The Journal of Chemical Physics

Compact atomic descriptors enable accurate predictions via linear models

Cite as: J. Chem. Phys. **154**, 224112 (2021); https://doi.org/10.1063/5.0052961
Submitted: 02 April 2021 . Accepted: 24 May 2021 . Published Online: 11 June 2021

Claudio Zeni, Kevin Rossi, Aldo Glielmo, and Stefano de Gironcoli

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Compact atomic descriptors enable accurate predictions via linear models

Cite as: J. Chem. Phys. 154, 224112 (2021); doi: 10.1063/5.0052961
Submitted: 2 April 2021 • Accepted: 24 May 2021 •
Published Online: 11 June 2021

Claudio Zeni,[1,a)] Kevin Rossi,[2] Aldo Glielmo,[1] and Stefano de Gironcoli[1]

AFFILIATIONS

[1] Physics Area, International School for Advanced Studies, Trieste, Italy
[2] Laboratory of Nanochemistry, Institute of Chemistry and Chemical Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, CH, Switzerland

[a)] **Author to whom correspondence should be addressed:** czeni@sissa.it

ABSTRACT

We probe the accuracy of linear ridge regression employing a three-body local density representation derived from the atomic cluster expansion. We benchmark the accuracy of this framework in the prediction of formation energies and atomic forces in molecules and solids. We find that such a simple regression framework performs on par with state-of-the-art machine learning methods which are, in most cases, more complex and more computationally demanding. Subsequently, we look for ways to sparsify the descriptor and further improve the computational efficiency of the method. To this aim, we use both principal component analysis and least absolute shrinkage operator regression for energy fitting on six single-element datasets. Both methods highlight the possibility of constructing a descriptor that is four times smaller than the original with a similar or even improved accuracy. Furthermore, we find that the reduced descriptors share a sizable fraction of their features across the six independent datasets, hinting at the possibility of designing material-agnostic, optimally compressed, and accurate descriptors.

I. INTRODUCTION

The advent of machine learning (ML) methods in atomistic simulation and modeling is benefiting a wide array of disciplines, e.g., it has become an important tool in the field of structure–property predictions.[1–3] As paradigmatic examples, accurate data-driven prediction of properties from structures has been developed for NMR shieldings in molecules and molecular crystals,[4,5] polarization of small to medium molecular systems,[6] solvation and efficacy of drugs,[7–9] and activity of homogeneous and heterogeneous catalysts.[10–12] By the same token, and in relation to the development of force fields (FFs), representative achievements may be found in the simulation of reactions in solutions explicitly accounting for the solvent,[13,14] the assessment of the stability of multi-phase materials relevant, e.g., to storage and conversion,[15–17] electronic devices,[18–20] geology,[21] and the realistic modeling of complex systems in soft matter and biophysics.[22–24]

An open issue of particular importance in data-driven approaches for atomistic systems lies in the choice of the representation of the atomistic system itself. As a witness of the relevance of this problem, a multitude of atomic environment descriptors have been proposed in the last 15 years.[25–31] Among the most successful representations in the field, we find local density representations. In a nutshell, these representations hinge on a construction where atom-centered distributions are represented in a vector form using a many-body expansion.[25–30] Recently, a general formulation of such a local density representation, named "atomic cluster expansion" (ACE), has been proposed by Drautz.[32–34] The ACE representation is symmetric with respect to rotation, translation, and permutation of identical atoms. It is furthermore differentiable with respect to atomic coordinates and complete, that is, in its generalized formulation, it leads to a descriptor body-order, which is iteratively expanded up to the desired one, hence satisfying the uniqueness principle.

In this work, we discuss the performance on two benchmark datasets of three-body representations following the ACE representation, used in conjunction with a ridge regression fitting procedure. In Sec. II, we present the ACE descriptor,[32] the scaled (SC) and non-scaled (NSC) versions of the Chebyshev radial basis functions, and the simplified spherical Bessel (SSB) radial basis

functions, first introduced by Kocer *et al.* for local atomic environments.[35] We then introduce the regression algorithm used to predict atomic forces and total energies throughout the manuscript in Subsection II B. The proposed descriptor-regression framework resembles the Spectral Neighbor Analysis Potentials (SNAPs) first introduced by Thompson *et al.* but relies on power spectrum coefficients rather than bispectrum coefficients, making it a three-body potential in the sense of Ref. 36 rather than a four-body potential (4 + 7-body in the case of quadratic SNAPs). Our regression framework is then benchmarked on two publicly available datasets in Sec. III. First, in Subsection III A, we consider the QM9 dataset, which contains atomic structures and properties, such as formation energy, of 134k small molecules.[27,37] We show that a simple ridge regression framework yields predictions for molecular systems that display an accuracy comparable to the one of more complex, and computationally demanding, regression methods. Similar to other local density representation methods, we observe a trade-off between the computational cost and accuracy, where accurate enough predictions are found only for a sufficiently large dimension of the descriptor; this verifies regardless of whether we employ SC, NSC, or SSB polynomials as the set of radial basis functions in the descriptor. Nevertheless, we observe that SSB functions enable more accurate predictions than the other two radial basis functions when a low number of radial basis functions are employed. Second, in Subsection III B, we look at the fitting of a FF for six single-element crystalline systems utilizing the dataset of forces and energies introduced by Ref. 38. We find again that the proposed learning framework can perform on par with other state-of-the-art approaches and that its accuracy depends on the dimension of the representation. Interestingly, we find that employing SSB radial basis functions is often optimal for compact representations. In Sec. IV, we discuss methods to reduce the dimension of descriptors employed to fit energies in the example case of the database containing six single-element periodic systems.[38] We find that through both principal component analysis (PCA) dimensionality reduction and least absolute shrinkage operator (LASSO) regression feature selection, we are able to match, and sometimes outperform, the accuracy obtained when using the full descriptor while reducing its dimension by a factor of ~4. The features selected by PCA and LASSO across the six single-element datasets are furthermore partially redundant, revealing an underlying material-agnostic structure to the relevant directions in the data space. This insight could guide the informed design of optimally compact, and computationally efficient, local atomic environment descriptors. Finally, the conclusions summarize the results and offer an outlook for future research aimed at improving the algorithm proposed in this manuscript.

II. METHODS

A. Atomic environment representation

To construct the local atomic environment descriptor $\mathbf{q}(\rho)$ used throughout this manuscript, we first define the local atomic density $\rho(\mathbf{r})$, through a standard procedure, as a sum of Dirac delta functions $\delta(\mathbf{r}_{ji} - \mathbf{r})$ centered on each atom surrounding a central atom i within a cutoff r_c,

$$\rho_i(\mathbf{r}) = \sum_{j|r_{ji}\leq r_c} \delta(\mathbf{r}_{ji} - \mathbf{r}), \tag{1}$$

where $\mathbf{r}_{ji}$ indicates the vector $(\mathbf{r}_j - \mathbf{r}_i)$ and r_{ji} is the magnitude of $\mathbf{r}_{ji}$. The local atomic environment representation in Eq. (1) is already invariant to permutations of identical atoms and translations, but not to rotation; it is, moreover, not trivial to transform such a representation into a finite-size descriptor. To overcome these problems, the local atomic density is first approximated via a truncated expansion in spherical harmonics and radial basis functions,

$$\rho_i(\mathbf{r}) \sim \sum_{j\in\rho_i} \sum_{n=0}^{n_{MAX}} \sum_{l=0}^{l_{MAX}} \sum_{m=-l}^{l} c^j_{nlm} g_n(r_{ji}) Y_{lm}(\hat{\mathbf{r}}_{ji}), \tag{2}$$

where $\hat{\mathbf{r}}_{ji}$ is the unit vector of $\mathbf{r}_{ji}$, g_n are the elements of a set of n_{MAX} radial basis functions, Y_{lm} are the elements of a set of spherical harmonics, n_{MAX} indicates the truncation limit for the radial basis set, and l_{MAX} is the truncation limit for the angular basis set. We note that the elements g_n should also depend on the angular expansion coefficient l. We, here, remove the coupling between angular and radial parts following the approach of Ref. 35, as it was shown that such simplification significantly reduces the complexity of evaluating $g(r_{ij})$ without noticeable decreases in the prediction accuracy.[35,39] In principle, one could use the array of coefficients $C_{nlm} = \sum_{j\in\rho_i} c^j_{nlm}$ as a descriptor, but it would not be invariant to rotations of the local atomic environment. To solve this issue, products of N coefficients c^j_{nlm} that correspond to a reducible representation of the identity of the rotation group are taken. The resulting descriptors are of order $(N + 1)$, i.e., they can encode the interaction of up to $N + 1$ atoms at once.[30,32] One advantage of ACE descriptors is given by the linear scaling of the computational cost for their evaluation in the number of atoms M in the neighborhood of i for any order N. This is not the case, e.g., for explicit N-body descriptors, where summations over groups of N neighbors have to be considered, therefore causing the computational cost of their evaluation to scale as $\mathcal{O}(M^{N-1})$, i.e., more than linearly in M whenever $N > 2$. The linear scaling in M of ACE descriptors, therefore, enables a large computational speed-up when compared to explicit N-body descriptors, especially for densely packed systems. In this manuscript, we employ three-body descriptors, where components $q_{n_1,n_2,l}$ of $\mathbf{q}$ are computed as

$$q_{n_1,n_2,l}(\rho_i) = \sum_{j\in\rho_i} \sum_{k\in\rho_i} \sum_{m=-l}^{m=+l} (-1)^m c^j_{n_1 lm} c^k_{n_2 l-m}. \tag{3}$$

The representation in Eq. (3) is expected to strike a good balance between descriptiveness and efficiency,[19,40] as the computational complexity of evaluating the three-body descriptor is $\mathcal{O}(M \cdot (n_{MAX} \cdot l^2_{MAX} + n^2_{MAX} \cdot l_{MAX}))$. Nevertheless, we expect that the inclusion of four- and higher-body descriptors, following the procedure in Ref. 32, will enable us to reach even higher accuracies, albeit at an increased computational cost. The equations reported so far hold for single-element systems. If $S > 1$ atomic species are present, we employ S^2 independent descriptors $\mathbf{q}_{a,b}(\rho_i)$, where a refers to the type of the central atom i, and only surrounding atoms of type b contribute to the value of $\mathbf{q}_{a,b}(\rho_i)$.

In the first manuscript introducing the so-called atomic cluster expansion, Drautz proposes an ensemble of SC polynomials as the orthonormal radial basis set $g_n(r)$,[32] which had been also previously used to expand 2- and 3-body correlation functions and chart

structure-to-property mappings in Ref. 41. We report here the SC radial basis set as defined in Ref. 32,

$$\begin{aligned} g_0(x) &= 1, \\ g_1(x) &= \frac{1}{2}[1+\cos(\pi r/r_c)], \\ g_n(x) &= \frac{1}{2}[1-T_{n-1}(x)]\frac{1}{2}[1+\cos(\pi r/r_c)], \end{aligned} \tag{4}$$

where the Chebyshev polynomials of the first kind $T_n(x)$ are defined recursively as

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x), \end{aligned} \tag{5}$$

and the scaled distance function is

$$x = 1 - 2\left(\frac{e^{-\lambda(r/r_c-1)}}{e^{\lambda}-1}\right), \tag{6}$$

where λ is a coefficient, set to 5 as in Ref. 32.

Beside the SC radial basis, we look at possible changes in performance originating from the use of different radial basis in the ACE expansion: a set of SSB functions of the first kind, introduced in Ref. 35, and a NSC radial basis set. The NSC radial basis set is defined by Eq. (4), where x is $x = 2r/r_c - 1$. The SSB functions of the first kind basis set $g_n(r_{ij})$, introduced in Ref. 35, is defined recursively as

$$\begin{aligned} g_n(r) &= \frac{1}{\sqrt{d_n}}\left(f_n(r) + \sqrt{\frac{e_n}{d_{n-1}}}g_{n-1}(r)\right), \\ d_n &= 1 - \frac{e_n}{d_{n-1}}, \\ e_n &= \frac{n^2(n+2)^2}{4(n+1)^4+1}, \end{aligned} \tag{7}$$

where $d_0 = 1$, $d_1 = 1$, $g_0(r) = 1$, $g_1(r) = f_0(r)$, and

$$\begin{aligned} f_n(r) = (-1)^n \frac{\sqrt{2}\pi}{r_c^{3/2}} \frac{(n+1)(n+2)}{\sqrt{(n+1)^2+(n+2)^2}} \\ \cdot\left[\mathrm{sinc}\left(r\frac{(n+1)\pi}{r_c}\right) + \mathrm{sinc}\left(r\frac{(n+2)\pi}{r_c}\right)\right]. \end{aligned} \tag{8}$$

In Fig. 1, we display the terms g_1 to g_5 for the three radial basis function sets, where r_c was set to 1.

Additionally, independent of the choice of the radial basis set, we have found that appending the element-wise squared descriptor $\mathbf{q}^2$ to $\mathbf{q}$ yields a sizable increase in prediction accuracy, with negligible computational cost. The further inclusion of $\mathbf{q}^3$, or higher-order powers of $\mathbf{q}$, element-wise square-root, or sigmoid, of the original descriptor $\mathbf{q}$, did not appear to yield any significant accuracy increase.

B. Regression algorithm

To carry out the supervised learning task, we adopt ridge regression (RR), one of the simplest and most computationally efficient fitting algorithms. RR recasts the learning problem into the following closed formula:

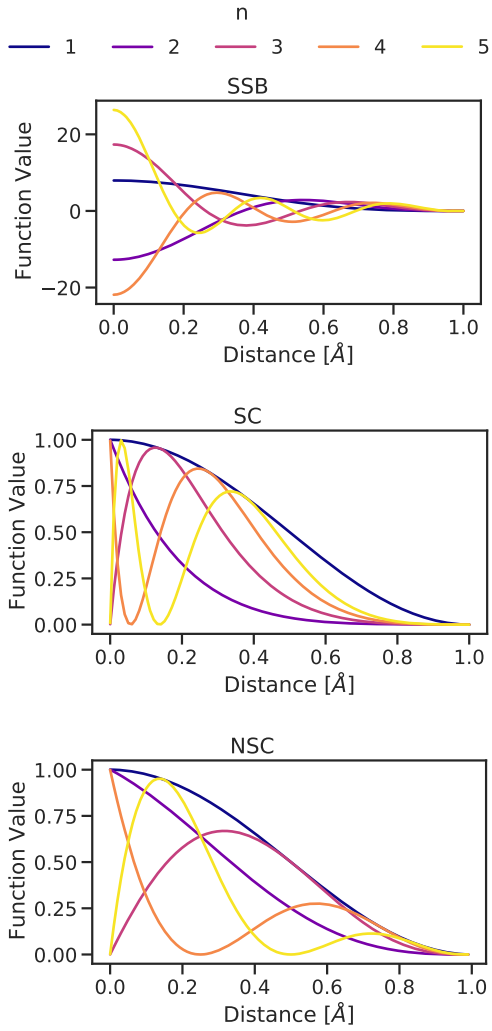

FIG. 1. Visualization of the terms g_1 to g_5 of (top to bottom) the SSB, SC, and NSC radial basis function sets. The cutoff radius was set to 1 Å for the three plots.

$$\mathbf{Y} = \mathbf{Q}\,\mathbf{W} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\mathbf{Y}$ is the matrix of dependent variables, $\mathbf{Q}$ is the matrix of explanatory variables, $\mathbf{W}$ is the parameter matrix that weights $\mathbf{Q}$, and $\boldsymbol{\varepsilon}$ is a vector of error terms, which accounts for possible hidden variables influencing $\mathbf{Y}$ that are not contained in $\mathbf{Q}$. Given a training set $\mathscr{D} = \{\mathbf{Y}_i, \mathbf{Q}_i\}$ $i = 1, \ldots, D$, we obtain the weights $\mathbf{W}$ analytically as

$$\mathbf{W} = (\mathbf{Q}^T\mathbf{Q} + \gamma\,\mathbb{I})^{-1}\mathbf{Q}^T\mathbf{Y}, \tag{10}$$

where γ is the ridge parameter. When both forces and energies are used to train the algorithm, $\mathbf{Y}$ is a 2D matrix with elements $\mathbf{Y}_d$ pertaining to structure d containing S atoms,

$$\mathbf{Y}_d = \left[E_d, f_1^x, f_1^y, f_1^z, \ldots, f_S^x, f_S^y, f_S^z\right], \tag{11}$$

where f_s^c indicates the c-component of the force vector acting on atom s of structure d. Similarly, the matrix of explanatory variables $\mathbf{Q}$ becomes a 3D tensor with elements $\mathbf{Q}_d$ pertaining to structure d,

$$\mathbf{Q}_d = \left[\mathbf{q}_d, -\frac{\partial \mathbf{q}_d}{\partial x_1}, -\frac{\partial \mathbf{q}_d}{\partial y_1}, -\frac{\partial \mathbf{q}_d}{\partial z_1}, \ldots, -\frac{\partial \mathbf{q}_d}{\partial x_S}, -\frac{\partial \mathbf{q}_d}{\partial y_S}, -\frac{\partial \mathbf{q}_d}{\partial z_S}\right], \tag{12}$$

where $\mathbf{q}_d$ is defined as the sum over all atoms i in structure d of the local atomic environment descriptor $\mathbf{q}(\rho_i)$. Whenever only energies are used to fit the algorithm, such as in the case of the QM9 dataset, both the elements of the matrix of explanatory variables and the elements of the matrix of dependent variables simplify to, respectively, $\mathbf{Q}_d = \mathbf{q}_d$ and $\mathbf{Y}_d = E_d$.

Two main advantages arise from the choice of employing RR over more complex learning algorithms, such as artificial neural networks (ANNs) or Gaussian Process (GP) regression. First, RR has a lower computational cost than ANNs or GP regression, since once the descriptor $\mathbf{Q}$ has been calculated, the prediction of $\mathbf{Y}$ requires a single matrix product. Second, RR models, similar to GPs, can be trained in closed form and, therefore, without the need for slow gradient descent algorithms, which also introduce additional hyper-parameters that require careful tuning.

III. RESULTS

A. Energy prediction in the QM9 dataset

As a first benchmark, we look at one of the most widely studied datasets in our community, the QM9 dataset,[27,37,42] and aim to predict the formation energy for each molecule in the database. The QM9 encompasses a relatively large number (133 885) of molecules with a total of up to five chemical species, with each molecule containing up to nine heavy atoms of C, N, O, or S, and any number of H atoms. In the top panel of Fig. 2, we report the mean absolute error (MAE) incurred on total energy predictions by three sets of RR models employing different radial basis functions for a fixed set of descriptor hyperparameters: $r_C = 4.5$, $n_{MAX} = 8$, $l_{MAX} = 10$, and, therefore, a fixed descriptor's dimensionality. A black dashed line in the plot indicates the target of 1 kcal/mol, often referred to as the target chemical accuracy for the prediction of formation energies for molecules. Among the three radial basis expansions under scrutiny, the SSB basis set performs best at any point on the training curve, and it reaches chemical accuracy, even when fewer than the maximum number of training structures (107 800, 80% of the total dataset) are used. We hypothesize that this result is a consequence of the higher spatial resolution of Bessel polynomials with respect to (scaled and non-scaled) Chebyshev polynomials for low values of n_{MAX}, as discussed in the supplementary material, Sec. A. The convergence error for the RR model here presented, employing SSB radial functions with $n_{MAX} = 8$ and $l_{MAX} = 10$, is 0.78 ± 0.02 Kcal/mol. Other methods are able to incur lower MAEs, such as neural networks, reaching 0.14 Kcal/mol in Ref. 43, or Gaussian process regression, reaching 0.14 Kcal/mol in Ref. 44. The lower error incurred by the aforementioned methods is, however, paralleled by a much higher computational cost and complexity. Furthermore, we are not aware of any linear method, which has been successfully used for the prediction of formation energies for the QM9 dataset, i.e., displaying a MAE on par with chemical accuracy.

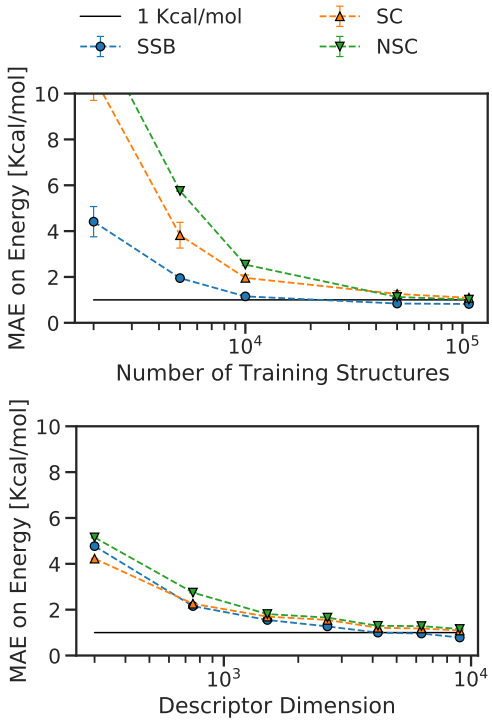

FIG. 2. Validation MAE on total formation energy for the QM9 dataset incurred by potentials, as a function of the number of training structures used with $n_{MAX} = 8$ and $l_{MAX} = 10$ (top panel) and as a function of the number of features in the representation, with $l_{MAX} = n_{MAX} + 2$ and $n_{MAX} = 2, \ldots, 8$ and using 107 800 training structures (bottom panel). The standard deviation of each measure across three independent runs is displayed, where the validation and training set were randomly selected and the size of the validation set was kept fixed at 26 777 (20% of the total dataset). The black dashed line indicates a MAE of 1 Kcal/mol, typically indicated as a target for chemical accuracy in structure energy prediction.

As stated in Sec. II A, the computational cost of the atomic cluster expansion descriptor strongly depends on the choice of n_{MAX} and l_{MAX}, which, in turn, affects the descriptor's dimension. For this reason, the investigation of the validation error as a function of the descriptor's dimension reveals the accuracy/cost trade-off of the algorithm and can lead to increased efficiency. The bottom panel of Fig. 2 displays the validation MAE incurred by RR trained on 107 800 structures for the QM9 dataset as a function of the descriptor's dimension when employing SSB, SC, and NSC radial basis functions. In this instance too, RR FFs employing the SSB basis set reach chemical accuracy (1 Kcal/mol) at smaller descriptor dimensions than the other two basis sets and, more specifically, at $n_{MAX} = 6$ and $l_{MAX} = 8$. This paradigmatic example shows that the choice of the most efficient basis may be key when developing surrogate models for databases, which encompass a large number of data and chemical species. For an analysis of the impact on prediction accuracy of the balance between n_{MAX} and l_{MAX}, the interested reader is directed to the supplementary material, Sec. B.

B. Force and energy prediction in materials

In the previous paragraph, we benchmarked the accuracy of our method while fitting on formation energies only. While energy prediction is of great importance, e.g., for structure search methods, it is often the case that *both* forces and energies are required, e.g., when running molecular dynamics (MD) simulations using a ML FF. In this second example, we consider the database containing forces and energies for six single-element periodic systems, first introduced in Ref. 38. The dataset contains perfect and deformed crystalline structures for two group IV semiconductors (Si and Ge), two body-centered-cubic (BCC) metals (Li and Mo), and two face-centered-cubic (FCC) metals (Ni and Cu); for additional details on the methods used to generate the data, the interested reader is referred to Ref. 38. We thus assess the accuracy of our framework hinging on RR fitting and a three-body ACE representation to produce a FF, given each of the six single-element datasets. Tables I and II report the performance of the proposed ML framework using a SSB radial basis employing $n_{MAX} = 8$ and $l_{MAX} = 10$ and using the same system dependent cutoffs r_c used in Ref. 38 (Mo = 5.2 Å, Si = 4.7 Å, Ge = 5.1 Å, Cu = 3.9 Å, Ni = 4.0 Å, Li = 5.1 Å). For reference, we also report the results from Ref. 38, which benchmarked other widespread state-of-the-art ML frameworks. Notwithstanding the simplicity and computational efficiency inherent to a linear fit, the proposed approach displays performances comparable to the most accurate methods discussed in the literature.

Similar to the case of the QM9 dataset, we look at the interplay between the number of radial and angular basis employed and the corresponding fitting accuracy. Figures 3 and 4 show the root mean squared error (RMSE) incurred by the proposed ML framework on forces and on energies, respectively, in each system, as a function of the descriptor's dimension. An increase in the descriptor extrinsic dimension corresponds to a decrease in the model RMSE, as expected. For most systems, the RMSE does reach a plateau around a descriptor's dimension of 10^2, indicating that more compact, and thus more computationally efficient, basis sets can be employed with negligible accuracy loss. These trends are in agreement with the ones previously reported in the literature for other formulations of the local atomic density representations.[45] In particular, descriptors employing $n_{MAX} = 5$ and $l_{MAX} = 7$ incur in RMSEs on forces and energies that are, on average, respectively, 0.003 ± 0.003 eV/Å and 1.77 ± 1.41 meV/atom higher than the ones incurred by the larger descriptor. In turn, using $n_{MAX} = 5$ and $l_{MAX} = 7$ is approximately four times faster than $n_{MAX} = 8$ and $l_{MAX} = 10$. Different to the case of the QM9 dataset, in Figs. 3 and 4, the difference in the performance of the descriptors employing the three radial basis sets is marginal, even for small descriptor sizes.

The current Python implementation of the algorithm favors code interpretability over efficiency. For a thorough discussion on the computational speed of the ACE framework, we refer the

TABLE I. Minimum RMSE on forces (eV/Å) incurred by our three-body RR potential employing SSB polynomials as radial basis functions for the six single-element datasets from Ref. 38. The symbol * indicates the results from Ref. 38, which are included for comparison.

Material	Our method	GAP*	MTP*	NN-BP*	SNAP*	qSNAP*
Ni	0.03	0.04	0.03	0.07	0.08	0.07
Cu	0.02	0.02	0.01	0.06	0.08	0.05
Li	0.01	0.01	0.01	0.06	0.04	0.04
Mo	0.16	0.16	0.15	0.20	0.37	0.33
Si	0.13	0.12	0.09	0.17	0.34	0.29
Ge	0.09	0.08	0.07	0.12	0.29	0.20

TABLE II. Minimum RMSE on energies (meV/atom) incurred by our three-body linear potential employing SSB polynomials as radial basis functions for the six single-element datasets from Ref. 38. The symbol * indicates the results from Ref. 38, which are included for comparison.

Material	Our method	GAP*	MTP*	NN-BP*	SNAP*	qSNAP*
Ni	1.74	0.62	0.74	2.25	1.17	1.04
Cu	1.19	0.56	0.52	1.68	0.87	1.16
Li	1.23	0.63	0.76	0.98	1.31	0.85
Mo	4.00	3.55	3.89	5.67	9.06	3.96
Si	5.16	4.18	3.02	9.95	8.06	6.28
Ge	11.62	4.47	3.68	10.95	10.96	10.55

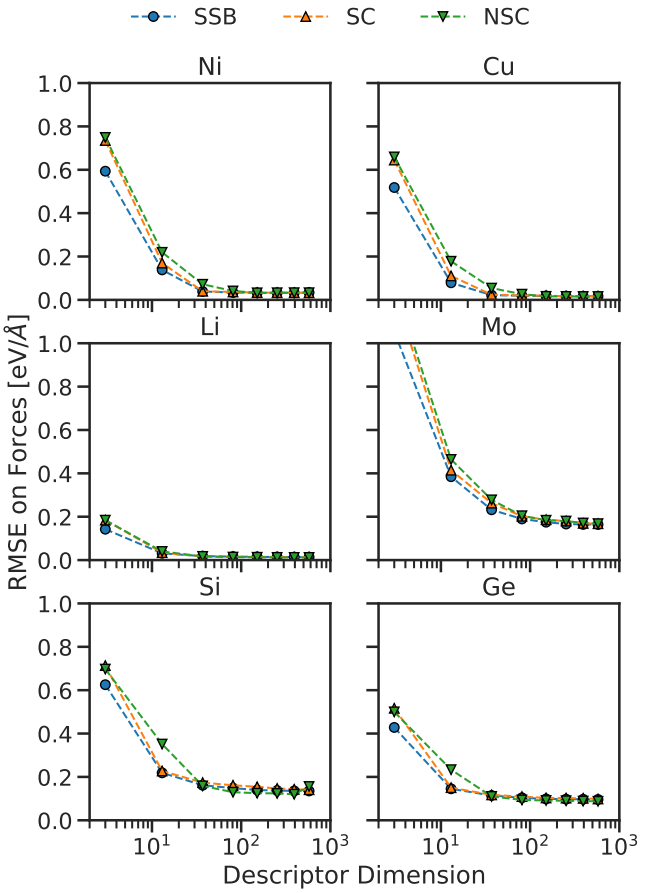

FIG. 3. RMSE on forces incurred by our RR potential trained and tested on data from Ref. 38 as a function of the number of features in the representation using $n_{MAX} = 2, \ldots, 8$ and $l_{MAX} = n_{MAX} + 2$. Color coding refers to the radial basis functions as in Fig. 2.

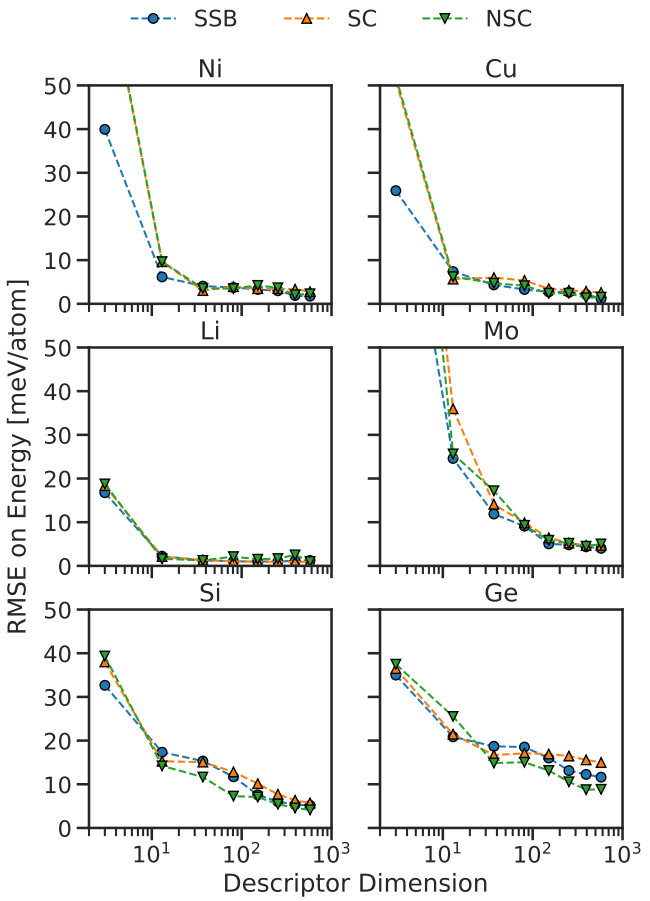

FIG. 4. RMSE on atomic energies incurred by our RR potential trained and tested on data from Ref. 38 as a function of the number of features in the representation using $n_{MAX} = 2, \ldots, 8$ and $l_{MAX} = n_{MAX} + 2$. Color coding refers to the radial basis functions as in Fig. 2.

interested reader to Ref. 46, where it is shown that an efficient C++ implementation of this representation generally leads to predictions whose accuracy and speed are both highly competitive with other state-of-the-art methods.

IV. DESCRIPTOR COMPRESSION

We have showcased how the use of efficient local atomic environment descriptors can yield a satisfying prediction accuracy, even when employing linear, and thus computationally cheap, regression algorithms. Nonetheless, the descriptors employed up to this point contain hundreds to thousands of elements, and the question of whether such a large number of variables are really necessary to describe the data naturally arises. Indeed, widely employed local atomic environment descriptors, such as the smooth overlap of atomic positions (SOAP) and atomic symmetry functions (ASF), can be compressed without loss of accuracy for the case of Gaussian process FFs and artificial neural networks FFs, respectively.[47] Here, we address this question by applying two different techniques, principal component analysis (PCA) and least absolute shrinkage and selection operator (LASSO) regression, to reduce the dimension of the descriptors employed for energy-only fitting on the six single-element datasets analyzed in Sec. III B. For all the six datasets, we compute descriptors $\mathbf{Q}$ using $n_{MAX} = 8$ and $l_{MAX} = 8$, we employ the SSB radial basis function, and we avoid augmenting the descriptor with the element-wise square of each element to simplify the analysis of the results. Figures mirroring the ones shown in the Secs. IV A and IV B, but for the case of SC and NSC radial basis functions, can be found in the supplementary material, Secs. C and D, respectively.

A. PCA dimensionality reduction

PCA is a well-known data analysis algorithm,[48] often used to draw low-dimensional projections of high-dimensional objects, such as the features derived from local density representations.[49] In a nutshell, PCA fits an ellipsoid to the data (in our case, the descriptors $\mathbf{Q}$), therefore allowing for the identification of the directions of highest variance in the dataset. Dimensionality reduction can then be performed by employing only the projections of the original data $\mathbf{Q}$ on the P orthogonal directions displaying the highest variance of the aforementioned ellipsoid. The reduced descriptor is, therefore, obtained as

$$\mathbf{Q}_P^{PCA} = \mathbf{Q} \cdot \mathbf{C}_P, \tag{13}$$

where $\mathbf{C}_P$ is a matrix with the P directions of maximum variance of the data as columns. In the top panel of Fig. 5, we showcase the

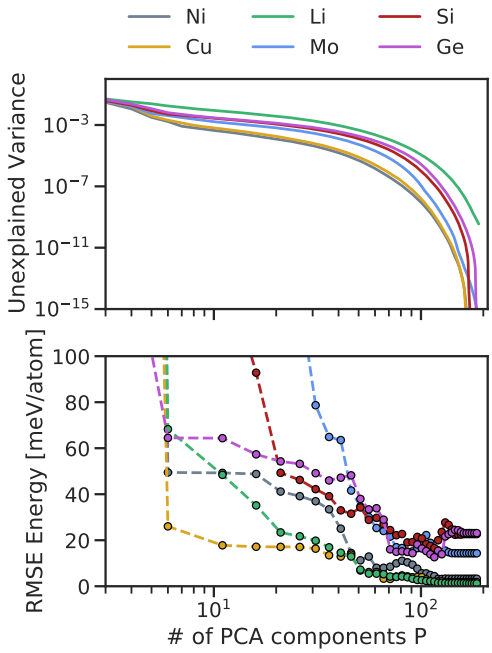

FIG. 5. Top panel: data unexplained variance as a function of the number of PCA components accounted for. Bottom panel: RMSE on energies incurred by RR potentials employing the reduced descriptor $\mathbf{Q}_P^{PCA}$ on the validation set, as a function of the number of PCA components P.

fraction of variance of the descriptors **Q**, which is not explained by the reduced descriptor $\mathbf{Q}_P^{PCA}$ as a function of the number P of PCA components used, in a log–log scale. We notice that around $P \sim 80$, for all systems, the curve of unexplained variance sharply changes slope, indicating that the inclusion of components over $P \sim 80$ in the reduced vector $\mathbf{Q}_P^{PCA}$ will yield a negligible improvement in the explained variance. This is confirmed by the bottom panel of Fig. 5, where we report the RMSE on the validation energy prediction of the RR potential employing reduced descriptors $\mathbf{Q}_P^{PCA}$ containing P PCA components (solid lines and circles) and the non-reduced descriptor **Q** (crosses) as a function of P. For all elements, the RMSEs have a minimum around $P \sim 80$. Moreover, for the elements displaying a higher RMSE, namely, Mo, Si, and Ge, the RMSE increases for $P > 80$; this suggests that the inclusion of components beyond $P \sim 80$ introduces noise in the descriptor, lowering the validation accuracy.

Using PCA feature selection, we are therefore able to construct an 80-dimensional descriptor that, for each material, performs on par with, and sometimes better than, the full 360-dimensional descriptor. To investigate the similarity between the matrices $\mathbf{C}_P$ among the six datasets, we look at the dimension of the intersection between the sub-spaces defined by the rows of $\mathbf{C}_P$; details on this procedure are available in the supplementary material, Sec. E. In Fig. 6, we report the fraction of shared dimensions between the sub-spaces defined by $\mathbf{C}_P$ with $P = 80$. Elements on the diagonal are 1, as a sub-space shares all of its dimensions with itself, while elements on the last row and column are 0, as these report the fraction of shared dimensions between the sub-spaces defined by $\mathbf{C}_P$ (with $P = 80$) and a sub-space defined by 80 360-dimensional randomly generated orthogonal vectors. The off-diagonal elements of all but the last rows have a mean value of 0.69, indicating that, on average, 55 of the 80 dimensions of the matrix $\mathbf{C}_P$ are shared among a couple of single-element descriptor sets. Similar results are observed when using SC and NSC basis functions, where the average fraction of dimensions of the matrix $\mathbf{C}_P$ that are shared among the couple of single-element descriptor sets is 0.93 and 0.78, respectively, as shown in Secs. C and D of the supplementary material. This indicates strong redundancy of the feature selection performed by PCA on the different materials, hinting at an underlying material-agnostic structure in the manifold the **Q** vectors live in.

B. LASSO LARS feature selection

We now look at LASSO least angle regression (LARS) as a way to inform the selection of sparse features in the descriptor vector.[50,51] LASSO LARS is a linear regression algorithm that employs L1 regularization, together with a tunable penalty term. Here, we refer to the features of the original descriptor, which have an associated non-zero weight vector in the LASSO LARS linear model, as the features that were "selected" by the model.[52–54] In the top panel of Fig. 7, we show the inverse correlation existing between the LASSO LARS penalty term and the number of selected features. A sharp transition is found for penalty terms around 10^{-5}, and after this transition, at most 132 out of the 360 features are selected, independently of the material and of the value of the penalty term. This behavior strongly suggests that a substantial fraction of the original descriptor contains redundant information or noise. This intuition is supported by the lower panel of Fig. 7, where the validation RMSE on atomic energy displays a minimum when ~80 features of the original descriptor are employed, for all six materials.

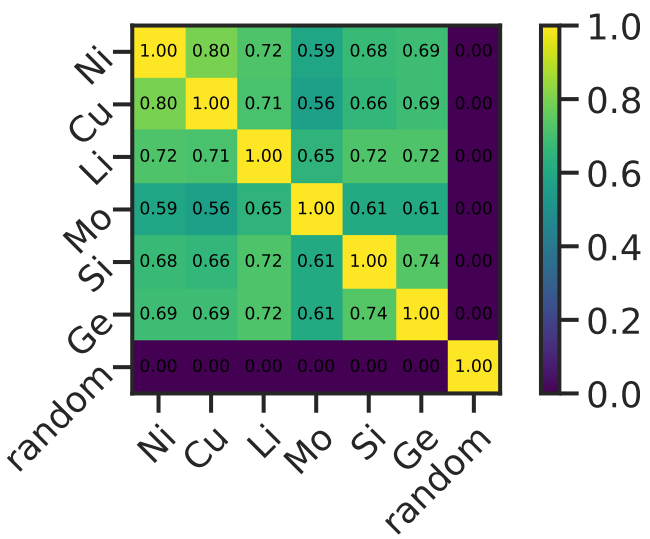

FIG. 6. Heatmap displaying the fraction of dimensions shared by the sub-spaces generated by the first 80 PCA-selected directions of the descriptors **Q** among a couple of single-element datasets. The random label indicates a sub-space generated by taking 80 random orthogonal vectors in the space of the **Q** vectors.

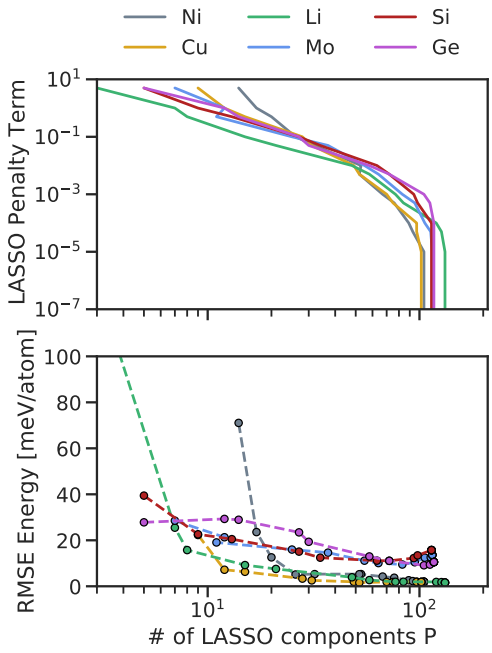

FIG. 7. Top panel: LASSO penalty term as a function of the number of LASSO components accounted for. Bottom panel: RMSE on energies incurred by RR potentials employing the reduced descriptor $\mathbf{Q}_P^{LASSO}$ on the validation set as a function of the number of LASSO components P.

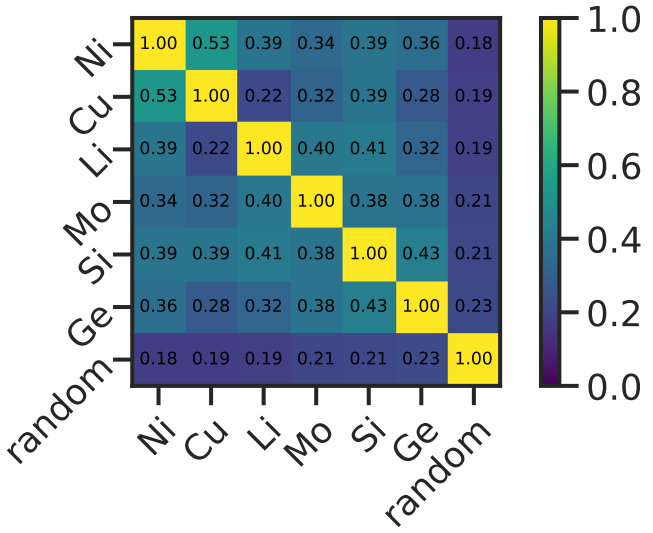

FIG. 8. Heatmap displaying the fraction of components selected via LASSO regression that are shared between two materials, with the penalty term set to $5 \cdot 10^{-4}$. The random label indicates a set of 65 randomly selected components out of the available 360.

To check whether the same set of features is used for L1 regression across the six materials, we calculate the number of times each descriptor feature is selected by the algorithm when the LASSO penalty term is set to $5 \cdot 10^{-4}$, i.e., when the average validation RMSE on atomic energy is lowest. In Fig. 8, we report the fraction of features selected by the LASSO algorithm shared between each pair of single-element datasets. On average, 37% of the LASSO-selected features are shared between each single-element dataset, while only 20% of features are shared between each single-element set and a randomly sampled set of 65 features. Similar results emerge also for the case of SC and NSC radial basis functions, where the percentage of LASSO-selected features that are shared between each single-element dataset is 42% and 32%, respectively (see the supplementary material, Secs. C and D, respectively). The above observations align with the outcome of PCA and indicate that the number of informative features is much smaller than the dimensionality of the descriptor and that these features are, at least partially, shared between different materials. These results, together with the observations drawn from the PCA dimensionality reduction analysis, suggest the possibility of constructing an efficient, low-dimensional descriptor that captures the most relevant features of a local atomic environment.

V. CONCLUSIONS

We systematically probe the accuracy of a three-body representation derived from the "atomic cluster expansion" to predict atomic forces and formation energies in solids and molecules. We furthermore expand the "atomic cluster representation" descriptor by employing Bessel polynomials as radial basis function sets and show that they often display better accuracy than non-scaled Chebyshev and scaled Chebyshev polynomials when a low number of radial basis functions are employed. We demonstrate that this representation, coupled with a simple linear regression algorithm, yields a satisfactory prediction accuracy on the QM9 dataset[27,37] and an accuracy on par with other state-of-the-art representations and statistical learning methods for six single-element datasets.[38]

In the second instance, we focus on methods to reduce the dimension of the representation. We consider both a dimensionality reduction scheme (PCA) and a regression algorithm encompassing a feature selection mechanism (LASSO LARS). We study the interplay between accuracy and representation dimensionality in the database of Ref. 38, which comprises FCC metals, BCC metals, and group IV semiconductors. We find that it is possible to obtain more compact local atomic environment descriptors with no loss in accuracy. Furthermore, we find that there exists an ideal number of PCA components or LASSO LARS selected features for which the accuracy in the prediction actually improves while yielding a fourfold decrease in the dimension of the descriptor. We then study the structure of the representation resulting from both PCA dimensionality reduction and LASSO LARS feature selection. We find that more than 64% of the first 80 directions of maximum variance of the descriptors are shared between each pair of single-element datasets. Similarly, we observe that several descriptor features are relevant (according to a LASSO LARS selection) for the representation of solids of different nature. While this result was drawn from databases containing only hundreds of structures, these were comprised of elements with diverse chemistry. In turn, we envision that our approach could inform the design of extremely compact and fast to compute, yet informative, atomic environment descriptors in a material-agnostic fashion.

SUPPLEMENTARY MATERIAL

A discussion on how well the Bessel, scaled Chebyshev, and Chebyshev radial basis sets approximate a Dirac delta function; details on the accuracy trade-off between radial and angular components of the employed local atomic density descriptor; and details on the computation of the dimension of the intersection of two subspaces can be found in the supplementary material. Furthermore, it contains the same plots shown in Figs. 5–8, but for the case of SC and NSC radial basis functions.

ACKNOWLEDGMENTS

C.Z., A.G., and S.d.G. gratefully acknowledge support from the European Union's Horizon 2020 research and innovation program (Grant No. 824143, MaX "MAterials design at the eXascale" Centre of Excellence).

DATA AVAILABILITY

The package for training ridge regression potentials is freely available under Apache 2.0 license at https://github.com/ClaudioZeni/Raffy. The QM9 dataset is freely available at https://doi.org/10.6084/m9.figshare.c.978904.v5.[42] The single-element materials dataset is freely available in the data directory at https://github.com/materialsvirtuallab/mlearn.[38]

REFERENCES

[1] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).
[2] N. Artrith, J. Phys.: Energy **1**, 032002 (2019).

[3]G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, J. Phys.: Mater. **2**, 032001 (2019).
[4]F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, Nat. Commun. **9**, 4501 (2018).
[5]A. Gupta, S. Chakraborty, and R. Ramakrishnan, Mach. Learn.: Sci. Technol. **2**, 035010 (2021).
[6]D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, Proc. Natl. Acad. Sci. U. S. A. **116**, 3401 (2019).
[7]C. Rauer and T. Bereau, J. Chem. Phys. **153**, 014101 (2020).
[8]Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, Chem. Sci. **9**, 513 (2018); arXiv:1703.00564.
[9]S. Axelrod and R. Gomez-Bombarelli, arXiv:2006.05531 (2020).
[10]M. O. J. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, and A. S. Foster, npj Comput. Mater. **4**, 37 (2018).
[11]B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, and C. Corminboeuf, Chem. Sci. **9**, 7069 (2018).
[12]G. H. Gu, J. Noh, S. Kim, S. Back, Z. Ulissi, and Y. Jung, J. Phys. Chem. Lett. **11**, 3185 (2020).
[13]K. Rossi, V. Jurásková, R. Wischert, L. Garel, C. Corminbœuf, and M. Ceriotti, J. Chem. Theory Comput. **16**, 5139 (2020).
[14]M. Yang, L. Bonati, D. Polino, and M. Parrinello, arXiv:2011.11455 (2020).
[15]M. Eckhoff, F. Schönewald, M. Risch, C. A. Volkert, P. E. Blöchl, and J. Behler, Phys. Rev. B **102**, 174102 (2020).
[16]J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, npj Comput. Mater. **6**, 20 (2020).
[17]C. Zeni, K. Rossi, A. Glielmo, and F. Baletto, Adv. Phys.: X **4**, 1654919 (2019).
[18]G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, Phys. Rev. B **85**, 174103 (2012); arXiv:1201.2026.
[19]C. Zeni, K. Rossi, A. Glielmo, Á. Fekete, N. Gaston, F. Baletto, and A. De Vita, J. Chem. Phys. **148**(24), 241739 (2018).
[20]V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Nature **589**, 59 (2021).
[21]Z. Zhang, G. Csányi, and D. Alfè, Geochim. Cosmochim. Acta **291**, 5 (2020).
[22]S.-L. J. Lahey and C. N. Rowley, Chem. Sci. **11**, 2362 (2020).
[23]J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, ACS Cent. Sci. **5**(5), 755 (2019).
[24]C. Scherer, R. Scheid, D. Andrienko, and T. Bereau, J. Chem. Theory Comput. **16**, 3194 (2020).
[25]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[26]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[27]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[28]A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, J. Comput. Phys. **285**, 316 (2015).
[29]A. V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).
[30]A. Glielmo, C. Zeni, and A. De Vita, Phys. Rev. B **97**, 184307 (2018).
[31]K. Rossi and J. Cumby, Int. J. Quantum Chem. **120**, e26151 (2020).
[32]R. Drautz, Phys. Rev. B **99**, 014104 (2019).
[33]M. Bachmayr, G. Csanyi, R. Drautz, G. Dusson, S. Etter, C. van der Oord, and C. Ortner, "Atomic cluster expansion: Completeness, efficiency and stability," arXiv:1911.03550 [math.NA] (2020).
[34]R. Drautz, Phys. Rev. B **102**, 024104 (2020).
[35]E. Kocer, J. K. Mason, and H. Erturk, J. Chem. Phys. **150**, 154102 (2019).
[36]A. Glielmo, C. Zeni, Á. Fekete, and A. De Vita, in *Machine Learning Meets Quantum Physics, Lecture Notes in Physics*, edited by K. Schütt, S. Chmiela, O. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K. R. Müller (Springer, Cham, 2020), Vol. 968.
[37]L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).
[38]Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood *et al.*, J. Phys. Chem. A **124**, 731 (2020).
[39]E. Kocer, J. K. Mason, and H. Erturk, AIP Adv. **10**, 015021 (2020).
[40]A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).
[41]N. Artrith, A. Urban, and G. Ceder, Phys. Rev. B **96**, 014112 (2017).
[42]R. Ramakrishnan, P. Dral, M. Rupp, and O. Anatole von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).
[43]O. T. Unke and M. Meuwly, J. Chem. Phys. **148**, 241708 (2018).
[44]M. J. Willatt, F. Musil, and M. Ceriotti, J. Chem. Phys. **150**, 154110 (2019).
[45]R. Jinnouchi, F. Karsai, C. Verdi, R. Asahi, and G. Kresse, J. Chem. Phys. **152**, 234102 (2020).
[46]Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner *et al.*, arXiv:2103.00814 (2021).
[47]A. Glielmo, C. Zeni, B. Cheng, G. Csanyi, and A. Laio, "Ranking the information content of distance measures," arXiv:2104.15079 [stat.ML] (2021).
[48]K. Pearson, London, Edinburgh Dublin Philos. Mag. J. Sci. **2**, 559 (1901).
[49]E. D. Cubuk, B. D. Malone, B. Onat, A. Waterland, and E. Kaxiras, J. Chem. Phys. **147**, 024104 (2017).
[50]B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, Ann. Stat. **32**, 407 (2004).
[51]M. Benoit, J. Amodeo, S. Combettes, I. Khaled, A. Roux, and J. Lam, Mach. Learn.: Sci. Technol. **2**, 025003 (2021) arXiv:1912.10761.
[52]L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[53]R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Phys. Rev. Mater. **2**, 083802 (2018).
[54]L. J. Nelson, G. L. Hart, F. Zhou, V. Ozoliņš *et al.*, Phys. Rev. B **87**, 035125 (2013).