特约评述

DOI: 10.12211/2096-8280.2025-041

DeepSeek模型分析及其在AI辅助蛋白质工程中的应用

李明辰[1,2]，钟博子韬[1]，余元玺[1]，姜帆[1]，张良[1]，谭扬[1,2]，虞慧群[2]，范贵生[2]，洪亮[1]
（[1] 上海交通大学张江高等研究院，上海 201203；[2] 华东理工大学信息科学与工程学院，上海 200237）

摘要： 2025年年初，杭州深度求索人工智能基础技术研究有限公司发布并开源了其自主研发的DeepSeek-R1对话大模型。该模型具备极低的推理成本和出色的思维链推理能力，在多种任务上能够媲美甚至超越闭源的GPT-4o和o1模型，引发了国际社会的高度关注。此外，DeepSeek模型在中文对话上的优异表现以及免费商用的策略，在国内引发了部署和使用的热潮，推动了人工智能技术的普惠与发展。本文围绕DeepSeek模型的架构设计、训练方法与推理机制进行系统性分析，探讨其核心技术在AI蛋白质研究中的迁移潜力与应用前景。DeepSeek模型融合了多项自主创新的前沿技术，包括多头潜在注意力机制、混合专家网络及其负载均衡、低精度训练等，显著降低了Transformer模型的训练和推理成本。尽管DeepSeek模型原生设计用于人类语言的理解与生成，但其优化技术对同样基于Transformer模型的蛋白质预训练语言模型具有重要的参考价值。借助DeepSeek所采用的关键技术，蛋白质语言模型在训练成本、推理成本等方面有望得到显著降低。

关键词： 大语言模型；AI蛋白质；深度自注意力变换网络；蛋白质语言模型；深度学习

中图分类号： Q816　　**文献标志码：** A

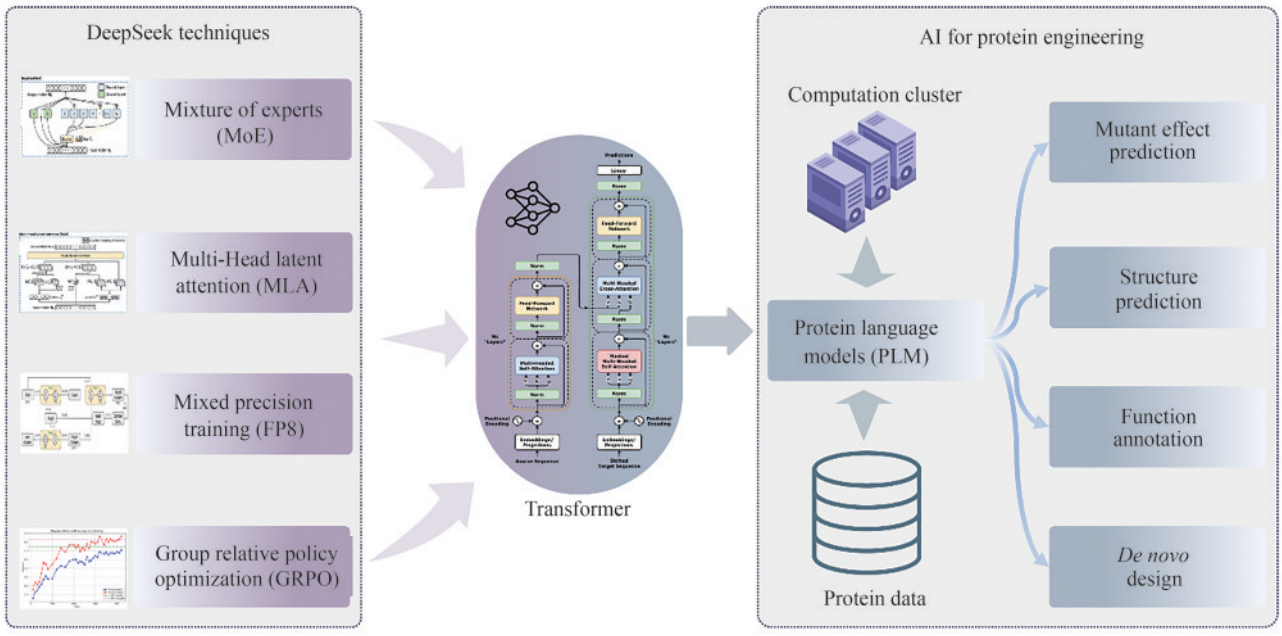
DeepSeek model analysis and its applications in AI-assistant protein engineering

LI Mingchen[1,2]，ZHONG Bozitao[1]，YU Yuanxi[1]，JIANG Fan[1]，ZHANG Liang[1]，TAN Yang[1,2]，YU Huiqun[2]，FAN Guisheng[2]，HONG Liang[1]
（[1]*Zhangjiang Institute for Advanced Study*，*Shanghai Jiao Tong University*，*Shanghai 201203*，*China*；[2]*School of Information Science and Engineering*，*East China University of Science and Technology*，*Shanghai 200237*，*China*）

Abstract: In early 2025, Hangzhou DeepSeek AI Foundation Technology Research Co., Ltd. released and open-sourced its independently developed DeepSeek-R1 conversational large language model. This model exhibits extremely low inference costs and outstanding chain-of-thought reasoning capabilities, performing comparably to, and in some tasks surpassing, proprietary models like GPT-4o and o1. This achievement has garnered significant international attention. Furthermore, DeepSeek's excellent performance in Chinese conversations and its free-for-commercial-use

收稿日期：2025-05-01 修回日期：2025-06-03
基金项目：上海市2023年度“科技创新行动计划”计算生物学重点专项（23JS1400600）
引用本文：李明辰，钟博子韬，余元玺，姜帆，张良，谭扬，虞慧群，范贵生，洪亮. DeepSeek模型分析及其在AI辅助蛋白质工程中的应用[J]. 合成生物学，2025，6(3)：636-650
Citation：LI Mingchen，ZHONG Bozitao，YU Yuanxi，JIANG Fan，ZHANG Liang，TAN Yang，YU Huiqun，FAN Guisheng，HONG Liang. DeepSeek model analysis and its applications in AI-assistant protein engineering[J]. Synthetic Biology Journal，2025，6(3)：636-650

strategy have ignited a wave of deployment and application within China, thereby promoting the widespread adoption and development of AI technology. This work systematically analyzes the architectural design, training methodology, and inference mechanisms of the DeepSeek model, exploring the transfer potential and application prospects of its core technologies in AI-assistant protein research. The DeepSeek model integrates several cutting-edge, independently innovated technologies, including a multi-head latent attention mechanism, mixture-of-experts (MoE) with load balancing, and low-precision training. These innovations have substantially reduced the training and inference costs for Transformer models. Although DeepSeek was originally designed for human language understanding and generation, its optimization techniques hold significant reference value for pre-trained language models with proteins, which are also based on the Transformer architecture. By leveraging the key technologies employed in DeepSeek, protein language models are expected to achieve substantial reductions in training and inference costs.



Keywords: large language models; AI-assistant protein; transformer; protein language model; deep learning

人工智能，尤其是深度学习与大模型的发展，正在深刻变革科学研究的范式。2024年诺贝尔化学奖和物理学奖分别颁发给了人工智能预测蛋白质结构预测与神经网络，标志着人工智能已成为推动基础科学进步的重要技术[1]。在生物工程领域，蛋白质的理解与设计是核心研究方向之一。近年来，受自然语言处理中预训练语言模型的启发，蛋白质预训练语言模型应运而生。该类模型通过对海量未标注蛋白质序列数据进行自监督学习，能够有效捕捉序列中的“语法”与“语义”特征，从而获得高质量的蛋白质表示，并可应用于结构预测、功能注释等下游任务，甚至具备生成全新蛋白质序列的能力[2]。当前主流的蛋白质预训练模型多基于深度自注意力变换网络[3]（Transformer），在模型设计与技术路径上与自然语言处理领域存在着高度的一致性。因此，自然语言模型研究领域针对Transformer所提出的各类优化策略，如注意力机制的改进、高效微调方法等，均可较为便捷地迁移至蛋白质语言模型的研究中。

杭州深度求索（DeepSeek）是国内领先的大语言模型研发企业，其发布的DeepSeek-V3[4]和DeepSeek-R1[5]等高性能语言模型在国内外均引发了广泛关注。这些模型同样基于经典的Transformer架构，并在此基础上引入多项关键技

术改进，包括多头潜在注意力机制（multi-head latent attention， MLA）、负载均衡的混合专家网络（mixture-of-experts， MoE）以及低精度FP8训练技术，有效降低了模型的训练与推理成本，显著提升了模型的计算效率。此外，DeepSeek-R1模型指出，强化学习算法是激发大模型深度思考（reasoning）能力的关键，打破了OpenAI发布的o1模型[6]对大模型深度思考能力的垄断地位，赢得了国际学术界与产业界的高度重视。

本文梳理了DeepSeek系列模型的发展历程及其提出的针对Transformer改进的关键技术，重点探讨其在蛋白质预训练语言模型中的应用潜力，分析相关技术在AI蛋白质领域的适用性、局限性以及未来发展，为蛋白质工程与人工智能交叉领域的研究提供新的思路与参考。

1 DeepSeek模型发展简述

DeepSeek系列重要的模型发展历史可以分为基础探索、深入研究和全面升级三个阶段（图1）。

1.1 基础探索阶段

2024年1月，DeepSeek公司正式发布了其自主研发的DeepSeek系列模型。在该月内，DeepSeek团队接连推出了三款初代模型：DeepSeek-LLM[7]、DeepSeek-Coder[8]和DeepSeekMoE[9]。其中，DeepSeek-LLM作为该系列的首款模型，虽然采用了与Llama模型[10]相同的基础架构，但其技术报告中指出，DeepSeek团队自主研发了大语言模型训练框架HAI-LLM，并依托“萤火超算”平台提供的万卡级算力资源，成功完成了模型的全流程自主训练，DeepSeek-LLM的发布标志着DeepSeek团队拥有了自主开发大语言模型的技术能力。同月发布的DeepSeekMoE则引入了负载均衡的混合专家网络（MoE），通过在推理过程中选择性地激活部分参数（即“专家”模块），在保持甚至提升模型性能的同时，显著降低了计算资源消耗。

1.2 深入研究阶段

2024年2月，DeepSeek团队推出了专注于数学推理任务的DeepSeekMath模型[11]。该模型首次引入了团队自主研发的群体相对策略优化算法（group relative policy optimization， GRPO），对传统强化学习中的PPO优化算法进行了创新性改进。GRPO通过省去额外价值模型的引入，直接使用多个采样输出的平均奖励作为基线，显著降低了训练中计算资源的消耗，在大规模强化学习的应用中实现了关键性的成本优化突破，充分体现了DeepSeek在算法设计和资源效率控制方面的深厚技术积累。2024年5月，DeepSeek发布了DeepSeek-V2模型[12]，首次提出多头潜在注意力（multi-head latent attention， MLA）技术。该技术通过将键值缓存压缩为低维潜在向量，在保持模型生成质量的同时减少了推理过程中的KV-Cache的显存占用，降低了模型的推理成本。随后在2024年9月，DeepSeek推出DeepSeek-V2.5版本[13]，将预训练数据集规模扩展至10万亿高质量token[13]，保留了Chat模型的对话功能和Coder模型的代码处理功能，并更符合人类偏好。随着负载均衡的MoE架构、GRPO算法以及MLA技术的相继推出，DeepSeek团队在降低大模型训练与推理成本的关键技术领域取得了系统性突破，为后续DeepSeek-V3模型的研发奠定了坚实基础。

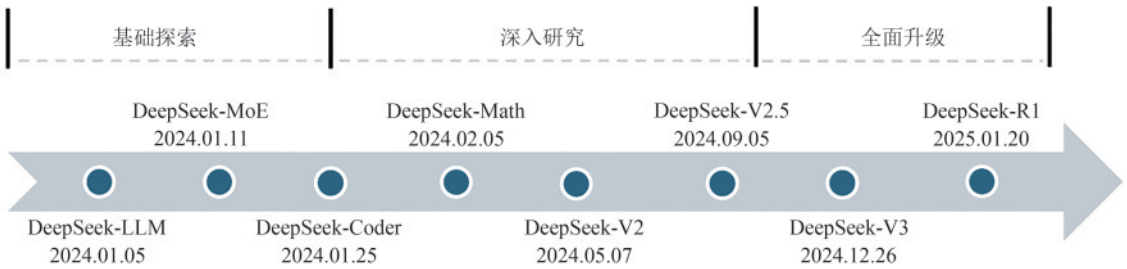


图1 DeepSeek系列模型的发展历史

Fig. 1 History with the development of DeepSeek models

1.3 全面升级阶段

以DeepSeek-V3模型和DeepSeek-R1模型的发布为标志，DeepSeek实现了将前期多项核心技术的整合，不仅显著提升了模型的生成质量，而且加速了模型的推理，相比于同样表现的其他模型具备更低的部署成本，引发了DeepSeek模型私有部署的热潮。DeepSeek-V3模型与DeepSeek-R1模型的主要区别在于R1模型具备深度思考（reasoning）的能力，回答准确度的表现更优。DeepSeek-V3和DeepSeek-R1拥有共同的训练起点——DeepSeek-V3-Base，DeepSeek-V3系列之间的关系如图2所示。

DeepSeek-V3-Base是基础模型，该模型为后续提供了预训练的参数。以DeepSeek-V3-Base模型的预训练参数为起点，DeepSeek-R1-Zero引入了源自DeepSeek-Math提出的GRPO强化学习算法，首次实现了一个具备“深度思考”能力的模型，即模型在回答问题的过程中能够进行自我反思与改进，从而带来显著性能提升。然而，DeepSeek-R1-Zero的指令遵循能力较弱，要求用户输入必须严格遵循预设格式。为了弥补这一不足，开发团队利用DeepSeek-R1-Zero生成的高质量回答数据，并在标注数据上采用监督学习与强化学习相结合的方法，完成了对模型的指令对齐训练，最终得到了指令遵循能力强、回答质量高的DeepSeek-R1。该模型在多种任务上已达到甚至超越OpenAI推出的深度思考模型o1，打破了OpenAI在大模型深度思考能力领域的技术垄断。

基于DeepSeek-V3-Base，结合DeepSeek-R1合成的数据与指令问答数据集，研发团队构建出了面向通用对话场景的DeepSeek-V3。相较于DeepSeek-R1，DeepSeek-V3舍弃了深度思考机制，减少了推理过程中所需的token数量。虽然在部分复杂任务上的表现略逊于DeepSeek-R1，但在大多数常规文本处理任务中，DeepSeek-V3凭借更高的响应速度和更低的资源消耗，成为更具性价比的选择。除了自研模型外，DeepSeek还通过知识蒸馏的方式，将DeepSeek-R1的推理思考能力迁移到其他主流开源模型上。例如，DeepSeek-R1-Distill-Llama-70B是在Llama 70B参数版本基础上蒸馏得到，具备接近GPT-4o的性能水平，且参数规模适中，可在普通服务器设备上部署运行；而DeepSeek-R1-Distill-Qwen-1.5B则主要面向学习和实验用途，参数仅为15亿，可在移动设备或者嵌入式平台上运行，胜任轻量级推理任务。

2 DeepSeek关键技术分析

2.1 潜在注意力

注意力机制作为Transformer架构的核心组件，凭借其高效的建模能力，使得该结构在序列建模任务中长期占据主导地位[14]。然而，标准多头注意力（multi-head attention， MHA）的计算复杂度随生成序列长度呈立方级增长（在应用KV-Cache机制时呈平方级增长），导致其在大规模语言模型推理过程中面临显著的效率瓶颈，尤其是在长序列生成和高并发请求场景下，推理成本高昂的问题尤为突出。为缓解上述问题，当前主流的大语

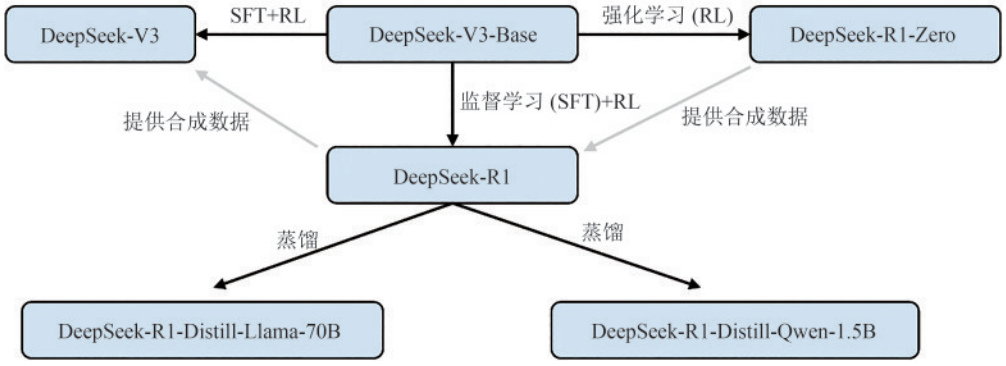


图2 DeepSeek-V3，V3-Base，R1与R1-Zero系列模型之间的关系

Fig. 2 Relationship between DeepSeek-V3, V3-Base, R1 and R1-Zero series models

言模型推理框架广泛采用键值缓存（KV-Cache）机制。KV-Cache通过将每个已生成token对应的键（key）和值（value）向量缓存于内存之中，实现“以空间换时间”的优化策略，显著提升了自回归生成过程的效率。然而，这种机制带来的存储开销同样不容忽视：KV-Cache的内存占用随着输出序列长度线性增长，在生成长文本时可能迅速耗尽显存资源，从而成为限制模型扩展性和服务吞吐量的关键瓶颈。在实际部署中，KV-Cache所占用的内存甚至可能超过模型参数本身的大小[15]。

为降低KV-Cache所带来的存储压力，研究者提出了多种优化方案，其中具有代表性的包括多查询注意力[16]（multi-query attention，MQA）与分组查询注意力[17]（grouped query attention，GQA）。MQA通过不同注意力头之间共享同一份键和值，大幅压缩了KV缓存的规模，但其简化结构可能导致模型生成质量下降；GQA则在此基础上引入分组机制，允许每组查询头共享一组键值头，在一定程度上恢复了模型表达能力，从而在推理效率与生成效果之间取得折中。尽管GQA相比MQA性能更优，但仍难以完全达到原始MHA的生成质量。

针对上述挑战，DeepSeek团队提出了一种新的注意力机制——多头潜在注意力机制（MLA）。该机制通过将键值向量压缩至一个固定维度的隐空间中，并在注意力计算前将其还原至原始表示空间，从而有效降低了KV-Cache的显存占用。实验表明，MLA能够在保持高质量生成的同时，显著减少模型推理时的显存消耗。尽管该方法在训练阶段引入了额外的压缩编码成本，但在推理过程中，得益于矩阵乘法的结合律特性，可以将压缩与注意力计算融合为一次高效的矩阵操作，从而不会增加额外的计算负担。图3以注意力头数64，模型隐层大小8192，层数80的Transformer模型为例展示了不同注意力机制（MQA、GQA与MLA）的KV-Cache显存占用随输出长度变化的趋势，其中GQA的分组数为8，计算公式为DeepSeek-V2中给出的注意力复杂度计算公式[12]。可以看出，随着序列长度的增加，MLA所需的缓存空间明显低于MHA及GQA，略高于MQA。然而，MLA在生成质量方面优于MQA。

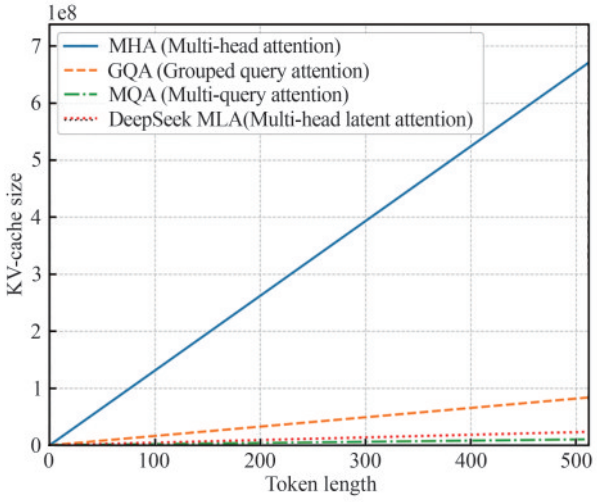


图3 不同注意力机制的KV-Cache内存占用与输出长度的关系示意图

Fig. 3 Comparison of KV-Cache sizes among different attention mechanisms

值得注意的一点是，对于采用双向注意力结构的Transformer模型（如自然语言处理中的BERT模型[18]，或蛋白质语言模型中的ESM-2模型[19]），由于其上下文表征是并行生成的，不依赖于KV-Cache机制，因此应用潜在注意力机制的意义有限，难以实现明显的资源节省效果。

2.2 负载均衡的混合专家网络

扩展定律[20-21]（scaling laws）表明，在模型参数规模不断扩大的过程中，语言模型的性能通常呈现出可预测的提升趋势。然而，参数规模的增长也必然带来计算资源消耗的增加，提高了模型的训练和推理成本。为缓解这一问题，研究者提出了多种混合专家网络（MoE）架构[22-24]。MoE的核心思想在于将模型划分为多个“专家”子网络，并通过一个门控机制（gating network）动态选择激活其中部分专家，从而实现稀疏计算，降低整体计算开销。以DeepSeek-V3为例，其采用的MoE架构包含256个独立专家子网络。在推理阶段，每个输入token仅激活其中8个专家，最多路由至4个计算节点。此外，该模型还引入了共享专家机制，即固定激活若干通用性较强的专家模块，用于捕捉跨任务的共性知识，减少冗余参数的激活频率。

在传统的MoE结构中，负载不均衡问题较为常见：部分专家被频繁调用，而其他专家则利用率较低，影响了计算资源的有效分配。为解决这一问题，DeepSeek提出了一种无辅助损失的负载均衡策略（auxiliary-loss-free balancing strategy）。该策略通过为每个专家引入可学习偏置项，依据其历史激活频率动态调整其被选中的概率，从而实现更均匀的专家激活分布。得益于MoE结构的优势，DeepSeek-V3在维持模型输出质量的同时，拥有着相对较低的训练与推理成本。尽管该模型总参数规模高达6710亿，但在实际生成过程中，每个token的处理仅需激活约370亿参数。其训练一次模型的成本约为278.8万H800 GPU小时（约557.6万美元），远低于相同性能级别的全量参数模型（405 B参数版本的Llama 3的一次训练成本约为3080万GPU小时，约9240万美元至1.23亿美元之间[25]）。

需要指出的是，“混合专家网络”这一术语虽暗示多个“专家”共同参与决策过程，但实际上更多是代表一种拟人化的命名方式。正如“注意力机制”并不意味着模型真正具备人类注意力功能一样，MoE中的“专家”本质上是不同的神经网络模块，其激活与否由门控机制决定。因此，MoE的设计初衷并非模拟多专家协同决策，而是通过稀疏激活机制提升模型的计算效率，从而在大规模模型中实现性能与成本之间的有效平衡。

2.3 低精度训练

在计算机中，实数通常通过浮点数进行近似表示。常见的浮点数格式包括FP64、FP32、FP16[26]以及FP8[27]，其中数字表示该格式所占用的比特位数。例如，FP64和FP32分别占用64和32位的存储空间。通常情况下，位数越多，数值表示的精度越高，但带来的计算开销也越大，运算速度相应下降。FP64是当前具有较高数值精度的标准浮点格式。图4展示了使用FP16、FP8（E4M3版本）表示标准正态分布时相对FP64所产生的误差。可以看出，相比FP16，FP8的表示误差明显更大，精度显著下降，但是其计算成本相对较低。

在以往的大语言模型的训练过程中，为了在精度与效率之间取得平衡，通常采用FP16或针对深度学习定制的BF16格式[28]。由于FP8数值精度较低，普遍认为其不易稳定地训练大模型[29]，主要原因包括数值溢出和量化误差过大等问题。然而，DeepSeek研发团队通过一系列精细化的工程优化策略，成功引入了FP8混合精度训练机制[4]，有效维持了训练过程中的数值稳定性。具体而言，模型将激活值（activations）按1×128元素分组、权重（weights）按128×128元素分组，以降低异常值引起的量化误差；在反向传播过程中，激活值以FP8格式缓存，注意力输出及SwiGLU输入则采用自定义的E5M6格式进行表示；而优化器状态、主权重及其梯度仍保留在FP32精度下，以确保关

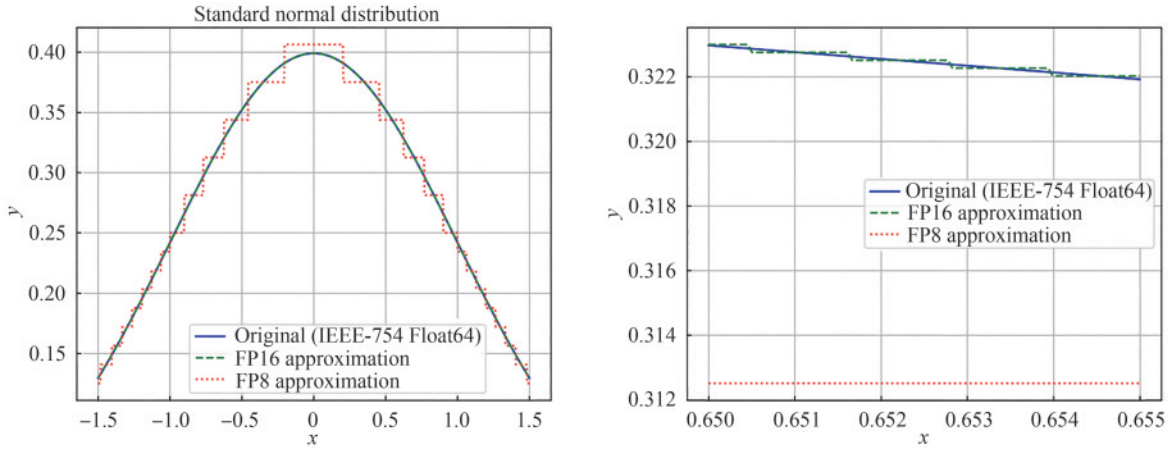


图4 FP8、FP16与FP64表示的标准正态分布

Fig. 4 Standard normal distribution represented by FP8, FP16 and FP64

键部分的稳定性。通过这种设计，DeepSeek研发团队基于FP8成功训练了DeepSeek-V3大模型，显著降低训练与推理成本。

2.4 强化学习与深度思考能力的激发

2024年9月，OpenAI发布了具备深度思考（reasoning）能力的新一代模型o1[6]，其在回答问题前会生成结构化的“思考”过程，展现出类人推理行为，但OpenAI并未公开其训练方法。DeepSeek的研究表明，通过强化学习即可有效激发模型的深度思考能力。其推出DeepSeek-R1-Zero模型在无监督微调数据的前提下，直接对基础模型DeepSeek-V3-Base进行强化学习训练，采用群体相对策略优化算法，并引入基于规则的奖励函数（如准确性与格式规范性）引导模型学习推理能力，验证了无需训练奖励模型，仅依靠人工定义的规则函数即可显著提升模型的深度思考能力。训练过程中，模型展现出若干高级推理行为，包括：“自我进化”，即在无监督条件下自发延长思考时间并反思解题路径；以及“顿悟时刻”，即通过重新评估初始方法来分配更多思考资源解决问题，表现出类似人类的认知跃迁。谷歌后续研究进一步证实，强化学习有助于模型实现知识泛化，而监督学习则更倾向于记忆拟合。这些发现表明，强化学习在增强大语言模型推理能力方面具有巨大潜力，为构建具备深度思考能力的智能系统提供了新的技术路径。

3 DeepSeek技术在蛋白质语言模型的应用

3.1 蛋白质语言模型概述

蛋白质语言模型借鉴了自然语言处理的预训练技术，通过使用神经网络大型的蛋白质序列数据库上以无监督训练的方式学习蛋白质序列的分布与排列规律，从而理解序列的深层表征和进化模式，以及生成新的蛋白质序列。

从应用场景来看，可以将蛋白质语言模型分为理解模型、生成模型两大类。理解模型侧重于计算蛋白质序列的深层表征（即通过数学形式量化序列特征），能够将氨基酸序列转化为可被计算机识别的编码，为后续的功能预测、结构分析等提供数据基础；生成模型则基于学习到的序列规律进行蛋白质序列的创新，通过模拟生物演化规律、结构逆折叠规律或结合人工设计目标，生成具备特定功能的蛋白质序列。此外，还有部分蛋白质语言模型专门被设计用于蛋白质突变预测，即根据预训练过程中学习到的蛋白质序列中的氨基酸分布规律，预测氨基酸的进化带来的影响。表1列举了目前较为常见的蛋白质语言模型的架构、参数量、预训练任务和应用场景。

3.2 DeepSeek技术在蛋白质语言模型中的应用

目前，蛋白质预训练语言模型大多基于Transformer架构及其变体，因此DeepSeek模型中

表1 目前常见的蛋白质语言模型的架构、参数量、预训练任务和应用场景

Table 1 Architecture, parameter count, pre-training task, and application scenario of currently available protein language models

模型	架构	参数量	预训练任务	应用场景
UniRep[30]	LSTM	18.2 M	自回归生成	理解
ESM-2[19]	Transformer	650 M/15 B	自回归生成	理解
ESM-3[31]	Transformer	98 B	掩码预测	生成
ProGEN[32]	Transformer	6.4 B	自回归生成	生成
xTrimoPGLM[33]	Transformer	100 B	掩码预测 & 自回归生成	生成
ProtT5[34]	Transformer	11 B	掩码预测	理解
SaProt[35]	Transformer	650 M	掩码预测	理解
ProSST[36]	Transformer	110 M	掩码预测	理解
ESM-1v[37]	Transformer	650 M	掩码预测	突变预测
Tranception[38]	Transformer	700 M	自回归生成	突变预测

提出的大部分关键技术可被直接借鉴并应用于蛋白质语言模型的优化中，降低其训练与推理成本。DeepSeek所采用的强化学习算法，具备在多种下游任务中提升语言模型性能的潜力，有望拓展至蛋白质序列建模与功能预测等生物应用场景。表1列举的模型中，用于理解的模型引用量最高的为ESM-2，用于生成的模型引用量最高的为ProGEN模型（ESM-2被引用3089次，ProGEN被引用1210次，数据来源Google Scholor，截至2025年5月31日）。因此，本节以引用量较高的ESM-2[19]和ProGEN[32]两种具有代表性的蛋白质语言模型为例，分析DeepSeek相关技术在该领域的可行性与应用价值。

ESM-2是由Meta公司Lin等[19]提出的一种基于掩码语言建模（masked language modeling，MLM）的蛋白质语言模型，专注于蛋白质表示学习与结构预测等下游任务，是目前应用广泛的AI蛋白质表示模型之一。作为典型的Transformer编码器架构，ESM-2基于双向自注意力机制构建，最大版本参数量约为150亿。由于其推理过程采用并行解码方式，能够一次性完成整个蛋白质序列的编码，无需依赖KV-Cache等机制，因此DeepSeek提出的多头潜在注意力技术在该模型中的应用空间较为有限。混合专家网络及其负载均衡策略仍可在大规模特征提取场景下发挥作用，通过引入稀疏激活机制降低推理成本，但是这需要对模型进行重新训练。此外，FP8低精度训练作为一种基础性计算优化手段，可直接应用于ESM-2，有效降低计算资源消耗。

ProGEN是一种基于自回归语言建模（auto-regressive language modeling）的蛋白质生成模型，由Madani等提出[32]，主要用于蛋白质与酶的设计与生成，并已在实验层面验证了其生成序列的功能性。ProGEN模型采用Transformer解码器结构，最大版本参数量约为12亿。由于其推理过程中采用串行解码方式，每一步生成一个氨基酸残基，涉及重复的注意力计算，DeepSeek提出的多头潜在注意力机制在此类模型中具有良好的适配性，有助于提升蛋白质序列生成效率。同时，混合专家网络及其负载均衡策略也可在大规模候选蛋白生成任务中发挥降本增效作用，但同样需通过对ProGEN模型重新训练。类似地，FP8低精度训练同样可以集成至ProGEN训练与推理过程中，降低计算开销。

总体而言，DeepSeek所提出的模型优化技术在蛋白质语言模型中的应用价值与其具体用途相关。对于以表征学习为主的模型（如ESM-2），其优化空间相对有限；而对于以生成任务为核心的模型（如ProGEN），这些技术展现出较高的实用价值，特别是在需要生成大量候选蛋白质序列的场景下，能够在不牺牲模型性能的前提下显著降低计算成本，具有工程应用价值。

3.3 强化学习及GRPO算法的应用

强化学习是一种通过与环境交互学习优化决策的机器学习算法，善于处理较为复杂的决策问题，例如序列生成、创新策略发现等任务。研究表明，强化学习可用于涉及复杂的蛋白质结构和序列。面向蛋白质结构设计，华盛顿大学David Baker教授团队的Lutz等的研究[39]展示了强化学习在设计复杂蛋白质结构中的应用。他们首先根据目标的功能，通过专家的经验设计了多种针对蛋白质结构的约束条件，并给出了奖励函数。随后他们使用基于蒙特卡洛树搜索（MCTS）的强化学习方法优化结构生成模型，最终设计出了具有预定功能约束的蛋白质结构，包括盘状纳米孔和超紧凑二十面体。冷冻电镜验证表明，设计结构与计算模型高度一致。该模型体现了强化学习在复杂场景下的效果，适用于疫苗开发和信号分子展示。面向蛋白质序列设计，Wang等提出了EvoPlay模型[40]用于通过突变优化蛋白质的序列。EvoPlay首先利用大量的绿色荧光蛋白突变后的荧光强度数据，训练了一个较为准确的代理模型，用于提供奖励，随后通过自我对弈和蒙特卡洛树搜索在蛋白质序列空间中不断搜索和优化序列，最终使模型能够生成奖励较高的序列。实验数据表明，该模型设计出的36个突变体中有26个比野生型突变体发出更强的生物荧光，其中最好的突变体荧光强度比野生型提高了6倍。该模型体现了强化学习在序列突变设计和序列决策场景下的效果，适用于蛋白质和酶的定向进化。GRPO算法作

为强化学习算法的一种，虽然目前未被直接应用于蛋白质设计，但是该算法本身能够减少训练所需的计算资源，提升模型的奖励期望，在基于强化学习的蛋白质序列和结构设计方面具有广阔的应用空间。

4 应用局限性分析

DeepSeek模型提出的一系列优化技术主要通过对Transformer架构进行改进，在训练与推理阶段显著降低了大规模语言模型的计算成本。正如第3节中的分析，这些技术原则上也可被引入蛋白质语言模型领域，用于提升其效率与性能。然而，由于蛋白质语言模型的应用场景、用户规模以及任务特性与通用自然语言模型存在显著差异，DeepSeek的相关技术在该领域的适用性仍存在一定局限。DeepSeek提出的优化路径是否适用于蛋白质语言模型，主要取决于两个关键假设：

假设1：蛋白质语言模型的性能随参数量、数据量的增加而提升，即其满足扩展定律（scaling laws）并具有涌现（emergent）现象。

假设2：蛋白质语言模型具有较高的推理计算需求，模型推理速度优化具备实际意义。

以下将围绕两个假设对DeepSeek模型优化技术在蛋白质语言模型中应用的局限性展开分析。

4.1 扩展定律与涌现现象

在自然语言处理大模型领域，扩展定律[20]（scaling laws）和涌现[41]（emergent）现象是推动大模型参数量愈来愈多的基础。扩展定律表明，随着模型参数量、训练数据量及计算资源的增加，语言模型的性能会呈现出可预测的、持续性的提升。而“涌现”则描述了一种非线性增强行为——当模型参数达到某一临界点后，其在某些下游任务上会突然展现出小模型不具备的能力。自然界中也有类似的行为：当温度达到某一临界点后，一些材料会突然出现电阻消失的现象。扩展定律和涌现现象共同支撑了“只要资源足够，模型就更强”的理念，推动业界构建参数量更大的语言模型。

在蛋白质语言模型中，已有研究初步验证了扩展定律的存在性。例如，Cheng等[42]分别对自回归式、掩码建模式架构下的蛋白质语言模型进行了扩展行为分析，结果普遍显示：预训练任务的损失确实随模型参数量和训练计算量的增加而下降，表明蛋白质语言模型确实存在着扩展定律。然而，这种训练损失的降低并不一定转化为下游任务性能的提升。即在蛋白质相关的下游任务中，模型性能与参数量之间并非一致正相关关系：Cheng等[42]在xTrimoPGLM的评估中发现，仅约44%的下游任务性能随模型参数量的增大、预训练损失降低而提升。Cheng等还发现，甚至有12%的任务甚至出现模型参数量越大下游任务性能表现越差的现象。Hesslow等[43]在提出的RITA模型的研究中发现，在酶功能预测和突变功能预测任务中，模型性能随着参数量的增大呈渐进式增长，未表现出“涌现”现象。Lin等[19]评估ESM-2模型的表征能力时发现，尽管模型参数量增加带来精度提升，但提升幅度缓慢，仍为渐进式变化。此外，Vieira等[44]的研究指出，在数据受限条件下，部分中小模型的表现优于大参数模型，说明数据质量可能比模型参数量更为关键。

综上所述，当前尚未有证据表明蛋白质语言模型存在“涌现”现象。因此，“模型越大越好”的假设尚未得到充分验证，这表明蛋白质语言模型的构建仍有较大的研究与优化空间。此外，蛋白质序列与自然语言文本之间的本质差异可能会影响扩展定律和涌现现象的可迁移性。主要差异包括：蛋白质语言模型的词表大小通常远小于自然语言模型词表的大小；蛋白质具有三维结构，序列上相距较远的氨基酸在空间上距离可能很近，存在长程依赖效应；下游任务的标记数据稀疏性；实验数据本身的系统性误差。这些因素可能导致仅通过扩大模型规模难以提升效果，因为在数据受限的情况下，更大的模型未必表现更好。因此，虽然扩展模型规模可以带来益处，但在设计和评估蛋白质语言模型时，需充分考虑这些独特的特性。

4.2 推理成本占比问题

蛋白质语言模型生成或编码一条序列所需的计算机运行成本低于实验室中实现表达、纯化与

功能检测的成本。因此，在当前AI蛋白质设计流程中，模型推理成本在整个研发链条中的占比较低，似乎表明**假设2**同样不成立，限制了DeepSeek提出的推理优化技术的直接应用空间。然而，随着模型即服务[45]（model-as-a-service，MaaS）模式在蛋白质建模领域的推广，DeepSeek的优化技术能够发挥较大的作用，特别是在高并发、多用户使用场景下，推理效率的提升将有助于显著降低服务器运行与维护成本。

4.3 数据屏障问题

自然语言和蛋白质语言模型的研究都表明，模型扩展必须伴随着高质量数据的同步增长。虽然近年来测序技术的发展为蛋白质语言模型提供了大量潜在训练数据，但与自然语言动辄数十万亿的数据量相比，蛋白质语言模型所能使用的高质量序列仍有限。更严重的问题在于，不同数据库对蛋白质序列的定义标准尚未统一。例如，RefSeq数据库[46]根据RNA确定蛋白序列，而其他数据库，例如Ensembl数据库[47]，可能依据DNA或实验数据进行识别。在UniProt数据库[48]中，真正经过实验验证的蛋白质序列仅有约百万条[49]，其余数亿条多源于宏基因组拼接或同源比对，缺乏足够的功能验证支持。这种数据数量与质量的双重瓶颈，严重制约了蛋白质语言模型的发展。而DeepSeek提出的模型优化策略难以解决数据稀缺的问题。

4.4 奖励函数设计的挑战

DeepSeek模型通过强化学习激发模型深度思考能力的成功经验，为蛋白质语言模型的训练提供了一种新思路。然而，将强化学习应用于蛋白质生成任务的关键难点在于奖励函数的设计。目前，常见的奖励函数包括基于规则的奖励函数、基于模拟计算的奖励函数以及基于模型的奖励函数三类。基于规则的奖励函数依赖人工专家知识定义评分标准，设计难度大。而基于模拟计算的奖励函数通常依赖分子动力学模拟或能量计算等方法评估生成序列的功能性。然而，这类方法在大规模强化学习中因计算开销过高而不具实用性；若采用简化近似，则可能失去指导意义。基于模型的奖励函数通过额外训练一个打分模型来为生成序列打分。此类方法依赖大量标注数据。因此，DeepSeek提出的GRPO算法，或者强化学习算法本身，在应用于蛋白质语言模型的优化方面仍有较大的进步空间。

5 其他增强蛋白质语言模型的方法

除优化蛋白质的训练和推理成本之外，还可以从其他多个层面提升蛋白质语言模型的性能。如图5所示，目前主流的研究方向主要集中在两个方面：引入外部知识增强蛋白质语言模型或者是通过改进内部的模型架构提升蛋白质语言模型的性能。以下对这两种方法做具体的分析。

5.1 引入外部知识增强蛋白质语言模型

引入外部知识指通过向蛋白质语言模型添加其他类型的特征输入或者预训练任务，增强蛋白

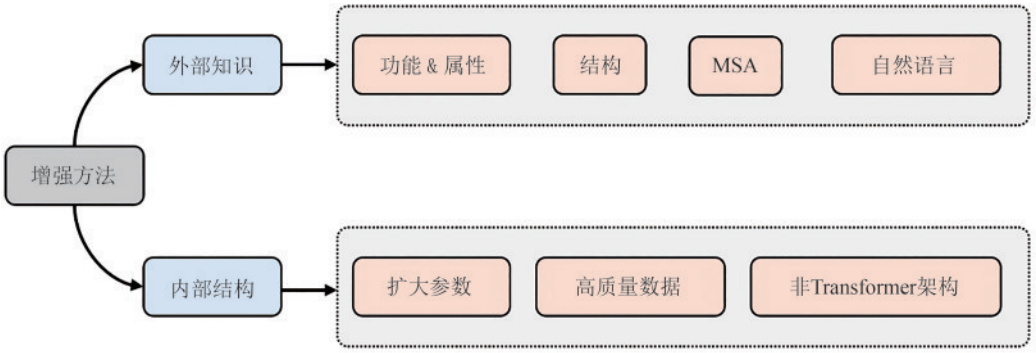


图5 增强蛋白质语言模型方法

Fig. 5 Methods for enhancing protein language models

质语言模型的表示能力。外部知识包括：①蛋白质的功能属性标签；②蛋白质的结构；③多序列比对；④自然语言。在功能属性标签层面，蛋白质基因本体注释（GO）、蛋白质的温度属性，甚至是物理模型计算出的蛋白质的理化性质均可以作为额外的知识用于增强蛋白质语言模型。蛋白质的结构决定了蛋白质的功能，因此使用结构信息增强蛋白质语言模型的性能是一个非常热门的研究，可以使用蛋白质的二级结构、三级结构或者是序列化的结构来增强蛋白质语言模型的性能。此外，自然语言作为一种人类使用的对蛋白质功能的理解符号，也可以用于增强蛋白质语言模型的性能，尤其是引入蛋白质的功能约束，例如ESM-3模型就采用了此方案。

外部知识整合进模型的方式主要包括通过多任务学习作为学习目标（方式1）、作为模型额外的输入（方式2）两种。其中方式1将多任务学习作为学习目标的优势在于，仅需要在训练模型时提供标签即可，在推理时无需提供标签。因此，大多数基于功能属性注释增强的蛋白质语言模型均属于方式1，因为推理的场景下序列不一定会有标签。而方式2的优势在于引入的方式更为直接，在推理时模型能够高效利用推理数据的外部知识。因此，结构和MSA这种能够通过AlphaFold预测和序列比对软件获取的知识大多数可以作为额外的整合进入模型，以达到更好的模型性能。表2列举了一些目前常见的引入外部知识增强的蛋白质语言模型。

表2 引入外部知识增强的蛋白质语言模型

Table 2 Protein language models enhanced by introducing external knowledge

模型	引入的外部知识	模型整合的方式	应用场景
ProteinBERT[50]	结构、功能	训练目标	理解
PromptProtein[51]	结构、属性	训练目标	理解
OntoProtein[52]	结构	训练目标	理解
MELT[53]	功能	训练目标	理解
Regression Transformer[54]	功能、属性	训练目标	理解
Prime[55]	温度	训练目标	突变设计
PeTriBERT[56]	结构	额外输入	理解
MIF-ST[57]	结构	额外输入	理解
ProteinMPNN[58]	结构	额外输入	生成（逆折叠）
ESM-IF[59]	结构	额外输入	生成（逆折叠）
ESM-S[60]	结构	额外输入	理解
ProstT5[34]	结构	训练目标	理解
LM-Design[61]	结构	额外输入	生成
SaProt[35]	结构	额外输入	理解
PST[62]	结构	额外输入	理解
SES-Adapter[63]	结构	额外输入	理解
ProSST[36]	结构	额外输入	理解
AlphaMissense[64]	结构、MSA	训练目标	突变设计
PoET[65]	结构、MSA	额外输入	理解
ProMEP[66]	结构	额外输入	突变设计

5.2 通过内部架构改进提升蛋白质语言模型

在内部架构方面，可以从应用扩展定律、构建高质量数据集与使用非Transformer架构三个方面来改进蛋白质语言模型。

根据扩展定律，扩大模型的参数可以获得预期的性能提升。目前，蛋白质语言模型中最大的模型为百图生科Chen等提出的xTrimoPGLM模型[33]，包含1000亿个参数。其次，EvolutionaryScale公司推出的ESM-3模型，包含了980亿个参数。蛋白质语言模型已经迈进千亿级别。但是训练这些模型往往需要海量算力支撑，耗费大量的资金，因此资源不足的情况下难以从事此类研究工作。

在构建高质量的数据集方面，Fournier等[49]提出，参数量大的模型不一定表现得好，开发更好的蛋白质语言模型除了扩大参数规模外，还可以通过构造高质量的数据来提升蛋白质语言模型。其开发的AMPLIFY模型的预训练数据集来源于UniRef100、SCOP与OAS[67]等多个数据库，能够提升蛋白质语言模型的表示能力，其提出，在训练模型时使用降重的方法对性能有害。在参数量大小远小于ESM-2的情况下，其模型完成了对ESM-2性能的超越。

在非Transformer架构方面，研究者针对超长序列的问题，提出了多种架构。例如，Nguyen等提出了Evo模型[68]。该模型尽管同样采用自回归语言模型，其网络架构采用了StripedHyena架构，在长序列的建模上优于Transformer模型，在基因组预训练模型上达到了较好的性能。实验结果表明，经过在CRISPR类蛋白上进一步微调后，Evo模型能够生成新的CRISPR-Cas蛋白。

6 总结与展望

蛋白质语言模型作为AI蛋白质设计领域的重要工具，正逐步展现出其在多种应用场景中的核心价值。在蛋白质功能注释方面，这类模型通过学习序列的深层隐空间表征，有效摆脱了对传统手工特征工程的依赖，其所生成的表示能够捕捉序列中蕴含的潜在生物学信息，并在多项功能预测与注释任务中展现出显著优势。在蛋白质及酶的工程改造中，蛋白质语言模型借助大规模预训练，掌握了自然序列的分布规律，从而理解进化过程中所体现的选择偏好[37, 69-70]。这种能力使得模型能够从最大似然估计的角度识别潜在有益突变，为功能优化提供有效指导，推动蛋白质向更具生物合理性的方向演化，缩小了搜索空间，加速了实验进程。例如，多项研究工作已经表明蛋白质语言模型能够提供有效的突变位点和突变后的氨基酸，提升酶活性和酶的稳定性。

在从头蛋白质设计任务中，蛋白质语言模型亦显示出强大潜力。随着“扩展定律”的验证，模型规模不断增大，逐步具备根据功能提示生成具有特定生物功能蛋白质序列的能力。在酶的发现与功能挖掘中，蛋白质语言模型通过计算蛋白质序列的高维特征嵌入（Embeddings），能够发现远程同源性。这些高维特征蕴含着蛋白质的结构与功能属性。因此，在高维特征空间中的距离远近，能够反映蛋白质在功能上的亲疏关系，为识别功能相似但序列分化显著的酶提供了重要线索。强化学习技术可优化人工智能模型，以设计具有特定结构和功能的蛋白质，通过模拟分子相互作用和折叠过程，为疫苗开发等生物医学领域提供关键应用。

近年来，DeepSeek系列模型提出了一系列降低Transformer模型训练与推理成本的优化技术，其中多项方法可迁移至蛋白质语言模型中加以应用。例如，混合专家网络及其负载均衡策略、多头隐注意力机制等，有望降低蛋白质语言模型在生成大量候选序列时的计算开销；低精度训练技术则有望降低模型的训练和推理计算成本。

尽管如此，DeepSeek系列技术在蛋白质语言模型中的应用仍面临若干挑战。首先，虽然AI辅助蛋白质设计在实验层面显著缩小了搜索空间，降低了实验成本，但在大多数实际场景中，计算开销仍远低于湿实验验证的成本。因此，与自然语言处理领域不同，模型推理效率的提升在此并非首要目标。在蛋白质设计任务中，模型的精度往往比推理速度和计算成本更为关键。其次，目前尚无明确证据表明蛋白质语言模型中存在“涌现现象”，即模型性能不会在参数规模扩大时出现非线性跃升。事实上，模型性能通常随参数量增长呈现渐进式变化，甚至在某些任务中出现“反涌现”现象，即大模型的性能反而弱。在下游任务中，数据质量的重要性远超模型的参数量，表明高质量数据仍是提升AI蛋白质设计效果的核心因素，而这一点是DeepSeek类技术难以解决的问题，尤其是其难以解决生物实验的系统性误差问题。

当前，蛋白质语言模型仍有多个方面亟待改进：一方面，应进一步挖掘现有生物数据库的潜力，扩大高质量数据资源的覆盖范围；另一方面，通过引入多模态建模手段（如整合结构、功能和进化信息），有望提升模型的表示能力和生成性能。此外，还需积极探索蛋白质语言模型的独特应用场景。目前多数任务仍未能充分体现其相较于传统生物信息学方法的不可替代性，因此，发掘并强化其独特优势将是未来研究的重要方向。

综上所述，尽管蛋白质语言模型已在多个方向取得积极进展，但其“ChatGPT时刻”尚未真正到来。如何进一步提升模型性能、拓展应用边界、挖掘其潜在科学价值，仍是未来AI辅助蛋白质设计领域的重要研究课题。这一进程不仅将推动合成生物学的发展，也将为生命健康等相关领域的创新与突破提供有力支撑。

参 考 文 献

[1] 余元玺, 钟博子韬, 洪亮. 人工智能的诺奖时刻: 重塑科学的未来[J]. 物理, 2025, 54(01): 25-29.

[2] FAN W Q, ZHOU Y, WANG S J, et al. Computational protein science in the era of large language models (LLMs)[EB/OL]. arXiv, 2025: 2501.10282. (2025-01-25)[2025-06-03]. https://arxiv.org/abs/2501.10282v2.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C/OL]// Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017: 5998-6008[2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2017/

hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[4] DeepSeek-AI, LIU A X, FENG B, et al. DeepSeek-V3 technical report[EB/OL]. arXiv, 2024: 2412.19437v1. (2024-12-27)[2025-06-03]. https://doi.org/10.48550/arXiv.2412.19437.

[5] GUO D Y, YANG D J, ZHANG H W, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs *via* reinforcement learning[EB/OL]. arXiv, 2025: 2501.12948. (2025-01-22) [2025-06-03]. https://arxiv.org/abs/2501.12948v1.

[6] JAECH A, KALAI A, LERER A, et al. Openai o1 system card [EB/OL]. arXiv, 2024: 2412.16720. (2024-12-21) [2025-06-03]. https://doi.org/10.48550/arXiv.2412.16720.

[7] BI X, CHEN D L, CHEN G T, et al. DeepSeek LLM: scaling open-source language models with longtermism[EB/OL]. arXiv, 2024: 2401.02954. (2024-01-05) [2025-06-03]. https://arxiv.org/abs/2401.02954v1.

[8] GUO D Y, ZHU Q H, YANG D J, et al. DeepSeek-Coder: when the large language model meets programming—the rise of code intelligence[EB/OL]. arXiv, 2024: 2401.14196. (2024-01-26)[2025-06-03]. https://arxiv.org/abs/2401.14196v2.

[9] DAI D M, DENG C Q, ZHAO C G, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models[C/OL]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand. Stroudsburg, PA, USA: ACL, 2024: 1280-1297[2025-06-03]. https://doi.org/10.18653/v1/2024.acl-long.70.

[10] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[EB/OL]. arXiv, 2023: 2307.09288. (2023-07-18) [2025-06-03]. https://doi. org/10.48550/arXiv.2307.09288.

[11] SHAO Z H, WANG P Y, ZHU Q H, et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models[EB/OL]. arXiv, 2024: 2402.03300. (2024-02-05)[2025-06-03].https://arxiv.org/abs/2402.03300v3.

[12] DEEPSEEK-AI, LIU A X, FENG B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model[EB/OL]. arXiv, 2024: 2405.04434. (2024-06-19) [2025-06-03].https://arxiv.org/abs/2405.04434v5.

[13] DeepSeek-V2.5: a new open-source model combining general and coding capabilities[EB/OL]. (2024-09-05) [2025-06-03]. https://api-docs.deepseek.com/news/news0905.

[14] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[EB/OL]. arXiv, 2023: 2303.18223. (2025-03-11)[2025-06-03]. https://doi.org/10.48550/arXiv.2303.18223.

[15] KWON W, LI Z H, ZHUANG S Y, et al. Efficient memory management for large language model serving with PagedAttention[C/OL]//Proceedings of the 29th Symposium on Operating Systems Principles. October 23-26, 2023, Koblenz, Germany. ACM, 2023: 611-626[2025-06-03]. https://doi.org/10.1145/3600006.36131.

[16] SHAZEER N. Fast transformer decoding: one write-head is all you need[EB/OL]. arXiv, 2019: 1911.02150. (2019-11-06) [2025-06-03].https://arxiv.org/abs/1911.02150v1.

[17] AINSLIE J, LEE-THORP J, DE JONG M, et al. GQA: training generalized multi-query transformer models from multi-head checkpoints[C/OL]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, 2023: 4895-4901[2025-06-03]. https://doi. org/10.18653/v1/2023.emnlp-main.298.

[18] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C/OL]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota, USA: ACL, 2019: 4171-4186[2025-06-03]. https://doi.org/10.18653/v1/N19-1423.

[19] LIN Z M, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[20] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB/OL]. arXiv, 2020: 2001.08361. (2020-01-23) [2025-06-03]. https://arxiv. org/abs/2001.08361v1.

[21] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[EB/OL]. arXiv, 2022: 2203.15556. (2022-03-29) [2025-06-03]. https://doi.org/10.48550/arXiv.2203.15556.

[22] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79-87.

[23] JORDAN M I, XU L. Convergence results for the EM approach to mixtures of experts architectures[J]. Neural Networks, 1995, 8(9): 1409-1431.

[24] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[C/OL]// 5th International Conference on Learning Representations ICLR 2017. (2017-02-06) [2025-06-03]. https://openreview.net/forum?id=B1ckMDqlg.

[25] WEI MING T. DeepSeek V3 Training cost: here's how it compares to Llama 3.1 (405B) [EB/OL]. (2025-01-26) [2025-06-03]. https://apxml. com/posts/training-cost-deepseek-v3-vs-llama-3.

[26] KAHAN W. IEEE standard 754 for binary floating-point arithmetic[EB/OL]. Lecture Notes on the Status of IEEE, 1996, 754(94720-1776): 11[2025-06-03]. https://people.eecs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF.

[27] MICIKEVICIUS P, STOSIC D, BURGESS N, et al. FP8 formats for deep learning[EB/OL]. arXiv, 2022: 2209.05433.

(2022-09-12) [2025-06-03]. https://doi. org/10.48550/arXiv.2209.05433.

[28] ZAMIRAI P, ZHANG J, ABERGER C R, et al. Revisiting BFloat16 training[EB/OL]. arXiv, 2020: 2010.06192. (2020-10-13)[2025-06-03]. https://arxiv.org/abs/2010.06192v2.

[29] FUJII K, NAKAMURA T, YOKOTA R. Balancing speed and stability: the trade-offs of FP8 *vs*. BF16 training in LLMs[EB/OL]. arXiv, 2024: 2411.08719. (2024-11-01) [2025-06-03]. https://arxiv.org/abs/2411.08719v1.

[30] ALLEY E C, KHIMULYA G, BISWAS S, et al. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature Methods, 2019, 16(12): 1315-1322.

[31] HAYES T, RAO R, AKIN H, et al. Simulating 500 million years of evolution with a language model[J]. Science, 2025, 387(6736): 850-858.

[32] MADANI A, KRAUSE B, GREENE E R, et al. Large language models generate functional protein sequences across diverse families[J]. Nature Biotechnology, 2023, 41(8): 1099-1106.

[33] CHEN B, CHENG X Y, LI P, et al. xTrimoPGLM: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins[J]. Nature Methods, 2025, 22(5): 1028-1039.

[34] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: toward understanding the language of life through self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7112-7127.

[35] SU J, HAN C C, ZHOU Y Y, et al. SaProt: protein language modeling with structure-aware vocabulary[C/OL]// The Twelfth International Conference on Learning Representations ICLR 2024, 2024[2025-06-03]. https://openreview. net/forum?id=6MRm3G4NiU.

[36] LI M C, TAN Y, MA X Z, et al. ProSST: protein language modeling with quantized structure and disentangled attention [C/OL]// Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024: 35700-35726[2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2024/hash/3ed57b293db0aab7cc30c44f45262348-Abstract-Conference.html.

[37] MEIER J, RAO R, VERKUIL R, et al. Language models enable zero-shot prediction of the effects of mutations on protein function[C/OL]// Advances in Neural Information Processing Systems 34 (NeurIPS 2021), 2021: 29287-29303 [2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

[38] NOTIN P, DIAS M, FRAZER J, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval[C/OL]// Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, 162: 16990-17017[2025-06-03]. https://proceedings. mlr. press/v162/notin22a.html.

[39] LUTZ I D, WANG S Z, NORN C, et al. Top-down design of protein architectures with reinforcement learning[J]. Science, 2023, 380(6642): 266-273.

[40] WANG Y, TANG H, HUANG L C, et al. Self-play reinforcement learning guides protein engineering[J]. Nature Machine Intelligence, 2023, 5(8): 845-860.

[41] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J/OL]. Transactions on Machine Learning Research, 2022. (2022-08-31) [2025-06-03]. https://openreview.net/forum?id=yzkSU5zdwD.

[42] CHENG X Y, CHEN B, LI P, et al. Training compute-optimal protein language models[C/OL]// Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024 [2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2024/hash/8066ae1446b2bbccb5159587cc3b3bcc-Abstract-Conference.html.

[43] HESSLOW D, ZANICHELLI N, NOTIN P, et al. RITA: a study on scaling up generative protein sequence models [EB/OL]. arXiv, 2022: 2205.05789. (2022-07-14) [2025-06-03]. https://arxiv.org/abs/2205.05789v2.

[44] VIEIRA L C, HANDOJO M L, WILKE C O. Scaling down for efficiency: medium-sized protein language models perform well at transfer learning on realistic datasets[EB/OL]. bioRxiv, 2025: 2024.11.22.624936. (2025-05-08) [2025-06-03]. https://doi.org/10.1101/2024.11.22.624936.

[45] GAN W S, WAN S C, YU P S. Model-as-a-service (MaaS): a survey[C/OL]//2023 IEEE International Conference on Big Data (BigData). December 15-18, 2023, Sorrento, Italy. IEEE, 2023: 4636-4645[2025-06-03]. https://ieeexplore. ieee. org/document/10386351.

[46] GOLDFARB T, KODALI V K, PUJAR S, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation[J]. Nucleic Acids Research, 2025, 53(D1): D243-D257.

[47] DYER S C, AUSTINE-ORIMOLOYE O, AZOV A G, et al. Ensembl 2025[J]. Nucleic Acids Research, 2025, 53(D1): D948-D957.

[48] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025[J]. Nucleic Acids Research, 2025, 53 (D1): D609-D617.

[49] FOURNIER Q, VERNON R M, VAN DER SLOOT A, et al. Protein language models: is scaling necessary? [EB/OL]. bioRxiv, 2024: 09.23.614603. (2024-09-23)[2025-06-03]. https://doi.org/10.1101/2024.09.23.614603.

[50] BRANDES N, OFER D, PELEG Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function [J]. Bioinformatics, 2022, 38(8): 2102-2110.

[51] WANG Z Y, ZHANG Q, HU S W, et al. Multi-level protein structure pre-training *via* prompt learning[C/OL]// The Eleventh International Conference on Learning Representations ICLR

2023. (2023-02-02) [2025-06-03]. https://openreview. net/forum?id=XGagtiJ8XC.

[52] ZHANG N Y, BI Z, LIANG X Z, et al. OntoProtein: protein pretraining with gene ontology embedding[C/OL]// The Tenth International Conference on Learning Representations ICLR 2022. (2022-01-29)[2025-06-03]. https://openreview.net/forum?id=yfe1VMYAXa4.

[53] GELMAN S, JOHNSON B, FRESCHLIN C, et al. Biophysics-based protein language models for protein engineering[EB/OL]. bioRxiv, 2025: 2024.03.15.585128. (2025-04-24) [2025-06-03]. https://doi.org/10.1101/2024.03.15.585128.

[54] BORN J, MANICA M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling[J]. Nature Machine Intelligence, 2023, 5(4): 432-444.

[55] JIANG F, LI M C, DONG J J, et al. A general temperature-guided language model to design proteins of enhanced stability and activity[J]. Science Advances, 2024, 10(48): eadr2641.

[56] DUMORTIER B, LIUTKUS A, CARRÉ C, et al. PeTriBERT: augmenting BERT with tridimensional encoding for inverse protein folding and design[EB/OL]. bioRxiv, 2022: 08.10.503344. (2022-08-13)[2025-06-03]. https://doi.org/10.1101/2022.08.10.503344.

[57] YANG K K, ZANICHELLI N, YEH H. Masked inverse folding with sequence transfer for protein representation learning[J]. Protein Engineering, Design and Selection, 2022, 36: gzad015.

[58] DAUPARAS J, ANISHCHENKO I, BENNETT N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. Science, 2022, 378(6615): 49-56.

[59] HSU C, VERKUIL R, LIU J, et al. Learning inverse folding from millions of predicted structures[C/OL]// Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, 162: 8946-8970[2025-06-03]. https://proceedings.mlr.press/v162/hsu22a.html.

[60] ZHENG Z X, DENG Y F, XUE D Y, et al. Structure-informed language models are protein designers[EB/OL]. arXiv, 2023: 2302.01649. (2023-02-09) [2025-06-03]. https://arxiv. org/abs/2302.01649v2.

[61] ZHANG Z B, LU J R, CHENTHAMARAKSHAN V, et al. Structure-informed protein language model[EB/OL]. arXiv, 2024: 2402.05856. (2024-02-07)[2025-06-03]. https://arxiv.org/abs/2402.05856v1.

[62] CHEN D X, HARTOUT P, PELLIZZONI P, et al. Endowing protein language models with structural knowledge[EB/OL]. arXiv, 2024: 2401.14819. (2024-01-26) [2025-06-03]. https://arxiv.org/abs/2401.14819v1.

[63] TAN Y, LI M C, ZHOU B X, et al. Simple, efficient, and scalable structure-aware adapter boosts protein language models[J]. Journal of Chemical Information and Modeling, 2024, 64(16): 6338-6349.

[64] CHENG J, NOVATI G, PAN J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense[J]. Science, 2023, 381(6664): eadg7492.

[65] TRUONG T F JR, BEPLER T. PoET: a generative model of protein families as sequences-of-sequences[C/OL]// Advances in Neural Information Processing Systems(NeurIPS 2023), 2023: 77379-415[2025-06-03]. https://proceedings. neurips. cc/paper_files/paper/2023/hash/f4366126eba252699b280e8f93c0ab2f-Abstract-Conference.html.

[66] CHENG P, MAO C, TANG J, et al. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering[J]. Cell Research, 2024, 34(9): 630-647.

[67] OLSEN T H, BOYLES F, DEANE C M. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences[J]. Protein Science, 2022, 31(1): 141-146.

[68] NGUYEN E, POLI M, DURRANT M G, et al. Sequence modeling and design from molecular to genome scale with Evo [J]. Science, 2024, 386(6723): eado9336.

[69] RIESSELMAN A J, INGRAHAM J B, MARKS D S. Deep generative models of genetic variation capture the effects of mutations[J]. Nature Methods, 2018, 15(10): 816-822.

[70] YU Y X, JIANG F, ZHONG B, et al. Entropy-driven zero-shot deep learning model selection for viral proteins[J]. Physical Review Research, 2025, 7: 013229.

通讯作者：洪亮（1981—），男，教授，博士生导师。研究方向为分子生物物理，人工智能功能蛋白质设计以及药物分子设计。

E-mail：hongl3liang@sjtu.edu.cn

第一作者：李明辰（1998—），男，博士研究生。研究方向为人工智能。

E-mail：lmc@mail.ecust.edu.cn