Machine learning wishlist

1. Optimal result is not guaranteed. How do we determine the ceiling of the reaction yield? The goal is to ensure that the yield reaches target goal without plateauing at an undesired level.最优结果：如果确定反应产率到天花板了？这样我们就不用再进行更多的反应了。

On a related note: How many iterations do we need to run to get conclusive results?

要跑多少个反应才能确定反应最终产率呢？

1. A good method to select next experiments. (Yang is working on this actively) For example, the algorithm gives many experiments that are closely related and would give similar results. We want to maximize the diversity of the subsequent experiments and don’t want to run redundant experiments. 基于欧式距离排序，得到最不相似的条件和最相似的条件

On a related note: What is the proper number of experiments to run in every iteration? We want to run the fewest number of experiments due to cost + time but obtain good results. 每次跑多少反应呢？

We have come up with two solutions: One is to calculate the distance between the top 1% of predicted reactions (two at a time but it will cycle through all possibilities). **This may not incorporate the previous experimental conditions.** The goal is to ensure that the experiments we select are diverse and as far apart from each other as possible. As mentioned earlier, we need to come up with a way to also consider conditions that are as distant from the training set as possible.

The second solution is to calculate the distance between the top 1% predicted reactions and the previous training set data. The most distant conditions will be selected.

1. Minimum number of experiments to make first generation training set? This will presumably be related number of our features. Our goal is to have a training set that is comprehensive but remains workable without consuming too much time or materials.

We want to have the initial training set be diverse and not randomly assigned. How do ensure that our training set is as diverse as possible? We want to screen as much chemical space as possible. We want a method to help us select the initial experimental conditions rather than relying on random distribution.第一次跑多少个实验合适？

1. A smarter algorithm to suggest expansion of feature range. We want indication of when and how we should expand the range of continuous features. 什么时候扩充反应条件范围？
2. How do we determine that the model is learning? We want a measure of confidence of the output (predictions) that encapsulates the uncertainty associated with the output. What data should we input the model to get a measure of uncertainty? Can we use the measure of uncertainty to determine that the model is learning? We have defined learning as the decreasing of RMSE (for now) over iterations. 如何确定模型在学习？
3. Can we integrate high throughput experimentation (HTE) with machine learning techniques? Could we develop a standardized procedure that incorporates aspects of discovery (HTE) with optimization using machine learning. This could be achieved using the robot in the HTE centre to perform large numbers of experiments (96 experiments in a training set) that would be tedious or challenging for a human to set up accurately. A larger initial training set could be achieved using this method. 是否可以结合THE？
4. We would like to incorporate discrete variables in our machine learning algorithms. Discrete variables are very important in early stage reaction development. Can we invite our stats collaborators to develop a method of description for discrete variables? 离散变量的融入
5. Can we merge HTE, machine learning and discrete variables to give a comprehensive experiment design. The dream would be to have a robust process which can follow a project from the initial discovery phase along with allowing for optimization of discrete and continuous variables. 机器学习+THE的实验设计