Unified Representation of Molecules and Crystals for Machine Learning

Haoyan Huo
School of Physics, Peking University, Beijing, China
present address: Department of Materials Science and Engineering,
University of California, Berkeley, CA 94720, USA

Matthias Rupp
*Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany**

(Dated: January 3, 2018)

Accurate simulations of atomistic systems from first principles are limited by computational cost. In high-throughput settings, machine learning can potentially reduce these costs significantly by accurately interpolating between reference calculations. For this, kernel learning approaches crucially require a single Hilbert space accommodating 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 is presented for energy prediction errors below 1 kcal/mol for 7 k organic molecules and 5 meV/atom for 11 k elpasolite crystals. Applicability is demonstrated for phase diagrams of Pt-group/transition-metal binary systems.

INTRODUCTION

Computational study of atomistic systems, including molecules and crystals, requires accurate treatment of interactions at the atomic and electronic scale. Existing first-principles methods, however, are limited by their high computational cost. In high-throughput settings, machine learning (ML) [1] might significantly reduce overall costs by accurately interpolating between reference calculations. [2, 3] For this, the problem of repeatedly solving a complex equation such as Schrödinger's equation *for many related inputs* is mapped onto a nonlinear regression problem: Instead of numerically solving new systems, they are statistically estimated based on a reference set of known solutions. [4,5] This ansatz potentially enables screening larger databases of molecules and materials [2,6], running longer dynamics simulations [3], investigating larger systems [7], and even increasing the accuracy of calculations [2,8].

Kernel-based ML models [9] for fast accurate prediction of ab initio properties require a single Hilbert space of atomistic systems in which regression is carried out. Representations, that is functions mapping atomistic systems to Hilbert space elements via a kernel [10], should be [5, 11, 12] (i) *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; (ii) *unique*, that is variant against transformations changing the property, as systems with identical representation but differing in property would introduce errors [13]; (iii) *continuous*, and ideally differentiable, as discontinuities work against the smoothness assumption of the ML model; (iv) *general* in the sense of being able to encode any atomistic system, including finite and periodic systems; (v) *fast* to compute, as the goal is to reduce computational cost; (vi) *efficient* in the sense of requiring few reference calculations to reach a given target error. Constant size is an advantage. [14]

Current representations such as Coulomb matrix (CM) [4], bag of bonds (BoB) [15], smooth overlap of atomic positions (SOAP) [11], symmetry functions [16], bonding angular machine learning [17], and others [18,19] fulfill these requirements partially. The descriptors used in cheminformatics [20] often violate (ii) and (iii), 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 interpolating experimental outcomes, which have strong measurement noise and are not functions of a single conformation. Representations can encode either atoms in their chemical environment (SOAP, symmetry functions) or systems as a whole (CM, BoB). The former are more suitable for predicting local properties such as forces or chemical shifts, but require partitioning of global properties such as energies. Whole-system representations are more suitable for global properties, but require modifications to represent local environments.

We introduce a many-body tensor representation (MBTR) derived from CM/BoB and concepts of many-body expansions that fulfills above requirements, is interpretable, allows visualization (Fig. 1), and describes finite and periodic systems. State-of-the-art empirical performance is demonstrated for organic molecules and inorganic crystals, as well as applicability to phase diagrams of Pt-group / transition metal binary systems.

METHOD

We start from the CM [4, 7, 21], which represents a molecule $\mathcal{M}$ as a symmetric atom-by-atom matrix $\boldsymbol{M}_{i,j}$ with off-diagonal elements $Z_i Z_j / d_{i,j}$, where Z_i are proton 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 (i), $\boldsymbol{M}$ is either

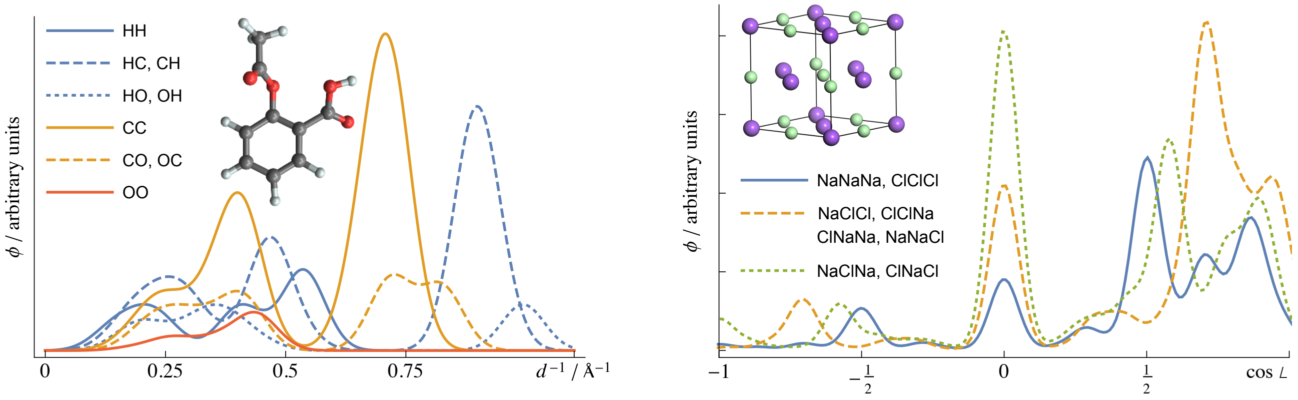

FIG. 1. *Visualization of many-body tensor representation.* Shown are distributions of inverse distances ($k = 2$, quadratic weighting) for aspirin ($C_9O_4H_8$, left), and, distributions of angles ($k = 3$, exponential weighting) for fcc salt (NaCl, right).

diagonalized, loosing information which violates (ii) [13], or sorted, causing discontinuities that violate (iii). Another shortcoming is the use of Z, which is not well suited for interpolation [22] as it overly decorrelates chemical elements from the same column of the periodic table. The related BoB [15] representation arranges the same terms differently. For each pair of chemical elements, corresponding Coulomb terms are stored in sorted order, which can be viewed as an $N_e \times N_e \times d$ tensor, where N_e is number of elements and d is sufficiently large. We retain stratification by elements, but avoid sorting by defining

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

where 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. Arranging distances on a real-space axis x removes the need for sorting. For a smoother measure, we replace Dirac's δ with another probability distribution $\mathcal{D}$, "broadening" or "smearing" it [11, 23], for which we use the normal distribution in this work. 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

$$f_2(x, z_1, z_2) = \sum_{i,j=1}^{N_a} w_2(i,j)\, \mathcal{D}\big(x, g_2(i,j)\big) C_{z_1,Z_i} C_{z_2,Z_j} \quad (2)$$

of which (1) is a special case. In general, g_2 describes a relation between atoms i and j, $\mathcal{D}$ broadens the result of g_2, and w_2 allows to weight down contributions, for example from far-away atoms. f_2 encodes two-body terms. Following recent work connecting ML with many-body expansions [17, 24], we generalize to the MBTR equation

$$f_k(x, \boldsymbol{z}) = \sum_{\boldsymbol{i}=1}^{N_a} w_k(\boldsymbol{i}) \mathcal{D}\big(x, g_k(\boldsymbol{i})\big) \prod_{j=1}^{k} C_{z_j, Z_{i_j}}, \quad (3)$$

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 $\mathcal{M}$. [25] Canonical choices of g_k for $k = 1, 2, 3, 4$ are atom counts, (inverse) distances, angles, and dihedral angles. We measure the similarity of two molecules as the Euclidean distance between their representations. In practice, we adjust (3) for symmetries. Discretizing the continuous axis results in a rank $k+1$ tensor of dimensions $N_e \times \cdots \times N_e \times N_x$ with $N_x = (x_{\max} - x_{\min})/\Delta x$. Linearizing element ranks yields $N_e^k \times N_x$ matrices, allowing for visualization (Fig. 1) and efficient numerical implementation via linear algebra routines.

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 (3) diverges. We prevent this by requiring an index of $\boldsymbol{i}$ to be in the (same) primitive unit cell. [26] This accounts for translational symmetry and prevents double-counting. Use of weighting functions w_k such as exponentially decaying weights [27] ensures convergence of the sum. Fig. 1 (right) presents the resulting distribution of angles for 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 $\frac{2}{3}\pi$.

RESULTS

To validate MBTR we demonstrate accurate predictions for properties of molecules and bulk crystals. Focusing on the representation, we employ plain kernel ridge regression models [5].

To demonstrate interpolation across *changes in the chemical structure of molecules* we utilize a benchmark dataset [21] of 7,211 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

TABLE I. *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 5 k molecules and evaluated on 2 k 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, MBTR = many-body tensor representation. Best performance in bold face.

		E / kcal mol^{-1}		α / Å^3	
Representation	Kernel	RMSE	MAE	RMSE	MAE
CM [4]	Laplacian	4.76	3.47	0.17	0.13
BoB [15]	Laplacian	2.86	1.79	0.12	0.09
BAML [17]	Laplacian	2.54	1.15	0.12	0.07
SOAP [31]	REMatch	1.61	0.92	0.07	0.05
MBTR	Linear	1.14	0.74	0.10	0.07
MBTR	Gaussian	**0.97**	**0.60**	**0.06**	**0.04**

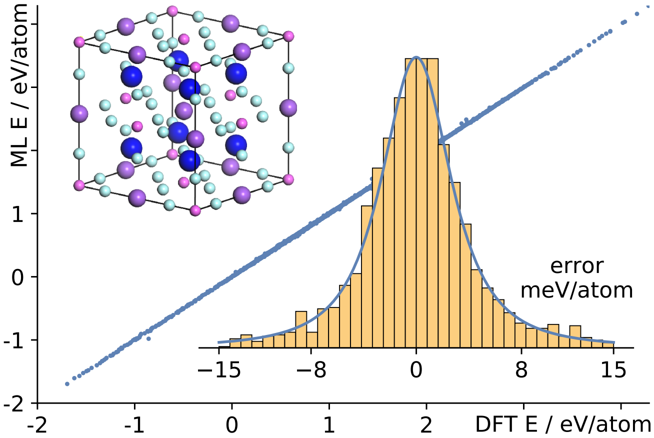

FIG. 2. *Total energy predictions for ABC_2D_6 elpasolite structures* containing 12 different elements. Shown are reference energies (DFT E) and predicted energies (ML E), as well as distribution of errors (inset) for 2 272 crystals, from an MBTR machine learning model trained on 9 086 other ones.

Perdew-Burke-Ernzerhof (PBE) [28] 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 I presents prediction errors for atomization energies and isotropic polarizabilities obtained from single point calculations with the hybrid PBE0 [29,30] functional. For 5 k training samples, prediction errors are below 1 kcal/mol ("chemical accuracy"), with the MBTR model's mean absolute error of 0.6 kcal/mol corresponding to thermal fluctuations at room temperature. Note that MBTR achieves state-of-the-art performance already with a linear regression model, allowing constant-time predictions.

Interpolation across *changes in chemistry of crystalline materials* is demonstrated for a dataset of 11 k elpasolite structures (ABC_2D_6, $AlNaK_2F_6$ prototype) [32, 33] composed of 12 different elements, with geometries and energies computed at DFT/PBE level of theory. Predicting total energies with MBTR yields an RMSE of 7.9 meV/atom and MAE of 4.1 meV/atom (Fig. 2) for a training set of 9 k 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 4 611 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) [34,35], with geometries and properties also computed using 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 (21 meV/atom, 14 meV/atom).

For interpolation of *changes in geometry*, we employ a benchmark dataset [36,37] of ab initio molecular dynamics trajectories of eight organic molecules. Each molecule was simulated at a temperature of 500 K for 150 k to 1 M time steps of 0.5 fs, with energies and forces computed at DFT/PBE level of theory and the Tkatchenko–Scheffler model [38] for van der Waals interactions. Table II presents prediction errors for MBTR, deep tensor neural networks [36] and gradient domain ML [37]. MBTR models (parametrized for dynamics data, see supplement) were trained on 10 k configurations and validated on 2 k other ones. The neural network was trained on a substantially larger set of 50 k configurations. The gradient domain ML model employs a modified CM (CM$^{\mathrm{md}}$), the Matérn kernel, and gradient of energy (forces) for training. The latter leads to kernel matrices of sizes between 27 k and 63 k. For comparison, we also show performance of CM$^{\mathrm{md}}$ using our simple models and 10 k training configurations. Non-linear MBTR regression performs best overall, with the linear kernel again being competitive.

We demonstrate *applicability* by identifying stable and meta-stable states of Pt-group/transition metal binary alloys, relevant for industrial applications. For this, we use a dataset [39] of 153 alloys computed at DFT/PBE level of theory. The task is to identify the lowest-energy compositions forming the convex hull in a phase diagram. We treat alloys separately, a challenging scenario due to small training sets of at most a few hundred structures. Due to resulting larger errors in predicted energies, their direct usage leads to wrong convex hulls. However, by employing a simple active learning [40] scheme the ML model can be used as a filter to exclude high-energy structures, saving up to 48 % of all calculations while still identifying the correct convex hull. Fig. 3 presents results for AgPt. The active learning model requested 357 DFT calculations and predicted energies of 331 (48 %) other structures, with a MAE of 39 meV/atom. The trade-off between number of saved calculations and probability of failing to identify the correct convex hull can be explicitly controlled.

TABLE II. *Energy prediction errors for changes in geometry of organic molecules.* MBTR models trained on 10 k random configurations from ab initio molecular dynamics simulations and evaluated on 2 k other ones. Shown are prediction errors for total energies in kcal/mol. MAE = mean absolute error, RMSE = root mean squared error, DTNN = deep tensor neural network [36] trained on 50 k configurations, GDML = gradient domain ML [37] trained on size 27 k–63 k kernel matrices, CM^{md} = Coulomb matrix variant, MBTR = many-body tensor representation. Best MAE in bold face.

	DTNN	GDML	CM^{md}		MBTR		MBTR	
Kernel	—	Matérn	Gaussian		linear		Gaussian	
Molecule	MAE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
benzene	**0.04**	0.07	0.05	**0.04**	0.06	0.05	0.05	**0.04**
uracil	–	0.11	0.09	0.06	0.14	0.10	0.06	**0.04**
naphthalene	–	0.12	0.16	0.12	0.15	**0.11**	0.15	**0.11**
aspirin	–	0.27	0.40	0.27	0.26	**0.18**	0.32	0.22
salicylic acid	0.50	0.12	0.15	0.11	0.17	0.12	0.11	**0.08**
malonaldehyde	0.19	0.16	0.26	0.19	0.28	0.21	0.13	**0.10**
ethanol	–	0.15	0.25	0.17	0.22	0.16	0.10	**0.07**
toluene	0.18	0.12	0.22	0.16	0.16	0.12	0.15	**0.11**

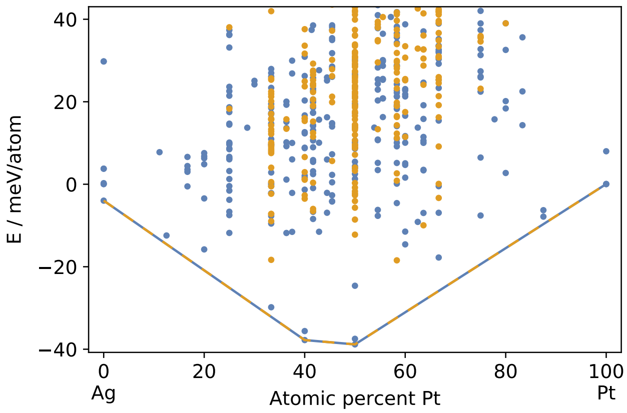

FIG. 3. *Phase diagrams of Pt-group/transition metal binary alloys.* Shown are coinciding convex hulls for Ag_xPt_{1-x} based on all DFT calculations (blue dashed) and from active ML as a filter (orange dashed). The ML model required 357 DFT calculations (blue dots), saving 331 (48 %) others (orange dots) and correctly identifying the convex hull. Low-energy structures are suggested for DFT, high-energy ones are excluded.

DISCUSSION AND OUTLOOK

MBTR is a general numerical description of atomistic systems for fast accurate interpolation between quantum-mechanical calculations via ML, based on distributions of k-atom terms stratified by elements. Despite, or because of, this simple principle it is connected to many other representations, including CM [4], BoB [15], histograms of distances, angles and dihedral angles [41], partial radial distribution functions [18], and moment tensor potentials [42], as well as cluster expansion [43].

For CM and BoB, the Laplacian kernel performs better than the Gaussian kernel. [15, 44] It has been hypothesized [45] that this is due to the Laplacian kernels better ability to deal with the discontinuities of these representations. In agreement with this, we observe that for the continuous MBTR, the Gaussian kernel consistently performs better than the Laplacian kernel (supplement).

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. [3, 46] This requires representing local chemical environments [11], which MBTR should accommodate with appropriate modifications.

A limitation of the simple ML models used here is that they were trained only on energies. For technical reasons [47, 48], differentiating such models can lead to errors in predicted forces. Including reference forces, often provided by electronic structure calculations at no additional cost, into training [3, 37, 49] can reduce number of reference calculations by an order of magnitude. [37] Training with forces would enable relaxation of structures, allowing to go from proof-of-principle presented here to applications such as virtual screening or crystal structure prediction.

We note in passing that problems in training of ML models, such as outliers, could often be traced back to problems in the underlying reference calculations, such as unconverged fast Fourier transform grids or use of different settings (violating the assumption that a single function is being fitted), a phenomenon also observed by others. [50] 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.

Advances in electronic structure codes and increasing availability of large-scale computing resources have led to big collections of ab initio calculations, such as Materials Project [51], AFLOWlib [52], Open Quantum Materials Database [35], and Novel Materials Discovery Laboratory [53]. Enabled by representations like MBTR, models combining quantum mechanics with machine learning (QM/ML) for fast accurate interpolation could be key to exploration and exploitation in such "big data" settings.

MR and HH thank Matthias Scheffler for helpful discussions and support. MR 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 Alessandro de Vita for extended discussions. MR acknowledges funding from EU Horizon 2020 program grant 676580, The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence.

Supplementary Materials

Table S1: Parametrization of MBTR for all experiments. Tables S2, S3, S4: Performance of MBTR, including results for Laplacian kernel, on small organic molecules, ABC_2D_6 and ABC_2 crystals, and dynamics of organic molecules.

Figure S1: Prediction errors of MBTR models as a function of training set size ("learning curves") for ABC_2D_6 and ABC_2 datasets, both complete and with one, two and three chemical element species removed.

The code for all computational experiments in this study is freely available as a Jupyter/IPython notebook.

* matthias.rupp@fhi-berlin.mpg.de; www.mrupp.info

[1] M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).

[2] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The δ-machine learning approach. *J. Chem. Theor. Comput.* **11**, 2087–2096 (2015).

[3] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).

[4] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).

[5] M. Rupp, Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.* **115**, 1058–1073 (2015).

[6] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

[7] M. Rupp, R. Ramakrishnan, O. A. von Lilienfeld, Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* **6**, 3309–3313 (2015).

[8] A. P. Bartók, M. J. Gillan, F. R. Manby, G. Csányi, Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **88**, 054104 (2013).

[9] B. Schölkopf, A. Smola, *Learning with Kernels* (MIT Press, Cambridge, 2002).

[10] 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.

[11] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

[12] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quant. Chem.* **115**, 1084–1093 (2015).

[13] J. E. Moussa, Comment on "Fast and accurate modeling of molecular atomization energies with machine learning". *Phys. Rev. Lett.* **109**, 059801 (2012).

[14] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, D. J. Yaron, Constant size molecular descriptors for use with machine learning. *arXiv*, 1701.06649 (2017).

[15] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).

[16] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).

[17] B. Huang, O. A. von Lilienfeld, Communication: Understanding molecular representations in machine learning: the role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).

[18] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller, E. K. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).

[19] Z. Li, J. R. Kermode, A. D. Vita, Molecular dynamics with on-the-fly machine learning of quantum mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).

[20] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors* (Wiley, Weinheim, Germany, 2009), 2nd edn.

[21] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).

[22] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).

[23] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, A. J. Liu, Identifying structural flow defects in disordered solids using machine-learning methods. *Phys. Rev. Lett.* **114**, 108001 (2015).

[24] K. Yao, J. E. Herr, J. Parkhill, The many-body expansion combined with neural networks. *J. Chem. Phys.* **146**, 014106 (2017).

[25] 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 $\mathcal{M}$.

[26] Effectively representing one unit cell, including influence of surrounding cells on it, in accordance with computed properties being reported per cell.

[27] Exponential weighting was motivated by the exponential decay of screened Coulombic interactions in solids.

[28] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

[29] J. P. Perdew, M. Ernzerhof, K. Burke, Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).

[30] C. Adamo, V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).

[31] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).

[32] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).

[33] Dataset ABC2D6-16, available at http://qmml.org.

[34] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *J. Miner. Met. Mater. Soc.* **65**, 1501–1509 (2013).

[35] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *Nat. Partn. J. Comput. Mater.* **1**, 15010 (2015).

[36] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks. *Nat. Comm.* **8**, 13890 (2017).

[37] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).

[38] A. Tkatchenko, M. Scheffler, Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).

[39] G. L. Hart, S. Curtarolo, T. B. Massalski, O. Levy, 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 (2013).

[40] B. Settles, *Active Learning*, vol. 18 of *Synthesis Lectures on Artificial Intelligence and Machine Learning* (Morgan & Claypool, 2012).

[41] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theor. Comput.*, in press (2017).

[42] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).

[43] J. M. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems. *Phys. Stat. Mech. Appl.* **128**, 334–350 (1984).

[44] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theor. Comput.* **9**, 3543–3556 (2013).

[45] G. Csányi: Consequences of Discontinuities, CECAM-Ψ_k workshop “From Many-Body Hamiltonians to Machine Learning and Back”, Berlin, Germany, May 11–13, 2015.

[46] J. Behler, Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).

[47] Differentiating an energy-based ML model can introduce noise due to small oscillations between training samples, whereas integrating a force-based model removes noise.

[48] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).

[49] A. Glielmo, P. Sollich, A. D. Vita, Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B* **95**, 214302 (2017).

[50] Independent personal communications by Jörg Behler, Gábor Csányi, and Ekin Doğuş Çubuk.

[51] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

[52] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).

[53] NOvel MAterials Discovery (NOMAD) Laboratory, a European Center of Excellence; https://nomad-coe.eu/.