Statistical Inference and Adaptive Design

Why we need?

The goal of material design is to minimize or maximize some desired properties of a material (y), through modifying features that control the chemistry, structure, or bonding (x). In the optimization process of a material, we firstly predict y and then select an x at the measured y. The (x,y) values is added to the known data. The most important difficulty here is the measurement of y, as synthesis and characterization of new materials is expensive and time consuming. To solve this problem, we have a statistical inference and adaptive design method as a surrogate to reduce the number of new materials we need to measure. The models used in the adaptive design are called selectors, which are EGO (efficient global optimization), and KG (knowledge gradient).

How it works?

Diagram

Description automatically generated

Fig. 1. The approach towards accelerated discovery

Step 1: build domain knowledge physics models by collecting known data

Step 2: constrain a statistical inference model trained on available data and predict a property with associated uncertainties

(The best score predicted by step 2 cannot be directly used in guiding the following calculations or experiments, as best score doesn’t mean an optimal option, considering the uncertainties)

Step 3: key step. Selectors are used as significant potential for designing new material with desired properties.

Step 4: the selected new materials are calculated or measured to get new results. If material with target property is proven in this step, the design ends. Otherwise, move to step 5.

Step 5: new results is added to known data and go to step 1, refreshing the domain knowledge physics model. The adaptive design is repeated.

Adaptive experimental design (selectors)

Adaptive experimental design is step 3 in Fig. 1. The uncertainties from the inference model in step 2 are used in selectors to balance the trade-off between exploiting the model predictions (current) and exploring the search space to find out new material for calculations and measurements (future).

The selectors we use are based on expected improvement or efficient global optimization (EGO). In EGO it’s assumed that a probability function P(y|x) of possible y values is given for each value of x. It suggests measuring the x that maximize the expected improvement in y over the current best values. Assuming µ\* is the current best measured values, the expected improvement achieved by measuring x’ is:

E (I) = E [ max (y, µ\*) - µ\*] = E [max (y - µ\*, 0) ] =

Here we use Gaussian distribution to describe P(y|x’), using mean µ and covariance σ2, then the expected improvement can be written as:

E(I) = σ [φ (z) + z Φ(z)]

z = (µ - µ\*) / σ

φ (z) is the standard normal density, and Φ(z) is the cumulative distribution function.

If σ is small, E(I) ~ µ - µ\*. To maximize E(I), we need the largest µ (exploitation).

If σ is large, E(I) ~ σ. Then we need the largest σ (exploration).

Another selector related to EGO is KG (knowledge gradient). It uses µ\* as the best of the predicted compounds.