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。

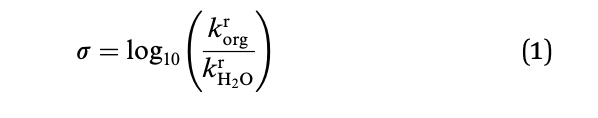
在这项工作中，我们开发了利用经典分子动力学模拟轨迹中获得的原子位置信息的3D卷积神经网络，以预测混合溶剂环境中液相酸催化生物质转化反应的速率。为了开发我们的训练程序，我们使用了在DIO–、GVL–和THF–水混合溶剂中作为标签的7种生物质衍生模型反应物的76个实验确定的反应速率。对于每个实验反应速率及其相关的溶剂混合物，我们记录了相应MD模拟轨迹中的构型（每个构型包含反应物、溶剂和水分子中原子的3D位置）。ChatGPT从模拟中收集的构型及其空间旋转被用于获得多个体素表示，这些表示映射到相同的实验反应速率，并被用作3D卷积神经网络的输入数据。这个过程使我们能够构建一个丰富的训练数据集，其中包括18,240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。

* 1. 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。我们使用spred和sexp之间的最佳斜率和根平均平方误差（RMSE）来评估模型的准确性；一个完美的模型的斜率为1，RMSE为零。由于实验数据的RMSE为0.10，故所有RMSE在0.1附近的预测模型都相当准确。线性模型的斜率为0.49，RMSE为0.58；大多数异常值是发生在THF-water（圆形符号）或90 wt%共溶剂系统（钻石型符号）中的反应。仅使用没有5倍交叉验证的DIO-水混合物数据拟合线性模型会导致RMSE为0.23,4，但这种方法对GVL-水（RMSE为0.36）和THF-水（RMSE为0.59）混合物（ESI，表S3†）的准确性较低，这表明线性模型的性能在很大程度上取决于共溶剂。这表明，线性模型在不同共溶剂的泛化能力较差。我们还测试了非线性模型是否可以使用相同的输入数据改进线性模型的预测。我们进行了5倍的交叉验证，以评估具有三个隐藏层的完全连接的神经网络，每个层有十个直线性单元（ReLU），然后是一个线性单元，用于使用三个人类选择的描述符作为输入来预测s的回归任务。图2d显示了使用完全连接的神经网络模型spred和sexp之间的奇偶校验图。完全连接的神经网络模型的行为与斜率为0.46、RMSE为0.62的线性模型的行为相当。我们使用这些多描述符模型作为与替代模型进行比较的基线。

这些预测表明，用多描述符模型预测的反应率位于正确的象限中，几乎没有假阳性或假阴性sigma值，并且对一些共溶剂（即DIO）是准确的。描述符强调了空间信息的重要性：gamma量化反应物附近区域的溶剂成分，tao描述反应物羟基和水的相对位置，delta与反应物亲水性和疏水性区域的相对表面积有关。然而，两个模型都对系统有显著的离群值，对应于更大的sexp值，这表明描述符无法捕获可能在反应物-溶剂环境的复杂几何（3D）特征中编码的重要信息。此外，这些描述符的识别需要领域专业知识，而且很耗时。

* 1. 生成输入数据集，供3D CNN使用经典MD数据进行解释

为了改进人为选择的多描述模型，我们假设3D CNN可用于在从经典MD（从反应物-溶剂、溶剂-共溶剂和反应物-共溶剂相互作用的组合中产生）采样的原子位置与实验反应速率之间建立映射。我们期望3D CNN适合分析这些系统，因为：（i）CNN可以提取反应物-溶剂的非直观特征，通过识别数据中的空间相关性，（ii）由于溶剂分子之间的分子间相互作用，等相系统表现出明显的空间相关性，（Iii）在人类选择的描述符中编码的空间信息的重要性表明，空间相关性与溶剂效应有关，（iv）3D CNN可以分析原子位置，而无需将域转换为二维空间（flattening），从而捕获反应物和局部溶解环境的详细几何形状。因此，3D CNNs为识别反应物-溶剂环境的复杂特征提供了一个自然框架，这些特征影响反应速率，但使用人类直觉可能不容易识别。

我们开发了一种协议，用于将从经典MD模拟中获得的反应物、溶剂和共溶剂分子原子位置的轨迹数据转换为适合3D CNN分析的数据表示。3D CNN解释由排列在3D网格中的一系列体素组成的数据，每个体素都包含几个单行通道的归一化强度。这些通道可以传达不同类型的信息。体素在网格中的相对定位赋予了空间信息。因此，我们将MD输出的空间连续原子位置转换为体素，以记录（0.2 nm）3体积元素内水、共溶剂和反应物氧原子的出现，并将其归一化。这种数据表示受到了人工选择的多描述符模型成功的物理直觉的启发：描述符G的重要性表明应记录水和共溶剂原子的位置，以量化溶剂分子在反应物附近的偏好富集，而描述符s和d表明应记录反应物氧原子的位置，以量化潜在的氢键形成和反应物的亲水性。每个体素关联的体积被选择为与典型的原子半径相当，以确保可以解析分子几何结构。

图3说明了将MD位置转换为体素网格的方法。对于与MD轨迹（即MD组合）期间采样的单个时间相一致的原子位置集，我们以反应物的质量中心为中心进行了三维直方图。直方图覆盖一个立方（4纳米）3体积（体积小于总模拟盒尺寸，以避免越过模拟盒边界），该体积被划分为对应于（0.2纳米）3体积元素的20x20x20阵列网格。对于每个箱，我们通过计算箱中的水原子数量和按任何垃圾箱内水原子的最大数量，进行归一化来统计水分子。分离地执行了相同的程序，以计算每个容器中共溶剂和反应物氧原子的归一化发生。然后，将正常化水、反应物和共溶剂分别存储在“红色、绿色和蓝色”颜色通道中（图3a）以从单个MD中获得20x20x20的体素网格。

为了捕获溶剂分子在立方体积内扩散时相对于反应物的优先位置，并防止体素被占用，我们平均了从多个连续的MD中获得的网格值，以生成一个单一的平均体素平均网格表示。具体来说，每个体素表示都是通过从2 ns的MD数据（对应于200个连续的MD配置）中平均网格值生成的，如图3b所示。选择2 ns模拟时间作为最大化模型精度和最小化计算成本之间的平衡。这个模拟时间远远短于每个训练标签关联的205纳秒轨迹。因此，我们将每条轨迹的前20纳秒分成10个独立的2纳秒分区，并为每个分区生成一个体素表示，从而为每个训练标签产生了10个体素表示。这些选择是基于广泛的稳健性测试，以确定最佳的输入数据表示，如在ESI第S5节（表格S4）中所描述。由于3D CNNs不具有旋转不变性，我们通过将每个体素表示旋转来进一步增强训练数据，生成每个体素表示的24个唯一的立方体旋转（图S1†），从而导致每个训练集标签有240个（增强的）体素表示。

* 1. 3D CNN以更少的模拟时间改善了反应率预测

接下来，我们开发了一个3D CNN模型，我们称之为SolventNet，该模型输入体素表示并输出预测的动力学溶剂参数。SolventNet架构由四个卷积层、两个最大池层和三个完全连接的层组成，如图4a所示。该架构基于之前开发的VoxNet 3D CNN，但用两个卷积层、一个最大池化层和三个全连接层取代了原先的全连接的层。我们还将SolventNet与VoxNet31和ORION30 3D CNN进行了比较（ESI，图。S4†）调查3D CNN架构的预测准确性。所有三个模型都包括一个带有线性激活单元的层，用于预测s的回归任务。

3D CNN使用5倍交叉验证进行评估，类似于用于人类选择的描述符模型的程序。对于3D CNN，训练数据包括80%标签的所有增强体素表示（14 400或14 640体素表示），验证数据仅包括其余20%标签的非增强体素表示（150或160体素表示）。图4b显示了使用SolventNet的spred和sxp之间的奇偶校验图。

Spred的值和错误条报告了每个标签10个非增强验证集体素表示中的每个预测值的平均值和标准差。与基线人类选择的多描述符模型相比（图2），SolventNet显著改善了动力学溶剂参数的预测，最佳斜率为0.89，RMSE为0.37。图4c进一步表明，使用所有三个3D CNN（SolventNet、VoxNet和ORION）获得的RMSE具有可比性，并优于多描述模型。此外，3D CNN仅使用每个反应物-溶剂组合的20纳斯MD数据进行训练，而用于计算eqn（2）中的三个描述符的每个反应剂-溶剂组合的MD数据为205纳秒。用一个GPU和一个CPU核心训练3D CNN需要1.6-2.4小时，而MD模拟需要216小时，使用一个GPU和28个CPU内核进行所有76个反应物-溶剂组合，时间要长得多。更多地，需要的MD数据相对原先减少了10倍。

训练CNN时，一个潜在的问题是大量学习的参数，这可能会导致过拟合。然而，我们注意到，这项工作中使用的3D CNN相对紧凑；SolventNet有172 417个参数，而VGG16有33 601 345个参数，VGG16是一个常见的2D CNN（方法和表S3†）。参数与SolventNet的增强训练体素表示数量之比为11.8-11.9；相比之下，完全连接神经网络模型的训练描述符集数量之比为4.4。SolventNet的参数与培训示例的比例很小，可以与替代的3D CNN架构相媲美，这表明我们可以不考虑训练中的过拟合。举个例子，3D CNN已用于评估蛋白质三级结构的质量，其参数与培训示例的比例为34（540万个参数，16万个体素网格）和85（1.7亿个参数，200万个体素网格）。我们还观察到，通过将200纳秒的分子动力学数据分割成100个体素表示，将训练体素表示的数量增加了10倍，并不影响3D卷积神经网络的性能（图S3†）。此外，ORION、VoxNet和SolventNet之间的RMSE差异表明，即使ORION的参数几乎是SolventNet的五倍，CNN架构对模型性能的影响很小。因此，具体的3D CNN架构并不是关键，而是使用3D CNN来分析分子动力学模拟，从而提高模型性能。我们还测试了替代的神经网络架构（见ESI†），包括同时使用描述符和体素表示作为输入的网络、2D CNN和具有替代体素表示的3D CNN，但没有观察到准确性的提高。综合这些结果表明，与基于人工选择的描述符的先前模型相比，3D CNN具有更高的准确性（图2），同时需要更少的分子动力学数据进行训练，所使用的训练数据量足够，并且预测准确性不随模型架构而显著变化。在本文的其余部分，我们将重点放在SolventNet上，因为它具有中位数数量的参数，并且在预测测试集时表现良好，详见下一节讨论（参见表S3†进行模型比较）。

2.5 SolventNet对新溶剂和反应物的可推广性

接下来，我们评估了SolventNet在测试集中包含的32种底物-溶剂组合的准确性。对于每个测试集中的底物-溶剂组合，我们进行了新的分子动力学模拟，以获取生成两个体素表示所需的4纳秒模拟数据（按照图3中所示的相同流程）。仅使用两个体素表示来评估SolventNet在使用极少的分子动力学模拟数据迅速推广到新的溶剂组合的能力。测试集的体素表示没有进行增强。我们使用所有的训练集数据（76个标签和18,240个增强的体素表示）重新训练了SolventNet，然后对测试集的体素表示预测了动力学溶剂参数。

图5展示了使用SolventNet对测试集进行的spred和sexp之间的平等图。spred的值和误差条报告了每个标签的2个测试集体素表示预测值的平均值和标准偏差。最佳拟合斜率为0.72，RMSE为0.48，表明与验证集的预测相比，虽然预测准确性略有下降，但SolventNet仍然具有良好的泛化性能。值得注意的是，测试集的准确性超过了基线多描述符模型的验证集准确性。平等图还显示与Pearson's r为0.80的高线性相关性，这是对测试集数据的描述。因此，即使对于SolventNet预测较不准确的系统，该模型仍然捕捉到了有关溶剂组成对反应速率有定性的预测能力。

测试集的结果显示，SolventNet在DMSO-水混合物中表现良好，RMSE为0.43。这个结果有些令人惊讶，因为DMSO比用于训练的共溶剂更碱性，而且已知在亲水底物的羟基周围的结合位点上与水竞争。此外，我们之前发现在DMSO-水混合物中，底物-DMSO相互作用可能优于底物-水相互作用，而在其他共溶剂-水系统中更倾向于底物-水相互作用。尽管存在这些独特的行为，SolventNet学到的特征能够将其转化为对DMSO-水混合物的反应速率预测，其准确性可与对所有其他测试系统的预测相媲美。在测试集中，GLU的预测准确性最差，RMSE为0.88。由于GLU不属于训练集，我们预期对于这种底物的预测准确性将低于FRU；此外，该系统使用了上述的盐酸催化剂。然而，尽管如此，对于GLU的转化，再次捕捉到了定性趋势，Pearson's r为0.93（仅计算包含GLU的系统）。

基于DMSO和GLU的分析，我们还使用留一法交叉验证来确定SolventNet的预测是否对包含在训练集中的特定底物或共溶剂敏感，这进一步得到了对基于描述符的模型在THF-和GVL-水混合物的预测性能较差的启示。在这个过程中（如图6a所示），我们保留了原始训练集中包含给定共溶剂的底物-溶剂组合（例如，所有DIO-水混合物）或给定底物的标签和相关的体素表示，并将这些数据用作测试集。SolventNet是使用剩余的数据进行训练，并用于预测每个测试集底物-溶剂组合的10个体素表示的动力学溶剂参数。通过迭代使用每个共溶剂或底物的数据作为测试集，重复执行此过程。图6b显示了在共溶剂之间进行留一法交叉验证的spred和sexp之间的平等图。RMSE在0.27–0.43之间变化，与对DMSO-水混合物的预测相当。这些结果表明，SolventNet对于广泛范围的混合溶剂环境的预测是可比较的，包括线性多描述符模型对THF-和GVL-水混合物的预测性能较差的情况。图6c显示了在底物之间进行留一法交叉验证的spred和sexp之间的平等图。RMSE在0.11–0.81之间变化，具体取决于底物，与在共溶剂之间进行留一法交叉验证的结果相当。最大的RMSE是对于LGA，与GLU的测试集结果相当。与GLU的结果一样，LGA的spred和sexp表现出强烈的线性相关性，r为0.90，表明捕捉到了反应性的定量趋势。LGA可能是一个离群值，因为它的亲水性被高估，由于体素表示考虑了LGA的所有氧，包括醚键中的氧。总体而言，来自独立测试集和留一法交叉验证的结果表明，SolventNet的预测在所有测试的共溶剂和除LGA外的所有底物中都具有良好的泛化性能。

* 1. SolventNet功能的物理解释

虽然SolventNet相对于基于人工选择的描述符的模型提供了更高的预测准确性和计算效率，但很难从物理上解释模型提取的特征。例如，不同底物-溶剂组合的代表性体素表示没有明显的视觉特征（图S2†）。为了初步解释SolventNet识别的特征，我们生成了显著性图以可视化SolventNet对不同体素的敏感性，从而对底物周围的特定空间区域进行分析。显著性图包含用于每个体素的显著性值（在0-1之间标准化），指示SolventNet预测对该体素中水、底物和共溶剂原子的规范化出现的敏感性。较大的显著性值表示增加的敏感性。图7显示了使用综合梯度方法39生成的显著性图（在方法中描述），通过将XYL在90 wt% DIO中的体素表示输入到完全训练的SolventNet模型中。显著性图被分成反应物、共溶剂和水通道的独立的三维体素网格，每个体素根据其规范化显著性值使用与输入体素表示相同的颜色方案着色（图3）。透明的体素对模型预测不重要（规范化显著性值小于0.10）。 对图7中的显著性图的检查确认SolventNet主要识别与人工选择描述符定义基础假设一致的在底物附近的局部区域内的溶剂环境特征。每个通道的显著性图通过沿z轴（任意选择）对显著性值进行平均并投影从3D到2D，生成图7中显示的等高线图。这些图清楚地显示了底物附近的区域对预测最为重要，模拟体积的大小足够大，远离底物的区域不重要，并且SolventNet提取了不受球对称假设约束的非直观几何特征。类似的显著性图可能对指导未来的从头计算非常有用，通过识别需要研究的重要溶剂区域，从而减少量子化学计算所需的分子数量。

1. 结论

在这项工作中，我们结合了机器学习工具和经典分子动力学模拟，开发了SolventNet，这是一个3D卷积神经网络，用于解释已转换为体素表示的分子动力学模拟数据，以预测在极性非质子共溶剂的水混合物中的酸催化反应速率。SolventNet不需要事先人工选择描述符，并且可以捕捉到空间相关性，而无需假设球对称性。SolventNet（以及其他3D CNNs）比基于人工选择的描述符的模型更准确地预测反应速率，并且推广到新的极性非质子共溶剂、共溶剂质量分数以及在模型训练期间未使用的底物。这些发现表明，SolventNet从经典分子动力学模拟中确定的底物、溶剂和共溶剂分子的空间构型中提取特征，这些特征包含足够的信息，可以预测溶剂对酸催化反应的影响，即使反应机制和可能的过渡态没有被明确建模。

这一新开发的方法的一个显著优势是计算效率。一旦训练完成，SolventNet仅需要4纳秒的分子动力学模拟数据就能够预测单一底物-溶剂组合的反应速率。对于本研究考虑的系统尺寸，在典型高性能计算集群的单个节点上，这些轨迹可以在不到一个小时内模拟完成，与从头计算方法相比，大大缩短了预测反应速率所需的时间。这降低的计算成本表明，SolventNet可以用于溶剂筛选，可能与过程模型结合，以设计更高效的生物质转化过程，而无需昂贵的实验试错。然而，到目前为止，所有研究的系统都涉及水和极性非质子共溶剂的混合物；将SolventNet扩展到在明显不同的溶剂体系中预测反应速率（例如，离子液体）可能需要额外的训练。

SolventNet的输入体素表示仅包含原子位置，因此忽略了重要的化学信息，如原子之间的共价键存在和原子电荷。这些数据可能通过替代的网络架构进行解释。例如，图神经网络可以将原子表示为节点，原子相互作用表示为边。因此，我们预计进一步的模型发展可以继续提高预测准确性。此外，输入到SolventNet的体素表示仅从底物状态的分子动力学模拟中生成；未来的工作将探索是否将来自产物状态模拟的体素表示纳入其中，能够提高反应速率预测的准确性或实现对反应选择性的预测。经典分子动力学还可以模拟比本研究中研究的小分子底物更大的分子，例如与生物质相关的聚合物（例如纤维素、半纤维素或木质素）。未来的工作将探索使用SolventNet与语义分离技术，该技术允许将输入到卷积神经网络的数据的各个区域分别分类，以基于生物质相关聚合物的分子动力学模拟同时预测多个反应位点的反应性。使用从头计算方法对这些更大的系统进行溶剂筛选是困难的。

1. 方法
   1. 经典分子动力学方法

经典分子动力学模拟使用2 fs的时间步长和leapfrog积分器进行。所有模拟中初始模拟盒的尺寸设置为(6 nm)3，并按照所需比例添加水和共溶剂分子。所有底物和共溶剂都是使用CGenFF/CHARMM36力场参数化的。水采用Single Point Charge/Extended (SPC/E) 模型进行建模。使用1.2 nm的邻居列表截断生成Verlet列表。范德瓦尔斯相互作用采用Lennard-Jones势，并在1.0 nm和1.2 nm之间平滑地过渡到零。电静力相互作用采用平滑粒子网格Ewald方法进行计算，短程截断为1.2 nm，网格间距为0.12 nm，4阶插值。使用LINCS算法对键进行约束。溶剂系统在300 K和1 bar的NPT模拟中平衡了5 ns，采用速度重标定恒温器和Berendsen恒压恒温器。然后向系统中添加单个底物分子，并使用相同的恒压和恒温器进行500 ps的平衡。然后在反应温度和1 bar的条件下使用Nose–Hoover恒温器和Parrinello–Rahman恒压恒温器进行NPT生产模拟。所有恒温器使用1.0 ps时间常数，所有恒压恒温器使用5.0 ps时间常数，等温压缩性为5.0 × 10^-5 bar^-1。每隔10 ps输出一次模拟构型。所有模拟使用GROMACS 2016进行，并使用VMD进行可视化。 使用该协议获得的底物-溶剂组合的模拟数据（包括图1c和d中用黑色绘制的分子）取自参考文献4。这些分子的生产模拟进行了200 ns。由于在参考文献4中描述的G的收敛较慢，因此需要这么长的模拟时间。最后的190 ns用于计算图2中所示的多描述符模型的G和d。额外进行了5 ns的生产模拟以计算s。描述符的计算方法如参考文献4中所述（数值列在表S1†中）。在这200 ns的第一部分用于生成训练和验证3D CNN的体素表示。对于包含在测试集中的底物-溶剂组合（包括图1c和d中用灰色绘制的分子），按照上述模拟协议进行了新的模拟。这些分子的生产模拟在反应温度（列在表S2†中）下进行了4 ns，并用于生成测试3D CNN的体素表示。

* 1. 模型训练、验证、测试总结

从文献来源中获得的108个实验确定的反应速率被转换为使用方程（1）得到的动力学溶剂参数，并用作标签。每个反应速率（标签）是针对唯一的底物-溶剂组合确定的，这被定义为在单一共溶剂和水（以不同重量百分比）的二元混合物中的单一底物。标签被分成一个训练集，其中包含72个标签（标签的70%），全部来自参考文献4，以及一个测试集，其中包含32个标签（标签的30%），来自参考文献3和33。每个标签与一个分子动力学轨迹相关联。所有模型都使用5倍交叉验证程序进行评估，其中80%的标签及其关联的描述符/体素表示（60-61个标签）被用作训练数据，剩余的20%的标签（15-16个标签）及其关联的描述符/体素表示被用作验证数据。该过程被迭代5次，以确保每个标签被用于验证一次。模型性能基于对验证集的动力学溶剂参数的预测值的均方根误差（RMSE）进行评估。 多描述符模型的训练数据是通过使用每个分子动力学轨迹计算205 ns的MD数据生成的，如上述所述，生成3个描述符。多描述符线性模型是通过对训练数据进行回归来进行训练的。多描述符全连接神经网络使用Keras深度学习库在TensorFlow的基础上进行了500个epochs的训练（对于每个epoch，神经网络对整个数据集进行一次训练周期）。训练使用Adam优化器进行，学习率为0.001，均方损失函数，训练批次大小为18（一个epoch相当于四个反向传播步骤）。

3D CNN的训练数据是通过将与训练集标签相关联的每个MD轨迹分割为包含2 ns（200个MD构型）连续MD数据的10个独立分区而生成的。对于每个分区，所有MD构型都被转换为体素的3D网格，然后进行平均以获得体素表示，如在结果和讨论中所述（图3）。用于训练的每个标签的10个体素表示通过将所有24个唯一的立方体旋转作为训练数据进行增强，从而每个标签有总共240个体素表示，即14,400-14,640个训练体素表示。用于验证的每个标签的体素表示没有被增强，导致150-160个验证体素表示。所有3D CNN都使用Keras深度学习库在TensorFlow的基础上进行了500个epochs的训练。模型使用Adam优化器进行训练，学习率为0.00001，均方损失函数，训练批次大小为18（一个epoch相当于814个反向传播步骤）。所有3D CNN的学习曲线显示在ESI†的图S5中。 SolventNet的泛化能力是通过使用测试集和对训练集的留一法交叉验证来评估的。对于每个测试集标签，连续的4 ns MD数据被转换为两个独立的未增强的体素表示。SolventNet是使用所有训练集数据（18,240个体素表示）的增强体素表示进行重新训练的，并用于预测每个测试集标签的2个体素表示（64个体素表示）的动力学溶剂参数。在留一法交叉验证过程中，将与训练集中的单个底物或共溶剂相关联的所有标签保留为测试集，然后使用剩余标签的所有数据对SolventNet进行训练，并用于预测每个保留标签的10个体素表示的动力学溶剂参数。在这两个过程中，模型的性能是基于测试集的动力学溶剂参数的预测值的均方根误差（RMSE）进行评估。

* 1. 3D CNN结构

这项工作考虑了三个在复杂性和参数数量上有所不同的3D CNN架构：VoxNet31（5层，150 689参数）、ORION30（8层，908 833参数）和SolventNet（9层，172 417参数），这是我们内部开发的。VoxNet和ORION的架构在ESI中描述（图S4）。† SolventNet有三个阶段（图4a）。第一阶段有两个卷积层，分别为8和16 3 3 3 3，以及2 2 2最大池层。第二阶段的结构与卷积层相同，除了两个卷积层分别有32和64卷。最后一个最大池层的结果被传递到批处理标准化层并被处理。然后将数据输入到具有三个完全连接层的神经网络中，每层有128个节点。ReLU激活函数用于完全连接的层。