Supplementary Material: Tree-based Quantile Active Learning for automated discovery of MOFs

1 QRT-AL - Algorithm

In this section, we provide the pseudo-code of the proposed algorithm. The number of samples to be labeled from each leaf k (representing a region $\mathcal{R}_k$) n_k^*, are distributed into the different leaves as:

$$n_k^* = n_{\text{act}} \frac{\sqrt{\pi_k \hat{\sigma}_k^2 \gamma_k}}{\sum_{\ell=1}^K \sqrt{\pi_\ell \hat{\sigma}_\ell^2 \gamma_k}};$$

where n_{act} are the total number of samples to be selected by QRT-AL, $\hat{\sigma}_k^2$ denotes the variance computed on the true labels in leaf k, π_k the proportion of unlabeled samples in leaf k, defined formally as follows: for $1 \leq k \leq K$,

$$\hat{\sigma}_k^2 = \frac{\sum_{i \in I_{\text{init}} : \mathbf{x}_i \in \mathcal{R}_k} (\hat{Y}_i^{I_{\text{init}}} - Y_i)^2}{|i \in I_{\text{init}} : \mathbf{x}_i \in \mathcal{R}_k| - 1}, \tag{1}$$

$$\pi_k = \frac{|i \notin I_{\text{init}} : \mathbf{x}_i \in \mathcal{R}_k|}{N}, \tag{2}$$

and γ_k specifies the quantile interval of interest: for each leaf $1 \leq k \leq K$,

$$\gamma_k = \frac{\sum_{q=1}^Q w^q n_k^q}{\sum_{q=1}^Q n_k^q}$$

The code has been written in python and will be made publicly available upon acceptance.

Algorithm 1 Quantile Regression Tree-based Active Learning (QRT-AL)

```
Input: Labeled set $(\mathbf{x}_i, y_i)_{i \in I_{\text{init}}}$ and unlabeled set $(\mathbf{x}_i)_{i \notin I_{\text{init}}}$; $n_{\text{act}}$ the maximum number of new samples to be labeled; quantile interval of interest, Q
Construct a standard regression tree with $K$ leaves using $(\mathbf{x}_i, y_i)_{i \in I_{\text{init}}}$
for $k = 1, \ldots, K$ do
    Compute $\pi_k$, $\hat{\sigma}_k^2$ and $\gamma_k$
    Calculate the number of samples $n_k^*$ to be labeled from leaf $k$
    Detect $I_{\text{act}}^k$, the set of $n_k^*$ observations from leaf $k$
end for
Output: The set $\cup_{k=1}^{K} (\mathbf{x}_i)_{i \in I_{\text{act}}^k}$ of observations to be labeled
```

2 Additional results

2.1 Different quantiles

Below are presented results for a different set of quantile intervals for the hMOF database. The high quantile intervals are of interest for efficient gas capture, so the quantiles chosen are [0,0.5), [0.5,0.7) and [0.7,1] with weights 0.05, 0.25 and 0.70 respectively. We see that even in this case, the conclusions from the proof of concept provided in the main paper hold.

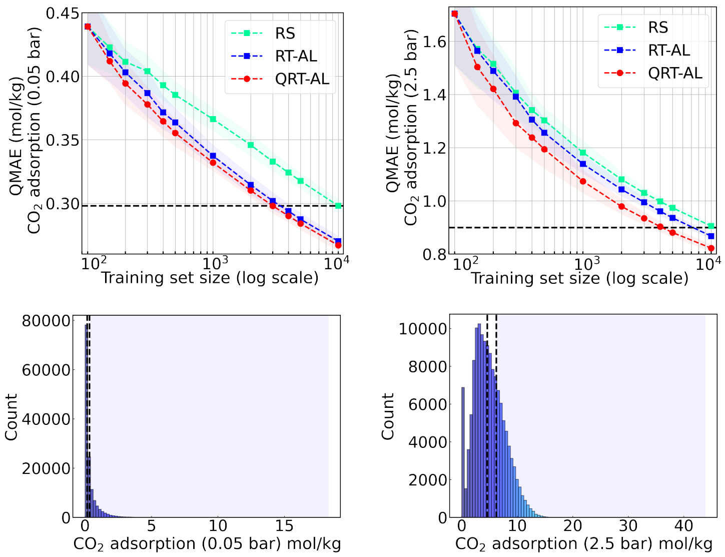

Figure 1: Quantile MAE (averaged over 100 runs) computed on the test set predicting band gaps for MOFs in the QMOF database, and CO_2 adsorption at 0.05 and 2.5 bar pressures for MOFs in the hMOF database, while sampling the training set with RS, RT-AL and QRT-AL. Distributions of the respective target properties has been shown as histograms below each case. The vertical dashed lines depict the quantile intervals chosen and the quantile interval of interest has been shaded.

2.2 Batch vs Sequential AL

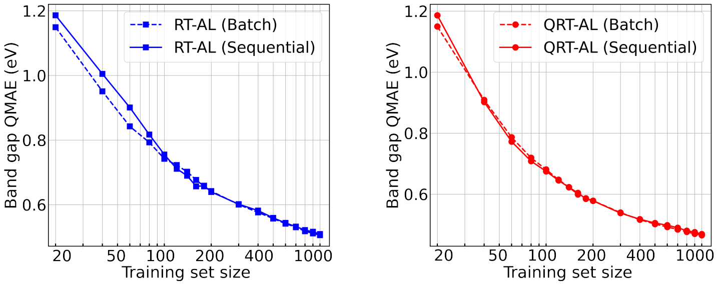

Figure 2: Quantile MAE (averaged over 100 runs) computed on the test set predicting band gaps for MOFs in the QMOF database while sampling the training set sequentially vs in batches with RT-AL and QRT-AL.

2.3 Parity plots for CO_2 adsorption

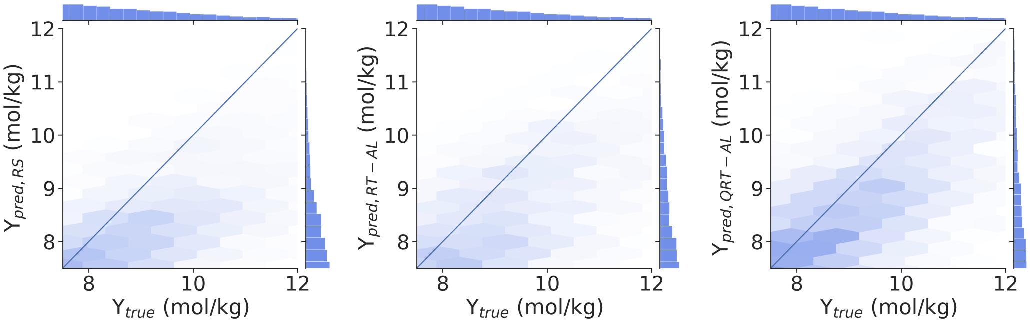

Figure 3: Parity plots for CO_2 adsorption at 2.5 bar pressure, with 1000 MOFs in training set, shown in quantile of interest.