nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-06169-2 OPEN

Towards exact molecular dynamics simulations with machine-learned force fields

Stefan Chmiela[1], Huziel E. Sauceda [2], Klaus-Robert Müller[1,3,4] & Alexandre Tkatchenko[5]

Molecular dynamics (MD) simulations employing classical force fields constitute the cornerstone of contemporary atomistic modeling in chemistry, biology, and materials science. However, the predictive power of these simulations is only as good as the underlying interatomic potential. Classical potentials often fail to faithfully capture key quantum effects in molecules and materials. Here we enable the direct construction of flexible molecular force fields from high-level ab initio calculations by incorporating spatial and temporal physical symmetries into a gradient-domain machine learning (sGDML) model in an automatic data-driven way. The developed sGDML approach faithfully reproduces global force fields at quantum-chemical CCSD(T) level of accuracy and allows converged molecular dynamics simulations with fully quantized electrons and nuclei. We present MD simulations, for flexible molecules with up to a few dozen atoms and provide insights into the dynamical behavior of these molecules. Our approach provides the key missing ingredient for achieving spectroscopic accuracy in molecular simulations.

[1] Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany. [2] Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany. [3] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea. [4] Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany. [5] Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg. Correspondence and requests for materials should be addressed to K.-R.M. (email: klaus-robert.mueller@tu-berlin.de) or to A.T. (email: alexandre.tkatchenko@uni.lu)

Molecular dynamics (MD) simulations within the Born-Oppenheimer (BO) approximation constitute the cornerstone of contemporary atomistic modeling. In fact, the 2013 Nobel Prize in Chemistry clearly highlighted the remarkable advances made by MD simulations in offering unprecedented insights into complex chemical and biological systems. However, one of the widely recognized and increasingly pressing issues in MD simulations is the lack of accuracy of underlying classical interatomic potentials, which hinders truly predictive modeling of dynamics and function of (bio)molecular systems. One possible solution to the accuracy problem is provided by direct ab initio molecular dynamics (AIMD) simulations, where the quantum-mechanical forces are computed on the fly for atomic configurations at every time step[1]. The majority of AIMD simulations employ the current workhorse method of electronic-structure theory, namely density-functional approximations (DFA) to the exact solution of the Schrödinger equation for a system of nuclei and electrons. Unfortunately, different DFAs yield contrasting results[2] for the structure, dynamics, and properties of molecular systems. Furthermore, DFA calculations are not systematically improvable. Alternatively, explicitly correlated methods beyond DFA could also be used in AIMD simulations, unfortunately this leads to a steep increase in the required computational resources, for example a nanosecond-long MD trajectory for a single ethanol molecule executed with the CCSD(T) method would take roughly a million CPU years on modern hardware. An alternative is a direct fit of the potential-energy surface (PES) from a large number of CCSD(T) calculations, however this is only practically achievable for rather small and rigid molecules[3–5].

To solve this accuracy and molecular size dilemma and furthermore to enable converged AIMD simulations close to the exact solution of the Schrödinger equation, here we develop an alternative approach using symmetrized gradient-domain machine learning (sGDML) to construct force fields with the accuracy of high-level ab initio calculations. Recently, a wide range of sophisticated machine learning (ML) models for small molecules and elemental materials[6–46] have been proposed for constructing PES from DFA calculations. While these results are encouraging, direct ML fitting of molecular PESs relies on the availability of large reference datasets to obtain an accurate model. Frequently, those ML models are trained on thousands or even millions of atomic configurations. This prevents the construction of ML models using high-level ab initio methods, for which energies and forces only for 100s of conformations can be practically computed.

Instead, we propose a solution that allows converged MD simulations with fully quantized electrons and nuclei for molecules with up to a few dozen atoms. This is enabled by two novel aspects: a reduction of the problem complexity through a data-driven discovery of relevant spatial and temporal physical symmetries, and enhancing the information content of data samples by exercising these identified static and dynamic symmetries, hence implicitly increasing the amount of training data. Using the proposed sGDML approach, we carry out MD simulations at the ab initio coupled cluster level of electronic-structure theory and provide insights into their dynamical behavior. Our approach contributes the key missing ingredient for achieving spectroscopic accuracy and rigorous dynamical insights in molecular simulations.

Results

Symmetrized gradient-domain machine learning. The sGDML model is built on the previously introduced gradient domain learning (GDML) model[47], but now incorporates all relevant physical symmetries, hence enabling MD simulations with high-level ab initio force field accuracy. One can classify physical symmetries of molecular systems into symmetries of space and time and specific static and dynamic symmetries of a given molecule (see Fig. 1). Global spatial symmetries include rotational and translational invariance of the energy, while homogeneity of time implies energy conservation. These global symmetries were already successfully incorporated into the GDML model[47]. Additionally, molecules possess well-defined rigid space group symmetries (i.e. reflection operation), as well as dynamic non-rigid symmetries (i.e., methyl group rotations). For example, the benzene molecule with only six carbon and six hydrogen atoms can already be indexed in $6!6! = 518400$ different, but physically equivalent ways. However, not all of these symmetric variants are accessible without crossing impassable energy barriers. Only the 24 symmetry elements in the D_{6h} point group of this molecule are relevant. While methods for identifying molecular point groups for polyatomic rigid molecules are readily available[48], Longuet-Higgins[49] has pointed out that non-rigid molecules have extra symmetries. These dynamical symmetries arise upon functional-group rotations or torsional displacements and they are usually not incorporated in traditional force fields and electronic-structure calculations. Typically, extracting nonrigid symmetries requires chemical and physical intuition about the system at hand. Here we develop a physically motivated algorithm for data-driven discovery of all relevant molecular symmetries from MD trajectories.

MD trajectories consist of smooth consecutive changes in nearly isomorphic molecular graphs. When sampling from these trajectories the combinatorial challenge is to correctly identify the same atoms across the examples such that the learning method can use consistent information for comparing two molecular conformations in its kernel function. While so-called bi-partite matching allows to locally assign atoms $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ for each pair of molecules in the training set, this strategy alone is not sufficient as it needs to be made globally consistent by multipartite matching in a second step[50–52].

We start with adjacency matrices as representation for the molecular graph[9,13,47,53,54]. To solve the pairwise matching problem we therefore seek to find the assignment τ which minimizes the squared Euclidean distance between the adjacency matrices $\mathbf{A}$ of two isomorphic graphs G and H with entries $(\mathbf{A})_{ij} = \left\|\mathbf{r}_i - \mathbf{r}_j\right\|$, where $\mathbf{P}(\tau)$ is the permutation matrix that realizes the assignment:

$$\underset{\tau}{\arg\min}\, \mathcal{L}(\tau) = \left\|\mathbf{P}(\tau)\mathbf{A}_G\mathbf{P}(\tau)^\top - \mathbf{A}_H\right\|^2. \tag{1}$$

Adjacency matrices of isomorphic graphs have identical eigenvalues and eigenvectors, only their assignment differs. Following the approach of Umeyama[55], we identify the correspondence of eigenvectors $\mathbf{U}$ by projecting both sets $\mathbf{U}_G$ and $\mathbf{U}_H$ onto each other to find the best overlap. We use the overlap matrix, after sorting eigenvalues and overcoming sign ambiguity

$$\mathbf{M} = \mathrm{abs}(\mathbf{U}_G)\mathrm{abs}(\mathbf{U}_H)^\top, \tag{2}$$

Then $-\mathbf{M}$ is provided as the cost matrix for the Hungarian algorithm[56], maximizing the overall overlap which finally returns the approximate assignment $\tilde{\tau}$ that minimizes Eq. (1) and thus provides the results of step one of the procedure. As indicated, global inconsistencies may arise, e.g., violations of the transitivity property $\tau_{jk} \circ \tau_{ij} = \tau_{ik}$ of the assignments, therefore a second step

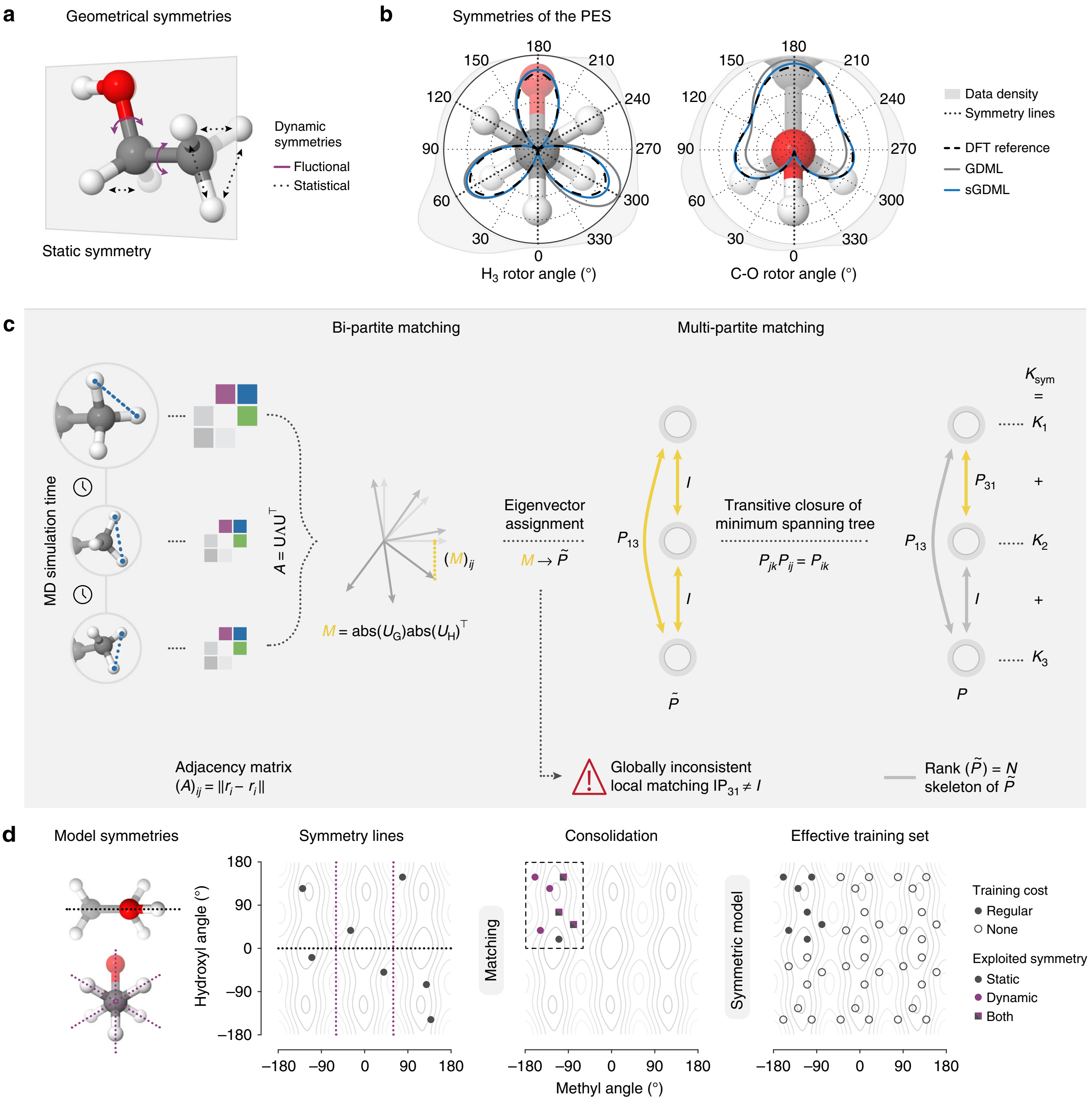

Fig. 1 Fully data-driven symmetry discovery. **a**, **b** Our multipartite matching algorithm recovers a globally consistent atom-atom assignment across the whole training set of molecular conformations, which directly enables the identification and reconstructive exploitation of relevant spatial and temporal physical symmetries of the molecular dynamics. **c** The global solution is obtained via synchronization of approximate pairwise matchings based on the assignment of adjacency matrix eigenvectors, which correspond in near isomorphic molecular graphs. We take advantage of the fact that the minimal spanning set of best bipartite assignments fully describes the multipartite matching, which is recovered via its transitive closure. Symmetries that are not relevant within the scope of the training dataset are successfully ignored. **d** This enables the efficient construction of individual kernel functions for each training molecule, reflecting the joined similarity of all its symmetric variants with another molecule. The kernel exercises the symmetries by consolidating all training examples in an arbitrary reference configuration from which they are distributed across all symmetric subdomains. This approach effectively trains the fully symmetrized dataset without incurring the additional computational cost

is necessary which is based on the composite matrix $\tilde{\mathcal{P}}$ of all pairwise assignment matrices $\tilde{\mathbf{P}}_{ij} \equiv \mathbf{P}(\tilde{\tau}_{ij})$ within the training set.

We propose to reconstruct a rank-limited $\mathcal{P}$ via the transitive closure of the minimum spanning tree (MST) that minimizes the bi-partite matching cost (see Eq. (1), Fig. 1) over the training set. The MST is constructed from the most confident bi-partite assignments and represents the rank N skeleton of $\tilde{\mathcal{P}}$, defining also $\mathcal{P}$.

The resulting consistent multipartite matching $\mathcal{P}$ enables us to construct symmetric kernel-based ML models of the form

$$\hat{f}(\mathbf{x}) = \sum_{ij}^{M} \alpha_{ij}\kappa\left(\mathbf{x}, \mathbf{P}_{ij}\mathbf{x}_i\right), \qquad (3)$$

by augmenting the training set with the symmetric variations of each molecule (see Supplementary Note 1 for a comparison with alternative symmetry-adapted kernel functions). A particular advantage of our solution is that it can fully populate all recovered permutational configurations even if they do not form a symmetric group, severely reducing the computational effort in evaluating the model. Even if we limit the range of j to include all S unique assignments only, the major downside of this approach is that a multiplication of the training set size leads to a drastic increase in the complexity of the cubically scaling kernel ridge regression learning algorithm. We overcome this drawback by exploiting the fact that the set of coefficients α for the symmetrized training set exhibits the same symmetries as the data, hence the linear system can be contracted to its original size, while still defining the full set of coefficients exactly.

For notational convenience we transform all training geometries into a canonical permutation $\mathbf{x}_i \equiv \mathbf{P}_{i1}\mathbf{x}_i$, enabling the use of uniform symmetry transformations $\mathbf{P}_j \equiv \mathbf{P}_{1j}$ (see Supplementary Note 2). Simplifying Eq. (3) accordingly, gives rise to the symmetric kernel that we originally set off to construct

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_i^M \alpha_i \sum_q^S \kappa\left(\mathbf{x}, \mathbf{P}_q\mathbf{x}_i\right)\\ &= \sum_i \alpha_i \kappa_{\mathrm{sym}}(\mathbf{x}, \mathbf{x}_i),\end{aligned} \qquad (4)$$

and yields a model with the exact same number of parameters as the original, non-symmetric one.

Our symmetric kernel is an extension to regular kernels and can be applied universally, in particular to kernel-based force fields. Here we construct a symmetric variant of the GDML model, sGDML. This symmetrized GDML force field kernel takes the form:

$$\mathrm{Hess}\left(\kappa_{\mathrm{sym}}\right)\left(\mathbf{x}, \mathbf{x}'\right) = \sum_q^S \mathrm{Hess}(\kappa)\left(\mathbf{x}, \mathbf{P}_q\mathbf{x}'\right)\mathbf{P}_q. \qquad (5)$$

Accordingly, the trained force field estimator collects the contributions of the partial derivatives $3N$ of all training points M and number of symmetry transformations S to compile the prediction for a new input $\mathbf{x}$. It takes the form

$$\hat{\mathbf{f}}_{\mathbf{F}}(\mathbf{x}) = \sum_i^M \sum_l^{3N} \sum_q^S \left(\mathbf{P}_q\alpha_i\right)_l \frac{\partial}{\partial x_l} \nabla\kappa\left(\mathbf{x}, \mathbf{P}_q\mathbf{x}_i\right) \qquad (6)$$

and a corresponding energy predictor is obtained by integrating $\hat{\mathbf{f}}_{\mathbf{F}}$ with respect to the Cartesian geometry. Due to linearity of integration, the expression for the energy predictor is identical up to second derivative operator on the kernel function.

Every (s)GDML model is trained on a set of reference examples that reflects the population of energy states a particular molecule visits during an MD simulation at a certain temperature. For our purposes, the corresponding set of geometries is subsampled from a 200 picosecond DFT MD trajectory at 500 K following the Boltzmann distribution. Subsequently, a globally consistent permutation graph is constructed that jointly assigns all geometries in the training set, providing a small selection of physically feasible transformations that define the training set specific symmetric kernel function. In the interest of computational tractability, we shortcut this sampling process to construct sGDML@CCSD(T) and only recompute energy and force labels at this higher level of theory.

The sGDML model can be trained in closed form, which is both quicker and more accurate than numerical solutions. Model selection is performed through a grid search on a suitable subset of the hyper-parameter space. Throughout, cross-validation with dedicated datasets for training, testing, and validation are used to estimate the generalization performance of the model.

Forces and energies from GDML to sGDML@DFT to sGDML@CCSD(T). Our goal is to demonstrate that it is possible to construct compact sGDML models that faithfully recover CCSD(T) force fields for flexible molecules with up to 20 atoms, by using only a small set of few hundred molecular conformations. As a first step, we investigate the gain in efficiency and accuracy of the sGDML model vs. the GDML model employing MD trajectories of ten molecules from benzene to azobenzene computed with DFT (see Fig. 2 and Supplementary Table 1). The benefit of a symmetric model is directly linked to the number of symmetries in the system. For toluene, naphthalene, aspirin, malonaldehyde, ethanol, paracetamol, and azobenzene, sGDML improves the force prediction by 31.3–67.4% using the same training sets in all cases (see Table 1). As expected, uracil and salicylic acid have no exploitable symmetries, hence the performance of sGDML is unchanged with respect to GDML. The inclusion of symmetries leads to a stronger improvement in force prediction performance compared to energy predictions. This is most clearly visible for the naphthalene dataset, where the force predictions even improve unilaterally. We attribute this to the difference in complexity of both quantities and the fact that an energy penalty is intentionally omitted in the cost function to avoid a tradeoff.

A minimal force accuracy required for reliable MD simulations is MAE = 1 kcal mol^{-1} Å^{-1}. While the GDML model can achieve

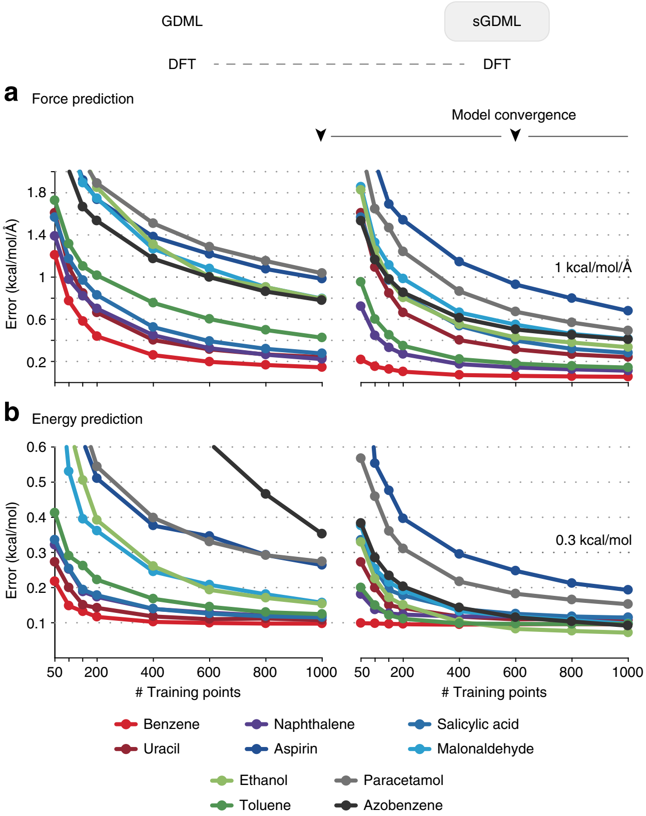

Fig. 2 Data efficiency gains using sGDML vs. GDML. Energy and force prediction accuracy (in terms of the mean absolute error (MAE)) as a function of training set size of both models trained on DFT forces: the gain in efficiency and accuracy is directly linked to the number of symmetries in the system

Table 1 Relative increase in accuracy of the sGDML@DFT vs. the non-symmetric GDML model: the benefit of a symmetric model is directly linked to the number of permutational symmetries in the system

Molecule	#Sym. in κ_{sym}	Δ MAE (%)	
		Energy	Forces
Benzene	12	−1.6	−62.3
Uracil	1	0.0	0.0
Naphthalene	4	0.0	−52.2
Aspirin	6	−29.6	−31.3
Salicylic acid	1	0.0	0.0
Malonaldehyde	4	−37.5	−48.8
Ethanol	6	−53.4	−58.2
Toluene	12	−16.7	−67.4
Paracetamol	12	−40.7	−52.9
Azobenzene	8	−74.3	−47.4

All symmetry counts include the identity transformation

this accuracy at around 800 training examples for all molecules except aspirin, sGDML only needs 200 training examples to reach the same quality. Note that energy-based ML approaches typically require two to three orders of magnitude more data[47].

Given that the novel sGDML model is data efficient and highly accurate, we are now in position to tackle CCSD(T) level of accuracy with modest computational resources. We have trained sGDML models on CCSD(T) forces for benzene, toluene, ethanol, and malonaldehyde. For the larger aspirin molecule, we used CCSD forces (see Supplementary Table 2). The sGDML@CCSD(T) model achieves a high accuracy for energies, reducing the prediction error of sGDML@DFT by a factor of 1.4 (for ethanol) to 3.4 (for toluene). This finding leads to an interesting hypothesis that sophisticated quantum-mechanical force fields are smoother and, as a convenient side effect, easier to learn. Note that the accuracy of the force prediction in both sGDML@CCSD(T) and sGDML@DFT is comparable, with the benzene molecule as the only exception. We attribute this aspect to slight shifts in the locations of the minima on the PES between DFT and CCSD(T), which means that the data sampling process for CCSD(T) can be further improved. In principle, we can envision a corrected resampling procedure for CCSD(T), using the sGDML@CCSD(T) model as future work.

MD with ab initio accuracy. The predictive power of a force field can only be truly assessed by computing dynamical and thermodynamical observables, which require sufficient sampling of the configuration space, for example by employing MD or Monte Carlo simulations. We remark that global error measures, such as mean average error (MAE) and root mean squared error are typically prone to overestimate the reconstruction quality of the force field, as they average out local topological properties. However, these local properties can become highly relevant when the model is used for an actual analysis of MD trajectories. As a demonstration, we will use the ethanol molecule; this molecule has three minima, gauche± ($M_{g\pm}$) and *trans* (M_t) shown in Fig. 3a, where experimentally it has been confirmed that M_t is the ground state and M_g is a local minimum[57]. The energy difference between these two minima is only 0.12 kcal mol^{-1} and they are separated by an energy barrier of 1.15 kcal mol^{-1}. Obviously, the widely discussed ML target accuracy of 1 kcal mol^{-1} is not sufficient to describe the dynamics of ethanol and other molecules.

This brings us to another crucial issue for predictive models: the reference data accuracy. Computing the energy difference between M_t and M_g using DFT(PBE-TS) we observe that M_g is 0.08 kcal mol^{-1} more stable than M_t, contradicting the experimental measurements. Repeating the same calculation using CCSD(T)/cc-pVTZ we find that M_t is more stable than M_g by 0.08 kcal mol^{-1}, in excellent agreement with experiment. From this analysis and subsequent MD simulations we conclude that CCSD(T) or sometimes even higher accuracy is necessary for truly predictive insights.

Additionally to requiring highly accurate quantum chemical approximations, the ethanol molecule also belongs to a category of fluxional molecules sensitive to nuclear quantum effects (NQE). This is because internal rotational barriers of the ethanol molecule ($M_g \leftrightarrow M_t$) are on the order of ~1.2 kcal mol^{-1} (see Fig. 3), which is neither low enough to generate frequent transitions nor high enough to avoid them. In a classical MD at room temperature the thermal fluctuations lead to inadequate sampling of the PES. By correctly including NQE via path-integral MD (PIMD), the ethanol molecule is able to transition between M_g and M_t configurations, radically increasing the transition frequency (see Supplementary Figure 1) and generating statistical weights in excellent agreement with experiment. Figure 3b shows the statistical occupations of the different minima for ethanol using classical MD and PIMD for the sGDML@CCSD(T) and sGDML@DFT models in comparison with the experimental results. Overall, our MD results for ethanol highlight the necessity of using a highly accurate force field with an equally accurate treatment of NQE for achieving reliable and quantitative understanding of molecular systems.

Having established the accuracy of statistical occupations of different states of ethanol, we are now in position to discuss for the first time the CCSD(T) vibrational spectrum of ethanol computed using the velocity–velocity autocorrelation function based on centroid PIMD (see Fig. 3c). As a reference, in Fig. 3c-top we compare the vibrational spectra from DFT and CCSD(T) sGDML models in the fingerprint zone, and as expected the sGDML@CCSD(T) model generates higher frequencies but both share similar shapes but slightly different peak intensities. Molecular vibrational spectra at finite temperature include anharmonic effects, hence anharmonicities can be studied by comparing the sGDML@CCSD(T) spectrum with the harmonic approximation. Figure 3c-middle shows such comparison and demonstrates that low-frequency and non-symmetric vibrations are most affected by finite-temperature contributions. The thermal frequency shift can be better seen in Fig. 3c-bottom, where the sGDML@CCSD(T) spectrum is compared at two different temperatures. We observe that each normal mode is shifted in a specific manner and not by a simple scaling factor, as typically assumed. The most striking finding from our simulations is the resolution of the apparent mismatch between theory and experiment explaining the origin of the torsional frequency for the hydroxyl group. Experimentally, the low frequency region of ethanol, around ~210 cm^{-1}, is not fully understood, but there are frequency measurements for the hydroxyl rotor ranging in between ~202[58,59] and ~207[60] cm^{-1} for gas-phase ethanol, while theoretically we found 243.7 cm^{-1} at the sGDML@CCSD(T) level of theory in the harmonic approximation. From the middle and bottom panels in Fig. 3c, we observe that by increasing the temperature the lowest peak shifts to substantially lower frequencies compared to the rest of the spectrum. The origin of such phenomena is the strong anharmonic behavior of the lowest normal mode 1, shown in Fig. 3c-middle, which mainly corresponds to hydroxyl group rotations. At room temperature the frequency of this mode drops to ~215 cm^{-1}, corresponding to a red-shift of 12% and getting closer to the experimental results, demonstrating the importance of dynamical anharmonicities.

Finally, we illustrate the wider applicability of the sGDML model to more complex molecules than ethanol by performing a

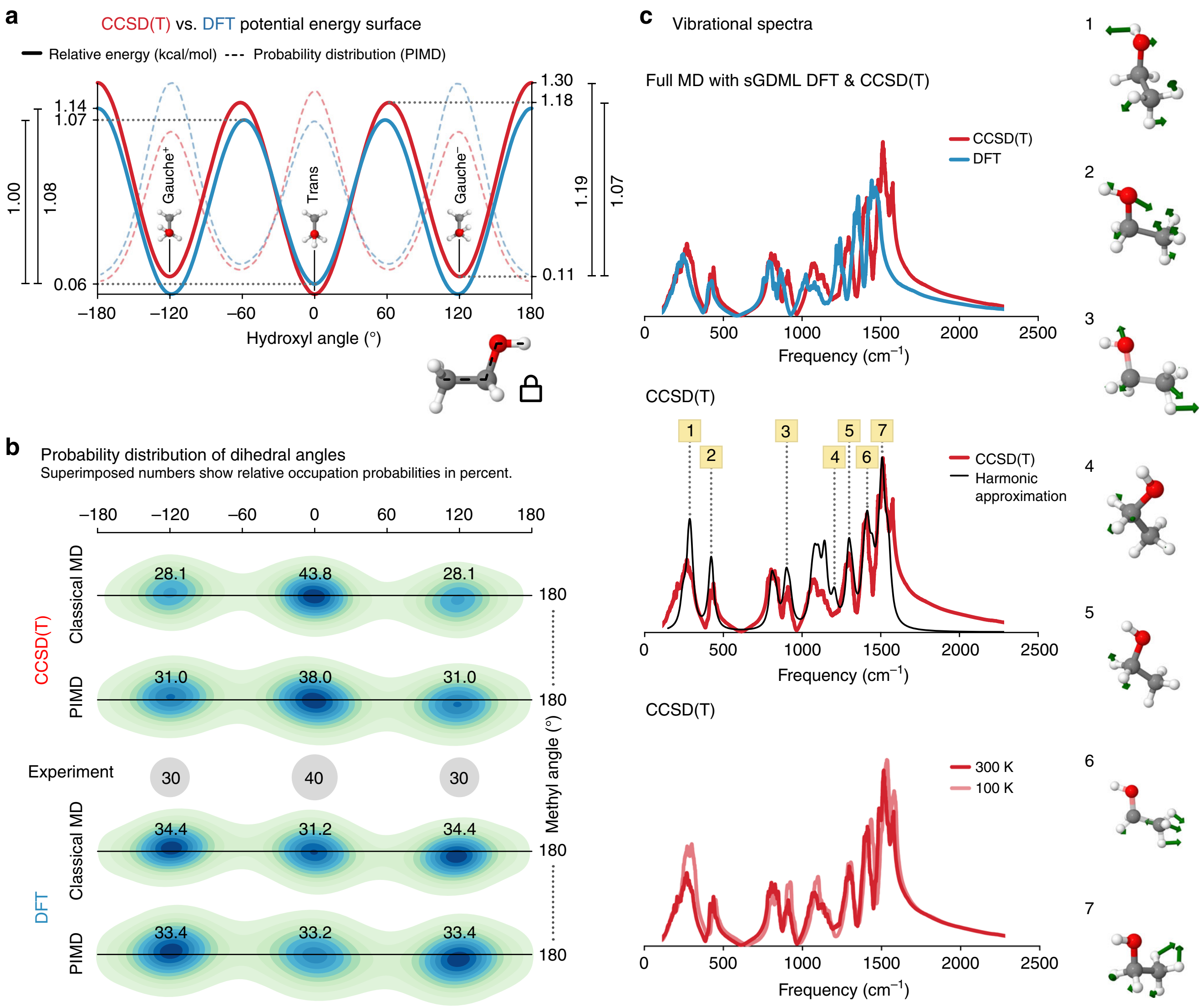

Fig. 3 Molecular dynamics simulations for ethanol. **a** Potential energy profile of the dihedral angle describing the rotation of the hydroxyl group for CCSD(T) (red) vs. DFT (blue). The energetic barriers predicted by sGDML@CCSD(T) are: $M_t \rightarrow M_g$: 1.18 kcal mol^{-1}, $M_{g-} \rightarrow M_{g+}$: 1.19 kcal mol^{-1}, and $M_g \rightarrow M_t$: 1.07 kcal mol^{-1}. The dashed lines show the probability distributions obtained from PIMD at 300 K. **b** Joint probability distribution function for the two dihedral angles of the methyl and hydroxyl functional groups. Each minimum is annotated with the occupation probability obtained from classical and path-integral MD in comparison with experimental values. **c** Analysis of vibrational spectra (velocity-velocity autocorrelation function). (top) Comparison between the vibrational spectrum obtained from PIMD simulations at 300 K for sGDML@CCSD(T) and its sGDML@DFT counterpart; (middle) comparison between the sGDML@CCSD(T) PIMD spectrum and the harmonic approximation based on CCSD(T) frequencies; (bottom) comparison of sGDML@CCSD(T) PIMD spectra at 300 and 100 K. The rightmost panel shows several characteristic normal modes of ethanol, where atomic displacements are illustrated by green arrows

detailed analysis of MD simulations for malonaldehyde and aspirin. In Fig. 4a, we show the joint probability distributions of the dihedral angles (PDDA) for the malonaldehyde molecule. This molecule has a peculiar PES with two local minima with a O···H···O symmetric interaction (structure (1)), and a shallow region where the molecule fluctuates between two symmetric global minima (structure (2)). The dynamical behavior represented in structure (2) is due to the interplay of two molecular states dominated by an intramolecular O···H interaction and a low crossing barrier of ~0.2 kcal mol^{-1}. An interesting result is the nearly unvisited structure (1) by sGDML@DFT in comparison to sGDML@CCSD(T) model regardless of the great similarities of their PES, which gives an idea of the observable consequences of subtle energy differences in the PES of molecules with several degrees of freedom. In terms of spectroscopic differences, the two approximations generate spectra with very few differences (Fig. 4a-right), but being the most prominent the one between the two peaks around 500 cm^{-1}. Such difference can be traced back to the enhanced sampling of the structure (1), and additionally it could be associated to the different nature between the methods in describing the intramolecular O···H coupling.

For aspirin, the consequences of proper inclusion of the electron correlation are even more significant. Figure 4b shows the PIMD generated PDDA for DFT and CCSD based models. By comparing the two distributions we find that sGDML@CCSD generates localized dynamics in the global energy minimum, whereas the DFT model yields a rather delocalized sampling of the PES. These two contrasting results are explained by the difference in the energetic barriers along the ester dihedral angle. The incorporation of electron correlation in CCSD increases the internal barriers by ~1 kcal mol^{-1}. This prediction was

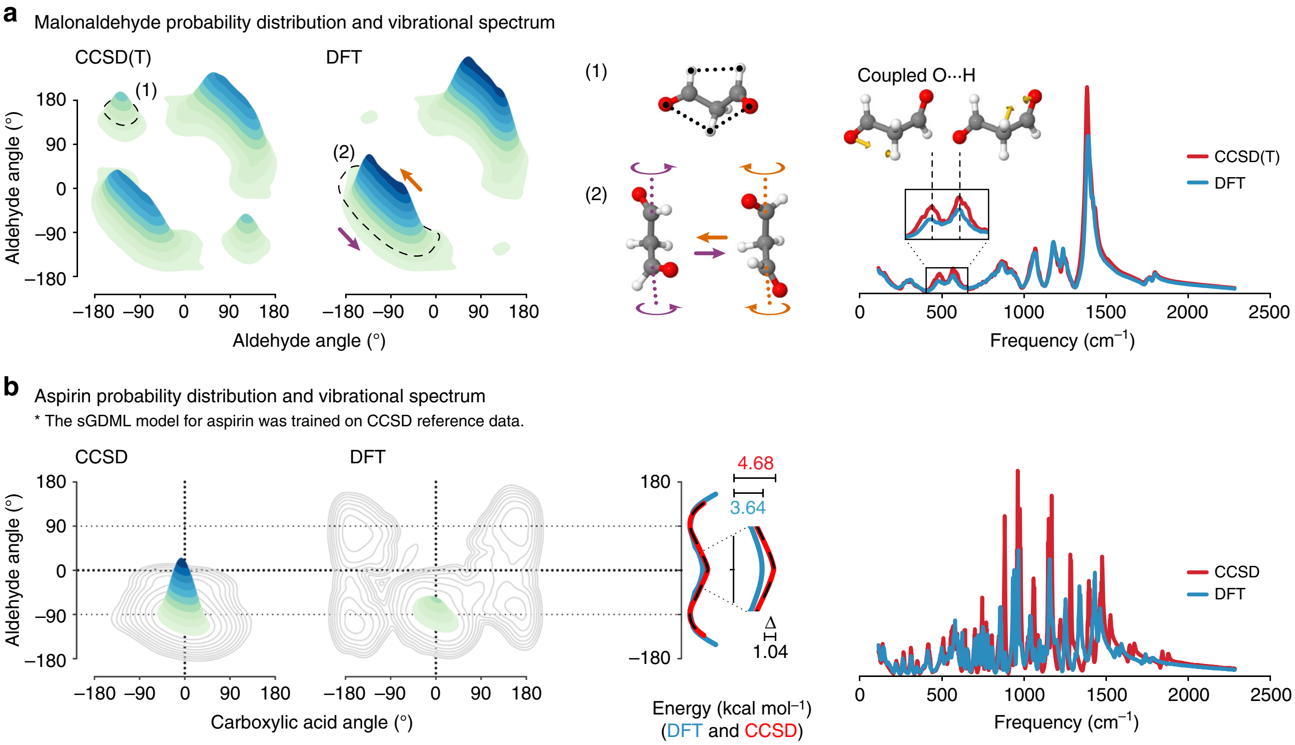

Fig. 4 Analysis of MD simulations with sGDML for malonaldehyde and aspirin. The MD simulations at 300 K were carried out for 500 ps. **a** Joint probability distributions of the dihedral angles in malonaldehyde, describing the rotation of both aldehyde groups based on classical MD simulations for sGDML@CCSD(T) and sGDML@DFT. The configurations (1) and (2) are representative structures of the most sampled regions of the PES. **b** Joint probability distributions of the dihedral angles in aspirin, describing the rotation of the ester and carboxylic acid groups based on PIMD simulations for sGDML@CCSD and sGDML@DFT using 16 beads at 300 K. The potential energy profile for the ester angle in kcal mol^{-1} is shown for sGDML@CCSD (red), sGDML@DFT (blue) and compared with the CCSD reference (black, dashed). Contour lines show the differences of both distributions on a log scale. Both panels also show a comparison of the vibrational spectra generated via the velocity-velocity autocorrelation function obtained with sGDML@CCSD (T)/CCSD (red) and sGDML@DFT (blue)

corroborated with explicit CCSD(T) calculations along the dihedral-angle coordinate (black dashed line in Fig. 4b-PES). Furthermore, the difference in the sampling is also due to the fact that the DFT model generates consistently softer interatomic interactions compared to CCSD, which leads to large and visible differences in the vibrational spectra between DFT and CCSD (Fig. 4b-right).

Discussion

The present work enables MD simulations of flexible molecules with up to a few dozen atoms with the accuracy of high-level ab initio quantum mechanics. Such simulations pave the way to computations of dynamical and thermodynamical properties of molecules with an essentially exact description of the underlying PES. On the one hand, this is a required step towards molecular simulations with spectroscopic accuracy. On the other, our accurate and efficient sGDML model leads to unprecedented insights when interpreting the experimental vibrational spectra and dynamical behavior of molecules. The contrasting demands of accuracy and efficiency are satisfied by the sGDML model through a rigorous incorporation of physical symmetries (spatial, temporal, and local symmetries) into a gradient-domain ML approach. This is a significant improvement over symmetry adaption in traditional force fields and electronic-structure calculations, where usually only (global) point groups are considered. Global symmetries are increasingly less likely to occur with growing molecule size, providing diminishing returns. Local symmetries on the other hand are system size independent and preserved even when the molecule is fragmented for large-scale modeling.

In many of the applications of machine-learned force fields the target error is the chemical accuracy or thermochemical accuracy (1 kcal mol^{-1}), but this value was conceived in the sense of thermochemical experimental measurements, such as heats of formation or ionization potentials. Consequently, the accuracy in ML models for predicting the molecular PES should not be tied to this value. Here, we propose a framework for the accuracy in force fields which satisfy the stringent demands of molecular spectroscopists, being typically in the range of wavenumbers (≈ 0.03 kcal mol^{-1}). Reaching this accuracy will be one of the greatest challenges of ML-based force fields. We remark that energy differences between molecular conformers are often on the order of 0.1–0.2 kcal mol^{-1}, hence reaching spectroscopic accuracy in molecular simulations is needed to generate predictive results.

A comparable accuracy is not obtainable with traditional force fields (see Fig. 5). In general, they miss most of the crucial quantum effects due to their rigid, handcrafted analytical form. For example, the absence of a term for electron lone pairs in AMBER leads to uncoupled rotors in ethanol. Furthermore the oversimplified harmonic description of bonded interactions generates an unphysical harmonic sampling at room temperature (see Fig. 5a). In the case of malonaldehyde (Fig. 5b), both distributions misleadingly resemble each other, however they emerge from different types of interactions. For AMBER, the dynamics are purely driven by Coulomb interactions, while the sampling with

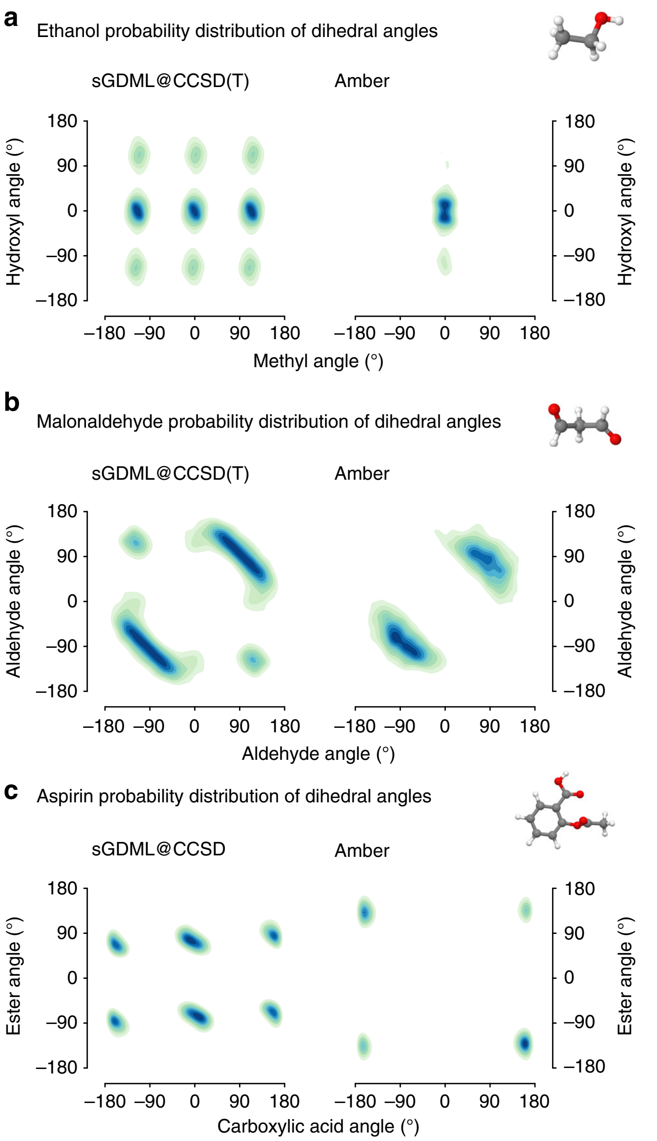

Fig. 5 Accuracy of the sGDML model in comparison to a traditional force field. We contrast the dihedral angle probability distributions of ethanol, malonaldehyde, and aspirin obtained from classical MD simulations at 300 K with sGDML (left column) vs. the AMBER[70] (right column) force field. The ethanol simulations were carried out at constant energy (NVE), whereas a constant temperature (NVT) was used for malonaldehyde and aspirin. **a** Ethanol: the coupling between the hydroxyl and methyl rotor is absent in AMBER. Moreover, the probability distribution shows an unphysical harmonic sampling at room temperature, revealing the oversimplified harmonic description of bonded interactions in that force field. **b**, **c** Malonaldehyde and aspirin: the formulation of the AMBER force field is dominated by Coulomb interactions, which can lead an incomplete description of the PES and even spurious global minima in the case of aspirin. The length of the simulations was 0.5 ns

sGDML@CCSD(T) (structure (2) in Fig. 4a) is mostly guided by electron correlation effects. Lastly, a complete mismatch between the regular force field and sGDML is evident for aspirin (see Fig. 5c), where the interactions dominated by Coulomb forces generate a completely different PES with spurious global and local minima. It is worth mentioning, that the observed shortcomings of the AMBER force field can be addressed for a particular molecule, however only at the cost of losing generality and computational efficiency.

In the context of ML, our work connects to recent studies on the usage of invariance constraints for learning and representations in vision. In the human visual system and also in computer vision algorithms the incorporation of invariances such as translation, scaling, and rotation of objects can in principle permit higher performance at more data efficiency[61]; learning theoretical bounds can furthermore show that the amount of data required is reduced by a factor: the number of parameters of the invariance transformation[62]. Interestingly, our study goes empirically beyond this factor, i.e., our gain in data efficiency is often more than two orders of magnitude when combining the invariances (physical symmetries). We speculate that our finding may indicate that the learning problem itself may become less complex, i.e., that the underlying problem structure becomes significantly easier to represent.

There is a number of challenges that remain to be solved to extend the sGDML model in terms of its applicability and scaling to larger molecular systems. Given an extensive set of individually trained sGDML models, an unseen molecule can be represented as a nonlinear combination of those models. This would allow scaling up and transferable prediction for molecules that are similar in size. Advanced sampling strategies could be employed to combine forces from different levels of theory to minimize the need for computationally intensive ab initio calculations. Our focus in this work was on intramolecular forces in small- and medium-sized molecules. Looking ahead, it is sensible to integrate the sGDML model with an accurate intermolecular force field to enable predictive simulations of condensed molecular systems (Ref.[63] presents an intermolecular model which would be particularly suited for coupling with sGDML). Many other avenues for further development exist[64], including incorporating additional physical priors, reducing dimensionality of complex PES, computing reaction pathways, and modeling infrared, Raman, and other spectroscopic measurements.

Methods

Reference data generation. The data used for training the DFT models were created running abinitio MD in the NVT ensemble using the Nosé-Hoover thermostat at 500 K during a 200 ps simulation with a resolution of 0.5 fs. We computed forces and energies using all-electrons at the generalized gradient approximation level of theory with the Perdew-Burke-Ernzerhof (PBE)[65] exchange-correlation functional, treating van der Waals interactions with the Tkatchenko-Scheffler (TS) method[66]. All calculations were performed with FHI-aims[67]. The final training data was generated by subsampling the full trajectory under preservation of the Maxwell-Boltzmann distribution for the energies.

To create the coupled cluster datasets, we reused the same geometries as for the DFT models and recomputed energies and forces using all-electron coupled cluster with single, double, and perturbative triple excitations (CCSD(T)). The Dunning's correlation-consistent basis set cc-pVTZ was used for ethanol, cc-pVDZ for toluene and malonaldehyde and CCSD/cc-pVDZ for aspirin. All calculations were performed with the Psi4[68] software suite.

Molecular dynamics. In order to incorporate the crucial effects induced by quantum nuclear delocalization, we used PIMD, which incorporates quantum-mechanical effects into MD simulations via the Feynman's path integral formalism. The PIMD simulations were performed with the sGDML model interfaced to the i-PI code[69]. The integration timestep was set to 0.2 fs to ensure energy conservation along the MD using the NVE and NVT ensemble. The total simulation time was 1 ns for ethanol (Fig. 3) to get a converged sampling of the PES using 16 beads in the PIMD.

Bipartite matching cost matrix. For the bipartite matching of a pair of molecular graphs, we solve the optimal assignment problem for the eigenvectors of their adjacency matrices using the Hungarian algorithm[56]. As input, this algorithm expects a matrix with all pairwise assignment costs $\mathbf{C}_{\mathbf{M}} = -\mathbf{M}$, which is constructed as the negative overlap matrix from Eq. (2). We add a penalty matrix with entries $(\mathbf{C}_{\mathbf{z}})_{ij} = \mathrm{abs}((\mathbf{z})_i - (\mathbf{z})_j)\varepsilon$ that prevents the matching of non-identical nuclei for sufficiently large $\varepsilon > 0$. The final const matrix is then $\mathbf{C} = \mathbf{C}_{\mathbf{M}} + \mathbf{C}_{\mathbf{z}}$.

Training sGDML. The symmetric kernel formulation approximates the similarities in the kernel matrix between different permutational configurations of the inputs, as they would appear with a fully symmetrized training set. We construct this object as the sum over all relevant atom assignments for each training geometry, such that the kernel matrix retains its original size. This procedure is used to symmetrize the GDML model[47], where the symmetric kernel function takes the form

$$\mathrm{Hess}\left(\kappa_{\mathrm{sym}}\right)\left(\mathbf{x},\mathbf{x}^{'}\right)=\frac{1}{S}\sum_{pq}^{S}\mathbf{P}_{p}^{\top}\mathrm{Hess}(\kappa)\left(\mathbf{P}_{p}\mathbf{x},\mathbf{P}_{q}\mathbf{x}^{'}\right)\mathbf{P}_{q}. \tag{7}$$

Note, that the rows and columns of the Hessian in the summand are permuted (using $\mathbf{P}_p^{\top}$ and $\mathbf{P}_q$) such that the corresponding partial derivatives align. When evaluating the model, the free variable $\mathbf{x}$ (first argument of the kernel function) is not permuted and the normalization factor is dropped (see Eq. (5). See Supplementary Note 3 for information on how to use the sGDML model, when the input is represented by a descriptor.

Data availability

All datasets used in this work are available at http://quantum-machine.org/datasets/. Additional data related to this paper may be requested from the authors.

Received: 24 March 2018 Accepted: 22 August 2018
Published online: 24 September 2018

References

1. Tuckerman, M. *Statistical Mechanics: Theory and Molecular Simulation* (Oxford University Press, Oxford, UK, 2010).
2. Koch, W. & Holthausen, M. C. *A Chemist's Guide to Density Functional Theory* (John Wiley & Sons, Hoboken, New Jersey, USA, 2015).
3. Partridge, H. & Schwenke, D. W. The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data. *J. Chem. Phys.* **106**, 4618–4639 (1997).
4. Mizukami, W., Habershon, S. & Tew, D. P. A compact and accurate semi-global potential energy surface for malonaldehyde from constrained least squares regression. *J. Chem. Phys.* **141**, 144310 (2014).
5. Schran, C., Uhl, F., Behler, J. & Marx, D. Highdimensional neural network potentials for solvation: the case of protonated water clusters in helium. *J. Chem. Phys.* **148**, 102310 (2018).
6. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
7. 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).
8. Jose, K. V. J., Artrith, N. & Behler, J. Construction of high-dimensional neural network potentials using environment-dependent atom pairs. *J. Chem. Phys.* **136**, 194111 (2012).
9. 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).
10. Montavon, G. et al. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
11. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
12. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
13. Hansen, K. et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
14. Rupp, M., Ramakrishnan, R. & von Lilienfeld, O. A. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* **6**, 3309–3313 (2015).
15. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).
16. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* **92**, 094306 (2015).
17. 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).
18. 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).
19. Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
20. De, S., Bartok, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
21. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
22. Artrith, N., Urban, A. & Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **96**, 014112 (2017).
23. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
24. Bartók, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
25. Glielmo, A., Sollich, P. & De Vita, A. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B* **95**, 214302 (2017).
26. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
27. 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).
28. Yao, K., Herr, J. E. & Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **146**, 014106 (2017).
29. Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **146**, 244108 (2017).
30. John, S. & Csányi, G. Many-body coarse-grained interactions using gaussian approximation potentials. *J. Phys. Chem. B* **121**, 10934–10949 (2017).
31. Huang, B. & von Lilienfeld, O. The "DNA" of chemistry: scalable quantum machine learning with "amons". Preprint at https://arxiv.org/abs/1707.04146 (2017).
32. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
33. Huan, T. D. et al. A universal strategy for the creation of machine learning-based atomistic force fields. *NPJ Comput. Mater.* **3**, 37 (2017).
34. Schütt, K. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **31**, 991–1001 (2017).
35. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
36. Glielmo, A., Zeni, C. & De Vita, A. Efficient nonparametric n-body force fields from machine learning. *Phys. Rev. B* **97**, 184307 (2018).
37. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
38. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, 241715 (2018).
39. 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).
40. 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).
41. Ryczko, K., Mills, K., Luchak, I., Homenick, C. & Tamblyn, I. Convolutional neural networks for atomistic systems. *Comput. Mater. Sci.* **149**, 134–142 (2018).
42. Kanamori, K. et al. Exploring a potential energy surface by machine learning for characterizing atomic transport. *Phys. Rev. B* **97**, 125124 (2018).
43. Pronobis, W., Tkatchenko, A. & Müller, K.-R. Manybody descriptors for predicting molecular properties with machine learning: analysis of pairwise and three-body interactions in molecules. *J. Chem. Theory Comput.* **14**, 2991–3003 (2018).
44. Hy, T. S., Trivedi, S., Pan, H., Anderson, B. M. & Kondor, R. Predicting molecular properties with covariant compositional networks. *J. Chem. Phys.* **148**, 241745 (2018).
45. Smith, J. S. et al. Outsmarting quantum chemistry through transfer learning. Preprint at https://chemrxiv.org/articles/Outsmarting_Quantum_Chemistry_Through_Transfer_Learning/6744440 (2018).
46. Yao, K., Herr, J. E., Toth, D. W., Mckintyre, R. & Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **9**, 2261–2269 (2018).
47. Chmiela, S. et al. Machine learning of accurate energyconserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
48. Wilson, E. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra* (McGraw-Hill Interamericana, São Paulo, Brasil, 1955).
49. Longuet-Higgins, H. The symmetry groups of non-rigid molecules. *Mol. Phys.* **6**, 445–460 (1963).

50. Pachauri, D., Kondor, R. & Singh, V. Solving the multi-way matching problem by permutation synchronization. *Adv. Neural Inf. Process. Syst.* **26**, 1860–1868 (2013)
51. Schiavinato, M., Gasparetto, A. & Torsello, A. *Transitive Assignment Kernels for Structural Classification* (Springer International Publishing, Cham, Switzerland, 2015).
52. Kriege, N. M., Giscard, P.-L. & Wilson, R. C. On valid optimal assignment kernels and applications to graph classification. *Adv. Neural Inf. Process. Syst.* **30**, 1623–1631 (2016).
53. Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R. & Borgwardt, K. M. Graph kernels. *J. Mach. Learn. Res.* **11**, 1201–1242 (2010).
54. Ferré, G., Haut, T. & Barros, K. Learning potential energy landscapes using graph kernels. *J. Chem. Phys.* **146**, 114107 (2017).
55. Umeyama, S. An eigendecomposition approach to weighted graph matching problems. *IEEE. Trans. Pattern Anal. Mach. Intell.* **10**, 695–703 (1988).
56. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist.* **2**, 83–97 (1955).
57. González, L., Mó, O. & Yáñez, M. Density functional theory study on ethanol dimers and cyclic ethanol trimers. *J. Chem. Phys.* **111**, 3855–3861 (1999).
58. Durig, J. & Larsen, R. Torsional vibrations and barriers to internal rotation for ethanol and 2, 2, 2-triuoroethanol. *J. Mol. Struct.* **238**, 195–222 (1990).
59. Wassermann, T. N. & Suhm, M. A. Ethanol monomers and dimers revisited: a Raman study of conformational preferences and argon nanocoating effects. *J. Phys. Chem. A* **114**, 8223–8233 (2010).
60. Durig, J., Bucy, W., Wurrey, C. & Carreira, L. Raman spectra of gases. XVI. Torsional transitions in ethanol and ethanethiol. *J. Phys. Chem. A* **79**, 988–993 (1975).
61. Poggio, T. & Anselmi, F. *Visual Cortex and Deep Networks: Learning Invariant Representations* (MIT Press, Cambridge, MA, 2016).
62. Anselmi, F., Rosasco, L. & Poggio, T. On invariance and selectivity in representation learning. *Inf. Inference* **5**, 134–158 (2016).
63. Bereau, T., DiStasio, R. A. Jr, Tkatchenko, A. & Von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: physicsbased potentials parametrized from machine learning. *J. Chem. Phys.* **148**, 241706 (2018).
64. De Luna, P., Wei, J., Bengio, Y., Aspuru-Guzik, A. & Sargent, E. Use machine learning to find energy materials. *Nature* **552**, 23 (2017).
65. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
66. Tkatchenko, A. & Scheffler, M. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).
67. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
68. Parrish, R. M. et al. Psi4 1.1: an open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *J. Chem. Theory Comput.* **13**, 3185–3197 (2017).
69. Ceriotti, M., More, J. & Manolopoulos, D. E. i-PI: a python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.* **185**, 1019–1026 (2014).
70. Case, D. et al. Amber 2018 (The Amber Project, 2018).

Acknowledgements

We thank Michael Gastegger for providing the AMBER force fields. S.C., A.T., and K.-R. M. thank the Deutsche Forschungsgemeinschaft (project MU 987/20-1) for funding this work. A.T. is funded by the European Research Council with ERC-CoG grant BeStMo. K.-R.M. gratefully acknowledges the BK21 program funded by the Korean National Research Foundation grant (no. 2012-005741). Part of this research was performed while the authors were visiting the Institute for Pure and Applied Mathematics, which is supported by the NSF.

Author contributions

S.C. conceived and constructed the sGDML models. S.C., H.E.S., A.T., and K.-R.M. developed the theory. H.E.S. and A.T. designed the analyses. H.E.S. performed the DFT and CCSD(T) calculations and MD simulations. S.C. and H.E.S. created the figures, with help from other authors. All authors wrote the paper, discussed the results and commented on the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-06169-2.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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