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

Liangliang Xu
Korea Advanced Institute of Science and Technology
Ning Xu
Zhejiang University
Yue Wang
Hanyang University
Xiaojuan Hu
Fritz-Haber-Institute of the Max Planck Society
Zhongkang Han (hanzk@zju.edu.cn)
Zhejiang University https://orcid.org/0000-0003-1489-6824

Article

Keywords:

Posted Date: June 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3042812/v1

Additional Declarations: There is **NO** Competing Interest.

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

Abstract

Ceria has been extensively utilized in different fields, with surface oxygen vacancies playing a central role. However, versatile oxygen vacancy regulation is still in its infancy. In this work, we propose an effective strategy to manipulate the oxygen vacancy formation energy via transition metal doping by combining first-principles calculations and analytical learning. We elucidate the underlying mechanism driving the formation of oxygen vacancies using combined symbolic regression and data analytics techniques. The results show that the Fermi level of the system, along with the electronegativity of the dopants, are the paramount parameters (features) influencing the formation of oxygen vacancies. These insights not only enhance our understanding of the oxygen vacancy formation mechanism in ceria-based materials to improve their functionality but also potentially lay the groundwork for future strategies in the rational design of other transition metal oxide-based catalysts.

Introduction

Ceria-based materials are the key components in catalysts used for automotive exhaust treatment and hydrogen purification, as well as in many other functional materials, such as oxide-ion conductors and sensors.[1-5] In all such applications, the formation of oxygen vacancy (O_v) is a critical step. These O_vs serve as anchor sites to stabilize small transition metal (TM) clusters and prevent them from aggregating in ceria-supported TM catalysts.[6-8] Two electrons remaining after the removal of one neutral oxygen localize in the empty 4*f* orbital of Ce,[9, 10] resulting in the reduction of two Ce cations from the +4 oxidation state to the +3 oxidation state; this reduction changes the surface chemistry of the ceria and consequently its catalytic performance.[11, 12] Thus, the effective regulation of the O_v formation energy is extremely important, especially in catalytic applications.

In recent decades, tremendous efforts have been devoted to designing and synthesizing TM oxides with optimized O_v populations and distributions. Some examples of such efforts include modifying the O_v dynamic behavior and concentration of CeO_2 by suitable doping and strong metal-support interactions;[13-18] tuning the oxygen storage capacity and O_v formation energy of CeO_2 through large biaxial strain;[19] and altering the formation, distribution, and mobility of O_v and ion defects of CeO_2 via dislocation.[20] TM doping has been thoroughly utilized to manipulate the surface chemistry and O_v behaviors of ceria-based catalysts.[21-23] Ke and coworkers proposed a robust approach to tune the CO oxidation activity over CeO_2 nanowires through careful modulation of local structure and surface states by doping lanthanides with continuous variation in the ionic radii.[21] Zhang *et al*. successfully synthesized CeO_2 solid solutions with an ultrahigh manganese doping concentration and found that highly reducible Mn^{4+} ions strongly facilitate the migration of O_v from the bulk to the surface.[22] Jiang and coworkers found that cobalt-doped ceria nanosheets possess high efficiency toward the hydrogen evolution reaction due to the increased concentration of O_v and the increased number of active sites.[23]

Theoretical calculations also play an essential role in investigating the origins of the increased performance of TM-doped ceria materials.[24-28] Guo *et al*. investigated the effects of Pd, Ru, and Cu dopants on the reduction selectivity of CO_2 on CeO_2(111) and found that the doped metals distorted the local geometry of the ceria surface by weakening the Ce-O bond and thus generating highly active O_v sites.[24] Yang *et al*. investigated the CO oxidation reaction on Cu-adsorbed and Cu-doped ceria systems and found that the Cu-adsorbed system inhibited the formation of O_v, while the Cu-doped system facilitated the formation of O_v.[26] Krcha and coworkers systematically calculated the O_v formation energies and reaction activity for hydrocarbon conversion on various TM-doped CeO_2(111) surfaces, and clear periodic trends in the O_v formation energies and C-H bond activations were observed.[25] Despite extensive efforts, the mechanisms underlying the changes in the dopant-induced O_v formation energy remain unclear.

In this work, we performed systematic calculations on the stabilities and surface O_v formation energies for 97 TM-doped ceria systems, with dopants occupying the surface cerium sites. An in-depth analysis was conducted to explore the correlations between the properties of these systems (including stabilities and surface O_v formation energies) and the features of the dopants, employing a combined approach of sure independence screening and sparsifying operator (SISSO)[29] and

subgroup discovery (SGD) methodologies.[30-35] These techniques facilitated the unveiling of the underlying mechanisms that drive the observed trends in the data.

Results and Discussion

The stability of catalysts is as significant as their catalytic performance. To begin, we conducted a comprehensive investigation into the stabilities of 97 TM-doped ceria systems, including those of $CeO_2(100)$, $CeO_2(110)$, and $CeO_2(111)$. We considered 30 different *d*-block TM dopants. As shown in Figure 1, the TM dopants were positioned at the surface cerium sites of the ceria substrates. The formation energies of the TM-doped ceria systems are collected in Table S1-S3. Given the existence of multiple localizations of Ce^{3+} polarons,[36] we have taken care to include a variety of Ce^{3+} polaron configurations in our dataset to ensure that our analytical learning approaches adequately capture these features (refer to Figure S1).

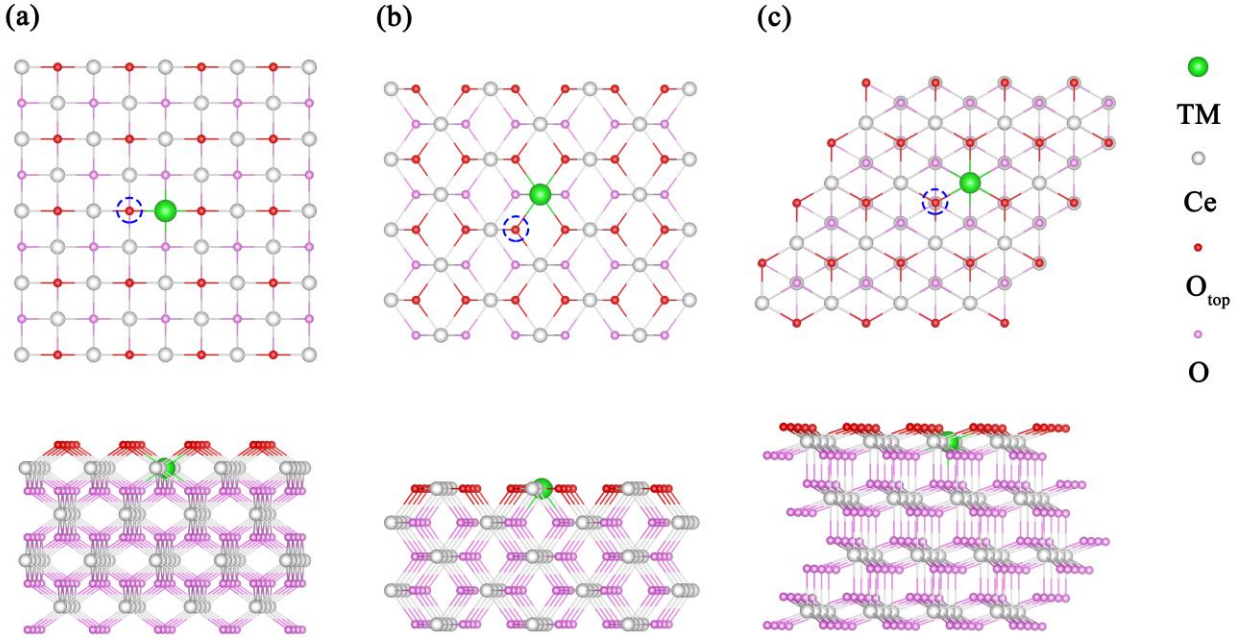

Figure 1. Structural models of the TM-doped ceria systems. The dopant is positioned at the top cerium site of (a) $CeO_2(100)$, (b) $CeO_2(110)$, and (c) $CeO_2(111)$ surface. The oxygen vacancy sites under consideration are highlighted with dashed circles.

To evaluate the impact of TM dopants on the stability of various TM-doped ceria systems, we systematically explored the correlations between the formation energies of the TM-doped ceria systems (the property) and the characteristics of the dopants (the features) using analytical learning techniques. The proper selection of initial features is of paramount importance in analytical learning. The guiding principles for initial feature selection are as follows: (a) the features can be obtained or calculated in convenient ways, and (b) the features are already evaluated and shown to have close correlations with properties investigated either experimentally or theoretically. We considered two classes of features: fourteen atomic features and five system-based features, as detailed in Table 1. The Pearson's correlation coefficient matrix among the considered features of the TM-doped ceria systems was checked first, with the results summarized in Figure 2a-d. The results reveal that certain features, including Ndsm, Ndm, Mm, and Rm, are closely correlated with each other for the subgroup data set encompassing 3*d*-, 4*d*-, and 5*d*-block metals (Figure 2a-

c). However, their correlation significantly diminished for the entire data set comprising all *d*-block metals (Figure 2d). This observation suggests a high degree of diversity within the primary feature space, which is well-suited for subsequent analytical learning analyses.

Table 1. The nineteen considered features of the TM dopants.

class	name	abbreviation
atomic	First ionization energy (eV)	EIm
	Highest occupied molecular orbital (eV)	Hm
	Lowest unoccupied molecular orbital (eV)	Lm
	Atomic radius (empirical) (Å)	Rm
	Bond dissociation energy of homo-diatomic molecules (eV)	BDEm
	Cohesive energy (eV)	ECm
	Radius of *s*-orbital (Å)	RSm
	Radius of *d*-orbital (Å)	RDm
	Electronegativity	ENm
	Number of *d* valence electrons	Ndm
	Total number of *d* and *s* valence electrons	Ndsm
	Miedema parameters (volt)	MdLm
	Miedema parameters (density unit)	MdNm
	Relative atomic mass (a.m.u.)	Mm
system	Bader charge of the dopant (e^-)	*q*
	Fermi level (eV)	FL
	Work function (eV)	WF
	d-band center of dopant (eV)	DC
	p-band center of oxygen (eV)	PC

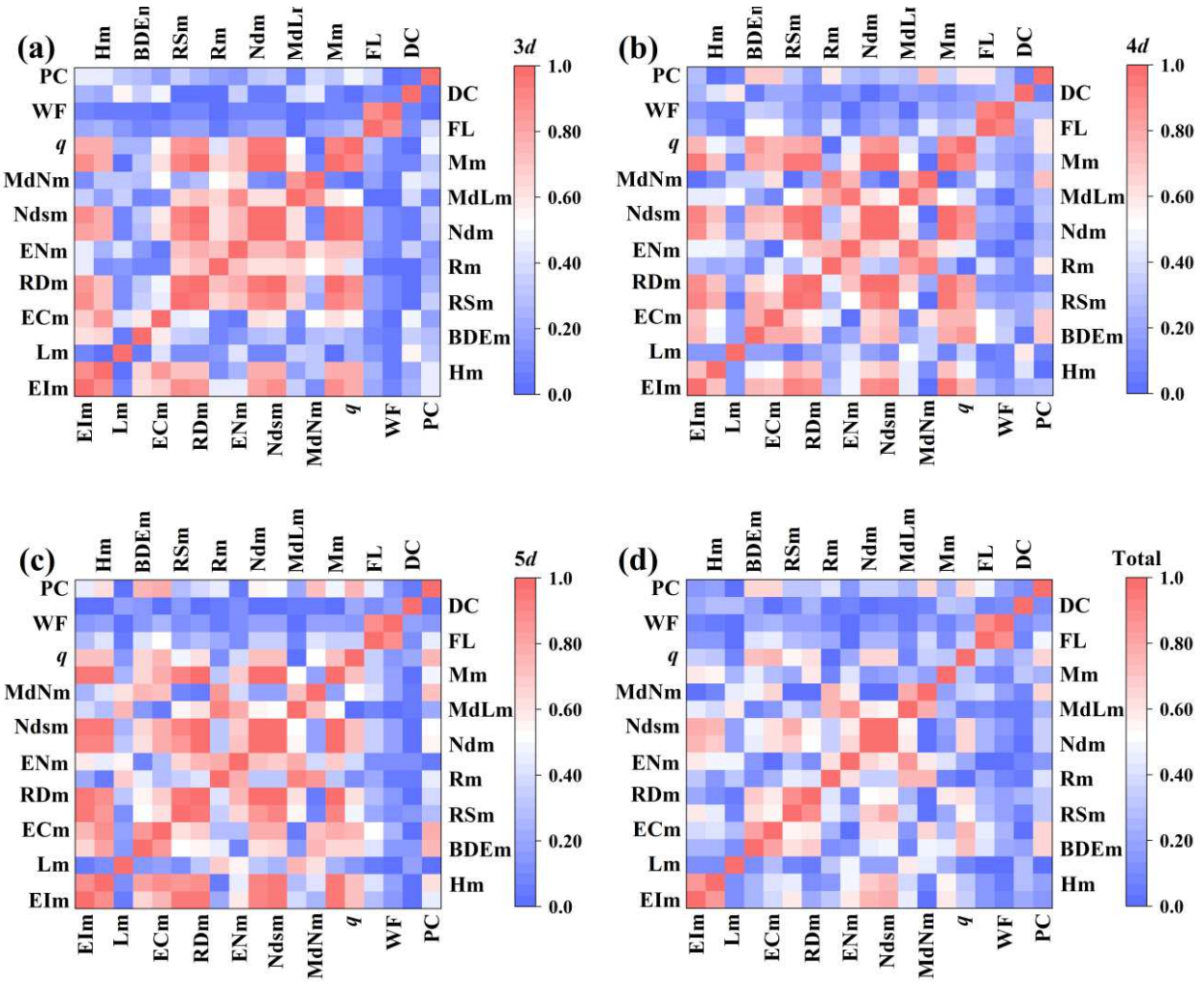

Figure 2. Heat map of the Pearson's correlation coefficient matrix among the nineteen features of the (a) 3*d*-, (b) 4*d*-, (c) 5*d*-, and (d) total *d*-block transition metals for the TM-doped ceria systems.

An in-depth investigation was conducted to explore the correlations between the primary features and the formation energies of the TM-doped ceria systems. The findings are presented in Figure 3 and Figure S2, which reveals that the formation energies exhibit a continuous distribution within a wide energy range (-12.0 eV to 3.0 eV) and display clear periodic trends that are associated with the number of *d* electrons (Ndm) and the radius of *d* orbitals (RDm) of the TM dopants.[37] The discerned linear relationship between Ndm and ΔE_F can be interpreted based on the significant influence the *d*-orbitals of TM atoms exert on their interaction with oxygen. Transition metals are characterized by their distinct numbers of *d*-electrons, and those with fewer *d*-electrons tend to have less strong correlations within their *d*-orbitals. This reduced correlation promotes the delocalization of the *d*-orbitals, which in turn makes it easier for other atoms to extract electrons and consequently form bonds. Furthermore, the formation of a covalent bond necessitates an effective overlap of orbitals with matching energies. Transition metals with fewer *d*-electrons generally have *d*-orbitals with energy levels closer to the 2*p* orbitals of oxygen. This facilitates the formation of stronger covalent bonds due to enhanced orbital overlap. These combined factors contribute to the strong binding between TM dopants with a small number of *d*-electrons and the lattice oxygen within ceria. This, in turn, stabilizes the TM-doped ceria systems. Similarly, an

increase in the size of *d*-orbitals can lead to enhanced orbital overlap with the oxygen's *p*-orbitals. If the *d*-orbitals of a metal atom are larger, they hold a greater potential for overlapping with the *p*-orbitals of the oxygen atom, leading to the creation of a stronger covalent bond. These observations provide a crucial understanding of the key factors contributing to the stabilities of TM-doped ceria systems.

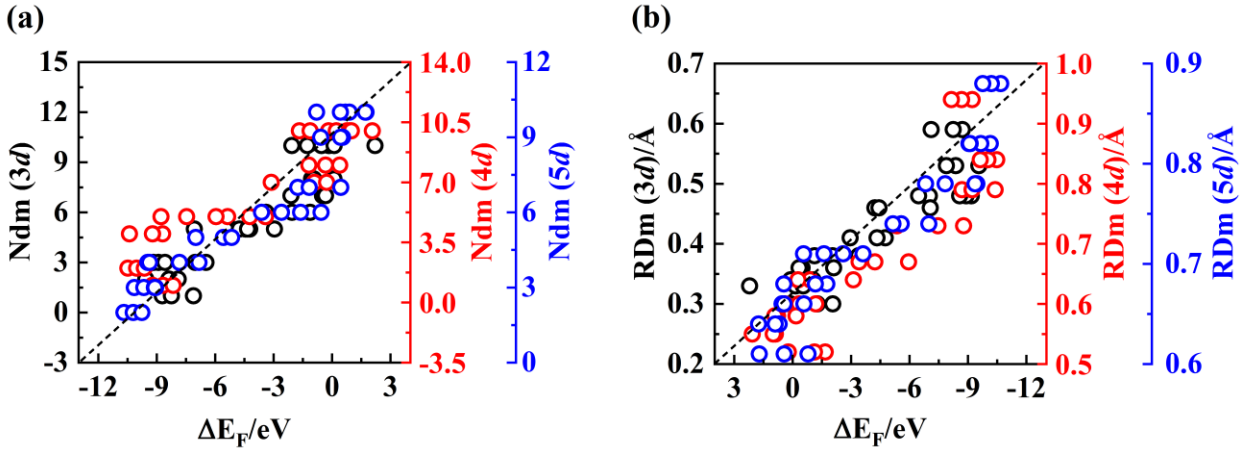

Figure 3. Correlation between the formation energies (ΔE_F) of TM-doped ceria systems and features of transition metal dopants. (a) The correlation between ΔE_F and Ndm, and (b) the correlation between ΔE_F and RDm.

We further explored the correlation between the surface O_v formation energies of the TM-doped ceria systems and the primary features. The corresponding results are displayed in Figure S3. It is evident that the surface O_v formation energies of the TM-doped ceria systems exhibit much weaker correlations with the primary features than that of the formation energies of the TM-doped ceria systems. Thus, more complex features are required to describe the O_v formation energies of the TM-doped ceria systems. To identify the key descriptive features, we employed the state-of-the-art compressed-sensing based approach SISSO. The SISSO approach enables us to identify the best low-dimensional descriptor in an immensity of offered candidates. The SISSO model comprises a linear combination of a few derived features, which are typically non-linear expressions produced by applying mathematical operations on and between the primary features. These derived features are selected from a massive set of more than one billion candidate derived features, with feature complexity (the number of operators in each derived feature) of up to five in this work. Figure S4 illustrates that the five-dimensional (5D) SISSO model, which incorporates five derived features, demonstrates a high level of accuracy and predictive power, as validated through 10 iterations of a leave-5%-out validation method (Table S4). Therefore, to maintain the simplicity of the model and to prevent overfitting, we decided not to explore higher-dimensional models. Interestingly, throughout the 10 iterations of the leave-5%-out validation, a similar set of primary features was consistently identified, especially the primary features WF, ECm, FL, PC, Hm, and BDEm, all of which were consistently selected (Table S5). It's not surprising that the systems-based features WF, FL, and PC were chosen, considering their strong correlation with the systems' redox abilities, a factor that directly influences the formation and healing of O_v. Moreover, features like ECm and BDEm encapsulate the interaction characteristics of the TM dopant with other species, while Hm captures the electron sharing capacity of the TM dopants with other species. All these features were consistently selected throughout the validation process, indicating

the SISSO model's capability to capture the underlying physics that govern the formation of O_v. The frequent selection of these descriptor components, despite changes in system size and composition, underscores the robustness of the model, affirming its reliability in comprehending such complex systems.

Table 2. Components of descriptors and their coefficient(C) and intercept(I) of the SISSO-identified best model.

Dimension	Descriptor	C	I
1D	(q+ECm)+(FL*ENm)*(FL+RSm)	-0.222	
2D	(ECm+MdLm)/((DC/ECm)+(WF−MdLm))	0.058	
3D	(BDEm−ENm)/((Hm+PC)−(q*WF))	0.134	-1.055
4D	(MdNm/q)/((Hm*BDEm)−(Ndsm/PC))	-0.136	
5D	(q+BDEm)/((Lm+Rm)+(FL*RSm))	-0.034	

The components of the optimal model identified by the SISSO approach are presented in Table 2. Additionally, Figure 4 displays the error distribution of the O_v formation energies as predicted by the optimal SISSO model, alongside the distribution of the O_v formation energies themselves. While the SISSO model provides mathematical formulas, gaining a deeper understanding of the mechanisms underlying the formation of O_vs at various TM-doped ceria systems using the SISSO model remains a challenging task. In an attempt to identify the key descriptive parameters and facilitate physical understanding of the actuating mechanisms, we applied the SGD local artificial intelligence approach.[32] SGD is a method designed to identify local patterns in data that maximize a specific quality function. The identified pattern/subgroup is defined as an intersection of simple inequalities involving provided features.[38] The so-called quality function measures the degree of distinctiveness of a pattern/subgroup is comparison to the entire data set (further details about the implementation of the SGD approach can be found in the Supporting Information).

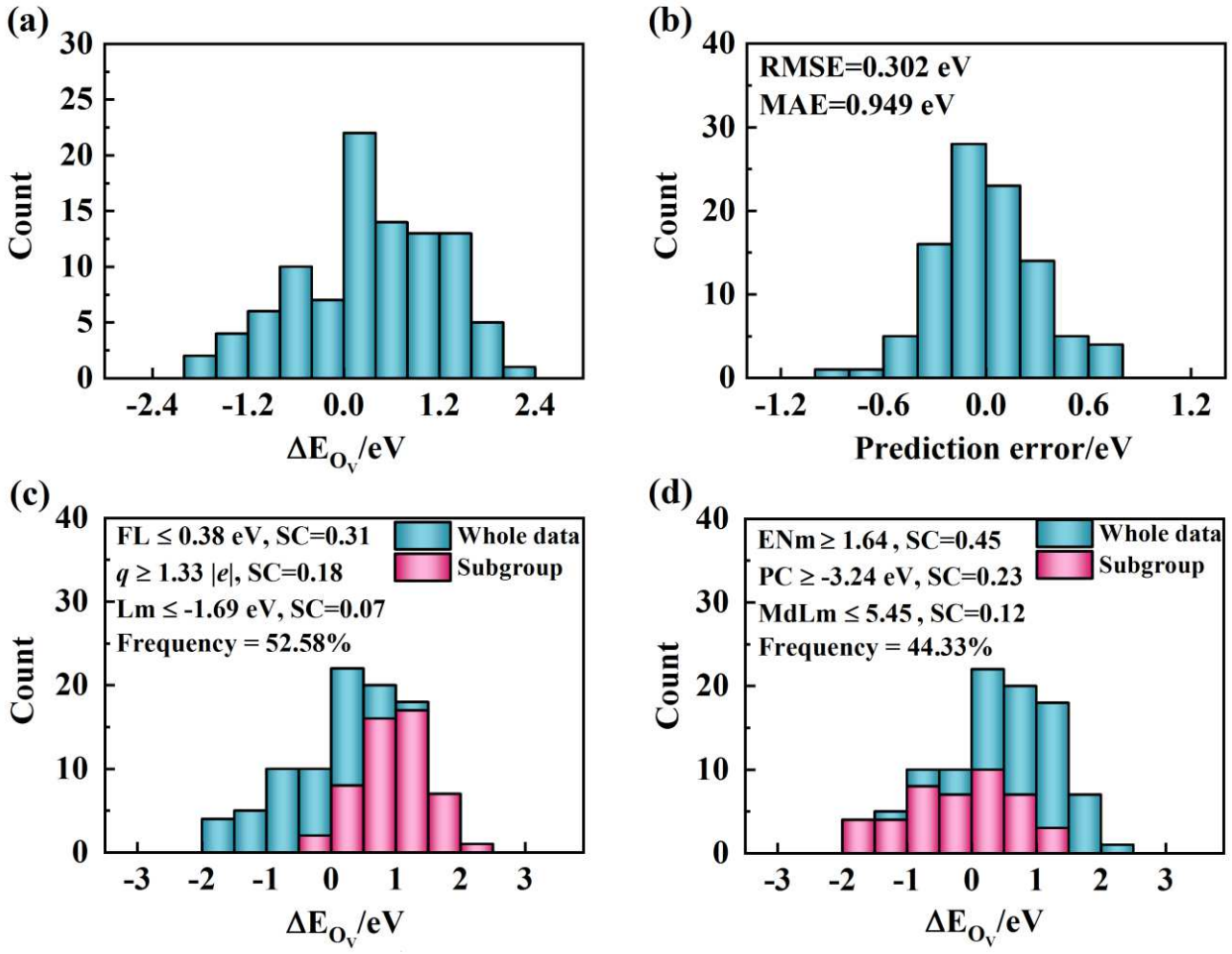

Figure 4. Analytical learning of surface O_v formation energies of the TM-doped ceria systems. (a) The histogram distribution of data samples for O_v formation energies of the TM-doped ceria systems, (b) the error distribution between the SISSO model predicted energies and DFT calculated energies for the TM-doped ceria systems, and the results of the SGD by maximizing (c) and minimizing (d) the surface O_v formation energies.

To gain better insight into the factors and mechanisms that facilitate the formation of O_v on various TM-doped ceria systems, we utilized the SGD approach. Starting with the identification of subgroups that maximize the O_v formation energies, one specific subgroup was singled out. This subgroup, which consists of 51 data points or 52.58% of the total dataset as depicted in Figure 4c, was distinguished by the conditions FL $\leqslant$ 0.38 eV, $q \geqslant$ 1.33 |e|, and Lm $\leqslant$ -1.69 eV. On the other hand, the identification of subgroups that minimize O_v formation energies led to the discovery of another subgroup. This subgroup contained 43 data points or 44.33% of the total dataset as shown in Figure 4d, and was characterized by the conditions ENm $\geqslant$ 1.64, PC $\geqslant$ -3.24 eV, and MdLm $\leqslant$ 5.45. The importance of each feature was determined by its respective support score (SC), which is derived from the quality function value (Q). Initially, a specific feature is removed from the entire feature set, then Q is recalculated with the remaining features. The score is calculated using the equation [1–Q (with all features)]/Q (with all features minus the removed one). A score of 0 suggests that removing this specific feature does not affect the value of Q, indicating that this feature does not influence the SGD result. Among these features, FL $\leq$ 0.38 eV and $q \geq$ 1.33 |e|

stood out for their relatively high SCs, which the SGD approach identified as key in inhibiting O_v formation. In contrast, features ENm $\geqslant$ 1.64 and PC $\geqslant$ -3.24 eV were noted to facilitate O_v formation. In a physical context, a lower Fermi lever (FL $\leq$ 0.38 eV) implies that the highest occupied 2*p* orbitals of lattice oxygen should be situated at relatively lower energy levels. As a result, electron transfer from the removed neutral oxygen to the empty *d* orbitals of the TM dopants or the *f* orbitals of the lattice cerium site becomes less feasible, inhibiting surface O_v formation. Similarly, the condition $q \geq 1.33$ |e| can be interpreted by the fact that four positive charges ($4e^+$) are retained when one neutral cerium atom is removed. If these positive charges are not adequately compensated by the TM dopant (as can occur when the *q* is small), the remaining positive charges will form empty hole states that can combine with the electrons left after oxygen removal, thereby facilitating O_v formation. Thus, a larger *q* is crucial to suppress O_v formation. Furthermore, a larger dopant electronegativity (ENm $\geqslant$ 1.64) and a larger *p* orbital center (PC $\geqslant$ -3.24 eV) of the lattice oxygen can facilitate O_v formation. Dopants with lower electronegativity, due to their large electronegativity difference, bind strongly with the lattice oxygen, inhibiting O_v formation. On the other hand, a higher *p* orbital center eases the electron transfer from oxygen to TM sites upon oxygen removal, promoting O_v formation. With the help of SGD, a sophisticated understanding was developed that would have been challenging to achieve without local artificial intelligence. SGD has proven useful in identifying multiple alternative subgroups, each corresponding to different mechanisms affecting the target properties. The target properties with desired values that lie in a specific range are also a factor affecting the identified subgroups.[38] Exploring the role of different factors can significantly improve the interpretability of the subgroups.

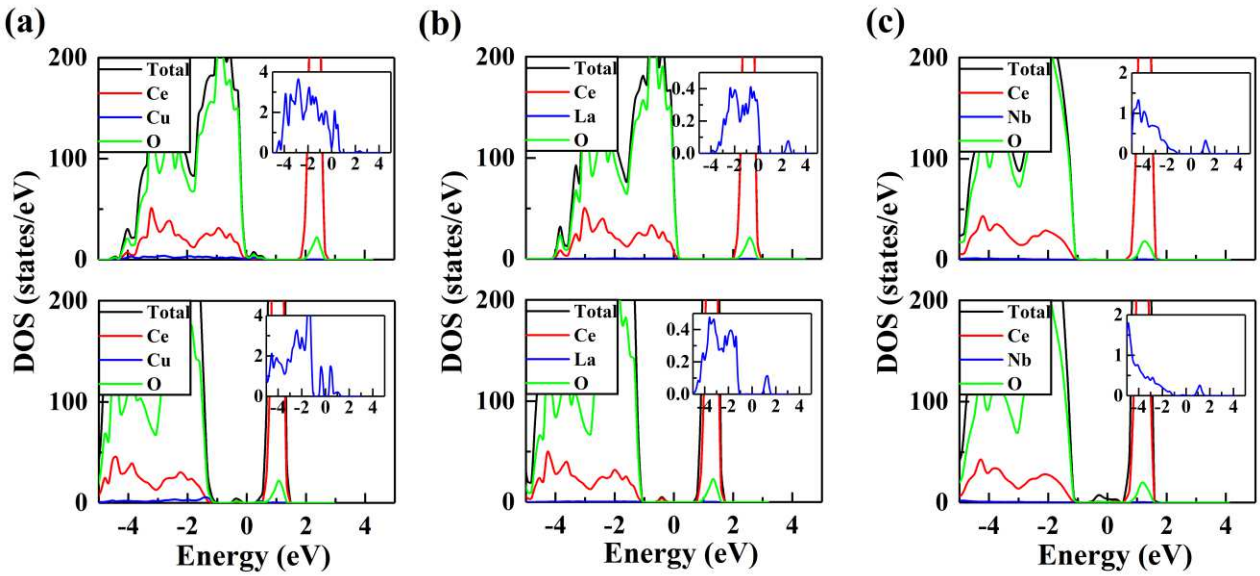

Figure 5. Species projected density of states for surface doped CeO_2(111) surfaces. The TM-doped CeO_2(111) surfaces with low O_v formation energy (a, Cu doping), moderate O_v formation energy (b, La doping), and high O_v formation energy (c, Nb doping).

The interpretation of SGD results is sometimes limited by our domain knowledge. To validate the abovementioned explanations, we further analyzed the species projected density of states of three represent CeO_2(111) systems with low (Cu-doped), moderate (La-doped), and high (Nb-doped) O_v formation energies (Figure 5). For the Cu-doped surface, pronounced 3*d* empty states of Cu exist

near the Fermi level due to the relatively large *q* of Cu (-1.10 e^-) (Figure 5a). These empty states will be occupied by extra electrons to facilitate the removal of neutral oxygen and the formation of O_vs. For the Nb-doped surface, there are no empty *d* orbitals near the Fermi level due to the relatively small *q* of Nb (-2.58 e^-). The electrons remaining after removal of neutral oxygen must occupy the 4*f* states of the cerium atoms, splitting from the stable unoccupied 4*f* states of cerium atoms, which makes O_v formation difficult (Figure 5c). The density of empty states decreases in the order of Cu-doped > La-doped > Nb-doped surfaces, resulting in O_v formation energies that increase in the order of Cu-doped < La-doped < Nb-doped surfaces.

Conclusion

In summary, by synergistically utilizing first-principles calculations and an interpretable artificial intelligence approach, we have successfully identified the key descriptive features and mechanisms that facilitate the formation of oxygen vacancies at various transition metal-doped CeO_2 systems. The Fermi level of the system, along with the electronegativity of the dopants, are the paramount parameters (features) influencing the formation of oxygen vacancies. By integrating domain knowledge, symbolic regression, subgroup discovery, and electronic structure analysis, the facilitated formation mechanism of oxygen vacancies was attributed to the distributions and behaviors of occupied 2*p* states of lattice oxygen and the empty states of the transition metal dopants near the Fermi level. Our methodology can be easily adapted to other transition metal oxide-based functional materials. It demonstrates that powerful potential of artificial intelligence to provide new understanding and insights into the physical principles underlying data.

Methods

Spin-polarized DFT calculations were carried out using the generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof (PBE) as implemented in the Vienna *ab initio* Simulation Package (VASP).[39, 40] A Hubbard-type term U = 5.0 eV was used for the Ce 4*f* states to describe the localization of the excess charges, which is within the range of suitable values to describe reduced ceria-based systems.[41] The Kohn-Sham valence states were expanded in a plane-wave basis set with a cutoff energy of 400 eV. The core-valence interactions were represented using the projector augmented wave (PAW) approach,[42] where the (4*f*, 5*s*, 5*p*, 5*d*, 6*s*) and (2*s*, 2*p*) states of the Ce and O atoms were treated as valence states. A periodic slab with a (4×4), (3×4), and (4×4) surface unit cell was employed to model the CeO_2(100), CeO_2(110), and CeO_2(111) surfaces, respectively. The slabs consisted of nine, five, and twelve atomic layers, respectively, with an additional vacuum layer of 15 Å. During the geometry optimizations, the bottom two layers (O-Ce), three layers (O-Ce-O), and three layers (O-Ce-O) were held fixed at their bulk positions for CeO_2(100), CeO_2(110), and CeO_2(111) surfaces, respectively. Due to the large supercell dimensions, the k-point sampling was restricted to the Γ point.

The formation energy (ΔE_F) of the TM-doped CeO_2 surfaces was calculated using formula (1):

$$\Delta E_F = E_{TM-slab} - E_{slab} + \frac{E_{bulk}^{CeO_2}}{N_{bulk}^{CeO_2}} - E_{O_2} - \frac{E_{bulk}^{TM}}{N_{bulk}^{TM}} \quad (1)$$

where $E_{TM-slab}$ is the total energy of the whole system, E_{slab} is the energy of the clean CeO_2 surfaces, $E_{bulk}^{CeO_2}$ is the energy of the conventional cell of ceria, $N_{bulk}^{CeO_2}$ is the number of O-Ce-O units in the conventional cell, E_{O_2} is the total energy of the gas phase O_2 molecule, E_{bulk}^{TM} is the

energy of the conventional cell of TM, and $N_{bulk}^{CeO_2}$ is the number of TM atoms in the conventional cell.

The O_v formation energy of the TM-doped CeO_2 surfaces was calculated by means of formula (2):

$$\Delta E_{O_v} = E_{O_v} - E_{TM-slab} + \frac{1}{2}E_{O_2} \quad (2)$$

where E_{O_v}, $E_{TM-slab}$, and E_{O_2} are the total energies of the TM-doped CeO_2 surfaces with oxygen vacancies, the TM-doped CeO_2 surfaces, and the gas-phase O_2 molecule, respectively.

Subgroup discovery (SGD) was performed using the RealKD package.[43] Each feature was split into 14 subsets using a 14-means clustering algorithm.[44] The candidate subgroups were built as conjunctions of obtained simple inequalities. The main idea of SGD is that the subgroups are unique if the distribution of the data they contain is as different as possible from the data distribution in the whole sample. Detailed information about the SGD approach can be found in the Supplementary Methods.

References

1. Fu, Q., Saltsburg, H. & Flytzani-Stephanopoulos, M. Active nonmetallic Au and Pt species on ceria-based water-gas shift catalysts. *Science* **301**, 935-938 (2003).
2. Deluga, G., Salge, J., Schmidt, L. & Verykios, X. Renewable hydrogen from ethanol by autothermal reforming. *Science* **303**, 993-997 (2004).
3. Campbell, C.T. & Peden, C.H. Oxygen vacancies and catalysis on ceria surfaces. *Science* **309**, 713-714 (2005).
4. Rodriguez, J.A. et al. Activity of CeO_x and TiO_x nanoparticles grown on Au(111) in the water-gas shift reaction. *Science* **318**, 1757-1760 (2007).
5. Park, S., Vohs, J.M. & Gorte, R.J. Direct oxidation of hydrocarbons in a solid-oxide fuel cell. *Nature* **404**, 265-267 (2000).
6. Bruix, A. et al. A New Type of Strong Metal–Support Interaction and the Production of H_2 through the Transformation of Water on $Pt/CeO_2(111)$ and $Pt/CeO_x/TiO_2(110)$ Catalysts. *J. Am. Chem. Soc.* **134**, 8968-8974 (2012).
7. Carrasco, J. et al. In situ and theoretical studies for the dissociation of water on an active Ni/CeO_2 catalyst: importance of strong metal–support interactions for the cleavage of O–H bonds. *Angew. Chem., Int. Ed.* **54**, 3917-3921 (2015).
8. Zhang, S. et al. Strong electronic metal-support interaction of Pt/CeO_2 enables efficient and selective hydrogenation of quinolines at room temperature. *J. Catal.* **359**, 101-111 (2018).
9. Ganduglia-Pirovano, M.V., Da Silva, J.L. & Sauer, J. Density-functional calculations of the structure of near-surface oxygen vacancies and electron localization on $CeO_2(111)$. *Phys. Rev. Lett.* **102**, 026101 (2009).
10. Li, H.-Y. et al. Multiple configurations of the two excess 4 f electrons on defective CeO_2 (111): origin and implications. *Phys. Rev. B* **79**, 193401 (2009).
11. Gao, Q., Hao, J., Qiu, Y., Hu, S. & Hu, Z. Electronic and geometric factors affecting oxygen vacancy formation on $CeO_2(111)$ surfaces: A first-principles study from trivalent metal doping cases. *Appl. Surf. Sci.* **497**, 143732 (2019).
12. Sravan Kumar, K.B. & Deshpande, P.A. On identification of labile oxygen in ceria-based solid solutions: which oxygen leaves the lattice? *J. Phys. Chem. C* **119**, 8692-8702 (2015).

13. Sen, S., Edwards, T., Kim, S.K. & Kim, S. Investigation of the Potential Energy Landscape for Vacancy Dynamics in Sc-Doped CeO_2. *Chem. Mater.* **26**, 1918-1924 (2014).
14. Wang, Z., Fu, H., Tian, Z., Han, D. & Gu, F. Strong metal–support interaction in novel core–shell Au–CeO_2 nanostructures induced by different pretreatment atmospheres and its influence on CO oxidation. *Nanoscale* **8**, 5865-5872 (2016).
15. Zheng, Y. et al. Ultra-Fine CeO_2 Particles Triggered Strong Interaction with $LaFeO_3$ Framework for Total and Preferential CO Oxidation. *ACS Appl. Mater. Interfaces* **12**, 42274-42284 (2020).
16. Han, Z.-K., Wang, Y.-G. & Gao, Y. Catalytic role of vacancy diffusion in ceria supported atomic gold catalyst. *Chem. Commun.* **53**, 9125-9128 (2017).
17. McFarland, E.W. & Metiu, H. Catalysis by doped oxides. *Chem. Rev.* **113**, 4391-4427 (2013).
18. Nolan, M. Enhanced oxygen vacancy formation in ceria (111) and (110) surfaces doped with divalent cations. *J. Mater. Chem.* **21**, 9160-9168 (2011).
19. Gopal, C.B. et al. Equilibrium oxygen storage capacity of ultrathin $CeO_{2-\delta}$ depends non-monotonically on large biaxial strain. *Nat. Commun.* **8**, 1-12 (2017).
20. Sun, L., Marrocchelli, D. & Yildiz, B. Edge dislocation slows down oxide ion diffusion in doped CeO_2 by segregation of charged defects. *Nat. Commun.* **6**, 1-10 (2015).
21. Ke, J. et al. Dopant-induced modification of active site structure and surface bonding mode for high-performance nanocatalysts: CO oxidation on capping-free (110)-oriented CeO_2: Ln (Ln= La–Lu) nanowires. *J. Am. Chem. Soc.* **135**, 15191-15200 (2013).
22. Zhang, P. et al. Mesoporous $MnCeO_x$ solid solutions for low temperature and selective oxidation of hydrocarbons. *Nat. Commun.* **6**, 1-10 (2015).
23. Jiang, S. et al. Promoting formation of oxygen vacancies in two-dimensional cobalt-doped ceria nanosheets for efficient hydrogen evolution. *J. Am. Chem. Soc.* **142**, 6461-6466 (2020).
24. Guo, C. et al. Initial reduction of CO_2 on Pd-, Ru-, and Cu-doped CeO_2(111) surfaces: effects of surface modification on catalytic activity and selectivity. *ACS Appl. Mater. Interfaces* **9**, 26107-26117 (2017).
25. Krcha, M.D., Mayernick, A.D. & Janik, M.J. Periodic trends of oxygen vacancy formation and C–H bond activation over transition metal-doped CeO_2(111) surfaces. *J. Catal.* **293**, 103-115 (2012).
26. Yang, Z., He, B., Lu, Z. & Hermansson, K. Physisorbed, chemisorbed, and oxidized CO on highly active Cu−CeO_2(111). *J. Phys. Chem. C* **114**, 4486-4494 (2010).
27. Riley, C. et al. Design of effective catalysts for selective alkyne hydrogenation by doping of ceria with a single-atom promotor. *J. Am. Chem. Soc.* **140**, 12964-12973 (2018).
28. Zhao, S. et al. Spontaneous formation of asymmetric oxygen vacancies in transition-metal-doped CeO_2 nanorods with improved activity for carbonyl sulfide hydrolysis. *ACS Catal.* **10**, 11739-11750 (2020).
29. 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. *Phys. Rev. Mater.* **2**, 083802 (2018).
30. S. Wrobel, An algorithm for multi-relational discovery of subgroups, European symposium on principles of data mining and knowledge discovery, Springer, pp. 78-87 (1997).
31. Friedman, J.H. & Fisher, N.I. Bump hunting in high-dimensional data. *Stat. Comput.* **9**, 123-143 (1999).

32. Atzmueller, M. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov.* **5**, 35-49 (2015).
33. Boley, M., Goldsmith, B.R., Ghiringhelli, L.M. & Vreeken, J. Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. *Data Min. Knowl. Discov.* **31**, 1391-1418 (2017).
34. Goldsmith, B.R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L.M. Uncovering structure-property relationships of materials by subgroup discovery. *New J. Phys.* **19**, 013031 (2017).
35. Mazheika, A. et al. Ab initio data-analytics study of carbon-dioxide activation on semiconductor oxide surfaces. *arXiv preprint arXiv:1912.06515* (2019).
36. Otero, G.S., Lustemberg, P.G., Prado, F. & Ganduglia-Pirovano, M.V. Relative stability of near-surface oxygen vacancies at the CeO_2(111) surface upon zirconium doping. *J. Phys. Chem. C* **124**, 625-638 (2019).
37. Hu, Z. & Metiu, H. Effect of dopants on the energy of oxygen-vacancy formation at the surface of ceria: Local or global? *J. Phys. Chem. C* **115**, 17898-17909 (2011).
38. Han, Z.-K. et al. Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence. *Nat. Commun.* **12**, 1-9 (2021).
39. Perdew, J.P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
40. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
41. Castleton, C., Kullgren, J. & Hermansson, K. Tuning LDA+ U for electron localization and structure at oxygen vacancies in ceria. *J. Chem. Phys.* **127**, 244704 (2007).
42. Blöchl, P.E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
43. Sutton, C. et al. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 1-9 (2020).
44. Kanungo, T. et al. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 881-892 (2002).

Acknowledgments

We acknowledge the financial support of the National Key Research and Development Program (2022YFA1505500), the Fundamental Research Funds for the Central Universities, and the National Research Foundation of Korea (NRF) grants funded by the Korean government (Grants RS-2023-00243788).

Author contributions

Z.-K.H. and X.J.H. created the idea and conceived the work. Z.-K.H. and X.J.H. designed and supervised the project. L.L.X. performed all the calculations. L.L.X. and N.X. wrote the manuscript with inputs from all the authors. All authors contributed to the analysis and interpretation of the results. All the authors commented on the manuscript and have given approval to the final version of the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary Information is available for this paper at http://www.nature.com/nature.

Correspondence and requests for materials should be addressed to Z.-K.H. or X.J.H.

Data Availability

All relevant data are available from the corresponding authors upon reasonable request.

Code Availability

SGD: https://bitbucket.org/realKD/realkd/
SISSO: https://github.com/rouyang2017/SISSO.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile1.dat
- SupportingInformation.pdf