On the Completeness of Atomic Structure Representations

Sergey N. Pozdnyakov,[1,*] Michael J. Willatt,[1,*] Albert P. Bartók,[2]
Christoph Ortner,[3,†] Gábor Csányi,[4,‡] and Michele Ceriotti[1,§]

[1] *Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland*
[2] *Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom*
[3] *Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom*
[4] *Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*

(Dated: September 10, 2020)

Many-body descriptors are widely used to represent atomic environments in the construction of machine learned interatomic potentials and more broadly for fitting, classification and embedding tasks on atomic structures. There is a widespread belief in the community that 3-body correlations are likely to provide an overcomplete description of the environment of an atom. We produce several counterexamples to this belief, with the consequence that any classifier, regression or embedding model for atom-centred properties that uses 3 (or 4)-body features will incorrectly give identical results for different configurations. Writing global properties (such as total energies) as a sum of many atom-centred contributions mitigates the impact of this fundamental deficiency – explaining the success of current "machine-learning" force fields. We anticipate the issues that will arise as the desired accuracy increases, and suggest potential solutions.

Over the past decade tremendous progress has been made in the use of statistical regression to sidestep computationally demanding electronic structure calculations, and obtain "machine-learning" models of materials and molecules, that use as inputs only the chemical nature and coordinates of the atoms [1–10]. A crucial driver of this progress has been the introduction of *representations* of atomic structures: A property associated with the i-th atom can be written as $F_i = \mathcal{F}(\mathcal{X}_i)$, where $\mathcal{X}_i = \{\mathbf{r}_{ij}\}_{j\neq i}$ describes the neighbour environment of the i-th atom. To preserve symmetries of the target property, the representation of $\mathcal{X}_i$ should be equivariant [11, 12] (often simply invariant [1, 2, 13–15]) with respect to translations, rotations, labelling of identical atoms, and often also reflections. Most of the invariant representations [1, 2, 13, 16, 17] can be seen as projections onto different bases of many-body correlation functions of the atom density[18]. To stress that our results apply equally to all these frameworks, we use the abstract notation $|\overline{\rho_i^{\otimes\nu}}\rangle$ to indicate the $(\nu+1)$-body correlation, which is centered on the i-th atom [18]. For instance, the 2-body correlation $|\overline{\rho_i^{\otimes 1}}\rangle$ corresponds to the histogram of interatomic distances r_{ij} – equivalent to the radial distribution function or the 2-body symmetry functions, G_2, of Ref. [1]. The 3-body correlation $|\overline{\rho_i^{\otimes 2}}\rangle$ is equivalent to the histogram of triangles, represented by the 3-tuples $(r_{ij}, r_{ij'}, \omega_{ijj'} = \hat{\mathbf{r}}_{ij}\cdot\hat{\mathbf{r}}_{ij'})$ – and to the power spectrum [2], or to the 3-body symmetry functions, G_3 [1]. Linear regression based on these features is equivalent to a body-ordered expansion of the target property[7, 18–22]. Given that computing higher-order terms is increasingly costly, the representation is typically truncated at 3 or 4 body correlations.

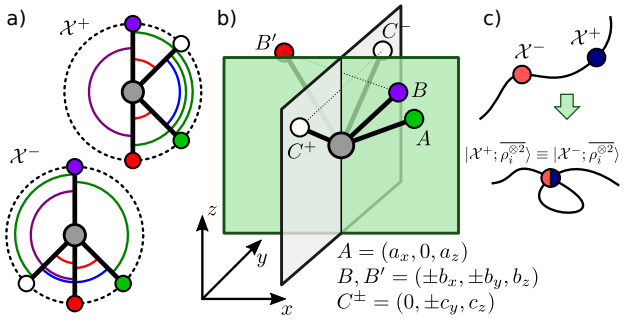

FIG. 1. (a) Two structures with the same histogram of triangles; (angles: 45°, 45°, 90°, 135°, 135°, 180°) (b) A manifold of degenerate pairs of environments: In addition to three points A, B, B' a fourth point C^+ or C^- is added leading to two degenerate environments, $\mathcal{X}^+$ and $\mathcal{X}^-$. (c) Degeneracies induce a transformation of feature space so that structures that should be far apart are brought close together.

Employing *non-linear* functions of low-order invariants, e.g. $F_i = \tilde{\mathcal{F}}(|\overline{\rho_i^{\otimes 2}}\rangle)$, incorporates information on higher-order correlations, and there is a widespread belief in the community [7, 23, 24], supported by numerical evidence [13], that the 3-body correlations likely provide an over-complete description of an atomic environment. The *completeness* (injectivity) of the structure-representation map would guarantee that any atom-centered property can be described by $\tilde{\mathcal{F}}$, which extends to any atom-centered decomposition of extensive properties, such as the total energy[7]. In this Letter, we present several counterexamples to this widely-held belief, discuss the implications for machine learning atomistic properties, and suggest directions towards the construction of complete representations.

Figure 1a exhibits a simple example of a pair of en-

vironments, $\mathcal{X}^+$ and $\mathcal{X}^-$, with four neighbouring atoms of the same species positioned on a circle around the central atom. The two structures cannot be superimposed by rotations and mirror symmetry, but they have the same list of distances and angles and hence cannot be distinguished by their 3-body correlations. To elucidate this example, and more generally understand the difficulty of reconstructing an atomic environment from a body order representations, consider the Gram matrix $G_{jj'} = \mathbf{r}_{ij} \cdot \mathbf{r}_{ij'}$, which contains sufficient information to reconstruct a configuration up to an arbitrary rotation or reflection. If all the distances r_{ij}, or the chemical identity of the neighbors, are distinct, one can unequivocally assign distances and angles to a specific atom, and reconstruct the Gram matrix from the unordered list $\{(r_{ij}, r_{ij'}, \omega_{ijj'})\}$. If some of the distances are the same, however, it becomes possible to swap some entries of $\mathbf{G}$, yielding two or more degenerate environments that are different, but have the same 3-body invariants.

As shown in Fig. 1b, one can generalize the construction to obtain a manifold of degenerate environment pairs parameterised by 7 continuous variables. The total dimensionality of the configuration space of 4 neighbours is $4 \times 3 - 3 = 9$. Thus, the degenerate manifold has a dimension of 7 and a codimension of 2. When going from the + to the − structure in the pair, the elements of the Gram matrix between C-type and B-type points are swapped, leading to non-equivalent structures that have the same 3-body description. This construction can be extended by adding further A or C-type points (increasing the codimension of the degenerate manifold by one) or pairs of B-type points (each pair increasing the codimension by three). Other counterexamples can be found, involving triplets of degenerate structures (see SI). Tight bounds on the codimension of degenerate manifolds and on the multiplicity of degenerate structures is a key aspect in understanding the success of incomplete environment descriptors, but is beyond the scope of the present work. However, the example of Fig. 1b is sharp in the sense that (i) for three or fewer neighbours the 3-body correlation suffices to reconstruct the enviroment and (ii) for four or more neighbours one can construct a manifold of co-dimension 2 which must contain all degenerate environments. These results, which build on those in Ref. [25], are detailed in the SI. It is unclear to us whether the increase of the co-dimension when neighbors are added in the example of Fig. 1b is specific to our construction, or reflects a general result.

Following the procedure in Fig. 1b, one can produce a pair of degenerate tetrahedral environments, that we label $\mathcal{X}^+$ and $\mathcal{X}^-$, corresponding to a CH_4 molecule. Figure 2a shows a portion of the two manifolds (blue and red surfaces, parameterised by two variables q and s) built as a principal component projection of the power spectrum space (details given in the SI). Structures within the two surfaces correspond to configurations that are

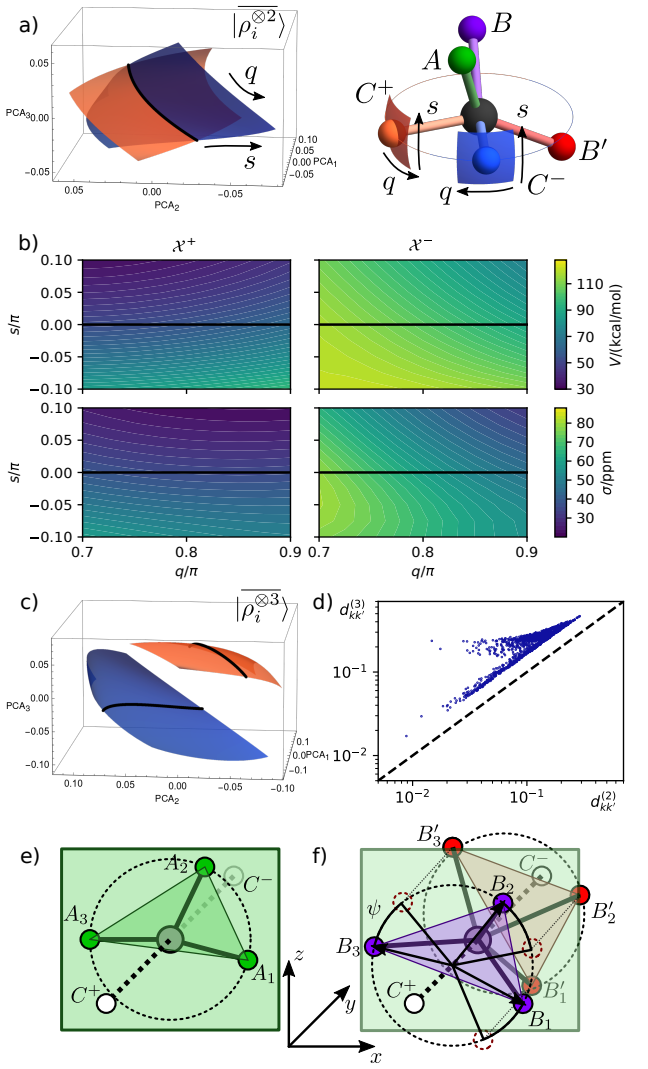

FIG. 2. (a) PCA projection of $|\mathcal{X}^+; \overline{\rho_i^{\otimes 2}}\rangle$ and $|\mathcal{X}^-; \overline{\rho_i^{\otimes 2}}\rangle$ for a continuous manifold of CH_4 environments $\mathcal{X}^+$ and $\mathcal{X}^-$, parameterised by q (that moves along the degenerate set, represented by a black line) and s (that breaks the degeneracy). (b) Energy (top) and ^{13}C chemical shieldings (bottom) of a CH_4 molecule that follows such manifolds; the zero of the two quantities is set to the values for the ideal geometry. (c) PCA projection of the bispectrum $|\overline{\rho_i^{\otimes 3}}\rangle$ space manifold. (d) Correlation plot of the distances between two points k and k' along both manifolds, computed based on the power spectrum ($d^{(2)}_{kk'}$) or the bispectrum ($d^{(3)}_{kk'}$). (e) Construction of a pair of environments that are mirror images but share identical chiral $|\overline{\rho_i^{\otimes 3}}\rangle$ features. A points lie in the xz plane, along a circle centred on the origin. $C^\pm$ points lie along the y axis, symmetric about the origin. (f) a pair of inequivalent structures with the same chiral $|\overline{\rho_i^{\otimes 3}}\rangle$ features. B and B' points lie on circles centred on the origin, and shifted by the same amount above and below the xz plane. One of the sets of points is twisted around y by an angle ψ.

different from each other, but those along the black line (corresponding to $s = 0$) have identical 2- and 3-body invariants, which therefore cannot distinguish $\mathcal{X}^+$ and $\mathcal{X}^-$, and the two manifolds intersect each other. As shown

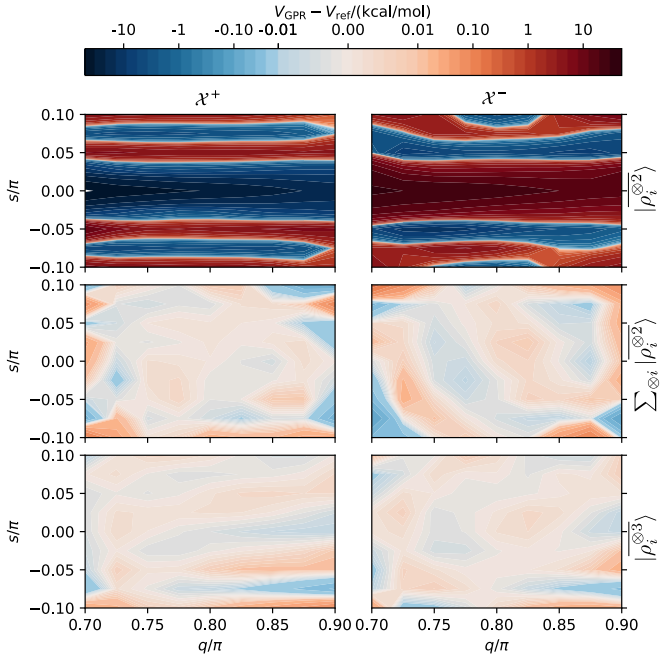

FIG. 3. Error in the prediction of the molecular energy for CH_4 configurations along the manifold depicted in Fig. 2c and d, using a GPR model based on a non-linear kernel built on the C-centred SOAP power spectrum (top, RMSE: 12kcal/mol), a combination of C and H-centred power spectra (middle, RMSE: 0.027 kcal/mol), and the C-centred bispectrum (bottom, RMSE: 0.011 kcal/mol).

in Fig. 2b, however, both atom-centred properties such as the ^{13}C NMR chemical shift, and extensive properties such as molecular energy, are very different as they cannot be described fully by 3-body correlations around the central atom. Higher body-order features can differentiate between $\mathcal{X}^+$ and $\mathcal{X}^-$. As shown in Fig. 2c, the feature-space degeneracy is lifted by the 4-body correlation (bispectrum), $|\overline{\rho_i^{\otimes 3}}\rangle$, which corresponds to the unordered list of tetrahedra formed by the central atom and three of its neighbors. The presence of a degeneracy can be revealed by comparing environment distances $d^{(2)}, d^{(3)}$ computed, respectively, from power spectrum coordinates $|\overline{\rho_i^{\otimes 2}}\rangle$ and bispectrum coordinates $|\overline{\rho_i^{\otimes 3}}\rangle$. One then observes that pairs of environments that are close in $d^{(2)}$ remain well separated by $d^{(3)}$ (Fig. 2d). However, the bispectrum is not complete either. While it does differentiate between the tetrahedral CH_4 environments in Fig. 2a, one can build pairs of inequivalent environments that have the same 4-body correlations. The environments in Fig. 2e are chiral (mirror) images of each other, but the bispectrum does not distinguish them because the tetrahedra it is composed of are not chiral. [26] Fig. 2f extends this construction to a pair of environments that have the same 4-body correlations ($\nu = 3$) and are *not* chiral images of each other.

A Gaussian process regression model based on a non-linear kernel built on the SOAP power spectrum (equivalent to the 3-body correlation, $|\overline{\rho_i^{\otimes 2}}\rangle$, see SI) results in

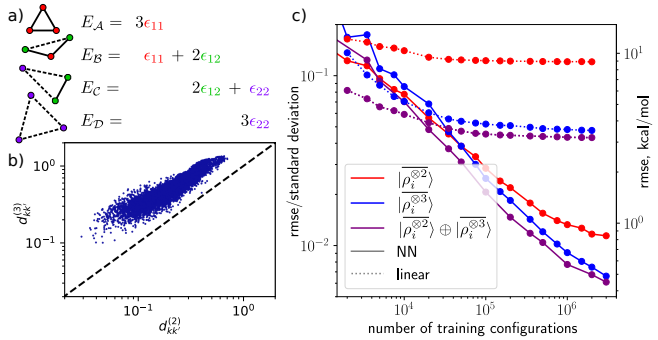

FIG. 4. (a) Four configurations distinguishable by the set of their atom-centered 2-body histograms. Only three different site energies occur in these configurations, hence fitting four total energies leads to overdetermined regression. (b) Correlation plot of powerspectrum and bispectrum distances between C environments in a database of random CH_4 configurations. (c) Learning curves for the atomization energy of random CH_4 configurations.

large errors, not just along the $s = 0$ line of degeneracy, but also for structures that are not exactly indistinguishable according to the power spectrum (top panels in Fig. 3). This underscores the fact that the existence of manifolds of degenerate structures introduces a distortion of the feature space (Fig. 1c), and hinders the ability to perform regression regardless of whether strictly degenerate pairs are included in the training. Because they are ultimately based on the same unordered sets of triangles, Behler-Parrinello "atom-centered symmetry functions" [1], the FCHL descriptors of von Lilienfeld and coworkers[27], the MBTR descriptor of Rupp [28], and the smooth version of the DeepMD framework [29] will also suffer from the same problem. The fact that a large manifold of CH_4 environments is un-learnable using 2- and 3-body features is a shortcoming that fundamentally limits the reliability of machine-learned models of atom-centred properties based on these descriptors.

When learning the decomposition of a *global* property, such as the total energy, one can hope to lift the degeneracy by using features centred on other atoms in the structure. For the construction in Fig. 1b, there is always at least one atom outside the bisecting A plane that breaks the indistinguishability of $\mathcal{X}^+$ and $\mathcal{X}^-$. Indeed, a model that combines C and H-centred non-linear kernels can approximate the molecular energy to excellent accuracy, also along the degenerate manifold (see Fig. 3, middle panels). In general, however, such a mechanism does not guarantee that efficient models can be constructed based on incomplete atom-centred features. For the sake of simplicity, we demonstrate this for the case of 2-body descriptors $|\overline{\rho_i^{\otimes 1}}\rangle$. It is well-known that the list of distances from the centre of an environment, or even the list of distances in a structure [25], are not complete representations. It has, however, been speculated [23] that simultaneous knowledge of all atom-centred lists of distances in a structure would provide a complete representation

of the configuration, and that one could use this representation to predict arbitrary potentials using an additive model based on non-linear functions of $\left|\overline{\rho_i^{\otimes 1}}\right\rangle$. Both conjectures are false. We present a counter-example to the first conjecture in the SI. The counterexample to the second statement, cf. Figure 4a, is far more concerning though: even if, in a training set, all configurations can be uniquely identified by the collection of the atom-centered 2-body histograms, it does not follow that a total energy represented in terms of these histograms can be learned. The breakdown of the purely 2-body models in these limiting cases has practical implications, as they translate into instability and data inefficiency in real-life scenarios – which is the ultimate reason why models based on purely radial information have been superseded by those incorporating 3-body features.

Proving the existence of similar counterexamples for the learning of global properties using $|\overline{\rho_i^{\otimes 2}}\rangle$ is more challenging. It is possible, however, to numerically demonstrate how a model based on 3-body features suffers from a degradation of learning efficiency, provided that one pushes it to sufficiently high accuracy. Figure 4b,c show results for a data set of about 3 million CH_4 configurations obtained by randomly distributing the atoms and discarding structures with too close contacts (details in the SI). The distance-distance correlations (panel b) show that there are configurations that approach the degenerate manifolds, but there are no fully-degenerate pairs. We then built an additive model that includes contributions from both the C and the H atoms, converging the discretization of $\left|\overline{\rho_i^{\otimes 2}}\right\rangle$ and using a neural network to ensure maximal flexibility in the feature-property mapping. The learning curves (Fig. 4c) exhibit clear signs of saturation – usually considered indicative of lack of information in the features or model [30–32] – suggesting that even though each pair of environments (and therefore structures) in the data set can be distinguished based on $\left|\overline{\rho_i^{\otimes 2}}\right\rangle$, the presence of near-degeneracies affects the stability and efficiency of the regression.

Using the higher-body order features to differentiate between $\mathcal{X}^+$ and $\mathcal{X}^-$ does indeed lead to a more efficient model (Fig. 3, bottom panel), that predicts the energy along the degenerate manifold with an error that is roughly a third of that obtained by a multi-center, power-spectrum-based model. Substantial improvements are also seen for the random CH_4 configurations. A NN based on $|\overline{\rho_i^{\otimes 3}}\rangle$ reduces the full-train-set error by 40%, down to $\approx$ 0.5 kcal/mol. Similar to what was observed for $\left|\overline{\rho_i^{\otimes 2}}\right\rangle$-based models that combine multiple cutoff distances [32], there is a data/complexity tradeoff. For small training set sizes a simpler powerspectrum model can outperform one based on the bispectrum, and linear regression outperforms a deep neural network. The best balance between data efficiency, computational cost and ultimate accuracy might involve a combination of different kinds of features, as demonstrated by the hybrid model in Fig. 4. Approaches such as the moment tensor potentials [22], permutationally invariant polynomials[33, 34] and the atomic cluster expansion [20] allow, if necessary, to further resolve degeneracies by including arbitrary body-orders of correlation. We show in the SI that similar considerations apply also to a database of bulk silicon structures [35]. The cutoff distance, however, complicates the picture, because the number of neighbors included in the environments influences the proximity of structures to the degenerate manifold, and because the model accuracy is also affected by the truncation of long-range interactions [36]. Descriptors such as eigenspectra of matrices constructed from the atomic configuration (distance matrix, Laplacian, orbital overlap, etc.)[37, 38] also contain information on high body order correlations, and as such are not expected to be degenerate for the present examples. Their completeness properties are not understood at present.

Overall, the results we have shown indicate that despite the remarkable success of ML models that describe atomic structures in terms of n-body correlations features, there is still work to do to understand fully how the configuration space of a set of atoms is mapped onto symmetry-adapted representations. The problem is to construct a representation which is (i) complete; (ii) smooth with smooth inverse; (iii) and invariant under isometries and permutations. An obvious, but ineffective, solution is to use the union of *all* n-point correlations [20, 22]. Pragmatically, one can proceed as we do here for the CH_4 dataset, increasing the correlation order until all configurations in a given training set are distinguishable, possibly reducing the cost of computing high-order features using a sparsification procedure along the lines of [39, 40]. It is, however, desirable to know *a priori* which features are required to guarantee (i–iii). For example, we may ask whether there is a fixed finite $\bar{n}$ such that all higher-order n-points correlations can be recovered from the $\bar{n}$-point correlation. There are at least two perspectives from which to pursue questions of this kind: signal processing and invariant theory.

In the signal processing literature it has long been known that the power spectrum is insufficient to reconstruct *most* signals, while the bi-spectrum uniquely identifies translation-invariant and compact signals [41–43]. On the other hand, Ref. [41] provides a range of elementary examples establishing that no correlation order suffices to reconstruct all periodic signals. Nevertheless, stable bispectrum inversion has been shown to work well in practice due to the fact the *most* signals can be reconstructed from it; see e.g. [44, 45] and references therein. These results have a striking parallel to our own observations regarding the reconstruction of an atomic environment and in particular suggest that *in theory* no $\bar{n}$-point correlation may suffice to reconstruct the environment.

Still, since atomic environments can be thought of as a very restrictive class of signals, the invariant theory perspective may shed additional light on our questions. The perspective of Boutin and Kemper [25] appears to be particularly useful, establishing conditions under which a points cloud can be reconstructed from the histogram of distances. The problem we tackle here is closely related: degeneracy of two centred environments with respect to n-body correlations implies degeneracy of the point clouds consisting of the neighbors with respect to $n-1$ body correlations. For example, Fig. 1a, implies that the length-histogram of the neighbours lying on the circle are degenerate (indeed, this is the example given in Fig. 4 in Ref.[25] and in Fig 2 of [23]). Similarly, Fig 2f, shows *environments* that are degenerate with respect to the 4-body correlation (tetrahedron histograms) are also degenerate with respect to the 3-body correlations (triangle histograms) of the *entire* structure. A similar approach may therefore help determine tight bounds on the codimension of the degenerate manifold although, as far as we are aware, there are no rigorous results in this direction.

The problem of formulating a complete feature map is of fundamental importance – particularly when considering the use for generative models that require inverting the relation between a representation and the underlying structure – and has practical implications, particularly when one wants to achieve high accuracy with the minimum amount of data. The presence of many neighbors or of different species (that provide distinct "labels" to associate groups of distances and angles to specific atoms), and the possibility of using representations centred on nearby atoms to lift the degeneracy of environments reduces the detrimental effects of the lack of uniqueness of the power spectrum when learning extensive properties such as the energy. We show, however, that the learning rate of this kind of models reduces dramatically in the high accuracy regime, revealing the limitations of a description based on 3-body features. Diagnostic tools such as the joint distance histogram that we introduce here can help identify problematic parts of datasets, give more confidence in the reliability of simple-to-compute low-order invariants, and guide the choice of a small number of higher-order features to improve the accuracy and efficiency of models.

MJW, SP and MC acknowledge funding by the Swiss National Science Foundation (Project No. 200021-182057). CO acknowledges funding by the Leverhulme trust, RPG-2017-191.

* These two authors contributed equally
† c.ortner@warwick.ac.uk
‡ gc121@cam.ac.uk
§ michele.ceriotti@epfl.ch

[1] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[3] W. J. Szlachta, A. P. Bartók, and G. Csányi, Phys. Rev. B **90**, 104108 (2014).
[4] R. Kobayashi, D. Giofré, T. Junge, M. Ceriotti, and W. A. Curtin, Phys. Rev. Mater. **1**, 053604 (2017).
[5] M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).
[6] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[7] A. Glielmo, C. Zeni, and A. De Vita, Phys. Rev. B **97**, 184307 (2018).
[8] N. Lubbers, J. S. Smith, and K. Barros, J. Chem. Phys. **148**, 241715 (2018).
[9] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, ACS Cent. Sci. **5**, 57 (2019).
[10] 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).
[11] A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).
[12] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. **120**, 036002 (2018).
[13] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[14] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, J. Chem. Phys. **139**, 184118 (2013).
[15] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, and S. Goedecker, J. Chem. Phys. **144**, 034203 (2016).
[16] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017), arXiv:1611.04678.
[17] O. A. von Lilienfeld, Angew. Chem. Int. Ed. **57**, 4164 (2018).
[18] M. J. Willatt, F. Musil, and M. Ceriotti, J. Chem. Phys. **150**, 154110 (2019).
[19] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, J. Comput. Phys. **285**, 316 (2015).
[20] R. Drautz, Phys. Rev. B **99**, 014104 (2019).
[21] C. van der Oord, G. Dusson, G. Csanyi, and C. Ortner, (2019), arXiv:1910.06010.
[22] A. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).
[23] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. **115**, 1084 (2015).
[24] E. Kocer, J. K. Mason, and H. Erturk, AIP Advances **10**, 015021 (2020).
[25] M. Boutin and G. Kemper, Adv. Appl. Math. **32**, 709 (2004).
[26] All body-order correlations with $\nu \geq 3$, when defined as an average over proper rotations, are sensitive to chirality and can differentiate between enantiomers. When learning non-chiral properties, such as the energy, one can average over inversion. Unless otherwise specified, in this work $|\overline{\rho_i^{\otimes 3}}\rangle$ indicates the non-chiral version.
[27] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).

[28] H. Huo and M. Rupp, (2017), arXiv:1704.06439.
[29] L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and W. E, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018) pp. 4436–4446.
[30] B. Huang and O. A. Von Lilienfeld, J. Chem. Phys. **145** (2016), 10.1063/1.4964627.
[31] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).
[32] M. J. Willatt, F. Musil, and M. Ceriotti, Phys. Chem. Chem. Phys. **20**, 29661 (2018).
[33] B. J. Braams and J. M. Bowman, International Reviews in Physical Chemistry **28**, 577 (2009).
[34] C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, Mach. Learn.: Sci. Technol. **1**, 015004 (2020).
[35] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X **8**, 041048 (2018), arXiv:1805.01568.
[36] A. Grisafi and M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019).
[37] F. Pietrucci and W. Andreoni, Phys. Rev. Lett. **107**, 085504 (2011).
[38] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, and S. Goedecker, J Chem Phys **144**, 034203 (2016).
[39] M. W. Mahoney and P. Drineas, Proc. Natl. Acad. Sci. U. S. A. **106**, 697 (2009).
[40] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[41] J. I. Yellott and G. J. Iverson, J. Opt. Soc. Am. A, JOSAA **9**, 388 (1992).
[42] R. Kondor, , 1 (2018), arXiv:1803.01588.
[43] R. Kakarala, J. Math. Imaging Vis. **44**, 341 (2012), arXiv:0902.0196.
[44] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, IEEE Trans. Signal Process. **66**, 1037 (2018).
[45] A. S. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. S. Wein, (2017), arXiv:1712.10163 [math.ST].