# 基于递归神经网络的分子水溶性和抗 HIV 活性预测

潘高翔，付思杰，蔡丹杨

北京大学化学与分子工程学院

**2020 年 1 月 6 日**

**摘 要：**本研究借鉴 Alessandro Lusci 等人开发的水溶性预测深度学习模型，利用 pytorch 架构，以递归神经网络学习抽取分子特征，实现了小分子水溶性和抗 HIV 活性预测。其中 SolNet 模型仅依靠递归特征，在测试集上达到了$R^2 = 0.83$的精度，而 PlusSolNet 模型训练不足，精度不高，预测抗 HIV 活性的模型因训练难度过大而失败。如经过科学调参，递归神经网络模型将会具有更优秀的表现。GitHub 地址：https://github.com/ChrisChemHater/ADMET_RNN

## 1　背景介绍

药物分子的吸收(Absorption)、分配(Distribution)、代谢(Metabolism)、排泄(Excretion)、毒性(Toxicity) 是在当代药物设计和药物筛选中十分重要的药物动力学性质。在药物设计研究中，针对新靶点的药物设计常常需经过蛋白分子动力学模拟、虚拟筛选、合成修饰、细胞试验、动物试验、临床试验等步骤，若在分子设计早期能够合理预测分子的 ADMET 性质，可显著提高药物设计的成功率，减少研发成本和劳动量，缩短研发周期。然而，与分子的能量、构象、电荷分布等直接物理性质不同，ADMET 性质与高度复杂的生理环境相关，造成普通计算手段难以预测。近年来兴起的机器学习方法，尤其是深度学习，从统计学的角度给出了解决方案。

分子结构信息不仅存在于各个原子中，还在于原子间的键连方式和空间结构，分子的拓扑和三维信息对分子的生物功能至关重要。因此，如何抽取分子特征成为了分子性质预测机器学习模型开发的关键之处。2013 年，Alessandro Lusci 等人开发出了基于递归神经网络预测分子水溶性的模型[1]，该模型仅依据分子结构就可给出精度基本令人满意的预测值，在缺少实验数据的虚拟筛选中有广阔的应用前景。2019 年，Kevin Yang 等人和 Zhaoping Xiong 等人分别开发了 Directed MPNN 和 Attentive FP 方法，实现了深度学习方法预测分子 ADMET 性质领域内的最佳精度。

本研究借鉴 Alessandro Lusci 等人的递归模型架构并加以改进，以及 Kevin Yang 等人的分子信息采集方法，对递归神经网络模型进行深入探索，旨在实现仅需分子基本结构信息的预测模型，而不需要任何实验数据。

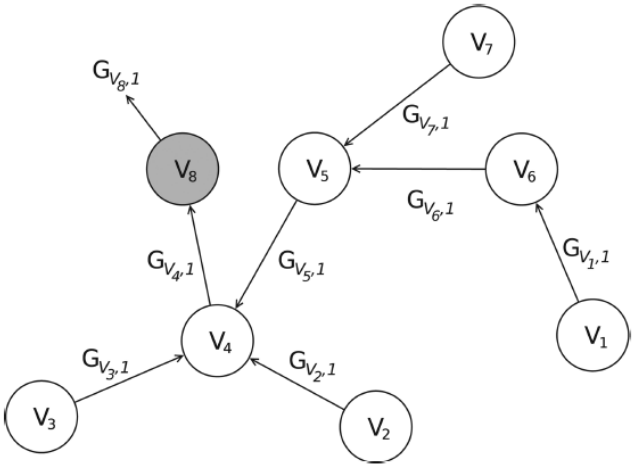
## 2　文献模型原理

需要注意的是，分子结构一般是用一个无向图来表示，而 RNN 方法基于树结构。为了解决这一问题，作者们考虑了一系列分子树的整合，将与分子有关的所有可能的有向图相联系起来。此做法的一个优点是，因为合适的分子表示数据可以自动从数据中学习出来，所以此法可以很大程度上减少识别合适的分子描述器的工作。

有几种类似于此法的方法已经被用于预测分子水溶性问题上，并已经在四组标志性数据集上做了测试。从实验性结果来看，通过一些参数的评估，深度学习的预测能力可以持平甚至超过其他最先进办法， 同时深度学习还能解决部分由于数据集太小或者噪音太大带来的限制。

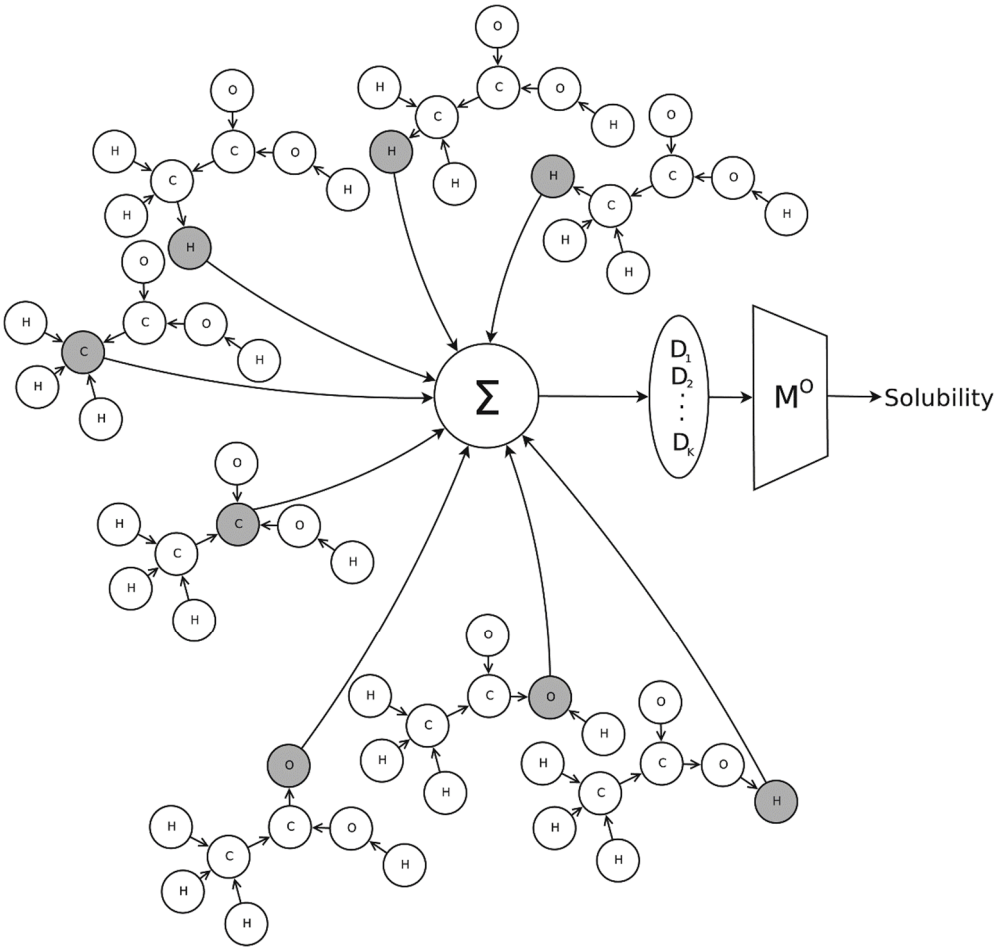
模型有向图表示分子结构数据 M 介绍(以 $CH_3COOH$ 为例)：

Directed Acyclic Graph Recursive Neural Networks (DAG-RNN)




首先，将分子的所有原子分别看作一个单独的节点(node)，选定分子内一个特定原子作为根节点，与其直接键连的原子结点为其子节点，继续往下，与该子节点键连的原子节点(除去前面已出现的原子节点)又被当做该子结点的子节点，如此反复往下建立树结构，直至所有原子结点被遍历完毕。每个结点所囊括的信息可以包括原子序数、键的类型、原子质量、原子半径等，结点与结点之间通过神经网络连接，如此形成一个 DAG-RNN 结构，得到一个长度固定的向量 D。

多次选取特定原子作为根节点，得出所有原子对应的向量 $D_i$，求和得到整个分子结构的特征向量数据$G_{structure}$，用作下一步训练。（如下图所示）




# 3 方法原理

## 3-1 输入数据结构

首先利用 rdkit 包提供的 smiles 解析器，将以 smiles 表达式格式存储的分子结构转化为分子图 Graph 数据结构(自定义无向图)，Graph 由节点(Node)构成，Node 对象对应原子，原子信息封装在 Node 中。Node 对象还存储了与之连接的其他 Node 对象引用，即“边”，对应化学键，以及这些引用的权值，即化学键的信息，同样存储于 Node 中。Node 中，原子信息用 125 维向量编码，键信息用 12 维向量编码，除相对原子质量外，其余自然特征全部用 one-hot 格式编码。其中 125 维数据包括原子类型(100)、成键数量(6)、杂化类型(5)、芳香性(1)、原子质量的十分之一 (1)、键型(4)、共轭(1)、成环(1)及立体构型(6)，其中，原子质量除以 10 是为了使之大小与其他特征相近，避免训练开始时对优化器的冲击，降低训练难度。

预测与训练中需应用分子树，可理解为具有根-枝-叶结构的有向无环分子图，同样由 Node 构成。Graph 对象通过最小深度生成树算法(定义在 Graph 类下的 build_tree 方法中)，可建立以任意 Node 为根节点的分子树，用于递归神经网络模型的预测和训练。上述自定义无向图可以按照背景介绍中的规则转化为 N(N 为总原子个数)个有向无环图(DAG-RNN)，并得到特征向量数据$G_{structure}$，完成分子结构数据的输入处理。

额外地，为了将更多的分子信息代入训练数据中，我们还搭建了 Plus 模型。在 Plus 模

型中，预测网络除接受递归网络返回的特征外，还接受 rdkit 计算出的 196 个分子全局特征，如分子量、氢键受体数、氢键给体数以及一些量子计算特征等。Plus 模型一方面增加全局特征，使模型对分子整体关注度上升，预测精度可能因此提高；另一方面，Plus 模型也作非 Plus 模型的对照，以观察递归网络得到的分子特征是否能够完全反映该分子的性质。但同时，由于数据维度的大幅提升，计算复杂度急剧提升，所以此模型目前仅做参考，不作为此大作业项目重点。

## 3-2 最小深度生成树算法

分子结构为无向图，而递归算法处理对象为树。最小深度生成树算法以指定节点为根节点，从无向图中生成具有最小深度的树。算法本身具有递归性质，每层递归中，遍历当前节点在无向图中的邻接节点，若该节点未被标记，则将该节点作为当前节点的子节点，并标记之，移动到该节点上重复以上步骤，直至原无向图中所有节点均被标记，算法结束。易证该算法可得到最小深度的树，以使递归网络深度最小，降低运算量。

## 3-3 递归神经网络

模型输入分子图(DAG-RNN)，输出 ADMET 预测值。模型可分为性质预测和特征抽取两部分，即：

$$Activity = M^O(E(strcture))$$

E 为特征抽取函数，返回一个定长特征向量，$M^O$为单隐层经典神经网络映射，返回预测值。

E 函数为求一组递归网络返回值的加和，即：

$$E(structure) = \sum_{k=1}^{N} M^E(Tree_k) = \sum_{k=1}^{N} G_{r_k,k} = (D_1, \dots, D_K)$$

$M^E$为递归神经网络映射，$Tree_k$为以原子 k 为根节点的分子树的根节点，$G_{r_k,k}$为以原子 k 为根节点的分子树中，k 原子节点的返回值向量。

$M^E$计算当前节点所有子节点在$M^E$映射下返回值向量，将这些向量拼接上键连信息向量后利用隐层循环采样，并将采样向量求和得到定长向量，与当前节点的原子信息向量拼接，通过单隐层神经网络映射得到返回值向量，数学描述如下：

$$M^E(Node_i) = Net^E\left(input_i, \sum_j Net^C\left(bond_{i+1,j}, M^E\left(Node_{i+1,j}\right)\right)\right)$$

改进版本模型简化了以上递归方式，将二重递归循环改为直接递归：

$$M^E(Node_i) = \sum_j Net^E\left(input_i, bond_{i+1,j}, M^E\left(Node_{i+1,j}\right)\right)$$

注意到这是一个递归函数，当递归进行至分子树的叶节点时，由于叶节点无子节点，网络开始逐级向上返回向量值。第一代模型仅循环键连信息，得到最终结果与原子信息合并进入下一层递归，而改进后的第二代模型，一次递归中的每一轮循环中都包含原子信息，强调键和原子的作用，同时降低网络深度，减小运算量。

与 Alessandro Lusci 等人开发出的递归神经网络原理略有不同：

$$M^E(Node_i) = Net^E\left(input_i, M^E\left(Node_{i+1,1}\right), \dots, M^E\left(Node_{i+1,m}\right), zero_{m+1}, \dots, zero_n\right)$$

由于神经网络算法常要求向量长度确定，然而分子树中节点的子节点数目不定，此递归神经网络利用零向量将剩余的特征空间补齐。本文对这一点进行了改进，利用类似于卷积神经网络的方式，对各个子节点循环抽取向量并最终加和，得到定长向量。由于各子节点共享权值，此方法大幅减少了数据维度，且此方法将各子节点放在相同地位上，更加符合实际，

也更优雅。

# 4 模型训练

## 4-1 数据集介绍

| 名称 | 描述 | 规模 | 任务 | 来源 |
|---|---|---|---|---|
| ESOL | 水溶性 | 1128 | 回归 | J. S. Delaney, J. Chem. Inf. Model., 2004, 44, 1000–1005 |
| HIV | 抗艾滋病病毒活性 | 34092 | 分类(二元) | https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data |

输入数据：分子 SMILES 表示式
目标值：分子性质，如水溶性、抗艾滋病活性
训练集/测试集比例：0.80:0.20
注：预处理阶段，HIV 数据集中 RDKit 无法解析的约 2000 条 SMILES 数据被直接删除，此外，原始数据集中分子结构采用索引编号，部分编号无法在官方数据库中找到，因此也不在训练范围内。
可以看到在 model.py 模块中有四个函数，SolNet 和 PlusSolNet 用于预测溶解度，HIVNet 和 PlusHIVNet 用于预测抗艾滋病活性，Plus 代表额外地输入数据特征(见前文“输入数据结构”)。

## 4-2 参数设置

在训练过程中参数经过多次修改，在此仅展示两代参数：
第一代参数：

1. $Net^C$网络(递归内层网络)输入 23 维(20 维+3 维键向量)，输出 20 维，激活函数 Tanh；
2. $Net^E$网络(递归外层网络)输入 34 维(20 维内层输出+14 维元素向量)，对于叶节点，输入仅 14 维，无内层输入，输出 20 维，激活函数 LeakyReLU；
3. $M^O$网络(性质预测网络)，输入 20 维，隐层 25 维，输出 1 维，激活函数 LeakyReLU，因执行拟合任务，输出无激活函数；
4. Loss(损失函数)，采用 MSE 损失函数；
5. Optimizer(优化器)，使用 Adam 算法，学习率 0.001，$\beta_0 = 0.9, \beta_1 = 0.999$；
6. 训练集共 858 组数据，测试集 286 组，占比 0.2，训练进行 8 轮。使用类似自助法采样，训练数据随机。计算机无 GPU，训练过程全部在 CPU 上进行。

第二代参数：

1. $Net^E$网络(递归外层网络)：输入 187 维(125 维原子信息+12 维键信息+50 维子节点递归特征输出)，含一个具有 150 个神经元的隐层，输出 50 维递归特征，激活函数采用 LeakyReLU。
2. $M^O$网络(性质预测网络)，非 Plus 网络输入 50 维，Plus 网络输入 246 维，隐层 100 维，

回归任务输出 1 维，激活函数 LeakyReLU，因执行拟合任务，输出无激活函数；分类任务输出 2 维，使用 Softmax 激活函数。

采用 5 折交叉验证监控训练进程，设置每批训练量为 64 条数据，此外，训练还设置了简单的提早结束规则，当交叉验证 MSE 损失小于 0.5 时提前结束训练，但事实上，模型拟合较慢，在实验范围内未触发早停机制。

碍于时间精力，我们只进行了手动调参，调参的一些改进措施在最后“讨论分析与改进建议”部分依旧会提到。

# 5　模型评估与分析

概况：采用 5-fold 交叉验证训练，分别使用 RMSE、R2 等 evaluation metrics 初步评估模型；同时采用学习率曲线等评估模型是否能合理拟合(无欠拟合或过拟合)，并采用合适的训练数据/测试数据比例，进行更深入的拟合。由于模型复杂度比较高，我们主要评估模块为 SolNet。

由于模型复杂度较大，程序训练难度较高，SolNet 和 PlusSolNet 模型对于个人电脑以达算力极限，因此我们能得出的有效模型结果并不多，所以评估结果也实在有限。后期如有必要，我们将使用超算资源进行调参和评估。

大致评估结果如下：

1. SolNet 和 PlusSolNet 完成回归任务，受限于结构复杂度与计算复杂度，PluSolNet 准确度更差一些；
2. HIVNet 勉强完成分类任务；PlusHIVNet 训练难度过大，计算资源不足而训练失败

为提供一个更清晰的评估结果，同时考虑到本模型的主要目标是预测水溶性，下述评估主要是针对 SolNet，并附加上少量 PlusSolNet。

## 5-1 RMSE 与 $R^2$

1. 首先看 SolNet 跑 10 个 epoch 和 20 个 epoch 的结果(见图 1 与图 2)：可以看出随着数据集被反复使用的次数增加，模型所能达到的准确度有所提升。在 10 epoch 运行条件下，R2 score 为 0.718，RMSE 则为 1.028；而在 20 epoch 的运行条件之下，R2 score 达到了 0.828，还算可以；RMSE 则为 0.802，同时可以从图二中看出，散点图大部分依旧在基准红线附近。因此也能说 SolNet 完成了回归任务，但目前效果一般，还需要进一步增加 epoch 数量和计算复杂度。

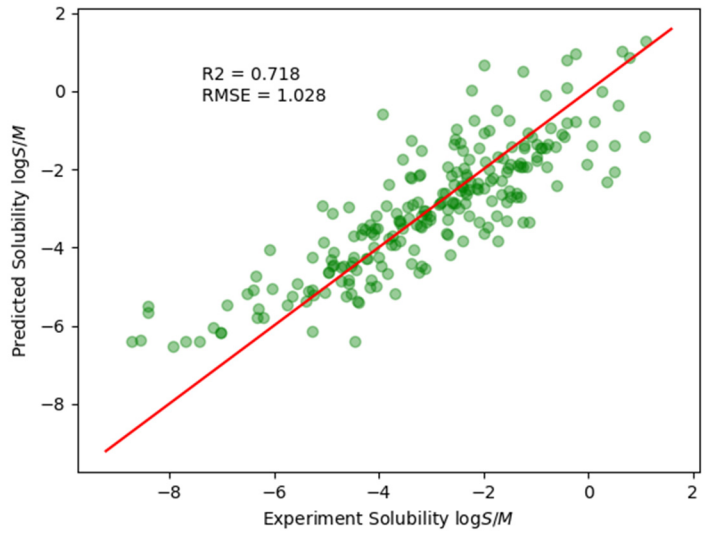



**Figure 1: SolNet Results (10 epoch)**

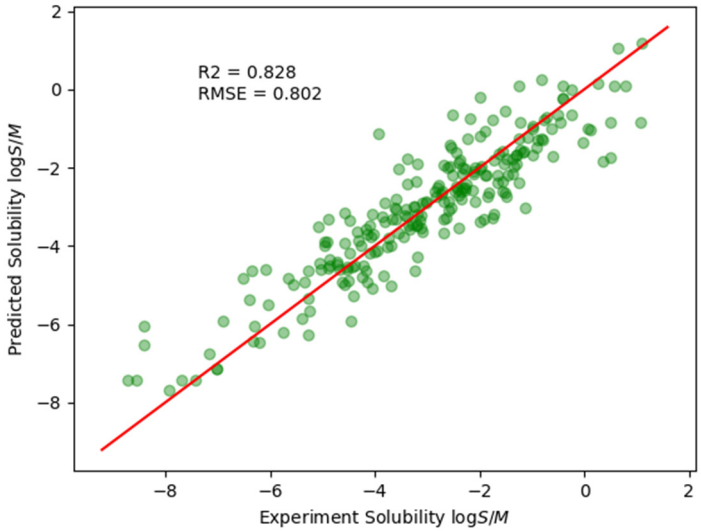



**Figure 2: SolNet Results (20 epoch)**

2. 再来看 PlusSolNet 的结果(见图 3，只跑了 10 epoch)：
可以很清楚地看到，R2 score 只有 0.573，相比于 SolNet 在 10 epoch 下 0.718 的结果差了一些；RMSE 也更大，为 1.265；且预测值在 target 值较低偏离平均值时，表现较差。可见，对于更高维度的 PlusSolNet，10 epoch 的训练量远远不够，尤其是图三中预测数据具有明显下界，说明众多神经元未被训练到。

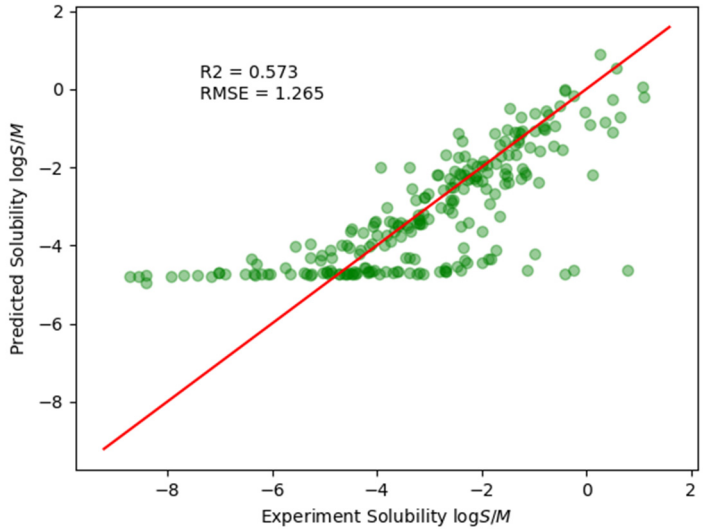



**Figure 3: PlusSolNet Results (10 epochs)**

## 5-2 学习曲线 Learning Score Curve

学习曲线就是通过画出不同训练集大小时训练集和交叉验证的准确率，可以看到模型在新数据上的表现，进而来判断模型是否方差偏高或偏差过高，以及增大训练集是否可以减小过拟合(如图 4 所示)。

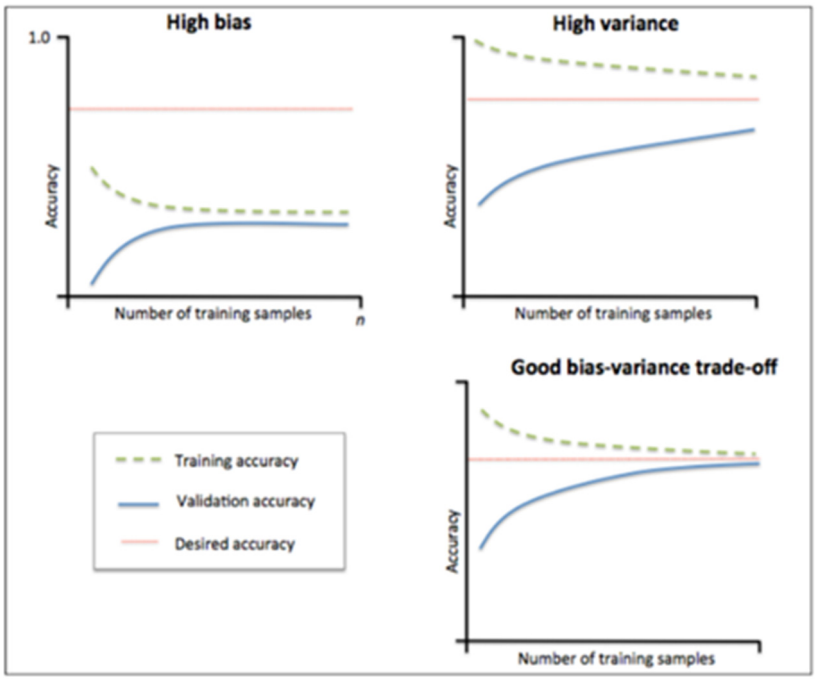



**Figure 4: Learning Curves**

我们在模型中只使用训练集与测试集。通过改变训练集与测试集之间的比例，比较模型在训练集与测试集之间的 R2 score，来看我们的模型 fit 程度如何，是否有过拟合或者欠

拟合。最终结果如图 5 所示，只能算粗略结果不过。

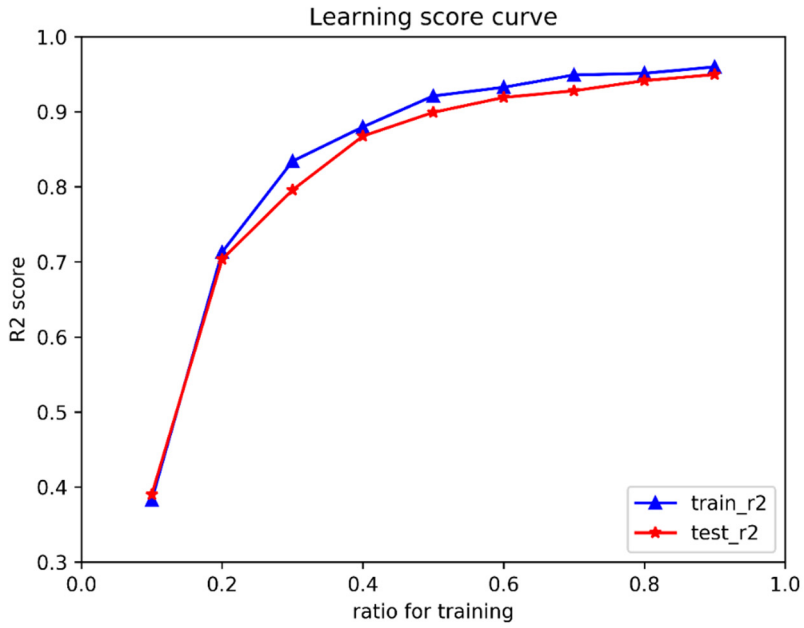



**Figure 5: Learning Score Curve**

每个 ratio 对应的 score 是两次运行得到的结果的平均值，受限于算力，只选取了跑两次选平均值，每次只跑 10 epoch(这段评估程序在电脑上跑下来需要一天多，再加上调 bug…)。

将常规的欠拟合/过拟合/拟合得到的曲线结果与此结果对比，可以看出此学习曲线比较迷——训练集与测试集的 r2 score 相关性过于紧密。这可能与模型没有完全收敛有关，导致即使在训练集很少的情况下，训练集的准确度反而更低。当训练集比例在 0.40 以上时，可以看到模型准确度已经在 0.88 一线附近了，说明模型能够在训练数据量偏少的情况很好地进行学习。目前暂时无法评估模型是否过拟合(可能性较小)，就姑且先看作能较好地拟合吧。

根据上述学习曲线，可知在训练集比例大致在 0.6 时已经能较好地完成拟合任务，因此选取 training rate = 0.60，针对 SolNet 以 10 个 epoch 为一个循环进行多次循环，观察 R 2 score 的变化，进一步评估模型(如图 6 所示)。

可知随着 epoch 次数的增加，模型的收敛程度增加，模型的准确度也有所上升，最终 t rain_r2 达到 0.94，test_r2 也达到了 0.88；在跑了 80 个 epoch 之后仍有上升趋势。预估最终 test_r2 能达到 0.90，这也与原文献中的 0.91 相接近。从这些数据来看，在模型能够充分收敛的情况下，模型的准确度还是挺好的。

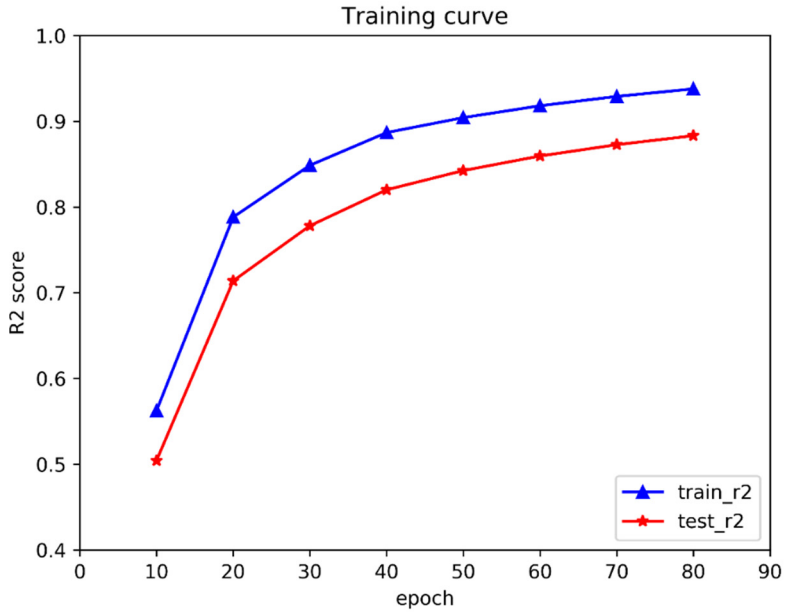



**Figure 6: SolNet Results after Multiple Epoches**

总的来说，虽然在前两部分评估测试中，模型受限于计算精度与收敛成都，所得到的结果不怎么让人满意，但第三部分的评估测试则让我们看到了希望。我们也有理由相信，在利用更大的计算算力的基础之上，我们能将模型完善得更加优秀！

# 6 改进建议

## 6-1 针对调参

1. 本工作的调参任务包括神经元数量、激活函数、外层网络深度等，但由于深度网络计算量大，我们只进行了一轮手动调参，效果有待改进；
2. 网格调参和随机调参：网格调参在参数空间内取等间距点进行试验，随即调参在参数空间内随机取点，每轮调参之间相互独立，在大量试验情况下可获得科学的结果，hyperopt 模块提供了调参接口；
3. 贝叶斯调参：调参过程本质与机器学习相同，尽可能使得分最优，每轮调参对下一轮有指导，相比于上述过程更加高效

## 6-2 针对分子表示方法

关于 Molecular Descriptors，除本文献中所提供的方法之外，还有很多其他的方法可以尝试。描述分子的方案主要分两类，一类是静态的分子指纹，通过确定规则对分子进行计算，得到唯一性序列描述；另一类是基于机器学习的特征抽取，得到分子特征的过程需要学习：

1. 比如，Gilmer 等人总结的 MPNN 方法，以及 Kevin Yang 等人开发出改进型 Directed MPNN，与此处的模型十分相近。与本工作不同的是，MPNN 的模型并非通过递归顺序运算，而是通过时序迭代的方式，通过控制迭代次数调节网络输出特征的整体性。这样的作法优点是对于较大规模分子，可控制网络深度而节约计算量，而递归网络的复杂度和深度将会因分子规模变大而爆炸式增长，但 MP

NN 若迭代次数过小，得到的特征将具有强烈的局部特点，不能正确反映分子整体的状况。

2. 此外，也可尝试使用径向基函数的方法，抛开深度学习学习网络，用简单的机器学习算法(NN, SVM 等)尝试预测分子性质。

### 6-3 针对计算复杂度与深度

为尽量减小计算量，本模型可能收敛得并不完全就直接输出结果，尤其是在训练数据量小的情况下，因此模型准确度并不高。条件允许的话(超算)，可以多跑一些 epoch 提高准确度；同时也可以调参增加网络复杂度，从而提高准确度。

### 6-4 针对训练数据

1. HIVNet 和 PlusHIVNet 的失败，一方面是训练难度过大，另一方面是高度不平衡数据集引入的额外难度。训练中采用了不平衡权重的损失函数，即调高了阳性数据的影响，导致过于稀疏的阳性数据对优化器造成不连续的冲击，模型难以收敛。通过调整训练集阳性数据比例可使这一情况得到改善，一种可行方法为过采样，即大量复制阳性数据并添加到数据集中。
2. 由于本模型用到的是分子拓扑特征，不能反映三维结构。通过增加分子空间构型描述可能提高模型的表现。

### 6-5 针对评估

1. 对于回归任务，R2 score 与 RMSE 侧重点不同。前者是拟合度的体现，越接近 1 拟合程度越高，后者是评估结果标准差的估计，展示预测值的分布。
2. 本人在搭建模型过程中对 N 折交叉验证的理解有误，错将该方法当作训练手段，而另外单独设置测试集进行评估。事实上，这一方法是模型评估方法，不需要另外设置测试集。发现这一点时评估工作已基本完成，重新进行评估工作量很大，因此权且以训练集上的交叉验证得分作为评估结果；
3. 对于 HIVNet 和 PlusHIVNet，使用交叉熵作为损失函数进行训练，评估时应着重考察回收率而不是精度。具有抗 HIV 潜力的分子较为宝贵，应尽力避免漏过。

## 7 致谢

感谢刘志荣教授在本课程上的亲情奉献，以及助教林康杰学长在前沿领域的热情指导，林学长提供了该领域的前沿内容讲解，并对我们的工作进行了认真负责的指导。若非与林学长的讨论，改进后的模型将不会出现。此外，也感谢小组成员通力合作，在算力不足的情况下，仍耗费大量精力和时间对该模型进行了积极改进和完善。

## 8 小组成员及分工

| 姓名 | 学号 | 院系 | 主要任务 |
| --- | --- | --- | --- |
| **潘高翔 (组长)** | 1700011756 | 化学院 | 收集数据库，模型的搭建和维护，以 |

| | | | |
|---|---|---|---|
| | | | 及相关训练和部分评估方法的开发，最终报告的修改 |
| **付思杰** | 1600011073 | 工学院 | 模型训练和评估，最终报告的撰写 |
| **蔡丹杨** | 1700011774 | 化学院 | 模型的训练和调参工作 |

**注：①主要分工并不代表只参与该项工作；②全员参与讨论；③蔡丹杨和付思杰的任务还包括向潘高翔学习深度学习和模型接口等。**

# 9 参考文献

[1] Alessandro Lusci; Gianluca Pollastri; and Pierre Baldi, Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules, *J. Chem. Inf. Model.* **2013**, 53, 1563−1575;

[2] Kevin Yang, Kyle Swanson, Wengong Jin, Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, 59, 3370−3388;

[3] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* acs.jmedchem.9b00959