28 May 2026

Activity Descriptors of Mo2C-based Catalysts for C-OH Bond Activation

Raghavendra Meena[1], Michael J. Purcell[1], Wouter Kluijtmans[1], Han Zuilhof[1,2], Johannes H. Bitter[1], Runhai Ouyang[3], Guanna Li[1]

1. Wageningen University & Research

2. Jiaxing University

3. Sanghai University

Abstract

Deriving structure-activity relationships is crucial for designing efficient catalysts. To aid in this quest, data-driven methods such as machine learning (ML) are emerging to unravel the complicated multi-dimensional structure–activity relationships in heterogeneous catalysis. In this work, we incorporate ML tools with accurate density functional theory (DFT) energetics, electronic and geometric features to establish the structure-activity relationships in the context of designing efficient orthorhombic molybdenum carbide (Mo2C)-based catalysts for hydrodeoxygenation (HDO) reaction. Our previous studies show that C-OH activation is the rate-determining step in the HDO reaction over Mo2C (101) surface. In this work, DFT is used to obtain accurate barriers (Ea), reaction energy (#E) for the C-OH activation in HDO over the most stable and meta-stable facets of Mo2C and transition metal doped Mo2C. Then we constructed the feature database of all structures of Mo2C based catalysts and employed scikit-learn's ML models to obtain the best primary features correlating with the activity (Ea). The results show that Kernal Ridge Regression (KRR) gives the best ML model for predicting Ea of C-OH bond breaking (on unseen data) with a mean test RMSE of 0.25 eV. SHapley Additive exPlanations (SHAP) analyses for all the ML models was combined in a systematic manner to reveal that #E, d-band center, and d_{MC} are the most important features governing the activity of Mo2C-based catalysts for C-OH bond activation. Finally, SISSO was used to validate the SHAP results, and to obtain a 2-dimensional generic descriptor which is also physically interpretable to some extent. This SISSO model yielded test RMSE score of 0.14 eV, which is well below the DFT level error margin (±0.20 eV).

Keywords

molybdenum carbide, catalytic activity, descriptors, machine learning, hydrodeoxygenation

Activity Descriptors of Mo_2C-based Catalysts for C-OH Bond Activation

Raghavendra Meena,[†,‡] Michael J. Purcell,[†] Wouter Kluijtmans,[†] Han Zuilhof,[‡,¶] Johannes H. Bitter,[†] Runhai Ouyang,[§] and Guanna Li[*,†]

†*Biobased Chemistry and Technology, Wageningen University & Research, Bornse Weilanden 9, 6708WG Wageningen, The Netherlands*

‡*Laboratory of Organic Chemistry, Wageningen University & Research, Stippeneng 4, 6708 WE Wageningen, The Netherlands*

¶*College of Biological and Chemical Engineering, Jiaxing University, Jiaxing 314001, China*

§*Materials Genome Institute, Shanghai University, No. 99 Shangda Road, Shanghai 200444, China*

E-mail: guanna.li@wur.nl

Abstract

Deriving structure-activity relationships is crucial for designing efficient catalysts. To aid in this quest, data-driven methods such as machine learning (ML) are emerging to unravel the complicate multi-dimensional structure–activity relationships in heterogeneous catalysis. In this work, we incorporate ML tools with accurate density functional theory (DFT) energetics, electronic and geometric features to establish the structure-activity relationships in the context of designing efficient orthorhombic molybdenum carbide (Mo_2C)-based catalysts for hydrodeoxygenation (HDO) reaction. Our previous studies show that C-OH activation is the rate-determining step in the HDO reaction over Mo_2C (101) surface. In this work, DFT is used to obtain accurate barriers (E_a), reaction energy (ΔE) for the C-OH activation in HDO over the most stable and meta-stable facets of Mo_2C and transition metal doped Mo_2C. Then we constructed the feature database of all structures of Mo_2C based catalysts and employed scikit-learn's ML models to obtain the best primary features correlating with the activity (E_a). The results show that Kernal Ridge Regression (KRR) gives the best ML model for predicting E_a of C-OH bond breaking (on unseen data) with a mean test RMSE of 0.25 eV. SHapley Additive exPlanations (SHAP) analyses for all the ML models was combined in a systematic manner to reveal that ΔE, d-band center, and d_{MC} are the most important features governing the activity of Mo_2C-based catalysts for C-OH bond activation. Finally, SISSO was used to validate the SHAP results, and to obtain a 2-dimensional generic descriptor which is also physically interpretable to some extent. This SISSO model yielded test RMSE score of 0.14 eV, which is well below the DFT level error margin ($\pm$0.20 eV).

Introduction

Understanding structure–activity relationships of catalysts is a fundamental challenge in heterogeneous catalysis, crucial for rational catalyst design. High-throughput computational chemistry and data-driven methods have been developed to address this challenge. Most

notably, Nørskov et al. pioneered the use of the Brønsted–Evans–Polanyi (BEP) principle and linear scaling (LS) relationships to correlate energy features of metal catalysts with reaction barriers and overall catalytic performance.[1–3] These relationships have since been applied extensively in designing metal catalysts for various reactions.[4,5] However, with rapidly evolving catalyst space, complex catalysts such as transition metal oxides and transition metal carbides (TMCs) often deviate from BEP and LS relationships.[6–9] This breakdown arises because the linear, few-descriptor form is specific to site geometry and mechanism,[10,11] however the nature of the active site in heterogeneous catalysis is often too complex to be captured by one or two descriptors in a simple linear regression framework.

A more compelling case for going beyond BEP relationship in TMCs builds from our own recent work on the hydrodeoxygenation (HDO) of butyric acid over the stable orthorhombic phase of molybdenum carbide (Mo_2C).[12,13] HDO is an industrially relevant reaction for upgrading bio-derived oxygenates, and Mo_2C has attracted significant attention as an earth-abundant, tunable catalyst with platinum group metals (PGMs)-like catalytic properties for a range of such chemical reactions.[14–17] While it is known that hexagonal and cubic phase of molybdenum carbide are commonly studied, the orthorhombic phase provide more corrugated surface structures which is a much better representation of the catalyst under working conditions.[18,19] Additionally, orthorhombic phase has been shown to be experimentally more active for hydrogenation reactions compared to hexagonal phase.[20] In our previous work, butanol dissociation (C–OH bond cleavage) was identified as the rate-determining step (RDS) for HDO reaction of *n*-butyric acid to *n*-butane.[12] We subsequently showed that C–OH bond dissociation plays an equally critical role over molybdenum oxycarbide (Mo_2CO_x) catalyst.[13] Across these catalysts, microkinetic modeling consistently identified C–OH bond dissociation in butanol as among the most challenging steps, establishing the C–OH activation barrier as a reliable descriptor for overall catalytic activity in Mo carbide-based systems for HDO reaction.".

However, accurately estimating the C-OH bond activation barrier and locating the tran-

sition state by DFT calculation is particularly challenging and time-consuming due to the structural complexity of both the catalyst surfaces and the reactants. To identify the structure-activity relationship in this context, we tried to correlate the C–OH activation barrier, E_a(C–OH), with various electronic and geometric descriptors of the TMC-based catalysts, and examined the BEP relationship between E_a(C–OH) and reaction energy ΔE ($R^2 = 0.79$). While these correlations offered qualitative insights, they are insufficient for quantitative prediction. This limitation motivates us to look for more robust, data-driven approaches to identify reliable structure-activity relationships.

To this end, exponential growth in computational capacity coupled with efficient machine learning (ML) algorithms has enabled the derivation of complex yet interpretable expressions capturing structure–activity relationships in heterogeneous catalysis.[21–24] Among the available approaches, the Sure Independence Screening and Sparsifying Operator (SISSO) has emerged as a particularly powerful method for constructing interpretable descriptors and capturing non-linear correlations, even with small training sets.[25,26] SISSO has been successfully applied across heterogeneous catalytic systems. For example, Shu et al.[24] identified a 2-dimensional descriptor combining topological, electronic, and energetic features to capture structure sensitivity in predicting reaction barriers. Xu et al.[27] trained a SISSO model to predict OER intermediate adsorption enthalpies over doped RuO_2 and IrO_2 surfaces with test RMSE as low as 0.12 eV, and Lin et al.[8] derived an optimal descriptor for CO dissociation barriers in iron-based Fischer–Tropsch catalysis. Collectively, these studies demonstrate the versatility and robustness of SISSO in identifying material-specific and physically interpretable descriptors that go beyond simple linear scaling relations.

Motivated by these advances, in this work we aimed to obtain a generic descriptor correlating structure and activity for C-OH bond activation of HDO reactions on Mo_2C-based catalysts. Regarding the structural database, we covered the most stable (111) facet with 3 unique terminations and three metastable facets (101, 010, and 110) known in the literature,[28] each facet were doped with 11 transition metals (Ti, V, Cr, Fe, Co, Ni, Zr, Nb, W,

Pt, and Au). The activity of these model catalysts for the C–OH activation step of HDO reaction were evaluated. To reveal the structure-activity relationship, we first evaluated the performance of scikit-learn ML models (12 models in total), and then applied SHapley Additive exPlanations (SHAP) for a quantitative assessment of feature importance from each of the ML models.[29–31] Building on these results, we then applied SISSO[26] to derive a partially interpretable mathematical model based on primary electronic, geometric, and energy features of the model catalysts under study. The important features identified by the ML and SHAP analysis were subsequently compared with the descriptors identified by the best SISSO model. Overall, the 2-dimensional SISSO model is able to predict E_a(C–OH) for Mo_2C-based catalysts with test RMSE score of 0.14 eV, which is well below the DFT lever error margin of ±0.20 eV.

Computational Details

DFT

All DFT calculations have been performed using the Vienna Ab Initio Simulation Package (VASP) and the Perdew-Burke-Ernzerhof functional,[32,33] with Grimme's DFT-D3BJ dispersion corrections.[34,35] The kinetic energy cut-off of the plane wave basis set was set to 500 eV. The convergence criterion for energy calculation and structure relaxation was set to a self-consistent field threshold of 10^{-5} eV, and a maximum force threshold of 0.05 eV/Å per atom. Γ-centered k-meshes of the size of $6 \times 6 \times 6$ and $2 \times 2 \times 1$ were used for sampling the Brillouin zone in the case of bulk and slab models, respectively. Gaussian-type smearing with a width of 0.05 eV was applied for the electronic energy density of states. For identifying the transition states, the climbing-image nudged elastic band (CI-NEB) method was used, and frequency analysis was done on the obtained transition state to confirm that there was only one imaginary frequency along the reaction coordinate.[36] For CI-NEB calculations, the maximum force threshold of 0.10 eV/Å was implemented. Dipole corrections were applied in

the vacuum (z) direction. The bulk structure of orthorhombic Mo_2C (mp-1552), experimentally reported most-stable phase, was obtained from the Materials Project website and was fully relaxed. The optimized lattice parameters for Mo_2C: a = 4.75 Å, b = 5.23 Å, c = 6.05 Å (from experiments: a = 4.74 Å, b = 5.21 Å, c = 6.03 Å),[37] are in good agreement with the experimentally reported values. From the optimized bulk(s), we cleaved the most stable (111) surface and metastable (010), (110), and (101) surfaces. Three different terminations of (111) surfaces were chosen based on 3 different surface Mo:C composition.[28] Depending on the chosen facet we built slab models, deemed to be a big enough surface for the butanol C-OH activation reaction, with two or three stoichiometric layers of Mo_2C. For all the slab models, a vacuum distance of 15 Å was introduced in the z-direction to minimize interaction with the periodic images. The bottom one or two stoichiometric layers, depending on the chosen facet, of the supercell were fixed to reduce the computational cost of the calculations and to mimic the bulk. Further, the metal active site in each of these facets was doped with relevant metals listed in the results section. Active sites for C-OH bond dissociation were defined as the hollow/bridge sites involving the dopant and adjacent *Mo* and *C* atoms, as shown in Figure 1. Furthermore, spin-polarized calculations were performed to ensure the ground state spin configuration was located for each system. The lattice parameters, top view, and side view of the slab models used in this work are provided in Table S1 in the supporting information (SI).

The adsorption energies (E_{ads}), reaction energies (ΔE), and activation barriers (E_a) were calculated as follows:

$$E_{ads} = E_{slab+reactant} - E_{slab} - E_{reactant} \tag{1}$$

$$\Delta E = E_{product} - E_{reactant} \tag{2}$$

$$E_a = E_{transition\ state} - E_{reactant} \tag{3}$$

Here, $E_{slab+reactant}$ is the total energy of the slab with a reactant adsorbed on it, E_{slab} is the

total energy of the clean slab, $E_{reactant}$ and $E_{product}$ are the total energies of the reactants and products of each elementary reaction step, and $E_{transition\ state}$ is the total energy of the transition state (TS). The electronic structure parameters, such as dopant's *d*-band center and *d*-band filling, are derived from the density of states (DOS) and were calculated using Python's *pymatgen* package.[38]

X-band filling (X = s, p, d) was calculated as:

$$f_x = \frac{\int_{-\infty}^{Fermi} \rho(\varepsilon)}{\int_{-\infty}^{\infty} \rho(\varepsilon)} \tag{4}$$

Here, ε means energy and $\rho(\varepsilon)$ means the density of states.

The unoccupied *d*-band center was calculated as:

$$\varepsilon_{d-un} = \frac{\int_{Fermi}^{\infty} \varepsilon\rho(\varepsilon)}{\int_{-\infty}^{\infty} \rho(\varepsilon)} \tag{5}$$

SISSO

The sure independence screening and sparsifying operator (SISSO)[26] is a Fortran-based code used to efficiently extract relevant material descriptors from huge and strongly correlated feature spaces, even when only small training sets are available. The SISSO approach finds the best descriptor by combining material's features (electronic, geometric, and energy properties), identifies the most correlated features and discards the irrelevant ones, and expresses the descriptor-property relationship in the form of a mathematical function, also known as the SISSO-derived model. A generic SISSO model to predict the material's property (P^{SISSO}) can be expressed as a linear combination of N-dimensional descriptors (Φ_i):

$$P^{SISSO} = \sum_{i=0}^{N} c_i \Phi_i \tag{6}$$

Here c_i's are the fitting coefficients. N denotes the number of terms/dimensions included in the SISSO model, $N \epsilon$ (1,5). Φ_i is a descriptor engineered by combining a number of primary features (up to 6 considered). The Φ_i is defined as:

$$\Phi_i = \cup_{i=1}^{n} \hat{H}^{(m)}[\phi_1, \phi_2], \forall \phi_1, \phi_2 \in \Phi_{i-1} \tag{7}$$

Here, $\mathbf{n} \epsilon$ (1,6). ϕ_i are the primary features. $\hat{H}$ is a set of mathematical operators considered for constructing complex features by combining primary features, e.g. ϕ_1 and ϕ_2 in eqn. 7. The $\hat{H}$ in this work contains the following operators:

$$\hat{H}^{(m)} = I, +, -, *, /, exp, log, |-|, \sqrt{}, ^{-1}, ^{2}, ^{3} \tag{8}$$

The superscript m indicates that when applying $\hat{H}^{(m)}$ to primary features ϕ_1 and ϕ_2 a dimensional analysis is performed, which ensures that only physically meaningful combinations are retained, i.e., only primary features with the same unit are added or subtracted. Therefore, the complexity of a SISSO model depends on i) dimensionality: the number of linear terms in the model P^{SISSO} (eqn. 6), and ii) feature complexity: the number of operators included in Φ_i (eqn. 7).

Regression Model's Hyperparameters

Table 1 shows the XGB model and scikit-learn[29] ML models used in this study, and their corresponding hyperparameters as optimized using the grid method. 5-fold cross validation approach is used for evaluating all the ML models. The models are then ranked based on their mean test RMSE scores (based on all 5-folds). Additionally, each model's best fold in terms of test RMSE score is plotted in Figure 3. The performance plots of all the ML models based on their R^2 score is plotted in Figures S2, S3, S4, S5.

Table 1: Regression Models and their optimized hyperparameters

Regressor	**Optimized hyperparameters**
RR	$\alpha = 10.0$
XGBR	learning rate = 0.2, max depth = 3, n estimators = 50
RFR	max depth = None, min samples split = 2, n estimators = 200
ETR	max depth = 10, min samples split = 2, n estimators = 50
SVR	$C = 0.1$, $\epsilon = 0.5$, kernel = linear
KNN	n neighbors = 5
GBR	learning rate = 0.2, max depth = 5, n estimators = 200
ElasticNet	$\alpha = 0.1$, l1 ratio = 0.2
Lasso	$\alpha = 0.01$
MLP Neural Net	hidden layer sizes = (50, 50), learning rate init = 0.01
KRR	$\alpha = 1.0$, $\gamma = 0.1$, kernel = polynomial

SHAP

Built on the foundational work by Kononenko and Štrumbelj,[39,40] the SHAP (SHapley Additive exPlanations)[30,41] is a widely used tool for interpreting ML predictions. The Shapley value for a feature i represents its contribution to the prediction, calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{9}$$

where, **N** is the set of all features, **S** is a subset of features that does not include i, $f(S)$ is the model's prediction using only features in **S**, and the fraction is a weighting term ensuring fairness across all possible feature orders. The term $[f(S \cup \{i\}) - f(S)]$ is the marginal contribution of feature i when added to the subset **S**. The sum averages this marginal contribution over all possible subsets of features, so no feature is favored just because it's considered earlier or later. For any single instance x, SHAP produces an additive explanation:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{10}$$

where ϕ_0 is the baseline value (average model output over the dataset) and each ϕ_i is the contribution of feature i for that specific prediction. By simulating the effect of removing

each feature (and considering all possible combinations), SHAP assigns each feature a value representing its contribution to pushing the prediction up or down. This allows to explain both individual predictions (local explanations) and the model's general behavior across the dataset (global explanations). Finally, summary plots were obtained which include ranking features by importance and showing the direction of their effect. SHAP analysis is done for all the ML models and presented in SI.

Results and Discussion

Data Generation

The C-OH bond breaking barriers of *n*-butanol ($C_4H_9OH \rightarrow C_4H_9$+OH, the rate-determining step of HDO reaction) were calculated over the stable orthorhombic Mo_2C-based catalysts. For this purpose, the most stable facets of the orthorhombic Mo_2C catalyst reported in the literature were used,[28] as shown in Table S1, i.e., 010, 101, 110, 111 (ter1), 111 (ter2), and 111 (ter3). The Wulff construction in Figure S1 is obtained at a carbon chemical potential of -10.1 eV, i.e., at a carburization ability similar to CH_4/H_2, which is used for preparing Mo_2C catalysts. All the facets used were Mo/C mixed-terminated to provide multiple active sites for the reaction to occur, making them more active surfaces in general.[18] The active site for C-OH bond dissociation were defined as a site involving the dopant and adjacent Mo and C atoms, as shown in Figure 1.

Table 2 (**Mo** column) shows that for pure Mo_2C catalysts, 101 surface is the most active, followed by 010, 111-ter1, 111-ter3, 111-ter2, and 110 surfaces. Further, in all the facets, the active site highlighted in Figure 1 was doped with the following transition metals: Ti, V, Cr, Fe, Co, Ni, Zr, Nb, W, Pt, or Au. The C-OH activation barriers of butanol were evaluated upon heteroatom doping of the active metal site, and reported in Table 2. These results show that doping the active metal site with Zr, Nb, and W reduces the C-OH activation barrier over all facets, irrespective of the specific surface structures. This further supports

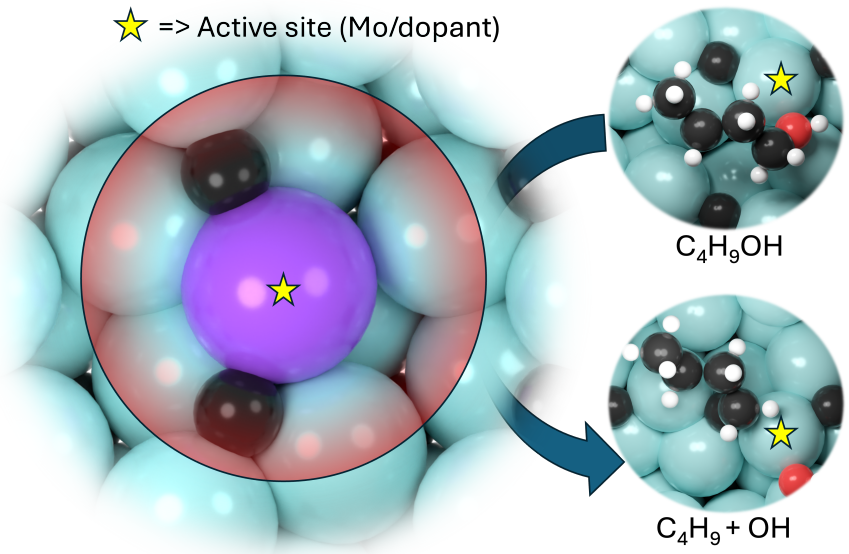


Figure 1: Butanol dissociation ($C_4H_9OH \rightarrow C_4H_9 + OH$) over the active site of Mo_2C catalyst denoted with a star. The local environment of the active is highlighted with a red circle. The active site, i.e., Mo in case of pristine Mo_2C catalyst, is doped with other transition metals and is highlighted in purple.

our previous work on butanol dissociation over Mo_2C (101) surface, where it was found that doping with Zr, Nb, and W improves the activity for C-OH bond activation.[12]

Table 2: The C-OH activation barriers as a function of Mo_2C facets and dopants (in eV).

Facet/dopant	Ti	V	Cr	Fe	Co	**Mo**(ref)	Ni	Zr	Nb	W	Pt	Au
010	0.8	0.88	0.99	1.16	0.97	0.94	1.01	0.71	0.71	0.78	1.28	N.A.
110	0.7	1.26	1.6	N.A.	N.A.	1.58	1.07	1.01	0.91	1.01	1.64	2.34
101 (Ref.[12])	0.95	0.77	1	1.32	1.63	0.83	1.59	0.54	0.58	0.78	1.6	N.A.
111-ter1	1.02	1.04	1.34	N.A.	N.A.	1.25	1.53	0.85	0.82	0.81	1.7	1.88
111-ter2	1.2	1.24	1.44	N.A.	1.94	1.54	N.A.	1.05	0.93	0.96	2.35	N.A.
111-ter3	1.23	1.21	1.46	1.9	N.A.	1.26	N.A.	0.98	0.88	0.97	2.35	2.5

*N.A. denotes configurations where a stable transition state could not be located.

Further, electronic structure analysis and geometry analysis of the active site as shown in Figure 1 is performed to obtain the potential primary features of activity in Mo_2C-based catalysts for C-OH activation. The data we obtained contain the C-OH activation barrier (E_a), reaction energy (ΔE), bond distance between the dopant (M) and the closest Carbon atom (d_{MC}), bond distance between the dopant and the closest molybdenum atom (d_{MM}),

Table 3: List of energy, electronic, and geometric features considered for ML models and SISSO analysis.

Electronic/Geometric feature	Symbol	Unit
Atomic radius	R_d	Å
Coordination number (dopant with metal)	CN_M	dimensionless
Coordination number (dopant with carbon)	CN_C	dimensionless
Coordination number (dopant)	CN_{total}	dimensionless
Common coordination number (in stable oxides)	CN_{oxides}	dimensionless
Closest dopant-metal bond distance	d_{MM}	Å
Closest dopant-carbon bond distance	d_{MC}	Å
d-band centre (up/down spin)	$\varepsilon_{d_{c\uparrow}}/\varepsilon_{d_{c\downarrow}}$	eV
d-band filling (up/down spin)	$\varepsilon_{d_{f\uparrow}}/\varepsilon_{d_{f\downarrow}}$	dimensionless
d-band skewness (up/down spin)	$\varepsilon_{d_{s\uparrow}}/\varepsilon_{d_{s\downarrow}}$	dimensionless
d-band kurtosis (up/down spin)	$\varepsilon_{d_{k\uparrow}}/\varepsilon_{d_{k\downarrow}}$	dimensionless
d-band width (up/down spin)	$\varepsilon_{d_{w\uparrow}}/\varepsilon_{d_{w\downarrow}}$	eV
Number of valence electrons	V_e	dimensionless
Pauling's electronegativity (dopant in vacuum)	χ_P	dimensionless
Electron affinity (dopant in vacuum)	EA	eV
1st ionization energy (dopant)	IP^{1st}	eV
Binding energy (dopant-oxygen) in stable oxides	BE_{M-O}	eV
Reaction energy (C-OH activation)	ΔE	eV

coordination number of the active metal site with metal or carbon in the first-shell ($CN_{M/C}$), and the atomic radius of the active metal site (R_d). Furthermore, clean surface properties such as dopant's/active metal site's d-band filling, d-band center, d-band kurtosis, d-band skewness, d-band width, electronegativity, first-ionization energy, electron affinity, and binding energy with Oxygen (BE_{M-O}) are evaluated. In total, this implies that a set containing 61 samples, each containing information about 23 primary features, were investigated (Table 2). The correlation (Pearson's) matrix for the different primary features is shown in Figure 2. It can be noted that d-band center ($R = 0.74$) and d-band filling ($R = 0.68$) have a better correlation with E_a, compared to ΔE ($R = 0.65$) with E_a. However, no single rigorous feature has a significant correlation ($R^2 \geq 0.50$) with the targeted C-OH activation barrier (E_a).

From the correlation matrix (in Figure 2) and linear regression,[42] we derive the BEP relationship between the activation energy E_a and reaction energy ΔE for all the facets and

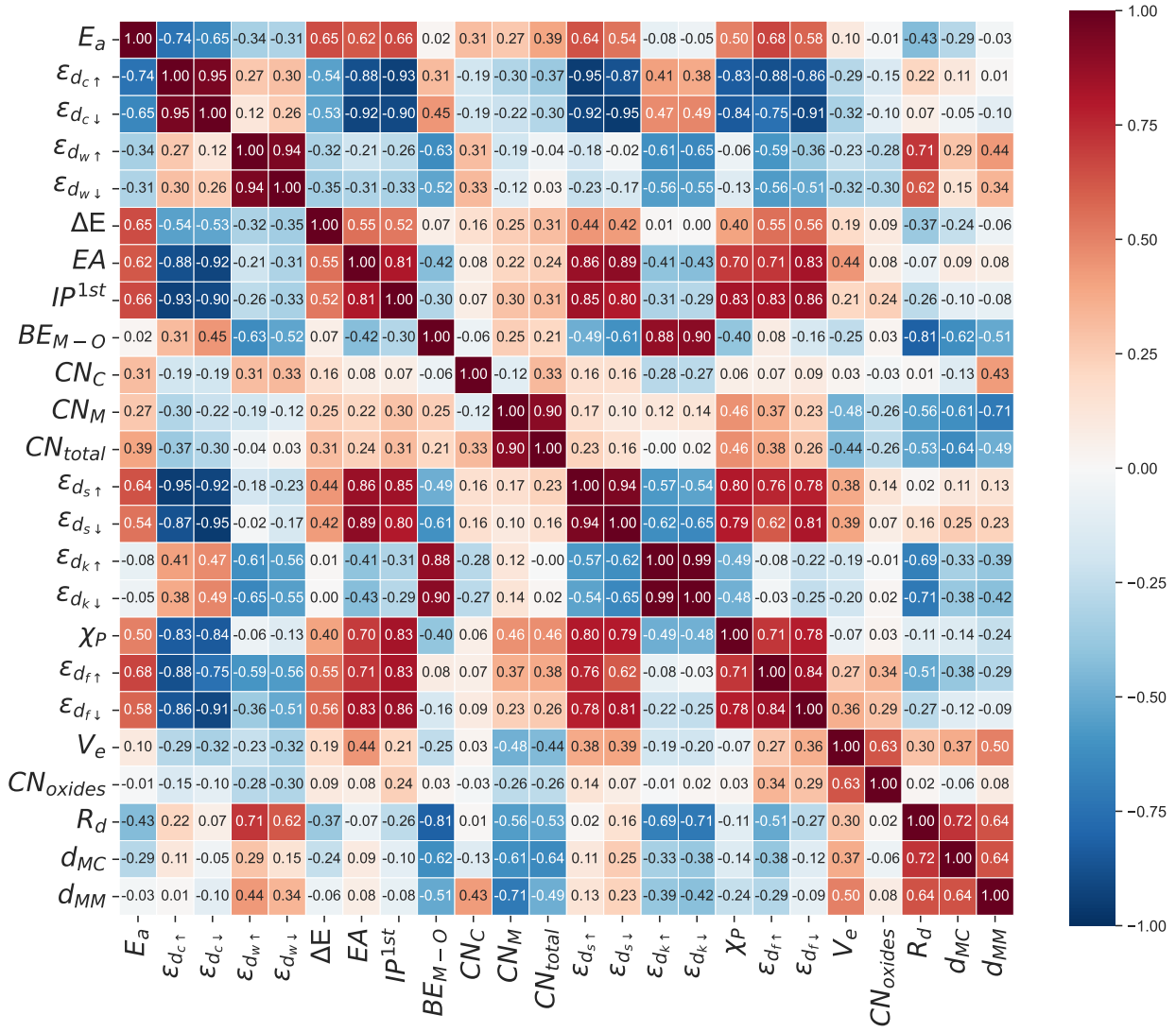


Figure 2: Correlation matrix between the activation barrier of C-OH bond activation E_a and primary energy, electronic, and geometric features.

dopants combined. The R^2 score of 0.42 ($R = 0.65$) indicates that the different facets of Mo_2C behave differently, and confirm that the BEP relationship is broken, as only facets 111-ter3 ($R^2 = 0.99$), 101 ($R^2 = 0.79$), and 111-ter1 ($R^2 = 0.80$) show a strong BEP relationship, while others facets such as 111-ter2 ($R^2 = 0.66$) and 110 ($R^2 = 0.38$) show a very weak linear BEP relationship. In contrast, the 010 facet ($R^2 = 0.04$) does not show a BEP relationship at all. Hence, in the following section more rigorous scikit-learn-based regressors were applied.

Prediction using scikit-learn based ML methods

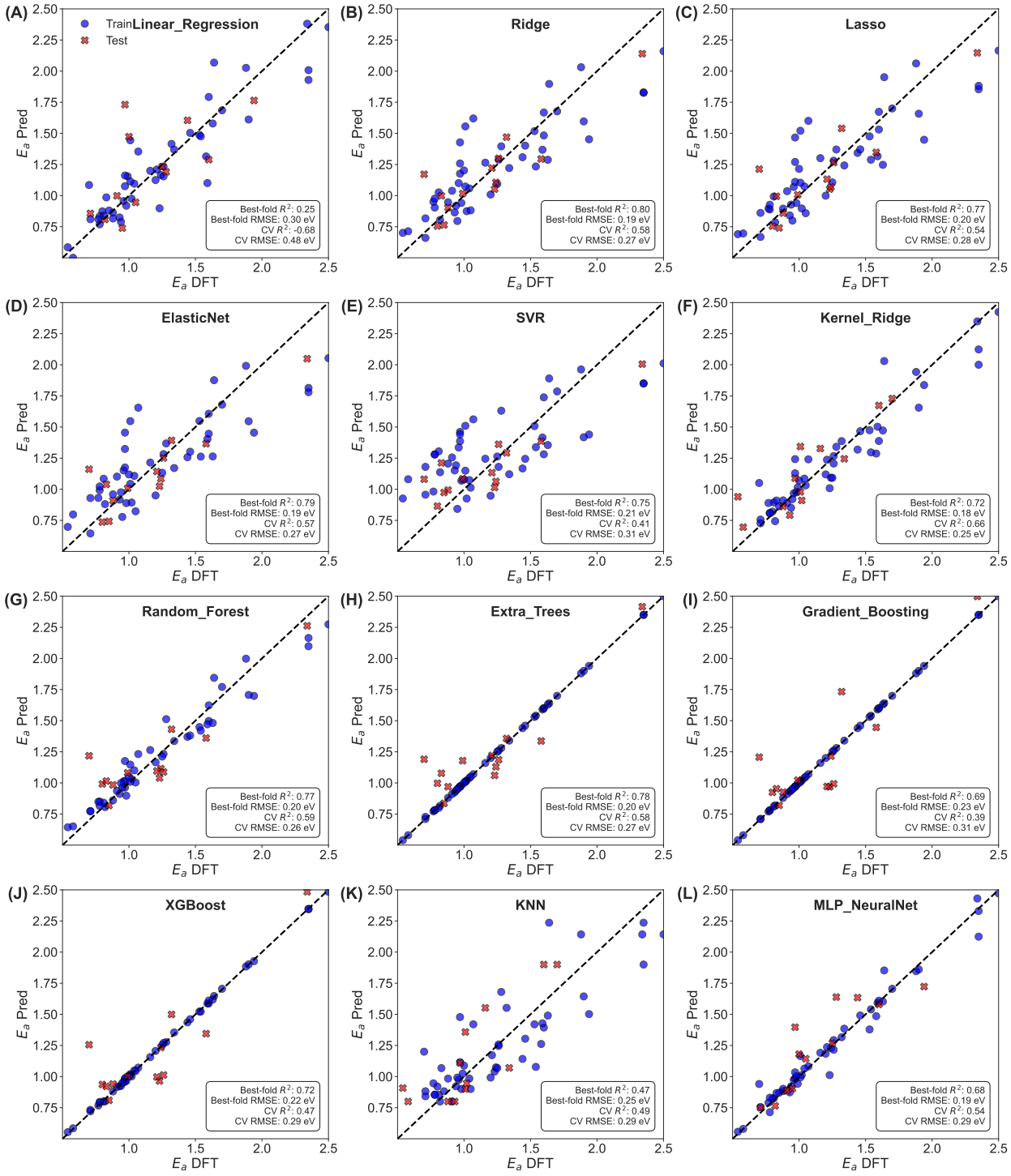


Figure 3: Test RMSE for the different ML regressors. The 5-fold CV mean test RMSE scores are reported to evaluate the robustness of each ML regressor. The best fold (with the lowest test RMSE score) obtained for each ML regressor is plotted.

The regression problem was first approached using the ML algorithms present within Python's scikit-learn package. The ML models used here include linear methods (Linear Regression (LR), Ridge Regressor (RR), Lasso, and ElasticNet), tree-based ensemble methods (Random Forest Regressor (RFR), Extra Trees Regressor (ETR), Gradient Boosting Regressor (GBR), and Extreme Gradient Boosting Regressor (XGBR)), kernel-based methods (Kernel Ridge Regressor (KRR) and Support Vector Regressor (SVR)), a distance-based method (K-Nearest Neighbors Regressor (KNN)), and a neural network (MLP). For all these regressors, the 5-fold cross-validation (CV5) approach was used. The results of these algorithms for the prediction of E_a (C-OH) are shown in Figure 3. Out of all the ML models, the kernel-based KRR emerges as the best performer ($R^2 = 0.66$, mean test RMSE = 0.25 eV). The superior performance of KRR over linear models (such as RR, $R^2 = 0.58$ and mean test RMSE = 0.27 eV) indicates that the relationship between features and DFT activation energies is complex and non-linear, requiring the kernel trick (applied to RR) to capture the underlying non-linear relationships. Further, tree-based bagging methods (RFR (R^2 0.59), ETR (R^2 0.58)) and the MLP neural network (R^2 0.54) also perform reasonably well, as they are inherently capable of modeling non-linear trends. In contrast, tree-based boosting methods (GBR, XGBR), distance-based models (KNN), and other kernel methods (SVR) perform worse ($R^2 < 0.50$), likely reflecting sensitivity to hyperparameters and the limited dataset size. Ultimately, all the ML models, except LR and GBR, outperform a simple linear BEP relationship (where $R^2 = 0.42$), confirming the necessity of capturing non-linear trends to accurately predict activity. Even though these ML models are better than linear BEP relationships, none of them breach the target DFT level error margin of ± 0.20 eV.

While the scikit-learn based ML models predict the activation barrier with reasonable accuracy, there are several limitations. First, most of these models typically do not provide an explicit mathematical expression for the prediction model, except linear models, as they are often based on decision boundaries/trees and hence are often referred to as *blackbox* models.

Second, the analysis does not incorporate dimensional considerations, making the resulting models purely mathematical rather than physically meaningful. Finally, the limited size of the dataset (61 samples) weakens the predictive performance, as such ML models generally require large amounts of data to achieve robustness. Therefore, the results of these ML models are then used for calculating the feature importance to identify the most important features governing the activity. Hence, first, the best ML model (KRR) based on the lowest mean test RMSE (0.25 eV), was selected for further analysis (Figure 4). However, since different ML models capture different relationships with geometry and electronic features, the feature importance analysis is done for all the ML models using SHapely Additive exPlanations (SHAP) analysis. These results are discussed in the next section and the analysis plots are provided in SI.

It is important to note here that an attempt was made to improve the performances of the ML models by excluding the PGMs (Pt and Au) from the dataset, as the barriers for PGMs were significantly larger than transition metals. The resulting dataset contain 52 samples. These results are provided in Figure S2-S5, which shows performance comparison of datasets with 61 samples and 52 samples. Based on a smaller sample size, it was observed that the performance of the ML models was degraded as all the models overfitted the data (evident from negative R^2 scores). These observations are expected since the ML models are known to be largely dependent on dataset size. Contrastingly, it also shows that 61 samples are enough to get meaningful correlations.

Feature Importance – SHAP analysis

Next we applied SHAP analysis which is originally developed for the biomedical field, but now applied to catalysis to gain insights into the nature of the active site by understanding the synergistic interactions of different electronic, energy, and geometric features to render catalytic activity.[43–47] The SHAP summary plots (Figure 4) generated from the best ML model (KRR) provide a detailed view of how different descriptors contribute to the model's

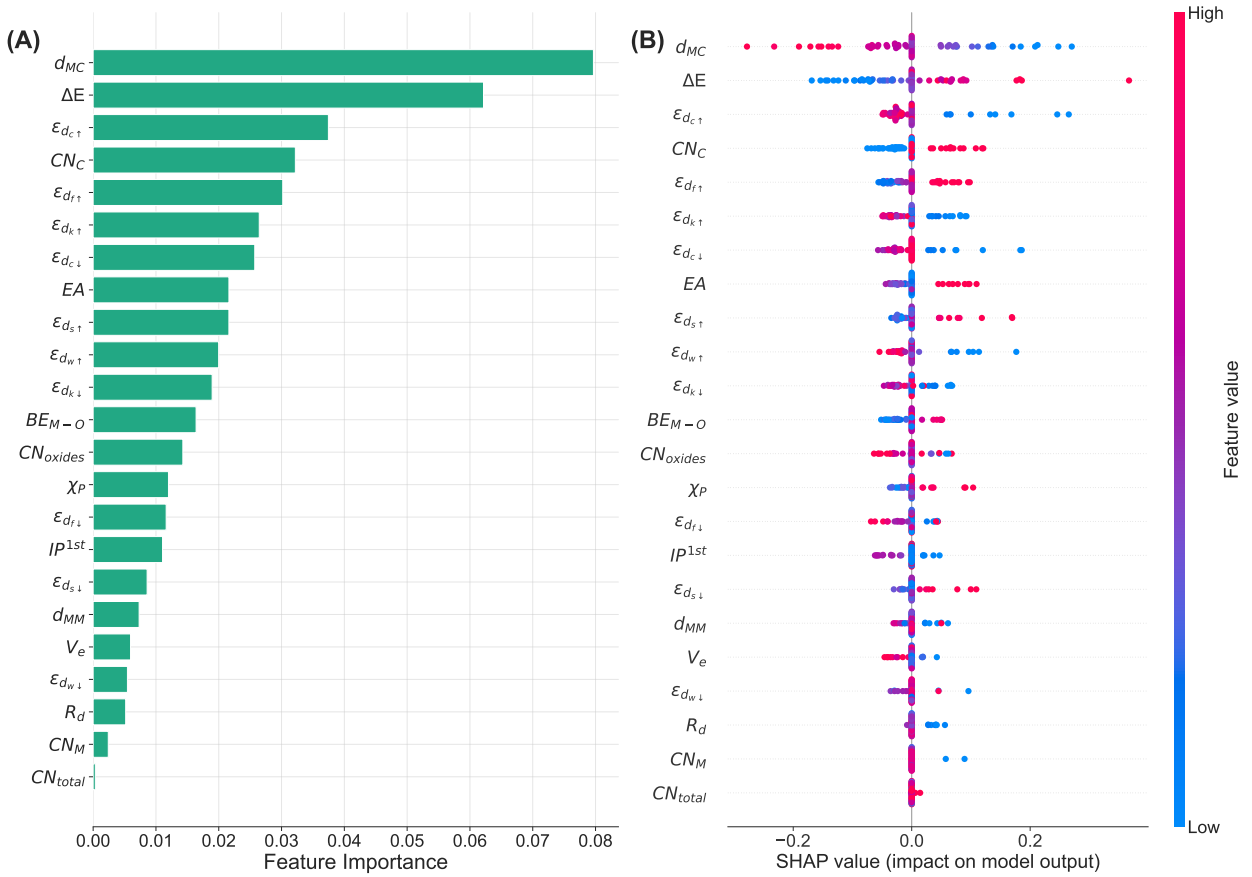


Figure 4: SHAP analysis. (A) Feature importance based on the KRR model. (B) Spread of the feature values and its influence on the activation barrier prediction.

predictions, i.e., C-OH activation barrier (E_a). Each feature is represented on the y-axis and is ranked by its average absolute SHAP value, which reflects its importance in the model (Figure 4A). In Figure 4B, the x-axis shows the SHAP values for each feature across the dataset, indicating whether a particular feature increases or decreases the model's output E_a for a given sample. Each point represents a single observation, and the color corresponds to the actual value of the feature — red for high, blue for low.

We did the similar SHAP analyses for all the ML models presented in the previous section, the results are in the SI (Figure S6 - S17). Comparing the feature importance of different ML models showed us that different ML models treat different primary features as more/less important in predicting the C–OH activation energy (E_a). In fact, features extremely important for one ML model can be at the bottom in feature importance plots in

other ML models. For example, ΔE appears to be a very important feature (ranked within top 3) in every model (Figure S7 - S17) except in LR model (Figure S6) where ΔE is in the bottom 3 of feature importance. In summary, the feature importance analysis results obtained were not consistent across different category of ML models. Therefore, to treat these inconsistencies in the feature importance, we took weighted sum of feature importance across all the ML models based on their performance (test R^2) in predicting E_a. The results are presented in Figure 5, which gives a clearer picture of the importance of each feature contributing to the output of C-OH bond activation barrier.

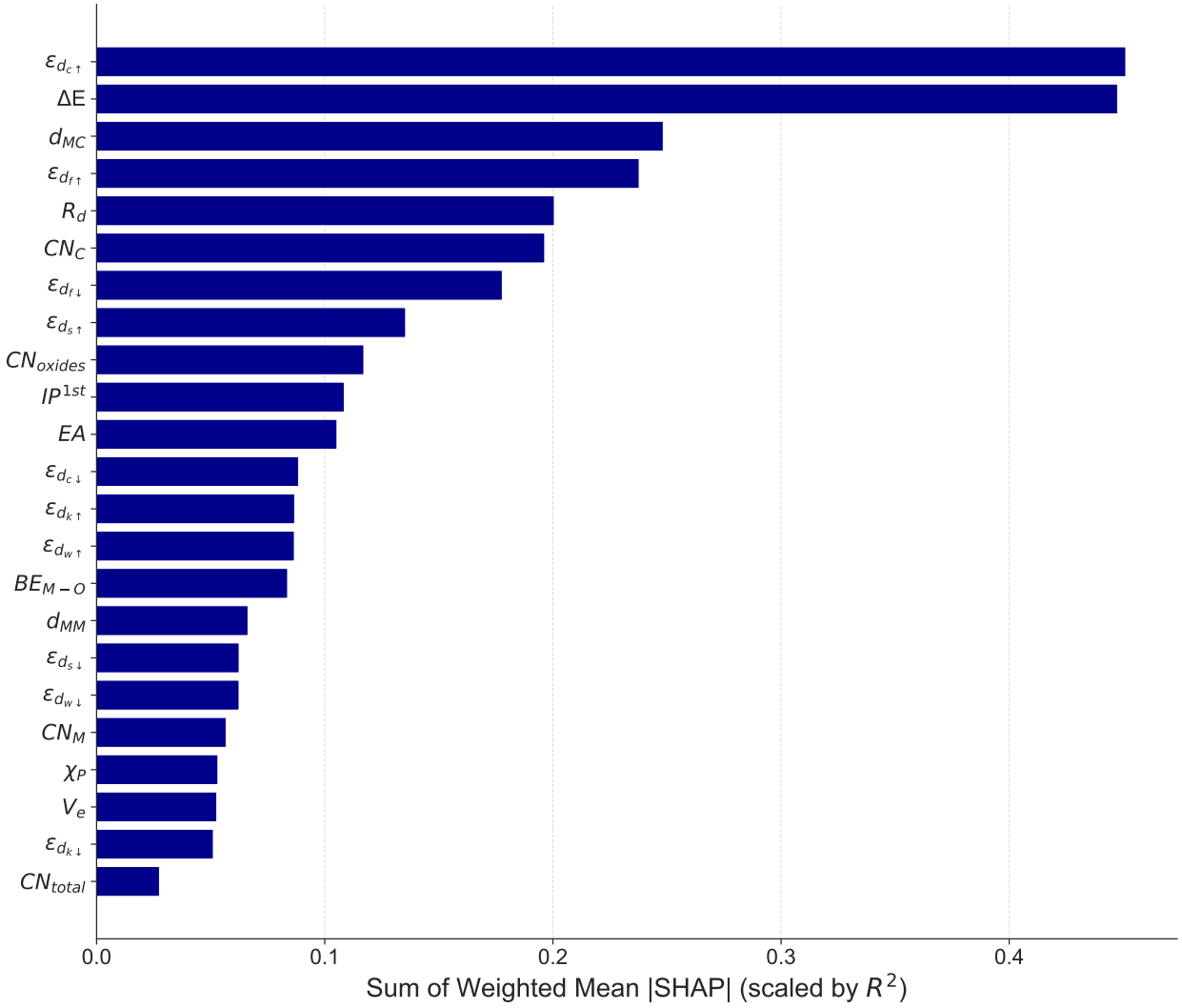


Figure 5: Combined weighted feature importance calculated using SHAP analysis, based on all the ML models used.

The combined SHAP analysis highlights the relative importance of different descriptors in predicting the C–OH activation energy (E_a). As in Figure 5, among all descriptors, the most impactful features (with mean |SHAP| > 0.3) are the up-spin *d*-band center ($\varepsilon_{d_{c\uparrow}}$) and the reaction energy (ΔE). These descriptors show a strong significant contribution to the model's predictions. Moderately important features (0.3 > mean |SHAP| > 0.15) include the dopant's distance with the closest Carbon atom (d_{MC}), the up-spin *d*-band filling ($\varepsilon_{d_{f\uparrow}}$), the dopant's atomic radius (R_d), the dopant's coordination number with surface Carbons (CN_C), and the down-spin *d*-band filling ($\varepsilon_{d_{f\downarrow}}$). Here, larger R_d values are associated with lower E_a, while the lower values $\varepsilon_{d_{f\uparrow}}$ renders lower E_a. These results are consistent with our earlier findings.[12] Similarly, lower CN_C tends to reduce E_a by creating a softer metal coordination environment, and these softer coordination environment are also a result of larger d_{MC}. These top 7 features (Figure 5) represent a mix of electronic, energy, and geometric effects that largely control the E_a. Additional descriptors, where 0.15 > mean |SHAP| > 0.1, such as the dopant's up-spin *d*-band skewness ($\varepsilon_{d_{s\uparrow}}$), dopant's CN in its most stable oxide form (CN_{oxides}), dopant's 1st Ionization potential (IP^{1st}), and dopant's electron affinity (EA) also contributes significantly to the catalytic activity. Except the d-band feature ($\varepsilon_{d_{s\uparrow}}$), the other 3 features (CN_{oxides}, IP^{1st}, and EA) are intrinsic, empirical properties of the dopant, indicating that fundamental atomic characteristics also play a role in modulating catalytic activity.

Further, there are less important features such as down-spin *d*-band center ($\varepsilon_{d_{c\downarrow}}$), up/down-spin *d*-band kurtosis (ε_{d_k}), up/down-spin *d*-band width (ε_{d_w}), binding energy with oxygen ($BE_{M\text{–}O}$), distance with the closest metal (d_{MM}), and up/down-spin *d*-band skewness (ε_{d_s}) also show noticeable influence, but markedly smaller (0.1 > mean |SHAP|). Many of these features are also directly positively or negatively correlated with the top features. Then some of the least influential descriptors include the dopant's electronegativity (χ_P), valence electron count (V_e), coordination number with nearby metals (CN_M), total coordination number (CN_{total}). These features (with mean |SHAP| < 0.1) contribute minimally, as their SHAP

values cluster around zero. These features are likely redundant or correlated with stronger descriptors, and are candidates for feature reduction.

In conclusion, the SHAP analysis not only ranked the importance of various catalysts' features, but also revealed how specific values of each descriptor alter the prediction of E_a (shown in Figure 4B) and effectively highlighted the most sensitive and important features in our dataset. Furthermore, SHAP analysis demonstrates that both electronic descriptors ($\varepsilon_{d_{c\uparrow}}$, $\varepsilon_{d_{f\uparrow}}$, $\varepsilon_{d_{f\downarrow}}$) and geometric descriptors (CN_C, d_{MC}, R_d), combined with energy descriptor ΔE, significantly influence E_a predictions, with ΔE and $\varepsilon_{d_{c\uparrow}}$ emerging as the dominant factors. It is important to note that while SHAP analysis provides a quantitative evaluation, it provides little understanding of the synergistic effects between different primary features. Therefore, in the following section we use SISSO to obtain the complex but partially interpretable descriptors of activity.

Prediction using SISSO method

The SISSO method is then employed to capture the synergistic effects of electronic and geometric features, and to develop a quantitative model for predicting the C-OH activation barriers. SISSO also offers dimensional analysis functionality, ensuring that the engineered features (descriptors) are physically interpretable. Similar to the scikit-learn based ML models, SISSO was employed on dataset with 61 samples and 52 samples excluding PGMs. The SISSO results for 61 samples are presented in SI (Figure S18). The results obtained using SISSO over 52 samples are presented in Figure 6A–D. The RMSE and MaxAE (maximum absolute error) metric is used on the test set to identify the best SISSO models. Figure 6A shows the the training and testing RMSE scores. It can be seen from Fig. 5A that SISSO models perform better as a function of dimension (denoted by D). Naturally and fundamentally, the more terms (descriptors) there are in the linear combination model, the better the model can capture the behavior of a catalyst. Similarly, each SISSO model

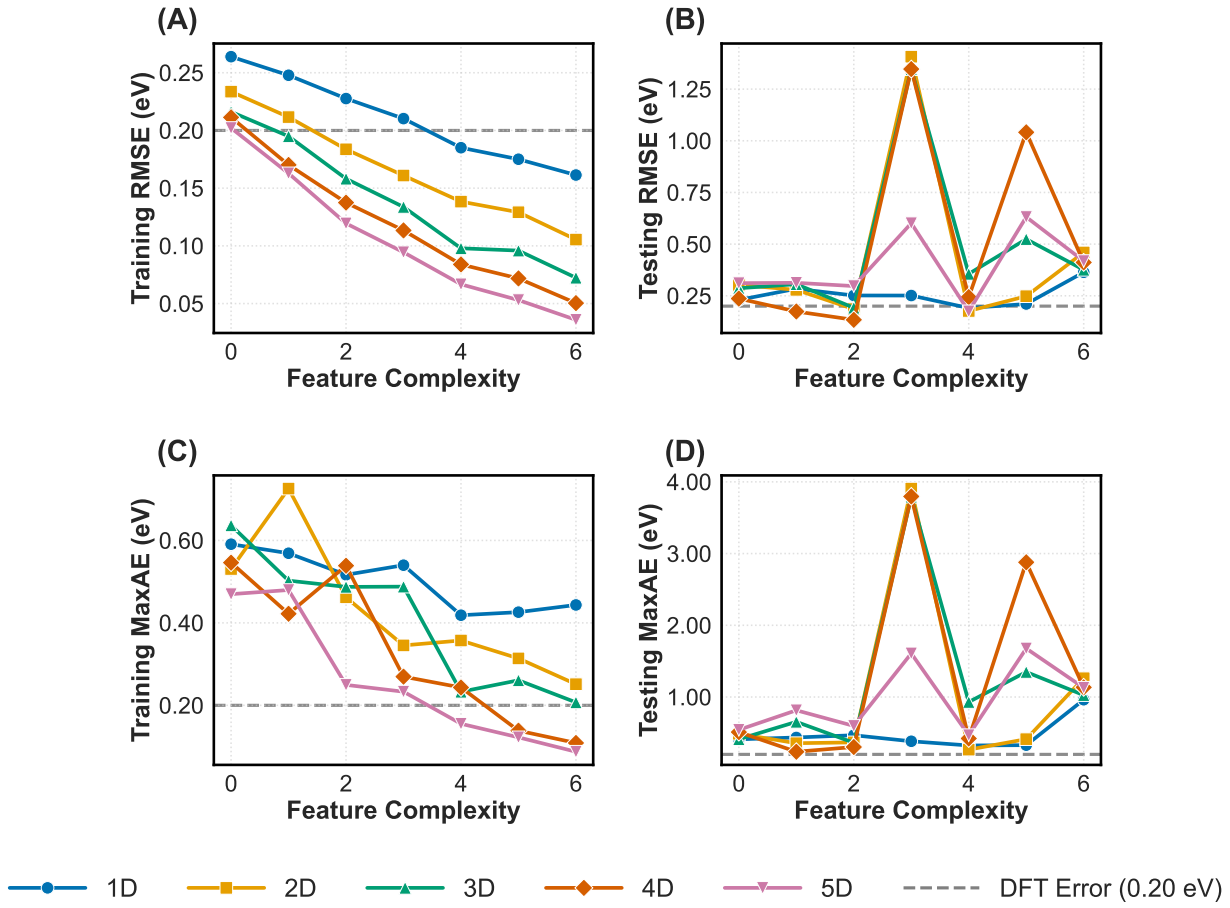


Figure 6: Training and testing RMSE using SISSO as a function of feature complexity per dimension.

improves as a function of feature complexity, as the interaction between different electronic and geometric parameters is accounted for within the newly engineered descriptor. In these models, RMSE scores as low as 0.03 eV were obtained on the training set, which is way below the DFT-level error margin of ±0.20 eV. In addition, Figure 6C shows the MaxAE obtained in the prediction of C–OH activation energy (in training set) at a given dimension and feature complexity. For all the dimensions, the MaxAE goes down as a function of feature complexity, with a few exceptions. The low values of MaxAE ($\leq 3 \times$ RMSE is considered a well-behaved model) shows the robustness of the models obtained by SISSO.

Although SISSO models perform better on seen (training) data, they do not necessarily guarantee a good prediction over unseen (test) data. This can be seen from testing RMSE

scores in Figure 6B and testing MaxAE in Figure 6D in which there is no linear/monotonous trend in test RMSE's as a function of dimension and feature complexity. In fact, in some cases, like for 4- and 5-dimension with 3-feature complexity, it yields a very poor test RMSE of 1.34 eV. At the same time, the MaxAE also blows up to $\sim$ 4.00 eV. High values of MaxAE exposes the inability of a SISSO model to treat and predict outliers and also hints toward possible overfitting. However, all of the obtained models still hold MaxAE $\leq 3 \times$ RMSE, which shows that models are well-behaved and robust.

Nevertheless, SISSO provided us with 7 models (all reported in SI) in total where the DFT-level error margin ($\pm$0.20 eV) is breached in predicting the C–OH activation energies. The raw data of the performances of these SISSO models is provided in SI. From these results, the best SISSO model is chosen based on test RMSE and test MaxAE scores. Additionally, we obtained multiple models with similar performance, so an attempt is made to strike a good balance in dimension and feature complexity. Based on such an analysis, the best SISSO model obtained is the 2D model (test RMSE = 0.18 eV) with 2 feature complexity:

$$E_a = -0.015 \cdot [\frac{BE_{M-O}}{\varepsilon_{d_{s\uparrow}}} \cdot \varepsilon_{d_{w\downarrow}}] - 0.269 \cdot (|\varepsilon_{d_{c\downarrow}} \cdot \varepsilon_{d_{f\uparrow}}| - \Delta E) + 0.984 \tag{11}$$

Naturally, the performance of the 2D SISSO model above is only improved with additions of geometry features such as dopant's disatnce with carbon(d_{MC}), as evident from the 2D model below with 4 feature complexity (test RMSE = 0.14 eV):

$$E_a = 0.066 \cdot (\frac{BE_{M-O}}{|(\varepsilon_{d_{c\downarrow}} \cdot \varepsilon_{d_{k\uparrow}}) - (\Delta E \cdot CN_{total})|}) + 2.505 \cdot \frac{e^{\varepsilon_{d_{w\downarrow}}}}{(\varepsilon_{d_{w\downarrow}})^2 \cdot d_{MC}} - 1.489 \tag{12}$$

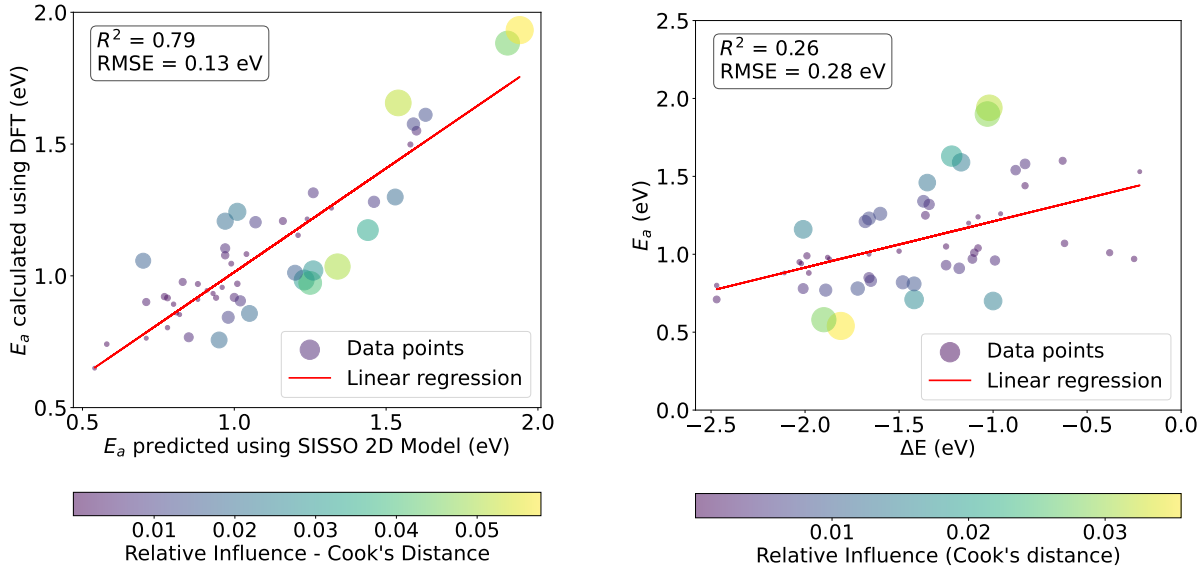
It should be noted that while the primary features in Table 3 possess varying units, the SISSO fitting coefficients (c_i) carry the necessary compensating units to ensure strict dimensional consistency across all terms in the linear combination.

It is clear from the mathematical expression in eqn. 11 and eqn. 12 that any generic descriptor derived using SISSO contains contributions combined from the primary electronic,

geometry, and energy features. Additionally, upon analyzing the expressions of 7 of the best SISSO models (reported in SI) that breach DFT-error margin, it is clear that these two analysis (SHAP and SISSO) complement each other with some exceptions. It is important to note here that fundamentally SHAP and SISSO results do not necessarily need to complement each other as they are independent methods. Nevertheless, the primary features included in eqn. 11 and eqn. 12, such as ΔE, ε_{d_c}, d_{MC}, and ε_{d_f}, also dominate the feature importance (from SHAP) in Figure 5 as well. While there are features like BE_{M-O}, ε_{d_w}, and $\mathrm{CN}_{\mathrm{total}}$ that do not appear as important from the SHAP analysis in Figure 5, but appear in eqn. 11 and eqn. 12. This discrepancy could be explained based on the collinearity of different primary features. For example, BE_{M-O} is correlated with d_{MC} (R = -0.62), as seen in Figure 2, and d_{MC} is present in the SISSO model (eqn. 12). Additionally, even though the feature importance from SHAP is low for some primary features, they can be combined with a more important feature as a scaling factor. This is evident in eqn. 11 and eqn. 12. Overall, from these results, it can be established that the reaction energy (ΔE) combined with the dopant's *d*-band features (ε_{d_c}, ε_{d_w}, ε_{d_f}) and its local environment (d_{MC}) strongly influence the activity of a given $\mathrm{Mo_2C}$-based catalyst.

As highlighted in the introduction, the purpose of using SISSO and SISSO-like multidimensional analysis tools is to obtain the least complex descriptor of activity that is physically interpretable as well (to an extent). Hence, the 2D model with 2-feature complexity obtained from SISSO is selected as the best generic descriptor. It is very important to mention here that although these results offer some physical insight, the underlying model remains challenging to interpret. Nevertheless, the obtained 2D descriptor was then used to recalculate the activation barriers for all the training and testing data, as in Figure 7a, and then compared with the popular BEP relationship (Figure 7b). It is clear that SISSO-obtained 2D descriptor performs significantly better in predicting the activation barrier ($\mathrm{R}^2 = 0.79$, RMSE = 0.13 eV) compared to the BEP relationship ($\mathrm{R}^2 = 0.26$, RMSE = 0.28 eV). Figure 7a also

shows the influence of each value on x-axis on the model's prediction of E_a. This influence is quantified by Cook's distance - a statistical measure used in regression analysis to identify influential data points, or outliers, that have a large impact on the regression model's coefficients. The greater number of such larger points (or outliers) heavily influences the model's prediction, rendering a bad BEP relationship, as is the case in Figure 7b.



(a) (b)

Figure 7: (A) Correlation of the SISSO-predicted E_a (using eqn. 12) with the DFT-calculated E_a. (B) BEP relationship using DFT-calculated E_a.

On the other hand, comparing the SISSO-2D results with the scikit-learn's ML models shows that the best scikit-learn model (KRR) does an okay job in predicting the activation barrier (R^2 = 0.66, mean RMSE = 0.25 eV). However, it fails to breach the DFT-level error margin of ±0.20 eV. Additionally, KRR does not provide an explicit mathematical expression for the underlying model, unlike SISSO. Overall, it is shown in this work that despite the small dataset, containing 52 samples, SISSO is able to produce a generic and, to some degree, physically interpretable descriptor for evaluating the C-OH activation barrier for Mo_2C-based catalysts. It is also shown (Figure S2 and S3) that scikit-learn based ML models are heavily data-dependent and fails when reducing the sample size from 61 to 52,

i.e., upon removing the PGMs data. Additionally, The SHAP feature importance (Figure 5) identifies ΔE and d-band center as dominant features. Consistently, the best SISSO model (eqn. 12) incorporates ΔE and d-band features directly, confirming that the mathematically derived descriptor aligns with the non-linear feature importance ranking.

Ultimately, Figure 8 is an annotated and modified version of Figure 1 that highlights the most important local electronic and geometric primary features, as identified with SHAP and SISSO analyses. The features in Figure 8 governs the catalytic activity of Mo_2C-based catalysts for C-OH bond activation. These features also provide a more accurate description of local environment of the active site. Additionally, it is important to acknowledge that under realistic HDO conditions, surface oxidation (Mo_2CO_x) occurs, as suggested in our previous work.[12,13] However, this study isolates the intrinsic descriptors of the carbide phase to establish a baseline. Future work will extend this descriptor analysis to oxycarbide surfaces.

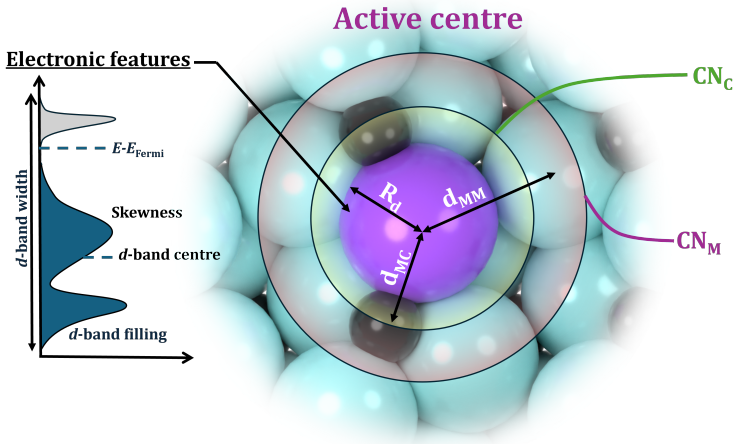


Figure 8: Scheme to highlight the majorly contributing primary descriptors controlling the activity of Mo_2C-based catalysts for C-OH bond activation.

Conclusions

This study demonstrates that for complex surface catalysts (such as TMCs) the linear scaling BEP relationship does not hold, and that machine learning (ML) tools can be effectively

used to extract physically meaningful descriptors that allow for an accurate prediction of the transition state energy. Scikit-learn's ML models generate a descriptor-based prediction of E_a that outperforms the traditional BEP relationship, but not with sufficient precision and without providing interpretable descriptors. Analysis of the all the scikit-learn's ML models by SHAP reveals that the reaction energy (ΔE) and the dopant's d-band center contribute most in predicting the E_a. Finally, SISSO is used to obtain low-dimensional physically interpretable descriptors, and to validate the findings from scikit-learn's ML models and SHAP analysis. It is found that a 2D descriptor containing contributions from the electronic features (d-band filling, d-band skewness) and geometric features (d_{MC}) can predict the activity of Mo_2C-based catalysts ($R^2 = 0.79$, RMSE = 0.13 eV), which is within DFT-level error margin of $\pm$0.20 eV and better than any of the above. Our results predict that the local environment of active metal sites plays a key role in C-OH bond activation. Further, the majorly contributing electronic (d-band center), geometric (d_{MC}), and energy (ΔE) features are also identified with two methods (SHAP and SISSO), independently. Overall, it is found that meta-stable 101 and 010 surfaces are the most active, and surface heteroatom doping with Zr, Nb, and W is a promising strategy to improve the performance of Mo_2C-based catalysts. For next step, this descriptor-activity relationship needs to be further validated for other transition metal carbide-based system for its generalized applicability.

Acknowledgement

This research was part of the Sector Plan Engineering II, funded by the Dutch Ministry of Education, Culture, and Science (OCW). The authors acknowledge the Dutch Organization for Scientific Research (NWO) for access to the Dutch national e-infrastructure (NWO-2025.019/L1, EINF-11982/L1, and EINF-7987).

Supporting Information Available

The supporting information is available free of charge. The lattice parameters, top view, and side view of the slab models used in this work are provided in the Table S1.

References

(1) Bligaard, T.; Nørskov, J.; Dahl, S.; Matthiesen, J.; Christensen, C.; Sehested, J. The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of Catalysis* **2004**, *224*, 206–217.

(2) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Physical Review Letters* **2007**, *99*, 016105.

(3) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chemistry* **2009**, *1*, 37–46.

(4) Michaelides, A.; Liu, Z.-P.; Zhang, C. J.; Alavi, A.; King, D. A.; Hu, P. Identification of General Linear Relationships between Activation Energies and Enthalpy Changes for Dissociation Reactions at Surfaces. *Journal of the American Chemical Society* **2003**, *125*, 3704–3705.

(5) Göltl, F.; Mavrikakis, M. Generalized Brønsted-Evans-Polanyi Relationships for Reactions on Metal Surfaces from Machine Learning. *ChemCatChem* **2022**, *14*.

(6) Vojvodic, A.; Calle-Vallejo, F.; Guo, W.; Wang, S.; Toftelund, A.; Studt, F.; Martínez, J. I.; Shen, J.; Man, I. C.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. K.; Abild-Pedersen, F. On the behavior of Brønsted-Evans-Polanyi relations for transition metal oxides. *The Journal of Chemical Physics* **2011**, *134*, 244509.

(7) Calle-Vallejo, F. What we talk about when we talk about breaking scaling relations. *Applied Physics Reviews* **2024**, *11*, 021305.

(8) Lin, Y.; Ushna; Lin, Q.; Wei, C.; Wang, Y.; Huang, S.; Chen, X.; Ma, X. Machine learning descriptors for CO activation on iron-based FischerTropsch catalysts. *Journal of Catalysis* **2025**, *442*, 115921.

(9) Prats, H.; Stamatakis, M. Breaking linear scaling relationships with transition metal carbides. *Catalysis Science & Technology* **2023**, *13*, 4635–4639.

(10) Logadottir, A.; Rod, T.; Nørskov, J.; Hammer, B.; Dahl, S.; Jacobsen, C. The Brønsted–Evans–Polanyi Relation and the Volcano Plot for Ammonia Synthesis over Transition Metal Catalysts. *Journal of Catalysis* **2001**, *197*, 229–231.

(11) Nørskov, J.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. Universality in Heterogeneous Catalysis. *Journal of Catalysis* **2002**, *209*, 275–278.

(12) Meena, R.; Bitter, J. H.; Zuilhof, H.; Li, G. Toward the Rational Design of More Efficient Mo2C Catalysts for Hydrodeoxygenation–Mechanism and Descriptor Identification. *ACS Catalysis* **2023**, *13*, 13446–13455.

(13) Meena, R.; Draijer, K. M.; Dam, B. v.; Zuilhof, H.; Bitter, J. H.; Li, G. Rationalizing Catalytic Performances of Mo/W-(Oxy)Carbides for Hydrodeoxygenation Reaction. *ChemCatChem* **2025**,

(14) Levy, R. B.; Boudart, M. Platinum-Like Behavior of Tungsten Carbide in Surface Catalysis. *Science* **1973**, *181*, 547–549.

(15) Iglesia, E.; Ribeiro, F. H.; Boudart, M.; Baumgartner, J. E. Synthesis, characterization, and catalytic properties of clean and oxygen-modified tungsten carbides. *Catalysis Today* **1992**, *15*, 307–337.

(16) Oyama, S. Preparation and catalytic properties of transition metal carbides and nitrides. *Catalysis Today* **1992**, *15*, 179–200.

(17) Shi, Y.; Yang, Y.; Li, Y.-W.; Jiao, H. Theoretical study about Mo 2 C(101)-catalyzed hydrodeoxygenation of butyric acid to butane for biomass conversion. *Catalysis Science & Technology* **2016**, *6*, 4923–4936.

(18) Tacey, S. A.; Jankousky, M.; Farberow, C. A. Assessing the role of surface carbon on the surface stability and reactivity of -Mo2C catalysts. *Applied Surface Science* **2022**, *593*, 153415.

(19) Politi, J. R. d. S.; Viñes, F.; Rodriguez, J. A.; Illas, F. Atomic and electronic structure of molybdenum carbide phases: bulk and low Miller-index surfaces. *Physical Chemistry Chemical Physics* **2013**, *15*, 12617–12625.

(20) Qi, K.-Z.; Wang, G.-C.; Zheng, W.-J. A first-principles study of CO hydrogenation into methane on molybdenum carbides catalysts. *Surface Science* **2013**, *614*, 53–63.

(21) Kempen, L. H. E.; Nielsen, M. J.; Andersen, M. Breaking Scaling Relations with Inverse Catalysts: A Machine Learning Exploration of Trends in CO2-to-Formate Energy Barriers. *ACS Catalysis* **2025**, 17635–17644.

(22) Cheula, R.; Andersen, M. Transition States Energies from Machine Learning: An Application to Reverse Water–Gas Shift on Single-Atom Alloys. *ACS Catalysis* **2025**, *15*, 11377–11388.

(23) Xin, H.; Mou, T.; Pillai, H. S.; Wang, S.-H.; Huang, Y. Interpretable Machine Learning for Catalytic Materials Design toward Sustainability. *Accounts of Materials Research* **2024**, *5*, 22–34.

(24) Shu, W.; Li, J.; Liu, J.-X.; Zhu, C.; Wang, T.; Feng, L.; Ouyang, R.; Li, W.-X. Structure Sensitivity of Metal Catalysts Revealed by Interpretable Machine Learning and First-

Principles Calculations. *Journal of the American Chemical Society* **2024**, *146*, 8737–8745.

(25) Ghiringhelli, L. M.; Vybiral, J.; Ahmetcik, E.; Ouyang, R.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Learning physical descriptors for materials science by compressed sensing. *New Journal of Physics* **2017**, *19*, 023017.

(26) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2018**, *2*, 083802.

(27) Xu, W.; Andersen, M.; Reuter, K. Data-Driven Descriptor Engineering and Refined Scaling Relations for Predicting Transition Metal Oxide Reactivity. *ACS Catalysis* **2021**, *11*, 734–742.

(28) Wang, T.; Luo, Q.; Li, Y.-W.; Wang, J.; Beller, M.; Jiao, H. Stable surface terminations of orthorhombic Mo2C catalysts and their CO activation mechanisms. *Applied Catalysis A: General* **2014**, *478*, 146–156.

(29) Bisong, E. Building Machine Learning and Deep Learning Models on Google Cloud Platform, A Comprehensive Guide for Beginners. **2019**, 215–229.

(30) Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman, S.-F.; Kim, J.; Lee, S.-I. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2018**, *2*, 749–760.

(31) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2020**, *2*, 56–67.

(32) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6*, 15–50.

(33) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54*, 11169–11186.

(34) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(35) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

(36) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics* **2000**, *113*, 9901–9904.

(37) Parthé, E.; Sadogopan, V. The structure of dimolybdenum carbide by neutron diffraction technique. *Acta Crystallographica* **1963**, *16*, 202–205.

(38) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.

(39) Štrumbelj, E.; Kononenko, I. Adaptive and Natural Computing Algorithms, 10th International Conference, ICANNGA 2011, Ljubljana, Slovenia, April 14-16, 2011, Proceedings, Part II. **2011**, 21–30.

(40) Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* **2014**, *41*, 647–665.

(41) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**,

(42) Marill, K. A. Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. *Academic Emergency Medicine* **2004**, *11*, 94–102.

(43) Mine, S.; Takao, M.; Yamaguchi, T.; Toyao, T.; Maeno, Z.; Siddiki, S. M. A. H.; Takakusagi, S.; Shimizu, K.; Takigawa, I. Analysis of Updated Literature Data up to 2019 on the Oxidative Coupling of Methane Using an Extrapolative Machine-Learning Method to Identify Novel Catalysts. *ChemCatChem* **2021**, *13*, 3636–3655.

(44) Li, Z. et al. Machine learning-assisted Ru-N bond regulation for ammonia synthesis. *Nature Communications* **2025**, *16*, 7818.

(45) Vidal-Lopez, A.; Mahringer, J.; Comas-Vives, A. Key Descriptors of Single-Atom Catalysts Supported on MXenes (Mo2C, Ti2C) Determining CO2 Activation. *The Journal of Physical Chemistry C* **2025**, *129*, 8556–8569.

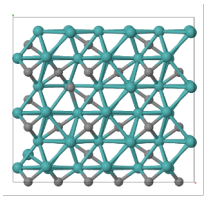
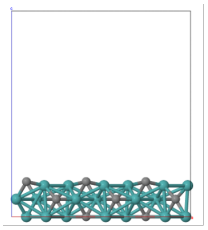
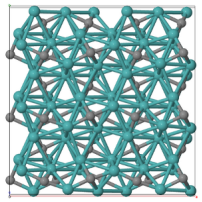
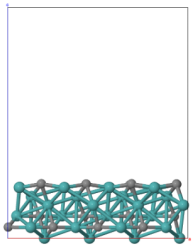
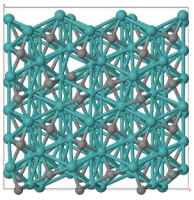
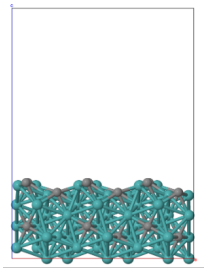
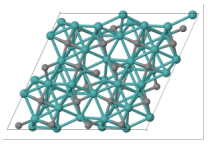
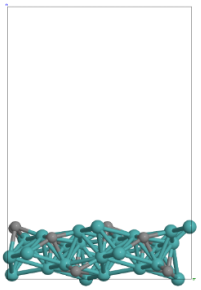
(46) Praveen, C. S.; Comas-Vives, A. Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces. *ChemCatChem* **2020**, *12*, 4611–4617.

(47) Zhao, S.; Mine, S.; Wang, G.; Zhang, W.; Fakir, A. A. E.; Yang, B.; Qin, Z.; Dostagir, N. H. M.; Matsushita, K.; Takigawa, I.; Shimizu, K.-i.; Toyao, T. Development of highly active catalysts for low-temperature CO2 hydrogenation to methanol using machine learning approach. **2025**,

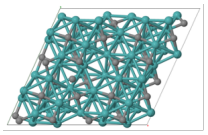
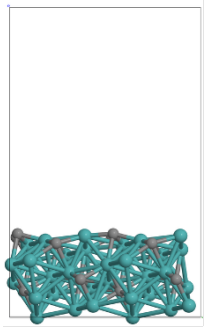
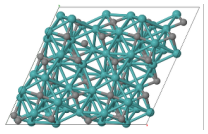
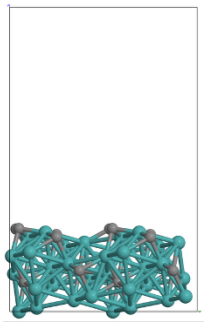
Supporting Information Available

Lattice parameters of chosen surface slabs

Table S1: Lattice parameters and surface representations of orthorhombic Mo_2C facets studied.

Facet	a **(Å)**	b **(Å)**	c **(Å)**	α **(°)**	β **(°)**	γ **(°)**	**Top view**	**Side view**
(010)	15.575	14.119	18.010	90	90	90		
(101)	15.390	15.699	19.682	90	90	90		
(110)	15.575	15.272	21.508	90	90	90		
(111)-ter1	15.272	14.014	20.805	90	90	65.548		

Continued on next page

Facet	a **(Å)**	b **(Å)**	c **(Å)**	α (°)	β (°)	γ (°)	**Top view**	**Side view**
(111)-ter2	15.272	14.014	20.805	90	90	65.548		
(111)-ter3	15.272	14.014	20.805	90	90	65.548		

Wuff construction

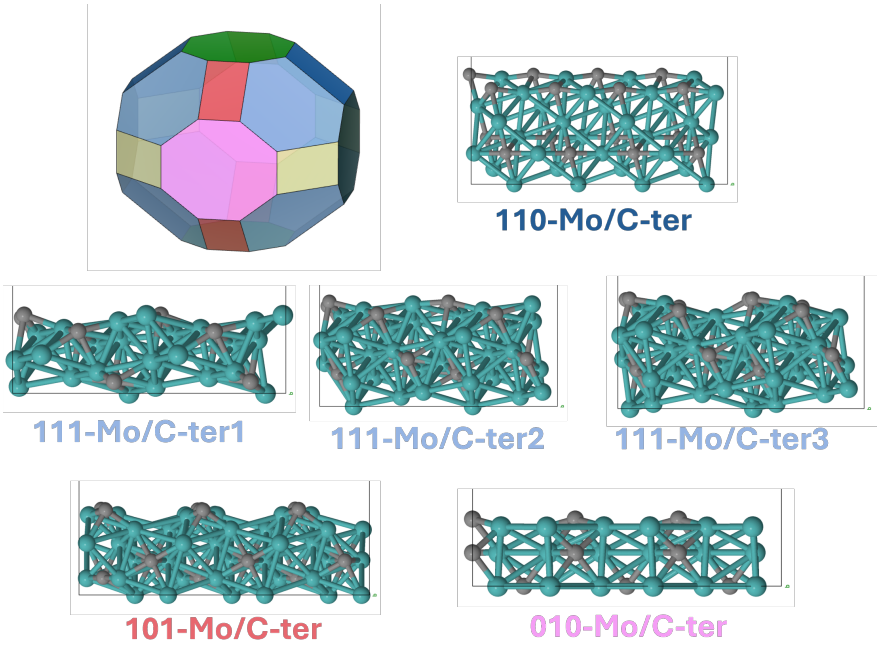


Figure S1: The Wulff construction obtained at a carbon chemical potential of -10.1 eV, i.e., at a carburization ability similar to CH_4/H_2, which is used for preparing Mo_2C catalysts.

ML model's performance comparison over 61 samples vs 52 samples

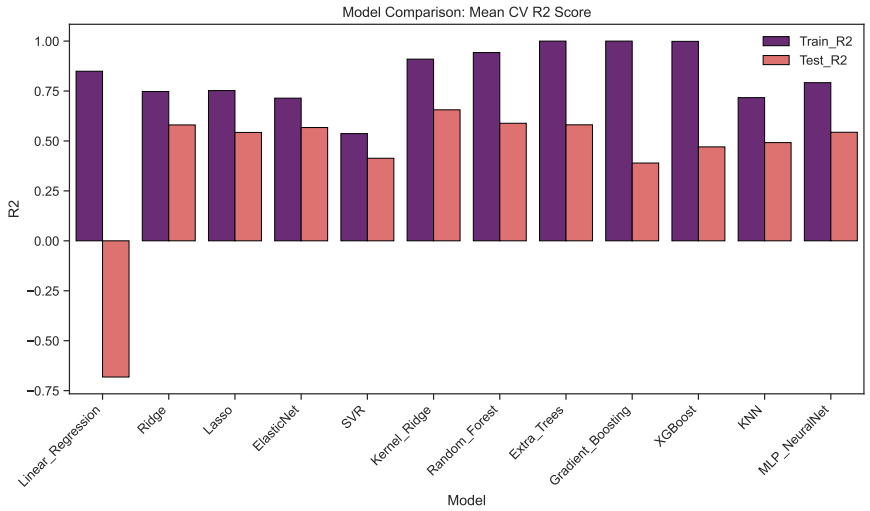


Figure S2: 61 sample size. 5 fold CV mean R^2 scores. The negative R^2 score for LR imply overfitting.

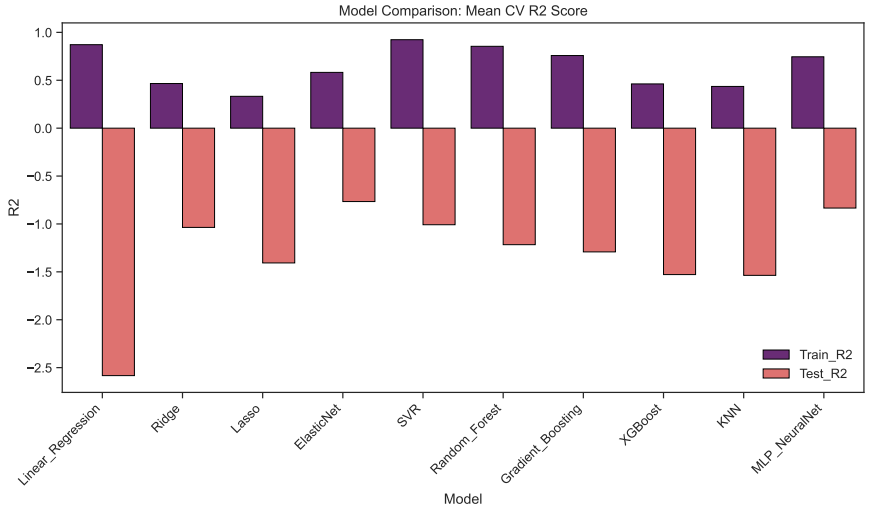


Figure S3: 52 sample size. 5 fold CV mean R^2 scores. The negative R^2 scores imply overfitting.

ML model's performance comparison over 61 samples vs 52 samples

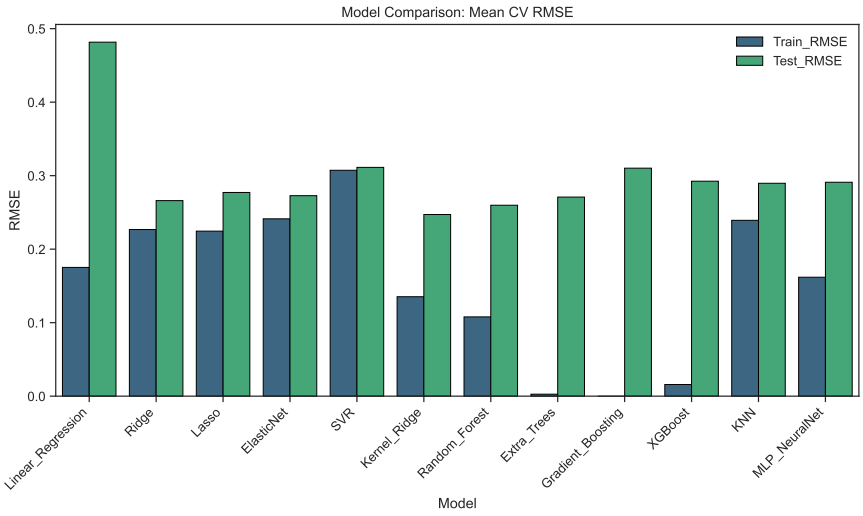


Figure S4: 61 sample size. 5 fold CV mean RMSE scores.

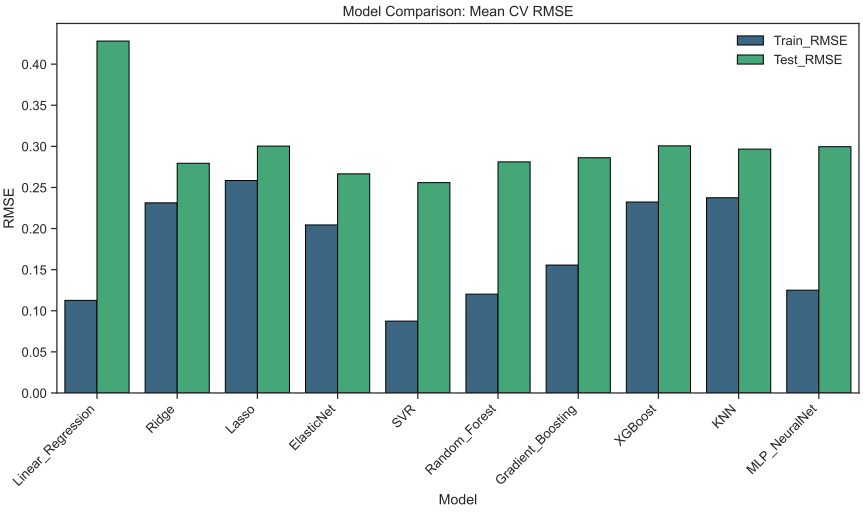


Figure S5: 52 sample size. 5 fold CV mean RMSE scores.

SHAP analyses for different ML models

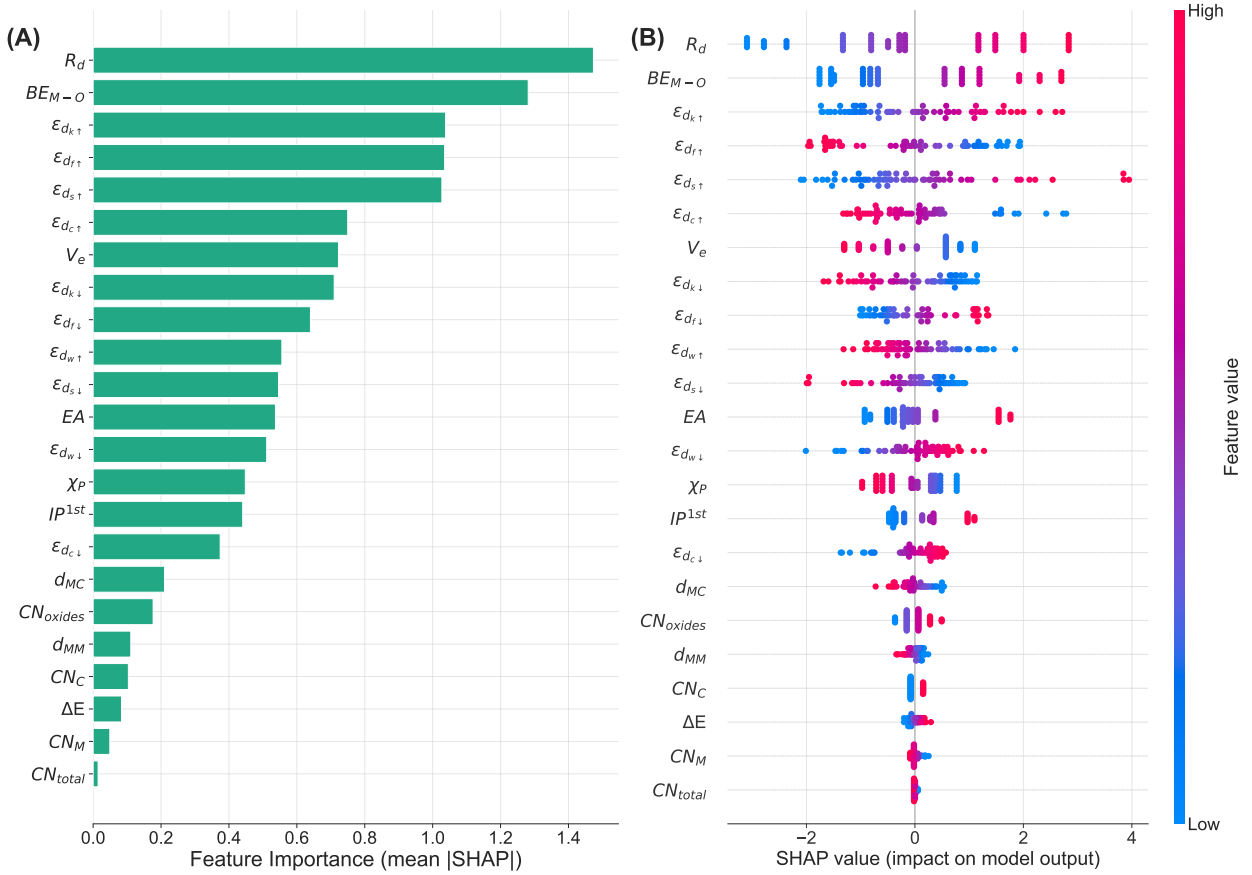


Figure S6: SHAP analysis of linear regression model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

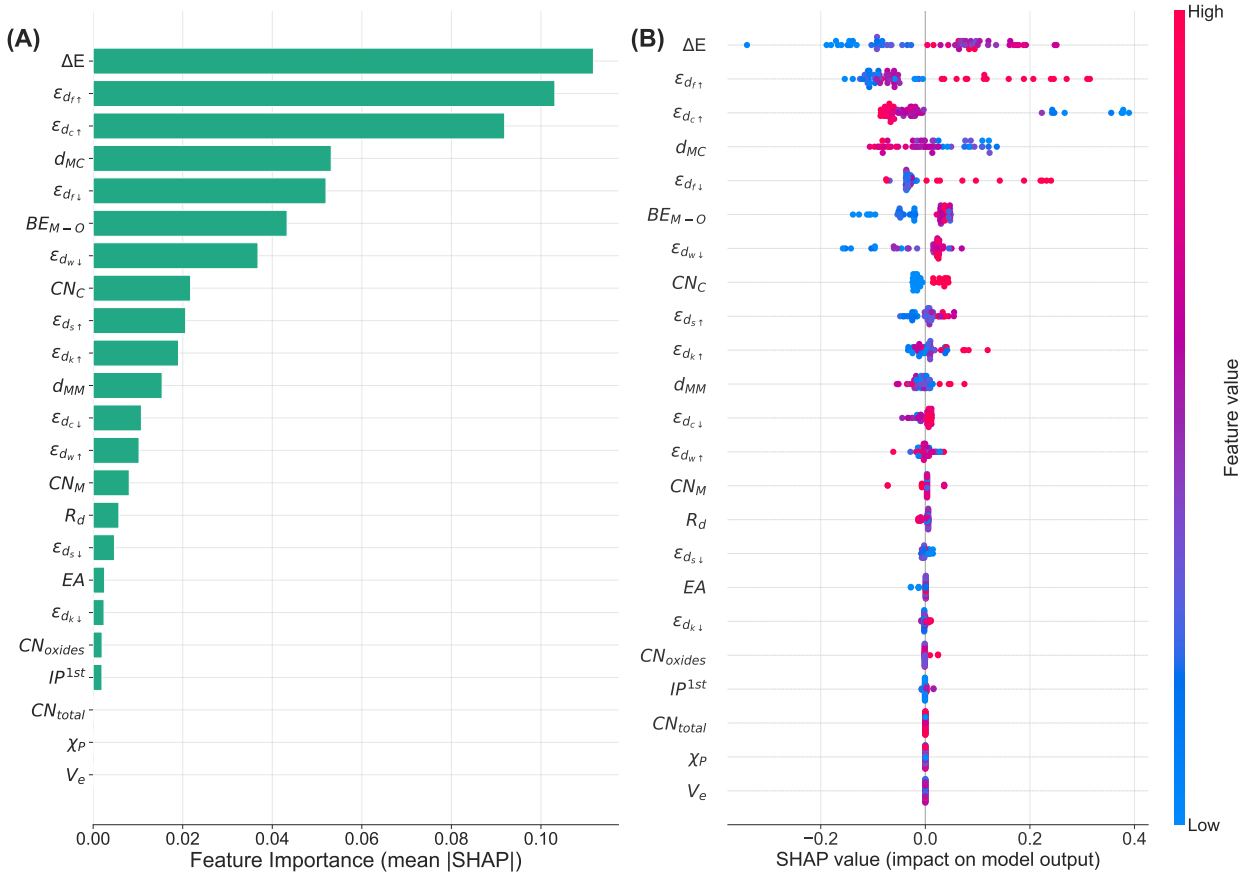


Figure S7: SHAP analysis of XGB ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

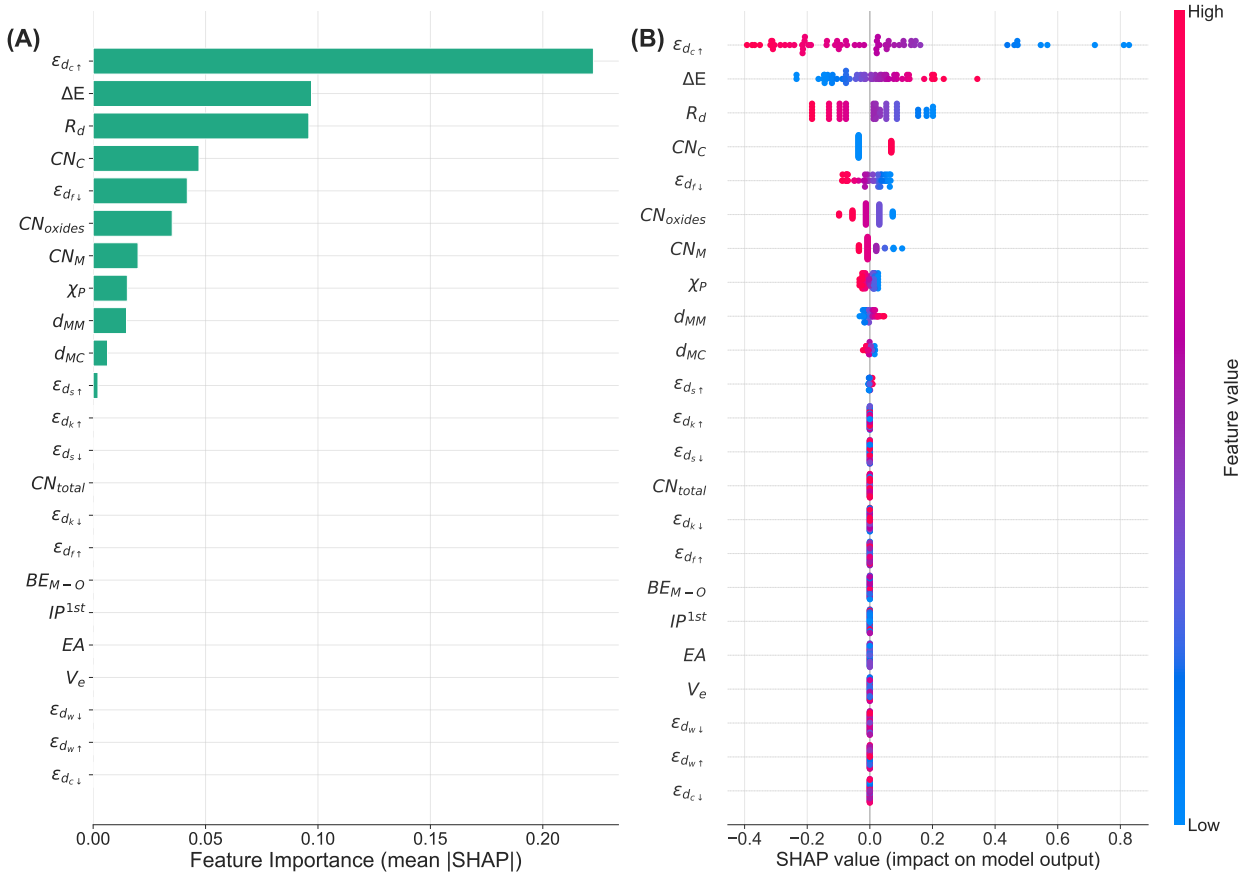


Figure S8: SHAP analysis of lasso regression model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

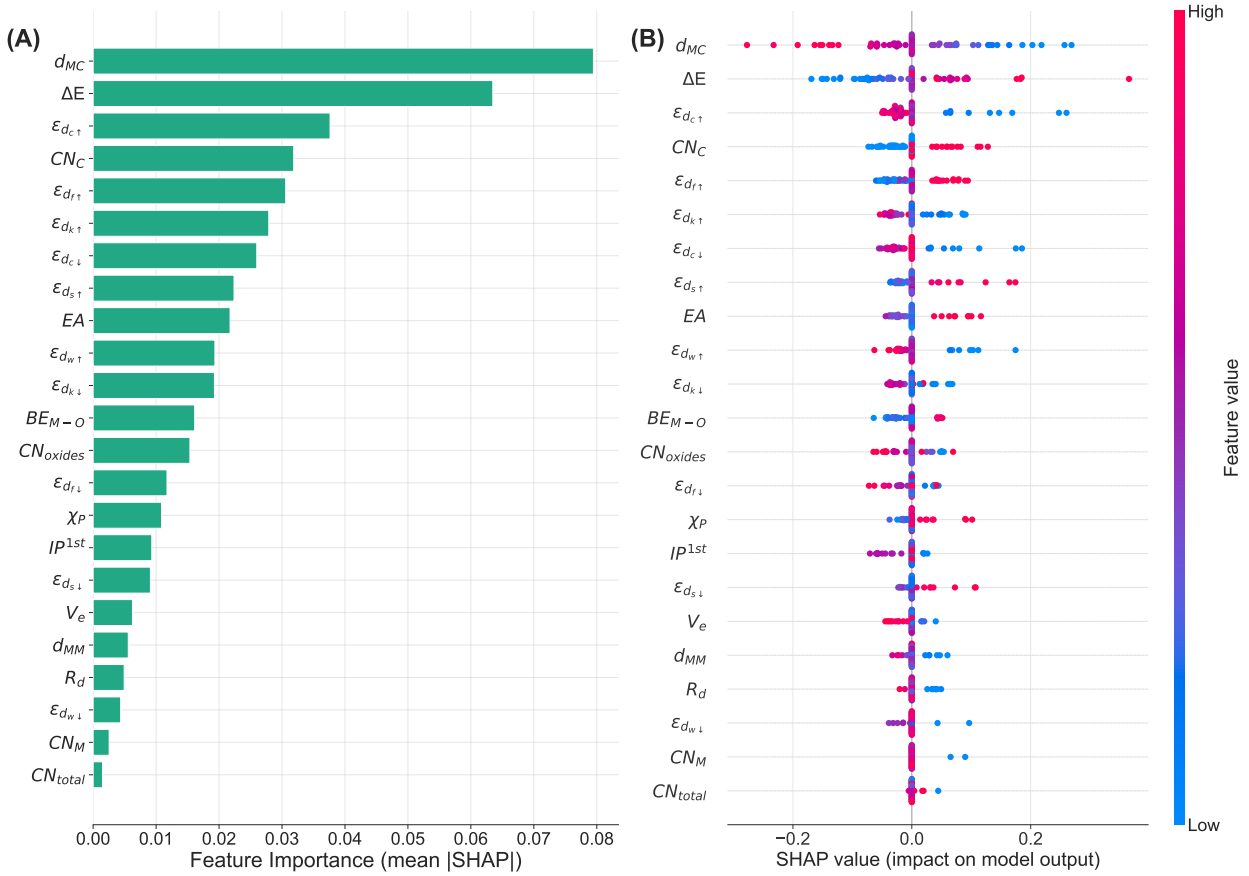


Figure S9: SHAP analysis of KRR ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

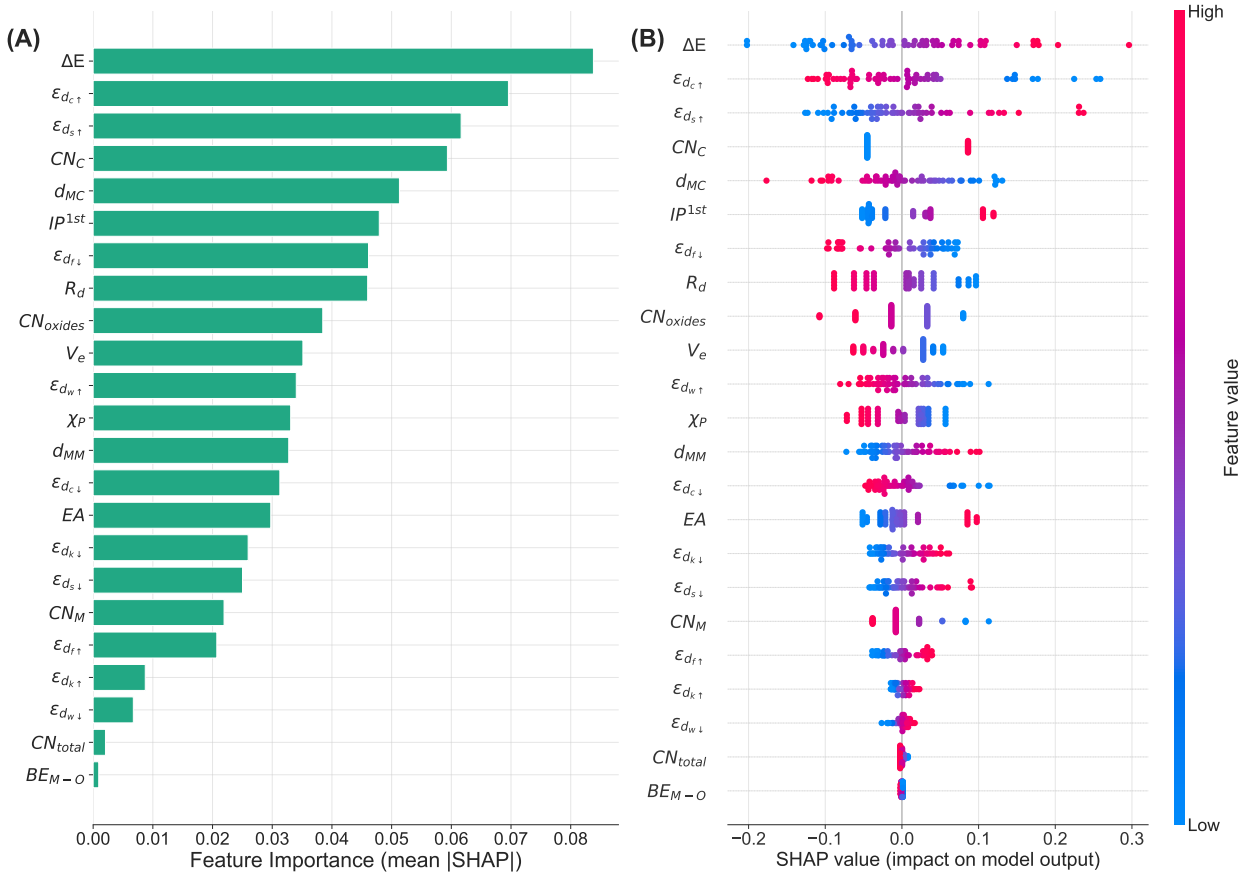


Figure S10: SHAP analysis of ridge regression model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

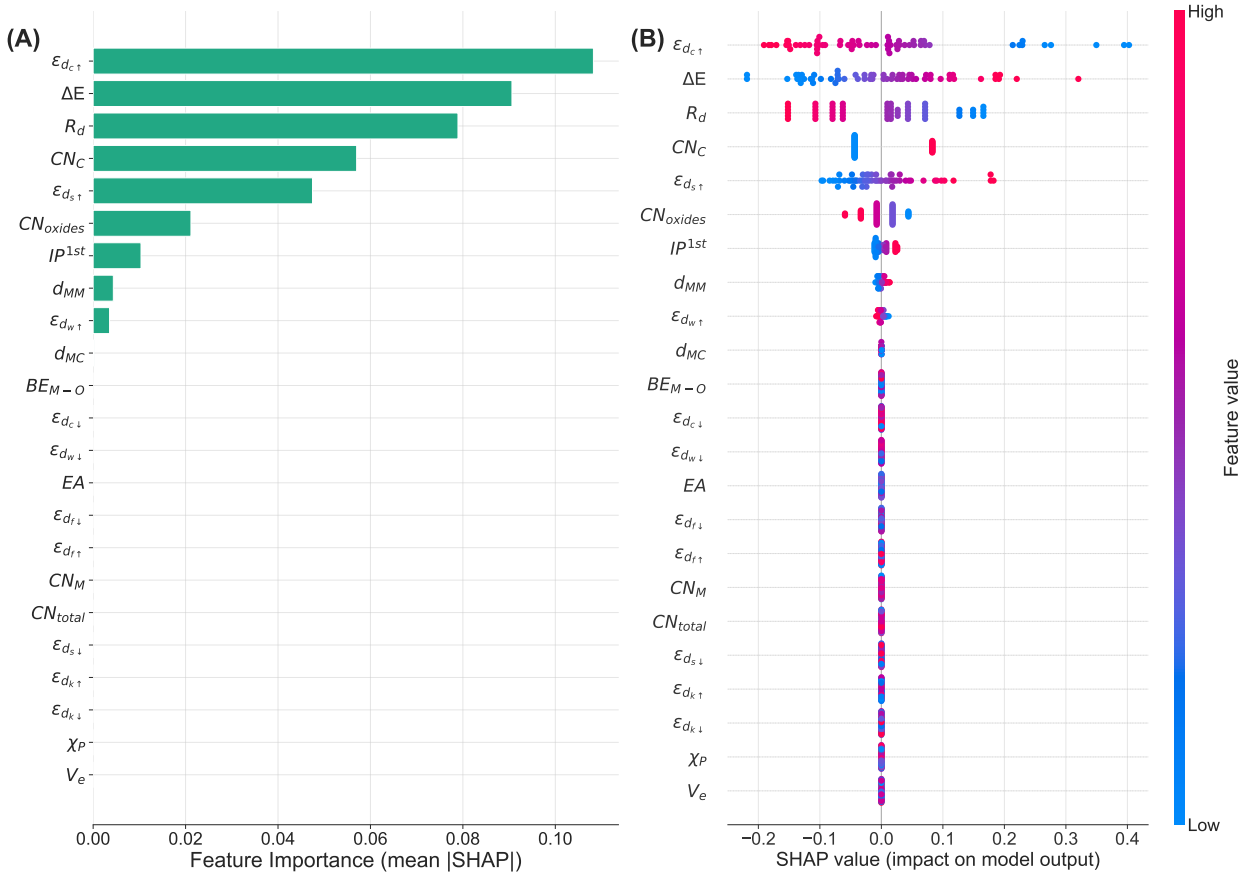


Figure S11: SHAP analysis of elasticnet regression model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

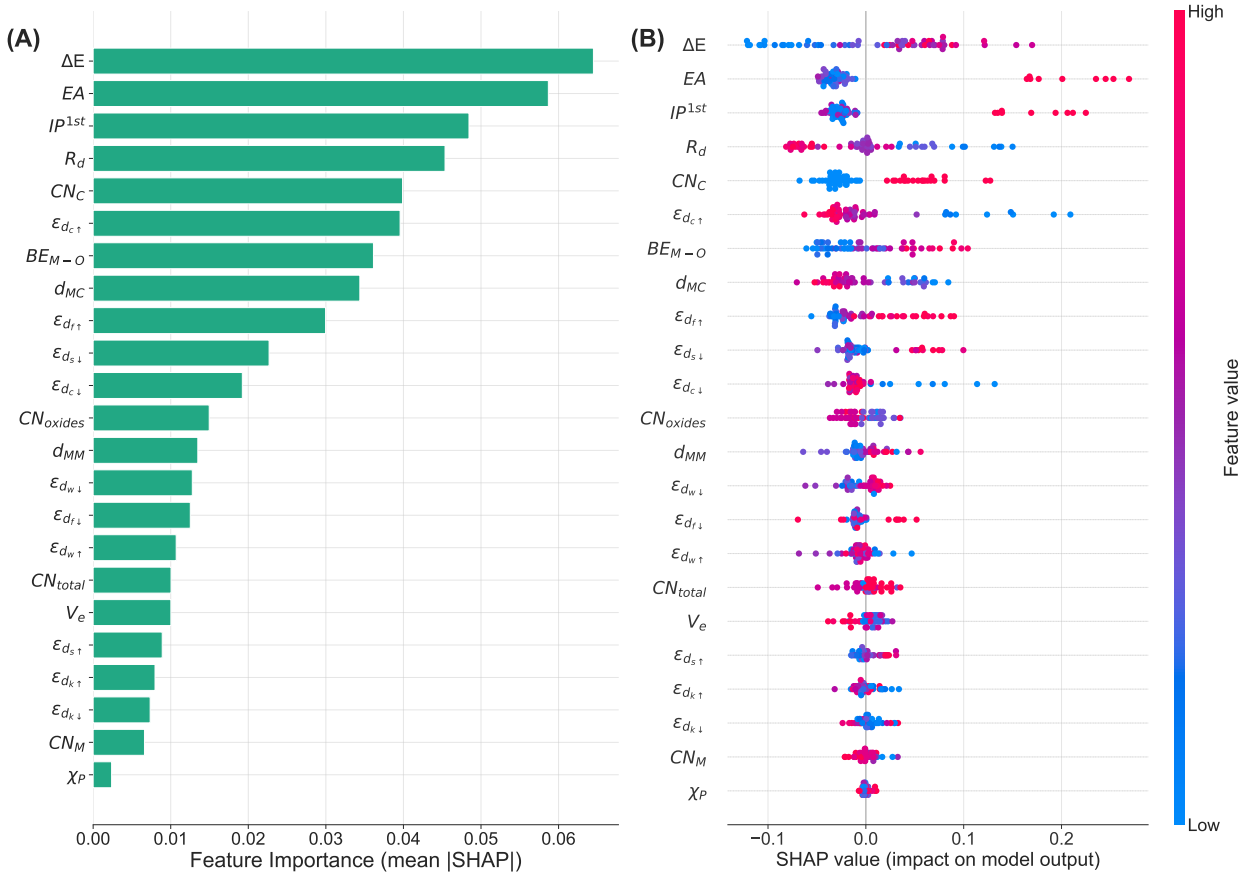


Figure S12: SHAP analysis of ETR ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

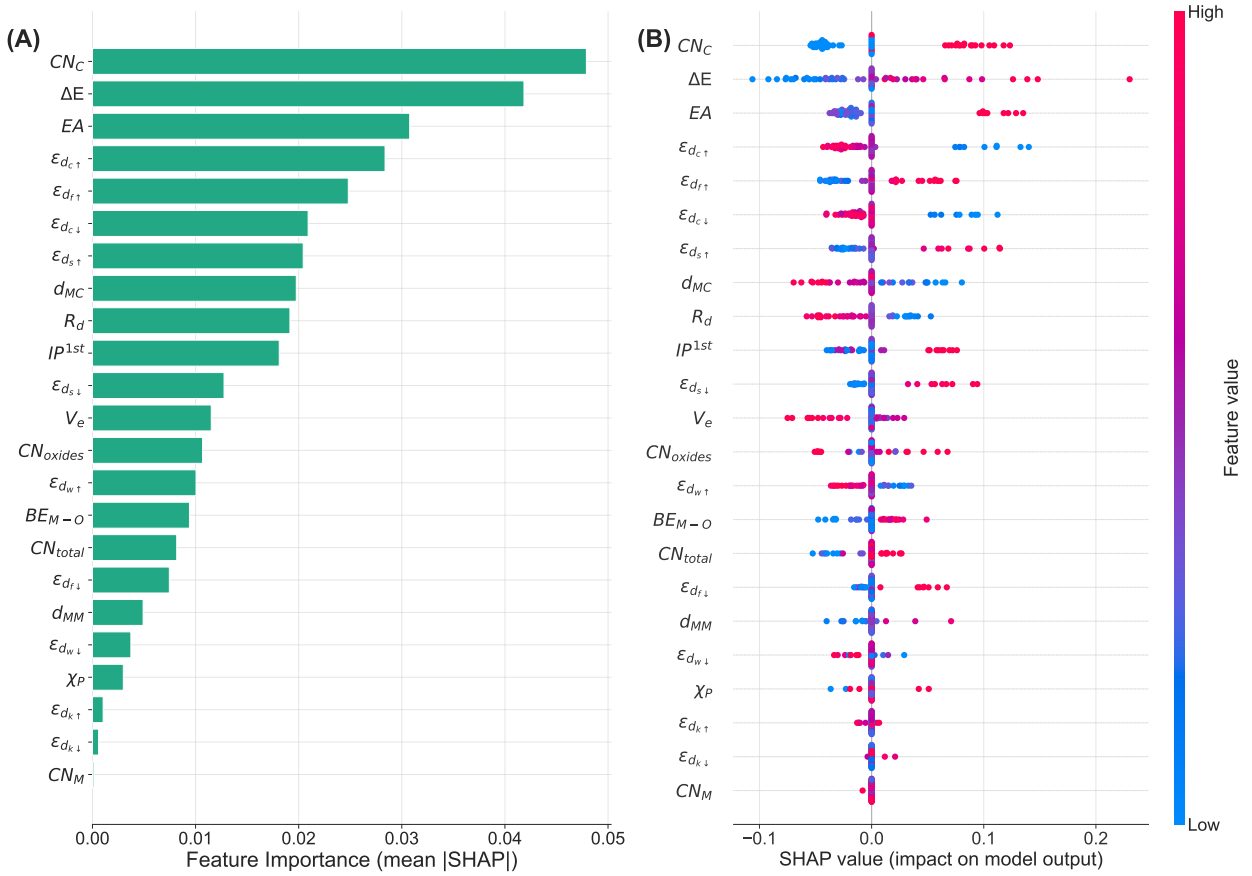


Figure S13: SHAP analysis of KNN ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

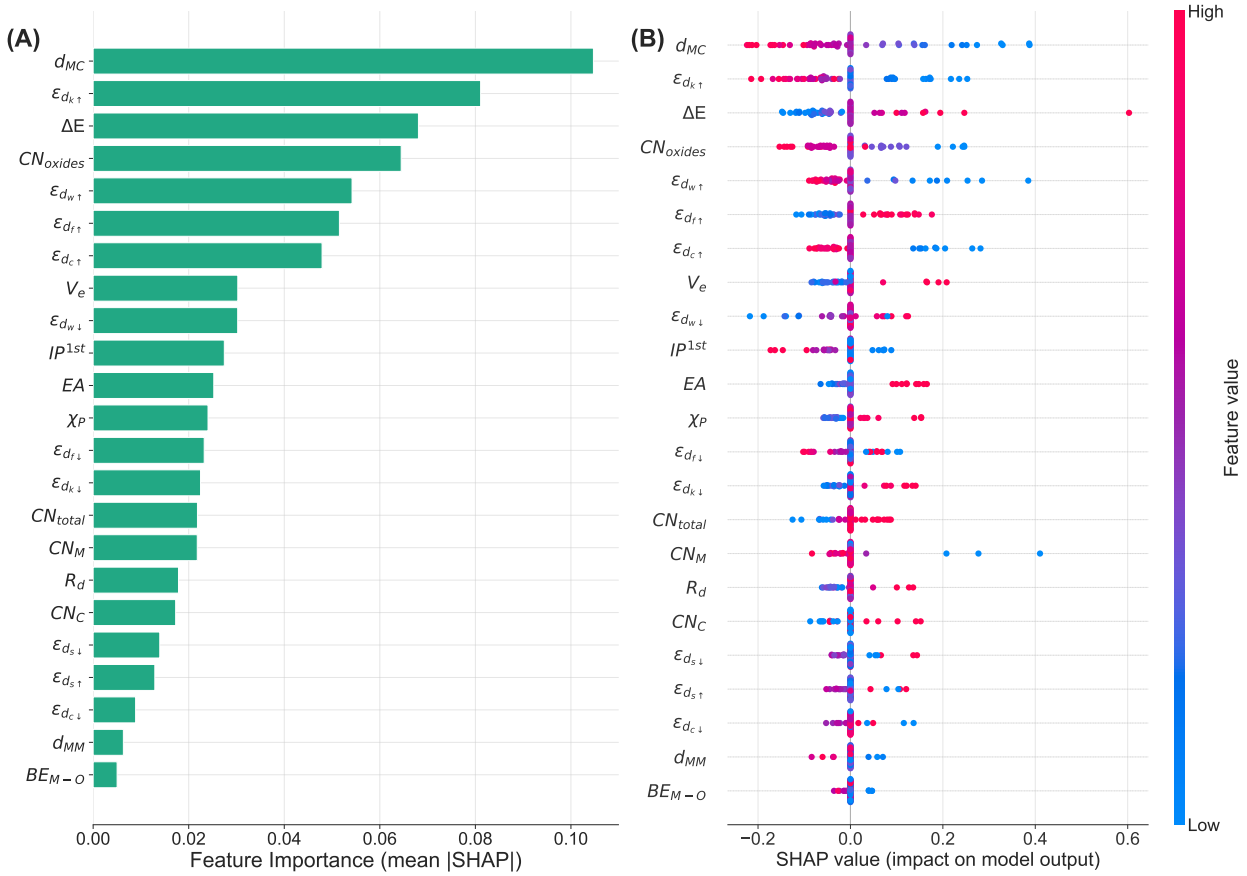


Figure S14: SHAP analysis of MLP–NN model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

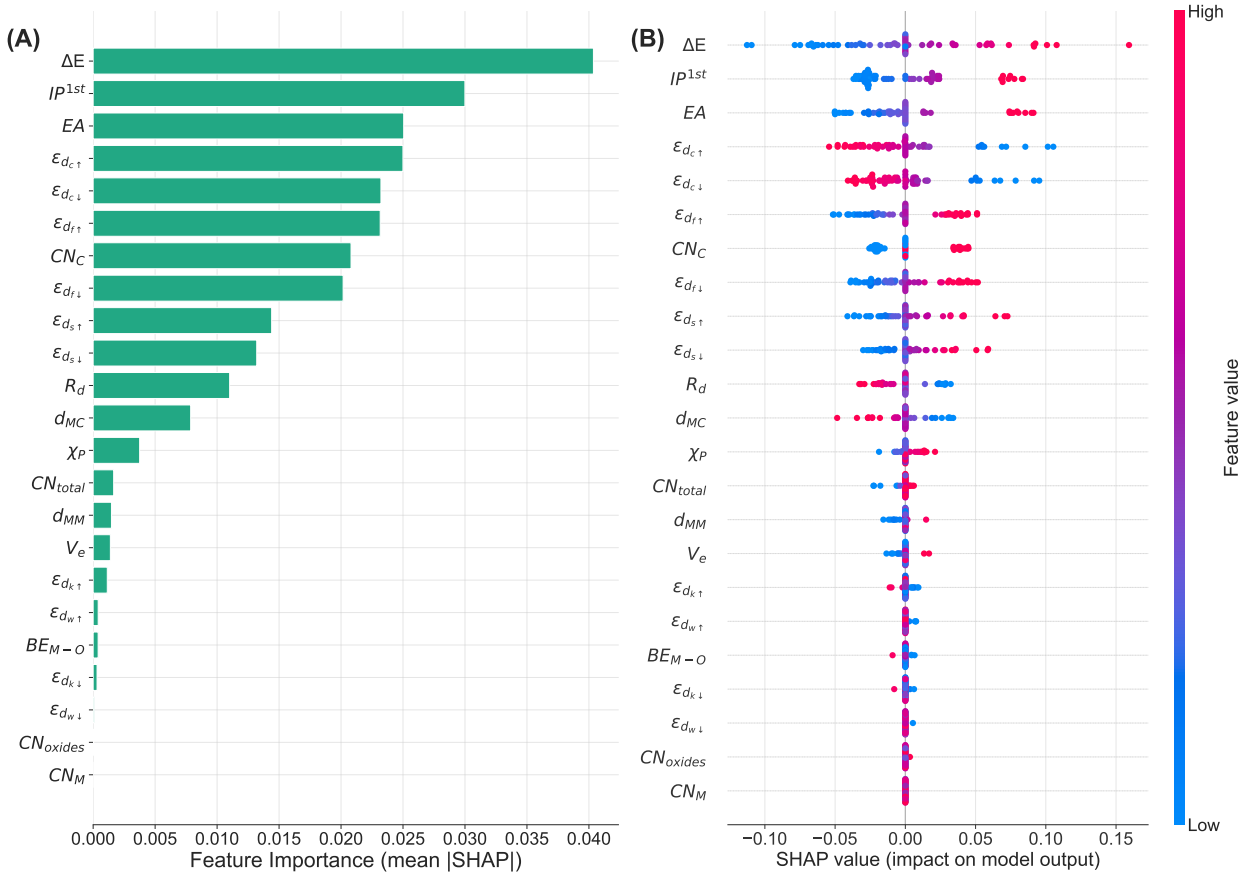


Figure S15: SHAP analysis of SVR ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

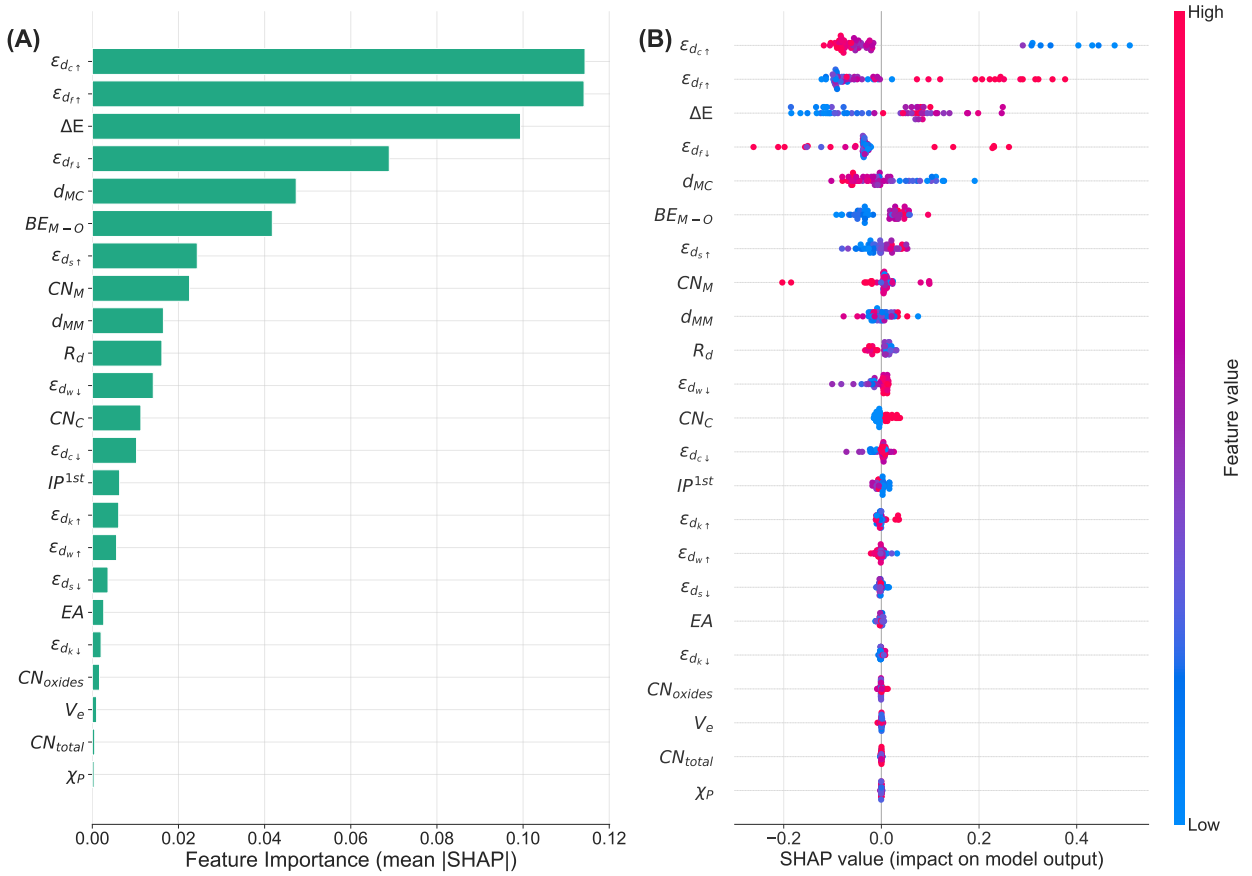


Figure S16: SHAP analysis of GBR ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

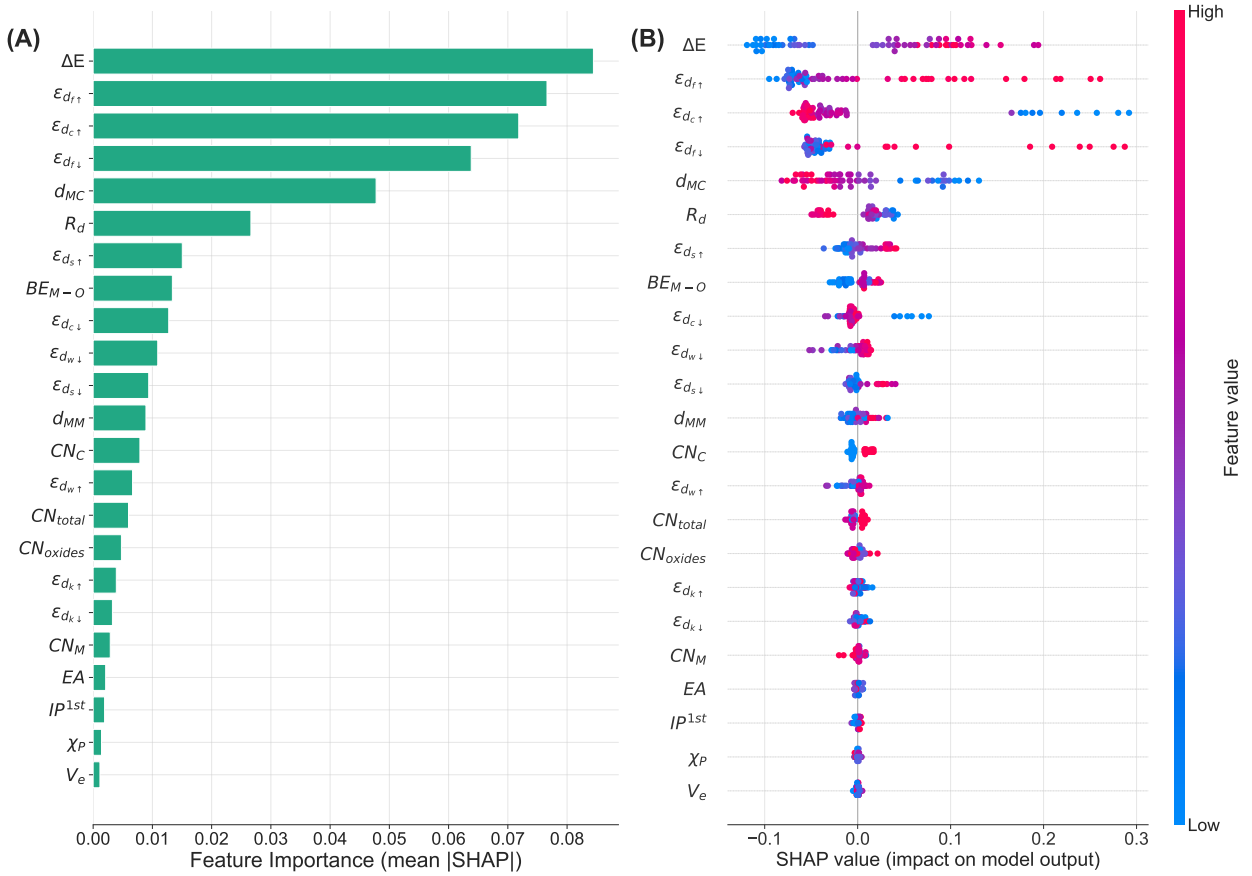


Figure S17: SHAP analysis of RFR ML model. (A) Feature importance. (B) SHAP values as a function of feature value(s).

Best SISSO models identified using 52 samples

Following are the 7 SISSO models that breaches the DFT-level error margin (±0.20 eV) on the test set:

1D model (test RMSE = 0.19 eV) with 4 feature complexity:

$$E_a = 0.001 \cdot \frac{BE_{M-O}}{|\frac{\varepsilon_{d_{c\downarrow}}}{CN_{total}} - \frac{\Delta E}{\varepsilon_{d_{k\uparrow}}}|} + 0.848 \tag{13}$$

2D model (test RMSE = 0.18 eV) with 2 feature complexity:

$$E_a = -0.015 \cdot [\frac{BE_{M-O}}{\varepsilon_{d_{s\uparrow}}} \cdot \varepsilon_{d_{w\downarrow}}] - 0.269 \cdot (|\varepsilon_{d_{c\downarrow}} \cdot \varepsilon_{d_{f\uparrow}}| - \Delta E) + 0.984 \tag{14}$$

2D model (test RMSE = 0.14 eV) with 4 feature complexity:

$$E_a = 0.066 \cdot (\frac{BE_{M-O}}{|(\varepsilon_{d_{c\downarrow}} \cdot \varepsilon_{d_{k\uparrow}}) - (\Delta E \cdot CN_{total})|}) + 2.505 \cdot \frac{e^{\varepsilon_{d_{w\downarrow}}}}{(\varepsilon_{d_{w\downarrow}})^2 \cdot d_{MC}} - 1.489 \tag{15}$$

3D model (test RMSE = 0.16 eV) with 2 feature complexity:

$$E_a = 0.381 \cdot log(BE_{M-O} \cdot \varepsilon_{d_{f\uparrow}}) - 0.417 \cdot |\varepsilon_{d_{c\uparrow}} - \Delta E| - 0.149 \cdot \frac{\Delta E}{\varepsilon_{d_{f\uparrow}}} \cdot \varepsilon_{d_{w\uparrow}} + 0.464 \tag{16}$$

4D model (test RMSE = 0.17 eV) with 1 feature complexity:

$$E_a = 0.011 \cdot (BE_{M-O} \cdot CN_C) - 0.717 \cdot (\varepsilon_{d_{k\uparrow}} - \varepsilon_{d_{f\uparrow}}) + 0.38 \cdot (\varepsilon_{d_{c\downarrow}} \cdot EA) - 0.004 \cdot (\varepsilon_{d_{k\uparrow}})^3 + 3.86 \tag{17}$$

4D model (test RMSE = 0.13 eV) with 2 feature complexity:

$$E_a = -0.10 \cdot \frac{BE_{M-O}}{CN_C - \varepsilon_{d_{k\uparrow}}} - 0.29 \cdot |(\varepsilon_{d_{c\downarrow}} \cdot \varepsilon_{d_{f\uparrow}}) - \Delta E| + 0.0001 \cdot \frac{e^{\varepsilon_{d_{k\uparrow}}}}{CN_M} + 0.49 \cdot |(\varepsilon_{d_{f\uparrow}} - \varepsilon_{d_{s\downarrow}}) - \chi_P| + 0.65 \tag{18}$$

5D model (test RMSE = 0.17 eV) with 4 feature complexity:

$$\begin{aligned} y = {} & 0.001 \cdot \frac{\mathrm{BE}_{M-O}}{\left| \left(\frac{d_{MM}}{CN_{total}} \right) - \left(\frac{\Delta E}{\varepsilon_{d_{k\uparrow}}} \right) \right|} \\ & + 0.957 \cdot \frac{\left| (CN_{total} \cdot \varepsilon_{d_{f\uparrow}}) - (\varepsilon_{d_{k\downarrow}} \cdot \varepsilon_{d_{f\downarrow}}) \right|}{CN_{oxides}} \\ & - 0.003 \cdot \frac{(CN_{oxides} - V_e)}{\left(\frac{\Delta E}{V_e} \right) - d_{MM}} \\ & - 0.030 \cdot \left| \left((EA - \Delta E) \cdot \varepsilon_{d_{k\downarrow}} \right) - \left(\frac{\mathrm{BE}_{M-O}}{\chi_P} \right) \right| \\ & + 0.139 \cdot \left| e^{(\Delta E + EA)} - (CN_C - \chi_P) \right| \\ & + 0.815 \end{aligned}$$

SISSO results over 61 samples

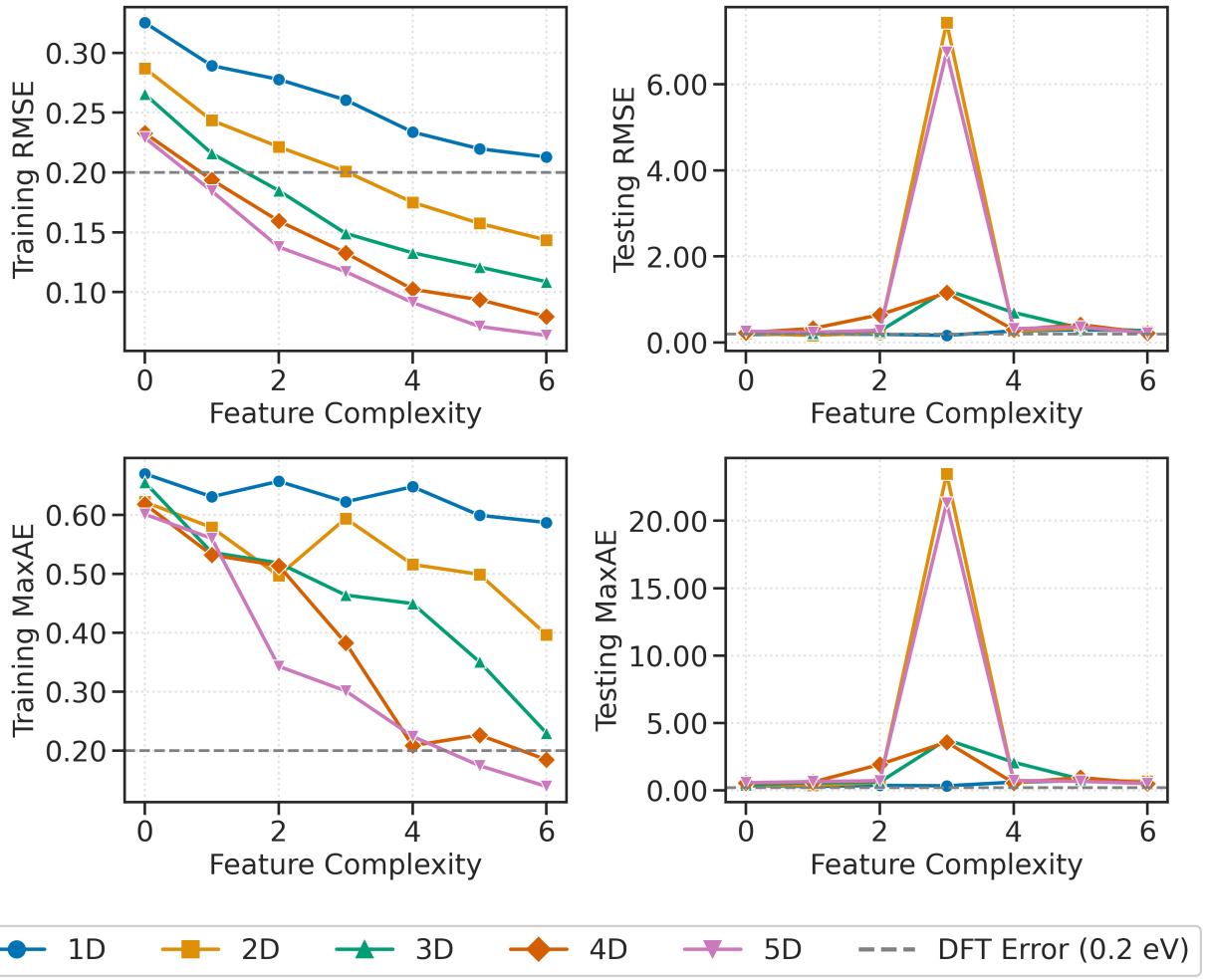
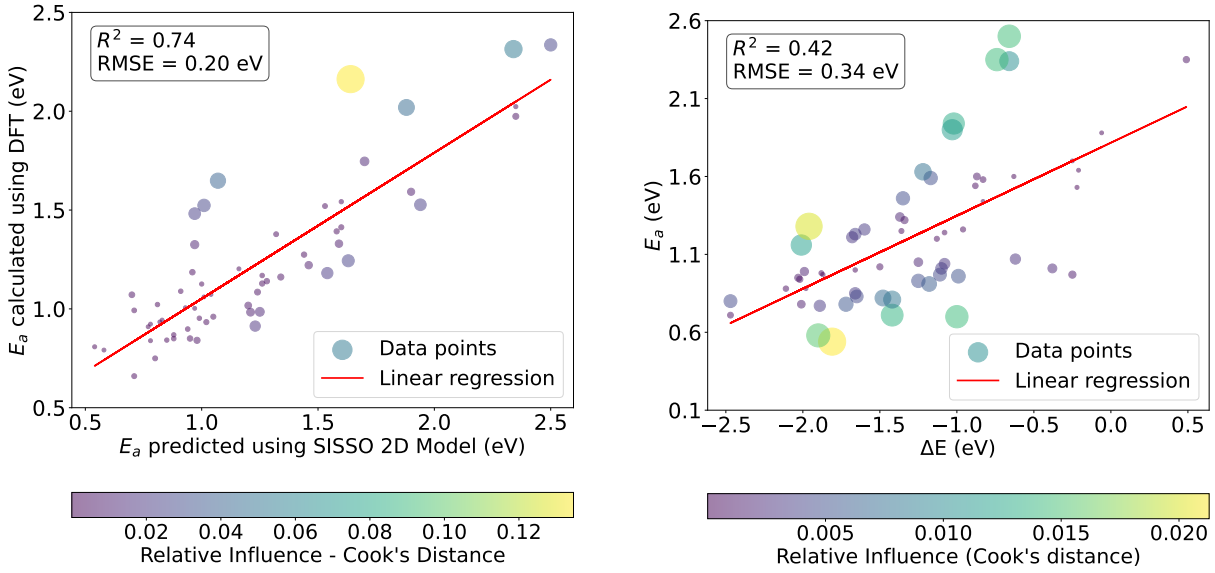


Figure S18: SISSO training and testing results for a dataset containing 61 samples. Dataset is split into 51 training samples and 10 test samples.



(a) (b)

Figure S19: (a) Correlation of the SISSO-predicted E_a with the DFT-calculated E_a. These are results for 61 sample size. (b) BEP relationship using DFT-calculated E_a.

Table S2: Training Data

Sample No.	E_a	$\varepsilon_{d_{c\uparrow}}$	$\varepsilon_{d_{c\downarrow}}$	$\varepsilon_{d_{w\uparrow}}$	$\varepsilon_{d_{w\downarrow}}$	ΔE	EA	IP^{1st}	$BE_{M\text{-}O}$	CN_C	CN_M	CN_{total}	$\varepsilon_{d_{s\uparrow}}$	$\varepsilon_{d_{s\downarrow}}$	$\varepsilon_{d_{k\uparrow}}$	$\varepsilon_{d_{k\downarrow}}$	χ_P	$\varepsilon_{d_{f\uparrow}}$	$\varepsilon_{d_{f\downarrow}}$	V_e	CN_{oxides}	R_d	d_{MC}	d_{MM}
#1	2.340	-4.700	-4.701	1.861	1.861	-0.660	2.309	9.225	3.681	3.000	5.000	8.000	0.346	0.344	5.137	5.139	2.540	0.954	0.954	11.000	6.000	1.740	2.096	2.835
#2	1.880	-4.334	-4.331	1.679	1.676	-0.060	2.309	9.225	3.681	2.000	6.000	8.000	0.305	0.310	5.911	5.905	2.540	0.961	0.961	11.000	6.000	1.740	2.137	2.787
#3	2.500	-4.765	-4.757	1.859	1.854	-0.660	2.309	9.225	3.681	3.000	5.000	8.000	0.307	0.318	5.215	5.214	2.540	0.957	0.955	11.000	6.000	1.740	2.080	2.855
#4	0.970	-1.597	-1.538	1.705	1.708	-0.250	0.660	7.881	33.821	2.000	5.000	7.000	-1.526	-1.525	8.706	8.669	1.880	0.826	0.811	9.000	9.000	1.520	1.893	2.553
#5	1.940	-1.708	-1.152	1.612	1.716	-1.020	0.660	7.881	33.821	3.000	5.000	8.000	-1.486	-1.772	9.195	9.680	1.880	0.875	0.758	9.000	9.000	1.520	1.895	2.509
#6	0.990	-0.955	-0.632	2.030	2.129	-1.990	0.666	6.767	25.001	2.000	6.000	8.000	-1.585	-1.691	7.371	7.507	1.660	0.607	0.514	6.000	6.000	1.660	1.946	2.535
#7	1.600	-1.194	-0.368	2.035	2.294	-0.630	0.666	6.767	25.001	3.000	5.000	8.000	-1.420	-1.730	6.744	7.075	1.660	0.685	0.433	6.000	6.000	1.660	1.975	2.630
#8	1.340	-1.132	0.076	1.809	2.191	-1.370	0.666	6.767	25.001	2.000	6.000	8.000	-1.429	-1.922	7.671	8.355	1.660	0.718	0.363	6.000	6.000	1.660	1.961	2.516
#9	1.440	-1.113	-0.042	1.833	2.203	-0.830	0.666	6.767	25.001	2.000	6.000	8.000	-1.566	-1.972	8.119	8.579	1.660	0.698	0.389	6.000	6.000	1.660	1.948	2.562
#10	1.460	-1.207	-0.589	2.146	2.357	-1.350	0.666	6.767	25.001	3.000	5.000	8.000	-1.494	-1.665	6.631	6.703	1.660	0.665	0.479	6.000	6.000	1.660	1.883	2.664
#11	1.160	-2.074	-0.815	1.742	1.975	-2.010	0.163	7.903	30.829	2.000	5.000	7.000	-1.167	-1.758	7.333	8.374	1.830	0.867	0.576	8.000	8.000	1.560	1.894	2.529
#12	1.900	-2.132	-0.881	1.795	2.075	-1.030	0.163	7.903	30.829	3.000	5.000	8.000	-1.112	-1.684	6.828	7.626	1.830	0.872	0.581	8.000	8.000	1.560	1.891	2.621
#13	0.800	-0.253	-0.166	2.302	2.332	-2.470	0.079	6.828	19.881	2.000	5.000	7.000	-1.905	-1.905	8.081	8.036	1.540	0.434	0.404	4.000	6.000	1.760	2.064	2.663
#14	0.700	-0.273	-0.264	2.420	2.420	-1.000	0.079	6.828	19.881	3.000	5.000	8.000	-1.816	-1.817	7.194	7.200	1.540	0.410	0.409	4.000	6.000	1.760	2.065	2.796
#15	1.020	-0.043	-0.029	2.276	2.276	-1.500	0.079	6.828	19.881	2.000	6.000	8.000	-1.952	-1.954	8.354	8.351	1.540	0.373	0.370	4.000	6.000	1.760	2.064	2.662
#16	1.200	-0.011	0.001	2.300	2.303	-1.130	0.079	6.828	19.881	2.000	4.000	6.000	-2.064	-2.063	8.838	8.835	1.540	0.377	0.373	4.000	6.000	1.760	2.049	2.734
#17	1.230	-0.242	-0.221	2.478	2.477	-1.660	0.079	6.828	19.881	3.000	5.000	8.000	-1.868	-1.872	7.349	7.372	1.540	0.407	0.404	4.000	6.000	1.760	2.060	2.821
#18	0.820	-0.753	-0.749	2.129	2.129	-1.480	0.892	6.759	8.984	2.000	5.000	7.000	-1.363	-1.366	5.818	5.817	1.600	0.521	0.518	13.000	8.000	1.980	2.072	2.739
#19	0.930	-0.722	-0.721	2.131	2.131	-1.250	0.892	6.759	8.984	2.000	4.000	6.000	-1.408	-1.406	5.970	5.962	1.600	0.527	0.527	13.000	8.000	1.980	2.079	2.766
#20	0.910	-0.950	-0.972	2.293	2.293	-1.180	0.892	6.759	8.984	3.000	4.000	7.000	-1.236	-1.238	4.982	4.992	1.600	0.531	0.534	13.000	8.000	1.980	2.089	2.829
#21	0.880	-1.083	-1.050	2.347	2.347	-1.980	0.892	6.759	8.984	3.000	4.000	7.000	-1.292	-1.289	5.164	5.152	1.600	0.567	0.563	13.000	8.000	1.980	2.042	2.876
#22	0.710	-1.058	-1.030	2.188	2.189	-1.420	0.892	6.759	8.984	2.000	5.000	7.000	-1.300	-1.298	5.618	5.609	1.600	0.597	0.582	13.000	8.000	1.980	2.078	2.686
#23	1.010	-1.920	-1.865	1.515	1.517	-0.380	1.161	7.640	37.012	2.000	5.000	7.000	-1.444	-1.435	9.774	9.710	1.910	0.899	0.891	10.000	7.000	1.490	1.928	2.615
#24	1.070	-1.927	-1.927	1.496	1.496	-0.620	1.161	7.640	37.012	3.000	5.000	8.000	-1.224	-1.222	8.832	8.831	1.910	0.891	0.891	10.000	7.000	1.490	2.008	2.626
#25	1.530	-1.733	-1.714	1.463	1.463	-0.220	1.161	7.640	37.012	2.000	6.000	8.000	-1.498	-1.507	10.358	10.372	1.910	0.902	0.898	10.000	7.000	1.490	1.919	2.528
#26	1.280	-3.360	-3.304	2.041	2.025	-1.960	2.128	9.017	3.248	2.000	5.000	7.000	-0.449	-0.421	4.812	4.754	2.280	0.914	0.902	10.000	7.000	1.770	2.076	2.656
#27	1.640	-3.652	-3.653	2.193	2.193	-0.210	2.128	9.017	3.248	3.000	5.000	8.000	-0.343	-0.344	4.277	4.278	2.280	0.909	0.909	10.000	7.000	1.770	2.049	2.741
#28	2.350	-3.735	-3.720	2.202	2.197	-0.740	2.128	9.017	3.248	3.000	5.000	8.000	-0.385	-0.378	4.348	4.338	2.280	0.914	0.911	10.000	7.000	1.770	2.023	2.743
#29	1.700	-3.230	-3.223	2.017	2.015	-0.250	2.128	9.017	3.248	2.000	6.000	8.000	-0.474	-0.472	5.015	5.006	2.280	0.913	0.911	10.000	7.000	1.770	2.042	2.658

Continued on next page

Sample No.	E_a	$\varepsilon_{d_{c\uparrow}}$	$\varepsilon_{d_{c\downarrow}}$	$\varepsilon_{d_{w\uparrow}}$	$\varepsilon_{d_{w\downarrow}}$	ΔE	EA	IP^{1st}	$BE_{M\text{-}O}$	CN_C	CN_M	CN_{total}	$\varepsilon_{d_{s\uparrow}}$	$\varepsilon_{d_{s\downarrow}}$	$\varepsilon_{d_{k\uparrow}}$	$\varepsilon_{d_{k\downarrow}}$	χ_P	$\varepsilon_{d_{f\uparrow}}$	$\varepsilon_{d_{f\downarrow}}$	V_e	CN_{oxides}	R_d	d_{MC}	d_{MM}
#30	2.350	-3.356	-3.354	2.129	2.129	0.490	2.128	9.017	3.248	2.000	5.000	7.000	-0.636	-0.634	5.107	5.105	2.280	0.916	0.916	10.000	7.000	1.770	2.009	2.679
#31	1.040	-0.485	-0.407	2.106	2.124	-1.080	0.524	6.746	22.426	2.000	6.000	8.000	-1.659	-1.672	7.414	7.392	1.630	0.478	0.455	5.000	5.000	1.710	2.014	2.457
#32	1.240	-0.425	-0.429	2.150	2.150	-1.080	0.524	6.746	22.426	2.000	5.000	7.000	-1.835	-1.840	8.115	8.133	1.630	0.467	0.467	5.000	5.000	1.710	1.989	2.570
#33	1.260	-0.647	-0.646	2.300	2.299	-0.960	0.524	6.746	22.426	3.000	5.000	8.000	-1.639	-1.639	6.743	6.743	1.630	0.486	0.486	5.000	5.000	1.710	2.005	2.651
#34	0.880	-0.542	-0.498	2.212	2.218	-2.110	0.524	6.746	22.426	2.000	6.000	8.000	-1.743	-1.737	7.549	7.520	1.630	0.497	0.486	5.000	5.000	1.710	1.996	2.562
#35	1.210	-0.652	-0.682	2.397	2.383	-1.680	0.524	6.746	22.426	3.000	5.000	8.000	-1.682	-1.674	6.716	6.723	1.630	0.488	0.502	5.000	5.000	1.710	1.932	2.725
#36	0.810	-1.616	-1.574	2.316	2.314	-1.420	0.815	7.981	1.544	2.000	5.000	7.000	-1.046	-1.043	4.855	4.808	2.360	0.664	0.647	6.000	7.000	1.930	2.055	2.556
#37	0.960	-1.583	-1.579	2.384	2.384	-0.990	0.815	7.981	1.544	2.000	5.000	7.000	-1.156	-1.155	5.010	5.006	2.360	0.648	0.647	6.000	7.000	1.930	2.043	2.626
#38	0.970	-2.019	-2.015	2.594	2.594	-1.110	0.815	7.981	1.544	3.000	5.000	8.000	-0.968	-0.967	4.101	4.099	2.360	0.678	0.678	6.000	7.000	1.930	2.053	2.692
#39	0.970	-2.016	-1.981	2.677	2.679	-1.870	0.815	7.981	1.544	3.000	5.000	8.000	-1.014	-1.008	4.119	4.103	2.360	0.689	0.686	6.000	7.000	1.930	1.993	2.763
#40	0.780	-1.867	-1.852	2.491	2.487	-2.010	0.815	7.981	1.544	2.000	6.000	8.000	-1.075	-1.068	4.716	4.705	2.360	0.696	0.692	6.000	7.000	1.930	2.037	2.576
#41	0.850	-0.265	-0.261	2.232	2.231	-1.660	0.426	6.634	7.909	2.000	3.000	5.000	-1.600	-1.603	6.487	6.497	1.330	0.430	0.427	12.000	7.000	2.060	2.153	2.905
#42	1.050	-0.292	-0.264	2.247	2.248	-1.250	0.426	6.634	7.909	2.000	2.000	4.000	-1.627	-1.629	6.567	6.569	1.330	0.437	0.431	12.000	7.000	2.060	2.148	2.960
#43	1.010	-0.644	-0.635	2.378	2.377	-1.100	0.426	6.634	7.909	3.000	2.000	5.000	-1.401	-1.402	5.384	5.387	1.330	0.489	0.488	12.000	7.000	2.060	2.165	2.996
#44	0.980	-0.672	-0.622	2.403	2.403	-1.880	0.426	6.634	7.909	3.000	1.000	4.000	-1.457	-1.463	5.598	5.612	1.330	0.500	0.492	12.000	7.000	2.060	2.179	2.983
#45	0.710	-0.677	-0.628	2.262	2.265	-2.470	0.426	6.634	7.909	2.000	3.000	5.000	-1.482	-1.486	6.109	6.113	1.330	0.517	0.499	12.000	7.000	2.060	2.147	2.859
#46	0.940	-1.602	-1.551	2.212	2.201	-2.020	0.745	7.092	10.107	2.000	5.000	7.000	-1.265	-1.245	5.578	5.531	2.160	0.690	0.679	6.000	6.000	1.900	2.046	2.582
#47	1.580	-1.728	-1.729	2.467	2.467	-0.830	0.745	7.092	10.107	3.000	5.000	8.000	-1.066	-1.067	4.463	4.464	2.160	0.652	0.652	6.000	6.000	1.900	2.053	2.709
#48	1.250	-1.423	-1.403	1.972	1.961	-1.360	0.745	7.092	10.107	2.000	5.000	7.000	-0.886	-0.882	4.599	4.594	2.160	0.636	0.630	6.000	6.000	1.900	2.064	2.560
#49	1.540	-1.694	-1.669	2.145	2.138	-0.880	0.745	7.092	10.107	2.000	5.000	7.000	-0.812	-0.820	4.152	4.149	2.160	0.648	0.638	6.000	6.000	1.900	2.046	2.653
#50	1.260	-1.803	-1.757	2.270	2.259	-1.600	0.745	7.092	10.107	3.000	5.000	8.000	-0.969	-0.963	4.607	4.576	2.160	0.694	0.677	6.000	6.000	1.900	1.986	2.794
#51	0.540	-0.109	-0.098	1.967	1.967	-1.810	0.426	6.634	7.909	2.000	2.000	4.000	-1.620	-1.610	7.034	7.003	1.330	0.423	0.421	12.000	7.000	2.060	2.249	2.853
#52	0.580	-0.647	-0.625	1.925	1.925	-1.900	0.892	6.759	8.984	2.000	4.000	6.000	-1.311	-1.308	5.869	5.846	1.600	0.525	0.521	13.000	8.000	1.980	2.153	2.702
#53	1.000	-1.104	0.124	1.601	2.028	-1.660	0.666	6.767	25.001	2.000	5.000	7.000	-1.435	-2.018	8.232	9.154	1.660	0.732	0.361	6.000	6.000	1.660	2.034	2.578
#54	0.950	0.054	0.043	1.996	1.997	-2.030	0.079	6.828	19.881	2.000	4.000	6.000	-2.043	-2.044	9.351	9.353	1.540	0.373	0.375	4.000	6.000	1.760	2.162	2.678
#55	0.770	-0.367	-0.364	1.957	1.957	-1.890	0.524	6.746	22.426	2.000	5.000	7.000	-1.814	-1.810	8.322	8.307	1.630	0.472	0.471	5.000	5.000	1.710	2.045	2.566
#56	1.320	-2.059	-0.406	1.387	1.665	-1.340	0.163	7.903	30.829	2.000	5.000	7.000	-0.839	-1.866	8.014	9.734	1.830	0.895	0.518	8.000	8.000	1.560	1.981	2.603
#57	1.590	-1.585	-1.585	1.203	1.203	-1.170	1.161	7.640	37.012	2.000	5.000	7.000	-1.096	-1.092	10.185	10.176	1.910	0.893	0.893	10.000	7.000	1.490	2.092	2.574
#58	1.630	-1.306	-1.295	1.447	1.448	-1.220	0.660	7.881	33.821	2.000	4.000	6.000	-1.363	-1.367	9.028	9.033	1.880	0.808	0.806	9.000	9.000	1.520	1.966	2.581
#59	0.780	-1.578	-1.570	2.246	2.246	-1.720	0.815	7.981	1.544	2.000	5.000	7.000	-1.083	-1.080	4.950	4.939	2.360	0.674	0.673	6.000	7.000	1.930	2.069	2.602
#60	0.830	-1.364	-1.371	2.112	2.112	-1.650	0.745	7.092	10.107	2.000	4.000	6.000	-1.123	-1.128	5.245	5.259	2.160	0.667	0.668	6.000	6.000	1.900	2.093	2.611

Continued on next page

Sample No.	E_a	$\varepsilon_{d_{c\uparrow}}$	$\varepsilon_{d_{c\downarrow}}$	$\varepsilon_{d_{w\uparrow}}$	$\varepsilon_{d_{w\downarrow}}$	ΔE	EA	IP^{1st}	$BE_{M\text{-}O}$	CN_C	CN_M	CN_{total}	$\varepsilon_{d_{s\uparrow}}$	$\varepsilon_{d_{s\downarrow}}$	$\varepsilon_{d_{k\uparrow}}$	$\varepsilon_{d_{k\downarrow}}$	χ_P	$\varepsilon_{d_{f\uparrow}}$	$\varepsilon_{d_{f\downarrow}}$	V_e	CN_{oxides}	R_d	d_{MC}	d_{MM}
#61	1.600	-3.319	-3.190	1.811	1.811	-0.870	2.128	9.017	3.248	2.000	4.000	6.000	-0.215	-0.222	4.909	4.923	2.280	0.905	0.906	10.000	7.000	1.770	2.100	2.659