npj | Computational Materials www.nature.com/npjcompumats

ARTICLE OPEN

Materials structure genealogy and high-throughput topological classification of surfaces and 2D materials

Lauri Himanen[1], Patrick Rinke[1] and Adam Stuart Foster[1,2,3]

Automated and verifiable structural classification for atomistic structures is becoming necessary to cope with the vast amount of information stored in various computational materials databases. Here we present a general recursive scheme for the structural classification of atomistic systems and introduce a structural materials map that can be used to organize the materials structure genealogy. We also introduce our implementation for the automatic classification of two-dimensional structures, especially focusing on surfaces and 2D materials. This classification procedure can automatically determine the dimensionality of a structure, further categorize the structure as a surface or a 2D material, return the underlying unit cell and also identify the outlier atoms, such as adsorbates. The classification scheme does not require explicit search patterns and works even in the presence of defects and dislocations. The classification is tested on a wide variety of atomistic structures and provides a high-accuracy determination for all of the returned structural properties. A software implementation of the classification algorithm is freely available with an open-source license.

npj Computational Materials (2018)4:52 ; doi:10.1038/s41524-018-0107-6

INTRODUCTION

Materials science is entering the data age. This transition is spearheaded by projects such as the Materials Genome Initiative,[1] the Novel Materials Discovery Laboratory[2] and Marvel[3] that combine high-throughput screening with data storage, systematic data curation and machine learning. Such projects produce computational materials databases that contain information extracted from atomistic simulations, e.g., system geometries, details of the applied theory, electronic structures, methodology and implementation and their number is increasing rapidly.[4–14]

A common problem in these databases is materials classification. Often database users would like to search for specific material types, specific functions or structural classes, such as crystals, molecules, surfaces or 2D materials. To facilitate such searches, the database entries should be tagged according to a classification system. Unfortunately, classifications are not always provided when the data are uploaded to the database, and when they are, they are often based on custom or unspecified definitions. To cope with large heterogeneous datasets from atomistic calculations, automated and verifiable methods for analyzing and categorizing atomistic structures have become a necessity.

Previous work on automated classification of atomistic structures has focused on very specific areas and often required an explicit structural search pattern. For example, defect identification and detection schemes have been developed for crystals that are based on neighbourhood analysis.[15–20] In another example, a more automated workflow was applied to identify lower dimensional stable structures in crystals, such as layered solids.[21] Conversely, many tools are available for the inverse problem that generates an atomistic representation from a given structure definition. Tool sets such as the atomic structure environment (`ase`)[22] and `pymatgen`[23] include routines for automating tasks like creating a surface given a lattice, orientation and number of layers, generating crystal structures with desired symmetry properties or generating a system representing surface adsorption, given an adsorbate and an adsorbant.

In this work, we focus on structural classification and present a generic structure classification scheme that encompasses all possible structure types. We then introduce a materials structure genealogy presented as an intuitive and human-readable materials structure 'tree of life'. After this general introduction, we present an automatic and accurate classification scheme for two-dimensional structures, including surfaces and 2D materials, that requires no explicit search patterns. This classification process also returns the underlying unit cell and works even in the presence of defects, dislocations and additional atoms. By being able to identify the unit cell, these structures can be meaningfully characterized and often linked to their bulk counterpart. We also present methods that can be used to accurately identify the outlier atoms that are not part of the underlying structure. The classification does not make assumptions about the used cell or the positioning of the structure within the cell.

The NOMAD Archive[2] is used as a benchmark for testing the classification accuracy and the applicability of our method in a realistic database environment containing heterogeneous data. The classification tools are implemented as a python library licensed under the open-source Apache 2[24] license, and the source code together with installation instructions can be found from https://github.com/SINGROUP/matid. This library is directly compatible with the popular atomic structure manipulation library ase.

[1]Department of Applied Physics, Aalto University, P.O. Box 11100, 00076 Aalto, Espoo, Finland; [2]Graduate School Materials Science in Mainz, Staudinger Weg 9, Mainz 55128, Germany and [3]WPI Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan
Correspondence: Lauri Himanen (lauri.himanen@aalto.fi)

Received: 25 April 2018 Revised: 8 August 2018 Accepted: 22 August 2018
Published online: 11 September 2018

npj nature partner journals

RESULTS

Topological classification

Often materials are classified based on a set of functional and/or chemical properties. This kind of classification can be based on a single property, such as the bandgap in the case of conductors, semiconductors and insulators, or it can be based on a collection of multiple functional and chemical properties like in the case of material types, such as ceramics, metals, polymers and composites.

In an analogous way, materials can be classified based on their structural properties. This is here referred to as topological classification. By topological classification we mean a classification that is based on structural features and properties that can be deduced from static atomistic geometries. This thus excludes structural phenomena that are dynamic in nature or depend on some external perturbation. Unlike functional or chemical properties, structural properties are often not so easily quantifiable. Some combinations of specific structural or electronic quantities have been shown to separate different structural types to provide 'structure maps'.[25,26] Such maps are, however, too crude to distinguish the full variety of the structural space. The task of structural classification is for these reasons often left to humans. As the systematic exploration of different materials is increasing, human labelling becomes inefficient and error-prone. For this reason, it is important to develop systematic concepts and tools for this task.

The general rules surrounding structural classification have not been discussed systematically before in the literature, and to fill this gap we introduce a recursive scheme based on primary building blocks and their higher level combinations. Primary building blocks are the lowest-level structural building units. One can recursively combine these building blocks to create higher order combinations. An example of common primary building blocks and higher order combinations is given in Fig. 1. For example, a polymer can be combined with water molecules to form a polymer–water solution, which can again be combined with a surface to form the final system of a polymer–water solution on a surface. This systematically extendable classification scheme can be used to describe arbitrarily complex structures.

Structures can be further grouped by the presence of certain building blocks and additional structural properties, such as dimensionality and long-range order. For example, a heterostructure, that consists of multiple stacked two-dimensional building blocks, can be categorized as a 2D or 3D heterostructure depending on whether there is a vacuum separating the periodic copies or not. This grouping forms structural families that can be better visualized in a tree structure. This materials genealogy is presented in Fig. 2.

Given an arbitrary atomic structure, an ideal classifier would detect the different building blocks, their interrelation and other structural features and be able to place the structure on the correct branch in the materials genealogy. The classifier should have the following generic properties:

Accuracy—The classifier should be able to perform accurate classification on a wide range of samples that have not been explicitly tested.

Invariance—The classifier should be able to correctly classify structures independent of the chosen unit cell, the basis vector directions and the relative position of the structure with respect to the unit cell.

Robustness—Small displacements of atoms from their expected positions should not disrupt the classification.

Efficiency—The time taken for the classification should be small enough that processing of a large data set is possible in a reasonable time.

It is worth noting that some of the features are not independent. For example, in order to reach good accuracy for a wide range of structures, the computational efficiency may suffer.

Classification for surfaces and 2D materials

Here we demonstrate the application of a topological classifier for identifying structures in the two-dimensional branch of the materials genealogy. Figure 3 shows a high-level overview of this classification procedure, and the reader is referred to the methods section for an in-depth discussion of the steps shown in the figure.

We apply the classification routine to atomic structures from the NOMAD Archive.[2] We chose the NOMAD Archive because it contains atomic structures and calculation results from a variety of electronic structure codes. The structures we used in this work originate from density functional theory simulations and were

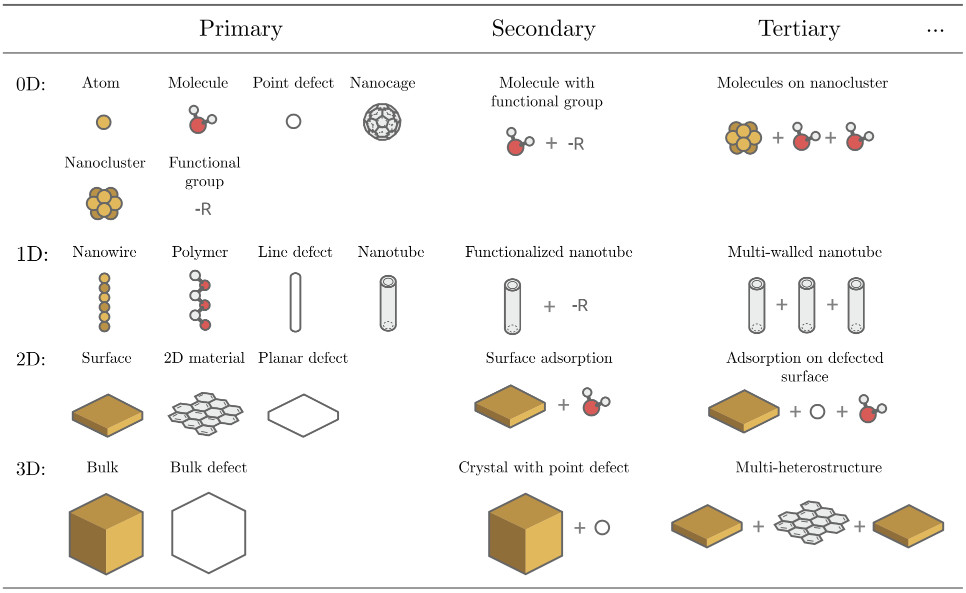


Fig. 1 Illustration of commonly used primary building blocks and their higher order combinations. The structures are here organized by dimensionality in the vertical direction and the cardinality of the used building blocks horizontally. The higher order combinations are specified by the involved building blocks and their relation to each other

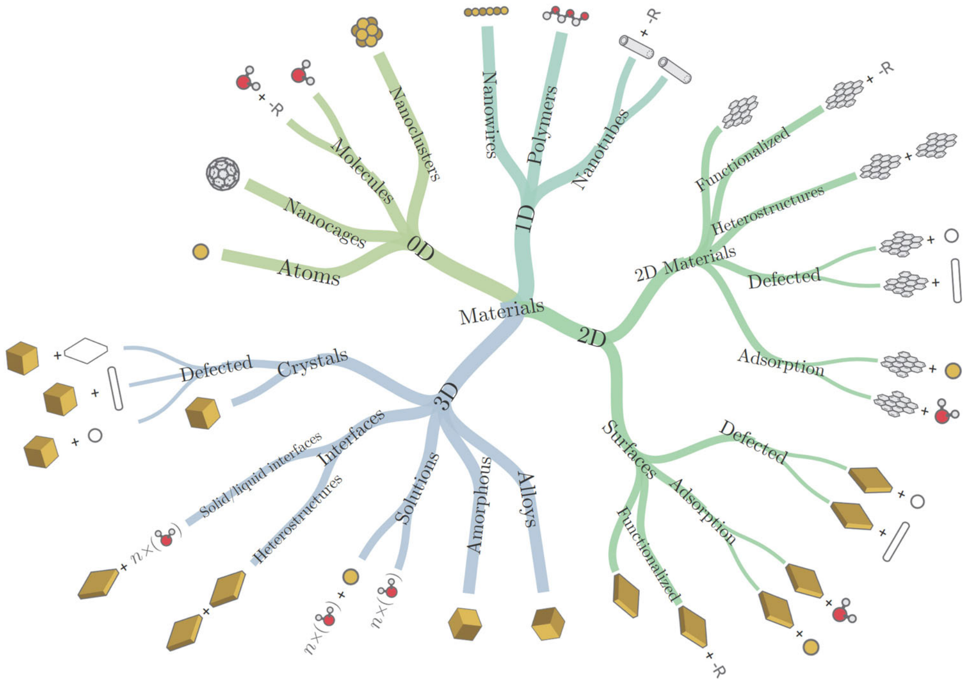


Fig. 2 Illustration of our topological materials genealogy. The first-classification level is based on the dimensionality of the structures, and further branches distinguish structural building blocks and other structural properties, like the presence of long-range order. The branches have been labelled with more common expressions for certain structure types. This map is not exhaustive, but illustrates the concept and includes many of the most commonly encountered structural categories

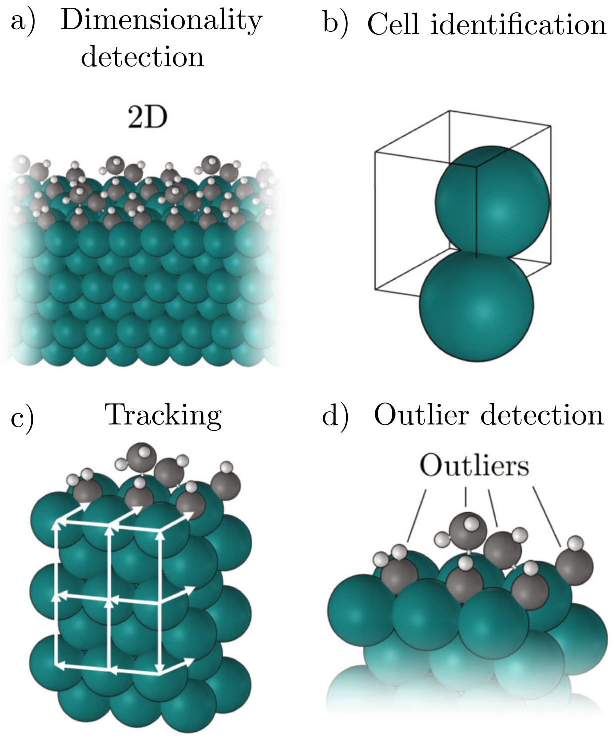


Fig. 3 Overview of the classification procedure for two-dimensional structures. **a** The classification starts with detecting the dimensionality of the system. **b** If a two-dimensional structure is detected, the classification proceeds to find a unit cell that best describes the structure. **c** This unit cell is then used to track all atoms that belong to the material. **d** In the final step, the outlier atoms are identified

specifically calculated either with the electronic structure program Exciting[27] or FHI-aims.[28–30] Both these codes have a significant number of calculations present in the database for two-dimensional systems, but the classification is generally applicable.

To cover a wide range of different structures, a cross-validation of the parameters was performed for each structure individually as described in section Parameters and generalizability. For each atomic element a representative atom, which is nearest to the centre of mass, was tested as a seed atom. The values $0.25d_{\mathrm{min}}$ and $0.75d_{\mathrm{min}}$ were tested for the position tolerance Δr. Here d_{min} is the minimum distance between two atoms in the structure. The value 12 Å was used for the maximum cell size r_{max}.

From an original data set of 394,332 structures, 35 were recognized as zero-dimensional, 13 as one-dimensional, 14,959 as two-dimensional, 376,756 as three-dimensional and 2569 could not be classified due to multiple disconnected components as described in section Dimensionality detection. For each two-dimensional structure, we run the full-classification procedure including the cell and outlier detection. The classified structures were then grouped by the chemical formula of the possible outliers and the chemical formula of the rest of the system. For each unique pair of these two chemical formulas, a representative sample was chosen resulting in 192 unique structures. These structures were visually inspected to assess the correctness of the classification, the material type, the found cells and the detected outliers were checked. A breakdown of the classification results is shown in Fig. 4.

In the visual inspection, we follow the same material definitions as used in the algorithm, surfaces should have at least two repetitions of the unit cell in three linearly independent directions and 2D materials should have at least two repetitions in two linearly independent directions restricted by a maximum thickness $h^{\mathrm{2D}}_{\mathrm{max}}$. Seven structures were found to be unclassifiable into surfaces or 2D materials both by the algorithm and by visual inspection. From the remaining set, 183/185 ≈ 99% were correctly classified into a surface or 2D material and had a correct cell and correct outliers. Two structures were misclassified as false positives by the algorithm. In one, an incorrect cell and thus a wrong outlier was detected, and in the other, all the correct outliers were not identified. Examples of the correctly classified, incorrectly classified and unknown structures are given in Figs. 5 and 6.

DISCUSSION

When a cross-validation of the parameters is performed for each structure individually, a very good accuracy for both the cell identification and outlier identification is achieved. The found structures cover many different cell geometries and types of structures, include optimized structures with deviations from perfect lattice symmetry, and we were able to analyze large quantities of structures efficiently.

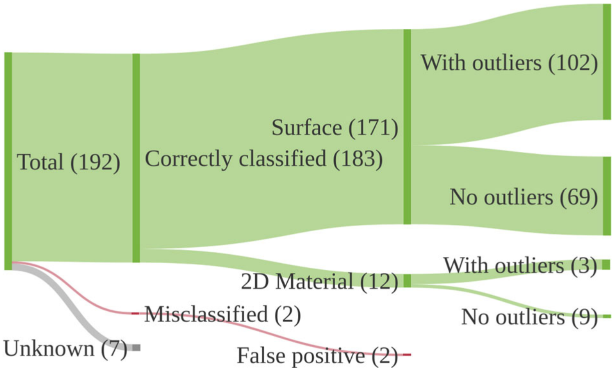


Fig. 4 Distribution of the classification results for 192 visually inspected two-dimensional systems from the NOMAD Archive

The structure in Fig. 6a was misclassified because a wrong unit cell was selected and thus also a wrong outlier was detected. Further analysis reveals that the unit cell is missing one carbon atom because it is rejected from the unit cell due to its too low average degree ⟨deg(*v*)⟩ in the corresponding connected component. One can decrease the threshold value n^{e}_{min} to get this atom included in the cell, but this will lead to problems in distinguishing real outliers. This issue could be avoided by requiring more repetitions of the cell to gain better statistics about the cell contents, but that will prevent the classification of many small structures.

Our chemical similarity measure proved to be highly accurate in detecting outliers. The data set contained multiple systems, one example shown in Fig. 5e, where adsorbates would not have been correctly detected if the chemical environment would not have been checked. Only in the system shown in Fig. 6b, the carbon atom of the carbon dioxide adsorbate was not fully detected. The carbon in question is directly on top of an expected location for a carbon in the lattice and is highly connected to the surface itself. Tweaking the similarity threshold ΔC or using a more sophisticated chemical environment measure, such as SOAP[34] or ACSF,[35] might resolve the problem partially. It is, however, likely that perfect outlier detection cannot be reached without explicit information of the expected outlier structures.

The structures in Fig. 6c, d represent surface-like structures that do not meet our classification criteria. Both are thicker than the height threshold $h^{\text{2D}}_{\text{max}}$ set for 2D materials and do not have two full repetitions of a unit cell in three linearly independent directions.

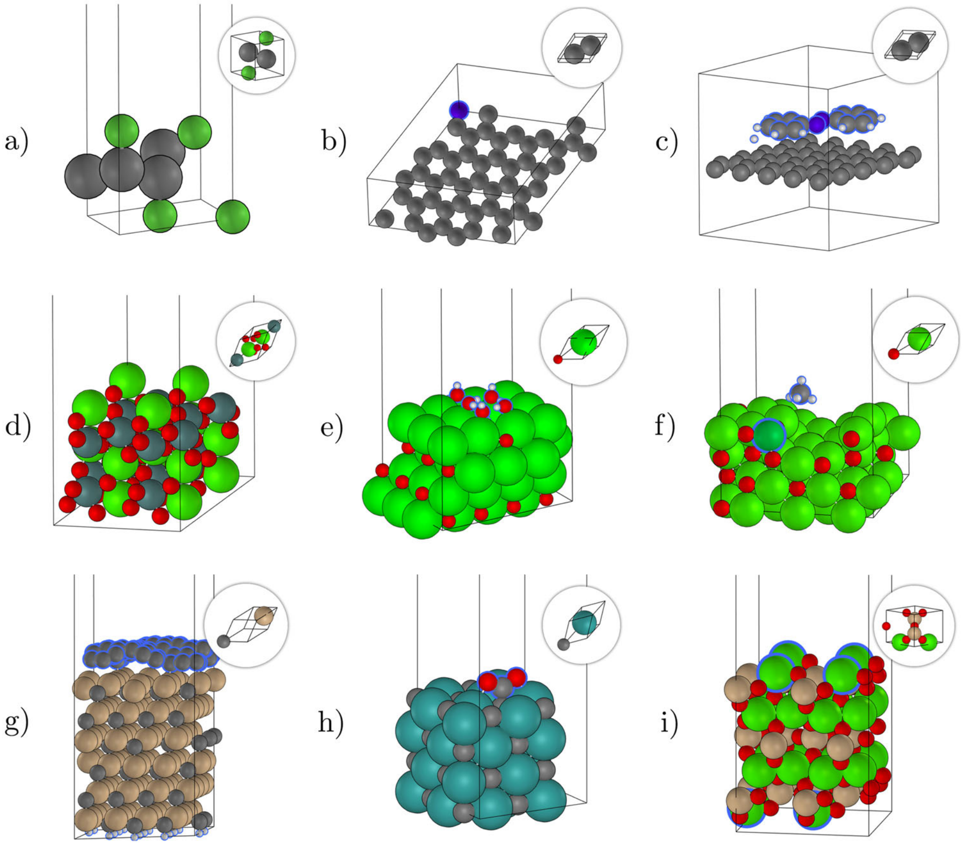


Fig. 5 Examples of correctly classified structures. The top right corner in each image shows a primitive cell corresponding to the unit cell identified by the algorithm. All detected outlier atoms are highlighted in blue. **a** Pristine 2D material. **b** 2D material with substitution. **c** 2D material with adsorbate. **d** Pristine surface. **e** Surface with multiple adsorbates. **f** Stepped surface with a substitution and adsorbate. **g** Surface with two different terminations. **h** Chemisorption. **i** Surface with a reconstruction

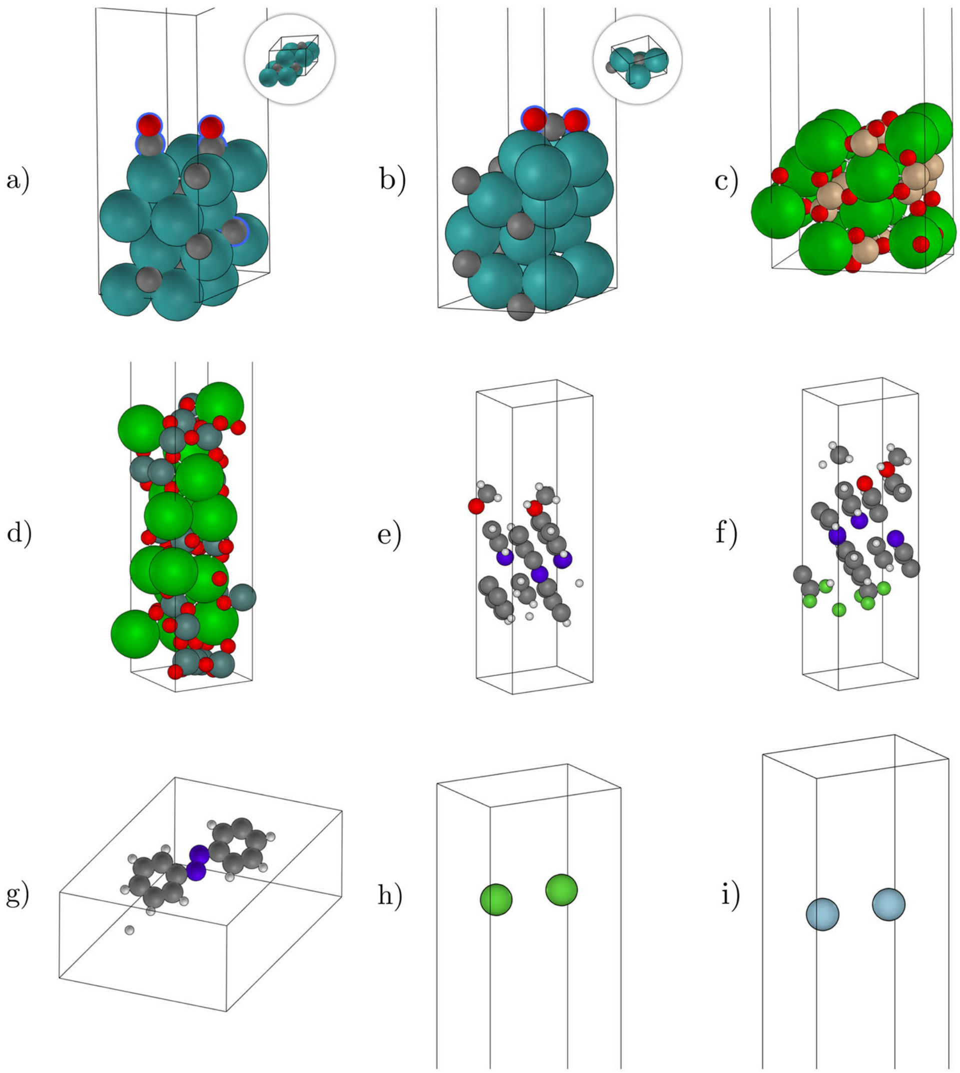


Fig. 6 Structures that were misclassified or labelled unknown. **a** Wrong cell and a wrong outlier detected. **b** All outliers not detected. **c**–**i** Unknown classification

Figure 6e–g shows two-dimensional structures that are comprising multiple networked molecules. These molecular networks fail the cell validity checks, as mentioned in section Cell validation, because they have a sparse unit cell. The classification of such sparse structures depends on the threshold r_{bond}, and depending on its value they can be regarded as 2D materials if other criteria are fulfilled. For now these molecular networks are given a generic two-dimensional classification both by the algorithm and by visual inspection. Figure 6h, i shows two additional structures that could not be classified by the algorithm or by visual inspection due to having a too sparse unit cell.

The centre of mass proves to be a good starting point for selecting the seed atoms, and by default the algorithm automatically tries multiple seed atoms with different atomic elements near the centre of mass. It is important to bear in mind that the choice of the seed atom is critical to the success of the algorithm, and generally multiple seed atoms should be tested. For example, if the seed atom is chosen to be a substitutional atom in a lattice, the algorithm will stop almost immediately without finding the correct unit cell. Also, if the seed atom is chosen to be located within a region with multiple substitutions or dislocations, the correct unit cell cannot be obtained.

To robustly detect the most common material classes, the classification procedure considers displacements that are below the user-specified threshold as noise. Some material characteristics may, however, arise from systematic displacements that are smaller than this displacement threshold. The implementation can be extended in the future to keep track of the displacement vectors for individual atoms, or to track distortions in the shapes of individual unit cells. This would allow building vector maps where different displacement patterns could be used to classify structures in more detail.

To reduce the need for manual parameter tuning, a wide range of parameter combinations are tested by default, and the best parameters are chosen by minimizing the number of found outlier atoms with some additional restrictions. So far the use of a systematic grid-search has been enough to find a good parameter combination, but in cases where a large parameter space is required, this optimization problem could be more efficiently solved with probabilistic optimization methods, such as Bayesian optimization.

Providing automatic and tractable topological classification for large quantities of atomic structures has become a challenge, as the number of new materials databases and their data volume is steadily increasing. We have introduced a general and systematic approach for the topological classification of atomistic structures and have developed an automatic, robust and accurate way for identifying two-dimensional structures, detecting the underlying

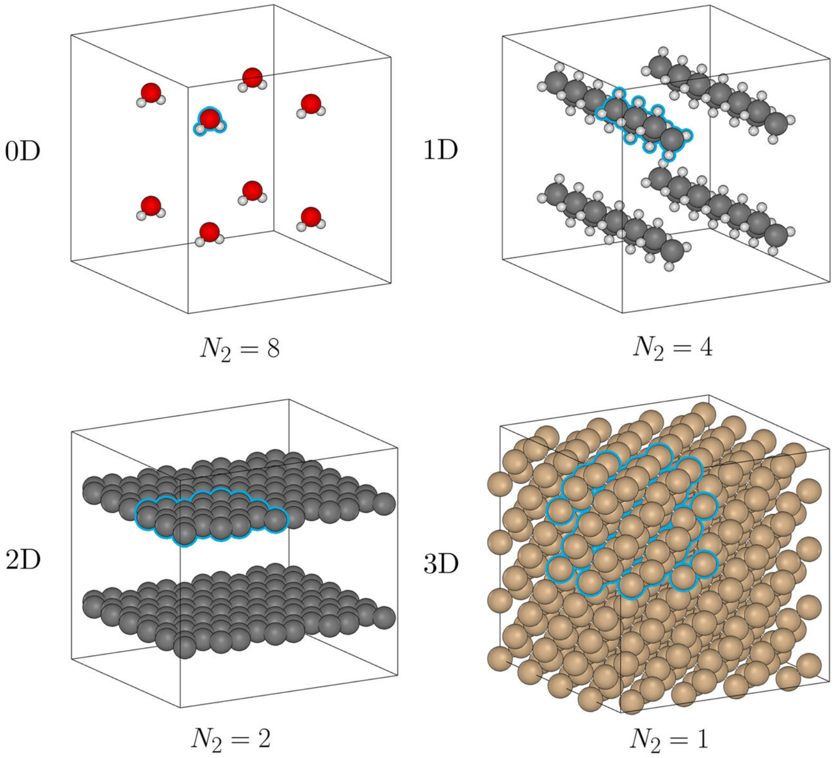


Fig. 7 Illustration of dimensionality detection. The images show a 2 × 2 × 2 supercell for a system with three periodic directions ($n_{\text{pbc}} = 3$) where atoms in the original simulation cell are highlighted in blue. The analysis of how the number of spatially separated clusters changes when going from the original cell to the supercell, reveals that zero-dimensional structures will have eight separate clusters, one-dimensional have four, two-dimensional have two and three-dimensional structures will have only one cluster. The dimensionality depends on this scaling and is given by $D = n_{\text{pbc}} - \log_n(N_n)$, where n is the number of repetitions for the system, N_n is the number of clusters in the supercell and n_{pbc} is the number of periodic directions

unit cell and outlier atoms in them. The method has been validated on realistic data taken from a database of DFT calculations performed with different programs and computational methods and containing a wide variety of structures. Our implementation can be easily integrated into any existing database that provides atomistic geometries.

One important remaining task in the two-dimensional branch is the further categorization of outlier atoms into more specific groups, such as adsorbates, substitutions or interstitials. This would provide one more level of detail when classifying structures according to the structural genealogy. Also, extending the automatic classification for other branches in the materials genealogy is a major remaining challenge. One possibility is to apply the concepts introduced for the identification of unit cells in classifying multi-component systems such as heterostructures or crystals with defects.

An interesting alternative for structural classification is using machine learning with pre-labelled data as training material. The level of detail our method is able to achieve, including the identification of the unit cell and outlier atoms, might be hard to match by a machine learning approach. However, a supervised learning technique could be envisioned for the classification of different structural families of the materials genealogy, which does not require the detailed knowledge of the unit cell or the exact identification of outliers. The introduced methods also offer a more automated approach for linking material properties to structural features by correlating structural building blocks and structural features with materials phenomena. By using automated tools, the user does not require extensive knowledge of the original simulation setup to investigate how different physical properties, such as adsorption energies or electronic features, relate to different structural properties, such as the location and type of adsorbate, the unit cell or different defects.

METHODS

Here we introduce methods for the automatic classification of two-dimensional structures especially focusing on surfaces and 2D materials, with or without outlier atoms, such as adsorbates.

Dimensionality detection

Dimensionality detection is in principle easy, as a zero-dimensional system has no periodicity, one-dimensional system is periodic in only one direction and so forth. The dimensionality is thus given by the number of lattice vectors with periodic boundary conditions. However, in many electronic structure theory codes, it is common to represent even non-periodic structures such as molecules, clusters, nanowires or surfaces with periodic boundary conditions, in particular if plane waves are used as a basis set. For this reason, we can no longer use the number of periodic directions as a criterion. Instead we must analyze the extent of vacuum that is added to the simulation cell to decouple the original structure from its periodic images in directions that are not meant to be periodic (see Fig. 7 for examples).

To determine the dimensionality of a system, we use a modified version of the topological scaling algorithm (TSA).[21] The basic idea behind the algorithm is illustrated in Fig. 7, and it is based on analyzing the size scaling of atomic clusters when going from the original system to a bigger supercell of the same system. With TSA, the dimensionality D is given by

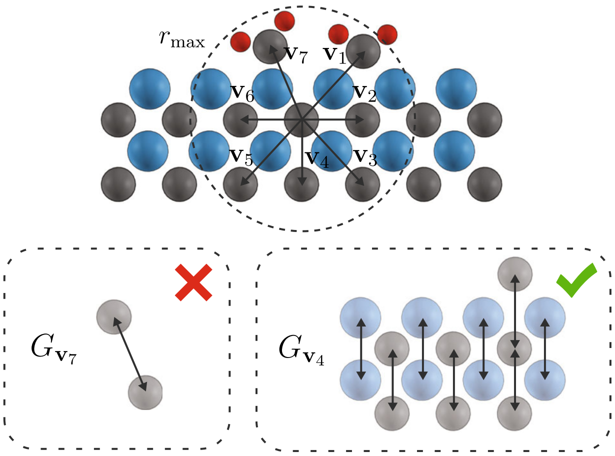
$$D = \begin{cases} n_{\text{pbc}} - \log_n(N_n), \text{when } n_{\text{pbc}} \neq 0 \\ 0, \text{when } n_{\text{pbc}} = 0 \end{cases}, \qquad (1)$$

where N_n is the number of clusters in a supercell that is repeated n times in each periodic direction and n_{pbc} is the number of periodic dimensions.

For the clustering, we use the density-based spatial clustering of applications with noise (DBSCAN)[31] data clustering algorithm. The advantage of this algorithm is that it does not require an initial guess for the number of clusters, and it can find arbitrarily shaped clusters. The clustering requires that we define a metric for the distance between the atoms. We use the following metric:

$$d_{ij} = \left|\mathbf{r}_i - \mathbf{r}_j\right|^{\mathrm{MIC}} - r_i^{\mathrm{cov}} - r_j^{\mathrm{cov}}, \qquad (2)$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the cartesian positions of atom i and j, respectively, and r_i^{cov} and r_j^{cov} are their covalent radii.[32] It is important to notice that in this metric the distances always follow the minimum image convention (MIC), i.e., the distance is calculated between the two closest periodic neighbours. By using the distance to the closest periodic neighbour, we obtain the correct clusters regardless of what shape of cell is used in the original simulation.



$$v_{\mathrm{a}} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6, \mathbf{v}_7\}$$
$$v_{\mathrm{b}} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6\}$$

Fig. 8 Illustration of finding the candidate set v_{a} of basis vectors and filtering this set into v_{b} based on the connectivity of atoms corresponding to these vectors. The neighbourhood of the seed atom is searched within a radius $r_{\max}$ for atoms of the same species as the seed atom. The vectors connecting the seed atom to such atoms form a candidate set v_{a} of basis vectors for the unit cell. For each of these vectors $\mathbf{v}$, a graph $G_{\mathbf{v}}$ is created. The set v_{b} is then formed by choosing vectors with enough connections in the graph as illustrated for the vectors $\mathbf{v}_4$ and $\mathbf{v}_7$

The clustering uses two parameters: the minimum cluster size $n_{\min}$ and the neighbourhood radius ϵ. We set $n_{\min}$ to 1 to allow clusters consisting of even single atoms and ϵ to 3.5 Å. At present, a system, in which there is more than one cluster in the original non-repeated system ($N_1 > 1$), is classified as unknown. Such a case corresponds to systems with multiple components that are spatially separated, such as a molecule far above a surface, low density gases, widely spaced clusters in vacuum, etc.

Determining the unit cell basis vectors

After we have detected a structure as being two-dimensional, we will try to find an underlying unit cell. The shape of this unit cell, its chemical elements and atomic positions identify a specific material and can be used to identify the bulk material that the structure originates from. By our definition, both surfaces and 2D materials should consist of multiple repetitions of this unit cell. In our classification scheme, surfaces have an unit cell that is repeated at least twice in three linearly independent directions. There are no limitations to the thickness of a surface. Similarly, 2D materials consists of a unit cell that is repeated at least twice in two linearly independent directions, but they have a controllable maximum thickness $h_{\max}^{\mathrm{2D}}$. According to this definition, there is no such thing as a single-layer surface, but even very thin structures with two-layers will be labelled as surfaces.

Traditional, symmetry-based unit cell reduction algorithms like the Niggli reduction[33] cannot be used to find a repeating primitive cell when the system contains symmetry breaking vacuum, vacancies or atoms from another structural component, such as an adsorbate. Here we introduce a more general cell identification algorithm for detecting a unit cell and the structure expanded by it in complex environments.

Cell identification starts by determining the basis vectors of the unit cell. The search for the basis vectors starts by selecting a seed atom that serves as a starting point for finding a unit cell. There are multiple approaches for selecting this seed atom, and depending on the expected complexity of the analyzed structures, one may need to use several spatially distributed seed points. A good initial guess is to use atoms that are close to the centre of mass of the system.

When a seed atom has been chosen, we collect all neighbouring atoms within a radius $r_{\max}$. From this set, we then choose only those atoms that belong to the same species as the seed atom. The displacement vectors from the seed atom to these atoms form the first possible set of unit cell basis vectors, v_{a}. This process is illustrated in Fig. 8. All of the basis vectors of the original simulation cell that are shorter than the defined maximum cell size $r_{\max}$ are automatically included as possible bases.

For each vector $\mathbf{v}$ in the set v_{a}, we form a graph $G_{\mathbf{v}}$ of atoms that are connected by the periodicity defined by the vector. The graph is formed by first including all the atoms within $r_{\max}$ from the seed atom as nodes in the network. Next for each atom in the graph, we look for an atom with the same element in the two directions ±**v**. If such an atom exists, it is added to the graph, if not already present, the two atoms are marked as being

Table 1. Summary of parameters used in the classification

Name	Description	Default
i_{seed}	Index of the seed atom	—
$r_{\max}$	Maximum cell basis vector length	—
ε	Clustering radius used in dimensionality detection	3.5 Å
Δr	Distance tolerance for finding atoms	—
α_{par}	Angle below which basis vectors are considered to be parallel	20°
r_{bond}	The distance parameter that is used to find bonded atoms	0.75 Å
$n_{\min}^{\mathrm{b}}$	Minimum number of edges in graph $G_{\mathbf{v}}$ for the vector $\mathbf{v}$ to be valid	$0.75n_{\mathrm{neigh}}$
$n_{\min}^{\mathrm{v}}$	Minimum number of nodes for subgraph to be valid	$0.5n_{\mathrm{seed}}$
$n_{\min}^{\mathrm{e}}$	Minimum number of edges for subgraph to be valid	$2(d+1)$
ΔC	Similarity threshold for detecting outliers	0.40
$h_{\max}^{\mathrm{2D}}$	Maximum thickness of a 2D material. Measured from the centre of the top-most atom to the centre of the bottom-most atom	5 Å
$l_{\max}^{\mathrm{2D}}$	Maximum allowed basis vector length for 2D materials with only one unit cell in the original simulation	5 Å
$n_{\max}^{\mathrm{outliers}}$	Maximum accepted number of outlier atoms	$0.5n_{\mathrm{atoms}}$

Variable d is the number of directions with repetitions of the cell (three for surface, two for 2D material), n_{seed} is the number of edges in the subgraph in which the seed atom is located, n_{neigh} is the number of atoms within the radius $r_{\max}$ from the seed atom and n_{atoms} is the number of atoms in the original system

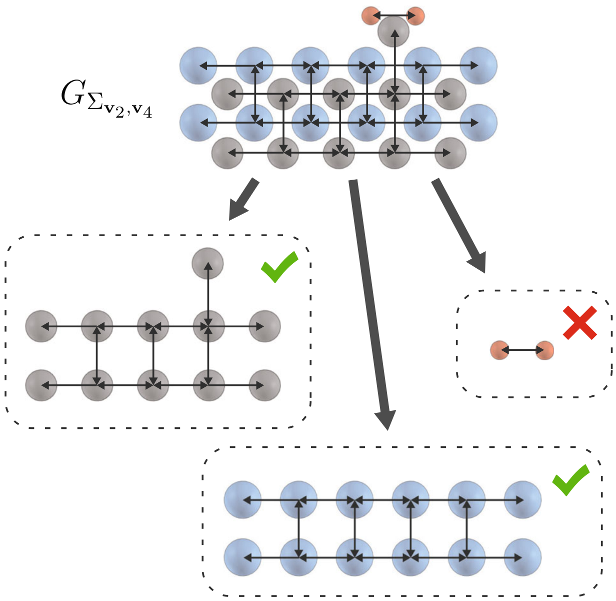


Fig. 9 Illustration of the connected components of the graph G_Σ corresponding to two vectors $\mathbf{v}_2$ and $\mathbf{v}_4$ shown in Fig. 8. The connected components will be filtered based on their size and connectivity. In this example, only two of the connected components are valid and thus the final unit cell will have two atoms. The positions of these two atoms will be based on an average gathered from the connected components. Notice that an outlier atom from an adsorbate is here included in a connected component, but this does not have a large impact on the final averaged positions

connected by an edge. The directionality of the edges are taken into account, meaning that an edge $E(i, j)$ is distinguished from $E(j, i)$. The search extends also beyond the periodic boundary conditions, and it is possible that an atom is connected to a periodic copy of itself in a neighbouring cell.

The possible basis vectors v_a are next filtered by checking the number of edges corresponding to the vector $\mathbf{v}$ in the graph $G_\mathbf{v}$. If the number of edges is less than the tunable parameter n^b_{min}, $\mathbf{v}$ is rejected, as there are not enough repetitions of atoms corresponding to that vector near the seed atom. The default value of n^b_{min} is given in Table 1. The remaining set of vectors is labelled as v_b. This filtering prevents the usage of candidate vectors that result only from the presence of defects or adsorbates, making the search robust against such outlier atoms. An illustration of the connections in a valid and an invalid graph $G_\mathbf{v}$ are shown in Fig. 8.

Next, we determine how many linearly independent vectors there are in the set v_b, that is we calculate the rank of the space spanned by this set. Exactly linearly dependent vectors could be identified by storing them as rows of a matrix and reducing this matrix to row echelon form. Because the atomic structures are rarely perfectly aligned, the vectors will in reality be parallel only up to some threshold. For this reason, we use angles to measure the linear dependence of vectors and specify an angle threshold α_{par} to identify the parallel ones. To do this, we calculate the following variables

$$\mathbf{n}_{ij} = \hat{\mathbf{v}}_i \times \hat{\mathbf{v}}_j \tag{3}$$

$$\alpha_{ijk} = \left|\hat{\mathbf{v}}_i \cdot \hat{\mathbf{n}}_{jk}\right| \tag{4}$$

Vector $\mathbf{n}_{ij}$ is a normal vector for the plane defined by vectors $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{v}}_j$, and its length is the sine of the angle between these two vectors. Variable α_{ijk} is equal to the sine of the angle between the vector $\hat{\mathbf{v}}_i$ and the plane defined by vectors $\hat{\mathbf{v}}_j$ and $\hat{\mathbf{v}}_k$. The values α_{ijk}, α_{jki} and α_{kij} are calculated for all combinations of vector triplets $\mathbf{v}_i$, $\mathbf{v}_j$ and $\mathbf{v}_k$ in the set v_b. First, we find all triplets for which α_{ijk}, α_{jki} and $\alpha_{kij} \geq \sin(\alpha_{par})$. If at least one such triplet is found, it represents a unit cell with repetitions in three directions, i.e., a surface. If no such triplet is found, we instead find all pairs $\mathbf{v}_i$, $\mathbf{v}_j$ for which $|\mathbf{n}_{ij}| \geq \sin(\alpha_{par})$. If at least one such pair is found, then it represents a unit cell with repetitions in two directions, i.e., a 2D material.

If more than one combination of valid vectors is found, a graph G_Σ is formed for each of them by combining the individual graphs $G_\mathbf{v}$ that correspond to vectors in the combination. The combinations are filtered so that only the ones which have a number of edges e in the graph G_Σ equal to the maximum found value of e are kept. This filtering is done to avoid choosing basis vector combinations that correspond to symmetric substructures inside the correct unit cell.

The remaining cells are all valid and equal in terms of the repetitions of the seed atom. From this remaining set, we then select cells with size similar to the smallest found size and then choose the combination with the highest orthogonality between the basis vectors. These properties are desirable when the cell is used to track the structure as explained in section Tracking. With small cells, the tracking can better adapt to local changes, and orthogonality makes it easier to find the atoms that belong to a cell. Cell volumes V or areas A are readily available from the already calculated quantities α_{ijk} and $\mathbf{n}_{ij}$

$$A = \left|\mathbf{v}_i \times \mathbf{v}_j\right| = |\mathbf{v}_i|\left|\mathbf{v}_j\right|\left|\mathbf{n}_{ij}\right| \tag{5}$$

$$V = \left|\mathbf{v}_i \cdot \left(\mathbf{v}_j \times \mathbf{v}_k\right)\right| = |\mathbf{v}_i|\left|\mathbf{v}_j\right||\mathbf{v}_k|\left|\mathbf{n}_{jk}\right|\alpha_{ijk} \tag{6}$$

The orthogonality between the vectors is measured by the sum of the squared cross-products between the normalized vectors. By using the identity $|\mathbf{a}\cdot\mathbf{b}|^2 = |\mathbf{a}|^2|\mathbf{b}|^2 - |\mathbf{a}\times\mathbf{b}|^2$, we can quantify the orthogonality as follows:

$$\left|\hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_j\right|^2 + \left|\hat{\mathbf{v}}_j \cdot \hat{\mathbf{v}}_k\right|^2 + \left|\hat{\mathbf{v}}_k \cdot \hat{\mathbf{v}}_i\right|^2 \tag{7}$$

$$= 3 - \left|\mathbf{n}_{ij}\right|^2 - \left|\mathbf{n}_{jk}\right|^2 - \left|\mathbf{n}_{ki}\right|^2 \tag{8}$$

Determining unit cell contents

The unit cell shape is determined by the unit cell vectors, but the atoms belonging to the unit cell and their positions still need to be determined. Not every atom within the radius r_{max} has to belong to the unit cell, as there might be adsorbates or defects within the neighbourhood. To correctly identify atoms within the cell, the graph G_Σ, corresponding to the best basis vectors, is divided into its connected components, i.e., subgraphs that are connected internally, but not to each other. Figure 9 illustrates the separation into these connected components. Atoms that are not part of the cell can now be filtered by discarding subgraphs that have too few edges or nodes. This makes the detection robust against any possible outlier atoms that are not part of the correct unit cell. Filtering can be done by discarding the subgraphs where $\langle \deg(v)\rangle \leq n^e_{min}$ and $n \leq n^v_{min}$, where $\langle \deg(v)\rangle$ is the average degree of the subgraph nodes that are within the radius r_{max} and n is the number of nodes in the subgraph. The default values of n^e_{min} and n^v_{min} can be seen in Table 1.

Each remaining valid connected component now represents a set of periodically repeated atoms corresponding to the same relative position within a repetition of the unit cell. For each atom in a valid connected component, a relative position in its respective unit cell repetition is calculated. These relative positions, which are in the interval [0, 1], are wrapped to the periodic repetition nearest to the origin of the cell and averaged to reach a robust estimate for a final relative atom position in the unit cell.

Cell validation

Sometimes the unit cell found in this way cannot represent a valid surface or 2D material because it is too sparse. This can happen, for example, when the cell contains many outliers. To detect these cases, we run the dimensionality detection routine as described in section Dimensionality detection on the found unit cell with a clustering threshold $\epsilon = r_{bond}$. The default value for r_{bond} is given in Table 1. If during this dimensionality detection, multiple clusters are detected for the original cell or the detected dimensionality is incorrect, the structure cannot be classified as a surface or a 2D material. In these cases, the structure is classified as a generic 2D structure.

The quality of cell identification depends heavily on the number of repetitions of the cell in the original system. If the algorithm finds that the best basis vectors correspond to the vectors of the original simulation cell, the correct classification cannot be guaranteed as we then only have information from one repetition. By default, these systems are classified as being generic 2D structures, but without the cell information. An exception to this can be made for 2D materials, in which it is quite typical to perform the simulation by using a single repetition of a primitive cell. To allow the identification of these systems, a maximum allowed basis vector length l^{2D}_{max} for 2D materials with only one unit cell in the original simulation can be specified. When a relatively small value for $l^{\mathrm{2D}}_{\mathrm{max}}$ is allowed, it is a reasonable assumption that the unit cell does not contain defects or adsorbates, and it can be accepted. The parameter $l^{\mathrm{2D}}_{\mathrm{max}}$ is not applicable to surfaces because by our definition, valid surfaces must have a repetition in three linearly independent directions, and thus systems with single repetition of a unit cell can never represent a surface.

Tracking

During unit cell detection, only the area within the radius r_{max} from the seed atom is taken into account. Usually the unit cell is also repeated beyond this radius. To correctly identify all atoms that are a part of the underlying structure, a more extensive tracking of unit cell repetitions is needed. This becomes especially important for detecting outlier atoms that do not match the atoms in any repetition of the unit cell.

Tracking works by identifying atoms belonging to periodic repetitions of the unit cell. This search is initiated by finding periodic copies of the seed atom near the positions $\mathbf{r}_{\mathrm{seed}} + \mathbf{v}$, where $\mathbf{r}_{\mathrm{seed}}$ is the seed atom position and $\mathbf{v}$ is any of the unit cell basis vectors or their corresponding opposite vectors. If such atom locations are found, they are used as new seed locations $\mathbf{r}'_{\mathrm{seed}}$, otherwise $\mathbf{r}'_{\mathrm{seed}} = \mathbf{r}_{\mathrm{seed}} + \mathbf{v}$. Using $\mathbf{r}'_{\mathrm{seed}}$ as a new cell origin, the atoms belonging to the neighbouring cell are then identified by searching for atoms in the known positions of the unit cell. In larger systems, the orientation and size of individual unit cells can be changed locally by external factors, such as adsorbates or pressure. To better adapt to these local distortions and reorientations of the lattice, the unit cell basis vectors $\mathbf{v}$ are updated as $\mathbf{v}\prime = \mathbf{r}'_{\mathrm{seed}} - \mathbf{r}_{\mathrm{seed}}$. If new atoms are identified as belonging to a neighbouring unit cell, the search is extended in a breadth-first manner, i.e., the search continues recursively by using the new origin $\mathbf{r}'_{\mathrm{seed}}$ and new basis vectors $\mathbf{v}'$, and the neighbouring cells closest to the original seed location are searched before advancing to the next layer of neighbouring cells. The branches of the search are stopped when no new atoms are identified within a neighbouring cell.

Outlier detection

Often 2D systems include atoms that cannot be attributed to the unit cell. Such atoms include adsorbates, vacancies, substitutions, interstitials and surface reconstructions. Being able to distinguish these outlier atoms is useful because it enables the search for structural combinations that go beyond bulk-terminated surfaces or ideal 2D materials.

During the tracking of the unit cells, the atoms that can be attributed to periodic repetitions of a unit cell are identified. However, in many cases, some of these atoms lie outside the extent of the surface and are in fact outliers that happen to be near a position where an atom is expected to be found based on the translational symmetry. This is also seen in Figs. 8 and 9, where one atom from the adsorbate happens to be near a position where a surface atom is expected. To better distinguish these outliers, we compare the chemical environments of all atoms in the unit cells to the ideal environment, as found in the detected unit cell. There are many ways to accurately quantify the local chemical environment near an atom, including the smooth overlap of atomic positions (SOAP)[34] and atom-centered symmetry functions (ACSF).[35] However, for the purpose of detecting outliers, only a rough estimate for the chemical environment is needed, and we apply a simpler and faster-to-compute quantity that is based on counting the occurrence of atomic elements within a finite radius. The chemical environment of atom a thus consists of a vector.

$$\mathbf{C}_a = [n_1, n_2, \dots] \tag{9}$$

where n_i is the number of atoms with atomic number i within a distance $r_a^{\mathrm{cov}} + r_i^{\mathrm{cov}} + r_{\mathrm{bond}}$. We only consider atomic numbers that are found in the unit cell. Here r_a^{cov} and r_i^{cov} are covalent radii and r_{bond} is a tunable parameter. The similarity between a real chemical environment $\mathbf{C}_b$ and the ideal chemical environment $\mathbf{C}_a$ as found in the unit cell is then measured as:

$$T(\mathbf{C}_a, \mathbf{C}_b) = \sum_i \min\left(n_i^a, n_i^b\right) / \sum_i n_i^a \tag{10}$$

This measure represents the ratio of common elements in the neighbourhoods of $\mathbf{C}_a$ and $\mathbf{C}_b$ to the number of neighbouring elements in the ideal environment $\mathbf{C}_a$. This form is similar in style to the Tanimoto similarity measure,[36] but is modified so that the comparison is always done to the ideal environment instead of the combined environments $\mathbf{C}_a$ and $\mathbf{C}_b$. If this similarity is above a controllable parameter ΔC, the atom is assigned to a unit cell. Otherwise the atom is labelled as an outlier. The default value for ΔC is given in Table 1.

After this analysis, the outlier atoms are flagged as atoms that were not matched to any repetition of the unit cell or were rejected from a unit cell by the chemical similarity check. If the number of outliers is bigger than $n^{\mathrm{outliers}}_{\mathrm{max}}$, which defaults to 50% of the total atoms, the found unit cell is not accepted. This is done to avoid using cells that are a part of a larger heterostructure or which have an exceptionally large fraction of outliers.

Parameters and generalizability

Several parameters control the classification and allow the user to customize the routine to different datasets. In Table 1, all the controllable parameters are introduced together with an explanation and the default values.

The given default parameters are a good starting point for most systems, but can be adjusted to specific environments. The optimal value for parameters Δr, r_{max} and i_{seed} are highly dependent on the system, and producing good estimates for them can be hard. One approach is to find optimal parameter values for each structure separately by trying out multiple parameter values. The optimal parameters can be selected by specifying a list of possible values for each parameter, performing the classification with all possible combinations of these parameters and then keeping the result that gives the least amount of outlier atoms. This parameter cross-validation requires that the position tolerance is kept below the minimum distance between the two atoms in the structure so that thr outliers are properly detected.

DATA AVAILABILITY

An overview of the classification results together with the original geometries are provided in Supplementary Note 1. In Supplementary Note 2, we provide descriptions for the different fields in the classification overview, instructions on accessing the classification code and a guide for accessing the original calculations in the NOMAD Archive by using the unique NOMAD Archive identifiers and the REpresentational State Transfer (REST) interface provided by the NOMAD Archive.[37] The source code and installation instructions for the classification tools can be found in https://github.com/SINGROUP/matid.

ACKNOWLEDGEMENTS

We thank Amber Geurts for insightful discussions on database-driven materials science. Computing resources from the Aalto Science-IT project are gratefully acknowledged. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Centre of Excellence and from the Jenny and Antti Wihuri Foundation. This work was furthermore supported by the Academy of Finland through its Centres of Excellence Programme 2015–2017 under the project number 284621, as well as its Key Project Funding scheme under project number 305632 and by the World Premier International Research Center Initiative (WPI), MEXT, Japan.

AUTHOR CONTRIBUTIONS

L.H. developed the model and performed all the calculations. P.R. and A.F. supervised the project. All authors reviewed and commented on the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-018-0107-6).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Materials Genome Initiative. https://www.mgi.gov/.
2. The Novel Materials Discovery (NOMAD) Laboratory. https://nomad-coe.eu/.
3. Materials' Revolution: Computational Design and Discovery of Novel Materials. http://nccr-marvel.ch/en/project.

4. NOMAD Archive. https://metainfo.nomad-coe.eu/nomadmetainfo_public/archive.html.
5. Kirklin, S. et al. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
6. Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
7. Curtarolo, S. et al. Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
8. Materials Cloud. https://www.materialscloud.org/.
9. Open Materials Database. http://openmaterialsdb.se/.
10. Theoretical Crystallography Open Database. http://www.crystallography.net/tcod/.
11. The Electronic Structure Project. http://gurka.fysik.uu.se/ESP/.
12. NIMS Materials Database (MatNavi). http://mits.nims.go.jp/index_en.html.
13. NREL MatDB. https://materials.nrel.gov/.
14. Computational Materials Repository. https://cmr.fysik.dtu.dk.
15. Kelchner, C. L., Plimpton, S. J. & Hamilton, J. C. Dislocation nucleation and defect structure during surface indentation. *Phys. Rev. B* **58**, 11085–11088 (1998).
16. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28**, 784–805 (1983).
17. Honeycutt, J. D. & Andersen, H. C. Molecular dynamics study of melting and freezing of small lennard-jones clusters. *J. Phys. Chem.* **91**, 4950–4963 (1987).
18. Ackland, G. J. & Jones, A. P. Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B* **73**, 054104 (2006).
19. Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Model. Simul. Mater. Sci. Eng.* **20**, 045021 (2012).
20. Stukowski, A. & Albe, K. Extracting dislocations and non-dislocation crystal defects from atomistic simulation data. *Model. Simul. Mater. Sci. Eng.* **18**, 085001 (2010).
21. Ashton, M., Paul, J., Sinnott, S. B. & Hennig, R. G. Topology-scaling identification of layered solids and stable exfoliated 2d materials. *Phys. Rev. Lett.* **118**, 106101 (2017).
22. Larsen, A. H. et al. The atomic simulation environment – a python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002 (2017).
23. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
24. Apache License Version 2.0. https://www.apache.org/licenses/LICENSE-2.0.
25. Jenke, J. et al. Chemistry informed structure map for measuring the similarity between atomic environments (2018). [Unpublished manuscript].
26. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
27. Gulans, A. et al. exciting: a full-potential all-electron package implementing density-functional theory and many-body perturbation theory. *J. Phys. Condens. Matter* **26**, 363202 (2014).
28. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175 (2009).
29. Levchenko, S. V. et al. Hybrid functionals for large periodic systems in an all-electron, numeric atom-centered basis framework. *Comput. Phys. Commun.* **192**, 60–69 (2015).
30. Ren, X. et al. Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2, and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **14**, 053020 (2012).
31. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).
32. Cordero, B. et al. Covalent radii revisited. *Dalton. Trans*. **21**, 2832–2838 (2008).
33. Santoro, A. & Mighell, A. D. Determination of reduced cells. *Acta Cryst*. **26**, 124–127 (1970).
34. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
35. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
36. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115–1118 (1960).
37. Luca, M. G. et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj Comput. Mater.* **3**, 46 (2017).