机器学习在大分子领域的应用

本文提及的大分子主要指生物大分子（蛋白等）和各种复杂有机物。在科研领域，生物大分子的折叠结构预测和有机大分子的逆合成分析一直是耗时的难题。本文将从这两个方向上叙述其进展。

一、生物大分子折叠结构预测

生物大分子折叠结构的研究主要集中在蛋白质晶体上。这主要是因为蛋白质的结构对生物化学反应的过程产生影响，而不像如核酸序列那样序列为反应的主要影响因素。蛋白质虽然序列很容易通过各种谱甚至是传统的电泳法来确定，但蛋白质分子具有二级（肽链局部的相互作用产生的构向）、三级（在二级基础上整条链的大致形状）以及四级结构（因肽链形状使亚基之间所成的相对位置的关系）。由于不同的构象（即前文所述二级结构的另一种表述）在不同条件下的转化是十分容易的，许多蛋白晶体存在不同条件下可改变的外观结构，存在无规律部分（loop）。因此，试图从理论的角度计算预测三级和四级结构是比较困难。过去为了使理论计算得以继续进行，通常需要与其他课题组合作，通过诸如冷冻电镜的方法。但冷冻电镜需要耗费大量资源并且可重复性并不高。在计算机模拟方面，过去通过I-TASSER，基于过去对于相似结构的研究进行预测，也就是同源建模。

目前在折叠结构预测方向上，最领先的项目为AlphaFold。在AlphaFold中，使用了一个二维矩阵来描述蛋白质的结构，矩阵的边为残基序列（残基可理解为缩合后的氨基酸本体），交叉点为残基间的距离，这样的数据便于卷积神经网络的训练，优于传统上描述蛋白质结构所使用的三维坐标系外加一个温度相关变量的方法。在生成了二维矩阵之后，通过神经网络预测蛋白质骨架的折叠角度，通过梯度下降法得出最终结果。当然，在实际程序中，AlphaFold一代几乎没有像AlphaZero那样“大力出奇迹”，实际运算的许多步骤都使用了传统算法。而在AlphaHold二代使用了“Attetion Model”（注意力模型），一种计算相邻数据间影响权重的模型。评估集则使用了前述二维矩阵的计算结果。

尽管AlphaFold并不基于模板，但从现阶段的结果表明，AlphaFold往往和晶体学结果一致，但如果实际实验使用核磁共振方法得出结构，则结果不太准确。AlphaFold可能已经出现了过拟合的情况，因此该项目对较小的核酸差异带来的结构功能改变不够敏感。同时为了便于计算得出最优解，该项目运用了物理上的“能量最小“原则，但在实际生命体中，蛋白必然不可能在能量最小的情况下发挥作用。当然在人工蛋白的设计方面，AlphaFold目前的准确度已经足够升任这一工作，可以有效的减少实验验证次数。

二、随机森林模型在有机化学中的应用

随机森林模型是一种以决策树为基础的算法。决策树是机器学习中的一种简单的分类模型，一层层的对输入进行分类，某种程度上类似于if-then。随机森林模型在决策树用到了bagging的方法。Bagging的全称是bootstrap aggregating。这个方法的核心思想是用不同的模型在同一个训练集上不可能产生完全相同的结果，也就不会产生完全相同的误差。而随机森林模型通过给各个决策树投喂不太相同的原始数据，使得不同的决策树训练出不同的结果，再通过投票选用占比最高的模型结果。从这样看，也不难想明白为什么说随机森林模型像“民主集中制”。

在美国普林斯顿大学和美国默克公司合作研究成果——具有多维变量的Buchwald-Hartwig偶联反应收率预测模型中，团队挑选出与体系相关的120种描述符，包括分子描述符：轨道能量（EHOMO和ELUMO）、偶极矩、电负性、硬度、体积、质量、椭圆度及表面积等；原子描述符：原子静电荷和核磁位移等；振动描述符：振动频率和强度等。其中用于描述异噁唑的参数有19个，描述卤化物的参数有27个，描述有机碱的参数有10个，而描述钯配体的参数则多达64个。该团队同时测试了多种线性回归模型、k-近邻算法、支持向量机、贝叶斯广义线性模型、神经网络模型和随机森林算法。结果表明随机森林模型的结果最好，预测产率值与实际产率值间的线性相关度最高，标准误差值为7.8%。而神经网络的标准误差值达到了16.9%。并且相对于其他模型需要使用数据中的70%才达到的预测准确度，随机森林算法只需5%的数据就能达到一样的预测准确度。这说明近年来最火的神经网络，并不是机器学习中万能的模型。目前，随机森林模型在有机化学领域中还在判定分子的反应性上的应用见诸报道。

三、人工智能分子合成设计

在这里选取Chematica这一软件来介绍目前机器学习在有机合成过程设计的进展。

Chematica软件的开发基础在于NOC，有机化学网络，一个用有向二分图结构来保存化合物和化学反应信息。在2009年，这个软件的团队负责人Grzybowski对NOC的网络结果进行了分析，认为有机反应的网络是一个无尺度网络，也就是说，大部分有机合成只会用到极小的一部分的分子，剩下的大部分有机物很难出现在合成路线上，因此这一软件的设计过程可以类比于设计搜索算法。在2016年Grzybowski用剪枝算法优化后的代价函数模型这一机器学习模型做了一个深度优先搜索算法。之后在2018年，Grzybowski在软件中加入了神经网络和分布式计算。