npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Using statistical learning to predict interactions between single metal atoms and modified MgO(100) supports

Chun-Yen Liu[1,4], Shijia Zhang[2,4], Daniel Martinez[1], Meng Li[3]✉ and Thomas P. Senftle[1]✉

Metal/oxide interactions mediated by charge transfer influence reactivity and stability in numerous heterogeneous catalysts. In this work, we use density functional theory (DFT) and statistical learning (SL) to derive models for predicting how the adsorption strength of metal atoms on MgO(100) surfaces can be enhanced by modifications of the support. MgO(100) in its pristine form is relatively unreactive, and thus is ideal for examining ways in which its electronic interactions with metals can be enhanced, tuned, and controlled. We find that the charge transfer characteristics of MgO are readily modified either by adsorbates on the surface (e.g., H, OH, F, and NO_2) or dopants in the oxide lattice (e.g., Li, Na, B, and Al). We use SL methods (i.e., LASSO, Horseshoe prior, and Dirichlet–Laplace prior) that are trained against DFT data to identify physical descriptors for predicting how the adsorption energy of metal atoms will change in response to support modification. These SL-derived feature selection tools are used to screen through more than one million candidate descriptors that are generated from simple chemical properties of the adsorbed metals, MgO, dopants, and adsorbates. Among the tested SL tools, we demonstrate that Dirichlet–Laplace prior predicts metal adsorption energies on MgO most accurately, while also identifying descriptors that are most transferable to chemically similar oxides, such as CaO, BaO, and ZnO.

npj Computational Materials (2020)6:102 ; https://doi.org/10.1038/s41524-020-00371-x

INTRODUCTION

Transition metals (TMs) supported on oxide surfaces are ubiquitous heterogeneous catalysts in chemical processes that produce fuel and value-added chemicals, as well as in processes central to renewable energy and environmental protection technologies. Among supported TM catalysts, single-atom catalysts (SACs) have attracted attention due to their enhanced performance in many applications[1], including CO oxidation[2–10], water–gas shift[11–13], selective hydrogenation[14–17], dehydrogenation[18–21], and photocatalytic[22] reactions. SACs with uniform TM dispersion expose every TM atom to the reaction environment, thus maximizing utilization of the expensive TM component. Key to achieving such uniform dispersion is the ability to control metal cluster sizes, which in turn is largely dictated by the TM's strength of interaction with the support. Predicting the strength of metal/support interactions is a challenging task because there are multiple factors at play, such as the reducibility of the oxide, the electronegativity of the metal, and the structure of the interface. Herein, we use density functional theory (DFT) and statistical learning (SL) to identify physical descriptors that build a predictive model for describing metal atom binding on MgO(100) surfaces. Focus is placed on understanding and predicting how modifications of the unreactive MgO(100) surface, through the introduction of surface adsorbates or dopants, can enhance metal binding energies. We also focus on comparing the performance of various SL approaches, where we show that the physical descriptors identified with Dirichlet–Laplace prior[23] are strong descriptors for predicting metal binding on MgO(100). Furthermore, we show that descriptors identified with MgO training sets also can be used to describe TM binding on modified CaO(100), BaO(100), and ZnO (100) surfaces—demonstrating the transferability of SL-derived descriptors beyond the MgO system they were trained to describe, albeit for very closely related systems.

Metal/support interactions impact both the morphology of the catalyst surface and oxidation states at the active site, which both influence catalytic performance[24,25]. The size of metal clusters on oxide supports is controlled by thermodynamic driving forces that usually cause small clusters to agglomerate (e.g., through Ostwald ripening)[26]. Metal/support interactions alter the chemical potential of each metal atom in the cluster, and thus influence thermodynamic stability with respect to cluster size[27,28]. These interactions also influence the kinetics of cluster formation, where an increase in the metal adsorption energy on the oxide surface generally reduces sintering rates leading to smaller particle sizes[29,30]. Metal/support interactions affect not only particle size distributions, but also alter the electronic state of the adsorbed metal[31–34]. This phenomenon, known as electronic metal–support interaction[35], is caused by charge transfer between the metal and the support, which can alter reaction rates by affecting how strongly intermediates adsorb at the active site.

Adsorbates from the reaction environment, or dopants present in the oxide structure, can enhance charge transfer at the metal/support interface[31,36–40]. Addou et al.[41] demonstrated that adsorbed hydrogen atoms on the rutile TiO_2(011)-(2 × 1) surface alters the binding energy of Pd atoms to the support, which stabilizes small Pd_n clusters (i.e., $n \sim 1$–3 atoms). Babucci et al.[42] used Fourier-transform infrared spectroscopy to show that the electronic state of $Ir(CO)_2$ can be altered on oxide supports that are modified with different ligands. They demonstrated that Ir complexes exhibit different reactivity toward 1,3-butene hydrogenation if the surface modifications shift the Ir oxidation state at the active site. Kumar et al.[43] used Hammett reactivity studies, together with DFT, to show that the electronic state of Au during

[1]Department of Chemical and Biomolecular Engineering, Rice University, Houston, TX 77005, USA. [2]Department of Computer Science and Engineering, School of Electrical Engineering and Computer Science, Pennsylvania State University, University Park, PA 16802, USA. [3]Department of Statistics, Rice University, Houston, TX 77005, USA. [4]These authors contributed equally: Chun-Yen Liu, Shijia Zhang. ✉email: meng@rice.edu; tsenftle@rice.edu

benzyl alcohol oxidation is influenced by electron donation from the supporting oxide. Similar effects have been observed on surfaces that are modified by introducing dopants in the lattice of the oxide. For example, Shao et al.[44] reported that the electronic structure of CaO surfaces can be controlled by introducing Mo dopants that replace Ca atoms in the surface lattice. The additional valence electrons supplied by Mo migrate toward adsorbed Au clusters, thus enhancing Au binding on the modified CaO surfaces. This effect was shown to be less pronounced over Cr-doped MgO surfaces[45,46], which demonstrates that the dopant effect varies between different types of dopants and oxide supports.

Given the impact that metal/support interactions can have on catalyst morphology and activity, it is clear that identifying physical descriptors for predicting interaction trends will be of high value. Toward this end, Campbell and Sellers[47] proposed that the adsorbed metal's enthalpy of oxide formation computed relative to the isolated metal atom ($\Delta H_{f,ox,atom}$) should be an effective descriptor for predicting metal adsorption energies on oxide surfaces; $\Delta H_{f,ox,atom}$ captures the bonding strength between the metal atom and oxygen, and it should therefore correlate with the metal's binding strength to oxide surfaces. They demonstrated a linear correlation between $\Delta H_{f,ox,atom}$ and metal adhesion energies measured by adsorption calorimetry, thus deriving a simple model for predicting metal adsorption energies based on readily available reference data[27]. Using both isothermal titration calorimetry and DFT, Strayer et al.[30] verified that $\Delta H_{f,ox,atom}$ is also an effective descriptor for predicting metal adsorption energies on $HCa_2Nb_3O_{10}$ perovskite surfaces. Although successful in these cases, identifying physical descriptors purely from chemical intuition is not a simple task. Complex charge transfer between the metals, the oxide supports, and surface modifiers (i.e., adsorbates and dopants) is challenging to describe with closed physical forms motivated from chemical intuition alone, as multiple phenomena are occurring simultaneously. As such, in this work we apply SL to identify useful physical descriptors that capture charge transfer at complex, multicomponent interfaces. We choose MgO for our initial study to demonstrate the effectiveness of the SL approach, since charge transfer characteristics can be anticipated from the non-reducible characteristics of MgO. Although we do not expect all surface modifications of MgO studied here to be experimentally feasible because dopants/adsorbates can generate charge-compensating defects that would counteract charge transfer to the adsorbed metal[45], the idealized nature of pristine MgO is a desirable testbed for evaluating the performance of various SL approaches and for identifying physical descriptors to predict how charge transfer affects metal adsorption in the idealized case. Once the strengths and weaknesses of the SL approaches are known, then we can pursue studies that use well-chosen SL approaches to understand more complex, but experimentally realizable, oxide systems containing defects.

Multiple feature selection (FS) methods have been introduced in the materials community to identify physical descriptors from extremely large sets of candidate descriptors. FS methods are preceded by gathering the chemical properties of the system's components from various databases and using these properties to generate a large feature space of candidate descriptors. This step, referred to as feature engineering, is achieved by applying a series of mathematical operations on all descriptor pairs to enumerate millions (or even billions) of possible descriptor combinations and functional forms. FS methods are then applied to identify the features from this pool that are the strongest predictors of the property of interest (Fig. 1). There is a rich menu of FS methods in SL, ranging from traditional principle analysis component (PCA)[48] or kernel ridge regression (KRR)[49,50] to large scale methods, such as least absolute shrinkage and selection operator with l_0 norm regression (LASSO + l_0)[51,52] or sure independence screening and sparsifying operator (SISSO)[53,54]. The success of these FS approaches in the materials community is demonstrated by many examples[55,56], including the work of Ghiringhelli et al.[51], who used LASSO + l_0 regression to identify physical descriptors that can predict crystal structures based on material composition. Similarly, Andersen et al.[54] used SISSO to derive descriptor subsets that can predict molecular adsorption energies on metal alloys. Alternative to these frequentist approaches, Bayesian FS has emerged in SL as a primary tool that provides a coherent and principled framework to quantify uncertainties[57–60]. In addition to inheriting the advantage of Bayesian methods to conveniently incorporate domain knowledge via prior distributions, state-of-the-art Bayesian FS is able to adapt to unknown sparsity levels in the feature space[61,62] and to achieve automatic multiplicity correction[63,64]. In this work, we will explore the performance of both LASSO-based FS methods and state-of-the-art Bayesian FS methods[23,65] for finding descriptors that can predict changes in metal binding energy caused by MgO surface modification.

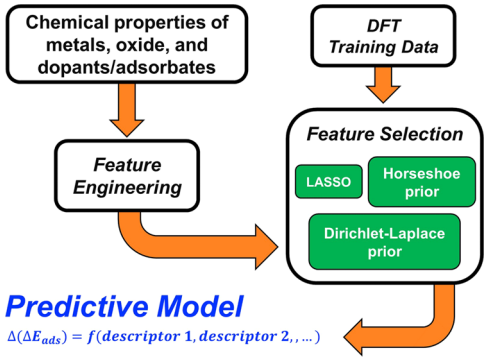

Fig. 1 Statistical learning workflow. The predictive model is built from statistical learning methods using fundamental chemical properties as the descriptors and the DFT energy as the training set.

In the present work, DFT binding energies of single metal atoms on modified MgO surfaces are used to train our SL models. Training and validation sets are built from data collected for MgO (100) surfaces that are modified by either adsorbates or dopants. MgO(100) is an irreducible oxide that, in its pristine form, binds TMs weakly compared to other supports[52]. It therefore is ideal for investigating the effects of support modifications, as changes in its binding interaction with the supported metal can be attributed solely to the effects of the surface modification. Indeed, MgO has been used as a template for many investigations of metal/support interactions, such as demonstrating how the oxidation state of gold varies upon surface modification[39,66] and how linear scaling relationships can be derived for a wide variety of adsorbates at Au/oxide interfaces[67]. We modified the MgO substrate with surface dopants and adsorbates to induce electron-poor and electron-rich conditions, where we find that both lead to an enhancement of the metal binding energies due to an increase in charge transfer between the metal and the support, as expected from the irreducible nature of pristine MgO.

We identify descriptors for predicting the effects of surface modifications by using multiple FS methods: LASSO[68], Dirichlet–Laplace prior[23], and Horseshoe prior[65] (Fig. 1). The methods are applied to identify physical descriptors for predicting the enhancement of metal binding energies on modified MgO based on readily available chemical properties of the system's components (i.e., the adsorbed metal, the MgO surface, and the adsorbate or dopant). The selected feature spaces are then refined with l_0 norm[51] regression to build predictive models that describe single metal atom binding on MgO surfaces. We find that Dirichlet–Laplace prior shows outstanding FS properties, as it mitigates many disadvantages evident in the performance of LASSO and Horseshoe prior. We evaluate the transferability of the

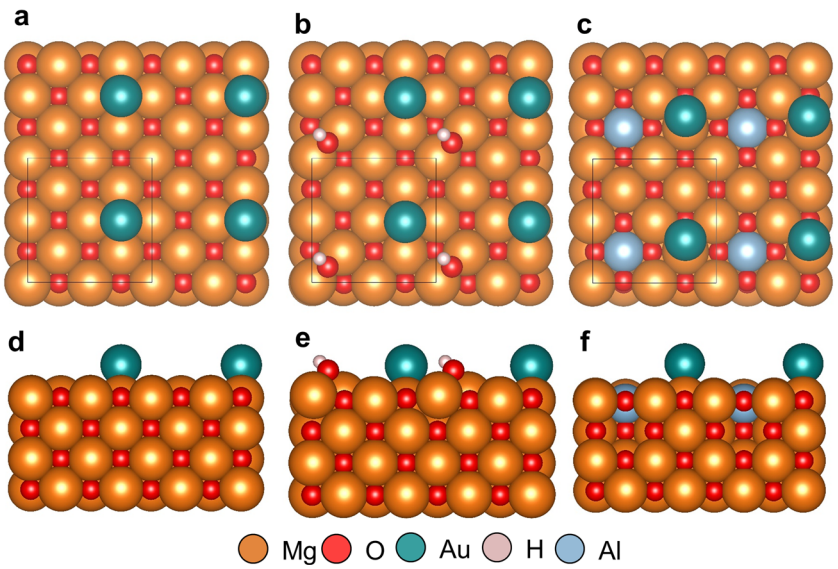

Fig. 2 Metal atom adsorption on MgO surfaces. Top/side view geometries of Au adsorption on **a**/**d** pristine, **b**/**e** OH-modified, and **c**/**f** Al-doped MgO(100) surfaces.

identified features of each method by applying them to predict changes in metal adsorption over CaO(100), BaO(100), and ZnO (100) surfaces. We find that features identified by Dirichlet–Laplace prior using the MgO training data are effective descriptors for predicting metal binding on these irreducible surfaces as well, which again demonstrates the robust nature of Dirichlet–Laplace prior because no CaO, BaO, and ZnO data were included in the FS procedure. It also demonstrates the transferability of SL-derived features within oxide families, which widens model applicability to systems that were not used explicitly to generate the training set.

RESULTS

Effects of dopants and adsorbates on metal adsorption energy

Our objective is to build a model that captures the relationship between metal adsorption energy and readily available chemical properties of the adsorbed metals, MgO, dopants, and adsorbates. A range of TMs (Ag, Au, Cd, Co, Cr, Cu, Fe, Ir, Mn, Mo, Nb, Ni, Pd, Pt, Rh, Ru, V, W, and Zn) were adsorbed on MgO(100) surfaces to generate our DFT training sets. The metal adsorption energy (ΔE_{ads}) is calculated using Eq. 1, where $E_{Metal/Surface}$ is the total DFT energy of the metal atom adsorbed on the MgO(100) surface at its most stable site, E_{Metal} is the total DFT energy of the isolated metal atom, and $E_{Surface}$ is the total DFT energy of the clean MgO(100) surface.

$$\Delta E_{ads} = E_{Meta/Surface} - E_{Metal} - E_{Surface}. \quad (1)$$

There are four unique adsorption sites on the MgO(100) surface: an atop site on O, an atop site on Mg, a bridge site between O and Mg, and a hollow site. Adsorption at all four sites was tested to find the most favorable adsorption site for each metal. The atop site above the oxygen anion was found to be the most stable adsorption site for every metal, in agreement with previous reports[52,69,70]. A representative structure of Au adsorbed on MgO (100) is shown in Fig. 2; similar geometries were obtained for all other metals.

We modified the MgO(100) surface by introducing dopants and adsorbates to enhance the reducibility of the support, which we find generally increases metal adsorption strengths. We first considered H and OH adsorbates, which are commonly encountered species under catalytic reaction conditions. Adsorption of H and OH on four unique surface sites was tested to identify the most stable site, which we found was the atop site on surface O for H and the surface hollow site for OH. The metal atoms were then placed on the optimized structures at the atop O site and then the entire system was optimized to calculate the metal binding energies. A representative structure of Au adsorbed on the OH-MgO(100) surface is shown in Fig. 2; other adsorbate-modified surfaces are shown in Supplementary Fig. 1. F and NO_2 were also introduced to expand the DFT training data, where F has a higher electron affinity (EA) (3.40 eV) than OH (1.83 eV) and thus is expected to yield to a stronger metal binding enhancement compared to OH. NO_2 was included because it can readily act as a neutral, positive, or negative species based on the chemical environment, thus diversifying the training set. Each metal atom was placed as far from the adsorbate as possible within the cell to prevent direct interaction between the metal and the adsorbate. This isolates the adsorbate's impact on the properties of the support from effects related to direct metal–adsorbate interactions. Ag, Nb, and Ru formed direct bonds to NO_2 during all optimization attempts, so their binding energies were excluded from the data set so that only surface-mediated effects were considered. Doping the oxide surface alters the electronic structure of the oxide in a manner similar to the adsorbates[37,39,44–46]. Li, Na, B, and Al were chosen as dopants in this study, as they contain either one fewer or one additional valance electron compared to Mg. The dopants replaced one Mg atom in the first layer of the oxide (Fig. 2c, f).

All modifications of the MgO(100) surface enhanced charge transfer between the surface and the supported metal, as shown by the charge distributions in Fig. 3. When H is adsorbed on Au/MgO(100), one electron is transferred from H to the metal adatom (Fig. 3b). Conversely, OH is electron withdrawing causing charge to be drawn away from Au (Fig. 3c). For surface dopants, Al has one more valence electron than Mg and donates charge to the adsorbed Au atom (Fig. 3d). Na reverses this trend, causing Au to donate electrons to balance the charge depletion in the surface (Fig. 3e). There is a clear similarity in the effects of surface dopants and adsorbates, and therefore we broadly classify surfaces with additional electrons compared to pristine MgO as electron-rich and surfaces with fewer electrons as electron-poor.

Metal adatom oxide formation enthalpy ($\Delta H_{f,ox,atom}$) is a widely used descriptor for predicting metal adsorption on different oxides[27,30,47,52]. However, we found that this descriptor was not sufficient for predicting changes in adsorption energy on the

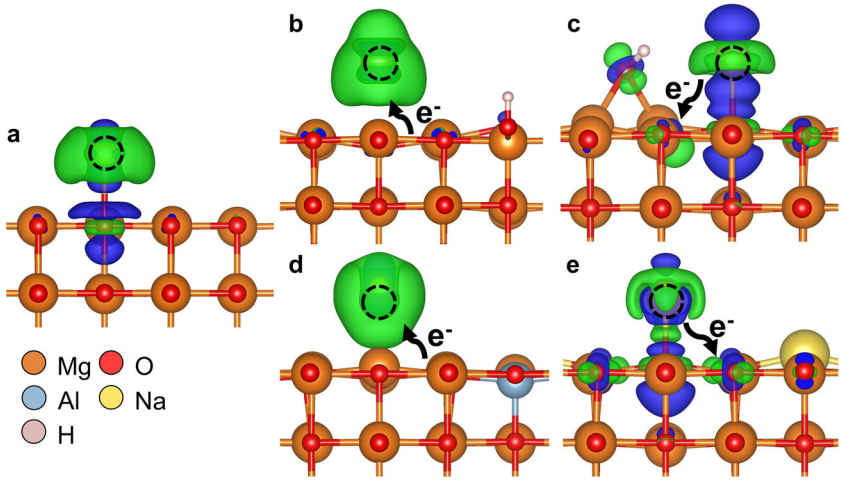

Fig. 3 Electron density distribution of Au on MgO surfaces. Isostructural charge density difference plots of **a** Au/MgO(100), **b** Au/H-MgO (100), **c** Au/OH-MgO(100), **d** Au/Al-MgO(100), and **e** Au/Na-MgO(100). Blue represents depletion of electron density and green represents accumulation. The black arrows depict the flow of electrons. The isosurface level is ±0.003 e^- $Bohr^{-3}$.

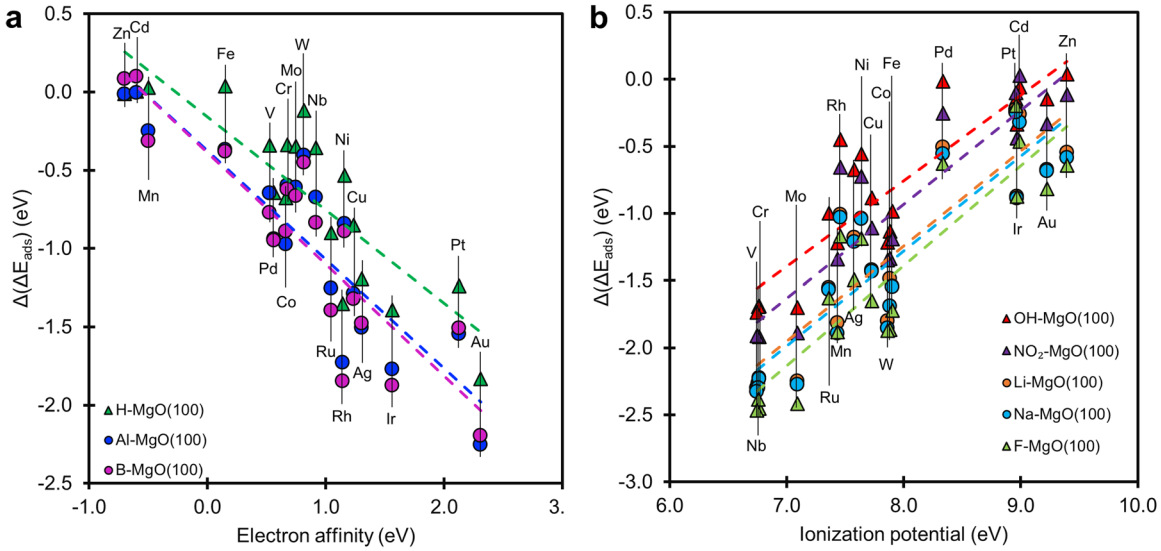

Fig. 4 Metal binding correlation to electron affinity and ionization potential. The **a** electron affinity (EA) and **b** ionization potential (IE_1) are used as descriptors for predicting the difference in metal adsorption energy between clean MgO versus electron-rich MgO and electron-poor MgO, respectively.

modified MgO(100) surfaces shown in Fig. 3, as it only captures how easily the metal can be oxidized by donating its electrons to the surface and fails to represent situations where electrons are donated to the metal (i.e., as seen in Fig. 3b, d for Au on electron-rich surfaces). The enhancement in metal adsorption energy in such situations is largely controlled by the ability of the adsorbed metal to accept excess charge from the modified surface, where $\Delta H_{\mathrm{f.ox,atom}}$ is no longer a suitable descriptor because it only captures the ability of the metal to donate electrons. Therefore, we propose that the ionization potential (i.e., IE_1) and EA can be used as broad descriptors for predicting the change in metal adsorption energy ($\Delta(\Delta E_{\mathrm{ads}})$, as defined by Eq. 2) caused by bidirectional charge transfer.

$$\Delta(\Delta E_{\mathrm{ads}}) = \Delta E_{\mathrm{ads}}(\mathrm{Modified\ MgO}) - \Delta E_{\mathrm{ads}}(\mathrm{Clean\ MgO}). \quad (2)$$

These descriptors capture the atom's propensity for both electron donation (versus IE_1) and acceptance (versus EA).

We plot the EA and IE_1 against $\Delta(\Delta E_{\mathrm{ads}})$ for electron-rich and electron-poor surfaces in Fig. 4. The correlation is separated into two categories: a negative correlation for electron-rich surfaces (Fig. 4a) and a positive correlation for electron-poor surfaces (Fig. 4b). On the electron-rich surfaces, Au adsorption has been enhanced significantly because Au has the highest EA and readily accepts the extra electronic charge. The adsorption of V is increased the most on the electron-poor surfaces because V has the lowest IE_1 and can readily donate charge to the surface. The direction of charge transfer in these systems is further confirmed by the charge density difference analyses for Au and Mn shown in Supplementary Fig. 2 and the corresponding Bader charge analysis in Supplementary Table 1. Moreover, the density of states (DOS) analysis for Au and Mn on doped MgO(100) are provided in Supplementary Fig. 3. The shift in the DOS at the Fermi level of the electron-poor MgO(100) surface clearly demonstrates a transfer of electrons to the surface from the adsorbed metal (and vice versa for the electron-rich surface).

SL for identifying physical descriptors

Although useful for predicting broad trends, EA and IE_1 are limited in their ability to quantitatively describe metal binding on

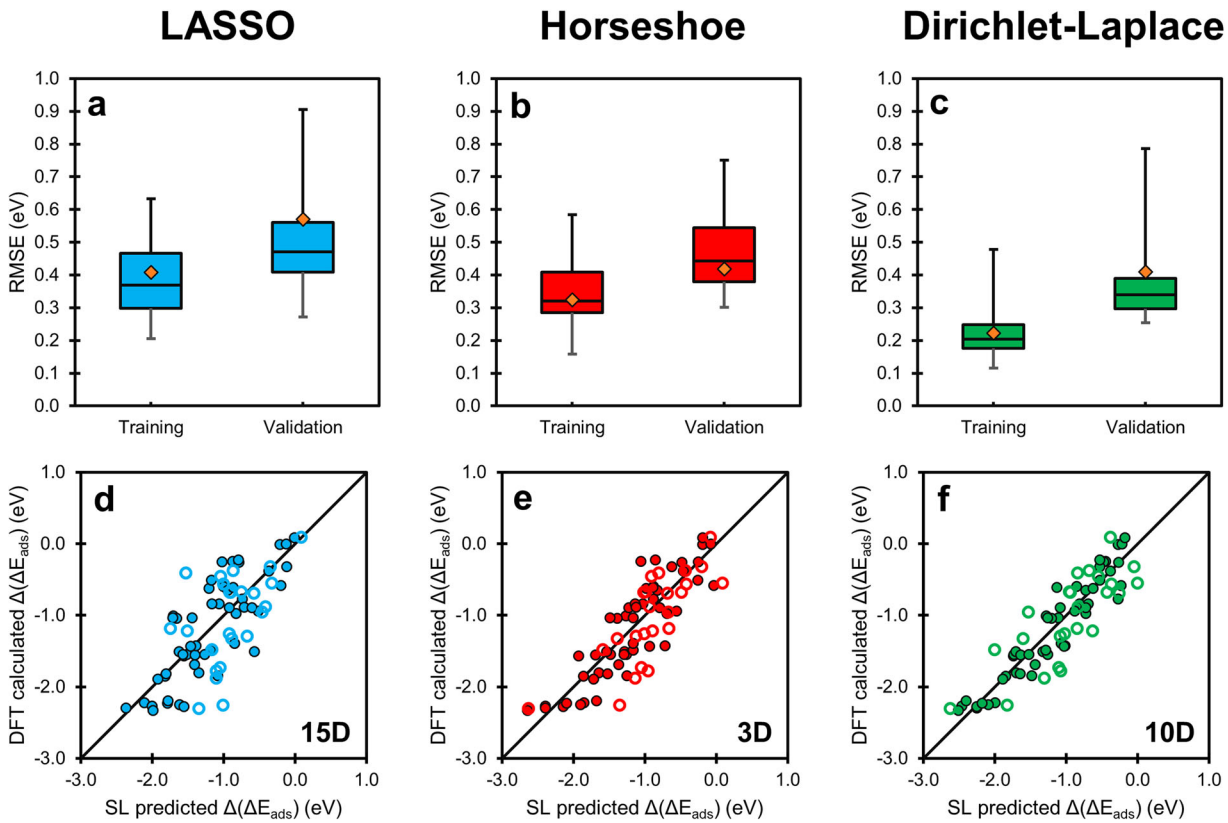

Fig. 5 Descriptor performance for doped MgO surfaces. Box plots represent the overall training and validation RMSE, computed using the l_0 norm method to refine the total descriptor set identified with each method to a model containing the number of descriptors that yields the lowest validation error, of the predictive models obtained from feature selection methods, i.e., **a** LASSO, **b** Horseshoe prior, and **c** Dirichlet–Laplace prior. Box plots reflect data from 50 trials of randomly split training and validation sets. Parity plots of models generated with **d** LASSO, **e** Horseshoe prior, and **f** Dirichlet–Laplace prior, where training and validation data are presented as solid and hollow points, respectively. The training and validation *RMSE* for **d**–**f** are indicated in **a**–**c** as orange diamonds. 15D, 3D, and 10D denote the number of descriptors in each model (i.e., the number of free coefficients used to fit the linear regression). The center line, upper bound, lower bound, upper whisker, and lower whisker in box plots represent median, 75th percentile, 25th percentile, maximum, and minimum, respectively.

modified MgO. This is evident in the scatter seen in Fig. 4. Quantitative descriptors that can unravel the effects of multiple charge transfer phenomena simultaneously will likely require more complex functional forms than what is captured by the simple electronic properties. In this section, we identify such descriptors for MgO using FS via LASSO, Horseshoe prior, and Dirichlet–Laplace prior. The feature spaces were also tested for transferability to CaO(100), BaO(100), and ZnO(100) in the next section, where DFT data for the additional oxide surfaces were not included in any training data used for feature selection. We find that Dirichlet–Laplace prior yields the best models, both in terms of lowest error and highest transferability.

We show the performance of the selected descriptor sets in Figs 5 and 6 using 50 trials of randomly separated training/validation data sets for dopant- and adsorbate-modified MgO, respectively. We use the l_0 norm method to refine each model by systematically decreasing the number of descriptors in the model (i.e., the model dimension corresponds to the number of descriptors in the model, *n*D). The root mean square error (RMSE) is calculated by Eq. 3 to quantify the prediction accuracy, where y_i is the ith predicted value, $\widehat{y_i}$ is the ith estimated value, and n is the number of predicted values.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}. \quad (3)$$

The model with trained coefficients is then applied to the validation set, and the dimension that provides the lowest validation RMSE is used to determine the final predictive model. RMSE for the final models derived from 50 randomly chosen training set points is shown in Fig. 5a–c and Fig. 6a–c for dopant- and adsorbate-modified MgO, respectively. Although LASSO tends to select the most descriptors, which will be shown in the "Methods" section, the derived models are less accurate than the ones built from the Bayesian methods. In addition, the total data set derived from Horseshoe prior contains less than five descriptors on average using either dopant-modified or adsorbate-modified MgO data, which is much smaller in size than the total data sets derived by the other two methods. Although Horseshoe prior constructs more accurate models than LASSO, it is highly aggressive in its selection and generates a very small pool of candidate features that are available for refining the model. Dirichlet–Laplace prior exhibits the best performance, as it is a balance between the stringent selection tendencies of Horseshoe prior and the permissive selection tendencies of LASSO. Although Horseshoe usually performs similarly to Dirichlet–Laplace (Supplementary Fig. 4) at a given model dimension, the model cannot be further improved because we rapidly exhaust the pool of candidate descriptors. Conversely, Dirichlet–Laplace prior can be systematically improved by adding more descriptors until the validation error of the test set increases, thus allowing us to balance model accuracy with model simplicity. The predictive performance of these models is shown by the parity plots in Figs 5d–f and 6d–f, where each plot was generated with a

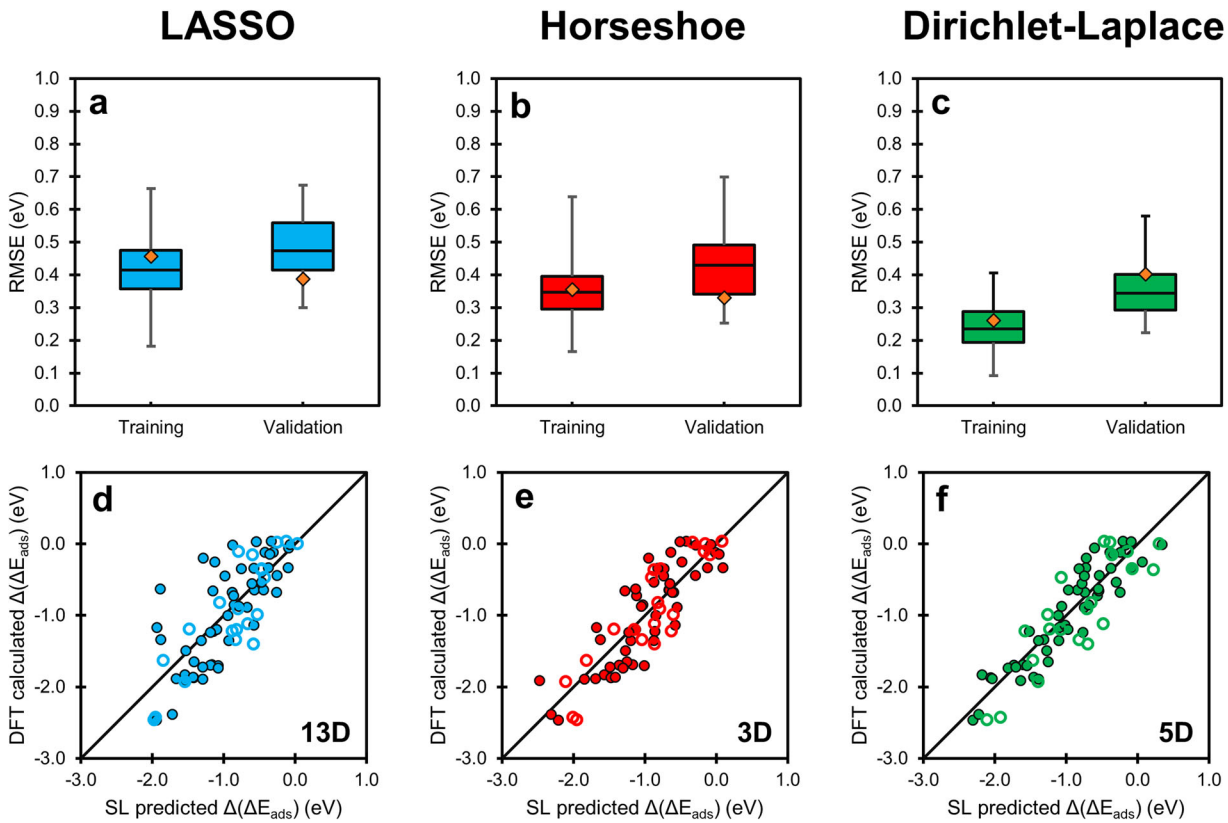

Fig. 6 Descriptor performance for adsorbate-modified MgO surfaces. Box plots represent the overall training and validation RMSE, computed using the l_0 norm method to refine the total descriptor set identified with each method to a model containing the number of descriptors that yields the lowest validation error, of the predictive models obtained from feature selection methods, i.e., **a** LASSO, **b** Horseshoe prior, and **c** Dirichlet–Laplace prior. Box plots reflect data from 50 trials of randomly split training and validation sets. Parity plots of models generated with **d** LASSO, **e** Horseshoe prior, and **f** Dirichlet–Laplace prior, where training and validation data are presented as solid and hollow points, respectively. The training and validation *RMSE* for **d**–**f** are indicated in **a**–**c** as orange diamonds. 13D, 3D, and 5D denote the number of descriptors in each model (i.e., the number of free coefficients used to fit the linear regression). The center line, upper bound, lower bound, upper whisker, and lower whisker in box plots represent median, 75th percentile, 25th percentile, maximum, and minimum, respectively.

representative training set that yields a training *RMSE* close to the median of each method (i.e., the same training set was used for all three methods to ensure comparability). The corresponding training and validation *RMSE* of these plots is labeled as an orange diamond in Figs 5a–c and 6a–c. The performance of the models is improved compared to our previous work on clean oxides using LASSO + l_0[52], where in that work we achieved an *RMSE* of 0.41 eV on the training set using five descriptors to predict 92 data points (i.e., compared to errors of 0.26 and 0.25 eV on training sets with five descriptors for the dopant and adsorbate data in Supplementary Fig. 4, respectively).

Tables 1 and 2 present the descriptors and relevant coefficients of the best models for dopant-modified and adsorbate-modified MgO surfaces that were shown in Figs 5f and 6f. The physical descriptors for doped MgO are composed of EN_P, EN_{MB}, IE_1, IE_2, EA, NVal, Z, and wr_p, and for adsorbate-modified MgO are composed of EN_P, EN_{MB}, IE_1, IE_2, and NVal. The formula of each descriptor is more complex than those identified in our previous study of clean oxides[52], reflecting the increased complexity of the problem once surface modifiers are introduced. The importance of charge transfer phenomena is evident in the character of the descriptors. For instance, $\left|\frac{IE_1^m - IE_2^s}{IE_2^m - IE_2^o}\right| \times \left|\frac{IE_1^m - IE_2^{dn}}{IE_2^m - IE_1^{dn}}\right|$ in Table 1 plays a major role in predicting the enhanced binding of Au adsorption on Al-MgO compared to Na-MgO. The appearance of a term capturing the difference between the ionization energies of the metal and the dopant suggests that the favorability of charge transfer between these two components plays a major role in determining the overall metal binding energy. For adsorbate-modified surfaces, $\left(\frac{EN_P^m - EN_P^s}{EN_P^s - EN_P^o}\right)^2 \times \left|\frac{IE_2^m - IE^a}{IE_1^o - IE^a}\right|$ is the most effective descriptor for Au adsorption. $\left|\frac{IE_2^m - IE^a}{IE_1^o - IE^a}\right|$ qualifies the difference in the ability of the parent metal, oxygen, and the adsorbate to donate charge. Although some terms in these models may be interpreted, this is in general a difficult task given the fact that interrelated phenomena controlling charge transfer cannot be described by a simple physical framework[71].

Robustness and transferability of the selected descriptors

The descriptors identified in the previous section were applied to predict $\Delta(\Delta E_{ads})$ on various surfaces to test feature transferability and robustness. The metals were placed on CaO(100), BaO(100), and ZnO(100) doped with the same dopants as used for MgO (Al, B, Li, and Na). We used the SL-derived descriptor sets from the MgO training data to predict the behavior on the abovementioned modified surfaces, where we use l_0 norm regression to refine the model dimensionality (Fig. 7). Note that this entails refitting the coefficients in front of each descriptor, so here we are only testing the transferability of the descriptors and not the transferability of the entire model. The performance of features derived from 50 trials of randomly separated MgO training sets are

Table 1. Descriptors, model coefficients, and responding values determined by Dirichlet–Laplace for dopant-modified MgO.

	Descriptors	Coefficients	Au/Na-MgO	Au/Al-MgO
1	$wr_p^m \times \left\|\frac{IE_2^m - IE_2^s}{IE_2^s - IE_1^d}\right\|$	11.99	0.98	0.30
2	$(NVal^m)^3 \times \left\|\frac{IE_2^{dn}}{IE_1^m - IE_2^s}\right\|$	0.00003	0.34	0.14
3	$\left(Z^m - Z^d\right)^2 \times \left\|\frac{IE_2^m - IE_2^s}{IE_1^o - IE_2^{dn}}\right\|$	−0.0002	−0.15	−0.89
4	$\left\|\frac{EN_p^m - EN_p^s}{EN_p^o - EN_p^d}\right\| \times \left\|\frac{IE_2^m - IE_1^d}{IE_1^s - IE_2^d}\right\|$	−1.69	−0.31	−0.69
5	$\sqrt{\left\|\frac{EN_{MB}^m - EN_{MB}^o}{EN_{MB}^s - EN_{MB}^o}\right\|} \times \left\|\frac{IE_1^s - IE_1^{dn}}{IE_1^m}\right\|$	−6.30	−1.76	−1.17
6	$\left\|\frac{IE_1^m - IE_2^s}{IE_2^m - IE_2^d}\right\| \times \left\|\frac{IE_1^m - IE_2^{dn}}{IE_2^m - IE_1^{dn}}\right\|$	−3.09	−0.88	−0.09
7	$\left\|\frac{IE_1^m - IE_2^d}{IE_2^s - IE_1^d}\right\| \times \left(NVal^m - NVal^d\right)$	0.04	0.83	0.47
8	$\left\|\frac{IE_2^d}{IE_2^m}\right\| \times \left\|\frac{EA^m}{EA^{dn}}\right\|$	−0.002	−0.03	−0.09
9	$\left\|\frac{IE_1^o - IE_1^d}{IE_2^m}\right\| \times \left\|\frac{IE_2^m - IE_2^s}{IE_1^{dn} - IE_2^{dn}}\right\|$	−0.16	−0.03	−0.34
10	$\left\|\frac{IE_2^o - IE_2^d}{IE_2^m}\right\| \times \left\|\frac{IE_2^m - IE_2^s}{IE_2^s - IE_1^{dn}}\right\|$	−0.10	−0.09	−0.33
11	Intercept	0.84	0.84	0.84
$\Delta(\Delta E_{ads})$	–	–	−0.26	−1.83

Table 2. Descriptors, model coefficients, and responding values determined by Dirichlet–Laplace for adsorbate-modified MgO.

	Descriptors	Coefficients	Au/OH-MgO	Au/H-MgO
1	$\sqrt{\left\|\frac{EN_p^s}{EN_p^m - EN_p^o}\right\|} \times \left(NVal^m - NVal^o\right)$	0.17	0.95	0.95
2	$\left\|\frac{EN_p^s - EN_p^a}{EN_p^m}\right\| \times \left\|\frac{IE_1^m - IE^a}{IE_2^m - IE_2^o}\right\|$	−0.99	−0.22	−0.11
3	$\left(\frac{EN_p^m - EN_p^s}{EN_p^s - EN_p^o}\right)^2 \times \left\|\frac{IE_2^m - IE^a}{IE_1^o - IE^a}\right\|$	−0.03	−0.09	−2.62
4	$\left(\frac{EN_{MB}^s}{EN_{MB}^m - EN_{MB}^o}\right)^2 \times \left\|\frac{IE_2^m - IE_2^s}{IE_1^m - IE_1^o}\right\|$	0.92	0.41	0.41
5	$\left\|\frac{IE_1^m - IE_2^o}{IE_2^m - IE_1^s}\right\| \times \left\|\frac{IE_2^m - IE_2^o}{IE_1^o - IE^a}\right\|$	0.0004	0.02	0.59
6	Intercept	−1.40	−1.4	−1.4
$\Delta(\Delta E_{ads})$	–	–	−0.34	−2.18

used to verify feature transferability, as shown by the box plots in Fig. 7. Since Horseshoe typically selects less than five descriptors, we restrict all of the predictive models to have at most five descriptors for comparison in Fig. 7a–c, as indicated by a dashed line in Fig. 7d–f. Comparing RMSE in Fig. 7a–c, the transferability of selected feature spaces is ranked as CaO ≈ BaO > ZnO. MgO and CaO have many similar properties and are commonly compared in the literature[45,46,72], and therefore it is expected that descriptors that are important for MgO should also be applicable to CaO. Since Ba is in the same group as Mg and Ca in the periodic table, it is intuitive that BaO data also correlates with the MgO-derived features. We find that the features are less transferable in the case of ZnO, as expected given the fact ZnO is in a different group. In the comparison between the SL methods, Horseshoe performs the worst. Although it can pick meaningful descriptors, it will only select the ones that are dominant for MgO and will aggressively throw out ones that, although important for other oxides, are less important for MgO. For a more detailed comparison, we summarize the difference in *RMSE* for LASSO and Dirichlet–Laplace (ΔRMSE) in Supplementary Fig. 5, which shows that LASSO has higher average prediction errors than Dirichlet–Laplace when comparing models derived with the same training set and the same model dimension. Thus, Dirichlet–Laplace shows the highest transferability among the SL methods. In summary, we find that descriptors identified using Dirichlet–Laplace prior for one oxide are applicable to other related oxides, demonstrating the robust nature of the selected features and their transferability outside the training set used for feature selection. We note that the RMSE for all of the testing oxides with the same feature space is generally similar to the MgO models.

Stability of modified MgO(100) surfaces

In the previous sections we evaluated the performance of various SL approaches for predicting the behavior of idealized MgO surfaces, where it is assumed that the dopants or adsorbates do not induce accompanying defects to compensate the additional charge and therefore charge transfer to the adsorbed metal atom was forced to occur. In real systems there inevitably will be a tendency for charge-compensating defects to compete for this charge transfer, which will be examined in this section. Numerous experiments demonstrate that the presence of the adsorbed metal can in certain cases suppresses the tendency to form charge-compensating defects in these irreducible oxides[73–75]. Shao et al.[44] showed that the charge transfer from Mo dopants in CaO enhances Au binding strength because excess electrons from Mo are transferred to the electronegative Au atoms. This effect was also present, but less pronounced, on Cr-doped MgO(100), where Stavale et al.[45] found that the enhancement of Au binding was inhibited by charge-compensating Mg vacancies. In addition to MgO, Tran et al.[76] induced Fe as dopants in ZnO to modulate the Pt/ZnO interaction. The X ray photoelectron spectroscopy analysis of Pt 4*f* core levels revealed the change in the oxidation state of Pt, which is direct evidence of the charge transfer between Fe and Pt. The enhanced metal–support interaction in this system stabilizes Pt nanoparticles and suppresses metal sintering during CO oxidation. For instance, the average Pt particle size increased from 2.2 to 5.7 nm on the bare ZnO during CO oxidation but only increased from 2.4 to 3.2 nm on the Fe-doped ZnO. Furthermore, the turnover frequency was raised from 0.60 to 5.37 s^{-1} with the assistance of the Fe dopants. Thus, the predictive model developed in this work is relevant for screening system compositions where one can expect significant charge transfer to induce an enhancement of the metal binding energy. Once identified, the stability of the required surface modification can be assessed, by computing free energies of formation as demonstrated below, to determine if the candidate system is a viable target for experimental synthesis attempts.

Here we compute phase diagrams for the formation of charge-compensating O or Mg vacancy defects in electron-poor (Na-doped and OH-modified) and electron-rich (Al-doped and H-modified) surfaces, respectively (Fig. 8). We expanded the simulation cell to twice the size of the original unit cell to accommodate multiple dopants in the surface (Supplementary Fig. 6). The O and Mg vacancy formation free energies are calculated by Eqs. 4 and 5, respectively, where $\Delta G_{O\text{-}vac}$ is the O vacancy formation free energy, $\Delta G_{Mg\text{-}vac}$ is the Mg vacancy formation free energy, $E_{O\text{-}vac}$ is the total DFT energy of the modified surface with one O vacancy, $E_{Mg\text{-}vac}$ is the total DFT energy of the modified surface with one Mg vacancy, E_{2MgO} is the

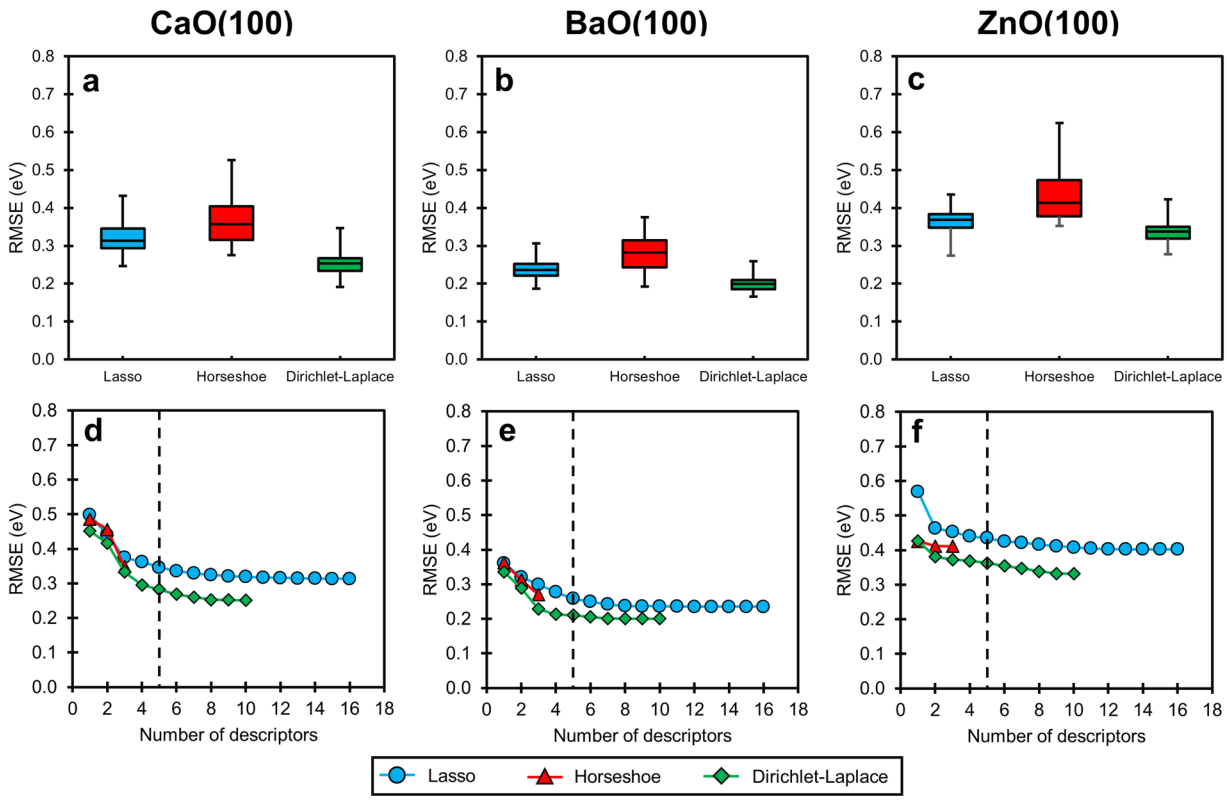

Fig. 7 Feature transferability. Box plots illustrate *RMSE* for **a** CaO, **b** BaO, and **c** ZnO applying the MgO-derived features. We note that RMSE is obtained from models with five descriptors even if the method selects more than five descriptors to have a fair comparison between the SL methods because Horseshoe prior usually selects less than five features (Fig. 10). Comparison of models for describing **d** CaO, **e** BaO, and **f** ZnO data built from features selected by LASSO, Horseshoe prior, and Dirichlet–Laplace prior using the MgO training data in Fig. 5. RMSE is computed using the l_0 norm method to refine the total descriptor set identified with each method to a model containing the number of descriptors shown on the *x*-axis. The center line, upper bound, lower bound, upper whisker, and lower whisker in box plots represent median, 75th percentile, 25th percentile, maximum, and minimum, respectively.

total DFT energy difference between the geometry with two MgO units on the surface (Supplementary Fig. 7) and the pristine MgO (100) surface, $E_{surface}$ is the total DFT energy of the modified surface without any charge-compensating defect, T is the temperature, and P_{O2} is the partial pressure of oxygen molecule. μ_{O2} is the chemical potential of an oxygen molecule in the gas phase, which is computed by the total DFT energy of the O_2 molecule with enthalpy and entropy corrections provided by the NIST webbook[77]. We used the energy of a small supported MgO cluster with two MgO formula units as the reference for Mg vacancy formation in Eq. 5 because the formation of this small cluster will be the first nucleation step once Mg atoms leave the MgO lattice. The energy of bulk MgO can also be used as the reference, which will systematically shift all Mg vacancy formation energies to be more negative by 2.59 eV. This will not impact the relative difference in vacancy formation energy when comparing the clean surfaces to the surfaces with adsorbed metal atoms.

$$\Delta G_{O-vac}(T,P_{O_2}) = E_{O-vac} + \frac{1}{2}\mu_{O_2}(T,P_{O_2}) - E_{surface}, \quad (4)$$

$$\Delta G_{Mg-vac}(T,P_{O_2}) = E_{Mg-vac} + \frac{1}{2}E_{2MgO} - \frac{1}{2}\mu_{O_2}(T,P_{O_2}) - E_{surface}. \quad (5)$$

We use the above equations to compute the boundary in (P_{O2}, T) space where the defect formation energy is zero, which is used to generate the phase diagrams shown in Fig. 8. In each case, we compute the defect formation boundary on the clean surface without any adsorbed metal, and then we compare it to the boundary on surfaces with the adsorbed metals. We find that the metal atom on modified MgO generally inhibits defect formation because the metal atom acts as an electron source or sink to alleviate the excess surface charge caused by the dopant or adsorbate. The extent of the boundary shift correlates with IE_1 and EA for electron-poor and electron-rich surfaces, respectively, which indicates that it is controlled by the ability of the adsorbed metal to donate or accept excess charge (Supplementary Fig. 8). Comparing vacancy formation energy of modified MgO(100) surfaces (the dashed lines in Supplementary Fig. 8) to the pristine surfaces (Supplementary Table 2), it is evident that the surface modification will lead to compensating defects, as discussed above. However, the favorability of compensating defect formation will be suppressed by the adsorbed metals. The suppression of the area of the phase diagram in which a compensating defect would occur, as seen in Fig. 8, suggests that it may thermodynamically be feasible to prepare some of the adsorbed metal/modified MgO surfaces considered in this work if the surface modification and metal adsorbate are introduced simultaneously before charge-compensating defects can form.

DISCUSSION

We have shown the capability of SL methods for constructing predictive models for computing $\Delta(\Delta E_{ads})$ in oxide-supported SAC systems. The SL methods yield useful relationships between the fundamental chemical properties of dopants/adsorbates and overall metal/support interactions, which are often used to control surface morphology but in many cases are not well-understood[31,36–38,41,44–46]. The resulting models provide an estimation tool for quantifying how metal binding energy can

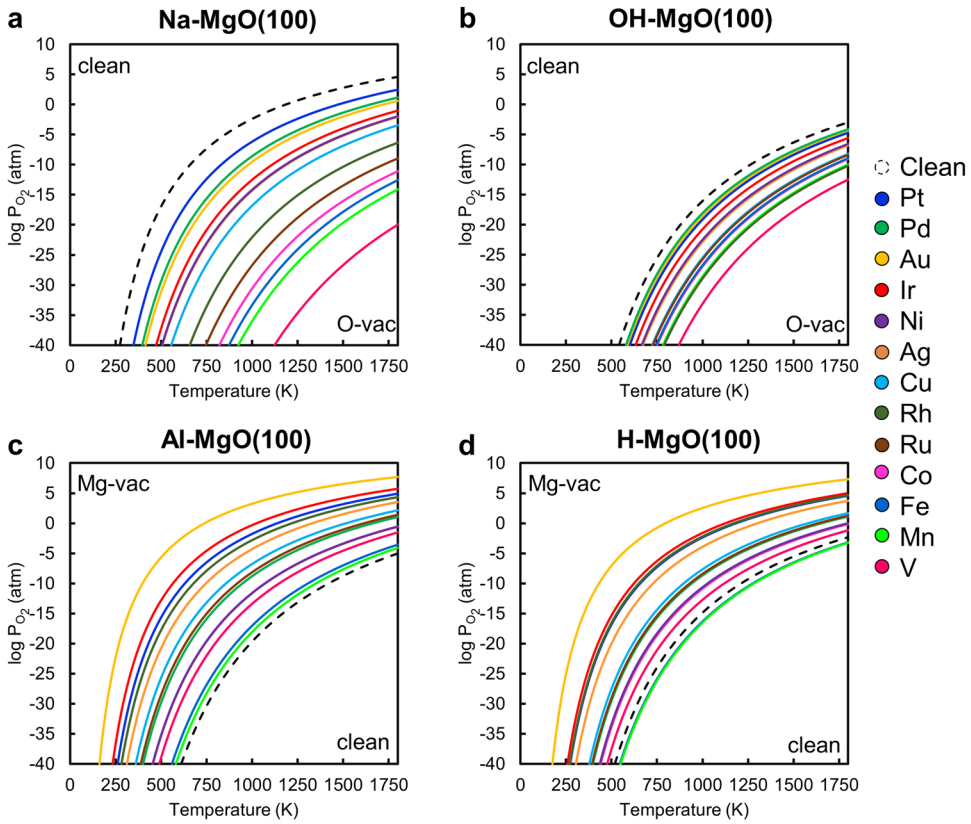

Fig. 8 Phase diagram of charge-compensating defect formation. The curve represents the boundary where the defect, e.g., O or Mg vacancies, formation free energy equals zero on modified MgO(100) surfaces as a function of temperature and oxygen partial pressure. For electron-poor surfaces, i.e., **a** Na-MgO and **b** OH-MgO, the O vacancy will form in the bottom-right region. For electron-rich surfaces, i.e., **c** Al-MgO and **d** H-MgO, the Mg vacancy will form in the top-left region.

be tuned through surface modification. This in turn can provide guidelines for designing and controlling the particle size of the TMs (e.g., by finding modifications that stabilize single adsorbed metal atoms relative to the bulk metal). Indeed, we predict that Mn may form stable SACs on Li-doped and Na-doped ZnO surfaces because the formation energy of single Mn atoms on these surfaces is negative relative to the formation of bulk Mn (Supplementary Section 1). We also introduced two state-of-the-art Bayesian methods that have several advantages over other common FS methods (e.g., PCA, KRR, LASSO, SISSO, etc.), such as flexible prior distributions, providing automatic error estimation on selected features, adaptivity to arbitrary sparsity in the feature space, and the ability to utilize the full data set for inference without requiring cross validation.

The robustness of our methodology was verified by applying the final predictive models to validation data sets, as well as by applying the selected features to CaO, BaO, and ZnO surfaces. The transferability of descriptors reveals the power of the Dirichlet–Laplace prior method, since it can use training data on one particular oxide to identify features that are applicable to similar oxides. We are currently exploring combined FS approaches based on training data derived from multiple oxides to build models that are fully general within oxide families. This will include extensions to more reducible oxides, where it is expected that the parent metal of the oxide will play a more direct role as a center of charge transfer. For example, our previous study[52] showed that metal–metal and redox interactions can occur on highly reducible oxides, such as CeO_2. The parent metals in the support are expected to interact with dopants and adsorbates as well, which will further complicate the nature of charge transfer. Finding suitable descriptors from chemical intuition alone will be nearly impossible for such systems, but likely will be feasible using the SL tools developed in this study.

Transition metal adsorption on MgO is enhanced by surface doping and adsorbate modification, which can be predicted using physical descriptors identified by SL (namely, LASSO, Horseshoe prior, and Dirichlet–Laplace prior methods). In particular, MgO can be transformed by dopants or adsorbates to exhibit either electron-rich or electron-poor characteristics. The extent of charge transfer between metal and modified MgO surfaces correlates with the ionization potential and the EA of the metal, but scatter in these correlations preclude quantitative prediction. Therefore, we used SL to derive predictive models that go beyond these simple descriptors, which were built from the fundamental chemical properties of the adsorbed metals, Mg, O, dopants, and adsorbates. Multiple FS tools, including state-of-the-art Bayesian methods, were applied together with l_0 norm regression to derive simple and effective models for predicting metal adsorption on modified MgO surfaces. Of the tested SL methods, we found that Dirichlet–Laplace prior yields the most accurate and transferable models. The descriptors identified for MgO(100) by Dirichlet–Laplace prior were also shown to be effective for estimating metal binding on CaO(100), BaO(100), and ZnO(100) surfaces with comparable accuracy, which demonstrates the robust nature of this FS approach and the transferability of our models to related oxides.

METHODS

DFT calculations

Metal adsorption energies were calculated with DFT using the Vienna ab initio simulation package (VASP 5.4.4)[78]. The Perdew–Burke–Ernzerhof exchange-correlation functional[79] was applied with spin polarization and

the projector augmented-wave method[80] was used to treat core electrons with VASP default potentials[81]. The valence electrons treated self-consistently for each atom type are listed in Supplementary Table 3. Planewave basis sets were expanded to a kinetic energy cutoff of 600 eV. Gaussian smearing was employed with a smearing width of 0.05 eV. A Monkhorst–Pack (MP)[82] *k*-point mesh was used with $4 \times 4 \times 4$ sampling on the bulk MgO structure and $3 \times 3 \times 1$ sampling on the (2×2) MgO(100) surface models. The Grimme empirical dispersion correction was used to treat van der Waals dispersion[83]. Geometries were optimized to a convergence criterion of 0.05 eV Å^{-1}. Metal binding energies computed with a force convergence criterion of 0.05 eV Å^{-1} were found to be nearly identical to those computed with a tighter criterion of 0.01 eV Å^{-1} (Supplementary Table 4).

Surface models contained four layers of MgO(100) with all layers relaxed to avoid spurious surface dipoles that can arise from frozen layers in oxide surfaces[84]. The vacuum distance between layers in the direction perpendicular to the surface was at least 15 Å to avoid interactions between MgO slabs, and a dipole correction[85] was applied perpendicular to the MgO(100) surface. Calculations on CaO(100), BaO(100), and ZnO (100) surfaces followed the same settings as those applied for the MgO (100) surfaces, except the MP *k*-points sampling is $5 \times 5 \times 1$ for ZnO(100). The ground state energy of single metal atoms is computed in a $15 \times 16 \times 17$ Å^3 unit cell, which was our reference for computing binding energies. The magnetization states and energies of the isolated metal atom are listed in Supplementary Table 5. Multiple magnetization states were tested when considering metal adsorption on the oxide surfaces, examining all probable spin configurations as listed in Supplementary Table 6, and the ground state with the lowest energy was selected for use in the adsorption energy training set. Metal binding energies on MgO(100), CaO(100), BaO (100), and ZnO(100) are listed in Supplementary Tables 7–10, and the corresponding system magnetizations are reported in Supplementary Tables 11–14. The Bader method[86,87] was used to calculate partial charges on atoms, which were only used for qualitatively assessing the direction of charge transfer (Supplementary Table 1). The defect formation energy of the pristine MgO(100) surface is listed in Supplementary Table 2. Coordinates of all energy-minimized structures are provided in the Supplementary Information.

Statistical learning

Our previous study[52] applied LASSO + l_0[51] to derive descriptors that can predict metal adsorption energies across a wide range of metal/oxide pairs. While successful in that work, in the present study we found that LASSO + l_0 was limited in its ability to handle correlated features when we introduced properties of the surface modifiers in the feature space (vide infra). As such, here we apply two additional state-of-the-art Bayesian methods for feature selection, Horseshoe prior[65] and Dirichlet–Laplace prior[23], that are implemented alongside the LASSO[68] approach for comparison. Bayesian methods are particularly well-suited for FS because they allow users to incorporate domain knowledge when applicable, they offer uncertainty quantification, and, most notably, they provide a coherent framework to infer model probabilities. Modern Bayesian FS methods can adapt to unknown sparsity levels in the feature space and perform well when handling correlated features[88]. Bayesian methods operate by first specifying an initial distribution of values for each descriptor coefficient (i.e., a prior distribution) and then updating this distribution based on available training data through Bayes' theorem (i.e., solving for the posterior distribution). These methods offer flexibility through the choice of the prior distribution, as different prior distributions will yield different FS characteristics. The DFT training data for FS in this work consisted of metal atom binding energies on MgO surfaces that were modified by either dopants or adsorbates, where FS was separate for "dopant" and "adsorbate" data sets. Details regarding the implementation of each FS method are provided in the Supplementary Information.

Primary descriptors. Our feature space was built from a primary descriptor set that contained chemical properties of the adsorbed metals, the parent atoms in the oxide surface (i.e., Ba, Ca, Mg, O, and Zn), and dopants that were available in the CRC Handbook of Chemistry and Physics[89]. These atomic properties include the atomic number (*Z*), electronegativity in Pauling and Martynov–Batsanov[90] scales (EN_P and EN_{MB}), first and second ionization energy (IE_1 and IE_2), EA, standard sublimation enthalpy (ΔH_{sub}), standard molar enthalpy of oxide formation of the metal adatoms ($\Delta H_{f,ox,bulk}$), standard molar enthalpy of formation of the adsorbed metal's most stable oxide referenced to the isolated metal atom ($\Delta H_{f,ox,atom}$), Zunger and Cohen orbital radii of *s* and *p* orbitals (zr_s and zr_p)[91], Waber and Cromer orbital radii of *s* and *p* orbitals (wr_s and wr_p)[92], number of valence electrons (NVal), Miedema metal alloy formation parameters ($\eta^{1/3}$and ϕ)[93], and absolute electronegativity (AEN)[94]. The following data were not available in CRC Handbook of Chemistry and Physics and instead were taken from the provided references: IE_2 of Ir[95], EA of Cd, Mg, Mn, and Zn[96], and $\Delta H_{f,ox,bulk}$ of Au and Pt[47]. Following our previous notation scheme[52], superscripts *m*, *s*, *o*, *d*, and *a* are used to indicate adsorbed metals, the parent metal in the oxide surface (which is Mg, Ca, Ba, or Zn in this study), oxygen, dopants, and adsorbates, respectively. The IE_1, IE_2, and *EA* of dopants include separate entries for the neutral (*n*) dopant (IE_1^{dn}, IE_2^{dn}, and EA^{dn}) and for the dopant in its most stable oxidation state (IE_1^{d}, IE_2^{d}, and EA^{d}). For adsorbate-modified systems, we also included EN_P, EN_{MB}, IE, EA, NVal, coordination number between adsorbates and the MgO(100) surface (CN), and bond dissociation energy (BD). We define BD as the binding energy between the atom in the adsorbate and the atom in the oxide surface to which the adsorbate binds (e.g., O–H for H*, Mg–O for OH* and NO_2*, and Mg–F for F*, where * indicates a species adsorbed on the surface). The EN of adsorbates is taken as the EN of the atom in the adsorbate that is attached to the support. We also considered polarization inside the adsorbed molecules by including electronegativity differences (ΔEN_P and ΔEN_{MB}) between the different atoms in multi-atom molecules (e.g., $|EN_N - EN_O|$ for NO_2 and $|EN_H - EN_O|$ for OH). All data described above are provided in the Supplementary Information.

Feature selection. We applied LASSO[68], Horseshoe prior[65], and Dirichlet–Laplace prior[23] methods, implemented in R version 3.6.0 on Linux[97] to identify correlations between the chemical properties of the system's components and the enhancement in metal binding energy that results from surface modification. There are a total of 76 and 73 calculated binding energies for dopant-modified and adsorbate-modified MgO(100), respectively. A total of 52 data points are randomly selected from this set to build a training set used to develop the predictive model, while the remaining data points are isolated from all SL procedures to use as a validation set. The feature engineering procedures used to generate the feature space of candidate descriptors are summarized in Fig. 9. First, we expanded our feature space of candidate descriptors by applying a series of mathematical operators (described in detail in the Supplementary Information) on the set of primary descriptors described in the previous section. This procedure introduces secondary descriptors that can capture nonlinear correlations between the fundamental properties of each component in the system and the metal binding energy. All categorical

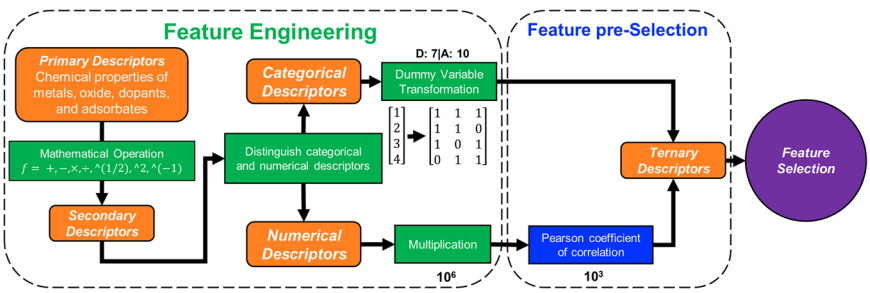

Fig. 9 Feature engineering and feature preselection procedures. Descriptor sets are indicated in orange boxes, feature engineering steps are indicated in green boxes, feature preselection steps are indicated in blue boxes, and the size of the feature space at each step is indicated in black text. There were seven and ten categorical descriptors for doped and adsorbate-modified MgO surfaces, respectively.

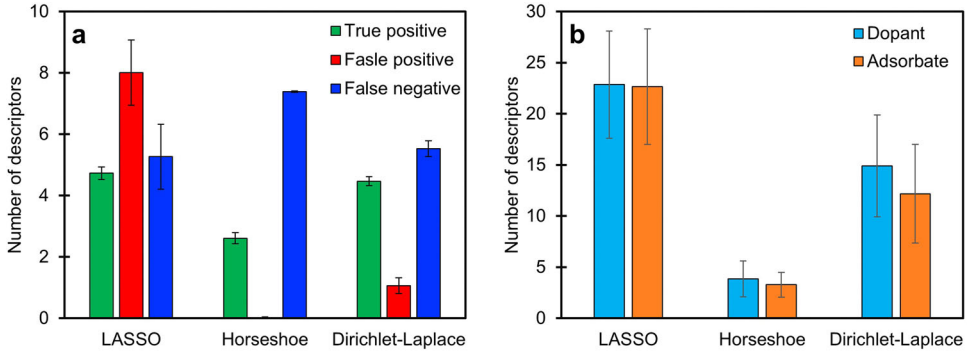

Fig. 10 Behavior and performance of feature selection methods. a Simulated average number of descriptors selected by each statistical learning method over 100 trials with 100 observations, 1000 candidate descriptors, and a correlation of $\rho = 0.5$ between the descriptors. **b** The average number of descriptors selected by each method using 50 randomly separated training set of metal binding energies on dopant-modified or adsorbate-modified MgO surfaces. The error bars represent the range of one standard deviation.

descriptors (i.e., descriptors for which all elements in the feature vector fall in a common category; see Supplementary Information for detailed explanation and definition of categorical descriptors) in the secondary descriptor set were converted into numerical descriptors using dummy variables so that they can be treated with linear methods[98]. The numerical descriptors from the secondary descriptor set were then cross-multiplied to mix different properties, which builds an engineered feature space containing ~10^6 descriptors. We prescreened the feature space by ranking the Pearson coefficient of correlation to identify the top 1,000 descriptors. We found that this data preprocessing step stabilizes the subsequent analysis, as well as improves calculation speed. We formed the final ternary feature space by adding the transformed categorical descriptors back into the secondary feature space, where all descriptor sets were normalized. All feature engineering procedures and the resulting feature space subsets are provided in the Supplementary Information. The algorithms of Dirichlet–Laplace prior and Horseshoe prior are summarized in Supplementary Tables 16 and 17, respectively.

The FS process yields a reduced feature space with an order of ~10 descriptors. To build the final model, we systematically test the performance of all combinations of features in this reduced space, where the size of the model (i.e., number of features in the combination) is predetermined, and select the combination that yields the lowest RMSE. We then systematically decrease the number of features in the final model until we reach a model with only one descriptor. This strategy, called l_0 norm regression, helps to systematically test the performance of the model with respect to the number of tunable parameters. Given the selected descriptors and the corresponding coefficients, we use the validation set to calculate the validation error to determine whether the predictive model is overfit, which occurs when the error of the validation set begins to rise as more descriptors are added to the model. We choose the final model size by selecting the model that yields the lowest error on the validation data set, thus ensuring that the model is not overfit. To test the FS processes, we applied the l_0 norm regression approach to evaluate feature spaces selected from MgO data on data derived from CaO, BaO, or ZnO modified with the same dopants (Al, B, Li, and Na). Each of these sets contained 52 data points for each oxide.

Evaluation of FS methods using synthetic data

It is crucial to understand the properties of each SL method by evaluating its performance characteristics on simulated data sets, where the ground truth is known a priori. This was done by populating a simulated data set with 100 observations and a feature space of 1,000 candidate descriptors, out of which ten descriptors truly correlate with the response. We repeated the simulation 100 times, each generating a random data set following the same model. We evaluated each method by reporting the average number of true positives (i.e., descriptors that truly correlate with the training set and are selected by the method), true negatives (i.e., descriptors that do not correlate with the training set and are not selected by the method), false positives (i.e., descriptors that do not correlate with the training set but are selected by the method), and false negatives (i.e., descriptors that correlate with the training set but are not selected by the method). This gives us insight into the characteristics of each SL approach.

Figure 10a shows the simulation result for a candidate feature space that has an average feature correlation of $\rho = 0.5$ (i.e., a test case with a high rate of correlation among the candidate features). As seen in the figure, LASSO suffers from a high rate of false positives when candidate features are highly correlated. This issue is largely mitigated by the Bayesian methods, which motivates their use in this work. Although Horseshoe prior shows a low rate of false positive selection, it also misses many true descriptors evident from its high rate of false negatives. Simulation tests demonstrate that the high rate of false positives in LASSO and false negatives in Horseshoe is consistent across various correlation parameters (see Supplementary Information for simulations with $\rho = 0$, $\rho = 0.5$, and $\rho = 0.9$ in Supplementary Tables 18–20), which suggests that this behavior will also be present given the expected range of correlation in our feature space. This is verified here by showing the number of features selected by each method on our MgO data set (Fig. 10b). Horseshoe selects the fewest descriptors in both dopant-modified and adsorbate-modified data sets, which suggests that Horseshoe is the most aggressive FS methods. We note that LASSO does not always select the most descriptors, but its performance is mostly worse than Dirichlet–Laplace. The feature spaces selected based on the "real" MgO data are analyzed in the "Results" section.

DATA AVAILABILITY

Coordinates of all energy-minimized DFT structures and all SL training data are provided in the following online repository: https://github.com/tsenftle/MgO_SL/.

CODE AVAILABILITY

All SL codes are provided in the following online repository: https://github.com/tsenftle/MgO_SL/Scripts.

Received: 25 October 2019; Accepted: 2 July 2020;
Published online: 21 July 2020

REFERENCES

1. Liu, L. & Corma, A. Metal catalysts for heterogeneous catalysis: from single atoms to nanoclusters and nanoparticles. *Chem. Rev.* **118**, 4981–5079 (2018).
2. Qiao, B. et al. Highly efficient catalysis of preferential oxidation of CO in H_2-rich stream by gold single-atom catalysts. *ACS Catal.* **5**, 6249–6254 (2015).
3. Qiao, B. et al. Single-atom catalysis of CO oxidation using Pt_1/FeO_x. *Nat. Chem.* **3**, 634–641 (2011).
4. Moses-DeBusk, M. et al. CO oxidation on supported single Pt atoms: experimental and ab initio density functional studies of CO interaction with Pt atom on θ-Al_2O_3(010) surface. *J. Am. Chem. Soc.* **135**, 12634–12645 (2013).
5. DeRita, L. et al. Catalyst architecture for stable single atom dispersion enables site-specific spectroscopic and reactivity measurements of CO adsorbed to Pt atoms, oxidized Pt clusters, and metallic Pt clusters on TiO_2. *J. Am. Chem. Soc.* **139**, 14150–14165 (2017).
6. Jones, J. et al. Thermally stable single-atom platinum-on-ceria catalysts via atom trapping. *Science* **353**, 150–154 (2016).
7. Zhang, Z. et al. Thermally stable single atom Pt/m-Al_2O_3 for selective hydrogenation and CO oxidation. *Nat. Commun.* **8**, 16100 (2017).

8. Abbet, S., Heiz, U., Häkkinen, H. & Landman, U. CO oxidation on a single Pd atom supported on magnesia. *Phys. Rev. Lett.* **86**, 5950–5953 (2001).
9. Liang, J.-X. et al. Theoretical and experimental investigations on single-atom catalysis: Ir_1/FeO_x for CO oxidation. *J. Phys. Chem. C.* **118**, 21945–21951 (2014).
10. Spezzati, G. et al. Atomically dispersed Pd–O species on CeO_2(111) as highly active sites for low-temperature CO oxidation. *ACS Catal.* **7**, 6887–6891 (2017).
11. Yang, M. et al. A common single-site Pt(II)–$O(OH)_x$—species stabilized by sodium on "active" and "inert" supports catalyzes the water-gas shift reaction. *J. Am. Chem. Soc.* **137**, 3470–3473 (2015).
12. Lin, J. et al. Remarkable performance of Ir_1/FeO_x single-atom catalyst in water gas shift reaction. *J. Am. Chem. Soc.* **135**, 15314–15317 (2013).
13. Yang, M. et al. Catalytically active Au-$O(OH)_x$—species stabilized by alkali ions on zeolites and mesoporous oxides. *Science* **346**, 1498–1501 (2014).
14. Wei, H. et al. FeO_x-supported platinum single-atom and pseudo-single-atom catalysts for chemoselective hydrogenation of functionalized nitroarenes. *Nat. Commun.* **5**, 5634 (2014).
15. Kwak, J. H., Kovarik, L. & Szanyi, J. CO_2 reduction on supported Ru/Al_2O_3 catalysts: cluster size dependence of product selectivity. *ACS Catal.* **3**, 2449–2455 (2013).
16. Liu, P. et al. Photochemical route for synthesizing atomically dispersed palladium catalysts. *Science* **352**, 797–800 (2016).
17. Guzman, J. & Gates, B. C. Structure and reactivity of a mononuclear gold-complex catalyst supported on magnesium oxide. *Angew. Chem. Int. Ed.* **115**, 714–717 (2003).
18. Wang, C. et al. Low-temperature dehydrogenation of ethanol on atomically dispersed gold supported on $ZnZrO_x$. *ACS Catal.* **6**, 210–218 (2016).
19. Guo, X. et al. Direct, Nonoxidative conversion of methane to ethylene, aromatics, and hydrogen. *Science* **344**, 616–619 (2014).
20. Gu, X.-K. et al. Supported single Pt_1/Au_1 atoms for methanol steam reforming. *ACS Catal.* **4**, 3886–3890 (2014).
21. Hu, B. et al. Isolated Fe^{II} on silica as a selective propane dehydrogenation catalyst. *ACS Catal.* **5**, 3494–3503 (2015).
22. Li, Y. H., Xing, J., Yang, X. H. & Yang, H. G. Cluster size effects of platinum oxide as active sites in hydrogen evolution reactions. *Chem. Eur. J.* **20**, 12377–12380 (2014).
23. Bhattacharya, A., Pati, D., Pillai, N. S. & Dunson, D. B. Dirichlet–Laplace priors for optimal shrinkage. *J. Am. Stat. Assoc.* **110**, 1479–1490 (2015).
24. Tauster, S. J., Fung, S. C. & Garten, R. L. Strong metal-support interactions. Group 8 noble metals supported on titanium dioxide. *J. Am. Chem. Soc.* **100**, 170–175 (1978).
25. Chandler, B. D. An extra layer of complexity: strong metal-support interactions. *Nat. Chem.* **9**, 108–109 (2017).
26. Dai, Y., Lu, P., Cao, Z., Campbell, C. T. & Xia, Y. The physical chemistry and materials science behind sinter-resistant catalysts. *Chem. Soc. Rev.* **47**, 4314–4331 (2018).
27. Hemmingson, S. L. & Campbell, C. T. Trends in adhesion energies of metal nanoparticles on oxide surfaces: understanding support effects in catalysis and nanotechnology. *ACS Nano* **11**, 1196–1203 (2017).
28. Campbell, C. T. & Mao, Z. Chemical potential of metal atoms in supported nanoparticles: dependence upon particle size and support. *ACS Catal.* **7**, 8460–8466 (2017).
29. Campbell, C. T. The energetics of supported metal nanoparticles: relationships to sintering rates and catalytic activity. *Acc. Chem. Res.* **46**, 1712–1719 (2013).
30. Strayer, M. E. et al. Charge transfer stabilization of late transition metal oxide nanoparticles on a layered niobate support. *J. Am. Chem. Soc.* **137**, 16216–16224 (2015).
31. Chen, G. et al. Interfacial electronic effects control the reaction selectivity of platinum catalysts. *Nat. Mater.* **15**, 564–569 (2016).
32. Wang, Y.-G., Yoon, Y., Glezakou, V.-A., Li, J. & Rousseau, R. The role of reducible oxide–metal cluster charge transfer in catalytic processes: new insights on the catalytic mechanism of CO oxidation on Au/TiO_2 from ab initio molecular dynamics. *J. Am. Chem. Soc.* **135**, 10673–10683 (2013).
33. Matsubu, J. C. et al. Adsorbate-mediated strong metal–support interactions in oxide-supported Rh catalysts. *Nat. Chem.* **9**, 120–127 (2017).
34. Hu, P. et al. Electronic metal–support interactions in single-atom catalysts. *Angew. Chem. Int. Ed.* **53**, 3418–3421 (2014).
35. Campbell, C. T. Catalyst–support interactions: electronic perturbations. *Nat. Chem.* **4**, 597–598 (2012).
36. Pacchioni, G. Electronic interactions and charge transfers of metal atoms and clusters on oxide surfaces. *Phys. Chem. Chem. Phys.* **15**, 1737 (2013).
37. Schlexer, P., Puigdollers, A. R. & Pacchioni, G. Tuning the charge state of Ag and Au atoms and clusters deposited on oxide surfaces by doping: a DFT study of the adsorption properties of nitrogen- and niobium-doped TiO_2 and ZrO_2. *Phys. Chem. Chem. Phys.* **17**, 22342–22360 (2015).
38. Hu, C. H. et al. Modulation of catalyst particle structure upon support hydroxylation: *ab initio* insights into Pd_{13} and Pt_{13}/γ-Al_2O_3. *J. Catal.* **274**, 99–110 (2010).
39. Ghosh, S., Mammen, N. & Narasimhan, S. Descriptor for the efficacy of aliovalent doping of oxides and its application for the charging of supported Au clusters. *J. Phys. Chem. C.* **123**, 19794–19805 (2019).
40. Rahmani Didar, B. & Balbuena, P. B. Reactivity of Cu and Co nanoparticles supported on Mo-doped MgO. *Ind. Eng. Chem. Res.* **58**, 18213–18222 (2019).
41. Addou, R. et al. Influence of hydroxyls on Pd atom mobility and clustering on rutile TiO_2(011)-2×1. *ACS Nano* **8**, 6321–6333 (2014).
42. Babucci, M. et al. Controlling catalytic activity and selectivity for partial hydrogenation by tuning the environment around active sites in iridium complexes bonded to supports. *Chem. Sci.* **10**, 2623–2632 (2019).
43. Kumar, G. et al. Evaluating differences in the active-site electronics of supported Au nanoparticle catalysts using Hammett and DFT studies. *Nat. Chem.* **10**, 268–274 (2018).
44. Shao, X. et al. Tailoring the shape of metal Ad-particles by doping the oxide support. *Angew. Chem. Int. Ed.* **50**, 11525–11527 (2011).
45. Stavale, F. et al. Donor characteristics of transition-metal-doped oxides: Cr-doped MgO versus Mo-doped CaO. *J. Am. Chem. Soc.* **134**, 11380–11383 (2012).
46. Prada, S., Giordano, L. & Pacchioni, G. Charging of gold atoms on doped MgO and CaO: identifying the key parameters by DFT calculations. *J. Phys. Chem. C.* **117**, 9943–9951 (2013).
47. Campbell, C. T. & Sellers, J. R. V. Anchored metal nanoparticles: effects of support and size on their energy, sintering resistance and reactivity. *Faraday Discuss.* **162**, 9–30 (2013).
48. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91**, 135503 (2003).
49. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
50. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
51. Ghiringhelli, L. M. et al. Learning physical descriptors for materials science by compressed sensing. *N. J. Phys.* **19**, 023017 (2017).
52. O'Connor, N. J., Jonayat, A. S. M., Janik, M. J. & Senftle, T. P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **1**, 531–539 (2018).
53. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: a compressed-sensing method for systematically identifying efficient physical models of materials properties. https://arxiv.org/abs/1710.03319 (2017).
54. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
55. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
56. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
57. Raftery, A. E. Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
58. Casella, G. & Moreno, E. Objective Bayesian variable selection. *J. Am. Stat. Assoc.* **101**, 157–167 (2006).
59. Park, T. & Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
60. Claeskens, G. & Hjort, N. L. *Model Selection and Model Averaging* (Cambridge University Press, 2008).
61. Castillo, I., Schmidt-Hieber, J. & van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Stat.* **43**, 1986–2018 (2015).
62. Zhang, Y. & Bondell, H. D. Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors. *Bayesian Anal.* **13**, 823–844 (2018).
63. Scott, J. G. & Berger, J. O. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Stat.* **38**, 2587–2619 (2010).
64. Li, M. & Dunson, D. B. Comparing and weighting imperfect models using D-probabilities. *J. Am. Stat. Assoc.* 1–26, https://doi.org/10.1080/01621459.2019.1611140 (2019).
65. Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480 (2010).
66. Brown, M. A. et al. Oxidation of Au by surface OH: nucleation and electronic structure of gold on hydroxylated MgO(001). *J. Am. Chem. Soc.* **133**, 10668–10676 (2011).
67. Choksi, T., Majumdar, P. & Greeley, J. P. Electrostatic origins of linear scaling relationships at bifunctional metal/oxide interfaces: a case study of Au nanoparticles on doped MgO substrates. *Angew. Chem. Int. Ed.* **57**, 1–6 (2018).
68. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
69. Yudanov, I., Pacchioni, G., Neyman, K. & Rösch, N. Systematic density functional study of the adsorption of transition metal atoms on the MgO(001) surface. *J. Phys. Chem. B* **101**, 2786–2792 (1997).

70. Risse, T., Shaikhutdinov, S., Nilius, N., Sterrer, M. & Freund, H.-J. Gold supported on thin oxide films: from single atoms to nanoparticles. *Acc. Chem. Res.* **41**, 949–956 (2008).
71. Lipton, Z. C. The mythos of model interpretability. https://arxiv.org/abs/1606.03490 (2016).
72. Cui, Y., Stiehler, C., Nilius, N. & Freund, H.-J. Probing the electronic properties and charge state of gold nanoparticles on ultrathin MgO versus thick doped CaO films. *Phys. Rev. B* **92**, 075444 (2015).
73. Lin, X. et al. Charge-mediated adsorption behavior of CO on MgO-supported Au clusters. *J. Am. Chem. Soc.* **132**, 7745–7749 (2010).
74. Pacchioni, G. & Freund, H. Electron transfer at oxide surfaces. The MgO paradigm: from defects to ultrathin films. *Chem. Rev.* **113**, 4035–4072 (2013).
75. Pacchioni, G. & Freund, H.-J. Controlling the charge state of supported nanoparticles in catalysis: lessons from model systems. *Chem. Soc. Rev.* **47**, 8474–8502 (2018).
76. Tran, S. B. T., Choi, H. S., Oh, S. Y., Moon, S. Y. & Park, J. Y. Iron-doped ZnO as a support for Pt-based catalysts to improve activity and stability: enhancement of metal–support interaction by the doping effect. *RSC Adv.* **8**, 21528–21533 (2018).
77. Linstrom, P. J. & Mallard, W. G. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (National Institute of Standards and Technology, 2020).
78. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
79. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
80. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
81. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
82. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
83. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).
84. Hinnemann, B. & Carter, E. A. Adsorption of Al, O, Hf, Y, Pt, and S atoms on α-Al_2O_3(0001). *J. Phys. Chem. C.* **111**, 7105–7126 (2007).
85. Neugebauer, J. & Scheffler, M. Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al(111). *Phys. Rev. B* **46**, 16067–16080 (1992).
86. Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **18**, 7 (1985).
87. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Comput. Mater. Sci.* **36**, 354–360 (2006).
88. Ishwaran, H. & Rao, J. S. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* **33**, 730–773 (2005).
89. Rumble, J. R. *CRC Handbook of Chemistry and Physics*, 99th (Internet Version 2018) (CRC Press/Taylor & Francis, Boca Raton, FL.).
90. Villars, P. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *J. Less Common Met.* **92**, 215–238 (1983).
91. Zunger, A. Systematization of the stable crystal structure of all AB -type binary compounds: a pseudopotential orbital-radii approach. *Phys. Rev. B* **22**, 5839–5872 (1980).
92. Waber, J. T. & Cromer, D. T. Orbital radii of atoms and ions. *J. Chem. Phys.* **42**, 4116–4123 (1965).
93. Miedema, A. R., de Châtel, P. F. & de Boer, F. R. Cohesion in alloys—fundamentals of a semi-empirical model. *Phys. B* **100**, 1–28 (1980).
94. Pearson, R. G. Absolute electronegativity and absolute hardness of Lewis acids and bases. *J. Am. Chem. Soc.* **107**, 6801–6806 (1985).
95. Finkelnburg, W. & Humbach, W. Ionisierungsenergien von Atomen und Atomionen. *Naturwissenschaften* **42**, 35–37 (1955).
96. Bratsch, S. G. & Lagowski, J. J. Predicted stabilities of monatomic anions in water and liquid ammonia at 298.15 K. *Polyhedron* **5**, 1763–1770 (1986).
97. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
98. Suits, D. B. Use of dummy variables in regression equations. *J. Am. Stat. Assoc.* **52**, 548–551 (1957).

ACKNOWLEDGEMENTS

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing (HPC) resources that have contributed to the research results reported within this paper. C.-Y.L. and T.P.S. would like to acknowledge startup funding provided by Rice University.

AUTHOR CONTRIBUTIONS

T.P.S. developed the concept of this project and supervised it together with M.L. C.-Y. L. and D.M. computed the DFT energies, conducted the electronic analysis, and engineered the primary features. S.Z. implemented the Horseshoe prior and Dirichlet–Laplace prior in the scripts written in *R* for feature selection. All authors discussed and modified the paper together. C.-Y.L. and S.Z. contributed equally in this work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00371-x.

Correspondence and requests for materials should be addressed to M.L. or T.P.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.