# Supporting Material for A diversity maximizing active learning strategy for graph neural network models of chemical properties

Bowen Li and Srinivas Rangarajan*

*Department of Chemical and Biomolecular Engineering, Lehigh University, 111 Research Dr, Bethlehem, 18015, PA, USA*

E-mail: SR:srr516@lehigh.edu

## S1:Representation of the molecules in different areas of the t-SNE plot

As shown in Figure S1, we colored different parts of the t-SNE plots for the areas in the edge mostly and labeled the points for the molecules they represent. Since t-SNE plots visualize the points with their relative distance remains unchanged, it could be anticipated that the clustered molecules are similar while the molecules in the far distance are different in structures and atoms, if the latent features capture the main features of different types of molecules. Such trend is shown in Figure S1 that in each colored area, the molecules share some features. The same type of substructures, (for example, branches in the blue part and rings in the red part) or the types of the main atoms presented can be observed. On the other hand, for the colored areas that are far in the distance in the figure, the molecules they

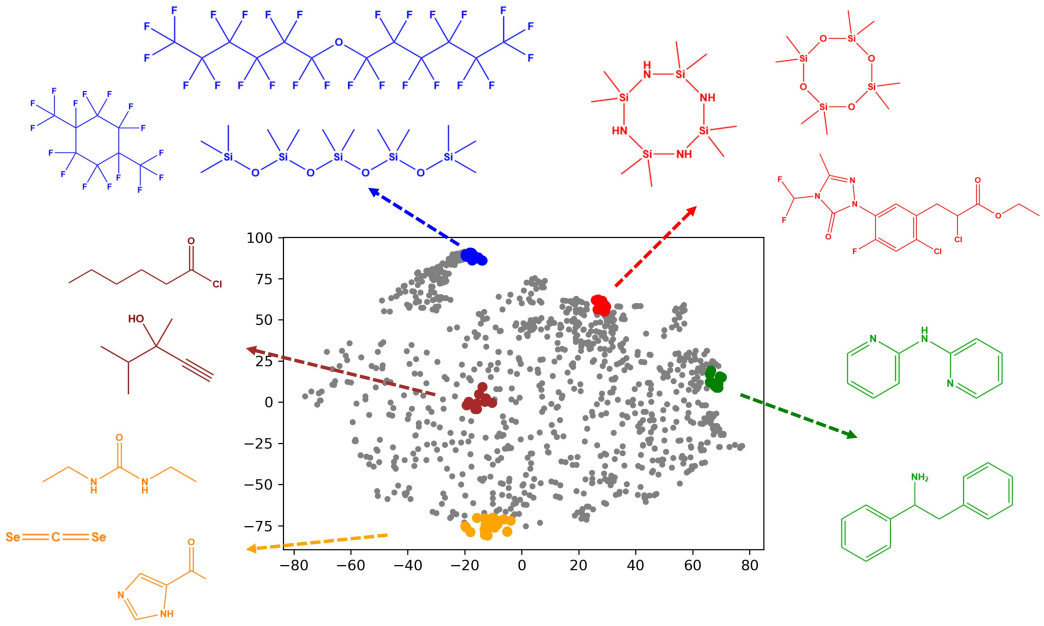


Figure S1: Visualization of the molecules in various parts of the t-SNE plot, the features for the t-SNE transformation are taken from the latent features for each molecule.

represent are notably different, the molecules on the edge are complicated in structures and large in size, while the molecules in the center are much simpler. Therefore, we can conclude that the clustered behavior of the molecules from the t-SNE visualization suggests that the numerical latent features from the embedded deep learning model succeed in capturing the features of the molecules, and those features are retained during the dimension reduction from t-SNE transformation.

## S2: T-SNE visualization for the batches of the added points in each iteration

In Figure S2 we color coded the points based on the iteration where they were added. Each color represents a batch of the added points in an iteration. Basically, for the random methods in the BP dataset and the two methods in the QM7 dataset, we can observe that the distribution of each added batch covers the dataset uniformly. While for the max-min

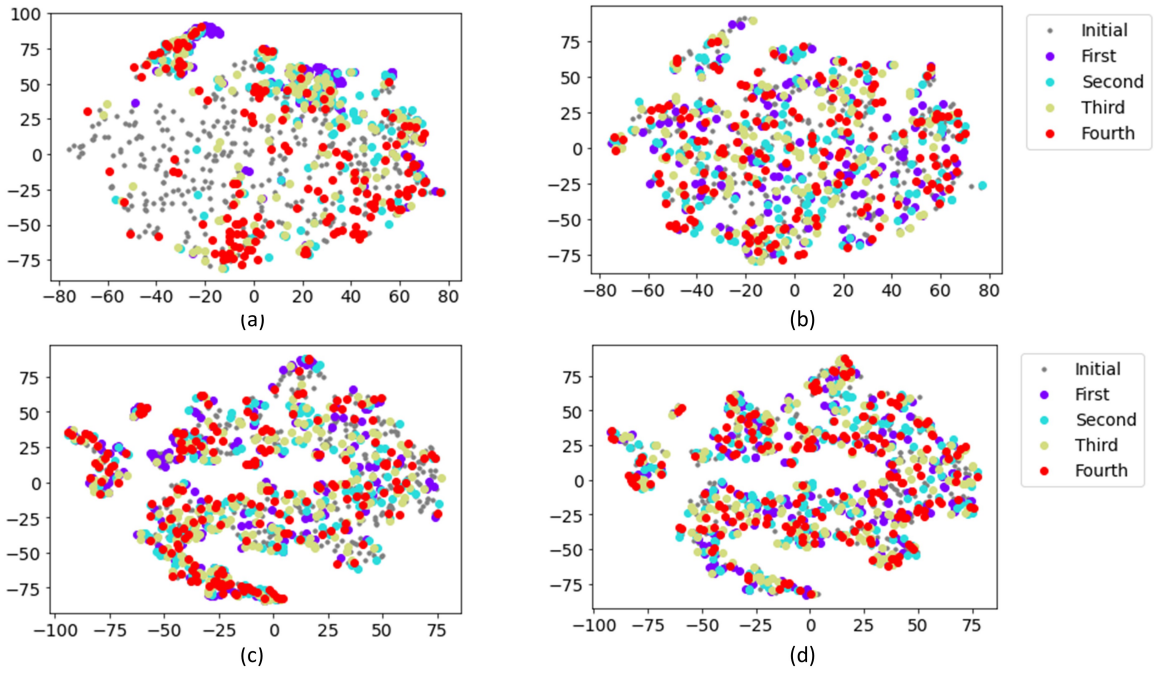


Figure S2: The t-SNE visualization for the batches of the added points in each iteration during the active learning process. Each color represents one batch of the added points in one iteration, while the grey color shows the randomly selected initial training set. The distribution of the points selected by the max-min method or randomly are compared for BP and QM7 datasets. Specifically, (a) BP dataset and max-min method, (b) BP dataset and random method, (c) QM7 dataset and max-min method, (d) QM7 dataset and random method.

method in the BP dataset, the batch mostly covers the edge of the dataset, and has the most points to represent the cluster that far from the center. During the iteration, the first batch covers the farthest points from the center, while the last batch starts to shift into the center. This indicates that for the BP dataset, the max-min method succeeds in capturing more information during the training process by selecting points in the clusters that lack enough points to represent; on the other hand, both methods sample the QM7 dataset more or less uniformly.

Note that in Figure S1, we showed that the points on the edges represent the molecules that are complicated in the structure, or the size (large branches and rings) compared to the molecules in the center, which potentially requires many points in that area to exist in the training set for a model to fully captures their property information. Thus, it is anticipated

that these points that are far from the center are prone to error if the model is designed only for most of the molecules present in the dataset, as come from the uniform selection. In Figure S2 we observed that the points selected from the max-min methods cover these areas in the edge during the first few iterations (iteration 1 and 2), while the random methods tend to pick molecules uniformly, the relative better performance for the model with training points from the max-min method could be expected.

## S3: Histograms showing the distribution of the two axes for the t-SNE plot for each iteration

In Figure S3 and Figure S4, we further plotted the histograms showing how all the added points distributed in the two axes from the t-SNE visualization during the training set selection process, extending from the 4 iterations in Figure S2. Each color represents one iteration, while the brown/grey colors showing together in one plot among the iterations represent the distribution of the training/remaining set at the corresponding iteration. From the figures, we can observe that for both the random methods in the BP dataset and QM7 dataset, the distribution of the added points all approximately centered in the plots and resembles a uniform distribution. Meanwhile, the distribution of the training/remaining set is almost the same and mirrors each other in those iterations, which suggests the uniform sampling from the random methods. On the other hand, for the max-min method in the BP dataset, the distribution either centered at the edge (i.e. iteration 1  3, BP max-min axis 2), or have a few peaks that dominated the distribution (i.e. iteration 1 & 4, BP max-min axis 1), which indicates that the max-min method brings in the types of molecules that not represented sufficiently from the random methods, thus makes the difference in the performance of the model. The difference is also reflected in the completely different training/remaining set distribution in both axes. However, for the QM7 dataset, as shown in Figure S4, the max-min method overall doesn't make a significant difference from the uniform

distribution; on axis 1, the max-min method only centered slightly to the left compared to the corresponding iterations in the random plots. While on axis 2, even though there is a peak in the edge of the distribution in most of the iterations for the max-min method, the other part in the distribution remains similar, thus the overall training set in Figure S4 is distributed similarly compared to the remaining set.

Also note that during the process of the iterations, the distribution of the max-min method for the BP dataset changed, the types of the molecules added in the early iterations are significantly different from those in the later iterations. On axis 1, the peaks shifted from the right to the left, while on axis 2 the peaks shifted from the right to a few peaks uniformly distributed in the plot. However, such a trend is not observed in the QM7 max-min method; the location of the peaks remains the same and the shape of the distribution doesn't vary, wherein it could imply that the types of the added points remain similar in the overall iterations.

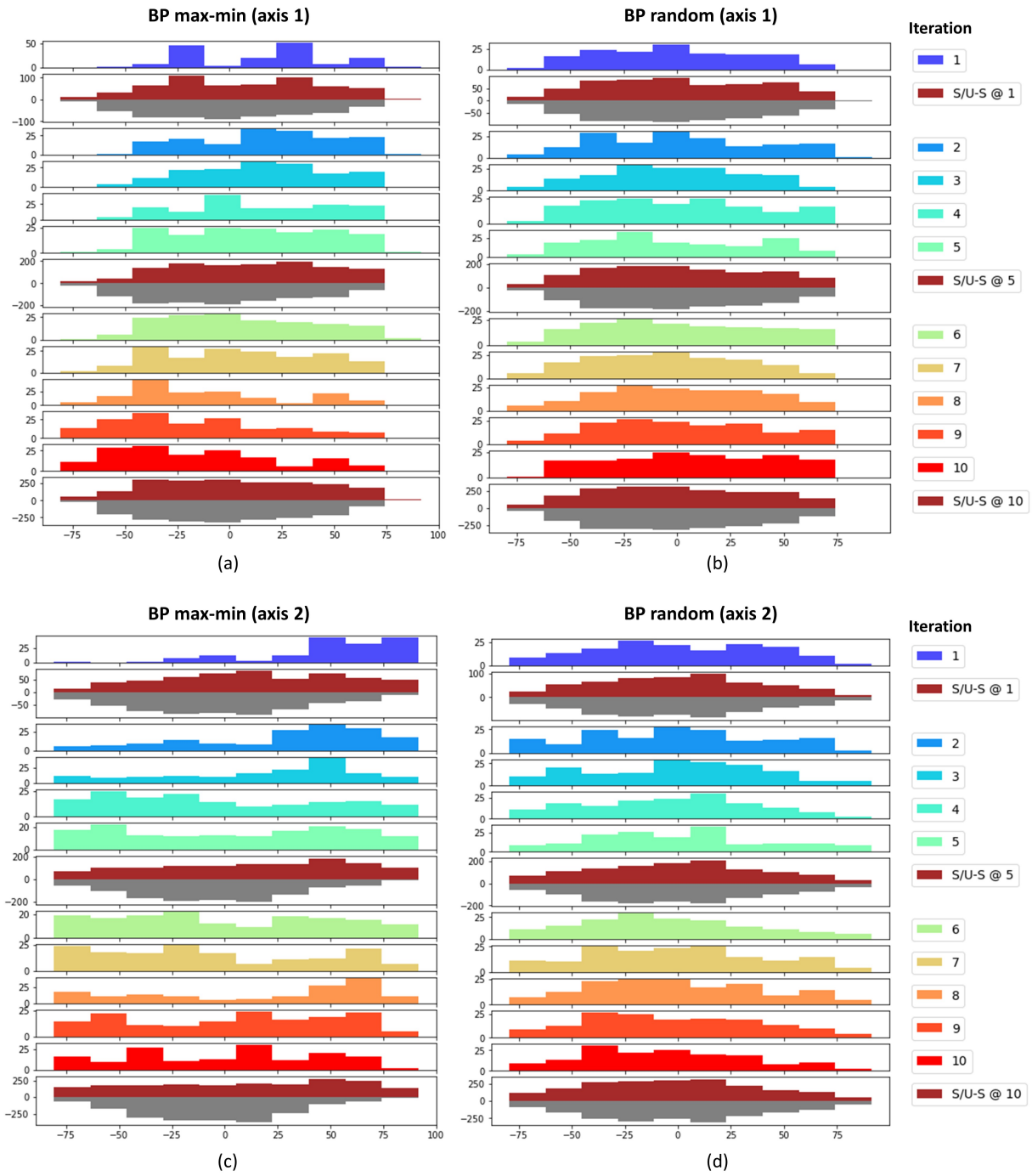


Figure S3: Histograms show how the added points are distributed in the two axes from t-SNE visualization in the BP dataset. Each color represents one batch of the added points, except the brown/grey colors that show together which represents the comparison between the distribution of the training set (S, brown) and the remaining set (U-S, grey) after a certain iteration, the axis of the histogram for the remaining set is normalized in respect to the training set and multiplied with negative values to direct downwards. The histograms of the added points via the max-min method and random method are compared for the BP dataset. Specifically, axis 1 of the t-SNE plot for (a) max-min method, (b) random method, and axis 2 of t-SNE plot for (c) max-min method, (d) random method.

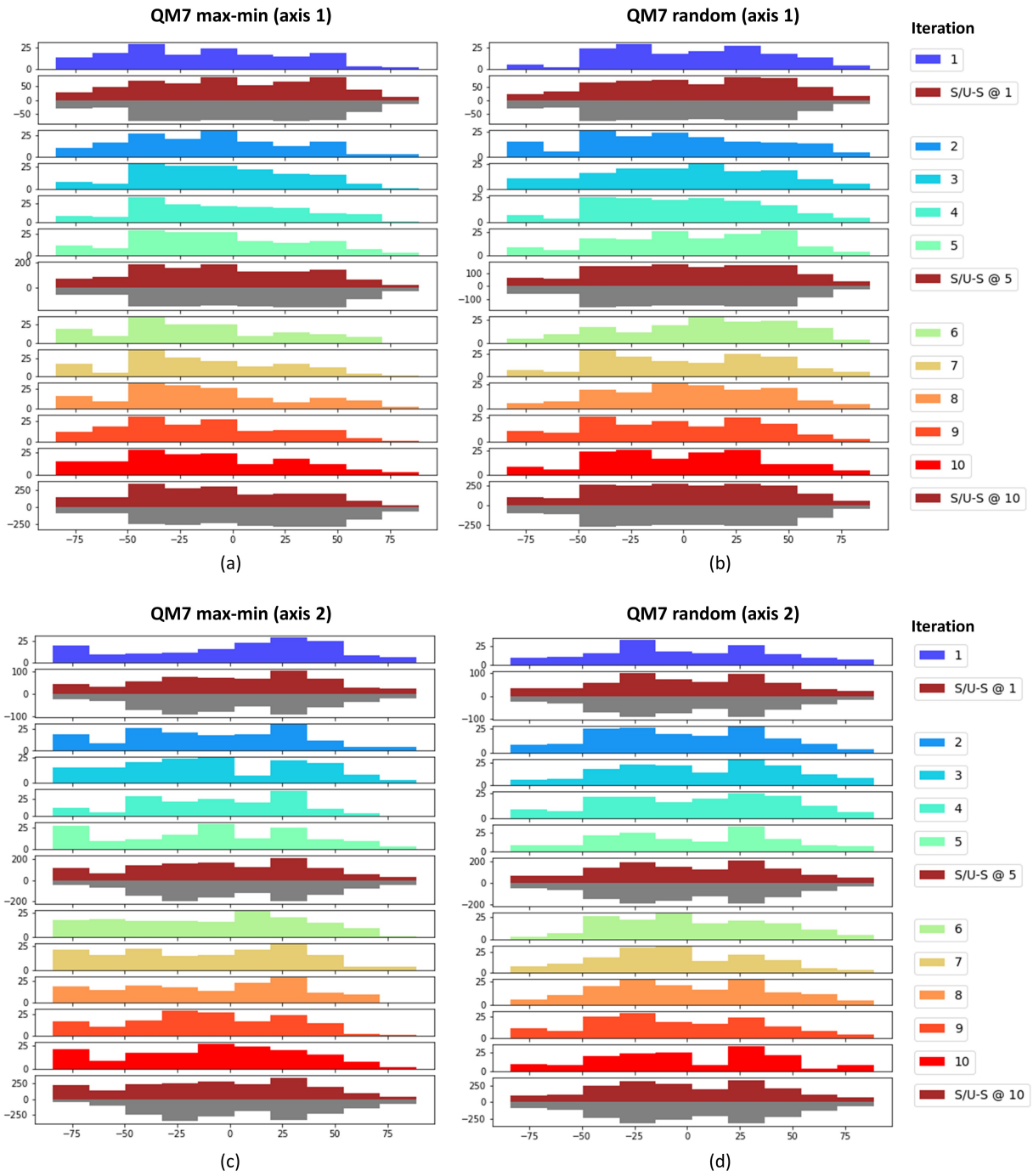


Figure S4: Histograms show how the added points are distributed in the two axes from t-SNE visualization in the BP dataset. Each color represents one batch of the added points, except the brown/grey colors that show together which represents the comparison between the distribution of the training set (S, brown) and the remaining set (U-S, grey) after a certain iteration, the axis of the histogram for the remaining set is normalized in respect to the training set and multiplied with negative values to direct downwards. The histograms of the added points via max-min method and random method are compared for the QM7 dataset. axis 1 of t-SNE plot for (a) max-min method, (b) random method, and axis 2 of t-SNE plot for (c) max-min method, (d) random method.