# Language models enable zero-shot prediction of the effects of mutations on protein function

语言模型能够零样本预测突变对蛋白质功能的影响

Joshua Meier Roshan Rao Robert Verkuil Jason Liu

约书亚·迈尔 罗尚·拉奥 罗伯特·维尔库伊尔 杰森·刘

Tom Sercu Alexander Rives

汤姆·塞尔库 亚历山大·里夫斯

# Abstract

# 摘要

Modeling the effect of sequence variation on function is a fundamental problem for understanding and designing proteins. Since evolution encodes information about function into patterns in protein sequences, unsupervised models of variant effects can be learned from sequence data. The approach to date has been to fit a model to a family of related sequences. The conventional setting is limited, since a new model must be trained for each prediction task. We show that using only zero-shot inference, without any supervision from experimental data or additional training, protein language models capture the functional effects of sequence variation, performing at state-of-the-art.

建模序列变异对功能的影响是理解和设计蛋白质的基本问题。由于进化将功能信息编码到蛋白质序列的模式中，因此可以从序列数据中学习变异效应的无监督模型。迄今为止的方法是将模型拟合到一组相关序列上。传统设置存在局限性，因为每个预测任务都需要训练一个新模型。我们展示了仅使用**零样本推理**，无需实验数据或额外训练的监督，蛋白质语言模型就能捕捉序列变异的功能效应，达到最先进的水平。

# 1 Introduction

# 1 引言

Proteins have a myriad of diverse functions that underlie the complexity of life. Protein sequences encode function via structure through the spontaneous folding of the sequence into the three dimensional structure of the protein [1]. The effects of sequence mutations on function form a landscape that reveals how function constrains sequence. Alterations at some sites in a protein sequence cannot be tolerated because they are essential to the protein’s function. Other sites evolve together because the structure and function is determined by them collectively. Mutations can enhance the activity of a protein, attenuate it, or leave it unchanged.

蛋白质具有多种多样的功能，这些功能构成了生命复杂性的基础。蛋白质序列通过自发折叠成蛋白质的三维结构来编码功能[1]。序列突变对功能的影响形成了一个揭示功能如何约束序列的景观。蛋白质序列中的某些位点的改变是不可容忍的，因为它们对蛋白质的功能至关重要。其他位点共同进化，因为结构和功能由它们共同决定。突变可以增强蛋白质的活性，减弱其活性，或使其保持不变。

The functional effect of sequence variations can be measured through deep mutational scanning experiments [2]. Consisting of thousands to hundreds of thousands of measurements of protein function, deep mutational scans give insight into the intrinsic constraints on a protein’s structure and function. Due to the cost and difficulty of implementing such experiments, compilations of deep mutational scanning data include experiments on a few dozens of proteins at most, relative to the tens of thousands of proteins encoded in the human genome, and the millions more across the tree of life that we would like to understand.

序列变异的功能效应可以通过深度突变扫描实验来测量[2]。深度突变扫描由数千到数十万个蛋白质功能测量组成，揭示了蛋白质结构和功能的内在约束。由于实施此类实验的成本和难度，深度突变扫描数据的汇编最多包括对几十种蛋白质的实验，相对于人类基因组中编码的数万种蛋白质，以及我们希望在生命树中理解的数百万种蛋白质。

A model that learns the landscape linking sequence to function can provide insight into function without having to do experiments. Unsupervised models of mutational effects can be learned from sequences . Statistical patterns in a family of evolutionarily related protein sequences contain information about structure and function [5-7]. This is because the properties of a protein act as constraints on the selection of sequences through evolution [8].

一个学习将序列与功能联系起来的景观的模型可以在不进行实验的情况下提供对功能的洞察。变异效应的无监督模型可以从序列中学习 。进化相关蛋白质序列家族中的统计模式包含有关结构和功能的信息[5-7]。这是因为蛋白质的特性通过进化对序列的选择起到了约束作用[8]。

In the natural language modeling community, there has been interest in zero-shot transfer of models to new tasks. Massive language models can solve tasks they haven’t been directly trained on [9-11]. Recently protein language models have achieved state-of-the-art in various structure prediction tasks [12-14]. Work to date has mainly focused on transfer in the classical representation learning setting, using pre-trained features with supervision on the downstream task.

在自然语言建模社区中，人们对模型零样本迁移到新任务产生了兴趣。大规模语言模型可以解决它们没有直接训练过的任务[9-11]。最近，蛋白质语言模型在各种结构预测任务中达到了最先进的水平[12-14]。迄今为止的工作主要集中在经典表示学习设置中的迁移，使用预训练特征并在下游任务上进行监督。

[[1]](#footnote-1)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.450648; this version posted November 17, 2021. 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 In this work we show that language models trained on large and diverse protein sequence databases can predict experimental measurements of protein function without further supervision. Prior work has focused on transferring the representations using supervision from experimental data . We find that language models can transfer to predict functional measurements without supervision. Language models perform zero-shot and few-shot prediction of mutational effects across a variety of proteins with widely differing functions. We perform experiments with state-of-the-art protein language models ESM-1b [12] and MSA Transformer [13]. We introduce a new protein language model, ESM-1v, with zero-shot performance comparable to state-of-the-art mutational effect predictors. Performance can be further improved by fine-tuning the model with sequences from the protein family. Predictions capture the functional landscape of the protein, correlate with amino acid conservation patterns in the core and surface, and identify residues responsible for binding and activity.

在这项工作中，我们展示了在大型多样化蛋白质序列数据库上训练的语言模型可以在没有进一步监督的情况下预测蛋白质功能的实验测量结果。先前的工作主要集中在使用实验数据的监督来迁移表示 。我们发现语言模型可以在没有监督的情况下迁移以预测功能测量结果。语言模型在各种功能差异很大的蛋白质中执行零样本和少样本的突变效应预测。我们使用最先进的蛋白质语言模型 ESM-1b [12] 和 MSA Transformer [13] 进行了实验。我们引入了一种新的蛋白质语言模型 ESM-1v，其零样本性能可与最先进的突变效应预测器相媲美。通过使用蛋白质家族的序列对模型进行微调，可以进一步提高性能。**预测捕捉了蛋白质的功能景观，与核心和表面的氨基酸保守模式相关，并识别出负责结合和活性的残基。**

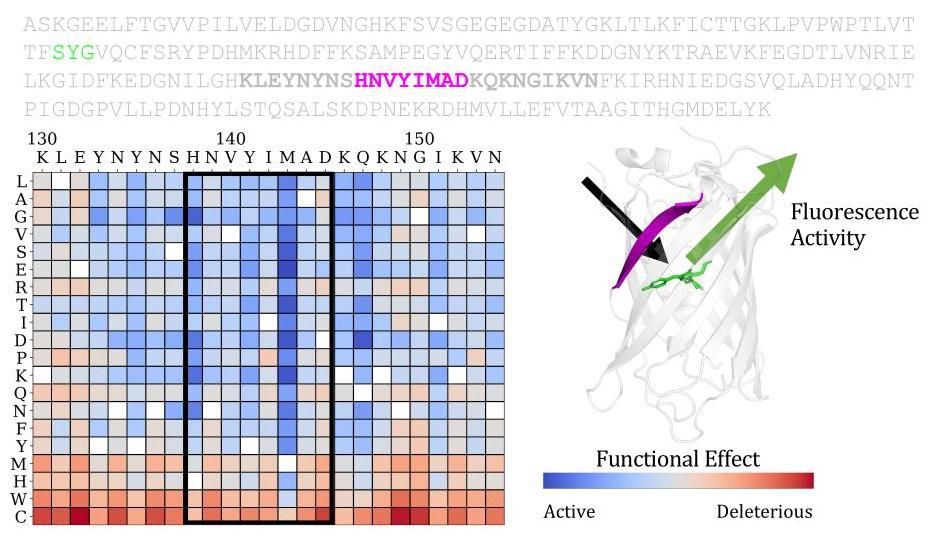


Figure 1: Depiction of a mutational effect prediction task. The objective is to score the effect of sequence mutations on the function of a protein. Deep mutational scanning experiments provide ground truth experimental measurements of the protein’s function (fluorescence activity in the example here) for a large set of single mutations or combinations of mutations. For each protein, the prediction task is to score each possible mutation and rank its relative activity. Predictions for single substitutions can be described in a score matrix. The columns are the positions in the sequence. The rows are the possible variations at each position.

图1:突变效应预测任务的示意图。目标是评估序列突变对蛋白质功能的影响。深度突变扫描实验提供了大量单突变或突变组合的蛋白质功能(此处示例为荧光活性)的真实实验测量值。对于每个蛋白质，预测任务是对每个可能的突变进行评分并排序其相对活性。单替换的预测可以用评分矩阵来描述。列表示序列中的位置，行表示每个位置可能的变异。

# 2 Zero-shot transfer

# 2 零样本迁移

Zero-shot learning has classically described the extension of a classifier to a new set of classes that have not been seen in training [17]. In natural language processing this idea has been extended to describe the transfer of models to entirely new tasks without further training. Proposed as zero-data learning by Larochelle et al. [18], this perspective on transfer has been at the center of recent work understanding the generalization capabilities of large language models [9-11, 19]. The distinction from representation learning is that the models are used directly without additional supervision for the task. This means that the tasks must be learned purely from pre-training.

零样本学习传统上描述的是将分类器扩展到训练中未见的新类别集合[17]。在自然语言处理中，这一概念被扩展为描述模型在无需进一步训练的情况下迁移到全新任务的能力。Larochelle等人[18]提出的零数据学习观点，已成为近期理解大型语言模型泛化能力工作的核心[9-11, 19]。与表示学习的区别在于，模型直接用于任务而无需额外监督。这意味着任务必须完全通过预训练来学习。

In this work we take a similar perspective on zero-shot transfer to that of GPT-3, described in Brown et al. [10]. We define zero-shot transfer to be transfer of a model to a new task without any further supervision to specialize the model to the task. We also consider the closely related idea of few-shot transfer. Here as in Brown et al. [10] we define the few-shot setting to be one in which a few positive examples are given to the model as inputs at inference time. As in the zero-shot setting, no gradient updates are performed to specialize the model. Similar to Brown et al. [10], the claim is not one of out-of-distribution generalization. The assumption is that in the pre-training stage, the model learns information relevant to the tasks to which it will later be transferred. In the case of protein language models, the pre-training dataset includes sequences from across evolution, which implies the model may see examples of sequences from protein families on which it will be evaluated. The essential

在本工作中，我们采用了与GPT-3类似的零样本迁移视角，如Brown等人[10]所述。我们将零样本迁移定义为模型在没有任何进一步监督的情况下迁移到新任务。我们还考虑了与之密切相关的少样本迁移概念。如Brown等人[10]所述，我们将少样本设置定义为在推理时向模型提供少量正例作为输入。与零样本设置一样，不进行梯度更新来专门化模型。与Brown等人[10]类似，这里的声明并非关于分布外泛化。假设在预训练阶段，模型学习了与后续迁移任务相关的信息。对于蛋白质语言模型，预训练数据集包含来自整个进化过程的序列，这意味着模型可能会看到来自其评估蛋白质家族的序列示例。

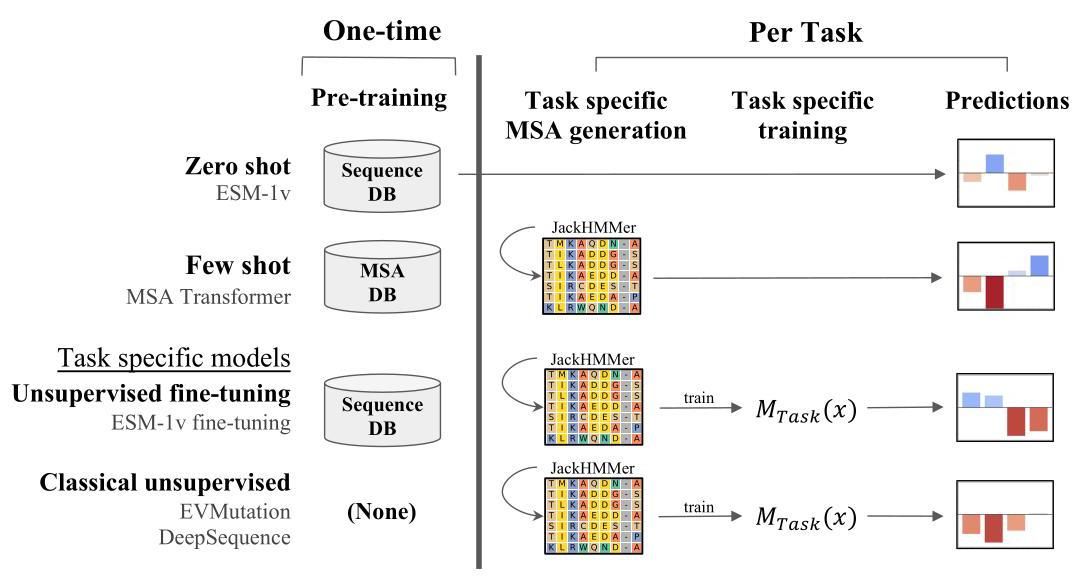


Figure 2: Steps involved in variant effect prediction methods. Compared with EVMutation [4] and DeepSequence [20], MSA Transformer and ESM-1v require no task-specific model training for inference. Moreover, ESM-1v does not require MSA generation.

图2:变异效应预测方法的步骤。与EVMutation[4]和DeepSequence[20]相比，MSA Transformer和ESM-1v在推理时无需任务特定的模型训练。此外，ESM-1v不需要MSA生成。

Measurements of function, a property of central importance to the understanding and design of proteins, are a practical ground for studying the generalization capability of protein language models. Deep mutational scanning experiments measure the effects of thousands to hundreds of thousands of mutations on a single protein, and have been performed on a variety of proteins having different functions and using various forms of experimental measurement. We study zero-shot and few-shot transfer of protein language models to function prediction using this data.

功能测量是理解和设计蛋白质的核心属性，为研究蛋白质语言模型的泛化能力提供了实际基础。深度突变扫描实验测量了数千到数十万突变对单个蛋白质的影响，并已在具有不同功能和采用各种实验测量形式的多种蛋白质上进行了实验。我们利用这些数据研究了蛋白质语言模型在功能预测中的零样本和少样本迁移。

Supervised methods trained with data from experimental measurements , and unsupervised methods trained only on sequences have been developed for prediction of mutational effects. Unsupervised mutational effect predictors are trained as task specific models on sequences from an individual protein family. In this view every protein is an independent prediction task where the objective is to score the effect of mutations on the protein’s function. While mutational effect predictors trained on multiple sequence alignments (MSAs) are typically described as unsupervised, they can also be seen as weakly supervised. Hsu et al. [15] observe that such models have weak supervision on the task through the MSA, which describes the fitness landscape of the protein through positive examples.

已开发出基于实验测量数据训练的监督方法 和仅基于序列训练的无监督方法 用于突变效应预测。无监督突变效应预测器作为任务特定模型在单个蛋白质家族的序列上进行训练。在这种观点下，每个蛋白质都是一个独立的预测任务，目标是评估突变对蛋白质功能的影响。虽然基于多序列比对(MSAs)训练的突变效应预测器通常被描述为无监督的，但它们也可以被视为弱监督的。Hsu等人[15]观察到，此类模型通过MSA对任务具有弱监督，MSA通过正例描述了蛋白质的适应度景观。

If protein language models can learn the information necessary to solve a task from pre-training, then they can be applied directly to new instances of the task, without specialization. This would mean that in practice a single general purpose model can be trained once and then applied to a variety of possible tasks. Thus zero-shot and few-shot transfer represent fundamentally new unsupervised learning capabilities that protein language models can bring to the computational biology toolkit.

如果蛋白质语言模型能够从预训练中学习解决任务所需的信息，那么它们可以直接应用于任务的新实例，而无需专门化。这意味着在实践中，可以训练一个单一的通用模型，然后应用于各种可能的任务。因此，零样本和少样本迁移代表了蛋白质语言模型可以为计算生物学工具包带来的全新无监督学习能力。

# 3 Method

# 3 方法

Protein language models trained with the masked language modeling objective are supervised to output the probability that an amino acid occurs at a position in a protein given the surrounding context. We use this capability to score sequence variations. For a given mutation we can consider the amino acid in the wildtype protein as a reference state, comparing the probability assigned to the mutated amino acid with the probability assigned to the wildtype.

通过掩码语言建模目标训练的蛋白质语言模型被监督输出在给定周围上下文的情况下，某个氨基酸在蛋白质中某个位置出现的概率。我们利用这一能力来评分序列变异。对于给定的突变，我们可以将野生型蛋白质中的氨基酸视为参考状态，比较分配给突变氨基酸的概率与分配给野生型的概率。

|  |  |  |
| --- | --- | --- |
| Models | Full | Test |
| PSSM | 0.460 | 0.460 |
| EVMutation (published) | 0.508 | 0.495 |
| EVMutation (replicated) | 0.511 | 0.498 |
| DeepSequence (published) | 0.514 | 0.499 |
| DeepSequence (replicated) | 0.520 | 0.506 |
| MSA Transformer | 0.542 | 0.524 |
| ESM-1v (zero shot) | 0.509 | 0.482 |
| ESM-1v (+further training) | 0.538 | 0.519 |
| 模型 | 完整 | 测试 |
| PSSM | 0.460 | 0.460 |
| EVMutation(已发表) | 0.508 | 0.495 |
| EVMutation(复现) | 0.511 | 0.498 |
| DeepSequence(已发表) | 0.514 | 0.499 |
| DeepSequence(复现) | 0.520 | 0.506 |
| MSA Transformer | 0.542 | 0.524 |
| ESM-1v(零样本) | 0.509 | 0.482 |
| ESM-1v(进一步训练) | 0.538 | 0.519 |

Table 1: Comparison of protein language models to state-of-the-art methods. Average ISpearman | on full and test sets. DeepSequence and ESM-1v models are each ensembles of 5 models. MSA Transformer is a single model, but is ensembled across 5 random samples of the MSA.

表1:蛋白质语言模型与最先进方法的比较。平均ISpearman | 在完整和测试集上的表现。DeepSequence和ESM-1v模型均为5个模型的集成。MSA Transformer是单一模型，但在MSA的5个随机样本上进行集成。

We score mutations using the log odds ratio at the mutated position, assuming an additive model when multiple mutations exist in the same sequence:

我们使用突变位置的log odds ratio对突变进行评分，假设在同一序列中存在多个突变 时采用加性模型:

Here the sum is over the mutated positions, and the sequence input to the model is masked at every mutated position.

这里的求和是针对突变位置进行的，模型的序列输入在每个突变位置都被掩码。

# 3.1 Zero-shot and few-shot transfer

# 3.1 零样本和少样本迁移

In the zero-shot setting, inference is performed directly on the sequence to be evaluated. Since the MSA Transformer can take multiple sequences as input at inference time, we use this model in the few-shot setting, where additional sequences from the protein family are provided along with the sequence to be evaluated. In both the zero-shot and few-shot settings, only forward passes of the models are performed during inference; no gradient updates are taken. Fig. 2 illustrates the approach in comparison to the current practice of fitting a new model for each task.

在零样本设置中，推理直接在被评估的序列上进行。由于MSA Transformer在推理时可以接受多个序列作为输入，我们在少样本设置中使用该模型，其中提供了来自蛋白质家族的额外序列以及被评估的序列。在零样本和少样本设置中，推理期间仅执行模型的前向传递；不进行梯度更新。图2展示了与当前为每个任务拟合新模型的实践相比的方法。

# 3.2 Inference efficiency

# 3.2 推理效率

Inference with ESM-1v is more efficient than current state-of-the-art methods. This is a result of two important differences: (i) the effect of mutations can be inferred directly without training a task-specific model; (ii) fitness landscapes can be predicted with a single forward pass. Time requirements are summarized in Fig. 7.

使用ESM-1v进行推理比当前最先进的方法更高效。这是由于两个重要差异:(i) 突变的影响可以直接推断，无需训练特定任务模型；(ii) 适应性景观可以通过单次前向传递预测。时间需求总结在图7中。

# 3.3 Scoring with MSA Transformer

# 3.3 使用MSA Transformer进行评分

We score mutations with MSA Transformer using the log odds ratio and additive model in Eq. (1). However, since MSA Transformer uses a set of sequences for inference, we input the sequence to be evaluated as the first sequence, and provide additional sequences from the MSA as context. Masking and scoring are performed on the first sequence only.

我们使用MSA Transformer和公式(1)中的log odds ratio及加性模型对突变进行评分。然而，由于MSA Transformer使用一组序列进行推理，我们将被评估的序列作为第一个序列输入，并提供来自MSA的额外序列作为上下文。掩码和评分仅在第一个序列上进行。

# 4 Results

# 4 结果

# 4.1 Experimental setup

# 4.1 实验设置

Prediction Models We compare to state-of-the-art unsupervised variant prediction methods, EV-Mutation [4] and DeepSequence [20]. We also examine performance of a variety of protein language models that have been recently introduced in the literature.

预测模型 我们与最先进的无监督变异预测方法EV-Mutation [4]和DeepSequence [20]进行比较。我们还检查了最近文献中引入的各种蛋白质语言模型的性能。

The position specific scoring matrix (PSSM), EVmutation [4], and DeepSequence [20] methods are all MSA based. The PSSM treats each position in the sequence independently, factorizing the likelihood into one term per sequence position. EVmutation is a Potts model, which adds pairwise terms modeling the interactions between positions. DeepSequence introduces a latent code, allowing potential higher-order interactions between positions.

位置特异性评分矩阵(PSSM)、EVmutation [4]和DeepSequence [20]方法均基于MSA。PSSM将序列中的每个位置独立处理，将似然分解为每个序列位置的一个项。EVmutation是一个Potts模型，它添加了建模位置之间相互作用的成对项。DeepSequence引入了潜在代码，允许位置之间的潜在高阶相互作用。

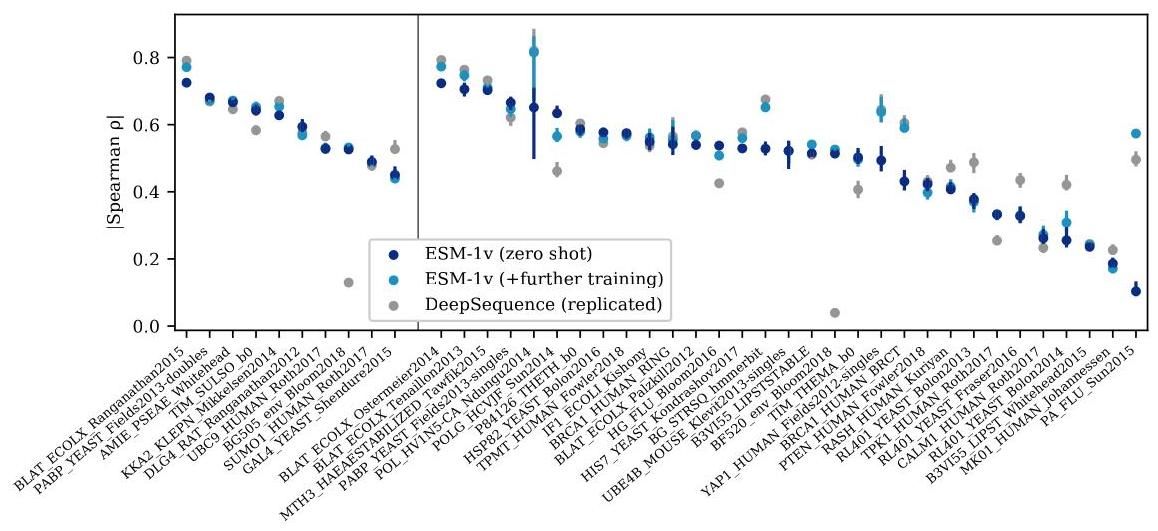


Figure 3: Per task performance. Comparison across 41 deep mutational scanning datasets. Points are ISpearman I on each dataset, error bars show standard deviation of 20 bootstrapped samples. Validation proteins are shown to the left of the dividing line and test proteins to the right. In 17 out of the 41 tasks, ESM-1v zero-shot has a higher ISpearman I than DeepSequence.

图3:每个任务的性能。在41个深度突变扫描数据集上的比较。点是每个数据集上的ISpearman I，误差条显示20个自举样本的标准差。验证蛋白质显示在分隔线左侧，测试蛋白质显示在右侧。在41个任务中的17个任务中，ESM-1v零样本的ISpearman I高于DeepSequence。

UniRep [21], TAPE [22], ProtBERT-BFD [14], ESM-1b [12], and ESM-1v (introduced here), are all single-sequence language models trained on large databases of unaligned and unrelated protein sequences (e.g. Pfam [23] or UniRef [24]). With the exception of UniRep, which is trained using next token prediction, all models are trained with masked language modeling [25].

UniRep [21]、TAPE [22]、ProtBERT-BFD [14]、ESM-1b [12]和ESM-1v(本文引入)都是在未对齐且不相关的蛋白质序列(例如Pfam [23]或UniRef [24])的大型数据库上训练的单序列语言模型。除了使用下一个令牌预测训练的UniRep外，所有模型都使用掩码语言建模[25]进行训练。

Finally, the MSA Transformer [13] is a combination of both approaches; it is trained on a large database of MSAs using masked language modeling and takes an MSA as input during inference.

最后，MSA Transformer [13]是两种方法的结合；它是在使用掩码语言建模的大型MSA数据库上训练的，并在推理期间将MSA作为输入。

ESM-1v We train ESM-1v, a 650M parameter transformer language model for prediction of variant effects, on 98 million diverse protein sequences across evolution. The model is trained only on sequences, without any supervision from experimental measurements of function. We use Uniref90 2020-03 [24], employing the ESM-1b architecture and masked language modeling approach of Rives et al. [12]. The model attains a perplexity of 7.29 on a set of held-out Uniref90 sequences (Table 10). We train five models with different seeds to produce an ensemble.

我们训练了ESM-1v，这是一个拥有6.5亿参数的Transformer语言模型，用于预测变异效应，训练数据涵盖了9800万条来自不同进化阶段的蛋白质序列。该模型仅基于序列进行训练，未使用任何实验测量功能的监督数据。我们使用了Uniref90 2020-03 [24]，并采用了Rives等人[12]提出的ESM-1b架构和掩码语言建模方法。该模型在一组保留的Uniref90序列上达到了7.29的困惑度(表10)。我们使用不同的种子训练了五个模型，以生成一个集成模型。

Evaluation Models are evaluated on a set of 41 deep mutational scans collected by Riesselman et al. [20], which comprise a variety of tasks assessing a diverse set of proteins. Across tasks, the experiments differ in the functions tested and in the measurements performed. We treat each deep mutational scanning dataset as a separate prediction task, scoring each of the variants in the dataset with the model. The tasks are split into a validation set of ten mutational scanning datasets and a test set consisting of the remaining datasets. We evaluate performance by comparing the scores with the experimental measurements using Spearman rank correlation.

评估模型在Riesselman等人[20]收集的41个深度突变扫描数据集上进行评估，这些数据集包含了对多种蛋白质进行的不同任务。在不同任务中，实验测试的功能和进行的测量各不相同。我们将每个深度突变扫描数据集视为一个独立的预测任务，使用模型对数据集中的每个变异进行评分。这些任务被分为一个包含十个突变扫描数据集的验证集和一个由剩余数据集组成的测试集。我们通过使用Spearman秩相关系数将评分与实验测量结果进行比较来评估性能。

Comparisons Since the published versions of EVMutation and DeepSequence use MSAs generated from an earlier version of Uniref100, we generate new MSAs using EVMutation methodology and the version of Uniref100 concurrent with our pretraining dataset. We train replications of EVMutation and DeepSequence using their open source code. The same MSAs are also used in few-shot experiments with MSA Transformer and unsupervised fine-tuning experiments with ESM-1v.

比较由于EVMutation和DeepSequence的已发布版本使用了从早期版本的Uniref100生成的多序列比对(MSA)，我们使用EVMutation的方法和与我们预训练数据集同时期的Uniref100版本生成了新的MSA。我们使用它们的开源代码训练了EVMutation和DeepSequence的复制版本。这些相同的MSA也被用于MSA Transformer的少样本实验和ESM-1v的无监督微调实验。

# 4.2 Language models enable zero-shot and few-shot prediction of the effects of mutations

# 4.2 语言模型能够实现零样本和少样本的突变效应预测

ESM-1v and MSA Transformer models make state-of-the-art predictions. Table 1 compares overall performance of the models across the 41 mutational scanning datasets. Fig. 3 presents a comparison between ESM-1v and DeepSequence on each of the tasks. Zero-shot inference with ESM-1v has a better correlation with experimental measurements than DeepSequence on 17 of the 41 datasets. The two methods are not statistically distinguishable via a paired -test.

ESM-1v和MSA Transformer模型做出了最先进的预测。表1比较了这些模型在41个突变扫描数据集上的整体性能。图3展示了ESM-1v和DeepSequence在每个任务上的比较。在41个数据集中，ESM-1v的零样本推断在17个数据集上与实验测量结果的相关性优于DeepSequence。通过配对 检验，这两种方法在统计上无法区分。

|  |  |  |
| --- | --- | --- |
| Models | Full | Test |
| UniRep | 0.156 | 0.151 |
| TAPE | 0.171 | 0.175 |
| ProtBERT-BFD | 0.428 | 0.399 |
| ESM-1b | 0.459 | 0.424 |
| ESM-1v† | 0.484 | 0.457 |
| ESM-1v\* | 0.509 | 0.482 |
| 模型 | 完整 | 测试 |
| UniRep | 0.156 | 0.151 |
| TAPE | 0.171 | 0.175 |
| ProtBERT-BFD | 0.428 | 0.399 |
| ESM-1b | 0.459 | 0.424 |
| ESM-1v† | 0.484 | 0.457 |
| ESM-1v\* | 0.509 | 0.482 |

Table 2: Zero-shot performance. Average ISpearman I on full and test sets. Average performance of five ESM-1v models. \*Ensemble of the five ESM-1v models.

表2:零样本性能。完整数据集和测试集上的平均ISpearman I。 五个ESM-1v模型的平均性能。\*五个ESM-1v模型的集成。

Table 2 compares protein language models in the zero-shot setting. ESM-1v outperforms existing protein language models TAPE [22], UniRep [21], ProtBERT-BFD [14], and ESM-1b [12]. Fig. 8 breaks down performance across each of the tasks.

表2比较了零样本设置下的蛋白质语言模型。ESM-1v优于现有的蛋白质语言模型TAPE [22]、UniRep [21]、ProtBERT-BFD [14]和ESM-1b [12]。图8分解了每个任务的性能。

Pre-training data We examine the effect of the clustering level of pre-training data. Fig. 4 compares models pre-trained on datasets clustered at increasing sequence identity thresholds. ESM- 1b is trained on sequences clustered at a identity threshold. Improvements are seen using a threshold with greatest improvement at 90%. Uniref100 performance appears to deteriorate early in training despite being the largest of the datasets. These results establish a link between model performance and the data distribution, highlighting the importance of training data in the design of protein language models.

预训练数据 我们研究了预训练数据的聚类水平的影响。图4比较了在逐渐增加的序列同一性阈值下聚类的数据集上预训练的模型。ESM-1b是在 同一性阈值下聚类的序列上训练的。使用 阈值可以看到改进，其中90%的改进最大。尽管Uniref100是数据集中最大的，但其性能在训练早期似乎有所下降。这些结果建立了模型性能与数据分布之间的联系，强调了训练数据在蛋白质语言模型设计中的重要性。

Scoring methods We compare four scoring methods on the validation set - masked marginals, wildtype marginals, mutant marginals, and psuedolikelihood. Table 5 shows that the masked marginal approach described in Eq. (1) outperforms other scoring methods, including ones in which the likelihood changes at non-mutated positions are considered. The scoring methods are described in detail in Appendix A.

评分方法 我们在验证集上比较了四种评分方法——掩码边际、野生型边际、突变边际和伪似然。表5显示，公式(1)中描述的掩码边际方法优于其他评分方法，包括考虑非突变位置似然变化的方法。评分方法在附录A中详细描述。

Parameter count Previous work with protein language models has established a link between model scale and learning of protein structure [12, 26]. We examine zero-shot transfer performance as a function of parameter count. We train models using the same width, depth, and learning rate as described in Henighan et al. [27], observing improvements with scale (Fig. 9). These findings suggest that continued scaling of the models will further improve results.

参数数量 先前关于蛋白质语言模型的工作已经建立了模型规模与蛋白质结构学习之间的联系 [12, 26]。我们研究了零样本迁移性能作为参数数量的函数。我们使用与Henighan等人 [27] 中描述的相同宽度、深度和学习率训练模型，观察到随着规模的增加而改进(图9)。这些发现表明，继续扩展模型将进一步改善结果。

# 4.3 MSA Transformer

# 4.3 MSA Transformer

We examine how the sequences provided to MSA Transformer affect few-shot transfer. Table 8 compares sequence selection methods that vary the diversity of the sequences. Providing a more diverse set of sequences improves few-shot performance. Selecting a set of sequences to maximize diversity outperforms selecting a diversity minimizing set of sequences. Random sampling performs even better, and sampling sequences according to sequence weights [28] performs best.

我们研究了提供给MSA Transformer的序列如何影响少样本迁移。表8比较了不同序列多样性的序列选择方法。提供更多样化的序列集可以提高少样本性能。选择一组序列以最大化多样性优于选择一组最小化多样性的序列。随机采样表现更好，而根据序列权重 [28] 采样序列表现最佳。

We also vary the number of sequences used for inference. Fig. 11 shows few-shot performance as a function of the number of sequences given as input. The model performs well using only a few sequences, but performs best with 384 total sequences. In the main tables we report results sampling 384 sequences using sequence reweighting and ensembling predictions over five different subsamples from the MSA.

我们还改变了用于推理的序列数量。图11显示了少样本性能作为输入序列数量的函数。模型仅使用少量序列时表现良好，但在使用384个序列时表现最佳。在主表中，我们报告了使用序列重加权和从MSA中抽取五个不同子样本进行集成预测的结果。

# 4.4 Unsupervised fine-tuning on MSAs

# 4.4 在MSA上的无监督微调

While ESM-1v performs well when evaluated in the zero-shot setting, we explore whether results can be improved by fine-tuning on the MSA. Fine-tuning on MSAs has been used in previous work as a stage in transfer learning to specialize a pre-trained model to a protein family, before applying supervision with labeled data. Here we consider using the fine-tuned model to make unsupervised predictions directly, without adding supervision from experimental data.

虽然ESM-1v在零样本设置下表现良好，但我们探讨是否可以通过在MSA上微调来改进结果。在先前的工作中，MSA上的微调已被用作迁移学习的一个阶段，以将预训练模型专门化到蛋白质家族，然后再使用标记数据进行监督。在这里，我们考虑使用微调模型直接进行无监督预测，而不添加实验数据的监督。

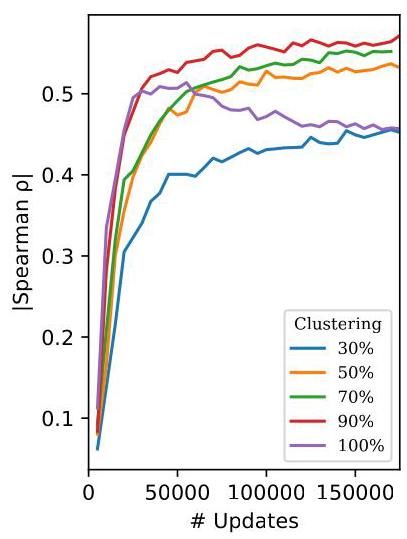


Figure 4: Comparison of pre-training datasets. Average |Spearman | on the single-mutation validation set. While a clustering threshold was used for ESM-1b, training with clustering results in a significant improvement on variant prediction tasks. Notably, models trained on Uniref100, the largest dataset in this figure, appear to deteriorate early in training. These results establish a link between model performance and the data distribution, and highlight the importance of training data in the design of protein language models.

图4:预训练数据集的比较。单突变验证集上的平均|Spearman |。虽然ESM-1b使用了 聚类阈值，但使用 聚类进行训练在变异预测任务上显著改进。值得注意的是，训练在Uniref100上的模型(图中最大的数据集)在训练早期似乎有所下降。这些结果建立了模型性能与数据分布之间的联系，并强调了训练数据在蛋白质语言模型设计中的重要性。

We observe that naively fine-tuning the model on the MSA results in rapid overfitting and poor performance on the prediction tasks (Fig. 12). While we experiment with a variety of approaches to freezing parameters during fine-tuning, detailed in Appendix B, none produce significant improvements. We find that an approach using pre-training sequences to regularize the fine-tuning performs well and enables training of all parameters without overfitting (Fig. 13). Spiked fine-tuning improves average absolute Spearman rho on the full dataset from 0.510 for zero-shot evaluation to 0.537 with fine-tuning.

我们观察到，在MSA上简单微调模型会导致快速过拟合和预测任务上的较差性能(图12)。虽然我们尝试了在微调期间冻结参数的各种方法(详见附录B)，但没有一种方法产生显著改进。我们发现，使用预训练序列来正则化微调的方法表现良好，并且能够训练所有参数而不会过拟合(图13)。尖峰微调将完整数据集上的平均绝对Spearman rho从零样本评估的0.510提高到微调后的0.537。

# 5 Analysis of models

# 5 模型分析

Protein structure and function ESM-1v probabilities reflect the functional properties of sites within the protein. We use the entropy of the model’s predictions for a position as a measure of its estimation of conservation. The lowest entropy predictions cluster at binding sites. Fig. 14 compares the distribution of the model’s entropy between binding sites and non-binding sites. A significant difference is observed between the entropy assignment to binding and non-binding site residues. Fig. 5 visualizes the side chains of the 10 lowest entropy residues as predicted by the model on the crystal structure of DNA methyltransferase M.HaeIII interacting with its DNA substrate. In the crystal structure a cytosine of the substrate is inserted into the active site of the enzyme. The low entropy residues cluster in the active site and interact with the cytosine. Additional examples are visualized in Fig. 18.

蛋白质结构与功能 ESM-1v 概率反映了蛋白质内部位点的功能特性。我们使用模型对某个位置预测的熵作为其保守性估计的度量。最低熵的预测集中在结合位点。图14比较了模型在结合位点与非结合位点之间熵的分布。在结合位点与非结合位点残基的熵分配之间观察到显著差异。图5可视化了模型预测的10个最低熵残基的侧链，这些残基位于DNA甲基转移酶M.HaeIII与其DNA底物相互作用的晶体结构中。在晶体结构中，底物的胞嘧啶插入到酶的活性位点中。低熵残基集中在活性位点并与胞嘧啶相互作用。其他示例在图18中可视化。

The model probabilities also correspond to structure. Fig. 15 compares the entropy assigned to sites that are buried in the core of the protein vs. exposed on the surface. The model assigns significantly lower entropy to sites that are in the core of the protein, consistent with the idea that tight packing in the core places greater constraints on the selection of residues. Fig. 5B visualizes the entropy assigned by the model to each position overlayed on the structure of Indole-3-glycerolphosphate Synthase, a TIM barrel protein. Higher entropy is assigned to residues having outward facing side chains on the alpha helices, while lower entropy is assigned to the inward facing positions. Fig. 17 compares the probability assigned to hydrophobic, polar, and charged amino acids for buried sites vs. non-buried sites. The model prefers hydrophobic residues in the core and hydrophilic residues on the surface. The model probabilities closely match the empirical probabilities and those from the PSSM. Fig. 5C visualizes probability assigned to hydrophobic amino acids on the structure of

模型概率也与结构相对应。图15比较了埋藏在蛋白质核心中的位点与暴露在表面的位点的熵分配。模型对位于蛋白质核心的位点分配了显著较低的熵，这与核心紧密堆积对残基选择施加更大限制的观点一致。图5B可视化了模型对每个位置分配的熵，这些熵叠加在吲哚-3-甘油磷酸合成酶(一种TIM桶蛋白)的结构上。较高熵分配给α螺旋上侧链朝外的残基，而较低熵分配给侧链朝内的位置。图17比较了埋藏位点与非埋藏位点上疏水性、极性和带电氨基酸的概率分配。模型倾向于在核心选择疏水性残基，在表面选择亲水性残基。模型概率与经验概率和PSSM的概率高度匹配。图5C可视化了疏水性氨基酸在结构上的概率分配。

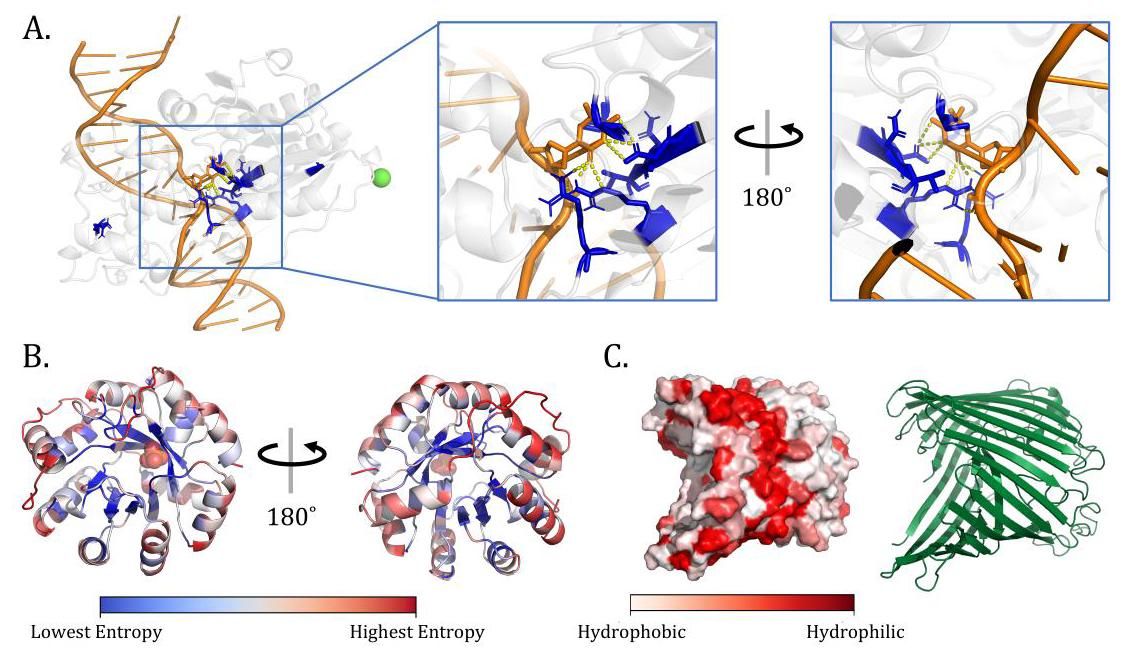


Figure 5: ESM-1v reflects the molecular basis of function in proteins. (A) DNA methylase HaeIII (pdbid: 1DCT [29]). Side chains for the top 10 positions with lowest prediction entropy shown in blue. Low-entropy positions cluster in the active site. (B) TIM Barrel (pdbid: 1IGS [30]) with residues colored by entropy. The model’s predictions for residues on the surface have highest entropy (red) while those in the core have lower entropy (blue). Notably, residues on the alpha helices show a clear gradient from high to low entropy as residues transition from surface-facing to core-facing. (C) Sucrose-specific Porin (pdbid: 1A0T [31]), a transmembrane protein. The model predicts a hydrophobic band where the protein is embedded in the membrane.

图5:ESM-1v反映了蛋白质功能的分子基础。(A) DNA甲基化酶HaeIII(pdbid: 1DCT [29])。预测熵最低的前10个位置的侧链以蓝色显示。低熵位置集中在活性位点。(B) TIM桶(pdbid: 1IGS [30])，残基按熵着色。模型对表面残基的预测具有最高熵(红色)，而对核心残基的预测具有较低熵(蓝色)。值得注意的是，α螺旋上的残基在从表面朝向核心过渡时显示出从高熵到低熵的明显梯度。(C) 蔗糖特异性孔蛋白(pdbid: 1A0T [31])，一种跨膜蛋白。模型预测了一个疏水带，该带位于蛋白质嵌入膜的位置。

Sucrose-specific Porin, a transmembrane protein. The model predicts a hydrophobic band in the center where the protein embeds in the membrane.

蔗糖特异性孔蛋白，一种跨膜蛋白。模型预测在蛋白质嵌入膜的中心位置有一个疏水带。

Calibration We evaluate model calibration using 15008 sequences with length from the trRosetta [32] dataset. ESM-1v probabilities for each amino acid at each position are calculated with the masked marginal probability in Eq. (1). Fig. 6 shows that the model is generally well calibrated for all amino acids except Methionine. ESM-1v always predicts Methionine as the first position in the sequence since full protein sequences always start with it, so care must be used when applying the model to subsequences. When excepting the first residue, the model achieves an average calibration error (defined for the multi-class setting in Appendix D.4) of 0.006 .

校准 我们使用来自trRosetta [32]数据集的15008条长度为 的序列来评估模型校准。ESM-1v对每个位置每个氨基酸的概率通过公式(1)中的掩码边缘概率计算。图6显示，除甲硫氨酸外，模型对所有氨基酸的校准效果普遍良好。ESM-1v总是将甲硫氨酸预测为序列的第一个位置，因为完整蛋白质序列总是以它开头，因此在将模型应用于子序列时必须小心。当排除第一个残基时，模型的平均校准误差(在附录D.4中为多类设置定义)为0.006。

We also explore the relationship between conservation (entropy of the PSSM) and the model’s predicted entropy. Fig. 16 shows that these are well correlated (Pearson’s ), suggesting the model is able to identify conserved positions.

我们还探讨了保守性(PSSM的熵)与模型预测熵之间的关系。图16显示，这两者具有良好的相关性(皮尔逊相关系数 )，表明模型能够识别保守位置。

# 6 Related Work

# 6 相关工作

# 6.1 Protein language models

# 6.1 蛋白质语言模型

In the past few years, a number of groups have developed language models for protein sequences . These models have been used for many tasks, including supervised low-N function prediction [16, 12], remote homology detection [22, 12], and protein generation [35]. The approach to the tasks typically involves transfer learning, where a pretrained language model is fine-tuned for a particular problem. Vig et al. [36] and Rao et al. [26] found that transformer attention corresponds to known biological properties such as structure and binding sites and can be used to predict contacts.

过去几年中，许多研究小组开发了用于蛋白质序列的语言模型 。这些模型已被用于许多任务，包括监督下的低N功能预测[16, 12]、远程同源检测[22, 12]和蛋白质生成[35]。这些任务的方法通常涉及迁移学习，即对预训练的语言模型进行微调以解决特定问题。Vig等人[36]和Rao等人[26]发现，Transformer的注意力机制与已知的生物学特性(如结构和结合位点)相对应，并可用于预测接触。

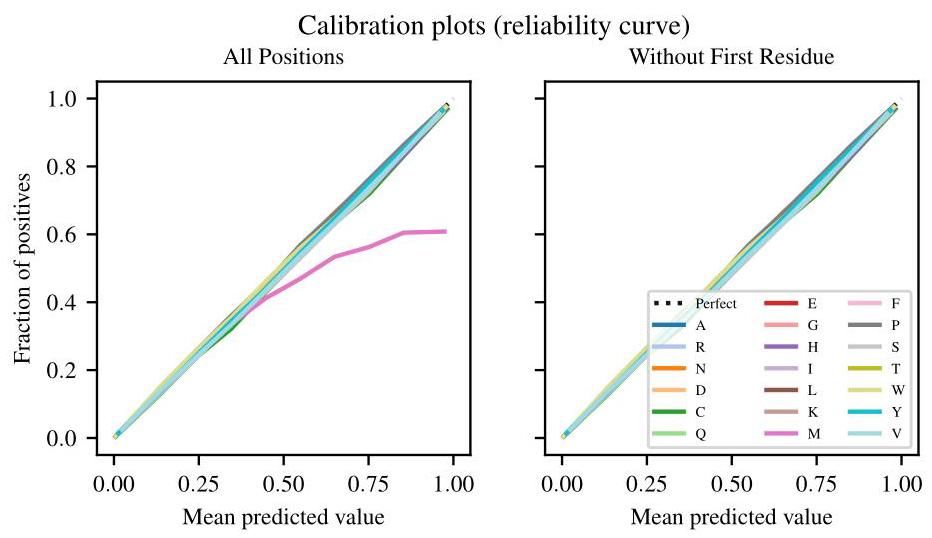


Figure 6: Calibration plot for ESM-1v predictions on each of the 20 naturally occurring amino acids on the trRosetta dataset. The multi-class classification is converted into a set of 20 one-versus-all classifications for the purpose of this analysis. Left and right plots show calibration of all positions and positions excluding the first residue, respectively. Since full sequences always start with Methionine, the model overwhelmingly predicts it in the first position. When evaluating the model on subsequences, such as those in the trRosetta dataset, this causes a miscalibration at the first residue. Including the first residue, the model has an average calibration error (ACE) of 0.011 in the first case and 0.006 in the second.

图6:ESM-1v在trRosetta数据集上对20种天然氨基酸的预测校准图。为了进行分析，多类分类被转换为20个一对多的分类。左图和右图分别显示了所有位置和排除第一个残基的位置的校准。由于完整序列总是以甲硫氨酸(Methionine)开始，模型在第一个位置几乎总是预测它。当在子序列(如trRosetta数据集中的子序列)上评估模型时，这会导致第一个残基的校准错误。包括第一个残基时，模型在第一种情况下的平均校准误差(ACE)为0.011，在第二种情况下为0.006。

# 6.2 Mutation effect prediction

# 6.2 突变效应预测

Supervised and unsupervised methods have been developed for prediction of mutational effects. Supervised methods train models using experimental measurements or labels from databases of clinical variants. Standard machine learning tools including linear regression, random forests, and support-vector machines can be used [37]. Models have been designed specifically for proteins, using feature engineering such as Envision [38] and PolyPhen-2 [39], ensemble methods such as Revel [40], MPC [41], CADD [42], and M-CAP [43], language models such as UniRep [21, 16] and ESM [12], and other representation learning approaches [44, 45].

已经开发了有监督和无监督的方法来预测突变效应。有监督方法使用实验测量或临床变异数据库中的标签来训练模型。可以使用标准的机器学习工具，包括线性回归、随机森林和支持向量机[37]。专门为蛋白质设计的模型使用了特征工程，如Envision[38]和PolyPhen-2[39]，集成方法如Revel[40]、MPC[41]、CADD[42]和M-CAP[43]，语言模型如UniRep[21, 16]和ESM[12]，以及其他表示学习方法[44, 45]。

Unsupervised mutation effect predictors work by inferring the likelihood of a mutation from the evolutionary landscape of the original protein. A density model fit to related sequences is used for scoring. SIFT [46] is a first order approach using a position-specific-scoring-matrix. EVMutation [4] extends this to a second-order approach by training a Potts model on the MSA. DeepSequence [20] includes higher-order interactions by training a VAE on the MSA instead, using the ELBO to score mutations. Riesselman et al. [47] proposes using an autoregressive model that does not require the sequences to be aligned.

无监督的突变效应预测器通过从原始蛋白质的进化景观中推断突变的可能性来工作。使用与相关序列拟合的密度模型进行评分。SIFT[46]是一种使用位置特异性评分矩阵的一阶方法。EVMutation[4]通过在MSA上训练Potts模型将其扩展到二阶方法。DeepSequence[20]通过在MSA上训练VAE来包含高阶相互作用，使用ELBO对突变进行评分。Riesselman等人[47]提出使用自回归模型，该模型不需要序列对齐。

Hsu et al. [15] show that unsupervised mutational effect predictors can be extended to perform supervised predictions, with better unsupervised predictors generally resulting in better supervised predictors. This suggests improving unsupervised prediction can drive progress in both settings. Concurrent with our work, Hie et al. [48] use open-source protein language models ESM-1b and TAPE to predict the direction of evolution in protein fitness landscapes.

Hsu等人[15]表明，无监督的突变效应预测器可以扩展到执行有监督的预测，通常更好的无监督预测器会带来更好的有监督预测器。这表明改进无监督预测可以推动这两种情况下的进展。与我们的工作同时，Hie等人[48]使用开源蛋白质语言模型ESM-1b和TAPE来预测蛋白质适应性景观中的进化方向。

# 7 Discussion

# 7 讨论

Advances in language modeling at scale are bringing the goal of a general purpose model for proteins closer to realization. This line of work aspires to a model that learns to read and write biology in its native language, that can be directly applied across a range of protein understanding and design tasks. For scalability, learning from sequences is important: while there are no central databases of high-throughput functional measurements, and few compilations exist, billions of sequences are available to learn from in sequence databases . Sequences give an unparalleled view into the vast diversity and complexity of molecular parts invented by nature through billions of years of evolution.

大规模语言建模的进展使得通用蛋白质模型的目标更接近实现。这项工作旨在开发一个能够以生物学的原生语言读取和编写生物学的模型，该模型可以直接应用于一系列蛋白质理解和设计任务。为了可扩展性，从序列中学习是重要的:虽然没有高通量功能测量的中央数据库，且现有的汇编很少，但序列数据库中有数十亿个序列可供学习 。序列提供了对自然界通过数十亿年进化发明的分子部分的巨大多样性和复杂性的无与伦比的视角。

Unsupervised structure and function learning methods first effectively realized the idea that biological properties could be read directly from sequences without supervision from experimental measurements. However these methods are not general purpose in the sense that a specialized model must be trained for every protein for which a prediction is to be made. We show that the same performance can be realized by a general purpose model that has been trained across many diverse protein families. Similar to observations on the learning of tertiary protein structure in large language models [12, 26], we find that increasing the scale of models leads to improvements in function learning. The understanding of mutational landscapes in the models correlates with the molecular basis of function in proteins, capturing binding sites and amino acid preferences that are determined by the folded structure.

无监督的结构 和功能 学习方法首次有效地实现了从序列中直接读取生物特性的想法，而无需实验测量的监督。然而，这些方法并不是通用的，因为必须为每个要预测的蛋白质训练一个专门的模型。我们表明，通过一个在许多不同蛋白质家族上训练的通用模型可以实现相同的性能。类似于在大语言模型中对三级蛋白质结构学习的观察[12, 26]，我们发现增加模型的规模会带来功能学习的改进。模型中对突变景观的理解与蛋白质功能的分子基础相关，捕捉了由折叠结构决定的结合位点和氨基酸偏好。

Zero-shot transfer is an interesting capability of large scale language models, and represents a major point of departure from the unsupervised learning methods that are the basis for current state-of-the-art inference of protein structure and function. The capability for zero-shot transfer implies that a model can be trained once and then applied to perform inference for many tasks. It is also a window into deeper questions about the forms of generalization that are possible in learning from sequences. Reading structural and functional design principles from sequences is a necessary capability for writing new biologically active sequences. Generalization in the zero-shot setting suggests the potential for large language models to capture knowledge that can be transferred to generating new functional proteins.

零样本迁移是大规模语言模型的一个有趣能力，代表了与当前最先进的蛋白质结构和功能推断基础的无监督学习方法的主要区别。零样本迁移的能力意味着模型可以训练一次，然后应用于执行许多任务的推断。它也是关于从序列学习中可能实现的泛化形式的更深层次问题的窗口。从序列中读取结构和功能设计原则是编写新的生物活性序列的必要能力。零样本设置中的泛化表明，大语言模型有可能捕捉到可以转移到生成新功能蛋白质的知识。

# Acknowledgments and Disclosure of Funding

# 致谢和资金披露

We thank Naman Goyal, Chloe Hsu, Zeming Lin, Ishan Misra, Myle Ott, Sergey Ovchinnikov, and Ethan Perez for valuable input on the paper. Roshan Rao was supported by funding from Facebook, Berkeley Deep Drive, and DARPA XAI.

我们感谢Naman Goyal、Chloe Hsu、Zeming Lin、Ishan Misra、Myle Ott、Sergey Ovchinnikov和Ethan Perez对本文的宝贵意见。Roshan Rao获得了Facebook、Berkeley Deep Drive和DARPA XAI的资助。

# References

# 参考文献

[1] Charles J Epstein, Robert F Goldberger, and Christian B Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In Cold Spring Harbor symposia on quantitative biology, volume 28, pages 439-449. Cold Spring Harbor Laboratory Press, 1963.

Charles J Epstein, Robert F Goldberger, 和 Christian B Anfinsen。三级蛋白质结构的遗传控制:模型系统研究。在《冷泉港定量生物学研讨会》第28卷，第439-449页。冷泉港实验室出版社，1963年。

[2] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. Nature methods, 11(8):801, 2014.

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

[3] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. Molecular biology and evolution, 33(1):268-280, 2016.

Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, 和 Martin Weigt。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。

[4] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. Nature Biotechnology, 35(2):128-135, feb 2017. ISSN 15461696. doi: 10.1038/nbt.3769.

Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, 和 Debora S. Marks。从序列共变异预测突变效应。《自然生物技术》，35(2):128-135，2017年2月。ISSN 15461696。doi: 10.1038/nbt.3769。

[5] C Yanofsky, V Horn, and D Thorpe. Protein Structure Relationships Revealed By Mutational Analysis. Science (New York, N.Y.), 146(3651):1593-4, 12 1964. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/14224506.

C Yanofsky, V Horn, 和 D Thorpe。通过突变分析揭示的蛋白质结构关系。《科学》(纽约)，146(3651):1593-4，1964年12月。ISSN 0036-8075。URL http://www.ncbi.nlm.nih.gov/pubmed/14224506。

[6] D. Altschuh, A.M. Lesk, A.C. Bloomer, and A. Klug. 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.

D. Altschuh, A.M. Lesk, A.C. Bloomer, 和 A. Klug。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年。ISSN 0022-2836。doi: 10.1016/0022-2836(87)90352-4。

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

D Altschuh, T Vernet, P Berti, D Moras, 和 K Nagai。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年。

[8] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Genetics, 18(4):309-317, 4 1994. ISSN 0887-3585. doi: 10.1002/prot.340180402. URL http://www.ncbi.nlm.nih.gov/pubmed/8208723http://doi.wiley.com/10.1002/prot.340180402.

Ulrike Göbel, Chris Sander, Reinhard Schneider, 和 Alfonso Valencia。蛋白质中的相关突变和残基接触。《蛋白质:结构、功能与遗传学》，18(4):309-317，1994年4月。ISSN 0887-3585。doi: 10.1002/prot.340180402。URL http://www.ncbi.nlm.nih.gov/pubmed/8208723http://doi.wiley.com/10.1002/prot.340180402。

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, 和 Ilya Sutskever。语言模型是无监督的多任务学习者。2019年。

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. CoRR, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, 和 Dario Amodei。语言模型是少样本学习者。CoRR, abs/2005.14165，2020年。URL https://arxiv.org/abs/2005.14165。

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, 和 Ilya Sutskever。从自然语言监督中学习可迁移的视觉模型，2021年。

[12] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. bioRxiv, page 622803, 4 2020. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v3.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, 和 Rob Fergus. 通过将无监督学习扩展到2.5亿个蛋白质序列，生物结构和功能得以涌现。bioRxiv, 页码622803, 2020年4月. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v3.

[13] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. bioRxiv, page 2021.02.12.430858, feb 2021. doi: 10.1101/2021.02.12.430858.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, 和 Alexander Rives. MSA Transformer. bioRxiv, 页码2021.02.12.430858, 2021年2月. doi: 10.1101/2021.02.12.430858.

[14] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing. bioRxiv, 72020. URL http://arxiv.org/abs/2007.06225.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, 和 Burkhard Rost. ProtTrans: 通过自监督深度学习和高性能计算破解生命密码的语言。bioRxiv, 2020年7月. URL http://arxiv.org/abs/2007.06225.

[15] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Combining evolutionary and assay-labelled data for protein fitness prediction. bioRxiv, page 2021.03.28.437402, mar 2021. doi: 10.1101/2021.03.28.437402.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, 和 Jennifer Listgarten. 结合进化和实验标记数据进行蛋白质适应性预测。bioRxiv, 页码2021.03.28.437402, 2021年3月. doi: 10.1101/2021.03.28.437402.

[16] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-N protein engineering with data-efficient deep learning. bioRxiv, page 2020.01.23.917682, 2020. doi: 10.1101/2020.01.23.917682.

Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, 和 George M Church. 使用数据高效的深度学习进行低N蛋白质工程。bioRxiv, 页码2020.01.23.917682, 2020年. doi: 10.1101/2020.01.23.917682.

[17] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 951-958. IEEE, 2009.

Christoph H Lampert, Hannes Nickisch, 和 Stefan Harmeling. 通过类间属性转移学习检测未见过的对象类别。在2009年IEEE计算机视觉与模式识别会议上, 页码951-958. IEEE, 2009年.

[18] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In , volume 1, page 3,2008.

Hugo Larochelle, Dumitru Erhan, 和 Yoshua Bengio. 零数据学习新任务。在 ,第1卷,第3页,2008年.

[19] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, and Sutskever I. Zero-shot text-to-image generation.

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, 和 Sutskever I. 零样本文本到图像生成.

[20] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10):816-822, 10 2018. ISSN 15487105. doi: 10.1038/s41592-018-0138-4.

Adam J. Riesselman, John B. Ingraham, 和 Debora S. Marks. 深度生成模型捕捉遗传变异的影响。自然方法, 15(10):816-822, 2018年10月. ISSN 15487105. doi: 10.1038/s41592-018-0138-4.

[21] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-only deep representation learning. Nature Methods, 12:1315-1322, 3 2019. ISSN 15487105. doi: 10.1101/589333. URL https://www.biorxiv.org/content/10.1101/589333v1.

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, 和 George M. Church. 仅使用序列的深度表示学习进行统一的理性蛋白质工程。自然方法, 12:1315-1322, 2019年3月. ISSN 15487105. doi: 10.1101/589333. URL https://www.biorxiv.org/content/10.1101/589333v1.

[22] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. In Neural Information Processing Systems. Cold Spring Harbor Laboratory, 62019. doi: 10.1101/676825. URL https://doi.org/10.1101/676825http://arxiv.org/abs/1906.08230.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, 和 Yun S. Song. 使用TAPE评估蛋白质迁移学习。在神经信息处理系统中. 冷泉港实验室, 2019年6月. doi: 10.1101/676825. URL https://doi.org/10.1101/676825http://arxiv.org/abs/1906.08230.

[23] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L.L. Sonnhammer, John Tate, and Marco Punta. Pfam: The protein families database, 1 2014. ISSN 03051048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965110/.

Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L.L. Sonnhammer, John Tate, 和 Marco Punta. Pfam: 蛋白质家族数据库, 2014年1月. ISSN 03051048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965110/.

[24] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23(10):1282- 1288, 5 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm098. URL http://www.uniprot.org.

Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, 和 Cathy H. Wu. 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 13674803. doi: 10.1093/bioinformatics/btm098. URL http://www.uniprot.org.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171- 4186, Minneapolis, Minnesota, 6 2019. Association for Computational Linguistics. doi: 10. 18653/v1/N19-1423. URL http://arxiv.org/abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, 和 Kristina Toutanova. BERT:用于语言理解的深度双向Transformer预训练。在2019年北美计算语言学协会会议:人类语言技术, 第1卷(长篇和短篇论文), 页4171-4186, 明尼苏达州明尼阿波利斯, 2019年6月. 计算语言学协会. doi: 10.18653/v1/N19-1423. URL http://arxiv.org/abs/1810.04805.

[26] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. ICLR, page 2020.12.15.422761, 12 2021. doi: 10.1101/2020.12.15.422761.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, 和 Alexander Rives. Transformer蛋白质语言模型是无监督的结构学习器。ICLR, 页2020.12.15.422761, 2021年12月. doi: 10.1101/2020.12.15.422761.

[27] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. CoRR, abs/2010.14701, 2020. URL https://arxiv.org/abs/2010.14701.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, 和 Sam McCandlish. 自回归生成模型的缩放定律。CoRR, abs/2010.14701, 2020. URL https://arxiv.org/abs/2010.14701.

[28] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America, 108(49):E1293-E1301, 12 2011. ISSN 00278424. doi: 10.1073/pnas.1111471108.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, 和 Martin Weigt. 残基共进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年12月. ISSN 00278424. doi: 10.1073/pnas.1111471108.

[29] KM Reinish, L Chen, GL Verdine, and WN Lipscomb. The crystal structure of haeiii methyl-transferase covalently complexed to dna: an extrahelical cytosine and rearranged base pairing. Cell, 82(1):143-153, 1995.

KM Reinish, L Chen, GL Verdine, 和 WN Lipscomb. HaeIII甲基转移酶与DNA共价复合物的晶体结构:一个螺旋外的胞嘧啶和重排的碱基配对。细胞, 82(1):143-153, 1995.

[30] Michael Hennig, Beatrice Darimont, Reinhard Sterner, Kasper Kirschner, and Johan N Jansonius. 2.0 "a structure of indole-3-glycerol phosphate synthase from the hyperthermophile sulfolobus solfataricus: possible determinants of protein stability. Structure, 3(12):1295-1306, 1995.

Michael Hennig, Beatrice Darimont, Reinhard Sterner, Kasper Kirschner, 和 Johan N Jansonius. 来自超嗜热菌Sulfolobus solfataricus的吲哚-3-甘油磷酸合酶的2.0 Å结构:蛋白质稳定性的可能决定因素。结构, 3(12):1295-1306, 1995.

[31] Doris Forst, Wolfram Welte, Thomas Wacker, and Kay Diederichs. Structure of the sucrose-specific porin scry from salmonella typhimurium and its complex with sucrose. Nature structural biology, 5(1):37-46, 1998.

Doris Forst, Wolfram Welte, Thomas Wacker, 和 Kay Diederichs. 来自鼠伤寒沙门氏菌的蔗糖特异性孔蛋白ScrY及其与蔗糖的复合物的结构。自然结构生物学, 5(1):37-46, 1998.

[32] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, David Baker, and John Harvard. Improved protein structure prediction using predicted inter-residue orientations. bioRxiv, page 846279, 2019. doi: 10.1101/846279. URL https://www.biorxiv.org/content/10.1101/846279v1.

Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, David Baker, 和 John Harvard. 使用预测的残基间方向改进蛋白质结构预测。bioRxiv, 页846279, 2019. doi: 10.1101/846279. URL https://www.biorxiv.org/content/10.1101/846279v1.

[33] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure, 2 2019. URL http://arxiv.org/abs/1902.08661https://arxiv.org/ abs/1902.08661.

Tristan Bepler 和 Bonnie Berger. 利用结构信息学习蛋白质序列嵌入, 2019年2月. URL http://arxiv.org/abs/1902.08661https://arxiv.org/abs/1902.08661.

[34] Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. bioRxiv, page 2020.09.04.283929, 9 2020. doi: 10.1101/2020.09.04.283929. URL https://doi.org/10.1101/2020.09.04.283929.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, 和 Alan Moses. 通过互信息最大化进行蛋白质表示的自监督对比学习. bioRxiv, 页码 2020.09.04.283929, 2020年9月. doi: 10.1101/2020.09.04.283929. URL https://doi.org/10.1101/2020.09.04.283929.

[35] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language Modeling for Protein Generation. bioRxiv, 3 2020. URL http://arxiv.org/abs/2004.03497.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, 和 Richard Socher. ProGen: 用于蛋白质生成的语言建模. bioRxiv, 2020年3月. URL http://arxiv.org/abs/2004.03497.

[36] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models. bioRxiv, page 2020.06.26.174417, 6 2020. doi: 10.1101/2020.06.26.174417. URL http: //arxiv.org/abs/2006.15222.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, 和 Nazneen Fatema Rajani. BERTology 遇见生物学: 解释蛋白质语言模型中的注意力机制. bioRxiv, 页码 2020.06.26.174417, 2020年6月. doi: 10.1101/2020.06.26.174417. URL http://arxiv.org/abs/2006.15222.

[37] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering, aug 2019. ISSN 15487105.

Kevin K. Yang, Zachary Wu, 和 Frances H. Arnold. 机器学习引导的定向进化用于蛋白质工程, 2019年8月. ISSN 15487105.

[38] Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, and Douglas M. Fowler. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Systems, 6(1):116-124.e3, jan 2018. ISSN 24054720. doi: 10.1016/j.cels.2017.11.003.

Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, 和 Douglas M. Fowler. 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.e3, 2018年1月. ISSN 24054720. doi: 10.1016/j.cels.2017.11.003.

[39] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations, apr 2010. ISSN 15487091.

Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, 和 Shamil R. Sunyaev. 一种预测有害错义突变的方法和服务器, 2010年4月. ISSN 15487091.

[40] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. Revel: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics, 99(4):877-885, 2016.

Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, 等. Revel: 一种预测罕见错义变异致病性的集成方法. 美国人类遗传学杂志, 99(4):877-885, 2016年.

[41] Kaitlin E Samocha, Jack A Kosmicki, Konrad J Karczewski, Anne H O’Donnell-Luria, Emma Pierce-Hoffman, Daniel G MacArthur, Benjamin M Neale, and Mark J Daly. Regional missense constraint improves variant deleteriousness prediction. BioRxiv, page 148353, 2017.

Kaitlin E Samocha, Jack A Kosmicki, Konrad J Karczewski, Anne H O’Donnell-Luria, Emma Pierce-Hoffman, Daniel G MacArthur, Benjamin M Neale, 和 Mark J Daly. 区域错义约束改善变异有害性预测. BioRxiv, 页码 148353, 2017年.

[42] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O’Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics, 46(3):310-315, 2014.

Martin Kircher, Daniela M Witten, Preti Jain, Brian J O’Roak, Gregory M Cooper, 和 Jay Shendure. 一种估计人类遗传变异相对致病性的通用框架. 自然遗传学, 46(3):310-315, 2014年.

[43] Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein, and Gill Bejerano. M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature genetics, 48(12): 1581, 2016.

Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein, 和 Gill Bejerano. M-cap 在高灵敏度下消除了临床外显子组中大多数意义不确定的变异. 自然遗传学, 48(12): 1581, 2016年.

[44] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. Nature genetics, 50(8):1161-1170, 2018.

Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, 等. 使用深度神经网络预测人类突变的临床影响. 自然遗传学, 50(8):1161-1170, 2018年.

[45] Haicang Zhang, Michelle S Xu, Wendy K Chung, and Yufeng Shen. Predicting functional effect of missense variants using graph attention neural networks. bioRxiv, 2021.

Haicang Zhang, Michelle S Xu, Wendy K Chung, 和 Yufeng Shen. 使用图注意力神经网络预测错义变异的功能效应. bioRxiv, 2021年.

[46] Ngak Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Research, 40(W1):W452, jul 2012. ISSN 03051048. doi: 10.1093/nar/gks539.

Ngak Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, 和 Pauline C. Ng. SIFT 网络服务器:预测氨基酸替换对蛋白质的影响。核酸研究, 40(W1):W452, 2012年7月. ISSN 03051048. doi: 10.1093/nar/gks539.

[47] Adam Riesselman, Jung-Eun Shin, Aaron Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew Kruse, and Debora Marks. Accelerating Protein Design Using Autoregressive Generative Models. bioRxiv, page 757252, 2019. doi: 10.1101/757252. URL https://www.biorxiv.org/content/10.1101/757252v1.

Adam Riesselman, Jung-Eun Shin, Aaron Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew Kruse, 和 Debora Marks. 使用自回归生成模型加速蛋白质设计。bioRxiv, 页 757252, 2019. doi: 10.1101/757252. URL https://www.biorxiv.org/content/10.1101/757252v1.

[48] Brian L Hie, Kevin K Yang, and Peter S Kim. Evolutionary velocity with protein language models. bioRxiv, 2021.

Brian L Hie, Kevin K Yang, 和 Peter S Kim. 使用蛋白质语言模型进行进化速度分析。bioRxiv, 2021.

[49] Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the department of energy joint genome institute: 2014 updates. Nucleic acids research, 42(D1): D26-D31, 2014.

Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, 和 Inna Dubchak. 能源部联合基因组研究所的基因组门户:2014年更新。核酸研究, 42(D1): D26-D31, 2014.

[50] M. Steinegger, M. Mirdita, and J. Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods, 16(7):603-606, 07 2019.

M. Steinegger, M. Mirdita, 和 J. Söding. 蛋白质级组装显著提高了从宏基因组样本中恢复蛋白质序列的效率。自然方法, 16(7):603-606, 2019年7月.

[51] Alan S. Lapedes, Bertrand G. Giraud, LonChang Liu, and Gary D. Stormo. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. Lecture Notes-Monograph Series, 33:236-256, 1999. doi: 10.2307/4356049. URL http://www.jstor.org/stable/ 4356049.

Alan S. Lapedes, Bertrand G. Giraud, LonChang Liu, 和 Gary D. Stormo. 蛋白质序列模型中的相关突变:系统发育和结构效应。讲义-专论系列, 33:236-256, 1999. doi: 10.2307/4356049. URL http://www.jstor.org/stable/ 4356049.

[52] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. Graphical models of residue coupling in protein families, 42008. ISSN 15455963. URL https://pubmed.ncbi.nlm.nih.gov/18451428/.

John Thomas, Naren Ramakrishnan, 和 Chris Bailey-Kellogg. 蛋白质家族中残基耦合的图形模型, 42008. ISSN 15455963. URL https://pubmed.ncbi.nlm.nih.gov/18451428/.

[53] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences of the United States of America, 106(1):67-72, 1 2009. ISSN 00278424. doi: 10.1073/pnas.0805923106. URL https://www.pnas.org/content/ 106/1/67https://www.pnas.org/content/106/1/67.abstract.

Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, 和 Terence Hwa. 通过消息传递识别蛋白质-蛋白质相互作用中的直接残基接触。美国国家科学院院刊, 106(1):67-72, 2009年1月. ISSN 00278424. doi: 10.1073/pnas.0805923106. URL https://www.pnas.org/content/ 106/1/67https://www.pnas.org/content/106/1/67.abstract.

[54] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics, 79(4):1061-1078, 4 2011. ISSN 08873585. doi: 10.1002/prot.22934. URL http://doi.wiley.com/10.1002/prot.22934.

Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, 和 Christopher James Langmead. 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年4月. ISSN 08873585. doi: 10.1002/prot.22934. URL http://doi.wiley.com/10.1002/prot.22934.

[55] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSI-COV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, 28(2):184-190, 1 2012. ISSN 1460-2059. doi: 10.1093/bioinformatics/btr638. URL https://academic.oup.com/bioinformatics/ article-lookup/doi/10.1093/bioinformatics/btr638.

David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, 和 Massimiliano Pontil. PSI-COV:使用稀疏逆协方差估计进行大规模多序列比对中的精确结构接触预测。生物信息学, 28(2):184-190, 2012年1月. ISSN 1460-2059. doi: 10.1093/bioinformatics/btr638. URL https://academic.oup.com/bioinformatics/ article-lookup/doi/10.1093/bioinformatics/btr638.

[56] Emily E Wrenbeck, Matthew S Faber, and Timothy A Whitehead. Deep sequencing methods for protein engineering and design. Current opinion in structural biology, 45:36-44, 2017.

Emily E Wrenbeck, Matthew S Faber, 和 Timothy A Whitehead. 用于蛋白质工程和设计的深度测序方法。结构生物学当前观点, 45:36-44, 2017.

[57] Justin R Klesmith, John-Paul Bacik, Ryszard Michalczyk, and Timothy A Whitehead. Comprehensive sequence-flux mapping of a levoglucosan utilization pathway in e. coli. ACS synthetic biology, 4(11):1235-1243, 2015.

Justin R Klesmith, John-Paul Bacik, Ryszard Michalczyk, 和 Timothy A Whitehead. 在大肠杆菌中左旋葡聚糖利用途径的全面序列-通量图谱。ACS合成生物学, 4(11):1235-1243, 2015.

[58] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of hiv envelope. Elife, 7:e34420, 2018.

Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, 和 Jesse D Bloom。绘制HIV包膜蛋白进化景观中的突变效应图。Elife, 7:e34420, 2018.

[59] Philip A Romero, Tuan M Tran, and Adam R Abate. Dissecting enzyme function with microfluidic-based deep mutational scanning. Proceedings of the National Academy of Sciences, 112(23):7159-7164, 2015.

Philip A Romero, Tuan M Tran, 和 Adam R Abate。利用微流控深度突变扫描解析酶功能。美国国家科学院院刊, 112(23):7159-7164, 2015.

[60] Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A comprehensive, high-resolution map of a gene’s fitness landscape. Molecular biology and evolution, 31(6): 1581-1592, 2014.

Elad Firnberg, Jason W Labonte, Jeffrey J Gray, 和 Marc Ostermeier。基因适应性景观的全面高分辨率图谱。分子生物学与进化, 31(6): 1581-1592, 2014.

[61] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, and Timothy Palzkill. Deep sequencing of systematic combinatorial libraries reveals -lactamase sequence constraints at high resolution. Journal of molecular biology, 424(3-4):150-167, 2012.

Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G Brown, Carolyn J Adamski, Kacie Rice, Donna Muzny, Richard A Gibbs, 和 Timothy Palzkill。系统组合文库的深度测序揭示了 -内酰胺酶序列的高分辨率限制。分子生物学杂志, 424(3-4):150-167, 2012.

[62] Michael A Stiffler, Doeke R Hekstra, and Rama Ranganathan. Evolvability as a function of purifying selection in tem-1 -lactamase. Cell,160(5):882-892,2015.

Michael A Stiffler, Doeke R Hekstra, 和 Rama Ranganathan。TEM-1 -内酰胺酶的进化能力作为纯化选择的函数。细胞,160(5):882-892,2015.

[63] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, et al. Capturing the mutational landscape of the beta-lactamase tem-1. Proceedings of the National Academy of Sciences, 110(32):13067-13072, 2013.

Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, 等. 捕捉β-内酰胺酶TEM-1的突变景观. 美国国家科学院院刊, 110(32):13067-13072, 2013.

[64] Scott D Findlay and Lynne-Marie Postovit. Comprehensive characterization of transcript diversity at the human nodal locus. BioRxiv, page 254409, 2018.

Scott D Findlay 和 Lynne-Marie Postovit. 人类Nodal位点转录多样性的全面表征. BioRxiv, 页 254409, 2018.

[65] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ran-ganathan. The spatial architecture of protein function and adaptation. Nature, 491(7422): 138-142, 2012.

Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, 和 Rama Ranganathan. 蛋白质功能和适应的空间结构. 自然, 491(7422): 138-142, 2012.

[66] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. Nature methods, 12(3):203-206, 2015.

Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, 和 Jay Shendure. 大规模并行单氨基酸突变. 自然方法, 12(3):203-206, 2015.

[67] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. Viruses, 8(6):155, 2016.

Michael B Doud 和 Jesse D Bloom. 准确测量所有氨基酸突变对流感血凝素的影响. 病毒, 8(6):155, 2016.

[68] K Pokusaeva, C Johnson, B Luk, G7 Uribe, Y Fu, N Oezguen, RK Matsunami, M Lugo, A Major, Y Mori-Akiyama, et al. Gaba-producing bifidobacterium dentium modulates visceral sensitivity in the intestine. Neurogastroenterology & Motility, 29(1):e12904, 2017.

K Pokusaeva, C Johnson, B Luk, G7 Uribe, Y Fu, N Oezguen, RK Matsunami, M Lugo, A Major, Y Mori-Akiyama, 等. 产生GABA的双歧杆菌dentium调节肠道内脏敏感性. 神经胃肠病学与运动性, 29(1):e12904, 2017.

[69] Parul Mishra, Julia M Flynn, Tyler N Starr, and Daniel NA Bolon. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. Cell reports, 15(3):588-598, 2016.

Parul Mishra, Julia M Flynn, Tyler N Starr, 和 Daniel NA Bolon. 系统突变分析阐明Hsp90功能的一般和客户特异性方面. 细胞报告, 15(3):588-598, 2016.

[70] Eric D Kelsic, Hattie Chung, Niv Cohen, Jimin Park, Harris H Wang, and Roy Kishony. Rna structural determinants of optimal codons revealed by mage-seq. Cell systems, 3(6):563-571, 2016.

Eric D Kelsic, Hattie Chung, Niv Cohen, Jimin Park, Harris H Wang, 和 Roy Kishony. RNA结构决定最优密码子的MAGE-seq揭示. 细胞系统, 3(6):563-571, 2016.

[71] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic acids research, 42(14):e112-e112, 2014.

Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, 和 Tarjei S Mikkelsen. 体内激酶全面突变扫描揭示底物依赖的适应性景观. 核酸研究, 42(14):e112-e112, 2014.

[72] Lisa Brenan, Aleksandr Andreev, Ofir Cohen, Sasha Pantel, Atanas Kamburov, Davide Cac-chiarelli, Nicole S Persky, Cong Zhu, Mukta Bagul, Eva M Goetz, et al. Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants. Cell reports, 17(4):1171-1183, 2016.

Lisa Brenan, Aleksandr Andreev, Ofir Cohen, Sasha Pantel, Atanas Kamburov, Davide Cacchiarelli, Nicole S Persky, Cong Zhu, Mukta Bagul, Eva M Goetz, 等. MAPK1/ERK2错义突变的全面表型表征. 细胞报告, 17(4):1171-1183, 2016.

[73] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, and Dan S Tawfik. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. PLoS computational biology, 11(8):e1004421, 2015.

Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, 和 Dan S Tawfik. 通过长期漂移系统映射蛋白质突变空间揭示看似中性突变的有害影响. PLoS计算生物学, 11(8):e1004421, 2015.

[74] Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. PLoS genetics, 11(7):e1005310, 2015.

Nicholas C Wu, C Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, 等. 病毒蛋白质功能约束分析揭示进化保守性与功能性的不一致. PLoS遗传学, 11(7):e1005310, 2015.

[75] Christopher D Aakre, Julien Herrou, Tuyen N Phung, Barrett S Perchuk, Sean Crosson, and Michael T Laub. Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell, 163(3):594-606, 2015.

Christopher D Aakre, Julien Herrou, Tuyen N Phung, Barrett S Perchuk, Sean Crosson, 和 Michael T Laub. 通过混杂中间体进化新的蛋白质-蛋白质相互作用特异性. 细胞, 163(3):594-606, 2015.

[76] Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. PLoS pathogens, 10(4):e1004064, 2014.

Hangfei Qi, C Anders Olson, Nicholas C Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, 等. 定量高分辨率遗传谱快速识别丙型肝炎病毒适应性和药物敏感性的序列决定因素. PLoS病原体, 10(4):e1004064, 2014.

[77] Kenneth A Matreyek, Lea M Starita, Jason J Stephany, Beth Martin, Melissa A Chiasson, Vanessa E Gray, Martin Kircher, Arineh Khechaduri, Jennifer N Dines, Ronald J Hause, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature genetics, 50(6):874-882, 2018.

Kenneth A Matreyek, Lea M Starita, Jason J Stephany, Beth Martin, Melissa A Chiasson, Vanessa E Gray, Martin Kircher, Arineh Khechaduri, Jennifer N Dines, Ronald J Hause, 等。通过大规模并行测序对蛋白质变体丰度进行多重评估。《自然遗传学》，50(6):874-882, 2018。

[78] Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, et al. Deconstruction of the ras switching cycle through saturation mutagenesis. Elife, 6:e27810, 2017.

Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, 等。通过饱和诱变解构Ras转换周期。《Elife》，6:e27810, 2017。

[79] Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, and Daniel NA Bolon. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. Journal of molecular biology, 425(8):1363-1377, 2013.

Benjamin P Roscoe, Kelly M Thayer, Konstantin B Zeldovich, David Fushman, 和 Daniel NA Bolon。分析所有泛素点突变对酵母生长速率的影响。《分子生物学杂志》，425(8):1363-1377, 2013。

[80] Benjamin P Roscoe and Daniel NA Bolon. Systematic exploration of ubiquitin sequence, e1 activation efficiency, and experimental fitness in yeast. Journal of molecular biology, 426(15): 2854-2870, 2014.

Benjamin P Roscoe 和 Daniel NA Bolon。系统探索泛素序列、E1激活效率和酵母中的实验适应性。《分子生物学杂志》，426(15): 2854-2870, 2014。

[81] David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A Barad, Alain R Bonny, Clinton L Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D Axen, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. Elife, 5: e15802, 2016.

David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A Barad, Alain R Bonny, Clinton L Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D Axen, 等。在课堂环境中确定不同化学压力下的泛素适应性景观。《Elife》，5: e15802, 2016。

[82] Yvonne H Chan, Sergey V Venev, Konstantin B Zeldovich, and C Robert Matthews. Correlation of fitness landscapes from three orthologous tim barrels originates from sequence and structure constraints. Nature communications, 8(1):1-12, 2017.

Yvonne H Chan, Sergey V Venev, Konstantin B Zeldovich, 和 C Robert Matthews。来自三个同源TIM桶的适应性景观的相关性源于序列和结构约束。《自然通讯》，8(1):1-12, 2017。

[83] Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. Rna, 19(11):1537-1551, 2013.

Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, 和 Stanley Fields。对酿酒酵母多聚(A)结合蛋白的RRM域进行深度突变扫描。《RNA》，19(11):1537-1551, 2013。

[84] Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. Proceedings of the National Academy of Sciences, 110(14):E1263-E1272, 2013.

Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, 和 Rachel E Klevit。通过高通量诱变鉴定的E3泛素连接酶中的活性增强突变。《美国国家科学院院刊》，110(14):E1263-E1272, 2013。

[85] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the National Academy of Sciences, 109 (42):16858-16863, 2012.

Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, 和 Stanley Fields。一个基本的蛋白质特性，热力学稳定性，仅从大规模蛋白质功能测量中揭示。《美国国家科学院院刊》，109 (42):16858-16863, 2012。

[86] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Pseudolikelihood reranking with masked language models. CoRR, abs/1910.14659, 2019. URL http://arxiv.org/abs/ 1910.14659.

Julian Salazar, Davis Liang, Toan Q. Nguyen, 和 Katrin Kirchhoff。使用掩码语言模型进行伪似然重排序。《CoRR》，abs/1910.14659, 2019。URL http://arxiv.org/abs/ 1910.14659。

[87] Kevin K. Yang, Zachary Wu, Claire N. Bedbrook, and Frances H. Arnold. Learned protein embeddings for machine learning. Bioinformatics, 34(15):2642-2648, 8 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty178.

Kevin K. Yang, Zachary Wu, Claire N. Bedbrook, 和 Frances H. Arnold。用于机器学习的蛋白质嵌入学习。《生物信息学》，34(15):2642-2648, 8 2018。ISSN 14602059。doi: 10.1093/bioinformatics/bty178。

[88] Andrew Leaver-Fay, Matthew J O’meara, Mike Tyka, Ron Jacak, Yifan Song, Elizabeth H Kellogg, James Thompson, Ian W Davis, Roland A Pache, Sergey Lyskov, et al. Scientific benchmarks for guiding macromolecular energy function improvement. Methods in enzymology, 523:109-143, 2013.

Andrew Leaver-Fay, Matthew J O’meara, Mike Tyka, Ron Jacak, Yifan Song, Elizabeth H Kellogg, James Thompson, Ian W Davis, Roland A Pache, Sergey Lyskov, 等。指导大分子能量函数改进的科学基准。《酶学方法》，523:109-143, 2013。

[89] Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research, 45(D1):D170-D176, 1 2017. ISSN 13624962. doi: 10.1093/nar/gkw1081.

Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, 和 Martin Steinegger。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, 1 2017。ISSN 13624962。doi: 10.1093/nar/gkw1081。

[90] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection. Technical report, oct 2018.

Lukas Neumann、Andrew Zisserman 和 Andrea Vedaldi。松弛Softmax:用于安全行人检测的高效置信度自动校准。技术报告，2018年10月。

[91] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring Calibration in Deep Learning. apr 2019. URL http: //arxiv.org/abs/1904.01685.

Jeremy Nixon、Mike Dusenberry、Ghassen Jerfel、Timothy Nguyen、Jeremiah Liu、Linchuan Zhang 和 Dustin Tran。深度学习中的校准测量。2019年4月。URL http://arxiv.org/abs/1904.01685。

[92] L. Steven Johnson, Sean R. Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics, 11(1):431, 8 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-431. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-431.

L. Steven Johnson、Sean R. Eddy 和 Elon Portugaly。隐马尔可夫模型速度启发式和迭代HMM搜索程序。BMC生物信息学，11(1):431，2010年8月。ISSN 14712105。doi: 10.1186/1471-2105-11-431。URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-431。

[93] Didier Nurizzo, Steven C Shewry, Michael H Perlin, Scott A Brown, Jaydev N Dholakia, Roy L Fuchs, Taru Deva, Edward N Baker, and Clyde A Smith. The crystal structure of aminoglycoside- 3-phosphotransferase-iia, an enzyme responsible for antibiotic resistance. Journal of molecular biology, 327(2):491-506, 2003.

Didier Nurizzo、Steven C Shewry、Michael H Perlin、Scott A Brown、Jaydev N Dholakia、Roy L Fuchs、Taru Deva、Edward N Baker 和 Clyde A Smith。氨基糖苷-3-磷酸转移酶-IIa的晶体结构，这是一种导致抗生素抗性的酶。分子生物学杂志，327(2):491-506，2003年。

# A Extraction methods

# A 提取方法

ESM-1v is pre-trained to output the probability for each possible amino acid at a masked position. We explore four methods of scoring the effects of mutations using the model:

ESM-1v 经过预训练，能够输出在掩码位置每个可能氨基酸的概率。我们探索了四种使用该模型对突变效应进行评分的方法:

* Masked marginal: Probabilities are extracted according to the mask noise during pretraining. At each position, we introduce a mask token and record the model’s predicted probabilities of the tokens at that position.
* 掩码边际:根据预训练期间的掩码噪声提取概率。在每个位置，我们引入一个掩码标记，并记录模型在该位置对标记的预测概率。
* Mutant marginal: Probabilities are extracted according to the random token noise during pre-training. Among the 15% predicted positions in the sequence during pre-training, 10% of those are randomly mutated and 10% retain their original identities. The model is tasked to predict the correct token at those positions. Therefore, in this extraction method, we follow the pre-training methodology by passing in mutated tokens and recording the model’s probability that they are correct.
* 突变边际:根据预训练期间的随机token噪声提取概率。在预训练期间预测的序列中15%的位置中，10%的位置被随机突变，10%的位置保留其原始身份。模型的任务是预测这些位置上的正确标记。因此，在这种提取方法中，我们遵循预训练方法，传入突变的标记并记录模型认为它们正确的概率。
* Wildtype marginal: We perform a single forward pass using the wildtype sequence. This method enables fast scoring as just a single forward pass is used.
* 野生型边际:我们使用野生型序列进行单次前向传递。这种方法能够快速评分，因为只使用了一次前向传递。
* Pseudolikelihood: This method is proposed in the literature for scoring with masked language models [86].
* 伪似然:这种方法在文献中提出，用于使用掩码语言模型进行评分 [86]。

In all cases, we assume an additive model when multiple mutations are present in a sequence. Results

在所有情况下，当序列中存在多个突变时，我们假设一个加性模型。结果

are summarized in Tables 5 and 7.

总结在表5和表7中。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Formal setting | Task level supervision | Representative ods |  |
| Supervised mutation prediction | Supervised | Direct supervision from ex- perimental measurements | Reviewed in [87] |
| Model trained on se- quences from individ- ual family | Unsupervised | Weak positive supervision from MSA | [4, 20] |
| Fine-tuning on experi- mental data | Semi-supervised trans- fer | Supervision from experi- mental measurements | [21, 16, 12] |
| Fine-tuning on MSA | Transfer learning with weak-positive supervi- sion | Weak positive supervision from MSA | Introduced for trans- fer learning in [21] |
| Direct forward pass | Zero-shot learning | None | This work |
| 方法 | 正式设置 | 任务级监督 | 代表性 方法 |  |
| 监督突变预测 | 监督 | 来自实验测量的直接监督 | 在[87]中回顾 |
| 在单个家族的序列上训练的模型 | 无监督 | 来自MSA的弱正监督 | [4, 20] |
| 在实验数据上微调 | 半监督迁移 | 来自实验测量的监督 | [21, 16, 12] |
| 在MSA上微调 | 带有弱正监督的迁移学习 | 来自MSA的弱正监督 | 在[21]中引入用于迁移学习 |
| 直接前向传递 | 零样本学习 | 无 | 本工作 |

Table 3: Zero-shot learning is a natural extension of the various approaches that have been used for mutational effect prediction to date. Rather than training a new model for every task, a single general purpose model is trained and can be directly applied across multiple tasks. The approach is fully unsupervised, no information from experimental measurements of function is used.

表3:零样本学习是迄今为止用于突变效应预测的各种方法的自然延伸。与其为每个任务训练一个新模型，不如训练一个通用的单一模型，并可以直接应用于多个任务。该方法完全无监督，不使用来自功能实验测量的任何信息。

Let and represent the mutant and wildtype sequences. We refer to as the sequence with a mask introduced at position . We refer to the set of mutations that are introduced as the set . For example, if mutations are introduced at positions 3 and 6, then .

设 和 分别代表突变体和野生型序列。我们将 称为在位置 引入掩码的序列 。我们将引入的突变集合称为集合 。例如，如果在位置3和6引入突变，则 。

Masked marginal probability (L forward passes) This method performs best among the four. We introduce masks at the mutated positions and compute the score for a mutation by considering its probability relative to the wildtype amino acid (Strategy a):

掩码边缘概率(L次前向传递) 该方法在四种方法中表现最佳。我们在突变位置引入掩码，并通过考虑突变相对于野生型氨基酸的概率来计算突变得分(策略a):

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.450648; this version posted November 17, 2021. 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/10.1101/2021.07.09.450648；此版本发布于2021年11月17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。

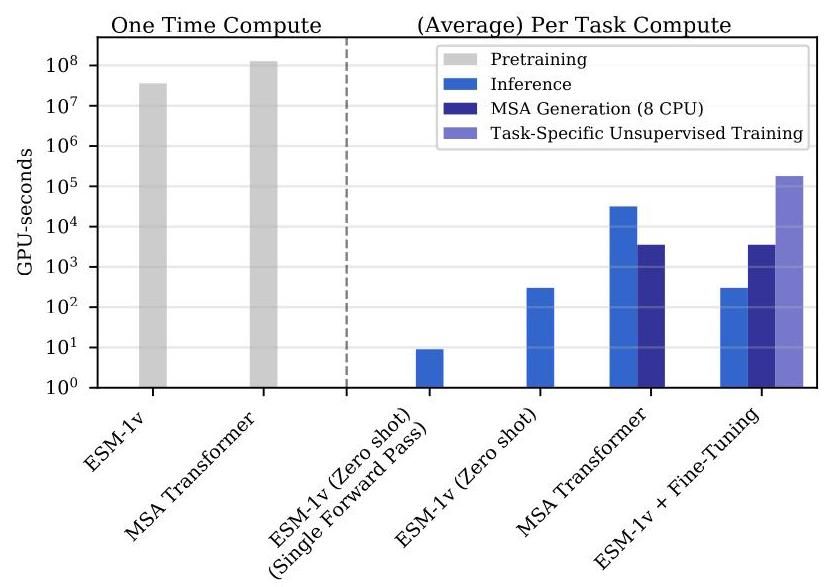


Figure 7: Compute requirements in GPU-seconds for (left) pre-training and (right) average task. With open-sourced pre-trained models, end users bypass the pre-training phase and only incur inference costs. ESM-1v and MSA Transformer amortize compute cost into a single expensive pre-training run. After pre-training, inference is fast. On average, it takes 10 seconds to label a deep mutational scan from Riesselman et al. [20] with ESM-1v (Zero-shot, Single Forward Pass). Performance improves marginally with the more expensive scoring scheme (Table 5).

图7:(左)预训练和(右)平均任务的GPU秒计算需求。使用开源的预训练模型，最终用户绕过预训练阶段，仅产生推理成本。ESM-1v和MSA Transformer将计算成本分摊到一次昂贵的预训练运行中。预训练后，推理速度很快。平均而言，使用ESM-1v(零样本，单次前向传递)标记Riesselman等人[20]的深度突变扫描需要10秒。使用更昂贵的评分方案，性能略有提升(表5)。

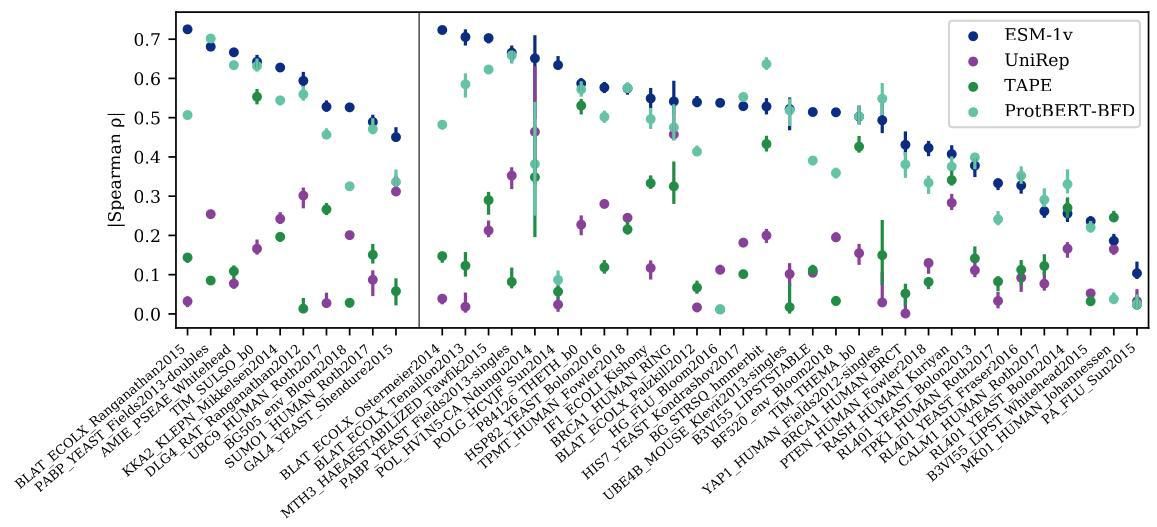


Figure 8: Zero-shot performance of ESM-1v compared to earlier protein language models on all 41 deep mutational scans. Points are lSpearman I on each dataset, error bars show standard deviation of 20 bootstrapped samples. Validation proteins are shown to the left of the dividing line and test proteins to the right. ESM-1v is the best performing method on 30 of the 41 deep mutational scans.

图8:ESM-1v与早期蛋白质语言模型在所有41个深度突变扫描中的零样本性能比较。点表示每个数据集上的lSpearman I，误差条显示20个自举样本的标准差。验证蛋白质显示在分界线的左侧，测试蛋白质显示在右侧。ESM-1v在41个深度突变扫描中的30个上表现最佳。

This formulation assumes an additive model, consistent with the training objective. We show that this assumption is justified empirically by evaluating the model with different choices at the non-mutated positions. First, the wildtype sequence (Strategy b):

该公式假设了一个加性模型，与训练目标一致。我们通过评估模型在非突变位置的不同选择，证明这一假设在经验上是合理的。首先，野生型序列(策略b):

and the mutant sequence (Strategy c):

以及突变序列(策略c):

|  |  |
| --- | --- |
| Clustering | | Spearman | |
| 30% | 0.456 |
| 50% | 0.537 |
| 70% | 0.552 |
| 90% | 0.564 |
| 100% | 0.458 |
| 聚类 | | 斯皮尔曼 | |
| 30% | 0.456 |
| 50% | 0.537 |
| 70% | 0.552 |
| 90% | 0.564 |
| 100% | 0.458 |

Table 4: Average ISpearman I on the single-mutation validation set after training a parameter Transformer model for 170,000 updates on various sequence identity clusterings of Uniref.

表4:在Uniref的不同序列同一性聚类上训练一个 参数Transformer模型进行170,000次更新后，单突变验证集上的平均ISpearman I。

|  |  |
| --- | --- |
| Method | | Spearman | |
| Masked marginal | 0.582 |
| Mutant marginal | 0.578 |
| Wildtype marginal | 0.572 |
| Pseudo-likelihood | 0.552 |
| 方法 | | 斯皮尔曼 | |
| 掩码边缘 | 0.582 |
| 突变体边缘 | 0.578 |
| 野生型边缘 | 0.572 |
| 伪似然 | 0.552 |

Table 5: Benchmarking scoring schemes on the single-mutation validation set. The means across the validation set are listed. The masked marginal scheme performs best.

表5:在单突变验证集上对评分方案进行基准测试。列出了验证集的平均值。掩码边缘方案表现最佳。

Strategy (a), where we mask all positions at the same time, performs best on the PABP Yeast Doubles validation dataset (Table 7).

策略(a)，即同时掩码所有位置，在PABP酵母双突变验证数据集上表现最佳(表7)。

Mutant marginal probability This method is analogous to the wildtype marginal probability, except we use the mutant sequence instead.

突变体边缘概率 该方法与野生型边缘概率类似，只是我们使用突变体序列代替。

This method requires a single forward pass for every mutation.

该方法需要对每个突变进行一次前向传递。

Wildtype marginal probability ( 1 forward pass) In the fastest scheme, we perform a single forward pass using the wildtype sequence as input. For a set of mutations at positions , the score is:

野生型边缘概率(1次前向传递) 在最快的方案中，我们使用野生型序列作为输入进行一次前向传递。对于位置 上的一组突变，得分为:

We find that the method performs well with a minor 1% decrease in absolute performance, while requiring very limited computational resources. The strong performance indicates that the masked language modeling objective causes the model to capture the fitness landscape of the protein in its outputs.

我们发现该方法表现良好，绝对性能仅下降1%，同时所需的计算资源非常有限。其强劲的表现表明，掩码语言建模目标使模型在其输出中捕捉到了蛋白质的适应度景观。

Pseudolikelihood Psuedolikelihood has been proposed in the literature as a method to score sequences using masked language models [86]. We compute the score as follows:

伪似然 伪似然已在文献中被提出作为一种使用掩码语言模型对序列进行评分的方法[86]。我们按如下方式计算得分:

As mutation prediction is a ranking task and as the contribution from the second term is constant throughout the deep mutation scan (i.e. the wildtype sequence is always the same), we can safely drop it from the computation.

由于突变预测是一个排序任务，并且第二项的贡献在整个深度突变扫描中是恒定的(即野生型序列始终相同)，我们可以安全地将其从计算中省略。

# A.1 Evaluation

# A.1 评估

We compare the methods described above on the validation set, finding that the masked marginal scheme performs best. To determine the specific mode of inference when multiple mutations are present, we examine each method on the "doubles" component of the PABP Yeast dataset finding the masked marginal (a) strategy performs best. This scoring method is used across the results.

我们在验证集上比较了上述方法，发现掩码边缘方案表现最佳。为了确定存在多个突变时的具体推理模式，我们在PABP酵母数据集的“双突变”部分上检查了每种方法，发现掩码边缘(a)策略表现最佳。该评分方法在结果中广泛使用。

|  |  |  |
| --- | --- | --- |
| Input | Consensus columns only | | Spearman | |
| MSA seed | Yes | 0.573 |
| MSA seed | No | 0.567 |
| Uniprot | N/A | 0.582 |
| 输入 | 仅共识列 | | 斯皮尔曼 | |
| MSA种子 | 是 | 0.573 |
| MSA种子 | 否 | 0.567 |
| Uniprot | 不适用 | 0.582 |

Table 6: ESM-1v performs better when including the full protein sequence as listed in Uniprot, compared to using the seed sequence of the MSA corresponding to the deep mutational scan. Results on single-mutation validation set. The means across the validation set are listed. We experiment with a number of strategies for inference: (i) the consensus columns only; (ii) the aligned part of the query sequence; and (iii) the complete Uniprot sequence. The complete Uniprot sequence performs best, possibly because the model was pre-trained on complete Uniprot sequences. We use the MSA seed sequence from the MSAs released by [20] corresponding to the deep mutational scans.

表6:与使用对应于深度突变扫描的MSA种子序列相比，ESM-1v在包含Uniprot中列出的完整蛋白质序列时表现更好。单突变验证集的结果。列出了验证集的平均值。我们尝试了多种推理策略:(i) 仅使用共识列；(ii) 查询序列的对齐部分；(iii) 完整的Uniprot序列。完整的Uniprot序列表现最佳，可能是因为模型是在完整的Uniprot序列上预训练的。我们使用了由[20]发布的对应于深度突变扫描的MSA种子序列。

|  |  |
| --- | --- |
|  | | Spearman | |
| Masked marginal (a) | 0.692 |
| Masked marginal (b) | 0.482 |
| Masked marginal (c) | 0.483 |
| Mutant marginal | 0.694 |
| Wildtype marginal | 0.672 |
| Pseudo-likelihood | 0.608 |
|  | | 斯皮尔曼 | |
| 掩蔽边缘 (a) | 0.692 |
| 掩蔽边缘 (b) | 0.482 |
| 掩蔽边缘 (c) | 0.483 |
| 突变边缘 | 0.694 |
| 野生型边缘 | 0.672 |
| 伪似然 | 0.608 |

Table 7: Ablating scoring schemes on the PABP Yeast Doubles dataset. The masked marginal scheme performs best when masking all mutated sites together. Mean absolute Spearman across the single-mutation validation tasks is reported.

表7:在PABP酵母双突变数据集上消融评分方案。当同时掩码所有突变位点时，掩码边缘方案表现最佳。报告了单突变验证任务中的平均绝对斯皮尔曼 。

# A.2 Evaluating ESM-1v on subsequences

# A.2 评估ESM-1v在子序列上的表现

DeepSequence, EVMutation, and the MSA Transformer use the consensus columns of a MSA as input. We construct MSAs using the seed sequences from the DeepSequence paper, which usually correspond to a subsequence of the protein capturing the domain where the deep mutational scan was performed.

DeepSequence、EVMutation和MSA Transformer使用MSA的共识列作为输入。我们使用DeepSequence论文中的种子序列构建MSA，这些序列通常对应于蛋白质的子序列，捕获了进行深度突变扫描的域。

Table 6 explores using the MSA seed sequence vs. the full Uniprot sequence for inference on the validation set. We find that the full Uniprot sequence performs best, possibly because the model was pre-trained on Uniprot sequences. We note in Figure Fig. 6 that the model captures some bias in the Uniprot dataset, for example that most proteins begin with a methionine (corresponding to the start codon).

表6探讨了在验证集上使用MSA种子序列与完整Uniprot序列进行推理的效果。我们发现完整的Uniprot序列表现最佳，可能是因为模型在Uniprot序列上进行了预训练。我们在图6中注意到，模型捕捉到了Uniprot数据集中的一些偏差，例如大多数蛋白质以甲硫氨酸(对应于起始密码子)开头。

# B Unsupervised fine-tuning ESM-1v

# B 无监督微调ESM-1v

Experimental setup We assess a number of approaches for fine-tuning ESM-1v on task-specific MSAs. We evaluate modeling decisions by fine-tuning on tasks from the validation set and examining the mean change in Spearman over the course of training. For efficiency, we compute Spearman using the wildtype marginal strategy, as this requires just a single forward pass. After the final modeling decisions are selected, we train all models for 7500 updates and evaluate on all proteins using the masked marginals strategy. All models in this section were trained with a constant learning rate of using the masked language modeling objective. For reference, the ESM-1v pre-training was performed with a target batch size of tokens.

实验设置 我们评估了多种在任务特定MSA上微调ESM-1v的方法。我们通过在验证集任务上进行微调并检查训练过程中斯皮尔曼 的平均变化来评估建模决策。为了提高效率，我们使用野生型边缘策略计算斯皮尔曼 ，因为这只需要一次前向传递。在最终建模决策选定后，我们对所有模型进行7500次更新训练，并使用掩码边缘策略对所有蛋白质进行评估。本节中的所有模型均以恒定学习率 使用掩码语言建模目标进行训练。作为参考，ESM-1v的预训练目标批量大小为 个标记。

Unsupervised fine-tuning baselines The concept of unsupervised fine-tuning of an MSA has been previously proposed [21, 16]. Fig. 12 studies a basic fine-tuning setup on the consensus columns of the MSA. Each model is fine-tuned on a single MSA with a target batch size of 8192 tokens. We first observe that that models overfit quickly if the entire model is trained. This results in a decrease in Spearman compared to initialization. As the fine-tuning is performed on the consensus columns of the MSA, we sought to regularize the model by fine-tuning only the embeddings. As a PSSM already captures information relevant to the task, we hypothesize that tuning the embeddings could

无监督微调基线 先前已经提出了MSA无监督微调的概念[21, 16]。图12研究了在MSA共识列上的基本微调设置。每个模型在单个MSA上进行微调，目标批量大小为8192个标记。我们首先观察到，如果训练整个模型，模型会迅速过拟合。这导致斯皮尔曼 相比初始化时有所下降。由于微调是在MSA的共识列上进行的，我们试图通过仅微调嵌入来正则化模型。由于PSSM已经捕获了与任务相关的信息，我们假设调整嵌入可以

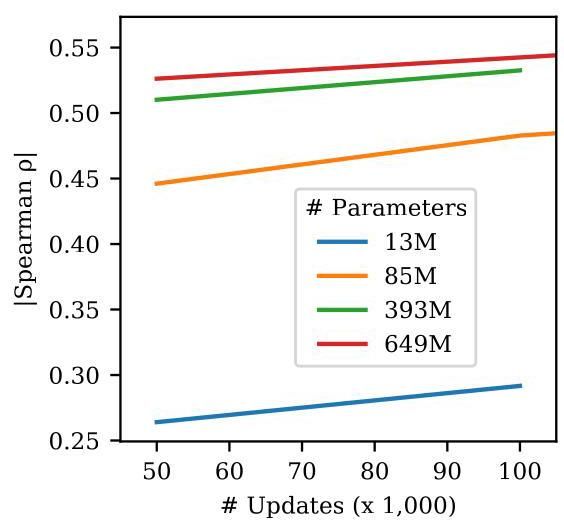


Figure 9: Larger models perform better on variant prediction. We trained four models of various scales, following the hyperparameters listed in Henighan et al. [27]. Results on single-mutation validation set.

图9:更大规模的模型在变异预测上表现更好。我们训练了四种不同规模的模型，遵循Henighan等人[27]列出的超参数。单突变验证集上的结果。

|  |  |  |
| --- | --- | --- |
| MSA Subsample Strategy | Context size | | Spearman | |
| Diversity minimizing | 256 sequences | 0.255 |
| Random | 256 sequences |  |
| HHFilter | 256 sequences |  |
| Sequence reweighting | 256 sequences |  |
| MSA子样本策略 | 上下文大小 | | 斯皮尔曼 | |
| 多样性最小化 | 256个序列 | 0.255 |
| 随机 | 256个序列 |  |
| HHFilter | 256个序列 |  |
| 序列重加权 | 256个序列 |  |

Table 8: Subsampling strategies for MSA Transformer evaluated on the single-mutation validation set. Sequence reweighting performs best. When sampling methods are stochastic, 5 seeds are run and the mean and standard deviation is reported. With HHFilter, we run with the -diff M parameter and randomly subsample the output if more than sequences are returned. We use a coverage parameter of 75 and a sequence identity parameter of 99. Mean absolute Spearman across the single-mutation validation tasks is reported.

表8:在单突变验证集上评估的MSA Transformer子采样策略。序列重加权表现最佳。当采样方法是随机的时，运行5个种子并报告均值和标准差。使用HHFilter时，我们使用-diff M参数运行，并在返回超过 条序列时随机子采样输出。我们使用覆盖参数75和序列一致性参数99。报告了单突变验证任务的平均绝对Spearman 。

capture similar information and boost performance. Similarly, we experiment with tuning only the layer normalizations, as these have also been recently shown to enable transfer to new tasks. In both cases, we found no improvement to the average Spearman . We also assessed label smoothing and replacing the gap token with a mask token or a pad token finding no significant impact; for simplicity, we omit label smoothing and use the mask token for future experiments.

捕获相似信息并提升性能。同样，我们尝试仅调整层归一化，因为这些最近也被证明能够迁移到新任务。在这两种情况下，我们发现平均Spearman 没有改善。我们还评估了标签平滑和用掩码标记或填充标记替换间隙标记，发现没有显著影响；为简单起见，我们省略了标签平滑并在未来的实验中使用掩码标记。

Minimal models We also examine a set of minimal models, in which we freeze all parameters in the Transformer and learn a projection from the ESM-1v outputs onto a PSSM, taking the sum of the projection and the PSSM. We experiment with freezing the PSSM or allowing it to train. We did not see a change in Spearman of more than 0.01 .

最小模型我们还检查了一组最小模型，其中我们冻结了Transformer中的所有参数，并学习从ESM-1v输出到PSSM的投影，取投影和PSSM的总和。我们尝试冻结PSSM或允许其训练。我们没有看到Spearman 的变化超过0.01。

Spiked unsupervised fine-tuning Next, we examine a new strategy, which we call spiked fine-tuning. In spiked fine-tuning, we regularize the fine-tuning by continuing to spike pretraining sequences into the fine-tuning batch. In this setting, we train on the entire MSA, including non-consensus positions. We find that spiked fine-tuning with a small ratio (0.01) of MSA tokens to pre-training tokens performs best and enables training of all parameters without overfitting.

尖峰无监督微调接下来，我们研究了一种新策略，我们称之为尖峰微调。在尖峰微调中，我们通过继续将预训练序列尖峰到微调批次中来正则化微调。在这种情况下，我们在整个MSA上训练，包括非共识位置。我们发现，使用MSA标记与预训练标记的小比例(0.01)进行尖峰微调表现最佳，并且能够训练所有参数而不会过拟合。

The final models were trained for 7500 updates using spiked fine-tuning with a batch size of tokens. To produce an ensemble, we perform the fine-tuning scheme on five models that were pre-trained with different seeds. Each model was also fine-tuned with a unique seed.

最终模型使用尖峰微调训练了7500次更新，批次大小为 个标记。为了生成集成，我们在五个使用不同种子预训练的模型上执行微调方案。每个模型也使用唯一的种子进行微调。

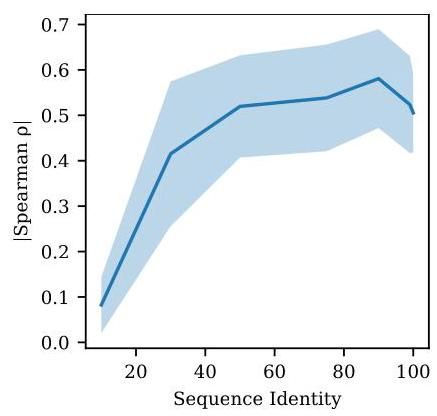


Figure 10: Filtering sequences with high sequence identity to the query improves performance. The curve illustrates mean standard deviation across the 9 validation proteins. HHFilter is used to filter the MSAs with coverage of 75 and various sequence identity values as shown on x-axis. After filtering,384 sequences are sampled for inference. Each sequence identity value refers to using sequences with no more than sequence identity to the seed sequence. The MSA Transformer appears to primarily use sequences that are close to the seed sequence, yet performance drops if sequences that are too similar remain in the MSA. Results are broken down across the single-mutation validation set in Table 9.

图10:过滤与查询具有高序列一致性的序列可以提高性能。该曲线显示了9个验证蛋白的平均 标准差。HHFilter用于过滤MSA，覆盖率为75，x轴上显示了各种序列一致性值。过滤后，采样384条序列进行推理。每个序列一致性值 指的是使用与种子序列的序列一致性不超过 的序列。MSA Transformer似乎主要使用接近种子序列的序列，但如果MSA中保留过于相似的序列，性能会下降。结果在表9中按单突变验证集进行了分解。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sequence Identity (%) | 10 | 30 | 50 | 75 | 90 | 99 | 100 |
| AMIE\_PSEAE\_Whitehead | 0.025 | 0.461 | 0.467 | 0.365 | 0.665 | 0.654 | 0.622 |
| BG505\_env\_Bloom2018 | 0.055 | 0.055 | 0.417 | 0.482 | 0.452 | 0.450 | 0.457 |
| BLAT\_ECOLX\_Ranganathan2015 | 0.060 | 0.630 | 0.745 | 0.776 | 0.795 | 0.662 | 0.478 |
| DLG4\_RAT\_Ranganathan2012 | 0.008 | 0.431 | 0.416 | 0.431 | 0.457 | 0.418 | 0.400 |
| GAL4\_YEAST\_Shendure2015 | 0.080 | 0.287 | 0.366 | 0.441 | 0.576 | 0.542 | 0.388 |
| SUMO1\_HUMAN\_Roth2017 | 0.131 | 0.430 | 0.516 | 0.541 | 0.500 | 0.492 | 0.495 |
| TIM\_SULSO\_b0 | 0.026 | 0.581 | 0.633 | 0.625 | 0.649 | 0.324 | 0.632 |
| UBC9\_HUMAN\_Roth2017 | 0.165 | 0.376 | 0.557 | 0.583 | 0.494 | 0.555 | 0.475 |
| KKA2\_KLEPN\_Mikkelsen2014 | 0.192 | 0.484 | 0.560 | 0.601 | 0.637 | 0.616 | 0.602 |
| 序列一致性(%) | 10 | 30 | 50 | 75 | 90 | 99 | 100 |
| AMIE\_PSEAE\_Whitehead | 0.025 | 0.461 | 0.467 | 0.365 | 0.665 | 0.654 | 0.622 |
| BG505\_env\_Bloom2018 | 0.055 | 0.055 | 0.417 | 0.482 | 0.452 | 0.450 | 0.457 |
| BLAT\_ECOLX\_Ranganathan2015 | 0.060 | 0.630 | 0.745 | 0.776 | 0.795 | 0.662 | 0.478 |
| DLG4\_RAT\_Ranganathan2012 | 0.008 | 0.431 | 0.416 | 0.431 | 0.457 | 0.418 | 0.400 |
| GAL4\_YEAST\_Shendure2015 | 0.080 | 0.287 | 0.366 | 0.441 | 0.576 | 0.542 | 0.388 |
| SUMO1\_HUMAN\_Roth2017 | 0.131 | 0.430 | 0.516 | 0.541 | 0.500 | 0.492 | 0.495 |
| TIM\_SULSO\_b0 | 0.026 | 0.581 | 0.633 | 0.625 | 0.649 | 0.324 | 0.632 |
| UBC9\_HUMAN\_Roth2017 | 0.165 | 0.376 | 0.557 | 0.583 | 0.494 | 0.555 | 0.475 |
| KKA2\_KLEPN\_Mikkelsen2014 | 0.192 | 0.484 | 0.560 | 0.601 | 0.637 | 0.616 | 0.602 |

Table 9: Filtering MSAs from the single-mutation validation set with HHFilter coverage 75 and various sequence identity values. Filtering sequences to an identity threshold of 75% or 90% consistently performs best. The Spearman rank correlation between MSA Transformer predictions and experimental data is shown for each deep mutational scan.

表9:使用HHFilter覆盖率为75和不同序列同一性值从单突变验证集中过滤MSA。将序列过滤到75%或90%的同一性阈值始终表现最佳。每个深度突变扫描中MSA Transformer预测与实验数据之间的Spearman等级相关性如图所示。

# C Datasets

# C 数据集

# C.1 Evaluation Tasks

# C.1 评估任务

We evaluate models on a set of 41 deep mutational scans collected by Riesselman et al. [20], which comprise a variety of tasks assessing a diverse set of proteins. Across tasks, the experimental data differ widely in the functions tested and in the experimental measurements performed. Of the 41 datasets, 37 are single-mutation only, 1 is double-mutation only, and the rest contain a variable number of mutations per sequence between 1 and 28 . The median number of mutations is 2979, and the average is 16822; the smallest dataset has 37 mutations, and the largest has 496137. We randomly select 9 single-mutation experiments as a validation set. We also ablate the multiple mutation scoring approach on the double mutations from the PABP Yeast deep mutational scan. We exclude the 10 tasks used for validation and ablations from the test set. These datasets are reported in results for the full set. While the original compilation has 43 datasets, we exclude the tRNA (which is not a protein) and the toxin-antitoxin complex (which comprises multiple proteins).

我们在Riesselman等人[20]收集的41个深度突变扫描上评估模型，这些扫描包括评估多种蛋白质的各种任务。在不同任务中，实验数据在测试的功能和进行的实验测量方面差异很大。在41个数据集中，37个仅为单突变，1个仅为双突变，其余每个序列的突变数量在1到28之间不等。突变的中位数为2979，平均数为16822；最小的数据集有37个突变，最大的有496137个突变。我们随机选择9个单突变实验作为验证集。我们还在PABP酵母深度突变扫描中的双突变上消融了多重突变评分方法。我们从测试集中排除了用于验证和消融的10个任务。这些数据集在完整集合的结果中报告。虽然原始汇编有43个数据集，但我们排除了tRNA(不是蛋白质)和毒素-抗毒素复合物(包含多种蛋白质)。

We treat each deep mutational scanning dataset as a separate prediction task, scoring each of the variants in the dataset with the model. We evaluate performance by comparing the scores with the experimental measurements using Spearman rank correlation. Results are broken out between the test set, which excludes the validation set, as well as the full set of 41 datasets. All ablations are performed on the single mutant validation set or the PABP Yeast doubles experiment. Only the final models are evaluated on the test set.

我们将每个深度突变扫描数据集视为一个单独的预测任务，用模型对数据集中的每个变体进行评分。我们通过使用Spearman等级相关性将评分与实验测量结果进行比较来评估性能。结果在排除验证集的测试集以及完整的41个数据集之间进行区分。所有消融均在单突变验证集或PABP酵母双突变实验上进行。只有最终模型在测试集上进行评估。

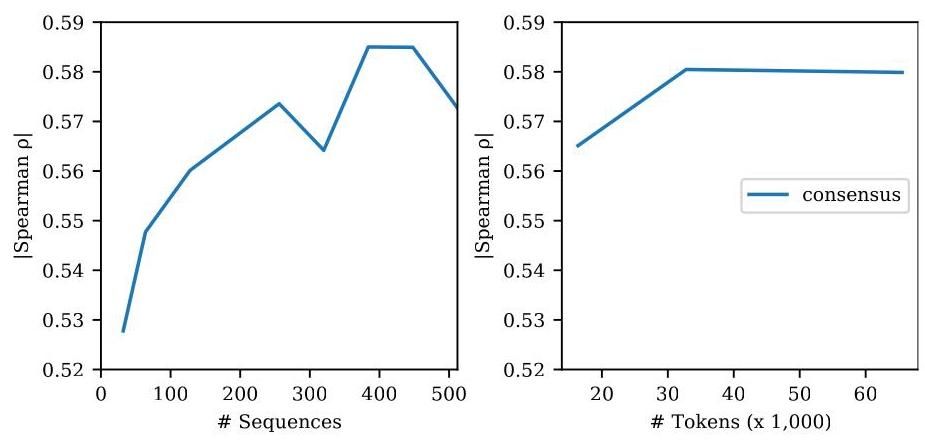


Figure 11: Few-shot performance of the MSA Transformer is robust to the number of sequences used for inference. Left: Varying the number of sequences used in inference. Right: Varying the number of tokens used for inference. Since the number of sequences in each MSA varies, we assess the effect of fixing the total number of tokens sampled from each MSA and drawing the corresponding number of sequences to fill the context. Results on single-mutation validation set.

图11:MSA Transformer的少样本性能对用于推理的序列数量具有鲁棒性。左图:改变用于推理的序列数量。右图:改变用于推理的标记数量。由于每个MSA中的序列数量不同，我们评估了固定从每个MSA中采样的总标记数量并绘制相应数量的序列以填充上下文的效果。单突变验证集上的结果。

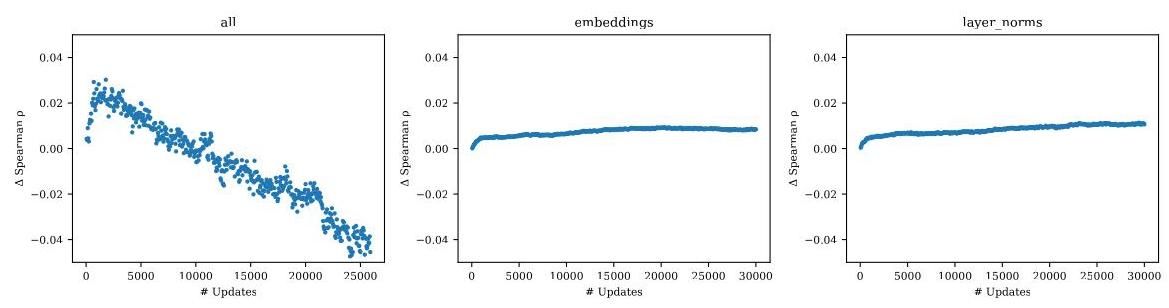


Figure 12: Unsupervised fine-tuning baselines. Mean change in Spearman across 9 models trained on the single-mutation validation set tasks. The title of each plot denotes the parameters that are trained. We find that fine-tuning the entire model results in overfitting, but limiting the training to just the embeddings or just the layer norms does not improve performance with respect to the pre-trained initialization. The choice of gap token and label smoothing has limited effect.

图12:无监督微调基线。在单突变验证集任务上训练的9个模型的Spearman 平均变化。每个图的标题表示训练的参数。我们发现微调整个模型会导致过拟合，但将训练限制在仅嵌入或仅层归一化并不会相对于预训练初始化提高性能。间隙标记和标签平滑的选择影响有限。

# C.2 Pre-training datasets

# C.2 预训练数据集

For the clustering sweep in Fig. 4, we use the Uniref50 and Uniref90 databases from the 2020\_03 release of Uniref [24], a publicly available database of proteins, clustered respectively to 50% and 90% sequence identity. For the 30% sequence identity dataset, we use Uniclust30 2020\_03 [89]. For the sequence identity dataset, we Uniref100 is hierarchically clustered to the , then 70% sequence identity level. MMseqs settings are those used by Uniref: 80% overlap with longest sequnece in the cluster, which translates to mmseqs-cluster -min-seq-id 90,70 -cov-mode 0 -alignment-mode 3 -c 0.8 . In order to compute pre-training perplexities on a heldout validation set, we randomly select of sequences each from Uniref30, Uniref50, and Uniref90. We then exclude sequences that are similar to the validation sequences by removing all sequences found with MMSeqs search (-min-seq-id 0.xx) for validation set xx. We use the most sensitive settings in MMSeqs -alignment-mode 3 -max-seqs , taking the train set as the query database and the validation set as the target database. We use settings -c 0.8 -cov-mode 0 to match the settings of Uniref. Pretraining perplexities on the validation sets are reported in Table 10. bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.450648; this version posted November 17, 2021. 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.

对于图4中的聚类扫描，我们使用了Uniref 2020\_03版本中的Uniref50和Uniref90数据库[24]，这是一个公开可用的蛋白质数据库，分别聚类到50%和90%的序列相似性。对于30%序列相似性的数据集，我们使用了Uniclust30 2020\_03 [89]。对于 序列相似性的数据集，我们将Uniref100分层聚类到 ，然后到70%的序列相似性水平。MMseqs的设置与Uniref相同:与聚类中最长序列的80%重叠，这对应于mmseqs-cluster -min-seq-id 90,70 -cov-mode 0 -alignment-mode 3 -c 0.8。为了在保留的验证集上计算预训练困惑度，我们随机从Uniref30、Uniref50和Uniref90中各选择 的序列。然后，我们通过移除所有在MMSeqs搜索(-min-seq-id 0.xx)中找到的与验证序列相似的序列来排除这些序列。我们使用MMSeqs中最敏感的设置-alignment-mode 3 -max-seqs ，将训练集作为查询数据库，验证集作为目标数据库。我们使用设置-c 0.8 -cov-mode 0以匹配Uniref的设置。验证集上的预训练困惑度在表10中报告。bioRxiv预印本doi: https://doi.org/10.1101/2021.07.09.450648; 此版本发布于2021年11月17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。该预印本在CC-BY-NC-ND 4.0国际许可下提供。

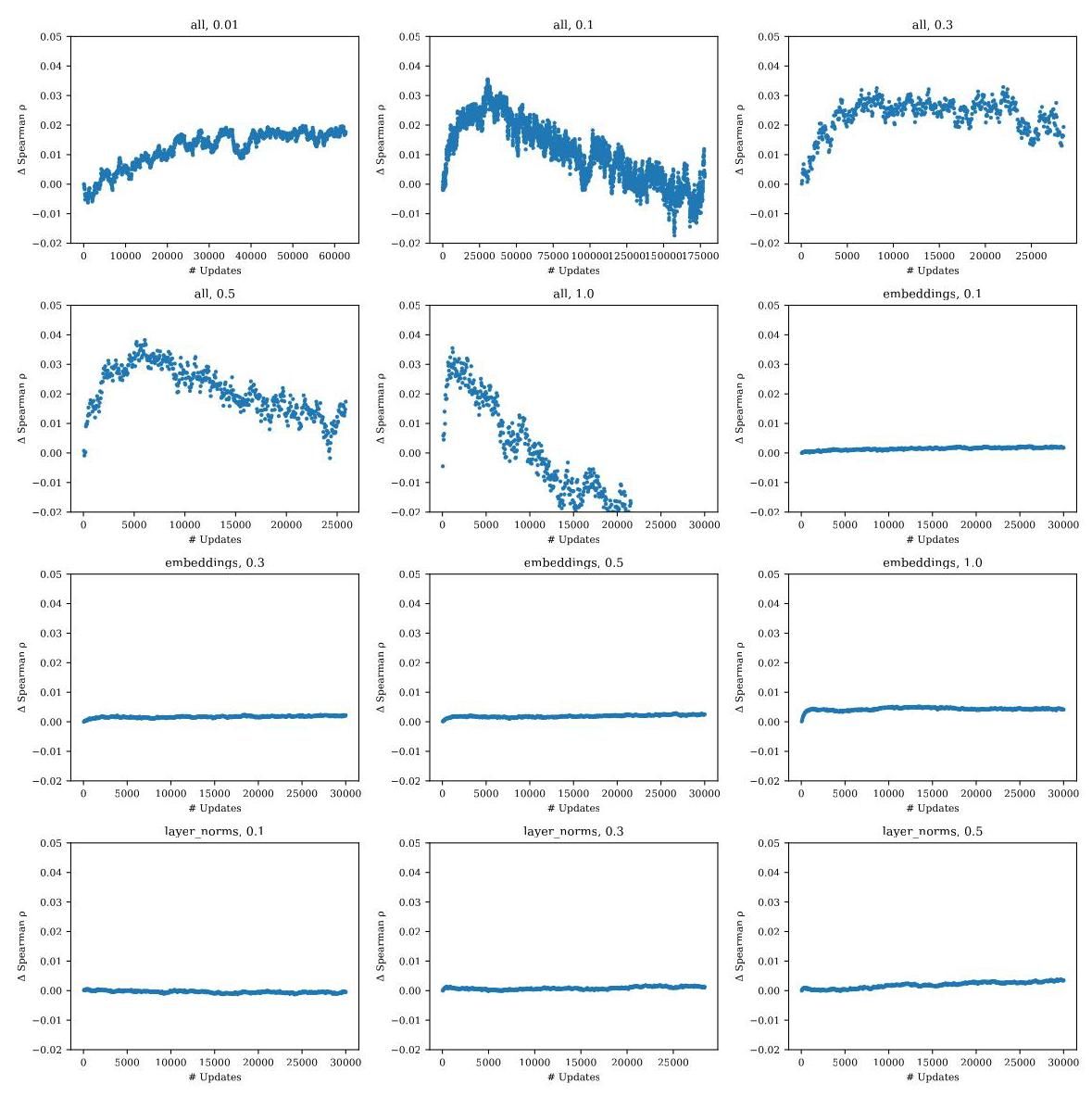


Figure 13: Spiked unsupervised fine-tuning. Mean change in Spearman across 9 models trained on the single-mutation validation tasks. The title of each plot denotes the parameters that are trained; and the ratio of MSA tokens to pre-training tokens. We find that a small ratio performs well and reduces the tendency for the model to overfit, while preserving strong performance. Performance is not improved if the fine-tuning is limited to just the embeddings or just the layer norms.

图13:尖峰无监督微调。在单突变验证任务上训练的9个模型的Spearman 平均变化。每个图的标题表示训练的参数；以及MSA标记与预训练标记的比例。我们发现，较小的比例表现良好，并减少了模型过拟合的趋势，同时保持了强大的性能。如果微调仅限于嵌入或仅限层归一化，性能不会提高。

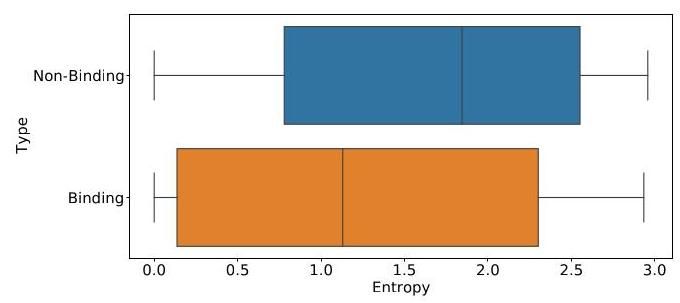


Figure 14: Box plot comparing entropy scores for binding vs non-binding positions in structures labeled in the Provis validation dataset (as described in Appendix B.4 of [36]). A Welch’s -test determines that the difference between the two means is statistically significant .

图14:箱线图比较了Provis验证数据集中标记的结构中结合位点与非结合位点的熵分数(如[36]附录B.4所述)。Welch的 检验确定两个均值之间的差异具有统计学意义 。

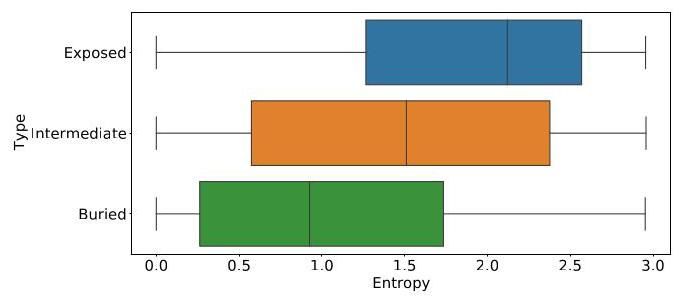


Figure 15: Box plot comparing entropy scores across residue depths in structures from the trRosetta dataset. Residue depths are categorized based on the number of neighboring residues with C-beta distance angstroms. (exposed , buried [88]). A one way Anova test determines that the differences between all three means are statistically significant .

图15:箱线图比较了trRosetta数据集中不同残基深度的熵分数。残基深度根据C-beta距离 埃的邻近残基数量进行分类(暴露 ，埋藏 [88])。单因素方差分析检验确定所有三个均值之间的差异具有统计学意义 。

# C.3 Baselines

# C.3 基线

The MSAs used for training DeepSequence and EVMutation are generated from the 2017-10 version of Uniref100, whereas the models we study are trained on sequences from Uniref90 2020- 03. In the case of MSA Transformer, the model is pre-trained on the 2018-03 Uniref, but we use 2020-03 MSAs for inference. In order to provide a fair comparison, we regenerate MSAs against the 2020-03 Uniref according to the methodology in Hopf et al. [4] and retrain EVMutation (replication) and DeepSequence (replication) on these datasets using their open-source codebases. For the viral proteins BF520\_env\_Bloom2018, BG505\_env\_Bloom2018, HG\_FLU\_Bloom2016, PA\_FLU\_Sun2015, POLG\_HCVJF\_Sun2014, POL\_HV1N5-CA\_Ndungu2014, we compute the sequence weights with (versus default ) following Riesselman et al. [20]. In the replication of the DeepSequence ensemble, for BF520\_env\_Bloom2018, BG505\_env\_Bloom2018, one of the five runs failed so we reran with a different random seed.

用于训练DeepSequence和EVMutation的多序列比对(MSA)是从2017-10版本的Uniref100生成的，而我们研究的模型是在2020-03版本的Uniref90序列上训练的。对于MSA Transformer，该模型是在2018-03版本的Uniref上进行预训练的，但我们在推理时使用2020-03版本的MSA。为了提供公平的比较，我们根据Hopf等人的方法[4]重新生成了针对2020-03版本Uniref的MSA，并使用其开源代码库在这些数据集上重新训练了EVMutation(复制)和DeepSequence(复制)。对于病毒蛋白BF520\_env\_Bloom2018、BG505\_env\_Bloom2018、HG\_FLU\_Bloom2016、PA\_FLU\_Sun2015、POLG\_HCVJF\_Sun2014、POL\_HV1N5-CA\_Ndungu2014，我们按照Riesselman等人的方法[20]计算了序列权重(使用 而非默认的 )。在DeepSequence集成的复制过程中，对于BF520\_env\_Bloom2018和BG505\_env\_Bloom2018，五次运行中有一次失败，因此我们使用不同的随机种子重新运行。

# C.4 Validation and test set

# C.4 验证集和测试集

The single-mutation validation set consists of the following deep mutational scans: AMIE\_PSEAE\_Whitehead, BG505\_env\_Bloom2018, BLAT\_ECOLX\_Ranganathan2015, BRCA1\_HUMAN\_RING, DLG4\_RAT\_Ranganathan2012, GAL4\_YEAST\_Shendure2015,

单突变验证集包括以下深度突变扫描:AMIE\_PSEAE\_Whitehead、BG505\_env\_Bloom2018、BLAT\_ECOLX\_Ranganathan2015、BRCA1\_HUMAN\_RING、DLG4\_RAT\_Ranganathan2012、GAL4\_YEAST\_Shendure2015、

POLG\_HCVJF\_Sun2014, SUM01\_HUMAN\_Roth2017, TIM\_SULSO\_b0, UBC9\_HUMAN\_Roth2017, KKA2\_KLEPN\_Mikkelsen2014.

POLG\_HCVJF\_Sun2014、SUM01\_HUMAN\_Roth2017、TIM\_SULSO\_b0、UBC9\_HUMAN\_Roth2017、KKA2\_KLEPN\_Mikkelsen2014。

For ablations studies with multiple mutations the following dataset is used: PABP\_YEAST\_Fields2013-doubles

对于包含多个突变的消融研究，使用以下数据集:PABP\_YEAST\_Fields2013-doubles

The test set consists of the following deep mutational scans: B3VI55\_LIPSTSTABLE, B3VI55\_LIPST\_Whitehead2015, BF520\_env\_Bloom2018, BG\_STRSQ\_hmmerbit, BLAT\_ECOLX\_Ostermeier2014, BLAT\_ECOLX\_Palzkil12012, BLAT\_ECOLX\_Tenaillon2013, BRCA1\_HUMAN\_BRCT, CALM1\_HUMAN\_Roth2017, HG\_FLU\_Bloom2016, HIS7\_YEAST\_Kondrashov2017, HSP82\_YEAST\_Bolon2016,

测试集包括以下深度突变扫描:B3VI55\_LIPSTSTABLE、B3VI55\_LIPST\_Whitehead2015、BF520\_env\_Bloom2018、BG\_STRSQ\_hmmerbit、BLAT\_ECOLX\_Ostermeier2014、BLAT\_ECOLX\_Palzkil12012、BLAT\_ECOLX\_Tenaillon2013、BRCA1\_HUMAN\_BRCT、CALM1\_HUMAN\_Roth2017、HG\_FLU\_Bloom2016、HIS7\_YEAST\_Kondrashov2017、HSP82\_YEAST\_Bolon2016、

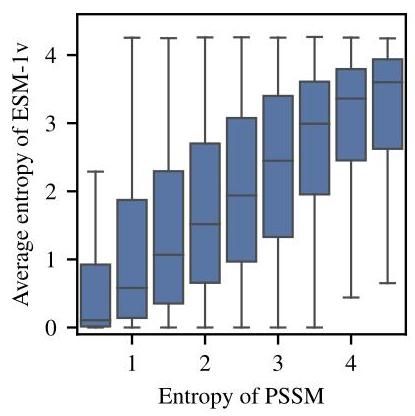


Figure 16: Entropy of PSSM versus ESM-1v predicted entropy on the trRosetta dataset. PSSM entropy determines the level of conservation at a given position in a protein family. ESM-1v entropy is well correlated with PSSM entropy (Pearson’s ), suggesting the model is able to identify conserved positions.

图16:PSSM的熵与ESM-1v在trRosetta数据集上预测的熵的比较。PSSM熵决定了蛋白质家族中给定位置的保守程度。ESM-1v的熵与PSSM熵高度相关(皮尔逊系数 )，表明该模型能够识别保守位置。

Ground Truth ESM (Uniref90) PSSM Hydrophobic Polar Charged Hydrophobic Polar Charge Hydrophobic Polar Charged

真实值 ESM (Uniref90) PSSM 疏水性 极性 带电 疏水性 极性 带电 疏水性 极性 带电

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Exposed | 0.36 | | 0.29 | | 0.35 |
| Intermediate | 0.55 | | 0.28 | | 0.17 |
| Buried | 0.69 | | 0.24 | | 0.068 |
| 暴露 | 0.36 | | 0.29 | | 0.35 |
| 中间 | 0.55 | | 0.28 | | 0.17 |
| 埋藏 | 0.69 | | 0.24 | | 0.068 |
| 0.38 | | 0.29 | | 0.33 | |
| 0.52 | | 0.28 | | 0.19 | |
| 0.67 | | 0.25 | | 0.088 | |
| 0.38 | | 0.29 | | 0.33 | |
| 0.52 | | 0.28 | | 0.19 | |
| 0.67 | | 0.25 | | 0.088 | |
| 0.39 | | 0.29 | | 0.32 | |
| 0.53 | | 0.28 | | 0.19 | |
| 0.67 | | 0.24 | | 0.086 | |
| 0.39 | | 0.29 | | 0.32 | |
| 0.53 | | 0.28 | | 0.19 | |
| 0.67 | | 0.24 | | 0.086 | |

Figure 17: Predicted distribution of hydrophobic, polar and charged amino acids at the surface and core of proteins in the trRosetta dataset. We compare to the actual proportion in the protein structure. We classify residues into buried, intermediate or exposed by residue depths based on the number of neighboring residues with C-beta distance angstroms (exposed , buried [88]. ESM-1v and PSSM both see increased hydrophobicity predictions for buried residues, in correspondence with the ground truth data. Predicted probabilities are produced by introducing a mask token at each position.

图17:trRosetta数据集中蛋白质表面和核心的疏水性、极性和带电氨基酸的预测分布。我们将其与蛋白质结构中的实际比例进行比较。我们根据相邻残基的C-beta距离 埃(暴露 ，埋藏 [88])将残基分类为埋藏、中间或暴露。ESM-1v和PSSM都看到埋藏残基的疏水性预测增加，与真实数据一致。预测概率是通过在每个位置引入掩码标记生成的。

IF1\_ECOLI\_Kishony, MKO1\_HUMAN\_Johannessen, MTH3\_HAEAESTABILIZED\_Tawfik2015, P84126\_THETH\_b0, PABP\_YEAST\_Fields2013-singles, PA\_FLU\_Sun2015,

IF1\_ECOLI\_Kishony, MKO1\_HUMAN\_Johannessen, MTH3\_HAEAESTABILIZED\_Tawfik2015, P84126\_THETH\_b0, PABP\_YEAST\_Fields2013-singles, PA\_FLU\_Sun2015,

POL\_HV1N5-CA\_Ndungu2014, PTEN\_HUMAN\_Fowler2018, RASH\_HUMAN\_Kuriyan,

POL\_HV1N5-CA\_Ndungu2014, PTEN\_HUMAN\_Fowler2018, RASH\_HUMAN\_Kuriyan,

RL401\_YEAST\_Bolon2013, RL401\_YEAST\_Bolon2014, RL401\_YEAST\_Fraser2016,

RL401\_YEAST\_Bolon2013, RL401\_YEAST\_Bolon2014, RL401\_YEAST\_Fraser2016,

TIM\_THEMA\_b0, TPK1\_HUMAN\_Roth2017, TPMT\_HUMAN\_Fowler2018,

TIM\_THEMA\_b0, TPK1\_HUMAN\_Roth2017, TPMT\_HUMAN\_Fowler2018,

UBE4B\_MOUSE\_Klevit2013-singles, YAP1\_HUMAN\_Fields2012-singles.

UBE4B\_MOUSE\_Klevit2013-singles, YAP1\_HUMAN\_Fields2012-singles.

# D Methodology

# D 方法论

# D.1 Model selection

# D.1 模型选择

ESM-1b and MSA Transformer model checkpoints are selected based on performance on the single mutation validation set. Open sourced checkpoints are used for ESM-1b and other protein language model baselines.

ESM-1b和MSA Transformer模型检查点是根据单突变验证集上的性能选择的。ESM-1b和其他蛋白质语言模型基线使用开源检查点。

# D.2 Treatment of synonymous mutations

# D.2 同义突变的处理

Synonymous mutations are mutations in DNA that do not change the protein sequence that is expressed. The deep mutational scanning datasets that we evaluate here can therefore include DNA mutations that do not change the protein sequence itself. Synonymous mutations are excluded from results.

同义突变是指DNA中不改变所表达蛋白质序列的突变。因此，我们在这里评估的深度突变扫描数据集可以包括不改变蛋白质序列本身的DNA突变。同义突变被排除在结果之外。

# D.3 Bootstraps

# D.3 自举法

To compute bootstraps for the pointplots, we randomly resample each deep mutational scan (with replacement) and compute the Spearman between the experimental data and model predictions.

为了计算点图的自举法，我们随机重新采样每个深度突变扫描(有替换)，并计算实验数据与模型预测之间的Spearman 。

# D.4 Average calibration error

# D.4 平均校准误差

The standard expected calibration error (ECE) performs poorly for highly imbalanced data [90]. Following Neumann et al. [90] and Nixon et al. [91] we adapt average calibration error for the multi-class setting as follows:

标准预期校准误差(ECE)对于高度不平衡的数据表现不佳[90]。根据Neumann等人[90]和Nixon等人[91]的研究，我们将多类设置的平均校准误差调整如下:

where is the number of classes, is the number of non-empty bins for class , and acc and conf are the accuracy and confidence for bin and class .

其中 是类别数， 是类别 的非空箱数，acc和conf是箱 和类别 的准确性和置信度。

# E Performance by MSA depth

# E 按MSA深度评估性能

We examine the relationship between the number of related sequences in the pre-training set and performance on the task. We use Jackhmmer [92] version 3.3.1 with a bitscore threshold of 27 and 8 iterations to construct MSAs from the ESM-1v training set. We do not observe a strong correlation between MSA depth and the observed absolute value of Spearman (Figure Fig. 19).

我们研究了预训练集中相关序列数量与任务性能之间的关系。我们使用Jackhmmer [92] 3.3.1版本，以27的比特分数阈值和8次迭代，从ESM-1v训练集中构建MSA。我们没有观察到MSA深度与Spearman 观察到的绝对值之间存在强相关性(见图Fig. 19)。

# F Compute costs

# F 计算成本

ESM-1v models are pre-trained for 6 days on 64 V100 GPUs. Weights for the MSA Transformer were retrieved from the open-source repository released by the authors; the model was pre-trained for 13 days on 128 V100 GPUs. Once trained, the models can be used directly for function prediction tasks. Forward inference is efficient, meaning that for applications of the models, the additional compute is minimal. In total, five ESM-1v models were trained on various Uniref clustering thresholds to five different levels: , and . For the sequence identity level, five total models with different random seeds were trained, for use in an ensemble. As illustrated in Fig. 7, inference is inexpensive by comparison. Batch inference was performed with preemptible, short (shorter than one hour), single V100 GPU jobs on a shared compute cluster.

ESM-1v模型在64个V100 GPU上预训练了6天。MSA Transformer的权重从作者发布的开源仓库中获取；该模型在128个V100 GPU上预训练了13天。一旦训练完成，这些模型可以直接用于功能预测任务。前向推理效率高，这意味着在应用这些模型时，额外的计算量是最小的。总共训练了五个ESM-1v模型，分别在不同的Uniref聚类阈值下训练到五个不同的水平: 和 。对于 序列同一性水平，训练了五个不同随机种子的模型，用于集成。如图7所示，相比之下，推理成本较低。批量推理是在共享计算集群上使用可抢占的、短时间(少于一小时)的单个V100 GPU作业进行的。

|  |  |  |  |
| --- | --- | --- | --- |
| Clustering | Valid (30%) | Valid (50%) | Valid (90%) |
| 30% | 8.93 | 8.33 | 7.29 |
| 50% | 8.90 | 7.77 | 6.27 |
| 70% | 9.05 | 7.80 | 5.85 |
| 90% | 9.37 | 8.10 | 5.56 |
| 100% | 9.89 | 8.65 | 6.05 |
| 聚类 | 有效(30%) | 有效(50%) | 有效(90%) |
| 30% | 8.93 | 8.33 | 7.29 |
| 50% | 8.90 | 7.77 | 6.27 |
| 70% | 9.05 | 7.80 | 5.85 |
| 90% | 9.37 | 8.10 | 5.56 |
| 100% | 9.89 | 8.65 | 6.05 |

Table 10: Perplexities on heldout pre-training validation sequences after training a parameter Transformer model for 170,000 updates on various sequence identity clusterings of Uniref. bioRxiv preprint doi: https://doi.org/10.1101/2021.07.09.450648; this version posted November 17, 2021. 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

表10:在Uniref的各种序列同一性聚类上训练一个 参数的Transformer模型进行170,000次更新后，在保留的预训练验证序列上的困惑度。bioRxiv预印本 doi: https://doi.org/10.1101/2021.07.09.450648；此版本发布于2021年11月17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。它已制作

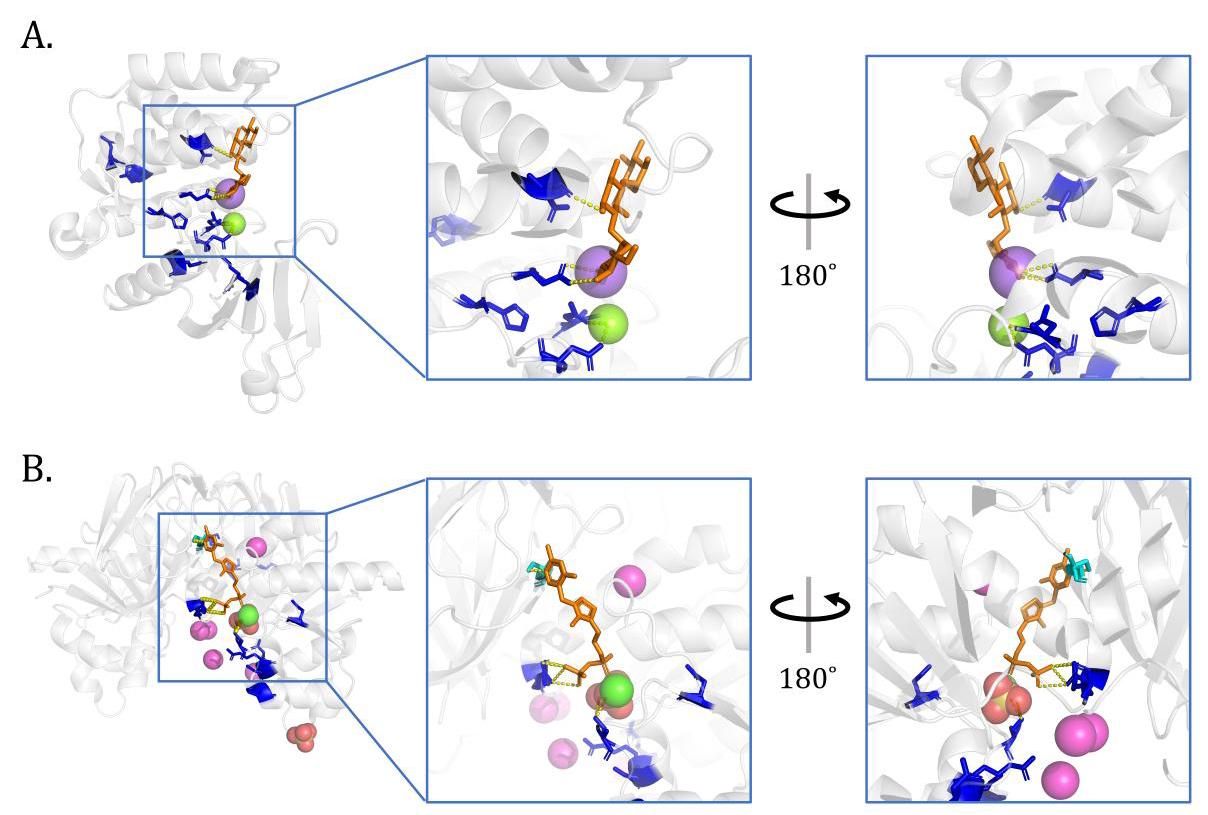


Figure 18: ESM-1v accurately captures functional properties. Further examples. The ten positions with lowest predicted entropy highlighted in blue. (A) Kanamycin kinase APH(3’)-II (pdbid: 1ND4 [93]). The highlighted residues interact with the kanamycin aminoglycoside, as well as the magnesium and sodium ions. (B) Thiamin pyrophosphokinase 1 (pdbid: 3S4Y). Residue 216 is one of the 10 lowest entropy residues, and we highlight it on the other chain (in cyan) to show both chains of the dimer interacting with the thiamine diphosphate.

图18:ESM-1v准确捕捉功能特性。更多示例。预测熵最低的十个位置以蓝色突出显示。(A) 卡那霉素激酶APH(3’)-II(pdbid: 1ND4 [93])。突出显示的残基与卡那霉素氨基糖苷以及镁和钠离子相互作用。(B) 硫胺素焦磷酸激酶1(pdbid: 3S4Y)。残基216是10个最低熵残基之一，我们在另一条链上(以青色)突出显示它，以显示二聚体的两条链与硫胺素二磷酸的相互作用。

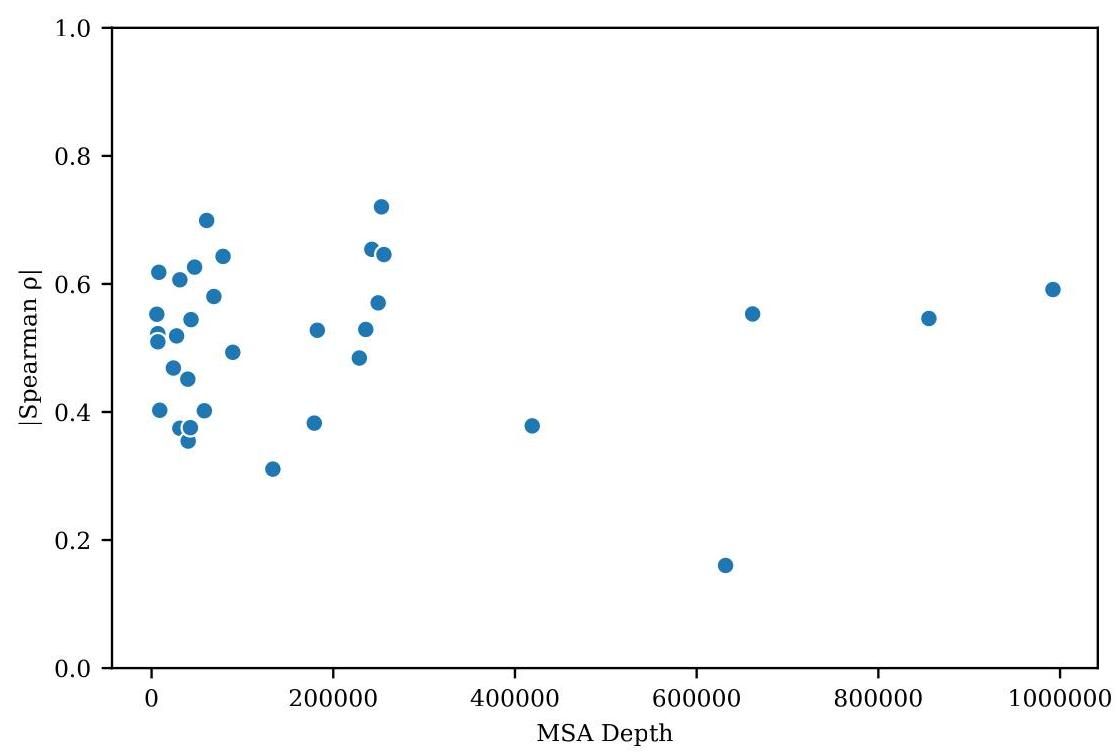


Figure 19: Relation between MSA depth and zero-shot performance of ESM-1v. We use JackHMMer [92] version 3.3.1 with a bitscore threshold of 27 and 8 iterations to construct MSAs from the ESM-1v training set. We do not observe a strong correlation between MSA depth and the observed ISpearman I.

图19:MSA深度与ESM-1v零样本性能之间的关系。我们使用JackHMMer [92] 3.3.1版本，以27的比特分数阈值和8次迭代从ESM-1v训练集构建MSA。我们没有观察到MSA深度与观察到的ISpearman I之间的强相关性。

1. Facebook AI Research New York University UC Berkeley. ESM-1v is available at <https://github.com/facebookresearch/esm>.Correspondence to: Alexander Rives <arives@fb.com>.

   Facebook AI 研究 纽约大学 加州大学伯克利分校。ESM-1v 可在 <https://github.com/facebookresearch/esm> 获取。通讯作者:亚历山大·里夫斯 <arives@fb.com>。 [↑](#footnote-ref-1)