
科研人员通过反应描述语言连接化学与人工智能

作者：writer 来源：中国科学院

本文原地址：<https://www.iikx.com/news/progress/33306.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！

科研人员通过反应描述语言连接化学与人工智能

。近日，中国科学院上海药物研究所郑明月团队报道了名为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在逆合成预测、反应表征和检索方面表现得更为出色。

以大语言模型为代表的人工智能技术在自然语言处理方面取得了进展，影响了科学研究范式。在生命科学领域，语言模型现已被用于在蛋白质和基因序列中挖掘隐藏信息，并取得了成果。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）也逐渐兴起。与自然语言、蛋白质和基因不同，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注释来学习和生成分子结构。常用的分子线性注释是简化分子输入线输入系统（SMILES）。

近年来，为提升CLMs在特定任务中的表现，有研究设计了一些新的分子线性注释。然而，这些语言都是为了描述化学分子的静态结构，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。这限制了语言模型在化学反应预测和表示中的应用。目前用于化学反应预测的语言模型包括正向和反向合成预测，通常直接将产物和反应物的线性表示相互转换，但在可解释性与交互性方面存在不足。此外，尽管预训练语言模型在多种序列数据的表示学习方面表现出色，但其在化学反应表示方面的进展相对有限。

受逆合成分析过程的启发，ReactSeq定义了从产物结构出发并将其转化为反应物分子所需的一系列分子编辑操作（MEO）。这些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，反应物不是从头开始逐个生成。相反地，它是通过这些MEO从产物分子转化而来。这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。研究显示，利用ReactSeq，只需要简单的Transformer模型便能在逆合成预测中实现先进的性能。

ReactSeq具有表示MEO的显式令牌，可对人类指令进行编码。结果表明，人类专家的提示可提高模型性能，甚至指导它探索新的反应。同时，这些MEO令牌利于提取反应表示。与聚合整个ReactSeq的嵌入相比，专注于这些MEO令牌的嵌入可以产生更加忠实且具有内在化学意义的反应表示。基于这一策略并结合自监督学习，研究人员构建了通用且可靠的反应表示方法。该方法能够自然区分反应类型并评估其相似性，从而促进相似反应的检索、实验流程的推荐及反应收率的预测。

该研究通过创新的化学语言设计，为垂直领域的大语言模型赋予了新能力。同时，这一成果提升了自然语言处理模型应对复杂化学问题的能力，为化学人工智能基础模型的发展提供了新方向。

5月13日，相关研究成果在线发表在《自然-机器智能》（*Nature Machine Intelligence*）上。研究工作得到国家自然科学基金、国家重点研发计划、中国科学院战略性先导科技专项等的支持。

[论文链接](#)

基于SMILES的传统反应预测语言模型与基于ReactSeq的方法的对比

研究团队单位：上海药物研究所

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](#)转发