



PAPER

OPEN ACCESS

RECEIVED
30 April 2020REVISED
19 July 2020ACCEPTED FOR PUBLICATION
24 August 2020PUBLISHED
6 April 2021

Original Content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



An assessment of the structural resolution of various fingerprints commonly used in machine learning

Behnam Parsaeifard^{1,2} , Deb Sankar De^{1,2}, Anders S Christensen³, Felix A Faber³, Emir Kocer⁴ , Sandip De^{2,5,6} , Jörg Behler⁴, O Anatole von Lilienfeld^{2,3} and Stefan Goedecker^{1,2,*}

¹ Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

² National Center for Computational Design and Discovery of Novel Materials (MARVEL), Basel, Switzerland

³ Institute of Physical Chemistry, Department of Chemistry, University of Basel, Klingelbergstr. 80, CH-4056 Basel, Switzerland

⁴ Universität Göttingen, Institut für Physikalische Chemie, Theoretische Chemie, Tammannstr. 6, 37077 Göttingen, Germany

⁵ Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

⁶ Present address: BASF SE, 67056 Ludwigshafen am Rhein, Germany

* Author to whom any correspondence should be addressed.

E-mail: stefan.goedecker@unibas.ch

Keywords: machine learning, symmetry functions, structural fingerprints, atomic descriptor, sensitivity matrix, overlap matrix fingerprint, SOAP

Supplementary material for this article is available [online](#)

Abstract

Atomic environment fingerprints are widely used in computational materials science, from machine learning potentials to the quantification of similarities between atomic configurations. Many approaches to the construction of such fingerprints, also called structural descriptors, have been proposed. In this work, we compare the performance of fingerprints based on the overlap matrix, the smooth overlap of atomic positions, Behler–Parrinello atom-centered symmetry functions, modified Behler–Parrinello symmetry functions used in the ANI-1ccx potential and the Faber–Christensen–Huang–Lilienfeld fingerprint under various aspects. We study their ability to resolve differences in local environments and in particular examine whether there are certain atomic movements that leave the fingerprints exactly or nearly invariant. For this purpose, we introduce a sensitivity matrix whose eigenvalues quantify the effect of atomic displacement modes on the fingerprint. Further, we check whether these displacements correlate with the variation of localized physical quantities such as forces. Finally, we extend our examination to the correlation between molecular fingerprints obtained from the atomic fingerprints and global quantities of entire molecules.

1. Introduction

Materials sciences and chemistry are becoming data-driven sciences [1–9]. Both experimental and theoretical data often contain similar, or duplicate structures which differ only by the noise which is present in any experimental measurements as well as in theoretical structure predictions [10–14]. Such structures can be eliminated based on fingerprint distances. If the structures differ by more than just noise, one frequently wants to quantify their dissimilarity. This is particularly important for applications of supervised machine learning in materials science [15–19], where fingerprints form in most schemes the input for neural networks or other machine learning schemes, but also for eliminating redundant structures, e.g. in the global exploration of potential-energy surfaces. Both, for the detection of duplicate structures as well as for machine learning various atomic environment descriptors have been proposed to date. In the pioneering work of Behler and Parrinello [20, 21] so-called symmetry functions have been introduced to explore the chemical environment of each atom and to form the input to atomic neural networks. Two schemes related to the original Behler–Parrinello atom-centered symmetry functions (ACSF) will also be used here and denoted as modified Behler–Parrinello symmetry functions (MBSF) [22] and Faber–Christensen–Huang–Lilienfeld

(FCHL) [23]. The numerically more efficient discretized version of the FCHL fingerprint [24] is used in our study. Another fingerprint that is widely used in the context of machine learning is the smooth overlap of atomic positions (SOAP) atomic environment descriptor [25]. The last fingerprint that is included in our tests is the overlap matrix (OM) fingerprint [26] that has been used to find duplicate structures in minima hopping based structures predictions [27] and to bias the potential energy landscape to find chemical reaction pathways [28], as well as in machine learning [29, 30]. Many other types of fingerprints have been proposed in the literature to date [31–41]. In the following all these descriptors will be called fingerprints, Cartesian coordinates of atoms in structures, augmented in the crystalline case with the vectors describing the unit cell, form an elementary representation of a configuration or atomic environment. However such Cartesian descriptors are problematic since they are not invariant under translations, rotation and atomic index permutations. So, other descriptors are needed which must be invariant under translations, rotations, and other symmetry operations as well as permutation of identical atoms [20]. All the fingerprints considered in this work are invariant under these operations. The fingerprint distance between two structures can for instance be calculated as the Euclidean norm of the difference between the two fingerprint vectors. In this work, we compare the structural resolution of various fingerprints, i.e. their ability to recognize and quantify differences in atomic environments based on such fingerprint distances.

2. Description of fingerprints used

In this section we give a very brief summary of the fingerprints used in this study. For a complete description of the fingerprints, the reader is referred to the original publications on OM [26], SOAP [25], FCHL [42], ACSF [21], and MBSF [22].

The OM method is inspired by the experimental approach to identify structures. Experimental approaches typically use some spectrum such as a vibrational spectrum or an electronic excitation spectrum to identify structures. Both are related to the eigenvalues of certain matrices. As was shown by Sadeghi *et al* [43] eigenvalues of the Hessian matrix or of the Kohn Sham Hamiltonian matrix are excellent fingerprints for molecular structures, but these matrices are quite expensive to calculate. Fortunately, it turns out that the eigenvalues of a matrix that is extremely fast to calculate, namely the overlap matrix which contains the full structural information are of comparable quality. To calculate the fingerprint of an atom k in the OM scheme, a sphere of radius R_c is centered on it. We place a minimal basis set of four Gaussian type orbitals (GTOs) $G_\nu(\mathbf{r} - \mathbf{R}_i)$ (i.e. radial Gaussians times spherical harmonics) on each atom i in the sphere, namely one s-type GTO ($\nu = 1$), and 3 p-type GTOs ($\nu = 2, 3, 4$) shown by OM[sp]. The width of the radial Gaussian is given by the covalent radius of the element. Then the overlap between all atoms in the sphere is calculated as $S_{i,j}^k = \int G_\nu(\mathbf{r} - \mathbf{R}_i) G_\mu(\mathbf{r} - \mathbf{R}_j) d\mathbf{r}$.

The off-diagonal elements of the overlap matrix decay quite fast with respect to distance from the central atom. This decay is also exploited in the linear electronic structure calculation [44]. Such a fast decay has been shown in a similar context to be advantageous compared to a slower inverse power law decay [45]. Each element $S_{i,j}^k$ of this matrix is then multiplied by two amplitudes $f_c(|\mathbf{R}_k - \mathbf{R}_i|)$ and $f_c(|\mathbf{R}_k - \mathbf{R}_j|)$ where $f_c(r) = (1 - \frac{1}{4}(\frac{r}{w})^2)^2$ is a cutoff function which smoothly tends to zero at $r = 2w = R_c$. So the width w which determines the cutoff radius is the only parameter in this scheme.

The vector \mathbf{F}^k containing all the eigenvalues of this matrix is then the fingerprint of atom k . The fingerprint distance between two atoms I and J is defined to be the Euclidean distance between their fingerprint vectors [43]: $\Delta_{IJ} = |\mathbf{F}^I - \mathbf{F}^J|$.

The above defined fingerprint distance has a discontinuity in the first derivative when two eigenvalues cross. This is an extremely rare event [46] and does not cause problems in most applications. If a completely continuous distance is desired the following post-processing of the eigenvalues can be used to generate a new set \tilde{F} of fingerprints that gives rise to completely continuous fingerprint distances:

$$\tilde{F}_i = \frac{\sum_l F_l \exp\left(-\frac{1}{2}\left(\frac{F_l - F_i}{a}\right)^2\right)}{\sum_l \exp\left(-\frac{1}{2}\left(\frac{F_l - F_i}{a}\right)^2\right)}. \quad (1)$$

In the SOAP scheme, a Gaussian of width σ is centered on each atom within the cutoff distance around the central atom k at position \mathbf{r} . The resulting density of atoms $\rho^k(\mathbf{r}) = \sum_i \exp\left(-\frac{(\mathbf{r} - \mathbf{R}_{ki})^2}{2\sigma^2}\right) \times f_{cut}(|\mathbf{r} - \mathbf{R}_{ki}|)$, multiplied with a cutoff function, which goes smoothly to zero at the cutoff radius over a characteristic width r_δ , is then expanded in terms of orthogonal radial functions $g_n(r)$ and spherical harmonics $Y_{lm}(\theta, \phi)$ as $\rho^k(\mathbf{r}) = \sum_{nlm} c_{nlm}^k g_n(r) Y_{lm}(\theta, \phi)$, where $c_{nlm}^k = \langle g_n Y_{lm} | \rho^k \rangle$. $p_{nn'l}^k = \sqrt{\frac{8\pi^2}{2l+1}} \sum_m c_{nlm}^k (c_{n'l m}^k)^*$ is invariant under rotations and the vector \mathbf{F}^k containing all $p_{nn'l}^k$'s with $n, n' \leq n_{\max}$ and $l \leq l_{\max}$ is the SOAP fingerprint

vector of atom k . The fingerprint distance between atoms I and J can then either be defined as $\Delta_{IJ} = |\mathbf{F}^I - \mathbf{F}^J|$ or $\Delta_{IJ} = (1 - \mathbf{F}^I \cdot \mathbf{F}^J)^{1/2}$. Since the second definition is used in the majority of machine learning applications and since we could not find any difference in preliminary tests, for SOAP we use the second definition of the fingerprint distance. This definition requires the fingerprint vector to be normalized to 1 such that $\sum_i F_i^2 = 1$.

This has the strange side effect that the N fingerprints of a system of N atoms remain identical if N additional atoms are placed on top of the original N atoms. Further, the fingerprint vectors are the same for a dimer where the two atoms are at a very large and zero distance.

The QUIPPY [47] software was used to generate the SOAP fingerprints, with the following parameters: $n_{\max} = l_{\max} = 12$ and $\sigma = 0.5$, $r_s = 4.0 \text{ \AA}$.

The atom-centered symmetry functions (ACSF) proposed by Behler and Parrinello in 2007 have been the first descriptors suitable as input for ML methods for the description of high-dimensional multi-atom systems [20, 21]. They form atomic fingerprint vectors consisting of sets of atom-centered many-body radial and angular functions, which describe the chemical environments of the atoms in the system.

Radial functions are the sum of two-body terms and describe the radial environment of an atom i . They have, for instance, the analytical form $G_i^2 = \sum_j e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij})$.

The angular functions are sums of three-body terms and describe the angular environment of the atom. Two examples are defined below

$$G_i^4 = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{all}} (1 + \lambda \cos(\theta_{ijk}))^\zeta e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \quad (2)$$

$$G_i^5 = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{all}} (1 + \lambda \cos(\theta_{ijk}))^\zeta e^{-\eta(R_{ij}^2 + R_{ik}^2)} f_c(R_{ij}) f_c(R_{ik}) \quad (3)$$

where θ_{ijk} is the angle between \mathbf{R}_{ij} and \mathbf{R}_{ik} and $f_c(r)$ is a smooth cutoff function [21]. The vector \mathbf{F}^i containing all the G_i 's for various values of η , λ , R_s , and ζ is the fingerprint vector of atom i . In the present work, we used 10 radial symmetry functions of type G_2 and 48 angular symmetry functions of type G_4 , which have been generated with the software RuNNer [19, 48]. We have used CUR to find the most relevant symmetry functions [49], as we found that larger sets did not lead to significant improvements.

Isayev *et al* made two modifications to the original Behler–Parrinello angular symmetry functions to obtain MBSFs [22] while retaining the form of the radial functions. These modifications are the addition of a reference angle θ_s to the term $\cos(\theta_{ijk})$ which allows an arbitrary number of shifts in the angular environment and R_s to the exponential term in the angular symmetry functions. The R_s addition allows the angular environment to be considered within radial shells based on the average of the distance from the neighboring atoms [22] similar to the radial shift R_s in the original Behler–Parrinello radial functions. So their modified angular symmetry function is

$$G_i^A = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{all}} (1 + \lambda \cos(\theta_{ijk} - \theta_s))^\zeta e^{-\eta(\frac{R_{ij} + R_{ik}}{2} - R_s)^2} f_c(R_{ij}) f_c(R_{ik}). \quad (4)$$

In this approach, a single η and multiple values of R_s and θ_s are used to generate the fingerprint vector \mathbf{F}^i . We used 32 evenly spaced radial shifting parameters for the radial part, and a total of eight radial and eight angular shifting parameters for the angular part for the MBSF resulting in a total 96 symmetry functions. The QML [50] software package was then used to generate the MBSF fingerprints.

The last fingerprint that we study is the discretized FCHL fingerprint introduced by Faber *et al* [42]. FCHL encodes geometric elemental information into the fingerprint with up to three-body terms included. The 2-body terms consist of sums of log-normal radial functions on the form

$$G^{2\text{-body}} = \xi_2(r_{IJ}) f_{\text{cut}}(r_{IJ}) \frac{1}{R_s \sigma(r_{ij}) \sqrt{2\pi}} \exp\left(-\frac{(\ln R_s - \mu(r_{ij}))^2}{2\sigma(r_{ij})^2}\right) \quad (5)$$

where $f_{\text{cut}}(r_{IJ})$ is a smooth cut-off function, $\xi_2(r_{IJ})$ is a weight function on the form $\frac{1}{r_{ij}^2}$ which serves to put a higher weight in the regression to effects from atoms at closer distances, $\mu(r_{ij}) = \ln\left(\frac{r_{ij}}{\sqrt{1 + \frac{\sigma}{r_{ij}}}}\right)$, and

$\sigma(r_{ij})^2 = 1 + \frac{w}{r_{ij}^2}$. The three-body term in FCHL is the product of a radial part, but uses a (truncated) Fourier expansion for the angular spectrum on the form

$$G^{3\text{-body}} = \xi_3 G_{\text{Radial}}^{3\text{-body}} G_{\text{Angular}}^{3\text{-body}} f_{\text{cut}}(r_{IJ}) f_{\text{cut}}(r_{JK}) f_{\text{cut}}(r_{KI}) \quad (6)$$

where

$$G_{\text{Radial}}^{3\text{-body}} = \sqrt{\frac{\eta_3}{\pi}} \exp\left(-\eta_3 \left(\frac{1}{2}(r_{IJ} + r_{IK}) - R_s\right)^2\right) \quad (7)$$

and $G_{\text{Angular}}^{3\text{-body}}$ contains the below sine and cosine terms with $n = 1$:

$$G_n^{\text{cos}} = \exp\left(-\frac{(\zeta n)^2}{2}\right) (\cos(n\theta_{KI}) - \cos(n(\theta_{KI} + \pi))) \quad (8)$$

$$G_n^{\text{sin}} = \exp\left(-\frac{(\zeta n)^2}{2}\right) (\sin(n\theta_{KI}) - \sin(n(\theta_{KI} + \pi)))$$

where θ_{KI} is the angle between the atoms I, J and K. Furthermore, the three-body symmetry functions are weighted with an Axilrod-Teller-Muto term [51, 52] defined as

$$\xi_3 = c_3 \frac{1 + 3 \cos(\theta_{KI}) \cos(\theta_{JK}) \cos(\theta_{KI})}{(r_{IK} r_{JK} r_{KI})^{N_3}} \quad (10)$$

This again serves to attribute a higher weight to atomic configuration that likely to more strongly interacting [23, 45]. We used the default parameters described in [23] and [24] and the QML [50] software to generate the FCHL fingerprints.

For all fingerprints related to the Behler–Parrinello symmetry functions, i.e. for ACSE, MBSF and FCHL we use the Euclidean norm of the difference of the fingerprint vectors as the fingerprint distance.

For a fair comparison we have chosen for all fingerprints the same cutoff radius, namely 6.0 Å. This or very similar values were used in numerous studies [22, 24, 48, 53]. So all the methods see exactly the same environment and could therefore in principle encode the same information in their resulting fingerprint vectors. With this choice of parameters, the length of the fingerprints was 240 for OM, 1015 for SOAP, 58 for ACSE, 96 for MBSF and 64 for FCHL.

3. Results

In this section we will introduce some criteria to assess the performance of the various fingerprints. First, we derive a formalism that allows to check the behavior of the different fingerprints under infinitesimal changes of the atomic coordinates. We show that there is a matrix, that we baptize sensitivity matrix, that describes this behavior. In particular, the displacement modes of this matrix that belong to zero eigenvalues give rise to constant fingerprints for movements along these modes and indicate therefore a failure of the fingerprint to detect geometry changes. Next we will compare for a test set the distances obtained by different fingerprints. This test helps us to find cases where a certain fingerprint can not recognize differences between different chemical environments. In addition we will correlate in both cases changes in fingerprint distances with changes of physical quantities such as forces, energies and densities of states.

3.1. Behavior of fingerprints under infinitesimal displacements

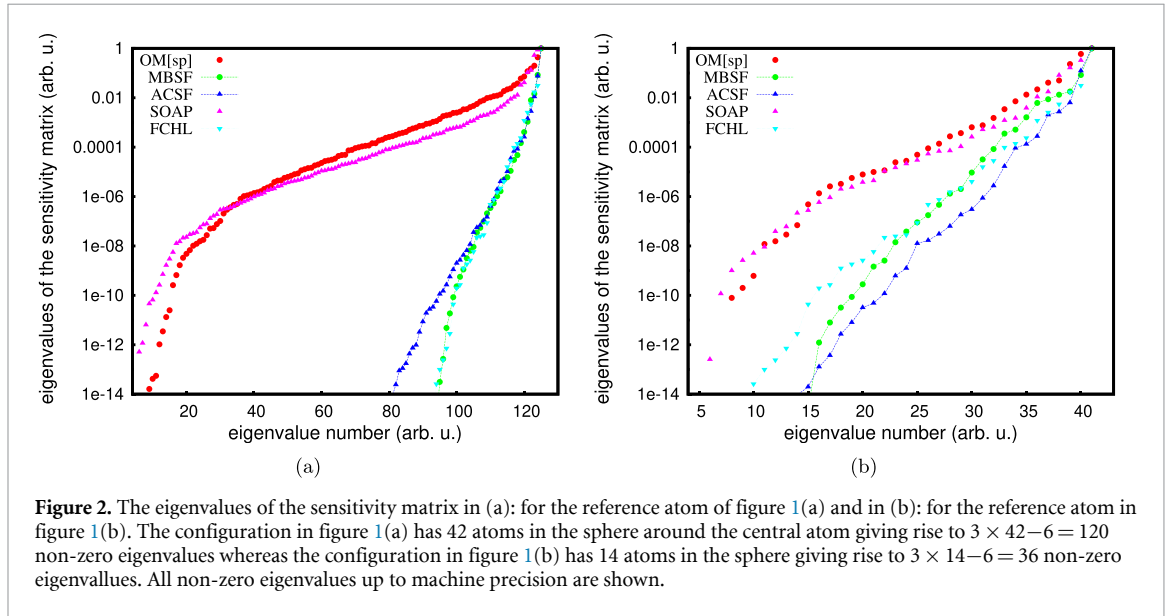
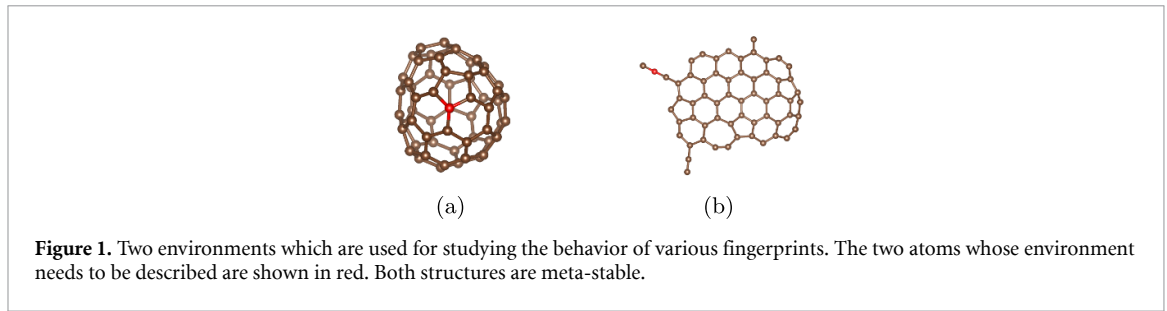
To study the evolution of fingerprint distances under small displacements, we consider the change of the squared fingerprint distance up to second order in a Taylor expansion around a reference configuration. Denoting the fingerprint of the reference configuration by \mathbf{F}^0 and the fingerprint of a configuration displaced by $\Delta\mathbf{R}$ by $\mathbf{F}(\mathbf{R})$ we get

$$(\mathbf{F}(\mathbf{R}) - \mathbf{F}^0)^2 = \sum_{\alpha, \beta} \Delta R_\alpha \left(\sum_i g_{i, \alpha} g_{i, \beta} \right) \Delta R_\beta \quad (11)$$

Table 1. The parameters used for each fingerprint.

Fingerprint type	Name	Number	Value	Unit	Description
MBSF	$R_s(G^R)$	32	a	Å	Two-body radial bins
	$R_s(G^A)$	8	b	Å	Three-body radial bins
	θ_s	8	c		Three-body angular bins
	r_{cut}		6.0	Å	Radial cutoff (two-body)
	a_{cut}		6.0	Å	Radial cutoff (three-body)
	$\eta(G^R)$		1.0	Å ⁻²	Two-body width parameter
	$\eta(G^A)$		1.0	Å ⁻²	Three-body width parameter
	ζ		1.0		Angular exponent
	n_{R2}	24	d	Å	Two-body radial bins
	n_{R3}	20	e	Å	Sin three-body radial bins
	n_{R3}	20	f	Å	Cos three-body radial bins
		w	0.32	Å ²	Two-body width parameter
		η_2	2.7	Å ⁻²	Three-body width parameter
		N_2	1.8		Two-body scaling exponent
	N_3	0.57		Three-body scaling exponent	
SOAP	c_3		13.4	Å ^{N₃}	Three body-weight
	ζ		π		Angular exponent
	r_{cut}		6.0	Å	Radial cutoff (two-body)
	a_{cut}		6.0	Å	Radial cutoff (three-body)
	σ		0.5	Å	atom sigma
	l_{max}		12		
	n_{max}		12		
	r_δ		4.0	Å	Characteristic decay length
	R_c		6.0	Å	Cutoff radius
	w		3.0	Å	Gaussian width
	$R_c = 2w$		6.0	Å	Cutoff radius
	s-type orbitals		s		
	p-type orbitals		p_x, p_y, p_z		
	ACSF	$\eta(G_2)$	10	g	Å ⁻²
$\eta(G_4)$		6	h	Å ⁻²	Three-body width parameter
λ		2	-1,1		Angular exponent
ζ		4	1,2,4,16		Cutoff radius
R_c			6.0	Å	

^a [0.8, 0.968, 1.135, 1.303, 1.471, 1.639, 1.806, 1.974, 2.142, 2.31, 2.477, 2.645, 2.813, 2.981, 3.148, 3.316, 3.484, 3.652, 3.819, 3.987, 4.155, 4.323, 4.490, 4.658, 4.826, 4.994, 5.161, 5.329, 5.497, 5.665, 5.832, 6.0].^b [0.8, 1.543, 2.286, 3.0286, 3.771, 4.514, 5.257, 6.0].^c [0.0, 0.449, 0.898, 1.346, 1.795, 2.244, 2.693, 3.142].^d [0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5, 3.75, 4.0, 4.25, 4.5, 4.75, 5.0, 5.25, 5.5, 5.75, 6.0].^e [0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 3.3, 3.6, 3.9, 4.2, 4.5, 4.8, 5.1, 5.4, 5.7, 6.0].^f [0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 3.3, 3.6, 3.9, 4.2, 4.5, 4.8, 5.1, 5.4, 5.7, 6.0].^g [0.003, 0.018, 0.036, 0.054, 0.071, 0.089, 0.125, 0.161, 0.214, 0.285].^h [0.0, 0.004, 0.018, 0.071, 0.214, 0.285].



where $g_{i,\alpha}$ is the gradient of the i th component of the fingerprint vector with respect to the three Cartesian components α (x , y , and z) of the position vector \mathbf{R} , i.e.

$$g_{i,\alpha} = \left. \frac{\partial F_i}{\partial R_\alpha} \right|_{\mathbf{R}=\mathbf{R}_0}. \quad (12)$$

In taking this derivative we have to consider only the atomic positions within the sphere around the central atom because by construction atoms outside the sphere have no influence on the fingerprint. We call this matrix $\sum_i g_{i,\alpha} g_{i,\beta}$ sensitivity matrix. It has the dimension $3N \times 3N$ where N is the number of atoms within the cutoff sphere around the reference atom. In the following, we will examine its eigenvalues and eigenvectors. To allow a meaningful comparison of the fingerprints obtained by different methods we have scaled all the eigenvalues such that the largest eigenvalue is one. Since the fingerprint is invariant under a uniform translation and rotation of all the atoms in the sphere, the sensitivity matrix has always at least 6 zero eigenvalues. More than 6 zero eigenvalues indicate that there are other displacement modes which will leave the fingerprint invariant. This is highly problematic since it indicates that one can generate different atomic environments which will not change the fingerprint. By calculating iteratively these zero eigenvalue displacement modes and then moving the system by an infinitesimal amount along those consecutive modes one can construct from a sequence of infinitesimal small displacements a finite displacement which will leave the fingerprint invariant [43]. Equally problematic are eigenvalues that are very small. In this case the fingerprint variation will not exactly be zero, but will be extremely small. We now study the sensitivity matrix for the two configurations of 60 carbon atoms shown in figure 1. An analogous analysis will be presented in the supplementary information (stacks.iop.org/MLST/2/015018/mmedia) for two more structures.

In figure 1(a) the reference atom forms three bonds with its three nearest neighbors and is surrounded by one pentagon and two hexagons, while in figure 1(b) the atom of interest resides on a chain and has fewer neighbors compared to the atom in figure 1(a).

In figure 2(a) we show the eigenvalues of the sensitivity matrix of configuration figure 1(a) for all the fingerprints examined in our study. The eigenvalues of the sensitivity matrix for ACSF, MBSF, and FCHL decrease much more rapidly to zero than the eigenvalues of SOAP and OM[sp]. This means that in ACSF, MBSF, and FCHL, there exist only a few modes that have a strong influence on the fingerprint. It is also of

interest to look at the associated modes shown in figure 3 and 4. In the context of machine learning one might hope that the modes that are associated to the largest eigenvalues and will therefore lead to the strongest variation in the fingerprint will also lead to the largest variation of physical properties such as forces [21]. Since movements of atoms close to the central atom will in general lead to a strong variation of the environment of the reference atom, this means that modes belonging to large eigenvalues should be localized around the central atom. The movement that will lead to the strongest variation of the energy for the configurations shown in figure 3 is clearly a bond stretching mode where the three neighboring atoms either move towards the central atom or away from it (figure 3(a)–(c)). Then follows a movement where two bonds of the central are compressed and one is stretched and finally an out of plane movement of the central atom. These three modes are exactly the modes associated to the three largest eigenvalues of the OM sensitivity matrix. SOAP and FCHL also describe the physically important modes with reasonably large eigenvalues. In the ACSF and MBSF fingerprints however only an out of plane mode has a reasonably large eigenvalue. The modes belonging to the few largest eigenvalues are always localized on the reference atom and a few surrounding atoms. As the eigenvalues become smaller the modes should get more delocalized, and this is indeed true in most cases. There are however some exceptions such as the modes of the ACSF shown in the panels (l) of figure 3 and figure 4, the modes of MBSF in the panels (p) of figure 3 and 4 and a mode of SOAP shown panel (h) of figure 4.

This discussion, which was based on some physical insight into which modes are important, can also be made more quantitative. We do this by plotting the change in the force acting on the central atom when the system is moved along the different modes against the eigenvalue of this mode. This is shown in figure 5. A clear correlation is found for OM and SOAP, while for ACSF, MBSF and FCHL the correlation is substantially weaker, with FCHL showing at least the correct trend. This means that movements along modes associated to large and small eigenvalues have almost the same influence on the force on the reference atom.

Even though the environment of figure 1(b) is quite different, the performance of the fingerprints is quite similar. Only OM and SOAP detect the physically important modes (figure 2(b)), i.e. assign a large eigenvalue to these modes. They are also the only two fingerprints that give a good correlation between the eigenvalues and the change in the force (figure 5).

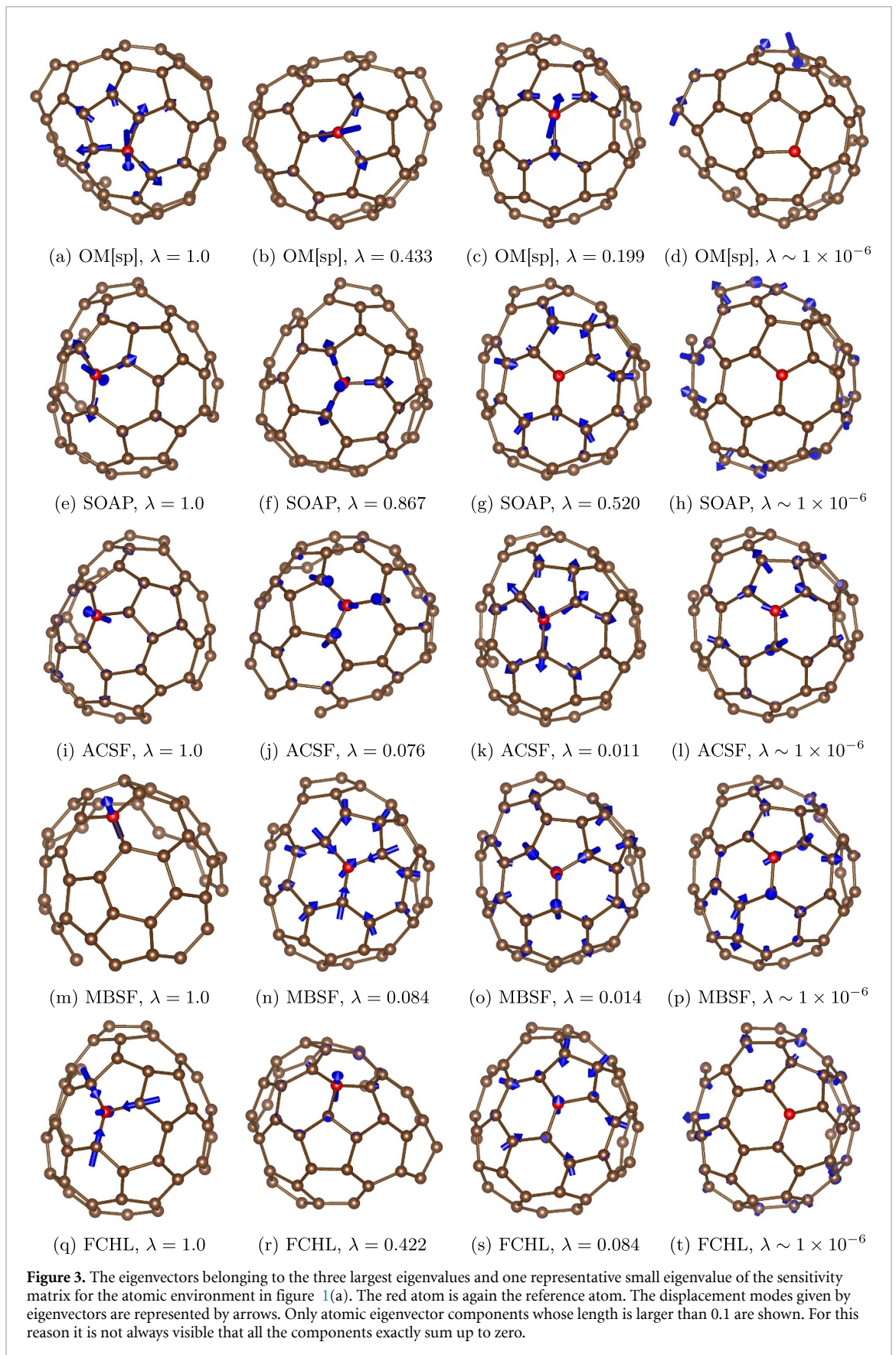
While SOAP is performing well in our test case where many atoms are contained in the sphere, it was recently shown [54] that for a methane molecule there are movements that leave the SOAP fingerprint of the carbon invariant. We detected the same deficiency also for ACSF, MBSF and FCHL. We have also tested the OM fingerprint for these configurations and did not find any small or even zero eigenvalues. This is to be expected since the OM fingerprint is based on a matrix diagonalization scheme that is similar to the diagonalization of the Hamiltonian matrix in a quantum-mechanical calculation. Hence the scheme is not restricted to the information obtained only from the radial and angular distribution of the atoms in the sphere.

3.2. Correlation of fingerprint distances

In this section, we are going to compare the resolution power of different fingerprints, i.e. their numerical sensitivity to small dissimilarities between atomic environments. To perform the tests we have generated a set of 1000 C₆₀ structures using minima hopping [27] coupled to DFTB [58]. In this way we have obtained 60 × 1000 environments arising from a large variety of structural motifs such as chains, planar structures and cages. We will in the following correlate all the $\frac{60000 \times (60000-1)}{2}$ pairwise atomic fingerprint distances obtained from different fingerprint types. Obviously large fingerprint distances should be obtained for environments that are quite different whereas small distances correspond to similar environments. Since the absolute value of a fingerprint distance is arbitrary, we scale all our fingerprint distances such that a distance of one corresponds to the noise level. We define the noise level as the fingerprint distance between identical structures, whose atoms were randomly displaced by an amount of up to ±0.02 Å.

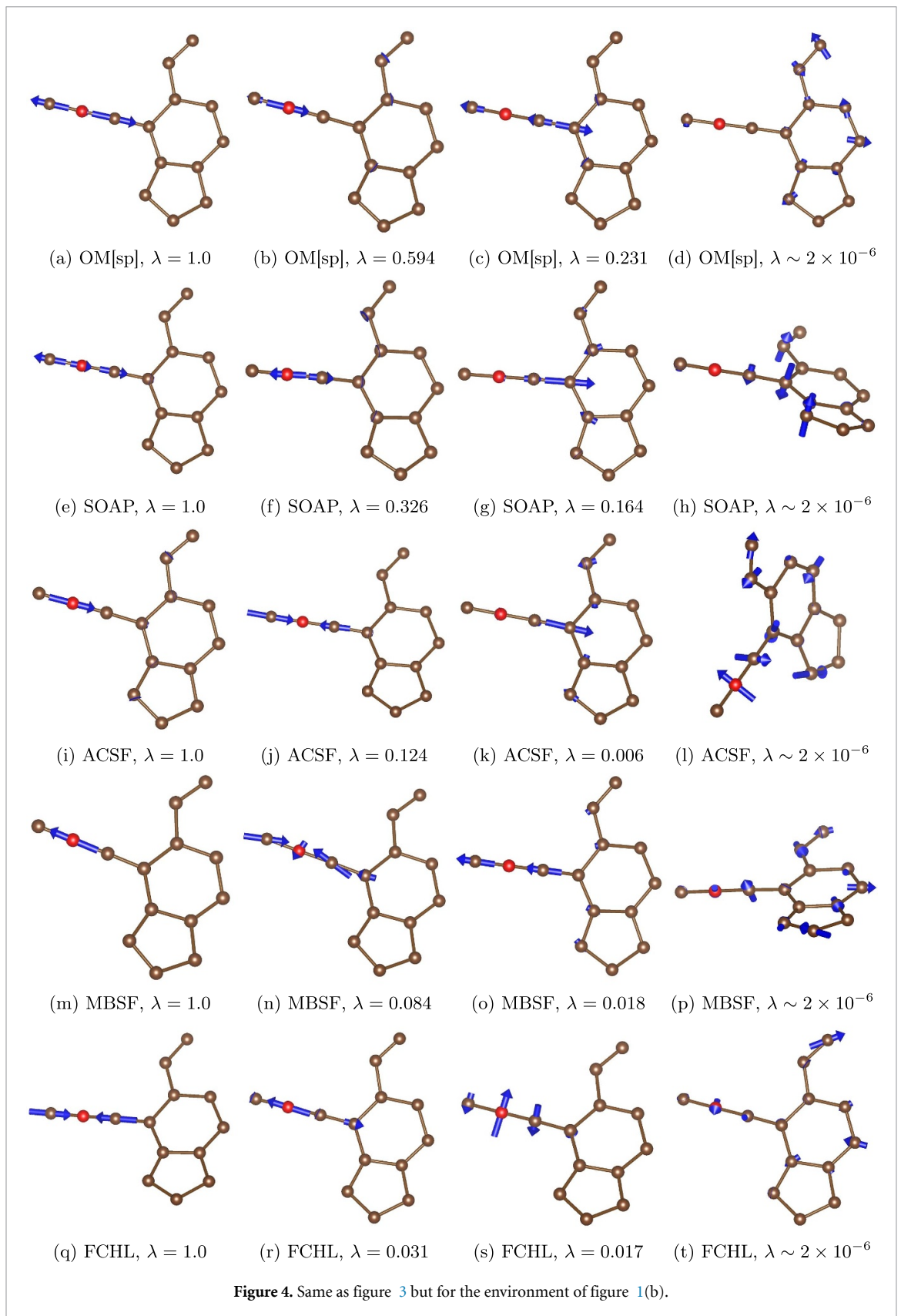
Since the number of environment pairs is huge we would not be able to resolve each pair in a simple correlation plot where we would plot the fingerprint distances $\Delta_{I,J}^A$ according to fingerprint A versus the distance $\Delta_{I,J}^B$ according to fingerprint B. However this large number of data allows us to generate a histogram. This histogram tells us how many environments have fingerprint distances $\Delta_{I,J}^A$ and $\Delta_{I,J}^B$. These two distances are plotted along the x and y axis and the height of the bins of the histogram is indicated by the color in this plot shown in figure 6.

As can be seen in figure 6, in most cases, the intensity is peaked around the diagonal which implies that both fingerprints agree on the degree of similarity or dissimilarity between the environment pairs. It can not be expected that all the points lie directly on the diagonal since different fingerprints weight different types of similarity or dissimilarity in different ways. There is however a problem if a point lies exactly on or very close to the x or y axis which means that the Δ is either zero or very small. This means that one fingerprint categorizes this pair of environments as identical whereas the other fingerprint can detect differences, i.e. it is

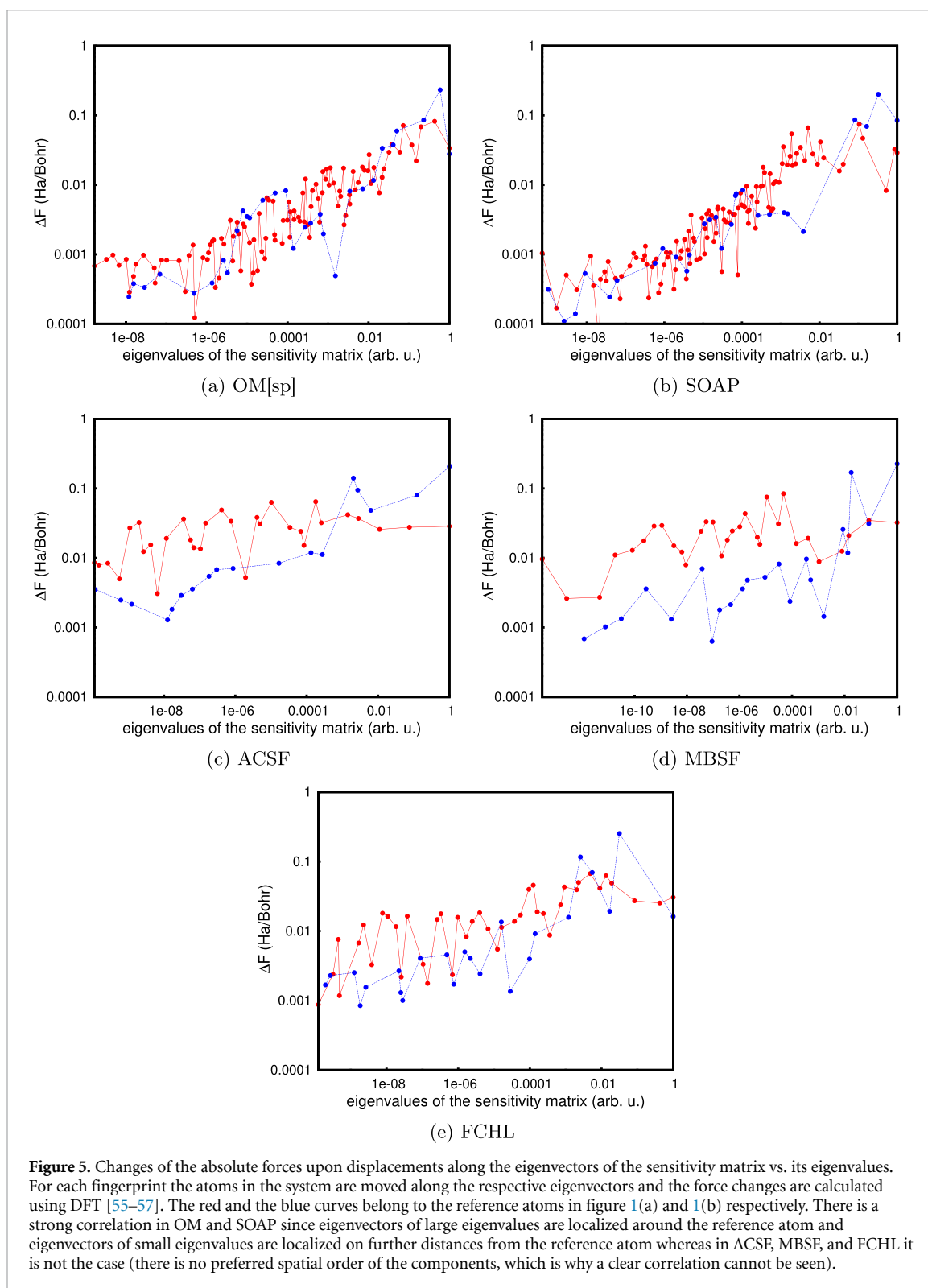


Δ value is large. In table 2 we show several pairs of environments that correspond to such problematic points in the correlation plot.

In table 2(a) we show the two most distinct environments in the data according to OM[sp]. One environment is at the end of a chain and the other is 3-fold coordinated. So OM recognizes the atoms with



the highest and lowest coordination number found in this data set as being the most different. The fingerprint distance is $\Delta^{OM[sp]} = 317$. Diamond-like environments were not in our MH generated data set. Due to their large number of surface dangling bonds such structures are considerably higher in energy than the structures arising from sp² and sp¹ hybridized carbon atoms. However, when we add by hand such a diamond derived cluster, OM predicts the central 4-fold coordinated atom of this cluster together with the previous atom at the end of the chain as the most distinct atoms. So again it classifies the two environments



with the highest and lowest coordination as the most different ones. ACSF, SOAP, FCHL, and MBSF predict the environments in table 2(b) and (c) to be the most distinct environments in the data. The fingerprint distances are $\Delta^{SOAP} = 214$, $\Delta^{FCHL} = 315$, $\Delta^{ACSF} = 822$, and $\Delta^{MBSF} = 1224$ respectively. This is not in agreement with our basic chemical concepts of what structural differences are important. According to these concepts the coordination number is the most important quantity in the chemistry of carbon, since it is related to the hybridization state. When adding the four fold coordinated carbon from the diamond-like cluster, then ACSF, MBSF and FCHL correctly identify this fourfold coordinated environment and the one from the end of the chain as the most different ones. The assignment of the largest fingerprint distance in

SOAP is however unchanged by the addition of this fourfold coordinated environment. So the assignments of the symmetry-function-related fingerprints are at least partly compatible with chemical concepts, whereas for SOAP this is not the case. It is unclear whether a fingerprint that is compatible with chemical concepts gives better performance in machine learning schemes. By choosing a shorter r_s in the case of SOAP and shorter cutoff radii for ACSF-related fingerprints, it is however expected that the immediate environment gets more weight and that then the other fingerprints can also better distinguish different coordinations. We note that also for the cutoff employed in the present work individual components of the fingerprint vectors in ACSF-related fingerprints adopt different values for varying coordinations, while this effect is much less visible in the combined fingerprint distances. In the following we look at the correlation plots of fingerprint distances obtained with different fingerprints. We check whether some fingerprints can not recognize structural differences.

Figure 6(a) shows the resolution plot between the OM and SOAP fingerprints. In this case, both OM[sp] and SOAP fingerprints agree quite well on similarities and dissimilarities between the environments.

Figure 6(b) shows the resolution intensity plot between OM[sp] and ACSF. There exist some points with significant values on the OM[sp] axis. These points represent different environments where ACSF cannot resolve the differences between them since the ACSF FP distance is close to zero. In table 2(d) we show two atomic environments which are obviously quite different, but whose ACSF distance is very small. The two environments are very different since the central atom in the left panel makes one bond with its nearest neighbor while the central atom in the right panel is two-fold coordinated. In table 2(e) we also show another example where the difference vectors of the ACSF are rather small.

Figure 6(c) shows the correlation intensity plot between OM[sp] and FCHL. There is not any point on the axes with significant values. So both fingerprints agree on similarities.

Figure 6(d) shows the correlation plot between OM[sp] and MBSF. In table 2(f) and (g) we show two examples in which the MBSF does not recognize the differences between the two environments. In table 2(f) left, the central environment is in the middle of the chain and has two nearest neighbors while on right, it is at the end of the chain and has one nearest neighbor. In table 2(g) left, the reference atom is again at the end of a chain while on right it is three-fold coordinated.

Figure 6(e) shows the correlation intensity between SOAP and ACSF. We can also see problematic points where the fingerprint distance is very small according to ACSF but not according to SOAP. In table 2(h) we show an example of two different environments where ACSF predicts a very small fingerprint distance. Although the central atom in both cases have one nearest neighbour, but the second and third shells are different. Table 3(a) shows another example in which ACSF does not recognize the differences in the local environment.

The correlation intensity between SOAP and FCHL is shown in figure 6(f). There is not any point on either axes with significant values and both fingerprints therefore agree on similarities and differences between environments.

Correlation intensity between SOAP and the MBSF is shown in figure 6(g). There exist again some problematic points on the SOAP axis which indicates that there are some different environments that MBSF predicts to be the same or very similar. In table 3(b) and (c) we show two such examples.

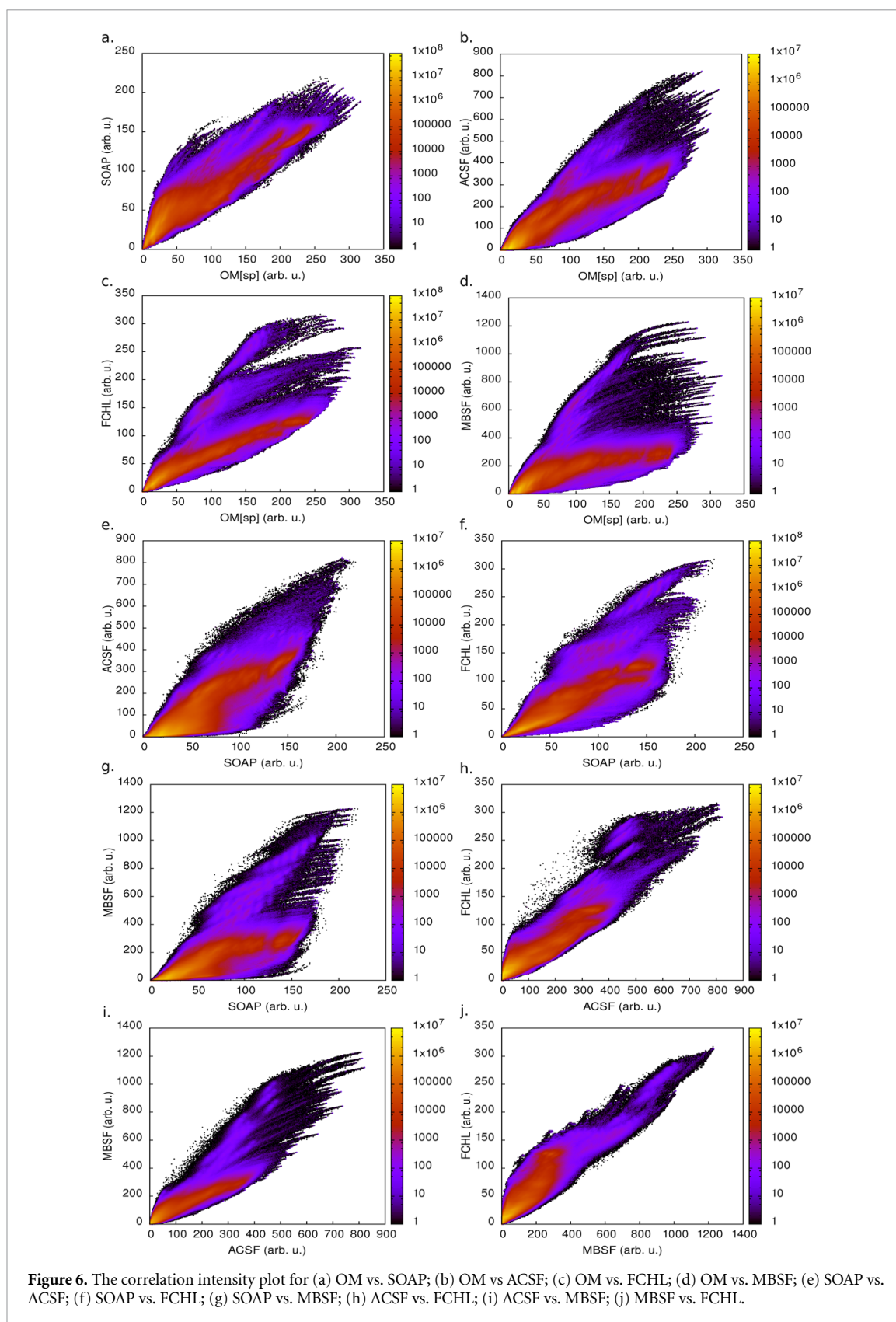
The correlation intensity between ACSF and FCHL is shown in figure 6(h). There are also some points lying on and very close to the FCHL axis (points with fingerprint distances up to 50 near the FCHL axis). These points indicate environments which are different according to FCHL and very similar according to ACSF. In table 3(d) and (e) we show two such examples where the two environments are different while fingerprint distance according to ACSF is very small. The reference atom is in one case two-fold coordinated while it is three-fold coordinated in the other case.

In figure 6(i) we show the correlation intensity between ACSF and the MBSF. The two fingerprints agree on most similarities and there are no points on axes with significant values.

As a last illustration we show the correlation plot between the MBSF and FCHL in figure 6(j). In table 3(f), (g), and (h) we show examples where the MBSF does not recognize differences between the local environments and predicts very small fingerprint distances compared to FCHL. To summarize, our analysis of the eigen modes of the sensitivity matrix shows that ACSF, MBSF, and partly FCHL are quite insensitive to certain displacements of the neighbouring atoms and have in this way an unsatisfactory structural resolution power. SOAP and OM perform significantly better in this respect.

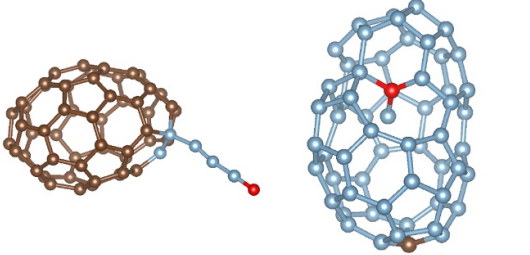
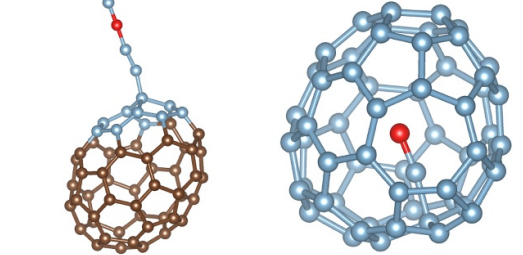
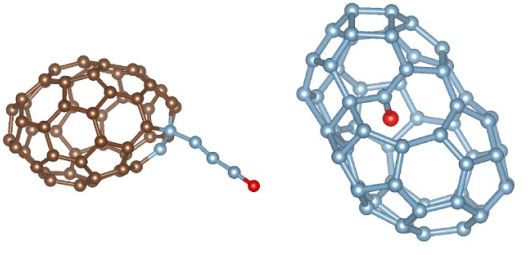
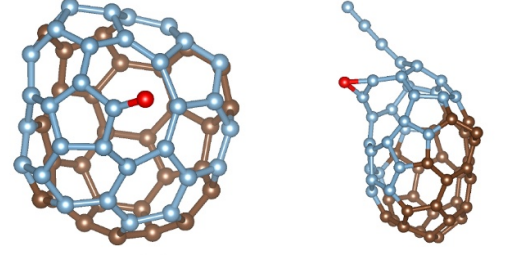
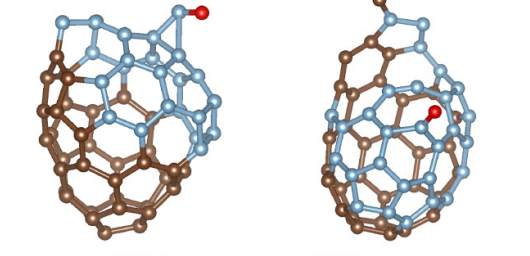
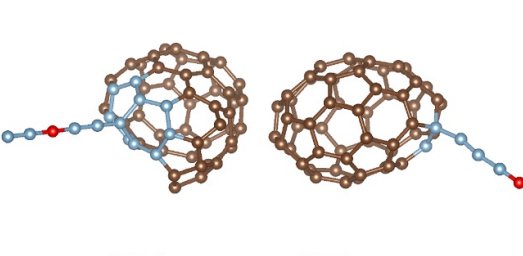
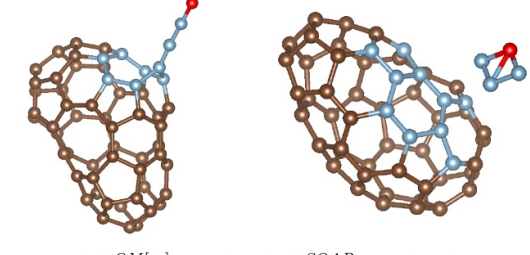
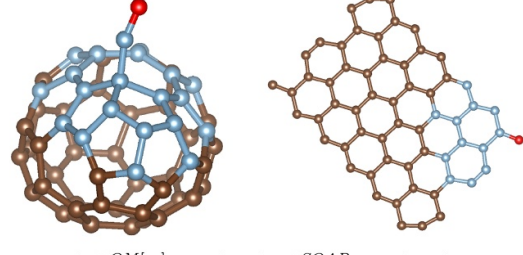
4. Correlation between molecular fingerprints and global physical properties

According to our analysis reported above several fingerprints that are widely and successfully used for instance in machine learning schemes are apparently sometimes unable to distinguish between different



chemical environments. One would thus expect that this gives rise to errors in the prediction of physical properties. One typical application that in principle could be affected is the development of machine learning potentials [59], which predict the energy and forces as a function of the atomic positions. Most of these ML potentials rely on a construction of the total energy as a sum of environment-dependent atomic energies [20, 35, 60] and thus should be sensitive to deficiencies in the discrimination of these environments. In this section we will discuss possible implications of our findings with respect to such applications of ML.

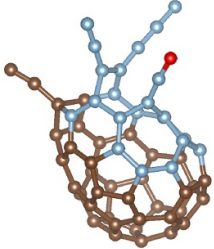
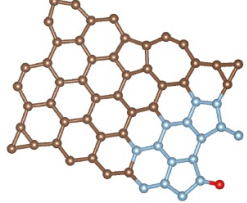
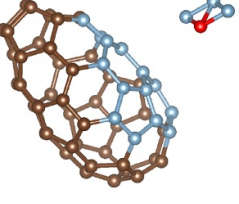
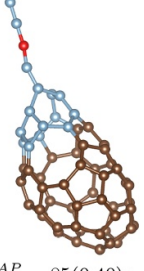
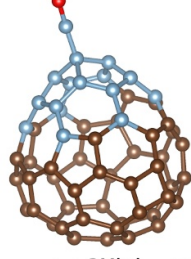
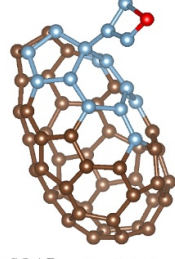
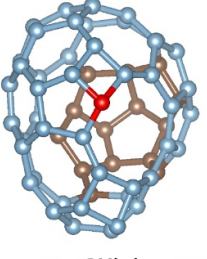
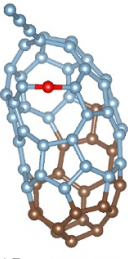
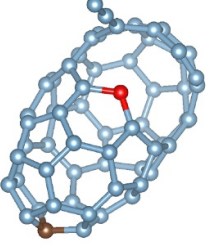
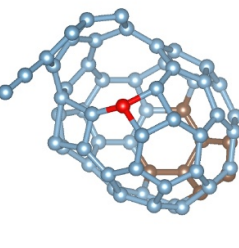
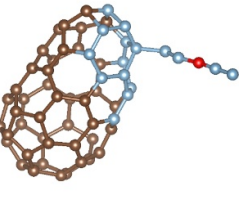
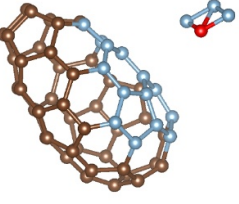
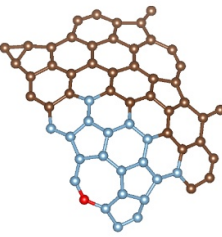
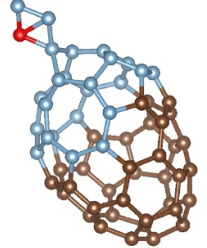
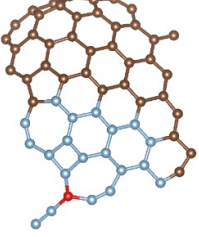
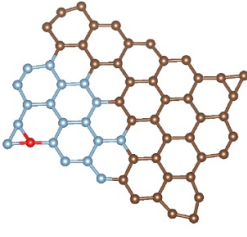
Table 2. The most distinct atomic environments according to (a) OM; (b) SOAP, FCHL, and MBSF; and (c) ACSF. The rest of the panels are problematic atomic environments in which one fingerprint predicts a large fingerprint distance whereas the other fingerprint predicts a small one. The first number is the absolute fingerprint distance whereas the number in parenthesis is the percentage of the largest distance. The reference atom whose environment we want to describe, is red colored, the atoms in the vicinity of the reference atom are blue colored and the remaining atoms in the structure which are outside of the cutoff sphere and do not affect the fingerprint are shown in brown.

 <p>a) $\Delta^{OM[sp]} = 317(1.0)$; $\Delta^{SOAP} = 189(0.88)$; $\Delta^{ACSF} = 738(0.90)$; $\Delta^{FCHL} = 256(0.82)$; $\Delta^{MBSF} = 844(0.69)$</p>	 <p>b) $\Delta^{OM[sp]} = 251(0.79)$; $\Delta^{SOAP} = 214(1.0)$; $\Delta^{ACSF} = 802(0.98)$; $\Delta^{FCHL} = 315(1.0)$; $\Delta^{MBSF} = 1224(1.0)$</p>
 <p>c) $\Delta^{OM[sp]} = 292(0.92)$; $\Delta^{SOAP} = 206(0.96)$; $\Delta^{ACSF} = 822(1.0)$; $\Delta^{FCHL} = 292(0.93)$; $\Delta^{MBSF} = 1119(0.91)$</p>	 <p>d) $\Delta^{OM[sp]} = 38(0.12)$; $\Delta^{SOAP} = 67(0.32)$; $\Delta^{ACSF} = 3(0.0)$; $\Delta^{FCHL} = 33(0.11)$; $\Delta^{MBSF} = 5(0.0)$</p>
 <p>e) $\Delta^{OM[sp]} = 34(0.11)$; $\Delta^{SOAP} = 43(0.2)$; $\Delta^{ACSF} = 2(0.0)$; $\Delta^{FCHL} = 17(0.05)$; $\Delta^{MBSF} = 8(0.01)$</p>	 <p>f) $\Delta^{OM[sp]} = 79(0.25)$; $\Delta^{SOAP} = 66(0.31)$; $\Delta^{ACSF} = 34(0.04)$; $\Delta^{FCHL} = 46(0.15)$; $\Delta^{MBSF} = 13(0.01)$</p>
 <p>g) $\Delta^{OM[sp]} = 78(0.25)$; $\Delta^{SOAP} = 79(0.37)$; $\Delta^{ACSF} = 22(0.03)$; $\Delta^{FCHL} = 60(0.19)$; $\Delta^{MBSF} = 11(0.01)$</p>	 <p>h) $\Delta^{OM[sp]} = 37(0.12)$; $\Delta^{SOAP} = 74(0.35)$; $\Delta^{ACSF} = 7(0.01)$; $\Delta^{FCHL} = 23(0.07)$; $\Delta^{MBSF} = 14(0.01)$</p>

For our investigation, we need to distinguish between local and global properties. While local properties like forces are observables that can be uniquely assigned to individual atoms, the total energy of the system is not an observable, and there is no physically unique definition of atomic energies. While ML potentials are supposed to represent both, forces and energies, with high accuracy and consistently, their analysis requires different approaches.

We will now investigate the role of the total energy as a global property. It has been shown for instance for the distribution of atomic energies within extended systems [61], that atomic energies determined by ML can compensate each other to yield the correct total energy if there is enough flexibility in the system. For many systems this flexibility can be reduced by adding constraints on the energy distribution in form of different stoichiometries [61], but in general there is no way to extract unique atomic energies for arbitrary systems

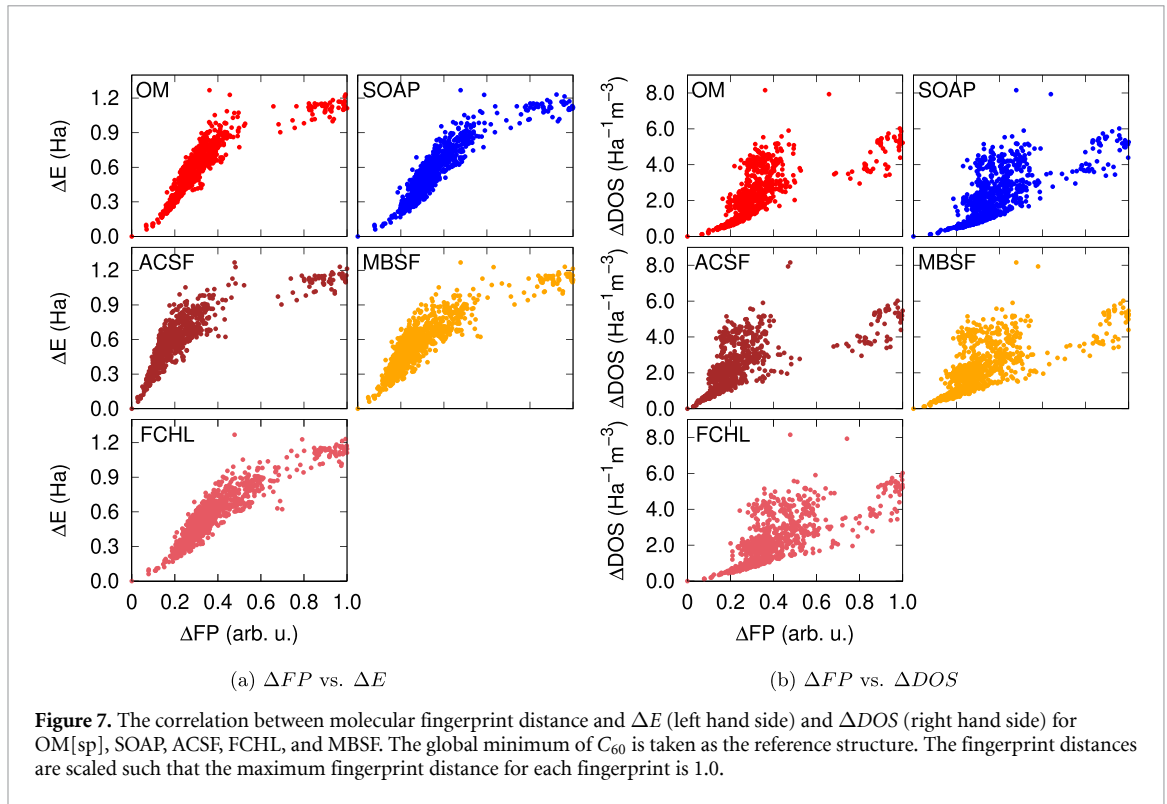
Table 3. Further problematic environments.

  <p>a) $\Delta^{OM[sp]} = 34(0.11)$; $\Delta^{SOAP} = 74(0.35)$; $\Delta^{ACSF} = 6(0.01)$; $\Delta^{FCHL} = 18(0.06)$; $\Delta^{MBSF} = 18(0.01)$</p>	  <p>b) $\Delta^{OM[sp]} = 55(0.18)$; $\Delta^{SOAP} = 85(0.40)$; $\Delta^{ACSF} = 37(0.05)$; $\Delta^{FCHL} = 35(0.11)$; $\Delta^{MBSF} = 7(0.01)$</p>
  <p>c) $\Delta^{OM[sp]} = 37(0.12)$; $\Delta^{SOAP} = 73(0.34)$; $\Delta^{ACSF} = 16(0.02)$; $\Delta^{FCHL} = 26(0.08)$; $\Delta^{MBSF} = 7(0.01)$</p>	  <p>d) $\Delta^{OM[sp]} = 46(0.15)$; $\Delta^{SOAP} = 60(0.28)$; $\Delta^{ACSF} = 8(0.01)$; $\Delta^{FCHL} = 46(0.15)$; $\Delta^{MBSF} = 25(0.02)$</p>
  <p>e) $\Delta^{OM[sp]} = 36(0.12)$; $\Delta^{SOAP} = 50(0.24)$; $\Delta^{ACSF} = 8(0.01)$; $\Delta^{FCHL} = 44(0.14)$; $\Delta^{MBSF} = 29(0.02)$</p>	  <p>f) $\Delta^{OM[sp]} = 52(0.16)$; $\Delta^{SOAP} = 76(0.36)$; $\Delta^{ACSF} = 28(0.4)$; $\Delta^{FCHL} = 33(0.11)$; $\Delta^{MBSF} = 5(0.0)$</p>
  <p>g) $\Delta^{OM[sp]} = 34(0.11)$; $\Delta^{SOAP} = 43(0.20)$; $\Delta^{ACSF} = 14(0.02)$; $\Delta^{FCHL} = 31(0.10)$; $\Delta^{MBSF} = 5(0.0)$</p>	  <p>h) $\Delta^{OM[sp]} = 21(0.07)$; $\Delta^{SOAP} = 34(0.16)$; $\Delta^{ACSF} = 15(0.02)$; $\Delta^{FCHL} = 29(0.09)$; $\Delta^{MBSF} = 5(0.0)$</p>

using ML. This finding is independent of the ability of the fingerprint vectors to distinguish chemically inequivalent atomic environments.

Here, we now go one step further and investigate if even a few ‘wrong’ environment descriptions, which cannot resolve some structural differences as reported above, might be tolerable as the total energy could still be well represented due to some error cancellation. To check the correlation of global properties with various atomic fingerprints we first have to construct a global, i.e. molecular fingerprint from our local atomic fingerprints. We do this by finding the optimal matching between all the atomic environments in the two structures [26], i.e. the matching that minimizes the root-mean-square distance (RMSD) between the two molecules [43]. In this approach the fingerprint distance between two molecules p and q is defined as

$$\Delta^{p, q} = \min_p \left(\sum_i^N |\mathbf{F}_p^i - \mathbf{F}_q^{P(i)}|^2 \right)^{1/2} \quad (13)$$



where \mathbf{F}_p^i is the fingerprint vector for atom i in configuration p and $\mathbf{F}_q^{P(i)}$ is the fingerprint of the best matching atom $P(i)$ in configuration q . The permutation function P which gives the best overall match is found with the Hungarian algorithm [62] in polynomial time. We note, however, that this construction of a global molecular fingerprint is different from the procedure that is usually applied in the construction of ML potentials, and here we use it primarily as a tool to detect correlations between global properties and the entire structure of a system.

While the atomic fingerprint distance shows how different two atomic environments are, the molecular fingerprint distance indicates the difference between two entire molecules. In the next step, we calculate the correlation between molecular fingerprints and two global properties, namely the total energy and the density of states (DOS). If two molecules have different energies or DOSs, they have to be different and so the fingerprint distance should be non-zero. On the other hand, if two molecules have nearly the same energies or DOS they can be similar (in case of degeneracy) or different. So the fingerprint distance does not need to be necessarily non-zero.

The density of states for molecule p , $D_p(\epsilon)$ is

$$D_p(\epsilon) = \sum_i \delta(\epsilon - \epsilon_i^p) \quad (14)$$

where ϵ_i^p are the Kohn-Sham eigenvalues for molecule p . We replace $\delta(\epsilon - \epsilon_i^p)$ with $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\epsilon - \epsilon_i^p)^2}{2\sigma^2}\right)$ with σ some smearing parameter. We define the difference between the density of states to be

$$\Delta DOS_{p,q} = \sqrt{\int d\epsilon (D_p(\epsilon) - D_q(\epsilon))^2}. \quad (15)$$

Taking advantage of the properties of Gaussian functions, we can calculate the integral analytically. Hence, $\Delta DOS_{p,q}$ can be calculated as

$$\Delta DOS_{p,q} = \sqrt{\sum_{i,j} \left(e^{-(\epsilon_i^p - \epsilon_j^p)^2/4\sigma^2} + e^{-(\epsilon_i^q - \epsilon_j^q)^2/4\sigma^2} - e^{-(\epsilon_i^p - \epsilon_j^q)^2/4\sigma^2} - e^{-(\epsilon_i^q - \epsilon_j^p)^2/4\sigma^2} \right)}. \quad (16)$$

We chose $\sigma = 0.01$ Ha in this work. The molecule with the lowest energy is taken as reference structure and fingerprint distances and energy differences are calculated with respect to it. In figure 7 we see the

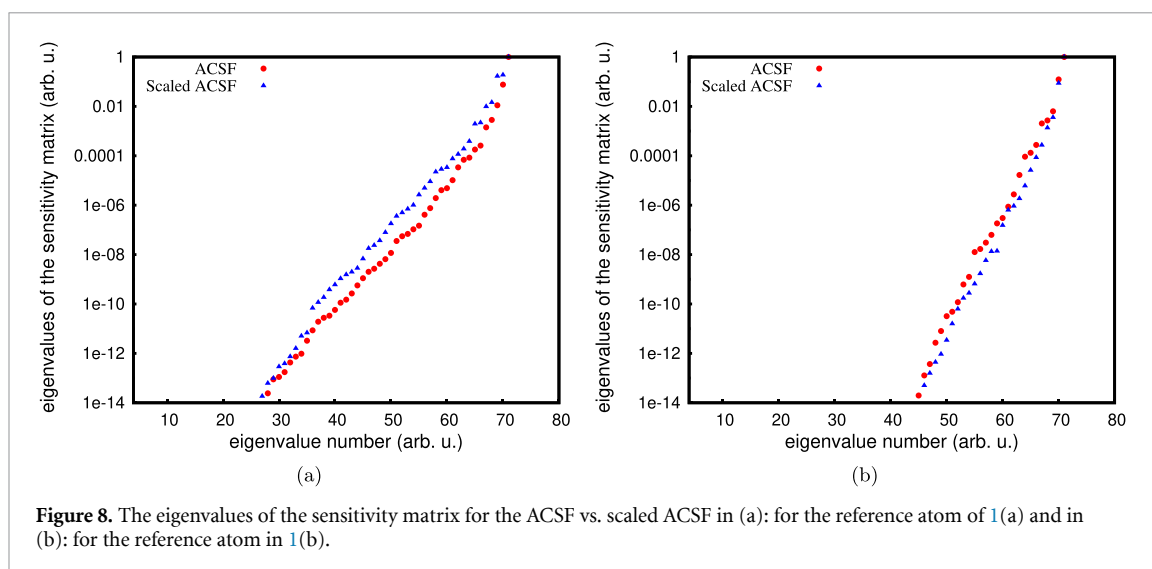


Figure 8. The eigenvalues of the sensitivity matrix for the ACSF vs. scaled ACSF in (a): for the reference atom of 1(a) and in (b): for the reference atom in 1(b).

correlation between the molecular fingerprint distance ΔFP and ΔE and ΔDOS with respect to the global minimum for OM[sp], SOAP, ACSF, FCHL, and MBSE.

Remarkably, all fingerprints show a quite similar behavior in these tests. In particular we could not find any pair of molecules that has a very small molecular fingerprint distance, but different energy or DOS. As also noted in a study highlighting difficulties in the structural description of methane [54], the fingerprints of neighboring atoms usually change under displacements even if the fingerprint of the central atom remains invariant. Through this effect machine learning schemes may compensate the deficiencies of a fingerprint, and the quality of the machine learning results for global quantities based on different fingerprints can become very similar in practice.

However, these findings are strictly true only if fingerprint vectors of different environments are exactly the same and have to be treated with care in the context of machine learning for several reasons, if fingerprint vectors are only similar. While correlations between physical properties and fingerprints are certainly supporting the construction of a ML model, most ML algorithms are highly non-linear methods, which are able to distinguish fingerprint vectors even if they are overall very similar, as measured by the fingerprint difference, but are sufficiently different in at least one or a few components. For instance, this is the case for the ACSF fingerprint vectors of the reference atoms shown in table 2(d). In this case the radial symmetry functions with large η parameters are rather sensitive to the local coordination and provide different numerical values for the exemplified one- and two-fold coordination of the reference atom. This is usually sufficient to distinguish these environments. Further, in ML applications fingerprint vectors are commonly scaled such that the values of each individual fingerprint component are normalized to a range between zero and one. We have not done this in the present work to avoid any bias in the comparison of the performance of different fingerprints. Further, any scaling, although common practice, depends on the fingerprint values in the available data set. We observed in figure 8 that scaling has some effect on ACSFs in terms of increasing the eigenvalues and therefore enhancing the sensitivity of the fingerprint overall, and similar effects are expected also for the other fingerprint types.

Finally, for instance in case of ML potentials, usually not only the total energy as a rather insensitive global property but also the atomic forces are used in the fitting process, which contain local atomic information about the potential energy surface. The inability to distinguish chemically different atomic environments thus results in large force errors, which can be used to improve the fingerprint set [21].

Irrespective of these aspects of ML applications, which reduce the effect of similar fingerprint vectors, it has been demonstrated in this work and elsewhere [54], that the detection of fingerprint vectors remaining exactly invariant upon structural changes is a major challenge and of utmost importance for many applications.

5. Conclusions

We have introduced stringent tests for the resolution power of atomic fingerprints describing the environment around a reference atoms. First we introduced the sensitivity matrix that can detect atomic displacement modes that leave the fingerprint invariant. Based on a large data set of carbon structures we then investigated the correlation between fingerprint distances calculated with various fingerprints. For

SOAP, ACSF, MBSF and FCHL, there exist atomic movements that leave the fingerprints invariant. This behavior can apparently only be found for some small molecules and it did not occur in our study of larger systems. For the symmetry function-related fingerprints, we found many movement modes that leave the fingerprint nearly invariant and we found many cases where environments that were classified as nearly identical were actually quite different. In all the tests we saw an improvement when going from the ACSF and MBSF to the FCHL fingerprint. The OM fingerprint is the only fingerprint for which no atomic displacement was ever found that leaves the fingerprint invariant. It is also the fingerprint whose distance assignments corresponds best to basic chemical concepts. This comes from the fact that the OM fingerprint is obtained from a matrix diagonalization that is akin to the solution of the Schrödinger equation and therefore naturally incorporates the full many-body character of the atomic environment. However, the limited resolution of some atomic fingerprints for some environments is most critical for structural discrimination, while there is still a good correlation of global molecular fingerprints in case of the prediction of extensive properties such as total energies of systems that are composed of a large number of environments. Also applications like machine learning are less affected, as they are able to resolve even subtle differences in the fingerprints.

5. Acknowledgment

This research was performed within the NCCR MARVEL funded by the Swiss National Science Foundation. The calculations were done using the computational resources of the Swiss National Supercomputer (CSCS) under project s963 and on the Scicore computing center of the University of Basel. JB thanks the Deutsche Forschungsgemeinschaft for support (Be3264/13-1, project number 411538199). We thank Gábor Csányi for help in finding good parameters for the SOAP fingerprints, Michele Ceriotti for providing us with the problematic methane configurations and Jonas Finkler for the careful reading of the manuscript.

ORCID iDs

Behnam Parsaeifard  <https://orcid.org/0000-0001-8095-4493>

Emir Kocer  <https://orcid.org/0000-0003-4207-8220>

Sandip De  <https://orcid.org/0000-0001-8434-3497>

References

- [1] Morgan D, Ceder G and Curtarolo S 2004 *Meas. Sci. Technol.* **16** 296
- [2] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501
- [3] Curtarolo S et al 2012 *Comput. Mater. Sci.* **58** 218
- [4] Jain A et al 2013 *Apl Materials* **1** 011002
- [5] De Jong M, Chen W, Geerlings H, Asta M and Persson K A et al 2015 A database to enable discovery and design of piezoelectric materials *Scientific Data* **2** 150053
- [6] Qu X et al 2015 *Comput. Mater. Sci.* **103** 56
- [7] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computational Mater.* **1** 15010
- [8] Blum L C and Reymond J-L 2009 *J. Am. Chem. Soc.* **131** 8732
- [9] Rupp M, Tkatchenko A, Müller K-R and Von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
- [10] Lyakhov A O, Oganov A R and Valle M 2010 *Modern Methods Crystal Structure Prediction* **147**
- [11] Goedecker S 2004 *J. Chem. Phys.* **120** 9911
- [12] Amsler M and Goedecker S 2010 *J. Chem. Phys.* **133** 224104
- [13] Neumann M A, Leusen F J and Kendrick J 2008 *Angew. Chem., Int. Ed.* **47** 2427
- [14] Oganov A R and Valle M 2009 *J. Chem. Phys.* **130** 104504
- [15] Handley C M and Popelier P L A 2010 *J. Phys. Chem. A* **114** 3371
- [16] Behler J 2011 *Phys. Chem. Chem. Phys.* **13** 17930
- [17] Botu V, Batra R, Chapman J and Ramprasad R 2017 *J. Phys. Chem. C* **121** 511
- [18] Ward L and Wolverton C 2017 *Current Opinion Solid State Mater. Sci.* **21** 167
- [19] Behler J 2017 *Angew. Chem. Int. Ed.* **56** 12828
- [20] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
- [21] Behler J 2011 *J. Chem. Phys.* **134** 074106
- [22] Smith J S, Isayev O and Roitberg A E 2017 *Chem. sci.* **8** 3192
- [23] Faber F A, Christensen A S, Huang B and Von Lilienfeld O A 2018 *J. Chem. Phys.* **148** 241717
- [24] Christensen A S, Bratholm L A, Faber F A and 2020 Anatole von Lilienfeld *J. Chem. Phys.* **152** 044107
- [25] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev. B* **87** 184115
- [26] Zhu L et al 2016 *J. Chem. Phys.* **144** 034203
- [27] Goedecker S 2004 *J. Chem. Phys.* **120** 9911
- [28] De D S, Krummenacher M, Schaefer B and Goedecker S 2019 *Phys. Rev. Lett.* **123** 206102
- [29] Schütt O and VandeVondele J 2018 *J. chem. Theory computation* **14** 4168
- [30] Babaei M, Azar Y T and Sadeghi A 2020 *Phys. Rev. B* **101** 115132
- [31] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
- [32] Gastegger M, Schwiedrzik L, Bittermann M, Berzsényi F and Marquetand P 2018 *J. Chem. Phys.* **148** 241709

- [33] Jindal S, Chiriki S and Bulusu S S 2017 *J. Chem. Phys.* **146** 204301
- [34] Jenke J, Subramanyam A P A, Densow M, Hammerschmidt T, Pettifor D G and Drautz R 2018 *Phys. Rev. B* **98** 144102
- [35] Shapeev A V 2016 *Multiscale Model. Simul.* **14** 1153
- [36] Thompson A P, Swiler L P, Trott C R, Foiles S M and Tucker G J 2015 *J. Comp. Phys.* **285** 316
- [37] Kocer E, Mason J K and Erturk H 2019 *J. Chem. Phys.* **150** 154102
- [38] Rupp M, Ramakrishnan R and Von Lilienfeld O A 2015 *J Phys. Chem. Lett.* **6** 309
- [39] Huang B and von Lilienfeld O A 2017 The 'dna' of chemistry: Scalable quantum machine learning with 'amons (arXiv: 1707.04146)
- [40] Eickenberg M, Exarchakis G, Hirn M, Mallat S and Thiry L 2018 *J. Chem. Phys.* **148** 241732
- [41] Huan T D, Batra R, Chapman J, Kim C, Chandrasekaran A and Ramprasad R 2019 *J. Phys. Chem. C* **123** 20715
- [42] Christensen A S, Bratholm L A, Faber F A, Glowacki D R, and von Lilienfeld O A 2019 (arXiv: 1909.01946)
- [43] Sadeghi A, Ghasemi S A, Schaefer B, Mohr S, Lill M A and Goedecker S 2013 *J. Chem. Phys.* **139** 184118
- [44] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
- [45] Huang B and Von Lilienfeld O A 2016 Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity
- [46] von Neumann J and Wigner E 1929 *Phys. Z.* **30** 467
- [47] Bernstein N, Csanyi G and Kermode J, [Quip and quippy documentation](#)
- [48] Behler J 2015 *Int. J. Quantum Chem.* **115** 1032
- [49] Imbalzano G, Anelli A, Giofre D, Klees S, Behler J and Ceriotti M 2018 *J. Chem. Phys.* **148** 241730
- [50] Christensen A, Faber F, Huang B, Bratholm L, Tkatchenko A, Muller K and von Lilienfeld O 2017 (<https://github.com/qmlcode/qml>)
- [51] Muto Y 1943 *J. Phys.-Math. Soc. Japan* **17** 629
- [52] Axilrod B M and Teller E 1943 *J. Comp. Phys.* **11** 299
- [53] Dragoni D, Daff T D, Csányi G and Marzari N 2018 *Phys. Rev. Mater.* **2** 013808
- [54] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G, and Ceriotti M 2020 (arXiv: 2001.11696)
- [55] Genovese L et al 2008 *J. Chem. Phys.* **129** 014109
- [56] Willand A, Kvashnin Y O, Genovese L, Vázquez-Mayagoitia A, Deb A K, Sadeghi A, Deutsch T and Goedecker S 2013 *J. Chem. Phys.* **138** 104109
- [57] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
- [58] Aradi B, Hourahine B and Frauenheim T 2007 *J. Phys. Chem. A* **111** 5678
- [59] Behler J 2016 *J. Chem. Phys.* **145** 170901
- [60] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
- [61] Eckhoff M and Behler J 2019 *J. Chem. Theory Comput.* **15** 3793
- [62] Kuhn H W 1955 *Nav. Res. Logist. Q.* **2** 83