Supplementary Information for A General Atomic Neighborhood Fingerprint for Machine Learning Based Methods

Rohit Batra[1], Huan Doan Tran[1], Chiho Kim[1], James Chapman[1], Lihua Chen[1], Anand Chandrasekaran[1], and Rampi Ramprasad[1]

[1]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, USA

1 Classification

1.1 Performance of Scalar, Vector and Tensorial Components

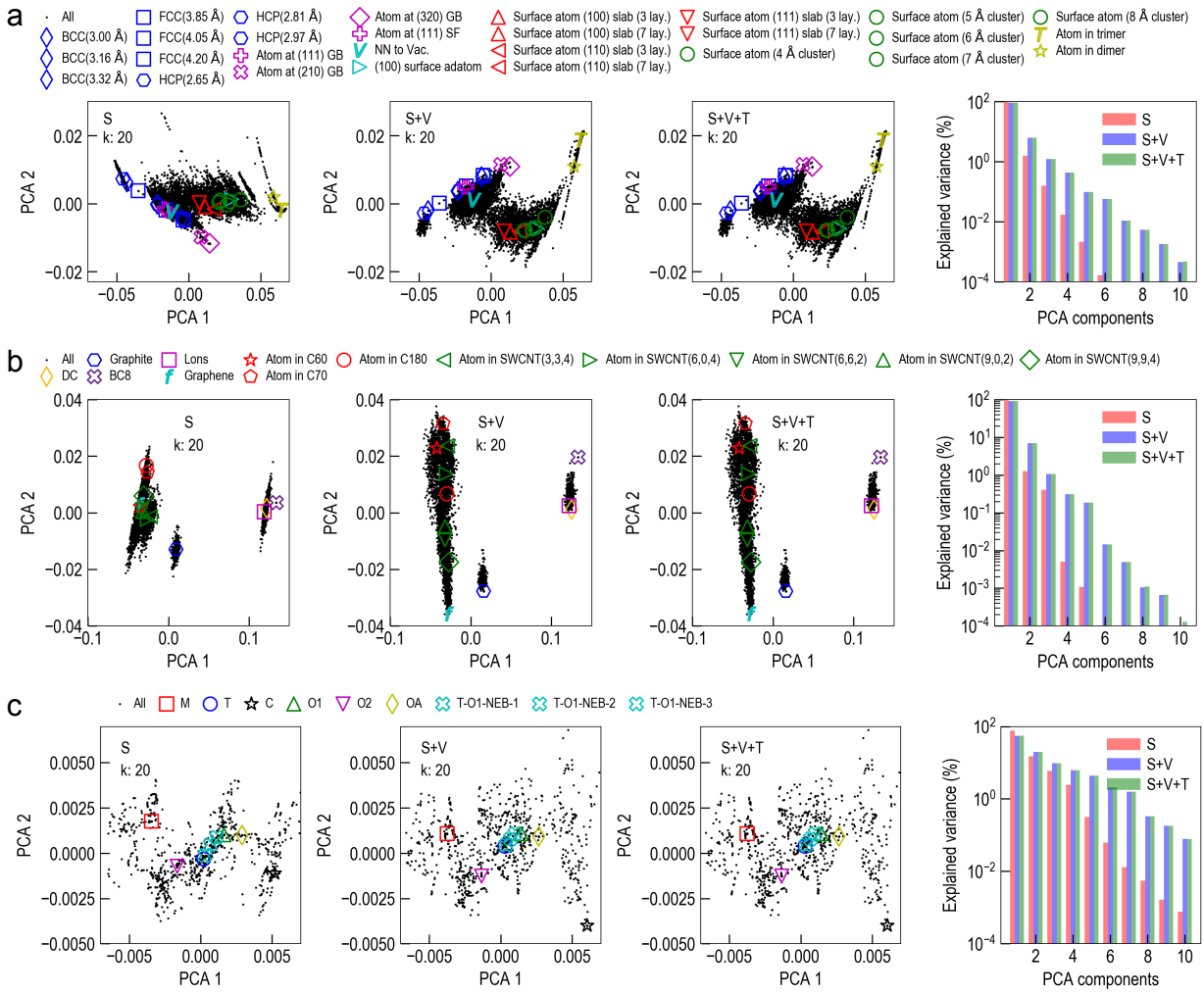

Figure S1: Atomic fingerprints of (a) Al, (b) C and (c) hafnia data sets illustrated using the first two principal components. While important atomic environments are highlighted using large colored symbols, thermal fluctuations in atomic configurations obtained from DFT-based MD simulations are presented using small black symbols. The first three panels in each case demonstrate the systematic improvement in the classification performance when moving from scalar (S), to scalar and vector (S+V), and to finally, scalar, vector and tensorial components (S+V+T). Also, notice the increase in the spread of the PCA space moving from S, to S+V, to S+V+T panels in each case. All of the right-most panels demonstrate the systematic increase in the information captured by the fingerprint on addition of more complex vector and tensorial components.

1.2 Classification of Carbon Dataset

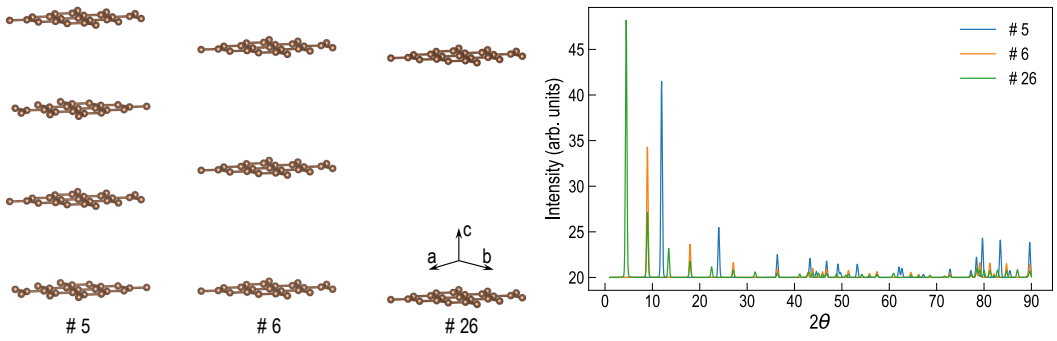

Figure S2: Three graphene structures (#5, 6 and 26 in the main manuscript) identified from the carbon data set using the structure fingerprint. Although physically they all resemble graphene, the structures differ in the length of the vacuum region, resulting in different X-ray diffraction pattern as shown in the right-most panel.

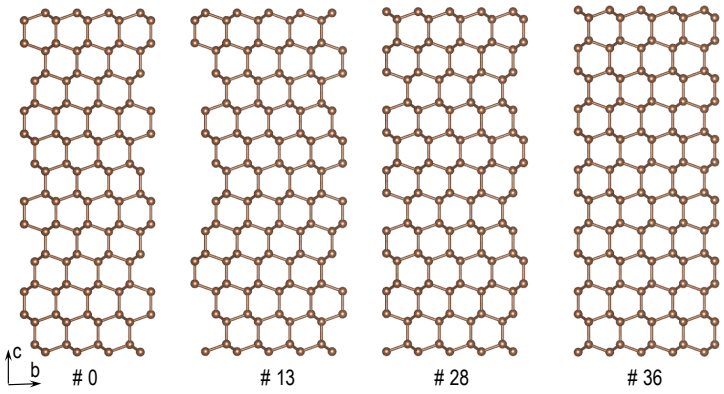

Figure S3: Four C phases (# 0, 13, 28 and 36 in the main manuscript) with space group $P6_3/mmc$ identified to be structurally close using the structure fingerprint definition. Only subtle difference in the bonding pattern is evident, in agreement with the findings based on the structure fingerprint.

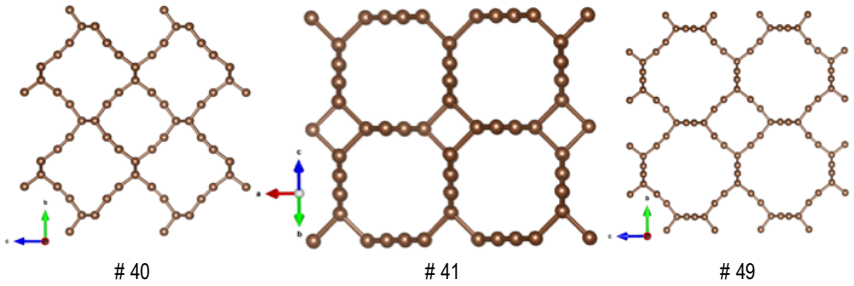

Figure S4: Three C phases (# 40, 41 and 49 in the main manuscript) with space group $I4/mmm$ identified to be structurally different using the structure fingerprint definition. Large variations in the bonding pattern is evident, in agreement with the findings based on the structure fingerprint.

2 Regression: Energy Model for Aluminum

2.1 Strategies for Mapping the Energy Model

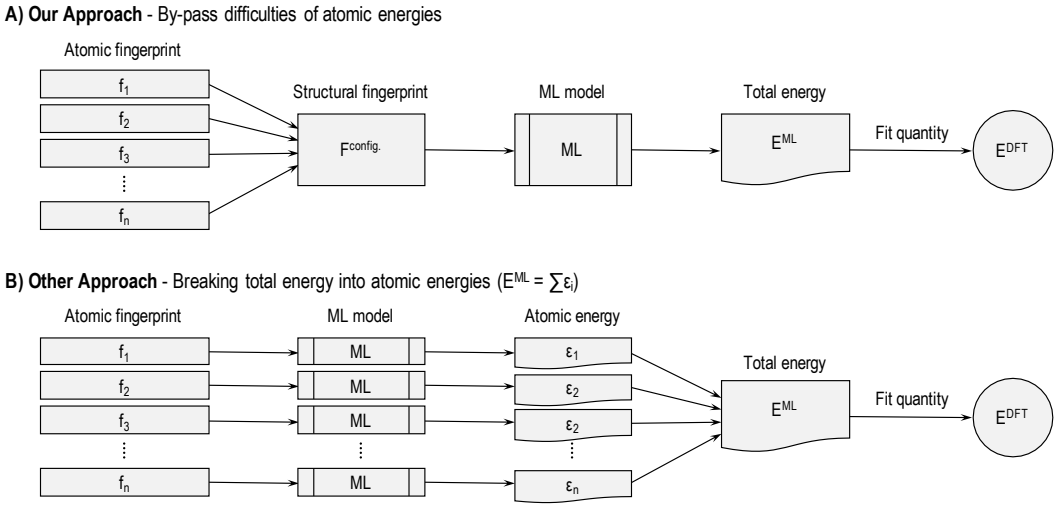

Figure S5: Two general strategies for building ML-based energy models. In (a), the atomic fingerprints are combined in some fashion to arrive at a structure fingerprint, which is then fit to the reference total energy (usually available from DFT computations). In (b), a common ML model is applied to atomic fingerprints to obtain fictitious atomic energies, which are then summed to obtain the total energy of the system. In this approach too, the ML predicted total energy, obtained from summing over the atomic energies, is fit to the reference total energy.

2.2 Aluminum Energy Model: Overall Scheme

Fig. S6 displays the different stages involved in the construction of the Al energy model. Starting from a diverse reference dataset of Al configurations and associated energies, generated using DFT-based MD simulations, a pool of ~ 100 low training error kernel ridge regression (KRR) based Al models are built within step 1, 2 and 3. For this, the structure fingerprint definition is used to numerically represent an Al configuration in step 2. In step 4, all of these energy models are subjected to a series of pre-defined tests (*e.g.* vacancy formation energy), wherein the models are evaluated for their prediction accuracy. Finally, the energy model that has the best overall performance is selected.

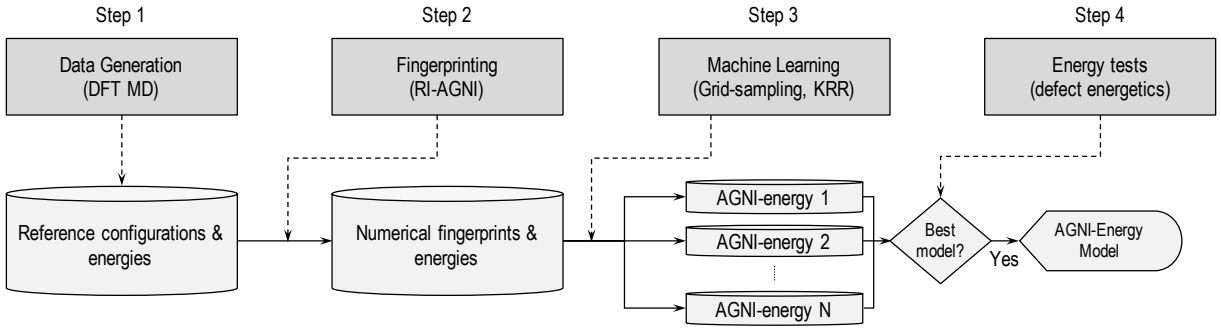

Figure S6: The overall scheme adopted to build the Al energy model using the structure fingerprint.

2.3 Learning Curves

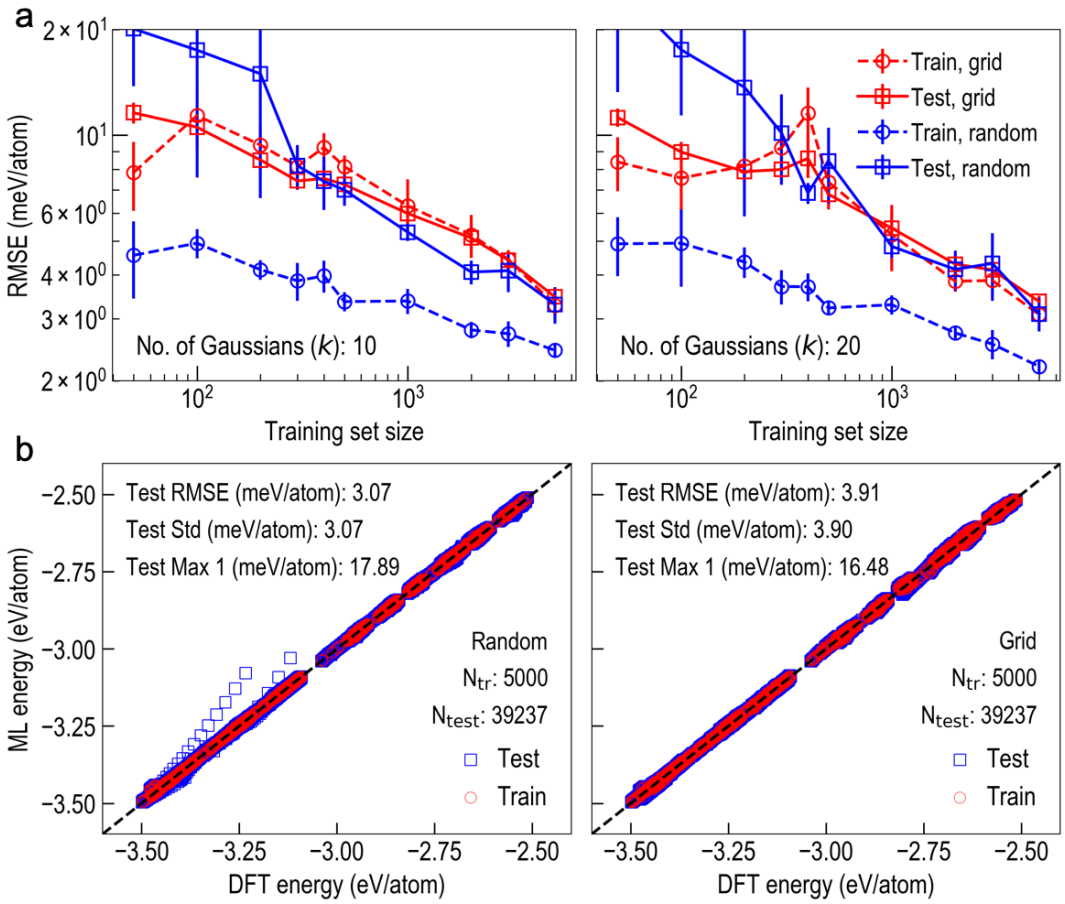

Figure S7: (a) Effect of for training set sampling methods (random and grid), and the number of Gaussians (k) on the learning curves of Al energy models. (b) Parity plots comparing the effect of random and grid sampling methods on the prediction accuracy of Al energy models.

Learning curves of the Al energy models presented in Fig. S7(a) show the similarity in the model performance for (1) both the random and the grid sampling methods for training set selection, and (2) using either 10 or 20 number of Gaussians (k) during fingerprinting. This suggests that evaluating root mean square error (RMSE) alone may not be sufficient, and other errors metrics should be employed to further evaluate these cases. Parity plots presented in Fig. S7(b) clearly illustrate the superiority of the grid sampling approach over the random sampling. Notice how both sampling methods result in models with similar test errors, further corroborating the problem of relying on limited error metrics. Based on these results, grid sampling was used in this work to construct Al energy model.

2.4 Model Selection

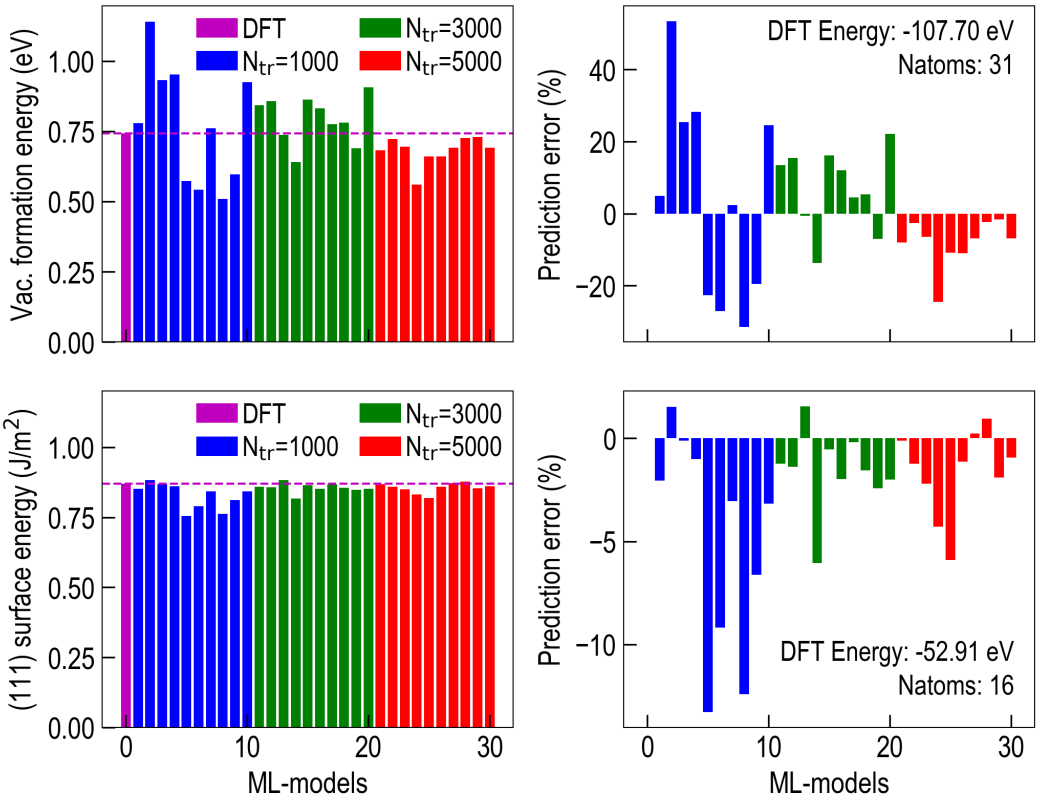

Figure S8: Vacancy formation and (111) surface energy tests to evaluate the performance of Al energy models built using different number of training examples.

As part of step 4 in the overall scheme, a pool of Al energy models are evaluated for various defect formation energies. Fig. S8 presents the results for few of such tests for Al energy models built using 1000, 3000 and 5000 training examples. The model with best overall performance is selected based on these tests. As expected, models with 5000 training points can be seen to outperform the other models.

2.5 Computation of Elastic Constants

Table 1: Performance of the Al energy model to capture various elastic constants, in comparison to DFT. μ' and μ'' represent the two definitions of shear modulus.

Elastic constant	DFT	ML
c_{11} (GPa)	117.16	121.03
c_{12} (GPa)	60.44	60.17
$\mu'' = c_{44}$ (GPa)	34.17	33.50
$\mu' = (c_{11} - c_{12})/2$ (GPa)	28.36	29.93

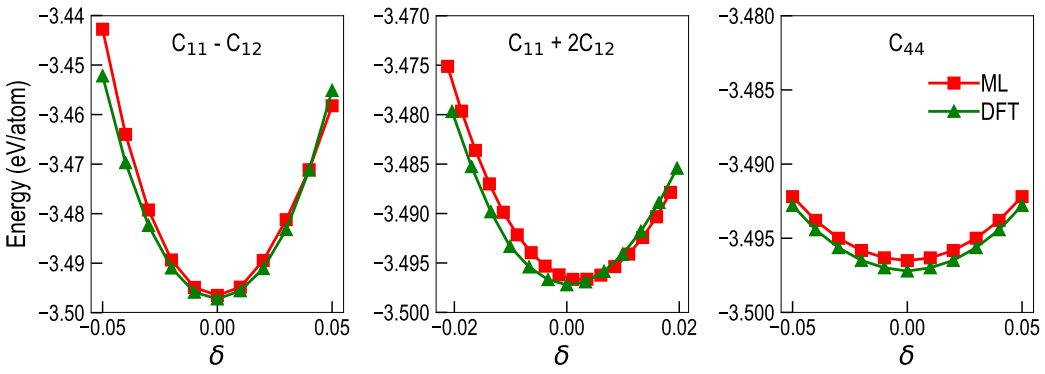

Figure S9: Energy as a function of strain introduced in the Al unit cell. See text for the definition of strain in each case.

In order to further test the energy model, elastic constants of Al in the fcc phase ($Fm\bar{3}m$) were determined. These were obtained by fitting following mechanical relations to the energy (evaluated using the energy model) of the deformed unit cells, (i) $3(c_{11}-c_{12})\delta^2$, when $e_{11}=e_{22}=\delta, e_{33}=-2\delta$, (ii) $\frac{3}{2}(c_{11}+2c_{12})\delta^2$, when $e_{11}=e_{22}=e_{33}=\delta$, and (iii) $(c_{44}\delta^2)/2$, when $e_{12}=e_{21}=\delta/2$, where c_{ij} are the elastic constants, and e_i are the components of the strain tensor given in Voigt notation. The theory behind these relations could be find elsewhere (Ding, Wen-Jiang, et al., Solid State Sciences, 14.5, 555 (2012)).