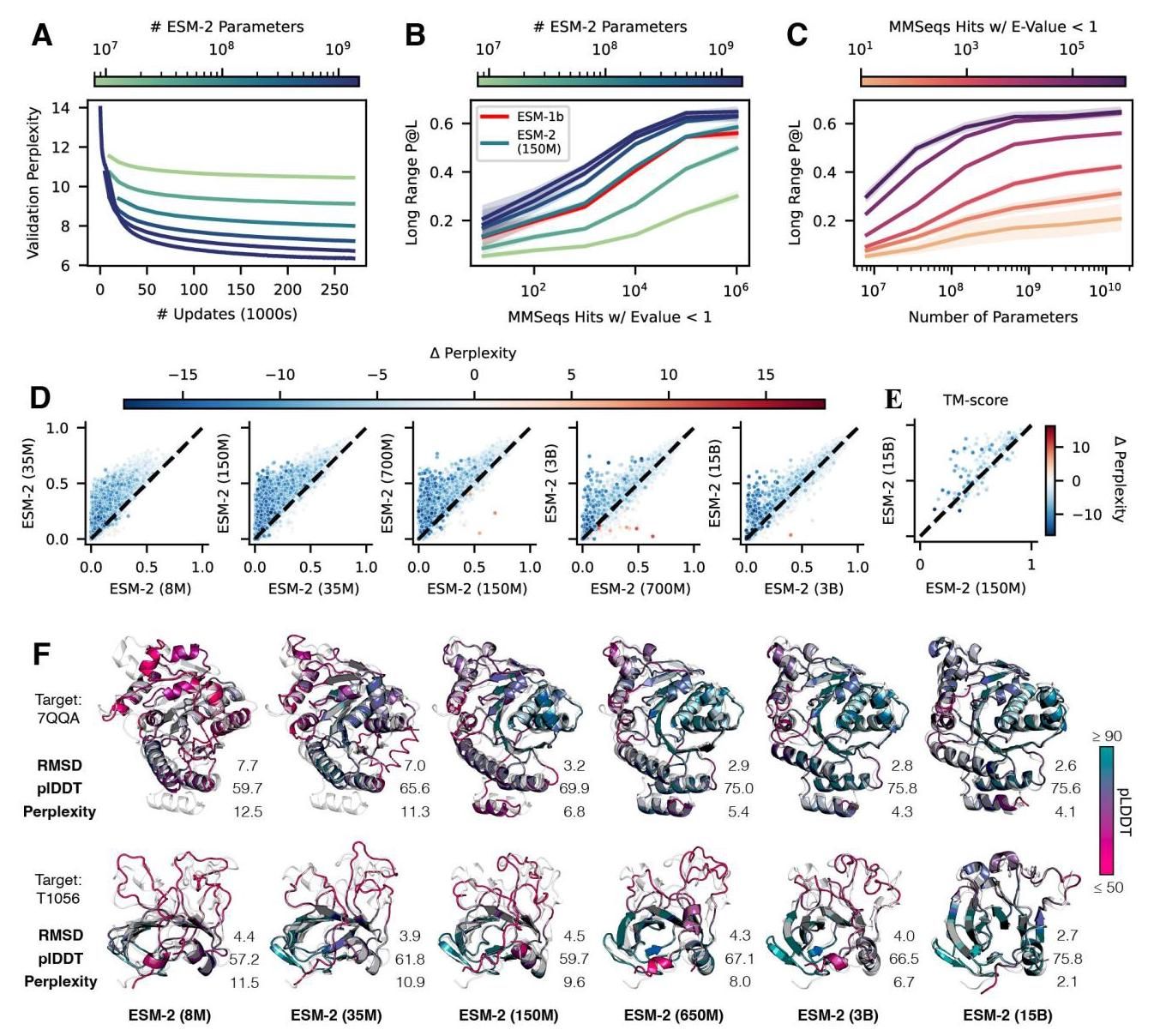


Figure 1: Emergence of structure when scaling language models to 15 Billion parameters. (A) Training curves for ESM-2 models at all scales, through 270,000 updates. (B, C, D) Unsupervised contact prediction performance (long range precision @ L) for different scales of the ESM-2 model. (B) Performance is binned by the number of MMseqs hits when searching the training set. Larger models perform better at all levels, and the parameter ESM-2 model performs comparably with the parameter ESM-1b model. (C) Trajectory of improvement as model scale increases for sequences with different numbers of MMseqs hits is shown. The largest improvement is seen for sequences with MMseqs hits. (D) Left-to-right shows models from 8M to 15B parameters, consecutively comparing the smaller model (x-axis) against the next larger model (y-axis) in terms of unsupervised contact precision. Points correspond to PDB proteins and are colored by the change in pseudo-perplexity for the sequence between the smaller and larger model. Sequences with large changes in contact prediction performance also exhibit large changes in language model understanding measured by pseudo-perplexity. (E) TM-score on combined CASP14 and CAMEO test sets. Models are structure module trained on 150M parameter (x-axis) and 15B parameter (y-axis) ESM-2. Points are colored by the change in pseudo-perplexity between the models. (F) Left-to-right structure predictions on CAMEO structure 7QQA and CASP target 1056 at all ESM-2 model scales, colored by pLDDT (pink = low, teal = high). For 7QQA, prediction accuracy improves suddenly at the parameter threshold, and slowly thereafter. For T1056, prediction accuracy improves suddenly at the 15B parameter threshold.

图1:将语言模型扩展到150亿参数时结构的出现。(A) 所有规模的ESM - 2模型在270,000次更新过程中的训练曲线。(B、C、D) ESM - 2模型不同规模下的无监督接触预测性能(长程精度@L)。(B) 性能按在训练集中搜索时MMseqs匹配数进行分组。更大的模型在所有水平上表现更好， 参数的ESM - 2模型的表现与 参数的ESM - 1b模型相当。(C) 展示了不同MMseqs匹配数的序列随着模型规模增加的性能提升轨迹。 个MMseqs匹配的序列提升最大。(D) 从左到右依次为800万到150亿参数的模型，连续比较较小模型(x轴)与下一个较大模型(y轴)的无监督接触精度。点对应PDB蛋白质，并根据较小和较大模型之间序列的伪困惑度变化进行着色。接触预测性能变化大的序列在通过伪困惑度衡量的语言模型理解方面也表现出较大变化。(E) 在CASP14和CAMEO组合测试集上的TM分数。模型是在1.5亿参数(x轴)和150亿参数(y轴)的ESM - 2上训练的结构模块。点根据模型之间的伪困惑度变化进行着色。(F) 在所有ESM - 2模型规模下对CAMEO结构7QQA和CASP目标1056的从左到右的结构预测，按pLDDT着色(粉色 = 低，蓝绿色 = 高)。对于7QQA，预测精度在 参数阈值处突然提高，此后缓慢提高。对于T1056，预测精度在150亿参数阈值处突然提高。

# Training and evaluating 15B parameter protein language models.

# 训练和评估150亿参数的蛋白质语言模型。

The ESM-2 language models are the most performant language models of proteins developed to date. Relative to our previous generation model ESM-1b we improve model architecture, training parameters, and increase computational resources and data. Addition of relative positional embeddings enables generalization to arbitrary length sequences. These modifications lead to a significantly better model. We observe that the ESM-2 model with parameters performs better than the ESM-1b model with parameters. On structure prediction benchmarks it also outperforms other recent protein language models (Table 1, table S3). This performance increase is consistent with scaling laws established in the large language modeling field (31). For context, the 15B parameter ESM-2 model is only one order of magnitude smaller than the largest state-of-the-art language models of text that have been trained such as Chinchilla (70 billion parameters), GPT3 and OPT-175B (both 175 billion parameters), and PALM (540 billion parameters)(11,14,32,33).

ESM - 2语言模型是迄今为止开发的性能最佳的蛋白质语言模型。相对于我们上一代模型ESM - 1b，我们改进了模型架构、训练参数，并增加了计算资源和数据。相对位置嵌入的加入使模型能够推广到任意长度的序列。这些改进导致了一个显著更优的模型。我们观察到， 参数的ESM - 2模型的表现优于 参数的ESM - 1b模型。在结构预测基准测试中，它也优于其他近期的蛋白质语言模型(表1，表S3)。这种性能提升与大语言模型领域确立的缩放定律一致(31)。作为参考，150亿参数的ESM - 2模型仅比已训练的最大的最先进文本语言模型小一个数量级，如Chinchilla(700亿参数)、GPT3和OPT - 175B(均为1750亿参数)以及PALM(5400亿参数)(11,14,32,33)。

ESM-2 is trained on protein sequences from the UniRef database (28). Given an input protein (represented as a character sequence of amino acids), 15% of amino acids are masked and ESM-2 is tasked with predicting these missing positions (34). Although this training objective only directly involves predicting missing amino acids, achieving a high degree of success requires the model to learn complex internal representations of its input. In natural language processing, these representations contain information about parts of speech, dependency parsing, semantic relatedness and textual entailment (34-36). In biology, these representations learn secondary structure prediction, binding site prediction, and contact prediction(15,37,38).

ESM - 2在UniRef数据库(28)中的蛋白质序列上进行训练。给定一个输入蛋白质(表示为氨基酸字符序列)，15%的氨基酸被掩码，ESM - 2的任务是预测这些缺失的位置(34)。尽管这个训练目标仅直接涉及预测缺失的氨基酸，但要取得高度成功需要模型学习其输入的复杂内部表示。在自然语言处理中，这些表示包含关于词性、依存句法分析、语义相关性和文本蕴含的信息(34 - 36)。在生物学中，这些表示学习二级结构预测、结合位点预测和接触预测(15,37,38)。

As we increase the scale of ESM-2, we observe large improvements in the fidelity of language modeling. Language model performance is evaluated using perplexity, which measures the model’s performance on the task of predicting amino acids from their context in a sequence. Perplexity ranges from 1 for a perfect model to 20 for a model that makes predictions at random. Intuitively, perplexity describes the number of amino acids the model is uncertain between when it makes a prediction. We hold out UniRef50 clusters from the ESM-2 training set for evaluation. Fig. 1A shows perplexity as a function of the number of updates for the ESM-2 family models. After 270k training steps the parameter model has a perplexity of 10.45, and the 15B model reaches a perplexity of 6.37.

随着我们扩大ESM - 2的规模，我们观察到语言建模的保真度有了显著提高。语言模型的性能通过困惑度(perplexity)来评估，困惑度衡量的是模型在根据序列上下文预测氨基酸任务中的表现。困惑度的范围从完美模型的1到随机预测模型的20。直观地说，困惑度描述了模型在进行预测时不确定的氨基酸数量。我们从ESM - 2训练集中留出 个UniRef50聚类用于评估。图1A展示了ESM - 2系列模型的困惑度随更新次数的变化情况。经过270k次训练步骤后， 参数模型的困惑度为10.45，而15B模型的困惑度达到了6.37。

Next, we look at the emergence of protein structure as the model scales. ESM-2 is a transformer-based language model, and uses an attention mechanism to learn interaction patterns between pairs of amino acids in the input sequence. Standard methods in computational biology use correlations between mutations in amino acids at different sites to learn protein contact maps, since amino acids that are in contact in the three dimensional structure are not free to evolve independently(5 - 8,39). More recent work has shown the pairwise interaction patterns learned by protein language models also correspond to protein contact maps because the correlations between different sites is useful for predicting missing amino acids(37,38). Importantly, this correspondence emerges solely from the language modeling objective, without direct supervision on protein structures.

接下来，我们研究随着模型规模的扩大，蛋白质结构是如何出现的。ESM - 2是一个基于Transformer的语言模型，它使用注意力机制来学习输入序列中氨基酸对之间的相互作用模式。计算生物学中的标准方法利用不同位点氨基酸突变之间的相关性来学习蛋白质接触图，因为在三维结构中相互接触的氨基酸不能独立进化(5 - 8,39)。最近的研究表明，蛋白质语言模型学习到的成对相互作用模式也与蛋白质接触图相对应，因为不同位点之间的相关性有助于预测缺失的氨基酸(37,38)。重要的是，这种对应关系完全是从语言建模目标中产生的，无需对蛋白质结构进行直接监督。

Throughout Fig. 1 we compare different ESM-2 parameter scales at 270,000 training updates, and find that scaling leads to large improvements in the unsupervised learning of structure (Fig. 1B). We also look at the accuracy of the predicted contacts as a function of the number of evolutionarily related sequences in the language model’s training set. Proteins with more related sequences in the training set have steeper learning trajectories with respect to model scale (Fig. 1C). In other words, improvement on sequences with high evolutionary depth saturates at lower model scales, and improvement on sequences with low evolutionary depth continues as models increase in size.

在图1中，我们比较了270,000次训练更新时不同ESM - 2参数规模的模型，发现扩大规模能显著提高结构的无监督学习效果(图1B)。我们还研究了预测接触的准确性与语言模型训练集中进化相关序列数量的关系。训练集中相关序列较多的蛋白质在模型规模方面具有更陡峭的学习轨迹(图1C)。换句话说，进化深度高的序列在较低的模型规模下改进效果就会饱和，而进化深度低的序列随着模型规模的增大仍会持续改进。

For individual proteins, we often observe non-linear improvements as a function of scale. Fig. 1D plots the distribution of long range contact precision as we scale from one model to the next. At each level of scale we see the overall distribution shift toward better performance. Also at each transition, there is a subset of proteins that undergo significant improvement. In Fig. 1D these are concentrated in the upper left of each plot, far from the diagonal. There is a link between contact accuracy and perplexity, with proteins undergoing large changes in contact accuracy also undergoing large changes in perplexity. This link indicates that the language modeling objective is directly correlated with the materialization of the structure in the attention maps.

对于单个蛋白质，我们经常观察到其性能随模型规模呈非线性改进。图1D绘制了从一个模型规模过渡到下一个模型规模时长程接触精度的分布情况。在每个规模级别上，我们都看到整体分布向更好的性能方向移动。而且在每次过渡时，都有一部分蛋白质会有显著的改进。在图1D中，这些蛋白质集中在每个图的左上角，远离对角线。接触准确性和困惑度之间存在联系，接触准确性变化较大的蛋白质其困惑度也会有较大变化。这种联系表明，语言建模目标与注意力图中结构的具体化直接相关。

Given that the folded structure, in the form of the contact pattern, develops in the attention patterns of the model with scale, we investigate whether information also develops that enables prediction of the structure at the full atom resolution. We develop a method to project the atom level structure from the internal representations of ESM-2 using the equivariant structure module of Alphafold. Using the same network architecture of the structure module alongside different versions of ESM-2 during training, we can quantitatively compare the language models on how much information their representations contain that is useful for making a full atom prediction. To train the full atom projection we use supervision from a dataset of experimentally determined protein structure from PDB (40). We evaluate on temporally held out CAMEO (41) and CASP14 (42) test sets consisting respectively of 194 and 51 proteins.

鉴于折叠结构以接触模式的形式随着模型规模在注意力模式中发展，我们研究是否也会产生能够实现全原子分辨率结构预测的信息。我们开发了一种方法，利用AlphaFold的等变结构模块从ESM - 2的内部表示中投影出原子水平的结构。在训练过程中，使用相同的结构模块网络架构和不同版本的ESM - 2，我们可以定量比较这些语言模型的表示中包含多少对全原子预测有用的信息。为了训练全原子投影，我们使用了来自PDB的实验确定的蛋白质结构数据集的监督信息(40)。我们在时间上留出的CAMEO(41)和CASP14(42)测试集上进行评估，这两个测试集分别包含194个和51个蛋白质。

Atomic resolution predictions can be extracted from the representations of the ESM-2 language models and improve with scale. The 15 billion parameter ESM-2 model achieves a TM-score (43) of 71.3 on the CAMEO test set and 53.9 on the CASP14 test set, 6.4 points higher than the 150 million parameter ESM-2 model on both (Fig. 1E). Similarly to the results with contact maps, at each increase in scale a subset of proteins see large changes in accuracy. For example, the protein 7QQA sees an improvement in RMSD from 7.0 to 3.2 when scale is increased from to parameters, and the CASP target T1056 sees an improvement in RMSD from 4.0 to 2.6 when scale is increased from 3B to 15B parameters (Fig. 1F). Before and after these jumps, changes in RMSD are much smaller. Across all models (table S3) there is a correlation of -0.99 between validation perplexity and CASP14 TM-score, and -1.00 between validation perplexity and CAMEO TM-score indicating a strong link between language model understanding of a sequence measured by perplexity and the atomic resolution structure prediction. Additionally there are strong correlations between the low resolution picture of the structure that can be extracted from the attention maps and the atomic resolution prediction ( 0.96 between long range contact precision and CASP14 TM-score, and 0.99 between long range contact precision and CAMEO TM-score). bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

可以从ESM - 2语言模型的表征中提取原子分辨率预测结果，并且预测精度会随着模型规模的增大而提高。拥有150亿参数的ESM - 2模型在CAMEO测试集上的TM分数(43)达到71.3，在CASP14测试集上达到53.9，比拥有1.5亿参数的ESM - 2模型在两个测试集上的分数均高出6.4分(图1E)。与接触图的结果类似，每次模型规模增大时，一部分蛋白质的预测精度会发生显著变化。例如，当模型参数从 增加到 时，蛋白质7QQA的均方根偏差(RMSD)从7.0改善到3.2；当模型参数从30亿增加到150亿时，CASP目标T1056的均方根偏差从4.0改善到2.6(图1F)。在这些显著改善前后，均方根偏差的变化要小得多。在所有模型中(表S3)，验证困惑度与CASP14的TM分数之间的相关性为 - 0.99，验证困惑度与CAMEO的TM分数之间的相关性为 - 1.00，这表明通过困惑度衡量的语言模型对序列的理解与原子分辨率结构预测之间存在很强的联系。此外，从注意力图中提取的结构低分辨率信息与原子分辨率预测之间也存在很强的相关性(长程接触精度与CASP14的TM分数之间的相关性为0.96，长程接触精度与CAMEO的TM分数之间的相关性为0.99)。生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议(CC - BY - NC - ND 4.0)提供。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | #Params | Validation Perplexity | LR P@L | CASP14 | CAMEO |
| ESM-2 | 8M | 10.33 | 0.17 | 0.37 | 0.48 |
| 35M | 8.95 | 0.30 | 0.41 | 0.56 |
| 150M | 7.75 | 0.44 | 0.49 | 0.65 |
| 650M | 6.95 | 0.52 | 0.51 | 0.70 |
| 3B | 6.49 | 0.54 | 0.52 | 0.72 |
| 15B | 6.37 | 0.54 | 0.55 | 0.72 |
| ESM-1b1 | 650M | - | 0.41 | 0.42 | 0.64 |
| Prot-T5-XL-UR50 (19) | 3B | - | 0.48 | 0.50 | 0.69 |
| Prot-T5-XL-BFD (19) | 3B | - | 0.36 | 0.46 | 0.63 |
| CARP (44) | 640M | - | - | 0.42 | 0.59 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | #参数 | 验证困惑度 | LR P@L | CASP14(第十四届蛋白质结构预测技术评估) | CAMEO(持续蛋白质结构模型评估) |
| ESM - 2(进化尺度模型2) | 8M | 10.33 | 0.17 | 0.37 | 0.48 |
| 35M | 8.95 | 0.30 | 0.41 | 0.56 |
| 150M | 7.75 | 0.44 | 0.49 | 0.65 |
| 650M | 6.95 | 0.52 | 0.51 | 0.70 |
| 3B | 6.49 | 0.54 | 0.52 | 0.72 |
| 15B | 6.37 | 0.54 | 0.55 | 0.72 |
| ESM - 1b1(进化尺度模型1b1) | 650M | - | 0.41 | 0.42 | 0.64 |
| Prot - T5 - XL - UR50 (19)(蛋白质T5超大型通用表示50 (19)) | 3B | - | 0.48 | 0.50 | 0.69 |
| Prot - T5 - XL - BFD (19)(蛋白质T5超大型大生物序列数据库 (19)) | 3B | - | 0.36 | 0.46 | 0.63 |
| CARP (44)(连续自注意力表示预测 (44)) | 640M | - | - | 0.42 | 0.59 |

Table 1: Evaluation Metrics for converged ESM-2 models compared with baselines.

表1:融合的ESM - 2模型与基线模型的评估指标。

Comparisons of the final structure predictions using models trained out to updates (except the parameter model which is trained to updates). All numbers are reported with only a structure module trained on top of various language models. Despite the shorter training time, the 15B parameter ESM-2 model has lowest validation perplexity and highest TM-score on CASP14. The 150M parameter ESM-2 model outperforms the parameter ESM-1b model on structure-based tasks, while the parameter ESM-2 model is comparable with the 3B parameter Prot-T5-XL-UniRef50 model, suggesting ESM-2 models are far more parameter efficient. bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

使用训练至 次更新的模型(除了训练至 次更新的 参数模型)进行最终结构预测的比较。所有数据均为仅在各种语言模型之上训练一个结构模块的结果。尽管训练时间较短，但150亿参数的ESM - 2模型在CASP14上具有最低的验证困惑度和最高的TM分数。1.5亿参数的ESM - 2模型在基于结构的任务上优于 参数的ESM - 1b模型，而 参数的ESM - 2模型与30亿参数的Prot - T5 - XL - UniRef50模型相当，这表明ESM - 2模型的参数效率要高得多。生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议提供。

[[1]](#footnote-31)

# End-to-end single-sequence structure prediction with ESMFold

# 使用ESMFold进行端到端单序列结构预测

Predicting protein structure from its amino acid sequence is a long-standing grand challenge in the natural sciences (10). AlphaFold2 (26), arguably the most successful evolutionary-based approach to the problem, presented a breakthrough achievement by training an end-to-end neural network on inputs of sequence, aligned sequences of evolutionary homologs, and optional structural templates. These advances built on earlier work learning using deep learning on sets of aligned sequences to prediction structure (45, 46). Here, we use the information in the large protein language model to train an atomic-resolution structure prediction model that requires only a single sequence input, which we refer to as ESMFold.

从蛋白质的氨基酸序列预测其结构是自然科学领域一个长期存在的重大挑战(10)。AlphaFold2(26)可以说是解决该问题最成功的基于进化的方法，它通过在序列输入、进化同源物的比对序列以及可选的结构模板上训练一个端到端的神经网络，取得了突破性的成果。这些进展建立在早期利用深度学习对一组比对序列进行学习以预测结构的工作基础之上(45, 46)。在这里，我们利用大型蛋白质语言模型中的信息来训练一个仅需单序列输入的原子分辨率结构预测模型，我们将其称为ESMFold。

A key difference between ESMFold and AlphaFold2 is the use of language model representations to remove the need for explicit homologous sequences (in the form of an MSA) as input. Language model representations are provided as input to ESMFold’s folding trunk (Fig. 2A), which simplifies the Evoformer in AlphaFold2 by replacing the computationally expensive network modules that process the MSA with a transformer module which processes a sequence (47). This simplification means that ESMFold is substantially faster than the MSA-based model. The output of the folding trunk is in turn processed by a structure module, which outputs the final atomic-level structure and predicted confidences (Fig. 2A). Additional architectural details can be found in Methods. We train ESMFold on a diverse subset of PDB chains, further augmented with a dataset of 12M structures predicted by AlphaFold2 (48).

ESMFold与AlphaFold2的一个关键区别在于使用语言模型表示来消除对明确的同源序列(以多序列比对(MSA)的形式)作为输入的需求。语言模型表示作为输入提供给ESMFold的折叠主干(图2A)，它通过用一个处理序列的变压器模块取代AlphaFold2中处理MSA的计算成本高昂的网络模块，简化了AlphaFold2中的进化模块(Evoformer)(47)。这种简化意味着ESMFold比基于MSA的模型快得多。折叠主干的输出依次由一个结构模块处理，该模块输出最终的原子级结构和预测的置信度(图2A)。更多的架构细节可在方法部分找到。我们在蛋白质数据库(PDB)链的一个多样化子集中训练ESMFold，并进一步用AlphaFold2预测的1200万个结构的数据集进行增强(48)。

We compare ESMFold to AlphaFold2 and RoseTTAFold on held out CAMEO (April 2022 to June 2022) and CASP14 test sets consisting of structures released after our training data cutoff date (May 2020). As an ablation, we also remove MSA and template information from AlphaFold2 and RoseTTAFold and compare against "single-sequence" versions of these methods. ESMFold achieves an average TM-score of 82.8 on CAMEO and 67.8 on CASP14, significantly higher than single-sequence versions of AlphaFold2 and RoseTTAFold (Fig. 2, B and C). With the full pipeline, including MSAs and templates, AlphaFold2 achieves 88.3 and 84.7 on CAMEO and CASP14 respectively. ESMFold achieves competitive accuracy with RoseTTAfold on CAMEO, which averages a TM-score of 82.0.

我们将ESMFold与AlphaFold2和RoseTTAFold在保留的CAMEO(2022年4月至2022年6月)和CASP14测试集上进行比较，这些测试集由我们的训练数据截止日期(2020年5月)之后发布的结构组成。作为消融实验，我们还从AlphaFold2和RoseTTAFold中去除MSA和模板信息，并与这些方法的“单序列”版本进行比较。ESMFold在CAMEO上的平均TM分数达到82.8，在CASP14上达到67.8，显著高于AlphaFold2和RoseTTAFold的单序列版本(图2B和图2C)。使用包括MSA和模板的完整流程，AlphaFold2在CAMEO和CASP14上分别达到88.3和84.7。ESMFold在CAMEO上与RoseTTAfold达到了相当的准确性，RoseTTAfold的平均TM分数为82.0。

Because the language model is a critical component of ESMFold, we test how well differences in the language model correspond to changes in structure prediction performance. In particular, the performance of ESMFold on both test sets is well correlated with the perplexity of the language model. On the CAMEO test set, language model perplexity has a Pearson correlation of -0.55 with the TM-score between the predicted and experimental structures; on CASP14, the correlation is -0.67 (Fig. 2, B and C). The relationship between perplexity and structure prediction suggests that improving the language model is key to improving single-sequence structure prediction accuracy, consistent with observations from the scaling analysis (Fig. 1, D and E). Additionally, this makes it possible to predict the performance of ESMFold from how well the language model understands the input sequence as quantified by perplexity.

由于语言模型是ESMFold的关键组成部分，我们测试了语言模型的差异与结构预测性能变化之间的对应程度。特别是，ESMFold在两个测试集上的性能与语言模型的困惑度(perplexity)密切相关。在CAMEO测试集上，语言模型的困惑度与预测结构和实验结构之间的TM分数的皮尔逊相关系数为 -0.55；在CASP14上，相关系数为 -0.67(图2，B和C)。困惑度与结构预测之间的关系表明，改进语言模型是提高单序列结构预测准确性的关键，这与缩放分析的观察结果一致(图1，D和E)。此外，这使得我们能够根据语言模型对输入序列的理解程度(通过困惑度量化)来预测ESMFold的性能。

We further conduct ablation studies to understand how different model components (Fig. 2A) affect the performance of ESMFold. With a much smaller folding trunk of 8 blocks, performance degrades to 0.74 IDDT (baseline). Without the language model, the single sequence performance on the CAMEO test set degrades substantially, to 0.58 IDDT. When removing the folding trunk entirely (i.e. only using the language model and the structure module), performance degrades to 0.66 IDDT. Any one ablation of (a) only 1 block of a structure module, (b) turning off recycling, (c) not using AlphaFold2 predicted structures as distillation targets, or (d) not using triangular updates all have similar performance degradation (change in IDDT of -0.01 to -0.04). This leads us to conclude that the language model makes the largest contribution to atomic-resolution structure prediction performance on unseen proteins.

我们进一步进行消融研究，以了解不同的模型组件(图2A)如何影响ESMFold的性能。当折叠主干(folding trunk)只有8个模块时，性能下降到0.74 IDDT(基线)。如果没有语言模型，CAMEO测试集上的单序列性能会大幅下降，降至0.58 IDDT。当完全移除折叠主干(即仅使用语言模型和结构模块)时，性能下降到0.66 IDDT。以下任何一种消融操作:(a)仅移除结构模块的1个模块，(b)关闭循环机制，(c)不使用AlphaFold2预测的结构作为蒸馏目标，或(d)不使用三角更新，都会导致类似的性能下降(IDDT变化在 -0.01至 -0.04之间)。这使我们得出结论，语言模型对未见过的蛋白质的原子分辨率结构预测性能贡献最大。

ESMFold matches AlphaFold2 performance on a majority of proteins (Fig. 2C). We find that this is true even on large proteins - T1076 is an example with 0.98 TM-score and 540 residues despite being trained on a maximum crop size of 384 (Fig. 2D). Parts of structure with low accuracy do not differ significantly between ESMFold and AlphaFold, suggesting that language models are learning information similar to what is provided by MSAs. We also observe that ESMFold is able to make good predictions for components of homo- and heterodimeric protein-protein complexes (Fig. 2D).

ESMFold在大多数蛋白质上的性能与AlphaFold2相当(图2C)。我们发现，即使对于大蛋白也是如此——T1076就是一个例子，尽管训练时的最大裁剪尺寸为384，但它的TM分数为0.98，包含540个残基(图2D)。ESMFold和AlphaFold在低精度结构部分没有显著差异，这表明语言模型学习到的信息与多序列比对(MSAs)提供的信息相似。我们还观察到，ESMFold能够对同源和异源二聚体蛋白质 - 蛋白质复合物的组件做出良好的预测(图2D)。

# Language models enable more efficient predictions of protein structure

# 语言模型使蛋白质结构预测更高效

A notable advantage of ESMFold is its computational efficiency. The use of MSAs and templates in AlphaFold2 and RoseTTAFold creates two bottlenecks which ESMFold avoids. First, potentially expensive CPU-based search is required to retrieve and align the MSAs and templates. This process can average up to 30 minutes per input sequence with the standard AlphaFold2 pipeline , although runtime can be reduced to the order of seconds with newer approximate search methods like MMseqs (49). The main speed benefits come from improvements in model architecture. Instead of a two-dimensional sequence embedding state, AlphaFold2 and RoseTTAFold operate on three-dimensional internal states corresponding to the MSA using axial attention, which is expensive even when using a GPU.

ESMFold的一个显著优势是其计算效率。AlphaFold2和RoseTTAFold中使用多序列比对(MSAs)和模板产生了两个瓶颈，而ESMFold避免了这些问题。首先，检索和比对MSAs和模板需要进行可能代价高昂的基于CPU的搜索。使用标准的AlphaFold2流程，每个输入序列的这个过程平均可能需要长达30分钟 ，不过使用像MMseqs(49)这样的新型近似搜索方法，运行时间可以缩短到秒级。主要的速度提升来自模型架构的改进。AlphaFold2和RoseTTAFold不是使用二维序列嵌入状态，而是使用轴向注意力对与MSA对应的三维内部状态进行操作，即使使用GPU，这种操作的代价也很高。

By contrast, ESMFold is a fully end-to-end sequence to structure predictor and can be run entirely on GPU without access to any external databases. On a single NVIDIA V100 GPU, ESMFold makes a prediction on a protein with 384 residues in 14.2 seconds, 6X faster than a single AlphaFold2 model. On shorter sequences we see a improvement (fig. S1). Note that this excludes the CPU time for MSA and template search, as well as the from the default ensemble of models. ESMFold can be run reasonably quickly on CPU, and an Apple M1 Macbook Pro makes the same prediction in just over 5 minutes (fig. S1).

相比之下，ESMFold是一个完全端到端的序列到结构预测器，可以完全在GPU上运行，无需访问任何外部数据库。在单个NVIDIA V100 GPU上，ESMFold对一个包含384个残基的蛋白质进行预测只需14.2秒，比单个AlphaFold2模型快6倍。在较短的序列上，我们看到了 的提升(图S1)。请注意，这不包括MSA和模板搜索的CPU时间，以及默认模型集成中的 。ESMFold在CPU上也能相当快速地运行，一台苹果M1 MacBook Pro只需5分多钟就能完成相同的预测(图S1)。

[[2]](#footnote-36)

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv预印本 doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。此预印本的版权所有者(未经同行评审认证)是作者/资助者，他已授予bioRxiv永久展示该预印本的许可。该预印本根据CC - BY - NC - ND 4.0国际许可协议提供。

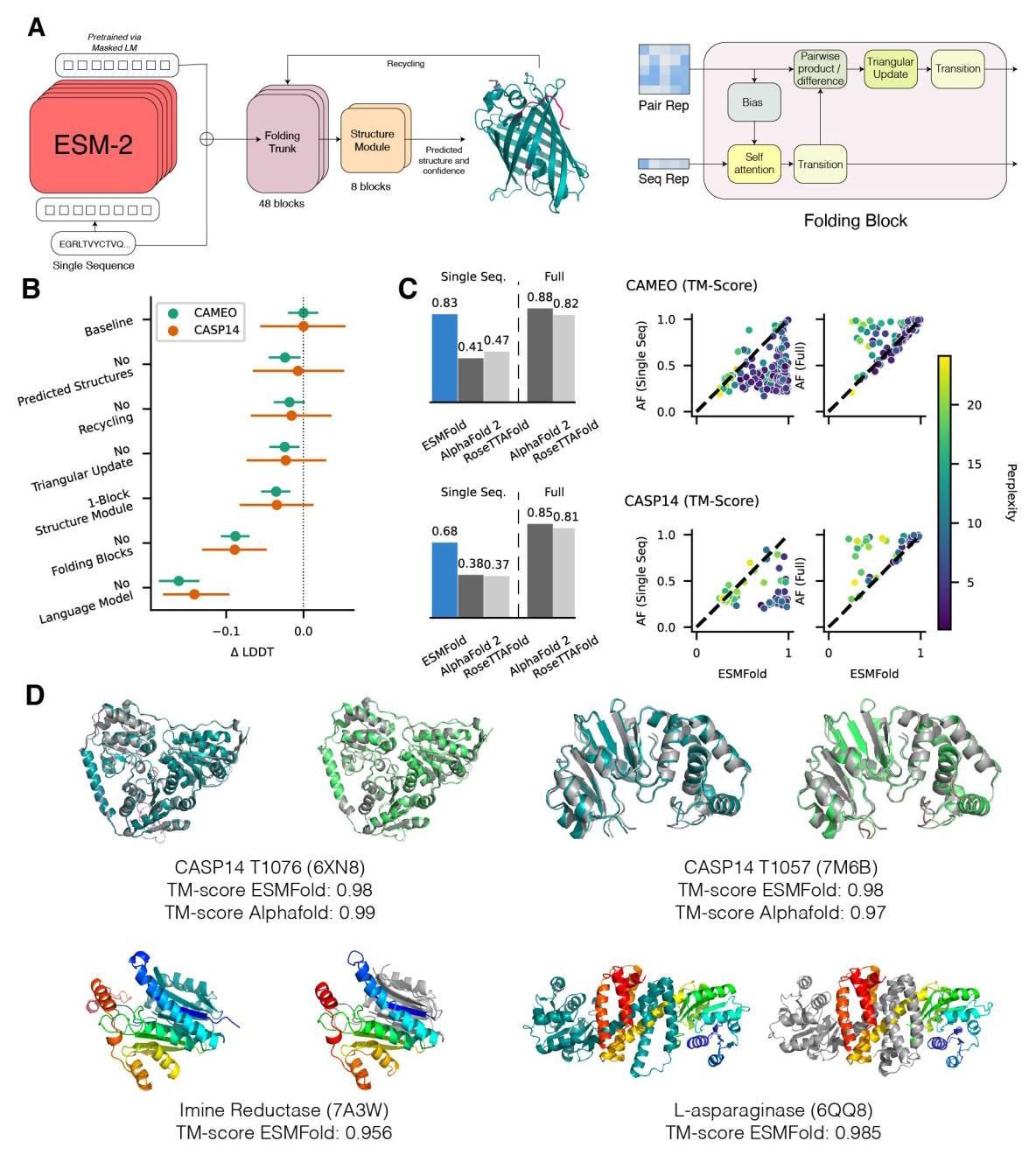


Figure 2: ESMFold enables accurate structure prediction from a single sequence. (A) ESMFold model architecture. Arrows show the information flow in the network from the language model to the folding trunk to the structure module which outputs 3D coordinates and confidences. The folding trunk is a simplified single-sequence version of the EvoFormer described in AlphaFold2. (B) Effect of various ablations on ESMFold test-time performance. Language models are by far the biggest contributor. (C) ESMFold outperforms both RoseTTAFold and AlphaFold2 when given a single sequence as input, and is competitive with RoseTTAFold even when given full MSAs on CAMEO. Scatter-plots show ESMFold (x-axis) against AlphaFold2 (y-axis) performance, colored by perplexity. Proteins with low perplexity under our model score similarly to AlphaFold2. (D) Top shows test-set predictions of ESMFold in teal, ground truth in gray, and AlphaFold2 predictions in green. Pink shows low predicted 1DDT for both ESMFold and AlphaFold2. Bottom shows complex predictions; chain B is colored teal for ESMFold (left) and gray for ground truth (right); chain A is colored rainbow from blue (N-terminal) to red (C-Terminal).

图2:ESMFold能够从单一序列进行准确的结构预测。(A) ESMFold模型架构。箭头展示了网络中的信息流，从语言模型到折叠主干，再到输出三维坐标和置信度的结构模块。折叠主干是AlphaFold2中描述的EvoFormer的简化单序列版本。(B) 各种消融操作对ESMFold测试时性能的影响。到目前为止，语言模型的贡献最大。(C) 当以单一序列作为输入时，ESMFold的性能优于RoseTTAFold和AlphaFold2；在CAMEO数据集上，即使提供完整的多序列比对(MSA)，ESMFold也能与RoseTTAFold相媲美。散点图展示了ESMFold(x轴)与AlphaFold2(y轴)的性能对比，颜色由困惑度表示。在我们的模型下困惑度较低的蛋白质得分与AlphaFold2相似。(D) 上图展示了ESMFold的测试集预测结果(青色)、真实结构(灰色)和AlphaFold2的预测结果(绿色)。粉色表示ESMFold和AlphaFold2预测的1DDT值都较低。下图展示了复合物的预测结果；链B在ESMFold预测中为青色(左)，在真实结构中为灰色(右)；链A从蓝色(N端)到红色(C端)呈彩虹色。

# Exploring metagenomic structural space

# 探索宏基因组结构空间

The order of magnitude improvement in speed is a unique advantage of ESMFold over AlphaFold2 (fig. S1), enabling us to construct large sets of predicted structures in much shorter timescales than existing methods. This is especially important considering the scale of available sequence data. For example, the initial version of the AlphaFold2 Protein Structure Database (50) was released with predicted structures, and as of July 2022 contains predictions, which is orders of magnitude smaller than many databases of protein sequences.

速度上的数量级提升是ESMFold相对于AlphaFold2的独特优势(图S1)，这使我们能够在比现有方法短得多的时间尺度内构建大量预测结构。考虑到可用序列数据的规模，这一点尤为重要。例如，AlphaFold2蛋白质结构数据库的初始版本(50)发布时包含 个预测结构，截至2022年7月，该数据库包含 个预测结构，这比许多蛋白质序列数据库小了几个数量级。

Because rapid structure prediction may enable new insight into large collections of poorly studied or annotated proteins, we use ESMFold to predict the structures of 1 million randomly-sampled, non-redundant metagenomic protein sequences from the MGnify database (29) (Methods; Fig. 3 and fig. S3). Of these 1 million sequences, ESMFold assigns good confidence (mean pLDDT > 0.7) to 2̃9% of structures and high confidence (mean pLDDT ) to . (Fig. 3A). We are able to compute these predictions in less than a day leveraging 1̃600 GPU hours of compute, indicating that ESMFold can produce structure predictions for a much larger set of metagenomic sequences within a computationally tractable timeline.

由于快速的结构预测可能为深入研究大量研究不足或注释不充分的蛋白质提供新的见解，我们使用ESMFold对从MGnify数据库(29)中随机抽取的100万个非冗余宏基因组蛋白质序列的结构进行预测(方法；图3和图S3)。在这100万个序列中，ESMFold对约29%的结构赋予了较高的置信度(平均pLDDT > 0.7) ，对 个结构赋予了高置信度(平均pLDDT )(图3A)。我们利用约1600个GPU小时的计算资源，在不到一天的时间内完成了这些预测，这表明ESMFold可以在计算可行的时间范围内为更大规模的宏基因组序列生成结构预测。

Of the high-confidence structures, we identify with low structural similarity (TM-score ) to any known protein chain in the PDB (Fig. 3B), most of which also agree well with the predictions from full AlphaFold2 with MSA: 80% of corresponding ESMFold-AlphaFold2 predictions have TM-score greater than 0.7, and 50% have TM-score greater than 0.87 (fig. S4). We also find that of the high-confidence structures have low sequence similarity to any sequence in UniRef90 (Methods). Interestingly, we find 317 high-confidence structures that have both low structural and sequence similarity to proteins in these databases (for example, MGYP000706186022; Fig. 3C and fig. S2; table S4), indicating that ESMFold can identify regions of the protein landscape that are distant from existing knowledge.

在高置信度结构中，我们鉴定出 个与蛋白质数据库(PDB)中任何已知蛋白质链的结构相似度较低(TM分数 )的结构(图3B)，其中大多数结构也与使用多序列比对(MSA)的完整AlphaFold2的预测结果高度一致:相应的ESMFold - AlphaFold2预测结果中，80%的TM分数大于0.7，50%的TM分数大于0.87(图S4)。我们还发现， 个高置信度结构与UniRef90数据库中的任何序列的序列相似度都较低(方法)。有趣的是，我们发现有317个高置信度结构与这些数据库中的蛋白质在结构和序列上的相似度都较低(例如，MGYP000706186022；图3C和图S2；表S4)，这表明ESMFold能够识别出与现有知识差异较大的蛋白质结构区域。

Many high-confidence structures with low similarity to UniRef90 sequences, on the other hand, do have similar structures in the PDB, which can enable structure-based functional insight that would be difficult to obtain from sequence information alone. For example, MGnify sequence MGYP000936678158 has no significant matches to any entry in UniRef90, nor any significant matches via a jackhmmer (51) reference proteome search, but has a predicted structure conserved across many nucleases (PDB 5YET\_B, TM-score 0.68; PDB 3HR4\_A, TM-score 0.67) (Fig. 3D and table S4); similarly, MGnify sequence MGYP004000959047 has no significant UniRef90 or jackhmmer reference proteome matches but its predicted structure has high similarity to experimental structures of lipid binding domains (PDB 6BYM\_A, TM-score 0.80; PDB 5YQP\_B, TM-score 0.78) (Fig. 3E and table S4). These results illustrate how ESMFold’s efficient structural prediction capabilities can augment our ability to explore large and poorly-understood regions of the metagenomic protein universe.

另一方面，许多与UniRef90序列相似度较低但置信度较高的结构，在蛋白质数据库(PDB)中确实存在相似结构，这有助于从结构层面洞察功能，而仅依靠序列信息则难以实现这一点。例如，MGnify序列MGYP000936678158与UniRef90中的任何条目均无显著匹配，通过jackhmmer(51)参考蛋白质组搜索也未发现显著匹配，但预测结构显示其与多种核酸酶的结构保守(PDB 5YET\_B，模板匹配分数(TM-score)为0.68；PDB 3HR4\_A，TM-score为0.67)(图3D和表S4)；同样，MGnify序列MGYP004000959047与UniRef90或jackhmmer参考蛋白质组均无显著匹配，但其预测结构与脂质结合域的实验结构高度相似(PDB 6BYM\_A，TM-score为0.80；PDB 5YQP\_B，TM-score为0.78)(图3E和表S4)。这些结果表明，ESMFold高效的结构预测能力能够增强我们探索宏基因组蛋白质领域中未知区域的能力。

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。本预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予bioRxiv永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议(CC-BY-NC-ND 4.0 International license)提供。

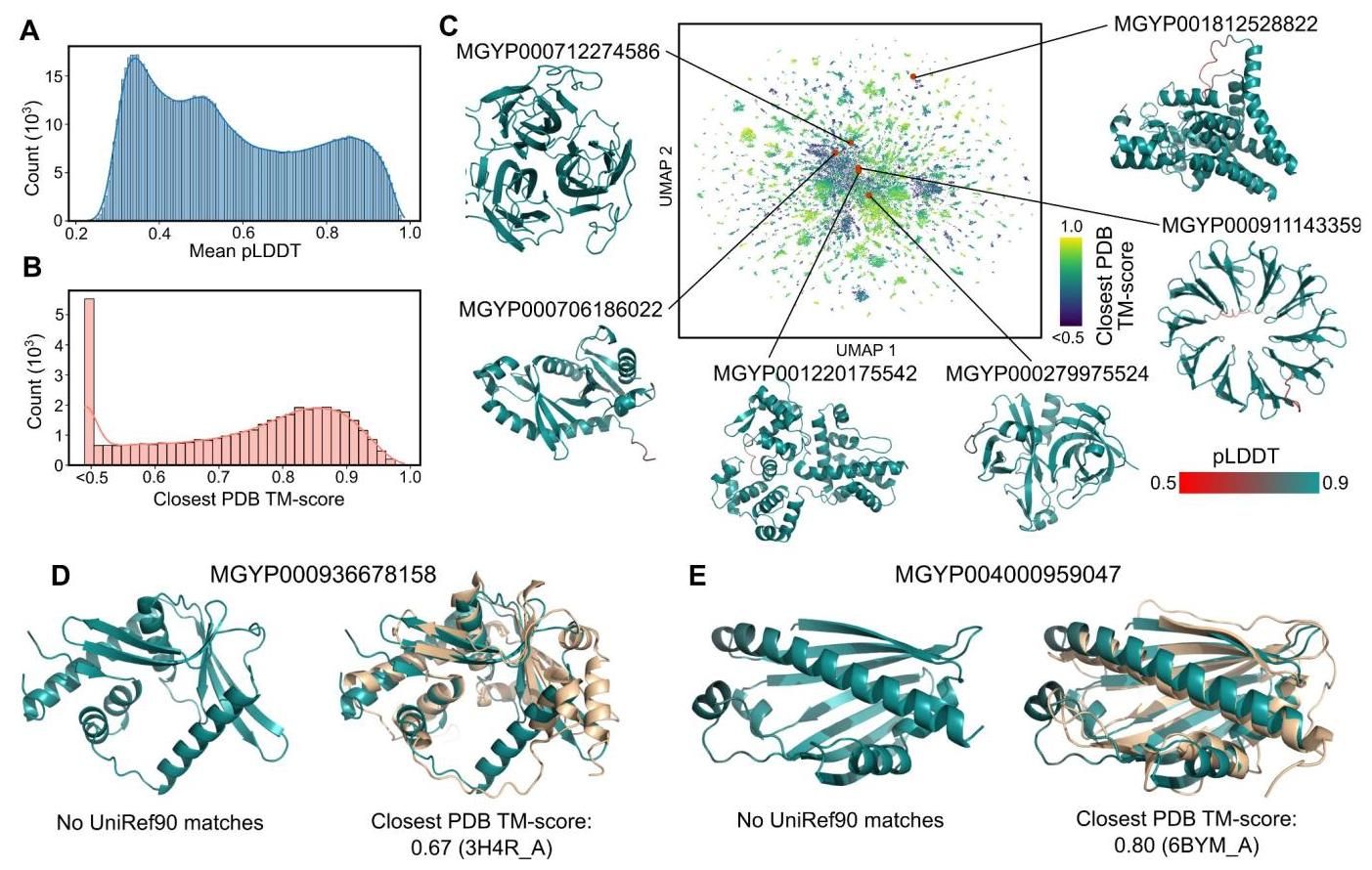


Figure 3: Exploring metagenomic structural space

图3:探索宏基因组结构空间

(A) The distribution of mean pLDDT values computed for each of 1 million ESMFold-predicted structures from the MGnify database. (B) The distribution of the TM-score to the most similar PDB structure for each of highest confidence (mean pLDDT ) structures. Values were obtained by a Foldseek (52) search that does not report structures below a TM-score cutoff of 0.5 . (C) High-confidence protein structures are visualized in two dimensions using the UMAP algorithm and colored according to distance from nearest PDB structure, where regions with low similarity to known structures are colored in dark blue. Example protein structures and their locations within the sequence landscape are provided; see also figure S2 and table S4. (D, E) Examples of two ESMFold-predicted structures that have good agreement with experimental structures in the PDB but that have low sequence identity to any sequence in UniRef90, potentially enabling structure-based functional insight when sequence information is inadequate. (D) The predicted structure of MGYP000936678158 aligns to an experimental structure from a bacterial nuclease (light brown, PDB: 3H4R), while (E) the predicted structure of MGYP004000959047 aligns to an experimental structure from a bacterial sterol binding domain (light brown, PDB: 6BYM).

(A)对MGnify数据库中100万个ESMFold预测结构计算得到的平均预测局部距离差异测试(pLDDT)值的分布。(B) 个最高置信度(平均pLDDT )结构与最相似PDB结构的TM-score分布。这些值通过Foldseek(52)搜索获得，该搜索不报告TM-score低于0.5的结构。(C)使用均匀流形近似与投影(UMAP)算法将高置信度蛋白质结构进行二维可视化，并根据与最近PDB结构的距离进行着色，其中与已知结构相似度较低的区域用深蓝色表示。图中展示了示例蛋白质结构及其在序列图谱中的位置；另见图S2和表S4。(D、E)两个ESMFold预测结构的示例，它们与PDB中的实验结构高度吻合，但与UniRef90中的任何序列的序列同一性较低，当序列信息不足时，这可能有助于从结构层面洞察功能。(D)MGYP000936678158的预测结构与一种细菌核酸酶的实验结构对齐(浅棕色，PDB: 3H4R)，而(E)MGYP004000959047的预测结构与一种细菌甾醇结合域的实验结构对齐(浅棕色，PDB: 6BYM)。

# Conclusions

# 结论

We find that language models trained with an unsupervised learning objective across a large database of evolutionarily diverse protein sequences enable atomic resolution prediction of protein structure. Scaling language models up to parameters enables systematic study of the effect of scale on the learning of protein structure. We see non-linear improvements in protein structure predictions as a function of model scale, and observe a strong link between how well the language model understands a sequence (as measured by perplexity) and the structure prediction that emerges.

我们发现，在包含大量进化上不同的蛋白质序列的数据库中，通过无监督学习目标训练的语言模型能够实现蛋白质结构的原子分辨率预测。将语言模型的参数扩展到 ，可以系统地研究模型规模对蛋白质结构学习的影响。我们观察到，蛋白质结构预测的性能随着模型规模的增加呈现非线性提升，并且发现语言模型对序列的理解程度(通过困惑度衡量)与结构预测结果之间存在紧密联系。

The ESM-2 family of models are the largest protein language models trained to date, with just an order of magnitude fewer parameters than the largest models of text that have recently been developed. ESM-2 has substantial improvements over prior models and even at parameters ESM-2 captures a more accurate picture of structure than ESM-1 generation language models at parameters. ESM-2 compares favorably to other recent, significant protein language models.

ESMFold-2系列模型是迄今为止训练的最大的蛋白质语言模型，其参数数量仅比最近开发的最大文本模型少一个数量级。与之前的模型相比，ESMFold-2有了显著改进，即使在 个参数的情况下，ESMFold-2对结构的描述也比具有 个参数的ESMFold-1代语言模型更准确。ESMFold-2与其他近期重要的蛋白质语言模型相比具有优势。

We show that the biggest driver of performance for ESMFold is the language model. With the strong link between language modeling perplexity and accuracy of structure prediction, we find that comparable predictions to state-of-the-art models can be obtained when the sequence is well understood by ESM-2.

我们发现，ESMFold性能的最大驱动因素是语言模型。由于语言建模的困惑度与结构预测的准确性之间存在紧密联系，我们发现，当ESMFold-2能够很好地理解序列时，其预测结果可与最先进的模型相媲美。

ESMFold obtains accurate atomic resolution structure predictions with up to an order of magnitude improvement in inference time over AlphaFold2. In practice the speed advantages are even greater as ESMFold removes the need to search for evolutionarily related sequences to construct an MSA. The time for this search can be substantial, although new faster methods(49,53)can reduce this.

ESMFold能够实现准确的原子分辨率结构预测，与AlphaFold2相比，推理时间最多可提高一个数量级。实际上，由于ESMFold无需搜索进化相关序列来构建多序列比对(MSA)，其速度优势更加明显。尽管新的更快的方法(49,53)可以减少搜索时间，但这一搜索过程可能仍然相当耗时。

The inference time advantage makes it possible to efficiently map the structural space of large metagenomics sequence databases. Alongside structure-based tools for identifying remote homology and conservation, rapid and accurate structure prediction with ESMFold can help to play a role in structural and functional analysis of large collections of novel sequences. Obtaining millions of predicted structures within practical timescales can help reveal new insights into the breadth and diversity of natural proteins, and enable the discovery of new protein structures and functions.

推理时间优势使得高效映射大型宏基因组序列数据库的结构空间成为可能。除了用于识别远缘同源性和保守性的基于结构的工具外，使用ESMFold进行快速准确的结构预测有助于在大量新序列的结构和功能分析中发挥作用。在实际时间尺度内获得数百万个预测结构有助于揭示对天然蛋白质的广度和多样性的新见解，并推动新蛋白质结构和功能的发现。

# Acknowledgements

# 致谢

We would like to thank Justas Dauparas, Laurens van der Maaten, Sergey Ovchinnikov, Ammar Rizvi, Jon Shepard, Joe Spisak, Sainbayar Sukhbaatar and Robert Verkuil for technical help, feedback, program support and discussions that helped shape this project.

我们要感谢尤斯塔斯·道帕拉斯(Justas Dauparas)、劳伦斯·范德马滕(Laurens van der Maaten)、谢尔盖·奥夫钦尼科夫(Sergey Ovchinnikov)、阿马尔·里兹维(Ammar Rizvi)、乔恩·谢泼德(Jon Shepard)、乔·斯皮萨克(Joe Spisak)、赛恩巴亚尔·苏赫巴塔尔(Sainbayar Sukhbaatar)和罗伯特·韦尔库尔(Robert Verkuil)提供的技术帮助、反馈、程序支持以及讨论，这些都有助于塑造这个项目。

References

参考文献

1. Z. S. Harris, Distributional Structure. WORD. 10, 146-162 (1954).

1. Z. S. 哈里斯(Z. S. Harris)，分布结构。《单词》(WORD)，10，146 - 162(1954)。

2. D. Altschuh, T. Vernet, P. Berti, D. Moras, K. Nagai, Coordinated amino acid changes in homologous protein families. Protein Eng. Des. Sel. 2, 193-199 (1988).

2. D. 阿尔楚(D. Altschuh)、T. 韦尔内(T. Vernet)、P. 贝尔蒂(P. Berti)、D. 莫拉斯(D. Moras)、K. 永井(K. Nagai)，同源蛋白质家族中的氨基酸协同变化。《蛋白质工程、设计与筛选》(Protein Eng. Des. Sel.)，2，193 - 199(1988)。

3. C. Yanofsky, V. Horn, D. Thorpe, Protein Structure Relationships Revealed By Mutational Analysis. Science. 146, 1593-4 (1964).

3. C. 亚诺夫斯基(C. Yanofsky)、V. 霍恩(V. Horn)、D. 索普(D. Thorpe)，通过突变分析揭示的蛋白质结构关系。《科学》(Science)，146，1593 - 4(1964)。

4. U. Göbel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins. Proteins Struct. Funct. Bioinforma. 18, 309-317 (1994).

4. U. 戈贝尔(U. Göbel)、C. 桑德(C. Sander)、R. 施奈德(R. Schneider)、A. 巴伦西亚(A. Valencia)，蛋白质中的相关突变和残基接触。《蛋白质结构、功能与生物信息学》(Proteins Struct. Funct. Bioinforma.)，18，309 - 317(1994)。

5. A. S. Lapedes, B. G. Giraud, L. Liu, G. D. Stormo, Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. Lect. Notes-Monogr. Ser. 33, 236-256 (1999).

5. A. S. 拉佩德斯(A. S. Lapedes)、B. G. 吉罗(B. G. Giraud)、L. 刘(L. Liu)、G. D. 斯托莫(G. D. Stormo)，蛋白质序列模型中的相关突变:系统发育和结构效应。《讲义笔记 - 专题系列》(Lect. Notes - Monogr. Ser.)，33，236 - 256(1999)。

6. J. Thomas, N. Ramakrishnan, C. Bailey-Kellogg, Graphical models of residue coupling in protein families. IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008), pp. 183-197.

6. J. 托马斯(J. Thomas)、N. 拉马克里什南(N. Ramakrishnan)、C. 贝利 - 凯洛格(C. Bailey - Kellogg)，蛋白质家族中残基耦合的图形模型。《电气与电子工程师协会/美国计算机协会计算生物学与生物信息学汇刊》(IEEE/ACM Trans. Comput. Biol. Bioinform.)，5(2008)，第183 - 197页。

7. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. U. S. A. 106, 67-72 (2009).

7. M. 韦格特(M. Weigt)、R. A. 怀特(R. A. White)、H. 祖尔曼特(H. Szurmant)、J. A. 霍赫(J. A. Hoch)、T. 华(T. Hwa)，通过消息传递识别蛋白质 - 蛋白质相互作用中的直接残基接触。《美国国家科学院院刊》(Proc. Natl. Acad. Sci. U. S. A.)，106，67 - 72(2009)。

8. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U. S. A. 108, E1293-E1301 (2011).

8. F. 莫尔科斯(F. Morcos)、A. 帕尼亚尼(A. Pagnani)、B. 伦特(B. Lunt)、A. 贝尔托利诺(A. Bertolino)、D. S. 马克斯(D. S. Marks)、C. 桑德(C. Sander)、R. 泽奇纳(R. Zecchina)、J. N. 奥努奇克(J. N. Onuchic)、T. 华(T. Hwa)、M. 韦格特(M. Weigt)，残基协同进化的直接耦合分析揭示了许多蛋白质家族中的天然接触。《美国国家科学院院刊》(Proc. Natl. Acad. Sci. U. S. A.)，108，E1293 - E1301(2011)。

9. T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, D. S. Marks, Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35, 128-135 (2017).

9. T. A. 霍普夫(T. A. Hopf)、J. B. 英格拉姆(J. B. Ingraham)、F. J. 波埃尔维耶克(F. J. Poelwijk)、C. P. I. 沙尔费(C. P. I. Schärfe)、M. 施普林格(M. Springer)、C. 桑德(C. Sander)、D. S. 马克斯(D. S. Marks)，从序列共变预测突变效应。《自然生物技术》(Nat. Biotechnol.)，35，128 - 135(2017)。

10. C. B. Anfinsen, Principles that Govern the Folding of Protein Chains. Science. 181, 223-230 (1973).

10. C. B. 安芬森(C. B. Anfinsen)，支配蛋白质链折叠的原则。《科学》(Science)，181，223 - 230(1973)。

11. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners. CoRR. abs/2005.14165 (2020) (available at https://arxiv.org/abs/2005.14165).

11. T. B. 布朗(T. B. Brown)、B. 曼(B. Mann)、N. 赖德(N. Ryder)、M. 苏比亚(M. Subbiah)、J. 卡普兰(J. Kaplan)、P. 达里瓦尔(P. Dhariwal)、A. 尼尔卡坦(A. Neelakantan)、P. 夏姆(P. Shyam)、G. 萨斯特里(G. Sastry)、A. 阿斯凯尔(A. Askell)、S. 阿加瓦尔(S. Agarwal)、A. 赫伯特 - 沃斯(A. Herbert - Voss)、G. 克鲁格(G. Krueger)、T. 亨尼根(T. Henighan)、R. 蔡尔德(R. Child)、A. 拉梅什(A. Ramesh)、D. M. 齐格勒(D. M. Ziegler)、J. 吴(J. Wu)、C. 温特(C. Winter)、C. 黑塞(C. Hesse)、M. 陈(M. Chen)、E. 西格勒(E. Sigler)、M. 利特温(M. Litwin)、S. 格雷(S. Gray)、B. 切斯(B. Chess)、J. 克拉克(J. Clark)、C. 伯纳德(C. Berner)、S. 麦坎德利什(S. McCandlish)、A. 拉德福德(A. Radford)、I. 苏茨克弗(I. Sutskever)、D. 阿莫迪(D. Amodei)，语言模型是少样本学习者。《计算机研究存储库》(CoRR)，abs/2005.14165(2020)(可在https://arxiv.org/abs/2005.14165获取)。

12. J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS, 46 (2022).

12. 魏(J. Wei)、博斯马(M. Bosma)、赵(V. Y. Zhao)、古(K. Guu)、余(A. W. Yu)、莱斯特(B. Lester)、杜(N. Du)、戴(A. M. Dai)、乐(Q. V. Le)，《微调语言模型是零样本学习者》，46(2022)。

13. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain of Thought Prompting Elicits Reasoning in Large Language Models (2022), (available at http://arxiv.org/abs/2201.11903).

13. 魏(J. Wei)、王(X. Wang)、舒尔曼斯(D. Schuurmans)、博斯马(M. Bosma)、伊希特(B. Ichter)、夏(F. Xia)、池(E. Chi)、乐(Q. Le)、周(D. Zhou)，《思维链提示在大语言模型中引发推理》(2022)，(可在http://arxiv.org/abs/2201.11903获取)。

14. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: Scaling Language Modeling with Pathways (2022),, doi:10.48550/arXiv.2204.02311.

14. 乔杜里(A. Chowdhery)、纳朗(S. Narang)、德夫林(J. Devlin)、博斯马(M. Bosma)、米什拉(G. Mishra)、罗伯茨(A. Roberts)、巴哈姆(P. Barham)、钟(H. W. Chung)、萨顿(C. Sutton)、格尔曼(S. Gehrmann)、舒(P. Schuh)、施(K. Shi)、茨维亚申科(S. Tsvyashchenko)、梅内兹(J. Maynez)、拉奥(A. Rao)、巴恩斯(P. Barnes)、泰(Y. Tay)、沙泽尔(N. Shazeer)、普拉巴卡兰(V. Prabhakaran)、赖夫(E. Reif)、杜(N. Du)、哈钦森(B. Hutchinson)、波普(R. Pope)、布拉德伯里(J. Bradbury)、奥斯汀(J. Austin)、伊萨德(M. Isard)、古尔 - 阿里(G. Gur - Ari)、尹(P. Yin)、杜克(T. Duke)、列夫斯卡娅(A. Levskaya)、格马瓦特(S. Ghemawat)、德夫(S. Dev)、米哈列夫斯基(H. Michalewski)、加西亚(X. Garcia)、米斯拉(V. Misra)、罗宾逊(K. Robinson)、费杜斯(L. Fedus)、周(D. Zhou)、伊波利托(D. Ippolito)、栾(D. Luan)、林(H. Lim)、佐夫(B. Zoph)、斯皮里多诺夫(A. Spiridonov)、塞帕西(R. Sepassi)、多汉(D. Dohan)、阿格拉瓦尔(S. Agrawal)、奥梅尼克(M. Omernick)、戴(A. M. Dai)、皮莱(T. S. Pillai)、佩拉特(M. Pellat)、莱科维茨(A. Lewkowycz)、莫雷拉(E. Moreira)、蔡尔德(R. Child)、波洛佐夫(O. Polozov)、李(K. Lee)、周(Z. Zhou)、王(X. Wang)、塞塔(B. Saeta)、迪亚兹(M. Diaz)、菲拉特(O. Firat)、卡斯塔塔(M. Catasta)、魏(J. Wei)、迈尔 - 赫尔斯滕(K. Meier - Hellstern)、埃克(D. Eck)、迪安(J. Dean)、彼得罗夫(S. Petrov)、菲德尔(N. Fiedel)，《PaLM:通过路径扩展语言建模》(2022)，doi:10.48550/arXiv.2204.02311。

15. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. 118 (2021), doi:10.1073/pnas.2016239118.

15. 里夫斯(A. Rives)、迈尔(J. Meier)、塞尔库(T. Sercu)、戈亚尔(S. Goyal)、林(Z. Lin)、刘(J. Liu)、郭(D. Guo)、奥特(M. Ott)、齐特尼克(C. L. Zitnick)、马(J. Ma)、弗格斯(R. Fergus)，《将无监督学习扩展到2.5亿个蛋白质序列揭示生物结构和功能》。《美国国家科学院院刊》118(2021)，doi:10.1073/pnas.2016239118。

16. R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, Y. S. Song, "Evaluating Protein Transfer Learning with TAPE" in Neural Information Processing Systems (Cold Spring Harbor Laboratory, 2019; https://doi.org/10.1101/676825 http://arxiv.org/abs/1906.08230).

16. 拉奥(R. Rao)、巴塔查里亚(N. Bhattacharya)、托马斯(N. Thomas)、段(Y. Duan)、陈(X. Chen)、坎尼(J. Canny)、阿比尔(P. Abbeel)、宋(Y. S. Song)，《用TAPE评估蛋白质迁移学习》，载于《神经信息处理系统大会》(冷泉港实验室，2019；https://doi.org/10.1101/676825 http://arxiv.org/abs/1906.08230)。

17. T. Bepler, B. Berger, "Learning protein sequence embeddings using information from structure" in International Conference on Learning Representations (2019;

17. 贝普勒(T. Bepler)、伯杰(B. Berger)，《利用结构信息学习蛋白质序列嵌入》，载于《国际学习表征会议》(2019；

https://openreview.net/forum?id=SygLehCqtm).

18. T. Bepler, B. Berger, Learning the protein language: Evolution, structure, and function. Cell Syst. 12, 654-669.e3 (2021).

18. 贝普勒(T. Bepler)、伯杰(B. Berger)，《学习蛋白质语言:进化、结构和功能》。《细胞系统》12，654 - 669.e3(2021)。

19. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, D. Bhowmik, B. Rost, ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. IEEE Trans. Pattern Anal. Mach. Intell. 14 (2021), doi:10.1109/TPAMI.2021.3095381.

19. 埃尔纳加尔(A. Elnaggar)、海因津格(M. Heinzinger)、达拉戈(C. Dallago)、里哈维(G. Rihawi)、王(Y. Wang)、琼斯(L. Jones)、吉布斯(T. Gibbs)、费赫尔(T. Feher)、安杰勒(C. Angerer)、鲍米克(D. Bhowmik)、罗斯特(B. Rost)，《ProtTrans:通过自监督深度学习和高性能计算破解生命代码语言》。《IEEE模式分析与机器智能汇刊》14(2021)，doi:10.1109/TPAMI.2021.3095381。

20. A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, R. Socher, ProGen: Language Modeling for Protein Generation. bioRxiv (2020) (available at http://arxiv.org/abs/2004.03497).

20. 马丹尼(A. Madani)、麦坎(B. McCann)、奈克(N. Naik)、凯斯卡尔(N. S. Keskar)、阿南德(N. Anand)、江口(R. R. Eguchi)、黄(P. - S. Huang)、索切尔(R. Socher)，《ProGen:用于蛋白质生成的语言建模》。生物预印本(2020)(可在http://arxiv.org/abs/2004.03497获取)。

21. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-only deep representation learning. Nat. Methods. 12, 1315-1322 (2019)

21. E. C. 艾利(Alley)、G. 基姆利亚(Khimulya)、S. 比斯瓦斯(Biswas)、M. 阿尔库赖希(AlQuraishi)、G. M. 丘奇(Church)，仅基于序列的深度表征学习实现统一的理性蛋白质工程。《自然·方法》，12 卷，1315 - 1322 页(2019 年)

22. M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, B. Rost, Modeling the language of life - Deep Learning Protein Sequences. bioRxiv, 614313 (2019).

22. M. 海因津格(Heinzinger)、A. 埃尔纳加尔(Elnaggar)、Y. 王(Wang)、C. 达拉戈(Dallago)、D. 涅恰耶夫(Nechaev)、F. 马泰斯(Matthes)、B. 罗斯特(Rost)，模拟生命语言——深度学习蛋白质序列。生物预印本服务器(bioRxiv)，614313 号(2019 年)。

23. B. Hie, E. D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape. Science. 371, 284-288 (2021).

23. B. 希(Hie)、E. D. 钟(Zhong)、B. 伯杰(Berger)、B. 布赖森(Bryson)，学习病毒进化与逃逸的语言。《科学》，371 卷，284 - 288 页(2021 年)。

24. R. Chowdhury, N. Bouatta, S. Biswas, C. Rochereau, G. M. Church, P. K. Sorger, M. AlQuraishi, "Single-sequence protein structure prediction using language models from deep learning" (preprint, Bioinformatics, 2021),, doi:10.1101/2021.08.02.454840.

24. R. 乔杜里(Chowdhury)、N. 布阿塔(Bouatta)、S. 比斯瓦斯(Biswas)、C. 罗舍罗(Rochereau)、G. M. 丘奇(Church)、P. K. 索尔格(Sorger)、M. 阿尔库赖希(AlQuraishi)，“利用深度学习语言模型进行单序列蛋白质结构预测”(预印本，《生物信息学》，2021 年)，doi:10.1101/2021.08.02.454840。

25. J. A. Ruffolo, J. J. Gray, Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Biophys. J. 121, 155a-156a (2022).

25. J. A. 拉福洛(Ruffolo)、J. J. 格雷(Gray)，基于大量天然抗体深度学习的快速、准确抗体结构预测。《生物物理学杂志》，121 卷，155a - 156a 页(2022 年)。

26. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583-589 (2021).

26. J. 江珀(Jumper)、R. 埃文斯(Evans)、A. 普里策尔(Pritzel)、T. 格林(Green)、M. 菲古尔诺夫(Figurnov)、O. 龙内贝格尔(Ronneberger)、K. 通亚苏武纳库尔(Tunyasuvunakool)、R. 贝茨(Bates)、A. 齐德克(Zidek)、A. 波塔彭科(Potapenko)、A. 布里格兰德(Bridgland)、C. 迈耶(Meyer)、S. A. A. 科尔(Kohl)、A. J. 巴拉德(Ballard)、A. 考伊(Cowie)、B. 罗梅拉 - 帕雷德斯(Romera - Paredes)、S. 尼科洛夫(Nikolov)、R. 贾因(Jain)、J. 阿德勒(Adler)、T. 巴克(Back)、S. 彼得森(Petersen)、D. 赖曼(Reiman)、E. 克兰西(Clancy)、M. 齐林斯基(Zielinski)、M. 施泰内格(Steinegger)、M. 帕霍尔茨卡(Pacholska)、T. 贝格哈默(Berghammer)、S. 博登斯坦(Bodenstein)、D. 西尔弗(Silver)、O. 维尼亚尔斯(Vinyals)、A. W. 西尼尔(Senior)、K. 卡武库奥卢(Kavukcuoglu)、P. 科利(Kohli)、D. 哈萨比斯(Hassabis)，利用阿尔法折叠(AlphaFold)进行高精度蛋白质结构预测。《自然》，596 卷，583 - 589 页(2021 年)。

27. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. Science. 373, 871-876 (2021).

27. M. 贝克(Baek)、F. 迪马约(DiMaio)、I. 阿尼申科(Anishchenko)、J. 道帕拉斯(Dauparas)、S. 奥夫钦尼科夫(Ovchinnikov)、G. R. 李(Lee)、J. 王(Wang)、Q. 丛(Cong)、L. N. 金奇(Kinch)、R. D. 舍费尔(Schaeffer)、C. 米兰(Millán)、H. 朴(Park)、C. 亚当斯(Adams)、C. R. 格拉斯曼(Glassman)、A. 德乔瓦尼(DeGiovanni)、J. H. 佩雷拉(Pereira)、A. V. 罗德里格斯(Rodrigues)、A. A. 范戴克(van Dijk)、A. C. 埃布雷希特(Ebrecht)、D. J. 奥珀曼(Opperman)、T. 萨格迈斯特(Sagmeister)、C. 布尔黑勒(Buhlheller)、T. 帕夫科夫 - 凯勒(Pavkov - Keller)、M. K. 拉蒂纳斯瓦米(Rathinaswamy)、U. 达尔瓦迪(Dalwadi)、C. K. 叶(Yip)、J. E. 伯克(Burke)、K. C. 加西亚(Garcia)、N. V. 格里辛(Grishin)、P. D. 亚当斯(Adams)、R. J. 里德(Read)、D. 贝克(Baker)，利用三轨道神经网络准确预测蛋白质结构和相互作用。《科学》，373 卷，871 - 876 页(2021 年)。

28. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, U. Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 31, 926-932 (2014).

28. B. E. 苏泽克(Suzek)、Y. 王(Wang)、H. 黄(Huang)、P. B. 麦加维(McGarvey)、C. H. 吴(Wu)、U. 联盟(Consortium)，UniRef 聚类:一种全面且可扩展的改进序列相似性搜索的替代方法。《生物信息学》，31 卷，926 - 932 页(2014 年)。

29. A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, E. Sakharova, M. Scheremetjew, A. Korobeynikov, A. Shlemov, O. Kunyavskaya, A. Lapidus, R. D. Finn, MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res., gkz1035 (2019).

29. A. L. 米切尔(Mitchell)、A. 阿尔梅达(Almeida)、M. 贝拉克奥切亚(Beracochea)、M. 博兰(Boland)、J. 伯金(Burgin)、G. 科克伦(Cochrane)、M. R. 克鲁索(Crusoe)、V. 卡莱(Kale)、S. C. 波特(Potter)、L. J. 理查森(Richardson)、E. 萨哈罗娃(Sakharova)、M. 舍列梅捷夫(Scheremetjew)、A. 科罗贝尼科夫(Korobeynikov)、A. 什廖莫夫(Shlemov)、O. 库尼亚夫斯卡娅(Kunyavskaya)、A. 拉皮多斯(Lapidus)、R. D. 芬恩(Finn)，MGnify:2020 年的微生物组分析资源。《核酸研究》，gkz1035 号(2019 年)。

30. M. Steinegger, M. Mirdita, J. Söding, Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods. 16, 603-606 (2019).

30. M. 施泰内格(Steinegger)、M. 米尔迪塔(Mirdita)、J. 瑟丁(Söding)，蛋白质水平组装使宏基因组样本中的蛋白质序列回收率提高数倍。《自然·方法》，16 卷，603 - 606 页(2019 年)。

31. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models. ArXiv Prepr: ArXiv200108361 (2020).

31. J. 卡普兰(J. Kaplan)、S. 麦坎德利什(S. McCandlish)、T. 亨尼根(T. Henighan)、T. B. 布朗(T. B. Brown)、B. 切斯(B. Chess)、R. 蔡尔德(R. Child)、S. 格雷(S. Gray)、A. 拉德福德(A. Radford)、J. 吴(J. Wu)、D. 阿莫迪(D. Amodei)，神经语言模型的缩放定律。预印本:ArXiv200108361 (2020)。

32. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training Compute-Optimal Large Language Models (2022), doi:10.48550/arXiv.2203.15556.

32. J. 霍夫曼(J. Hoffmann)、S. 博尔戈奥德(S. Borgeaud)、A. 门施(A. Mensch)、E. 布恰茨卡娅(E. Buchatskaya)、T. 蔡(T. Cai)、E. 卢瑟福(E. Rutherford)、D. 德·L·卡萨斯(D. de L. Casas)、L. A. 亨德里克斯(L. A. Hendricks)、J. 韦尔布尔(J. Welbl)、A. 克拉克(A. Clark)、T. 亨尼根(T. Hennigan)、E. 诺兰(E. Noland)、K. 米利坎(K. Millican)、G. 范登德里舍(G. van den Driessche)、B. 达莫克(B. Damoc)、A. 盖伊(A. Guy)、S. 奥辛德罗(S. Osindero)、K. 西莫扬(K. Simonyan)、E. 埃尔森(E. Elsen)、J. W. 雷(J. W. Rae)、O. 维尼亚尔斯(O. Vinyals)、L. 西弗雷(L. Sifre)，训练计算最优的大语言模型 (2022)，doi:10.48550/arXiv.2203.15556。

33. S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: Open Pre-trained Transformer Language Models (2022), doi:10.48550/arXiv.2205.01068.

33. S. 张(S. Zhang)、S. 罗勒(S. Roller)、N. 戈亚尔(N. Goyal)、M. 阿泰特克塞(M. Artetxe)、M. 陈(M. Chen)、S. 陈(S. Chen)、C. 德万(C. Dewan)、M. 迪亚布(M. Diab)、X. 李(X. Li)、X. V. 林(X. V. Lin)、T. 米哈伊洛夫(T. Mihaylov)、M. 奥特(M. Ott)、S. 施莱弗(S. Shleifer)、K. 舒斯特(K. Shuster)、D. 西米格(D. Simig)、P. S. 库拉(P. S. Koura)、A. 斯里达尔(A. Sridhar)、T. 王(T. Wang)、L. 泽特尔莫耶(L. Zettlemoyer)，OPT:开放预训练的Transformer语言模型 (2022)，doi:10.48550/arXiv.2205.01068。

34. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" in Proceedings of the 2019 Conference of the North merican Chapter of the Association for Computational Linguistics: Human Language

34. J. 德夫林(J. Devlin)、M.-W. 张(M.-W. Chang)、K. 李(K. Lee)、K. 图托纳娃(K. Toutanova)，“BERT:用于语言理解的深度双向Transformer预训练”，载于2019年北美计算语言学协会会议论文集:人类语言

Technologies, Volume 1 (Long and Short Papers) (Association for Computational Linguistics, Minneapolis, Minnesota, 2019; http://arxiv.org/abs/1810.04805), pp. 4171-4186.

技术，第1卷(长论文和短论文)(计算语言学协会，美国明尼苏达州明尼阿波利斯，2019；http://arxiv.org/abs/1810.04805)，第4171 - 4186页。

35. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding" in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (Association for Computational Linguistics, Brussels, Belgium, 2018; https://aclanthology.org/W18-5446), pp. 353-355.

35. A. 王(A. Wang)、A. 辛格(A. Singh)、J. 迈克尔(J. Michael)、F. 希尔(F. Hill)、O. 利维(O. Levy)、S. 鲍曼(S. Bowman)，“GLUE:自然语言理解的多任务基准和分析平台”，载于2018年EMNLP研讨会“黑盒NLP:分析和解释用于NLP的神经网络”论文集(计算语言学协会，比利时布鲁塞尔，2018；https://aclanthology.org/W18-5446)，第353 - 355页。

36. A. Rogers, O. Kovaleva, A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works. Trans. Assoc. Comput. Linguist. 8, 842-866 (2020).

36. A. 罗杰斯(A. Rogers)、O. 科瓦列娃(O. Kovaleva)、A. 鲁姆希斯基(A. Rumshisky)，BERT学入门:我们对BERT工作原理的了解。《计算语言学协会汇刊》8，842 - 866 (2020)。

37. J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, N. Rajani, "BERTology Meets Biology: Interpreting Attention in Protein Language Models" in (2020; https://openreview.net/forum?id=YWtLZvLmud7).

37. J. 维格(J. Vig)、A. 马达尼(A. Madani)、L. R. 瓦尔什尼(L. R. Varshney)、C. 熊(C. Xiong)、R. 索切尔(R. Socher)、N. 拉贾尼(N. Rajani)，“BERT学遇见生物学:解读蛋白质语言模型中的注意力机制”，载于 (2020; https://openreview.net/forum?id=YWtLZvLmud7)。

38. R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, "Transformer protein language models are unsupervised structure learners" in International Conference on Learning Representations (Cold Spring Harbor Laboratory, 2021), p. 2020.12.15.422761.

38. R. 拉奥(R. Rao)、J. 迈尔(J. Meier)、T. 塞尔库(T. Sercu)、S. 奥夫钦尼科夫(S. Ovchinnikov)、A. 里夫斯(A. Rives)，“Transformer蛋白质语言模型是无监督的结构学习者”，载于国际学习表征会议(冷泉港实验室，2021)，第2020.12.15.422761页。

39. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmed, Learning generative models for protein fold families. Proteins Struct. Funct. Bioinforma. 79, 1061-1078 (2011).

39. S. 巴拉卡里什南(S. Balakrishnan)、H. 卡米塞蒂(H. Kamisetty)、J. G. 卡博内尔(J. G. Carbonell)、S.-I. 李(S.-I. Lee)、C. J. 兰梅德(C. J. Langmed)，学习蛋白质折叠家族的生成模型。《蛋白质结构、功能与生物信息学》79，1061 - 1078 (2011)。

40. S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranoví, D. Guzenko, B. P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prli, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 47 (2019), doi:10.1093/nar/gky1004.

40. S. K. 伯利(S. K. Burley)、H. M. 伯曼(H. M. Berman)、C. 比卡迪亚(C. Bhikadiya)、C. 毕(C. Bi)、L. 陈(L. Chen)、L. 迪·科斯坦佐(L. Di Costanzo)、C. 克里斯蒂(C. Christie)、K. 达伦伯格(K. Dalenberg)、J. M. 杜阿尔特(J. M. Duarte)、S. 杜塔(S. Dutta)、Z. 冯(Z. Feng)、S. 戈什(S. Ghosh)、D. S. 古德塞尔(D. S. Goodsell)、R. K. 格林(R. K. Green)、V. 古兰诺维(V. Guranoví)、D. 古岑科(D. Guzenko)、B. P. 哈德森(B. P. Hudson)、T. 卡尔罗(T. Kalro)、Y. 梁(Y. Liang)、R. 洛(R. Lowe)、H. 南孔(H. Namkoong)、E. 佩萨克(E. Peisach)、I. 佩里斯科娃(I. Periskova)、A. 普里利(A. Prli)、C. 兰德尔(C. Randle)、A. 罗斯(A. Rose)、P. 罗斯(P. Rose)、R. 萨拉(R. Sala)、M. 塞卡兰(M. Sekharan)、C. 邵(C. Shao)、L. 谭(L. Tan)、陶一鹏(Y.-P. Tao)、Y. 瓦拉萨塔瓦(Y. Valasatava)、M. 沃伊特(M. Voigt)、J. 韦斯特布鲁克(J. Westbrook)、J. 吴(J. Woo)、H. 杨(H. Yang)、J. 扬(J. Young)、M. 朱拉夫列娃(M. Zhuravleva)、C. 扎德茨基(C. Zardecki)，RCSB蛋白质数据库:助力基础生物学、生物医学、生物技术和能源领域研究与教育的生物大分子结构。《核酸研究》47 卷(2019 年)，doi:10.1093/nar/gky1004。

41. J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, T. Schwede, Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins Struct. Funct. Bioinforma. 86, 387-398 (2018).

41. J. 哈斯(J. Haas)、A. 巴尔巴托(A. Barbato)、D. 贝林格(D. Behringer)、G. 施图德(G. Studer)、S. 罗斯(S. Roth)、M. 贝尔托尼(M. Bertoni)、K. 莫斯塔吉尔(K. Mostaguir)、R. 古米耶尼(R. Gumienny)、T. 施韦德(T. Schwede)，连续自动模型评估(CAMEO)对第12届蛋白质结构预测技术评估(CASP12)中结构预测的关键评估的补充。《蛋白质:结构、功能与生物信息学》86, 387 - 398 (2018)。

42. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins Struct. Funct. Bioinforma. 89, 1607-1617 (2021).

42. A. 克里斯塔福维奇(A. Kryshtafovych)、T. 施韦德(T. Schwede)、M. 托普夫(M. Topf)、K. 菲德利斯(K. Fidelis)、J. 莫尔特(J. Moult)，蛋白质结构预测方法的关键评估(CASP)——第十四轮。《蛋白质结构、功能与生物信息学》89 卷，1607 - 1617 页(2021 年)。

43. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710 (2004).

43. 张Y.，斯科尔尼克J.，用于自动评估蛋白质结构模板质量的评分函数。《蛋白质》，57卷，702 - 710页(2004年)。

44. K. K. Yang, A. X. Lu, N. Fusi, Convolutions are competitive with transformers for protein sequence pretraining (2022), doi:10.1101/2022.05.19.492714.

44. 杨 K. K.(K. K. Yang)、陆 A. X.(A. X. Lu)、富西 N.(N. Fusi)，卷积在蛋白质序列预训练方面可与Transformer相媲美(2022年)，doi:10.1101/2022.05.19.492714。

45. S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Comput. Biol. 13, 1-34 (2017).

45. 王S.、孙S.、李Z.、张R.、徐J.，通过超深度学习模型准确从头预测蛋白质接触图。《公共科学图书馆·计算生物学》(PLOS Comput. Biol.)，第13卷，第1 - 34页(2017年)。

46. Y. Liu, P. Palmedo, Q. Ye, B. Berger, J. Peng, Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Syst. 6, 65-74 (2018).

46. 刘(Liu)Y.、帕尔梅多(Palmedo)P.、叶(Ye)Q.、伯杰(Berger)B.、彭(Peng)J.，利用深度卷积神经网络增强进化耦合。《细胞系统》(Cell Syst.)6, 65 - 74 (2018)。

47. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention Is All You Need" in Advances in Neural Information Processing Systems (2017; https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf), pp. 5998-6008.

47. A. 瓦斯瓦尼(A. Vaswani)、N. 沙泽尔(N. Shazeer)、N. 帕尔马(N. Parmar)、J. 乌兹科雷特(J. Uszkoreit)、L. 琼斯(L. Jones)、A. N. 戈麦斯(A. N. Gomez)、Ł. 凯泽(Ł. Kaiser)、I. 波洛苏金(I. Polosukhin)，《注意力即所需一切》(“Attention Is All You Need”)，发表于《神经信息处理系统进展》(Advances in Neural Information Processing Systems)(2017 年；https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf)，第 5998 - 6008 页。

48. C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, A. Rives, "Learning inverse folding from millions of predicted structures" in Proceedings of the 39th International Conference on Machine Learning (PMLR, 2022; https://proceedings.mlr.press/v162/hsu22a.html), pp. 8946-8970.

48. 徐(Hsu)、R. 费库尔(Verkuil)、刘(Liu)、林(Lin)、B. 希(Hie)、T. 塞尔库(Sercu)、A. 勒雷尔(Lerer)、A. 里夫斯(Rives)，“从数百万个预测结构中学习反向折叠”，载于《第39届国际机器学习会议论文集》(PMLR，2022年；https://proceedings.mlr.press/v162/hsu22a.html)，第8946 - 8970页。

49. M. Hauser, M. Steinegger, J. Söding, MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. 32, 1323-1330 (2016).

49. M. 豪泽(M. Hauser)、M. 施泰内格(M. Steinegger)、J. 瑟丁(J. Söding)，MMseqs软件套件用于快速深度聚类和搜索大型蛋白质序列集。《生物信息学》(Bioinformatics)，32卷，1323 - 1330页(2016年)。

50. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zidek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural

50. M. 瓦拉迪(M. Varadi)、S. 安扬戈(S. Anyango)、M. 德什潘德(M. Deshpande)、S. 奈尔(S. Nair)、C. 纳塔西亚(C. Natassia)、G. 约尔达诺娃(G. Yordanova)、D. 袁(D. Yuan)、O. 斯特罗埃(O. Stroe)、G. 伍德(G. Wood)、A. 莱登(A. Laydon)、A. 齐德克(A. Zidek)、T. 格林(T. Green)、K. 通亚苏武纳库尔(K. Tunyasuvunakool)、S. 彼得森(S. Petersen)、J. 江珀(J. Jumper)、E. 克兰西(E. Clancy)、R. 格林(R. Green)、A. 沃拉(A. Vora)、M. 卢特菲(M. Lutfi)、M. 菲古尔诺夫(M. Figurnov)、A. 考伊(A. Cowie)、N. 霍布斯(N. Hobbs)、P. 科利(P. Kohli)、G. 克莱韦格特(G. Kleywegt)、E. 伯尼(E. Birney)、D. 哈萨比斯(D. Hassabis)、S. 韦兰卡尔(S. Velankar)，AlphaFold蛋白质结构数据库:利用高精度模型大幅扩展蛋白质序列空间的结构覆盖范围

coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439-D444 (2022).

核酸研究(Nucleic Acids Res.)50, D439 - D444 (2022)。

51. S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, R. D. Finn, HMMER web server: 2018 update. Nucleic Acids Res. 2, W200-W204 (2018).

51. S. C. 波特(S. C. Potter)、A. 卢恰尼(A. Luciani)、S. R. 埃迪(S. R. Eddy)、Y. 朴(Y. Park)、R. 洛佩斯(R. Lopez)、R. D. 芬恩(R. D. Finn)，HMMER网络服务器:2018年更新。核酸研究(Nucleic Acids Res.)2, W200 - W204 (2018)。

52. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, C. L. M. Gilchrist, J. Söding, M. Steinegger, Foldseek: fast and accurate protein structure search (2022), doi:10.1101/2022.02.07.479398.

52. M. 范肯彭(M. van Kempen)、S. S. 金(S. S. Kim)、C. 图梅谢特(C. Tumescheit)、M. 米尔迪塔(M. Mirdita)、C. L. M. 吉尔克里斯特(C. L. M. Gilchrist)、J. 瑟丁(J. Söding)、M. 施泰内格(M. Steinegger)，Foldseek:快速准确的蛋白质结构搜索 (2022)，doi:10.1101/2022.02.07.479398。

53. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. Nat. Methods. 19, 679-682 (2022).

53. M. 米尔迪塔(M. Mirdita)、K. 舒策(K. Schütze)、Y. 森脇(Y. Moriwaki)、L. 许(L. Heo)、S. 奥夫钦尼科夫(S. Ovchinnikov)、M. 施泰内格(M. Steinegger)，ColabFold:让所有人都能进行蛋白质折叠预测。自然方法(Nat. Methods)19, 679 - 682 (2022)。

54. J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function" (preprint, Synthetic Biology, 2021),, doi:10.1101/2021.07.09.450648.

54. J. 迈尔(J. Meier)、R. 饶(R. Rao)、R. 韦尔库尔(R. Verkuil)、J. 刘(J. Liu)、T. 塞尔库(T. Sercu)、A. 里夫斯(A. Rives)，“语言模型实现对蛋白质功能突变影响的零样本预测”(预印本，合成生物学，2021)，doi:10.1101/2021.07.09.450648。

55. R. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, "MSA Transformer" in Proceedings of the 38th International Conference on Machine Learning (PMLR, 2021; https://proceedings.mlr.press/v139/rao21a.html), pp. 8844-8856.

55. R. 饶(R. Rao)、J. 刘(J. Liu)、R. 韦尔库尔(R. Verkuil)、J. 迈尔(J. Meier)、J. 坎尼(J. Canny)、P. 阿比尔(P. Abbeel)、T. 塞尔库(T. Sercu)、A. 里夫斯(A. Rives)，“MSA变压器(MSA Transformer)”，见第38届国际机器学习会议论文集(PMLR，2021；https://proceedings.mlr.press/v139/rao21a.html)，第8844 - 8856页。

56. S. D. Dunn, L. M. Wahl, G. B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 24, 333-340 (2008).

56. S. D. 邓恩(S. D. Dunn)、L. M. 瓦尔(L. M. Wahl)、G. B. 格洛尔(G. B. Gloor)，不受系统发育或熵影响的互信息显著改善残基接触预测。生物信息学(Bioinformatics)24, 333 - 340 (2008)。

57. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830 (2011).

57. F. 佩德雷戈萨(F. Pedregosa)、G. 瓦罗夸克斯(G. Varoquaux)、A. 格拉姆福特(A. Gramfort)、V. 米歇尔(V. Michel)、B. 蒂里翁(B. Thirion)、O. 格里塞尔(O. Grisel)、M. 布隆德尔(M. Blondel)、P. 普雷滕霍费尔(P. Prettenhofer)、R. 韦斯(R. Weiss)、V. 迪布尔格(V. Dubourg)、J. 范德普拉斯(J. Vanderplas)、A. 帕索斯(A. Passos)、D. 库尔纳波(D. Cournapeau)、M. 布鲁彻(M. Brucher)、M. 佩罗(M. Perrot)、E. 迪谢纳(E. Duchesnay)，Scikit - 学习:Python中的机器学习。机器学习研究杂志(J. Mach. Learn. Res.)12, 2825 - 2830 (2011)。

58. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. 117, 1496-1503 (2020).

58. J. 杨(J. Yang)、I. 阿尼申科(I. Anishchenko)、H. 朴(H. Park)、Z. 彭(Z. Peng)、S. 奥夫钦尼科夫(S. Ovchinnikov)、D. 贝克(D. Baker)，利用预测的残基间取向改进蛋白质结构预测。美国国家科学院院刊(Proc. Natl. Acad. Sci.)117, 1496 - 1503 (2020)。

59. J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, RoFormer: Enhanced Transformer with Rotary Position Embedding (2021), (available at http://arxiv.org/abs/2104.09864).

59. J. 苏(J. Su)、Y. 陆(Y. Lu)、S. 潘(S. Pan)、B. 文(B. Wen)、Y. 刘(Y. Liu)，RoFormer:采用旋转位置嵌入的增强型变压器 (2021)(可在http://arxiv.org/abs/2104.09864获取)。

60. Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, C.-J. Hsieh, LARGE BATCH OPTIMIZATION FOR DEEP LEARNING: TRAINING BERT IN 76 MINUTES, 38 (2020).

60. Y. 尤(Y. You)、J. 李(J. Li)、S. 雷迪(S. Reddi)、J. 许(J. Hseu)、S. 库马尔(S. Kumar)、S. 博贾纳帕利(S. Bhojanapalli)、X. 宋(X. Song)、J. 德梅尔(J. Demmel)、K. 科伊策(K. Keutzer)、C. - J. 谢(C. - J. Hsieh)，深度学习的大批量优化:76分钟训练BERT，38 (2020)。

61. S. Rajbhandari, J. Rasley, O. Ruwase, Y. He, "ZeRO: Memory optimizations Toward Training Trillion Parameter Models" in SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (2020), pp. 1-16.

61. S. 拉杰班达里(S. Rajbhandari)、J. 拉斯利(J. Rasley)、O. 鲁瓦塞(O. Ruwase)、Y. 何(Y. He)，“ZeRO:面向万亿参数模型训练的内存优化”，见SC20:国际高性能计算、网络、存储和分析会议(2020)，第1 - 16页。

62. J. Ho, N. Kalchbrenner, D. Weissenborn, T. Salimans, Axial Attention in Multidimensional Transformers. arXiv (2019) (available at http://arxiv.org/abs/1912.12180).

62. J. 霍(J. Ho)、N. 卡尔克布伦纳(N. Kalchbrenner)、D. 魏森伯恩(D. Weissenborn)、T. 萨利曼斯(T. Salimans)，多维变换器中的轴向注意力。预印本服务器arXiv(2019 年)(可在 http://arxiv.org/abs/1912.12180 查看)。

63. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018), doi:10.48550/ARXIV.1802.03426.

63. L. 麦金尼斯(L. McInnes)、J. 希利(J. Healy)、J. 梅尔维尔(J. Melville)，UMAP:用于降维的均匀流形近似与投影(2018 年)，doi:10.48550/ARXIV.1802.03426。

64. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15 (2018).

64. F. A. 沃尔夫(F. A. Wolf)、P. 安杰勒(P. Angerer)、F. J. 泰斯(F. J. Theis)，SCANPY:大规模单细胞基因表达数据分析。《基因组生物学》19 卷，15 期(2018 年)。

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于 2022 年 7 月 21 日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎 4.0 国际许可协议(CC - BY - NC - ND 4.0)提供。

# Supplementary Figures

# 补充图

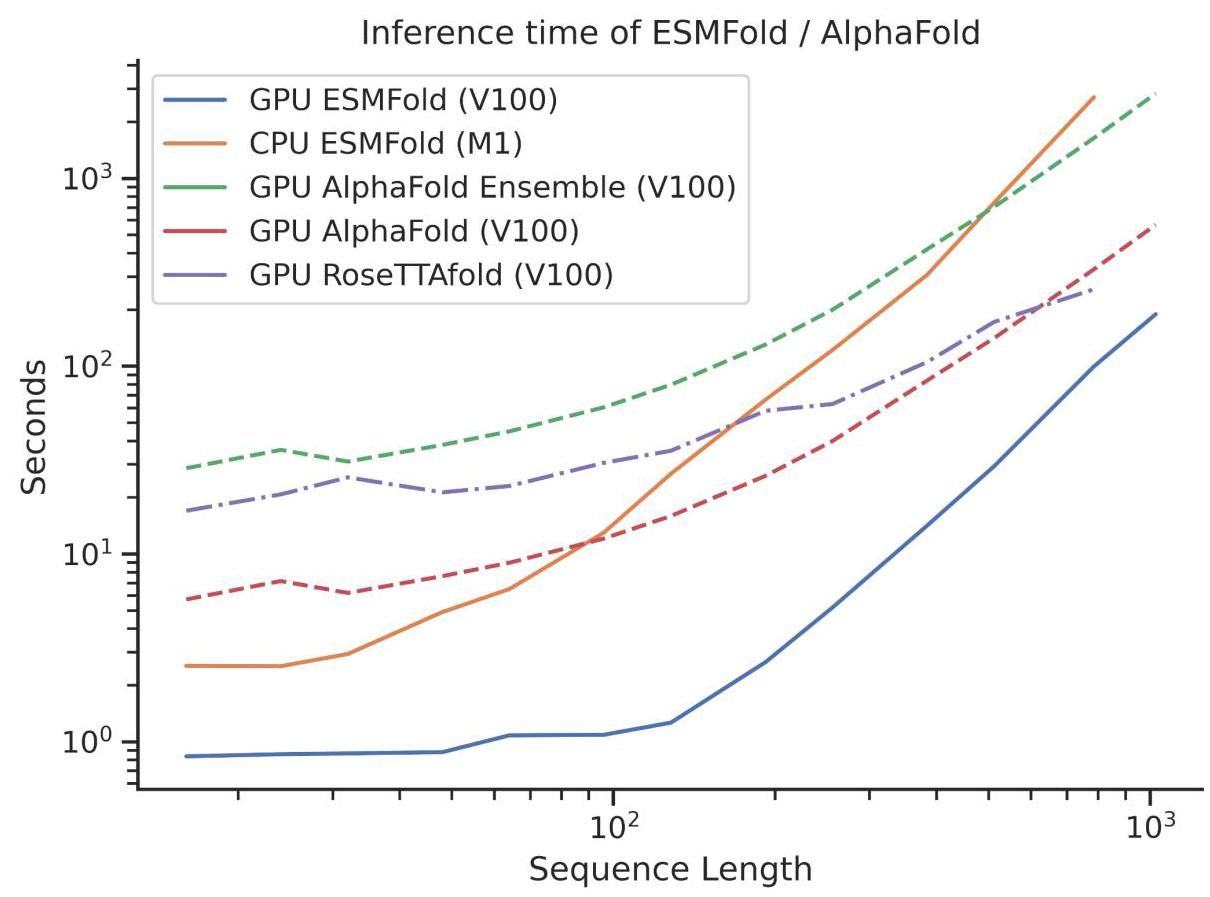


Figure S1: ESMFold vs AlphaFold2 and RoseTTAfold timing experiments

图 S1:ESMFold 与 AlphaFold2 和 RoseTTAfold 的计时实验

We test the speed of ESMFold vs AlphaFold2 on sequence lengths up to 1024. At low sequence lengths, ESMFold is dominated by language model performance, while the computation of pairwise representations takes over at high sequence lengths. Most of the speed advantage of ESMFold comes from not needing to process the MSA branch. We see an over speed advantage for shorter protein sequences, and a reasonable speed advantage for longer protein sequences. We also do not count Jax graph compilation times or MSA search times for AlphaFold2 - meaning in practice there is a larger performance advantage in the cold start case. We also use an optimized Colabfold 1.3.0 (53) to do speed comparison. No significant optimization has been performed on ESMFold, and we suspect that further gains can be made by optimizing ESMFold as well. For RoseTTAfold, the speed of the SE(3) Transformer dominates, especially at low sequence lengths. The number of SE(3) max-iterations are artificially limited to 20 (default 200) and no MSAs are used as input for these measurements. Additionally, this only measures the network forward time, and does not include the time taken to compute sidechains with PyRosetta or search for MSAs. These comparisons are much more favorable towards AlphaFold2 and RoseTTAfold. bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

我们测试了 ESMFold 与 AlphaFold2 在序列长度达 1024 时的速度。在低序列长度下，ESMFold 的速度主要受语言模型性能的限制，而在高序列长度下，成对表示的 计算起主导作用。ESMFold 的大部分速度优势来自于无需处理多序列比对(MSA)分支。我们发现，对于较短的蛋白质序列，ESMFold 有超过 的速度优势，对于较长的蛋白质序列也有相当的速度优势。我们也未将 AlphaFold2 的 Jax 图编译时间或多序列比对搜索时间计算在内，这意味着在实际冷启动情况下，ESMFold 有更大的性能优势。我们还使用优化后的 Colabfold 1.3.0(53)进行速度比较。ESMFold 尚未进行显著优化，我们推测对 ESMFold 进行优化也能进一步提升其性能。对于 RoseTTAfold，SE(3) 变换器的速度起主导作用，尤其是在低序列长度时。在这些测量中，SE(3) 的最大迭代次数被人为限制为 20 次(默认 200 次)，且未使用多序列比对作为输入。此外，这仅测量了网络前向传播时间，不包括使用 PyRosetta 计算侧链或搜索多序列比对的时间。这些比较对 AlphaFold2 和 RoseTTAfold 更为有利。生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于 2022 年 7 月 21 日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎 4.0 国际许可协议(CC - BY - NC - ND 4.0)提供。

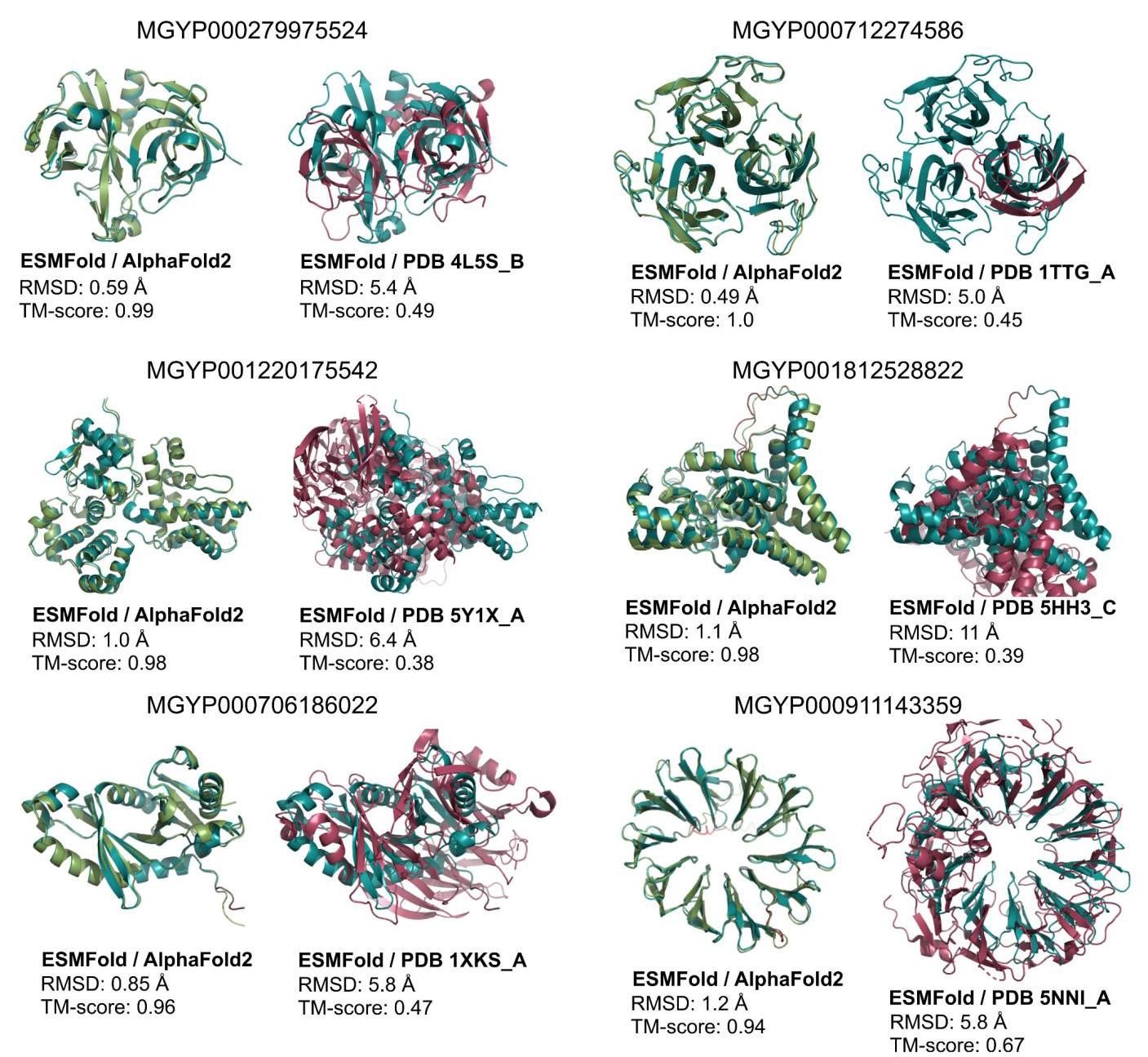


Figure S2: Highlighted ESMFold structure predictions, comparison to AlphaFold2, and comparison to closest PDB structure, related to Figure 3.

图 S2:ESMFold 结构预测亮点、与 AlphaFold2 的比较以及与最接近的蛋白质数据库(PDB)结构的比较，与图 3 相关。

Example predicted structures from six different metagenomic sequences; also see table S4. Left of each subfigure: The prediction is displayed with the AlphaFold2 prediction (light green). Right of each subfigure: The prediction is displayed with the Foldseek-determined nearest PDB structure according to TM-score.

来自六个不同宏基因组序列的预测结构示例；另见表 S4。每个子图左侧:预测结构与 AlphaFold2 的预测结果(浅绿色)一同展示。每个子图右侧:预测结构与根据模板匹配分数(TM - score)由 Foldseek 确定的最接近的蛋白质数据库结构一同展示。

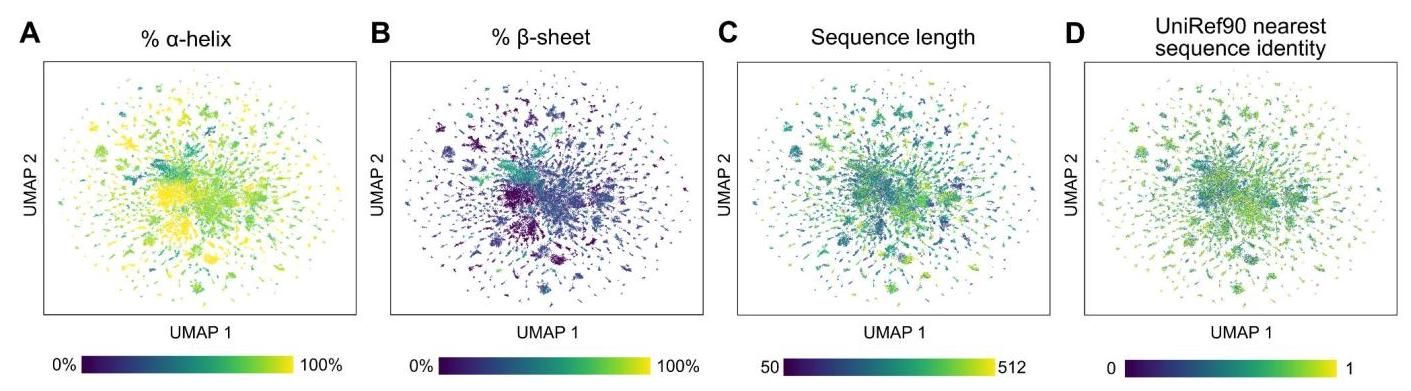


Figure S3: Sequence landscape UMAP visualizations, related to Figure 3.

图 S3:序列景观的均匀流形近似与投影(UMAP)可视化，与图 3 相关。

Additional UMAP plots in which MGnify sequences are plotted according to the same coordinates as in Figure 3C, but colored by secondary structure percentage (A, B), sequence length (C), or the sequence identity to the most similar entry in UniRef90 according to a blastp search (D).

额外的均匀流形近似与投影图，其中宏基因组序列数据库(MGnify)序列按照与图 3C 相同的坐标绘制，但根据二级结构百分比(A、B)、序列长度(C)或根据蛋白质基本局部比对搜索工具(blastp)搜索得到的与 UniRef90 中最相似条目的序列同一性(D)进行着色。

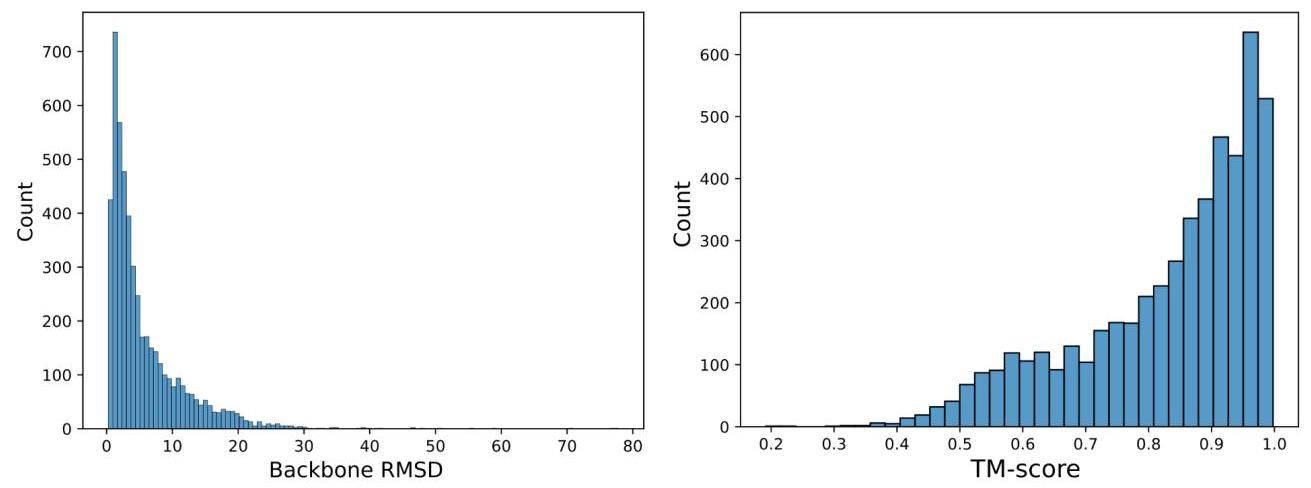


Figure S4: Comparison to AlphaFold2 of structurally remote ESMFold predictions, related to Figure 3.

图 S4:结构上与 AlphaFold2 差异较大的 ESMFold 预测结果的比较，与图 3 相关。

Distributions of backbone RMSDs (left) and TM-scores (right) of ESMFold-AlphaFold2 predictions of the same sequence, where the ESMFold prediction has both high confidence (mean pLDDT ) and relatively low structural similarity to the PDB (Foldseek closest PDB TM-score < 0.5). Supplementary Tables bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

相同序列的ESMFold - AlphaFold2预测结果的主链均方根偏差(RMSD，左图)和模板建模得分(TM - score，右图)分布，其中ESMFold预测结果具有高置信度(平均预测局部距离差异测试值[pLDDT ])，且与蛋白质数据库(PDB)的结构相似性相对较低(Foldseek最接近的PDB模板建模得分 < 0.5)。补充表格见生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。本预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。本预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议(CC - BY - NC - ND 4.0 International license)提供。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 8M | 35M | 150M | 650M | 3B | 15B |
| Dataset | UR50/D | UR50/D | UR50/D | UR50/D | UR50/D | UR50/D |
| Number of layers | 6 | 12 | 30 | 33 | 36 | 48 |
| Embedding dim | 320 | 480 | 640 | 1280 | 2560 | 5120 |
| Attention heads | 20 | 20 | 20 | 20 | 40 | 40 |
| Training steps | 500K | 500K | 500K | 500K | 500K | 270K |
| Learning rate | 4e-4 | 4e-4 | 4e-4 | 4e-4 | 4e-4 | 1.6e-4 |
| Weight decay | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.1 |
| Clip norm | 0 | 0 | 0 | 0 | 1.0 | 1.0 |
| Distributed backend | DDP | DDP | DDP | DDP | FSDP | FSDP |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 8M | 35M | 150M | 650M | 3B | 15B |
| 数据集 | UR50/D | UR50/D | UR50/D | UR50/D | UR50/D | UR50/D |
| 层数 | 6 | 12 | 30 | 33 | 36 | 48 |
| 嵌入维度 | 320 | 480 | 640 | 1280 | 2560 | 5120 |
| 注意力头数 | 20 | 20 | 20 | 20 | 40 | 40 |
| 训练步数 | 500K | 500K | 500K | 500K | 500K | 270K |
| 学习率 | 4e-4 | 4e-4 | 4e-4 | 4e-4 | 4e-4 | 1.6e-4 |
| 权重衰减 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.1 |
| 裁剪范数 | 0 | 0 | 0 | 0 | 1.0 | 1.0 |
| 分布式后端 | 分布式数据并行(Distributed Data Parallel) | 分布式数据并行(Distributed Data Parallel) | 分布式数据并行(Distributed Data Parallel) | 分布式数据并行(Distributed Data Parallel) | 完全分片数据并行(Fully Sharded Data Parallel) | 完全分片数据并行(Fully Sharded Data Parallel) |

Table S1: ESM-2 model parameters at different scales

表S1:不同尺度下的ESM - 2模型参数

|  |  |  |  |
| --- | --- | --- | --- |
|  | LRP@L | LR P@L/5 | Validation Perplexity |
| Baseline | 0.381 | 0.626 | 8.42 |
| No RoPE | 0.365 | 0.599 | 8.62 |
| Older UniRef Data | 0.368 | 0.599 | 7.98 |
| No UR90 Sampling | 0.387 | 0.631 | 8.40 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | LRP@L | LR P@L/5 | 验证困惑度 |
| 基线 | 0.381 | 0.626 | 8.42 |
| 无旋转位置编码(RoPE) | 0.365 | 0.599 | 8.62 |
| 旧版UniRef数据集 | 0.368 | 0.599 | 7.98 |
| 无UR90采样 | 0.387 | 0.631 | 8.40 |

Table S2: ESM-2 Architecture Ablations

表S2:ESM - 2架构消融实验

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

生物预印本服务器(bioRxiv)预印本，doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。本预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予生物预印本服务器永久展示该预印本的许可。本预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议(CC - BY - NC - ND 4.0 International license)提供。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | #Params | #Updates | Validation Perplexity | LRP@L |  | CASP14 | CAMEO |
| ESM-2 | 8M | 270k | 10.45 | 0.16 | 0.28 | 0.37 | 0.48 |
| 35M | 270k | 9.12 | 0.29 | 0.49 | 0.41 | 0.56 |
| 150M | 270k | 8.00 | 0.42 | 0.68 | 0.47 | 0.63 |
| 650M | 270k | 7.23 | 0.50 | 0.77 | 0.51 | 0.68 |
| 3B | 270k | 6.73 | 0.53 | 0.80 | 0.51 | 0.71 |
| 8M | 500k | 10.33 | 0.17 | 0.29 | 0.37 | 0.48 |
| 35M | 500k | 8.95 | 0.30 | 0.51 | 0.41 | 0.56 |
| 150M | 500k | 7.75 | 0.44 | 0.70 | 0.49 | 0.65 |
| 650M | 500k | 6.95 | 0.52 | 0.79 | 0.51 | 0.70 |
| 3B | 500k | 6.49 | 0.54 | 0.81 | 0.52 | 0.72 |
| 15B | 270k | 6.37 | 0.54 | 0.82 | 0.55 | 0.72 |
| ESM-1b3 | 650M | - | - | 0.41 | 0.66 | 0.42 | 0.64 |
| Prot-T5-XL (UR50) (19) | 3B | - | - | 0.48 | 0.72 | 0.50 | 0.69 |
| Prot-T5-XL (BFD) (19) | 3B | - | - | 0.36 | 0.58 | 0.46 | 0.63 |
| CARP (44) | 640M | - | - | - | - | 0.42 | 0.59 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 参数数量 | 更新次数 | 验证困惑度 | 长程精度@L(LRP@L) |  | 第14届蛋白质结构预测技术评估(CASP14) | 持续蛋白质结构预测评估(CAMEO) |
| 进化尺度模型2(ESM - 2) | 8M | 270k | 10.45 | 0.16 | 0.28 | 0.37 | 0.48 |
| 35M | 270k | 9.12 | 0.29 | 0.49 | 0.41 | 0.56 |
| 150M | 270k | 8.00 | 0.42 | 0.68 | 0.47 | 0.63 |
| 650M | 270k | 7.23 | 0.50 | 0.77 | 0.51 | 0.68 |
| 3B | 270k | 6.73 | 0.53 | 0.80 | 0.51 | 0.71 |
| 8M | 500k | 10.33 | 0.17 | 0.29 | 0.37 | 0.48 |
| 35M | 500k | 8.95 | 0.30 | 0.51 | 0.41 | 0.56 |
| 150M | 500k | 7.75 | 0.44 | 0.70 | 0.49 | 0.65 |
| 650M | 500k | 6.95 | 0.52 | 0.79 | 0.51 | 0.70 |
| 3B | 500k | 6.49 | 0.54 | 0.81 | 0.52 | 0.72 |
| 15B | 270k | 6.37 | 0.54 | 0.82 | 0.55 | 0.72 |
| 进化尺度模型1b3(ESM - 1b3) | 650M | - | - | 0.41 | 0.66 | 0.42 | 0.64 |
| 蛋白质T5大模型(通用表示50)(Prot - T5 - XL (UR50))(19) | 3B | - | - | 0.48 | 0.72 | 0.50 | 0.69 |
| 蛋白质T5大模型(大数据集)(Prot - T5 - XL (BFD))(19) | 3B | - | - | 0.36 | 0.58 | 0.46 | 0.63 |
| 对比自注意力表示预测模型(CARP)(44) | 640M | - | - | - | - | 0.42 | 0.59 |

# Table S3: Detailed language model comparison on structure prediction and unsupervised contact prediction.

# 表S3:语言模型在结构预测和无监督接触预测方面的详细比较。

Table S3 shows a comparison of ESM-2 language models with other language models on structure prediction and unsupervised contact prediction. ESM-2 language models are compared at different numbers of parameters and at different numbers of training updates. Training updates and validation perplexity are not reported for baseline models, since there is no straightforward comparison. For the number of training updates, different models use different batch sizes, so the number of sequences seen can vary even if the number of updates are the same. For validation perplexity, baseline models are not trained on the same dataset, and do not share a common heldout validation set with ESM-2. Unsupervised contact precision results, in the form of long range precision at and at , do allow us to compare all transformer language models despite variance in training data. However, CARP, a convolution based language model, does not have attention maps with which to identify protein contacts. Despite this, supervised training manages to extract reasonable contact prediction results approximately on par with ESM-1b.

表S3展示了ESM - 2语言模型与其他语言模型在结构预测和无监督接触预测方面的比较。我们在不同的参数数量和不同的训练更新次数下对ESM - 2语言模型进行了比较。由于无法直接进行比较，因此未报告基线模型的训练更新次数和验证困惑度。对于训练更新次数，不同的模型使用不同的批量大小，因此即使更新次数相同，所看到的序列数量也可能不同。对于验证困惑度，基线模型并非在相同的数据集上进行训练，并且与ESM - 2没有共同的保留验证集。无监督接触精度结果，以 和 处的长程精度形式呈现，尽管训练数据存在差异，但仍使我们能够比较所有的Transformer语言模型。然而，基于卷积的语言模型CARP没有用于识别蛋白质接触的注意力图。尽管如此，有监督训练仍能提取出与ESM - 1b大致相当的合理接触预测结果。

[[3]](#footnote-65)

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv预印本doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予bioRxiv永久展示该预印本的许可。该预印本根据CC - BY - NC - ND 4.0国际许可协议提供。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MGnify ID | pLDDT | Foldseek server closest TM-score | Foldseek server closest | Closest blastp sequence identity (UniRef90) | Closest blastp sequence (UniRef90) |
| MGYP000712274586 | 0.96 | 0.45 | 1ttg\_A | 54% | UniRef90 A0A539E457 Uncharacterized protein (Acidimicrobiaceae bacterium) |
| MGYP000911143359 | 0.90 | 0.67 | 5nni\_A | 43% | UniRef90 A0A7Y5V7P8 Uncharacterized protein (Flavobacteriales bacterium) |
| MGYP001220175542 | 0.94 | 0.38 | 5ylx\_A | 98% | UniRef90 UPI0013011942 Helix-turn-helix domain-containing protein (Caenibacillus caldisaponilyticus) |
| MGYP001812528822 | 0.93 | 0.39 | 5hh3\_C | 50% | UniRef90 A0A545U581 Fatty acid desaturase (Exilibacterium tricleocarpae) |
| MGYP000706186022 | 0.92 | 0.47 | 1xks\_A | 29% | UniRef90 A0A6N6S1Z1 Uncharacterized protein (Candidatus brocadia) |
| MGYP000279975524 | 0.93 | 0.49 | 415s B | 38% | UniRef90 A0A1F4EWL6 Uncharacterized protein (Betaproteobacteria bacterium) |
| MGYP004000959047 | 0.90 | 0.80 | 6bym\_A | No significant matches | NA |
| MGYP000936678158 | 0.95 | 0.68 | 5yet\_B | No significant matches | NA |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MGnify标识符 | 预测局部距离差异测试值(pLDDT) | Foldseek服务器最近邻模板匹配分数(TM - score) | Foldseek服务器最近邻 | 最近邻蛋白基本局部比对搜索工具(blastp)序列一致性(UniRef90) | 最近邻蛋白基本局部比对搜索工具(blastp)序列(UniRef90) |
| MGYP000712274586 | 0.96 | 0.45 | 1ttg\_A | 54% | UniRef90 A0A539E457 未表征蛋白(酸微菌科细菌) |
| MGYP000911143359 | 0.90 | 0.67 | 5nni\_A | 43% | UniRef90 A0A7Y5V7P8 未表征蛋白(黄杆菌目细菌) |
| MGYP001220175542 | 0.94 | 0.38 | 5ylx\_A | 98% | UniRef90 UPI0013011942 含螺旋 - 转角 - 螺旋结构域蛋白(热解皂凯尼芽孢杆菌) |
| MGYP001812528822 | 0.93 | 0.39 | 5hh3\_C | 50% | UniRef90 A0A545U581 脂肪酸去饱和酶(三果外杆菌) |
| MGYP000706186022 | 0.92 | 0.47 | 1xks\_A | 29% | UniRef90 A0A6N6S1Z1 未表征蛋白(暂定布罗卡德氏菌) |
| MGYP000279975524 | 0.93 | 0.49 | 415s B | 38% | UniRef90 A0A1F4EWL6 未表征蛋白(β - 变形菌纲细菌) |
| MGYP004000959047 | 0.90 | 0.80 | 6bym\_A | 无显著匹配 | 不适用 |
| MGYP000936678158 | 0.95 | 0.68 | 5yet\_B | 无显著匹配 | 不适用 |

# Table S4: Information on highlighted MGnify proteins, related to Figure 3.

# 表S4:与图3相关的高亮MGnify蛋白质信息。

MGnify sequence identifiers corresponding to predicted structures highlighted throughout this study, including the PDB chain and corresponding TM-score of the closest structure identified by the Foldseek webserver as well as the UniRef90 entry and sequence identity of the closest sequence identified by blastp (Methods).

本研究中高亮显示的预测结构对应的MGnify序列标识符，包括PDB链、通过Foldseek网络服务器识别出的最接近结构的对应TM分数，以及通过blastp识别出的最接近序列的UniRef90条目和序列同一性(方法部分)。

# Methods

# 方法

# 1.1 Data

# 1.1 数据

# 1.1.1 ESM

# 1.1.1 ESM

UniRef50, April 2021 version, is used for the training of ESM models. We partition the training dataset by randomly selecting sequences to form our validation set. The training set has sequences removed via the procedure described in (54). MMseqs search (-min-seq-id 0.5 -alignment-mode 3 -max-seqs -cov-mode 0 ) is run using the train set as query database and the validation set as target database. All train sequences which match a validation sequence with 50% sequence identity under this search are removed from the train set.

2021年4月版的UniRef50用于ESM模型的训练。我们通过随机选择 条序列来划分训练数据集，以形成我们的验证集。训练集按照文献(54)中描述的程序去除了部分序列。使用训练集作为查询数据库，验证集作为目标数据库运行MMseqs搜索(-min-seq-id 0.5 -alignment-mode 3 -max-seqs -cov-mode 0)。在该搜索下，所有与验证序列具有50%序列同一性的训练序列都从训练集中移除。

We also filter out de-novo designed proteins from the pretraining dataset via two filters. First, we remove any sequence in UniRef50 and UniRef90 that was annotated as "artificial sequence" by a taxonomy search on the UniProt website, when 2021\_04 was the most recent release (1,027 proteins). Second, we use jackhmmer to remove all hits around a manually curated set of 81 de-novo proteins. jackhmmer was run with ‘–num-iter 1 -max‘ flags, with each of the 81 de-novo proteins as a query and UniRef100 as a search database. All proteins returned by jackhmmer were removed from both UniRef50 and UniRef90 via their UniRef IDs (58,462 proteins). This filtering is performed to enable future work evaluating the generalization of language models to de-novo sequences.

我们还通过两个过滤器从预训练数据集中过滤掉从头设计的蛋白质。首先，我们移除UniRef50和UniRef90中所有在UniProt网站上通过分类学搜索被注释为“人工序列”的序列(2021\_04是最新版本时，共1027个蛋白质)。其次，我们使用jackhmmer移除围绕一组手动筛选的81个从头设计蛋白质的所有匹配项。jackhmmer使用‘–num-iter 1 -max‘标志运行，以81个从头设计蛋白质中的每一个作为查询，UniRef100作为搜索数据库。jackhmmer返回的所有蛋白质都通过其UniRef ID从UniRef50和UniRef90中移除(共58462个蛋白质)。进行这种过滤是为了便于未来评估语言模型对从头设计序列的泛化能力的研究。

To increase the amount of data and its diversity, we sampled a minibatch of UniRef50 sequences for each training update. We then replaced each sequence with a sequence sampled uniformly from the corresponding UniRef90 cluster. This allowed ESM-2 models to train on over 60M protein sequences.

为了增加数据量及其多样性，我们为每次训练更新采样一个UniRef50序列的小批量。然后，我们用从相应UniRef90簇中均匀采样的序列替换每个序列。这使得ESM - 2模型能够在超过6000万个蛋白质序列上进行训练。

# 1.1.2 Structure Training Sets

# 1.1.2 结构训练集

For training ESMFold, we closely follow the training procedure outlined in (26). We find all PDB chains until 2020-05-01 with resolution greater than or equal to and length greater than 20 . All proteins where over 20% of the sequence is the same residue is not considered. We use MMseqs easy-cluster with default parameters to cluster resulting sequences at sequence identity. Only individual chains are used during training, even when the chain is part of a protein complex.

为了训练ESMFold，我们严格遵循文献(26)中概述的训练程序。我们找出截至2020年5月1日所有分辨率大于或等于 且长度大于20的PDB链。不考虑序列中超过20%为同一残基的所有蛋白质。我们使用默认参数的MMseqs easy - cluster以 的序列同一性对所得序列进行聚类。训练期间仅使用单个链，即使该链是蛋白质复合物的一部分。

At training time, we sample each cluster evenly, and then sample a random protein from each cluster. We also do the same rejection sampling technique, where protein chains are accepted with probability . This means longer proteins are trained on more frequently.

在训练时，我们对每个簇进行均匀采样，然后从每个簇中随机采样一个蛋白质。我们还采用相同的拒绝采样技术，即蛋白质链以概率 被接受。这意味着更长的蛋白质更频繁地被用于训练。

The predicted structures dataset is the same generated from Hsu et al. 2022 (48). This dataset is a set of 13,477,259 structures predicted using AlphaFold2 on MSAs generated via the process in (55). We then filter this dataset by predicted IDDT greater than 70 . Because of the way the dataset is constructed, we lose only 1.5% of the dataset with this filter. Additionally, during backpropagation, we do not backpropagate residues where predicted IDDT is less than 70 . We found that this is necessary to obtain increased performance using predicted structures. We sample from predicted structures 75% of the time, and real structures 25% of the time during training.

预测结构数据集与Hsu等人2022年(48)生成的数据集相同。该数据集是一组通过文献(55)中的过程生成的多序列比对(MSAs)，使用AlphaFold2预测的13477259个结构。然后，我们通过预测的IDDT大于70来过滤该数据集。由于数据集的构建方式，使用此过滤器我们仅损失了1.5%的数据集。此外，在反向传播期间，我们不对预测IDDT小于70的残基进行反向传播。我们发现这对于使用预测结构获得更高性能是必要的。训练期间，我们75%的时间从预测结构中采样，25%的时间从真实结构中采样。

# 1.1.3 Structure Validation and Test Sets

# 1.1.3 结构验证集和测试集

During method development (e.g. hyperparameter selection), we used a temporally held out validation set obtained from the Continuous Automated Model EvaluatiOn (CAMEO) server (41) by filtering from August 2021 to January 2022.

在方法开发期间(例如超参数选择)，我们使用了一个从连续自动模型评估(CAMEO)服务器(41)获得的时间上保留的验证集，该验证集是通过筛选2021年8月至2022年1月的数据得到的。

We report results by testing 3D structure prediction models on two test sets, both chosen to be temporally held out from our supervised training set. The first test is from CAMEO, consisting of all 194 test proteins from April 01, 2022 through June 25, 2022. Our second test set consists of 51 targets from the CASP14 competition (42). For both test sets, metrics are computed on all modeled residues in the PDB file. The full CASP14 target list is:

我们通过在两个测试集上测试三维结构预测模型来报告结果，这两个测试集在时间上均与我们的有监督训练集相互独立。第一个测试集来自CAMEO，包含从2022年4月1日到2022年6月25日的所有194个测试蛋白质。我们的第二个测试集由来自CASP14竞赛(42)的51个目标组成。对于这两个测试集，指标是在PDB文件中所有建模的残基上计算的。完整的CASP14目标列表如下:

49, T1050, T1053, T1054, T1055, T1056, T1057, T1058, T1064, T1065s1, T1065s2, T1067, T1070, T1073, T1073, T107 4, T1076, T1078, T1079, T1080, T1082, T1089, T1090, T1091, T1099.

49, T1050, T1053, T1054, T1055, T1056, T1057, T1058, T1064, T1065s1, T1065s2, T1067, T1070, T1073, T1073, T107 4, T1076, T1078, T1079, T1080, T1082, T1089, T1090, T1091, T1099.

These are all publicly available CASP14 targets as of July 2022.

截至2022年7月，这些都是公开可用的CASP14目标。

No filtering is performed on these test sets, as ESMFold is able to make predictions on all sequences, including the length-2166 target T1044.

这些测试集未进行任何过滤，因为ESMFold能够对所有序列进行预测，包括长度为2166的目标T1044。

# 1.2 Language Models

# 1.2 语言模型

# 1.2.1 Unsupervised Contact Prediction

# 1.2.1 无监督接触预测

The unsupervised contact prediction methodology used throughout this work is taken from Rao et al. 2021 (38). They show that transformer language models trained on large databases of protein sequences learn to predict protein contacts in their attention maps using little to no supervision. We exploit this fact as a very computationally inexpensive method for measuring a language model’s knowledge of protein structure.

本文通篇使用的无监督接触预测方法取自Rao等人2021年的研究(38)。他们表明，在大型蛋白质序列数据库上训练的Transformer语言模型能够在很少或没有监督的情况下，在其注意力图中学习预测蛋白质接触。我们利用这一事实，将其作为一种计算成本极低的方法来衡量语言模型对蛋白质结构的了解程度。

Rather than training an atomic-level structure predictor for each checkpoint of each language model (a process which can take several days for the largest language models), we use the logistic regression described in Rao et al. 2021 for contact prediction. The probability of a contact is defined as:

我们没有为每个语言模型的每个检查点训练一个原子级结构预测器(对于最大的语言模型，这个过程可能需要几天时间)，而是使用Rao等人2021年描述的逻辑回归进行接触预测。接触的概率定义如下:

Let be a boolean random variable which is true if amino acids are in contact. Suppose our transformer has layers and attention heads per layer. Let be the symmetrized and APC-corrected (56) attention map for the -th attention head in the -th layer of the transformer, and be the value of that attention map at position . Then

设 为一个布尔随机变量，如果氨基酸 处于接触状态，则该变量为真。假设我们的Transformer有 层，每层有 个注意力头。设 为Transformer第 层中第 个注意力头的对称化且经过APC校正(56)的注意力图， 为该注意力图在位置 处的值。那么

The parameters are fit in scikit-learn (57) using L1-regularized logistic regression with . The regression is fit using the 20 protein training set from Rao et al. 2021 (38), which was simply a random selection from the trRosetta (58) training set. We performed a variability analysis using 20 bootstrapped samples of 20 training proteins from the total set of 14862 proteins. The average long range P@L was 0.4287 with a standard deviation of 0.0028 . We also performed experiments using larger training sets, but observed no significant performance change. Given these results, we are confident that selecting a subset of 20 proteins for training provides a good estimate of contact precision performance.

参数 使用scikit - learn(57)中的L1正则化逻辑回归进行拟合，其中 。回归使用Rao等人2021年研究(38)中的20个蛋白质训练集进行拟合，该训练集只是从trRosetta(58)训练集中随机选取的。我们使用从14862个蛋白质的总集中抽取的20个自举样本(每个样本包含20个训练蛋白质)进行了变异性分析。平均长程P@L为0.4287，标准差为0.0028。我们还使用更大的训练集进行了实验，但未观察到显著的性能变化。鉴于这些结果，我们有信心选择20个蛋白质的子集进行训练能够很好地估计接触精度性能。

Unsupervised contact prediction results are reported for the 14842 protein test set used in Rao et al. 2021, which is also derived from the trRosetta training set. For both training and test a contact is defined as two amino acids with -alpha distance .

无监督接触预测结果是针对Rao等人2021年使用的14842个蛋白质测试集报告的，该测试集也源自trRosetta训练集。对于训练和测试，接触定义为两个氨基酸的 - 阿尔法距离为 。

# 1.2.2 Perplexity Calculation

# 1.2.2 困惑度计算

Perplexity is a measure of a language model’s uncertainty of a sequence and is defined as the exponential of the negative log-likelihood of the sequence. Unfortunately, there is no efficient method of computing the log-likelihood of a sequence under a masked language model. Instead, there are two methods we can use for estimating perplexity.

困惑度是衡量语言模型对序列不确定性的指标，定义为序列负对数似然的指数。不幸的是，在掩码语言模型下没有有效的方法来计算序列的对数似然。相反，我们可以使用两种方法来估计困惑度。

First, let the mask be a random variable denoting a set of tokens from input sequence . Each token has a 15% probability of inclusion. If included the tokens have an 80% probability of being replaced with a mask token, a 10% probability of being replaced with a random token, and a 10% probability of being replaced with an unmasked token. Let denote the set of modified input tokens. The perplexity is then defined as

首先，设掩码 为一个随机变量，表示来自输入序列 的一组标记。每个标记有 15% 的概率被包含在内。若被包含，这些标记有 80% 的概率被替换为掩码标记，有 10% 的概率被替换为随机标记，还有 10% 的概率被替换为未掩码标记。设 表示修改后的输入标记集。困惑度定义如下

As the set is a random variable, this expression is non-deterministic. This makes it a poor estimate of the perplexity of a single sequence. However, it requires only a single forward pass of the model to compute, so it is possible to efficiently obtain an estimate of the expectation of this expression over a large dataset. When reporting the perplexity over a large dataset (such as our UniRef validation set), this estimate is used.

由于集合 是一个随机变量，这个表达式是不确定的。这使得它对单个序列的困惑度估计效果不佳。然而，计算它只需要模型进行一次前向传播，因此可以在大型数据集上高效地获得该表达式期望值的估计。在报告大型数据集(如我们的 UniRef 验证集)的困惑度时，使用此估计值。

The second perplexity calculation is the pseudo-perplexity, which is the exponential of the negative pseudo-log-likelihood of a sequence. This estimate provides a deterministic value for each sequence, but requires forward passes to compute, where is the length of the input sequence. It is defined as

第二种困惑度计算方法是伪困惑度，它是序列负伪对数似然的指数。此估计为每个序列提供一个确定的值，但需要 次前向传播来计算，其中 是输入序列的长度。其定义如下

When reporting the perplexity for an individual sequence (e.g. on CASP14 or CAMEO), this estimate is used. For brevity, we refer to both of these estimates as the "perplexity," as they can be interpreted in a similar manner. bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which

在报告单个序列的困惑度时(例如在 CASP14 或 CAMEO 上)，使用此估计值。为简洁起见，我们将这两种估计都称为“困惑度”，因为它们可以以类似的方式进行解释。生物预印本 doi: https://doi.org/10.1101/2022.07.20.500902；此版本于 2022 年 7 月 21 日发布。此预印本的版权所有者(

# 1.2.3 ESM-2 Architecture

# 1.2.3 ESM - 2 架构

We use a BERT (34) style encoder only transformer architecture (47) with modifications. We change the number of layers, number of attention heads, hidden size and feed forward hidden size as we scale the ESM model (table S1).

我们使用经过修改的 BERT(34)风格的仅编码器的Transformer架构(47)。在扩展 ESM 模型时，我们改变了层数、注意力头的数量、隐藏层大小和前馈隐藏层大小(表 S1)。

The original transformer paper uses absolute sinusoidal positional encoding to inform the model about token positions. These positional encodings are added to the input embeddings at the bottom of the encoder stack. In ESM-1b (15), we replaced this static sinusoidal encoding with a learned one. Both static and learned absolute encodings provide the model a very cheap way of adding positional information. However, absolute positional encoding methods don’t extrapolate well beyond the context window they are trained on. In ESM-2, we used Rotary Position Embedding (RoPE) (59) to allow the model extrapolate beyond the context window it is trained on. RoPE slightly increases the computational cost of the model, since it multiplies every query and key vector inside the self attention with a sinusoidal embedding. In our experiments, we observed that this improves model quality for small models. However, we observed that the performance improvements start to disappear as the model size and training duration get bigger.

原始的Transformer论文使用绝对正弦位置编码来告知模型标记的位置。这些位置编码被添加到编码器堆栈底部的输入嵌入中。在 ESM - 1b(15)中，我们用学习得到的位置编码替换了这种静态正弦编码。静态和学习得到的绝对编码都为模型提供了一种非常简便的添加位置信息的方法。然而，绝对位置编码方法在超出其训练的上下文窗口时泛化能力不佳。在 ESM - 2 中，我们使用旋转位置嵌入(RoPE)(59)，使模型能够在超出其训练的上下文窗口时进行泛化。RoPE 会略微增加模型的计算成本，因为它会将自注意力内的每个查询向量和键向量与正弦嵌入相乘。在我们的实验中，我们观察到这对小型模型的质量有提升。然而，我们也观察到，随着模型规模和训练时长的增加，性能提升开始消失。

# 1.2.4 Training Details

# 1.2.4 训练细节

In ESM-2, we have made multiple small modifications to our model with the goal of increasing the effective capacity of our models. ESM-1b had dropout both in hidden layers and attention which we removed completely to free up more capacity. In our experiments, we did not observe any significant performance regressions with this change.

在 ESM - 2 中，我们对模型进行了多项小修改，目的是提高模型的有效容量。ESM - 1b 在隐藏层和注意力层都有丢弃层(dropout)，我们将其完全移除，以释放更多容量。在我们的实验中，我们没有观察到这种改变导致任何显著的性能下降。

We trained most of our models on a network with multiple nodes connected via a network interface. As the models get bigger, the amount of communication becomes the fundamental bottleneck for the training speed. Since BERT style models have been shown to be amenable to very large batch sizes (60), we increased our effective batch size to tokens.

我们在一个通过网络接口连接多个节点的网络上训练了我们的大部分模型。随着模型规模的增大，通信量成为训练速度的基本瓶颈。由于 BERT 风格的模型已被证明适用于非常大的批量大小(60)，我们将有效批量大小增加到 个标记。

For model training optimization, we used Adam with and weight decay of 0.01 for all models except the 15 billion parameter model, where we used a weight decay of 0.1 . The learning rate is warmed up over the first 2,000 steps to a peak value of 4e-4 (1.6e-4 for the 15B parameter model), and then linearly decayed to one tenth of its peak value over the 90% of training duration. We trained all models for updates except the model which we trained for steps. All models used 2 million tokens as batch size except the model where we used 3.2 million tokens batch size. In order to efficiently process large proteins, we cropped long proteins to random 1024 tokens. We used BOS and EOS tokens to signal the beginning and end of a real protein, to allow the model to separate a full sized protein from a cropped one.

为了优化模型训练，除了150亿参数模型使用0.1的权重衰减外，我们对所有模型都使用了Adam优化器，其中 和 的权重衰减为0.01。学习率在前2000步内预热到峰值4e - 4(150亿参数模型为1.6e - 4)，然后在90%的训练时长内线性衰减到峰值的十分之一。除了 模型训练了 步外，我们对所有模型都进行了 次更新。除了 模型使用320万个标记作为批量大小外，所有模型都使用200万个标记作为批量大小。为了高效处理大蛋白，我们将长蛋白裁剪为随机的1024个标记。我们使用起始标记(BOS)和结束标记(EOS)来表示真实蛋白的开始和结束，以便模型能够区分完整大小的蛋白和裁剪后的蛋白。

We used standard distributed data parallelism for models up to parameters and used sharded data parallelism (FSDP) (61) for the 2.8B and 15B parameter models. FSDP shards model weights and optimization parameters across multiple GPUs, allowing us to train models that can’t fit into a single GPU memory.

对于参数最多为 的模型，我们使用标准的分布式数据并行；对于28亿和150亿参数的模型，我们使用分片数据并行(FSDP)(61)。FSDP在多个GPU之间分片存储模型权重和优化参数，使我们能够训练无法放入单个GPU内存的模型。

# 1.2.5 ESM-2 Ablation Details

# 1.2.5 ESM - 2消融实验细节

We ran ablation experiments using parameter models trained for steps. Ablations were performed for RoPE, new data version, and UniRef90 sampling (table S2).

我们使用经过 步训练的 参数模型进行了消融实验。对旋转位置编码(RoPE)、新数据版本和UniRef90采样进行了消融实验(表S2)。

Unsupervised contact prediction results show that both RoPE and newer data significantly improve the results. We do observe a slight regression when sampling from UniRef90 clusters, however we believe this difference is small and the UniRef90 cluster sampling is likely to help for the larger models.

无监督接触预测结果表明，旋转位置编码(RoPE)和较新的数据都显著改善了结果。然而，当从UniRef90聚类中采样时，我们确实观察到了轻微的性能下降，但我们认为这种差异很小，并且UniRef90聚类采样可能对更大的模型有帮助。

# 1.3 ESMFold

# 1.3 ESMFold

The AlphaFold2 architecture is split into two major sections, the Evoformer and the structure module. The structure module processes the final representations into 3D coordinates for atomic-level structure predictions and requires no changes to be used with ESM-2. The Evoformer, however, builds separate MSA and residue-pairwise embedding spaces.

AlphaFold2架构分为两个主要部分:进化模块(Evoformer)和结构模块。结构模块将最终表示处理为原子级结构预测的三维坐标，并且在与ESMFold - 2一起使用时无需更改。然而，进化模块构建了单独的多序列比对(MSA)和残基对嵌入空间。

The major change that needs to be made in order to adapt the Evoformer block to language model features is to remove its dependence on MSAs. Since MSAs are two dimensional, the Evoformer employs axial attention (62) over the columns and rows of the MSA. The language model features are one dimensional, so we can replace the axial attention with a standard attention over this feature space. All other operations in the Evoformer block are kept the same. We call this simplified architecture the Folding block.

为了使进化模块块适应语言模型特征，需要进行的主要更改是消除其对多序列比对(MSA)的依赖。由于多序列比对是二维的，进化模块在多序列比对的列和行上采用轴向注意力机制(62)。语言模型特征是一维的，因此我们可以用对该特征空间的标准注意力机制取代轴向注意力机制。进化模块块中的所有其他操作保持不变。我们将这种简化的架构称为折叠块。

The second change involves the removal of templates. Template information is passed to the model as pairwise distances, input to the residue-pairwise embedding. We simply omit this information, passing instead the attention maps from the language model, as these have been shown to capture structural information well (38).

第二个更改涉及去除模板。模板信息以成对距离的形式传递给模型，并作为残基对嵌入的输入。我们简单地省略了此信息，而是传递语言模型的注意力图，因为这些注意力图已被证明能够很好地捕捉结构信息(38)。

Our final architecture, which we call ESMFold, has 48 folding blocks. It was trained for an initial 125k steps on protein crops of size 256, and then fine-tuned with the structural violation loss for steps, on crop sizes of 384. We use the Frame Aligned Point Error (FAPE) and distogram losses introduced in AlphaFold2, as well as heads for predicting IDDT and the pTM score. We omit the masked language modeling loss. Language model parameters are frozen for training ESMFold.

我们最终的架构称为ESMFold，有48个折叠块。它首先在大小为256的蛋白裁剪片段上进行了125000步的训练，然后在大小为384的裁剪片段上使用结构违规损失进行了 步的微调。我们使用了AlphaFold2中引入的帧对齐点误差(FAPE)和距离图损失，以及用于预测交互域距离差(IDDT)和预测模板建模得分(pTM)的头部。我们省略了掩码语言建模损失。在训练ESMFold时，语言模型的参数被冻结。

The folding block is as follows, and shown in Figure 2A:

折叠块如下所示，并在图2A中展示:

Algorithm 1:

FoldingBlock(s, z)

b = Linear(z)

s = s + MultiHeadSelfAttention (s, bias=b)

s = s + MLP(s)

z = z + Linear(Concat([OuterProduct(s), OuterDifference(s)]))

z = z + TriangularMultiplicativeUpdateOutgoing(z)

z = z + TriangularMultiplicativeUpdateIncoming(z)

z = z + TriangularSelfAttentionOutgoing(z)

z = z + TriangularSelfAttentionIncoming(z)

return s, z

The folding block is extremely similar to the Evoformer described in AlphaFold2. There are two major differences. Because we do not have an MSA, the modules responsible for processing the MSA are replaced with a simple transformer. The self-attention here still uses a bias derived from the pairwise representations. Secondly, the sequence representations communicate with pairwise representation via both an outer product and outer difference.

折叠块与AlphaFold2中描述的进化模块极为相似。有两个主要区别。因为我们没有多序列比对(MSA)，负责处理多序列比对的模块被一个简单的变压器取代。这里的自注意力机制仍然使用从成对表示中导出的偏差。其次，序列表示通过外积和外差与成对表示进行通信。

And ESMFold is as follows:

ESMFold如下所示:

Algorithm 2 :

esm\_c\_s: number of channels in ESM hidden representation

c\_s = 1024

c\_z = 128

ESMFold (sequence)

s = ESM\_hiddens(sequence) # num\_layers x Length x esm\_c\_s

s = (softmax(layer\_weights) \* s).sum(0)

s = MLP(s)

z = PairwiseRelativePositionalEncoding(Length)

for b in folding\_blocks:

return StructureModule ( s, z )

We use a learned weighted sum of ESM embeddings to produce the initial hidden state into the model. This is then fed through an MLP. The initial pairwise state is simply the pairwise relative positional encoding described in (26). We found that using the attention maps initially gives a boost in performance, but this disappears during training. For experiments that do not use any folding blocks, we use an MLP applied to the ESM attention maps as input, and add the pairwise relative positional encoding to the attention map scores. Finally, the StructureModule parses these results into coordinates.

我们使用学习得到的进化尺度模型(ESM)嵌入的加权和来生成模型的初始隐藏状态。然后将其输入多层感知器(MLP)。初始成对状态就是(26)中描述的成对相对位置编码。我们发现，最初使用注意力图可以提升性能，但在训练过程中这种提升会消失。对于不使用任何折叠模块的实验，我们将应用于ESM注意力图的MLP作为输入，并将成对相对位置编码添加到注意力图得分中。最后，结构模块(StructureModule)将这些结果解析为坐标。

The predicted IDDT head is output from the hidden representation of the StructureModule. The predicted TM head uses the pairwise representation z. Finally, we also predict the distogram, from the same representation.

预测的全原子距离离散化得分(IDDT)头从结构模块的隐藏表示中输出。预测的模板建模得分(TM)头使用成对表示z。最后，我们还从相同的表示中预测距离直方图。

To predict complexes shown in Figure 2D, we give a residue index break of 1000 to ESMFold and link chains with a 25-residue poly-glycine linker, which we remove before displaying. Note that this is using ESMFold out of distribution since single chains are used during training.

为了预测图2D中所示的复合物，我们为进化尺度模型折叠(ESMFold)提供一个残基索引断点1000，并使用一个25个残基的聚甘氨酸接头连接各链，在展示前我们会移除该接头。请注意，这是在分布外使用ESMFold，因为训练期间使用的是单链。

# 1.4 Metagenomics experiments

# 1.4 宏基因组学实验

MGnify (29) version 2022 was clustered at 50% sequence similarity using parameters corresponding to high sensitivity ( –min-seq-id 0.5 –kmer-per-seq 100 –cluster-mode 2 –cov-mode 1 -c 0.6 ). Of these cluster representatives, we filtered out any sequences with less than 50 residues or greater than 512 residues, from which we obtained a uniformly sampled set of 1 million sequences. We then used ESMFold to obtain structure predictions and corresponding pLDDT values for each of these sequences. bioRxiv preprint doi: https://doi.org/10.1101/2022.07.20.500902; this version posted July 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

使用对应于高灵敏度的参数(–min - seq - id 0.5 –kmer - per - seq 100 –cluster - mode 2 –cov - mode 1 - c 0.6)，将MGnify(29)2022版以50%的序列相似度进行聚类。在这些聚类代表中，我们过滤掉所有少于50个残基或多于512个残基的序列，从中获得了一组均匀采样的100万条序列。然后，我们使用进化尺度模型折叠(ESMFold)为这些序列中的每一条获得结构预测和相应的预测局部距离差异测试(pLDDT)值。生物预印本服务器(bioRxiv)预印本doi: https://doi.org/10.1101/2022.07.20.500902；此版本于2022年7月21日发布。此预印本的版权持有者(未经同行评审认证)是作者/资助者，其已授予bioRxiv永久展示该预印本的许可。该预印本根据知识共享署名 - 非商业性使用 - 禁止演绎4.0国际许可协议(CC - BY - NC - ND 4.0)提供。

We further analyzed the high-confidence subset of the 1 million structures with mean pLDDT greater than 0.9, corresponding to structures. For each high-confidence structure, we used Foldseek (52) easy-search (–alignment-type 1) to identify similar structures in the PDB (as of April 12, 2022) based on TM-score, with a cutoff of 0.5 (lower than which Foldseek would return no structures) to enable more efficient search. For high-confidence structures that also have no structures with TM-score greater than 0.5 returned by Foldseek, we used full AlphaFold2 with MSAs to also obtain structure predictions (we picked the top of five relaxed models ranked by mean pLDDT). We then compute RMSD values of aligned backbone coordinates and all-atom TM-score between the ESMFold- and AlphaFold2-predicted structures. For each sequence corresponding to a high-confidence structure, we also used blastp version 2.10.0+ to search for similar sequences in UniRef90 to compute sequence identity. We defined sequences as having low sequence identity to UniRef90 if all of the returned hits have an E-value greater than 1 or if the closest entry in UniRef90 has sequence identity of less than 30%. For sequences with no significant matches in UniRef90, we also used the jackhmmer web server

我们进一步分析了100万个结构中平均预测局部距离差异测试(pLDDT)大于0.9的高置信度子集，对应于 个结构。对于每个高置信度结构，我们使用Foldseek(52)的easy - search(–alignment - type 1)，基于模板建模得分(TM - score)在蛋白质数据库(PDB，截至2022年4月12日)中识别相似结构，截断值为0.5(低于该值Foldseek将不返回任何结构)以实现更高效的搜索。对于Foldseek未返回任何模板建模得分大于0.5的结构的高置信度结构，我们使用带有多序列比对(MSAs)的完整AlphaFold2也获得结构预测(我们选择按平均预测局部距离差异测试排名前五的松弛模型中的最佳模型)。然后，我们计算进化尺度模型折叠(ESMFold)和AlphaFold2预测结构之间的对齐主链坐标的均方根偏差(RMSD)值和全原子模板建模得分。对于对应于高置信度结构的每个序列，我们还使用BLASTP 2.10.0+版本在UniRef90中搜索相似序列以计算序列同一性。如果所有返回的匹配结果的E值都大于1，或者UniRef90中最接近的条目序列同一性小于30%，我们将这些序列定义为与UniRef90具有低序列同一性。对于在UniRef90中没有显著匹配的序列，我们还使用JackHMMER网络服务器

(https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer) (51) to manually query four reference proteomes for similar sequences.

(https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer) (51) 手动查询四个参考蛋白质组以查找相似序列。

To construct the landscape of MGnify sequences, we first used ESM-1b to embed each sequence as a 1280-dimensional vector. These embeddings were then visualized using the umap version 0.5.3, scanpy version 1.9.1, and anndata 0.8.0 Python packages (63, 64), where dimensionality reduction was applied directly to the embedding vectors (use\_rep=’X’ in scanpy.tl.umap) with default parameters (15-nearest-neighbors graph via approximate Euclidean distance, UMAP min\_dist=0.5). Highlighted structure predictions with low similarity to known structures were manually selected and are summarized in Figures 4 and fig. S2. For these structures, we performed an additional structural similarity search using the Foldseek webserver (https://search.foldseek.com/search) with default parameters to identify the closest structures in PDB100 211201 beyond the TM-score cutoff of 0.5 .

为了构建MGnify序列的图谱，我们首先使用进化尺度模型1b(ESM - 1b)将每个序列嵌入为一个1280维的向量。然后使用umap 0.5.3版、scanpy 1.9.1版和anndata 0.8.0 Python包(63, 64)对这些嵌入进行可视化，其中直接对嵌入向量应用降维(在scanpy.tl.umap中use\_rep=’X’)，使用默认参数(通过近似欧几里得距离构建15近邻图，UMAP最小距离=0.5)。手动选择与已知结构相似度较低的突出结构预测，并总结在图4和图S2中。对于这些结构，我们使用Foldseek网络服务器(https://search.foldseek.com/search)以默认参数进行了额外的结构相似性搜索，以识别蛋白质数据库100(PDB100 211201)中超出模板建模得分截断值0.5的最接近结构。

1. ESM-1b evaluated only on sequences of length , due to constraints with position embedding.

   由于位置嵌入的限制，ESM - 1b仅在长度为 的序列上进行评估。 [↑](#footnote-ref-31)
2. Average computed across all CASP14 sequences.

   对所有CASP14序列计算的平均值。 [↑](#footnote-ref-36)
3. ESM-1b evaluated only on sequences of length , due to constraints with position embedding.

   由于位置嵌入的限制，ESM - 1b仅对长度为 的序列进行评估。 [↑](#footnote-ref-65)