Supplementary Information

A critical examination of compound stability predictions from machine-learned formation energies

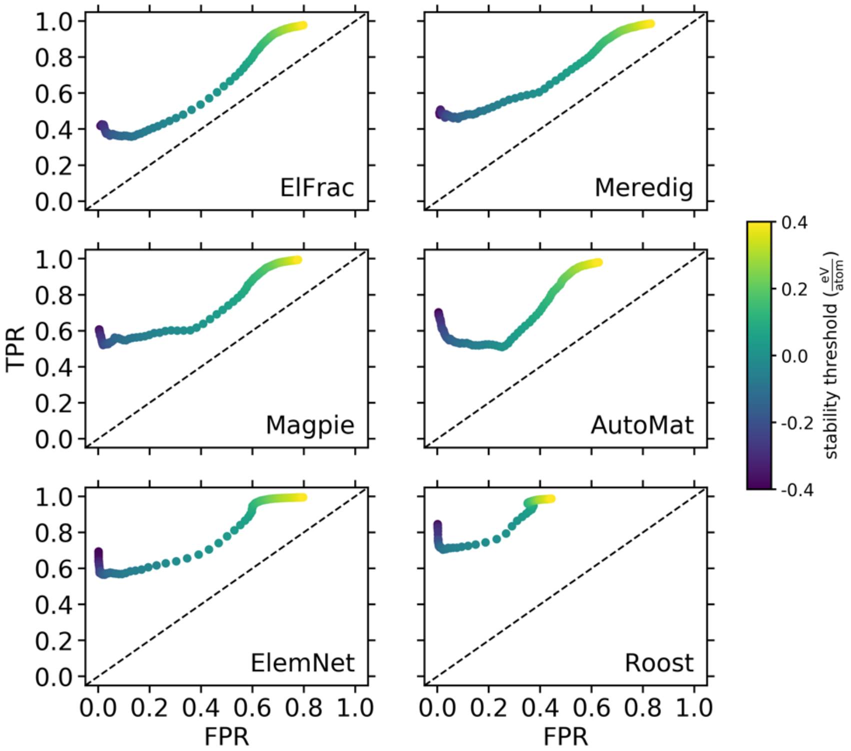
Christopher J. Bartel[1*], Amalie Trewartha[1], Qi Wang[2], Alexander Dunn[1,2], Anubhav Jain[2], Gerbrand Ceder[1,3*]

[1]Department of Materials Science & Engineering, University of California, Berkeley, Berkeley, CA 94720, USA

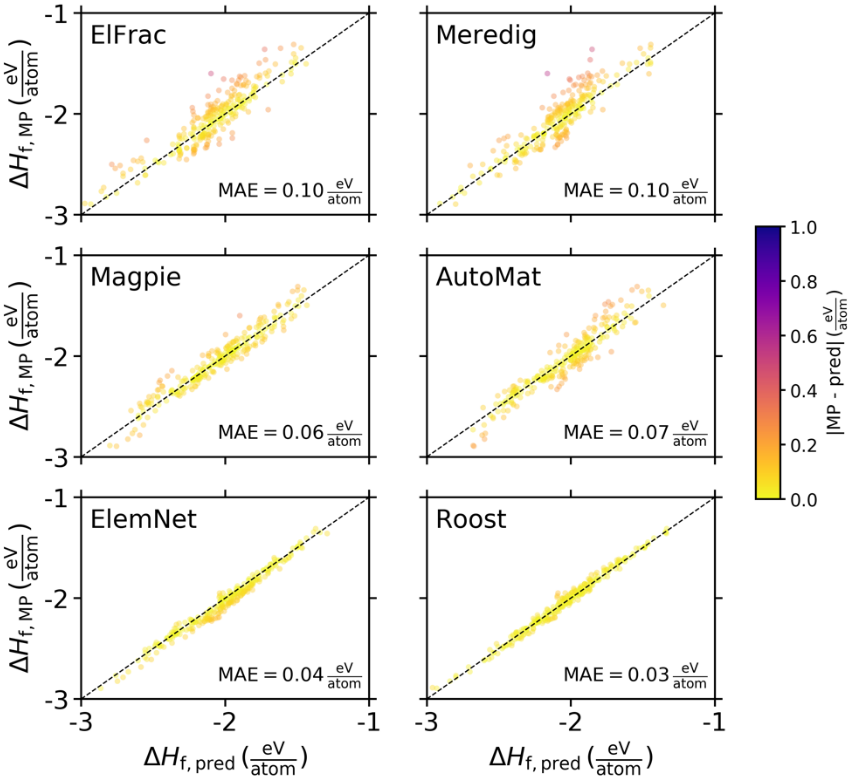
[2]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

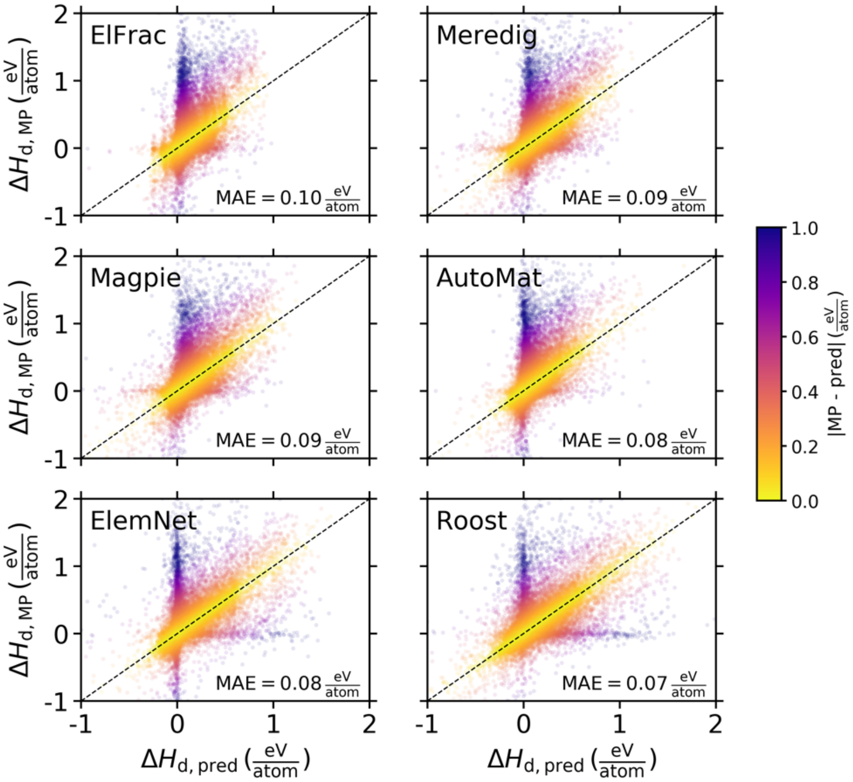
[*]correspondence to cbartel@berkeley.edu, gceder@berkeley.edu



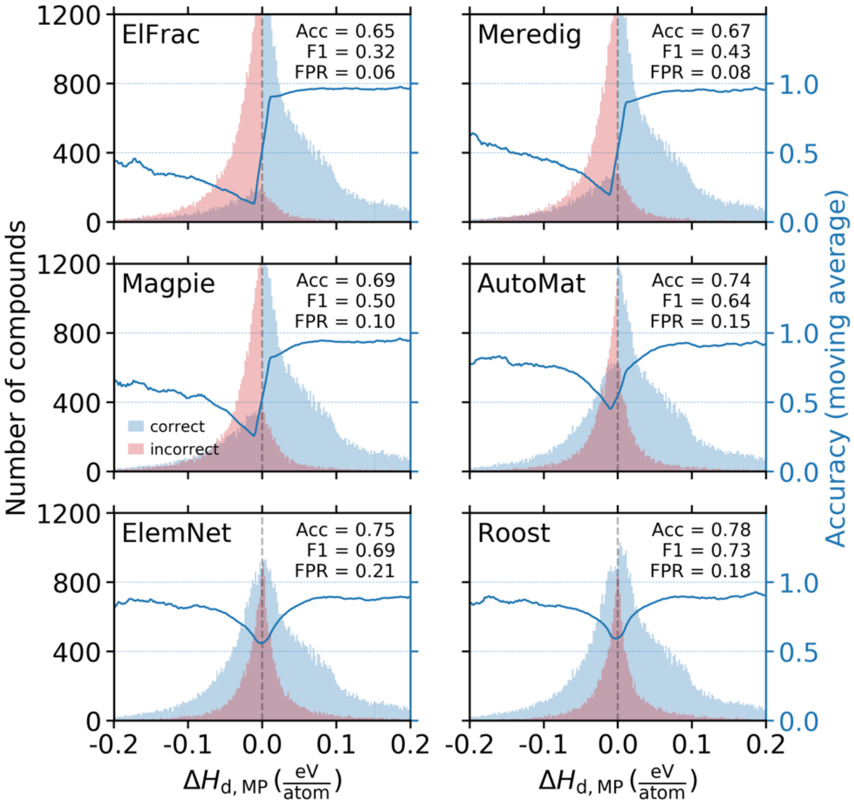
Supplementary Figure 1. Receiver operating characteristic (ROC) curves for each model trained on ΔH_f. TPR is the true positive rate and FPR the false positive rate. The colorbar indicates the stability threshold – i.e., a compound is classified as "stable" if ΔH_d is less than the stability threshold. Note that the models are trained on ΔH_f and are therefore insensitive to this changing threshold. Instead, the choice of threshold simply allows for an expanded analysis of the ΔH_f model performance on ΔH_d predictions.



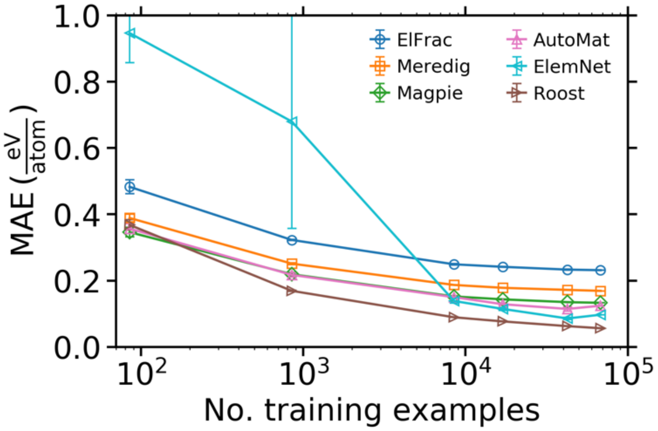
Supplementary Figure 2. Re-training each model on all of MP minus 267 quaternary compounds in the Li-Mn-TM-O chemical space (TM ∈ {Ti, V, Cr, Fe, Co, Ni, Cu}) and predicting ΔH_f for each of the excluded compounds ($\Delta H_{f,pred}$) and comparing to MP, $\Delta H_{f,MP}$. All annotations are the same as in **Figure 2**.



Supplementary Figure 3. Reproducing **Figure 3** but training on ΔH_d instead of ΔH_f. All annotations are the same as in **Figure 3**.



Supplementary Figure 4. Reproducing **Figure 4**, but training on ΔH_d instead of ΔH_f. All annotations are the same as in **Figure 4**.



Supplementary Figure 5. Learning curves for all compositional models. The MAE on predicting ΔH_f as a function of number of compounds used for training. Performance is shown on the test set, which is all MP compounds except those used for training. The MAE is averaged over five random splits of the training/testing compounds with the standard deviation in MAE over these five splits shown as the error bar. The final data point for each model at 68,011 training examples was taken from the 5-fold cross validation shown in **Figure 2**.

Supplementary Table 1. Reproducing **Table 1** but training on ΔH_d instead of ΔH_f.

	ElFrac	***Meredig***	***Magpie***	***AutoMat***	***ElemNet***	***Roost***
candidate compounds	13,659	13,659	13,659	13,659	13,659	13,659
stable compounds in MP	9	9	9	9	9	9
compounds predicted stable	0	0	0	0	58	299
% predicted stable	0	0	0	0	0.4	2.2
pred. stable and stable in MP	0	0	0	0	0	0

Supplementary Table 2. The performance of each compositional representation trained to classify compounds as stable ($\Delta H_d \leq 0$) or unstable ($\Delta H_d > 0$). Note that the *Roost* representation is excluded from this analysis as described in Methods.

	Accuracy	***F_1 score***	***False positive rate***
ElFrac	0.723	0.631	0.191
Meredig	0.745	0.666	0.180
Magpie	0.759	0.683	0.170
AutoMat	0.792	0.732	0.153
ElemNet	0.744	0.683	0.219

Supplementary Table 3. Training and inference times for learning and predicting ΔH_f. Training time is the time required to train the models on 80% of the MP dataset (68,011 compounds). Inference time is the time required to predict ΔH_f for the remaining 20% of the MP dataset (17,013 compounds). Note that for *AutoMat*, the training time is a user-specified input.

	Training time (h)	***Inference time (s)***
ElFrac	0.02	15
Meredig	0.06	15
Magpie	0.05	15
AutoMat	10.00	2719
ElemNet	2.35	8
Roost	3.47	38
CGCNN	20.90	926