## 2024.11.1

1. 在Aviv的数据集上跑一下bagging，跑起来比较慢，已经一天了。

2. 师兄提供了算力，找jiaqi师兄要L40S的算力，可以联网，但不清楚是否只能在fit楼的网上使用。

3. 算法需要修改一下，计算overlap时只在validation上计算overlap，保证每一次实验都在相同的数据上计算overlap。

4. 不要把DKL和bagging加在一起，轮子套轮子没有好结果，对算法进行优化要比对模型进行优化来的好。

5. 科学问题：在序列数据上个使用主动学习的方法进行表达量预测。另外，之前没有工作总结过在序列数据上做主动学习的benchmark。

6. Enamine的数据本身不是序列数据，而是fingerprint，不具有说明性。

## 2024.11.2

1. 把summary/bagging中增添了Aviv数据集的接口，同时将在all\_data上做重合度计算改成了在valid\_data上做重合度计算，删去训练过程中在validation上计算pcc的部分简化训练过程。更新：在summary的三个文件夹中main函数均做了这项修改。

2. problem：关键点不在于找到一个合适的不确定度量化方法，而在于找一个合适的指标来指导主动学习，这个指标可能是不确定度，也可能不是。即便能完全量化出不确定度，这也不一定能提高主动学习的性能。**已有的实验内容不能说明ensemble得到的不确定度是否有用，也不能说明不确定度对于主动学习是否有性能上的提升。**

3. EI方法：按照道理来说，使用EI筛选出来的数据应该有较大的预测均值和较小的预测不确定性。

4. 重大bug：筛选数据时没有把valid\_data从数据池中分离出去，已经对summary中的代码进行了修改。除此之外的没有修改。

## 2024.11.5

1. 在Aviv上使用GP跑出现了问题，原因是cpu内存太小，导致无法顺利完成训练。具体出问题的点在于训练后对train pool预测的地方，或许可以将这一过程进行分布计算。

## 2024.11.6

1. 把之前在所有数据上计算overlap的图像更名为summary\_overall，在validation上计算overlap的图像命名为summary\_valid。

2. 为了便于同时训练，在summary文件夹下新建了bagging\_no\_aviv，主要区别在于去掉了aviv数据集用到的seq2onehot接口，防止对原有三个数据集训练的影响。

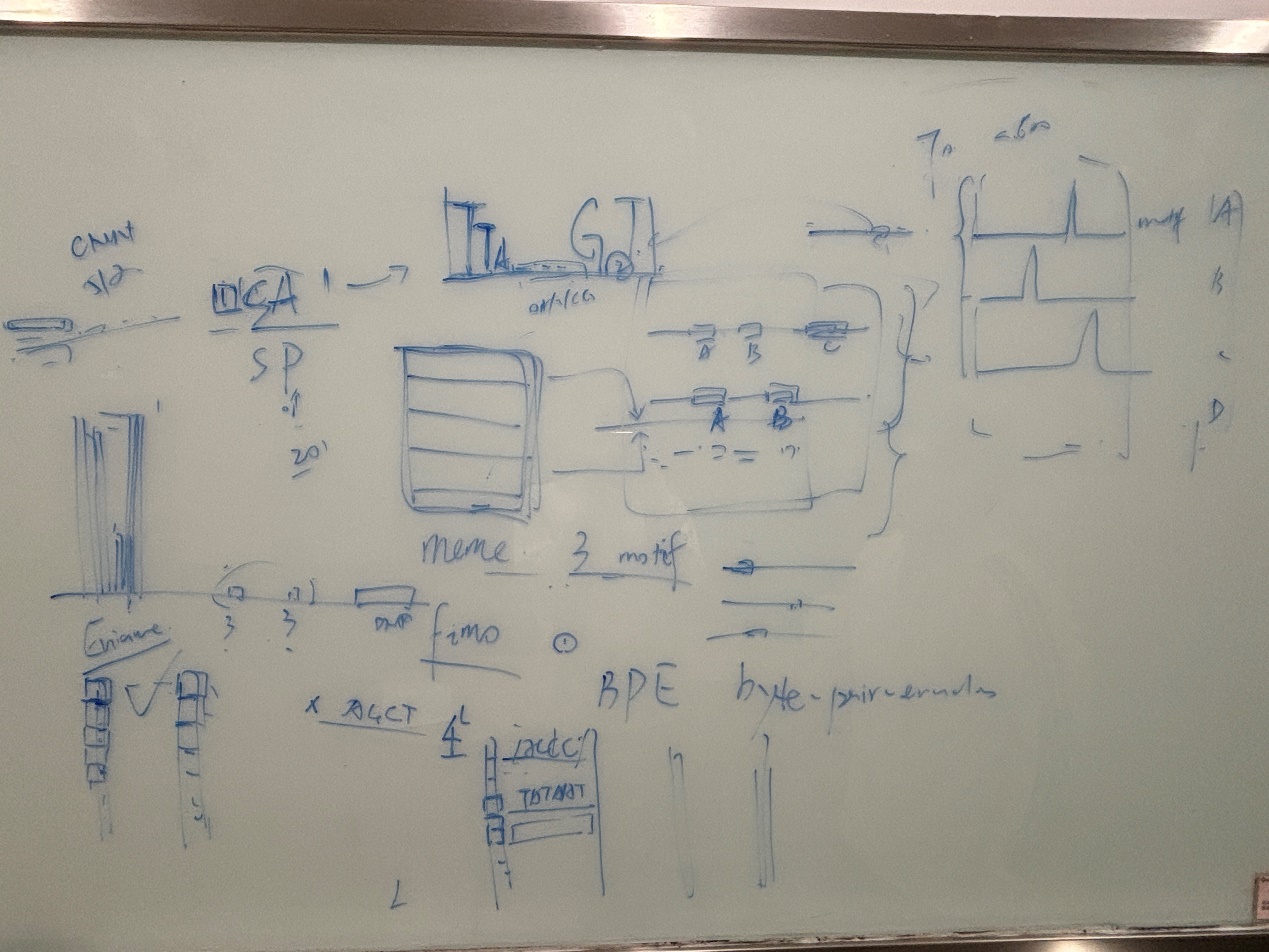
3. 更改了Aviv数据集的命名，改为Yeast\_aviv\_2022，在所有txt结果文件中也进行了修改。

## 2024.11.8

1. 和汪老师交流：

* 确定问题是解决根上的问题（设计一种适合所有数据集的主动学习算法，对数学要求高），还是解决枝上的问题（设计只在生物数据集上有价值的算法）。生物数据由进化决定是一簇一簇的，这不是随机采样能采出来的。生物主动学习的实现形式（人机闭环）决定了其迭代次数小，迭代数据少。
* 多读文章，每天抽出两小时读文章。边看边试方法，有目的地试方法，形成对主动学习的认识。
* 把毕设当做解决问题中的一环，而不是简简单单一个设计。

2. 和杜师兄交流：



* Byte pair encodings，在NLP中找一些可能的对序列空间表征的方法。
* Meme和fimo，两个搜索序列中重复度高motif的模型，或许可以用于提取语义。
* Enamine在DKL方法下很好，可能是因为fingerprint对序列空间进行了降维表征（GP和DKL对于高维数据的学习性能不好）