# 全新反应描述语言可编码化学反应中分子编辑操作

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/33252.html

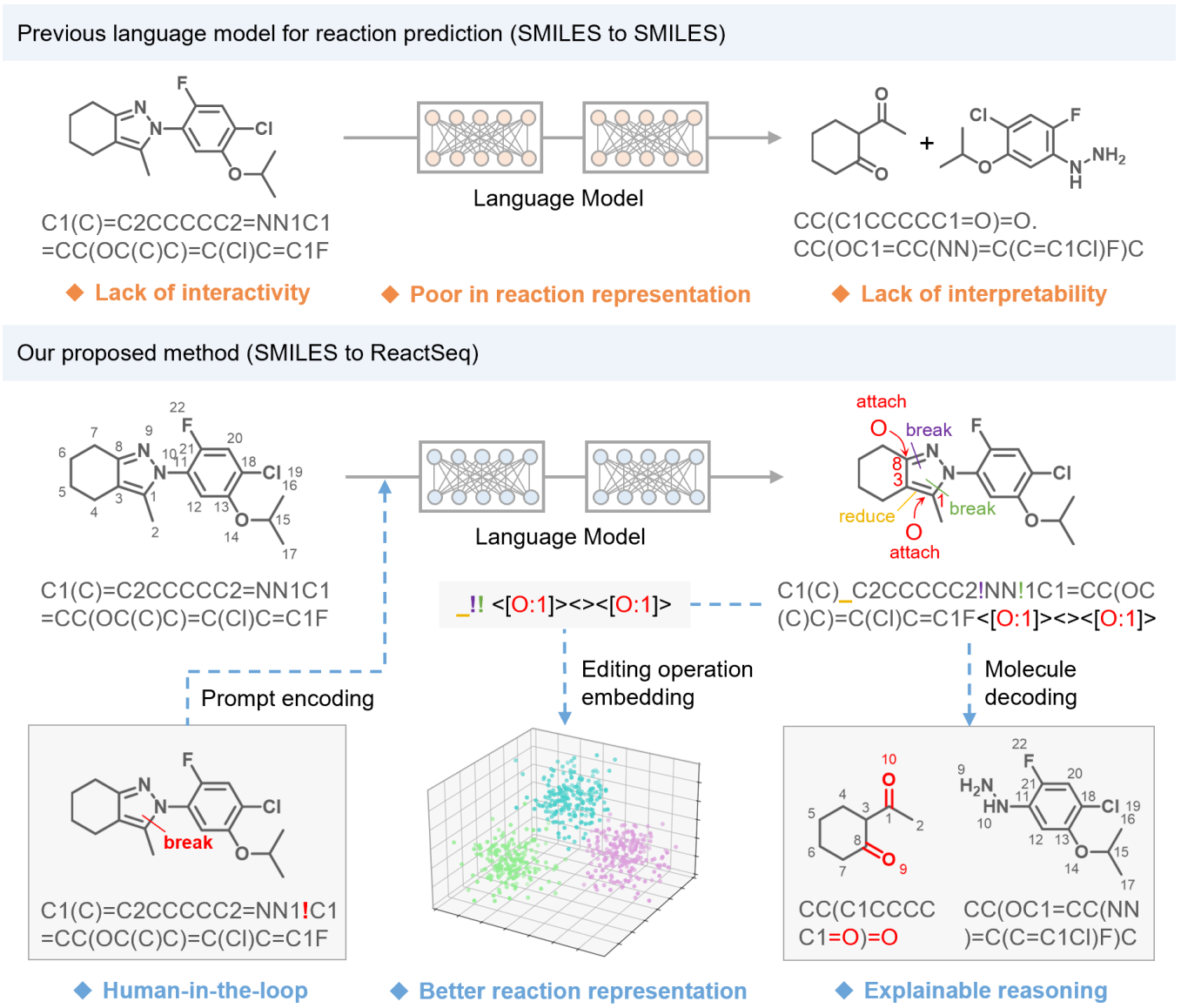


**全新反应描述语言可编码化学反应中分子编辑操作**。中国科学院上海药物研究所研究员郑明月团队，报道了一种名为ReactSeq反应描述语言，该语言可以编码化学反应中的分子编辑操作，使自然语言处理模型（NLP）在逆合成预测、反应表征检索、交互问答等方面表现得更为出色。5月13日，相关研究发表于《自然-机器智能》。

以大语言模型为代表的人工智能（AI）技术在自然语言处理方面取得了前所未有的突破，正在深刻改变科学研究的范式。近年来，在化学与药物研发领域，处理化学分子与反应的化学语言模型（CLMs）逐渐兴起。由于化学分子缺乏固有的顺序表示，CLM利用化学家定义的分子线性编码来学习和生成分子结构，目前最常用的分子线性编码是简化分子输入线输入系统（SMILES）。

为了提升CLMs在特定任务中的表现，学界设计了一些新的分子线性编码语言，用于描述化学分子的静态结构。然而，这些语言无法明确描述化学反应过程中分子中原子和键的变化过程，严重限制了语言模型在化学反应预测和表示中的应用。

为了克服上述挑战，研究团队设计了一种新的化学反应描述语言ReactSeq。ReactSeq定义了从产物结构出发，将其转化为反应物分子所需的一系列分子编辑操作（MEO），包括化学键的断裂和变化、原子电荷的改变以及离去基团的附着。在基于ReactSeq的逆合成模型中，反应物通过这些MEO从产物分子转化而来，确保了预测反应物和产物之间的精确原子映射，增强了模型的可解释性。




基于SMILES的传统反应预测语言模型与基于ReactSeq的方法之间的对比。图片由研究团队提供

?

利用ReactSeq，在不改变基本变换器（Transformer）架构的情况下便能在逆合成预测中实现最先进的性能。同时，ReactSeq具有表示MEO的显式令牌，可以对人类指令进行编码和上下文提示。测试结果表明，人类专家的提示可以显著提高模型的性能，甚至指导语言模型探索新的反应，这些MEO令牌也有利于提取反应表示，且可以产生更加精准且具有内在化学意义的反应表示。

基于该策略并结合自监督学习，研究团队构建了一种通用且可靠的反应表示方法，能够自然地区分反应类型并评估其相似性，从而提升相似反应检索、实验流程推荐以及反应收率预测等一系列下游任务上的表现。

研究团队表示，这项研究为垂直领域的大语言模型赋予了多项涌现的新能力，显著提升了自然语言处理模型应对复杂化学问题的能力，为化学领域的人工智能基础模型开发提供了新的思路。

相关论文信息：https://doi.org/10.1038/s42256-025-01032-8

作者：江庆龄 来源：中国科学报