1.数据预处理文件[0\_Data\_PreProcessing]:

数据库总共有去重滋味肽483条，去除没有经过验证的剩下474条。再丢弃滋味矛盾组[比如又bitter 又umami】，剩下437条。其中2\_3肽 有 203条【2肽99条（31条鲜味，68条苦），3肽104条（53条鲜味，51条苦味）】，4肽及以上234条。其中2-3肽中鲜味x条，苦味x条，不成比例，所以采用数据增强。

然后制备对应的分子描述符【Rdkit，的208个+人为指定的一些+氨基酸前后顺序类】。

明天的任务：

①特征可视化+做数据EDA探索——顺带特征筛选

全部特征：279 → 方差筛选：208 →K-Q分布检验：剩下57个：→卡方检验：剩下51个→特征递归消除得到8个特征

②数据增强【数据增强一定要在特征筛选后面做，是因为这样可以保留原始特征的信息】

这个明天有空也可以做成对应的包，直接调用

③模型选择-已经做到pip install 里面啦

④超参数搜索

⑤模型评价，模型鲁棒性评价【用之后的来检验模型】

和另外两个模型比较，那几个指标 + 马修斯矩阵进行计算，sklearn可以算

⑥模型可解释性分析

SHAP ELIF 还想试试新的方法 还有一个方法需要看 pptx

#####。和其他类似的预测网站进行结果对比：

对比从两个角度：

1⃣️是： 准确度

2⃣️是 便捷度：Iumami这种预测前，还需要单独写一行无所谓的注释，实在是很无聊。但是他支持传递多一个序列，这个值得学习

模型比较方面，由于两个服务器都是

网页展示思路

图形用户界面, 应用程序

描述已自动生成

基于官能团的分析方法：[Ido Nissim](https://iubmb.onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Nissim%2C+Ido)等人研究认为黄酮类化合物、生物碱、硫代葡萄糖苷、磺酰胺、萜烯等物质与成分常常与苦味有关[3]。【下图放入PPT】

图示

描述已自动生成