Learning design rules for selective oxidation catalysts from high-throughput experimentation and artificial intelligence

Lucas Foppa,[a,b*] Christopher Sutton,[a†] Luca M. Ghiringhelli,[a,b] Sandip De,[c*] Patricia Löser,[d] Stephan A. Schunk,[c,d] Ansgar Schäfer,[c] and Matthias Scheffler [a,b]

[a]The NOMAD Laboratory at the Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany. [b]The NOMAD Laboratory at the Humboldt-Universität zu Berlin, Zum Großen Windkanal 6, D-12489 Berlin, Germany, [c]BASF SE, Carl-Bosch-Straße 38, D-67065 Ludwigshafen, Germany. [d]hte GmbH, Kurpfalzring 104, D-69123, Heidelberg, Germany.

Abstract: The design of heterogeneous catalysts is challenged by the complexity of materials and processes that govern reactivity and by the fact that the number of good catalysts is very small compared to the number of possible materials. Here, we show how the subgroup-discovery (SGD) artificial-intelligence approach can be applied to an experimental plus theoretical data set to identify constraints on key physicochemical parameters, the so-called SG *rules*, which exclusively describe materials and reaction conditions with outstanding catalytic performance. By using high-throughput experimentation, 120 SiO_2-supported catalysts containing ruthenium, tungsten and phosphorus were synthesized and tested in the catalytic oxidation of propylene. As candidate descriptive parameters, the temperature and ten parameters related to the composition and chemical nature of the catalyst materials, derived from calculated free-atom properties, were offered. The temperature, the phosphorus content, and the composition-weighted electronegativity are identified as key parameters describing high yields towards the value-added oxygenate products acrolein and acrylic acid. The SG rules not only reflect the underlying processes particularly associated to high performance but also guide the design of more complex catalysts containing up to five elements in their composition.

Keywords: artificial intelligence, subgroup discovery, high-throughput experimentation, selective oxidation.

Introduction

Heterogeneous catalysis is governed by an intricate interplay of multiple processes[1] such as the surface reaction networks and the typically unknown dynamic restructuring of the catalyst material under reaction conditions. Thus, the design of new materials is challenging. While theoretical approaches attempt to address the complexity of heterogeneous catalysis,[2] the explicit atomistic modelling of the full catalytic progression by first-principles methods is impractical. Another approach for identifying novel catalysts consists on the use of high-throughput experimentation (HTE) to test large amounts of materials.[3] However, utilizing the information obtained by the experiments to decide on the next promising materials to investigate is not straightforward.[4] As the number of possible materials is practically infinite and the number of good catalysts is very small, the direct approach is unlikely to identify the needed catalyst material.

Firstly, when large libraries of materials are tested, the detailed characterization of each material is typically not feasible. Thus, only little information on the structure and physicochemical properties of each compound might be available. This hinders the in-depth understanding of the underlying processes governing reactivity, which could be used for rational catalyst design. Secondly, distinct catalytic mechanisms might operate depending on the materials and reaction conditions, and only very few situations result in good catalytic performance. This leads to an unbalanced distribution between high- and low-performance scenarios and brings into question the usefulness of *global* models to help deciding on the next materials to be tested. These models are trained to describe all materials and reaction conditions simultaneously by minimizing the expected average prediction error over all samples. While this approach may provide an accurate prediction on average, it does not necessarily allow for a focused modelling of the most interesting materials and mechanisms. Alternative approaches for catalyst design are therefore required.

Here, we apply the subgroup-discovery (SGD) artificial-intelligence local approach[5] to a hybrid data set obtained from HTE & theory to identify key physicochemical descriptive parameters and constraints on their values, i.e., rules, which are particularly associated with high performance. The reactivity measured by HTE is used as target in the SGD analysis. The temperature and composition-dependent physicochemical properties evaluated with density-functional-theory (DFT) calculations are used as candidate descriptive parameters.

The SGD approach has been applied in computational catalysis[6] and materials science.[5e, 7] It starts with the generation of a pool of propositions (π), statements about the data that apply only to a portion of the data set. For the case of continuous candidate descriptive parameters, the propositions are inequalities constraining their values. Then, SGD identifies selectors (σ), i.e., statements formed by a number of propositions and the "AND" connector (denoted "$\wedge$"), that result in the selection of subgroups of materials and conditions with the most outstanding distributions of the target values with respect to the whole data set (Fig. 1A). The propositions entering these selectors can be seen as rules describing the exceptional SG behavior. The parameters entering these propositions are, in turn, the key, most relevant, descriptive parameters, out of all the offered ones, associated with the desired reactivity. Because the SG search is performed by maximizing a quality function that measures how outstanding specific subselections of data points are, this approach identifies a local behavior.

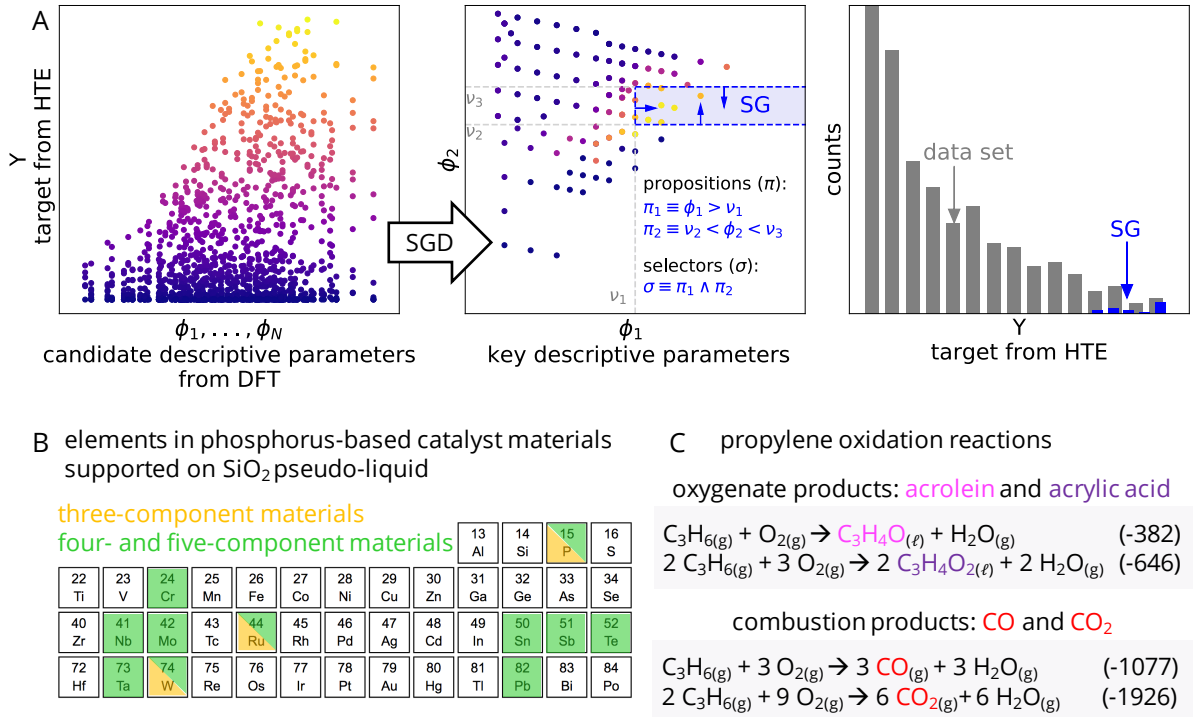

Figure 1. A: SGD approach for identifying key descriptive parameters and rules associated to materials and reaction conditions with outstanding catalytic performance. The rules are given by the propositions and consist on constraints on the values of key descriptive parameters. "∧" denotes the "AND" connector. B: Elements entering the composition of the SiO_2-supported materials. C: Competing reactions in propylene oxidation lead to the desired oxygenates but also to the combustion by-products. The values shown in parenthesis in C are the reaction enthalpies, in kJ/mol.[8]

Thus, the identified rules reflect the specific underlying processes resulting in outstanding performance.

We apply the SGD-HTE-&-theory approach to the selective oxidation of propylene on SiO_2-supported catalysts based on ruthenium, tungsten, and phosphorus. By using the product yield measured by HTE as target, we circumvent the need for the explicit modelling of the full catalytic progression. Additionally, because the candidate descriptive materials parameters can be calculated by first-principles methods, extensive materials characterization by experiment is not required and the resulting SG rules can be used to identify promising catalyst candidates which were not yet synthesized by experiment.

Selective oxidation reaction and high-throughput-experimentation

The selective partial oxidation of light alkanes to value-added olefins or oxygenates is an efficient technology for feedstock upgrading.[9] However, the intricate surface reaction networks[10] typically lead to product mixtures containing up to 20 different molecules, including undesirable by-products such as CO and CO_2. In order to selectively produce the olefins or the oxygenates, mixed-metal oxide or phosphate heterogeneous catalysts based on molybdenum and vanadium redox-active species have been used, such as the $MoVTeNbO_x$ and the state-of-the-art industrial catalyst for n-butane selective oxidation, vanadyl pyrophosphate. Several recent investigations have explored the catalytic activity of mixed-metal phosphates in a systematic way.[11]

In this study, we investigate materials based on ruthenium combined with tungsten and phosphorus (Fig. 1B) as an alternative class of catalysts for selective oxidation. Platinum-group-metal-based catalysts commonly result in hydrocarbon combustion products (Fig. 1C). The combination of these metals with tungsten and phosphorus, in a tungsten-phosphate-like matrix, however, could favor the selectivity towards the desired olefins and oxygenates, following a catalyst design strategy based on the dilution of highly active metal sites. With the aim of studying these systems, HTE measurements were performed using 120 different *three-component* catalyst compositions containing ruthenium, tungsten and phosphorus in different proportions. At each catalyst composition, several reaction temperatures between 200ºC and 400ºC were examined. The detailed preparation, characterization and reactivity analysis of these catalysts in the selective oxidation of n-butane, propane and propylene is discussed in a separate contribution.[12] In this paper, we only provide details on the propylene selective oxidation reaction.

All the reactions were carried out in tubular, fixed-bed reactors with the following reaction feed: Ar, H_2O, N_2, O_2 and propylene (C_3H_6) with molar rates 4.015, 4.015, 104.40, 20.08, and 1.57 mmol/h, respectively. The same mass of catalyst was used in all reactions, so that the contact time, in terms of volumetric flow per mass of catalyst, was kept fixed across experiments. These three-component catalysts are prepared on a SiO_2 pseudo-liquid support and might present a disordered, possibly amorphous, structure. The atomic structures of all the tested catalysts are not known in detail. However, similar catalytic performance was found for crystalline and disordered phases at the same composition.[12] This indicates that the composition is more crucial for the catalytic performance than the degree of crystallinity.

In HTE, a large materials space is accessible for catalyst design by changing the relative amount of each component and the specific elements on the catalyst composition. Approaches to guide the efficient exploration of such materials space, indicating the most promising compositions to be tested next, are thus desirable. The

most interesting compositions are those that display both considerable activity, i.e., those providing significant propylene conversion, and selectivity, i.e., those that specifically form the desired oxygenates (acrolein and acrylic acid, Fig. 1C) from propylene. This is motivated by using the yield of oxygenates $Y_{\text{oxygenates}}$ as target in our SGD analysis, defined as

$$Y_{\text{oxygenates}} = Y_{\text{acrolein}} + Y_{\text{acrylic acid}} = \frac{\dot{F}_{\text{acrolein,out}}}{\dot{F}_{\text{propylene,in}}} + \frac{\dot{F}_{\text{acrylic acid,out}}}{\dot{F}_{\text{propylene,in}}}. \qquad (1)$$

In Eq. 1, $\dot{F}_{A,\text{in}}$ and $\dot{F}_{A,\text{out}}$ denote the molar rate, in mmol/h, of species A in the reactor feed and outlet, respectively. Our goal is to identify key parameters and rules describing materials and reaction conditions that give high yields of oxygenates.

Subgroup discovery approach

The two main crucial aspects in SGD are the offered candidate descriptive parameters and the quality function. In this work, we use the reaction temperature (T) and the phosphorus molar content (x_{P}) as experimental candidate parameters. In addition, we include a set of free-atom properties as candidate descriptive parameters to characterize the catalyst material in terms of the proportion and chemical nature of the elements entering the composition. The following elemental properties are used:

- the radius of maximum electron density of *s*, *p*, *d,* and *valence* orbitals (r_s, r_p, r_d, and r_{val}, respectively);
- the Kohn-Sham single-particle eigenvalue of the highest occupied and lowest unoccupied states (ϵ_H and ϵ_L, respectively);
- electron affinity (EA);
- ionization potential (IP);
- electronegativity (EN), defined as $EN = \frac{EA+IP}{2}$.

These properties were calculated for the isolated atoms using DFT-PBEsol[13] and the FHI-aims[14] code (further calculation details and values for the elemental properties used in the work available in the electronic supporting information, ESI). r_{val} is defined as the radius of the highest-occupied state. For a given catalyst composition, the per-element free-atom properties are converted into system-specific properties by taking the composition-weighted average:

$$\overline{\varphi_a} = \sum^{M} \varphi_{a,i} x_i, \qquad (2)$$

where φ_a is an arbitrary elemental property, x_i is the molar content of element i in the material, and i runs over all M elements in the composition. For the three-component materials, $M = 3$. We note that oxygen is also present in all materials, but its proportion is not known from the catalysts formulation nor measured for all materials. Therefore, the oxygen content is not included in the material's characterization. Properties that can be readily calculated for the free atom are advantageous to structure-based properties because they do not have to be re-evaluated for each new material. In total, 11 descriptive parameters are used in our SGD analysis: T, x_{P}, $\overline{r_s}$, $\overline{r_p}$, $\overline{r_d}$, $\overline{r_{val}}$, $\overline{\epsilon_H}$, $\overline{\epsilon_L}$, $\overline{EA}$, $\overline{IP}$, and $\overline{EN}$.

As SGD quality function, we use

$$Q(P,SG) = \frac{s(SG)}{s(P)} * D_{\text{cJS}}(P,SG), \qquad (3)$$

where the *coverage* $\frac{s(SG)}{s(P)}$ is the ratio between the number of data points in the subgroup, $s(SG)$, and the total number of data points in the whole data set, $s(P)$, and $D_{\text{cJS}}(P,SG)$ is the cumulative Janson-Shannon divergence between the distribution of the target values in the SG and the distribution of the target values in the whole data set.[15] The coverage term controls the subgroup size and prevents that very small SGs with little statistical significance are selected. The second term, D_{cJS}, is the cumulative-distribution-function formulation[15] of the Jensen-Shanon divergence, which is a properly symmetrized version of the information-theoretic Kulback-Leibler divergence. D_{cJS} measures the dissimilarity between two distributions: It assumes close-to-zero values for similar distributions and increases, for instance, as the distributions have different standard deviations or different means. Thus, the second term in Eq. 3 favors the identification of SGs presenting target values as "unusual" as possible compared to the majority of observations. It also favors distributions which are contained in narrower value ranges compared to the whole data set. When most of the data points at hand contain low-performing materials and conditions, the use of D_{cJS} in the quality function allows focusing on the exceptional high-performing materials. Further SGD details, including the detailed description of the approach and of the Jensen-Shanon divergence are available in ESI.

Subgroup of three-component catalysts with exceptional performance

The propylene conversion vs. oxygenates selectivity profiles and the distribution of yield of oxygenates in the dataset (Fig. 2A and B, respectively) show that the vast majority of observations correspond to low performance. Indeed, 50% of the measured materials and conditions result in less than 2% oxygenates yield and only 41 measurements, out of 1220, are associated to yields of oxygenates above 20%. The average oxygenates yield over the whole data set is equal to 4.83% and the maximum $Y_{\text{oxygenates}}$ value is 26.85 %.

By applying the SGD, we identified several SGs providing near-optimal quality-function values (Fig. S3). Among the SGs displaying quality-function values within 40% of the optimal value (see Fig. S3), we selected, for further analysis and discussion, the SG that presents the highest value of cumulative Janson-Shannon divergence (0.693). This SG contains only 15 data points, i.e., approximately 1.2 % of the data set, which all have high yield of oxygenates (Fig. 2A and B, in black). The average yield of oxygenates in this SG is equal to 24.15%, i.e., five times higher than the average on the whole data set. This SG is described by rules on three descriptive parameters: $280 \leq T \leq 300$°C, $x_{\text{P}} > 0.55$, and $3.910 \leq \overline{EN} < 4.002$ eV (Fig. 2C). The rule on the temperature highlights that the highest yields of oxygenates are observed for intermediate temperatures within the tested range of 200-400°C (Fig. 2C). The rule on the phosphorus content shows that a relatively high phosphorus content is needed to achieve outstanding performance (Fig. 2D). This could be related to the dilution of metal active sites on a phosphate matrix that occurs at high phosphorus loadings.[12] Finally, the rule on the composition-averaged electronegativity (Fig. 2F) effectively limits the range of Ru contents, as shown in the ternary diagram of Fig. 3B. Indeed, Ru is needed to achieve propylene conversion (Fig. S5A) but too much of this element in the composition leads to undesired combustion products (Fig. S5B).

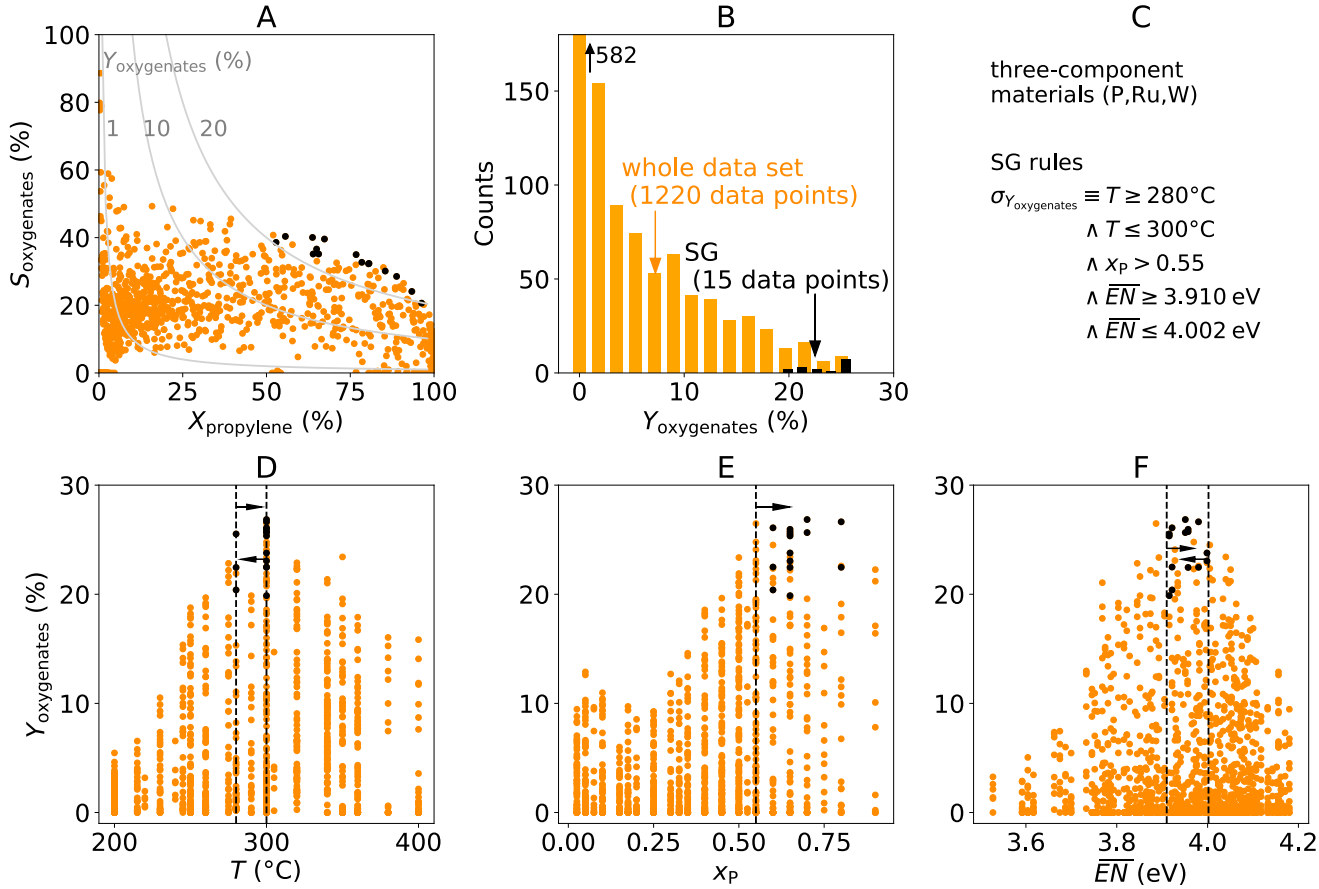

Figure 2. SGD analysis of propylene selective oxidation using three-component materials with ruthenium, tungsten, and phosphorus. A: Overview of reactivity measured by HTE. B: Distribution of oxygenates yield over the data set of 1220 measurements. C: Identified rules describing the SG. D, E, and F: SG rules (indicated by the black dashed lines and arrows) on the identified key descriptive parameters: temperature (T), phosphorus molar content (x_P), and composition-averaged electronegativity ($\overline{EN}$), respectively. The data points corresponding to the identified SG are displayed in black. The propylene conversion and the selectivity towards oxygenates are defined by $X_{propylene} = \frac{\dot{F}_{propylene,in} - \dot{F}_{propylene,out}}{\dot{F}_{propylene,in}}$ and $S_{oxygenates} = \frac{\dot{F}_{acrolein,out} + \dot{F}_{acrylic\ acid,out}}{\dot{F}_{propylene,in} - \dot{F}_{propylene,out}}$, respectively.

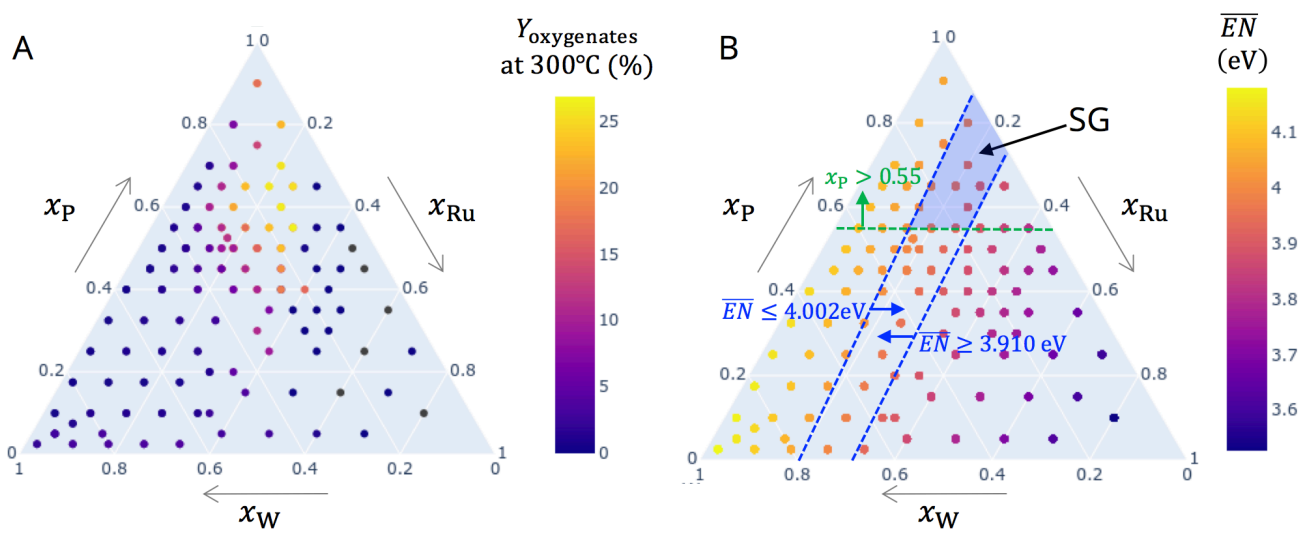

Figure 3. Ternary diagrams for three-component materials with ruthenium, tungsten, and phosphorus tested in propylene selective oxidation using HTE. A: Measured yield of oxygenates at 300° C. B: composition-averaged electronegativity $\overline{EN}$ for each tested composition. The SG rules are shown by the dashed lines and arrows in B and the portion of the ternary diagram selected by the SG rules is shown in blue.

Similar SG rules are obtained when the training is performed with randomly-selected 90% of the data points (see cross-validation study in ESI) or when the data points presenting yield of oxygenates lower than 3% are excluded from training (see details in ESI). SG rules constraining the $\overline{EN}$ parameter to an intermediate range, for instance, are always observed when only 90% of the data is used for training. Furthermore, the ranges of variation of minimum and maximum thresholds are [3.882, 3.910 eV] and [3.989, 4.031], respectively, i.e., similar to the thresholds shown in Fig. 2F. These results indicate that the SG rules are not strongly affected by variations on the data used for their derivation.

Overall, these results demonstrate the ability of the HTE & theory SGD approach to detect interpretable, chemically meaningful, and complex patterns associated to very few data points with exceptional catalytic performance.

Exploiting the subgroup rules for the design of four- and five-component catalysts

Using the rules defining the SG of outstanding oxygenate production for the three-component data, we designed more complex materials containing additional elements. We start by considering four-component materials containing ruthenium, tungsten, phosphorus, and one additional E_1 element. For this analysis, we fix the phosphorus content to 0.60 according to the rule identified in Fig. 2C. To further reduce the number of possible variables determining the catalyst composition, we also fix the ruthenium molar content to 0.05. We focus on such relatively low ruthenium loadings to ensure that the formation of combustion products is not significant. In this way, the compositions of the four-component materials are determined solely by the choice of E_1 element and its molar content. Materials with E_1 molar content of 0.35, which do not contain tungsten and are thus composed by three elements, are also referred to as four-component materials in our analysis to highlight that they contain different chemical elements compared to ruthenium, tungsten and phosphorus, the elements used to derive the rules. We concentrate on E_1 elements that show octahedral coordination patterns among reported phosphorus-containing materials structures[16] and that have a maximum atomic radius difference compared to tungsten of 0.10 Å (see details in ESI). This is to ensure that only elements that are compatible with tungsten, i.e., that could possibly replace tungsten in the material structure, are taken into account. The following E_1 are considered: niobium, tantalum, chromium, molybdenum, tin, antimony, and tellurium. These elements have atomic radii of 1.45, 1.45, 1.40, 1.45, 1.45, 1.45, and 1.40 Å, respectively. The atomic radius of tungsten is 1.45 Å. We have also included lead in this analysis, since materials containing this element were also experimentally tested (see below).

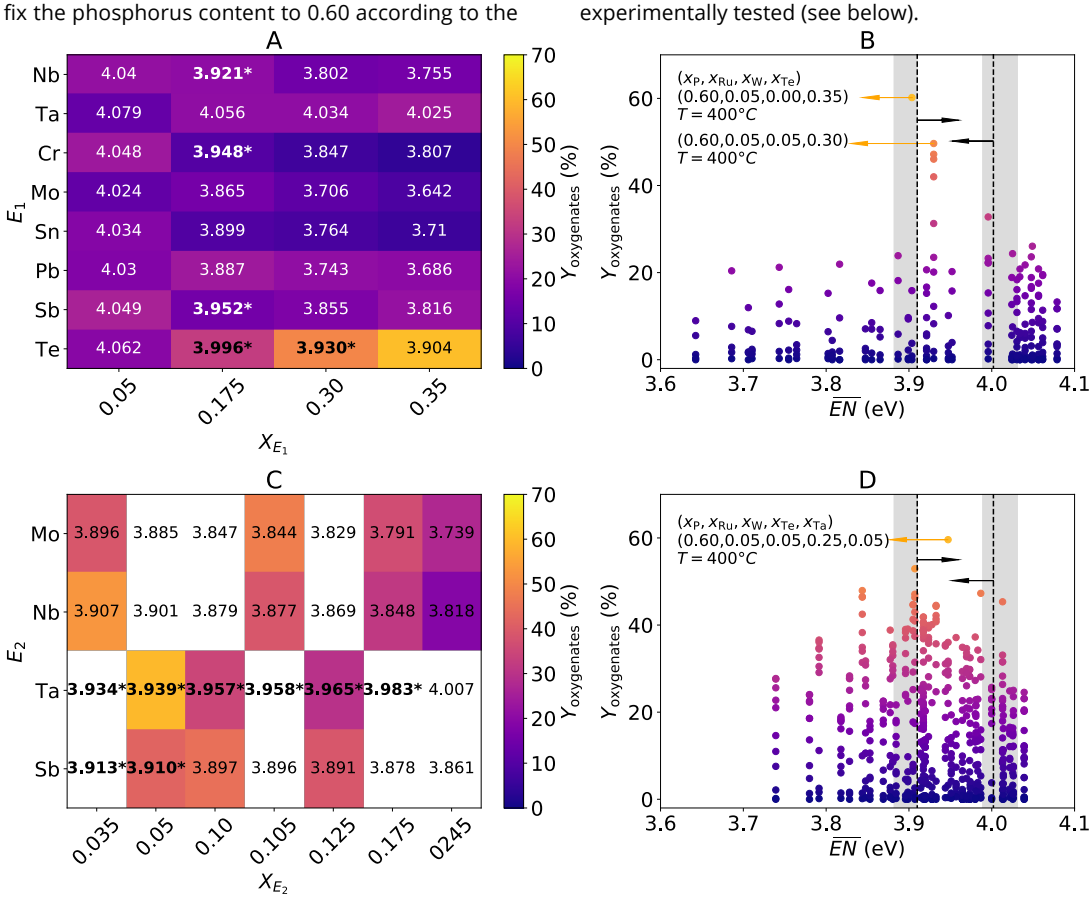

Figure 4. SG rules applied to the design of four- and five-component materials for propylene selective oxidation. A and C: Composition-averaged electronegativity ($\overline{EN}$) for different elements and molar contents in four- and five-component materials, respectively. B and D: Distribution of all measured yield of oxygenates (214 and 533 data points) for four- and five-component materials, respectively. In A and C, the $\overline{EN}$ values are shown in bold and are marked with starts if they satisfy the SG rules on $\overline{EN}$ identified based on the three-component materials. The colors in A and C indicate the highest measured yield of oxygenates for each material. The SG rules identified based on the three-component materials are indicated by the dashed lines and arrows in B and D. The shaded areas in B and D indicate the variability of $\overline{EN}$ thresholds observed when using only 90% of the data set for training (Table S2). The oxygenate yields shown in C correspond to materials with $x_W = 0.035$ for the cases $E_2 = \mathrm{Mo, Nb}$, and with $x_W = 0.050$ for the cases $E_2 = \mathrm{Ta, Sb}$. The white cells in C indicate materials not measured by HTE.

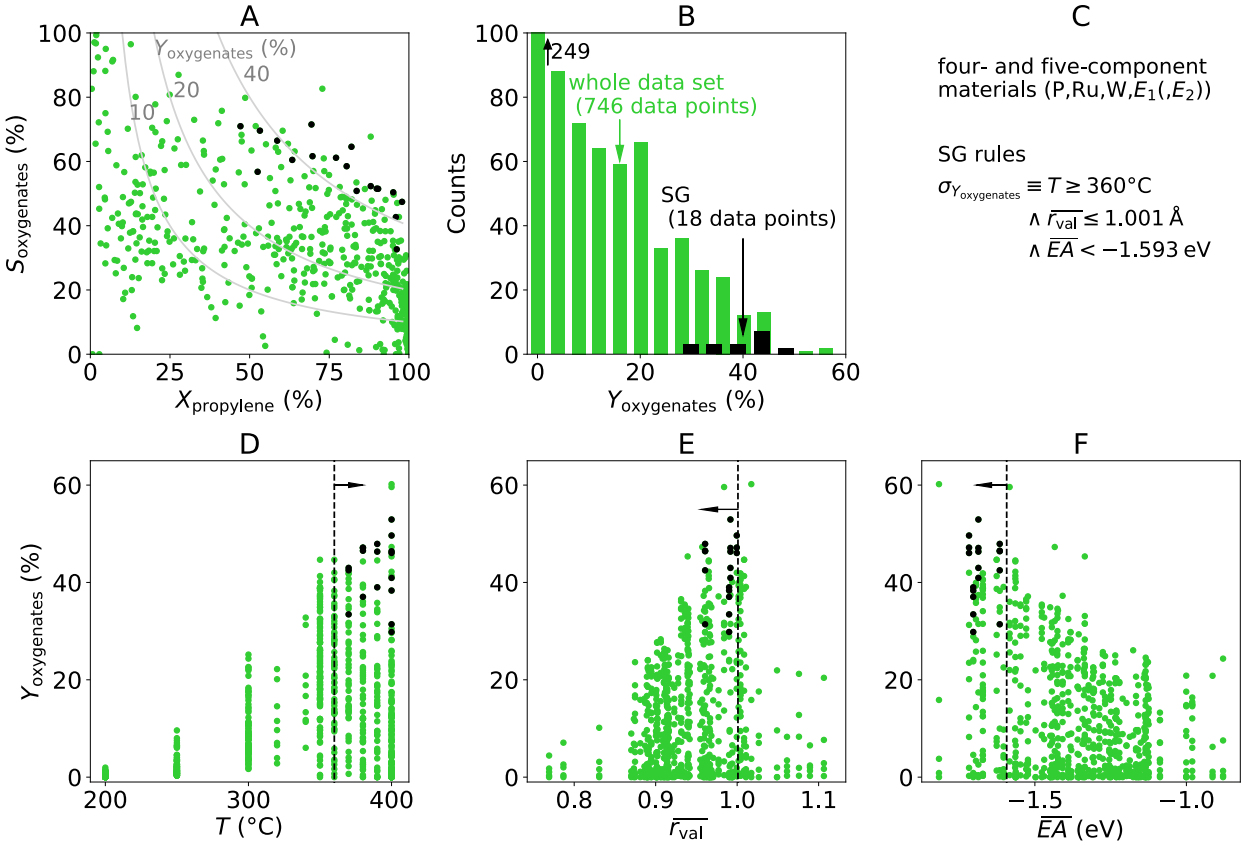

Figure 5. SGD analysis of propylene selective oxidation on four and five-component materials. A: Overview of reactivity measured by HTE. B: Distribution of oxygenates yield over the data set of 746 measurements. C: Identified rules describing the SG. D, E, and F: SG rules (indicated by the black dashed lines and arrows) on the identified key descriptive parameter: temperature (T), composition-averaged valence radius ($\overline{r_{val}}$), and electron affinity ($\overline{EA}$), respectively. The data points corresponding to the SG are displayed in black.

We evaluated the composition-averaged electronegativity ($\overline{EN}$) for the selected E_1 and the molar contents 0.05, 0.175, 0.30 and 0.35 in Fig. 4A. In this figure, the $\overline{EN}$ values for the new four-component materials are shown in bold and marked with starts if they satisfy the SG rule $3.910 \leq \overline{EN} < 4.002$ eV. This *catalyst map* suggests that the use of 5.0 to 17.5 % (molar) of niobium, chromium, molybdenum, tin, lead and antimony, in the catalyst composition in addition to ruthenium, tungsten and phosphorus, results in catalysts which are part of the identified SG, and thus likely high-performant materials. For the case of tantalum and tellurium, 17.5 % (molar) or more of these elements is needed for the resulting materials to present the $\overline{EN}$ values compatible with the SG.

The four-component catalyst compositions shown in Fig. 4A were tested in propylene oxidation using HTE at the same reaction conditions compared to those used for testing the three-component materials. The highest yield of oxygenates achieved for each composition is shown by the colors in Fig. 4A. The comparison of the experimental results with the SG rules on $\overline{EN}$ shows that the catalyst design rules derived by SGD correctly describe the experimental trend. In particular, the materials based on niobium, chromium, molybdenum, tin, lead and antimony achieve the highest oxygenate yields at relative lower E_1 molar fractions compared to the tantalum and tellurium-based materials, in line with the optimal ranges of $\overline{EN}$ values indicated by the SG rules.

All measured yield of oxygenates, corresponding to the materials shown in the catalyst map of Fig. 4A at all tested temperatures, are plotted as a function of $\overline{EN}$ in Fig. 4B. In this figure, the SG rules on $\overline{EN}$ are shown as vertical black lines and arrows. The variability of $\overline{EN}$ thresholds in the SG rule with respect to the input data set is indicated by the ranges of $\overline{EN}$ values in the grey shaded areas. These ranges correspond to the variations of the thresholds observed when using only 90% of the data set for training (see Table S2). The catalyst achieving the highest yield of oxygenates (60.19 % at 400°C) contains 0.35 molar fraction of tellurium as E_1 element and lies within the window of $\overline{EN}$ values suggested by the SG rules.

We have also used the SG rules derived from the three-component materials to address five-component materials, which were tested experimentally (Fig. 4C and D). For this purpose, E_1 was fixed as tellurium based on the best four-component catalysts and molybdenum, niobium, tantalum, and antimony were evaluated as E_2. Thus, ruthenium, tungsten, phosphorus, tellurium, and E_2 enter in the composition of the considered five-component materials. The agreement between the SG rule and the measured oxygenate yield is reasonable in spite of the much higher materials complexity with respect to the catalysts used for training. In particular, the five-component catalyst corresponding to the highest yield of oxygenates (59.60 % at 400°C) contains tantalum as E_2 element and the composition-averaged electronegativity for this material is 3.947 eV. Such $\overline{EN}$ value lies within the threshold defined by the SG rule. These results demonstrate the potential of the SGD-HTE-&-theory approach to identify generalizable rules describing exceptional performance. Indeed, the identified four- and five-

component catalysts are significantly more complex than those of the training data set (three-component materials). Moreover, the four- and five-component outstanding catalysts achieve oxygenate yields (60.19 and 59.60%, respectively) up to twice as large as those obtained with three-component materials (highest value of 26.85 %). Thus, the SG rules hinted at materials that are significantly more performant that any of the observations used in training.

We note that the four- and five-component materials achieve the highest yields of oxygenates at higher temperatures (400 °C) compared to the the three-component systems (300ºC) - see Fig. S4. The SG rule on the reaction temperature derived based on the three-component materials data set is thus not transferable to the four-component ones.

Finally, we applied the SGD approach to the four- and five-component HTE data (746 data points, Fig. 5A and B) using the same candidate descriptive parameters used for the previous SGD analysis of three-component materials. The identified SG presenting the highest D_{clS} (0.355) contains 18 data points, i.e., ca. 2.4 % of the data set (black points in Fig. 5A and 5B). The selected data points correspond to one four-component material with tellurium as E_1 element as well as different compositions of five-component materials with $E_1 = \mathrm{Te}$ and $E_2 = \mathrm{Mo, Nb}$. The rules describing this SG (Fig. 5C) constraint the values of three parameters: $T \geq 360$ ℃, $\overline{r_{\mathrm{val}}} \leq 1.001$ Å, and $\overline{EA} < -1.593$ eV, (Fig. 5D, E, and F, respectively). The comparison of these SG rules with that for the SG obtained with the three-component materials data set (Fig. 2C) highlights the higher temperatures needed for the four- and five-component materials to achieve outstanding performance. Moreover, different composition-dependent parameters ($\overline{r_{\mathrm{val}}}$ and $\overline{EA}$) are required to describe this SG compared to the case of three-component materials (x_{P} and $\overline{EN}$), even though the electronegativity and the electron affinity are related by $EN = \frac{EA+IP}{2}$.

The SG rules derived in this study are expected to describe outstanding materials whose performance is governed by the same processes governing the reactivity of the exceptional materials in the input data sets used for training. The analysis of four- and five-component materials was focused, nevertheless, on low ruthenium contents and on E_1 and E_2 elements compatible with tungsten, i.e., with similar atomic radii. Thus, it is unclear if the SG rules presented in Fig. 5C can identify exceptional materials and conditions for any arbitrary ruthenium content or for E_1 and E_2 elements which have significantly different radii compared to tungsten. This is because different mechanisms may operate on these materials which could also lead to exceptional performance. Therefore, the SGD analysis might need to be performed including new data points covering such so-far unexplored portions of the materials space for enlarging the domain in which the SG rules can detect exceptional catalysts and reaction conditions.

Conclusions

In this paper, we applied the SGD approach to the design of selective oxidation phosphorus-containing supported catalysts based on data from HTE and DFT calculations. The yield of value-added oxygenate product measured by HTE was used as target, and parameters obtained from DFT-calculated free-atom properties were offered as candidate descriptive parameters. The composition-weighted electronegativity, the phosphorus content and the temperature are identified as key parameters associated to outstanding production of acrolein and acrylic acid from propylene in three-component catalysts containing ruthenium, tungsten and phosphorus. The SG rules on these key parameters not only rationalize a local reactivity pattern particularly associated with exceptional catalytic performance, but also guide the design of more complex catalysts. In particular, a five-component material containing ruthenium, tungsten, phosphorus, tellurium and tantalum in the composition, which presents an oxygenate yield more than twice as large as any observation in data set used for training, is captured by the SG rules. This local modelling approach is suitable for the search of exceptional materials whose structures and functions are hardly modelled explicitly by theory.

Electronic supplementary information

DFT calculation details, additional SGD details, and details on the choice of compatible elements for the four- and five-component materials are available as ESI. The SGD analysis described in this publication can be found in a Jupyter notebook at the *NOMAD Artificial-Intelligence Toolkit* (https://nomad-lab.eu/AItutorials/), where it can be repeated and modified directly in a web browser.

Author information

Corresponding authors

*foppa@fhi-berlin.mpg.de, *sandip.de@basf.com

Current addresses

[†]Christopher Sutton: University of South Carolina, Department of Chemistry and Biochemistry. Columbia, South Carolina, United States.

Acknowledgments

Mario Boley is acknowledged for helpful discussions. L. F. acknowledges the funding from the NOMAD CoE (European Union's Horizon 2020 research and innovation program under the grant agreement Nº 951786). Funding by BASF SE is gratefully acknowledged. We would also like to acknowledge the productive cooperation and fruitful interaction with BASF SE and hte GmbH. This collaboration was supported by members of the consortia NFDI4Cat (German Research-Data Infrastructure for Catalysis) and FAIRmat (FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids).

Conflict of interest statement

The authors declare no competing interests.

Data availability

All data generated and analyzed during this study are included in this published article as supplementary information files.

References

1.(a) Freund, H.-J.; Meijer, G.; Scheffler, M.; Schlögl, R.; Wolf, M., CO Oxidation as a Prototypical Reaction for Heterogeneous Processes. Angew. Chem. Int. Ed. 2011, 50 (43), 10064-10094; (b) Schlögl, R., Heterogeneous Catalysis. Angew. Chem. Int. Ed. 2015, 54 (11), 3465-3520; (c) Foppa, L.; Ghiringhelli, L. M.; Girgsdies, F.; Hashagen, M.; Kube, P.; Hävecker, M.; Carey, S. J.; Tarasov, A.; Kraus, P.; Rosowski, F.; Schlögl, R.; Trunschke, A.; Scheffler, M., Materials genes of heterogeneous catalysis from

clean experiments and artificial intelligence. MRS Bull. 2021, 46, doi:10.1557/s43577-021-00165-6.

2. Reuter, K.; Stampf, C.; Scheffler, M., Ab Initio Atomistic Thermodynamics and Statistical Mechanics of Surface Properties and Functions. In Handbook of Materials Modeling: Methods, Yip, S., Ed. Springer Netherlands: Dordrecht, 2005; pp 149-194.
3. (a) Hattrick-Simpers, J.; Wen, C.; Lauterbach, J., The materials super highway: integrating high-throughput experimentation into mapping the catalysis materials genome. Catal. Lett. 2015, 145 (1), 290-298; (b) Farrusseng, D., High-throughput heterogeneous catalysis. Surf. Sci. Rep. 2008, 63 (11), 487-513; (c) Tompos, A.; Sanchez-Sanchez, M.; Végvári, L.; Szijjártó, G. P.; Margitfalvi, J. L.; Trunschke, A.; Schlögl, R.; Wanninger, K.; Mestl, G., Combinatorial optimization and synthesis of multiple promoted MoVNbTe catalysts for oxidation of propane to acrylic acid. Catal. Today 2019.
4. (a) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L., Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 2018, 559 (7714), 377-381; (b) Williams, T.; McCullough, K.; Lauterbach, J. A., Enabling Catalyst Discovery through Machine Learning and High-Throughput Experimentation. Chem. Mater. 2020, 32 (1), 157-165; (c) Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J., Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. Adv. Mater. 2020, 32 (14), 1907801.
5. (a) Wrobel, S. In An algorithm for multi-relational discovery of subgroups, European symposium on principles of data mining and knowledge discovery, Springer: 1997; pp 78-87; (b) Friedman, J. H.; Fisher, N. I., Bump hunting in high-dimensional data. Statistics and Computing 1999, 9 (2), 123-143; (c) Atzmueller, M., Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2015, 5 (1), 35-49; (d) Boley, M.; Goldsmith, B. R.; Ghiringhelli, L. M.; Vreeken, J., Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. Data Min. Knowl. Discov. 2017, 31 (5), 1391-1418; (e) Goldsmith, B. R.; Boley, M.; Vreeken, J.; Scheffler, M.; Ghiringhelli, L. M., Uncovering structure-property relationships of materials by subgroup discovery. New J. Phys. 2017, 19 (1), 013031.
6. (a) Foppa, L.; Ghiringhelli, L. M., Identifying Outstanding Transition-Metal-Alloy Heterogeneous Catalysts for the Oxygen Reduction and Evolution Reactions via Subgroup Discovery. Top. Catal. 2021, doi:10.1007/s11244-021-01502-4; (b) Mazheika, A.; Wang, Y.; Valero, R.; Ghiringhelli, L. M.; Vines, F.; Illas, F.; Levchenko, S. V.; Scheffler, M., Ab initio data-analytics study of carbon-dioxide activation on semiconductor oxide surfaces. arXiv:1912.06515 2019.
7. Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M., Identifying domains of applicability of machine learning models for materials science. Nat. Commun. 2020, 11 (1), 4428.
8. CRC Handbook of Chemistry and Physics. Cleveland, Ohio : CRC Press: 2020; Vol. 101.
9. (a) Grasselli, R. K.; Burrington, J. D., Oxidation of Low-Molecular-Weight Hydrocarbons. Handbook of Heterogeneous Catalysis 2008, 3479-3489; (b) Grasselli, R. K., Fundamental Principles of Selective Heterogeneous Oxidation Catalysis. Top. Catal. 2002, 21 (1), 79-88; (c) Schlögl, R., Selective Oxidation: From a Still Immature Technology to the Roots of Catalysis Science. Top. Catal. 2016, 59 (17), 1461-1476.
10. (a) Li, X.; Teschner, D.; Streibel, V.; Lunkenbein, T.; Masliuk, L.; Fu, T.; Wang, Y.; Jones, T.; Seitz, F.; Girgsdies, F.; Rosowski, F.; Schlögl, R.; Trunschke, A., How to control selectivity in alkane oxidation? Chem. Sci. 2019, 10 (8), 2429-2443; (b) Kube, P.; Frank, B.; Schlögl, R.; Trunschke, A., Isotope Studies in Oxidation of Propane over Vanadium Oxide. ChemCatChem 2017, 9 (18), 3446-3455.
11. (a) Schulz, C.; Roy, S. C.; Wittich, K.; d'Alnoncourt, R. N.; Linke, S.; Strempel, V. E.; Frank, B.; Glaum, R.; Rosowski, F., αII-(V1-xWx)OPO4 catalysts for the selective oxidation of n-butane to maleic anhydride. Catal. Today 2019, 333, 113-119; (b) Lister, S. E.; Soleilhavoup, A.; Withers, R. L.; Hodgkinson, P.; Evans, J. S. O., Structures and Phase Transitions in (MoO2)2P2O7. Inorg. Chem. 2010, 49 (5), 2290-2301; (c) Roy, S. C.; Raguž, B.; Assenmacher, W.; Glaum, R., Synthesis and crystal structure of mixed metal(III) tungstenyl(VI) ortho-pyrophosphates. Solid State Sciences 2015, 49, 18-28.
12. Machado, R.; Dimitrakopoulou, M.; Girgsdies, F.; Löser, P.; Xie, J.; Wittich, K.; Weber, M.; Geske, M.; Glaum, R.; Karbstein, A.; Rosowski, F.; Titlbach, S.; Skorupska, K.; Tarasov, A.; Schlögl, R.; Schunk, S. A., Platinum group metal phosphates as catalysts for selective C-H activation of lower alkanes. In preparation.
13. Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G., Assessing the performance of recent density functionals for bulk solids. Phys. Rev. B 2009, 79 (15), 155107.
14. Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M., Ab initio molecular simulations with numeric atom-centered orbitals. Comput. Phys. Commun. 2009, 180 (11), 2175-2196.
15. Nguyen, H.-V.; Vreeken, J. In Non-parametric jensen-shannon divergence, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer: 2015; pp 173-189.
16. Waroquiers, D.; Gonze, X.; Rignanese, G.-M.; Welker-Nieuwoudt, C.; Rosowski, F.; Göbel, M.; Schenk, S.; Degelmann, P.; André, R.; Glaum, R.; Hautier, G., Statistical Analysis of Coordination Environments in Oxides. Chem. Mater. 2017, 29 (19), 8346-8360.