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 (1–3) for 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.

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.

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.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

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

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

This article contains supporting information online at www.pnas.orglookup/supp/doi/10.1073/pnas.1816132116/-/DCSupplemental.
corpore 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 \(~7,000\) 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 polarization 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, \(\mu\), with respect to an external electric field \(E\) 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 \(\alpha\) (21–23, 39), we used the d-aug-cc-pVDZ basis set (39) for all calculations herein. Although this double-\(\zeta\) basis set has only a moderate number of polarization functions, augmentation with an additional set of diffuse functions almost always increases the convergence of \(\alpha\) 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 \(\alpha\) 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, \(n\), contained within a given molecule. On the QM7b database, the popular B3LYP hybrid DFT functional predicts \(\alpha\) 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 \(\sigma_{CCSD} = \left[ \frac{1}{N} \sum_i \left( \alpha_i^{CCSD} - \langle \alpha^{CCSD} \rangle \right)^2 \right]^{1/2}\). The large MSE value obtained with B3LYP indicates a systematic overestimation of \(\alpha\) 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 \(\alpha\) 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 \(\lambda\)-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 \(\lambda\)-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 \(\lambda\)-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 present 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, \(r_e\), 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 \(\alpha\) 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 \(\sigma_{CCSD} = 2.216\) 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 \(\sigma_{CCSD}\) 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 \(\sigma_{CCSD}\) 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 \(\sigma_{CCSD}\). As seen in other cases (29, 30), highly accurate quantum chemistry calculations are smoother
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 ~2 relative to the direct learning of αCCSD (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 α 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 CnHn (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 α 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 λ = 0 and λ = 2 components of α; 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 ~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, α(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 C60), the nonlocality is less pathological, and AlphaML performs remarkably well. For this prototypical nanotechnological

![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, σCCSD.](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE (λ = 0)</th>
<th>RMSE (λ = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSD/DFT</td>
<td>0.573</td>
<td>0.348</td>
</tr>
<tr>
<td>CCSD/ML</td>
<td>0.244</td>
<td>0.120</td>
</tr>
<tr>
<td>DFT/ML</td>
<td>0.302</td>
<td>0.143</td>
</tr>
<tr>
<td>Δ(CCSDDFT)/ML</td>
<td>0.181</td>
<td>0.083</td>
</tr>
</tbody>
</table>

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(CCSDDFT)/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 (λ = 0) and tensorial (λ = 2) components of α.
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 $\alpha$ into a sum of local terms, $\sum_i \alpha_i$, 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 $\alpha_i$, 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 $\alpha_i$ 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 $\alpha_i$. 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 $\alpha_i$; a large fraction of the polarizability of $\text{−OH}$ and $\text{−COOH}$ 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 $\alpha$ 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 $\alpha$ 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 $\Delta$ 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 $\lambda$-SOAP kernels and CCSD reference calculations on small molecules, allows us to sidestep these expensive calculations and obtain results with an accuracy that almost
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 α, and their extent is proportional to the square root of the corresponding eigenvalue. The principal axes are shown, and they can be interpreted in terms of physical-chemistry considerations, revealing for instance the large and nonlocal physics into the model. The need for improvements will make it possible to predict the polarizability which are predicted by DFT and CCSD to have divergent components of order r.

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 δE = 1.8897261250 × 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 α = ΔU/ΔE². 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 ||α||₂ = \( \sum_{i,j} |α_{ij}|^2 \), to assess the accuracy of a polarizability estimate α in a way that is rotationally invariant and includes both scalar and anisotropic components. Given two sets of polarizabilities, αi and α🗜, for N structures (each containing ni atoms), we define the following quantities: MSE = \( \frac{1}{N} \sum_{i} \left( \frac{\sum_{j} |α_{ij} - α_{i'j'}|}{n_i} \right) \), MAE = \( \frac{1}{N} \sum_{i} |α_{ij} - α_{i'j'}| / n_i \) and RMSE = \( \left( \frac{1}{N} \sum_{i} \left( |α_{ij} - α_{i'j'}| \right)^2 / n_i \right)^{1/2} \). 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, αi, is decomposed into its irreducible (real spherical) components: the scalar α(0) = (α + 2αzz + αxx + αyy) / 3 and the five-vector α(1) = \( \sqrt{2} \left[ 2α_{xx} - 2α_{yy} - 2α_{zz} + \sqrt{2} \left( α_{zz} - α_{xx} - α_{yy} \right) \right] \). One can compute the RMSE separately on these two components, since ||α(0)||₂ = ||α(1)||₂ = ||α(1)||₂. (ii) λ-SOA

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 αi, 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.

\[ α ≈ \begin{pmatrix} 16 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \]


