
“知识面”更广的人工智能蛋白质语言大模型来了

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/30116.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！

“知识面”更广的人工智能蛋白质语言大模型来了。蛋白质和DNA相互作用在大多数生命活动中起到基础性作用。传统基于生物湿实验研究蛋白质-DNA相互作用的方法周期长、费用高，成功率低。

过去十年中，基于深度学习的技术已广泛应用于蛋白质-DNA相互作用预测研究。然而，这些方法大多严重依赖于训练数据集中有限的初级序列和高质量的多序列比对信息，限制了其泛化性和准确性。大规模蛋白质语言模型的出现为克服这些限制提供了重要机会。

湖南大学国家超算长沙中心副主任、信息科学与工程学院教授彭绍亮课题组调研发现，国内外研究者使用的通用蛋白质语言模型没有特别关注特定功能领域（比如DNA结合蛋白质）的知识且通常缺乏可解释性。

为解决这些问题，该团队提出了基于大型通用蛋白质语言模型和领域自适应预训练的DNA结合蛋白语言模型ESM-DBP，系统研究了如何从蛋白质初级序列出发有效地预测DNA结合蛋白质和残基这一生物信息学和人工智能领域的挑战性问题，探索发现了基于大型蛋白质语言模型的高质量表征学习技术，为研究复杂的DNA-蛋白质相互作用机制提供新思路。



彭绍亮（最右）和团队成员探讨科学问题。受访者供图

?

团队在DNA结合蛋白质相关的四个下游任务上系统评估了ESM-DBP的预测性能，多个基准测试集上和现有预测方法的比较展示了ESM-DBP优异的预测精度。ESM-DBP还展现了在只有少量相似同源序列蛋白质序列上较好的预测性能，远超过通用蛋白质语言模型和国内外其他预测方法。

团队还通过对预测模型的可解释性分析发现神经网络对DNA结合域的高度关注，从而导致了ESM-DBP在DNA结合蛋白质预测任务上的高准确率，大大提升了蛋白质语言模型黑箱的可解释性。

这一成果近日在线发表于Nature Communications。研究得到了国家自然科学基金、科技部重点研发计划、湖南省创新群体等项目的支持。（来源：中国科学报 王昊昊）

相关论文信息：<https://www.nature.com/articles/s41467-024-52293-7>

作者：彭绍亮等 来源：《自然—通讯》

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

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