Supporting Information

Unraveling the Formation Mechanism of Oxygen Vacancy on the Surface of Transition Metal-Doped Ceria Utilizing Artificial Intelligence

Liangliang Xu[‡1] Ning Xu[‡2,4], Yue Wang[5], Xiaojuan Hu*[2,3], Zhong-Kang Han*[2]

[1]Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-Ro, Yuseong-Gu, Daejeon 34141, Republic of Korea

[2]School of Materials Science and Engineering, Zhejiang University, Hangzhou, 310027, China

[3]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany

[4]Department of Physics, School of Physical Science and Technology, Ningbo University, Ningbo, 315211, China

[5]Department of Electrical Engineering, Hanyang University, Seoul 04763, Republic of Korea

‡These authors contributed equally.

*Emails: xhu@fhi-berlin.mpg.de; hanzk@zju.edu.cn

Contents

1. Supplementary Methods

1.1 Subgroup discovery

1.2 SISSO approach

2. Supplemental Figures and Tables

Table S1. The formation energies of the transition metal doped-CeO_2(100) systems.

Table S2. The formation energies of the transition metal doped-CeO_2(110) systems.

Table S3. The formation energies of the transition metal doped-CeO_2(111) systems.

Figure S1. Calculated spin charges (top and side views) of V-doped CeO_2(100) surface (a) (b), and W-doped CeO_2(100) surface (c). Nb-doped CeO_2(110) surface (d) (e) (f). Os-doped CeO_2(111) surface (g), and Ta-doped CeO_2(111) surface (h) (i) (j).

Figure S2. Correlation between the formation energies and low-cost features of the 3*d*-, 4*d*-, and 5*d*-block transition metal dopants.

Figure S3. Correlation between the oxygen vacancy formation energies and low-cost features of the 3*d*-, 4*d*-, and 5*d*-block transition metal dopants.

Figure S4. The error distributions for all the lower-dimensional models of the surface O_v formation energies: (a) 1D, (b) 2D, (c) 3D, (d) 4D, and the SISSO selected best 5D model (e).

Table S4. Descriptors components, coefficient(C) and intercept(I) of the SISSO-identified 1D to 5D models for leave-5%-out validation.

Table S5. The selected feature set of the 5D-SISSO model during validation process.

Supplementary References

1. Supplementary Methods

1.1 Subgroup discovery

The subgroup discovery was performed using RealKD package.[1] Each feature was split to 14 subsets using 14-means clustering algorithm.[2] The borders between adjacent data clusters (a1, a2, ...) are applied further for construction of inequalities (feature1 < a1), (feature2 ≥ a2), etc.. While final result might depend on the number of considered clusters, in our previous study we found that relatively high numbers of considered clusters provide essentially the same result.[3] The candidate subgroups are built as conjunctions of obtained simple inequalities. The main idea of SGD is that the subgroups are unique if the distribution of the data in them is as different as possible from the data distribution in the whole sampling. Here the data distribution is the distribution of a target property (ΔE_F & ΔE_V). The uniqueness is evaluated with a quality function. In this study we used the following function:

$$Q(S) = \frac{s(S)}{s(P)}\left(\frac{\mathrm{med}(P)-\mathrm{med}(S)}{\mathrm{med}(P)-\min(P)}\right)\left(1-\frac{\mathrm{amd}(S)}{\mathrm{amd}(P)}\right) \quad \text{(S1)}$$

with S – subgroup, P – whole sampling, s – size, med and min – median and minimal values of a target property, amd – absolute average deviation of the data around the median of target property. With this function the algorithm is searching for subgroups with lower values of target properties. The search was done with an adapted for such purposes Monte-Carlo algorithm,[4] in which first a certain number of trial conjunctions (seeds) is generated. Afterwards, for each seed (accompanied with pruning of inequalities) the quality function is calculated. We have tested here several numbers of initial seeds: 10 000, 30 000, 50 000 and 100 000. The subgroups with the overall high quality function value were selected.

1.2 SISSO calculation

In SISSO, the feature construction was performed by creating all the possible mathematical combinations, within certain feature complexity, between the primary features using the operators $\{+, -, \cdot, /, \log, \exp, \exp-, ^{-1}, ^{2}, ^{3}, \sqrt{}, \sqrt[3]{}, |-|\}$. A list of the primary features for model training data sets can be found in the file "Supplementary file1." The code for SISSO and user guide are available at: https://github.com/rouyang2017/SISSO.

Table S1. The formation energies of the transition metal doped-CeO_2(100) surfaces.

Systems	Formation energy (eV)	O_v formation energy (eV)
Ce/CeO_2(100)	/	1.3248
Ag/CeO_2(100)	0.8806	-0.7649
Au/CeO_2(100)	0.6889	0.1344
Cd/CeO_2(100)	-1.6771	-0.3077
Co/CeO_2(100)	-2.1040	0.0723
Cr/CeO_2(100)	-7.0689	0.3446
Cu/CeO_2(100)	-0.1574	-0.4801
Fe/CeO_2(100)	-3.3845	0.3278
Hf/CeO_2(100)	-10.6819	1.1832
Hg/CeO_2(100)	-0.7927	0.3417
Ir/CeO_2(100)	-1.7595	1.4700
Mn/CeO_2(100)	-4.7566	0.5709
Mo/CeO_2(100)	-8.7790	1.4765
Nb/CeO_2(100)	-10.3926	1.2638
Ni/CeO_2(100)	-1.0478	0.1203
Os/CeO_2(100)	-3.6163	1.6679
Pd/CeO_2(100)	-0.1797	0.0607
Pt/CeO_2(100)	-0.5841	0.7017
Re/CeO_2(100)	-6.9910	1.7273
Rh/CeO_2(100)	-1.1927	0.5603
Ru/CeO_2(100)	-3.1164	0.9742
Sc/CeO_2(100)	-8.7130	0.1713
Ta/CeO_2(100)	-10.1491	0.5735
Tc/CeO_2(100)	-5.9561	0.7993
Ti/CeO_2(100)	-9.5551	0.1794
V/CeO_2(100)	-9.1747	0.6571
W/CeO_2(100)	-9.4552	1.4776
Y/CeO_2(100)	-9.2190	1.0793
Zn/CeO_2(100)	-2.0586	-0.6811
Zr/CeO_2(100)	-10.4769	1.2418
V/CeO_2(100)-a	-8.9109	0.5511
V/CeO_2(100)-b	-8.5756	0.8518
W/CeO_2(100)-a	-9.3645	1.9251

Table S2. The formation energies of the transition metal doped-CeO_2(110) surfaces.

Systems	Formation energy (eV)	O_v formation energy (eV)
Ce/CeO_2(110)	/	1.3628
Ag/CeO_2(110)	0.9887	-0.4652
Au/CeO_2(110)	0.8930	0.1603
Cd/CeO_2(110)	-1.1249	-0.7864
Co/CeO_2(110)	-0.4911	-1.1726
Cr/CeO_2(110)	-4.2249	-1.7270
Cu/CeO_2(110)	-0.5484	-0.3005
Fe/CeO_2(110)	-1.1243	-1.4013
Hf/CeO_2(110)	-10.1998	0.3974
Hg/CeO_2(110)	0.4333	-0.9371
Ir/CeO_2(110)	-1.1739	0.5169
Mn/CeO_2(110)	-4.3381	0.7255
Mo/CeO_2(110)	-7.4664	0.0180
Nb/CeO_2(110)	-9.2039	0.3260
Ni/CeO_2(110)	-0.9371	-0.6781
Os/CeO_2(110)	-2.5914	0.1456
Pd/CeO_2(110)	0.7702	-1.5715
Pt/CeO_2(110)	0.5383	-1.2349
Re/CeO_2(110)	-5.5769	0.2676
Rh/CeO_2(110)	-0.3306	0.0979
Ru/CeO_2(110)	-0.9040	-1.0278
Sc/CeO_2(110)	-8.2404	-0.1570
Ta/CeO_2(110)	-9.6603	0.1103
Tc/CeO_2(110)	-3.4133	-1.5863
Ti/CeO_2(110)	-8.3696	-0.5202
V/CeO_2(110)	-7.0235	-0.6643
W/CeO_2(110)	-7.8333	-0.0696
Y/CeO_2(110)	-8.6798	0.3024
Zn/CeO_2(110)	-1.2691	-0.9377
Zr/CeO_2(110)	-10.0310	0.7530
Nb/CeO_2(110)-a	-9.2141	0.0056
Nb/CeO_2(110)-b	-9.2119	-0.0201
Nb/CeO_2(110)-c	-9.2179	0.0222

Table S3. The formation energies of the transition metal doped-CeO_2(111) surfaces.

Systems	Formation energy (eV)	O_v formation energy (eV)
Ce/CeO_2(111)	/	2.1325
Ag/CeO_2(111)	2.0690	-0.2579
Au/CeO_2(111)	1.7476	0.6469
Cd/CeO_2(111)	0.2174	-0.8739
Co/CeO_2(111)	-0.3335	0.8430
Cr/CeO_2(111)	-4.4318	0.1756
Cu/CeO_2(111)	2.2034	-1.7354
Fe/CeO_2(111)	-2.0558	0.5071
Hf/CeO_2(111)	-9.7577	1.2544
Hg/CeO_2(111)	1.7124	-0.5947
Ir/CeO_2(111)	0.4524	1.2674
La/CeO_2(111)	-8.3518	0.8178
Mn/CeO_2(111)	-2.9553	1.6282
Mo/CeO_2(111)	-5.3516	-0.3176
Nb/CeO_2(111)	-8.6914	1.6545
Ni/CeO_2(111)	0.1041	0.7988
Os/CeO_2(111)	-0.5615	-0.4504
Pd/CeO_2(111)	0.9071	0.8672
Pt/CeO_2(111)	0.4440	1.5505
Re/CeO_2(111)	-5.1546	0.9927
Rh/CeO_2(111)	0.3927	1.2897
Ru/CeO_2(111)	-0.2753	1.2498
Sc/CeO_2(111)	-7.1139	0.1236
Ta/CeO_2(111)	-9.1165	1.5726
Tc/CeO_2(111)	-4.2207	1.0218
Ti/CeO_2(111)	-7.8932	1.2228
V/CeO_2(111)	-6.4701	1.1966
W/CeO_2(111)	-6.8311	1.0265
Y/CeO_2(111)	-8.1550	0.6026
Zn/CeO_2(111)	0.1080	-1.0411
Zr/CeO_2(111)	-9.6291	1.2621
Os/CeO_2(111)-a	-1.6138	2.2512
Ta/CeO_2(111)-a	-9.0905	0.8003
Ta/CeO_2(111)-b	-9.0504	1.0707
Ta/CeO_2(111)-c	-9.0928	1.0128

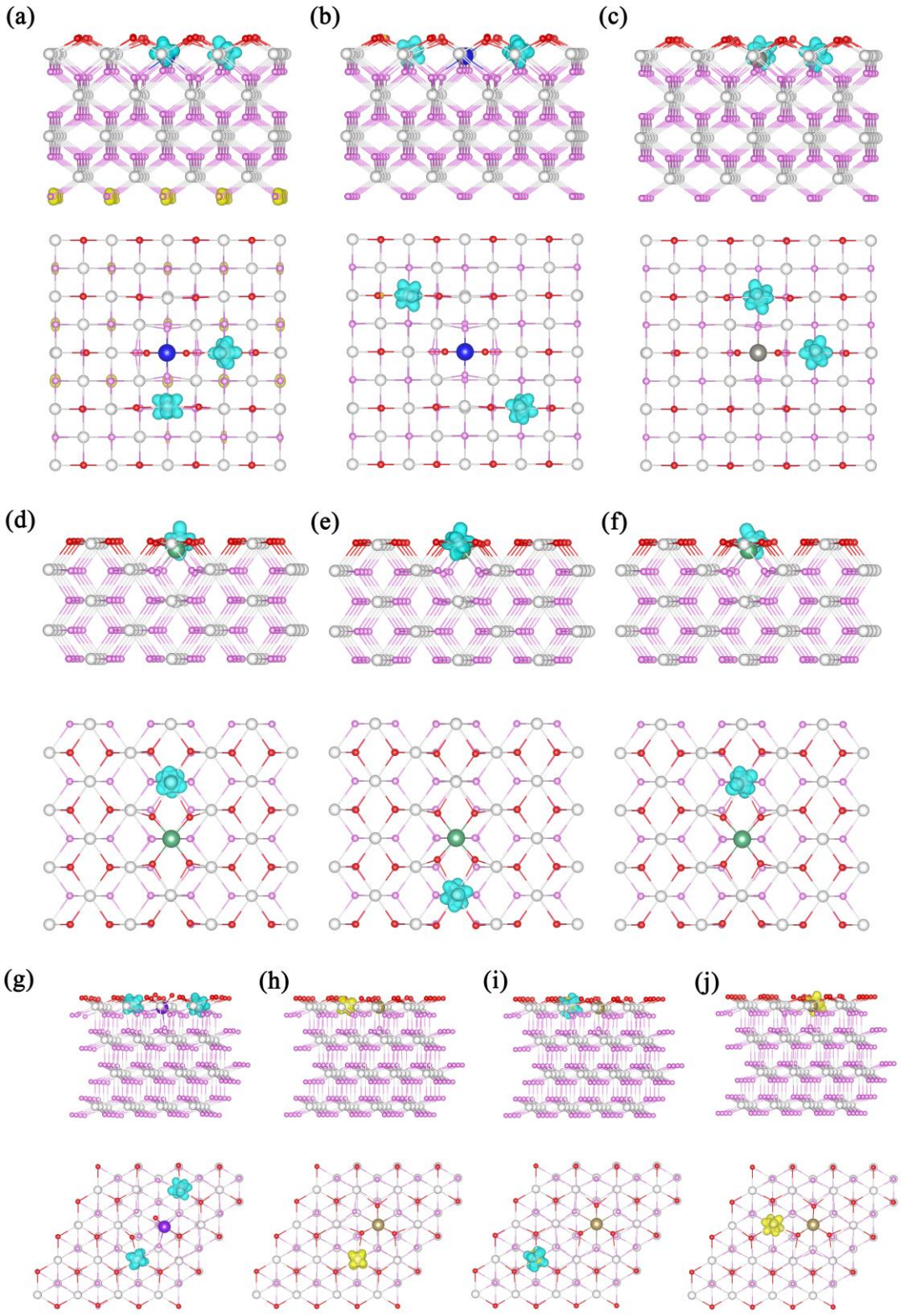
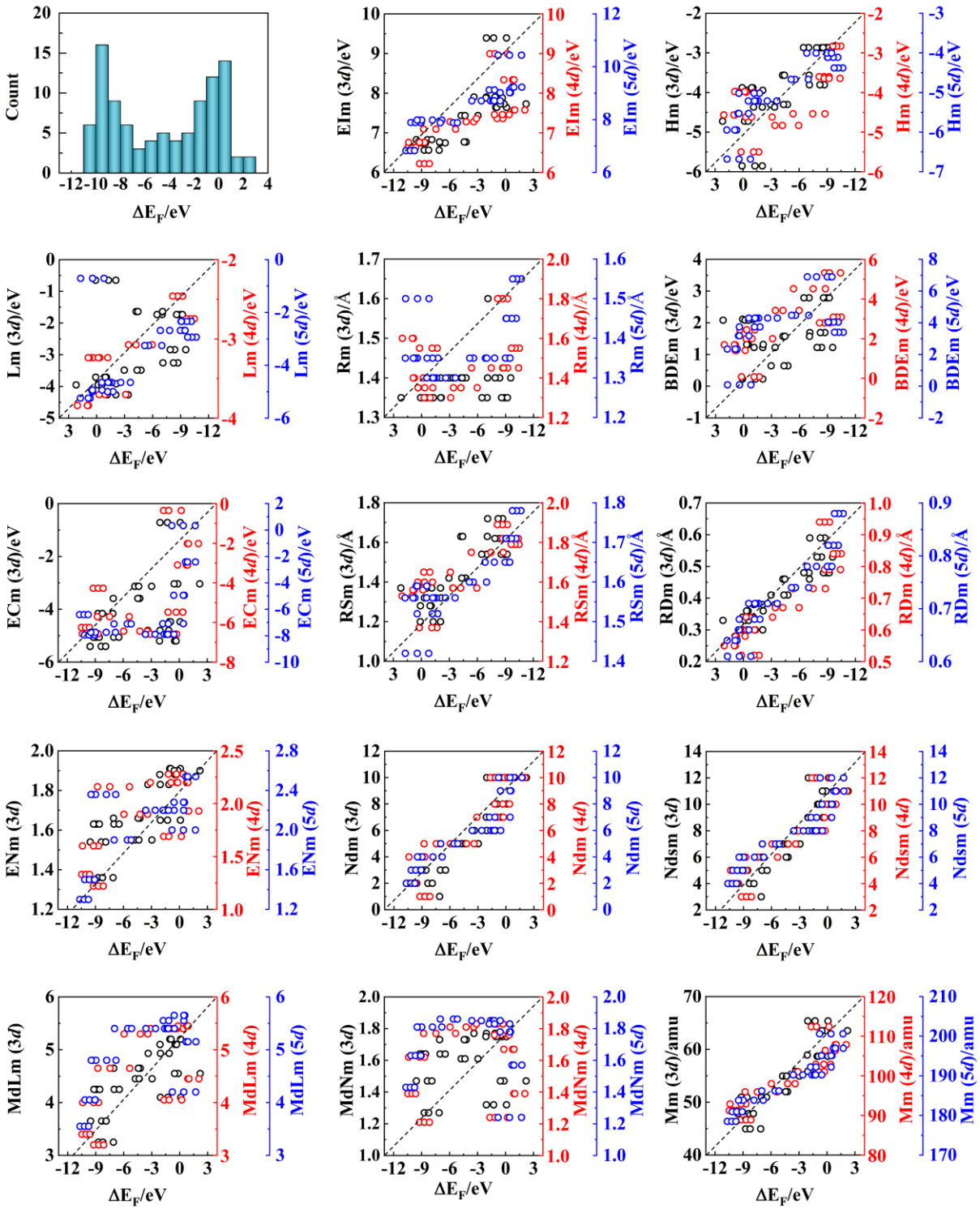


Figure S1. Calculated spin charges (top and side views) of V-doped CeO_2(100) surface (a) (b), and W-doped CeO_2(100) surface (c). Nb-doped CeO_2(110) surface (d) (e) (f). Os-doped CeO_2(111) surface (g), and Ta-doped CeO_2(111) surface (h) (i) (j).


Count
ΔE_F/eV
EIm (3d)/eV
EIm (4d)/eV
EIm (5d)/eV
Hm (3d)/eV
Hm (4d)/eV
Hm (5d)/eV
Lm (3d)/eV
Lm (4d)/eV
Lm (5d)/eV
Rm (3d)/Å
Rm (4d)/Å
Rm (5d)/Å
BDEm (3d)/eV
BDEm (4d)/eV
BDEm (5d)/eV
ECm (3d)/eV
ECm (4d)/eV
ECm (5d)/eV
RSm (3d)/Å
RSm (4d)/Å
RSm (5d)/Å
RDm (3d)/Å
RDm (4d)/Å
RDm (5d)/Å
ENm (3d)
ENm (4d)
ENm (5d)
Ndm (3d)
Ndm (4d)
Ndm (5d)
Ndsm (3d)
Ndsm (4d)
Ndsm (5d)
MdLm (3d)
MdLm (4d)
MdLm (5d)
MdNm (3d)
MdNm (4d)
MdNm (5d)
Mm (3d)/amu
Mm (4d)/amu
Mm (5d)/amu

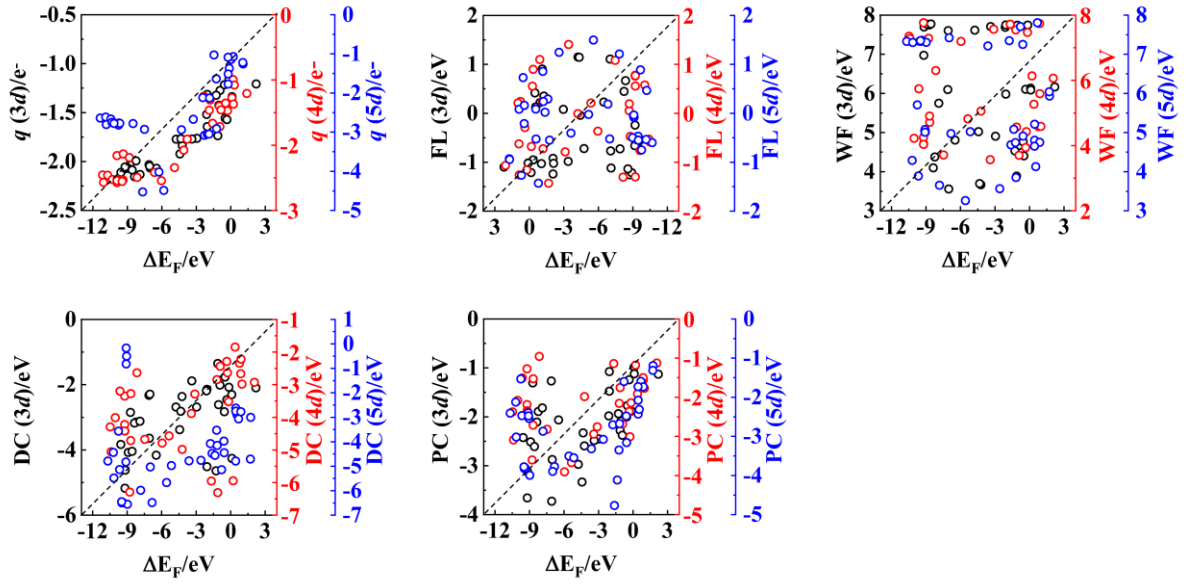
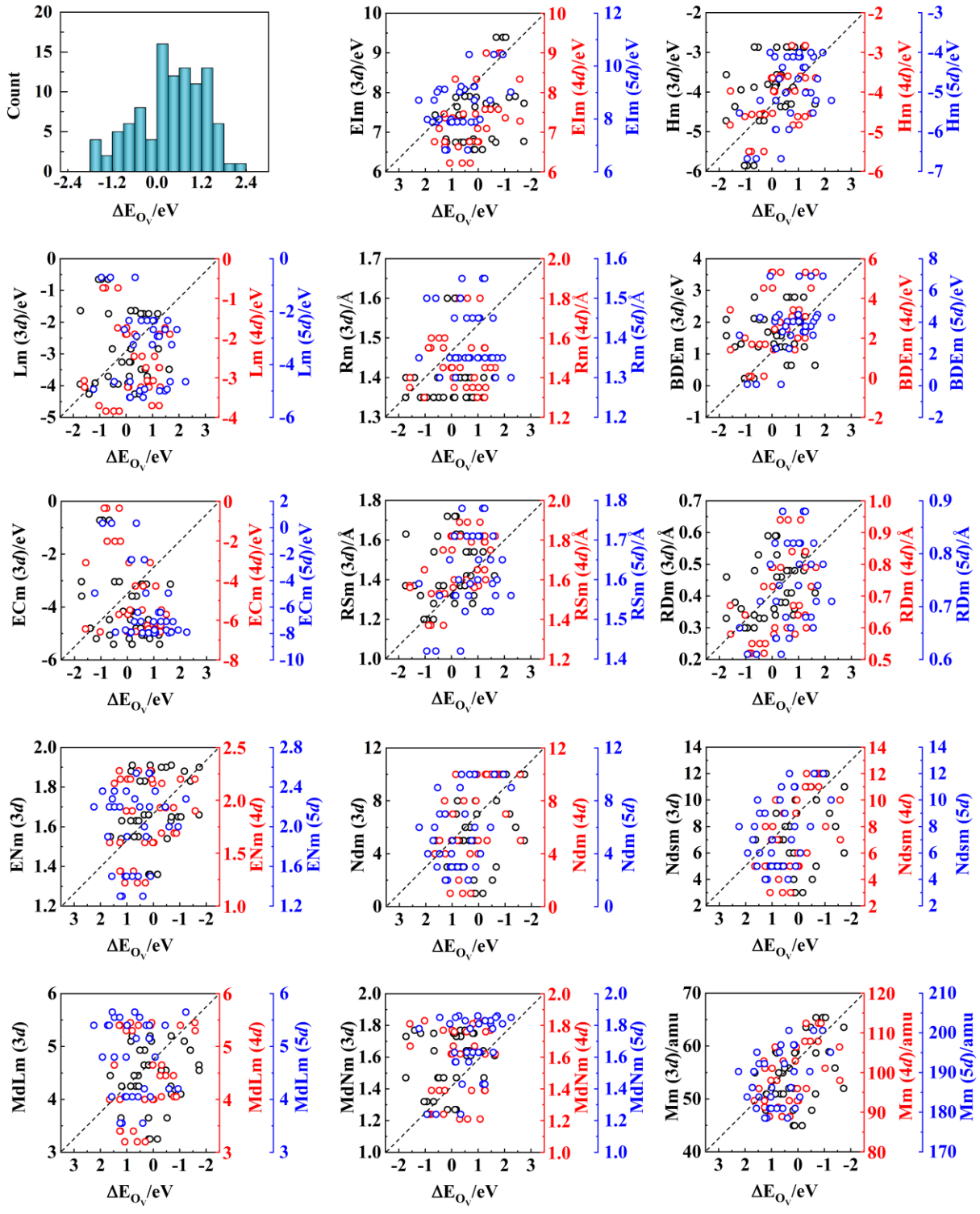


Figure S2. Correlation between the formation energies and low-cost features of the 3*d*-, 4*d*-, and 5*d*-block transition metal dopants.


Count
ΔE$_{O_v}$/eV
EIm (3d)/eV
EIm (4d)/eV
EIm (5d)/eV
Hm (3d)/eV
Hm (4d)/eV
Hm (5d)/eV
Lm (3d)/eV
Lm (4d)/eV
Lm (5d)/eV
Rm (3d)/Å
Rm (4d)/Å
Rm (5d)/Å
BDEm (3d)/eV
BDEm (4d)/eV
BDEm (5d)/eV
ECm (3d)/eV
ECm (4d)/eV
ECm (5d)/eV
RSm (3d)/Å
RSm (4d)/Å
RSm (5d)/Å
RDm (3d)/Å
RDm (4d)/Å
RDm (5d)/Å
ENm (3d)
ENm (4d)
ENm (5d)
Ndm (3d)
Ndm (4d)
Ndm (5d)
Ndsm (3d)
Ndsm (4d)
Ndsm (5d)
MdLm (3d)
MdLm (4d)
MdLm (5d)
MdNm (3d)
MdNm (4d)
MdNm (5d)
Mm (3d)/amu
Mm (4d)/amu
Mm (5d)/amu

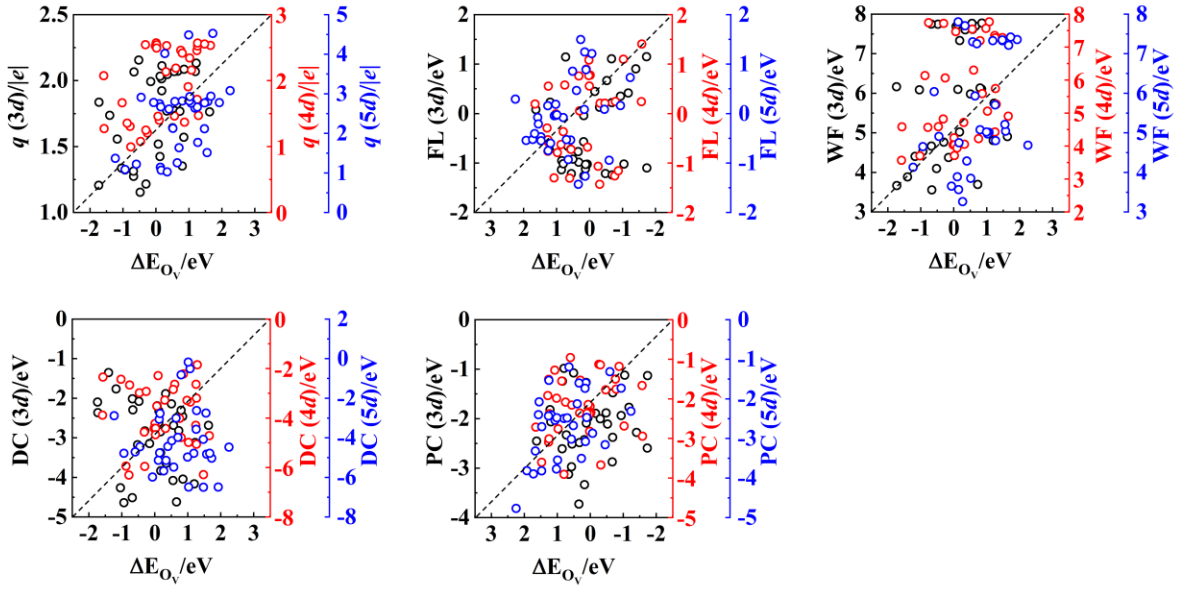


Figure S3. Correlation between the oxygen vacancy formation energies and low-cost features of the 3*d*-, 4*d*-, and 5*d*-block transition metal dopants.

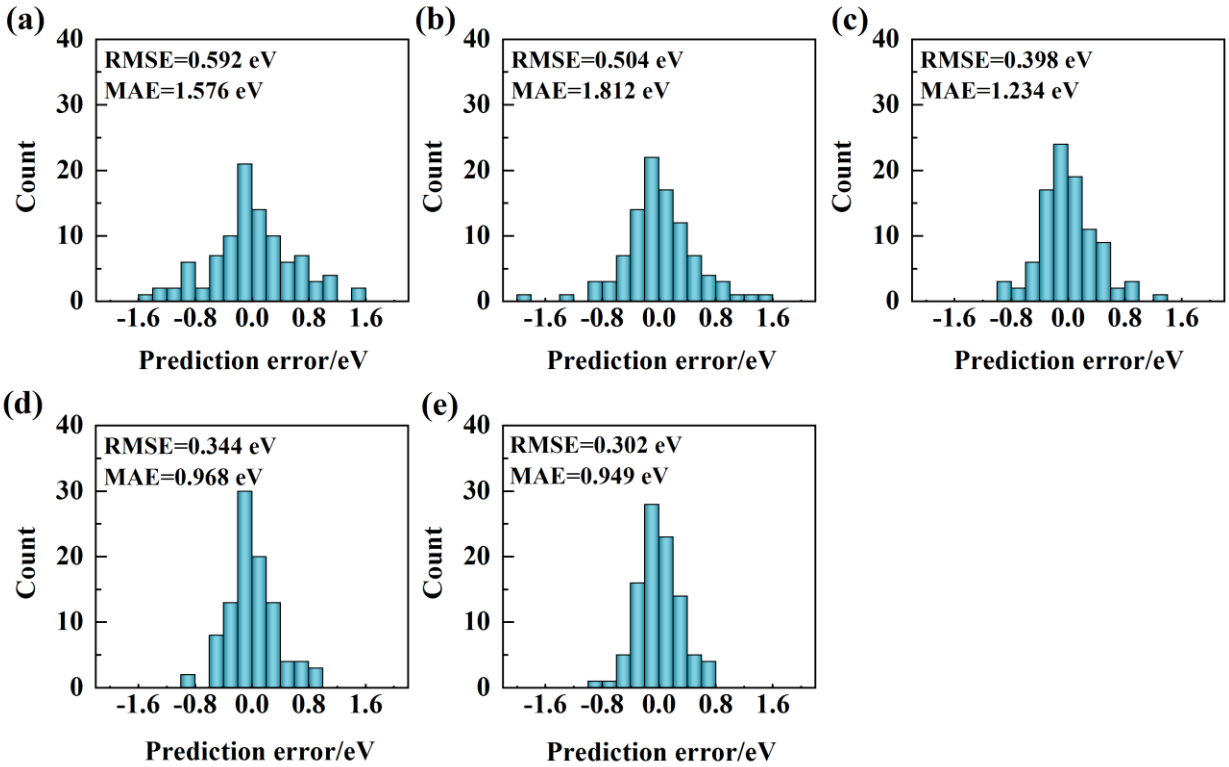


Figure S4. The error distributions for all the lower-dimensional models of the surface O_v formation energies: (a) 1D, (b) 2D, (c) 3D, and (d) 4D, and the SISSO selected best model (e) 5D.

Table S4. Descriptors components, coefficient(C) and intercept(I) of the SISSO-identified 1D to 5D models for leave-5%-out validation.

Iteration	Dimension	Descriptor	C	I
01	1D	(RDm−Rm)/q)+(RSm/(WF−FL))	-3.095	2.760
	1D	(RDm−Rm)/q)+(RSm/(WF−FL))	-3.151	2.829
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.043	
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/q)	-0.192	-0.011
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.044	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.078	
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/q)	-0.185	-0.027
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.045	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.076	
	4D	(PC*ECm)/((Lm+RDm)+(FL*Ndsm))	-0.011	
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/q)	-0.185	-0.013
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.044	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.070	
	4D	(PC/EIm)/((Lm+RDm)+(FL*Ndsm))	0.501	
	5D	(Rm/PC)/((PC−Rm)−(q*ENm))	-0.039	
	1D	((ECm/WF)*(FL−RDm))−(PC*RDm)	0.655	-0.959
	1D	(FL−(RDm*MdNm))−(WF/(q+PC))	-1.215	0.663
	2D	((FL−MdNm)−PC)/(Rm+(ECm*RDm))	0.111	
	1D	(FL−(RDm*MdNm))−(WF/(q+PC))	-1.193	0.632
	2D	((FL−BDEm)−PC)/(RDm+(Rm/ECm))	-0.019	
	3D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.037	

02	1D	(FL−(RDm*MdNm))−(WF/(*q*+PC))	-1.124	0.644
	2D	(FL/(BDEm−FL))/(WF+(ECm*ENm))	0.533	
	3D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.041	
	4D	(DC+MdLm)/((Ndsm/WF)−(Rm+RDm))	-0.024	
	1D	(FL−(RDm*MdNm))−(WF/(*q*+PC))	-1.102	0.624
	2D	(FL/(BDEm−FL))/(WF+(ECm*ENm))	0.493	
	3D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.047	
	4D	(DC+MdLm)/((Ndsm/WF)−(Rm+RDm))	-0.023	
	5D	((WF/PC)−Hm)/(RDm+(BDEm/Hm))	-0.005	
03	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.120	2.781
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.231	-1.098
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.068	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.235	-1.087
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	-0.068	
	3D	(MdLm/Ndm)/((Ndsm/Hm)−(PC*BDEm))	-0.093	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.240	-1.099
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.068	
	3D	(RSm/DC)/((Hm*BDEm)−(Ndsm/PC))	-0.235	
	4D	(BDEm/(FL+BDEm))/((DC/FL)−PC)	-0.228	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.245	-1.088
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.067	
	3D	(ENm/Ndm)/((Ndsm/Hm)−(PC*BDEm))	-0.255	
	4D	(WF+(EIm/PC))/((DC/FL)−PC)	-0.165	
	5D	(*q*−Lm)/((WF+ECm)+(Ndsm/EIm))	-0.029	
	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.074	2.739

	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.130	2.809
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.043	
	1D	(ECm+(ECm/*q*))+(FL*(FL+RSm))	-0.384	-0.450
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.065	
	3D	(MdNm/Ndm)/((Ndsm/Hm)−(PC*BDEm))	-0.254	
04	1D	(ECm+(ECm/*q*))+(FL*(FL+RSm))	-0.385	-0.708
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.067	
	3D	(Lm/Ndm)/((Ndsm/Hm)−(PC*BDEm))	0.112	
	4D	(DC+(FL*BDEm))/((*q**WF)−ECm)	0.443	
	1D	(ECm+(ECm/*q*))+(FL*(FL+RSm))	-0.415	-0.730
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.070	
	3D	(MdNm/Ndm)/((Ndsm/Hm)−(PC*BDEm))	-0.295	
	4D	(DC+(FL*BDEm))/((*q**WF)−ECm)	0.395	
	5D	(DC−Lm)/((ENm−MdNm)−(FL/Ndm))	0.014	
	1D	(RDm+(FL/*q*))*((EIm*MdNm)−Ndsm)	0.266	-0.585
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/*q*)	-0.200	-0.018
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.046	
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/*q*)	-0.194	-0.004
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.044	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.078	
05	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/*q*)	-0.188	-0.021
	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.045	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.076	
	4D	(PC*ECm)/((Lm+RDm)+(FL*Ndsm))	-0.010	
	1D	((FL−ENm)*(RDm*MdLm))−(Ndm/*q*)	-0.176	0.017

	2D	(ECm+MdLm)/((Hm+MdLm)−(Lm+WF))	-0.045	
	3D	(DC−Hm)/((EIm−BDEm)−(Rm*WF))	-0.073	
	4D	(Hm*ECm)/((Lm+RDm)+(FL*Ndsm))	-0.008	
	5D	(Hm+BDEm)/((DC+MdLm)−(RSm−PC))	-0.019	
	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.163	2.797
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.230	-1.089
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.067	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.240	-1.123
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.071	
	3D	(ENm/Lm)/((Hm*BDEm)−(Ndsm/PC))	-0.333	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.234	-1.130
06	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.058	
	3D	(MdNm/*q*)/((Hm*BDEm)−(Ndsm/PC))	-0.135	
	4D	(BDEm−ENm)/((Hm+PC)−(*q*WF))	0.135	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.240	-1.180
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.057	
	3D	(MdNm/*q*)/((Hm*BDEm)−(Ndsm/PC))	-0.137	
	4D	(BDEm−ENm)/((Hm+PC)−(*q*WF))	0.136	
	5D	(DC*(BDEm−WF))/((PC*Rm)−Hm)	0.006	
	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.103	2.756
	1D	((FL/*q*)−RSm)*(ECm+(FL*MdLm))	0.142	-0.635
	2D	(Rm/(PC+WF))/((ECm−BDEm)−Hm)	0.760	
	1D	((FL/*q*)−RSm)*(ECm+(FL*MdLm))	0.146	-0.630
	2D	(Rm/(PC+WF))/((ECm−BDEm)−Hm)	0.817	
	3D	(Ndsm/(FL−PC))/(FL+(RDm/ECm))	0.003	

07	1D	((FL/*q*)−RSm)*(ECm+(FL*MdLm))	0.147	-0.627
	2D	(Rm/(PC+WF))/((ECm−BDEm)−Hm)	0.727	
	3D	(Ndm/(FL−PC))/(FL+(RDm/ECm))	0.003	
	4D	(Ndm−WF)/((FL+Hm)+(MdLm−RSm))	-0.034	
	1D	((FL/*q*)−RSm)*(ECm+(FL*MdLm))	0.147	-0.656
	2D	(Rm/(PC+WF))/((ECm−BDEm)−Hm)	0.636	
	3D	(Ndm/(FL−PC))/(FL+(RDm/ECm))	0.004	
	4D	(Ndm−WF)/((FL+Hm)+(MdLm−RSm))	-0.035	
	5D	(FL+Rm)/((WF/*q*)+(Rm+RSm))	0.032	
08	1D	((RDm−Rm)/*q*)+(RSm/(WF−FL))	-3.045	2.735
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.225	-1.038
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.067	
	1D	(*q*+ECm)+((FL*ENm)*(FL+RSm))	-0.232	1.049
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.068	
	3D	(ECm/DC)/((Hm*BDEm)−(Ndsm/PC))	0.123	
	1D	(*q*+ECm)+((FL*RSm)*(FL+ENm))	-0.245	-1.167
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.063	
	3D	(ECm/DC)/((Hm*BDEm)−(Ndsm/PC))	0.110	
	4D	(FL/(FL+MdNm))/((DC/PC)−FL)	0.161	
	1D	(*q*+ECm)+((FL+Rm)*(FL*ENm))	-0.256	-1.151
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.063	
	3D	(ECm/RDm)/((Hm*BDEm)−(Ndsm/PC))	-0.016	
	4D	((MdLm−WF)/ECm)/((FL*PC)−DC)	-0.661	
	5D	(ECm/Rm)/((FL+ECm)+(Lm+EIm))	-0.031	
	1D	((FL*BDEm)/(FL−BDEm))−(FL+ECm)	0.283	-1.074

	1D	(q+ECm)+((FL*ENm)*(FL+RSm))	-0.220	-0.980
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.068	
	1D	((FL*BDEm)/(q+FL))−(ECm/Rm)	0.381	-0.913
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.061	
	3D	((PC+BDEm)/PC)/(WF+(Lm/BDEm))	-1.051	
09	1D	((FL*BDEm)/(q+FL))−(ECm/Rm)	0.393	-0.933
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.063	
	3D	((PC+BDEm)/PC)/(WF+(Lm/BDEm))	-1.102	
	4D	(MdLm−Ndsm)/((PC−Hm)−(PC/q))	0.021	
	1D	((FL*BDEm)/(q+FL))−(ECm/Rm)	0.405	-0.918
	2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.063	
	3D	((PC+BDEm)/PC)/(WF+(Lm/BDEm))	-1.119	
	4D	(MdLm−Ndsm)/((PC−Hm)−(PC/q))	0.019	
	5D	(q+Ndm)/((Lm*ENm)+(WF−MdNm))	0.068	
	1D	(WF/(DC+ECm))/(PC+(FL*MdNm))	-4.011	1.709
	1D	(WF/(q*ENm))/(PC+(FL*MdNm))	-1.723	1.743
	2D	(Ndsm/(FL−PC))/(q+(MdLm/MdNm))	-0.005	
	1D	(WF/(DC+ECm))/(PC+(FL*MdNm))	-4.029	1.760
	2D	((DC−PC)/ECm)/((DC*MdNm)−ECm)	0.232	
	3D	(Lm+BDEm)/((EIm−Ndsm)−(FL*Ndsm))	0.075	
10	1D	(WF/(DC+ECm))/(PC+(FL*MdNm))	-3.987	1.776
	2D	((DC−PC)/ECm)/((ECm/MdNm)−DC)	-0.152	
	3D	(Lm+BDEm)/((EIm−Ndsm)−(FL*Ndsm))	0.078	
	4D	(Ndm/Lm)/((ECm*MdNm)−(Lm−WF))	0.123	
	1D	(WF/(DC+ECm))/(PC+(FL*MdNm))	-4.079	1.765

2D	((DC−PC)/ECm)/((DC*MdNm)−ECm)	0.244
3D	(Lm+BDEm)/((EIm−Ndsm)−(FL*Ndsm))	0.073
4D	(Ndsm−WF)/((FL*ECm)−(Ndsm/PC))	-0.135
5D	(Lm−PC)/((DC+FL)−(Ndm/Hm))	0.098

Table S5. The selected feature set of the 5D-SISSO model during validation process.

Iteraction	5D
01	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, Rm, RDm, Ndm, EIm, RSm
02	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, Rm, RDm, Ndm, EIm
03	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, Rm, RDm, EIm
04	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, Rm, RDm, MdNm
05	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, Rm, Ndm, MdNm
06	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, Lm, EIm, Ndm, RSm
07	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, MdNm, Ndm, RSm
08	WF, ECm, FL, PC, Hm, BDEm, *q*, DC, Ndsm, ENm, MdLm, MdNm, Rm, RSm
09	WF, ECm, FL, PC, Hm, BDEm, *q*, RDm, Rm, Ndm, MdLm, RSm
10	WF, ECm, FL, PC, Hm, BDEm, DC, Ndsm, Ndm, MdNm, Lm, EIm

Supplementary References

1. Sutton, C. et al. Identifying domains of applicability of machine learning models for materials science. *Nature Communications* **11**, 1-9 (2020).
2. Kanungo, T. et al. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence* **24**, 881-892 (2002).
3. Han, Z.-K. et al. Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence. *Nature Communications* **12**, 1-9 (2021).
4. Goldsmith, B.R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L.M. Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics* **19**, 013031 (2017).