**MLKR**

Supervised machine learning techniques can be used to predict a function (y = y(**x**)) using a set of training examples ((**x**1, y1), (**x**2, y2), … , (**x**N\_train, yN\_train)). In this study we predict the TOF of a catalytic site on an amorphous support as a function of the local site geometry (**x**). At present, the geometry of the site is defined as an abstract quantity, which is related to the local support structure around the site. It will be defined more specifically in section XX.

We assume the rate law can be written in the form of a power law with a single activation barrier, as described in section XX (Eq. XX). In this context, we can use machine learning to learn the activation barrier () of a site as a function of its local geometry (**x**). We use kernel regression for this purpose. In kernel regression, the quantity of interest () is estimated as a weighted average of the training data, Eq. XX.

where is the prediction, represents the values in the training set, and are the weights. The weights are given by

where are kernels, which are functions of the distance between data points in the **x** space. We adopt the Metric Learning for Kernel Regression (MLKR) framework of Weinberger and Tesauro.1They use a Gaussian kernel and the Mahalanobis metric to define distance between data points. Gaussian kernels are defined as follows.

where d is the Mahalanobis distance (generalized Euclidean distance) between data points i and j in the **x** space, and is given as follows.

where M is a dim() x dim() dimensional matrix. The matrix M is constrained to be positive definite and symmetric. The objective of MLKR is to find the M which best fits the training data, eq XX. The leave-one-out loss function, Eq. XX, is minimized for this purpose.

Where N\_train is the number of training points. Each data point is a weighted average of all the other data points in the training set excluding itself. We use MLKR as implemented in the metric-learn python library. The library uses the conjugate gradient method with analytical derivatives of the leave-one-out loss to minimize Eq. XX.

**Add line to yellowed portion in biased sampling**

**Sampling Error with Unbiased Sampling**

The k-weighted activation barrier is given by

If the sites are sampled using the unbiased density (, the sampled k-weighted activation barrier is given by

The variance of the sampled k-weighted mean can be computed as follows:

The variance of the random variables is the same, hence the expression simplifies to

**Sampling Error with Biased Sampling**

Rearranging equation XX, the average can be recast with a biased probability density as follows

If sites are sampled according to a biased density, equation XX, the k-weighted activation barrier can be estimated as follows

Equation XX, can be simplified as follows

The variance of all random variables is the same, hence the expression simplifies to

This leads to the central limit theorem result

**Write results**

**1**

**Figure 5a** shows the parity plot of the activation barriers predicted by the trained ML model compared with their true values. The model is able to predict the activation barriers with extremely good accuracy for this initial pool of sites, with all of the errors being <<2.5 kJ/mol. Following this, the trained model was used to predict the activation barriers of all sites (~20K). **Figure 5b** shows the distribution of model predicted errors on all sites. The errors are approximately distributed as a Gaussian with most of the errors within kJ/mol.



**Figure 5**: Parity plot of model predicted activation barriers (trained on 50 randomly sampled sites) vs true activation barriers (left). Distribution of model predicted activation barrier errors (Ea,true – Ea,model) on all sites (right).

**2**

Sites were also randomly sampled to compare with the Importance Learning algorithm. **Figure 7** compares the convergence of the k-weighted activation barrier using Importance Sampling and Random Sampling. Error bars on the averages were computed using methods described in section XX SI.



**Figure 7**: Convergence of the k-weighted activation barrier using Importance Sampling (left). Convergence of the k-weighted activation barrier using Random Sampling (right).

It should be noted that the 100 importance sampled sites also include an initial training pool of 50 sites. Hence, 150 randomly sampled sites are compared with 100 importance sampled sites. Sampling errors on the randomly sampled averages are significantly larger than the importance sampled errors. The importance sampled errors decay to within kJ/mol after 30 iterations. However, for the random sampling case the errors are kJ/mol after 150 iterations.

**Write in unbiased and biased errors**

**Make proper mlkr section**

**Write yellow highlighted section**