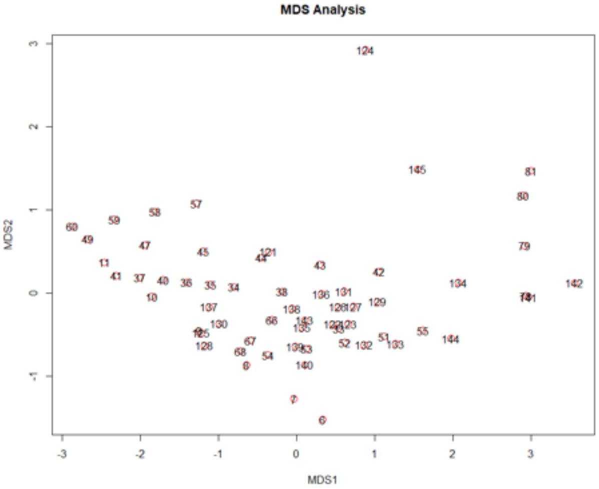
Active Learning in Materials Science with Emphasis on Adaptive Sampling Using Uncertainties for Targeted Design

Turab Lookman,[1] Prasanna V. Balachandran,[1] Dezhen Xue,[1] and Ruihao Yuan[1]

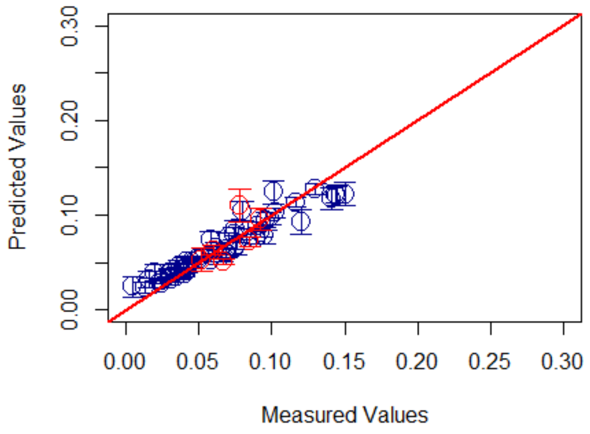
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

(Dated: 14 December 2018)

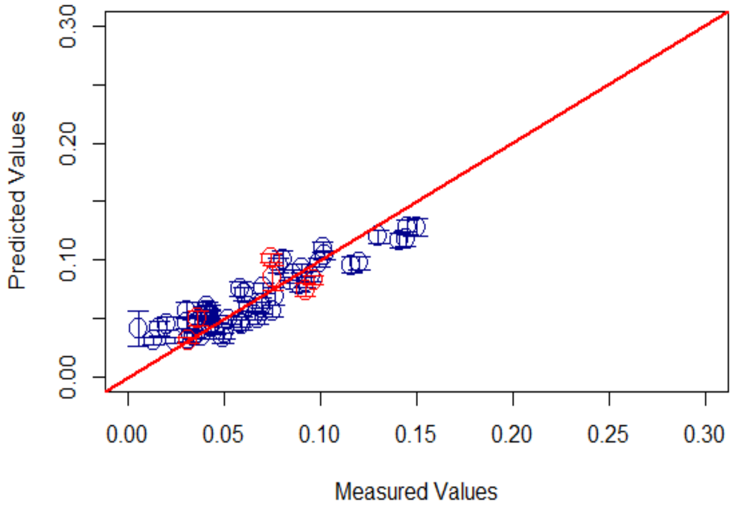
The data, which includes the electrostrain and 7 features, namely, electronegativity, polarizability, ionic displacement, ionic radius and volume, as well as features which capture the direction (increase, decrease or no change) of the dependence of the cubic to tetragonal ferroelectric transition temperature and tetragonal to orthorhombic ferroelectric transition temperature, respectively, on the doping elements, has been uploaded in a csv file.



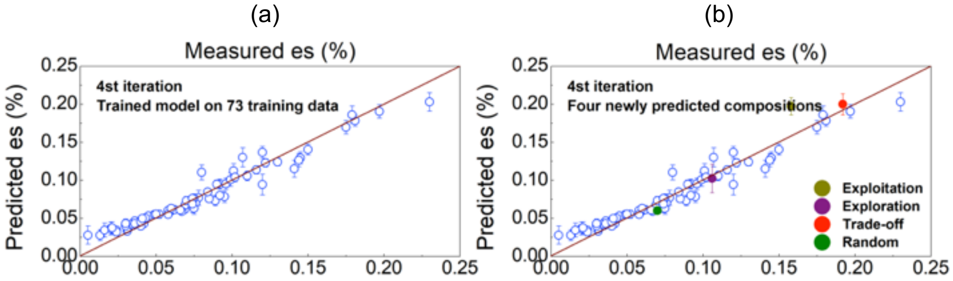
Supplementary Figure 1 | Visualization of the top two dimensions (MDS1 and MDS2) from the multidimensional scaling analysis of the training data used for the electrostrain problem.



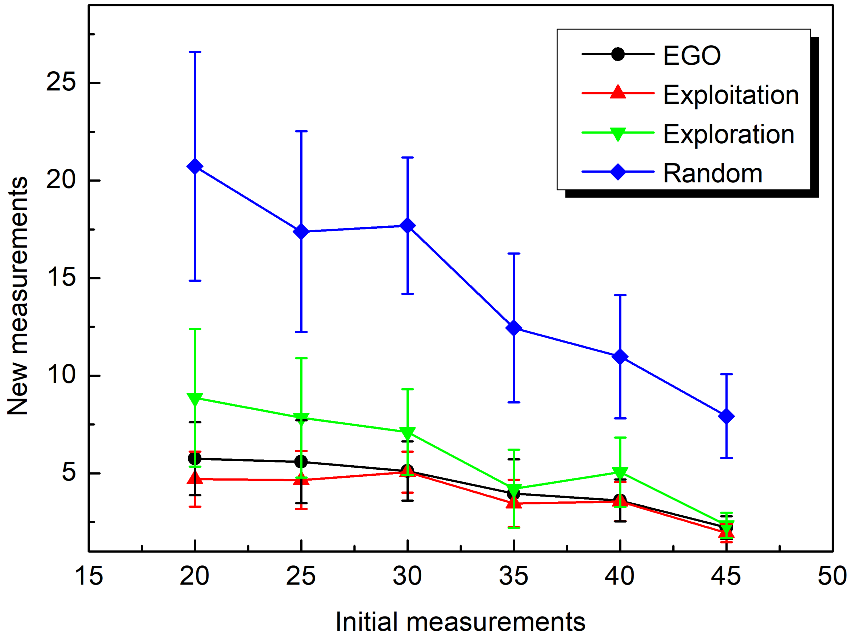
Supplementary Figure 2 | Trained SVR.RBF model for Example 1 in the manuscript, where 55 data points were chosen for training (blue circles) and 6 for testing (red circles). It can be seen that the trained model predicts the electrostrain value for the test data points with a high accuracy, indicating that our trained model is not an overfit to the data.



Supplementary Figure 3 | Trained Gradient Tree Boosting method for Example 1 in the manuscript. This is in addition to the data shown in Supplementary Figure 2 indicating that our trained model is not an overfit to the data.



Supplementary Figure 4 | (a) Performance of the trained SVR.RBF models on the 4th iteration (which now has 73 training data). es stands for Electrostrain (in %). (b) The predictions for the four new materials to be tested according to the four AL strategies are shown as colored (filled) points. The predictions are consistent with the experimental measurements (x-axis).



Supplementary Figure 5 | The dependence of the number of new measurements on the size of randomly chosen data from the original training data set. An SVR.RBF model was trained on 1000 bootstrap samples and the statistics of the number of new measurements are computed over 100 trials. AL strategies outperform random selection for Example 1 (compare to Figure 12b for a DFT data set).