使用分子动力学模拟和卷积神经网络快速预测液相酸催化反应速率

Alex K.Chew, SHengli Jiang, Weiqi Zhang, Victor M. Zavala, Reid C. Van Lehn

摘要：用于将生物质制造为高价值化学品的液相酸催化反应的反应速率对溶剂的成分非常敏感。而确定合适的混合溶剂在理论和实验上都是较为困难的。我们研究表明，三维卷积神经网络（3D-CNN）可以利用经典的分子动力学模拟数据生成的反应物-溶剂环境复杂构型，来准确预测生物质的Brønsted酸催化反应速率。我们开发了一种三维卷积神经网络（称为SolventNet），使用七种生物质衍生氧化物在水-共溶剂体系内的反应速率实验值和相关的分子动力学模拟数据来训练模型，以达到预测酸催化反应速率的目的。研究表明，SolventNet可以预测更多的反应物和溶剂系统的反应速率，比之前的模拟方法快一个数量级。这种将机器学习和分子动力学模拟结合的方法能快速、高通量地筛选溶剂体系，并且确定改进的生物质转化条件。

1. 引言

木质纤维素生物质得催化转化过程，是从可再生原料中获取交通运输所需燃料、高价值化学品等产品的一种很有前景的方法。生物质衍生分子转化通常由液相、酸催化的反应来促进（如图1.a例），而这些反应经常因为在水中受阻，故需要酸催化。一种加快酸催化反应速率的方法是修改溶剂组成，通过将极性有机共溶剂和水混合，创造混合溶剂环境。这种强化方法相比于在纯水中的速率，可以提高大约100倍。遗憾的事，通过试错实验来确认可以提升反应速率的溶剂环境的时间和金钱消耗都很高，并且对溶剂效应的物理内涵了解相当有限。相对地，计算工具已经被用于了解溶剂对化学反应的影响，并成功指导了溶剂混合物的设计，来实现高效的生物质转化过程。

在过去的十年中，从头算的量子化学方法被用来量化溶剂对生物质转化反应中基元步骤的势垒产生的影响。用从头算的分子动力学模拟(MD)，Mellmer等人发现有机共溶剂可以通过改变酸性质子催化剂、反应物和带电离子周围的溶剂环境，来降低活化能，从而增加生物质转化反应速率。（图1b）。这些模拟显示，亲水反应物会在混合溶剂环境中促进富含水的小区域形成，这些区域会优先溶解酸催化剂，并且稳定随后的碳正离子类似过渡态。这些发现表明，反应物周围水富集的程度和酸催化反应性能直接相关。类似的从头算分子动力学模拟可用于理解溶剂环境如何改变关键酸催化反应的动力学性质，比如将果糖转化为5-羟甲基糠醛（一种用于聚合物前体和交通燃料的中间化学品）的反应。但是，尽管从头算分子动力学方法可以直接探究反应机理的细节，它们的极高计算成本使得筛选多种溶剂变得不可行。

相比从头算分子动力学方法，经典分子动力学模拟可以获取更长时间尺度和更大空间尺度的信息，并且显著降低计算开销，故可以更快速地表征复杂的溶剂环境，甚至是更大的反应物的周围环境。经典分子动力学适用于模拟由反应物-溶剂-共溶剂相互作用相互影响产生的混合溶剂环境中的空间结构。这种环境可能影响反应动力学（例如，由于上述的反应物优先溶解作用），但是这种影响很难被大部分溶剂描述符（例如，介电常数）来捕捉到。从另一方面来说，经典分子动力学模拟的一个重要限制是，它们不能直接建模化学反应。尽管如此，我们最近利用了经典分子动力学模拟来了解和预测溶剂效应对水性体系中生物质衍生物转化的反应速率的影响，衍生物包括1,4-二氧六环(DIO)，γ-戊内酯(GVL)和四氢呋喃(THF)。根据一个假设，即经典分子动力学（MD）模拟能够量化在先前的从头算模拟中决定反应性的底物-水-共溶剂相互作用，我们开发了一个仅包含底物、水和共溶剂分子的MD模型，并计算了三个模拟得出的描述符，用于量化在底物周围的水富集程度、底物-水氢键形成情况以及底物的亲水性。我们随后设计了一个线性回归模型，使用这三个描述符来预测实验反应速率，并且在DIO-水体系中获得了良好的验证。这些成果展示了经典分子动力学模拟可以用来预测溶剂效应对反应的影响，不需要精确建模算催化剂或者反应机理。这个回归模型在GVL-水和THF-水体系中精度下降，这显示要么这些在经典分子动力学计算出的描述符不能量化这些系统中的反应速率，要么表示需要定义更复杂的描述符来捕捉反应趋势。但是，通过人工设计新的反应动力学是很有挑战性的，经常需要复杂、耗时的数据分析工具（例如溶液自由能或者三维溶剂图），并且很难推广到不同的溶剂体系中。

作为人工设计描述符的替代，机器学习方法正在被广泛用来推测分子性质，来自动地从复杂数据源中提取信息。举个例子，卷积神经网络(CNN)可以被用来识别和量化二维空间的数据集，比如图像。通过在合适的打过标签的图像数据之上训练，卷积神经网络可以提取空间特征，不需要人为干扰，并且可以用这些特征来分辨图像内容。CNN在分类一百二十万张图片的ImageNet大规模视觉识别挑战中，相对其他机器学习模型展示出了更突出的性能（相对比如全连接神经网络和支持向量机）。CNN还可被进一步泛化，用来概括三维数据，这可以用来促进三维分子结构的分析。举例，3D CNN最近被用于在蛋白质结构数据库上训练，来探测蛋白质功能位点，估计配体蛋白的结合点，并且量化配体蛋白的结合偏向，给予这些例子和我们之前用传统分子动力学模拟来预测酸催化反应的成功成果，我们假设3D CNN可以更精确的从分子动力学模拟中获取数据，来预测酸催化反应速率。

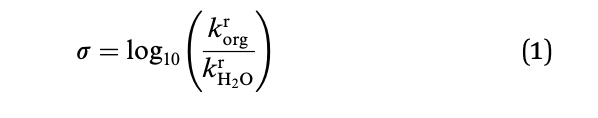
在这项工作中，我们开发了利用经典分子动力学模拟轨迹中获得的原子位置信息的3D卷积神经网络，以预测混合溶剂环境中液相酸催化生物质转化反应的速率。为了开发我们的训练程序，我们使用了在DIO–、GVL–和THF–水混合溶剂中作为标签的7种生物质衍生模型反应物的76个实验确定的反应速率。对于每个实验反应速率及其相关的溶剂混合物，我们记录了相应MD模拟轨迹中的构型（每个构型包含反应物、溶剂和水分子中原子的3D位置）。ChatGPT从模拟中收集的构型及其空间旋转被用于获得多个体素表示，这些表示映射到相同的实验反应速率，并被用作3D卷积神经网络的输入数据。这个过程使我们能够构建一个丰富的训练数据集，其中包括18,240个体素表示（超过240个不同的体素表示映射到每个76个实验反应速率）。这种方法旨在证明MD轨迹数据嵌入了能够解释反应速率的丰富信息，并且这些信息在MD模拟的早期阶段就得到了发展，从而大幅减少计算时间。我们发现所有的3D CNN，包括我们设计的全新3D CNN，我们称为SolventNet，以及两种以前设计的3D CNN(ORION和VoxNet)，预测反应速率都比人工选择描述符准确很多。SolventNet的预测推广到一个测试集，该测试集包括从文献中获得的32个实验确定的反应速率，其中包括三种额外的极性无水共溶剂中的反应物的速率，这三种共溶剂没有包含在模型训练中，分别是二甲基亚砜、乙腈和丙酮，这些共溶剂具有不同的性质（例如官能团、碱性和极化能）与用于训练模型的共溶剂不同。我们进一步发现，使用SolventNet进行准确的反应速率预测仅需要4纳秒的经典MD轨迹数据，这比基于人工选择描述符的模型中使用的原始205纳秒的MD数据提高了50倍。因此，我们得出结论认为3D原子构型包含解释反应速率的重要信息，并且这样的构型在MD模拟的早期阶段就得到了发展。我们预计3D CNNs和经典MD模拟相结合的计算效率将使这些工具能够与过程模型结合，用于筛选溶剂并优化生物质转化过程的反应器条件。

论文的组织结构如下。我们首先描述了一组108个实验确定的反应速率，每个速率与一组单独的反应物和混合溶剂环境相对应，这些速率被用作训练集（76个反应速率）和测试集（32个反应速率）的标签。然后，我们描述了使用从MD轨迹计算得到的人工选择的描述符组成的数据对基线线性和非线性模型进行训练和验证，每个标签对应一个轨迹和相应的描述符集。接下来，我们描述了通过将每个MD轨迹分割为10个独立的体素表示来生成的另一种输入数据集，以供3D CNNs解释。经过数据增强，这个过程为每个标签产生了240个体素表示，用于训练和验证3D CNNs，以与基线模型进行比较。然后，我们评估SolventNet的测试集和留一交叉验证准确性，以建立模型的泛化能力。最后，我们使用显著性图可视化影响SolventNet准确性的空间特征。

1. 结果和讨论
   1. 用于训练和验证模型的实验反应数据

共使用了108个先前报道的在混合溶剂环境中进行的酸催化脱水或水解反应的实验确定的反应速率作为酸催化反应速率预测模型的训练和测试的标签。每个反应速率是针对唯一的反应物-溶剂组合（即在混合溶剂环境中的反应物）测量的。共有76个实验确定的反应速率（标签的70%）取自参考文献4，并被用作训练集的标签。训练集包括7个与生物质相关的模型反应物的速率：乙基叔丁基醚（ETBE）、叔丁醇（TBA）、1,2-丙二醇（PDO）、乐禾葡聚糖（LGA）、果糖（FRU）、麦芽二糖（CEL）和木糖醇（XYL）。溶剂系统包括含有三种极性无水共溶剂之一（1,4-二氧六环（DIO）、γ-戊内酯（GVL）和四氢呋喃（THF））的25 wt%、50 wt%、75 wt%或90 wt%的水混合物。32个实验确定的反应速率（标签的30%）取自参考文献3和33，并被用作测试集的标签。测试集包括4个模型反应物的速率：叔丁醇（TBA）、果糖（FRU）、1,2-丙二醇（PDO）和葡萄糖（GLU）。溶剂系统包括含有二甲基亚砜（DMSO）、乙腈（MeCN）和丙酮（ACE）的水混合物。训练集实验在343–433 K的温度范围内进行，而测试集实验在363–433 K的温度范围内进行。所有实验均使用三氟乙酸作为催化剂，但参考文献33中的测试集实验（对于丙酮水混合物中的果糖和葡萄糖）使用了盐酸。因此，测试集的标签由来自不同文献来源的独立反应速率组成，这些反应速率对应于训练集中未包含但在可比较的实验条件下的反应物和共溶剂。所有反应物和共溶剂的化学结构分别显示在图1c和d中。表S1和S2†列出了每个标签的反应条件。

为了对比不同混合溶剂环境对反应速率的影响，定义方程（1）将动力学溶剂参数（sigma）定义为反应物r在给定混合溶剂环境中脱水或水解的表观速率常数(Krorg)与相同反应在纯水中的表观速率常数之间(krh2o)的对数比率。



Sigma>1表示在混合体系中的反应速率快于在水中的，sigma<1表示小于在水中的。所有训练集和测试集的标签都定义为sigmaexp来促进回归模型的训练。

2.2 使用从经典分子动力学中人工选择描述符的基线预测模型，用于反应速率。

在先前的工作中，我们表征了一个反应物分子在混合体系的分子动力学模拟，生成了76种反应物的205ns的模拟轨迹，这些轨迹被纳入了训练集。图2a说明了从这些分子动力学（MD）模拟中提取描述符以预测实验动力学溶剂参数的一般方法。从每个分子动力学轨迹中，我们计算了三个在物理上有意义且由人选择的描述符，用于捕捉反应物的亲水性和溶剂相互作用：优先排斥系数(gamma), 这描述了在与反应物附近的空间区域中水的局部富集程度的量化；反应物和水之间的氢键结合寿命，这描述了通过附近水分子对假设的带电过渡态的稳定程度的量化；以及可访问的羟基分数（d），通过将反应物的羟基组的可访问表面积（ASA）除以整个分子的ASA来量化反应物的亲水性。描述符计算在引用.4中，描述符的值在表S.1中列出。

描述符的值被用在作为线性回归的输入（等式2），来预测溶质动力学参数（sigmapred）：

文本

低可信度描述已自动生成

ABCD是回归系数，A、B、C 和 D 是回归系数，而带有波浪符号的描述符经过最小-最大缩放，重新缩放在 0 到 1 之间。图2b说明了用于评估线性模型以及本工作中描述的所有后续模型是否能够推广到未包含在训练数据中的新的反应物-溶剂组合的1/5交叉验证过程。在这个过程中，训练集中的 76 个标签被随机分成五个折叠，每个折叠包含大约 20% 的标签。与五个折叠之一中的标签相关联的所有输入数据（即描述符值）被用作验证集，与其余四个折叠中的标签相关联的输入数据被用于训练模型（即回归系数被拟合），并为验证集计算sigmapred的值。这个过程迭代了五次，确保每个训练集标签在验证集中出现一次。

图2c展示了线性回归模型的sigmapred和sigmaexp相关关系图，每一个sigmapred值都对应了一个测试集的交叉验证。