

Machine learning-based prediction for stability of hydrogen species on metal-doped CeO_2(111) surfaces

Kai Wu(吴凯), P. Hu(胡培君), Xue-Qing Gong(龚学庆)

Citation: Chin. Phys. B. 35, 096801 (2026). DOI: 10.1088/1674-1056/ae3132 |
CSTR: 32038.14.CPB.ae3132
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Epitaxial growth of Bi nanowires on Pb-7×3 surface

Siyu Huo(霍思宇), Jieying Li(李洁莹), Yuzhou Liu(刘宇舟), Desheng Cai(蔡德胜), Yitong Gu(谷易通), Haoen Chi(迟浩恩), Wenhui Pang(庞文慧), Gan Yu(于淦), Xiaoying Shi(史晓影), Wenguang Zhu(朱文光), Shengyong Qin(秦胜勇)

Chin. Phys. B. 34, 106801 (2025). DOI: 10.1088/1674-1056/addeb9 | CSTR: 32038.14.CPB.addeb9

Combining machine learning algorithms with traditional methods for resolving the atomic-scale dynamic structure of monolayer MoS_2 in high-resolution transmission electron microscopy

Yu Meng(蒙宇), Shuya Wang(王淑雅), Xibiao Ren(任锡标), Han Xue(薛涵), Xuejun Yue(岳学军), Chuanhong Jin(金传洪), Shanggang Lin(林上港), Fang Lin(林芳)

Chin. Phys. B. 34, 016802 (2025). DOI: 10.1088/1674-1056/ad9ba3 | CSTR: 32038.14.CPB.ad9ba3

Reconstruction and stability of Fe_3O_4 (001) surface: An investigation based on particle swarm optimization and machine learning

Hongsheng Liu(柳洪盛), Yuanyuan Zhao(赵圆圆), Shi Qiu(邱实), Jijun Zhao(赵纪军), Junfeng Gao(高峻峰)

Chin. Phys. B. 32, 056802 (2023). DOI: 10.1088/1674-1056/acb9e4

Atomistic insights into early stage corrosion of bcc Fe surfaces in oxygen dissolved liquid lead-bismuth eutectic (LBE-O)

Ting Zhou(周婷), Xing Gao(高星), Zhiwei Ma(马志伟), Hailong Chang(常海龙), Tielong Shen(申铁龙), Minghuan Cui(崔明焕), Zhiguang Wang(王志光)

Chin. Phys. B. 32, 036801 (2023). DOI: 10.1088/1674-1056/ac834b

Machine learning identification of impurities in the STM images

Ce Wang(王策), Haiwei Li(李海威), Zhenqi Hao(郝镇齐), Xintong Li(李昕彤), Changwei Zou(邹昌炜), Peng Cai(蔡鹏), Yayu Wang(王亚愚), Yi-Zhuang You(尤亦庄), Hui Zhai(翟荟)

Chin. Phys. B. 29, 116805 (2020). DOI: 10.1088/1674-1056/abc0d5

SPECIAL TOPIC — AI + Physical Science

Machine learning-based prediction for stability of hydrogen species on metal-doped $CeO_2(111)$ surfaces

Kai Wu(吴凯)[1], P. Hu(胡培君)[1,2], and Xue-Qing Gong(龚学庆)[3,†]

[1] *Centre for Computational Chemistry and Research Institute of Industrial Catalysis, School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai* 200237, *China*

[2] *School of Physical Science and Technology, Shanghai Tech University, Shanghai* 201210, *China*

[3] *State Key Laboratory of Synergistic Chem-Bio Synthesis, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*

(Received 30 July 2025; revised manuscript received 12 December 2025; accepted manuscript online 26 December 2025)

Tuning the stability of hydrogen species on metal-doped CeO_2 surfaces is a crucial aspect in the rational design of more efficient CeO_2-based catalysts. However, traditional theoretical and experimental approaches are limited in their capacity to explore the complex chemical space arising from doping and defect introduction. To address this limitation, we establish a comprehensive dataset using high-precision density functional theory (DFT) calculations for the $CeO_2(111)$ surfaces doped with 43 different metals, encompassing three types of surface structures and five representative hydrogen adsorption configurations. A multi-level feature engineering strategy is then employed, combining elemental properties and multi-center smooth overlap of atomic positions (SOAP) features. The ensemble ExtraTrees regression model can achieve reliable prediction of adsorption energies ($R^2 = 0.89$, RMSE = 0.55 eV), and its robustness is assessed through repeated cross-validation. Shapley additive explanations (SHAP) analysis shows that model predictions rely mainly on the principal component of local geometric features. Analysis across principal component intervals indicates that the SOAP-derived features differentiate coordination symmetry, anisotropy, and cooperative effects, and reveal their influences on adsorption energies. The results of principal component analysis (PCA) and cross-validation suggest that representing samples within a physicochemical feature space provides a more nuanced understanding of microenvironmental effects on hydrogen stability compared with the traditional configuration-based approaches. This study provides a data-driven framework for high-throughput screening and mechanistic exploration of structure–property relationships in complex multi-component oxide catalysts.

Keywords: machine learning, CeO_2, surface hydrogen species, smooth overlap of atomic positions, density functional theory

PACS: 68.43.–h, 82.20.Wt, 71.15.Mb, 68.43.Fg

DOI: 10.1088/1674-1056/ae3132 **CSTR:** 32038.14.CPB.ae3132

1. Introduction

The efficient activation and selective transformation of hydrogen molecules (H_2) is a fundamental scientific issue in catalysis and energy materials science. The related processes range from traditional technologies such as Haber–Bosch ammonia synthesis[1,2] and petroleum hydrodesulfurization,[3,4] to emerging methods such as CO_2 reduction to hydrocarbons[5,6] and selective hydrogenolysis of lignin.[7,8] The reaction kinetics and product distribution are primarily influenced by two factors: (i) the energy barrier for heterolytic dissociation of H_2 molecules at catalysts; and (ii) the adsorption strength, migration, and conversion dynamics (proton/hydride transfer) of active hydrogen species (H^*) on the surface.[4,5,9] Therefore, the rational design of catalysts that simultaneously exhibit high activity, stability, and selectivity remains a significant challenge.

Cerium dioxide (CeO_2) is widely used in heterogeneous catalysis due to its excellent oxygen storage capacity and thermal stability, which largely derive from the reversible Ce^{3+}/Ce^{4+} redox couple.[10–13] Its surface plays a key role in facilitating the dissociation of H_2 to form hydroxyl (O–H), cerium hydride (Ce–H), and other active hydrogen species.[14–18] For optimal catalytic activity, the adsorption strength of these intermediates must be precisely tuned in accordance with the Sabatier principle.[19–22] Furthermore, metal doping enables additional atomic-scale modulation of CeO_2's electronic structure and defect chemistry. For example, dopant atoms can alter the oxygen vacancy formation energy and influence the distribution of active sites on the surface.[15,17,23–26] However, the high-dimensional chemical space arising from multiple dopants and their complex interactions with the host lattice greatly complicates the mapping between structure and performance, rendering both experimental and density functional theory (DFT)-based first-

†Corresponding author. E-mail: xqgong@sjtu.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

principles methods costly and less efficient.[22,27–30] Therefore, addressing the challenge posed by a multidimensional parameter space under the constraints of limited computational resources remains a key bottleneck in rational design of CeO_2-based catalysts.

Machine learning (ML) and other data-driven approaches have become important tools in catalyst design,[27,31–35] achieving notable progress in predicting catalytic properties, identifying active sites, and elucidating reaction mechanisms.[36–40] For example, Esterhuizen *et al.*[41] developed interpretable ML models that provide mechanistic insight into structure–performance relationships in heterogeneous catalysis. Tejaswini *et al.* [42] investigated the hydrogen adsorption energy on conventional transition metal dichalcogenide (TMD) monolayers with the general formula MX_2 and on Janus-type TMD monolayers with the formula MXY, revealing that the adsorption sites, hydrogen coverages, and built-in electric fields can significantly affect the activity of the hydrogen evolution reaction. Xu *et al.*[43] proposed a physics-inspired machine learning model capable of accurately predicting the adsorption enthalpy and binding modes of complex adsorbates on the surfaces of transition metal catalysts. In general, these examples demonstrate the effectiveness of machine learning methods for predicting catalytic properties. To address the aforementioned challenges in exploring complex doped CeO_2 systems, we therefore employed ML-based methods in this study.

Herein, we constructed an extensive dataset based on high-precision DFT calculations to examine hydrogen adsorption on $CeO_2(111)$ surfaces doped with 43 different metals, encompassing three types of surface structures and five representative adsorption configurations. We then constructed and optimized an ExtraTrees ensemble regression model using features that integrate local atomic environments and intrinsic elemental properties. This model allowed for efficient and accurate prediction of hydrogen adsorption energies. Furthermore, interpretability methods such as Shapley additive explanations (SHAP) revealed the key physical factors governing the stability of hydrogen species, thereby facilitating the rational design and data-driven high-throughput screening of complex multi-component catalysts, and enabling deeper analysis of structure–property relationships.

2. Methodology

2.1. DFT calculations

All spin-polarized DFT calculations in this study were performed using the Vienna *Ab initio* Simulation Package (VASP). The projector augmented wave (PAW) method was used to treat the electron–ion interactions, and the exchange–correlation effects were described by the Perdew–Burke–Ernzerhof (PBE) functional within the framework of the generalized gradient approximation (GGA). The on-site Coulombic interaction correction (DFT+U) was applied to appropriately account for the strong correlation of Ce 4f electrons. Consistent with previous studies on CeO_2-based systems, the effective Hubbard U parameter was set to $U = 5.0$ eV, which has been shown to yield accurate results for the atomic, electronic, and catalytic properties of CeO_2 bulk and surface structures.[44,45] The full details of DFT computational settings, convergence criteria, and surface model construction are provided in the supporting information.

The hydrogen adsorption energy (E_{ads}) on the metal-doped $CeO_2(111)$ surface was defined as

$$E_{ads} = -(E(2H/M\text{-}CeO_2) - E(M\text{-}CeO_2) - E(H_2)), \quad (1)$$

where $E(2H/M\text{-}CeO_2)$ is the energy of the surface with two adsorbed H atoms, $E(M\text{-}CeO_2)$ is the total energy of the clean doped surface, and $E(H_2)$ is the total energy of a gas-phase H_2 molecule. In this convention, a more positive adsorption energy (E_{ads}) indicates a more exothermic and thus more stable adsorption state.

2.2. Machine learning model construction

2.2.1. Dataset construction and feature engineering

The dataset was generated based on DFT calculations of the $CeO_2(111)$ surfaces doped with 43 different metal elements. Considering the possible conditions of the surface in practical applications, three primary structure types were systematically constructed: the pristine doped surface (pristine), the surface with an oxygen vacancy (Ov-surf), and the surface with a sub-surface oxygen vacancy (Ov-sub). In combination with the most common hydrogen adsorption sites, five representative surface configurations were also constructed (see Fig. 1(a)) to systematically explore the effects of defects and adsorption locations. Figure 1(b) displays two representative adsorption geometries optimized by DFT. For each case, preference was given to configurations in which one H atom is adsorbed at a metal site or vacancy and another at a surface oxygen site, thereby covering the main active centers. If such a configuration was unstable, a dihydroxyl adsorption motif was selected instead. In total, 215 distinct DFT-calculated adsorption energies were collected, serving as the basis for subsequent modeling. As shown in Fig. 1(c), while all the five configurations exhibited similar trends in adsorption energies for the pristine surface, clear periodicity and group-dependent differences emerge among various dopants. In particular, the transition metals (such as Fe, Co, Ni) generally gave more positive adsorption energies, favoring hydrogen stability, whereas the main group and certain p-block elements gave adsorption energies near zero or negative, indicating weaker adsorptions.

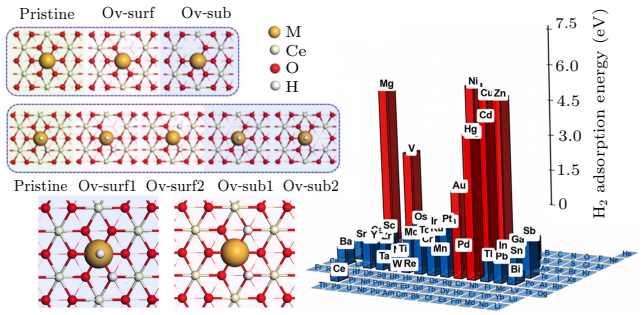


Fig. 1. Atomic structures and DFT results for metal-doped $CeO_2(111)$ surfaces: (a) atomic models of three fundamental metal-doped $CeO_2(111)$ surface structures (pristine surface, surface with a surface oxygen vacancy, and surface with a subsurface oxygen vacancy), combined with five representative hydrogen adsorption site configurations (top view); (b) DFT-optimized local atomic geometries of two typical hydrogen adsorption environments (top view); (c) distribution of hydrogen adsorption energies for different dopant metals on the pristine surfaces.

In order to transfer complex atomic structures into high-quality features suitable for machine learning, we developed a systematic feature engineering workflow that integrates both local atomic environments and elemental properties, as detailed below. To quantitatively describe each structure, we applied a systematic feature engineering approach that integrates both atomic-level and local environment information. For each key atom (M, O, Ce), 9 fundamental physicochemical features — including Pauling electronegativity, atomic radius, and first ionization energy — were extracted to characterize elemental properties (see Table S1 in the supporting information for the full descriptor list and values). To comprehensively describe the local geometric environments, a multi-center smooth overlap of atomic positions (SOAP) framework was adopted,[46] whereby the SOAP features were calculated for the dopant metal atom both before and after hydrogen adsorption, as well as for the adsorbed hydrogen atoms themselves. The construction, parameterization (e.g., cutoff radius $r_{\mathrm{cut}} = 6$ Å, $n_{\mathrm{max}} = 4$, $l_{\mathrm{max}} = 6$), and detailed theoretical background of the SOAP descriptor are described in the supporting information. Given the high dimensionality introduced by this multi-center SOAP method, we further adopted a physics-guided principal component analysis (PCA) strategy: the SOAP features were grouped by atomic pairings (such as M–O, M–Ce, and O–H), and each group was independently reduced via PCA, retaining components that together explained over 90% of the total variance (see supporting information for further benchmarking and implementation details).

The final feature vector for each sample comprised both elemental and PCA-reduced multi-center SOAP features, which ensured high information density and accurate structural representation.

2.2.2. Model training and performance evaluation

The dataset, consisting of 215 samples, was randomly divided into training and test sets in an 8:2 ratio. The training set was used for model fitting and hyperparameter optimization. The test set was reserved exclusively for unbiased final model evaluation. Several standard regression algorithms were evaluated, including linear models, Support Vector Regression (SVR), RandomForest, AdaBoost, XGBoost, and ExtraTrees. Model performance was assessed using the coefficient of determination (R^2) and root mean squared error (RMSE), calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3}$$

3. Results and discussion

3.1. Model performance

With the features obtained via PCA dimensionality reduction, we systematically evaluated 13 regression algorithms — including ExtraTrees, XGBoost, RandomForest, and SVR — through five-fold cross-validation. Model performance was assessed using R^2 and RMSE on an independent test set (Fig. 2(a)). ExtraTrees exhibited the highest predictive accuracy among all tested models, with $R^2 = 0.89$ and RMSE = 0.50 eV, outperforming SVR ($R^2 = 0.64$, RMSE = 0.90 eV) and XGBoost ($R^2 = 0.86$, RMSE = 0.56 eV). Parity plots for representative data splits (Fig. 2(b)) demonstrated good agreement between the ExtraTrees predictions and the DFT- calculated reference values (test set $R^2 = 0.89$, RMSE = 0.55 eV), indicating satisfactory generalization ability. In order to systematically assess model robustness and reduce the impact of random data splits, we performed 100 rounds of random resampling cross-validation, with the resulting prediction distributions shown in Figs. 2(c) and 2(d). Across different data partitions, ExtraTrees consistently yielded R^2 values around 0.85 and RMSEs below 0.6 eV, demonstrating reliable model performance. Detailed specifications of the ExtraTrees model, including the optimization of key hyperparameters, cross-validation protocols, and the complete training workflow, are provided in the supporting information. Accordingly, ExtraTrees was selected as the primary model for adsorption energy prediction and subsequent mechanistic analysis.

A detailed comparison of adsorption energy prediction accuracy with representative studies in related fields is provided in the supporting information (Table S2). The results indicate that the performance of our ExtraTrees model is comparable to those reported for similar systems, confirming that the proposed approach achieves a competitive level of accuracy.

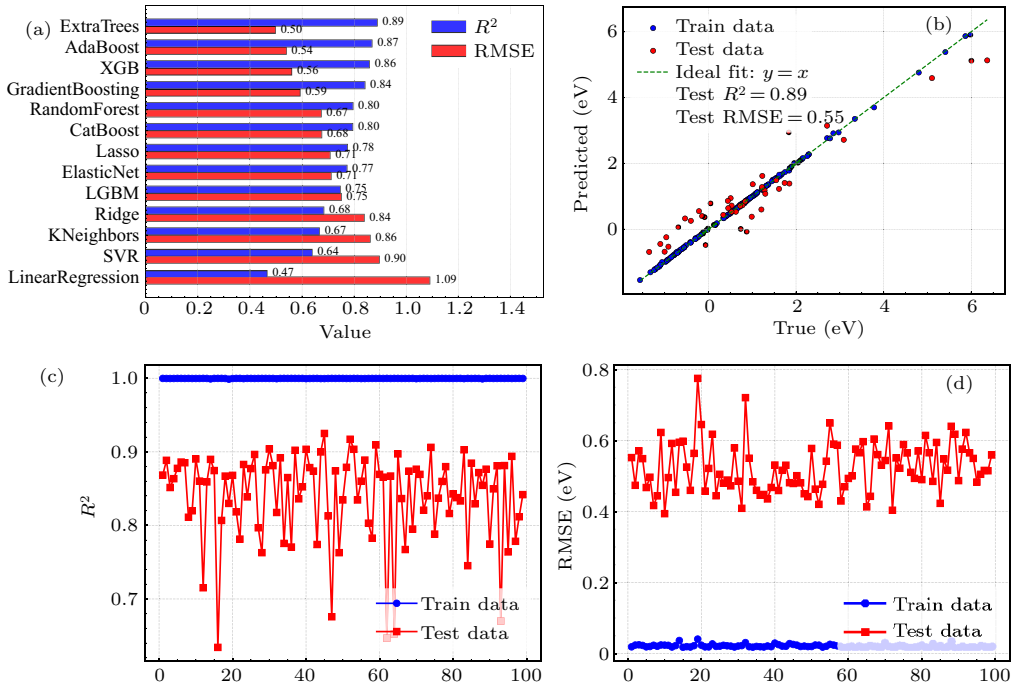


Fig. 2. Performance evaluation and generalization testing of regression models for hydrogen adsorption energy prediction: (a) Five-fold CV performance comparison of 13 regressors; (b) ExtraTrees parity plot (DFT versus model prediction based on a representative fold); (c) R^2 distribution for ExtraTrees over 100 random splits; (d) RMSE distribution for ExtraTrees over 100 random splits.

3.2. Feature importance and model interpretation

SHAP is a method for interpreting machine learning models by quantifying the contribution of each input feature to the model's predictions.[47] In this work, the SHAP analysis indicated that the model's predictions of hydrogen adsorption energies are primarily influenced by the SOAP-based geometric feature (2H/H_OCe/01), which reflects the local environment of H, O, and Ce atoms. This feature had the highest SHAP value among all input features (Fig. 3(a)). It should be noted that higher values of this feature (2H/H_OCe/01) are generally associated with lower SHAP values, which correspond to weaker hydrogen adsorption, whereas lower feature values are linked to higher SHAP values and greater adsorption stability. Compared with the conventional features such as electronegativity and atomic radius, the SOAP-based geometric features ranked higher in importance, highlighting the importance of the three-dimensional structural information. A baseline model built solely on the elemental properties (electronegativity and radius) exhibited no predictive capability ($R^2 \approx 0$), confirming that the SOAP features are crucial for representing the adsorption energetics (see Fig. S3, supporting information).

Furthermore, the inverse mapping of the most important SOAP-based feature back to the original SOAP components revealed that its contribution is distributed relatively evenly among the top ten contributing components, which span angular momentum numbers $l = 0$ to 4 and several radial basis functions ($n = 0$ to 3). Figure 3(b) shows that the weights of the SOAP components are concentrated in the low-to-intermediate angular range and the lowest radial orders, implying that the most relevant structural correlations are those describing short- to medium-range distortions of the Ce–O framework. The low-l components capture symmetry breaking associated with oxygen vacancies or dopant-induced relaxation, consistent with the Ce^{3+}/Ce^{4+} redox flexibility. The intermediate-l components encode directional Ce–O–M bond correlations reflecting anisotropic d/f–2p interactions. Radially, the dominant low-n terms correspond to interactions within the first few coordination shells (≈ 3–5 Å), aligning with the spatial extent of vacancy-induced perturbations. Overall, this pattern indicates that both short-range geometric distortions and medium-range angular correlations are the key physicochemical features governing adsorption on doped CeO_2 surfaces. This indicates that the model can integrate multi-scale directional and spatial information to capture geometric effects on the adsorption structures. The even distribution of PCA weights confirmed that multiple geometric features may jointly contribute to the predictions, which reflects the diversity of local atomic environments. Thus, the key physical features used by the model are not isolated chemical or structural factors, but comprehensive geometric representations that encode both interatomic distances and spatial arrangements over different length scales. These results suggest that the stability of hydrogen species on the metal-doped $CeO_2(111)$ surfaces is strongly governed by the fine local geometric environment and cooperative multi-atom effects, rather than by singular atomic attributes or simplified configuration classifications.

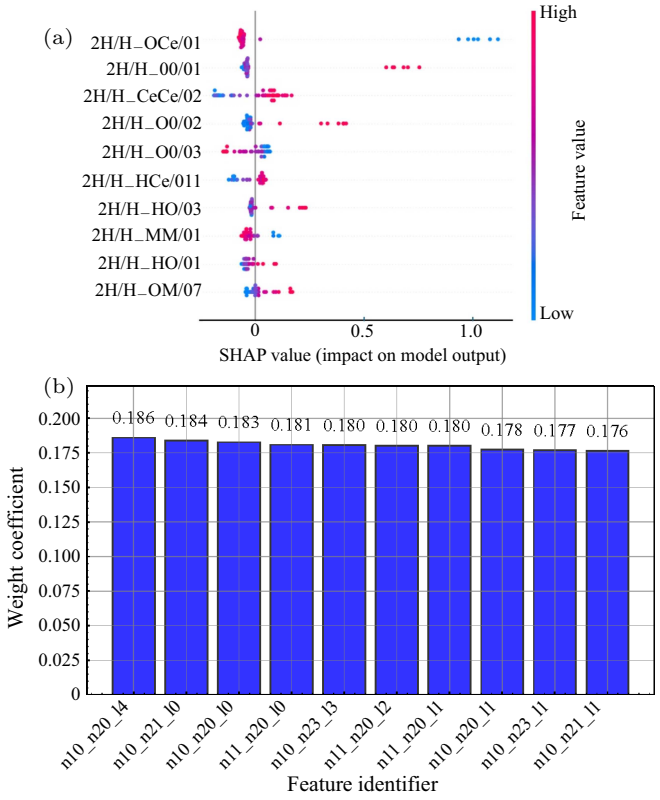


Fig. 3. Feature importance of principal components and PCA decomposition: (a) SHAP feature importance ranking for hydrogen adsorption energy prediction; (b) weights of the top 10 SOAP components in the most important principal component (2H/H_OCe/01).

To elucidate the relationship between machine learning features and actual surface structural distortions, we selected the representative structures corresponding to both high and low values of the most significant SOAP feature. The displacements of O and Ce atoms, as well as the coordination modes of H adsorption centers, were visualized in the hexagonal plane (Figs. 4(a) and 4(b)). For structures with high values of the most important SOAP-based feature — such as Os_Ov-sub1 (feature value = 8.18) and Tc_Ov-sub2 (feature value = 7.82) — the hexagonal adsorption center exhibits pronounced distortion, characterized by uneven vertex distances and significant local atomic displacements, indicating strong local anisotropy. SHAP analysis further indicated that these structures exhibit lower predicted adsorption energies, consistent with their reduced hydrogen adsorption stability. In contrast, the two adsorption structures with low values of this SOAP-based feature — Zn_Ov-sub1 (feature value = −15.64) and Mg_pristine (feature value = −15.90) — exhibit nearly regular hexagonal coordination and high symmetry, closely resembling the ideal structure of pristine CeO_2(111). These configurations are indicative of highly stable predicted hydrogen adsorption and are characterized by relatively high adsorption energy values.

Further analysis in Fig. 4(c) compared structures with mildly distorted (feature value = 3.07) and strongly distorted (feature value = −7.09) coordination geometries and systematically related the most important SOAP-based feature values to the degree of coordination distortion. Both quantitative and visual analyses confirmed that the value of this feature reliably reflects the extent of coordination environment distortion: greater distortion (higher feature values) corresponds to weaker (more negative) adsorption energies, whereas structures closer to the ideal configuration (lower feature values) are associated with stronger (more positive) hydrogen adsorption energies.

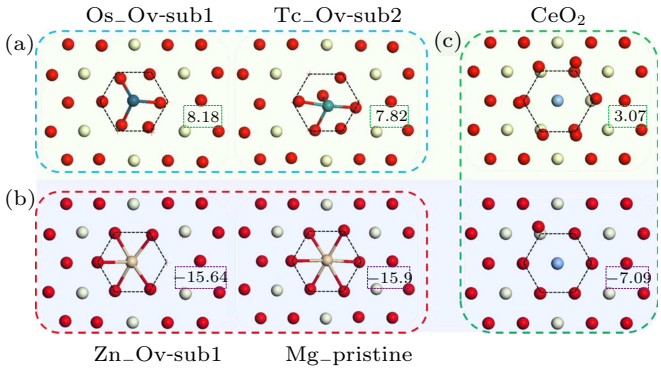


Fig. 4. Relationship between SOAP feature values and local coordination environment at hydrogen adsorption centers: (a) high SOAP feature value representative of distorted structure; (b) low SOAP feature value representative of symmetric structure; (c) structures with mild (top) versus severe (bottom) distortion.

Overall, these structural comparisons visually elucidate the physical significance of the SOAP-based features and demonstrate that anisotropic coordination plays a significant role in determining hydrogen adsorption stability. This provides compelling evidence of structure–property correlations, supporting the rational design of catalysts.

3.3. Comparison of feature space and configuration-based classification

To assess the predictive ability of the configuration labels regarding hydrogen adsorption stability, we initially applied two-dimensional PCA to visualize the SOAP features of all samples (Fig. 5(a)). This analysis revealed substantial overlap among different intuitive configurations in the feature space, reflecting similar local atomic environments. These findings indicated that relying solely on configuration labels may be insufficient for distinguishing essential differences in adsorption behaviors.

To further evaluate the generalization capability of the configuration-based classification, a leave-one-configuration-out (LOCO) cross-validation experiment was conducted (Fig. 5(b)). In this approach, each configuration was used in turn as the test set, while the remaining classes served as the training data. The results revealed substantial variation in predictive performance: high accuracy ($R^2 > 0.65$) was

achieved for Ov-surf1, Ov-sub1, and Ov-sub2, while performance dropped sharply for Ov-surf2 ($R^2 = 0.08$). This indicated that the configuration-based labels alone are insufficient to reliably capture the hydrogen adsorption stability across all surface types.

In addition, kernel density estimation (KDE) was performed to analyze the distributions of key SOAP features across the different configurations (Figs. 5(c) and 5(d)). The most important feature (2H/H_OCe/01) showed considerable overlap among all configurations (Fig. 5(c)). In contrast, while the distribution of another key feature (2H/H_OO/01) for Ov-surf2 overlapped significantly with the other configurations, it also exhibited a distinct region unique to this configuration (Fig. 5(d)). This uniqueness may account for the poor predictive performance for this configuration.

Overall, classifying samples according to their positions in high-dimensional feature space can more accurately reflect the physicochemical factors underlying hydrogen adsorption strengths. Notably, clear separation in feature distributions may provide valuable guidance for model generalization, data set construction, and optimization strategies.

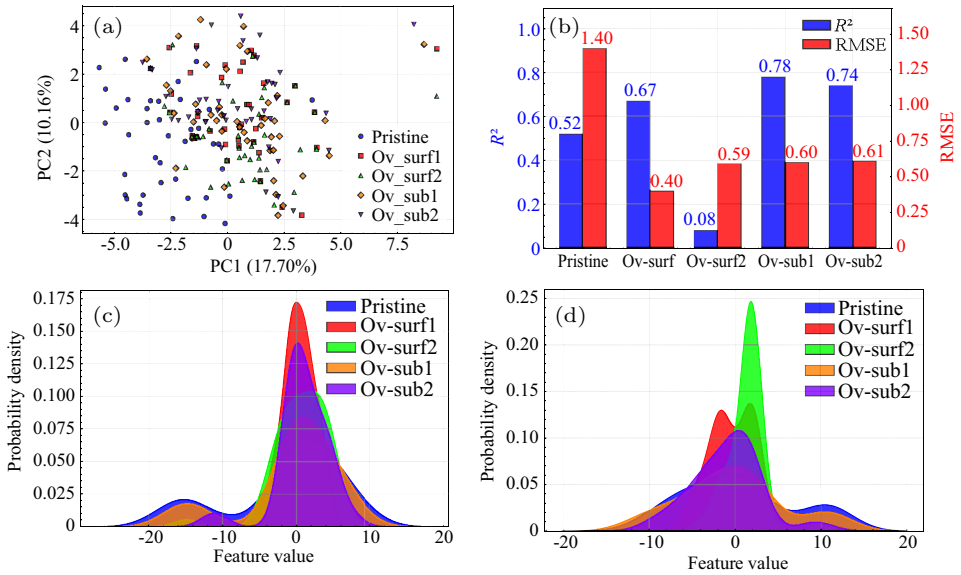


Fig. 5. Influence of feature space and configuration-based classification on hydrogen adsorption energy prediction: (a) two-dimensional PCA visualization of SOAP features for all configurations; (b) leave-one-configuration-out cross-validation performance (R^2 and RMSE) grouped by configuration (five-fold CV); (c) kernel density estimation (KDE) distribution of the most important feature (2H/H_OCe/01); (d) KDE distribution of the second most important feature (2H/H_OO/01).

4. Conclusion

By integrating DFT calculations with machine learning methods, we systematically evaluated hydrogen adsorption on the $CeO_2(111)$ surfaces doped with 43 different metals and adopting a range of surface configurations. By employing both atomic features and multi-center SOAP environment features, we designed an effective approach for dimensionality reduction and feature integration. The ExtraTrees regression model achieved high accuracy and satisfactory generalization, while SHAP analysis showed the significant importance of the local three-dimensional structure. Our analysis revealed that the higher and lower values of the most important SOAP feature correspond to structures exhibiting coordination distortion and high symmetry, respectively. This relationship may transcend conventional defect types and dopant categories, directly capturing the essential impact of the local geometry on adsorption stability. Furthermore, cross-validation and distributional analyses suggested that the high-dimensional physical features can facilitate the understanding of catalyst structure–activity relationships and provide practical technical pathways and theoretical guidance for the machine learning-driven rational design of complex oxide-based catalysts.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2024YFA1509901). Computational resources were provided by Shanghai Jiao Tong University.

References

[1] van Rooij A 2005 *Hist. Technol.* **21** 345
[2] Chisalita D A, Petrescu L and Cormos C C 2020 *Renewable Sustainable Energy Rev.* **130** 109964
[3] Zeng L and Cheng C 2009 *Hydrometallurgy* **98** 1
[4] Doukeh R, Bombos M, Trifoi A, Mihai O, Popovici D, Bolocan I and Bombos D 2018 *C. R. Chim.* **21** 277
[5] Song H and Ye J 2022 *Chem* **8** 1181
[6] Kar S, Goeppert A and Prakash G K S 2019 *J. Am. Chem. Soc.* **141** 12518
[7] Fang R, Dhakshinamoorthy A, Li Y and Garcia H 2020 *Chem. Soc. Rev.* **49** 3638
[8] Park J H, Noh Y H, Kim J S, Song G S, Park S J, Choi J W, Choi Y C and Lee Y J 2025 *Green Chem.* **27** 4750

[9] Zhong H, Shevlin M and Chirik P J 2020 *J. Am. Chem. Soc.* **142** 5272
[10] Zhang X, Wei Y, Song Z, Liu W, Gao C and Luo J 2021 *Chemosphere* **263** 128129
[11] He J, Zhao Z H, Li J, Ren Z, Jiang R, Zhang J and Zhang J 2024 *Chem. Eng. J.* **487** 150427
[12] Hao X, Yoko A, Chen C, Inoue K, Saito M, Seong G, Takami S, Adschiri T and Ikuhara Y 2018 *Small* **14** 1870195
[13] Yoon S, Kim J and An K 2025 *Chem. Catal.* **5** 101423
[14] Werner K, Weng X, Calaza F, Sterrer M, Kropp T, Paier J, Sauer J, Wilde M, Fukutani K, Shaikhutdinov S and Freund H J 2017 *J. Am. Chem. Soc.* **139** 17608
[15] Lee J, Tieu P, Finzel J, Zang W, Yan X, Graham G, Pan X and Christopher P 2023 *JACS Au* **3** 2299
[16] Cao T, You R, Li Z, Zhang X, Li D, Chen S, Zhang Z and Huang W 2020 *Appl. Surf. Sci.* **501** 144120
[17] Liu T, Wu X, Liu K and Liu L 2024 *Mol. Catal.* **569** 114518
[18] Wang Z Q, Chu D R, Zhou H, Wu X P and Gong X Q 2021 *ACS Catal.* **12** 624
[19] Feng Y, Xu W, Huang B, Shao Q, Xu L, Yang S and Huang X 2019 *J. Am. Chem. Soc.* **142** 962
[20] Kwon S G, Krylova G, Sumer A, Schwartz M M, Bunel E E, Marshall C L, Chattopadhyay S, Lee B, Jellinek J and Shevchenko E V 2012 *Nano Lett.* **12** 5382
[21] Wang D J, Sun X C, Yin H J, Dong H, Liu H and Zhang Y W 2024 *ACS Catal.* **14** 10060
[22] Zhao J W, Wang H Y, Feng L, Zhu J Z, Liu J X and Li W X 2023 *Chem. Rev.* **124** 164
[23] Zhou Y, Chen A, Ning J and Shen W 2020 *Chin. J. Catal.* **41** 928
[24] Tada S, Ogata S, Nishikawa S, Yamaguchi H, Nakashima Y, Honma T, Hiramatsu H, Nishijima M, Joutsuka T and Kikuchi R 2025 *Int. J. Hydrogen Energy* **125** 202
[25] Riley C, Zhou S, Kunwar D, Riva A D L, Peterson E, Payne R, Gao L, Lin S, Guo H and Datye A 2018 *J. Am. Chem. Soc.* **140** 12964
[26] Wexler R B, Martirez J M P and Rappe A M 2018 *J. Am. Chem. Soc.* **140** 4678
[27] Raabe D, Mianroodi J R and Neugebauer J 2023 *Nat. Comput. Sci.* **3** 198
[28] Freeze J G, Kelly H R and Batista V S 2019 *Chem. Rev.* **119** 6595
[29] Shen Y, Fu C, Luo W, Liang Z, Wang Z R and Huang Q 2023 *Green Chem.* **25** 7605
[30] Ma F, Chen F, Xu P, Liu X and Zhang W 2025 *J. Mater. Chem. A* **13** 12553
[31] Abraham B M, Jyothirmai M V, Sinha P, Viñes F, Singh J K and Illas F 2024 *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **14** e1730
[32] Kang M, Huang K, Li J, Liu H and Lian C 2024 *J. Mater. Inf.* **4** N/A
[33] Li H, Jiao Y, Davey K and Qiao S Z 2023 *Angew. Chem. Int. Ed.* **62** e202216383
[34] Masood H, Toe C Y, Teoh W Y, Sethu V and Amal R 2019 *ACS Catal.* **9** 11774
[35] Shi Y F, Yang Z X, Ma S, Kang P L, Shang C, Hu P and Liu Z P 2023 *Engineering* **27** 70
[36] Chen D, Kang P L and Liu Z P 2021 *ACS Catal.* **11** 8317
[37] Yang Y, Zhou J, Zhao Z, Sun G, Moniri S, Ophus C, Yang Y, Wei Z, Yuan Y, Zhu C, Liu Y, Sun Q, Jia Q, Heinz H, Ciston J, Ercius P, Sautet P, Huang Y and Miao J 2024 *Nat. Catal.* **7** 796
[38] Kunz M R, Yonge A, Fang Z, Batchu R, Medford A J, Constales D, Yablonsky G and Fushimi R 2021 *Chem. Eng. J.* **420** 129610
[39] Wang X, Ye S, Hu W, Sharman E, Liu R, Liu Y, Luo Y and Jiang J 2020 *J. Am. Chem. Soc.* **142** 7737
[40] Monreal-Corona R, Pla-Quintana A and Poater A 2023 *Trends Chem.* **5** 935
[41] Esterhuizen J A, Goldsmith B R and Linic S 2022 *Nat. Catal.* **5** 175
[42] Tejaswini G, Sudheer A E, Kumar A, Perepu P K, Vallinayagam M, Kamal C, Prakash S M, Posselt M, Zschornak M and Murali D 2025 *J. Mater. Chem. A* **13** 19896
[43] Xu W, Reuter K and Andersen M 2022 *Nat. Comput. Sci.* **2** 443
[44] Bai Z, Lan M, Yu W and Shen S 2025 *Int. J. Mech. Sci.* **294** 110234
[45] Wang Z Q, Chu D R, Zhou H, Wu X P and Gong X Q 2022 *ACS Catal.* **12** 624
[46] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev. B* **87** 184115
[47] Lundberg S and Lee S I 2017 *Advances in Neural Information Processing Systems* **30** 4768