

Machine learning prediction of the electronic property of binary transition metal alloys

Yixuan Che,^{*,†} Junyi Zhao,^{*,†} and Hao Wang^{*,‡}

[†]*University of Science and Technology of China, 96 Jinzhai Road, Hefei 230026, China*

[‡]*Peking University, 5 Yiheyuan Road, Beijing 100871, China*

E-mail: zz0312@mail.ustc.edu.cn; junyi98@mail.ustc.edu.cn; tony.wanghao@stu.pku.edu.cn

Abstract

Machine learning methods have garnered much attention and use in computational catalysis. Previous studies have demonstrated rapid and accurate prediction of a variety of catalytic properties as well as the underlying potential energy landscapes. In particular, d-band center, defined as the first moment of the d-projected density of states, has been widely used as the key descriptor of activity trends for reactions catalyzed by metal surfaces. In this work, we construct a gradient boosting regression (GBR) model for prediction of the d-band center of bulk binary transition metal alloys. An accurate model is obtained using a dataset of over 1200 alloys from the Materials Project database spanning the entire d-block of the periodic table. The d-band centers, periodic groups, and relative compositions of the constituent metals are determined to have the highest feature importance scores, consistent with the underlying physics of the alloy. The regression model presented here offer a promising strategy of rapid property prediction with physical interpretability to aid the optimization and discovery of efficient heterogeneous catalysts.

Introduction

Over 90% of energy-intensive chemical production relies on heterogeneous catalysis. The ultimate goal in catalytic science is to accurately predict trends in catalytic activity using the electronic structure of metals, which will enable rational design of surfaces with specific catalytic properties without extensive trial-and-error experiments.¹ Previous studies have established a good correlation between catalytic activity and the d-band center.^{2,3} As such, computed d-band centers can be used to predict the activity trends of different metal surfaces.^{4,5} To model surfaces, first-principles calculations based on density functional theory (DFT) are employed. However, given the large computational cost and the static nature of the method, alternative data-driven approaches have the potential to accelerate the rational catalyst design efforts.

Machine learning (ML) methods are widely employed in materials science⁶ and engineering.^{7,8} Given sufficient training set and proper training method, ML models can ultimately bypass expensive DFT calculations while maintaining high accuracy. As such, they can be applied to prediction of various physical and chemical properties that remain unknown or too complex for analytical modeling. In computational catalysis, ML methods have been applied to screen the materials space in terms of structure and composition of bulk materials for desired catalytic properties. For example, Takigawa *et al.* used ML methods to predict the d-band center for eleven metals (Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Ir, Pt and Au) and their pairwise bimetals.¹ Andersen *et al.* employed SISSO (Sure Independence Screening and Scarifying Operator) to predict adsorption energies of atomic and molecular species on mono- and multi-metallic surfaces with root mean square errors within 0.2 eV.⁹ Surrogate models have also been used to accelerate the exploration of energy landscapes, such as the optimization of atomic structures. Hammer *et al.* employed supervised learning to enhance global optimization algorithms of molecular structures.¹⁰ Peterson *et al.* utilized convoluted neural networks to alleviate the cost of nudged elastic band calculations.¹¹ Overall, fast combinatorial materials screening and acceleration of traditional optimization algorithms are two

main trends of ML in computational catalysis.¹²

In transition metals, unoccupied d-orbitals can act as electrophiles, whereas occupied d-orbitals provide electrons as nucleophiles in chemical reactions to form intermediate surface products, thereby reducing the activation energy and promoting the reaction kinetics. In transition metal oxides, metal cations have strong redox properties due to the ease of losing or seizing the d-electrons. Both transition metal oxides and transition metals can be used as redox catalysts. In chemical industry, heterogeneous catalysts typically consist of a combination of a metal with other metals, nonmetals, or metalloids. Specific physical and chemical properties can be enhanced in multimetallic alloys compared to their monometallic counterparts. For example, alloy steel, which contain significant amounts of elements other than iron and carbon, has enhanced toughness, hardness, corrosion resistance, and ductility compared to the non-alloy steel. In catalysis, single-atom alloy catalysts have garnered much attention due to the promotion of the activity by a dilute amount of an active metal and the selectivity by an inert host metal.^{13,14}

In this work, we construct a regression model for prediction of the d-band center of binary transition metal alloys as a baseline for future research on the performance of alloy catalysts. Building on the previous study of eleven metals,¹ we use a dataset of over 1200 alloys from the Materials Project database spanning the entire d-block of the periodic table. We discover three most important descriptors of the constituent metals that determine the electronic property of the corresponding alloy: (1) the d-band center, (2) relative composition, and (3) periodic group. The proposed model offers a strategy for fast, accurate, and physically interpretable prediction of a key catalytically relevant property for binary alloys.

Methods and data

Dataset

For atomic descriptors, we use the data of Pauling electronegativity,^{15–17} d-band center (ε_d), the number of electrons, and the atomic radii, which can be obtained from the periodic table, the CRC handbook,¹⁸ and the Materials Project database¹⁹ (**Table 1**). The d-band center is calculated as the first moment of the d-projected density of states:

$$\varepsilon_d = \frac{\int_{-\infty}^{E_F} \varepsilon n_d(\varepsilon) d\varepsilon}{\int_{-\infty}^{E_F} n_d(\varepsilon) d\varepsilon}, \quad (1)$$

where E_F is the Fermi energy, ε is the energy, and $n_d(\varepsilon)$ is the density of states. The Pearson correlation coefficients for the descriptors (**Fig. 1**) are calculated using MATLAB.

Table 1: Input features (descriptors) of transition metals used for prediction, including atomic number (AN), period (P), group (G), Pauling electronegativity (EN), covalent radii (R_{cov}) in Å, and d-band center (ε_d) in eV. All d-band centers are obtained from the density of states available in the Materials Project database.¹⁹

Metal	AN	P	G	EN	R_{cov} (Å)	ε_d (eV)	Metal	AN	P	G	EN	R_{cov} (Å)	ε_d (eV)
Sc	21	4	3	1.36	1.59	1.59	Ru	44	5	8	2.20	1.36	−1.33
Ti	22	4	4	1.54	1.48	0.96	Rh	45	5	9	2.28	1.34	−1.82
V	23	4	5	1.63	1.44	1.15	Pd	46	5	10	2.20	1.30	−1.89
Cr	24	4	6	1.66	1.30	−0.10	Ag	47	5	11	1.93	1.36	−4.11
Mn	25	4	7	1.55	1.29	−0.59	Cd	48	5	12	1.69	1.40	−8.78
Fe	26	4	8	1.83	1.24	−1.65	Hf	72	6	4	1.30	1.64	0.98
Co	27	4	9	1.88	1.18	−2.11	Ta	73	6	5	1.50	1.58	1.55
Ni	28	4	10	1.91	1.17	−1.75	W	74	6	6	2.36	1.50	0.28
Cu	29	4	11	1.90	1.22	−2.48	Re	75	6	7	1.90	1.41	−0.09
Zn	30	4	12	1.65	1.20	−7.29	Os	76	6	8	2.20	1.36	−1.47
Zr	40	5	4	1.33	1.64	0.94	Pt	78	6	10	2.28	1.30	−2.27
Nb	41	5	5	1.60	1.56	1.00	Au	79	6	11	2.54	1.30	−3.53
Mo	42	5	6	2.16	1.46	0.10	Hg	80	6	12	2.00	1.32	−5.35
Tc	43	5	7	1.90	1.38	−0.22							

1211 binary alloys (A_mB_n) involving all 29 transition metal elements (Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au and Hg) were obtained from the Materials Project database (**Fig. 2a**).¹⁹ After imposing the

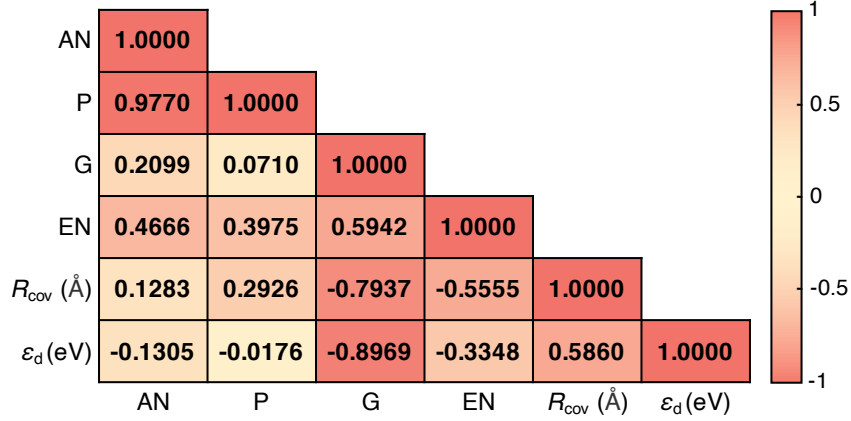


Figure 1: The correlation matrix of 6 descriptors for transition metals in **Table 1**. The values range from -1 to $+1$; the larger the magnitude, the redder the color of the block.

criterion of energy above hull (E_{hull}) per atom to be less than 10^{-6} eV, 580 thermodynamically stable alloys were obtained (**Fig. 2b**). Only 1004 and 504 alloys had DOS data available for calculation of the d-band center, without and with the stability criterion, respectively. For ternary alloys ($A_mB_nC_p$), we obtained 1396 and 386 without and with the restriction, respectively.

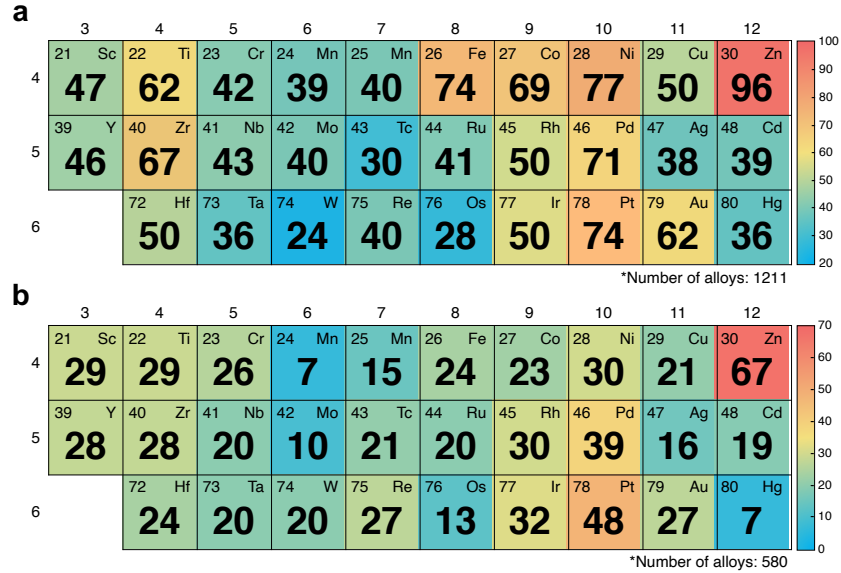


Figure 2: Number of binary alloys (A_mB_n), indicated at the center of each block as well as the color (blue to red), for each element as the majority host. If m and n are equal, both are counted as the host. (a) Total search results from the Materials Project database. (b) Search results after imposing stability criterion based on the energy above hull.

Machine learning method

We build our machine-learning models using the gradient boosting regression (GBR)²⁰ method as implemented in the scikit-learn package.²¹ GBR learns from its mistakes by integrating poor learning algorithms with small confidence interval and concentrated data. As an ensemble model of boosted regression trees, GBR can provide accurate and stable predictions.¹ The model parameters are set to `n_estimators = 1000`, `max_depth = 6`, `min_samples_split = 5`, and `learning_rate = 0.01`. The parameter `random_state` is set to 42 to maintain the same training and test set composition across all cases.

Results and discussion

Model prediction with different test/training split

Using datasets without and with the stability criterion (**Fig. 2a and b**), we first determine the optimal size of the training set. For a binary alloy AmBn, there are total 14 descriptors for prediction: six for A, six for B, and the relative composition of each element (x_m , x_n). We made predictions for the cases without and with the stability criterion, respectively. The parity plots of GBR-predicted vs. DFT-computed d-band centers with different test/training ratios (25/75, 50/50, and 75/25) are shown in **Fig. 3**. Without the stability criterion, we have total of 1004 data points, which provide test/training set sizes of 251/753, 502/502, and 753/251 for the three ratios, respectively. With the stability criterion, 504 data points

Table 2: The accuracy, mean squared error (MSE), and goodness of fit (R^2) of the GBR model on the test set, using the dataset (a) without and (b) with the stability criterion (**Fig. 2a and b**, respectively). The corresponding parity plots are shown in **Fig. 3**.

Case	(a) Full dataset			(b) Dataset with the stability criterion		
Test/Training	Model accuracy	MSE	R^2	Model accuracy	MSE	R^2
25/75	0.9778	0.0905	0.9780	0.9652	0.1360	0.9658
50/50	0.9667	0.1455	0.9671	0.9337	0.2447	0.9358
75/25	0.9515	0.2063	0.9516	0.9258	0.2953	0.9266

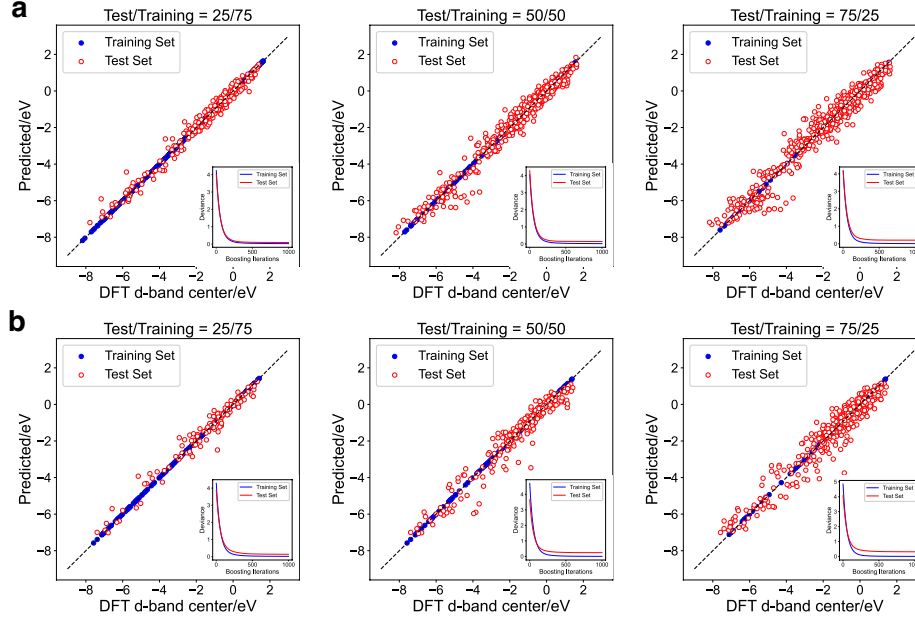


Figure 3: GBR vs. DFT-computed d-band center of binary alloys without the stability criterion using all 14 descriptors. The subplot is the convergence of the model accuracy over boosting iterations. Blue and red points represent training and test set, respectively. The best performance is achieved with test/training split of 25/75. (a) Full dataset. (b) Dataset with the stability criterion.

are available, providing the respective sizes of 126/378, 202/202, and 378/126. The model accuracy, mean squared error (MSE), and goodness of fit (R^2) on the test set are listed in **Table 2**. The larger the training set, the more accurate the model prediction. As such, all following analyses are performed using the test/training ratio of 25/75.

Descriptor and permutation importance for GBR prediction

Next, we investigate the importance of the 14 descriptors listed in **Table 1**. GBR is based on "boosting" technique, which adaptively combines a large number of relatively simple regression-tree models that recursively partition the data using a single selected descriptor. From this process, feature-importance score can be obtained as a weighted average of the number of times that a given descriptor is selected for the partitioning. The statistical importance of descriptors may vary with the ML method used.

The feature-importance scores of all 14 descriptors are shown in **Fig. 4**. Without the

stability criterion, the top three sets of descriptors from mean decrease in impurity (MDI) are: (1) d-band center, (2) relative composition, and (3) periodic group of the two constituent metals. With the stability criterion, d-band center remains the most important descriptor, but the order of the other two descriptors is flipped. The variation of a given feature importance between the host and the guest metal is rather small. The top descriptors remain the same from permutation importance. Overall, the observed trend is consistent with the underlying physics: the alloy d-band center is determined by the energetic proximity of the d-band centers of individual components. These values are influenced by the number of electrons determined by the periodic group, i.e. the more to the right of the periodic table an element is located, the more electrons it has, and the lower the d-band center becomes.

Model prediction with different numbers of descriptors

To determine the optimal number of descriptors for GBR prediction, we train different models using top 6, 4, and 2 descriptors from **Fig. 4** with test/training ratio of 25/75 and compare against the model with all 14 descriptors. The three-performance metrics for each model are listed in **Table 3**, and the corresponding parity plots are shown in **Fig. 5**. Using the full dataset, the model accuracy decreases from 0.9778 with all 14 descriptors to 0.9761, 0.9722, and 0.7646 with the top 6, 4, and 2 descriptors, respectively. The model accuracy remains

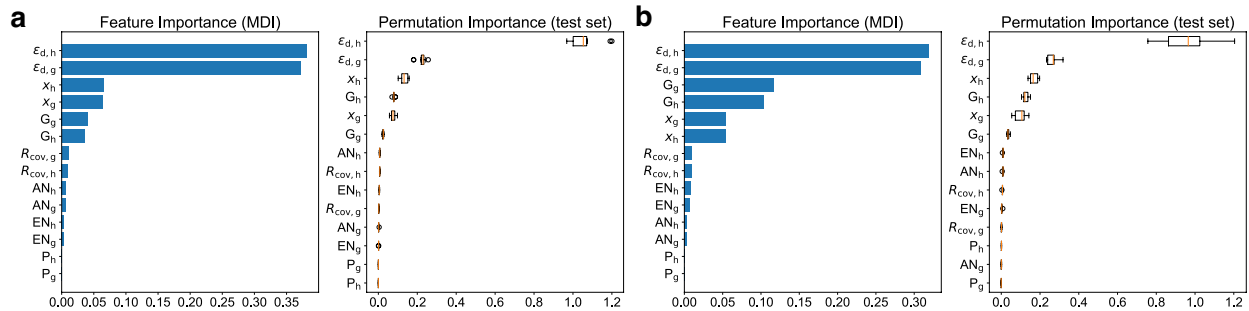


Figure 4: Feature importance from mean decrease in impurity (MDI) and the permutation importance of the descriptors (**Table 1**, including the relative composition x of each component) on the test set, using the full dataset (**Fig. 2a**) with the test/training split of 25/75 (**Table 2a**). The subscripts "h" and "g" refer to the host and guest metal, respectively.

high (~ 0.97) down to the top 4 descriptors but suffers when only the top 2 descriptors are used. Using the smaller dataset with the stability criterion, the prediction performance remains high down to the top 6 descriptors.

Table 3: The accuracy, mean squared error (MSE), and goodness of fit (R^2) of the GBR model on the test set, using the dataset (a) without and (b) with the stability criterion (**Fig. 2a** and **b**, respectively). The corresponding parity plots are shown in **Fig. 5**.

Case	(a) Full dataset			(b) Dataset with the stability criterion		
Number of descriptors	Model accuracy	MSE	R^2	Model accuracy	MSE	R^2
All (14)	0.9778	0.0905	0.9780	0.9652	0.1360	0.9658
Top 6	0.9761	0.0972	0.9764	0.9702	0.1166	0.9705
Top 4	0.9722	0.1130	0.9725	0.8132	0.7306	0.8153
Top 2	0.7646	0.9580	0.7723	0.7993	0.7849	0.8040

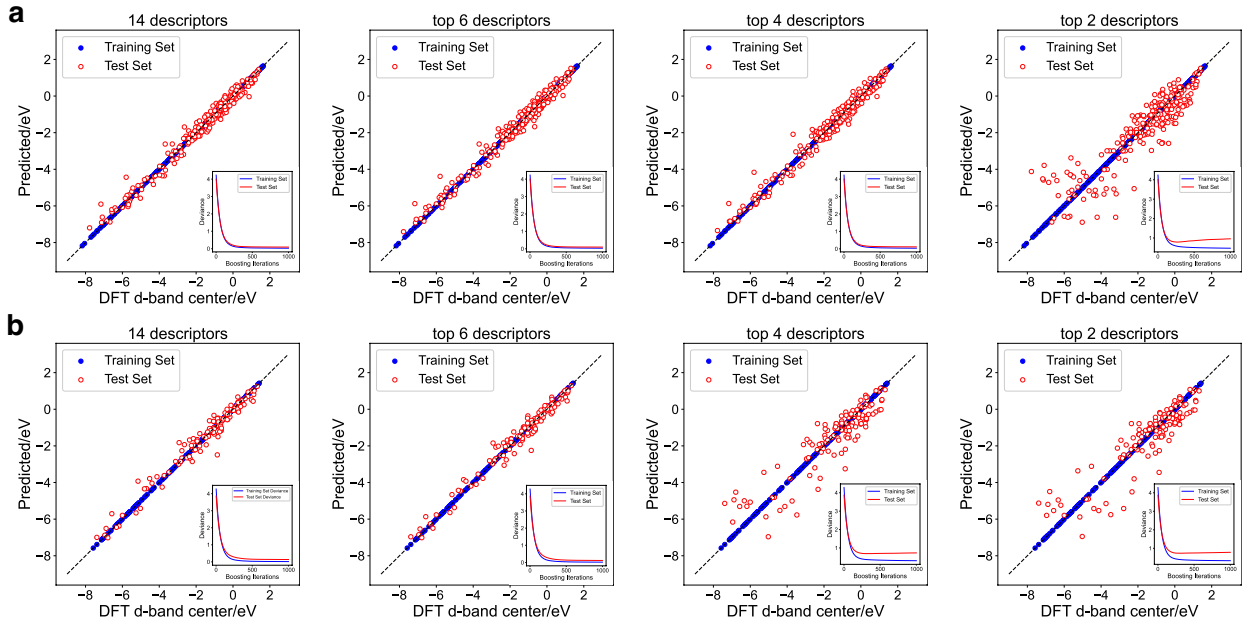


Figure 5: GBR vs. DFT-computed d-band center of binary alloys without the stability criterion using all 14, top 6, top 4, and top 2 descriptors. The subplot is the convergence of the model accuracy over boosting iterations. Blue and red points represent training and test set, respectively. (a) Full dataset. (b) Dataset with the stability criterion.

Conclusion

We propose a fast, accurate, and physically interpretable gradient boosting regression (GBR) model for prediction of the d-band center of bulk binary transition metal alloys based on three sets of readily available descriptors – d-band center, relative composition, and periodic group. We use a dataset of over 1200 alloys from the Materials Project database spanning all 29 transition metals, with and without the stability criterion of energy above hull; the full dataset exhibits better prediction performance. The method presented here can also be applied to catalytically relevant properties of surface alloys, which include not only the d-band center, but also other structural and chemical properties such as lattice constant, miscibility, and short- and long-range order parameters. Other more data-intensive methods such as neural networks and reinforcement learning can also be considered for larger datasets involving other types of materials beyond alloys.

Acknowledgement

This project was supported by Touch Education Technology, Inc. We acknowledge scientific and editorial support from the Project Lead, J. S. Lim, of Harvard University; technical support from the Project Support, A. Louis of Panthéon-Sorbonne University; and administrative support from M. Clarke and X. Wang of Touch Education Technology, Inc.

Author contributions

All three authors contributed equally to this work. J.Z. and H.W. conducted literature reviews. The coding and visualization were performed by Y.C. for the d-band center calculation and correlation matrix; by H.W. for the regression and validation; and by J.Z. for feature importance. All authors contributed to manuscript preparation.

Competing Financial Interests

The authors declare no competing financial interests.

References

- (1) Takigawa, I.; Shimizu, K.; Tsuda, K.; Takakusagi, S. Machine-learning prediction of the d-band center for metals and bimetals. *RSC advances* **2016**, *6*, 52587–52595.
- (2) Acerbi, N.; Tsang, S. E.; Jones, G.; Golunski, S.; Collier, P. Rationalization of interactions in precious metal/ceria catalysts using the d-band center model. *Angewandte Chemie* **2013**, *125*, 7891–7895.
- (3) others,, et al. Correlation between the surface electronic structure and CO-oxidation activity of Pt alloys. *Physical Chemistry Chemical Physics* **2015**, *17*, 4879–4887.
- (4) Hammer, B.; Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surface science* **1995**, *343*, 211–220.
- (5) Hammer, B.; Nørskov, J. K. Why gold is the noblest of all the metals. *Nature* **1995**, *376*, 238–240.
- (6) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
- (7) Tsai, C. W.; Lai, C. F.; Chiang, M. C.; Yang, L. T. Data mining for internet of things: A survey. *IEEE Communications Surveys & Tutorials* **2013**, *16*, 77–97.
- (8) Liu, H.; Xu, C.; Liang, J. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews* **2017**, *21*, 171–193.
- (9) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catalysis* **2019**, *9*, 2752–2759.

- (10) Jørgensen, M. S.; Larsen, U. F.; Jacobsen, K. W.; Hammer, B. Exploration versus exploitation in global atomistic structure optimization. *The Journal of Physical Chemistry A* **2018**, *122*, 1504–1509.
- (11) Peterson, A. A. Acceleration of saddle-point searches with machine learning. *The Journal of chemical physics* **2016**, *145*, 074106.
- (12) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine learning for computational heterogeneous catalysis. *ChemCatChem* **2019**, *11*, 3581–3601.
- (13) Luneau, M.; Lim, J. S.; Patel, D. A.; Sykes, E. C. H.; Friend, C. M.; Sautet, P. Guidelines to achieving high selectivity for the hydrogenation of α , β -unsaturated aldehydes with bimetallic and dilute alloy catalysts: A review. *Chemical reviews* **2020**, *120*, 12834–12872.
- (14) Hannagan, R. T.; Giannakakis, G.; Flytzani-Stephanopoulos, M.; Sykes, E. C. H. Single-atom alloy catalysis. *Chemical Reviews* **2020**, *120*, 12044–12088.
- (15) Pauling, L. THE NATURE OF THE CHEMICAL BOND. IV. THE ENERGY OF SINGLE BONDS AND THE RELATIVE ELECTRONEGATIVITY OF ATOMS. *Journal of the American Chemical Society* **1932**, *54*, 3570–3582.
- (16) Allred, A. L. Electronegativity values from thermochemical data. *Journal of Inorganic and Nuclear Chemistry* **1961**, *17*, 215–221.
- (17) Speight, J. G. *Lange’s handbook of chemistry*; McGraw-Hill Education, 2017.
- (18) Rumble, J. R. CRC Handbook of Chemistry and Physics (102nd edition). *CRC Press* **2021**,

- (19) others,, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling* **2020**, 1751–1784.
- (20) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, 1189–1232.
- (21) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

Graphical TOC Entry

