PAPER • OPEN ACCESS

The role of feature space in atomistic learning

To cite this article: Alexander Goscinski *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 025028

View the article online for updates and enhancements.

MACHINE LEARNING
Science and Technology

The role of feature space in atomistic learning

Alexander Goscinski [ID], **Guillaume Fraux** [ID], **Giulio Imbalzano** and **Michele Ceriotti** [ID]

Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

E-mail: michele.ceriotti@epfl.ch

Abstract

Efficient, physically-inspired descriptors of the structure and composition of molecules and materials play a key role in the application of machine-learning techniques to atomistic simulations. The proliferation of approaches, as well as the fact that each choice of features can lead to very different behavior depending on how they are used, e.g. by introducing non-linear kernels and non-Euclidean metrics to manipulate them, makes it difficult to objectively compare different methods, and to address fundamental questions on how one feature space is related to another. In this work we introduce a framework to compare different sets of descriptors, and different ways of transforming them by means of metrics and kernels, in terms of the structure of the feature space that they induce. We define diagnostic tools to determine whether alternative feature spaces contain equivalent amounts of information, and whether the common information is substantially distorted when going from one feature space to another. We compare, in particular, representations that are built in terms of n-body correlations of the atom density, quantitatively assessing the information loss associated with the use of low-order features. We also investigate the impact of different choices of basis functions and hyperparameters of the widely used SOAP and Behler–Parrinello features, and investigate how the use of non-linear kernels, and of a Wasserstein-type metric, change the structure of the feature space in comparison to a simpler linear feature space.

1. Introduction

The construction of efficient and insightful descriptors of atomic configurations has been one of the focal points of the development of data-driven applications for atomic-scale modeling [1–17]. Two of the core ideas that underlie most of the existing schemes are the use of an atom-centred description that is particularly well-suited to model additive, extensive properties, and the incorporation of geometric and atom permutation symmetries. While incorporation of symmetries makes representations much more data efficient, it raises subtle issues of whether the mapping from structure to descriptor is injective or not [4, 18, 19]. Many of the structural representations that fulfill these symmetry requirements are closely related to one another, corresponding to projections of n-body correlations of the atom density [11, 20]. Yet, comparing them is not straightforward. When used to build an interatomic potential, or to predict another atomic-scale property, representations are used together with different supervised learning schemes, so it is difficult to disentangle the interplay of descriptor, regression method, and target property that combine to determine the accuracy and computational cost of the different methods [21]. Juxtaposing alternative choices of representations is complicated by the fact that non-linear transformations are often applied as a part of the data processing algorithm, and so it would be equally important to be able to analyze the effect of these transformations. Efforts to compare different choices of descriptors have been mostly focused this far on a comparison of compressibility [22, 23], their ability to represent atomic structures uniquely [4, 19, 24, 25], their role in constructing a metric [5, 16] and their sensitivity to perturbations of the atomic structure [23, 26].

Figure 1. A schematic representation of the different measures of feature-space dissimilarity we introduce in this work, discussed from left to right. The figure considers a dataset containing five samples, embedded in a two-dimensional feature space $\mathcal{F}$ and a one-dimensional feature space $\mathcal{F}'$. As shown, the relationship between the two embeddings can involve arbitrary linear and non-linear transformations (panel (a)). The global feature space reconstruction error (GFRE) defined in equation (2) amounts to finding the best linear mapping between the two feature spaces. This measure is not symmetric: in this example the GFRE$(\mathcal{F},\mathcal{F}')$ (panel (b)) is smaller than the GFRE$(\mathcal{F}',\mathcal{F})$ (panel (c)), since $\mathcal{F}$ contains an additional nonzero dimension. The local version of the reconstruction error (LFRE) defined in equation (7) makes it possible to probe whether a non-linear map exists between the two spaces: in this case, the sinus function can be approximated in neighbourhood of each sample $\mathbf{x}_i$ by a linear map $\mathbf{P}^{(i)}_{\mathcal{F}\mathcal{F}'}$ defined in equation (6); by treating each neighbourhood separately it is possible to achieve a low LFRE$(\mathcal{F}',\mathcal{F})$ (panel (d)). Finally, the global feature space reconstruction distortion (GFRD) defined in equation (5) determines whether the two featurizations are connected by an orthogonal transformation $\mathbf{Q}_{\mathcal{F}\mathcal{F}'}$ defined in equation (4), by finding the best alignment between $\mathcal{F}$ and the approximation of $\mathcal{F}'$ that can be obtained as a linear projection of $\mathcal{F}$. Even though GFRE$(\mathcal{F},\mathcal{F}')$ is small, one of the components of $\mathcal{F}$ is scaled down to zero, resulting in a large value of GFRD$(\mathcal{F},\mathcal{F}')$ (panel (e)).

Here we propose a strategy to compare feature spaces both in terms of their mutual information content—which we define transparently as the ability to linearly or non-linearly reconstruct each other—and in terms of the amount of deformation that has to be applied to match the common information between the two. Note that the definition we use here differs from that used in information-theoretical treatments, based on Shannon entropy—which is difficult to compute in high dimensions [27], and does not reflect as naturally the behavior of different features when used in the context of atomistic machine learning.

We demonstrate this strategy by applying it to elucidate several issues related to the behavior of density-based representations. First, we investigate the role of the basis and of the density smearing in the practical implementation of three-body density correlation features; we then estimate the loss of information that one incurs by truncating the description to low body-order of correlations; finally, we discuss the role of the metric used to compare two structures, by testing the commonly used Euclidean distance against kernel-induced and Wasserstein-type metrics. We provide an open-source implementation of functions to compute these quantities at https://github.com/cosmo-epfl/scikit-cosmo.

2. Comparing feature spaces

Consider a dataset $\mathcal{D} = \{x_i\}$ containing n items. For a given choice of features $\mathcal{F}$, each item is described by an $m_\mathcal{F}$-dimensional feature vector $\mathbf{x}_i$. As a whole, the dataset is described by a feature matrix $\mathbf{X}^{\mathcal{D}}_{\mathcal{F}} \in \mathbb{R}^{n \times m_\mathcal{F}}$. We consider all of the feature matrices in this work to be standardized, i.e. centred and scaled so as to have zero mean and unit variance for the selected data set. Consider a second featurization $\mathcal{F}'$. We want to be able to compare the behavior of different choices of feature spaces when representing the dataset $\mathcal{D}$, e.g. which of two sets of features has more expressive power, and how much distorted is one representation relative to the other. A schematic representation of the different measures we introduce in this section is shown in figure 1.

2.1. Global feature space reconstruction error
As a simple, easily-interpretable measure of the relative expressive power of $\mathcal{F}$ and $\mathcal{F}'$, we introduce the global feature space reconstruction error GFRE$^{\mathcal{D}}(\mathcal{F}, \mathcal{F}')$, defined as the root mean-square error that one incurs when using the feature matrix $\mathbf{X}_\mathcal{F}$ to linearly regress $\mathbf{X}_{\mathcal{F}'}$. In this work we compute the GFRE by a twofold split of the dataset, i.e. compute the regression weights $\mathbf{P}_{\mathcal{F}\mathcal{F}'}$ over a train set $\mathcal{D}_{\text{train}}$ composed of half the entries in $\mathcal{D}$,

$$\mathbf{P}_{\mathcal{F}\mathcal{F}'} = \arg\min_{\mathbf{P} \in \mathbb{R}^{m_\mathcal{F} \times m_{\mathcal{F}'}}} \| \mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}'} - \mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}} \mathbf{P} \| = \left(\left(\mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}} \right)^T \mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}} \right)^{-1} (\mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}})^T \mathbf{X}^{\mathcal{D}_{\text{train}}}_{\mathcal{F}'} \qquad (1)$$

and then compute the error over the remaining test set $\mathcal{D}_{\text{test}}$,

$$\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') = \sqrt{\| \mathbf{X}_{\mathcal{F}'}^{\mathcal{D}_{\text{test}}} - \mathbf{X}_{\mathcal{F}}^{\mathcal{D}_{\text{test}}} \mathbf{P}_{\mathcal{F}\mathcal{F}'} \|^2 / n_{\text{test}}}, \tag{2}$$

averaging, if needed, over multiple random splits. The GFRE is a positive quantity, which is equal to zero when there is no error in the reconstruction, and that is usually bound by one[1]. For numbers of features larger than n_{train}, the covariance matrix is not full rank, and one needs to compute a pseudoinverse. Without loss of generality, one can regularize the regression to stabilize the calculation. In this paper, we computed the pseudoinverse by means of a singular value decomposition, and we determined the optimal regularization in terms of the truncation of the singular value spectrum, using twofold cross-validation over the training set to determine the optimal truncation threshold. Often, it is also useful to observe the behavior of the GFRE in the absence of any regularization: overfitting is in itself a signal of the instability of the mapping between feature spaces. In general, $\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}')$ is not symmetric. If $\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') \approx \text{GFRE}^{\mathcal{D}}(\mathcal{F}', \mathcal{F}) \approx 0$, $\mathcal{F}$ and $\mathcal{F}'$ contain similar types of information; if $\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') \approx 0$, while $\text{GFRE}^{\mathcal{D}}(\mathcal{F}', \mathcal{F}) > 0$, one can say that $\mathcal{F}$ is more descriptive than $\mathcal{F}'$: this is the case, for instance, one would observe if $\mathcal{F}'$ consists of a sparse version of $\mathcal{F}$, with some important and linearly-independent features removed; finally, if $\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') \approx \text{GFRE}^{\mathcal{D}}(\mathcal{F}', \mathcal{F}) > 0$, the two feature spaces contain different, and complementary, kinds of information and it may be beneficial to combine them to achieve a more thorough description of the problem.

2.2. Global feature space reconstruction distortion

The feature space reconstruction error gives insights into whether a feature space can be inferred by knowledge of a second one. However, having both a small $\text{GFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}')$ and $\text{GFRE}^{\mathcal{D}}(\mathcal{F}', \mathcal{F})$ does not imply two feature spaces are identical. Even though they contain similar amounts of information, one feature space could give more emphasis to some features compared to the other, which can eventually result in different performance when building a model. To assess the amount of distortion of $\mathcal{F}'$ relative to $\mathcal{F}$, we introduce the global feature space reconstruction distortion $\text{GFRD}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}')$. To evaluate it, we first compute the singular value decomposition of the projector obtained through equation (1), $\mathbf{P}_{\mathcal{F}\mathcal{F}'} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and then use it to reduce the two feature spaces to a common basis, in which the reconstruction error is zero, because the residual has been discarded:

$$\tilde{\mathbf{X}}_{\mathcal{F}} = \mathbf{X}_{\mathcal{F}}\mathbf{U} \quad \tilde{\mathbf{X}}_{\mathcal{F}'} = \tilde{\mathbf{X}}_{\mathcal{F}}\mathbf{\Sigma}. \tag{3}$$

When the second feature space $\mathcal{F}'$ has a lower dimensionality than $\mathcal{F}$, some combinations of the starting features are not used to compute $\tilde{\mathcal{F}}'$. In this case, we pad $\mathbf{\Sigma}$ with zeros, so that $\tilde{\mathcal{F}}'$ has the same dimensionality $m_{\mathcal{F}}$ as the starting space. This choice ensures that the GFRD takes the same value it would have in the case $\mathcal{F}'$ had the same dimensionality as $\mathcal{F}$, but lower rank. In the opposite case, with $m_{\mathcal{F}} < m'_{\mathcal{F}}$, padding $\mathbf{\Sigma}$ and $\mathbf{U}$ with zeros, or truncating $\mathbf{V}$, yields the same GFRD.

We can then address the question of whether $\tilde{\mathbf{X}}_{\mathcal{F}}$ and $\tilde{\mathbf{X}}_{\mathcal{F}'}$ are linked by a unitary transformation (in which case the GFRD should be zero), or there is a distortion involved. A possible answer involves solving the orthogonal Procrustes problem [28], i.e. finding the orthogonal transformation that 'aligns' as well as possible $\tilde{\mathbf{X}}_{\mathcal{F}}$ to $\tilde{\mathbf{X}}_{\mathcal{F}'}$:

$$\mathbf{Q}_{\mathcal{F}\mathcal{F}'} = \text{argmin}_{\mathbf{Q} \in \mathbb{U}^{m \times m}} \| \tilde{\mathbf{X}}_{\mathcal{F}'}^{\mathcal{D}_{\text{train}}} - \tilde{\mathbf{X}}_{\mathcal{F}}^{\mathcal{D}_{\text{train}}} \mathbf{Q} \| = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T, \tag{4}$$

where $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T = (\tilde{\mathbf{X}}_{\mathcal{F}}^{\mathcal{D}_{\text{train}}})^T \tilde{\mathbf{X}}_{\mathcal{F}'}^{\mathcal{D}_{\text{train}}}$. The amount of distortion can then be computed by assessing the residual on the test set,

$$\text{GFRD}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') = \sqrt{\| \tilde{\mathbf{X}}_{\mathcal{F}'}^{\mathcal{D}_{\text{test}}} - \tilde{\mathbf{X}}_{\mathcal{F}}^{\mathcal{D}_{\text{test}}} \mathbf{Q}_{\mathcal{F}\mathcal{F}'} \|^2 / n_{\text{test}}}. \tag{5}$$

If desired, the error can be averaged over multiple random splits of the reference data set $\mathcal{D}$.

2.3. Local feature space reconstruction error

A downside of the global feature comparison schemes introduced above is that the linear nature of the regression means that they cannot detect if $\mathcal{F}$ and $\mathcal{F}'$ contain analogous information, but differ by a non-linear transformation. In the next section we discuss how one can generalize the schemes to use kernel features, that can also be used to detect non-linear relationships between the original feature spaces. An

---

[1] This is due to the fact that feature matrices are standardized, and so $\| \mathbf{X}_{\mathcal{F}'}^{\mathcal{D}_{\text{test}}} \|^2 / n_{\text{test}}$ is of the order of one.

alternative approach is to compute a local version of the feature space reconstruction error, $\text{LFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}')$, loosely inspired by locally-linear embedding [29]. To compute the LFRE, a local regression is set up, computed in the k-neighbourhood $\mathcal{D}_k^{(i)}$ around sample i—the set of k nearest neighbours of sample i, based on the Euclidean distance between the samples in $\mathcal{F}$—to reproduce the $\mathcal{F}'$ features using $\mathcal{F}$ as input features, centred around their mean values $\bar{\mathbf{x}}_{\mathcal{F}'}$ and $\bar{\mathbf{x}}_{\mathcal{F}}$.

A local embedding of $\mathbf{x}_i$ is determined as

$$\tilde{\mathbf{x}}_i' = \bar{\mathbf{x}}_{\mathcal{F}'} + (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{F}})\mathbf{P}_{\mathcal{F}\mathcal{F}'}^{(i)}, \tag{6}$$

where $\mathbf{P}_{\mathcal{F}\mathcal{F}'}^{(i)}$ contains the regression weights computed from $\mathcal{D}_k^{(i)}$. The local feature space reconstruction error is given by the residual discrepancy between the $\mathcal{F}'$ counterpart of the ith point and its local embedding (6):

$$\text{LFRE}^{\mathcal{D}}(\mathcal{F}, \mathcal{F}') = \sqrt{\sum_i \| \mathbf{x}_i' - \tilde{\mathbf{x}}_i' \|^2 / n_{\text{test}}}. \tag{7}$$

Inspecting the error associated with the reconstruction of individual points can reveal regions of feature space for which the mapping between $\mathcal{F}$ and $\mathcal{F}'$ is particularly problematic. Similarly, one can compute a local version of GFRD, that could be useful to detect strong local distortions that might indicate the presence of a singularity in the mapping between two feature spaces.

2.4. Bending space: comparing induced feature spaces

It is often possible to substantially improve the performance of regression or dimensionality reduction algorithms, without explicitly changing the feature vectors. This can be achieved by introducing a (non-linear) similarity measure to compare $\mathbf{x}_i$, which takes the form of a kernel function $k(\mathbf{x}, \mathbf{x}')$, or a dissimilarity measure which takes the form of a distance $d(\mathbf{x}, \mathbf{x}')$.

Let us recall that a positive-definite kernel induces a kernel distance by the relation [30]

$$d_k(\mathbf{x}, \mathbf{x}')^2 = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}'), \tag{8}$$

and that any negative-definite distance can be used to build positive-definite kernels such as the substitution kernel [31]

$$k_d^{\mathbf{x}_0}(\mathbf{x}, \mathbf{x}') = -\frac{1}{2}(d(\mathbf{x}, \mathbf{x}')^2 - d(\mathbf{x}, \mathbf{x}_0)^2 - d(\mathbf{x}_0, \mathbf{x}')^2), \mathbf{x}_0 \in \mathcal{F} \tag{9}$$

or the radial basis function (RBF) kernel

$$k_d^{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp(-\gamma d(\mathbf{x}, \mathbf{x}')^2), \quad \gamma \in \mathbb{R}_+. \tag{10}$$

A positive definite kernel induces a feature space $\mathcal{H}$, commonly known as reproducing kernel Hilbert space (RKHS), in which the similarity measure can be expressed as a dot product:

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle, \quad \mathbf{x}, \mathbf{x}' \in \mathcal{F}, \, \phi : \mathcal{F} \to \mathcal{H}. \tag{11}$$

While in general $\phi(\mathbf{x})$ is not known, for a given dataset $\mathcal{D}$ it is possible to approximate the RKHS features by using a kernel principal component analysis [32]. Since linear regression in RKHS features is equivalent to kernel ridge regression, we simply use kernel features computed on the training dataset $\mathcal{D}_{\text{train}}$ to reduce the problem of comparing kernel (or distance) induced features to that of comparing explicit features, and use GFRE and GFRD as defined in equations (2) and (5). It is possible to re-formulate these measures in an explicit kernelized form, as well as to compute low-rank approximations of the kernel to reduce the computational cost for very large datasets (see e.g. [22] for a pedagogic discussion). In this paper we simply use the explicit RKHS features, that can be obtained by diagonalizing the kernel matrix $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and defining

$$\mathbf{X}_{\mathcal{H}} = \mathbf{U}\mathbf{\Lambda}^{-1/2}, \tag{12}$$

which is then standardized as we do for any other set of features. To define a feature space associated with a metric, rather than a kernel, we first center the squared distance matrix (which is equivalent to computing a substitution kernel analogous to equation (9)) and then proceed similarly by diagonalizing the resulting matrix.

2.5. Dataset selection

We use four different datasets, chosen to emphasize different aspects of the problem of representing atomic structures: A *random methane* dataset consisting of different random displacements of the four hydrogen atoms around the central carbon atom to cover the complete configurational space of CH_4 structures; A *carbon* dataset of approximately 10 000 minimum energy carbon structures, obtained as the result of *ab initio* random structure search [33, 34], as an example for a realistic dataset of condensed phase structures; A *degenerate methane* dataset composed of two groups of methane structures (which we refer to as $\mathcal{X}^+$ and $\mathcal{X}^-$), each associated with a 2D manifold parameterised by two parameters (u, v): structures with $v = 0$ in the two manifolds have exactly the same C-centred three-body correlations, despite being different (as discussed in [19]); A *displaced methane* dataset, which consists of an ideal, tetrahedral CH_4 geometry with one hydrogen atom pulled away from the central carbon atom, as an example of a set of structures that are distinguished by a clearly identifiable structural feature, here the C–H distance.

3. Comparing atom-centred representations

Atom-centred representations that are based on a symmetrized expansion of the atom density constitute one of the most successful and widely adopted classes of features for atomistic machine learning [1, 2, 4, 11, 12, 35]. The construction begins by describing a structure A in terms of a sum of localized functions $g(\mathbf{x}) \equiv \langle \mathbf{x}|g \rangle$ (e.g. a Gaussian with variance $\sigma_G^2/2$) centred on the atom positions $\mathbf{r}_i$

$$\langle \mathbf{x}|A; \rho \rangle = \sum_i \langle \mathbf{x} - \mathbf{r}_i|g \rangle. \tag{13}$$

Symmetrizing over translations and rotations leads to a description of the structure in terms of a sum of *environment* features

$$\langle \mathbf{x}_1; \dots \mathbf{x}_\nu|A; \overline{\rho_i^{\otimes \nu}} \rangle \tag{14}$$

that describe ν-point correlations of the density centred on atom i (effectively corresponding to a $(\nu + 1)$-body correlation function in the sense used e.g. in statistical mechanics of liquids). Different values of ν correspond to conceptually distinct descriptions of the system—higher body order terms being more complicated, but potentially more information-rich—while different discretizations of the abstract vectors on a basis are a matter of computational convenience and affect the computational cost of different approaches [21], but their descriptive power should become equivalent in the limit of a complete basis set. We demonstrate the use of the GFRE, LFRE and GFRD to assess with quantifiable measures the effect of some of the different choices one can make when designing a representation.

3.1. SOAP and symmetry functions

We begin by considering two practical realizations of atom-centred symmetrized features of order $\nu = 2$: smooth overlap of atomic positions (SOAP) features [4], as implemented in librascal [36], and Behler–Parrinello symmetry functions (BPSF)[37] as implemented in the n2p2 package [38]. In the SOAP representation the atom-centred density is written as a sum of Gaussians with finite width σ_G, and the density is expanded in a basis that is a product of spherical harmonics $Y_m^l(\hat{\mathbf{x}}) \equiv \langle \hat{\mathbf{x}}|lm \rangle$ and a radial basis $R_n(x) \equiv \langle x|n \rangle$,

$$\langle nlm|A; \rho_i \rangle = \sum_j \int d\mathbf{x} \langle n|x \rangle \langle lm|\hat{\mathbf{x}} \rangle \langle \mathbf{x} - \mathbf{r}_{ji}|g \rangle, \tag{15}$$

where $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$. We consider two different basis sets here, Gaussian-type orbitals (GTOs)

$$\langle r|n; \text{GTO} \rangle = N_n r^n \exp(-b_n r^2), \tag{16a}$$

$$\text{with } N_n = \frac{2}{\sigma_n^{2n+3}\Gamma((n+3)/2)}, \tag{16b}$$

$$b_n = 1/(2\sigma_n), \quad \sigma_n = r_c \max(\sqrt{n}, 1)/n_{\max}, \tag{16c}$$

that are orthogonalized with respect to each other, and a discrete variable representation (DVR) basis

$$\langle r|n; \text{DVR} \rangle = \sqrt{w_n}\delta(r - r_n) \tag{17}$$

where r_n are Gaussian quadrature points and w_n their corresponding weights. For both bases, the integral (15) can be evaluated analytically, and the density coefficient computed as a sum over the neighbours of the ith atom. The invariant SOAP features are then computed as a contraction of the density coefficients,

$$\langle nn'l|\rho_i^2 \rangle = \sum_m \frac{(-1)^m}{\sqrt{2l+1}} \langle nlm|\rho_i \rangle \langle n'l(-m)|\rho_i \rangle. \tag{18}$$

Even though they can be seen as a projection on an appropriate basis of the symmetrized atom density that underlies SOAP [11], Behler–Parrinello symmetry functions (BPSF) are usually computed as a sum over tuples of neighbouring atoms of functions of interatomic angles and distances. Among the many functional forms that have been proposed [39] we consider the two-body functions

$$G_i^{(2)} = \sum_j e^{-\eta(r_{ji}-R_s)^2} \cdot f_c(r_{ji}), \tag{19}$$

and the three-body functions

$$G_i^{(3)} = 2^{1-\zeta} \sum_j \sum_{k \neq j} (1 + \lambda \cdot \cos \hat{r}_{ij} \cdot \hat{r}_{ik})^\zeta \cdot$$

$$e^{-\eta(r_{ji}^2+r_{ik}^2+r_{jk}^2)} \cdot f_c(r_{ji})f_c(r_{ik})f_c(r_{jk}), \tag{20}$$

where f_c is a cutoff function, and $\eta, \zeta, \lambda, R_s$ are parameters that define the shape of each BPSF. We generate systematically groups of symmetry functions of different size by varying the values of these parameters following the prescriptions discussed in [40]. The list of values for the BPSF parameters we used are supplied in the supplementary information (available online at stacks.iop.org/MLST/2/025028/mmedia).
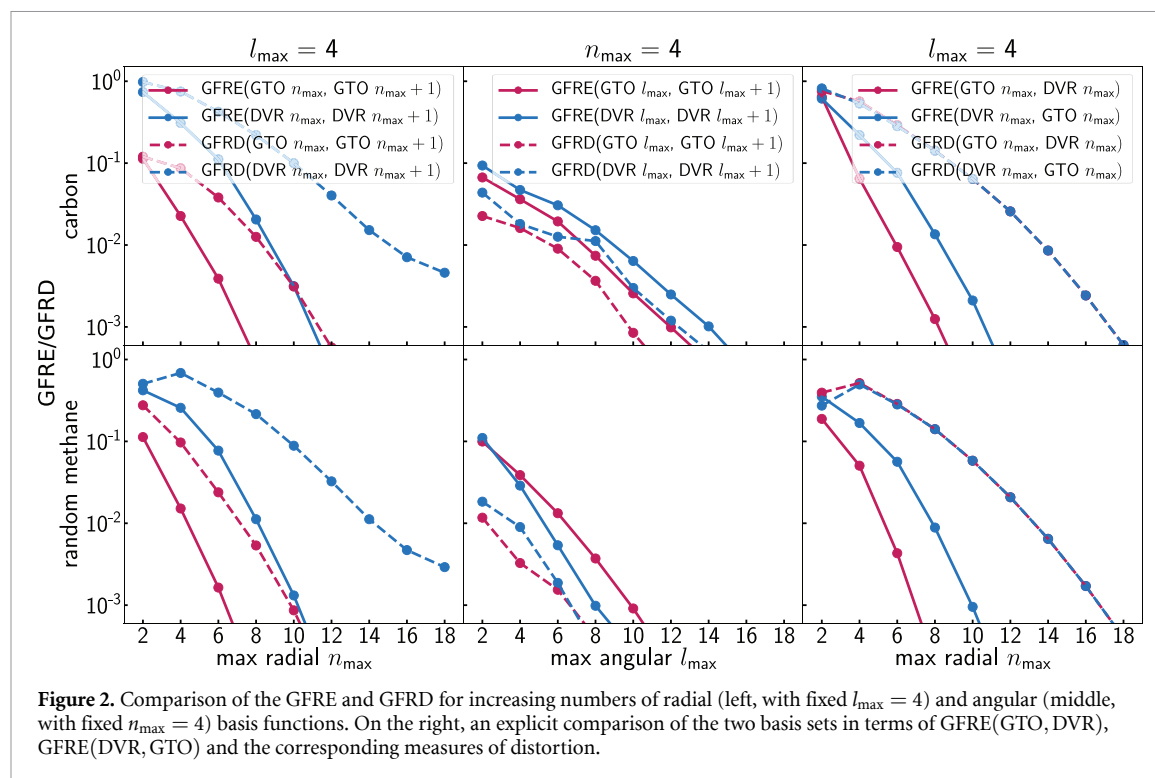
3.1.1. GTO and DVR radial basis
We start by considering the convergence of the SOAP representation with different choices of radial basis. Figure 2 demonstrates the convergence with the number of radial functions $n_{\max}$ and angular momentum channels $l_{\max}$ (in a Cauchy sense, i.e. comparing results for successive increments of these parameters). Overall, the GTO basis converges faster than DVR for most cases, both in terms of GFRE and GFRD. The slower radial convergence of the GFRD indicates that even as the discretization approaches convergence, the changing position of peaks and nodes of the basis functions gives different emphasis to interatomic correlations over different ranges. This is consistent with the observation that, particularly for small $(n_{\max}, l_{\max})$, regression accuracy depends on the number of basis functions in a way that is not necessarily monotonic. When considering the convergence of the angular component $l_{\max}$, GTO and DVR show nearly identical error decay, indicating that the convergence of the radial and angular basis are largely independent of each other.

The faster convergence of the GTO basis suggests that, for a given $n_{\max}$, a representation expanded on this basis should contain a greater amount of information on the structure. This is reflected in the direct comparison of the two bases, $\text{GFRE}(\text{GTO}\,n_{\max}, \text{DVR}\,n_{\max}) < \text{GFRE}(\text{DVR}\,n_{\max}, \text{GTO}\,n_{\max})$ for small $n_{\max}$. When both basis set have converged, they become essentially equivalent. Since the two representations are related to each other by a unitary transformation, $\text{GFRD}(\text{GTO}\,n_{\max}, \text{DVR}\,n_{\max}) \to 0$ as $n_{\max} \to \infty$.

3.1.2. Gaussian smearing
The Gaussian smearing used in SOAP features works as a parameter controlling the balance between local resolution and the smoothness of the mapping between Cartesian coordinates and symmetrized density features. A small σ_G value can identify minute changes more accurately, but a too small value for σ_G can lead to ill-conditioned regression, as the features associated with different structures show little overlap with each other. In fact, there is a tight interplay between the density smearing, the choice of the basis set, and the regularization of a regression model. As seen in figures 3(a) and (b), in the case of the smooth GTO basis set there is relatively little reconstruction error, and in general smaller σ_G values give a better reconstruction of large-σ_G features than vice versa. The opposite is true for the δ-like DVR basis: the GFRE for DVR is larger than in the case of GTO, and it is harder to reconstruct large-σ_G features from their sharp-Gaussian counterparts than vice versa. It should be also added that, without an automatic choice of regularization, results depend greatly on the way the feature mapping is executed. In particular, sharp-to-smooth mapping can lead to major overfitting problems, with GFRE becoming much larger than one for the test set. Even in cases where the GFRE is small, the feature space distortion is large, which highlights the fact that the
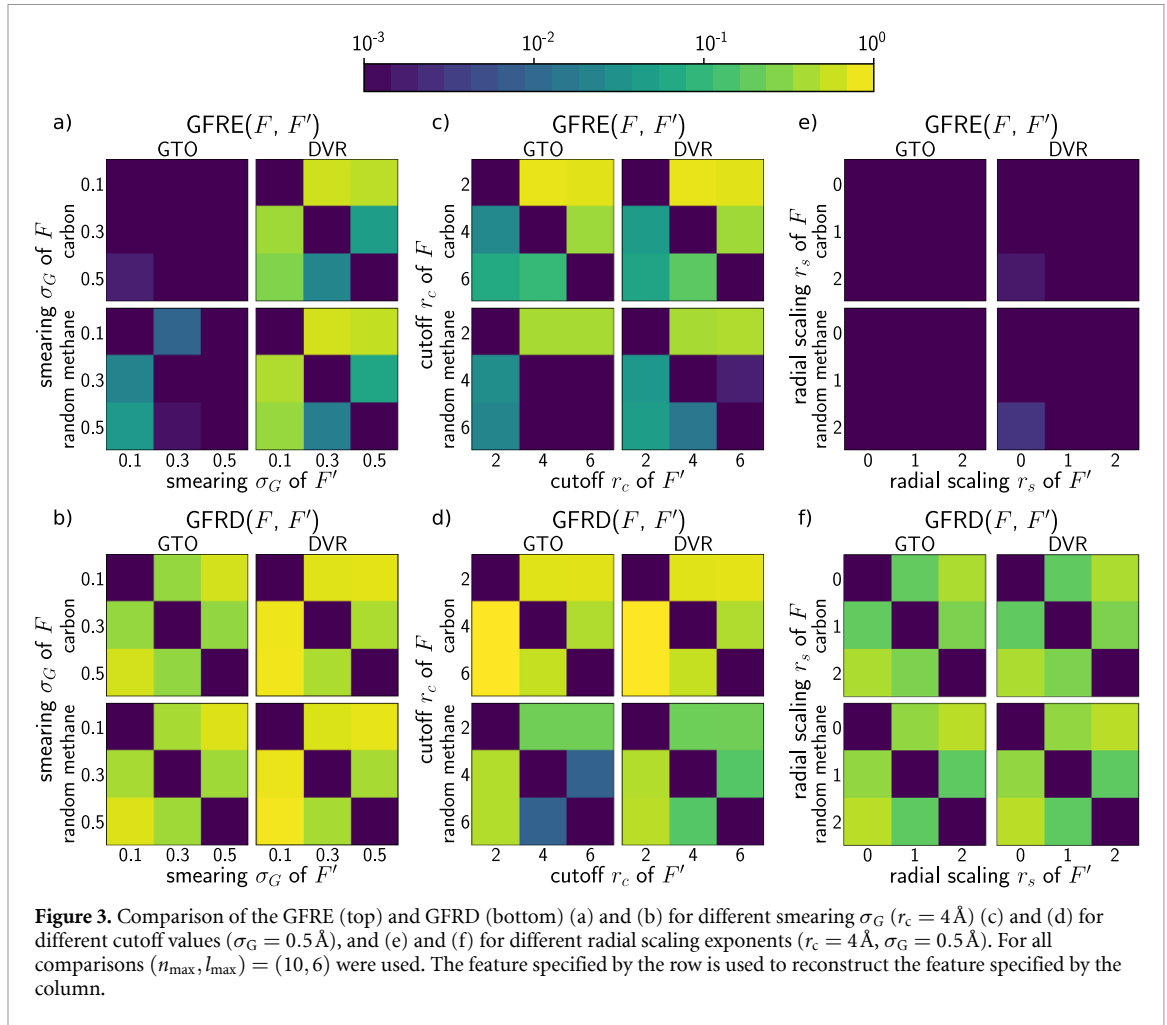
Figure 2. Comparison of the GFRE and GFRD for increasing numbers of radial (left, with fixed $l_{max} = 4$) and angular (middle, with fixed $n_{max} = 4$) basis functions. On the right, an explicit comparison of the two basis sets in terms of GFRE(GTO, DVR), GFRE(DVR, GTO) and the corresponding measures of distortion.

Gaussian smearing changes significantly the emphasis given to different structural correlations, and can therefore affect the accuracy of regression models.

3.1.3. Radial cutoff and scaling

One of the most important hyperparameters when defining an atom-centred representation is the cutoff distance, which restricts the contributions to the density to the atoms with $r_{ji} < r_c$. Figures 3(c) and (d) shows that the GFRE captures the loss of information associated with an aggressive truncation of the environment, with very similar behavior between GTO and DVR bases. The figure also reflects specific features of the different data sets: for instance, GFRE($r_c = 4$ Å, $r_c = 6$ Å) is close to zero for the random methane data set, because there are no structures where atoms are farther than 4 Å from the centre of the environment. GFRE > 0 also when mapping long-cutoff features to short-range features, although the reconstruction error is much smaller than in the opposite direction. This indicates the need for an increase in n_{max} to fully describe the structure of an environment when using a large value of r_c, which is consistent with the greater amount of information encoded within a larger environment. The GFRD plot also underscores the strong impact of the choice of r_c on the emphasis that is given to different parts of the atom-density correlations. This effect explains the strong dependency of regression performance on r_c, and the success of multi-scale models that combine features built on different lengthscales [41]. A similar modulation of the contributions from different radial distances can be achieved by scaling the neighbour contribution to the atom-centred density by a decaying function, e.g. $1/(1 + (r_{ji}/r_0)^s)$. This approach has proven to be very effective in fine-tuning the performance of regression models using density-based features [8, 20, 42]. As shown in figures 3(e) and (f), this is an example of a transformation of the feature space that entails essentially no information loss—resulting in a very small GFRE between different values of the scaling exponent s. However, it does result in substantial GFRD, providing additional evidence of how the emphasis given by a set of features to different inter-atomic correlations can affect regression performance even if it does not remove altogether pieces of structural information.

3.1.4. Behler–Parinello symmetry functions

BPSF can be seen as projections of the same, abstract symmetrized density features that underlies the construction of SOAP features. While the latter representation is usually implemented using an orthogonal set of basis functions, BPSFs are non-orthogonal, and are usually chosen based on a careful analysis of the inter-atomic correlations that are relevant for a given system [37, 43, 44], or selected automatically out of a large pool of candidates [40]. Figure 4 shows clearly that an orthogonal basis set provides a more effective strategy to converge a representation than the grid-based enumeration of the non-linear hyperparameters of non-orthogonal basis functions. GFRE(SOAP, BPSF) < GFRE(BPSF, SOAP) for all feature set sizes and both

Figure 3. Comparison of the GFRE (top) and GFRD (bottom) (a) and (b) for different smearing σ_G ($r_c = 4\,\text{Å}$) (c) and (d) for different cutoff values ($\sigma_G = 0.5\,\text{Å}$), and (e) and (f) for different radial scaling exponents ($r_c = 4\,\text{Å}$, $\sigma_G = 0.5\,\text{Å}$). For all comparisons $(n_{max}, l_{max}) = (10, 6)$ were used. The feature specified by the row is used to reconstruct the feature specified by the column.

data sets. As usual, we remark that zero reconstruction error does not imply equivalence for regression purpose: the GFRD remains very high even for the largest feature set sizes.

Given that, in real scenarios, one would usually combine systematic enumeration of BPSF features with an automatic selection method [40], we also use the feature reconstruction framework to investigate the convergence of the automatic screening procedure, i.e. the error in reconstructing the full vector based on the first m features chosen with a CUR decomposition-based procedure [40, 45]. Figure 5 shows that a few dozens CUR-selected features allow to almost-perfectly reconstruct the full feature vector. The convergence is particularly fast for BPSF, where $m = 50$ leads to a minuscule GFRE, indicating that the non-orthogonal features are highly redundant, and explaining the saturation in model performance that was observed in [40].

3.1.5. Computational cost

In this work we focus on the comparison of different kinds of features in terms of their information content, without commenting on the computational overhead associated with their evaluation, or the application of non-linear transformations. Computational cost depends on implementation choices, and can be optimized for usage patterns that differ from those that we apply here. However, the effort associated with the evaluation of a model plays an important role in determining its ultimate usability. To provide some context for our experiments, we report in table 1 the timings for the evaluation of SOAP and BPSF features with the same parametres used in this section. These timings show clearly that the computational savings afforded by the use of a DVR basis are not sufficient to offset the reduced information content with respect to the GTO basis. When comparing BPSF and SOAP, the clearest difference is that the cost of evaluating the former scales linearly with the number of features, while the cost of evaluating SOAP features is sublinear since it is dominated by the calculation of the density expansion coefficients $\langle nlm|\rho_i\rangle$ and is therefore sublinear with respect to the invariants. This difference in scaling is due to the different mechanism for evaluating 3-body terms, that scales quadratically with the number of neighbours for BPSF, and only linearly for SOAP, underscoring how a careful selection of the most relevant features is very important in the context of a BPSF framework.
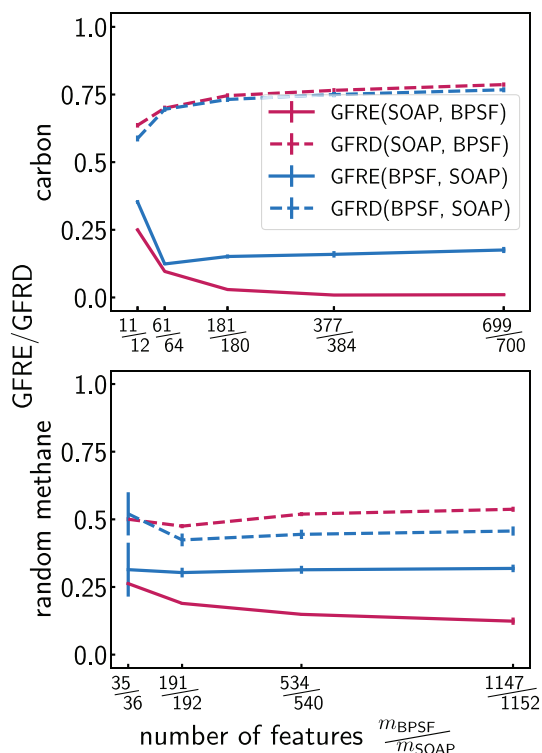
Figure 4. Comparison of the GFRE and the GFRD between SOAP(GTO) and BPSF features with systematically-increasing sizes of the feature vectors. BPSF features are generated by varying over a grid the hyperparameters entering the definitions of $G^{(2)}$ and $G^{(3)}$, following [40]. SOAP expansion truncation parameters (n_{max}, l_{max}) are adjusted to approximately match the number of BPSF features.
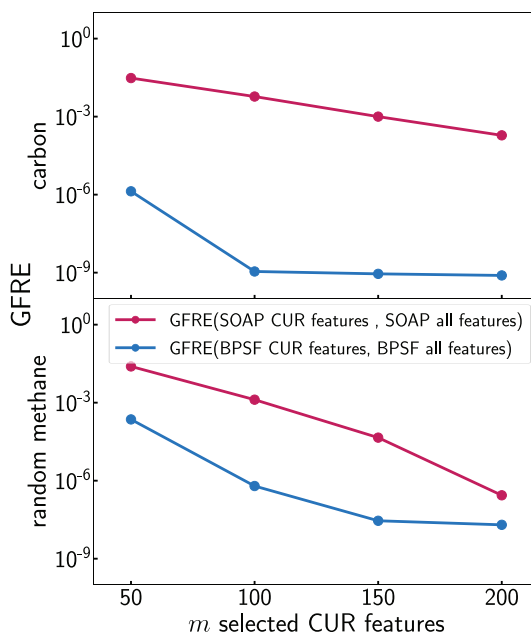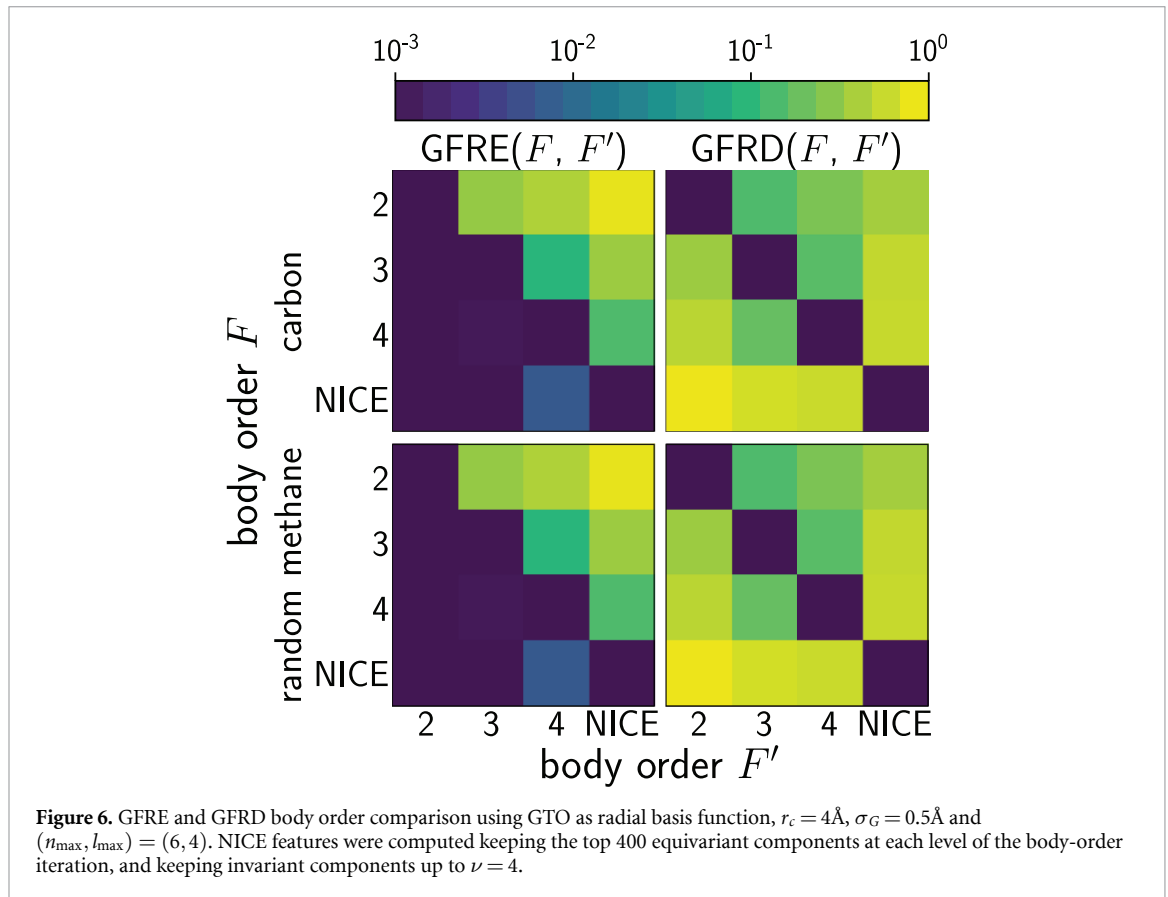

Figure 5. Convergence of a CUR approximation of the full SOAP/BPSF feature vectors (the largest size considered in figure 4) with number of retained features.

3.2. Body order feature truncation

The examples in section 3.1 demonstrate the impact of implementation details and hyperparameters choices on the information content of features that are all equivalent to a three-body correlation of the atom density. A more substantial issue is connected to the use of representations based on different ν-body correlations of a decorated atom density, which is equivalent to the pair correlation function (two-body, $\nu = 1$), to the SOAP power spectrum (three-body, $\nu = 2$) or to the bispectrum (four-body, $\nu = 3$). Different orders incorporate

Table 1. Timings in seconds for the evaluation of SOAP features using GTO ($t_{\text{SOAP-GTO}}$) and DVR ($t_{\text{SOAP-DVR}}$) as radial basis, and of BPSF (t_{BPSF}), using $r_c = 4\,\text{Å}$, on the entire carbon dataset. The SOAP discretization parameters are chosen to approximately match the number of features m_{SOAP} and m_{BPSF}. The measurements have been conducted on a single Intel® Xeon® CPU E3-1245 v6 @ 3.70GHz core, using librascal [36] for SOAP features and n2p2 [46] for BPSF.
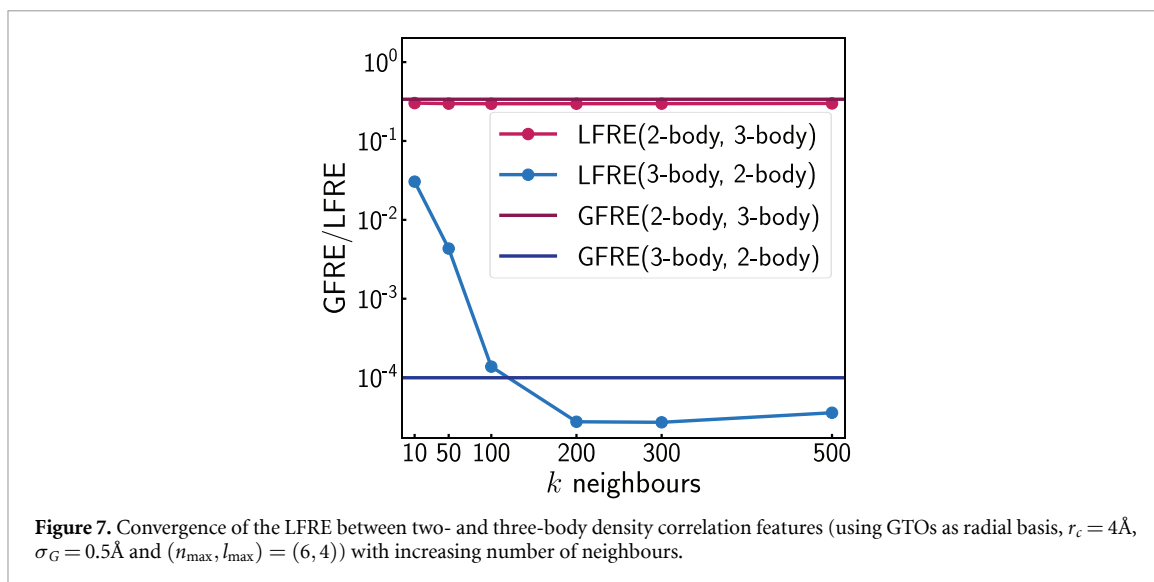
$(n_{\max}, l_{\max})$	m_{SOAP}	$t_{\text{SOAP-GTO}}/\text{s}$	$t_{\text{SOAP-DVR}}/\text{s}$	m_{BPSF}	t_{BPSF}/s
(2,2)	12	8.08 ± 0.30	7.95 ± 0.03	11	3.90 ± 0.14
(4,3)	64	10.58 ± 0.03	9.57 ± 0.02	61	10.43 ± 0.22
(6,4)	180	13.87 ± 0.03	11.87 ± 0.14	181	30.20 ± 0.41
(8,5)	384	17.58 ± 0.05	14.50 ± 0.02	377	66.55 ± 0.55
(10,6)	700	23.02 ± 0.04	18.57 ± 0.03	699	124.60 ± 0.78

Figure 6. GFRE and GFRD body order comparison using GTO as radial basis function, $r_c = 4\,\text{Å}$, $\sigma_G = 0.5\,\text{Å}$ and $(n_{\max}, l_{\max}) = (6, 4)$. NICE features were computed keeping the top 400 equivariant components at each level of the body-order iteration, and keeping invariant components up to $\nu = 4$.

conceptually distinct kinds of information: when used in linear regression, different density correlation orders correspond to a body-order expansion of the target property [11, 12, 47–49], and the link between the convergence of the body-order expansion and the injectivity of the structure-feature map is an open problem, with known counter-examples showing that low values of ν are insufficient to achieve a complete representation of an atomic environment [19].

Figure 6 shows that high-order features cannot be recovered as linear functions of lower-order features, while an approximate (if not complete) reconstruction of lower-ν components based on high-ν components is possible. Reconstructing features of different order entails a large amount of distortion, with the GFRD approaching one in most cases. We also include in the comparison features obtained with the recently-developed N-body iterative contraction of equivariants (NICE) framework, that identifies the most important features for each ν value, and uses them to compute $(\nu + 1)$-order features [50]. Keeping 400 features for each body order is sufficient to achieve perfect reconstruction of two and three-body features, but not for the four-body (bispectrum) term, which cannot be reconstructed fully with 400 NICE features. Considering however that GFRE(NICE, $\nu = 3$) $\ll$ GFRE($\nu = 3$, NICE), one can infer that the information loss associated with truncating the body order expansion is more severe than when restricting the number of four-body features.

The comparison of features of different order can also be used to elucidate the role of the (non-)linearity of the mapping between feature spaces. Figure 7 compares global and local feature reconstruction errors between two and three-body density correlation features, for the random CH_4 data set. In the case of the
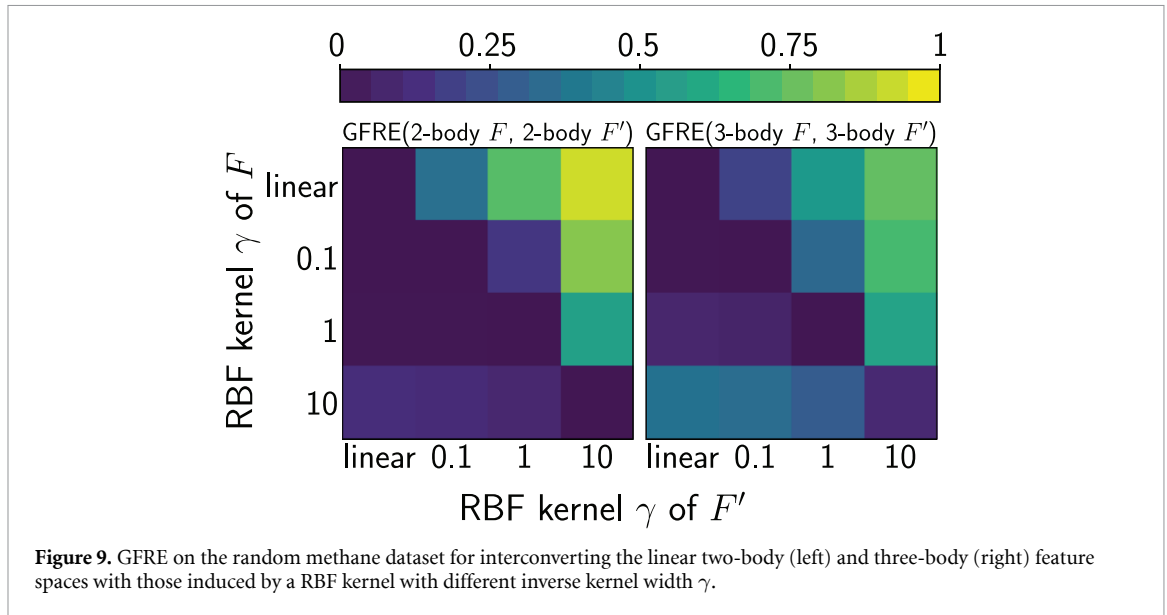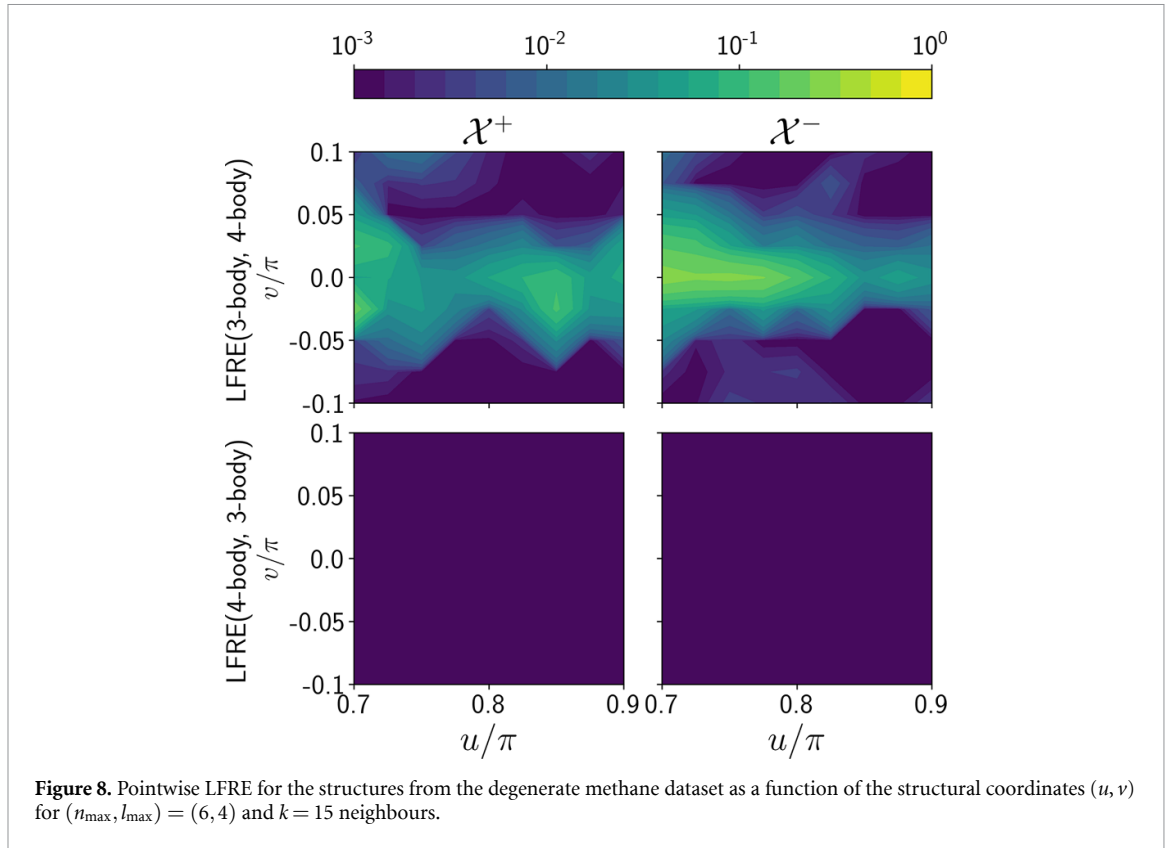
Figure 7. Convergence of the LFRE between two- and three-body density correlation features (using GTOs as radial basis, $r_c = 4$Å, $\sigma_G = 0.5$Å and $(n_{max}, l_{max}) = (6, 4)$) with increasing number of neighbours.

low-to-high body order reconstruction, the LFRE is only marginally lower than its global counterpart, indicating that the large GFRE$(\nu = 1, \nu = 2)$ is a consequence of lower information content and not only of the linear nature of the map. The reverse case is also revealing: for small k-neighbourhood sizes, LFRE$(\nu = 2, \nu = 1) >$ GFRE$(\nu = 2, \nu = 1)$, because the small number of neighbours included in the model reduce the accuracy of the feature reconstruction map. When the number of neighbours approaches the intrinsic dimensionality of the $\nu = 2$ features, instead, LFRE $<$ GFRE—because the reconstruction is based on a locally-linear map that can approximate a non-linear relationship between features. As k approaches the full train set size, the LFRE approaches the GFRE, as the locality of the mapping is lost.
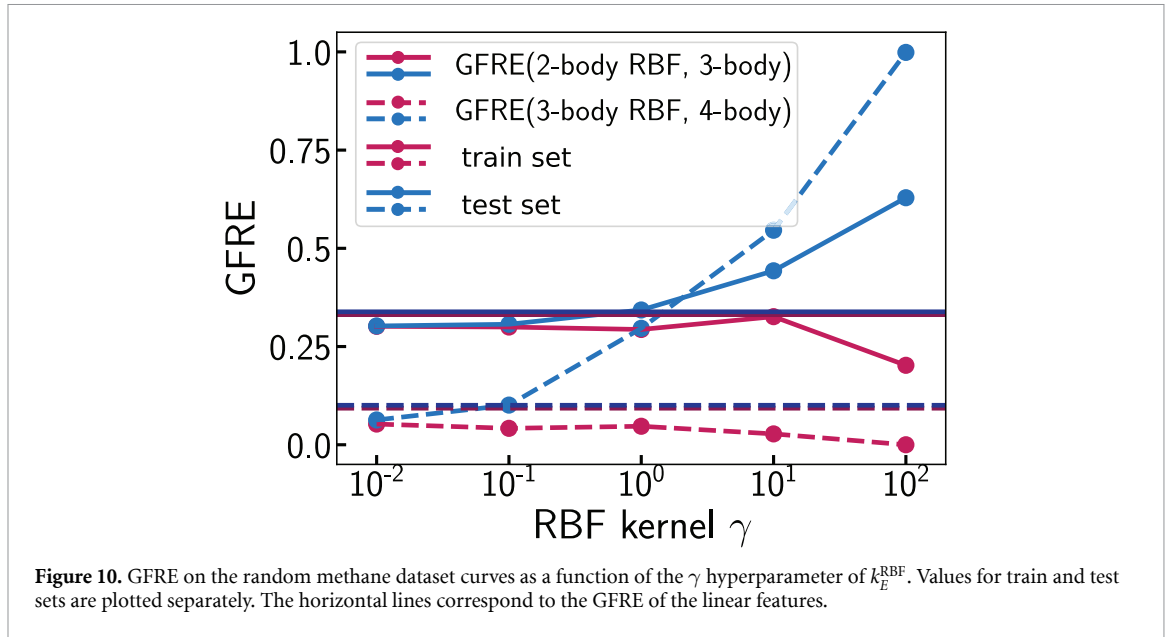
The LFRE also makes it possible to identify regions of phase space for which the construction of a mapping between feature spaces is difficult or impossible. Consider the case of the degenerate manifold discussed in [19]. The dataset includes two sets of CH_4 environments, and those parameterised by $\nu = 0$ cannot be distinguished from each other using three-body $(\nu = 2)$ features. Figure 8 shows the LFRE for each point along the two manifolds. When trying to reconstruct three-body features using as inputs four-body features (that take different values for the two manifolds) the LFRE is essentially zero. When using the three-body features as inputs, instead, one observes a very large error for points along the degenerate line, while points that are farther along the manifold can be reconstructed well. This example demonstrates the use of the LFRE to identify regions of feature space for which a simple, low-body-order representation is insufficient to fully characterize the structure of an environment, and can be used as a more stringent, explicit test of the presence of degeneracies than the comparison of pointwise distances discussed in [19].

3.3. Kernel-induced feature spaces

With the exception of the trivial, scalar-product form, a kernel introduces a non-linear transformation of the feature space, potentially allowing to obtain more accurate regression models. A crucial aspect of kernel methods is the fact that this non-linear transformation gives rise to a linear feature space that is defined by the combination of the kernel and the training samples—or the active samples in the case of sparse kernel methods. We can then use our feature-space reconstruction framework to compare quantitatively the linear feature space with the kernel-induced features. We do so using a radial basis function kernel, varying the γ parameter. In the $\gamma \to 0$ limit the RBF kernel becomes roughly linear, and the non-linearity increases with growing γ. The use of standardized input features means that γ is effectively unitless, we also standardize the kernel-induced features. To reduce noise, given that the kernel matrices are often very ill-conditioned, we only retain the RKHS features that are associated with the largest eigenvalues, preserving those that together contribute to approximately 99% of the variance. Figure 9 plots the GFRE and GFRD for the mapping of linear and RBF features computed for two and three-body density correlations. The non-linear nature of the transformation is apparent in the increase in the GFRE(linear,RBF) for larger values of γ, for both $\nu = 1$ and $\nu = 2$. The transformation is not entirely lossless, as evidenced by the fact that the reverse GFRE is also non-zero. The GFRE(RBF,linear) becomes particularly large for very large values of the γ parameter. This can be understood from the fact that the decay of the kernel becomes very sharp, and it only provides information about the nearest neighbours of each point—effectively leading to an ill-conditioned regression problem as we show in more detail below.

Figure 8. Pointwise LFRE for the structures from the degenerate methane dataset as a function of the structural coordinates (u, v) for $(n_{\max}, l_{\max}) = (6, 4)$ and $k = 15$ neighbours.

Figure 9. GFRE on the random methane dataset for interconverting the linear two-body (left) and three-body (right) feature spaces with those induced by a RBF kernel with different inverse kernel width γ.

Having assessed the impact of non-linear kernel features on a single body order representation, we can then investigate whether a non-linear transformation helps inferring high-body order correlations from low-body-order features. This is relevant because the use of non-linear kernels has been proposed [48] (and used in practice for a long time [2, 4]) as a strategy to describe many-body effects on atomistic properties. We compute the GFRE for promoting $\nu = 1$ (two-body) to $\nu = 2$ (three-body) and $\nu = 2$ to $\nu = 3$ features for different values of the RBF kernel γ. In figure 10 we show these curves for both the usual GFRE definition (that involves a separate test set) and for a prediction carried out on the train set. These results show that while a non-linear kernel does allow a low-body-order model to discern higher body-order features, it does so in a poorly transferable way: high-γ models show much reduced GFRE for train-set predictions, but lead to a degradation in the feature reconstruction for the test set. Only low-γ models show a small improvement in the test-set GFRE compared to an entirely linear mapping. In this regime, the RBF kernel is dominated by

Figure 10. GFRE on the random methane dataset curves as a function of the γ hyperparameter of k_E^{RBF}. Values for train and test sets are plotted separately. The horizontal lines correspond to the GFRE of the linear features.

the low-exponent components of the Gaussian expansion, vindicating the choice of low-order polynomial kernels, that are used in most of the published SOAP-based potentials. A better understanding of the effect of a non-linear feature space transformation can be obtained by analyzing the distribution of reconstruction errors for individual samples:

$$\mathrm{GFRE}^{(i)}(\mathcal{F}, \mathcal{F}') = \| \mathbf{x}_i' - \mathbf{x}_i \mathbf{P}_{\mathcal{F}\mathcal{F}'} \| . \tag{21}$$

The histograms for this 'pointwise GFRE' (figure 11) show that increasing the non-linearity of the kernel does indeed allow to reconstruct more accurately a fraction of both the test and the train set. When extrapolating the mapping to points that have not been seen before, however, there is an increasingly large fraction of outliers for which the reconstruction is catastrophically poor.
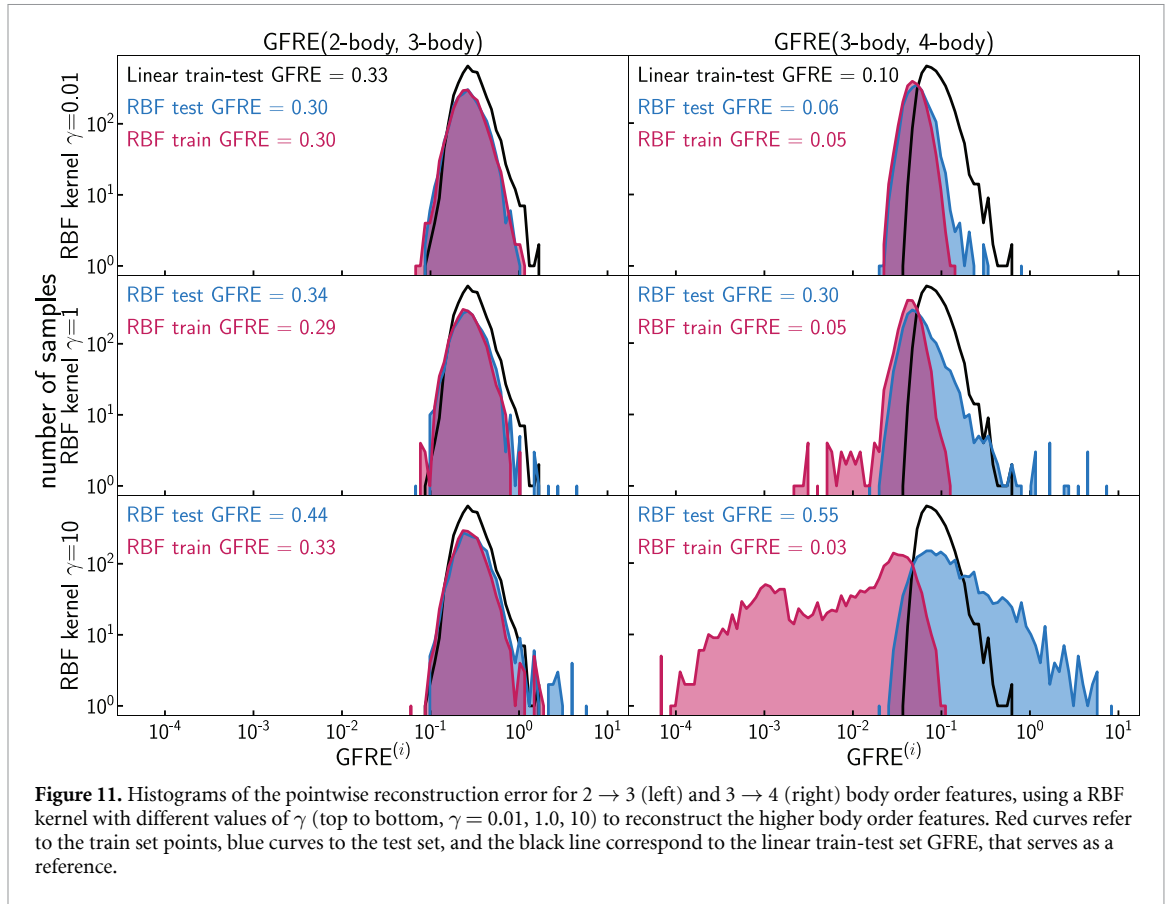
The pointwise errors are also revealing of the different nature of the $\nu = 1 \to \nu = 2$ and $\nu = 2 \to \nu = 3$ cases. In the former case, the clear lack of information in the two-body descriptor makes it impossible, even for a highly non-linear kernel, to obtain an accurate linear reconstruction of higher body-order features. In the latter case, instead, the train set reconstruction become nearly perfect with large γ—indicating that despite the existence of degenerate manifolds of configurations [19] it is possible to reconstruct four-body features using only three-body inputs, for structures that are not exactly on the degenerate manifold. However, the increasingly large tail of very high test-set GFRE samples suggests that this mapping is not smooth, and rather unstable. When building a regression model for a property that depends strongly on four-body terms, this instability may translate into poor extrapolative power for a non-linear model based on three-body features.

3.4. Wasserstein metric

As an example of the transformation induced by a non-Euclidean metric we consider the effect of using a Wasserstein distance to compare $\nu = 1$ density correlation features. The Wasserstein distance (also known as the earth mover distance, EMD) is defined as the minimum 'work' that is needed to transform one probability distribution into another—with the work defined as the amount of probability density multiplied by the extent of the displacement [51–53]. The EMD has been used to define a 'regularized entropy match' kernel to combine local features into a comparison between structures [5], to obtain permutation-invariant kernels based on Coulomb matrices [54], and has been shown to be equivalent to the Euclidean distance between vectors of sorted distances [11]. Here we use the Wasserstein distance to compare two-body ($\nu = 1$) features, that can be expressed on a real-space basis and take the form of one-dimensional probability distributions.

The formal definition of the Wasserstein distance of order 2 between two probability distributions $p(r)$ and $p'(r)$ defined on a domain M reads

$$W(p, p')^2 = \inf_{\gamma \in \Gamma(p, p')} \int_{M \times M} d(r, r')^2 \, \mathrm{d}\gamma(r, r'), \tag{22}$$

Figure 11. Histograms of the pointwise reconstruction error for $2 \rightarrow 3$ (left) and $3 \rightarrow 4$ (right) body order features, using a RBF kernel with different values of γ (top to bottom, $\gamma = 0.01, 1.0, 10$) to reconstruct the higher body order features. Red curves refer to the train set points, blue curves to the test set, and the black line correspond to the linear train-test set GFRE, that serves as a reference.

where $\Gamma(p, p')$ is the set of all joint distributions with marginals p and p'. For one dimensional distributions, $W(p, p')$ can be expressed as the two norm of the difference between the associated inverse cumulative distribution function (ICDF) P^{-1} of two environments, $W(p, p')^2 = \int_0^1 \left| P^{-1}(s) - P'^{-1}(s) \right|^2 \mathrm{d}s$, with $P(r) = \int_0^r p(r) \mathrm{d}r$.

In order to express the symmetrized two-body correlation function as a probability density, we first write it on a real-space basis $\langle r |$, and evaluate it on 200 Gaussian quadrature points, that we also use to evaluate the CDF and its inverse. We then proceed to normalize it, so that it can be interpreted as a probability density. We estimate the integral of the distribution (that effectively counts the number of atoms within the cutoff distance)

$$Z_i = \int_0^{r_c} \langle r | A; \overline{\rho_i^{\otimes 1}} \rangle \, \mathrm{d}r, \tag{23}$$

and the maximum value of the integral over the entire dataset $Z_{\mathcal{D}}$. A simple scaling of the correlation function

$$p_i^s(r) = \frac{1}{Z_i} \langle r | A; \overline{\rho_i^{\otimes 1}} \rangle \tag{24}$$

distorts the comparison between environments with different numbers of atoms. To see how, we use the *displaced methane* dataset, in which three atoms in a CH_4 molecule are held fixed in the ideal tetrahedral geometry, at a distance of 1 Å from the carbon centre. The fourth atom, aligned along the z axis, is displaced along it, so that each configuration is parameterised by a single coordinate z_H. Figure 12(a) shows the distance computed between pairs of configurations with different z_H, demonstrating the problem with the renormalized probability (24): p^s loses information on the total number of atoms within the cutoff, and so once the tagged atom moves beyond r_c the remaining CH_3 environment becomes indistinguishable from an ideal CH_4 geometry.

One can obtain a more physical behavior when atoms enter and leave the cutoff by introducing a δ-like 'sink' at the cutoff distance, defining

$$p_i^\delta(r) = \frac{1}{Z_{\mathcal{D}}} \left[\langle r | \mathcal{A}; \overline{\rho_i^{\otimes 1}} \rangle + (Z_{\mathcal{D}} - Z_i) \delta(r - r_c) \right]. \tag{25}$$

Figure 12. Distance between two *displaced methane* configurations with different values of z_H, computed using a Wasserstein distance using (a) scaling normalization; (b) cutoff δ normalization; (c) Euclidean distance between sorted interatomic distance vectors.
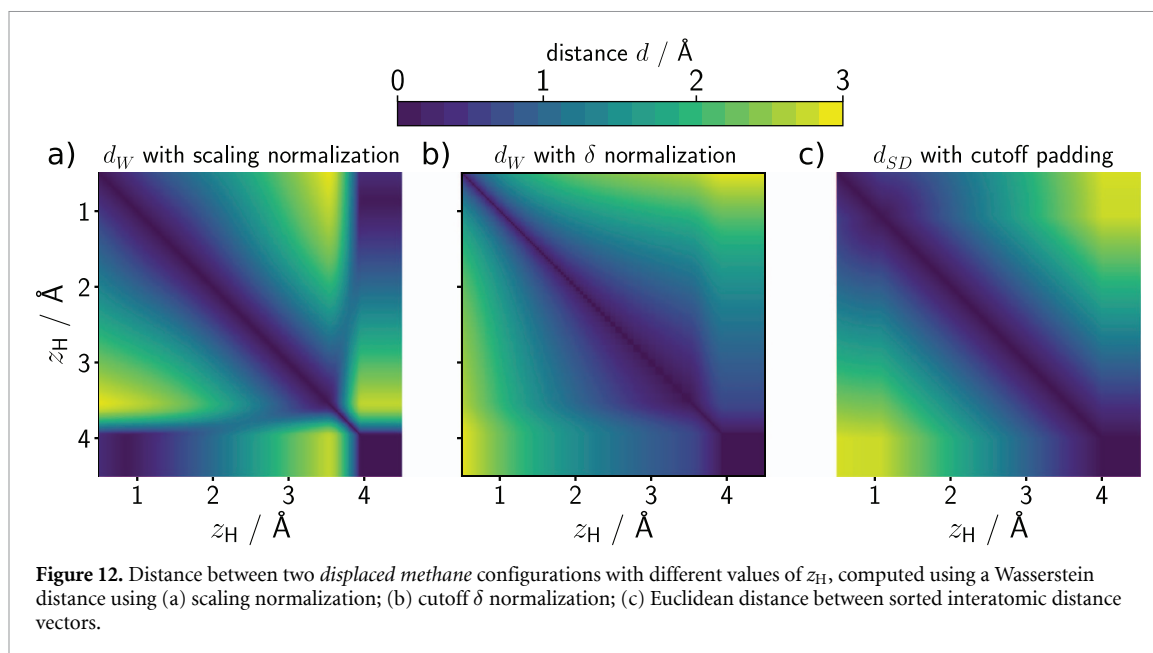
Figure 12(b) shows that with this choice the Wasserstein metric between $p_i^\delta(r)$ reflects the distance between the moving atoms. With this normalization, in fact, the Wasserstein metric corresponds to a smooth version of the Euclidean metric computed between vectors of sorted interatomic distances [11], shown in figure 12(c). The distortions that can be seen in the comparison between figures 12(b) and (c) are a consequence of the Gaussian smearing, the smooth cutoff function, and the $SO(3)$ integration that modulates the contribution to $\langle r | \mathcal{A}; \rho_i^1 \rangle$ coming from atoms at different distances.

Having defined a meaningful normalization and a probabilistic interpretation of the radial density correlation features, we can investigate how the feature space induced by a Wasserstein metric relates to that induced by an Euclidean distance. Figure 13 shows the error in the reconstruction of z_H for the *displaced methane* dataset when restricting the training set to 0.05 Å and 1.0 Å spaced grids. Using a Euclidean distance with a sharp σ_G leads to a highly non-linear mapping between the displacement coordinate and feature space, and a linear model cannot interpolate accurately between the points of a sparse grid. A Wasserstein metric, on the other hand, measures the minimal distortion needed to transform one structure into another, and so provides a much more natural interpolation along z_H, which is robust even with a sharp density and large spacing between training samples. It is worth stressing that the sorted distance metric—which effectively corresponds to the δ density limit of the Wasserstein metric—performs rather poorly, and cannot even reproduce the training points. This is because the mapping between feature space and z_H is not exactly linear, changing slope when z_H crosses 1 Å (because the sorting of the vector changes) and 4 Å (because one atom exits the cutoff). The sorted-distances feature space does not have sufficient flexibility to regress this piecewise linear map, as opposed to its smooth Wasserstein counterpart.

Having rationalized the behavior of the Wasserstein metric for a toy model, we can test how it compares to the conventional Euclidean metric on a more realistic data set. We consider in particular the AIRSS *carbon* data set, and compare different levels of density smearing as well as Euclidean and Wasserstein metrics. Figure 14 paints a rather nuanced picture of the relationship between the linear and the Wasserstein-induced feature spaces. The GFRE is non-zero in both directions, meaning that (in a linear sense) Wasserstein and Euclidean features provide complementary types of information. Smearing of the density has a small effect on the Wasserstein metric, so that both GFRE($W(\sigma_G = 0.1\,\text{Å})$, $W(\sigma_G = 0.5\,\text{Å})$) and GFRD($W(\sigma_G = 0.1\,\text{Å})$, $W(\sigma_G = 0.5\,\text{Å})$) are small, whereas for Euclidean features—as observed in section 3.1—changing σ_G induces small information loss, but a large distortion of feature space. Overall, there is no sign of the pathological behavior seen in figure 13, which is an indication that (at least for two-body features) the *carbon* dataset is sufficiently dense, and that the better interpolative behavior of the EMD does not lead to a more informative feature space.

4. Conclusion

Applications of machine learning to atomistic modelling suggest that the featurization that is chosen to represent a molecule or material can be equally or more important than the choice of regression scheme [8].
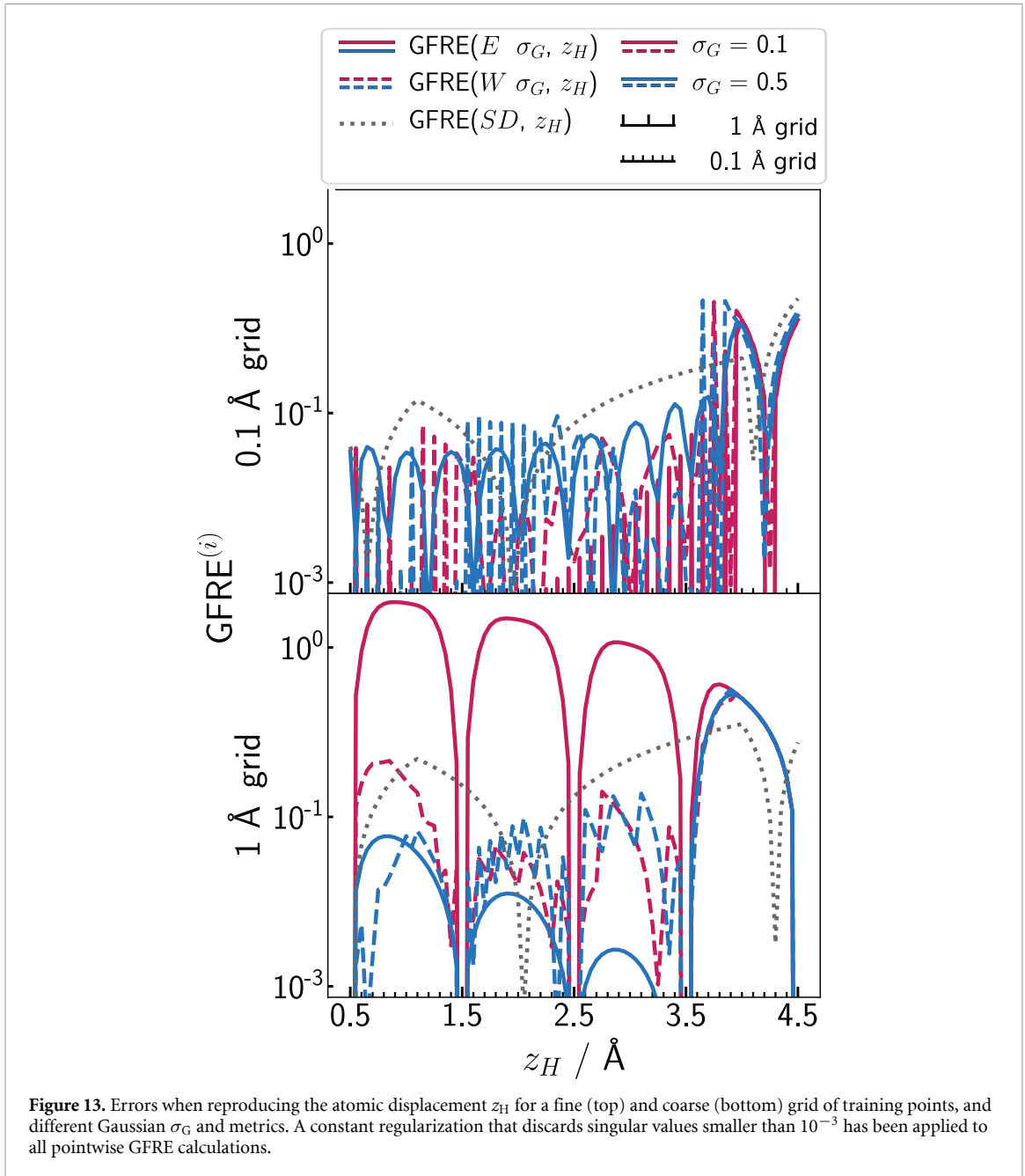
Figure 13. Errors when reproducing the atomic displacement z_H for a fine (top) and coarse (bottom) grid of training points, and different Gaussian σ_G and metrics. A constant regularization that discards singular values smaller than 10^{-3} has been applied to all pointwise GFRE calculations.

This has led to the proliferation of approaches to build descriptors, that often differ from each other only in implementation details. The framework we introduce here in this work enables a comparison of alternative choices of representations that does not depend on the target property, and makes it possible to determine objectively which of two features contains more information—based on a feature-space reconstruction error—and how much distortion is present in the way they describe the information that is common between the pair—based on a measure of feature-space distortion. Even though the framework is linear in nature, it can be generalized to account for non-linear relationships between feature spaces, either by using kernel-induced features, or by decomposing the feature comparison problem into a collection of local mappings.

Using this framework we demonstrate that the choice of basis set can affect substantially the convergence of SOAP features, and that for instance Gaussian type orbitals are more stable in the limit of small density smearing than the discrete variable representation basis. In practice the convergence of the representation with the number of basis functions should be considered together with the computational cost of the basis. The analytical expressions for GTOs involve special functions that are harder to evaluate than those that appear for a DVR basis. This overhead, however, is not sufficient to compensate for the reduced information content, and can be avoided altogether by using a spline approximation for the special functions. In general, computational cost depends substantially on the details of the implementation, and should be assessed in an
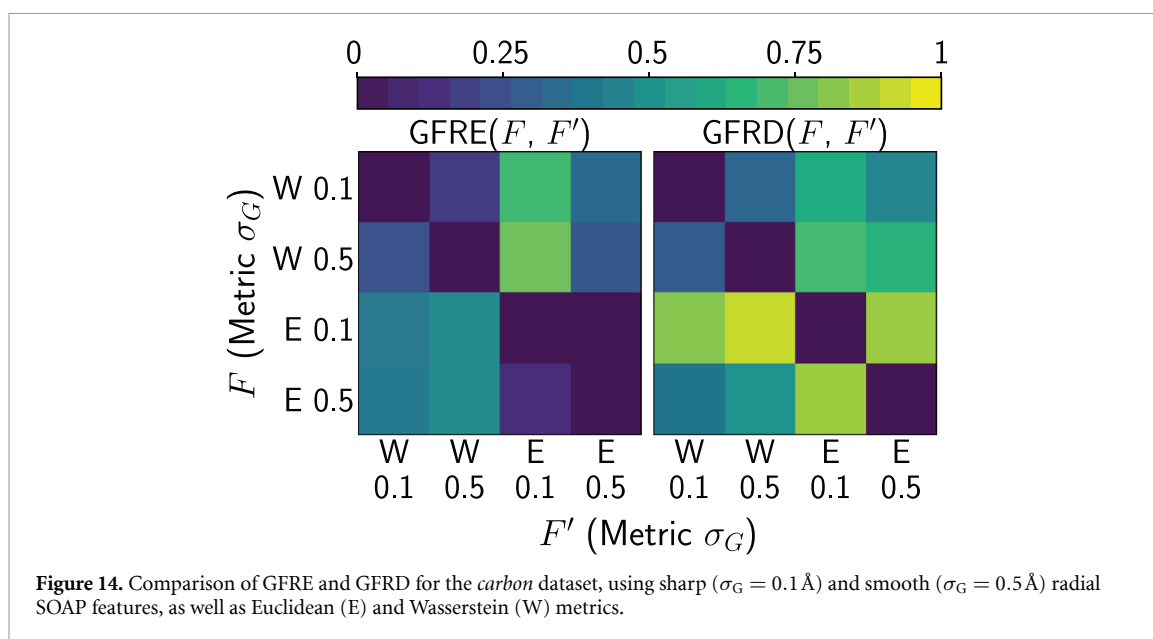
Figure 14. Comparison of GFRE and GFRD for the *carbon* dataset, using sharp ($\sigma_{\mathrm{G}} = 0.1\,\text{Å}$) and smooth ($\sigma_{\mathrm{G}} = 0.5\,\text{Å}$) radial SOAP features, as well as Euclidean (E) and Wasserstein (W) metrics.

end-to-end manner as a function of the specific use case. We also show quantitatively that a systematic orthogonal basis is much more effective in describing the atom density than the heuristic symmetry functions of the Behler–Parrinello kind—notwithstanding the considerable success that the latter approach has had in the construction of neural-network-based interatomic potentials [55].

A more systematic difference between atomistic machine-learning frameworks arises from the choice of the order of inter-atomic correlations that underlies the representation. We show that atom density correlation features of high body order make it possible to approximately reconstruct low-body order features, while the opposite is not true. Even when using a non-linear (or locally-linear) mapping, reconstructing three-body features from two-body information is virtually impossible. The three- to four-body mapping is more subtle: an overall reconstruction based on a linear model is not possible, but a local mapping works well, provided that the structures are far from the manifold of structures for which the three-body description is not injective. The associated transformation, however, is highly non-linear, and a kernel model that can reconstruct four-body features shows poor transferability outside of the training set, which hints at similar shortcomings whenever one wanted to use it to learn a property that depends strongly on four-body correlations. Even though an overall linear reconstruction is not possible, the ν-to-$(\nu + 1)$-body mapping error decreases with increasing ν, indicating that less information is added with higher body-orders. This is consistent with the satisfactory results that have been obtained in the regression of atomistic properties using only three-body information [42, 48], even though the fundamental incompleteness of three-body features has been shown to have implications for the asymptotic learning performance [19]. An analysis based on the GFRE might help determine the high-order correlations that provide the highest amount of information, and combined with an iterative scheme to evaluate the corresponding features [50] provide a strategy to increase model accuracy with an affordable computational effort.

We also investigate the effect of changing the metric used to compare features, by juxtaposing the Euclidean distance (that is induced by a linear description of the feature space) with a Wasserstein metric, that can be applied to the comparison of *n*-body correlation features when expressed as real-space distributions. We find that—with an appropriate normalization—the Wasserstein distance can be seen as a proxy of the minimal amount of distortion needed to transform an environment into another, and that this behavior induces smooth interpolation between sparse reference points, contrary to what is observed for the Euclidean distance. However, both an aggressive smearing of the atom density, and the use of a more realistic data set cure the pathological behavior of the linear featurization, so that the Wasserstein metric should not be regarded as superior to the Euclidean one, but complementary. Generalizing the Wasserstein metric to higher body-order correlations, which induce a higher-dimensional feature space that is more likely to be sparsely populated, would be an interesting further research direction.

The analysis in this work can be extended to compare atom-density representation with a broader class of descriptors based on topological [56–58] or physicochemical information [59–61] as well as property-dependent representations induced by neural network frameworks [62, 63]. Even more broadly, an objective measure of the relative effectiveness of features can guide the development of machine learning

schemes for any problem that depends strongly on the strategy used to obtain a mathematical description of its inputs. The feature space reconstruction error and distortion can be incorporated into any machine learning frameworks to drive feature selection algorithms [40, 61, 64] or to ensure that implementation choices that improve computational efficiency do not cause a degradation in the resolving power of the resulting features.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

ORCID iDs

Alexander Goscinski ● https://orcid.org/0000-0001-8076-215X
Guillaume Fraux ● https://orcid.org/0000-0003-4824-6512
Michele Ceriotti ● https://orcid.org/0000-0003-2571-2832

References

[1] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
[2] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[3] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[4] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[5] De S, Bartók A P, Csányi G and Ceriotti M 2016 *Phys. Chem. Chem. Phys.* **18** 13754
[6] Eickenberg M, Exarchakis G, Hirn M and Mallat S 2017 *Adv. Neural Inf. Process. Syst.* **2017-December** 6541
[7] Huo H and Rupp M 2017 **13754** (arXiv:1704.06439)
[8] Faber F A *et al* 2017 *J. Chem. Theory Comput.* **13** 5255
[9] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 *Nat. Commun.* **9** 1
[10] Zhang L, Han J, Wang H, Car R and Weinan E 2018 *Phys. Rev. Lett.* **120** 143001
[11] Willatt M J, Musil F and Ceriotti M 2019 *J. Chem. Phys.* **150** 154110
[12] Drautz R 2019 *Phys. Rev.* B **99** 014104
[13] Christensen A S, Bratholm L A, Faber F A and Anatole von Lilienfeld O 2020 *J. Chem. Phys.* **152** 044107
[14] van der Oord C, Dusson G, Csányi G and Ortner C 2020 *Machine Learning: Science and Technology* **1** 015004
[15] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114** 105503
[16] Zhu L *et al* 2016 *J. Chem. Phys.* **144** 034203
[17] Gallet G A and Pietrucci F 2013 *J. Chem. Phys.* **139** 074101
[18] von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 *Int. J. Quantum Chem.* **115** 1084
[19] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G and Ceriotti M 2020 *Phys. Rev. Lett.* **125** 166001
[20] Willatt M J, Musil F and Ceriotti M 2018 *Phys. Chem. Chem. Phys.* **20** 29661
[21] Zuo Y *et al* 2020 *J. Phys. Chem.* A **124** 731–45
[22] Helfrecht B, Cersonsky R K, Fraux G and Ceriotti M 2020 *Mach. Learn.: Sci. Technol.* **1** 045021
[23] Onat B, Ortner C and Kermode J R 2020 *J. Chem. Phys.* **153** 144106
[24] Moussa J E 2012 *Phys. Rev. Lett.* **109** 059801
[25] Sadeghi A, Ghasemi S A, Schaefer B, Mohr S, Lill M A and Goedecker S 2013 *J. Chem. Phys.* **139** 184118
[26] Parsaeifard B, De D S, Christensen A S, Faber F A, Kocer E, De S, Behler J, von Lilienfeld A and Goedecker S 2020 *Mach. Learn.: Sci. Technol.* (https://doi.org/10.1088/2632-2153/abb212)
[27] Torkkola K 2003 *J. Mach. Learn. Res.* **3** 1415
[28] Schönemann P H 1966 *Psychometrika* **31** 1
[29] Roweis S T and Saul L K 2000 *Science* **290** 2323
[30] Schölkopf B 2001 The kernel trick for distances *13th Int. Conf. on Neural Information Processing Systems* 283–9
[31] Haasdonk B and Bahlmann C 2004 *Joint Pattern Recognition Symposium* (Berlin: Springer) pp 220–7
[32] Schölkopf B, Smola A and Müller K-R 1997 *Nternational Conference on Artificial Neural Networks* (Berlin: Springer) pp 583–8
[33] Pickard C J and Needs R J 2011 *J. Phys. Condens. Matter* **23** 053201
[34] Pickard C J 2020 AIRSS data for carbon at 10 gpa and the C+N+H+O system at 1 gpa
[35] Thompson A, Swiler L, Trott C, Foiles S and Tucker G 2015 *J. Comput. Phys.* **285** 316
[36] Musil F *et al* librascal—A scalable and versatile library to generate representations for atomic-scale learning (available at: https://github.com/cosmo-epfl/librascal)
[37] Behler J 2011a *J. Chem. Phys.* **134** 074106
[38] Singraber A, Morawietz T, Behler J and Dellago C 2019 *J. Chem. Theory Comput.* **15** 3075
[39] Behler J 2011b *Phys. Chem. Chem. Phys. PCCP* **13** 17930
[40] Imbalzano G, Anelli A, Giofré D, Klees S, Behler J and Ceriotti M 2018 *J. Chem. Phys.* **148** 241730

[41] Bartók A P, De S, Poelking C, Bernstein N, Kermode J R, Csányi G and Ceriotti M 2017 *Sci. Adv.* **3** e1701816
[42] Paruzzo F M, Hofstetter A, Musil F, De S, Ceriotti M and Emsley L 2018 *Nat. Commun.* **9** 4501
[43] Jose K V, Artrith N and Behler J 2012 *J. Chem. Phys.* **136** 194111
[44] Behler J 2015 *Int. J. Quantum Chem.* **115** 1032
[45] Mahoney M W and Drineas P 2009 *Proc. Natl Acad. Sci.* **106** 697
[46] Singraber A N2P2 (available at: https://github.com/CompPhysVienna/n2p2)
[47] Shapeev A V 2016 *Multiscale Model. Simul.* **14** 1153
[48] Glielmo A, Zeni C and De Vita A 2018 *Phys. Rev.* B **97** 184307
[49] Jinnouchi R, Karsai F, Verdi C, Asahi R and Kresse G 2020 *J. Chem. Phys.* **152** 234102
[50] Nigam J, Pozdnyakov S and Ceriotti M 2020 *J. Chem. Phys.* **153** 121101
[51] Vallender S 1974 *Theory Probab. Its Appl.* **18** 784
[52] Cohen S D and Guibas L 1997 **1** (available at: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA358270)
[53] Cuturi M 2007 *Int. Jt. Conf. Artif. Intell. IJCAI* 732–7
[54] Çaylak O, von Lilienfeld A and Baumeier B 2020 *Mach. Learn.: Sci. Technol.* **1** 03LT01
[55] Behler J 2016 *J. Chem. Phys.* **145** 170901
[56] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S and Tropsha A 2017 *Nat. Commun.* **8** 1
[57] Sutton C *et al* 2019 *npj Comput. Mater.* **5** 1
[58] Liu S, Demirel M F and Liang Y 2019 *Advances in Neural Information Processing Systems* (Red Hook, NY: Curran Associates, Inc.) pp 8466–78
[59] Pilania G, Wang C, Jiang X, Rajasekaran S and Ramprasad R 2013 *Sci. Rep.* **3** 1
[60] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *npj Comput. Mater.* **2** 1
[61] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 *Phys. Rev. Mater.* **2** 083802
[62] Schütt K T, Sauceda H E, Kindermans P-J, Tkatchenko A and Müller K-R 2018 *J. Chem. Phys.* **148** 241722
[63] Cohen T S, Geiger M, Köhler J, and Welling M 2018 (arXiv: 1801.10130)
[64] Paleico M L and Behler J 2020 (arXiv: 2008.10977)