Electronic Supplementary Material (ESI) for Journal of Materials Chemistry C.
This journal is © The Royal Society of Chemistry 2022

Machine learning assisted hierarchical filtering: A strategy for designing magnets with large moment and anisotropy energy

Supplementary Information

Arijit Dutta * and Prasenjit Sen

Harish-Chandra Research Institute, HBNI, Chhatnag Road, Jhunsi, Allahabad(Prayagraj), UP 211019, India.

* Current affiliation: Institut für Theoretische Physik, Goethe-Universität, 60438 Frankfurt am Main, Germany.

I. h_{form} REGRESSION FOR 3D AND COMBINED 2D3D MATERIALS

TABLE S1: Training and test set performance of three h_{form} regression models on the *3D-perovskites* and *2D3D-combined* datasets. MAE are in units of eV/atom.

	Training set		**Test set**		
	R^2	MAE	R^2	r	MAE
3D perovskites dataset					
KRR-rbf	0.79	1.03	0.61	0.75	1.37
SVM-rbf	0.81	0.78	0.63	0.76	1.24
RF	0.86	0.83	0.63	0.80	1.40
Combined 2D-3D dataset					
KRR-rbf	0.86	0.90	0.76	0.77	1.10
SVM-rbf	0.88	0.71	0.77	0.78	1.01
RF	0.92	0.66	0.78	0.79	1.09

II. STRUCTURE AND CHEMISTRY OF MATERIALS

Since the dataset contains many different types of materials, it is difficult to give a comprehensive discussion on their chemistry. Some basic information has already been given in the main manuscript. Here we would like to add that the elemental compounds contained in our data set are C, Si, Ge, P, As and Sn. The compounds are composed of one or two TM atoms in the unit cell along with various combinations of pnictogens, chalcogens, halogens, H etc. There are no materials with more that two types of TM atoms in the unit cell, and there are no metallic alloys either.

For the sake of illustration, we give structure figures for four compounds that are elemental, binary, ternary and quaternary respectively.

III. DISTRIBUTION OF SCORING PARAMETERS

In the following subsections, we show the distribution of values of the scoring parameters used to train each of the classification and regression models for 100 different train-test

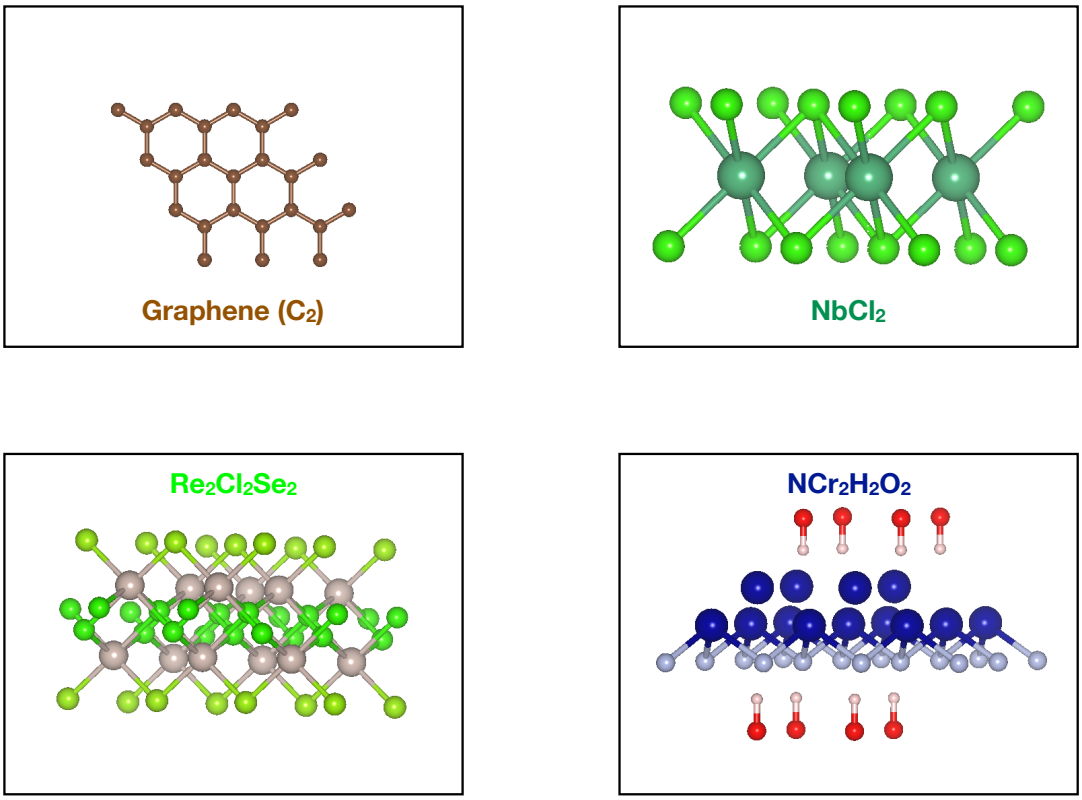

FIG. S1: Structure of a few materials contained in the data set. One each of elemental, binary, ternary and quaternary materials has been shown.

splits. f_1 score was used as the scoring parameter for the classification models, and mean absolute error (MAE, defined in the main manuscript) was used for the regression models. **Training set values for the scoring parameter for each model are also given.** Narrow, unimodal distribution of the scoring parameters (except for SVM-rbf and KRR for moment regression) in each case shows that models are trained reliably even with the small datasets we have.

A. Stability classification models

Training set f_1 scores:

SVM-linear: 0.92; SVM-rbf: 0.98; RF: 0.99.

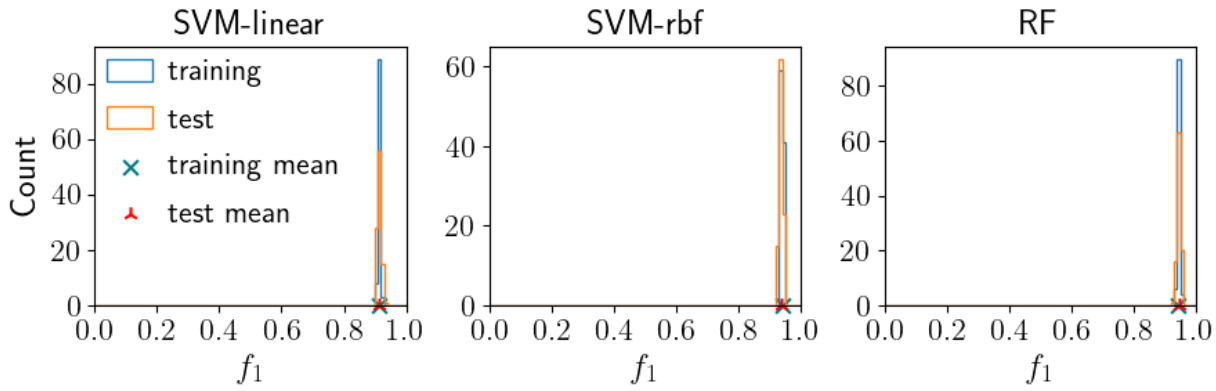

FIG. S2: Distribution of training and test f_1 scores for different classification models over 100 training instances.

B. Regression models for h_{form}

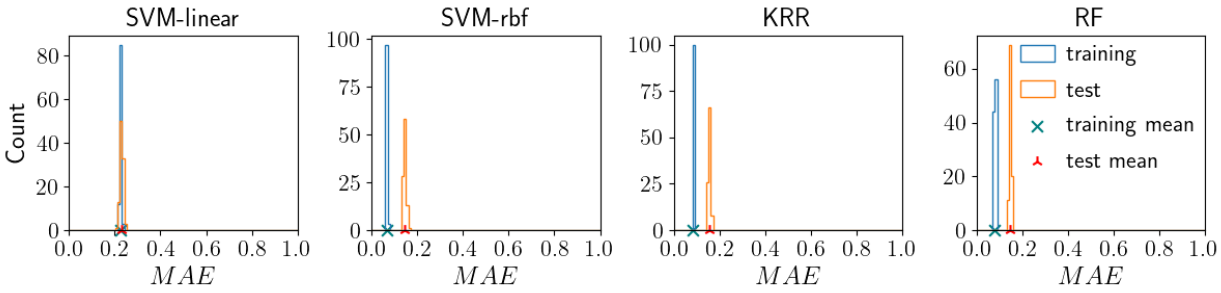

FIG. S3: Distribution of training and test MAE scores for different h_{form} regression models over 100 training instances.

Training set MAE values (eV/atom):

SVM-linear: 0.23; SVM-rbf: 0.07; KRR: 0.08; RF: 0.08.

C. Magnetic-NM classification models

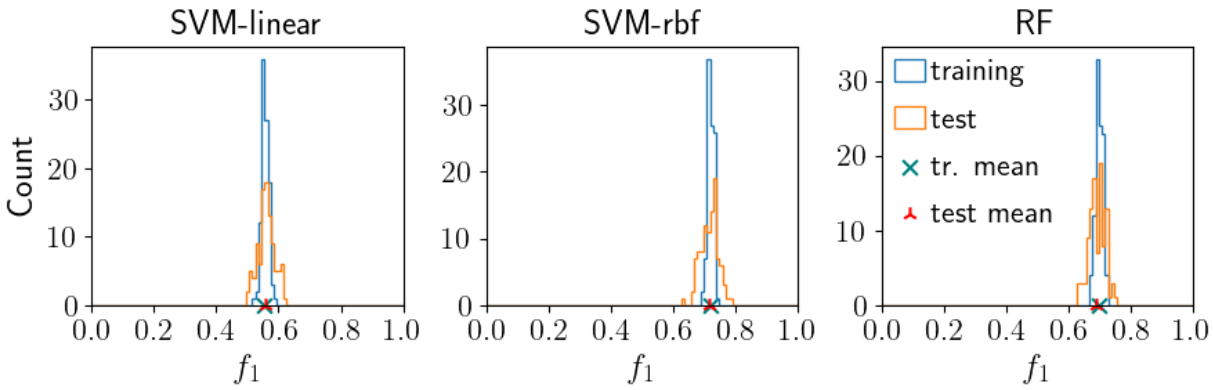

FIG. S4: Distribution of training and test f_1 scores for different magnetic-NM classification models over 100 training instances.

Training set f_1 scores:

SVM-linear: 0.56; SVM-rbf: 0.72; RF: 0.70.

D. Regression models for magnetic moment

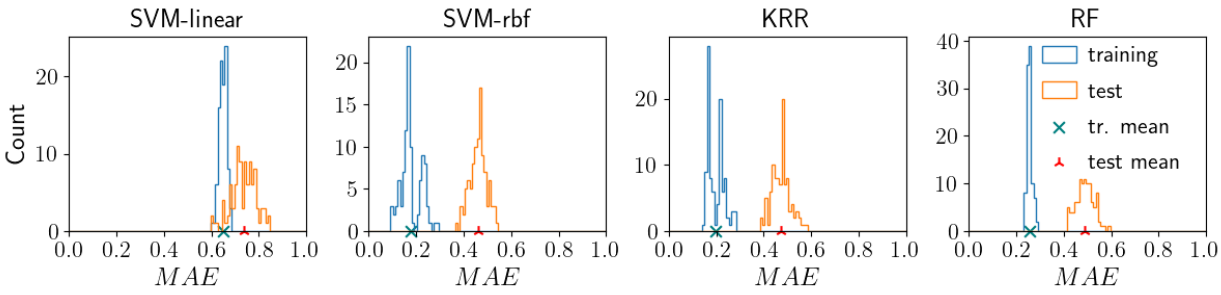

FIG. S5: Distribution of training and test MAE scores for different magnetic moment regression models over 100 training instances.

Training set MAE values (eV/atom):

SVM-linear: 0.65; SVM-rbf: 0.15; KRR: 0.21; RF: 0.26.

E. Anisotropy energy classification models

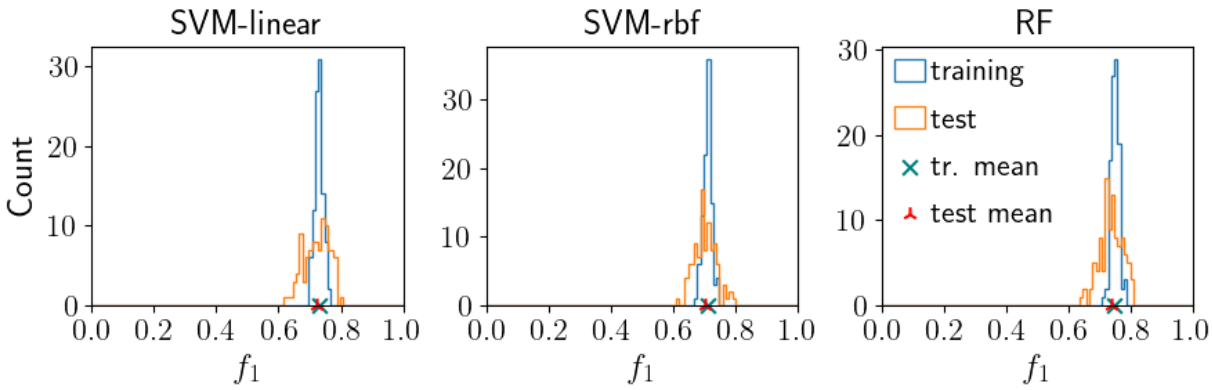

FIG. S6: Distribution of training and test f_1 scores for different low-high magnetic anisotropy classification models over 100 training instances.

Training set f_1 scores:

SVM-linear: 0.73; SVM-rbf: 0.71; ; RF: 0.75.

IV. REGRESSION MODELS FOR h_{form} ON THE *DATA-FULL* DATASET

In this section we present results of training the RF regression model for h_{form} on the entire *data-full* dataset.

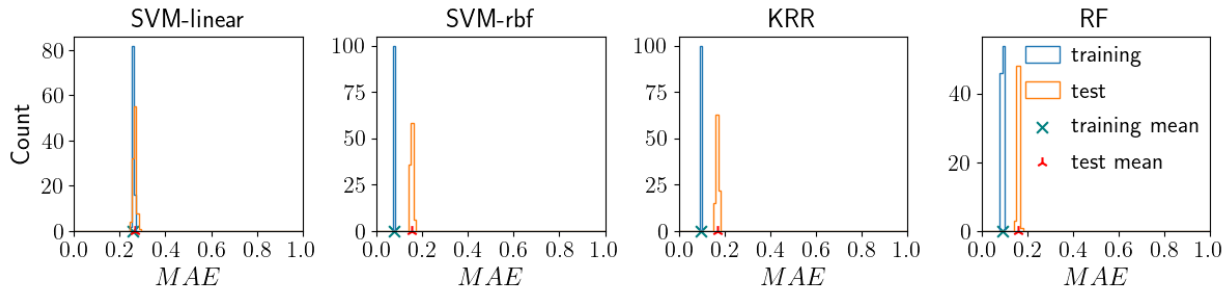

FIG. S7: Distribution of training and test MAE scores for different regression models over 100 training instances using the *data-full* dataset.

TABLE S2: Test set R^2, correlation (r) and mean absolute error (MAE, eV/atom) of different regression models for prediction of heat of formation. *data-full* dataset is used for training and testing.

Model	R^2	r	MAE
KRR-rbf	0.83	0.91	0.17
SVM-linear	0.62	0.80	0.27
SVM-rbf	0.85	0.91	0.15
RF	0.85	0.93	0.16

Training set MAE values (eV/atom):

SVM-linear: 0.26; SVM-rbf: 0.08; KRR: 0.10; RF: 0.10.