CHEMISTRY

A unified picture of the covalent bond within quantum-accurate force fields: From organic molecules to metallic complexes' reactivity

Alessandro Lunghi* and Stefano Sanvito*

Computational studies of chemical processes taking place over extended size and time scales are inaccessible by electronic structure theories and can be tackled only by atomistic models such as force fields. These have evolved over the years to describe the most diverse systems. However, as we improve the performance of a force field for a particular physical/chemical situation, we are also moving away from a unified description. Here, we demonstrate that a unified picture of the covalent bond is achievable within the framework of machine learning–based force fields. Ridge regression, together with a representation of the atomic environment in terms of bispectrum components, can be used to map a general potential energy surface for molecular systems at chemical accuracy. This protocol sets the ground for the generation of an accurate and universal class of potentials for both organic and organometallic compounds with no specific assumptions on the chemistry involved.

INTRODUCTION

The covalent bond represents the essential building block for chemistry, and its rationalization has been one of the main driving forces for the development of chemical sciences. The reorganization of electrons among atoms and the formation of molecules are complex and multifaceted processes, whose full description is only possible within the boundaries of quantum mechanics (QM). In this respect, density functional theory (DFT) represents the most common choice for routine ground-state calculations, but it becomes readily unsuitable when one needs to sample extended size and time scales. This situation is often encountered in important modern chemistry–related fields and in biological sciences, and it requires the development of computational approaches capable of capturing a restricted but essential number of chemical features in favor of a reduced computational cost.

Multiscale approaches are essential for an efficient implementation of high-throughput (*1*) and molecular docking screening frameworks (*2*). In these frameworks, the exact ground-state potential energy surface (PES) is represented in terms of simplified atomistic models called force fields (FFs). An FF consists of an analytical function of the atoms' positions, which in general depends on a number of unknown parameters to be determined. An ideal FF offers an exact representation of the quantum mechanical PES. The most crucial part in the development of an FF is the initial choice of its own mathematical form. The construction of a general FF, able to describe on an equal footing any chemical system, represents a long-standing open problem in quantum chemistry. Its solution would represent a fundamental step forward in narrowing the existing accuracy gap between a first-principles and an FF representation of a general PES.

FFs are traditionally conceived for a restricted portion of the configurational space and use simple functions, mainly inspired by the chemical understanding of the system under investigation. Such an approach, when possible, can be quite efficient. However, it generally lacks the necessary flexibility and accuracy to describe the broad spectrum of interactions that falls under the definition of covalent bond. In particular, the coordination bond represents a challenge for FFs due to its nature, which is intermediate between that of covalent directional bonds, commonly encountered in organic materials, and that of the more spherically symmetric metallic bonds. To date, a satisfactory model to describe coordination complexes is still missing (*3*), a shortfall that hinders an efficient description of highly relevant compounds and phenomena such as organometallic complexes, metalloproteins, enzymes, and catalytic reactions. Such a deficiency can be only resolved by an FF, whose analytical form is able to account for the three-dimensional and many-body nature of the chemical bond. These are features ultimately shared by any covalent bond, but they reach their maximum complexity with metallic complexes.

In recent years, machine learning frameworks have witnessed an increasing attention as possibly revolutionary computational approaches. Chemistry makes no exception to this trend (*4*). In particular, machine learning has been demonstrated applicable to the prediction of complex PESs by means of generalized regression methods such as neural networks and kernel regression (*5*–*20*). These methods share the fundamental feature of being able to represent a general continuous function with no limitations. They require an arbitrarily large number of parameters and a suitable representation of the atomic chemical environment.

In this work, we demonstrate that ridge regression–based FFs combined with a bispectrum function representation of the atomic distribution (*5*), also called the spectral neighbor analysis potentials (SNAPs) (*21*), are able to account for any fundamental feature of the covalent bond in a natural fashion and without any assumption on the chemistry of the bonds considered. The FFs predict a smooth PES and thus stable molecular dynamics.

We will first give an overview of the method used to generate and test the FF. Then, we will demonstrate its performances on model systems composed of organic and inorganic molecules comprising σ bonds, π bonds, and the most common geometries encountered in metallic complexes. The natural emergence of Jahn-Teller distortion in the model will also be shown. We will apply the method to generate three FFs for three chemical molecules of fundamental importance for materials science and biology. These are, respectively, ferrocene, a metallorganic molecule with applications in catalysis, fuel additives, and nanomedicine; dioxo-Fe^{2+}(porphyrin), the hemoglobin functional unit, regulating oxygen delivery in all vertebrates and alanine, the fundamental chemical unit for the alanine-glucose cycle that regulates the

School of Physics and CRANN Institute, Trinity College, Dublin 2, Ireland.
*Corresponding author. Email: lunghia@tcd.ie (A.L.); sanvitos@tcd.ie (S.S.)

glucose metabolism of the human body. Last, we will use the MD-17 benchmark set to compare our model with the state-of-the-art deep learning machine learning potentials.

RESULTS

PES parameterization

The representation of a quantum mechanical PES, $E_0^{QM}(\{R_i\})$, depending on the set of nuclear coordinates, $\{R_i\}$, by means of an FF, $E^{FF}(\{R_i\}, \{\alpha_j\})$, requires three fundamental steps: (i) the definition of a feature vector to describe the system geometry, (ii) the definition of the relation between this vector and the energy of the system, and (iii) the determination of the unknown coefficients, $\{\alpha_i\}$, that minimize the difference between $E_0^{QM}(\{R_i\})$ and $E^{FF}(\{R_i\}, \{\alpha_j\})$. Let us now describe how our method addresses these three requirements.

The chemical environment of a given atom contained within a radial cutoff can be conveniently described in terms of the bispectrum components $B_j(\{R_i\})$ (*5*). These functions are efficiently calculated from the sole atoms' Cartesian coordinates. The values of the coefficients associated to each term of the $B_j(\{R_i\})$ series comprise the feature vector. This represents the fingerprint of the *N*-body atomic environment, and thus, it overcomes the common assumption on the 2- or 3-body terms of the interaction. The number of functions in the series (N_{2J}) can be tuned by increasing their maximum order, 2*J*. Expanding the chemical environment over a growing number of bispectrum components improves the accuracy of the description in a variational fashion. This definition of the feature vector satisfies the fundamental symmetries of the system such as rotational and translational invariance and the swap invariance of chemically equivalent atoms.

The SNAPs used here use these functions as building blocks for the FF. The crucial assumption is that the total energy of a system containing N_{atom} atoms can be decomposed into the sum of atomic energies, $E_i(\{R_i\}, \{\alpha_j\})$, which in turn can be written as a linear function of their bispectral components, namely

$$E^{FF}(\{R_i\}, \{\alpha_j\}) = \sum_{l}^{N_{atom}} E_l(\{R_i\}, \{\alpha_j\}) = \sum_{l}^{N_{atom}} \sum_{j}^{N_{2J}} \alpha_j^l B_j^l(\{R_i\}) \quad (1)$$

There is no rigorous mathematical proof that the total energy can be written as a sum of atomic energies. However, this formulation provides a highly flexible way to represent the PES, whose accuracy can be validated with the accuracy of the predictions made. Similar to first-principles calculations, this class of FFs does not attempt to parameterize the chemical bond but rather to predict its existence as a consequence of the knowledge of the PES's shape. As a consequence, a large degree of flexibility in the FF is required. This translates to a need to use a large number of degrees of freedom, that is, a large number of coefficients α_i. Crucially, since the energy is linearly dependent on the α_is (see Eq. 1), their determination can proceed by simple least-square fit, at variance with conventional FFs, which do not depend linearly on the parameters.

Ridge regression, expressed by Eq. 2, is here used to solve the problem. The constant λ is the regularization parameter, and it has to be determined to minimize the error on the validation set while using the training set for the fitting. The introduction of a regularization term has the effect of selecting the smoothest solution among the many quasi-degenerate solutions of the simple linear least-square problem. This becomes crucial when the number of parameters to be determined is larger compared to the number of QM reference energy values.

$$\underset{\{\alpha_l\}}{\text{Min}} \left[||E_0^{QM}(\{R_i\}) - E^{FF}(\{R_i\}, \{\alpha_l\})||^2 + \lambda ||\{\alpha_l\}||^2 \right] \quad (2)$$

Equation 2 can be identically used to train the model against either conformations' total energy or atomic forces. These reference quantities can be computed by means of any quantum mechanical method able to describe a smooth PES. The DFT functional or the post Hartree-Fock method used will be chosen according to the complexity of the problem at hand. Our fitting strategy then follows a relatively canonical four-step procedure: (i) the generation of the training, validation, and test sets by quantum mechanical methods (here DFT); (ii) the fitting of the FF by solving Eq. 2 for energies and/or forces; (iii) the benchmarking of the FF on the test set, which comprises only configurations not contained in the training and validation sets (configurations that the FF has never seen before); and (iv) the enlargement of the number of configurations in the training set until the results on the test set are satisfactory.

In particular, this last step is fundamental as machine learning–based FFs are able to interpolate to a high degree of accuracy those areas of the phase space included in the training set, but they know little of chemical structures not included in the training set (*22*). This means that a low error over the training and test sets does not ensure the FF to be of good quality. One must ensure that thermal fluctuations will not bring the system in region of the phase space totally unexplored by the training set. A very powerful method to enforce the self-consistency of the training set consists of using the FF obtained after the third step to generate a molecular dynamics trajectory. Then, one can extract a number of configurations that are classified distant enough from those included in the training set (*23*, *24*) and reintroduce these in the training set itself. The newly enlarged training set is then used to construct a new FF (second step). The Gaussian metric of Eq. 3 has been chosen to determine the distance between two local environments of an atom *l*

$$d\left(||B^l(\{R_i\}) - B^l(\{R_j\})||\right) = \exp\left(-\sum_{\nu}(B_\nu^l(\{R_i\}) - B_\nu^l(\{R_j\}))^2/2\sigma_l^2\right) \quad (3)$$

where σ_l is a hyperparameter that sets the procedure selectivity.

General covalent bonds description

We now demonstrate that the SNAPs can provide an extremely accurate description of the covalent bond without relying on any assumption on the bond geometry. To achieve this goal, we have selected eight prototypical examples of covalent bond geometries, namely two simple organic molecules and six transition metal complexes: methane (CH_4), benzene (C_6H_6), $[FeCl_6]^{3-}$, $[MnCl_6]^{3-}$, $[MnCl_5]^{2-}$, $[NiCl_4]^{2-}$, $[ZnCl_4]^{2-}$, and $[VOCl_4]^{2-}$. For each of these systems, we generate 800 distorted configurations by applying random atomic displacements to the vacuum-optimized structure. The displacements are applied to every atom, and their magnitude is constrained to be smaller than 0.2 Å for 400 configurations and smaller than 0.1 Å for the remaining 400. We take 200 geometries of the training set and use them as validation and test sets. In the case of benzene and methane, we apply displacements of 0.1 and 0.05 Å, by virtue of the fact that they present smaller interatomic distances. We find that 56 coefficients for each chemical species,

corresponding to $2J = 8$, are enough to achieve sufficient accuracy (see details below). This choice results in moderate computational overheads, since a larger number of parameters will require a substantially larger training set, that is, more DFT calculations.

Figure 1 shows the quality of the SNAP description of the PES for $[FeCl_6]^{3-}$. The figure reports the mean error on the total energy for the training, validation, or test set as function of the number of configurations included in the training set, a curve called the training curve. The overall error on both the training and the test sets is outstandingly small and shows clear convergence as the number of configurations gets larger. At full convergence, one expects that the training set error will be equal to that on the test one, a condition that allows us to extrapolate an asymptotic error in the region of 0.015 kcal/mol. This is as small as the DFT error, meaning that the SNAP PES is indistinguishable from the DFT one. Figure 2 reports the error on the energy calculated for the training, validation, and test sets for all the eight molecules considered, demonstrating that extremely high accuracy can be achieved regardless of the molecule geometry. The training curves for all molecules look similar to that in Fig. 1 and are reported in the Supplementary Materials for completeness.

Note that, in this formalism, the coordination bond is naturally described without any approximation coming from the introduction of the notion of bonds and topology. For instance, Jahn-Teller distortion is here automatically included in the model and arises as consequence of the symmetry and shape of the PES. In many FFs, it is only possible to introduce Jahn-Teller effects by artificially imposing a lower symmetry on the metal's coordination shell or by explicitly introducing the d electrons in the model (*25*, *26*). Figure 3 shows the energy profile of $[FeCl_6]^{3-}$ and $[MnCl_6]^{3-}$, along a distortion going from the perfect O_h symmetry toward configurations that break the E_g d orbitals' degeneracy. As expected, the complex $[FeCl_6]^{3-}$, having a d^5 valence electronic configuration, does not show any Jahn-Teller distortion. In contrast, $[MnCl_6]^{3-}$, having four electrons in the d shell, is correctly predicted to undergo a spontaneous distortion along the selected normal modes to minimize the crystal field energy.

Hapticity and coordination: Ferrocene

As the first example of a chemically relevant complex, we choose ferrocene. Ferrocene and its derivatives have no wide-scale applications yet, but they have been intensively investigated as possible catalysts (*27*), fuel additives, and in nanomedicine (*28*). In ferrocene, the Fe^{+2} ion is sandwiched between two cyclopentadienyl (Cp) molecules, $[C_5H_5]^-$. The Fe^{+2} ion has a coordination geometry where the π electron cloud of the aromatic rings acts as a single ligand. To make the nature of this bond even more interesting and complex, there is a virtually absent rotational barrier (1.0 kcal/mol from DFT) between two different conformers: one where the two $[C_5H_5]^-$ rings are eclipsed and one where they are staggered. This is a typical situation where FFs, to be effective, need to be constructed by making specific assumptions on the bond geometry.

The same recipe described in the previous section has been used here to generate a starting training set for the isolated FeCp molecule for both the staggered and eclipsed conformations for a total of 1600 configurations. The errors on the training, validation, and test sets are 2.94, 4.41, and 3.78 kcal/mol, respectively. Furthermore, 100 configurations from a molecular dynamics run at 300 K and 12 configurations from one at 500 K are included to guarantee the stability of the structure against the high-energy fluctuations encountered in the molecular dynamics. To emphasize the ability of this approach to predict a smooth and accurate PES, we show in Fig. 4 the energy profile for the reciprocal rotation of the two Cp rings around the metal ion. The rather small energy difference between the staggered and eclipsed configurations is well reproduced with an error of only 0.3 kcal/mol. Note that configurations specifically exploring the rotation were not explicitly included in the training set, so that the knowledge of these fine details of the PES is provided by the configurations generated by molecular dynamics.

Dioxo-Fe^{2+}(porphyrin)

Next, we want to demonstrate the ability of SNAP to describe chemical reactions. This is another hard challenge for conventional FFs, which generally fail in describing bond breaking. Reactive FFs are currently available (*29*), but they typically require a large number of parameters. These FF parameters enter in the definition of the energy in a highly nonlinear form, so the construction of accurate potentials for molecules containing several atomic species is often an insurmountable task (*30*). The system chosen for this task is the Fe^{2+}(porphyrin) molecule coordinated with histidine (Hys) and molecular oxygen. This coordination

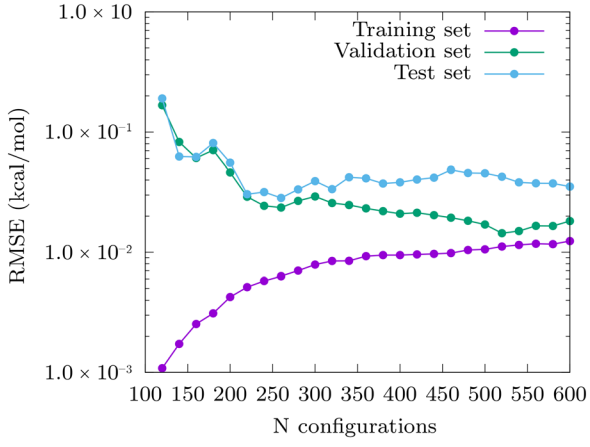

Fig. 1. FF training curve for the octahedral $[FeCl_6]^{3-}$complex. The deviation between the DFT and the FF total energy of $[FeCl_6]^{3-}$ is plotted as function of the number of configurations included in the training set. The validation and test sets are always composed of the same 200 configurations. Note that an asymptotic error of the order of 0.015 kcal/mol is achieved. RMSE, root mean square error.

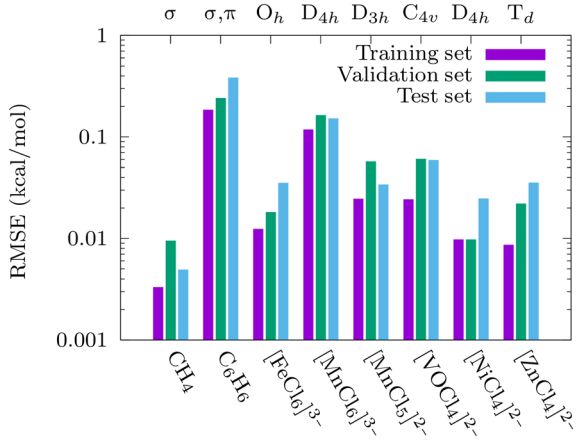

Fig. 2. SNAP performances for organic and coordination bonds with respect to DFT. For each molecule, we report the RMSE on the total energy for the training, validation, and test sets. The molecules span a broad range of bond type. The symmetry of the bond and the symmetry of the coordination environment are reported on top of the plot for the organic and inorganic molecules, respectively.

mimics the active site of the heme group. The impact of the ability to model the O_2 binding and dissociation is enormous, since it is the mechanism at play for oxygen exchange in all vertebrates.

Upon O_2 binding to the Fe^{2+}(porphyrin)(Hys) complex, one electron is transferred from the Fe^{2+} ion to O_2, and at the same time, a spin crossover transition is observed. This transforms the starting $2S + 1 = 7$ spin state into a $2S + 1 = 1$ ground state (*31*, *32*). Predicting this spin crossover phenomenon is already rather challenging at the DFT level (*33*), and a precise estimation of the energetics involved is beyond the scope of this work. Here, we restrict our study to the reaction occurring along the singlet Born-Oppenheimer surface, selected as representing ground state for dioxo-Fe^{+2}(porphyrin)(Hys) and correctly showing the charge transfer reversibility along the oxygen dissociation.

A set of 150:20:30 configurations (training:validation:test) has been generated by applying random displacements of 0.05 and 0.1 Å to the optimized structure of dioxo-Fe^{2+}(porphyrin)(Hys). To include information about the oxygen reactivity, we have also scanned the reaction pathway by performing eight DFT constraint optimizations with Fe-O distances between 1.6 and 3.0 Å. The structures optimized along the reaction path have been used to generate additional 750:120:120 configurations with random displacements of 0.05 and 0.1 Å. Last, the molecular dynamics refinement at 100, 200, and 300 K has been carried out for a total of ~400 new structures, spanning both the energy minima of the reaction and the transition state. The final model contains 448 parameters and several chemical species, but it has been successfully trained with less than ~1500 configurations. The overall errors on the training, validation, and test sets are 1.18, 1.38, and 2.42 kcal/mol, respectively. This is an accuracy comparable with DFT.

Figure 5 shows that the energetics of the reaction are perfectly reproduced. From a structural point of view, upon dissociation, the O_2 bond length goes from 1.27 to 1.20 Å, and at the same time, the Fe^{+2} ion slightly moves out of the porphyrin plane, passing from an octahedral coordination to a square pyramidal one. This is in perfect agreement with DFT results. All these fine structural rearrangements are of many-body nature, and they are well reproduced by the SNAP with great accuracy and without any a priori knowledge of the model.

The same methodology applied to predict energy changes upon conformational distortion can be applied to the prediction of any scalar molecular quantity such as the atomic charges and the spin densities. Here, we test this possibility by building a predictive model for the Fe spin density. For any DFT total energy calculation included in the training or validation set, we also perform a Mülliken population analysis to extract the spin density of the Fe ion. Then, using Eq. 1, we train a model, which expresses the Fe spin density as a function of the bispectrum components, in total analogy to the method used to predict the energy. The comparison between the DFT and SNAP predictions for the spin density upon O_2 dissociation is also reported in Fig. 5, showing the local Fe spin crossover process.

Molecular dynamics for alanine

The amino acid molecule alanine has been chosen as a prototypical organic molecule because of its ubiquitous relevance in biology. The various alanine derivatives are used in the biosynthesis of proteins and occur both in polypeptides in some bacterial cell walls and in peptide antibiotics. In mammals, alanine is key for the glucose-alanine cycle, which links the glucose production in the liver to the energy production in other tissues. Alanine presents a challenge for conventional FFs as it contains chemical species with different local environments and degrees of freedom, whose dynamics spans quite different time scales.

The starting configurations for constructing the FF are 1000 for the training set and 100 for the validation and test sets, respectively. These are constructed by applying random displacements of 0.05 and 0.1 Å to the optimized alanine structure in four different conformers. Additional 691:150:150 (training:validation:test) configurations extracted from molecular dynamics trajectory runs at temperatures between 100 and 400 K are also included in the training set to guarantee a complete sampling of the configurational space of the system. This time, because of the large number of molecular dynamics configurations included at the refinement stage, we have also extended the validation and test sets to some molecular dynamics configurations. This guarantees an unbiased estimation of the FF error. The total energy regression root mean

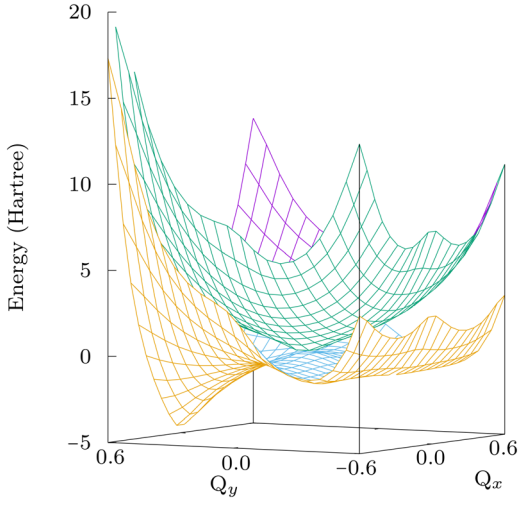

Fig. 3. FF prediction of the energy profile for Jahn-Teller–active and Jahn-Teller–inactive octahedral complexes. The concave (purple/green) and the convex (blue/yellow) surfaces represent the PESs for $[FeCl_6]^{3-}$ and $[MnCl_6]^{3-}$, respectively. The PES has been scanned in the two-dimensional space defined by the normal modes of vibration with symmetry representation E_g of the O_h group. The point $(Q_x,Q_y) = (0,0)$ corresponds to a structure optimized under the constraint of having a perfect O_h symmetry. Note that $[FeCl_6]^{3-}$ maintains a minimum at the undistorted O_h symmetry, while $[MnCl_6]^{3-}$ undergoes Jahn-Teller distortion.

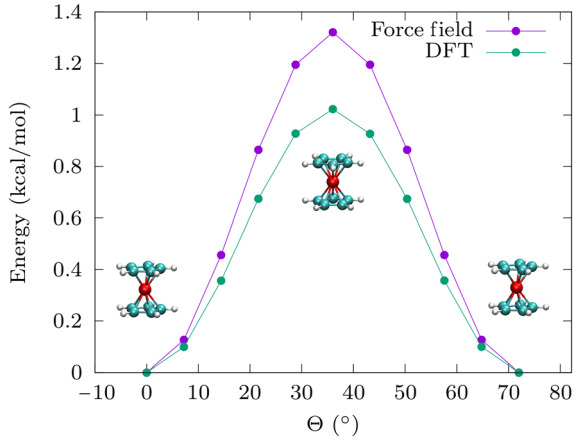

Fig. 4. Rotational barrier energy profile for the Cp rings of ferrocene. The energy profile for the complete reciprocal rotation of the two Cp rings around the Fe^{+2} ion is calculated with both DFT (green line and points) and FF (purple line and points). Note that SNAP describes the energy barrier with an accuracy of about 0.3 kcal/mol.

square error (RMSE) is only 2.34 kcal/mol for all the three configuration sets, that is similar to that obtained previously for the simpler benzene and methane molecules.

Machine learning potentials, although successfully applied to organic compounds, are still in their infancy, and they are not widely used for molecular dynamics acquisition runs. Instead, the AMBER family of FFs represents the common choice for the chemistry community involved in molecular dynamics of biological systems, and a comparison with it is crucial. Here, we compare the accuracy of SNAP against that of the general AMBER FF (GAFF) (*34*). The test is performed by comparing the FF total energy against the DFT one for 450 structures randomly sampled from molecular dynamics at 200, 300, and 400 K. These are all configurations not included in either the training or the validation set. A linear regression analysis, reported in fig. S8, shows that the SNAP is in good agreement with DFT. The RMSE is 1.82 kcal/mol, and the slope of the DFT versus FF energy curve is 1.032. In contrast, the linear regression of GAFF has a slope of 0.812, highlighting a systematic energy overestimation. This is particularly true as the size of the distortions increases, that is, as one considers largely anharmonic displacements. SNAP is able to account for anharmonic contributions to the PES, which are relevant even at relatively low temperatures. This benchmark also offers the chance to compare the computational performance of SNAP with respect to broadly used harmonic potentials. Testing the two potentials on the alanine benchmark, the SNAP potential is only ~100× more expensive than GAFF. Moreover, the time needed for a single SNAP potential energy evaluation scales linearly with the number of central processing units in parallel runs for large systems (*21*), suggesting that the extended size and time scales of common harmonic potentials are still within the reach of this machine learning FF.

Comparison with deep learning potentials

Last, we want to present a comparison between the model discussed here and the state-of-the-art machine learning potentials. In recent years, several variants of neural networks and kernel regression methods have been developed and tested on organic molecular compounds (*9*, *12*, *13*, *16*). Along with these studies, benchmark sets have been developed to assess the accuracy of the machine learning potentials. The MD-17 benchmark set contains 0.1 to 1 M conformations for eight small organic molecules, as obtained from ab initio molecular dynamics at 500 K (*9*). Figure 6 shows the accuracy of the SNAP and the SchNet potentials in predicting the conformational energies of the MD-17 benchmark set. The SchNet potential represents the state-of-the-art deep learning models (*13*). Training SNAP is rather fast, and the model reaches its optimal accuracy of 1 kcal/mol or below using only 1000 or less energy values for the training. Using training sets of such a small size, SNAP outperforms the SchNet potential. However, the latter, being more complex, is able to reach higher accuracies when trained over many more conformations or when atomic forces are used besides energies (*13*). The same consideration applies to the energy-conserving gradient-domain potentials (*9*). The possibility to reach the threshold accuracy (1 kcal/mol) using just conformational energies is an important aspect. It enables the possibility to parameterize the PESs by means of quantum chemistry methods for whom the calculation of analytical forces is not available. This is the common case of high-accuracy post Hartree-Fock methods.

DISCUSSION AND CONCLUSIONS

The search for a universal FF capable of describing any chemical system within the same formalism has been a long-standing open problem in chemistry, physics, biology, and materials science. The advent of machine learning methods raises the expectation that such an achievement is possible when a simple and chemically sound form of FF is replaced with a complex and flexible one. The gain in accuracy achieved by this complex formalism, however, comes at the price of requiring a large number of first-principles reference points to train the model. These are often more than 10^4 even for one- or two-element materials (*6*, *13*, *35*, *36*), posing severe limitations to the widespread use of machine learning FFs to systems containing several chemical species.

In this arena, SNAP has shown a very fast convergence with respect to the size of the training set even for complex systems containing several chemical species and multiple-minima PESs. The accuracy obtained with SNAP is always comparable with that of more demanding models such as neural networks or kernel regression (*9*, *12*). Moreover, we here showed that a protocol almost free of human intervention,

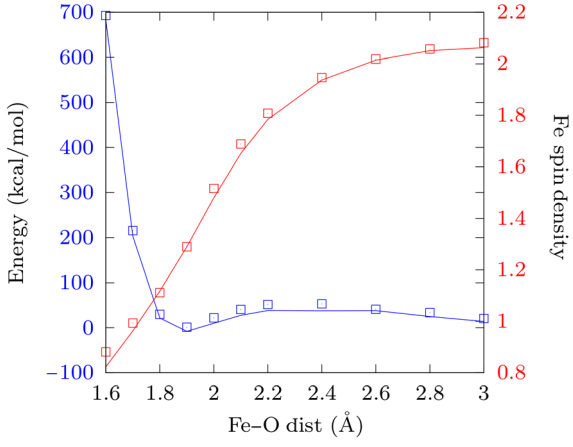

Fig. 5. Energy and Fe spin density profile for oxygen bonding reaction with Fe^{2+}(porphyrin). The lines correspond to the SNAP-predicted total energy (blue line and left-hand side scale) and Fe spin density (red line and right-hand side scale), while the symbols are for the DFT results. Here, energy and spin density are computed along the O_2 dissociation path.

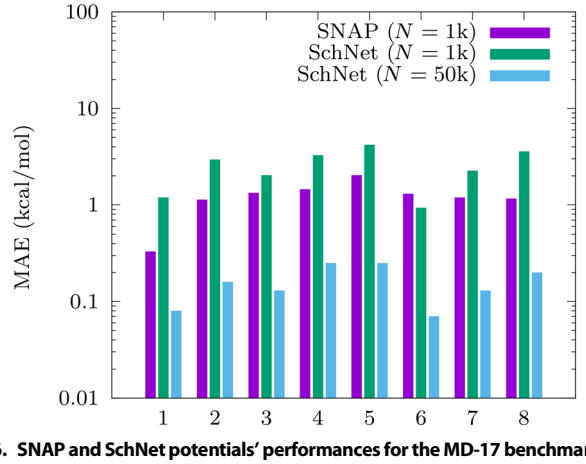

Fig. 6. SNAP and SchNet potentials' performances for the MD-17 benchmark set. The MD-17 set contains the molecules benzene (1), toluene (2), malonaldehyde (3), salicylic acid (4), aspirin (5), ethanol (6), uracil (7), and naphthalene (8). The mean absolute error (MAE) is reported for three different scenarios: the training of SNAP and that of SchNet on small-size training sets (*13*) and the training of SchNet on large training sets (*13*).

based on a random sampling of configurations near the ground state followed by a selective molecular dynamics refinement, is able to deliver an FF robust enough to withstand molecular dynamics runs. This is an essential property that must be ensured if any real application of the model is intended.

Given the great capabilities of SNAP, further development of this potential should be pursued. In this study, we derived molecule-dedicated potentials without seeking for the exportability or the universality of the SNAP parameters. A thorough study of SNAP's capabilities in describing different chemical environments with the same set of parameters and bond breaking represents the natural extension of this work. The current definition of SNAP only includes local interactions. This has been proved to be ideal to represent covalent bonds, but it needs to be merged to the description of long-range electrostatic and dispersion interactions to describe condensed-phase materials. The task can be achieved through several strategies. A possibility is to combine SNAP with point charges and Lennard-Jones parameters, as it is commonly done for organic compounds (*34*). A more refined strategy would involve the use of SNAP to predict local charges, dipoles, and dispersion coefficients and use them to compute long-range interactions on the fly (*16*). Our test, predicting the local spin density for Fe^{2+}(porphyrin), suggests that this is possible.

In conclusion, we have demonstrated that machine learning potentials, under the SNAP framework, make it possible to describe covalent bonds and their reactivity, as occurring in both organic molecules and coordination compounds, under a unified picture and with an accuracy comparable to first-principles calculations. To the best of our knowledge, this is the first time that the universality of an FF has been demonstrated upon this variety of molecular geometries, including subtle features such as Jahn-Teller distortions and many-body effects. In particular, while the application of neural networks to a general class of organic compounds has been recently demonstrated (*12*, *16*), the possibility to accurately describe coordination compounds has been proved here for the first time. The training of the model is particularly simple and automatized and only requires a limited number of quantum chemistry calculations. These properties, together with the fact that SNAP potentials are already implemented in high-performance simulation packages (*37*), make this approach extremely appealing for a fast development of unified classes of FFs for both biological and materials sciences.

MATERIALS AND METHODS

DFT calculations

All the DFT calculations were performed with the ORCA software (*38*). The dioxo-Fe^{+2}(porphyrin)(Hys) was simulated with the B3LYP functional and the def2-TZVP basis set for all the elements and the RIJCOSX approximation. All other systems were simulated with the Perdew-Burke-Ernzerhof (PBE) functional, including Grimme's D3 van der Waals corrections (*39*, *40*), basis set def2-TZVP and by applying the resolution of identity (RI) approximation.

SNAP fitting and simulations

The fitting and refinement of the SNAP potentials were done through the Fortran code fitsnap. Such a code uses large-scale atomic/molecular massively parallel simulator (LAMMPS) (*37*) as an external library to generate the bispectrum components for all atoms and to accordingly calculate molecular dynamics runs. In all cases, the order $2J = 8$ for the bispectrum components, corresponding to 56 elements for each atomic kind, was used. Thus, except for Fe^{2+}(porphyrin), we defined the atomic kinds as identical to the chemical elements, even when the same element appears in a different chemical environment. The dioxo-Fe^{+2}(porphyrin)(Hys) system was simulated with a redundant number of kinds. In particular, chemically inequivalent elements in the porphyrin ring and the hystidine moiety were defined as different kinds to give more flexibility to the model. The radial cutoff used to build the bispectrum components was optimized to minimize the overall error on the training and validation set. The radial cutoff for hydrogen was always reduced by a factor of 0.6 with respect to the other elements. The definition of bispectrum components gives the possibility to differentiate different atomic kinds with weights (*21*). In this work, we set all the weights to unity.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/5/eaaw2210/DC1

Fig. S1. FF training curve for methane.
Fig. S2. FF training curve for benzene.
Fig. S3. FF training curve for the distorted octahedral $[MnCl_6]^{3-}$ complex.
Fig. S4. FF training curve for the trigonal bipyramidal $[MnCl_5]^{2-}$ complex.
Fig. S5. FF training curve for the square pyramidal $[VOCl_4]^{2-}$ complex.
Fig. S6. FF training curve for the square planar $[NiCl_4]^{2-}$ complex.
Fig. S7. FF training curve for the tetrahedral $[ZnCl_4]^{2-}$ complex.
Fig. S8. Alanine molecular dynamics blind test.

REFERENCES AND NOTES

1. S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
2. D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
3. P. Li, K. M. Merz Jr., Metal ion modeling using classical mechanics. *Chem. Rev.* **117**, 1564–1686 (2017).
4. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
5. 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).
6. G. C. Sosso, G. Miceli, S. Caravati, J. Behler, M. Bernasconi, Neural network interatomic potential for the phase change material GeTe. *Phys. Rev. B* **85**, 174103 (2012).
7. J. Behler, Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **115**, 1032–1050 (2015).
8. A. P. Bartók, G. Csányi, Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).
9. S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2016).
10. A. P. Bartok, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
11. J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chemie Int. Ed.* **56**, 12828–12840 (2017).
12. J. S. Smith, O. Isayev, A. E. Roitberg, ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
13. K. T. Shutt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
14. T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, A universal strategy for the creation of machine learning-based atomistic force fields. *npj Comput. Mater.* **3**, 37 (2017).
15. L. Zhang, J. Han, R. Car, Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
16. K. Yao, J. E. Herr, D. W. Toth, R. McKintyre, J. Parkhill, The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics. *Chem. Sci.* **9**, 2261–2269 (2018).

17. A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Phys. Rev. Lett.* **120**, 36002 (2018).
18. S. Chmiela, H. E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
19. A. Glielmo, C. Zeni, A. De Vita, Efficient nonparametric *n*-body force fields from machine learning. *Phys. Rev. B* **97**, 184307 (2018).
20. M. Rupp, O. A. V. von Lilienfeld, K. Burke, Guest editorial: Special topic on data-enabled theoretical chemistry. *J. Chem. Phys.* **148**, 241401 (2018).
21. A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2014).
22. J. Behler, Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
23. V. Botu, R. Ramprasad, Adaptive machine learning framework to accelerateab initiomolecular dynamics. *Int. J. Quantum Chem.* **115**, 1074–1083 (2015).
24. Z. Li, J. R. Kermode, A. De Vita, Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
25. Q. Liao, S. C. L. Kamerlin, B. Strodel, development and application of a nonbonded Cu^{2+} model that includes the Jahn–Teller effect. *J. Phys. Chem. Lett.* **6**, 2657–2662 (2015).
26. J. Y. Xiang, J. W. Ponder, An angular overlap model for Cu(II) Ion in the AMOEBA polarizable force field. *J. Chem. Theory Comput.* **10**, 298–311 (2013).
27. A. Togni, *Ferrocenes: Homogeneous Catalysis, Organic Synthesis, Materials Science* (John Wiley & Sons, 2008).
28. D. Astruc, Why is ferrocene so exceptional? *Eur. J. Inorg. Chem.* **2017**, 6–29 (2017).
29. T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, A. C. T. van Duin, The ReaxFF reactive force-field: Development, applications and future directions. *npj Comput. Mater.* **2**, 15011 (2016).
30. M. Dittner, J. Müller, H. M. Aktulga, B. Hartke, Efficient global optimization of reactive force-field parameters. *J. Comput. Chem.* **36**, 1550–1561 (2015).
31. J. Ribas-Ariño, J. J. Novoa, The mechanism for the reversible oxygen addition to heme. A theoretical CASPT2 study. *Chem. Commun.* **14**, 3160–3162 (2007).
32. M. E. Ali, B. Sanyal, P. M. Oppeneer, Electronic structure, spin-states, and spin-crossover reaction of heme-related Fe-porphyrins: A theoretical perspective. *J. Phys. Chem. B* **116**, 5849–5859 (2012).
33. D. A. Scherlis, M. Cococcioni, P. Sit, N. Marzari, Simulation of heme using DFT + U: A step toward accurate spin-state energetics. *J. Phys. Chem. B* **111**, 7384–7391 (2007).
34. J. Wang, R. R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
35. R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, M. Parrinello, Graphite-diamond phase coexistence study employing a neural-network mapping of theab initio potential energy surface. *Phys. Rev. B* **81**, 100103 (2010).
36. S. Chiriki, S. Jindal, S. S. Bulusu, Neural network potentials for dynamics and thermodynamics of gold nanoparticles. *J. Chem. Phys.* **146**, 084314 (2017).
37. S. Plimpton, Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
38. F. Neese, The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).
39. J. P. J. Perdew, K. Burke, Y. Wang, Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Phys. Rev. B* **54**, 16533 (1996).
40. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).

Acknowledgments: We acknowledge G. Sosso, E. Bosoni, and J. Nelson for the useful discussions. **Funding:** This work was sponsored by the Science Foundation Ireland (grant 14/IA/2624). Computational resources were provided by the Trinity Centre for High Performance Computing (TCHPC) and the Irish Centre for High-End Computing (ICHEC). **Author contributions:** All the authors contributed to the discussion of the results and to the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The code used to perform the SNAP fittings is available at https://github.com/lunghiale/fitsnap. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 29 November 2018
Accepted 17 April 2019
Published 31 May 2019
10.1126/sciadv.aaw2210

Citation: A. Lunghi, S. Sanvito, A unified picture of the covalent bond within quantum-accurate force fields: From organic molecules to metallic complexes' reactivity. *Sci. Adv.* **5**, eaaw2210 (2019).

A unified picture of the covalent bond within quantum-accurate force fields: From organic molecules to metallic complexes' reactivity

Alessandro Lunghi and Stefano Sanvito

Sci Adv **5** (5), eaaw2210.
DOI: 10.1126/sciadv.aaw2210