

Uncovering electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning

Jacques A. Esterhuizen,^{1,2} Bryan R. Goldsmith,^{1,2*} and Suljo Linic^{1,2,3*}

Summary

We show that unsupervised machine learning (ML) using principal component analysis (PCA) provides a straightforward pathway for developing accurate and interpretable electronic-structure descriptors of the chemical and catalytic properties of materials. We demonstrate the approach by finding chemisorption descriptors for metal alloys and surface oxygens on metals and metal oxides. In both cases, the principal component (PC) descriptors yield ML models that predict the material's chemical properties with competitive accuracy compared to ML models built using established descriptors. Importantly, interpreting the electronic-structure patterns captured by each PC descriptor via signal reconstruction suggests potential design motifs for future electronic-structure descriptor design and allows us to identify links between a material's geometric and catalytic properties. Ultimately, we show that the unsupervised ML approach provides a route to find electronic-structure descriptors of the catalytic properties of materials that readily connect to geometric structure and composition.

Introduction

In the field of heterogeneous catalysis, electronic-structure descriptors serve as an invaluable link between the geometric structure of catalysts and their chemisorption properties.¹⁻⁷ One of the most

¹ Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136, USA

² Catalysis Science and Technology Institute, University of Michigan, Ann Arbor, Michigan 48109-2136, USA

³ Lead contact

* Correspondence: bgoldsm@umich.edu, linic@umich.edu

widely used descriptors of the chemisorption energy of atoms or molecules on a transition metal surface site is the site's d -band center. The d -band center, calculated as the first statistical moment of the d -projected density of states (DOS) at the adsorption site, describes the average energy of a surface site's d -electronic states.^{3,4,8-10} One approach to improve the predictive capacity of electronic-structure chemisorption descriptors beyond the d -band center is to consider higher-order moments of the projected DOS (e.g., the second-order moment which describes the d -band width⁵), or even other fine-structure descriptors (i.e., local changes to the maxima, minima, and tails of the distribution not captured by low-degree statistical momenta) of the DOS such as the position of the upper d -band edge relative to the Fermi level.^{6,11} While these targeted descriptor development efforts have led to augmented descriptors for some materials, there exists no unified framework to identify these descriptors across a range of different materials systematically.

In this contribution, we present a data-driven workflow that allows us to identify accurate and interpretable electronic-structure-based chemisorption descriptors. We use these electronic-structure descriptors as a bridge to relate a catalyst's chemisorption properties to its geometric structure and composition. To accomplish this, we employ principal component analysis (PCA) to derive descriptors of a material's electronic structure (**Schematic 1**). We show that these principal component (PC) descriptors are minimal and robust features for building accurate chemisorption models using supervised ML algorithms. We emphasize that there are more efficient approaches for simply predicting chemisorption energies than using electronic structure descriptors, such as graph convolutional neural networks trained on large adsorption datasets like OC20.^{12,13} However, the utility of electronic-structure descriptor approaches lie in their ability to yield scientific insights in an interpretable fashion. We interpret the electronic-structure effects captured by each of the PC descriptors using signal reconstruction and show that these effects map to local changes in the

geometric structure of a site (**Schematic 1b–c**). We demonstrate the approach by finding PC descriptors for chemisorption on transition-metal alloy surfaces and validate it by comparing our findings to the results of physics-based chemisorption descriptors. We extend the approach to study surface oxygen reactivity for metals, rutile metal-oxides, and perovskite metal-oxides, uncovering new electronic-structure descriptor motifs for quantifying oxygen reactivity for these materials. We expect that the approach will generally be applicable to identify electronic-structure descriptors for a range of problems in the physical and chemical sciences (e.g., solid-state materials,^{14–16} organometallic catalysts,¹⁷ and enzymes¹⁸).

Results and discussion

Modeling Chemisorption with Density Functional Theory

We chose a layered alloy model system with well-defined ligand and strain effects to elucidate the interplay between electronic-structure and geometric effects on chemisorption energies.^{3,9,10,19} An overview of the surface and ligand metals considered for the layered alloy (111) model systems is shown in **Figure 1a**. In the layered alloys, the ligand metal composes the layer immediately beneath the surface, and the surface metal composes the rest of the slab (**Figure 1b**). A total of 245 layered alloys were considered (in all cases, strain = -2%, -1%, 0%, 1%, 2%): 55 Rh alloys (ligand metal = Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Ir, Pt, and Au), 60 Pd alloys (Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Os, Ir, Pt, and Au), 45 Ir alloys (Ni, Cu, Rh, Pd, Ag, Os, Ir, Pt, and Au), and 75 Pt alloys (Fe, Co, Ni, Cu, Mo, Ru, Rh, Pd, Ag, W, Re, Os, Ir, Pt, and Au). The considered surface metals are commonly used in catalytic applications and span different groups and periods on the periodic table. Ligand atoms for each surface metal were selected based on having favorable alloy formation character according to the Hume-Rothery rules and previous reports of their synthesis.^{20,21} Strained layered alloys were considered because it is well-established that strain

engineering can mediate the chemisorption strength through electronic-structure effects.^{4,8} We emphasize that the alloy-induced ligand and strain effects present in these layered alloy model systems are also present in intermetallic and disordered alloys.

We modeled chemisorption of C, O, N, and H atomic adsorbates on the layered alloys using Kohn-Sham density functional theory (DFT) with the PBE functional. These four adsorbates were selected because their chemisorption energies have been used as descriptors for numerous chemical transformations, including water oxidation,²² hydrogen evolution,²³ nitrogen dissociation,²⁴ and hydrocarbon transformations.^{25,26} Adsorbates were placed at the same adsorption sites on the strained and unstrained layered alloys as on the pure surface metal. The considered adsorption sites have been previously reported as the most stable sites for the studied adsorbates and surface metals,²⁷⁻³⁰ and are shown in **Figure 1c**. The adsorption of C atoms was modeled at the hcp hollow site on all surfaces except for Pt, where it was modeled at the fcc hollow site. The O species was placed at the fcc hollow site on all alloys. The N atom was placed at the fcc hollow site on all surfaces, except for Rh, where it was placed at the hcp hollow site. H was placed at the fcc hollow site on Rh and Pd surfaces and the atop site on Ir and Pt. The **Methods** section contains additional modeling details regarding the calculation parameters and model systems used in this study.

Comparative Analysis of Principal Component Descriptor Performance

Principal component analysis (PCA) is an unsupervised machine learning technique that reduces a dataset's dimensionality (i.e., the number of variables describing the dataset) by projecting the data onto a reduced orthogonal basis, called the principal components (PCs), that describes the maximum variance within the data.³¹ Projecting the complete data onto the vectors that make up this orthogonal basis yields a dataset of reduced dimensionality such that all data

points are described by a set of values called PC scores rather than the original complete feature space.³² The resulting PC descriptors are adsorbate independent electronic-structure descriptors of the alloy. In contrast to previous ML studies that applied PCA to atomic, geometric, or energetic features,^{33–39} we apply PCA to the atom projected *d*-electronic DOS at the bare catalyst surface. For example, García-Muelas and López used PCA to find adsorption descriptors of C₁-C₂ species on various close-packed metal surfaces, including layered alloys.³³ However, they applied PCA to a matrix of thermochemical data rather than electronic-structure data.

We first analyze the predictive performance of the PC descriptors when used in a regression application to predict chemisorption energies. Models built with the PC descriptors exhibit good out-of-sample performance when predicting the DFT-computed adsorption energies of a hold-out test set, with an average test error of 0.062 eV across all adsorbates (**Figure S1**). To compare the performance between ML models built using PC-based descriptors and models built using traditional electronic-structure descriptors and the full DOS,^{6,40–46} we performed rigorous nested cross-validation (CV) analysis of various ML model classes, shown in **Figure 2**. Specifically, we constructed Gaussian process regression (GP), random forest regression (RF), explainable boosting regression (EBM), gradient boosted regression (GB), support vector regression (SVR), and ridge regression (RR) models built using traditional electronic-structure descriptors (*d*-band center with respect to the Fermi level, *d*-band width, *d*-band upper edge with respect to the Fermi level, and *d*-band filling), the full density of states, and the top ten PC descriptors. We give additional details regarding the traditional descriptor calculations in the Supporting Information. The **Methods** section contains details regarding feature selection, ML model training, and cross-validation.

The data in **Figure 2** show that ML models using ten PCs as input features consistently yield CV root-mean-square errors (RMSEs) about a factor of two smaller than the models built using traditional descriptors for predicting chemisorption. The PC models are competitive with models built using the complete DOS, with average RMSEs <0.01 eV larger for all adsorbates studied. Of the ML models examined, GP models yield the highest predictive accuracy for all adsorbates. Therefore, we used GP models to examine singular value decomposition⁴⁷ and kernel PCA⁴⁸ as alternative dimensionality reduction methods (**Figure S2**) but found neither of these to yield considerable performance improvements over PCA. The ML models using PCs as inputs perform similarly to the full DOS models based on average CV RMSE, even outperforming the complete DOS in some cases. Typically, the PC descriptor models have an RMSE that is slightly larger than using the complete DOS, likely due to the inherent loss of information from dimensionality reduction.

Linking the Principal Component Descriptors to Chemisorption Energy and Geometric Structure and Composition of Alloys

To analyze how the adsorption behavior changes as a function of the PC scores, we have generated partial dependence plots (PDPs) for the first two PCs, displayed in **Figure 3a,b**.⁴⁹ We have selected the first two PCs because they describe 73.8% of the training data variance (the first PC describes 47.4% of the variance, the second PC describes 26.4% of the variance). From the third PC onwards, each PC captures <10% of the variance; thus, interpretation of these PCs was not considered (**Figure S3**).

The PDPs in **Figure 3** summarize how the adsorption trends for C, O, N, and H change for different scores of the first and second PCs. The data in **Figure 3a** show that the adsorption energies of all four adsorbates become more exothermic as the score of the first PC increases.

Conversely, **Figure 3b** shows that all four adsorbates bind weaker as the score of the second PC increases. In general, the electronic-structure trends captured by the PC descriptors affect the adsorption process similarly across different atomic adsorbates.

To determine the electronic-structure effects captured by each PC descriptor, we analyze the DOS reconstructions as a function of each PC descriptor's score. Despite being well-established in the cognitive neuroscience and machine intelligence communities for decades,⁵⁰ interpretation of the PCs via reconstruction has yet to see use in the field of catalysis. Previous efforts were unable to leverage this aspect of PCA because they used atomic, geometric, or energetic features,^{33–39} resulting in difficult-to-interpret PCs that are high-dimensional linear combinations of the input features. In contrast, an electronic DOS is like a signal, and therefore the impact of the individual PC on the overall reconstruction can be visualized as a signal, which enables us to explain the significant information packaged in each PC descriptor.

The data in **Figure 4a,b** reveal how the first and second PC scores affect the DOS reconstructions. The reconstructions are robust and remain unchanged when learning the PCs on only a subset of the data (**Figure S4**). **Figure 4c–f** show the behavior of summary statistics related to the DOS reconstructions' statistical moments (*d*-band center or first moment, width or second moment, skewness or third moment, and kurtosis or fourth moment) as a function of the scores of the first two PCs. The *d*-band center describes the average energy of the DOS relative to the Fermi level, and the *d*-band width describes the width of the DOS. The skewness describes the DOS's asymmetry, whereas the kurtosis is related to the fourth moment of the DOS and is a measure of the fractional occupation of the states in the tails of the DOS. Additional details regarding the calculation of these summary statistics are given in the Supporting Information. We note that the skewness and kurtosis are unitless because they are scaled by the standard deviation.

The data in **Figure 4c** show that the *d*-band center shifts up in energy and becomes narrower as the first PC score increases. The correlation between the first PC and the *d*-band center is strong, with a Pearson correlation coefficient of 0.85. We recall that increasing the first PC score leads to more exothermic chemisorption. Furthermore, the kurtosis increases nearly monotonically with the first PC score (**Figure 4e**). This result would indicate that a *d*-band with a higher first PC score has higher occupations in the DOS distributions' tails based on the definition of kurtosis. In addition to changing the DOS distribution and its low-degree momenta, **Figure 4a** shows that the first PC describes fine-structure effects not captured by the low-degree DOS momenta. For example, we observe a shift in the lower *d*-band edge with the first PC score, suggesting that the lower *d*-band edge may be one descriptor motif to explore in future work. Prior studies have attributed similar electronic-structure effects to the degree of orbital overlap experienced by surface atoms.^{3,4,8,51} Connecting this to the first PC, we can see that widening and downshifting the *d*-states of a surface site, which corresponds to lowering the first PC scores, corresponds with a higher degree of orbital overlap.

The data in **Figure 4b** show the effect of changing the second PC score on the DOS, and **Figure 4d,f** shows the second PC score's influence on the statistical moments. The *d*-band widens and shifts down in energy as the second PC score increases, **Figure 4d**. However, the correlation between the *d*-band center and the second PC score is much weaker than that between the first PC score and the *d*-band, with a Pearson correlation coefficient of -0.37 . Additionally, the shape of the DOS changes based on the value of the second PC score. For example, the data in **Figure 4f** show that the skewness increases along with the second PC score. Based on the definition of skewness, this indicates that the occupation at the upper *d*-band edge also increases with the second PC score. A marginal decrease in the kurtosis occurs as the second PC score increases, indicating

fewer d -states in the distributions' tails. Like the first PC, the second PC also captures fine-structure effects that are not well-described by the low-degree d -band momenta. For example, the upper band edge shifts in conjunction with the second PC, indicating that the position of the upper band edge relative to the Fermi level may be descriptive of adsorption trends. Previous works have attributed similar electronic-structure effects to a metal surface's valence d -electron character.⁵² This prior result would suggest that surface sites with higher upper band edges and lower second PC scores correspond to metals with a lower number of valence d -electrons. Conversely, surface sites with lower upper band edges and higher second PC scores correspond to metals with a higher number of valence d -electrons.

We demonstrated above that the PC descriptors yield accurate chemisorption energy predictions and provide insight into the electronic-structure differences between different surface sites. However, connecting alloys' geometric structure and composition to their chemisorption strengths is critical to move beyond electronic-structure models towards building predictive geometry-chemisorption strength models. We emphasize that electronic structure serves as an invaluable link between geometric structure and chemisorption and can aid in identifying geometric descriptors. To leverage this, we created box and swarm plots, shown in **Figure 5a-f**, which reveal how the first and second PC score distributions, and therefore the electronic structure, change due to variations in the geometric structure and composition of the alloys.

We first analyze how the PC scores change as a function of changing an alloys' surface metal. The data in **Figure 5a** show that the $4d$ surface metals (i.e., Rh and Pd) have similar first PC score distributions. Likewise, the $5d$ surface metals (i.e., Ir and Pt) also have similar first PC score distributions, which are lower in value than the $4d$ surface metal distributions. The primary difference between $4d$ metals and $5d$ metals of the same group is increased atomic size. Because

the increase in the first PC leads to stronger chemisorption, these results indicate that alloys with $5d$ surface metals (larger atomic radius) bind adsorbates weaker than $4d$ surface metals (smaller atomic radius) for similar d -valence electronic structures, i.e., chemisorption strength decreases moving down a given group of the periodic table. Additionally, the first PC score distributions in **Figure 5a** do not show a discernible trend concerning the surface metal group, which is descriptive of the number of valence d -electrons. This finding suggests that the surface metal's size accounts for the surface metal's primary geometric effect on the first PC score.

The data in **Figure 5b** show that the second PC score depends primarily on the number of valence d -electrons in the surface metal. Surface metals in a higher group have higher second PC scores, indicating that the second PC score increases as the number of valence d -electrons in the surface metal increases. Since increasing the second PC score leads to weaker chemisorption (**Figure 3b**), these results indicate the chemically intuitive result that transition metals with more valence d -electrons bind adsorbates less strongly.

Figure 5c,d show the impact of the subsurface ligand atoms on the first and second PC scores. **Figure 5c** shows that alloy surfaces with late transition metal ligands like Cu, Ag, and Au lead to higher first PC scores (and stronger chemisorption) than earlier transition metal ligands such as Rh, Ir, and Ni. Furthermore, we also observe trends in the first PC score within a given group of subsurface atoms. For example, a Cu subsurface ligand atom forms alloys characterized by a lower first PC score and more endothermic chemisorption than alloys containing Ag or Au subsurface ligands. The data in **Figure 5d** suggest that, in general, there is only a weak relationship between the second PC and the ligand metal identity (**Figure S5d,f**).

By comparing the findings associated with **Figure 5a,b**, which shed light on the impact of the surface metal on the PC descriptors, and **Figure 5c,d**, which shed light on the role of ligand

atoms, we can arrive at simple but powerful conclusions about how the character of the surface and ligand metals impact the chemisorption strength in metal alloys, summarized in **Schematic 2**. The number of valence *d*-electrons and the metal atoms' sizes are two critical and easily accessible parameters that govern the chemisorption strength. In general, we find that as the filling of the surface metal atom's *d*-band increases, indicating a higher number of valence *d*-electrons, chemisorption becomes more endothermic. Based on the data in **Figure 5c**, we observe that the *d*-electron character of the subsurface ligand atoms has precisely the opposite effect on the chemisorption strength compared to the surface atoms. Ligand atoms with fuller *d*-bands lead to more exothermic chemisorption on the surface metal site. We also observe that chemisorption becomes more endothermic as the atom size increases for a fixed number of *d*-electrons in both the ligand and surface metal atoms.

It is critical to compare these findings with established chemisorption models and electronic-structure theories.⁵³⁻⁵⁵ For most adsorbates (including those analyzed in this work), metals with lower energy *d*-band centers bind adsorbates weaker than metals with higher energy *d*-band centers. Changing an alloy's composition can tune the surface site's *d*-band center. For example, for atoms with an equivalent number of *d*-electrons, changing the degree of orbital overlap experienced by surface atoms changes the position of the *d*-band center. Alloys containing larger surface and ligand metals have more orbital overlap and therefore have wider and lower-in-energy *d*-band centers. This effect is captured by the first PC.^{3,4,51} Additionally, surface metal atoms with a higher number of valence *d*-electrons, and therefore fuller *d*-bands, generally have lower energy *d*-bands. This effect appears to be well-described by the second PC, with surface metals with more valence *d*-electrons having higher second PC scores. The role of the number of *d*-electrons in the ligand metal on chemisorption captured by the PC descriptors is also consistent

with theories of chemisorption on metals as explained by bond-order conservation arguments.^{56,57} Surface atoms in alloys with more noble ligands (containing more valence *d*-electrons), characterized by higher first PC scores, compensate for weaker metal-metal interactions with the ligand metal by interacting with adsorbates more strongly. In sum, these results indicate that the conclusions we derived from the PCA descriptors regarding the impact of the character of the surface and ligand metals on the chemisorption strength in metal alloys are entirely consistent with prior chemisorption theories.

We have also used PC descriptors to analyze the impact of geometric strain on the electronic structure and chemisorption strength. The data in **Figure 5e** show that strain has a significant effect on the first PC score (**Figure S5e**), with increasing tensile strain resulting in larger first PC scores. This trend suggests that the first PC primarily describes the degree of orbital overlap between metal atoms near the surface. The data reveals that increasing compressive strain leads to a monotonic downward shift in the first PC score distribution, corresponding to a widening of the *d*-band and a downshift in the *d*-band center. Consequently, the adsorption energy weakens for C, O, N, and H adsorbates under compressive strain. This observation is consistent with prior reports showing that compressive strain modifies alloys' electronic structures by widening their *d*-band and shifting them down in energy to maintain their filling.^{4,8} On the other hand, **Figure 5f** shows that the second PC scores barely change due to strain (**Figure S5f**). This trend is consistent with the idea that the second PC primarily captures the surface metal's valence *d*-electron character, which should be largely unaffected by strain.

This PCA analysis allowed us to identify electronic-structure descriptors for alloys that shed light on the relationships between a surface site's geometric structure (including the direct adsorption site and ligand atoms) and the chemisorption strength of different adsorbates on that

site. We show that that the first two PCs connect well to prior efforts to develop electronic-structure descriptors and are primarily descriptive of the degree of orbital overlap at the surface and the number of valence d -electrons in the surface metal. We show that these effects can be well-captured by two atomic characteristics of the alloys' constituent metals, namely the surface and subsurface atoms' number of d -electrons and sizes. Importantly, our PCA approach arrives at these descriptors without engaging any existing chemisorption theory beforehand.

Using PCA to Find an Electronic-Structure Descriptor for Oxygen Reactivity on Metals and Metal Oxides

To establish that this approach can be extended beyond chemisorption models for metal alloys, we employed this workflow to identify electronic and geometric descriptors of surface oxygen reactivity on metals, rutile metal oxides, and perovskite metal-oxides (ABO_3). The reactivity of a surface oxygen species is defined in terms of its ability to bind a hydrogen atom ($\Delta E_O - \Delta E_{OH} = E_{O^*} + 1/2E_{H_2(g)} - E_{OH^*}$). Finding electronic-structure descriptors for reactive oxygen species has been a long-standing challenge with broad implications for understanding a material's catalytic performance for oxidation chemistries, such as the oxidative coupling of methane,⁵⁸ oxygen evolution,⁵⁹ propylene epoxidation,⁶⁰ and many others.⁶¹ Dickens et al. have proposed the O $2p$ states' average energy (ϵ_{2p}) as an electronic-structure descriptor for quantifying the surface oxygen atom's reactivity.⁶² The O $2p$ states typically have a bimodal distribution because they segment into bonding and anti-bonding states, so the average energy may not describe this distribution with a high degree of accuracy.

We applied PCA to the DOS of surface oxygen species on metals, rutile metal oxides, and perovskite metal-oxides to construct a descriptor of the O $2p$ states. We used a dataset from Dickens et al.⁶² that contains the $2p$ DOS between -10 eV and 2 eV relative to the Fermi level for

O species on 97 pure fcc metals, 32 rutile metal-oxides, and 166 perovskite metal-oxides (**Figure 6a**). The states of both lattice O and superstoichiometric adsorbed O (similar to the adsorbate evolution mechanism of the oxygen evolution reaction⁵⁹) are considered for oxides. Using the top seven PC descriptors for constructing Gaussian process (GP) models (**Figure S8**) results in a test RMSE of 0.486 eV. We note that including *d* or *f* states that might accept localized or delocalized electrons from the hydrogen atom may be one avenue to improve the predictive capacity of oxygen reactivity descriptors. Nonetheless, only O 2*p* states were considered in this study to enable direct comparison between the PCA descriptors to the previously reported ϵ_{2p} descriptor from Dickens et al. Additionally, we observe that even stronger predictive performance is obtained when using the non-linear descriptors identified using kernel PCA (**Figure S11**). However, it is not possible to interpret the individual effects of these non-linear descriptors using reconstruction; therefore, we proceed with standard PCA.

We analyze the O 2*p* DOS reconstructions as a function of the PC descriptor's score to interpret the electronic-structure effects captured by our descriptor, shown in **Figure 6b**. Only the first PC is analyzed because it describes 36.7% of the total variance of the data, more than twice as much as any of the remaining PCs (**Figure S6**). The data show that DOSs with low first PC scores are characterized by a high and discrete atom-like occupancy in the bonding states and a lower metal-like occupancy in the anti-bonding states, whereas the converse is true for DOSs with high first PC scores. This finding suggests that the first PC describes the relative occupancy of the bonding and anti-bonding states, a physical quantity that generally correlates with the ϵ_{2p} descriptor (Pearson correlation coefficient of 0.93, **Figure S7**).

To analyze how the relative occupancies of the bonding and anti-bonding states affect the surface oxygen reactivity, $E_O - \Delta E_{OH}$, we generated a PDP showing how $E_O - \Delta E_{OH}$ changes as a

function of the first PC score (**Figure 6c**). The data show that, in general, $\Delta E_O - \Delta E_{OH}$ increases as the first PC score increases. This result is consistent with a physical chemistry description of bonding since it suggests that surface oxygen species with lower (higher) PC scores, which are more stable due to a higher (lower) relative filling of the bonding orbital, are less (more) likely to abstract hydrogen atoms or be active for oxidation chemistries.

We analyze trends in the first PC regarding material composition (**Figure 6d,e**), which reveals that the trends concerning the base metal bonded to the oxygen are broadly consistent with what we observed for alloys. The base metal's number of *d*-electrons and size are descriptive of the oxygen reactivity for metals and rutile oxides. As the base metal becomes nobler (with nobility increasing with atomic radius from 3*d* metals to 5*d* metals and increasing with the number of valence *d*-electrons²) the oxygen species becomes more unstable, and therefore more reactive. This insight rationalizes the fact that the most active materials for oxidation chemistries are typically platinum-group metal oxides (e.g., RuO₂ and IrO₂).^{63,64}

For perovskite oxides, we observe that changing the valence *d*-electron character of the B-site cation (the site that interacts most strongly with the oxygen atoms) leads to similar electronic structure effects as changing the valence *d*-electron character of the base metal in a pure metal or rutile oxide (**Figure S9b**). As the number of valence *d*-electrons increases, the O 2*p* states move up in energy and become characterized by higher first PC scores. However, we observe an opposite trend to pure metals and rutile oxides concerning changes in the B-site cation's atomic size (**Figure S9a**). As the size of the B-site cation increases, the oxygen states shift down in energy and become characterized by lower first PC scores. To explain this counter-intuitive result at the dataset level, we investigated local trends in the O 2*p* states of perovskite oxides with a fixed A-site cation, focusing on materials where the A-site cation is La (**Figure S10**). Visualization of the La-based

perovskite O $2p$ states suggests that they generally behave similarly to how they do for metals and rutile oxides. As the size of the B-site cation increases, the metal-oxygen interaction becomes weaker, leading to a lower degree of hybridization and more atom-like O $2p$ states. While this leads to states with higher average energy for rutile oxides and metals, the average energy of the O $2p$ states becomes lower as the states become atom-like in perovskites; this leads to an opposite trend in the first PC score with respect to the B-site cation's atomic size.

In summary, we extended our PCA approach to identify electronic and geometric descriptors of surface oxygen reactivity for metals, rutile metal-oxides, and perovskite metal-oxides. Analysis of the first PC descriptor via reconstruction suggests that the descriptors capture trends that are consistent with a physical chemistry description of bonding. For example, low first PC scores are characterized by a high and discrete atom-like occupancy in the bonding states and a lower metal-like occupancy in the anti-bonding states, whereas the converse is true for DOSs with high first PC scores, suggesting that the first PC score captures the relative occupations of the O bonding and anti-bonding orbitals. Analysis of the surface oxygen reactivity trends with respect to the PC descriptor suggests that materials with high first PC scores, characterized by unstable surface oxygen species with a higher relative filling of the anti-bonding orbital, are active towards oxidation chemistries.

Conclusions

Unsupervised learning represents a promising paradigm for the efficient discovery of descriptors for materials. We use a simple approach that leverages PCA to identify electronic-structure descriptors based on the site-projected d -band density of states of alloys containing late transition metals. The PC descriptors yield accurate ML models for predicting chemisorption that outperform models built using traditional electronic-structure descriptors. Importantly, the PC descriptors are interpretable, thus giving insight into how a material's electronic structure is connected to surface geometry and composition, and ultimately chemisorption strength. We demonstrate that the approach presented herein could be applied to diverse catalytic systems through a case study examining surface oxygen reactivity for metals, rutile metal-oxides, and perovskite metal-oxides. Therefore, we expect this approach can extend readily to other catalytic systems such as intermetallic and random alloys, nitrides, and sulfides, as well as other application fields across chemistry and materials science, such as describing activity trends for homogeneous catalysts and developing relationships between electronic structure and functionality for photovoltaic materials.

Experimental procedures

Resource availability

Lead contact

Further requests for information and resources should be directed to the Lead Contact, Suljo Linic (linic@umich.edu).

Materials Availability

This study did not generate new materials or reagents.

Data and code availability

The data and codes generated during this study are available on GitHub at: https://github.com/jesterhui/pca_electronic_structure_descriptors. This repository contains all DFT calculated structures and representative INCAR files, as well as representative Python scripts for carrying out PCA analysis. Structures are also available on the NOMAD repository at <https://dx.doi.org/10.17172/NOMAD/2021.06.22-1>.

DFT modeling of layered alloys and their chemisorption properties

The Vienna Ab initio Simulation Package (VASP) was used to perform all electronic-structure theory calculations.^{65–68} We used the Perdew-Burke-Ernzerhof (PBE) functional to describe the electron exchange and correlation,^{69,70} and the projector augmented wave method to describe electron-ion interactions. A plane-wave kinetic-energy cutoff of 450 eV was selected, and a 4×4×1 Monkhorst-Pack k -point grid was used for the Brillouin zone integration.⁷³ Benchmarking calculations indicate that the DOS are converged using a 4×4×1 Monkhorst-Pack k -point grid (**Figure S12**).

Layered alloy surfaces were studied using the slab model illustrated in **Figure 1b–c**, which has been used in prior studies to probe the ligand and strain effects.^{3,4,9,10} Each layered-alloy slab model consists of five layers of metal atoms composed entirely of the base metal, except for the second layer of atoms directly below the surface layer that is composed of the ligand metal. The bottom three layers of the model system were fixed at the bulk lattice constant of the corresponding metal and was uniformly adjusted for strain. We studied slabs with strains of –2% to +2% in intervals of one percent. Conjugate gradient geometry optimization was performed on the upper two layers of the slab and the adsorbate, with all other layers fixed in their bulk lattice positions. Geometry optimization was stopped when the maximum forces on all atoms in the system were less than 0.03 eV/Å. Alloys containing Fe, Co, and Ni were studied with collinear spin-

polarization. Model systems were constructed and manipulated using the Atomic Simulation Environment.⁷⁴ For all adsorbates, we considered electronic adsorption energies (zero-point energies were not included) defined relative to the corresponding free atom:

$$\Delta E_i = E_{i/M} - (E_M + E_i)$$

Where M denotes the slab and i denotes the adsorbate. All alloy structures studied are available on the NOMAD repository at <https://dx.doi.org/10.17172/NOMAD/2021.06.22-1>.

Identifying electronic-structure descriptors with PCA

PCA is an unsupervised learning approach that has seen use in previous works to extract catalysis descriptors, which were then used in conjunction with supervised learning methods to yield accurate and computationally efficient models.³³⁻³⁹ However, this work represents the first example of PCA explicitly applied to the electronic-structure data of surfaces. For the alloy dataset, the first step was to construct a DOS matrix containing the d -projected DOS for all the alloys, which is visualized in **Schematic 1a**. The VASP output samples different energy values for different systems when outputting the DOS. However, for this analysis the DOS must be sampled at uniform energy intervals. A one-dimensional interpolating spline was fit to the DOS of each alloy, and the DOS was sampled uniformly for 300 points between -10 eV to 10 eV. The DOS were normalized such that the integral of each DOS was unity. This analysis yielded the data matrix shown in **Schematic 1a**. Next, dimensionality reduction (including PCA, Kernel PCA, singular value decomposition) on the DOS matrix (containing both the training and test set splits) was performed using the scikit-learn package to find a minimal set of descriptors for the data.⁷⁵

In general, the segmenting of the cross-validation folds usually occurs before feature and model selection procedures apply. However, Hastie, Friedman, and Tibshirani note that initial unsupervised screening steps can be done before cross-validation fold segmentation.³¹ Since this

filtering does not involve the data labels, it should not bias the predictors nor give an unfair advantage. Testing this assumption empirically indeed suggests that selecting descriptors prior to performing the supervised learning step does not yield noticeable improvements in predictive performance (**Table S1**). For the alloy dataset, ten PCs were selected because this was the minimum number of PCs that the GP models required to converge in RMSE for adsorption energy predictions based on nested cross-validation (**Figure S2**).

For our case study of surface oxygen reactivity, a dataset from Dickens et al. was used.⁶² This dataset contained hydrogen binding energies and the O 2*p* DOS for 97 oxygen species on fcc (111) surfaces, 32 rutile metal-oxide oxygen species, and 166 perovskite-oxide oxygen species. Like alloys, an interpolating spline was fit to the DOS and then the DOS was sampled uniformly over 600 points between -10 eV and 2 eV.

Machine learning model construction

Explainable Boosting Machine regression models (EBM, also known as iGAM models) and partial dependence plots were built using Microsoft Research's InterpretML package.⁷⁶ All other ML models were learned using the scikit-learn package.⁷⁵ Nested cross-validation,⁷⁷ with ten folds in both the inner and outer loops, was performed three times and the performance averaged for evaluating model generalization performance as a function of the ML model and the dimensionality reduction routine used. All ML models were constructed with ten PCs, except for the EBM models, which were constructed with five PCs due to a drop in cross-validation performance upon the addition of more PCs into the model. Partial dependence plots were built using were built using the InterpretML package.⁷⁶ The data used for machine learning model construction for the alloys and metal oxides, as well as scripts used for model training and testing

are available from the authors on GitHub at https://github.com/jesterhui/pca_electronic_structure_descriptors.

Acknowledgements

This work was supported by the US DOE Office of Basic Energy Sciences, Division of Chemical Sciences (DE-SC0021008). Secondary support for DFT calculations on alloy surfaces was provided by the US DOE Office of Basic Energy Sciences, Division of Chemical Sciences (FG-02-05ER15686). This work was partially supported by faculty start-up funds of Goldsmith from the University of Michigan, Ann Arbor. J.A.E. also acknowledges support from the University of Michigan J. Robert Beyster Computational Innovation Graduate Fellows Program. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

Author Contributions

Conceptualization, J.A.E., B.R.G., and S.L.; Methodology, J.A.E., B.R.G., and S.L.; Software, J.A.E.; Formal analysis, J.A.E.; Investigation, J.A.E.; Resources, B.R.G. and S.L.; Data Curation, J.A.E.; Writing – Original Draft, J.A.E.; Writing – Review & Editing, J.A.E., B.R.G., and S.L.; Visualization, J.A.E.; Supervision, B.R.G. and S.L.; Project Administration, B.R.G. and S.L.; Funding Acquisition, B.R.G. and S.L.

Declaration of Interests

B.R.G. is a member of the Chem Catalysis advisory board.

References

1. Zhao, Z.-J., Liu, S., Zha, S., Cheng, D., Studt, F., Henkelman, G., and Gong, J. (2019). Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. *Nat. Rev. Mater.* *4*, 792–804.
2. Hammer, B., and Nørskov, J.K. (1995). Why gold is the noblest of all the metals. *Nature* *376*, 238–240.
3. Kitchin, J.R., Nørskov, J.K., Barteau, M.A., and Chen, J.G. (2004). Modification of the surface electronic and chemical properties of Pt(111) by subsurface 3d transition metals. *J. Chem. Phys.* *120*, 10240–10246.
4. Kitchin, J.R., Nørskov, J.K., Barteau, M.A., and Chen, J.G. (2004). Role of Strain and Ligand Effects in the Modification of the Electronic and Chemical Properties of Bimetallic Surfaces. *Phys. Rev. Lett.* *93*, 156801.
5. Vojvodic, A., Nørskov, J.K., and Abild-Pedersen, F. (2014). Electronic Structure Effects in Transition Metal Surface Chemistry. *Top Catal* *57*, 25–32.
6. Xin, H., Vojvodic, A., Voss, J., Nørskov, J.K., and Abild-Pedersen, F. (2014). Effects of d-band shape on the surface reactivity of transition-metal alloys. *Phys. Rev. B* *89*.
7. Andersen, M., and Reuter, K. (2021). Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Acc. Chem. Res.* *54*, 2741–2749.
8. Mavrikakis, M., Hammer, B., and Nørskov, J.K. (1998). Effect of Strain on the Reactivity of Metal Surfaces. *Phys. Rev. Lett.* *81*, 2819–2822.
9. Xin, H., Holewinski, A., and Linic, S. (2012). Predictive Structure–Reactivity Models for Rapid Screening of Pt-Based Multimetallic Electrocatalysts for the Oxygen Reduction Reaction. *ACS Catal.* *2*, 12–16.

10. Xin, H., and Linic, S. (2010). Communications: Exceptions to the d-band model of chemisorption on metal surfaces: The dominant role of repulsion between adsorbate states and metal d-states. *J. Chem. Phys.* *132*, 221101.
11. Fung, V., Hu, G., Ganesh, P., and Sumpter, B.G. (2021). Machine learned features from density of states for accurate adsorption energy prediction. *Nature Communications* *12*, 88.
12. Back, S., Yoon, J., Tian, N., Zhong, W., Tran, K., and Ulissi, Z.W. (2019). Convolutional Neural Network of Atomic Surface Structures to Predict Binding Energies for High-Throughput Screening of Catalysts. *J. Phys. Chem. Lett.*, 4401–4408.
13. Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., et al. (2021). Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* *11*, 6059–6072.
14. Yan, J., Gorai, P., Ortiz, B., Miller, S., Barnett, S.A., Mason, T., Stevanović, V., and Toberer, E.S. (2015). Material descriptors for predicting thermoelectric performance. *Energy Environ. Sci.* *8*, 983–994.
15. Wang, Z., Chu, I.-H., Zhou, F., and Ong, S.P. (2016). Electronic Structure Descriptor for the Discovery of Narrow-Band Red-Emitting Phosphors. *Chem. Mater.* *28*, 4024–4031.
16. Kirchartz, T., and Rau, U. (2018). Linking structural properties with functionality in solar cell materials – the effective mass and effective density of states. *Sustain. Energy Fuels* *2*, 1550–1560.
17. Fey, N. (2010). The contribution of computational studies to organometallic catalysis: descriptors, mechanisms and models. *Dalton Trans.* *39*, 296–310.
18. Kirchmair, J., Williamson, M.J., Afzal, A.M., Tyzack, J.D., Choy, A.P.K., Howlett, A., Rydberg, P., and Glen, R.C. (2013). FAsT METabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *J. Chem. Inf. Model.* *53*, 2896–2907.
19. Esterhuizen, J.A., Goldsmith, B.R., and Linic, S. (2020). Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem* *6*, 1–18.
20. Jr, W.D.C., and Rethwisch, D.G. (2012). *Fundamentals of Materials Science and Engineering: An Integrated Approach* (John Wiley & Sons).
21. Vines, R.F. (1941). *The Platinum Metals and their Alloys* (The International Nickel Company).
22. T. Hong, W., Risch, M., A. Stoerzinger, K., Grimaud, A., Suntivich, J., and Shao-Horn, Y. (2015). Toward the rational design of non-precious transition metal oxides for oxygen electrocatalysis. *Energy Environ. Sci.* *8*, 1404–1427.

23. Nørskov, J.K., Bligaard, T., Logadottir, A., Kitchin, J.R., Chen, J.G., Pandelov, S., and Stimming, U. (2005). Trends in the Exchange Current for Hydrogen Evolution. *J. Electrochem. Soc.* *152*, J23.
24. Logadottir, A., Rod, T.H., Nørskov, J.K., Hammer, B., Dahl, S., and Jacobsen, C.J.H. (2001). The Brønsted–Evans–Polanyi Relation and the Volcano Plot for Ammonia Synthesis over Transition Metal Catalysts. *J. Catal.* *197*, 229–231.
25. Jones, G., Studt, F., Abild-Pedersen, F., Nørskov, J.K., and Bligaard, T. (2011). Scaling relationships for adsorption energies of C₂ hydrocarbons on transition metal surfaces. *Chem. Eng. Sci.* *66*, 6318–6323.
26. Liu, B., and Greeley, J. (2011). Decomposition Pathways of Glycerol via C–H, O–H, and C–C Bond Scission on Pt(111): A Density Functional Theory Study. *J. Phys. Chem. C* *115*, 19702–19709.
27. Mavrikakis, M., Rempel, J., Greeley, J., Hansen, L.B., and Nørskov, J.K. (2002). Atomic and molecular adsorption on Rh(111). *J. Chem. Phys.* *117*, 6737–6744.
28. Herron, J.A., Tonelli, S., and Mavrikakis, M. (2012). Atomic and molecular adsorption on Pd(111). *Surf. Sci.* *606*, 1670–1679.
29. Krekelberg, W.P., Greeley, J., and Mavrikakis, M. (2004). Atomic and Molecular Adsorption on Ir(111). *J. Phys. Chem. B* *108*, 987–994.
30. Ford, D.C., Xu, Y., and Mavrikakis, M. (2005). Atomic and molecular adsorption on Pt(111). *Surf. Sci.* *587*, 159–174.
31. Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning* T. Hastie, J. Friedman, and R. Tibshirani, eds. (Springer New York).
32. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Springer New York).
33. García-Muelas, R., and López, N. (2019). Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. *Nat Commun* *10*, 4687.
34. Saadun, A.J., Pablo-García, S., Paunović, V., Li, Q., Sabadell-Rendón, A., Kleemann, K., Krumeich, F., López, N., and Pérez-Ramírez, J. (2020). Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* *10*, 6129–6143.
35. Sieg, S.C., Suh, C., Schmidt, T., Stukowski, M., Rajan, K., and Maier, W.F. (2007). Principal Component Analysis of Catalytic Functions in the Composition Space of Heterogeneous Catalysts. *QSAR Comb. Sci.* *26*, 528–535.

36. Mostad, H.B., Riis, T.U., and Ellestad, O.H. (1990). Use of principal component analysis in catalyst characterization: Catalytic cracking of decalin over Y-zeolites. *Appl. Catal.* *64*, 119–141.
37. Abdelfatah, K., Yang, W., Vijay Solomon, R., Rajbanshi, B., Chowdhury, A., Zare, M., Kundu, S.K., Yonge, A., Heyden, A., and Terejanu, G. (2019). Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning. *J. Phys. Chem. C* *123*, 29804–29810.
38. Chowdhury, A.J., Yang, W., Walker, E., Mamun, O., Heyden, A., and Terejanu, G.A. (2018). Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J. Phys. Chem. C* *122*, 28142–28150.
39. Smith, A., Keane, A., Dumesic, J.A., Huber, G.W., and Zavala, V.M. (2020). A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Appl. Catal. B* *263*, 118257.
40. Hammer, B., and Nørskov, J.K. (2000). Theoretical surface science and catalysis—calculations and concepts. In *Advances in Catalysis Impact of Surface Science on Catalysis*. (Academic Press), pp. 71–129.
41. Lambert, R.M., and Pacchioni, G. eds. (1997). *Chemisorption and Reactivity on Supported Clusters and Thin Films* (Springer Netherlands).
42. Hammer, B., and Nørskov, J.K. (1995). Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* *343*, 211–220.
43. Andersen, M., Levchenko, S., Scheffler, M., and Reuter, K. (2019). Beyond scaling relations for the description of catalytic materials. *ACS Catal.* *9*, 2752–2759.
44. Pankajakshan, P., Sanyal, S., de Noord, O.E., Bhattacharya, I., Bhattacharyya, A., and Waghmare, U. (2017). Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.* *29*, 4190–4201.
45. Li, Z., Ma, X., and Xin, H. (2017). Feature engineering of machine-learning chemisorption models for catalyst design. *Catal. Today* *280*, 232–238.
46. Li, Z., Wang, S., Chin, W.S., Achenie, L.E., and Xin, H. (2017). High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* *5*, 24131–24138.
47. Kalman, D. (1996). A Singularly Valuable Decomposition: The SVD of a Matrix. *College Math. J.* *27*, 2–23.
48. Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* *10*, 1299–1319.

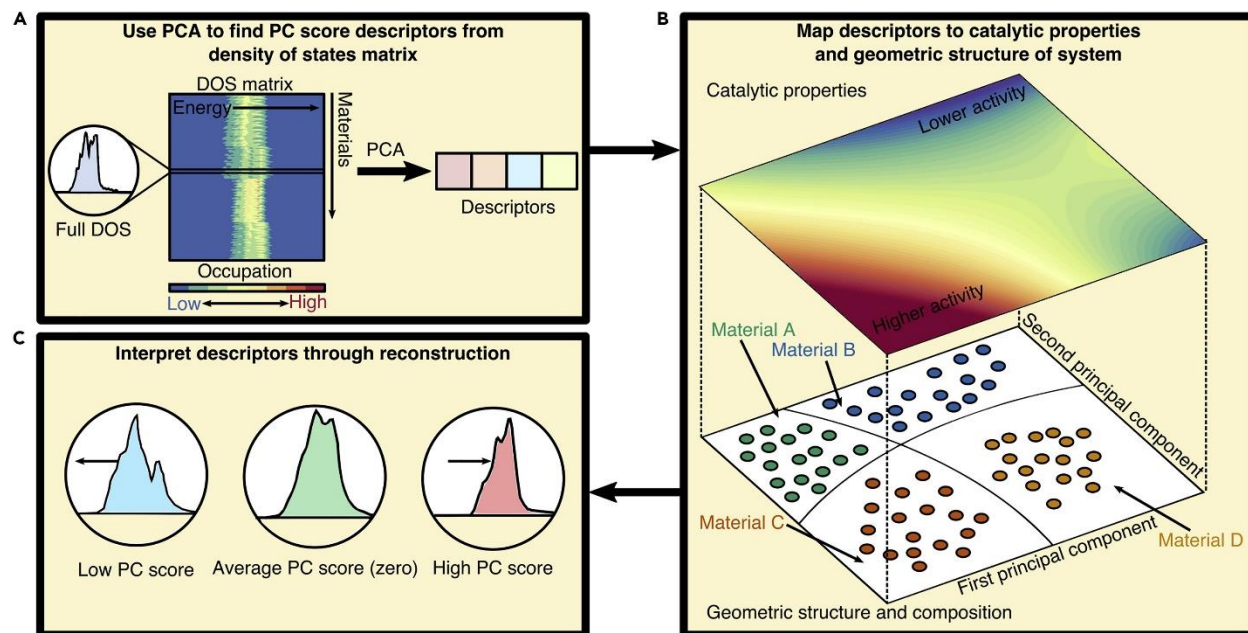
49. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* *29*, 1189–1232.
50. Turk, M., and Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* *3*, 71–86.
51. Schweitzer, N., Xin, H., Nikolla, E., Miller, J.T., and Linic, S. (2010). Establishing Relationships Between the Geometric Structure and Chemical Reactivity of Alloy Catalysts Based on Their Measured Electronic Structure. *Top Catal* *53*, 348–356.
52. Calle-Vallejo, F., Inoglu, N.G., Su, H.-Y., Martínez, J.I., Man, I.C., Koper, M.T.M., Kitchin, J.R., and Rossmeisl, J. (2013). Number of outer electrons as descriptor for adsorption processes on transition metals and their oxides. *Chem. Sci.* *4*, 1245–1249.
53. Newns, D.M. (1969). Self-Consistent Model of Hydrogen Chemisorption. *Phys. Rev.* *178*, 1123–1135.
54. İnoğlu, N., and Kitchin, J.R. (2010). Simple model explaining and predicting coverage-dependent atomic adsorption energies on transition metal surfaces. *Phys. Rev. B* *82*, 045414.
55. İnoğlu, N., and Kitchin, J.R. (2010). New solid-state table: estimating d-band characteristics for transition metal atoms. *Mol. Simul.* *36*, 633–638.
56. Shustorovich, E. (1984). Activation barrier for adsorbate surface diffusion, heat of chemisorption, and adsorbate registry: theoretical interrelations. *J. Am. Chem. Soc.* *106*, 6479–6481.
57. Shustorovich, E. (1988). Chemisorption theory: in search of the elephant. *Acc. Chem. Res.* *21*, 183–189.
58. Farrell, B.L., Igenegbai, V.O., and Linic, S. (2016). A Viewpoint on Direct Methane Conversion to Ethane and Ethylene Using Oxidative Coupling on Solid Catalysts. *ACS Catal.* *6*, 4340–4346.
59. Man, I.C., Su, H.-Y., Calle-Vallejo, F., Hansen, H.A., Martínez, J.I., Inoglu, N.G., Kitchin, J., Jaramillo, T.F., Nørskov, J.K., and Rossmeisl, J. (2011). Universality in Oxygen Evolution Electrocatalysis on Oxide Surfaces. *ChemCatChem* *3*, 1159–1165.
60. Torres, D., Lopez, N., Illas, F., and Lambert, R.M. (2007). Low-Basicity Oxygen Atoms: A Key in the Search for Propylene Epoxidation Catalysts. *Angewandte Chemie International Edition* *46*, 2055–2058.
61. Capdevila-Cortada, M., Vilé, G., Teschner, D., Pérez-Ramírez, J., and López, N. (2016). Reactivity descriptors for ceria in catalysis. *Applied Catalysis B: Environmental* *197*, 299–312.

62. Dickens, C.F., Montoya, J.H., Kulkarni, A.R., Bajdich, M., and Nørskov, J.K. (2019). An electronic structure descriptor for oxygen reactivity at metal and metal-oxide surfaces. *Surf. Sci.* *681*, 122–129.
63. Lee, Y., Suntivich, J., May, K.J., Perry, E.E., and Shao-Horn, Y. (2012). Synthesis and Activities of Rutile IrO₂ and RuO₂ Nanoparticles for Oxygen Evolution in Acid and Alkaline Solutions. *J. Phys. Chem. Lett.* *3*, 399–404.
64. McCrory, C.C.L., Jung, S., Peters, J.C., and Jaramillo, T.F. (2013). Benchmarking Heterogeneous Electrocatalysts for the Oxygen Evolution Reaction. *Journal of the American Chemical Society* *135*, 16977–16987.
65. Kresse, G., and Hafner, J. (1993). Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* *47*, 558–561.
66. Kresse, G., and Hafner, J. (1994). Ab initio molecular-dynamics simulation of the liquid-metal--amorphous-semiconductor transition in germanium. *Phys. Rev. B* *49*, 14251–14269.
67. Kresse, G., and Furthmüller, 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.
68. Kresse, G., and Furthmüller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* *54*, 11169–11186.
69. Perdew, J.P., Burke, K., and Ernzerhof, M. (1996). Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* *77*, 3865–3868.
70. Perdew, J.P., Burke, K., and Ernzerhof, M. (1997). Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. *77*, 3865 (1996)]. *Phys. Rev. Lett.* *78*, 1396–1396.
71. Blöchl, P.E. (1994). Projector augmented-wave method. *Phys. Rev. B* *50*, 17953–17979.
72. Kresse, G., and Joubert, D. (1999). From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* *59*, 1758–1775.
73. Monkhorst, H.J., and Pack, J.D. (1976). Special points for Brillouin-zone integrations. *Phys. Rev. B* *13*, 5188–5192.
74. Larsen, A.H., Mortensen, J.J., Blomqvist, J., Castelli, I.E., Christensen, R., Dułak, M., Friis, J., Groves, M.N., Hammer, B., Hargus, C., et al. (2017). The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* *29*, 273002.
75. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* *12*, 2825–2830.

76. Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 [cs, stat].

77. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Ser. B. Stat. Methodol. 36, 111–147.

Figure and scheme titles



Schematic 1. Workflow for automating electronic-structure descriptor identification using PCA. (a) Principal component analysis (PCA) identifies a lower-dimensional basis (i.e., the principal components) of a density of states (DOS) matrix to yield PC score descriptors. (b) These descriptors allow exploration of the links between a material's electronic structure, geometry, and catalytic properties (e.g., activity). (c) Notably, the electronic-structure effects captured in each descriptor can be analyzed and interpreted by reconstructing the DOS from the descriptors.

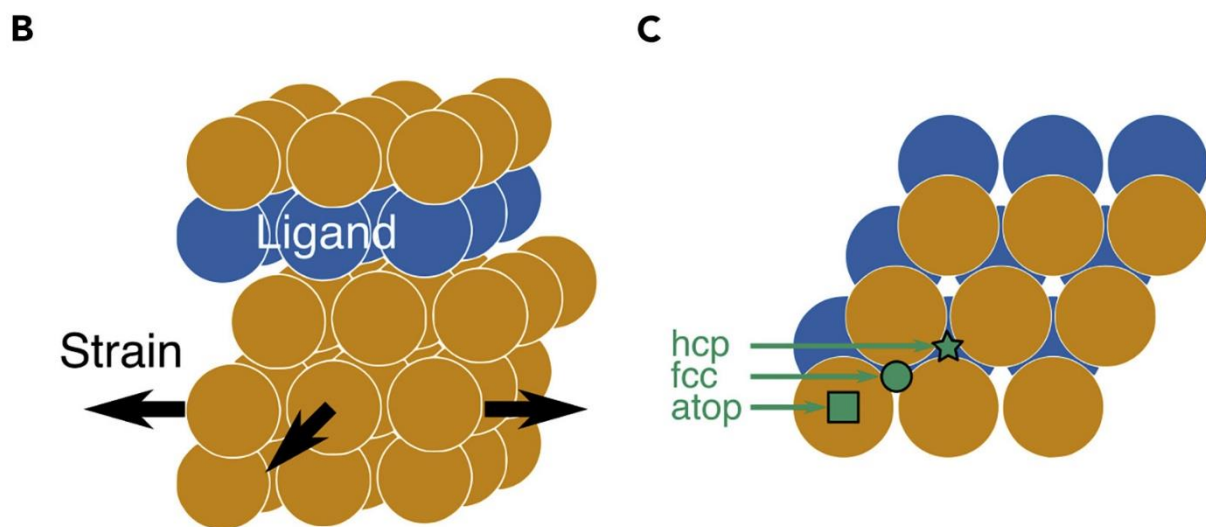
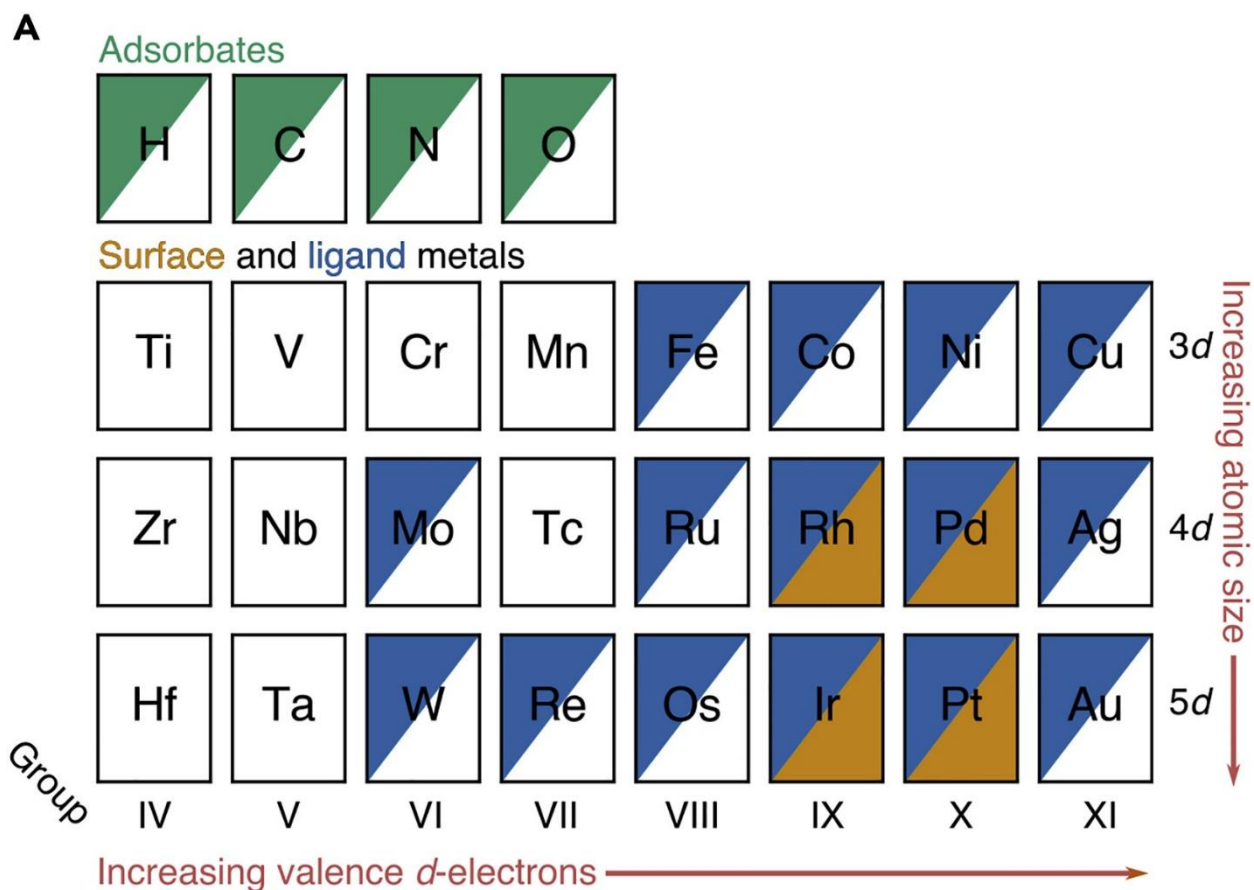


Figure 1. Layered alloy structure and composition. (a) Map of the elements considered as surface metals (orange), ligand metals (blue), and adsorbates (green). Not all ligand metals were studied with all surface metals—see main text for details. (b) A generalized example of the layered alloy model (111) surface. (c) A top-down view of the layered alloy (111) surface with adsorption sites labeled. The green star denotes the face-centered cubic (fcc) hollow binding site, the green circle denotes the hexagonal close-packed (hcp) hollow site, and the green square denotes the atop site.

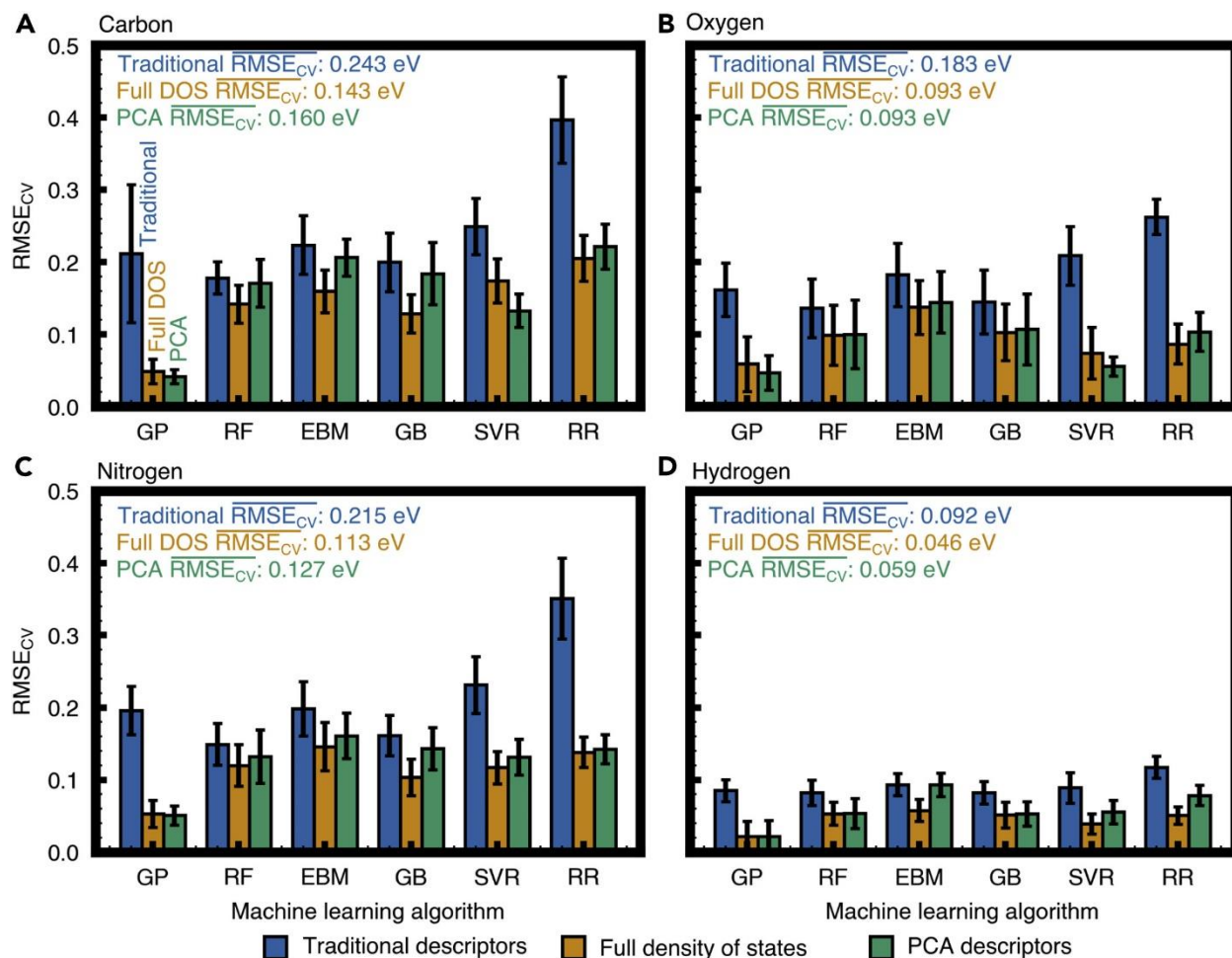


Figure 2. Cross-validation confirms PCA generalizability for different machine learning algorithms. Nested cross-validation, with ten folds in both the inner and outer loops, was performed for predicting (a) C, (b) O, (c) N, and (d) H adsorption energies on layered alloys. Colored bars show the cross-validation error for Gaussian process regression (GP), random forest regression (RF), explainable boosting regression (EBM), gradient boosted regression (GB), support vector regression (SVR), and ridge regression (RR) models built using the traditional electronic-structure descriptors (blue), the full density of states (orange), and the top ten PC descriptors (green). The error intervals denote one standard deviation in the cross-validation RMSE across the outer loops of the nested cross-validation procedure. The average cross-validation RMSE (\overline{RMSE}_{CV}) across all models for each feature set is reported inset.

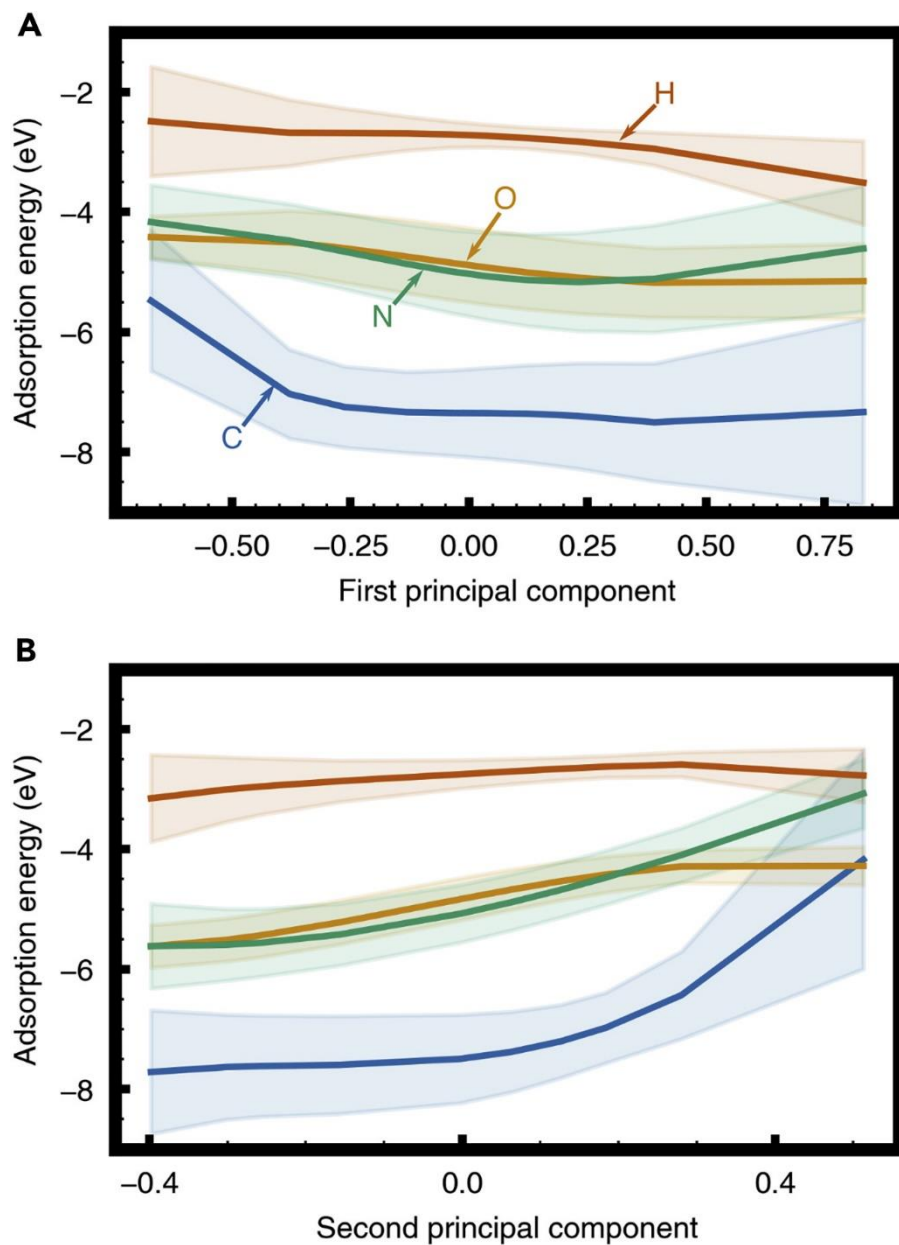


Figure 3. Effect of the principal components on the adsorption behavior of C, O, N, and H. Partial dependence plots of Gaussian process regression models are constructed for the (a) first and (b) second principal components. The shaded region denotes one standard deviation around the mean (solid line).

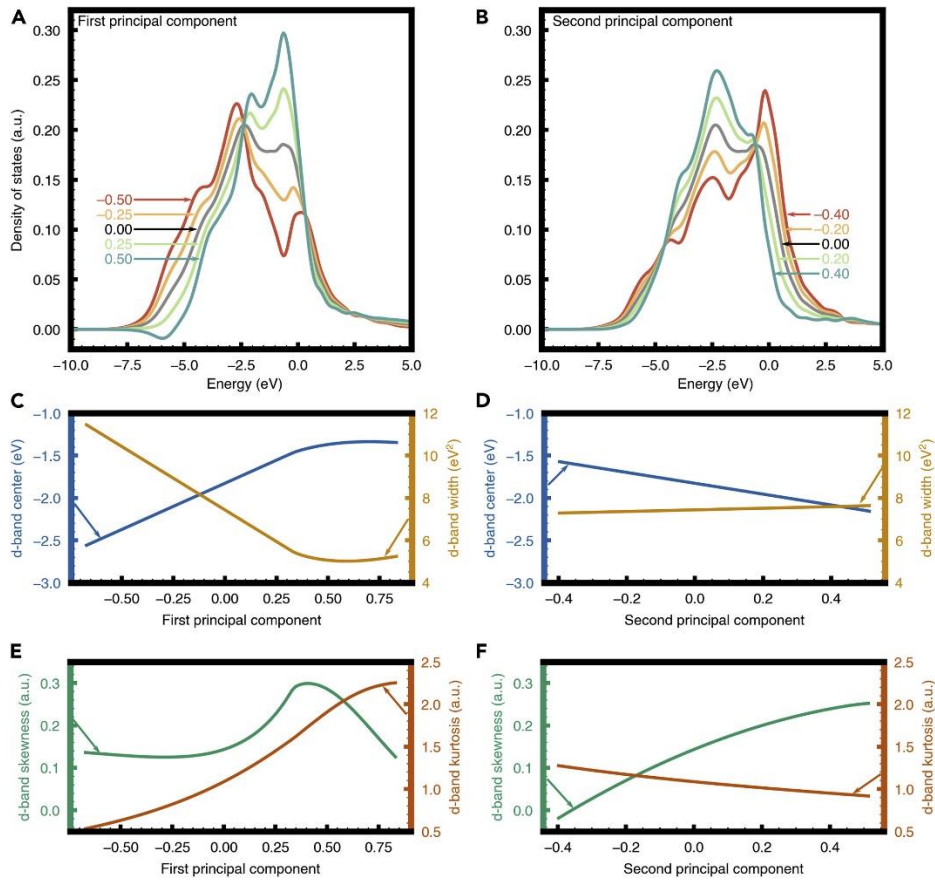


Figure 4. Analysis of individual principal components on density of states reconstruction. The data in the first row show how the (a) first and (b) second principal components affect the DOS reconstruction. The principal component values for each DOS reconstruction are provided inset. The data in the second and third rows show how the statistical moments of the DOS reconstructions depend on the first and second principal components. The second row of figures (c, d) displays the *d*-band center and width, and the third row of figures (e, f) shows the *d*-band skewness and kurtosis.

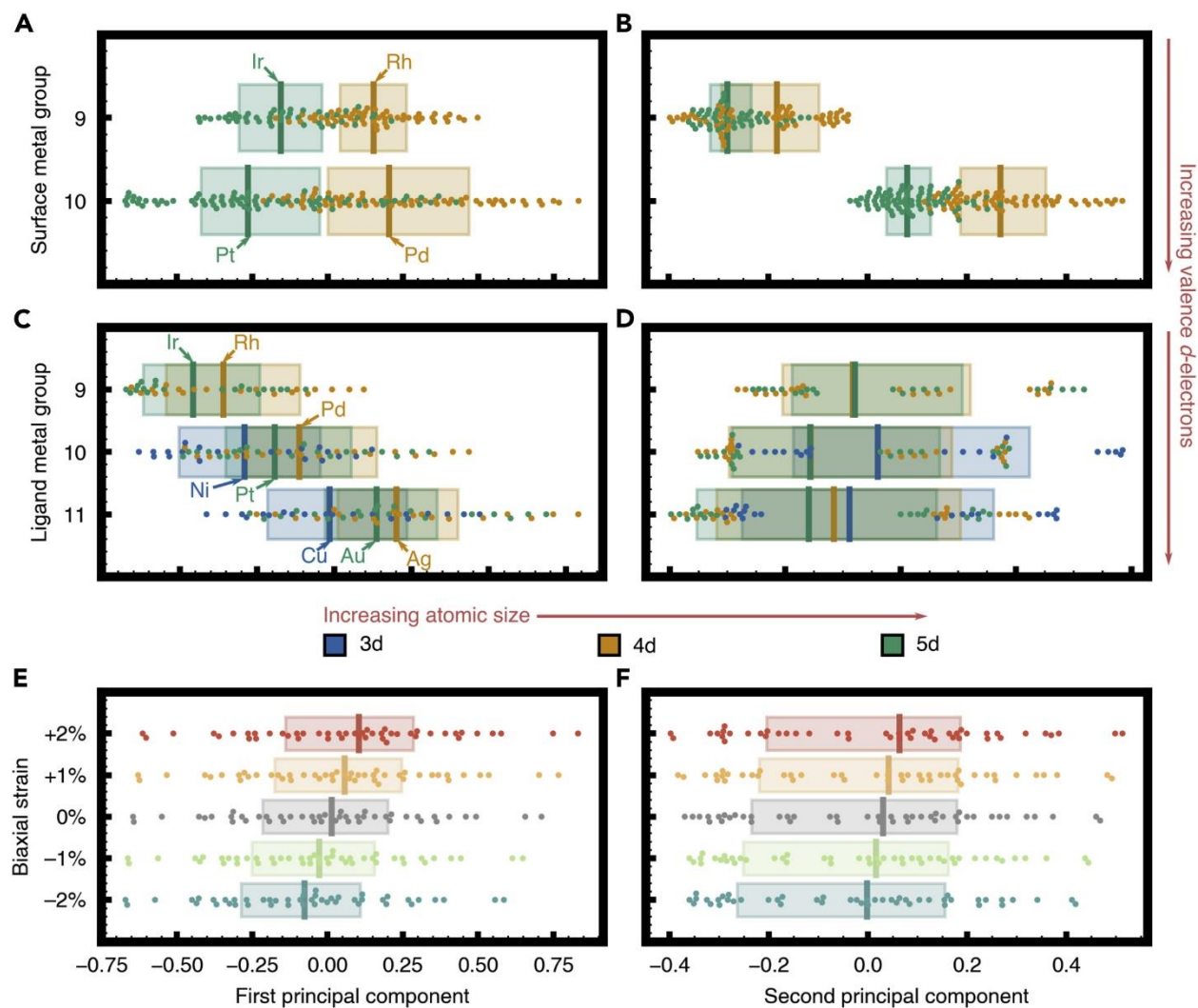
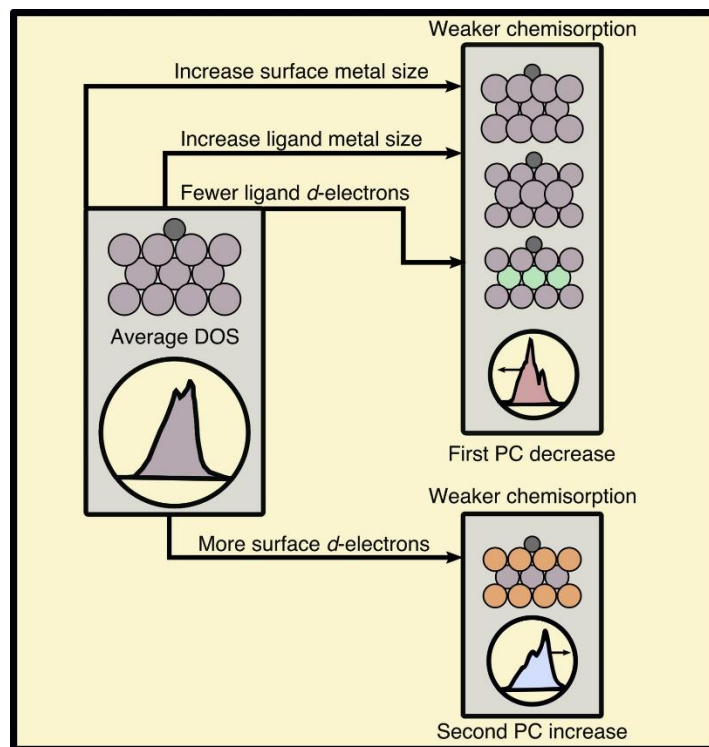


Figure 5. Linking the principal components to the alloys' geometric structure and composition. Box and swarm plots decompose the distributions of the first and second principal components as a function of changing (a, b) the surface metal, (c, d) the ligand metal, and (e, f) the alloy-induced biaxial strain in the x - y plane. Solid lines in the middle of each box denote the median value, and the box edges denote the interquartile range. In (a–d), the periodic row is color-coded such that 3d metals are blue, 4d metals are orange, and 5d metals are green. Labels of specific surface and ligand metals are displayed inset.



Schematic 2. How the character of the surface and ligand metals impact the chemisorption strength in metal alloys. Increasing the ligand and surface metal sizes and decreasing the number of *d*-electrons in the ligand metal correspond with a decrease in the first PC and weaker chemisorption. Increasing the number of surface *d*-electrons corresponds with an increase in the second PC and weaker chemisorption.

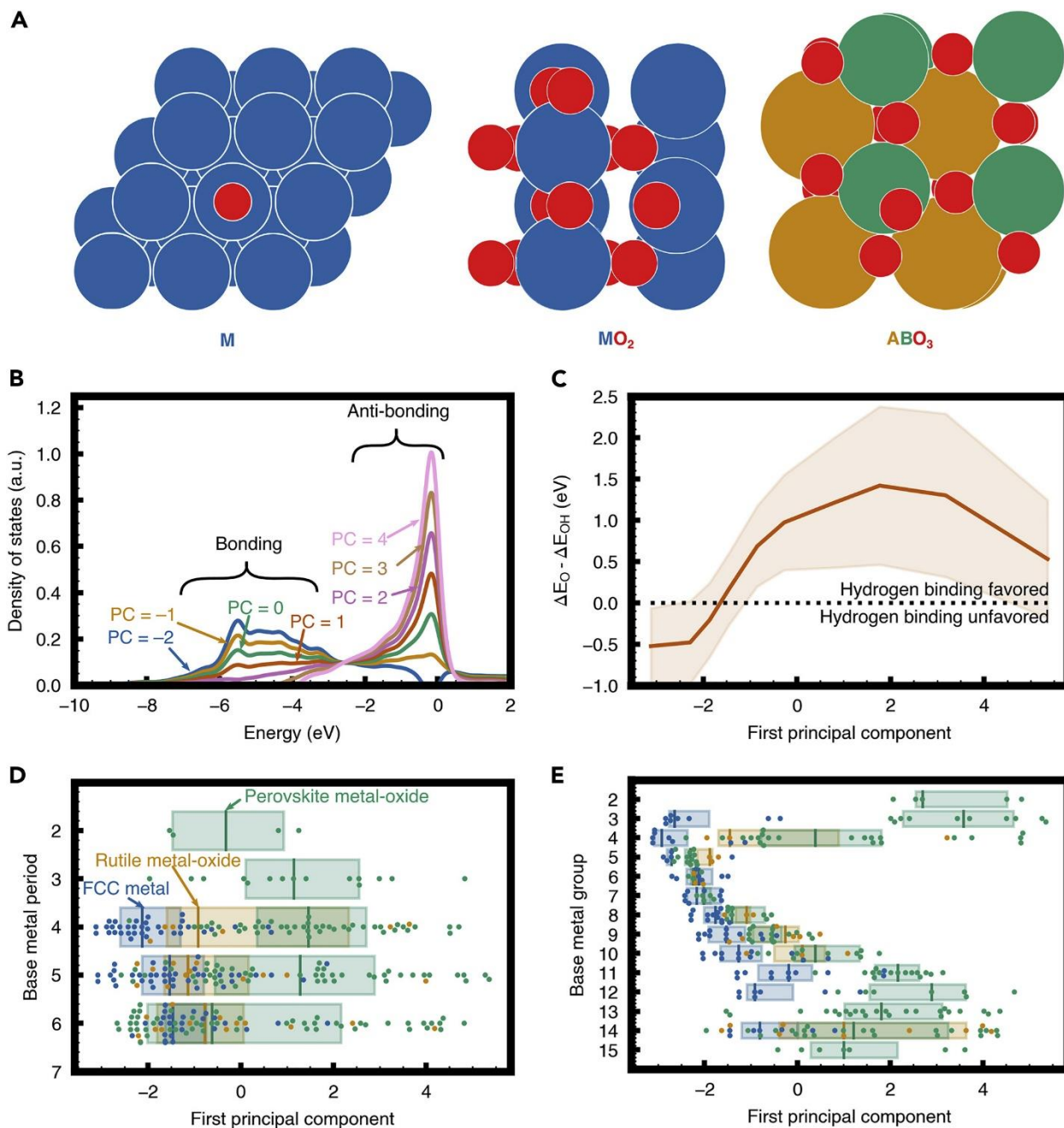


Figure 6. Electronic-structure descriptors from PCA for O reactivity on metals and metal oxides. (a) Representative structures making up the dataset of oxygen species on fcc metals, rutile metal oxides, and perovskite oxides from Dickens et al.⁶² (b) How the O 2*p* bonding and anti-bonding orbitals of surface oxygen change as a function of the first principal component score. (c) A partial dependence plot showing the effect of the first principal component score on surface oxygen reactivity. The shaded region denotes plus/minus one standard deviation of confidence around the mean. Box and swarm plots decompose the distributions of the first principal component score as a function of changing the base metal (d) period and (e) group.

Graphical abstract

