# Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

# 生物结构和功能通过将无监督学习扩展到2.5亿个蛋白质序列中涌现

Alexander Rives Joshua Meier Tom Sercu Siddharth Goyal Zeming Lin Jason Liu Demi Guo Myle Ott C. Lawrence Zitnick Jerry Ma Rob Fergus

亚历山大·里夫斯 约书亚·迈尔 汤姆·塞尔库 西达尔特·戈亚尔 林泽明 杰森·刘 郭德米 迈尔·奥特 C. 劳伦斯·齐特尼克 杰里·马 罗布·弗格斯

# Abstract

# 摘要

In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In the life sciences, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Protein language modeling at the scale of evolution is a logical step toward predictive and generative artificial intelligence for biology. To this end we use unsupervised learning to train a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity. The resulting model contains information about biological properties in its representations. The representations are learned from sequence data alone. The learned representation space has a multi-scale organization reflecting structure from the level of biochemical properties of amino acids to remote homology of proteins. Information about secondary and tertiary structure is encoded in the representations and can be identified by linear projections. Representation learning produces features that generalize across a range of applications, enabling state-of-the-art supervised prediction of mutational effect and secondary structure, and improving state-of-the-art features for long-range contact prediction.

在人工智能领域，数据规模和模型能力的结合，通过无监督学习，推动了表示学习和统计生成的重大进展。在生命科学中，测序技术的预期增长有望带来关于自然序列多样性的前所未有的数据。在进化规模上进行蛋白质语言建模，是迈向生物学预测和生成人工智能的逻辑步骤。为此，我们使用无监督学习在跨越进化多样性的2.5亿个蛋白质序列中的860亿个氨基酸上训练了一个深度上下文语言模型。生成的模型在其表示中包含了关于生物特性的信息。这些表示仅从序列数据中学习得到。学习到的表示空间具有多尺度组织，反映了从氨基酸的生化特性到蛋白质远程同源性的结构。关于二级和三级结构的信息被编码在表示中，并可以通过线性投影识别。表示学习生成的特征在一系列应用中具有泛化能力，使得突变效应和二级结构的监督预测达到最先进水平，并改进了远程接触预测的最先进特征。

# 1. Introduction

# 1. 引言

Growth in the number of protein sequences in public databases has followed an exponential trend over decades, creating a deep view into the breadth and diversity of protein sequences across life. This data is a promising ground for studying predictive and generative models for biology using artificial intelligence. Our focus here will be to fit a single model to many diverse sequences from across evolution. Accordingly we study high-capacity neural networks, investigating what can be learned about the biology of proteins from modeling evolutionary data at scale.

几十年来，公共数据库中蛋白质序列数量的增长呈指数趋势，为我们提供了对生命体中蛋白质序列广度和多样性的深入视角。这些数据为使用人工智能研究生物学的预测和生成模型提供了有前景的基础。我们的重点是将一个单一模型拟合到来自整个进化的多种序列中。因此，我们研究高容量神经网络，探讨从大规模进化数据建模中可以学到哪些关于蛋白质生物学的知识。

The idea that biological function and structure are recorded in the statistics of protein sequences selected through evolution has a long history (Yanofsky et al., 1964; Altschuh et al., . Out of the possible random perturbations to a sequence, evolution is biased toward selecting those that are consistent with fitness (Göbel et al., 1994). The unobserved variables that determine a protein’s fitness, such as structure, function, and stability, leave a record in the distribution of observed natural sequences (Göbel et al., 1994).

生物功能和结构记录在通过进化选择的蛋白质序列统计中的观点由来已久(Yanofsky等，1964；Altschuh等， )。在可能的随机扰动中，进化偏向于选择那些与适应性一致的序列(Göbel等，1994)。决定蛋白质适应性的未观察变量，如结构、功能和稳定性，在观察到的自然序列分布中留下了记录(Göbel等，1994)。

Unlocking the information encoded in protein sequence variation is a longstanding problem in biology. An analogous problem in the field of artificial intelligence is natural language understanding, where the distributional hypothesis posits that a word’s semantics can be derived from the contexts in which it appears (Harris, 1954).

解锁蛋白质序列变异中编码的信息是生物学中的一个长期问题。在人工智能领域，一个类似的问题是自然语言理解，其中分布假设认为一个词的语义可以从其出现的上下文中推导出来(Harris，1954)。

Recently, techniques based on self-supervision, a form of unsupervised learning in which context within the text is used to predict missing words, have been shown to materialize representations of word meaning that can generalize across natural language tasks (Collobert & Weston, 2008; Dai & Le, 2015; Peters et al., 2018; Devlin et al., 2018). The ability to learn such representations improves significantly with larger training datasets (Baevski et al., 2019; Radford et al., 2019).

最近，基于自监督的技术，一种无监督学习形式，其中文本中的上下文用于预测缺失词，已被证明可以生成能够泛化到自然语言任务中的词义表示(Collobert & Weston，2008；Dai & Le，2015；Peters等，2018；Devlin等，2018)。学习此类表示的能力随着训练数据集的增大而显著提高(Baevski等，2019；Radford等，2019)。

Protein sequences result from a process greatly dissimilar to natural language. It is uncertain whether the models and objective functions effective for natural language transfer across differences between the domains. We explore this question by training high-capacity Transformer language models on evolutionary data. We investigate the resulting unsupervised representations for the presence of biological organizing principles, and information about intrinsic biological properties. We find metric structure in the representation space that accords with organizing principles at scales from physicochemical to remote homology. We also find that secondary and tertiary protein structure can be identified in representations. The structural properties captured by the representations generalize across folds. We apply the representations to a range of prediction tasks and find that they improve state-of-art features across the applications.

蛋白质序列的产生过程与自然语言大不相同。目前尚不确定对自然语言有效的模型和目标函数是否能够跨领域转移。我们通过在进化数据上训练高容量Transformer语言模型来探索这个问题。我们研究了生成的无监督表示中是否存在生物组织原则和内在生物特性的信息。我们发现表示空间中的度量结构与从物理化学到远程同源性的组织原则一致。我们还发现二级和三级蛋白质结构可以在表示中识别。表示捕获的结构特性在折叠之间具有泛化能力。我们将这些表示应用于一系列预测任务，并发现它们在这些应用中改进了最先进的特征。

[[1]](#footnote-1)

# 2. Background

# 2. 背景

Sequence alignment and search is a longstanding basis for comparative and statistical analysis of biological sequence data. (Altschul et al., 1990; Altschul & Koonin, 1998; Eddy, 1998; Remmert et al., 2011). Search across large databases containing evolutionary diversity assembles related sequences into a multiple sequence alignment (MSA). Within sequence families, mutational patterns convey information about functional sites, stability, tertiary contacts, binding, and other properties (Altschuh et al., 1987; 1988; Göbel et al., 1994). Conserved sites correlate with functional and structural importance (Altschuh et al., 1987). Local biochemical and structural contexts are reflected in preferences for distinct classes of amino acids (Levitt, 1978). Covarying mutations have been associated with function, tertiary contacts, and binding (Göbel et al., 1994).

序列比对和搜索是生物序列数据比较和统计分析的基础(Altschul等，1990；Altschul & Koonin，1998；Eddy，1998；Remmert等，2011)。在包含进化多样性的大型数据库中进行搜索，可以将相关序列组装成多序列比对(MSA)。在序列家族中，突变模式传递了关于功能位点、稳定性、三级接触、结合和其他特性的信息(Altschuh等，1987；1988；Göbel等，1994)。保守位点与功能和结构的重要性相关(Altschuh等，1987)。局部生化和结构环境反映在对不同类别氨基酸的偏好中(Levitt，1978)。共变突变与功能、三级接触和结合相关(Göbel等，1994)。

The prospect of inferring biological structure and function from evolutionary statistics has motivated development of machine learning on individual sequence families. Direct coupling analysis (Lapedes et al., 1999; Thomas et al., 2008; Weigt et al., 2009) infers constraints on the structure of a protein by fitting a generative model in the form of a Markov Random Field (MRF) to the sequences in the protein’s MSA. Various methods have been developed to fit the MRF (Mor-cos et al., 2011; Jones et al., 2011; Balakrishnan et al., 2011; Ekeberg et al., 2013b). The approach can also be used to infer functional constraints (Figliuzzi et al., 2016; Hopf et al., 2017), and the generative picture can be extended to include latent variables (Riesselman et al., 2018).

从进化统计中推断生物结构和功能的前景，推动了机器学习在单个序列家族中的应用。直接耦合分析(Lapedes等，1999；Thomas等，2008；Weigt等，2009)通过将马尔可夫随机场(MRF)形式的生成模型拟合到蛋白质的MSA序列中，推断蛋白质结构的约束。已经开发了多种方法来拟合MRF(Morcos等，2011；Jones等，2011；Balakrishnan等，2011；Ekeberg等，2013b)。该方法也可用于推断功能约束(Figliuzzi等，2016；Hopf等，2017)，并且生成模型可以扩展到包括潜在变量(Riesselman等，2018)。

Recently, self-supervision has emerged as a core direction in artificial intelligence research. Unlike supervised learning which requires manual annotation of each datapoint, self-supervised methods use unlabeled datasets and thus can exploit far larger amounts of data. Self-supervised learning uses proxy tasks for training, such as predicting the next word in a sentence given all previous words (Bengio et al., 2003; Dai & Le, 2015; Peters et al., 2018; Radford et al., or predicting words that have been masked from their context (Devlin et al., 2018; Mikolov et al., 2013).

最近，自监督学习已成为人工智能研究的核心方向。与需要手动标注每个数据点的监督学习不同，自监督方法使用未标注的数据集，因此可以利用更大量的数据。自监督学习使用代理任务进行训练，例如在给定所有先前单词的情况下预测句子中的下一个单词(Bengio等，2003；Dai & Le，2015；Peters等，2018；Radford等， )，或预测从其上下文中被掩盖的单词(Devlin等，2018；Mikolov等，2013)。

Increasing the dataset size and the model capacity has shown improvements in the learned representations. In recent work, self-supervision methods used in conjunction with large data and high-capacity models produced new state-of-the-art results approaching human performance on various question answering and semantic reasoning benchmarks (Devlin et al., 2018), and coherent natural text generation (Radford et al., 2019).

增加数据集大小和模型容量已显示出在学习表示方面的改进。在最近的工作中，自监督方法与大数据和高容量模型结合使用，在各种问答和语义推理基准测试中产生了接近人类表现的最新结果(Devlin等，2018)，以及连贯的自然文本生成(Radford等，2019)。

This paper explores self-supervised language modeling approaches that have demonstrated state-of-the-art performance on a range of natural language processing tasks, applying them to protein data in the form of unlabeled amino acid sequences. Since protein sequences use a small vocabulary of twenty canonical elements, the modeling problem is more similar to character-level language models (Mikolov et al., 2012; Kim et al., 2016) than word-level models. Like natural language, protein sequences also contain long-range dependencies, motivating use of architectures that detect and model distant context (Vaswani et al., 2017).

本文探讨了在一系列自然语言处理任务中表现出最先进性能的自监督语言建模方法，并将其应用于未标注的氨基酸序列形式的蛋白质数据。由于蛋白质序列使用二十种标准元素的小词汇表，建模问题更类似于字符级语言模型(Mikolov等，2012；Kim等，2016)，而不是单词级模型。与自然语言一样，蛋白质序列也包含长程依赖关系，这促使使用能够检测和建模远距离上下文的架构(Vaswani等，2017)。

# 3. Scaling language models to 250 million diverse protein sequences

# 3. 将语言模型扩展到2.5亿个多样化的蛋白质序列

Large protein sequence databases contain diverse sequences sampled across life. In our experiments we explore datasets with up to 250 million sequences of the Uniparc database (The UniProt Consortium, 2007) which has 86 billion amino acids. This data is comparable in size to large text datasets that are being used to train high-capacity neural network architectures on natural language (Devlin et al., 2018; Radford et al., 2019). To model the data of evolution with fidelity, neural network architectures must have capacity and inductive biases to represent its breadth and diversity.

大型蛋白质序列数据库包含从生命体中采样的多样化序列。在我们的实验中，我们探索了包含多达2.5亿个序列的Uniparc数据库(The UniProt Consortium，2007)，该数据库包含860亿个氨基酸。这些数据的大小与用于训练高容量神经网络架构的自然语言大型文本数据集相当(Devlin等，2018；Radford等，2019)。为了忠实地建模进化数据，神经网络架构必须具有容量和归纳偏差，以表示其广度和多样性。

We investigate the Transformer (Vaswani et al., 2017), which has emerged as a powerful general-purpose model architecture for representation learning and generative modeling, outperforming recurrent and convolutional architectures in natural language settings. We use a deep Transformer (Devlin et al., 2018), taking as input amino acid character sequences.

我们研究了Transformer(Vaswani等人，2017)，它已成为一种强大的通用模型架构，用于表示学习和生成建模，在自然语言环境中优于循环和卷积架构。我们使用深度Transformer(Devlin等人，2018)，以氨基酸字符序列作为输入。

The Transformer processes inputs through a series of blocks that alternate self-attention with feed-forward connections. Self-attention allows the network to build up complex representations that incorporate context from across the sequence. Since self-attention explicitly constructs pairwise interactions between all positions in the sequence, the Transformer architecture directly represents residue-residue interactions.

Transformer通过一系列交替进行自注意力和前馈连接的块来处理输入。自注意力使网络能够构建复杂的表示，这些表示结合了序列中的上下文信息。由于自注意力显式地构建了序列中所有位置之间的成对交互，Transformer架构直接表示了残基-残基相互作用。

We train models using the masked language modeling objective (Devlin et al., 2018). Each input sequence is corrupted by replacing a fraction of the amino acids with a special mask token. The network is trained to predict the missing tokens from the corrupted sequence: bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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.

我们使用**掩码语言建模目标**(Devlin等人，2018)来训练模型。每个输入序列通过**用特殊掩码标记替换一部分氨基酸来进行破坏**。网络**被训练从破坏的序列中预测缺失的标记**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model |  | Params | Training | ECE |
|  | Oracle |  |  |  | 1 |
| (a) | Uniform Random |  |  |  | 25 |
| (b) | n-gram | 4-gram |  | UR50/S | 17.18 |
| (c) | LSTM | Small | 28.4M | UR50/S | 14.42 |
| LSTM | Large | 113.4M | UR50/S | 13.54 |
| (d) | Transformer | 6-layer | 42.6M | UR50/S | 11.79 |
| Transformer | 12-layer | 85.1M | UR50/S | 10.45 |
| (e) | Transformer | 34-layer | 669.2M | UR100 | 10.32 |
| Transformer | 34-layer | 669.2M | UR50/S | 8.54 |
| Transformer | 34-layer | 669.2M | UR50/D | 8.46 |
| (f) | Transformer | 10% data | 669.2M | UR50/S | 10.99 |
| Transformer | 1% data | 669.2M | UR50/S | 15.01 |
| Transformer | 0.1% data | 669.2M | UR50/S | 17.50 |
|  | 模型 |  | 参数 | 训练 | ECE |
|  | Oracle |  |  |  | 1 |
| (a) | 均匀随机 |  |  |  | 25 |
| (b) | n-gram | 4-gram |  | UR50/S | 17.18 |
| (c) | LSTM | 小型 | 28.4M | UR50/S | 14.42 |
| LSTM | 大型 | 113.4M | UR50/S | 13.54 |
| (d) | Transformer | 6层 | 42.6M | UR50/S | 11.79 |
| Transformer | 12层 | 85.1M | UR50/S | 10.45 |
| (e) | Transformer | 34层 | 669.2M | UR100 | 10.32 |
| Transformer | 34层 | 669.2M | UR50/S | 8.54 |
| Transformer | 34层 | 669.2M | UR50/D | 8.46 |
| (f) | Transformer | 10% 数据 | 669.2M | UR50/S | 10.99 |
| Transformer | 1% 数据 | 669.2M | UR50/S | 15.01 |
| Transformer | 0.1% 数据 | 669.2M | UR50/S | 17.50 |

Table 1. Evaluation of language models for generalization to held-out UniRef50 clusters. (a) Exponentiated cross-entropy (ECE) ranges from 25 for a random model to 1 for a perfect model. (b) Best n-gram model across range of context sizes and Laplace-smoothing settings. (c) State-of-the-art LSTM bidirectional language models (Peters et al., 2018). (d) Transformer model baselines with 6 and 12 layers. Small Transformer models have better performance than LSTMs despite having fewer parameters. (e) 34-layer Transformer models trained on datasets of differing sequence diversity. Increasing the diversity of the training set improves generalization. High-capacity Transformer models outperform LSTMs and smaller Transformers. (f) 34-layer Transformer models trained on reduced fractions of data. Increasing training data improves generalization.

表1. 语言模型在保留的UniRef50聚类上的泛化能力评估。(a) 指数化交叉熵(ECE)范围从随机模型的25到完美模型的1。(b) 在不同上下文大小和拉普拉斯平滑设置下的最佳n-gram模型。(c) 最先进的LSTM双向语言模型(Peters等，2018)。(d) 具有6层和12层的Transformer模型基线。尽管参数较少，小型Transformer模型的表现优于LSTM。(e) 在不同序列多样性数据集上训练的34层Transformer模型。增加训练集的多样性提高了泛化能力。高容量Transformer模型优于LSTM和较小的Transformer。(f) 在减少数据量上训练的34层Transformer模型。增加训练数据提高了泛化能力。

For each sequence we sample a set of indices to mask, replacing the true token at each index with the mask token. For each masked token, we independently minimize the negative log likelihood of the true amino acid given the masked sequence as context. Intuitively, to make a prediction for a masked position, the model must identify dependencies between the masked site and the unmasked parts of the sequence.

对于每个序列 ，我们采样一组索引 进行掩码，用掩码标记替换每个索引 处的真实标记。对于每个被掩码的标记，我们独立地最小化给定掩码序列 作为上下文的真实氨基酸 的负对数似然。直观地说，为了对被掩码的位置进行预测，模型必须识别被掩码位点与序列未掩码部分之间的依赖关系。

Evaluation of language models We begin by training a series of Transformers on all the sequences in UniParc (The UniProt Consortium, 2007), holding out a random sample of sequences for validation. We use these models throughout to investigate properties of the representations and the information learned during pre-training.

语言模型评估 我们首先在UniParc(UniProt Consortium，2007)中的所有序列上训练一系列Transformer，保留 个序列的随机样本用于验证。我们使用这些模型来研究表示特性和预训练期间学习到的信息。

To comparatively evaluate generalization performance of different language models we use UniRef50 (Suzek et al., 2015), a clustering of UniParc at sequence identity. For evaluation, a held-out set of 10% of the UniRef50 clusters is randomly sampled. The evaluation dataset consists of the representative sequences of these clusters. All sequences belonging to the held-out clusters are removed from the pre-training datasets.

为了比较评估不同语言模型的泛化性能，我们使用UniRef50(Suzek等，2015)，这是UniParc在 序列同一性上的聚类。为了评估，随机采样了10%的UniRef50聚类作为保留集。评估数据集由这些聚类的代表序列组成。所有属于保留聚类的序列都从预训练数据集中移除。

We explore the effect of the underlying sequence diversity in the pre-training data. Clustering UniParc shows a power-law distribution of cluster sizes (Suzek et al., 2007), implying the majority of sequences belong to a small fraction of clusters. Training using a clustering of the sequences results in a re-weighting of the masked language modeling loss toward a more diverse set of sequences. We use UniRef (Suzek et al., 2015) to create three pre-training datasets with differing levels of diversity: (i) the low-diversity dataset (UR100) uses the UniRef100 representative sequences; (ii) the high-diversity sparse dataset uses the UniRef50 representative sequences; (iii) the high-diversity dense dataset (UR50/D) samples the UniRef100 sequences evenly across the UniRef50 clusters.

我们探讨了预训练数据中基础序列多样性的影响。聚类UniParc显示了聚类大小的幂律分布(Suzek等，2007)，这意味着大多数序列属于一小部分聚类。使用序列聚类进行训练会导致掩码语言建模损失向更多样化的序列集重新加权。我们使用UniRef(Suzek等，2015)创建了三个具有不同多样性水平的预训练数据集:(i) 低多样性数据集(UR100)使用UniRef100代表序列；(ii) 高多样性稀疏数据集 使用UniRef50代表序列；(iii) 高多样性密集数据集(UR50/D)在UniRef50聚类中均匀采样UniRef100序列。

Table 1 presents modeling performance on the held-out UniRef50 sequences across a series of experiments exploring different model classes, number of parameters, and pretraining datasets. Models are compared using the exponentiated cross entropy (ECE) metric, which is the exponential of the model’s loss averaged per token. In the case of the Transformer this is . ECE describes the mean uncertainty of the model among its set of options for every prediction: ranging from 1 for an ideal model to 25 (the number of unique amino acid tokens in the data) for a completely random prediction. To measure the difficulty of generalization to the evaluation set, we train a series of n-gram models across a range of context lengths and settings of Laplace smoothing on UR50/S. The best n-gram model has an ECE of 17.18 with context size of 4 .

表1展示了在一系列探索不同模型类别、参数数量和预训练数据集的实验中，保留的UniRef50序列上的建模性能。模型使用指数化交叉熵(ECE)指标进行比较，该指标是模型每个标记的平均损失的指数。在Transformer的情况下，这是 。**ECE描述了模型在每个预测选项中的平均不确定性**:从理想模型的1到完全随机预测的25(数据中唯一氨基酸标记的数量)。为了衡量泛化到评估集的难度，我们在UR50/S上训练了一系列n-gram模型，跨越不同的上下文长度和拉普拉斯平滑设置。最佳n-gram模型的ECE为17.18，上下文大小为4。

As a baseline we train recurrent LSTM bidirectional language models (Peters et al., 2018), which are state-of-the-art for recurrent models in the text domain. Unlike standard left-to-right autoregressive LSTMs, these models use the whole sequence context, making them comparable to the Transformers we study. We evaluate a small model with approximately parameters, and a large model with approximately parameters. Trained on the UR50/S dataset, the small and large LSTM models have an ECE of 14.4 and 13.5 respectively.

作为基线，我们训练了循环LSTM双向语言模型(Peters等，2018)，这是文本领域中最先进的循环模型。与标准的从左到右自回归LSTM不同，这些模型使用整个序列上下文，使它们与我们研究的Transformer具有可比性。我们评估了一个大约有 个参数的小型模型，以及一个大约有 个参数的大型模型。在UR50/S数据集上训练的小型和大型LSTM模型的ECE分别为14.4和13.5。

We also train two small Transformers, a 12-layer (85.1M parameters) and 6-layer Transformer (42.6M parameters) on the UR50/S dataset. Both Transformer models have better ECE values (10.45, and 11.79 respectively) than the small and large LSTM models, despite the large LSTM having more parameters. These results show the Transformer enables higher fidelity modeling of protein sequences for a comparable number of parameters.

我们还在UR50/S数据集上训练了两个小型Transformer模型，分别是12层(8500万参数)和6层Transformer(4260万参数)。尽管大型LSTM模型拥有更多参数，但这两个Transformer模型的ECE值(分别为10.45和11.79)均优于小型和大型LSTM模型。这些结果表明，在参数数量相当的情况下，Transformer能够实现更高保真度的蛋白质序列建模。

We train high-capacity 34-layer Transformers (approx 670M parameters) across the three datasets of differing diversity. The high-capacity Transformer model trained on the dataset outperforms the smaller Transformers indicating an improvement in language modeling with increasing model capacity. Transformers trained on the two high-diversity datasets, UR50/S and UR50/D, improve generalization over the UR100 low-diversity dataset. The best Transformer trained on the most diverse and dense dataset reaches an ECE of 8.46, meaning that intuitively the is model choosing among approximately 8.46 amino acids for each prediction.

我们在三个不同多样性的数据集上训练了高容量的34层Transformer模型(约6.7亿参数)。在 数据集上训练的高容量Transformer模型表现优于较小的Transformer模型，表明随着模型容量的增加，语言建模能力有所提升。在UR50/S和UR50/D这两个高多样性数据集上训练的Transformer模型，相较于UR100低多样性数据集，其泛化能力有所提高。在最多样化和最密集的数据集上训练的最佳Transformer模型的ECE值为8.46，这意味着该模型在每次预测时直观上大约在8.46个氨基酸中进行选择。

We also train a series of 34-layer Transformer models on , and of the UR50/S dataset, seeing the expected relationship between increased data and improved generalization performance. Underfitting is observed even for the largest models trained on of UR50/S suggesting potential for additional improvements with higher capacity models.

我们还在UR50/S数据集的 和 上训练了一系列34层Transformer模型，观察到数据增加与泛化性能提升之间的预期关系。即使在UR50/S的 上训练的最大模型也观察到欠拟合现象，这表明更高容量的模型仍有改进空间。

ESM-1b Transformer Finally we perform a systematic optimization of model hyperparameters on parameter models to identify a robust set of hyperparameters. The hyperparameter search is described in detail in Appendix B. We scale the hyperparameters identified by this search to train a model with approximately parameters (33 layers) on the UR50/S dataset, resulting in the ESM-1b Transformer.

ESM-1b Transformer 最后，我们在 参数模型上对模型超参数进行了系统优化，以确定一组稳健的超参数。超参数搜索的详细描述见附录B。我们将通过此搜索确定的超参数扩展到在UR50/S数据集上训练一个约 参数(33层)的模型，最终得到ESM-1b Transformer。

# 4. Multi-scale organization in sequence representations

# 4. 序列表示中的多尺度组织

The variation observed in large protein sequence datasets is influenced by processes at many scales, including properties that affect fitness directly, such as activity, stability, structure, binding, and other properties under selection (Hor-moz, 2013; Hopf et al., 2017) as well as by contributions from phylogenetic bias (Gabaldon, 2007), experimental and selection biases (Wang et al., 2019; Overbaugh & Bang-ham, 2001), and sources of noise such as random genetic drift (Kondrashov et al., 2003).

大型蛋白质序列数据集中观察到的变异受到多尺度过程的影响，包括直接影响适应性的属性，如活性、稳定性、结构、结合和其他选择下的属性(Hormoz, 2013; Hopf et al., 2017)，以及系统发育偏差(Gabaldon, 2007)、实验和选择偏差(Wang et al., 2019; Overbaugh & Bangham, 2001)和随机遗传漂变等噪声来源(Kondrashov et al., 2003)的贡献。

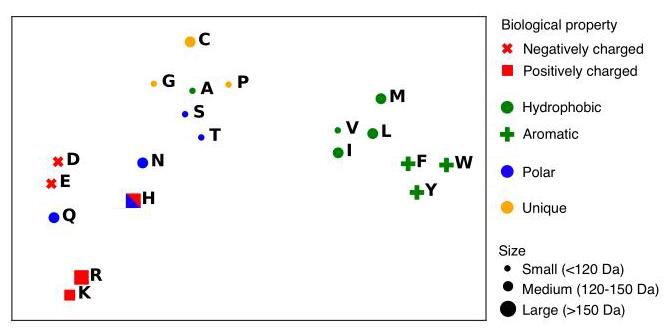


Figure 1. Biochemical properties of amino acids are represented in the Transformer model’s output embeddings, visualized here with t-SNE. Through unsupervised learning, residues are clustered into hydrophobic, polar, and aromatic groups, and reflect overall organization by molecular weight and charge. Visualization of 36-layer Transformer trained on UniParc.

图1. 氨基酸的生化特性在Transformer模型的输出嵌入中表示，此处通过t-SNE进行可视化。通过无监督学习，残基被聚类为疏水、极性和芳香族基团，并反映了分子量和电荷的整体组织。在UniParc上训练的36层Transformer的可视化。

Unsupervised learning may encode underlying factors that, while unobserved, are useful for explaining the variation in sequences seen by the model during pre-training. We investigate the representation space of the network at multiple scales from biochemical to evolutionary homology to look for signatures of biological organization.

无监督学习可能编码了潜在因素，这些因素虽然未被观察到，但对于解释模型在预训练期间看到的序列变异是有用的。我们从生化到进化同源性的多尺度上研究网络的表示空间，以寻找生物组织的特征。

Neural networks contain inductive biases that impart structure to representations. Randomly initialized networks can produce features that perform well without any learning (Jarrett et al., 2009). To understand how the process of learning shapes the representations, it is necessary to compare representations before and after they have been trained. Furthermore, a basic level of intrinsic organization is expected in the sequence data itself as a result of biases in amino acid composition. To disentangle the role of frequency bias in the data we also compare against a baseline that maps each sequence to a vector of normalized amino acid counts.

神经网络包含归纳偏差，这些偏差赋予表示结构。随机初始化的网络可以生成在没有学习的情况下表现良好的特征(Jarrett et al., 2009)。为了理解学习过程如何塑造表示，有必要比较训练前后的表示。此外，由于氨基酸组成的偏差，序列数据本身预计会存在一定程度的固有组织。为了解开数据中频率偏差的作用，我们还与一个将每个序列映射到归一化氨基酸计数向量的基线进行比较。

Learning encodes biochemical properties The Transformer neural network represents the identity of each amino acid in its input and output embeddings. The input em-beddings project the input amino acid tokens into the first Transformer block. The output embeddings project the final hidden representations back to logarithmic probabilities. The interchangeability of amino acids within a given structural or functional context in a protein depends on their biochemical properties (Hormoz, 2013). Self-supervision can be expected to capture these patterns to build a representation space that reflects biochemical knowledge. under aCC-BY-NC-ND 4.0 International license. To investigate if the network has learned to encode physicochemical properties in representations, we project the weight matrix of the final embedding layer of the network into two dimensions with t-SNE (Maaten & Hinton, 2008). In Figure 1 the structure of the embedding space reflects biochemical interchangeability with distinct clustering of hydrophobic and polar residues, aromatic amino acids, and organization by molecular weight and charge.

学习编码生化特性 Transformer神经网络在其输入和输出嵌入中表示每个氨基酸的身份。输入嵌入将输入氨基酸标记投影到第一个Transformer块中。输出嵌入将最终的隐藏表示投影回对数概率。在蛋白质的给定结构或功能上下文中，氨基酸的可互换性取决于它们的生化特性(Hormoz, 2013)。自监督学习可以预期捕捉这些模式，以构建反映生化知识的表示空间。为了研究网络是否学会了在表示中编码物理化学特性，我们使用t-SNE(Maaten & Hinton, 2008)将网络最终嵌入层的权重矩阵投影到二维空间中。在图1中，嵌入空间的结构反映了生化可互换性，疏水和极性残基、芳香族氨基酸以及分子量和电荷的组织有明显的聚类。

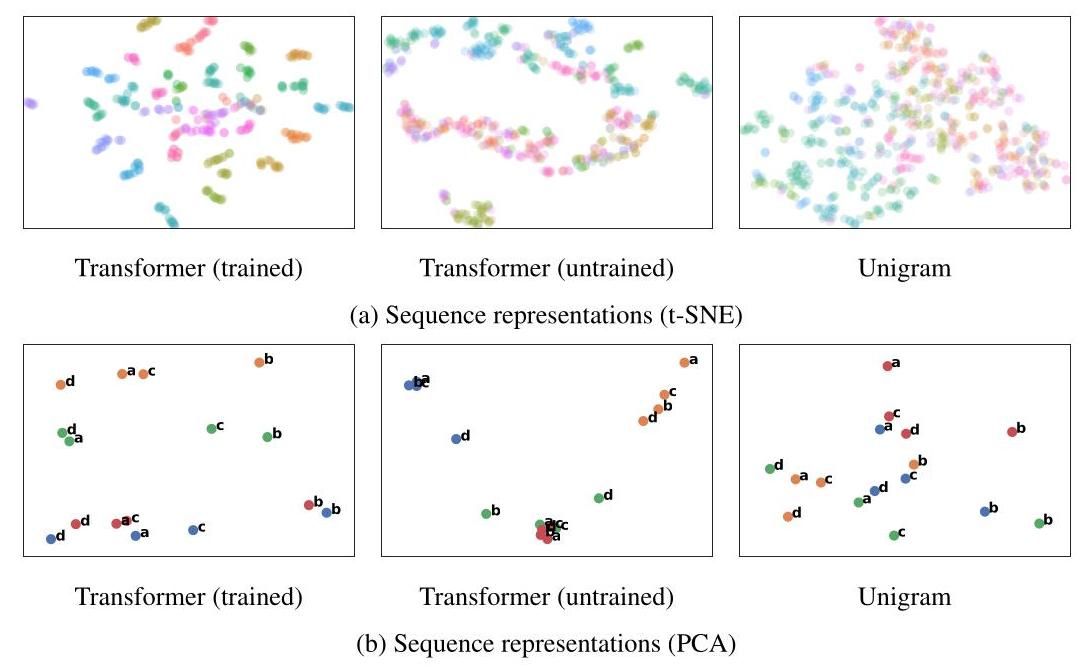


Figure 2. Protein sequence representations encode and organize biological variations. (a) Each point represents a gene, and each gene is colored by the orthologous group it belongs to (dimensionality is reduced by t-SNE). Orthologous groups of genes are densely clustered in the trained representation space. By contrast, the untrained representation space and unigram representations do not reflect strong organization by evolutionary relationships. (b) Genes corresponding to a common biological variation are related linearly in the trained representation space. Genes are colored by their orthologous group, and their species is indicated by a character label. PCA recovers a species axis (horizontal) and orthology axis (vertical) in the trained representation space, but not in the untrained or unigram spaces. Representations are from the 36-layer Transformer model trained on UniParc.

图2. 蛋白质序列表示法编码和组织生物变异。(a) 每个点代表一个基因，每个基因按其所属的直系同源组着色(通过t-SNE降维)。在训练后的表示空间中，基因的直系同源组紧密聚集。相比之下，未训练的表示空间和单字表示法并未反映出由进化关系驱动的强组织性。(b) 在训练后的表示空间中，对应于共同生物变异的基因呈线性相关。基因按其直系同源组着色，其物种由字符标签表示。PCA在训练后的表示空间中恢复了物种轴(水平)和直系同源轴(垂直)，但在未训练或单字空间中则未恢复。表示法来自在UniParc上训练的36层Transformer模型。

Biological variations are encoded in representation space Each protein can be represented as a single vector by averaging across the hidden representation at each position in its sequence. Protein embeddings represent sequences as points in a high dimensional space. Each sequence is represented as a single point and sequences assigned to similar representations by the network are mapped to nearby points. We investigate how homologous genes are represented in this space.

生物变异在表示空间中编码 每个蛋白质可以通过对其序列中每个位置的隐藏表示进行平均来表示为单个向量。蛋白质嵌入将序列表示为高维空间中的点。每个序列表示为单个点，网络将分配给相似表示的序列映射到附近的点。我们研究了同源基因在此空间中的表示方式。

The structure and function of orthologous genes is likely to be retained despite divergence of their sequences (Huerta-Cepas et al., 2018). We find in Figure 2a that training shapes the representation space so that orthologous genes are clustered. Figure 2a shows a two-dimensional projection of the model’s representation space using t-SNE. Prior to training the organization of orthologous proteins in the model’s representation space is diffuse. Orthologous genes are clustered in the learned representation space.

尽管序列存在分歧，直系同源基因的结构和功能可能得以保留(Huerta-Cepas等，2018)。我们在图2a中发现，训练塑造了表示空间，使得直系同源基因聚集在一起。图2a展示了使用t-SNE对模型表示空间的二维投影。在训练之前，模型中直系同源蛋白质的组织是分散的。直系同源基因在学习的表示空间中聚集。

We examine whether unsupervised learning encodes biological variations into the structure of the representation space. We apply principal component analysis (PCA), to recover principal directions of variation in the representations, selecting 4 orthologous genes across 4 species to look for directions of variation. Figure indicates that linear dimensionality reduction recovers species and orthology as primary axes of variation in the representation space after training. This form of structure is absent from the representations prior to training.

我们研究了无监督学习是否将生物变异编码到表示空间的结构中。我们应用主成分分析(PCA)来恢复表示中的主要变异方向，选择4个物种中的4个直系同源基因来寻找变异方向。图 表明，线性降维在训练后恢复了物种和直系同源作为表示空间中的主要变异轴。这种结构形式在训练前的表示中不存在。

To quantitatively investigate the structure of the representation space, we assess nearest neighbor recovery under vector similarity queries. If biological properties are encoded along independent directions in the representation space, then proteins corresponding with a unique biological variation are related by linear vector arithmetic. In Figure S1 we find that learning improves recovery of target proteins under queries encoded as linear transformations along the species or gene axes.

为了定量研究表示空间的结构，我们评估了在向量相似性查询下的最近邻恢复。如果生物特性沿表示空间中的独立方向编码，那么对应于独特生物变异的蛋白质通过线性向量运算相关。在图S1中，我们发现学习提高了在沿物种或基因轴编码的线性变换查询下目标蛋白质的恢复。

Learning encodes remote homology Remotely homologous proteins have underlying structural similarity despite divergence of their sequences. If structural homology is encoded in the metric structure of the representation space, then the distance between proteins reflects their degree of structural relatedness.

学习编码远程同源性 尽管序列存在分歧，远程同源蛋白质具有潜在的结构相似性。如果结构同源性在表示空间的度量结构中编码，那么蛋白质之间的距离反映了它们的结构相关程度。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | training | Hit-10 | | AUC | |
| Fold | SF | Fold | SF |
| HHblits |  | .584 | .965 | .831 | .951 |
| LSTM (S) | UR50/S | .558 | .760 | .801 | .863 |
| LSTM (L) | UR50/S | .574 | .813 | .805 | .880 |
| Transf-6 | UR50/S | .653 | .878 | .768 | .901 |
| Transf-12 | UR50/S | .639 | .915 | .778 | .942 |
| Transf-34 | None | .481 | .527 | .755 | .807 |
| Transf-34 | UR100 | .599 | .841 | .753 | .876 |
| Transf-34 | UR50/D | .617 | .932 | .822 | .932 |
| Transf-34 | UR50/S | .639 | .931 | .825 | .933 |
| ESM-1b | UR50/S | .532 | .913 | .770 | .880 |
|  | 训练 | 命中率-10 | | 曲线下面积(AUC) | |
| 折叠 | SF | 折叠 | SF |
| HHblits |  | .584 | .965 | .831 | .951 |
| 长短期记忆网络(S) | UR50/S | .558 | .760 | .801 | .863 |
| 长短期记忆网络(L) | UR50/S | .574 | .813 | .805 | .880 |
| Transformer-6 | UR50/S | .653 | .878 | .768 | .901 |
| Transformer-12 | UR50/S | .639 | .915 | .778 | .942 |
| Transformer-34 | 无 | .481 | .527 | .755 | .807 |
| Transformer-34 | UR100 | .599 | .841 | .753 | .876 |
| Transformer-34 | UR50/D | .617 | .932 | .822 | .932 |
| Transformer-34 | UR50/S | .639 | .931 | .825 | .933 |
| ESM-1b | UR50/S | .532 | .913 | .770 | .880 |

Table 2. Remote homology detection. Structural homology at the fold and superfamily (SF) level is encoded in the metric structure of the representation space. Results for unsupervised classifier based on distance between vector sequence embeddings. Hit-10 reports the probability that a remote homolog is included in the ten nearest neighbors of the query sequence. Area under the ROC curve (AUC) is reported for classification by distance from the query in representation space. Transformer models have higher performance than LSTMs and similar performance to HMMs at the fold level. Best neural models are indicated in bold. blits (Remmert et al., 2011), a state-of-the-art HMM-based method for remote homology detection, using 3 iterations of sequence search.

表2. 远程同源性检测。折叠和超家族(SF)级别的结构同源性在表示空间的度量结构中编码。基于向量序列嵌入之间距离的无监督分类器的结果。Hit-10报告了远程同源物包含在查询序列的十个最近邻中的概率。ROC曲线下面积(AUC)报告了表示空间中与查询距离的分类。Transformer模型在折叠级别上比LSTM表现更好，与HMM表现相似。最佳神经模型以粗体表示。 blits(Remmert等，2011)，一种基于HMM的远程同源性检测的最先进方法，使用3次序列搜索迭代。

We investigate whether the representation space enables detection of remote homology at the superfamily (proteins that belong to different families but are in the same superfamily) and fold (proteins that belong to different superfamilies but have the same fold) level. We construct a dataset to evaluate remote homology detection using SCOPe (Fox et al., 2014). Following standard practices (Söding & Remmert, 2011) we exclude Rossman-like folds (c.2-c.5, c.27 and 28, c.30 and 31) and four- to eight-bladed -propellers (b.66-b.70).

我们研究了表示空间是否能够在超家族(属于不同家族但属于同一超家族的蛋白质)和折叠(属于不同超家族但具有相同折叠的蛋白质)级别上检测远程同源性。我们构建了一个数据集，使用SCOPe(Fox等，2014)评估远程同源性检测。遵循标准实践(Söding & Remmert，2011)，我们排除了类似Rossman的折叠(c.2-c.5，c.27和28，c.30和31)以及四到八叶 -螺旋桨(b.66-b.70)。

An unsupervised classifier on distance from the query measures the density of homologous proteins in the neighborhood of a query sequence. For each domain, a vector similarity query is performed against all other domains, ranking them by distance to the query domain. For evaluation at the fold level, any domain with the same fold is a positive; any domain with a different fold is a negative; and domains belonging to the same superfamily are excluded. For evaluation at the superfamily level, any domain with the same superfamily is a positive; any domain with a different superfamily is a negative; and domains belonging to the same family are excluded. We report AUC for the classifier, and Hit-10 (Ma et al., 2014) which gives the probability of recovering a remote homolog in the ten highest ranked results.

基于查询距离的无监督分类器测量查询序列邻域中同源蛋白质的密度。对于每个域，对所有其他域执行向量相似性查询，按与查询域的距离进行排序。在折叠级别评估时，任何具有相同折叠的域为正；任何具有不同折叠的域为负；属于同一超家族的域被排除。在超家族级别评估时，任何具有相同超家族的域为正；任何具有不同超家族的域为负；属于同一家族的域被排除。我们报告了分类器的AUC，以及Hit-10(Ma等，2014)，它给出了在十个最高排名结果中恢复远程同源物的概率。

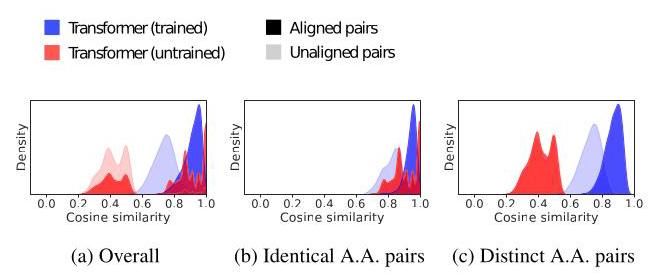


Figure 3. Final representations from trained models implicitly align sequences. Cosine similarity distributions are depicted for the final representations of residues from sequences within PFAM family PF01010. The differences between the aligned (dark blue) and unaligned (light blue) distributions imply that the trained Transformer representations are a powerful discriminator between aligned and unaligned positions in the sequences. In contrast representations prior to training do not separate the aligned (dark red) and unaligned positions (light red). AUCs across 128 PFAM families are reported in Table S1.

图3. 训练模型的最终表示隐含地对齐序列。描绘了来自PFAM家族PF01010序列残基的最终表示的余弦相似性分布。对齐(深蓝色)和未对齐(浅蓝色)分布之间的差异表明，训练后的Transformer表示是序列中对齐和未对齐位置的有力区分器。相比之下，训练前的表示并未区分对齐(深红色)和未对齐位置(浅红色)。128个PFAM家族的AUC在表S1中报告。

Table 2 indicates that vector nearest neighbor queries using the representations can detect remote homologs that are distant at the fold level with similar performance to HH-blits (Remmert et al., 2011) a state-of-the-art HMM-HMM alignment-based method. At the superfamily level, where sequence similarity is higher, HMM performance is better, but Transformer embeddings are close. Fast vector nearest neighbor finding methods allow billions of sequences to be searched for similarity to a query protein within milliseconds (Johnson et al., 2017).

表2表明，使用表示的向量最近邻查询可以检测到在折叠级别上距离较远的远程同源物，其性能与HH-blits(Remmert等，2011)相似，后者是一种基于HMM-HMM对齐的最先进方法。在序列相似性较高的超家族级别，HMM表现更好，但Transformer嵌入接近。快速向量最近邻查找方法允许在毫秒内搜索数十亿个序列与查询蛋白质的相似性(Johnson等，2017)。

# Learning encodes alignment within a protein family

# 学习编码蛋白质家族内的对齐

An MSA identifies corresponding sites across a family of related sequences (Ekeberg et al., 2013a). These correspondences give a picture of evolutionary variation at different sites within the sequence family. The model receives as input individual sequences and is given no access to the family of related sequences except via learning. We investigate whether the final hidden representations of a sequence encode information about the family it belongs to.

MSA识别了一组相关序列中的对应位点(Ekeberg等，2013a)。这些对应关系提供了序列家族内不同位点的进化变异图。模型接收单个序列作为输入，并且除了通过学习外，无法访问相关序列家族。我们研究了序列的最终隐藏表示是否编码了其所属家族的信息。

Family information could appear in the network through assignment of similar representations to positions in different sequences that are aligned in the family’s MSA. Using the collection of MSAs of structurally related sequences in Pfam (Bateman et al., 2013), we compare the distribution of cosine similarities of representations between pairs of residues that are aligned in the family’s MSA to a background distribution of cosine similarities between unaligned pairs of residues. A large difference between the aligned and unaligned distributions implies that the representations use shared features for related sites within all the sequences of the family.

家族信息可能通过网络为家族MSA中对齐的不同序列中的位置分配相似表示而出现。使用Pfam(Bateman等，2013)中结构相关序列的MSA集合，我们比较了家族MSA中对齐的残基对表示的余弦相似性分布与未对齐残基对的余弦相似性背景分布。对齐和未对齐分布之间的较大差异表明，表示使用了家族所有序列中相关位点的共享特征。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | training | SSP | | | Contact | | |
| Family | Superfamily | Fold | Family | Superfamily | Fold |
| HMM Profile | - |  |  |  |  |  |  |
| CCMpred | - | - | - | - |  |  |  |
| Transformer-6 | UR50/S |  |  |  |  |  |  |
| Transformer-12 | UR50/S |  |  |  |  |  |  |
| Transformer-34 | (None) |  |  |  |  |  |  |
| Transformer-34 | UR100 |  |  |  |  |  |  |
| Transformer-34 | UR50/S |  |  |  |  |  |  |
| Transformer-34 | UR50/D |  |  |  |  |  |  |
| ESM-1b | UR50/S |  |  |  |  |  |  |
| 模型 | 训练 | SSP | | | 联系 | | |
| 家族 | 超家族 | 折叠 | 家族 | 超家族 | 折叠 |
| HMM 配置文件 | - |  |  |  |  |  |  |
| CCMpred | - | - | - | - |  |  |  |
| Transformer-6 | UR50/S |  |  |  |  |  |  |
| Transformer-12 | UR50/S |  |  |  |  |  |  |
| Transformer-34 | (无) |  |  |  |  |  |  |
| Transformer-34 | UR100 |  |  |  |  |  |  |
| Transformer-34 | UR50/S |  |  |  |  |  |  |
| Transformer-34 | UR50/D |  |  |  |  |  |  |
| ESM-1b | UR50/S |  |  |  |  |  |  |

Table 3. Linear projections. Five-fold cross validation experiment for generalization at the family, superfamily, and fold level. 8-class accuracy (secondary structure), Top-L long-range precision (contacts), mean and standard deviation across test sets for the five partitions. Minimal information about structure is present in representations prior to training. Information about secondary and tertiary structure emerges in representations as a result of unsupervised learning on sequences with the language modeling objective. Increasing diversity of sequences improves learning of structure. (Higher diversity UR50 datasets improve over UR100). Learned representations enable linear projections to generalize to held-out folds, outperforming projections of the sequence profile, and contacts identified by the CCMpred (Seemayer et al., 2014) implementation of direct coupling analysis.

表3. 线性投影。在家族、超家族和折叠层次上进行五折交叉验证实验以评估泛化能力。8类准确率(二级结构)、Top-L长程精度(接触)，以及五个分区的测试集上的均值和标准差。在训练前，表示中包含的结构信息极少。通过以语言建模为目标的无监督学习，序列中的二级和三级结构信息在表示中显现。序列多样性的增加提高了结构的学习效果。(更高多样性的UR50数据集优于UR100)。学习到的表示使得线性投影能够泛化到保留的折叠，优于序列谱的投影，以及通过CCMpred(Seemayer等，2014)实现的直接耦合分析所识别的接触。

Figure 3a depicts the distribution of cosine similarity values between aligned and unaligned positions within a representative family for the trained model and baselines. Unsupervised learning produces a marked shift between the distributions of aligned and unaligned pairs. Figure and Figure 3c indicate that these trends hold under the constraints that the residue pairs (1) share the same amino acid identity or (2) have different amino acid identities.

图3a展示了训练模型和基线在代表性家族中对齐和未对齐位置之间余弦相似度值的分布。无监督学习在对齐和未对齐对的分布之间产生了显著的变化。图 和图3c表明，在残基对(1)具有相同的氨基酸身份或(2)具有不同氨基酸身份的约束下，这些趋势仍然成立。

We estimate differences between the aligned and unaligned distributions across 128 Pfam families using the area under the ROC curve (AUC) as a metric of discriminative power between aligned and unaligned pairs. Table S1 shows a quantitative improvement in average AUC after unsupervised training, supporting the idea that self-supervision encodes information about the MSA of a sequence into its representation of the sequence.

我们使用ROC曲线下面积(AUC)作为对齐和未对齐对之间区分能力的度量，估计了128个Pfam家族中对齐和未对齐分布之间的差异。表S1显示了无监督训练后平均AUC的定量改进，支持了自监督将序列的多序列比对(MSA)信息编码到其表示中的观点。

# 5. Prediction of secondary structure and tertiary contacts

# 5. 二级结构和三级接触的预测

There is reason to believe that unsupervised learning will cause the model’s representations to contain structural information. The underlying structure of a protein is a hidden variable that influences the patterns observed in sequence data. For example local sequence variation depends on secondary structure (Levitt, 1978); and tertiary structure introduces higher order dependencies in the choices of amino acids at different sites within a protein (Marks et al., 2011; Anishchenko et al., 2017). While the model cannot observe protein structure directly, it observes patterns in the sequences of its training data that are determined by structure. In principle, the network could compress sequence variations by capturing commonality in structural elements across the data, thereby encoding structural information into the representations.

有理由相信无监督学习会使模型的表示包含结构信息。蛋白质的底层结构是一个隐藏变量，它影响序列数据中观察到的模式。例如，局部序列变异依赖于二级结构(Levitt，1978)；而三级结构在蛋白质不同位点的氨基酸选择中引入了更高阶的依赖性(Marks等，2011；Anishchenko等，2017)。虽然模型无法直接观察蛋白质结构，但它观察到了由其训练数据序列中由结构决定的模式。原则上，网络可以通过捕捉数据中结构元素的共性来压缩序列变异，从而将结构信息编码到表示中。

Linear projections We begin by identifying information about protein structure that is linearly encoded within the representations. The use of linear projections ensures that the information originates in the Transformer representations, enabling a direct inspection of the structural content of representations. By comparing representations of the Transformer before and after pre-training, we can identify the information that emerges as a result of the unsupervised learning.

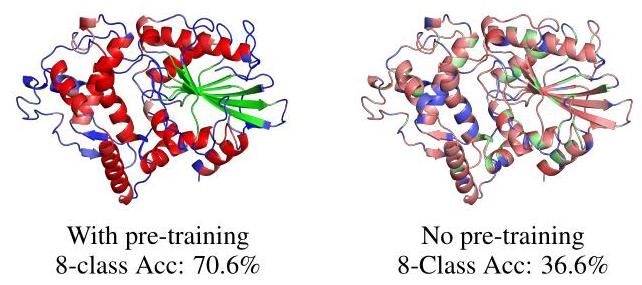
线性投影 我们首先识别表示中线性编码的蛋白质结构信息。使用线性投影确保信息源自Transformer表示，从而能够直接检查表示的结构内容。通过比较预训练前后的Transformer表示，我们可以识别出无监督学习所产生的新信息。

We perform a five-fold cross validation experiment to study generalization of structural information at the family, superfamily, and fold level. For each of the three levels, we construct a dataset of 15,297 protein structures using the SCOPe database. We partition the structures into five parts, splitting by family, superfamily, and fold accordingly. Fivefold cross validation is performed independently for each of the levels of structural hold-out.

我们进行了五折交叉验证实验，以研究在家族、超家族和折叠层次上结构信息的泛化能力。对于每个层次，我们使用SCOPe数据库构建了一个包含15,297个蛋白质结构的数据集。我们将结构分为五个部分，分别按家族、超家族和折叠进行划分。对于每个结构保留层次，独立进行五折交叉验证。

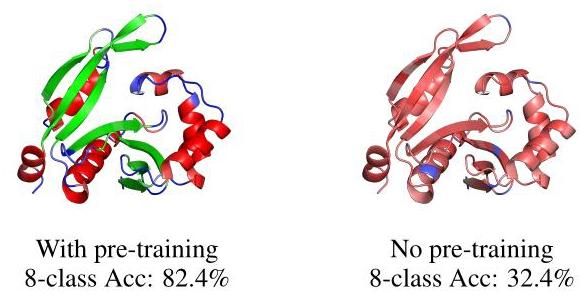
To detect information about secondary structure we fit a logistic regression to the hidden representations using the 8-class secondary structure labels. To detect information about tertiary structure, we fit two separate linear projections to the hidden representations of pairs of positions in the sequence, taking their dot product to regress a binary variable indicating whether the positions are in contact in the protein’s 3-dimensional structure. The neural representations are compared to (i) projections of the sequence profile;

为了检测二级结构信息，我们使用8类二级结构标签对隐藏表示进行逻辑回归。为了检测三级结构信息，我们对序列中位置对的隐藏表示进行两个独立的线性投影，取其点积来回归一个二元变量，该变量指示这些位置在蛋白质的三维结构中是否接触。神经表示与(i)序列谱的投影进行比较；



(a) d1nt4a\_ (Phosphoglycerate mutase-like fold)

(a) d1nt4a\_(磷酸甘油酸变位酶样折叠)



(b) d3wr7a\_ (Acyl-CoA N-acyltransferases fold)

(b) d3wr7a\_(酰基-CoA N-酰基转移酶折叠)

Figure 4. Secondary structure (linear projections). Example predictions for held-out folds. Unsupervised pre-training encodes secondary structure into representations. Following pre-training, linear projections recover secondary structure (left column). Without pre-training little information is recovered (right column). Colors indicate secondary structure class identified by the projection: helix (red), strand (green), and coil (blue). Color intensities indicate confidence. Representations from ESM-1b Transformer are used.

图4. 二级结构(线性投影)。对保留折叠的示例预测。无监督预训练将二级结构编码到表示中。预训练后，线性投影恢复了二级结构(左列)。没有预训练时，恢复的信息很少(右列)。颜色表示投影识别的二级结构类别:螺旋(红色)、链(绿色)和卷曲(蓝色)。颜色强度表示置信度。使用了ESM-1b Transformer的表示。

(ii) unsupervised contacts predicted by the CCMpred implementation (Seemayer et al., 2014) of direct coupling analysis. MSAs for the baselines are generated from the UniClust30 (Mirdita et al., 2017) database using 3 iterations of search by Hhblits. For secondary structure, we report 8-class accuracies. For contact precision we report Top-L long-range precision, i.e. the precision of the L (length of the protein) highest ranked predictions for contacts with sequence separation of at least 24 residues.

(ii) 由CCMpred实现(Seemayer等，2014)的直接耦合分析预测的无监督接触。基线的多序列比对(MSA)使用Hhblits进行3次搜索，从UniClust30(Mirdita等，2017)数据库中生成。对于二级结构，我们报告8类准确率。对于接触精度，我们报告Top-L长程精度，即对于序列分离至少24个残基的接触，L(蛋白质长度)个最高排名预测的精度。

Table 3 shows results of the cross validation experiments. Prior to pre-training, minimal information about secondary structure and contacts can be detected. After pre-training, projections recover information about secondary structure and long-range contacts that generalizes across families, superfamilies, and folds. Secondary structure prediction 8-class accuracy distributions (Figure S2), and long-range contact prediction Top-L precision distributions (Figure S3) demonstrate that pre-training produces an increase in structural information across the entire distribution of test domains. Table 3 shows that projections of the Transformer representations recover more structure than projections of the sequence profile. For long-range contacts, projections of the best Transformer models have higher precision than contacts predicted by CCMpred across all levels of structural generalization. As the level of structural split becomes more remote, there is little degradation for secondary structure, with performance at the family level similar to the fold level. For long-range contacts, while generalization is reduced at the fold level in comparison to the family level, the best models still capture more structure than the unsupervised baseline. Training with higher-diversity sequences (UR50 datasets) improves learning of both secondary structure and long-range contacts, with a more pronounced effect on long-range contacts.

表3展示了交叉验证实验的结果。在预训练之前，只能检测到关于二级结构和接触的极少信息。预训练后，投影恢复了关于二级结构和远程接触的信息，这些信息在家族、超家族和折叠之间具有普遍性。二级结构预测的8类准确率分布(图S2)和远程接触预测的Top-L精度分布(图S3)表明，预训练在整个测试域分布中增加了结构信息。表3显示，Transformer表示的投影比序列谱的投影恢复了更多的结构。对于远程接触，最佳Transformer模型的投影在所有结构泛化水平上的精度都高于CCMpred预测的接触。随着结构分割水平变得更加遥远，二级结构的性能几乎没有下降，家族水平的性能与折叠水平相似。对于远程接触，尽管与家族水平相比，折叠水平的泛化有所减少，但最佳模型仍然比无监督基线捕获了更多的结构。使用更高多样性的序列(UR50数据集)进行训练，提高了二级结构和远程接触的学习效果，对远程接触的影响更为显著。

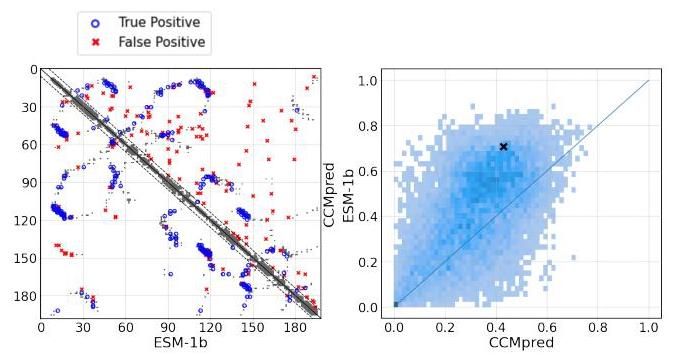


Figure 5. Residue-residue contacts (linear projections). Left: Top-L predictions for fold level held-out example d1n3ya\_, with vWA-like fold. True positives in blue, false positives in red, superimposed on ground-truth contact map in grey. ESM-1b Transformer projections below the diagonal, CCMpred predictions above the diagonal. Right: Precision distribution (Top-L long-range) comparing ESM-1b projections with CCMpred across all domains in the five test partitions with structural hold-out at the fold level. Visualized domain marked by .

图5. 残基-残基接触(线性投影)。左:折叠水平保留示例d1n3ya\_的Top-L预测，具有类似vWA的折叠。真阳性为蓝色，假阳性为红色，叠加在灰色背景的真实接触图上。ESM-1b Transformer投影在对角线下方，CCMpred预测在对角线上方。右:精度分布(Top-L远程)，比较ESM-1b投影与CCMpred在五个测试分区中所有域的折叠水平结构保留情况。可视化域由 标记。

Figure 4 visualizes three-class secondary structure projections for two domains belonging to held-out folds. Prior to pre-training, projections produce an incoherent prediction of secondary structure. After pre-training projections recover a coherent prediction with most errors occurring at the boundaries of secondary structure regions. Figure 5 compares a projected contact map to predictions from CCMpred. Transformer projections recover complex contact patterns, including long-range contacts. Further visualizations of projected contacts for eight randomly selected test proteins are shown in Figure S7.

图4展示了属于保留折叠的两个域的三类二级结构投影。在预训练之前，投影产生了不连贯的二级结构预测。预训练后，投影恢复了连贯的预测，大多数错误发生在二级结构区域的边界。图5将投影的接触图与CCMpred的预测进行了比较。Transformer投影恢复了复杂的接触模式，包括远程接触。图S7展示了八个随机选择的测试蛋白质的投影接触的进一步可视化。

Deep neural network We train deep neural networks to predict secondary structure and contacts from the representations. We choose state-of-the-art neural architectures for both tasks. These downstream models are trained with a supervised loss to predict either the secondary structure or contact map from the pre-trained representations. The architecture of the downstream model is kept fixed across experiments with different representations and baselines to enable comparison. bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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.

深度神经网络 我们训练深度神经网络从表示中预测二级结构和接触。我们为这两个任务选择了最先进的神经架构。这些下游模型通过监督损失进行训练，以从预训练表示中预测二级结构或接触图。下游模型的架构在不同表示和基线的实验中保持不变，以便进行比较。bioRxiv预印本 doi: https://doi.org/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本根据aCC-BY-NC-ND 4.0国际许可提供。

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Training | CB513 | CASP13 |
| HMM Profile | |  |  |
| LSTM (S) | UR50/S |  |  |
| LSTM (L) | UR50/S |  |  |
| Transf-6 | UR50/S |  |  |
| Transf-12 | UR50/S |  |  |
| Transf-34 | None |  |  |
| Transf-34 | UR100 |  |  |
| Transf-34 | UR50/S |  |  |
| Transf-34 | UR50/D |  |  |
| ESM-1B | UR50/S |  |  |
| 模型 | 训练 | CB513 | CASP13 |
| HMM 轮廓 | |  |  |
| LSTM (S) | UR50/S |  |  |
| LSTM (L) | UR50/S |  |  |
| Transf-6 | UR50/S |  |  |
| Transf-12 | UR50/S |  |  |
| Transf-34 | 无 |  |  |
| Transf-34 | UR100 |  |  |
| Transf-34 | UR50/S |  |  |
| Transf-34 | UR50/D |  |  |
| ESM-1B | UR50/S |  |  |

Table 4. Eight-class secondary structure prediction accuracy on the CB513 and CASP13 test sets. A fixed neural architecture is trained to predict the secondary structure label from the language model representation of the input sequence. The Transformer has higher performance than the comparable LSTM baselines. Pretraining with the high-diversity UR50 datasets increases accuracy significantly. Features from ESM-1b Transformer are competitive with HMM profiles for supervised secondary structure prediction.

表4. CB513和CASP13测试集上的八类二级结构预测准确率。使用固定的神经网络架构从输入序列的语言模型表示中预测二级结构标签。Transformer的性能优于可比的LSTM基线。使用高多样性UR50数据集进行预训练显著提高了准确率。ESM-1b Transformer的特征在监督二级结构预测中与HMM谱图具有竞争力。

To predict secondary structure we replace the linear layer with a deep neural network, using the model architecture introduced by the Netsurf method (Klausen et al., 2019). For tertiary structure, we predict the binary contact map from the hidden representation of the sequence. We use a dilated convolutional residual network similar to recent state-of-the-art methods for tertiary structure prediction (Xu, 2018; Jones & Kandathil, 2018; Senior et al., 2018).

为了预测二级结构，我们使用Netsurf方法(Klausen等，2019)引入的模型架构，将线性层替换为深度神经网络。对于三级结构，我们从序列的隐藏表示中预测二元接触图。我们使用类似于最近最先进的三级结构预测方法的扩张卷积残差网络(Xu，2018；Jones & Kandathil，2018；Senior等，2018)。

Table 4 compares the representations for secondary structure prediction. We evaluate models on the CB513 test set (Cuff & Barton, 1999) and the CASP13 domains (Kryshtafovych et al., 2019). For comparison we also re-implement the Netsurf method. The models are trained on the Netsurf training dataset which applies a 25% sequence identity holdout with CB513, and a temporal hold-out with CASP13. The Transformer features are compared before and after unsupervised pre-training to features from the LSTM baselines. They are also compared to the HMM profiles used by Netsurf. The best Transformer features (71.6%) match the performance of the HMM profiles (71.2%), and exceed the published performance of RaptorX (70.6)% on the same benchmark (Klausen et al., 2019), implying that protein langauge models can produce features that are directly competitive with sequence profiles for secondary structure prediction.

表4比较了二级结构预测的表示。我们在CB513测试集(Cuff & Barton，1999)和CASP13域(Kryshtafovych等，2019)上评估模型。为了比较，我们还重新实现了Netsurf方法。模型在Netsurf训练数据集上进行训练，该数据集对CB513应用了25%的序列同一性保留，对CASP13应用了时间保留。Transformer特征在无监督预训练前后与LSTM基线特征进行了比较。它们还与Netsurf使用的HMM谱图进行了比较。最佳Transformer特征(71.6%)与HMM谱图(71.2%)的性能相当，并超过了RaptorX在同一基准上发布的性能(70.6%)(Klausen等，2019)，这表明蛋白质语言模型可以生成与序列谱图直接竞争的特征，用于二级结构预测。

Table 5 shows performance of the various representations for predicting Top-L long-range contacts across a panel of benchmarks using the RaptorX train set (Wang et al., 2017). For comparison we train the same architecture using features from RaptorX (Wang et al., 2017; Xu, 2018). The Test (Wang et al., 2017) and CASP11 (Moult et al., 2016) test sets evaluate with sequence identity hold-out at 25%; the CASP12 (Moult et al., 2018) test set implements a temporal hold-out with the structural training data; and the CASP13 (Kryshtafovych et al., 2019) experiment implements a full temporal hold-out of both the pre-training and training data. For contact prediction, the best features from representation learning do not achieve comparable performance to the state-of-the-art RaptorX features (50.2 vs 59.4 respectively on the RaptorX test set).

表5显示了使用RaptorX训练集(Wang等，2017)在一系列基准上预测Top-L长程接触的各种表示的性能。为了比较，我们使用RaptorX的特征(Wang等，2017；Xu，2018)训练了相同的架构。测试集(Wang等，2017)和CASP11(Moult等，2016)测试集以25%的序列同一性保留进行评估；CASP12(Moult等，2018)测试集对结构训练数据应用了时间保留；CASP13(Kryshtafovych等，2019)实验对预训练和训练数据都应用了完全的时间保留。对于接触预测，表示学习的最佳特征未能达到与最先进的RaptorX特征相当的性能(在RaptorX测试集上分别为50.2和59.4)。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | CASP | | |
| Model | training | Test | 11 | 12 | 13 |
| LSTM (S) | UR50/S | 24.1 | 23.6 | 19.9 | 15.3 |
| LSTM (L) | UR50/S | 27.8 | 26.4 | 24.0 | 16.4 |
| Transf-6 | UR50/S | 30.2 | 29.9 | 25.3 | 19.8 |
| Transf-12 | UR50/S | 37.7 | 33.6 | 27.8 | 20.7 |
| Transf-34 | (None) | 16.3 | 17.7 | 14.8 | 13.3 |
| Transf-34 | UR100 | 32.7 | 28.9 | 24.3 | 19.1 |
| Transf-34 | UR50/S | 50.2 | 42.8 | 34.7 | 30.1 |
| Transf-34 | UR50/D | 50.0 | 43.0 | 33.6 | 28.0 |
| ESM-1b | UR50/S | 56.9 | 47.4 | 42.7 | 35.9 |
|  |  |  | CASP | | |
| 模型 | 训练 | 测试 | 11 | 12 | 13 |
| LSTM (S) | UR50/S | 24.1 | 23.6 | 19.9 | 15.3 |
| LSTM (L) | UR50/S | 27.8 | 26.4 | 24.0 | 16.4 |
| Transf-6 | UR50/S | 30.2 | 29.9 | 25.3 | 19.8 |
| Transf-12 | UR50/S | 37.7 | 33.6 | 27.8 | 20.7 |
| Transf-34 | (无) | 16.3 | 17.7 | 14.8 | 13.3 |
| Transf-34 | UR100 | 32.7 | 28.9 | 24.3 | 19.1 |
| Transf-34 | UR50/S | 50.2 | 42.8 | 34.7 | 30.1 |
| Transf-34 | UR50/D | 50.0 | 43.0 | 33.6 | 28.0 |
| ESM-1b | UR50/S | 56.9 | 47.4 | 42.7 | 35.9 |

Table 5. Top-L long-range contact precision. A deep dilated convolutional residual network is trained to predict contacts using the representations from the pre-trained language model. The pre-trained Transformer representations outperform the LSTM representations in all cases. Pre-training on the high-diversity UR50 datasets boosts precision of representations over pre-training on UR100. High-capacity Transformers (34 layer) outperform lower capacity models (6/12 layer).

表5. Top-L长程接触精度。训练了一个深度扩张卷积残差网络，使用预训练语言模型的表示来预测接触。在所有情况下，预训练的Transformer表示都优于LSTM表示。在高多样性UR50数据集上进行预训练比在UR100上进行预训练提高了表示的精度。高容量Transformer(34层)优于低容量模型(6/12层)。

In the secondary structure benchmarks Transformer representations produce higher accuracy than the LSTM baselines with comparable numbers of parameters. For contact prediction Transformer reprentations yield higher precision than LSTMs, with even the smallest Transformer representations exceeding LSTMs with more parameters. Diversity in the pre-training data also has a strong effect, with the high-diversity datasets providing significant improvements over the low-diversity dataset. Relative performance of the representations is consistent across all four of the contact benchmarks using different hold-out methodology.

在二级结构基准测试中，Transformer表示在参数数量相当的情况下比LSTM基线产生更高的准确性。对于接触预测，Transformer表示比LSTM产生更高的精度，即使是最小的Transformer表示也超过了具有更多参数的LSTM。预训练数据的多样性也有很强的影响，高多样性数据集比低多样性数据集提供了显著的改进。在使用不同保留方法的四个接触基准测试中，表示的相对性能是一致的。

Relationship between language modeling and structure learning To investigate the relationship between the language modeling objective and information about structure in the model, linear projections for secondary structure and contacts are fit using the representations from Transformer models taken from checkpoints across their pre-training trajectories. We use the Transformers trained on UR50/S. We bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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. fit the projections and evaluate with the train and test split implemented by the first partition of the fold level structural hold-out dataset. For each model, Figure 6 shows a linear relationship between the language modeling objective and information about structure, which is maintained over the course of pre-training. The linear fit is close to ideal for both secondary structure and contacts. A similar experiment is also performed for secondary structure with a deep neural network instead of linear projection, using the Netsurf training sequences and CB513 test set. A linear relationship between secondary structure accuracy and language modeling ECE is also observed for the deep prediction head (Figure S4). Thus, for a given model and pre-training dataset, language modeling fidelity measured by ECE is a good proxy for the structural content of the representations. Since performance on the language modeling objective improves with model capacity, this suggests further scale may improve results on structure prediction tasks.

语言建模与结构学习之间的关系 为了研究语言建模目标与模型中结构信息之间的关系，使用从预训练轨迹中提取的Transformer模型的表示来拟合二级结构和接触的线性投影。我们使用在UR50/S上训练的Transformer。我们拟合投影并使用由折叠级别结构保留数据集的第一个分区实现的训练和测试分割进行评估。对于每个模型，图6显示了语言建模目标与结构信息之间的线性关系，这种关系在预训练过程中保持不变。对于二级结构和接触，线性拟合都接近理想。还使用Netsurf训练序列和CB513测试集对二级结构进行了类似的实验，使用深度神经网络代替线性投影。对于深度预测头，也观察到二级结构准确性与语言建模ECE之间的线性关系(图S4)。因此，对于给定的模型和预训练数据集，通过ECE测量的语言建模保真度是表示结构内容的良好代理。由于语言建模目标的性能随着模型容量的提高而提高，这表明进一步的扩展可能会改善结构预测任务的结果。

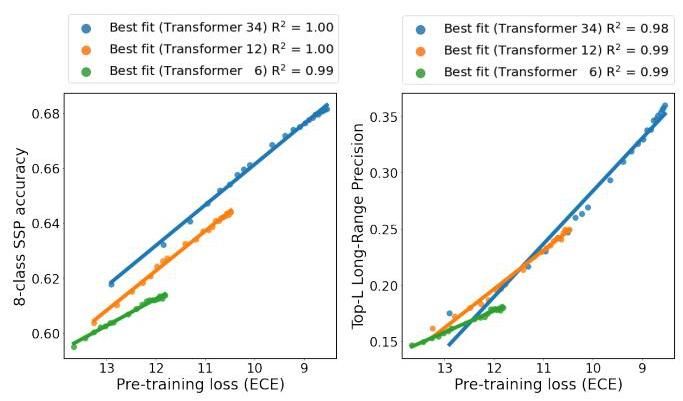


Figure 6. Relationship between the language modeling objective and structure learning. Eight-class secondary structure prediction accuracy (left) and contact prediction Top-L long-range precision (right) both as a function of pre-training ECE. Performance is evaluated on held-out folds. Linear projections are fit using model checkpoints over the course of pre-training on UR50/S. The linear relationship for each model indicates that for a given model and pre-training dataset, the language modeling ECE is a good proxy for the structural content of the representations. Improvement of the model’s ECE leads to an increase in information about structure. This establishes a link between the language modeling objective and unsupervised structure learning.

图6. 语言建模目标与结构学习之间的关系。八类二级结构预测准确性(左)和接触预测Top-L长程精度(右)均作为预训练ECE的函数。性能在保留的折叠上进行评估。使用在UR50/S上预训练过程中的模型检查点拟合线性投影。每个模型的线性关系表明，对于给定的模型和预训练数据集，语言建模ECE是表示结构内容的良好代理。模型ECE的提高导致结构信息的增加。这建立了语言建模目标与无监督结构学习之间的联系。

Single versus multi-family pre-training We compare training across evolutionary statistics to training on single protein families. We pre-train separate 12-layer Transformer models on the Pfam multiple sequence alignments of the three most common domains in nature longer than 100 amino acids, the ATP-binding domain of the ABC transporters, the protein kinase domain, and the response regulator receiver domain. We test the ability of models trained on one protein family to generalize secondary structure information within-family and out-of-family by evaluating on sequences with ground truth labels from the family the model was trained on or from the alternate families. The models are evaluated using linear projections. In all cases, the model trained on within-family sequences has higher accuracy than models trained on out-of-family sequences (Table S2), indicating poor generalization when training on single MSA families. More significantly, the model trained across the full UniParc sequence diversity has a higher accuracy than the single-family model accuracies, even on the same-family evaluation dataset. This suggests that the representations learned from the full dataset are generalizing information about secondary structure learned outside the sequence family.

单家族与多家族预训练 我们比较了跨进化统计数据的训练与单一蛋白质家族上的训练。我们在自然界中最常见的三个长度超过100个氨基酸的结构域的Pfam多序列比对中预训练了单独的12层Transformer模型，这些结构域包括ABC转运蛋白的ATP结合域、蛋白激酶域和响应调节器接收域。我们通过评估来自模型训练家族或来自其他家族的具有真实标签的序列，测试了在一个蛋白质家族上训练的模型在家族内和家族外泛化二级结构信息的能力。使用线性投影对模型进行评估。在所有情况下，在家族内序列上训练的模型比在家族外序列上训练的模型具有更高的准确性(表S2)，表明在单一MSA家族上训练时泛化能力较差。更重要的是，在完整UniParc序列多样性上训练的模型比单一家族模型的准确性更高，即使在相同家族的评估数据集上也是如此。这表明从完整数据集中学习到的表示正在泛化序列家族之外学习的二级结构信息。

|  |  |  |
| --- | --- | --- |
| Features | CB513 | CASP13 |
| RaptorX | 70.6 |  |
| Netsurf | 72.1 | 74 |
| (a) Netsurf (reimpl.) |  |  |
| (b) +direct |  |  |
| (c) |  |  |
| 特性 | CB513 | CASP13 |
| RaptorX | 70.6 |  |
| Netsurf | 72.1 | 74 |
| (a) Netsurf (重新实现) |  |  |
| (b) +直接 |  |  |
| (c) |  |  |

Table 6. Feature combination (secondary structure prediction). Eight-class accuracy. The language model improves state-of-the-art features for secondary structure prediction. Features from a reimplementation of Netsurf (Klausen et al., 2019) are combined with 34-layer Transformer (UR50/S) embeddings using a two layer BiLSTM architecture. (a) Performance of Netsurf features alone. (b) Direct adds the Transformer representation of the input sequence. (c) adds the average of Transformer features for each position in the MSA of the input sequence. Results exceed those for state-of-the-art methods RaptorX (70.6%) and Netsurf (72.1%) on the CB513 test set, and for Netsurf (74.0%) on the CASP13 evaluation set used here.

表6. 特征组合(二级结构预测)。八类准确率。语言模型改进了二级结构预测的最先进特征。来自Netsurf重新实现(Klausen等，2019)的特征与34层Transformer(UR50/S)嵌入相结合，使用双层BiLSTM架构。(a) 单独使用Netsurf特征的性能。(b) 直接添加输入序列的Transformer表示。(c) 添加输入序列MSA中每个位置的Transformer特征的平均值。结果超过了CB513测试集上最先进方法RaptorX(70.6%)和Netsurf(72.1%)的结果，以及在此使用的CASP13评估集上Netsurf(74.0%)的结果。

# 6. Feature combination

# 6. 特征组合

Features discovered by unsupervised protein language modeling can be combined with state-of-the-art features to improve them further. Current state-of-the-art methods use information derived from MSAs. We combine this information with features from the Transformer model.

通过无监督蛋白质语言建模发现的特征可以与最先进的特征相结合，以进一步改进它们。当前最先进的方法使用来自MSA的信息。我们将这些信息与Transformer模型的特征相结合。

We explore three approaches for incorporating information from representation learning. For each input sequence : (i) direct uses the final hidden representation from the Transformer directly; (ii) avg takes the average of the final hidden representation at each position across the sequences from the MSA of ; (iii) cov produces features for each pair of positions, using the uncentered covariance across sequences from the MSA of , after dimensionality reduction of the final hidden representations by PCA. Note that (i) and (ii) produce features for each position in , while (iii) produces bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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. features for each pair of positions.

我们探索了三种结合表示学习信息的方法。对于每个输入序列 :(i) 直接使用Transformer的最终隐藏表示；(ii) avg取 的MSA序列中每个位置的最终隐藏表示的平均值；(iii) cov为每对位置生成特征，使用 的MSA序列中未中心化的协方差，在通过PCA对最终隐藏表示进行降维后。注意，(i)和(ii)为 中的每个位置生成特征，而(iii)为每对位置生成特征。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Test | CASP11 | CASP12 | CASP13 |
| #domains | 500 | 105 | 55 | 34 |
| (a) RaptorX |  |  |  |  |
| (b) +direct |  |  |  |  |
| (c) |  |  |  |  |
| (d) +cov |  |  |  |  |
|  | 测试 | CASP11 | CASP12 | CASP13 |
| #域 | 500 | 105 | 55 | 34 |
| (a) RaptorX |  |  |  |  |
| (b) +直接 |  |  |  |  |
| (c) |  |  |  |  |
| (d) +覆盖 |  |  |  |  |

Table 7. Feature combination (contact prediction). Top-L long-range contact precision. The language model improves state-of-the-art features for contact prediction. A deep ResNet with fixed architecture is trained on each feature set to predict binary contacts. (a) performance of state-of-the-art RaptorX (Xu, 2018) features including PSSM, predicted secondary structure, predicted accessibility, pairwise APC-corrected Potts model couplings and mutual information, and a pairwise contact potential. (b) Adds Transformer representation of the input sequence to the feature set. (c) Adds the average Transformer representation at each position of the MSA. (d) Adds the uncentered covariance over the MSA of a low-dimenensional projection of the Transformer features. Features are from the 34-layer Transformer pre-trained on UR50/S.

表7. 特征组合(接触预测)。Top-L长程接触精度。语言模型提升了接触预测的最先进特征。在每个特征集上训练一个固定架构的深度ResNet以预测二元接触。(a) 最先进的RaptorX(Xu, 2018)特征的表现，包括PSSM、预测的二级结构、预测的可及性、成对APC校正的Potts模型耦合和互信息，以及成对接触势。(b) 将输入序列的Transformer表示添加到特征集中。(c) 将MSA每个位置的平均Transformer表示添加到特征集中。(d) 将Transformer特征的低维投影在MSA上的未中心化协方差添加到特征集中。特征来自在UR50/S上预训练的34层Transformer。

Secondary structure Current state-of-the-art methods for secondary structure prediction have high accuracies for the eight-class prediction problem. We investigate whether performance can be improved by combining Transformer features with sequence profiles. Table 6 shows that combining the representations with profiles further boosts accuracy, resulting in state-of-the-art performance on secondary structure prediction.

二级结构 当前最先进的二级结构预测方法在八类预测问题上具有高精度。我们研究了通过将Transformer特征与序列谱结合是否可以提高性能。表6显示，将表示与谱结合进一步提高了精度，从而在二级结构预测上达到了最先进的性能。

We establish a baseline of performance by reimplementing the Klausen et al. (2019) method using the same features, resulting in an accuracy of (vs. published performance of ) on the CB513 test set. Then we add the the Transformer features using the direct and avg combination methods; these achieve and absolute improvement in accuracy respectively. This suggests that the Transformer features contain information not present in the MSA-derived features.

我们通过重新实现Klausen等人(2019)的方法，使用相同的特征，在CB513测试集上建立了性能基线，结果为 (与已发布的性能 相比)。然后我们使用直接和平均组合方法添加Transformer特征；这些方法分别在精度上实现了 和 的绝对提升。这表明Transformer特征包含了MSA衍生特征中不存在的信息。

Residue-residue contacts Deep neural networks have enabled recent breakthroughs in the prediction of protein contacts and tertiary structure (Xu, 2018; Senior et al., 2018). State-of-the-art neural networks for tertiary structure and contact prediction use deep residual architectures with two-dimensional convolutions over pairwise feature maps to output a contact prediction or distance potential for each pair of residues (Wang et al., 2017; Xu, 2018; Senior et al., 2018).

残基-残基接触 深度神经网络在蛋白质接触和三级结构预测方面取得了最近的突破(Xu, 2018; Senior等, 2018)。用于三级结构和接触预测的最先进神经网络使用深度残差架构，在成对特征图上进行二维卷积，以输出每对残基的接触预测或距离势(Wang等, 2017; Xu, 2018; Senior等, 2018)。

A variety of input features, training datasets, and supervision signals are used in state-of-the-art methods. To make a controlled comparison, we fix a standard architecture, training dataset, multiple sequence alignments, and set of base input features for all experiments, to which we add pre-trained features from the Transformer model. For the base features we use the RaptorX feature set which includes PSSM, 3- state secondary structure prediction, one-hot embedding of sequence, APC corrected Potts model couplings, mutual information, pairwise contact potential, and predicted accessibility. RaptorX was the winning method for contact prediction in the CASP12 and CASP13 competitions (Xu, 2018). The training and evaluation sets are the same as used in the previous section.

最先进的方法使用了多种输入特征、训练数据集和监督信号。为了进行受控比较，我们为所有实验固定了标准架构、训练数据集、多序列比对和基本输入特征集，并添加了来自Transformer模型的预训练特征。对于基本特征，我们使用了RaptorX特征集，包括PSSM、三态二级结构预测、序列的独热编码、APC校正的Potts模型耦合、互信息、成对接触势和预测的可及性。RaptorX是CASP12和CASP13竞赛中接触预测的获胜方法(Xu, 2018)。训练和评估集与前一节相同。

Table 7 indicates that addition of Transformer features from the 34-layer model trained on UR50/S consistently produces an improvement across the test sets. The table shows Top-L long-range precisions reporting mean and standard deviation over 5 different model seeds. Direct gives a modest improvement on some test sets. Avg improves over direct, and cov provides further gains. For example, cov produces an absolute improvement of 3.9% on the RaptorX test set, and 1.8% improvement on the CASP13 test set evaluated with temporal hold-outs on both fine-tuning and pre-training data. Additional results and metrics for contact prediction are reported in Table S3.

表7表明，添加来自在UR50/S上训练的34层模型的Transformer特征在测试集上一致地产生了改进。该表显示了Top-L长程精度，报告了5个不同模型种子的平均值和标准差。直接方法在某些测试集上略有改进。平均方法优于直接方法，协方差方法进一步提高了性能。例如，协方差方法在RaptorX测试集上实现了3.9%的绝对提升，在CASP13测试集上实现了1.8%的提升，评估时在微调和预训练数据上使用了时间保留。接触预测的更多结果和指标见表S3。

# 7. Prediction of mutational effects

# 7. 突变效应预测

The mutational fitness landscape provides deep insight into biology. Coupling next generation sequencing with a mutagenesis screen allows parallel readout of tens of thousands of variants of a single protein (Fowler & Fields, 2014). The detail and coverage of these experiments provides a view into the mutational fitness landscape of individual proteins, giving quantitative relationships between sequence and protein function. We adapt the Transformer protein language model to predict the quantitative effect of mutations.

突变适应度景观提供了对生物学的深入洞察。将下一代测序与诱变筛选相结合，可以并行读取单个蛋白质的数万个变体(Fowler & Fields, 2014)。这些实验的细节和覆盖范围提供了单个蛋白质的突变适应度景观视图，给出了序列与蛋白质功能之间的定量关系。我们调整了Transformer蛋白质语言模型以预测突变的定量效应。

First we investigate intra-protein variant effect prediction, where a limited sampling of mutations is used to predict the effect of unobserved mutations. This setting has utility in protein engineering applications (Yang et al., 2019). We evaluate the representations on two deep mutational scanning datasets used by recent state-of-the-art methods for variant effect prediction, Envision (Gray et al., 2018)

首先，我们研究了蛋白质内变体效应预测，其中使用有限的突变样本来预测未观察到的突变的影响。这种设置在蛋白质工程应用中具有实用性(Yang等, 2019)。我们在两个深度突变扫描数据集上评估了表示，这些数据集被最近最先进的变体效应预测方法Envision(Gray等, 2018)使用。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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. and DeepSequence (Riesselman et al., 2018). Collectively the data includes over 700,000 variant effect measurements from over 100 large-scale experimental mutagenesis datasets.

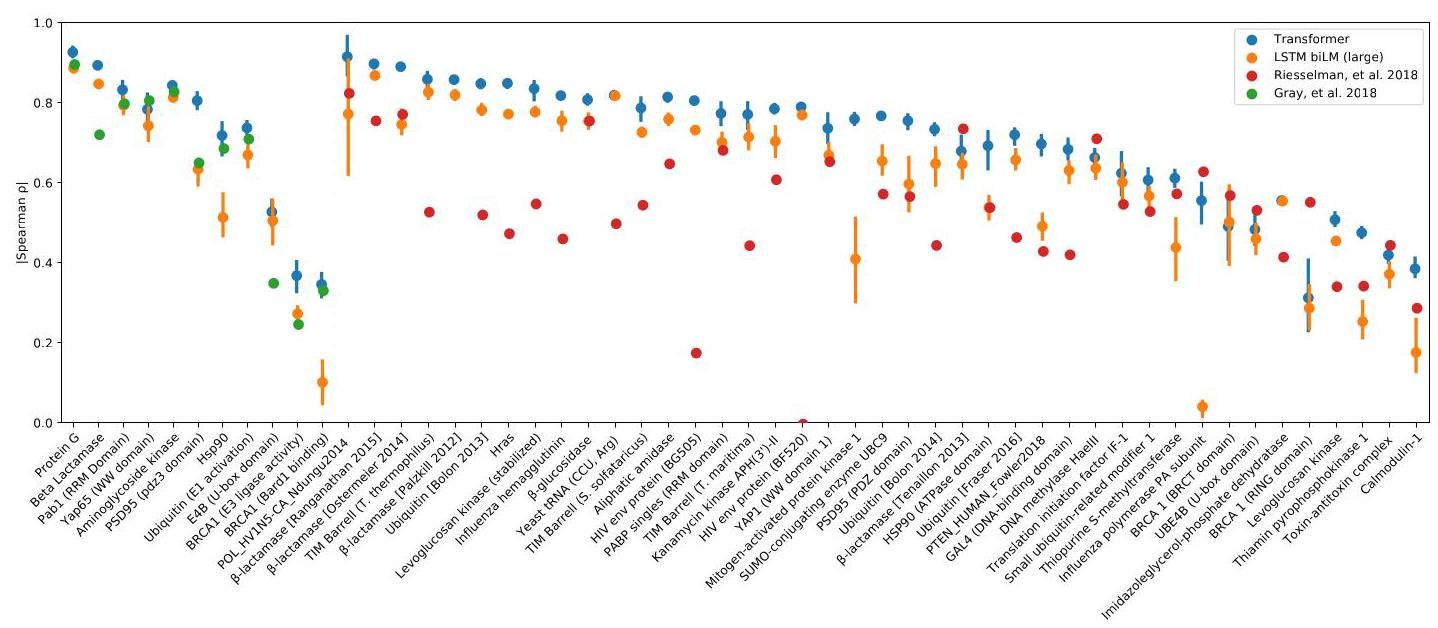


Figure 7. Representation learning enables state-of-the-art supervised prediction of the quantitative effect of mutations. Left panel: Envision dataset (Gray et al., 2018); right panel: DeepSequence dataset (Riesselman et al., 2018). Transformer representations (34-layer, UR50/S) are compared to the LSTM bidirectional language model (large model, UR50/S). The result of five-fold cross validation is reported for each protein. For each partition, supervised fine-tuning is performed on of the mutational data for the protein, and results are evaluated on the remaining . Transformer representations outperform baseline LSTM representations on both datasets. State-of-the-art methods are also shown for each dataset. Gray et al. (2018) is a supervised method using structural, evolutionary, and biochemical features, trained with the same protocol as used for the Transformer. Riesselman et al. (2018) is an unsupervised method trained on the MSA of each protein.

图7. 表示学习实现了对突变定量效应的最先进监督预测。左面板:Envision数据集(Gray等，2018)；右面板:DeepSequence数据集(Riesselman等，2018)。Transformer表示(34层，UR50/S)与LSTM双向语言模型(大模型，UR50/S)进行比较。每个蛋白质的五折交叉验证结果均被报告。对于每个分区，对蛋白质的突变数据进行 的监督微调，并在剩余的 上进行评估。Transformer表示在两个数据集上均优于基线LSTM表示。每个数据集的最先进方法也显示出来。Gray等(2018)是一种使用结构、进化和生化特征的监督方法，训练协议与Transformer相同。Riesselman等(2018)是一种无监督方法，训练于每个蛋白质的多序列比对(MSA)。

Fine-tuning the Transformer yields a mutational effect predictor that is comparable to the results of Envision. Envision (Gray et al., 2018) relies on protein structural and evolutionary features to generalize. We assess whether the Transformer can achieve similar generalization results, without direct access to structural features. The same methodology for partitioning data for training and evaluation is used as in Gray et al. (2018) to allow a comparison of the results. We use the 34-layer Transformer trained on UR50/S. Figure 7 shows the fine-tuned Transformer exceeds the performance of Envision on 10 of the 12 proteins. For each protein a fraction of the data is used for training and the remaining data is used for testing. We report mean and standard deviations for five-fold cross-validation in Table S5. Results varying the fraction of data that is used for training are reported in Figure S5.

微调Transformer得到的突变效应预测器与Envision的结果相当。Envision(Gray等，2018)依赖于蛋白质结构和进化特征进行泛化。我们评估Transformer是否能在不直接访问结构特征的情况下实现类似的泛化结果。使用与Gray等(2018)相同的数据划分方法进行训练和评估，以便比较结果。我们使用在UR50/S上训练的34层Transformer。图7显示，微调后的Transformer在12个蛋白质中的10个上超过了Envision的性能。对于每个蛋白质，使用 的数据进行训练，剩余数据用于测试。我们在表S5中报告了五折交叉验证的均值和标准差。图S5报告了不同训练数据比例的结果。

We also evaluate using the same five-fold cross validation methodology on the deep mutational scanning experiments assembled for DeepSequence (Riesselman et al., 2018). The fine-tuned Transformer model outperforms the fine-tuned LSTM baselines. While not directly comparable, we also include the performance of the original DeepSequence method which is unsupervised, and represents state-of-the-art for this dataset.

我们还使用相同的五折交叉验证方法评估了为DeepSequence(Riesselman等，2018)组装的深度突变扫描实验。微调后的Transformer模型优于微调的LSTM基线。虽然不直接可比，但我们还包括了原始DeepSequence方法的性能，该方法是无监督的，代表了该数据集的最先进水平。

Generalization to a new fitness landscape We analyze the Transformer’s ability to generalize to the fitness landscape of a new protein. Following the protocol introduced in Envision, we use a leave-one-out analysis: to evaluate performance on a given protein, we train on data from the remaining proteins and test on the held-out protein. Figure S6 shows that the Transformer’s predictions from raw sequences perform better than Envision on 5 of the 9 tasks.

泛化到新的适应度景观 我们分析了Transformer泛化到新蛋白质适应度景观的能力。遵循Envision引入的协议，我们使用留一法分析:为了评估给定蛋白质的性能，我们在剩余的 个蛋白质的数据上进行训练，并在保留的蛋白质上进行测试。图S6显示，Transformer从原始序列进行的预测在9个任务中的5个上优于Envision。

# 8. Related Work

# 8. 相关工作

Contemporaneously with the preprint of this work, Rives et al. (2019), related preprints Alley et al. (2019) and Heinzinger et al. (2019), also proposed protein language modeling, albeit at a smaller scale. These works, along with Rao et al. (2019), evaluated on a variety of downstream tasks. Rives et al. (2019) first proposed protein language modeling with Transformers. Alley et al. (2019) and Heinzinger et al. (2019) train LSTMs on UniRef50. Rao et al. (2019) trained a 12-layer Transformer model (38M parameters) on Pfam (Bateman et al., 2013). The baselines in this paper are bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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. comparable to these models. The large Transformer models trained in this paper are considerably larger than in these related works.

与本工作的预印本同时，Rives等(2019)、相关预印本Alley等(2019)和Heinzinger等(2019)也提出了蛋白质语言建模，尽管规模较小。这些工作与Rao等(2019)一起，在各种下游任务上进行了评估。Rives等(2019)首次提出了使用Transformer进行蛋白质语言建模。Alley等(2019)和Heinzinger等(2019)在UniRef50上训练LSTM。Rao等(2019)在Pfam(Bateman等，2013)上训练了一个12层Transformer模型(3800万参数)。本文中的基线模型与这些模型相当。本文中训练的大型Transformer模型比这些相关工作中的模型要大得多。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Training | Params |  | SSP | Contact |
|  |  | 18M | .527 | 58.4 | 21.9 |
| Seq |  | 93M | .545 | 62.1 | 29.0 |
| Tape3† |  | 38M | .581 | 58.0 | 23.2 |
| LSTM (S) | UR50/S | 28.4M | .558 | 60.4 | 24.1 |
| LSTM (L) | UR50/S | 113.4M | .574 | 62.4 | 27.8 |
| Transformer-6 | UR50/S | 42.6M | .653 | 62.0 | 30.2 |
| Transformer-12 | UR50/S | 85.1M | .639 | 65.4 | 37.7 |
| Transformer-34 | UR100 | 669.2M | .599 | 64.3 | 32.7 |
| Transformer-34 | UR50/S | 669.2M | .639 | 69.2 | 50.2 |
| ESM-1b | UR50/S | 652.4M | .532 | 71.6 | 56.9 |
| 模型 | 训练 | 参数 |  | SSP | 联系 |
|  |  | 18M | .527 | 58.4 | 21.9 |
| 序列 |  | 93M | .545 | 62.1 | 29.0 |
| Tape3† |  | 38M | .581 | 58.0 | 23.2 |
| LSTM (S) | UR50/S | 28.4M | .558 | 60.4 | 24.1 |
| LSTM (L) | UR50/S | 113.4M | .574 | 62.4 | 27.8 |
| Transformer-6 | UR50/S | 42.6M | .653 | 62.0 | 30.2 |
| Transformer-12 | UR50/S | 85.1M | .639 | 65.4 | 37.7 |
| Transformer-34 | UR100 | 669.2M | .599 | 64.3 | 32.7 |
| Transformer-34 | UR50/S | 669.2M | .639 | 69.2 | 50.2 |
| ESM-1b | UR50/S | 652.4M | .532 | 71.6 | 56.9 |

Table 8. Comparison to related protein language models. (RH) Remote Homology at the fold level, using Hit-10 metric on SCOP. (SSP) Secondary structure Q8 accuracy on CB513. (Contact) Top-L long range contact precision on RaptorX test set from Wang et al. (2017). Results for additional test sets in Table S6. Alley et al. (2019) Heinzinger et al. (2019) Rao et al. (2019). The pre-training datasets for related work have differences from ours.

表8. 与相关蛋白质语言模型的比较。(RH) 远程同源性在折叠水平上，使用SCOP上的Hit-10指标。(SSP) 在CB513上的二级结构Q8准确率。(Contact) 在Wang等人(2017)的RaptorX测试集上的Top-L长程接触精度。其他测试集的结果见表S6。 Alley等人(2019) Heinzinger等人(2019) Rao等人(2019)。 相关工作的预训练数据集与我们的有所不同。

We benchmark against related work in Table 8. Heinzinger et al. (2019), Alley et al. (2019), and Rao et al. (2019), evaluate models on differing downstream tasks and test sets. We retrieve the weights for the above models, evaluating them directly in our codebase against the panel of test sets used in this paper for remote homology, secondary structure prediction, and contact prediction, with the same training data and model architectures. This allows a direct comparison between the representations. Table 8 shows that high-capacity Transformers have strong performance for secondary structure and contact prediction significantly exceeding Alley et al. (2019), Heinzinger et al. (2019), and Rao et al. (2019). The small Transformer models trained as baselines also have higher performance than the methods with comparable parameter numbers.

我们在表8中对相关工作进行了基准测试。Heinzinger等人(2019)、Alley等人(2019)和Rao等人(2019)在不同的下游任务和测试集上评估了模型。我们检索了上述模型的权重，直接在我们的代码库中针对本文使用的远程同源性、二级结构预测和接触预测的测试集进行评估，使用相同的训练数据和模型架构。这使得我们能够直接比较这些表示。表8显示，高容量Transformer在二级结构和接触预测方面表现出色，显著超过了Alley等人(2019)、Heinzinger等人(2019)和Rao等人(2019)。作为基线训练的小型Transformer模型也比参数数量相当的方法表现更好。

Protein sequence embeddings have been the subject of recent investigation for protein engineering (Yang et al., 2018). Bepler & Berger (2019) pre-trained LSTMs on protein sequences, adding supervision from contacts to produce em-beddings. Subsequent to our preprint, related works have built on its exploration of protein sequence modeling, exploring generative models (Riesselman et al., 2019; Madani et al., 2020), internal representations of Transformers (Vig et al., 2020), and applications of representation learning and generative modeling such as classification (Elnaggar et al., 2019; Strodthoff et al., 2020), mutational effect prediction (Luo et al., 2020), and design of sequences (Repecka et al., 2019; Hawkins-Hooker et al., 2020; Amimeur et al., 2020).

蛋白质序列嵌入最近成为蛋白质工程研究的主题(Yang等人, 2018)。Bepler & Berger (2019)在蛋白质序列上预训练了LSTM，并添加了来自接触的监督以生成嵌入。在我们的预印本之后，相关工作基于其对蛋白质序列建模的探索，探索了生成模型(Riesselman等人, 2019; Madani等人, 2020)、Transformer的内部表示(Vig等人, 2020)，以及表示学习和生成模型的应用，如分类(Elnaggar等人, 2019; Strodthoff等人, 2020)、突变效应预测(Luo等人, 2020)和序列设计(Repecka等人, 2019; Hawkins-Hooker等人, 2020; Amimeur等人, 2020)。

# 9. Discussion

# 9. 讨论

One of the goals for artificial intelligence in biology could be the creation of controllable predictive and generative models that can read and generate biology in its native language. Accordingly, research will be necessary into methods that can learn intrinsic biological properties directly from protein sequences, which can be transferred to prediction and generation.

人工智能在生物学中的一个目标可能是创建可控的预测和生成模型，这些模型能够以其原生语言读取和生成生物学。因此，有必要研究能够直接从蛋白质序列中学习内在生物学特性的方法，这些方法可以转移到预测和生成中。

We investigated deep learning across evolution at the scale of the largest protein sequence databases, training contextual language models across 86 billion amino acids from 250 million sequences. The space of representations learned from sequences by high-capacity networks reflects biological structure at multiple levels, including that of amino acids, proteins, and evolutionary homology. Information about secondary and tertiary structure is internalized and represented within the network. Knowledge of intrinsic biological properties emerges without supervision - no learning signal other than sequences is given during pre-training.

我们在最大的蛋白质序列数据库规模上研究了跨进化的深度学习，训练了来自2.5亿个序列的860亿个氨基酸的上下文语言模型。高容量网络从序列中学习到的表示空间反映了多个层次的生物结构，包括氨基酸、蛋白质和进化同源性。二级和三级结构的信息被内化并在网络中表示。内在生物学特性的知识在没有监督的情况下出现——在预训练期间除了序列外没有提供任何学习信号。

We find that networks that have been trained across evolutionary data generalize: information can be extracted from representations by linear projections, deep neural networks, or by adapting the model using supervision. Fine-tuning produces results that match state-of-the-art on variant activity prediction. Predictions are made directly from the sequence, using features that have been automatically learned by the language model, rather than selected by domain knowledge.

我们发现，经过进化数据训练的网络具有泛化能力:信息可以通过线性投影、深度神经网络或通过使用监督调整模型从表示中提取。微调产生的结果与变异活性预测的最新水平相匹配。预测直接来自序列，使用语言模型自动学习的特征，而不是通过领域知识选择的特征。

We find that pre-training discovers information that is not present in current state-of-the-art features. The learned features can be combined with features used by state-of-the-art structure prediction methods to improve results. Empirically we find that features discovered by larger models perform better on downstream tasks. The Transformer outperforms LSTMs with similar capacity across benchmarks. Increasing diversity of the training data results in significant improvements to the representations.

我们发现预训练发现了当前最新特征中不存在的信息。学习到的特征可以与最新结构预测方法使用的特征结合以提高结果。经验上，我们发现由更大模型发现的特征在下游任务上表现更好。Transformer在基准测试中优于容量相似的LSTM。增加训练数据的多样性显著改善了表示。

While the protein language models we study are of comparable scale to those used in the text domain, our experiments have not yet reached the limit of scale. We observed that even the highest capacity models we trained (with approximately parameters) under-fit the sequence datasets, due to insufficient model capacity. The relationship we find between language modeling fidelity and the information about structure encoded into the representations suggests that higher capacity models will yield better representations. These findings imply potential for further model scale and data diversity incorporating sequences from metagenomics.

虽然我们研究的蛋白质语言模型与文本域中使用的模型规模相当，但我们的实验尚未达到规模的极限。我们观察到，即使是我们训练的最高容量模型(大约有 参数)也由于模型容量不足而未能充分拟合序列数据集。我们发现语言建模保真度与编码到表示中的结构信息之间的关系表明，更高容量的模型将产生更好的表示。这些发现暗示了进一步扩大模型规模和数据多样性的潜力，包括来自宏基因组学的序列。

Combining high-capacity generative models with gene synthesis and high throughput characterization can enable generative biology. The models we have trained can be used to generate new sequences (Wang & Cho, 2019). If neural networks can transfer knowledge learned from protein sequences to design functional proteins, this could be coupled with predictive models to jointly generate and optimize sequences for desired functions. The size of current sequence data and its projected growth point toward the possibility of a general purpose generative model that can condense the totality of sequence statistics, internalizing and integrating fundamental chemical and biological concepts including structure, function, activity, localization, binding, and dynamics, to generate new sequences that have not been seen before in nature but that are biologically active.

将高容量生成模型与基因合成和高通量表征相结合，可以实现生成生物学。我们训练的模型可用于生成新序列(Wang & Cho, 2019)。如果神经网络能够将从蛋白质序列中学到的知识转移到设计功能性蛋白质上，这可以与预测模型结合，共同生成和优化具有所需功能的序列。当前序列数据的规模及其预计增长表明，有可能开发出一种通用生成模型，该模型可以浓缩序列统计的整体，内化和整合包括结构、功能、活性、定位、结合和动力学在内的基本化学和生物学概念，从而生成自然界中从未见过但具有生物活性的新序列。

# Pre-trained models

# 预训练模型

Transformer models and baselines are available at: https://github.com/facebookresearch/esm

Transformer模型和基线可在以下网址获取:https://github.com/facebookresearch/esm

# Acknowledgments

# 致谢

We thank Tristan Bepler, Richard Bonneau, Yilun Du, Vladimir Gligorijevic, Anika Gupta, Omer Levy, Ian Peikon, Hetunandan Kamisetty, Laurens van der Maaten, Ethan Perez, Oded Regev, Neville Sanjana, and Emily Wang for feedback on the manuscript and insightful conversations. We thank Jinbo Xu for sharing RaptorX features and help with CASP13. We thank Michael Klausen for providing Netsurf training code. Alexander Rives was supported at NYU by NSF Grant #1339362.

我们感谢Tristan Bepler、Richard Bonneau、Yilun Du、Vladimir Gligorijevic、Anika Gupta、Omer Levy、Ian Peikon、Hetunandan Kamisetty、Laurens van der Maaten、Ethan Perez、Oded Regev、Neville Sanjana和Emily Wang对稿件的反馈和富有洞察力的讨论。我们感谢Jinbo Xu分享RaptorX特征并帮助CASP13。我们感谢Michael Klausen提供Netsurf训练代码。Alexander Rives在纽约大学的工作得到了NSF Grant #1339362的支持。

# References

# 参考文献

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. 统一的理性蛋白质工程

with sequence-based deep representation learning. Nature methods, 16(12):1315-1322, 2019.

基于序列的深度表示学习。自然方法，16(12):1315-1322, 2019。

Altschuh, D., Lesk, A., Bloomer, A., and Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. Journal of Molecular Biology, 193(4):693-707, 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90352-4.

Altschuh, D., Lesk, A., Bloomer, A., and Klug, A. 与烟草花叶病毒相关的病毒中协调氨基酸替换与功能的相关性。分子生物学杂志，193(4):693-707, 1987。ISSN 0022-2836。doi: 10.1016/0022-2836(87)90352-4。

Altschuh, D., Vernet, T., Berti, P., Moras, D., and Nagai, K. Coordinated amino acid changes in homologous protein families. Protein Engineering, Design and Selection, 2 (3):193-199, 1988.

Altschuh, D., Vernet, T., Berti, P., Moras, D., and Nagai, K. 同源蛋白质家族中的协调氨基酸变化。蛋白质工程、设计与选择，2(3):193-199, 1988。

Altschul, S. F. and Koonin, E. V. Iterated profile searches with psi-blast - a tool for discovery in protein databases. Trends in Biochemical Sciences, 23(11):444-447, 11 1998. ISSN 0968-0004. doi: 10.1016/S0968-0004(98) 01298-5.

Altschul, S. F. and Koonin, E. V. 使用psi-blast进行迭代轮廓搜索——蛋白质数据库中的发现工具。生物化学趋势，23(11):444-447, 1998年11月。ISSN 0968-0004。doi: 10.1016/S0968-0004(98)01298-5。

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology, 215(3):403-410, 1990. ISSN 0022- 2836. doi: 10.1016/S0022-2836(05)80360-2.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 基本局部比对搜索工具。分子生物学杂志，215(3):403-410, 1990。ISSN 0022-2836。doi: 10.1016/S0022-2836(05)80360-2。

Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., Van Citters, D., Siska, C. C., Smidt, P., Sprague, M., et al. Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. bioRxiv, 2020.

Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., Van Citters, D., Siska, C. C., Smidt, P., Sprague, M., 等。使用生成对抗网络设计特征控制的人形抗体发现库。bioRxiv, 2020。

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. Origins of coevolution between residues distant in protein structures. Proceedings of the National Academy of Sciences, 114(34):9122-9127, 2017.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. 蛋白质 结构中远距离残基共进化的起源。美国国家科学院院刊，114(34):9122-9127, 2017。

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. Cloze-driven pretraining of self-attention networks. CoRR, abs/1903.07785, 2019. URL http://arxiv.org/abs/1903.07785.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. 自注意力网络的填空驱动预训练。CoRR, abs/1903.07785, 2019。URL http://arxiv.org/abs/1903.07785。

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics, 79(4):1061-1078, 2011.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. 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(4):1061-1078, 2011。

Bateman, A., Heger, A., Sonnhammer, E. L. L., Mistry, J., Clements, J., Tate, J., Hetherington, K., Holm, L., Punta, M., Coggill, P., Eberhardt, R. Y., Eddy, S. R., and Finn, R. D. Pfam: the protein families database. Nucleic Acids Research, 42(D1):D222-D230, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1223.

Bateman, A., Heger, A., Sonnhammer, E. L. L., Mistry, J., Clements, J., Tate, J., Hetherington, K., Holm, L., Punta, M., Coggill, P., Eberhardt, R. Y., Eddy, S. R., and Finn, R. D. Pfam: 蛋白质家族数据库。核酸研究，42(D1):D222-D230, 2013年11月。ISSN 0305-1048。doi: 10.1093/nar/gkt1223。

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137-1155, 2003.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. 一种神经概率语言模型。机器学习研究杂志，3(2月):1137-1155, 2003年。

Bepler, T. and Berger, B. Learning protein sequence embeddings using information from structure. In - ternational Conference on Learning Representations, 2019. URL https://openreview.net/forum? id=SygLehCqtm.

Bepler, T. 和 Berger, B. 利用结构信息学习蛋白质序列嵌入。在 - 国际学习表示会议，2019年。URL https://openreview.net/forum? id=SygLehCqtm。

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pp. 160-167. ACM, 2008.

Collobert, R. 和 Weston, J. 自然语言处理的统一架构:多任务学习的深度神经网络。在第25届国际机器学习会议论文集，第160-167页。ACM, 2008年。

Cuff, J. A. and Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioin-formatics, 34(4):508-519, 1999.

Cuff, J. A. 和 Barton, G. J. 评估和改进蛋白质二级结构预测的多序列方法。蛋白质:结构、功能和生物信息学，34(4):508-519, 1999年。

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In Advances in Neural Information Processing Systems, pp. 3079-3087, 2015.

Dai, A. M. 和 Le, Q. V. 半监督序列学习。在神经信息处理系统进展，第3079-3087页，2015年。

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Devlin, J., Chang, M., Lee, K., 和 Toutanova, K. BERT:用于语言理解的深度双向Transformer预训练。CoRR, abs/1810.04805, 2018年。URL http://arxiv.org/abs/1810.04805。

Eddy, S. R. Profile hidden markov models. Bioinformatics, 14(9):755-763, 10 1998. ISSN 1367-4803. doi: 10.1093/ bioinformatics/14.9.755.

Eddy, S. R. 轮廓隐马尔可夫模型。生物信息学，14(9):755-763, 1998年10月。ISSN 1367-4803。doi: 10.1093/ bioinformatics/14.9.755。

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Au-rell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. Phys. Rev. E, 87:012707, Jan 2013a. doi: 10.1103/PhysRevE.87. 012707. URL https://link.aps.org/doi/10.1103/PhysRevE.87.012707.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., 和 Au-rell, E. 改进的蛋白质接触预测:使用伪似然推断Potts模型。物理评论E，87:012707, 2013年1月。doi: 10.1103/PhysRevE.87. 012707。URL https://link.aps.org/doi/10.1103/PhysRevE.87.012707。

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. Improved contact prediction in proteins: using pseu-dolikelihoods to infer potts models. Physical Review E, 87(1):012707, 2013b.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., 和 Aurell, E. 改进的蛋白质接触预测:使用伪似然推断Potts模型。物理评论E，87(1):012707, 2013年。

Elnaggar, A., Heinzinger, M., Dallago, C., and Rost, B. End-to-end multitask learning, from protein language to protein features without alignments. bioRxiv, pp. 864405, 2019.

Elnaggar, A., Heinzinger, M., Dallago, C., 和 Rost, B. 端到端多任务学习，从蛋白质语言到蛋白质特征，无需对齐。bioRxiv，第864405页，2019年。

Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., and Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. Molecular biology and evolution, 33(1):268-280, 2016.

Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., 和 Weigt, M. 共进化景观推断和β-内酰胺酶TEM-1突变的上下文依赖性。分子生物学与进化，33(1):268-280, 2016年。

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. Nature methods, 11(8):801, 2014.

Fowler, D. M. 和 Fields, S. 深度突变扫描:蛋白质科学的新风格。自然方法，11(8):801, 2014年。

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research, 42(D1): D304-D309, January 2014. ISSN 0305-1048, 1362- 4962. doi: 10.1093/nar/gkt1240. URL https:// academic.oup.com/nar/article-lookup/ doi/10.1093/nar/gkt1240.

Fox, N. K., Brenner, S. E., 和 Chandonia, J.-M. SCOPe:蛋白质结构分类扩展，整合SCOP和ASTRAL数据及新结构分类。核酸研究，42(D1): D304-D309, 2014年1月。ISSN 0305-1048, 1362- 4962。doi: 10.1093/nar/gkt1240。URL https:// academic.oup.com/nar/article-lookup/ doi/10.1093/nar/gkt1240。

Gabaldon, T. Evolution of proteins and proteomes: a phy-logenetics approach. Evol Bioinform Online, 1:51-61, 2007.

Gabaldon, T. 蛋白质和蛋白质组的进化:系统发育方法。进化生物信息学在线，1:51-61, 2007年。

Göbel, U., Sander, C., Schneider, R., and Valencia, A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics, 18(4):309-317, 1994.

Göbel, U., Sander, C., Schneider, R., 和 Valencia, A. 蛋白质中的相关突变和残基接触。蛋白质:结构、功能和生物信息学，18(4):309-317, 1994.

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell systems, 6(1):116-124, 2018.

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., 和 Fowler, D. M. 使用大规模诱变数据进行定量错义变异效应预测。细胞系统，6(1):116-124, 2018.

Harris, Z. S. Distributional structure. Word, 10(2-3):146- 162, 1954.

Harris, Z. S. 分布结构。词，10(2-3):146-162, 1954.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. Generating functional protein variants with variational autoencoders. BioRxiv, 2020.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., 和 Bikard, D. 使用变分自编码器生成功能性蛋白质变体。BioRxiv, 2020.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. bioinformatics,20(1):723,2019.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., 和 Rost, B. 通过迁移学习蛋白质序列模拟生命语言的各个方面。 生物信息学，20(1):723,2019.

Hopf, T., Ingraham, J., Poelwijk, F., Scharfe, C., Springer, M., Sander, C., and Marks, D. Mutation effects predicted from sequence co-variation. Nature Biotechnology, 35: 128-135, 1 2017.

Hopf, T., Ingraham, J., Poelwijk, F., Scharfe, C., Springer, M., Sander, C., 和 Marks, D. 从序列共变异预测突变效应。自然生物技术，35: 128-135, 1 2017.

Hormoz, S. Amino acid composition of proteins reduces deleterious impact of mutations. Scientific reports, 3: 2919, 2013.

Hormoz, S. 蛋白质的氨基酸组成减少了突变的有害影响。科学报告，3: 2919, 2013.

Huerta-Cepas, J., Forslund, S. K., Bork, P., Hernández-Plaza, A., von Mering, C., Szklarczyk, D., Heller, D., Cook, H., Jensen, L., Mende, D. R., Letunic, I., and Rattei, T. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47(D1):D309-D314, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1085.

Huerta-Cepas, J., Forslund, S. K., Bork, P., Hernández-Plaza, A., von Mering, C., Szklarczyk, D., Heller, D., Cook, H., Jensen, L., Mende, D. R., Letunic, I., 和 Rattei, T. eggNOG 5.0:基于5090种生物和2502种病毒的分层、功能和系统发育注释的直系同源资源。核酸研究，47(D1):D309-D314, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1085.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In 2009 IEEE 12th international conference on computer vision, pp. 2146-2153. IEEE, 2009.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., 和 LeCun, Y. 什么是物体识别的最佳多阶段架构？在2009年IEEE第12届国际计算机视觉会议上，pp. 2146-2153. IEEE, 2009.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. CoRR, abs/1702.08734, 2017. URL http://arxiv.org/abs/1702.08734.

Johnson, J., Douze, M., 和 Jégou, H. 使用GPU进行十亿级相似性搜索。CoRR, abs/1702.08734, 2017. URL http://arxiv.org/abs/1702.08734.

Jones, D. T. and Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics, 34(19):3308-3315, 2018.

Jones, D. T. 和 Kandathil, S. M. 使用全卷积神经网络和最小序列特征进行高精度蛋白质接触预测。生物信息学，34(19):3308-3315, 2018.

Jones, D. T., Buchan, D. W., Cozzetto, D., and Pontil, M. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, 28(2):184-190, 2011.

Jones, D. T., Buchan, D. W., Cozzetto, D., 和 Pontil, M. Psicov:使用稀疏逆协方差估计在大规模多序列比对中进行精确结构接触预测。生物信息学，28(2):184-190, 2011.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. Character-aware neural language models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., pp. 2741-2749, 2016. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489.

Kim, Y., Jernite, Y., Sontag, D., 和 Rush, A. M. 字符感知神经语言模型。在第三十届AAAI人工智能会议论文集，2016年2月12-17日，美国亚利桑那州凤凰城，pp. 2741-2749, 2016. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sonderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., and Marcatili, P. NetSurfP- 2.0: Improved prediction of protein structural features by integrated deep learning. Proteins, 87(6):520-527, 06 2019.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sonderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., 和 Marcatili, P. NetSurfP- 2.0:通过集成深度学习改进蛋白质结构特征预测。蛋白质，87(6):520-527, 06 2019.

Kondrashov, F. A., Bork, P., Sunyaev, S., and Ramensky, V. Impact of selection, mutation rate and genetic drift on human genetic variation. Human Molecular Genetics, 12(24):3325-3330, 12 2003. ISSN 0964-6906. doi: 10.1093/hmg/ddg359. URL https://doi.org/10.1093/hmg/ddg359.

Kondrashov, F. A., Bork, P., Sunyaev, S., and Ramensky, V. 选择、突变率和遗传漂变对人类遗传变异的影响。人类分子遗传学，12(24):3325-3330, 2003年12月。ISSN 0964-6906。doi: 10.1093/hmg/ddg359。URL https://doi.org/10.1093/hmg/ddg359。

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. Critical assessment of methods of protein structure prediction (casp)-round xiii. Proteins: Structure, Function, and Bioinformatics, 87(12):1011-1020, 2019.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. 蛋白质结构预测方法的关键评估(CASP)-第十三轮。蛋白质:结构、功能和生物信息学，87(12):1011-1020, 2019。

Lapedes, A. S., Giraud, B. G., Liu, L., and Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. Lecture Notes-Monograph Series, pp. 236-256, 1999.

Lapedes, A. S., Giraud, B. G., Liu, L., and Stormo, G. D. 蛋白质序列模型中的相关突变:系统发育和结构效应。讲座笔记-专著系列，第236-256页，1999。

Levitt, M. Conformational preferences of amino acids in globular proteins. Biochemistry, 17(20):4277-4285, 1978.

Levitt, M. 球状蛋白质中氨基酸的构象偏好。生物化学，17(20):4277-4285, 1978。

Luo, Y., Vo, L., Ding, H., Su, Y., Liu, Y., Qian, W. W., Zhao, H., and Peng, J. Evolutionary context-integrated deep sequence modeling for protein engineering. In - ternational Conference on Research in Computational Molecular Biology, pp. 261-263. Springer, 2020.

Luo, Y., Vo, L., Ding, H., Su, Y., Liu, Y., Qian, W. W., Zhao, H., and Peng, J. 用于蛋白质工程的进化背景整合深度序列建模。在 - 国际计算分子生物学研究会议，第261-263页。Springer, 2020。

Ma, J., Wang, S., Wang, Z., and Xu, J. Mrfalign: protein homology detection through alignment of markov random

Ma, J., Wang, S., Wang, Z., and Xu, J. Mrfalign:通过马尔可夫随机场对齐进行蛋白质同源性检测

fields. In International Conference on Research in Computational Molecular Biology, pp. 173-174. Springer, 2014.

在国际计算分子生物学研究会议，第173-174页。Springer, 2014。

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. Journal of machine learning research, 9(Nov): 2579-2605, 2008.

Maaten, L. v. d. and Hinton, G. 使用t-SNE可视化数据。机器学习研究杂志，9(11月): 2579-2605, 2008。

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation. arXiv preprint arXiv:2004.03497, 2020.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen:用于蛋白质生成的语言建模。arXiv预印本 arXiv:2004.03497, 2020。

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3d structure computed from evolutionary sequence variation. PloS one, 6(12):e28766, 2011.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. 从进化序列变异计算的蛋白质3D结构。PloS one，6(12):e28766, 2011。

Mikolov, T., Sutskever, I., Deoras, A., Le, H.-S., Kom-brink, S., and Cernocky, J. Subword language modeling with neural networks. preprint (http://www.fit.vutbr.cz/imikolov/rnnlm/char.pdf), 8, 2012.

Mikolov, T., Sutskever, I., Deoras, A., Le, H.-S., Kom-brink, S., and Cernocky, J. 使用神经网络进行子词语言建模。预印本 (http://www.fit.vutbr.cz/imikolov/rnnlm/char.pdf), 8, 2012。

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2- 4, 2013, Workshop Track Proceedings, 2013. URL http://arxiv.org/abs/1301.3781.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 向量空间中词表示的高效估计。在第一届国际学习表示会议，ICLR 2013，美国亚利桑那州斯科茨代尔，2013年5月2-4日，研讨会论文集，2013。URL http://arxiv.org/abs/1301.3781。

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research, 45(D1):D170-D176, 2017.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust数据库:聚类和深度注释的蛋白质序列和比对。核酸研究，45(D1):D170-D176, 2017。

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences, 108(49):E1293-E1301, 2011.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. 残基共演化的直接耦合分析捕捉了许多蛋白质家族中的天然接触。美国国家科学院院刊，108(49):E1293-E1301, 2011。

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round xi. Proteins: Structure, Function, and Bioinformatics, 84: 4-14, 2016.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. 蛋白质结构预测方法的关键评估:第十一轮的进展和新方向。蛋白质:结构、功能和生物信息学，84: 4-14, 2016。

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. Critical assessment of methods of protein structure prediction (casp)-round xii. Proteins: Structure, Function, and Bioinformatics, 86:7-15, 2018.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., 和 Tramontano, A. 蛋白质结构预测方法的关键评估(CASP)-第十二轮。蛋白质:结构、功能与生物信息学，86:7-15, 2018。

Overbaugh, J. and Bangham, C. Selection forces and constraints on retroviral sequence variation. Science, 292: 1106-1109, 5 2001. doi: 10.1126/science.1059128.

Overbaugh, J. 和 Bangham, C. 逆转录病毒序列变异的选择压力与限制。科学，292: 1106-1109, 2001年5月。doi: 10.1126/science.1059128。

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark,

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark,

C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 2227- 2237,2018.URL https://aclanthology.info/ papers/N18-1202/n18-1202.

C., Lee, K., 和 Zettlemoyer, L. 深度上下文词表示。在2018年北美计算语言学协会人类语言技术分会会议论文集(NAACL-HLT 2018)中，美国路易斯安那州新奥尔良，2018年6月1-6日，第1卷(长论文)，pp. 2227-2237, 2018。URL https://aclanthology.info/papers/N18-1202/n18-1202。

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pretraining. 2018.

Radford, A., Narasimhan, K., Salimans, T., 和 Sutskever, I. 通过生成预训练提升语言理解。2018。

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., 和 Sutskever, I. 语言模型是无监督多任务学习者。2019。

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. In Advances in Neural Information Processing Systems, pp. 9686-9698, 2019.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., 和 Song, Y. 使用TAPE评估蛋白质迁移学习。在神经信息处理系统进展中，pp. 9686-9698, 2019。

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hh-blits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nature Methods, 9:173 EP, 12 2011. doi: 10.1038/nmeth.1818.

Remmert, M., Biegert, A., Hauser, A., 和 Söding, J. HH-blits:通过HMM-HMM比对进行闪电般快速的迭代蛋白质序列搜索。自然方法，9:173 EP, 2011年12月。doi: 10.1038/nmeth.1818。

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Zrimec, J., Poviloniene, S., Rokaitis, I., Laurynenas, A., Abuajwa, W., Savolainen, O., et al. Expanding functional protein sequence space using generative adversarial networks. bioRxiv, pp. 789719, 2019.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Zrimec, J., Poviloniene, S., Rokaitis, I., Laurynenas, A., Abuajwa, W., Savolainen, O., 等。使用生成对抗网络扩展功能性蛋白质序列空间。bioRxiv, pp. 789719, 2019。

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10):816-822, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4.

Riesselman, A. J., Ingraham, J. B., 和 Marks, D. S. 深度生成模型捕捉遗传变异的影响。自然方法，15(10):816-822, 2018。ISSN 1548-7105。doi: 10.1038/s41592-018-0138-4。

Riesselman, A. J., Shin, J.-E., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Accelerating protein design using autoregressive generative models. bioRxiv, pp. 757252, 2019.

Riesselman, A. J., Shin, J.-E., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., 和 Marks, D. S. 使用自回归生成模型加速蛋白质设计。bioRxiv, pp. 757252, 2019。

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv, pp. 622803, 2019.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., 和 Fergus, R. 生物结构与功能通过将无监督学习扩展到2.5亿蛋白质序列而涌现。bioRxiv, pp. 622803, 2019。

Seemayer, S., Gruber, M., and Söding, J. Compred-fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics, 30(21):3128- 3130, 2014.

Seemayer, S., Gruber, M., 和 Söding, J. Compred:从相关突变中快速精确预测蛋白质残基-残基接触。生物信息学，30(21):3128-3130, 2014。

Senior, A., Jumper, J., and Hassabis, D. AlphaFold: Using AI for scientific discovery, 12 2018. URL https:// deepmind.com/blog/alphafold/.

Senior, A., Jumper, J., 和 Hassabis, D. AlphaFold:使用AI进行科学发现，2018年12月。URL https://deepmind.com/blog/alphafold/。

Söding, J. and Remmert, M. Protein sequence comparison

Söding, J. 和 Remmert, M. 蛋白质序列比较

and fold recognition: progress and good-practice benchmarking. Current opinion in structural biology, 21(3): 404-411, 2011.

与折叠识别:进展与良好实践基准。结构生物学当前观点，21(3): 404-411, 2011。

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. Udsmprot: universal deep sequence models for protein classification. Bioinformatics, 36(8):2401-2409, 2020.

Strodthoff, N., Wagner, P., Wenzel, M., 和 Samek, W. Udsmprot:用于蛋白质分类的通用深度序列模型。生物信息学, 36(8):2401-2409, 2020.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23(10):1282-1288, May 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm098. URL https: //academic.oup.com/bioinformatics/ article/23/10/1282/197795.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., 和 Wu, C. H. UniRef:全面且非冗余的UniProt参考聚类。生物信息学, 23(10):1282-1288, 2007年5月. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm098. URL https://academic.oup.com/bioinformatics/article/23/10/1282/197795.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31(6):926-932, 2015.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., 和 Consortium, U. Uniref聚类:用于改进序列相似性搜索的全面且可扩展的替代方案。生物信息学, 31(6):926-932, 2015.

The UniProt Consortium. The universal protein resource (uniprot). Nucleic Acids Research, 36(suppl\_1):D190- D195, 11 2007. ISSN 0305-1048. doi: 10.1093/nar/ gkm895.

UniProt Consortium. 通用蛋白质资源(uniprot)。核酸研究, 36(suppl\_1):D190-D195, 2007年11月. ISSN 0305-1048. doi: 10.1093/nar/gkm895.

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. Graphical models of residue coupling in protein families. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5(2):183-197, 2008.

Thomas, J., Ramakrishnan, N., 和 Bailey-Kellogg, C. 蛋白质家族中残基耦合的图模型。IEEE/ACM计算生物学与生物信息学汇刊, 5(2):183-197, 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polosukhin, I. 注意力机制就是你所需要的一切。在神经信息处理系统进展中, pp. 5998-6008, 2017.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. Bertology meets biology: Interpreting attention in protein language models. arXiv preprint arXiv:2006.15222, 2020.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., 和 Rajani, N. F. Bertology与生物学的结合:解释蛋白质语言模型中的注意力机制。arXiv预印本 arXiv:2006.15222, 2020.

Wang, A. and Cho, K. BERT has a mouth, and it must speak: BERT as a markov random field language model. CoRR, abs/1902.04094, 2019. URL http://arxiv.org/abs/1902.04094.

Wang, A. 和 Cho, K. BERT有嘴，它必须说话:BERT作为马尔可夫随机场语言模型。CoRR, abs/1902.04094, 2019. URL http://arxiv.org/abs/1902.04094.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1): e1005324, 2017.

Wang, S., Sun, S., Li, Z., Zhang, R., 和 Xu, J. 通过超深度学习模型准确预测蛋白质接触图。PLoS计算生物学, 13(1): e1005324, 2017.

Wang, S.-W., Bitbol, A.-F., and Wingreen, N. Revealing evolutionary constraints on proteins through sequence analysis. PLoS Comput Biol, 15(4), 2019. doi: 10.1371/ journal.pcbi.1007010.

Wang, S.-W., Bitbol, A.-F., 和 Wingreen, N. 通过序列分析揭示蛋白质的进化约束。PLoS计算生物学, 15(4), 2019. doi: 10.1371/journal.pcbi.1007010.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences, 106(1):67-72, 2009.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., 和 Hwa, T. 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, 106(1):67-72, 2009.

Xu, J. Distance-based protein folding powered by deep learning. arXiv preprint arXiv:1811.03481, 2018.

徐, J. 基于距离的蛋白质折叠由深度学习驱动. arXiv预印本 arXiv:1811.03481, 2018.

Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. Learned protein embeddings for machine learning. Bioin-formatics, 34(15):2642-2648, 2018.

杨, K. K., 吴, Z., Bedbrook, C. N., 和 Arnold, F. H. 用于机器学习的蛋白质嵌入. 生物信息学, 34(15):2642-2648, 2018.

Yang, K. K., Wu, Z., and Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. Nature methods, 16(8):687-694, 2019.

杨, K. K., 吴, Z., 和 Arnold, F. H. 机器学习引导的定向进化用于蛋白质工程. 自然方法, 16(8):687-694, 2019.

Yanofsky, C., Horn, V., and Thorpe, D. Protein structure relationships revealed by mutational analysis. Science, 146(3651):1593-1594, 1964.

Yanofsky, C., Horn, V., 和 Thorpe, D. 通过突变分析揭示的蛋白质结构关系. 科学, 146(3651):1593-1594, 1964.

# A. Approach & Data

# A. 方法与数据

# 1. Pre-training datasets

# 1. 预训练数据集

UniParc pre-training dataset. A series of development models are trained on UniParc (The UniProt Consortium, 2007) Release 2019\_01 which contains approximately sequences. sequences are held-out randomly for validation. These models were used in the preprint of this paper, and representations from the models are used in Figures 1, 2, and 3.

UniParc预训练数据集. 一系列开发模型在UniParc(UniProt Consortium, 2007)2019\_01版本上进行训练，该版本包含大约 条序列。 条序列被随机保留用于验证。这些模型用于本论文的预印本，模型的表示用于图1、2和3。

UniRef pre-training datasets. Datasets are based on UniRef (Suzek et al., 2015) dated March 28, 2018 to permit a temporal hold-out with CASP13. of UniRef50 clusters are randomly selected as a held-out evaluation set, yielding 3.02 million representative sequences for evaluation. Three training datasets are used, removing all sequences belonging to clusters selected for the evaluation set: (i) UR100, 124.9M UniRef100 representative sequences; (ii) UR50/S, 27.1M UniRef50 representative sequences; (iii) UR50/D, 124.9M UniRef50 cluster members sampled evenly by cluster. To ensure a deterministic validation set, we removed sequences longer than 1024 amino acids from the validation set.

UniRef预训练数据集. 数据集基于2018年3月28日的UniRef(Suzek等, 2015)，以便与CASP13进行时间上的保留。 的UniRef50簇被随机选为保留评估集，产生302万条代表性序列用于评估。使用了三个训练数据集，移除了所有属于评估集所选簇的序列:(i) UR100, 1.249亿条UniRef100代表性序列；(ii) UR50/S, 2710万条UniRef50代表性序列；(iii) UR50/D, 1.249亿条UniRef50簇成员按簇均匀采样。为确保确定性验证集，我们从验证集中移除了长度超过1024个氨基酸的序列。

# 2. Downstream tasks

# 2. 下游任务

Remote Homology. A dataset of remote homolog pairs is derived from SCOPe (Fox et al., 2014) containing 256,806 pairs of remote homologs at the fold level and 92,944 at the superfamily level, consisting of 217 unique folds and 366 unique superfamilies. Creation of the dataset is detailed below in the section on remote homology.

远程同源性. 远程同源对数据集源自SCOPe(Fox等, 2014)，包含256,806对折叠级别的远程同源和92,944对超家族级别的远程同源，由217个独特的折叠和366个独特的超家族组成。数据集的创建详见下文关于远程同源性的部分。

Linear projections. Five-fold cross validation datasets implementing structural hold-outs at the family, superfamily, and fold level are constructed using SCOPe (Fox et al., 2014). Independently for each level of structural hold-out, the domains are split into 5 equal sets, i.e. five sets of folds, superfamilies, or families. This ensures that for each of the five partitions, structures having the same classification do not appear in both the train and test sets. For a given classification level each structure appears in a test set once, so that in the cross validation experiment each of the structures will be evaluated exactly once. Scores reported are the mean and standard deviation over each of the five test sets. Further details on construction of the dataset are given below in the section on linear projections.

线性投影. 使用SCOPe(Fox等, 2014)构建了在家族、超家族和折叠级别上实施结构保留的五折交叉验证数据集。对于每个结构保留级别，域被分成5个相等的集合，即五组折叠、超家族或家族。这确保了在五个分区中的每一个中，具有相同分类的结构不会同时出现在训练集和测试集中。对于给定的分类级别，每个结构在测试集中出现一次，因此在交叉验证实验中，每个结构将被评估一次。报告的分数是五个测试集的平均值和标准差。关于数据集构建的更多细节见下文关于线性投影的部分。

Secondary structure prediction. All downstream models are trained using the Netsurf (Klausen et al., 2019) training dataset containing 10,837 examples with labels and HMM profiles. Netsurf features are replicated for CASP13 domains using MMseqs2 (Steinegger & Söding, 2017) on the Uniclust90 (Mirdita et al., 2017) dataset released April 2017. For test sets we use (i) the standard CB513 (Cuff & Barton, 1999) test set of 513 sequences with sequence identity hold-out at 25% identity; and (ii) the 34 publicly available CASP domains, using DSSP (Kabsch & Sander, 1983) to label secondary structure, with temporal hold-out for both pre-training and downstream data.

二级结构预测. 所有下游模型均使用Netsurf(Klausen等, 2019)训练数据集进行训练，该数据集包含10,837个带有标签和HMM谱的示例。Netsurf特征在2017年4月发布的Uniclust90(Mirdita等, 2017)数据集上使用MMseqs2(Steinegger & Söding, 2017)为CASP13域复制。对于测试集，我们使用(i) 标准的CB513(Cuff & Barton, 1999)测试集，包含513条序列，序列同一性保留在25%；以及(ii) 34个公开的CASP域，使用DSSP(Kabsch & Sander, 1983)标记二级结构，预训练和下游数据均进行时间保留。

Contact prediction. All downstream models are trained using the training and test sets of Wang et al. (2017). Comparisons with RaptorX features use features from Wang et al. (2017) and Xu (2018). The following test sets are used: (i) RaptorX Test, 500 domains (25% sequence identity holdout); (ii) CASP11, 105 domains (25% sequence identity hold-out); (iii) CASP12, 55 domains (temporal hold-out from training data but not pre-training data); (iv) CASP13, 34 publicly released domains (temporal hold-out from training data and pre-training data). The training set consists of 6,767 sequences with contact map targets, a subset of PDB created in February 2015 (Wang et al., 2017). The use of an earlier version of the PDB ensures a temporal holdout w.r.t. both CASP12 and CASP13. Additionally, Wang et al. (2017) implemented a sequence identity hold-out for Test and CASP11 by removing proteins from the training set which share sequence identity or have BLAST E-value with the proteins in these test sets.

接触预测。所有下游模型均使用Wang等人(2017)的训练集和测试集进行训练。与RaptorX特征的比较使用了Wang等人(2017)和Xu(2018)的特征。使用的测试集如下:(i)RaptorX测试集，500个域(25%序列同一性保留)；(ii)CASP11，105个域(25%序列同一性保留)；(iii)CASP12，55个域(从训练数据中时间保留，但不包括预训练数据)；(iv)CASP13，34个公开的域(从训练数据和预训练数据中时间保留)。训练集由6,767个具有接触图目标的序列组成，这些序列是2015年2月创建的PDB子集(Wang等人，2017)。使用早期版本的PDB确保了与CASP12和CASP13的时间保留。此外，Wang等人(2017)通过从训练集中移除与这些测试集中的蛋白质共享 序列同一性或具有BLAST E值 的蛋白质，实现了测试集和CASP11的序列同一性保留。

Mutational effect prediction. The model is fine-tuned on deep mutational scanning datasets compiled by Gray et al. (2018) and Riesselman et al. (2018).

突变效应预测。该模型在Gray等人(2018)和Riesselman等人(2018)编译的深度突变扫描数据集上进行了微调。

# 3.The Transformer

# 3.Transformer

We use a deep Transformer encoder model (Vaswani et al., 2017b; Devlin et al., 2018b), processing input as character sequences of amino acids. In contrast to recurrent and convolutional neural networks, the Transformer makes no assumptions on the ordering of the input and instead uses position embeddings. Particularly relevant to protein sequences is the Transformer’s ability to model long range dependencies, which are not effectively captured by RNNs or LSTMs (Khandelwal et al., 2018). A key factor affecting the performance of LSTMs on these tasks is the path lengths that must be traversed by forward activation and backward gradient signals in the network (Hochreiter et al., 2001).

我们使用了一个深度Transformer编码器模型(Vaswani等人，2017b；Devlin等人，2018b)，将输入处理为氨基酸的字符序列。与循环神经网络和卷积神经网络不同，Transformer不对输入的顺序做任何假设，而是使用位置嵌入。特别与蛋白质序列相关的是Transformer建模长程依赖关系的能力，这是RNN或LSTM无法有效捕捉的(Khandelwal等人，2018)。影响LSTM在这些任务上表现的一个关键因素是网络中前向激活和后向梯度信号必须遍历的路径长度(Hochreiter等人，2001)。

It is well known that structural properties of protein sequences are reflected in long-range dependencies. Direct coupling analysis (Lapedes et al., 1999; Thomas et al., 2008; Weigt et al., 2009) which aims to detect pairwise dependencies in multiple sequence alignments uses a Markov Random Field (Potts Model) which models the complete sequence with pairwise coupling parameters. Similarly, the Transformer builds up a representation of a sequence by alternating self-attention with non-linear projections. Self-attention structures computation so that each position is represented by a weighted sum of the other positions in the sequence. The attention weights are computed dynamically and allow each position to choose what information from the rest of the sequence to integrate at every computation step.

众所周知，蛋白质序列的结构特性反映在长程依赖关系中。直接耦合分析(Lapedes等人，1999；Thomas等人，2008；Weigt等人，2009)旨在检测多序列比对中的成对依赖关系，使用马尔可夫随机场(Potts模型)通过成对耦合参数对完整序列进行建模。同样，Transformer通过交替自注意力和非线性投影来构建序列的表示。自注意力结构计算使得每个位置由序列中其他位置的加权和表示。注意力权重是动态计算的，允许每个位置在每个计算步骤中选择从序列的其余部分集成哪些信息。

Developed to model large contexts and long range dependencies in language data, self-attention architectures currently give state-of-the-art performance on various natural language tasks, mostly due to the Transformer’s scalability in parameters and the amount of context it can integrate (Devlin et al., 2018b). The tasks include token-level tasks like part-of-speech tagging, sentence-level tasks such as textual entailment, and paragraph-level tasks like question-answering.

自注意力架构最初是为建模语言数据中的大上下文和长程依赖关系而开发的，目前在各种自然语言任务中提供了最先进的性能，这主要归功于Transformer在参数数量和上下文集成量方面的可扩展性(Devlin等人，2018b)。这些任务包括词性标注等标记级任务、文本蕴含等句子级任务以及问答等段落级任务。

Scaled dot-product attention. Self-attention takes a sequence of vectors and produces a sequence of vectors by computing interactions between all elements in the sequence. The Transformer model uses scaled dot-product attention (Vaswani et al., 2017b):

缩放点积注意力。自注意力通过计算序列中所有元素之间的交互，将向量序列 转换为向量序列 。Transformer模型使用缩放点积注意力(Vaswani等人，2017b):

Here the query , key , and value , are projections of the input sequence to matrices where is the length of the sequence and is the inner dimension of the matrix outer product between and . This outer product parameterizes an map of attention logits, which are rescaled, and passed through the softmax function row-wise, thereby representing each position of the sequence in the output as a convex combination of the sequence of values . One step of self-attention directly models possible pairwise interactions between all positions in the sequence simultaneously. Note the contrast to recurrent and convolutional models which can only represent long-range context through many steps, and the parallel in inductive bias with the explicit pairwise parameterization of Markov Random Fields in widespread use for modeling protein MSAs.

这里查询 、键 和值 是输入序列到 矩阵的投影，其中 是序列的长度， 是 和 之间矩阵外积的内维度。此外积参数化了注意力逻辑的 映射，这些逻辑被重新缩放，并通过逐行的softmax函数传递，从而在输出中将序列的每个位置表示为值序列 的凸组合。自注意力的一个步骤直接同时建模序列中所有位置之间可能的成对交互。请注意，这与循环和卷积模型形成对比，后者只能通过多个步骤表示长程上下文，并且与广泛用于建模蛋白质MSA的马尔可夫随机场的显式成对参数化在归纳偏差上具有相似性。

Multi-headed self-attention concatenates the output of independent attention heads:

多头自注意力将 个独立注意力头的输出连接起来:

Use of multiple heads enables representation of different inter-position interaction patterns.

使用多头机制能够表示不同位置间的交互模式。

Architecture The Transformer models (Vaswani et al., 2017b) in this work take a sequence of tokens and output a sequence of probabilities which are optimized using the masked language modeling objective. The computation proceeds through a series of residual blocks producing hidden states, each a sequence of vectors with embedding dimension .

架构 本工作中的Transformer模型(Vaswani等，2017b)接收一个标记序列 并输出一个 概率序列 ，这些概率通过掩码语言建模目标进行优化。计算过程通过一系列残差块进行，生成隐藏状态，每个隐藏状态是一个向量序列 ，其嵌入维度为 。

The Transformer model architecture consists of a series of encoder blocks interleaving two functions: a multiheaded self-attention computing position-position interactions across the sequence, and a feed-forward network applied independently at each position.

Transformer模型架构由一系列编码器块组成，这些块交替执行两个功能:多头自注意力计算序列中位置-位置间的交互，以及在前馈网络中独立应用于每个位置。

The attention unit:

注意力单元:

Applies one step of multi-headed scaled dot-product attention to the normalized input, denoted by , projecting the result into the residual path.

对归一化输入应用一步多头缩放点积注意力，记为 ，并将结果投影到残差路径中。

The feed-forward network (with the output state of defining the "MLP dimension"):

前馈网络(以 的输出状态定义“MLP维度”):

Passes the normalized input through a position-independent multi-layered perceptron (MLP) with activation function .

将归一化输入通过一个位置无关的多层感知机(MLP)，其激活函数为 。

The full Transformer block:

完整的Transformer块:

Successively applies the self-attention unit, and the feed-forward network on a residual path.

依次在残差路径上应用自注意力单元和前馈网络。

The Transformer model:

Transformer模型:

Transformer(x):

Transformer(x):

Consists of an embedding step with token and positional embeddings, followed by layers of Transformer blocks, before a projection to log probabilities. The raw input sequence is represented as a sequence of 1- hot vectors of dimension 25, which is passed through the learned embedding layer before being presented to the first Transformer layer.

包括一个嵌入步骤，使用标记 和位置 嵌入，随后是 层Transformer块，最后投影到对数概率 。原始输入序列表示为维度为25的独热向量序列，在传递给第一个Transformer层之前通过 学习到的嵌入层。

The models trained in the paper use pre-activation blocks (He et al., 2016), where the layer normalization (Ba et al., 2016) is applied prior to the activation as in Radford et al. (2019b), enabling stable training of deep Transformer networks. No dropout is used. All projections include biases, except for the token and positional embeddings. We use learned token embeddings, and harmonic positional em-beddings as in (Vaswani et al., 2017b). The feed-forward network uses the Gaussian error linear unit (Hendrycks & Gimpel, 2016) activation function. We initialize all layers from a zero centered normal distribution with standard deviation 0.02, and re-scale the initialization of the projections into the residual path by where is the number of residual layers. All biases are initialized to zero. The query, key, and value projections are to dimensions, and the hidden dimension of the feed-forward network is .

本文训练的模型使用**预激活块**(He等，2016)，其中层归一化(Ba等，2016)在激活之前应用，如Radford等(2019b)所述，从而实现深度Transformer网络的稳定训练。未使用dropout。所有投影包括偏置，除了标记和位置嵌入。我们使用学习到的标记嵌入，以及如(Vaswani等，2017b)中的谐波位置嵌入。前馈网络使用**高斯误差线性单元**(Hendrycks & Gimpel，2016)激活函数。我们从均值为零、标准差为0.02的正态分布初始化所有层，并通过 重新缩放残差路径中的投影初始化，其中 是残差层数。所有偏置初始化为零。查询、键和值投影到 维度，前馈网络的隐藏维度为 。

# 4. Pre-trained Transformer Models

# 4. 预训练的Transformer模型

UniParc development models We experimented with Transformer models of various depths, including a 36-layer Transformer with 708.6 million parameters, and a 12-layer model with parameters trained. Development models were trained on UniParc. Details are in Table A.1.

UniParc开发模型 我们实验了不同深度的Transformer模型，包括一个36层、7.086亿参数的Transformer，以及一个12层、 参数的模型。开发模型在UniParc上训练。详细信息见表A.1。

UniRef models We train 34-layer models with 669.2M parameters across different datasets and fractions of training data. Additionally we train 6 and 12-layer models. These models are detailed in Table A.2.

UniRef模型 我们在不同数据集和训练数据比例上训练了34层、6.692亿参数的模型。此外，我们还训练了6层和12层的模型。这些模型的详细信息见表A.2。

ESM-1b The ESM-1b hyperparameter sweep and model is described in detail in Appendix B. In brief, ESM-1b is the result of an extensive hyperparameter sweep that was performed on smaller 12-layer models. ESM-1b is the result of scaling up that model to 33 layers. Compared to the Uniref models, the main changes in ESM-1b are: higher learning rate; dropout after word embedding; learned positional embeddings; final layer norm before the output; and tied input/output word embedding.

ESM-1b ESM-1b的超参数扫描和模型在附录B中有详细描述。简而言之，ESM-1b是在较小的12层模型上进行广泛超参数扫描的结果。ESM-1b是将该模型扩展到33层的结果。与Uniref模型相比，ESM-1b的主要变化包括:更高的学习率；词嵌入后的dropout；学习到的位置嵌入；输出前的最终层归一化；以及绑定的输入/输出词嵌入。

Pre-training task The masked language modeling pretraining task follows Devlin et al. (2018b). Specifically, we select as supervision 15% of tokens randomly sampled from the sequence. For those of tokens, we change the input token to a special "masking" token with probability, a randomly-chosen alternate amino acid token with 10% probability, and the original input token (i.e. no change) with probability. We take the loss to be the whole batch average cross entropy loss between the model’s predictions and the true token for these of amino acid tokens. In contrast to Devlin et al. (2018b), we do not use any additional auxiliary prediction losses. The ESM-1b models, as well as the UniParc development models used in visualizations and in the supplemental results are trained with the masking procedure above. The UniRef models used across the experiments of the main text are trained similarly, except that for the of tokens selected as prediction targets, all are replaced by the mask token.

预训练任务 掩码语言建模预训练任务遵循Devlin等人(2018b)的方法。具体来说，我们从序列中随机抽取15%的标记作为监督。对于这些 的标记，我们以 的概率将输入标记更改为特殊的“掩码”标记，以10%的概率更改为随机选择的替代氨基酸标记，并以 的概率保留原始输入标记(即不更改)。我们将损失计算为模型预测与这些 氨基酸标记的真实标记之间的整个批次的平均交叉熵损失。与Devlin等人(2018b)不同，我们没有使用任何额外的辅助预测损失。ESM-1b模型以及用于可视化和补充结果的UniParc开发模型都使用上述掩码程序进行训练。主文本实验中使用的UniRef模型也以类似方式训练，只是对于被选为预测目标的 标记，所有标记都被替换为掩码标记。

Pre-training details Our model was pre-trained using a context size of 1024 tokens. As most Uniparc sequences (96.7%) contain fewer than 1024 amino acids, the Transformer is able to model the entire context in a single model pass. For those sequences that are longer than 1024 tokens, we sampled a random crop of 1024 tokens during each training epoch. The model was optimized using Adam with learning rate . We trained with 131,072 tokens per batch (128 gpus x 1024 tokens). The models follow a warm-up period of 16000 updates, during which the learning rate increases linearly. Afterwards, the learning rate follows an inverse square root decay schedule. All models were trained using the fairseq toolkit (Ott et al., 2019) on 128 NVIDIA V100 GPUs.

预训练细节 我们的模型使用1024个标记的上下文大小进行预训练。由于大多数Uniparc序列(96.7%)包含少于1024个氨基酸，Transformer能够在一次模型传递中建模整个上下文。对于那些超过1024个标记的序列，我们在每个训练周期中随机裁剪1024个标记。模型使用Adam 进行优化，学习率为 。我们以每批次131,072个标记(128个GPU x 1024个标记)进行训练。模型在16000次更新的预热期内，学习率线性增加。之后，学习率遵循逆平方根衰减计划。所有模型都使用fairseq工具包(Ott等人，2019)在128个NVIDIA V100 GPU上进行训练。

# 5. Evaluating the models for downstream tasks

# 5. 评估下游任务的模型

After pre-training the model with unsupervised learning, we can adapt the parameters to supervised tasks. By passing the input sequence through our pre-trained model, we obtain a final vector representation of the input sequence . During pre-training, this representation is projected to log probabilities . Recall that a soft- over represents the model’s posterior for the amino acid at position . These final representations are used directly, or fine-tuned in a task-dependent way by adding additional layers to the model and allowing the gradients to backpropagate through the weights of the pre-trained model to adapt them to the new task. Hidden representations from intermediate layers rather than the final layer can also be used.

在使用无监督学习预训练模型后，我们可以将参数适应到有监督任务中。通过将输入序列 传递到我们的预训练模型中，我们获得输入序列 的最终向量表示。在预训练期间，该表示被投影为对数概率 。回想一下，soft- over 表示模型对位置 氨基酸的后验概率。这些最终表示 可以直接使用，或者通过向模型添加额外的层并以任务依赖的方式进行微调，允许梯度通过预训练模型的权重反向传播，以适应新任务。也可以使用中间层而不是最终层的隐藏表示。

# 6. Language Modeling Baselines

# 6. 语言建模基线

In addition to comparing to past work, we also implemented deep learning baselines for our experiments.

除了与过去的工作进行比较外，我们还为实验实现了深度学习基线。

Frequency (n-gram) models To establish a meaningful performance baseline on the sequence modeling task (Section 3), we construct n-gram frequency-based models for context sizes , applying optimal Laplace smoothing for each context size. The Laplace smoothing hyperparameter in each case was tuned on the validation set. ECE is reported for the best left-conditioning n-gram model.

频率(n-gram)模型 为了在序列建模任务(第3节)上建立有意义的性能基线，我们为上下文大小 构建了基于n-gram频率的模型，对每个上下文大小应用最优的拉普拉斯平滑。每种情况下的拉普拉斯平滑超参数都在验证集上进行了调整。报告了最佳左条件n-gram模型的ECE。

Bidirectional LSTM language models We trained state-of-the-art LSTM (Hochreiter & Schmidhuber, 1997) language models on the UR50 dataset. We use the ELMo model of Peters et al. (2018) which concatenates two independent autoregressive language models with left-to-right and right-to-left factorization. Unlike standard LSTM language models, the ELMo model receives context in both directions and is therefore comparable to the Transformers we train that also use the whole context of the sequence. We train two models: (i) the small model has approximately parameters across 3 layers, with an embedding dimension of 512 and a hidden dimension of 1024; (ii) the large model has approximately parameters across 3 layers, with an embedding dimension of 512 and a hidden dimension of 4096 . The models are trained with a nominal batch size of 32,768, with truncated backpropagation to 100 tokens, dropout of 0.1, learning rate of , using the Adam optimizer with betas of(0.9,0.999), clip norm 0.1 and warmup of 1500 updates using an inverse square root learning rate schedule. We searched across a range of learning rates and found 8e-4 to be optimal.

双向LSTM语言模型 我们在UR50数据集上训练了最先进的LSTM(Hochreiter & Schmidhuber, 1997)语言模型。我们使用了Peters等人(2018)提出的ELMo模型，该模型将两个独立的自回归语言模型与从左到右和从右到左的因子分解进行拼接。与标准的LSTM语言模型不同，ELMo模型接收双向的上下文，因此与我们训练的同样使用序列完整上下文的Transformer模型具有可比性。我们训练了两个模型:(i)小型模型在3层中大约有 个参数，嵌入维度为512，隐藏维度为1024；(ii)大型模型在3层中大约有 个参数，嵌入维度为512，隐藏维度为4096。模型训练时，名义批量大小为32,768，截断反向传播至100个词元，dropout为0.1，学习率为 ，使用Adam优化器，betas为(0.9,0.999)，梯度裁剪范数为0.1，并使用逆平方根学习率调度进行1500次更新的预热。我们在一定范围内搜索了学习率，发现8e-4为最优值。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)为作者/资助者，其已授予bioRxiv永久展示该预印本的许可。本预印本在CC-BY-NC-ND 4.0国际许可下提供。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| #Layers | #Heads | Embedding Dim | MLP Dim | #Params | Steps |
| 12 | 12 | 768 | 3072 | 85.1M | 1.6M |
| 24 | 12 | 768 | 3072 | 170.2M | 300k |
| 36 | 20 | 1280 | 5120 | 708.6M | 300k |
| #层数 | #头数 | 嵌入维度 | MLP维度 | #参数 | 步骤 |
| 12 | 12 | 768 | 3072 | 85.1M | 1.6M |
| 24 | 12 | 768 | 3072 | 170.2M | 300k |
| 36 | 20 | 1280 | 5120 | 708.6M | 300k |

Table A.1. Hyperparameters for development Transformer models trained on UniParc. Embedding dim is the dimension of the hidden states at the output of each transformer block. MLP Dim refers to the width of hidden layer in the Transformer’s MLPs.

表 A.1. 在 UniParc 上训练的 Transformer 开发模型的超参数。嵌入维度(Embedding dim)是每个 Transformer 块输出时隐藏状态的维度。MLP 维度(MLP Dim)指的是 Transformer 的 MLP 中隐藏层 的宽度。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| #Layers | #Heads | Embedding Dim | MLP Dim | #Params | Data | Steps |
| 6 | 12 | 768 | 3072 | 42.6M | UR50/S | 840K |
| 12 | 12 | 768 | 3072 | 85.1M | UR50/S | 840K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR100 | 275K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR50/S | 840K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR50/D | 906K |
| #层数 | #头数 | 嵌入维度 | MLP维度 | #参数 | 数据 | 步骤 |
| 6 | 12 | 768 | 3072 | 42.6M | UR50/S | 840K |
| 12 | 12 | 768 | 3072 | 85.1M | UR50/S | 840K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR100 | 275K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR50/S | 840K |
| 34 | 20 | 1280 | 5120 | 669.2M | UR50/D | 906K |

Table A.2. Hyperparameters for UniRef Transformer models. Note: UR100 model stopped making significant progress on valid loss and was stopped at updates.

表A.2. UniRef Transformer模型的超参数。注意:UR100模型在验证损失上停止显著进展，并在 次更新时停止。

# 7. Metric structure experiments

# 7. 度量结构实验

Dataset An orthologous group dataset was constructed from eggNOG 5.0 (Huerta-Cepas et al., 2018) by selecting 25 COG orthologous groups toward maximizing the size of the intersected set of species within each orthologous group. Through a greedy algorithm, we selected 25 COG groups with an intersecting set of 2,609 species. We shrank the dataset above by selecting only one species from each of 24 phyla in order to ensure species-level diversity.

数据集 从eggNOG 5.0(Huerta-Cepas等，2018)构建了一个直系同源组数据集，通过选择25个COG直系同源组以最大化每个直系同源组内物种的交集大小。通过贪心算法，我们选择了25个COG组，其交集包含2,609个物种。为了确保物种水平的多样性，我们从24个门中各选择一个物种来缩小上述数据集。

# 8. Remote Homology

# 8. 远程同源性

Dataset We used the database of SCOP 2.07e filtered to sequence similarity, provided by the ASTRAL compendium (Fox et al., 2014). Following standard practices (Söding & Remmert, 2011), we exclude folds that are known to be related, specifically Rossman-like folds (c.2- c. 5, c. 27 and 28, c. 30 and 31) and four- to eight-bladed - propellers (b.66-b.70). This yields 256,806 pairs of remote homologs at the fold level and 92,944 at the superfamily level, consisting of 217 unique folds and 366 unique superfamilies. We then perform an 80-20 split, and tune our hyperparameters on the "training set" and report results on the held out of the data.

数据集 我们使用了SCOP 2.07e数据库，过滤到 序列相似性，由ASTRAL汇编提供(Fox等，2014)。遵循标准做法(Söding & Remmert，2011)，我们排除了已知相关的折叠，特别是Rossman样折叠(c.2-c.5，c.27和28，c.30和31)以及四到八叶 螺旋桨(b.66-b.70)。这产生了256,806对折叠水平的远程同源物和92,944对超家族水平的远程同源物，包括217个独特的折叠和366个独特的超家族。然后我们进行80-20分割，并在“训练集”上调整超参数，并在保留的 数据上报告结果。

Metrics Given a protein sequence , with final hidden representation , we define the embedding of the sequence to be a vector which is the average of the hidden representations across the positions in the sequence:

度量 给定一个蛋白质序列 ，其最终隐藏表示为 ，我们将序列的嵌入定义为一个向量 ，它是序列中所有位置的隐藏表示的平均值:

We can compare the similarity of two protein sequences, and having embeddings and using a metric in the embedding space.

我们可以比较两个蛋白质序列 和 的相似性，它们分别具有嵌入 和 ，使用嵌入空间中的度量。

We evaluate the L2 distance and the cosine distance . Additionally we evaluated the L2 distance after projecting the vectors to the unit sphere.

我们评估了L2距离 和余弦距离 。此外，我们评估了将 向量投影到单位球后的L2距离。

Evaluation To evaluate HHblits (Remmert et al., 2011), first we construct HMM profiles for each sequence using default parameters for ’hhblits’, except we use 3 iterations. Then, we do an all-to-all alignment using ’hhalign’ with default parameters, and use the resulting E-value as a measure of similarity. Given a query sequence, a sequence is more similar with a smaller E-value.

评估 为了评估HHblits(Remmert等，2011)，首先我们使用’hhblits’的默认参数为每个序列构建HMM配置文件，除了我们使用3次迭代。然后，我们使用’hhalign’进行全对全比对，并使用默认参数，将得到的E值作为相似性的度量。给定一个查询序列，E值越小，序列越相似。

The two metrics reported are Hit-10 as introduced in Ma et al. (2014) and AUC. For both metrics, for each sequence, we treat it as a query and we rank each other sequence according to the distance metric used. Following Ma et al. (2014), for evaluation at the fold level, any domain with the same fold is a positive; any domain with a different fold is a negative; and domains belonging to the same superfamily are excluded. For evaluation at the superfamily level, any domain with the same superfamily is a positive; any domain with a different superfamily is a negative; and domains belonging to the same family are excluded. This ensures bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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. we specifically measure how well our models do on finding remote homologs.

报告的两个度量是Ma等(2014)引入的Hit-10和AUC。对于这两个度量，对于每个序列，我们将其视为查询，并根据使用的距离度量对其他序列进行排序。遵循Ma等(2014)，在折叠水平评估时，任何具有相同折叠的域为正；任何具有不同折叠的域为负；属于同一超家族的域被排除。在超家族水平评估时，任何具有相同超家族的域为正；任何具有不同超家族的域为负；属于同一家族的域被排除。这确保了我们专门测量模型在寻找远程同源物方面的表现。

For Hit-10, we consider the query a success if any of the top 10 sequences is a remote homolog. We report the proportion of successes averaged across all queries. For AUC, we first compute the AUC under the ROC curve when classifying sequences by vector similarity to the query. Then, we average the AUC across all query sequences.

对于Hit-10，如果前10个序列中有任何一个序列是远程同源物，则认为查询成功。我们报告所有查询的平均成功率。对于AUC，我们首先计算在通过向量相似性对序列进行分类时的ROC曲线下的AUC。然后，我们计算所有查询序列的AUC平均值。

We found that cosine similarity results in the best Hit-10 scores, while the L2 with unnormalized vectors result in the best AUC scores, so we report this in Table 2.

我们发现余弦相似性在Hit-10得分上表现最佳，而未归一化向量的L2距离在AUC得分上表现最佳，因此我们在表2中报告了这一点。

Implementation We used the FAISS similarity search engine (Johnson et al., 2017).

实现 我们使用了FAISS相似性搜索引擎(Johnson等，2017)。

# 9. Representational similarity-based alignment of sequences within MSA families

# 9. 基于表示相似性的MSA家族内序列对齐

Family selection We use the Pfam database (Bateman et al., 2013). We first filtered out any families whose longest sequence is less than 32 residues or greater than 1024 residues in length. We then ranked the families by the number of sequences contained in each family and selected the 128 largest families and associated MSAs. Finally, we reduced the size of each family to 128 sequences by uniform random sampling.

家族选择 我们使用Pfam数据库(Bateman等，2013)。我们首先过滤掉任何最长序列长度小于32个残基或大于1024个残基的家族。然后我们按每个家族中包含的序列数量对家族进行排序，并选择128个最大的家族及其相关的MSA。最后，我们通过均匀随机抽样将每个家族的大小减少到128个序列。

Aligned pair distribution For each family, we construct an empirical distribution of aligned residue pairs by enumerating all pairs of positions and indices that are aligned within the MSA and uniformly sampling 50,000 pairs.

对齐对分布 对于每个家族，我们通过枚举MSA中对齐的所有位置和索引对，并均匀抽样50,000对，构建对齐残基对的经验分布。

Unaligned pair distribution We also construct for each family a background empirical distribution of unaligned residue pairs. This background distribution needs to control for within-sequence position, since the residues of two sequences that have been aligned in an MSA are likely to occupy similar positions within their respective unaligned source sequences. Without controlling for this bias, a difference in the distributions of aligned and unaligned pairs could arise from representations encoding positional information rather than actual context. We control for this effect by sampling from the unaligned-pair distribution in proportion to the observed positional differences from the aligned-pair distribution. Specifically, the following process is repeated for each pair in the empirical aligned distribution:

未对齐对分布 我们还为每个家族构建了一个未对齐残基对的背景经验分布。这个背景分布需要控制序列内位置，因为在MSA中对齐的两个序列的残基很可能在它们各自的未对齐源序列中占据相似的位置。如果不控制这种偏差，对齐对和未对齐对分布的差异可能来自于表示编码的位置信息，而不是实际上下文。我们通过从未对齐对分布中按与对齐对分布中观察到的位置差异成比例地抽样来控制这种影响。具体来说，对于经验对齐分布中的每一对，重复以下过程:

1. Calculate the absolute value of the difference of each residue’s within-sequence positions in the aligned pair.

1. 计算对齐对中每个残基在序列内位置的差异的绝对值。

2. Select a pair of sequences at random.

2. 随机选择一对序列。

3. For that pair of sequences, select a pair of residues at random whose absolute value of positional difference equals the one calculated above.

3. 对于那对序列，随机选择一对残基，其位置差异的绝对值等于上面计算的值。

4. Verify that the residues are unaligned in the MSA; if so, add the pair to the empirical background distribution.

4. 验证这些残基在MSA中未对齐；如果是，将该对添加到经验背景分布中。

5. Otherwise, return to step 2 .

5. 否则，返回步骤2。

This procedure suffices to compute a empirical background distribution of 50,000 unaligned residue pairs.

此过程足以计算50,000个未对齐残基对的经验背景分布。

Similarity distributions Finally, for each family and each distribution, we apply the cosine similarity operator to each pair of residues to obtain the per-family aligned and unaligned distribution of representational cosine similarities.

相似性分布 最后，对于每个家族和每个分布，我们对每对残基应用余弦相似性算子，以获得每个家族的对齐和未对齐表示余弦相似性分布。

# 10. Linear projections

# 10. 线性投影

Dataset We construct five-fold cross validation datasets with structural hold-outs at the family, superfamily, and fold level using SCOPe 2.07 (Fox et al., 2014). We use the full version of SCOPe 2.07, clustered at 90% sequence identity, generated on January 23, 2020, and extract the domain annotations with labels. There are 19,695 domains. Then, independently for each hold-out level, we split the domains at the hold-out level into 5 equal sets, i.e. five sets of folds, superfamilies, or families. This ensures that for each partition, structures having the same classification do not appear in both the train and test sets. For a given classification level each structure appears in a test set once, so that in the cross validation experiment each of the structures will be evaluated once. Scores reported are the mean and standard deviation over each of the five test sets.

数据集 我们使用SCOPe 2.07(Fox等，2014)构建了在家族、超家族和折叠级别上具有结构保留的五折交叉验证数据集。我们使用2020年1月23日生成的90%序列同一性聚类的SCOPe 2.07完整版本，并提取带有标签的域注释。共有19,695个域。然后，对于每个保留级别，我们将保留级别的域分成5个相等的集合，即五组折叠、超家族或家族。这确保了对于每个分区，具有相同分类的结构不会同时出现在训练集和测试集中。对于给定的分类级别，每个结构在测试集中出现一次，因此在交叉验证实验中，每个结构将被评估一次。报告的分数是五个测试集的平均值和标准差。

For each domain, we first obtain the sequence whose residues align with the domain specification. To construct the secondary structure labels, we take each CIF file (pulled from PDB), and run DSSP (Joosten et al., 2010; Kabsch &Sander,1983). If has any residues where DSSP has not provided a secondary structure label, we mark them as missing data and do not supervise for those positions.

对于每个域，我们首先获得与域规范对齐的序列 。为了构建二级结构标签，我们获取每个CIF文件(从PDB中提取)，并运行DSSP(Joosten等，2010；Kabsch & Sander，1983)。如果 有任何残基DSSP未提供二级结构标签，我们将其标记为缺失数据，并不对这些位置进行监督。

To construct the contact map, we obtain coordinates from the structure portion of the CIF file in the case of glycine), defaulting to NaN where information is missing, and finally calculating pairwise distances and thresholding at . Similar to secondary structure, we do not supervise over NaNs.

为了构建接触图，我们从CIF文件的结构部分获取 坐标(在甘氨酸的情况下为 )，在信息缺失时默认为NaN，最后计算成对距离并在 处进行阈值处理。与二级结构类似，我们不监督NaN。

We discard any domains where (1) DSSP fails, (2) we are unable to align the sequence to the structure, or (3) the domain is longer than 1023 residues. This leaves 15,297 domains.

我们丢弃了以下情况下的任何域:(1) DSSP失败，(2) 我们无法将序列与结构对齐，或 (3) 域的长度超过1023个残基。这样剩下15,297个域。

For the MSA baselines, we query each sequence against the Uniclust30, 2017 database (Mirdita et al., 2017) with HH-blits (Remmert et al., 2011) using the default settings with additional parameters . For secondary structure prediction, we construct HMM profiles using HHmake (default settings). For contact prediction, we apply CCMpred (Seemayer et al., 2014) implementation of pseudolikelihood maximization (Balakrishnan et al., 2011; Ekeberg et al., 2013) using default settings with to each MSA, from which we extract both an output matrix , as well as a sequence profile where is the length of the sequence and is the size of the amino acid vocab, i.e. 25.

对于MSA基线，我们使用HH-blits(Remmert等，2011)在Uniclust30, 2017数据库(Mirdita等，2017)中查询每个序列，使用默认设置和附加参数 。对于二级结构预测，我们使用HHmake(默认设置)构建HMM配置文件。对于接触预测，我们使用CCMpred(Seemayer等，2014)实现的伪似然最大化(Balakrishnan等，2011；Ekeberg等，2013)，使用默认设置和 对每个MSA进行处理，从中提取输出矩阵 以及序列配置文件 ，其中 是序列长度， 是氨基酸词汇表的大小，即25。

Representations To obtain sequence representations, we provide the sequence of the domain as input to a forward pass of the Transformer model. We retrieve the activations into the final multi-head attention (after the layer normalization), using this as the matrix of sequence representations where is the hidden dimension of the model.

表示 为了获得序列表示，我们将域的序列作为输入提供给Transformer模型的前向传递。我们检索最终多头注意力(在层归一化之后)的激活，将其用作序列表示矩阵 ，其中 是模型的隐藏维度。

Secondary structure projections For secondary structure, we fit a multi-class logistic regression taking as input an individual representation and as output the secondary structure label from DSSP. We observe that the logistic regression model’s performance does not change with penalty settings (L1, L2, no penalty); therefore we report the result where the L2 penalty is applied during training.

二级结构投影 对于二级结构，我们拟合一个多类逻辑回归，将单个表示 作为输入，DSSP的二级结构标签作为输出。我们观察到逻辑回归模型的性能不随惩罚设置(L1、L2、无惩罚)而变化；因此我们报告在训练期间应用L2惩罚的结果。

Contact projections For contacts, we fit two linear projections. Given two representations at positions and , we regress to whether residues at positions and are in contact:

接触投影 对于接触，我们拟合两个线性投影。给定位置 和 的两个表示 ，我们回归到位置 和 的残基是否接触:

With learned projections , vector biases , and scalar bias . We use the AdamW optimizer (Loshchilov & Hutter, 2017) to fit the projections and bias term.

使用学习的投影 、向量偏差 和标量偏差 。我们使用AdamW优化器(Loshchilov & Hutter，2017)来拟合投影和偏差项。

For each partition we set aside approximately of the training set as validation. We sweep over a range of projection dimensions (32 to 512), learning rates (1e-6 to 1e-2) and weight decay values ( 0 to 0.9 ). Based on the best validation Top-L, long-range precision score, we set the projection dimension to 128, the learning rate to 1e-4, and the optimizer weight decay to 0.2 . We observe that the precision score does not improve with an increased projection dimension over 128 .

对于每个分区，我们将训练集的大约 作为验证集。我们扫描了一系列投影维度(32到512)、学习率(1e-6到1e-2)和权重衰减值(0到0.9)。基于最佳验证Top-L、长程精度得分，我们将投影维度设置为128，学习率设置为1e-4，优化器权重衰减设置为0.2。我们观察到，当投影维度超过128时，精度得分没有提高。

The above setup for contact projections only applies to the Transformer models and the sequence profile baseline. No supervision is applied to the CCMpred output.

上述接触投影的设置仅适用于Transformer模型和序列配置文件基线。CCMpred输出没有应用监督。

# 11. Single-family data and analysis

# 11. 单家族数据和分析

For each of the three domains used, we extracted all domain sequences from the Pfam dataset (Bateman et al., 2013) and located the subset of PDB files containing the domain, using the latter to derive ground truth secondary structure labels (Kabsch & Sander, 1983).

对于使用的三个域中的每一个，我们从Pfam数据集(Bateman等，2013)中提取了所有域序列，并定位了包含该域的PDB文件子集，使用后者来推导真实的二级结构标签(Kabsch & Sander，1983)。

Pre-training is with the masked language modeling objective, using the same hyperparameters as used to train the UniParc development models. The domain sequences were randomly partitioned into training, validation, and testing datasets. For each family, the training dataset comprises 65,536 sequences, the validation dataset comprises either 16,384 sequences (PF00005 and PF00069) or 8,192 sequences (PF00072), and the test dataset comprises the remainder.

预训练使用掩码语言建模目标，使用与训练UniParc开发模型相同的超参数。域序列被随机划分为训练、验证和测试数据集。对于每个家族，训练数据集包含65,536个序列，验证数据集包含16,384个序列(PF00005和PF00069)或8,192个序列(PF00072)，测试数据集包含其余序列。

Each Pfam family also forms an evaluation dataset for linear projection; from the sequences with corresponding crystal structures, the training dataset comprises 80 sequences and the test dataset comprises the remainder.

每个Pfam家族也形成线性投影的评估数据集；从具有相应晶体结构的序列中，训练数据集包含80个序列，测试数据集包含其余序列。

# 12. Secondary structure prediction. Deep neural networks and feature combination

# 12. 二级结构预测。深度神经网络和特征组合

We use features from the final hidden representations of the models. We removed the final embedding layer, added layer norm, and applied a top-level architecture following (Klausen et al., 2019). In particular, this top-level architecture consists of two parallel convolution layers and an identity layer, whose outputs are concatenated in the feature dimension and fed to a two layer bidirectional LSTM containing 1024 hidden units and dropout . The output is then projected to an 8-dimensional feature vector at each position and the model is trained with a categorical cross-entropy loss with the Q8 labels. The training data was obtained from the (Klausen et al., 2019). Secondary structure labels for the CASP13 test set were constructed using DSSP.

我们使用模型最终隐藏表示的特征。我们移除了最终的嵌入层，添加了层归一化，并应用了(Klausen等，2019)提出的顶层架构。具体来说，该顶层架构由两个并行的卷积层和一个恒等层组成，其输出在特征维度上连接并输入到一个包含1024个隐藏单元和dropout的双向LSTM中 。然后，输出在每个位置投影为一个8维特征向量，并使用Q8标签的类别交叉熵损失训练模型。训练数据来自(Klausen等，2019)。CASP13测试集的二级结构标签使用DSSP构建。

In feature combination experiments, we used the features provided by (Klausen et al., 2019) which were generated using MMseqs2 on the Uniclust90 dataset released April 2017. For CASP13 experiments, we generated these features using code provided by (Klausen et al., 2019) on CASP13 domains.

在特征组合实验中，我们使用了(Klausen等，2019)提供的特征，这些特征是使用MMseqs2在2017年4月发布的Uniclust90数据集上生成的。对于CASP13实验，我们使用(Klausen等，2019)提供的代码在CASP13域上生成了这些特征。

As a baseline, we reimplemented (Klausen et al., 2019) by replacing the Transformer features with the MMseqs2 features and keeping the top-level architecture. For feature combination experiments, we projected (a) the features from this baseline and (b) the features from the Transformer to the same dimension (512 units), concatenated along the feature dimension, and fed the resulting tensor to a two layer bidirectional LSTM with 512 hidden units and dropout . To check our dataset construction, we used the pretrained weights provided by (Klausen et al., 2019) and evaluated their model directly in our evaluation pipeline. We were able to reproduce the values reported in (Klausen et al., 2019).

作为基线，我们重新实现了(Klausen等，2019)的方法，用MMseqs2特征替换了Transformer特征，并保留了顶层架构。在特征组合实验中，我们将(a)该基线的特征和(b)Transformer的特征投影到相同的维度(512个单位)，在特征维度上连接，并将结果张量输入到一个包含512个隐藏单元和dropout的双向LSTM中 。为了检查我们的数据集构建，我们使用了(Klausen等，2019)提供的预训练权重，并在我们的评估管道中直接评估了他们的模型。我们能够复现(Klausen等，2019)中报告的值。

# 13. Contact prediction. Deep neural networks and feature combination

# 13. 接触预测。深度神经网络和特征组合

Data We use the datasets and features distributed with Wang et al. (2017) and Xu (2018). The base features are those used by RaptorX (Xu, 2018) a state-of-the-art method in CASP13, including sequence features, PSSM, 3-state secondary structure prediction, predicted accessibility, one-hot embedding of sequence, and pairwise features APC-corrected Potts model couplings, mutual information, pairwise contact potential.

数据 我们使用了Wang等(2017)和Xu(2018)发布的数据集和特征。基础特征是RaptorX(Xu，2018)在CASP13中使用的最先进方法，包括序列特征、PSSM、三态二级结构预测、预测的可及性、序列的独热编码以及成对特征APC校正的Potts模型耦合、互信息、成对接触势。

We use the training, standard test set, and CASP11 test set from Wang et al. (2017). We use the CASP12 test set from (2018). For the CASP13 test set we use the 34 publicly released domains.

我们使用了Wang等(2017)的训练集、标准测试集和CASP11测试集。我们使用了 (2018)的CASP12测试集。对于CASP13测试集，我们使用了34个公开发布的域。

Wang et al. (2017) established training and test sets as follows. The train (6,367 proteins), valid (400 proteins) and test (500 proteins) datasets were selected as subsets of PDB25 (each protein having sequence similarity). Proteins having sequence similarity or BLAST E-value with any test or CASP11 protein were excluded from training data.

Wang等(2017)按以下方式建立了训练和测试集。训练集(6,367个蛋白质)、验证集(400个蛋白质)和测试集(500个蛋白质)被选为PDB25的子集(每个蛋白质具有 序列相似性)。与任何测试或CASP11蛋白质具有序列相似性 或BLAST E值 的蛋白质被排除在训练数据之外。

All our MSAs (used for the avg and cov combination methods) are constructed by running HHblits (Remmert et al., 2011) with 3 iterations and E-value 0.001 against Uniprot20 released on 2016-02; except for CASP12 and CASP13 where we used the four different MSAs released with and described in Xu (2018). Note that for the Transformer pretraining UniRef50 from 2018-03 was used; hence no data which was not already available prior to the start of CASP13 was present during either pre-training or contact prediction training.

我们所有的MSA(用于avg和cov组合方法)都是通过运行HHblits(Remmert等，2011)构建的，使用3次迭代和E值0.001，针对2016年2月发布的Uniprot20；除了CASP12和CASP13，我们使用了Xu(2018)发布和描述的四种不同的MSA。请注意，对于Transformer预训练，使用了2018年3月的UniRef50；因此，在预训练或接触预测训练期间，没有使用在CASP13开始之前不可用的数据。

Model architecture On top of the sequence and pairwise features we use a depth-32 residual network (ResNet) model to predict binary contacts. The ResNet model architecture is similar to Wang et al. (2017) and Xu (2018).

模型架构 在序列和成对特征之上，我们使用了一个深度为32的残差网络(ResNet)模型来预测二元接触。ResNet模型架构与Wang等(2017)和Xu(2018)相似。

The first component of the ResNet is a learned sequence pipeline which maps sequence features to with the length of the protein. Though could be a 1D convolutional network or residual network as in Wang et al. (2017), we chose our sequence net to be a simple linear projection from input dimension to dimensions. The input dimension is either 46 (RaptorX only), 1280 (Transformer hidden state), or 1326 (feature combination). We varied and empirically determined 128 to be optimal.

ResNet的第一个组件是一个学习到的序列管道 ，它将序列特征 映射到 ，其中 是蛋白质的长度。尽管 可以像Wang等人(2017年)那样是一维卷积网络或残差网络，但我们选择我们的序列网络为从输入维度 到 维度的简单线性投影。输入维度 可以是46(仅RaptorX)、1280(Transformer隐藏状态)或1326(特征组合)。我们改变了 ，并通过经验确定128为最优值。

The 128-D output of the sequence net gets converted to pairwise matrix features with 256 feature maps, by an outer concatenation operation; i.e. at position we concatenate and along the feature dimension, giving rise to feature maps. This is then concatenated in the first (feature map or channel) dimension, with the pairwise features i.e. the pairwise RaptorX features described in previous subsection and/or the msa embedding covariance features described in the next subsection. As such the concatenated or .

序列网络的128维输出 通过外连接操作转换为具有256个特征图的成对矩阵特征 ；即在位置 处，我们沿特征维度连接 和 ，从而产生 个特征图。然后，这个 在第一个(特征图或通道)维度上与成对特征 连接，即前一节中描述的成对RaptorX特征和/或下一节中描述的msa嵌入协方差特征 。因此，连接后的 或 。

The final component is the actual 2D residual network operating in , which computes the binary contact probability with and the continuous predicted probability of position and of the protein being in contact. The ResNet has an initial convolutional layer going to feature maps, followed by MaxOut over the feature maps with stride 2, reducing to 64 feature maps. After this, there are 32 residual blocks. Each residual block has on its weight path consecutively BatchNorm - ReLU - Conv (64 feature maps) - Dropout (0.3) - ReLU - Conv (64 feature maps). The residual blocks have consecutive dilation rates of 1,2,4. This follows Adhikari (2019). The final output is computed with a Conv (1 output feature map) and sigmoid to produce probability of contact . As such there are 66 convolutional layers in the main 2D ResNet.

最后一个组件是在 中操作的实际二维残差网络，它计算二进制接触概率 ，其中 和 是蛋白质位置 和 处于接触状态的连续预测概率。ResNet有一个初始的 卷积层，输出到 个特征图，随后在特征图上进行步幅为2的MaxOut，减少到64个特征图。之后，有32个残差块。每个残差块在其权重路径上依次为BatchNorm - ReLU - Conv (64个特征图)- Dropout(0.3)- ReLU - Conv (64个特征图)。残差块的连续膨胀率为1,2,4。这遵循Adhikari(2019年)。最终输出通过Conv (1个输出特征图)和sigmoid计算，以产生接触概率 。因此，主二维ResNet中有66个卷积层。

Note that a number of shortcomings exist from our pipeline to CASP13 winners (Senior et al., 2018; Xu, 2018); most importantly we use an earlier training dataset of PDB structures compiled from PDB dated Feb 2015 by Wang et al. (2017), additionally we do not incorporate more recent developments like distance distribution prediction, sliding window on small crops allowing deeper ResNets, auxiliary losses like torsion angles, or data augmentation.

需要注意的是，我们的管道与CASP13获胜者(Senior等人，2018年；Xu，2018年)相比存在一些不足；最重要的是，我们使用了由Wang等人(2017年)从2015年2月的PDB结构中编译的早期训练数据集，此外我们没有纳入更近期的进展，如距离分布预测、在小裁剪上滑动窗口以允许更深的ResNets、辅助损失如扭转角或数据增强。

For reference, the officially released AlphaFold (Senior et al., 2018) predictions achieve a top-L/5, LR and top-L, LR precision on the same subset of CASP-13 targets of 75.2% and 52.2% respectively. The discrepancies in the pipeline explain why our best precisions using RaptorX features are about 7-9% lower (compare CASP13-AVG (a): 68.0% / 43.4%)

作为参考，官方发布的AlphaFold(Senior等人，2018年)预测在CASP-13目标的同一子集上实现了top-L/5, LR和top-L, LR的精度，分别为75.2%和52.2%。流程中的差异解释了为什么我们使用RaptorX特征的最佳精度大约低7-9%(比较CASP13-AVG (a): 68.0% / 43.4%)。

MSA Embedding feature combination. We construct features based on the embedding of the MSA of a protein sequence in our training data. We denote the original protein in our labeled dataset, i.e. query sequence of length , to have corresponding embedding Transformer , and the embedding of the -th position to be . Typically is the last hidden state from the pre-trained Transformer model. The th sequence in the MSA is , with corresponding embedding with the MSA depth. The embeddings are computed by embedding the original sequence without inserting gaps (there is no gap character in our vocabulary), then realigning the embedding according to the alignment between and query sequence by inserting 0 -vectors at position if the is the gap character; ie . We also use indicator variable if is non-gap (match state), or if is gap. We further compute sequence weights as the commonly used debiasing heuristic to reduce the influence of the oversampling of many similar sequences. The weights are defined in the usual way with 70% sequence similarity threshold: sequence weight which is the inverse of the number of sequences that are more than similar to the sequence i.e. hamming distance less than .

MSA嵌入特征组合。我们基于训练数据中蛋白质序列的MSA嵌入构建特征。我们将标记数据集中的原始蛋白质，即长度为 的查询序列 ，表示为具有相应嵌入 Transformer ，并且第 个位置的嵌入为 。通常 是预训练Transformer模型的最后一个隐藏状态。MSA中的第 个序列是 ，具有相应的嵌入 ，其中 是MSA深度。嵌入是通过嵌入原始序列 而不插入间隙(我们的词汇表中没有间隙字符)来计算的，然后根据 和查询序列 之间的对齐重新对齐嵌入，如果 是间隙字符，则在位置 插入0向量；即 。我们还使用指示变量 如果 是非间隙(匹配状态)，或 如果 是间隙。我们进一步计算序列权重 作为常用的去偏启发式方法，以减少许多相似序列过采样的影响。权重以通常的方式定义，具有70%的序列相似性阈值:序列权重 ，它是与序列 相似度超过 的序列数 的倒数，即汉明距离小于 。

Now we introduce the average embedding over an MSA:

现在我们介绍MSA上的平均嵌入:

with per-position denominator This is effectively a weighted average over the sequence embeddings in the MSA. Note that if the embeddings were one-hot encodings of AA identities, we would recover the position probability matrix (except the absence of a pseudo-count).

每个位置的分母 这实际上是MSA中序列嵌入的加权平均。请注意，如果嵌入是AA身份的单热编码，我们将恢复位置概率矩阵(除了缺少伪计数)。

Similarly; we introduce the (uncentered) covariance of the embeddings, with PCA-projected :

类似地；我们引入嵌入的(未中心化)协方差，具有PCA投影的 :

With pairwise position denominator

具有成对位置分母

Note that to make above covariance embedding feasible, we first reduce the dimensionality of the embeddings by projecting onto the first 16 PCA directions: with , giving rise to a covariance per pair of positions and pair of interacting PCA components of . The 256 different pairs of will now become the feature maps of , such that . We tried training (rather than fixed PCA) the projection of the features before covariance (learned linear projection or training a 3-layer MLP). We also varied the formulation to center the embeddings over the MSA (normal covariance) and to rescale the feature maps with a pre-computed mean and standard deviation for each feature map corresponding to a pair of . We found no gains from these variations over the current formulation. Note that centering with the average as in normal empirical covariance calculation, introduces a shift that is independent per protein (because specific to the MSA), and independent per position. Therefore it is not unexpected that the uncentered covariance gives better (more consistent) features.

请注意，为了使上述协方差嵌入可行，我们首先通过投影到前16个PCA方向来降低嵌入的维度: 与 ，从而产生每对位置 和每对相互作用的PCA组件 的协方差 。256个不同的 对 现在将成为 的特征图，使得 。我们尝试在协方差之前训练(而不是固定PCA)特征的投影 (学习的线性投影 或训练一个3层MLP)。我们还改变了公式，将嵌入集中在MSA上(正常协方差)，并使用预先计算的均值和标准差重新缩放特征图，每个特征图对应于一对 。我们发现这些变化在当前公式中没有带来任何增益。请注意，像正常经验协方差计算中那样使用平均值 进行中心化，会引入一个独立于每个蛋白质(因为特定于MSA)和每个位置的偏移。因此，未中心化的协方差提供更好(更一致)的特征并不意外。

# 14. Mutational Effect

# 14. 突变效应

Datasets We used two datasets of variant effect measurements compiled by Gray et al. (2018) and Riesselman et al. (2018). The first dataset is a collection of 21,026 measurements from nine experimental deep mutational scans. The second dataset contains 712,218 mutations across 42 deep mutational scans.

数据集 我们使用了Gray等人(2018)和Riesselman等人(2018)编制的两个变异效应测量数据集。第一个数据集包含来自九个实验性深度突变扫描的21,026个测量结果。第二个数据集包含42个深度突变扫描中的712,218个突变。

Fine-tuning procedure To fine-tune the model to predict the effect of changing a single amino acid or combination of amino acids we regress the scaled mutational effect with:

微调程序 为了微调模型以预测单个氨基酸或氨基酸组合变化的影响，我们使用以下公式对缩放的突变效应进行回归:

Where is the mutated amino acid at position , and wt(i)is the wildtype amino acid. The sum runs over the indices of the mutated positions. As an evaluation metric, we report the Spearman between the model’s predictions and experimentally measured values.

其中 是位置 处的突变氨基酸，wt(i)是野生型氨基酸。求和运行在突变位置的索引上。作为评估指标，我们报告了模型预测值与实验测量值之间的Spearman 。

# 15. Area under the ROC curve

# 15. ROC曲线下面积

For a binary classification task, the ROC curve plots the true positive rate against the false positive rate at various classification thresholds. The area under the ROC curve gives a measure that quantifies the model’s ability to distinguish between classes. Intuitively a perfect classifier has an AUC of 1, while a uniform random classifier has an AUC of 0.5 .

对于二分类任务，ROC曲线绘制了在不同分类阈值下的真阳性率与假阳性率。ROC曲线下面积给出了一个量化模型区分类别能力的指标。直观上，完美分类器的AUC为1，而均匀随机分类器的AUC为0.5。

# B. ESM-1b Hyperparameter Optimization

# B. ESM-1b超参数优化

Experimental setup We perform a systematic analysis on Transformer models with parameters. We train models on the UniRef50 dataset following the same methodology described in the rest of this work.

实验设置 我们对具有 参数的Transformer模型进行了系统分析。我们在UniRef50数据集上训练模型，遵循本文其余部分描述的相同方法。

After identifying the best performing settings on parameter models, we explore scale by training parameter models. All models are trained with a target batch size of tokens. To accommodate the large batch size, we use gradient accumulation and distributed data parallel. Under this setup, each epoch of a parameter model completes in 1.8 hours on 32 GPUs. Each epoch of a parameter model completes in 8.5 hours on 64 GPUs.

在确定 参数模型的最佳性能设置后，我们通过训练 参数模型来探索规模。所有模型都以 个标记为目标批量大小进行训练。为了适应大批量大小，我们使用梯度累积和分布式数据并行。在此设置下， 参数模型的每个epoch在32个GPU上完成需要1.8小时。 参数模型的每个epoch在64个GPU上完成需要8.5小时。

When studying architectural variants, we assess the quality of representations from each model after 10 or 12 epochs of pre-training. We observed that relative performance ranking of the models does not change after this point. Notably, this is still early in training; our best performing model, ESM-1b, is trained for 56 epochs.

在研究架构变体时，我们在10或12个epoch的预训练后评估每个模型的表示质量。我们观察到，在此之后，模型的相对性能排名不会改变。值得注意的是，这仍然是在训练的早期；我们表现最好的模型ESM-1b训练了56个epoch。

Hyperparameter: Masking and data setup Protein language models are trained with a masked language modeling objective, wherein each input sequence is corrupted by replacing a fraction of the tokens with a special mask token. We train parameter models for 10 epochs and compare their performance on the CB513 test set. We investigate four masking strategies:

超参数:掩码和数据设置 蛋白质语言模型通过掩码语言建模目标进行训练，其中每个输入序列通过用特殊掩码标记替换一部分标记来破坏。我们训练 参数模型10个epoch，并比较它们在CB513测试集上的性能。我们研究了四种掩码策略:

* All masks: following supplemental section 4, 15% of the input tokens are replaced with a mask token and predicted.
* 所有掩码:根据补充部分4，15%的输入标记被替换为掩码标记并进行预测。
* All random (uniform): 15% of the input tokens are replaced with an amino acid selected uniform randomly and predicted.
* 全部随机(均匀):15%的输入标记被随机均匀选择的氨基酸替换并预测。
* All random (frequency): 15% of the input tokens are replaced with an amino acid selected according to their frequency in the dataset and predicted.
* 全部随机(频率):15%的输入标记根据它们在数据集中出现的频率被选择的氨基酸替换并预测。
* BERT: of the input tokens are selected and predicted. Of these, are replaced with mask token; with a uniform random amino acid; not changed.
* BERT: 的输入标记被选择并预测。其中， 被替换为掩码标记； 被替换为均匀随机的氨基酸； 保持不变。

In all cases, we follow Liu et al. (2019) and dynamically mask the sequences, such that a new mask is randomly selected at each epoch. Since the input data changes across runs, language modeling perplexities cannot be fairly compared. Therefore, we evaluate the downstream performance of the models on a secondary structure benchmark. Table B. 1 finds that the BERT masking pattern performs better than the other masking patterns and is therefore used for all model variations below.

在所有情况下，我们遵循Liu等人(2019)的方法，动态地对序列进行掩码处理，使得在每个epoch随机选择一个新的掩码。由于输入数据在每次运行中都会变化，语言模型的困惑度无法公平比较。因此，我们在一个二级结构基准上评估模型的下游性能。表B.1发现BERT掩码模式比其他掩码模式表现更好，因此在下面的所有模型变体中使用。

|  |  |
| --- | --- |
| Masking pattern | CB513 |
| All masks | 60.4 |
| All random (uniform) | 59.3 |
| All random (frequency) | 59.0 |
| BERT | 60.8 |
| 掩码模式 | CB513 |
| 所有掩码 | 60.4 |
| 完全随机(均匀) | 59.3 |
| 完全随机(频率) | 59.0 |
| BERT | 60.8 |

Table B.1. Comparison of masking patterns, 8-class secondary structure prediction accuracy. Models are pre-trained for 10 epochs on Uniref50. The model and all other hyperparameters remain fixed between experiments. The BERT masking pattern performs best and is used for future experiments.

表 B.1. 掩码模式比较，8类二级结构预测准确率。模型在Uniref50上预训练了10个周期。模型和所有其他超参数在实验之间保持不变。BERT掩码模式表现最佳，并用于后续实验。

Hyperparameter: Dynamic batching Models in section 4 were trained with dynamic batching, which results in a single sequence in each sample in the batch. This design choice contrasts with existing language modeling works. For example, in NLP, Liu et al. (2019) and Devlin et al. (2018a), use a static batching approach, wherein multiple proteins are concatenated in the same batch along the sequence dimension. This approach is common in NLP, as sentences that are nearby in a corpus generally relate to the same topic. As this situation is not the case in protein language modeling, we analyze the impact of static batching schemes in protein language models, finding that they reduce model performance (Table B.2). parameter models are evaluated on the secondary structure downstream task after 10 epochs.

超参数:动态批处理 第4节中的模型使用动态批处理进行训练，这导致批次中的每个样本只有一个序列。这种设计选择与现有的语言建模工作形成对比。例如，在自然语言处理(NLP)中，Liu等人(2019)和Devlin等人(2018a)使用静态批处理方法，其中多个蛋白质在序列维度上连接在同一批次中。这种方法在NLP中很常见，因为语料库中相邻的句子通常与同一主题相关。由于这种情况在蛋白质语言建模中并不存在，我们分析了静态批处理方案对蛋白质语言模型的影响，发现它们降低了模型性能(表B.2)。 参数模型在10个周期后对二级结构下游任务进行评估。

|  |  |
| --- | --- |
| Batching mode | CB513 |
| Static | 56.2 |
| Static [cropping long sequences] | 57.0 |
| Dynamic | 60.8 |
| 批处理模式 | CB513 |
| 静态 | 56.2 |
| 静态 [裁剪长序列] | 57.0 |
| 动态 | 60.8 |

Table B.2. Comparison of batching modes, 8-class secondary structure accuracy. Batching modes are investigated on parameter models and trained for 10 epochs. All models are trained with a context size of 1024 tokens. In the static batching mode, sequences longer than 1024 can span multiple batches. We crop long sequences for the other batching modes considered. The dynamic batching scheme we used performs significantly better than static batching modes.

表 B.2. 批处理模式比较，8类二级结构准确率。批处理模式在 参数模型上进行研究，并训练了10个周期。所有模型均在上下文大小为1024个标记的情况下进行训练。在静态批处理模式中，长度超过1024的序列可以跨多个批次。对于其他考虑的批处理模式，我们对长序列进行裁剪。我们使用的动态批处理方案显著优于静态批处理模式。

Hyperparameters: further sweeps After fixing the data distribution to use the BERT masking scheme with dynamic batching, we next perform a hyperparameter sweep to identify the best learning rates, initializations and layer norm placement. We also propose a token dropout scheme which further improves performance on downstream tasks.

超参数:进一步扫描 在确定使用BERT掩码方案和动态批处理的数据分布后，我们接下来进行超参数扫描，以确定最佳学习率、初始化和层归一化位置。我们还提出了一种标记丢弃方案，进一步提高了下游任务的性能。

Initializations We compare our initializations to the initializations presented in Liu et al. (2019), finding that they perform similarly (Table B.3).

初始化 我们将我们的初始化与Liu等人(2019)提出的初始化进行比较，发现它们的表现相似(表 B.3)。

|  |  |
| --- | --- |
| Initializations | CB513 |
| Radford et al. (2019a) | 60.9 |
| Liu et al. (2019) | 60.8 |
| 初始化 | CB513 |
| Radford 等人 (2019a) | 60.9 |
| 刘等人 (2019) | 60.8 |

Table B.3. Comparison of initializations, 8-class secondary structure accuracy. parameter models are trained for 10 epochs and evaluated on CB513.

表 B.3. 初始化方法比较，8类二级结构准确率。 参数模型训练 10 个周期并在 CB513 上评估。

Layer norm placement Recent works Child et al. (2019) and Shoeybi et al. (2020) have suggested that pre-activation layer norm results in more stable training for larger models. We investigate the impact of this choice in a smaller controlled setting using parameter transformer models, finding that pre-activation layer norm improves performance on downstream tasks (Table B.4). To account for the lack

层归一化位置 最近的研究 Child 等人 (2019) 和 Shoeybi 等人 (2020) 表明，预激活层归一化对于较大模型的训练更稳定。我们在较小的受控环境中使用 参数变压器模型研究了这一选择的影响，发现预激活层归一化提高了下游任务的性能(表 B.4)。为了弥补

of a final layer norm, we additionally add a final layer norm before the linear output projection, improving performance (Table B.5).

缺少最终层归一化的问题，我们在线性输出投影之前额外添加了一个最终层归一化，从而提高了性能(表 B.5)。

|  |  |
| --- | --- |
| Model | CB513 |
| Post-activation | 60.8 |
| Pre-activation | 61.1 |
| 模型 | CB513 |
| 激活后 | 60.8 |
| 激活前 | 61.1 |

Table B.4. Comparison of layer norm placement, 8 class secondary structure accuracy. parameter models are trained for 10 epochs and evaluated on CB513.

表 B.4. 层归一化位置比较，8类二级结构准确率。 参数模型训练10个周期，并在CB513上进行评估。

|  |  |
| --- | --- |
| Final layer norm | CB513 |
| No | 61.3 |
| Yes | 61.9 |
| 最终层归一化 | CB513 |
| 否 | 61.3 |
| 是 | 61.9 |

Table B.5. Including a final layer norm before the language modeling head performs best, 8-class secondary structure accuracy. parameter models are trained for 12 epochs.

表 B.5。在语言建模头之前包含最终层归一化效果最佳，8类二级结构准确率。 参数模型训练了12个周期。

Token dropout Usually, masked language models are pre-trained with corrupted inputs and fine-tuned with complete sequences. We hypothesize that this shift in data distribution reduces performance on downstream tasks. Therefore, we propose a token dropout scheme which replaces the mask token embedding with a fixed tensor of zeros. As of positions are masked, the zero tensors cause a change in the mean statistics of the word embeddings. We therefore adjust the distribution during fine-tuning by multiplying the word embeddings by a fixed constant. Formally, if a mask token is introduced during pre-training with probability , then during fine-tuning, we multiply the embeddings by . We find that this significantly improves performance (Table B.6).

标记丢弃 通常，掩码语言模型使用损坏的输入进行预训练，并使用完整序列进行微调。我们假设这种数据分布的变化会降低下游任务的性能。因此，我们提出了一种标记丢弃方案，将掩码标记嵌入替换为固定的零张量。由于 的位置被掩码，零张量会导致词嵌入的平均统计量发生变化。因此，我们在微调期间通过将词嵌入乘以一个固定常数来调整分布。正式地，如果在预训练期间以概率 引入掩码标记，则在微调期间，我们将嵌入乘以 。我们发现这显著提高了性能(表 B.6)。

|  |  |
| --- | --- |
| Masking | CB513 |
| Mask Tokens | 61.9 |
| Token Dropout | 62.6 |
| 掩码 | CB513 |
| 掩码标记 | 61.9 |
| 标记丢弃 | 62.6 |

Table B.6. Token dropout performs better than mask tokens, 8- class secondary structure accuracy. parameter models are trained for 12 epochs.

表 B.6。Token dropout 比 mask tokens 表现更好，8 类二级结构准确率。 参数模型训练了 12 个周期。

Learning rate Next, we investigate learning rate, finding that a peak learning rate of performs best (Table B.7). Higher learning rates result in instability, while lower learning rates result in lower performance. In all cases, learning rate is warmed up linearly for 16000 steps and then decayed following an inverse square-root schedule (Raffel et al., 2020).

学习率 接下来，我们研究了学习率，发现峰值学习率为 时表现最佳(表 B.7)。较高的学习率会导致不稳定，而较低的学习率会导致性能下降。在所有情况下，学习率在前 16000 步线性预热，然后按照逆平方根计划衰减(Raffel 等，2020)。

|  |  |
| --- | --- |
| Learning rate | CB513 |
|  | 61.9 |
|  | 63.7 |
|  | 65.2 |
| 学习率 | CB513 |
|  | 61.9 |
|  | 63.7 |
|  | 65.2 |

Table B.7. A peak learning rate of performs best,8 class secondary structure accuracy. parameter models are trained for 10 epochs.

表 B.7。峰值学习率为 时表现最佳，8类二级结构准确率。 参数模型训练了10个周期。

|  |  |  |
| --- | --- | --- |
| Positional embeddings | Weights | CB513 |
| Harmonic | Untied | 63.7 |
| Harmonic | Tied | 64.8 |
| Learned | Untied | 65.7 |
| Learned | Tied | 66.0 |
| 位置嵌入 | 权重 | CB513 |
| 谐波 | 未绑定 | 63.7 |
| 谐波 | 绑定 | 64.8 |
| 学习到的 | 未绑定 | 65.7 |
| 学习到的 | 绑定 | 66.0 |

Table B.8. Comparison of tied embeddings and l.earned or harmonic positional embeddings, 8 class secondary structure accuracy. parameter models are trained for 17 epochs.

表 B.8. 比较固定嵌入与学习或谐波位置嵌入，8类二级结构准确率。 参数模型训练了17个周期。

Tying embeddings Although all experiments above are performed with tied input and output embeddings, we investigate whether learning these separately could improve performance. Our results (Table B.8) indicate that sharing the weights negatively impacts performance. Therefore, we maintain our initial setup.

嵌入绑定 尽管上述所有实验均使用绑定的输入和输出嵌入进行，但我们研究了分别学习这些嵌入是否能提高性能。我们的结果(表 B.8)表明，共享权重会对性能产生负面影响。因此，我们保持初始设置。

Positional embeddings We also investigate the impact of learning positional embeddings compared to fixed harmonic embeddings (Vaswani et al., 2017a), finding that learned positional embeddings perform better (Table B.8).

位置嵌入 我们还研究了学习位置嵌入与固定谐波嵌入(Vaswani 等，2017a)的影响，发现学习的位置嵌入表现更好(表 B.8)。

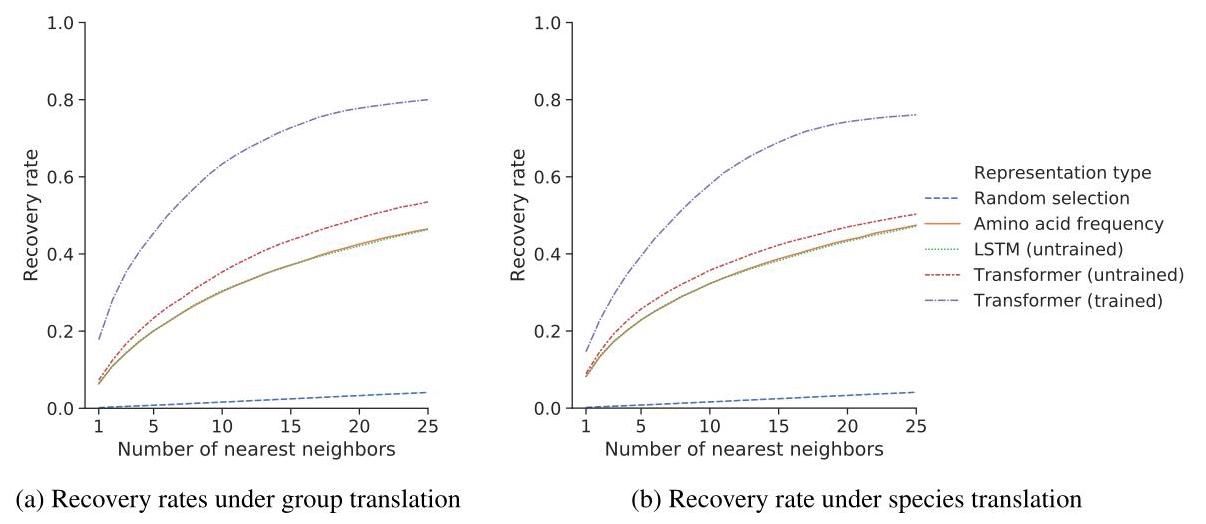


Figure S1. Learned sequence representations can be translated between orthologous groups and species. Depicted is the recovery rate of nearest-neighbor search under (a) orthologous group translation, and (b) species translation. In both settings, the trained Transformer representation space has a higher recovery rate. Results shown for 36-layer dev Transformer pre-trained on UniParc. To define a linear translation between protein and protein of the same species, we define the source and target sets as the average of protein or protein across all 24 diverse species. If representation space linearly encodes orthology, then adding the difference in these averages to protein of some species will recover protein in the same species. We use an analogous approach to translate a protein of a source species to its ortholog in the target species . Here, we consider the average representation of the proteins in and in . If representation space is organized linearly by species, then adding the difference in average representations to a protein in species will recover the corresponding protein in species .

图 S1. 学习的序列表示可以在同源群和物种之间进行转换。描绘的是在(a)同源群转换和(b)物种转换下最近邻搜索的恢复率。在这两种设置中，经过训练的 Transformer 表示空间具有更高的恢复率。结果显示的是在 UniParc 上预训练的 36 层 dev Transformer。为了定义同一物种中蛋白质 和蛋白质 之间的线性转换，我们将源集和目标集定义为所有 24 个不同物种中蛋白质 或蛋白质 的平均值。如果表示空间线性编码同源性，那么将这些平均值的差异添加到某个物种的蛋白质 中，将恢复同一物种中的蛋白质 。我们使用类似的方法将源物种 中的蛋白质转换为其在目标物种 中的同源物。在这里，我们考虑 和 中蛋白质的平均表示。如果表示空间按物种线性组织，那么将平均表示的差异添加到物种 中的蛋白质中，将恢复物种 中的相应蛋白质。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803；此版本发布于 2020 年 12 月 15 日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予 bioRxiv 永久展示预印本的许可。本预印本根据 CC-BY-NC-ND 4.0 国际许可提供。

|  |  |  |  |
| --- | --- | --- | --- |
| Representation type | Overall | Identical amino acid pairs | Distinct amino acid pairs |
| Transformer (trained) | 0.841 | 0.870 | 0.792 |
| Transformer (untrained) | 0.656 | 0.588 | 0.468 |
| 表示类型 | 总体 | 相同氨基酸对 | 不同氨基酸对 |
| Transformer(已训练) | 0.841 | 0.870 | 0.792 |
| Transformer(未训练) | 0.656 | 0.588 | 0.468 |

Table S1. Area under the ROC curve (AUC) of per-residue representational cosine similarities in distinguishing between aligned and unaligned pairs of residues within a Pfam family. Results displayed are averaged across 128 families.

表S1. 在区分Pfam家族中对齐和未对齐残基对时，每个残基表示余弦相似度的ROC曲线下面积(AUC)。显示的结果是128个家族的平均值。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可提供。

|  |  |  |  |
| --- | --- | --- | --- |
| Representation | PF00005 | PF00069 | PF00072 |
| Amino acid identity | 0.516 | 0.506 | 0.536 |
| 12-layer (untrained) | 0.818 | 0.719 | 0.835 |
| 12-layer (PF00005) | 0.864 | 0.725 | 0.842 |
| 12-layer (PF00069) | 0.816 | 0.842 | 0.850 |
| 12-layer (PF00072) | 0.789 | 0.688 | 0.888 |
| 12-layer (UniParc) | 0.900 | 0.872 | 0.906 |
| 36-layer (UniParc) | 0.902 | 0.884 | 0.902 |
| 表示 | PF00005 | PF00069 | PF00072 |
| 氨基酸一致性 | 0.516 | 0.506 | 0.536 |
| 12层(未训练) | 0.818 | 0.719 | 0.835 |
| 12层(PF00005) | 0.864 | 0.725 | 0.842 |
| 12层(PF00069) | 0.816 | 0.842 | 0.850 |
| 12层(PF00072) | 0.789 | 0.688 | 0.888 |
| 12层(UniParc) | 0.900 | 0.872 | 0.906 |
| 36层(UniParc) | 0.902 | 0.884 | 0.902 |

Table S2. Three-class secondary structure prediction accuracy by linear projection. Learning across many protein families produces better representations than learning from single protein families. Transformer models are trained on three PFAM families: ATP-binding domain of the ABC transporters (PF00005), Protein kinase domain (PF00069), and Response regulator receiver domain (PF00072). The single-family models are contrasted with models trained on the full UniParc data. Comparisons are relative to the family (columnwise), since each of the families differ in difficulty. Underline indicates models trained and evaluated on the same family. Representations learned from single families perform well within the family, but do not generalize as well to sequences outside the family. Representations trained on UniParc outperform the single-family representations in all cases.

表S2. 通过线性投影的三类二级结构预测准确率。跨多个蛋白质家族学习比从单一蛋白质家族学习产生更好的表示。Transformer模型在三个PFAM家族上训练:ABC转运蛋白的ATP结合域(PF00005)、蛋白激酶域(PF00069)和响应调节器接收域(PF00072)。单一家族模型与在完整UniParc数据上训练的模型进行了对比。比较是相对于家族(按列)进行的，因为每个家族的难度不同。下划线表示在同一家族上训练和评估的模型。从单一家族学习的表示在家族内表现良好，但在家族外的序列上泛化能力较差。在UniParc上训练的表示在所有情况下都优于单一家族表示。

under aCC-BY-NC-ND 4.0 International license.

根据CC-BY-NC-ND 4.0国际许可协议。

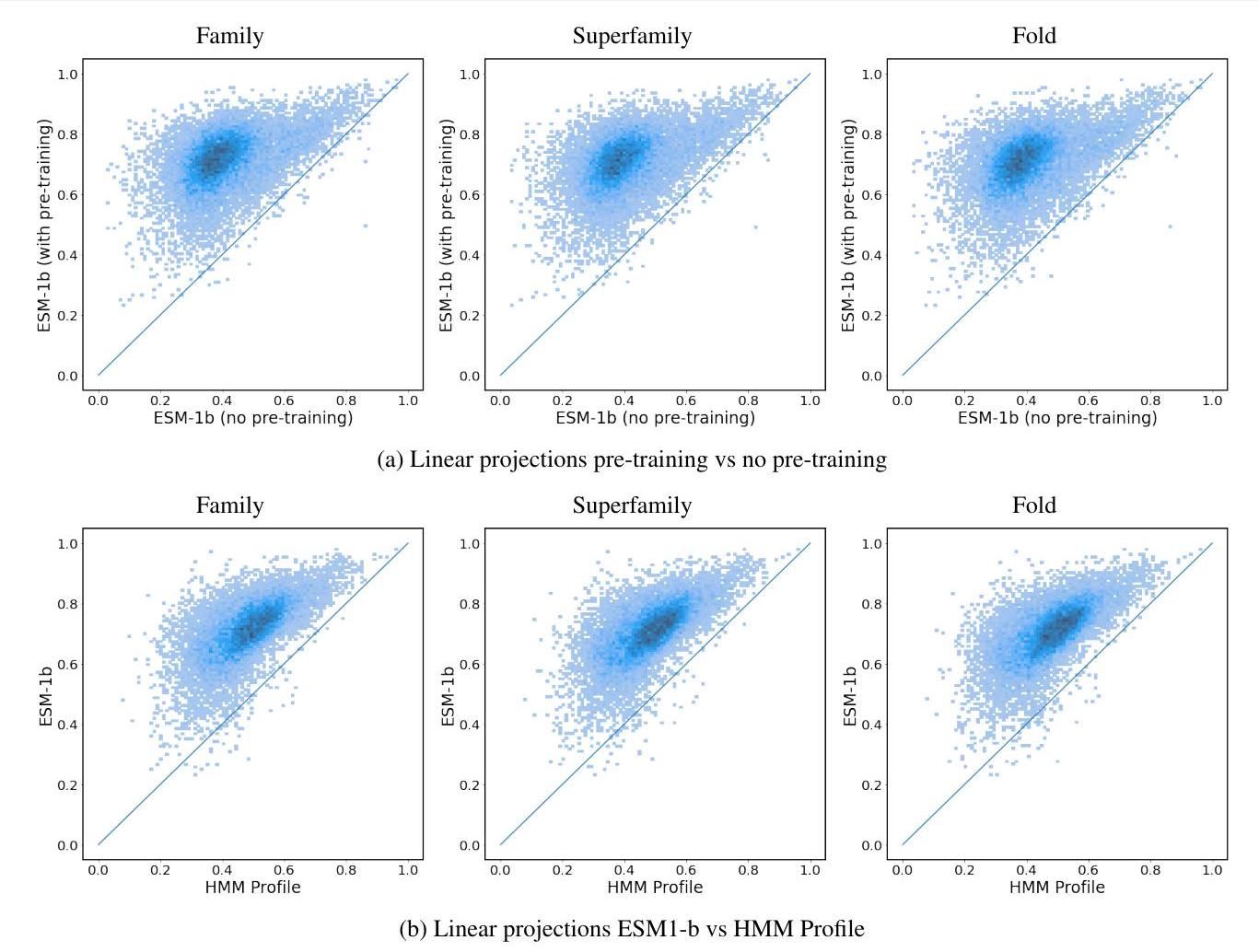


Figure S2. Secondary structure prediction 8-class accuracy distributions for linear projections. (a) Comparison with and without pretraining; (b) comparison of ESM-1b Transformer representations with HMM sequence profiles. Density is indicated by color.

图S2. 线性投影的八类二级结构预测准确率分布。(a) 有预训练和无预训练的比较；(b) ESM-1b Transformer表示与HMM序列谱的比较。密度由颜色表示。

under aCC-BY-NC-ND 4.0 International license.

根据CC-BY-NC-ND 4.0国际许可协议。

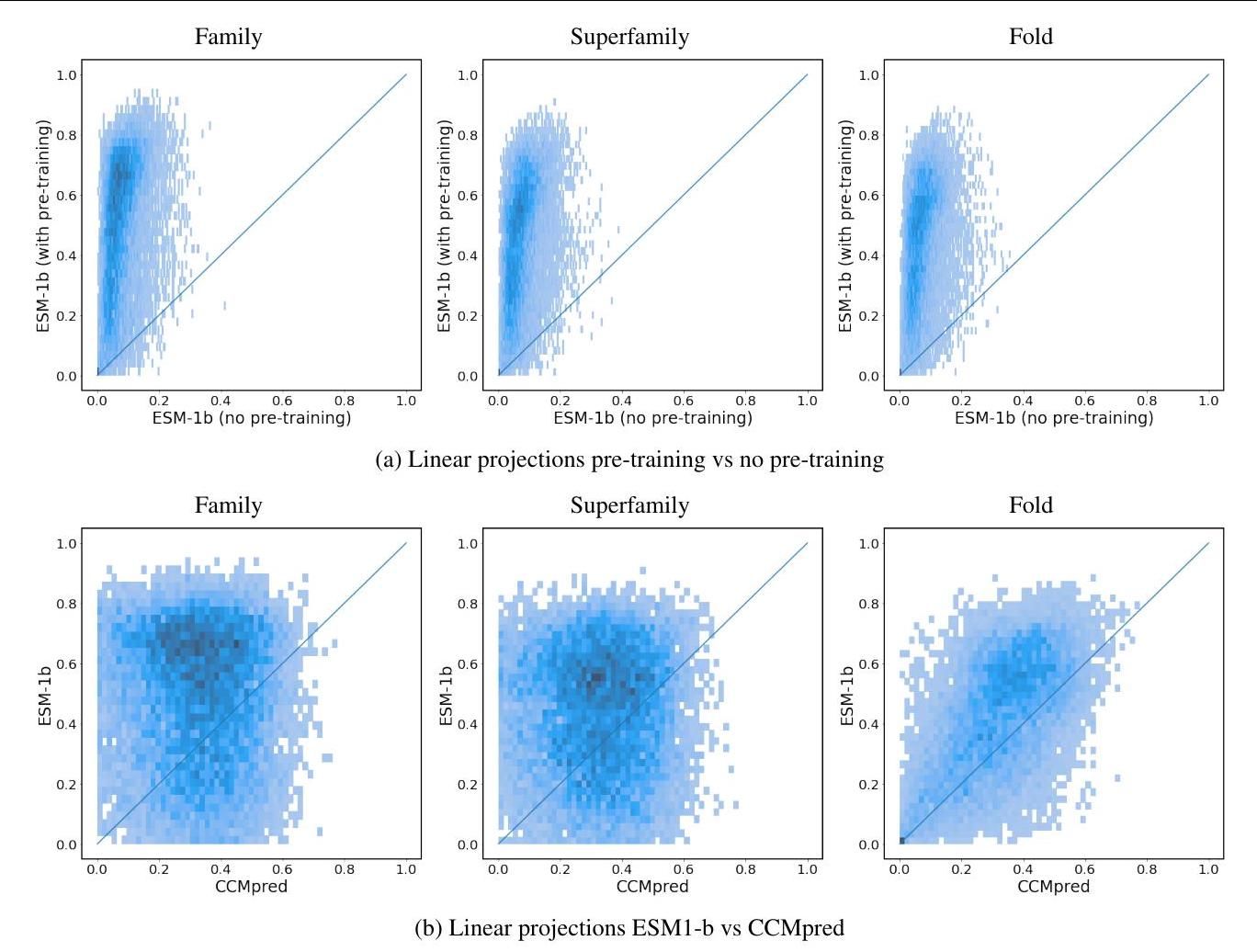


Figure S3. Contact prediction Top-L long-range precision distributions for linear projections. (a) Comparison with and without pretraining; (b) comparison of ESM-1b Transformer representations with CCMpred predictions. Density is indicated by color.

图S3. 线性投影的Top-L长程接触预测精度分布。(a) 有预训练和无预训练的比较；(b) ESM-1b Transformer表示与CCMpred预测的比较。密度由颜色表示。

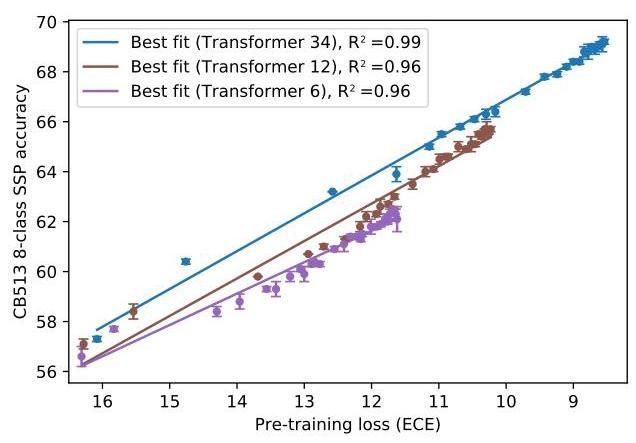


Figure S4. Eight-class secondary structure prediction accuracy as a function of pre-training ECE. A deep secondary structure predictor is trained using features from Transformer models over the course of pre-training on UR50/S. The Netsurf training sequences and CB513 test set are used. Averages across three seeds of the downstream model per pre-training checkpoint are plotted, with line of best fit for each Transformer.

图S4. 八类二级结构预测准确率作为预训练ECE的函数。在UR50/S上预训练过程中，使用Transformer模型的特征训练了一个深度二级结构预测器。使用了Netsurf训练序列和CB513测试集。绘制了每个预训练检查点的下游模型三个种子的平均值，并为每个Transformer绘制了最佳拟合线。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可协议提供。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Metric: top- | | L/5, LR | L, LR | L/5, MR | L, MR | L/5, SR | L, SR |
| Test | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |
| CASP11 | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) |  |  |  |  |  |  |
| (d) +cov |  |  | 76.6 ± .4 |  |  |  |
| CASP12-AVG | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov | 76.6 ± .7 |  |  |  |  |  |
| CASP12-ENS | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |
| CASP13-AVG | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  | 71.2 ± .6 |  |
| (d) +cov |  |  |  |  |  |  |
| CASP13-ENS | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |
| 指标:top- | | L/5, LR | L, LR | L/5, MR | L, MR | L/5, SR | L, SR |
| 测试 | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |
| CASP11 | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) |  |  |  |  |  |  |
| (d) +cov |  |  | 76.6 ± .4 |  |  |  |
| CASP12-AVG | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov | 76.6 ± .7 |  |  |  |  |  |
| CASP12-ENS | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |
| CASP13-AVG | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  | 71.2 ± .6 |  |
| (d) +cov |  |  |  |  |  |  |
| CASP13-ENS | (a) RaptorX |  |  |  |  |  |  |
| (b) +direct |  |  |  |  |  |  |
| (c) +avg |  |  |  |  |  |  |
| (d) +cov |  |  |  |  |  |  |

Table S3. Additional metrics. Feature combination for supervised contact prediction. AVG corresponds to the average of the metrics over the different MSAs, while in ENS the probabilities are averaged (ensembled) over the different MSA predictions before computing the metrics.

表S3. 附加指标。用于监督接触预测的特征组合。AVG对应于不同MSA(多序列比对)上指标的平均值，而在ENS中，概率在不同MSA预测上进行平均(集成)，然后计算指标。

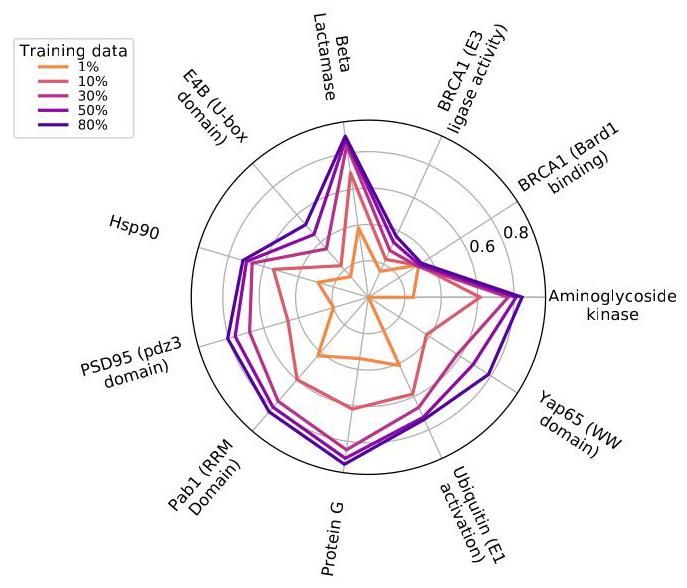


Figure S5. After pre-training, the Transformer can be adapted to predict mutational effects on protein function. The 34-layer Transformer model pre-trained on UR50/S is fine-tuned on mutagenesis data. Spearman on each protein when supervised with smaller fractions of the data.

图S5. 预训练后，Transformer(变压器模型)可以适应预测突变对蛋白质功能的影响。在UR50/S上预训练的34层Transformer模型在诱变数据上进行微调。当使用较少数据监督时，每个蛋白质的Spearman 。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予bioRxiv永久展示该预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可提供。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Amount of training data Protein | 1% data | 10% data | 30% data | 50% data | 80% data |
| Aminoglycoside kinase |  |  |  |  |  |
| BRCA1 (Bard1 binding) |  |  |  |  |  |
| BRCA1 (E3 ligase activity) |  |  |  |  |  |
| Beta Lactamase |  |  |  |  |  |
| E4B (U-box domain) |  |  |  |  |  |
| Hsp90 |  |  |  | 0.70 ± 0.02 |  |
| PSD95 (pdz3 domain) |  |  |  | 0.76 ± 0.01 |  |
| Pab1 (RRM Domain) |  |  |  |  |  |
| Protein G |  |  |  |  |  |
| Ubiquitin (E1 activation) |  |  |  |  |  |
| Yap65 (WW domain) |  |  |  |  |  |
| 训练数据量 蛋白质 | 1% 数据 | 10% 数据 | 30% 数据 | 50% 数据 | 80% 数据 |
| 氨基糖苷激酶 |  |  |  |  |  |
| BRCA1(Bard1 结合) |  |  |  |  |  |
| BRCA1(E3 连接酶活性) |  |  |  |  |  |
| β-内酰胺酶 |  |  |  |  |  |
| E4B(U-box 结构域) |  |  |  |  |  |
| Hsp90 |  |  |  | 0.70 ± 0.02 |  |
| PSD95(pdz3 结构域) |  |  |  | 0.76 ± 0.01 |  |
| Pab1(RRM 结构域) |  |  |  |  |  |
| 蛋白质 G |  |  |  |  |  |
| 泛素(E1 激活) |  |  |  |  |  |
| Yap65(WW 结构域) |  |  |  |  |  |

Table S4. Aggregate spearman measured across models and datasets. Mean and standard deviations of spearman performance for the fine-tuned Transformer-34 on intraprotein tasks. Performance was assessed on five random partitions of the validation set. Model pre-trained on UR50/S.

表S4. 跨模型和数据集的聚合斯皮尔曼 测量。微调后的Transformer-34在蛋白质内任务上的斯皮尔曼 性能的均值和标准差。性能在验证集的五个随机分区上进行了评估。模型在UR50/S上进行了预训练。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)为作者/资助者，其已授予bioRxiv永久展示该预印本的许可。本预印本根据aCC-BY-NC-ND 4.0国际许可提供。bioRxiv预印本 doi: https://doi.org/10.1101/622803; 此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)为作者/资助者，其已授予bioRxiv永久展示该预印本的许可。本预印本根据aCC-BY-NC-ND 4.0国际许可提供。

|  |  |  |  |
| --- | --- | --- | --- |
| dataset model | spearmanr Envision | Envision (LOPO) | DeepSequence |
| Transformer |  | 0.51 |  |
| LSTM biLM (Large) |  |  |  |
| Gray, et al. 2018 |  | 0.45 |  |
| Riesselman, et al. 2018 |  |  |  |
| 数据集模型 | 斯皮尔曼相关系数 Envision | Envision (LOPO) | DeepSequence |
| Transformer |  | 0.51 |  |
| LSTM 双向语言模型(大型) |  |  |  |
| Gray 等人 2018 |  | 0.45 |  |
| Riesselman 等人 2018 |  |  |  |

Table S5. Aggregate spearman measure across models and datasets. 34-layer Transformer pre-trained on UR50/S. For intraprotein models, the train/valid data was randomly partitioned five times. The mean standard deviation across the five runs is reported. No standard deviations are reported for LOPO experiments, as the evaluation is performed across all proteins.

表 S5. 跨模型和数据集的聚合斯皮尔曼(Spearman) 度量。在 UR50/S 上预训练的 34 层 Transformer。对于蛋白质内模型，训练/验证数据被随机划分五次。报告了五次运行的平均值 和标准差。LOPO 实验未报告标准差，因为评估是在所有蛋白质上进行的。

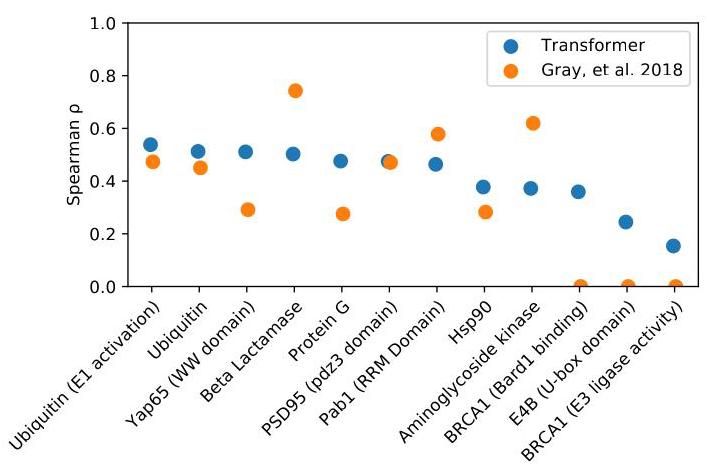


Figure S6. Leave-one-out experiment on Envision dataset (Gray et al., 2018). Pre-training improves the ability of the Transformer to generalize to the mutational fitness landscape of held-out proteins. All mutagenesis data from the protein selected for evaluation are held out, and the model is supervised with data from the remaining proteins. For each evaluation protein, a comparison is shown for the 34-layer Transformer pre-trained on UR50/S.

图 S6. 在 Envision 数据集(Gray 等，2018)上进行的留一法实验。预训练提高了 Transformer 对保留蛋白质的突变适应度景观的泛化能力。评估所选蛋白质的所有突变数据被保留，模型使用其余蛋白质的数据进行监督。对于每个评估蛋白质，展示了在 UR50/S 上预训练的 34 层 Transformer 的比较结果。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  |  | SSP | | | Contact | | |
| Training | Params | CB513 | CASP13 | Test | CASP11 | CASP12 | CASP13 |
| Transformer-34 | (None) | 669.2M | 56.8 | 60.0 | 16.3 | 17.7 | 14.8 | 13.3 |
| UniRep (LSTM) |  | 18.2M | 58.4 | 60.1 | 21.9 | 21.4 | 16.8 | 14.3 |
| SeqVec (LSTM) |  | 93M | 62.1 | 64.0 | 29.0 | 25.5 | 23.6 | 17.9 |
| TAPE (Transformer) |  | 38M | 58.0 | 61.5 | 23.2 | 23.8 | 20.3 | 16.0 |
| LSTM (S) | UR50/S | 28.4M | 60.4 | 63.2 | 24.1 | 23.6 | 19.9 | 15.3 |
| LSTM (L) | UR50/S | 113.4M | 62.4 | 64.1 | 27.8 | 26.4 | 24.0 | 16.4 |
| Transformer-6 | UR50/S | 42.6M | 62.0 | 64.2 | 30.2 | 29.9 | 25.3 | 19.8 |
| Transformer-12 | UR50/S | 85.1M | 65.4 | 67.2 | 37.7 | 33.6 | 27.8 | 20.7 |
| Transformer-34 | UR100 | 669.2M | 64.3 | 66.5 | 32.7 | 28.9 | 24.3 | 19.1 |
| Transformer-34 | UR50/S | 669.2M | 69.1 | 70.7 | 50.2 | 42.8 | 34.7 | 30.1 |
| ESM-1b | UR50/S | 652.4M | 71.6 | 72.5 | 56.9 | 47.4 | 42.7 | 35.9 |
| 模型 |  |  | SSP | | | 联系 | | |
| 训练 | 参数 | CB513 | CASP13 | 测试 | CASP11 | CASP12 | CASP13 |
| Transformer-34 | (无) | 669.2M | 56.8 | 60.0 | 16.3 | 17.7 | 14.8 | 13.3 |
| UniRep(LSTM) |  | 18.2M | 58.4 | 60.1 | 21.9 | 21.4 | 16.8 | 14.3 |
| SeqVec(LSTM) |  | 93M | 62.1 | 64.0 | 29.0 | 25.5 | 23.6 | 17.9 |
| TAPE(Transformer) |  | 38M | 58.0 | 61.5 | 23.2 | 23.8 | 20.3 | 16.0 |
| LSTM(S) | UR50/S | 28.4M | 60.4 | 63.2 | 24.1 | 23.6 | 19.9 | 15.3 |
| LSTM(L) | UR50/S | 113.4M | 62.4 | 64.1 | 27.8 | 26.4 | 24.0 | 16.4 |
| Transformer-6 | UR50/S | 42.6M | 62.0 | 64.2 | 30.2 | 29.9 | 25.3 | 19.8 |
| Transformer-12 | UR50/S | 85.1M | 65.4 | 67.2 | 37.7 | 33.6 | 27.8 | 20.7 |
| Transformer-34 | UR100 | 669.2M | 64.3 | 66.5 | 32.7 | 28.9 | 24.3 | 19.1 |
| Transformer-34 | UR50/S | 669.2M | 69.1 | 70.7 | 50.2 | 42.8 | 34.7 | 30.1 |
| ESM-1b | UR50/S | 652.4M | 71.6 | 72.5 | 56.9 | 47.4 | 42.7 | 35.9 |

Table S6. Comparison to related methods. Top-L long-range contact precision. Predictions are directly from protein sequence, no coevolutionary features or MSAs used. Test is RaptorX test set of Wang et al. (2017). Model weights for related work are obtained and evaluated in our codebase with same downstream architecture, training, and test data.

表S6。与相关方法的比较。Top-L长程接触精度。预测直接来自蛋白质序列，未使用共进化特征或多序列比对(MSA)。测试使用的是Wang等人(2017)的RaptorX测试集。相关工作的模型权重在我们的代码库中获取，并使用相同的下游架构、训练和测试数据进行评估。

under aCC-BY-NC-ND 4.0 International license.

根据CC-BY-NC-ND 4.0国际许可协议。

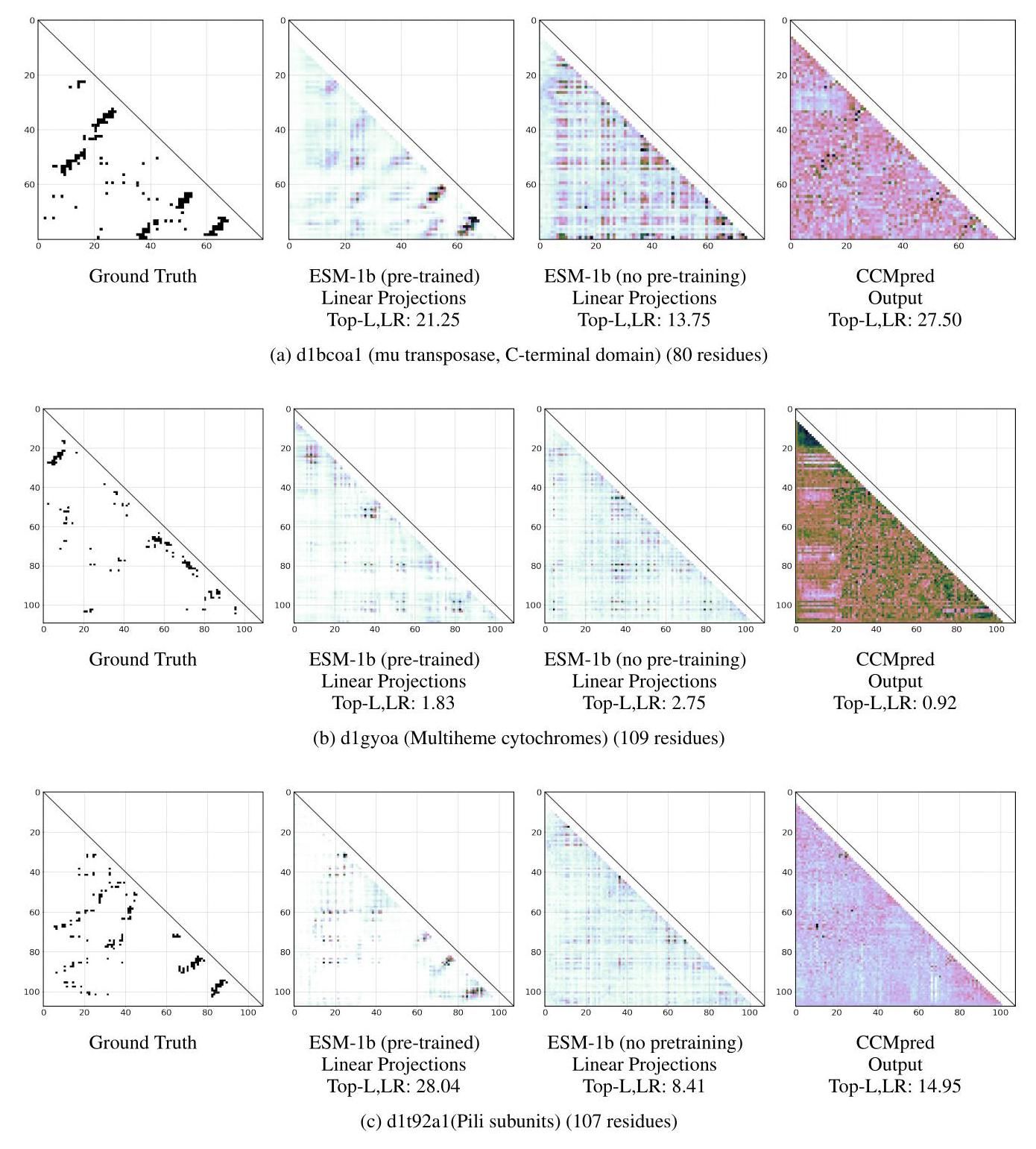
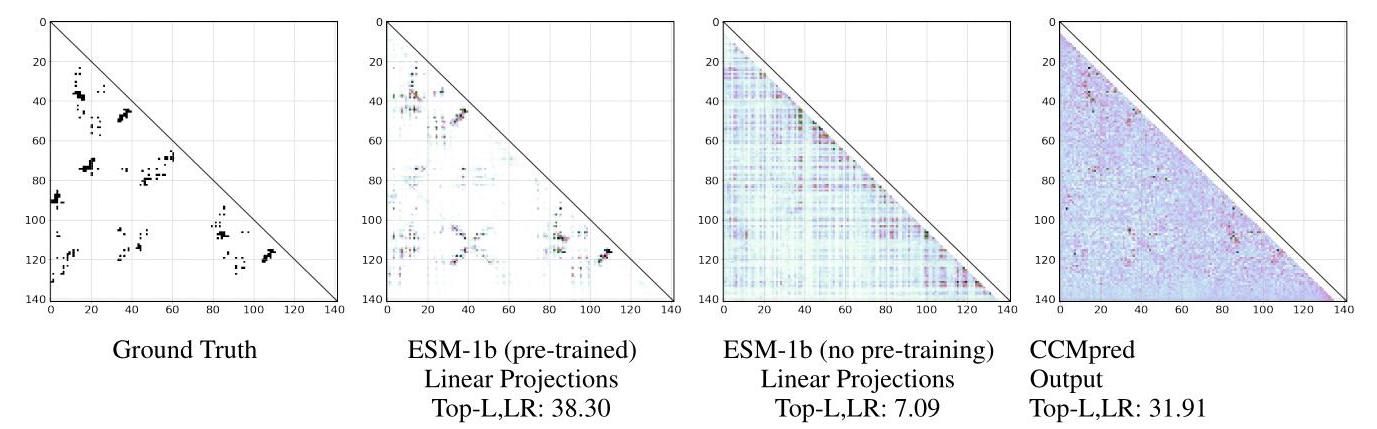


Figure S7. Comparison of ground truth contact map, projections from ESM-1b with and without pre-training, and CCMpred output. Labels indicate SCOPe domain, fold name, and number of residues. Eight domains randomly sampled from fold-level test sets are shown.

图S7。真实接触图、ESM-1b在有和没有预训练情况下的投影以及CCMpred输出的比较。标签显示了SCOPe结构域、折叠名称和残基数量。展示了从折叠级别测试集中随机采样的八个结构域。

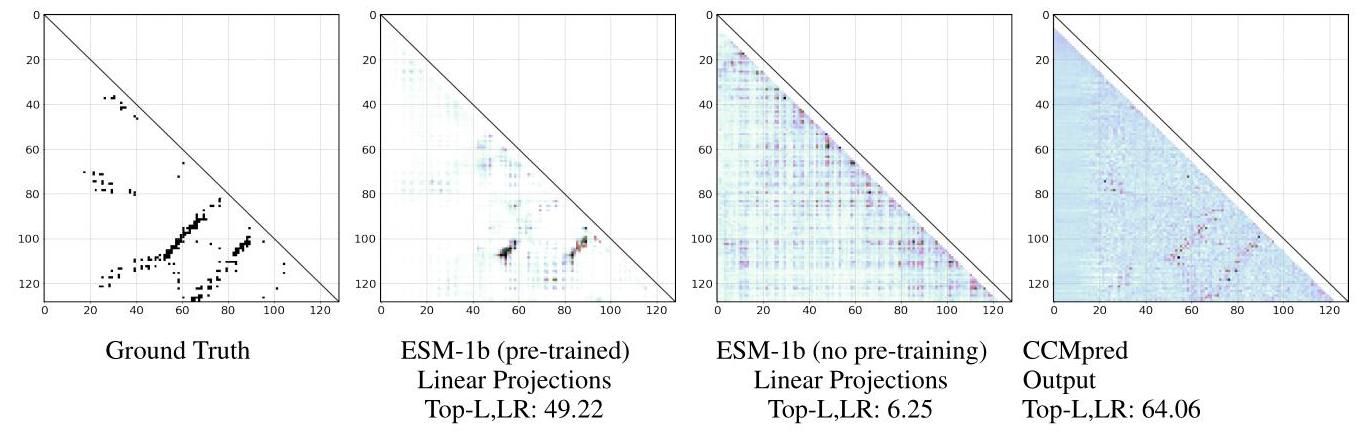
bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license. under aCC-BY-NC-ND 4.0 International license.

bioRxiv预印本doi: https://doi.org/10.1101/622803；此版本发布于2020年12月15日。该预印本的版权持有者(未经同行评审认证)是作者/资助者，他们已授予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。根据CC-BY-NC-ND 4.0国际许可协议。



(d) d3ddja1 (CBS-domain pair) (141 residues)

(d) d3ddja1(CBS结构域对)(141个残基)



(e) d3paja1 (alpha/beta-Hammerhead) (128 residues)

(e) d3paja1(α/β-锤头结构域)(128个残基)

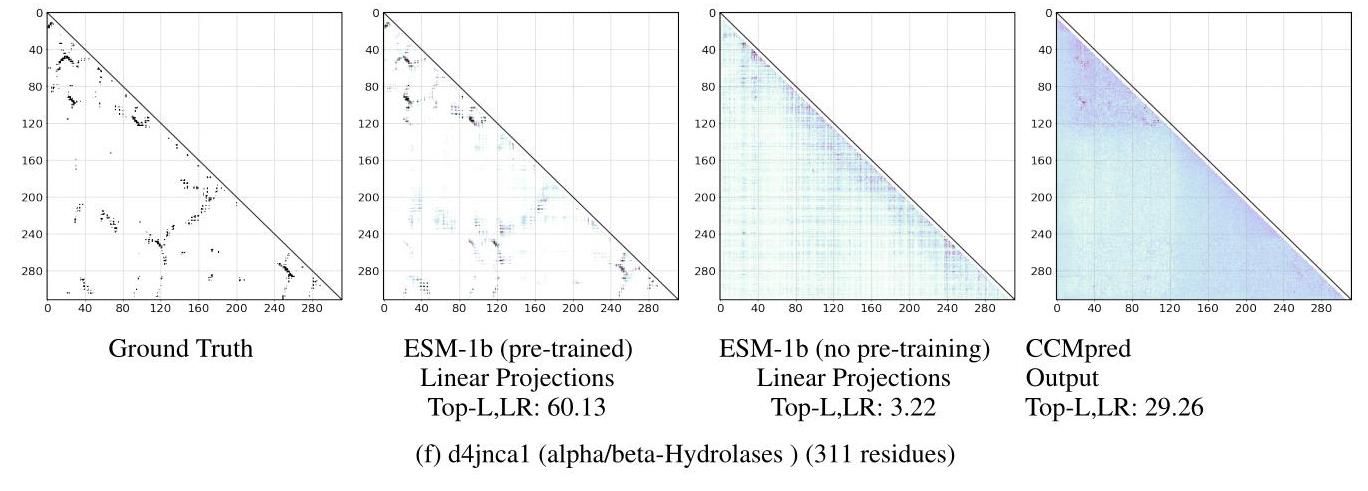
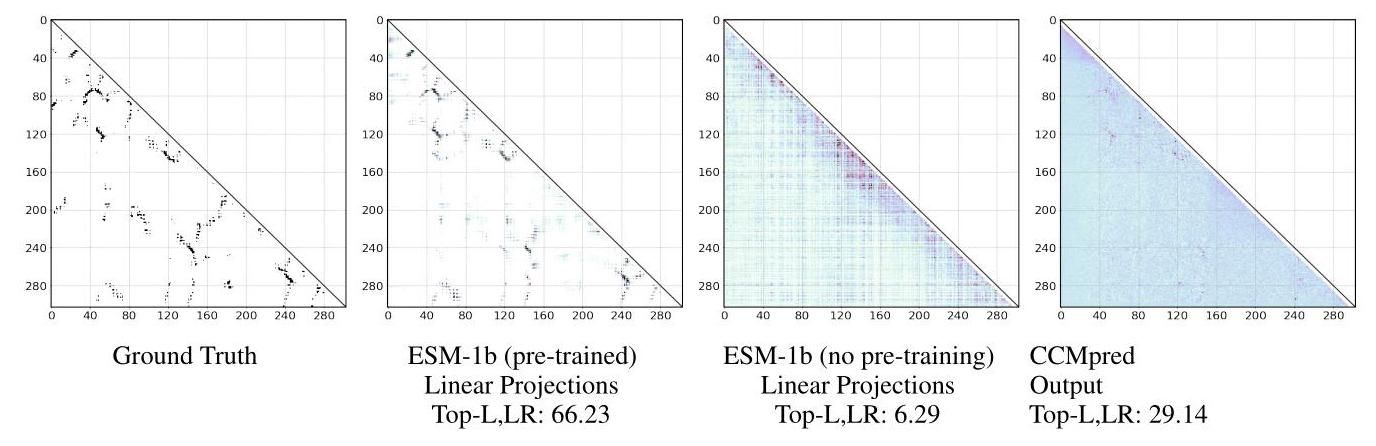


Figure S7. (continued from above)

图S7。(接上文)



(g) d5esra1 (alpha/beta-Hydrolases) (302 residues)

(g) d5esra1(α/β-水解酶)(302个残基)

100

Ground Truth ESM-1b (pre-trained)

真实接触图 ESM-1b(预训练)

Linear Projections Linear Projections

线性投影 线性投影

Top-L, LR: 46.43 Top-L, LR: 4.46 Top-L, LR: 27.68

Top-L, LR: 46.43 Top-L, LR: 4.46 Top-L, LR: 27.68

(h) d3b64a (Tautomerase/MIF) (112 residues)

(h) d3b64a(互变异构酶/MIF)(112个残基)

Figure S7. (continued from above)

图S7。(接上文)

Supplemental References

补充参考文献

Adhikari, B. DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout. Bioinformatics, 36(2):470-477, 07 2019.

Adhikari, B. DEEPCON:使用带有dropout的扩张卷积神经网络进行蛋白质接触预测。生物信息学，36(2):470-477，2019年7月。

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

Ba, J. L., Kiros, J. R., 和 Hinton, G. E. 层归一化。arXiv预印本arXiv:1607.06450，2016年。

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics, 79(4):1061-1078, 2011.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., 和 Langmead, C. J. 学习蛋白质折叠家族的生成模型。蛋白质:结构、功能和生物信息学，79(4):1061-1078, 2011.

Bateman, A., Heger, A., Sonnhammer, E. L. L., Mistry, J., Clements, J., Tate, J., Hetherington, K., Holm, L., Punta, M., Coggill, P., Eberhardt, R. Y., Eddy, S. R., and Finn, R. D. Pfam: the protein families database. Nucleic Acids Research, 42(D1):D222-D230, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1223.

Bateman, A., Heger, A., Sonnhammer, E. L. L., Mistry, J., Clements, J., Tate, J., Hetherington, K., Holm, L., Punta, M., Coggill, P., Eberhardt, R. Y., Eddy, S. R., 和 Finn, R. D. Pfam:蛋白质家族数据库。核酸研究，42(D1):D222-D230, 2013年11月。ISSN 0305-1048. doi: 10.1093/nar/gkt1223.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. URL https://openai.com/blog/sparse-transformers, 2019.

Child, R., Gray, S., Radford, A., 和 Sutskever, I. 使用稀疏变压器生成长序列。URL https://openai.com/blog/sparse-transformers, 2019.

Cuff, J. A. and Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioin-formatics, 34(4):508-519, 1999.

Cuff, J. A. 和 Barton, G. J. 评估和改进蛋白质二级结构预测的多序列方法。蛋白质:结构、功能和生物信息学，34(4):508-519, 1999.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018a. URL http://arxiv.org/abs/1810.04805.

Devlin, J., Chang, M., Lee, K., 和 Toutanova, K. BERT:用于语言理解的深度双向变压器预训练。CoRR, abs/1810.04805, 2018a. URL http://arxiv.org/abs/1810.04805.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018b. URL http://arxiv.org/abs/1810.04805.

Devlin, J., Chang, M., Lee, K., 和 Toutanova, K. BERT:用于语言理解的深度双向变压器预训练。CoRR, abs/1810.04805, 2018b. URL http://arxiv.org/abs/1810.04805.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. Improved contact prediction in proteins: using pseu-dolikelihoods to infer potts models. Physical Review E, 87(1):012707, 2013.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., 和 Aurell, E. 改进蛋白质中的接触预测:使用伪似然推断Potts模型。物理评论E，87(1):012707, 2013.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research, 42(D1): D304-D309, January 2014. ISSN 0305-1048, 1362- 4962. doi: 10.1093/nar/gkt1240. URL https:// academic.oup.com/nar/article-lookup/ doi/10.1093/nar/gkt1240.

Fox, N. K., Brenner, S. E., 和 Chandonia, J.-M. SCOPe:蛋白质结构分类扩展，整合SCOP和ASTRAL数据并分类新结构。核酸研究，42(D1): D304-D309, 2014年1月。ISSN 0305-1048, 1362- 4962. doi: 10.1093/nar/gkt1240. URL https:// academic.oup.com/nar/article-lookup/ doi/10.1093/nar/gkt1240.

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell systems, 6(1):116-124, 2018.

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., 和 Fowler, D. M. 使用大规模诱变数据定量预测错义变异效应。细胞系统，6(1):116-124, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings

He, K., Zhang, X., Ren, S., 和 Sun, J. 深度残差网络中的恒等映射

in deep residual networks. In European conference on computer vision, pp. 630-645. Springer, 2016.

在欧洲计算机视觉会议上，第630-645页。Springer, 2016.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

Hendrycks, D. 和 Gimpel, K. 高斯误差线性单元(GELUs)。arXiv预印本 arXiv:1606.08415, 2016.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation, 9(8):1735-1780, 1997. doi: 10. 1162/neco.1997.9.8.1735.

Hochreiter, S. 和 Schmidhuber, J. 长短期记忆。神经计算，9(8):1735-1780, 1997. doi: 10. 1162/neco.1997.9.8.1735.

Hochreiter, S., Bengio, Y., and Frasconi, P. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 2001.

Hochreiter, S., Bengio, Y., 和 Frasconi, P. 循环网络中的梯度流:学习长期依赖性的困难。2001.

Huerta-Cepas, J., Forslund, S. K., Bork, P., Hernández-Plaza, A., von Mering, C., Szklarczyk, D., Heller, D., Cook, H., Jensen, L., Mende, D. R., Letunic, I., and Rattei, T. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47(D1):D309-D314, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1085.

Huerta-Cepas, J., Forslund, S. K., Bork, P., Hernández-Plaza, A., von Mering, C., Szklarczyk, D., Heller, D., Cook, H., Jensen, L., Mende, D. R., Letunic, I., 和 Rattei, T. eggNOG 5.0:基于5090种生物和2502种病毒的分层、功能和系统发育注释的直系同源资源。核酸研究，47(D1):D309-D314, 2018年11月。ISSN 0305-1048. doi: 10.1093/nar/gky1085.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. CoRR, abs/1702.08734, 2017. URL http://arxiv.org/abs/1702.08734.

Johnson, J., Douze, M., 和 Jégou, H. 使用GPU进行十亿级相似性搜索。CoRR, abs/1702.08734, 2017. URL http://arxiv.org/abs/1702.08734.

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., and Vriend, G. A series of pdb related databases for everyday needs. Nucleic acids research, 39(suppl\_1):D411-D419, 2010.

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., 和 Vriend, G. 一系列用于日常需求的PDB相关数据库。核酸研究, 39(suppl\_1):D411-D419, 2010.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12):2577-2637, 1983.

Kabsch, W. 和 Sander, C. 蛋白质二级结构词典:氢键和几何特征的模式识别。生物聚合物, 22(12):2577-2637, 1983.

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. Sharp nearby, fuzzy far away: How neural language models use context. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 284-294, 2018.

Khandelwal, U., He, H., Qi, P., 和 Jurafsky, D. 近处清晰，远处模糊:神经语言模型如何使用上下文。在第56届计算语言学协会年会论文集(第1卷:长论文)中，第284-294页，2018年。

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sonderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., and Marcatili, P. NetSurfP- 2.0: Improved prediction of protein structural features by integrated deep learning. Proteins, 87(6):520-527, 06 2019.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sonderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., 和 Marcatili, P. NetSurfP-2.0:通过集成深度学习改进蛋白质结构特征预测。蛋白质, 87(6):520-527, 2019年6月。

Lapedes, A. S., Giraud, B. G., Liu, L., and Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. Lecture Notes-Monograph Series, pp. 236-256, 1999.

Lapedes, A. S., Giraud, B. G., Liu, L., 和 Stormo, G. D. 蛋白质序列模型中的相关突变:系统发育和结构效应。讲义-专论系列，第236-256页，1999年。

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., 和 Stoyanov, V.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. 7 2019. URL http://arxiv.org/abs/ 1907.11692.

RoBERTa:一种稳健优化的BERT预训练方法。2019年7月。URL http://arxiv.org/abs/1907.11692.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101, 2017年。

Ma, J., Wang, S., Wang, Z., and Xu, J. Mrfalign: protein homology detection through alignment of markov random fields. In International Conference on Research in Computational Molecular Biology, pp. 173-174. Springer, 2014.

Ma, J., Wang, S., Wang, Z., 和 Xu, J. Mrfalign:通过马尔可夫随机场对齐进行蛋白质同源性检测。在国际计算分子生物学研究会议中，第173-174页。Springer, 2014年。

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research, 45(D1):D170-D176, 2017.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., 和 Steinegger, M. Uniclust数据库:聚类和深度注释的蛋白质序列和比对。核酸研究, 45(D1):D170-D176, 2017年。

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., 和 Auli, M. fairseq:一个快速、可扩展的序列建模工具包。在NAACL-HLT 2019年会议演示中，2019年。

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 2227- 2237, 2018. URL https://aclanthology.info/ papers/N18-1202/n18-1202.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., 和 Zettlemoyer, L. 深度上下文化词表示。在2018年北美计算语言学协会年会:人类语言技术会议论文集，NAACL-HLT 2018，美国路易斯安那州新奥尔良，2018年6月1-6日，第1卷(长论文)，第2227-2237页，2018年。URL https://aclanthology.info/papers/N18-1202/n18-1202.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019a.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., 和 Sutskever, I. 语言模型是无监督的多任务学习者。2019a.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019b.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., 和 Sutskever, I. 语言模型是无监督的多任务学习者。2019b.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 探索迁移学习的极限:统一的文本到文本转换器，2020。

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hh-blits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nature Methods, 9:173 EP, 12 2011. doi: 10.1038/nmeth.1818.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hh-blits:通过HMM-HMM比对进行闪电般快速的迭代蛋白质序列搜索。《自然方法》，9:173 EP，2011年12月。doi: 10.1038/nmeth.1818。

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10):816-822, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. 深度生成模型捕捉遗传变异中突变的影响。《自然方法》，15(10):816-822，2018。ISSN 1548-7105。doi: 10.1038/s41592-018-0138-4。

Seemayer, S., Gruber, M., and Söding, J. Compred-fast

Seemayer, S., Gruber, M., and Söding, J. Compred-快速

and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics, 30(21):3128- 3130, 2014.

并精确预测蛋白质残基-残基接触，基于相关突变。《生物信息学》，30(21):3128-3130，2014。

Senior, A., Jumper, J., and Hassabis, D. AlphaFold: Using AI for scientific discovery, 12 2018. URL https:// deepmind.com/blog/alphafold/.

Senior, A., Jumper, J., and Hassabis, D. AlphaFold:利用AI进行科学发现，2018年12月。URL https://deepmind.com/blog/alphafold/。

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multibillion parameter language models using model parallelism, 2020.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm:使用模型并行训练数十亿参数的语言模型，2020。

Söding, J. and Remmert, M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Current opinion in structural biology, 21(3): 404-411, 2011.

Söding, J. and Remmert, M. 蛋白质序列比较和折叠识别:进展与良好实践基准。《结构生物学当前观点》，21(3):404-411，2011。

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology, 35(11):1026-1028, 2017.

Steinegger, M. and Söding, J. Mmseqs2实现了敏感蛋白质序列搜索，用于大规模数据集分析。《自然生物技术》，35(11):1026-1028，2017。

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31(6):926-932, 2015.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. Uniref聚类:一种全面且可扩展的替代方案，用于改进序列相似性搜索。《生物信息学》，31(6):926-932，2015。

The UniProt Consortium. The universal protein resource (uniprot). Nucleic Acids Research, 36(suppl\_1):D190- D195, 11 2007. ISSN 0305-1048. doi: 10.1093/nar/ gkm895.

The UniProt Consortium. 通用蛋白质资源(UniProt)。《核酸研究》，36(suppl\_1):D190-D195，2007年11月。ISSN 0305-1048。doi: 10.1093/nar/gkm895。

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. Graphical models of residue coupling in protein families. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5(2):183-197, 2008.

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. 蛋白质家族中残基耦合的图形模型。《IEEE/ACM计算生物学与生物信息学汇刊》，5(2):183-197，2008。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008, 2017a.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 注意力机制就是一切。《神经信息处理系统进展》，pp. 5998-6008，2017a。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008, 2017b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 注意力机制就是一切。《神经信息处理系统进展》，pp. 5998-6008，2017b。

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1): e1005324, 2017.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. 通过超深度学习模型准确预测蛋白质接触图。《PLoS计算生物学》，13(1):e1005324，2017。

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences, 106(1):67-72, 2009.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. 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》，106(1):67-72，2009。

bioRxiv preprint doi: https://doi.org/10.1101/622803; this version posted December 15, 2020. 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/10.1101/622803；此版本发布于2020年12月15日。本预印本的版权持有者(未经同行评审认证)为作者/资助者，其已授予bioRxiv永久展示该预印本的许可。本预印本根据CC-BY-NC-ND 4.0国际许可提供。

Xu, J. Distance-based protein folding powered by deep learning. arXiv preprint arXiv:1811.03481, 2018.

徐，J. 基于距离的蛋白质折叠，由深度学习驱动。arXiv预印本arXiv:1811.03481，2018年。

1. Equal contribution Work performed while at Facebook AI Research Facebook AI Research Dept. of Computer Science, New York University Harvard University Booth School of Business, University of Chicago & Yale Law School. Correspondence to: Alexander Rives <arives@cs.nyu.edu>. Pre-trained models available at: <https://github.com/facebookresearch/esm>.

   同等贡献 在Facebook人工智能研究院(Facebook AI Research)工作期间完成 Facebook人工智能研究院 纽约大学计算机科学系 哈佛大学 芝加哥大学布斯商学院与耶鲁法学院。通讯作者:Alexander Rives <arives@cs.nyu.edu>。预训练模型可在以下网址获取:<https://github.com/facebookresearch/esm>。 [↑](#footnote-ref-1)