

A fingerprint based metric for measuring similarities of crystalline structures

Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S. Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker'

Citation: *The Journal of Chemical Physics* **144**, 034203 (2016); doi: 10.1063/1.4940026

View online: <http://dx.doi.org/10.1063/1.4940026>

View Table of Contents: <http://aip.scitation.org/toc/jcp/144/3>

Published by the [American Institute of Physics](#)

Articles you may be interested in

[Metrics for measuring distances in configuration spaces](#)

The Journal of Chemical Physics **139**, 184118184118 (2013); 10.1063/1.4828704

[Computationally efficient characterization of potential energy surfaces based on fingerprint distances](#)

The Journal of Chemical Physics **145**, 034101034101 (2016); 10.1063/1.4956461

[Crystal structure prediction using the minima hopping method](#)

The Journal of Chemical Physics **133**, 224104224104 (2010); 10.1063/1.3512900

[Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems](#)

The Journal of Chemical Physics **120**, (2004); 10.1063/1.1724816



**COMPLETELY
REDESIGNED!**

**PHYSICS
TODAY**

Physics Today Buyer's Guide
Search with a purpose.

A fingerprint based metric for measuring similarities of crystalline structures

Li Zhu,¹ Maximilian Amsler,^{1,2} Tobias Fuhrer,¹ Bastian Schaefer,¹ Somayeh Faraji,³ Samare Rostami,³ S. Alireza Ghasemi,³ Ali Sadeghi,⁴ Migle Grauzinyte,¹ Chris Wolverton,² and Stefan Goedecker^{1,a)}

¹Department of Physics, Universität Basel, Klingelbergstr. 82, 4056 Basel, Switzerland

²Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA

³Institute for Advanced Studies in Basic Sciences, P.O. Box 45195-1159, Zanjan, Iran

⁴Physics Department, Shahid Beheshti University, G. C., Evin, 19839 Tehran, Iran

(Received 23 July 2015; accepted 30 December 2015; published online 21 January 2016)

Measuring similarities/dissimilarities between atomic structures is important for the exploration of potential energy landscapes. However, the cell vectors together with the coordinates of the atoms, which are generally used to describe periodic systems, are quantities not directly suitable as fingerprints to distinguish structures. Based on a characterization of the local environment of all atoms in a cell, we introduce crystal fingerprints that can be calculated easily and define configurational distances between crystalline structures that satisfy the mathematical properties of a metric. This distance between two configurations is a measure of their similarity/dissimilarity and it allows in particular to distinguish structures. The new method can be a useful tool within various energy landscape exploration schemes, such as minima hopping, random search, swarm intelligence algorithms, and high-throughput screenings. © 2016 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4940026>]

I. INTRODUCTION

Large data sets of crystalline structures are nowadays available in two major contexts. On one hand, databases of materials have been created containing structural information of both experimental and theoretical compounds from high-throughput calculations, which are the basis for data-mining techniques in materials discovery projects.^{1–7} On the other hand, *ab initio* structure predictions^{8–15} can produce a huge number of new structures that have either not yet been found experimentally or are metastable.^{16–21} In both cases, it is essential to quantify similarities and dissimilarities between structures in the data sets, requiring a configurational distance that satisfies the properties of a metric. Databases frequently contain duplicates and insufficiently characterized structures which need to be identified and filtered. In experimental data, the representation of identical structures as obtained from different experiments will always slightly differ due to noise in the measurements, such that the configurational distance is never exactly zero. Noise is also present in theoretical calculations where a geometry relaxation is, for instance, stopped once a certain, possibly insufficient convergence threshold is reached. In *ab initio* structure prediction schemes, it is typically necessary to maintain some structural diversity which can be quantified as a certain minimal configurational distance. All these examples clearly show the need for a metric that allows to measure configurational distances and local structures in a reliable and efficient way.

Crystalline structures are typically given in a dual representation. The first part specifies the cell and the second part the atomic positions within the cell. The former can, for instance,

be given by the three lattice vectors **a**, **b**, and **c**, or by their lengths *a*, *b* and *c*, and the intermediate angles α , β , and γ . The atomic positions can either be specified by cartesian coordinates or by reduced coordinates with respect to the lattice vectors. However, such representations are not unique, since any choice of lattice points can serve as cell vectors of the same crystalline structure. Unique and preferably standardized cell parameters are required for comparison and analysis of different crystals.²² Algorithms to transform unit cells to a reduced form are frequently used in crystallography, such as the Niggli-reduction^{23–27} which produces cells with shortest possible vectors ($|\mathbf{a} + \mathbf{b} + \mathbf{c}| = \text{minimal}$). Unfortunately, in the presence of noisy lattice vectors, cells can change discontinuously within the Niggli-reduction algorithm. Symmetry analysis and the corresponding classification in the 230 crystallographic space groups are another tool to compare crystal structures. However, the outcome of a symmetry analysis algorithm strongly depends on a tolerance parameter such that the introduction of some noise can change the resulting space group in a discontinuous manner. Because of the above described problems, it is difficult to quantify similarities based directly on dual representations.

Within the structure prediction community, fingerprints that are not directly based on such a dual representation have been proposed. Radial distribution functions have been used as a fingerprint to measure the distances between crystal structures,^{28,29} as well as methods based on comparison of the calculated powder diffraction patterns.^{30,31} In the same spirit, Oganov and Valle³² used element resolved radial distribution functions as a crystal fingerprint. For a crystal containing one element, only a single function is obtained for the entire system. The difference between the radial distribution functions of two crystals is then taken as the configurational

^{a)}Electronic mail: stefan.goedecker@unibas.ch

distance. By definition, the radial distribution function contains only radial information, but no information about the angular distribution of the atoms. Such angular information has been added in the bond characterization matrix (BCM) fingerprint.^{13,33} In this fingerprint, spherical harmonic and exponential functions are used to set up modified bond-orientational order descriptors³⁴ of the entire configuration. The distance between two configurations can be measured by the Euclidean distance between their BCMs. For molecular crystals, Chisholm *et al.*³⁵ used intermolecular contact distances and a matching algorithm to characterize and compare structures. Atomic and molecular environment descriptors are also needed in the context of machine learning schemes for force fields,^{36–38} bonding pattern recognition,³⁹ or to compare vacancy, interstitial, and intercalation sites.⁴⁰ These descriptors could also be used to measure similarities between structures. Even though they have never been used in this context, we will present a comparison with such a descriptor.

When humans decide by visual inspection whether two structures are similar, they proceed typically in a different way. They try to find matching atoms which have the same structural environment. If all the atoms in one structure can be matched with the atoms of the other structure, the two structures are considered to be identical. Such a matching approach based on the Hungarian algorithm⁴¹ has already turned out to be useful for the distinction of clusters.^{42,43}

In this paper, we will present a fingerprint for crystalline structures which is based on such a matching approach. The environment of each atom is described by an atomic fingerprint which is calculated in real space for an infinite crystal and represents some kind of environmental scattering properties observed from the central atom. Therefore, all the ambiguities of a dual representation do not enter into the fingerprint, allowing an efficient and precise comparison of structures.

II. FINGERPRINT DEFINITION

Recently, we have proposed a configurational fingerprint for clusters.⁴³ In this approach, an overlap matrix is calculated for an atom centered Gaussian basis set. The vector formed by the eigenvalues of this matrix forms a global fingerprint that characterizes the entire structure. The Euclidean norm of the difference vector between two structures is the configurational distance between them and satisfies the properties of a metric.

Since there is no unique representation of a crystal by a group of atoms (e.g., the atoms in some unit cell), we will use atomic fingerprints instead of global fingerprints in the crystalline case. However, this atomic fingerprint is closely related to our global fingerprint for non-periodic systems. For each atom k in a crystal located at \mathbf{R}_k , we obtain a cluster of atoms by considering only those contained in a sphere centered at \mathbf{R}_k . For this cluster, we calculate the overlap matrix elements $S_{i,j}^k$, as described in Ref. 43 for a non-periodic system, i.e., we put on each atom one or several Gaussian type orbitals and calculate the resulting overlap integral. The orbitals are indexed by the letters i and j and the index $w(i)$ gives the index of the atom on which the Gaussian $G_i(\mathbf{r})$ is

centered, i.e.,

$$S_{i,j}^k = \int d\mathbf{r} G_i(\mathbf{r} - \mathbf{R}_{w(i)}) G_j(\mathbf{r} - \mathbf{R}_{w(j)}). \quad (1)$$

In this first step, the amplitudes of the Gaussians c_{norm} are chosen such that the Gaussians are normalized to one, and the width of each Gaussian $G_i(\mathbf{r})$ is given by the covalent radius of the atom $w(i)$ on which it is centered. To avoid discontinuities in the eigenvalues when an atom enters into or leaves the sphere, we construct in a second step another matrix T^k such that

$$T_{i,j}^k = f_c(|\mathbf{R}_{w(i)} - \mathbf{R}_k|) S_{i,j}^k f_c(|\mathbf{R}_{w(j)} - \mathbf{R}_k|). \quad (2)$$

The cutoff function f_c smoothly goes to zero on the surface of the sphere with radius $\sqrt{2n}\sigma_c$,

$$f_c(r) = \left(1 - \frac{r^2}{2n\sigma_c^2}\right)^n. \quad (3)$$

In the limit where n tends to infinity, the cutoff function converges to a Gaussian of width σ_c . The characteristic length scale σ_c is typically chosen to be the sum of the two largest covalent radii in the system.

The value n determines how many derivatives of the cutoff function are continuous on the surface of the sphere, and $n = 3$ was used in the following. One can consider the modified matrix T^k to be the overlap matrix of the cluster where the amplitude of the Gaussian at atom i is determined by $c_{\text{norm}} f_c(|\mathbf{R}_i - \mathbf{R}_k|)$. In this way atoms close to the surface of the sphere give rise to very small eigenvalues of T^k and are thus weighted less than the atoms closer to the center. The eigenvalues of this matrix T^k are sorted in descending order and form the atomic fingerprint vector \mathbf{V}_k . Since we cannot predict exactly how many atoms will be in the sphere, we estimate a maximum length for the atomic fingerprint vector. If the number of atoms is too small to generate enough eigenvalues to fill up the entire vector, the entries at the end of the fingerprint vector are filled up with zeros. This also guarantees that the fingerprint is a continuous function with respect to the motion of the atoms when atoms might enter or leave the sphere. If an atom enters into the sphere, some zeros towards the end of the fingerprint vector are transformed in a continuous way into some very small entries which only contribute little to the overall fingerprint. The Euclidean norm $|\mathbf{V}_k - \mathbf{V}_l|$ measures the dissimilarity between the atomic environments of atoms k and l .

The atomic fingerprints \mathbf{V}_k^p and \mathbf{V}_k^q of all the N_{at} atoms in two crystalline configurations p and q can now be used to define a configurational distance $d(p, q)$ between the two crystals,

$$d(p, q) = \min_P \left(\sum_k^{N_{\text{at}}} |\mathbf{V}_k^p - \mathbf{V}_{P(k)}^q|^2 \right)^{1/2}, \quad (4)$$

where P is a permutation function which matches a certain atom k in crystal p with atom $P(k)$ in crystal q . The optimal permutation function which minimizes $d(p, q)$ can be found with the Hungarian algorithm⁴¹ in polynomial time. If the two crystals p and q are identical, the Hungarian algorithm will in this way assign corresponding atoms to each other. The

Hungarian algorithm needs as its input only the cost matrix C given by

$$C_{k,l} = |\mathbf{V}_k^p - \mathbf{V}_l^q|^2.$$

In the following, it will be shown that $d(p, q)$ satisfies the properties of a metric, namely,

- positiveness: $d(p, q) \geq 0$,
- symmetry: $d(p, q) = d(q, p)$,
- coincidence axiom: $d(p, q) = 0$ if and only if $p = q$,
- triangle inequality: $d(p, r) + d(r, q) \geq d(p, q)$.

From the definition (Eq. (4)), it is obvious that the positiveness and symmetry conditions are fulfilled. The coincidence theorem is satisfied if the individual atomic fingerprints are unique, i.e., if there are not two different atomic environments that give rise to identical atomic fingerprints. In our work on fingerprints for clusters, we have shown that the fingerprints can be considered to be unique if they have a length larger or equal to 3 per atom. The triangle inequality can be established in this way,

$$\begin{aligned} d(p, r) + d(r, q) &= \left(\sum_k^{Nat} |\mathbf{V}_k^r - \mathbf{V}_{P(k)}^p|^2 \right)^{1/2} \\ &\quad + \left(\sum_k^{Nat} |\mathbf{V}_k^q - \mathbf{V}_{P'(k)}^r|^2 \right)^{1/2} \\ &\geq \left(\sum_k^{Nat} |\mathbf{V}_{P(k)}^p - \mathbf{V}_{P'(k)}^q|^2 \right)^{1/2} \\ &\geq \left(\sum_k^{Nat} |\mathbf{V}_k^p - \mathbf{V}_{Q(k)}^q|^2 \right)^{1/2} \\ &= d(p, q), \end{aligned}$$

where P , P' , and Q are assumed to be the permutations that minimize respectively the Euclidean vector norms associated to $d(p, r)$, $d(r, q)$, and $d(p, q)$.

III. CONTRACTED FINGERPRINTS

Since the \mathbf{R}_k -centered spheres contain typically about 50 atoms, an atomic fingerprint has at least length 50 if only s -type Gaussian orbitals or length 200 if both s and p orbitals are used. Since a configuration is characterized by the ensemble of all the atomic fingerprints of all the atoms in the cell, the amount of data needed to characterize a structure is quite large even though it is certainly manageable for crystals with a small number of atoms per unit cell. Storage requirements might however become too high in certain cases such as large molecular crystals. We will, therefore, introduce contraction schemes that allow to considerably reduce the amount of data necessary to characterize a crystalline structure. Two such schemes will briefly be discussed below.

A. Contractions by properties

Let us introduce a function $\tau(i)$ that designates a certain property of the Gaussian orbital i and encodes it in form of a contiguous integer index. In case of a multicomponent

crystal, it can indicate on which kind of chemical element the Gaussians are centered and whether the orbital is of s or p type. The principal vector is thus chopped into pieces whose elements all carry the same value $\tau(i)$. In the following presentation of numerical results, we have always considered the central atom to be special, independent of its true chemical type. Having m atomic species in the unit cell and using atomic Gaussian orbitals with a maximum angular momentum l_{\max} , $\tau(i)$ runs from 1 to $(m+1)(l_{\max}+1)$. Now we can construct a contracted matrix t^k ,

$$t_{\nu, \mu}^k = \sum_{i, j} \delta_{\nu, \tau(i)} u_i^k T_{i, j}^k u_j^k \delta_{\mu, \tau(j)},$$

together with its metric tensor s^k ,

$$s_{\nu, \mu}^k = \sum_i \delta_{\nu, \tau(i)} u_i^k u_j^k \delta_{\mu, \tau(j)},$$

where \mathbf{u}^k is the principal vector of the matrix T^k of Eq. (2). The eigenvalues λ of the generalized eigenvalue problem,

$$t^k \mathbf{v} = \lambda s^k \mathbf{v},$$

form again an atomic fingerprint of length $(m+1)(l_{\max}+1)$ which is much shorter than the non-contracted fingerprint \mathbf{V}_k .

B. Contractions to form molecular orbitals for molecular crystals

The fingerprints described so far can in principle also be used for molecular crystals. However, the amount of data needed to characterize such crystals can be quite large if the molecules forming the crystal contain many atoms. By creating molecular orbitals in analogy with standard methods in electronic structure calculations, the required amount of data can be considerably reduced. The eigenvalues arising from the overlap matrix in this molecular basis set will then form a fingerprint for the molecular crystal. The molecular orbitals can be obtained in the following way: for each molecule k in our unit cell, we cut out a cluster of molecules within a sphere of a certain radius. For each molecule α in this sphere, we set up the overlap matrix by putting Gaussian type orbitals on all its constituent atoms. Then, we calculate for this matrix the eigenvalues and eigenvectors. The principal vectors $\mathbf{W}^{\alpha, \mu}$ belonging to several of the largest eigenvalues λ_μ^α are subsequently used for the contraction,

$$S_{\alpha, \mu; \beta, \nu}^k = \sum_{i, j} W_i^{\alpha, \mu} S_{i, j}^k W_j^{\beta, \nu}. \quad (5)$$

No metric tensor is required since the set of vectors used for the contraction is orthogonal. The molecular orbitals have characteristic patterns, such that the orbital corresponding to the first principal vector has no nodes, while the orbitals of the following principal vectors have increasing number of nodes. They are therefore similar to the atomic orbitals of s , p and higher angular momentum character, which were used for the fingerprints in the ordinary crystals. In Fig. 1, these orbitals are shown for the case of the paracetamol molecule.

By multiplying S with some cutoff function as in Eq. (3), we can then obtain molecule centered overlap matrices in this molecular basis which is free of discontinuities with

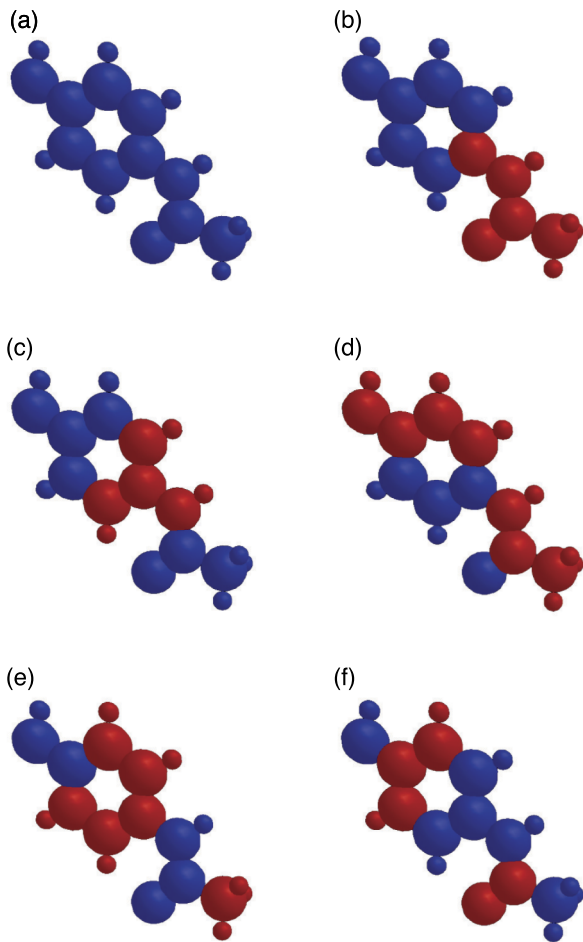


FIG. 1. The nodal character of the first six principal vectors for the paracetamol molecule. The atoms are colored according to the sign of the elements of the first six principal vectors $\mathbf{W}^{\alpha,\mu}$. A systematic colour pattern can be observed. The first principal eigenvector never changes sign and has therefore no nodes (a). Higher principal vectors exhibit more and more nodes ((b)-(f)).

respect to the motion of the atoms. In the molecular case, the value of the cutoff function depends on some short range pseudo-interaction between the central and the surrounding molecules. This interaction $U_{k,\alpha}$ between the central molecule k and another molecule α is given by

$$U_{k,\alpha} = \sum_{i,j} \left(1 - \frac{d_{i,j}^2}{2n(\sigma_{i,j}^{van})^2} \right)^n, \quad (6)$$

where the sum over i runs over all the atoms in the central molecule k and the sum over j over all the atoms in the surrounding molecule α . $d_{i,j}$ is the distance between the atoms i and j and $\sigma_{i,j}^{van}$ is the sum of the van der Waals radii of the two atoms. The interaction is taken to vanish beyond its first zero. Because of the short range of the interaction, molecules sharing a large surface will be coupled strongly. The analytical form of the cutoff function is identical to the one for the atomic case (Eq. (3)). However, since a cartesian distance between molecules is ill defined, the argument in Eq. (3) is modified. The scaled distance $r\sigma_c$ between the atoms is replaced by the normalized interaction between the molecules,

$$\frac{U_{k,\alpha}}{U_{k,k}}.$$

The eigenvalues of this final overlap matrix form now a fingerprint describing the environment of this molecule k with respect to the other molecules. To compare two structures, this procedure is done for all molecules contained in the corresponding unit cell. A configurational distance is calculated then as in Eq. (4) by using the Hungarian algorithm.⁴¹

IV. APPLICATION OF FINGERPRINT DISTANCES TO EXPERIMENTAL STRUCTURES

Structural data found in various material databases are frequently obtained from measurements at different temperatures which results in thermal expansion. Similarly, measurements at different pressures or low quality x-ray diffraction patterns can lead to slight cell distortions. Obviously, our fingerprint distances among such expanded or distorted but otherwise identical structures are different from zero. For these reasons, we have introduced a scheme where the six degrees of freedom associated to the cell are optimized while keeping the reduced atomic coordinates fixed such as to obtain the smallest possible distance to a reference configuration. The gradient of our fingerprint distance with respect to the lattice vectors can be calculated analytically using the Hellmann Feynman theorem. An application of the lattice optimization scheme was applied to a subset of ZrO_2 structures taken from the Open Quantum Materials Database (OQMD),^{2,7} as will be discussed in further detail later in Sec. V.

V. NUMERICAL TESTS

Fig. 2 shows all possible pairwise configurational distances obtained with several fingerprints for various data sets. Different fingerprints are plotted along the x and y axis. Hence, for two different fingerprints which performed identically well all data points would lie on a diagonal line. Furthermore, an ideal fingerprint produces a clear gap in the fingerprint distances, separating structures that are either classified as identical or different. LFP stands for the uncontracted long fingerprint and in square parenthesis it is indicated whether only s or both s and p orbitals were used to set up the overlap matrix, SFP[s] stands for the short contracted fingerprint with s orbitals only where the properties used for the contraction are central atom and the element type of the neighboring atoms in the sphere. For materials that have only one type of element (Si in our case), the atomic fingerprint has only length two and the coincidence theorem is not satisfied. Even though there are hyperplanes in the configurational space where different configurations have identical fingerprints, it is very unlikely that different local minima lie on such hyperplanes and the fingerprint can therefore nevertheless well distinguish between identical and distinct structures. If both s and p orbitals are used (SFP[sp]), the atomic fingerprint has at least length 4 and no problem with the coincidence theorem arises. In addition, we also show the configurational distances arising from the Oganov and Valle and BCM^{13,33} fingerprints as well

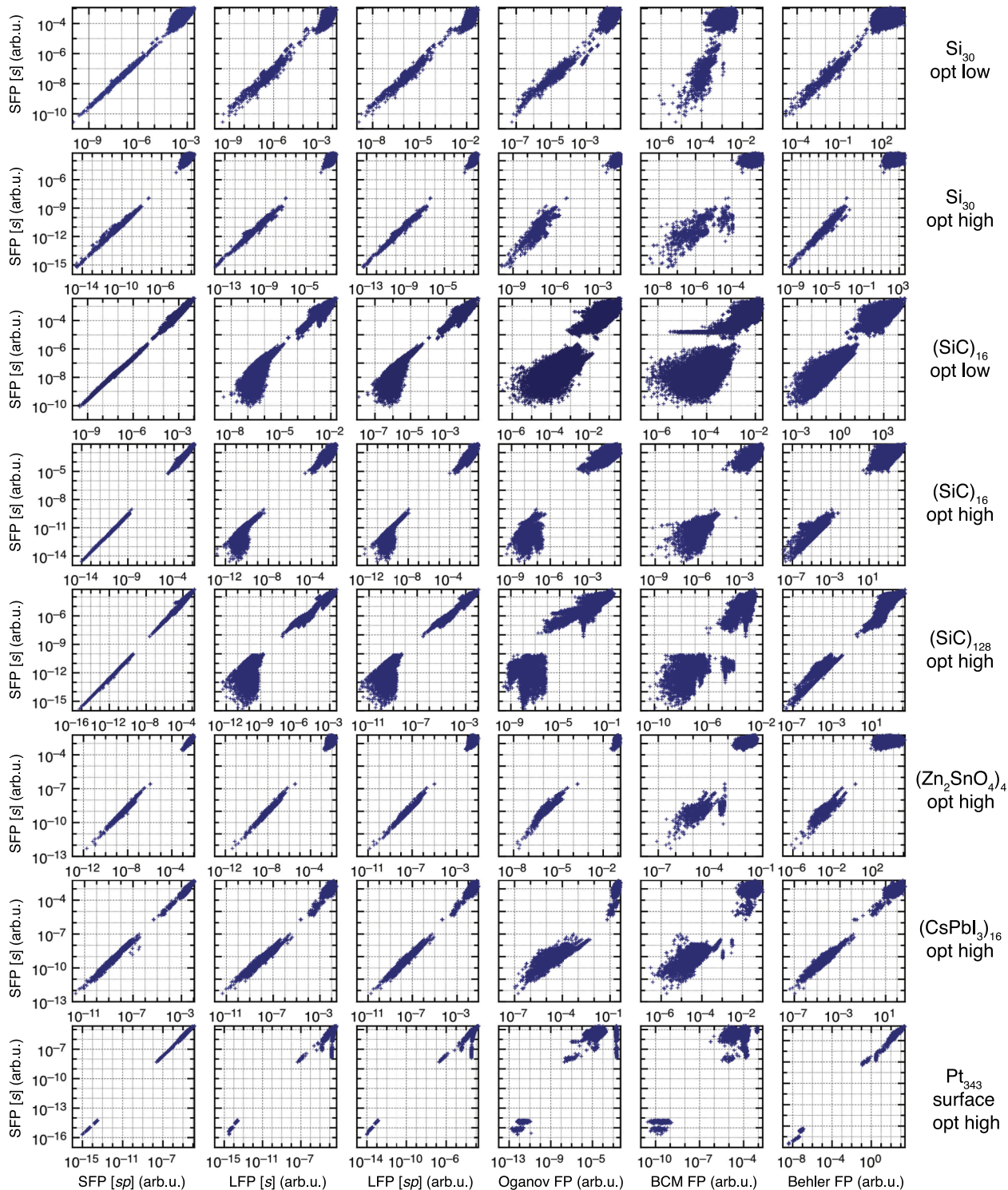


FIG. 2. Correlation between different fingerprints for all the 8 test sets obtained during structure prediction runs. SFP[x] and LFP[x] indicate short and long fingerprints with x orbitals, respectively. “Opt high” and “Opt low” indicate the quality of the geometry relaxation.

as from a fingerprint based on the amplitudes of symmetry functions.³⁶ All our data sets contain both the global minimum (geometric ground state) as well as local minima (metastable) structures, obtained from minima hopping runs.⁹ Energies and forces were calculated with the Density Functional based Tight binding method (DFTB+)⁴⁴ method for SiC and the molecular crystals, and the Lenosky tight-binding scheme was used for Si.⁴⁵ For the CsPbI₃ perovskite and the transparent

conductive oxide Zn₂SnO₄, plane wave density functional theory (DFT) calculations were used as implemented in the Quantum Espresso code.^{46,47}

The first test set consists of clathrate like structures of low density silicon allotropes.⁴⁸ Low density silicon gives rise to a larger number of low energy crystalline structures than silicon at densities of diamond silicon and thus poses an ideal benchmark system. In the first line of the figure,

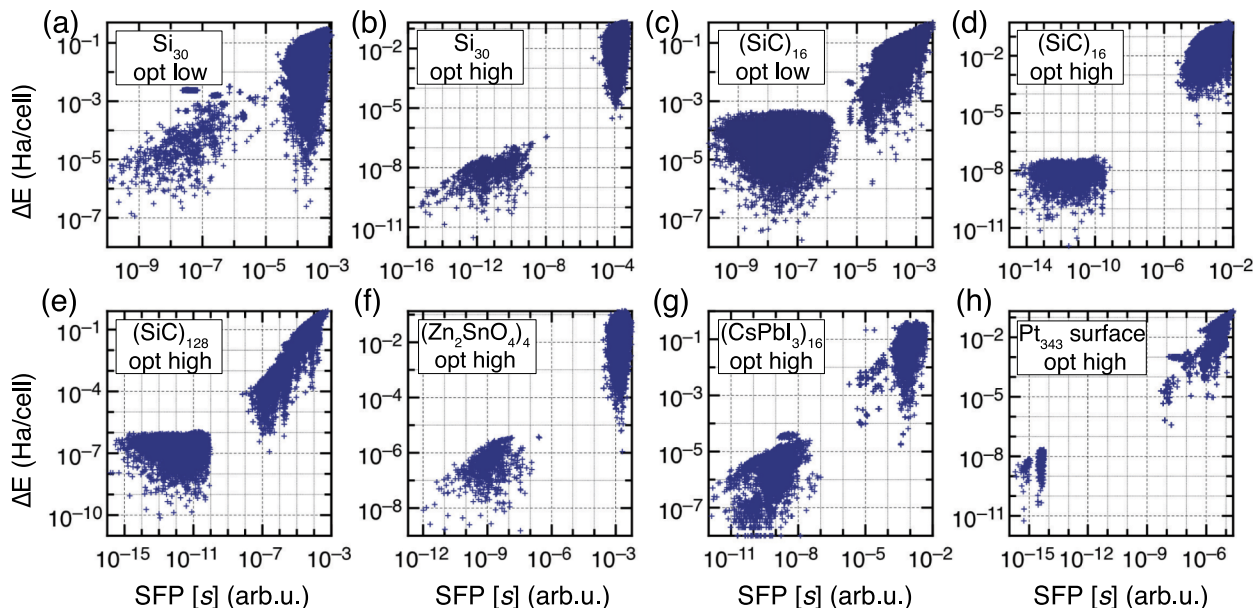


FIG. 3. Correlation between the energy difference and the fingerprint distance for all the 8 theoretical test setting of Fig. 2.

we show the results of a relatively sloppy local geometry optimization, where the relaxation is stopped once the forces are smaller than 0.05 eV/\AA . Gaps separating identical from distinct structures are hardly visible for all fingerprints. Once a very accurate geometry optimization with a force threshold of 5 meV/\AA is performed, gaps become visible for all the fingerprints.

The second data set is silicon carbide, a material well known for its large number of polytypes. Our fingerprint gives rise to a small gap whereas the configurational distances based on all other fingerprints do not show any gap at all. The opening of a gap can again be observed once the geometry optimization is done with high accuracy. For this case, all fingerprints result in a gap, but like for all test sets, it is the least pronounced for the BCM fingerprint. Both the Oganov and BCM fingerprints are global ones and in the required averaging process over all the atoms more and more information is lost as the system gets larger. Therefore, it is not surprising that the gap again disappears even for the high quality geometry optimization once one goes to large cells.

The next two test sets consist of an oxide material and a perovskite with their characteristic building blocks of octahedra and tetrahedra which can be arranged in a very large number of different ways. All our fingerprints give rise to clear gaps separating identical from distinct structures. The Oganov fingerprint also gives rise to clear gaps whereas the BCM fingerprint only weakly indicates some gap. The Behler fingerprint gives a well pronounced gap for Zn_2SnO_4 but only a blurred gap for CsPbI_3 .

The last theoretical test system is a platinum surface. In this case, the energies were calculated with the Morse potential.^{49,50} The geometry optimization was done with high accuracy and therefore a big gap is visible in all cases.

Fig. 3 shows the correlation between the energy difference and the fingerprint distance for all the test cases of Fig. 2. Except for the very large 256 atoms system, there exists always

a clear energy gap if the geometry optimization was done with high accuracy. Even though there is of course the possibility of nearly degenerate structures, this seems to happen rarely in practice and energy is thus a rather good and simple descriptor for small unit cells.

To test our molecular fingerprint (MFP), two test systems were employed, namely, crystalline formaldehyde and paracetamol. The formaldehyde system comprised 240 structures with 8 molecules per cell and the paracetamol

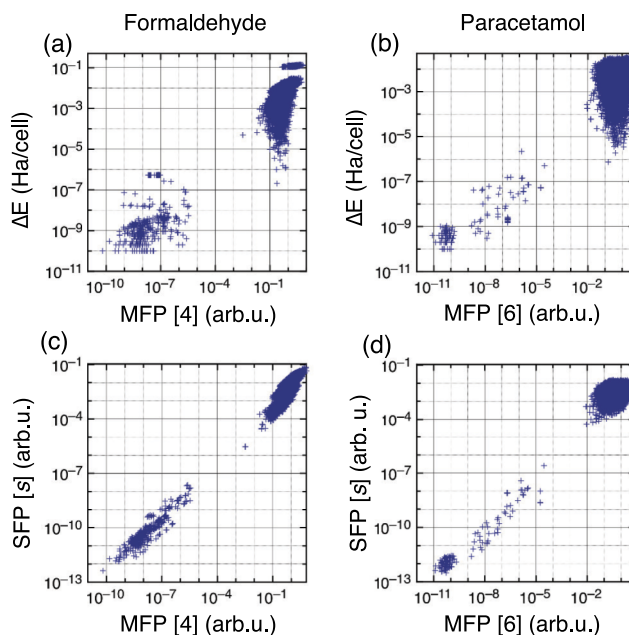


FIG. 4. Top panels: Correlation between the energy difference and the molecular fingerprint (MFP) distance for formaldehyde (a) and paracetamol (b). Bottom panels: Correlation between molecular fingerprint distance and standard fingerprint distance (short contracted fingerprint with s orbitals only, SFP[s]) for formaldehyde (c) and paracetamol (d).

system 300 structures with 4 molecules per cell. The two top panels of Fig. 4 show the molecular fingerprint distance versus the energy difference of different structures of paracetamol and formaldehyde, respectively. The two bottom panels show the correlation of the standard fingerprint against the molecular fingerprint for both systems. The existence of a gap in the pairwise distance distributions clearly indicates that identical and distinct structures can be identified by both fingerprints. However, the molecular fingerprint vector is considerably shorter because only six principal vectors were used (shown in Fig. 1). Since six is the number of degrees of freedom of a rigid rotator, it is expected that this fingerprint is long enough to satisfy the coincidence theorem.

COMPACK is a program developed by Chisholm *et al.*³⁵ to compare molecular crystal structures. It is interfaced with the Cambridge Structural Database System through the CSD Python API.⁵¹ Both the formaldehyde and paracetamol test sets were used to compare the accuracy of the new and the COMPACK molecular fingerprint methods. Fig. 5 shows the COMPACK fingerprint distances plotted against our molecular fingerprint. In the panels (a) and (b), the default tolerance values of COMPACK was used, namely, distance tolerance = 20% and angle tolerance = 20° whereas for panels (c) and (d), we used larger values of 40% and 40°, respectively. In all 4 cases, there is a gap, showing that both methods are capable to distinguish between identical and different structures. Overall, the gaps for the distances computed by COMPACK are significantly smaller than with our new fingerprint. Especially for paracetamol with the default tolerances, the gap almost vanishes. Increasing the tolerances improves the gap (see panels (c) and (d)), and both fingerprints perform similarly well.

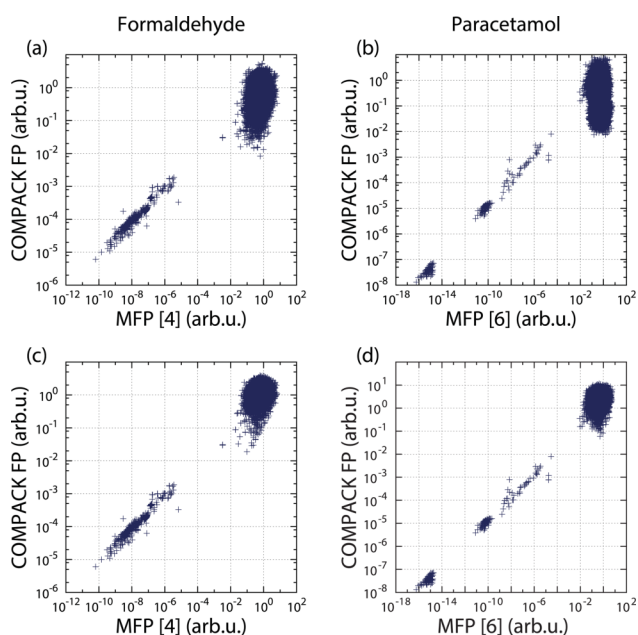


FIG. 5. Comparison of the new molecular fingerprint (MFP) distance with the COMPACK structural comparison engine as interfaced through the CSD Python API for formaldehyde in (a) and (c) and paracetamol in (b) and (d). Panels (a) and (b) and (c) and (d) show the COMPACK distances computed with distance/angle tolerances of 20%/20° and 40%/40°, respectively.

Next, we applied our fingerprint to ZrO_2 structures contained in the OQMD. 115 different entries were available at this composition. The structures were either based on experimental data originating from the Inorganic Crystal Structure Database (ICSD) or on binary structural prototypes. When the OQMD was initially created, duplicate entries were identified with the structure comparison algorithm as implemented in the Materials Interface (MINT) software package⁵² which employs a 6-level test that includes cell reduction as well as an analysis of the lattice symmetry. Structures classified as identical to an existing entry in OQMD were mapped to that entry without performing a structural relaxation. Therefore, the structural data set contains both DFT optimized and experimental structures, resulting in noise on the atomic and cell coordinates arising from the numerical calculations as well as from the different experiments and thermal effects. In Fig. 6(a), we show the ordinary and the lattice vector optimized fingerprint distances for all 115 structures from the database. We can see that the optimization process can reduce the fingerprint distance down to about 10^{-7} for many structures. For some of them, the initial fingerprint

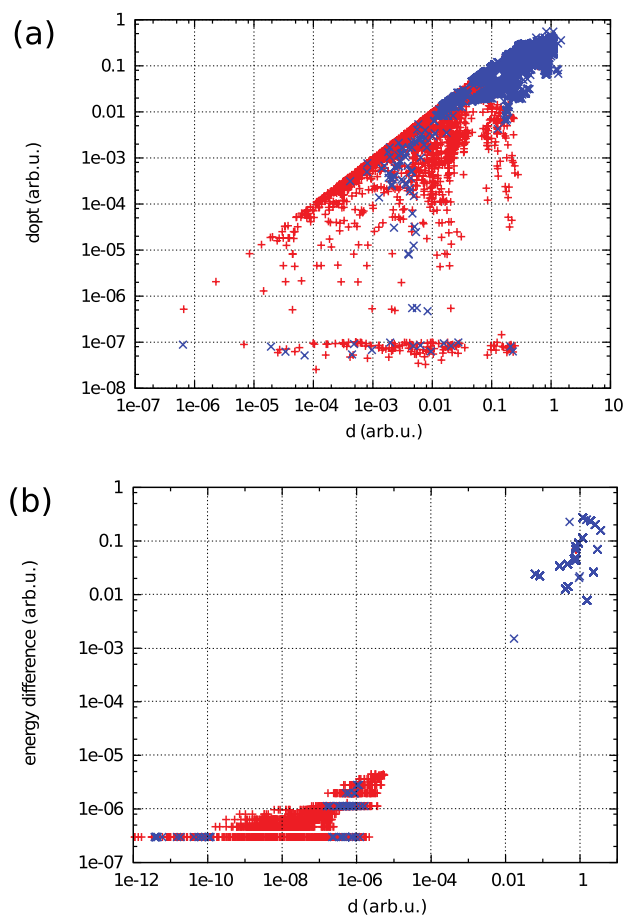


FIG. 6. Panel (a) shows along the x-axis the ordinary fingerprint distances and along the y-axis the lattice optimized fingerprint distances for the ZrO_2 structures retrieved from the OQMD. Distances between two structures that were identified as identical by the structural comparison algorithm implemented in MINT are shown in red and structures that were identified as distinct are shown in blue. Panel (b) shows the correlation between the DFT energy differences among all relaxed structures and the ordinary fingerprint distances.

distances were as large as 0.1. This allows to detect some identical structures whose initial large fingerprint distance was only due to thermal expansion. However, even with lattice vector optimization, it was not possible to decide for the whole data set in an unambiguous way which structures are identical and which were not. Therefore, local geometry optimizations were performed at the DFT level for all structures using the VASP code.^{53–55} A plane wave cutoff energy of 520 eV was used together with a dense k -point mesh. Both the atomic and cell variables were relaxed until the maximal force component was less than 2 meV/Å and the stress below 0.01 GPa. Panel (b) of Fig. 6 shows the DFT energy differences of the relaxed structures against the fingerprint distances, showing a clear gap that allows to distinguish between identical and different structures. Applying the lattice vector optimization scheme on these relaxed structures was not able to further lower the fingerprint distances of identical structures. The coloring in Fig. 6 indicates how the two structures belonging to a fingerprint distance were classified by MINT. Assuming that there are no different structures with degenerate DFT energies, one can conclude that MINT was not able to extract from the non-relaxed data set the information whether structures are identical or not and has erroneously assigned numerous identical structures as distinct and vice versa to a lesser extent.

Since both Oganov and BCM methods are global fingerprints that discard crucial information, they can fail to describe structural differences, a problem that becomes especially apparent when considering defect structures in complex materials. As an example, a $2 \times 2 \times 2$ supercell was constructed of the cubic perovskite structure of LaAlO_3 .⁵⁶ Half of the Al atoms on the B-sites were replaced by Mn. Then, single oxygen vacancies were introduced on symmetrically inequivalent X-sites. Obviously, the structural symmetry was reduced from the initial space group $Pm\bar{3}m$ of LaAlO_3 to the orthorhombic space group $Amm2$ of the supercell $\text{La}(\text{Al},\text{Mn})\text{O}_3$, and the oxygen vacancies resulted in structures with Cm and Pm symmetry. Both MINT and our fingerprint confirm that the structures are clearly different, whereas the Oganov and BCM fingerprint erroneously classify both structures as identical.

VI. CONCLUSIONS

Atomic fingerprints that describe the scattering properties as obtained from an overlap matrix are well suited to characterize atomic environments. An ensemble of atomic fingerprints forms a global fingerprint that allows to identify crystalline structures and to define configurational distances satisfying the properties of a metric. The widely used Oganov and BCM fingerprints do not have these properties and do also in practice not allow a reliable way to distinguish identical from distinct structures. Symmetry function based fingerprints are of similar quality as our scattering fingerprints. However, they are much more costly to calculate. Both fingerprints have a cubic scaling with respect to the number of atoms within the cutoff range, but our prefactor of the matrix diagonalization is much smaller than the prefactor for the 3-body terms required for the calculation of the symmetry functions. In contrast to

“true”–“false” schemes such as employed in MINT which rely on a threshold and affirm that two structures are either identical or distinct, our fingerprint gives a distance between configurations. The appearance of a gap in the distance distribution indicates that a reliable assignment of identical and distinct structures can be performed. In addition, strong reductions in the fingerprint distances upon lattice vector optimization can detect and eliminate thermal noise on the data set, rendering our fingerprint ideal to scan for duplicates in large structural databases. Furthermore, the fingerprint distance can be used as a simple measure of the similarity between two structures with a higher likelihood that they are identical if the fingerprint distance is small. In this context, machine learning methods (such as Bayesian techniques,⁵⁷ support vector machines,⁵⁸ and neural network^{59–62}) could be trained with attributes derived from our fingerprint to better quantify crystal structure similarities. The new fingerprint can also accurately explore local environments to create atomic and structural attributes for machine learning techniques trained to predict material properties. Our scheme can easily be extended to molecular crystals by introducing quantities that are analogous to molecular orbitals. Even for large molecular crystals, the molecular fingerprint is short and still very accurate. They could thus be used in the context of crystal structure prediction, where the potential energy surface is explored and a large number of structures need to be compared to eliminate duplicates. In summary, we have demonstrated that our novel approach allows to characterize crystalline structures by rather short fingerprint vectors and to decide more reliably whether structures are identical or not than previously proposed methods.

ACKNOWLEDGMENTS

We thank Vinay Hegde and Antoine Emery for valuable expert discussions. This work was done within the NCCR MARVEL project. M.A. gratefully acknowledges support from the Novartis Universität Basel Excellence Scholarship for Life Sciences and the Swiss National Science Foundation. C.W. acknowledges financial support from the U.S. Department of Energy under Grant No. DE-FG02-07ER46433. Computer resources were provided at the CSCS under project No. s499 and the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

¹D. Morgan, G. Ceder, and S. Curtarolo, *Meas. Sci. Technol.* **16**, 296 (2004).

²J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *JOM* **65**, 1501 (2013).

³S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, *Comput. Mater. Sci.* **58**, 218 (2012).

⁴A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, *APL Mater.* **1**, 011002 (2013).

⁵M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, and M. Asta, *Sci. Data* **2**, 150009 (2015).

⁶X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss, and K. A. Persson, *Comput. Mater. Sci.* **103**, 56 (2015).

- ⁷S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies," *npj Comput. Mater.* **1**, 15010 (2015).
- ⁸S. Goedecker, *J. Chem. Phys.* **120**, 9911 (2004).
- ⁹M. Amsler and S. Goedecker, *J. Chem. Phys.* **133**, 224104 (2010).
- ¹⁰C. W. Glass, A. Oganov, A. R. Oganov, and N. Hansen, *Comput. Phys. Commun.* **175**, 713 (2006).
- ¹¹C. J. Pickard and R. J. Needs, *J. Phys.: Condens. Matter* **23**, 053201 (2011).
- ¹²Y. Wang, J. Lv, L. Zhu, and Y. Ma, *Phys. Rev. B* **82**, 094116 (2010).
- ¹³Y. Wang, J. Lv, L. Zhu, and Y. Ma, *Comput. Phys. Commun.* **183**, 2063 (2012).
- ¹⁴J. C. Schön and M. Jansen, *Z. Kristallogr. - Cryst. Mater.* **216**, 307 (2001).
- ¹⁵S. M. Woodley, P. D. Battle, J. D. Gale, C. Richard, and A. Catlow, *Phys. Chem. Chem. Phys.* **1**, 2535 (1999).
- ¹⁶M. Amsler, J. Flores-Livas, L. Lehtovaara, F. Balima, S. Ghasemi, D. Machon, S. Pailhès, A. Willand, D. Caliste, S. Botti, A. San Miguel, S. Goedecker, and M. Marques, *Phys. Rev. Lett.* **108**, 065501 (2012).
- ¹⁷M. Amsler, J. A. Flores-Livas, T. D. Huan, S. Botti, M. A. L. Marques, and S. Goedecker, *Phys. Rev. Lett.* **108**, 205505 (2012).
- ¹⁸M. Amsler, S. Botti, M. A. L. Marques, and S. Goedecker, *Phys. Rev. Lett.* **111**, 136101 (2013).
- ¹⁹L. Zhu, Z. Wang, Y. Wang, G. Zou, H.-k. Mao, and Y. Ma, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 751 (2012).
- ²⁰L. Zhu, H. Liu, C. J. Pickard, G. Zou, and Y. Ma, *Nat. Chem.* **6**, 644 (2014).
- ²¹W. Zhang, A. Oganov, A. R. Oganov, A. F. Goncharov, Q. Zhu, S. E. Boulfelfel, A. O. Lyakhov, E. Stavrou, M. Somayazulu, V. B. Prakapenka, and Z. Konopkova, *Science* **342**, 1502 (2013).
- ²²D. C. Lonie and E. Zurek, *Comput. Phys. Commun.* **183**, 690 (2012).
- ²³P. Niggli, *Kristallographische und Strukturtheoretische Grundbegriffe*, Handbuch der Experimentalphysik Vol. 7 (Akademische Verlagsgesellschaft, 1928).
- ²⁴R. W. Grosse-Kunstleve, N. K. Sauter, and P. D. Adams, *Acta Crystallogr., Sect. A: Found. Crystallogr.* **60**, 1 (2004).
- ²⁵K. J. McGill, M. Asadi, M. T. Karakasheva, L. C. Andrews, H. J. Bernstein, and IUCr, *J. Appl. Cryst.* **47**, 360 (2014).
- ²⁶L. C. Andrews and H. J. Bernstein, e-print [arXiv:1203.5146](https://arxiv.org/abs/1203.5146) (2012).
- ²⁷L. C. Andrews and H. J. Bernstein, *J. Appl. Cryst.* **47**, 346 (2014).
- ²⁸B. P. Van Eijck and J. Kroon, *J. Comput. Chem.* **18**, 1036 (1997).
- ²⁹P. Verwer and F. J. J. Leusen, *Computer simulation to predict possible crystal polymorphs*, Reviews in Computational Chemistry Vol. 12 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998).
- ³⁰H. R. Karfunkel, B. Rohde, F. J. J. Leusen, R. J. Gdanitz, and G. Rihs, *J. Comput. Chem.* **14**, 1125 (1993).
- ³¹R. de Gelder, R. Wehrens, and J. A. Hageman, *J. Comput. Chem.* **22**, 273 (2001).
- ³²A. R. Oganov and M. Valle, *J. Chem. Phys.* **130**, 104504 (2009).
- ³³Y. Wang, J. Lv, L. Zhu, S. Lu, K. Yin, Q. Li, H. Wang, L. Zhang, and Y. Ma, *J. Phys.: Condens. Matter* **27**, 203203 (2015).
- ³⁴P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, *Phys. Rev. B* **28**, 784 (1983).
- ³⁵J. A. Chisholm, S. Motherwell, and IUCr, *J. Appl. Crystallogr.* **38**, 228 (2005).
- ³⁶J. Behler, *J. Chem. Phys.* **134**, 074106 (2011).
- ³⁷A. P. Bartok, R. Kondor, and G. Csanyi, *Phys. Rev. B* **87**, 184115 (2013).
- ³⁸O. A. von Lilienfeld, *Int. J. Quantum Chem.* **113**, 1676 (2013).
- ³⁹P. Gasparotto and M. Ceriotti, *J. Chem. Phys.* **141**, 174110 (2014).
- ⁴⁰L. Yang, S. Dacek, and G. Ceder, *Phys. Rev. B* **90**, 054102 (2014).
- ⁴¹H. W. Kuhn, *Nav. Res. Logist. Q.* **2**, 83 (1955).
- ⁴²J. M. Vásquez-Pérez, G. U. G. Martínez, A. M. Köster, and P. Calaminici, *J. Chem. Phys.* **131**, 124126 (2009).
- ⁴³A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, *J. Chem. Phys.* **139**, 184118 (2013).
- ⁴⁴B. Aradi, B. Hourahine, and T. Frauenheim, *J. Phys. Chem. A* **111**, 5678 (2007).
- ⁴⁵T. J. Lenosky, J. D. Kress, I. Kwon, A. F. Voter, B. Edwards, D. F. Richards, S. Yang, and J. B. Adams, *Phys. Rev. B* **55**, 1528 (1997).
- ⁴⁶P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- ⁴⁷M. Ceriotti, J. More, and D. E. Manolopoulos, *Comput. Phys. Commun.* **185**, 1019 (2014).
- ⁴⁸M. Amsler, S. Botti, M. A. L. Marques, T. J. Lenosky, and S. Goedecker, *Phys. Rev. B* **92**, 014101 (2015).
- ⁴⁹P. M. Morse, *Phys. Rev.* **34**, 57 (1929).
- ⁵⁰D. W. Bassett and P. R. Webber, *Surf. Sci.* **70**, 520 (1978).
- ⁵¹F. H. Allen, *Acta Crystallogr., Sect. B* **58**, 380 (2002).
- ⁵²See <https://github.com/materials/mint> for Mint (materials interface).
- ⁵³G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- ⁵⁴P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).
- ⁵⁵G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- ⁵⁶S. A. Hayward, F. D. Morrison, S. A. T. Redfern, E. K. H. Salje, J. F. Scott, K. S. Knight, S. Tarantino, A. M. Glazer, V. Shuvaeva, P. Daniel, M. Zhang, and M. A. Carpenter, *Phys. Rev. B* **72**, 054110 (2005).
- ⁵⁷A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Taylor & Francis, 2014), Vol. 2.
- ⁵⁸C. Cortes and V. Vapnik, *Mach. Learn.* **20**, 273 (1995).
- ⁵⁹W. McCulloch and W. Pitts, *Bull. Math. Biophys.* **5**, 115 (1943).
- ⁶⁰J. Behler, *Int. J. Quantum Chem.* **115**, 1032 (2015).
- ⁶¹A. P. Bartók and G. Csányi, *Int. J. Quant. Chem.* **115**, 1051–1057 (2015).
- ⁶²M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld, *J. Phys. Chem. Lett.* **6**(16), 3309–3313 (2015).