Revisiting the $N(N+1)/2$-site *s*-type Gaussian charge model for permutationally invariant polynomial fitting of global molecular tensor surfaces

Sophia A. Kudratov,[1] Martina Kaledin,[1] Alexey L. Kaledin[2,*]

[1] *Department of Chemistry & Biochemistry, Kennesaw State University, 370 Paulding Ave NW, Box # 1203, Kennesaw, Georgia, 30144*

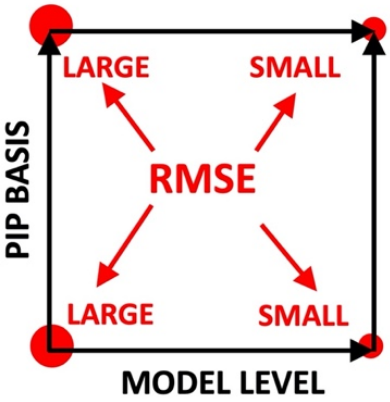
[2] *Cherry L. Emerson Center for Scientific Computation and Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, Georgia, 30322*

ABSTRACT: The exact expressions for the dipole, quadrupole, and octupoles of a collection of N point charges involve summations of corresponding tensors over the N sites weighted by their charge magnitudes. When the point charges are atoms (in a molecule) the N-site formula is an approximation, and one must integrate over the electron density to recover the exact multipoles. In the present work we revisit the $N(N+1)/2$-site point charge density model of Hall [*Chem. Phys. Lett.* **6**, *501*, 1973] for the purpose of fitting *ab initio* derived multipole moment hypersurfaces using permutationally invariant polynomials (PIP). We examine new approaches in PIP-fitting procedures for the dipole, quadrupole, octupole moments, and polarizability tensor surfaces (DMS, QMS, OMS and PTS, respectively) for a non-polar CCl_4 and a polar $CHCl_3$ and show that compared to the primitive N-site model the $N(N+1)/2$-site model appreciably improves the relative RMSE of the DMS and does much more substantially so, by an order of magnitude, for the corresponding ones of QMS and OMS. Training datasets are obtained by sampling potential energies up to 18000 cm^{-1} above the global minima, generated by molecular dynamics simulations at the DFT B3LYP/aug-cc-pVDZ level of theory.

Keywords: permutationally invariant polynomials, multipole tensor surfaces, distributed multipoles, molecular dynamics, linear regressions

E-mail: akaledi@emory.edu

Graphical Abstract



Molecular dipole, quadrupole and octupole tensor surfaces are fitted to *ab initio* data via the internuclear distances and permutationally invariant polynomials (**PIP**). It is shown that the conventional approach of expressing the multipoles using the N atomic sites (**Model N1**) is dramatically improved by the $N(N-1)/2$ internuclear barycenters added as point charges (**Model N2**) while using them in the fitting procedure. The root mean square errors (**RMSE**) are reduced by an order of magnitude.

1. INTRODUCTION

Molecular multipole moments are known to play a crucial role in describing intermolecular interactions in the gas and liquid phase.[1-3] The multipoles, typically up to the quadrupole and octupole, and in some cases higher moments, form the basis of non-bonding interactions in such molecular mechanics models as AMOEBA,[4,5] SIBFA (Sum of Interactions Between Fragments Ab initio computed),[6-8] GMM (the Gaussian Multipole Model),[9] and others[10-18]. The first order multipole moment (the permanent dipole) and its polarizability are key components in simulations of infrared and, respectively, Raman spectra of molecules, liquids and solids.[19,20] In recent applications, use of both the dipole and quadrupole moments was demonstrated in direct molecule-field coupled simulations of vibrational spectra of biological macromolecules.[21]

Computational models for the dipole moment vector and occasionally its polarizability tensor, of various degrees of approximation exist for a range of systems,[9,22-25] being most prominent for water.[4,5,14,15,26-31] More advanced methodologies, such as those based on polynomial functions fitted to *ab initio* data, bring up a family of alternative approaches.[32-37] Representation of the molecular dipole moment and polarizability tensor, as *global* functions of nuclear coordinates fitted to an extensive set of high quality *ab initio* data is a rather challenging task; however, recent developments in the permutationally invariant polynomial (PIP) theory have produced some major new advances.[32-36,38-40] In a related manner, machine leaning techniques, in particular those based on artificial neural networks (ANN),[41] have shown themselves to be competitive in accurate representation of multipole moments and dipole polarizability tensor trained on *ab initio* derived data.[13-16,42,43] The latter approaches, i.e. the collective family of PIP/ANN, are particularly relevant to the present study as we aim to express the multipoles as

global functions of internal nuclear coordinates, as have previously been done in the fitting of global potential energy surfaces (PES).[36,44-46]

We note that in addition to a high-quality description of the dipole moment, it is essential for centrosymmetric, tetrahedral and other highly symmetric molecules to also have a high quality description of the quadrupole and octupole moments in the same manner due to their prominent contribution to the electrostatic energy. For instance, the respective dipole-quadrupole and quadrupole-quadrupole energy terms decay as d^{-4} and d^{-5} with the intermolecular separation distance d. Furthermore, the dipole polarizability is responsible for the induction and dispersion forces, such as the dipole-induced-dipole and induced-dipole-induced-dipole interactions.[47] These phenomena have previously been studied by various ANN methods.[13-16] Presently, however, we wish to also incorporate invariance with respect to cyclic permutations of any number of like nuclei since it is crucial for a physically correct description of chemical properties. Yet despite these important characteristics, we have not encountered reports of high-quality *global* quadrupole and higher rank multipole fits using the PIP approach. This apparent lack of research in the computational field adds to the motivation of the present study, in which we will attempt to improve on the existing methods of fitting multipole moment surfaces in the hyperdimensional space of nuclear coordinates, specifically by combining a version of the DMA theory with the method of PIPs.

For the purposes of defining the goals of the present work, to be outlined in detail in the following sections, we emphasize that aside from a few earlier studies using DMA-type approaches,[31,48] the more recent PIP/ANN fitting models for multipole moments have so far been realized at the level of the 'classical' N-atom point charge formula, akin to the Mulliken population analysis.[1,49] This approach is identified here as (PIP) **Model N1**, i.e. a sum over the N atomic

centers weighted by their configuration dependent effective charges $w(\mathbf{R})$. The latter are the actual quantities fitted by PIP/ANN model.[36,43] However, given the success of the distributed multipole analysis (DMA)[1] in many practical applications and other related theories[6-9,50] in describing electrostatic interactions and electron density distribution in complex systems that go beyond the 'classical' representation, we seek to examine and apply a DMA formulation in the present work. To this end we explore its simplest form, namely by placing a single *s*-type Gaussian on each of the atomic sites and using the PIP approach to constructing global fits of molecular multipole moment surfaces based on extensive *ab initio* data. We note that others have recently utilized s-type Gaussian density in polarizable force fields.[51] Understandably, the addition of the higher angular momentum functions, i.e. *p*, *d*, *etc.*, on the atomic centers is expected to be important due to an axial distortion of electron density. It is a natural extension of a 'minimal basis' s-function approximation, and has been considered in a variety of DMA approaches in modeling of molecular electrostatics.[2,3] This will be a subject of future investigations in our development of the DMA-PIP approach.

One may recall the original exploration of Hall of the electron density in terms of a set of spherical Gaussians placed on the atoms, which produces an $N(N+1)/2$-site point charge expression, identified here as (PIP) **Model N2**: for the dipole, quadrupole, and octupole.[52-55] Specifically, **Model N2** consists of the N atomic sites augmented by the $N(N-1)/2$ non-atom sites, alternatively called bond centroids or barycenters[50] due to their locations being determined by the Gaussian exponents. As will be shown below, this model is ideally suited for the PIP representation since the latter uses the $N(N-1)/2$ internuclear distances as a basis for PIP formation.[36] The major practical issue to be addressed below is whether **Model N2** can improve upon **Model N1** in the

PIP fitting of the multipoles and dipole polarizability to *ab initio* data, that is, whether it can significantly reduce the fitting error in the least squares sense: the root mean square error (RMSE).

To test our supposition of the role of the barycenters in the PIP representation we consider two systems with sufficiently diffuse and structured electron densities while having appreciably dissimilar spatial symmetries, namely, an isotropic non-polar CCl_4 with no permanent (static) dipole and quadrupole, and secondly a highly anisotropic and polar $CHCl_3$. These molecules are known to readily form liquids at room temperature, in pure and mixed compositions, and are used as organic solvents. They have been subjects of numerous investigations, some of which examined the dynamics of C-Cl/C-H stretch using 2D-Raman spectroscopy.[56,57] More recently, hydrogen bonding was studied in $CHCl_3$ dimers at the level of the GMM theory employing the DMA approach.[9]

In the work presented here, we calculate and fit the aforementioned systems' multipoles up to the octupole moment, along with their dipole polarizabilities employing the **N1** and **N2** models via a conventional PIP representation.[36] We then present a series of extensive error analyses, compare performances of the two models and discuss the effect of the PIP order on the quality of the fits.

2. THEORETICAL METHODS

2.1 Hall's point charge model

The charge density of an *N*-atom molecule at point **r** in the laboratory frame can be approximated in terms of functions $\varphi_k(\mathbf{r})$, centered at the *N* atomic sites *k*,[2,3]

$$\rho(\mathbf{r}) = \sum_{kk'} \rho_{kk'}\, \varphi_k(\mathbf{r})\varphi_{k'}(\mathbf{r}) \qquad (1)$$

where $\rho_{kk'}$ are elements of the density matrix. Using a basis consisting of *N* primitive normalized spherical Gaussian functions $\varphi_k(\mathbf{r}) = (\zeta_k^3/\pi^3)^{1/4} e^{-\zeta_k|\boldsymbol{r}-\boldsymbol{r}_k|^2/2}$ (others have used Slater-type Gaussian contractions[9]) and substituting for the product of two Gaussians another Gaussian, the latter centered at $\mathbf{r}_{kk'} = (\zeta_k \mathbf{r}_k + \zeta_{k'} \mathbf{r}_{k'})/(\zeta_k + \zeta_{k'})$ with the exponent $\zeta_{kk'} = \zeta_k + \zeta_{k'}$, the resultant density is a sum of the $N(N+1)/2$ spherical Gaussians at the centers $\mathbf{r}_{kk'}$.

For the purposes of present calculations, we find it informative to express the moments using the exact density $\rho(\mathbf{r})$ by separating them into the sum of the "classical", e.g., from the point charge model, and overlap originating "quantum" contributions. Namely, we have for the total charge,

$$Z = \sum_k \rho_k \; + 2 \sum_{k<k'} \rho_{kk'} S_{kk'} \qquad (2a)$$

the dipole moment,

$$\boldsymbol{\mu} = \sum_k \rho_k \mathbf{r}_k + 2 \sum_{k<k'} \rho_{kk'} S_{kk'} \mathbf{r}_{kk'} \qquad (2b)$$

the traceless quadrupole moment tensor,

$$\mathbf{Q} = \sum_k \rho_k (3\mathbf{r}_k \mathbf{r}_k^T - r_k^2 \mathbf{1}) + 2 \sum_{k<k'} \rho_{kk'} S_{kk'} (3\mathbf{r}_{kk'} \mathbf{r}_{kk'}^T - r_{kk'}^2 \mathbf{1}) \qquad (2c)$$

and the symmetry unique elements of the octupole moment tensor,

$$\Omega_{xxx} = \sum_k \rho_k x_k \left(x_k^2 + \frac{3}{2\zeta_k} \right) + 2 \sum_{k<k'} \rho_{kk'} S_{kk'} x_{kk'} \left(x_{kk'}^2 + \frac{3}{2\zeta_{kk'}} \right) \qquad (2d)$$

$$\Omega_{xyy} = \sum_k \rho_k x_k \left(y_k^2 + \frac{1}{2\zeta_k} \right) + \frac{1}{2} \sum_{k<k'} \rho_{kk'} S_{kk'} x_{kk'} \left(y_{kk'}^2 + \frac{1}{2\zeta_{kk'}} \right) \qquad (2e)$$

$$\Omega_{xyz} = \sum_{k} \rho_k x_k y_k z_k + 2 \sum_{k<k'} \rho_{kk'} S_{kk'} x_{kk'} y_{kk'} z_{kk'} \qquad (2f)$$

where the Gaussian overlap $\langle \varphi_k | \varphi_{k'} \rangle$ is

$$S_{kk'} = \left(\frac{4\zeta_k \zeta_{k'}}{(\zeta_k + \zeta_{k'})^2} \right)^{\frac{3}{4}} exp\left[-\frac{\zeta_k \zeta_{k'}}{\zeta_k + \zeta_{k'}} |\mathbf{r}_k - \mathbf{r}_{k'}|^2 \right] \qquad (2g)$$

and $\zeta_{kk'} = (\zeta_k + \zeta_{k'})/2$. The leading terms in Eq. (2a-f) may be interpreted as originating from the sum of the *N* 'classical' point charges ρ_k over the atomic centers *k*. The second terms, on the other hand, are purely quantum in nature due to the wavefunction overlap factors $S_{kk'}$, which decay to zero at large *k-k'* separations. However, they do behave physically in exactly the same manner as the classical terms by being sums over point charges $\rho_{kk'}$ located at the *N*(*N*-1)/2 interatomic distance barycenters $\mathbf{r}_{kk'}$.

To facilitate calculation of the electrostatic potential and its derivatives, Hall originally proposed an approximation to Eq. 1 that has the form,

$$\rho_{\text{Hall}}(\mathbf{r}) = \sum_{kk'} \rho_{kk'} \, S_{kk'} \delta(\mathbf{r} - \mathbf{r}_{kk'}) \qquad (3)$$

The products $\rho_{kk'} S_{kk'}$, which are key elements in the Mulliken population analysis, can formally be interpreted as point charges $Z_{kk'}$. Hall first showed that the approximate Eq. 3 conserves the total charge and the dipole moment of the exact density. The form of Eq. 3 also happens to conserve the traceless quadrupole moment. The model deviates from the exact formula for the octupole moment, namely for the normal (unadjusted) O_{xxx} and O_{xyy} elements, as can be seen in Eqs. 2d, 2e. Others have shown that conversion of the octupole (and higher moments) to a traceless form is straightforward, although presently we consider the normal octupole definition.[9]

The utility of Hall's approximation in conjunction with the PIP representation becomes more apparent when one examines the dipole polarizability. By one definition, the dipole

polarizability of a molecule is a $3N \times 3N$ relay matrix $\mathbf{G}$ interconnecting the induced atomic dipoles caused by an external electric field,[58,59]

$$\mathbf{G} = (\mathbf{A}^{-1} + \mathbf{T})^{-1} \quad (4)$$

In the above, **A** is a diagonal matrix with the isotropic atomic polarizabilities a_k as its elements, and **T** is an $N \times N$ tensor with each element being a 3×3 Cartesian block connecting a pair of atoms *k, k'* via the dipole field,

$$(\mathbf{T}_{kk'})_{ij} = \frac{\delta_{ij}}{r_{kk'}^3} - 3\frac{r_i r_j}{r_{kk'}^5} \quad (5)$$

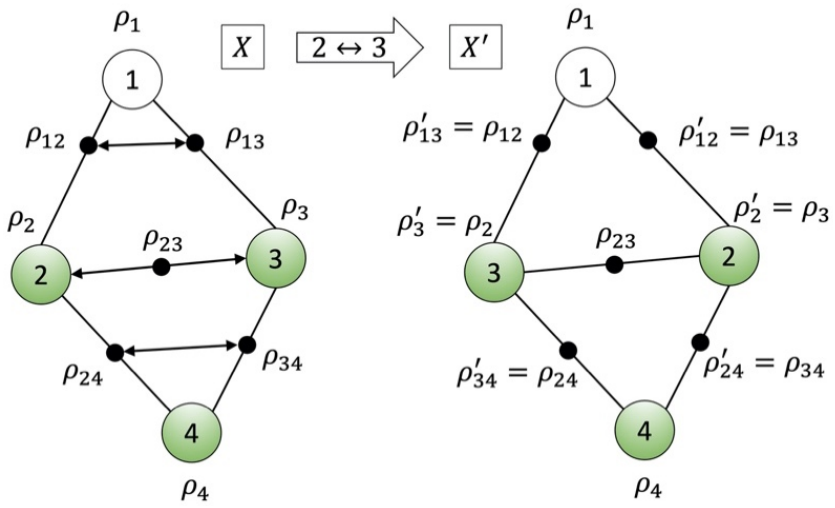
where r_i, r_j, are the Cartesian components *x, y, z* of the $\mathbf{r}_{kk'}$ vector. The matrix **G** is related to the actual 3×3 polarizability matrix $\boldsymbol{\alpha}$ by summing over all atom sites (the 3×3 blocks) k, k',

$$\alpha_{ij} = \sum_{k \neq k'}^{N} [\mathbf{G}_{kk'}]_{ij} \quad (6)$$

With the present interpretation of Eqs. 4-6 being strictly limited to the *N* atomic sites *k, k'*, i.e. the natural sources of (induced) dipoles, it is particularly intriguing in the present PIP approach to evoke Hall's model to extend the currently used polarizability theory to a super atomic representation with *k*, *k'*=1,..., *N*(*N*+1)/2 with the elements $\rho_{kk'}$ as the sources of polarizable electron density. Details of this approach are described in the Supplementary Information (SI). For completeness, we note that Stone[47] and others[60] have proposed more advanced theories of polarizability representation beyond the classic Applequist model,[58] yet it is presently unclear if they can be applied for PIP fitting in a conventional linear regression way.

A major effort of the present investigation is thus to evaluate the quantum contribution, stemming from the atomic orbital overlaps to the quality of the PIP fits of molecular tensor properties. Explicit dependence of the overlap values and barycenter point positions on the Gaussian exponents may further be used to tune the accuracy of the fit by mimicking the exact

electron density, that is, by placing diffuse Gaussians at the more electronegative atomic sites, e.g. C and Cl atoms, and compact Gaussians at the hydrogens. These effects are explored below.



Scheme 1. Details of transformation of a nuclear configuration *X* into an identical configuration *X'* as a result of a permutation of the identical nuclei 2 and 3. The two-sided arrows indicate the elements being transformed upon the permutation. The black dots mark the barycenters. In this example, nucleus 1 is of one kind (white), and nuclei 2, 3 and 4 are of another kind (green).

2.2 A PIP representation

In a first principles calculation the density matrix elements are determined by minimizing the total electronic energy at a nuclear configuration **R**. Presently, however, we follow the example of the preceding work[36] and introduce an explicit nuclear-configuration-dependence into the elements of the density matrix using a PIP ansatz. From now on the collective coordinate **R** is assumed to be a set of 3N Cartesian coordinates completely determining the molecular geometry. For compactness, let us consider the atomic charge densities corresponding to a particular nuclear group class, in this case the chlorines in CCl_4,

$$\rho_k(\mathbf{R}) = \rho_k^{(0)} + \sigma_k(\mathbf{R}') \sum_{i=1}^{L} c_i^{\text{atom}} u_{k,i}(\mathbf{R}) \qquad (7)$$

where $\rho_k^{(0)}$ is the permanent charge (polarizability) on atom *k*, c_i^{atom} is a single set of linear expansion coefficients expressing all the atomic charges (polarizabilities) in the given group (Cl); $u_{k,i}$ is the *i*-th PIP of the total power $0 < m_i \leq M$ describing atom *k*. It is expressed in the usual *N*(*N*-1)/2 internuclear distances employing a permutationally symmetrized sum of monomial products. The latter products are determined by cycling over all possible permutations *J* within the like nucleus groups, that is, 1! 4! permutations for CCl_4 and 1! 1! 3! permutations for $CHCl_3$,

$$u_{k,i}(\mathbf{R}) = \sum_{j=1}^{J} b_{k,i,j} y_{12}^{p_{i,j}} y_{13}^{q_{i,j}} \cdots y_{(N-1)N}^{z_{i,j}} \qquad (8)$$

Specifically for the atom pair (k, k') our choice for the internuclear distance function is a Morse variable, $y_{kk'}^{p_{i,j}} = exp(-p_{i,j} d_{kk'}/d_0)$ where $d_{kk'} = |\mathbf{r}_k - \mathbf{r}_{k'}|$ and d_0 is a constant; and so on for all the other unique atom pairs. The integer powers satisfy the relation $p_{i,j} + q_{i,j} + \cdots + z_{i,j} = m_i$. All combinations of a given integer set, which satisfy the like atom permutational symmetry rules, are considered for each m_i. For a chosen maximal polynomial power *M* this determines the PIP basis size *L*.[36]

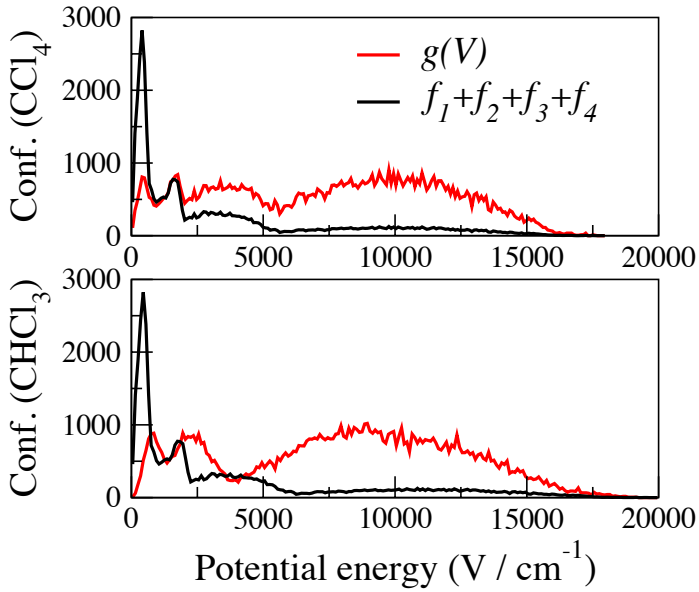


Figure 1. Distributions (not normalized) of potential energies of the four NVE trajectories $f_1+f_2+f_3+f_4$ with energies E_1=18225, E_2=6075, E_3=2025, E_4=675 cm^{-1} for CCl_4 and E_1=20000, E_2=12891, E_3=4297, E_4=1432 cm^{-1} $CHCl_3$, and the corresponding pruned sets $g(V)$, calculated at the B3LYP/aug-cc-pVDZ level. The primitive sets $f_i(V)$ contain 10000 configurations each (40000 configurations in total), while the pruned set $g(V)$ contains ~1000 configurations, i. e., the respective numbers of points selected from the primitive NVE sets are: 714, 195, 76, 16 for CCl_4, and 606, 191, 151, 52 for $CHCl_3$.

The phase factors $b_{k,i,j} = \pm 1$ for atom k in the monomial j of polynomial i are determined based on the number of times, F_k, the integer k shows up in the numeral sequence of the atom pairs (1,2)(1,3)...(N-1,N) that have *only non-zero* powers among $p_{i,j}$, $q_{i,j}$, ..., $z_{i,j}$. For example, in the monomial $y_{12}^2 y_{13}^1 y_{24}^0 y_{25}^3$ we have F_1=2, F_2=2, F_3=1, F_4=0, F_5=1. If F_k is greater than a certain threshold F*, then we set $b_{k,i,j} = -1$, otherwise $b_{k,i,j} = 1$. F* depends critically on the number of non-zero powers in each monomial. (A complete procedure for its determination is described in the SI.) The above-described phase assignment ensures covariance of $u_{k,i}(\mathbf{R})$, meaning that a

permutation of two identical nuclei, e.g. 1 and 2, results in a corresponding exchange of ρ_1 and ρ_2 so that the overall tensor is invariant under the said permutation.

Finally, the quantity σ_k introduced in Eq. 7 is a damping operator acting on an *N*-1 subset of internuclear distances contained in $\mathbf{R}'$. It must be invariant under a permutation of an identical pair k' and k'', but must be covariant with respect to a permutation of an identical pair k and k' and at the same time must decay to zero as atom *k* is removed to infinity away from the other *N*-1 atoms. A simple, although not unique choice, fulfilling these requirements is $\sigma_k(\mathbf{R}') = \sum_{k' \neq k}^{N} y_{kk'}$. This damping function is a new feature in Eq. (7) since in all formulations published previously it was assumed to be unity, and the actual decay of $\rho_k(\mathbf{R})$ to $\rho_k^{(0)}$ in the removed-atom-*k*-limit was not strictly enforced.

To represent the *N*(*N*-1)/2 off-atom (barycenter) point charges, we use an expression similar to Eq. (7) for a particular pair class, e.g., the C-Cl pairs in CCl_4,

$$\rho_{kk'}(\mathbf{R}) = \rho_{kk'}^{(0)} + \sigma_{kk'}(\mathbf{R}') \sum_{i=1}^{L} c_i^{\text{pair}} u_{kk',i}(\mathbf{R}) \tag{9}$$

Here, for the multipoles we set $\rho_{kk'}^{(0)} = 0$ and $\sigma_{kk'}(\mathbf{R}') = 1$. For the polarizability we use a somewhat different ansatz, as described in the SI. Additionally, the pair polynomials must transform properly upon like nuclei permutations. For instance, a permutation of like nuclei 1 and 2 must result in $\rho_{12} \leftrightarrow \rho_{21}$ and simultaneously in $\rho_{1k} \leftrightarrow \rho_{2k}$ for *k* = 3, ..., *N*; and so on for all other permutations of this sort. This may be achieved with little effort by constructing pair-polynomials using direct products of the singles phases, as below,

$$u_{kk',i}(\mathbf{R}) = \sum_{j=1}^{J} b_{k,i,j} b_{k',i,j} y_{12}^{p_{i,j}} y_{13}^{q_{i,j}} \cdots y_{(N-1)N}^{z_{i,j}} \tag{10}$$

(An illustration of this is provided in Scheme 1.) The coefficients c_i^{atom} and c_i^{pair} (for $i = 1, L$) of each of the symmetry-unique species, i.e. for CCl_4 they are C, Cl, C-Cl, Cl-Cl thus $4L$ in total and for $CHCl_3$ they are C, H, Cl, C-H, C-Cl, H-Cl, Cl-Cl thus $7L$ in total, are determined by formulating and solving a linear least squares problem. In this work we use linear regression by means of the singular value decomposition treatment[61] as implemented in the DGESVD subroutine of the Intel Math Kernel Library.[62] This procedure was done by us previously in similar applications.[39]

Table 1. The total number of expansion coefficients for the various types of linear regression problems and the PIP order M. The number in parentheses is the size of the corresponding PIP basis L.

	Model N1		**Model N2**	
	$M = 3$	$M = 4$	$M = 3$	$M = 4$
CCl_4 (μ,Q,Ω)	58 (29)	164 (82)	116 (29)	328 (82)
CCl_4 (α)	203 (29)	574 (82)	261 (29)	738 (82)
$CHCl_3$ (μ,Q,Ω)	222 (74)	690 (230)	518 (74)	1610 (230)
$CHCl_3$ (α)	592 (74)	1840 (230)	888 (74)	2760 (230)

3. RESULTS AND DISCUSSION

3.1 Generation of training data for CCl_4 and $CHCl_3$

In recent works we proposed an approach to constructing well balanced training sets by pruning NVE trajectories.[63] Briefly, our motivation is to generate a training set of configurations such that the distribution of their potential energies is as close to uniform as possible. In other words, the general requirement is that all configurations with potential energies up to some

maximal value V_{max} be equally represented in the set. In the present work, the same strategy is used. We start with setting a high total energy level E_1, e.g., a multiple of the harmonic ZPVE, and apply an exponential cooling schedule to bring it down to a low energy level E_n in several steps. The high energy level describes the quantum vibrationally accessible range of the potential, while the low energy level describes the classically relevant range. The energies E_2,...,E_{n-1} in between form a smooth connection of the two ranges. Following this initial step, we run NVE trajectories, with energies and forces computed on the fly, at each of the E_j total energies with zero total angular momentum using the B3LYP functional[64] with aug-cc-pVDZ basis set.[65] As was noted by us recently in applications involving PIP fitting of polarizability tensors[38,39] and discussed by others in different contexts,[66,67] augmenting orbital basis sets with diffuse functions plays a very important role in polarizability calculations. In the SI we provide the equilibrium geometries (Tables S1, S2), multipole and polarizability tensors (Tables S3, S4, S5) and vibrational frequencies (Table S6). All electronic structure calculations were carried out using the Gaussian[68] and MOLPRO[69] software packages.

At this level of theory, for CCl_4, the ZPVE is 2025 cm^{-1} and the first dissociation limit, $CCl_2 + Cl_2$, is 25172 cm^{-1} above the global minimum. This allows us to start with E_1 = 9ZPVE = 18225 cm^{-1} and cool to E_2 = 3ZPVE = 6075 cm^{-1}, E_3 = ZPVE = 2025 cm^{-1} and E_4 = ZPVE/3 = 675 cm^{-1}. For $CHCl_3$, the ZPVE is 4297 cm^{-1} and the dissociation energy into CCl_2 + HCl is 19473 cm^{-1}. Applying the same cooling schedule creates E_1 = 38673 cm^{-1}, E_2 = 12891 cm^{-1}, E_3 = 4297 cm^{-1}, E_4 = 1432 cm^{-1}. However, the trajectory with the energy E_1 produced fragmentation into CCl_2 + HCl within 2 ps. Lowering the total energy to 30000 cm^{-1} and then to 25000 cm^{-1} also led to dissociation. We finally pinned E_1 = 20000 cm^{-1} as approximately the highest energy corresponding to a non-fragmenting NVE trajectory.

Following this procedure for each system, we ran 4 NVE trajectories for 10 ps with a 1 fs time step, generating a total of 40000 "primitive" configurations. Then, based on this initial step, we contract the primitive configurations into a single set while maximizing the criterion of the potential energy uniformity as mentioned above,

$$g(V) = \sum_{j=1}^{n} d_j f_j(V) \tag{11}$$

That is, the contracted set $g(V)$ is a linear combination of the primitive NVE sets $f_j(V)$ with the coefficients chosen to make $g(V)$ uniform in a least-squares sense. The coefficients d_j in Eq. (10) are solved using a linear regression.[63] The primitive 40000 configurations for each molecule are pruned to sets of 1000 configurations, each. These distributions are shown in Figure 1. As is usually the case, a simple (unweighted) superposition of the primitive distributions favors the low energy region. Pruning these sets by optimizing the coefficients in Eq. (11) yields a much more balanced distribution.

Table 2. The relative RMSE (%) of the dipole (μ), quadrupole (Q), octupole (Ω) moments and the polarizability (α) of CCl_4 and $CHCl_3$ calculated using the **N1** and **N2** models with PIP orders M = 3 and 4. The optimized Gaussian exponents for the multipoles are ζ_{Cl}=0.2 bohr^{-2}, ζ_C = 0.2 bohr^{-2} for CCl_4, and ζ_{Cl}=0.09 bohr^{-2}, ζ_C = 0.4 bohr^{-2}, ζ_H = 0.7 bohr^{-2} for $CHCl_3$. The optimized Gaussian exponents for CH_4 polarizability are ζ_{Cl}=ζ_C=0.2 bohr^{-2}, and for $CHCl_3$ polarizability are ζ_{Cl}=ζ_C=ζ_H=0.1 bohr^{-2}. The polynomial range parameter is d_0=2.0 bohr.

RMSE	**Model N1**				**Model N2**			
	μ	Q	Ω	α	μ	Q	Ω	α
CCl_4 M=3	2.4	7.4	12.0	1.0	0.5	1.2	2.0	0.7
CCl_4 M=4	0.5	5.9	11.2	0.9	0.1	0.3	0.8	0.6
$CHCl_3$ M=3	1.6	6.9	26.7	2.2	0.4	0.9	4.3	1.9
$CHCl_3$ M=4	0.5	6.2	26.6	1.9	0.1	0.2	2.6	1.7

3.2 Fitting of the DMS*,* QMS*,* OMS *for CCl_4 and $CHCl_3$*

To fit the dipole, quadrupole, octupole moment surfaces (DMS, QMS, OMS respectively), and additionally the polarizability tensor surface (PTS) which is necessarily required for constructing intermolecular induction forces, we used the 3rd and 4th order PIPs with $L = 29$ and 82 terms for CCl_4 and $L = 74$ and 230 terms for $CHCl_3$, excluding the constant term, as mentioned previously in Section 2.2. For each of the multipoles we fit the unique tensor elements: 3 for the dipole, 6 for the quadrupole/polarizability, 10 for the octupole, to their respective datasets. Briefly, for the PIP order $M = 4$, (**Model N1**) for the multipoles, and incidentally for the PTS representation, requires 164 linear variables for CCl_4 and 690 linear variables for $CHCl_3$; by the same token, **Model N2** requires 328 linear variables for CCl_4 and 1610 linear variables for $CHCl_3$. The full dimensions are reported in Table 1.

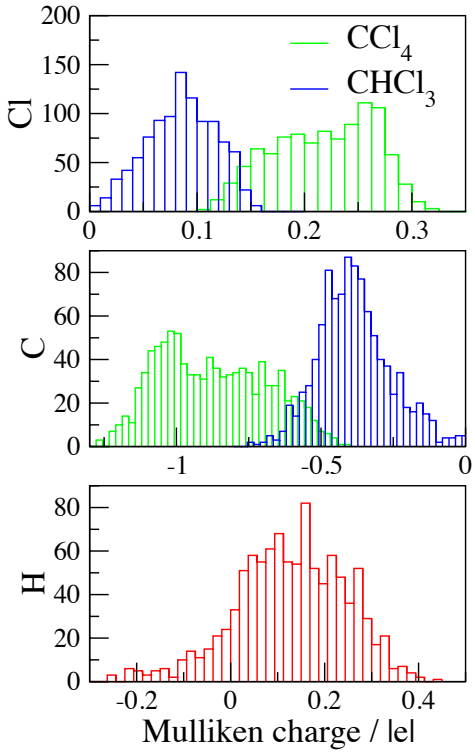


Figure 2. Distribution of Mulliken atomic charges in the training sets for CCl_4 (green histograms) and $CHCl_3$ (blue and red histograms), calculated at the B3LYP/aug-cc-pVDZ level.

In addition to the linear variables there is the exponential range parameter d_0 and the Gaussian exponents describing compactness/diffuseness of the atomic density centers. We first optimized d_0 by minimizing the RMSE of the PESs, recording the best fits produced by setting d_0 = 2 bohr for both molecules. (See Section 3.3 for details of PES fitting.) We then optimized the octupole RMSEs by manually varying the Gaussian exponents in the 2D and 3D spaces and following a simplex-like search, for CCl_4 and $CHCl_3$, respectively. The results for the multipole moments and the polarizability tensor RMSEs are presented in Table 2. Presently, the relative RMSE is defined as

$$RMSE(\%) = 100\left[\frac{\sum_i \sum_\nu \left(\tau_{\nu,i}^{\mathrm{DFT}} - \tau_{\nu,i}^{\mathrm{FIT}}\right)^2}{\sum_i \sum_\nu \left(\tau_{\nu,i}^{\mathrm{DFT}}\right)^2}\right]^{1/2} \qquad (12)$$

where $\tau_{\nu,i}^{\mathrm{DFT}}$ and $\tau_{\nu,i}^{\mathrm{FIT}}$ are the respective moments with Cartesian component ν at configuration i. For the dipole moment, ν = x, y, z; for the quadrupole moment ν = xx, xy, xz, yx, yy, yz, zx, zy, zz; for the octupole moment ν = xxx, xxy, ..., zzz.

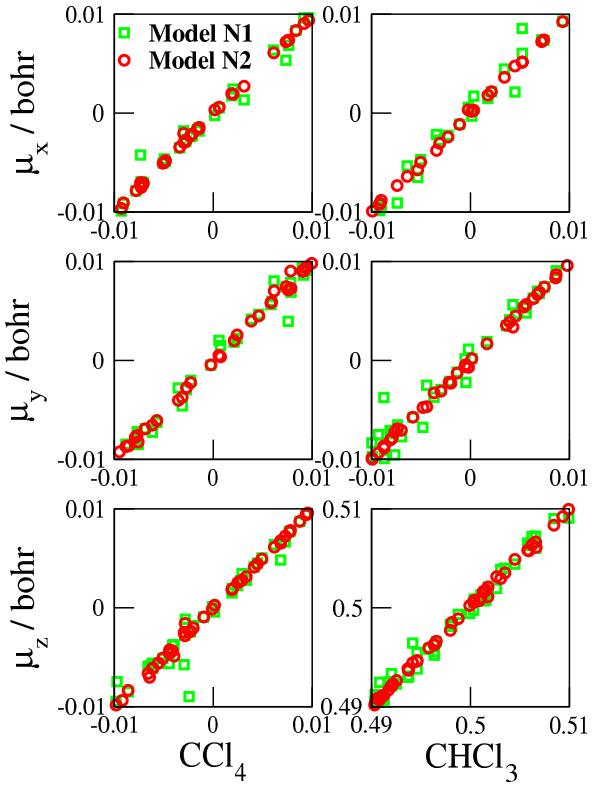


Figure 3A. Correlation diagrams of the fitted dipole moments (vertical axis) vs. the B3LYP/aug-cc-pVDZ training data (horizontal axis), for CCl_4 and $CHCl_3$ using **Model N1** (green dots) and **Model N2** (red dots) with PIP order $M = 4$. In the training sets the molecular z-axis is approximately aligned with a CCl and the CH bonds in CCl_4 and $CHCl_3$, respectively. The permanent $CHCl_3$ dipole moment at the equilibrium is $\mu_z = 0.44$ a.u. Only a narrow range of ± 0.01 bohr of the dipole data is shown to better emphasize point scattering in the fits.

Optimization of the exponents produced the following values for CCl_4 : ζ_{Cl}=0.2 bohr^{-2} and ζ_C = 0.2 bohr^{-2}, suggesting both atom types are roughly equally electronegative and that the $\mathbf{r}_{C,Cl}$ barycenters are located precisely at the bond centers. For $CHCl_3$, on the other hand, the optimized heavy atom exponents are ζ_{Cl}=0.09 bohr^{-2} and ζ_C = 0.4 bohr^{-2}, with the hydrogen's exponent ζ_H = 0.7 bohr^{-2}, the latter being a strong electron donor. The latter exponents point to a more diffuse, or electronegative Cl and a more compact C. This behavior is well reflected by a visual analysis of

Figure 2 where one observes that the average (Mulliken) charge on the chlorine transforms from +0.25 |e| in CCl_4 to +0.1 |e| in $CHCl_3$, i.e., acquiring electron density), while the Carbon atom evolves in the opposite direction, from -0.8 to -0.4 |e| , (losing electron density). The disparity in the $CHCl_3$ exponents places the $\mathbf{r}_{H,Cl}$ barycenters much closer to H, reflecting the dynamics of the trajectories where the light H is much more mobile than the chlorines.

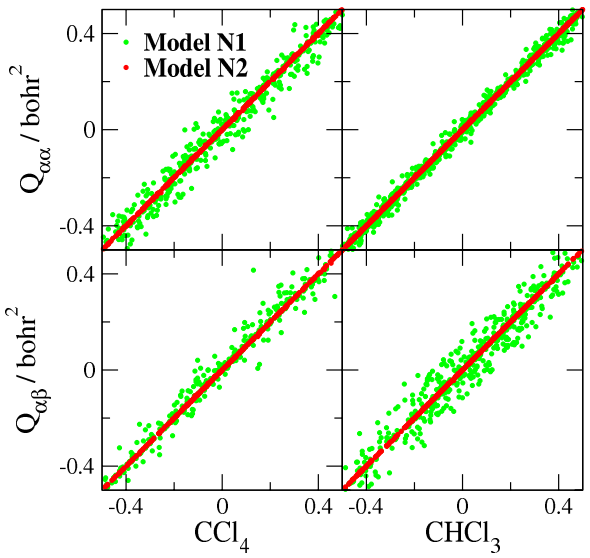


Figure 3B. Correlation diagrams of the fitted quadrupole moments (vertical axis) vs. B3LYP/aug-cc-pVDZ training data (horizontal axis), for CCl_4 and $CHCl_3$ using **Model N1** (green dots) and **Model N2** (red dots) with PIP order $M = 4$. A narrowed range of [-0.5,0.5] bohr^2 of the quadrupole data is shown for a better visual perspective.

Concurrently, we observe that **Model N2** clearly outperforms the traditional **Model N1** for all multipoles: by a factor of 5 for the dipole, and in the case of the quadrupole and octupole it performs dramatically better, by factors of 20 and 14 for CCl_4 and 31 and 10 for $CHCl_3$, respectively, for the $M = 4$ PIP order. This result indicates that while **Model N1** is quite suitable for describing the dipole, which has previously been demonstrated by some of us,[70,71] it is in fact

inadequate for the quadrupole, octupole and likely higher order multipole surfaces. A more detailed examination of the data in Table 2 further reveals that increasing the PIP basis order (i.e. from $M = 3$ to $M = 4$) has a markedly smaller effect on the RMSE than does increasing the model level from **N1** to **N2** while using a bigger set of linear parameters. For instance, the RMSE of the quadrupole moment of CCl_4 improves from 7.4% to 5.9% by increasing the PIP basis from $M = 3$ to $M = 4$ (from 58 to 164 PIPs, a 183% basis increase, cf. Table 1.) At the same time, upgrading the model from **N1** to **N2** reduces the RMSE to 1.2% while increasing the linear basis from 58 to 116 PIPs (a 100% basis increase). Similarly, the RMSE of the octupole moment of $CHCl_3$ is corrected only slightly, from 26.7% to 26.6%, by using the larger PIP basis ($M = 4$) while at the same time the fitting basis increases from 222 to 690 (a 211% increase). Yet, keeping $M = 3$ and employing **Model N2** suppresses the error down to 4.3% with an increased basis (222 to 518 PIPs, a 133% increase). Similar observations can be made for the other instances in Table 2. In other words, **Model N2** coupled with a low order PIP basis provides a much better PIP representation of the multipoles than does **Model N1** coupled with a higher order PIP basis, and is shown to achieve this level with substantially fewer resources. The above statistics suggests that in fitting the multipole moment effective charges and atomic polarizabilities with PIPs, the treatment of the electron density with a well-structured spatial model is more important than using an arbitrarily high order PIP basis. We note however that the degree of efficiency may be different for molecules with compact or less structured electron densities, such as electron deficient molecules and ions; we leave this point for future investigations. For a visual inspection of the performance of the quadrupole and octupole fits, we refer the reader to the correlation plots of individual tensor elements in Figures 3A, 3B and 3C.

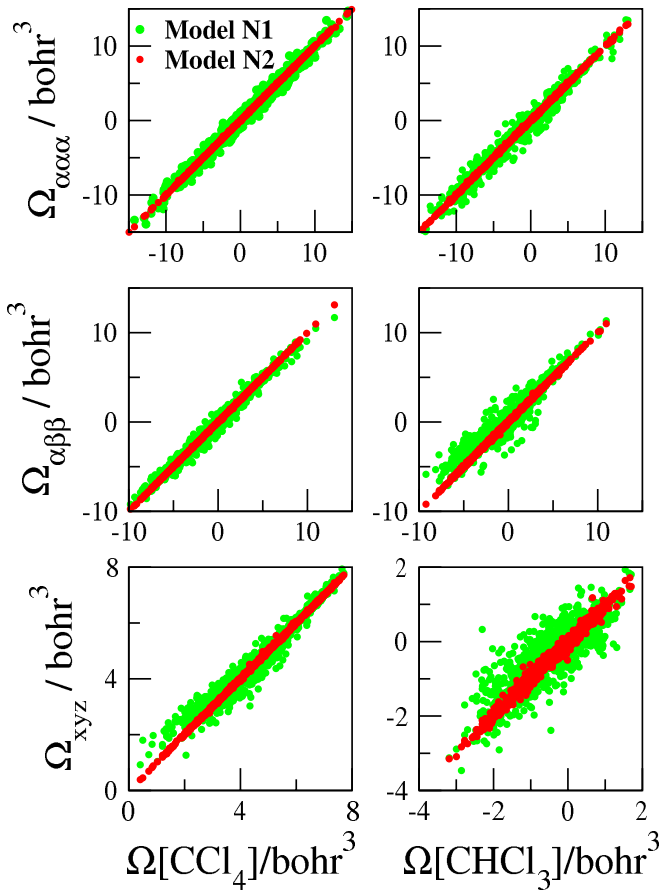


Figure 3C. Correlation diagrams of the fitted octupole moments (vertical axis) vs. B3LYP/aug-cc-pVDZ training data (horizontal axis), for CCl_4 and $CHCl_3$ using **Model N1** (green dots) and **Model N2** (red dots) with PIP order $M = 4$ (with $\alpha, \beta = x, y, z$)

As a separate but related exercise, it is instructive to examine the quality of representation of the total molecular charge. We note that for each of the fitted multipoles we do constrain the total charge to its true molecular value, $Z = 0$ for both molecules, by least-squares fitting of Eq. (2a). This is an important property that determines translational symmetry of the multipoles, the dipole in particular. These results are summarized in Figure 4 by plotting the charges as functions of configuration number in the corresponding training set, arranged by increasing configuration number from high to low potential energies. It is immediately clear that **Model N1** preserves the

total charge, to ~0.001 |e|, for both CCl_4 and $CHCl_3$ dipoles. Similar behavior was reported previously by Braams and Bowman in fitting of dipole moment surfaces.[44,72-74] However, the model completely fails to preserve the total charge to the same order of accuracy for the quadrupole and octupole, with the average errors of ~0.1 |e| and ~1 |e|, respectively. On the other hand, **Model N2** performs remarkably well constraining the total charge for all multipoles. The biggest deviations are observed for the high energy configurations of $CHCl_3$ where the error occasionally reaches ~0.1 |e|.

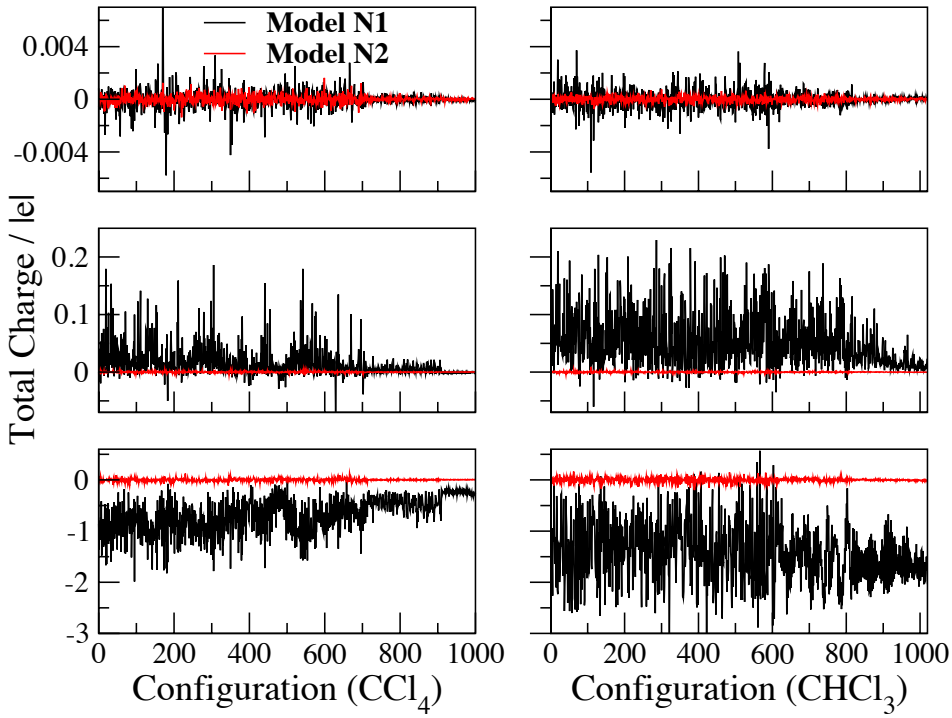


Figure 4. Fits of the total molecular charge $Z(\mathbf{R})$ (Eq. (2a)), where **R** represents a nuclear configuration, using the two models with $M = 4$ and considered separately in the dipole (upper), quadrupole (middle), octupole (lower) procedures for the CCl_4 and $CHCl_3$ training data. The exact molecular charge of the two systems is $Z = 0$.

The presently observed case of the PTS follows the same trend as that of the multipole surfaces, but with a much less dramatic variation in RMSE in terms of the PIP basis and the point charge model level. First of all, for CCl_4 we see in Table 2 that the RMSE of the [*M*=4]@**N1** level (0.9%) is corrected only moderately by the [*M*=4]@**N2** level to 0.6%. For $CHCl_3$ we observe that the RMSE of the [*M*=4]/**N1** level of 1.9% is reduced to 1.7% by application of the [*M*=4]/**N2** level. These are marginal improvements relative even to those of the DMS, which appears to be described well by the standard **Model N1**. (For a visual examination of the fits one is referred to Figures S1, S2 in the SI.) A possible explanation is that since the PTS is defined as the dipole polarizability, it possesses spatial properties of the dipole moment. That is, upon application of a small external field $\delta\mathbf{F}$ the resultant molecular dipole moment is a sum of the permanent dipole and the induced dipole: $\boldsymbol{\mu}(\delta\mathbf{F}) = \boldsymbol{\mu}(0) + \boldsymbol{\mu}^{\mathrm{ind}}(\delta\mathbf{F})$. For a uniform electric field applied arbitrarily along the space-fixed axis *z*, the components of the induced dipole are proportional to the components of the polarizability: $\mu_i^{\mathrm{ind}}/\delta F_z = \alpha_{iz}$. In other words, if one were to fit $\boldsymbol{\mu}(\delta\mathbf{F})$ to *ab initio* data generated with the applied field using the same nuclear configurations as for fitting $\boldsymbol{\mu}(0)$, one could recover the PTS by simple difference of the field-applied and field-free DMSs with an RMSE comparable to that of the DMS fit. (It is to be understood, however, that a PTS derived in this way using an explicit electric field would not be useful as a generic standalone surface.) This marginal improvement of PTS fit using **Model N2** coupled with the fact that the computational complexity of the model scales as $O(N^6)$ compared to $O(N^3)$ for **Model N1**, suggests that the classic Applequist model (**N1**) is actually better suited for PTS PIP representation. The excellent performance of the **Model N1** for PTS has been tested extensively on a variety of systems, including H_2, H_2O, $H_5O_2^+$, CH_4, N_4H^+,[38,39,63] and is not revisited here.

3.3 High temperature "stress test" of the multipole fits

On the other hand, testing the quality of **Model M2** for the multipole surfaces is of paramount importance in the present work. In order to rigorously test the DMS, QMS and OMS, which were fitted to a rather small set (only 1000 configurations) of training data, we run a much longer trajectory (100 ps) than the primitive NVE samples and set it to a high temperature, namely 2000K. Such a trajectory is expected to yield a set of configurations different from those of the training set while sampling regions of the potential energy roughly corresponding to 6000 cm^{-1}, i. e. sufficiently far from the equilibrium configurations, cf. Figure 1. For the purpose of running the extensive simulation we fit the PESs of the two systems using the same training set of configurations as for the other quantities. To fit the PESs we take a 5th order PIP resulting in 208 and 636 linear terms for CCl_4 and $CHCl_3$, respectively, with the corresponding RMSEs of 2.8 and 1.9 cm^{-1}; Table S6 in the SI provides further evidence of the high quality of the PES fits. Thus derived PESs contain both the energy value and its analytic gradient for a given input nuclear configuration. For the subsequent analysis, we save the test data every 0.1 ps resulting in a total of 1000 loosely correlated configurations per system and recalculate the multipoles at the B3LYP/aug-cc-pVDZ level. The results for the OMSs are summarized in Figure 5. Analysis of the DMS and QMS as well as the potential energy and nuclear configuration distributions are provided in the SI in Figures S3A, S3B, S4A, S4B and S5A, S5B.

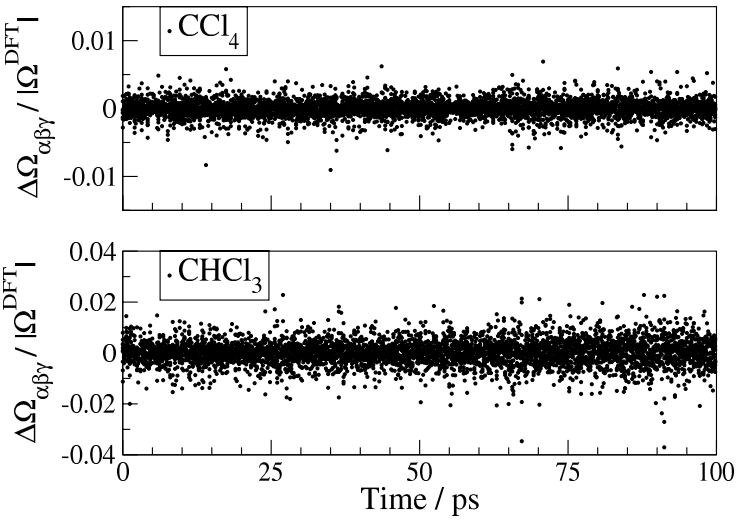


Figure 5. Test of the CCl_4 and $CHCl_3$ octupole moment **Model N2** PIP representations against B3LYP/aug-cc-pVDZ data (DFT) using an extensive MD simulation at 2000K. The ten unique octupole components at each time step are shown as black dots. The trajectories were propagated on the respective PESs.

For a better emphasis of the error we find it useful to plot the ratio of the individual components of the difference tensor to the norm of the test data tensor calculated along the trajectory, i.e. for the octupole moment the quantity $\Delta\Omega_{\alpha\beta\gamma}/|\boldsymbol{\Omega}^{\mathrm{DFT}}|$ is examined, taking the Frobenius norm of the tensor.[75] Visual inspection of the octupole errors, for instance, shows that they stay uniform along the trajectory and roughly within the bounds of the test set errors, 0.8% for CCl_4 and 2.6% for $CHCl_3$ (cf. Table 2). Similar behavior can be described for the quadrupole and dipole errors, shown in Figures S5A, S5B of the SI. These observations point to two important qualities, (*i*) the training data is not overfitted by **Model N2** despite the additional fitted charge sites situated at the $N(N\text{-}1)/2$ additional barycenters since for a generic $(A_1)_{n1}(A_2)_{n2}...(A_\lambda)_{n\lambda}$ molecule the number of symmetry unique pair-species is bound on $[\lambda(\lambda\text{-}1)/2, \lambda(\lambda+1)/2]$ where λ is the number of atom classes ($\lambda = 2$ for CCl_4, $\lambda = 3$ for $CHCl_3$), and (*ii*) the fits show an excellent degree of fidelity considering that only 1000 configurations were used to train the model. In other

words, the **Model N2** multipole surfaces developed in the present work appear to be suitable for a wide range of MD simulations from low to high temperatures, as well as for constructing well-balanced, polarizable electrostatic force fields for non-covalently bound systems relevant for simulation of liquids, in the spirit of recent works of Handley *et al*.[15,16]

4. CONCLUSIONS

We have described an approach to fitting global molecular multipole moment surfaces using permutationally invariant polynomials in conjunction with a basic concept of the distributed multipole theory. The method, named **Model N2** after Hall's original treatment of an *s*-type electron density, involves placing $N(N\text{-}1)/2$ virtual charges (or isotropic polarizabilities) at the bond barycenters in addition to the N atomic sites and using them for the fitting just as is done in the classical N-atom charge model (**Model N1**). The calculations, employing CCl_4 and $CHCl_3$ as test cases, show a major improvement (more than an order of magnitude reduction of the RMSE) of **Model N2** over **Model N1** for the quadrupole and octupole moment surfaces, and a smaller but still significant improvement (about a factor of 5 reduction of the RMSE) for the dipole moment surface. Furthermore, unlike **Model N1**, **Model N2** was found to conserve the molecular charge for all multipoles, which is important for maintaining the translational symmetry of the dipole and the higher multipoles.

Another imperative observation, one specifically pertaining to the role of the PIP order, is that **Model N2** combined with a low-order PIP is better in the RMSE sense than **Model N1** combined with a high order PIP. This finding points to some very practical computational advantages of **Model N2**, namely: (*i*) possibility to use a lower order PIP basis without sacrificing accuracy of the fit implies generating fewer *ab initio* training points to accomplish a reasonable

quality fit. This is extremely important for calculations of large systems with high permutational symmetry where both the generation of a PIP basis[37] and concurrently of *ab initio* data are particularly expensive; (*ii*) since the number of linear parameters in **Model N2** grows quadratically with the number of nuclear classes λ, which rarely exceeds a few units in a typical case, as opposed to the total number of atoms, which can be high, the problem of overfitting is effectively controlled as was demonstrated above.

We note in closing that the polarizability tensor surfaces considered here using the classic Applequist dipole polarizability formulation show only a marginal improvement (about 50% reduction of the RMSE) by the **Model N2** level of treatment. At the same time the computational cost involved in expressing the polarizability with **Model M2** increases quite dramatically from $O(N^3)$ to $O(N^6)$. Therefore, our present conclusion is that **Model N1** appears the more favorable of the two for polarizability fitting using PIPs.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CHE–1855583. This work was also supported in part by research computing resources and technical expertise via a partnership between Kennesaw State University's Office of the Vice President for Research and the Office of the CIO and Vice President for Information Technology.[76] ALK acknowledges the use of computational resources of the Cherry L. Emerson Center.

ORCID

Martina Kaledin; http://orcid.org/0000-0003-1763-3552

Alexey L. Kaledin; http://orcid.org/0000-0003-3112-3989

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

All the data used in study is available from corresponding author.

SUPPLEMENTARY INFORMATION

Formulation of polarizability representation; a covariant symmetrization scheme; CCl_4 and $CHCl_3$ properties at the equilibrium geometry; details of the test set.

REFERENCES

[1] A. J. Stone, *Chem. Phys. Lett.* **1981***, 83*, 233.

[2] A. J. Stone, S. L. Price, *J. Phys. Chem.* **1988**, *92*, 3325.

[3] C. H. Faermant, S. L. Price, *J. Am. Chem. Soc.* **1990**, *112*, 4915.

[4] P. Ren, J. W. Ponder, *J. Phys. Chem. B* **2003**, *107*, 5933.

[5] J.-P. Piquemal, L. Perera, G. A. Cisneros, P. Ren, L. G. Pedersen, T. A. Darden, *J. Chem. Phys.* **2006**, *125*, 054511.

[6] N. Gresh, G. A. Cisneros, T. A. Darden, J.-P. Piquemal, *J. Chem. Theory Comput.* **2007**, *3*, 1960.

[7] N. Gresh, *J. Comput. Chem.* **1995**, *16*, 856.

[8] J.-P. Piquemal, H. Chevreau, N. Gresh, *J. Chem. Theory Comput.* **2007**, *3*, 824.

[9] D. M. Elking, G. A. Cisneros, J.-P. Piquemal, T. A. Darden, L. G. Pedersen, *J. Chem. Theory Comput.* **2010**, *6*, 190.

[10] S. Patel, C. L. Brooks III, *J. Comput. Chem.* **2004**, *25*, 1.

[11] S. Patel, A. D. MacKerell, C. L. Brooks III, *J. Comput. Chem.* **2004**, *25*, 1504.

[12] G. A. Kaminski, H. A. Stern, B. J. Berne, R. Friesner, *J. Phys. Chem. A* **2004**, *108*, 621.

[13] S. Houlding, S. Y. Liem, P. L. A. Popelier, *Int. J. Quantum Chem.* **2007**, *107*, 2817.

[14] M. G. Darley, C. M. Handley, P. L. A. Popelier, *J. Chem. Theory Comput.* **2008**, *4*, 1435.

[15] C. M. Handley, P. L. A. Popelier, *J. Chem. Theory Comput.* **2009**, *5*, 1474.

[16] C. M. Handley, G. I. Hawe, D. B. Kell, P. L. A. Popelier, *Phys. Chem. Chem. Phys.* **2009**, *11*, 6365.

[17] V. R. Saunders, C. Freyria-Fava, R. Dovesi, L. Salasco, C. Roetti, *Mol. Phys.* **1992**, 77, 629.

[18] L. Joubert, P. L. A. Popelier, *Mol. Phys.* **2002**, *100*, 3357.

[19] P. M. A. Sherwood, *Vibrational spectroscopy of solids*; Cambridge University Press, **1972**.

[20] D. A. McQuarrie, *Statistical Mechanics*; Harper & Row: New York, **1976**.

[21] J. Kessler, P. Bouř, *J. Chem. Theory Comput.* **2022**, *18*, 1780.

[22] S. Iuchi, S. Izvekov, G. A. Voth, *J. Chem. Phys.* **2007**, *126*, 124505.

[23] F. Paesani, S. Iuchi, G. A. Voth, *J. Chem. Phys.* **2007**, *127*, 074506.

[24] F. Wang, K. D. Jordan, *J. Chem. Phys.* **2002**, *116*, 6973.

[25] H. Jiang, K. D. Jordan, C. E. Taylor, *J. Phys. Chem. B* **2007**, *111*, 6486.

[26] C. J. Burnham, J. C. Li, S. S. Xantheas, M. Leslie, *J. Chem. Phys.* **1999**, *110*, 4566.

[27] C. J. Burnham, S. S. Xantheas, *J. Chem. Phys.* **2002**, *116*, 1500.

[28] C. J. Burnham, S. S. Xantheas, *J. Chem. Phys.* **2002**, *116*, 5115.

[29] V. Babin, C. Leforestier, F. Paesani, *J. Chem. Theory Comput.* **2013**, *9*, 5395.

[30] G. R. Medders, F. Paesani, *J. Chem. Theory Comput.* **2015**, *11*, 1145.

[31] M. in het Panhuis, P. L. A. Popelier, R. W. Munn, J. G. Angyan, *J. Chem. Phys.* **2001**, *114*, 7951.

[32] A. Brown, A. B. McCoy, B. J. Braams, Z. Jin, J. M. Bowman, *J. Chem. Phys.* **2004**, *121*, 4105.

[33] X. Huang, A. B. McCoy, J. M. Bowman, L. M. Johnson, C. Savage, F. Dong, D. J. Nesbitt, *Science* **2006**, *311*, 60.

[34] A. B. McCoy, B. J. Braams, A. Brown, X. Huang, Z. Jin, J. M. Bowman, *J. Phys. Chem. A* **2004**, *108*, 4991.

[35] X. Huang, L. M. Johnson, J. M Bowman, A. B. McCoy, *J. Amer. Chem. Soc.* **2006**, *128*, 3478.

[36] B. J. Braams, J. M. Bowman, *Int. Rev. Phys. Chem.* **2009**, *28*, 577.

[37] Z. Xie, J. M. Bowman, *J. Chem. Theory Comput.* **2010**, *6*, 26.

[38] O. Omodemi, S. Sprouse, D. Herbert, M. Kaledin, A. L. Kaledin, *J. Chem. Theory Comput.* **2022**, *18*, 37.

[39] O. Omodemi, M. Kaledin, A. L. Kaledin, *J. Comput. Chem*. **2022**, *43*, 1495.

[40] C. Qu, J. M. Bowman, *J. Chem. Phys.* **2019**, *150*, 141101.

[41] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press: Oxford, **1996**.

[42] G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, R. Car, *Phys. Chem. Chem. Phys.* **2020**, *22*, 10592.

[43] Y. Zhang, S. Ye, J. Zhang, C. Hu, J. Jiang, B. Jiang, *J. Phys. Chem. B* **2020**, *124*, 7284.

[44] C. Qu, Q. Yu, J. M. Bowman, *Annu. Rev. Phys. Chem.* **2018**, *69*, 151.

[45] S. Manzhos, T. Carrington, *Chemical Reviews* **2021**, *121*, 10187.

[46] Q. Yu, C. Qu, P. L. Houston, R. Conte, A. Nandi, J. M. Bowman, *J. Phys. Chem. Lett.* **2022**, *13*, 5068.

[47] A. J. Stone, *The Theory of Intermolecular Forces*; Oxford University Press: Oxford, **1996**.

[48] J. G. Angyan, G. Jansen, M. Loos, C. Haettig, B. A. Hess, *Chem. Phys. Lett.* **1994**, *219*, 267.

[49] R. S. Mulliken, *J. Chem. Phys.* **1955**, *23*, 1833.

[50] F. Vigné-Maeder, P. Claverie, *J. Chem. Phys.* **1988**, *88*, 4934.

[51] M. M. Ghahremanpour, P. J. van Maaren, C. Caleman, G. R. Hutchison, D. van der Spoel, *J. Chem. Theory Comput.* **2018**, *14*, 5553.

[52] G. G. Hall, *Chem. Phys. Lett.* **1973**, *20*, 501.

[53] A. D. Tait, G. G. Hall, *Theoret. Chim. Acta (Berl.)* **1973**, *31*, 311.

[54] G. G. Hall, *Int. J. Quantum. Chem.* **1975**, *9*, 279.

[55] D. Martin, G. G. Hall, *Theoret. Chim. Acta (Berl.)* **1981**, *59*, 281.

[56] A. Tokmakoff, M. J. Lang, D. S. Larsen, G. R. Fleming, V. Chernyak, S. Mukamel, *Phys. Rev. Lett*. **1997**, *79*, 2702.

[57] K. Tominaga, K. Yoshihara, *Phys. Rev. A* **1997**, *55*, 831.

[58] J. Applequist, J. R. Carl, K.-K. Fung, *J. Am. Chem. Soc.* **1972**, *94*, 2952.

[59] B. T. Thole, *Chem. Phys.* **1981**, *59*, 341.

[60] I. Harczuk, O. Vahtras, H. Ågren, *Phys. Chem. Chem. Phys.* **2016**, *18*, 8710.

[61] E. Anderson, Z. Bai, C. Bischof, et al., LAPACK Users' Guide, 3rd ed. SIAM, Philadelphia, PA, **1999**.

[62] https://www.intel.com/content/www/us/en/developer/articles/release-notes/onemkl-release-notes.html

[63] O. Omodemi, R. Revennaugh, J. Riley, A. L. Kaledin, M. Kaledin, *J. Chem. Phys.* **2022**, *157*, 154303.

[64] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[65] T. H. Dunning Jr., *J. Chem. Phys.* **1989**, *90*, 1007.

[66] Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, R. A. DiStasio Jr., *Sci. Data* **2019**, *6*, 152.

[67] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio Jr., M. Ceriotti, *PNAS* **2019**, *116*, 3401.

[68] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T.Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 16, Revision A.03 (Gaussian, Inc., Wallingford, CT, **2016**).

[69] MOLPRO, version 2019.2, a package of ab initio programs, H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, S. J. Bennie, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C.

Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, S. J. R. Lee, Y. Liu, A. W. Lloyd, Q. Ma, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, T. F. Miller III, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, and M. Welborn, see https://www.molpro.net.

[70] T. Esser, H. Knorke, K. R. Asmis, W. Schollkopf, Q. Yu, C. Qu, J. M. Bowman, M. Kaledin, *J. Phys. Chem. Lett*. **2018**, *9*, 798.

[71] M. Kaledin, A. L. Kaledin, J. M. Bowman, *J. Phys. Chem A* **2006**, *110*, 2933.

[72] X. C. Huang, B. J. Braams, J. M. Bowman, *J. Chem. Phys.* **2005**, *122*, 044308.

[73] X. Huang, B. J. Braams, J. M. Bowman, *J. Phys. Chem. A* **2006**, *110*, 445.

[74] C. Qu, J. M. Bowman, *J. Chem. Phys.* **2018**, *148*, 241713.

[75] K. E. Atkinson "An Introduction to Numerical Analysis" 2nd ed.; Wiley, New York, **1989**.

[76] Research Computing, Kennesaw State University, 2021, Digital Commons Training Materials. 10. https://digitalcommons.kennesaw.edu/training/10 (accessed November 3, 2022).

Supplementary Information for
Revisiting the $N(N+1)/2$-site s-type Gaussian charge model for permutationally invariant polynomial fitting of molecular tensor properties

Sophia A. Kudratov,[1] Martina Kaledin,[1] Alexey L. Kaledin[2,*]
[1] *Department of Chemistry & Biochemistry, Kennesaw State University, 370 Paulding Ave NW, Box # 1203, Kennesaw, Georgia, 30144*
[2] *Cherry L. Emerson Center for Scientific Computation and Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, Georgia, 30322*

Table of contents:

S1. Formulation of polarizability representation

Previously it was shown by us that **G** may be functionalized using PIPs by introducing nuclear configuration **R** dependence into its elements, such as the isotropic atomic polarizabilities, i.e. $a_k \equiv a_k(\mathbf{R})$ for an atomic site *k*. This was achieved by expanding Eq. 4 in a power series

$$(\mathbf{A}^{-1} + \mathbf{T})^{-1} \cong \mathbf{A} - \lambda_2(\mathbf{r})\mathbf{ATA} + \lambda_3(\mathbf{r})\mathbf{ATATA} - \lambda_4(\mathbf{r})\mathbf{ATATATA} + \cdots \quad (S1)$$

with $\lambda_n(\mathbf{r})$ being introduced as configuration dependent short-range scalar correction factors that tend to 1 at long range, with the requirement of $\lambda_1 \equiv 1$. **A** is a diagonal matrix with elements a_k. We then express the atomic polarizabilities and the correction factors using PIPs that decay to zero at large separation,

$$\alpha_k(\mathbf{R}) = \alpha_k^{(0)} + \sigma_k(\mathbf{R}') \sum_{i=1}^{L} c_i^{\text{atom}} u_{k,i}(\mathbf{R}) \quad (S2.a)$$

$$\alpha_{kk'}(\mathbf{R}) = S_{kk'}(\mathbf{R})\alpha_{kk'}^{(0)} + S_{kk'}(\mathbf{R}) \sum_{i=1}^{L} c_i^{\text{pair}} u_{kk',i}(\mathbf{R}) \quad (S2.b)$$

$$\lambda_n(\mathbf{R}) = 1 + \sum_{i=1}^{L} c_{n,i} u_i(\mathbf{R}) \quad (S2.c)$$

As can be seen, in the limit of separated atoms, $\alpha_k(\mathbf{R} \to \infty)$ tends to the atomic polarizability $a_k^{(0)}$ for $k = 1, N$, which is a naturally required limit. At the same time, $\alpha_{kk'}(\mathbf{R} \to \infty) = 0$, due to the vanishing of overlaps. The choice of $\alpha_{kk'}^{(0)}$ is completely arbitrary, but should reflect physical properties of the molecule, in particular the atomic pair *kk'*. Several possible forms have been considered, such as (i) $\alpha_{kk'}^{(0)} = 1$, (ii) $\alpha_{kk'}^{(0)} = (\alpha_k^{(0)} + \alpha_{k'}^{(0)})/2$, (iii) $\alpha_{kk'}^{(0)} = \sqrt{\alpha_k^{(0)}\alpha_{k'}^{(0)}}$. The best performance was found with (iv) the 'reduced mass' formula, namely,

$$\frac{1}{\alpha_{kk'}^{(0)}} \equiv \frac{1}{\alpha_k^{(0)}} + \frac{1}{\alpha_{k'}^{(0)}} \qquad (S3)$$

which is also the formula for the exponent in the Gaussian overlap integral (cf. Eq. 2g). Back-substitution of Eq. S2 into Eq. S1 yields a non-linear function in the expansion coefficients {c}. In our recent paper we demonstrated that it is possible to linearize Eq. S1 with only a small penalty in computational accuracy, as follows,

$$(\mathbf{A}^{-1} + \mathbf{T})^{-1} \cong (\mathbf{A}_0^{-1} + \mathbf{T})^{-1} + \Delta\mathbf{A}(\mathbf{R}) + \sum_{n=2}^{l\to\infty} \Delta\lambda_n(\mathbf{R})\mathbf{M}_n(\mathbf{R}) \qquad (S4)$$

where $\mathbf{M}_n = (-1)^{n-1}\mathbf{A}_0^{1/2}\left(\mathbf{A}_0^{1/2}\mathbf{T}\mathbf{A}_0^{1/2}\right)^{n-1}\mathbf{A}_0^{1/2}$, $\Delta\lambda_n(\mathbf{r}) = \lambda_n(\mathbf{r}) - 1$ and $\Delta\mathbf{A}(\mathbf{R})$ is a correction matrix. Relaying the above linear approximation to the 3x3 form gives the *l*-order Thole-linearized (TL) polarizability tensor. In the below we have switched to cumulative *p*, *q* labels as those of atom sites and bond barycenters, in the order of 1,...,*N*,...,*N*(*N*+1)/2,

$$\begin{aligned}\boldsymbol{\alpha}_{\mathrm{TL}}^{(l)}(\mathbf{R}) = & \sum_{p,q}^{N(N+1)/2} [(\mathbf{A}_0^{-1} + \mathbf{T}')^{-1}]_{p,q} + \mathbf{1} \sum_{p}^{N(N+1)/2} \Delta\alpha_p(\mathbf{R}) - \Delta\lambda_2(\mathbf{R}) \sum_{p\neq q}^{\frac{N(N+1)}{2}} \alpha_p^{(0)}\alpha_q^{(0)}\mathbf{T}'_{pq} \\ & + \Delta\lambda_3(\mathbf{R}) \sum_{p,q}^{\frac{N(N+1)}{2}} \alpha_p^{(0)}\alpha_q^{(0)} \sum_{s\neq p,q}^{\frac{N(N+1)}{2}} \alpha_s^{(0)}\mathbf{T}'_{ps}\mathbf{T}'_{sq} \\ & - \Delta\lambda_4(\mathbf{R}) \sum_{p,q}^{\frac{N(N+1)}{2}} \alpha_p^{(0)}\alpha_q^{(0)} \sum_{\substack{s(\neq p) \\ s'(\neq s,q)}}^{\frac{N(N+1)}{2}} \alpha_s^{(0)}\alpha_{s'}^{(0)}\mathbf{T}'_{ps}\mathbf{T}'_{ss'}\mathbf{T}'_{s'q} + \cdots \qquad (S5)\end{aligned}$$

with $\Delta\alpha_p(\mathbf{R}) = \alpha_p(\mathbf{R}) - \alpha_p^{(0)}$, and the index *p* running over all the sites, i.e. *p* = 1, ..., *N*, ..., *N*(*N*+1)/2. In the present calculations, as in the previous ones, we take *l* = 6. The Thole-modified dipole tensors in the internuclear distances r_{pq} are

$$\left(\mathbf{T}'_{pq}\right)_{ij} = f_1(r_{pq})\frac{\delta_{ij}}{r_{pq}^3} - 3f_2(r_{pq})\frac{r_i r_j}{r_{pq}^5} \qquad (S6)$$

with $r_i = r_{pq,x}, r_{pq,y}, r_{pq,z}$ for the respective i = 1, 2, 3 and

$$f_1(r_{pq}) = 1 - (a^2u^2/2 + au + 1)e^{-au} \qquad (S7a)$$

$$f_2(r_{pq}) = 1 - (a^3u^3/6 + a^2u^2/2 + au + 1)e^{-au} \qquad (S7b)$$

with $u \equiv \left(\alpha_p^{(0)}\alpha_q^{(0)}\right)^{-1/6} r_{pq}$. The value of a = 0.5 was used in the calculations.

It is important to note that the dimensions of the linear parameter space for the aforementioned polarizability are different from those of the multipole representations. Common for both, the atomic point charges and atomic polarizabilities for each atom and atom-pair group is represented by its L elements (Eq. S2). However, for the polarizability tensor there are 5 additional $\Delta\lambda_n$ (n = 2 - 6) quantities with L elements each.

The isotropic and anisotropic polarizability components are defined as

$$\alpha^{\text{iso}} = \frac{1}{3}\text{tr}[\boldsymbol{\alpha}] \qquad (S8a)$$

$$\alpha^{\text{aniso}} = \frac{1}{\sqrt{2}}\left[\left(\alpha_{xx} - \alpha_{yy}\right)^2 + \left(\alpha_{yy} - \alpha_{zz}\right)^2 + (\alpha_{zz} - \alpha_{xx})^2 + 6\left(\alpha_{xy}^2 + \alpha_{xz}^2 + \alpha_{yz}^2\right)\right]^{1/2} \qquad (S8b)$$

and they are shown in the figures below.

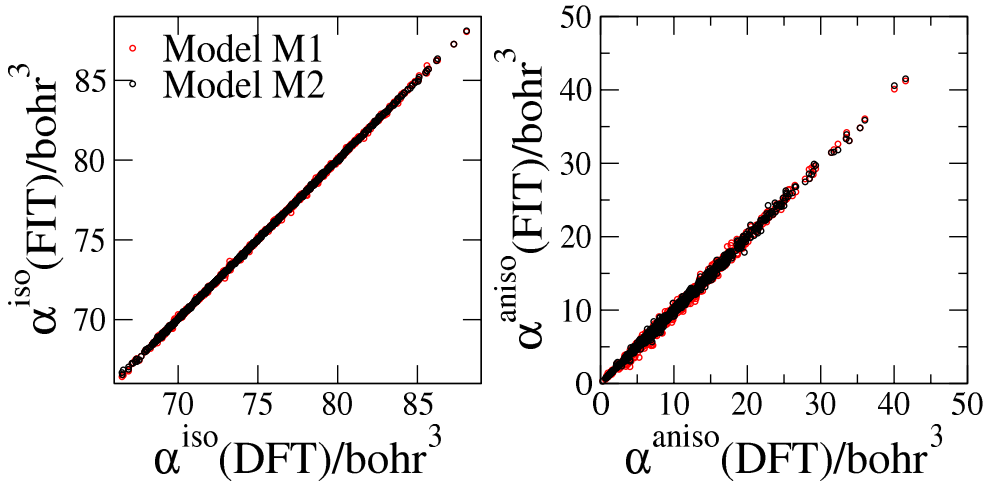


Figure S1. CCl_4 polarizabilities computed with an M = 4 PIP using the two models. DFT = B3LYP/aug-cc-pVDZ.

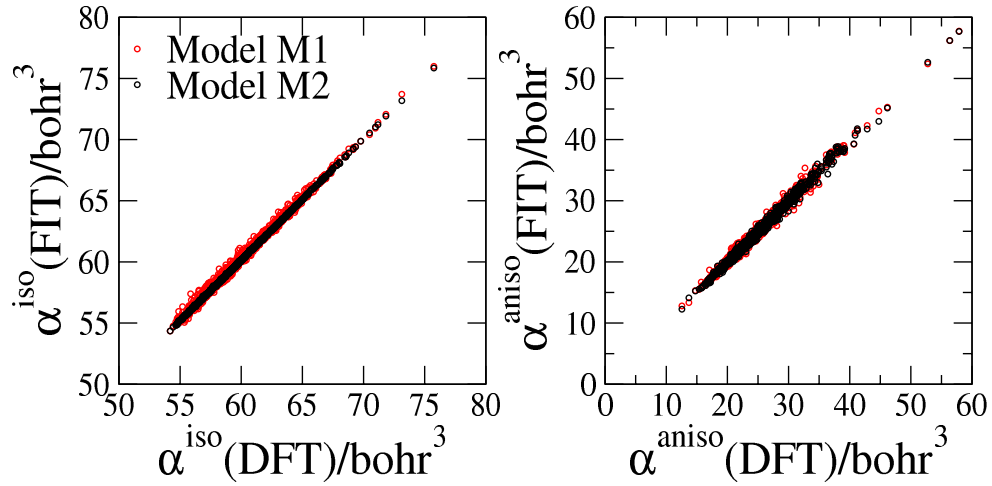


Figure S2. $CHCl_3$ polarizabilities computed with an M = 4 PIP using the two models. DFT = B3LYP/aug-cc-pVDZ.

S2. A covariant symmetrization scheme

1. determine the number of non-zero powers in a monomial: M_NZ{monomial}

   Ex: M_NZ$\{y_{12}^{2}y_{13}^{1}y_{14}^{0}y_{15}^{0}y_{23}^{1}y_{24}^{0}y_{25}^{3}\}$ = 4

2. define quantity MAX_SYM = min[N_ATOM - 1, M_NZ]

   where N_ATOM is the number of atoms

3. define quantity MIN_SYM = M_NZ - MAX_SYM

4. compute the number of times (F_k) the atom index (k) shows up in the monomial

   Ex: for $y_{12}^{2}y_{13}^{1}y_{14}^{0}y_{15}^{0}y_{23}^{1}y_{24}^{0}y_{25}^{3}$

   F_1=2, F_2=3, F_3=2, F_4=0, F_5=1

5. Compare F_k against the quantity F^* = max[MAX_SYM - 1, 1]:

   a) set by default b_k = 1

   b) if F_k > MIN_SYM; then

   if $F_k \geq F^*$; then b_k = -1

S3. CCl_4 and $CHCl_3$ properties at the B3LYP/aug-cc-pVDZ equilibrium geometries.

Table S1. XYZ geometry of CCl_4 in Å.

C	0.000000	0.000000	0.000000
Cl	1.035088	1.035088	1.035088
Cl	-1.035088	-1.035088	1.035088
Cl	-1.035088	1.035088	-1.035088
Cl	1.035088	-1.035088	-1.035088

Table S2. XYZ geometry of $CHCl_3$ in Å.

Cl	0.000000	1.703706	-0.084668
Cl	1.475453	-0.851853	-0.084668
Cl	-1.475453	-0.851853	-0.084668
H	0.000000	0.000000	1.551118
C	0.000000	0.000000	0.461157

Table S3. Elements of the dipole and quadrupole moments at the equilibrium geometry in a.u. calculated using DFT=B3LYP/aug-cc-pVDZ and the fitting procedure.

		z	xy	xz	yz
DFT	CCl_4	0	0	0	0
FIT	CCl_4	0	0	0	0
DFT	$CHCl_3$	0.436915	-0.5347746	-0.5347746	1.0695493
FIT	$CHCl_3$	0.43701897	-0.5347614	-0.5347614	1.0695225

Table S4. Polarizability tensor components at the equilibrium geometry in a.u. calculated using DFT=B3LYP/aug-cc-pVDZ and the fitting procedure.

		xx	yy	zz	xy	xz	yz
DFT	CCl_4	70.12	70.12	70.12	0.00	0.00	0.00
FIT	CCl_4	70.17	70.17	70.17	0.00	0.00	0.00
DFT	$CHCl_3$	63.09	63.09	43.74	0.00	0.00	0.00
FIT	$CHCl_3$	63.35	63.35	43.94	0.00	0.00	0.00

Table S5. Octupole moment components at the equilibrium geometry in a.u. calculated using DFT=B3LYP/aug-cc-pVDZ and the fitting procedure.

		xxx	yyy	zzz	xyy	xxy	xxz	xzz	yzz	yyz	xyz
G16	CCl_4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	8.376
FIT	CCl_4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	8.378
G16	$CHCl_3$	0.000	5.449	4.573	0.000	-5.449	-4.160	0.000	0.000	-4.160	0.000
FIT	$CHCl_3$	0.000	5.456	4.572	0.000	-5.456	-4.161	0.000	0.000	-4.161	0.000

Table S6. Vibrational frequencies in cm^{-1} of the PES fits and the DFT=B3LYP/aug-cc-pVDZ data.

ω_n	CCl4(DFT)	CCl4(PES)	CHCl3(DFT)	CHCl3(PES)
1	214.4	214.4	256.6	256.6
2	214.4	214.4	256.6	256.6
3	311.5	311.6	361.4	362.4
4	311.5	311.6	660.8	661.1
5	311.5	311.6	732.9	732.7
6	451.1	451.2	732.9	732.7
7	744.5	744.7	1206.4	1206.2
8	744.5	744.7	1206.4	1206.2
9	744.5	744.7	3180.6	3184.4

S4. Details of the test set

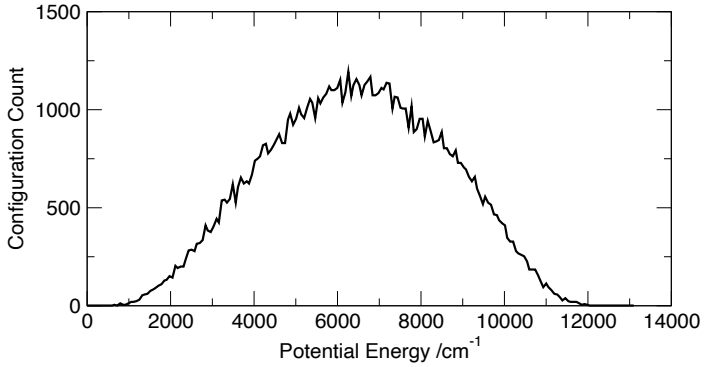


Figure S3A. The potential energy distributions of the CCl_4 test set generated with a 2000 K trajectory propagated on the fitted PES.

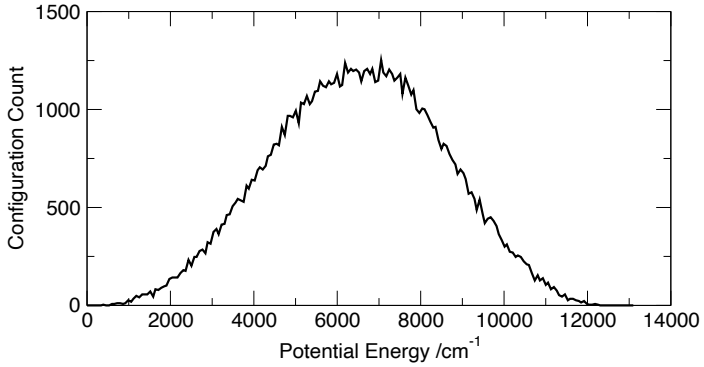


Figure S3B. The potential energy distributions of the $CHCl_3$ test set generated with a 2000K trajectory propagated on the fitted PES.

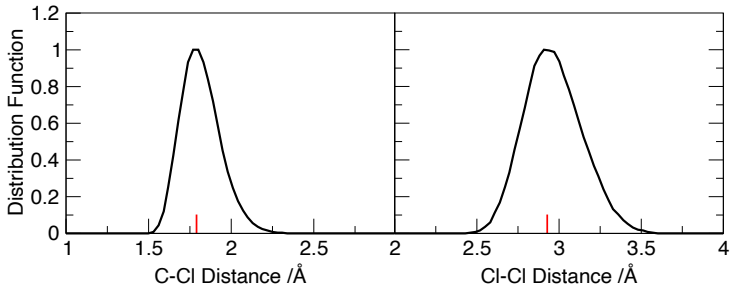


Figure S4A. Internuclear distance distributions of CCl_4 in the test set generated with a 2000 K trajectory propagated on the fitted PES. The red sticks mark the equilibrium geometry values.

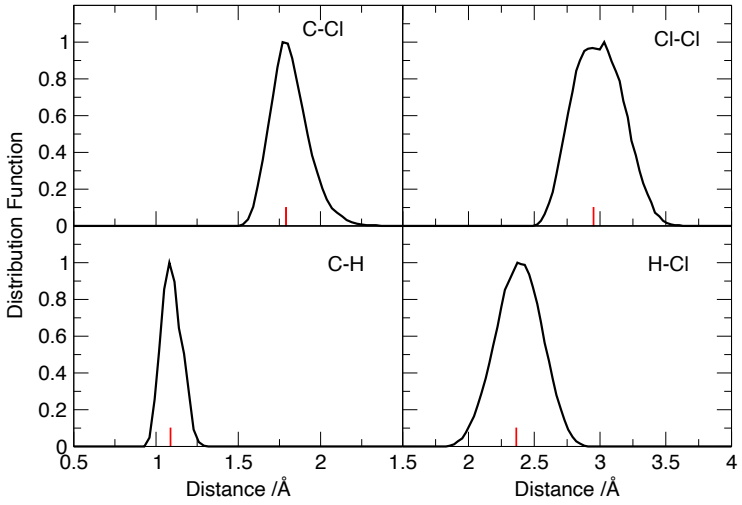


Figure S4B. Internuclear distance distributions of $CHCl_3$ in the test set generated with a 2000 K trajectory propagated on the fitted PES. The red sticks mark the equilibrium geometry values.

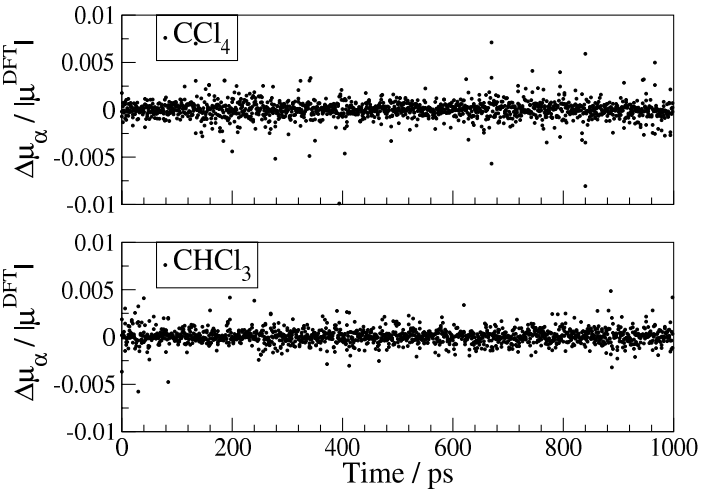


Figure S5A. Test of the CCl_4 and $CHCl_3$ dipole moment **Model N2** PIP representation errors against DFT=B3LYP/aug-cc-pVDZ data using an extensive MD simulation at 2000 K. The three dipole components at each time step are shown as black dots. The trajectories were propagated on the respective PESs.

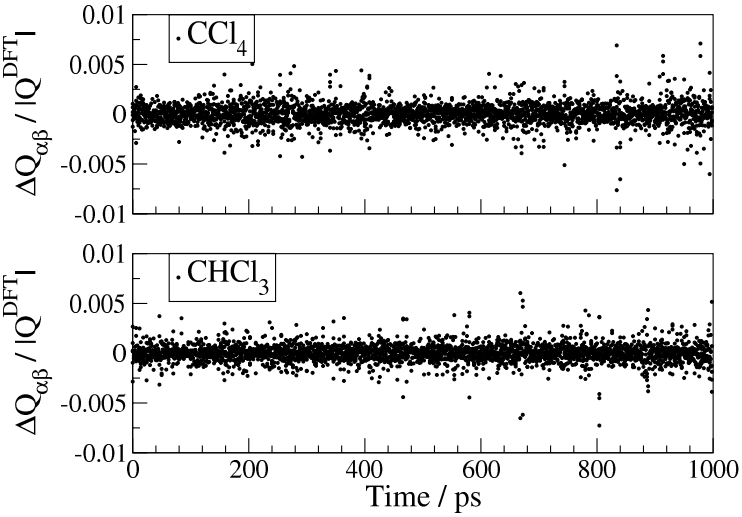


Figure S5B. Test of the CCl_4 and $CHCl_3$ quadrupole moment **Model N2** PIP representation errors against DFT=B3LYP/aug-cc-pVDZ data using an extensive MD simulation at 2000 K. The six unique quadrupole components at each time step are shown as black dots. The trajectories were propagated on the respective PESs.