JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

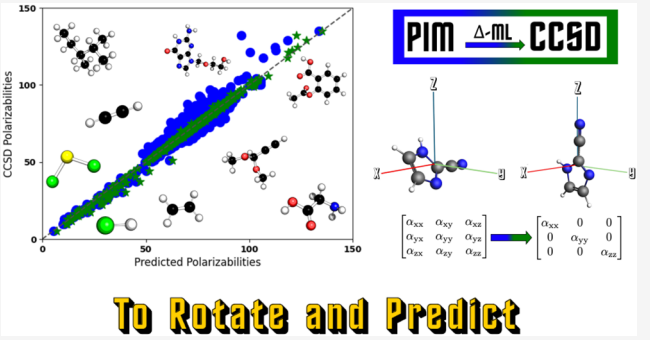
# Δ-Machine Learning of Polarizability Tensors Using a Dipole Interaction Model

Imran Chaudhry, Mark J. Bronson, Jr., and Lasse Jensen*

Cite This: J. Chem. Theory Comput. 2025, 21, 7040−7050 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** As a fundamental response property, the molecular polarizability is responsible for a wide variety of physical phenomena relevant for understanding light-matter and intermolecular interactions. Therefore, it is important to have accurate and efficient methods to calculate the polarizability tensor. In this work, we introduce a model that combines a polarizable dipole interaction model (PIM) with Δ-machine learning to predict polarizability tensors, which we refer to as $\Delta_{PIM}^{CCSD}$. To describe the rotational symmetry, we adapt a unique reference geometry obtained by diagonalizing the PIM polarizability tensor. A major benefit is that only the diagonal elements of the polarizability tensor needs to be learned. The model was parameterized to the coupled cluster singles and doubles (CCSD) polarizabilities from the QM7b data set and was used to predict the polarizabilities for various systems, such as small molecules and molecules from the QM9 data set. We show that the $\Delta_{PIM}^{CCSD}$ is comparable in accuracy to density-functional theory with the B3LYP exchange correlation functional (DFT/B3LYP) at a lower cost for molecules with similar chemical composition as the QM7b data set. For the QM9 data set, this was also found, although only after correcting for the smaller basis set used for calculating the polarizabilities in this data set. For molecules smaller and chemically more diverse than the training set, we find that the model performs worse than DFT/B3LYP. Ultimately, our work suggests that larger and more chemically diverse data sets with polarizabilities obtained at a high level of theory are needed. Finally, our results suggest that the model can be improved by incorporating atom-specific polarizabilities into PIM to better account for local environments. In summary, the combination of PIM with Δ-machine learning provides a simple and promising approach for predicting polarizability tensors.




## 1. INTRODUCTION

The polarizability describes the response of a molecule to an external electric field, and thus is central to understanding light-matter and intermolecular interactions. For instance, Raman scattering is determined by changes in the polarizability due to molecular vibrations.[1−3] Another example comes from Rayleigh scattering of small particles in the atmosphere, where the intensity is proportional to the square of the polarizability of the particle.[4−6] Furthermore, the description of intermolecular interactions has been shown to improve by including polarization through the use of polarizable force fields[7−11] and quantum mechanical/molecular mechanical (QM/MM) methods.[12−15] However, calculating accurate polarizabilities typically requires electronic structure methods and large basis sets, which is computationally expensive.[16,17] For that reason, there is significant effort put into devising methods that can both produce polarizabilities with sufficient accuracies at low computational cost.[18−20]

Semiempirical methods have been used to calculate polarizabilities at a lower computational cost than *ab initio* methods.[20−25] These methods achieve their efficiency by adapting parameterized integrals in combination with minimal basis sets.[18,26] However, the use of minimal basis sets leads to a sacrifice in accuracy that manifests as an underestimation of molecular polarizabilities, especially in describing the anisotropy.[18,27−29] Alternatively, dipole interaction methods utilize classical electrostatics by treating atoms as a set of interacting induced point dipoles.[30,31] These methods are parameterized to a training set consisting of either experimental or *ab intio* calculated molecular polarizabilities. However, these methods have trouble accurately describing the polarizabilities charged species as well as highly conjugated systems due to adapting atom type parameters.[32−37] The main advantage of these methods is that they can be applied to describe the polarizability of larger systems.[38−41]

Received: May 9, 2025
Revised: June 26, 2025
Accepted: July 1, 2025
Published: July 9, 2025







https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

An alternative to physics-based methods that has emerged is machine learning models which have been shown to be more efficient than *ab initio* methods and thus has shown great promise for modeling molecular polarizabilities. Machine learning (ML)[42−45] has emerged as a popular approach to improve computational efficiency[46−52] and thus shows great promise for modeling molecular polarizabilities.[53−60] To model molecular polarizabilities with ML, these models have to account for the orientation of the molecule in order to describe the rotational symmetry of the rank-2 tensor. One approach combines a symmetry-adapted Gaussian process regression model with the smooth overlap of atomic positions (SOAP) representation such that the kernel explicitly accounts for the rotational symmetry.[61] This $\lambda$-SOAP kernel has demonstrated its ability to model the full polarizability tensor with better accuracy than hybrid DFT for a set of larger organic molecules.[54] Other approaches exist that preserve the rotational symmetry of the polarizability tensor, such as mapping molecules to a reference geometry[62] or to its canonical orientation using the tensor of inertia,[55] expanding the polarizability tensor in terms of the atomic spherical harmonics,[58] or by using equivariant neural networks to encode the rotational symmetry of the system.[58,63,64] Besides preserving rotational symmetry, one can also use ML to predict the atomic polarizabilities, which can be incorporated into a physical model such as a dipole interaction model to describe the polarizability tensor.[65] One major benefit of using ML models is that they can be used to generate Raman spectra by predicting the polarizability tensor of a system of interest, thereby lowering the computational cost.[55−57,64,66−69] Another benefit of these models is that they can be used to calculate the polarizability of large systems where *ab initio* approaches can become prohibitively expensive and have been applied to such systems as silicon,[70] water,[63] and gold clusters.[49] A different approach to predicting polarizabilities is based on parameterizing the ML model to the difference in computational methods, where one method is more accurate and expensive and the other is cheaper and less accurate, called the Δ-ML model.[71] This method can be used to achieve more accurate descriptions of molecular properties at a lower computational cost and has been shown to give good predictions on systems that are outside of the training set.[72−79] For instance, a Δ-ML model was used to correct DFT/B3LYP polarizability tensors to the level of CCSD accuracy, which had a lower overall error when compared to the direct ML model parameterized to the CCSD polarizability tensors.[54] In a separate study, a Δ-ML model was used to simulate the Raman spectra of various systems, such as molecules and extended solids, from molecular dynamics simulations.[57] The Δ-ML model used was trained on the difference in the polarizabilities from a parameterized linear-response model and density-functional perturbation theory calculations. This Δ-ML model was shown to provide Raman spectra of similar quality to DFT and gave more accurate predictions than the direct ML model.[57]

In this work, we demonstrate how a Δ-ML model can use a dipole interaction model to learn the polarizability tensor of organic molecules. A major advantage of combining these models is that the dipole interaction model offers a unique reference geometry obtained by diagonalizing the polarizability tensor. This further has the benefit that only the diagonal tensor elements has to be learned. The model was parameterized to the QM7b data set and was used to predict the polarizabilities for various systems, such as small organic molecules and molecules from the QM9 data set. Overall, we show that a simple Δ-ML model can be used to predict the polarizability tensor.

## 2. COMPUTATIONAL METHODS AND DETAILS

**2.1. Polarizability Interaction Model.** In the polarizability interaction model (PIM), each atom, $i$, in a system of N-interacting atoms is characterized by an isotropic atom-type polarizability $\alpha_{i,\alpha\beta} = \delta_{\alpha\beta}\alpha_i$, where the Greek indices represent Cartesian coordinates.[31] In PIM, the molecular polarizability is defined as

$$\alpha_{\alpha\beta}^{\mathrm{mol}} = \sum_{i,j} B_{ij,\alpha\beta} \tag{1}$$

where the $B_{ij,\alpha\beta}$ elements are found by inverting the interaction matrix

$$\boldsymbol{\mu}^{\mathrm{ind}} = \mathbf{B}\mathbf{E}^{\mathrm{ext}} = \mathbf{A}^{-1}\mathbf{E}^{\mathrm{ext}} \tag{2}$$

The elements of the interaction matrix, **A**, are defined as

$$A_{ij,\alpha\beta} = \begin{cases} \alpha_{i,\alpha\beta}^{-1} & \text{for} \quad i = j \\ -T_{ij,\alpha\beta}^{(2)} & \text{for} \quad i \neq j \end{cases} \tag{3}$$

where $T_{ij,\alpha\beta}^{(2)}$ is an element of the screened dipole−dipole interaction tensor[80]

$$T_{ij,\alpha\beta}^{(2)} = \frac{3r_{ij,\alpha}r_{ij,\beta} - \delta_{\alpha\beta}r_{ij}^2}{r_{ij}^5}\left[\mathrm{erf}\left(\frac{r_{ij}}{R_{\mu\mu}^{ij}}\right) - \frac{2}{\sqrt{\pi}}\frac{r_{ij}}{R_{\mu\mu}^{ij}}\mathrm{e}^{-(r_{ij}/R_{\mu q\mu}^{ij})^2}\right] - \left[\frac{4}{\sqrt{\pi}(R_{\mu\mu}^{ij})^3}\frac{r_{ij,\alpha}r_{ij,\beta}}{r_{ij}^2}\mathrm{e}^{-(r_{ij}/R_{\mu\mu}^{ij})^2}\right] \tag{4}$$

Here, $r_{ij}$ is the distance between atoms $i$ and $j$, and $R_{\mu\mu}^{ij}$ is the screening radius defined as $R_{\mu\mu}^{ij} = (R_\mu^i)^2 + (R_\mu^j)^2$ with $R_\mu^i$ given in terms of the atomic polarizability as[39]

$$R_\mu^i = \left(\sqrt{\frac{2}{\pi}}\frac{\alpha_i}{3}\right)^{1/3} \tag{5}$$

The atomic polarizabilites are parameterized to the CCSD polarizabilities taken from the QM7b data set.[54] This is done by minimizing the root mean squared difference between the PIM and CCSD polarizabilities as

$$\mathrm{error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha}\left(\frac{\alpha_{\alpha\alpha}^{\mathrm{PIM},i} - \alpha_{\alpha\alpha}^{\mathrm{CCSD},i}}{\alpha_{\alpha\alpha}^{\mathrm{CCSD},i}}\right)^2} \tag{6}$$

using simulated annealing.[81] The optimized atomic parameters are given in the Supporting Information.

**2.2. Rotationally Invariant Polarizability Tensors.** To describe rotationally invariant polarizability tensors, we will adapt a reference geometry obtained by diagonalizing the PIM polarizability tensor as


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

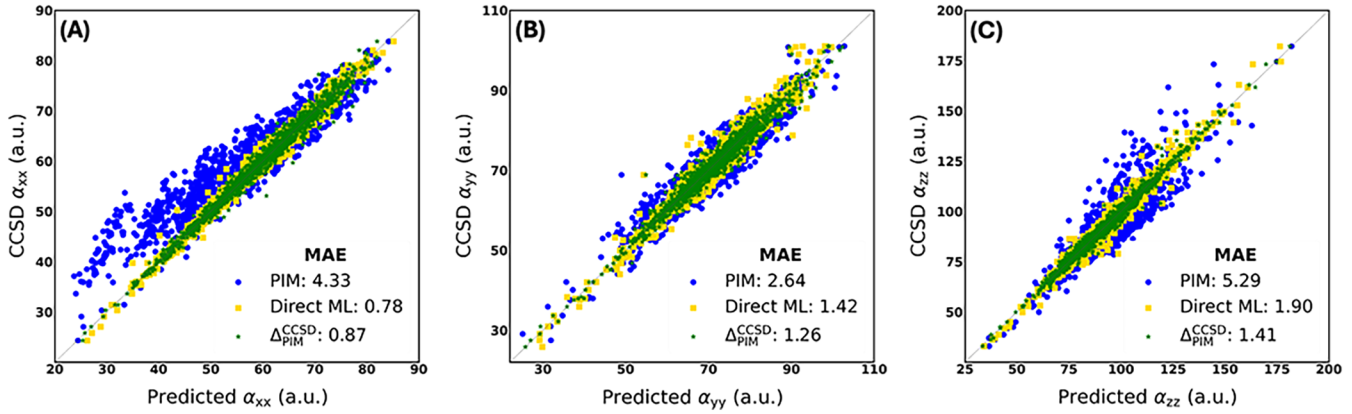



**Figure 1.** Comparison between the molecular polarizabilities predicted by PIM (blue circle), direct ML (gold square), and $\Delta^{CCSD}_{PIM}$ (green star) on the QM7b testing set. (A)−(C) show model predictions with respect to each tensor element (*xx*, *yy*, *zz*).

$$\mathbf{Q}^{\dagger}\begin{bmatrix} \alpha_{xx}^{PIM}(\mathbf{R}) & \alpha_{xy}^{PIM}(\mathbf{R}) & \alpha_{xz}^{PIM}(\mathbf{R}) \\ \alpha_{yx}^{PIM}(\mathbf{R}) & \alpha_{yy}^{PIM}(\mathbf{R}) & \alpha_{yz}^{PIM}(\mathbf{R}) \\ \alpha_{zx}^{PIM}(\mathbf{R}) & \alpha_{zy}^{PIM}(\mathbf{R}) & \alpha_{zz}^{PIM}(\mathbf{R}) \end{bmatrix}\mathbf{Q} = \begin{bmatrix} \alpha_{xx}^{PIM}(\mathbf{R}_{ref}) & 0 & 0 \\ 0 & \alpha_{yy}^{PIM}(\mathbf{R}_{ref}) & 0 \\ 0 & 0 & \alpha_{zz}^{PIM}(\mathbf{R}_{ref}) \end{bmatrix} \tag{7}$$

where $\mathbf{Q}$ is the rotation matrix that sorts the diagonal tensor such that the *xx*-element has the smallest value and the *zz*-element has the largest value. The reference geometry, $\mathbf{R}_{ref}$, is then obtained by rotating the geometry using $\mathbf{R}_{ref} = \mathbf{RQ}$. An advantage of adapting this specific reference geometry is that it will only be necessary to machine learn the three diagonal components rather than the six components of the full polarizability tensor. This simplifies the ML problem although it introduces the requirement that a set of polarizability tensors needed to generate the reference geometries are available. However, within the context of Δ-ML, this is naturally provided by the lower level method.

Since an independent Δ-ML model will be used for each diagonal component, the polarizability tensor will not reflect the point group symmetry of the molecule. Therefore, we will use the symmetry of the diagonalized PIM polarizability tensor to symmetrize the Δ-ML model. Specifically, if all three diagonal elements of the PIM tensor are equal, we will average the three predictions. Similarly, if two PIM tensor elements are equal, then the corresponding predictions will be averaged. The tensor is considered symmetric if the polarizability tensor elements are similar within a threshold of 0.01 atomic units.

The molecular polarizability is often characterized by two rotational invariant quantities. Namely, the isotropic polarizability, given by

$$\alpha^{iso} = \frac{1}{3}[\alpha_{xx} + \alpha_{yy} + \alpha_{zz}] \tag{8}$$

and the anisotropic polarizability

$$\alpha^{aniso} = \frac{\sqrt{2}}{2}\{[\alpha_{xx} - \alpha_{yy}]^2 + [\alpha_{yy} - \alpha_{zz}]^2 + [\alpha_{zz} - \alpha_{xx}]^2\}^{1/2} \tag{9}$$

which here is given in a form that assumes the polarizability tensor is diagonal.[82]

**2.3. The Δ-Machine Learning Model.** To describe the molecular polarizability using a Δ-ML model, we approximate the diagonal elements of the polarizability as

$$\alpha_{\alpha\alpha}^{Pred}(\mathbf{x}) = \alpha_{\alpha\alpha}^{PIM}(\mathbf{x}) + \sum_{i=1}^{N} \gamma_i^{\alpha\alpha} K(\mathbf{x}, \mathbf{x}_i) \tag{10}$$

where $\gamma_i^{\alpha\alpha}$ is a regression coefficient for the $\alpha\alpha$ polarizability element, $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel, $\mathbf{x}_i$ is the feature vector of a molecule in the training set, and $\mathbf{x}$ represents the feature vector of the molecule of interest. The model has been parameterized using kernel ridge regression (KRR) as implemented in the sklearn software package.[83] The kernel used is the Gaussian kernel

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}||^2}{\sigma^2}\right) \tag{11}$$

which uses the Euclidean norm and the variance of the kernel, $\sigma^2$, to characterize the distance between the feature vectors. We have also tested the Laplacian kernel, but only small differences in model performance were found.

To represent the molecular geometry, we used the Many-Body Tensor Representation (MBTR)[84] as implemented in the DScribe 2.1.x python package.[85,86] We described the geometry using the inverse distance between two atoms ($k$ = 2) as well as applying an exponential decay $\exp(-sx)$, where $x$ is the pairwise distance between atoms, and $s$ is a hyperparameter. Previous work has shown that $k$ = 2 representation is able to capture most of the relevant geometric information.[87] To optimize the hyperparameters, namely the regularization, $\lambda$, the kernel width, $\sigma$, MBTR decay, $s$, and the MBTR width, $\omega$, we utilize a random search algorithm with 300 iterations.[88] In each iteration, the selected hyperparameters are tested using 5-fold cross-validation over different ranges: $\lambda \in [1, 10^{-10}]$, $\sigma \in [10^{-3}, 10^{-10}]$, $s \in [0.3, 1.5]$, and $\omega \in [1, 10^{-3}]$. From the training set, 40% of the molecules were randomly chosen for use in hyperparameter selection. The optimal hyperparameters


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

Table 1. Mean Absolute Polarizability Errors for Three Datasets in a.u.

| | QM7b | | | | | showcase | | | small molecules | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| model | $\alpha_{xx}$ | $\alpha_{yy}$ | $\alpha_{zz}$ | $\alpha^{iso}$ | $\alpha^{aniso}$ | $\alpha_{xx}$ | $\alpha_{yy}$ | $\alpha_{zz}$ | $\alpha_{xx}$ | $\alpha_{yy}$ | $\alpha_{zz}$ |
| PIM | 4.33 | 2.64 | 5.29 | 2.40 | 5.92 | 5.51 | 3.22 | 7.27 | 2.84 | 3.08 | 4.81 |
| $\Delta^{CCSD}_{PIM}$ | 0.86 | 1.26 | 1.41 | 0.27 | 1.51 | 2.09 | 2.44 | 3.27 | 3.39 | 3.68 | 3.89 |
| direct ML | 0.78 | 1.42 | 1.90 | 0.42 | 1.89 | 2.83 | 3.13 | 6.11 | 6.94 | 8.08 | 13.1 |
| DFT/B3LYP | 0.70 | 1.30 | 3.88 | 1.91 | 3.03 | 1.11 | 1.43 | 4.63 | 0.84 | 0.85 | 1.01 |
| $\alpha^{ML}$ | | | | | | 1.57 | 1.77 | 1.99 | 5.18 | 4.97 | 5.53 |

for each model are given in the Supporting Information. To train our Δ-ML model, we used the QM7b data set, which consists of 7211 molecules containing up to seven non-hydrogen, or "heavy", atoms (i.e., C, N, O, S, Cl).[16,54,89,90] The polarizabilities for this data set were calculated using linear response coupled cluster singles and doubles (LR-CCSD) with the d-aug-cc-pVDZ basis set. The data set was randomly split into a training and testing set, with 80% of the data set in the training set and 20% of the data set in the testing set. A Δ-ML model for each diagonal polarizability component was fit to the differences in PIM and CCSD polarizabilities. In optimizing the hyperparameters, we chose to run the random search procedure three times, after which the model was chosen that best performed on the QM7b testing set.

To test our Δ-ML model we used the $\alpha^{ML}$ showcase molecules, which consists of 52 drug-like molecules ranging from 16 to 27 atoms, for which polarizability tensors at the CCSD level are available.[16,54] We also tested our model on a small molecule data set consisting of 47 molecules with polarizabilities calculated at the coupled cluster singles and doubles with approximated triples (CCSD(T)) level using a combination of the aug-cc-pCVQZ/5Z and aug-cc-pCVTZ/QZ basis sets.[91] This data set is derived from a larger data set consisting of 132 molecules that was then filtered to exclude atomtypes dissimilar to the QM7b data set. Lastly, we tested our model on a subset of the QM9 data set,[92,93] where fluorinated molecules were excluded. For simplicity, we will refer to this reduced data set containing 131k molecules as the QM9r data set. The molecules consist of up to nine heavy atoms (i.e., C, N, O) with DFT isotropic polarizabilities calculated using B3LYP exchange-correlation functional and the 6−31G(2df,p) basis set.

## 3. RESULTS AND DISCUSSION

**3.1. Learning the Polarizabilities from the QM7b Data Set.** Figure 1 shows the comparison between PIM, direct ML, and Δ-ML predictions relative to the CCSD diagonal tensor elements of the QM7b testing set. The CCSD diagonal tensor elements were obtained by diagonalizing the reported polarizability tensors.[16,54] The mean absolute errors (MAE) for each model are collected in Table 1. Although PIM was parameterized against the full QM7b data set, we assess the error only on the QM7b testing set to be consistent with the other models. For context, the training set consists of 5769 molecules whereas the testing set consists of 1442 molecules. From the figure, we see that PIM provides a reasonable description of the polarizabilities, though the predictions of $\alpha_{xx}$ shown in Figure 1(A) has a noticeably larger deviation. The prediction error for the polarizability is 4.33, 2.64, and 5.29 atomic units (a.u.) for the *xx*, *yy*, and *zz* components, respectively. Although the MAE for the zz component is the largest, it simply reflects that the molecules are oriented with the largest polarizability component in that direction. For comparison, the MAE of DFT/B3LYP on the testing set is 0.70, 1.30, and 3.88 a.u. for the *xx*, *yy*, and *zz* components, respectively. Thus, PIM has an error that is roughly twice that of DFT/B3LYP except for the *xx* component, which is more than a factor of 5 larger. This likely reflects that PIM tends to overestimate the out-of-plane polarization by favorably describe the in-plane polarization.

Further, to assess the performance of PIM, we compared it to previously reported parametrization of dipole interaction models.[65,94] The first model is the SMARTS-based dipole interaction model.[94] This model was trained on 700 molecules from the QM7b data set, and was tested on 6k molecules, yielding a root mean squared error (RMSE) of 5.26 a.u. for the diagonal elements of the polarizability tensor. For comparison, PIM has an RMSE of 5.47 a.u. for the diagonal elements of the polarizability tensors fit to the entire QM7b data set. The difference in performance can be attributed to the use of atomic polarizabilities that take into account the local environment based on SMARTS patterns, whereas PIM simply uses one parameter for each element. Another dipole-interaction model is the D-REANN model,[65] which uses a neural network to predict the chemical environment-dependent atomic polarizability. The RMSE for the D-REANN model trained on over 5k molecules from the QM7b data set is 0.88 a.u., which is significantly better than PIM. However, this can simply be attributed to the use of an optimized neural network for predicting the atomic polarizabilities for a given molecule.

We parameterized a Δ-ML model on the difference between PIM and CCSD polarizability tensor elements, which we refer to as the $\Delta^{CCSD}_{PIM}$ model. It can be seen in Figure 1 that the $\Delta^{CCSD}_{PIM}$ model gives reasonable predictions for the *xx*, *yy*, and *zz* polarizability tensor elements. From the figure, it is clear that the $\Delta^{CCSD}_{PIM}$ model provides a more accurate description of the polarizabilities than PIM. Specifically, $\Delta^{CCSD}_{PIM}$ has MAEs of 0.86, 1.26, and 1.41 a.u. for the *xx*, *yy*, and *zz* components, respectively. Figure 1(A) shows that, compared to PIM, the $\Delta^{CCSD}_{PIM}$ model reduces the error of the *xx* component, which is done by about a factor of 5. Furthermore, Figure 1(B),(C) show the *yy* and *zz* component errors are reduced, which is done by around a factor of 2 and four, respectively. Most notably, we see the largest improvement in the *xx* and *zz* components, meaning that the ML model overcomes the intrinsic shortcomings of PIM. Compared to DFT/B3LYP, the $\Delta^{CCSD}_{PIM}$ performs similarly except for the *zz* component, where the error is lowered by around a factor of 2. The larger error in the *zz* component obtained by DFT/B3LYP is due to an overestimation of molecules with larger polarizability values. This can be attributed to DFT tending to overestimate the polarizability for larger conjugated molecules due to the delocalization error.[95,96]

Further comparison was also made with a direct-ML model that was parameterized to the QM7b training set. This model was trained using the CCSD polarizabilities on the same


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

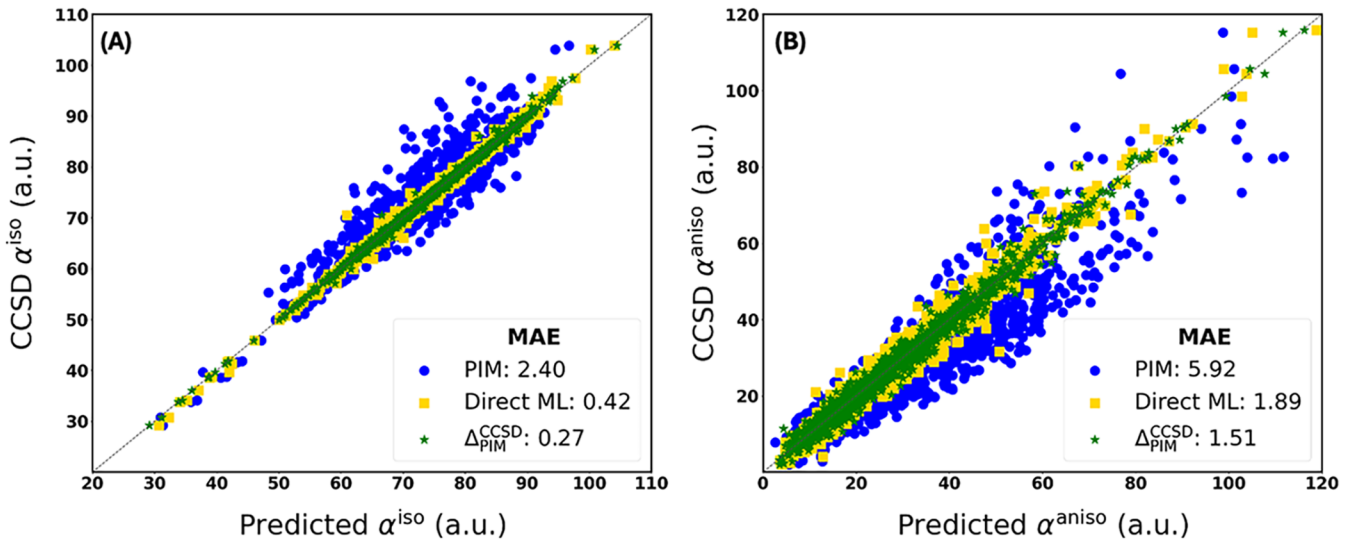



**Figure 2.** Comparison between molecular polarizabilities predicted by PIM (blue circle), direct ML (gold square), and $\Delta_{PIM}^{CCSD}$ (green star) on the QM7b testing set. (A) and (B) show model predictions with respect to isotropic and anisotropic polarizabilities, respectively.

training set used to parametrize the $\Delta_{PIM}^{CCSD}$ model. Here, we used the geometries corresponding to the diagonalized CCSD polarizabilities to train the direct-ML model. Thus, this model will not in general describe the rotational symmetry of the polarizability tensor as it requires the knowledge of the CCSD polarizability tensor beforehand. Figure 1 shows that direct ML offers a similar description to the $\Delta_{PIM}^{CCSD}$ model, with MAEs of 0.78, 1.42, and 1.90 a.u. for the *xx*, *yy*, and *zz* components, respectively. The direct ML model has the lowest MAE for the *xx* component, but the $\Delta_{PIM}^{CCSD}$ model has lower errors for the yy and zz components. Thus, it can be seen that the $\Delta_{PIM}^{CCSD}$ model only provides a slight improvement over the direct ML model. This points to that there are unsystematic errors in the PIM model that prevents a easier learning of the polarizabilities. This is well-known in Δ-ML models, where the quality of the underlying model determines the learning ability.[71] However, the major advantages of the $\Delta_{PIM}^{CCSD}$ model is that it only needs to learn the three diagonal components as well as accounting for the rotational symmetry. Since each component is learned independently, the predictions from the $\Delta_{PIM}^{CCSD}$ model could, in principle, be skewed for molecules with near-degenerate polarizability components. To examine this, we retrained the model by removing those molecules that had polarizability components with 95−100% similarity. This removed about 900 molecules from the training set. The retrained model performed similarly to the original $\Delta_{PIM}^{CCSD}$ model, indicating that this issue is minor. To further characterize the predictions of the models for the different components, we also considered a normalized error metric. The error metric chosen was the RMSE divided by the standard deviation of the reference polarizabilities, which can be found in the Supporting Information. This normalized error metric describes the prediction error relative to the natural variation of the reference data for each component, thus providing a direct indication of model performance. From the data, we see that model performance across components becomes more similar. This suggests that the different MAE for the each polarizability component simply reflects the natural variance of the data sets.

Because the model has been parameterized to the three components independently, it becomes important to quantify the relationship among the components. To do this we characterize the model predictions of the iso- and anisotropic polarizabilities, which can be seen in Figure 2. The errors for these polarizabilities can be seen in Table 1. It can be seen in Figure 2(A) that both $\Delta_{PIM}^{CCSD}$ and direct ML models perform comparably well in predicting the isotropic polarizabilities, whereas PIM tends to deviate more. In particular, PIM tends to underestimate the isotropic polarizabilities. We find that $\Delta_{PIM}^{CCSD}$ lowers the error of PIM by an order of magnitude and performs slightly better than the direct ML model. Furthermore, in Figure 2(B) we see that there is a greater deviation of the anisotropic polarizability using all models and especially using PIM. From the figure, we see that PIM overestimates the anisotropic polarizabilities, especially for molecules with large anisotropic polarizabilities. Both $\Delta_{PIM}^{CCSD}$ and direct ML models improves upon the anisotropic polarizability by a factor of about three to four, with $\Delta_{PIM}^{CCSD}$ again performing slightly better than direct ML. Compared to DFT/B3LYP, we note that both direct ML and $\Delta_{PIM}^{CCSD}$ performs better, especially for predicting the isotropic polarizabilities. As mentioned above, the worse performance of B3LYP is due to the overestimation of the *zz* component of the polarizability tensor. This leads to a larger error especially in the anisotropic polarizability. The improved isotropic polarizability for the direct ML and $\Delta_{PIM}^{CCSD}$ models must imply that some polarizability components are overestimated, while others are underestimated. We find that the direct ML model tends to overestimate the *yy* and *zz* component of the polarizability and underestimate the *xx* component, whereas the $\Delta_{PIM}^{CCSD}$ model overpredicts the yy component while underpredicting the *xx* and *zz* components. This highlights that learning the isotropic polarizability is easier and thus a better assessment is found by considering the individual components or the anisotropic polarizability. Overall, we find that $\Delta_{PIM}^{CCSD}$ is comparable to DFT/B3LYP at a lower cost. Although comparable to a direct ML model, $\Delta_{PIM}^{CCSD}$ naturally incorporates the rotational symmetry of the polarizability through the diagonalization/rotation scheme, thereby providing a simple ML framework for predicting polarizability tensors.


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

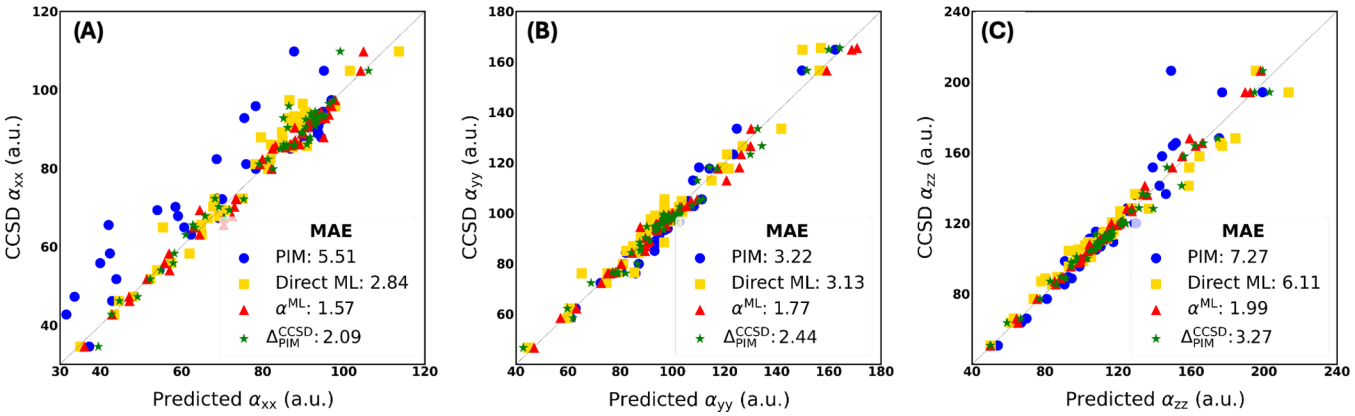



**Figure 3.** Comparison of the molecular polarizabilities predicted by PIM, direct ML, $\Delta^{CCSD}_{PIM}$, and $\alpha^{ML}$ (red triangle) on the showcase data set. (A)–(C) show model predictions with respect to each tensor element.

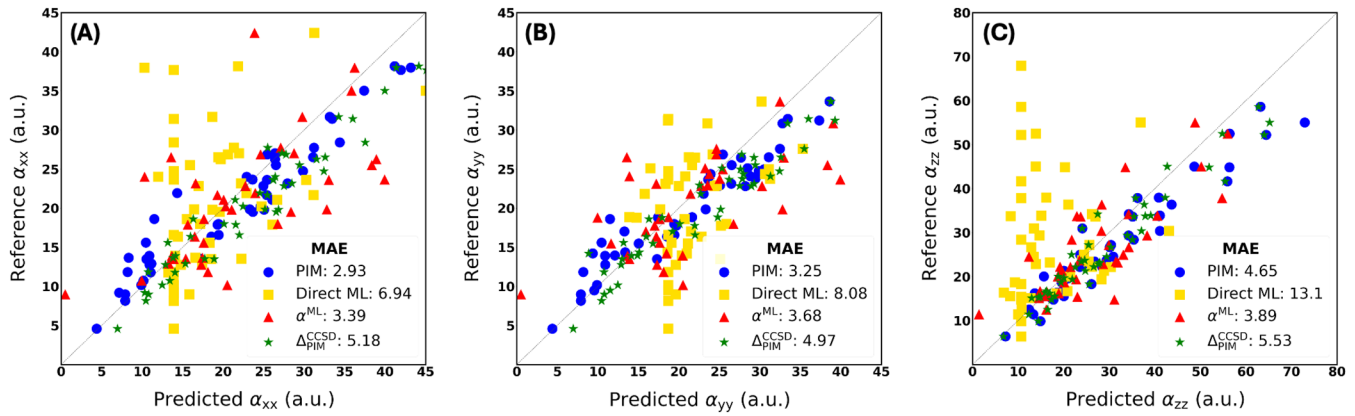



**Figure 4.** Comparison between polarizability tensor elements predicted by PIM, direct ML, $\Delta^{CCSD}_{PIM}$, and $\alpha^{ML}$ on the small molecule data set.

**3.2. Predicting the Polarizabilities of the Showcase Molecules.** The previous results have shown the model performance on the QM7b data set. To further assess the performance of each model, we consider the showcase molecules. This has previously been shown to provide a stringent test for the predictions of the polarizability tensor as the showcase molecules belong to the borders of the chemical space spanned by QM7b.[54] To visualize this, we performed kernel principal component analysis (KPCA), which can be seen in the Supporting Information. Figure 3 shows the comparison in performance between the PIM, direct ML, and $\Delta^{CCSD}_{PIM}$ models for the polarizability components of the showcase molecules. We have also included the $\alpha^{ML}$ model which has been parameterized using $\lambda$-SOAP and previously shown to provide accurate predictions of polarizability tensors.[54] The $\alpha^{ML}$ polarizability tensors were obtained from the rotated CCSD geometries using the web interface.[54] From the figure, we see that the machine learned models give reasonable predictions whereas PIM performs worse, especially for the *xx* component, which is significantly underestimated as shown in Figure 3(A). We also note that the direct ML model tends to slightly overestimate the *zz* component for molecules with larger polarizability values, which is shown in Figure 3(C). The MAEs for these models are collected in Table 1. We find that that the $\Delta^{CCSD}_{PIM}$ model improves upon PIM by a factor of about two for the *xx* and *zz* components, while the error in the yy component is lowered by about 50%, which can be seen in Figure 3(B). In contrast, the direct ML model only provides moderate improvements for the *yy* and *zz* components relative to PIM. Although the *xx* component is improved in the direct ML model, the performance is worse than $\Delta^{CCSD}_{PIM}$. Thus, we see the advantage of using the Δ-ML model over the direct ML model for these larger molecules beyond the QM7b training set. For comparison, the $\alpha^{ML}$ model provides a better performance for all components, especially for the *zz* polarizability elements. With respect to the $\Delta^{CCSD}_{PIM}$ model, the $\alpha^{ML}$ model has a lower MAE by about 30% for the *xx* and *yy* components, and about 60% for the *zz* components. However, none of these models are better than DFT/B3LYP except in their *zz* component predictions, although, as mentioned previously, this is simply due to delocalization error.

The difference in performance of the direct ML and $\alpha^{ML}$ models is likely due to the local representation used in $\alpha^{ML}$ which allows for a finer representation of the chemical space and thus the prediction of the polarizability tensor. For the $\Delta^{CCSD}_{PIM}$ model, the quality of the lower level method is expected to be reflected in the accuracy of the prediction. For example, PIM may have unsystematic errors that lead to greater difficulty in learning. To verify this, we parameterized a Δ-ML model to the DFT/B3LYP polarizability tensors obtained with the d-aug-cc-pVDZ basis set for the QM7b data set. These tensors were diagonalized using the same procedure as mentioned above. Using this model, we found the MAE in $\Delta^{CCSD}_{B3LYP}$ to be 0.43, 0.73, and 1.53 a.u. for the *xx*, *yy*, and *zz* components, respectively, for the showcase molecules. This model improves significantly upon $\Delta^{CCSD}_{PIM}$ and DFT/B3LYP, as expected, and performs better than $\alpha^{ML}$. However, a more fair comparison would be to use a $\Delta\alpha^{ML}$ model, which predicts the


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

corrections to DFT/B3LYP utilizing the $\lambda$-SOAP kernel.[54] This $\Delta\alpha^{ML}$ model provides polarizability tensors with errors of 0.70, 0.67, and 1.18 a.u., for the *xx*, *yy*, and *zz* polarizability components, respectively, for the showcase molecules. This result is slightly better than the $\Delta^{CCSD}_{B3LYP}$ model, again indicating a potential advantage of using a local representation. However, the difference is now much smaller, which points to improving the PIM model will lead to greater performance. This can be done by better accounting for the local environment in PIM by extending the current per atom type polarizability to a per atom polarizability similar to what is done in the SMARTS mode.[94]

**3.3. Predicting the Polarizabilities of Small Molecules.** The predictions of the showcase molecules demonstrated how the model performs on systems slightly larger than the QM7b training set. It would also be interesting to test the performance of the models on systems that are smaller and chemically diverse, yet contains the same atom types as the QM7b data set. From the KPCA analysis, shown in the Supporting Information, we see that these molecules reside at the periphery of the QM7b data set. These molecules are also part of a data set that have been used for benchmarking DFT functionals, and thus provide an interesting comparison.[91] Figure 4 shows the performance for the direct ML, $\Delta^{CCSD}_{PIM}$, and $\alpha^{ML}$, and PIM models on these small molecules. The MAEs for these models are found in Table 1. From the figure, we see that the direct ML model fails to predict the polarizability tensors, especially by underestimating both the *xx* and the *zz* components, as seen in Figure 4(A),(B), respectively. Furthermore, the large errors for these molecules are particularly significant due to their smaller size relative to the showcase molecules, thus pointing to a much worse performance. This is perhaps not so surprising as these molecules are more chemically distinct than those found in the QM7b training set. The $\Delta^{CCSD}_{PIM}$ model performs better than the direct ML model, though it does not perform as well as PIM for the *xx* and *yy* components but slightly improves the *zz* component. From the figure, we see that $\Delta^{CCSD}_{PIM}$ tends to overestimate the polarizability tensors for these molecules. This is likely due to a large positive correction since PIM was found to underestimate the polarizabilities of the QM7b testing set. However, for the smaller molecules, PIM was found to overestimate the polarizabilities since it was also parameterized to the QM7b datset, giving atomic polarizabilities that are less representative of the chemical environment of these molecules.

Similarly, from the figure and Table 1, we see that the $\alpha^{ML}$ performs worse than $\Delta^{CCSD}_{PIM}$, although not as bad as the direct ML model. Furthermore, we found that $\alpha^{ML}$ for a few small molecules provided predictions that were severely over- or underestimated, and even in some cases negative. Thus, these molecules were excluded from the comparisons in both Figure 4 and Table 1. The excluded molecules are noted in the Supporting Information (SI). Again, this emphasizes that these small molecules are chemically distinct from the QM7b data set, and thus it is not that surprising that both direct ML models mentioned provide less accurate predictions. For contrast, we also report in Table 1 the errors for the DFT/B3LYP polarizability tensor elements taken from ref 91. We see that the performance of DFT/B3LYP for small molecules is comparable to the predictions on the QM7b and showcase data sets. Furthermore, the error of the *zz* component is comparable to the other two components, which reflects that these molecules are more spherical than the molecules in the QM7b data set. Overall, we find that all the models perform worse than DFT/B3LYP for this data set. This points to that it would be important to also include smaller and more chemically diverse molecules in the training set.

**3.4. Predicting the Polarizabilities of the QM9r Molecules.** As a final test, we assessed the performance of the $\Delta^{CCSD}_{PIM}$ model on the QM9r data set which is composed of organic molecules with up to nine heavy atoms (C, N, O). This datset contains significantly more molecules and a larger size range than the QM7b data set, for which isotropic polarizabilities at the DFT/B3LYP level of theory is available.[92,93] This is also illustrated in KPCA analysis found in the Supporting Information. Figure 5 compares the isotropic polarizabilities between DFT/B3LYP and the $\Delta^{CCSD}_{PIM}$ model. Since the polarizability tensors at the CCSD level of theory are unavailable for this data set, the comparison is mainly made to assess the trend in the predicted polarizability. From the figure, we see that $\Delta^{CCSD}_{PIM}$ predicts polarizabilities that are larger than what is found from DFT/B3LYP. We found the difference in the isotropic polarizabilites, as measured by the MAE, of the two models to be 12.2 a.u. This is surprising as we would expect a better correlation between DFT/B3LYP and CCSD polarizabilities as, for example, found in the QM7b data set (see Table 1). Similarly, we found the error between the isotropic polarizabilities of the $\Delta^{CCSD}_{PIM}$ model and DFT/B3LYP for the QM7b data set to be 1.92 a.u., thus we would have expected a similar performance for the QM9r data set.

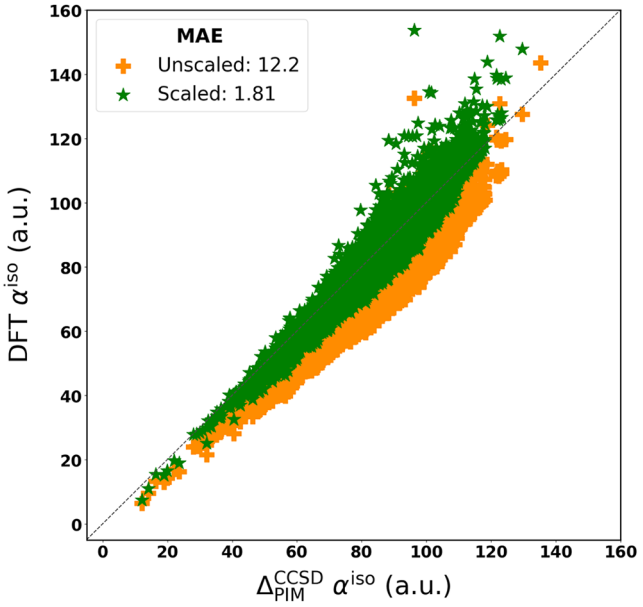



**Figure 5.** Comparison between $\Delta^{CCSD}_{PIM}$ predictions and DFT/B3LYP isotropic polarizabilities. Unscaled DFT/B3LYP polarizabilities are shown as orange crosses, and the scaled polarizabilities are shown as green stars.

Since the molecules in the two data sets are not significantly different, we would not expect that this would cause the discrepancy and therefore the most likely reason is due to the difference in basis sets used for each set of DFT/B3LYP calculations. Specifically, for the QM9 data set the smaller 6−31G(2df,p) was used, while for the QM7b datset the larger d-aug-cc-pVDZ basis set was used. To account for the difference in both basis set and method, we identified the set of common molecules between the two data sets and determined a scaling


https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050

factor to best match the CCSD d-aug-cc-pVDZ polarizabilities and DFT/B3LYP 6−31G(2df,p) polarizabilities (see Supporting Information). From this, the best scaling factor was determined to be 1.16, which is consistent with the use of a smaller basis set for the QM9 data set leading to underestimation of the isotropic polarizabilities. This scaling factor was then applied to the full QM9r datset, with the comparison between $\Delta_{PIM}^{CCSD}$ and the scaled DFT/B3LYP isotropic polarizabilities are plotted in Figure 5. From the figure, we see that this leads to a better correlation between the two models with the MAE difference lowered to 1.81 a.u. This error is consistent with what we found between $\Delta_{PIM}^{CCSD}$ and DFT/B3LYP for the QM7b data set, but larger than the MAE between $\Delta_{PIM}^{CCSD}$ and the CCSD isotropic polarizabilites reported in Table 1. This points to that the simple scaling factor applied accounts to the differences in basis set and, to a lesser degree, the difference in using B3LYP versus CCSD for predicting polarizabilities. Further, it can be seen that the scaled B3LYP polarizabilities tend to be larger than the $\Delta_{PIM}^{CCSD}$ model, especially as the polarizability values increase. A similar trend was also reported[16] in the comparison of the B3LYP polarizabilities with the CCSD polarizabilities for the QM7b data set, which indicates that B3LYP is prone to overestimating the polarizabilities for larger and more conjugated molecules. However, we found that the molecules with the largest differences between $\Delta_{PIM}^{CCSD}$ and the scaled DFT/B3LYP polarizabilities are either zwitterionic or alkyne-like molecules. These types of molecules tend also be difficult for PIM to predict, thus leading to discrepancies in describing their polarizabilities. Ultimately, this suggests that larger and more chemically diverse data sets with polarizabilities obtained at a high level of theory are needed.

## 4. CONCLUSIONS

In this work, we have introduced a simple ML model to predict polarizability tensors by combining a polarizable dipole interaction model with a Δ-machine learning approach. This model incorporates the rotational symmetry of the polarizability tensor by diagonalizing the PIM tensor to obtain a unique reference geometry. The major advantage of adapting this unique reference geometry is that it is only necessary to learn the diagonal elements of the tensor. The model was parameterized to the CCSD polarizabilities from the QM7b data set. We have shown that the $\Delta_{PIM}^{CCSD}$ is comparable to DFT/B3LYP at a lower cost for molecules with similar chemical composition as the QM7b data set. We further show that both iso- and anisotropic polarizabilities are well described, even though the model was not directly trained on these quantities. For the QM9r data set, we showed that good agreement for the isotropic polarizability can be achieved. However, this required introducing a scaling of the QM9 polarizabilities to account for the smaller basis set compared to the one used for the QM7b data set as well as the differences in the level of theory. Finally, for molecules smaller and chemically more diverse than the training set, we find that models trained on the QM7b data set performed worse than DFT/B3LYP. Therefore, to generate accurate machine learned models for polarizabilities, it is likely necessary to construct larger and more chemically diverse data sets using a high level of theory, although this is computationally challenging. For larger, conjugated molecules, the $\Delta_{PIM}^{CCSD}$ model performed slightly worse. This is likely attributed to shortcomings in PIM in describing these molecules, and thus can be improved by better accounting for the local environment in PIM such that conjugated and nonconjugated molecular systems can be better differentiated. In summary, the combination of PIM with Δ-machine learning provides a simple and promising approach for predicting polarizability tensors and could therefore serve as models to describe Raman spectroscopy, where the efficient calculation of polarizability tensors is paramount.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00752.

Specific ML model hyperparameters used in this work, the molecules from the small molecule data set, and a brief description of the QM9 data set scaling procedure; There is also a comparison of model performance using the root mean squared error, as well as a kernel principal component analysis plot comparing the feature space between the QM7b training set and the QM9r data set (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Lasse Jensen** − *Department of Chemistry, Penn State University, University Park, Pennsylvania 16802, United States*; orcid.org/0000-0003-1237-5021; Email: jensen@chem.psu.edu

### Authors

**Imran Chaudhry** − *Department of Chemistry, Penn State University, University Park, Pennsylvania 16802, United States*

**Mark J. Bronson, Jr.** − *Department of Chemistry, Penn State University, University Park, Pennsylvania 16802, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00752


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from National Science Foundation Grant CHE-2312222. Simulations in this work were in part performed on the Pennsylva- nia State University's Institute for Computational and Data Sciences' Roar supercomputer (https://icds.psu.edu/).

## ■ REFERENCES


(1) Diem, M. *Introduction to Modern Vibrational Spectroscopy*; Wiley-Interscience, 1993.

(2) Wilson, E. B.; Decius, J.; Cross, P. C. *Molecular Vibrations: The Theory If Infrared and Raman Vibrational Spectra*; Dover Publications, 1980.

(3) Müller, A.; Mohan, N. Raman Intensities, Depolarization Ratios, and Polarizability Derivatives: Matrix Notation, Relation to Other Molecular Constants and Simple Isotopic Rules. *J. Chem. Phys.* **1977**, *67*, 1918−1925.

(4) Hasager, F.; Nielsen, O. J.; Mikkelsen, K. V. Geometries, Molecular Rayleigh Scattering, Raman and Infrared Frequencies of Polycyclic Aromatic Hydrocarbons and Subunits of Graphite Studied by DFT Methods. *Environ. Sci.: Atmos.* **2022**, *2*, 1023−1031.

(5) Ni, S.; Meng, T.-T.; Huang, G.-Q.; Tang, Y.-Z.; Bai, F.-Y.; Zhao, Z. Roles of Amides on the Formation of Atmospheric HONO and the Nucleation of Nitric Acid Hydrates. *J. Phys. Chem. A* **2023**, *127*, 5402−5413.
(6) Harczuk, I.; Vahtras, O.; Ågren, H. Modeling Rayleigh Scattering of Aerosol Particles. *J. Phys. Chem. B* **2016**, *120*, 4296−4301.
(7) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034−2045.
(8) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D. J. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983−5013.
(9) Wang, H.; Yang, W. Determining Polarizable Force Fields with Electrostatic Potentials from Quantum Mechanical Linear Response Theory. *J. Chem. Phys.* **2016**, *144*, No. 224107.
(10) van Duijnen, P. T.; Swart, M. Molecular and Atomic Polarizabilities: Thole's Model Revisited. *J. Phys. Chem. A* **1998**, *102*, 2399−2407.
(11) Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations II: Induced Dipole Models Significantly Improve Accuracy of Intermolecular Interaction Energies. *J. Phys. Chem. B* **2011**, *115*, 3100−3111.
(12) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable Embedding QM/MM: The Future Gold Standard for Complex (Bio)Systems? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433−14448.
(13) Giovannini, T.; Puglisi, A.; Ambrosetti, M.; Cappelli, C. Polarizable QM/MM Approach with Fluctuating Charges and Fluctuating Dipoles: The QM/FQF$\mu$ Model. *J. Chem. Theory Comput.* **2019**, *15*, 2233−2245.
(14) Loco, D.; Polack, É.; Caprasecca, S.; Lagardère, L.; Lipparini, F.; Piquemal, J.-P.; Mennucci, B. A QM/MM Approach Using the AMOEBA Polarizable Embedding: From Ground State Energies to Electronic Excitations. *J. Chem. Theory Comput.* **2016**, *12*, 3654−3661.
(15) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.
(16) Yang, Y.; Lao, K. U.; Wilkins, D. M.; Grisafi, A.; Ceriotti, M.; DiStasio, R. A. Quantum Mechanical Static Dipole Polarizabilities in the QM7b and AlphaML Showcase Databases. *Sci. Data* **2019**, *6*, No. 152.
(17) Lao, K. U.; Jia, J.; Maitra, R.; DiStasio, R. A., Jr. 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.* **2018**, *149*, No. 204303.
(18) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116*, 5301−5337.
(19) de Wergifosse, M.; Grimme, S. Perspective on Simplified Quantum Chemistry Methods for Excited States and Response Properties. *J. Phys. Chem. A* **2021**, *125*, 3841−3851.
(20) Löffelsender, S.; Beaujean, P.; de Wergifosse, M. Simplified Quantum Chemistry Methods to Evaluate Non-Linear Optical Properties of Large Systems. *WIREs Comput. Mol. Sci.* **2024**, *14*, No. e1695.
(21) Beran, J. A.; Kevan, L. Semiempirical Calculation of Molecular Polarizabilities and Diamagnetic Susceptibilities of Fluorocarbons, Substituted Fluorocarbons, Ethers, Esters, Ketones, and Aldehydes. *J. Phys. Chem. A* **1969**, *73*, 3860−3866.
(22) Miller, K. J.; Savchik, J. A New Empirical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(23) Adant, C.; Brédas, J. L.; Dupuis, M. An Ab Initio and Semiempirical Study of the First- and Third-Order Polarizabilities in Benzene and Thiophene Derivatives: Electron Correlation Effects. *J. Phys. Chem. A* **1997**, *101*, 3025−3031.
(24) Praveen, P. A.; Babu, R. R.; Ramamurthi, K. In *Validation of PM6 & PM7 Semiempirical Methods on Polarizability Calculations*, AIP Conference Proceedings; AIP, 2015.
(25) de Wergifosse, M.; Grimme, S. Nonlinear-Response Properties in a Simplified Time-Dependent Density Functional Theory (sTD-DFT) Framework: Evaluation of the First Hyperpolarizability. *J. Chem. Phys.* **2018**, *149*, No. 024108.
(26) Thiel, W. Semiempirical Quantum−Chemical Methods. *WIREs Comput. Mol. Sci.* **2014**, *4*, 145−157.
(27) Dewar, M. J. S.; Yamaguchi, Y.; Suck, S. H. MNDO Calculations of Molecular Electric Polarizabilities, Hyperpolarizabilities, and Nonlinear Optical Coefficients. *Chem. Phys. Lett.* **1978**, *59*, 541−544.
(28) Parkinson, W. A.; Zerner, M. C. Hyperpolarizability Determined from the Intermediate Neglect of Differential Overlap Model. *J. Chem. Phys.* **1991**, *94*, 478−483.
(29) Fiedler, L.; Gao, J.; Truhlar, D. G. Polarized Molecular Orbital Model Chemistry. 1. Ab Initio Foundations. *J. Chem. Theory Comput.* **2011**, *7*, 852−856.
(30) Applequist, J. An Atom Dipole Interaction Model for Molecular Optical Properties. *Acc. Chem. Res.* **1977**, *10*, 79−85.
(31) Thole, B. T. Molecular Polarizabilities Calculated with a Modified Dipole Interaction. *Chem. Phys.* **1981**, *59*, 341−350.
(32) Olson, M. L.; Sundberg, K. R. An Atom Monopole-Dipole Interaction Model with Charge Transfer for the Treatment of Polarizabilities of $\pi$-Bonded Molecules. *J. Chem. Phys.* **1978**, *69*, 5400−5404.
(33) Applequist, J. Atom Charge Transfer in Molecular Polarizabilities: Application of the Olson-Sundberg Model to Aliphatic and Aromatic Hydrocarbons. *J. Phys. Chem. A* **1993**, *97*, 6016−6023.
(34) Shanker, B.; Applequist, J. Polarizabilities of Fullerenes C20 through C240 from Atom Monopole-Dipole Interaction Theory. *J. Phys. Chem. A* **1994**, *98*, 6486−6489.
(35) Bode, K. A.; Applequist, J. Improved Theoretical $\Pi - \pi^*$ Absorption and Circular Dichroic Spectra of Helical Polypeptides Using New Polarizabilities of Atoms and NC̀O Chromophores. *J. Phys. Chem. A* **1996**, *100*, 17825−17834.
(36) Shanker, B.; Applequist, J. Atom Monopole-Dipole Interaction Model with Limited Delocalization Length for Polarizabilities of Polyenes. *J. Phys. Chem. A* **1996**, *100*, 10834−10836.
(37) Jensen, L.; Åstrand, P.-O.; Mikkelsen, K. V. An Atomic Capacitance-Polarizability Model for the Calculation of Molecular Dipole Moments and Polarizabilities. *Int. J. Quantum Chem.* **2001**, *84*, 513−522.
(38) Jensen, L.; Åstrand, P.-O.; Osted, A.; Kongsted, J.; Mikkelsen, K. V. Polarizability of Molecular Clusters as Calculated by a Dipole Interaction Model. *J. Chem. Phys.* **2002**, *116*, 4001−4010.
(39) Jensen, L. L.; Jensen, L. Electrostatic Interaction Model for the Calculation of the Polarizability of Large Noble Metal Nanoclusters. *J. Phys. Chem. C* **2008**, *112*, 15697−15703.
(40) Nénon, S.; Champagne, B. SCC-DFTB Calculation of the Static First Hyperpolarizability: From Gas Phase Molecules to Functionalized Surfaces. *J. Chem. Phys.* **2013**, *138*, No. 204107.
(41) Nénon, S.; Champagne, B. Origin of the Surface-Induced First Hyperpolarizability in the C60/SiO2 System: SCC-DFTB Insight. *J. Phys. Chem. Lett.* **2014**, *5*, 149−153.
(42) Bishop, C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, NY, 2006.
(43) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, 2009.
(44) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336−2347.
(45) Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.
(46) Kulichenko, M.; Smith, J. S.; Nebgen, B.; Li, Y. W.; Fedik, N.; Boldyrev, A. I.; Lubbers, N.; Barros, K.; Tretiak, S. The Rise of Neural

Networks for Materials and Chemical Dynamics. *J. Phys. Chem. Lett.* **2021**, *12*, 6227−6243.
(47) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, No. 5024.
(48) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(49) Ojha, A.; Bulusu, S. S.; Banerjee, A. Machine Learning Approaches for Modelling of Molecular Polarizability in Gold Nanoclusters. *Artif. Intell. Chem.* **2024**, *2*, No. 100080.
(50) Heinen, S.; Schwilk, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Machine Learning the Computational Cost of Quantum Chemistry. *Mach. Learn.: Sci. Technol.* **2020**, *1*, No. 025002.
(51) Cheng, Z.; Zhao, D.; Ma, J.; Li, W.; Li, S. An On-the-Fly Approach to Construct Generalized Energy-Based Fragmentation Machine Learning Force Fields of Complex Systems. *J. Phys. Chem. A* **2020**, *124*, 5007−5014.
(52) Chen, Y.; Yan, W.; Wang, Z.; Wu, J.; Xu, X. Constructing Accurate and Efficient General-Purpose Atomistic Machine Learning Model with Transferable Accuracy for Quantum Chemistry. *J. Chem. Theory Comput.* **2024**, *20*, 9500−9511.
(53) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(54) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 3401−3406.
(55) Martinka, J.; Pederzoli, M.; Barbatti, M.; Dral, P. O.; Pittner, J. A Simple Approach to Rotationally Invariant Machine Learning of a Vector Quantity. *J. Chem. Phys.* **2024**, *161*, No. 174104.
(56) Berger, E.; Komsa, H.-P. Polarizability Models for Simulations of Finite Temperature Raman Spectra from Machine Learning Molecular Dynamics. *Phys. Rev. Mater.* **2024**, *8*, No. 043802.
(57) Grumet, M.; von Scarpatetti, C.; Bučko, T.; Egger, D. A. Delta Machine Learning for Predicting Dielectric Properties and Raman Spectra. *J. Phys. Chem. C* **2024**, *128*, 6464−6470.
(58) Nguyen, V. H. A.; Lunghi, A. Predicting Tensorial Molecular Properties with Equivariant Machine Learning Models. *Phys. Rev. B* **2022**, *105*, No. 165131.
(59) Zhang, Y.; Ye, S.; Zhang, J.; Hu, C.; Jiang, J.; Jiang, B. Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties. *J. Phys. Chem. B* **2020**, *124*, 7284−7290.
(60) Góger, S.; Karimpour, M. R.; Tkatchenko, A. Four-Dimensional Scaling of Dipole Polarizability: From Single-Particle Models to Atoms and Molecules. *J. Chem. Theory Comput.* **2024**, *20*, 6621−6631.
(61) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, No. 036002.
(62) Raimbault, N.; Grisafi, A.; Ceriotti, M.; Rossi, M. Using Gaussian Process Regression to Simulate the Vibrational Raman Spectra of Molecular Crystals. *New J. Phys.* **2019**, *21*, No. 105001.
(63) Sommers, G. M.; Andrade, M. F. C.; Zhang, L.; Wang, H.; Car, R. Raman Spectrum and Polarizability of Liquid Water from Deep Neural Networks. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10592−10602.
(64) Schütt, K.; Unke, O.; Gastegger, M. In *Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra*, Proceedings of the 38th International Conference on Machine Learning; PMLR, pp 9377−9388.
(65) Feng, C.; Xi, J.; Zhang, Y.; Jiang, B.; Zhou, Y. Accurate and Interpretable Dipole Interaction Model-Based Machine Learning for Molecular Polarizability. *J. Chem. Theory Comput.* **2023**, *19*, 1207−1217.
(66) Kebabsa, A.; Maurel, F.; Brémond, É. Boosting the Modeling of Infrared and Raman Spectra of Bulk Phase Chromophores with Machine Learning. *J. Chem. Theory Comput.* **2024**, *20*, 7009−7015.
(67) Fang, M.; Tang, S.; Fan, Z.; Shi, Y.; Xu, N.; He, Y. Transferability of Machine Learning Models for Predicting Raman Spectra. *J. Phys. Chem. A* **2024**, *128*, 2286−2294.
(68) Berger, E.; Niemelä, J.; Lampela, O.; Juffer, A. H.; Komsa, H.-P. Raman Spectra of Amino Acids and Peptides from Machine Learning Polarizabilities. *J. Chem. Inf. Model.* **2024**, *64*, 4601−4612.
(69) Xu, N.; Rosander, P.; Schäfer, C.; Lindgren, E.; Österbacka, N.; Fang, M.; Chen, W.; He, Y.; Fan, Z.; Erhart, P. Tensorial Properties via the Neuroevolution Potential Framework: Fast Simulation of Infrared and Raman Spectra. *J. Chem. Theory Comput.* **2024**, *20*, 3273−3284.
(70) Zauchner, M. G.; Forno, S. D.; Cśanyi, G.; Horsfield, A.; Lischner, J. Predicting Polarizabilities of Silicon Clusters Using Local Chemical Environments. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 045029.
(71) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(72) Pauletti, M.; Rybkin, V. V.; Iannuzzi, M. Subsystem Density Functional Theory Augmented by a Delta Learning Approach to Achieve Kohn−Sham Accuracy. *J. Chem. Theory Comput.* **2021**, *17*, 6423−6431.
(73) Krug, S. L.; Khan, D.; von Lilienfeld, O. A. Alchemical Harmonic Approximation Based Potential for All Iso-Electronic Diatomics: Foundational Baseline for $Δ$-Machine Learning. *J. Chem. Phys.* **2025**, *1*, No. 044101, DOI: 10.1063/5.0241872.
(74) Nandi, A.; Pandey, P.; Houston, P. L.; Qu, C.; Yu, Q.; Conte, R.; Tkatchenko, A.; Bowman, J. M. Δ-Machine Learning to Elevate DFT-Based Potentials and a Force Field to the CCSD(T) Level Illustrated for Ethanol. *J. Chem. Theory Comput.* **2024**, *20*, 8807−8819.
(75) Chen, X.; Li, P.; Hruska, E.; Liu, F. Δ-Machine Learning for Quantum Chemistry Prediction of Solution-Phase Molecular Properties at the Ground and Excited States. *Phys. Chem. Chem. Phys.* **2023**, *25*, 13417−13428.
(76) Bowman, J. M.; Qu, C.; Conte, R.; Nandi, A.; Houston, P. L.; Yu, Q. Δ-Machine Learned Potential Energy Surfaces and Force Fields. *J. Chem. Theory Comput.* **2023**, *19*, 1−17.
(77) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ-Machine Learning for Potential Energy Surfaces: A PIP Approach to Bring a DFT-based PES to CCSD(T) Level of Theory. *J. Chem. Phys.* **2021**, *154*, No. 051102.
(78) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum Chemical Accuracy from Density Functional Approximations via Machine Learning. *Nat. Commun.* **2020**, *11*, No. 5223.
(79) Dral, P. O. *Advances in Quantum Chemistry*; Elsevier, 2020; Vol. *81*, pp 291−324.
(80) Mayer, A. Formulation in Terms of Normalized Propagators of a Charge-Dipole Model Enabling the Calculation of the Polarization Properties of Fullerenes and Carbon Nanotubes. *Phys. Rev. B* **2007**, *75*, No. 045407.
(81) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes, The Art of Scientific Computing*; Cambridge University Press, 2007.
(82) Bonin, K. D.; Kresin, V. V. *Electric-Dipole Polarizabilities of Atoms, Molecules, and Clusters*; World Scientific Publishing Co. Pte. Ltd.: NJ, 1997.
(83) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.


https://doi.org/10.1021/acs.jctc.5c00752
*J. Chem. Theory Comput.* 2025, 21, 7040−7050

(84) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, No. 045017.
(85) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, *247*, No. 106949.
(86) Laakso, J.; Himanen, L.; Homm, H.; Morooka, E. V.; Jäger, M. O. J.; Todorović, M.; Rinke, P. Updates to the DScribe Library: New Descriptors and Derivatives. *J. Chem. Phys.* **2023**, *158*, No. 234802.
(87) Langer, M. F.; Goeßmann, A.; Rupp, M. Representations of Molecules and Materials for Interpolation of Quantum-Mechanical Simulations via Machine Learning. *npj Comput. Mater.* **2022**, *8*, No. 41.
(88) Stüke, A.; Rinke, P.; Todorović, M. Efficient Hyperparameter Tuning for Kernel Ridge Regression with Bayesian Optimization. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 035022.
(89) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(90) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, No. 095003.
(91) Hait, D.; Head-Gordon, M. How Accurate Are Static Polarizability Predictions from Density Functional Theory? An Assessment over 132 Species at Equilibrium Geometry. *Phys. Chem. Chem. Phys.* **2018**, *20*, 19800−19810.
(92) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, No. 140022.
(93) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(94) Litman, J. M.; Liu, C.; Ren, P. Atomic Polarizabilities for Interactive Dipole Induction Models. *J. Chem. Inf. Model.* **2022**, *62*, 79−87.
(95) Sekino, H.; Maeda, Y.; Kamiya, M.; Hirao, K. Polarizability and Second Hyperpolarizability Evaluation of Long Molecules by the Density Functional Theory with Long-Range Correction. *J. Chem. Phys.* **2007**, *126*, No. 014107.
(96) Bryenton, K. R.; Adeleke, A. A.; Dale, S. G.; Johnson, E. R. Delocalization Error: The Greatest Outstanding Challenge in Density-Functional Theory. *WIREs Comput. Mol. Sci.* **2023**, *13*, No. e1631.





https://doi.org/10.1021/acs.jctc.5c00752
J. Chem. Theory Comput. 2025, 21, 7040−7050