Accurate molecular polarizabilities with coupled-cluster theory and machine learning
Supplementary Information

David M. Wilkins,[1] Andrea Grisafi,[1] Yang Yang,[2] Ka Un Lao,[2] Robert A. DiStasio Jr.,[2] and Michele Ceriotti[1]

[1] Laboratory of Computational Science and Modeling, IMX,
École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[2] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA

www.pnas.org/cgi/doi/10.1073/pnas.1816132116

OPTIMIZATION OF KERNEL HYPERPARAMETERS

Power-Spectrum Sparsification

It has been shown recently that the efficiency of SOAP-based methods can be improved by sparsification: rather than retaining several tens of thousands of spherical harmonic components of the power spectrum, a farthest-point sampling method allows us to choose a set of components among which the difference in the values observed within the training set is as large as possible, meaning that we can retain instead a few hundred components with negligible loss in accuracy [S1]. Fig. S1 shows learning curves for the $\lambda = 0$ and $\lambda = 2$ polarizability components retaining either the full power spectrum or some subset thereof, using base power spectra with 8 radial functions and an l-cutoff of 6, with nonlinearity parameter $\zeta = 2$ and radial cutoff $r_c = 4$ Å. In the $\lambda = 2$ case it is prohibitively expensive to use the full power spectrum in building the kernel, so instead the error is shown to reach a plateau when a large enough number of components are retained. We see that both the $\lambda = 0$ and $\lambda = 2$ kernels remain accurate when 400 components are retained (amounting to ∼2% of the 16,128 components in the un-sparsified $\lambda = 0$ power spectrum and ∼0.7% of the 59,904 components in the original $\lambda = 2$ power spectrum). It is worth noting that even if only 10 components of the power spectra are kept, amounting to 0.06% for $\lambda = 0$ and 0.02% for $\lambda = 2$, then the error in predicting the polarizability is ∼15%, which is comparable to the error incurred when using DFT to predict the CCSD polarizability.

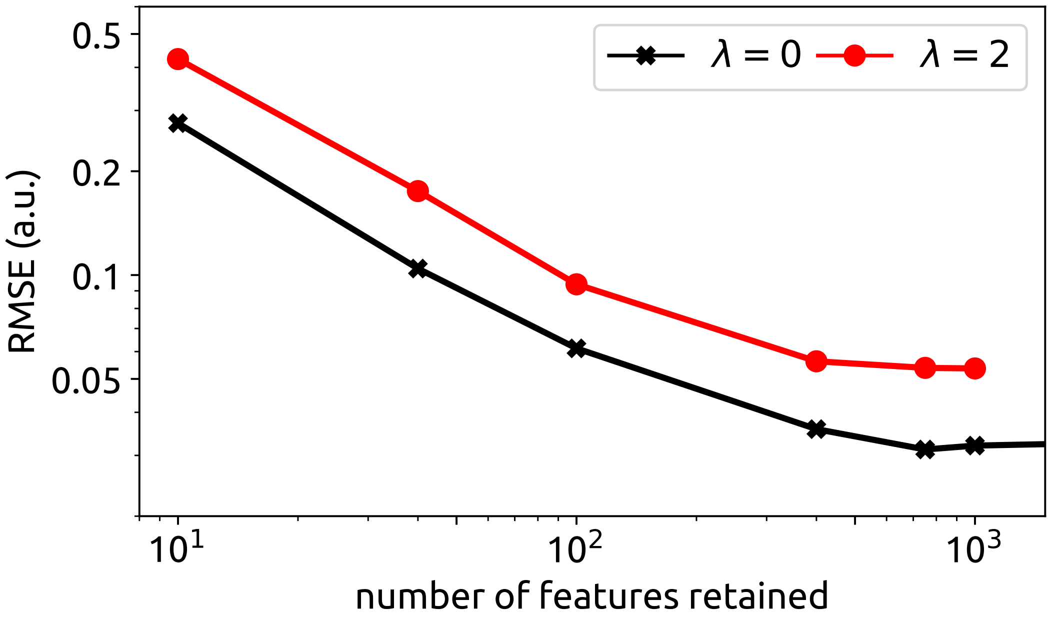

FIG. S1. Error in learning the $\lambda = 0$ and $\lambda = 2$ components of the per-atom polarizability for the QM7b dataset, with different percentages of the spherical harmonic components retained in calculating the kernels. Polarizabilities were calculated using CCSD. The training set was chosen to contain 5400 molecules, and the testing set in all cases consists of 1811 molecules. We use a nonlinearity parameter $\zeta = 2$ and radial cutoff $r_c = 4$ Å.

Nonlinear Kernels

Fig. S2 shows learning curves for the $\lambda = 0$ and $\lambda = 2$ components of the polarizability per atom, using kernels with varying values of the nonlinearity parameter ζ. In both cases, the linear kernel with $\zeta = 1$ saturates at large training set size, while the nonlinear kernels do not saturate. $\zeta = 2$ provides a clear advantage over the linear kernel, while larger values of ζ give no further improvement. The advantage of nonlinear kernels is apparent in both polarizability components: the accuracy in learning the $\lambda = 0$ component increases twofold when nonlinearity is included, and the accuracy of the $\lambda = 2$ component increases threefold.

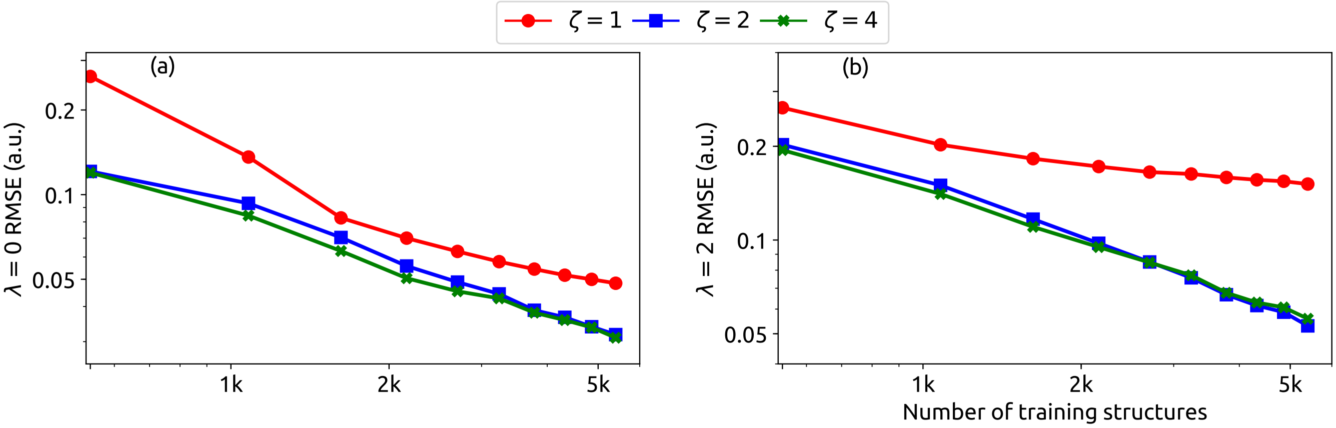

FIG. S2. Learning curves for (a) the $\lambda = 0$ and (b) the $\lambda = 2$ components of the per-atom polarizability for the QM7b dataset, with linear ($\zeta = 1$) and nonlinear ($\zeta = 2, 4$) kernels. Polarizabilities were calculated using CCSD, and the testing set in all cases consists of 1811 molecules. We use a radial cutoff $r_c = 4$ Å.

Multiscale Kernels

The cutoff radius r_c also affects the accuracy of learning. Fig. S3 shows learning curves for kernels with several different cutoff radii. In the case of the scalar SOAP kernel, it has previously been shown that a combination of kernels with different cutoffs can perform better than any individual kernel [S2]. To test whether this is also true of the λ-SOAP kernel, we optimized a multiscale $\lambda = 0$ kernel $k^{(\lambda=0)}_{\mathrm{MS}} = \sum_{i=2}^{5} c_i \, k^{(\lambda=0)}_{r_c=i\text{ Å}}$ to minimize the prediction error on the QM7b set, using 2-fold cross-validation on a training set of 5400 configurations. The coefficients c_i were used to build a multiscale kernel for both the $\lambda = 0$ and $\lambda = 2$ polarizability components. Fig. S3 also shows results for the optimum multiscale kernel, with $c_2 = 0.04053$, $c_3 = 0.00997$, $c_4 = 0.02250$ and $c_5 = 0.01560$. It can be clearly seen that, as has been previously observed for the scalar case only, for both components the learning curves with $r_c = 2$ Å kernels saturate fairly quickly, whereas kernels with larger cutoff radii do not lead to saturation. The MS kernel performs better than any single kernel, and as in Ref. [S2] the greatest contribution to the optimum multiscale kernel is from the $r_c = 2$ Å kernel.

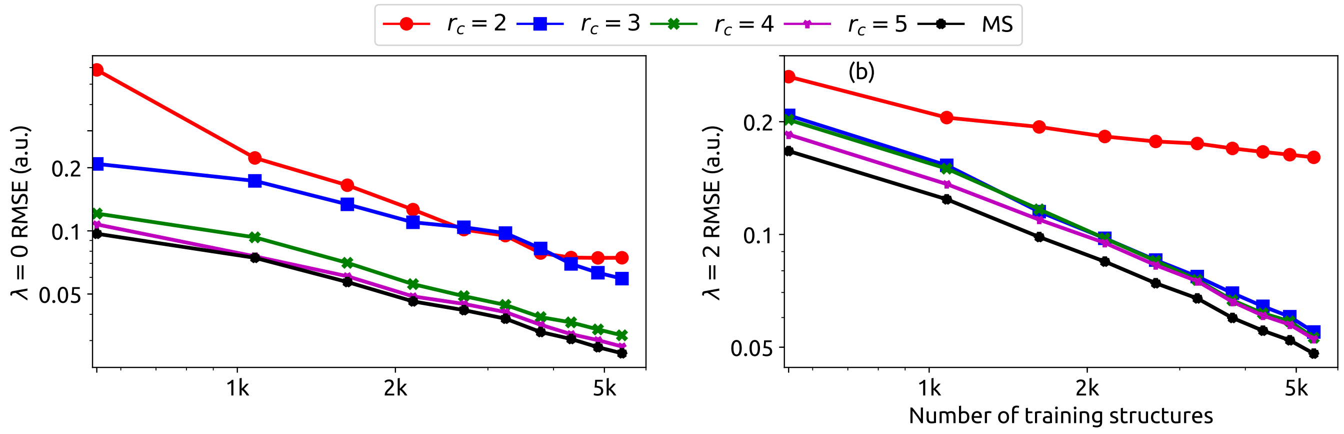

FIG. S3. Learning curves for (a) the $\lambda = 0$ and (b) the $\lambda = 2$ components of the per-atom polarizability for the QM7b dataset, using kernels built with different cutoffs and a multiscale (MS) kernel built by combining individual kernels. Polarizabilities were calculated using CCSD, and the testing set in all cases consists of 1811 molecules.

SHOWCASE MOLECULES

Fig. S4 shows a numbered key of the 52 showcase molecules used in the manuscript.

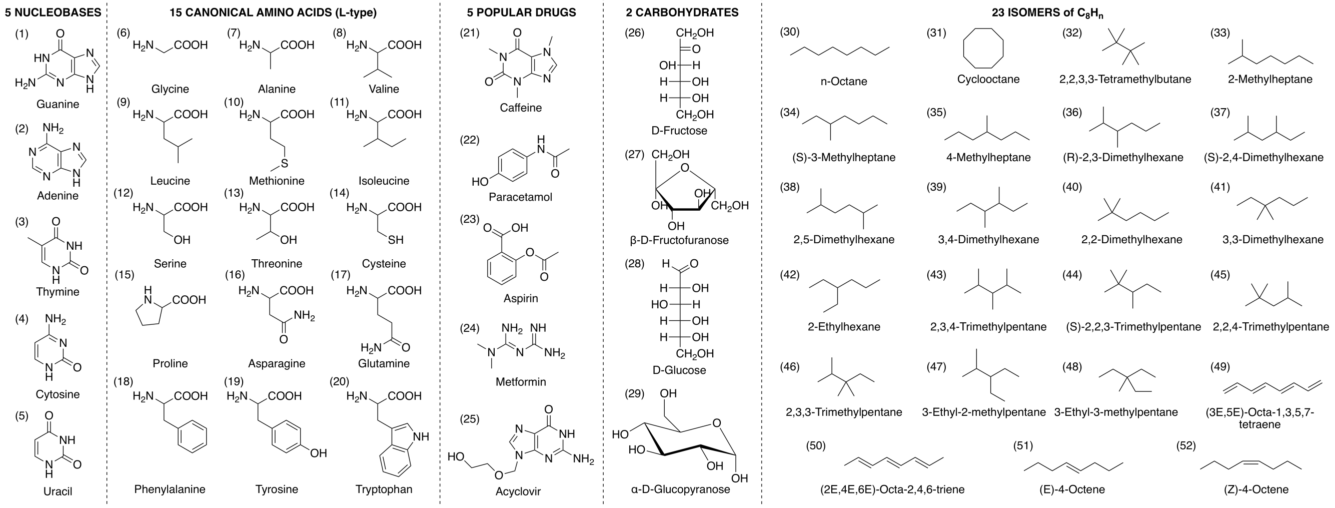

FIG. S4. Names and chemical structures of the 52 molecules included in the showcase dataset. The numbers refer to the position of each molecule in the dataset and are used for reference in the text and other figures.

LEARNING DIFFERENT LEVELS OF THEORY

QM7b Set

In Fig. S5 we show learning curves for both the SCAN0-DFT and B3LYP-DFT functionals as well as for the difference between the two. We see that although these two levels of theory provide different predictions for the polarizability, in each case the polarizability is learned with essentially the same accuracy. The difference between the

SCAN0 and B3LYP results is predicted with an accuracy of 7.92×10^{-3} a.u.. Performing Δ-learning on this difference allows $\boldsymbol{\alpha}_{\mathrm{SCAN0}}$ to be predicted with a relative error that is 0.44% of the 1.820 a.u. intrinsic deviation of the SCAN0 polarizabilities, a significant improvement over the 7.01×10^{-2} a.u. error (3.8% relative error) obtained by directly learning the SCAN0 polarizability. Fig. S5 also shows that Δ-learning of the difference between CCSD and SCAN0 polarizabilities can be performed with a very similar accuracy to that of the B3LYP-CCSD difference.

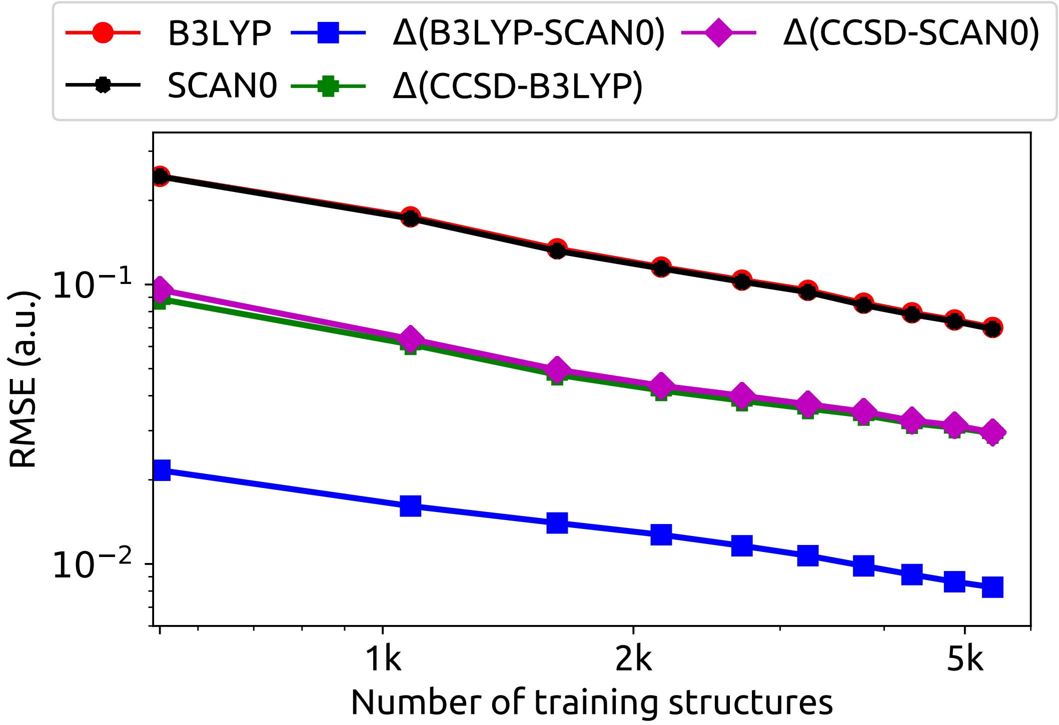

FIG. S5. Learning curves for the per-atom polarizability of the QM7b molecules, calculated using DFT with the B3LYP or SCAN0 functionals and the double-ζ basis set, as well as the differences between SCAN0 and B3LYP, between CCSD and B3LYP and between CCSD and SCAN0. In all cases the testing set consists of 1811 molecules.

Showcase Set

Table I shows the error in learning the SCAN0 polarizability of the members of the showcase set, extending Table I in the main text. The error incurred in using the SCAN0-DFT polarizability to predict the CCSD result is higher than using B3LYP-DFT for the showcase set, as shown in more detail by Fig. S6. The accuracy of the $\lambda = 2$ component given by both functionals is very similar for all showcase molecules, with B3LYP performing slightly worse in a small number of cases, the accuracy of the $\lambda = 0$ component varies much more: for the nucleobases, amino acids and sugar molecules the SCAN0 systematically under-predicts this component, while B3LYP generally slightly over-predicts it. On the other hand, for the hydrocarbons both functionals over-predict α_{iso}, with SCAN0 giving the more accurate prediction. As in the QM7b set, an AlphaML model of the SCAN0 polarizabilities performs about as well as that of the B3LYP polarizabilities, although the difference between SCAN0 and CCSD is more difficult to learn than that between B3LYP and CCSD, increasing the error by about 25%. Although Δ-learning from either of these functionals is more accurate than simply predicting the CCSD polarizability, the B3LYP polarizability is a better starting point for doing so.

TABLE I. Root mean square errors in machine-learning of the per-atom polarizabilities of the showcase molecules. CCSD/SCAN0 denotes the RMSE between CCSD and SCAN0 calculations, while CCSD/ML and SCAN0/ML give the errors in predicting CCSD and SCAN0 $\boldsymbol{\alpha}_n$ respectively, using a machine-learning model. Δ(CCSD-SCAN0)/ML gives the error in predicting the difference between CCSD and SCAN0 polarizability. In all cases, the full QM7b database is used as a training set. For comparison, the errors from Table I in the main text are also reproduced here.

Method	RMSE	RMSE($\lambda = 0$)	RMSE($\lambda = 2$)
CCSD/SCAN0	0.579	0.363	0.451
CCSD/B3LYP	0.573	0.348	0.456
CCSD/ML	0.244	0.120	0.212
SCAN0/ML	0.321	0.134	0.291
B3LYP/ML	0.302	0.143	0.266
Δ(CCSD-SCAN0)/ML	0.171	0.113	0.128
Δ(CCSD-B3LYP)/ML	0.181	0.083	0.161

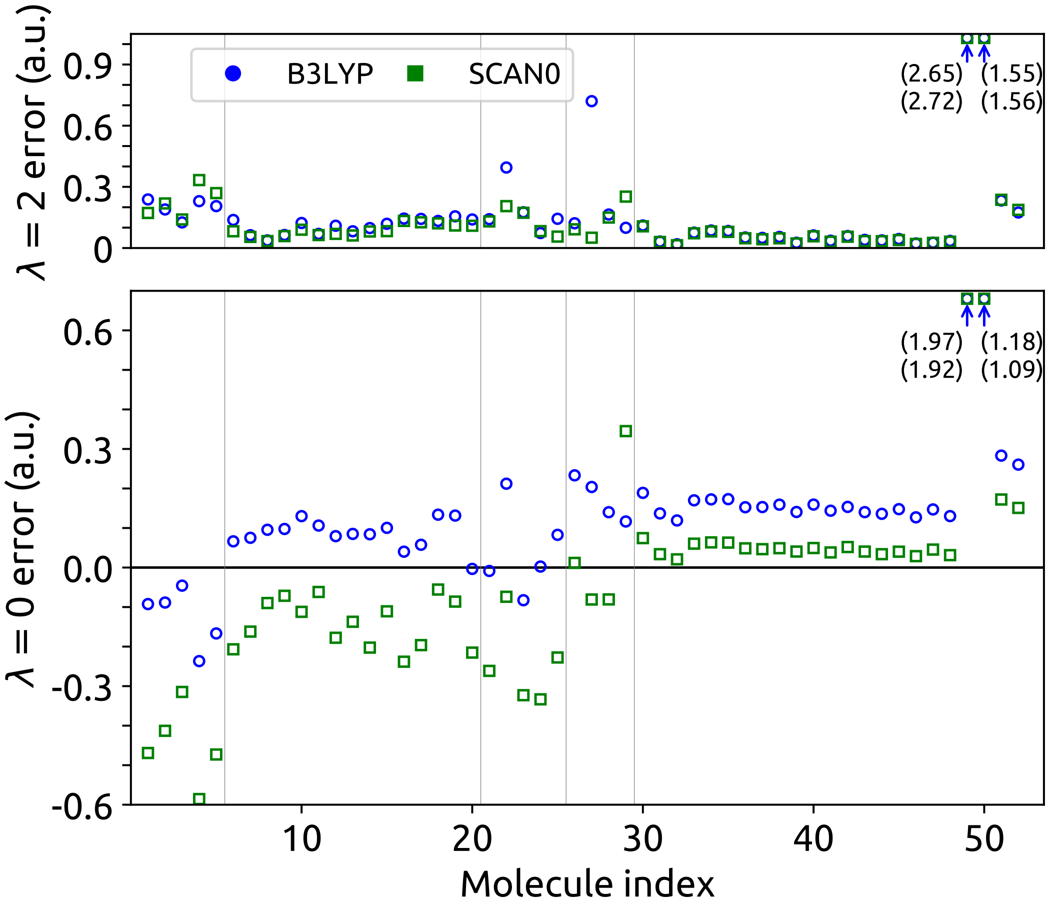

FIG. S6. Error made in approximating the $\lambda = 0$ (bottom panel) and $\lambda = 2$ (top panel) components of the average polarizability per atom for the 52 showcase molecules, as a function of the molecule indices in Fig. 2 of the main text. Vertical lines show the partitioning of these molecules into different groups. Blue circles show the error made in using the B3LYP polarizability to approximate the CCSD polarizability and green squares show the error when the SCAN0 polarizability is used to approximate the CCSD polarizability. Where components are outside of the graph, the top bracketed number refers to the B3LYP and the bottom number to the SCAN0 value.

REPRESENTATION OF CHEMICAL COMPOUND SPACE

Figure S7 shows a kernel principal component analysis (KPCA) [S3] for the QM7b dataset, wherein each point corresponds to a molecule and the positions of the points correspond to the two principal eigenvectors of the kernel matrix. We used the same kernel that we employed for the scalar part of AlphaML. Red dots correspond to the projection of the showcase molecules on these KPCA axes. One can see that most molecules lie at the periphery of the dataset, underscoring the difficulties associated with predicting their properties. The showcase molecules span a broad portion of the QM7b chemical compound space, indicating the diversity of this dataset. The distance between showcase molecules and QM7b points is roughly equivalent to the spacing between the first 52 farthest point sampling (FPS) points of the QM7b. In fact, the errors in the showcase predictions are indeed of the same order of magnitude as the accuracy of the QM7b model when trained on about 100 FPS reference compounds.

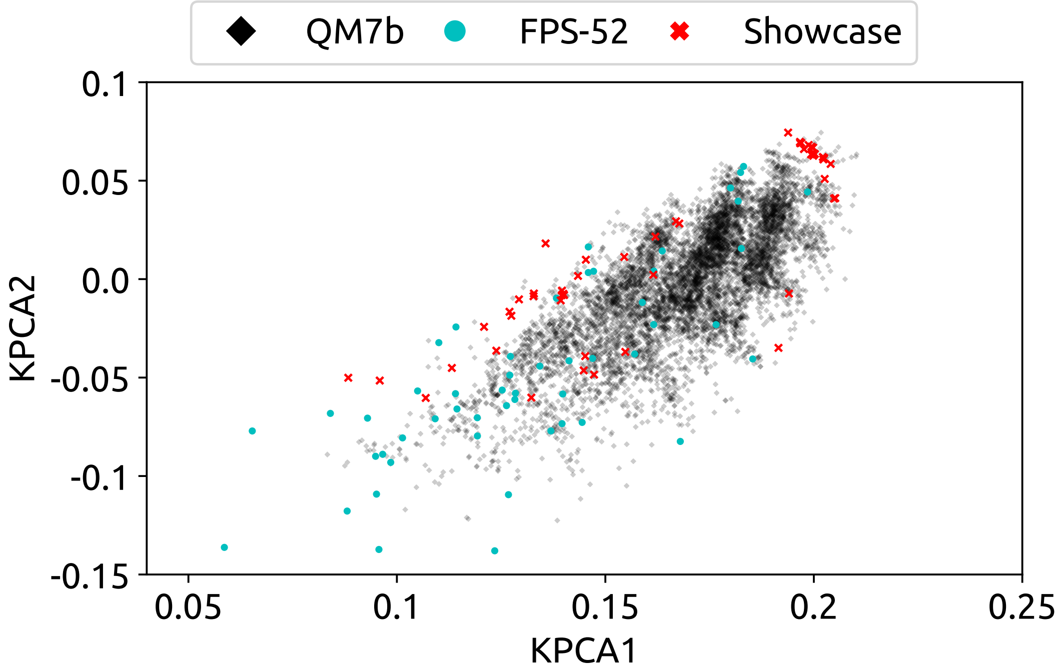

FIG. S7. A KPCA representation of the chemical space covered by the QM7b dataset (black points). Red points correspond to the projection of the showcase molecules, and cyan points to the first 52 FPS points from the QM7b set.

POLARIZABILITIES OF ATOM-CENTERED ENVIRONMENTS

The prediction of the polarizability as a sum of environmental contributions means that we can use AlphaML to better understand the origins of the difference between CCSD and DFT polarizabilities. By writing $(\boldsymbol{\alpha}_{\text{CCSD}} - \boldsymbol{\alpha}_{\text{DFT}}) = \Delta\boldsymbol{\alpha} = \sum_i \Delta\boldsymbol{\alpha}_i$, where i is an atom label, we can predict the contribution of each atom to this difference. Fig. S8 shows the distribution of the predicted difference between $\lambda = 0$ components of the CCSD and DFT polarizability, decomposed into atom-centered contributions for each member of the QM7b set. Aside from sulfur, which is considerably more polarizable than the other elements considered, each distribution is centered around some constant value. This suggests a simpler model for predicting the $\lambda = 0$ polarizability component, in which,

$$\alpha^{(0)}_{i,\text{CCSD}} = \alpha^{(0)}_{i,\text{DFT}} + \sum_j n_{ij} \Delta\alpha^{(0)}_{i,\text{eff}}, \tag{S1}$$

where the label i refers to a molecule and j an element, n_{ij} is the stoichiometry of element j in molecule i and $\Delta\alpha^{(0)}_{i,\text{eff}}$ is an effective $\lambda = 0$ polarizability difference for element i, which can be found by regression on the training set. Using this dressed-atom model we obtain a relative error of 11% (compared to the intrinsic deviation of the polarizability for the respective functional) in predicting the $\lambda = 0$ component of the B3LYP or SCAN0 polarizability. While this is an improvement over the ∼20% error between DFT and CCSD, this prediction is still far worse than that obtained from AlphaML. The effective differences we obtain are given in Table II, with the two different functionals giving quite different values for most of the elements. These $\Delta\alpha^{(0)}_{i,\text{eff}}$ values also afford some further insight into the results shown in Fig. S6: Both B3LYP and SCAN0 over-predict the $\lambda = 0$ polarizability of the hydrocarbons, but this is more pronounced for B3LYP. These molecules contain CH_2 and CH_3 groups, whose total $\Delta\alpha^{(0)}_{i,\text{eff}}$ are respectively 0.71 a.u. and 0.44 a.u. from B3LYP, and 0.31 a.u. and -0.05 a.u. from SCAN0, on average. That is, for SCAN0 there is much greater compensation between the carbon and hydrogen contributions to the total polarizability difference. On the other hand, the effective polarizability differences predicted for N,O, and S are significantly more negative for SCAN0 than for B3LYP, which is consistent with the systematic under-prediction by SCAN0 of the polarizability of nucleobases and amino acids, which contain these atoms.

While the effective $\lambda = 0$ polarizability predictions for C, H and N are quite reasonable compared to the distributions of Fig. S8, the predictions for O, Cl and S are much worse. While a simple dressed-atom model can capture the essential features of the difference between electronic structure methods, a full machine-learning model provides more detailed information, being capable of picking up the difference in behavior for the same element in different functional groups.

TABLE II. Effective $\lambda = 0$ polarizability difference $\Delta\alpha^{(0)}_{i,\text{eff}}$ for elements in the QM7b set, as defined in Eq. (S1).

Element	B3LYP (a.u.)	SCAN0 (a.u.)
H	−0.27	−0.36
C	+1.25	+1.03
N	−0.05	−0.69
O	−0.59	−1.15
S	−1.72	−2.60
Cl	−2.16	−3.01

This atom-centered ML model of Δ-learning can also be built based on the predictions on the showcase dataset,

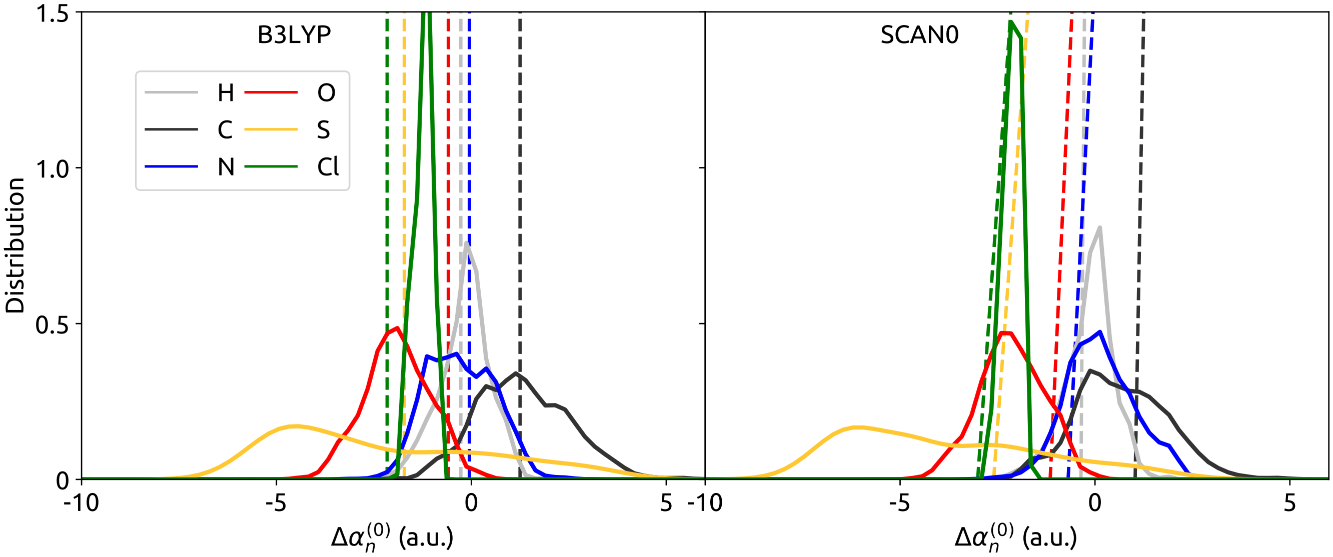

FIG. S8. Predicted contributions of each atomic species to the difference between the $\lambda = 0$ component of CCSD and B3LYP-DFT (left panel) or SCAN0-DFT (right panel) polarizabilities in the QM7b set. Vertical lines show the polarizability differences predicted by fitting to the dressed-atom model of Eq. (S1), as shown in Table II.

revealing further the sources of the discrepancies between electronic structure methods. As shown in Fig. S9, the Δ-learning predictions between CCSD and DFT systematically attribute positive (negative) contributions to the C (O) atoms. This confirms the observation made on the QM7b data set that DFT tends to overestimate the polarizability of carbon-centered groups and underestimate that of oxygen-containing moieties. Inspecting individual $\Delta\boldsymbol{\alpha}_i$ provides even more detailed insight into the differences between an accurate and a more approximate method. In the case of DFT, we observe an inherent asymmetry in the treatment of carbon atoms in different hybridization states. On one hand, DFT tends to substantially overestimate $\boldsymbol{\alpha}$ for conjugated systems such as octatetraene, an error which is attributed to an inaccurate description of delocalized π electrons. On the other hand, DFT provides relatively low errors for saturated carbon atoms, as seen in dimethylhexane and fructose. In contrast to the anisotropic and environment-dependent errors observed for carbon, the ML-based decomposition of $\Delta\boldsymbol{\alpha}$ suggests that the DFT underestimation of polarizability contributions from oxygen-containing groups is isotropic and relatively insensitive to the environment.

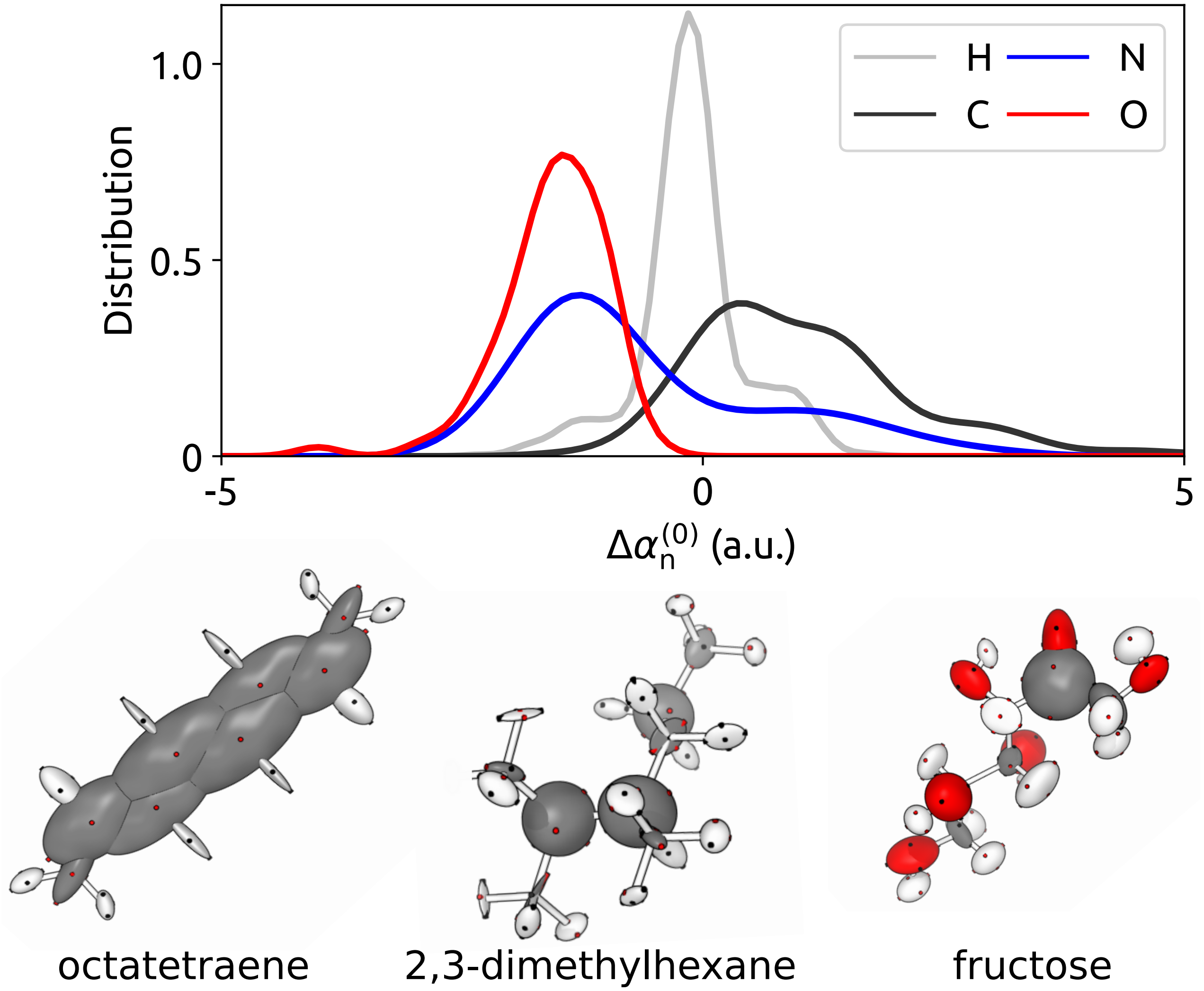

FIG. S9. Top: distributions of the predicted atomic contribution to the $\lambda = 0$ component of the difference between DFT and CCSD polarizabilities. Bottom: example decompositions of the polarizability difference. The ellipsoids represent the magnitude and principal axes of $\Delta\boldsymbol{\alpha}_i$. Black (red) axes indicate that DFT polarizability is larger (smaller) than CCSD.

DETAILS OF THE REFERENCE ELECTRONIC STRUCTURE CALCULATIONS

All calculations used the Dunning-style d-aug-cc-pVDZ basis set [S4], which was obtained from the EMSL Basis Set Library [S5, S6]. For all DFT calculations in `Q-Chem`, a tight threshold with `scf_convergence=10` and `thresh=13` was used for all molecules in the QM7b database, while `scf_convergence=8` combined with `thresh=14` was used for the molecules in the showcase dataset. For all DFT calculations in `Psi4`, the convergence criteria for the SCF density and energy were both set to 10^{-10}. For the LR-CCSD calculations, all convergence criteria were set to the default values in `Psi4`. For the finite-field CCSD calculations, SCF convergence criteria on the density and energy were both set to 5.0×10^{-10}, and a much tighter convergence criterion of 5.0×10^{-9} was used for the CC amplitudes to minimize numerical error.

BASIS SET CONVERGENCE OF THE LR-CCSD CALCULATIONS

Since d-aug-cc-pVDZ (daDZ) and aug-cc-pVTZ (aTZ) are arguably the largest Dunning-style basis sets that one can currently employ (without significant dedicated supercomputer resources) to compute the polarizabilities of every molecule in the QM7b dataset, we investigate the accuracy and convergence properties of these necessarily incomplete basis sets in this section. We begin our discussion by focusing on their performance with respect to a series of reference values provided in the literature. In Table III, we compare polarizability values for 15 different atoms, ions, and small molecules obtained with a number of different quantum chemical methods and Dunning-style basis sets (that are more converged with respect to augmentation and/or angular momentum). With mean absolute errors (mean absolute percent errors) of 0.27 (2.50%) for daDZ and 0.85 (6.15%) for aTZ, statistical error analysis indicates that daDZ outperforms aTZ in the calculation of polarizabilities. However, the relative performance of these basis sets in this rather limited dataset is dominated by the presence of highly diffuse systems such as the F^- and Cl^- anions, for which the doubly-augmented daDZ basis set yields considerably smaller errors than aTZ. Excluding these anionic systems from the statistical analysis yields mean absolute errors (mean absolute percent errors) of 0.14 (1.98%) for daDZ and 0.12 (3.33%) for aTZ, which indicates that these basis sets are able to yield atomic and molecular polarizabilities (such as those found in the QM7b dataset) of comparable quality.

These two basis sets are also comparable when considering the $\boldsymbol{\alpha}$ values corresponding to the 19 smallest molecules contained within the most diverse 100 configurations (obtained *via* FPS) of the QM7b dataset. When compared against LR-CCSD predictions of this quantity using the larger d-aug-cc-pVTZ (daTZ) basis set (*i.e.*, the largest feasible Dunning-style basis for computing the polarizabilities of molecules of this size without significant dedicated supercomputer resources), we find that the difference between LR-CCSD/daDZ and LR-CCSD/aTZ is minuscule with respect to the corresponding DFT error (see Table IV). In other words, even though the error of aTZ (RMSPE of 1.1% σ_{CCSD}) on this subset is smaller than that of daDZ (2.8%), both of these errors are inconsequential when compared to the B3LYP error (60%). Since this difference is comparable to the asymptotic accuracy of AlphaML (2-3%), additional computational effort (with the objective of building a more accurate ML model) would be better invested on increasing the size of the training set rather than using a slightly more converged basis set. We note in passing that the FPS strategy chooses fairly extreme and challenging molecules from the QM7b dataset, which thereby provides a stringent test of the quality of these basis sets; this is evidenced by the larger intrinsic variability in this subset (6.72 a.u. per atom) when compared to the entirety of the QM7b dataset (2.20 a.u. per atom).

In making the decision to employ the daDZ basis set in this work, it is also important to note that the increased computational effort and memory requirements associated with the slightly larger aTZ basis set would have prevented us from performing LR-CCSD calculations for $\sim$ 1000 of the larger QM7b structures (on the hardware at our disposal). This would have forced us to use finite-field methods to compute $\boldsymbol{\alpha}$, as was deemed necessary for the largest molecules in the showcase dataset. In the training set, this would have introduced unnecessary inconsistencies that could potentially interfere with our ability to learn the polarizability tensors. While the tests above are by no means exhaustive, they still demonstrate that the level of theory chosen for our reference polarizability calculations (*i.e.*, LR-CCSD/daDZ) is more than appropriate for obtaining substantial improvements in accuracy relative to the predictions of hybrid DFT. Future extensions to AlphaML will have to reassess the level of theory to strike the best balance between thorough sampling of chemical compound space, inclusion of larger reference molecules, and basis set convergence.

TABLE III. Performance of the d-aug-cc-pVDZ (daDZ) and aug-cc-pVTZ (aTZ) basis sets in the *ab initio* determination of the polarizabilities in a series of 15 atoms, ions, and small molecules. Statistical error analysis with respect to the reference values (and methods) indicated below includes: mean signed errors (MSE), mean absolute errors (MAE), root mean square errors (RMSE), mean signed percent errors (MSPE), mean absolute percent errors (MAPE), and root mean square percent errors (RMSPE). Error quantities dressed with an asterisk (*) denote that the anion data (F^- and Cl^-) were not included in the statistical error analysis. All polarizability values and errors are reported in a.u.

System	Property	daDZ	aTZ	Ref. Value	Ref. Method	Ref.
He	$\alpha_{\rm iso}$	1.39	1.38	1.38	CCSD/x-aug-cc-pVQZ limit	S4
Ne	$\alpha_{\rm iso}$	2.64	2.39	2.63	CCSD/q-aug-cc-pVQZ	S4
Ne	$\alpha_{\rm iso}$	2.71	2.43	2.67	CCSD/d-aug-cc-pV5Z	S7
Ne	$\alpha_{\rm iso}$	2.68	2.42	2.66	CC3/d-aug-cc-pV6Z	S8
Ne	$\alpha_{\rm iso}$	2.67	2.42	2.68	CCSD(T)/q-aug-cc-pVQZ	S4
Ar	$\alpha_{\rm iso}$	10.96	10.81	11.16	MP2/x-aug-cc-pVQZ limit	S4
Ar	$\alpha_{\rm iso}$	11.05	10.84	11.12	CCSD(T)/x-aug-cc-pVQZ limit	S4
F^-	$\alpha_{\rm iso}$	14.52	8.89	16.73	MP2/x-aug-cc-pVTZ limit	S4
F^-	$\alpha_{\rm iso}$	14.84	8.77	17.15	CCSD(T)/x-aug-cc-pVTZ limit	S4
Cl^-	$\alpha_{\rm iso}$	35.73	27.81	37.09	MP2/x-aug-cc-pVTZ limit	S4
Cl^-	$\alpha_{\rm iso}$	36.63	28.11	37.43	CCSD(T)/x-aug-cc-pVTZ limit	S4
N_2	α_{xx}	10.25	10.16	10.19	MP2/x-aug-cc-pVTZ limit	S4
N_2	α_{zz}	14.79	14.41	14.45	MP2/x-aug-cc-pVTZ limit	S4
N_2	α_{xx}	10.19	10.13	10.08	CCSD/d-aug-cc-pV5Z	S7
N_2	α_{zz}	14.74	14.58	14.52	CCSD/d-aug-cc-pV5Z	S7
N_2	$\alpha_{\perp}$	10.19	10.13	10.11	CCSD/d-aug-cc-pVQZ	S9
N_2	$\alpha_{\parallel}$	14.74	14.58	14.55	CCSD/d-aug-cc-pVQZ	S9
N_2	α_{xx}	10.42	10.26	10.29	CCSD(T)/x-aug-cc-pVTZ limit	S4
N_2	α_{zz}	15.44	14.96	14.99	CCSD(T)/x-aug-cc-pVTZ limit	S4
N_2	$\alpha_{\perp}$	10.26	10.19	10.22	CCSDT/d-aug-cc-pVTZ	S9
N_2	$\alpha_{\parallel}$	14.94	14.75	14.78	CCSDT/d-aug-cc-pVTZ	S9
CO	α_{xx}	11.91	11.82	11.79	CCSD/d-aug-cc-pV5Z	S7
CO	α_{zz}	15.91	15.70	15.57	CCSD/d-aug-cc-pV5Z	S7
BH	α_{xx}	20.76	21.01	21.04	CCSD/d-aug-cc-pV5Z	S7
BH	α_{zz}	23.33	22.90	22.75	CCSD/d-aug-cc-pV5Z	S7
CH^+	α_{xx}	6.96	7.05	7.06	CCSD/d-aug-cc-pV5Z	S7
CH^+	α_{zz}	8.44	8.34	8.27	CCSD/d-aug-cc-pV5Z	S7
HF	α_{xx}	5.29	4.90	5.17	CCSD/d-aug-cc-pV5Z	S7
HF	α_{zz}	6.46	6.35	6.34	CCSD/d-aug-cc-pV5Z	S7
HF	α_{xx}	5.23	4.89	5.19	CC3/d-aug-cc-pV5Z	S8
HF	α_{zz}	6.40	6.33	6.33	CC3/d-aug-cc-pV5Z	S8
HCl	α_{xx}	16.70	16.30	16.67	MP2/x-aug-cc-pVTZ limit	S4
HCl	α_{zz}	18.56	18.15	18.39	MP2/x-aug-cc-pVTZ limit	S4
HCl	α_{xx}	16.86	16.45	16.86	CCSD(T)/x-aug-cc-pVTZ limit	S4
HCl	α_{zz}	18.69	18.28	18.53	CCSD(T)/x-aug-cc-pVTZ limit	S4
H_2O	$\alpha_{\rm iso}$	9.79	9.48	9.56	CCSD/t-aug-cc-pV5Z	S10
H_2O	$\alpha_{\rm aniso}$	0.45	0.78	0.59	CCSD/t-aug-cc-pV5Z	S10

TABLE III. Continued

System	Property	daDZ	aTZ	Ref. Value	Ref. Method	Ref.
H_2O	α_{iso}	9.72	9.47	9.60	CCSDT/t-aug-cc-pVQZ	S10
H_2O	α_{aniso}	0.44	0.74	0.53	CCSDT/t-aug-cc-pVQZ	S10
C_2H_4	α_{xx}	24.89	24.73	24.88	CC3/t-aug-cc-pVTZ	S8
C_2H_4	α_{yy}	22.26	21.89	21.93	CC3/t-aug-cc-pVTZ	S8
C_2H_4	α_{zz}	34.12	34.03	34.04	CC3/t-aug-cc-pVTZ	S8
CH_3CN	α_{xx}	24.48	24.36	24.45	MP2/d-aug-cc-pVTZ	S11
CH_3CN	α_{yy}	38.88	38.83	38.84	MP2/d-aug-cc-pVTZ	S11
CH_3CN	α_{zz}	29.28	29.18	29.25	MP2/d-aug-cc-pVTZ	S11
C_6H_6	α_{LL}	80.53	80.26	80.35	CCSD/d-aug-cc-pVTZ	S12
C_6H_6	α_{NN}	45.00	44.51	44.49	CCSD/d-aug-cc-pVTZ	S12
	MSE (MSPE)	−0.05 (−0.82%)	−0.81 (−2.90%)			
	MAE (MAPE)	0.27 (2.50%)	0.85 (6.15%)			
	RMSE (RMSPE)	0.55 (5.23%)	2.55 (13.78%)			
	MSE* (MSPE*)	0.10 (−0.15%)	−0.08 (0.22%)			
	MAE* (MAPE*)	0.14 (1.98%)	0.12 (3.33%)			
	RMSE* (RMSPE*)	0.20 (4.60%)	0.17 (8.48%)			

TABLE IV. Performance of the daDZ and aTZ basis sets in the determination of the polarizability ($\boldsymbol{\alpha}$) of the 19 smallest molecules contained within the first 100 FPS structures in the QM7b dataset. All statistical errors were computed with respect to the $\boldsymbol{\alpha}$ values obtained using LR-CCSD and the d-aug-cc-pVTZ (daTZ) basis set, and reported as a percentage of the intrinsic variability of the full QM7b dataset ($\sigma_{\text{CCSD}} = 2.20$ a.u. per atom) as defined in the main text.

Test Set	Ref. Set	MSPE	MAPE	RMSPE
CCSD/daDZ	CCSD/daTZ	1.39%	2.59%	2.83%
CCSD/aTZ	CCSD/daTZ	−0.70%	1.01%	1.10%
B3LYP/daDZ	CCSD/daTZ	39.70%	42.57%	60.29%

[S1] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).

[S2] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).

[S3] B. Schölkopf, A. Smola, and K.-R. Müller, Neur. Comp. **10**, 1299 (1998).

[S4] D. E. Woon and T. H. Dunning Jr., J. Chem. Phys. **100**, 2975 (1994).

[S5] D. Feller, J. Comput. Chem. **17**, 1571 (1996).

[S6] K. L. Schuchardt, B. T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T. L. Windus, J. Chem. Inf. Model. **47**, 1045 (2007).

[S7] O. Christiansen, C. Hättig, and J. Gauss, J. Chem. Phys. **109**, 4745 (1998).

[S8] K. Hald, F. Pawowski, P. Jørgensen, and C. Hättig, J. Chem. Phys. **118**, 1292 (2003).

[S9] J. R. Hammond, W. A. de Jong, and K. Kowalski, J. Chem. Phys. **128**, 224102 (2008).

[S10] J. R. Hammond, N. Govind, K. Kowalski, J. Autschbach, and S. S. Xantheas, J. Chem. Phys. **131**, 214103 (2009).

[S11] H. Reis, M. G. Papadopoulos, and A. Avramopoulos, J. Phys. Chem. A **107**, 3907 (2003).

[S12] J. R. Hammond, K. Kowalski, and W. A. de Jong, J. Chem. Phys. **127**, 144105 (2007).