Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Supporting Material for Designing compact training sets for data-driven molecular property prediction through optimal exploitation and exploration

Bowen Li and Srinivas Rangarajan*

Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem

E-mail: SR:srr516@lehigh.edu

S1: Generalized pathway fingerprints for molecular representation

Figure 1 shows the illustration of the modified pathway fingerprints. The fingerprints consist of two parts, linear substructures (FP) enumerates paths of length one to seven atoms emanating from each atom of the molecule and correction terms (Correction) introduce additional ring information into the fingerprints; ring information is lumped into its presence of aromaticity and its size, while fused ring information is lumped into the size of two adherent rings. The lumped ring information is appended to the feature vector of linear substructures.

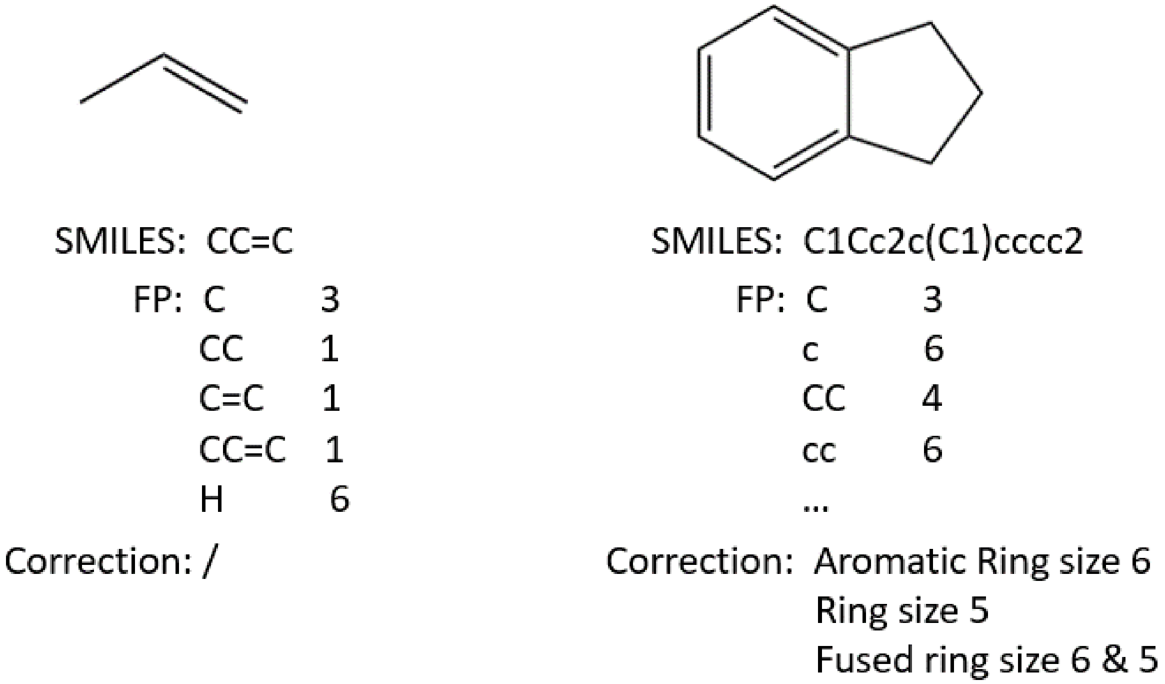


Figure 1: Illustrations of linear substructures and correction terms of modified pathway fingerprints on propane and a fused ring molecule. The notation includes (1) SMILES scheme, (2) linear substructures (FP), and (3) correction terms.

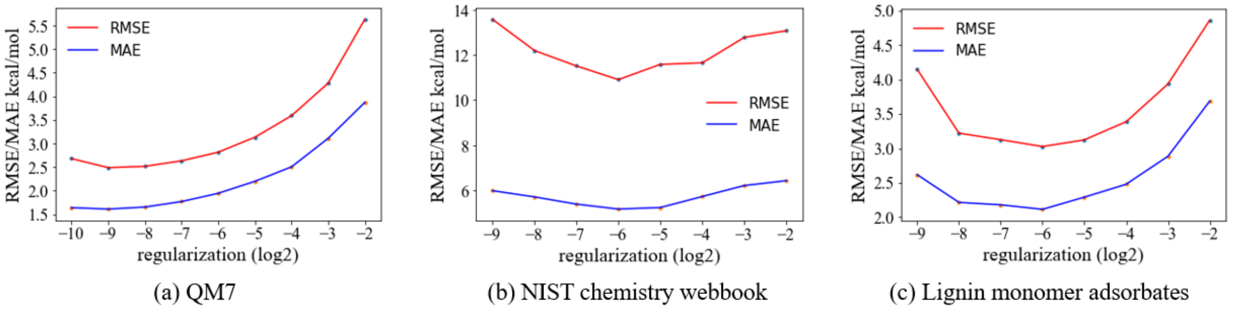
S2: Regularization value selection for the proposed algorithm

When the initial training set is built or the molecule is updated into the training set, LASSO regression is performed to build the sparse model. The regularization value λ determines the number of the selected features and its optimal value is chosen from the range $\{(2^i)\,|i = i_{upper} : i_{low} : -1\}$, where the upper bond i_{upper} is a predefined large value that allows only a few features to be selected when the training set size is small, and the lower bond i_{low} is the ten-fold cross validation optimal regularization value based on the initial training set. At each iteration, the regularization value is keeping decreasing from $2^{i_{upper}}$ to $2^{i_{low}}$ by the 2-base logarithmic step and stops at the value where the number of the feature selected is larger than the number of current training molecules or the value reaches the lower bond $2^{i_{low}}$. In the former case, we multiply the value that stops at by 2 and choose the result value for LASSO to allow maximum features to be selected (to prevent potential overfitting if the final feature number is only 2 less than the training number, we further mulitply the regularization value by 2.), and in the latter case we choose the lower bound $2^{i_{low}}$. The ten-fold cross validation of the initial training set is performed with scikit-learn[1] LASSOCV python module.

Sometimes the number of features determined by the regularization value is slightly less than the molecule number in the early iteration steps, which causes severe overfitting and dramatic increase of the RMSE. To prevent such scenario an additional restriction relying on prediction variance function ($x_i^T \left(X^T X\right)^{-1} x_i$) is added that in two consecutive iteration steps, if the selected regularization value leads to a huge increase of the average prediction variance function in the remaining set, we multiply the regularization value by 2 once to select fewer features. In this work, the multiplication is performed if the average prediction variance function is increased by more than half of the value in the previous iteration. On the other hand, while more features can be included when the training set size gets large, in several cases the restriction will prevent regularization value getting smaller in later stage even though the average prediction variance function is very small. Thus the restriction is removed to allow more features to be selected after the largest prediction variance function value reaching a small value which indicates a stable model. In this work, the restriction is removed when the largest prediction variance function value is smaller than 1.5.

For some datasets, the lower bound $2^{i_{low}}$ of the regularization range determined by ten-fold cross validation for Max-Min selected initial training set is unreasonably large, which makes feature selected by LASSO too sparse to build an accurate model in the whole iteration. When such a scenario happens, the regularization value decreases to the lower bound within a few iterations, and remains the same for the remaining iterations. One possible explanation is that due to the diversity of the initial training set, during ten-fold cross validation only features common to most molecules are selected by LASSO, while other specific features contained by a few molecules are likely to be pushed to zero, which leads to a large regularization value being favored. In this work, the scenario happens when applying the algorithm on the surface intermediates set started with Max-Min method for initial training set building, the solution so far is to simply choose the lower bound $2^{i_{low}}$ determined by the randomly selected initial training set instead for iterations with the initial training set selected by Max-Min method. In principle, we can do another round of cross validation during the middle stage if the too sparse model scenario happens.

S3: Five-fold Cross validation results of the datasets



(a) QM7

(b) NIST chemistry webbook

(c) Lignin monomer adsorbates

Figure 2: Comparison of the average RMSE/MAE with different regularization values for five models during five-fold cross validation for (a) 6.5k gaseous molecules of QM7[2,3] dataset with their heats of atomization, (b) 598 hydrocarbons of NIST chemistry webbook[4,5] with their heats of formation, (c) 591 surface intermediates set[6] with their heat of formation.

We approach the baseline error based on five-fold cross-validation. Cross-validation is a widely used method for estimating prediction error, especially when data are scarce.[7] When applying five-fold cross-validation, we randomly split the data into five folds that are similar in size. One of the folds is used as validation set for prediction and the others are used as training sets to fit the model. The process is repeated for five times until each fold is used as validation set for exactly once, and the prediction performance is averaged. The (hyper)parameters that lead to the minimum average error is selected and the corresponding error is reported as the estimated prediction error.

Figure 2 shows the average RMSE/MAE corresponding to different regularization values for five models during five-fold cross validation for three datasets. The lowest RMSE value in each dataset is selected as the optimal value, which is used as the baseline value for the learning rate compare. Ordinary least square (OLS) regression is performed on the LASSO selected features to obtain the prediction error of each fold. For the three datasets, the selected hyperparameter regularization values and the optimal RMSE/MAE are (a) 2^{-9}, 2.5/1.6 kcal/mol for QM7 dataset, (b) 2^{-6}, 11.0/5.2 kcal/mol for NIST dataset, and (c) 2^{-6}, 3.0/2.1 kcal/mol for lignin monomer adsorbates respectively.

S4: 10 runs of the variance sampling strategy with a randomly selected initial training set for the kernel model

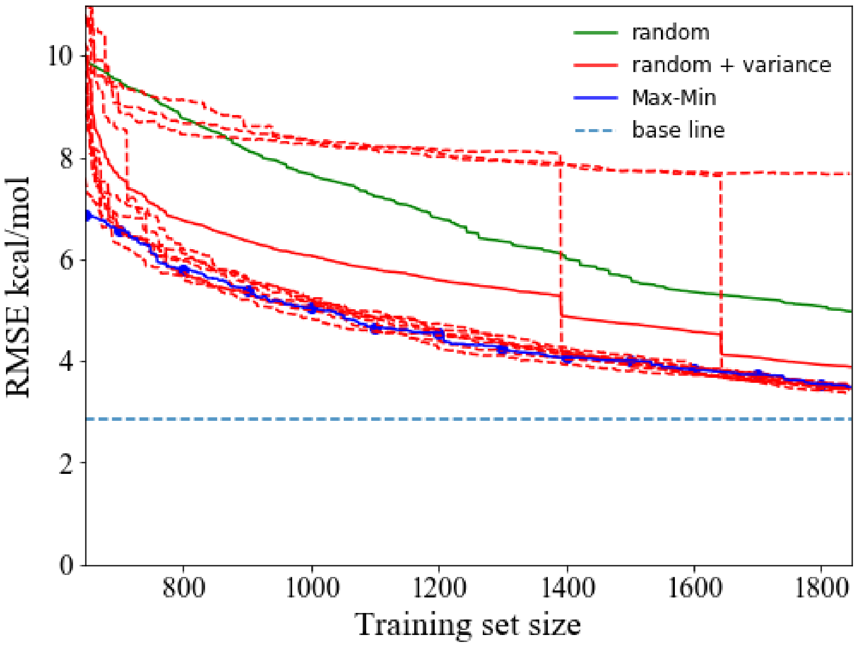
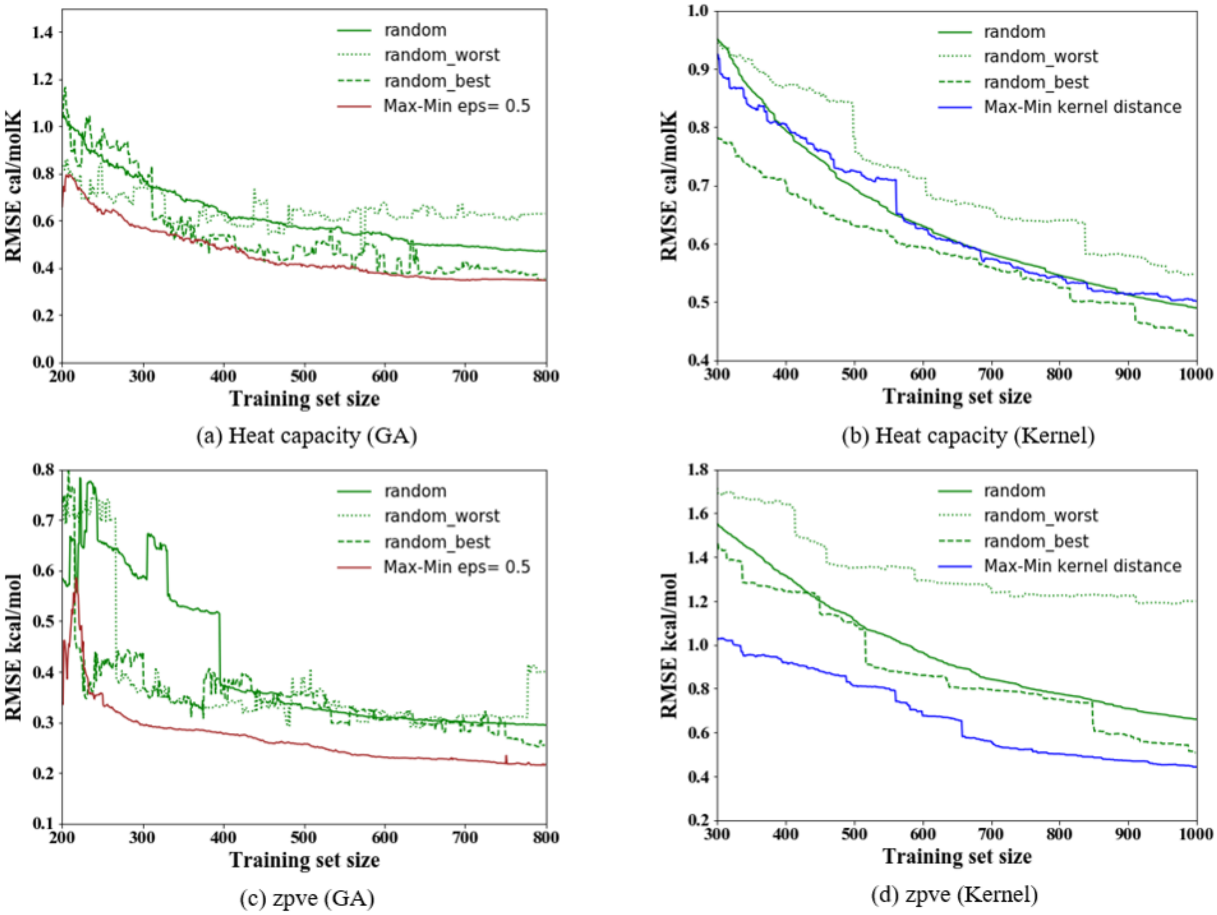


Figure 3: Comparison of learning rates for 10 executions of the variance sampling strategy started with different initial training sets, each dashed curve represents a run. The solid curves represent the learning rates of (1) random sampling, (2) pure Max-Min method, and (3) average variance sampling. The performance of random sampling and variance sampling is averaged from 10 runs.

Figure 3 shows the learning rates of the variance sampling strategy that starts with different randomly selected initial training sets. 10 runs based on 10 different initial training set are performed. Each dashed curve represents a run. We can observe that there are three runs that their errors drop perpendicularly after a molecule is added at a certain step, while for the other executions within a few iterations the curves converge to the kernel distance diversity maximization curve. The result indicates that the randomly selected initial training set may fail to represent some parts of the molecule space due to the absence of a specific type of molecule.



(a) Heat capacity (GA)

(b) Heat capacity (Kernel)

(c) zpve (GA)

(d) zpve (Kernel)

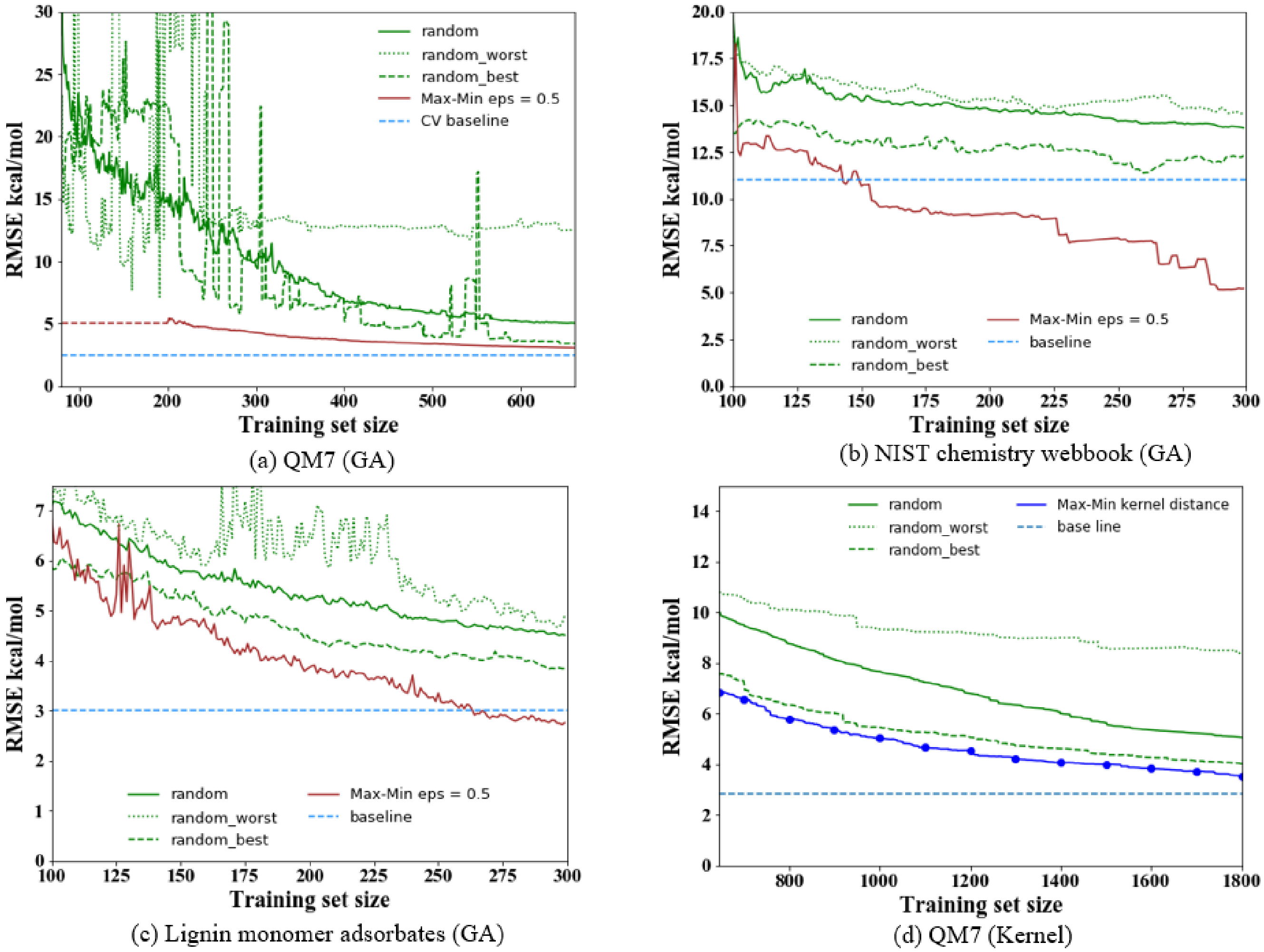
Figure 4: Comparison of the learning rates of different molecule selection strategies on the QM9 dataset for GA-based and kernel-based prediction of heat capacity and zero-point vibrational energy. The y-axis measures the root mean square error of the prediction on the remaining set during the updating process with (1) random sampling (green) and (2) epsilon greedy method with ϵ set to 0.5 (brown) for GA-based model or the pure Max-Min method based on kernel distance for kernel-based model (blue). The random worst curve (dotted) represents one random iteration with the highest final model RMSE, and the random best curve (dashed) represents the lowest. The reported performance of the random sampling and epsilon-greedy method is averaged over multiple runs.

S5: Applying the algorithm on heat capacity and zero-point vibrational energy on QM9 dataset

We extend the algorithm application from energetics of molecules to other properties. In Figure 4 we show the performance of the algorithm tested on the subset of QM9[8] dataset for predicting heat capacity and zero-point vibrational energy (zpve) compared to random sampling. QM9 dataset contains ∼134k molecules consisted of C,N,O,F,H atoms with number of non hydrogen atoms less than 9 and their various properties from DFT calculation. In this case, we select ∼3.1k molecules (with non hydrogen atoms no more than 7 and excluding fused ring molecules) as our benchmark dataset. From the figure we can observe that most of the learning rate agrees with the study of the energetics of molecules. For heat capacity and zpve with GA-based model, the epsilon-greedy method shows superior performance to the random sampling. For zpve with kernel-based model, the performance of pure exploration approach based on kernel distance is better than the random sampling as well. However, for the kernel model applied to the heat capacity, there is no clear advantage over the average performance of the random sampling. Nevertheless, the performance is more reliable compared to the worst iteration of the random sampling strategy.

S6: Comparison of the highest and the lowest RMSE of the final training model built with random sampling strategy

In Figure 5 We included the best and worst iteration of multiple runs of the random sampling strategy for different datasets considered in the manuscript. The best curve represents the case wherein the final model has the lowest RMSE; the worst curve represents the case wherein the final model has the highest RMSE. Clearly, there is a finite chance that random sampling can give comparable performance as the epsilon-greedy approach but there is an equal chance that the models may have about 10-15 kcal/mol higher errors.



(a) QM7 (GA)

(b) NIST chemistry webbook (GA)

(c) Lignin monomer adsorbates (GA)

(d) QM7 (Kernel)

Figure 5: Comparison of the learning rates of the best and worst random sampling on different datasets. The y-axis measures the root mean square error of the prediction on the remaining set during the updating process with (1) random sampling (green) and (2) epsilon greedy method with ϵ set to 0.5 (brown) for GA-based model or the pure Max-Min method based on kernel distance for kernel-based model (blue). The random worst curve (dotted) represents one random iteration with the highest final model RMSE, and the random best curve (dashed) represents the lowest. The reported performance of the random sampling and epsilon-greedy method is averaged over multiple runs.

References

(1) Pedregosa, F. et al. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(2) Blum, L. C.; Reymond, J.-L. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

(3) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(4) Buerger, P.; Akroyd, J.; Martin, J. W.; Kraft, M. *Combust. Flame* **2017**, *176*, 584–591.

(5) Linstrom, P.; Mallard, W. *NIST standard reference database* **2005**, 20899.

(6) Gu, G. H.; Vlachos, D. G. *J. Phys. Chem. C* **2016**, *120*, 19234–19241.

(7) Hastie, T.; Tibshirani, R.; Friedman, J.; Franklin, J. *The Mathematical Intelligencer* **2005**, *27*, 83–85.

(8) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. *Scientific data* **2014**, *1*, 140022.