记录：

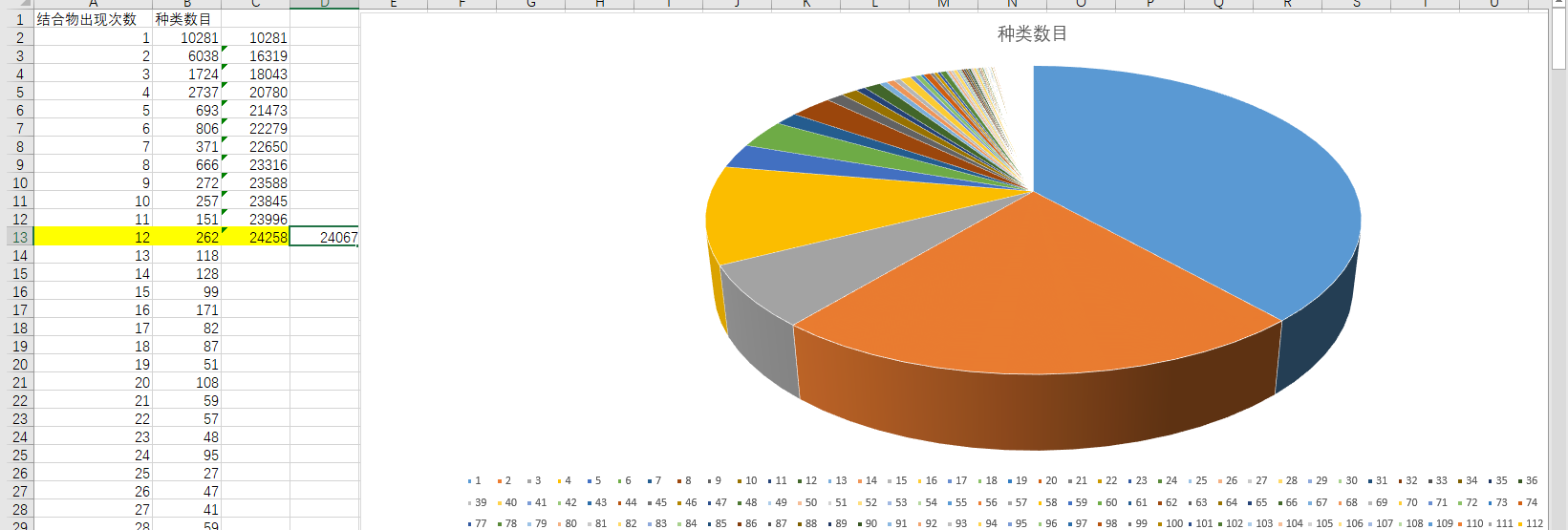
1. 数据提取过程及思路。
2. 数据提取相关代码。
3. 数据整理及分析。

首先，在pdb20190529版数据库中，约有15万条数据。其中约有3.7万个没有包含“site”相关的点和信息。

分析REMARK800 中，description中的有关点的记录。统计有哪些结合物、以及每种结合物出现过几次。

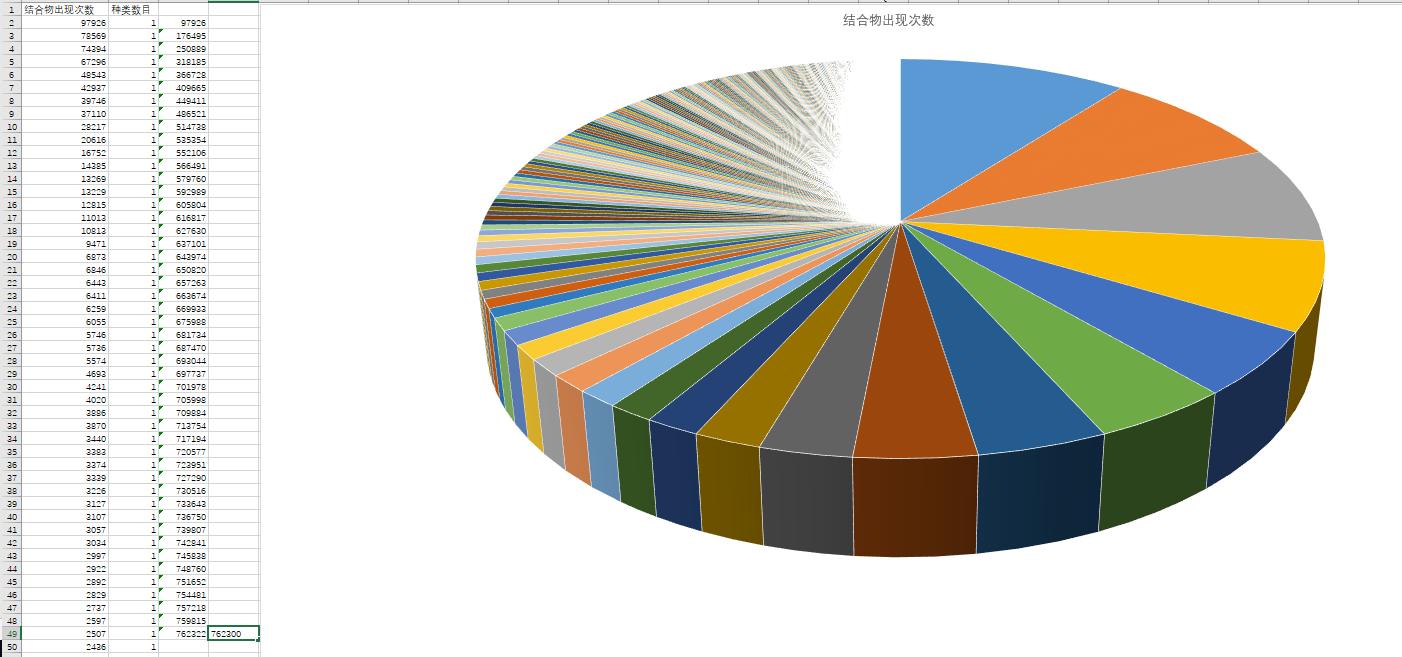
针对所有 含有“site”记录的蛋白质，我们提取了其中每一条序列中的结合物信息。提取的序列有114859个序列。还有1673个序列的description为“NULL”

所有提取到的”binding site”种类有26782种。每个种类都有多次记录，共计1089001次记录（平均每种结合物出现了40.742次）

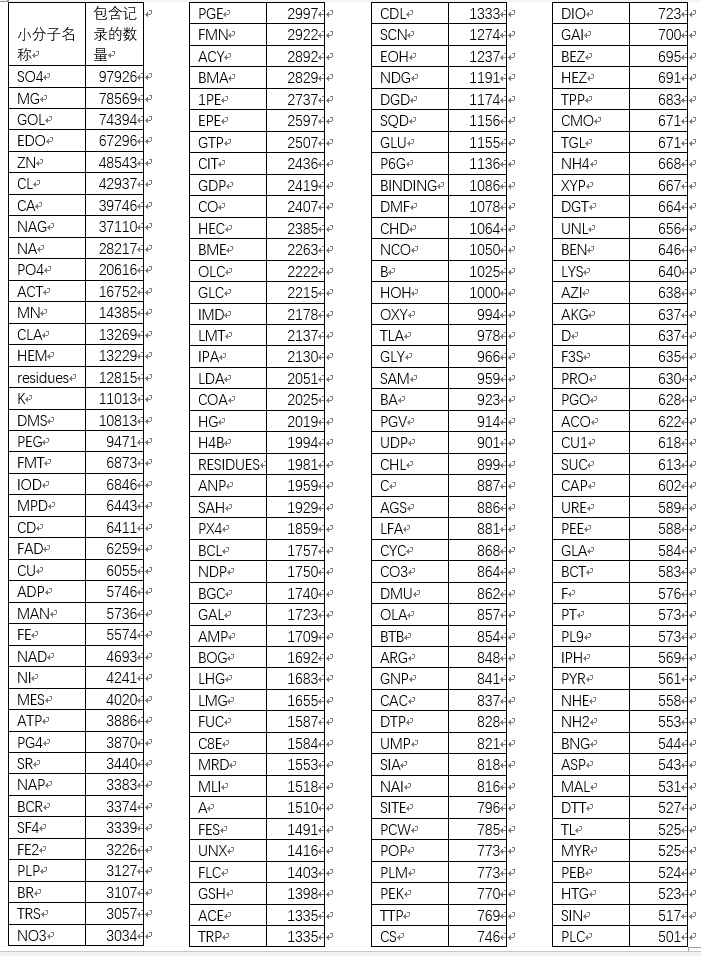


根据饼状图我们可以得出的结论是：

1. 在26742个种类的结合物中，有大概超过90%的种类，出现次数少于12次（26742\*0.9 = 24067），也就意味着，大部分的化合物记录是小众的，它们仅仅被记录了少数的几次。

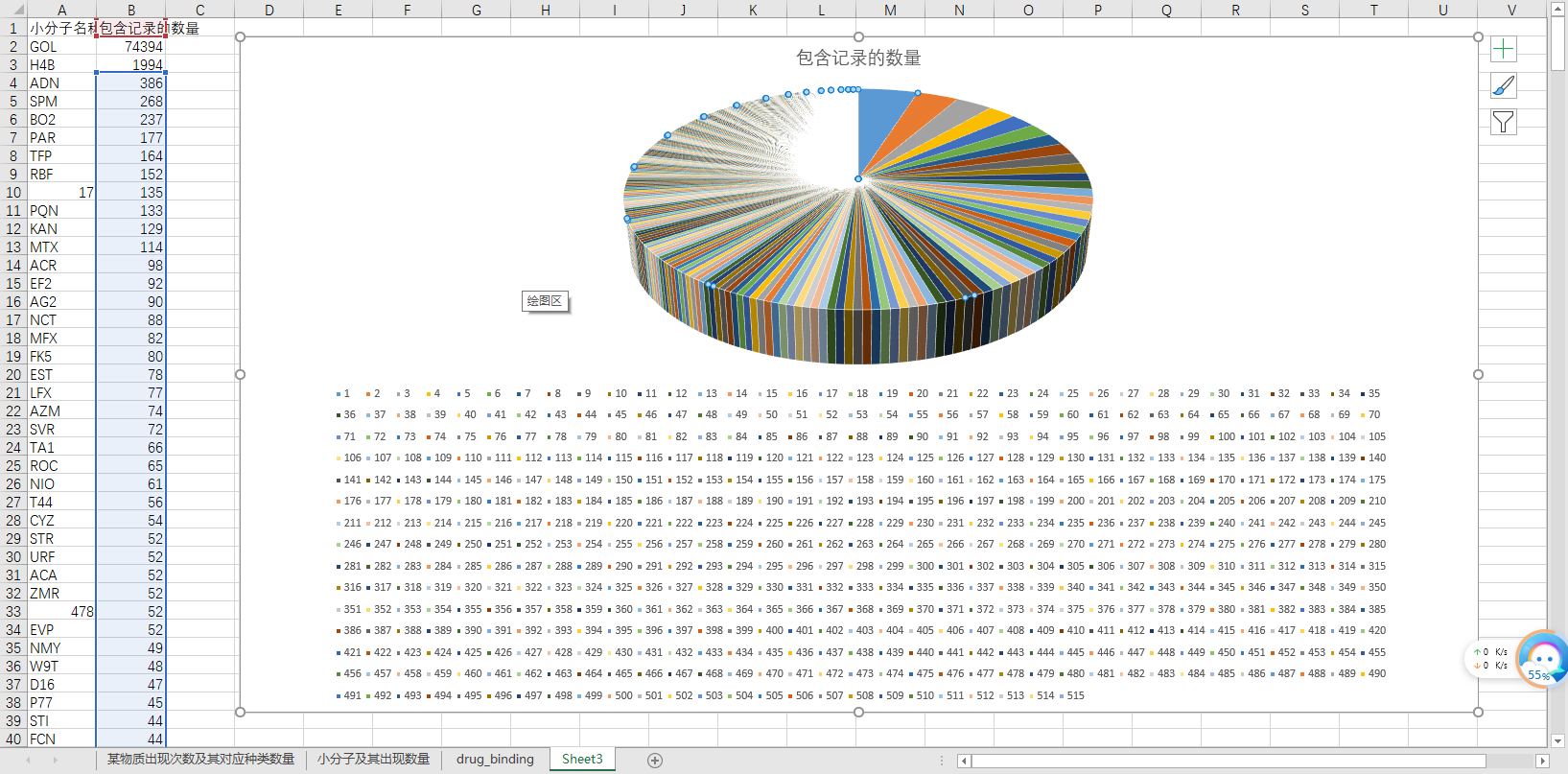


1. 而根据结合物出现次数排序时，根据饼状图我们可以看出：记录结合点最多的50种化合物，占据了约70%的结合物出现记录。这意味着少数结合物，拥有充足的结合位点信息。



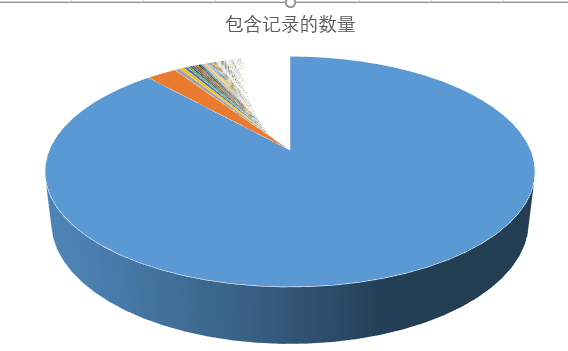
列出靠前的分子。

若是出现在528种药物列表中的小分子：如图所示。



在drugbank中列出的528种小分子标识，出现在pdb中的记录只有516个（还有12个pdb中无记录）共计出现84246次记录。

其中出现了极大的数据不平衡显现：GOL有74394次记录，占全部的88%。第二名1994占全部记录的2%.



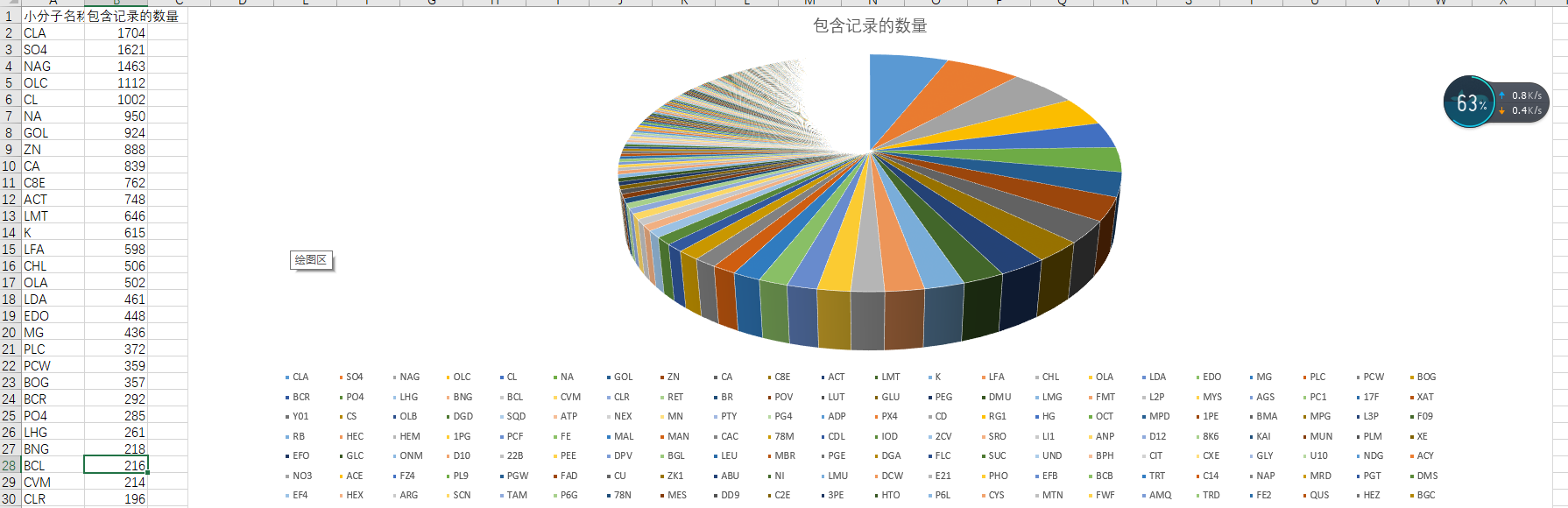
画风变成这个样子（如果增加前两位的话。）

膜蛋白点位提取记录

在所有信息中，筛选膜蛋白并进行点位记录。

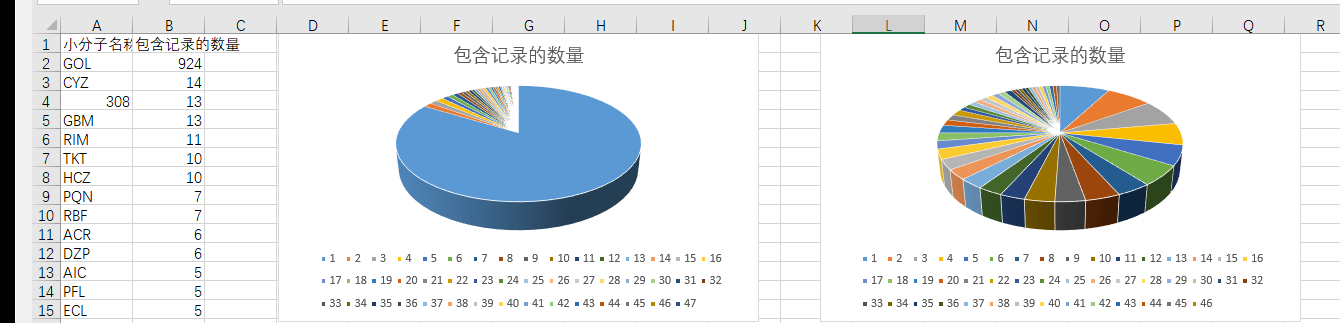
一共在1873条蛋白质序列中，找到784个种类的结合物。共计28186次结合物的记录。

（全部的是114859条蛋白质序列中，找到26782个种类的结合物。共计1089001次结合物的记录。）



他们的分布大致如此。。

如果根据药物来筛选以下，得到此图：



如果预测GOL的话，则数量还是很多的：924次记录。要是预测膜蛋白其他小分子，则数据量非常少。左图是包含GOL的记录，右图是不包含GOL的记录。

可选方案：

1. 延续之前思路，预测膜蛋白小分子药物：将会遇到数据量少、正样本点少的问题。
2. 转换一下，预测膜蛋白结合物，则数据充裕，但是和药物不沾边。
3. 若预测蛋白和药物结合位点，这部分数据充裕，有2个方向：第一，预测GOL结合位点（但我猜测GOL是溶剂，甘油，需要针对其意义进行探究）。第二排除GOL数据，寻找其他小分子结合物。
4. 预测蛋白质结合位点：这个数据量可是很大的。

结合物记录出现次数：

|  |  |  |
| --- | --- | --- |
|  | 蛋白质 | 膜蛋白 |
| 全部种类小分子 | 1089001次记录 | 28186次记录 |
| 药物小分子 | 84246次  （74394次记录是GOL）  还有9000次各种小分子记录 | 1104次  （924次记录是GOL） |

我的想法是这样：1在全蛋白中寻找药物小分子。2在膜蛋中寻找GOL。