Software News and Updates

OPEP: A Tool for the Optimal Partitioning of Electric Properties

JÁNOS G. ÁNGYÁN,[1] CHRISTOPHE CHIPOT,[1] FRANÇOIS DEHEZ,[1] CHRISTOF HÄTTIG,[2] GEORG JANSEN,[3] CLAUDE MILLOT[1]

[1]*Equipe de chimie et biochimie théorique, Unité Mixte de Recherche CNRS/UHP 7565, Institut nancéien de chimie moléculaire, Université Henri Poincaré—Nancy I, B.P. 239, 54506 Vandœuvre-lès-Nancy, France*

[2]*Forschungszentrum Karlsruhe GmbH, Institut für Nanotechnologie, Postfach 3640, 76021 Karlsruhe, Germany*

[3]*LASIR—Laboratoire de Spectrochimie Infrarouge et Raman, Unité Mixte de Recherche 8516, Université des Sciences et Technologies de Lille, Bâtiment C5, 59655 Villeneuve d'Ascq cedex, France*

Received 13 August 2002; Accepted 11 November 2002

Abstract: OPEP is a suite of FORTRAN programs targeted at the optimal partitioning of molecular electric properties. It includes an interactive module for the construction of Cartesian grids of points, on which either the molecular electrostatic potential or the induction energy is mapped. The generation of distributed multipoles and polarizabilities is achieved using either the formalism of the normal equations of the least-squares problem, which restates the fitting procedure in terms of simple matrix operations, or a statistical approach, which provides a pictorial description of the distributed models of multipoles and polarizabilities, thereby allowing the pinpointing of pathological cases. Molecular symmetry is accounted for by means of local atomic frames, which are generated in an automated fashion. A Tcl/Tk graphical user interface wraps the suite of programs, thereby making OPEP a user-friendly package for building models of distributed multipoles and polarizabilities. OPEP is an open-source suite of programs distributed free of charge under the GNU general public license (GPL) at http://www.lctn.uhp-nancy.fr/Opep.

© 2003 Wiley Periodicals, Inc. J Comput Chem 24: 997–1008, 2003

Key words: distributed multipoles; distributed polarizabilities; local frames; open-source program

Introduction

One stumbling block of the road towards the easy and routine use of statistical mechanical simulations to investigate a variety of chemical systems, including molecular liquids, solutions, and biological macromolecules, is the difficulty of determining reliable intermolecular interaction parameters. Whereas, the van der Waals component of molecular force fields can be shown to be relatively transferable, which is illustrated by the countless sets of parameters available for atoms in various chemical environments, parameterization of both the electrostatic and the induction contributions usually requires a more careful, case-by-case treatment.

Parameterization of electrostatic forces in terms of atomic quantities is conceptually straightforward. It relies upon a partitioning of the total charge density into pieces that correspond to the atoms, and a selection of the leading terms of the classical electrostatic forces between these atomic distributions. Among the numerous propositions that have flourished in the literature, point charges fitted to the electrostatic potential have proven to be the most widely accepted method to obtain electrostatic interaction parameters. Following the pioneering work of Cox and Williams,[1] numerous variants of this method have been suggested in the past 20 years.[2,3] The treatment of induction forces, related to the electric polarization of the interacting molecules or fragments, raises much more technical difficulties. On the one hand, the representation of these forces by means of atomic parameters inevitably leads to the tricky issue of partitioning an essentially nonlocal quantity, namely the polarizability density,[4,5] into atomic

Correspondence to: C. Chipot; e-mail: chipot@lctn.uhp-nancy.fr

© 2003 Wiley Periodicals, Inc.

contributions. On the other hand, the many-body nature[6] of these forces increases considerably the cost of large-scale, statistical simulations,[7–10] thus encouraging the molecular modeling community to replace this term with pairwise additive, effective potentials obtained by rescaling the electrostatic component. A number of explicit parameterization schemes of induction forces[11–13] follow essentially the method of Applequist,[14] based on interacting atomic dipole polarizabilities. Among the noteworthy alternatives, an interesting approach has been devised by Nakagawa for fitting distributed polarizabilities and hyperpolarizabilities to the perturbed potential of a molecule polarized by a point charge.[15–17] Similar in spirit, the approach of Alkorta et al. relies upon the knowledge of the induction energy to derive atomic polarizabilities.[18]

The significant decrease of the computer price/performance ratio witnessed in recent years has not only made accessible the treatment of increasingly larger chemical systems, but it has also opened the way to the use of more sophisticated and more accurate intermolecular interaction models. The most important concepts have been formulated by Claverie,[19] who showed the advantages of a "piece-wise" parameterization of the physically distinct contributions to the intermolecular forces, and by Stone,[20,21] who developed an elegant and compact formalism to obtain distributed electrostatic and induction interaction parameters from basis-set represented, monomer wave functions. Distributed multipoles and distributed polarizabilities[21–23] allow us, in principle, to compute highly accurate electrostatic and induction energies. The resulting models, however, soon become very complicated, involving a cumbersome plethora of terms. Furthermore, because sufficiently precise wave functions cannot be obtained for too large a molecular fragment, resorting to the concept of transferability appears to be difficult to avoid.

Very few procedures for partitioning the molecular volume into atomic, or somewhat larger, fragments possess transferability properties. The definition of atoms in molecules,[24] according to Bader, based on the topological properties of the molecular charge densities, is one of them. A straightforward combination of the partitioning of the physical space into atomic regions with the formalism of distributed polarizabilities[25] and multipoles led to the approach of topologically partitioned electric properties (TPEP),[26–28] which, at the price of relatively costly numerical integrations over atomic basins, provides accurate and transferable electrostatic models.[29]

TPEP models, involving several hundreds to thousands of independent parameters, can be used for benchmark purposes, but are obviously impractical in routine molecular simulations. This shortcoming has naturally raised the question as to whether or not a compromise between accuracy and tractability could be found easily. In other words, can the quality of TPEP models be attained, yet without wasting computational time in tedious, lengthy multipole expansions. The latter corresponds to countless, tiny, nonlocal components that result in models of distributed polarizability extremely difficult to handle in a practical fashion. The route that we have decided to explore consists of increasing moderately the complexity of the model, guided by our experience gained from far more sophisticated TPEP models, and determining multipole and polarizability components by means of a fitting procedure. The resulting model of optimally partitioned electrostatic properties (OPEP) appears to be a very promising alternative for obtaining reliable electrostatic and induction parameters that can be utilized safely in the statistical simulations of simple liquids and more complex biological systems.

The underlying concept of OPEP applies in a straightforward manner when the electrostatic and induction energy data sets on a grid of points around the molecule are generated by some other means than the TPEP model. The electrostatic potential can be generated on a grid of points from the solution of the Poisson equation with the charge density of a quantum chemically determine wave function. Besides the "standard" way of proceeding,[1,30–34] other possible choices involve the fit to the exact multipolar part of the electrostatic potential[35] that satisfies the Laplace equation and is generated easily for charge densities expanded in Gaussian basis functions, through an expansion of overlap multipole moments. High-rank distributed multipole moment expansions, following the description of Stone[36] or using the Mulliken partitioning according to Sokalski and Poirier[37] provide good approximations of the exact multipolar potential. Direct generation of induction energy data bases by a finite perturbation (FP) approach, via quantum chemical calculations in the presence of point charges, constitutes a straightforward, but also expensive, approach.[38] Special care should be taken to avoid spurious effects resulting from nonlinearity problems, as well as penetration of charge densities. Alternatively, approximate, infinitesimal second-order perturbation can be employed to obtain induction energies from restricted Hartree-Fock (RHF) wave functions.[39] The latter method allows one to avoid nonlinearity artifacts, but penetration should be taken into account in the construction of the grid. Finally, the TPEP scheme constitutes an attractive procedure to evaluate the linear response part of the approximate multipolar induction energy[40] at the time-dependent Hartree-Fock (TDHF), the time-dependent second-order Møller-Plesset[41] (TDMP2), or, in principle, any higher level of quantum chemical theory.

OPEP is nothing else than a detailed description of how to obtain electrostatic and induction parameters from virtually any of the aforementioned electrostatic and induction energy data sets. These interaction parameters, ranging from simple sets of net atomic charges to more sophisticated models of distributed multipoles and polarizabilities, can be directly plugged into a molecular simulation package. Starting from a molecular geometry and running quantum chemical calculations, such parameters can be determined easily through various fitting procedures that make use of local, atomic frames of reference, and, whenever necessary, properly chosen sets of constraints. In order to assist the user in carrying out such a series of calculations in a seamless, possibly error-free fashion, the suite of programs developed and improved over the years has been wrapped in a Tcl/Tk graphical user interface (GUI), which allows the user to go through the necessary steps interactively. This is what the OPEP package is about.

The emphasis of this contribution has been put in a general overview of all those methods and procedures available in the OPEP suite of programs. The following section introduces the working equations, together with the necessary computational details for the construction of the grids of points on which electric properties, that is, electrostatic potentials or induction energies,

will be evaluated, the definition of local atomic frames of reference, the standard least-squares fitting procedures of distributed multipoles and polarizabilities, and, finally, the alternative, statistical approach that may be employed to determine the latter quantities, that is, statistical analysis of distributed multipoles[42] (SADM) and statistical analysis of distributed polarizabilities[43] (SADP). Illustrative case examples are provided next, before drawing conclusions and outlining future developments of OPEP.

Methods and Computational Details

mkgrid: Constructing the Grid of Points

Quantum mechanical electric properties are evaluated on a Cartesian grid of regularly spaced points in the x-, y-, and z-directions around the molecule. The number of points, N_p, forming the grid is determined by two parameters, that is, ξ, a multiplicative factor for the van der Waals radius of the participating atoms, and Δr, the separation between consecutive points. Points found to lie further than ξ times the van der Waals radius[44] from any nucleus are discarded. Points inside the so-called van der Waals envelope are also rejected. On account of the short-range effects of electron cloud penetration, which generally contaminates the "exact," multipolar part of the quantum mechanical electrostatic potential, thereby preventing its correct representation by models of distributed multipoles, the van der Waals radii defining the envelope have been doubled. In the light of a detailed analysis, it has been concluded that, insofar as the space surrounding the molecule is sampled adequately and $\xi \leq 4$, modifying the grid step does not affect the charge distribution significantly.[32,33] Increasing ξ results in exaggerating the weight of those points lying far from the nuclei, to such extent that the subtle contributions of the short-range multipole moments will be concealed almost exclusively by the leading terms of the multipole expansion, that is, by the molecular charge and/or dipole. In contrast, comparison of distinct protocols for mapping the induction energy have emphasized the necessity to sample regions of the space far enough from the nuclei to warrant an appropriate reproduction of molecular dipole polarizabilities, that is, typically $\xi = 5$. Compared to the spatial extent of the grid, characterized by ξ, the density of points dictated by Δr appears to be of lesser importance in this case. It is, therefore, clear that grids aimed at mapping either the molecular electrostatic potential or the induction energy necessarily possess distinct geometrical characteristics, namely a high density of points lying close enough to the nuclei for the former, *versus* a lesser compactness of points, the distribution of which extends sufficiently far from the molecule.

mkgrid offers the possibility to generate Cartesian grids of points in an interactive fashion. This feature will prove to be particularly useful when the number of grid points, N_p, should be limited, for example, mapping the induction energy by means of an FP approach, wherein each point k of the grid requires a separate QM calculation of the energy corresponding to the interaction of the molecule with a nonpolarizable, polarizing charge located at $\mathbf{r}_k$. Three possible choices are available, for which two of the $\{\xi, \Delta r, N_p\}$ parameters are frozen. By fixing either ξ or Δr, and N_p, mkgrid will explore iteratively the unconstrained parameter to attain the desired number of points within a tolerance threshold. Freezing ξ and Δr simultaneously yields the number of points corresponding to these values.

grid: Fitting Distributed Multipoles and Polarizabilities

The electrostatic potential created by the molecule at point $\mathbf{r}_k$ on a grid of N_p points

$$\mathcal{V}_k = \sum_A Z_A T_{00}(\mathbf{r}_A - \mathbf{r}_k) - \sum_\mu \sum_\nu P_{\mu\nu} \int \phi^*_\mu(\mathbf{r})\phi_\nu(\mathbf{r}) T_{00}(\mathbf{r} - \mathbf{r}_k)\mathrm{d}\mathbf{r} \tag{1}$$

where $\phi_\mu(\mathbf{r})$ is an atomic orbital, $P_{\mu\nu}$ the density matrix, and $T_{00}(\mathbf{r} - \mathbf{r}_k) = (1/|\mathbf{r} - \mathbf{r}_k|)$ is the Coulomb interaction function. The reference, quantum mechanical (QM) potential, $\mathcal{V}_k$, can be approximated by

$$\tilde{\mathcal{V}}_k = \sum_a \sum_{\ell=0}^{N_\ell(a)} \sum_\kappa T^{ka}_{\ell\kappa} Q^a_{\ell\kappa} \quad \forall k = 1, \ldots, N_p \tag{2}$$

where $Q^a_{\ell\kappa}$ is the $\ell\kappa$ real spherical harmonic component of a multipole situated at the point $\mathbf{r}_a$, and $T^{ka}_{\ell\kappa} = T_{\ell\kappa}(\mathbf{r}_k - \mathbf{r}_a)$ is an element of the charge-multipole interaction tensor.[45] The optimal series of distributed multipoles is obtained by minimizing the functional:

$$f(\{Q^a_{\ell\kappa}\}) = \sum_{k=1}^{N_p} (\mathcal{V}_k - \tilde{\mathcal{V}}_k)^2 \tag{3}$$

The number of distributed multipoles to be determined is given by $N_c = \Sigma_a\, 2N_\ell(a) + 1$. The solution, $\mathbf{Q} \equiv (Q^a_1, \ldots, Q^a_{N_c})$, which minimizes $f(\{Q^a_{\ell\kappa}\})$, satisfies the normal equations of the least-squares problem:

$$\mathbf{T}^\mathrm{T}\mathbf{T}\mathbf{Q} = \mathbf{T}^\mathrm{T}\mathbf{V} \tag{4}$$

Here, $\mathbf{T}$ is an $N_p \times N_c$ matrix containing the $T^{ka}_{\ell\kappa}$ elements. $\mathbf{V}$ is an N_p vector containing the $\mathcal{V}_k$ elements. Provided that the rank of $\mathbf{T}$ is equal to N_c, the matrix product $\mathbf{T}^\mathrm{T}\mathbf{T}$ can, in principle, be inverted, so that the optimal, least-squares solution is

$$\mathbf{Q} = (\mathbf{T}^\mathrm{T}\mathbf{T})^{-1}\mathbf{T}\mathbf{V} \tag{5}$$

Molecular moments at the origin, $\mathbf{r}_0$, can be regenerated subsequently by translating the distributed quantities, $Q^a_{\ell_a\kappa_a}$ to $\mathbf{r}_0$:

$$Q_{\mathscr{L}\mathscr{K}} = \sum_a \sum_\ell \sum_\kappa \mathcal{W}_{\mathscr{L}\mathscr{K},\ell\kappa}(\mathbf{r}_0 - \mathbf{r}_a) Q^a_{\ell\kappa} \tag{6}$$

$\mathcal{W}_{\mathscr{L}\mathscr{K},\ell\kappa}(\mathbf{r})$ is a translation function defined by Stone.[45]

In some cases, one may wish to constrain the total sum of partial charges to a predefined value. Although the total charge can be reproduced reasonably well by choosing a grid of good quality, small systematic errors can be avoided by imposing the charge conservation. Constraining fragment charges may be necessary, for example, in the piecewise construction of macromolecular charge density models. The constraints are implemented in grid employing the method of Lagrange multipliers. Writing a set of N_s constraint equations in a matrix form

$$\mathbf{C}^{\mathrm{T}}\mathbf{Q} = \mathbf{K} \tag{7}$$

where $\mathbf{C}$ is an $N_c \times N_s$ matrix, the constrained least-squares problem is given by:

$$\begin{pmatrix} \mathbf{T}^{\mathrm{T}}\mathbf{T} & \mathbf{C} \\ \mathbf{C}^{\mathrm{T}} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{Q} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{T}^{\mathrm{T}}\mathbf{V} \\ \mathbf{K} \end{pmatrix} \tag{8}$$

Here, λ is the vector containing the Lagrange multipliers for the N_s constraints.

In addition to the molecular multipole moments, which can provide a valuable assessment of the quality of the fitted distributed components, when compared to the QM expectation values, grid computes the root-mean square deviation (RMSD) and the mean relative deviation, δ, between the QM determined and the recovered electrostatic potentials:

$$\begin{cases} \mathrm{RMSD} = \left[\dfrac{1}{N_p}\displaystyle\sum_k (\mathcal{V}_k - \tilde{\mathcal{V}}_k)^2\right]^{1/2} \\ \delta = \dfrac{1}{N_p}\displaystyle\sum_k \left|\dfrac{\mathcal{V}_k - \tilde{\mathcal{V}}_k}{\mathcal{V}_k}\right| \end{cases} \tag{9}$$

The current version of OPEP allows one to fit distributed multipoles to the electrostatic potential up to hexadecapoles, that is, $\ell = 4$.

The induction energy, $\mathcal{U}_{\mathrm{ind},k}$, resulting from the polarization of the molecule by a nonpolarizable charge, q_k, located at point $\mathbf{r}_k$, can be expressed by

$$\mathcal{U}_{\mathrm{ind},k} = \mathcal{E}_{\mathrm{total},k} - \mathcal{E}^0 - q_k\mathcal{V}_k \tag{10}$$

Here, $\mathcal{E}_{\mathrm{total},k}$ denotes the energy of the molecule interacting with point charge q_k. $\mathcal{E}^0$ is the energy of the isolated molecule, and $\mathcal{V}_k$ is the electrostatic potential created by the latter at point $\mathbf{r}_k$. The default value of the polarizing charge adopted in OPEP to map the induction energy over a discrete set of points is equal to $+1.0$. This arbitrary choice is not expected to affect significantly the derived models of distributed polarizabilities.[15,16] In principle, the induction energy is invariant with respect to the sign of the polarizing charge, q_k. In practice, however, the electron cloud is polarized somewhat differently with $+q_k$ and $-q_k$—a consequence of the FP approach. Polarizing charges of the same magnitude, but of opposite sign, that is, ± 1.0, are, therefore, anticipated to yield slightly different results, yet of comparable quality and accuracy. Clearly, eq. (10) should be applied with particular care to avoid possible contamination resulting from nonlinear contributions to the induction energy, which is considered here as a pure linear-response effect. Of course, the role of this equation is taken over by any other appropriate expression, if the induction energy data base is not determined by the EP method. The reference induction energy, $\mathcal{U}_{\mathrm{ind},k}$, can be approximated at any given order of a distributed polarizability expansion:

$$\tilde{\mathcal{U}}_{\mathrm{ind},k} = -\frac{1}{2}\sum_{a,b}\sum_{\ell=0}^{N_\ell(a)}\sum_{\kappa}\sum_{\ell'=0}^{N_{\ell'}(b)}\sum_{\kappa'} q_k T^{ka}_{\ell\kappa}\alpha^{ab}_{\ell\kappa,\ell'\kappa'}T^{bk}_{\ell'\kappa'}q_k \qquad \forall k = 1,\ldots,N_p \tag{11}$$

where $\alpha^{ab}_{\ell\kappa,\ell'\kappa'}$ are the components of the many-body polarizability tensor α^{ab}, corresponding to sites a and b. $N_\ell(a)$ and $N_{\ell'}(b)$ stand for the orders of the multipole expansion at centers a and b, respectively. The number of distinct components of the complete array α, containing all the possible one- and two-center distributed polarizabilities, is $N_\zeta = \sum_{a\le b}[2N_\ell(a) + 1] \times [2N_\ell(b) + 1]$. It should be noted that expression (11) corresponds to the total induction energy, because the perturbing charge, q_k, is nonpolarizable. In practice, one attempts to keep the number of nonzero components, $N_c \lll N_\zeta$, as low as possible while preserving the quality of the fit. It can be argued that the sampling of the space surrounding the molecule could be somewhat enhanced by considering in eq. (11) doublets, $\{k, k'\}$, at which the polarizing charge is located.[15–17] The latter increases the data base of induction energy values to $N_p \times (N_p - 1)/2$.

Considering that the number of components of α to be determined is markedly smaller than N_p, the number of points of the grid, the overdetermined system of linear equations expressed in (11) can be conveniently restated in the following matrix form:

$$\mathbf{B}\alpha = \mathbf{U}_{\mathrm{ind}} \tag{12}$$

Here, $\mathbf{U}_{\mathrm{ind}}$ is the vector that contains the induction energies evaluated for various locations of the polarizing charge, q_k. $\mathbf{B}$ is an $N_p \times N_c$ matrix constructed using the $N_p \times N_\zeta$ electrostatic tensor, $\mathbf{T}$. The elements of $\mathbf{B}$ are defined by

$$B_{k,s} = -\frac{1}{2}q_k^2 T^{ka}_{\ell\kappa}T^{bk}_{\ell'\kappa'} \qquad \forall k, s = 1,\ldots,N_c \tag{13}$$

where $s \equiv \{ab;\ \ell\kappa;\ \ell'\kappa'\}$. The relationship $\alpha^{ab}_{\ell\kappa,\ell',\kappa'} = \alpha^{ba}_{\ell'\kappa',\ell\kappa}$ suggests that the computational effort can be reduced significantly by considering only the upper part of the matrix $\mathbf{B}$, thereby lowering the dimensionality of the least-squares problem to, at the most, $N_p \times N_\zeta\ (N_\zeta + 1)/2$.

The matrix elements corresponding to charge flow polarizabilities deserve a special attention. An important criterion for charge-flow polarizabilities relies on the physical property that there should not be any net charge induced in the molecule of interest. This pivotal property can be written as

$$\sum_a \alpha^{ab}_{00,00} = 0 \quad \text{or} \quad -\sum_{b(\neq a)} \alpha^{ab}_{00,00} = \alpha^{aa}_{00,00} \tag{14}$$

Substitution of the latter form of the sum rule in the charge-flow contribution of the induction energy

$$-\frac{1}{2} q_k^2 \sum_a \sum_b T^{ka}_{00} \alpha^{ab}_{00,00} T^{bk}_{00} = \frac{1}{2} q_k^2 \sum_a \sum_{b(\neq a)} T^{ka}_{00} \alpha^{ab}_{00,00} (T^{bk}_{00} - T^{ak}_{00}) \tag{15}$$

makes it possible to exclude the "diagonal," $\alpha^{aa}_{00,00}$ charge-flow terms from the fitting procedure and write the corresponding matrix elements as

$$B_{k,s} = -\frac{1}{2} q_k^2 T^{ka}_{00} (T^{bk}_{00} - T^{ak}_{00}) \quad s = \{ab; 00:00\}, a \neq b \tag{16}$$

In most circumstances, that through-space charge flows are sufficiently small to be ignored constitutes a reasonable assumption. This rationalizes the neglect of charge-flow polarizabilities between atoms that are not bonded chemically, as has been done in OPEP.

The system of overdetermined linear equations, (11), can be solved by minimizing the functional

$$g(\{\alpha^{ab}_{\ell\kappa,\ell'\kappa'}\}) = \sum_k (\mathcal{U}_{\text{ind},k} - \tilde{\mathcal{U}}_{\text{ind},k})^2 \tag{17}$$

The optimal solution, in a least-squares sense, satisfies the normal equations, which can be expressed as

$$\mathbf{B}^{\mathrm{T}}\mathbf{B}\alpha = \mathbf{B}^{\mathrm{T}}\mathbf{U}_{\text{ind}} \tag{18}$$

Assuming that the rank of matrix $\mathbf{B}$ is equal to N_c, the matrix product $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ can, in principle, be inverted. The optimal, least-squares solution takes the form

$$\alpha = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{U}_{\text{ind}} \tag{19}$$

Depending on the model chosen, singularities in matrix $\mathbf{B}$ may be encountered, thereby preventing the inversion of $\mathbf{B}^{\mathrm{T}}\mathbf{B}$. Of particular concern, two-center components, $\alpha^{ab}_{\ell\kappa,\ell'\kappa'}$, for which $a \neq b$ and $\ell, \ell' > 0$, have proven to be a source numerical instabilities. For this reason, two-center polarizabilities other than charge-flow polarizabilities are omitted in grid.

Just like for multipole moments, eq. (6) can be generalized to recover molecular polarizabilities from the fitted distributed components:

$$\alpha_{\mathcal{LK},\mathcal{L}'\mathcal{K}'} = \sum_{a,b} \sum_{\ell,\kappa} \sum_{\ell',\kappa'} \mathcal{W}_{\mathcal{LK},\ell\kappa}(\mathbf{r}_0 - \mathbf{r}_a) \alpha^{ab}_{\ell\kappa,\ell'\kappa'} \mathcal{W}_{\mathcal{L}'\mathcal{K}',\ell'\kappa'}(\mathbf{r}_0 - \mathbf{r}_b) \tag{20}$$

Finally, grid assesses the quality of the model of distributed polarizabilities in terms of two quantities, that is, the RMSD between the QM computed and the regenerated induction energies on the grid of N_p points, together with the corresponding mean relative deviation, δ:

$$\begin{cases} \text{RMSD} = \left[\dfrac{1}{N_p} \displaystyle\sum_k (\mathcal{U}_{\text{ind},k} - \tilde{\mathcal{U}}_{\text{ind},k})^2 \right]^{1/2} \\ \delta = \dfrac{1}{N_p} \displaystyle\sum_k \left| \dfrac{\mathcal{U}_{\text{ind},k} - \tilde{\mathcal{U}}_{\text{ind},k}}{\mathcal{U}_{\text{ind},k}} \right| \end{cases} \tag{21}$$

The current version of OPEP allows one to fit distributed polarizabilities to the induction energy up to quadrupole-quadrupole polarizabilities, that is, $\ell = \ell' = 2$.

SADM and SADP: A Statistical Approach for Fitting Distributed Multipoles and Polarizabilities

The SADM and SADP schemes constitute an alternative to the least-squares fitting procedure for determining distributed multipoles and polarizabilities, respectively. In essence, these formalisms restate the normal equations of the least-squares problem [eqs. (4, 18)]. The input data are identical to that required for the least-squares fitting procedure, that is, a grid of N_p points, over which the electrostatic potential, or the induction energy, has been mapped, and the definition of a model of distributed multipoles, or polarizabilities, consisting of N_c components. Next, N_c points are chosen randomly amongst the total N_p points of the grid, and a set of distributed multipoles, or polarizabilities, which reproduce exactly the electrostatic potential, or the induction energy, for this subset of N_c points is obtained by solving a system of N_c linear equations. In the case of distributed multipoles, the latter can be expressed as

$$\mathcal{V}_k = \sum_a \sum_{\ell=0}^{N_\ell(a)} \sum_\kappa T^{ka}_{\ell\kappa} Q^a_{\ell\kappa}, \quad \forall k = 1, \ldots, N_c \tag{22}$$

For distributed polarizabilities, the system of N_c linear equations is written as

$$\mathcal{U}_{\text{ind},k} = -\frac{1}{2} \sum_{a,b} \sum_{\ell=0}^{N_\ell(a)} \sum_\kappa \sum_{\ell'=0}^{N_{\ell'}(b)} \sum_{\kappa'} q_k T^{ak}_{\ell\kappa} \alpha^{ab}_{\ell\kappa,\ell'\kappa'} T^{bk}_{\ell'\kappa'} q_k \quad \forall k = 1, \ldots, N_c \tag{23}$$

Determination of an individual set of N_c components corresponds to *one experiment.* The value of these components inherently depends upon the subset of points taken from the grid. Consequently, the result of a single experiment is physically meaningless *per se.* By carrying out this process over a sufficiently large number, $N_e \leq C^{N_c}_{N_p}$, of distinct subsets of N_c points, the probability distribution of occurence of each distributed multipole, $\mathcal{S}(Q^a_{\ell\kappa})$, or polarizability, $\mathcal{S}(\alpha^{ab}_{\ell\kappa,\ell'\kappa'})$, can be constructed. The optimal set of

distributed multipoles, or polarizabilities, is found by extracting the maximum value of the $\mathcal{S}(Q^a_{\ell\kappa})$, or $\mathcal{S}(\alpha^{ab}_{\ell\kappa,\ell'\kappa'})$, function, which corresponds to the most probable value of a given component.

Furthermore, the probability distribution, $\mathcal{S}(Q^a_{\ell\kappa})$, or $\mathcal{S}(\alpha^{ab}_{\ell\kappa,\ell'\kappa'})$, constitutes a valuable source of information, not only about the value of the multipolar component itself, $Q^a_{\ell\kappa}$ or $\alpha^{ab}_{\ell\kappa,\ell'\kappa'}$, but also about the accuracy of its determination. Examination of the probability distributions allows one to assess in a very efficient manner the suitability and the reliability of a particular model, and pinpoint ill-defined components. In the case of distributed multipoles, pathological models generally correspond to wide, poorly resolved distributions, characteristic, for instance, of buried atoms, the contribution to the electrostatic potential of which is marginal.[42] For distributed polarizabilities, a pathological model can be synonymous with redundant components.[43]

locfrm: Constructing Local Frames of Reference

Fitting multipole and polarizability parameters to *families* of atoms instead of individual atoms opens the possibility of tackling two important issues, namely (1) the molecular symmetry, which imposes exact transformations between tensor properties of symmetry related atoms, and (2) transferability, which exploits the approximate equivalence of the tensor properties shared by those atoms that have similar chemical environments. While the transformation of zeroth-order properties, that is, atomic charge and charge flow polarizabilities, is trivial, equivalence of higher order properties cannot be appreciated without transforming them into a common frame of reference. The most natural choice for such a common frame of reference for a family of exactly—that is, symmetry—or approximately—that is, transferability—equivalent atoms is based upon an analysis of the local environment, as outlined below.

Before detailing the algorithm implemented in OPEP for identifying the axes of local frames of reference attached to each atom, their use for fitting distributed multipoles or polarizabilities over families rather than individual atoms should first be rationalized. This can be done most conveniently by adopting the formalism of charge-multipole interaction tensors, $T^{ka}_{00,\ell\kappa}(\Omega_a)$, that depend intrinsically upon the orientation, Ω_a, of the local frame of reference in which the multipole moments, $Q^a_{\ell\kappa}$, are described.[45] Explicit expressions,[46] as well as general recursion formulae,[47] are available to calculate such orientation-dependent interaction tensors.

When fitting multipoles to the electrostatic potential, eq. (2) can be rewritten as

$$\tilde{\mathcal{V}}_k = \sum_{a=1}^{N_a} \sum_{\ell}^{N_\ell(a)} \sum_{\kappa} T^{ka}_{\ell\kappa}(\Omega_a) Q^a_{\ell\kappa} = \sum_{A=1}^{N_f} \sum_{\ell}^{N_\ell(A)} \sum_{\kappa} Q^A_{\ell\kappa} \sum_{a\in A}^{N_a(A)} T^{ka}_{\ell\kappa}(\Omega_a) \quad \forall k = 1, \ldots, N_p \tag{24}$$

where N_f denotes the number of families of symmetry-related atoms and $N_a(A)$ the number of atoms belonging to family A.

In an analogous fashion, when fitting polarizabilities to the induction energy, eq. (11) can be explicited using the formalism of the local frames of reference:

$$\tilde{\mathcal{U}}_{\mathrm{ind},k} = -\frac{1}{2} q_k^2 \sum_a \sum_{\ell} \sum_{\kappa} \sum_b \sum_{\ell'} \sum_{\kappa'} T^{ka}_{\ell\kappa}(\Omega_a) \alpha^{ab}_{\ell\kappa,\ell'\kappa'} T^{bk}_{\ell'\kappa'}(\Omega_b) \tag{25}$$

After applying the restrictions discussed previously, we are left with two-center charge-flow polarizabilities connecting chemically bonded atoms, together with one-center multipole-multipole terms, $\alpha^{aa}_{\ell\kappa,\ell'\kappa'}$:

$$\tilde{\mathcal{U}}_{\mathrm{ind},k} = -\frac{1}{2} q_k^2 \sum_a \sum_b \beta(a,b) T^{ka}_{00} \alpha^{ab}_{00} T^{bk}_{00} - \frac{1}{2} q_k^2 \sum_a \sum_{\ell=1}^{N_\ell(a)} \sum_{\kappa} \sum_{\ell'=1}^{N_\ell(a)} \sum_{\kappa'} T^{ka}_{00,\ell\kappa}(\Omega_a) \alpha^{aa}_{\ell\kappa,\ell'\kappa'} T^{ak}_{\ell'\kappa',00}(\Omega_a) \tag{26}$$

where function $\beta(a, b)$ is 1 for bonded pairs of atoms and 0 otherwise.

Using the equivalence relationship between atoms and the connectivity matrix of the molecule, one can establish a list of unique bond families, $\{AB\}$, each consisting of several bonds, $\{ab\}$, and separate the summation over bond families. The one-center multipole polarizability parameters can be handled in a similar manner as in the case of multipoles. The expression for the model induction energy then becomes

$$\tilde{\mathcal{U}}_{\mathrm{ind},k} = -\frac{1}{2} q_k^2 \sum_{\{AB\}} \alpha^{AB}_{00,00} \sum_{\{ab\}\in\{AB\}} T^{ka}_{00}(T^{bk}_{00} - T^{ak}_{00}) - \frac{1}{2} q_k^2 \sum_A \sum_{\ell\ell'}^{N_\ell(a)} \sum_{\kappa\kappa'} \alpha^{AA}_{\ell\kappa,\ell'\kappa'} \sum_{a\in A} T^{ka}_{\ell\kappa}(\Omega_a) T^{ak}_{\ell'\kappa'}(\Omega_a) \tag{27}$$

The condensed form of the system of equations to be solved can then be written as

$$\tilde{\mathcal{U}}_{\mathrm{ind},k} = \sum_r B^{kr} \alpha^r \quad \forall k = 1, \ldots, N_p \tag{28}$$

where index r is either $r = \{AB\}$, in the case of charge-flow polarizabilities, or $r = \{A, \ell\kappa, \ell'\kappa'\}$. The coefficient matrix elements, B^{kr}, can be expressed as

$$B^{k,\{AB\}} = -\frac{1}{2} q_k^2 \sum_{a-b\in A-B} T^{ka}_{00}(T^{bk}_{00} - T^{ak}_{00}) q_k^2$$

$$B^{k,\{A\ell\kappa\ell'\kappa'\}} = -\frac{1}{2} q_k^2 \sum_{a\in A} T^{ka}_{\ell\kappa}(\Omega_a) T^{ak}_{\ell'\kappa'}(\Omega_a) q_k^2 \tag{29}$$

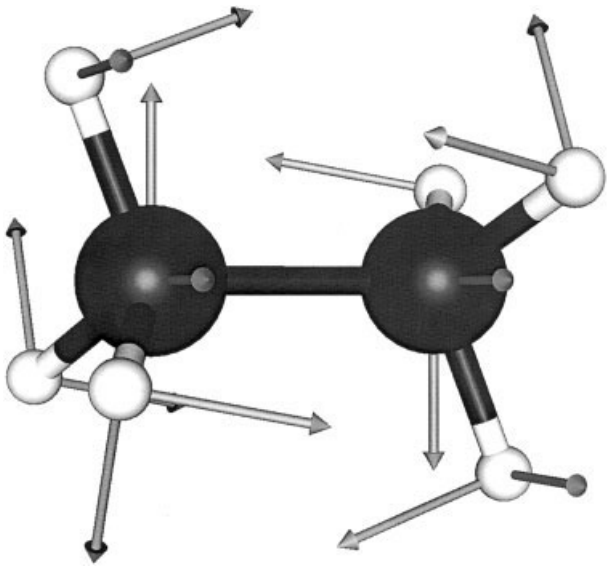


Figure 1. XMakemol representation of the ethane molecule, featuring the frames of reference on the carbon and the hydrogen atoms. The z-axes for the latter are oriented along the C—H bond, pointing towards the carbon atom. Orientation of the x-axes follows the $\mathscr{C}_3$ symmetry of the molecule.

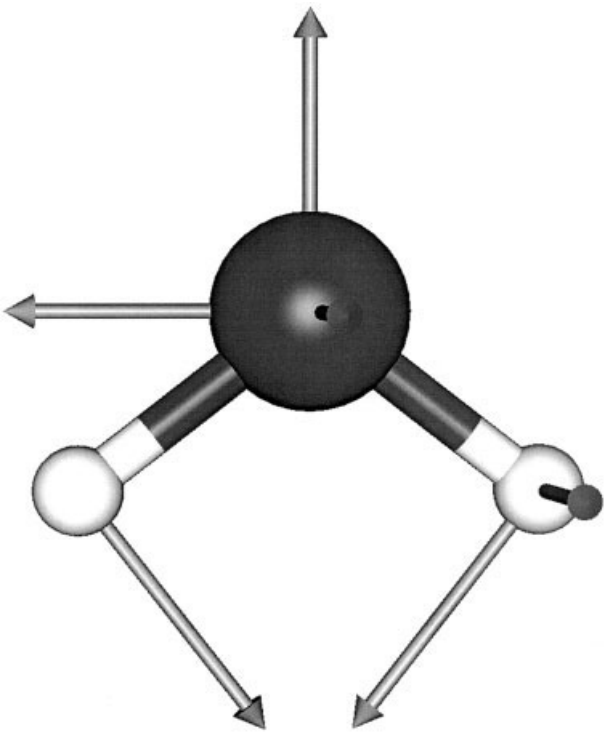


Figure 2. XMakemol representation of the water molecule, featuring the frames of reference on the oxygen and the hydrogen atoms. The z-axes for the latter, that is, z_{H_1} and z_{H_2}, are oriented along the O—H bond, pointing towards the oxygen atom.

In the OPEP suite of programs, the local frames of reference, right-handed, orthonormal sets of reference axes, are computed by locfrm. This reference frame can be described in terms of four well-delineated rules:

1. The first reference axis—x, y, or z—is defined by selecting one, two, or three neighbors. The direction of the axis is given, respectively, by the vector pointing towards the neighbor, the bisector between the two neighbors, or the vector resulting from the sum of the unit vectors pointing towards the three neighbors.
2. The sign of this axis is defined by an additional entry, as $+1$ or -1.
3. An identical syntax is adopted for the definition of the second axis, which will be orthogonalized to the first one.
4. The third axis is obtained as an ordered cross-product of the two other axes.

By default, locfrm generates the local frames of reference in an automated fashion. A list of connectivity is built by comparing the distance separating two atoms in the molecule with the sum of their covalent radii. Next, the number of nearest neighbors of a given atom a is sought by applying the following rules:

1. If the number of first neighbors is 1, that is, atom b, the molecular graph is explored by searching amongst the neighbors of b for those that form an angle different than π with $\mathbf{r}_{ab}$. The z-axis is chosen to point towards b, and the x-axis is the orthogonal projection of the vector pointing towards the immediate neighbor of b.

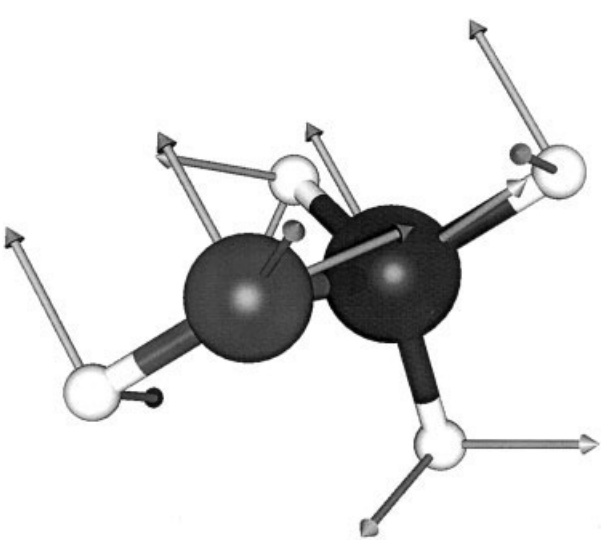


Figure 3. XMakemol representation of the methanol molecule, featuring the frames of reference on the carbon, the oxygen, and the hydrogen atoms. The z-axes for the aliphatic hydrogen atoms are oriented along the C—$H_{C,i}$ bonds, pointing towards the carbon atom. The frames of reference for these atoms reflect the local $\mathscr{C}_3$ symmetry of the methyl group.

Table 1. Models of Distributed Multipoles and Regenerated Molecular Multipole Moments of Ethane and Water, at the MP2/Sadlej//MP2/6-311++G(2d, 2p) Level of Approximation.

		Distributed multipoles		Molecular multipoles		MP2/Sadlej
Ethane	Model a	Q_{00}^{C}	0.4598	Q_{20}	−0.6792	−0.6398
		$Q_{00}^{H_i}$	−0.1533	Q_{40}	−17.0838	−15.4833
		Q_{10}^{C}	0.1149	Q_{43s}	−31.1861	−26.1049
		$Q_{10}^{H_i}$	−0.1937			
		Q_{20}^{C}	0.3666			
		RMSD (10^{-3} a.u.)	0.038			
		$\Delta\varepsilon$ (%)	24.741			
	Model b	Q_{00}^{C}	−0.3651	Q_{20}	−0.75590	−0.6398
		$Q_{00}^{H_i}$	0.1217	Q_{40}	−27.3720	−15.4833
		Q_{10}^{C}	0.2383	Q_{43s}	−23.9359	−26.1049
		RMSD (10^{-3} a.u.)	0.089			
		$\Delta\varepsilon$ (%)	39.276			
	Model c	Q_{00}^{C}	0.0881	Q_{20}	−0.1499	−0.6398
		$Q_{00}^{H_i}$	−0.0294	Q_{40}	5.2184	−15.4833
				Q_{43s}	5.7767	−26.1049
		RMSD (10^{-3} a.u.)	0.472			
		$\Delta\varepsilon$ (%)	139.152			
Water	Model a	Q_{00}^{O}	−0.6000	Q_{10}	−0.7323	−0.7321
		$Q_{00}^{H_i}$	0.3000	Q_{20}	−0.2318	−0.2305
		Q_{10}^{O}	0.0246	Q_{22c}	2.1727	2.1790
		$Q_{10}^{H_i}$	−0.0877			
		$Q_{11c}^{H_i}$	−0.0111			
		Q_{20}^{O}	−0.0174			
		Q_{22c}^{O}	0.7348			
		RMSD (10^{-3} a.u.)	0.026			
		$\Delta\varepsilon$ (%)	0.656			
	Model b	Q_{00}^{O}	−1.1899	Q_{10}	−0.7339	−0.7321
		$Q_{00}^{H_i}$	0.5950	Q_{20}	−0.0720	−0.2305
		Q_{10}^{O}	0.5886	Q_{22c}	2.1053	2.1790
		RMSD (10^{-3} a.u.)	0.325			
		$\Delta\varepsilon$ (%)	13.064			
	Model c	Q_{00}^{O}	−0.6724	Q_{10}	−0.7474	−0.7321
		$Q_{00}^{H_i}$	0.3362	Q_{20}	−0.1885	−0.2305
				Q_{22c}	1.1898	2.1790
		RMSD (10^{-3} a.u.)	1.308			
		$\Delta\varepsilon$ (%)	54.273			

2. If the number of first neighbors is 2, that is, atoms b and c, the valence angle formed by these two atoms is evaluated. Should the latter be different than π, the z-axis is defined as the bisector between b and c, and the x-axis is assigned to one of the neighbors. Alternatively, the x-axis is assigned to the first neighbor of b and c, and the z-axis to either b or c.
3. If the number of first neighbors is 3, that is, atoms b, c, and d, the three valence angles formed by these atoms are computed. Should the sum of the angles be equal to 2π, the z-axis is defined by the central atom and atom $s = b$, c, or d, and the x-axis by the central atom, a, and atom $s' = b$, c, or d, so that $s' \neq s$. Alternatively, the z-axis is defined as the sum of the three vectors connecting the central atom to b, c, or d, and the x-axis is assigned to one of the neighbors.
4. If the number of first neighbors is 4, the sum of the valence angles formed by the four neighbors and the central atom, a, is evaluated. Should this sum be equal to 2π, the z-axis is defined as the bisector of two vectors amongst the four, and the x-axis is assigned to one of the neighbors. Alternatively, the z-axis is defined as the sum of the three vectors pointing from central atom a towards three of the four neighbors, and the x-axis is assigned to the fourth neighbor.

Applications and Discussion

To illustrate the performance of OPEP, distributed multipole and polarizability analyses of the prototypical ethane, water, and methanol molecules were carried out—see Figures 1–3. Preliminary investigations on these model systems suggested that the choice of

a large enough basis set is pivotal to warrant an appropriate agreement with the experimentally determined molecular moments and polarizabilities. In this contribution, all calculations were carried out at the MP2 level of theory, using the basis set developed by Sadlej.[48] The latter is targeted at the description of electric properties, and has proven to constitute an optimal choice in terms of accuracy and number of basis functions. The electrostatic potential created *in vacuo* by ethane and water was computed on a grid of 892 and 2106 points, respectively, employing the GAUSSIAN98 suite of programs—here $\xi = 3$. Interestingly enough, the sole information required by OPEP for such calculations is the geometry of the molecule in an xyz format. Here, the lowest energy structure for the two molecules was obtained at the MP2/6-311++G($2d$, $2p$) level of approximation. Once the grid of points is constructed, OPEP automatically generates the relevant input file for GAUSSIAN98, thus simplifying the execution of the latter. The Tcl/Tk GUI makes the selection of the distributed multipole components to be fitted particularly easy. Molecular symmetry, which appears in eq. (24), can be turned on by updating the list of families of symmetry-related atoms. By default, each atom of the molecule belongs to a distinct family. As can be seen in Table 1, models of distributed multipoles consisting of point charges borne by all atoms, supplemented by atomic dipoles on heavy atoms only (compare model b), constitute an interesting compromise between complexity and accuracy.

The strength of the statistical approach, either SADM or SADP, lies in the possibility of detecting pathological models by a simple inspection of the normalized distributions, which characterize the constituent components—for example $\mathcal{S}(Q^a_{\ell\kappa})$ in the case of a model of distributed multipoles, $Q^a_{\ell\kappa}$. The conspicuous and chemically counter-intuitive $C^{\delta-}$—$X^{\delta+}$ (X = F or Cl) bond polarity observed in the simple point charge model of difluorodichloromethane[49] corresponds to a poorly resolved probability distribution for the buried carbon atom, as shown in Figure 4—details of the quantum chemical calculations are given in ref. 42. A closer examination at $\mathcal{S}(Q_{00_C})$ brings to light a secondary peak, a shoulder in reality, around $Q^C_{00} = 0$, often symptomatic of an ill-defined fitted model. This example further illustrates that, under certain circumstances, models limited to net atomic charges can be mathematically consistent while not physically so. Circumventing this shortcoming by assigning *a priori* a chemically meaningful point charge on buried atoms constitutes the philosophy of the well-behaved electrostatic potential based method for deriving net atomic charges by means of hyperbolic restraints (RESP).[50] This solution, although tempting on account of its simplicity, allowing the user to remain in the framework of a monopole description of the molecular electrostatic potential, may not always be desirable, especially if an accurate reproduction of the latter is sought. In particular, for those atoms bearing large local moments of higher order than Q^a_{00}, hoping to attain a precise and faithful representation of the electrostatic potentials using simple point charges is illusory. In fact, enforcing Q^C_{00} to be positive, so that the differences of electronegativities are respected, increases the RMSD between the QM calculated and the point charge derived electrostatic potentials from 0.692×10^{-3} a.u. for the unrestrained fit, to 1.280×10^{-3} a.u., when $Q^C_{00} = 0.5$. In sharp contrast, as indicated in Figure 4, incorporation of atomic dipoles offers a more satisfactory solution to the problem, reverting the bond polarities to the expected $C^{\delta+}$—$X^{\delta-}$, while decreasing the RMSD to 0.092×10^{-3} a.u.

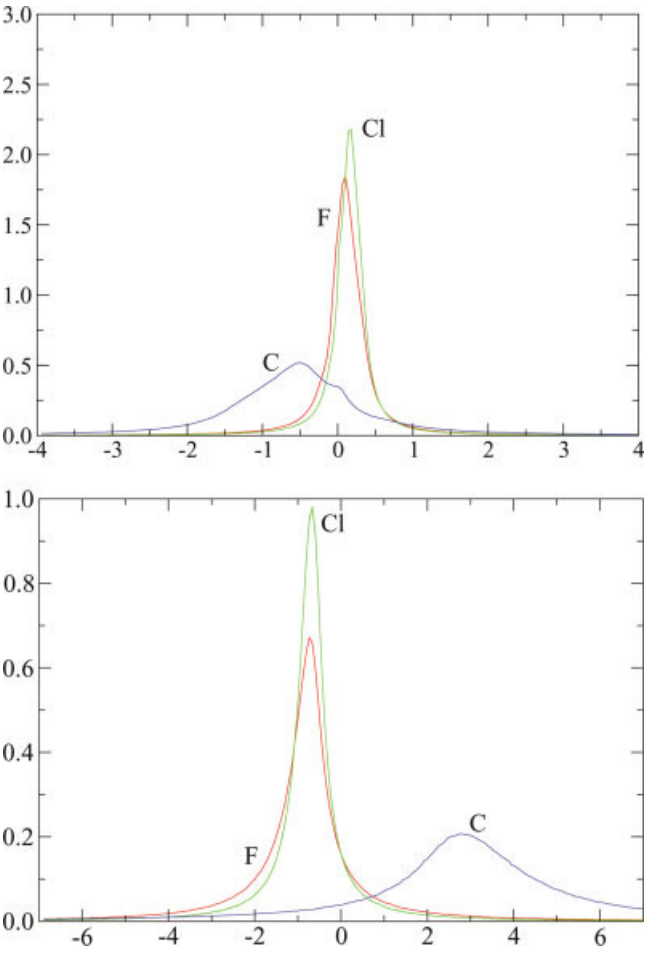


Figure 4. sadm probability distributions of the point charges, Q^a_{00}, born by difluorodichloromethane in a model limited to simple net atomic charges (above), and in a model that also includes atomic dipoles, Q^a_{1m}, (below). The counter-intuitive $C^{\delta-}$—$X^{\delta+}$ (X = F or Cl) bond polarity in the rudimentary point charge description is inverted when atomic dipoles are incorporated in the model.

Determination of distributed polarizabilities in OPEP relies on the knowledge of the induction energy mapped on a grid of points around the molecule of interest. Just like for the electrostatic potential, this grid can be defined using the mkgrid algorithm, keeping in mind that the spatial extent of the grid, characterized by ξ, and the density of points, controlled by Δr, should reflect the necessity to sample regions far enough from the nuclei. Once the grid of points is built, OPEP offers the possibility to generate a script appropriate for computing FP induction energies. This script drives the evaluation of: the interaction energy, $\mathcal{E}_{\text{total},k}$, for each point, k, of the grid, where the polarizing charge is located; the energy, $\mathcal{E}^0$, of the isolated molecule; and the electrostatic potential, $\mathcal{V}_k$, created by the latter. This approach, which requires N_p distinct

Table 2. Models of Distributed Polarizabilities and Regenerated Molecular Polarizabilities of Methanol and Water, at the MP2/Sadlej//MP2/6-311++G(2d, 2p) Level of Approximation.

		Distributed polarizabilities		Molecular polarizabilities		MP2/Sadlej
Methanol	Model a	$\alpha^{CC}_{10,10}$	13.642	$\alpha_{10,10}$	20.961	19.770
		$\alpha^{CC}_{10,11c}$	0.139	$\alpha_{11c,11c}$	23.655	20.250
		$\alpha^{CC}_{11c,11c}$	21.191	$\alpha_{11c,11s}$	0.969	0.590
		$\alpha^{CC}_{11s,11s}$	11.590	$\alpha_{11s,11s}$	25.960	23.080
		$\alpha^{OO}_{10,10}$	7.794			
		$\alpha^{OO}_{10,11c}$	−6.383			
		$\alpha^{OO}_{11c,11c}$	7.466			
		$\alpha^{OO}_{11s,11s}$	7.051			
		$\alpha^{H_OH_O}_{10,10}$	2.555			
		$\alpha^{H_OH_O}_{10,11c}$	1.977			
		$\alpha^{H_OH_O}_{11c,11c}$	−3.032			
		$\alpha^{H_OH_O}_{11s,11s}$	2.320			
		RMSD (10^{-3} a.u.)	0.063			
		$\Delta\varepsilon$ (%)	8.636			
	Model b	$\alpha^{CO}_{00,00}$	−1.333	$\alpha_{10,10}$	21.298	19.770
		$\alpha^{CH_{C,i}}_{00,00}$	−1.174	$\alpha_{11c,11c}$	23.592	20.250
		$\alpha^{OH_O}_{00,00}$	−0.819	$\alpha_{11c,11s}$	0.487	0.590
		$\alpha^{CC}_{1m,1m}$	7.989	$\alpha_{11s,11s}$	26.423	23.080
		$\alpha^{OO}_{1m,1m}$	6.705			
		RMSD (10^{-3} a.u.)	0.067			
		$\Delta\varepsilon$ (%)	7.982			
	Model c	$\alpha^{CO}_{00,00}$	−3.676	$\alpha_{10,10}$	17.292	19.770
		$\alpha^{CH_{C,i}}_{00,00}$	−3.075	$\alpha_{11c,11c}$	23.635	20.250
		$\alpha^{OH_O}_{00,00}$	−2.257	$\alpha_{11c,11s}$	1.385	0.590
				$\alpha_{11s,11s}$	32.091	23.080
		RMSD (10^{-3} a.u.)	0.162			
		$\Delta\varepsilon$ (%)	14.363			
Water	Model a	$\alpha^{OO}_{10,10}$	7.561	$\alpha_{10,10}$	9.805	9.751
		$\alpha^{OO}_{10,11c}$	−0.056	$\alpha_{11c,11c}$	10.073	10.063
		$\alpha^{OO}_{11c,11c}$	7.177	$\alpha_{11s,11s}$	9.483	9.542
		$\alpha^{OO}_{11s,11s}$	8.532			
		$\alpha^{OO}_{20,20}$	−0.997			
		$\alpha^{OO}_{20,21c}$	1.667			
		$\alpha^{OO}_{20,22c}$	5.086			
		$\alpha^{OO}_{21c,21c}$	−31.659			
		$\alpha^{OO}_{21c,22c}$	2.982			
		$\alpha^{OO}_{21s,21s}$	2.817			
		$\alpha^{OO}_{21s,22s}$	2.751			
		$\alpha^{OO}_{22c,22c}$	−9.061			
		$\alpha^{OO}_{22s,22s}$	−1.885			
		$\alpha^{HH}_{10,10}$	2.215			
		$\alpha^{HH}_{10,11c}$	−0.068			
		$\alpha^{HH}_{11c,11c}$	0.355			
		$\alpha^{HH}_{11s,11s}$	0.475			
		RMSD (10^{-3} a.u.)	0.012			
		$\Delta\varepsilon$ (%)	0.584			
	Model b	$\alpha^{OH_i}_{00,00}$	−1.604	$\alpha_{10,10}$	9.906	9.751
		$\alpha^{OO}_{10,10}$	5.945	$\alpha_{11c,11c}$	10.223	10.063
		$\alpha^{OO}_{11c,11c}$	3.671	$\alpha_{11s,11s}$	9.367	9.542
		$\alpha^{OO}_{11s,11s}$	9.367			
		RMSD (10^{-3} a.u.)	0.032			
		$\Delta\varepsilon$ (%)	1.501			
	Model c	$\alpha^{OH_i}_{00,00}$	−3.256	$\alpha_{10,10}$	8.044	9.751
				$\alpha_{11c,11c}$	13.305	10.063
				$\alpha_{11s,11s}$	0.000	0.9542
		RMSD (10^{-3} a.u.)	0.622			
		$\Delta\varepsilon$ (%)	37.311			

QM calculations of $\mathscr{E}_{\text{total},k}$ to map the induction energy on a grid of N_p points, is CPU-intensive and costly. Here, the induction energies were determined employing the FP[38] approach for methanol, and from the partitioning of the response charge density[51] (TPEP) in the case of water. The choice of these two approaches is, however, not restrictive, and alternative methods, among which that based on second-order perturbation theory,[39,52] can be adopted. Grids of 3905 and 1175 points were used for water and methanol, respectively—here $\xi = 5$.

As can be observed in Table 2, a model of methanol consisting of charge-flow polarizabilities supplemented by isotropic dipole polarizabilities on heavy atoms offers a reasonable compromise between complexity and accuracy. Clearly, a simple model of water limited to charge-flow polarizabilities not too unexpectedly cannot reproduce the out-of-plane component of the molecular dipole polarizability, exaggerating the $\alpha^{\text{OO}}_{10,10}$ and the $\alpha^{\text{OO}}_{11c,11c}$ contributions. Inclusion of an anisotropic local dipole polarizability on the oxygen atom reduces dramatically the RMSD between the QM calculated and the distributed polarizability derived induction energies—albeit an isotropic dipole polarizability, which yields an RMSD of 0.127×10^{-3} a.u., constitutes a satisfactory approximation, considering the symmetry of water.

Conclusions and Perspectives

The OPEP suite of FORTRAN programs, featuring a Tcl/Tk GUI, constitutes a convenient and userfriendly toolkit for developing models of distributed multipoles and polarizabilities that can be subsequently plugged in molecular modeling softwares. Design of an optimal model, a compromise between accuracy and compactness, is greatly facilitated by the use of two powerful, alternative fitting procedures, which allow the user to apply judiciously chosen sets of constraints and exploit both symmetry and transferability properties of the molecules by means of local atomic frames of reference. To our knowledge, OPEP is the first free, open-source code of its kind that offers a wide range of algorithms gathered in a userfriendly environment. What distinguishes OPEP from other similar programs lies in its ability to derive distributed polarizabilities following a nonempirical scheme, based on a reliable determination of the quantum chemical induction energy mapped on a grid of points.

For the most part, this first version of OPEP is self-contained, and includes all the necessary tools for building models of distributed multipoles and polarizabilities—that is, from the construction of a grid of points to the fitting of quantum chemically calculated electric properties. There is still room, however, for further enhancements and inclusion of novel, useful features. On-going developments of this package essentially focus on the procedures to generate grids of induction energies, as an alternative to the costly EP method, already available in OPEP. Upcoming versions will incorporate the determination of TPEP models and the generation of the corresponding grids of induction energies. Extension to the derivation of dispersion interaction models, through the computation of distributed dynamic polarizabilities, also constitutes another promising feature of OPEP. In addition, in the context of force field development, the current fitting procedures will be generalized to multiple molecules and conformations, thereby allowing transferable, conformationally averaged parameters to be derived.

Acknowledgments

The authors are indebted to Matt Hodges for having facilitated the inclusion in OPEP of his visualization program, XMakemol, and for having incorporated in the latter the possibility of visualizing the local frames of reference generated by OPEP. The Centre Charles Hermite (CCH), Vandœuvre-lès-Nancy, France, and the CINES, Montpellier, France, are gratefully acknowledged for provision of generous amounts of CPU time on their SGI Origin 2000 and Origin 3800.

References

1. Cox, S. R.; Williams, D. E. J Comput Chem 1981, 2, 304.
2. Bachrach, S. M. In Reviews in Computational Chemistry, Vol. 5; Lipkowitz, K., Boyd, D. B., Eds.; VCH: New York, 1994; p 171.
3. Cornell, W. D.; Chipot, C. In Encyclopedia of Computational Chemistry, Vol. 1; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollmann, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; Wiley and Sons: Chichester, 1998; pp 258–263.
4. Sipe, J. E.; van Kranendonk, J. Mol Phys 1978, 35, 1579.
5. Hunt, K. L. J Chem Phys 1983, 78, 6149.
6. Stone, A. J. Chem Phys Lett 1989, 155, 102.
7. Ahlström, P.; Wallquist, A.; Engströms, S.; Jönsson, B. Mol Phys 1989, 68, 563.
8. Dang, L. X.; Rice, J. E.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1991, 113, 2481.
9. Bernardo, D. N.; Ding, Y.; Krogh-Jesperesen, K.; Levy, R. M. J Phys Chem 1994, 98, 4180.
10. Soetens, J. C.; Millot, C. Chem Phys Lett 1995, 235, 22.
11. Birge, R. R. J Chem Phys 1980, 72, 5312.
12. Thole, B. T. J Chem Phys 1981, 59, 341.
13. Voisin, C.; Cartier, C. J Mol Struct Theochem 1993, 286, 35.
14. Applequist, J. Acc Chem Res 1977, 10, 79.
15. Nakagawa, S.; Kosugi, N. Chem Phys Lett 1993, 210, 180.
16. Nakagawa, S. Chem Phys Lett 1997, 278, 272.
17. Nakagawa, S. Chem Phys Lett 1995, 246, 256.
18. Alkorta, I.; Bachs, M.; Perez, J. J. Chem Phys Lett 1994, 224, 160.
19. Claverie, P. In Intermolecular Interactions: From Diatomics to Biopolymers; Pullman, B., Ed.; Wiley–Interscience: New York, 1978; p 69.
20. Stone, A. J.; Alderton, M. Mol Phys 1985, 56, 1047.
21. Stone, A. J. Mol Phys 1985, 56, 1065.
22. Le Sueur, C. R.; Stone, A. J. Mol Phys 1993, 78, 1267.
23. Le Sueur, C. R.; Stone, A. J. Mol Phys 1994, 83, 293.
24. Bader, R. F. W. Atoms in Moleucles—A Quantum Theory; Oxford University Press: London, 1990.
25. Ángyán, J. G.; Jansen, G.; Loos, M.; Hättig, C.; Heß, B. A. Chem Phys Lett 1994, 219, 267.
26. Hättig, C.; Jansen, G.; Heß, B. A.; Ángyán, J. G. Can J Chem 1996, 74, 976.
27. Jansen, G.; Hättig, C.; Heß, B. A.; Ángyán, J. G. Mol Phys 1996, 88, 69.

28. Hättig, C.; Jansen, G.; Heß, B. A.; Ángyán, J. G. Mol Phys 1997, 91, 145.
29. Stone, A. J.; Hättig, C.; Jansen, G.; Ángyán, J. G. Mol Phys 1996, 89, 595.
30. Singh, U.; Kollman, P. J Comput Chem 1984, 5, 129.
31. Chirlian, L. E.; Francl, M. M. J Comput Chem 1987, 8, 894.
32. Chipot, C.; Maigret, B.; Rivail, J. L.; Scheraga, H. A. J Phys Chem 1992, 96, 10276.
33. Ángyán, J. G.; Chipot, C. Int J Quantum Chem 1994, 52, 17.
34. Francl, M. M.; Chirlian, L. E. Rev Comput Chem 2000, 14, 1.
35. Vigné-Maeder, F.; Claverie, P. J Chem Phys 1988, 88, 4934.
36. Stone, A. J. Chem Phys Lett 1981, 83, 233.
37. Sokalski, W. A.; Poirier, R. A. Chem Phys Lett 1983, 98, 86.
38. Celebi, N.; Ángyán, J. G.; Dehez, F.; Millot, C.; Chipot, C. J Chem Phys 2000, 112, 2709.
39. Chipot, C.; Luque, F. J. Chem Phys Lett 2000, 332, 190.
40. Jansen, G. Theor Chem Acc 2000, 104, 499.
41. Hättig, C.; Heß, B. A. Chem Phys Lett 1995, 233, 359.
42. Chipot, C.; Ángyán, J. G.; Millot, C. Mol Phys 1998, 94, 881.
43. Dehez, F.; Soetens, J. C.; Chipot, C.; Ángyán, J. G.; Millot, C. J Phys Chem A 2000, 104, 1293.
44. Bondi, A. J Phys Chem 1964, 68, 441.
45. Stone, A. J. In Theoretical Models of Chemical Bonding, Vol. 4; Maksić, H., Ed.; Springer-Verlag: Berlin, 1991; pp 103–131.
46. Hättig, C.; Heß, B. A. Mol Phys 1994, 81, 813.
47. Hättig, C. Chem Phys Lett 1996, 260, 341.
48. Sadlej, A. J. Collec Czech Chem Commun 1988, 53, 1995.
49. Chipot, C.; Wilson, M. A.; Pohorille, A. J Phys Chem B 1997, 101, 782.
50. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. J Phys Chem 1993, 97, 10269.
51. Dehez, F.; Chipot, C.; Millot, C.; Ángyán, J. G. Chem Phys Lett 2001, 338, 180.
52. Chipot, C.; Dehez, F.; Ángyán, J. G.; Millot, C.; Orozco, M.; Luque, F. J. J Phys Chem A 2001, 105, 11505.