Uncovering structure-property relationships of materials by subgroup discovery

2017 New J. Phys. 19 013031

(http://iopscience.iop.org/1367-2630/19/1/013031)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
30 November 2016

REVISED
5 January 2017

ACCEPTED FOR PUBLICATION
9 January 2017

PUBLISHED
25 January 2017

PAPER

Uncovering structure-property relationships of materials by subgroup discovery

Bryan R Goldsmith[1,3], **Mario Boley**[1,2,3], **Jilles Vreeken**[2], **Matthias Scheffler**[1] **and Luca M Ghiringhelli**[1]

[1] Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany
[2] Max Planck Institute for Informatics, Campus Mitte, D-66123 Saarbrücken, Germany
[3] These authors have contributed equally to this manuscript.

E-mail: goldsmith@fhi-berlin.mpg.de **and** ghiringhelli@fhi-berlin.mpg.de

Keywords: big-data analytics, data mining, pattern discovery, machine learning, octet binary semiconductors, gold clusters

Supplementary material for this article is available online

Abstract

Subgroup discovery (SGD) is presented here as a data-mining approach to help find interpretable local patterns, correlations, and descriptors of a target property in materials-science data. Specifically, we will be concerned with data generated by density-functional theory calculations. At first, we demonstrate that SGD can identify physically meaningful models that classify the crystal structures of 82 octet binary (OB) semiconductors as either rocksalt or zincblende. SGD identifies an interpretable two-dimensional model derived from only the atomic radii of valence *s* and *p* orbitals that properly classifies the crystal structures for 79 of the 82 OB semiconductors. The SGD framework is subsequently applied to 24 400 configurations of neutral gas-phase gold clusters with 5–14 atoms to discern general patterns between geometrical and physicochemical properties. For example, SGD helps find that van der Waals interactions within gold clusters are linearly correlated with their radius of gyration and are weaker for planar clusters than for nonplanar clusters. Also, a descriptor that predicts a local linear correlation between the chemical hardness and the cluster isomer stability is found for the even-sized gold clusters.

1. Introduction

Rational design of advanced functional materials, e.g., active and selective catalysts [1], efficient thermoelectrics [2], and high-temperature superconductors [3], requires an understanding of the underlying fundamental physical mechanisms. Identifying interpretable[4] [4] rule-based models that describe materials phenomena is therefore critical. For example, Brønsted–Evans–Polanyi relations allow for an efficient approach to estimate activation energies of similar reactions [5], the Goldschmidt tolerance factor is an indicator for the stability and structure of ionic crystals [6, 7], and the thermoelectric figure of merit, as well as semi-empirical models, guide the design of thermoelectrics [8]. However, in general it remains difficult to extract insights from materials-science data and to discover rules for desired materials properties and function.

Big-data analytics tools, e.g., statistical/machine learning, compressed sensing, and data mining methods, are becoming widely applied in the materials-science community [9–19]. Efficient algorithms for model selection can be used for the estimation of alloy formation energies [20], and machine-learning algorithms trained on reaction data can help predict the crystallization outcomes of materials [21]. Data analytics tools can identify descriptors [22, 23] that characterize properties such as hole traps in amorphous silicon [24] and the intrinsic dielectric breakdown field of insulators [25]. Importantly, the application of big-data analytics to obtain

[4] A model is interpretable when it is not a 'black box', i.e., the role of the descriptor (the input variables) in the descriptor-to-property mapping is explicit. In subgroup discovery (SGD) the model is a selector, i.e., a set of 'human readable' Boolean statements about the descriptor. Another example is the compressed-sensing work of some of us (see [22]), where the model is an analytical formula that maps the descriptor onto the property. Examples of less interpretable models are those based on artificial neural networks [4], Kernel Ridge Regression [4] or Gaussian processes [4].

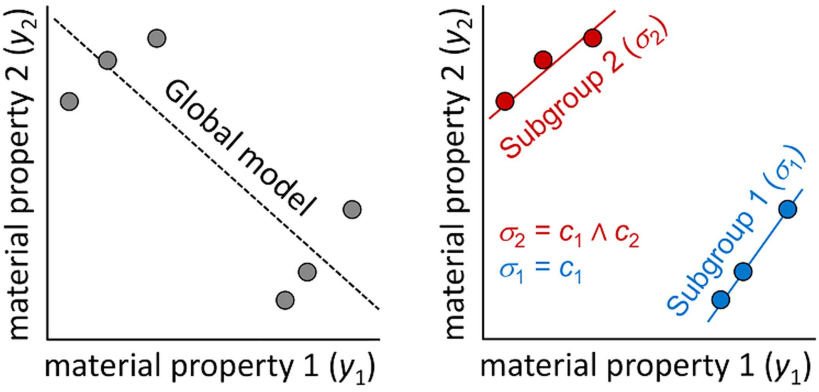

Figure 1. A schematic of an optimal global regression model predicting a negative relationship between material properties y_1 and y_2 (with nonzero error), whereas two subpopulations allow a nearly perfect fit using a linear regression function with positive slope. Subgroup discovery aims to describe such subpopulations by Boolean selector functions (σ_1 and σ_2) defined as conjunctions (logical AND, denoted as $\wedge$) of basic selectors (the c_i).

material insights and to predict novel materials can be enhanced by the availability of large materials repositories, e.g., AFLOWLIB, Computational Materials Repository, Electronic Structure Project, Materials Project, Novel Materials Discovery (NOMAD), Open Quantum Materials Database, and Pauling file [26]. Our objective is to develop and exploit big-data analytics tools to discover materials insights and to predict advanced materials from large collections of materials data stored within the NOMAD Archive [27].

Big-data analytics applied to materials-science data often focuses on the inference of a global prediction model for some property of interest for a given class of materials. However, the underlying mechanism for some target property could differ for different materials within a large pool of materials-science data. Consequently, a global model fitted to the entire dataset may be difficult to interpret and may well hide or incorrectly describe the actuating physical mechanisms [28]. In these situations, local models describing subgroups would be advantageous to global models. For illustration (see figure 1), a globally optimal regression model could predict a negative relationship between two material properties, whereas among subgroups there exists a positive relationship. As a more physical example, the transition metals of the periodic table are a subgroup, and the actinides, lanthanides, and halogens are other subgroups. Thus, identification of subgroups is useful to gain an understanding of similarities and differences between materials.

In this paper we demonstrate a multipurpose data-mining algorithm called SGD to identify and describe local patterns, correlations, and descriptors in materials-science data according to some desired target property (or properties) [29–32]. At first, we begin by formulating the SGD algorithm for materials-science applications. Next, we demonstrate that SGD can identify physically meaningful models that classify the crystal structures of 82 octet binary (OB) semiconductors as either rocksalt (RS) or zincblende (ZB) from only information of its chemical composition. The OB compounds have long been studied [22, 33–40], and we consider it an exemplary dataset for the demonstration of SGD to find descriptors of materials. Notably, SGD helps us to find a two-dimensional model derived from only the atomic radii of valence *s* and *p* orbitals that properly classify the crystal structures for 79 of the 82 OB semiconductors. Subsequently, we apply SGD to 24 400 configurations of neutral gas-phase gold clusters with 5–14 atoms. Small gold clusters have different physical and chemical properties than their bulk counterpart, and they exhibit a diverse array of physicochemical properties depending on their size and shape [41–50]. The aim of investigating gold clusters here is two-fold: (1) to search for general structure-property relationships holding across gold clusters of different sizes and vastly different configurations; and (2) to demonstrate the versatility of SGD on a large and heterogeneous dataset. It is established that SGD can help identify unexpected and general, size-independent, patterns within the dataset of gold cluster configurations.

2. Subgroup discovery

The concepts of SGD originate from the early 90s, when the advent of large databases motivated the development of explorative and descriptive analytics tools as an interpretable complement to the supervised learning (or global modeling) paradigm [28, 30, 32, 51–54]. Below we will start with a discussion of the three main components of SGD: (i) the use of a description language for identifying subpopulations within a given pool of data (section 2.1); (ii) the definition of utility functions that formalize the interestingness (quality) of subpopulations (section 2.2); and (iii) the design of a Monte Carlo search algorithm to find selectors that describe interesting subpopulations (section 2.3). As notational convention, we write $|X|$ to refer to the number

Table 1. Examples of basic statements constructed from categorical, ordinal, and metric features.

Categorical	The material has a rocksalt crystal structure	The gold cluster has a planar geometry
Ordinal	The material has $\geqslant N$ atoms per unit cell	The gold cluster has a mean atom coordination $\leqslant X$
Metric[a]	The band gap of the material is *high*	The chemical hardness of the gold cluster is *low*

[a] The notion of *high* and *low* is based on cut-off values of the metric features determined via *k*-means clustering.

of elements contained in X (the cardinality of X), and $\mathscr{P}(X)$ to refer to the set of all possible subsets of X (its power set). Moreover, we denote by $X \setminus Y$ the set of elements of X that are not contained in Y (its set difference). Logical conjunctions AND, OR, and NOT, are represented by the symbols $\wedge$, $\vee$, and $\neg$.

2.1. Description language

Consider some population P of materials that contains subgroups corresponding to subpopulations $P' \subseteq P$ that exhibit yet unknown regularities with respect to some properties of interest. The population of materials is assumed to be represented by a set of features, where each feature is a map (a function) $a\colon P \to V$ of the population into some value domain V. The implicit assumption is that features can be measured and are comparable across all materials in the population. In the context of a specific analysis question, the features are partitioned into two subsets: *description features* A and *target variables* T. The description features are used to describe subpopulations, whereas the target variables determine how important specific subpopulations are to our question being analyzed. For example, if we are interested in examining the band gaps of materials, then the target variable is the band gap and the description features are the material's properties, such as the number of atoms in a crystal unit cell, the composition, and the radii of its atomic *s* and *p* orbitals. For lists of features that we consider in our study, see tables S1 and S2 in the supporting information.

The next aspect of our analysis is the *basic selectors*. The definition of basic selectors is an important design step that determines the interpretability and interestingness of the subgroup descriptions [32]. Basic selectors are statements regarding the features such as 'the band gap of the material is large' or 'the material has a rocksalt crystal structure'. A set of basic selectors C are constructed from the description features $a\colon P \to V$ according to different rules depending on their type: categorical, ordinal and metric (see table 1 for examples). For *categorical* features, i.e., when V is a discrete set with no relevant internal order, basic selectors of the form $c(p) \equiv a(p) = v$ for all values $v \in V$ are constructed. For *ordinal* features, i.e., when V contains a set of discrete and ordered values, but the scale cannot be used to interpolate between values in a meaningful way, inequality constraints $c(p) \equiv a(p) \leqslant v$ and $c(p) \equiv a(p) \geqslant v$ for all $v \in V$ are used. For *metric* features, i.e., the values are from a continuous ordered scale that adheres to a meaningful notion of distance, selectors similar to the ordinal case are constructed. In this case, however, we cannot simply use all possible cut-off values, but instead have to find a small computationally feasible subset. Ideally, we would like to allow the SGD algorithm to either completely select or to completely deselect all groups of materials with very similar feature values. This goal can be approximated by finding cut-off values through *k*-means clustering [55]. That is, for a desired number k of cut-off values we find a set $R \subseteq V$ of $k + 1$ representative values that minimize the sum of squared differences $\sum_{p \in P}(a(p) - r_{a(p)})^2$, where r_v minimizes $|v - r|$ among $r \in R$ for some a-value v. In this way, each a-value in the population P is assigned to a cluster represented by one element in R. The cut-off values are then given as the arithmetic mean between the maximum and the minimum a-value of neighboring clusters.

Based on a final set of basic selectors C, subgroup descriptions are formed as complex Boolean *selectors* $\sigma\colon P \to \{\text{true}, \text{false}\}$ defined through conjunctions

$$\sigma(\cdot) \equiv c_1(\cdot) \wedge \ldots \wedge c_l(\cdot) \tag{1}$$

of basic statements $c_1, \ldots, c_l \in C$. Analogously to previous work by some of the authors [22], we define the *descriptor* induced by σ as the set of descriptive features that are referenced in σ.[5] The subpopulation of P that is defined by σ is called the *extension* of σ and is written as

$$\text{ext}(\sigma) = \{p \in P\colon \sigma(p) = \text{true}\}. \tag{2}$$

Although this definition yields $2^{|C|}$ possible subgroup selectors for a given set of basic selectors, usually only a few of those describe distinct and interesting subpopulations. To algorithmically determine those of interest, we have to formalize the notion of interestingness (quality) of subpopulations. As indicated above, this definition refers to the target variables T. In particular, let $Y = V_1 \times \ldots \times V_k$ denote the joint domain of all target variables. Then the utility of a selector σ depends on the collection of Y values in its extension.

[5] In Ghiringhelli *et al* the term 'descriptor' refers to the set of features referenced in a linear model. Note that, in contrast to our usage here, the SGD literature often uses the term descriptor to refer to the selector itself and not to the set of variables it contains.

2.2. Subgroup quality

The SGD literature utilizes different notions of subgroup quality depending on the type and number of target variables, as well as on the kind of patterns to be discovered [32, 51]. A shared characteristic between quality functions is that they are usually a weighted product of two factors corresponding to the relative size of the selected subpopulation and the utility of the selection. Formally, for a weight parameter $\alpha \in [0,\ 1]$ we consider quality functions of the form

$$q(\sigma) = \mathrm{cov}(\sigma)^{\alpha} u(\mathrm{ext}(\sigma))^{1-\alpha}, \tag{3}$$

where $\mathrm{cov}(\sigma) = |\mathrm{ext}(\sigma)|/|P|$ is the *coverage* of σ and u: $\mathscr{P}(P) \to \mathbb{R}$ is some *utility function*. The combination of these two factors is required because individually each of them is trivially maximized by either extremely general selectors (maximizing coverage) or extremely specific selectors (maximizing utility). An alternative view on considering size is that it plays a similar role as the regularization term in Ridge Regression or LASSO [56, 57]. We use an α-value of 0.5 unless mentioned otherwise, which puts equal importance on the generality and the utility of findings.

Regarding the utility function, the traditional focus of SGD is to look for subgroups that exhibit target values with a distribution that differs as much as possible from the distribution of the target variables in the global population. A representative example is the *(absolute) mean shift function* $u_m(S) = |m_S - m_P|$ for a single metric target variable, i.e., $T = \{a\}$ with a: $P \to \mathbb{R}$, where $m_U = \sum_{p \in U} a(p)/|U|$ is the mean value of a in a population U. Although focusing on large deviation can lead to interesting findings, it has some problems in our application context of materials datasets: (1) Deviation in itself neglects *consistency* in the sense that subgroups with a high target deviation might have a poor model fit of the target variable. For example, in a heterogeneous materials-science data set a subgroup might have a large mean shift but have a local standard deviation that is higher than the global standard deviation. This is problematic for our goal of uncovering physical relations between material structure and properties, for which we want our findings to be highly consistent. (2) Focusing on deviation has the implicit assumption that the global reference distribution is already well understood and distance from it is therefore meaningful. However, the global distribution of properties in big-data of materials is often an effect of a large number of mixed factors and therefore often too complex to describe in a compact way in fact this complexity is one of the reasons to resort to local modeling via SGD in the first place.

Therefore, the utility functions we define for this study aim for consistent findings by formalizing different notions of purity in the distribution of target values. In particular, we consider the following utility functions:

- The *(normalized) information gain* $u_{\mathrm{ig}}(S) = (H_P - H_S)/H_P$ for categorical target variables $T = \{a_1, \ldots, a_k\}$ with joint domain Y, where $H_U = -\sum_{y \in Y} \pi_U(y) \log \pi_U(y)$ is the Shannon entropy [58] of the empirical probabilities $\pi_U(y) = |\{p \in U\colon (a_1(p), \ldots, a_k(p)) = y\}|/|U|$ (defining $\pi_U(y) \log \pi_U(y) = 0$ for $\pi_U(y) = 0$). This measure is maximized by populations with point distributions of target values (i.e., such that there is a $y \in Y$ with $\pi_U(y) = 1$) and minimized by those that have a uniform target distribution.
- The *(standard) variation reduction* $u_{\mathrm{vr}}(S) = (s_P - s_S)/s_P$ where

$$s_U = \sqrt{\sum_{p \in U} \frac{(m_U - t(p))^2}{|U| - 1}} \tag{4}$$

is the sample standard deviation in the case of a single metric target variable $T = \{a\}$, a: $P \to \mathbb{R}$, with empirical mean $m_U = \sum_{p \in U} a(p)/|U|$ (and in the case of multiple metric target variables, the squared differences can for example be replaced by the squared norm of the difference vectors between sample values and the mean vector). Similar to the information gain u_{ig}, this utility function favors subgroups where the target values are as close as possible to some localized value over groups with uniform distribution, only this time the deviation from a point distribution is measured in a metric sense.

- The *(Pearson) correlation gain* $u_{\mathrm{cg}}(S) = (|r_S| - |r_P|)/(1 - |r_P|)$ between pairs of numeric target variables $T = \{a, b\}$, a, b: $P \to \mathbb{R}$, where

$$r_U = \frac{1}{|U| - 1} \sum_{p \in U} \left(\frac{m_U^a - a(p)}{s_U^a} \right) \left(\frac{m_U^b - b(p)}{s_U^b} \right) \tag{5}$$

is the sample Pearson product-moment correlation coefficient of the paired a and b-values in the population U (with m_U^x and s_U^x being the sample mean and the standard deviation of target variable x as in the definition of the variation reduction utility function). This utility function is maximized (having a value of 1) for subpopulations where the paired target values all lie on a line ($|r_S| = 1$). Hence, it is used to find subgroups where there is an

approximately linear relationship between two metric target variables. This is in contrast to traditional variants where subgroups with an unusual correlation (e.g., inverse effects) are sought [28].

The usage of these utility functions for uncovering interpretable local patterns and descriptors is demonstrated in section 3. Beforehand, we describe a simple and robust algorithm for finding subgroups with high quality values.

2.3. Search strategy

Optimizing any of the above quality functions is a computationally hard problem with no known polynomial time approximation algorithm (note that the size of the search space has an exponential relation to the number of basic selectors considered). The standard algorithmic approach to find optimal subgroups are exponential time branch-and-bound algorithms, which can be effective for certain input datasets if a good bounding function for the employed quality function is known [59]. Although deriving such bounding functions is an interesting research problem, here we follow a different route and utilize a heuristic two-step Monte Carlo sampling approach, which works well for many practical problems [60, 61]. The following procedure is repeated iteratively for as many random result patterns as desired:

1. *Seed generation*: Sample a random seed conjunction σ_0 with generation probability $\mathbb{P}(\sigma_0 = \sigma)$ proportional to the size of the extension $|\mathrm{ext}(\sigma)|$. This can be implemented by a direct sampling approach in time $O(mk)$, where m denotes the number of data points and $k = |C|$ the number of basic selector functions [61]. The idea is to first sample a member of the global population with a probability proportional to the number of conjunctions of basic selectors that are true for that population member, and then to sample uniformly a selector from those conjunctions.
2. *Opportunistic pruning*: Starting from $\sigma_0(\cdot) \equiv c_1(\cdot) \wedge \ldots \wedge c_l(\cdot)$, with the basic selectors in random order, consider each c_i for $i \in \{1, \ldots, l\}$ and remove it if the quality $q(\sigma'_{i-1}) \geqslant q(\sigma_{i-1})$, where σ'_{i-1} results from σ_{i-1} by removing c_i (in this case define $\sigma_i = \sigma'_{i-1}$, otherwise $\sigma_i = \sigma_{i-1}$). Since all quality functions considered here can be computed in time $O(m)$, the worst-case time complexity of this step is $O(mk)$.

For our analysis, subgroup selectors are chosen based on having the highest value of the quality function. At least 10 000 random seeds are used in the Monte Carlo search of subgroup selectors. Upon reapplication of the Monte Carlo algorithm, the same optimal subgroup selectors are found for the patterns described below. Nevertheless, due to its nonexhaustive nature, the Monte Carlo procedure does not guarantee that superior unfound patterns do not exist. The SGD algorithm was implemented in the Creedo web application with the realKD library [62].

3. Application

3.1. OB semiconductors: toward predicting crystal structures

Predicting the crystal structure of a material from only knowledge of its chemical composition is a long-standing goal to facilitate the design of materials [2, 63, 64]. Over four decades ago, van Vechten and Phillips [35, 36, 65] analyzed the classification of OB semiconductors and proposed a descriptor to classify the zincblende (ZB), wurtzite (WZ) and rocksalt (RS) structures. Since the studies by Van Vechten and Phillips, many researchers have sought to identify superior descriptors to classify the OB compounds (see [22, 39, 40] and the references within). Descriptors used to classify the OB crystal structures were typically introduced based on understanding by the authors of the bonding nature of these materials.

As a more general and less biased approach, recently a two-step feature selection algorithm, which consists of using the least absolute shrinkage and selection operator method followed by ℓ_0-constrained minimization (LASSO + ℓ_0), was utilized to systematically discover an interpretable descriptor that provides a classification and even quantitative energy differences of the OB compounds (as either ZB or RS) [22]. Figure 2 shows the main conclusion. The two-dimensional descriptor is derived from solely combinations of the radii of the maximum probability density of the valence s (r_s) and valence p (r_p) orbitals of the free atoms A and B that make up the OB compounds, as well as their ionization potential (IP) and electron affinity (EA).

As a complementary local modeling approach to the LASSO + ℓ_0 global modeling paradigm, here we examine the crystal-structure classification of OB semiconductors using SGD. The dataset of the OB compounds is obtained from Ghiringhelli *et al* [22] (the full list of the OB compounds is provided in their supporting information and all the input and output files can be downloaded from the NOMAD Repository using http://link.aps.org/doi/10.1103/PhysRevLett.114.105503 as an external DOI reference). The feature space is restricted to atomic properties of the free atoms within the OB compounds; in total, 55 features are used in the SGD algorithm (crystal structure information, 14 atomic properties, and 40 features derived from those; see

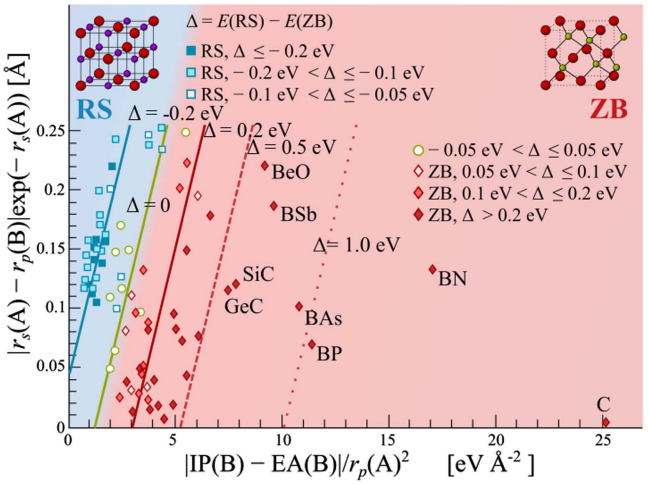

Figure 2. Computed energy differences (Δ) between rocksalt and zincblende crystal structures of the 82 octet AB-type binary compounds organized according to the two-dimensional descriptor found using LASSO + ℓ_0 [22].

table S1 of the supporting information). For the application of SGD, the information gain utility function u_{ig} is used with the sign of the energy difference between RS and ZB as the target variable ($T = \{\text{sign}(\Delta)\}$). The difference in energy between ZB and WZ structures for these materials is very small: the maximum absolute difference is 0.04 eV and the average over the dataset is 0.01 eV. Therefore, as in [22], we do not distinguish between ZB and WZ and we use the energy of ZB as the reference for Δ. A set C of 1 576 basic selectors is generated from the remaining 54 description features according to the rules outlined in section 2.1.

Application of SGD identifies several pure subgroups that exclusively contain either only RS or only ZB structures. Figure 3(a) shows the simplest (smallest number of basic selectors) maximum quality selectors found for each of the two crystal structures. The RS subgroup and ZB subgroup are described by the selectors $\sigma_1^{\text{RS}} \equiv |r_p^{\text{A}} - r_p^{\text{B}}| \geqslant 0.91$ Å $\wedge\ r_s^{\text{A}} \geqslant 1.22$ Å and $\sigma_1^{\text{ZB}} \equiv |r_p^{\text{A}} - r_p^{\text{B}}| \leqslant 1.16$ Å $\wedge\ r_s^{\text{A}} \leqslant 1.27$ Å, respectively. Here superscripts A and B denote the free atoms that make up the octet AB-type semiconductors. Since both subgroups are pure, they have the maximum entropy gain ($u_{\text{ig}}(\sigma_1^{\text{RS}}) = u_{\text{ig}}(\sigma_1^{\text{ZB}}) = 1$, where for notational convenience we write $u(\sigma)$ for $u(\text{ext}(\sigma))$. The RS subgroup covers all but two compounds of its respective crystal structure group ($\text{cov}(\sigma_1^{\text{RS}}) = 39/82$), whereas the ZB subgroup covers all but one ($\text{cov}(\sigma_1^{\text{ZB}}) = 40/82$).

A global model can be created by combining the underlying descriptors of both subgroups shown in figure 3(a), i.e., the descriptive features that are referenced in both σ_1^{RS} and σ_1^{ZB}, which describe 79 of the 82 OB compounds correctly as either RS or ZB—and that is agnostic about only three structures that have a nearly degenerate energy difference between RS and ZB (36.5 meV for AgI, 19.0 meV for CuF, and 4.5 meV for MgTe). This two-dimensional descriptor consists of only linear combinations of atomic radii of the valence s and p orbitals, i.e., $\{r_s^{\text{A}}, |r_p^{\text{A}} - r_p^{\text{B}}|\}$. SGD helps find that OB semiconductors with relatively larger values of r_s^{A} and $|r_p^{\text{A}} - r_p^{\text{B}}|$ favor RS structures, whereas smaller values favor ZB structures. Large valence p radii differences between atomic elements suggests ionic character within compounds, whereas smaller atomic radii differences suggest covalent character, which is in agreement with reports that ionic OB compounds typically form RS structures [35, 36, 39, 65].

Figure 3(b) depicts RS and ZB subgroups described by selectors consisting of previously reported descriptors; namely, St. John and Bloch's r_σ and r_π descriptors are used, which approximate the s–p contribution to the electronegativity and the s–p hybridization [33]. Interestingly, the descriptors that make up the optimally found subgroup selectors (figure 3(a)) are similar to r_π and r_σ (figure 3(b)). Selectors using r_σ and r_π as descriptive features describe 38 of the 41 ZB structures and 35 of the 41 RS structures. On the other hand, subgroups described by selectors that consist of the two-dimensional descriptor found by Ghiringhelli *et al* using LASSO + ℓ_0 contain only 71 of the RS and ZB structures (see figure S2 of the supporting information). This illustrates the different modeling strategies used by LASSO + ℓ_0 and SGD, i.e., LASSO + ℓ_0 optimizes for a single linear separating hyperplane, whereas SGD uses a (nonlinear) intersection of axis-parallel hyperplanes. The LASSO + ℓ_0 algorithm is suited to find a globally optimal and sparse descriptor for crystal-structure classification [22], which is complementary to the SGD methodology. However, SGD could find multiple local models that, when combined, span a large and relevant portion of the dataset. This could be a useful strategy to build structure maps for crystal-structure prediction [64].

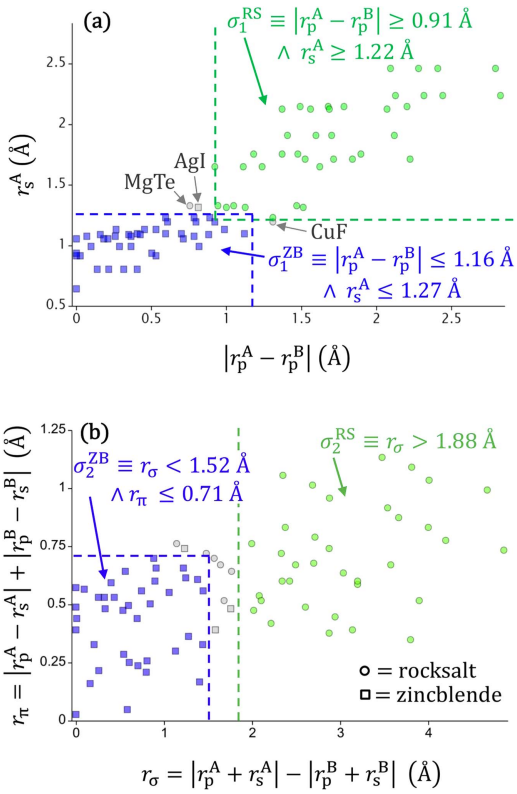

Figure 3. Application of subgroup discovery to the 82 octet binary semiconductors helps us identify interpretable selectors that describe subgroups of the rocksalt (RS) and zincblende (ZB) structures. (a) The subgroups described by selectors σ_1^{RS} (describing the RS subgroup) and σ_1^{ZB} (describing the ZB subgroup) with the highest quality value for a two-dimensional descriptor. (b) The subgroups described by selectors consisting of St. John and Bloch's r_σ and r_π descriptors [33]. The dashed blue and green lines denote the (nonlinear) intersection of axis-parallel hyperplanes that contain the ZB and RS subgroups. The squares and circles denote zincblende and rocksalt crystal structures, respectively. Green: rocksalt subgroup described by σ_i^{RS}; Blue: zincblende subgroup described by σ_i^{ZB}; gray: points described by neither selector. Another representation of σ_1^{RS} and σ_1^{ZB} is shown in figure S1 of the supporting information, where instead the axes are chosen to be the two-dimensional descriptor found by Ghiringhelli and coworkers using LASSO + ℓ_0 [22].

3.2. Neutral gas-phase gold clusters: the search for structure-property relationships

SGD is applied here to ascertain interpretable and general patterns between physicochemical and geometrical properties for 24 400 neutral gas-phase gold clusters (sizes of 5–14 atoms). Gold clusters have been a topic of sustained interest due to their various important and unique electronic, optical, and catalytic properties [66–74]. However, the majority of past computational studies on such clusters focused on a static, monostructure, description at 0 K, but increasing amounts of evidence indicate dynamic structural disorder is a common feature among clusters and that numerous isomers can coexist at finite temperature. General patterns holding across multiple cluster isomers of various sizes at finite temperature may be missed by this standard approach, therefore below we analyze gold clusters generated from the canonical ensemble at various temperatures.

Ab initio replica-exchange molecular dynamics (REMD) [75] simulations are performed with the FHI-aims electronic-structure code [76] to generate the gold cluster configurations based on uniform sampling of the canonical ensemble at temperatures from 100 to 814 K. REMD simulations utilized *light-tier* 1 (excluding the hydrogenic 6 h basis functions) numerical settings with energies and forces obtained from spin-polarized DFT with the PBE exchange-correlation functional [77] corrected for many-body dispersion (MBD) [78] (which we denote as PBE + MBD). Relativistic effects are treated using the 'atomic ZORA' scalar correction [79, 80]. See [74] for validation of the used REMD[6] settings and the choice of exchange-correlation functional, as well as the importance of including van der Waals corrections. Gold cluster geometries and their features are sampled from each replica in the REMD simulation every 1.0–2.3 ps, yielding 2 440 configurations per gold cluster size (24 400 configurations total for Au_5–Au_{14}). Because the quality function depends on coverage, we chose to uniformly sample the gold cluster configurations as a function of size to ensure that one gold cluster size is not biased over

[6] REMD simulations used a 10 fs molecular dynamics (MD) time step with 8–15 replicas exponentially distributed over a temperature range of at least 100–814 K. Between replica exchanges, canonical ensemble MD is carried out for 500 fs. The total MD simulation time is at least 3 ns for each gold cluster size. The stochastic velocity rescaling thermostat with a time-scale parameter of 100 fs is used to ensure proper sampling of the canonical distribution.

others when searching for subgroups. Additionally, all patterns identified using SGD were preserved upon subsampling of the gold cluster configurations.

The features computed for each cluster geometry are: relative total energy ΔE (relative to the most stable structure at each size), normalized radius of gyration R_{g0}, IP, EA, HOMO–LUMO energy gap E_{HL}, cluster size N, replica temperature T, atom coordination histogram (a vector containing the number of atoms with a certain bond coordination number) [81], and relative intramolecular van der Waals energy ΔE_{vdW} (relative to the structure with the largest van der Waals energy at each size), among others. A list of the gold cluster features is provided in table S2 of the supporting information and all the configurations can be downloaded from the NOMAD Repository (http://dx.doi.org/10.17172/NOMAD/2016.11.02-1). Since at finite temperatures a gold cluster will never be exactly at a minimum on the potential energy surface, all features of the gold clusters are computed from the unrelaxed structures generated from the canonical ensemble.

3.2.1. Finding patterns of the HOMO–LUMO energy gap

The HOMO–LUMO energy gap (E_{HL}) of neutral gold clusters is known to oscillate depending on whether the cluster has an even or odd number of atoms [82, 83]. The even–odd oscillatory behavior of E_{HL} is due to spin pairing for the even-sized neutral gold clusters and the lack of spin pairing for the odd-sized clusters. As a tutorial example, we first demonstrate that the SGD algorithm can help rediscover the known phenomenon that neutral clusters with an odd number of atoms have a small E_{HL} relative to neutral clusters with an even number of atoms. Algorithmically, subgroups of gold cluster configurations with a low standard deviation of the HOMO–LUMO energy gap are sought. That is, the target variable is specified as $T = \{E_{HL}\}$ and the standard variation reduction utility function u_{vr} is used (see section 2.2 for the definition of u_{vr}). A set C of 338 basic selectors is generated (using 6 cut-off values for metric variables). Examples of basic selectors are statements such as 'the number of atoms in the gold cluster is an even number', 'the number of the atoms in the gold cluster is $\geqslant 8$,' or 'the energy of the gold cluster configuration is low (i.e., an energetically favorable configuration).'

Upon application, SGD finds that the highest quality subgroup is described by the selector $\sigma_1^{HL} \equiv \text{odd}(N)$, which has a coverage of $\text{cov}(\sigma_1^{HL}) = 0.50$ (the subgroup covers 50% of the total population) and a utility value of $u_{vr}(\sigma_1^{HL}) = 0.89$ (89% of the standard deviation of the HOMO–LUMO energy gap is reduced in this subgroup relative to the global data). In other words, the clusters with an odd number of atoms ($N = 5, 7, 9, 11$ and 13) form a well-defined subgroup, figure 4(a). The second highest quality subgroup found is described by $\sigma_2^{HL} \equiv \text{even}(N) \wedge N \geqslant 7$, which has $\text{cov}(\sigma_2^{HL}) = 0.40$ and $u_{vr}(\sigma_2^{HL}) = 0.61$. Interestingly, Au_6 is excluded from this subgroup because it has an unusually large E_{HL} due to its highly stable ground state structure as a result of its σ-aromaticity [42, 46, 83]. As shown in figures 4(b) and (c), the distribution of the HOMO–LUMO energy gap values indicates two high quality subgroups; namely, odd(N)and even(N) $\wedge$ $N \geqslant 7$, and the lower quality subgroup $N = 6$. In figure 4(c), the oscillation in the HOMO–LUMO gap with respect to cluster size is observed, as well as the fact that larger even-sized gold clusters generally have a decreasing average HOMO–LUMO gap (due to their increasingly metallic nature). Moreover, the HOMO–LUMO energy gap varies dramatically depending on cluster configuration at a given size.

3.2.2. Structural and electronic properties of planar/quasi-planar and nonplanar gold clusters

SGD is next applied to discern general patterns among the structural and electronic features of planar/quasi-planar and nonplanar (compact, three-dimensional) structures. Small gold clusters often adopt stable planar geometries as a consequence of relativistic effects [84, 85], and planar and nonplanar structures can coexist simultaneously at finite temperature [86–88]. It is important to emphasize that theoretical studies using traditional generalized gradient functionals (without van der Waals corrections) are biased towards planar structures for gold clusters [47, 49]; however, our benchmark studies using PBE + MBD with *tight-tier* 2 settings have reasonable agreement with isomer energetics predictions using HSE06 + MBD as well as RPA@PBE. Geometries and energy differences between the lowest energy planar and nonplanar clusters using PBE + MBD are provided in figure S3 of the supporting information. Detailed analysis of the choice of exchange-correlation functional on relative isomer stabilities between planar and nonplanar neutral gold clusters as well as the importance of temperature on cluster isomer probabilities based on analysis of free energy surfaces will be reported in a separate article.

The minimum thickness of a gold cluster configuration has been used as an order parameter to monitor the planarity of a cluster [86]. Instead, here the planar/quasi-planar (from here on called planar or 2D) and nonplanar (3D) gold clusters are approximately discriminated based on their normalized radius of gyration. The radius of gyration is computed as the root mean square distance of the cluster's parts from its center of mass. The normalized radius of gyration is obtained by dividing the radius of gyration of each cluster configuration by the radius of gyration of the lowest energy planar structure at each size. Planar clusters have a larger radius of gyration compared with nonplanar structures due to their less compact nature. Cut-off values for discriminating between planar and nonplanar clusters and are chosen based on examining the probability distribution of the

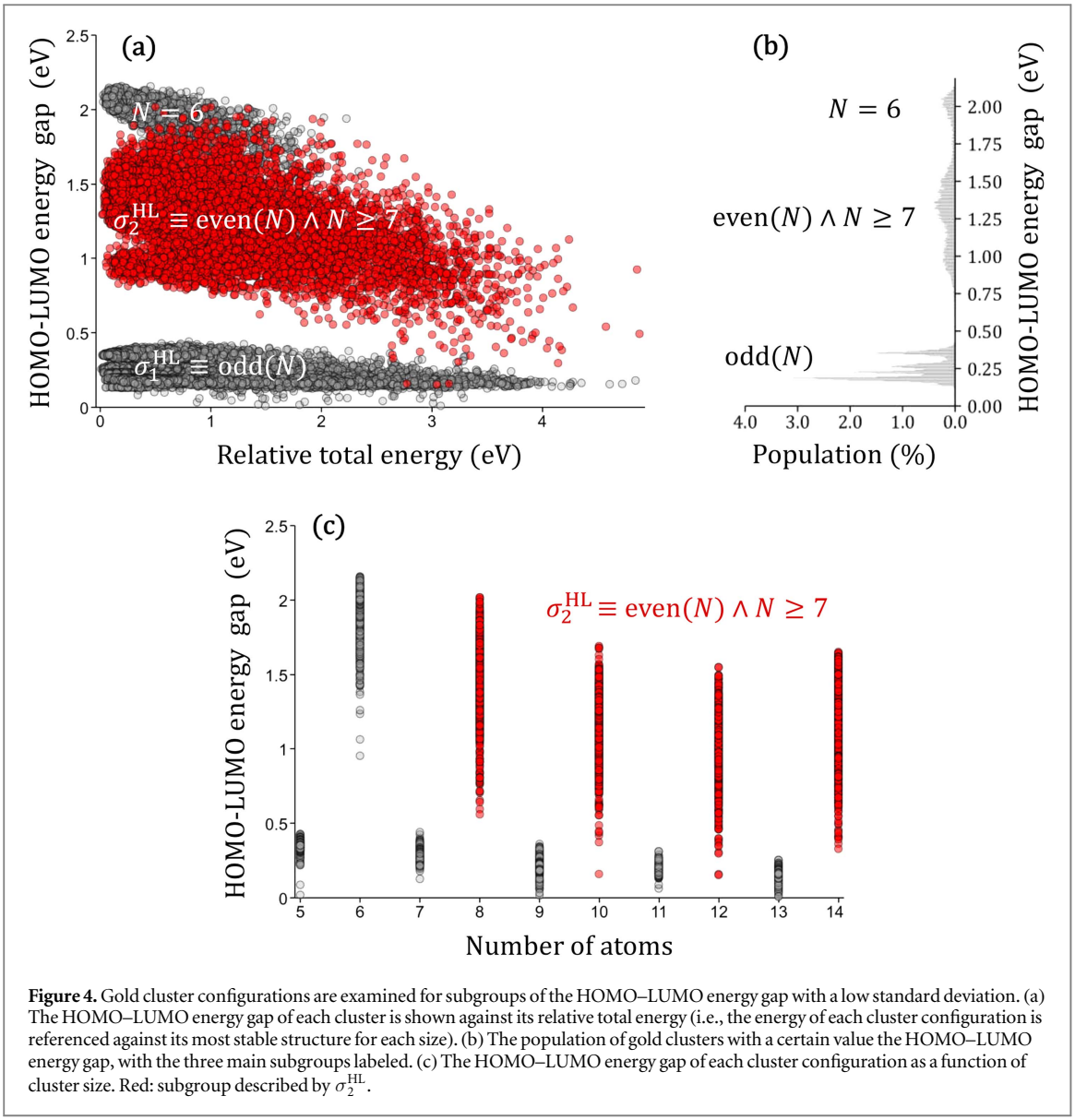

Figure 4. Gold cluster configurations are examined for subgroups of the HOMO–LUMO energy gap with a low standard deviation. (a) The HOMO–LUMO energy gap of each cluster is shown against its relative total energy (i.e., the energy of each cluster configuration is referenced against its most stable structure for each size). (b) The population of gold clusters with a certain value the HOMO–LUMO energy gap, with the three main subgroups labeled. (c) The HOMO–LUMO energy gap of each cluster configuration as a function of cluster size. Red: subgroup described by σ_2^{HL}.

radius of gyration (see table S3 and figure S4 of the supporting information). We refer to this categorical feature as 'Shape', i.e., planar or nonplanar.

Although there is a broad distribution of HOMO–LUMO energy gaps depending on cluster size and geometry, no statistically significant local patterns between E_{HL} and planar or nonplanar structures are found using SGD. The Monte Carlo procedure does not guarantee that unfound patterns do not exist, and thus we cannot ascribe significance to the lack of a found pattern. However, application of SGD using the information gain utility function u_{ig} with planar or nonplanar categorization as the target variable $T = \{\mathrm{Shape}\}$ elucidates that the majority of low energy (stable) planar structures have an average atom coordination number of less than or equal to three (see figure 5). The identified subgroup selector is $\sigma_1^{\mathrm{Sh}} \equiv \Delta E \leqslant 1.88$ eV $\wedge$ #mean $\leqslant 3$. The extension $S = \mathrm{ext}(\sigma_1^{\mathrm{Sh}})$ described by σ_1^{Sh} contains 14 097 configurations (coverage of 58%) of sizes 5–14 atoms that are almost exclusively planar ($\pi_S(2\mathrm{D}) = 0.98$ and $u_{\mathrm{ig}}(\sigma_1^{\mathrm{Sh}}) = 0.86$). In other words, low-energy planar clusters with 5–14 atoms typically have a coordination number less than or equal to three (although the subgroup is dominated by clusters of Au_5–Au_{10}). The high coverage of this subgroup reinforces the notion that gold cluster isomers within this size range typically adopt planar geometries due to their energetic stability [49]. Note, Au_{13} and Au_{14} rarely adopt planar gold cluster configurations [49, 88] at finite temperature and therefore are not widely included in this subgroup (figure 5(b)).

One can examine the importance of the individual descriptive features making up the selectors by removing them and examining the generalized selector's properties. For example, if the descriptive feature $\Delta E \leqslant 1.88$ eV is removed from σ_1^{Sh} then the generalized selector $\sigma_2^{\mathrm{Sh}} \equiv$ #mean $\leqslant 3$ is produced, which describes planar structures with less purity ($u_{\mathrm{ig}}(\sigma_2^{\mathrm{Sh}}) = 0.77$). It is unexpected that the average atom coordination number of low-energy planar gold clusters remains $\leqslant 3$ across a broad range of clusters sizes; for an infinite fcc(111) the

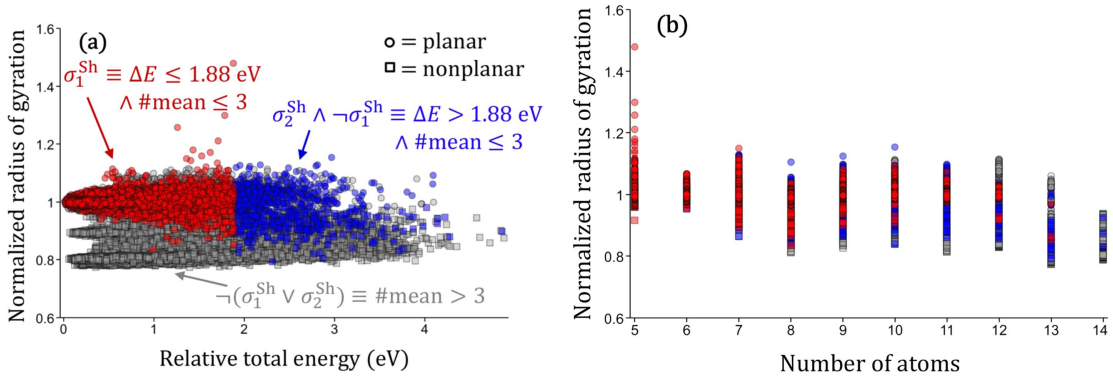

Figure 5. Subgroup selectors are identified that describe the structural and electronic features of planar/quasi-planar structures. (a) The normalized radius of gyration of each cluster configuration is shown against its relative total energy. (b) The normalized radius of gyration of each cluster configuration as a function of size. Red: subgroup described by σ_1^{Sh}; Blue: additional points included in generalized variant $\text{ext}(\sigma_2^{Sh} \wedge \neg\sigma_1^{Sh}) = \text{ext}(\sigma_2^{Sh}) \setminus \text{ext}(\sigma_1^{Sh})$; gray: points described by neither selector $\neg(\sigma_1^{Sh} \vee \sigma_2^{Sh}) = \neg\, \sigma_2^{Sh}$.

average atom coordination is six, and the 'inner' atoms of planar gold clusters are also six-fold coordinated. In some cases, the low average atom-coordination number of planar clusters relative to their more compact, nonplanar, isomer counterparts may result in increased reactivity due to the presence of more under-coordinated sites and better electron-accepting capabilities [89–91].

3.2.3. Intra-cluster van der Waals interactions and its shape dependence

Van der Waals (vdW) interactions are critical to include in electronic-structure theory calculations for determining the correct energetic ordering of material configurations, e.g., adsorption configurations of molecules on surfaces [92] and peptide conformers [93]. It has been suggested that nonplanar cluster configurations are more stabilized by intramolecular vdW interactions than planar cluster configurations [88], but this has yet to be quantified or demonstrated across a large range of cluster sizes and geometries at finite temperature. To test this hypothesis, SGD is applied using the variation reduction utility function u_{vr} with the intramolecular vdW energy (referenced to the maximum at each size) as the target variable ($T = \{\Delta E_{\text{vdW}}\}$), figure 6(a). Note, recall here that ΔE_{vdW} is just the long-range correlation energy within the MBD framework [78]. The highest quality selector found is $\sigma_1^{\text{vdW}} \equiv N \geqslant 8 \wedge N \leqslant 10 \wedge \text{Shape(2D)}$ with $\text{cov}(\sigma_1^{\text{vdW}}) = 0.26$ and $u_{\text{vr}}(\sigma_1^{\text{vdW}}) = 0.84$. This subgroup describes the phenomena that planar (2D) gold clusters generally exhibit weaker vdW interactions than nonplanar clusters for Au_8–Au_{10} clusters. Although the Au_8–Au_{10} pattern is highlighted in figure 6(a), subgroups describing similar phenomena for sizes 11–12 are also found.

The pattern found by SGD is further supported based on examination of the vdW energy difference between the lowest energy planar and nonplanar structures as a function of size. The vdW interaction per atom becomes increasingly larger in nonplanar clusters relative to planar clusters as the cluster size increases, figure 6(b). Additionally, the visualization in figure 6(a) suggests the existence of an unexpected higher order pattern: for certain cluster sizes there appears to be a linear relationship between the normalized radius of gyration and the vdW energy. SGD is used to test this hypothesis by augmenting the set of target variables to $T' = \{R_{g0}, \Delta E_{\text{vdW}}\}$ and using the correlation gain utility function u_{cg}. SGD finds $\sigma_3^{\text{vdW}} \equiv N \geqslant 8$ with $\text{cov}(\sigma_3^{\text{vdW}}) = 0.70$ and a correlation gain of $u_{\text{cg}}(\sigma_3^{\text{vdW}}) = 0.49$, which corresponds to a local Pearson product-moment correlation coefficient of $r_S = 0.84$. By modifying the weight parameter α, smaller groups with an even more linear relationship can be found, e.g., $\sigma_4^{\text{vdW}} \equiv N \geqslant 13$ with $r_S = 0.97$ using $\alpha = 1/8$. The compact nature of the nonplanar gold clusters increases its intra-cluster vdW interactions relative to planar clusters, even though the nonplanar clusters are less polarizable [94]. The strong influence of vdW interactions on stabilizing nonplanar gold structures relative to planar structures suggests that an accurate treatment of vdW interactions is required for predicting the correct isomer energetic ordering of polarizable nanoclusters.

3.2.4. Analyzing relationships between chemical hardness and cluster stability

The concept of chemical hardness is typically understood as the resistance of a system's chemical potential to a change in the number of electrons, and thus it is often used as a reactivity index [95–98]. Correlations have been found between the chemical hardness, stability, polarizability, and size of different systems [99–101]. Statistical mechanics in the grand canonical ensemble suggests that the ground state structure of a system has the maximum hardness of all the possible states at 0 K [102–104]. As a manifestation of the principle of maximum hardness, relatively more stable lithium clusters (those having a magic number of atoms) were predicted to have

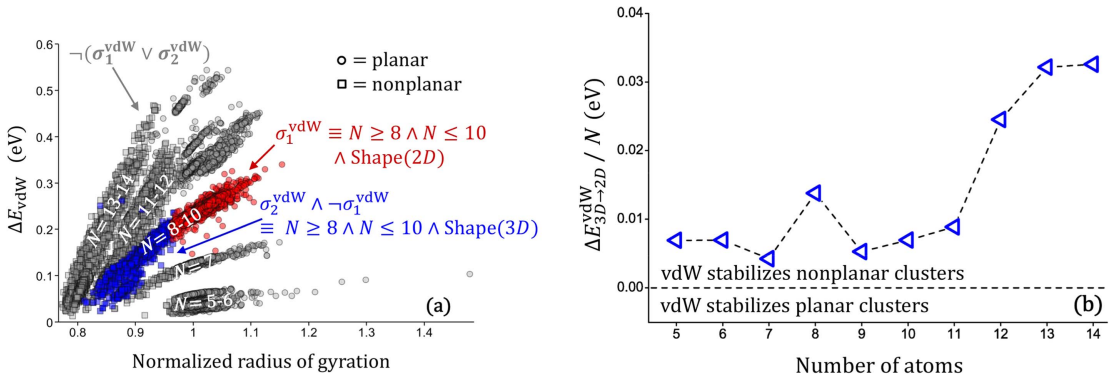

Figure 6. Gold cluster configurations are examined for patterns involving intra-cluster van der Waals interactions. (a) SGD finds subgroup selectors using the variation reduction utility function with the relative intramolecular vdW interaction energy ΔE_{vdW} (referenced to its maximum value at each size) as the target variable. Red: subgroup described by σ_1^{vdW}; Blue: additional points included in generalized variant $\text{ext}(\sigma_2^{vdW} \wedge \neg\sigma_1^{vdW}) = \text{ext}(\sigma_2^{vdW}) \setminus \text{ext}(\sigma_1^{vdW})$; gray: points described by neither selector $\neg(\sigma_1^{vdW} \vee \sigma_2^{vdW}) = \neg\, \sigma_2^{vdW}$. (b) The intramolecular vdW energy difference per atom between the lowest energy planar and nonplanar gold cluster geometries as a function of size (based on structures in figure S3 of the Supporting Information).

a local maximum in their chemical hardness [101]. However, to what degree correlations between chemical hardness and stability are present in a large set of cluster configurations in the canonical ensemble is not known.

To formalize this question for SGD the target variables are set to the relative total energy (referenced to the lowest energy cluster at each size) and the chemical hardness (referenced to the maximum hardness cluster at each size) $T = \{\Delta E,\ \Delta\eta\}$ and the utility function to the correlation gain u_{cg}. The chemical hardness is calculated at 0 K, which is sufficient because thermal corrections to the hardness are small even above room temperature [105, 106]. Although the global linear correlation between cluster stability and chemical hardness is small ($r_P = -0.27$), SGD finds a selector that describes a strong local linear trend, figure 7. The highest quality subgroup selector identified is $\sigma_1^{hd} \equiv \Delta E_{vdW} \leqslant 0.178$ eV $\wedge$ even$(N) \wedge$ #mode $\leqslant 5$, where #mode is the mode of the distribution of the atom coordination number. This subgroup has a coverage of $\text{cov}(\sigma_1^{hd}) = 0.20$ (20% of the total population) and correlation gain of $u_{cg}(\sigma_1^{hd}) = 0.54$ corresponding to a local Pearson correlation coefficient of $r_S = -0.81$.

It appears that the chemical hardness cannot generally predict the correct trend for the stability of multiple isomers [107]. However, within the subgroup described by σ_1^{hd} the hardness principle qualitatively holds (with coverage of 40% of the even-sized gold clusters in the dataset). The selector σ_1^{hd} is quite complex and relates seemingly unrelated features, i.e., van der Waals energy, atom size, and coordination number, to find a local descriptor that linearly correlates chemical hardness and stability, i.e., less stable cluster geometries are less chemically hard. Clusters with an odd number of atoms are excluded from the subgroup description because their hardness is nearly constant with geometry, which results from their unpaired electron. The importance of the constraint on the atom coordination mode can be illustrated by considering the nonoptimal selector σ_2^{hd} obtained by removing #mode $\leqslant 5$ from σ_1^{hd}. This subgroup described by σ_2^{hd} has a substantially weaker linear relationship between chemical hardness and energy ($r_S = -0.66$). Further investigation into other systems for trends between chemical hardness, reactivity, and stability is required for a deeper understanding.

4. Conclusions

Advances in big-data analytics, i.e., statistical and machine learning, compressed sensing, and data mining, alongside the exponential growth of materials-science repositories are opening innovative avenues for identifying advanced functional materials for use in applications such as batteries, thermoelectrics, and superconductors. Nevertheless, it remains challenging to screen databases of hypothetical and known materials for anomalies and to predict materials with superior properties than existing ones. Additionally, finding predictive descriptors of materials from high-dimensional data manually is laborious, error-prone, and typically subjective. Consequently, the development and application of sophisticated big-data analytics tools to extract materials insights is required.

One promising approach is to use SGD, which is a descriptive data-mining technique to find interpretable local patterns, correlations, and descriptors of properties according to some target property (or properties) of interest. As a complement to global modeling techniques, here SGD is formulated in the context of materials science, and two exemplary systems are examined: (1) 82 OB semiconductors to find physically meaningful rule-based models that predict their crystal structure as either zincblende or rocksalt (RS); and (2) 24 400

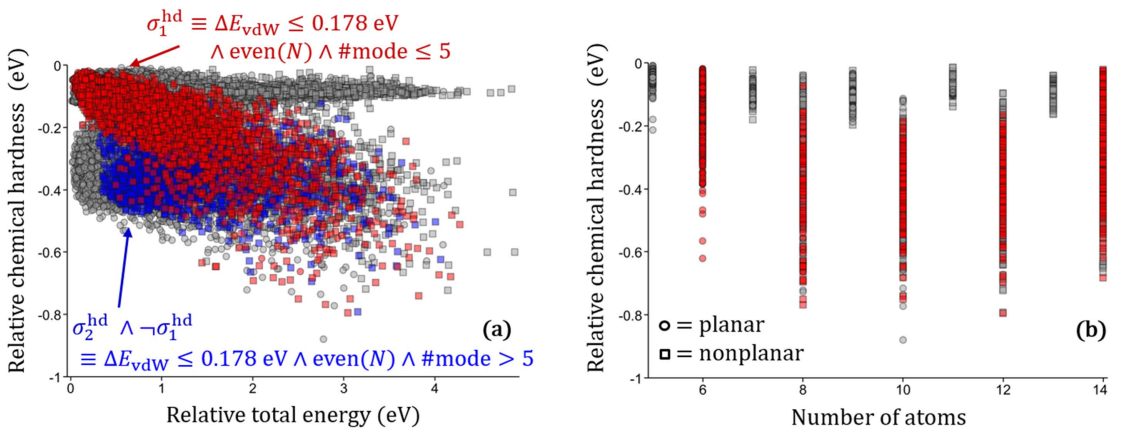

Figure 7. Gold cluster configurations are examined for linear correlations between the chemical hardness and the stability of the cluster. (a) The relative chemical hardness of each cluster is shown against its relative total energy. The selector σ_1^{hd} describes a subgroup with a local linear correlation between chemical hardness and cluster stability. (b) The relative chemical hardness of each cluster configuration as a function of cluster size. Red: subgroup described by σ_1^{hd}; Blue: additional points included in generalized variant $\text{ext}(\sigma_2^{hd} \wedge \neg\sigma_1^{hd}) = \text{ext}(\sigma_2^{hd}) \setminus \text{ext}(\sigma_1^{hd})$. The points of $\text{ext}(\sigma_2^{hd}) \setminus \text{ext}(\sigma_1^{hd})$ are displayed in figure 7(a) only; Gray: points described by neither selector $\neg(\sigma_1^{hd} \vee \sigma_2^{hd}) = \neg\sigma_2^{hd}$.

configurations of neutral gas-phase gold clusters with 5–14 atoms to search for general structure-property relationships holding across numerous gold cluster isomers of various sizes.

In this paper, SGD is demonstrated to find an interpretable two-dimensional model consisting only of atomic radii of valence *s* and *p* orbitals that properly classifies 79 of the 82 OB structures as either rocksalt or zincblende. Since the octets are only part of over 550 known AB_n binary solids [38], SGD can likely be used to find descriptors in the broader class of binary solids as well as construct structure maps, and these are directions for further investigation. For the gold clusters, unexpected and general trends are found upon application of SGD. For example, the intramolecular van der Waals interactions within planar clusters are typically significantly weaker compared with nonplanar, compact, clusters. This suggests that van der Waals interactions can be critical for accurately predicting the isomer energetic ordering of gold nanoclusters, especially for the planar to nonplanar geometrical transition.

Data analytics tools applied to materials-science data will continue to facilitate the understanding of structure-property relationships and the rational design of advanced materials. Nonetheless, limitations in both machine learning and data-mining approaches remain. In particular, for SGD an important open problem is the design of efficient optimal solvers for the quality function variants proposed in this work. Although the Monte Carlo algorithm discovers interesting patterns, it remains heuristic in nature. This prevents us to draw conclusions from the absence of certain patterns in the result set. For instance, are there hitherto undiscovered relations between the HOMO–LUMO energy gap, gold cluster stability, and gold cluster geometry? The design of optimal solvers for the proposed variants of SGD is currently underway to address this question, among others. We posit that SGD will serve as a useful tool for the extraction of insights from big data of materials, and its continued development will help pave the way toward materials discovery.

Acknowledgments

The project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. BRG acknowledges support from the Alexander von Humboldt-Foundation with a Postdoctoral Fellowship. The authors thank Christopher Sutton, Matthias Rupp and Runhai Ouyang for stimulating discussions and for providing feedback on the manuscript.

Notes

All the examples reported in this paper can be run via web-based tutorials accessible via: https://analytics-toolkit.nomad-coe.eu/Creedo/index.htm, where the users can also interactively change the input settings and compare the outcome with our results. These tutorials are part of the NOMAD analytics toolkit (https://analytics-toolkit.nomad-coe.eu/), which is developed in the context of the NOMAD Laboratory (https://nomad-coe.eu/).

References

[1] Nørskov J K, Bligaard T, Rossmeisl J and Christensen C H 2009 *Nat. Chem.* **1** 37
[2] Curtarolo S, Hart G L W, Nardelli M B, Mingo N, Sanvito S and Levy O 2013 *Nat. Mater.* **12** 191
[3] Jain A, Shin Y and Persson K A 2016 *Nat. Rev. Mater.* **1** 15004
[4] Behler J 2015 *Int. J. Quantum Chem.* **115** 1032
Rupp M 2015 *Int. J. Quantum Chem.* **115** 1058
Bartók A P and Csányi G 2015 *Int. J. Quantum Chem.* **115** 1051
[5] Bligaard T, Nørskov J K, Dahl S, Matthiesen J, Christensen C H and Sehested J 2004 *J. Catal.* **224** 206
[6] Goldschmidt V M 1926 *Naturwissenschaften* **14** 477
[7] Sato T, Takagi S, Deledda S, Hauback B C and Orimo S-I 2016 *Sci. Rep.* **6** 23592
[8] Yan J, Gorai P, Ortiz B, Miller S, Barnett S A, Mason T, Stevanovic V and Toberer E S 2015 *Energy Environ. Sci.* **8** 983
[9] Curtarolo S, Morgan D, Persson K, Rodgers J and Ceder G 2003 *Phys. Rev. Lett.* **91** 135503
[10] Fischer C C, Tibbetts K J, Morgan D and Ceder G 2006 *Nat. Mater.* **5** 641
[11] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[12] Saad Y, Gao D, Ngo T, Bobbitt S, Chelikowsky J R and Andreoni W 2012 *Phys. Rev.* B **85** 104104
[13] Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, Scheffler M, Von Lilienfeld O A, Tkatchenko A and Müller K-R 2013 *J. Chem. Theory Comput.* **9** 3404
[14] Rajan K 2015 *Annu. Rev. Mater. Res.* **45** 153
[15] Mueller T, Kusne A and Ramprasad R 2015 *Reviews in Coputational Chemistery* ed A L Parrill and K B Lipkowitz (New York: Wiley)
[16] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 *J. Chem. Theory Comput.* **11** 2087
[17] Calfa B A and Kitchin J R 2016 *AIChE J.* **62** 2605
[18] de Jong M, Chen W, Notestine R, Persson K, Ceder G, Jain A, Asta M and Gamst A 2016 *Sci. Rep.* **6** 34256
[19] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *Npj Comput. Mater.* **2** 16028
[20] Nelson L J, Hart G L W, Zhou F and Ozoliņš V 2013 *Phys. Rev.* B **87** 035125
[21] Raccuglia P *et al* 2016 *Nature* **533** 73
[22] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114** 105503
[23] Meredig B and Wolverton C 2014 *Chem. Mater.* **26** 1985
[24] Mueller T, Johlin E and Grossman J C 2014 *Phys. Rev.* B **89** 115202
[25] Kim C, Pilania G and Ramprasad R 2016 *Chem. Mater.* **28** 1304
[26] See http://nomad-repository.eu/cms/index.php?page=other-repositories for a list of online materials repositories
[27] Ghiringhelli L M, Carbogno C, Levchenko S V, Mohamed F, Huhs G, Lüders M, Oliveira M and Scheffler M 2016 *Psi-k Scientific Highlight Of The Month*
[28] Duivesteijn W, Feelders A J and Knobbe A 2016 *Data Min. Knowl. Discovery* **30** 47
[29] Fayyad U, Piatetsky-Shapiro G and Smyth P 1996 *AI Mag.* **17** 37
[30] Klösgen W 1996 *Advanced Techniques in Knowledge Discovery and Data Mining* (Menlo Park, CA: American Association for Artificial Intelligence) pp 249
[31] Gamberger D, Lavrač N, Železný F and Tolar J 2004 *J. Biomed. Inf.* **37** 269
[32] Atzmueller M 2015 *WIREs Data Min. Knowl. Discovery* **5** 35
[33] John J and Bloch A N 1974 *Phys. Rev. Lett.* **33** 1095
[34] Mooser E and Pearson W 1959 *Acta Cryst.* **12** 1015
[35] Van Vechten J A 1969 *Phys. Rev.* **182** 891
[36] Phillips J C and Van Vechten J A 1969 *Phys. Rev. Lett.* **22** 705
[37] Zunger A 1980 *Phys. Rev.* B **22** 5839
[38] Pettifor D G 1984 *Solid State Commun.* **51** 31
[39] Pilania G, Gubernatis J E and Lookman T 2015 *Sci. Rep.* **5** 17504
[40] Pilania G, Gubernatis J E and Lookman T 2015 *Phys. Rev.* B **91** 214302
[41] Grönbeck H and Andreoni W 2000 *Chem. Phys.* **262** 1
[42] Wang J, Wang G and Zhao J 2002 *Phys. Rev.* B **66** 035418
[43] Li J, Li X, Zhai H-J and Wang L-S 2003 *Science* **299** 864
[44] Fernández E M, Soler J M, Garzón I L and Balbás L C 2004 *Phys. Rev.* B **70** 165403
[45] Chandrakumar K R S, Ghanty T K and Ghosh S K 2004 *J. Phys. Chem.* A A **108** 6661–6
[46] Li X-B, Wang H-Y, Yang X-D, Zhu Z-H and Tang Y-J 2007 *J. Chem. Phys.* **126** 084505
[47] Johansson M P, Lechtken A, Schooss D, Kappes M M and Furche F 2008 *Phys. Rev.* A **77** 053202
[48] Mingos D M P 2014 *Gold Clusters, Colloids and Nanoparticles I* (Cham: Springer) vol 161, p 282
[49] Johansson M P, Warnke I, Le A and Furche F 2014 *J. Phys. Chem.* C **118** 29370
[50] Bhattacharya S, Sonin B H, Jumonville C J, Ghiringhelli L M and Marom N 2015 *Phys. Rev.* B **91** 241115
[51] Herrera F, Carmona C J, González P and del Jesus M J 2011 *Knowl. Inf. Syst.* **29** 495
[52] Siebes A 1995 *KDD* (Montreal, Canada: AAAI Press) pp 269
[53] Wrobel S 1997 *Principles of Data Mining and Knowledge Discovery: First European Symp., PKDD '97 (Trondheim, Norway, 24–27 June 1997)* ed J Komorowski and J Zytkow (Berlin: Springer) pp 78
[54] Friedman J H and Fisher N I 1999 *Stat. Comput.* **9** 123
[55] MacQueen J 1967 *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* (Berkeley, CA: University of California Press) pp 281
[56] Tibshirani R 1996 *J. R. Stat. Soc.* **58** 267
[57] Vovk V 2013 *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* ed B Schölkopf *et al* (Berlin: Springer) pp 105
[58] Gray R M 1990 *Entropy and Information Theory* (Berlin: Springer) pp 21
[59] Grosskreutz H, Rüping S and Wrobel S 2008 *ECML PKDD* (Berlin: Springer) pp 440
[60] Boley M, Lucchese C, Paurat D and Gärtner T 2011 *17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'11* (San Diego, CA: ACM) pp 582
[61] Boley M, Moens S and Gärtner T 2012 *18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'12* (Beijing: ACM) pp 69

[62] Creedo is a web application that provides an intuitive graphical user interface for real knowledge discovery algorithms and allows to rapidly design, deploy, and conduct user studies. See http://realkd.org/creedo-webapp/for additional information. See also the NOMAD analytics-toolkit for a tutorial (https://labdev-nomad.esc.rzg.mpg.de/home/)
[63] Hawthorne F C 1990 *Nature* **345** 297
[64] Bialon A F, Hammerschmidt T and Drautz R 2016 *Chem. Mater.* **28** 2550
[65] Phillips J 1970 *Rev. Mod. Phys.* **42** 317
[66] Sanchez A, Abbet S, Heiz U, Schneider W D, Häkkinen H, Barnett R N and Landman U 1999 *J. Phys. Chem.* A **103** 9573
[67] Häkkinen H, Moseler M and Landman U 2002 *Phys. Rev. Lett.* **89** 033401
[68] Xing X, Yoon B, Landman U and Parks J H 2006 *Phys. Rev.* B **74** 165423
[69] Bulusu S, Li X, Wang L-S and Zeng X C 2006 *Proc. Natl Acad. Sci. USA* **103** 8326
[70] Kryachko E S and Remacle F 2007 *Int. J. Quantum Chem.* **107** 2922
[71] Olson R M and Gordon M S 2007 *J. Chem. Phys.* **126** 214310
[72] Häkkinen H 2008 *Chem. Soc. Rev.* **37** 1847
[73] Gruene P, Rayner D M, Redlich B, van der Meer A F G, Lyon J T, Meijer G and Fielicke A 2008 *Science* **321** 674
[74] Ghiringhelli L M, Gruene P, Lyon J T, Rayner D M, Meijer G, Fielicke A and Scheffler M 2013 *New J. Phys.* **15** 083003
[75] Earl D J and Deem M W 2005 *Phys. Chem. Chem. Phys.* **7** 3910
[76] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 *Comput. Phys. Commun.* **180** 2175
[77] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[78] Ambrosetti A, Reilly A M, DiStasio R A Jr and Tkatchenko A 2014 *J. Chem. Phys.* **140** 18A508
[79] van Lenthe E, Ehlers A and Baerends E-J 1999 *J. Chem. Phys.* **110** 8943
[80] van Lenthe E, Snijders J and Baerends E 1996 *J. Chem. Phys.* **105** 6505
[81] The atom coordination histogram was computed following the procedure used in Ceriotti M, Tribello G A and Parrinello M 2013 *J. Chem. Theory Comput.* **9** 1521
[82] Häkkinen H and Landman U 2000 *Phys. Rev.* B **62** R2287
[83] Woodruff D P 2007 *Atomic Clusters: From Gas Phase to Deposited* (Amsterdam: Elsevier)
[84] Pyykko P 1988 *Chem. Rev.* **88** 563
[85] Arratia-Perez R, Ramos A F and Malli G L 1989 *Phys. Rev.* B **39** 3005
[86] Koskinen P, Häkkinen H, Huber B, von Issendorff B and Moseler M 2007 *Phys. Rev. Lett.* **98** 015701
[87] Santarossa G, Vargas A, Iannuzzi M and Baiker A 2010 *Phys. Rev.* B **81** 174205
[88] Beret E C, Ghiringhelli L M and Scheffler M 2011 *Faraday Discuss.* **152** 153
[89] Martínez A 2010 *J. Phys. Chem.* C **114** 21240
[90] Sekhar De H, Krishnamurty S and Pal S 2010 *J. Phys. Chem.* C **114** 6690
[91] Staykov A, Nishimi T, Yoshizawa K and Ishihara T 2012 *J. Phys. Chem.* C **116** 15992
[92] Liu W, Ruiz V, Zhang G-X, Santra B, Ren X, Scheffler M and Tkatchenko A 2013 *New J. Phys.* **15** 053046
[93] Tkatchenko A, Rossi M, Blum V, Ireta J and Scheffler M 2011 *Phys. Rev. Lett.* **106** 118102
[94] Idrobo J C, Walkosz W, Yip S F, Öğüt S, Wang J and Jellinek J 2007 *Phys. Rev.* B **76** 205422
[95] Ayers P W and Parr R G 2000 *J. Am. Chem. Soc.* **122** 2010
[96] Chattaraj P K 2009 *Chemical Reactivity Theory: A Density Functional View* (Boca Raton: CRC Press) p 610
[97] Malek A and Balawender R 2015 *J. Chem. Phys.* **142** 054104
[98] Pan S, Sola M and Chattaraj P K 2013 *J. Phys. Chem.* A **117** 1843
[99] Ghanty T K and Ghosh S K 1993 *J. Phys. Chem.* **97** 4951
[100] Zhou Z and Parr R G 1989 *J. Am. Chem. Soc.* **111** 7371
[101] Harbola M K and Natl P 1992 *Acad. Sci. USA* **89** 1036
[102] Yang W and Parr R G 1985 *Proc. Natl Acad. Sci. USA* **82** 6723
[103] Pearson R G 1993 *Acc. Chem. Res.* **26** 250
[104] Parr R G and Chattaraj P K 1991 *J. Am. Chem. Soc.* **113** 1854
[105] Franco-Pérez M, Gázquez J L and Vela A 2015 *J. Chem. Phys.* **143** 024112
[106] Franco-Pérez M, Gázquez J L, Ayers P W and Vela A 2015 *J. Chem. Phys.* **143** 154103
[107] Noorizadeh S 2005 *J. Mol. Struct.: THEOCHEM* **713** 27