We show that, contrary to popular assumptions, predictions from machine learning potentials built upon atom-density representations almost exclusively occur in an extrapolation regime — i.e., in regions of the representation space which lie outside the convex hull defined by the training set points. We then propose a perspective to rationalise the domain of robust extrapolation and accurate prediction of atomistic machine learning potentials in terms of the probability density induced by training points in the representation space.

Keywords: local density representation, extrapolation, interpolation, machine learning

Machine learning (ML) potentials for atomistic systems infer the mapping between configurations and a target objective function, e.g., the total energy of the system and/or the forces acting on each atom. These potentials are trained on a database of configurations whose objective function has been calculated via a computationally expensive, yet accurate, reference, e.g., density functional theory (DFT) methods. Following the training procedure, ML potentials offer predictions that are accurate with respect to, and much faster to compute than, the reference method A key aspect towards the making of accurate and efficient ML potentials lies in the choice of the representation function, which maps atomic coordinates to a set of numerical features. Among the most successful ones, we find expansions of local densities around atoms in the systems. In a nutshell, these representations are built upon the description of an atomic environment in terms of atom-centred distributions (encoding N-body correlations up to a desired order of N), which are approximated via a truncated expansion in radial and angular basis sets.

One notable characteristic of these representations is that they originate a high-dimensional feature space. The accuracy of the ML predictions is moreover generally observed to correlate with the representation’s dimensionality, when the other free parameters in the expansion are fixed. For this reason, there is an incentive to employ high-dimensional atomic-density representations when training ML potentials. Recent publications by Balestriero et al. and Yousefzadeh showcased that the predictions made by image recognition models whose inputs are high-dimensional, happen in an extrapolation regime, where interpolation and extrapolation regimes are formally defined according to a geometric criterion, in particular. Interpolation occurs for a point \( x^* \), whenever the latter belongs to the convex hull (CH) of a set of training points \( X \triangleq (x_1, x_2, ... x_M) \), if not, extrapolation occurs.

The above definition is employed also in the community of scientists applying ML methods to atomistic systems, and a common assumption is that the accuracy of ML potentials is strongly dependent on the fact that their predictions take place in an interpolation regime. When a ML potential accurately predicts the objective function for a structure outside the training database, this result is often interpreted as a sign that the atomic environment representations in the out-of-sample structures are "contained in", "covered by", or "interpolated between" points in the training set.

In this letter, we clarify the difference between the geometric definition of interpolation, and the alternative concept of a well-sampled region of the representation space. We furthermore showcase that, while the knowledge of whether the test point lies in an interpolation or extrapolation regime yields little information on the test accuracy of a ML potential, we can establish a strong connection between the test error incurred by ML potentials and the probability density function induced by the training set and computed on the test set.

First, we show that, for the example case of datasets and benchmarks widely quoted in the literature, extrapolation is the most common prediction regime for ML potentials. We then propose a protocol to measure the sampling density induced by the training data on points belonging to the test data set. This quantity is computed on a test point’s features as the log probability density of training points estimated via an adaptive k-nearest-neighbours algorithm. We show that such measure strongly correlates with the prediction errors incurred by ridge-regression potentials on test sets, thus providing an effective tool to identify low-accuracy regions in the representation space, and to rationalise the accuracy of ML potentials beyond whether they fall in an extrapolation or interpolation regime.

To draw general conclusions, we consider three datasets in our investigation. These comprise periodic and finite-size systems with different chemistry:

- The ice-water dataset by Monserrat et al., which was employed to test the transferability of a ML force-field trained on water configurations to the case of ice crystals. It encompasses a training set of forces and energies in 1000 liquid water configurations — corresponding to 192000 atomic environments — and a test set contain-
ing structures corresponding to 54 known ice phases – comprising 2847 atomic environments – which also includes all the experimentally verified ice structures.

- The Li, Mo, Ge, Si, Ni, Cu dataset by Zuo et al.\textsuperscript{12}, which was used to benchmark cost and accuracy of several ML force-field flavours. It gathers energies and forces in systems of the six different elements for their ground state crystalline bulk configuration, strained crystals, low Miller index surfaces, bulk structures sampled during ab initio molecular dynamics (MD) trajectories at different temperatures, and bulk structures with a vacancy, also sampled during ab initio MD trajectories. Configurations are then organised, according to a random 90:10 split, into a training and testing set.

- The Au\textsubscript{13} database, which was custom-built to probe the likelihood of extrapolation during a MD trajectory. It comprises five subsets of 1000 configurations of planar Au\textsubscript{13} nanocluster, with energies and forces labels, sampled every 3 ps during finite-temperature (50K, 100K, 200K, 300K, and 400K) MD runs where no structural rearrangements were observed; for further details we refer the reader to the Supplementary Information (SI), section C.

To associate features to atomic environments, we employ the Atom-Centered Symmetry Functions (ACSF)\textsuperscript{6,24} the Smooth Overlap of Atomic Positions (SOAP)\textsuperscript{12,28} or the Atomic Cluster Expansion (ACE)\textsuperscript{11,26} representation (see SI, sections A and B, and Tables S1–S6 for further detail), and adopt previously-reported set-ups, when these are available in the literature\textsuperscript{11,26,27,28}. We then transform the high-dimensional representations in a linear fashion via principal components analysis (PCA), and construct sets of P-dimensional representations which employ the first P principal components, using the same procedure as reported in Zeni et al.\textsuperscript{28} We then systematically investigate whether test points are contained within the CH of the training set, also as a function of the number of employed PCA components. This approach parallels emerging protocols in the literature where low-dimensional embedding of atom-density representations are employed as a tool to probe the similarity among structures\textsuperscript{20,21} and rationalise ML accuracy and transferability\textsuperscript{11,26,28}. Rather than computing the CH – a time-consuming task in high dimensions\textsuperscript{11} – we verify whether a test point \(x'\) can be expressed as a linear combination of the points in the training set \(\{x_i\}\)\textsubscript{i=1} constrained to non-negative coefficients \(\lambda_i\) summing up to one:

\[
x' = \sum_{i=1}^{M} \lambda_i x_i
\]

with \(\sum_{i=1}^{M} \lambda_i = 1\) \& \(\lambda_i \geq 0 \ \forall \ i = 1,...,M\).

The test point \(x'\) is in the CH of the training set if and only if the above can be satisfied; this can be verified efficiently via a linear programming approach\textsuperscript{22,23}.

Figure 1 reports the number of test atomic environments which fall within the CH induced by the training set as a function of the dimensionality \(P\) of the PCA representations, for the three databases described. For reference we show in Figure S1 the cumulative variance explained by 2 to 25 principal components. The test points are completely contained in the CH determined by the training set, when considering a projection in the space of the first two PCA components. Increasing the dimensionality of the embedding, the number of test points enclosed within the CH diminishes rapidly. Regardless of the database design, chemical nature of the system, and choice of representation (see also Figure S2 for additional benchmarks), embedding on a low yet sizeable \((P \sim 10-20)\), number of PCA
components results in an almost complete separation between each of the test points and the CH associated to the training points.

Following Figure 1, we note that low-dimensional projections of atom-density features fail in faithfully preserving the information about whether an atomic environment is contained within the CH determined by a set of other ones. At a more fundamental level, we highlight that ML potentials based on high-dimensional representations are very likely to carry out predictions in extrapolation regions. This is true not only when testing the transferability of the ML potential from one phase to another (i.e., the ice-water database), but also in apparently trivial MD trajectory where no structural rearrangements take place (i.e., the Zuo et al. and the Au13 database). Notably, the non-negligible portion of atomic environments found within the training points convex hull in the Zuo et al. dataset (Cu, Ni, Ge and Si curves in Figure 1B) actually consists of local atomic environments identical to those also present in the training set. In light of this, we conclude that accurate predictions do not imply that test points are contained within the high-dimensional CH enclosing the set of available training points.

After determining that extrapolation is likely to happen when using a high-dimensional input representation, we test whether a ML potential’s prediction effectively operates in such a high-dimensional space. To investigate if an accurate supervised learning potential projects data into a low-dimensional interpolation space, we train a regularised linear regression potential trained on the Zuo et al. dataset (see SI, section G for further info). We analyse the weights assigned to each feature by the regression algorithm following training, and map the relevance of each PCA feature towards diminishing the per-atom energy mean absolute error (MAE). We report the MAE of the potential as a function of the number of PCA components accounted for in the regression in Figure 2A. We refer the interested reader to Figure S1 for a report on the amount of variance explained, and to Figure S2 for the number of test points which are enclosed by the training points CH, as a function of the first P principal components. We observe that ridge regression potentials for the six single-element datasets do need to operate in a feature space where the majority of the test set atomic environments lie in a region of extrapolation to reach their best accuracy.

To generalise our conclusion, we focus on the Au13 database. We fit energies gathered from each single finite-temperature MD trajectory, and test the ML potential predictions on the structures observed during the other four finite-temperature MD sampling. Figures S3, S6 show that the MAE on energy reaches a plateau at \( P \sim 20 \) PCA components. We verify that, when accounting these many components to probe whether test points fall into the training CH, none appears therein. This analysis provides additional evidence that accurate ML potentials require a large enough number of features, which, in turn, make out-of-sample data almost surely lie in an extrapolation regime.

In so far, our analysis shows that ML potentials exploiting atom-density representation generally work in an extrapolation regime and yet they provide accurate predictions. To rationalise the effectiveness of ML potentials we hypothesise that the test points often lie in regions of the representation space that are sufficiently sampled by the training set distribution. To test this hypothesis, we estimate the probability density generated by the training set points, which we call sampling density, on the locations of representation space where test points lie.

We do so by using an adaptive k-nearest-neighbour density estimation, which works as follows. The test point \( x^* \) is virtually added to the training set and its \( k^* \) nearest training points are found. The density is then computed as

\[
\rho(x^*) = \frac{k^* - 1}{MV^*},
\]

where \( M \) is the size of the training set and \( V^* \) is the volume occupied by the first \( k^* \) training point neighbours.

The number \( k^* \) of training neighbours considered is chosen adaptively for each test point as to maximise the accuracy of the estimate as described in Rodriguez et al.\(^{[33]}\) The volume...
$V^*$ is computed as the volume of a hypersphere in dimension $d$, where $d$ is the intrinsic dimension of the training data manifold computed via the TwoNN estimator. This way, we obtain a measure that indicates how much test set points are well-sampled by the training set density.

We assess the relationship existing between such measure and the error incurred by a ridge-regression potential based on ACE representations trained on the data that generates the density distribution when tested on points in the test set. We do so for the Zuo et al. training and test databases and for the Au$_{13}$ database introduced previously. In the latter case, each potential is trained on data from a MD trajectory at a single temperature, and tested on data coming from the other four MD trajectories. In Figure 3 we report the difference between the MAE found in test sets and an asymptotic minimum MAE. The minimum achievable MAE is approximately estimated as the bottom percentile error that is incurred by the potential on the training set. The MAE is averaged and binned over the log probability density of local atomic environment representations in the test sets and using an averaging window of size 5 and bin spacing of 5 (only bins containing at least 2% of the data are displayed). Figure 4 highlights how the log probability density correlates with the binned difference between test and convergence training MAE on forces for points outside the training set, for all databases considered (see also Figures S7-S9). The metric we introduce can therefore offer an estimate of the degree by which an out-of-sample atomic environment lies within a well-sampled region of the representation space and, more importantly, correlates with the error incurred by trained regression potentials. We nevertheless find that this metric is dependent on the choice of the representation (Figure S8-S9) and that the precise relationship with the $\Delta$MAE is model- and system-dependent.

To conclude, we tested whether ML potentials exploiting high dimensional atom-density representations generally make predictions in an interpolation or extrapolation regime, for the case of popular databases in the literature as well as custom-built ones. We follow the definition spelled out by Balestriero et al., where extrapolation occurs if the test structure lies outside the high-dimensional CH which encloses the set of training structure, and find - contrary to the assumptions in our community - that the large majority of the test predictions take place in an extrapolation regime when employing high-dimensional local atomic density representations (ACSF, SOAP, ACE). We envision future endeavours to probe whether other popular representations and frameworks in the context of ML for atomistic systems work in an extrapolation regime.

At a more fundamental level, we demonstrate the need to revisit previous interpretations relating ML potentials’ accuracy and the geometrical relationships between test and training atomic environments. To understand why ML potentials exploiting atom-density representations work in an extrapolation regime, we thus propose an alternative perspective, which relates their accuracy to the probability density induced by training points in the representation space.

We believe that the criterion suggested to verify whether a training set is well-suited to enable accurate predictions at a target point in representation space will promote the development of novel database generation routines, data point selection algorithms, and a more rigorous understanding of why established protocols work. Finally, we highlight the generality of the training dataset sampling density analysis, which could be applied in other domains where high-dimensional ML models are widespread and successful, e.g., image recognition, diagnostics, therapeutic development and healthcare delivery.

ACKNOWLEDGEMENTS

C.Z. gratefully acknowledges support by the European Union’s Horizon 2020 research and innovation program (Grant No. 824143, MaX ‘MAterials design at the eXascale’ Centre of Excellence). K.R. has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Marie Curie Individual Fellowship Grant agreement No. 890414). The authors thank Matteo Carli, Stefano De Gironcoli, Piero Gasparotto, Federico Grasselli, Andrea Grisafi, Giulio Imbalzano, and Natasha Lopanitsyna for useful discussions. In memory of Alessandro De Vita.

DATA AVAILABILITY

Inputs and outputs to test whether a test atomic environment falls into the CH of the training dataset will be deposited on the Materials Cloud after acceptance.

CONFLICT OF INTEREST

A.A. is currently employed at Hoffman La Roche and has contributed to this work outside of his Roche Post Doctoral Fellowship.
Fellowship project.

BIBLIOGRAPHY

1. F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Chemical Reviews 121, 9759–9815 (2021).
2. V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Chemical Reviews 121, 10073–10141 (2021).
3. O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K. R. Müller, Chemical Reviews 121, 10142–10186 (2021).
4. J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K. R. Müller, and A. Tkatchenko, Chemical Reviews 121, 9816–9872 (2021).
5. C. Zeni, K. Rossi, A. Glielmo, and F. Baletto, Advances in Physics: X 4, 1654919 (2019).
6. J. Behler and M. Parrinello, Physical Review Letters 98, 146401 (2007).
7. A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters 104, 136403 (2010).
8. A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, Journal of Computational Physics 285, 316 (2015).
9. A. V. Shapeev, Multiscale Modeling & Simulation 14, 1153 (2016).
10. M. J. Willatt, F. Musil, and M. Ceriotti, Journal of Chemical Physics 150, 154110 (2019).
11. R. Drautz, Phys. Rev. B 99, 014104 (2019).
12. 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., The Journal of Physical Chemistry A 124, 731 (2020).
13. R. Balesctiero, J. Pesenti, and Y. LeCun, arXiv:2101.09485 (2021).
14. R. Yousefzadeh, Arxiv arXiv:2101.09849 (2021).
15. A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X 8, 041048 (2018).
16. K. Rossi, V. Jurášková, R. Wischert, L. Garel, C. Cournimourot, and M. Ceriotti, Journal of Chemical Theory and Computation 16, 5139–5149 (2020).
17. A. Tkatchenko, Nat Commun 11, 4125 (2020).
18. B. Monserrat, J. G. Brandenburg, E. A. Engel, and B. Cheng, Nature Communications 11, 5757 (2020).
19. T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, J. Chem. Phys. 148, 241725 (2018).
20. A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Science Advances 3, e1701816 (2017).
21. C. Zeni, K. Rossi, T. Pavlouidis, K. Kioseoglou, S. DeGrironcoli, R. E. Palmer, and F. Baletto, Nature Communications 12, 6065 (2021).
22. E. Engel, A. Anelli, A. Hofstetter, F. Paruzzo, L. Emsley, and M. Ceriotti, Phys Chem Chem Phys 21, 23385 (2019).
23. Y. Shao, F. M. Dietrich, C. Nettelblad, and C. Zhang, The Journal of Chemical Physics 155, 204108 (2021).
24. A. Singrauer, J. Behler, and C. Dellago, J. Chem. Theory Comput. 15, 1827 (2019).
25. L. Himanen, M. O. J. Jäger, E. V. Moroooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Computer Physics Communications 247, 106949 (2020).
26. G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, arXiv, 1911.03550 (2021).
27. B. Cheng, E. Engel, A. J. Behler, C. Dellago, and M. Ceriotti, Proc. Natl Acad. Sci. 116, 1110 (2019).
28. C. Zeni, K. Rossi, A. Glielmo, and S. D. Gironcoli, Journal of Chemical Physics 154, 224112 (2021).
29. E. A. Engel, A. Anelli, M. Ceriotti, C. J. Pickard, and R. J. Needs, Nature Communications 9, 2173 (2018).
30. A. Anelli, E. A. Engel, C. J. Pickard, and M. Ceriotti, Physical Review Materials 2, 103804 (2018).
31. F. Nielsen, Information Processing Letters 59, 255 (1996).
32. P. Pardalos, Y. Li, and W. Hager, Computers & Mathematics with Applications 29, 23 (1995).
33. A. Rodriguez, M. d’Errico, E. Facco, and A. Laio, J. Chem. Theory Comput. 14, 1206 (2018).
34. E. Facco, M. d’Errico, A. Rodriguez, and A. Laio, Scientific Reports 7 (2017).
35. M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters 108, 058301 (2012).
36. K. T. Schütt, H. E. Sauceda, P. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem Phys 148, 241722 (2018).
37. S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, Computer Physics Communications 240, 38 (2019).
38. T. Xie and J. C. Grossman, Phys. Rev. Lett. 120, 145301 (2018).
39. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, in Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML’17 (JMLR.org, 2017) pp. 1263–1272.
40. A. Glielmo, C. Zeni, and A. De Vita, Physical Review B 97, 1 (2018).
41. G. Seifert, J. Phys. Chem. A 111, 5609 (2007).
42. B. Aradi, B. Hourahine, and T. Frauenheim, J. Phys. Chem. A 111, 5678 (2007).
43. V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Cournimourot, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, Comput. Phys. Commun. 236, 214 (2019).
SUPPLEMENTARY INFORMATION

A. Representation for atomistic systems

In the context of ML potentials, information about the configuration of atoms is often represented through atomic environment representations. The atomic environment representation for atom $i$ encodes information about the position and atomic species of all atoms $j$ contained in a ball of radius $r_c$ centred on $i$. In this investigation, we considered three frameworks to represent atomic environments that found a widespread and successful use in the literature. A brief resume of their salient characteristic is discussed in detail in the next corresponding subsections.

1. Atom-centered Symmetry Functions representation

Atom-centered Symmetry Functions representation (ACSF) encode information about the arrangement of atoms within an atomic environment through a discrete set of features which map 2- and 3-body correlations. Behler-Parrinello\textsuperscript{[5]} Radial symmetry functions $G_r^2$ are calculated as sums of Gaussians of width $1/\eta$ and centred at $r_i$:

$$G_r^2 = \sum_{j|\rho_j \leq r_c} e^{-\eta(r_{ij} - r_c)^2} f_c(r_{ij}).$$

Angular symmetry functions are built as sums of cosine functions of the angles $\theta_{ijk} = acos(r_{ik}/r_{jk})$:

$$G_3 = 2^{1-\xi} \sum_{j,k | \rho_{jk} \leq r_c} (1 + \lambda \cos(\theta_{ijk}))^\xi e^{-\eta(r_{ij} + r_{ik} + r_{jk})^2} f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}),$$

where $\lambda$, whose value ranges between 1 and -1, tunes the maximum of the cosine function, and $\xi$ tunes the angular resolution. We compute ACSF features using the n2p2 package.\textsuperscript{[26]}

2. Atom density representations

Atom-density representations describe the atomic density $\rho_i(r)$ as a sum of functions $u(r_{ij} - r)$ centred on each atom $j$ within a cut-off $r_c$ around the atom $i$:

$$\rho_i(r) = \sum_{j|\rho_j \leq r_c} u(r_{ij} - r),$$

where $r_{ij}$ indicates the vector $(r_j - r_i)$.

Under the Smooth Overlap of Atomic Positions (SOAP) representation framework\textsuperscript{[21]} $u(r_{ij} - r)$ takes the form of a Gaussian of width $\sigma$:

$$\rho_i(r) = \sum_{j|\rho_j \leq r_c} e^{-|r_{ij} - r|^2 / \sigma^2}.$$
B. Representation parameters

Table S1 gathers the SOAP representation parameters employed for the Monserrat et al. [18] dataset, and mirror the ones reported in the original work.

| system | $r_c$ | $n_{max}$ | $l_{max}$ | $\sigma$ |
|--------|-------|-----------|-----------|---------|
| H2O    | 6     | 6         | 6         | 0.5     |

TABLE S1. Resume of the SOAP representation parameters adopted to expand atom-densities around atom centred environments in the Monserrat et al. [18] database.

Tables S2 and S3, respectively, gather the 2- and 3-body ACSF parameters employed for the Monserrat et al. [18] dataset, and mirror the ones reported in the original work.

| central atom | neighbour 1 | neighbour 2 | $\eta$ | $\lambda$ | $\xi$ | $r_c$ |
|--------------|-------------|-------------|-------|----------|------|-------|
| H            | O           | H           | 0.2   | 1.0      | 1.0  | 12.00 |
| O            | H           | H           | 0.07  | 1.0      | 1.0  | 12.00 |
| H            | O           | H           | 0.07  | 1.0      | 1.0  | 12.00 |
| O            | H           | H           | 0.07  | 1.0      | 1.0  | 12.00 |
| H            | O           | H           | 0.03  | 1.0      | 1.0  | 12.00 |
| O            | H           | H           | 0.03  | 1.0      | 1.0  | 12.00 |
| H            | O           | H           | 0.03  | -1.0     | 1.0  | 12.00 |
| O            | H           | H           | 0.03  | -1.0     | 1.0  | 12.00 |
| H            | O           | H           | 0.03  | 1.0      | 4.0  | 12.00 |
| O            | O           | H           | 0.03  | 1.0      | 4.0  | 12.00 |
| H            | O           | O           | 0.03  | 1.0      | 4.0  | 12.00 |
| O            | O           | O           | 0.03  | -1.0     | 4.0  | 12.00 |
| H            | O           | O           | 0.03  | 1.0      | 4.0  | 12.00 |
| H            | O           | O           | 0.03  | -1.0     | 4.0  | 12.00 |
| H            | O           | O           | 0.03  | 1.0      | 4.0  | 12.00 |
| O            | O           | O           | 0.03  | -1.0     | 4.0  | 12.00 |

TABLE S2. Resume of the 2-body ACSF representation parameters adopted to expand atom-densities around atom centred environments in the Monserrat et al. [18] database.

| central atom | neighbour 1 | neighbour 2 | $\eta$ | $r_c$ |
|--------------|-------------|-------------|-------|------|
| H            | H           | H           | 0.001 | 0.0  |
| H            | H           | H           | 0.01  | 0.0  |
| H            | H           | H           | 0.03  | 0.0  |
| H            | H           | H           | 0.06  | 0.0  |
| H            | H           | H           | 0.15  | 1.9  |
| H            | H           | H           | 0.30  | 1.9  |
| H            | H           | H           | 0.60  | 1.9  |
| H            | H           | H           | 1.50  | 1.9  |
| H            | O           | H           | 0.001 | 0.0  |
| O            | H           | H           | 0.01  | 0.0  |
| O            | H           | H           | 0.03  | 0.0  |
| O            | H           | H           | 0.06  | 0.0  |
| O            | H           | H           | 0.15  | 0.9  |
| O            | H           | H           | 0.30  | 0.9  |
| O            | H           | H           | 0.60  | 0.9  |
| H            | O           | O           | 1.50  | 0.9  |
| O            | H           | O           | 0.001 | 0.0  |
| O            | O           | H           | 0.01  | 0.0  |
| O            | O           | H           | 0.03  | 0.0  |
| O            | H           | O           | 0.06  | 0.0  |
| O            | H           | O           | 0.15  | 4.0  |
| O            | O           | O           | 0.30  | 4.0  |
| O            | O           | O           | 0.60  | 4.0  |
| O            | O           | O           | 1.50  | 4.0  |

TABLE S3. Resume of the 3-body ACSF representation parameters adopted to expand atom-densities around atom centred environments in the Monserrat et al. [18] database.

A resume of the SOAP and ACE parameters for the case of the Zuo et al. [12] dataset are reported in Table S4 and S5.

| element | $r_c$ | $n_{max}$ | $l_{max}$ | $\sigma$ |
|---------|-------|-----------|-----------|---------|
| Cu      | 3.9   | 8         | 8         | 0.5     |
| Ni      | 3.9   | 8         | 8         | 0.5     |
| Mo      | 5.2   | 8         | 8         | 0.5     |
| Li      | 4.8   | 8         | 8         | 0.5     |
| Si      | 5.4   | 8         | 8         | 0.5     |
| Ge      | 5.4   | 8         | 8         | 0.5     |

TABLE S4. Resume of the SOAP representation parameters adopted to expand atom-densities around atom centered environments in the Zuo et al. [12] database.

| element | $r_c$ | $n_{max} + l_{max}$ | $N$ |
|---------|-------|----------------------|-----|
| Cu      | 3.9   | 10                   | 5   |
| Ni      | 3.9   | 10                   | 5   |
| Mo      | 5.2   | 10                   | 5   |
| Li      | 4.8   | 10                   | 5   |
| Si      | 5.4   | 10                   | 5   |
| Ge      | 5.4   | 10                   | 5   |

TABLE S5. Resume of the ACE representation parameters adopted to expand atom-densities around atom centered environments in the Zuo et al. [12] database.
Table S6 recap the ACE representation parameters employed for the Au\textsubscript{13} dataset.

| element | \( r_c \) | \( n_{\text{max}} + l_{\text{max}} \) | \( N \) |
|---------|-----------|-----------------|-----|
| Au      | 5         | 10              | 5   |

**TABLE S6. Resume of the ACE representation parameters adopted to expand atom-densities around atom centered environments for the Au configurations.**

**C. Au\textsubscript{13} DATASET GENERATION DETAILS**

Data is gathered from MD simulations of an Au\textsubscript{13} nanocluster ran for 3ns (configurations are stored every 3ps, for a total of 1000 de-correlated structures) at temperatures of 50 K, 100 K, 200 K, 300 K, and 400 K. The initial frame of each trajectory corresponds to the relaxed global minimum structure for Au\textsubscript{13}. To simulate the trajectories, we integrate the dynamics with a 3 fs timestep, enforce an NVT ensemble via a velocity rescaling thermostat, with a time-scale of 200 fs, model interatomic interactions within the Density functional tight binding framework\textsuperscript{43,44}, and use i-P\textsuperscript{45} as the MD driver.

**D. PCA explained variance**

![Fraction of unexplained variance](image)

**FIG. S4.** Fraction of unexplained variance in the data explained by the first \( P \) PCA for feature spaces deriving from the system-dependent atom-density representations. Data relative to both the SOAP (cyan pentagones) and ACSF (grey diamonds) features are displayed in Panel A. Panels B and C report ACE features with \( n_{\text{max}} + l_{\text{max}} = 10 \) and \( N = 5 \).
E. Additional Convex Hull Benchmarks

We report about tests regarding the number of training points inside the CH defined by the test set for the case of Ni, Li, and Si databases when exploiting a SOAP and an ACE representation with a different number of radial and angular basis set components w.r.t. the ones employed in the main text. In particular, Figure S2 reports the same plots as in Figure 1B, but for ACE representations with $n_{\text{max}} + l_{\text{max}} = 6, 8$, and for SOAP representations with $n_{\text{max}}, l_{\text{max}} = 4, 6, \text{and } 8$. We find that the number of test points within the convex hull enclosing the training set is similar across the different SOAP and ACE set-ups. This result holds both for the analysis in the full-dimensional space and in the low-dimensional PCA embedding.

FIG. S5. Fraction of atomic environments in the test set which are contained in the N-dimensional CH enclosing the points in the training set, featurised according to the first $P$ PCA components of high dimensional features. Panel A: SOAP representation using $n_{\text{max}} = 8, l_{\text{max}} = 8$. Panel B: SOAP representation using $n_{\text{max}} = 6, l_{\text{max}} = 6$. Panel C: SOAP representation using $n_{\text{max}} = 4, l_{\text{max}} = 4$. Panel D: ACE representation using $n_{\text{max}} + l_{\text{max}} = 8$. Panel E: ACE representation using $n_{\text{max}} + l_{\text{max}} = 6$. 
F. Effect of identical atomic environments on the convex hull computation

Figure S3 shows the relative abundance of Euclidean distances between ACE representations in each Zuo et al.12 test and training set. For the case of the Cu, Ni, Si, and Ge datasets, respectively 6%, 7%, 18% and 17% of such distances are smaller than $10^{-12}$, whereas this value goes to less than 0.1% for the Li and Mo datasets. We remark that identical atomic environments can be found in different configurations; this happens e.g., when comparing atoms in relaxed crystalline bulk with atoms in relaxed surfaces.

Figure S4 highlights the fact that test atomic environments for which exists an almost identical atomic environment in the training set are likely to be counted as within the CH defined by the training set. When this anomalous atomic environments are not accounted for, the fraction of points inside the CH rapidly and monotonically goes to 0 as the number of PCA components increases. This is even clearer in Figure S5, which shows the fraction of test ACE representations inside the CH defined by the training ACE representation for test representations with no distance smaller than $10^{-12}$ from test representations. We observe that, once we remove from the test set environments that are almost identical to ones in the training set, the fraction of points inside the CH goes consistently to zero for all six elemental datasets at around 15 PCA components.

![Histogram of distances between training and test set descriptors](image1)

FIG. S6. Histogram of distances between training and test set descriptors. The ACE representation employs $N = 5$ and $n_{\text{max}} + l_{\text{max}} = 10$ as hyper parameters.

![Fraction of atomic environments in the Ni test set which are contained in the CH](image2)

FIG. S7. Fraction of atomic environments in the Ni test set which are contained in the N-dimensional CH enclosing the points in the training set, featurised according to the first $P$ PCA components of an ACE representation using $n_{\text{max}} + l_{\text{max}} = 10$. The dotted line refers to atomic environments in the test set that are distant less than $10^{-12}$ from at least one environment in the training set, the full line to the other environments in the test set.

![Fraction of atomic environments in the test set which are contained in the CH](image3)

FIG. S8. Fraction of atomic environments in the test set which are contained in the $P$-dimensional CH enclosing the points in the training set, featurised according to the first $P$ PCA components of an ACE representation using $n_{\text{max}} + l_{\text{max}} = 10$. Atomic environments in the test set are less than $10^{-12}$ distant from at least one environment in the test set have been excluded from the CH evaluation.
G. Ridge Regression for Energy and Forces

To train a machine learning model on forces and energies we build a 2D matrix \( Y \), whose elements \( Y_i \) correspond to atomic environments \( S \) within a configuration \( i \):

\[
Y_i = [E_i, f_x^i, f_y^i, f_z^i, \ldots, f_x^S, f_y^S, f_z^S].
\]

In the above \( f_c^S \) labels the \( c \)-component of the force vector acting on the atom \( s \) in the structure \( i \).

The matrix of explanatory variables \( Q \) then takes the form of a 3D tensor with elements \( Q_i \) pertaining to structure \( i \):

\[
Q_i = \left[ q_i, -\frac{\partial q_i}{\partial x_1}, -\frac{\partial q_i}{\partial y_1}, \ldots, -\frac{\partial q_i}{\partial x_S}, -\frac{\partial q_i}{\partial y_S}, -\frac{\partial q_i}{\partial z_S} \right],
\]

where \( q_i \) is the sum over all atoms \( j \) in configuration \( i \) of the atomic environment representation \( q(\rho_j) \).

When training a ridge regression machine learning model on energy and forces we translate the learning problem into the following closed formula:

\[
Y = Q W + \varepsilon,
\]

where \( Y \) is the matrix of dependent variables, \( Q \) is the matrix of explanatory variables, \( W \) is the parameter matrix that weights \( Q \), and \( \varepsilon \) is a vector of error terms which takes into consideration the possible presence of hidden variables which affect \( Y \) while not being encoded in \( Q \). For a given training set \( D = \{Y_i, Q_i\} \ i = 1, \ldots, M \), one finds the regression weights \( W \) analytically by solving:

\[
W = (Q^T Q + \gamma I)^{-1} Q^T Y,
\]

where \( \gamma \) is the ridge parameter.

H. Regression weights for the \( \text{Au}_{13} \) database

FIG. S9. Cumulative Mean Absolute Error on energy prediction when accounting for the contribution of the first \( P \) PCA of the high-dimensional feature spaces deriving from an ACE representation for a ridge-regression potential trained on the \( \text{Au}_{13} \) database at 50 K (Panel A), 100 K (Panel B), 300 K (Panel C), 400 K (Panel D).
I. Δ MAE and log density correlation histograms

**FIG. S10.** Two-dimensional histogram reporting the correlation between the MAE difference between the force predictions incurred by ridge regression potentials in the validation set and in the most densely sampled regions of the test set, and the probability density estimate for the training points in representation space, computed on test points. Darker colours indicate larger number of points in that bin. Points mark the average density bin. The representation is an ACE one, with $N = 5$ and $n_{\text{max}} + l_{\text{max}} = 10$.

**FIG. S11.** Two-dimensional histogram reporting the correlation between the MAE difference between the force predictions incurred by ridge regression potentials in the validation set and in the most densely sampled regions of the test set, and the probability density estimate for the training points in representation space, computed on test points. Darker colours indicate larger number of points in that bin. Points mark the average density bin. The representation is an ACE one, with $N = 5$ and $n_{\text{max}} + l_{\text{max}} = 8$. 
FIG. S12. Two-dimensional histogram reporting the correlation between the MAE difference between the force predictions incurred by ridge regression potentials in the validation set and in the most densely sampled regions of the test set, and the probability density estimate for the training points in representation space, computed on test points. Darker colours indicate larger number of points in that bin. Points mark the average density bin. The representations is an ACE one, with $N = 5$ and $n_{\text{max}} + l_{\text{max}} = 6$. 