Supporting information

An Adaptive Design Approach for Defects Distribution Modeling in Materials from First Principle Calculations

Maicon Pierre Lourenço,[1] Alexandre dos Santos Anastácio.[1] Andreia L. da Rosa[2,3], Thomas Frauenheim[3,4] and Maurício Chagas da Silva[3,5].*

1- *Departamento de Química e Física – Centro de Ciências Exatas, Naturais e da Saúde – CCENS – Universidade Federal do Espírito Santo, 29500-000, Alegre, Espírito Santo, Brazil*
2- *Instituto de Física, Universidade Federal de Goiás, Av. Esperança, s/n - Campus Samambaia, Goiânia - GO, 74690-900, Brazil*
3- *Bremen Center for Computational Materials Science, University of Bremen, P.O. Box 330440, 28334 Bremen, Germany*
4- *Computational Science Research Center, No.10 East Xibeiwang Road, Beijing 100193, Computational Science and Applied Research Institute Shenzhen, China*
5- *Max Planck Institute for the Structure and Dynamics of Matter; Luruper Chaussee 149, Geb. 99 (CFEL), 22761 Hamburg, Germany*

** Address correspondence to:* maiconpl01@gmail.com(MPL).

1. Density of states of pure and modified goethite

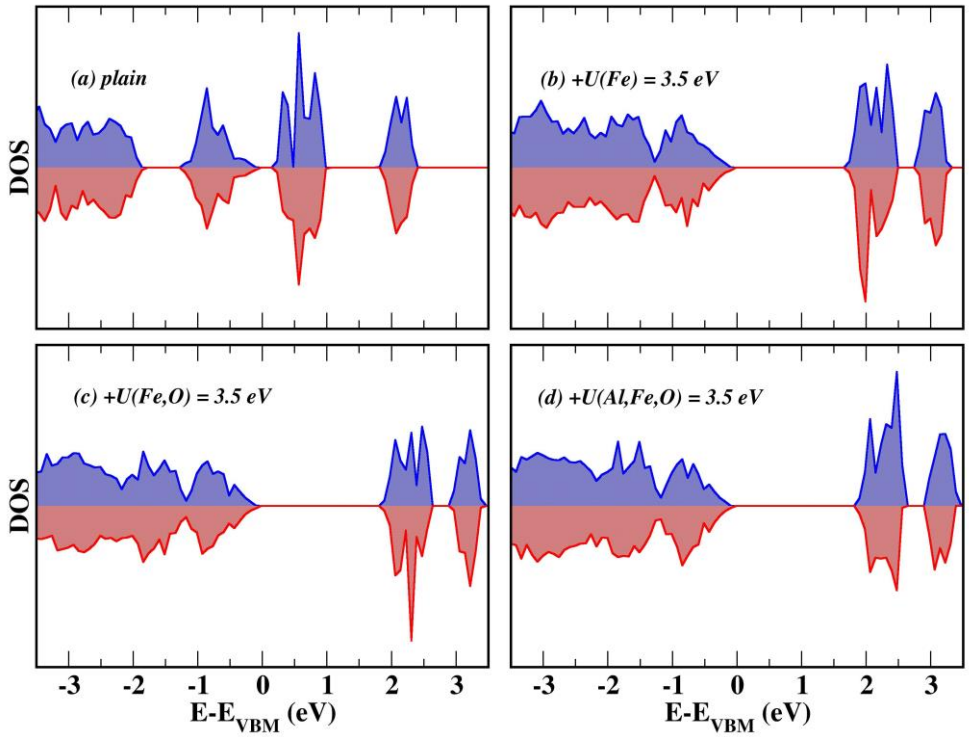

Figure S1- The GGA+U effect to the density of states (DOS) of the modified goethite $Fe_{0.875}Al_{0.125}OOH$. The blue is spin-up and red is the spin-down DOS.

2. Statistical Regression

The idea behind the statistical regression is to obtain the N observed properties **y** = ($y^{(1)}$, …,$y^{(n)}$), i = 1, …., N, to describe statistically **y**, the descriptor $\mathbf{x}^{(i)}$ = ($x_1^{(i)}$, …,$x_k^{(i)}$), with K variables is required. This result in a matrix **X** of dimension (N × K), called feature matrix which is associated with the one-dimension vector **y** (the objective function) of dimension (N).

By modelling the desired problem in this manner, several surrogate models, such as the Multiplayer Perceptron Regressor (MLP) (an Artificial Neural Network, ANN, regressor), can be used by exploring high level libraries, such as the scikit-learn[1].

After performing the regression, the statistical model is obtained and represented as:

$$\hat{y} = \hat{f}(X), \qquad \textbf{(1)}$$

where $\hat{y}$ is the vector with the predicted properties and $\hat{f}(X)$ is the statistical model (the predictor)designed from the MLP regressor, for instance.

Usually, to obtain a model without data bias, the matrix **X** and **y** – which defines the initial data to obtain and test the ML models ($\mathbf{X}^{l}$,$\mathbf{y}^{l}$) – is split in two other matrices: ($\mathbf{X}^{train}$, $\mathbf{y}^{train}$), which is used to train the statistical model and the ($\mathbf{X}^{test}$, $\mathbf{y}^{test}$) to validate it. Moreover, the matrix ($\mathbf{X}^{train}$, $\mathbf{y}^{train}$) is partitioned by Cross Validation (CV) and for each partition p (in a total space of P) it is obtained a statistical model, hence: $\hat{y}^{p} = \hat{f}(X)$, p = 1, ..., P. Then, for each descriptor in the observed data set $\mathbf{x}^{(i)}$ (or for each descriptor j non-observed: $\mathbf{x}^{(j)}$) it is obtained the average $\mu(\mathbf{x}^{(i)})$ and (or $\mu(\mathbf{x}^{(j)})$) and the standard deviation $\sigma(\mathbf{x}^{(i)})$ (or $\sigma(\mathbf{x}^{(j)})$), as illustrated in figure S1 for the data in the space. The same description is valid when dealing with Gaussian Process (GP) surrogate model, but the Cross Validation part, since the $\mu(\mathbf{x}^{(i)})$ and (or $\mu(\mathbf{x}^{(j)})$) and the standard deviation $\sigma(\mathbf{x}^{(i)})$ (or $\sigma(\mathbf{x}^{(j)})$), as illustrated in figure S1, is obtained directly from the GP regression, not requiring the CV or any other method to obtain it.

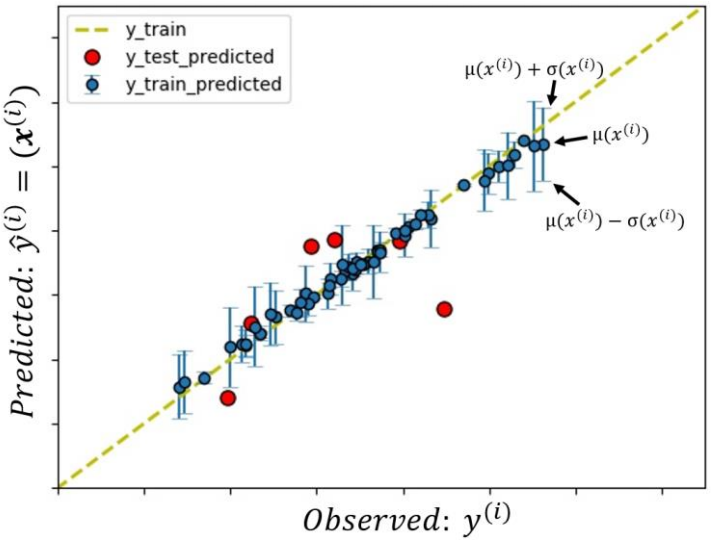

Figure S2- Plot of the observed $y^{(i)}$ and predicted $\hat{y}(x^{(i)})$ target property. The use of Cross-Validation, by splitting the (X^{train}, y^{train}) K times, allows us to have K regression models and their performance in each data point $x^{(i)}$ is represented from the mean

$\mu(x^{(i)})$ and the standard deviation $\sigma(x^{(i)})$ of the predicted target. The abscissa is the observed property and the ordinate the predicted one.

As it will be discussed, the mean $\mu(\mathbf{x^{(i)}})$ and the standard deviation $\sigma(\mathbf{x^{(i)}})$ for each descriptor entry will be used to obtain the acquisition function[2, 3] which is used to indicate the next candidate to be evaluated from computational simulation or even experiment. The next candidate is, then, incorporated in the initial descriptor matrix: ($\mathbf{X}^{\text{train+1}}$, $\mathbf{y}^{\text{train+1}}$) and the iteration and the iteration process continue one step more until the optimization of the target property.

3. Expected improvement for minimum search and a 3D plot

It is important to highlight that if one wants to search for the global or local minimum, the expected improvement should be written in the following way[3]:

$$E\left[I\left(x^{(j)}\right)\right] = \left(f_{min} - \mu\left(x^{(j)}\right)\right)\Phi\left(\frac{f_{min}-\mu\left(x^{(j)}\right)}{\sigma\left(x^{(j)}\right)}\right) + \sigma\left(x^{(j)}\right)\phi\left(\frac{f_{min}-\mu\left(x^{(j)}\right)}{\sigma\left(x^{(j)}\right)}\right), \quad \textbf{(1)}$$

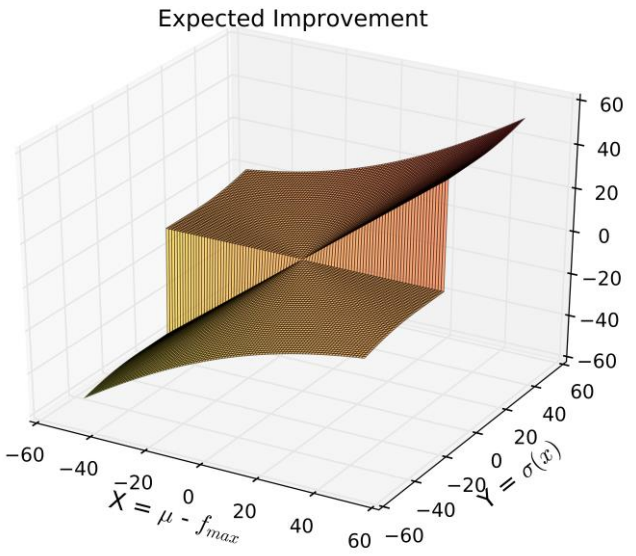

Figure S3- The expected improvement (EI) plot as a function of (μ – f_{max}) and the standard error $\sigma(x)$.

4. Results of the modified goethite

Stability and structure

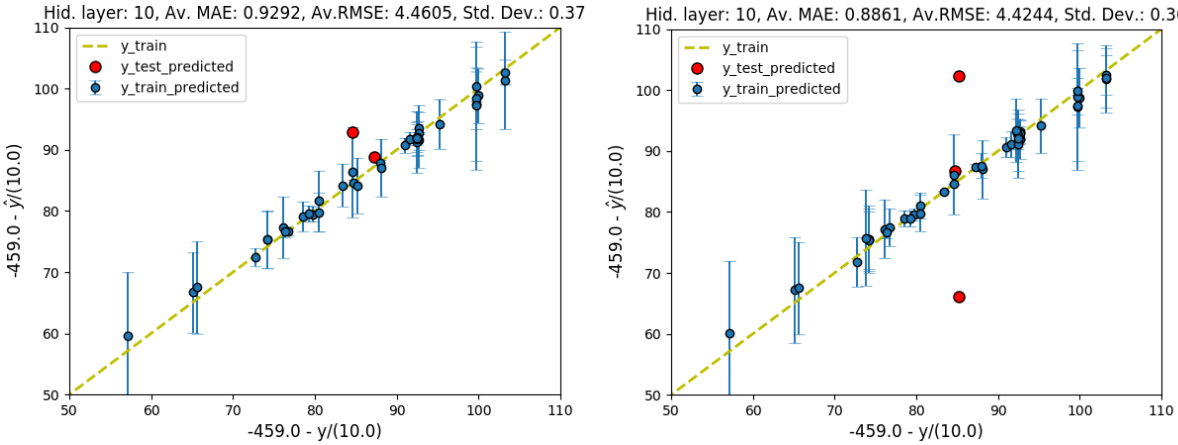

Figure S4- Result of the ANN-10 regression within the AD cycle with four indications (EI-4). Blue points are from the train set and the red from the test. Left: plot of the predicted y (from ANN-10 algorithm) and observed scaled energy of goethite for the initial data with 40 random configurations. Right: plot of the predicted y in the 2nd iteration, which the most stable structure was found by EI. Since the total energy was *scaled* during the AD design, the abscissa is the *observed scaled energy* in eV and the ordinate the *predicted* one. The actual total energy value (the non scaled one in eV) is obtained from the formula present in the graph axes: -459.0 eV – y/10.0, where "y" is the *scaled energy* as seen in any axes. The above statistical labels mean: the average Mean Absolute Error (MAE), the average Root Mean Squared Error (RMSE) and the standard deviation (σ) in the *scaled energy space. They were obtained* from cross-validation. The actual MAE, RMSE and σ can are obtained by dividing their scaled values by 10.0.

Table S1- The AD method applied for the energy optimization of modified goethite by applying the AD with four EI indications (EI-4). The percentage of the of data is related to the initial (computed) sample size, compared to the 2024 possible defect distributions, and to the size of sample when the convergence is achieved. Configuration is the atomic index, from the cartesian coordinate system, where the Al^{3+} replaced the Fe^{3+}. The energy difference between the most stable structure in the sample for certain iteration (E_{sample}) and the global minimum one ($E^{GM}_{min.}$, Figure 8b) is: $E_{sample} - E^{GM}_{min.}$ (in meV).

Initial Sample	# iterations	% of data	Configuration	$E_{sample} - E^{GM}_{min.}$ (meV)
40	0	2.0	74_83_85	79.52
	2	2.4	92_93_94	79.49
50	0	2.5	73_91_94	426.95
	3	3.0	92_93_94	79.49
60	0	3.0	73_90_96	79.90
	30	8.9	84_85_86	79.47
70	0	3.4	78_80_86	401.60
	3	4.0	77_92_94	79.51
80	0	4.0	75_84_88	408.04
	3	4.5	77_92_94	79.51
90	0	4.4	79_87_95	395.19
	3	5.0	92_93_94	79.49
100	0	4.9	78_89_94	149.25
	4	5.7	92_93_94	79.49

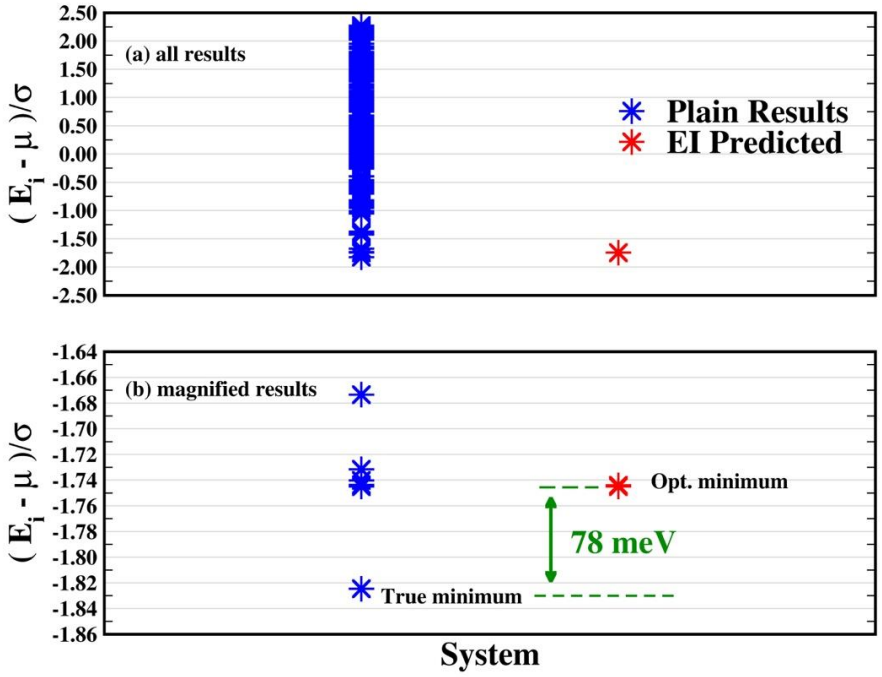

Figure S5- EI minimum energy prediction compared with the true minimum by using the AD method with four EI indications (EI-4).

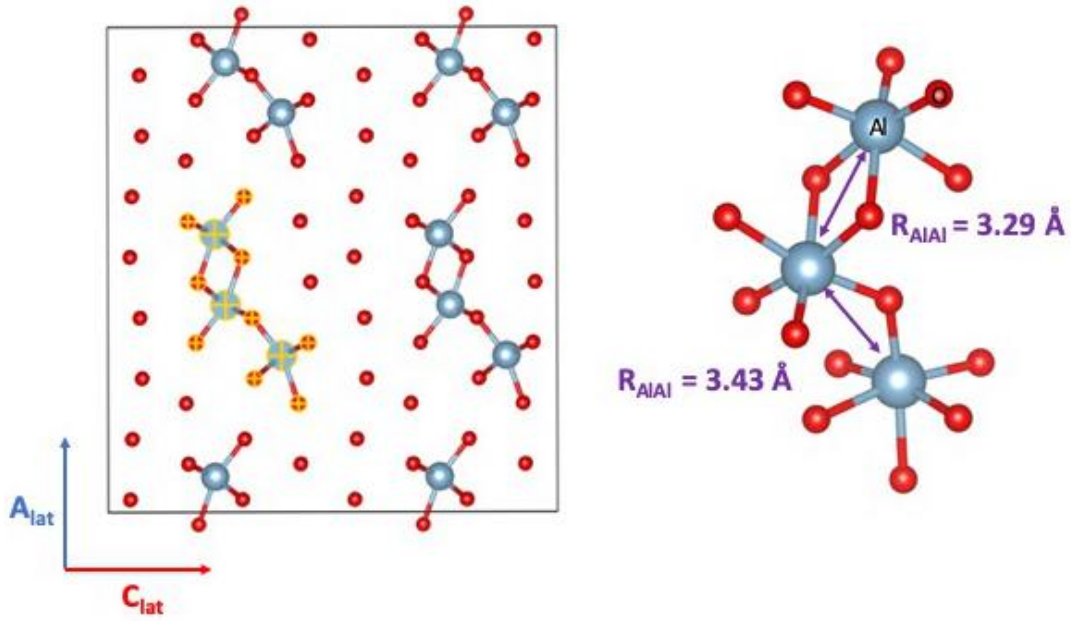

Figure S6- N=40. AD optimized structure with four EI indications (EI-4): index 84_85_86. The supercell model was replicated two times in all directions and the

positions of Fe and H were supressed. In yellow is shown the main Al microstructure present in the system.

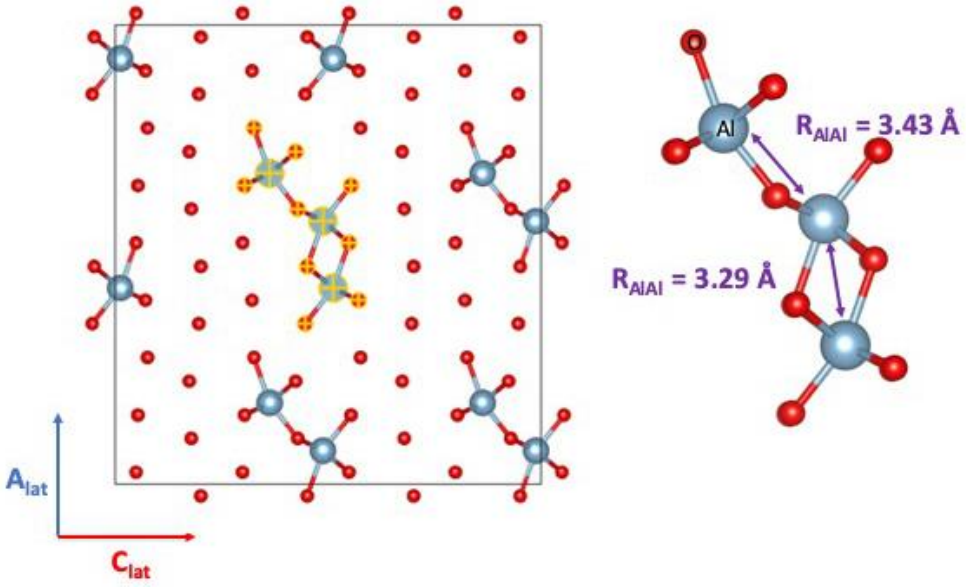

Figure S7- N= 40, 90, 100. AD optimized structure with four EI indications (EI-4): index 92_93_94. The supercell model was replicated two times in all directions and the positions of Fe and H were supressed. In yellow is shown the main Al microstructure present in the system.

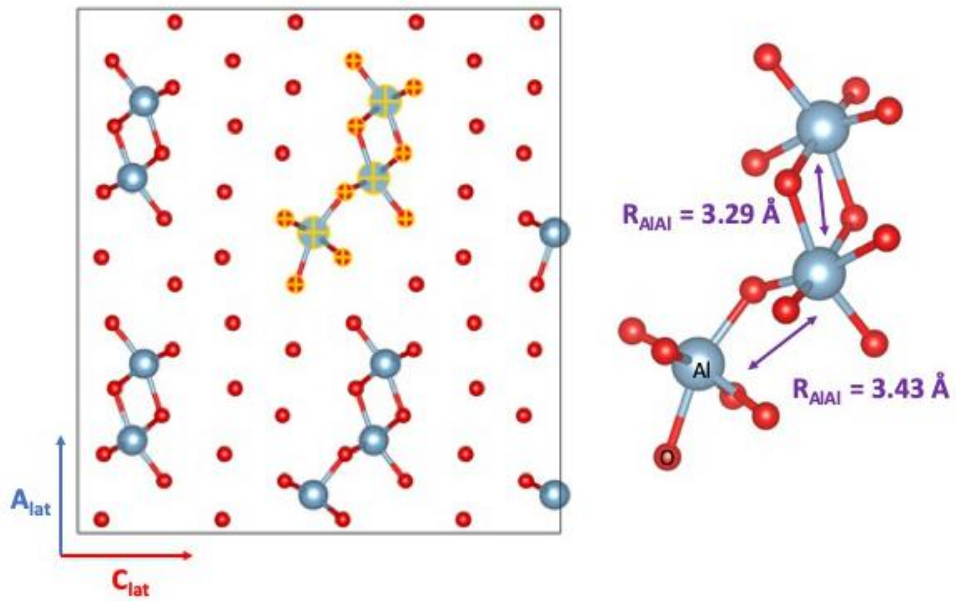

Figure S8- N=60. AD optimized structure with four EI indications (EI-4): index 90_91_93. The supercell model was replicated two times in all directions and the

positions of Fe and H were supressed. In yellow is shown the main Al microstructure present in the system.

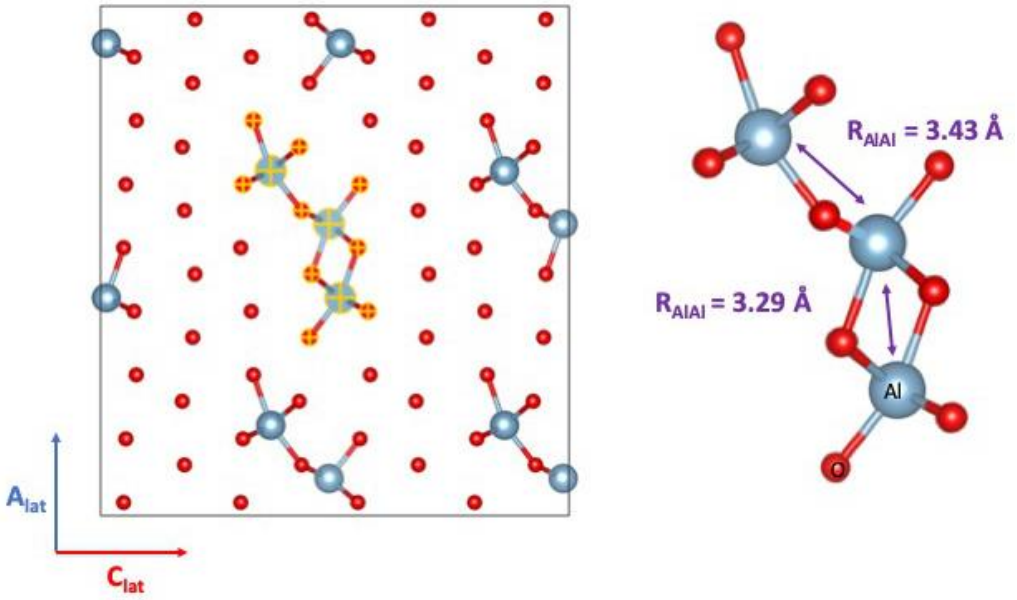

Figure S9- N=70,80. AD optimized structure with four EI indications (EI-4): index 77_92_94. The supercell model was replicated two times in all directions and the positions of Fe and H were supressed. In yellow is shown the main Al microstructure present in the system.

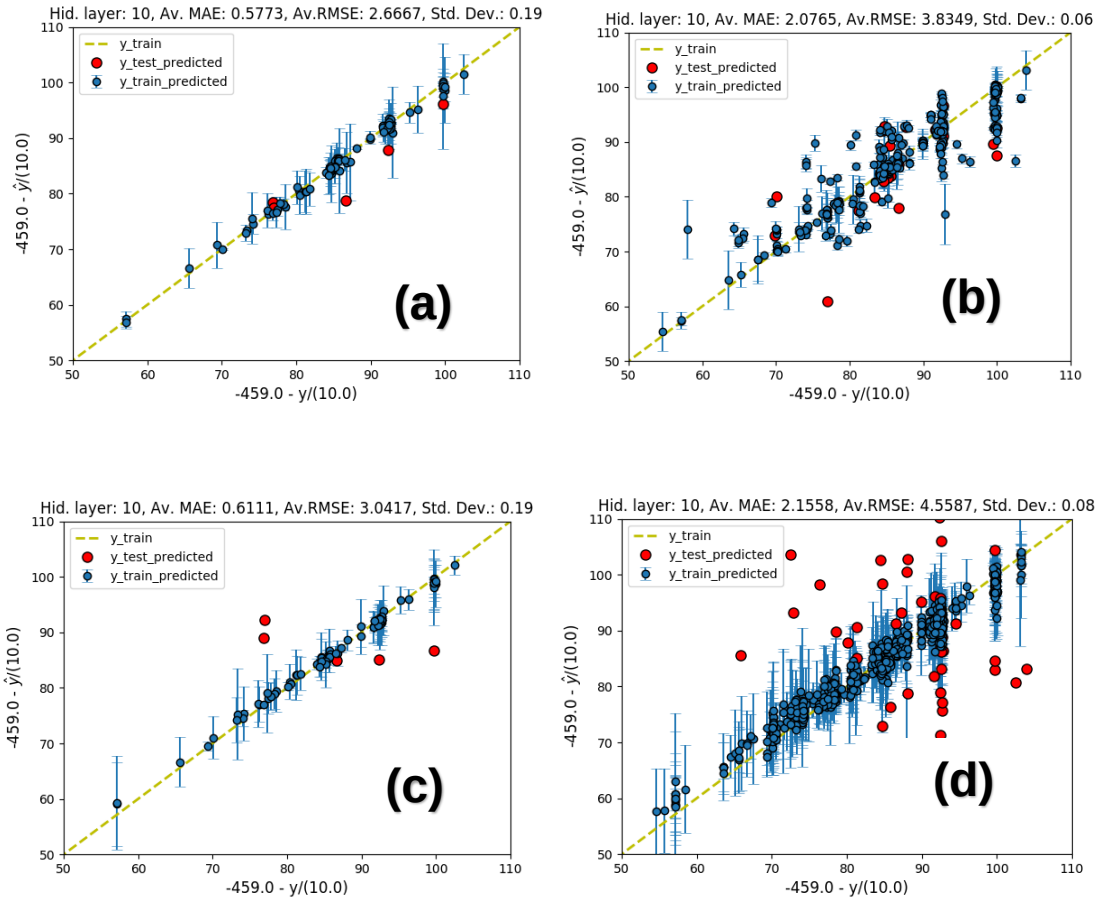

Figure S10- Result of the ANN-10 regression within the AD cycle. Blue points are from the train set and the red from the test. (a) First iteration of the AD of the energy by using ANN-10 with the Ewald Sum Matrix (ESM) descriptor as implemented in Dscribe library[4]. (b) The 66 th iteration where the convergence was achieved by using the ESM descriptor. (c) First iteration of the AD of the energy by using ANN with the Distances descriptor. (d) The 142nd iteration where the convergence was achieved by using the Distances descriptor. For all cases the initial data size is 100. Since the total energy was *scaled* during the AD design, the abscissa is the *observed scaled energy* in eV and the ordinate the *predicted* one. The actual total energy value (the non scaled one in eV) is obtained from the formula present in the graph axes: -459.0 eV – y/10.0, where "y" is the *scaled energy* as seen in any axes. The above statistical labels mean: the average Mean Absolute Error (MAE), the average Root Mean Squared Error (RMSE) and the standard deviation (σ) in the *scaled energy space. They were obtained* from cross-validation. The actual MAE, RMSE and σ can are obtained by dividing their scaled values by 10.0.

4. References

[1] F. Pedregosa, Ga, #235, l. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, #201, d. Duchesnay, J. Mach. Learn. Res., 12 (2011) 2825-2830.
[2] T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, npj Computational Materials, 5 (2019) 21.
[3] D.R. Jones, M. Schonlau, W.J. Welch, Journal of Global Optimization, 13 (1998) 455-492.
[4] L. Himanen, M.O.J. Jäger, E.V. Morooka, F. Federici Canova, Y.S. Ranawat, D.Z. Gao, P. Rinke, A.S. Foster, Computer Physics Communications, 247 (2020) 106949.