上海交大发布蛋白质设计模型“ Venus ”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/32376.html

上海交大发布蛋白质设计模型“ Venus ”。中新网上海3月22日电 (记者 许婧)上海交通大学特聘教授洪亮团队22日发布最新成果：团队将AI与蛋白质设计与改造相结合，建立了全球最大的蛋白质数据集，基于该数据集训练的模型，可以精准、高效地预测、设计蛋白质的功能，把蛋白质生产由“ 缓慢的试错 ”变为“ 高效率的精准设计 ”。

蛋白质是由氨基酸序列构成的，氨基酸序列的长度从数百个到上千个不等。AI时代，数据是推动技术进步的核心资源，庞大的蛋白质序列数据集能帮助模型更好地理解蛋白质的序列、结构和功能关系。洪亮团队建立的蛋白质序列数据集Venus-Pod(Venus-Protein Outsize Dataset)含有近90亿条蛋白质序列，包含数亿个功能标签，是全球数据规模最大、功能批注标签最多的数据集，也是另一行业知名模型——美国ESM-C模型训练用的21亿蛋白质序列的4倍体量。

3月22日，洪亮教授在上海交通大学蛋白质功能预测Venus系列模型发布暨产业合作峰会上发布该成果。上海交通大学供图

洪亮表示，该数据集构成了巨大的“蛋白质矿藏”，使得人类有可能挖掘新的蛋白或者生物催化剂，助力生物医药和合成生物学的快速发展；其次，AI大模型有望通过海量数据的学习和掌握自然界蛋白质的进化模式，为AI设计优异的蛋白质产品提供宝贵的学习资料。

蛋白质是由20种氨基酸组成的一条高分子链，这个高分子链会扭曲并折叠成独特的三维结构，正是这种独特结构赋予了特定蛋白质的生物功能。要设计出一款成功的蛋白质产品，不能只关注它的三维结构，而是要能成功预测和设计它的功能。洪亮团队直接瞄准“功能预测”这一终极目标，将复杂的蛋白质设计变成以需求为导向，配合少量实验输出结果的简单过程。

“我们训练了Venus(启明星)系列模型，与DeepMind团队的AlphaFold预测蛋白质结构不同，这个模型学习自然界蛋白质序列的组织规则以及它与功能之间的关系，其预测蛋白质突变功能的精度位居行业榜单之首。”洪亮说，Venus系列模型具备两大核心功能：“AI定向进化”与“AI挖酶”。这些超常规功能的蛋白质在生物技术、医药研发和工业生产中具有巨大的应用潜力，能够为相关领域带来创新和突破。

同时，配合Venus系列模型的全球首款低通量大体积蛋白质表达、纯化与功能检测自动化一体机，可在24小时内不间断地完成100余个蛋白质的表达、纯化与检测任务，较人力效率提高近10倍，将大大减少研发过程中的人力、物力和时间成本投入，显著提高蛋白质工程与合成生物学研究的效率。

据介绍，一款功能过硬的蛋白质产品的诞生，通常需要丰富的专家经验配合数以万计的实验试错。长期以来，蛋白质设计改造的时间长、成本高、试错密集问题，一直是业界难题。

洪亮介绍，该成果配合行业领先的自动化设备，已经进行产业化落地，比如Venus系列模型对某体外诊断头部公司碱性磷酸酶(ALP)的改造项目。Venus系列模型成功优化 ALP，使其分子活性超国际头部公司产品3倍，为超敏检测诊断(如心肌梗塞、阿尔兹海默症)带来巨大价值。目前，改造后的 ALP已进入200L规模放大生产阶段，标志着Venus系列模型成功实现产业转化。(完)

作者：许婧 来源：中国新闻网