npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Compressing local atomic neighbourhood descriptors

James P. Darby[1✉], James R. Kermode[2] and Gábor Csányi[1]

Many atomic descriptors are currently limited by their unfavourable scaling with the number of chemical elements S e.g. the length of body-ordered descriptors, such as the SOAP power spectrum (3-body) and the (ACE) (multiple body-orders), scales as $(NS)^{\nu}$ where $\nu+1$ is the body-order and N is the number of radial basis functions used in the density expansion. We introduce two distinct approaches which can be used to overcome this scaling for the SOAP power spectrum. Firstly, we show that the power spectrum is amenable to lossless compression with respect to both S and N, so that the descriptor length can be reduced from $\mathcal{O}(N^2S^2)$ to $\mathcal{O}(NS)$. Secondly, we introduce a generalised SOAP kernel, where compression is achieved through the use of the total, element agnostic density, in combination with radial projection. The ideas used in the generalised kernel are equally applicably to any other body-ordered descriptors and we demonstrate this for the (ACSF).

npj Computational Materials (2022)8:166 ; https://doi.org/10.1038/s41524-022-00847-y

INTRODUCTION

Increases in computational power have made it possible to use quantum mechanical methods[1–3] to model materials at a level of accuracy where chemical transformations can be usefully studied. This ability has opened the door for in silico materials modelling to usefully compliment, and sometimes even replace, the need for costly experiments. A major drawback of such computational techniques is the limited simulation sizes and timescales that can be used. Empirical potentials do away with the electrons and model the total energy as a sum of local atomic energies which depend only on the local environment of each atom. These simplifications lead to potentials that are typically many orders of magnitude faster and, crucially, a computational cost that scales linearly with the number of atoms. Moving from hand-crafted potentials with tens of parameters to machine-learned potentials[4,5] has led to a substantial increase in the accuracy and transferability of such potentials in recent years[6–10]. This improvement has enabled vast, and accurate, simulations involving hundreds of thousands of atoms[11] that would otherwise be completely inaccessible.

Another area which has seen rapid development is the application of both supervised and unsupervised learning techniques to large scale materials data[12–14]. Such techniques have already seen great success and will undoubtedly prove useful in expediting material design and discovery[15] as well as understanding observed trends. Both potential fitting and, more broadly, materials machine learning require mathematical descriptions of material structures that can be used as inputs to models. Many different types of descriptors have been proposed[16], almost all of which are invariant to translations, rotations, and permutation of equivalent atoms. Incorporating these symmetries within the descriptor avoids models having to learn them, leading to greater data efficiency. Another commonality is the rapid increase in descriptor size for environments composed of multiple elements. For instance, the size of the SOAP power spectrum[17] scales quadratically with the number of elements S whilst the length of the bispectrum scales as S^3. This increase poses challenges for interatomic potentials, both in terms of evaluation speed and the memory required for fitting, as well as more generally for the storage of descriptors in databases. Before introducing some of the existing approaches it is worth noting that the problem can be somewhat mitigated by exploiting sparsity where possible. For instance, the SOAP power spectrum is sparse with respect to elements, so that even if there are S_{total} elements present across a given dataset, only those present in a given environment S_{env} need be considered when computing an individual descriptor. Whilst this can facilitate a storage saving, it does not always lead to a reduction in the number of model parameters, e.g. in a linear model, nor does it fully solve the problem for higher-body order descriptors; the corresponding SOAP bispectrum for a high entropy alloy liquid environment with $S_{\text{env}}=8$ is ~500 times larger than if a single element were present.

A great deal of effort has already gone into ways of reducing this scaling, with many interesting approaches being used. In refs. [18,19] constant complexity in S is achieved by concatenating two descriptors, one of which is element agnostic and another where the contributions from each element are weighted. This effectively amounts to embedding the elemental information into two dimensions, rather than keeping different elements entirely distinct. A similar strategy was used in refs. [20,21], except here the element embeddings were optimised during model fitting, so that the final embeddings contained a data-driven measure of chemical similarity between the elements. Approaching the problem from an information content perspective, the recent work of ref. [22], demonstrated that a model fitted to as few as 10% of the power spectrum components led to negligible degradation in force errors on an independent test set, suggesting that significant compression can be achieved. Similar results were seen in ref. [23], where descriptors were selected using CUR matrix decomposition[24], in ref. [25], where state-of-the-art model performance was achieved on the QM9 dataset with a heavily compressed descriptor obtained though repeated tensor products and reduction using Principal Component Analysis (PCA), and also in ref. [26] where a data-driven approach to constructing an optimal radial basis was employed with great success.

A complementary strategy is to focus on general, non-data driven, ways of reducing the scaling with S whilst minimising the loss of information. Such compression strategies could prove useful in situations where the dataset evolves over time, e.g.

[1]Engineering Dept, University of Cambridge, Trumpington St, Cambridge CB2 1PZ, UK. [2]Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, UK. ✉email: jpd47@cam.ac.uk

adding configurations with new elements, or for the storage of descriptors in databases such as NOMAD[14,27], where the use-case for the descriptors is not yet known. Analytic compression strategies could also be used before applying data-driven compressions, allowing for larger and more diverse datasets to be treated in this way. The recently introduced Moment Invariant Local Atomic Descriptors descriptors[19] follow this philosophy by using a minimal set of rotational invariants as the descriptor, such that the environment can be recovered by inverting the descriptor. Whilst the primary focus of that work was not to reduce the scaling with S, we stress that the core ideas used are highly pertinent and that similar themes are present here.

In this work we introduce two non-data driven approaches for compressing the SOAP power spectrum; the power spectrum is available through the quippy[28,29], dscribe[30] and librascal[31] packages. Firstly, by considering the ability to recover the density expansion coefficients from the power spectrum, we introduce a compressed power spectrum and show that, under certain conditions, it is possible to recover the original descriptor from the compressed version. Secondly, we introduce a generalisation of the SOAP kernel which affords compression with respect to both S and the number of radial basis functions N used in the density expansion. This kernel retains a useful physical interpretation and the ideas used are applicable to all body-ordered descriptors, which is demonstrated using the ACSFs[32]. Finally, we evaluate the performance of the compressed descriptors across a variety of datasets using numerical tests which probe the information content, sensitivity to small perturbations and the accuracy of fitted energy models and force-fields.

RESULTS

SOAP

The SOAP kernel, introduced in ref. [17], provides a way of computing the similarity between a pair of atomic environments. Invariance to permutation of equivalent atoms is achieved by forming densities,

$$\rho^{\alpha}(\mathbf{r}) = \sum_i \delta_{\alpha s_i} \exp\left[\frac{-|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma^2}\right] f_{\mathrm{cut}}(\mathbf{r}) \tag{1}$$

where a separate density is constructed for each element α, σ controls the width of the Gaussians used to construct the density, $f_{\mathrm{cut}}(\mathbf{r})$ is a cutoff function that ensures atoms enter the environment smoothly and $\mathbf{r}_i$ and s_i denote the position and element of the ith neighbour atom respectively. The kernel is made invariant to the orientation of environments by integrating over all possible rotations $\hat{R} \in SO(3)$ with the full multi-species kernel[33] defined as

$$k(\rho,\rho') = \int d\hat{R} \left| \sum_{\alpha} \int d\mathbf{r} \rho^{\alpha}(\mathbf{r}) \rho'^{\alpha}(\hat{R}\mathbf{r}) \right|^{\nu} \tag{2}$$

where $k(\rho,\rho')$ is the linear SOAP kernel and

$$K(\rho,\rho') = \left(\frac{k(\rho,\rho')}{\sqrt{k(\rho,\rho)k(\rho',\rho')}} \right)^{\zeta} \tag{3}$$

is the polynomial SOAP kernel typically used in models.

Expanding the density using the spherical harmonics $Y_{lm}(\hat{\mathbf{r}})$, and a set of orthogonal radial basis functions $g_n(r)$,

$$\rho^{\alpha}(\mathbf{r}) = \sum_{nlm} c^{\alpha}_{nlm} Y_{lm}(\hat{\mathbf{r}}) g_n(r) \tag{4}$$

and substituting into the definition of the SOAP kernel with $\nu = 2$ yields,

$$k(\rho,\rho') = \sum_{\alpha\beta} \sum_{nn'l} p^{\alpha\beta}_{nn'l} p'^{\alpha\beta}_{nn'l} = \mathbf{p}\cdot\mathbf{p}' \tag{5}$$

$$p^{\alpha\beta}_{nn'l} = \sum_m c^{\alpha *}_{nlm} c^{\beta}_{n'lm} = \mathbf{c}^{\alpha *}_{nl} \cdot \mathbf{c}^{\beta}_{n'l} \tag{6}$$

where $\mathbf{p}$ and $\mathbf{p}'$ are the SOAP power spectrums for the environments. In this form the rotational invariance of the power spectrum can be seen by noting that density expansion coefficients transform under rotations as

$$\mathbf{c}^{\alpha}_{nl} \rightarrow \mathbf{D}^{l}(\hat{R}) \mathbf{c}^{\alpha}_{nl} \tag{7}$$

where D is a unitary Wigner-D matrix, so that the terms $\mathbf{c}^{\alpha *}_{nl} \cdot \mathbf{c}^{\beta}_{n'l}$ are individually invariant[17]. Taking $\sigma \rightarrow 0$ (for convienience) in Eq. (2) with $\nu = 2$ reveals that the power spectrum is a 3-body descriptor,

$$k(\rho,\rho') = \int d\hat{R} \sum_{ijpq} \delta_{s_i s_p} \delta(\mathbf{r}_i - \hat{R}\mathbf{r}_p) \delta_{s_j s_q} \delta(\mathbf{r}_j - \hat{R}\mathbf{r}_q) \tag{8}$$

where i,j and p,q index atoms in the first and second environment respectively. The integration over all rotations ensures that each matching pair of triangles, where one vertex is the central atom and the others are neighbour atoms, between the environments contributes to the kernel, so that the power spectrum is effectively a histogram of triangles. Increasing ν increases the body order of the descriptor, so that the bispectrum ($\nu = 3$) corresponds to a histogram of tetrahedra and so on. With multiple elements, the neighbour atoms at the corner of each triangle must also match, so that there is a histogram of triangles for each pair of elements. In general, the length of the descriptor scales as S^{ν} where S is the number of elements and the body-order is $\nu + 1$, rendering these descriptors impractical for large S. In this work we investigate a number of alternatives which aim to circumvent this scaling, with a particular focus on compressing the power spectrum.

Information content

After exploiting symmetry, $p^{\alpha\beta}_{nn'l} = p^{\beta\alpha}_{n'nl}$, the length of the power spectrum is $\frac{1}{2}NS(NS+1)(L+1)$, where N, L and S are the number of radial basis functions, highest order of spherical harmonic and total number of elements respectively. Here we show that the power spectrum is amenable to lossless compression, with the final descriptor having length $NS(L+1)^2$. We start by highlighting that the sum over m in Eq. (6) can instead be viewed as a dot product between density expansions coefficient vectors $\mathbf{c}^{\alpha}_{nl}$ of length $2l+1$. Then, because all products between coefficient vectors with equal l index are taken, the power spectrum can be re-shaped from a vector into a sequence of l-slices

$$P_l = \begin{pmatrix} \mathbf{c}^{\alpha}_1 \cdot \mathbf{c}^{\alpha}_1 & \mathbf{c}^{\alpha}_1 \cdot \mathbf{c}^{\alpha}_2 & \dots & \mathbf{c}^{\alpha}_1 \cdot \mathbf{c}^{S}_N \\ \mathbf{c}^{\alpha}_2 \cdot \mathbf{c}^{\alpha}_2 & \mathbf{c}^{\alpha}_2 \cdot \mathbf{c}^{\alpha}_2 & \dots & \mathbf{c}^{\alpha}_2 \cdot \mathbf{c}^{S}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}^{\beta}_1 \cdot \mathbf{c}^{\alpha}_1 & \mathbf{c}^{\beta}_1 \cdot \mathbf{c}^{\alpha}_2 & \dots & \mathbf{c}^{\beta}_1 \cdot \mathbf{c}^{S}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}^{S}_N \cdot \mathbf{c}^{\alpha}_1 & \mathbf{c}^{S}_N \cdot \mathbf{c}^{\alpha}_2 & \dots & \mathbf{c}^{S}_N \cdot \mathbf{c}^{S}_N \end{pmatrix} \tag{9}$$

where the l index has been suppressed as it is the same for all vectors in a given slice. Viewed as such, P_l is a Gram matrix between the coefficient vectors, $P_l = X_l^T X_l$ where $X_l = (\mathbf{c}^{\alpha}_1, \mathbf{c}^{\alpha}_2, \dots, \mathbf{c}^{S}_N)$. As P_l only contains information on the relative orientations of the $\mathbf{c}^{\alpha}_n$, the $\mathbf{c}^{\alpha}_n$ can, at best, be recovered from P_l up to a global rotation in the $2l+1$ dimensional space. A direct consequence of this is that simultaneously rotating all coefficient vectors with equal l index does not affect the power

spectrum, so that there are many different densities which share the same power spectrum. Fortunately, this issue is mitigated as atomic descriptors do not have to distinguish all possible infinite dimensional densities, but only atomic densities constructed from $\mathcal{O}(10)$ neighbour atoms as in Eq. (1). Furthermore, we are typically interested in physically relevant atomic configurations where the atomic nuclei do not overlap, which further restricts the region of configuration space that must be described. In combination, these effects make the problem of finding concise descriptors for atomic densities significantly more tractable than for general densities.

Returning to the power spectrum, it is interesting that whilst P_l is an $NS \times NS$ matrix, its rank is at most $2l+1$, as the rank of the $(2l+1) \times NS$ matrix X_l is at most $2l+1$. This means that provided the first $2l+1$ columns in X_l form a basis, then all terms in P_l can be recovered from only the first $2l+1$ rows. This observation suggests that the power spectrum is amenable to non-data-driven lossless compression when $2l+1 < NS$. For situations with many different elements $NS \gg 2l+1$ so that the potential compression factor is large. Rather than simply storing the first $2l+1$ rows, we propose storing $2l+1$ random linear combinations of all rows,

$$W^T P_l = \begin{pmatrix} \mathbf{b}_1 \cdot \mathbf{c}_1 & \mathbf{b}_1 \cdot \mathbf{c}_2 & \dots & \mathbf{b}_1 \cdot \mathbf{c}_{NS} \\ \mathbf{b}_2 \cdot \mathbf{c}_1 & \mathbf{b}_2 \cdot \mathbf{c}_2 & \dots & \mathbf{b}_2 \cdot \mathbf{c}_{NS} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}_{2l+1} \cdot \mathbf{c}_1 & \mathbf{b}_{2l+1} \cdot \mathbf{c}_2 & \dots & \mathbf{b}_{2l+1} \cdot \mathbf{c}_{NS} \end{pmatrix} \tag{10}$$

where W is an $NS \times (2l+1)$ matrix of randomly chosen weights.

A set of density expansion coefficients consistent with the original power spectrum can then be recovered from $W^T P_l$ by diagonalising $W^T P_l W = U^T \Lambda U$, taking $(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{2l+1}) = \Lambda^{\frac{1}{2}} U$ and then forming $X_l = \Lambda^{-\frac{1}{2}} U W^T P_l$, where U is a unitary matrix whose columns are the eigenvectors of $W^T P_l W$ and Λ is a diagonal matrix of the eigenvalues[34]. This procedure fails if $(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{2l+1})$ does not form a basis, so that Λ has one or more zero eigenvalues. This could arise because $\mathrm{rank}(W^T X_l) < \mathrm{rank}(X_l) \leq 2l+1$, which is highly unlikely because of the random weights, or because $\mathrm{rank}(W^T X_l) = \mathrm{rank}(X_l) = r < 2l+1$, which occurs frequently as explained below. From a recovery perspective the latter is not problematic as the same procedure can be carried out using $(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r)$ as a basis by discarding rows and columns from $W^T P_l W$ as required.

By compressing all l-slices in this way only $NS(L+1)^2$ invariants need be stored, compared to $\frac{1}{2}NS(NS+1)(L+1)$ for the uncompressed power spectrum (Fig. 1). Interestingly, slightly more compression can be achieved by forming the symmetric matrix $Q^T P_l Q$, where Q is an $NS \times NS$ matrix of random weights and then storing only the upper right triangular portion—the original power spectrum can still be recovered in an entirely analogous fashion. The symmetry of $Q^T P_l Q$ means that there are actually fewer invariants then density expansion coefficients; per l slice, there are $l(2l+1)$ fewer invariants, which corresponds exactly to the number of distinct rotations in $2l+1$ dimensions. However, the cost of this additional compression is a loss of sparsity. Whereas only the entries of $W^T P_l$ (and P_l) corresponding to elements present in the local atomic environment will be non-zero, $Q^T P_l Q$ will be a dense matrix as all the $\mathbf{c}^a_{nl}$ are mixed together. Retaining this sparsity is exceptionally important for efficient descriptor storage as often $S \gg S_{\mathrm{env}}$, where S is the total number of elements present in a dataset whereas S_{env} is the typical number present in any given environment.

Finally, we note that there is an additional restriction on the rank of P_l, namely $\mathrm{rank}(P_l) \leq n_{\mathrm{neighb}}$ where n_{neighb} is the number of neighbouring atoms contained within the cutoff. This occurs because the density expansion vectors for the density of a *single neighbour atom* corresponding to different radial basis functions are all parallel. This is shown explicitly in the Supporting Information but can be understood by noting that the projections of a Gaussian onto different radial shells will vary only in magnitude, so that the angular part of the expansion, which determines the direction of $\mathbf{c}^a_{nl}$, will be identical. This observation is consistent with intuition—the information content should depend on the number of neighbours—and may prove useful when constructing concise descriptors for datasets where n_{neighb} is known to be bounded.

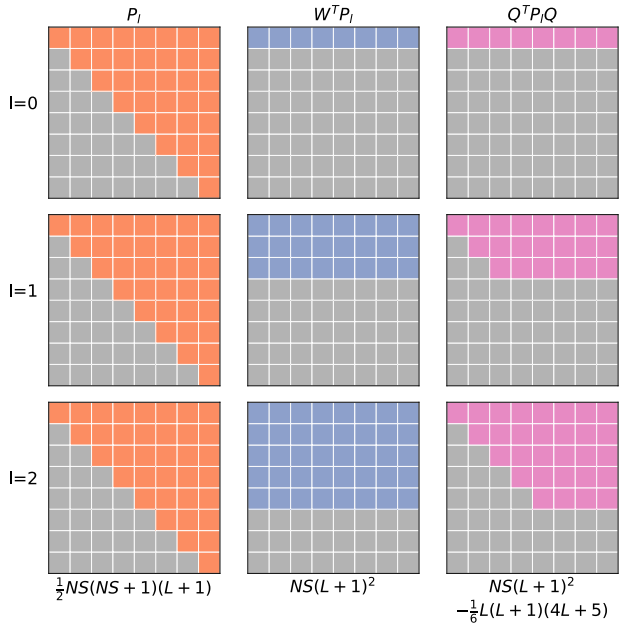

Fig. 1 Illustration of the $W^T P_l$ compression scheme. Moving from left to right P_l, $W^T P_l$ and $Q^T P_l Q$ are shown for $l = 0, 1, 2$ and $NS = 8$. Typically NS will be significantly larger than shown here. Only the elements shown in colour need to be stored to determine all remaining elements which are shown in grey. The total length of each descriptor is listed underneath.

Generalised kernel

In the previous section we showed that it is possible to compress the SOAP power spectrum in a lossless manner. Here we introduce a family of physically interpretable compression options based on generalising the SOAP kernel. The first step is to generalise Eq. (2) to

$$k(\rho, \rho') = \int d\hat{R} \left| \sum_\alpha \int d\mathbf{r} \rho^\alpha(\mathbf{r}) \rho'^\alpha(\hat{R}\mathbf{r}) \right|^\nu \left| \int d\mathbf{r} \rho(\mathbf{r}) \rho'(\hat{R}\mathbf{r}) \right|^\mu \tag{11}$$

where the first factor is as before and the second factor involves the total density $\rho(\mathbf{r}) = \sum_\alpha \rho(\mathbf{r})^\alpha$. The advantage of this approach is that the total body order is now partitioned into element-sensitive and element-agnostic terms, so that the scaling with S can be controlled separately from the overall body-order. For instance, $\nu, \mu = 1, 1$ results in a modified power spectrum $p^\alpha_{nn'l}$ with a length that scales linearly with S and is related to the original power spectrum, $\nu = 2$, via,

$$p^\alpha_{nn'l} = \sum_m c^{\alpha *}_{nlm} \left(\sum_\beta c^\beta_{n'lm} \right) = \sum_\beta p^{\alpha\beta}_{nn'l} \tag{12}$$

The $\nu, \mu = 1, 1$ power spectrum still corresponds to a histogram of triangles, but now only one vertex of each triangle is element-sensitive, as shown in Fig. 2. Clearly, this idea can be applied just as well for higher body orders to provide compression with respect to S. In the previous section we also achieved compression with respect to N. Following a similar approach as for S, we further

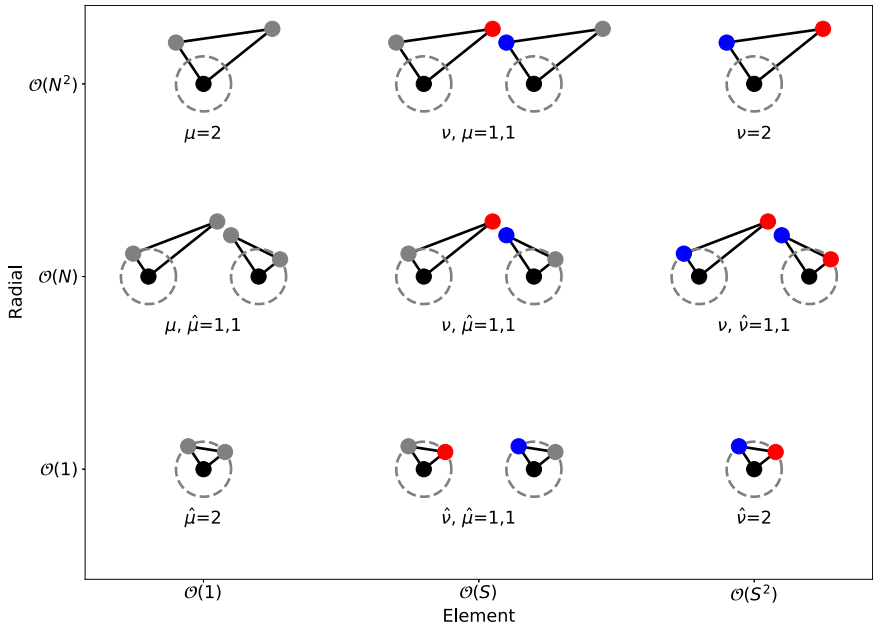

Fig. 2 Generalised SOAP power spectrum. Schematic showing the physical interpretation of the generalised power spectrum for various choices of ν, $\hat{\nu}$, μ and $\hat{\mu}$. For such 3-body terms $\nu+\hat{\nu}+\mu+\hat{\mu}=2$; indices which are zero are not listed. The vertices shown in blue and red are element-sensitive whilst those shown in grey are not. The grey dashed line indicates the unit sphere. When the projection results in two distinct triangles both are shown, otherwise only one is shown.

generalise the kernel to

$$k(\rho,\rho') = \int d\hat{R}\left|\int d\mathbf{r}\rho(\mathbf{r})\rho'(\hat{R}\mathbf{r})\right|^{\mu}\left|\int d\mathbf{r}\hat{\rho}(\hat{\mathbf{r}})\hat{\rho}'(\hat{R}\hat{\mathbf{r}})\right|^{\hat{\mu}} \left|\sum_{\alpha}\int d\mathbf{r}\rho^{\alpha}(\mathbf{r})\rho'^{\alpha}(\hat{R}\mathbf{r})\right|^{\nu}\left|\sum_{\alpha}\int d\mathbf{r}\hat{\rho}^{\alpha}(\hat{\mathbf{r}})\hat{\rho}'^{\alpha}(\hat{R}\hat{\mathbf{r}})\right|^{\hat{\nu}} \tag{13}$$

where $\hat{\rho}(\hat{\mathbf{r}})$ is the projection of the density onto the surface of the unit sphere. As before, Eq. (13) still corresponds to comparing all possible triangles for $\nu+\hat{\nu}+\mu+\hat{\mu}=2$. However now some of the vertices may be projected onto the surface of the unit sphere, as well as potentially being element insensitive. The full range of 3-body compression options is depicted in Fig. 2, with the interpretation of the original power spectrum, $\nu=2$ (only non-zero index values are listed), shown in the upper right corner.

Utilising these compressions offers a significant reductions in descriptor size but does result in a descriptor that is typically less informative. An example of such an information loss is shown in Fig. 3, where the two environments shown are distinguished by $\nu=2$ but not by $\mu,\nu=1,1$, as now there is only 1 element-sensitive vertex per triangle, rather than two. Returning to the idea of the previous section, we see that $\nu,\mu=1,1$ corresponds to using $\{\mathbf{C}_n\}$, where $\mathbf{C}_n=\sum_\alpha \mathbf{c}_n^\alpha$, in place of the random basis $(\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_{2l+1})$. In this case, the mirror symmetry of the total density means that $\text{rank}(\mathbf{C}_1,\mathbf{C}_2,\ldots,\mathbf{C}_N)=2<\text{rank}(\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_{NS})=4$, so that the full power spectrum cannot be recovered from the compression. However, this pair of environments are a handcrafted example and more generally, in the absence of similar symmetries, we would expect $\{\mathbf{C}_n\}$ to span the required space, provided $N\geq 2l+1$, so that the compression would be lossless. This line of thought motivates the introduction of a further compression, denoted $\nu,\mu^*=1,1$, where the kernel is as defined in Eq. (11), but $2l+1$ radial basis functions are used for the total density expansion. This choice achieves the same level of compression as W^TP_l and, provided $\text{rank}(\mathbf{C}_1,\mathbf{C}_2,\ldots,\mathbf{C}_{2l+1})=\text{rank}(\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_{NS})$, it will also be lossless. Such arguments suggest that compressing beyond $\nu,\mu^*=1,1$ will necessarily be lossy, with the analysis for higher body orders being left to future work. We also note that $\nu,\mu^*=1,1$ is slightly faster to compute than W^TP_l as fewer operations are required to form $\mathbf{C}_n$ than $\mathbf{b}_i$. However, the difference is small and the latter is more flexible as it can be applied when $NS\geq 2l+1$ whilst the former requires $N\geq 2l+1$. At first glance the degeneracies introduced by choosing $\nu,\mu=1,1$ appear catastrophic. However, it should be noted that the original power spectrum $\nu=2$ has similar degeneracies. A simple, single-element example was given in ref. [35], which is trivially modified to contain multiple elements in Supplementary Fig. 1. Despite this, the full $\nu=2$ power spectrum has been widely used with great success across a wide variety of applications[6,36,37], as have other descriptors which are known to be mathematically incomplete[32,38]. As such, whilst it is useful to understand the origin of any additional degeneracies in compressed descriptors, the most practically interesting question is whether or not they lead to a significant degradation in model performance across typical datasets. We assess this for the variants of the SOAP power spectrum listed in Table 1 by performing numerical tests which are discussed in the results section.

Finally, whilst we only test compressions of the power spectrum in this work, these compression ideas can be applied just as easily to higher body-orders. For instance, choosing $\nu,\hat{\nu},\mu=1,1,1$ would result in a compressed version of the bispectrum that scales quadratically with S and N, rather than cubically. In general, the length of the descriptor will scale as $S^{\nu+\hat{\nu}}N^{\nu+\mu}$ with the total body order given by $\nu+\hat{\nu}+\mu+\hat{\mu}$, so that the scaling with S and N can be chosen independently from the overall body-order. We anticipate these compressions being particularly useful for descriptors such as ACE[39,9] where $\nu\geq 4$ is commonly used.

Element embedding

An alternative approach to constructing concise descriptors for systems with large S was used in refs. [18,19,40]. Rather than using a separate density for each element they instead used two density channels; the total, element agnostic, density and an element weighted density defined as $\rho^Z(\mathbf{r})=\sum_\alpha w_\alpha\rho_\alpha(\mathbf{r})$ where w_α is an element dependent weight. The descriptor can then be formed using these two density channels, so that the length of the power spectrum is $N(2N+1)(L+1)$, independent of S.

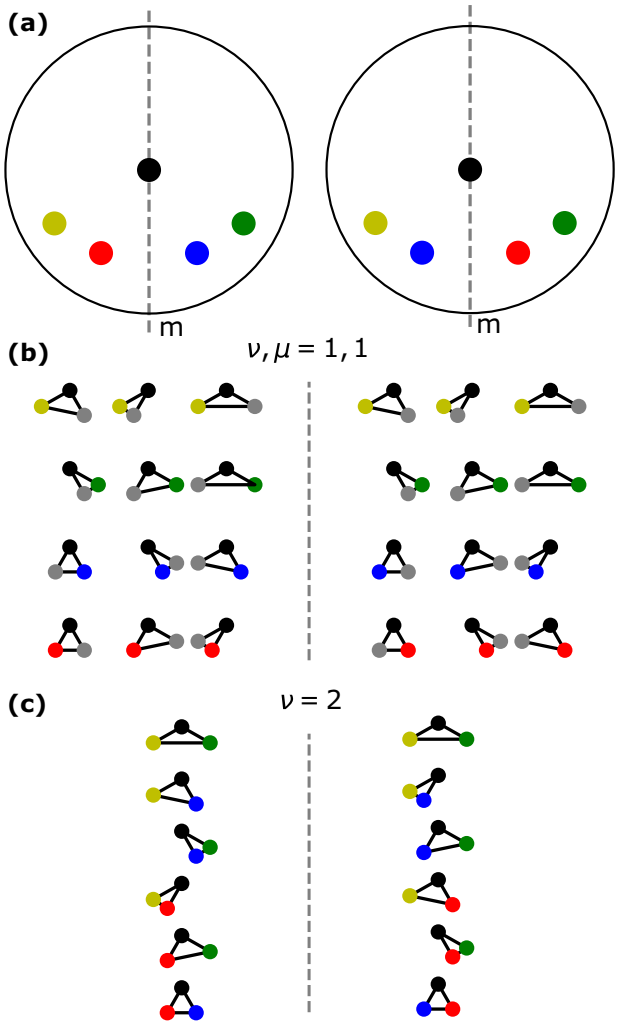

Fig. 3 Pair of degenerate environments. Two environments which are distinct using $\nu = 2$ but identical according to $\nu, \mu = 1, 1$, because of the mirror symmetry of the total density, are shown. Elements are distinguished by colour whilst the histograms of triangles shown on the right - element agnostic vertices are shown in grey.

The element-weighted power spectrum is an instance of a more general type of constant complexity approach, where each element α is represented by a vector $\mathbf{u}^\alpha$, where $\dim(\overrightarrow{u_\alpha}) = d_J < S$, so that the chemical elements are effectively embedded into a lower dimensional space. The $\mathbf{u}^\alpha$ can then be optimised during model fitting so that the alchemical similarity between different elements, $k^{\alpha\beta} = \mathbf{u}_\alpha \cdot \mathbf{u}_\beta$[21], is learned from the data. This approach was used in refs. [21,20], where in both cases the optimised embedding was consistent with known chemical trends, and more recently, similar, learnable mappings have been used in refs. [41,42] and [43–45].

This approach was taken further still in ref. [26] where PCA was used to determine a reduced optimal radial basis for a given dataset. By allowing basis changes, followed by truncation's, that also mix different elemental channels this approach can be seen as a simultaneous embedding of both the elemental and radial information into a lower dimensional space. Interestingly, the random weight matrix W in Eq. (10) can be interpreted as performing an analogous embedding, $X_l = (\mathbf{c}_1^a, \mathbf{c}_2^a, \ldots, \mathbf{c}_N^S) \to (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{2l+1}) = X_l W$. This identification connects the approaches and motivates further work where W is optimised for a given dataset.

Using element embedding, with fixed or optimisable embedding vectors, offers an alternative route to reducing the scaling with S and we include a corresponding compressed power spectrum, denoted as $\nu = 2, d_J$, in our numerical tests for comparison. A summary of all ways of compressing the power spectrum is given in Table 1. An alternative way of evaluating the full SOAP kernel is outlined in the next section with the numerical results following afterwards. The datasets used are described briefly in 'Datasets'.

Alternative evaluation of the SOAP kernel

For typical (KRR) SOAP models a significant fraction of the time taken to evaluate the model is spent computing the SOAP kernel between pairs of environments. The kernel is typically evaluated as $k(\rho, \rho') = \mathbf{p} \cdot \mathbf{p}'$ where computing the dot product requires $\mathcal{O}(N^2 S^2 L)$ operations. An alternative way to evaluate the kernel is

$$k(\rho, \rho') = \mathbf{p} \cdot \mathbf{p}' = \sum_{\alpha\beta nn' l} p^{\alpha\beta}_{nn'l} p'^{\alpha\beta}_{nn'l} \tag{14}$$

$$= \sum_{\alpha\beta nn' l} \sum_m c^{*\alpha}_{nlm} c^{\beta}_{n'lm} \sum_{m'} c'^{\alpha}_{nlm'} c'^{*\beta}_{n'lm'} \tag{15}$$

$$= \sum_{lmm'} \left(\sum_{an} c^{*\alpha}_{nlm} c'^{\alpha}_{nlm'} \right) \left(\sum_{\beta n'} c^{\beta}_{n'lm} c'^{*\beta}_{n'lm'} \right) \tag{16}$$

$$= \sum_{lmm'} \left| \sum_{an} c^{*\alpha}_{nlm} c'^{\alpha}_{nlm'} \right|^2 \tag{17}$$

where evaluating the final line requires $\mathcal{O}(NSL^3)$ operations, so becomes competitive for $NS > L^2$. The evaluation of a single kernel could also employ both methods, with this approach used only for terms where $NS \gtrsim 2(2l+1)^2$. An advantage of computing kernels in this way is the reduction in the memory requirement, as only the density expansion coefficients need be stored. For completeness we note that the matrix form of the above is $k(\rho, \rho') = \sum_l \mathrm{tr}\left(X_l^\dagger X_l X_l'^\dagger X_l'\right) = \sum_l ||X_l' X_l^\dagger||_2^2$, where X_l is defined as in 'Information Content'.

Distance–distance correlation and information imbalance

The similarity between a pair of environments can be quantified by the Euclidean distance between their respective descriptor vectors, so that different descriptors give different measures of similarity. These distances can be used to quantify the relative information content of descriptors[22] and, more generally, to compare how different descriptors encode a given dataset. We follow ref. [46] using distance-distance correlation plots to compare the distances implied by the compressed descriptors to those of the full power spectrum, where it is desirable for a compression to preserve the distances as faithfully as possible. In Fig. 4 the distance–distance (left) and ranked distance-ranked distance (right) correlations between the $\nu = 2$ power spectrum and two of the compressed alternatives are shown for all pairs of environments within liquid configurations, $S \geq 3$, in the HEA dataset. The advantage of comparing the ranked distances is that the ranking process eliminates scaling and monotonic transformations of the distances, leaving only the correlation structure behind[22,47].

The correlations between both the distances and ranked distances for the $\nu, \mu = 1, 1$ compression and $\nu = 2$ are strong, with a notable absence of points in the top left and bottom-right of each plot. The same is not true of both the constant complexity alternatives, where there are many environments deemed well separated by $\nu = 2$ but which are poorly distinguished by

Table 1. Overview of compression schemes.

Label	References	Descriptor	Length
$\nu = 2$	Full SOAP kernel[17,33]	$p^{\alpha\beta}_{nn'l} = \sum_m c^{\alpha *}_{nlm} c^{\beta}_{n'lm}$	$\frac{1}{2} NS(NS+1)(L+1)$
$\nu, \mu = 1, 1$	-	$p^{\alpha}_{nn'l} = \sum_m c^{\alpha *}_{nlm} \left(\sum_\beta c^{\beta}_{n'lm} \right)$	$N^2S(L+1)$
$\nu, \mu^* = 1, 1$	-	$p^{\alpha}_{nn'l} = \sum_m c^{\alpha *}_{nlm} \left(\sum_\beta c^{\beta}_{n'lm} \right)$	$NS(L+1)^2$
$\nu, \hat{\nu} = 1, 1$	Similar to $p^{\alpha\beta}_{1n'l} = \sum_m c^{\alpha *}_{1lm} c^{\beta}_{n'lm}$ used in TurboGAP[56,57]	$p^{\alpha\beta}_{nl} = \sum_m c^{\alpha *}_{nlm} \left(\sum_{n'} c^{\beta}_{n'lm} \right)$	$NS^2(L+1)$
$\nu, \hat{\mu} = 1, 1$	-	$p^{\alpha}_{nl} = \sum_m c^{\alpha *}_{nlm} \left(\sum_{n'\beta} c^{\beta}_{n'lm} \right)$	$NS(L+1)$
$\mu = 2$	Element agnostic SOAP kernel[17]	$p_{nn'l} = \sum_m \left(\sum_\alpha c^{\alpha *}_{nlm} \right) \left(\sum_\beta c^{\beta}_{n'lm} \right)$	$\frac{1}{2} N(N+1)(L+1)$
$\nu = 2$, d_J	Element weighting[18] Element embedding[21]	$p^{JJ'}_{nn'l} = \sum_m c^{J *}_{nlm} c^{J'}_{n'lm}$	$\frac{1}{2} Nd_J(Nd_J+1)(L+1)$
—	Optimal Radial Basis[26] with mixed-species basis	$p_{uu'l} = \sum_m c^{*}_{ulm} c_{u'lm}$	$\frac{1}{2} u_{\max}(u_{\max}+1)(L+1)$
$W^T P_l$	-	$W^T P_l$	$NS(L+1)^2$

Various compressions of the SOAP power spectrum are listed, N, S, L, d_J and $u_{\max}$ are the number of radial basis functions, elements, maximum order of spherical harmonics, number of embedding dimensions and the number of optimised radial basis functions respectively. The circumflex indicates projection onto the unit sphere so that $\hat{\rho}(\mathbf{r}) = \hat{\rho}(\hat{\mathbf{r}})$. The kernel for $\nu, \mu^* = 1, 1$ is identical as for $\nu, \mu = 1, 1$, except that $2l+1$ radial basis functions are used in the total density expansion, so that $n' = 1, 2, \ldots, 2l+1$.

$\nu, d_J = 2, 2$ (and $\mu = 2$). An example of such an environment is shown in Supplementary Fig. 2, with the distances according to each descriptor listed alongside. The most striking difference is seen in the ranked-distance correlation plots which are near uniform for $\nu, d_J = 2, 2$ (and $\mu = 2$), demonstrating that the original correlation structure is almost completely lost.

We also compute the information imbalance (introduced in ref. [22]) between the different descriptors; the information imbalance is a way of measuring the relative information content of different distances measures, see 'Information Imbalance' for details. The information imbalance planes for the HEA and amino acid datsets are shown in Figs. 5 and 6, where points in the bottom left, top left and along the diagonal indicate descriptors which encode the same, less and orthogonal (different) information to $\nu = 2$ respectively. The results for the HEA dataset provide quantitative evidence for the trends seen in Fig. 4, demonstrating that for this dataset, the $\nu, d_J = 2, 2$ (and $\mu = 2$) descriptors are significantly less informative than the others. Conversely, these descriptors carry almost identical information to $\nu = 2$ on the amino acid dataset. We believe this is because all the amino acid molecules are geometry optimised, so that the atom type information can be inferred from the atomic positions alone; this is backed up by $\Delta(\mu = 2 \to \nu = 2) \ll 1$. The stark differences seen between these two datasets suggests that whilst low-dimensional element embeddings are undoubtedly useful, they are not suitable as information preserving compressions for all datasets.

Fitting to energies

The ultimate test of any descriptor is the accuracy and overall performance of the models which use it. In this work we test the proposed SOAP compressions by fitting interatomic potentials to total energies for four separate datasets using both Ridge Regression (RR) and KRR models, see 'KRR Models' for details.

Learning curves for all models and datasets are shown in Fig. 7 whilst the total length of each descriptor is shown in the bottom left of each plot. It is important to note that all descriptors other than $\nu = 2, d_J$ are sparse with respect to the elements, so that in practice their storage requirements scale with S_{env}, the typical number of elements present within each environment, rather than S_{total}, the total number of elements present in the dataset. For the elpasolite dataset $S_{\text{env}} = 4$ for all structures, so in this case the lengths of the sparse descriptors are indicated in white. However, whilst the descriptors themselves are sparse the number of model parameters is dependent on the full length of the descriptor. As such, we stress that both sparsity in S_{env} and reducing the scaling with respect to S_{total} are highly desirable. As a final point, the amino acids and Li-TM datasets were also used in ref. [18] to test the accuracy achievable combining a constant complexity descriptor and a neural network potential. We stress that the errors reported there are on the training set, acting as a proof of principle, whilst the larger errors reported here are on a distinct test set, so that the numbers are not comparable.

The trends seen across the QM9, Li-TM and amino acid datasets are broadly similar, with $\nu, \mu = 1, 1$ achieving the same accuracy as $\nu = 2$ with a significantly reduced descriptor size. The results with $W^T P_l$ and $\nu, \mu^* = 1, 1$ are ~5–15% less accurate then with $\nu = 2$ but, crucially, both these descriptors offer additional compression with respect to the radial channels and, as outlined in the previous section, contain sufficient information to recover the full power spectrum under certain conditions. Compressing beyond the recoverable limit with $\nu, \hat{\mu} = 1, 1$—projecting the element agnostic vertex onto the unit sphere—offers an additional factor of ~5 in terms of compression. On the elpasolite dataset this does not compromise the accuracy at all whilst for QM9 this compression incurs a 66% increase in the MAE, which is limited to ~30% for both the Li-TM's and amino acids. Comparing these results to the errors achieved using $\mu = 2$ and $\nu = 2$ hints at the relative importance of atom type and geometric information for the different datasets. Unsurprisingly, knowledge of the atom types is much more important for the elapsolites, where all structures are in the same crystal prototype, whilst for the chemically reasonable organic molecules found in the QM9 and the amino acid datasets a reasonable model can be fit using geometric information alone —MAE of 0.62 kcal mol^{-1} for $\mu = 2$ on QM9. This is consistent with the information imbalance analysis on the amino acids and suggests that for these molecules, in their equilibrium geometries, it is possible to use known bond lengths and coordination numbers to infer the atom type information from the geometry alone.

The (unoptimised) embedding $\nu, d_J = 2, 2$ performs relatively well with errors 27%, 57% and 26% larger then $\nu = 2$ for the QM9, Li-TM, and amino acids datsets respectively. This performance is most comparable to $\nu, \hat{\mu} = 1, 1$, which, particularly after exploiting sparsity, offers a greater level of compression. A clear, qualitative difference in behaviour is seen for the embedding approaches on the elpasolite dataset, which contains a much larger number of elements. Here, the unoptimised embedding performs no better than using $\mu = 2$ and has a greatly diminished final learning rate compared to $\nu = 2$. For comparison, the optimised embeddings, denoted using a $\overline{d}_J$, from ref. [21], which use similar SOAP parameters ($r_{\text{cut}} = 5$Å, $N = 12$, $L = 9$), are overlaid. This shows that optimising the embedding with only two dimensions leads to minimal improvement although a significant gain can be made

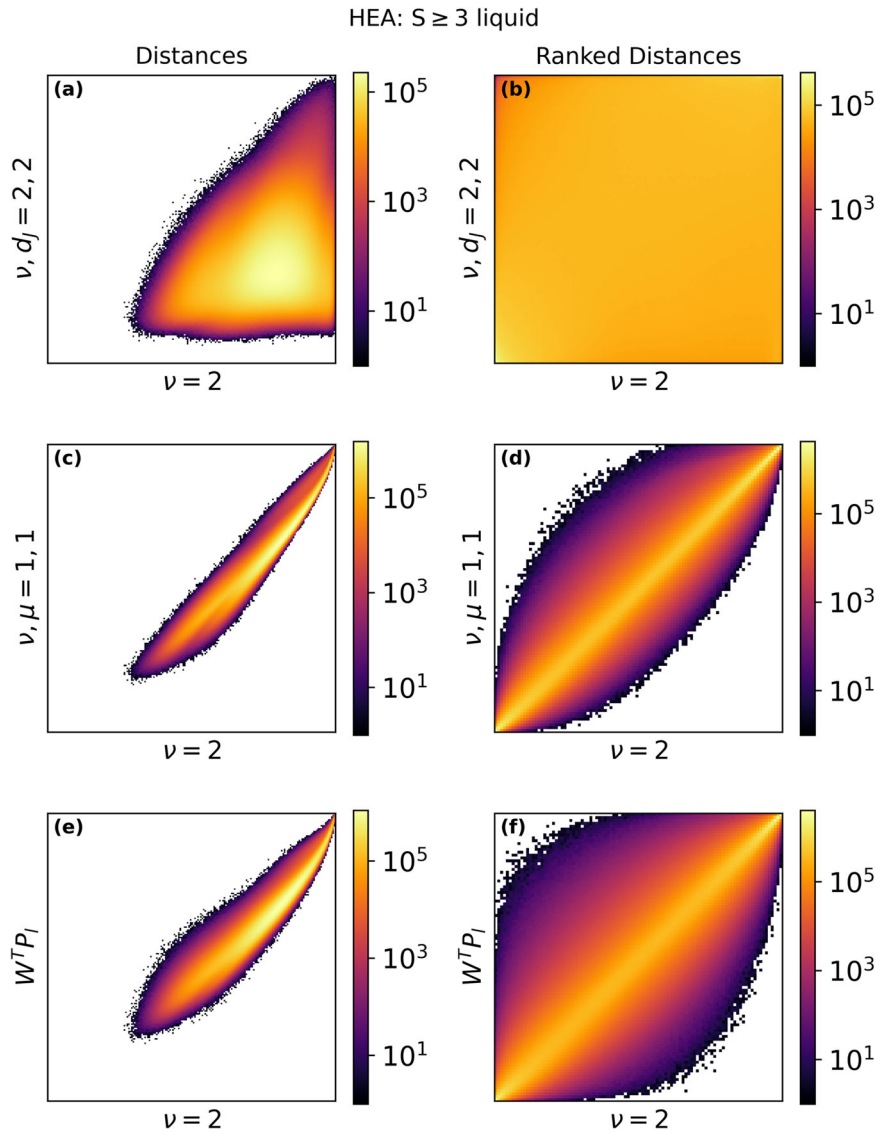

Fig. 4 Comparison between descriptor distances. The correlation between the distances (left column) and ranked distances (right column) between all pairs of environments within liquid configurations with $S \geq 3$ in the HEA dataset are shown. See Supplementary Fig. 5 for the correlations for $\mu = 2$, $\nu, \hat{\mu} = 1, 1$ and $\nu, \mu^* = 1, 1$.

using $\overline{d}_J = 4$. However, the final learning rate for $\overline{d}_J = 4$ is still diminished, relative to $\nu = 2$, again indicating a qualitative difference in behaviour from the compressions which keep separate densities for each element. It is also worth noting that as $S_{\text{env}} = 4$ for the elpasolites, using $d_J = 4$ does not offer any descriptor compression compared to $\nu = 2$ once sparsity is exploited. The embedding approaches do however offer a clear advantage in the low-data regime, which we believe is due to the reduced dimensionality of the descriptor space. If each data point occupies a certain volume of descriptor space then such embeddings greatly increase the relative fraction of descriptor space occupied by physically reasonable configurations. As such, the relevant descriptor space is covered faster so that shortest distance from a test data point to any point in the training set decreases faster with new training data.

Fitting a force-field

The compression options illustrated in Fig. 2 were tested further by training sparse KRR models on energies, forces and stresses for the HEA dataset using the gap_fit program[48], with the same parameters used in ref. [49]. This dataset is of particular interest as in previous work, ref. [49], it was found that using a model fitted using a 2-body + 3-body descriptor performed better than using a 2-body descriptor + the full multi-element power spectrum, $\nu = 2$. This result is explained by the power spectrum being a higher dimensional descriptor, so that much more training data is required to span the full descriptor space corresponding to the relevant atomic environments. In the low data regime, i.e. before this condition is satisfied, the model will generalise very poorly outside of the training set. However, provided sufficient data, one would expect a model trained on a more informative descriptor to provide greater accuracy. This effect can be seen in the left hand panel of Fig. 8 where the force errors for $\nu, \hat{\mu} = 1, 1$ are initially better than for $\nu = 2$, but then plateau quickly. In contrast the errors for the $\nu = 2$ model continue to decrease, however, here the very large descriptor meant that only 4000 sparse points could be used due to a memory restriction of 1.5 TB of RAM.

Interestingly the best model, subject to the practical restrictions on memory and quantity of training data, makes use of the

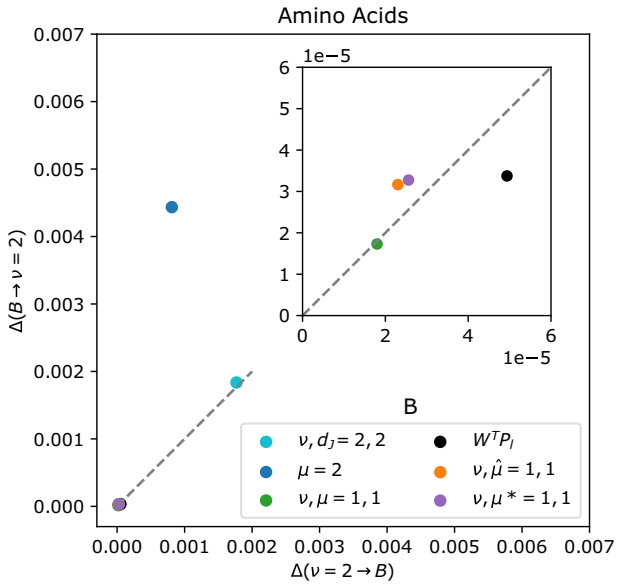

Fig. 5 Information imbalance plane for amino acids. Information imbalance plane for atomic environments in the amino acid dataset. The central atom was not included in the density expansion.

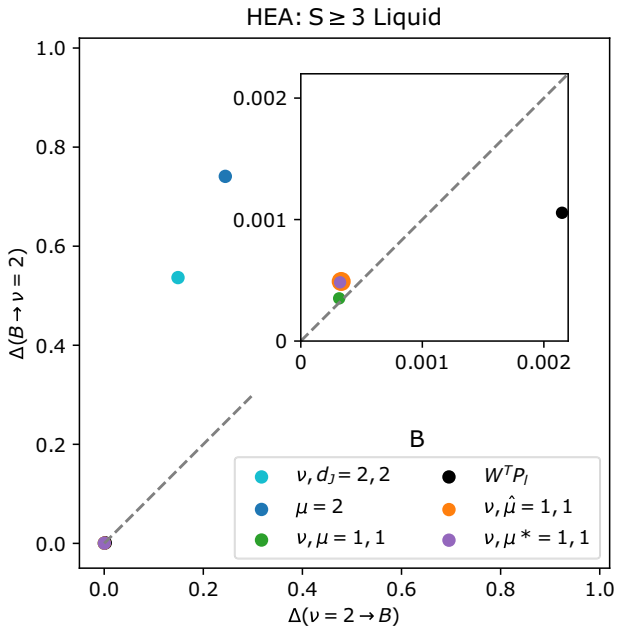

Fig. 6 Information imbalance plane for HEA liquid. Information imbalance plane for the environments from liquid configurations with $S \geq 3$ in the HEA dataset. Note the difference in scale relative to the amino acids.

$\nu, \mu = 1, 1$ compression. Whilst the increase in performance, compared to the 2b + 3b, is modest across the entire test set it is much more dramatic on the quinary alloy liquid configurations where the energy RMSE is reduced from 96 meV atom^{-1} to 12.3 meV atom $^{-1}$, Supplementary Fig. 3.

To assess whether the effectiveness of the compression schemes depends on the diversity of the dataset we also fit models using a subset of the original dataset where the liquid, surface and defect configurations were removed. The trends seen in Fig. 8 are replicated in Supplementary Fig. 6, suggesting that the applicability of these compression schemes is not strongly dependent on the intrinsic dimensionality of the dataset.

Sensitivity analysis

When using atomic descriptors for tasks such as regression it is highly desirable that they are sensitive to small perturbations of a given atomic environment, with in depth analysis of well known descriptors carried out in refs. [23,46]. The latter introduced the concept of the sensitivity matrix Λ, see 'Sensitivity Matrix', where the eigenvalues λ can be used to determine if there are any perturbations of the local environment which a descriptor is insensitive to.

The results in Fig. 9 and Supplementary Fig. 4 show that the proposed compressions do not significantly affect the sensitivity. In particular, no perturbations were found for which any of the compressed descriptors had near-zero sensitivity, whilst the original $\nu = 2$ did not. This behaviour is not entirely unexpected, as for instance, choosing $\mu = 2$ is equivalent to using the original power spectrum with a single element environment. Perhaps more surprising is that using $\nu, \hat{\mu} = 1, 1$—projecting one atom onto the unit sphere—did not reduce the sensitivity more drastically. Of course, these tests are not exhaustive and it is probable that special environments, likely closely related to any additional degeneracies, exist where this is not the case[50]. However, we find these results promising and leave further investigation to future work.

Atom-centred symmetry functions

Whilst the compression presented in 'Information Content' is specific to the SOAP power spectrum, the ideas used to form the generalised SOAP kernel are equally applicable to all body-ordered descriptors. To demonstrate this we fit KRR models to the QM9 and Li-TM datasets using compressed versions of the popular ACSFs introduced in ref. [32]. We also fit models based on element agnostic and element-weighted ACSFs[18,40] for comparison. As before a polynomial kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)^\zeta$, was employed and the regularisation strength was chosen using k-fold cross-validation with $k = 10$. The G2 (2-body) and G4 (3-body) ACSF functions were used with the traditional parameters[23,32] scaled to cutoffs of 4 Å and 6 Å for the QM9 and Li-TM datasets respectively. Compression was applied to the 3-body terms only and the descriptors are labelled using the notation outlined in Fig. 2, so that $\nu = 2$ is the conventional $\overrightarrow{\mathrm{G4}}^{\alpha\beta}$ which has length $\mathcal{O}(S^2)$, $\nu, \mu = 1, 1$ corresponds to $\overrightarrow{\mathrm{G4}}^{\alpha} = \sum_\beta \overrightarrow{\mathrm{G4}}^{\alpha\beta}$ so has length $\mathcal{O}(S)$ and $\mu = 2$ is element agnostic, $\overrightarrow{\mathrm{G4}} = \sum_{\alpha\beta} \overrightarrow{\mathrm{G4}}^{\alpha\beta}$. For consistency with the existing literature we label the element-weighted ACSF as wACSF[40], rather than $\nu, d_J = 2, 2$. The dscribe[30] python package was used to compute the full ACSF descriptor, $\nu = 2$, and the compressed variants were formed by summing over elements as required.

Learning curves are shown in Fig. 10, where it can be seen that using the $\nu, \mu = 1, 1$ compression does not cause a noticeable decrease in model accuracy on either dataset. For the Li-TM dataset models were fit using $\zeta = 1$, 2, 4, 8 and 16 to assess the effect of varying the flexibility of the model. As expected, the shorter descriptors, $\mu = 2$ and wACSF, benefited more from increasing ζ but still never achieved comparable accuracy to the others. More interestingly, using $\nu, \mu = 1, 1$ provided the same accuracy as $\nu = 2$ even with $\zeta = 1$, despite the full descriptor being more than 5× as long. These results clearly show how the ideas behind the generalised SOAP kernel can be successfully applied to other body-ordered descriptors.

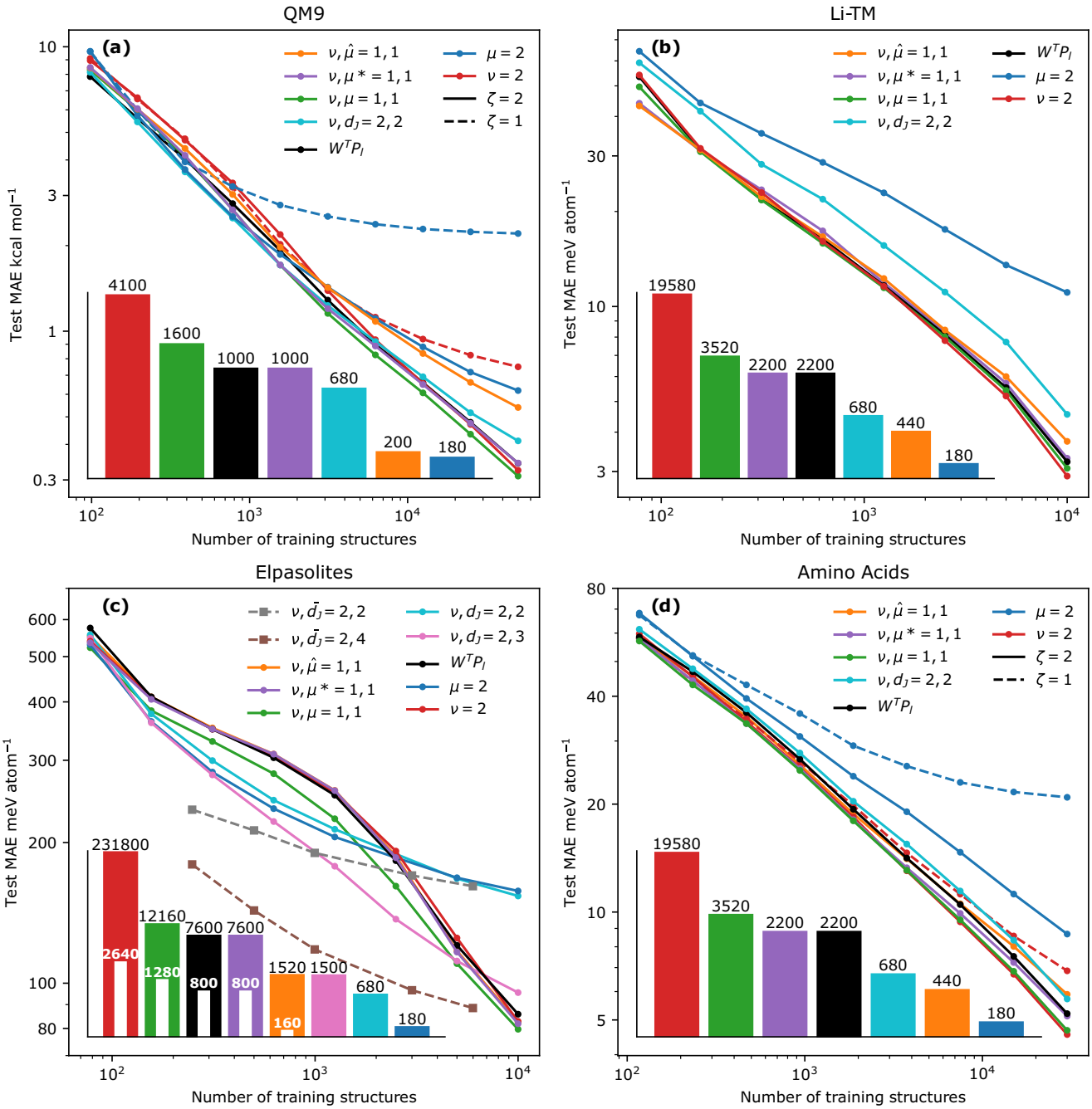

Fig. 7 Learning curves for the total energy fits. For the Li-TM dataset $\zeta = 1$, 2, 4 and 8 were all tested but only the models with $\zeta = 8$ are shown. The $\overline{d}_J = 2$ and $\overline{d}_J = 4$ curves for the Elpasolites are taken from ref. [21] where the embedding was optimised. All models for the Elpasolites were fitted using $\zeta = 1$. The total length of the descriptors is indicated in the bar chart. The overlaid white bars indicate the number of non-zero elements present, computed using $S_{\text{env}} = 4$.

DISCUSSION

As the number of elements S increases, the length of many atomic descriptors increases drastically, with S^ν scaling common for $\nu + 1$-body order descriptors. In this work we have sought non-data-driven ways to reduce this scaling, with a focus on the SOAP power spectrum. We started by investigating the degree to which the density expansion coefficients can be recovered from the power spectrum. This analysis revealed that the power spectrum can be viewed as a collection of Gram matrices P_l, one for each total angular momentum number l, and that storing only a subset of components $W^T P_l$ is sufficient to preserve full information. This compression reduces the descriptor size from $\frac{1}{2}NS(NS+1)(L+1)$ to $NS(L+1)^2$, where N and L are the number of radial basis functions and highest order of spherical harmonic used in the density expansion respectively. Compressing the power spectrum in this way requires a single matrix of random weights W to be stored and used consistently to compress all descriptors for a given dataset.

Next, we introduced the generalised SOAP kernel. The standard power spectrum, $\nu = 2$, is a 3-body descriptor corresponding to a histogram of triangles, with one histogram for each pair of species. In the generalised power spectrum the number of triangle vertices which are element-sensitive can be varied, as can the number of vertices which are projected onto the surface of the unit sphere.

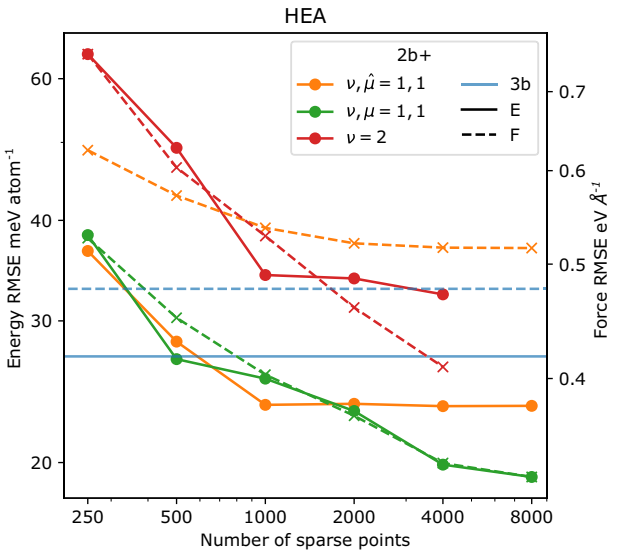

Fig. 8 Model errors on HEA test set. The left hand panel shows how the energy and force errors on the test set change with the number of sparse points. Note that the 2b + 3b model is from ref. [49] where it was reported that increasing the number of sparse points did not improve performance.

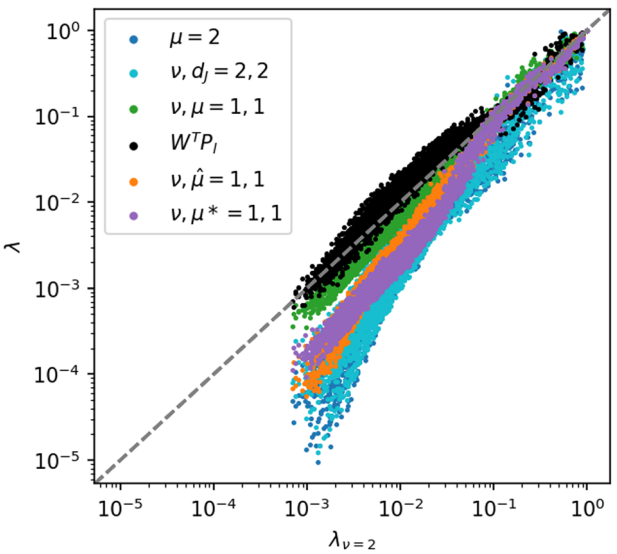

Fig. 9 Eigenvalues of the sensitivity matrix. A scatter plot showing eigenvalues, λ, of the sensitivity matrix for 50 randomly chosen liquid environments from the HEA dataset. The first six eigenvalues are not shown. The grey dashed line indicates $y = x$. The eigenvalues have been scaled, so that the largest eigenvalue for each environment is 1.

These modifications allow the scaling of the descriptor size with S and N to be set independently of the overall body-order and are applicable to all body-ordered descriptors. By again considering the ability to reconstruct the original power spectrum from the generalised one we showed that, subject to certain conditions, choosing $\nu, \mu^* = 1, 1$ (or $\nu, \mu = 1, 1$ where $N \geq 2l + 1$) does not lead to any loss of information. We also stress that descriptors based on the generalised kernel retain the element-wise sparsity of the original kernel, so that the number of non-zero components scales only with the number of elements present in a given environment, rather than total number present across a dataset.

The real-world performance of the compressions was tested using multiple numerical tests across a total of five pre-existing datasets. First, the information content, relative to the original power spectrum was analysed using the information imbalance approach of ref. [22]. This analysis indicated that retaining element sensitivity on only one vertex was sufficient to ensure a minimal loss of information across all tested datasets. The constant complexity compression approaches performed well for the geometry optimised amino acids but incurred severe information loss on the liquid environments within the high-entropy alloy dataset. We believe this is because the atom type information for the QM9 and amino acid datasets is effectively encoded in the equilibrium geometry of the molecules, which suggests that such datasets, in particular QM9, are not well suited to assess the ability of a given descriptor to encode multi-element information.

Secondly, models were fitted to the total energies for the QM9, Li-TM, amino acid and elpasolite datasets using linear and kernel ridge regression. The most promising compressions achieved very similar results to the full power spectrum across all datasets, whilst being significantly shorter. A notable deviation in behaviour was seen for the element-embedding fits to the elpasolite dataset, highlighting the differences between compression approaches. KRR models were also fitted to the QM9 and Li-TM datasets using compressed versions of ACSF's, where, as before, the errors achieved with the $\nu, \mu = 1, 1$ compression was almost identical to those achieved with the full descriptor. Following this, KRR models using the generalised SOAP power spectrum were fitted to energies, forces and virials for the HEA dataset using the gap_fit program[48]. The most accurate model that could be fitted, subject to practical restrictions on the quantity of training data and available memory, made use of the $\nu, \mu = 1, 1$ compression, providing concrete evidence that these compressions will be useful when fitting force-fields. Finally, the sensitivity of all descriptors to small perturbations was evaluated using the sensitivity matrix introduced in ref. [46]. None of the compressions were found to significantly reduce the sensitivity of the descriptor, which is unsurprising given their relation to the single element power spectrum.

We anticipate that these compressions will prove useful across a wide range of applications and that some of the ideas may be applicable to other body-ordered atomic descriptors, as was shown explicitly for ACSFs. In particular, the generalised kernel could provide compression for approaches such as ACE[9,39], where the high body-orders, $\nu \geq 4$, that are needed limit the number of different elements that can be treated. Furthermore, we stress that the compression ideas presented here can often be combined with pre-existing techniques, such as element-embedding, and that, in general, choosing the appropriate compression methods for a given situation is crucial.

METHODS

Datasets

The datasets used for the numerical tests are:

1. **QM9**: The QM9 dataset[51] contains ~140,000 geometry optimised organic molecules containing only H, C, N, O and F. In this work we fit the internal energy at 0 K, reported as U0, and hereafter refer to this as energy.
2. **HEA**: A quinary (Mo, Nb, Ta, V and W) high entropy alloy dataset from ref. [49] containing 2329 configurations including bulk crystals, surfaces, vacancies, alloys and liquid structures. The training set was used for the distance-distance correlation, information imbalance and sensitivity analysis whilst the independent test from ref. [49] was used to assess the accuracy of the force-fields.
3. **Li-TM**: This dataset from ref. [18] consists of 16,407 Lithium Transition Metal Oxides formed from 11 differnet elements, (Li, Ti, O, Mn, Ni, Sc, V, Cr, Mn, Fe, Co and Cu).
4. **Amino Acids**: A collection of 45,701 geometry optimised amino

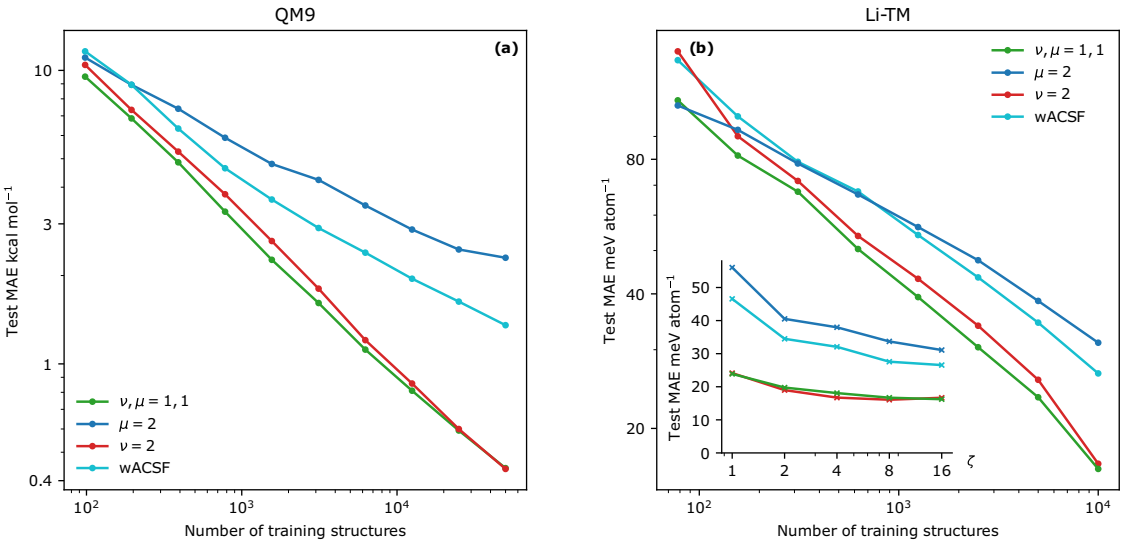

Fig. 10 Learning curves for ACSF energy models. Energy errors for KRR models with ACSF's as the atomic descriptors for the QM9 ($\zeta = 2$) and Li-TM ($\zeta = 16$ shown in main panel) datasets. The inset on the right shows how the error changed with ζ for a training set with 10^4 configurations.

acids from ref. [52] containing a total of 11 different elements. The dataset covers a total of 280 molecular systems - product of 20 proteingenic amino acids with 2 different backbone types (N-terminally acetylated or C-terminally amino-methylated) and 7 different cation additions (None, Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+} or Hg^{2+}).

5. **Elpasolites**: A collection of ~10,500 geometry optimised structures from ref. [53]. All main group elements up to Bi are present, 39 elements in total, and all structures share the elpasolite structural prototype.

Element embedding parameters

For the element embeddings, $d_J = 2$ was used for most datasets with an additional test performed using $d_J = 3$ for the elpasolite dataset. These embeddings were not optimised, as in refs. [18,19], and were both constructed using $w_a^1 = 1$, so that the first density channel is element agnostic. For $d_J = 2$ the weights for the second channel were chosen by ordering the elements according to atomic number and then assigning weights of 1, 2, 3,... so that for QM9 the w_a^2 used were 1, 2, 3, 4 and 5 for H,C,N,O and F respectively. For $d_J = 3$ the weights for the second and third density channels were assigned using the group and period of each element in the periodic table, so that for sulphur $w_{1a} = 1$, $w_{2a} = 3$ and $w_{3a} = 16$. This was done in an attempt to capture the chemical similarity encoded in the periodic table and is similar to the encoding used in ref. [53]. For the elpasolite dataset the results of using optimised embeddings, denoted using $\overline{d}_J = 2$ and $\overline{d}_J = 4$, from ref. [21] are also shown.

Information imbalance

The information imbalance is a way of measuring the relative information content of different distance measures. Whilst the distance–distance correlation compares all pairs of distances, the information imbalance is only concerned that the nearest neighbours of each environment are the same according to descriptor A and descriptor B. More precisely, for each environment the distances to all other environments are computed using both distance A and distance B. These distances are then sorted and ranked from $0 - N$ so that each environment has a rank according to A, r_A and according to B, r_B. Then the information imbalance from A to B is defined as

$$\Delta_{A\to B} = \frac{2}{N}\langle r_A | r_B = 1\rangle \tag{18}$$

where N is the number of environments in the dataset and $\langle r_B | r_A = 1\rangle$ is the conditional average of r_B given that $r_A = 1$. Defined in this way $\Delta_{A\to B}$ is statistically confined to lie between 0, A contains the information in B, and 1, A is not informative about B. By comparing ranks, rather than distances, $\Delta_{A\to B}$ is insensitive to changes in scale and by considering only nearest neighbour distances $\Delta_{A\to B}$ is also well suited to handle non-linear relationships. Please refer to ref. [22] for more details.

Sensitivity matrix

Here we give a brief explanation of how the sensitivity matrix Λ is constructed, please refer to ref. [46] for full details. The distance d between the original environment and the perturbed environment is given by

$$d^2 = \sum_i (\Delta x_i)^2$$

where Δx_i is the change in component i of the descriptor $\boldsymbol{x}$. In terms of the atomic displacements this can be re-written as

$$d^2 = \sum_{jk} \Delta R_j \left(\sum_i \frac{\partial x_i}{\partial R_j}\frac{\partial x_i}{\partial R_k} \right) \Delta R_k = \Delta \mathbf{R}^T \Lambda \Delta \mathbf{R} \tag{19}$$

where $\Delta\mathbf{R}$ is a vector of length $3N$ containing the small perturbations to the atomic positions. Defined as such, the distance between the original environment and one perturbed along an eigenvector $\boldsymbol{u}$ of Λ is given by $d = \sqrt{\lambda}|\boldsymbol{u}|$ where $\Lambda\boldsymbol{u} = \lambda\boldsymbol{u}$. Thus, by examining the eigenvalues of Λ we can detect if there are any perturbations that a given descriptor is insensitive to. We note that there will always be six zero eigenvalues, corresponding to three translations and three rotations, and that we expect additional zero eigenvalues for perturbations about symmetric atomic configurations[50]. This is demonstrated in Supplementary Fig. 4 where there are only 6 zero eigenvalues for the asymmetric liquid environments from the HEA dataset but many more zero eigenvalues for the symmetric environments found in the elpasolite dataset.

KRR models

A simple linear model was used to fit the average chemical potential μ_a for each element so that the predicted energy $\hat{E}_j$ for configuration j, denoted by A_j, was given by,

$$\hat{E}_j = \sum_{ja} n_{ja}\mu_a + \boldsymbol{\beta}\cdot\mathbf{k}(A_j) \tag{20}$$

where n_{ja} is the number of atoms of type a in S_j, $\boldsymbol{\beta}$ is a coefficient vector, and $\mathbf{k}(A_j)$ is shorthand for the vector of kernels between A_j and the structures in the training set,

$$\left[\mathbf{k}(A_j)\right]_i = k(A_i, A_j) = \sum_{\substack{x_j \in A_j \\ x_i \in A_i}} k(\mathbf{x}_i, \mathbf{x}_j) \tag{21}$$

where $\mathbf{x}_i$ is the descriptor of atom i, so that the kernel between structures is the sum over all pairwise atomic kernels. A polynomial kernel was used

throughout so that $k(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)^\zeta$ where $\zeta = 1$ is equivalent to RR and $\zeta = 2$, 4, 8 or 16 were used for KRR. We follow ref. [54] in using the following loss function, motivated by the Gaussian Process Regression view,

$$L = \left|\Sigma^{-1}(\mathbf{E} - \hat{\mathbf{E}})\right|^2 + \boldsymbol{\beta}^T K \boldsymbol{\beta} \tag{22}$$

where $K_{ij} = k(A_i, A_j)$, $\mathbf{E}$ is the vector of total energies for structures in the training set, $\hat{\mathbf{E}}$ are the predicted energies and $\Sigma_{ij} = \sigma n_i \delta_{ij}$ where n_i is the total number of atoms in S_i. Minimising L is equivalent to minimising the sum of the RMSE per atom on the training set and an L_2 regression penalty on $\boldsymbol{\beta}$. Doing so yields the following well known solution[55],

$$\boldsymbol{\beta} = (K + \Sigma)^{-1}\mathbf{E} \tag{23}$$

In all cases multiple models were trained using a randomly selected train:test split, with all data not in the training set used as test data. The average MAE achieved across these models is reported with error bars indicating one standard deviation, whilst the regularisation strength σ was chosen using k-fold cross validation, $k \sim 10-20$.

DATA AVAILABILITY

All datasets used are freely available and the data used to generate all plots are available at https://doi.org/10.5281/zenodo.5793851.

CODE AVAILABILITY

The compressions outlined in Fig. 2 have been implemented in the Gaussian Approximation Potential fitting code gap_fit[48]. A Jupyter notebook demonstrating how the $W^T P_l$ compressed power spectrum is computed, and how the original power spectrum can be recovered from it, is available at https://doi.org/10.5281/zenodo.5793851.

Received: 23 December 2021; Accepted: 1 July 2022;
Published online: 11 August 2022

REFERENCES

1. Burke, K. Perspective on density functional theory. *J. Chem. Phys.* **136**, 150901 (2012).
2. Nightingale, M. P. & Umrigar, C. J. Quantum Monte Carlo Methods in Physics and Chemistry. 525 (Springer Science & Business Media, 1998).
3. Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79**, 291 (2007).
4. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
5. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
6. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
7. Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
8. Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model Sim* **14**, 1153–1173 (2016).
9. van der Oord, C., Dusson, G., Csányi, G. & Ortner, C. Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials. *Mach. Learn. Sci. Technol.* **1**, 015004 (2020).
10. Faber, F. A., Christensen, A. S., Huang, B. & Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
11. Deringer, V. L. et al. Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
12. Jain, A., Persson, K. A. & Ceder, G. Research update: the materials genome initiative: data sharing and the impact of collaborative ab initio databases. *APL Mater.* **4**, 053102 (2016).
13. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *JOM* **65**, 1501–1509 (2013).
14. Draxl, C. & Scheffler, M. Nomad: The fair concept for big data-driven materials science. *Mrs Bull.* **43**, 676–682 (2018).
15. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **5**, 1–9 (2019).
16. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
17. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
18. Artrith, N., Urban, A. & Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **96**, 014112 (2017).
19. Uhrin, M. Through the eyes of a descriptor: constructing complete, invertible descriptions of atomic environments. *Phys. Rev. B* **104**, 144110 (2021).
20. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
21. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661–29668 (2018).
22. Glielmo, A., Zeni, C., Cheng, B., Csányi, G. & Laio, A. Ranking the information content of distance measures. *PNAS Nexus* **1**, 1–8 (2022).
23. Onat, B., Ortner, C. & Kermode, J. R. Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials. *J. Chem. Phys.* **153**, 144106 (2020).
24. Mahoney, M. W. & Drineas, P. Cur matrix decompositions for improved data analysis. *Proc. Natl Acad. Sci.* **106**, 697–702 (2009).
25. Nigam, J., Pozdnyakov, S. & Ceriotti, M. Recursive evaluation and iterative contraction of n-body equivariant features. *J. Chem. Phys.* **153**, 121101 (2020).
26. Goscinski, A., Musil, F., Pozdnyakov, S., Nigam, J. & Ceriotti, M. Optimal radial basis for density-based atomic representations. *J. Chem. Phys.* **155**, 104106 (2021).
27. Draxl, C. & Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
28. Csányi, G. et al. Expressive programming for computational physics in fortran 95+. *IoP Comput. Phys. Newsletter* 1–24 (2007).
29. Kermode, J. R. f90wrap: an automated tool for constructing deep python interfaces to modern fortran codes. *J. Phys.: Condens. Matter* **32**, 305901 (2020).
30. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
31. Musil, F. et al. Efficient implementation of atom-density representations. *J. Chem. Phys.* **154**, 114109 (2021).
32. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
33. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
34. Riley, K. F., Hobson, M. P. & Bence, S. J. Mathematical Methods for Physics and Engineering (American Association of Physics Teachers, 1999).
35. Pozdnyakov, S. N. et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **125**, 166001 (2020).
36. Deringer, V. L., Caro, M. A. & Csányi, G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat. Commun.* **11**, 1–11 (2020).
37. Bartók, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
38. Artrith, N. & Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: performance for tio2. *Comp. Mater. Sci.* **114**, 135–150 (2016).
39. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
40. Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F. & Marquetand, P. wacsf-weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.* **148**, 241709 (2018).
41. Batzner, S. et al. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 1–11 (2022).
42. Anderson, B., Hy, T. S. & Kondor, R. Cormorant: covariant molecular neural networks. In *Proc. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (2019).
43. Geiger, M. et al. Euclidean neural networks: e3nn. https://doi.org/10.5281/zenodo.5292912 (2020).
44. Weiler, M., Geiger, M., Welling, M., Boomsma, W. & Cohen, T. S. 3d steerable cnns: learning rotationally equivariant features in volumetric data. In *Proc. 32nd Conference on Neural Information Processing Systems* (2018).
45. Thomas, N. et al. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. Preprint at https://doi.org/10.48550/arXiv.1802.08219 (2018).
46. Parsaeifard, B. et al. An assessment of the structural resolution of various fingerprints commonly used in machine learning. *Mach. Learn. Sci. Technol.* **2**, 015018 (2021).
47. Calsaverini, R. S. & Vicente, R. An information-theoretic approach to statistical dependence: Copula information. *Europhys. Lett.* **88**, 68003 (2009).
48. The gap_fit code. https://github.com/libAtoms/GAP.

49. Byggmästar, J., Nordlund, K. & Djurabekova, F. Modeling refractory high-entropy alloys with efficient machine-learned interatomic potentials: Defects and segregation. *Phys Rev B* **104**, 104101 (2021).
50. Pozdnyakov, S. N., Zhang, L., Ortner, C., Csányi, G. & Ceriotti, M. Local invertibility and sensitivity of atomic structure-feature mappings [version 1; peer review: 2 approved]. Open Res. Europe 2021 (2021).
51. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1−7 (2014).
52. Ropo, M., Schneider, M., Baldauf, C. & Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **3**, 1–13 (2016).
53. Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
54. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
55. Williams, C. K. & Rasmussen, C. E. Gaussian Processes for Machine Learning, vol. 2 (MIT press Cambridge, MA, 2006).
56. Caro, M. A. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Phys. Rev. B* **100**, 024112 (2019).
57. Caro, M. A. Turbogap compression. https://github.com/mcaroba/turbogap/tree/master/tools/compress_indices.

ACKNOWLEDGEMENTS

The authors would like to thank Luca Ghiringhelli, Ioan-Bogdan Magdău and Albert Bartók-Pártay for helpful discussions. We acknowledge support from the NOMAD Centre of Excellence, funded by the European Commission under grant agreement 951786. J.R.K. acknowledges additional support provided by the Leverhulme Trust under grant RPG-2017-191. We are grateful for computational support from the UK national high performance computing service, ARCHER, for which access was obtained via the UKCP consortium and funded by EPSRC grant reference EP/P022065/1.

AUTHOR CONTRIBUTIONS

J.R.K. and G.C. jointly designed the research. JPD developed the compression strategies, performed the calculations and wrote the paper, with input from J.R.K. and G.C. at all stages. All authors revised the paper and approved its final version.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00847-y.

Correspondence and requests for materials should be addressed to James P. Darby.

Reprints and permission information is available at http://www.nature.com/reprints

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