Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00174a

Coherent collections of rules describing exceptional materials identified with a multi-objective optimization of subgroups†

Lucas Foppa * and Matthias Scheffler

Useful materials are often statistically exceptional and they might be overlooked by artificial intelligence (AI) models that attempt to describe all materials simultaneously. These global models perform well for the majority of materials, but they do not necessarily capture the useful ones. Subgroup discovery (SGD) identifies descriptions of subsets of materials associated with exceptional values of a chosen property. Thus, SGD can better capture exceptional materials compared to widely used AI techniques. Previous studies focused on the SG that maximizes an objective function establishing a tradeoff between the SG size and the exceptionality of the distribution of property values within the SG. However, this optimization does not give a unique solution, but many SGs typically have similar objective-function values. Here, we identify a "Pareto region" of SGD solutions presenting a multitude of size-exceptionality tradeoffs. The approach is demonstrated by learning descriptions of perovskites with a high bulk modulus.

Received 28th April 2025
Accepted 17th June 2025

DOI: 10.1039/d5dd00174a

rsc.li/digitaldiscovery

1 Introduction

Materials serve as the cornerstone of critical economic sectors and they play a pivotal role in driving the transition to a sustainable economy and to renewable energy.[1,2] Thus, there is an urgent need for the discovery of appropriate and more efficient materials. However, the materials that are useful for a given application are often statistically exceptional. These materials might present, for instance, extremely high values of materials properties compared to other known compounds. Exceptional materials are very few compared to the practically infinite space of possible materials, which remains largely unknown.[3,4] Artificial intelligence (AI) has been increasingly applied to identify correlations and patterns in data in materials science and engineering.[5–8] Indeed, AI might describe materials properties and functions governed by intricate mechanisms better than previous theoretical and computational approaches because it targets correlations and does not assume a single underlying physical model.[9] Thus, AI holds the potential to accelerate the exploration of the immense materials space, leading to the discovery of new materials. However, capturing exceptional materials is a challenging task for most widely used AI methods.

AI methods often fail in describing the exceptional materials first because the training data are typically not well distributed over (or representative of) the huge, unknown materials space. Therefore, interpolation schemes are unable to generalize to potentially interesting portions of the materials space that were disregarded in the training data.[10–14] This issue, which might be referred to as an "out of distribution" issue, can be alleviated by AI approaches that can better extrapolate compared to methods that are inherently interpolative.[12,15] Besides, AI model training can be combined with the systematic acquisition of new data corresponding to portions of the materials space that were not covered by the initial training data using sequential-learning approaches such as active learning or Bayesian optimization.[16–19] However, the efficiency of sequential learning often relies on the quality of uncertainty estimates, which is in some cases problematic.[20–22] A second key reason that can explain the inability of current AI approaches to capture exceptional materials is the focus on global models. These models attempt to describe all materials simultaneously. They are obtained by optimizing an objective (loss) function that reflects the average performance, *e.g.*, the mean prediction error. Thus, global models are designed to perform well in average for the majority of (uninteresting) materials, but do not necessarily perform well for exceptional ones.[23] Objective functions can be adapted to give more importance to the description of specific property values, *e.g.*, high values.[24] However, different groups of materials could operate according to different mechanisms. This might render a global description not only inaccurate, but also inappropriate.

Alternative AI methods for materials discovery include strategies based on similarity among materials[25,26] or among

The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, D-14195 Berlin, Germany. E-mail: foppa@fhi-berlin.mpg.de

† Electronic supplementary information (ESI) available: Comparison of approaches for defining the Pareto region and analysis of the Pareto region of SG solutions based on similarity between SG rules and hierarchical clustering. See DOI: https://doi.org/10.1039/d5dd00174a

their constituents, *e.g.*, ions in solids,[27] and the subgroup-discovery (SGD)[28,29] approach. In particular, SGD tackles the limitations of global descriptions and it has the potential to better capture exceptional materials. Indeed, SGD has been recently put forward in materials science.[30–35] SGD is a supervised, descriptive rule-induction[36] technique and it identifies subsets of a dataset associated with exceptional values of a target quantity of interest, for instance a materials property or performance indicator. Crucially, SGD identifies these subsets (SGs) of data along with the descriptions of these subsets, referred to as rules. The SGD analysis starts with the choice of many features that relate to possibly relevant mechanisms governing the target materials property. Then, SGD creates a number of statements about the features that are satisfied only for a part of the dataset. These statements are, for instance, inequalities constraining the values of the features. Finally, a search algorithm[37–39] identifies the combination of typically few statements that results in a SG that maximizes an objective function. This objective (or quality) function is a product of the relative SG size and the so-called utility function. The relative SG size is the fraction of data points that satisfies the statements describing the SG. The higher the relative SG size, the more general the description. The utility function quantifies the "exceptionality" of the distribution of target values in the SG with respect to the entire dataset. The positive mean shift is one example of a utility function often utilized when the target values of interest are high. This utility function measures the shift of the mean value of the target in the SG with respect to the mean value of the target in the entire dataset. The higher the value of the positive-mean-shift utility function, the higher the target values in the identified SG. Thus, such a utility function favors the identification of SGs associated with high target values. We will discuss this utility function in more detail in the Results section. We will also discuss a second example of the utility function, namely the Jensen–Shannon divergence between the distributions of target values in the SG and in the entire dataset.

SG rules typically constrain the values of only a few key features, out of the many initially offered ones. Thus, SGD learns a (low-dimensional) representation. The SGD approach is illustrated in Fig. 1 (top). The aim of SGD is to find descriptions of portions of the materials space that are exceptional. Thus, it accepts that the mechanisms governing the materials' properties might vary across the materials space and that not all of these mechanisms need to be described for the discovery of useful materials. Indeed, SGD was used to identify rules

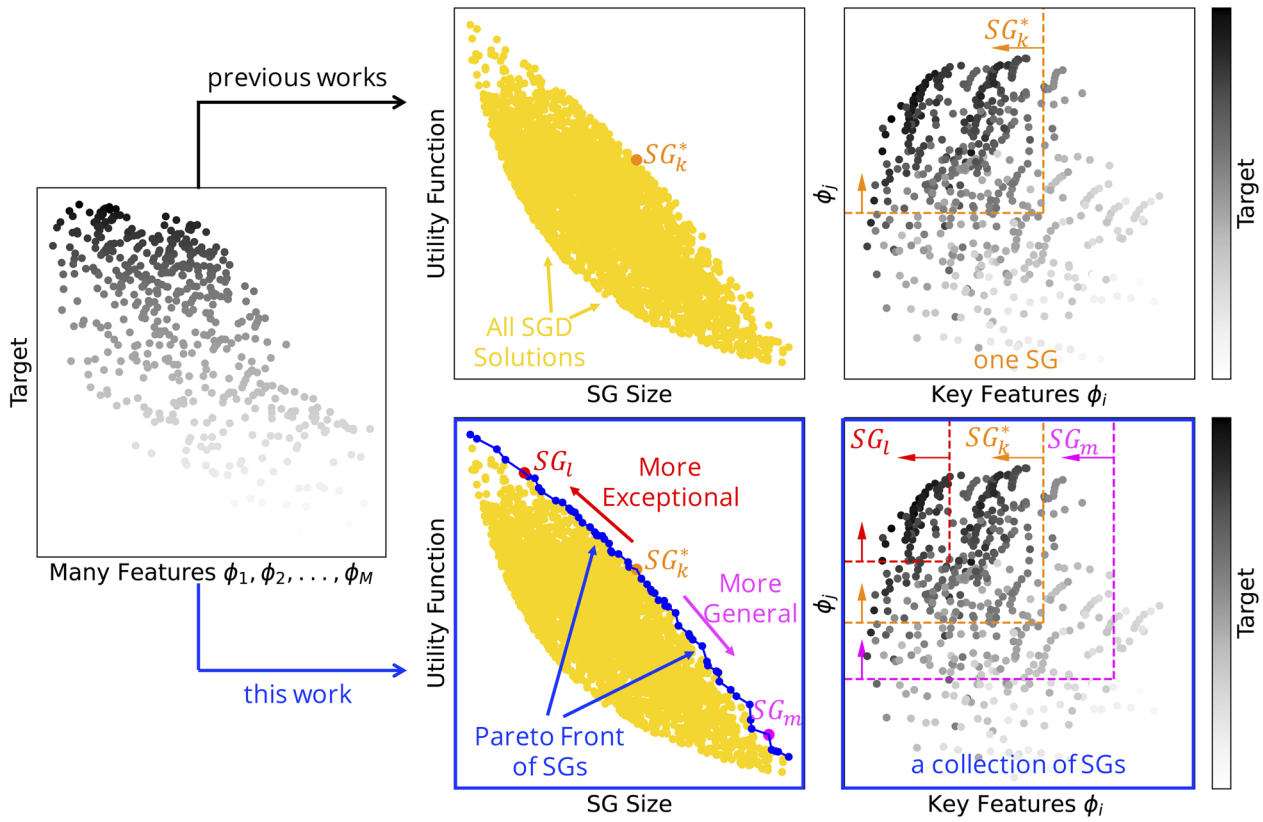

Fig. 1 Subgroup discovery (SGD) identifies subsets of materials that are outstanding with respect to a certain materials property, the target of interest. The subsets, or subgroups (SGs), are described by rules that typically constrain key features characterizing the materials and mechanisms governing the target, out of many initially offered features. The subgroups are obtained by maximizing an objective function that is a product of the (relative) SG size and the utility function. These two terms reflect the generality and the exceptionality of the SG, respectively. Top: Previous studies focused on the one SG that maximizes the objective function, denoted as SG_k^*. Bottom: In this work, we identify a collection of SGs containing, *e.g.*, SG_l, SG_k^*, and SG_m, and presenting multiple tradeoffs between the SG size and the utility function.

associated with high-performance materials even based on datasets dominated by low-performance situations.[35] Additionally, SGD is an exploratory analysis that can identify unexpected patterns and anomalies. We note that SGD is significantly different from clustering techniques because it does not aim at describing the entire dataset. Moreover, SGD is a supervised approach, and it explicitly identifies rules indicating why the data points belong to the SG. Clustering is an unsupervised technique that groups data points into clusters based on similarity, without considering any target quantity. Besides, clustering does not explicitly identify why the data points are clustered together.

Previous SGD studies[30–33] focused on the identification of the SG (and rules) that maximizes the objective function, as illustrated in Fig. 1 (top). However, the SG that maximizes the objective function does not reflect all possible tradeoffs between the relative SG size and utility function that could be relevant for a given application. Additionally, the definitions of some utility functions assume that the distributions of target values in the entire dataset and in the SG are appropriately characterized by one single summary-statistics value, such as the mean value in the case of the positive-mean-shift utility function. However, this assumption may be questioned in materials science, as the distributions can significantly deviate from the normal distribution. Indeed, distributions related to materials can be skewed or even bimodal. Finally, utility functions often assume that the summary-statistics values appropriately reflect the huge, unknown materials space, *e.g.*, the mean value in the dataset is a good approximation for the mean value in the entire materials space. This assumption does not hold if the datasets are created according to certain selection biases. Thus, the datasets can be highly unbalanced compared to the materials space.

In this manuscript, we introduce a multi-objective optimization of SGs that identifies coherent collections of SGs and rules in the "Pareto region" of optimal SGD solutions. These SGs present a multitude of tradeoffs between the relative SG size and the utility function, the two conflicting objectives in SGD. This concept is schematically shown in Fig. 1 (bottom). The multi-objective optimization of SGs is demonstrated for the identification of ABO_3 perovskites with a high bulk modulus as an example of a target. We compare the SGs obtained with two different utility functions, the positive mean shift and the cumulative Jensen–Shannon divergence. The latter does not make assumptions on the shape of the distributions of target values. We also analyze the sensitivity of the results with respect to the offered set of features. Finally, we exploit the rules trained on a dataset of 504 single ABO_3 perovskites to identify high-bulk-modulus perovskites out of a candidate space of 12 096 single ABO_3 and double $A_2BB'O_6$ perovskites. Our results show that rules focusing on perovskites with a high bulk modulus do not necessarily correspond to the single SG which maximizes the objective function, but they can be systematically derived with the Pareto-region concept. These rules identify perovskites of the candidate space that present the bulk modulus up to 13% higher than the highest value of the training set.

2 Methods

2.1 Subgroup discovery

The SGD approach is based on a dataset of materials, which we denote as $\tilde{P}$. This dataset is part of the huge materials space, the full population, P. Each material of the full population is associated with a set of features, namely physical parameters that are potentially related to a target quantity of interest y, for instance, a materials property. The target of interest is only known for the materials in the dataset. SGD starts by systematically constructing statements about the features. Each statement is only verified for a part of the materials in the dataset. Thus, the statements select part of the dataset. The construction of these statements follows different approaches depending on the type of feature: categorical, ordinal, or metric. For categorical features, *i.e.*, when the feature values are a discrete set with no relevant order, all possible statements of the form $\phi = c_i$ are constructed, where c_i are the categories in the dataset. For ordinal features, *i.e.*, when the feature values contain a set of discrete and ordered values, all possible inequality constraints such as $\phi \geq z_i$ and $\phi \leq z_i$ are generated, where z_i represents the integer values in the dataset. For metric features, *i.e.*, when the feature values are from a continuous ordered scale, statements similar to those of the ordinal case are constructed, *i.e.*, $\phi \geq \nu_i$ and $\phi \leq \nu_i$. In this case, however, one cannot simply use all possible ν_i values, but instead has to find a small computationally feasible subset of ν_i values. This is accomplished with the aid of k-means clustering. First, the clustering algorithm is applied to identify $k + 1$ values representing the center of clusters corresponding to range of values for each of the features in the dataset. Then, the arithmetic means between the centers of two neighboring clusters are taken as possible ν_i. Thus, the possible ν_i values are closer to each other when the concentration of data is higher. In this work, we use $k = 10$. Further details on the construction of statements and on the choice of k are discussed elsewhere.[30,40] Then, SGD uses a search algorithm, for instance Monte Carlo-based[37,38] or branch-and-bound,[39] to identify conjunctions of statements constructed with the "AND" operator ($\wedge$), that result in SGs that maximize an objective (quality) function Q of the form

$$Q(\mathrm{SG}, \tilde{\mathrm{P}}) = \left(\frac{s(\mathrm{SG})}{s(\tilde{\mathrm{P}})}\right)^{\alpha} \left(u(\mathrm{SG}, \tilde{\mathrm{P}})\right)^{\beta}. \quad (1)$$

Here, $s(\mathrm{SG})$ and $s(\tilde{\mathrm{P}})$ are the sizes of the SG and of the dataset $\tilde{\mathrm{P}}$, respectively, *i.e.*, the number of data points that satisfy the statements defining the SG and the number of data points in the entire dataset. The ratio between the size of the SG and the size of the dataset, $s(\mathrm{SG})/s(\tilde{\mathrm{P}})$, is referred to as the relative SG size. $u(\mathrm{SG}, \tilde{\mathrm{P}})$ is the utility function describing how exceptional the distribution of the target in the SG is compared to the entire dataset. The utility function is chosen according to the question to be addressed, and there are many possibilities.[31] The positive shift of the mean value of the target in the SG compared to the mean value of the target in the entire dataset and the Jensen–Shannon divergence between the distribution of target values in

the SG and the distribution of target values in the entire dataset[41] are two examples of utility functions that we will consider in this work. Finally, α and β are tunable parameters controlling the tradeoff between the relative SG size, *i.e.*, the generality of the description, and the utility function, *i.e.*, the exceptionality of the description. Usually, $\alpha = \beta = 1$ or $\beta = 1 - \alpha$, with $\alpha \in [0.1, 0.9]$. The Monte Carlo search algorithm[37,38] randomly generates conjunctions of the previously generated statements with probability proportional to $\frac{s(\mathrm{SG})}{s(\tilde{\mathrm{P}})}$. Then, an opportunistic pruning algorithm refines these conjunctions by removing statements that result in the increase of $Q(\mathrm{SG}, \tilde{\mathrm{P}})$ values. The iterative removal of statements leads to the maximization of the objective function of eqn (1). We note that SG search algorithms such as the branch-and-bound approach[39] are more systematic than the stochastic Monte Carlo algorithm. However, the computational cost of the branch-and-bound approach increases more rapidly with the number of statements compared to the cost of the Monte Carlo search. Finally, it should be noted that optimizing the SGD objective function over the full set of possible conjunctions of statements about the data is an NP-hard combinatorial optimization problem and that SG searches are extensive but not exhaustive. Thus, there is no guarantee that all possible SGs will be identified in the Pareto front of SGD solutions. In this work, we used the SGD algorithm implemented in the realkd version 0.7.2. A Monte Carlo-based SG search algorithm[37,38] was used with 50 000 seeds for the initialization.

The inputs to the SGD analysis are the datasets containing a target quantity of interest (*e.g.*, a materials property) and the features that characterize the materials. Additionally, one has to choose an appropriate quality function, which determines the desired distribution of target values in the SGs. The outputs are the subsets of data (SGs) and the rules (statements) that describe these subsets of data. These rules typically depend only on key features, out of all initially offered features. In analogy to genes in biology, these key features might be called materials genes,[9] as they correlate with the mechanisms governing the materials properties. The rules can be exploited to efficiently identify the few exceptional materials in the huge materials space P, for which the target property is unknown.

2.2 Approach for identifying the Pareto region of SGD solutions

In order to identify the pursued coherent collections of SGs with multiple generality-exceptionality tradeoffs, we first run the SGD algorithm using the objective function of eqn (1) with $\alpha = \beta = 1$. Thus, relative SG size and the utility function are given the same importance. Then, we collect a number of SGD solutions identified by this algorithm that display high objective-function values. Among these top-ranked SGD solutions, we identify a Pareto front with respect to the two objectives relative SG size and the utility function. In multi-objective optimization, a Pareto front is the set of solutions for which no single objective can be improved without deteriorating at least one other objective. Thus, the solutions in the Pareto front reflect an optimal tradeoff between competing objectives. To ensure that no interesting SGD solution is left out, we included in our analysis not only solutions that are part of the Pareto front but also solutions within a fixed threshold distance (in this work equal to 0.01) to the Pareto front in the relative SG size-utility function space, *i.e.*, solutions which are near the Pareto front. We refer to the solutions at the Pareto front plus the solutions near the Pareto front as the *Pareto region*. The definition of a Pareto region *via* a fixed distance to the Pareto front ensures that all SGD solutions of the Pareto region have objective-function values within the range determined by the chosen threshold distance. However, this approach is sensitive to the form of the Pareto front and the distribution of SGD solutions in the Pareto front. This issue can be alleviated by defining the Pareto region based on subsequent Pareto fronts. This alternative approach to define the Pareto region is discussed in detail and compared with the distance-based approach in the ESI.†

3 Results and discussion

3.1 Identification of perovskites with a high bulk modulus

The identification of coherent collections of SGs and the usefulness of our approach will be demonstrated for the learning of rules describing the bulk modulus (B_0) of ABO_3 perovskites. More specifically, the problem that we will tackle is the identification of materials that exhibit a high bulk modulus. The bulk modulus quantifies the resistance of the material to compression and it correlates with the materials' hardness. We will use SGD to identify rules based on basic physical parameters which describe subsets of materials presenting a high bulk modulus. Thus, the bulk modulus is the target of our SGD analysis. The rules obtained by SGD using a training dataset of 504 materials will then be used to identify promising materials with a high bulk modulus from a pool of 12 096 candidate materials. Perovskites are a promising materials class[42,43] for energy-related applications such as photovoltaics and catalysis[44–46] and they have been the subject of a number of AI and machine-learning studies.[47–51]

The dataset[52] used to train the SG rules contains 504 perovskites composed of *A* elements from the alkali, alkaline-earth, and scandium groups and lanthanides. *B* elements include transition metals and main-group elements such as bismuth, antimony, and germanium. The choice of *A* and *B* elements reflects common elements reported in perovskites.[44–46] We only consider the cubic, highly symmetric perovskite structure in our dataset, which is often only stable at high temperatures. Thus, our analysis focuses on diversity of the chemical elements entering the material rather than on the diversity of structures. However, it is straightforward to extend the SGD approach to other, less symmetric crystal structures. We used 24 features characterizing the perovskites (Table 1). Two of the features are properties of the solid perovskite materials (denoted S), the equilibrium lattice constant (a_0) and the cohesive energy (E_0). The equilibrium lattice constant is the only structural degree of freedom of the cubic structure. The cohesive energy corresponds to the energy required to atomize the materials' crystal.

Table 1 Features used to characterize the ABO_3 perovskites in the SGD analysis

Type	Name	Symbol	Unit
S[a]	Equilibrium lattice constant[d]	a_0	Å
S[a]	Cohesive energy[de]	E_0	eV per atom
A[b]	Radii of the valence-s orbitals of the A and B neutral atoms[d]	$r_{s,A}$ and $r_{s,B}$	Å
A[b]	Radii of the valence-s orbitals of the A and B +1 cations[d]	$r_{s,A}^{cat}$ and $r_{s,B}^{cat}$	Å
A[b]	Radii of the highest-occupied orbitals of A and B neutral atoms[d]	$r_{val,A}$ and $r_{val,B}$	Å
A[b]	Radii of the highest-occupied orbitals of the A and B +1 cations[d]	$r_{val,A}^{cat}$ and $r_{val,B}^{cat}$	Å
A[b]	Electron affinity of the A and B atoms[d]	EA_A and EA_B	eV
A[b]	Ionization potential of the A and B atoms[d]	IP_A and IP_B	eV
A[b]	Electronegativity of the A and B atoms[d]	EN_A and EN_B	eV
A[b]	Kohn–Sham single-particle eigenvalue of the highest-occupied orbital of the A and B atoms[d]	$\varepsilon_{H,A}$ and $\varepsilon_{H,B}$	eV
A[b]	Kohn–Sham single-particle eigenvalue of the lowest-unoccupied orbital of the A and B atoms[d]	$\varepsilon_{L,A}$ and $\varepsilon_{L,B}$	eV
A[b]	Atomic numbers of A and B elements	Z_A and Z_B	$\mathbb{Z}$
C[c]	Expected oxidation states of the elements A and B in the perovskite formula[f]	n_A and n_B	$\mathbb{Z}$

[a] Properties of the solid material. [b] Properties of free atoms of elements constituting the material. [c] Properties of the composition of the material. [d] Evaluated using DFT-PBEsol. [e] Energy needed per atom to atomize the crystal. [f] Defined based on the periodic-table group of the A element and on the charge neutrality of the ABO_3 composition, *i.e.*, $n_A + n_B = 6$.

Ten of the features are atomic properties of free atoms of the elements A or B (denoted A), such as orbital radii, ionization potential, and electronegativity. Finally, we included two features that depend on the composition of the material (denoted C), the expected oxidation states of A and B elements in the compound (n_A and n_B, respectively). The bulk modulus and the features (except the atomic numbers of A and B, n_A and n_B) were calculated using density-functional theory (DFT) with the PBEsol exchange-correlation functional. The bulk modulus is evaluated by fitting the Birch–Murnaghan equation of state to a series of energies of the crystal calculated using structures that present slightly larger or smaller volumes than the equilibrium volume. Further calculation details are provided elsewhere.[52] We note that some of the features in Table 1 are correlated with each other. This is not a limitation for SGD. However, the presence of correlated features might result in similar SGs defined by slightly different rules.

3.1.1 Collection of SG rules obtained with the positive-mean-shift utility function. We start by analyzing the results obtained with the positive-mean-shift utility function, defined as

$$u(\mathrm{SG}, \tilde{\mathrm{P}}) = \frac{\bar{y}(\mathrm{SG}) - \bar{y}(\tilde{\mathrm{P}})}{y_{\max}(\tilde{\mathrm{P}}) - \bar{y}(\tilde{\mathrm{P}})}. \quad (2)$$

Here, $\bar{y}(\mathrm{SG})$ and $\bar{y}(\tilde{\mathrm{P}})$ are the mean values of the distribution of the target in the SG and in the entire dataset and $y_{\max}(\tilde{\mathrm{P}})$ is the maximum value that the target assumes in the dataset. In our application, the target is the bulk modulus and $y_{\max}(\tilde{\mathrm{P}}) = 1.49$ eV Å^{-3} for the $ScMnO_3$ perovskite. The utility function in eqn (2) requests that the values of the target within the SG are high with respect to the mean value of the target in the dataset. It assumes that the distributions of target values in the SG and in the entire dataset are properly described by the mean values.

The 5000 SGs with the highest objective-function values identified in the analysis using the positive-mean-shift utility function are shown as grey points in Fig. 2(A). The Pareto region is shown in blue in this plot: the 60 and 49 SGs belonging to the Pareto front and to the near-Pareto-front region are displayed in dark and light blue, respectively. This plot shows, in orange, the SG that maximizes the objective function, denoted as $\mathrm{SG}_{55}^{\mathrm{m}*}$. The curve corresponding to the constant value of $Q(\mathrm{SG}_{55}^{\mathrm{m}*}, \tilde{\mathrm{P}})$ is shown as a dashed orange line. Note that we assign the i indices to the SGs of the Pareto region $\mathrm{SG}_i^{\mathrm{m}}$ according to increasing values of relative SG size. The star in $\mathrm{SG}_{55}^{\mathrm{m}*}$ indicates that this is the SG associated with the maximum objective-function value. The Pareto region contains many SGs with objective-function values close to the maximum at relative SG sizes in the range [0.4, 0.6]. Conversely, SGs in the Pareto region with relative sizes lower than 0.4 and higher than 0.6 present relatively lower objective-function values compared to the maximum value.

The Pareto-region concept leads to the identification of not one, but a collection of 109 SGs presenting multiple generality-exceptionality tradeoffs. However, it is unclear how to choose which of these SGs should be considered for a detailed analysis of physical insights or for materials discovery in larger candidate materials spaces. Many of the SGs of the Pareto region might be similar to each other and they might contain redundant information. In order to assess the variability of the SG rules of the Pareto region and to facilitate further analysis of these rules, we established a measure of similarity between SGs and used it to identify clusters of SGs containing similar SGs.[53] This analysis is described in detail in the ESI.† In summary, the similarity is assessed using Jaccard indices. These indices consider that the similarity between two SGs is proportional to the overlap of their elements, *i.e.*, to the number of data points that satisfy the rules defining both SGs. Thus, this similarity will be high between SG rules that result in a similar selection of materials, even though the rules themselves might be different, *e.g.*, due to correlated key features or due to different thresholds. To obtain the clusters, we applied agglomerative hierarchical clustering.[54]

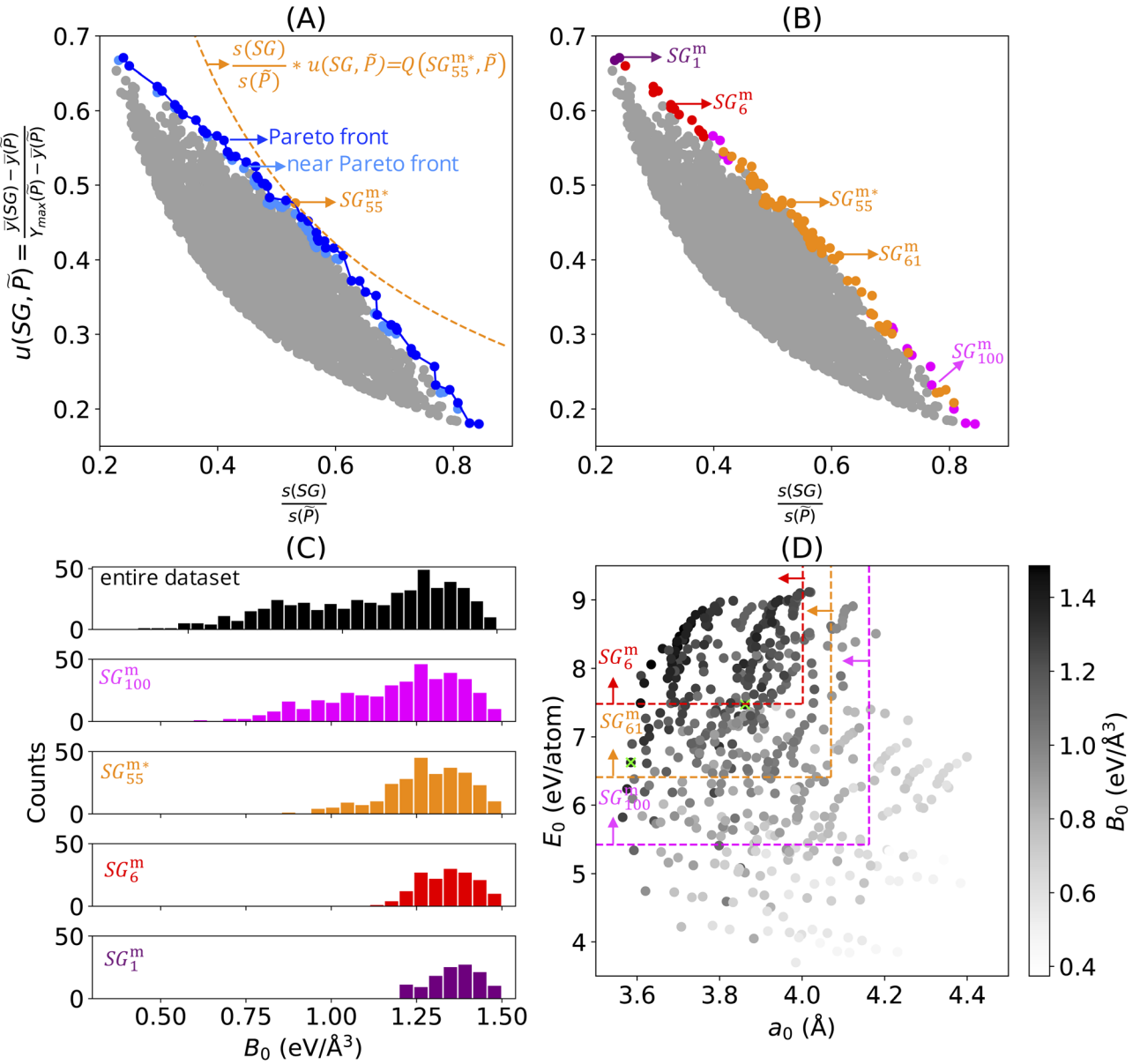

Fig. 2 Collections of SGs describing perovskites with a high bulk modulus (B_0) obtained using the multi-objective optimization approach and the positive-mean-shift utility function. The results for the full feature set containing the 24 features in Table 1 are shown. (A) The 5000 SGD solutions with high values of the objective function are shown in grey and the Pareto region of SGD solutions is shown in blue. The SG associated with the maximum value of the objective function is shown in orange. (B) The 109 SGs of the Pareto region are clustered according to their similarity *via* hierarchical clustering. Each cluster is shown in a different color. (C) Distributions of B_0 in the entire training dataset and in some examples of SGs of the Pareto region. The rules associated with these SGs are shown in Table 2. (D) SG rules for some examples of SGs of the Pareto region that constrain the values of the equilibrium lattice constant (a_0) and cohesive energy (E_0).

The results of the similarity analysis and clustering of the Pareto region in Fig. 2(A) for a chosen number of four clusters are displayed in Fig. 2(B). In this figure, the four identified clusters are displayed in four different colors. In general, the clusters of SGs correspond to different ranges of relative sizes. The cluster shown in orange, for instance, can be related to the SGD solutions with objective-function values close to the maximum in the relative size range of [0.4, 0.6]. Interestingly, a small cluster containing only three SGs is identified at low relative sizes. This indicates that these three SG rules are unique compared to the remaining ones. We note that the SGs of the Pareto region are spread in a continuous manner in the utility-function *vs.* relative-size plot in Fig. 2(A). The aim of the clustering technique is not to identify clusters of SGs that preexist in the utility-function *vs.* relative-size space, but rather to partition the pool of SGs of the Pareto region into clusters containing similar rules. This partitioning aims to facilitate the analysis of the many SGD solutions identified with the multi-objective-optimization approach.

Table 2 Characteristics of some of the SGs identified with the Pareto-region approach. Information on all other SGs of the Pareto region is provided in the ESI

Features	Index[b]	$Q(\mathrm{SG}, \tilde{\mathrm{P}})$	$u(\mathrm{SG}, \tilde{\mathrm{P}})$	$s(\mathrm{SG})/s(\tilde{\mathrm{P}})$	$\bar{y}(\mathrm{SG})$[c]	$y_{\mathrm{std}}(\mathrm{SG})$[d]	Rules
S, A, and C[a]	$\mathrm{SG}_{1}^{\mathrm{m}}$	0.16	0.67	0.24	1.36	0.07	$E_0 > 7.48$ eV per atom $\wedge$ $r_{\mathrm{s},B}^{\mathrm{cat}} \leq 1.44$ Å
S, A, and C[a]	$\mathrm{SG}_{6}^{\mathrm{m}}$	0.19	0.63	0.31	1.34	0.08	$E_0 > 7.48$ eV per atom $\wedge$ $a_0 \leq 4.00$ Å
S, A, and C[a]	$\mathrm{SG}_{55}^{\mathrm{m}*}$	0.25	0.53	0.48	1.28	0.12	$E_0 \geq 6.41$ eV per atom $\wedge$ $a_0 \leq 4.07$ Å $\wedge$ $r_{\mathrm{s},B}^{\mathrm{cat}} \geq 0.94$ Å
S, A, and C[a]	$\mathrm{SG}_{61}^{\mathrm{m}}$	0.25	0.45	0.55	1.27	0.13	$E_0 \geq 6.41$ eV per atom $\wedge$ $a_0 \leq 4.07$ Å
S, A, and C[a]	$\mathrm{SG}_{100}^{\mathrm{m}}$	0.20	0.26	0.77	1.19	0.18	$E_0 \geq 5.42$ eV per atom $\wedge$ $a_0 \leq 4.16$ Å
S, A, and C[a]	$\mathrm{SG}_{5}^{\mathrm{JS}}$	0.06	0.61	0.10	1.42	0.04	$-4.55 \leq \varepsilon_{\mathrm{L},B} < -4.33$ eV $\wedge$ $a_0 < 3.85$ Å $\wedge$ $n_B < 3.5$
S, A, and C[a]	$\mathrm{SG}_{96}^{\mathrm{JS}}$	0.10	0.45	0.22	1.37	0.06	$r_{\mathrm{val},A} \leq 1.51$ Å $\wedge$ $\mathrm{EA}_B \geq -1.84$ eV $\wedge$ $r_{\mathrm{s},B}^{\mathrm{cat}} \leq 1.50$ Å $\wedge$ $r_{\mathrm{val},B} \leq 1.14$ Å $\wedge$ $E_0 > 7.11$ eV per atom
S, A, and C[a]	$\mathrm{SG}_{129}^{\mathrm{JS}*}$	0.11	0.29	0.39	1.32	0.09	$\mathrm{EA}_B \leq 0.31$ eV $\wedge$ $\varepsilon_{\mathrm{L},B} \leq -3.50$ eV $\wedge$ $r_{\mathrm{s},B}^{\mathrm{cat}} \geq 1.09$ Å $\wedge$ $E_0 \geq 6.77$ eV per atom
S, A, and C[a]	$\mathrm{SG}_{169}^{\mathrm{JS}}$	0.09	0.16	0.55	1.27	0.13	$E_0 \geq 6.41$ eV per atom $\wedge$ $a_0 \leq 4.07$ Å
A and C[a]	$\mathrm{SG}_{1}^{\mathrm{JS}'}$	0.05	0.67	0.08	1.43	0.03	$\mathrm{EA}_B \geq -1.03$ eV $\wedge$ $-4.55 \leq \varepsilon_{\mathrm{L},B} < -4.33$ eV $\wedge$ $n_B < 4$
A and C[a]	$\mathrm{SG}_{4}^{\mathrm{JS}'}$	0.05	0.67	0.08	1.43	0.04	$\mathrm{IP}_B \leq 7.82$ eV $\wedge$ $\varepsilon_{\mathrm{L},B} < -4.33$ eV $\wedge$ $r_{\mathrm{val},B} < 0.68$ Å $\wedge$ $n_B < 3.5$
A and C[a]	$\mathrm{SG}_{5}^{\mathrm{JS}'}$	0.06	0.65	0.09	1.42	0.04	$r_{\mathrm{val},A} < 1.28$ Å $\wedge$ $\mathrm{EA}_B \leq 0.31$ eV $\wedge$ $Z_B < 36$ $\wedge$ $r_{\mathrm{s},B} \geq 1.26$ Å
A and C[a]	$\mathrm{SG}_{51}^{\mathrm{JS}'}$	0.10	0.43	0.22	1.37	0.06	$-5.11 \leq \varepsilon_{\mathrm{L},B} \leq -3.50$ eV $\wedge$ $r_{\mathrm{val},B}^{\mathrm{cat}} \leq 0.94$ Å $\wedge$ $n_A > 2$
A and C[a]	$\mathrm{SG}_{66}^{\mathrm{JS}'*}$	0.10	0.35	0.30	1.34	0.08	$r_{\mathrm{val},A}^{\mathrm{cat}} \leq 1.38$ Å $\wedge$ $\varepsilon_{\mathrm{L},B} \leq -3.49$ eV $\wedge$ $r_{\mathrm{val},B} \leq 1.14$ Å $\wedge$ $n_A \geq 1.5$
A and C[a]	$\mathrm{SG}_{179}^{\mathrm{JS}'}$	0.06	0.12	0.52	1.26	0.15	$\mathrm{EN}_A \geq 2.76$ eV $\wedge$ $\mathrm{EN}_B \leq 4.85$ eV $\wedge$ $r_{\mathrm{s},B} \geq 1.09$ Å

[a] S, A, and C correspond to solid, atomic, and compositional, respectively (see Table 1). [b] The star indicates the SG with the maximum value of objective (quality) function Q obtained with a given utility function and feature set. The superscripts "m" and "JS" correspond to the utility functions positive mean shift and cumulative Jensen–Shannon divergence, respectively. [c] Mean value of the target within the SG, in eV Å^{-3}. [d] Standard deviation of the target within the SG, in eV Å^{-3}.

We analyzed in more detail one SG per identified cluster. $\mathrm{SG}_{1}^{\mathrm{m}}$, $\mathrm{SG}_{6}^{\mathrm{m}}$, $\mathrm{SG}_{55}^{\mathrm{m}*}$, and $\mathrm{SG}_{100}^{\mathrm{m}}$ are examples of SGs belonging to the purple, red, orange, and magenta clusters, respectively. The distributions of bulk-modulus values in the entire dataset and in the mentioned SGs are shown in Fig. 2(C). As the relative SG size decreases, the SGs of the Pareto region have higher mean bulk-modulus values and narrower bulk-modulus distributions. For the goal of identifying perovskites with an extremely high bulk modulus, the rules associated with SGs with low relative SG size and high mean bulk-modulus values, *e.g.*, associated with $\mathrm{SG}_{1}^{\mathrm{m}}$ or $\mathrm{SG}_{6}^{\mathrm{m}}$, are useful, since they provide a more focused description. Such SGs would not be detected based solely on the maximization of the objective function.

Next, we analyzed the rules defining $\mathrm{SG}_{1}^{\mathrm{m}}$, $\mathrm{SG}_{6}^{\mathrm{m}}$, $\mathrm{SG}_{55}^{\mathrm{m}*}$, and $\mathrm{SG}_{100}^{\mathrm{m}}$, shown in Table 2. The rules constrain the values of 3 key features, out of the 24 offered features: the equilibrium lattice constant (a_0), cohesive energy (E_0), and the radius of valence-s orbitals of +1 cations (cat) of the B element ($r_{\mathrm{s},B}^{\mathrm{cat}}$). In particular, the a_0 and E_0 values are always constrained to maximum and minimum thresholds, respectively. Thus, perovskites with a short lattice constant and high cohesive energy tend to present a high bulk modulus. This reflects the inverse relationship of the bulk modulus with the lattice constant and the direct relationship of the bulk modulus with cohesive energy.[52] This analysis illustrates how physical insights can be obtained from the key features identified by SGD.

The rules associated with $\mathrm{SG}_{6}^{\mathrm{m}}$ and $\mathrm{SG}_{100}^{\mathrm{m}}$ are presented in the coordinates of the key parameters a_0 and E_0 in Fig. 2(D). We also present, in this figure, the rules associated with $\mathrm{SG}_{61}^{\mathrm{m}}$, as an example of an SG that belongs to the orange cluster in Fig. 2(B) whose rules only depend on a_0 and E_0 – see Table 2. In this plot, the bulk-modulus values are indicated by the grey scale color of the circles. The figure shows graphically that a more focused description is achieved as the utility-function values increase (and the SG size decreases) within the Pareto region. This figure also highlights that more focused rules might arise at the expense of missing some high-bulk-modulus materials. For instance, some dark-grey circles corresponding to high-bulk-modulus materials are outside the limits of $\mathrm{SG}_{6}^{\mathrm{m}}$. Thus, these high-bulk-modulus materials are not captured by $\mathrm{SG}_{6}^{\mathrm{m}}$.

To understand which high-bulk-modulus materials might be "missed" in $\mathrm{SG}_{6}^{\mathrm{m}}$ as an example of an SG with a high utility function, we verified whether the 5% materials with the highest bulk moduli of the dataset are contained in $\mathrm{SG}_{6}^{\mathrm{m}}$. These are 26 perovskites presenting bulk moduli higher than 1.43 eV Å^{-3} and composed of the B elements chromium, manganese, iron, cobalt, and tungsten, and the A elements scandium, praseodymium, neodymium, cerium, promethium, yttrium, samarium, and beryllium. 24 of these 26 materials satisfy the rules associated with $\mathrm{SG}_{6}^{\mathrm{m}}$. The two high-bulk-modulus materials that do not satisfy the rules of $\mathrm{SG}_{6}^{\mathrm{m}}$ are $BeMnO_3$ and $BeWO_3$, with bulk moduli of 1.43 and 1.45 eV Å^{-3}, respectively. These two materials present the lowest cohesive energies (6.63 and 7.47 eV per atom, respectively) among the 26 materials. They are shown as lime crosses in Fig. 2(D). Thus, they do not satisfy the inequality $E_0 > 7.48$ eV per atom, which is part of the rules of $\mathrm{SG}_{6}^{\mathrm{m}}$ (Table 2). The bulk modulus for these two materials could be governed by a different mechanism compared to the materials that are part of $\mathrm{SG}_{6}^{\mathrm{m}}$. We will analyze this

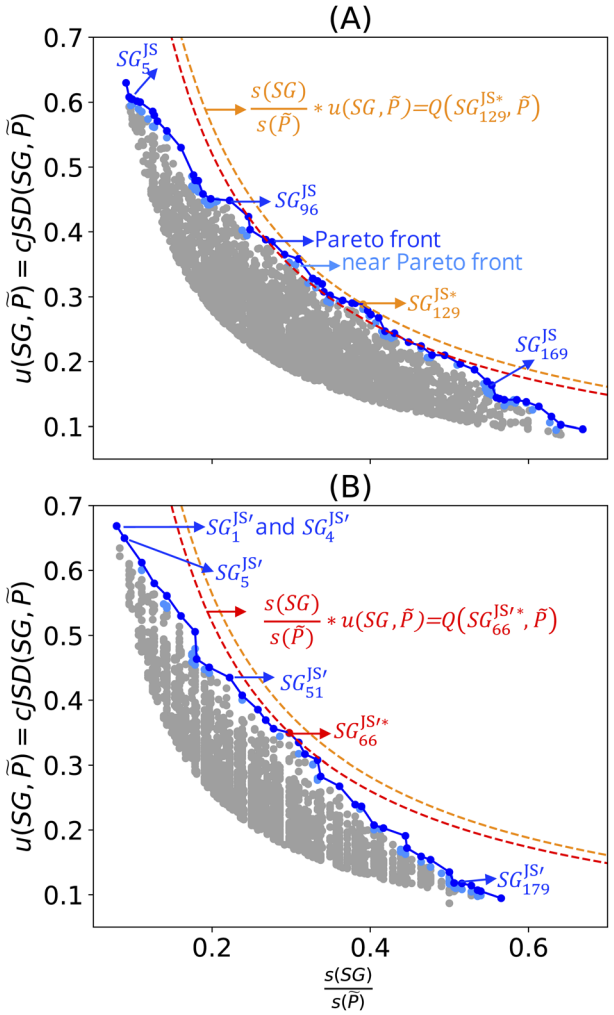

Fig. 3 Collections of SGs describing perovskites with a high bulk modulus using the cumulative formulation of the Jensen–Shannon divergence as the utility function. The 5000 SGD solutions with high values of the objective function are shown in grey and the Pareto region is displayed in blue. The SG associated with the maximum value of the objective function is shown in orange or red. The two panels show the results for two different sets of features. (A) Results for the full feature set containing the 24 features in Table 1. (B) Results for the reduced feature set containing 22 atomic and compositional features (see Table 1). The rules associated with the SGs indicated in the figure are shown in Table 2.

unexpected pattern in more detail in a dedicated subsection of the manuscript (see below).

We evaluated the variability of the identified SGs with respect to the dataset size by training the SGD with random selections of 75%, 50%, and 25% of the dataset. Even though the similarity between the SG identified based on the entire dataset and the SG identified based on a fraction of the dataset (measured using the Jaccard similarity index) decreases with decreasing data-set size, the SGs obtained with only 25% of the dataset and presenting large relative sizes are significantly similar compared to the SGs obtained with the entire dataset. The SGs obtained with 25% of the dataset and presenting low relative sizes, however, are significantly different from the SGs obtained with such relative sizes using the entire dataset. Thus, for the problem under consideration, SGD is efficient with up to 50% less data. More details of this analysis can be found in the ESI.†

3.1.2 Collection of SG rules obtained with the cumulative-Jensen–Shannon-divergence utility function. We now turn our attention to the results obtained with a utility function based on the Jensen–Shannon divergence. The information-theoretic Jensen–Shannon divergence (D_{JS}) is a symmetrized version of the Kullback–Leibler divergence (D_{KL}), also known as relative entropy. The Jensen–Shannon divergence between the discrete distributions R and S is defined as

$$D_{JS}(R,S) = \frac{1}{2}D_{KL}(R,M) + \frac{1}{2}D_{KL}(S,M) = \frac{1}{2}\sum_{x\in\chi}R(x)\log\left(\frac{R(x)}{M(x)}\right) + \frac{1}{2}\sum_{x\in\chi}S(x)\log\left(\frac{S(x)}{M(x)}\right), \quad (3)$$

where $M(x) = \frac{R(x)+S(x)}{2}$, and χ indicates the sample space. The Jensen–Shannon divergence measures the dissimilarity between two distributions. It assumes small values for similar distributions and increases as the distributions are shifted with respect to each other. The value of the Jensen–Shannon divergence also increases if two distributions have different narrownesses. In our SGD analysis, the cumulative formulation of the Jensen–Shannon divergence,[41] denoted D_{cJS}, is used as the utility function. The divergence is evaluated between the distribution of the target values in the SG and the distribution of the target values in the entire dataset. This utility function favors the selection of SGs presenting distributions of target values that are shifted and narrower compared to the distribution of the entire dataset. However, it does not explicitly require the values of the target in the SG to be high or low. Moreover, the cumulative Jensen–Shannon divergence does not make assumptions on the shape of the distributions. Thus, it can handle distributions that deviate significantly from a Gaussian more efficiently than utility functions based on the mean shift. We stress that other measures of similarity between distributions such as the Bhattacharyya distance can be used as utility functions in SGD.

The results obtained with the cumulative Jensen–Shannon-divergence utility function and with the full set of the 24 features are shown in Fig. 3(A) and in Table 2. The Pareto front and region contain 101 and 189 SGs, respectively. The SGs identified in the Pareto region contain materials with a high bulk modulus and they present narrow distributions of target values. We analyzed the selectors defining some of the SGs of this Pareto region. The rules associated with SG_5^{JS}, SG_{96}^{JS}, SG_{129}^{JS*}, and SG_{169}^{JS} constrain the values of the key features: the equilibrium lattice constant (a_0), cohesive energy (E_0), the radius of valence-s orbitals of +1 cations (cat) of the B element ($r_{s,B}^{cat}$), the Kohn–Sham single-particle eigenvalue of the lowest-unoccupied orbital of the B atom ($\varepsilon_{L,B}$), the expected oxidation state of B in the perovskite (n_B), the radius of the highest-occupied orbital of A and B neutral atoms ($r_{val,A}$ and $r_{val,B}$, respectively), and the electron affinity of the element B (EA_B). Therefore, the rules highlight that the lattice constant and the cohesive energy are

important parameters for describing high-bulk-modulus materials, along with atomic and compositional properties that mainly reflect the nature of the B element of the ABO_3 perovskites.

Let us now compare the results obtained with the positive-mean-shift with the results obtained with the cumulative-Jensen–Shannon-divergence utility functions. The Pareto region of SGs identified based on the positive-mean-shift utility function is associated with relative SG sizes in the approximate range [0.20, 0.80] (Fig. 2(A)). The range of relative SG sizes in the Pareto region identified for the case of the cumulative-Jensen–Shannon-divergence utility function is, in turn, *ca.* [0.10, 0.70] (Fig. 3(A)). Thus, optimal SGs with relative SG sizes below 0.20, *i.e.*, SGs that contain less than 20% of the dataset, could only be obtained with the utility function based on the Jensen–Shannon divergence. These SGs with small size are associated with distributions of bulk-modulus values that are narrower and more shifted towards high values than those associated with the SGs identified with the positive-mean-shift utility function. For instance, the materials selected in SG_5^{JS} and SG_{96}^{JS} present standard deviations of bulk-modulus values of 0.04 and 0.06 eV Å^{-3}, respectively. The mean values of the bulk modulus among the perovskites in these two SGs are 1.42 and 1.37 eV Å^{-3}, respectively. None of the SGs identified with the positive mean shift present higher mean values or lower standard deviation values (see Table 2). Thus, the rules corresponding to small SGs identified with the cumulative Jensen–Shannon divergence are more focused on the high bulk modulus. This can be related to the fact that only this utility function explicitly favors narrow SGs.

The SG rules identified using the cumulative-Jensen–Shannon-divergence utility function contain, in general, more statements and more features compared to the SG rules identified using the positive mean shift (Table 2). However, we note that the equilibrium lattice constant, the cohesive energy, and the radii of B atoms are identified as key features by both approaches. The cumulative Jensen–Shannon-divergence utility function provides more focused rules for the present dataset and it is thus better than the positive-mean-shift for the purpose of identifying exceptional perovskites with very high bulk moduli. However, we stress that this utility function does not explicitly require low or high values of the target. This is a disadvantage compared to the positive-mean-shift utility function. Dispersion-corrected utility functions simultaneously take into account positive or negative shifts of the mean (or of the medians) of target values and the narrowness of the distributions of targets in the SG. These utility functions were proposed in order to simultaneously incorporate the requirements for a shift in a specific direction and for small narrowness.[31]

3.1.3 SG rules obtained with a reduced set of features. So far, we have used the entire set of the 24 features in Table 1 to obtain SGs of perovskites with a high bulk modulus. SGD identified the equilibrium lattice constant (a_0) and the cohesive energy (E_0) among the key features required for describing high-bulk-modulus materials. However, in order to calculate a_0 and E_0 using DFT, the geometry of the materials needs to be optimized. This optimization corresponds to the majority of the work needed to calculate the bulk modulus itself. Thus, from the standpoint of exploring a large materials space, a_0 and E_0 are impractical (expensive) features, since one needs to evaluate these quantities for the materials under consideration in order to apply the SG rules. In order to obtain SG rules that describe high-bulk-modulus perovskites based on easily accessible features, we have also considered a reduced set of 22 features by excluding a_0 and E_0. Thus, we only considered the atomic and compositional features. In addition to this crucial cost aspect, this analysis also illustrates how the SG rules change when the feature set changes and, in particular, when important features are not included in SGD. The identification of appropriate rules based solely on atomic and compositional features is a remarkable challenge for SGD, since the relationship between these basic features and the bulk modulus is significantly more indirect compared to the relationship between the bulk modulus and a_0 or E_0.[52]

We identified the Pareto region of SGs using the reduced set of 22 features and the cumulative Jensen–Shannon-divergence utility function. The SGs identified with this approach are denoted as $SG^{JS'}$, where $'$ indicates the reduced feature set. The results are shown in Fig. 3(B) and in Table 2. The identified Pareto front and near the Pareto front contain 66 and 136 SGs, respectively. The maximum value of the objective function obtained with the reduced feature set is slightly lower than that obtained with the full feature set (0.10 and 0.11, respectively, see orange and red dashed lines in Fig. 3(A) and (B)). This indicates that the quality of the SG description is lower with the reduced feature set. The SG identified based on the reduced set of features that displays the maximum objective function, denoted as $SG_{66}^{JS'^*}$ in Table 2 and Fig. 3(B), contains 150 materials. This SG contains fewer materials than the SG identified with all the features $SG_{129}^{JS^*}$ (197 materials). However, the overlap between the two SGs is significant. Indeed, 125 materials are present in both SGs. Thus, there is a high similarity between both descriptions.

SG rules focusing on high-bulk-modulus materials in the low relative size portion of the Pareto region were also obtained with this reduced feature set. For instance, $SG_1^{JS'}$, $SG_4^{JS'}$, and $SG_5^{JS'}$ display mean bulk-modulus values of 1.43, 1.43, and 1.42 eV Å^{-3} and standard deviations of 0.03, 0.04 and 0.04 eV Å^{-3}, respectively. These figures are similar to those associated with SG_5^{JS}, which was identified based on the entire set of the 24 features. The SGs $SG_5^{JS'}$ and SG_5^{JS} contain, respectively, 45 and 48 materials. 40 materials are present in both SGs. This reflects a significant similarity between both descriptions. Thus, the SGs with small relative size identified in the Pareto region with the reduced set of features are comparable to those obtained with the full feature set. The rules associated with the SGs identified using the reduced set of 22 features depend on some key parameters that were also identified by the analysis of the entire set of the 24 features and the cumulative-Jensen–Shannon-divergence utility function, *e.g.*, radii of A and B elements, Kohn–Sham single-particle eigenvalue of the lowest-unoccupied orbital of the B atom ($\varepsilon_{L,B}$), electron affinity of the B atom (EN_B), and expected oxidation state of the B element in

the perovskite. However, some additional key features are identified in the analysis of the reduced set of features, such as the ionization potential of the B atom (IP_B), the atomic charge of the B element (Z_B), the expected oxidation state of the A element in the perovskite (n_A), and the electronegativity of the A element (EN_A). This shows how SGD attempts to reconstruct the information contained in the important features a_0 and E_0 by using the information on other offered features. Overall, the results obtained with the reduced feature set are comparable to those obtained with all 24 features. The rules derived based on the reduced feature set can thus be applied to identify high-bulk-modulus materials in larger materials spaces compared to the training set.

3.1.4 Exploitation of the SG rules for the identification of perovskites with a high bulk modulus. We applied the SG rules trained on 504 single ABO_3 perovskites with the reduced set of 22 features to identify materials with a high bulk modulus (B_0) out of a candidate materials space of 12 096 compounds. This candidate material space was created by considering additional A and B elements that were not included in the training set (A: thorium and protactinium; B: hafnium, rhenium, osmium, iridium, gold, mercury, and thallium). Additionally, we combined two different B elements to form double perovskites with the formula $A_2BB'O_6$. In the case of the double perovskites, the features related to the B element are defined as the (composition) average of the features associated with the two different B and B' elements. In this analysis, we are explicitly considering a finite set of $A_2BB'O_6$ materials that contain a 1 : 1 B : B' stoichiometric ratio. However, the material space of double perovskites is practically infinite, since any proportion of B and B' is possible, *i.e.*, any $A_2B_{2-x}B'_xO_6$ formula with x in the range [0.0, 2.0] corresponds to a material in this space. In order to assess the usefulness of SG rules identified with the Pareto-region approach with respect to the SG rules associated with the maximum value of the objective function, we applied two different sets of SG rules to select materials from the candidate materials space that likely present high bulk moduli. The first set of rules corresponds to $SG_{55}^{JS'*}$, which are associated with the SG with maximum value of the objective function. These rules are satisfied by 4518 of the 12 096 candidates. The second set of rules corresponds to $SG_1^{JS'}$, $SG_4^{JS'}$, and $SG_5^{JS'}$, which are associated with the SGs in the Pareto region with high utility-function values, *i.e.*, high exceptionality. These rules are satisfied by 238 materials, out of the 12 096 candidates, *i.e.*, 1.97% of the candidate materials space. Then, we randomly selected 50 of the 4518 selected materials and 50 of the 238 selected materials and evaluated their B_0 using DFT-PBEsol calculations. For comparison, we have also calculated the B_0 of 50 perovskites that were randomly selected from the 12 096 materials.

The distribution of B_0 for the materials that were randomly selected from the candidate materials space (Fig. 4, in grey) has practically the same mean value compared to that of the distribution of B_0 in the training set (Fig. 4, in black), *i.e.*, 1.09 eV Å^{-3}. This indicates that the training data might be representative of this specific candidate materials space. The highest B_0 among the materials randomly selected from the candidate materials space is 1.36 eV Å^{-3}. This value is significantly lower than the highest B_0 in the training dataset, 1.49 eV Å^{-3} for $ScMnO_3$.

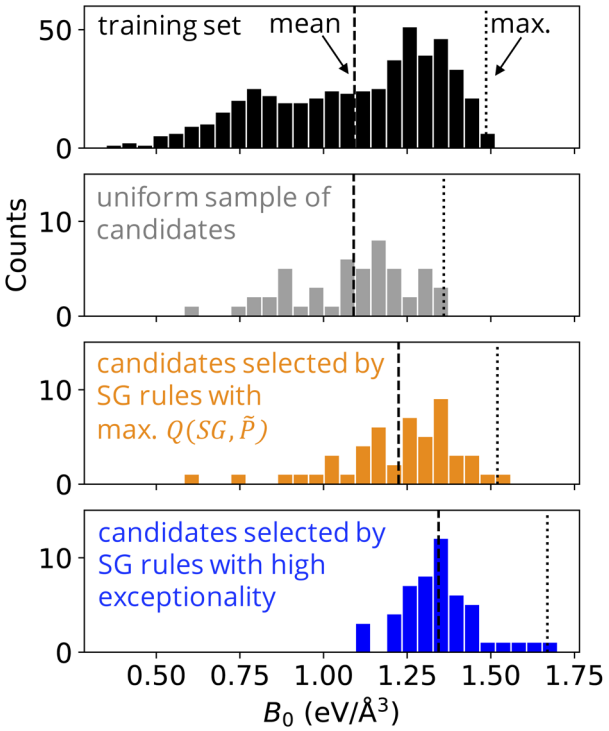

Fig. 4 The SG rules describing perovskites with a high bulk modulus (B_0) trained on 504 single ABO_3 perovskites are applied to identify promising single ABO_3 and double $A_2BB'O_6$ perovskites from a candidate space containing 12 096 materials. The histograms show the distribution of B_0 among the materials of the training dataset (in black), among 50 materials randomly selected from the candidate space (in grey), among 50 materials of the candidate space selected according to the SG rules $SG_{66}^{JS'*}$ (in orange), and among 50 materials of the candidate space suggested by the SG rules $SG_1^{JS'}$, $SG_4^{JS'}$, and $SG_5^{JS'}$ (in blue). The dashed and dotted lines indicate the mean and maximum (max.) B_0 values of each distribution, respectively.

The distribution of B_0 for the materials that were selected from the candidate materials space using the SG rules $SG_{55}^{JS'*}$ (Fig. 4, in orange) is concentrated in high B_0, with a mean B_0 value of 1.22 eV Å^{-3}. One material suggested by these SG rules has B_0 higher than the highest value in the training dataset, $PaCoOsO_6$, with a bulk modulus of 1.52 eV Å^{-3}, respectively. This value is slightly higher than the highest B_0 in the training dataset (1.49 eV Å^{-3}). However, we note that the rules suggested several materials that turned out to have relatively low values of B_0.

The distribution of B_0 for the materials that were selected from the candidate materials space using the SG rules $SG_1^{JS'}$, $SG_4^{JS'}$, and $SG_5^{JS'}$ (Fig. 4, in blue) is concentrated in higher values compared to the other distributions, with a mean B_0 value of 1.35 eV Å^{-3}. Additionally, four of the materials suggested by the SG rules have B_0 higher than the highest value in the training dataset. These are $PaMnO_3$, Pa_2CrFeO_6, Pa_2VCrO_6, and $PaVO_3$, with a bulk modulus of 1.67, 1.63, 1.59, and 1.55 eV Å^{-3}, respectively. Thus, the SG rules lead to the identification of materials with B_0 up to 13% higher than any observation in the

training dataset. This result indicates that the SGs identified by the Pareto-region analysis can point at more exceptional materials compared to the standard SGD approach. In particular, in the present use case, materials that present higher performance than the known compounds of the training set were more efficiently selected using the focused SG rules provided by the multi-objective optimization of SGs. We note that in previous studies SGD was compared with global approaches such as decision trees in the context of materials science applications.[34,55] The results indicate that SGD is more efficient in describing exceptional situations compared to the decision-tree approach. This can be related to the fact that the objective function of the decision tree aims at a good performance on average, while the objective function of SGD eqn (1) can focus on statistically exceptional situations.

The dataset utilized to train SGD and the verification of SGD suggestions are based on DFT-PBEsol calculations. Even though DFT-PBEsol presents an overall good accuracy for describing properties of solid materials such as the bulk modulus,[56] the errors of DFT-PBEsol should be taken into account when comparing the reported bulk moduli with experimental results.

3.1.5 Investigation of high-bulk-modulus perovskites that are not captured using the SG rules. When analyzing the SG rules, we highlighted that $BeMnO_3$ and $BeWO_3$ present bulk moduli above the 95%-ile of the bulk-modulus distribution in the training set (1.43 and 1.45 eV Å^{-3}, respectively) but they are not contained in the high-utility-function SG SG_6^m identified in Fig. 2. In the perovskite structure, the *A* cations are larger than the *B* cations and the relationships between the radii of the *A* and *B* cations and the anions determine the thermodynamic stability of the perovskite structure, as given by the tolerance factors.[48,57] Beryllium and magnesium are the only two *A* elements in our dataset for which the radius of *A* (*e.g.*, $r_{s,A}$) might be smaller than the radius of *B* (*e.g.*, $r_{s,B}$). This is the case for the materials $BeMnO_3$ and $BeWO_3$. Indeed, $r_{s,Be} < r_{s,Mn}$ and $r_{s,Be} < r_{s,W}$. Thus, beryllium would most likely occupy the *B* site of these cubic perovskite structures. Indeed, several reports discuss the properties of perovskites composed of beryllium at the *B* sites, *i.e.*, coordinated with 6 oxygen or halide anions.[58–60] This observation motivated us to evaluate the bulk modulus of the perovskites $MnBeO_3$ and $WBeO_3$, where beryllium sits at the *B* site and manganese or tungsten sits at the *A* sites. The calculated bulk moduli are equal to 1.44 and 1.82 eV Å^{-3}. The bulk modulus of $WBeO_3$ is 22% higher than the highest value in the training set and is the highest bulk modulus identified in this paper. We have also evaluated the bulk modulus of perovskites with *B* = Be and other transition metals of the third row of the periodic table as *A* elements, namely hafnium, tantalum, rhenium, osmium and iridium. The bulk modulus of the materials $HfBeO_3$, $TaBeO_3$, $ReBeO_3$, $OsBeO_3$, and $IrBeO_3$ are equal to 1.70, 1.81, 1.75, 1.63, and 1.57 eV Å^{-3}, respectively. These values are also relatively high compared to the values of the training set and they show that several beryllium based materials might be exceptional. This analysis illustrates how the exploratory nature of SGD analysis can identify unexpected patterns and anomalies, which might lead in turn to the identification of exceptional materials.

We note that the rules identified with SGD will be valid as long as the physical processes governing the materials in the training dataset also govern the behavior of the materials in the materials space to be explored. This is a crucial aspect, since the choice of the materials in the training dataset is often influenced by bias and the number of materials in the training set is very small compared to the practically infinite space of possible materials. In order to cover portions of the materials space where underlying processes different from those present in the training set are important, the incorporation of new data points and retraining of SG rules will be required. Indeed, exceptional SGs and phenomena might emerge in regions of the data space that are not sufficiently covered by the training dataset. Additionally, the identified SGs can be associated with genuinely exceptional phenomena, but they might also correspond to measurement artifacts when the data are generated by an experiment or calculation subjected to noise.[61] These two situations would not be distinguished by SGD. However, by analyzing the SG rules and key identified parameters, one might be able to judge whether SGD identified correlations that have a physical meaning. For instance, the rules derived in Fig. 2(D) reflect that stronger bonds between atoms in the crystal result in short lattice constants, high cohesive energy, and a high bulk modulus. Additionally, SGD models for materials rely on the fact that the offered features correlate with the underlying physical processes governing the materials. Thus, the choice of features is critical. The performance of SGD can be assessed by cross-validation, as described in ref. 35. Finally, useful materials might present unusual combinations of different materials properties. These could also be considered exceptional materials. Identifying such materials calls for multi-objective optimization of materials properties.[62–65] SGD can be adapted for this scenario and this aspect will be addressed in an upcoming contribution.

4 Conclusions

We introduced an approach for the identification of coherent collections of SGs of the "Pareto region" with respect to the SG size and exceptionality objectives of the SGD analysis. The concept was demonstrated by the learning of rules that describe perovskites with a high bulk modulus. Our results show that rules focused on exceptional materials do not necessarily correspond to the one SG that maximizes the objective function, but these rules can be identified with the Pareto-region concept. This analysis does not require additional computational effort, since the SGD solutions with high objective-function values are obtained on the fly during the optimization of the objective function. We used the SG rules obtained by the multi-objective approach to identify exceptional perovskites with the bulk modulus up to 13% higher than the highest value found in the training set of 504 materials, out of a materials space of more than 12 000 materials.

Data availability

All input and output files of the DFT calculations and datasets are available in ref. 52 and 66. The SGD analysis is available at

https://github.com/lfoppa/Multi-objective-optimization-of-subgroups.

Author contributions

L. F. conceived the project and performed the SGD analysis and DFT calculations. L. F. and M. S. wrote the manuscript jointly.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

This work was funded by the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233). Open Access funding was provided by the Max Planck Society.

Notes and references

1 N. S. Lewis and D. G. Nocera, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 15729–15735.
2 S. Chu and A. Majumdar, *Nature*, 2012, **488**, 294–303.
3 D. Davies, K. Butler, A. Jackson, A. Morris, J. Frost, J. Skelton and A. Walsh, *Chem*, 2016, **1**, 617–627.
4 J. Schrier, A. J. Norquist, T. Buonassisi and J. Brgoch, *J. Am. Chem. Soc.*, 2023, **145**, 21699–21716.
5 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput. Mater.*, 2017, **3**, 54.
6 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.
7 D. Raabe, J. R. Mianroodi and J. Neugebauer, *Nat. Comput. Sci.*, 2023, **3**, 198–209.
8 S. Bauer, P. Benner, T. Bereau, V. Blum, M. Boley, C. Carbogno, C. R. A. Catlow, G. Dehm, S. Eibl, R. Ernstorfer, Á. Fekete, L. Foppa, P. Fratzl, C. Freysoldt, B. Gault, L. M. Ghiringhelli, S. K. Giri, A. Gladyshev, P. Goyal, J. Hattrick-Simpers, L. Kabalan, P. Karpov, M. S. Khorrami, C. T. Koch, S. Kokott, T. Kosch, I. Kowalec, K. Kremer, A. Leitherer, Y. Li, C. H. Liebscher, A. J. Logsdail, Z. Lu, F. Luong, A. Marek, F. Merz, J. R. Mianroodi, J. Neugebauer, Z. Pei, T. A. R. Purcell, D. Raabe, M. Rampp, M. Rossi, J.-M. Rost, J. Saal, U. Saalmann, K. N. Sasidhar, A. Saxena, L. Sbailò, M. Scheidgen, M. Schloz, D. F. Schmidt, S. Teshuva, A. Trunschke, Y. Wei, G. Weikum, R. P. Xian, Y. Yao, J. Yin, M. Zhao and M. Scheffler, *Modell. Simul. Mater. Sci. Eng.*, 2024, **32**, 063301.
9 L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski, R. Schlögl, A. Trunschke and M. Scheffler, *MRS Bull.*, 2021, **46**, 1016–1026.
10 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.
11 S. K. Kauwe, J. Graser, R. Murdock and T. D. Sparks, *Comput. Mater. Sci.*, 2020, **174**, 109498.
12 E. S. Muckley, J. E. Saal, B. Meredig, C. S. Roper and J. H. Martin, *Digit. Discov.*, 2023, **2**, 1425–1435.
13 K. Li, B. DeCost, K. Choudhary, M. Greenwood and J. Hattrick-Simpers, *npj Comput. Mater.*, 2023, **9**, 55.
14 K. Li, A. N. Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A. B. Dieng and J. Hattrick-Simpers, *Commun. Mater.*, 2025, **6**, 9.
15 Y. Wang, N. Wagner and J. M. Rondinelli, *MRS Commun.*, 2019, **9**, 793–805.
16 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, *Proc. IEEE*, 2016, **104**, 148–175.
17 H. Zhang, W. W. Chen, J. M. Rondinelli and W. Chen, *Appl. Phys. Rev.*, 2023, **10**, 021403.
18 A. E. Siemenn, Z. Ren, Q. Li and T. Buonassisi, *npj Comput. Mater.*, 2023, **9**, 79.
19 A. Biswas, Y. Liu, N. Creange, Y.-C. Liu, S. Jesse, J.-C. Yang, S. V. Kalinin, M. A. Ziatdinov and R. K. Vasudevan, *npj Comput. Mater.*, 2024, **10**, 29.
20 L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay and C. W. Coley, *J. Chem. Inf. Model.*, 2020, **60**, 3770–3780.
21 G. Palmer, S. Du, A. Politowicz, J. P. Emory, X. Yang, A. Gautam, G. Gupta, Z. Li, R. Jacobs and D. Morgan, *npj Comput. Mater.*, 2022, **8**, 115.
22 K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing and Z. W. Ulissi, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 025006.
23 C. K. H. Borg, E. S. Muckley, C. Nyby, J. E. Saal, L. Ward, A. Mehta and B. Meredig, *Digit. Discov.*, 2023, **2**, 327–338.
24 G. Serraino and S. Uryasev, in *Conditional Value-at-Risk (CVaR)*, ed. S. I. Gass and M. C. Fu, Springer US, Boston, MA, 2013, pp. 258–266.
25 A. Seko, H. Hayashi, H. Kashima and I. Tanaka, *Phys. Rev. Mater.*, 2018, **2**, 013805.
26 M. Kuban, Š. Gabaj, W. Aggoune, C. Vona, S. Rigamonti and C. Draxl, *MRS Bull.*, 2022, **47**, 991–999.
27 H. Jia, M. Horton, Y. Wang, S. Zhang, K. A. Persson, S. Meng and M. Liu, *Adv. Sci.*, 2022, **9**, 2202756.
28 S. Wrobel, *European Conference on Principles of Data Mining and Knowledge Discovery*, 1997.
29 J. H. Friedman and N. I. Fischer, *Stat. Comput.*, 1999, **9**, 123.
30 B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler and L. M. Ghiringhelli, *New J. Phys.*, 2017, **19**, 013031.
31 M. Boley, B. R. Goldsmith, L. M. Ghiringhelli and J. Vreeken, *Data Min. Knowl. Discovery*, 2017, **31**, 1391–1418.
32 C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken and M. Scheffler, *Nat. Commun.*, 2020, **11**, 4428.
33 H. Li, Y. Liu, K. Chen, J. T. Margraf, Y. Li and K. Reuter, *ACS Catal.*, 2021, **11**, 7906–7914.
34 L. Foppa and L. M. Ghiringhelli, *Top. Catal.*, 2022, **65**, 196.
35 L. Foppa, C. Sutton, L. M. Ghiringhelli, S. De, P. Löser, S. A. Schunk, A. Schäfer and M. Scheffler, *ACS Catal.*, 2022, **12**, 2223.

36 P. K. Novak, N. Lavrač and G. I. Webb, in *Supervised Descriptive Rule Induction*, ed. C. Sammut and G. I. Webb, Springer US, Boston, MA, 2010, pp. 938–941.
37 M. Boley, C. Lucchese, D. Paurat and T. Gärtner, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2011, pp. 582–590.
38 M. Boley, S. Moens and T. Gärtner, *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012, pp. 69–77.
39 H. Grosskreutz, S. Rüping and S. Wrobel, *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, 2008, pp. 440–456.
40 H. I. Rivera-Arrieta and L. Foppa, *ACS Catal.*, 2025, **15**, 2916–2926.
41 H.-V. Nguyen and J. Vreeken, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II 15*, 2015, pp. 173–189.
42 M. A. Peña and J. L. G. Fierro, *Chem. Rev.*, 2001, **101**, 1981–2018.
43 W. Travis, E. N. K. Glover, H. Bronstein, D. O. Scanlon and R. G. Palgrave, *Chem. Sci.*, 2016, **7**, 4548–4556.
44 A. K. Jena, A. Kulkarni and T. Miyasaka, *Chem. Rev.*, 2019, **119**, 3036–3103.
45 J. Hwang, R. R. Rao, L. Giordano, Y. Katayama, Y. Yu and Y. Shao-Horn, *Science*, 2017, **358**, 751–756.
46 J. Y. Kim, J.-W. Lee, H. S. Jung, H. Shin and N.-G. Park, *Chem. Rev.*, 2020, **120**, 7867–7918.
47 G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Sci. Rep.*, 2016, **6**, 19375.
48 C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M. Scheffler, *Sci. Adv.*, 2019, **5**, eaav0693.
49 Q. Tao, P. Xu, M. Li and W. Lu, *npj Comput. Mater.*, 2021, **7**, 23.
50 A. Ihalage and Y. Hao, *npj Comput. Mater.*, 2021, **7**, 75.
51 G. H. Gu, J. Jang, J. Noh, A. Walsh and Y. Jung, *npj Comput. Mater.*, 2022, **8**, 71.
52 L. Foppa, T. A. R. Purcell, S. V. Levchenko, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Lett.*, 2022, **129**, 055301.
53 U. Niemann, M. Spiliopoulou, B. Preim, T. Ittermann and H. Völzke, *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 582–587.
54 F. Nielsen, in *Hierarchical Clustering*, Springer International Publishing, Cham, 2016, pp. 195–211.
55 A. Mazheika, Y.-G. Wang, R. Valero, F. Viñes, F. Illas, L. M. Ghiringhelli, S. V. Levchenko and M. Scheffler, *Nat. Commun.*, 2022, **13**, 419.
56 G.-X. Zhang, A. M. Reilly, A. Tkatchenko and M. Scheffler, *New J. Phys.*, 2018, **20**, 063020.
57 V. M. Goldschmidt, *Naturwissenschaften*, 1926, **14**, 477–485.
58 C. Li, X. Lu, W. Ding, L. Feng, Y. Gao and Z. Guo, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2008, **64**, 702–707.
59 Q. Mahmood, M. Hassan, M. Yaseen and A. Laref, *Chem. Phys. Lett.*, 2019, **729**, 11–16.
60 P. Kumari, V. Srivastava, R. Sharma, N. Kaur and H. Ullah, *Mater. Today Commun.*, 2024, **40**, 109571.
61 S. B. Harris, R. Vasudevan and Y. Liu, *npj Comput. Mater.*, 2025, **11**, 23.
62 Z. del Rosario, M. Rupp, Y. Kim, E. Antono and J. Ling, *J. Chem. Phys.*, 2020, **153**, 024112.
63 K. M. Jablonka, G. M. Jothiappan, S. Wang, B. Smit and B. Yoo, *Nat. Commun.*, 2021, **12**, 2312.
64 B. Shi, T. Lookman and D. Xue, *Mater. Genome Eng. Adv.*, 2023, **1**, e14.
65 A. K. Y. Low, F. Mekki-Berrada, A. Gupta, A. Ostudin, J. Xie, E. Vissol-Gaudin, Y.-F. Lim, Q. Li, Y. S. Ong, S. A. Khan and K. Hippalgaonkar, *npj Comput. Mater.*, 2024, **10**, 104.
66 L. Foppa, T. Purcell, S. Levchenko, M. Scheffler and L. M. Ghiringhelli, 2022, DOI: **10.17172/NOMAD/2022.02.21-3**.