### **应用传统机器学习和深度学习方法于蛋白质进化研究**

Essay 1:

Neural networks to learn protein sequence–function relationships from deep mutational scanning data.

背景：EVmutation和 DeepSequence等无监督学习方法在进化相关蛋白质序列的大比对上进行训练。这些方法可以模拟蛋白质家族的天然功能，但它们无法预测不受长期进化选择约束的特定蛋白质特性。我们新提出了一个有监督的深度学习框架，训练有监督的神经网络学习序列到功能的映射函数.

数据集：由数千到数百万个蛋白质序列变体组成，每个变体具有一个用于量化其在高通量检测中的活性或适应性的分数。数据集被分为81%训练集，9%调优集（用于优化超参），10%验证集

特征工程： 我们对蛋白质序列进行编码，编码信息包含了每个氨基酸在每个位置的物理和化学性质。One hot编码用于标记特定位置的氨基酸，还采用AAindex编码捕获氨基酸的物理性质和化学性质，运用PCA降到19维，最后将两个编码连接。在图卷积神经网络中，我们通过计算蛋白质的三维结构得到残基距离矩阵，对距离进行阈值处理,并将得到的矩阵转化为无向图，得到结构图。

团队共测试了线性回归和全连接、数列卷积、图卷积三种神经网络。卷积神经网络中，卷积层可以学习识别β链中常见的极性和非极性氨基酸，使网络能评估整个输入序列的β链倾向，并将这些信息与蛋白质功能关联。

应用：

1） 运用可视化，将网络中的连接层可视化成蛋白质的三维结构，并根据其得分给蛋白质着色，从而了解哪些序列位置对蛋白质的影响最大，了解哪些区域是突变不耐受的。

2 ）在野生型蛋白基础上设计新蛋白质。运用训练的模型对那些相对野生型仅有局部变化的蛋白质进行评估，运用线性回归和三种神经网络四个模型联合打分,这里采用随机重启爬山算法，最大化最低预测功能分数

讨论：

因为突变常以加法形式结合，所以线性模型的表现尚可，但无法表示突变之间的相互作用，而卷积网络可以泛化不同位置之间的突变影响来提高性能，图神经网络并没有比基于序列的神经网络表现更佳,可能是因为缺乏数据多样性等原因。

Essay 2:

Learning the local landscape of protein structures with convolutional neural networks.

适应性图谱将蛋白质势能描述为构象坐标的函数，用于研究进化、识别具有新的有用特性的蛋白质或量化可变性。另一方面，蛋白质晶体是蛋白质的稳定结构，通常是能量地形中的最低点，并在种群中占主导地位。

AlphaFold基本实现将蛋白质序列放置在正确的能量最小值中，但缺乏将深度学习应用到相反的问题：结构如何限制给定位点允许的蛋白质，这将限制有害突变的采样，加速靶向诱变和蛋白质工程工作。我们使用3DCNN对此进行研究，研究了从局部化学环境（微环境）预测隐蔽残基的CNN模型能否良好预测野生型残基和进化上分化的同源物中的残基，并关注那些高准确率预测的氨基酸分布。

具体方法：

网络构架共九层：六层特征提取块：两对3D卷积层,每对后是一个降维最大池化层,分别采用3\*3\*3和2\*2\*2 的滤波器，采用Relu函数应用到每一层的输出，最终生成的特征图的维度是400\*3\*3\*3,展平后进入三层分类块。分类块由三个全连接层构成，前两个连接层应用relu函数，第三层用softmax函数获得20个氨基酸的概率分数的向量.

对数据集的处理：

我们在蛋白质数据集中随机取样残基，以创建一个反映每种氨基酸自然丰度的微环境数据集.对每个蛋白质链，我们最多选取50个残基，并且选取的残基数不超过总数的50%，以免模型偏向那些分子量大的蛋白质。最终的微环境数据集由1455978个微环境构成，拆分成90%的训练集和10%的验证集。

利用这些元数据，我们生成了微环境的体素化表示（4D 张量），该表示以α指定残基的碳为中心，并相对于主链定向，使得侧链沿+z轴定向。体素化表示的分辨率为 1 Å，由 3D 空间 （x、y 、z ） 和 7 个辅助通道组成。辅助通道编码有关体素中存在的原子性质（C、H、O、N、S）以及部分电荷和溶剂可及表面积的信息。

CNN中共训练了5个epoch，初始学习率为0.05,每2000个批次的验证准确率如果没有增加0.1%，则学习率减半。

结果：

在包含130个结构的独立数据集PSICOV数据集上评估,野生型序列的预测准确率通常很高，平均达到60%.将生化相似的氨基酸分组后，我们预测氨基酸组别的能力高于预测特定氨基酸的能力，达到71%。

接下来，我们询问了该网络在多序列比对 （MSA） 中预测位点的共有氨基酸的能力。这种预测的可能程度取决于给定位点周围的微环境在同源结构中的保守程度。有40%的概率预测正确的共有氨基酸，55%的概率预测正确的类别.

氨基酸的预测正确率与同源物之间的序列差异有很大关系，对于与野生型分歧较大（相似度低于20%）的蛋白质，正确率低于30%.

讨论：

我们的研究结果强化了氨基酸与其局部生化环境有着密切关联的观点，并且作用在蛋白质中氨基酸上的限制往往会随着蛋白质的进化而随着时间的推移而变化。模型的错误预测可能为蛋白质工程提供了机会，因为它会指出为突变做好准备的位点。具体来说，我们预计，每当网络自信地预测与当前野生类型不同的氨基酸时，该位点就会为突变做好准备，并且特定错误预测的氨基酸是稳定或功能获得性突变的良好候选者。

Essay 3:

Efficiently predicting protein stability changes upon single-point mutation with large language models.

这份文章详细描述了一种新的方法，利用大型语言模型（LLM）特别是ESM（Evolutionary Scale Modeling）模型来高效预测蛋白质在单点突变后的稳定性变化。

背景：多年来，预测由单点突变引起的蛋白质稳定性变化一直是一个难题，吸引了众多研究人员的兴趣。准确预测蛋白质热稳定性对于药物开发、蛋白质进化分析和酶合成等生物化学领域的应用至关重要。

评估蛋白质的热稳定性：用ΔΔG=ΔG\_M-ΔG\_W (M:mutated;W:wild)评估,较高的ΔΔG意味着更高的热稳定性

基于机器学习的蛋白质热稳定性预测可以分为两类：（1）基于序列的方法，从蛋白质序列中提取有效特征；（2）基于结构的方法，提取结构方差以预测ΔΔG

方法：

数据集：仔细对齐序列和结构，后者表示每个氨基酸内部骨架原子的坐标.再从PROSTATA中得出我们的初始数据集，包含野生型和突变型序列，和相应的ΔΔG值,并且设置了TM-score阈值为0.5（以防出现相似的蛋白质），共10544个样本

模型：

模型接收野生型和突变型的序列-结构对作为输入。ESM辅助连续图卷积网络独立应用于每个输入(即野生型和突变型蛋白质)，将序列和结构特征整合到一个连贯的低维特征向量中。然后，二者相减以捕捉有突变引起的特征变化，再通过非线性分类器转换为一维输出，即ΔΔG值，从而预测蛋白质的稳定性变化

讨论：

该方法在拥有较高准确率的同时，处理方法是现有方法的15倍

作者指出需要进一步改进序列选择算法以及进行更多生物化学方面的研究以提高模型的鲁棒性和互操作性(interoperability)。

Essay 4:

Rapid protein stability prediction using deep learning representations.

介绍：

蛋白质语言模型PLM通过训练在综合蛋白序列数据库中完成掩蔽氨基酸来学习蛋白质多样性的功能,从而生成突出的生物学特征.但由于进化限制和有限的数据限制，PLM无法再零样本学习的环境中显著提高蛋白质活性,无法推广到新的环境.

这里我们提出一种新的蛋白质进化模型EVOLVEpro,通过小样本学习和最少的实验测试来进化高活性蛋白质变体,从而仅从序列中预测高活性突变体.

关于EVOLVEpro的模型开发：

这是一个基于深度学习的定向进化框架,涉及：

1. 一个蛋白质语言模型，将蛋白质序列编码至一个潜在的连续空间中以促进其活性优化

2）一个顶层回归模型，用于研究从潜在空间中的几个数据点到蛋白质活动之间的映射.

我们采用主动学习，使用回归模型根据预测的符合度对蛋白质序列进行排名，从中选出排名较前的序列进行实验验证.该循环迭代执行以进化确定的蛋白质活性。

以ESM-2蛋白质语言模型为例，我们的流程如下：

选择一组随机的第一轮变体,采用随机森林回归判别模型预测蛋白质活性,使用残基合并平均嵌入,在每轮进化中选取前N个选择策略。最后，通过捕捉EVOLVEpro在进化过程中的残基位点偏好,可以了解哪些位置对提高活性有利

模型的应用：

抗体优化,进化微型DNA引导的CRISPR核酸酶,改进设计的prime编辑器,进行Bxb1整合酶进化,进化T7RNA聚合酶

总结：

EVOLVEpro学习蛋白质活性的一般规则，只要几个进化周期便可以生成高活性的蛋白质突变体.PLM生成的潜在空间和顶层模块中存在强大的功能选择，它们对蛋白质的表示优于传统的编码方法，如OneHot编码和整数编码.

PLM学习跨进化多样性的掩蔽序列重建任务，其学习的适应度地形通常不与蛋白质活性的适应性地形相关，因为它排除了典型的定向进化策略，使EVOLVEpro可以从上万个可能的序列中选择高活性的单个突变体，并从数千亿个多突变体中选择高活性蛋白。

EVOLVEpro有以下优点：

1） 成功率高

2） 不需要蛋白质的特殊知识

3） 可用于多目标优化

4） 高度模块化,允许将任何可量化的的属性作为输入而无需微调

Essay 5:

Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation.

蛋白质稳定性极易收到外界环境和基因突变的扰动，即使是微小的扰动，如单点突变，也可以使得活性蛋白走向无功能、错误折叠或者聚合的形式。我们训练了一个自监督模型Pythia，用于零样本预测蛋白质突变后的自由能变化。

模型构建：

在假设蛋白质在未折叠状态下的能量在很大程度上不受突变影响的前提下，通过公式推导，我们可以得到自由能变化与野生型、突变型蛋白质氨基酸在所有Rotamer构象中的概率之和的正比关系：

\Delta\Delta G \propto -\ln \frac{P\_{\text{MUT}}}{P\_{\text{WT}}

因此，我们使用一种图神经网络架构，以蛋白质的局部结构作为输入，该结构被表示为一个k近邻图（k-NN graph），节点代表氨基酸残基，边则根据Cα原子之间的欧几里得距离来定义。每个节点的特征包括氨基酸类型以及三个二面角（φ, ψ, ω）,边特征包含主链原子间的距离、序列位置信息及链信息。

Pythia使用了消息传递神经网络（MPNN），使用基于注意力的消息传递和读出功能进行定制。在注意力消息传递层（AMPL）的每一层中，顶点表示使用 Attention 块进行更新，并与 edge 连接表示来获得消息表示。

模型的训练任务是预测中心节点正确的氨基酸类型。通过这种方式，Pythia能够解码给定蛋白质中残基之间的内在模式，从而精确地预测突变的影响。

Pythia模型的可解释性：

由于 Pythia 采用了注意力机制，我们可以利用该模型用于研究它是否确实成功地捕获了蛋白质内复杂的交互。结果显示 Pythia 对突变对的注意力得分较高，这表明 Pythia 对突变体结构的敏感性，并且可以有效地捕捉突变残基和周围环境之间的重要关系。

模型评估：

Pythia在预测相关性和计算速度方面都超过了现有的深度学习方法和其他传统方法,相比传统的基于力场的方法，Pythia实现了105倍的加速。

Pythia成功应用于预测柠檬烯环氧化物水解酶（LEH）的有效热稳定突变，并展示了其在探索2600万高质量蛋白质结构上的潜力。

Applying traditional machine learning and deep learning methods to the study of protein evolution

Essay 1

Neural networks to learn protein sequence-function relationships from deep mutational scanning data.

Background: Unsupervised learning methods such as EVmutation and DeepSequence train on large comparisons of evolutionarily relevant protein sequences. These methods can model the natural functions of protein families, but they cannot predict specific protein properties that are not constrained by long-term evolutionary selection. We newly propose a supervised deep learning framework that trains supervised neural networks to learn sequence-to-function mapping functions.

Dataset: consists of thousands to millions of protein sequence variants, each with a score used to quantify its activity or fitness in high-throughput assays. The dataset was divided into 81% training set, 9% tuning set (for optimization of hyper-references), and 10% validation set

Feature Engineering: We encode the protein sequence and the encoded information contains the physical and chemical properties of each amino acid at each position.One hot encoding is used to label the amino acid at a specific position and also AAindex encoding is used to capture the physical and chemical properties of the amino acid and PCA is applied to reduce it to 19 dimensions and finally the two encodings are connected. In Graph Convolutional Neural Network we get the residue distance matrix by calculating the 3D structure of the protein, thresholding the distances,and transforming the obtained matrix into an undirected graph to get the structure map.

The team tested a total of three neural networks: linear regression and fully connected, array convolution, and graph convolution. In a convolutional neural network, the convolutional layer learns to recognize common polar and nonpolar amino acids in the beta chain, allowing the network to assess the beta chain tendencies of the entire input sequence and correlate this information with protein function.

Applications:

1) Using visualization, visualize the connecting layers in the network as a 3D structure of the protein and color the protein according to its score, so as to understand which sequence positions have the greatest impact on the protein and to understand which regions are mutation-intolerant.

2 ) Designing new proteins based on wild-type proteins. The proteins with only localized changes relative to the wild type were evaluated using the trained model, and scored using a combination of four models, linear regression and three neural networks, where a random restart hill-climbing algorithm was used to maximize the minimum predictive function score.

Discussion:

Since mutations are often combined in an additive form, the linear model performs moderately well, but is unable to represent the interactions between mutations, whereas the convolutional network can generalize the effects of mutations between different locations to improve performance, and the graphical neural network does not outperform the sequence-based neural network, possibly due to a lack of data diversity, among other reasons.

Essay 2.

Learning the local landscape of protein structures with convolutional neural networks.

Adaptive mapping describes protein potentials as a function of conformational coordinates and is used to study evolution, identify proteins with new useful properties, or quantify variability. On the other hand, protein crystals are stable structures of proteins that are usually the lowest point in the energy terrain and dominate the population.

The basic AlphaFold implementation places protein sequences in the correct energy minima, but lacks the ability to apply deep learning to the opposite problem: how the structure restricts the proteins allowed for a given locus, which would limit the sampling of deleterious mutations and accelerate target mutagenesis and protein engineering efforts. We investigated this using 3DCNN, examining whether CNN models that predict hidden residues from local chemical environments (microenvironments) can predict residues well in wild-type residues and evolutionarily diverged homologs, and focusing on those amino acid distributions that are predicted with high accuracy.

Specific methods:

The network architecture has nine layers: six-layer feature extraction block: two pairs of 3D convolutional layers, each pair is followed by a dimension reduction maximum pooling layer, respectively, using 3\*3\*3 and 2\*2\*2 filters, using the Relu function applied to the output of each layer, the final dimension of the generated feature maps is 400\*3\*3\*3, and the spreading into the three-layer classification block. The classification block consists of three fully connected layers, the first two connected layers apply the relu function, and the third layer uses the softmax function to obtain a vector of probability scores for 20 amino acids.

Processing of the dataset:

We randomly sampled residues in the protein dataset to create a microenvironmental dataset that reflects the natural abundance of each amino acid. For each protein chain, we selected up to 50 residues and selected no more than 50% of the total number of residues to avoid biasing the model toward proteins with high molecular weights. The final microenvironment dataset consists of 1455978 microenvironments, split into 90% training set and 10% validation set.

Using these metadata, we generated a voxelized representation (4D tensor) of the microenvironment centered on the carbon of the α-specified residue and oriented with respect to the main chain such that the side chains are oriented along the +z axis. The voxelized representation has a resolution of 1 Å and consists of 3D space (x, y, z) and seven auxiliary channels. The auxiliary channels encode information about the nature of the atoms present in the voxel (C, H, O, N, S) as well as the partial charge and solvent accessible surface area.

A total of 5 epochs were trained in the CNN with an initial learning rate of 0.05,and the learning rate was halved if the validation accuracy did not increase by 0.1% for every 2000 batches.

Results:

Evaluated on the PSICOV dataset, an independent dataset containing 130 structures, the prediction accuracy of wild-type sequences was generally high, averaging 60%. After grouping biochemically similar amino acids, our ability to predict amino acid groupings was higher than our ability to predict specific amino acids, at 71%.

Next, we asked about the network's ability to predict the shared amino acids of a locus in a multiple sequence alignment (MSA). The degree of likelihood of this prediction depends on how well the microenvironment surrounding the given locus is conserved in the homologous structure. There is a 40% probability of predicting the correct shared amino acid, and a 55% probability of predicting the correct class ...

The percentage of correct amino acid predictions is highly dependent on sequence differences between homologs, and is less than 30% for proteins with large divergence from the wild type (less than 20% similarity).

Discussion:

Our results reinforce the idea that amino acids are closely linked to their local biochemical environment and that restrictions acting on amino acids in proteins tend to change over time as proteins evolve. The model's mispredictions may provide an opportunity for protein engineering, as it will point to sites that are ready for mutation. Specifically, we expect that whenever the network confidently predicts an amino acid that is different from the current wild type, that site will be ready for mutation, and that the specific incorrectly predicted amino acid is a good candidate for a stabilizing or function-acquiring mutation.

Essay 3.

Efficiently predicting protein stability changes upon single-point mutation with large language models.

This article describes in detail a new approach that utilizes Large Language Models and in particular ESM (Evolutionary Scale Modeling) models to efficiently predict changes in the stability of proteins after a single point mutation.

Background: Predicting changes in protein stability induced by single point mutations has been a challenge that has attracted the interest of many researchers for many years. Accurate prediction of protein thermal stability is crucial for applications in biochemistry such as drug development, protein evolution analysis and enzyme synthesis.

Assessment of protein thermal stability: assessed by

\Delta\Delta G = \Delta G\_{\text{M}} - \Delta G\_{\text{W}} (M: mutated; W: wild)

higher ΔΔG means higher thermal stability.Machine learning-based predictions of protein thermal stability can be divided into two categories: (1) sequence-based methods, which extract effective features from protein sequences, and (2) structure-based methods, which extract structural variance to predict ΔΔG

Methods:

Dataset: Sequences and structures were carefully aligned, with the latter representing the coordinates of the backbone atoms within each amino acid. Our initial dataset was then derived from PROSTATA, containing wild-type and mutant sequences, and corresponding ΔΔG values, with a TM-score threshold of 0.5 (to prevent the appearance of similar proteins), for a total of 10,544 samples.

Model:

The model receives wild-type and mutant sequence-structure pairs as inputs. an ESM-assisted continuous graph convolutional network is applied independently to each input (i.e., wild-type and mutant proteins), which integrates the sequence and structural features into a concatenated low-dimensional feature vector. The two are then subtracted to capture feature changes caused by the presence of mutations, which are then converted to a one-dimensional output, i.e., ΔΔG values, by a nonlinear classifier to predict changes in the stability of the protein

Discussion:

The method has a higher accuracy while processing 15 times more than existing methods

The authors point to the need for further refinement of the sequence selection algorithm as well as more biochemical studies to improve the robustness and interoperability of the model.

Essay 4.

Rapid protein stability prediction using deep learning representations.

Presentation:

The Protein Language Model (PLM) learns the function of protein diversity by training to accomplish masked amino acids in a comprehensive protein sequence database, thereby generating salient biological features. However, due to evolutionary constraints and limited data limitations, PLM is unable to significantly improve protein activity in zero-sample learning environments and cannot be generalized to new environments.

Here we propose a new protein evolution model, EVOLVEpro, to evolve highly active protein variants through small-sample learning and minimal experimental testing, thus predicting highly active mutants from sequences only.

About EVOLVEpro's model development:

This is a deep learning based framework for directed evolution that involves:

1. A protein language model that encodes protein sequences into a potentially continuous space to facilitate optimization of their activity

2) A top-level regression model for studying the mapping between several data points from the latent space to protein activity.

We employ active learning using a regression model to rank protein sequences based on predicted compliance, from which the top-ranked sequences are selected for experimental validation . The loop is executed iteratively to evolve the identified protein activities.

Taking the ESM-2 protein language model as an example, our process is as follows:

A random set of first round variants were selected, and a random forest regression discriminant model was used to predict protein activity, using residue merge average embedding to select the top N selection strategies in each round of evolution. Finally, by capturing the residue site preferences of EVOLVEpro during evolution, it is possible to understand which positions are favorable for increasing activity

Application of the model:

Antibody optimization, Evolution of micro-DNA-guided CRISPR nuclease, Improved design of prime editor for Bxb1 integrase evolution, Evolution of T7RNA polymerase

Summary:

EVOLVEpro learns the general rules of protein activity and can generate highly active protein mutants in just a few evolutionary cycles.There is a powerful selection of functions in the potential space and top-level modules generated by PLM that represent proteins better than traditional coding methods such as OneHot coding and integer coding.

PLM learns masked sequence reconstruction tasks across evolutionary diversity, and the fitness terrain of its learning does not typically correlate with the fitness terrain of protein activity because it precludes typical directed evolution strategies, allowing EVOLVEpro to select highly active single mutants from tens of thousands of possible sequences and highly active proteins from hundreds of billions of multi-mutants.

EVOLVEpro has the following advantages:

1) High success rate

2) No special knowledge of proteins is required

3) Can be used for multi-objective optimization

4) Highly modular, allowing any quantifiable property to be used as input without fine-tuning.

Essay 5.

Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation.

Protein stability is highly susceptible to perturbations from the external environment and genetic mutations, and even small perturbations, such as single point mutations, can move active proteins towards non-functional, misfolded, or aggregated forms. We trained a self-supervised model, Pythia, for zero-sample prediction of free energy changes after protein mutations.

Model building:

Under the assumption that the energy of the protein in its unfolded state is largely unaffected by mutation, the derivation of the formula allows us to obtain a positive relationship between the free energy change and the sum of the probabilities that the amino acids of the wild-type and mutant proteins are in all Rotamer conformations:

\Delta\Delta G \propto -\ln \frac{P\_{\text{MUT}}}}{P\_{\text{WT}}

Therefore, we use a graph neural network architecture that takes as input the local structure of a protein, which is represented as a k-nearest neighbor graph (k-NN graph), where nodes represent amino acid residues and edges are defined based on Euclidean distances between Cα atoms. The features of each node include the amino acid type as well as three dihedral angles (φ, ψ, ω), and the edge features contain distances between main chain atoms, sequence position information, and chain information.

Pythia uses a message-passing neural network (MPNN) that is customized using Attention-based message passing and readout functions. In each layer of the Attention Message Passing Layer (AMPL), vertex representations are updated using Attention blocks and connected to edge connection representations to obtain message representations.

The training task of the model is to predict the correct amino acid type for the center node. In this way, Pythia is able to decode the intrinsic patterns between residues in a given protein and thus accurately predict the effects of mutations.

Interpretability of Pythia models:

Since Pythia employs an attention mechanism, we can utilize this model for investigating whether it is indeed successful in capturing complex interactions within proteins. The results show that Pythia has a high attention score for mutant pairs, indicating that Pythia is sensitive to the structure of the mutant and can effectively capture important relationships between mutant residues and their surroundings.

Model Evaluation:

Pythia outperforms existing deep learning methods and other traditional methods in terms of predictive relevance and computational speed, and achieves a 105-fold speedup over traditional force-field based methods.

Pythia was successfully applied to predict effective heat-stable mutations in limonene epoxide hydrolase (LEH) and demonstrated its potential for exploring the structure of 26 million high-quality proteins.