Theoretical Chemistry Accounts (2024) 143:50
https://doi.org/10.1007/s00214-024-03124-x

BRIEF REPORT

DFT and machine learning for predicting hydrogen adsorption energies on rocksalt complex oxides

Adrian Domínguez-Castro[1]

Received: 6 March 2024 / Accepted: 15 May 2024

Abstract
The prediction of hydrogen adsorption energies on complex oxides by integrating DFT calculations and machine learning is considered. In particular, 14 descriptors for electronic and geometric properties evaluation are adapted within a 336 hydrogen adsorption energy dataset created. Supervised learning techniques were explored to establish an accurate predictive model. With the deep neural network results, a MAE of about 0.06 eV is achieved. This research highlights the synergistic potential of DFT and machine learning for accelerating the exploration of materials for catalysis.

Keywords Hydrogen adsorption energy · DFT · Machine learning · Deep neural network

1 Introduction

The unique properties of high-entropy oxides with the mixing of multiple cations in the complex structure significantly influence the structural, electronic, and thermal properties, making this class of materials promising candidates for a wide range of applications in materials science. [1] The study of hydrogen adsorption on high-entropy oxides is crucial for advancing hydrogen storage, catalysis, and material design, with potential implications for energy storage, conversion, and environmental sustainability. Additionally, predicting hydrogen adsorption energies (HAEs) is of paramount significance in advancing the field of sustainable energy and catalysis, particularly concerning catalytic processes involving C-H activation on surfaces. [2] A valuable tool to address the challenges and support experiments in materials science, providing a deeper understanding, is density functional theory (DFT) calculations. [3] DFT plays a pivotal role in modeling reaction pathways, understanding reaction kinetics, elucidating the electronic structure of materials, and guiding experimental efforts. [4–10] In specific, the exploration of chemical space, for practical purposes the sinergistic integration of machine learning with DFT, has impulsed materials science and computational chemistry. [11, 12] Machine learning techniques, such as neural networks and regression models, complement the computational capabilities of DFT by accelerating the prediction of complex materials properties and electronic structures. [13] Machine learning algorithms trained on DFT-derived datasets can discern intricate patterns and correlations within vast parameter spaces, leading to the prediction of energetics, reaction pathways, and adsorption properties with remarkable accuracy. [14] This powerful combination expedites materials discovery and optimization, making it possible to screen and prioritize candidates for experimental validation. The marriage of machine learning and DFT not only expedites materials design but also opens avenues for uncovering new insights into the fundamental principles governing catalytic processes, thereby advancing our ability to engineer materials with tailored properties for diverse applications. In this context, can the integration of DFT calculations with machine learning offer an efficient approach to accurately predict hydrogen adsorption energies on complex oxides materials? A starting point for the development of a study in this topic is the valid selection of an accurate representation model to mimic high-entropy oxides with a diverse cation arrangements. For this purpose, to address the critical challenge of predicting hydrogen adsorption energies, this research focuses on the complex oxide $Ni_{16}Mg_{16}Cu_{16}Zn_{16}O_{64}$. The presence of multiple cations in the model promotes the formation of complex structural motifs, similar to those observed in high-entropy oxides. Due to the presence of multiple cations, using a single complex composition like $Ni_{16}Mg_{16}Cu_{16}Zn_{16}O_{64}$ can reduce the number of individual models needed to explore a range of

✉ Adrian Domínguez-Castro
adriandc1989@gmail.com

[1] Department of Chemical and Biomolecular Engineering, Vanderbilt University, Nashville, TN 37235, USA

Published online: 04 June 2024

compositions and structures, thereby enhancing computational efficiency. The methodology employed here can represent an improvement in the screening of high-entropy oxides because the computation of the descriptors employed is obtained by a reduction on the computational cost needed, with only a DFT optimization geometry calculation of the slab in consideration. Making the accurate prediction of the hydrogen adsorption energy more accessible.

2 Computational methods

The database was made by the following workflow: First, randomly generated oxide slabs, with the $Ni_{16}Mg_{16}Cu_{16}Zn_{16}O_{64}$ system fixed composition. These systems present a fascinating exploration of complex materials with potential applications in various technological domains. Derived from a rock crystal salt of MgO, this high-entropy composition introduces multiple cationic species, including Ni, Mg, Cu, Zn, and combined with O. The incorporation of diverse metal cations into the oxide structure enhances the structural disorder and complexity, contributing to unique electronic, optical, and mechanical properties. From the initial MgO bulk, a 2x2x2 supercell exposing the -100 surface with four layers is created. Then, by using this MgO supercell, a $Ni_{16}Mg_{16}Cu_{16}Zn_{16}O_{64}$ slab is created and randomly distributed the Ni, Mg, Cu, and Zn cations, making a total of 21 different slab models. These are a good model system for studying surface reactions and capture a diversity in the local environment of the adsorption sites. Second, the hydrogen adsorption on oxygen sites was performed by DFT calculations using the Vienna Ab initio Simulation Package (VASP) code [15–18] and the PBE exchange-correlation functional [19] with the DFT-D3 correction scheme. The accuracy of PBE functional is a common practice based on its computational efficiency and accuracy for structural properties, such as bond lengths. It is well know that PBE functional underestimates band gaps and it is an inadequate choice for magnetic properties. [19] For the problem in consideration, it is still a valid approximation. In these calculations, periodic boundary conditions are employed. The projector-augmented wave (PAW) method is used. [20] The Brillouin zone is sampled using a 2x2x1 gamma-centered grid and an energy cutoff of 400 eV. Spin polarization is considered. The hydrogen adsorption energies were calculated as:

$$E_{ads-H} = E_{surf+H} - E_{surf} - \frac{1}{2}E_{H_2}$$

Where E_{surf+H}, E_{surf}, and E_{H_2} are the energies for the adsorbate/surface complex, the catalyst surface, and the adsorbate, respectively.

Third, a set of descriptors is considered to capture the electronic and geometric properties of the oxygen adsorption sites on the complex oxide structures under study. For a full list, the set of descriptors is summarized in Table 1. The Bader net charge [21] and average oxygen 2p-state energy descriptors are key metrics in understanding the electronic structure. The Bader net charge is a measure of the charge localized around a specific atomic nucleus according to the Bader analysis. It provides valuable information about the electron density and charge transfer within a molecule or material. On the other hand, the average oxygen 2p-state energy is a descriptor related to the energy levels of oxygen's 2p orbitals, and it is calculated as in ref [22]:

$$\bar{\epsilon}_{2p} = \frac{\int_{\epsilon_{\min}}^{\epsilon_{\max}} \rho_{2p}\epsilon \mathrm{d}\epsilon}{\int_{\epsilon_{\min}}^{\epsilon_{\max}} \rho_{2p}\mathrm{d}\epsilon}$$

The VASPKIT package was used to post-processing. [23]

To replicate the specific chemical surroundings, a new predictor is introduced here for the first time and implemented is the Bond Length-Valence Electron (BLVE) ratio descriptor:

$$\mathrm{BLVE} = \frac{\mathrm{distance(Metal\text{-}O)}}{\mathrm{Valence}_{e^-}\mathrm{(Metal)}}$$

This descriptor is designed to quantify the ratio between the distance of metal-oxygen and the valence electron numbers of the corresponding metal, for each case. Providing insights into the bond lengths weighted by valence electrons and offering a principle of combination of electronic structure and bonding properties. In specific, five BLVE metrics are described, one for each of the five metals around the oxygen adsorption site (four of the metals in the first layer and the number 5 belonging to the second layer of the slab). Another group of descriptors related to the chemical nature of the environment surrounded the oxygen adsorption site

Table 1 List of descriptors

Descriptor 1	Bader net charge
Descriptor 2	Average oxygen 2p (up)
Descriptor 3	Average oxygen 2p (down)
Descriptor 4	Bond length-valence electron ratio (neighbor metal 1)
Descriptor 5	Bond length-valence electron ratio (neighbor metal 2)
Descriptor 6	Bond length-valence electron ratio (neighbor metal 3)
Descriptor 7	Bond length-valence electron ratio (neighbor metal 4)
Descriptor 8	Bond length-valence electron ratio (neighbor metal 5)
Descriptor 9	Frequency of Ni around oxygen adsorption site
Descriptor 10	Frequency of Cu around oxygen adsorption site
Descriptor 11	Frequency of Mg around oxygen adsorption site
Descriptor 12	Frequency of Zn around oxygen adsorption site
Descriptor 13	Average difference in electronegativity between oxygen and neighboring metals
Descriptor 14	Average difference in ionization energy between oxygen and neighboring metals

is the frequency of metals around the active site allowing to a consideration into the structural factors. In this case, this metric for Ni, Cu, Mg, and Zn is an integer value from 0 to 5. Additionally, the average difference in electronegativity and ionization energy between oxygen and neighboring metals is considered. These last two metrics provide a valuable information about the bonding characteristics and the description of the polarity of the oxygen active site. Giving a total of 14 descriptors in consideration within 336 hydrogen adsorption energy points dataset. A more detailed analysis of the descriptors in consideration and one example about how those descriptors are calculated is referred in the **Supporting Information**.

3 Results and discussion

The random arrangement of these metal cations within the oxide lattice reflects the presence of different local chemical environments, as shown in Fig. 1b for one of the randomly generated structures. The distribution of DFT calculated hydrogen adsorption energies is displayed in an histogram, exposed in Fig. 1a with a range of approximately 2.2 eV. This effect has been reported before for high-entropy oxides and high-entropy alloys. [24, 25] The complex relationships and patterns within the data are investigated using supervised machine learning algorithms to efficiently predict hydrogen adsorption energies. The dataset is divided into training and testing sets. This division is structured as 80% of the dataset is allocated for training, while the remaining 20% is reserved for testing.

An exploratory data analysis based on the correlation matrix of the features of the dataset can reveal important insights into the relationships between the variables involved. (See the correlation matrix obtained in the Supporting Information) The correlation factors between the hydrogen adsorption energy and the descriptors of the dataset are of weak-to-moderate correlations, indicating a modest relationship. The highest values are −0.4 and −0.23 for the frequency of Cu atoms surrounding the oxygen adsorption site and the Bader net atomic charge of oxygen descriptors. The correlation factor of −0.4 between hydrogen adsorption energy and the frequency of Cu atoms surrounding the oxygen adsorption site indicates a moderate negative correlation. A negative correlation indicates that as the descriptor related to Cu atoms increases, the hydrogen adsorption energy tends to decrease, suggesting a stronger adsorption of hydrogen. Similar, as the Bader net atomic charge of oxygen increases, reflecting a higher charge density near the oxygen adsorption site, the hydrogen adsorption energy decreases, indicating a stronger adsorption of hydrogen. To the best of my knowledge, there is no single descriptor described in the literature capable of an accurate description of the hydrogen adsorption energy observable study here. From the point of view of feature engineering, it is possible that the descriptors explored in the present work, while not present high correlation with the hydrogen adsorption energy, could be combined in a way that makes them more relevant.

Figure 2 illustrates the model performances in a parity plot. Notably, the mean absolute error (MAE) values are reported, providing a quantitative measure of the model's accuracy. Linear regression models are presented on the top of Fig. 2. The one based on one descriptor (Bader net atomic charge of oxygen) in Fig. 2 top left and the multi-feature linear regression based on the full set of 14 descriptors in Fig. 2 top right have a testing MAE of approx 0.19 eV in both cases. Moving to the RF (Fig. 2 bottom left) decreased the testing MAE to lightly 0.17 eV. These algorithms were created using the scikit-learn

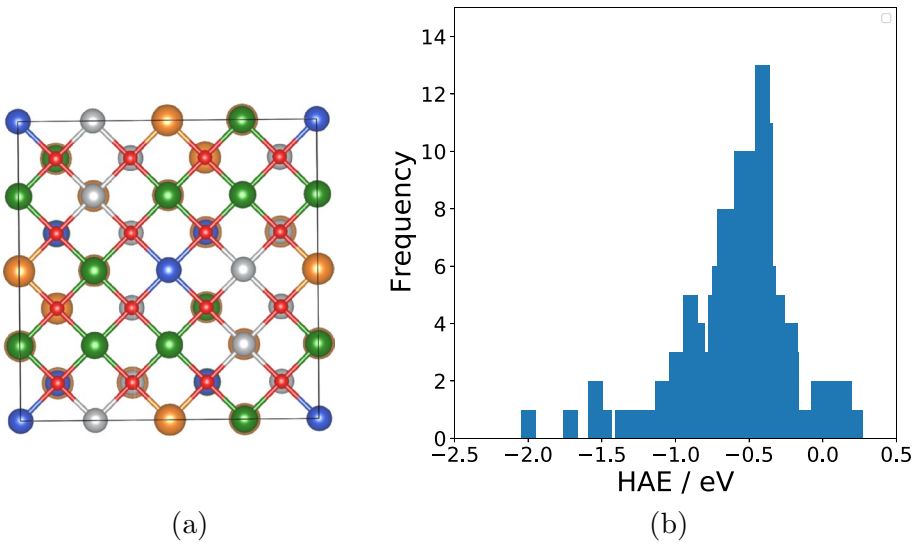

Fig. 1 **a** Top view of one of the structures generated of the complex oxide under study. Ni, Mg, Cu, Zn, and O atoms are represented in *green*, *orange*, *blue*, *gray*, and *red*, respectively. **b** Distribution of DFT calculated hydrogen adsorption energies

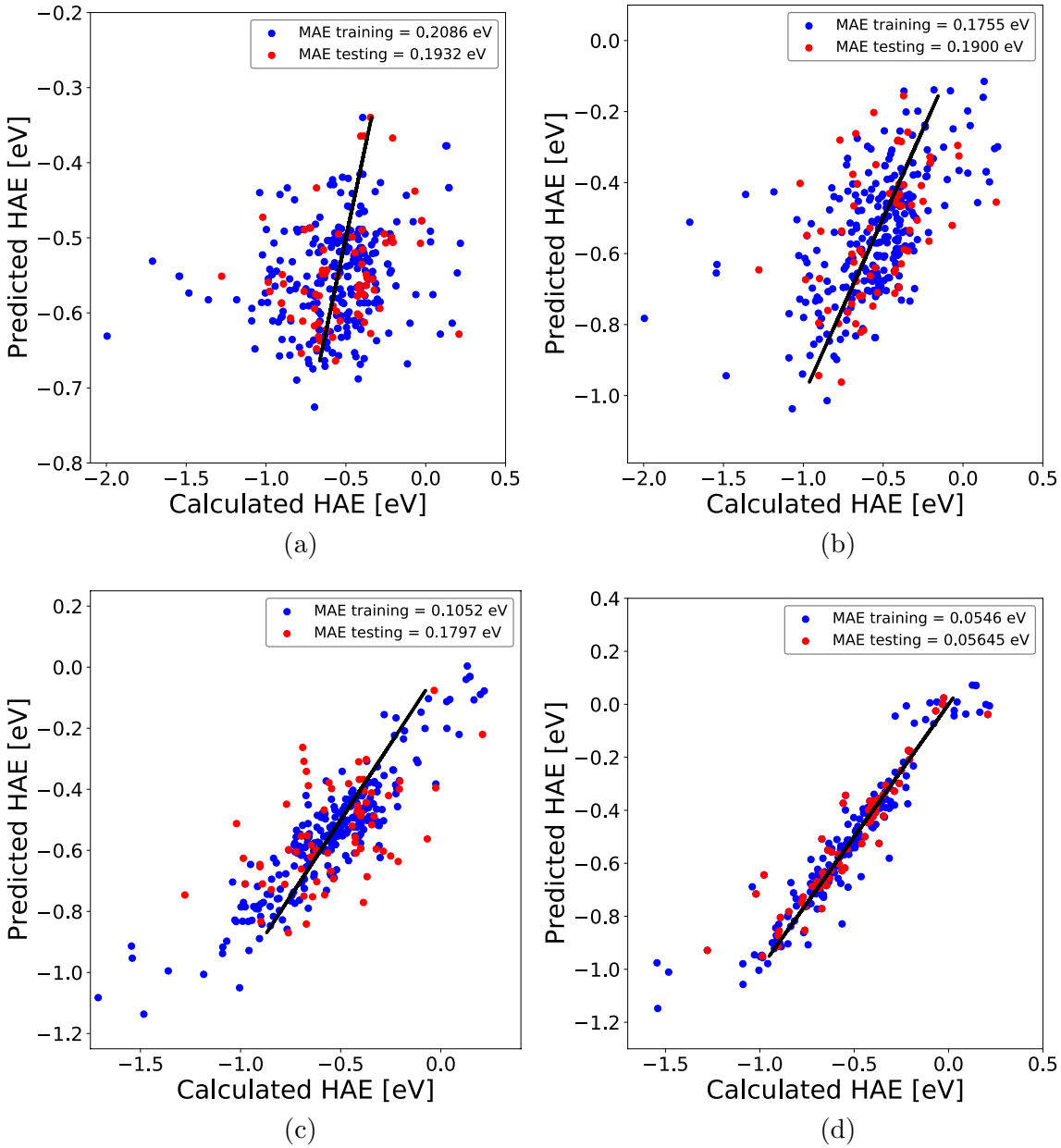

Fig. 2 Comparison of DFT calculated and Machine Learning predicted hydrogen adsorption energies: **a** LR-one-feature, **b** LR-multi-feature, **c** RF, and **d** DNN

Table 2 Performance comparison of machine learning methods in the prediction of hydrogen adsorption energies

	MAE training	MAE testing	R^2 training	R^2 testing
One-feature linear regression	0.2086	0.1932	0.0475	0.0653
Multi-feature linear regression	0.1755	0.1900	0.3218	0.1299
Random forest	0.1052	0.1797	0.7670	0.2550
Deep neural network	0.0546	0.0564	0.8950	0.8680

python library. [26] Moving to the DNN model (Fig. 2 bottom right), the testing MAE was reduced with significance, compared to the linear regression with multiple features, from a testing MAE of 0.19 eV to around 0.06 eV. The DNN was constructed using the Keras python library. [27] The architecture under consideration is composed of five

hidden layers, each layer houses 64 neurons. To mitigate overfitting, a dropout rate of 0.25 is applied in each layer, ensuring robust generalization. The incorporation of an L2 kernel regularizer further bolsters the network's ability to generalize by imposing a penalty on large weights. The rectified linear unit (ReLU) activation function, the Adam optimizer, and a learning rate of 0.0001 are employed. The model's robustness is further assessed through KFold cross-validation with five splits, providing a comprehensive evaluation of its performance across different subsets of the dataset.

The results of model performances are summarized in Table 2. The hydrogen adsorption energies and Bader net atomic charges of oxygen are basically not related; indeed is something expected from the analysis of the correlation factor mentioned above. This is also supported by the linear regression plot. The DNN model presents very good performance. The R2 value of 0.868 is relatively high and indicates that the model explains a significant portion (86.8%) of the variance in the predictions. This suggests that the model is performing well in terms of explaining the variability in the target variable (hydrogen adsorption energy) based on the input features (descriptors) used. It is important to mention that there are some cases with large errors in predicted hydrogen adsorption energies, mainly from the computed extreme values. (high and lower values of the distribution of hydrogen adsorption energies on the oxide slab.) Improving our model is possible through the addition of more data to the training dataset, particularly focusing on local chemical environments that are currently underrepresented in the dataset.

The findings of this study hold significant implications with a balance between predictive accuracy and model interpretability. The MAE can be reduced from 0.19 eV in the linear regression to 0.06 eV with a deep neural network. This improving in the accuracy represents a valid approach for a rapid computational screening of complex oxide materials, with a significance to assisting in the understanding of structure-reactivity relationship of high-entropy oxides. The present study adds further insights into the still open discussion about the exploration of a universal descriptor or features that have a stronger relationship with the target observable. With a set of descriptors related to electronic, structural and chemical information collectively capture relevant information when considered together underlying the powerful combination of DFT calculations and machine learning methods. Future work may involve the use of the methodology discussed to explore high-entropy oxides materials and efficiently predict hydrogen adsorption energies and advancing materials design for a sustainable environment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00214-024-03124-x.

Acknowledgements The author acknowledges Vanderbilt University for the Postdoctoral Fellowship and Prof. De-en Jiang for the insightful discussions.

Data Availability The full data and the processing Jupyter notebook used for the present paper are available at https://github.com/DrAdrianDC/DFT-and-ML.

Declarations

Conflict of interest The author declares no conflict of interest.

References

1. Sarkar A, Wang Q, Schiele A, Chellali MR, Bhattacharya SS, Wang D, Brezesinski T, Hahn H, Velasco L, Breitung B (2019) Adv Mater 31:1806236
2. Latimer AA, Kulkarni AR, Aljama H, Montoya JH, Yoo JS, Tsai C, Abild-Pedersen F, Studt F, Nørskov JK (2017) Understanding trends in C-H bond activation in heterogeneous catalysis. Nat Mater 16:225–229
3. Kohn W, Becke AD, Parr RG (1996) Density functional theory of electronic structure. J Phys Chem 100(31):12974–12980
4. Wagstaffe M, Wenthaus L, Dominguez-Castro A et al (2020) Ultrafast real-time dynamics of CO oxidation over an oxide photocatalyst. ACS Catalysis 10(22):13650–13658
5. Wagstaffe M, Dominguez-Castro A, Wenthaus L et al (2023) Photoinduced dynamics at the water/TiO2 (101) interface. Phys Rev Lett 130(10):108001
6. Domínguez-Castro A, Guzmán F, Novo-Fernández Y (2017) Adsorption on a nanoporous organic polymer for clean energy applications: a multiscale modeling study using density functional tight binding approach. Comput Theor Chem 1102:30–37
7. Domínguez-Castro A, Hernández D, Guzmán F (2017) Insights into the interactions of biomolecules with small gold clusters: a theoretical study from a DFTB perspective. Theor Chem Accounts 136:84
8. Dominguez-Castro A, Lien-Medrano CR, Maghrebi K, Messaoudi S, Frauenheim Th, Fihey A (2021) Photoinduced charge-transfer in chromophore-labeled gold nanoclusters: quantum evidence of the critical role of ligands and vibronic couplings. Nanoscale 13:6786–6797
9. Castañeda-Arriaga R, Domínguez-Castro A, Lee J, Alvarez-Idaboy JR, Mora-Diez N (2016) Chemical repair of protein carbon-centred radicals: long-distance dynamic factors. Can J Chem 94(12):1119–1126
10. Domínguez-Castro A, Frauenheim Th (2021) Impact of vibronic coupling effects on light-driven charge transfer in pyrene-functionalized middle and large-sized metalloid gold nanoclusters from Ehrenfest dynamics. Phys Chem Chem Phys 23(32):17129–17133
11. Singh AR, Rohr BA, Gauthier JA et al (2019) Predicting chemical reaction barriers with a machine learning model. Catal Lett 149:2347–2354
12. Takahashi K, Miyazato I (2018) Rapid estimation of activation energy in heterogeneous catalytic reactions via machine learning. J Comput Chem 39:2405–2408
13. Roy D, Mandal SCh, Pathak B (2022) Machine learning assisted exploration of high entropy alloy-based catalysts for selective CO2 reduction to methanol. J Phys Chem Lett 13(25):5991–6002
14. Göltl F, Mavrikakis M (2022) Generalized Brønsted–Evans–Polanyi relationships for reactions on metal surfaces from machine learning. ChemCatChem 14:e202201108

15. Kresse G, Hafner J (1993) Ab initio molecular dynamics for liquid metals. Phys Rev B Condens Matter Mater Phys 47:558(R)
16. Kresse G, Hafner J (1994) Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. Phys Rev B Condens Matter Mater Phys 49:14251
17. Kresse G, Furthmuller J (1996) Efficiency of Ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci 6:15–50
18. Kresse G, Furthmuller J (1996) Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B 54:11169
19. Perdew JP, Ernzerhof M, Burke K (1996) Rationale for mixing exact exchange with density functional approximations. J Chem Phys 105:9982–9985
20. Blöchl PE Projector augmented-wave method. Phys Rev B 50:17953
21. Bader RFW (1991) A quantum theory of molecular structure and its applications. Chem Rev 91(5):893–928
22. Dickens CF, Montoya JH, Kulkarni AR, Bajdich M, Nørskov JK (2019) An electronic structure descriptor for oxygen reactivity at metal and metal-oxide surfaces. Surf Sci 681:122–129
23. Wang V, Xu N, Liu JC, Tang G, Geng WT (2021) VASPKIT: a user-friendly interface facilitating high-throughput computing and analysis using VASP code. Comput Phys Commun 267:108033
24. Batchelor ThAA, Pedersen JK, Winther SH, Castelli IE, Jacobsen KW, Rossmeisl J (2019) High-entropy alloys as a discovery platform for electrocatalysis. Joule 3:834–845
25. Svane KL, Rossmeisl J (2022) Theoretical optimization of compositions of high-entropy oxides for the oxygen evolution reaction. Angew Chem Int Ed 61:e202201146
26. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825
27. Chollet F, et al (2015) Keras [Internet]. GitHub. Available from: https://github.com/fchollet/keras