Unified representation of molecules and crystals for machine learning

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 045017 https://doi.org/10.1088/2632-2153/aca005

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
2 August 2022

REVISED
27 October 2022

ACCEPTED FOR PUBLICATION
3 November 2022

PUBLISHED
21 November 2022

PAPER

Unified representation of molecules and crystals for machine learning

Haoyan Huo[1,4] **and Matthias Rupp**[2,3,5,*]

[1] School of Physics, Peking University, Beijing, People's Republic of China
[2] Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[3] Department of Computer and Information Science, University of Konstanz, Konstanz, Germany
[4] Present address: Department of Materials Science and Engineering, University of California, Berkeley, CA, United States of America.
[5] Present address: Materials Research and Technology Department, Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg.
[*] Author to whom any correspondence should be addressed.

E-mail: mrupp@mrupp.info

Keywords: many-body tensor representation, machine-learning potential, atomistic simulations

Supplementary material for this article is available online

Abstract

Accurate simulations of atomistic systems from first principles are limited by computational cost. In high-throughput settings, machine learning can reduce these costs significantly by accurately interpolating between reference calculations. For this, kernel learning approaches crucially require a representation that accommodates arbitrary atomistic systems. We introduce a many-body tensor representation that is invariant to translations, rotations, and nuclear permutations of same elements, unique, differentiable, can represent molecules and crystals, and is fast to compute. Empirical evidence for competitive energy and force prediction errors is presented for changes in molecular structure, crystal chemistry, and molecular dynamics using kernel regression and symmetric gradient-domain machine learning as models. Applicability is demonstrated for phase diagrams of Pt-group/transition-metal binary systems.

1. Introduction

The computational study of atomistic systems such as molecules and crystals requires accurate treatment of interactions at the atomic and electronic scale. Accurate first-principles methods, however, are limited by their high computational cost. In settings that require many calculations, such as dynamics simulations, phase diagrams, or high-throughput searches, machine learning (ML) [1, 2] can reduce overall costs by orders of magnitude via accurate interpolation between reference calculations [3–5]. For this, the problem of repeatedly solving a complex equation such as Schrödinger's equation *for many related inputs* is mapped onto a non-linear regression problem: instead of numerically solving new systems, they are statistically estimated based on a reference set of known solutions [6, 7]. This ansatz enables, among other applications, screening larger databases of molecules and materials [5, 8], running longer dynamics simulations [9], investigating larger systems [10], and increasing the accuracy of calculations [5, 11].

Kernel-based ML models [12–15] for data-efficient accurate prediction of *ab initio* properties require a single space in which regression is carried out. Representations [16] are functions that map atomistic systems to elements in such spaces, either directly or via a kernel [17]. Representations should be (a) *invariant* against transformations preserving the predicted property, in particular translations, rotations, and nuclear permutations of same elements, as learning these invariances from data would require many reference calculations; non-scalar properties can require equivariance instead of invariance; (b) *unique*, that is, variant against transformations changing the property, as systems with identical representation that differ in property would introduce errors [18]; (c) *continuous*, and ideally *differentiable*, as discontinuities work against the smoothness assumption of the ML model and model gradients are often useful; (d) *general* in the sense of being able to encode any atomistic system, including finite and periodic systems; (e) *fast* to compute,

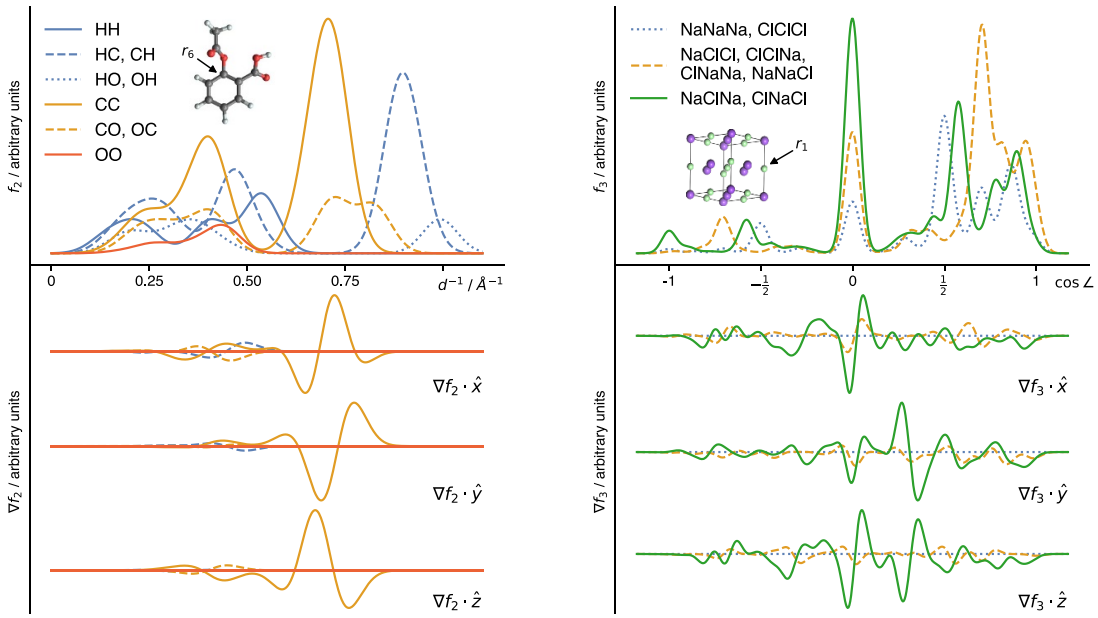

Figure 1. *Visualization of many-body tensor representation.* Top panels show distributions of inverse distances ($k = 2$, quadratic weighting) for aspirin ($C_9O_4H_8$, left) and distributions of angles ($k = 3$, exponential weighting) for fcc salt (NaCl, right). Bottom panels show derivatives of the representation, obtained by differentiating with respect to the Cartesian coordinates of C atom r_6 connecting the ester group (left) and the Na atom r_1 at lattice point (right).

as the goal is to reduce computational cost; (f) *data-efficient* in the sense of requiring few reference calculations to reach a given target error. Constant size is an advantage, [19] as is the ability to encode the whole system as well as local atomic environments. Requirements (e) and (f) are in practice determined empirically. See [16, 20–22] for details on these and further requirements.

Some representations fulfill these requirements only partially, such as the Coulomb matrix (CM) [6] and bag of bonds (BoB) [23] discussed below. State-of-the-art representations often fulfill these requirements in some limit, such as infinite expansion order. See [16] for a comprehensive and detailed discussion. The descriptors used in cheminformatics, [24] and sometimes in materials informatics, often violate (b) and (c), in particular if they do not include atomic coordinate information or rely on cutoff-based definitions of chemical bonds. Such descriptors serve the different purpose of directly predicting derived properties that are not functions of a single conformation, such as solubility or binding affinity to a macromolecule.

We introduce a many-body tensor representation (MBTR) derived from CM/BoB and concepts of many-body expansions. It is related [16] to Behler–Parrinello symmetry functions [25] and histograms of distances, angles, and dihedral angles [26]. MBTR fulfills the above requirements in the limit, is interpretable, allows visualization (figure 1), and describes finite and periodic systems. State-of-the-art empirical performance is demonstrated by us for organic molecules and inorganic crystals, as well as applicability to phase diagrams of Pt-group / transition metal binary systems, and by others for predicting and optimizing various molecular [27–33] and crystalline properties [34–37]. Implementations of MBTR are publicly available (see Code Availability section at the end).

2. Method

We start from the CM [6, 10, 38], which represents a molecule $\mathcal{M}$ as a symmetric atom-by-atom matrix

$$\boldsymbol{M}_{i,j} = \begin{cases} \frac{1}{2} Z_i^{2.4} & i = j \\ \frac{Z_i Z_j}{d_{i,j}} & i \neq j \end{cases}, \tag{1}$$

where Z_i are atomic numbers and $d_{i,j} = ||\boldsymbol{R_i} - \boldsymbol{R_j}||$ is Euclidean distance between atoms i and j. To avoid dependence on atom ordering (in the input), which would violate (a), $\boldsymbol{M}$ is either diagonalized, loosing information which violates (b) [18], or sorted, causing discontinuities that violate (c). Another shortcoming is the use of Z, which is not well suited for interpolation [39] as it overly decorrelates chemical elements from the same column of the periodic table.

The related BoB [23] representation uses the same terms, but arranges them differently. For each pair of chemical elements, corresponding CM terms are stored in sorted order, which can be viewed as an

$N_e \times N_e \times d$ tensor, or an $N_e \times (N_e+1)/2 \times d$ tensor if symmetry is taken into account, where N_e is number of elements and d is sufficiently large. Unlike the CM, it can not distinguish homometric molecules [21], which might distort its feature space [40]. While the BoB tensor itself does not suffer from discontinuities, its derivative does.

To derive MBTR, we retain stratification by elements, but avoid the sorting by arranging distances on a real-space axis:

$$f_{\mathrm{BoB}}(x,z_1,z_2) = \sum_{i,j=1}^{N_a} \delta\left(x - d_{i,j}^{-1}\right)\delta(z_1,Z_i)\delta(z_2,Z_j), \tag{2}$$

where x is a real number, z_1, z_2 are atomic numbers, N_a is number of atoms, $\delta(\cdot)$ is Dirac's delta, and $\delta(\cdot,\cdot)$ is Kronecker's delta. f_{BoB} has mixed continuous-discrete domain and encodes all (inverse) distances between atoms with elements z_1 and z_2. For a smoother measure, we replace Dirac's δ with another probability distribution $\mathcal{D}$, 'broadening' or 'smearing' it [20, 41]. In this work, we use the normal distribution. Other distributions can be used, in particular symmetric and short-tailed ones, for example, the Laplace distribution or the uniform distribution. We did not observe significant differences in performance, however. Adding a weighting function w_2 and replacing the Kronecker δ functions by an element correlation matrix $C \in \boldsymbol{R}^{N_e \times N_e}$ yields

$$f_2(x,z_1,z_2) = \sum_{i,j=1}^{N_a} w_2(i,j)\,\mathcal{D}\left(x, g_2(i,j)\right) C_{z_1,Z_i} C_{z_2,Z_j} \tag{3}$$

of which equation (2) is a special case. In general, g_2 describes a relation between atoms i and j, such as their inverse distance, $\mathcal{D}$ broadens the result of g_2, and w_2 allows to weight down contributions, for example, from far-away atoms. Building on the idea of many-body expansions, [42, 43] we generalize from f_2 in equation (3), which encodes two-body terms, to the MBTR equation

$$f_k(x,\boldsymbol{z}) = \sum_{\boldsymbol{i}=1}^{N_a} w_k(\boldsymbol{i})\mathcal{D}\left(x, g_k(\boldsymbol{i})\right) \prod_{j=1}^{k} C_{z_j,Z_{i_j}}, \tag{4}$$

where $\boldsymbol{z} \in \boldsymbol{N}^k$ are atomic numbers, $\boldsymbol{i} = (i_1,\ldots,i_k) \in \{1,\ldots,N_a\}^k$ are index tuples, and w_k, g_k assign a scalar to k atoms in $\mathcal{M}$ [44]. Canonical choices of g_k for $k=1,2,3,4$ are atom counts, (inverse) distances, angles, and dihedral angles. The element correlation matrices C allow exploitation of similarities between chemical element species ('alchemical learning'), for example, within the same column of the periodic table [45–47].

We measure the similarity of two atomistic systems $\mathcal{M}$ and $\mathcal{M}'$ as the Euclidean distance between their representations,

$$d_k^2(\mathcal{M},\mathcal{M}') = \sum_{\boldsymbol{z}} \int \left(f_k(x,\boldsymbol{z}) - f_k'(x,\boldsymbol{z})\right)^2 \mathrm{d}x. \tag{5}$$

In practice, we adjust equation (4) for symmetries. Discretizing the continuous axis as $(x_{\min}, x_{\min}+\Delta x,\ldots,x_{\max})$ results in a rank $k+1$ tensor of dimensions $N_e \times \cdots \times N_e \times N_x$ with $N_x = (x_{\max}-x_{\min})/\Delta x$, where $x_{\min}$ and $x_{\max}$ are the smallest and largest values for which $f_k(x,\boldsymbol{z}) \neq 0$ for all $\boldsymbol{z}$ and $\mathcal{M}$. Linearizing element ranks yields $N_e^k \times N_x$ matrices, allowing for visualization (figure 1) and efficient numerical implementation via linear algebra routines. For systems with many element species, discretization can lead to large matrices, requiring substantial amounts of memory. In such settings, memory-efficient implementation via sparse matrix formats or on-the-fly calculation of distances and inner products (see, e.g. [45]) of MBTR matrices might be preferable.

Periodic systems, used to model bulk crystals and surfaces, can be viewed as unit cells surrounded by infinitely many translated images of themselves. For such systems, $N_a = \infty$ and the sum in equation (4) diverges. We prevent this by requiring one index of $\boldsymbol{i}$ to be in the (same) primitive unit cell [48]. This accounts for translational symmetry and prevents double-counting. Use of weighting functions w_k such as exponentially decaying weights [49] then ensures convergence of the sum. Figure 1 (right) presents the resulting distributions of angles for face-centered cubic (FCC) NaCl as an example. Note that the k-body terms g_k do not depend on choice of unit cell geometry (lattice vectors). This ensures unique representation of Bravais lattices where the choice of basis vectors is not unique, for example 2D hexagonal lattices where the angle between lattice vectors can be $\frac{1}{3}\pi$ or $\frac{2}{3}\pi$.

Table 1. *Prediction errors for small organic molecules.* Machine-learning models of atomization energies E and isotropic polarizabilities α, obtained at hybrid density functional level of theory, were trained on 5k molecules and evaluated on 2k others using different representations. RMSE = root mean square error, MAE = mean absolute error, CM = Coulomb matrix, BoB = bag of bonds, BAML = bonding angular machine learning, SOAP = smooth overlap of atomic positions, FCHL19 = Faber–Christensen–Huang–Lilienfeld representation, MBTR = many-body tensor representation.

		E (kcal mol^{-1})		α (Å)^3	
Representation	Kernel	MAE	RMSE	MAE	RMSE
CM [6]	Laplacian	3.47	4.76	0.13	0.17
BoB [23]	Laplacian	1.78	2.86	0.09	0.12
BAML [42]	Laplacian	1.15	2.54	0.07	0.12
SOAP [55]	REMatch	0.92	1.61	0.05	0.07
FCHL19 [45, 47]	Gaussian	0.44	—	—	—
MBTR	Linear	0.74	1.14	0.07	0.10
MBTR	Gaussian	0.60	0.97	0.04	0.06

Many applications, including dynamics simulations and structural relaxation, require forces, the negative gradient of the energy with respect to atomic coordinates. The gradient of equation (4) is given by

$$\nabla f_k(x,\boldsymbol{z}) = \sum_{\boldsymbol{i}=1}^{N_a}\Bigg(\mathcal{D}\big(x,g_k(\boldsymbol{i})\big)\nabla w_k(\boldsymbol{i}) + w_k(\boldsymbol{i})\frac{\partial\mathcal{D}\big(x,g_k\big)}{\partial g_k}\nabla g_k(\boldsymbol{i})\Bigg)\prod_{j=1}^{k} C_{z_j,Z_{i_j}}. \tag{6}$$

The gradient $\nabla f_k(x,\boldsymbol{z})$ can be derived analytically if this is possible for ∇w_k, ∇g_k, and $\nabla\mathcal{D}$. Alternatively, automatic differentiation [50, 51] can be used, removing the need for manual derivation. Figure 1 visualizes MBTR gradients.

3. Results

To validate MBTR, we demonstrate accurate predictions for properties of molecules and crystals. Focusing on the representation, we employ plain kernel ridge regression (KRR) models [7] unless stated otherwise.

3.1. Changes in molecular structure

To demonstrate interpolation across changes in the chemical structure of molecules we utilize a benchmark dataset [38] of 7211 small organic molecules composed of up to seven C, N, O, S and Cl atoms, saturated with H. Molecules were relaxed to their ground state using the Perdew–Burke–Ernzerhof (PBE) [52] approximation to Kohn–Sham density functional theory (DFT). Restriction to relaxed structures projects out spatial variability and allows focusing on changes in chemical structure. Table 1 presents prediction errors for atomization energies and isotropic polarizabilities obtained from single point calculations with the hybrid PBE0 [53, 54] functional. For 5k training samples, prediction errors are below 1 kcal mol^{-1} ('chemical accuracy'), with the MBTR model's mean absolute error (MAE) of 0.6 kcal mol^{-1} corresponding to thermal fluctuations at room temperature. Note that MBTR achieves similar performance with a linear regression model, allowing constant-time predictions.

3.2. Changes in crystal chemistry

Interpolation across changes in the chemistry of crystalline solids is demonstrated for a dataset of 11k elpasolite structures (ABC_2D_6, $AlNaK_2F_6$ prototype) [56, 57] composed of 12 different elements, with geometries and energies computed at DFT/PBE level of theory. Predicting formation energies with MBTR yields a root-mean-squared-error (RMSE) of 8.1 meV/atom and MAE of 4.7 meV/atom (figure 2) for a training set of 9k crystals.

Adding chemical elements should increase the intrinsic dimensionality of the learning problem, and thus prediction errors. To verify this, we created a dataset of 4611 ABC_2 ternary alloys containing 22 non-radioactive elements from groups 1, 2, 13–15, spanning five rows and columns of the periodic table. Structures were taken from the Open Quantum Materials Database (OQMD) [58, 59], with geometries and properties also computed via DFT/PBE. As expected, energy predictions exhibit larger errors (RMSE 31 meV/atom, MAE 23 meV/atom) compared to an elpasolite model of same training set size (RMSE 23 meV/atom, MAE 15 meV/atom).

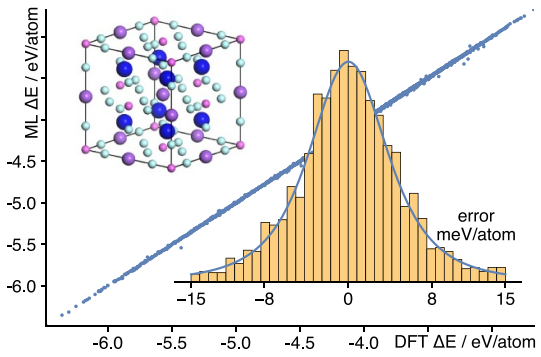

Figure 2. *Formation energy predictions for ABC_2D_6 elpasolite structures* containing 12 different elements. Shown are reference energies (DFT ΔE) and predicted energies (ML ΔE), as well as distribution of errors (inset) for 2272 crystals, from an MBTR ML model trained on 9086 other ones.

Table 2. *Energy and force prediction errors for changes in geometry of organic molecules.* Shown are prediction errors for total energies (kcal mol^{-1}) and atomic forces (kcal mol^{-1} Å^{-1}). MAE = mean absolute error, RMSE = root mean squared error, PaiNN = polarizable atom interaction neural network [60], FCHL19 = Faber–Christensen–Huang–Lilienfeld representation [45, 47], sGDML = symmetric gradient domain machine learning [13], sMatérn = Matérn kernel augmented with symmetric permutations for sGDML [61], CM^{md} = Coulomb matrix variant, MBTR = many-body tensor representation.

	Trained only on forces, 1k references								Trained only on energies, 10k references		
Kernel	PaiNN —		FCHL19 Gaussian		sGDML/CM^{md} sMatérn		sGDML/MBTR sMatérn		CM^{md} Gaussian	MBTR linear	MBTR Gaussian
Molecule	Energy MAE	Force MAE	Energy MAE	Force MAE	Energy MAE	Force MAE	Energy MAE	Force MAE	Energy MAE	Energy MAE	Energy MAE
Benzene	—	—	—	—	0.07[a]	0.06[a]	0.07	0.15	0.03	0.03	0.03
Uracil	0.11	0.13	0.10	0.10	0.11	0.24	0.11	0.17	0.05	0.10	0.03
Naphthalene	0.12	0.08	0.12	0.15	0.12	0.11	0.11	0.09	0.12	0.10	0.07
Aspirin	0.17	0.34	0.17	0.50	0.19	0.68	0.17	0.48	0.36	0.21	0.25
Salicylic acid	0.12	0.20	0.12	0.22	0.12	0.28	0.11	0.18	0.11	0.13	0.07
Malonaldehyde	0.10	0.34	0.08	0.25	0.10	0.41	0.09	0.36	0.18	0.21	0.10
Ethanol	0.06	0.22	0.05	0.14	0.07	0.33	0.06	0.26	0.17	0.17	0.06
Toluene	0.10	0.09	0.10	0.20	0.10	0.14	0.09	0.13	0.16	0.11	0.10

[a] We observed higher noise in predictions of benzene, whose reported prediction errors are also inconsistent in different publications. To make results comparable, we retrained the sGDML/CM^{md} model (originally reported values are 0.10 and 0.06).

3.3. Changes in molecular geometry

For interpolation of changes in molecular geometry, we employ a benchmark dataset [13, 62] of *ab initio* molecular dynamics trajectories of eight organic molecules. Each molecule was simulated at a temperature of 500 K for between 150k and 1M time steps of 0.5 fs, with energies and forces computed at the DFT/PBE level of theory and the Tkatchenko–Scheffler model [63] for van der Waals interactions. Table 2 presents results for models trained and evaluated only on energies (right-hand side) and only on forces (left-hand side).

Energy-only models were trained on 10k configurations and validated on 2k other ones, employing MBTR (parametrized for dynamics data, see supplement) and a similarly modified CM (CM^{md}, see supplement). Non-linear MBTR regression performs best overall, with the linear kernel again being competitive.

On the one hand, differentiating energy-based ML potentials can introduce errors, for example, from small oscillations between training samples due to insufficient regularization, and from insufficient model constraints in directions not covered by the training data [64]. On the other hand, electronic structure calculations often provide reference forces at not additional cost. It is therefore beneficial to include these in model training. This often reduces the required number of reference calculations by an order of magnitude [9, 13, 61, 65].

Force-only models require an adaptation of plain KRR. To accommodate forces into training and to demonstrate use of MBTR with other regression approaches, we employ MBTR in the framework of symmetrized gradient-domain machine-learning (sGDML) [61]. This approach uses the Matérn kernel, augmentation by symmetric molecular permutations, and reference forces for training, while providing energy and force predictions.

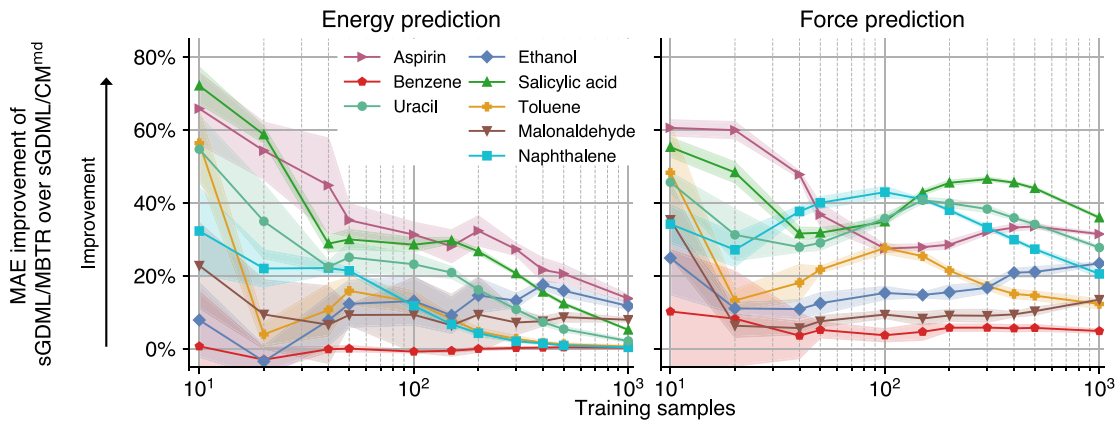

Figure 3. *Relative improvement in predictive accuracy on dynamics data of eight different organic molecules.* Shown are force and energy prediction MAE ratios of sGDML/MBTR over sGDML/CM^{md} as a function of training set size. Error bars show the standard deviation of the ratios over five runs with different random seeds.

Table 2 (left-hand side) compares performance of the original sGDML approach (based on the CM^{md} representation [13, 61]) and of sGDML based on a two-body MBTR representation. Both models were trained on 1000 configurations, leading to kernel matrices with dimensionalities between 27k and 63k. For reference, we also present results for the Faber–Christensen–Huang–Lilienfeld (FCHL) representation [45, 47] and the polarizable atom interaction neural network (PaiNN) [60].

sGDML/MBTR performs as good or better than sGDML/CM^{md} for energy and force predictions. Compared to PaiNN, sGDML/MBTR performs better for energy predictions, but worse for forces. For a more fine-grained comparison between the sGDML models, figure 3 presents learning curves of relative MAE ratios of sGDML/MBTR over sGDML/CM^{md}, together with standard deviations over five runs starting from different random seeds. sGDML/MBTR consistently outperforms sGDML/CM^{md} with error reductions up to 50%–60%, especially when less than 100 training samples are used. For more symmetric molecules such as benzene, malonaldehyde, and ethanol, use of MBTR is less beneficial but still an improvement.

3.4. Phase diagrams

We demonstrate applicability by identifying the convex hull of the phase diagram for Pt-group/transition metal binary alloys, relevant for industrial applications [66]. For a given dataset of candidate structures, we predict the energy of each structure and identify those with the lowest energy, which form the convex hull in a phase diagram. Compositions that lie on or slightly above the convex hull correspond to stable and meta-stable alloys, respectively.

To demonstrate this, we use a dataset [66] of 153 alloys computed at the DFT/PBE level of theory. This dataset contains at most a few hundred structures for each alloy. Due to this small amount of data direct application of ML models results in errors in predicted energies that are large enough to lead to wrong convex hulls. We address this by employing a simple active learning [67] scheme.

Starting with a few randomly selected structures, we iteratively train ML models on these and predict energies of candidate structures. In each iteration, we calculate (look up) DFT energies only for structures predicted to be low in energy and include these in the training dataset of the next iteration. This procedure prevents computationally expensive DFT calculations for high-energy structures that lie above the convex hull, saving up to 48% of all DFT calculations while still identifying the correct convex hull.

Figure 4 presents results for AgPt. The active learning model selected 357 DFT calculations for training and predicted energies of 331 (48%) other structures, with a MAE of 39 meV/atom. The trade-off between the number of saved calculations and the probability of failing to identify the correct convex hull can be explicitly controlled by adjusting the energy threshold below which DFT calculations are requested. In this simple demonstration, structures are given and not derived from composition by relaxation. While structural relaxation is possible with ML, it brings its own challenges [68–76].

3.5. Other uses

MBTR has been used to study structure and properties of molecules, clusters, crystals and other atomistic systems. Studies related to properties include predictions of

- gas-particle partition coefficients, such as saturation vapor pressure and equilibrium partitioning coefficients, of atmospheric molecules via KRR [27]

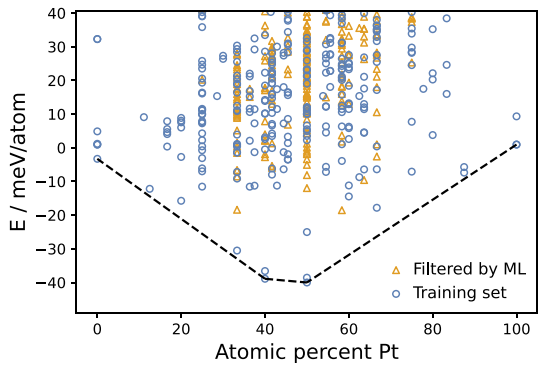

Figure 4. *Phase diagrams of Pt-group/transition metal binary alloys.* Shown are Ag_xPt_{1-x} structures (points) and their convex hull (dashed line) as given by DFT and identified by ML. All structures are shown at their DFT energy. Symbols indicate whether a structure was selected for training (blue circles) or predicted as high-energy (orange triangles). See main text for details.

- Heisenberg exchange spin coupling constants for dicopper complexes via Gaussian process regression [28]
- total and orbital energies of diverse larger organic molecules from the OE62 dataset [77] via a graph neural network [78]
- extrapolation of size-extensive properties [79] at the example of atomization energies of organic molecules in the QM9 [8] and OE62 [77] datasets
- formation energies of Al–Ni and Cd–Te binary compounds via support vector regression [34]
- band gaps and formation energies of perovskite-like materials [35]
- energetics of compositionally disordered compounds via KRR [80].

Studies related to structure include

- visualization of the conformational space of tannic acid molecules via principal component analysis [29]
- global optimization of atomic clusters, including electronic spin multiplicities, via active learning and Gaussian process regression [30, 31]
- derivative-free structural relaxation of water and small unbranched alkanes via KRR and simulated annealing [32]
- identification of low-energy point defects in solids via evolutionary algorithms, clustering, and Gaussian process regression [36]
- Monte Carlo simulations of the thermodynamics of thiolate-protected gold nanoclusters via minimal learning machine regression [37]
- developing data-efficient ML potentials at the example of Cs^+ in water via active learning and Gaussian process regression [33].

4. Discussion and outlook

MBTR is a general representation (numerical description, feature set) of atomistic systems for fast accurate interpolation between quantum-mechanical calculations via ML. It is based on distributions of k-atom terms stratified by chemical elements. Despite, or because of, this simple principle, it is connected to many other representations, including CM [6], BoB [23], histograms of distances, angles and dihedral angles [26], atom-centered symmetry functions [25], partial radial distribution functions [81], FCHL representation [45, 47], as well as cluster expansion [82]. See [16] for further details on these and other relationships.

MBTR represents whole molecules and crystals. With increasing number of atoms, and thus degrees of freedom, this approach is likely to degrade, and exploitation of locality via prediction of additive atomic energy contributions becomes appealing [9, 83]. This requires representing local chemical environments [20], for which MBTR can be modified [34, 84, 85].

We note in passing that problems in the training of ML models, such as outliers, can often be traced back to problems in the underlying reference calculations, such as unconverged fast Fourier transform grids or inconsistent settings (violating the assumption that a single function is being fitted), a phenomenon also observed by others [86]. This suggests that automated identification of errors in big datasets of electronic structure calculations via parametrization of ML models might be a general approach for validation of such datasets. We rationalize this hypothesis by ML models identifying regularity (correlations) in data, and faulty calculations deviating in some way from correct ones.

Continuing advances in electronic structure codes and increasing availability of large-scale computing resources have led to large collections of *ab initio* calculations, such as Materials Project [87], AFLOWlib [88], OQMD [59], and Novel Materials Discovery Laboratory [89]. Representations such as MBTR are key to combine quantum mechanics with machine learning (QM/ML) for fast, accurate and precise interpolation in these settings.

Data availability statement

All datasets used in this study are publicly available. Implementations of MBTR are available as part of the DScribe [85] and qmmlpack [90] libraries. Code to reproduce results of reported experiments is available at: https://github.com/hhaoyan/mbtr.

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.19567324.v1.

Acknowledgments

M R and H H thank Matthias Scheffler for helpful discussions and support. M R thanks the Institute for Pure and Applied Mathematics (IPAM) for hospitality and support, participants of its program on Understanding Many-Particle Systems with Machine Learning for feedback, and Albert Bartók-Pártay, Gábor Csányi, Alexander Shapeev, Alexandre Tkatchenko and the late Alessandro de Vita for extended discussions. M R acknowledges funding from EU Horizon 2020 program Grant 676580, The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence.

This article is an extended and updated version of the original arXiv preprint 1704.06439, whose formal publication was delayed for non-technical reasons.

ORCID iDs

Haoyan Huo https://orcid.org/0000-0003-2227-9121
Matthias Rupp https://orcid.org/0000-0002-2934-2958

References

[1] Ghahramani Z 2015 Probabilistic machine learning and artificial intelligence *Nature* **521** 452–9
[2] Jordan M I and Mitchell T M 2015 Machine learning: trends, perspectives and prospects *Science* **349** 255–60
[3] Jinnouchi R, Karsai F and Kresse G 2019 On-the-fly machine learning force field generation: application to melting points *Phys. Rev.* B **100** 014105
[4] Sendek A D, Cubuk E D, Antoniuk E R, Cheon G, Cui Y and Reed E J 2019 Machine learning-assisted discovery of solid Li-ion conducting materials *Chem. Mater.* **31** 342–52
[5] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 Big data meets quantum chemistry approximations: the Δ-machine learning approach *J. Chem. Theor. Comput.* **11** 2087–96
[6] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[7] Rupp M 2015 Machine learning for quantum mechanics in a nutshell *Int. J. Quant. Chem.* **115** 1058–73
[8] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
[9] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403
[10] Rupp M, Ramakrishnan R and von Lilienfeld O A 2015 Machine learning for quantum mechanical properties of atoms in molecules *J. Phys. Chem. Lett.* **6** 3309–13
[11] Bartók A P, Gillan M J, Manby F R and Csányi G 2013 Machine-learning approach for one- and two-body corrections to density functional theory: applications to molecular and condensed water *Phys. Rev.* B **88** 054104
[12] Rupp M 2015 Machine learning for quantum mechanics in a nutshell *Int. J. Quant. Chem.* **115** 1058–73
[13] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields *Sci. Adv.* **3** e1603015
[14] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073–141
[15] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K-R 2021 Machine learning force fields *Chem. Rev.* **121** 10142–86
[16] Langer M F, Goeßmann A and Rupp M 2022 Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning *npj Comput. Mater.* **8** 41
[17] Kernel methods use a positive definite function (kernel) to implicitly define the Hilbert space. We focus on explicit numerical representations as input for vector kernels.
[18] Moussa J E 2012 Comment on "Fast and accurate modeling of molecular atomization energies with machine learning" *Phys. Rev. Lett.* **109** 059801

[19] Collins C R, Gordon G J, von Lilienfeld O A and Yaron D J 2018 Constant size descriptors for accurate machine learning models of molecular properties *J. Chem. Phys.* **148** 241718
[20] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[21] von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties *Int. J. Quant. Chem.* **115** 1084–93
[22] Onat B, Ortner C and Kermode J R 2020 Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials *J. Chem. Phys.* **153** 144106
[23] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O A, Müller K-R and Tkatchenko A 2015 Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326–31
[24] Todeschini R and Consonni V 2009 *Handbook of Molecular Descriptors* 2nd edn (Weinheim: Wiley)
[25] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[26] Faber F A, Hutchison L, Huang B, Gilmer J, Schoenholz S S, Dahl G E, Vinyals O, Kearnes S, Riley P F and von Lilienfeld O A 2017 Prediction errors of molecular machine learning models lower than hybrid DFT error *J. Chem. Theor. Comput.* **13** 5255–64
[27] Lumiaro E, Todorović M, Kurten T, Vehkamäki H and Rinke P 2021 Predicting gas-particle partitioning coefficients of atmospheric molecules with machine learning *Atmos. Chem. Phys.* **21** 13227–46
[28] Bahlke M P, Mogos N, Proppe J and Herrmann C 2020 Exchange spin coupling from Gaussian process regression *J. Phys. Chem.* A **124** 8708–23
[29] Petry R, Focassio B, Schleder G R, Martinez D S T and Fazzio A 2021 Conformational analysis of tannic acid: environment effects in electronic and reactivity properties *J. Chem. Phys.* **154** 224102
[30] Lourenço M P, Herrera L B, Hostaš J, Calaminici P, Köster A M, Tchagang A and Salahub D R 2021 Taking the multiplicity inside the loop: active learning for structural and spin multiplicity elucidation of atomic clusters *Theor. Chem. Acc.* **140** 116
[31] Lourenço M P, Galvão B R L, Herrera L B, Hostaš J, Tchagang A, Silva M X and Salahub D R 2021 A new active learning approach for global optimization of atomic clusters *Theor. Chem. Acc.* **140** 62
[32] Iype E and Urolagin S 2019 Machine learning model for non-equilibrium structures and energies of simple molecules *J. Chem. Phys.* **150** 024307
[33] Zhai Y, Caruso A, Gao S and Paesani F 2020 Active learning of many-body configuration space: application to the Cs^+-water MB-nrg potential energy function as a case study *J. Chem. Phys.* **152** 144103
[34] Honrao S J, Xie S R and Hennig R G 2020 Augmenting machine learning of energy landscapes with local structural information *J. Appl. Phys.* **128** 085101
[35] Mayr F and Gagliardi A 2021 Global property prediction: a benchmark study on open-source, perovskite-like datasets *ACS Omega* **6** 12722–32
[36] Arrigoni M and Madsen G K H 2021 Evolutionary computing and machine learning for discovering of low-energy defect configurations *npj Comput. Mater.* **7** 71
[37] Pihlajamäki A, Hämäläinen J, Linja J, Nieminen P, Malola S, Kärkkäinen T and Häkkinen H 2020 Monte Carlo simulations of $Au_{38}(SCH_3)_{24}$ nanocluster using distance-based machine learning methods *J. Phys. Chem.* A **124** 4827–36
[38] Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A, Müller K-R and von Lilienfeld O A 2013 Machine learning of molecular electronic properties in chemical compound space *New J. Phys.* **15** 095003
[39] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 Big data of materials science: critical role of the descriptor *Phys. Rev. Lett.* **114** 105503
[40] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G and Ceriotti M 2020 Incompleteness of atomic structure representations *Phys. Rev. Lett.* **125** 166001
[41] Cubuk E D, Schoenholz S S, Rieser J M, Malone B D, Rottler J, Durian D J, Kaxiras E and Liu A J 2015 Identifying structural flow defects in disordered solids using machine-learning methods *Phys. Rev. Lett.* **114** 108001
[42] Huang B and von Lilienfeld O A 2016 Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity *J. Chem. Phys.* **145** 161102
[43] Yao K, Herr J E and Parkhill J 2017 The many-body expansion combined with neural networks *J. Chem. Phys.* **146** 014106
[44] We use scalar geometry functions g_k for convenience; assigning vectors would simply increase the rank of the tensor. The product structure $w_k(\boldsymbol{i})\mathcal{D}(x, g_k(\boldsymbol{i}))$ allows efficient implementation as $\mathcal{D}$ does not depend on $\mathcal{M}$.
[45] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717
[46] Herr J E, Koh K, Yao K and Parkhill J 2019 Compressing physics with an autoencoder: creating an atomic species representation to improve machine learning models in the chemical sciences *J. Chem. Phys.* **151** 455–72
[47] Christensen A S, Bratholm L A, Faber F A and von Lilienfeld O A 2020 FCHL revisited: faster and more accurate quantum machine learning *J. Chem. Phys.* **152** 044107
[48] Effectively representing one unit cell, including influence of surrounding cells on it, in accordance with computed properties being reported per cell.
[49] Exponential weighting was motivated by the exponential decay of screened Coulombic interactions in solids.
[50] Frostig R, Johnson M J and Leary C 2018 Compiling machine learning programs via high-level tracing *1st Conf. on Systems and Machine Learning (SysML 2018) (Stanford, California, 15–16 February 2018)* (available at: https://mlsys.org/Conferences/doc/2018/146.pdf)
[51] Paszke A *et al* 2019 PyTorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems 32 (NeurIPS 2019), (Montréal, Canada, 8–14 December 2019)*, ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox and R Garnett Curran Associates pp 8024–35 (available at: https://neurips.cc/Conferences/2019)
[52] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
[53] Perdew J P, Ernzerhof M and Burke K 1996 Rationale for mixing exact exchange with density functional approximations *J. Chem. Phys.* **105** 9982–5
[54] Adamo C and Barone V 1999 Toward reliable density functional methods without adjustable parameters: the PBE0 model *J. Chem. Phys.* **110** 6158–70
[55] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69

[56] Faber F A, Lindmaa A, von Lilienfeld O A and Armiento R 2016 Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals *Phys. Rev. Lett.* **117** 135502
[57] Rupp M 2022 Dataset ABC2D6-16 (available at http://qmml.org)
[58] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD) *J. Miner. Met. Mater. Soc.* **65** 1501–9
[59] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies *npj Comput. Mater.* **1** 15010
[60] Schütt K T, Unke O T and Gastegger M 2021 Equivariant message passing for the prediction of tensorial properties and molecular spectra *Proc. 38th Int. Conf. on Machine Learning (ICML 2021) (Virtual, 18–24 July 2021)*, ed M Meila and T Zhang (Proceedings of Machine Learning Research) pp 9377–88 (available at: https://proceedings.mlr.press/v139/schutt21a.html)
[61] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887
[62] Schütt K T, Arbabzadah F, Chmiela S, Müller K R and Tkatchenko A 2017 Quantum-chemical insights from deep tensor neural networks *Nat. Commun.* **8** 13890
[63] Tkatchenko A and Scheffler M 2009 Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data *Phys. Rev. Lett.* **102** 073005
[64] Snyder J C, Rupp M, Hansen K, Müller K-R and Burke K 2012 Finding density functionals with machine learning *Phys. Rev. Lett.* **108** 253002
[65] Glielmo A, Sollich P and De Vita A 2017 Accurate interatomic force fields via machine learning with covariant kernels *Phys. Rev.* B **95** 214302
[66] Hart G L W, Curtarolo S, Massalski T B and Levy O 2013 Comprehensive search for new phases and compounds in binary alloy systems based on platinum-group metals, using a computational first-principles approach *Phys. Rev.* X **3** 041035
[67] Settles B 2012 *Active Learning, Volume 18 of Synthesis Lectures on Artificial Intelligence and Machine Learning* (San Rafael, CA: Morgan & Claypool)
[68] Ulissi Z W, Singh A R, Tsai C and Nørskov J K 2016 Automated discovery and construction of surface phase diagrams using machine learning *J. Phys. Chem. Lett.* **7** 3931–5
[69] Kolsbjerg E L, Peterson A A and Hammer B 2018 Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles *Phys. Rev.* B **97** 195424
[70] Denzel A and Kästner J 2018 Gaussian process regression for geometry optimization *J. Chem. Phys.* **148** 094114
[71] Schmitz G and Christiansen O 2018 Gaussian process regression to accelerate geometry optimizations relying on numerical differentiation *J. Chem. Phys.* **148** 241704
[72] Yoon J and Ulissi Z W 2020 Differentiable optimization for the prediction of ground state structures (DOGSS) *Phys. Rev. Lett.* **125** 173001
[73] Mortensen H L, Meldgaard S A, Bisbo M K, Christiansen M-P V and Hammer B 2020 Atomistic structure learning algorithm with surrogate energy model relaxation *Phys. Rev.* B **102** 075427
[74] Huang D, Bao J L and Tristan J-B 2022 Geometry meta-optimization *J. Chem. Phys.* **156** 134109
[75] Hao D, He X, Roitberg A E, Zhang S and Wang J 2022 Development and evaluation of geometry optimization algorithms in conjunction with ANI potentials *J. Chem. Theor. Comput.* **18** 978–91
[76] Born D and Kästner J 2021 Geometry optimization in internal coordinates based on Gaussian process regression: comparison of two approaches *J. Chem. Theor. Comput.* **17** 5955–67
[77] Stuke A, Kunkel C, Golze D, Todorović M, Margraf J T, Reuter K, Rinke P and Oberhofer H 2020 Atomic structures and orbital energies of 61,489 crystal-forming organic molecules *Sci. Data* **17** 83–92
[78] Rahaman O and Gagliardi A 2020 Deep learning total energies and orbital energies of large organic molecules using hybridization of molecular fingerprints *J. Chem. Inf. Model.* **60** 5971–83
[79] Jung H, Stocker S, Kunkel C, Oberhofer H, Han B, Reuter K and Margraf J T 2020 Size-extensive molecular machine learning with global representations *ChemSystemsChem* **2** e1900052
[80] Yaghoobi M and Alaei M 2022 Machine learning for compositional disorder: a comparison between different descriptors and machine learning frameworks *Comput. Mater. Sci.* **207** 111284
[81] Schütt K T, Glawe H, Brockherde F, Sanna A, Müller K-R and Gross E K U 2014 How to represent crystal structures for machine learning: towards fast prediction of electronic properties *Phys. Rev.* B **89** 205118
[82] Sanchez J M, Ducastelle F and Gratias D 1984 Generalized cluster description of multicomponent systems *Phys. Stat. Mech. Appl.* **128** 334–50
[83] Behler J 2011 Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations *Phys. Chem. Chem. Phys.* **13** 17930–55
[84] Jäger M O J, Morooka E V, Federici-Canova F, Himanen L and Foster A S 2018 Machine learning hydrogen adsorption on nanoclusters through structural descriptors *npj Comput. Mater.* **4** 37
[85] Himanen L, Jäger M O J, Morooka E V, Canova F F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2019 DScribe: library of descriptors for machine learning in materials science *Comput. Phys. Comm.* **247** 106949
[86] Independent personal communications by Jörg Behler, Gábor Csányi, and Ekin Doğuş Çubuk
[87] Jain A *et al* 2013 Commentary: The materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[88] Curtarolo S *et al* 2012 AFLOW: An automatic framework for high-throughput materials discovery *Comput. Mater. Sci.* **58** 218–26
[89] Draxl C and Scheffler M 2019 The NOMAD laboratory: from data sharing to artificial intelligence *J. Phys. Materials* **2** 3
[90] Rupp M 2020 `qmmlpack` (quantum mechanics machine learning package) (available at: https://gitlab.com/qmml/qmmlpack) under the Apache 2.0 license