Accurate molecular polarizabilities with coupled cluster theory and machine learning

David M. Wilkins[a], Andrea Grisafi[a], Yang Yang[b], Ka Un Lao[b], Robert A. DiStasio Jr.[b,1], and Michele Ceriotti[a,1]

[a]Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; and [b]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853

Edited by Michael L. Klein, Institute of Computational Molecular Science, Temple University, Philadelphia, PA, and approved December 26, 2018 (received for review September 23, 2018)

The molecular dipole polarizability describes the tendency of a molecule to change its dipole moment in response to an applied electric field. This quantity governs key intra- and intermolecular interactions, such as induction and dispersion; plays a vital role in determining the spectroscopic signatures of molecules; and is an essential ingredient in polarizable force fields. Compared with other ground-state properties, an accurate prediction of the molecular polarizability is considerably more difficult, as this response quantity is quite sensitive to the underlying electronic structure description. In this work, we present highly accurate quantum mechanical calculations of the static dipole polarizability tensors of 7,211 small organic molecules computed using linear response coupled cluster singles and doubles theory (LR-CCSD). Using a symmetry-adapted machine-learning approach, we demonstrate that it is possible to predict the LR-CCSD molecular polarizabilities of these small molecules with an error that is an order of magnitude smaller than that of hybrid density functional theory (DFT) at a negligible computational cost. The resultant model is robust and transferable, yielding molecular polarizabilities for a diverse set of 52 larger molecules (including challenging conjugated systems, carbohydrates, small drugs, amino acids, nucleobases, and hydrocarbon isomers) at an accuracy that exceeds that of hybrid DFT. The atom-centered decomposition implicit in our machine-learning approach offers some insight into the shortcomings of DFT in the prediction of this fundamental quantity of interest.

dipole polarizability | machine learning | coupled cluster theory | density functional theory | Gaussian process regression

The last decade has seen great progress in the first principles evaluation of the structures, stabilities, and properties of molecules and materials. Kohn–Sham density functional theory (DFT) has played a pivotal role in this endeavor by providing ground-state properties with an accuracy that is sufficient for many useful applications at a manageable computational cost (1–3). However, DFT is not equally accurate for every property of interest. For instance, an accurate and reliable description of the molecular dipole polarizability $\boldsymbol{\alpha}$, a tensor that describes how the molecular dipole changes in the presence of an applied electric field $\mathbf{E}$, can be quite difficult to obtain (4). This is primarily due to the fact that $\boldsymbol{\alpha}$ is a response property that is particularly sensitive to the quantum mechanical description of the underlying electronic structure. As such, nontrivial electron correlation effects and basis set incompleteness error must be simultaneously accounted for when determining $\boldsymbol{\alpha}$. For these reasons and in light of the fact that $\boldsymbol{\alpha}$ is a fundamental quantity of interest that underlies induction and dispersion interactions (5–7) and Raman and sum frequency generation spectroscopy (8–11), and represents a key ingredient in the development of next generation polarizable force fields (12–16), it is important to provide benchmark values for $\boldsymbol{\alpha}$ beyond the accuracy of DFT. In this regard, linear response coupled cluster singles and doubles theory (LR-CCSD) (17–19) has been shown to provide considerably more accurate and reliable predictions for $\boldsymbol{\alpha}$ when used in conjunction with a sufficiently large (diffuse) one-particle basis set (20–23). However, such a prediction is accompanied by a substantially larger computational cost (scaling with the sixth power of the system size), which can become quite prohibitive even when treating molecules with as few as $10-15$ atoms.

In the last few years, machine learning (ML) has gained traction as an alternative approach to the prediction of molecular properties, substituting or complementing electronic structure methods (24–26). In particular, it has been shown that accuracy on par with (or even better than) DFT can be achieved in the prediction of many molecular properties (27, 28) and that DFT (29) or coupled cluster (30) accuracy can be reached more easily when using a less accurate but more computationally efficient electronic structure method as a stepping stone. The polarizability, however, poses an additional challenge to ML. Due to its tensorial nature, the predicted $\boldsymbol{\alpha}$ must transform according to the symmetries of the $SO(3)$ rotation group. For rigid molecules, this is easily achieved by learning the components of the tensor written in the reference frame of the molecule (31, 32). However, to obtain a transferable model that would also be suitable for flexible molecules—as well as different compounds—this line of thought would require a cumbersome and inelegant fragment decomposition. To avoid these complications, a symmetry-adapted Gaussian process regression

Significance

The dipole polarizability of molecules and materials is central to several physical phenomena, modeling techniques, and the interpretation of many experiments. Its accurate evaluation from first principles requires quantum chemistry methods that are often too demanding for routine use. The highly accurate calculations reported herein provide a much-needed benchmark of the accuracy of hybrid density functional theory (DFT) as well as training data for a machine-learning model that can predict the polarizability tensor with an error that is about 50% smaller than DFT. This framework provides an accurate, inexpensive, and transferable strategy for estimating the polarizabilities of molecules containing dozens of atoms, and therefore removes a considerable obstacle to accurate and reliable atomistic-based modeling of matter.

Author contributions: D.M.W., A.G., Y.Y., K.U.L., R.A.D., and M.C. designed research; D.M.W., A.G., Y.Y., and K.U.L. performed research; D.M.W., A.G., Y.Y., K.U.L., R.A.D., and M.C. analyzed data; and D.M.W., A.G., Y.Y., K.U.L., R.A.D., and M.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Coordinates and dipole polarizabilities of the QM7b database, the showcase dataset, and the alkenes and acenes have been deposited in the Materials Cloud Archive (https://archive.materialscloud.org/2019.0002/v1).

[1]To whom correspondence may be addressed. Email: distasio@cornell.edu or michele.ceriotti@epfl.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1816132116/-/DCSupplemental.

Published online February 7, 2019.

(SA-GPR) scheme has recently been derived to naturally incorporate this $SO(3)$ covariance into an ML scheme that is suitable to predict tensorial quantities of arbitrary order (33). In this paper, we present comprehensive coupled cluster-level benchmarks for the polarizabilities of the $\sim 7,000$ small organic molecules contained in the QM7b database (34). We use these reference calculations to assess the accuracy of different hybrid DFT schemes and train an SA-GPR–based ML scheme (AlphaML) that can accurately predict the polarizability tensor with nominal computational cost. We then test the extrapolative prediction capabilities of AlphaML on a showcase dataset composed of 52 larger molecules and demonstrate that this approach provides a viable alternative to state-of-the-art and computationally prohibitive electronic structure methods for predicting molecular polarizabilities.

Results

Electronic Structure Calculations. The QM7b database (26, 34, 35) includes $N = 7{,}211$ molecules containing up to seven "heavy" atoms (i.e., C, N, O, S, Cl) with varying levels of H saturation. This dataset is based on a systematic enumeration of small organic compounds (35) and contains a rich diversity of chemical groups, making it a challenging test of the accuracy associated with DFT and quantum chemical methodologies. DFT-based molecular polarizabilities were obtained by (numerical) differentiation of the molecular dipole moment, $\boldsymbol{\mu}$, with respect to an external electric field $\mathbf{E}$ using the hybrid B3LYP (36, 37) and SCAN0 (38) functionals. Reference molecular polarizabilities were obtained using LR-CCSD. To account for basis set incompleteness error, which can be even more important than higher-order electron correlation effects in an accurate and reliable determination of $\boldsymbol{\alpha}$ (21–23, 39), we used the d-aug-cc-pVDZ basis set (39) for all calculations herein. Although this double-ζ basis set has only a moderate number of polarization functions, augmentation with an additional set of diffuse functions almost always increases the convergence of $\boldsymbol{\alpha}$ with respect to aug-cc-pVDZ (21, 22, 39–41). The alternative choice of retaining a single set of diffuse functions and simply increasing the angular momentum by using the slightly larger aug-cc-pVTZ basis set yields $\boldsymbol{\alpha}$ values of comparable quality to d-aug-cc-pVDZ (*SI Appendix*) (21, 22, 39–41), albeit with a significant increase in the computational effort required to treat the entire QM7b dataset. A more detailed description of the electronic structure calculations performed in this work is in *Materials and Methods*.

To enable comparisons between molecules of different sizes, all error estimates (explicit expressions for which are given in *Materials and Methods*) are computed based on molecular polarizabilities divided by the number of atoms, n_i, contained within a given molecule. On the QM7b database, the popular B3LYP hybrid DFT functional predicts $\boldsymbol{\alpha}$ with a mean signed error (MSE) of 0.259 a.u., a mean absolute error (MAE) of 0.302 a.u., and a root mean square error (RMSE) of 0.404 a.u. with respect to the reference LR-CCSD values. These errors, which include both scalar and anisotropic contributions, are quite substantial and correspond to 18.3% of the intrinsic variability within the QM7b database, defined as $\sigma_{\mathrm{CCSD}} = [\frac{1}{N}\sum_i \|\boldsymbol{\alpha}_i^{(\mathrm{CCSD})} - \langle\boldsymbol{\alpha}^{(\mathrm{CCSD})}\rangle\|_F^2/n_i^2]^{1/2}$. The large MSE value obtained with B3LYP indicates a systematic overestimation of $\boldsymbol{\alpha}$ by this functional (4, 42); results from the SCAN0 hybrid functional show a substantially reduced MSE of 0.059 a.u. Despite the smaller systematic overestimation of $\boldsymbol{\alpha}$ in comparison with B3LYP, the statistical errors obtained with SCAN0 are still quite large, with computed MAE (RMSE) values of 0.217 (0.316) a.u. From the ML point of view, the AlphaML model presented herein performs almost equally well for B3LYP and SCAN0. For this reason, we focus our discussion on the B3LYP and LR-CCSD results, which will be referred to as DFT and coupled cluster singles and doubles theory (CCSD), respectively, throughout the remainder of the manuscript.

Improved SA-GPR. The formalism underlying the SA-GPR scheme in general and the λ-SOAP (smooth overlap of atomic positions) descriptors on which our model is based have been introduced elsewhere (33) and are summarized in *Materials and Methods*. In this work, we include several substantial improvements that increase the accuracy and speed of the SA-GPR model, and these are worth a separate discussion. For one, evaluation of the λ-SOAP representation is greatly accelerated by choosing the most significant few hundred spherical harmonic components (of several tens of thousands) using farthest point sampling (FPS) (43). The calculation of the kernel in Eq. **1** can be carried out with essentially the same result as if all components were retained but with a much lower computational cost. A second improvement is the generalization of the λ-SOAP kernels beyond the linear kernels used in ref. 33. It has been shown that, in many cases, taking an integer power of the scalar SOAP kernel improves the performance of the associated ML model. This can be understood in terms of the order (two body, three body,...) of the interatomic correlations that are described by different kernels (44, 45). In the tensorial case, one should be careful, as the linear nature of the kernel is essential to ensure the correct covariant behavior. To include nonlinearity and increase the order of the model without affecting the symmetry properties, we multiplied the $\lambda > 0$ kernels by the scalar $\lambda = 0$ kernel raised to the power of $\zeta - 1$ as in Eq. **2**. Finally, we combined multiple kernels computed with different environment radii, r_{c}, which have been shown to be beneficial in the scalar case (30). Together, these improvements halve the error on QM7b as discussed in detail in *SI Appendix*.

Learning on the QM7b Database. These highly accurate reference CCSD calculations and the SA-GPR scheme lay the foundation for a transferable model to predict molecular polarizabilities. In this first incarnation of the AlphaML model, we use the reference DFT and CCSD calculations on the QM7b set for training (34). As a first verification of its performance, we computed learning curves for the DFT and CCSD polarizabilities of the QM7b dataset. We used up to 5,400 structures for training with subsequent assessment of the accuracy and reliability of the AlphaML model in the prediction of $\boldsymbol{\alpha}$ for the 1,811 structures that were not included in the training set. The structures were added to the training set according to their FPS order (43) (i.e., starting from the most diverse configurations). This procedure is representative of an efficient learning strategy that aims to obtain uniform accuracy with the minimum number of reference calculations (30). Using the best kernel hyperparameters (as described in *SI Appendix*), we trained a model to learn the CCSD polarizabilities. We report ML errors in terms of the percentage of the intrinsic variability of the CCSD dataset ($\sigma_{\mathrm{CCSD}} = 2.216$ a.u. per atom) so as to provide a direct measure of the learning performance. As illustrated by the learning curves in Fig. 1, using up to 75% of the QM7b database for training yields a 2.5% RMSE with respect to σ_{CCSD} in predicting CCSD polarizabilities.

To get a clearer idea of the accuracy associated with these ML-based predictions, one can compare these values against hybrid DFT. Using the same metric, the intrinsic error of DFT is 18% of σ_{CCSD} in the prediction of CCSD polarizabilities. This demonstrates that an ML model based on SA-GPR can yield polarizabilities with an accuracy that is approximately one order of magnitude greater than DFT. At the same time, the corresponding DFT polarizabilities can be learned with an error of 3.2% of σ_{CCSD}. As seen in other cases (29, 30), highly accurate quantum chemistry calculations are smoother

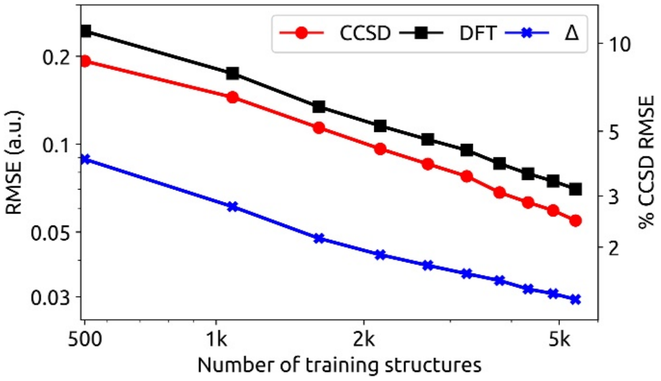

Fig. 1. Learning curves for the per atom polarizabilities of the molecules in the QM7b database calculated using either CCSD or DFT as well as for the difference (Δ) between the two. The testing set consists of 1,811 molecules, and the right-hand axis shows the RMSE as a fraction of the intrinsic variability of the CCSD polarizability, σ_{CCSD}.

and slightly easier to learn than more approximate methods, like DFT.

The AlphaML model can also be trained to evaluate the correction between different levels of theory, a correction commonly referred to as Δ learning that is often found to result in much smaller error than learning the raw quantity itself (29, 30). For instance, the use of DFT as a baseline to learn CCSD polarizabilities reduces the error by an additional factor of $\sim 2\times$ relative to the direct learning of $\boldsymbol{\alpha}_{\mathrm{CCSD}}$ (Fig. 1). Δ learning therefore provides a way to further reduce the prediction error at the cost of performing a baseline DFT calculation. In *SI Appendix*, we demonstrate that the performance of AlphaML is rather insensitive to the details of the target electronic structure method, showing similar accuracy for SCAN0 as that observed for B3LYP.

Extrapolation to Larger Molecules. Our definition of the kernel between two molecules as an average of environmental kernels means that the polarizabilities predicted by AlphaML are given as a sum of predicted polarizabilities for each environment (30). This feature allows one to predict $\boldsymbol{\alpha}$ for larger molecules. To test the behavior of AlphaML in this extrapolative regime, we trained this model on the entire QM7b database and then predicted the polarizabilities in a showcase dataset of 52 large molecules, which includes amino acids, nucleobases, drug molecules, carbohydrates, and 23 isomers of $\mathrm{C_8H}_n$ (the molecule key is in *SI Appendix*). As discussed in *SI Appendix*, many of these molecules are at the periphery of the portion of chemical compound space spanned by the QM7b dataset and therefore constitute a challenging test for AlphaML.

In Table 1, we show the RMSE errors in predicting $\boldsymbol{\alpha}$ for the showcase molecules using AlphaML as well as the error made when using DFT to approximate CCSD. Table 1 also breaks down the error into the $\lambda = 0$ and $\lambda = 2$ components of $\boldsymbol{\alpha}$; with an error in the anisotropic response comparable with that in the trace, this demonstrates that AlphaML learns both components with similar efficiency. As seen in the previous section, we again note that using the AlphaML model to predict CCSD polarizabilities is more accurate than simply using DFT. However, the use of DFT as the baseline in the Δ-learning sense leads to an additional reduction of $\sim 20-30\%$ in the error. In *SI Appendix*, we further discuss the behavior of the model when using the SCAN0 functional, which is similar to that observed here for B3LYP. While AlphaML predicts CCSD polarizabilities of the showcase molecules with better than DFT accuracy, we observe a substantial decrease in accuracy, which is to be expected when the model is extrapolated to the larger molecules in the showcase dataset.

We can investigate the performance of AlphaML in more detail by analyzing the errors of individual molecules in the showcase dataset. Fig. 2 shows that the errors are actually very small for most molecules. Large errors occur predominantly for highly polarizable compounds, particularly those that show a large degree of conjugation, such as long-chain alkenes and the purine nucleobases. For these systems, the underlying electronic structure is characterized by a high degree of delocalization, which requires larger cutoffs and more complex reference molecules to ensure accurate predictions. The ML predictions for the tensorial component of the polarizability, $\boldsymbol{\alpha}^{(2)}$, tend to be slightly less accurate than the DFT reference except for the highly polarizable alkenes, for which AlphaML dramatically outperforms DFT. Sulfur-containing structures, which are poorly represented in QM7b, also exhibit comparatively large errors.

The large discrepancy between DFT, CCSD, and AlphaML observed for alkenes (like octatetraene) reflects the nonlocal and collective nature of the underlying physics in these systems as well as the inherent structure of the AlphaML model. For DFT and CCSD, the narrowing HOMO-LUMO (highest occupied molecular orbital-lowest unoccupied molecular orbital) gaps in conjugated hydrocarbons lead to near-metallic states, which are known to exhibit strong multireference character (46). As such, these systems represent a significant challenge for electronic structure methods (like DFT and CCSD) that are not explicitly based on a multireference wavefunction. In practice, this leads to divergent polarizabilities (47, 48), and methods like CCSD are no longer reliable as the source of reference quantum chemical data for ML. An ML framework like AlphaML, which relies on local atomic environments to represent structures, tacitly disregards any collective (nonlocal) behavior that extends beyond the range of the local domains and the size of the molecules included in the training set. As shown in Fig. 3, the per carbon polarizabilities predicted by AlphaML therefore saturate to a constant value for the s-*trans* alkenes and acenes that are larger than those included in the QM7b dataset (i.e., hexatriene and benzene, respectively). Although this is a limitation when trying to learn collective and nonlocal physics, the local structure of AlphaML is also instrumental for obtaining the accurate and transferable predictions that we demonstrated on the showcase dataset.

Even when it comes to challenging conjugated systems with a vanishing HOMO-LUMO gap, the predictions of AlphaML are stable and completely avoid the unphysical and divergent predictions of costlier (but far from reference) quantum mechanical methods, like DFT and CCSD. For molecules with a sizable gap (like $\mathrm{C_{60}}$), the nonlocality is less pathological, and AlphaML performs remarkably well. For this prototypical nanotechnological

Table 1. RMSE in the prediction of the per atom polarizabilities of 52 showcase molecules

Method	RMSE	RMSE ($\lambda = 0$)	RMSE ($\lambda = 2$)
CCSD/DFT	0.573	0.348	0.456
CCSD/ML	0.244	0.120	0.212
DFT/ML	0.302	0.143	0.266
Δ(CCSD-DFT)/ML	0.181	0.083	0.161

CCSD/DFT denotes the discrepancy between CCSD and DFT values, while CCSD/ML and DFT/ML give the errors in predicting CCSD and DFT polarizabilities using AlphaML. Δ(CCSD-DFT)/ML gives the error in predicting the differences between the CCSD and DFT polarizabilities. All ML predictions are based on training on the full QM7b database. The total RMSE is expressed in atomic units (a.u.) per atom and broken down into the errors associated with the scalar ($\lambda = 0$) and tensorial ($\lambda = 2$) components of $\boldsymbol{\alpha}$.

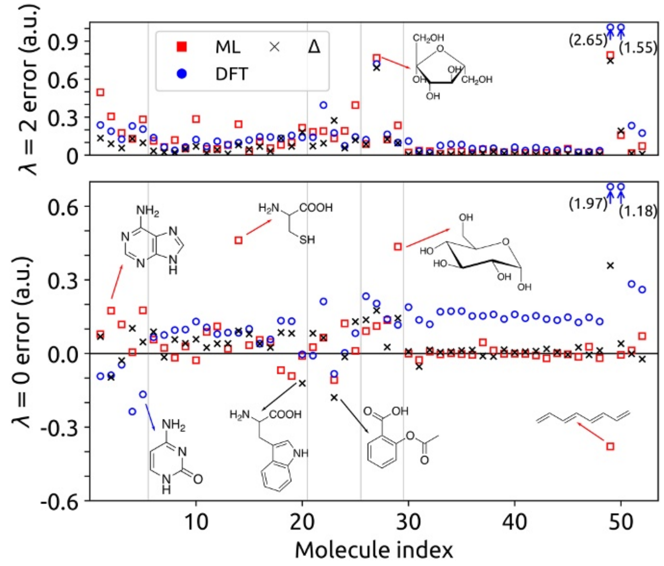

Fig. 2. RMSE made in approximating the $\lambda = 0$ (*Lower*) and $\lambda = 2$ (*Upper*) components of the per atom polarizability in the showcase dataset. The *x* axis corresponds to the numerical indices provided in the showcase molecule key in *SI Appendix*, and the vertical lines show the partitioning of the dataset into the different groups outlined in the same figure. Red squares show the ML error, blue circles show the error made in using DFT to approximate CCSD, and black crosses show the error made when Δ learning the CCSD correction with respect to DFT.

system, ML predictions are within 10% of DFT and CCSD results and within the range of experimental values, despite the extrapolation to a system size that is one order of magnitude larger than the molecules in the training set.

Atom-Centered Environmental Polarizabilities. The atom-centered structure of AlphaML provides a natural additive decomposition of $\boldsymbol{\alpha}$ into a sum of local terms, $\sum_i \boldsymbol{\alpha}_i$, which can be used to better understand how different functional groups contribute to the molecular polarizability. Unlike other methods for decomposing the polarizability, such as an atoms-in-molecules scheme (52) or a self-consistent decomposition (53), the approach used in this section does not require any additional calculations on top of the molecular polarizability, as the atom-centered polarizabilities are obtained as a by-product of the local nature of the SA-GPR scheme. When interpreting the $\boldsymbol{\alpha}_i$, one should keep in mind that each term corresponds to the contribution from the entire atom-centered environment, and the way that the polarizability is split between neighboring atoms is entirely inductive, reflecting the interplay between data, structure (as represented by the kernels), and regression rather than explicit physicochemical considerations. For instance, a few atoms within the showcase dataset (in particular, several H environments) have $\boldsymbol{\alpha}_i$ with negative eigenvalues, which reflects the fact that they reduce the dielectric response of the functional group to which they belong.

With this in mind, one can recognize physically meaningful features in the magnitude and anisotropy of the $\boldsymbol{\alpha}_i$. Fig. 4 depicts eight representative examples. Comparing saturated and unsaturated hydrocarbons (e.g., 2,2-dimethylhexane, cis-4-octene, and octatetraene), one sees that AlphaML predicts the contribution from the unsaturated carbon atoms to be large and very anisotropic, which is consistent with the higher degree of electron delocalization along conjugated molecules. Similarly large and anisotropic contributions are associated with aromatic systems as seen, for example, in guanine and the indole ring of tryptophan. Oxygen atoms are associated with a very anisotropic $\boldsymbol{\alpha}_i$; a large fraction of the polarizability of –OH and –COOH groups is assigned to the environments centered around nearby H and C atoms, but O atoms systematically contribute another anisotropic term oriented perpendicularly to the highly polarizable lone pairs (e.g., fructose as well as the carboxyl group in the amino acids). The sulfur-centered environments in cysteine and methionine have the largest contribution to the total polarizability in the showcase set and exhibit a strongly anisotropic response. All of these examples suggest that AlphaML can use relatively local structural information to determine an atom-centered decomposition of $\boldsymbol{\alpha}$ that encodes nontrivial quantum mechanical contributions from each functional group (or moiety) contained within a given molecule. It is this ability to predict such an environment-dependent decomposition of $\boldsymbol{\alpha}$ that underlies the observed better than DFT performance of AlphaML when faced with the often insurmountable challenge of transferability to a sector of chemical compound space that contains molecules that are quite distinct and notably larger than those included in the training set. A similar atom-centered decomposition can also be performed in the context of Δ learning, revealing the molecular features that are associated with the most substantial errors of the approximate methods. As shown in *SI Appendix*, this approach reveals how the large discrepancy between DFT and CCSD for alkenes is associated primarily with the extended conjugate system.

Discussion

Polarizability calculations with traditional quantum chemical methods have always implied a tradeoff between accuracy and computational cost. While CCSD calculations give more accurate predictions for the polarizabilities of molecules (especially large molecules) than DFT with various functionals (4, 21), the associated computational cost can be prohibitive. In our case, the CCSD calculations for the largest molecules in the showcase dataset required thousands of central processing unit (CPU) hours and approximately 500 GB of RAM. In this paper, we have demonstrated that the AlphaML framework, which combines SA-GPR with λ-SOAP kernels and CCSD reference calculations on small molecules, allows us to sidestep these expensive calculations and obtain results with an accuracy that almost

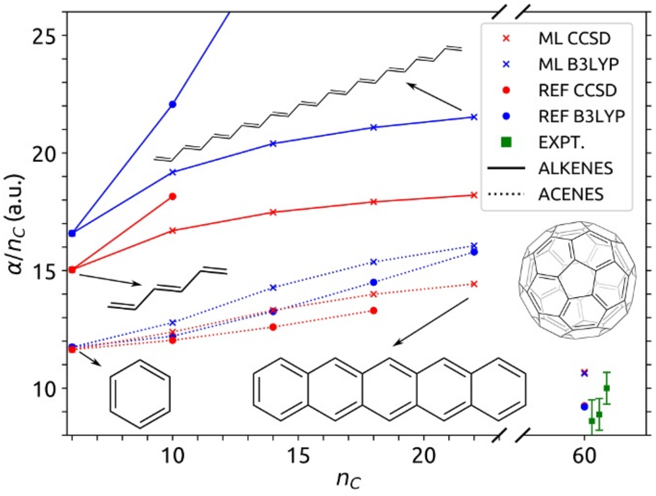

Fig. 3. Polarizability per C atom for the series of s-*trans* alkenes (from C_6H_8 to $C_{22}H_{24}$) and acenes (from benzene to pentacene) as well as fullerene (C_{60}). The reference CCSD results for anthracene and tetracene were taken from ref. 49, and the reference CCSD result for C_{60} was taken from ref. 50. The green squares (and error bars) indicate the experimental measurements for C_{60} (51). Results are provided from DFT and CCSD calculations as well as the corresponding AlphaML models.

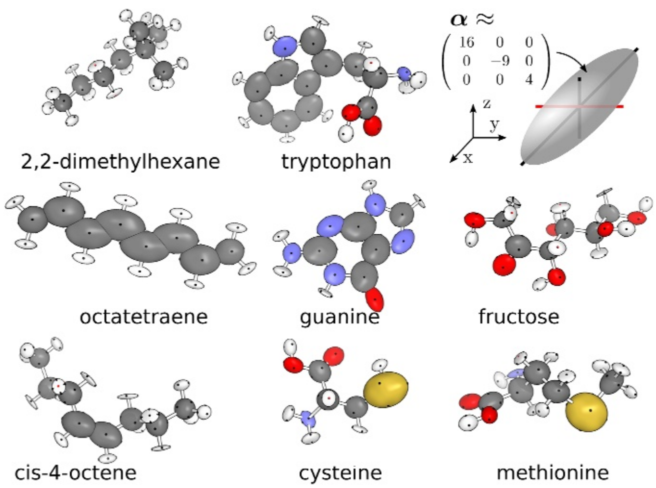
2,2-dimethylhexane
tryptophan
$\boldsymbol{\alpha} \approx \begin{pmatrix} 16 & 0 & 0 \\ 0 & -9 & 0 \\ 0 & 0 & 4 \end{pmatrix}$
z
y
x
octatetraene
guanine
fructose
cis-4-octene
cysteine
methionine

Fig. 4. Predicted atomic contributions to the total CCSD polarizability tensor for a selection of showcase molecules. The ellipsoids are aligned along the principal axes of $\boldsymbol{\alpha}_i$, and their extent is proportional to the square root of the corresponding eigenvalue. The principal axes are shown, and they are colored based on whether the corresponding eigenvalues are positive (black) or negative (red). The figure key (which is not drawn to scale) has additional details.

always exceeds DFT but at a fraction of the computational cost. Although this model was trained on a database of small organic molecules, it can be used to predict larger compounds with an accuracy that rivals DFT and can be systematically improved by extending the training set. The atom-centered decomposition of the ML predictions of $\boldsymbol{\alpha}$ can be interpreted in terms of physicochemical considerations, revealing for instance the large and anisotropic contributions that originate from delocalized π systems. In doing so, however, one should keep in mind that these contributions correspond to chemical environments rather than an atoms-in-molecules decomposition scheme.

Having shown the promise of the AlphaML framework by learning polarizabilities of small molecules, future work will focus on extensions of the training set to include larger molecules and oligomers, improvements in the accuracy of the underlying reference calculations, incorporation of methods to estimate the uncertainty in the predictions, as well as more efforts to include collective and nonlocal physics into the model. The need for these fundamental developments is underscored by an analysis of the behavior of the challenging series of conjugated alkenes, which are predicted by DFT and CCSD to have divergent polarizabilities due to vanishing HOMO-LUMO gaps. These improvements will make it possible to predict the polarizability for increasingly complex molecular systems and eventually, condensed phases. The availability of inexpensive atom-centered estimates of the fully anisotropic $\boldsymbol{\alpha}$ will be useful to design more accurate polarizable force fields for atomistic simulations as well as to computationally evaluate Raman and sum frequency generation spectroscopies, thereby improving the predictive power of simulations and increasing the insight that can be obtained from experiments.

Materials and Methods

First Principles Calculations. In this work, DFT calculations with the B3LYP functional (36, 37) and all LR-CCSD calculations were performed with Psi4 v1.1 (54), while DFT calculations with the SCAN0 functional (38) were performed with Q-Chem v5.0 (55). All of the molecular geometries used for ML were taken from the QM7b database (34, 35). All 52 showcase molecules (as well as the alkene series, acene series, and fullerene molecule in Fig. 3) were relaxed following the protocol used for QM7b (34). All DFT polarizabilities were computed with the finite-field method using a central difference formula with a step size of $\delta \boldsymbol{E} = 1.8897261250 \times 10^{-5}$ a.u. All CCSD polarizabilities were calculated using LR-CCSD, except for those of the eight largest molecules in the showcase dataset (which include molecules 18, 19, 20, 21, 23, 25, 26, and 28 as listed in *SI Appendix*). CCSD polarizabilities for these molecules were obtained using the (orbital unrelaxed) finite-field method due to the prohibitively large computational resources (memory and disk) needed by LR-CCSD. The frozen core approximation and direct scf_type were used during all CCSD calculations. In the cases where the finite-field method was used, CCSD polarizabilities were obtained as $\boldsymbol{\alpha} = \partial^2 U / \partial \boldsymbol{E}^2$. Additional details of the calculations are given in *SI Appendix*. The polarizability data generated can be found in Yang et al. (56).

Error Assessment. We use the Frobenius norm, defined as $\|\boldsymbol{\alpha}\|_F^2 = \sum_{i,j \in \{x,y,z\}} \alpha_{ij}^2$, to assess the accuracy of a polarizability estimate $\boldsymbol{\alpha}$ in a way that is rotationally invariant and includes both scalar and anisotropic components. Given two sets of polarizabilities, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_i'$, for N structures (each containing n_i atoms), we define the following quantities: $\text{MSE} \equiv \frac{1}{N} \sum_i (\|\boldsymbol{\alpha}_i\|_F - \|\boldsymbol{\alpha}_i'\|_F)/n_i$; $\text{MAE} \equiv \frac{1}{N} \sum_i \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i'\|_F / n_i$; and $\text{RMSE} \equiv [\frac{1}{N} \sum_i \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i'\|_F^2 / n_i^2]^{1/2}$. Errors are defined on a per atom basis to simplify the comparison between molecules of different sizes.

SA-GPR. The SA-GPR framework used herein to build an ML model for the polarizability is based on the following steps. (*i*) Each polarizability tensor, $\boldsymbol{\alpha}$, is decomposed into its irreducible (real spherical) components: the scalar $\alpha^{(0)} = (\alpha_{xx} + \alpha_{yy} + \alpha_{zz})/\sqrt{3}$ and the five-vector $\boldsymbol{\alpha}^{(2)} = \sqrt{2}\left[\alpha_{xy}, \alpha_{yz}, \alpha_{xz}, \frac{2\alpha_{zz} - \alpha_{xx} - \alpha_{yy}}{2\sqrt{3}}, \frac{\alpha_{xx} - \alpha_{yy}}{2}\right]$. One can compute the RMSE separately on these two components, since $\|\boldsymbol{\alpha}\|_F^2 = |\alpha^{(0)}|_F^2 + |\boldsymbol{\alpha}^{(2)}|_F^2$. (*ii*) λ-SOAP vector components $\langle \alpha n l \alpha' n' l' | \mathcal{X}_{j,\lambda\mu}^{(2)} \rangle$ are computed for each environment $\mathcal{X}_j$ and describe interatomic correlations within a prescribed cutoff radius, r_c, of the central atom j. The definition of these components is given in ref. 33. (*iii*) The base kernel between two environments, suitable to learn tensor components of order λ, is then defined as

$$k_{\mu j, \mu' k}^{\lambda} \equiv k_{\mu \mu'}^{\lambda}(\mathcal{X}_j, \mathcal{X}_k) = \sum_{\{J\}} \langle \mathcal{X}_{j,\lambda\mu} | J \rangle \langle J | \mathcal{X}_{k,\lambda\mu'} \rangle^{\star}, \qquad [1]$$

where we use the shorthand $\{J\}$ to indicate a subset of the possible spherical harmonic components of the descriptors, $|\alpha n l \alpha' n' l' \rangle$, that are automatically selected with a farthest-point sampling procedure (43). (*iv*) The linear SOAP kernel can describe atomic correlations up to three-body terms. Many-body correlations can be introduced by normalizing it and then raising it to an integer power. To preserve the linear nature of the λ-SOAP kernels, which is crucial to enforce the correct symmetry properties, we use

$$\begin{aligned} k_{\mu\mu'}^{\lambda,\zeta}(\mathcal{X}_j, \mathcal{X}_k) &\leftarrow k_{\mu\mu'}^{\lambda}(\mathcal{X}_j, \mathcal{X}_k)\, k_{00}^{0}(\mathcal{X}_j, \mathcal{X}_k)^{\zeta - 1}; \\ \mathbf{k}_{j,k}^{\lambda} &\leftarrow \mathbf{k}_{j,k}^{\lambda} \Big/ \sqrt{\left\|\mathbf{k}_{j,j}^{\lambda}\right\|_F \left\|\mathbf{k}_{k,k}^{\lambda}\right\|_F}. \end{aligned} \qquad [2]$$

(*v*) For each component of $\boldsymbol{\alpha}$, we build a kernel ridge regression model with weights $w_{k\mu}$ that are determined by optimizing the loss

$$\ell^2 = \sum_{\mu, \mathcal{A} \in N} \left| \alpha_{\mu}^{(\lambda)}(\mathcal{A}) - \sum_{\substack{k \in M \\ j \in \mathcal{A}}} w_{k\mu'} (\mathbf{k}_{j,k}^{\lambda})_{\mu\mu'} \right|^2 + \sigma^2 \mathbf{w}^T \mathbf{K}_{MM} \mathbf{w}, \qquad [3]$$

in which N is the training set, M is a (possibly sparse) set of representative environments used as the basis, and $\mathbf{K}_{MM}$ is the matrix of kernels between representative environments. An online prediction tool for $\boldsymbol{\alpha}$, based on the AlphaML framework, is also available at http://alphaml.org.

ACKNOWLEDGMENTS. The authors thank Felix Musil and Michael Willatt for helpful discussions. D.M.W. and M.C. acknowledge support from European Research Council Horizon 2020 Grant 677013-HBMAP. A.G. acknowledges funding from the Max Planck–École Polytechnique Fédérale de Lausanne (MPG-EPFL) Center for Molecular Nanoscience. Y.Y., K.U.L., and R.A.D. acknowledge partial support from Cornell University through startup funding. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by Office of Science of the US Department of Energy Contract DE-AC02-06CH11357. This work was also supported by a grant from Swiss National Supercomputing Center Project ID s843 and the EPFL Scientific Computing Center.

1. Engel E, Dreizler RM (2011) *Density Functional Theory: An Advanced Course* (Springer, Berlin).
2. Burke K (2012) Perspective on density functional theory. *J Chem Phys* 136:150901.
3. Lejaeghere K, et al. (2016) Reproducibility in density functional theory calculations of solids. *Science* 351:145–152.
4. Hait D, Head-Gordon M (2018) How accurate are static polarizability predictions from density functional theory? An assessment over 132 species at equilibrium geometry. *Phys Chem Chem Phys* 20:19800–19810.
5. Stone A (1997) *The Theory of Intermolecular Forces*, International Series of Monographs on Chemistry (Clarendon, Oxford, United Kingdom).
6. Hermann J, DiStasio RA, Jr, Tkatchenko A (2017) First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications. *Chem Rev* 117:4714–4758.
7. Grimme S (2014) Dispersion interaction and chemical bonding. *The Chemical Bond: Chemical Bonding Across the Periodic Table*, eds Frenking G, Shaik S (Wiley-VCH, Hoboken, NJ), pp 477–500.
8. Shen YR (1989) Surface properties probed by second harmonic and sum-frequency generation. *Nature* 337:519–525.
9. Luber S, Iannuzzi M, Hutter J (2014) Raman spectra from ab initio molecular dynamics and its application to liquid s-methyloxirane. *J Chem Phys* 141:094503.
10. Morita A, Hynes JT (2000) A theoretical analysis of the sum frequency generation spectrum of the water surface. *Chem Phys* 258:371–390.
11. Medders GR, Paesani F (2016) Dissecting the molecular structure of the air/water interface from quantum simulations of the sum-frequency generation spectrum. *Chem Phys Lett* 138:3912–3919.
12. Sprik M, Klein ML (1988) A polarizable model for water using distributed charge sites. *J Chem Phys* 89:7556–7560.
13. Fanourgakis GS, Xantheas SS (2008) Development of transferable interaction potentials for water. v. extension of the flexible, polarizable, thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *J Chem Phys* 128:074506.
14. Ponder JW, et al. (2010) Current status of the AMOEBA polarizable force field. *J Phys Chem B* 114:2549–2564.
15. Medders GR, Babin V, Paesani F (2014) Development of a "first-principles" water potential with flexible monomers. III. Liquid phase properties. *J Chem Theory Comput* 10:2906–2910.
16. Bereau T, DiStasio RA, Jr, Tkatchenko A, von Lilienfeld OA (2018) Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J Chem Phys* 148:241706.
17. Monkhorst HJ (1977) Calculation of properties with the coupled-cluster method. *Int J Quantum Chem* 12:421–432.
18. Koch H, Jørgensen P (1990) Coupled cluster response functions. *J Chem Phys* 93:3333–3344.
19. Christiansen O, Jørgensen P, Hättig C (1998) Response functions from Fourier component variational perturbation theory applied to a time-averaged quasienergy. *Int J Quantum Chem* 68:1–52.
20. Christiansen O, Gauss J, Stanton JF (1999) Frequency-dependent polarizabilities and first hyperpolarizabilities of CO and H_2O from coupled cluster calculations. *Chem Phys Lett* 305:147–155.
21. Hammond JR, de Jong WA, Kowalski K (2008) Coupled-cluster dynamic polarizabilities including triple excitations. *J Chem Phys* 128:224102.
22. Hammond JR, Govind N, Kowalski K, Autschbach J, Xantheas SS (2009) Accurate dipole polarizabilities for water clusters n=2-12 at the coupled-cluster level of theory and benchmarking of various density functionals. *J Chem Phys* 131:214103.
23. Lao KU, Jia J, Maitra R, DiStasio RA, Jr (2018) On the geometric dependence of the molecular dipole polarizability in water: A benchmark study of higher-order electron correlation, basis set incompleteness error, core electron effects, and zero-point vibrational contributions. *J Chem Phys* 149:204303.
24. Behler J, Parrinello M (2007) Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys Rev Lett* 98:146401.
25. Bartók AP, Payne MC, Kondor R, Csányi G (2010) Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys Rev Lett* 104: 136403.
26. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* 108:058301.
27. De S, Bartók AP, Csányi G, Ceriotti M (2016) Comparing molecules and solids across structural and alchemical space. *Phys Chem Chem Phys* 18:13754–13769.
28. Faber FA, et al. (2017) Prediction errors of molecular machine learning models lower than hybrid DFT error. *J Chem Theory Comput* 13:5255–5264.
29. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA (2015) Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J Chem Theory Comput* 11:2087–2096.
30. Bartók AP, et al. (2017) Machine learning unifies the modeling of materials and molecules. *Sci Adv* 3:e1701816.
31. Bereau T, Andrienko D, von Lilienfeld OA (2015) Transferable atomic multipole machine learning models for small organic molecules. *J Chem Theory Comput* 11:3225–3233.
32. Liang C, et al. (2017) Solvent fluctuations and nuclear quantum effects modulate the molecular hyperpolarizability of water. *Phys Rev B* 96:041407.
33. Grisafi A, Wilkins DM, Csányi G, Ceriotti M (2018) Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Phys Rev Lett* 120:036002.
34. Montavon G, et al. (2013) Machine learning of molecular electronic properties in chemical compound space. *New J Phys* 15:095003.
35. Blum LC, Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc* 131:8732–8733.
36. Becke AD (1993) Density-functional thermochemistry. III, the role of exact exchange. *J Chem Phys* 98:5648–5652.
37. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J Phys Chem* 98:11623–11627.
38. Hui K, Chai JD (2016) Scan-based hybrid and double-hybrid density functionals from models without fitted parameters. *J Chem Phys* 144:044114.
39. Woon DE, Dunning TH, Jr (1994) Gaussian basis sets for use in correlated molecular calculations. IV. calculation of static electrical response properties. *J Chem Phys* 100:2975–2988.
40. Christiansen O, Hättig C, Gauss J (1998) Polarizabilities of CO, N_2, HF, Ne, BH, and CH^+ from *ab initio* calculations: Systematic studies of electron correlation, basis set errors and vibrational contributions. *J Chem Phys* 109:4745–4757.
41. Reis H, Papadopoulos MG, Avramopoulos A (2003) Calculation of the microscopic and macroscopic linear and nonlinear optical properties of acetonitrile. I. Accurate molecular properties in the gas phase and susceptibilities of the liquid in onsager's reaction-field model. *J Phys Chem A* 107:3907–3917.
42. Karne AS, et al. (2015) Systematic comparison of DFT and CCSD dipole moments, polarizabilities and hyperpolarizabilities. *Chem Phys Lett* 635:168–173.
43. Imbalzano G, et al. (2018) Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J Chem Phys* 148:241730.
44. Bartók AP, Kondor R, Csányi G (2013) On representing chemical environments. *Phys Rev B* 87:184115.
45. Glielmo A, Zeni C, De Vita A (2018) Efficient nonparametric *n*-body force fields from machine learning. *Phys Rev B* 97:184307.
46. Voloshina E, Paulus B (2014) First multireference correlation treatment of bulk metals. *J Chem Theory Comput* 10:1698–1706.
47. Smith SM, et al. (2004) Static and dynamic polarizabilities of conjugated molecules and their cations. *J Phys Chem A* 108:11063–11072.
48. Grüning M, Gritsenko OV, Baerends EJ (2002) Exchange potential from the common energy denominator approximation for the Kohn–Sham Green's function: Application to (hyper)polarizabilities of molecular chains. *J Chem Phys* 116:6435–6442.
49. Huzak M, Deleuze MS (2013) Benchmark theoretical study of the electric polarizabilities of naphthalene, anthracene, and tetracene. *J Chem Phys* 138:024319.
50. Kowalski K, Hammond JR, de Jong WA, Sadlej AJ (2008) Coupled cluster calculations for static and dynamic polarizabilities of C_{60}. *J Chem Phys* 129:226101.
51. Sabirov DS (2014) Polarizability as a landmark property for fullerene chemistry and materials science. *RSC Adv* 4:44996.
52. Laidig KE, Bader RFW (1990) Properties of atoms in molecules: Atomic polarizabilities. *J Chem Phys* 93:7213–7224.
53. Applequist J, Carl JR, Fung KK (1972) Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *J Am Chem Soc* 94:2952–2960.
54. Parrish RM, et al. (2017) Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *J Chem Theory Comput* 13:3185–3197.
55. Shao Y, et al. (2015) Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol Phys* 113:184–215.
56. Yang Y, et al. (2019) Coupled-cluster polarizabilities in the QM7b and a showcase database. *Materials Cloud Archive (2019)*, doi:10.24435/materialscloud:2019.0002/v1.