Supporting Information

Simple local environment descriptors for accurate prediction of hydrogen absorption and migration in metal alloys

Vladislav Korostelev, James Wagner, Konstantin Klyukin[#]

Department of Mechanical and Materials Engineering, Auburn University, Auburn, AL 36849, USA

Corresponding author: klyukin@auburn.edu

Descriptors engineering: In this study, we examined the absorption energy of hydrogen in three types of interstitial sites: triangular, tetrahedral, and octahedral, which correspond to hydrogen being coordinated by 3, 4, and 6 neighboring atoms, respectively.

The average descriptor, denoted as $(\bar{x})$, was computed using the following formula: $(\bar{x}) = \frac{1}{i}\sum_{1}^{i} x_i$, where x_i represents the value of the descriptor for one of the surrounding atoms, and i is the number of neighbors (3, 4, or 6). The maximum descriptor (x_{max}) and minimum descriptor (x_{min}) were determined as the highest and lowest values, respectively, among the descriptor values of the 3, 4, or 6 neighboring atoms. The sum descriptor (x_{sum}) was computed as

the sum of the descriptor values from the neighboring atoms. It involved adding up the values of the descriptor for the 3, 4, or 6 neighboring atoms, depending on the type of interstitial.

Descriptors:

$\bar{\chi}$ - Average Electronegativity,

$\bar{N}_e$ - Average number of valence electrons,

$\bar{I}$ - Average Ionization energies,

ε_d - Average d-band center,

$\bar{f}_d$ - Average d-band filling,

$\bar{W}_d$ - Average d-band width,

ω_{max} - Maximum Phonon band center,

$\bar{\omega}$ - Average Phonon band center,

ω_{sum} - Sum of Phonon band centers,

$\bar{r}_{Ion}$ - Average ionic radius,

$\bar{r}_{Atom}$ - Average atomic radius,

$\bar{r}_{Waals}$ - Average Van der Waals radius,

$\bar{r}_{Metal}$ - Average metallic radius,

R_{pore}^{DFT} - represents the average pore radius relative to the metallic radius of each neighboring atom. To determine this value, we used the following procedure:

1. After performing a full structural optimization using density functional theory (DFT), we identified the center point for the 3, 4, or 6 atoms surrounding each interstitial site.
2. Next, we calculated the distance between each atom and subtracted the metallic radius of the respective atom.
3. Finally, we computed the average of these distances to obtain the R_{pore}^{DFT} value, representing the average pore radius with respect to the metallic radius of each neighboring atom.

R_{pore}^{ideal} - Average pore radius for perfect structures was determined by using the following formulas:

Triangular interstitial: $R_{pore}^{ideal} = \bar{r}_{Metal} \times \left(\frac{2\sqrt{3}}{3} - 1\right) \approx \bar{r}_{Metal} \times 0.1547$

Tetrahedral interstitials (FCC): $R_{pore}^{ideal} = \bar{r}_{Metal} \times \left(\frac{\sqrt{6}}{2} - 1\right) \approx \bar{r}_{Metal} \times 0.2247$

Tetrahedral interstitials (BCC): $R_{pore}^{ideal} = \bar{r}_{Metal} \times \left(\frac{\sqrt{5}}{\sqrt{3}} - 1\right) \approx \bar{r}_{Metal} \times 0.291$

Octahedral interstitials (FCC): $R_{pore}^{ideal} = \bar{r}_{Metal} \times (\sqrt{2} - 1) \approx \bar{r}_{Metal} \times 0.4142$

Octahedral interstitials (BCC): $R_{pore}^{ideal} = \bar{r}_{Metal} \times \left(\frac{2}{\sqrt{3}} - 1\right) \approx \bar{r}_{Metal} \times 0.1547$

Mean Phonon center and mean metallic radius:

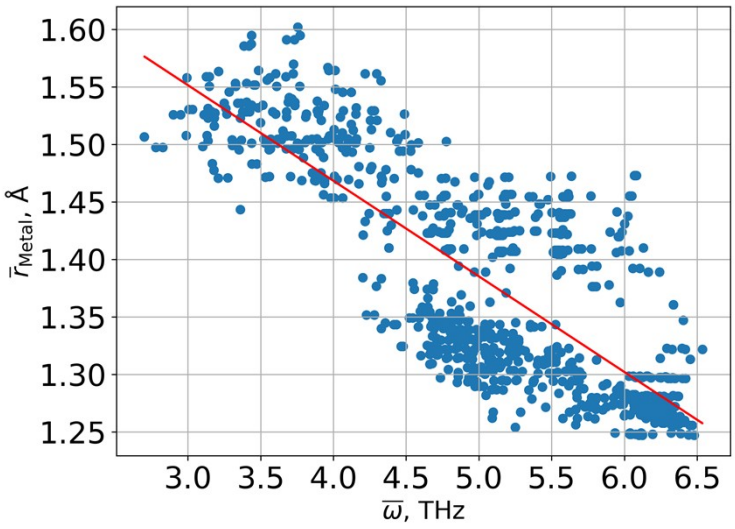

Figure S1. The correlation between the average metallic radius ($\bar{r}_{Metal}$) and the mean phonon center ($\bar{\omega}$).

SISSO optimization:

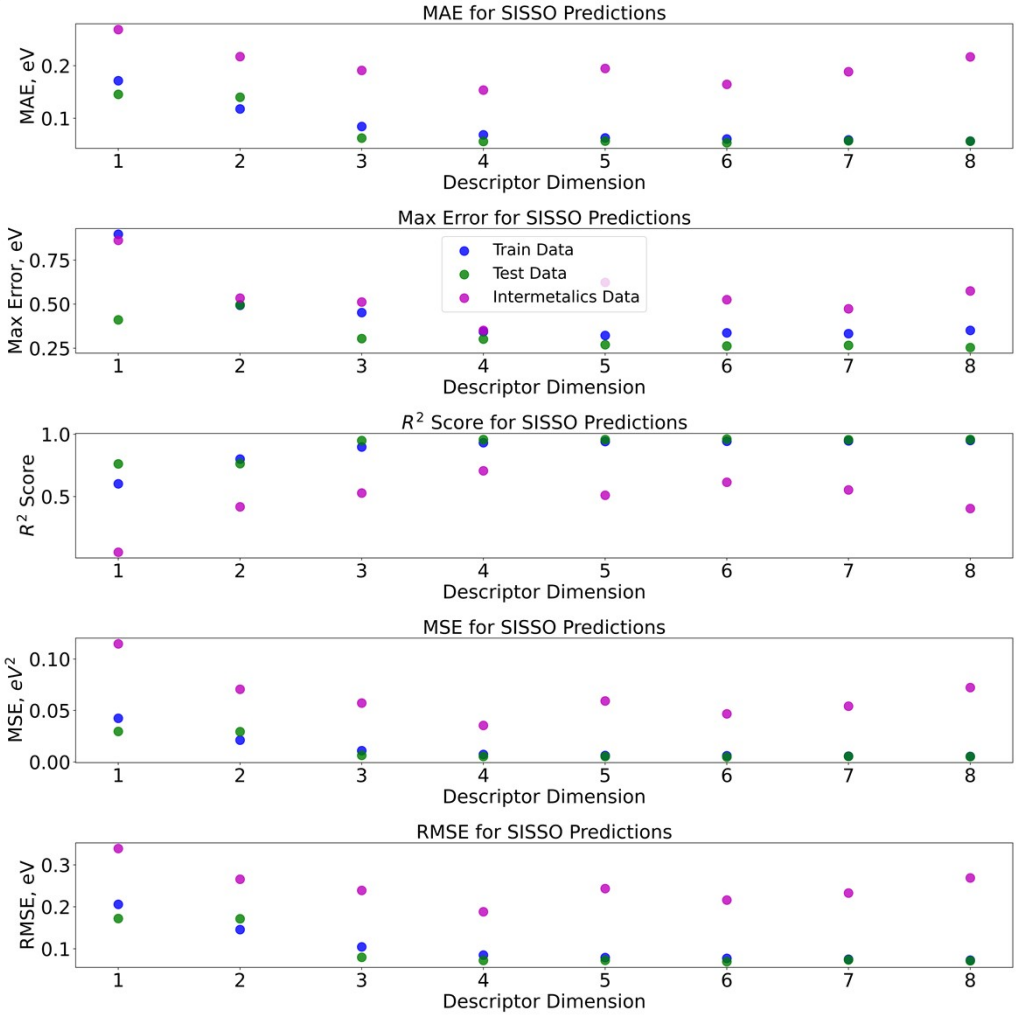

Figure S2. Performance of the $SISSO_{DFT}$ model (MAE, Max Error, R^2 score, MSE, and RMSE) as a function of descriptor dimensionality: training HEAs dataset (blue), testing HEAs dataset (green), and intermetallic compound dataset (magenta).

XGBoost Performance: Figure S3 showcases the performance of the XGBoost model in predicting hydrogen absorption energies for various scenarios. The plot displays the model's predictions for training, validation, and testing data in high-entropy alloys (HEA) interstitials, as well as for the tetrahedral interstitials of intermetallic compounds. The plot includes metrics such as the R^2 score, Mean Absolute Error (MAE), and Maximum Absolute Error (MaxE) to assess the accuracy of the predictions. The model was trained using 12 descriptors derived from the heat

matrix depicted in **Figure 2** of the main text. The results clearly demonstrate that the model performs well for interstitials in high-entropy alloys (HEA), as indicated by the metrics. However, when considering intermetallic compounds, the performance of the XGBoost model is lower compared to the $SISSO_{DFT}$ model discussed in the main text (Figure 3b).

The XGBoost model underwent optimization by adjusting the following hyperparameters: 'colsample_bytree' was set to 0.9, 'gamma' was assigned a value of 0.00, 'learning_rate' was established at 0.076, 'max_depth' was limited to 4, 'min_child_weight' was set to 5, 'n_estimators' was set to 200, 'reg_alpha' was assigned a value of 0.1, 'reg_lambda' was set to 0.1, and 'subsample' was determined as 0.9. These hyperparameters were carefully chosen and fine-tuned to optimize the performance and enhance the accuracy of the XGBoost model when evaluated on the validation dataset.

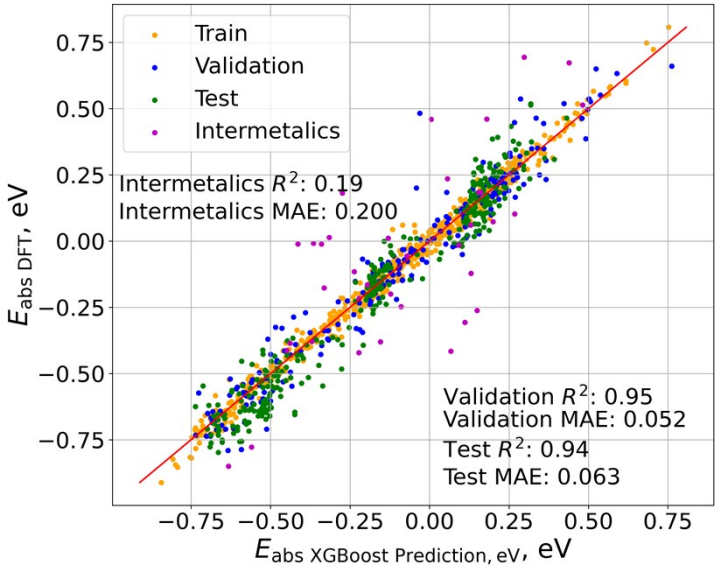

Figure S3. Comparison of DFT-calculated absorption energies with XGBoost-predicted energies for 857 interstitial sites in HEA.

The feature importance analysis results for hydrogen absorption energy prediction in HEA interstitials provide valuable insights into the significance of different features in the predictive XGBoost model. These results, shown in Figure S4, quantify the contribution of each feature to overall model performance. The feature importance analysis highlights the importance of 2 electronical structure features, the $\bar{\chi}$ and ε_{d} features, which have the highest significance in predicting hydrogen absorption energy in HEA interstitials. These features capture essential electronic structure characteristics of interstitial sites that strongly influence the absorption process. In addition to these key features, the analysis identifies those 2 structural descriptors, the R_{pore}^{DFT} and the R_{pore}^{ideal} as moderately important. These features provide information about the amount of free space available in the interstitial structure, and their values have a noticeable effect on the model's predictions.

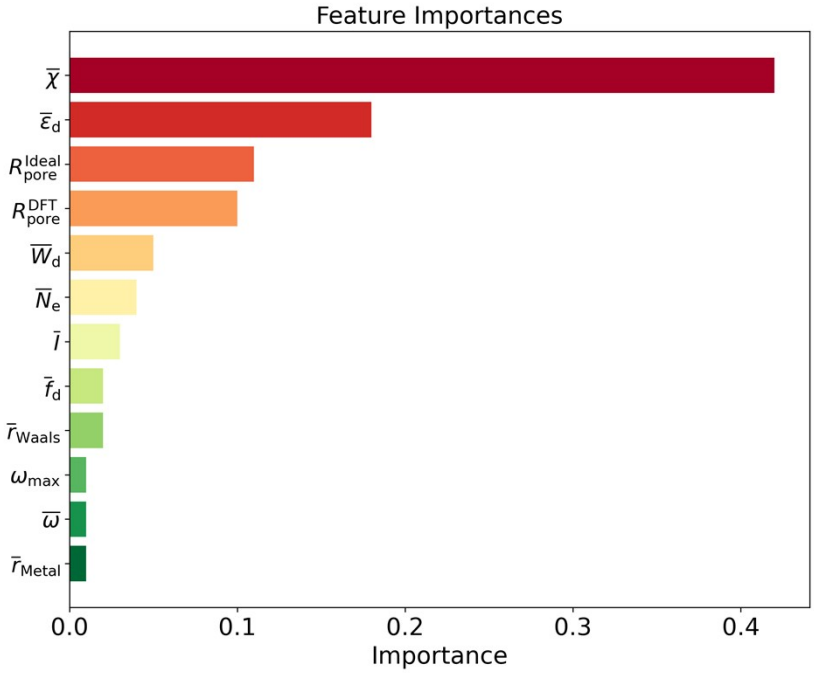

Figure S4. Feature importance analysis for the XGBoost Regression Model (Training set).

The permutation importance analysis (**Figure S5 and FigureS6**) is an alternative way to evaluate the impact of specific features on the target variable. Among the examined features, the top 5 contributors are: $\varepsilon_{\mathrm{d:}}$, R_{pore}^{DFT}, $\bar{\chi}$, $\bar{W}_{\mathrm{d}}$, R_{pore}^{ideal}.

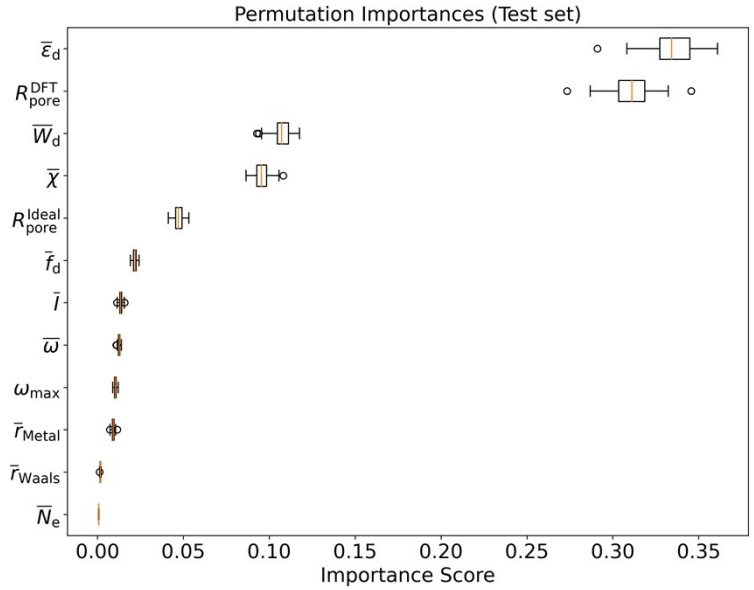

Figure S5. Permutation Importance Analysis for the XGBoost Regression Model (Testing set).

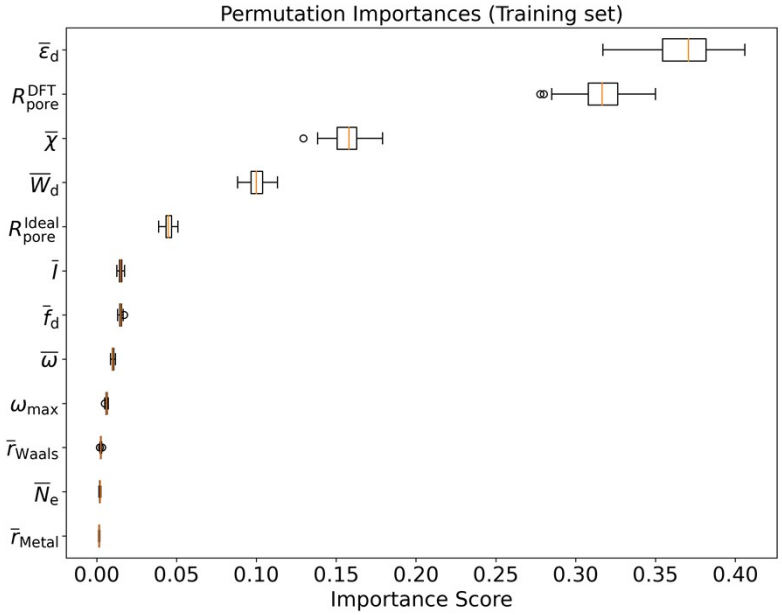

Figure S6. Permutation Importance Analysis for the XGBoost Regression Model (Training set).

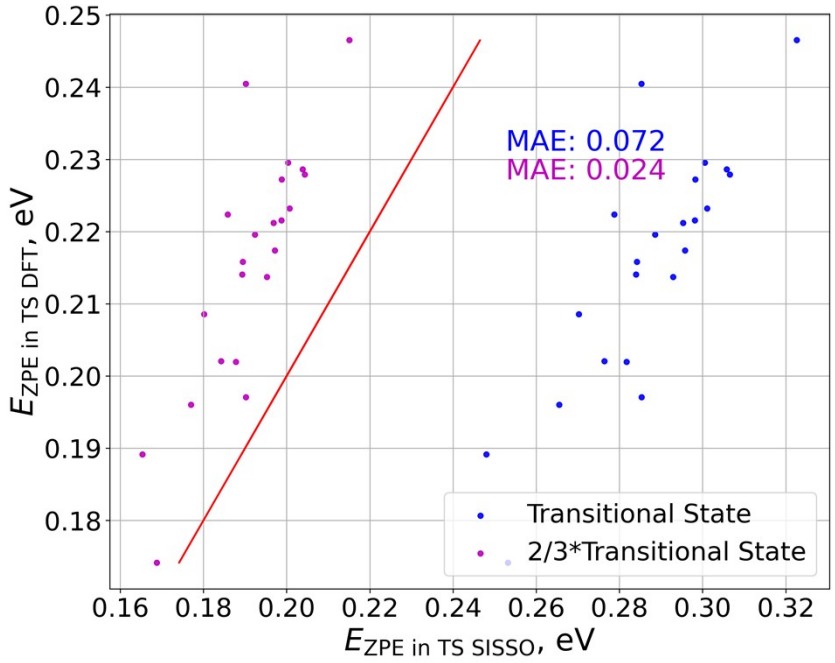

Figure S7. Performance of the $SISSO_{ZPE}$ model in predicting hydrogen zero-point energy compared to the actual values for 22 triangular interstitials (i.e., transitional state). The performance of $SISSO_{ZPE}$ model multiplied by 2/3 is given for comparison.

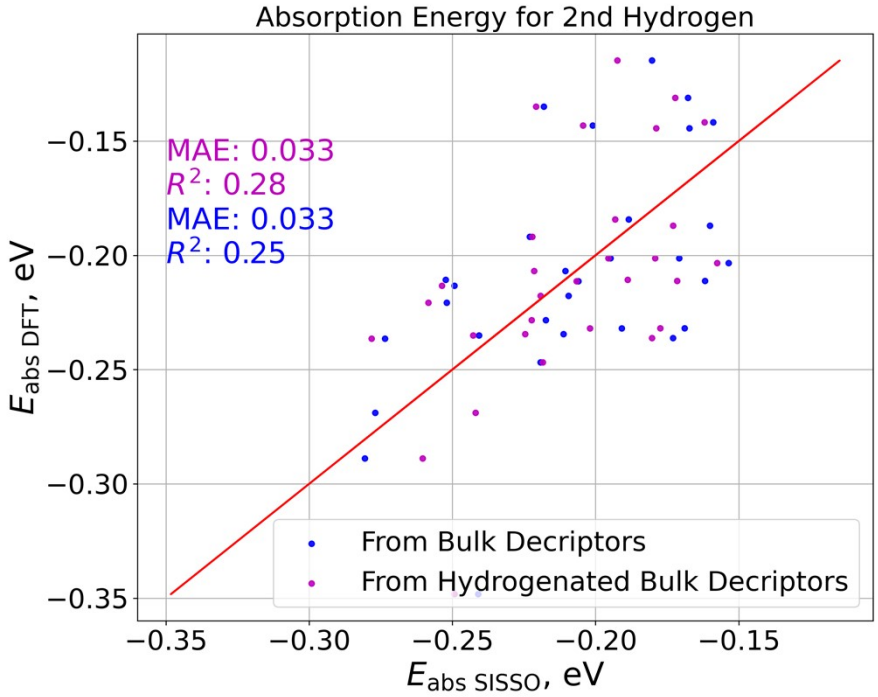

Figure S8. Performance of the $SISSO_{DFT}$ model in predicting absorption energy for the second hydrogen in CoNiV alloys. Predictions are made for a set of local descriptors of bulk structure without first hydrogen (blue) and for a set of modified descriptors calculated after the absorption of first hydrogen (magenta).

Table S1: Linear Fitted Equations and R^2 Values for Predicted Activation Energy and Absorption Energy Difference for Five FCC High-Entropy Alloys (HEA).

Alloy	Linear Fit Equation	R^2
CoNiV	-0.14x-0.19	0.18
CuNiPdPtRh	-0.41x-0.04	0.44
FeCoNiCrMn	-0.27x-0.12	0.42
CuNiFeCrMo	-0.27x-0.10	0.32
RhIrPdPtNiCu	-0.52x+0.02	0.58

Table S2: The selected number of valence electrons for each element (*Ne)* used in this work. Valence electron values were selected based on the electronic configurations in the outermost shell.

Element	*Ne*	Element	*Ne*	Element	*Ne*	Element	*Ne*
Cr	6	Ni	10	Co	9	Mn	7
Fe	8	Mo	6	V	5	Cu	11
Pd	10	Pt	10	Rh	9	Ti	4
Ir	9	Ta	5	Nb	5	Zr	4
Hf	4	Re	7				

Table S3: Performance of hydrogen absorption energy prediction for several ML machine learning models, obtained on training and two testing datasets. R^2 score and mean average error (MAE) are used as evaluation metrics.

Model	Train		Test			
	R^2	MAE, eV	R^2		MAE, eV	
			HEA	Intermetallics	HEA	Intermetallics
$SISSO_{DFT}$	0.93	0.068	0.96	0.71	0.056	0.154

$SISSO_{TABL}$	0.78	0.120	0.74	0.09	0.132	0.277
XGBoost	0.99	0.022	0.94	0.19	0.063	0.200
AdaBoost	0.92	0.066	0.91	-0.02	0.079	0.235
Support Vector Regression	0.94	0.059	0.96	0.33	0. 058	0.177
Feedforward Neural Network	0.83	0.091	0.93	-1.08	0.065	0.212

Table S4: Performance of hydrogen binding energy prediction in the transition state (e.g., triangular interstitials) for several ML models. R^2 score and mean average error (MAE) are used as evaluation metrics.

Model	Train		Test	
	R^2	MAE, eV	R^2	MAE, eV
$SISSO_{DFT}$	0.94	0.089	0.90	0.091
$SISSO_{TABL}$	0.70	0.168	0.65	0.203
XGBoost	0.96	0.036	0.85	0.117
AdaBoost	0.86	0.089	0.76	0.133
Support Vector Regression	0.88	0.091	0.94	0.079
Feedforward Neural Network	0.67	0.143	0.83	0.128