Nanoscale

PAPER

View Article Online
View Journal | View Issue

Cite this: *Nanoscale*, 2025, **17**, 4531

Decoupling many-body interactions in the CeO_2(111) oxygen vacancy structure with statistical learning and cluster expansion†

Yujing Zhang,[a,b] Zhong-Kang Han, [c] Beien Zhu, [d] Xiaojuan Hu,[e] Maria Troppenz,[f] Santiago Rigamonti, [f] Hui Li,*[a] Claudia Draxl, [f] M. Verónica Ganduglia-Pirovano *[g] and Yi Gao *[d,h]

Oxygen vacancies (V_O's) are of paramount importance in influencing the properties and applications of ceria (CeO_2). Yet, comprehending the distribution and nature of V_O's poses a significant challenge due to the vast number of electronic configurations and intricate many-body interactions among V_O's and polarons (Ce^{3+} ions). In this study, we established a cluster expansion model based on first-principles calculations and statistical learning to decouple the interactions among the Ce^{3+} ions and V_O's, thereby circumventing the limitations associated with sampling electronic configurations. By separating these interactions, we identified specific electronic configurations characterized by the most favorable V_O–Ce^{3+} attractions and the least favorable Ce^{3+}–Ce^{3+}/V_O–V_O repulsions, which are crucial in determining the stability of vacancy structures. Through more than 10^8 Metropolis Monte Carlo samplings of V_O's and Ce^{3+} ions in the near surface of CeO_2(111), we explored potential configurations within an 8 × 8 supercell. Our findings revealed that oxygen vacancies tend to aggregate and are abundant in the third oxygen layer with an elevated V_O concentration primarily due to extensive geometric relaxation, an aspect previously overlooked. This work introduces a novel theoretical framework for unraveling the complex vacancy structures in metal oxides, with potential applications in redox and catalytic chemistry.

Received 4th November 2024,
Accepted 23rd December 2024

DOI: 10.1039/d4nr04591b

rsc.li/nanoscale

1. Introduction

Cerium oxide (CeO_2), known for its high concentration of surface oxygen vacancies (V_O's), exhibits remarkable oxygen storage capacity and redox catalytic performance, finding applications in diverse areas such as automobile exhaust gas treatment, hydrogen production, purification, and fuel cells.[1–10] The CeO_2 (111) crystal facet, known for its stability and accessibility, has been extensively studied.[11–19] Experiments involving scanning tunneling microscopy (STM) and atomic resolution dynamic force microscopy (DFM) have provided insights into the various near-surface V_O structures at the CeO_2(111) surface.[20,21] The theoretical interpretation of the distributions of these vacancies has proven to be a challenge.[22–30] Early studies identified the formation of a single V_O, leaving two electrons localized on the 4f orbitals of two Ce^{4+} ions, converting them into Ce^{3+} ions.[22] Li *et al.*[23] and Ganduglia-Pirovano *et al.*[24,25] suggested that the energetically favored location for an isolated oxygen vacancy is in the subsurface layer, with the excess electrons localized in two next-nearest neighbor (NNN) cations. This configuration was found to be approximately 0.2 eV more stable than the NNN of the surface vacancy and significantly lower in energy than the V_O in deeper oxygen layers. Further research explained the observed high stability of an ordered (2 × 2) oxygen subsurface vacancy structure at a V_O concentration of 1/4.[21,25] However, precise sampling of the V_O structures remains challenging due to the varied stability introduced by Ce^{3+} ions at different coordination spheres around the V_O's.[31] As the number of V_O's increases, the number of Ce^{3+} also rises, leading to an expo-

[a]*Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: hli@buct.edu.cn*
[b]*Key Laboratory of Interfacial Physics and Technology, Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201800, China*
[c]*School of Materials Science and Engineering, Zhejiang University, Hangzhou 310027, China*
[d]*Photon Science Research Center for Carbon Dioxide, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China. E-mail: gaoyi@sari.ac.cn*
[e]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany*
[f]*Institut für Physik und Iris Adlershof, Humboldt-Universität zu Berlin, Zum Großen Windkanal 2, 12489 Berlin, Germany*
[g]*Instituto de Catálisis y Petroleoquímica, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain. E-mail: vgp@icp.csic.es*
[h]*Key Laboratory of Low-Carbon Conversion Science & Engineering, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China*
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4nr04591b

nential increase in possible electronic configurations. The spatial correlation among Ce^{3+}'s and V_O's and the mutual spatial correlation between Ce^{3+}'s and V_O's become exceedingly complex and are closely tied to the stability of configurations. Understanding the nature of V_O structures and making accurate predictions about V_O/Ce^{3+} configurations require a precise evaluation of the stabilities of numerous electronic configurations, considering the interactions among Ce^{3+}–Ce^{3+}, V_O–Ce^{3+}, V_O–V_O, and their couplings. However, this level of complexity exceeds the current computational capacity. In recent years, theoretical research has increasingly focused on combining machine learning with first-principles methods to address challenges related to metal oxide surfaces.[32,33]

In this study, we employed a combination of statistical learning methods, a cluster expansion (CE) model, and first-principles calculations to decouple the interactions among Ce^{3+}'s and V_O's. This approach allowed us to sample the electronic configurations of V_O's at the $CeO_2(111)$ surface more effectively. We identified specific configurations characterized by maximum V_O–Ce^{3+} attractions and minimum Ce^{3+}–Ce^{3+}/V_O–V_O repulsions. Notably, as the concentration of V_O's increases, they tend to distribute toward the third oxygen layer rather than concentrating near the surface, as observed with isolated V_O's. This unique behavior is attributed to a distinct geometric relaxation, a factor that was overlooked in earlier studies, which primarily considered lower oxygen vacancy concentrations and/or focused on the surface and subsurface layers.[23–25,34] The integration of statistical learning and computational models has significantly enhanced our understanding of the intricate interactions and configurations of oxygen vacancies (V_O's) and Ce^{3+}'s species on the $CeO_2(111)$ surface. It is evident that the stability of oxygen vacancies is crucial in understanding the surface physics and chemistry of reducible oxides, indicating a potential need for reinterpreting earlier data. Given that the surface chemistry of such oxides is largely influenced by defects, there is ample opportunity for unexpected discoveries.

2. Theoretical methods

2.1. DFT modeling

We performed spin-polarized DFT calculations using the Perdew–Burke–Ernzerhof (PBE) functional, implemented in the Vienna *ab initio* simulation package (VASP).[35,36] To account for the localized Ce 4f states (Ce^{3+}), we applied DFT+U with an effective U value of 5.0 eV.[37–39] Additionally, to inspect different configurations of the reduced Ce^{3+} sites, a two-step relaxation procedure was applied. In the first step, we replaced selected Ce^{4+} ions by La^{3+} ions, with a larger ionic radius, for coarse optimization. The obtained relaxed structure was further optimized using the regular Ce^{4+} projector augmented wave (PAW) potentials. All configurations were charge-compensated according to the number of oxygen vacancies. Valence states were considered for Ce (5s, 5p, 6s, 4f, 5d) and O (2s, 2p) electrons, and PAW potentials were employed to represent core–valence interactions. A plane-wave cutoff of 400 eV was used to expand the Kohn–Sham valence states.[40,41] The $CeO_2(111)$ surface was modeled as a periodic slab with a 4 × 4 surface supercell and included four O–Ce–O tri-layers. To avoid interactions between periodic images, a vacuum layer of 15 Å was included. The bottom three atomic layers (O–Ce–O) were held fixed to their bulk positions, and the Γ point was adopted for all calculations.

2.2. Statistical learning and the cluster expansion model

Statistical learning in conjunction with the CELL code (available at the Python Package Index (PyPI) repository **https://pypi.org/project/clusterX/**; documentation can be found at **https://sol.physik.hu-berlin.de/cell**) was employed to train energy data calculated using DFT and optimize the cluster expansion (CE) model.[34,42–46] The cluster pool consists of 3144 clusters, each containing no more than 4 points (where a point represents either a Ce^{3+} or a V_O), with the distance between the points determined according to the lattice constant of 15.45 Å in the x and y directions and periodicity. Increasing the cutoff from 8 Å to 9 Å raises the total number of features from 3144 clusters to 8527 clusters, but we did not identify any additional clusters with predominant contributions. Further increasing the cutoff exceeds the supercell limit. Thus, choosing 8 Å as the cutoff threshold in our work is reasonable and justified (Table S1†). Various configurations were considered, incorporating one to four vacancies randomly located at the surface, subsurface, and third oxygen layer, each corresponding to different concentrations (Θ = 1/16, 1/8, 3/16, 1/4). The Ce^{3+} positions were randomly located at the surface and subsurface. The V_O formation energy (E_f^{CE}) for a given configuration S is described as

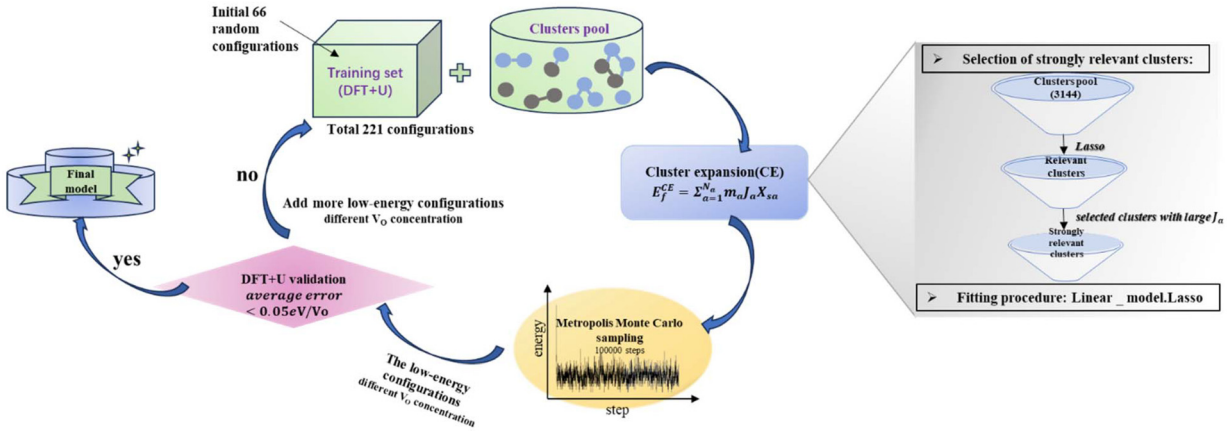
$$E_f^{CE} = \sum_{a=1}^{N_a} m_a J_a X_{sa} \quad (1)$$

where the sum is taken over all symmetrically inequivalent interactions.[34,47] Configuration S encompasses various clusters, with the multiplicity of cluster a represented by m_a. The probability of finding cluster a in configuration S is expressed by X_{sa}, and J_a represents the effective cluster interactions; for more details, refer to ref. 48. During the training process, an indicator-binary cluster basis was set.[42] For the selection of relevant clusters and fitting of the model, the LASSO algorithm was chosen, with the optimal sparsity parameter of 0.000294. Additionally, the leave-one-out-cross-validation method was employed to assess the model's predictive power and avoid overfitting.

3. Results

3.1. Decoupling of the many-body interaction using a statistical-learning model

To determine the V_O formation energy (E_f^{CE}) of any configuration of V_O's and their associated Ce^{3+}'s, we constructed a CE



Scheme 1 Statistical learning flowchart of the CE model.

model that accounts for interactions among Ce^{3+}–Ce^{3+}, V_O–Ce^{3+}, V_O–V_O, and their couplings (clusters). As shown in Scheme 1, the training data for our CE model consisted of DFT+*U* calculations involving diverse random configurations of V_O structures. These structures encompassed V_O concentrations (Θ) of 1/16, 1/8, 3/16, and 1/4. These V_O concentrations pertain to the monolayer oxygen atoms and are defined as $\Theta = n/N$, where $N = 16$ and $n = 1, 2, 3, 4$. Here, N represents the number of oxygen atoms in a non-reduced oxygen layer within the surface unit cell, and n is the number of V_O's in the unit cell. Due to the fact that the formation energies of bulk V_O's are much larger than those near the surface, as indicated by previous studies[24,49,50] and our preliminary tests (Fig. S1†), we have confined the V_O's to the three topmost oxygen layers: the surface, the subsurface, and the third oxygen layer. Similarly, Ce^{3+} ions are confined to the two topmost cationic layers: the surface and subsurface. Our objective is to address the possibility of V_O's in the deeper layers. We will demonstrate the high abundance of V_O's in the third layer using a combination of LASSO and cluster expansion methods in this work, which extends beyond previous understanding. While we do not exclude the possibility of V_O's in the 4th layer or even deeper layers, these were not included in the current work due to their relatively higher energies and much higher computational cost. Subsequently, the V_O formation energies were calculated,

$$E_f^{DFT} = E_{CeO_{2-x}}^{relax} + \frac{n}{2}E_{O_2} - E_{CeO_2}^{pristine} \quad (2)$$

where $E_{CeO_{2-x}}^{relax}$ is the energy of the configuration with n V_O's and relaxed geometry and E_{O_2} and $E_{CeO_2}^{pristine}$ are the total energies of the isolated O_2 molecule and the pristine slab. In the DFT+*U* calculation process, the oxidation state of a given Ce atom can be estimated by evaluating its local magnetic moment, defined as the difference between up and down spins on the atoms. This can be estimated by integrating the site- and angular momentum projected spin-resolved density of states over spheres with radii chosen as the Wigner–Seitz radii of the PAW potentials. In the case of reduced Ce ions, where the occupation of Ce f states is close to 1 and the magnetic moment is ~1μ_B, these ions are typically referred to as Ce^{3+}. Fig. S2† illustrates isosurfaces of the difference between spin-up and spin-down charges for the example of four subsurface oxygen vacancies and eight Ce^{3+} ions. The localization of the charge in f-like orbitals is clearly observable. Additionally, Ce^{3+}–O bond lengths (~2.49 Å) are larger than those of Ce^{4+}–O bonds (~2.37 Å).

In the cluster expansion method with an 8 Å cutoff, we considered all clusters consisting of no more than 4 points, where each point represents either a Ce^{3+} or a V_O. The initial pool included 3144 clusters: 1 zero-body, 5 one-body, 47 two-body, 430 three-body, and 2661 four-body clusters. These clusters account for the positions of V_O's and Ce^{3+}'s, and the distances between them. The key clusters were selected based on the vacancy formation energies of charge-balanced configurations obtained from DFT calculations, which inherently account for all the electronic interactions, including electrostatic interactions. Thus, charge balance is implicit in our model. These clusters serve as parameters within the CE model rather than as independent physical structures. To represent a complete and realistic configuration, multiple clusters must be combined (as shown in Fig. S3†). To disentangle the many-body interactions and identify the predominant features with the largest effective cluster interactions, we employed the least absolute shrinkage and selection operator (LASSO) approach.[8,51–55] LASSO is an effective algorithm to avoid overfitting and select the optimal cluster set. In this approach, the objective function minimized is as follows:

$$\eta = \sum_i^m \left(E_f^{CE}(J_a)_i - E_{f_i}^{DFT}\right)^2 + \lambda \| J_a \|_1 \quad (3)$$

Here, the first term represents the summation of the squared error between the predicted and calculated formation energies, and the second term, the l_1 regularization term, is the summation of $|J_a|$. The hyperparameter λ is obtained by minimizing a leave-one-out cross-validation error. The leave-one-out cross-validation method was employed to avoid overfit-

ting and evaluate the predictive power of the CE model. From the clusters identified by LASSO, we isolated those with significant effective cluster interactions (J_a) to establish a preliminary model. Subsequently, we performed Metropolis Monte Carlo (MC) simulations coupled with statistical learning predictions at 1000 K over 10^5 steps. At each step, the model provided energy predictions, enabling the identification of the 20 lowest-energy configurations across various V_O concentrations (4 concentrations with 5 configurations per concentration). Given the complexity of the reduced ceria surface, the primary objective of this work is not to develop a model with extremely high accuracy for predicting the lowest energy configuration directly. Instead, we aim to efficiently identify low-energy configurations with a sufficiently low error between the average predicted V_O formation energy, E_f^{CE}, and DFT+U calculations. If the error exceeded the threshold (0.05 eV per V_O), new configurations sampled by MC were incorporated into the training set, and the training process continued. The final model converged to standard reference training error metrics (RMSE, MAE), cross-validation error (CV_RMSE, CV_MAE), and DFT verification thresholds for the 20 low-energy configurations. Following 8 iterations of active training, wherein clusters were re-selected with the pool containing 3144 clusters at each iteration, we identified 15 key clusters involving only one-body and two-body interactions from the initial pool of 3144 clusters by analyzing errors across different clusters and the energy deviation of the lowest-energy configurations. This analysis is detailed in Tables S2 and S3.† A total of 221 configurations were used in the final training process of our CE model. The configurations for training, along with the diversity and uniformity of our training set, are shown in Fig. S4.†

The training maximum-absolute error (MAE) is 0.07 eV per V_O (Fig. 1a) and the cross-validation MAE is 0.08 eV per V_O (Fig. 1b). The detailed distributions of the absolute errors between the CE predicted formation energy and the DFT calculated formation energy of Vo's are displayed in Fig. 1c and d. The error distributions are concentrated below 0.1 eV, as indicated by the median line and the broad area of the violin plot. The largest deviations between the calculated and predicted energies for the training set and leave-one-out cross-validation arise from the high-energy unstable configurations. The formation energies of the 15 primary cluster features identified by decoupling the many-body interaction are shown in Fig. 2. In this notation, $V_{On}Ce_{m}_\ l$N, n and m denote the oxygen atomic layer and the Ce atomic layer where the V_O and Ce^{3+} are located, respectively. l denotes the positional relationship definition of V_O–V_O, V_O–Ce^{3+}, and Ce^{3+}–Ce^{3+} in the order of nearest neighbors (as shown in Fig. S5 and Table S4†).

The monomer configuration reveals that the isolated vacancy prefers the subsurface (V_{O2} = 0.22 eV), which is approximately 0.2 eV more stable than the surface vacancy (V_{O1} = 0.40 eV) and the 3rd layer vacancy (V_{O3} = 0.43 eV). In contrast, the Ce^{3+} ion prefers the surface (Ce_1 = 0.82 eV) over the subsurface (Ce_2 = 1.03 eV), which is in line with previous knowledge.[23] Regarding two-body interactions, there is a clear repulsion between the V_O's in nearest-neighbor positions in

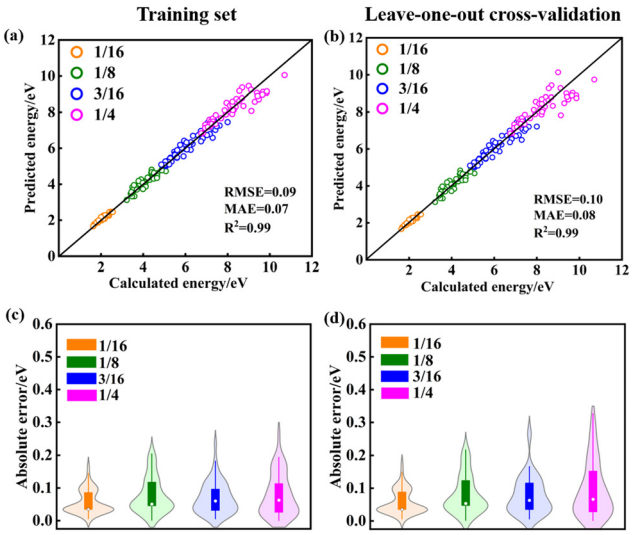


Fig. 1 Model evaluation and primary features from statistical learning. The error between the cluster-expansion model-predicted energy and the DFT energy for (a) training set and (b) leave-one-out cross-validation. Violin plots of the absolute error in the energy of configurations for (c) training set and (d) leave-one-out cross-validation. The upper and lower limits of the rectangles represent the 75th and 25th percentiles of the distribution, the internal white dot marks the median (50th percentile), and the upper and lower limits of the thin line extending from the rectangle bars indicate the minimum and maximum errors. The violin area represents the probability of data distribution.

different layers ($V_{O1}V_{O2}$_1N, $V_{O2}V_{O3}$_1N, $V_{O1}V_{O3}$_2N). Among these interactions, $V_{O1}V_{O2}$_1N exhibits the strongest repulsion (0.28 eV), while $V_{O2}V_{O3}$_1N is the weakest (0.19 eV). However, Vo's in the subsurface and the 3rd layers exhibit strong attraction when forming a specific pattern, denoted as $V_{O2}V_{O3}$_5N, with an interaction energy of −0.16 eV. When we replace this two-body cluster with a three-body cluster, $V_{O2}Ce_1V_{O3}$_5N (Fig. S6†), where one Ce^{3+} is added at a next-nearest neighbor position relative to V_{O2} and V_{O3}, the fitting accuracy remains similar (Table S5†). This suggests that the two-body cluster $V_{O2}V_{O3}$_5N effectively captures the three-body interaction in $V_{O2}Ce_1V_{O3}$_5N. In contrast, removing $V_{O2}V_{O3}$_5N decreases the accuracy of the model (Table S5†). Moreover, as shown in Table S6,† the prediction errors of model 3, which excluded the $V_{O2}V_{O3}$_5N two-body cluster, are significant for low-energy configurations. The $V_{O2}V_{O3}$_5N pattern allows the Ce^{3+} ions to occupy energetically favorable sites as next-nearest neighbors to both V_O's, thereby enhancing the system stability (Fig. S7†). Thus, this two-body cluster effectively represents the existing multi-body interactions within the model. For cerium ions, surface Ce^{3+} ions in nearest-neighbor positions exhibit stronger repulsion (Ce_1Ce_1_1N = 0.13 eV) than subsurface Ce^{3+} ions (Ce_2Ce_2_1N = 0.06 eV). Additionally, the next-nearest neighbor interaction between V_O and Ce^{3+} ($V_{On}Ce_n$_2N) is consistently attractive, in line with previous studies suggesting that Ce^{3+} ions prefer the NNN position to V_O's.[24] Specifically, for surface Ce^{3+} ions, the strongest attraction is with the formation of a NNN configuration with subsurface V_O's ($V_{O2}Ce_1$_2N =

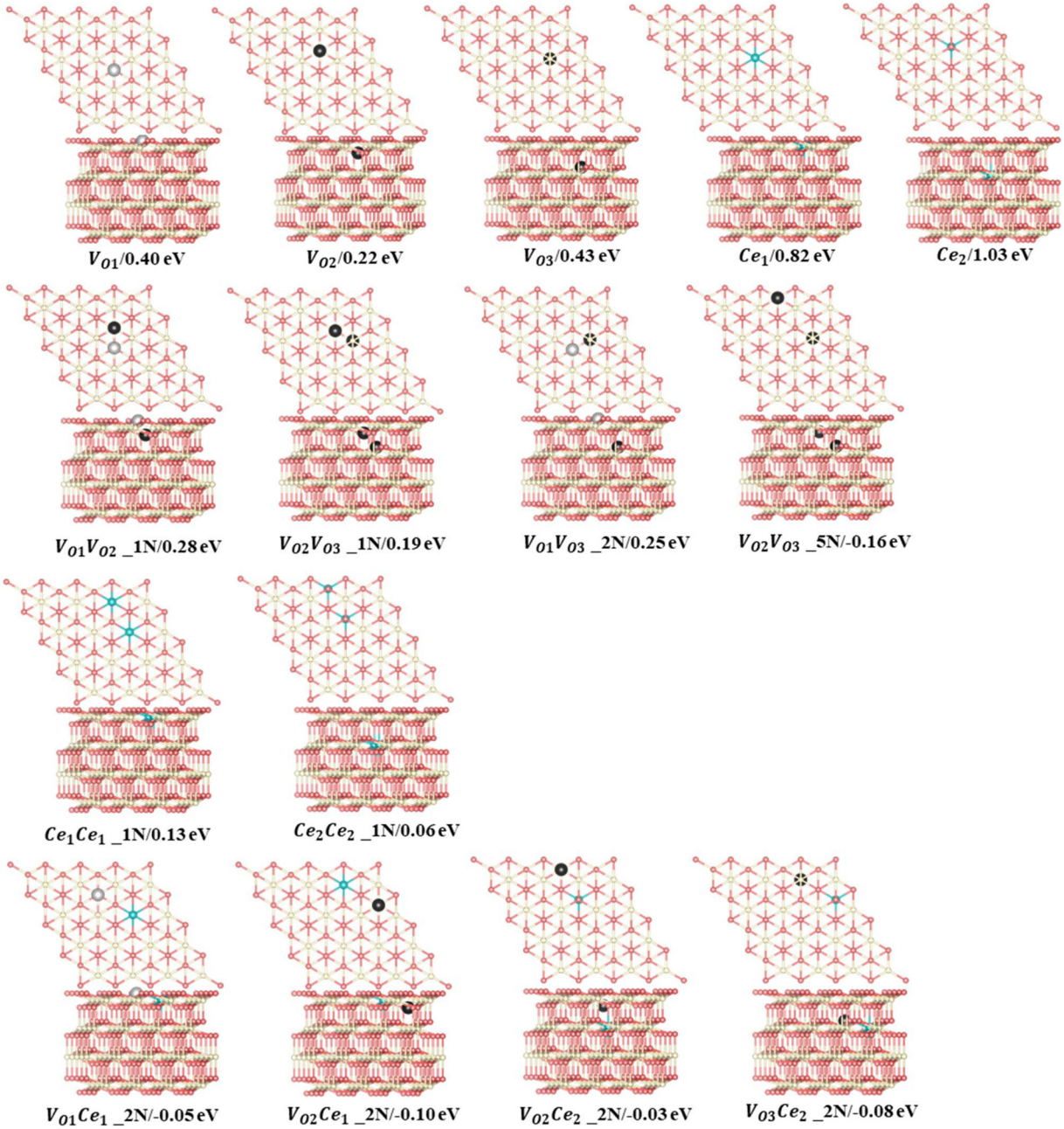


Fig. 2 15 primary features predicted by statistical learning and their energies, with structures depicted in the unrelaxed state. Ce^{4+}, Ce^{3+}, oxygen atoms, and surface oxygen vacancy are depicted in white, blue, pink, and light gray, respectively. Subsurface and third oxygen layer vacancies are depicted in black.

−0.10 eV), while for the subsurface Ce^{3+}, the strongest attractions are with 3rd layer V_O's ($V_{O3}Ce_2$_2N = −0.08 eV). The interaction (attraction/repulsion) is relative to the isolated V_O's or Ce^{3+}'s. Clusters with negligible interactions, such as the "missing" 3N and 4N clusters (and others), can be interpreted as having the same energy as isolated V_O's. This insight helps to understand the frequent occurrence of patterns with two V_O's along the same line (Fig. 3a), even when not identified as key clusters.

3.2. Statistical-learning model predicts the stability of configurations

The special motifs described above can be used to understand the specific stability of the local configurations, serving as building blocks in comprehending the stability of complete physical configurations. Leveraging on the 15 primary features obtained through decoupling, one can readily predict the stability of any vacancy structure concerning the positions of

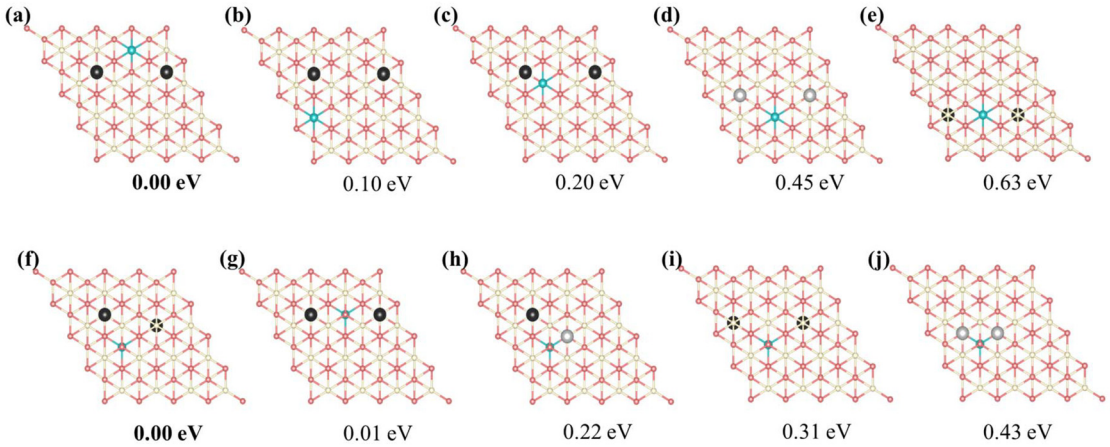


Fig. 3 Predicted relative energies of selected motifs of Vo's and Ce^{3+}'s based on the identified 15 primary features. (a)–(e) the Ce^{3+} ion is located in the first cerium layer, (f)–(j) the Ce^{3+} ion is located in the second cerium layer. Ce^{4+}, Ce^{3+}, oxygen atoms, and surface oxygen vacancies are depicted in white, blue, pink, and light gray, respectively; subsurface and third oxygen layer vacancies are depicted in black. Structures are depicted in the unrelaxed state.

V_O's and Ce^{3+}'s. Fig. S8† presents several configurations of a single V_O with two Ce^{3+} ions. Among these configurations, the subsurface V_O with two NNN surface Ce^{3+}'s is most stable, and the formation energy of this configuration can be used as a benchmark (0.00 eV) to measure the stability of other configurations. For example, the relative energies of the surface and 3rd layer Vo with two NNN surface Ce^{3+} ions are 0.28 eV and 0.42 eV, respectively. This finding is consistent with previous studies.[23,25] However, in the case of multi-vacancies, the situation becomes more complicated; again, the most stable motif is used as a benchmark. As shown in Fig. 3, for the surface Ce^{3+} ions (Fig. 3(a–e)), as the positions of V_O's and the relative position between the V_O's and the Ce^{3+} ions vary, it becomes evident that motif (a) involves the most stable subsurface V_{O2} and exhibits the most attractive interaction between Ce^{3+} and V_O ($V_{O2}Ce_1$_2N) compared to other configurations. The distance between the V_O's is that of 3rd neighbors (7.72 Å) in the oxygen layer, consistent with previous work.[25] On the other hand, for subsurface Ce^{3+} (Fig. 3 (f–j)), motif (f), which satisfies both strong $V_{O3}Ce_2$_2N and $V_{O2}V_{O3}$_5N attractions, is the most stable. This indicates that the 3rd layer V_O's are stable when subsurface Ce^{3+} are present. In summary, specific local configurations with high stability have been identified.

Based on the above analysis, we can understand the lowest-energy configurations at different concentrations (Θ = 1/16, 1/8, 3/16, and 1/4) using a 4 × 4 supercell obtained from the protocol in Scheme 1. As shown in Fig. S9 and S10,† these configurations were predicted by the CE model and subsequently verified through DFT+U calculations. For low V_O concentrations (Θ = 1/16 and 1/8), the low-energy configurations feature vacancies distributed in the surface and subsurface layers. This observation is consistent with previous literature,[23,25] indicating that subsurface V_O's are the most stable, with excess electrons localized on NNN cerium positions to the vacancies. Specifically, for Θ = 1/8, the most stable configuration has V_O's forming third-nearest-neighbor pairs in the oxygen plane, aligning with previous knowledge[25] (Fig. S10†). However, as the V_O concentration increases (Θ = 3/16), vacancies begin to appear in the 3rd oxygen layer in addition to the subsurface layer, rather than in the surface layer. A configuration featuring two V_O's in the subsurface and one V_O in the 3rd oxygen layer (*cf.* $V_{O2}V_{O3}$_5N in Fig. 2) becomes more stable than the configuration where all V_O's are located in the subsurface (Fig. S10†). These results extend the previous understanding that Vo's first distribute in the subsurface and then in the surface across a wide range of vacancy concentrations. With a further increase in V_O concentration to 1/4, more V_O's distribute to the 3rd layer. A stable structure emerges, composed of $V_{O2}V_{O3}$_5N and $V_{O2}Ce_1$_2N features (Fig. 4a). In this structure, termed A_{2233} (A_{mnpq}, m, n, p, and q denote the oxygen layers where the V_O's are located), two V_O's are in the subsurface layer, while two V_O's are in the 3rd oxygen layer. The Ce^{3+} ions are positioned in NNN positions relative to the vacancies. This configuration is energetically favorable by 0.09 eV compared to the proposed (2 × 2) structure from previous literature, where all V_O's are in the subsurface and are third-nearest neighbors in the oxygen plane, termed A_{2222} (Fig. 4b).[25] It has been suggested that vacancy-induced lattice relaxation effects play an essential role in determining the spacing between V_O's in the (2 × 2) structure. Moreover, although the outermost cerium layer is usually the energy-preferred location for Ce^{3+} ions, in both of these structures, they prefer to be in the deeper layer rather than adjacent to a vacancy. In the novel structure (Fig. 4a), two V_O's are located in the third oxygen layer rather than all in the subsurface, with a distance of 6.11 Å ($V_{O2}V_{O3}$_5N = 6.11 Å) between V_{O2} and V_{O3}.

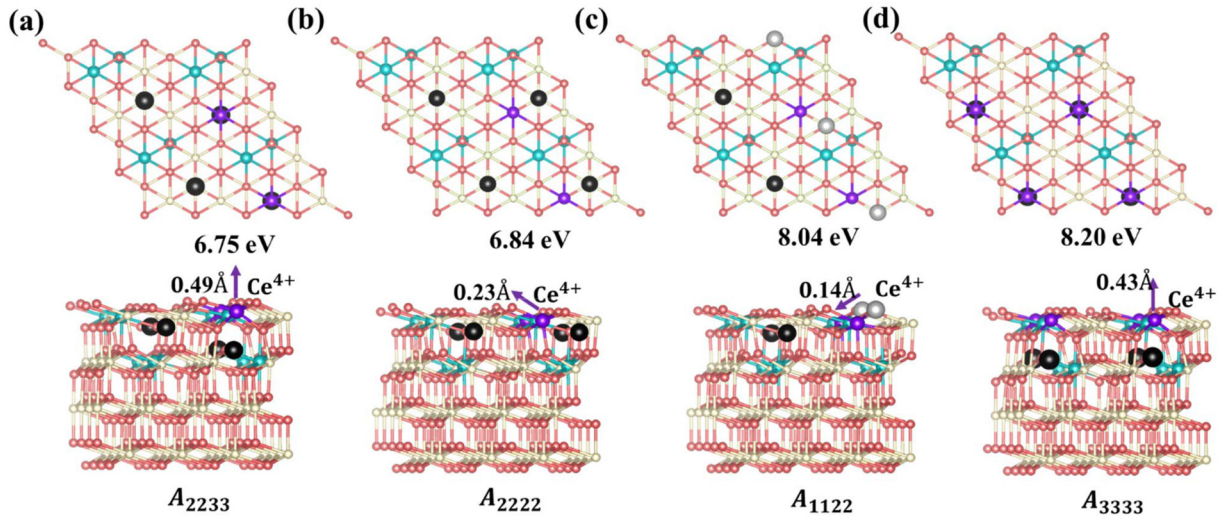


Fig. 4 Top views, side views, and the Vo's formation energies of the configurations A_{2233}, A_{2222}, A_{1122} and A_{3333}. (a) Configuration A_{2233}, with two V_O's in the subsurface layer and two V_O's in the 3rd oxygen layer, (b) configuration A_{2222}, with four V_O's in the subsurface layer, (c) configuration A_{1122}, with two V_O's in the subsurface layer and two V_O's in the surface layer and (d) configuration A_{3333}, with four V_O's in the 3rd layer. Ce^{4+}, Ce^{3+}, oxygen atoms and surface oxygen vacancies are depicted in white, blue, pink, and light gray, respectively; subsurface and third oxygen layer vacancies are depicted in black, and purple atoms are Ce^{4+}. Top and side views are depicted in the unrelaxed and relaxed states, respectively. The arrows indicate the relaxation direction of the purple-colored Ce^{4+} at the surface layer.

4. Discussion

4.1. Stability analysis based on DFT calculations

To gain a deeper understanding of the molecular mechanism behind the stability of V_O's in the 3rd layer rather than the surface layer at higher vacancy concentrations, we focus on the A_{2233}, A_{2222} and A_{1122} configurations at Θ = 1/4 (Fig. 4). In our analysis, we propose that the vacancy formation energy can be decomposed into two components:[24] bond-breaking energy (E_b) and relaxation energy (E_r):

$$E_b = E^{unrelax}_{CeO_{2-x}} + \frac{n}{2}E_{O_2} - E^{pristine}_{CeO_2} \quad (4)$$

$E^{unrelax}_{CeO_{2-x}}$ is the energy of a configuration with n V_O's and unrelaxed geometry and E_{O_2} and $E^{pristine}_{CeO_2}$ are the total energies of the isolated O_2 molecule and the pristine slab,

$$E_r = E^{relax}_{CeO_{2-x}} - E^{unrelax}_{CeO_{2-x}} \quad (5)$$

where $E^{relax}_{CeO_{2-x}}$ is the energy of a configuration with n V_O's and relaxed geometry, as shown in Fig. 5. The relaxation energies for A_{2233}, A_{2222}, and A_{1122} are −9.85 eV, −8.80 eV, −6.90 eV, respectively. When compared to E_b, it is evident that the larger E_r, resulting from substantial geometric relaxation, is the primary driver behind the stability of the A_{2233} configuration. To further elucidate the impact of the relaxation, we also calculated E_r and E_b for isolated V_O's (Θ = 1/16) and considered the sum of the corresponding energies of isolated single V_O's in the A_{mnpq} configurations. As shown in Table S7 and Fig. S11,† the deeper the V_O is located, the more significant E_r and E_b become. When comparing an isolated subsurface V_O to a surface V_O, the relaxation energy gain increases (ΔE_r = 0.41 eV)

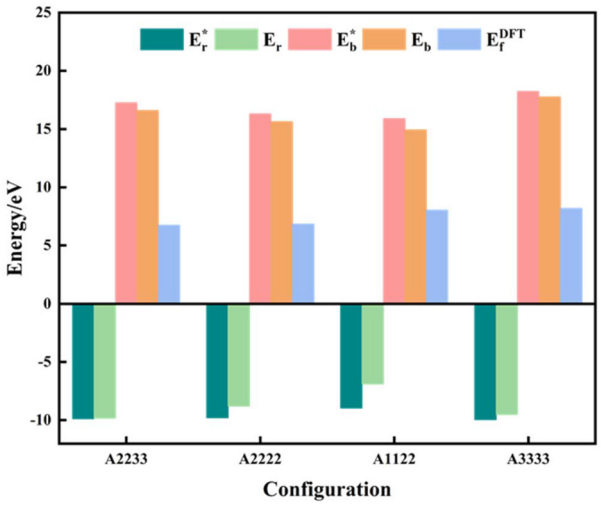


Fig. 5 Vacancy formation energy (E_f^{DFT}), relaxation energy (E_r), and bond-breaking energy (E_b) of the configurations with four V_O's (Θ = 1/4). E_r^*, E_b^* are the sum of the corresponding energies of isolated single V_O's in the A_{mnpq} configurations.

considerably more than bond-breaking energies (ΔE_b = 0.21 eV). This indicates that structure relaxation is necessary for subsurface vacancy stabilization, in line with a previous study.[24] Conversely, when the isolated V_O is positioned in the 3rd layer, E_b is significantly enhanced (0.69 eV) compared to the surface V_O, surpassing the increase in relaxation energy (0.45 eV). As a result, a single V_O is favored to be distributed in the subsurface. In summary, there exists a competitive relationship between E_r and E_b,[24] where extensive geometric

relaxation plays a pivotal role in determining the stability of V_O's at different locations. For configurations A_{mnpq} with four V_O's, the E_b for A_{2233}, A_{2222} and A_{1122} (Fig. 5) is approximately 1 eV lower than E_b^* (sum of the corresponding energies of isolated single V_O). Furthermore, the relaxation energy gains ($|E_r|$) are reduced compared to isolated single V_O's ($|E_r^*|$) due counteracting vacancy-induced relaxation effects with increasing coverage (from 1/16 to 1/4). However, the relaxation in configuration A_{2233} with 3rd layer V_O's is only minimally impeded (0.03 eV) compared to A_{2222} (1 eV) and A_{1122} (2.08 eV), which contributes to the stability of A_{2233}. The stronger relaxation in A_{2233} is reflected in the atomic displacements of surface Ce^{4+} ions nearest to the V_O's. This distance is much larger for A_{2233} (0.49 Å) on average (purple atoms in Fig. 4) compared to those for A_{2222} (0.23 Å) and A_{1122} (0.14 Å). This is because Ce^{4+} ions nearest neighbor to the V_O's tend to move away from V_O site;[25] A_{2233} provides more space for the upward relaxation of the 6-fold coordinated surface Ce^{4+} ions due to the absence of a Ce–O bond at the opposite position, which is a distinct feature compared to A_{2222} and A_{1122} (Fig. 4). However, it is not the case that larger relaxation energy always leads to a more stable corresponding configuration. Based on our CE model, we sampled configurations with four vacancies in the third layer and predicted the stable configuration A_{3333}, which was then verified by DFT (Fig. 4). Note that in all four configurations, Ce^{3+} ions are in next-nearest sites to the vacancies. Our analysis revealed that A_{2233} remains the most stable configuration, while A_{3333} is the least stable. Specifically, the E_b of A_{3333} (17.77 eV) is significantly higher compared to other configurations. Although E_r of A_{3333} (−9.53 eV) is higher than those of A_{2222} (−8.80 eV) and A_{1122} (−6.90 eV), its overall stability is compromised due to its substantial E_b. Moreover, the Ce^{4+} ions directly above the oxygen vacancy in A_{3333} exhibit a stronger relaxation (0.43 Å) compared to those in A_{2222} (0.23 Å) and A_{1122} (0.14 Å). This indicates that while surface Ce^{4+} relaxation contributes to stability, the large bond-breaking energy in A_{3333} reduces its overall stability compared to A_{2233}. Therefore, based on this analysis, at very low V_O concentration, V_O's tend to prefer the subsurface layer. As the concentration of V_O's increases (from 1/16 to 1/4), the repulsion between V_O's become significant, causing some V_O's to migrate to the 3rd layer, where more relaxation space is available. However, due to their high bond-breaking energy associated with V_O's in the 3rd layer, not all of V_O's will migrate there. Note that already at Θ = 3/16, vacancies begin populating the 3rd layer (Table S8 and Fig. S12†). We also consider additional configurations in which the positions of Ce^{3+} remain unchanged with respect to those in A_{2233}, A_{2222} and A_{1122}, but those of the vacancies do not. They are either distributed in the surface and subsurface (A_{1122a}) or scattered across the surface, subsurface, and third oxygen layer (A_{1223}) (Fig. S13†). In these cases, the formation energy is higher than that of the A_{2233} configuration, and the relaxation is smaller (Table S8†). To rule out the migration of V_O's to the fourth oxygen layer, we performed DFT+U calculations for some configurations with vacancies distributed in the fourth layer (Fig. S14†), and no configurations with lower formation energies were found. This further supports the stability of the A_{2233} configuration.

4.2. Monte Carlo sampling in a larger supercell based on the statistical-learning model

Furthermore, we performed simulations using a periodic slab with an 8 × 8 surface supercell of four O–Ce–O layers to sample the distribution of V_O's at different temperatures (T = 300 K, 500 K, 700 K, 900 K, and 1100 K) and concentrations (Θ=1/16, 1/8, 3/16, and 1/4) using the Metropolis Monte Carlo (MC) algorithm in the canonical ensemble, where the volume of the simulation cell, the composition, and the temperature are held constant. Periodic boundary conditions are used in the simulation. This sampling algorithm generates a sequence of configurations by altering the current configuration at each step through one or more swaps of two randomly selected species within the substitutional lattice. Following a swap, the new proposed configuration with energy E_j is accepted with a probability

$$P(E_i \rightarrow E_j) = \min\left[\left(e^{-\frac{E_j - E_i}{K_B T}}\right), 1\right] \quad (6)$$

where E_i is the energy of the configuration before the swap. The energy of each configuration was predicted by the statistical-learning model. Each sampling process involved no less than 5 million steps, and the results are shown in Fig. 6 and 7, and Fig. S15.† The configurations presented in Fig. 6 represent the final snapshots obtained from each trajectory, while distinct simulated configurations from MC sampling are supplemented in Fig. S15†, including five snapshots extracted from the entire trajectory at regular intervals. In these simulations, configurations featuring the aggregation of local motifs formed by Ce^{3+} ions and V_O's, in which V_O's are distributed in the subsurface and 3rd layers ($V_{O2}V_{O3}$_5N), were observed. This implies that V_O's prefer locating in the 3rd layer rather than the surface layer. Specifically, for Θ = 1/16, V_O's aggregate and form a linear configuration of third-neighbor vacancy pairs in the subsurface at 300 K, consistent with DFT predictions[25] and AFM observations.[21] As temperature increases to 500 K, V_O's can migrate to the third oxygen layer and form $V_{O2}V_{O3}$_5N clusters with subsurface V_O's. With rising temperatures, the V_O distribution becomes disordered, as reported in previous studies.[34] At 500 K and 1100 K, we observe the presence of surface V_O's (Fig. S15†). For Θ = 1/8, V_O's aggregate and form similar patterns to those at Θ = 1/16, and from 900 K, we also observed surface V_O's (Fig. 6 and Fig. S15†). As the V_O concentration further increases (Θ = 3/16, 1/4), third layer V_O's begin to appear already from 300 K, and the corresponding $V_{O2}V_{O3}$_5N motifs and third-neighbor vacancy pairs co-exist in the subsurface. Among these, for Θ = 1/4 at 300 K, we also observe a local 2 × 2 ordered pattern in the subsurface, as previously predicted.[25] At these high vacancy concentrations, as the temperature increases, some nearest-neighbor linear V_O patterns appear in the subsurface accompanied by few V_O's in the third layer. Surface Vo's

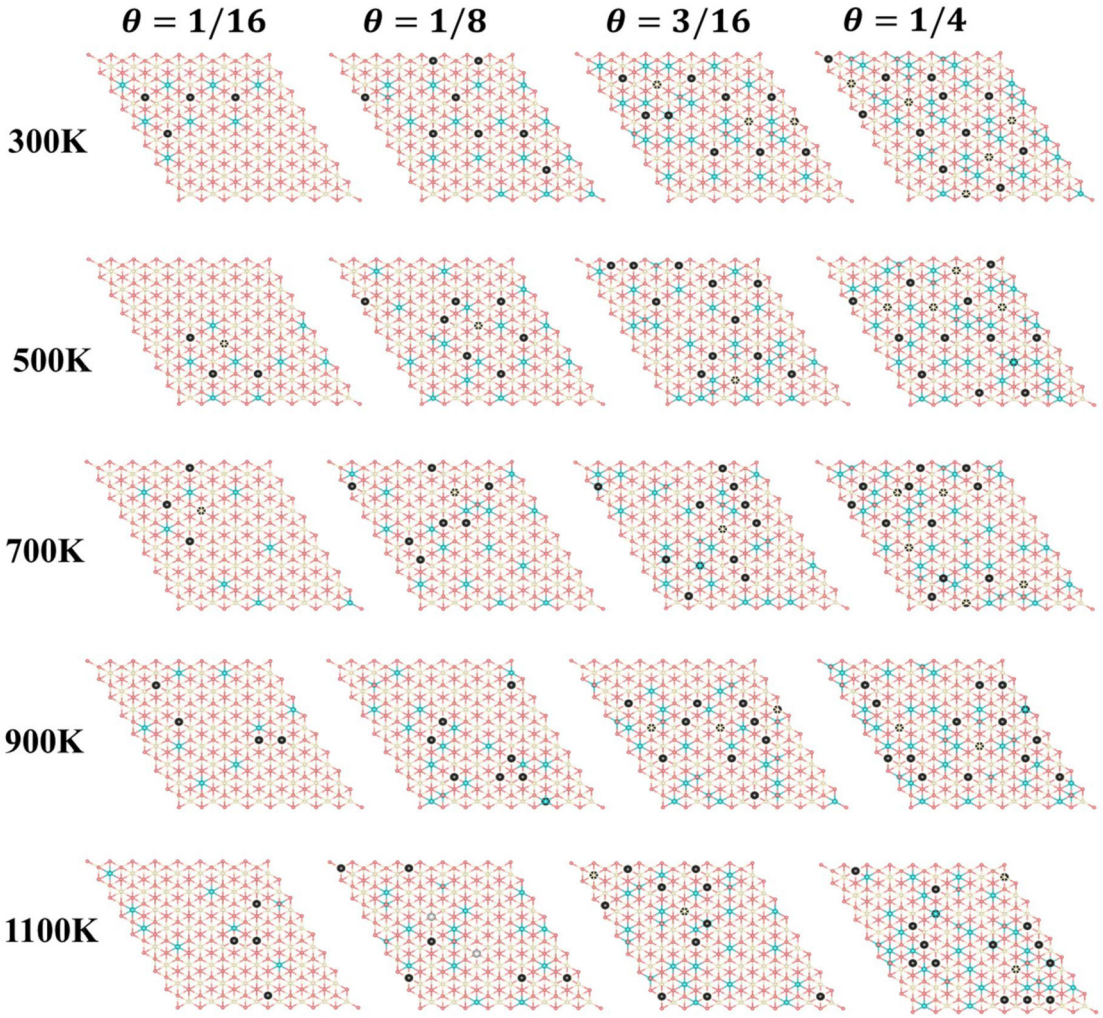


Fig. 6 Simulated configurations after 5 million MC steps. Top view of the final snapshots obtained from each trajectory of an 8 × 8 supercell at various temperatures (300 K, 500 K, 700 K, 900 K, and 1100 K) and different vacancy concentrations (1/16, 1/8, 3/16, and 1/4) sampled using MC methods. Structures are depicted in the unrelaxed state.

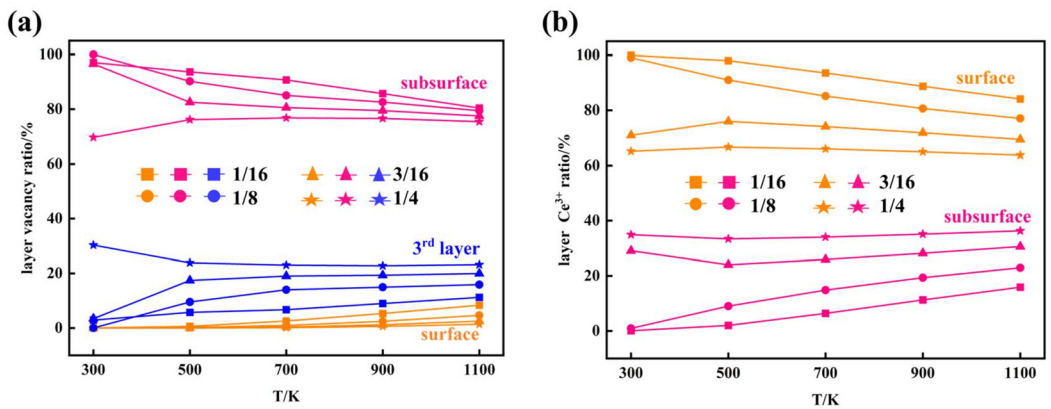


Fig. 7 Distribution of V_O's and Ce^{3+}'s at different temperatures and concentrations. The distribution ratio of (a) V_O's in the surface, the subsurface and the third layer and (b) the distribution ratio of Ce^{3+}'s in the surface and the subsurface at different V_O concentrations (Θ = 1/16, 1/8, 3/16, and 1/4) and different temperatures (T = 300 K, 500 K, 700 K, 900 K, and 1100 K).

appear very occasionally (Fig. S15†). According to the statistics of the MC trajectory, as the V_O concentration and temperature increase, V_O's are more likely to be distributed in the third layer than the surface. At a vacancy concentration where Ce^{3+} ions occupy approximately 25% of the Ce sites in the surface Ce layer, it becomes favorable for the remaining Ce^{3+} ions to occupy sites in the second Ce layer (Fig. 7). To verify the distribution of V_O and Ce^{3+} on larger supercells, we also performed simulations using a periodic slab with a 16 × 16 surface supercell (Fig. S16 and S17†). The analysis of the distribution of vacancies across two different supercell sizes reveals that at low temperatures and low concentrations (Θ = 1/16, T = 300 K), V_O's predominantly reside in the subsurface, forming linear configurations of third-neighbor oxygen vacancy pairs. This pattern is consistent regardless of the supercell size. As the temperature and vacancy concentration increase, the linear arrangement of third-neighbor oxygen vacancy pairs diminishes, particularly in the 16 × 16 supercell, where vacancies begin to exhibit a linear pattern of nearest neighbors, and a local 2 × 2 ordered pattern is also observed. Concurrently, similar to the 8 × 8 supercell, a small number of oxygen vacancies emerge in the third oxygen atom layer of the 16 × 16 supercell, leading to the formation of a $V_{O2}V_{O3}$_5N pattern with some vacancies of the subsurface, and the surface only occasionally features oxygen vacancies. For the distribution probability of Ce^{3+} ions in the 16 × 16 supercell, from Fig. S17(b),† we can observe that is also consistent with the 8 × 8 supercell. Previous studies considered only the two outermost oxygen layers for the presence of V_O's and suggested the higher stability of vacancy structures with Ce^{3+} ions in next-nearest neighbor cationic sites to the vacancies, even with some Ce^{3+} ions located in the second Ce layer.[24,25] This scenario was exemplified by the A_{2222} configuration, which was considered the most stable. However, our current work shows that when Ce^{3+} ions are present in the second Ce layer, they favor the stabilization of V_O's in the third oxygen layer (*i.e.*, A_{2233} configuration) while keeping the Ce^{3+} ions in next-nearest neighbor cationic sites to the vacancies. This behavior is attributed to additional geometric relaxation, thereby extending our previous understanding.

We also note that an earlier STM experiment observed the abundance of surface V_O's at moderate to high temperature.[20] These experiments cannot characterize the V_O's beyond the subsurface. The novel finding of the relatively higher stability of the 3rd layer V_O's compared to surface V_O's at the high V_O's concentrations underscores the potential requirement to reconsider previous interpretations that neglected the possible existence of vacancies in the third oxygen layer.[34] More sophisticated studies are needed in the future.

5. Conclusions

In summary, the reduced CeO_2 surface is a complex system featuring oxygen vacancies, V_O's, and Ce^{3+} ions with many-body interactions among them. Understanding these interactions is challenging without decoupling them into simpler one- and two-body terms. Through the application of the LASSO regression method and cluster expansion, we successfully achieved this decoupling with high accuracy. Among the total 3144 clusters representing many-body interactions, we identified 15 key clusters that play a predominant role in the configuration space. Our approach effectively addresses the challenge of sampling a large number of configurations and disentangling the many-body interactions among the Ce^{3+} ions and V_O's. Our investigation revealed a tendency for V_O's on the reduced CeO_2(111) surface to aggregate into specific configurations with Ce^{3+} ions, favoring the V_O positioning in the third oxygen layer. This phenomenon primarily stems from the significant lattice relaxation effect. This research reshapes the fundamental understanding of ceria surface structures and lays a solid foundation for further active learning when extending the approach to other Ce-related oxides. Moreover, the coupled statistical-learning and cluster expansion method demonstrated here can be readily adapted to study other metal oxides with well-defined lattice structures.

Author contributions

Y. G. conceived the original idea. H. L., M. V. G. P., and Y. G. supervised the project. Y. Z. performed the DFT calculations. Y. Z., Z.-K. H., and X. H. performed the statistical learning and CE calculations. Y. Z. analyzed the data. Z.-K. H., B. Z., and X. H. helped to analyze the data. M. T., S. R., and C. D. developed the program code. Y. Z. wrote the initial draft of the manuscript. M. V. G. P. and Y. G. assisted in interpreting the results and revising the manuscript. All authors contributed to discussions and provided critical feedback.

Data availability

The CELL code is available *via* the Python Package Index (PyPI) repository at **https://pypi.org/project/clusterX/**. Data of statistical learning can be found at **https://doi.org/:10.17172/NOMAD/2024.05.08-1**.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

Y. G. acknowledges the support of the National Key R&D Program of China (2023YFA1506900), the National Natural Science Foundation of China (12174408), the Natural Science Foundation of Shanghai (22JC1404200), and the Foundation of Key Laboratory of Low-Carbon Conversion Science & Engineering, Shanghai Advanced Research Institute, Chinese Academy of Sciences (KLLCCSE-202201Z, SARI, CAS). H. L.

acknowledges the funding support from the National Key R&D Program of China (2022YFA1504001) and the National Natural Science Foundation of China (22288102, 21935001). M. V. G. P. acknowledges the support of the Grant PID2021-128915NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, UE. All calculations were performed at National Supercomputer Centers in Tianjin, Shanghai and Guangzhou.

References

1 J. A. Rodriguez, S. Ma, P. Liu, J. Hrbek, J. Evans and M. Pérez, *Science*, 2007, **318**, 1757–1760.
2 S. Park, J. M. Vohs and R. J. Gorte, *Nature*, 2000, **404**, 265–267.
3 Q. Fu, H. Saltsburg and M. Flytzani-Stephanopoulos, *Science*, 2003, **301**, 935–938.
4 G. A. Deluga, J. R. Salge, L. D. Schmidt and X. E. Verykios, *Science*, 2004, **303**, 993–997.
5 C. T. Campbell and C. H. F. Peden, *Science*, 2005, **309**, 713–714.
6 X. Q. Wang, J. A. RodriguezJonathan, J. C. HansonDaniel, D. Gamarra, A. Martínez-Arias and M. Fernández-Garcíaet, *J. Phys. Chem. B*, 2006, **110**, 428–434.
7 R. J. Gorte, *AIChE J.*, 2010, **56**, 1126–1135.
8 J. A. Rodriguez, P. Liu, J. Hrbek, J. Evans and M. Pérez, *Angew. Chem., Int. Ed.*, 2007, **46**, 1329–1332.
9 A. Bruix, J. A. Rodriguez, P. J. Ramírez, S. D. Senanayake, J. Evans, J. B. Park, D. Stacchiola, P. Liu, J. Hrbek and F. Illas, *J. Am. Chem. Soc.*, 2012, **134**, 8968–8974.
10 M. Capdevila-Cortada, M. García-Melchor and N. López, *J. Catal.*, 2015, **327**, 58–64.
11 H. Nörenberg and G. A. D. Briggs, *Surf. Sci.*, 1999, **424**, L352–L355.
12 K.-I. Fukui, Y. Namai and Y. Iwasawa, *Appl. Surf. Sci.*, 2002, **188**, 252–256.
13 R. Olbrich, G. E. Murgida, V. Ferrari, C. Barth, A. M. Llois, M. Reichling and M. V. Ganduglia-Pirovano, *J. Phys. Chem. C*, 2017, **121**, 6844–6851.
14 S. Gritschneder and M. Reichling, *Nanotechnology*, 2007, **18**, 044024.
15 S. Gritschneder, Y. Namai, Y. Iwasawa and M. Reichling, *Nanotechnology*, 2005, **16**, S41–S48.
16 S. Torbrügge, M. Cranney and M. Reichling, *Appl. Phys. Lett.*, 2008, **93**, 073112.
17 F. Šutara, M. Cabala, L. Sedláček, T. Skála, M. Škoda, V. Matolín, K. C. Prince and V. Cháb, *Thin Solid Films*, 2008, **516**, 6120–6124.
18 M. Škoda, M. Cabala, I. Matolínová, K. C. Prince, T. Skála, F. Šutara, K. Veltruská and V. Matolínet, *J. Chem. Phys.*, 2009, **130**, 034703.
19 D. C. Grinter and G. Thornton, *J. Phys.: Condens. Matter*, 2022, **34**, 253001.
20 F. Esch, S. Fabris, L. Zhou, T. Montini, C. Africh, P. Fornasiero, G. Comelli and R. Rosei, *Science*, 2005, **309**, 752–755.
21 S. Torbrügge, M. Reichling, A. Ishiyama, S. Morita and Ó. Custance, *Phys. Rev. Lett.*, 2007, **99**, 056101.
22 N. V. Skorodumova, S. I. Simak, B. I. I. Lundqvist, A. Abrikosov and B. Johansson, *Phys. Rev. Lett.*, 2002, **89**, 166601.
23 H.-Y. Li, H.-F. Wang, X.-Q. Gong, Y.-L. Guo, Y. Guo, G. Z. Lu and P. Hu, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2009, **79**, 193401.
24 M. V. Ganduglia-Pirovano, J. L. F. D. Silva and J. Sauer, *Phys. Rev. Lett.*, 2009, **102**, 026101.
25 G. E. Murgida and M. V. Ganduglia-Pirovano, *Phys. Rev. Lett.*, 2013, **110**, 246101.
26 J. Kullgren, M. J. Wolf, C. W. M. Castleton, P. Mitev, W. J. Briels and K. Hermansson, *Phys. Rev. Lett.*, 2014, **112**, 156102.
27 X.-P. Wu and X.-Q. Gong, *Phys. Rev. Lett.*, 2016, **116**, 086102.
28 M. J. Wolf, J. Kullgren and K. Hermansson, *Phys. Rev. Lett.*, 2016, **117**, 279601.
29 X.-P. Wu and X.-Q. Gong, *Phys. Rev. Lett.*, 2016, **117**, 279602.
30 J. Kullgren, M. J. Wolf, P. D. Mitev, K. Hermansson and W. J. Briels, *J. Phys. Chem. C*, 2017, **121**, 15127–15134.
31 J.-F. Jerratsch, X. Shao, N. Nilius, H.-J. Freund, C. Popa, M. V. Ganduglia-Pirovano, A. M. Burow and J. Saueret, *Phys. Rev. Lett.*, 2011, **106**, 246801.
32 V. C. Birschitzky, F. Ellinger, U. Diebold, M. Reticcioli and C. Franchini, *npj Comput. Mater.*, 2022, **8**, 125.
33 V. C. Birschitzky, I. Sokolović, M. Prezzi, K. Palotás, M. Setvín, U. Diebold, M. Reticcioli and C. Franchini, *npj Comput. Mater.*, 2024, **10**, 89.
34 Z.-K. Han, Y.-Z. Yang, B. Zhu, M. V. Ganduglia-Pirovano and Y. Gao, *Phys. Rev. Mater.*, 2018, **2**, 035802.
35 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
36 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
37 S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys and A. P. Sutton, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **57**, 1505–1509.
38 M. Cococcioni and S. D. Gironcoli, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2005, **71**, 035105.
39 D. W. Zhang, Z.-K. Han, G. E. Murgida, M. V. Ganduglia-Pirovano and Y. Gao, *Phys. Rev. Lett.*, 2019, **122**, 096101.
40 P. E. Blöchl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953–17979.
41 G. Kresse and D. Joubert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 1758–1775.
42 Cell documentation, **https://sol.physik.hu-berlin.de/cell**.
43 S. Rigamonti, M. Troppenz, M. Kuban, A. Hübner and C. Draxl, *NPJ Comput. Mater.*, 2024, **10**, 195.
44 M. C. Nguyen, X. Zhao, C.-Z. Wang and K.-M. Ho, *J. Appl. Phys.*, 2015, **117**, 093905.
45 M. Troppenz, S. Rigamonti and C. Draxl, *Chem. Mater.*, 2017, **29**, 2414–2424.
46 Z.-K. Han, D. Sarker, M. Troppenzet, S. Rigamonti, C. Draxl, W. A. Saidi and S. V. Levchenko, *J. Appl. Phys.*, 2020, **128**, 145302.
47 J. M. Sanchez, F. Ducastelle and D. Gratias, *Phys. A*, 1984, **128**, 334–350.

48 D. Sarker, Z.-K. Han and S. V. Levchenko, *Phys. Rev. Mater.*, 2023, **7**, 055802.
49 G. E. Murgida, V. Ferrari, A. M. Llois and M. V. Ganduglia-Pirovano, *Phys. Rev. Mater.*, 2018, **2**, 083609.
50 G. E. Murgida, V. Ferrari, M. V. Ganduglia-Pirovano and A. M. Llois, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **90**, 115120.
51 R. Tibshirani, *J. R. Stat. Soc.: Ser. B (Methodol.)*, 1996, **58**, 267–288.
52 L. Barroso-Luque, P. C. Zhong, J. H. Yang, F. Y. Xie, T. Chen, B. Ouyang and G. Ceder, *Phys. Rev. B*, 2022, **106**, 144202.
53 Y. F. Wang, Y.-Q. Su, E. J. M. Hensen and D. G. Vlachos, *ACS Nano*, 2020, **14**, 13995–14007.
54 H. Bhattacharjee, N. Anesiadis and D. G. Vlachos, *Sci. Rep.*, 2021, **11**, 14372.
55 Y. J. Hu, G. Zhao, M. F. Zhang, B. Bin, T. D. Rose, Q. Zhao, Q. Zu, Y. Chen, X. K. Sun, M. D. Jong and L. Qi, *npj Comput. Mater.*, 2020, **6**, 25.