The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 13 2023

A fuzzy classification framework to identify equivalent atoms in complex materials and molecules

King Chun Lai ; Sebastian Matera ; Christoph Scheurer ; Karsten Reuter

J. Chem. Phys. 159, 024129 (2023)
https://doi.org/10.1063/5.0160369

AIP Publishing

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

A fuzzy classification framework to identify equivalent atoms in complex materials and molecules

Cite as: J. Chem. Phys. **159**, 024129 (2023); doi: 10.1063/5.0160369
Submitted: 1 June 2023 • Accepted: 26 June 2023 •
Published Online: 13 July 2023

View Online Export Citation CrossMark

King Chun Lai,[a] Sebastian Matera, Christoph Scheurer, and Karsten Reuter

AFFILIATIONS

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany

[a]**Author to whom correspondence should be addressed:** lai@fhi-berlin.mpg.de

ABSTRACT

The nature of an atom in a bonded structure—such as in molecules, in nanoparticles, or in solids, at surfaces or interfaces—depends on its local atomic environment. In atomic-scale modeling and simulation, identifying groups of atoms with equivalent environments is a frequent task, to gain an understanding of the material function, to interpret experimental results, or to simply restrict demanding first-principles calculations. However, while routine, this task can often be challenging for complex molecules or non-ideal materials with breaks in symmetries or long-range order. To automatize this task, we here present a general machine-learning framework to identify groups of (nearly) equivalent atoms. The initial classification rests on the representation of the local atomic environment through a high-dimensional smooth overlap of atomic positions (SOAP) vector. Recognizing that not least thermal vibrations may lead to deviations from ideal positions, we then achieve a fuzzy classification by mean-shift clustering within a low-dimensional embedded representation of the SOAP points as obtained through multidimensional scaling. The performance of this classification framework is demonstrated for simple aromatic molecules and crystalline Pd surface examples.

I. INTRODUCTION

When bound into molecules or materials, even atoms of the same chemical species can still possess very different properties and functions, e.g., different roles in a chemical reaction. A decisive factor in this is the local atomic environment of the atom, i.e., the relative positions of all other atoms in its vicinity. A natural question is then which of the atoms in one or several different bonded structures are equivalent in terms of this local environment and would correspondingly be attributed similar properties and functions. Indeed, such a grouping of equivalent atoms is common in materials science and chemistry. In atomic-scale modeling and simulation, it is, e.g., central to allocate computational effort to representative atoms of each equivalence class, to structure the data analysis, and to select building blocks in material design—to name but a few of the frequent use cases. A specific application that served as the original motivation for this work would, for instance, be adaptive kinetic Monte Carlo (kMC) simulations,[1,2] where the transition states of elementary processes need to be computed for every atom in a structure in a potentially huge number of sequential kMC steps. Good starting guesses for the transition states are based on recognizing that an atom has a similar local environment to previously calculated cases is there a pivotal efficiency driver.

Now, it is intuitively clear that a small perturbation of the local environment will generally not dramatically change the nature of an atom. Similarly, the nearsightedness of chemical interactions also tells us that neighboring atoms further and further away will typically play an ever decreasing role. In practice, the classification of equivalent atoms should, therefore, be fuzzy, up to such small perturbations and prioritizing nearby neighbors. In fact, the resolution, i.e., with up to which differences in their local environment atoms are still classified as equivalent, is a continuous function, and the optimum resolution will depend on the bonded structure and the task at hand. For instance, for organic molecules, the direct coordination of an atom may already be enough to obtain a qualitative understanding of its function. A carbon atom in an aromatic ring will have very different properties from a carbon atom in an alkyl chain, and a coarse representation of the local environment accounting only for the directly coordinated neighbors, their distances, and bond angles would suffice to distinguish the two cases. Similarly, at crystalline metal surfaces, a first distinction is generally made in terms of differently coordinated terrace atoms, step atoms, kink atoms, or adatoms. However, depending on the application, it may also be necessary to further branch these into sub-classes resolving e.g., the surface orientation (facet), the step type, combinations of multiple chemical species, nearby defects, or other increasingly more subtle variations in the local atomic environments.

Traditionally, the grouping of equivalent atoms is performed manually by the researcher and is often merely based on visual inspection of the atomic structure. This approach is obviously laborious and error-prone, and conflicts with increasing interest in high-throughput workflows,[3–6] e.g., for catalysis[7] or battery interfaces,[8,9] with interest in the generation of large and growing structural databases,[10,11] in global structure optimization problems,[12–16] or in the treatment of complex atomic arrangements, such as nanostructures[17] or amorphous materials.[18] In these tasks, identifying a complete set of equivalent local environments merely by visual inspection would either become a severe limitation or be completely intractable.

To address this issue, we here develop a general machine-learning (ML) framework to automatically identify the groups of (near-)equivalent atoms within any single or any set of bonded structures. These bonded structures may thereby comprise molecules, extended (crystalline or amorphous) materials, and their surfaces or interfaces. Emphasis is made to have a simple and continuous control of the resolution in the fuzzy classification. The starting point is to utilize one of the local descriptors,[19–23] which have been developed during the last years, to map the local environment of an atom onto a point in a high-dimensional space $\mathbb{R}^S$. After determining this vector for all atoms in the considered structure(s), we employ clustering on the resulting set of data points to obtain different classes of equivalent environments. Fuzziness is introduced in this approach by specifically employing the double smooth overlap of atomic positions (SOAP) descriptor,[19] which naturally emphasizes nearsightedness, and by employing multidimensional scaling (MDS)[24] to embed the SOAP-points in a lower-dimensional space $\mathbb{R}^{S'}$. This lower-dimensional space is then beneficial to obtain the fuzzy classes of approximately equivalent atoms by mean shift clustering (MSC).[25] Besides, the parameters of the SOAP representation, the framework, thus, has two key hyperparameters to control the resolution of the classification, i.e., the dimensionality of the MDS space and the bandwidth of the MSC.

We would like to emphasize that algorithms to categorize atoms on the basis of their local environments have been proposed for multiple application purposes before. A few prominent examples are the work of the Hammer group in the context of global geometry optimization[12,14] or the work of the Ceriotti group in the context of probabilistic analysis of molecular motifs (PAMM).[26,27] While following analog conceptual steps as e.g., the Hammer workflow, our framework differs in its attempt to avoid any predefinition and system specificity. Rather than *a priori* specifying the number of different equivalence classes, this number and the corresponding classes result automatically as a consequence of the chosen resolution as controlled by the MSC bandwidth. With a larger bandwidth, fewer classes will be distinguished, and atoms with wider variations in their local environment will still be classified as equivalent. Similarly, rather than imposing system-specific features of the local environment as central to the classification, these features again emerge naturally in the MDS dimensionality reduction step. With lower MDS dimensionality, the clustering will only be based on the most eminent components extracted from the SOAP descriptor, which are typically connected to the immediate neighboring shell. With this variability in the fuzziness, our generic algorithm is also geared to a later inclusion into larger workflows, where the resolution as defined by MDS dimensionality and MSC bandwith could, for instance, be adjusted in active learning cycles evaluating the suitability of the determined equivalence classes for the targeted application.

The following Sec. II will discuss details of the technical implementation of our approach. In Sec. III, the performance of the algorithm will be demonstrated by applying it to both finite molecules (Sec. III A) and extended materials surfaces (Sec. III B). On the basis of these results, we will then discuss limitations and possible extensions in Sec. III C.

II. METHODS

A. Environment representation through SOAP

In particular, within the booming field of ML interatomic potentials, much progress has recently been achieved in developing general representations of atomic environments that go beyond the mere recognition of (generalized) coordination numbers.[28,29] By construction, these representations encode for instance fundamental symmetries such as translational and rotational symmetry, as well as symmetry with respect to permutations of atoms of the same species. Among these representations, we choose for this work the vectorial SOAP descriptor, which for the present purposes offers a good compromise between flexibility and ease of use. The developed framework does not depend on this choice though, and any other environment descriptor, e.g., a graph-based one, could equally be employed.

Referring to the original literature for details,[19] Fig. 1 shows an illustration of the working principle of SOAP. In short, it places a Gaussian density function with variance σ^2_{SOAP} at the location of each atom within a sphere with radius $r_{\mathrm{SOAP,cut}}$ centered around the atom i for which the local environment shall be mapped. The overlapped local density of each chemical species is then expanded into a product basis of spherical harmonics for the angular dependence, and a set of orthogonal basis functions for the radial dependence. At this point, one set of expansion coefficients is obtained for each chemical species. To achieve rotational invariance, a normalized power spectrum is subsequently constructed between all combinatorial pairs of coefficient sets of the involved chemical species. This power spectrum is an abstract vector $\boldsymbol{\chi}_i \in \mathbb{R}^S$ describing the local environment like a fingerprint. The dimensionality S of the vector is thereby determined by the number of chemical species and the parameters for the SOAP expansion, namely the maximums, $n_{\mathrm{SOAP,max}}$ and $l_{\mathrm{SOAP,max}}$, for the radial and angular basis functions, respectively.

Here, we specifically use the so-called double SOAP approach,[30,31] which distinguishes two spheres around the central atom. Higher values for $n^{\mathrm{short}}_{\mathrm{SOAP,max}}$ and $l^{\mathrm{short}}_{\mathrm{SOAP,max}}$ are chosen

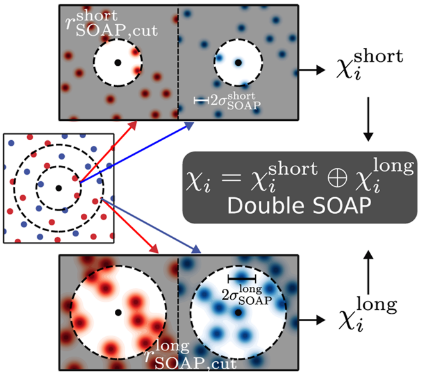
$r^{\text{short}}_{\text{SOAP,cut}}$
$2\sigma^{\text{short}}_{\text{SOAP}}$
χ_i^{short}
$\chi_i = \chi_i^{\text{short}} \oplus \chi_i^{\text{long}}$
Double SOAP
$2\sigma^{\text{long}}_{\text{SOAP}}$
$r^{\text{long}}_{\text{SOAP,cut}}$
χ_i^{long}

FIG. 1. Illustration of the double SOAP representation of the local atomic environment of a bonded structure consisting of two chemical species A (red) and B (blue). The panel on the left shows the locations of the atoms in the vicinity of the atom i (black central dot) for which the local environment is to be mapped. The upper and lower rows illustrate the short- and long-range parts of the double SOAP approach, respectively; see the text. For each part, a Gaussian density of width $\sigma^{\text{short/long}}_{\text{SOAP}}$ is placed at the position of each atom. Summing the densities of the same species gives the total overlapped atomic density for each species separately. Within the corresponding cut-off radius $r^{\text{short/long}}_{\text{SOAP, cut}}$, these densities are expanded into basis functions. The SOAP vector $\boldsymbol{\chi}_i^{\text{short/long}}$ is constructed from the power spectra of the coefficients of these expansions. The double SOAP vector $\boldsymbol{\chi}_i$ of atom i is finally formed by concatenating the short- and long-range SOAP vectors.

in a smaller sphere with a radius $r^{\text{short}}_{\text{SOAP,cut}}$, while neighboring atoms lye beyond $r^{\text{short}}_{\text{SOAP,cut}}$ but within a sphere of radius $r^{\text{long}}_{\text{SOAP,cut}}$ are less resolved with lower values $n^{\text{long}}_{\text{SOAP,max}}$ and $l^{\text{long}}_{\text{SOAP,max}}$. This way, the principle of nearsightedness is naturally built into the environment representation, placing less weight on more distant atoms in the outer sphere, and completely neglecting any neighboring atoms beyond $r^{\text{long}}_{\text{SOAP,cut}}$. The two SOAP vectors of the two spheres are then concatenated to form the final SOAP vector $\boldsymbol{\chi}_i$. We will specify the SOAP parameters used in the different examples in Sec. III below. The package DScribe[32] is used for the SOAP vector generation throughout this work.

B. MDS dimensionality reduction for clustering

The dimensionality S of double SOAP vectors is very high and easily exceeds several hundreds. This can be a hazard in the clustering process. In particular, to also achieve an easily tunable fuzziness, we next map the $\boldsymbol{\chi}_i$ first onto a low-dimensional space using MDS.[24] While the multiple SOAP parameters would, therefore, in general be chosen to achieve an accurate and non-system specific representation of the local environment, the truly distinctive features of this environment would then emerge naturally through this embedding. The dimensionality S' of the corresponding MDS space is thereby a tuning hyperparameter for the fuzziness, which, as will be seen below, can be as low as two.

MDS is a general technique to map data onto an abstract space while preserving the dissimilarity among data points.[24] In MDS, dissimilarity is interpreted directly as the distance between data points, and in the context of this work, the Euclidean distance between the double SOAP vectors of atoms i and j

$$D_{ij} = D(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2 \tag{1}$$

is an obvious choice for the dissimilarity of the two local environments. Other kernel forms[23,33] are conveniently available thanks to the vectorial nature of SOAP. For a total of N atoms in the bonded structure(s) under consideration, this yields a $(N \times N)$ matrix $\mathbf{D}$, for which classical MDS solves the eigenvalue problem of the Gram matrix $\mathbf{G}$,[24]

$$G_{ij} = \frac{1}{2N}\sum_k D_{ik}^2 + \frac{1}{2N}\sum_k D_{kj}^2 - \frac{1}{2N^2}\sum_{k,l} D_{kl}^2 - \frac{1}{2}D_{ij}^2. \tag{2}$$

The result is a set of eigenvalues λ_a with corresponding normalized orthogonal eigenvectors $\mathbf{v}_a = (v_{a,1}, v_{a,2}, \ldots, v_{a,N})^\top$, where a is the index of descendingly ordered λ_a.

The eigenspace of $\mathbf{G}$ can now be used to create a mapping from the SOAP space $\mathcal{S}$ to the abstract embedded space $\mathcal{S}'$. For a chosen dimensionality S' of this MDS space $\mathcal{S}'$, this starts by setting up the $(N \times S')$ matrix $\mathbf{P}$ from the first $1 \le a \le S'$ eigenvalue-weighted eigenvectors

$$P_{ak} = v_{a,k}/\sqrt{\lambda_a}. \tag{3}$$

Now consider any atom i. The ath component of the S'-dimensional mapped SOAP vector $\boldsymbol{\chi}'_i$ is then[34,35]

$$\chi'_{i,a} = \sum_{k=1}^{N} P_{ak} D^2(\boldsymbol{\chi}_k, \boldsymbol{\chi}_i). \tag{4}$$

In other words, Eq. (3) defines the embedding projector $\mathbf{P}$ from the high-dimensional SOAP space $\mathcal{S}$ to the low-dimensional MDS space $\mathcal{S}'$. With the choice of Euclidean distance as the dissimilarity measure, the MDS eigen problem is equivalent to that of a principal component analysis (PCA). This allows us to use the set of eigenvalues λ as guidance in selecting a suitable dimension S', e.g., through the broken-stick method,[36] which does not involve any extra hyperparameter. In particular, S' is estimated such that for all $a \le S'$, the normalized eigenvalues are larger than the broken-stick model series l_a, $\lambda_a/\sum_k^N \lambda_k \ge l_a$, where l_a[36] is

$$l_a = \frac{1}{N}\sum_{k=a}^{N} \frac{1}{k}. \tag{5}$$

This approach is also illustrated in Sec. III. Alternatively, S' may be seen as a tunable hyperparameter that may, e.g., be optimized within a larger workflow that assesses the performance of the classification for a targeted application. Note also that the embedding operator $\mathbf{P}$ works equally for atoms of any additional structure not contained in the original set. This allows us to conveniently analyze new structures in terms of a once-achieved fuzzy classification, as will be illustrated below. We would like to emphasize that besides classical MDS, there are, of course, other options for dimensionality reduction,[37] such as, e.g., kernel principal component analysis[38] or Sketch Map.[39,40] Each of them comes with pros and cons. With other reduction options in general, the number of intrinsic dimensions can also be estimated with other packages, e.g., DADApy.[41] Here, MDS is our primary choice because of its minimum number of hyperparameters, namely the embedding dimension. Other reduction methods are demonstrated in the supplementary material.

C. Clustering of atomic environments

Having mapped the dataset to the low-dimensional MDS space $\mathcal{S}'$, we finally cluster the atomic environments according to the geometric similarity reflected in the spatial distribution of the corresponding N data points in $\mathcal{S}'$. This grouping is achieved by mean shift clustering,[25] where we employ the implementation of the Scikit-learn package.[42] We chose MSC as it does not require to predefine the final number of groups of equivalent atoms. Instead, its only input parameter is a characteristic distance, the MSC bandwidth δ_{MSC}, which thus emerges as the second tunable hyperparameter of our framework. As with the choice of the SOAP representation before, we note that the developed framework is not restricted to the choice of MSC. As with the choice of the SOAP representation before, we note that the developed framework is not restricted to the choice of MSC. One could well substitute MSC with other density-based cluster algorithms with similar capabilities, e.g., DBSCAN,[43] HDBSCAN,[44,45] or spectral clustering.[46] Our current choice of MSC is motivated by its flexibility in handling both noisy and non-noisy datasets and its convenience in out-of-sample classification. For the case of very noisy datasets, HDBSCAN[44,45] might indeed be a more efficient choice. As shown in the supplementary material, its performance seems not so good for non-noisy datasets though.

A simple illustration of MSC is shown in Fig. 2(a). We start by considering the collection of N data points in $\mathcal{S}'$. A sphere with radius δ_{MSC} is drawn around any one chosen data point, and the mean position $\chi'_{\mathrm{mean},1}$ of all data points within the sphere is determined. Next, we calculate the mean position $\chi_{\mathrm{mean},2}$ of all data points within a new sphere around $\chi'_{\mathrm{mean},1}$ with the same radius δ_{MSC}, cf. Fig. 2(a). The mean position is now shifted from $\chi'_{\mathrm{mean},1}$ to $\chi'_{\mathrm{mean},2}$. This iteration goes on until the location of the mean converges, which finally gives the center location of a cluster. Starting the algorithm subsequently from all N data points, cf. Fig. 2(b), yields a complete list of cluster centers. The number of these clusters is generally lower than N, as cluster centers will have coincided during the iterative determination of their location. Each data point is finally assigned to its nearest cluster center.

The bandwidth δ_{MSC} crucially determines the resolution of the clustering algorithm. With a too large δ_{MSC}, the algorithm will fail to differentiate non-equivalent groups. With a too small δ_{MSC}, it isolates every atom (aka point in the MDS space) into its own group. As with the MDS dimension S', one may simply consider δ_{MSC} as a tunable hyperparameter of our framework that could, e.g., be optimized by a higher-level workflow into which the present framework is integrated and which evaluates the performance of the achieved fuzzy classification for the targeted application. Alternatively, a simple heuristic for the bandwidth may also be employed. Clusters in the MDS space $\mathcal{S}'$ distinguish themselves by having closer distances among their data points than distances to other data points. They, thus, manifest themselves as agglomerations in the distribution of pairwise distances $D(\chi'_i, \chi'_j) = \|\chi'_i - \chi'_j\|_2$. For the finite number of N data points, this distribution corresponds to a set of $N(N-1)/2$ δ-peaks. For larger numbers N, identifying distance regions with more or less δ-peaks may then become cumbersome. We, therefore, conveniently smear every δ-peak into a Gaussian of width σ_{smear} and add all Gaussians to arrive at a smooth distribution $D(\chi'_i, \chi'_j) = \|\chi'_i - \chi'_j\|_2$ that resembles a spectrum. In this spectrum, agglomerations of similar distances will simply show up as peaks. Choosing δ_{MSC} accordingly somewhere in the minimum after any dominant peak in the smoothed $D(\chi'_i, \chi'_j) = \|\chi'_i - \chi'_j\|_2$ spectrum should correspondingly yield a good heuristic to identify clusters. A δ_{MSC} chosen in the first minimum of $D(\chi'_i, \chi'_j)$ will thereby resolve a maximum of clusters, while with a δ_{MSC} chosen at later minima, less and less clusters will be resolved.

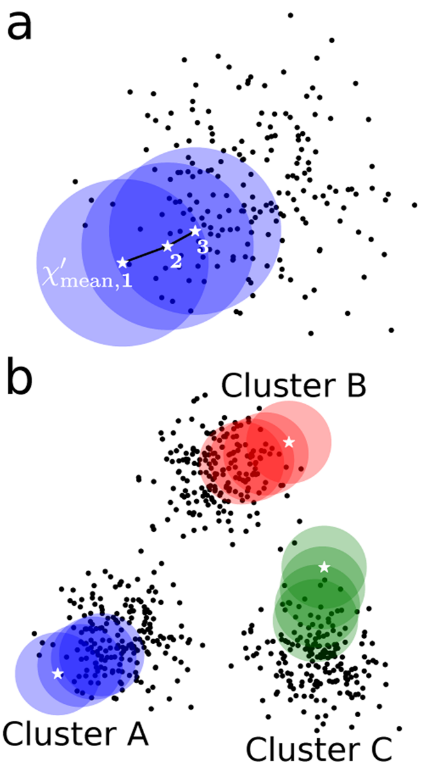

FIG. 2. (a) Illustration of iterations in a typical MSC algorithm. The white stars are the means in the first three iterations; see the text. (b) Examples of locating three cluster centers by starting the MSC iteration at different data points. The white stars are the starting points of three MSC runs.

However, while convenient, the smoothing admittedly adds in principle another empirical parameter σ_{smear} to our scheme. In practice, a suitable value for it may readily be found from visual inspection of the smoothened distribution $D(\chi'_i, \chi'_j)$. A more automatized approach recognizes that at any finite temperature, the vibrations of the atoms in the bonded structures will lead to small changes in the local environment of every atom. Time-averaged, these changes will broaden every point in MDS space and correspondingly every δ-peak in $D(\chi'_i, \chi'_j) = \|\chi'_i - \chi'_j\|_2$ into a finite Gaussian, too. A useful value for σ_{smear} may, therefore, naturally be determined by analyzing data from molecular dynamics (MD) simulations or by estimating the effect of harmonic displacements on the SOAP vectors. Using a Nose–Hoover thermostat, NVT MD data generated for a large Pd fcc bulk cell at room temperature with 0.5 fs time steps, e.g., gives the $\sigma_{\text{smear}} = 9.20 \times 10^{-3}$ that we employ in the examples below. The same MD setup was used to generate a 15 ps NVT trajectory for the island on the Pd(100) surface structure described below. Equilibration was reached after 2 ps, and 10 snapshots were extracted at random later times to analyze the performance of the framework in the case of finite temperature dynamics.

III. RESULTS

To demonstrate the versatility of the developed framework, we consider two largely different classes of structures. The first, molecular class comprises polycyclic aromatic hydrocarbons (PAHs), while the second class covers various crystalline Pd surfaces. All PAH structures are ideal and generated with nearest-neighbor C–C and C–H distances of 1.42 and 1.08 Å, respectively. C 1s Kohn–Sham values for these ideal structures were calculated with density-functional theory (DFT) using the FHI-aims package[47] and PBE functional.[48] The result is presented in the supplementary material. For the generation of the Pd surface, we employ an embedded atom potential,[49] which yields a bulk Pd–Pd nearest-neighbor distance of 2.75 Å. All surface structures are then relaxed until residual forces fall below 0.001 eV/Å, which already introduce some non-ideality requiring a fuzzy classification. All these data and the entire code used to achieve the fuzzy classifications of the examples discussed in this work can be retrieved from the EDMOND repository. Please refer to the data availability statement for the URL.

As already mentioned above, the purpose of the initial SOAP representation is to provide an accurate and non-system specific description of the local atomic environments, while the truly decisive features governing the fuzzy classification emerge in the subsequent MDS embedding step. As such, we simply set all SOAP specific parameters conservatively according to heuristics presently used in the field of ML interatomic potentials (for which SOAP was originally developed).[37] Namely, this is $n^{\text{short}}_{\text{SOAP,max}} = 8$, $l^{\text{short}}_{\text{SOAP,max}} = 4$, $n^{\text{long}}_{\text{SOAP,max}} = 4$, and $l^{\text{long}}_{\text{SOAP,max}} = 3$. $r^{\text{short}}_{\text{SOAP, cut}}$ is conveniently set to a value that corresponds to the mean between the first and the second coordination shell distance in a representative structure for the considered class, whereas $r^{\text{long}}_{\text{SOAP, cut}}$ is set at the middle of the third and fourth coordination shells. Here, the representative structures for the two classes are graphene and bulk fcc Pd, which then leads to $r^{\text{short}}_{\text{SOAP, cut}} = 1.940$ Å and $r^{\text{long}}_{\text{SOAP, cut}} = 3.550$ Å for the PAHs, and $r^{\text{short}}_{\text{SOAP, cut}} = 3.320$ Å and $r^{\text{long}}_{\text{SOAP, cut}} = 5.132$ Å for the Pd surfaces. The Gaussian width $\sigma^{\text{short/long}}_{\text{SOAP}} = r^{\text{short/long}}_{\text{SOAP, cut}}/8$ in the density representation is chosen proportional to the corresponding cutoff.

For the present illustration purposes, these SOAP settings are fully sufficient, and neither the SOAP determination step nor the entire algorithm imposes any significant computational burden. The latter could only start to change for excessively large structural databases with a huge total number N of atoms or if the classification needs to be repeated at very high frequencies. In such cases, optimizing the SOAP parameters would, of course, decrease the computational effort—at the risk of eventually leading to a too coarse initial representation of the environments. In principle, the SOAP settings may also be optimized for the classification task, e.g., following the ideas of Barnard *et al.*[50] For the present case studies, the successful classification achieved shows though that the general and simple heuristic settings provide a sufficient initial representation that is also not computationally demanding. The performance with different SOAP settings is further illustrated in the supplementary material.

A. Polycyclic aromatic hydrocarbons

Besides, the SOAP representation settings, there are only two relevant hyperparameters left in the framework, both of which tune the resolution of the final fuzzy classification, the MDS dimension S' and the MSC bandwidth δ_{MSC}. As will be seen below, the application to ideal PAH structures renders the determination of the MSC bandwidth trivial and, thus, provides a good starting point to illustrate the effect of the MDS dimensionality reduction step. The specific PAH set considered is shown in Fig. 4(b) below. It comprises benzene, naphthalene, anthracene, tetracene, phenanthren, and a graphene sheet, with a total of $N = 110$ C and H atoms. The normalized eigenvalues of the Gram matrix for this set, cf. Eq. (2), are shown in Fig. 3. Following Eq. (5), the estimated suitable MDS dimension S' is 2.

Figure 4(a) shows the 110 environments embedded in this two-dimensional space. Because of the ideal structures employed, they collapse into 11 visually easily distinguishable classes. The same classification would also be obtained with a wide range of δ_{smear} for the smoothed pair distance distribution $D(\boldsymbol{\chi}'_i, \boldsymbol{\chi}'_j)$ and the described heuristics to choose the value 0.0241 for the MSC bandwidth somewhere in the first minimum. In the PAH structures shown in Fig. 4(b), all atoms are colored according to the thus identified 11 equivalence classes. The automatized algorithm perfectly distinguishes the species and their direct coordinations, just as any human researcher would have done.

Groups 1–4 are carbon atoms with two carbon neighbors, and groups 5–7 are carbon atoms that have three carbon neighbors; groups 8–11 correspond to hydrogen atoms that all have one neighboring carbon atom. Obviously, any large enough choice of δ_{MSC}

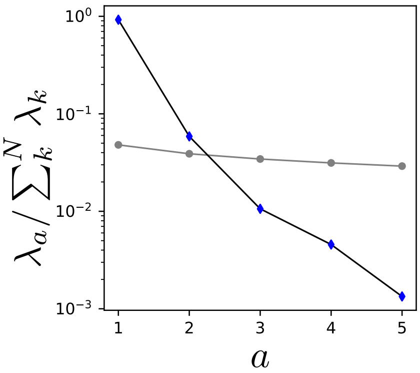

FIG. 3. The black line shows the normalized eigenvalues $\lambda_a / \sum_k^N \lambda_k$ of the Gram matrix in descending order, cf. Eq. (2), for the PAH structure set shown in Fig. 4(b) below. The gray line shows the broken-stick series. Following Eq. (5), the estimated suitable MDS dimension S' is 2.

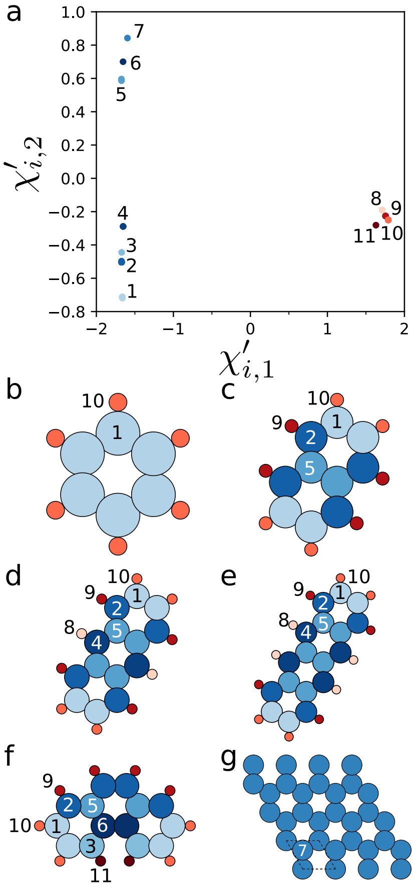

FIG. 4. (a) Two-dimensional embedded MDS space, in which the N = 110 atomic environments contained in the PAH structures collapse into 11 visually separable classes. (b) Structure models of the PAH set, with C atoms drawn as larger spheres and H atoms as smaller spheres. Each atom is colored according to the 11 equivalence classes in (a). For clarity, the corresponding class index is shown only once in each structure.

would have resulted in a clustering that only distinguishes these dominant direct C coordination differences between the three supergroups. However, in the finer resolution of eleven classes, differences in the arrangement of more distant neighboring atoms are also captured.

Which resolution in the classification is more suitable depends on the intended application. We illustrate this in the supplementary material with computed C1s Kohn–Sham levels for these molecules as a simple approximation for core-level spectroscopies.[51] Consistent with the strong difference in C coordination, the levels of each molecule are clustered into two main groups, with the (degenerate) individual peaks reflecting the subtle geometry variations between the C atoms of groups 1–4 and 5–7. It is a question of the experiment, if this substructure in the two main peaks is resolved or not, and correspondingly, which clustering bandwidth is more suitably utilized in an automatized computational spectroscopy workflow.

Intriguingly, the differences in the finely resolved 11 classes go beyond mere coordination. For instance, the H atoms in groups 9 and 10 are still resolved even though their local environments only differ in the arrangement of the second neighbor shell. Because of the subtlety of these differences, the corresponding clusters in the MDS space are admittedly very close, cf. Fig. 4(a). Yet, they are still automatically resolved by our framework— a task that would have been difficult to achieve with predefined symmetry parameters or other classification tools.

Figure 4(a) also nicely demonstrates the added benefit of an increased MDS dimension. The much larger size of the first eigenvalue of the Gram matrix in Fig. 3 could also have motivated us to just choose a one-dimensional MDS space ($S' = 1$). Then, the eleven points in Fig. 4(a) would have all collapsed onto the $\chi'_{i,1}$ axis in Fig. 4(a). At minute distances from each other, the different classes might in principle still have been distinguishable from each other. However, the second embedding dimension separates them much better. The latter would particularly become important if we consider small deviations from the ideal structures, e.g., induced by thermal vibrations. In that case, the 11 discrete points in Fig. 4(a) would spread into 11 dense groups of points (or 11 smeared out points if time-averaged MD data are used as discussed above). Then, the MSC clustering would indeed be needed. This is also the case for the Pd surface structure set, and we will conveniently discuss the effect of the corresponding δ_{MSC} hyperparameter for that class in Sec. III B.

B. Crystalline palladium surfaces

The crystalline Pd surface structure set comprises the low-index Pd(100) and Pd(111) surfaces, and the Pd(211) vicinal surface. To represent the extended surfaces, we employ periodic boundary conditions, and for convenience we use in all cases slab geometries as they would also occur in corresponding electronic structure supercell calculations. For the (100) and (111) surfaces, we, thus, use 4 slab layers, and for the (211) surface, we use 5 slab layers. To demonstrate the performance in differentiating surface environments also in more complex cases, we, furthermore, include two extra structures, namely a (13×13)-Pd(100) surface with a (7×7) square island and an adatom in a fourfold hollow site; as well as a $c(14 \times 7\sqrt{3})$-Pd(111) surface with a hexagonal island and two adatoms on fcc and hcp hollow sites. Figure 5 shows the atomic arrangement of these two nanostructures. In total, the Pd surface structure set then contains N = 1576 atoms.

Figure 6(a) shows the ordered eigenvalues of the Gram matrix for this set, cf. Eq. (2). Following Eq. (5), S' is estimated as 3. The representation of the N = 1576 atomic environments in this space is shown in Figs. 6(b) and 6(c). Not least due to the small geometric differences induced by the surface relaxation, these environments now spread out more than in the ideal PAH example. Yet, they still exhibit a substructure for which even visual inspection suggests some form of clustering. The heuristics to choose the MSC bandwidth in the first minimum of the smoothed distribution of pairwise distances in the MDS space lead to a value $\delta_{\mathrm{MSC}} = 0.0416$. The resulting clustering then identifies 17 different classes that are colored and numbered in Fig. 6(b). Analyzing these classes in more detail indicates that the primary MDS dimension χ'_1 predominantly distinguishes different

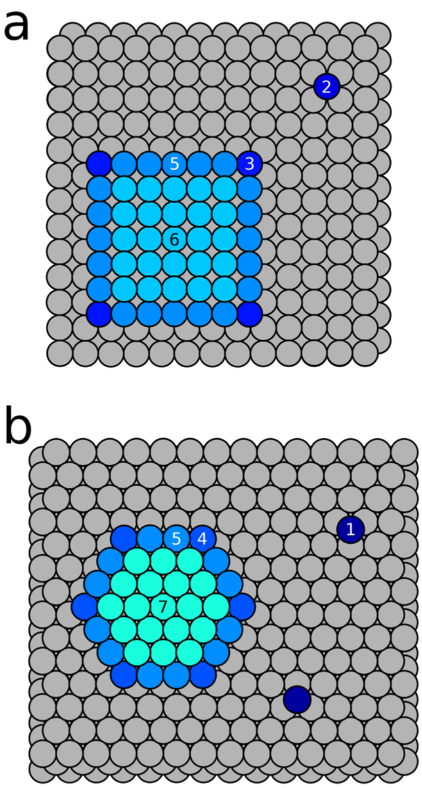

FIG. 5. Top view of the atomic arrangement of the two nanostructured surfaces contained in the crystalline Pd surface structure set: (a) a (13×13)-Pd(100) surface with a (7×7) square island and an adatom on top; (b) a $c(14 \times 7\sqrt{3})$-Pd(111) surface with a hexagonal island and two adatoms on fcc and hcp hollow sites. In both cases, groups of atoms discussed in the main text are highlighted with color according to the MSC classification of Fig. 6(b). For clarity, we restrict this coloring to the island atoms and the adatoms, and the corresponding class index is shown only once in each structure.

coordination numbers, while the other two MDS dimensions χ'_2, χ'_3 seem to more diffusively pick up longer-range arrangement.

The achieved fuzzy classification is demonstrated by color labeling all island surface atoms and adatoms in the two surface nanostructures in Fig. 5. Without relying on any human predefined symmetry parameters, the automatized classification recovers intuitive differences. Adatoms (groups 1 and 2), island corner atoms (groups 3 and 4), and island edge atoms (group 5) at the two surface symmetries are correctly distinguished. This performance also extends to the regular surface atoms of the Pd(100), Pd(111), and Pd(211) surfaces, which are all categorized as would be expected from visual inspection.

On the other hand, one also has to recognize that the resolution achieved by the heuristics is not perfect. This is most straightforwardly seen for the two adatoms on the Pd(111) surface shown in Fig. 5. Both adatoms are classified into the same group 1, even though one of them sits in a hcp hollow site and the other one sits in an fcc hollow site. At the present MDS dimension and MSC bandwidth, the framework is, thus, not able to distinguish the differences in the positioning of the second layer Pd atoms between these two sites. The same problem applies to group 5, which contains multiple types of edge atoms. As a result, the color coding of the island on the (111) surface suggests a sixfold symmetry, whereas in reality the symmetry should only be threefold (compare the position of the edge atoms to the underlying Pd terrace atoms).

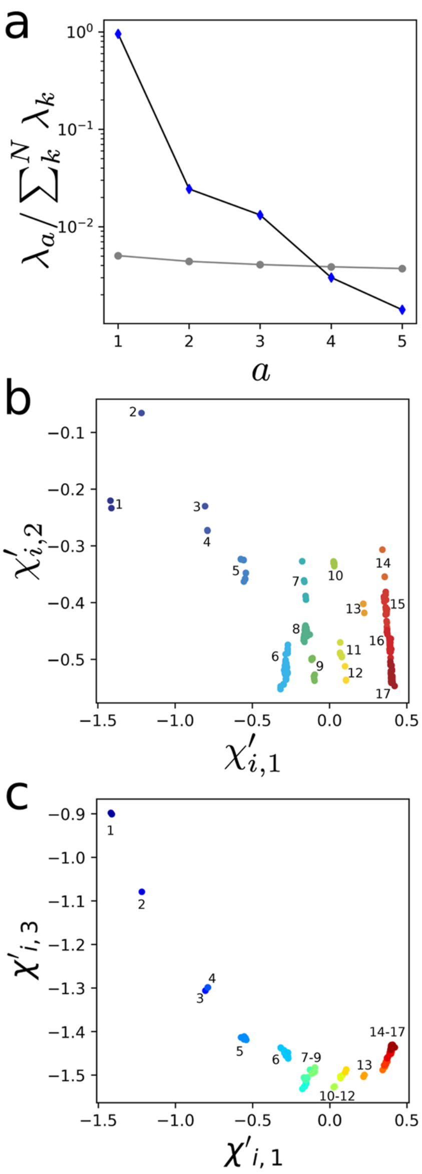

FIG. 6. (a) The black line shows the normalized eigenvalues $\lambda_a / \sum_k^N \lambda_k$ of the Gram matrix in descending order, cf. Eq. (2) for the crystalline Pd surface structure set shown in Fig. 5(b). The gray line shows the broken-stick series, following Eq. (5), S' is estimated as 3. (b) and (c) Three-dimensional embedded MDS space, in which the $N = 1576$ atomic environments contained in the Pd surface structures are drawn as individual points. The coloring of the points corresponds to the MSC clustering with a bandwidth of $\delta_{\text{MSC}} = 0.0416$ as determined by the simple heuristics; see the text. In total, 17 equivalent atom classes are identified.

C. Discussion

On the positive side, the developed framework achieves an automatized fuzzy classification even when resorting to simple heuristics for the determination of its two central hyperparameters, S' and δ_{MSC}. The performance shown for two completely different structural datasets attests to the versatility of the approach, which neither requires any system-specific input nor predefinition of the number of equivalence classes to be distinguished. On the negative side, already the second, somewhat more involved Pd surface case reveals that the resolution achieved with the heuristic hyperparameters is not perfect.

Depending on the targeted application, a classification distinguishing even finer details in the local atomic environments might be desirable. As stressed before, we see the main use of the presented fuzzy classification approach as part of a larger workflow in which active learning loops provide feedback on whether the achieved resolution is satisfactory or needs to be increased (or could even be decreased). Take, e.g., the initially mentioned application where a starting guess for a transition state search is deduced for an atom by recognizing that it has a similar local environment to another atom for which a transition state is already known. If the efficiency of such guided transition state searches turns out to be low, this indicates that atoms with too dissimilar environments are fuzzily categorized into the same equivalence class. Provided such feedback, the resolution can then be increased, which, in principle, should be achievable by increasing the MDS dimension S' and/or decreasing the MSC bandwidth δ_{MSC}.

Unfortunately, there are interdependencies between the two hyperparameters that render systematic tuning to gradually increase the resolution beyond the one achieved with the heuristic settings difficult. We illustrate this in Table I with the number of identified equivalence classes when further and further increasing the MDS dimension S' while maintaining the heuristics-based strategy to determine the MSC bandwidth from the smoothed pairwise distance distribution of the points in MDS space. As expected, the number of resolved equivalence classes does initially increase with a larger S'. However, it saturates quickly, and even in a six-dimensional space, the problematic adatom and edge atom cases discussed above are still not properly resolved. The reason for this is that the length scale of the ath dimension of the MDS space $\mathcal{S}'$ correlates with the corresponding eigenvalue $\sqrt{\lambda_a}$. Since the λ_a are ordered in descending order, the length scales of the higher MDS dimensions become smaller and smaller. This is already shown in the striped structure of the data points in the two-dimensional embeddings in Figs. 4(a) and 6(b). The length scale in the dimension χ_1' is much larger than in the dimension χ_2', and correspondingly, the data points are generally more distant from each other in the prior dimension than in the latter dimension. Now, the MSC clustering algorithm determines the mean of the data points within an S'-dimensional sphere of radius δ_{MSC}. If the distances between components of the higher MDS dimension become smaller and smaller, adding these dimensions will not help much to further distinguish clusters unless δ_{MSC} is also reduced. However, as shown in Table I the simple heuristics to determine δ_{MSC} from the smoothed pairwise distance distribution instead lead to a roughly constant value for this bandwidth in higher MDS dimensions. Indeed, simply reducing δ_{MSC} from the presently employed values to below 0.025 will immediately resolve 23 equivalence classes already in $S' = 2$.

TABLE I. Number of identified equivalence classes for the Pd surface structure set, when systematically increasing the dimension S' of the MDS space, while maintaining the heuristic determination of the MSC bandwidth δ_{MSC} described in Sec. II C. The number of 17 classes resolved for $S' = 3$ was the case discussed in Sec. III B.

S'	δ_{MSC}	No. of identified classes
1	0.0828	9
2	0.0394	17
3	0.0416	17
4	0.0416	17
5	0.0421	17
6	0.0425	18

On the other hand, just reducing δ_{MSC} is not a general purpose solution. A too small δ_{MSC} will start to distinguish atoms according to their larger distances in the primary MDS dimensions and maybe such distinction is not desired either. Take the example of the two non-resolved adatoms on the Pd(111) nanostructured surface, i.e., the two points in group 1 in Fig. 6(b). A sufficiently reduced δ_{MSC} would allow to distinguish the two. However, at such a small δ_{MSC}, the MSC algorithm will also start to differentiate the numerous bulk-like Pd atoms that currently make up the red stripe at the bottom right in Fig. 6(b)— and adding further MDS dimensions will not mitigate this problem at all. Alternatively, one could imagine re-scaling the MDS dimensions by their eigenvalue to achieve more comparable length scales in all MDS dimensions. However, the diminishing length scales of higher MDS dimensions have a meaning. They reflect that differences in these dimensions correspond to more and more subtle differences in the local atomic environments. Blowing up these differences by simply renormalizing the MDS length scales might, therefore, neither be a generally applicable remedy to increase the resolution in the desired way. In the end, the careful tuning of both hyperparameters, S' and δ_{MSC}, will be required if the default heuristics do not achieve a satisfactory fuzzy classification for a specific application. The classification performance with the heuristics utilizing a different σ_{smear} is further demonstrated in the supplementary material.

A noteworthy positive feature of the developed framework is that new structures may readily be evaluated within a once achieved fuzzy classification. As long as exactly the same SOAP settings to initially describe the atomic environments are employed, the embedding operator $\mathbf{P}$ of this classification will project any environment contained in the new structure to the low-dimensional MDS space $\mathcal{S}'$. The corresponding new data point $\boldsymbol{\chi}'_{\mathrm{new}}$ can then straightforwardly be assigned to the nearest cluster center.

We explore this idea for a room-temperature MD trajectory generated for the Pd(100) island structure. 10 snapshots are extracted at random times, and the environments of all surface atoms are categorized in terms of the 17 equivalence classes of the Pd surface structure set discussed above. As summarized in the supplementary material, adatoms, island edges, and surface atoms are correctly identified with a 90% or higher probability despite the thermal displacements. More problematic are only the island corner atoms with their larger anisotropic vibrations, which are miscategorized with a 50% probability. One option that we currently pursue to improve this could be to exploit correlations in the

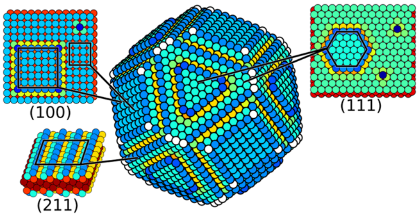

FIG. 7. Categorization performance for new structures within a once achieved fuzzy classification. The 8523 atoms of the shown Pd nanoparticle are colored according to the 17 equivalence classes of the Pd surface structure discussed above, cf. Fig. 6(b). For comparison, the Pd(211) surface and the two surface nanostructures of this original structure set are also shown in the same coloring, readily allowing the framework to identify similar local environments. White atoms close to facet edges of the Pd nanoparticle are classified as distinct from all classes of the existing classification, see the text.

classification of the individual atoms in successive snapshots along the trajectory.

An alternative ansatz, not only for the case of finite temperature dynamics, is to iteratively expand a given fuzzy classification with such new structures. In this case, an atomic environment would be identified as distinctly different from all previously considered environments if the corresponding new data point χ' is located further away from any existing cluster in MDS space than the MSC bandwidth δ_{MSC}. In Fig. 7, this is illustrated for the 8523 atoms of a Pd nanoparticle again within the 17 previously discussed equivalence classes. In this case, all edge atoms between the nanoparticle facets are categorized as a new environment. Once one or a sufficient number of such new environments are identified, a flag could be set in an iterative framework that initiates a new fuzzy classification now involving the entire, augmented structure set.

IV. CONCLUSIONS

We presented an automatized machine-learning framework to identify atoms with (near-)equivalent local atomic environments in any one or a set of given structures. The required fuzziness in the classification is achieved by embedding an initial high-dimensional representation of the local environment and a subsequent clustering in the resulting low-dimensional space. Emphasis was placed on the high versatility of the framework with minimum system specific input. As such, the framework is readily applicable to molecular structures, extended materials, or interfaces, to ideal or non-ideal, as well as to crystalline or amorphous geometries. Simple heuristics are provided for the two central hyperparameters, the dimension S' of the MDS embedding space and the bandwith δ_{MSC} of the MSC clustering in this space. If the resolution achieved with the heuristic settings is not optimum for a specific application, the two hyperparameters are tunable with understandable effects on the resulting classification. The framework could, therefore, readily be integrated into larger workflows that achieve optimum tuning of the hyperparameters, e.g., in active-learning iterations evaluating the classification performance for the targeted application. A sample implementation of the framework as a standalone application is provided in the repository stated below.

The versatility of the framework also extends to its capability to assess new structures within an achieved fuzzy classification. In addition to this end, one can, therefore, imagine an iterative usage in which the atoms of new structures are first assessed and a new fuzzy classification of the increased set of structures is initiated whenever a critical number of new, distinctly different atom classes has been identified. We also note that the initial high-dimensional fingerprint for each atom is not necessarily restricted to the structure-sensitive (double) SOAP vector employed in this work. This representation was chosen here within the focus on equivalence in the local atomic environments. Other fingerprints, such as partial charges or other electronic structure properties may, e.g., be added to further improve the resolving capabilities of the framework, or directly be used to base the fuzzy classification on aspects other than geometric similarity.

SUPPLEMENTARY MATERIAL

Please see the supplementary material for demonstrations of other dimensionality reduction methods (kPCA, Sketch Map); another clustering method (HDBSCAN). In addition, classification demonstration was applied to MD data of structured Pd(100) surfaces.

ACKNOWLEDGMENTS

K.C.L. gratefully acknowledges the Alexander von Humboldt Foundation research fellowship.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

King Chun Lai: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Sebastian Matera**: Conceptualization (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal). **Christoph Scheurer**: Conceptualization (equal); Investigation (equal); Methodology (equal). **Karsten Reuter**: Methodology (equal); Project administration (lead); Supervision (lead); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are openly available in EDMOND at http://doi.org/10.17617/3.U7VKBM.

REFERENCES

[1] M. Andersen, C. Panosetti, and K. Reuter, "A practical guide to surface kinetic Monte Carlo simulations," Front. Chem. 7, 202 (2019).

[2] K. Reuter, "Ab initio thermodynamics and first-principles microkinetics for surface catalysis," Catal. Lett. **146**, 541–563 (2016).

[3] J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C. J. Eom, J. R. Lunger, D. J. Zheng, R. R. Rao *et al.*, "Human-and machine-centred designs of molecules and materials for sustainability and decarbonization," Nat. Rev. Mater. **7**, 991–1009 (2022).

[4] W. Xu, K. Reuter, and M. Andersen, "Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation," Nat. Comput. Sci. **2**, 443–450 (2022).

[5] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[6] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).

[7] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-i. Shimizu, "Machine learning for catalysis informatics: Recent applications and prospects," ACS Catal. **10**, 2260–2297 (2019).

[8] A. Bhowmik, I. E. Castelli, J. M. Garcia-Lastra, P. B. Jørgensen, O. Winther, and T. Vegge, "A perspective on inverse design of battery interphases using multi-scale modelling, experiments and generative deep learning," Energy Storage Mater. **21**, 446–456 (2019).

[9] S. N. Steinmann and Z. W. Seh, "Understanding electrified interfaces," Nat. Rev. Mater. **6**, 289–291 (2021).

[10] S. Stocker, G. Csányi, K. Reuter, and J. T. Margraf, "Machine learning in chemical reaction space," Nat. Commun. **11**, 5505 (2020).

[11] C. Schober, K. Reuter, and H. Oberhofer, "Virtual screening for high carrier mobility in organic semiconductors," J. Phys. Chem. Lett. **7**, 3973–3977 (2016).

[12] K. H. Sørensen, M. S. Jørgensen, A. Bruix, and B. Hammer, "Accelerating atomic structure search with cluster regularization," J. Chem. Phys. **148**, 241734 (2018).

[13] S. A. Meldgaard, E. L. Kolsbjerg, and B. Hammer, "Machine learning enhanced global optimization by clustering local environments to enable bundled atomic energies," J. Chem. Phys. **149**, 134104 (2018).

[14] S. Chiriki, M.-P. V. Christiansen, and B. Hammer, "Constructing convex energy landscapes for atomistic structure optimization," Phys. Rev. B **100**, 235436 (2019).

[15] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," J. Chem. Phys. **120**, 9911–9917 (2004).

[16] C. J. Pickard and R. J. Needs, "*Ab initio* random structure searching," J. Phys.: Condens. Matter **23**, 053201 (2011).

[17] A. Banerjee, S. Maity, and C. H. Mastrangelo, "Nanostructures for biosensing, with a brief overview on cancer detection, IoT, and the role of machine learning in smart biosensors," Sensors **21**, 1253 (2021).

[18] S. Stegmaier, R. Schierholz, I. Povstugar, J. Barthel, S. P. Rittmeyer, S. Yu, S. Wengert, S. Rostami, H. Kungl, K. Reuter *et al.*, "Nano-scale complexions facilitate Li dendrite-free operation in LATP solid-state electrolyte," Adv. Energy Mater. **11**, 2100707 (2021).

[19] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[20] H. Huo and M. Rupp, "Unified representation of molecules and crystals for machine learning," Mach. Learn.: Sci. Technol. **3**, 045017 (2022).

[21] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).

[22] X. Chen, M. S. Jørgensen, J. Li, and B. Hammer, "Atomic energies from a convolutional neural network," J. Chem. Theory Comput. **14**, 3933–3942 (2018).

[23] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials," Chem. Rev. **121**, 9759–9815 (2021).

[24] J. D. Carroll and P. Arabie, *Multidimensional Scaling* (Elsevier, 1998), pp. 179–250.

[25] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans. Inf. Theory **21**, 32–40 (1975).

[26] P. Gasparotto and M. Ceriotti, "Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond," J. Chem. Phys. **141**, 174110 (2014).

[27] P. Gasparotto, R. H. Meißner, and M. Ceriotti, "Recognizing local and global structural motifs at the atomic scale," J. Chem. Theory Comput. **14**, 486–498 (2018).

[28] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From DFT to machine learning: Recent approaches to materials science–a review," J. Phys.: Mater. **2**, 032001 (2019).

[29] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg *et al.*, "QSAR without borders," Chem. Soc. Rev. **49**, 3525–3564 (2020).

[30] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv. **3**, e1701816 (2017).

[31] N. Bernstein, B. Bhattarai, G. Csányi, D. A. Drabold, S. R. Elliott, and V. L. Deringer, "Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon," Angew. Chem. **131**, 7131–7135 (2019).

[32] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," Comput. Phys. Commun. **247**, 106949 (2020).

[33] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).

[34] Y. Bengio, J.-f. Paiement, P. Vincent, O. Delalleau, N. Roux, and M. Ouimet, "Out-of-sample extensions for LLE, IsoMap, MDS, eigenmaps, and spectral clustering," Adv. Neural Inf. Process. Syst. **16**, 177–191 (2003).

[35] J. C. Gower, "Adding a point to vector diagrams in multivariate analysis," Biometrika **55**, 582–585 (1968).

[36] I. T. Jolliffe, "Choosing a subset of principal components or variables," in *Principal Component Analysis* (Springer, 2002), pp. 111–149.

[37] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csanyi, "Mapping materials and molecules," Acc. Chem. Res. **53**, 1981–1991 (2020).

[38] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput. **10**, 1299–1319 (1998).

[39] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," Proc. Natl. Acad. Sci. U. S. A. **108**, 13023–13028 (2011).

[40] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," J. Chem. Theory Comput. **9**, 1521–1532 (2013).

[41] A. Glielmo, I. Macocco, D. Doimo, M. Carli, C. Zeni, R. Wild, M. d'Errico, A. Rodriguez, and A. Laio, "DADApy: Distance-based analysis of data-manifolds in Python," Patterns **3**, 100589 (2022).

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[43] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd* (AAAI, 1996), Vol. 96, pp. 226–231.

[44] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," J. Open Source Software **2**, 205 (2017).

[45] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14–17, 2013, Proceedings, Part II* (Springer, 2013), Vol. 17, pp. 160–172.

[46] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 888–905 (2000).

[47] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "*Ab initio* molecular simulations with numeric atom-centered orbitals," Comput. Phys. Commun. **180**, 2175–2196 (2009).
[48] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).
[49] S. M. Foiles, M. I. Baskes, and M. S. Daw, "Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, and their alloys," Phys. Rev. B **33**, 7983 (1986).
[50] T. Barnard, S. Tseng, J. P. Darby, A. P. Bartók, A. Broo, and G. C. Sosso, "Leveraging genetic algorithms to maximise the predictive capabilities of the soap descriptor," Mol. Syst. Des. Eng. **8**, 300 (2023).
[51] G. S. Michelitsch and K. Reuter, "Efficient simulation of near-edge x-ray absorption fine structure (NEXAFS) in density-functional theory: Comparison of core-level constraining approaches," J. Chem. Phys. **150**, 074104 (2019).