大家好，我们组的主题是

(protein binder site prediction) 蛋白质结合位点预测

下面将会从一下几个方面介绍我们的工作

background

我们都学过：蛋白质是生命活动的主要承担者

蛋白质通过与**蛋白质**或者其他小分子的结合，以发挥生命活动的功能

因此**预测**蛋白质之间的结合位点尤为重要，可以帮助我们预测蛋白质的性质、起到药物研发(drug discovery)等一系列重要的工作的作用

所以我们要做的其实是AI4science 中**蛋白质结合位点预测**任务中**蛋白质口袋分类**的工作

也就是说，我们其实做的是就是对结合位点一共有7种类型的一个分类任务。

在大多数生物过程中，结合位点预测是预测**蛋白质与蛋白质**结合性质的重要一环

只根据蛋白质结构来预测这些相互作用仍然是结构生物学中最重要的挑战之一。

dataset

我们采用的蛋白质结构数据集为PDB(protein dataset bank)，里面有大量的蛋白质结构数据

我们可以看到里面存的都是蛋白质原子的**3维空间坐标**，以及各原子之间的连接关系

所以我们可以把蛋白质看作是一个3维的图结构

所以一个很intuitively的想法就是通过GCNN来进行预测

但是GCNN忽略了表面所带来的物化性质，所以效果较差

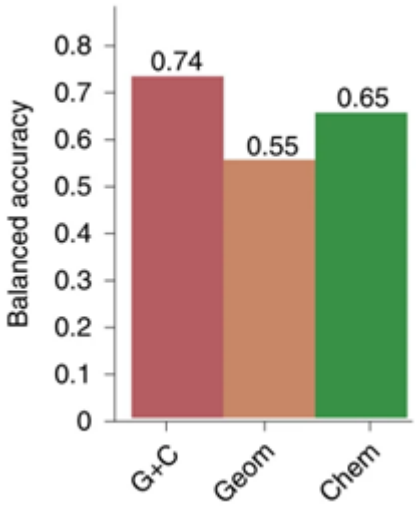
MaSIF

当前一个工作是MaSIF (molecular surface interaction fingerprinting)，通过对分子表面相互作用指纹识别，来利用蛋白质表面的几何特征以及物化性质进行结合位点的预测

分子表面是蛋白质结构的一种高级表示，它将蛋白质建模为具有几何和化学特征的连续形状

通过利用蛋白质表面的

fingerprint 作为 feature，masif做结合位点的分类任务取得了远好于GCNN的优秀效果



Ours

总结一下前面的两种方法

GCNN利用了空间结构，MaSIF利用了表面信息

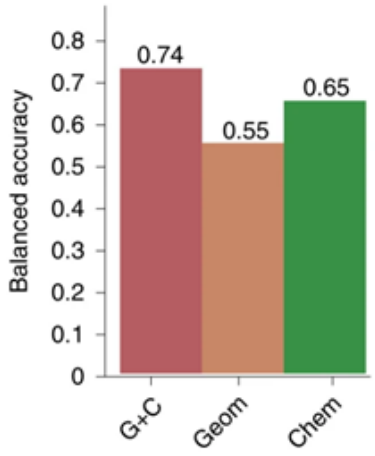
我们充分结合 GCNN所利用的结构信息 & MaSIF 所利用的表面信息)以追求更好的效果

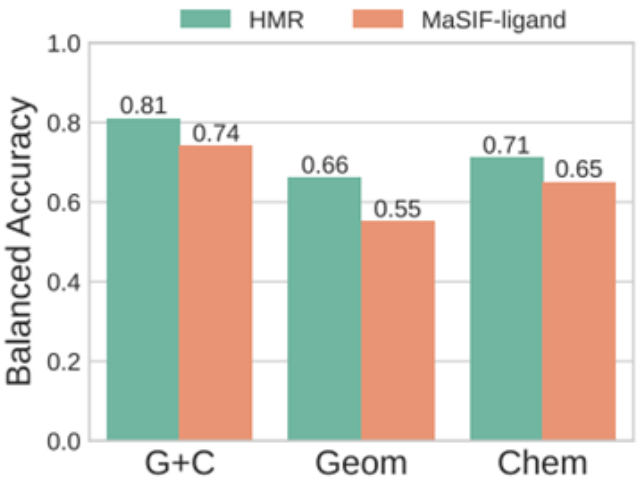
首先利用3维结构信息，结合SDF（有向距离场）建模出表面

然后利用表面信息：对表面做谐波分解，提取处他们的**几何化学特征，也就是表面函数**。

- 先前我们了解到，分子表面特征由几何特征和化学特征组成。那么我们如何学习分子表面的特征呢？以往的研究使用GNN在分子表面传播特征信息来编码在不同区域的特征。我们使用MLP来编码几何和化学特征，再最后通过另一个MLP处理获得一组新特征。这些特征被认为是表面函数。
- 在获取表面函数后，我们使用cross attention的方法来编码两个表面之间的关系，继而分类出两个表面上的不同结合位点，最后通过位点确认两个分子的结合方式。

这样我们就实现了将结构信息与表面信息的结合，并且得到了效果上的提升。





一共7类，远超random 14% ($\frac{1}{7}$)，比当前的baseline也有了近10%的准确度提升