文献综述

1 研究背景

离子液体(ILs)，在催化[1-1，2]、萃取和分离等领域具有广泛的应用，这是因为它们具有可忽略的蒸汽压力、高离子电导率、较大的温度稳定性和良好的溶剂性能。对于特定应用，ILs的性质可以通过不同的阳离子和阴离子轻松调控[1-3,4]。特别是，阳离子的烷基链长度在ILs的性质中起着重要作用，包括密度、电导率、表面张力和扩散系数等。然而，合成具有长烷基链的ILs通常需要很长的反应时间[1-5]。例如，含有长烷基链的苯胺衍生物的反应在70°C下需要2-7天的长时间，而且反应动力学参数信息不多[1-4]。

随着算法、算力的不断进步，分子动力学（Molecule Dynamics，MD）逐渐成为研究分子体系的强大工具。这种方法通过计算分子在时间序列上的运动，可以进一步计算体系的热力学、动力学等性质，同时可以大幅度减少需要的湿实验量，也可用于预测难以用实验表征的现象，直观展示实验现象发生的机理。[1-6789]

然而，分子动力学模拟生成的数据只包含物理信息，不直接包含反应，为了更全面地理解模拟结果，研究者们迫切需要有效的分析方法。特别是，关注分子动力学过程中的速率信息对于理解反应动力学至关重要。而传统的手工分析方法已经难以满足对大规模模拟数据的需求，这就迫使我们转向机器学习这一强大的工具。

卷积神经网络（Conclusive Neuron Network）由于其自动提取特征的能力，尤其适用于处理复杂的图像类信息。在分子动力学模拟中，其特征极多且难以选择合适的描述符。通过将机器学习引入分子动力学模拟的结果分析中，我们有望从复杂的数据集中提取出有关反应速率的关键信息。这种方法有望为科学家们提供更深入、更全面的了解分子动力学过程，并为设计新材料、研究生物分子结构和预测药物反应性等领域的研究提供有力支持。

因此，本研究旨在借助机器学习技术，通过对分子动力学模拟数据的智能分析，从已有的反应速率湿实验结果中，回归模拟数据的反应速率等关键信息，为分子动力学模拟结果的解释和应用提供新的途径。这一方法可以大幅度降低湿实验的量，同时给实验设计、工程设计等提供定量或定性数据。

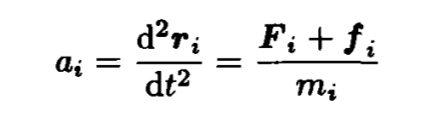
2 文献综述

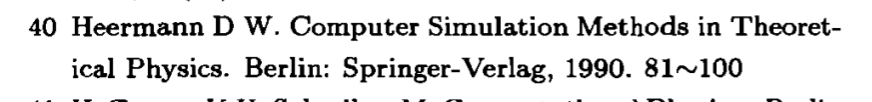
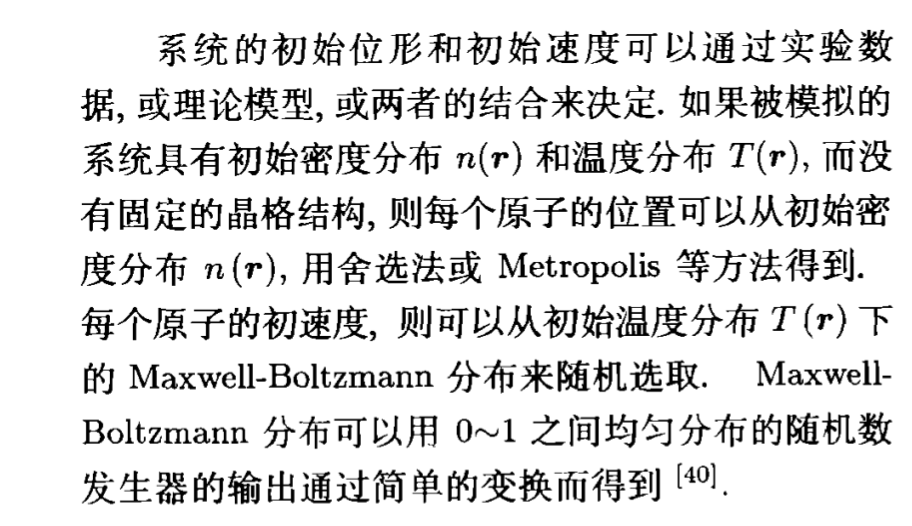
2.1 研究对象

本分析方法基于郑伟中等[2.1-1]对烷基咪唑(RIm)和烷基硫氰酸酯(RSCN)，其中R=甲基(M),乙基(E),正丁基(B)，的反应动力学和经典分子动力学数据进行。在实验中取得的结果表明，硫氰酸酯的较长烷基链导致显著较慢的速率常数和更高的活化能。在330 K下，MIm/BuSCN系统的速率常数相比于MIm/MeSCN系统降低了21倍。MIm/BuSCN系统的活化能在MIm/MeSCN系统中增加了27.5 kJ/mol。在明确的溶剂中进行的元动力学计算与实验得出的过渡态能垒呈一致的趋势。MD模拟表明，硫氰酸酯的长烷基链导致反应位点的空间分布较弱，扩散和重新取向速度较慢，这可以降低分子间碰撞的概率。

2.2 经典分子动力学(MD)和从头算分子动力学(AIMD)的基本原理简述

2.2.1 经典分子动力学

在分子动力学中，系统中原子的一系列的位形是通过对牛顿运动方程积分得到的。通过解牛顿第二定律的微分方程可以获得原子的运动细节。这里‘，，‘， ‘及了‘分别为第原子的质量、 位置矢量、 所受的可以根据势函数 的梯度求出的力 和其他的力。U原则上是所有原子位置的函数， 而 则原则上可 以是所有原子的位置和速度的己知函数，

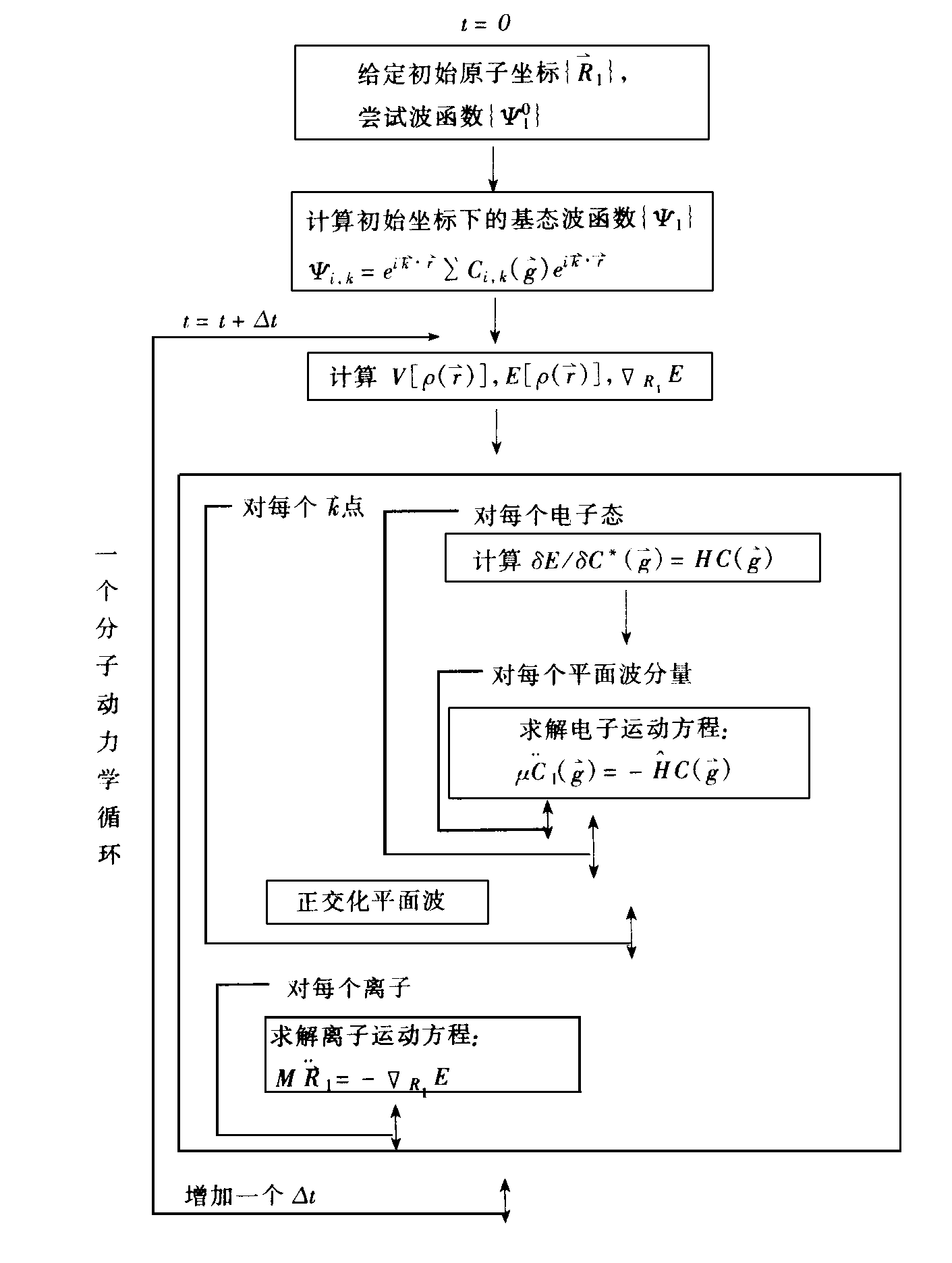


2.2.2 从头算分子动力学

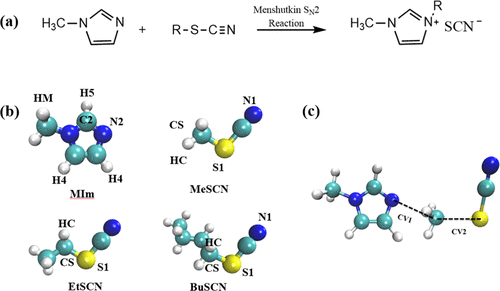
MD方法在处理很多化学问题中获得了极大的成功。但是，因为MD是基于力场的，因此忽略了电子极化效应，同时无法描述化学反应，无法得到化学键的生成和断裂的本质。如果采用经验方法、价键理论等设置各类近似处理，仍会偏离反应的路径。为了解决上述问题，从头算分子动力学方法(ab initio molecular dynamics, AIMD)被提出。

AIMD主要基于三个假设：1、忽略系统的核量子效应 2、认为系统满足轨道近似（即不按照电子云来处理） 3、认为系统满足绝热条件。

AIMD的基础是电子基态计算，这是一个非常复杂的量子多体问题，故引入密度泛函理论(Density Functional Theory, DFT) 来将多体问题转化为单电子轨道方程，即Kohn-Sham方程[2.2.2-1], 根据电子和原子核的相互作用对电子密度的影响程度,对交换势采用局域密度近似(Local Density Approximation)和广义梯度近似(General Gradient Approximation) 来使得方程可解。[2.2.2-2] 1985年Car-Parrinello提出的CP算法是从头算分子动力学的基本原理。[2.2.2-3]该算法的具体细节不在本文中赘述，而其大致框图如下图。[2.2.2-4]



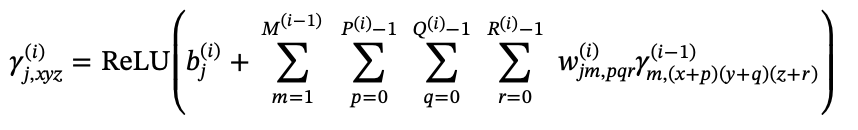
2.3 反应基础

本反应(Rim+RSCN)的反应方程式如图：

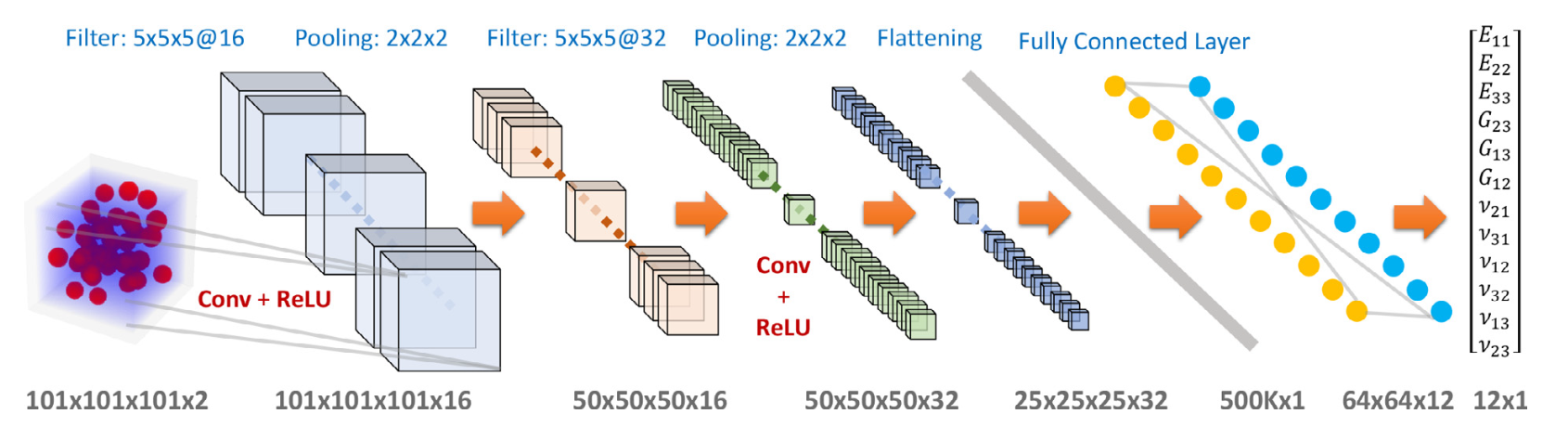
此反应机理为著名的Menshutkin SN2. 两种亲核试剂和底物反应生成了有明显点和分离的过渡态，然后产物形成明显的电荷分离，形成阳离子、阴离子。

2.4 卷积神经网络基本原理

自从2012年AlexNet在ILSVRC挑战的巨大成功之后，CNN（Convolutional Neural Network，卷积神经网络）以极快的速度在全球掀起研究热潮。CNN最初是为了解决计算机视觉问题而提出的。LeCun[2.4-1]等人在1990年设计了最早的CNN，成功识别手写数字。由于当时计算能力有限，CNN的应用受到了限制。近年来，随图形处理单元（GPU）等计算硬件的巨大进步，CNN方法得以复兴。CNN通过其权重共享机制与传统的全连接神经网络（FC-NN）有所不同。卷积神经网络(CNN)是深度学习中用于处理图像数据的强大工具，它能自动提取和学习图像中的特征。CNN通 过使用卷积层来处理图像，能够捕捉到图像中的局部特征，并通过池化层降低特征的空间维度。这种结构使得CNN在 图像分类、目标检测等领域表现出色。在本研究中，我们建立了一个3D-CNN，以一个三维的形式来分析三维的模拟盒。

3D-CNN采用预处理后的体素作为输入。随后的多个卷积层构成了3D卷积滤波器和池化操作的关键组成部分。如图 所示，3D滤波器在相位体素上进行扫描，并应用卷积操作（张量的点积）以生成特征图。每个滤波器的权重和偏置被训练以从输入中提取显著特征。步幅（Stride）、填充（Padding）和滤波器大小是定义卷积操作的一些常见超参数。步幅表示滤波器每次移动的步长大小。例如，步幅为1表示滤波器逐体素扫描体积。为了保持输出的空间大小，使用零值体素填充输入。如果使用步幅为1和2层零填充进行卷积操作，则图 中的输入和输出大小将相同。在CNN中，通常在连续的卷积层之间添加池化层。它通过对体素值进行下采样逐渐减小数据的空间大小。池化操作可以计算体积内的最大值或平均值。图5演示了具有体积大小的最大池化操作的工作方式。激活层用于在CNN中引入非线性。激活函数接收输入张量并执行逐元素的非线性转换。一些典型的激活函数包括修正线性单元（Rectified Linear Unit, ReLU f(x)=max(0, x), Sigmoid函数和tanh函数。在这些非线性函数中，由于ReLU廉价的算术运算（一阶线性，可以避免exp和三角函数运算，这些运算都需要大量的浮点操作，消耗很高的GPU算力）备受青睐[2.4-2]。Relu的输出可由下式表出[2.4-3]：

最后，使用一个全连接层，其中两个相邻层之间的神经元是相互连接的。FC层以前一隐藏层的展平张量作为输入，并将其映射到所需的输出，完成所需数据的回归。



2.5 由分子动力学模拟数据获取反应速率的前期尝试

Theodore W. Walker等[2.5-1]通过对酸催化的乙基叔丁基醚、叔丁醇、乐福葡萄糖三聚体、1,2-丙二醇、果糖、赤霉糖二聚体和木糖醇的反应速率在水与三种极性非质子共溶剂（γ-戊内酯、1,4-二氧六环和四氢呋喃）的混合溶剂中进行的测量，并且对这些体系进行了分子动力学模拟计算。从模拟中手动选择了三个描述量：

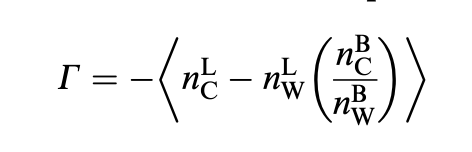
(1) 反应物局部溶剂领域中水富集的程度（Γ）；

(2) 水分子与反应物之间的平均氢键寿命（τ）；

(3) 反应物可及表面积中羟基占据的比例（δ）。

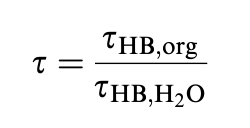
此三个描述量如下：

定义优先排除系数Γ为相对于溶剂领域的体积领域而言，在反应物的局部溶剂领域内的共溶剂分子的过量数量。优先排除系数根据MD模拟的结果，按照方程 计算，其中nC和nW分别是共溶剂和水分子的总数，上标L和B分别表示局部和整个体积盒内的分子。



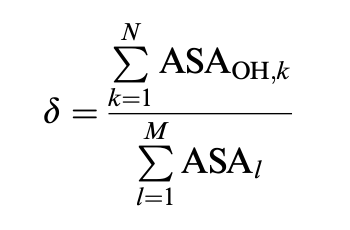
Γ的正值表明，在反应物的局部溶剂领域中，共溶剂的浓度低于体积溶剂领域。正的Γ也被称为优先水合，因为共溶剂的排除表明反应物对水有更高的亲和力。负值的Γ表明，在反应物的局部溶剂领域中，共溶剂的浓度高于体积溶剂领域，反应物对共溶剂有更高的亲和力。[2.5-2345]

定义平均反应物-水氢键寿命τ为第二个描述符，用于量化混合溶剂体系中水与反应物结合的强度。我们期望，在具有较大共溶剂浓度的混合溶剂环境中，由于无法形成水-水氢键，反应物-水氢键的偏好增加，从而水与反应物之间的氢键更强（即更持久）。水与反应物之间更强的相互作用可能导致过渡态自由能降低，从而增加反应速率。下式为定义[2.5-67]。



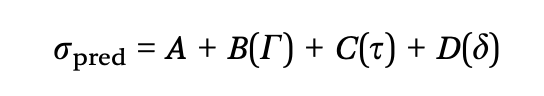
实验观察到随着共溶剂分数的增加，反应物在所有共溶剂环境中的氢键寿命比率单调增加，证实了随着混合物中可用水量的减少，反应物与水之间的氢键结合强度通常增加。

定义可及表面积中羟基占据的比例δ为反应物分子中（N）羟基占据的可及表面积除以分子中（M）总原子占据的表面积，公式为：

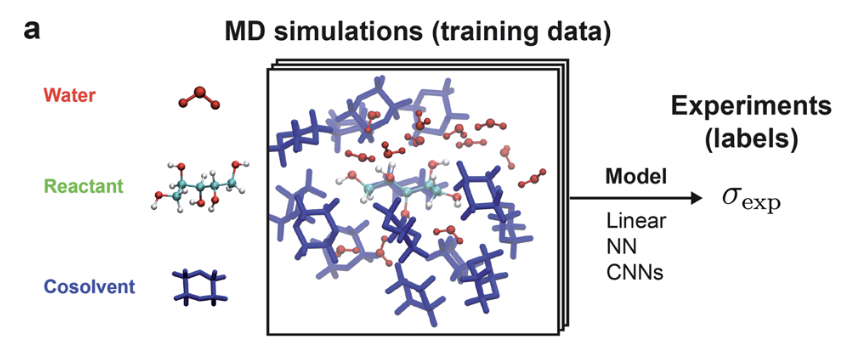


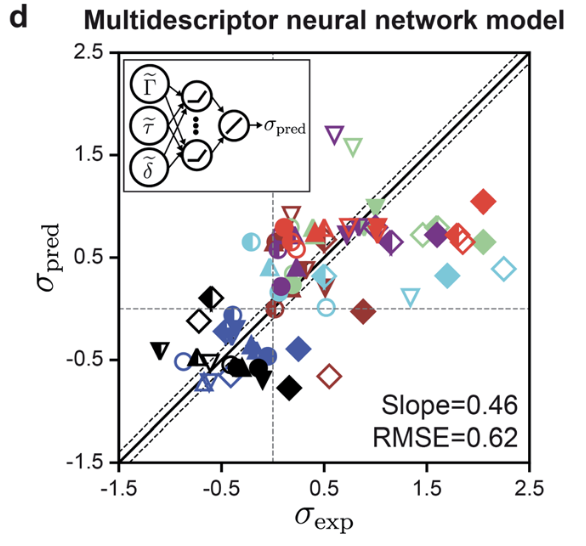
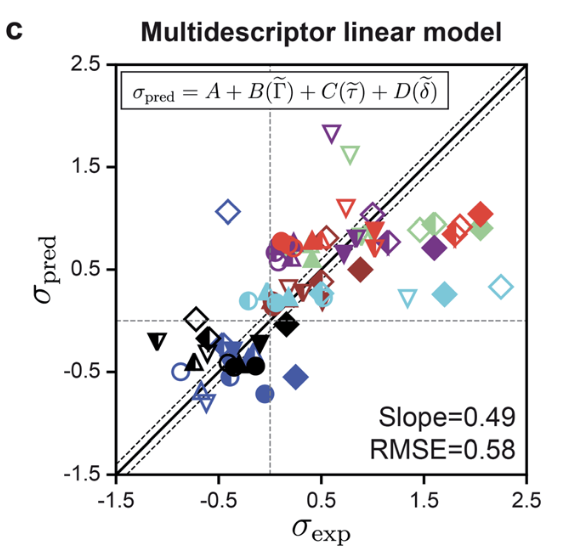
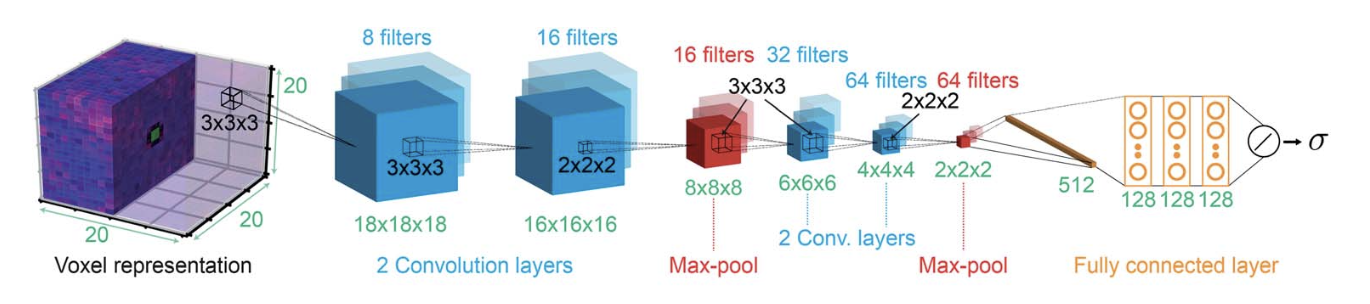
将这个无量纲的量解释为反应物亲水性的一种定性分子描述符，通过其分子大小进行归一化；较大的δ值表示更具亲水性的分子。

将这三个描述符用线性方法进行连接，得到公式：

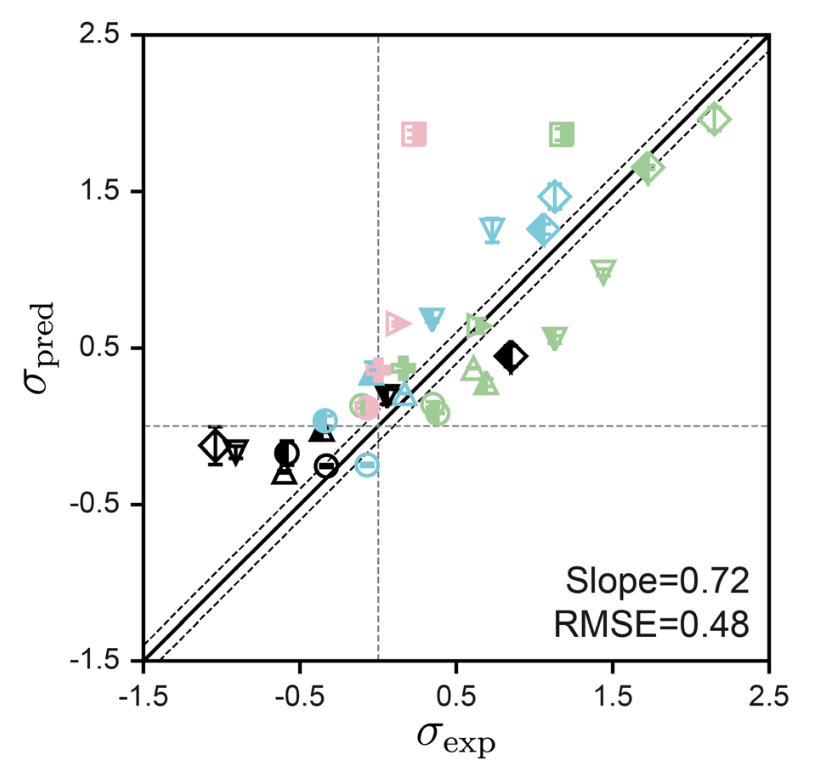


此方法可以得到一个初步的预测结果。而由人类选择的描述符总是存在问题的，线性方法来描述复杂的体系也是显然不精确的。

Alex K. Chew[2.5-8]等人创新性地使用了各类数据方法，继承和发展了此类描述符方法。其不仅使用线性模型，还使用神经网络、卷积神经网络等方法来分析了这三个描述符和反应速率的关系，如图 ，其线性结果为RMSE=0.58，神经网络结果为RMSE=0.62，情况甚至有所下降。

  
 作者团队使用了更多创新性的方法。分子动力学模拟数据可以认为是一个三维的点云，如果将其处理成类似(x,y,z,c)的信号，则可以处理成类似三维的图像，其中c代表颜色通道，也可以认为是一个多维数组，每个通道内就是每种分子的分布情况。作者团队首先设计了一个协议，将从经典分子动力学模拟中获得的反应物、溶剂和共溶剂分子的原子位置轨迹数据转换为适用于3D卷积神经网络（CNN）分析的数据表示。3D CNN解释由一系列体素组成的数据，这些体素排列成3D网格，每个体素包含在几个独立通道中的归一化强度。这些通道可以传递不同类型的场信息。网格中体素的相对位置提供了空间信息。因此，我们将由MD输出的空间连续的原子位置转换为体素，记录了在(0.2纳米)³体积元素内水、共溶剂和反应物氧原子的归一化出现次数。这种数据表示受到了人为选择的多描述符模型成功的物理直觉的启发：描述符Γ表明应记录水和共溶剂原子的位置，以量化溶剂分子在接近反应物时的偏好富集，而描述符τ和δ表明应记录反应物氧原子的位置，以量化潜在的氢键和反应物亲水性。选择每个体素关联的体积与典型原子半径相当，以确保可以解析分子几何结构。最后得出的20x20x20x3结构恰好满足CNN所需输入形式。下图为作者团队使用的网络的示意图。

在完成训练之后，得到了非常好的结果：

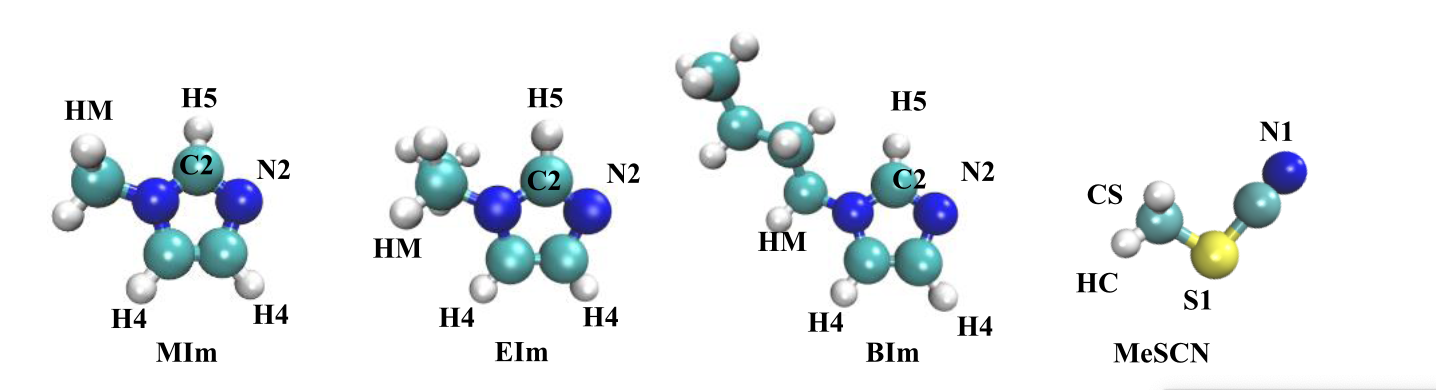


2.6 其他基于分子动力学和机器学习的研究

Kota Noda等[2.6-1]使用图神经网络建立了机器学习模型，建立了原子位置、速度、每个节点的边和相态（固体、液体）的模型，这个模型可以通过分子动力学数据来预测一些物理性质。Theodore W. Walker等[2.6-2] 总结了其工作组通过实验、经典分子动力学模拟和机器学习工具估计溶剂对液相酸催化的生物质转化反应速率和选择性的工作，将这些见解综合到一个工作流程中，通过经典分子动力学模拟和机器学习工具，其估计了溶剂组成对生物质衍生分子反应性的影响，通过分析反应物周围的溶剂环境。此外，分子动力学模拟可以估计溶剂诱导的生物质转化反应选择性对特定产物的影响，通过以溶剂组成为函数的形式量化反应物和产物状态的溶剂化自由能。这些基于分子动力学和机器学习的工具被结合成一个工作流程，用于选择在生物质转化应用中使用的混合溶剂系统，同时最小化实验的数量。Carlos Ramírez-Palacios[2.6-3]等使用深度学习和分子动力学方法预测了ω-转氨酶选择性

3 数据分析方法

3.1 分子动力学数据来源



图展示了MIm、EIm、BIm和MeSCN的分子结构。对于分子动力学模拟，分子由Packmol软件构建，烷基咪唑和MeSCN的分子数量分别为400和20，遵循20:1的分子比。MIm/MeSCN、EIm/MeSCN和BIm/MeSCN系统的盒子尺寸分别为35.0 Å × 34.3 Å × 34.0 Å、36.9 Å × 37.7 Å × 37.7 Å和41.2 Å × 41.7 Å × 43.0 Å。随后，使用10000步的能量最小化，来消除不合理的结构。而后使用Nose-Hoover恒定体积（NVT）方法进行了5 ns的模拟，使用Parrinello-Rahman恒压方法进行了等温等压（NPT）[3.1-1,2]集合的完全平衡模拟。这些MD模拟都使用GROMACS程序运行，时间步长为1fs.[3.1-3,4]

3.2 数据读取与voxel representation的生成

使用python mdtraj库来完成分子动力学模拟数据的读取。

3.3 数据增强方法

3.4 神经网络结构