Mapping materials and molecules

Bingqing Cheng,[*,†] Ryan-Rhys Griffiths,[‡] Simon Wengert,[¶] Christian Kunkel,[¶] Tamas Stenczel,[§] Bonan Zhu,[‖] Volker L. Deringer,[⊥] Noam Bernstein,[#] Johannes T. Margraf,[¶] Karsten Reuter,[¶] and Gabor Csanyi[§]

†*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom*

‡*Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom*

¶*Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany*

§*Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

‖*Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB3 0FS, United Kingdom*

⊥*Department of Chemistry, Inorganic Chemistry Laboratory, University of Oxford, Oxford OX1 3QR, United Kingdom*

#*Center for Materials Physics and Technology, U.S. Naval Research Laboratory,Washington, DC 20375, United States*

E-mail: bc509@cam.ac.uk

Conspectus

The visualization of data is indispensable in scientific research, from the early stages when human insight forms, to the final step of communicating results. In computational physics,

chemistry and materials science, it can be as simple as making a scatter plot, or as straightforward as looking through the snapshots of atomic positions manually. However, as a result of the "big data" revolution these conventional approaches are often inadequate. The widespread adoption of high-throughput computation for materials discovery and the associated community-wide repositories have given rise to data sets that contain an enormous number of compounds and atomic configurations. A typical data set contains thousands to millions of atomic structures, along with a diverse range of properties such as formation energies, band gaps, or bio-activities.

It would thus be desirable to have a data-driven and automated framework for visualizing and analyzing such structural datasets. The key idea is to construct a low-dimensional representation of the data, which facilitates navigation, reveals underlying patterns, and helps to identify data points with unusual attributes. Such data-intensive maps, often employing machine learning methods, are appearing more and more frequently in the literature. However, to the wider community, it is not always transparent how these maps are made and how they should be interpreted. Furthermore, while these maps undoubtedly serve a decorative purpose in academic publications, it is not always apparent what extra information can be garnered from reading or making them.

This Account attempts to answer such questions. We start with a concise summary of the theory of representing chemical environments, followed by the introduction of a simple yet practical conceptual approach for generating structure maps in a generic and automated manner. Such analysis and mapping is made nearly effortless by employing the newly developed software tool, ASAP. To showcase the applicability to a wide variety of systems in chemistry and materials science, we provide several illustrative examples, including crystalline and amorphous materials, interfaces, and organic molecules. In these examples, the maps not only help to sift through large datasets, but also reveal hidden patterns that could be easily missed using conventional analyses.

The explosion in the amount of computed information in chemistry and materials science

has made visualization into a science in itself. Not only have we benefited from exploiting these visualization methods in previous works, we also believe that the automated mapping of datasets will in turn stimulate further creativity and exploration, as well as ultimately feed back into future advances in the respective fields.

Key References

- Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. Machine Learning Unifes the Modeling of Materials and Molecules. Science Advances **2017**, 3, e1701816.[1]

  Showcases SOAP and Gaussian process regression for machine learning in a variety of material and molecular prediction tasks.

- Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. Physical Chemistry Chemical Physics **2020**, 22, 12697-12705.[2]

  Uses the automatic maps for crystal structure predictions.

- Stuke, A.; Kunkel, C.; Golze, D.; Todorovic, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. Sci. Data **2020**, 7, 58.[3]

  Example of emerging role of 'Big Data' in molecular chemistry.

Introduction

We are experiencing a dramatic growth of data in chemistry, physics and materials science, thanks to the ever-increasing computational power available, advances in electronic structure methods and algorithms, and community-wide data repositories. Exploiting the "big data" efficiently and effectively using traditional tools is not easy: datasets often contain

thousands to millions of atomistic structures, along with diverse properties. Consequently, machine learning (ML) methods are increasingly employed to handle the large and complex datasets[1,4–8]. Often, data visualization is an initial and a final step of these data-driven studies. A low-dimensional map shows a condensed view of the dataset and reveals underlying patterns, such as clusters, outliers and correlations, allowing researchers to gain first insights from visual inspections[9,10]. During the final stage, visualization is essential and efficient in communicating results.

However, most of these papers focus on data generation or ML predictions, while displaying visualizations without much interpretation or explanation. This is somewhat unsatisfactory, as many chemical representations and embedding methods for generating these maps are available. Furthermore, it may be unclear in what ways the maps are helpful and what kind of physical insights they provide. To fill in this gap, this Account summarises the underlying principles of the visualization and showcases its applicability to a wide variety of physical systems.

To largely automate the mapping task, we have developed user-friendly software packages: the Automatic Selection And Prediction tools for materials and molecules (ASAP) is a Python-based command-line tool that enables automatic analysis and mapping using just a couple of simple commands and options. We display such commands in snippets below when showing figures generated using the ASAP. To explore a dataset interactively, we rely on a web-browser based viewing tool that can display 3D-structures corresponding to each data point together with its attributes.

Essential concepts and methods for mapping atomic structure

Low-dimensional embeddings

The geometrical configuration of a molecule or material is intrinsically high dimensional, $3n$ for n atoms. To visualize the relationship between the structures in a dataset, we need to represent each structure as a point in a low-dimensional space, typically the two dimensions of paper or a computer screen. This high ($3n$) to low dimensional transformation is called *dimensionality reduction* or *embedding*. Such embedding is common and crucial for analyzing simulation results or structural databases. Traditionally, it usually requires human insights for selecting appropriate low-dimensional coordinates, often referred to as collective variables (CVs). A textbook embedding example is the Ramachandran plot that visualizes energetically favorable regions for backbone dihedral angles (Φ and Ψ) of amino acid residues in a protein structure. The plot in Figure 1(a) illustrates an alanine dipeptide molecule with 66 geometric degrees of freedom using just two torsion angles. Most configurations concentrate in three distinct clusters, associated with common secondary structure elements (the α-helix, β-sheet, and left-handed α-helix).

The Ramachandran and similar plots provide powerful insight into high-dimensional structural data, but they typically require domain knowledge to hand-craft the CVs for every specific system. In contrast, automatic and system-agnostic embeddings for atomistic structures do not rely on system-specific information. In general, embedding procedures preserve some relationships between the points in high and low dimensional space. Loosely speaking, points that are "close" to each other in high dimension should remain so on the low-dimensional map. Embedding methods differ in the definition of "closeness", whether calculated for all points or just a subset, and in the numerical algorithms employed. A particularly simple method is *principal component analysis* (PCA), which defines closeness as the scalar product between the vectors pointing to the points[12]. Consequently, the axes

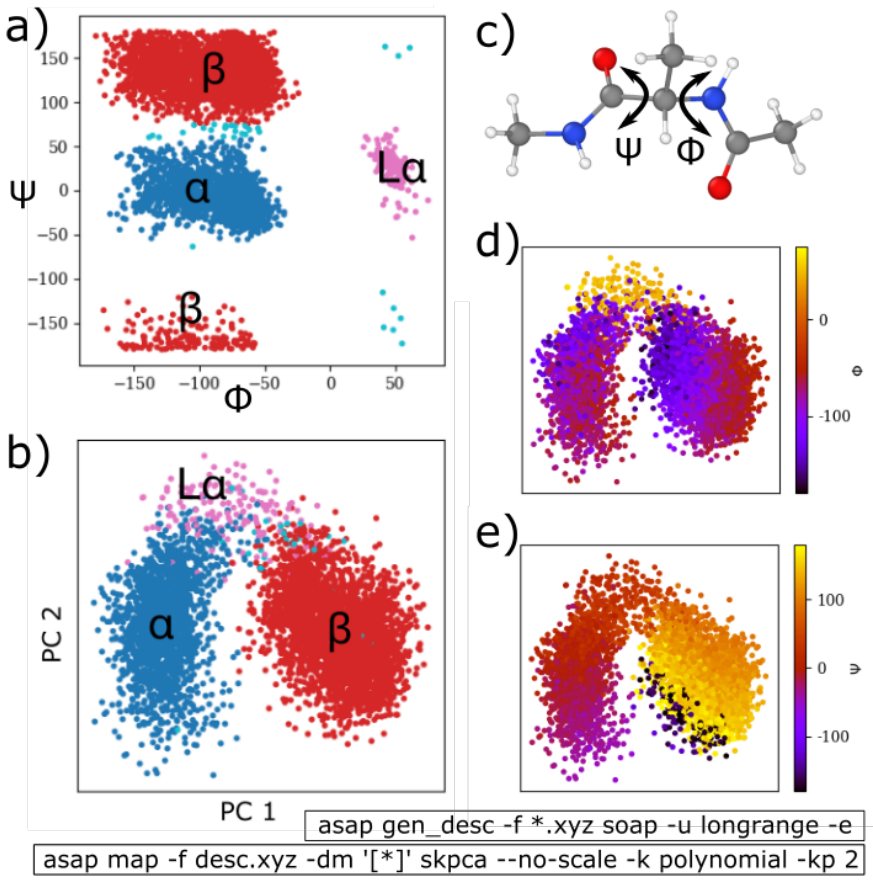


Figure 1: Panel (a) shows a Ramachandran plot of 5K configurations of an alanine dipeptide selected from a molecular dynamics trajectory[11], with respect to the two dihedral angles indicated in panel (c). The snapshots are classified according to the canonical structural motifs. Panel (b,d,e) show the KPCA projections using the SOAP descriptors, colored according to classifications, Φ, and Ψ, respectively.

of the low-dimensional map are just the first few eigenvectors of the *design matrix*, formed by concatenating the high dimensional coordinates. Alternatively, if the closeness is defined using pairwise Euclidean distances, the method is called multi-dimensional scaling[13]. Other definitions of closeness yield t-distributed stochastic neighbor embedding (t-SNE)[14], sketch-map[15], the uniform manifold approximation and projection (UMAP)[16], etc.

Therefore, the critical first step in designing a successful embedding method for materials and molecules is to decide how to compare atomic structures, i.e. by defining a distance metric. Several methods have been proposed over the past decade to describe structures, primarily for predicting atomic scale properties using machine learning[17–24]. They all respect

the appropriate physical symmetries; many are based on atomic densities, and these are essentially equivalent in some limit, differing only in the basis onto which the density is projected[25]. Here we focus on the Smooth Overlap of Atomic Positions (SOAP) descriptor[20], coupled with *kernel* PCA (KPCA)[26], which defines a scalar product in high dimensions with respect to a metric, as given by a user-supplied *kernel function*.

The computational cost of the whole process of constructing a map, computing descriptors and then using PCA or a sparse version of KPCA as implemented in ASAP, scales linearly with the total number of atoms in the dataset. The workflow usually takes only a few seconds on laptops for moderately sized datasets, and less than a few minutes even for the largest set considered in this Account. To make the method suitable for even larger sets, the ASAP code is made parallelizable, and contains tools to sparsify datasets (i.e. select a representative subset) as well.

Returning to the first example, an automatic mapping of the alanine dipeptide configurations using this method is shown in Figure 1(b). Similarly to the standard Ramachandran plot in Figure 1(a), the structures with different motifs are clearly separated on the KPCA map. (Note that in PCA or KPCA the first few eigenvectors of the design matrix, which form the axes of the plot, are also called "principal components", PCs.) Panels (d) and (e) show the same KPCA projection, but with the points colored according to Φ and Ψ. The strong horizontal color gradient in panel (e) suggests that the PC1 is essentially equivalent to Ψ, with the additional advantage that the β cluster does not split. The vertical (PC2) axis is well correlated with the Φ angle at the top of the plot where the Lα cluster is separated from the others. As such the KPCA map provides the same or even improved view compared with the conventional Ramachandran plot, but without relying on the prior domain knowledge.

Describing and comparing atomic environments

The automatic comparison and mapping of materials and molecules starts with describing each *atomic environment* $\mathcal{X}$, which consists of the atoms (chemical species and position)

within a sphere of radius r_{cut} centered at a specific atom. A good descriptor of $\mathcal{X}$ should be invariant to translation, rotation, and permutation of atoms of the same species, because these operations do not change physical properties. Many traditional descriptors used in cheminformatics are based on the covalent connectivity of atoms, such as simple valence counting and common neighbor analysis[27], the presence or absence of predefined atomic fragments (e.g. the Morgan fingerprints[28]) or orientational order parameters[29]. These are relatively low dimensional descriptors, and lose much geometric information. We opt to retain all geometric information when representing atomic environments and structures, and then rely on the dimensionality reduction of the embedding to arrive at a low-dimensional map.

To construct SOAP descriptors, first consider an atomic environment $\mathcal{X}$ that contains only one atomic species, and place a Gaussian function of width σ centered on each atom i in $\mathcal{X}$ to make an atomic density function:

$$\rho_{\mathcal{X}_i}(\mathbf{r}) = \sum_{i \in \mathcal{X}} \exp\left[-\frac{|\mathbf{r} - \mathbf{r}_i|^2}{2\sigma^2}\right] f_{\text{cut}}(|\mathbf{r}|), \tag{1}$$

where $\mathbf{r}$ denotes a point in Cartesian space, $\mathbf{r}_i$ is the position of atom i relative to the central atom of $\mathcal{X}$, and the cutoff function f_{cut} smoothly decays to zero beyond the radius r_{cut}. This density representation ensures invariance with respect to translations and permutations of atoms of the same species, but not rotations. To obtain a rotationally-invariant descriptor, we expand the density in a basis of spherical harmonics, $Y_{lm}(\hat{\mathbf{r}})$, and a set of orthogonal radial functions, $g_n(r)$, as

$$\rho_{\mathcal{X}}(\mathbf{r}) = \sum_{nlm} c_{nlm} g_n(|\mathbf{r}|) Y_{lm}(\hat{\mathbf{r}}), \tag{2}$$

and construct the *power spectrum* of the density using the expansion coefficients,

$$\psi_{nn'l}(\mathcal{X}) = \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^* c_{n'lm}. \tag{3}$$

Then we obtain a vector of descriptors $\boldsymbol{\psi} = \{\psi_{nn'l}\}$ by considering all components $l \leq l_{\max}$ and $n, n' \leq n_{\max}$, which act as band limits, controlling the spatial resolution with which the atomic density is resolved. The generalization to more than one chemical species is straightforward[4]: we construct separate densities for each of n_{sp} species α, and compute power spectra $\psi^{\alpha\alpha'}_{nn'l}(\mathcal{X})$ for each pair of elements α and α', where the two species indices correspond to the c^* and c coefficients, respectively. The $n_{\text{sp}}(n_{\text{sp}}+1)/2$ vectors corresponding to each of the α-α' pairs are then concatenated to obtain the descriptor vector of the complete environment. In some cases, we might choose to neglect the cross terms ($\alpha \neq \alpha'$) and obtain a much shorter descriptor vector. The ASAP tool uses the DScribe python library to compute the SOAP descriptors[30].

Subsequent dimensionality reduction needs a distance metric to compare atomic environments or, equivalently, a positive semi-definite similarity kernel K (the latter should take its maximum value for a pair of identical environments and be smaller but positive for different environments). A natural similarity kernel between atomic densities is the overlap integrated over all 3D rotations, and it turns out that computing it is easy once we have the SOAP vectors[20],

$$K(\mathcal{X}, \mathcal{X}') = \int_{\hat{R}\in \text{SO}(3)} d\hat{R} \left| \int dr \rho_{\mathcal{X}}(\mathbf{r}) \rho_{\mathcal{X}'}(\hat{R}\mathbf{r}) \right|^2 = \boldsymbol{\psi}^T \boldsymbol{\psi}. \tag{4}$$

When considering a large number of atomic environments, we collect their descriptor vectors into a *design matrix*, Ψ, whose rows are the descriptor vectors $\boldsymbol{\psi}$. For N environments, each described by a descriptor vector of length D, the design matrix has size $N \times D$. From the design matrix, we can form the *kernel matrix* of size $N \times N$, whose elements are given by the similarity kernel between each environment. The simplest linear kernel is $\mathbf{K} = \Psi\Psi^T$, for which PCA and KPCA are equivalent; other options are available[12]. A common choice together with the SOAP representation is to raise the above kernel elements to a small integer power, giving rise to a polynomial kernel. If ones needs an explicit distance between two environments, it can be defined by $d(\mathcal{X}, \mathcal{X}') = \sqrt{(\boldsymbol{\psi} - \boldsymbol{\psi}')^2} = \sqrt{K(\mathcal{X}, \mathcal{X}) + K(\mathcal{X}', \mathcal{X}') - 2K(\mathcal{X}, \mathcal{X}')}$. Notice that for nonlinear kernels, one can thus define

the distance using just the kernel, bypassing explicit descriptors entirely.

Universal SOAP hyper-parameters

The length-scale hyper-parameters (r_{cut} and σ) for constructing the SOAP vectors can be fine-tuned for any given application[31]. While to date this was done case by case, we have now formulated general heuristics for choosing the SOAP hyper-parameters for a system with arbitrary chemical composition. The radial resolution is related to σ and $r_{\mathrm{cut}}/n_{\mathrm{max}}$, and the angular resolution is determined by $2\pi/l_{\mathrm{max}}$ as well as σ/r at each shell of radius r. As such, using a set of fixed hyper-parameters is inefficient, because different systems have distinct length scales and varying spatial complexity. Furthermore, a system with many different chemical elements can contain a wide range of length scales, and so using multiple sets of SOAP descriptors with different hyper-parameters can be advantageous[1,32].

Our universal heuristics are based on the characteristic bond lengths in the system, which in turn depend on the chemical species involved. For each atomic species Z, we calculate six structures (dimer, graphite, diamond, β-Sn, body-centered cubic, and face-centered cubic) spanning coordination from 1 to 12, minimizing the total energy with respect to uniform isotropic strain of each structure. The bond length in the lowest energy structure is defined as r^Z_{typ}, and the shortest bond length of any local minimum structure is r^Z_{min}. We then use these species-specific bond lengths to choose the SOAP hyper-parameters for a given system with a set of species. The specific rules for doing this and the resulting length scales for two examples are included in the Supplementary Information. In the ASAP tool, the usage of these hyper-parameters is simply activated by the "–universal" or "-u" flag.

Comparing molecules and crystal structures

So far we have described how to represent atomic environments. Frequently, however, we would like to represent, compare, and map *entire structures*. This requires descriptors for whole structures instead of environments. To do this, for structure A, one can combine all

the descriptors for the environments $\mathcal{X}_i$ of all N_A atoms, and the most straightforward way is to simply take the average,

$$\Phi(A) = \frac{1}{N_A} \sum_{i \in A}^{N_A} \boldsymbol{\psi}(\mathcal{X}_i). \tag{5}$$

Alternative constructions that lose less information are described elsewhere[4,33]. In the presence of multiple chemical species, one can apply a single sum, or first average separately for each species and then concatenate the species-specific averaged vectors. From the descriptor vector for each structure, one can then construct the design matrix and the kernel matrix, analogously to the procedure for environments.

Examples

Amorphous carbon

Here we show an example application on tetrahedral amorphous carbon (*ta*-C) films, which have intricate local environments[34–36]. The KPCA maps in Figure 2 (b-d) show 2D projections based on the atomic SOAP descriptors of local environments in *ta*-C (illustrated in Figure 2 (a)). Carbon atoms with different coordination numbers are automatically separated into clusters on the maps, reminiscent of the traditional classification of carbon environments as "sp", "sp^2", and "sp^3". In addition, the KPCA maps show continuous distributions of different environments within the sp and sp^2 clusters: there is significant variability in bond lengths that is strongly correlated to the vertical axis. The implication of such variability is discussed in-depth by Caro et al. in terms of reactivity (hydrogenation energy) and the classification of carbon bonds[5]. KPCA does not further separate the points within the coordination clusters as shown by the single density peak of each cluster in Figure 2 (d), which suggests there is no clear-cut way to sub-divide the sp and sp^2 clusters.

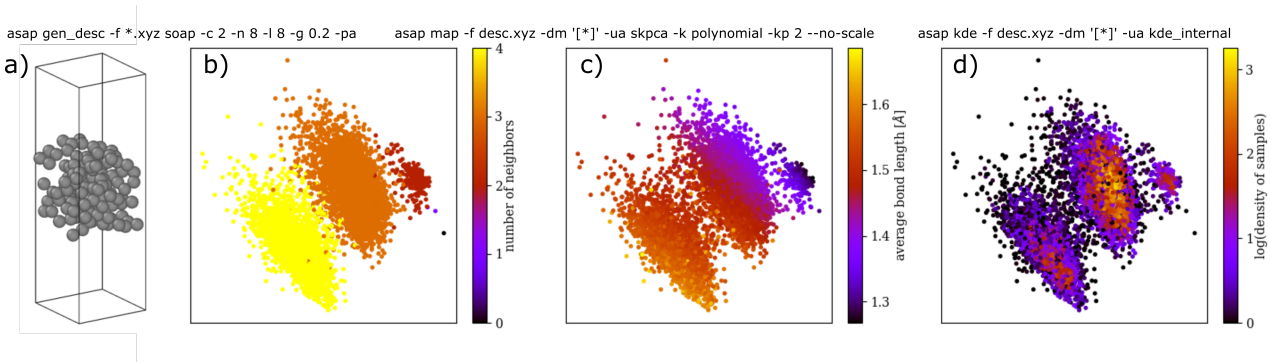


Figure 2: Panel (a) shows a snapshot of an amorphous carbon thin film[36]. Panels (b-d) show the KPCA projections of the atomic environments from 50 snapshots of the system with 125 carbon atoms[36], colored according to coordination number (b), average bond length (c), and the logarithm of the relative probability of each atomic environment (d). The rightmost point is absent in (c), as the corresponding atom has no neighbors.

The nucleation of a crystal from the liquid state

We now show the use of the automatic mapping in understanding the structural heterogeneity of nucleation. Solidification of materials starts with a small crystal nucleating from the melt. Despite a multitude of atomistic simulation studies, it is still a matter of debate whether body-centered cubic (bcc) ordering exists at the surface of the nuclei of face-centered cubic (fcc) crystals[37]. This controversy arises because the physical definition of bcc ordering is somewhat ambiguous, and also because the commonly used local bond order parameters[38] do not distinguish between bcc and interface atoms.

In Figure 3 we show the PCA map based on SOAP descriptors of each atom-centered environment inside a Lennard-Jones system consisting of a solid nucleus surrounded by undercooled liquid. Environments are colored according to how similar they are to fcc using a conventional fcc order parameter that was used for enhanced sampling[39,40]. Figure 3 reveals a smooth and gradual transition between the center of the nucleus and the bulk liquid, with two blobs of data points corresponding to the fcc and liquid-like motifs. There is no clear indication of an extra density peak that is associated with the bcc local ordering. Furthermore, the reference bcc environments are clearly separated on the map. The embedding thus severely questions the existence of bcc ordering at the surface of the forming nuclei.

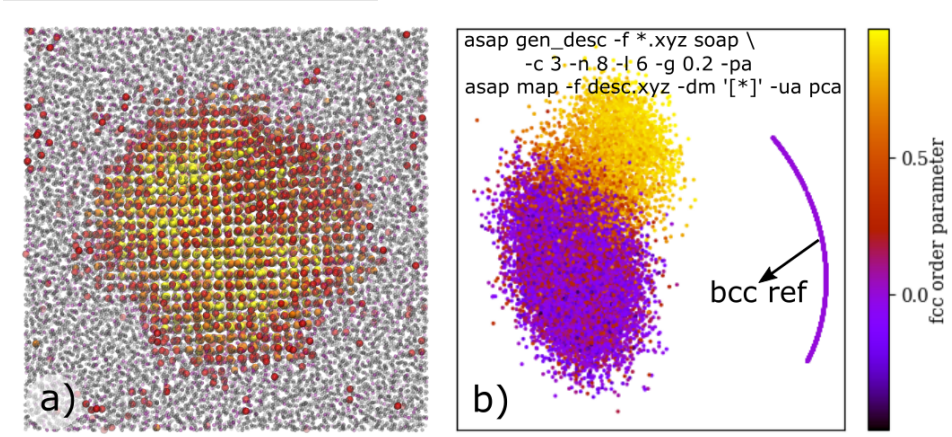


Figure 3: Snapshot of a Lennard-Jones system of 23,328 atoms containing a solid nucleus surrounded by undercooled liquid. Atoms are colored according to the similarity of their environment to fcc[39,40] (yellow means very similar, black and purple mean dissimilar). (a) Real space view, (b) PCA map for the atomic environments. In addition, we show the location of the bcc atomic environments (from perfect bcc crystals at a range of molar volumes) on the PCA map.

Liquid water structure

The compositions of training sets are crucial for the quality of machine learning models for chemistry and materials. Mapping atomic structures is useful for examining and understanding the training configurations, particularly when curating or expanding an existing data set. One such task is the fitting of machine learning interatomic potentials, which are increasingly popular as they can be both accurate and efficient[41].

Here we visualize the training set of a recent potential for bulk liquid water[42]. First 1,000 structures of liquid water were harvested from classical molecular dynamics simulations at 1000 K and densities between 0.7 and 1.2 g/mL, and augmented with lower energy configurations obtained after a few steps of geometry optimization. The remaining configurations were extracted from path-integral molecular dynamics (PIMD) simulations at ambient pressure and 300 K, which account for the quantum mechanical nature of hydrogen nuclei. The difference between classical and quantum mechanical water is not apparent from inspecting atomic snapshots by eye, cannot be captured using conventional metrics such as

oxygen radial distribution functions[42], and has only subtle manifestation in hydrogen bond analysis[43].

However, in the KPCA maps of the training set (Figure 4) the distinction is obvious: the classical and quantum water form two well-separated clusters. It is further revealed that the classical water configurations have a relatively wide spread in both energy and molar volume, and both quantities are correlated with the axes of the plot. Such spread in the training set is important for constructing a potential that is stable at a range of pressures and elevated temperatures.

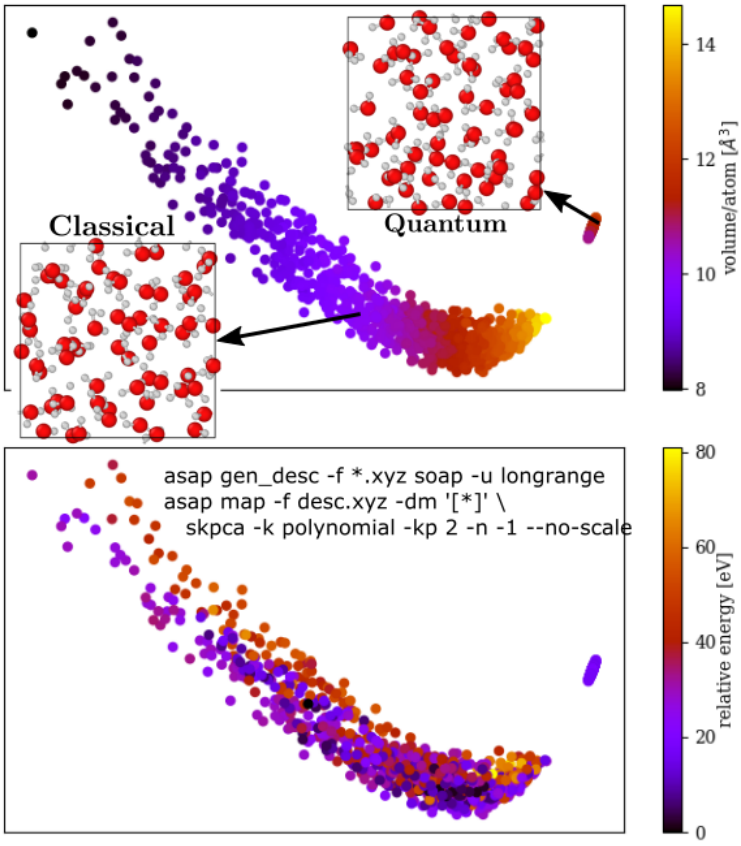


Figure 4: KPCA maps of liquid water configurations (1,000 classical and 593 quantum mechanical structures) from a training set[42], colored according to volume (upper panel) and the relative energy of each configuration (lower panel).

Crystal structure search: titanium dioxide

Ab initio random structure search[44] is a very productive tool of materials discovery. To demonstrate the use of visualization in this domain, we show an example of mapping the TiO_2 crystalline polymorphs[2] that were produced from random searches[44]. This dataset includes thousands of distinct TiO_2 structures with different atomic coordinates, cell shapes and numbers of formula units in the cell. Even though the knowledge of space groups, molar volumes and energies of the structures provides hints on how to classify them, it is still a formidable task to sort through them manually. The KPCA map in Figure 5 instead directly gives an overview of the structural similarities between 4,690 locally stable structures of titanium dioxide. Properties such as the relative enthalpy or unit cell volume vary smoothly across the figure, and regions of high density or stability are revealed. We project the known (marked in blue) and newly discovered phases (marked in green) of TiO_2 on the map[2], so one can immediately spot if a particular phase has been found in the random search, instead of having to rely on the traditional identifications such as the space groups. Indeed, as also shown on the map, different structures can adopt an identical space group, while atomic configurations that are structurally similar were classified to have distinct symmetries.

Structure of heterogeneous interfaces

Structure searches can be extended to systems with interfaces, to reveal the stable configurations that are hard to obtain otherwise[45]. The data analysis for this is even more challenging compared with bulk phases, because the presence of the interface breaks the crystallographic symmetry, so the traditional space group analysis is often ineffective. The extended, often low-symmetry nature of interfaces also makes visual inspections more difficult. Hence, automatic maps become extremely desirable in this case.

Figure 6 shows the PCA map of $SrTiO_3$ and CeO_2 (STO/CeO_2) (100)/(110) interface structures. Each point represents a configuration at a local energy minimum. The relative energies, used as the color scale, strongly correlate with the horizontal axis of the map.

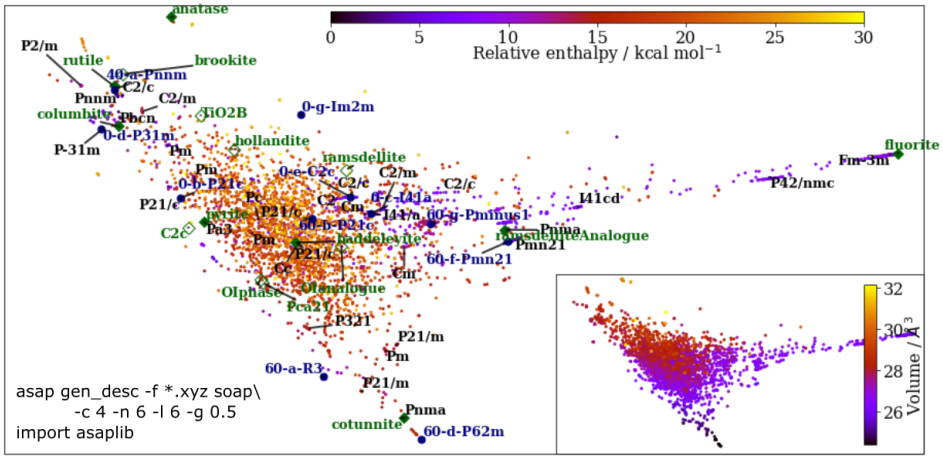


Figure 5: A KPCA map for TiO_2 structures generated from random structure searches at 20 GPa[2]: each dot indicates a crystal structure, with the known and new phases found in Ref.[2] shown using blue or green markers, respectively, and annotated by their names. If a certain phase is found in the search, it is marked as a solid symbol, and otherwise a hollow symbol (e.g. $C2c$, TiO_2B and ramsdellite). Only the space groups of structures that have low energy and have appeared multiple times are indicated. The plot is generated using a Python notebook, by importing ASAP as a library.

This means that while the interfacial energies are not used to construct the map, PCA identifies them automatically, presumably just from the distortion of the interface regions. We identified two clusters with low energies: group A consists of structures similar to the ideal interface that forms by simply joining the bulk phases, while group B contains the reconstructed structures.

Organic molecules

The QM9 data set[46], which contains 133,885 organic molecules composed of H and up to nine heavy atoms (C,N,O and F), has become a standard benchmark for ML-based property prediction. Here we compare molecular structures using average SOAP descriptors (Eqn. (5)), and then use a sparse version of KPCA for dimensionality reduction as the dataset is large. We use the resulting map (Figure 7) to navigate the QM9 set, and exploit the interactive

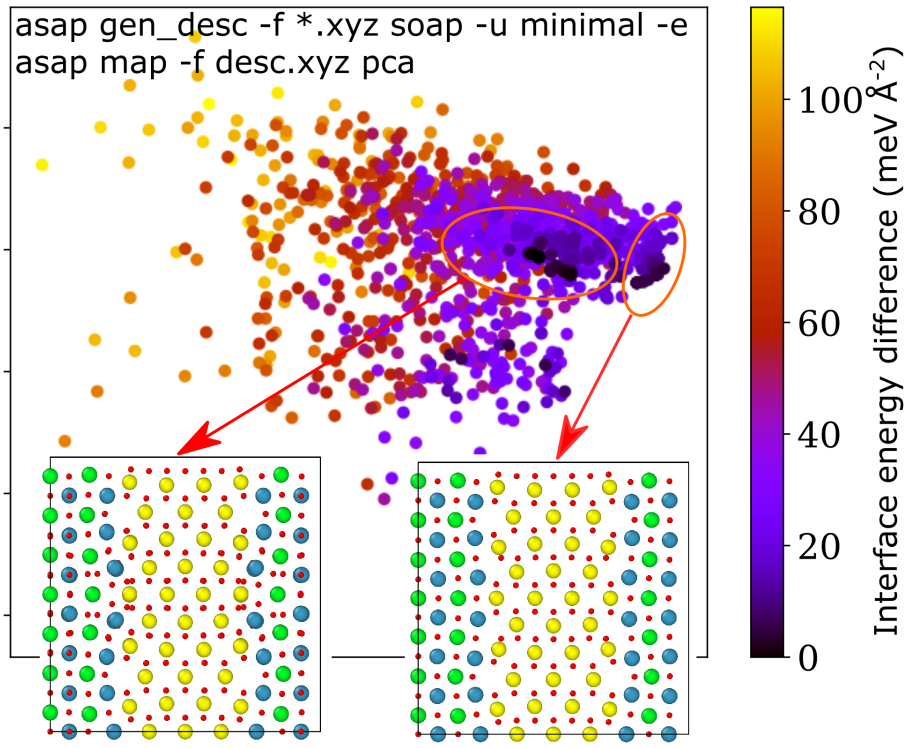


Figure 6: A PCA map for 1332 STO/CeO_2 (100)/(110) interface structures relaxed from randomly generated structures[45]. These structures consist of four layers on each side of the interface with a mirror plane at the middle. The red ovals indicate two distinct lowest-energy groups.

viewer to observe molecules along various "paths" through the map (illustrated in Figure 7a).

Color-coding the points on the map using elemental compositions (Fig. 7b) shows that pure hydrocarbons and other compositions (e.g. C,H,O or C,H,N,O) form separate clusters. Together Figs. 7c and d show that different carbon and total atom counts cause further splitting of these clusters. The key features of the map are thus mainly defined by molecular composition. Furthermore, systems with different numbers of rings also form distinct clusters across the map (Fig. 7g).

Molecular properties correlate both with the axes of the map, and with the molecular compositions. The atomization energy per atom (Fig. 7a) scales inversely with the total number of atoms (Fig. 7d)[4,47]. The reason is that most molecules in QM9 contain 9 non-

hydrogen atoms, so molecules with less overall atoms tend to have more double and triple bonds. This also explains the trend in the HOMO-LUMO gap ϵ_{gap} (Fig. 7e): unsaturated compounds tend to have lower gaps. On the other hand, HOMO energies ϵ_{HOMO} (Fig. 7e) are less systematic, presumably because the electronegativities of the contained elements and structural features like π-conjugation have a strong influence.

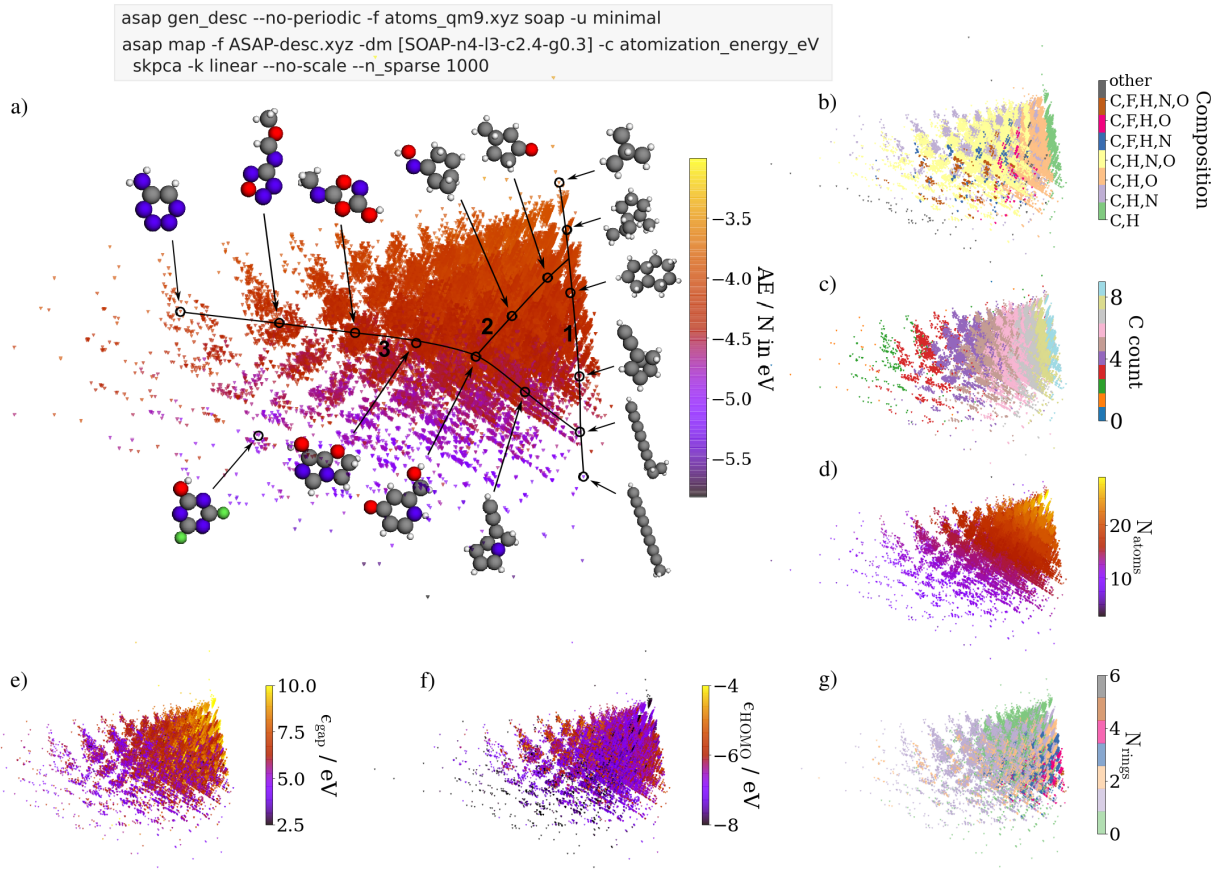


Figure 7: KPCA maps of the QM9 database using a global SOAP kernel. The frames are color-coded according to structural descriptors (b,c,d,g) and quantum mechanical properties (a,e,f).

As a complementary way to visualize QM9, we consider the atomic environments of all the carbon atoms (Figure 8). Upon inspection, the clusters of environments are found to reflect different numbers of neighboring carbon and hydrogen atoms (strongly correlated with the vertical and horizontal axes of the plot, respectively). The clusters thus correspond to atom-types, reflecting the fundamental concept behind classical bio-organic force-fields,

which define different atom-types according to the basic bonding topology of a molecule. Each cluster displays a fairly homogeneous Mulliken charge, with a large number of hydrogen neighbors leading to a negative partial charge on the carbon atom (and vice versa). Within each cluster, different realizations of the C/H neighborhoods cause further difference in the charges. For example, both a carboxylic acid and a $-CF_3$ group attached to a hydrocarbon contain a central carbon atom with a single carbon and no hydrogen neighbors. Such subclusters are illustrated in Figure 8b and c, according to the hybridization and whether a carbon atom is part of a ring. This also serves as a warning that dimensionality reduction may obscure some relevant structural features of the data. In this case, carbon and hydrogen are the most abundant elements in the dataset so they dominate the embedding, whereas the role of heteroatoms is not immediately clear. Color-coding using additional properties and inspecting representative structures can fill this gap.

Visualizing atomic environments also helps understand and interpret ML potentials. We consider a SOAP-based GAP model trained on QM9 energies[1], in which total energies are expressed as the sums of local atomic energies. Figure 8d is color-coded using these local energies and shows systematic trends of similar energies within each cluster and a smooth variation of energies between clusters. This shows how the local energies are related to the specific environments.

Polymorphs and Conformers of Oxalic Acid crystals

KPCA maps can also be used to emphasize differences in conformations or crystal polymorphs for systems with fixed composition. This is illustrated for oxalic acid (OA) in Figure 9. In the left panel, seven conformers of OA (each a (meta-)stable structure on the potential energy surface) are shown. The map intuitively arranges these structures according to the orientation of the protons (from left to right: in-in, in-out and out-out), and its vertical axis correlates with the relative energy.

Additionally, configurations sampled from a series of MD trajectories initialized from

each conformer geometry are shown. Note that the highest energy conformer is not thermally stable, and almost immediately rearranges during the MD. These configurations are arranged in larger basins separated by energetic barriers, while conformers within a basin are connected by low energy paths. In particular, the leftmost conformer (with corresponding points indicated by partial transparency) does not rearrange during the MD simulation due to the high kinetic stability afforded by the two intramolecular hydrogen bonds, whereas all other conformers are connected by the MD trajectories.

In the right panel, a similar KPCA plot is shown for 48 bulk crystal structures of OA, which were generated using random structure search. The initial random structures (small yellow circles) and the corresponding fully relaxed configurations (large colored circles) are connected by gray lines. All molecules in the structure search were initialized from the bottom-right conformer (out-out, trans) in the left panel, but in some cases the monomers in the relaxed structures belong to a different conformer (indicated by dashed arrows between the two panels). The random and optimized crystal structures stay on distinct regions of the plot. Almost all optimized structures are well-separated, indicating that there are many stable minima for the crystal, unlike in the gas-phase. Besides the two experimentally formed polymorphs, several other crystalline structures of OA with comparable energies are also found. This multitude of low energy local minima make organic crystal structure prediction difficult.

The two maps in Figure 9 highlight different aspects of molecular structure: the intramolecular aspects (mainly proton orientation) on the left, and the differences in intermolecular interactions on the right. Considering both thus allows a more complete understanding of the structural factors underpinning molecular crystal formation.

Conclusion

Automating the mapping of diverse classes of materials and molecules yields physical and chemical insights, saves human effort, and provides a data-driven perspective on large atomistic datasets. Because of this utility, and the software packages that are now available, we believe that these maps will become a standard tool for the wider computational chemistry and materials science community. From the perspective of methodology there is certainly room for improvement. For example, a systematic comparison of the maps produced using different descriptors and dimensionality reduction algorithms would be useful, as would be the development of new schemes that have better scaling with respect to the number of atomic species in the dataset. As in most works using ML for chemistry and materials to date, we have neglected long-range interactions and correlations. Incorporating descriptions of these may improve the ability of maps to discern and tease out such effects, for example in ionic solutions and large protein complexes.

All in all, beyond visualization being a valuable tool for molecular modelling, it is also becoming a science in itself. Without any doubt, this Account will not be the final word in this new science, and we anticipate exciting new developments.

Data availability The datasets and scripts for the visualization are uploaded to a public repository at

`https://github.com/BingqingCheng/Mapping-the-space-of-materials-and-molecules`.

The ASAP code and the interactive viewing tool are also available:

`https://github.com/BingqingCheng/ASAP`

`https://github.com/chkunkel/projection_viewer`

An alternative interactive viewing tool[50] developed by another research group is at

`https://chemiscope.org`.

The output of ASAP can be directly used as the input of either viewing tool.

Biographical Sketches

Bingqing Cheng is a junior research fellow in Cambridge, and her work focuses on theoretical predictions of material properties.

Ryan-Rhys Griffiths is a PhD student in Cambridge, working on machine learning methodology for scientific applications.

Simon Wengert is a PhD student at TU Munich, and his work focuses on machine-learning assisted crystal structure prediction.

Christian Kunkel is a PhD student at TU Munich, working on machine-learning based organic materials discovery.

Tamás K. Stenczel is a student at the University of Cambridge, and his research work focuses on machine-learning modelling of reactive chemical systems.

Bonan Zhu is a PhD student at the University of Cambridge, and his research work focuses on predicting and modelling oxides interfaces.

Volker Deringer is Associate Professor of Theoretical and Computational Inorganic Chemistry at the University of Oxford.

Noam Bernstein is a research physicist the U. S. Naval Research Laboratory.

Johannes T. Margraf is a research group leader at TU Munich, working on molecular machine learning.

Karsten Reuter is Professor of Theoretical Chemistry at TU Munich.

Gábor Csányi is Professor of Molecular Modelling at the Engineering Laboratory, University of Cambridge.

Acknowledgement

BC acknowledges funding from Swiss National Science Foundation (Project P2ELP2-184408). VLD acknowledges a Leverhulme Early Career Fellowship. SW acknowledges funding from Deutsche Forschungsgemeinschaft (DFG) and JM and CK are grateful for support by DFG

through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. The work of NB was supported by the U. S. Office of Naval Research through the U. S. Naval Research Laboratory's fundamental research base program.

Supporting information: Supporting Information Available: details on the universal SOAP heuristics. This material is available free of charge via the Internet at `http://pubs.acs.org`.

References

1. Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. Machine Learning Unifies the Modelling of Materials and Molecules. *Science Advances* **2017**, *3*, e1701816.

2. Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. *Physical Chemistry Chemical Physics* **2020**, *22*, 12697–12705.

3. Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **2020**, *7*, 58.

4. De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

5. Caro, M. A.; Aarva, A.; Deringer, V. L.; Csanyi, G.; Laurila, T. Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning. *Chemistry of Materials* **2018**, *30*, 7446–7455.

6. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

7. Bernstein, N.; Csányi, G.; Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Computational Materials* **2019**, *5*, 1–9.

8. Huang, J.-X.; Csányi, G.; Zhao, J.-B.; Cheng, J.; Deringer, V. L. First-principles study of alkali-metal intercalation in disordered carbon anode materials. *J. Mater. Chem. A* **2019**, *7*, 19070–19080.

9. Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials* **2015**, *27*, 735–743.

10. Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, between Predictions and Understanding. *Journal of Chemical Physics* **2019**, *150*, 150901.

11. Nüske, F.; Wu, H.; Prinz, J.-H.; Wehmeyer, C.; Clementi, C.; Noé, F. Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias. *The Journal of Chemical Physics* **2017**, *146*, 094104.

12. Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics New York, 2001; Vol. 1.

13. Cox, M. A.; Cox, T. F. *Handbook of data visualization*; Springer, 2008; pp 315–347.

14. Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.

15. Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011**, *108*, 13023–13028.

16. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**,

17. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics* **2011**, *134*, 074106.

18. Huo, H.; Rupp, M. Unified representation for machine learning of molecules and crystals. *arXiv preprint arXiv:1704.06439* **2017**, *13754*.

19. Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

20. Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

21. Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of chemical physics* **2018**, *148*, 241717.

22. Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *The Journal of chemical physics* **2013**, *139*, 184118.

23. Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. *Scientific Computing and Algorithms in Industrial Simulations*; Springer, 2017; pp 25–42.

24. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters* **2015**, *6*, 2326–2331.

25. Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-density representations for machine learning. *The Journal of chemical physics* **2019**, *150*, 154110.

26. Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **1998**, *10*, 1299–1319.

27. Tsuzuki, H.; Branicio, P. S.; Rino, J. P. Structural characterization of deformed crystals by analysis of common atomic neighborhood. *Computer physics communications* **2007**, *177*, 518–523.

28. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

29. Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of chemical physics* **2008**, *129*, 114707.

30. Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *arXiv e-prints* **2019**,

31. Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120*, 156001.

32. Bernstein, N.; Bhattarai, B.; Csanyi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angewandte Chemie International Edition* **2019**, *58*, 7057–7061.

33. Mavracic, J.; Mocanu, F. C.; Deringer, V. L.; Csanyi, G.; Elliott, S. R. Similarity Between Amorphous and Crystalline Phases: The Case of TiO2. *The Journal of Physical Chemistry Letters* **2018**, *9*, 2985–2990, PMID: 29763315.

34. Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **2017**, *95*, 094203.

35. Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csányi, G. Growth Mechanism

and Origin of High sp^3 Content in Tetrahedral Amorphous Carbon. *Phys. Rev. Lett.* **2018**, *120*, 166101.

36. Deringer, V. L.; Caro, M. A.; Jana, R.; Aarva, A.; Elliott, S. R.; Laurila, T.; Csanyi, G.; Pastewka, L. Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chemistry of Materials* **2018**, *30*, 7438–7445.

37. Ten Wolde, P. R.; Ruiz-Montero, M. J.; Frenkel, D. Numerical evidence for bcc ordering at the surface of a critical fcc nucleus. *Physical review letters* **1995**, *75*, 2714.

38. Lechner, W.; Dellago, C.; Bolhuis, P. G. Role of the prestructured surface cloud in crystal nucleation. *Phys. Rev. Lett.* **2011**, *106*, 085701.

39. Cheng, B.; Tribello, G. A.; Ceriotti, M. Solid-liquid interfacial free energy out of equilibrium. *Physical Review B* **2015**, *92*, 180102.

40. Cheng, B.; Ceriotti, M. Bridging the gap between atomistic and macroscopic models of homogeneous nucleation. *The Journal of chemical physics* **2017**, *146*, 034106.

41. Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Advanced Materials* **2019**, *31*, 1902765.

42. Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proceedings of the National Academy of Sciences* **2019**, *116*, 1110–1115.

43. Wang, L.; Ceriotti, M.; Markland, T. E. Quantum fluctuations and isotope effects in ab initio descriptions of water. *The Journal of chemical physics* **2014**, *141*, 104502.

44. Pickard, C. J.; Needs, R. Ab initio random structure searching. *Journal of Physics: Condensed Matter* **2011**, *23*, 053201.

45. Zhu, B.; Schusteritsch, G.; Lu, P.; MacManus-Driscoll, J. L.; Pickard, C. J. Determining Interface Structures in Vertically Aligned Nanocomposite Films. *APL Materials* **2019**, *7*, 061105.

46. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*.

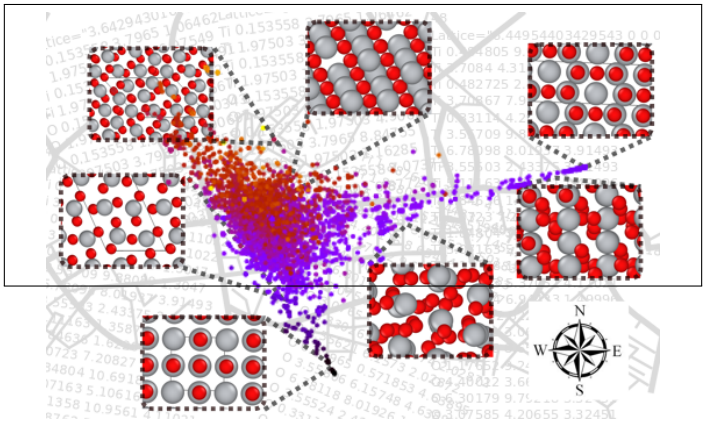
47. Jung, H.; Stocker, S.; Kunkel, C.; Oberhofer, H.; Han, B.; Reuter, K.; Margraf, J. T. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem* **2020**, *2*.

48. Thalladi, V. R.; Nüsse, M.; Boese, R. The Melting Point Alternation in α,ω-Alkanedicarboxylic Acids. *Journal of the American Chemical Society* **2000**, *122*, 9227–9236.

49. Derissen, J. L.; Smith, P. H. Refinement of the crystal structures of anhydrous α- and β-oxalic acids. *Acta Crystallographica Section B* **1974**, *30*, 2240–2242.

50. Fraux, G.; Cersonsky, R.; Ceriotti, M. Chemiscope: Interactive Structure-Property Explorer for Materials and Molecules. *Journal of Open Source Software* **2020**, *5*, 2117.

Graphical TOC Entry



```
asap gen_desc -f *.xyz --no-periodic soap -u minimal -pa
asap map -f ASAP-desc.xyz -dm '[*]' -ua --only_use_species 6 pca --no-scale
```

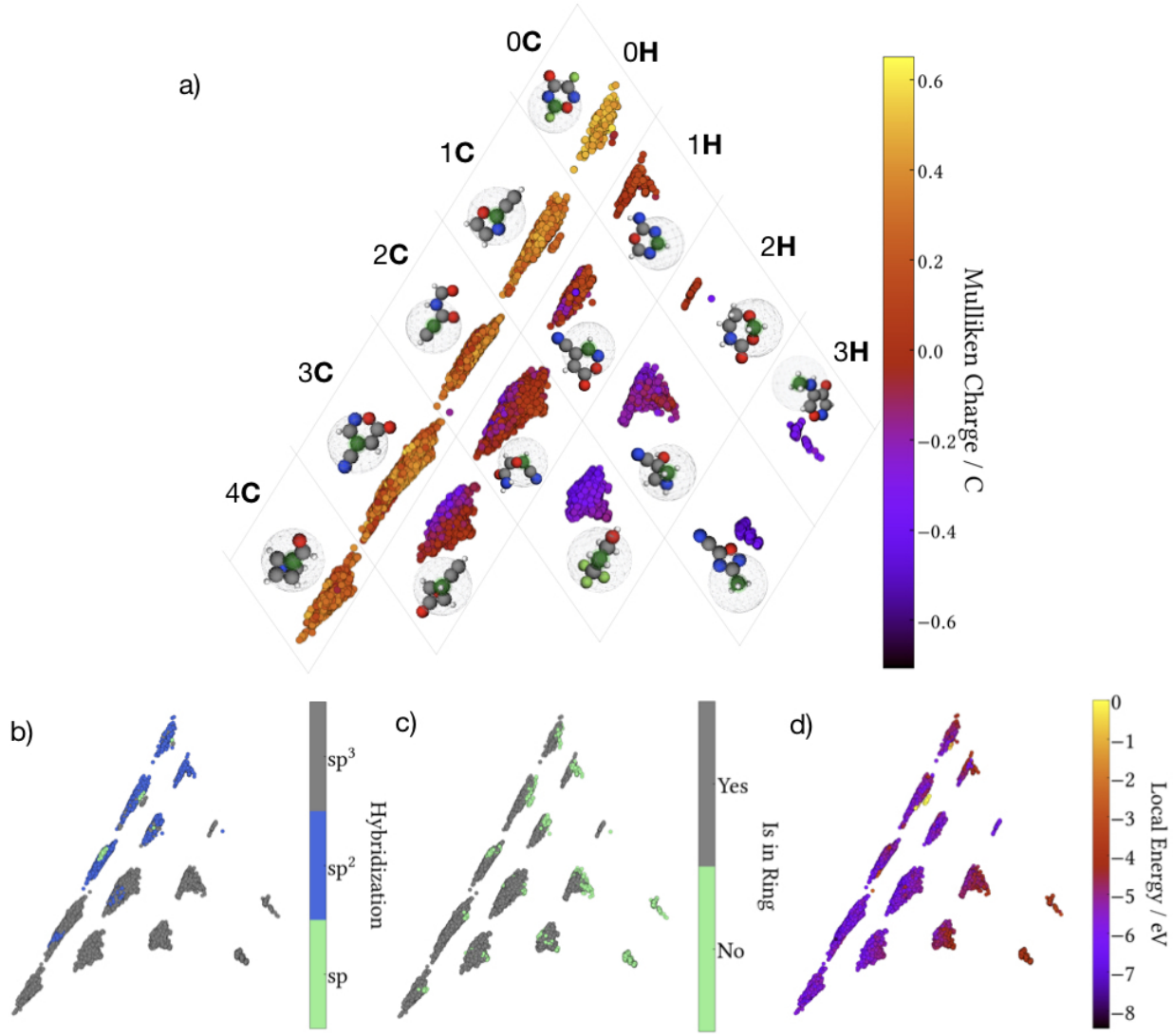


Figure 8: KPCA maps of carbon atom environments in the QM9 database. Maps are color-coded according to Mulliken charges (a), hybridization (b), whether the atoms are in rings (c) and according to local energies predicted by a machine learning potential (d).

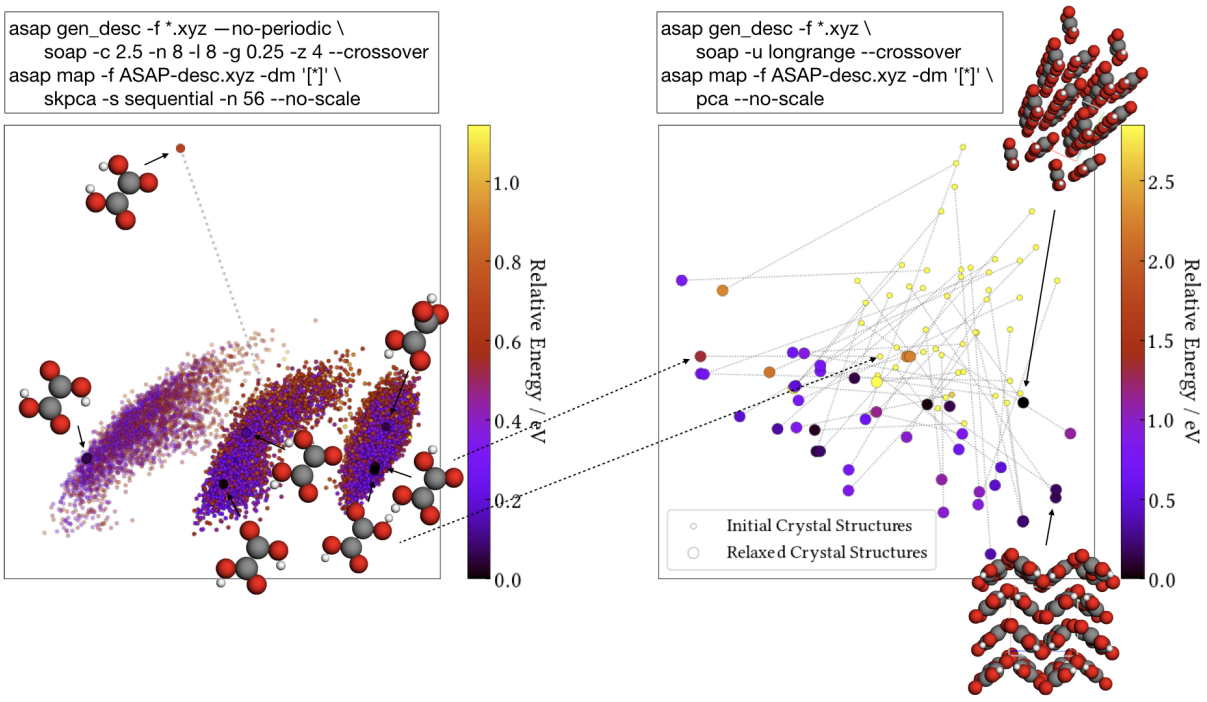


Figure 9: Left: KPCA map of oxalic acid conformers in the gas-phase (large points) and configurations from different MD simulations at 500 K initialized at the conformer geometries (small points). Configurations belonging to the MD for which no transitions to other basins are observed are shown as transparent points. Right: Randomly generated oxalic acid unit cells (small yellow circles) and the corresponding fully relaxed crystals (large colored circles). The experimentally known α[48] (lower structure) and β[49] (upper structure) polymorphs are highlighted. All random structures were initialized from the same gas-phase conformer, but in some cases the conformer changed upon relaxation (highlighted by arrows across the panels).