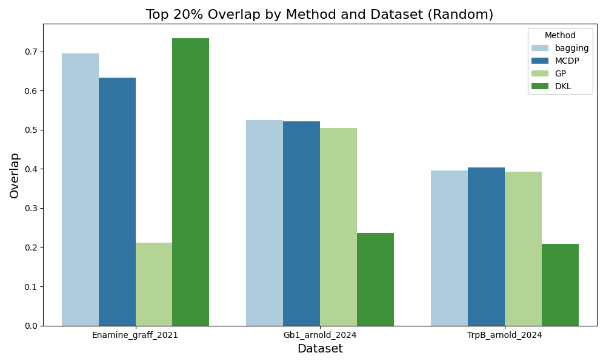
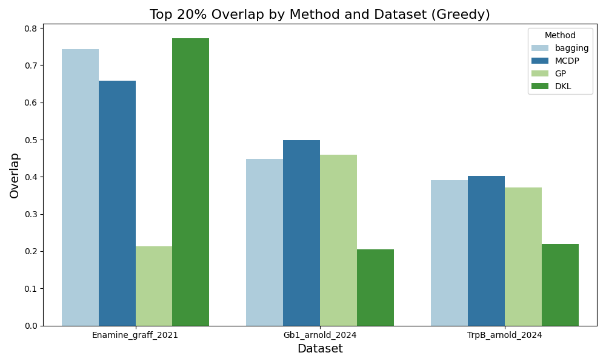
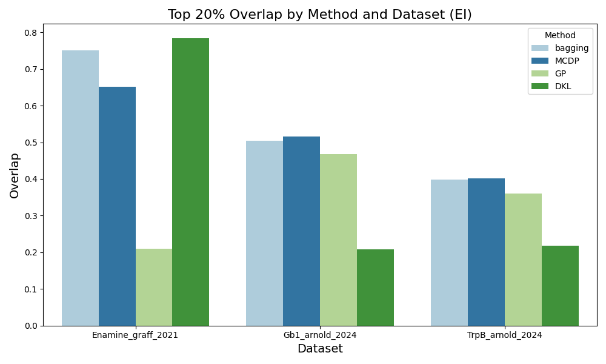
## 10.29.2024

完成了EI、Greedy、Random三种采样方法以及bagging、MCDP、GP、DKL四种算法小下在Enamine10k和Arnold两个蛋白集上个的主动学习代码编写，并开始运行。记录训练过程中overlaps和最后计算出来的收敛overlap，对收敛overlap进行做图，结果为result/summary/result\_EI(Greedy/Random).png。

## 10.30.2024

昨天出了一点点小bug，没有把图全部画出来。（更新使用Paired配色）



Bagging V.S. Monte Carlo Dropout：

MCDP相当于Bagging丢掉一部分神经元，猜测在复杂的landscape上（Enamine10k数据集）性能会下降，在简单的landscape上会有更好的性能。

Bagging V.S. DKL：

DKL相当于在Bagging中NN的基础上增加了GP，猜测其对于复杂的landscape有更好的学习性能。对于简单的landscape反而会增加feature map的复杂程度。

EI和Greedy在Enamine10k上的性能要明显优于Random。

接下来可能使用的算法：

1. 在数据集上动手脚，因为高表达量的数据较为稀疏，把低表达量的数据加一个低置信度然后视作高表达量用于训练，提高高表达空间预测的泛化性。
2. 将DKL和bagging的训练方法结合一下，训练10个tag的DKL模型进行集成。

## 2024.10.31

把三种采样方法的图像整合在一起了：

