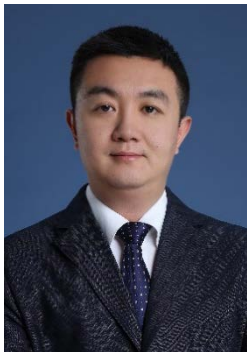
• AI 驱动底层技术 •



洪亮 上海交通大学自然科学研究院/物理与天文学院/药学院/张江高等研究院特聘教授，上海交通大学张江高研院人工智能生物医药中心主任。从事计算、人工智能和实验相结合的方式进行分子生物物理和蛋白质设计研究。2016 年入选国家高层次人才青年专家，2021 年入选教育部长江学者。在 *Nature*、*Proceedings of the National Academy of Sciences of the United States of America*、*Nature Physics* 等期刊上发表 SCI 论文 70 余篇。参与并主导开发了多个创新算法来提升功能蛋白和小分子药物的研发效率。

基于蛋白质语言模型的突变效应预测研究进展

张良[1]，谈攀[2,3]，洪亮[1,2,3,4*]

1 上海交通大学 物理与天文学院，上海 200240
2 上海应用数学中心&上海交通大学自然科学研究院，上海 200240
3 上海人工智能实验室，上海 200232
4 上海交通大学张江高等研究院，上海 201203

张良，谈攀，洪亮. 基于蛋白质语言模型的突变效应预测研究进展[J]. 生物工程学报, 2025, 41(3): 934-948.
ZHANG Liang, TAN Pan, HONG Liang. Research progress in mutation effect prediction based on protein language models[J]. Chinese Journal of Biotechnology, 2025, 41(3): 934-948.

摘 要：蛋白质突变效应预测是生物信息学和蛋白质工程领域的一个关键挑战。近年来，深度学习，特别是蛋白质语言模型的发展为该领域带来了新的机遇。本文综述了蛋白质语言模型在蛋白质突变效应预测中的应用，重点讨论了 3 类主要模型：基于序列的模型、基于结构的模型以及结合序列和结构信息的模型，详细分析了这些模型的原理、优势和局限性，并探讨了无监督学习和监督学习在模型训练中的应用。此外，还讨论了当前面临的主要挑战，包括高质量数据集的获取、数据噪声的处理等。最后，展望了未来研究方向，包括多模态融合、少样本学习等新兴技术的应用前景。本综述为研究者提供了一个全面的视角，以推动蛋白质突变效应预测领域的进一步发展。

关键词：蛋白质语言模型；突变效应预测；深度学习；序列模型；结构模型；多模态融合；无监督学习；监督学习

资助项目：国家自然科学基金(12104295)
This work was supported by the National Natural Science Foundation of China (12104295).
*Corresponding author. E-mail: hongl3liang@sjtu.edu.cn
Received: 2024-08-25; Accepted: 2025-02-13; Published online: 2025-02-14

Research progress in mutation effect prediction based on protein language models

ZHANG Liang[1], TAN Pan[2,3], HONG Liang[1,2,3,4*]

1 School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China
2 Shanghai National Center for Applied Mathematics (SJTU Center) & Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China
3 Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
4 Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 201203, China

Abstract: Predicting protein mutation effects is a key challenge in bioinformatics and protein engineering. Recent advancements in deep learning, particularly the development of protein language models (PLMs), have brought new opportunities to this field. This review summarizes the application of PLMs in predicting protein mutation effects, focusing on three main types of models: sequence-based models, structure-based models, and models that combine sequence and structural information. We analyze in detail the principles, advantages, and limitations of these models and discuss the application of unsupervised and supervised learning in model training. Furthermore, this paper discusses the main challenges currently faced, including the acquisition of high-quality datasets and the handling of data noise. Finally, we look ahead to future research directions, including the application prospects of emerging technologies such as multimodal fusion and few-shot learning. This review aims to provide researchers with a comprehensive perspective to further advance the prediction of protein mutation effects.

Keywords: protein language models; mutation effect prediction; deep learning; sequence models; structure models; multimodal fusion; unsupervised learning; supervised learning

蛋白质是生命活动的主要执行者，其功能高度依赖于氨基酸序列及其三维结构[1]。蛋白质中的氨基酸突变可能导致其功能、稳定性或相互作用能力的改变，进而影响生物体的表型，甚至导致疾病的发生[2]。因此，准确预测蛋白质突变的效应对于理解疾病机制、药物设计和蛋白质工程等领域具有重要意义。传统的蛋白质突变效应预测方法主要依赖于序列保守性分析[3]、物理化学特性计算[4]或统计学习方法[5]。这些方法虽然在某些情况下表现良好，但存在显著局限性：首先，它们难以捕捉蛋白质序列中的长程依赖关系，例如远距离残基之间的协同效应[6]；其次，这些方法通常无法有效利用蛋白质的三维结构信息，导致对局部环境变化的预测能力不足[7]。近年来，深度学习技术，尤其是以 Transformer[8]架构为基础的预训练语言模型，在蛋白质科学中展现出了巨大潜力。与传统的序列保守性分析和统计学习方法相比，这些模型通过自注意力机制能够有效捕捉蛋白质序列中的长程依赖关系。此外，基于 Transformer 的语言模型通过在大规模未标注的蛋白质序列数据上进行预训练，能够学习到丰富的进化信息和上下文相关的序列表示[9]，为突变效应预测提供了更强大的工具和新的视角。

本文系统地回顾和分析蛋白质语言模型在突变效应预测中的最新进展，重点关注 3 类主要模型：基于序列的模型、利用结构信息的模型，以及结合序列和结构信息的混合模型。本

文的主要内容包括：(1) 对 3 类模型的原理、优势和局限性进行了深入比较，为研究者选择合适模型提供指导；(2) 详细探讨了无监督预训练和有监督微调在模型训练中的应用，特别是多任务学习和少样本学习等新兴技术的进展；(3) 总结了蛋白质语言模型在蛋白质工程中的成功应用实例，展示了其在酶热稳定性改造、基因编辑工具优化等领域的实际价值；(4) 系统分析了当前面临的主要挑战，如高质量数据集的获取、数据噪声的处理，以及模型的可解释性问题；(5) 展望了未来研究方向，包括多模态融合、动态结构建模和跨物种泛化等。

本文期望能够为研究者提供一个全面的视角，帮助他们了解蛋白质语言模型在突变效应预测中的最新进展、现存挑战和未来机遇。相信随着这一领域的不断发展，蛋白质语言模型将在生物信息学、药物设计和蛋白质工程等多个领域发挥越来越重要的作用。

1 蛋白质突变适应性

1.1 蛋白质突变的生物学意义

蛋白质突变适应性(protein mutation fitness)是指突变对蛋白质功能和生物体适应能力的影响程度。从分子水平来看，突变适应性通常通过突变体与野生型在功能、稳定性或表达水平等方面的相对差异来衡量。正向的突变适应性表明突变提高了蛋白质的功能或稳定性，而负向的突变适应性则意味着突变导致了功能的丧失或稳定性的降低。蛋白质突变在生物学中具有深远的影响，其效应范围广泛，既可能导致蛋白质细微的功能变化，也可能导致严重的病理表型。首先，突变可能对蛋白质的结构产生显著影响。氨基酸的改变可能导致局部或全局结构的变化，进而影响蛋白质的稳定性和功能。例如，当疏水性氨基酸突变为亲水性氨基酸时，可能会破坏蛋白质核心区域的稳定性，导致整体折叠异常[10]。这种结构变化的经典案例是镰状细胞贫血症，其中 β-珠蛋白的单点突变(谷氨酸到缬氨酸)导致血红蛋白异常聚合，进而引发疾病[11]。

此外，突变可能发生在蛋白质的活性位点或结合位点，改变其催化活性或与其他分子的相互作用。例如，*CFTR* 基因中的 F508del 突变会导致蛋白质错误折叠和降解，这是囊性纤维化的主要病因[12]。某些突变在进化过程中可能为生物体提供选择优势。以人类免疫缺陷病毒(human immunodeficiency virus, HIV)感染为例，*CCR5* 基因中的 Δ32 突变能够赋予个体部分抗性，这种突变在特定人群中显著增加[13-14]。进一步来说，许多遗传疾病与特定的蛋白质突变密切相关。例如，亨廷顿舞蹈症是由亨廷顿蛋白中的 CAG 三核苷酸重复扩增引起的[15]。此外，蛋白质突变还可能影响药物反应。非小细胞肺癌(non small cell lung cancer, NSCLC)患者中的表皮生长因子受体(epidermal growth factor receptor, EGFR)基因突变，尤其是 19 号外显子缺失突变和 21 号外显子 L858R 点突变，与对吉非替尼等靶向药物的敏感性密切相关，这在癌症治疗中具有重要意义[16]。

蛋白质突变在工业酶的定向改造中也得到了广泛应用。例如，高度热稳定的地衣芽孢杆菌 α-淀粉酶及其衍生物广泛用于淀粉液化，Machius 等[17]在酶序列中引入 5 个点突变，使酶的解折叠温度提高了 13 °C，显著提高了其热稳定性。在改造酶的立体选择性方面，Reetz 团队[18]通过错误倾向 PCR 和饱和突变，成功将假单胞菌脂肪酶对 2-甲基癸酸酯的对映选择性从 2%显著提高到 90%–93%。Song 和 Rhee[19]通过随机突变和重组筛选获得了在有机溶剂中具有增强稳定性和活性的磷脂酶 A1 突变体。这些

成功案例充分展示了定向进化在工业酶改造中的巨大潜力。

1.2 蛋白质突变适应性预测方法的发展

蛋白质突变适应性预测方法的发展经历了几个关键阶段，每个阶段都有其特点和贡献。

在 20 世纪末至 21 世纪初，实验方法成为研究蛋白质突变适应性的主要手段。定点突变实验能在体外或体内直接检测突变对蛋白质功能的影响，是研究这一领域最直接的方法之一[20]。此外，高通量筛选技术的引入，使得研究者能够通过系统性的方法快速评估大量突变的功能性影响，例如 2000 年 Weiss 等[21]提出的组合氨基酸扫描(combinatorial alanine scanning)方法，通过将蛋白质中的每个氨基酸逐一突变为丙氨酸，来识别关键功能性残基，从而在短时间内提供了宝贵的实验数据。然而，这些方法由于耗时、成本高，且难以覆盖所有可能的突变，其应用范围受到限制。

21 世纪初至 2015 年，早期计算方法开始逐步发展，并成为研究的主流。这一阶段的方法主要包括序列保守性分析、基于物理化学性质的预测以及统计学习方法。序列保守性分析基于多序列比对(multiple sequence alignment, MSA)，通过评估氨基酸位点的保守程度来预测突变的影响。物理化学特性预测则利用氨基酸的物理化学性质，如疏水性、电荷等，来推测突变对蛋白质功能的影响。统计学习方法，如支持向量机(support vector machines, SVM)、随机森林等，被广泛应用于突变适应性预测中，代表性工具包括 SIFT[22]、PolyPhen-2[23]和 PROVEAN[24]。虽然这些方法提高了计算效率，能够快速评估大量突变，但其预测精度有限，难以捕捉蛋白质结构与功能之间的复杂关系。

自 2015 年以来，随着深度学习技术的快速发展，蛋白质突变适应性预测进入了一个全新的阶段。现代方法如蛋白质语言模型、卷积神经网络、图神经网络和多模态融合模型等大幅提升了预测的精度和能力。蛋白质语言模型利用自然语言处理技术，从大量蛋白质序列数据中学习潜在特征；基于卷积神经网络的模型可以从蛋白质 3D 微环境以及生物物理特征中提取蛋白质局部特征，高效地发现新的功能增益突变；图神经网络则将蛋白质结构表示为图，学习氨基酸之间的空间关系；多模态融合模型通过结合序列、结构等多种信息，实现了更加精确地预测蛋白质突变。其中纯序列语言模型代表性模型包括 ESM-1b[9]、ESM-1v[25]、ESM2[26]、ProteinBERT[27]、UniRep[28]和 EVE[29]等，卷积神经网络的代表模型有 MutCompute[30]、MutComputeX[31]等，使用图神经网络的模型包括 ProtSSN[32]、ESM-IF[33]等，结合序列和结构信息的多模态模型包括 ProSST[34]、ProstT5[35]和 SaProt[36]等。这些方法显著提高了预测的精度，能够捕捉到更复杂的非线性关系，然而，这些方法依赖于大量数据和计算资源，且模型的解释性仍有待进一步提高。

总的来说，蛋白质突变适应性预测方法的发展经历了从实验到计算再到深度学习的演变，每个阶段都有其独特的挑战和贡献。随着技术的不断进步，这一领域必将继续发展，为理解和应对蛋白质突变的生物学意义提供更为强大的工具和方法。

2 蛋白质语言模型

2.1 概述

蛋白质语言模型(protein language model, PLM)作为深度学习领域的前沿方法，通过引入自然语言处理(natural language processing, NLP)技术，为蛋白质序列的潜在特征和规律学

习提供了新的途径。从数学角度看，语言模型的核心是对序列的概率分布进行建模，给定一个长度为 n 的序列 $x=(x_1,\ldots,x_{n-1})$，语言模型的目标是对该序列的概率分布 $p(x)$进行建模[37]，计算方法如公式(1)所示：

$$p(x)=p(x_1)p(x_2|x_1)\ldots p(x_n|x_1,\ldots,x_{n-1}) \quad (1)$$

在蛋白质研究中，20 种氨基酸或者融入结构信息的 token 被视为“词汇”，由这些“词汇”构成的蛋白质序列被视为“句子”，蛋白质语言模型的目的就是学习这些序列的似然分布。在应用过程中，PLM 通常首先在大规模未标注的蛋白质序列数据上进行预训练，普遍采用掩码任务或自回归任务，以捕捉序列中的重要模式。通过这一过程，模型能够学习到蛋白质序列的上下文相关表示，有效捕捉氨基酸之间的长程依赖关系及其蕴含的进化信息。这些预训练后的模型可以通过微调或特征提取应用于具体的下游任务，如突变效应预测、功能注释和结构预测等，显著提升任务性能。此外，PLM 在多尺度建模方面的能力，使其能够同时捕捉局部残基环境和全局序列模式，为蛋白质特征表示提供了更丰富的层次。

在蛋白质突变适应性预测领域，PLM 展现出卓越的应用前景。首先，PLMs 通过生成上下文相关的向量表示，编码序列中的复杂模式和进化信号，这些表示能够有效捕捉序列之间的依赖关系，进而直接用于突变效应的高精度预测。其次，部分 PLM (如 ESM-1v)具有零样本预测能力，能够对未见过的突变进行预测，无需额外的任务特定训练。此外，PLM 生成的特征还可以与其他类型的特征(如结构信息)结合，进一步提高模型的整体预测性能。通过在大规模数据上进行预训练，PLM 能够将学习到的知识迁移到数据稀缺的任务中，从而显著提升预测的准确性和泛化能力。

在分类蛋白质语言模型时，可以根据其所利用的训练数据类型将其分为序列模型、结构模型以及序列+结构模型。序列模型主要依赖于蛋白质的氨基酸序列，通过学习序列中的模式来进行功能预测和突变分析。结构模型则通过捕捉蛋白质的三维构象信息，深入解析结构与功能之间的关系。序列+结构模型整合了序列和结构这 2 方面的数据，旨在通过融合多模态信息来提供更加全面和精准的预测结果。这种分类方法不仅有助于理解不同模型的优势和适用场景，还为研究者在特定应用中选择合适的模型提供了指导依据。在未来的研究中，这些不同类别的模型可能会进一步结合，形成更加复杂且有效的多模态学习框架，以应对蛋白质研究中的挑战性问题。

2.2 序列模型

序列模型是蛋白质语言模型中的关键类别，这类模型在训练和推理过程中均仅使用蛋白质的氨基酸序列作为输入。通过分析这些序列，模型能够学习其中的模式和进化信号，从而有效预测突变对蛋白质功能的影响。根据是否依赖 MSA，序列模型可以进一步细分为依赖 MSA 的模型和不依赖 MSA 的模型。依赖 MSA 的模型通过多序列比对获取进化信息，以捕捉序列之间的共演化关系；而不依赖 MSA 的模型则直接从单一序列中提取信息，利用深度学习技术识别序列中的复杂模式和长程依赖性。这 2 种类型的模型在突变效应预测中各有优劣，适用于不同的生物学研究场景。

2.2.1 基于 MSA 的蛋白质语言模型

MSA 是一种广泛应用于生物信息学的技术，通过将来自不同物种或同一物种中不同蛋白质的序列进行对齐，揭示序列中的保守性和变异性。基于 MSA 的基因变异效应预测研究

已积累了丰富的理论基础和实践经验[3]。早期的模型主要关注从比对中提取位点特异性的信息，而后续的研究则致力于捕捉更为复杂的模式。Hopf等[6]提出了使用基于能量的模型来模拟不同位点对之间的相互作用。Riesselman 等[38]进一步扩展了这一概念，提出了 DeepSequence：基于蛋白质特异性 MSA 训练的变分自编码器，用于学习氨基酸序列的分布，从而捕捉更高阶的相互作用。在预测与人类疾病相关的蛋白质变异致病性方面，EVE[18]进一步改进了 DeepSequence 架构，以实现更高的适应性预测性能。Rao 等[39]提出的 MSA Transformer 模型使用轴向注意力机制整合多序列比对(MSA)信息，通过在序列和位点之间交替进行注意力计算，显著提高了蛋白质结构预测的准确性和参数效率，在处理低深度 MSA 时尤其表现优异。

Tranception[40]是使用 MSA 信息的 PLM 代表之一，它采用了 Transformer 架构，并在推理过程中利用 MSA 信息有效提升了突变效应预测的准确性。其核心技术在于双模式推理策略的应用：首先，Tranception 模型利用基于 Transformer 的自回归架构在序列上进行预训练；随后，在推理过程中，除了自回归输出目标序列的似然得分，通过检索目标蛋白质家族的同源序列，利用这些序列中氨基酸的经验分布可以进一步提升突变体适应性得分的预测精度。最终的突变效应得分通过自回归推理模式和检索推理模式所得对数似然的加权算术平均计算得出。假设原始序列为 S，突变序列为 S'，则 Tranception 模型的得分如公式(2)所示[40]：

$$\begin{aligned}\mathrm{Score}(S') = (1-\alpha)\sum_{i=1}^{l}\ln P(S_i' \mid S_{<i}) + \\ \alpha\sum_{i=1}^{l}\ln P(S_i' \mid \text{retrieved sequences})\end{aligned} \tag{2}$$

其中，$P(S_i' \mid S_{<i})$ 是位置 i 上的氨基酸在自回归模式下的概率，$P(S_i' \mid \text{retrieved sequences})$ 表示基于检索到的同源序列对突变氨基酸 S_i' 的经验概率分布，l 是序列长度，α 是一个可调节的权重超参。

最近，DeepMind 团队提出的 AlphaMissense[41]模型在基于 MSA 的蛋白质语言模型领域取得了重要突破。该模型采用了 2 阶段训练策略：首先进行类似 AlphaFold2[42]的预训练，同时执行单链结构预测和蛋白质语言建模任务；随后在人类蛋白质数据上进行微调，引入变异致病性分类目标。AlphaMissense 的创新之处在于其独特的 MSA 处理方式：在预训练阶段采用随机位置掩码策略学习通用蛋白质知识，在微调阶段则针对性地掩码变异位置以专注于变异效应预测。模型通过在 MSA 的第 2 行呈现变异序列，并结合来自人类和灵长类群体的频率信息进行训练，实现了对蛋白质变异致病性的高精度预测。此外，AlphaMissense 还引入了自蒸馏技术来提升训练集质量，通过初步模型过滤可能的良性变异，进一步提高了预测准确性。

基于 MSA 的蛋白质语言模型尽管在捕捉进化信息方面具有显著优势，但仍存在若干局限性。首先，这些模型在处理插入和删除等与 MSA 坐标系统不兼容的序列时，无法有效预测，从而限制了其适用范围。此外，蛋白质组中存在大量无法通过 MSA 进行有效比对的无序区域，研究表明，约 50%的人类蛋白质含有至少 40 个氨基酸长的无序区域[43-44]，这些区域难以对齐使得基于 MSA 的模型在这些区域的预测能力受到限制。即使在可进行比对的区域，MSA 模型的性能也高度依赖于比对的质量和深度，尤其当蛋白质功能特定于某一分类群时，MSA 算法可能无法检索到足够多的同源序列用于模型训练。此外，这类模型在不同数据子集上独立训练，缺乏信息共享，导致难以充分利用多种数据来源的优势。这些限制揭示了基

于MSA的蛋白质语言模型在处理复杂和多变的蛋白质序列预测任务中的局限性。

2.2.2 不依赖MSA的蛋白质语言模型

不依赖MSA的序列模型通过直接利用蛋白质的原始序列进行训练，克服了传统依赖MSA方法的局限。ESM系列模型为该类PLM的一个重要进展，该系列模型从ESM-1b开始，通过在大规模未标注的蛋白质序列数据上进行预训练，成功学习了蛋白质序列中的隐含生物学特征。ESM-1b包含大约6.5亿个参数，展现了其在多个生物信息学任务中的有效性，尤其是在无法获取高质量MSA的情况下，其在蛋白质功能预测中的强大能力[9]。ESM-1v在ESM-1b的基础上进行了优化，采用了更大规模的训练数据集和改进的训练技术，提出了语言模型具有预测蛋白质突变适应性的零样本能力[25]。ESM2进一步扩展参数规模到150亿，并引入了更为复杂的模型架构和预训练策略，显著提升了蛋白质结构预测的分辨率和速度[26]。在突变效应的预测上，ESM系列模型可通过计算突变序列与野生型序列在突变位置处的对数概率之差来评估突变的适应性，打分如公式(3)所示：

$$\sum_{t\in T}[\ln p(x_t = x_t^{mt} \mid x_{\setminus T}) - \ln p(x_t = x_t^{wt} \mid x_{\setminus T})] \quad (3)$$

其中，求和代表对多点突变进行打分，$p(x_t = x_t^{mt} \mid x_{\setminus T})$和$p(x_t = x_t^{wt} \mid x_{\setminus T})$分别代表给模型输入将突变区掩码后的野生型序列的前提下，突变位点处突变氨基酸和野生型氨基酸出现的概率[25]。

2.3 结构模型

结构模型通过输入蛋白质的三维结构信息来预测突变效应，这类模型能够捕捉到序列模型难以获取的空间构象信息。早期的结构模型主要基于能量函数，如Rosetta[45]和FoldX[46]，它们依据物理化学原理计算突变前后蛋白质的能量变化。随着机器学习技术的发展，研究者开始利用结构特征训练SVM和随机森林(random forest, RF)等模型来预测突变效应。例如，Pires等[7]开发了基于SVM的mCSM方法，通过图理论来表示蛋白质结构环境，成功预测了突变对蛋白质稳定性的影响。近年来，深度学习在蛋白质科学中的应用日益广泛，3D卷积神经网络(3D convolutional neural network, 3DCNN)和图神经网络(graph neural network, GNN)等方法被用于处理蛋白质结构信息。Torng和Altman[47]提出了一种基于3DCNN的方法，直接从原子级别的3D表示中学习蛋白质特征。Rao等[48]开发了GVP-GNN模型，利用图神经网络处理蛋白质结构，在多个蛋白质设计任务中取得了优异成绩。

近年来，结合了语言模型和结构信息的逆折叠模型ESM-IF的出现标志着结构模型的新突破。ESM-IF是一个条件语言模型[33]，它以蛋白质结构作为条件c，建模序列x的条件概率分布$p(x|c)$。具体来说，给定蛋白质主链结构X，模型通过自回归方式预测氨基酸序列Y的似然概率，计算方法如公式(4)所示：

$$p(Y \mid X) = \prod_{i=1}^{n} p(y_i \mid y_{i-1},...,y_1; X) \quad (4)$$

其中，X表示蛋白质主链的空间坐标，Y表示预测的氨基酸序列，$p(y_i \mid y_{i-1},...,y_1; X)$表示在给定前序氨基酸和主链结构的情况下，第$i$个氨基酸的条件概率；ESM-IF采用了GVP+Transformer架构，GVP (geometric vector perceptron)是一种专门设计用于处理3D几何数据的神经网络架构，可以保持输入特征的旋转等变性，结合自然语言处理中常用的Transformer架构后，ESM-IF将蛋白质结构信息编码为一种“空间语言”，然后使用自回归方式预测输入主链坐标对应的似

然概率最大的氨基酸序列；这种设计最初主要是为了解决蛋白质设计中的逆折叠问题，但由于模型可以输出每个位置处各种氨基酸出现的概率，ESM-IF 也可以用于预测突变的适应性；得益于AlphaFold2的应用，ESM-IF利用了1 200万个AlphaFold2 预测的蛋白质结构进行训练，显著提高了模型的泛化能力[33]。

ESM-IF 的成功表明，结合大规模结构数据和先进的深度学习架构可以显著提升结构模型的性能。这种方法不仅能够从蛋白质结构反向预测序列，还可以为蛋白质突变适应性预测提供新的范式。然而，结构模型仍面临一些挑战，例如：如何处理蛋白质的动态性和柔性，以及如何更好地整合序列和结构信息等。未来的研究方向可能包括发展多尺度模型、整合分子动力学模拟等方法来进一步提高预测准确性。这些不同类型的结构模型，特别是融合了语言模型思想的方法，共同推动了蛋白质突变效应预测领域的进展。

2.4 序列+结构模型

序列+结构模型通过结合蛋白质的氨基酸序列信息和三维结构信息，在蛋白质突变效应预测任务中取得了显著进展。这类模型不仅继承了语言模型在序列建模方面的优势，还能捕捉蛋白质的空间构象信息，从而提高突变效应预测的准确性。

早期的序列+结构模型主要采用简单的特征组合方法。例如，Wang 等[49]开发的 LM-GVP 模型将 ProtBERT-BFD[50]的序列嵌入作为 GVP 模型的输入特征，实现了序列和结构信息的初步融合。近年来，随着 AlphaFold2 等大规模蛋白质结构预测技术的进步，基于预测结构的序列+结构模型在突变效应预测领域取得了重大突破。Li 等[34]开发的 ProSST 模型引入了创新的结构量化模块，将三维蛋白质结构序列化为离散的结构令牌，并与氨基酸序列信息结合进行联合训练。这种方法使得 ProSST 能够在保留语言模型优势的同时，有效捕捉蛋白质的局部结构特征，从而提高突变效应预测的准确性。Su 等[36]建立的 SaProt 模型提出了“结构感知词汇”的概念，将残基令牌与结构令牌相结合，通过 Foldseek 工具编码蛋白质的三维结构信息；这种创新的表示方法使 SaProt 能够在语言模型的框架下融入结构信息，在零样本突变效应预测任务中展现出强大的性能。SaProt 的成功表明，将结构信息引入语言模型可以显著提升模型对蛋白质突变影响的理解能力。最近，EvolutionaryScale 发布的 ESM3[51]代表了序列+结构模型的重大突破；ESM3 是一个多模态生成语言模型，能够同时对蛋白质的序列、结构和功能进行建模；与 ProSST 以及 SaProt 类似，ESM3 将三维原子结构编码为离散的 token，这种设计使模型能够以统一的方式处理序列和结构信息，同时保持了计算效率；在训练规模上，ESM3 使用了 27.8 亿个蛋白质序列和 7 710 亿个独特 token 进行训练，模型包含 980 亿个参数；如此大规模的训练不仅使模型在突变效应预测任务上展现出了优异的性能，更赋予了模型强大的蛋白质设计能力，例如，ESM3 能够根据功能需求生成全新的蛋白质序列，并成功设计出与已知序列差异显著的功能性蛋白质。这种突变预测与蛋白质设计相结合的能力，使 ESM3 成为一个更全面的蛋白质工程工具。ESM3 的成功表明，通过合理的结构编码方式和大规模多模态训练，语言模型可以超越单一的突变效应预测，为蛋白质工程提供更多可能性。

2.5 蛋白质突变效应预测模型性能评估数据集 ProteinGym

ProteinGym 是一个大规模、综合性的蛋白

质突变效应预测基准测试平台，旨在为不同模型提供标准化评估环境，推动蛋白质突变效应预测和设计领域的发展[52]；该平台通过整合超过250个标准化的深度突变扫描(deep mutational scanning, DMS)实验数据，涵盖数百万个突变序列，同时结合高质量的临床数据集，提供了丰富的突变效应注释；ProteinGym 的显著特点在于其规模性和全面性，不仅包括错义突变和插入/缺失突变，还涵盖了从基础研究到临床应用的多样化数据来源。其评估框架结合了突变效应预测和蛋白质设计的多种指标，考虑了实验方法的局限性，并支持零样本和监督学习这 2 种设置。通过将来自不同领域的 70 多个高性能模型(如基于多序列比对的模型、逆折叠模型等)纳入统一的基准测试，ProteinGym 为研究者提供了一个全面、透明的性能评估平台[52]；另外，ProteinGym 开源了相关代码库、数据集、多序列比对(MSA)信息、结构数据以及模型预测结果，并开发了用户友好的网站，便于数据访问和分析；表 1 展示了 ProteinGym 官方测试中部分模型打分与实验值的 Spearman 相关系数情况，涵盖了酶活性(activity)、分子结合(binding)、表达水平(expression)和蛋白质稳定性(stability)等功能分类，为模型性能的全面评估提供了重要参考。

3 训练方法

3.1 无监督学习

无监督学习方法通过大量未标注的蛋白质序列进行预训练，模型能够从中学习到序列的统计特性和进化模式。该方法的关键优势在于可以充分利用海量数据，而无需依赖昂贵且耗时的标注过程。无监督学习通常可以分为双向编码器表示(bidirectional encoder representations from transformers, BERT)模式和自回归(autoregressive)模式。

BERT 模式的无监督学习通过掩蔽语言模型(masked language modeling, MLM)来实现[53]；该方法通过随机掩蔽输入序列中的部分氨基酸，然后要求模型根据上下文预测这些被掩蔽的部分；这种方式使得模型能够学习到序列的双向依赖关系，即同时从前向和后向的上下文信息中获取知识。在 BERT 模式中，假设蛋白质序列为 X，模型参数为 θ，被掩蔽的位置集合为 M，则无监督学习的目标是最大化以下对数似然[53]，计算如公式(5)所示：

$$L_{\mathrm{MLM}}(\theta)=\sum_{i\in M}\ln\, p(x_i\,|\,\mathrm{X}_{\setminus M};\theta) \tag{5}$$

其中，$p(x_i\,|\,\mathrm{X}_{\setminus M};\theta)$ 表示在给定未掩蔽序列 $\mathrm{X}_{\setminus M}$ 的情况下，预测被掩蔽位置 i 上氨基酸 x_i 的条件概率。BERT 模式被广泛应用于 ESM 系列模型以及 ProSST 模型的预训练阶段。

表 1 ProteinGym 测试结果[52]

Table 1 ProteinGym test results[52]

Model name	Average	Activity	Binding	Expression	Stability	Model type
SaProt (650M)	0.457	0.458	0.378	0.488	0.592	Sequence+Structure
TranceptEVE L	0.456	0.487	0.376	0.457	0.500	Sequence+MSA
ProtSSN (ensemble)	0.449	0.466	0.366	0.449	0.568	Sequence+Structure
MSA Transformer	0.434	0.469	0.337	0.446	0.495	Sequence+MSA
ESM-IF	0.422	0.368	0.389	0.407	0.624	Structure
ESM2 (650M)	0.414	0.425	0.337	0.415	0.523	Sequence

自回归模式的无监督学习通过逐步生成序列的方式进行。模型以前序氨基酸为条件，依次预测序列中的每个氨基酸，这种方式使得模型能够学习到序列的单向依赖关系，即当前氨基酸仅依赖于之前的氨基酸[40]；自回归模式中，假设蛋白质序列为 $X=\{x_1,x_2,...,x_n\}$，模型参数为 θ，则无监督学习的目标是最大化以下对数似然，计算方法如公式(6)所示：

$$L_{\text{Autoregressive}}(\theta)=\sum_{i=1}^{n}\ln\, p(x_i \mid x_{<i};\theta) \tag{6}$$

其中，$p(x_i \mid x_{<i};\theta)$ 表示在给定前序氨基酸 $x_{<i}$ 的条件下，预测第 i 个氨基酸 x_i 的概率。自回归模式通常用于模型如 ESM IF 中的序列生成和逆折叠任务中[33]。Tranception 模型是自回归模式的一个典型代表[40]；过结合自回归变换器架构和推理时的同源序列检索功能，成功实现了对蛋白质突变效应的高精度预测；Tranception 模型利用序列的自回归特性逐步生成并评估突变序列，同时在推理时动态检索同源序列信息，从而显著提高了模型对复杂突变效应的预测能力。

3.2 监督学习

监督学习方法通过利用已标记的蛋白质突变数据来训练模型，从而提高模型预测突变效应的能力。近年来，多任务学习和少样本学习在蛋白质突变预测领域取得了显著进展。

多任务学习通过同时学习多个相关任务，利用任务间的共享信息来提高模型性能。Chen 等[53]提出的 GVP-MSA 模型在这一领域展现了出色的性能。该模型巧妙地结合了几何向量感知图神经网络和 MSA Transformer，分别用于提取蛋白质结构信息和捕获进化信息。通过多任务训练框架，GVP-MSA 能够同时学习多个蛋白质的突变效应预测任务，有效整合不同蛋白质的信息。在 87 个深度突变扫描数据集上的实验结果表明，该方法不仅提高了单蛋白质预测的准确性，还在外推到高阶突变和新位点突变方面表现出色[54]。这些结果突显了多任务学习在提升模型泛化能力方面的优势，为整合多样化蛋白质数据以改进突变效应预测提供了新的思路。

在实际应用中，获取大量标记数据往往成本高昂，这促使研究者们探索如何利用有限的标记数据来提高模型性能。eUniRep[55]作为一种创新的少样本学习方法，通过结合全局无监督预训练和局部微调(evotuning)，能够在仅使用少量功能表征序列(如 24 个突变体)的情况下，构建高质量的虚拟适应性景观，并通过计算机模拟的定向进化筛选出功能增强的蛋白质变体；eUniRep 首先在包含超过 2 000 万个氨基酸序列的数据库 UniRef50 上进行全局无监督预训练，学习所有功能蛋白质的全局特征；随后在目标蛋白质的局部序列邻域中进行微调，结合全局和局部信息，提升模型的表现[54]；通过这种两阶段的学习策略，eUniRep 能够在低数据量的情况下，设计出功能显著增强的蛋白质变体，如荧光蛋白 avGFP 和 TEM-1 β-内酰胺酶，其性能与经过多年高通量工程优化的变体相当。这一方法不仅减少了实验数据的需求，还为蛋白质工程提供了一种高效且可扩展的设计范式。

Zhou 等[56]提出的 few-shot learning for protein fitness prediction (FSFP)方法也在定向进化的少样本学习领域取得了重要突破。FSFP 通过创新性地结合元学习、排序学习和 LoRA[57]微调技术，实现了仅使用几十个标记的单位点突变样本就能显著提升蛋白质语言模型性能的目标。该方法利用其他蛋白质数据构建辅助任务，通过元学习获得更好的模型初始化；将突变效应预测转化为排序问题，更贴合蛋白质工程的实际需求；并采用 LoRA 技术进行

参数高效微调，有效避免了过拟合问题。在ProteinGym基准测试中，FSFP展现出优异的性能，能够在仅有 20 个训练样本的情况下，将蛋白质语言模型的性能提升 0.1 (以 Spearman 相关系数衡量)[56]；更值得注意的是，FSFP 在湿实验中成功将小样本学习方法应用于 Phi29 DNA 聚合酶的热稳定性改造工程，进一步验证了其在实际蛋白质设计中的实用价值。

在监督学习框架的创新方面，Notin 等[57]提出的ProteinNPT模型采用了半监督条件伪生成模型的设计；该模型的核心特征是其独特的三轴注意力机制：通过 MSA Transformer 层处理同源序列间的注意力，在ProteinNPT层中实现残基-标签间的行注意力和样本间的列注意力。在训练策略上，模型采用两阶段方法：利用预训练语言模型的嵌入应对标签稀缺问题，并通过结合输入去噪和目标预测的联合任务进行优化；模型还引入辅助标签机制提升预测性能，并支持基于目标属性值的条件序列采样。这种设计在标签稀缺情况下表现出色，在多个蛋白质属性预测基准测试中取得了领先结果。

总的来说，监督学习方法通过有效利用标记数据，显著提高了蛋白质语言模型在突变效应预测任务上的性能。多任务学习能够整合多个蛋白质的信息，提高模型的泛化能力；而少样本学习则为解决标记数据稀缺的问题提供了可行的解决方案。这些进展不仅推动了计算方法在蛋白质工程领域的应用，还为定向进化提供了有力的工具支持，为未来蛋白质功能的精确预测和定制化设计铺平了道路。

4 蛋白质语言模型在蛋白质工程中的应用实例

蛋白质语言模型已在蛋白质工程领域取得多项重要应用成果。Pro-PRIME[59]是一种新型的温度感知语言模型，模型基于 9 600 万个带有温度标签的蛋白质序列进行预训练，通过token 层面的掩码语言建模(MLM)和序列层面最优生长温度(optimal growth temperature, OGT)预测的多任务学习，使模型能够更好地理解蛋白质序列与温度的关系；其次，模型引入correlation loss项来对齐token和序列层面的任务信息，增强了对蛋白质温度特征的捕捉能力；最后，模型采用零样本预测和小样本微调相结合的策略，能够在极少量实验数据的情况下快速优化蛋白质性能。Pro-PRIME 的实际应用效果在多个案例中得到了验证：在肌酸酶(404 个氨基酸)的热稳定性改造中，模型预测的 top-45 个突变位点中约 32%表现出显著提升的热稳定性；在 T7 RNA 聚合酶的优化中，通过 4 轮干湿结合迭代，获得了 T_m 值提高 12.8 ℃且活性提升近 4 倍的多点突变体，性能超越现有商业化产品；在 VHH 抗体(142 个氨基酸)的改造中，约 38%的预测突变体表现出增强的碱性耐受能力；在非天然核酸聚合酶(773 个氨基酸)的底物特异性改造中，约 41%的预测突变体显示出增强的非天然底物聚合能力[59]。

另一个成功案例是Protein Mutational Effect Predictor (ProMEP)[60]模型。该模型通过多模态深度表示学习，结合序列和结构信息实现零样本突变效应预测，无需依赖多序列比对。ProMEP在基因编辑酶的工程化改造中展现出显著效果：通过 5 个位点突变将 *TnpB* 的基因编辑效率从 24.66%提升至 74.04%；基于 15 个位点突变优化的TadA碱基编辑工具，其A-to-G转换频率达到77.27%，同时显著降低了脱靶效应[60]。

这些成功案例不仅证明了蛋白质语言模型在不同类型蛋白质改造中的通用性，同时大大提升了蛋白质改造工作的效率，这对传统蛋白质工程具有重要的指导意义。

5 当前挑战

尽管蛋白质语言模型在预测突变效应方面取得了显著进展，但仍面临若干关键挑战。其中，获取高质量、大规模的蛋白质突变数据集对于训练和验证预测模型至关重要，然而这些数据的收集和质量控制仍然是一个重大难题。

Fowler 和 Fields 指出，尽管深度突变扫描(DMS)实验能够提供大量突变数据，但其实施仍然面临着挑战，他们强调了高通量检测面临的困难，并指出了细胞外检测的局限性[61]。这些因素可能限制了高质量可用数据的数量和多样性，进而影响基于这些数据训练的模型的效果和泛化能力。Livesey 等[62]强调，不同实验条件和测量方法可能导致数据间的不一致性。例如，不同实验室使用的蛋白质表达系统、功能测定方法和环境条件可能存在差异，这些差异可能会影响突变效应的测量结果。如何整合和标准化来自不同来源的数据，以确保数据的一致性和可比性，是一个亟待解决的问题。

此外，Schmiedel 和 Lehner 强调，深度突变扫描(DMS)数据中的噪声是一个需要认真考虑的因素[63]。他们指出，突变效应的实验测量存在噪声，且噪声的大小通常取决于野生型氨基酸、突变氨基酸和结构位置。为了应对这一挑战，他们开发了一个概率模型，明确考虑了 DMS 数据中的噪声，以及不同实验和蛋白质内不同位置之间的噪声水平差异。这些问题不仅影响实验数据的质量，也对基于这些数据训练的模型的准确性提出了挑战，因此需要进一步研究来改进数据处理和模型训练方法。

6 总结与展望

本文全面回顾了蛋白质语言模型在突变效应预测领域的最新进展。详细讨论了 3 类主要模型：基于序列的模型、基于结构的模型以及结合序列和结构信息的模型。这些模型各有优势，为蛋白质突变效应预测提供了多角度的解决方案。

基于序列的模型，如 ESM 系列，展现了强大的序列模式捕捉能力，能够有效处理长程依赖关系。基于结构的模型，如 ESM-IF，通过整合三维结构信息，提高了对蛋白质局部环境的理解。结合序列和结构的模型，如 ProSST 和 SaProt，则通过融合多模态信息，实现了更全面的蛋白质表征。这些模型的发展不仅推动了突变效应预测的准确性，也为蛋白质工程和药物设计等领域提供了有力工具。

在训练方法方面，无监督学习利用大规模未标注数据学习通用表示，而监督学习则通过标注数据进行任务特定优化。多任务学习和少样本学习等新兴技术的应用，进一步提高了模型的泛化能力和数据效率。

尽管取得了显著进展，该领域仍面临诸多挑战。高质量、大规模的蛋白质突变数据集的获取和标准化仍是一个亟待解决的问题。实验数据中的噪声和不一致性也对模型训练和评估提出了挑战。此外，如何更好地整合不同来源和类型的生物学知识，以提高模型的解释性和可靠性，也是未来研究的重要方向。

展望未来，以下几个方向值得关注：(1) 多模态融合。进一步探索序列、结构、功能等多种生物学信息的有效融合方法，以获得更全面的蛋白质表征。(2) 少样本和零样本学习。开发更先进的迁移学习和元学习技术，以应对特定蛋白质突变数据稀缺的问题。(3) 可解释性研究。提高模型的可解释性，使预测结果能够为生物学家提供更多参考。(4) 动态结构考虑。将蛋白质的动态性和柔性纳入模型考虑范围，以更准确地捕捉突变对蛋白质功能的影响。(5) 跨物种泛化。提高模型在不同物种间的泛化能

力,以应对生物多样性带来的挑战。(6) 与实验方法的结合。探索计算预测和实验验证的有效结合，建立反馈循环以持续改进模型性能。

总的来说，蛋白质语言模型为突变效应预测带来了新的机遇和挑战。随着技术的不断进步和跨学科合作的深入，有理由相信，这一领域将在推动生命科学研究和生物技术创新方面发挥越来越重要的作用。

作者贡献

张良：负责文献检索和筛选，撰写全文；谈攀：提出了综述主题和框架，语言润色；洪亮：提出了综述主题和框架，协调投稿。

利益冲突声明

作者声明，在本综述的撰写过程中不存在任何财务或个人利益冲突。

REFERENCES

[1] DILL KA, MacCALLUM JL. The protein-folding problem, 50 years on[J]. Science, 2012, 338(6110): 1042-1046.

[2] STENSON PD, MORT M, BALL EV, EVANS K, HAYDEN M, HEYWOOD S, HUSSAIN M, PHILLIPS AD, COOPER DN. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies[J]. Human Genetics, 2017, 136(6): 665-677.

[3] NG PC, HENIKOFF S. Predicting deleterious amino acid substitutions[J]. Genome Research, 2001, 11(5): 863-874.

[4] CAPRIOTTI E, FARISELLI P, CASADIO R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure[J]. Nucleic Acids Research, 2005, 33(Web Server issue): W306-W310.

[5] KUMAR P, HENIKOFF S, NG PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm[J]. Nature Protocols, 2009, 4(7): 1073-1081.

[6] HOPF TA, INGRAHAM JB, POELWIJK FJ, SCHÄRFE CPI, SPRINGER M, SANDER C, MARKS DS. Mutation effects predicted from sequence co-variation[J]. Nature Biotechnology, 2017, 35(2): 128-135.

[7] PIRES DEV, ASCHER DB, BLUNDELL TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures[J]. Bioinformatics, 2014, 30(3): 335-342.

[8] COHEN SB, COLLINS M. Advances in neural information processing systems 25[C]//Neural Information Processing Systems, 1999.

[9] RIVES A, MEIER J, SERCU T, GOYAL S, LIN ZM, LIU J, GUO DM, OTT M, LAWRENCE ZITNICK C, MA J, FERGUS R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[10] NICK PACE C, SCHOLTZ JM, GRIMSLEY GR. Forces stabilizing proteins[J]. FEBS Letters, 2014, 588(14): 2177-2184.

[11] INGRAM VM. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin[J]. Nature, 1956, 178(4537): 792-794.

[12] KEREM B, ROMMENS JM, BUCHANAN JA, MARKIEWICZ D, COX TK, CHAKRAVARTI A, BUCHWALD M, TSUI LC. Identification of the cystic fibrosis gene: genetic analysis[J]. Science, 1989, 245(4922): 1073-1080.

[13] LIU R, PAXTON WA, CHOE S, CERADINI D, MARTIN SR, HORUK R, MacDONALD ME, STUHLMANN H, KOUP RA, LANDAU NR. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection[J]. Cell, 1996, 86(3): 367-377.

[14] SAMSON M, LIBERT F, DORANZ BJ, RUCKER J, LIESNARD C, FARBER CM, SARAGOSTI S, LAPOUMEROULIE C, COGNAUX J, FORCEILLE C, MUYLDERMANS G, VERHOFSTEDE C, BURTONBOY G, GEORGES M, IMAI T, RANA S, YI Y, SMYTH RJ, COLLMAN RG, DOMS RW, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the *CCR*-5 chemokine receptor gene[J]. Nature, 1996, 382(6593): 722-725.

[15] MacDONALD ME, AMBROSE CM, DUYAO MP, MYERS RH, LIN C, SRINIDHI L, BARNES G, TAYLOR SA, JAMES M, GROOT N, MacFARLANE H, JENKINS B, ANDERSON MA, WEXLER NS, GUSELLA JF, BATES GP, BAXENDALE S, HUMMERICH H, KIRBY S, NORTH M, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes[J]. Cell, 1993, 72(6): 971-983.

[16] TANOUE LT. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy[J]. Yearbook of Pulmonary Disease, 2006, 2006: 112-114.

[17] MACHIUS M, DECLERCK N, HUBER R, WIEGAND G. Kinetic stabilization of *Bacillus licheniformis* α-amylase through introduction of hydrophobic residues at the surface[J]. Journal of Biological Chemistry, 2003, 278(13): 11546-11553.

[18] LIEBETON K, ZONTA A, SCHIMOSSEK K, NARDINI M, LANG D, DIJKSTRA BW, REETZ MT, JAEGER KE. Directed evolution of an enantioselective lipase[J]. Chemistry & Biology, 2000, 7(9): 709-718.

[19] SONG JK, RHEE JS. Enhancement of stability and activity of phospholipase A1 in organic solvents by directed evolution[J]. Biochimica et Biophysica Acta

(BBA)-Protein Structure and Molecular Enzymology, 2001, 1547(2): 370-378.
[20] SMITH M. *In vitro* mutagenesis[J]. Annual Review of Genetics, 19: 423-462.
[21] WEISS GA, WATANABE CK, ZHONG A, GODDARD A, SIDHU SS. Rapid mapping of protein functional epitopes by combinatorial alanine scanning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(16): 8950-8954.
[22] NG PC, HENIKOFF S. SIFT: Predicting amino acid changes that affect protein function[J]. Nucleic Acids Research, 2003, 31(13): 3812-3814.
[23] ADZHUBEI IA, SCHMIDT S, PESHKIN L, RAMENSKY VE, GERASIMOVA A, BORK P, KONDRASHOV AS, SUNYAEV SR. A method and server for predicting damaging missense mutations[J]. Nature Methods, 2010, 7: 248-249.
[24] CHOI Y, CHAN AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels[J]. Bioinformatics, 2015, 31(16): 2745-2747.
[25] MEIER J, RAO R, VERKUIL R, LIU J, SERCU T, RIVES A, MEIER J, RAO R, VERKUIL R, LIU J, SERCU T, RIVES A. Language models enable zero-shot prediction of the effects of mutations on protein function[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. ACM, 2021: 29287-29303.
[26] LIN ZM, AKIN H, RAO R, HIE B, ZHU ZK, LU WT, SMETANIN N, VERKUIL R, KABELI O, SHMUELI Y, dos SANTOS COSTA A, FAZEL-ZARANDI M, SERCU T, CANDIDO S, RIVES A. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.
[27] BRANDES N, OFER D, PELEG Y, RAPPOPORT N, LINIAL M. ProteinBERT: a universal deep-learning model of protein sequence and function[J]. Bioinformatics, 2022, 38(8): 2102-2110.
[28] ALLEY EC, KHIMULYA G, BISWAS S, AlQURAISHI M, CHURCH GM. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16(12): 1315-1322.
[29] FRAZER J, NOTIN P, DIAS M, GOMEZ A, MIN JK, BROCK K, GAL Y, MARKS DS. Disease variant prediction with deep generative models of evolutionary data[J]. Nature, 2021, 599(7883): 91-95.
[30] SHROFF R, COLE AW, DIAZ DJ, MORROW BR, DONNELL I, ANNAPAREDDY A, GOLLIHAR J, ELLINGTON AD, THYER R. Discovery of novel gain-of-function mutations guided by structure-based deep learning[J]. ACS Synthetic Biology, 2020, 9(11): 2927-2935.
[31] D'OELSNITZ S, DIAZ DJ, KIM W, ACOSTA DJ, DANGERFIELD TL, SCHECHTER MW, MINUS MB, HOWARD JR, DO H, LOY JM, ALPER HS, JESSIE ZHANG Y, ELLINGTON AD. Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme[J]. Nature Communications, 2024, 15(1): 2084.
[32] TAN Y, ZHOU BX, ZHENG LR, FAN GS, HONG L. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability[J]. eLife, 2024, 13: RP98033.
[33] HSU C, VERKUIL R, LIU J, LIN Z, HIE B, SERCU T, LERER A, RIVES A. Learning inverse folding from millions of predicted structures[C]. International Conference on Machine Learning (PMLR), 2022: 8946-8970.
[34] LI MC, TAN P, MA XZ, ZHONG B, YU HQ, ZHOU ZY, OUYANG WL, ZHOU BX, HONG L, TAN Y. ProSST: protein language modeling with quantized structure and disentangled attention. Advances in Neural Information Processing Systems, 2025, 37: 35700-35726.
[35] HEINZINGER M, WEISSENOW K, SANCHEZ JG, HENKEL A, MIRDITA M, STEINEGGER M, ROST B. Bilingual language model for protein sequence and structure[J]. NAR Genomics and Bioinformatics, 2024, 6(4): lqae150.
[36] SU J, HAN CC, ZHOU YY, SHAN JJ, ZHOU XB, YUAN FJ. SaProt: protein language modeling with structure-aware vocabulary[J/OL]. bioRxiv, 2005. Doi: 10.1101/2023.10.01.560349.
[37] BENGIO Y, DUCHARME R, VINCENT P, JANVIN C. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
[38] RIESSELMAN AJ, INGRAHAM JB, MARKS DS. Deep generative models of genetic variation capture the effects of mutations[J]. Nature Methods, 2018, 15(10): 816-822.
[39] RAO RM, LIU J, VERKUIL R, MEIER J, CANNY J, ABBEEL P, SERCU T, RIVES A. MSA transformer[C]. International Conference on Machine Learning (PMLR), 2021: 8844-8856.
[40] NOTIN P, DIAS M, FRAZER J, MARCHENA-HURTADO J, GOMEZ AN, MARKS D, GAL Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval[C]. Proceedings of the 39th International Conference on Machine Learning (PMLR), 2022, 162: 16990-17017
[41] CHENG J, NOVATI G, PAN J, BYCROFT C, ŽEMGULYTĖ A, APPLEBAUM T, PRITZEL A, WONG LH, ZIELINSKI M, SARGEANT T, SCHNEIDER RG, SENIOR AW, JUMPER J, HASSABIS D, KOHLI P, AVSEC Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense[J]. Science, 2023, 381(6664): eadg7492.
[42] JUMPER J, EVANS R, PRITZEL A, GREEN T, FIGURNOV M, RONNEBERGER O, TUNYASUVUNAKOOL K, BATES R, ŽÍDEK A, POTAPENKO A, BRIDGLAND A, MEYER C, KOHL SAA, BALLARD AJ, COWIE A, ROMERA-PAREDES B, NIKOLOV S, JAIN R, ADLER J, BACK T, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.
[43] RADIVOJAC P, OBRADOVIC Z, SMITH DK, ZHU G, VUCETIC S, BROWN CJ, DAVID LAWSON J, KEITH DUNKER A. Protein flexibility and intrinsic disorder[J]. Protein Science, 2004, 13(1): 71-80.
[44] TOTH-PETROCZY A, PALMEDO P, INGRAHAM J, HOPF TA, BERGER B, SANDER C, MARKS DS. Structured states of disordered proteins from genomic

sequences[J]. Cell, 2016, 167(1): 158-170.e12.
[45] ALFORD RF, LEAVER-FAY A, JELIAZKOV JR, O'MEARA MJ, DiMAIO FP, PARK H, SHAPOVALOV MV, DOUGLAS RENFREW P, MULLIGAN VK, KAPPEL K, LABONTE JW, PACELLA MS, BONNEAU R, BRADLEY P, DUNBRACK RL Jr, DAS R, BAKER D, KUHLMAN B, KORTEMME T, GRAY JJ. The Rosetta all-atom energy function for macromolecular modeling and design[J]. Journal of Chemical Theory and Computation, 2017, 13(6): 3031-3048.
[46] SCHYMKOWITZ J, BORG J, STRICHER F, NYS R, ROUSSEAU F, SERRANO L. The FoldX web server: an online force field[J]. Nucleic Acids Research, 2005, 33: W382-W388.
[47] TORNG W, ALTMAN RB. 3D deep convolutional neural networks for amino acid environment similarity analysis[J]. BMC Bioinformatics, 2017, 18(1): 302.
[48] RAO RM, LIU J, VERKUIL R, MEIER J, CANNY J, ABBEEL P, SERCU T, RIVES A. Learning from protein structure with geometric vector perceptrons[C]. International Conference on Learning Representations. 2020.
[49] WANG ZC, COMBS SA, BRAND R, CALVO MR, XU PP, PRICE G, GOLOVACH N, SALAWU EO, WISE CJ, PONNAPALLI SP, CLARK PM. LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction[J]. Scientific Reports, 2022, 12(1): 6832.
[50] ELNAGGAR A, HEINZINGER M, DALLAGO C, REHAWI G, WANG Y, JONES L. Prottrans: toward understanding the language of life through self-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(10): 7112-7127.
[51] HAYES T, RAO R, AKIN H, SOFRONIEW NJ, OKTAY D, LIN ZM, VERKUIL R, TRAN VQ, DEATON J, WIGGERT M, BADKUNDRI R, SHAFKAT I, GONG J, DERRY A, MOLINA RS, THOMAS N, KHAN YA, MISHRA C, KIM C, BARTIE LJ, et al. Simulating 500 million years of evolution with a language model[J]. Science, 2025: eads0018.
[52] NOTIN P, KOLLASCH A, RITTER D, VAN NIEKERK L, PAUL S, SPINNER H, ROLLINS N, SHAW A, ORENBUCH R, WEITZMAN R, FRAZER J, DIAS M, FRANCESCHI D, GAL Y, MARKS D. ProteinGym: large-scale benchmarks for protein fitness prediction and design[J]. Advances in Neural Information Processing Systems, 2024, 36.
[53] DEVLIN J, CHANG M W, LEE K, TOUTANOVA K. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Yechnologies. Volume 1 (long and short papers). 2019: 4171-4186.
[54] CHEN L, ZHANG ZH, LI ZH, LI R, HUO RF, CHEN LF, WANG DY, LUO XM, CHEN KX, LIAO CS, ZHENG MY. Learning protein fitness landscapes with deep mutational scanning data from multiple sources[J]. Cell Systems, 2023, 14(8): 706-721.e5.
[55] BISWAS S, KHIMULYA G, ALLEY EC, ESVELT KM, CHURCH GM. Low-N protein engineering with data-efficient deep learning[J]. Nature Methods, 2021, 18(4): 389-396.
[56] ZHOU ZY, ZHANG L, YU YX, WU BH, LI MC, HONG L, TAN P. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning[J]. Nature Communications, 2024, 15(1): 5566.
[57] HU E J, SHEN Y, WALLIS P, ALLEN-ZHU Z, LI Y, WANG S, WANG L, CHEN W. LoRA: low-rank adaptation of large language models[J/OL]. arXiv, 2021. https://doi.org/10.48550/arXiv.2106.09685.
[58] NOTIN P, WEITZMAN R, MARKS DS, GAL Y. ProteinNPT: improving protein property prediction and design with non-parametric transformers[J]. Advances in Neural Information Processing Systems, 2023, 36: 33529-33563.
[59] JIANG F, LI MC, DONG JJ, YU YX, SUN XY, WU BH, HUANG J, KANG LQ, PEI YF, ZHANG L, WANG SJ, XU WX, XIN JY, OUYANG WL, FAN GS, ZHENG LR, TAN Y, HU ZQ, XIONG Y, FENG Y, et al. A general temperature-guided language model to design proteins of enhanced stability and activity[J]. Science Advances, 2024, 10(48): eadr2641.
[60] CHENG P, MAO C, TANG J, YANG S, CHENG Y, WANG WK, GU QX, HAN W, CHEN H, LI SH, CHEN YF, ZHOU JL, LI WJ, PAN AM, ZHAO SW, HUANG XX, ZHU SQ, ZHANG J, SHU WJ, WANG SQ. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering[J]. Cell Research, 2024, 34: 630-647.
[61] FOWLER DM, FIELDS S. Deep mutational scanning: a new style of protein science[J]. Nature Methods, 2014, 11(8): 801-807.
[62] LIVESEY BJ, MARSH JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations[J]. Molecular Systems Biology, 2020, 16(7): e9380.
[63] SCHMIEDEL JM, LEHNER B. Determining protein structures using deep mutagenesis[J]. Nature Genetics, 2019, 51(7): 1177-1186.