npj | Computational Materials

www.nature.com/npjcompumats

PERSPECTIVE OPEN

Check for updates

Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning

Marcel F. Langer[1,2], Alex Goeßmann[1,3] and Matthias Rupp[1,4,5]✉

Computational study of molecules and materials from first principles is a cornerstone of physics, chemistry, and materials science, but limited by the cost of accurate and precise simulations. In settings involving many simulations, machine learning can reduce these costs, often by orders of magnitude, by interpolating between reference simulations. This requires representations that describe any molecule or material and support interpolation. We comprehensively review and discuss current representations and relations between them. For selected state-of-the-art representations, we compare energy predictions for organic molecules, binary alloys, and Al–Ga–In sesquioxides in numerical experiments controlled for data distribution, regression method, and hyperparameter optimization.

npj Computational Materials (2022)8:41 ; https://doi.org/10.1038/s41524-022-00721-x

INTRODUCTION

Quantitative modeling of atomic-scale phenomena is central for scientific insights and technological innovations in many areas of physics, chemistry, and materials science. Solving the equations that govern quantum mechanics (QM), such as Schrödinger's or Dirac's equation, allows accurate calculation of the properties of molecules, clusters, bulk crystals, surfaces, and other polyatomic systems. For this, numerical simulations of the electronic structure of matter are used, with tremendous success in explaining observations and quantitative predictions.

However, the high computational cost of these ab initio simulations (Supplementary Note 1) often only allows investigating from tens of thousands of small systems with a few dozen atoms to a few large systems with thousands of atoms, particularly for periodic structures. In contrast, the number of possible molecules and materials grows combinatorially with the number of atoms: 13 or fewer C, N, O, S, Cl atoms can form a billion possible molecules[1], and for 5-component alloys, there are more than a billion possible compositions when choosing from 30 elements (Supplementary Note 2). This limits systematic computational study and exploration of molecular and materials spaces. Similar considerations hold for ab initio dynamics simulations, which are typically restricted to systems with a few hundred atoms and sub-nanosecond timescales.

Such situations require many simulations of systems correlated in structure, implying a high degree of redundancy. Machine learning[2,3] (ML) exploits this redundancy to interpolate between reference simulations[4–7] (Fig. 1). This ansatz replaces most ab initio simulations by ML predictions, based on a small set of reference simulations. Effectively, it maps the problem of repeatedly solving a QM equation for many related systems onto a regression problem. This approach has been demonstrated in benchmark settings[4,8,9] and applications[5,10,11], with reported speed-ups between zero to six orders of magnitude[12–15]. It is currently regarded as a highly promising avenue towards extending the scope of ab initio methods.

The most relevant aspect of ML models for interpolation of QM simulations (QM/ML models) after data quality (Supplementary Note 3) is the definition of suitable input features, that is, *representations* of atomistic systems. Representations define how systems relate to each other for regression and are the subject of this perspective.

Scope and structure

QM/ML models require a space in which interpolation takes place. Such spaces can be defined explicitly, often as vector spaces, or implicitly, for example, via a kernel function in kernel-based machine learning[16,17]. *This work reviews and compares explicit Hilbert-space representations of finite and periodic polyatomic systems for accurate interpolation of QM observables via ML*, focusing on representations that satisfy the requirements discussed in section "Requirements" and energy predictions.

This excludes features that do not encode all input information, such as atomic numbers and coordinates, for example, descriptors or fingerprints used in cheminformatics and materials informatics to interpolate between experimental outcomes[18], and implicit representations learned by end-to-end deep neural networks[19–37] or defined via direct kernels between systems[38–42] (Supplementary Notes 4 and 10).

Characteristics and requirements of representations are discussed in the sections "Role and types of representations" and "Requirements" followed by a short description of a unified mathematical framework for representations ("A unified framework"). Specific representations are delineated (benchmarked ones in "Selected representations", others in "Other representations"), qualitatively compared ("Analysis"), and empirically benchmarked ("Empirical comparison"). We conclude with an outlook on open problems and possible directions for future research in section "Conclusions and outlook". See Table 1 for a glossary of covered representations and technical terms.

[1]NOMAD Laboratory, Fritz Haber Institute of the Max Planck Society, Berlin, Germany. [2]Machine Learning Group, Technische Universität Berlin, Berlin, Germany. [3]Institute of Mathematics, Technische Universität Berlin, Berlin, Germany. [4]Citrine Informatics, Redwood City, CA, USA. [5]Department of Computer and Information Science, University of Konstanz, Konstanz, Germany. ✉email: mrupp@mrupp.info

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

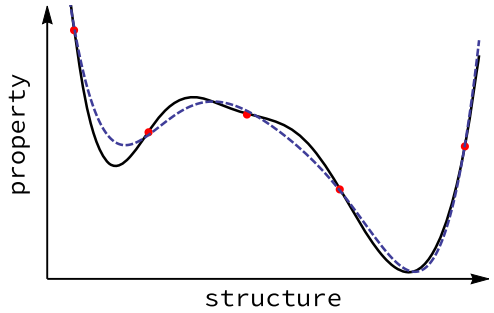


Fig. 1 Sketch illustrating the interpolation of quantum-mechanical simulations by machine learning. The horizontal axis represents chemical or materials space, the vertical axis the predicted property. Instead of conducting many computationally expensive ab initio simulations (solid line), machine learning (dashed line) interpolates between reference simulations (dots).

Table 1. Glossary.

Acronym	Meaning
Representations	
ACE	Atomic cluster expansion
BoB	Bag of bonds
BS	Bispectrum
CM	Coulomb matrix
DECAF	Density-encoded canonically-aligned fingerprint
FCHL	Faber-Christensen-Huang-von Lilienfeld
GM	Gaussian moments
HDAD	Histograms of distances, angles, and dihedral angles
IDMBR	Inverse-distance many-body representation
MBTR	Many-body tensor representation
MILAD	Moment invariants local atomic descriptors
MOB	Molecular orbital basis
MTP	Moment tensor potential
NICE	*N*-body iterative contraction of equivariants
OMF	Overlap matrix fingerprint
SF	Symmetry function
SOAP	Smooth overlap of atomic positions
WST	Wavelet scattering transform
Methodology	
GPR	Gaussian process regression
HP	Hyperparameter (free parameter)
KRR	Kernel ridge regression
MAE	Mean absolute error
ML	Machine learning
QM	Quantum mechanics
QM/ML	ML model for accurate prediction of QM data
RMSE	Root mean squared error
system	Polyatomic system, e.g., a molecule or a crystal

Related work

Studies of QM/ML models often compare their performance estimates with those reported in the literature. While such comparisons have value, they entertain considerable uncertainty due to different datasets, learning algorithms, including choice of hyperparameters (HPs, free parameters), sampling, validation procedures, and reported quantities. Accurate, reliable performance estimates require a systematic comparison that controls for the above factors, which we perform in this work.

Table 2. Related work. See Supplementary Note 5 for details.

	Reference											
	91	84	150	65	111	151	152	153	104	122	68	here
Finite systems	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓
Periodic systems	×	✓	✓	×	✓	×	✓	×	×	×	✓	✓
Other properties	✓	✓	×	×	×	✓	✓	✓	✓	×	✓	×
Numerical HPs	×	✓	✓	×	×	✓	×	×	×	✓	×	✓
Structural HPs	×	×	×	×	×	×	×	×	×	×	×	✓
Regression HPs	✓	✓	✓	✓	✓	✓	×	✓	×	✓	×	✓
Timings	×	×	✓	✓	×	×	×	×	×	×	×	✓

Finite systems: study uses datasets of finite systems, such as molecules or clusters; Periodic systems: uses datasets of periodic systems, such as crystalline materials; Other properties: evaluate properties other than energy or its derivatives; numerical/structural/regression HPs: whether numerical hyperparameters of representations, structural hyperparameters of representations, or regression hyperparameters are optimized automatically.

Several recent studies systematically measured and compared prediction errors of representations (Table 2). We distinguish between studies that automatically (as opposed to manually) optimize numerical HPs of representations, for example, the width of a normal distribution; structural HPs of representations, for example, choice of basis functions; and HPs of the regression method, for example, regularization strength. Supplementary Note 5 discusses the individual studies from Table 2.

ROLE AND TYPES OF REPRESENTATIONS

An N-atom system formally has $3N-6$ degrees of freedom. Covering those with M samples per dimension requires M^{3N-6} reference calculations, which is infeasible except for the smallest systems. How then is it possible to learn high-dimensional energy surfaces?

Part of the answer is that learning the whole energy surface is unnecessary, as configurations high in energy become exponentially unlikely—it is sufficient to learn low-energy regions. Another reason is that the regression space's formal dimensionality is less important than the data distribution in this space. (Supplementary Note 6) Representations can have thousands of dimensions, but their effective dimensionality[43] can be much lower if they are highly correlated. The role of representations is, therefore, to map atomistic systems to spaces amenable to regression. These spaces, together with the data's distribution, determine the efficiency of learning.

We distinguish between *local* representations that describe parts of an atomistic system, such as atoms in their environment[8,44], and *global* ones that describe the whole system. For global representations, represented systems are either *finite*, such as molecules and clusters, or *periodic*, such as bulk crystals and surfaces (Table 3).

Local representations are directly suitable for local properties, such as forces, nuclear magnetic resonance shifts, or core-level excitations[45], which depend only on a finite-size environment of an atom. Extensive global properties (Supplementary Note 7) such as energies can be modeled with local representations via additive approximations, summing over atomic contributions (Supplementary Note 8). Since local representations require only finite support, it does not matter whether the surrounding system is finite or periodic. Global representations are suited for properties of the whole system, such as energy, band gap, or the polarizability tensor. Since periodic systems are infinitely large,

Table 3. Types of representations.

Category	Representation
Local	ACE, BS, DECAF, FCHL, GM, MILAD, MTP, NICE, OMF, SF, SOAP, WST
Global (finite)	BoB, CM, HDAD, IDMBR, MBTR, MOB, OMF, WST
(periodic)	MBTR

We distinguish between local (atoms in their environment) and global (holistic, whole system) representations, as well as between representations for finite (molecules, clusters) and periodic systems (bulk crystals, surfaces). Local representations have finite support, and thus do not need to distinguish between finite and periodic systems. See Glossary (Table 1) for abbreviations.

global representations usually need to be designed for or adapted to these. Trade-offs between local and global representations are discussed in the section "Analysis".

Historically, interpolation has been used to reduce the effort of numerical solutions to quantum problems from the beginning. Early works employing ML techniques such as Tikhonov regularization and reproducing kernel Hilbert spaces in the late 1980s and throughout the 1990s were limited to small systems[46–49]. Representations for high-dimensional systems appeared a decade later[8,9,50], underwent rapid development, and constitute an active area of research today. Table 4 presents an overview.

REQUIREMENTS

Figures of merit for QM/ML models include computational efficiency, predictive accuracy, and sample efficiency, that is, the number of reference simulations required to reach a given target accuracy. Imposing physical constraints on representations improves their sample efficiency by removing the need to learn these constraints from the training data. The demands of speed, accuracy, and sample efficiency give rise to specific requirements, some of which depend on the predicted property:

(i) *Invariance* to transformations that preserve the predicted property, including (a) changes in atom indexing (input order, permutations of like atoms), and often (b) translations, (c) rotations, and (d) reflections. Predicting tensorial properties requires (e) *covariance* (equivariance) with rotations[6,25,26,29,51–54]. Dependence of the property on a global frame of reference, for example, due to the presence of a non-isotropic external field, can affect variance requirements.

(ii) *Uniqueness*, that is, variance against all transformations that change the predicted property: Two systems that differ in property should be mapped to different representations. Systems with equal representation that differ in property introduce errors[55–57]: Because the ML model cannot distinguish them, it predicts the same value for both, resulting in at least one erroneous prediction. Uniqueness is necessary and sufficient for reconstruction, up to invariant transformations, of an atomistic system from its representation[44,58].

(iii) (a) *Continuity*, and ideally (b) *differentiability*, with respect to atomic coordinates.

Discontinuities work against the regularity assumptions in ML models, which try to find the least complex function compatible with the training data. Intuitively, continuous functions require less training data than functions with jumps. Differentiable representations enable differentiable ML models. If available, reference gradients can further constrain the interpolation function ("force matching"), improving sample efficiency[59–61].

Table 4. Overview of representations.

		References		
Year	Repr.	Orig.	Dev.	Avail.
2007	SF	8	66,77–80	84,154
2010	BS	9	82,155,156	157
2012	CM	4	45,55,64,86,87	84,158
2013	SOAP	44	6,9,15,58,83,85,108,155,156,159	84,160
2013	OMF	62	103,104	–
2015	BoB	88	89	161
2015	WST	94	95–100	162
2016	MTP	74	127,163–165	166
2017	MBTR	73	111	84,158
2017	HDAD	91	–	–
2018	DECAF	106	–	167
2018	FCHL	92	93	161
2018	IDMBR	90	–	168
2018	MOB	63	105,169,170	–
2019	ACE	101	53,102	171,172
2020	NICE	71	–	173
2020	GM	75	–	–
2021	MILAD	174	–	175

For each representation (Repr.), year of publication (Year), original reference (Orig.), references for further methodological development (Dev.), and availability of implementations (Avail.) are shown. See Glossary (Table 1) for abbreviations.

(iv) *Computational efficiency* relative to the reference simulations. For an advantage over simulations alone (without ML), overall computational costs should be reduced by one or more orders of magnitude to justify the effort. The difference between running reference simulations and computing representations usually dominates costs. (Supplementary Note 9) Therefore, the results of computationally sufficiently cheaper simulations, for example, from a lower level of theory, can be used to construct representations[62,63] or to predict properties at a higher level of theory ("Δ-learning")[63–65].

(v) *Structure* of representations and the resulting data distribution should be suitable for regression. (Supplementary Notes 6 and 10) It is useful if feature vectors always have the same length[66,67]. Representations often have a Hilbert space structure, featuring an inner product, completeness, projections, and other advantages. Besides the formal space defined by the representation, the structure of the subspace spanned by the data is critical[57,68]. This requirement is currently less well understood than (i)–(iv) and evaluated mostly empirically (see section "Empirical comparison").

(vi) *Generality*, in the sense of being able to encode any atomistic system. While current representations handle finite and periodic systems, less work was done on charged systems, excited states, continuous spin systems, isotopes, and systems subjected to external fields.

Simplicity, both conceptually and in terms of implementation, is, in our opinion, a desirable quality of representations, albeit hard to quantify.

The above requirements preclude direct use of Cartesian coordinates, which violate requirement (i), and internal coordinates, which satisfy (i.b)–(i.d) but are still system-specific, violating (v) and possibly (i.a) if not defined uniquely. Descriptors and fingerprints from cheminformatics[18] and materials informatics violate (ii) and (iii.a).

Simple representations such as the Coulomb matrix (section "Other representations") either suffer from coarse-graining, violating (ii), or from discontinuities, violating (iii.a). In practice, representations do not satisfy all requirements exactly (section "Analysis") but can achieve high predictive accuracy regardless; for example, for some datasets, modeling a fraction of higher-order terms can be sufficiently unique already[69]. The optimal interaction orders to utilize in a representation also depend on the type and amount of data available[42].

A UNIFIED FRAMEWORK

Based on recent work[6,70–72] we describe concepts and notation towards a unified treatment of representations in order to highlight their common foundation. For this, we successively build up Hilbert spaces of atoms, k-atom tuples, local environments, and global structures, using group averaging to ensure physical invariants and tensor products to retain desired information and construct invariant features.

Representing atoms, environments, and systems

Information about a single atom, such as position and proton number, is represented as an abstract ket $|\alpha\rangle$ in a Hilbert space $\mathcal{H}_\alpha$. Relations between k atoms, where their order can matter, are encoded as k-body functions $g_k : \mathcal{H}_\alpha^{\times k} \to \mathcal{H}_g$. (Supplementary Note 11) These functions can be purely geometric, such as distances or angles, but could also be of (al)chemical or mixed nature. Tuples of atoms and associated many-body properties are thus elementary tensors of a space $\mathcal{H} \equiv \mathcal{H}_\alpha^{\otimes k} \otimes \mathcal{H}_g$,

$$|\mathcal{A}_{\alpha_1 \ldots \alpha_k}\rangle \equiv |\alpha_1\rangle \otimes \ldots \otimes |\alpha_k\rangle \otimes g_k(|\alpha_1\rangle, \ldots, |\alpha_k\rangle). \tag{1}$$

A local environment of an atom $|\alpha\rangle$ is represented via the relations to its $k-1$ neighbors by keeping $|\alpha\rangle$ fixed:

$$|\mathcal{A}_\alpha\rangle \equiv \sum_{\alpha_1, \ldots, \alpha_{k-1}} |\mathcal{A}_{\alpha, \alpha_1, \ldots, \alpha_{k-1}}\rangle. \tag{2}$$

Weighting functions can reduce the influence of atoms far from $|\alpha\rangle$; we include these in g_k. An atomistic system as a whole is represented by summing over the local environments of all its atoms:

$$|\mathcal{A}\rangle = \sum_{\alpha_i} |\mathcal{A}_{\alpha_i}\rangle = \sum_{\alpha_1, \ldots, \alpha_k} |\mathcal{A}_{\alpha_1, \ldots, \alpha_k}\rangle. \tag{3}$$

For periodic systems, this sum diverges, which requires either exploiting periodicity, for example, by working in reciprocal space, or employing strong weighting functions and keeping one index constrained to the unit cell[73].

Symmetries, tensor products, and projections

Representations incorporate symmetry constraints (section "Requirements") by using invariant many-body functions g_k, such as distances or angles, or through explicit symmetrization via group averaging[70]. Explicit symmetrization transforms a tensor $|T\rangle$ by integrating over a symmetry group $\mathcal{S}$ with right-invariant Haar measure dS,

$$|T\rangle_{\mathcal{S}} \equiv \int_{\mathcal{S}} S|T\rangle \mathrm{d}S, \tag{4}$$

where symmetry transformations $S \in \mathcal{S}$ act separately on each subspace of $\mathcal{H}$ or parts thereof. For example, for rotational invariance, only the atomic positions in $\mathcal{H}_\alpha$ change. Rotationally invariant features can be derived from tensor contractions[74], as any full contraction of contravariant with covariant tensors yields rotationally invariant scalars[75].

Sometimes group averaging can integrate out desired information encoded in $|T\rangle$. To counter this, one can perform tensor products of $|T\rangle$ with itself, effectively replacing $\mathcal{H}$ by $\mathcal{H}^{\otimes\nu}$. Together, this results in a generalized transform

$$|T^\nu\rangle_{\mathcal{S}} \equiv \int_{\mathcal{S}} (S|T\rangle)^{\otimes\nu} \mathrm{d}S. \tag{5}$$

To retain only part of the information in $\mathcal{A}$, one can project onto orthogonal elements $\{|h_l\rangle\}_{l=1}^m$ of $\mathcal{H}$ via an associated projection operator $\mathcal{P} = \sum_l |h_l\rangle\langle h_l|$. Inner products and induced distances between representations are then given by

$$\langle \mathcal{A}|\mathcal{P}|\mathcal{A}'\rangle \text{ and } d_{\mathcal{P}}(|\mathcal{A}\rangle, |\mathcal{A}'\rangle) = ||\mathcal{P}|\mathcal{A}\rangle - \mathcal{P}|\mathcal{A}'\rangle||_{\mathcal{H}}. \tag{6}$$

SELECTED REPRESENTATIONS

We discuss three representations that fulfill the requirements in section "Requirements" and for which an implementation not tied to a specific regression algorithm and supporting finite and periodic systems was openly available. These representations are empirically compared in section "Empirical comparison".

Symmetry functions

Symmetry functions[8,66] (SFs) describe k-body relations between a central atom and the atoms in a local environment around it. (Supplementary Notes 11 and 12) They are typically based on distances (*radial* SFs, $k = 2$) and angles (*angular* SFs, $k = 3$). Each SF encodes a local feature of an atomic environment, for example, the number of H atoms at a given distance from a central C atom.

For each SF and k-tuple of chemical elements, contributions are summed. Sufficient resolution is achieved by varying the HPs of an SF. For continuity (and differentiability), a cut-off function ensures that SFs decay to zero at the cut-off radius. Two examples of SFs from ref. [66] (see Table 4 and Supplementary Note 22 for further references and SFs) are

$$\begin{aligned} G_i^2 &= \sum_j \exp\left(-\eta(d_{ij} - \mu)^2\right) f_c(d_{ij}) \\ G_i^4 &= 2^{1-\zeta} \sum_{j,k\neq i} (1 + \lambda \cos\theta_{ijk})^\zeta \cdot \\ &\quad \exp\left(-\eta(d_{ij}^2 + d_{ik}^2 + d_{jk}^2)\right) f_c(d_{ij})\, f_c(d_{ik})\, f_c(d_{jk}) \end{aligned} \tag{7}$$

where $\eta, \mu, \zeta, \lambda$ are numerical HPs controlling radial broadening, shift, angular resolution, and angular direction, respectively, d_{ij} is a distance, θ_{ijk} is the angle between atoms i, j, k, and f_c is a cut-off function. Figure 2 illustrates the radial SFs in Eq. (7). The choice of which SFs to use is a structural HP. Variants of SFs include partial radial distribution functions[76], SFs with improved angular resolution[77] and reparametrizations for improved scaling with the number of chemical species[78–80].

In terms of the unified notation, SFs use invariant functions g_k based on distances and angles, multiplied by a cut-off function, to describe local environments $|\mathcal{A}_\alpha\rangle$. Projections $\mathcal{P}$ onto tuples of atomic numbers Z then separate contributions from different combinations of chemical elements. For instance, for G_i^2 in Eq. (7), the representation of atom i is

$$|\mathcal{A}_i, (\mu, \eta)\rangle = \sum_j (|Z_i\rangle \otimes |Z_j\rangle) G^2(d_{ij}, \mu, \eta). \tag{8}$$

with $G^2(d_{ij}, \mu, \eta) = \exp\left(-\eta(d_{ij} - \mu)^2\right) f_c(d_{ij})$.

Many-body tensor representation

The global many-body tensor representation[73] (MBTR) consists of broadened distributions of k-body terms, arranged by element combination. For each k-body function and k-tuple of elements, all corresponding terms (for example, all distances between C and H atoms) are broadened and summed up (Fig. 3). The resulting

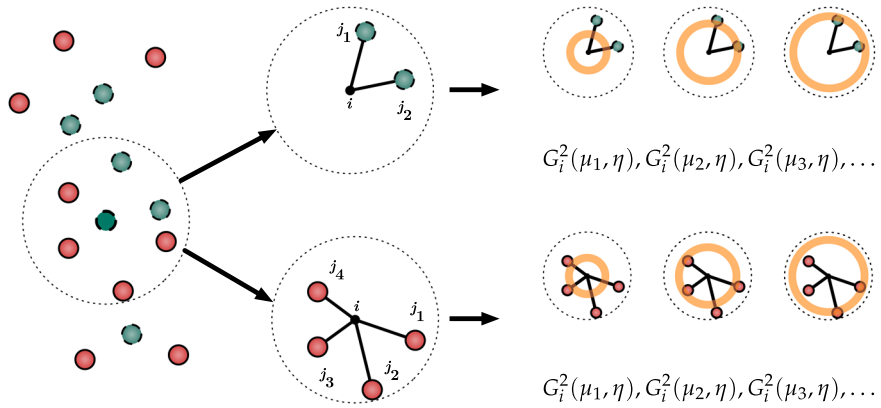


Fig. 2 Symmetry functions. Shown are radial functions $G_i^2(\mu,\eta)$ (Eq. (7)) for increasing values of μ. The local environment of a central atom is described by summing contributions from neighboring atoms separately by element.

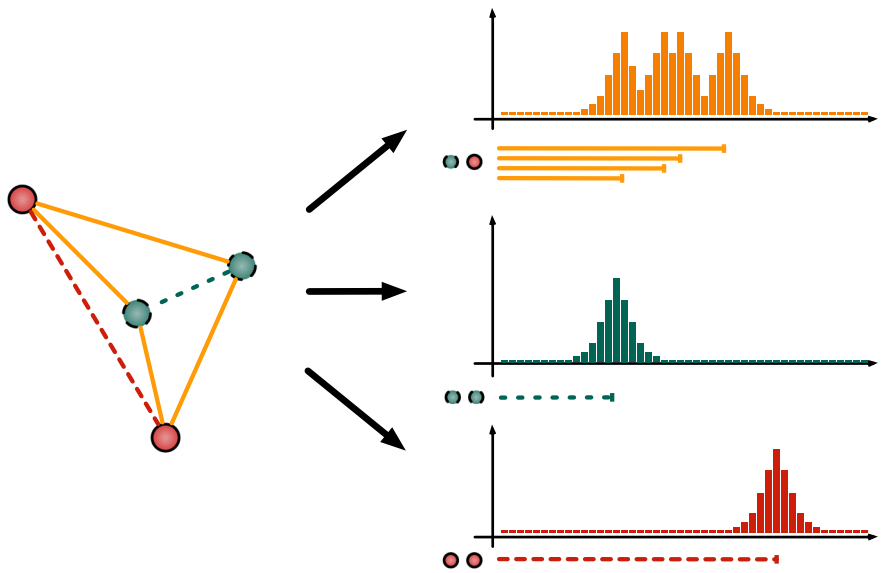


Fig. 3 Many-body tensor representation. Shown are broadened distances (no weighting) arranged by element combination.

distributions describe the geometric features of an atomistic system:

$$f_k(x,z_1,\ldots,z_k) = \sum_{i_1,\ldots,i_k} w_k\, \mathcal{N}(x|g_k,\sigma) \prod_{j=1}^{k} \delta_{z_j,Z_{i_j}}\,, \tag{9}$$

where w_k is a weighting function that reduces the influence of tuples with atoms far from each other, and g_k is a k-body function; both w_k and g_k depend on atoms $i_1,\ldots,i_k$. $\mathcal{N}(x|\mu,\sigma)$ denotes a normal distribution with mean μ and variance σ^2, evaluated at x. The product of Kronecker δ-functions restricts to the given element combination $z_1,\ldots,z_k$.

Periodic systems can be treated by using strong weighting functions and constraining one index to the unit cell. In practice, Eq. (9) can be discretized. Structural HPs include the choice of w_k and g_k; numerical HPs include variance σ of normal distributions. Requiring one atom in each tuple to be the central atom results in a local variant[81].

In terms of the unified notation, MBTR uses distribution-valued functions g_k, including weighting, with distributions centered on k-body terms such as (inverse) distances or angles. The outer-product structure of $|\mathcal{A}\rangle$ corresponds to the product of δ-functions in Eq. (9), which selects for specific k-tuples of chemical elements. For $k=2$, for example, the geometry and weighting functions depend on pairwise distances d_{ij}:

$$\begin{aligned} |\mathcal{A},x\rangle &= \sum_i |\mathcal{A}_i,x\rangle \\ |\mathcal{A}_i,x\rangle &\propto \sum_j \left(|Z_i\rangle \otimes |Z_j\rangle\right) G_i^2\left(g_2(d_{ij}),x,\frac{1}{2}\sigma^{-2}\right). \end{aligned} \tag{10}$$

Smooth overlap of atomic positions

Smooth overlap of atomic positions[44] (SOAP) representations expand a central atoms' local neighborhood density, a scalar function of position $\mathbf{r}$, approximated by Gaussian functions located at atom positions, in orthogonal radial and spherical harmonics basis functions (Fig. 4):

$$\rho(\mathbf{r}) = \sum_{n,l,m} c_{nlm}\, g_n(\mathbf{r})\, Y_{lm}(\mathbf{r}), \tag{11}$$

where c_{nlm} are expansion coefficients, g_n are radial, and Y_{lm} are (angular) spherical harmonics basis functions. From the coefficients, rotationally invariant quantities can be constructed, such as the power spectrum

$$p_{nn'\ell} = \sum_m c_{n\ell m} c^*_{n'\ell m} \tag{12}$$

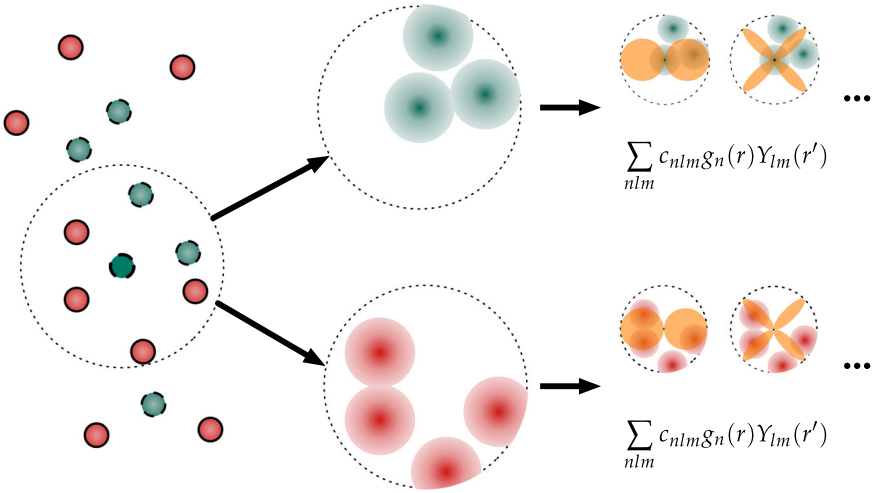


Fig. 4 Smooth overlap of atomic positions. The local density around a central atom is modeled by atom-centered normal distributions and expanded into radial and spherical harmonics basis functions.

which is equivalent to a radial and angular distribution function[15], and therefore captures up to three-body interactions. Numerical HPs are the maximal number of radial and angular basis functions, the broadening width, and the cut-off radius.

An alternative to the power spectrum is the *bispectrum*[9] (BS), a set of invariants that couples multiple angular momentum and radial channels. The Spectral Neighbor Analysis Potential (SNAP) includes quadratic terms in the BS components[82]. Extensions of the SOAP framework include recursion relations for faster evaluation[83] and alternative radial basis functions g_n, such as third- and higher-order polynomials[83], Gaussian functions[84], and spherical Bessel functions of the first kind [58,85].

In terms of the unified notation, SOAP uses vector-valued g_k to compute the basis set coefficients in Eq. (11). Analytic group-averaging (symmetry integration) then results in invariant features such as the power spectrum ($\nu = 2$, Eq. (5)) or bispectrum ($\nu = 3$). The SOAP ($\nu = 2$) representation is therefore

$$|\mathcal{A}_i, nn'l\rangle = \sum_j \left(|Z_i\rangle \otimes |Z_j\rangle\right) p_{nn'l}\,. \tag{13}$$

OTHER REPRESENTATIONS

Many other representations were proposed.

The Coulomb matrix[4] (CM) globally describes a system via inverse distances between atoms but does not contain higher-order terms. It is fast to compute, easy to implement, and in the commonly used sorted version (see footnote reference 25 in ref. [4]) allows reconstruction of an atomistic system via a least-squares problem. However, its direct use of atomic numbers to encode elements is problematic, and it suffers either from discontinuities in the sorted version or from information loss in the diagonalized version as its eigenspectrum is not unique[55,86]. A local variant exists[87].

The bag-of-bonds[88] (BoB) representation uses the same inverse-distance terms as the CM but arranges them by element pair instead of by atom pair. The "BA-representation"[89] extends this to higher-order interactions by using bags of dressed atoms, distances, angles, and torsions. The inverse-distance many-body representation[90] (IDMBR) employs higher powers of inverse distances and separation by element combinations.

Histograms of distances, angles, and dihedral angles[91] (HDAD) are histograms of geometric features organized by element combination. This global representation is similar to MBTR but typically uses fewer bins, without broadening or explicit weighting.

The Faber-Christensen-Huang-von Lilienfeld representation[92,93] (FCHL) describes atomic environments with normal distributions over row and column in the periodic table ($k = 1$), interatomic distances ($k = 2$), and angles ($k = 3$), scaled by power laws. In the FCHL18 variant[92], the full continuous distributions are used, requiring an integral kernel for regression. Among other optimizations, FCHL19[93] discretizes these distributions, similar to the approach taken by SFs, and can be used with standard vector kernels.

Wavelet scattering transforms[94–100] (WST) use a convolutional wavelet frame representation to describe variations of (local) atomic density at different scales and orientations. Integrating non-linear functions of the wavelet coefficients yields invariant features, where second- and higher-order features couple two or more length scales. Variations use different wavelets (Morlet[94,95], solid harmonic, or atomic orbital[96–98,100]) and radial basis functions (exponential[96], Laguerre polynomials[97,100]).

Moment-tensor potentials[74] (MTP) describe local atomic environments using a spanning set of efficiently computable, rotationally and permutationally invariant polynomials derived from tensor contractions. Related representations include Gaussian moments[75] (GM), based on contractions of tensors from (linear combinations of) Gaussian-type atomic orbitals; the *N*-body iterative contraction of equivariants (NICE) framework[71], which uses recursion relations to compute higher-order terms efficiently; atomic cluster expansion[53,101,102] (ACE), which employs a basis of isometry- and permutation-invariant polynomials from trigonometric functions and spherical harmonics; and, moment invariants as (local) atomic descriptors (MILAD), which are non-redundant invariants constructed from Zernike polynomials.

Overlap-matrix fingerprints[62,103,104] (OMF) and related approaches[30,35] employ the sorted eigenvalues (and derived quantities) of overlap matrices based on Gaussian-type orbitals as representation. Eigenvalue crossings can cause derivative discontinuities, requiring post-processing[104] to ensure continuity. Using a molecular orbital basis (MOB[63,105] and related approaches[36]) adds the cost of computing the basis, for example, localized molecular orbitals via a Hartree–Fock self-consistent field calculation. Other matrices can be used, such as Fock, Coulomb, and exchange matrices, or even the Hessian, for example, from a computationally cheaper reference method. Density-encoded canonically-aligned fingerprints[106] (DECAF) represent the local density in a canonical, invariant coordinate frame found by solving an optimization problem related to kernel principal component analysis.

Tensor properties require covariance (equivariance). Proposed solutions include local coordinates from eigendecompositions[45], which exhibit discontinuities when eigenvalues cross, related local coordinate systems[106], and internal vectors[107] (IV), based on inner products of summed neighbor vectors at different scales, as well as covariant extensions of SOAP[6,52] and ACE[53].

ANALYSIS

We discuss relationships between specific representations, to which degree they satisfy the requirements in section "Requirements", trade-offs between local and global representations, and relationships to other models and modeling techniques, including systematic selection and generation of features.

Relationships between representations

Most representations in sections "Selected representations" and "Other representations" are related through the concepts in section "A unified framework". We distinguish two primary strategies to deal with invariances, the use of invariant k-body functions (BoB, CM, FCHL, HDAD, IDMBR, MBTR, SF) and explicit symmetrization (ACE, BS, GM, MILAD, MOB, MTP, NICE, OMF, SOAP, WST). A similar distinction can be made for kernels[40]. Some representations share specific connections:

Comparing Eqs. (8) and (10) reveals that for suitable choices of hyperparameters, SFs can be identified with the local terms of distance-based MBTR, as both can be seen as histograms of geometric features, similar to HDAD. This suggests a local MBTR or HDAD variant by restricting summation to atomic environments[81], and a global variant of SFs by summing over the whole system.

ACE, BS, GM, MILAD, MTP, NICE, and SOAP share the idea of generating tensors that are then systematically contracted to obtain rotationally invariant features. These tensors should form an orthonormal basis, or at least a spanning set, for atomic environments. Formally, expressing a local neighborhood density in a suitable basis before generating derived features avoids asymptotic scaling with the number of neighboring atoms[101], although HPs, and thus runtime, still depend on it. Within a representation, recursive relationships can exist between many-body terms of different orders[71,83,102]. References [53,101,102] discuss technical details of the relationships between ACE and SFs, BS, SNAP, SOAP, MTP.

Requirements

Some representations, in particular early ones such as the CM, do not fulfill all requirements in section "Requirements". Most representations fulfill some requirements only in the limit, that is, absent practical constraints such as truncation of infinite sums, short cut-off radii, and restriction to low-order interaction terms. The degree of fulfillment often depends on HPs, such as truncation order, the length of a cut-off radius, or the highest interaction order k used. Effects can be antagonistic; for example, in Eq. (11), both *(ii) uniqueness* and *(iv) computational effort* increase with n, l, m[44]. In addition, not all invariances of a property might be known or require additional effort to model, for example, symmetries[51].

Mathematical proof or systematic empirical verification that a representation satisfies a requirement or related property are sometimes provided: The symmetrized invariant moment polynomials of MTPs form a spanning set for all permutationally and rotationally invariant polynomials[74]; basis sets can also be constructed[102]. For SOAP, systematic reconstruction experiments demonstrate the dependence of uniqueness on parametrization[44].

While *(ii) uniqueness* guarantees that reconstruction of a system up to invariances is possible in principle, accuracy and complexity of this task vary with representation and parametrization. For example, reconstruction is a simple least-squares problem for the global CM as it comprises the whole distance matrix $\boldsymbol{D}_{ij} = ||\boldsymbol{r}_i - \boldsymbol{r}_j||_2$, whereas for local representations, (global) reconstruction is more involved.

If a local representation comprises only up to 4-body terms then there are degenerate environments that it cannot distinguish[57], but that can differ in property. Combining representations of different environments in a system can break the degeneracy. However, by distorting feature space *(v) structure*, these degeneracies degrade learning efficiency and limit achievable prediction errors, even if the training set contains no degenerate systems[57]. It is currently unknown whether degenerate environments exist for representations with terms of order $k > 4$. The degree to which a representation is unique can be numerically investigated through the eigendecomposition of a sensitivity matrix based on a representation's derivatives with respect to atom coordinates[104].

Global versus local representations

Local representations can be used to model global properties by assuming that these decompose into atomic contributions. In terms of prediction errors, this tends to work well for energies. (Supplementary Note 7) Learning with atomic contributions adds technical complexity to the regression model and is equivalent to pairwise-sum kernels on whole systems, (Supplementary Note 8) with favorable computational scaling for large systems (see Supplementary Notes 9 and 27, and Table 5). Other approaches to creating global kernels from local ones exist[108].

Conversely, using global representations for local properties can require modifying the representation to incorporate locality and directionality of the property[45,84]. A general recipe for constructing local representations from global ones is to require interactions to include the central atom, starting from $k = 2$ [81].

Relationships to other models and techniques

Two modeling aspects directly related to representations are which subset of the features to use and the construction of derived features. Both modulate feature space dimensionality and *(v) structure*. Adding products of 2-body and 3-body terms as features, for example, can improve performance[69], as these features relate to higher-order terms, (Supplementary Note 11) but can also degrade performance if the features are unrelated to the predicted property, or if there is insufficient data to infer the relationship. Feature selection tailors a representation to a dataset by selecting a small subset of features that still predict the target property accurately enough. Optimal choices of features depend on the data's size and distribution.

In this work, we focus exclusively on representations. In kernel regression, however, kernels can be defined directly between two systems, without an explicit intermediate representation. For example, n-body kernels between atomic environments can be

Table 5. Computational cost of calculating representations.

Time in ms	Dataset		
Representation	qm9	ba10	nmd18
MBTR $k = 2$	0.76 ± 0.32	13 ± 5.1	340 ± 99
SF $k = 2$	1.4 ± 0.18	3.3 ± 1.4	8.2 ± 1.1
MBTR $k = 2, 3$	12 ± 6.9	290 ± 140	28k ± 4.4k
SF $k = 2, 3$	2.8 ± 0.85	27 ± 12	98 ± 89
SOAP	1.9 ± 0.54	9.1 ± 4.8	19 ± 8.6

Costs given in milliseconds of processor time. Shown are mean ± standard deviation over all training set sizes of a dataset for the time to compute the representation of a single molecule or unit cell. See Supplementary Note 27 for details.

systematically constructed from a non-invariant Gaussian kernel using Haar integration, or using invariant k-body functions (Supplementary Note 11), yielding kernels of varying body-order and degrees of freedom[40,42]. Similarly, while neural networks can use representations as inputs, their architecture can also be designed to learn implicit representations from the raw data (end-to-end learning). In all cases, the requirements in section "Requirements" apply.

EMPIRICAL COMPARISON

We benchmark prediction errors for all representations from section "Selected representations" on three benchmark datasets. Since our focus is exclusively on the representations, we control for other factors, in particular for data distribution, regression method, and HP optimization.

Datasets

The `qm9` consensus benchmarking dataset[109,110] comprises 133,885 organic molecules composed of H, C, N, O, F with up to 9 non-H atoms. (Supplementary Note 13) Ground state geometries and properties are given at the DFT/B3LYP/6-31G(2df,p) level of theory. We predict U_0, the atomization energy at 0 K.

The `ba10` dataset[110,111] (Supplementary Note 14) contains the ten binary alloys AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, and NbNi. For each alloy system, it comprises all structures with up to 8 atoms for face-centered cubic (FCC), body-centered cubic (BCC), and hexagonal close-packed (HCP) crystal types, 15,950 structures in total. Formation energies of unrelaxed structures are given at the DFT/PBE level of theory.

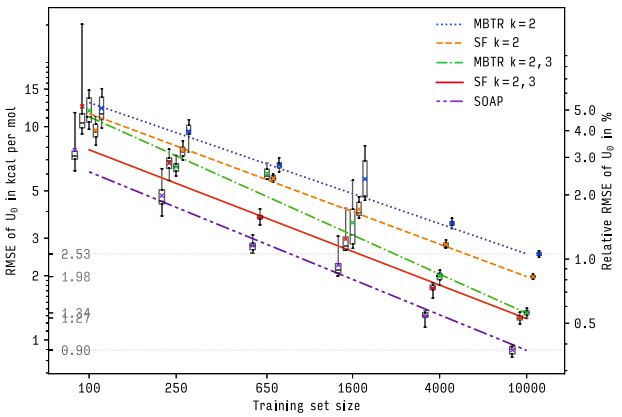
The `nmd18` challenge[112] dataset[113] (Supplementary Note 15) contains 3000 ternary $(Al_x\text{-}Ga_y\text{-}In_z)_2O_3$ oxides, $x+y+z=1$, of potential interest as transparent conducting oxides. Formation and band-gap energies of relaxed structures are provided at the DFT/PBE level of theory. The dataset contains both relaxed (`nmd18r`, used here) and approximate (`nmd18u`) structures as input. In the challenge, energies of relaxed structures were predicted from approximate structures.

Together, these datasets cover finite and periodic systems, organic and inorganic chemistry, and ground state as well as off-equilibrium structures. See Supplementary Notes 13–15 for details.

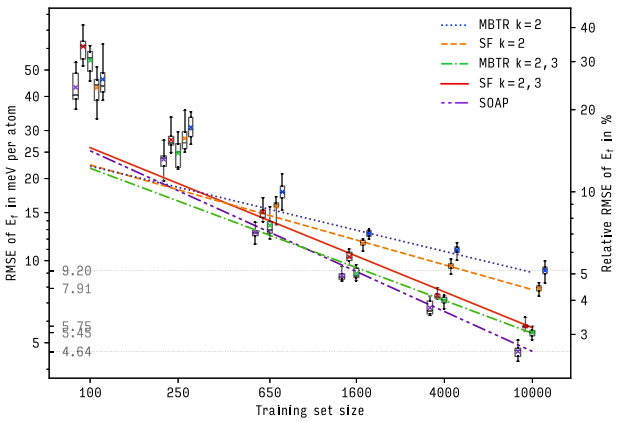
Benchmarking method

We estimate prediction errors as a function of training set size (learning curves, Supplementary Notes 16 and 17). To ensure that subsets are representative, we control for the distribution of elemental composition, size, and energy. (Supplementary Note 18) This reduces the variance of performance estimates and ensures the validity of the independent-and-identically-distributed data assumption inherent in ML. All predictions are on data never seen during training.

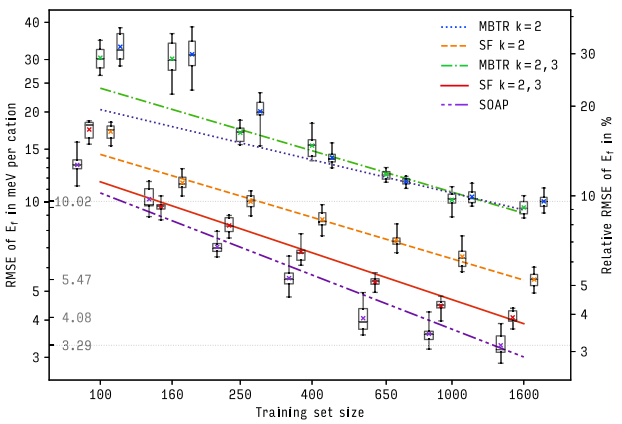
We use kernel ridge regression[114] (KRR; predictions are equivalent to those of Gaussian process regression[115], GPR) with a Gaussian kernel as an ML model. (Supplementary Note 19) KRR is a widely-used non-parametric non-linear regression method. There are two regression HPs, the length scale of the Gaussian kernel and the amount of regularization. (Supplementary Note 21) In this work, training is exclusively on energies; in particular, derivatives are not used. All HPs, that is, regression HPs, numerical HPs (e.g., a weight in a weighting function), and structural HPs (e.g., which weighting function to use), are optimized with a consistent and fully automatic scheme based on sequential model-based optimization and tree-structured Parzen estimators[116,117]. (Supplementary Note 20) This setup treats all representations on equal footing. See Supplementary Notes 21–24 for details on the optimized HPs.



(a) Dataset `qm9`.



(b) Dataset `ba10`.



(c) Dataset `nmd18r`.

Fig. 5 Learning curves for selected representations on datasets. Datasets **a** qm9, **b** ba10, and **c** nmd18r. Shown is root mean squared error (RMSE) of energy predictions on out-of-sample-data as a function of training set size. Boxes, whiskers, bars, crosses show interquartile range, total range, median, mean, respectively. Lines are fits to theoretical asymptotic RMSE. (Supplementary Note16). See Glossary (Table 1) for abbreviations.

Learning curves and compute times

Figure 5 presents learning curves for SF, MBTR, SOAP on datasets `qm9`, `ba10`, `nmd18r` (see Supplementary Note 25 for tabulated values). For each dataset, representation, and training set size, we trained a KRR model and evaluated its predictions on a separate

hold-out validation set of size 10k (qm9), 1k (ba10), and 0.6k (nmd18r). This procedure was repeated 10 times to estimate the variance of these experiments.

Boxes, whiskers, horizontal bars, and crosses show interquartile ranges, minimum/maximum value, median, and mean, respectively, of the root mean squared error (RMSE) of hold-out-set predictions over repetitions. We show RMSE as it is the loss minimized by least-squares regression such as KRR, and thus a natural choice. For other loss functions, see Supplementary Note 26. From statistical learning theory, RMSE decays as a negative power of training set size (a reason why learning curves are preferably shown on log-log plots)[118–120]. Lines show corresponding fits of mean RMSE, weighted by the standard deviation for each training set size.

Figure 6 reveals dependencies between the time to compute representations for a training set (horizontal axis) and RMSE (vertical axis). When comparing observations in two dimensions, here time t and error e, there is no unique ordering $<$, and we resort to the usual notion of dominance: Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$; then $\boldsymbol{x}$ dominates $\boldsymbol{x}'$ if $x_i \leq x_i'$ for all dimensions i and $x_i < x_i'$ for some i. The set of all non-dominated points is called the Pareto frontier, shown by a line, with numbers indicating training set sizes. Table 5 presents compute times for representations (see Supplementary Note 27 for kernel matrices).

Findings

Asymptotically, observed prediction errors for all representations on all datasets relate as

$$\begin{aligned} &\text{SF-2,3} \prec \text{SF-2}\,, && \text{MBTR-2,3} \preccurlyeq \text{MBTR-2}\,, \\ &\text{SOAP} \prec \text{SF-2,3}\,, && \text{SOAP} \prec \text{MBTR-2,3}\,, \\ &\text{SF-2,3} \preccurlyeq \text{MBTR-2,3}\,, && \text{SF-2} \prec \text{MBTR-2}\,, \end{aligned} \tag{14}$$

where $A \prec B$ ($A \preccurlyeq B$) indicates that A has lower (or equal) estimated error than B asymptotically. Except for MBTR-2,3 $\not\preccurlyeq$ SF-2 on dataset nmd18r,

$$\text{SOAP} \prec \text{SF-2,3} \preccurlyeq \text{MBTR-2,3} \prec \text{SF-2} \prec \text{MBTR-2}\,. \tag{15}$$

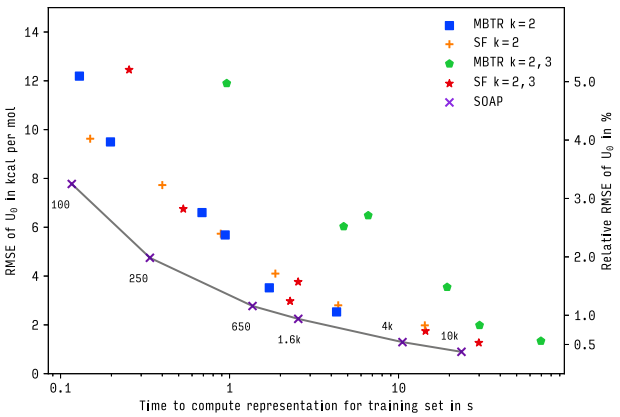
We conclude that, for energy predictions, accuracy improves with modeled interaction order and for local representations over global ones. The magnitude of, and between, these effects varies across datasets.

Dependence of predictive accuracy on interaction order has been observed by others[82,84,90,92,121] and might be partially due to a higher resolution of structural features[57]. The latter would only show for sufficient training data, such as for dataset ba10 in Fig. 5. We do not observe this for dataset qm9, possibly because angular terms might be immediately relevant for characterizing organic molecules' carbon scaffolds[90].

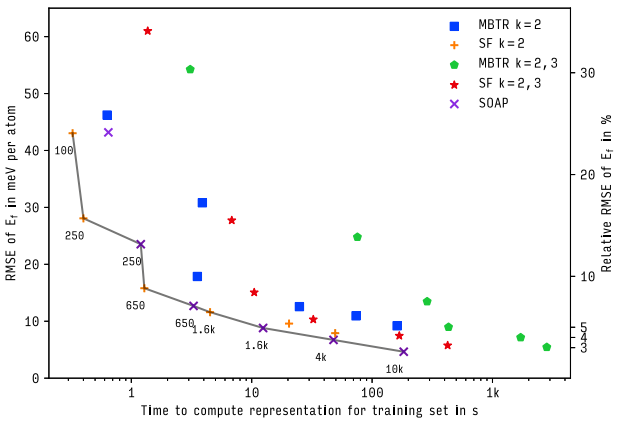
Better performance of local representations might be due to higher resolution and better generalization (both from representing only a small part of the whole structure), and has also been observed by others[122,123]. The impact of assuming additivity is unclear but likely depends on the structure of the modeled property. (Supplementary Note 7) Our comparison includes only a single global representation (MBTR), warranting further study of the locality aspect. For additional analysis details, see Supplementary Notes 28 and 29.

Computational costs tend to increase with predictive accuracy. Representations should therefore be selected based on a target accuracy, constrained by available computing resources.

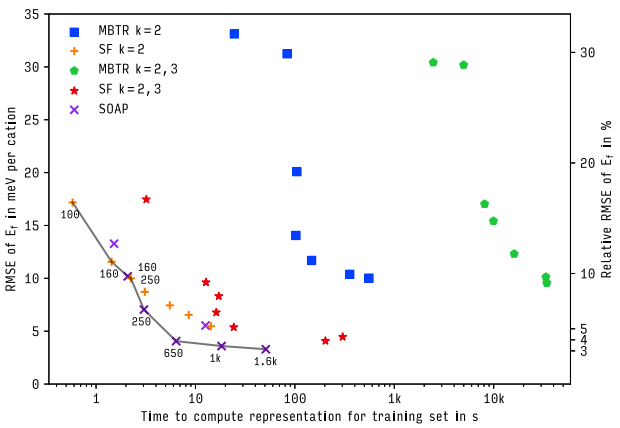
Converged prediction errors are in reasonable agreement with the literature (Supplementary Note 30) considering the lack of standardized conditions such as sampling, regression method, HP optimization, and reported performance statistics. In absolute terms, prediction errors of models trained on 10k samples are closer to the differences between DFT codes than the (systematic) differences between the underlying DFT reference and experimental measurements. (Supplementary Note 31).



(a) Dataset qm9.



(b) Dataset ba10.



(c) Dataset nmd18r.

Fig. 6 Compute times of selected representations for datasets. Datasets **a** qm9, **b** ba10, and **c** nmd18r. Shown is root mean squared error (RMSE) of energy predictions on out-of-sample-data as a function of the time needed to compute all representations in a training set. Lines indicate Pareto frontiers; inset numbers show training set sizes. See Glossary (Table 1) for abbreviations.

CONCLUSIONS AND OUTLOOK

We review representations of atomistic systems, such as molecules and crystalline materials, for machine learning of ab initio quantum-mechanical simulations. For this, we distinguish between local and global representations and between using invariant k-body functions and explicit symmetrization to deal with invariances. Despite their apparent diversity, many representations

can be formulated in a single mathematical framework based on k-atom terms, symmetrization, and tensor products. Empirically, we observe that when controlling for other factors, including distribution of training and validation data, regression method, and HP optimization, both prediction errors and compute time of SFs, MBTR and SOAP improve with interaction order, and for local representations over global ones.

Our findings suggest the following guidance:

- If their prediction errors are sufficient for an application, we recommend two-body versions of simple representations such as SF and MBTR as they are fastest to compute.
- For large systems, local representations should be used.
- For strong noise or bias on input structures, as in dataset `nmd18u`, performance differences between representations vanish, (Supplementary Note 29) and computationally cheaper features that do not satisfy the requirements in section "Requirements" (descriptors) suffice.

We conclude by providing related current research directions, grouped by topic.

Directly related to representations:

- *Systematic development of representations* via extending the mathematical framework (section "A unified framework") to include more state-of-the-art representations. This would enable deriving "missing" variants of representations (see Table 3), such as a global SOAP[108] and local MBTR[81], on a principled basis, as well as understanding and reformulating existing representations in a joint framework, perhaps to the extent of an efficient general implementation[124].
- *Representing more systems*. Develop or extend representations for atomistic systems currently not representable, or only to a limited extent, such as charged atoms and systems[28,53,79,125–129], excited states[130–134], spin systems, isotopes, and systems in an applied external field[135,136].
- *Alchemical learning*. Further understand and develop alchemical representations[92,137,138] that incorporate similarity between chemical species to improve sample efficiency. What are the salient features of chemical elements that need to be considered, also with respect to charges, excitations, spins, and isotopes?
- *Analysis of representations* to better understand structure and data distribution in feature spaces and how they relate to physics and chemistry concepts. Possible approaches include quantitative measures of structure and distribution of datasets in these spaces, dimensionality reduction methods, analysis of data-driven representations from deep neural networks, and construction, or proof of non-existence, of non-distinguishable environments for representations employing terms of order higher than four.
- *Explicit complexity control*. Different applications require different trade-offs between computational cost and predictive accuracy. This requires determination, and automatic adaptation as an HP, of the capacity (complexity, dimensionality) and computational cost of a representation to a dataset, for example, through selection, combination[139], or systematic construction of features[42,57].

Related to benchmarking of representations:

- *Extended scope*. We empirically compare one global and two local representations on three datasets to predict energies using KRR with a Gaussian kernel. For a more systematic coverage, further representations and datasets, training with forces[60,61], and more properties should be included while maintaining control over regression method, data distribution, and HP optimization. Deep neural networks[23,126,140,141] could be included via representation learning. Comparison with simple baseline models such as k-nearest neighbors[142] would be desirable.
- *Improved optimization of HPs:* The stochastic optimizer used in this work required multiple restarts in practice to avoid sub-optimal results, and reached its limits for large HP search spaces. It would be desirable to reduce the influence and computational cost of HP optimization. Possible means include reducing the number of HPs in representations, employing more systematic and thus more robust optimization methods, and providing reliable heuristics for HP default values.
- *Multi-objective optimization*. We optimize HPs for predictive accuracy on a single property. In practice, though, parametrizations of similar accuracy but lower computational cost would be preferable, and more than one property can be of interest. HPs should, therefore, be optimized for multiple properties and criteria, including computational cost and predictive uncertainties (see below). How to balance these is part of the problem[143].
- *Predictive uncertainties*. While prediction errors are frequently analyzed, and reasonable guidelines exist, this is not the case for predictive uncertainties. These are becoming increasingly important as applications of ML mature, for example, for human assessment and decisions, learning on the fly[144], and active learning. Beyond global analysis of uncertainty estimates, local characterization (in input or feature space) of prediction errors is relevant[143,145].

Related through context:

- *Long-range interactions*. ML models appear to be well-suited for short- and medium-ranged interactions, but problematic for long-ranged interactions due to the increasing degrees of freedom of larger systems and larger necessary cut-off radii of atomic environments. Two approaches are to integrate ML models with physical models for long-range interactions[125,128,146], and to adapt ML models to learn long-range interactions directly[147].
- *Relationships between QM and ML*. A deeper understanding of the relationships between QM and kernel-based ML could lead to insights and technical progress in both fields. As both share concepts from linear algebra, such relationships could be formal mathematical ones. For example, QM concepts such as matrix product states can parameterize non-linear kernel models[148].

DATA AVAILABILITY

The data that support the findings of this study are publicly available. The benchmark datasets `qm9`, `ba10`, `nmd18` were obtained from qmml.org. The repository gitlab.com/repbench/repbench-datasets contains the used data splits, the datasets in `cmlkit` format, as well as the data underlying all plots and tables, including the optimized models and the hyperparameter search spaces.

CODE AVAILABILITY

The code that was used to generate the results in this study is publicly available. It is based on the the `cmlkit` python package, which can be found at github.com/sirmarcel/cmlkit; a tutorial introduction to this package is part of the NOMAD Analytics Toolkit[149]. The repository gitlab.com/repbench/repbench-project contains all additional code specific to this project; an overview is available at marcel.science/repbench.

Received: 1 April 2020; Accepted: 3 February 2022;
Published online: 16 March 2022

REFERENCES

1. Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732 (2009).

2. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452 (2015).
3. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255 (2015).
4. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
5. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828 (2017).
6. Ceriotti, M., Willatt, M. J. & Csányi, G. in *Handbook of Materials Modeling. Methods: Theory and Modeling* (eds. Andreoni, W. & Yip, S.) (Springer, 2018).
7. Huang, B. & von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chem. Rev.* **121**, 10001 (2021).
8. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
9. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
10. Caro, M. A., Deringer, V. L., Koskinen, J., Laurila, T. & Csányi, G. Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon. *Phys. Rev. Lett.* **120**, 166101 (2018).
11. Jinnouchi, R., Lahnsteiner, J., Karsai, F., Kresse, G. & Bokdam, M. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference. *Phys. Rev. Lett.* **122**, 225701 (2019).
12. Kiyohara, S., Oda, H., Tsuda, K. & Mizoguchi, T. Acceleration of stable interface structure searching using a Kriging approach. *Jpn. J. Appl. Phys.* **55**, 045502 (2016).
13. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
14. Sendek, A. D. et al. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem. Mater.* **31**, 342 (2018).
15. Jinnouchi, R., Karsai, F. & Kresse, G. On-the-fly machine learning force field generation: application to melting points. *Phys. Rev. B* **100**, 014105 (2019).
16. Schölkopf, B. & Smola, A. *Learning with Kernels* (MIT Press, 2002). https://mitpress.mit.edu/books/learning-kernels
17. Hofmann, T., Schölkopf, B. & Smola, A. Kernel methods in machine learning. *Ann. Stat.* **36**, 1171 (2008).
18. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors* 2nd edn (Wiley, 2009).
19. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In: *Proc. 34th International Conference on Machine Learning (ICML)* 1263 (2017).
20. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K.-R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
21. Schütt, K. T. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)* (2017).
22. Kondor, R. *n*-body networks: a covariant hierarchical neural network architecture for learning atomic potentials. Preprint at https://arxiv.org/abs/1803.01588 (2018).
23. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. : SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
24. Zhang, L. et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 4436 (2018).
25. Thomas, N. et al. Tensor field networks: rotation- and translation-equivariant neural networks for 3D point clouds. In *NeurIPS Workshop on Machine Learning for Molecules and Materials* (2018).
26. Kondor, R., Li, Z., Trivedi, S. Clebsch-Gordan nets: a fully Fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems 31 (NeurIPS)* 10117 (2018).
27. Weiler, M., Geiger, M., Welling, M., Boomsma, W. & Cohen, T. S. 3D steerable CNNs: learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 10381 (2018).
28. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theor. Comput.* **15**, 3678 (2019).
29. Anderson, B., Hy, T.-S., Kondor, R.: Cormorant: covariant molecular neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS)* 14537 (2019).
30. Zhang, Y., Hu, C. & Jiang, B. Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation. *J. Phys. Chem. Lett.* **10**, 4962 (2019).
31. Mailoa, J. P. et al. A fast neural network approach for direct covariant forces prediction in complex multi-element extended systems. *Nat. Mach. Intell.* **1**, 471 (2019).
32. Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *Proc. 8th International Conference on Learning Representations (ICLR)* (2020).
33. Miller, B. K., Geiger, M., Smidt, T. E. & Noé, F. Relevance of rotationally equivariant convolutions for predicting molecular properties. In *NeurIPS Workshop on Machine Learning for Molecules* (2020).
34. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS)* (2020).
35. Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & Miller III, T. F. OrbNet: deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
36. Chen, Y., Zhang, L., Wang, H. & E, W. Ground state energy functional with Hartree-Fock efficiency and chemical accuracy. *J. Phys. Chem. A* **124**, 7155 (2020).
37. Hermann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **12**, 891 (2020).
38. Ho, T.-S. & Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations. *J. Chem. Phys.* **104**, 2584 (1996).
39. Unke, O. T. & Meuwly, M. A toolkit for the construction of reproducing kernel-based representations of data: application to multi-dimensional potential energy surfaces. *J. Chem. Inf. Model.* **57**, 1923 (2017).
40. Glielmo, A., Zeni, C. & Vita, A. D. Efficient non-parametric *n*-body force fields from machine learning. *Phys. Rev. B* **97**, 184307 (2018).
41. Koner, D. & Meuwly, M. Permutationally invariant, reproducing kernel-based potential energy surfaces for polyatomic molecules: from formaldehyde to acetone. *J. Chem. Theor. Comput.* **16**, 5474 (2020).
42. Glielmo, A., Zeni, C., Fekete, Á., De Vita, A.: Building nonparametric *n*-body force fields using Gaussian process regression. In Machine Learning Meets Quantum Physics, 67 (eds. Schütt, K. T. et al.) (Springer, 2020).
43. Braun, M. L., Buhmann, J. M. & Müller, K.-R. On relevant dimensions in kernel feature spaces. *J. Mach. Learn. Res.* **9**(Aug), 1875 (2008).
44. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
45. Rupp, M., Ramakrishnan, R. & von Lilienfeld, O. A. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* **6**, 3309 (2015).
46. Bowman, J. M., Bittman, J. S. & Harding, L. B. Ab initio calculations of electronic and vibrational energies of HCO and HOC. *J. Chem. Phys.* **85**, 911 (1986).
47. Darsey, J. A., Noid, D. W. & Upadhyaya, B. R. Application of neural network computing to the solution for the ground-state eigenenergy of two-dimensional harmonic oscillators. *Chem. Phys. Lett.* **177**, 189 (1991).
48. Heo, H., Ho, T.-S., Lehmann, K. K. & Rabitz, H. Regularized inversion of diatomic vibration-rotation spectral data: a functional sensitivity analysis approach. *J. Chem. Phys.* **97**, 852 (1992).
49. Hollebeek, T., Ho, T.-S. & Rabitz, H. Constructing multidimensional molecular potential energy surfaces from ab initio data. *Annu. Rev. Phys. Chem.* **50**, 537 (1999).
50. Li, G. et al. Random sampling-high dimensional model representation (RS-HDMR) and orthogonality of its different order component functions. *J. Phys. Chem. A* **110**, 2474 (2006).
51. Glielmo, A., Sollich, P. & De Vita, A. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B* **95**, 214302 (2017).
52. Grisafi, A., Wilkins, D. M., Csányi, G. & Ceriotti, M. Symmetry-adapted machine-learning for tensorial properties of atomistic systems. *Phys. Rev. Lett.* **120**, 036002 (2018).
53. Drautz, R. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B* **102**, 024104 (2020).
54. Hy, T. S., Trivedi, S., Pan, H., Anderson, B. M. & Kondor, R. Covariant compositional networks for learning graphs. In *Proc. International Workshop on Mining and Learning with Graphs (MLG)* (2019).
55. Moussa, J. E. Comment on "Fast and accurate modeling of molecular atomization energies with machine learning". *Phys. Rev. Lett.* **109**, 059801 (2012).
56. von Lilienfeld, O. A., Ramakrishnan, R., Rupp, M. & Knoll, A. Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quant. Chem.* **115**, 1084 (2015).
57. Pozdnyakov, S. N. et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **125**, 166001 (2020).
58. Kocer, E., Mason, J. K. & Erturk, H. Continuous and optimally complete description of chemical environments using spherical Bessel descriptors. *AIP Adv.* **10**, 015021 (2020).

59. Le, H. M., Huynh, S. & Raff, L. M. Molecular dissociation of hydrogen peroxide (HOOH) on a neural network ab initio potential surface with a new configuration sampling method involving gradient fitting. *J. Chem. Phys.* **131**, 014107 (2009).
60. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quant. Chem.* **116**, 1051 (2015).
61. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
62. Sadeghi, A. et al. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **139**, 184118 (2013).
63. Welborn, M., Cheng, L. & Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theor. Comput.* **14**, 4772 (2018).
64. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theor. Comput.* **11**, 2087 (2015).
65. Schmitz, G., Godtliebsen, I. H. & Christiansen, O. Machine learning for potential energy surfaces: an extensive database and assessment of methods. *J. Chem. Phys.* **150**, 244113 (2019).
66. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
67. Collins, C. R., Gordon, G. J., von Lilienfeld, O. A. & Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **148**, 241718 (2018).
68. Goscinski, A., Fraux, G., Imbalzano, G. & Ceriotti, M. The role of feature space in atomistic learning. *Mach. Learn. Sci. Tech.* **2**, 025028 (2021).
69. Jinnouchi, R., Karsai, F., Verdi, C., Asahi, R. & Kresse, G. Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *J. Chem. Phys.* **152**, 234102 (2020).
70. Willatt, M. J., Musil, F. & Ceriotti, M. Atom-density representations for machine learning. *J. Chem. Phys.* **150**, 154110 (2019).
71. Nigam, J., Pozdnyakov, S. & Ceriotti, M. Recursive evaluation and iterative contraction of *n*-body equivariant features. *J. Chem. Phys.* **153**, 121101 (2020).
72. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759 (2021).
73. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. Preprint athttps://arxiv.org/abs/1704.06439 (2017).
74. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153 (2016).
75. Zaverkin, V. & Kästner, J. Gaussian moments as physically inspired molecular descriptors for accurate and scalable machine learning potentials. *J. Chem. Theor. Comput.* **16**, 5410 (2020).
76. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
77. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192 (2017).
78. Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F. & Marquetand, P. wACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.* **148**, 241709 (2018).
79. Rostami, S., Amsler, M. & Ghasemi, S. A. Optimized symmetry functions for machine-learning interatomic potentials of multicomponent systems. *J. Chem. Phys.* **149**, 124106 (2018).
80. Artrith, N., Urban, A. & Ceder, G. Constructing first-principles phase diagrams of amorphous Li_xSi using machine-learning-assisted sampling with an evolutionary algorithm. *J. Chem. Phys.* **148**, 241711 (2018).
81. The DScribe code contains a local MBTR example of this. See https://github.com/SINGROUP/dscribe.
82. Wood, M. A. & Thompson, A. P. Extending the accuracy of the SNAP interatomic potential form. *J. Chem. Phys.* **148**, 241721 (2018).
83. Caro, M. A. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Phys. Rev. B* **100**, 024112 (2019).
84. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
85. Kocer, E., Mason, J. K. & Erturk, H. A novel approach to describe chemical environments in high-dimensional neural network potentials. *J. Chem. Phys.* **150**, 154102 (2019).
86. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Reply to the comment by J.E. Moussa. *Phys. Rev. Lett.* **109**, 059802 (2012).
87. Barker, J., Bulin, J., Hamaekers, J. & Mathias, S. in *Scientific Computing And Algorithms In Industrial Simulations* (eds. Griebel, M. et al.) 25, Springer, 2017).
88. Hansen, K. et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326 (2015).
89. Huang, B. & von Lilienfeld, O. A. Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).
90. Pronobis, W., Tkatchenko, A. & Müller, K.-R. Many-body descriptors for predicting molecular properties with machine learning: analysis of pairwise and three-body interactions in molecules. *J. Chem. Theor. Comput.* **14**, 2991 (2018).
91. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theor. Comput.* **13**, 5255 (2017).
92. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
93. Christensen, A. S., Bratholm, L. A., Faber, F. A. & von Lilienfeld, O. A. FCHL revisited: faster and more accurate quantum machine learning. *J. Chem. Phys.* **152**, 044107 (2020).
94. Hirn, M., Poilvert, N. & Mallat, S. Quantum energy regression using scattering transforms. Preprint athttps://arxiv.org/abs/1502.02077 (2015).
95. Hirn, M., Mallat, S. & Poilvert, N. Wavelet scattering regression of quantum chemical energies. *Multiscale Model. Simul.* **15**, 827 (2017).
96. Eickenberg, M., Exarchakis, G., Hirn, M. & Mallat, S. Solid harmonic wavelet scattering: predicting quantum molecular energy from invariant descriptors of 3D electronic densities. In *Advances in Neural Information Processing Systems 30 (NeurIPS),* 6522 (2017).
97. Brumwell, X., Sinz, P., Kim, K. J., Qi, Y. & Hirn, M. Steerable wavelet scattering for 3D atomic systems with application to Li-Si energy prediction. In *NeurIPS Workshop on Machine Learning for Molecules and Materials* (2018).
98. Eickenberg, M., Exarchakis, G., Hirn, M., Mallat, S. & Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *J. Chem. Phys.* **148**, 241732 (2018).
99. Homer, E. R., Hensley, D. M., Rosenbrock, C. W., Nguyen, A. H. & Hart, G. L. W. Machine-learning informed representations for grain boundary structures. *Front. Mater.* **6**, 168 (2019).
100. Sinz, P. et al. Wavelet scattering networks for atomistic systems with extrapolation of material properties. *J. Chem. Phys.* **153**, 084109 (2020).
101. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 249901 (2019).
102. Dusson, G. et al. Atomic cluster expansion: completeness, efficiency and stability. *J. Comput. Phys.* **454**, 110946 (2022).
103. Zhu, L. et al. A fingerprint based metric for measuring similarities of crystalline structures. *J. Chem. Phys.* **144**, 034203 (2016).
104. Parsaeifard, B. et al. An assessment of the structural resolution of various fingerprints commonly used in machine learning. *Mach. Learn. Sci. Tech.* **2**, 015018 (2020).
105. Cheng, L., Welborn, M., Christensen, A. S. & Miller III, T. F. A universal density matrix functional from molecular orbital-based machine learning: transferability across organic molecules. *J. Chem. Phys.* **150**, 131103 (2019).
106. Tang, Y.-H., Zhang, D. & Karniadakis, G. E. An atomistic fingerprint algorithm for learning ab initio molecular force fields. *J. Chem. Phys.* **148**, 034101 (2018).
107. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
108. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754 (2016).
109. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
110. Available at the QM/ML website (quantum mechanics/machine learning, https://qmml.org, publicly accessible).
111. Nyshadham, C. et al. Machine-learned multi-system surrogate models for materials prediction. *npj Comput. Mater.* **5**, 51 (2019).
112. Nomad2018 Predicting Transparent Conductors. Predict the key properties of novel transparent semiconductors. Available at https://www.kaggle.com/c/nomad2018-predict-transparent-conductors.
113. Sutton, C. et al. Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition. *npj Comput. Mater.* **5**, 111 (2019).
114. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.* **115**, 1058 (2015).
115. Rasmussen, C. & Williams, C. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
116. Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In: *Advances in Neural Information Processing Systems 24 (NeurIPS)*, 2546 (2011).
117. Bergstra, J. S., Yamins, D. & Cox, D. D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. 30th International Conference on Machine Learning (ICML)*, 115 (2013).
118. Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V. & Denker, J. S. Learning curves: asymptotic values and rate of convergence. In *Advances in Neural Information Processing Systems 6 (NeurIPS)* (1993).

119. Müller, K.-R., Finke, M., Murata, N., Schulten, K. & Amari, S. A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Comput.* **8**, 1085 (1996).
120. Huang, B., Symonds, N. O. & von Lilienfeld, O. A. in *Handbook of Materials Modeling. Methods: Theory and Modeling* (eds.W. Andreoni, W. & Yip, S.) (Springer, 2018).
121. Samanta, A. Representing local atomic environment using descriptors based on local correlations. *J. Chem. Phys.* **149**, 244102 (2018).
122. Jäger, M. O. J., Morooka, E. V., Federici-Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **4**, 37 (2018).
123. Honrao, S. J., Xie, S. R. & Hennig, R. G. Augmenting machine learning of energy landscapes with local structural information. *J. Appl. Phys.* **128**, 085101 (2020).
124. Musil, F. et al. Efficient implementation of atom-density representations. *J. Chem. Phys.* **154**, 114109 (2021).
125. Ghasemi, S. A., Hofstetter, A., Saha, S. & Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B* **92**, 045131 (2015).
126. Nebgen, B. et al. Transferable dynamic molecular charge assignment using deep neural networks. *J. Chem. Theor. Comput.* **14**, 4687 (2018).
127. Novikov, I. S. & Shapeev, A. V. Improving accuracy of interatomic potentials: more physics or more data? A case study of silica. *Mater. Today Commun.* **18**, 74 (2018).
128. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).
129. Parsaeifard, B., De, D. S., Finkler, J. A. & Goedecker, S. Fingerprint-based detection of non-local effects in the electronic structure of a simple single component covalent system. *Condens. Matter* **6**, 9 (2021).
130. Westermayr, J. & Marquetand, P. Machine learning and excited-state molecular dynamics. *Mach. Learn. Sci. Tech.* **1**, 043001 (2020).
131. Westermayr, J. & Marquetand, P. Deep learning for UV absorption spectra with SchNarc: first steps toward transferability in chemical compound space. *J. Chem. Phys.* **153**, 154112 (2020).
132. Westermayr, J., Gastegger, M. & Marquetand, P. Combining SchNet and SHARC: the SchNarc machine learning approach for excited-state dynamics. *J. Phys. Chem. Lett.* **11**, 3828 (2020).
133. Behler, J., Delley, B., Lorenz, S., Reuter, K. & Scheffler, M. Dissociation of O_2 at Al (111): the role of spin selection rules. *Phys. Rev. Lett.* **94**, 036104 (2005).
134. Westermayr, J., Faber, F. A., Christensen, A. S., von Lilienfeld, O. A. & Marquetand, P. Neural networks and kernel ridge regression for excited states dynamics of $CH_2NH_2^+$: from single-state to multi-state representations and multi-property machine learning models. *Mach. Learn. Sci. Tech.* **1**, 025009 (2020).
135. Gastegger, M., Schütt, K. T. & Müller, K.-R. Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* **12**, 11473 (2021).
136. Christensen, A. S., Faber, F. A. & von Lilienfeld, O. A. Operators in quantum machine learning: response properties in chemical space. *J. Chem. Phys.* **150**, 064105 (2019).
137. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661 (2018).
138. Herr, J. E., Koh, K., Yao, K. & Parkhill, J. Compressing physics with an autoencoder: creating an atomic species representation to improve machine learning models in the chemical sciences. *J. Chem. Phys.* **151**, 455 (2019).
139. Goryaeva, A. M., Maillet, J.-B. & Marinica, M.-C. Towards better efficiency of interatomic linear machine learning potentials. *Comput. Mater. Sci.* **166**, 200 (2019).
140. Schütt, K. T., Gastegger, M., Tkatchenko, A. & Müller, K.-R. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W. et al.) 311–330 (Springer, 2019).
141. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
142. Reveil, M. & Clancy, P. Classification of spatially resolved molecular fingerprints for machine learning applications and development of a codebase for their implementation. *Mol. Syst. Des. Eng.* **3**, 431 (2018).
143. del Rosario, Z., Rupp, M., Kim, Y., Antono, E. & Ling, J. Assessing the frontier: active learning, model accuracy, and multi-objective candidate discovery and optimization. *J. Chem. Phys.* **153**, 024112 (2020).
144. Csányi, G., Albaret, T., Payne, M. C. & De Vita, A. "Learn on the fly": a hybrid classical and quantum-mechanical molecular dynamics simulation. *Phys. Rev. Lett.* **93**, 175503 (2004).
145. Sutton, C. et al. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 4428 (2020).
146. Artrith, N., Morawietz, T. & Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **83**, 153101 (2011).
147. Grisafi, A. & Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **151**, 204105 (2019).
148. Stoudenmire, E. M. & Schwab, D. J. Supervised learning with tensor networks. In *Advances in Neural Information Processing Systems* 29 (NeurIPS), 4799 (2016).
149. Analytics Toolkit of the Novel Materials Discovery (NOMAD) Laboratory, https://analytics-toolkit.nomad-coe.eu.
150. Zuo, Y. et al. : Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731 (2020).
151. Stuke, A. et al. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J. Chem. Phys.* **150**, 204121 (2019).
152. Onat, B., Ortner, C. & Kermode, J. R. Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials. *J. Chem. Phys.* **153**, 144106 (2020).
153. Käser, S., Koner, D., Christensen, A. S., von Lilienfeld, O. A. & Meuwly, M. ML models of vibrating H_2CO: Comparing reproducing kernels, FCHL and PhysNet. *J. Phys. Chem. A* **124**, 8853 (2020).
154. Available as part of the software RuNNer at http://www.uni-goettingen.de/de/560580.html, GPL license, per email request).
155. Seko, A., Togo, A. & Tanaka, I. Group-theoretical high-order rotational invariants for structural representations: application to linearized machine learning interatomic potential. *Phys. Rev. B* **99**, 214108 (2019).
156. Seko, A. Machine learning potentials for multicomponent systems: The Ti-Al binary system. *Phys. Rev. B* **102**, 174104 (2020).
157. Available as part of the software LAMMPS (large-scale atomic/molecular massively parallel simulator, http://lammps.sandia.gov, GPL license, publicly accessible).
158. Available as part of the software qmmlpack (quantum mechanics machine learning package) at https://gitlab.com/qmml/qmmlpack, Apache 2.0 license, publicly accessible.
159. Bartók, A. P. et al. Machine learning unifies the modelling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
160. Available as part of the software libAtoms (http://www.libatoms.org, custom license, per webform request).
161. Available as part of the software QML (quantum machine learning, https://www.qmlcode.org/, MIT license, publicly accessible).
162. Andreux, M. et al. Kymatio: scattering transforms in Python. *J. Mach. Learn. Res.* **21**, 1 (2020).
163. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171 (2017).
164. Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. Machine learning of molecular properties: locality and active learning. *J. Chem. Phys.* **148**, 241727 (2018).
165. Shapeev, A. V. Applications of machine learning for representing interatomic interactions. In (eds. Oganov, A. R. et al.) *Computational Materials Discovery* Ch. 3, 66 (Royal Society of Chemistry, 2019).
166. Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. The MLIP package: Moment tensor potentials with MPI and active learning. *Mach. Learn. Sci. Tech.* **2**, 025002 (2021).
167. A reference implementation in Python can be found at https://doi.org/10.5281/ZENODO.1054550, CC BY-SA 4.0 license, publicly accessible.
168. Pseudo-code is available as part of the supporting information at http://pubs.acs.org/doi/abs/10.1021/acs.jctc.8b00110.
169. Husch, T., Sun, J., Cheng, L., Lee, S. J. R. & Miller III, T. F. Improved accuracy and transferability of molecular-orbital-based machine learning: organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **154**, 064108 (2021).
170. Lee, S. J. R., Husch, T., Ding, F. & Miller III, T. F. Analytical gradients for molecular-orbital-based machine learning. *J. Chem. Phys.* **154**, 124120 (2021).
171. Lysogorskiy, Y. et al. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Comput. Mater.* **7**, 97 (2021).
172. An implementation in Julia can be found at https://github.com/ACEsuit/ACE.jl, ASLv1 license, publicly accessible.
173. An implementation in Python can be found at https://github.com/cosmo-epfl/nice, MIT license, publicly accessible.
174. Uhrin, M. Through the eyes of a descriptor: constructing complete, invertible descriptions of atomic environments. *Phys. Rev. B* **104**, 144110 (2021).
175. An implementation in Python can be found at https://github.com/muhrin/milad, GPLv3 license, publicly accessible.

ACKNOWLEDGEMENTS

This work received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreements No. 676580, the NOMAD Laboratory CoE,

and No. 740233, ERC: TEC1P. It was funded in part by the German Ministry for Education and Research as BIFOLD—Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A). Part of the research was performed while the authors visited the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1440415). The authors thank Profs. Matthias Scheffler, Klaus-Robert Müller, Jörg Behler, Gábor Csányi, O. Anatole von Lilienfeld, Carsten Baldauf, Matthew Hirn, as well as Emre Ahmetcik, Lauri Himanen, Yair Litman, Dmitrii Maksimov, Felix Mocanu, Wiktor Pronobis, and Christopher Sutton for constructive discussions.

AUTHOR CONTRIBUTIONS

M.F.L. and M.R. designed numerical experiments and analyzed results. M.F.L. developed software and conducted numerical experiments. All authors contributed to writing, with emphasis by A.G. on the mathematical framework and M.F.L. on representations and benchmarking. M.R. supervised the study.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00721-x.

Correspondence and requests for materials should be addressed to Matthias Rupp.

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

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