蛋白质口袋配体结合分类

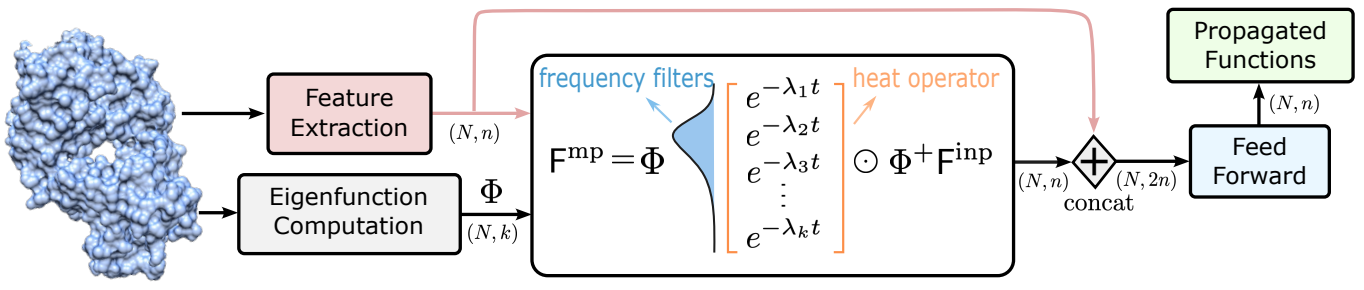
**Introduce:**

蛋白质口袋(protein pocket)用于描述蛋白质分子中特定的结构域，它能够与其他分子（例如配体）结合形成复合物。这个结合口袋通常是蛋白质分子的三维结构的一部分，具有特定的化学性质和几何构型，以便与特定类型的分子相互作用。研究其结构和性质对于理解蛋白质的功能以及开发药物设计等方面具有重要意义。

**Dataset:**

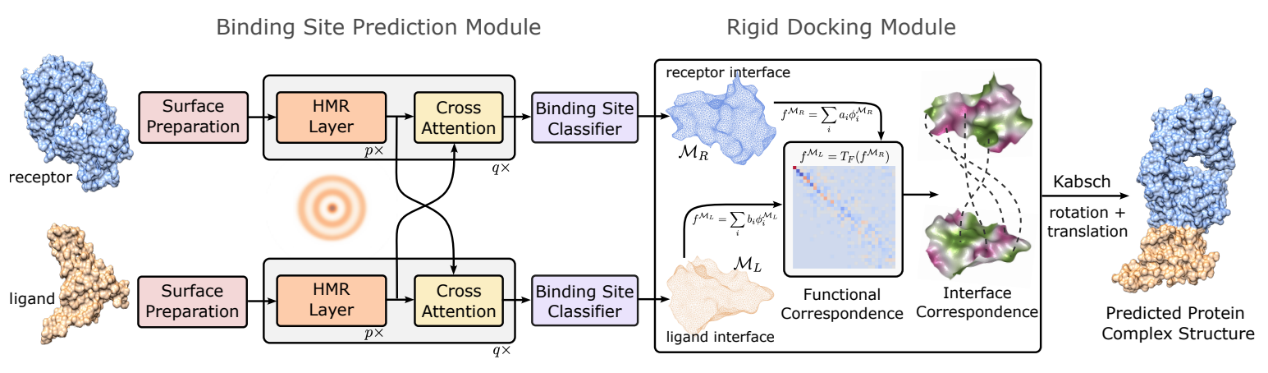
数据集来源于Gainza等人的研究（2020），其中包含了1,438个非冗余蛋白质结构。每个结构对应于一组带有其原子坐标的结合配体。  
HMR框架的原始输入仅为3D原子结构（例如，小分子配体的xyz文件以及蛋白质的PDB文件）。  
  
{   
"protein\_structure\_id": "ABC123",   
"protein\_structure": "PDB format data...",   
"ligands": [   
{ "ligand\_id": "LIG1",   
"atom\_coordinates": [   
{"atom": "A1", "x": 1.23, "y": 4.56, "z": 7.89},   
{"atom": "B1", "x": 2.34, "y": 5.67, "z": 8.90}, ...   
]   
},   
{ "ligand\_id": "LIG2",   
"atom\_coordinates": [   
{"atom": "X1", "x": 3.45, "y": 6.78, "z": 9.01},   
{"atom": "Y1", "x": 4.56, "y": 7.89, "z": 1.23}, ...   
]   
},  
 ...  
 ]   
}  
  
·HEADER: 描述文件的头部信息。  
·ATOM 或 HETATM: 描述一个原子。ATOM 用于描述蛋白质或核酸的原子，而 HETATM 用于描述非标准残基（如水分子等）的原子。  
·1: 序列号，唯一标识一个原子。  
·N: 原子名称。  
·GLY: 残基名称。  
·A: 链标识。  
·1: 残基序号。  
·11.234 45.678 23.456: 原子的 x、y、z 坐标。  
·1.00 20.00: 原子的温度因子和占有率。  
·N: 元素符号。  
·TER: 链的结束。  
·END: 文件结束。  
  
HEADER SAMPLE PDB FILE   
ATOM 1 N GLY A 1 5.234 8.678 12.456 1.00 20.00 N   
ATOM 2 CA GLY A 1 6.345 9.789 13.567 1.00 18.00 C   
ATOM 3 C GLY A 1 7.456 10.890 14.678 1.00 16.00 C   
ATOM 4 O GLY A 1 8.567 11.901 15.789 1.00 14.00 O   
ATOM 5 CB GLY A 1 6.789 9.234 14.901 1.00 16.00 C   
ATOM 6 N ALA A 2 7.123 10.567 11.890 1.00 20.00 N   
ATOM 7 CA ALA A 2 8.234 11.678 12.567 1.00 18.00 C   
ATOM 8 C ALA A 2 9.345 12.789 13.678 1.00 16.00 C   
ATOM 9 O ALA A 2 10.567 13.901 14.789 1.00 14.00 O   
TER END  
  
数据处理步骤：  
·生成蛋白质表面网格：通过描述的步骤生成蛋白质的表面网格。  
·提取配体结合口袋：通过在表面网格上识别与任何配体原子的距离在4 Å以内的口袋顶点，提取最大连接组件作为结合口袋。移除包含小于100个顶点的口袋。  
  
数据集统计：  
·共提取了2,254个结合口袋，分别与七种配体（ADP、HEM、FAD、COA、NAD、SAM、NAP）中的一种结合。  
·平均每个口袋表面获得1,199个顶点和190个特征。  
·数据集按照与Gainza等人（2020年）相同的训练、验证和测试划分，分别包含1,634个训练口袋（986个蛋白质复合物）、202个验证口袋（112个蛋白质复合物）和418个测试口袋（274个蛋白质复合物）。  
  
输入特征：  
·提供了化学信息（亲水性得分、部分电荷）以及局部几何信息（平均/高斯曲率和Heat Kernel Signatures）作为神经网络的输入特征。

**Algorithm:**



算法的第一部分是将分子表面转化成一系列表面函数。

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（多重感知机）来编码几何和化学特征。接着使用一个滤波器以及一种与热扩散模型类似的方法，来传播特征信息。最后通过另一个MLP处理几何和化学特征，以获得一组新特征。这些特征被认为是表面函数。

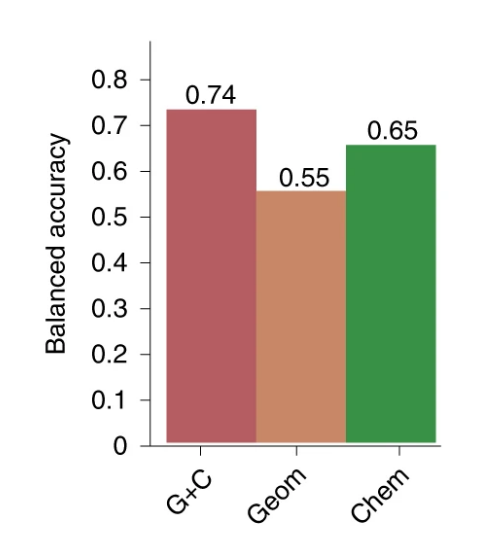


第二部分，我们通过学习表面关联度来进行分子匹配。

两个分子能相互作用源于它们有相似的几何化学特征，也就是表面函数。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。

**Baseline:**

MaSIF-ligand



MaSIF是一种基于卷积神经网络的模型。图上分别是考虑几何及化学特征、只考虑几何特征和只考虑化学特征的三种情况下对于测试集分类的准确率。我们希望能以此位基准来衡量我们之后工作的有效性。