

# QB期刊 上海科技大学郑杰课题组： ESM蛋白质语言模型下游应用全面综述

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/40226.html>

*本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！*

**QB期刊 上海科技大学郑杰课题组：ESM蛋白质语言模型下游应用全面综述。** 论文标题：A survey of downstream applications of evolutionary scale modeling protein language models

期刊：Quantitative Biology

作者：Qingyu Yang, Jiale Yu, Jie Zheng

发表时间：28 Oct 2025

DOI：10.1002/qub2.70013

微信链接：[点击此处阅读微信文章](#)

进化尺度建模（Evolutionary Scale Modeling, ESM）系列模型通过将大型语言模型（LLMs）与蛋白质表征相结合，为理解蛋白质序列、结构和功能之间的关系提供了强大的框架，从而有望彻底改变蛋白质科学和工程领域。ESM模型在大量未标记的蛋白质序列上进行训练，能够捕捉蛋白质序列和进化历史中的复杂模式，进而深入洞察蛋白质的结构和功能特性。尽管围绕ESM的文献日益增多，但现有的综述往往未能集中且全面地描述其最新进展或具体应用。

近日，上海科技大学郑杰课题组在 Quantitative Biology 期刊上发表了一篇题为A survey of downstream applications of evolutionary scale modeling protein language models的综述文章。该文章全面梳理了ESM系列模型的最新发展，系统分类了ESM的使用技术及其在生物学领域的下游应用，并深刻探讨了ESM当前的局限性与未来发展方向。该综述为探索ESM模型能力以及大语言模型在生物医药领域的应用提供了宝贵的资源。

# A survey of downstream applications of evolutionary scale modeling protein language models

Qingyu Yang<sup>1</sup> | Jiale Yu<sup>1</sup> | Jie Zheng<sup>1,2</sup>

## 全文概要

文章回顾了ESM系列模型的发展历程、核心架构和主要应用场景，同时也提出这些模型面临的一些挑战。

## 一、ESM系列模型全景回顾

ESM系列模型经历了从ESM-1b到ESM3的更新。具体内容如下。ESM-1b是一个通用基础模型，采用了BERT架构，在约2.5亿条蛋白质序列上使用掩码语言建模（MLM）目标进行了预训练。ESM-MSA-1b则结合多序列比对（MSA）信息，常用于推断蛋白质的接触图和二级结构。ESM-1v用于变异效应预测（VEP），能够对序列突变影响蛋白质功能的效果进行评分。ESM-IF1用于固定骨架的序列设计（逆向折叠）或针对给定结构预测序列变异的功能效应。ESM-2在ESM-1b基础上进一步扩展了模型规模和优化了训练细节，是目前在蛋白质科学领域最广泛使用的通用基础模型之一。ESMFold：实现了端到端的单序列三维结构预测，在特定场景下表现出可媲美AlphaFold2的性能。ESM3是最新的前沿模型，突破了传统设计边界，能够跨越序列、结构和功能三个维度进行联合推理与蛋白质设计。

## 二、ESM的核心应用技术方法

在多样化的下游任务中，如何有效利用ESM的能力是研究的关键。如图1所示，文章总结了以下几类主要技术：一是直接使用，即针对已有特定功能的模型，例如直接使用ESM-IF1进行固定主链的蛋白质设计，或使用ESMFold预测蛋白质三维结构。二是与任务特异性模型相结合，将ESM作为高质量特征提取器，其输出的嵌入（embeddings）输入到多层感知机（MLP）、长短期记忆网络（LSTM）、图神经网络（GNN）或卷积神经网络（CNN）等用于特定任务的深度学习网络中进行有监督训练。三是模型微调（Fine-tuning），为了在特定下游任务中获得更好性能，研究者常对ESM模型使用参数高效微调（PEFT）方法。其中适配器微调（adapter tuning）、提示微调（prompt tuning）和低秩自适应（LoRA）技术是最常用且有效的微调策略。四是多模态融合（Multimodality），由于单一ESM模型往往仅包含单一模态信息，研究者利用对比学习（如CLIP）或交叉注意力机制等技术，将ESM的序列信息与蛋白质结构、药物分子等多模态信息进行融合，以弥补结构信息的不足并提升预测性能。五是利用注意力图（Attention map）：利用Transformer中的注意力图（反映残差间的余弦相似度）来直接预测残基的接触图，或作为预测复合物结构的额外特征。六是用于评估与验证，例如利用ESM模型生成的分数或概率分布，辅助指导定向进化过程，或在模拟退火优化中作为评估生成蛋白质结构的能量函数。

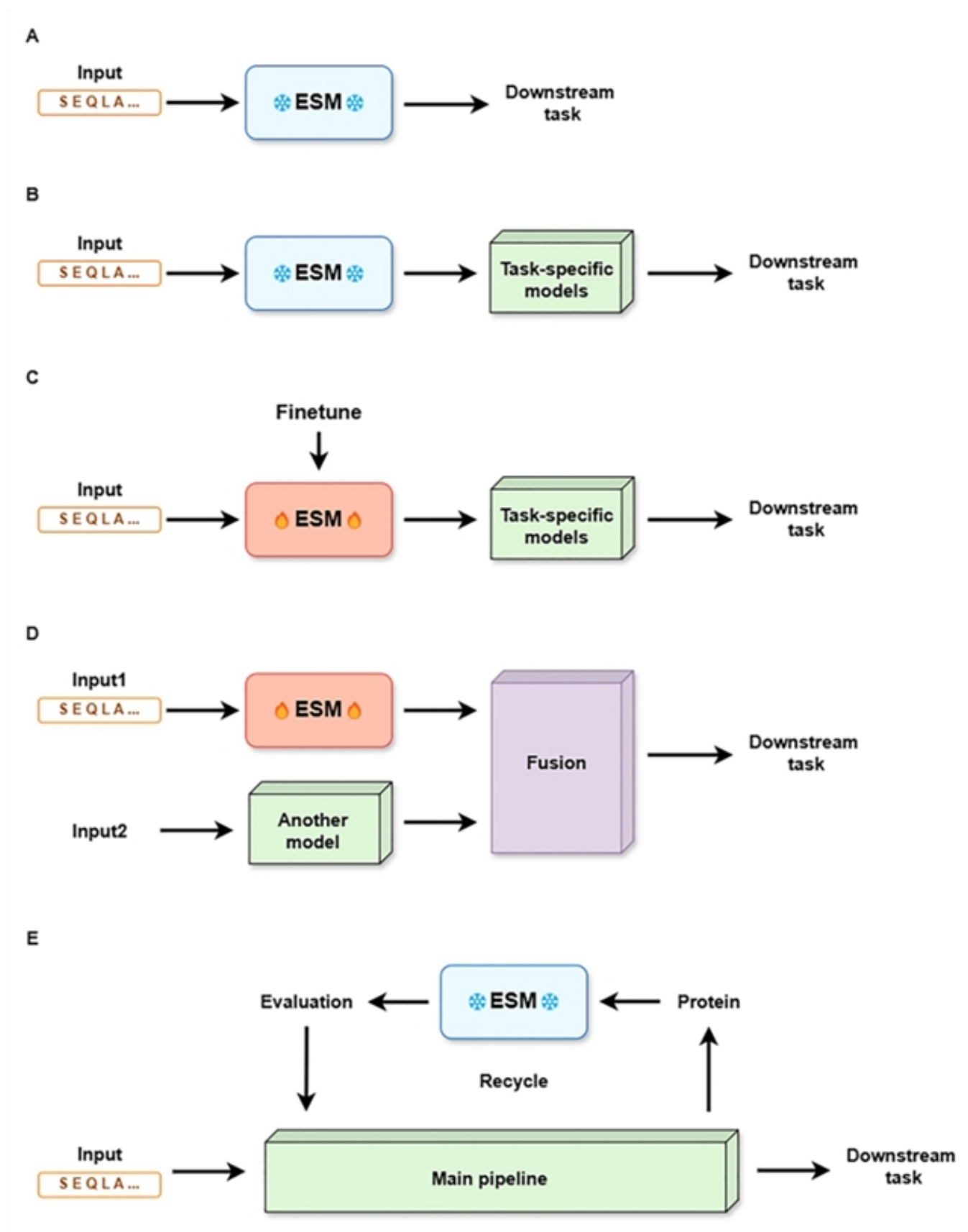


图1. 使用ESM的技术。(A) 直接使用针对特定任务的 ESM 模型（即 ESM?IF1 和 ESMFold）。(B) 利用 ESM 学习蛋白质表征并与其他模型融合。(C) 对 ESM 进行微调，而非

直接使用其嵌入向量。(D) 与其他模型的输出融合以实现多模态特性。(E) 对其他模型进行评估或验证。

### 三、广泛的下游生物学应用

得益于其强大的表征能力，ESM模型在多个主流蛋白质研究领域得到了广泛应用。在结构预测方面，从二级结构和接触图预测，到使用ESMFold等基于模型进行快速、高精度的三维结构预测，不仅速度远超传统MSA方法，还能有效处理孤儿蛋白和从头设计蛋白。在功能预测方面，涵盖酶委员会（EC）编号预测、基因本体（GO）预测、亚细胞定位预测、溶解度预测等，还出现了结合生物医学语言模型对文本定义的蛋白质功能进行零样本预测的创新应用。在相互作用预测方面，涉及蛋白质-蛋白质相互作用（PPI）的结合亲和力回归及残基级别接触预测、蛋白质-配体相互作用（PLI），以及对药物发现至关重要的药物-靶点相互作用（DTI）预测。在变异效应预测（VEP）与定向进化方面，ESM通过预训练隐式学习了序列的进化模式，一般无需显式的同源信息即能预测突变序列的功能变化；或可作为序列的评分工具应用于抗体优化、酶工程等定向进化任务中。在蛋白质设计方面，包含序列的无条件生成、针对特定功能及骨架的条件生成（如逆向折叠），以及ESM3支持的序列-结构-功能联合设计。

### 四、总结与展望

尽管ESM展现了强大的泛化性与迁移学习能力，但该综述也指出了其在实际应用中面临的几大挑战。首先是数据偏差，预训练数据（如UniProt）在不同物种和功能分类上存在严重的不平衡，且难以涵盖全新设计的非天然蛋白质，这种数据分布的偏移可能会限制ESM在特定下游任务中的实际表现。其次是高计算资源需求，基于Transformer架构的模型的时间复杂度随序列长度呈平方级增长，当序列极长、数据集庞大或需要对千亿级参数的大模型（如ESM3-98B）进行微调和训练时，昂贵的显存和时间成本制约了其规模化应用。第三是可解释性不足，如何打开深度学习的黑匣子，明确ESM是如何学习生物学知识并在复杂的生物学任务中发挥作用的，依然是一个亟待深入探索的领域。

随着相关技术的不断成熟，提高模型的可解释性、构建更可靠的数据集、将应用扩展至合成生物学与系统生物学等更广泛的领域，将是ESM未来的潜在研究方向。

### QB期刊介绍

Quantitative Biology（QB）期刊是由清华大学、北京大学、高等教育出版社联合创办的全英文学术期刊。由高等教育出版社和Wiley双平台出版和发行。QB主要刊登生物信息学、计算生物学、系统生物学、理论生物学和合成生物学的最新研究成果和前沿进展，并为生命科学与计算机、数学、物理等交叉研究领域打造一个学术水平高、可读性强、具有全球影响力的交叉学科期刊品牌。

QB期刊目前已被ESCI, PMC, Scopus, DOAJ, CSCD等国内外重要数据库收录。

来源：Quantitative Biology

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发