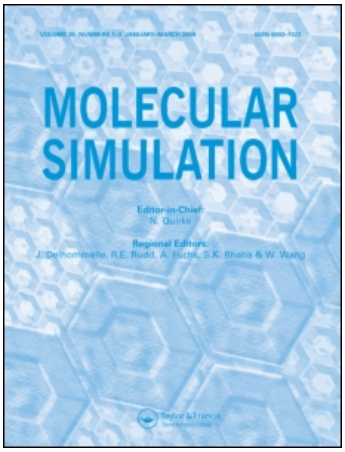


Molecular Simulation

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Optimal Site Charge Models for Molecular Electrostatic Potentials

Panagiotis G. Karamertzanis[a]; Constantinos C. Pantelides[a]
[a] Department of Chemical Engineering and Chemical Technology, Centre for Process Systems Engineering, Imperial College London, London, United Kingdom

To cite this Article Karamertzanis, Panagiotis G. and Pantelides, Constantinos C.(2004) 'Optimal Site Charge Models for Molecular Electrostatic Potentials', Molecular Simulation, 30: 7, 413 — 436
To link to this Article: DOI: 10.1080/08927020410001680769
URL: http://dx.doi.org/10.1080/08927020410001680769

Molecular Simulation, Vol. 30 (7), 15 June 2004, pp. 413–436

Taylor & Francis
Taylor & Francis Group

Optimal Site Charge Models for Molecular Electrostatic Potentials

PANAGIOTIS G. KARAMERTZANIS and CONSTANTINOS C. PANTELIDES*

Department of Chemical Engineering and Chemical Technology, Centre for Process Systems Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

(Received July 2003; In final form February 2004)

This paper presents a comprehensive methodology for the fitting of site charge models to molecular electrostatic fields computed using quantum mechanical calculations. Charges may be placed both at atomic positions and at "satellite" positions associated with selected atoms. The positions of these satellites is also optimised. Molecular symmetry is taken into account for the determination of physically meaningful charge magnitudes and satellite positions. Maximum tolerances may be imposed on the deviations of the total charge, dipole moment and quadrupole computed by the site charge model from the corresponding quantum mechanical values.

The fitting is achieved via the solution of a nonlinear least squares problem that incorporates specified upper bounds on the statistical error of the charge magnitudes; this automatically excludes satellite charge positions that lead to ill-conditioned and inaccurate charge estimates. The overall optimisation problem is non-convex, involving multiple local minima. Molecular symmetry is exploited to derive an exact reformulation of the original problem that involves a reduced number of optimisation decision variables. A multi-level single-linkage (MLSL) stochastic minimization algorithm, coupled with a quadratic programming local minimisation algorithm and analytical derivatives for both objective function and constraints, is used for the reliable determination of globally optimal solutions of the reformulated problem.

A number of examples illustrating the applicability of the methodology are presented. The use of the optimal site charge models in the area of *ab initio* crystal structure prediction is briefly discussed.

Keywords: Site charges; Electrostatic potentials; Global optimization; Confidence intervals; Cyclobutane

INTRODUCTION

Site charge models are employed for the description of the electrostatic part of the energy in most of the force fields that are used in practical applications. Unfortunately, site charges cannot be measured experimentally or computed directly through quantum mechanical calculations. A common approach is to estimate them through a least-squares fit to the electrostatic potential or the electrostatic field, subject to constraints on the total molecular charge or more generally, the first few multipole moments. The electrostatic potential field can be computed through *ab initio* or semiempirical calculations, or from experimental charge density analysis of X-ray diffraction data [1,2].

Since the first attempts for the construction of site charge models [3,4] based on quantum mechanically observable quantities, the majority of researchers have considered only charges located at atomic positions. Much attention has been directed to the efficient sampling of the electrostatic potential derived from the quantum mechanical calculations. Sampling points should not be too close to the nuclei as the electrostatic potential is not well defined by a point charge model inside the van der Waals radius; for this reason, points that are closer than 1.0–2.5 times the van der Waals radius from any atom are excluded from consideration. However, this constraint still leaves significant freedom regarding the selection of points, and methods such as CHELP [5], CHELPG [6], Merz-Kollman [7], Spackman [8], Woods [9] differ mainly in this aspect.

Sigfridsson and Ryde [10] proposed the CHELP-BOW method, in which the potential points are Boltzmann-weighted according to the frequency of their occurrence in actual simulations. Boltzmann

*Corresponding author. E-mail: c.pantelides@imperial.ac.uk

ISSN 0892-7022 print/ISSN 1029-0435 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/08927020410001680769

weighting of the various least-squares matrices was also used by Reynolds *et al.* [11] in their attempt to compute a set of charges that could be used for the whole conformational space of the molecule.

Optimal Atomic Charges and their Accuracy

An issue of concern in the literature has been the accuracy of site charges determined via least squares fitting [12]. In particular, charges in larger molecules have been found to exhibit a strong dependence on the orientation of the molecule with respect to the cartesian axes employed for the quantum mechanical calculations, and a non-smooth dependence on the molecular conformation. Moreover, the least squares procedure does not always assign equal charges to atoms related by molecular symmetry, while some atoms are assigned non-physically large charges. These problems were initially attributed to the sampling procedure used, and in particular, the insufficient number of sampling points.

Bayly *et al.* [13] reported that, in many cases, the linear least squares (LLS) solution yielded unreasonably high charges. A restrained model was devised to ensure that charge magnitudes stay within reasonable limits. The authors claimed that the introduction of suitable restraints (RESP) in the form of a penalty function alleviates the problem without affecting the quality of the fit. However, such devices bear little direct relation to the physics of the problem. They also involve several arbitrary decisions such as the choice of the weights of the restraint functions and the value at which the charges should be restrained. Woods and Chappelle [14] validated RESP-derived charges for use in condensed phase simulations through molecular dynamics studies of carbohydrate crystals. Their study demonstrated the importance of the electrostatic interactions in the successful modelling of crystals of polar molecules.

The computation of site charges can also take account of constraints due to molecular symmetry as well as the need to match exactly the electrostatic moments of the quantum mechanical field. Such constraints can be handled through standard techniques such as those making use of Lagrange multipliers. Alternatively, these linear constraints may be used to eliminate some of the optimisation decision variables through the use of generalized inverses [10,15] or projection operators [16,17]. It has been suggested that the increase in the dimensionality of the system caused by the use of Lagrange multipliers aggravates the difficulties associated with the determination of atomic charges, while, conversely, reducing the size of the problem by exploiting the constraint linearity has a beneficial effect.

Several researchers have judged the ability of least-squares fitting to yield an accurate solution through the condition number of the least squares matrix [18] or the matrix of the sensitivity coefficients [12]. It has also been suggested that the number of charges that can be determined unambiguously by such fitting procedures is equal to the number of singular values s_j whose value exceeds $10^k \min_i s_i$, where k is of the order 4–5. A difficulty with all such criteria is that both condition numbers and singular values depend on the scaling of the rows and columns of the matrix (e.g. the units of measurement of charges and potentials). Moreover, in practice, it is very difficult to relate the accuracy of the solution x of a linear system $\mathbf{Ax} = \mathbf{b}$ to the condition number of matrix $\mathbf{A}$—or indeed to determine a universally applicable threshold, above which accuracy is supposed to become problematic.

Other researchers have attempted to relate the quality of the computed solution to statistical quantities associated with least-squares fitting. In particular, Spackman [8] mentions that the standard deviations of the charges defined through the least squares matrix can provide a good estimate of the variation of the charges with molecular rotation. Also, Stouch and Williams [19] examined the least squares variance–covariance matrix and identified co-linearities that reduce the dimensionality of the data to a number well below the number of site charges.

In considering least squares fitting procedures, it is important to distinguish between numerical stability problems and those related to insufficient statistical significance of the estimated charges. The former class of problems (e.g. those that may arise from exceedingly large condition numbers) can often be remedied by better numerical methods and/or problem reformulations. However, such actions cannot resolve statistical accuracy difficulties which are intrinsic to the estimation of certain quantities from certain data. For example, the charge on an atom that is buried inside a tetrahedral arrangement of other atoms will always be more difficult to estimate because the effects of this charge on the electrostatic field are masked by those of the other atoms (cf. Bayly *et al.* [13], Spackman [8] and also the section "Molecules with only atomic charges: CH_3CN and $HCONH_2$" of this paper). In such cases, one may have to change the data set used for the estimation, e.g. by selecting sampling points closer to the molecular surface [18]. Alternatively, one can restrict the set of quantities that are being estimated, e.g. by fixing the charge of buried atoms to chemically reasonable values and applying least squares fitting to the remaining charges.

The Need for Non-atomic Site Charges

Because of the anisotropic nature of the charge distribution, models based exclusively on atomic charges may be unable to describe the intermolecular

interactions, unless one focuses on only long distances. One approach in addressing this difficulty is via the use of distributed multipoles [20] which allows each atom to carry higher moments (e.g. dipole, quadrupole, octupole etc.) in addition to its charge. This, however, also results in considerably higher computational cost.

An alternative approach which aims to improve the accuracy of distributed charge models without significant increase in computational cost is that of introducing additional non-atomic charges. A typical example is molecular nitrogen, which has zero total charge and dipole, but a non-zero quadrupole moment. If two atomic charges are used, then symmetry dictates that both of them must be zero. A solution to the problem [21] is to include a number of non-atomic charges on the axis connecting the two nitrogen atoms. Also in the case of cycloalkanes, the use of only atomic charges has been found to lead to large deviations from the quantum mechanical field [22]. The importance of including additional sites was also emphasised by Dixon and Kollman [23] who placed charges at the approximate locations of the electron lone pair. This addition resulted in better angular dependence of hydrogen bonds and improved accuracy in solvation free energies. Williams [24] reported that atomic charge models fail badly for alkanes higher than methane. The relative root mean square (RRMS) error in ethane with atomic charges only is 97%, while the inclusion of additional charges at the methylene bisector can reduce this to 7%. The optimum position of methylene bisector charges was found to be at a distance of approximately 0.6 Å from the carbons, suggesting an accumulation of charge between the hydrogen atoms. This was verified by an experimental electron density study of ethane [25].

For more complex molecules, a more rigorous approach for the positioning of the non-atomic charges is required. One cannot always rely on chemical intuition as site charges are not quantum mechanical observables or experimentally measurable quantities with a well understood physical meaning. For example, positioning a non-atomic charge at the location of a lone pair electron (identified as a local minimum in the electrostatic potential) does not necessarily lead to the optimal description of the electrostatic field. For example, Singh and Kollman [4] concluded that, in the case of water, the lone pairs in the optimal model are inverted from the tetrahedral direction and point in the direction of the hydrogens; this arrangement has been found to lead to excellent agreement with the true dipole and quadrupole moments, unlike the case where the lone pairs are placed in the chemically intuitive direction.

Outline of this Paper

Stone [26, page 122] noted the lack of a systematic procedure for choosing the positions of satellite charges in distributed charge models, and this continues to be the case to date. This paper aims to address this deficiency by establishing a rigorous and generally applicable methodology for the estimation of site charge models from a set of points sampling a quantum mechanical electrostatic field. Charges may be placed both at atomic positions and at satellite positions associated with a specified subset of the molecule's atoms. The procedure determines optimal estimates of all charge magnitudes as well as the positions of the satellite charges. This is achieved automatically. Chemical intuition is needed only in selecting the subset of the molecule's atoms with which satellite charges will be associated.

The second section of this paper presents the mathematical formulation of the problem being addressed. Particular emphasis is placed on molecular symmetry and its implications regarding charge magnitudes and satellite charge positions. Constraints on the acceptable deviations of the electrostatic field moments (i.e. total charge, dipole moment and quadrupole) are also incorporated in the formulation.

The mathematical formulation can be viewed as a bilevel optimisation problem comprising an inner and an outer optimisation problem. The former involves the determination of optimal charge magnitudes for given positions of satellite charges. This is equivalent to the problem considered in the literature reviewed in the "Optimal Atomic Charges and Their Accuracy" section. The third section demonstrates how the accuracy of the estimated charge magnitudes can be characterised by the confidence intervals of the least squares estimation. This is a useful result even in the context of models involving only atomic charges as it provides a more reliable and direct estimate of statistical significance than those proposed in earlier literature. Moreover, such error estimates are essential for models involving satellite charges to prevent the latter from being positioned in a way that leads to inaccurate estimates of certain charge magnitudes.

The fourth section deals with the outer optimisation problem that seeks to determine the satellite charge positions subject to constraints that limit the confidence intervals on the charge magnitudes (see above). The constraints on satellite positions arising from molecular symmetry are exploited to reformulate the optimisation problem to a mathematically equivalent form that involves a (often much) smaller number of decision variables. This significantly reduces the computational cost required to determine the global solution of the non-convex outer optimisation problem. The fifth section deals with the details of the global optimization approach used in this work.

The sixth section presents a detailed analysis of the problem of building a site charge model for cyclobutane. In this case, the reformulation procedure reduces the number of degrees of freedom of the outer optimisation problem to only two. This allows us to investigate the entire solution space to an extent that is not normally possible, leading to enhanced understanding of various important issues, especially those related to the statistical significance of charge estimates in site charge models. Some larger molecules are studied in the seventh section.

where $\boldsymbol{M}^{\mathcal{G}}$ is the matrix of the linear transformation associated with the $\mathcal{G}$ symmetry operation given by:

$$\mathbf{M}^{\sigma} \equiv \begin{bmatrix} n_y^2 + n_z^2 - n_x^2 & -2n_x n_y & -2n_x n_z \\ -2n_y n_x & n_x^2 + n_z^2 - n_y^2 & -2n_y n_z \\ -2n_z n_x & -2n_z n_y & n_x^2 + n_y^2 - n_z^2 \end{bmatrix} \tag{3}$$

where $\mathbf{n} \equiv [n_x, n_y, n_z]^T$ the unit vector vertical to the plane of symmetry σ,

$$\mathbf{M}^{\mathcal{C}_n^m} \equiv \begin{bmatrix} n_x^2 + \left(n_y^2 + n_z^2\right)\cos\psi & n_x n_y - n_z \sin\psi - n_x n_y \cos\psi & n_x n_z + n_y \sin\psi - n_x n_z \cos\psi \\ n_y n_x + n_z \sin\psi - n_y n_x \cos\psi & n_y^2 + \left(n_x^2 + n_z^2\right)\cos\psi & n_y n_z - n_x \sin\psi - n_y n_z \cos\psi \\ n_z n_x - n_y \sin\psi - n_z n_x \cos\psi & n_z n_y + n_x \sin\psi - n_z n_y \cos\psi & n_z^2 + \left(n_x^2 + n_y^2\right)\cos\psi \end{bmatrix} \tag{4}$$

PROBLEM STATEMENT

Molecular Structure, Symmetry and Atom Types

Consider a set $\mathcal{A}$ comprising N^A atoms at given positions $\mathbf{r}_{i0}$ with respect to an arbitrary cartesian coordinate system. Let $\mathbf{r}_0$ be the position of the molecular centre of mass.

Molecular symmetry is important in the derivation of optimal site charge models because it may imply certain relations between different charge magnitudes. Moreover, it may impose constraints on the position of the satellite charges. Two atoms i, i' can be related by an operation of one or more symmetry elements. The point group symmetry, and thus the symmetry operations that apply to the molecule under consideration, are functions of the molecular geometry. Let $\mathcal{S}$ denote the set of all symmetry relations present in the molecule:

$$\mathcal{S} \equiv \{(i, i', \mathcal{G}) | i, i' \in \mathcal{A}, \mathcal{G} \in \{\sigma, \mathcal{C}_n^m, \mathcal{S}_n^m, \mathcal{I}\}\} \tag{1}$$

where σ, $\mathcal{C}$, $\mathcal{S}$, $\mathcal{I}$ represent different types of symmetry elements, namely symmetry plane, proper axes of rotation, rotation–reflection axes and inversion centre, respectively. The subscript n refers to different orders of the rotation (-reflection) axis, while the superscript m refers to different operations of the same symmetry element.

If $(i, i', \mathcal{G}) \in \mathcal{S}$ is a symmetry relation, then the coordinates of the atoms i and i_0 are related by the expression:

$$\mathbf{r}_{i0} - \mathbf{r}_0 = \boldsymbol{M}^{\mathcal{G}}(\mathbf{r}_{i'0} - \mathbf{r}_0) \tag{2}$$

where $\mathbf{n} \equiv [n_x, n_y, n_z]^T$ the unit vector parallel to the proper axis of rotation $\mathcal{C}_n$ and $\psi = m(2\pi/n)$,

$$\mathbf{M}^{\mathcal{S}_n^m} = \left[\mathbf{M}^{\sigma}\mathbf{M}^{\mathcal{C}_n^1}\right]^m \tag{5}$$

$$\mathbf{M}^{\mathcal{I}} = -\mathbf{I} \tag{6}$$

If an atom i lies on the plane, axis or point that characterises a symmetry element present in the point group, then it is related to itself, and thus the set $\mathcal{S}$ may also contain entries of the form $\{i, i, \mathcal{G}\}$.

The number of unknown charges that need to be estimated can be reduced if we introduce the concept of atom types. An atom type $\mathcal{A}_t$ is the set of atoms that are related in pairs by at least one symmetry operation. Two atoms of the same type t must have the same charges Q_{t0}. Moreover, if they have satellite charges, then these must also be of the same magnitude. For example, all carbon atoms in the benzene molecule are related by a six-fold axis of rotation and should, therefore, be assigned equal charges.

More formally, we partition the set of atoms $\mathcal{A}$ into the *minimum* number N^T of disjoint subsets $\mathcal{A}_t$:

$$\mathcal{A} = \bigcup_{t=1}^{N^T} \mathcal{A}_t, \quad \mathcal{A}_t \cap \mathcal{A}_{t'} = \emptyset, \quad \forall t \neq t' \tag{7}$$

Two atoms i, i' belong to the same subset $\mathcal{A}_t$ for some $t \in [1, \ldots, N^T]$ if and only if there exists a symmetry operation $\mathcal{G}$ such that $(i, i', \mathcal{G}) \in \mathcal{S}$.

Description of the Electrostatic Field

For a given molecular geometry, quantum mechanical calculations are used to compute the following

molecular properties:

- a set of N^P pairs of positions and electrostatic potential values:

$$\left(\mathbf{r}_k^{\mathrm{QM}}, U_k^{\mathrm{QM}}\right), \quad k = 1, \ldots, N^P \tag{8}$$

- the first three moments of the molecular electrostatic field (i.e. the total charge Q^{QM} (usually zero), the dipole $\boldsymbol{\mu}^{\mathrm{QM}}$ and the quadrupole $\boldsymbol{\Theta}^{\mathrm{QM}}$) evaluated at the centre of mass $\mathbf{r}_0$.

The above constitute the basic data used for the fitting of the site charge model.

Satellite Charges

If two atoms belong to the same type, then their satellite charges should be equal in number and magnitude. Let $N_t^S, \forall t = 1, \ldots, N^T$ be the number of satellite charges for atoms of type t. For the purposes of this paper, we will assume that N_t^S is either 0 or 1. Let $N^{TS} = \sum_{t=1}^{N^T} N_t^S$ denote the number of distinct atom types that are allowed to have satellite charges. In the interests of simplicity of notation and without loss of generality, we assume that it is the first N^{TS} atom types (out of the total number of types N^T) that have a satellite charge. Let $\mathbf{r}_{i1}$ denote the position of the satellite charge of atom i, and Q_{t1} the magnitude of the satellite charge assigned to atoms of type t.

The distance between any satellite charge and the corresponding atom is not allowed to exceed a given upper bound $r^{\max}$. This prevents the satellites from approaching the molecular surface, outside of which the electrostatic potential is sampled: the electrostatic potential is not well defined in the area very close to the point charges. Moreover, the satellite charges are not allowed to lie at distances smaller than $r^{\min}$ from the parent atom, otherwise the estimates of the satellite and atomic charges may become unrealistically large and opposite in sign, as observed by Singh and Kollman [4]. Overall, we need to take account of constraints of the form:

$$\begin{aligned} &r^{\min} \le \|\mathbf{r}_{i1} - \mathbf{r}_{i0}\| \le r^{\max}, \quad \forall i \in \mathcal{A}_t, \\ &\forall t = 1, \ldots, N^{TS} \end{aligned} \tag{9}$$

Mathematical Formulation

The estimation procedure seeks to determine:

- the charge Q_{t0} associated with each atom type $t = 1, \ldots, N^T$;
- for each of the first N^{TS} types, t:
  - the satellite charge magnitude Q_{t1}
  - the position $\mathbf{r}_{i1}$ of this charge for each atom $i \in \mathcal{A}_t$

so as to minimise the deviation of the site charge model from an electrostatic field described by data 8. Thus, the vectors of decision variables $\mathbf{Q} \in \mathbb{R}^{NQ}$ and $\mathbf{r}^S \in \mathbb{R}^{3N^S}$ are defined as:

$$\mathbf{Q} \equiv \left\{ \bigcup_{t=1,\ldots,N^T} \{Q_{t0}\} \right\} \cup \left\{ \bigcup_{t=1,\ldots,N^{TS}} \{Q_{t1}\} \right\} \tag{10}$$

$$\mathbf{r}^S \equiv \bigcup_{i \in \mathcal{A}_t t=1,\ldots,N^{TS}} \{\mathbf{r}_{i1}\} \tag{11}$$

where N^Q is the number of independent charge magnitudes:

$$N^Q = N^T + N^{TS} \tag{12}$$

and N^S is the total number of atoms that have satellite charges, i.e.:

$$N^S = \sum_{t=1}^{N^{TS}} |\mathcal{A}_t| \tag{13}$$

where $|\mathcal{A}_t|$ denotes the cardinality of set $\mathcal{A}_t$.

Optimisation Objective

We seek to determine charges $\mathbf{Q}$ and satellite charge positions $\mathbf{r}^S$ that minimise the residual:

$$\min_{\mathbf{Q},\mathbf{r}^S} \Phi \equiv \frac{1}{2} \sum_{k=1}^{N^P} \left[U_k^{\mathrm{QM}} - U\left(\mathbf{r}_k^{\mathrm{QM}}; \mathbf{r}^S, \mathbf{Q}\right) \right]^2 \tag{14}$$

where $U(\mathbf{r}_k^{\mathrm{QM}}; \mathbf{r}^S, \mathbf{Q})$ is the potential of the electrostatic field that is generated by the site charge model at the point $\mathbf{r}_k^{\mathrm{QM}}$ given by:[†]

$$U(\mathbf{r}; \mathbf{r}^S, \mathbf{Q}) \equiv \sum_{t=1}^{N^T} \sum_{i \in \mathcal{A}_t} \sum_{j=0}^{N_t^S} \frac{Q_{tj}}{\|\mathbf{r} - \mathbf{r}_{ij}\|} \tag{15}$$

The right hand side of the above equation involves a triple summation over the atom types t, the atoms i that belong to each of these types and their satellites (if any).

In the literature, the quality of fit is usually measured by two indicators [8], the root mean square (RMS) residual (typically in kcal/mol) defined as:

$$\mathrm{RMS} \equiv \left[\frac{2\Phi}{N^P} \right]^{1/2} \tag{16}$$

[†]For simplicity of notation, we neglect the constant $1/4\pi\epsilon_0$.

and the dimensionless RRMS expressed as a percentage and given by:

$$\mathrm{RRMS} \equiv 100 \left[\frac{2\Phi}{\sum_{k=1}^{N^P} \left(U_k^{\mathrm{QM}} \right)^2} \right]^{1/2} \tag{17}$$

If the charge magnitudes are fitted at points that are closer to the molecular surface, the corresponding values of the electrostatic potential will be larger and thus the RMS will generally be larger as well. Thus, RMS-based comparisons of results of different researchers are meaningful only if the electrostatic field was sampled at sets of points separated by the same distance from the molecular surface. The RRMS is also affected by the choice of sampling points. Moreover, it conveys little information on the absolute deviation of the modelled electrostatic field from the exact one, an important consideration in molecular modelling where electrostatic forces are combined with others.

Matching the Moments of the Electrostatic Field

In addition to minimising Φ, we often desire the values of the first few molecular electrostatic moments computed through the site charge model to be within specified tolerances from the QM computed ones. The first three such moments computed at the molecular centre of mass $\mathbf{r}_0$ are given by:

$$\hat{Q}_{\mathrm{tot}}(\mathbf{Q}) \equiv \sum_{t=1}^{N^T} \sum_{i \in \mathcal{A}_t} \sum_{j=0}^{N_t^S} Q_{tj} \tag{18}$$

$$\hat{\boldsymbol{\mu}}(\mathbf{Q}, \mathbf{r}^S) \equiv \sum_{t=1}^{N^T} \sum_{i \in \mathcal{A}_t} \sum_{j=0}^{N_t^S} Q_{tj} (\mathbf{r}_{ij} - \mathbf{r}_o) \tag{19}$$

$$\hat{\boldsymbol{\Theta}}(\mathbf{Q}, \mathbf{r}^S) \equiv \frac{1}{2} \sum_{t=1}^{N^T} \sum_{i \in \mathcal{A}_t} \sum_{j=0}^{N_t^S} Q_{tj} \left[3(\mathbf{r}_{ij} - \mathbf{r}_o)(\mathbf{r}_{ij} - \mathbf{r}_o)^T - \|\mathbf{r}_{ij} - \mathbf{r}_o\|^2 \mathbf{I} \right] \tag{20}$$

where $\hat{Q}_{\mathrm{tot}} \in \mathbb{R}$, $\hat{\boldsymbol{\mu}} \in \mathbb{R}^3$ and $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^3 \times \mathbb{R}^3$. In order to enforce a certain degree of matching between these moments and the corresponding quantum mechanical values, we specify non-negative relative tolerances ϵ^Q, ϵ^μ, ϵ^Θ, and impose the constraints:

$$-\epsilon^Q |Q_{\mathrm{tot}}^{\mathrm{QM}}| \le \hat{Q}_{\mathrm{tot}} - Q_{\mathrm{tot}}^{\mathrm{QM}} \le \epsilon^Q |Q_{\mathrm{tot}}^{\mathrm{QM}}| \tag{21}$$

$$-\epsilon^\mu |\mu_a^{\mathrm{QM}}| \le \hat{\mu}_a - \mu_a^{\mathrm{QM}} \le \epsilon^\mu |\mu_a^{\mathrm{QM}}|, \; a = 1,\ldots,3 \tag{22}$$

$$-\epsilon^\Theta |\Theta_{aa'}^{\mathrm{QM}}| \le \hat{\Theta}_{aa'} - \Theta_{aa'}^{\mathrm{QM}} \le \epsilon^\Theta |\Theta_{aa'}^{\mathrm{QM}}|, \; a = 1,2, \; a' = a,\ldots,3 \tag{23}$$

where $Q_{\mathrm{tot}}^{\mathrm{QM}}$, $\boldsymbol{\mu}^{\mathrm{QM}}$ and $\boldsymbol{\Theta}^{\mathrm{QM}}$ are the moments obtained by the quantum-mechanical calculation.[‡] In some cases, we may insist on matching one or more of the electrostatic moments exactly, by setting the corresponding ϵ to zero.

Since, by its definition, $\boldsymbol{\Theta}$ is a symmetric 3×3 matrix with zero trace, it is sufficient to match only 5 of its elements. Also, molecular symmetry may impose further constraints on the higher moments (see, for example, Gelessus *et al.* [27]). For example, if the charge distribution is symmetric about the z-axis then $\Theta_{xx} = \Theta_{yy} = -(1/2)\Theta_{zz}$.

Statistical Accuracy of Estimated Charges

The quality of estimates of the charges Q depends on three main factors:

- the structure of the molecule: as mentioned in the "Optimal Atomic Charges and Their Accuracy" section of this paper, some charges may be intrinsically more difficult to estimate than others;
- the available set of QM results;
- the positions of the satellite charges.

The first two of the above are given data as far as this paper is concerned. However, the satellite charge positions are decision variables. This implies that, in trying to determine the optimal positioning of satellite charges, we should avoid positions that lead to poor accuracy of the corresponding charge estimates. For this reason, we will impose an upper bound $\delta Q^{\max}$ on the statistical error δQ_{tj} associated with each estimated charge Q_{tj}:

$$\delta Q_{tj} \le \delta Q^{\max} \quad \forall t = 1,\ldots,N^T, \forall j = 0,\ldots,N_t^S \tag{24}$$

In section "Confidence Intervals for Charge Estimates", we shall return to consider the estimation of the errors δQ_{tj} based on a statistical analysis of the least squares fitting procedure.

OPTIMAL CHARGES FOR GIVEN SATELLITE CHARGE POSITIONS

The optimisation problem of interest comprises the minimisation of the objective function (14) with

[‡]Alternatively, one could use the experimental values if they are available.

respect to the charges $\mathbf{Q}$ and satellite charge positions $\mathbf{r}^S$ (cf. Eqs. 10 and 11), subject to constraints (21)–(24), for given values of the quantities U_k^{QM}, $\mathbf{r}_k^{\mathrm{QM}}$, $\forall k = 1, \ldots, N^P$, and the tolerances ϵ^Q, ϵ^μ, ϵ^Θ and $\delta Q^{\max}$. This can be expressed as a bi-level optimisation problem of the form:

$$\min_{\mathbf{r}^S} \Phi^*(\mathbf{r}^S) \tag{25}$$

where the function $\Phi^*(\mathbf{r}^S)$ is itself the result of a minimisation problem, and is defined as:

$$\Phi^*(\mathbf{r}^S) \equiv \min_{\mathbf{Q}} \Phi(\mathbf{Q}, \mathbf{r}^S) \tag{26}$$

Thus, we have:

- an inner optimisation problem (26) which determines the optimal charges $\mathbf{Q}(\mathbf{r}^S)$ for any given satellite charge positions $\mathbf{r}^S$;
- an outer optimisation problem (25) that determines the optimal values of the satellite positions $\mathbf{r}^S$.

This section will deal with the inner minimisation problem. The least squares estimation of charges located at given positions has been addressed extensively in the literature (cf. section "Optimal atomic charges and their accuracy"), and is reviewed here only to the extent necessary for establishing the notation required for the definition of the outer problem. We also consider the formal description of the statistical accuracy of the determined charges through the use of confidence intervals in constrained parameter estimation problems.

Optimality Conditions for the Inner Minimisation Problem

The inner optimisation solves the problem:

$$\min_{\mathbf{Q}} \Phi = \frac{1}{2} \sum_{k=1}^{N^P} \left[U_k^{\mathrm{QM}} - U\left(\mathbf{r}_k^{\mathrm{QM}}, \mathbf{r}^S; \mathbf{Q}\right) \right]^2 \tag{27}$$

where the positions of the satellite charges $\mathbf{r}^S$ are now known quantities. This is a LLS problem and, thus, can be solved analytically. This property is preserved even if we augment the original minimisation with exact moment matching constraints, i.e. constraints (21)–(23) (or a subset thereof) with the corresponding ϵ quantities being set to zero. These constraints become equalities that are linear in $\mathbf{Q}$ and can be expressed as:

$$\mathbf{C}(\mathbf{r}^S)\mathbf{Q} = \mathbf{c} \tag{28}$$

The number N^C of such constraints typically ranges from 1 (if only the total charge is matched exactly) to 9 (if the total charge, dipole moment and quadrupole are all matched exactly). The matrix $\mathbf{C} \in \mathbb{R}^{N^C} \times \mathbb{R}^{N^Q}$ depends on the satellite charge positions $\mathbf{r}^S$ but not on the charges $\mathbf{Q}$; the vector $\mathbf{c} \in \mathbb{R}^{N^C}$ is constant.

If we define the auxiliary quantities:

$$w_{k,tj} \equiv \sum_{i \in \mathcal{A}_t} \frac{1}{\|\mathbf{r}_k^{\mathrm{QM}} - \mathbf{r}_{ij}\|}, \quad \forall k = 1, \ldots, N^P, \quad t = 1, \ldots, N^T, \quad j = 0, \ldots, N_t^S \tag{29}$$

then Eq. (15) can be rewritten as:

$$U(\mathbf{r}_k^{\mathrm{QM}}, \mathbf{r}^S; \mathbf{Q}) = \sum_{t=1}^{N^T} \sum_{j=0}^{N_t^S} Q_{tj} w_{k,tj} \tag{30}$$

By introducing Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{N^C}$ corresponding to the linear constraints (28), we can write the first-order optimality conditions for the inner minimisation problem as a square system of linear equations:

$$\mathbf{A}(\mathbf{r}_s) \begin{pmatrix} \mathbf{Q} \\ \boldsymbol{\lambda} \end{pmatrix} = \mathbf{b}(\mathbf{r}_s) \tag{31}$$

where the $(N^Q + N^C) \times (N^Q + N^C)$ square non-singular matrix $\mathbf{A}$ is given by:

$$\mathbf{A}(\mathbf{r}_s) \equiv \begin{bmatrix} \mathbf{H} & \mathbf{C}^T \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \tag{32}$$

The $N^Q \times N^Q$ symmetric matrix $\mathbf{H}$ is given by:

$$H_{l(t,j)l(t',j')} = \sum_{k=1}^{N^P} w_{k,tj} w_{k,t'j'}, \quad \forall t, t' = 1, \ldots, N^T, \quad j = 0, \ldots, N_t^S, \; j' = 0, \ldots, N_{t'}^S \tag{33}$$

where the index $l(t,j)$ is defined by:

$$l(t,j) \equiv \begin{cases} t & \text{if } j = 0 \\ N^T + t & \text{if } j \neq 0 \end{cases} \tag{34}$$

The right hand side vector $\mathbf{b} \in \mathbb{R}^{N^Q + N^C}$ of Eq. (31) is defined as:

$$b_t = \sum_{k=1}^{N^P} U_k^{\mathrm{QM}} w_{k,t0}, \quad \forall t = 1, \ldots, N^T \tag{35}$$

$$b_{N^T + t} = \sum_{k=1}^{N^P} U_k^{\mathrm{QM}} w_{k,t1}, \quad \forall t = 1, \ldots, N^{TS} \tag{36}$$

$$b_{N^T + N^{TS} + l} = c_l, \quad \forall l = 1, \ldots, N^C \tag{37}$$

Confidence Intervals for Charge Estimates

The error in the charges $\mathbf{Q}$ obtained through the solution of Eq. (31) can be related to the accuracy of the data used in that equation (i.e. the quantities derived from the quantum mechanical calculations) via the associated confidence intervals. Briefly, the x%

confidence interval is a quantity δQ such that the probability of the true value of the charge being between $Q - \delta Q$ and $Q + \delta Q$ is x%, where Q denotes the optimal estimate. The corresponding statistical significance α of the estimate is then $1 - x/100$.

The confidence intervals for LLS estimation are proportional to the square root of the diagonal elements of the variance covariance matrix [28]:

$$\delta Q_{tj}(\mathbf{r}^S) \approx t_{1-\alpha/2,N^P-N^Q+N^C}\sqrt{\hat{V}_{l(t,j),l(t,j)}(\mathbf{r}^S)},$$
$$t = 1,\ldots,N^T, \quad j = 0,\ldots,N_t^S \tag{38}$$

where $l(t, j)$ is defined by Eq. (34). The coefficient of proportionality $t_{1-\alpha/2,N^P-N^Q+N^C}$ is the value of the Student function for statistical significance α and $N^P - N^Q + N^C$ degrees of freedom.

For LLS problems subject to equality constraints (28), the variance covariance matrix $\hat{\mathbf{V}}$ is given by page 182 of Bard ([28]):

$$\hat{\mathbf{V}} = s^2\mathbf{T}\mathbf{H}^{-1}\mathbf{T}^T \tag{39}$$

where:

$$\mathbf{T} \equiv \mathbf{I} - \mathbf{H}^{-1}\mathbf{C}^T[\mathbf{C}\mathbf{H}^{-1}\mathbf{C}^T]^{-1}\mathbf{C} \tag{40}$$

and:

$$s^2 \equiv \frac{1}{N^P - N^Q + N^C}\sum_{k=1}^{N^P}\left[U_k^{\mathrm{QM}} - U\left(\mathbf{r}_k^{\mathrm{QM}}, \mathbf{r}^S; \mathbf{Q}\right)\right]^2 \tag{41}$$

OPTIMISATION OF SATELLITE CHARGE POSITIONS

The inner optimisation problem of section "Optimal Charges for Given Satellite Charge Positions" determines the optimal charge magnitudes and their accuracies as explicit functions of the satellite charge positions (cf. Eqs. 31 and 38). The outer optimisation problem considered in this section determines optimal values for these positions.

Basic Statement of the Outer Optimisation Problem

The outer optimisation problem can be formulated as:

$$\min_{\mathbf{r}^S} \Phi^*(\mathbf{r}^S) = \frac{1}{2}\sum_{k=1}^{N_P}\left[U_k^{\mathrm{QM}} - U\left(\mathbf{r}_k^{\mathrm{QM}}; \mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S\right)\right]^2 \tag{42}$$

subject to:

$$r^{\min} \le \|\mathbf{r}_{i1} - \mathbf{r}_{i0}\| \le r^{\max}, \quad \forall i \in \mathcal{A}_t,$$
$$\forall t = 1,\ldots,N^{TS} \tag{43}$$

$$\delta Q_{tj}(\mathbf{r}^S) \le \delta Q^{\max} \quad \forall t = 1,\ldots,N^T,$$
$$\forall j = 0,\ldots,N_t^S \tag{44}$$

When the electrostatic moment matching constraints are not enforced exactly as part of the inner problem (i.e. when the tolerances $\boldsymbol{\epsilon}$ are set to nonzero values), they are included as inequality constraints in the outer optimisation problem:

$$-\boldsymbol{\epsilon}^Q|Q_{\mathrm{tot}}^{\mathrm{QM}}| \le \hat{Q}_{\mathrm{tot}}(\mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S) - Q_{\mathrm{tot}}^{\mathrm{QM}} \le \boldsymbol{\epsilon}^Q|Q_{\mathrm{tot}}^{\mathrm{QM}}| \tag{45}$$

$$-\boldsymbol{\epsilon}^\mu|\mu_a^{\mathrm{QM}}| \le \hat{\mu}_a(\mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S) - \mu_a^{\mathrm{QM}} \le \boldsymbol{\epsilon}^\mu|\mu_a^{\mathrm{QM}}|, \quad a = 1,\ldots,3 \tag{46}$$

$$-\boldsymbol{\epsilon}^\Theta|\Theta_{aa'}^{\mathrm{QM}}| \le \hat{\Theta}_{aa'}(\mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S) - \Theta_{aa'}^{\mathrm{QM}} \le \boldsymbol{\epsilon}^\Theta|\Theta_{aa'}^{\mathrm{QM}}|, \quad a = 1,2, \quad a' = a,\ldots,3 \tag{47}$$

The symmetry relations between atoms of the same type also hold for their satellites.

This means that the positions of the latter should satisfy (cf. Eq. 2):

$$\mathbf{r}_{i1} - \mathbf{r}_0 = \mathbf{M}^{\mathcal{G}}(\mathbf{r}_{i'1} - \mathbf{r}_0), \quad \forall(i, i', \mathcal{G}) \in \mathcal{S} \tag{48}$$

The charge magnitudes $\mathbf{Q}(\mathbf{r}^S)$ used in the objective function (42) are computed by solving the linear system (31), while the confidence intervals $\delta\mathbf{Q}(\mathbf{r}^S)$ can be obtained from Eq. (38). The molecular electrostatic moments $\hat{Q}_{\mathrm{tot}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Theta}}$ in Eqs. (45)–(47) are computed using Eqs. (18)–(20).

Transformed Outer Optimisation Problem

Unlike the inner optimisation problem, the outer one is nonlinear and cannot be solved exactly. Moreover, it is a nonconvex optimisation problem with multiple local minima. Consequently, a global optimisation approach will have to be employed to determine the best possible satellite charge positions.

The computational complexity of currently available global optimisation techniques is a very strong function of the number of independent decision variables. This number can be reduced by exploiting the molecular symmetry constraints (48), which can be written in the form:

$$\mathbf{r}_{i1} - \mathbf{M}^{\mathcal{G}}\mathbf{r}_{i'1} = (\mathbf{I} - \mathbf{M}^{\mathcal{G}})\mathbf{r}_o, \quad \forall(i, i', \mathcal{G}) \in \mathcal{S} \tag{49}$$

or, in matrix notation:

$$\mathbf{M}\mathbf{r}^S = \mathbf{d} \tag{50}$$

This is a linear system with $3|\mathcal{S}|$ rows and $3N^S$ unknowns. Because of the existence of redundant

symmetry relations, $|\mathcal{S}|$ may exceed N^S. However, the rank of the linear system will always be less than the unknown satellite charge coordinates, i.e.:

$$\text{rank}\,(\mathbf{M}) < 3N^S \tag{51}$$

The singular value decomposition (SVD) [29] of the matrix $\mathbf{M}$[29] is of the form:

$$\mathbf{M} = \mathbf{U}\mathbf{S}\,\mathbf{V}^T \tag{52}$$

where the square matrix $\mathbf{S}$ will contain $N^X = 3N^S - \text{rank}\,(\mathbf{M})$ singular values, i.e. zero diagonal elements. Any vector of satellite charge positions that satisfies (48) can be expressed as:

$$\mathbf{r}^S = \mathbf{x}_0 + \sum_{h=1}^{N^X} e_h \mathbf{x}_h \tag{53}$$

where $\mathbf{x}_h \in \mathbb{R}^{3N^S}$, $h = 1, \ldots, N^X$ is the column of the matrix $\mathbf{V}$ that corresponds to the hth zero diagonal element of the matrix $\mathbf{S}$, and $\mathbf{x}_0$ is the solution of the linear system (50) with the smallest norm. Both $\mathbf{x}_0$ and $\mathbf{x}_h$, $h = 1, \ldots, N^X$ can be obtained directly from the application of standard SVD algorithms on the constant matrix $\mathbf{M}$.

The variable transformation (53) replaces the vector of $3N^S$ optimisation decision variables $\mathbf{r}^S$ by the shorter vector $\mathbf{e} \in \mathbb{R}^{N^X}$ and also eliminates the equality constraints (48). This results in a smaller outer optimisation problem to be solved to global optimality. It is worth noting that the SVD operation for the determination of N^X and the vectors $\mathbf{x}_h$, $h = 1, \ldots, N^X$ is performed only once for any particular molecule; hence the associated computational cost is negligible. Moreover, the results obtained are the same regardless of any linear dependence of the rows of matrix $\mathbf{M}$ caused by the inclusion of redundant symmetry relations in the set $\mathcal{S}$.

GLOBAL OPTIMISATION ALGORITHM

In practice, the first step for the construction of optimal site charge models is usually to assume that there are no satellite charges, i.e. set $N_t^S = 0$, $\forall t = 1, \ldots, N^T$ and to compute the minimum RRMS deviation for the atomic charge model. If this is acceptably low, there is no need for satellite charges. Otherwise, one has to decide on the chemically important atom types that will be assigned a satellite charge, before performing global optimization with respect to the decision variables $\mathbf{e}$ appearing in Eq. (53). This section presents an outline of the global optimisation algorithm used in this work.

The Quasi Monte-Carlo Multi-level Single Linkage (MLSL) Method

The global optimisation is performed by means of a quasi Monte-Carlo variant of the multi-level single linkage (MLSL) algorithm. This comprises of two phases. In the *global phase*, the objective function is evaluated at a number of suitably chosen points. In the *local phase*, some of the more promising points generated during the global phase are used as starting points for local minimisation. By the end of the algorithm, one or more local minima are identified; the one with the lowest value of the objective function is deemed to be the global minimum.

As the local minimisations are computationally the most expensive part of the algorithm, the latter attempts to reduce their number by "clustering". Throughout its operation, the algorithm maintains a set of clusters. Each cluster comprises a number of points, exactly one of which is a local minimum while the others represent initial guesses that lead to that minimum via a local minimisation procedure. A local minimisation is attempted from a particular point (initial guess) only if the latter is not too close to any point belonging to any of the existing clusters.

The steps of an MLSL algorithm with clustering are summarised in Fig. 1. Step 1 initialises various counters and the set of all local minima $\mathbf{E}^*$ that have been detected so far. It also computes lower and upper bounds $e_h^{\min}$, $e_h^{\max}$, $h = 1, \ldots, N^X$ on the decision variables $\mathbf{e}$ as described in section "Bounds on the Decision Variables". Step 2 marks the beginning of a new iteration; every iteration involves a global (step 3) and a local search (step 4).

The global search generates a set of N initial guesses (step 3.1) in the form of a low discrepancy sequence of the type proposed by Sobol' [30]. Such sequences are designed to extract the maximum amount of information regarding the objective function from a given number of points. The interested reader is referred to the literature on stochastic optimisation methods with quasi-random number generators [31]. The N points generated are sorted according to the value of the objective function (step 3.2); only a fraction $\beta \in (0, 1]$ of the most promising points are considered during the local search.

The local search (step 4) considers βN points as potential candidates for initiating local minimisations. A local minimisation is actually performed only if the distance of $\mathbf{e}_l$ from each and every point $\mathbf{e}$ in each of the existing clusters exceeds a certain threshold ρ. Once the local minimisation is performed (step 4.1), the algorithm checks whether the local minimum $\mathbf{e}_l^*$ has already been found at an earlier stage of the algorithm. If this is not the case, then a new cluster is created and both $\mathbf{e}_l$ and $\mathbf{e}_l^*$ are added to it (step 4.2.1); otherwise, the initial guess $\mathbf{e}_l$ is added to the cluster that already contains $\mathbf{e}_l^*$ (step 4.2.2).

0. Given:

   0.1 the number N of points to be generated per iteration

   0.2 the fraction $\beta \in (0, 1]$ of the above points to be considered as possible starting points for local minimisation

   0.3 the adjustable parameter σ used in equation 54

1. Perform initialisation

   1.1 Set the cluster counter $j := 0$

   1.2 Set the iteration counter $k := 0$

   1.3 Set the set of local minima $\mathbf{E}^* := \emptyset$

   1.4 Calculate the decision variable bounds e_h^{min}, e_h^{max}, $h = 1, \ldots, N^X$

2. Start a new iteration k:=k+1

3. Perform global search

   3.1 Create N points $\mathbf{e}_l \in \mathbb{R}^{N^X}$ such that $e_{lh} \in \left[e_h^{min}, e_h^{max}\right], \quad l = 1, \ldots, N, \; h - 1, \ldots, N^X$

   3.2 Sort the points in non-descending objective function value order, *i.e.* $\Phi(\mathbf{e}_l) \leq \Phi(\mathbf{e}_{l+1}), \quad \forall l = 1, \ldots, N - 1$

4. Perform local search

   FOR $l := 1$ TO βN DO

      IF $\|\mathbf{e}_l - \mathbf{e}\| > \rho(N; k), \forall \mathbf{e} \in \bigcup_{j'=1}^{j} C_{j'}$ THEN

      4.1 Perform a local minimisation using $\mathbf{e}_l$ as the initial guess to determine the local minimum $\mathbf{e}_l^*$

      4.2 IF $\mathbf{e}_l^* \notin \mathbf{E}^*$ THEN

         4.2.1 Add $\mathbf{e}_l^*$ to set of local minima: $\mathbf{E}^* := \mathbf{E}^* \bigcup \{\mathbf{e}_l^*\}$

         4.2.2 Create a new cluster: $j := j + 1$

         4.2.3 Add both $\mathbf{e}_l$ and $\mathbf{e}_l^*$ to new cluster: $C_j := \{\mathbf{e}_l, \mathbf{e}_l^*\}$

      ELSE

      4.3 add $\mathbf{e}_l$ to the cluster j^* that already contains $\mathbf{e}_l^*$: $C_{j^*} := C_{j^*} \bigcup \{\mathbf{e}_l\}$

      END IF

   END FOR

5. Check stopping criterion: IF $|\mathbf{E}^*| + 0.5 \leq \hat{N}^E$, THEN go to step 2

6. Determine optimal solution: $\mathbf{e}^* = \arg\min_{\mathbf{e} \in \mathbf{E}^*} \Phi(\mathbf{e})$

FIGURE 1 Outline of global optimisation algorithm.

The clustering threshold $\rho(N; k)$ is given by [32,33]:

$$\rho(N; K) \equiv \left(\frac{m\sigma \log(kN)}{\omega_{N^X} kN}\right)^{1/N^X} \tag{54}$$

where m is the Lebesgue measure of the feasible domain of the optimisation decision variables (equal to 1 if one uses normalised variables, i.e. the sampling is done in the unit hypercube), $\omega_{N^X} \equiv \pi^{N^X/2}/\Gamma(1 + (N^X/2))$ and σ is an adjustable parameter. Lower values of σ lead to fewer points being rejected at step 4 as belonging to existing clusters, and consequently, to more local minimisations. In our implementation, σ was set to 2.

Step 5 makes use of a Bayesian test developed by Boender [34] to determine whether iterations can be terminated. The test is based on the following estimate $\hat{N}^E$ for the number of distinct local minima in the problem:

$$\hat{N}^E \equiv \frac{N^E(k\beta N - 1)}{k\beta N - N^E - 2} \tag{55}$$

where $N^E = |\mathbf{E}^*|$ is the number of local minima that have already been identified. The search for the global minimum terminates if N^E essentially exceeds $\hat{N}^E$; in that case, the global minimum is simply the local minimum with the smallest value of the objective function Φ (step 6). Otherwise, a new iteration is started from step 2.

Bounds on the Decision Variables

The global phase of our algorithm requires the upper and lower bounds for the decision variables e_h,

$h = 1, \ldots, N^X$ in order to generate the candidate points at step 3.1. The bounds on e_h are related to the bounds on the distances of the satellite charges from the nuclei they are assigned to. In particular, they can be determined by solving $2N_X$ linear programming problems. As an illustration, the lower bound for the variable e_h can be computed by solving the problem:

$$e_h^{\min} = \min_{\mathbf{e}} e_h \tag{56}$$

subject to:

$$\mathbf{r}^S = \mathbf{x}_0 + \sum_{h'=1}^{N^X} e_{h'} x_{h'}, \quad -r^{\max} \le r_{i1}^a - r_{i0}^a \le r^{\max},$$
$$\forall i \in \mathcal{A}_t, \quad \forall t = 1, \ldots, N^{TS}, \quad \forall a \in \{x, y, z\} \tag{57}$$

The bounds were computed with the E04MFF linear programming code from NAg (http://www.nag.co.uk). The associated computational cost is negligible for all practical problems.

Solution of Local Minimisation Problems

The local minimisations are performed using the E04UFF successive quadratic programming code from NAg (http://www.nag.co.uk).

The derivatives of the objective function and the constraints with respect to the decision variables are computed exactly using analytical expressions derived in Appendix A. This is important for the reliable identification of local minima and the prevention of premature termination of the optimisation algorithm at non-optimal points. Moreover, it generally results in faster convergence of the local minimisation, thereby reducing the overall computational cost.

CASE STUDY: CYCLOBUTANE

Cyclobutane has two known conformations: the planar and puckered form [35,36]. Here we use the quantum mechanically optimised puckered form to illustrate some of the key concepts introduced in this paper.

Unless otherwise stated, all quantum mechanical calculations relating to the examples presented in this and the following section were performed with the quantum mechanical package GAMESS at the Hartree-Fock (HF)/6–31G** level [37]. The electrostatic field was sampled on a geodesic grid [8] of order $\{3,5+\}_{3,0}$. This is more isotropic than both cubic grids and grids generated by methods based on Connolly surfaces, thus leading to improved rotational invariance.

The *xyz* coordinates for the *ab initio* optimised geometry computed with gradient convergence tolerance of 10^{-6} are presented in Table I. The origin is located at the molecular centre of mass.

The puckered form of cyclobutane belongs to the point group D2d, having one S_4 axis, three C_2 and two symmetry planes σ_d. All four carbons are related by symmetry, but there are two different types of hydrogen atoms.

The quantum mechanical electrostatic potential was computed on a set of 16 layers. The first layer is at distance 1.4 times the van der Waals radii of the atoms, and the distance between consecutive shells is 0.05 times the van der Waals radii. This results in 4280 sampling points that will be used for the fitting of the site charge models.

Site Charge Model Based on Atomic Charges Only

Table II presents the optimal atomic charges obtained when no satellite site charges are used. The only constraint imposed is that the total molecular charge is zero. The error in the estimated charges is based on the 90% confidence intervals, corresponding to a value of the statistical significance parameter α of 0.1 (cf. Eq. 38).

We first perform a computation without imposing any symmetry constraints. The results are shown in columns 4 and 5 of Table II. We note that, even without explicit symmetry constraints, atoms of

TABLE I (HF)/6–31G** optimised cyclobutane (puckered form)

N^o	*Atom*	*x* (Å)	*y* (Å)	*z* (Å)
1	C	0.7622330060	0.7622330847	0.1251207796
2	H	0.9979163371	0.9979159663	1.1580500874
3	H	1.3796410932	1.3796415461	− 0.5178850495
4	C	− 0.7622329442	0.7622329376	− 0.1251207451
5	H	− 1.3796410120	1.3796414643	0.5178849915
6	H	− 0.9979162611	0.9979157414	− 1.1580500328
7	C	− 0.7622329761	− 0.7622328433	0.1251212064
8	H	− 0.9979166067	− 0.9979167970	1.1580501575
9	H	− 1.3796408891	− 1.3796406201	− 0.5178853772
10	C	0.7622329126	− 0.7622329598	− 0.1251210128
11	H	1.3796411914	− 1.3796409046	0.5178851478
12	H	0.9979161489	− 0.9979166156	− 1.1580501528

TABLE II Atomic charges of cyclobutane (puckered form)

			Symmetry constraints included			
			No		*Yes*	
N°	*Atom*	*Type*	*Q (e)*	*δQ (e)*	*Q (e)*	*δQ (e)*
1	C	1	− 0.057710	0.028895	− 0.057707	0.006777
2	H	2	0.044751	0.007556	0.044750	0.003898
3	H	3	0.012957	0.007714	0.012957	0.003272
4	C	1	− 0.057706	0.028895		
5	H	3	0.012956	0.007714		
6	H	2	0.044750	0.007556		
7	C	1	− 0.057708	0.028895		
8	H	2	0.044751	0.007556		
9	H	3	0.012956	0.007714		
10	C	1	− 0.057705	0.028895		
11	H	3	0.012957	0.007714		
12	H	2	0.044751	0.007556		

the same type are assigned almost equal charges; this is due to the isotropic sampling of the electrostatic field that is achieved with the geodesic grid [8]. However, the magnitude of the confidence intervals is approximately equal to 50% of the absolute value of the charges. This implies that the computed charges may be very sensitive to small changes in the quantum mechanical data used, e.g. the charges may change significantly if the quantum mechanical calculation is repeated with the molecule rotated with respect to the cartesian axes or if a different sampling scheme is applied to the electrostatic field. We also observe that the small deviation from planarity that characterises the puckered conformation of cyclobutane is enough to result in significantly different charges for the hydrogen atoms that are not symmetry related.

On the other hand, when symmetry constraints are explicitly imposed (see last two columns of Table II), there are only three independent charges to be estimated, and their statistical accuracy is improved considerably.

Williams [36] studied the planar form of cyclobutane, in which all hydrogens are related by symmetry. He computed the atomic charges to be equal to $Q_C = -0.024$ and $Q_H = 0.012$. Unfortunately, it is difficult to compare our results with these findings because, although the planar conformer appears to be a local minimum under molecular mechanics energy minimisation, it is transformed to the puckered form when it is minimised by *ab initio* quantum mechanics.

Table III presents the quality of the fit in terms of the RMS and RRMS errors. We note that these are not affected by the imposition of symmetry constraints. Moreover, the accuracy of the charges does not seem to be directly related to the condition number of the least squares matrix: for example, the imposition of symmetry constraints causes an increase in the condition number while, as expected, the confidence intervals are reduced significantly (cf. Table II).

Table III also compares the computed dipole and quadrupole moments at the center of mass $\mathbf{r}_0 = [0 \quad 0 \quad 0]^T$ with those obtained via the quantum mechanical calculation. The atomic site charge models have zero dipole moment, which is in agreement with the *ab initio* calculations. However, the components of the quadrupole tensor are not computed correctly.

TABLE III Quality of fit for the atomic site charge model of cyclobutane (puckered form)

	Symmetry constraints included		
	No	*Yes*	*ab initio QM*
RMS (kcal/mol)	1.1076	1.1076	–
RRMS %	87.4	87.4	–
Condition number	1.316×10^6	5.262×10^6	–
μ_x*	1.2391×10^{-5}	-5.7729×10^{-8}	1.3×10^{-5}
μ_y*	-1.4710×10^{-5}	-3.3476×10^{-7}	-1.5×10^{-5}
μ_z*	-3.2862×10^{-6}	-6.8364×10^{-8}	-3.0×10^{-6}
Θ_{xx}†	-2.5829×10^{-1}	-2.5829×10^{-1}	-6.5363×10^{-1}
Θ_{xy}†	-1.9547×10^{-5}	1.9849×10^{-7}	-2.0×10^{-5}
Θ_{xz}†	-6.4604×10^{-6}	1.5810×10^{-7}	-7.0×10^{-6}
Θ_{yy}†	-2.5829×10^{-1}	-2.5829×10^{-1}	-6.5364×10^{-1}
Θ_{yz}†	-5.5605×10^{-6}	7.3371×10^{-8}	-5.0×10^{-6}
Θ_{zz}†	5.1658×10^{-1}	5.1658×10^{-1}	1.3072

*Dipole moments in Debye. †Quadrupole moments in Buckingham.

This factor, together with the very high RRMS values, indicate that satellite charges may need to be considered in order to derive an acceptable site charge model for cyclobutane.

Cyclobutane Site Charge Model with Atomic and Satellite Charges

Here, we attempt to produce an improved description of the electrostatic field by introducing satellite charges. As in section "Site Charge Model Based on Atomic Charges Only", we impose a constraint of zero total charge but no constraints on the dipole and quadrupole moments. If all four carbon atoms were to be assigned a satellite charge, we would need to estimate 16 charge magnitudes and $4 \times 3 = 12$ cartesian coordinates. However, symmetry dictates that all satellite charges should be equal; it also leads to 37 linear constraints on the 12 coordinates of these satellite charges. Of course, not all of these constraints are linearly independent.

In fact, the SVD of the corresponding matrix shows that there are only two independent coordinates, e_1 and e_2 (cf. Eq. 53). Because of the low dimensionality of this particular problem, it is possible to compute the RRMS in the whole space of decision variables, thus gaining some insight on the characteristics of the underlying mathematical problem.

If we require the satellite charges to be kept within distances 0.15 Å and 0.90 Å from the carbon atoms, the method of section "Bounds on the Decision Variables" can be used to compute the following decision variable bounds:

$$-0.391089 \leq e_1 \leq 4.702581 \tag{58}$$

$$-1.550028 \leq e_2 \leq 2.053511 \tag{59}$$

The RRMS for the above solution space is shown in Fig. 2. The shaded area corresponds to points that violate constraints 9. For these points, the combination of values of e_1 and e_2 result in satellite charge positions (cf. Eq. 53) that are either too far from, or too close to the corresponding atoms.

No point in Fig. 2 has a RRMS smaller than 20%. This contrasts with the findings of Williams and Abraha [22] who reported RRMS values of less than 15% for the planar form of cyclobutane. This discrepancy can be attributed, at least partly, to the different regions around the molecule employed for the fit (cf. section "Optimisation Objective"). Williams and Abraha [22] computed the electrostatic field at distances 1.6 times the van der Waals radii and higher; in contrast, our computations placed the first shell of the geodesic grid at distances 1.4 times the van der Waals radii. In general, the long-range

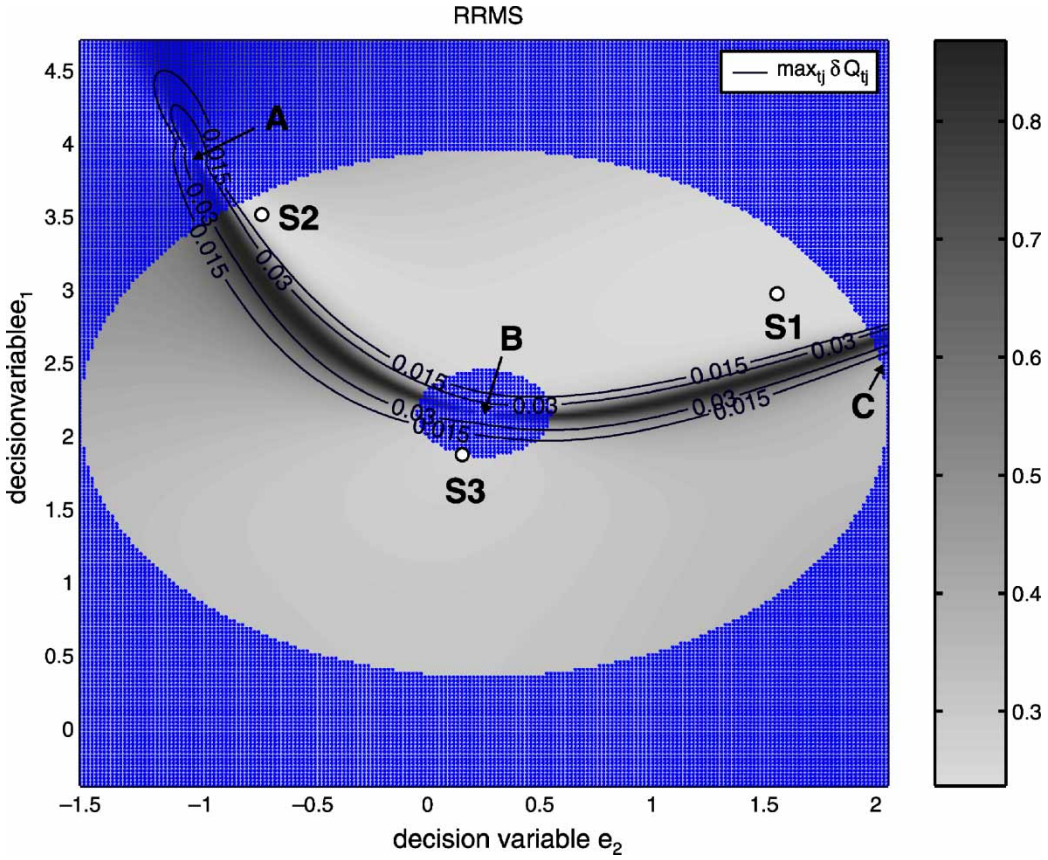


FIGURE 2 RRMS for various positions of the satellite charges for puckered cyclobutane. (Colour version available online.)

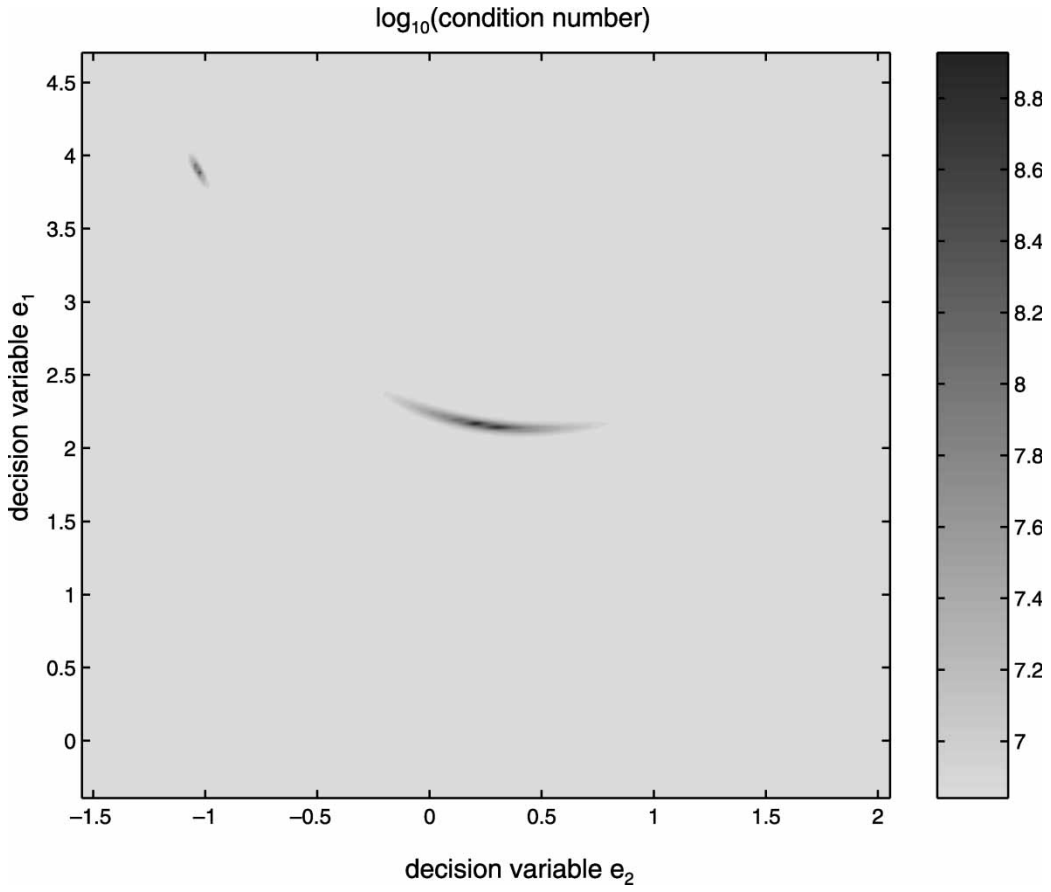


FIGURE 3 Logarithm of condition number of the LLS matrix for various positions of the satellite charges for puckered cyclobutane.

electrostatic potential is easier to fit the farther one gets from the molecular surface.

Figure 2 also shows the contours of the quantity $\max(\delta Q_{H_1}, \delta Q_{H_2}, \delta Q_C, \delta Q_{C_{sat}})$, where δQ_{H_1}, and δQ_{H_2}, are the 90% confidence intervals for the charges of the two different types of hydrogen atom, and δQ_C and $\delta Q_{C_{sat}}$ are the confidence intervals for the carbon atom charge and the corresponding satellite charge. Only points outside the 0.015e contour are likely to be sufficiently accurate to be of practical interest.

Figure 3 shows the condition number of the least squares matrix for the whole solution space. As can be seen, condition numbers are indeed highest in areas of low accuracy (i.e. where the confidence intervals are wide). However, it is very difficult to choose a condition number value as an appropriate threshold, above which charge estimates can be judged as "inaccurate": in fact, every single point in Fig. 3 has a condition number exceeding 10^7. This reinforces the argument that condition numbers are not reliable measures of the accuracy of the estimated charges.

The estimated charge magnitudes are shown in Fig. 4. We note that there are three points A, B and C, at which the carbon satellite charge attains unphysically large magnitudes. These actually correspond to situations under which the satellite charge coincides in space with a hydrogen atom of type 1 or 2 (points C and A, respectively) or the carbon atom (point B). This coincidence also leads to an unphysically large value for the atomic charge involved (see Figs. 4a–c). Points A, B and C are also marked in Fig. 2. We note that these correspond to areas of wide confidence intervals which are automatically excluded from consideration by the imposition of accuracy constraints (24) on the confidence intervals δQ_{tj}. The error in the estimation of the charge magnitudes for the whole solution space is shown in Fig. 5.

Satellite Charge Optimisation

The cyclobutane example is so simple that we can examine the entire solution space as done in "Cyclobutane Site Charge Model with Atomic and Satellite Charges" and then select the optimal satellite positions by inspection. Nevertheless, it is instructive to apply the optimisation algorithm to this problem. The results are shown in Table IV. The algorithm found three distinct local minima with only four local minimisations at step 4 of the algorithm of Fig. 1. The minima are shown in Fig. 2.

The details for all three local minima are shown in Table V. The global minimum is the point *S*1 (see Fig. 2) with a RRMS of 23.7%. In fact, this point

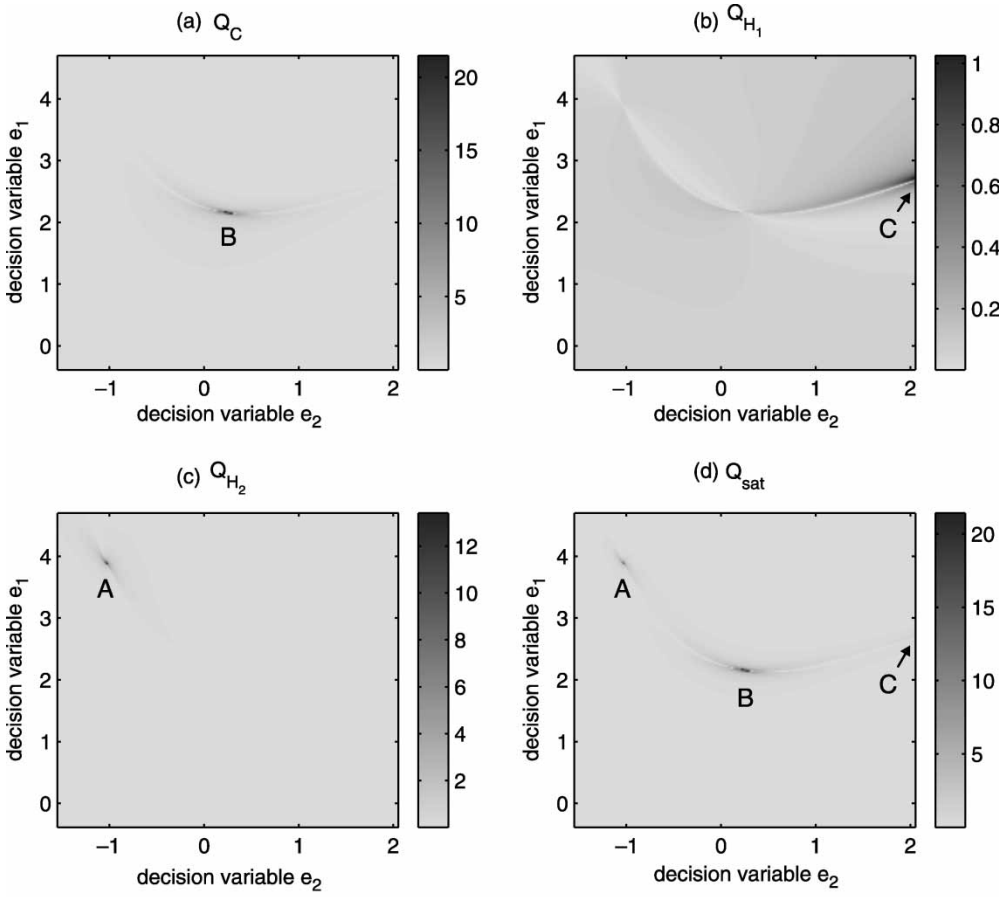


FIGURE 4 Charge magnitudes for various positions of the satellite charges for puckered cyclobutane.

is physically more realistic than the other two, as it is the only one with negatively charged carbon atoms and positively charged hydrogens. The computed quadrupole moment is in good agreement with the quantum mechanical values for all three local minima, despite the significant differences among the corresponding site charge models.

The solutions *S*1 and *S*2 are shown in Fig. 6, with the positions of the satellite charges being listed in Table VI. The carbon atoms are connected to their satellite charges with solid lines. It is interesting to note that solution *S*1 corresponds to the methylene bisector (MB) charges proposed by Williams and Abraha [22]. We determined the optimal distance of the satellite charges from the carbon atoms to be 0.771 Å rather than 0.6 Å; this is mainly due to the fact that we consider the puckered rather than the planar conformation of cyclobutane. The important point to note is that the global optimisation approach makes it possible to locate the exact optimum position of the satellite charges without relying on chemical intuition and inaccurate trial-and-error procedures.

As a final validation of the global optimisation algorithm, we performed local minimisations starting from approximately 400 evenly spaced feasible points selected from the 100×100 grid that was used for the construction of Figs. 2–4. We found the same three distinct local minima as those previously identified by the global optimisation algorithm. We can therefore see that the algorithm has succeeded in finding not only the global minimum, but also all of the local minima, with only a small number of local searches.

APPLICATION TO OTHER MOLECULES

This section considers the construction of optimal site charge models for a set of other molecules. For some of them, the description of the electrostatic field with atomic charges was found to be adequate, while for others it was necessary to use satellite charges. In all cases, the accuracy of the computed charge magnitudes is expressed in terms of 90% confidence intervals.

Molecules with only Atomic Charges: CH_3CN and $HCONH_2$

For both of these molecules, atomic charges were found to be sufficient for the accurate description of

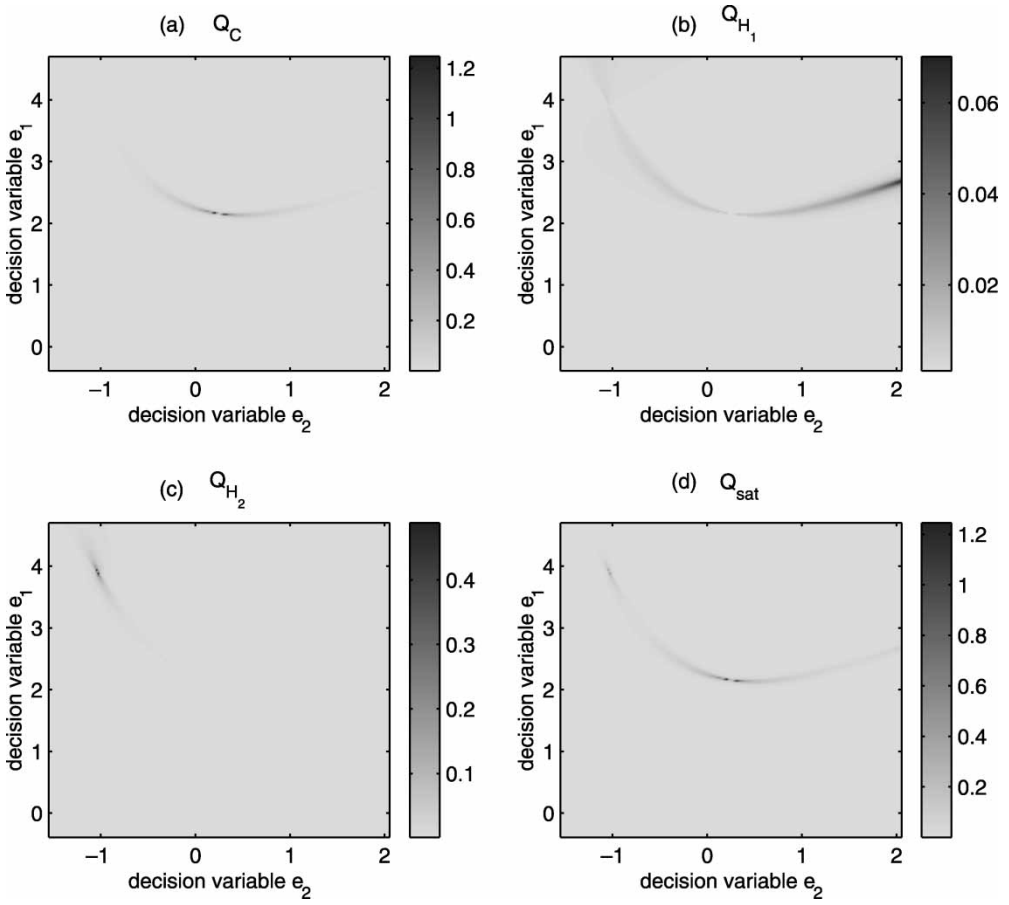


FIGURE 5 Errors in charge magnitudes for various positions of the satellite charges for puckered cyclobutane.

TABLE IV Global optimisation statistics for cyclobutane (puckered)

N	*β*	*Iterations*	*Local minimisations*	N^E	*Global minimum (% RRMS)*	*CPU (s)**
32	0.5	2	4	3	23.7	19.4

* On an Intel PIII 850 MHz processor.

TABLE V Quality of fit for the atomic site charge model of cyclobutane (puckered form); units as in Table III

	S1	*S2*	*S3*
e_1	2.9735	3.5131	1.8722
e_2	1.5580	− 0.7381	0.1539
RMS	0.3001	0.3052	0.3493
RRMS %	23.7	24.1	27.6
Condition number	6.9599×10^6	6.9840×10^6	6.9766×10^6
$Q_1 \pm \delta Q_1$ (e)	− 0.05499 ± 0.00184	0.11230 ± 0.00268	− 0.76014 ± 0.00945
$Q_2 \pm \delta Q_2$ (e)	0.16403 ± 0.00169	0.03497 ± 0.00108	0.07483 ± 0.00129
$Q_3 \pm \delta Q_3$ (e)	0.06467 ± 0.00106	0.39657 ± 0.004426	0.07864 ± 0.00134
$Q_4 \pm \delta Q_4$ (e)	− 0.17370 ± 0.00193	− 0.54384 ± 0.00614	0.60667 ± 0.00795
μ_x	-1.2285×10^{-7}	8.0900×10^{-7}	-1.6565×10^{-7}
μ_y	-1.1596×10^{-6}	2.0600×10^{-6}	-1.4276×10^{-7}
μ_z	-2.7670×10^{-7}	-9.4963×10^{-7}	-2.9146×10^{-7}
Θ_{xx}	-6.6260×10^{-1}	-6.1796×10^{-1}	-6.2133×10^{-1}
Θ_{xy}	7.6451×10^{-7}	-1.3248×10^{-6}	-4.6005×10^{-7}
Θ_{xz}	7.4020×10^{-7}	7.1209×10^{-6}	1.0597×10^{-6}
Θ_{yy}	-6.6260×10^{-1}	-6.1796×10^{-1}	-6.2133×10^{-1}
Θ_{yz}	2.4436×10^{-7}	6.7898×10^{-7}	3.4948×10^{-7}
Θ_{zz}	1.3252	1.2359	1.2427

* Units as in Table III.

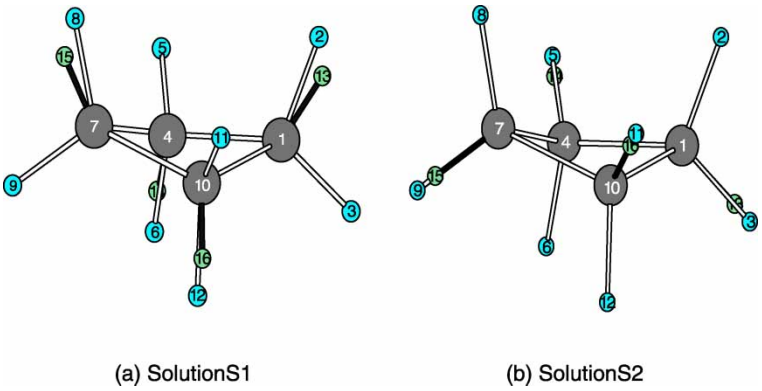


FIGURE 6 Satellite charge positions for cyclobutane. (Colour version available online.)

the electrostatic field. The results obtained are in agreement with the ones reported in the literature by Spackman [8] and are presented in Table VII. The only moment matching constraint imposed explicitly was zero total charge.

In the case of CH_3CN, the error δQ is significantly larger for the charge of the methyl carbon atom that for the other charges. This particular atom is located at the centre of a tetrahedral arrangement of atoms, which means that it is less exposed to the molecular surface and thus its charge is computed with less statistical accuracy. This accords with some observations reported in the literature regarding the methyl carbon in tetrahedral molecules, such as CH_3CN and CH_3OH. In particular, Bayly *et al.* [13] observed that the least-squares error is relatively insensitive to the charge on the methyl carbon, but highly sensitive to other charges. Moreover, Spackman [8] observed significantly higher dependence of the methyl carbon charge on the molecular orientation used in the quantum mechanical calculations.

In general, both dipole and quadrupole moments were found to be modelled accurately by the computed solutions. The exception was the computed quadrupole tensor for $HCONH_2$ which deviates significantly from the *ab initio* results. Satellite charges may need to be introduced to achieve better matching of this moment.

Molecules with Additional Satellite Charges: *n*-alkanes

The failure of atomic charge models to produce adequate representations of the electrostatic potential of *n*-alkanes has been reported in the literature [24]. Here we consider two such molecules. In both cases, the maximum allowable error for the charge magnitudes was set to 0.02, while the satellite charges were kept at distances ranging from 0.15–0.9 Å from the parent atoms.

n-*Butane*

Table VIII shows results for a conformation of butane with symmetry C_{2h}. The only exact moment matching constraint imposed is zero total charge. As can been seen, the RRMS error is very high, i.e. 69.6% (RMS = 0.767 kcal/mol). Due to symmetry, the computed dipole moment is practically zero. However, there is a large discrepancy in the quadrupole moment between the *ab initio* and atomic site charge model, which provides the motivation for the introduction of satellite charges.

Placing satellite charges on the two methyl carbons results in two independent decision variables due to the presence of a two-fold axis C_2, an inversion centre I and a symmetry plane σ_h in the point group C_{2h}. We solve this problem for different values of the tolerance ϵ^{Θ} within which

TABLE VI Position of the satellite charges for the minima S1 and S2

		S1			*S2*		
N°	*Description*	*x*	*y*	*z*	*x*	*y*	*z*
13	Sat. of C1	1.0517	1.0517	0.7779	1.2419	1.2419	− 0.3703
14	Sat. to C4	− 1.0517	1.0517	− 0.7779	− 1.2419	1.2419	0.3703
15	Sat. to C7	− 1.0517	− 1.0517	0.7779	− 1.2419	− 1.2419	− 0.3703
16	Sat. to C10	1.0517	− 1.0517	− 0.7779	1.2419	− 1.2419	0.3703

TABLE VII Optimal atomic charge models for CH_3CN and $HCONH_2$ (units as in Table III)

	CH_3CN, symmetry C_{3v}			$HCONH_2$, symmetry Cs		
RMS	0.4636			1.0640		
RRMS %	3.3			6.5		
		Q	δQ		Q	δQ
	C	− 0.493758	0.013662	H	0.009577	0.005204
	H	0.179032	0.003439	C	0.703760	0.016911
	C	0.476989	0.005183	O	− 0.589651	0.005414
	N	− 0.520326	0.002003	N	− 0.968544	0.018575
				H	0.438629	0.005950
				H	0.406230	0.006199
	Site charge model		*Ab initio*	Site charge model		*Ab initio*
μ_x	− 0.447031		− 0.444889	− 1.898095		− 1.879003
μ_y	− 0.357051		− 0.355340	2.832464		2.850584
μ_z	− 4.045652		− 4.026266	− 2.278694		− 2.258208
Θ_{xx}	1.171035		1.185201	− 1.102434		− 0.841116
Θ_{xy}	− 0.034839		− 0.035260	1.827344		1.816575
Θ_{xz}	− 0.394751		− 0.399527	1.230535		1.593397
Θ_{yy}	1.186827		1.201183	2.098676		1.492060
Θ_{yz}	− 0.315295		− 0.319108	1.824351		1.840613
Θ_{zz}	− 2.357861		− 2.386385	− 0.996242		− 0.650944

the quadrupole is matched (cf. Eq. 23). No constraints need to be imposed on the dipole moment because, due to symmetry, it is always zero. The results are summarised in Table IX. The solution behaviour with ϵ^{Θ} values down to 0.25 is the same, always locating two local minima within only five local minimisations; the global optimum is the same for all these cases, and its details are shown in Table X.

By comparing the values of the quadrupole reported in Table X with the *ab initio* values in Table VIII, it can be easily computed that the relative difference between the two is of the order of about 8%. Better matching of the quadrupole can be attained by sacrificing some of the quality of the fit, i.e. by accepting a higher value of the RRMS. In fact, the results for $\epsilon^{\Theta} = 0.05$, reported in the last row of Table IX, indicate that only a very slight deterioration in RRMS (i.e. an increase from 32.39 to 32.53%) is incurred in this context. From the computational point of view, most of the local minimisations attempted in this case could not locate a feasible point. Consequently, we do not report the corresponding CPU time or number of iterations.

The molecular conformation of butane, along with the positions of the satellite charges for the two minima are shown in Fig. 7. It is interesting to note that the methylene bisector charges proposed by Williams [24] correspond to a local minimum, but the RRMS error for the global minimum is about 6% smaller.

TABLE VIII Quality of fit for the atomic site charge model of *n*-butane (Units as in Table III)

	Atomic charge model	*ab initio QM*
RMS (kcal/mol)	0.767	–
RRMS %	69.6	–
μ_x	4.7629×10^{-6}	5.0×10^{-6}
μ_y	22998×10^{-5}	2.3×10^{-5}
μ_z	1.1180×10^{-5}	1.1×10^{-5}
Θ_{xx}	0.25492	0.22698
Θ_{xy}	0.34882	0.68310
Θ_{xz}	0.27247	− 0.13821
Θ_{yy}	− 0.56816	− 0.60719
Θ_{yz}	− 0.27216	− 0.41172
Θ_{zz}	0.31323	0.38022

n-*Pentane*

As in the case of butane, atomic charge models do not provide an adequate representation of the electrostatic field of *n*-pentane. When all three moments are matched exactly, the resulting RRMS has a very high value of 75.8% (RMS = 0.7871 kcal/mol). Consequently, we consider the use of satellite charges for the three methyl carbon atoms. This is a relatively

TABLE IX Global optimisation statistics for *n*-butane (approximate dipole, quadrupole)

ϵ^{Θ}	N	β	*Iterations*	*Local minimizations*	N^E	*Global minimum (% RRMS)*	*CPU (s)**
∞	32	0.5	2	5	2	32.39	14.9
0.50	32	0.5	2	5	2	32.39	14.3
0.25	32	0.5	2	5	2	32.39	13.7
0.05	32	0.5	–	–	–	32.53	–

* On an Intel PIII 850 MHz processor.

TABLE X Global minimum for *n*-butane (approximate dipole, quadrupole), Units as in Table III

*Site**	*x*	*y*	*z*	*Q*	*δQ*
C1	0.48184	− 1.70988	− 0.82513	− 0.50833	0.00748
C2	− 0.26488	− 0.39661	− 0.59772	− 0.63411	0.01572
C3	0.26489	0.39661	0.59772	− 0.63411	0.01572
C4	− 0.48185	1.70987	0.82514	− 0.50833	0.00748
H5	0.40350	− 2.36065	0.04155	0.09630	0.00167
H6	0.08207	− 2.24768	− 1.67943	0.10941	0.00190
H7	1.53800	− 1.53347	− 1.01007	0.09630	0.00167
H8	− 0.19730	0.21757	− 1.49394	0.07528	0.00243
H9	− 1.32338	− 0.60348	− 0.45010	0.07528	0.00243
H10	0.19731	− 0.21757	1.49394	0.07528	0.00243
H11	1.32338	0.60349	0.45100	0.07528	0.00243
H12	− 0.08268	2.24768	1.67943	0.10941	0.00190
H13	− 0.40352	2.36065	− 0.041553	0.09630	0.00167
H14	− 1.53800	1.53346	1.01009	0.09630	0.00167
(Sat. of C2) 15	0.03867	− 0.75594	− 0.55290	0.68988	0.01329
(Sat. of C3) 16	− 0.03867	0.75594	0.55290	0.68988	0.01329
RMS	0.3569				
RRMS%	32.4				
μ_x	6.55135×10^{-6}				
μ_y	4.99612×10^{-6}				
μ_z	-5.53731×10^{-6}				
Θ_{xx}	0.24676				
Θ_{xy}	0.71911				
Θ_{xz}	− 0.13070				
Θ_{yy}	− 0.65247				
Θ_{yz}	− 0.43928				
Θ_{zz}	0.40571				

* Numbering as in Fig. 7.

difficult global minimisation problem because of its dimensionality and the large number of local minima. The molecular conformation of pentane used in this study presents no symmetry (point group C_1) and thus the optimisation variables are all 9 coordinates of the satellite charges.

The globally optimal solution that matches the electrostatic field moments up to quadrupole is presented in Table XI. The corresponding computational statistics are presented in Table XII. The large number of local minima, some of which represent very poor descriptions of the electrostatic field, provide an indication of the difficulty of relying on chemical intuition alone for the exact positioning of satellite charges in such molecules.

Optimal Site Charge Models for Crystal Structure Prediction

One area in which the accuracy of the force field is of great importance is *ab initio* crystal structure prediction [38,39]. In typical cases, crystal structure prediction algorithms predict a very large number of packing motifs that differ only slightly in lattice energy. The error of the force field should be smaller than these differences for the approach to be able to

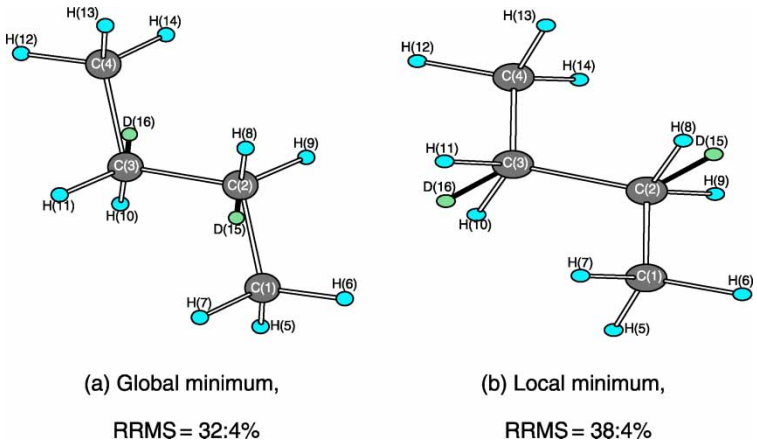


FIGURE 7 Optimal satellite charge positions for *n*-butane (approximate dipole, quadrupole). (Colour version available online.)

TABLE XI Global minimum for n-pentane (exact dipole, quadrupole), units as in Table III

Site	*X*	*Y*	*z*	*Q*	*δQ*
C1	− 0.86314	− 1.86672	0.90098	− 0.48607	0.00984
C2	− 0.40944	− 1.39659	− 0.48189	0.30337	0.00956
C3	0.52909	− 0.18499	− 0.46778	− 0.89625	0.02000
C4	− 0.10455	1.11284	0.03929	− 0.70232	0.02000
H5	− 1.46323	− 1.11840	1.40824	0.11140	0.002560
H6	− 1.46438	− 2.76759	0.82357	0.11220	0.00238
H7	− 0.01005	− 2.09310	1.53543	0.10662	0.00237
H8	0.09812	− 2.21933	− 0.98038	0.03206	0.00243
H9	− 1.28285	− 1.16718	− 1.08993	0.03228	0.00250
H10	1.40530	− 0.41697	0.13700	0.07979	0.00256
H11	0.89542	− 0.01991	− 1.47960	0.09737	0.00309
H12	− 0.42798	0.99191	1.06977	0.09728	0.00336
H13	− 1.00169	1.32380	− 0.54051	0.07612	0.00320
C14	0.84613	2.30629	− 0.04694	− 0.50921	0.01064
H15	0.37409	3.21303	0.31904	0.10721	0.00248
H16	1.74141	2.13756	0.54541	0.09901	0.00247
H17	1.15871	2.48625	− 1.07193	0.09380	0.00242
(Sat. of C2)18	− 0.65764	− 1.71131	− 1.22071	− 0.12551	0.00256
(Sat. of C3)19	0.33650	− 0.07318	− 0.30243	0.71931	0.01742
(Sat. of C4)20	0.26113	1.46539	− 0.05779	0.65156	0.01380
RMS	0.3216				
RRMS %	31.0				
$\mu_x^{\dagger}$	− 0.02669				
$\mu_y^{\dagger}$	− 0.02554				
$\mu_z^{\dagger}$	− 0.03243				
$\Theta_{xx}^{\dagger}$	0.08400				
$\Theta_{xy}^{\dagger}$	− 0.62020				
$\Theta_{xz}^{\dagger}$	0.34519				
$\Theta_{yy}^{\dagger}$	− 0.39588				
$\Theta_{yz}^{\dagger}$	0.19105				
$\Theta_{zz}^{\dagger}$	0.31188				

† *Ab initio* electrostatic moments (exact).

distinguish stable candidate polymorphs from less stable ones.

The methodology described in this paper has been tested extensively in the context of a novel *ab initio* crystal structure prediction algorithm [40] for the case of several compounds, including 3-aza-bicyclo[3.3.1]nonane-2,4-dione (molecule 6 of second blind test, see Motherwell *et al.* [38]) and allopurinol. In the case of 3-aza-bicyclo[3.3.1]nonane-2,4-dione, the use of 9 non-atomic charges (one on each carbon and nitrogen atom) reduced the RRMS error from 7.813 (RMS = 0.844 kcal/mol) to 3.063% leading to a successful prediction of the experimental crystal structure. Figure 8 compares the *ab initio* electrostatic potential with those computed by the atomic and satellite charge models; in all three cases, the potentials are projected at the surface that corresponds to 1.4 times the van der Waals radii, i.e. at the points of the maximum error. It can clearly be seen that certain regions of the surface are modelled poorly if only atomic charges are used. As expected, this has a detrimental effect in the accuracy of the intermolecular force field in the crystal structure prediction.

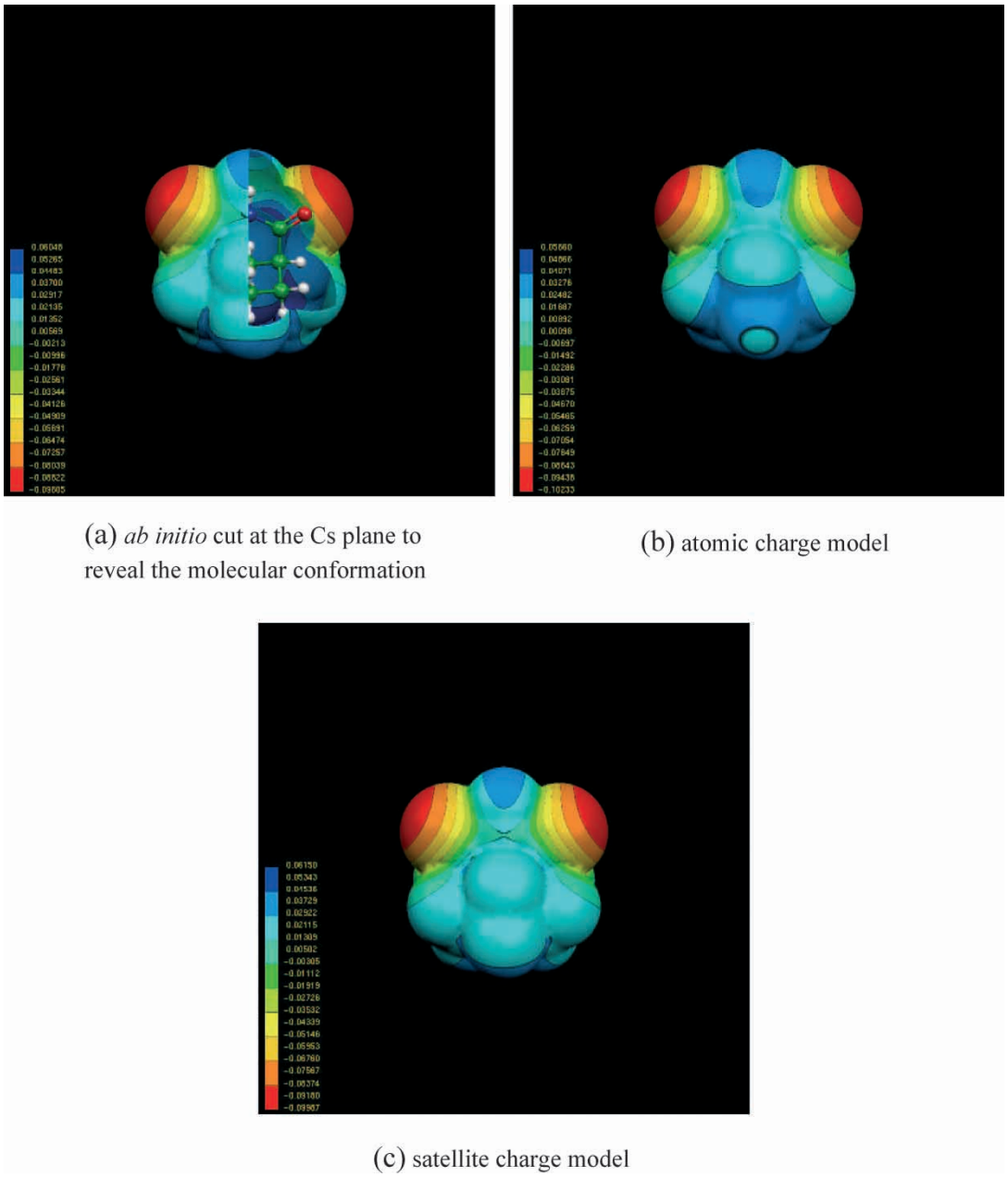
In the case of allopurinol, the error in the description of the electrostatic potential with only atomic charges was RRMS = 8.192% (RMS = 1.117 kcal/mol). The lattice energy minimisation of the experimental crystal [41] clearly showed that the use of only atomic charges is not sufficient, as it led to a break in symmetry and errors in the lattice lengths and angles that exceeded 20%. On the contrary, the use of one additional satellite charge on each nitrogen reduced the RRMS to 4.476% and the maximum error in the lattice energy parameters to 1.76%.

Concluding, the use of satellite charge models allows the accurate modelling of crystal structures without incurring the computational burden associated with more complicated force fields based on atomic multipoles. This is an important consideration for the extensive exploration of the solution space required for successful *ab initio* crystal structure prediction.

TABLE XII Global optimisation statistics for n-pentane (exact dipole, quadrupole)

N	*β*	*Iterations*	*Local minimizations*	N^E	*Global minimum (% RRMS)*	*CPU (s)**
256	0.25	209	4188	81	30.98	366.6

* On a DEC Alpha 733 MHz processor.



(a) *ab initio* cut at the Cs plane to reveal the molecular conformation

(b) atomic charge model

(c) satellite charge model

FIGURE 8 Electrostatic potential for 3-aza-bicyclo[3.3.1]nonane-2,4-dione (normalised units at 1.4 times the van der Waals surface).

CONCLUDING REMARKS

This paper has presented a comprehensive approach for the fitting of site charge models to electrostatic fields obtained using quantum mechanical calculations. Its main elements are as follows:

- Satellite charges may be associated with any subset of the atoms in the molecule. Their positions and magnitudes, as well as the magnitudes of all atomic charges, are determined using a mathematical optimisation procedure that does not rely on any *a priori* knowledge or assumption.
- The optimisation formally incorporates constraints that ensure that the accuracy of the computed charges, as measured by statistical confidence intervals, is kept below a specified bound. This prevents excessive sensitivity of these charges with respect to the problem data (e.g. the sampling points used for the quantum mechanical field).
- Additional constraints for matching the moments of the electrostatic field, either exactly or approximately, may also be included.
- A combined stochastic-deterministic optimisation technique is used to determine globally optimal solutions. Molecular symmetry is automatically

exploited, both to obtain physically realistic solutions and to produce a variable transformation that reduces the necessary computational effort.

- The quality of the derived site charge model is judged solely in terms of its ability to model the quantum mechanical electrostatic field. Chemical intuition is however, required in the choice of the chemically important atoms that will be assigned satellite charges. The presence of lone pairs or π electrons that are known to cause strong atomic dipole and quadrupole moments can provide some insight towards this decision. In any case, the suitability of any choice can be established *a posteriori* by the achieved RRMS global minimum and the corresponding confidence intervals.

The applicability of the above approach has been demonstrated using a number of different molecules. It is worth mentioning at this point that the validity of the statistical confidence intervals was tested by computing the values of the atomic charges of the orthorhombic molecular conformation of paracetamol [42] for three significantly different orientations of the molecule with respect to the cartesian frame. Optimal atomic charges were computed for each orientation, both with and without additional constraints for matching the dipole and quadrupole moments. The differences between each set of atomic charges and the corresponding mean values were always found to be within the 90% confidence intervals computed by our procedure.

Overall, the above results provide some further evidence of the ability of confidence intervals to provide an estimate of the accuracy of charge magnitudes with respect to perturbations in the data used for their estimation. The confidence intervals can be used to identify the cases where the fitting is not adequate, e.g. because of the existence of "buried" atoms whose charge magnitude has only a slight effect on the electrostatic potential outside the molecular surface. In such cases, it has to be recognised that the fitting to the electrostatic potential cannot yield a statistically significant solution and other methods that analyse the wavefunction itself, rather than the properties predicted by it, should be applied (see, for example, Swart *et al.* [43]).

This paper has not examined the "transferability" of the computed charge models between different molecules or different conformations of the same molecules. Instead, it focuses on the construction of the globally optimal site charge model for a given molecular conformation. However, preliminary studies with the orthorhombic and monoclinic form of paracetamol [42,44] show that the small effect of the crystalline environment on the molecular structure is sufficient to cause differences in the computed charge models that are larger than the corresponding confidence intervals.

In the interest of simplicity, the objective function used in this paper gives equal weighting to all parts of the electrostatic field. In large molecules, the deviation between the computed electrostatic field and the quantum mechanical one may be excessively high in certain important parts of the molecule (e.g. in the vicinity of hydrogen donors and acceptors) even if the overall RRMS appears to be acceptable. In such cases, it may be desirable to introduce different weights for the different terms in the summation of Eq. (14). This does not alter the proposed approach in any substantial manner.

Finally, as is well known, the description of the electrostatic field can be further improved by including site dipoles, quadrupoles or even higher moments in the model. In principle, the approach presented here can be extended to the optimisation of these more complex models.

Acknowledgements

The authors wish to thank Professor D. N. Theodorou of the National Technical University of Athens for useful discussions during the course of this work. The financial support of the Mitsubishi Chemical Corporation and the United Kingdom's Engineering and Physical Sciences Research Council (EPSRC) under Platform Grant GR/N08636 is gratefully acknowledged.

APPENDIX A: ANALYTICAL PARTIAL DERIVATIVES

The exact partial derivatives required by the local minimisation steps of the optimisation algorithm are computed via the chain rule. This involves multiplying the corresponding derivatives with respect to the cartesian coordinates of the satellite charges $\mathbf{r}^S$ by the derivatives of $\mathbf{r}^S$ with respect to the decision variables e_h. By virtue of Eq. (53), the latter derivatives are simply given by:

$$\frac{\partial \mathbf{r}^S}{\partial e_h} = \mathbf{x}_h \tag{A1}$$

Partial Derivatives of the Objective Function

The derivatives of the objective function (42), with respect to $\mathbf{r}^S$ are given by:

$$\frac{\partial \Phi}{\partial \mathbf{r}^S} = \sum_{k=1}^{N_p} \left[U_k^{\mathrm{QM}} - U\left(\mathbf{r}_k^{\mathrm{QM}}; \mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S\right)\right] \times \frac{\partial U}{\partial \mathbf{r}^S}\left(\mathbf{r}_k^{\mathrm{QM}}; \mathbf{Q}(\mathbf{r}^S), \mathbf{r}^S\right) \tag{A2}$$

The partial derivatives appearing on the right hand side of the above equation are given by:

$$\frac{\partial U}{\partial \mathbf{r}_{i1}}\left(\mathbf{r}_k^{\mathrm{QM}};\mathbf{Q}(\mathbf{r}^S),\mathbf{r}^S\right)$$

$$=\sum_{t=1}^{N^T}\sum_{j=0}^{N_t^S}\left\{\frac{\partial Q_{tj}}{\partial \mathbf{r}_{i1}}w_{k,tj}+Q_{t1}\frac{1}{\|\mathbf{r}_k^{\mathrm{QM}}-\mathbf{r}_{i1}\|^3}\left(\mathbf{r}_k^{\mathrm{QM}}-\mathbf{r}_{i1}\right)^T\right\},$$

$$\forall i\in\mathcal{A}_{t'},\quad \forall t'=1,\ldots,N^{TS} \tag{A3}$$

The sensitivities $(\partial\mathbf{Q}/\partial\mathbf{r}^S)$ of the estimated charges with respect to the position of the satellite charges can be found by differentiating relation 31 with respect to the position of the satellite charges $\mathbf{r}^S$ to obtain:

$$\mathbf{A}\begin{pmatrix}\frac{\partial\mathbf{Q}}{\partial\mathbf{r}^S}\\ \frac{\partial\boldsymbol{\lambda}}{\partial\mathbf{r}^S}\end{pmatrix}=\frac{\partial\mathbf{b}}{\partial\mathbf{r}^S}-\frac{\partial\mathbf{A}}{\partial\mathbf{r}^S}\begin{pmatrix}\mathbf{Q}\\ \boldsymbol{\lambda}\end{pmatrix} \tag{A4}$$

The above is a linear system in the required derivatives $(\partial\mathbf{Q}/\partial\mathbf{r}^S)$. The computational cost of its solution is minimal as the matrix $\mathbf{A}$ will have already been transformed to its LU-factorised form in the context of the solution of the linear system 31.

Partial Derivatives of the Confidence Intervals

The derivatives of the confidence interval for the charge Q_{tj} with respect to the a component of the position of the satellite charge of the atom i, r_{i1}^a are given by (cf. Eq. 38):

$$\frac{\partial\delta Q_{tj}}{\partial r_{i1}^a}=\frac{t_{1-\alpha/2,N^P-N^Q+N^C}}{2\sqrt{\hat{V}_{l(t,j),l(t,j)}}}\frac{\partial\hat{V}_{l(t,j),l(t,j)}}{\partial r_{i1}^a},$$

$$\forall t=1,\ldots,N^T,\quad \forall j=0,\ldots,N_t^S\quad \forall i\in\mathcal{A}_{t'},$$

$$\forall t'=1,\ldots,N^{TS},\quad \forall a\in\{x,y,z\} \tag{A5}$$

where l is defined in Eq. (34). The derivatives of the variance–covariance matrix (cf. Eq. 39) with respect to the satellite charge positions are:

$$\frac{\partial\hat{\mathbf{V}}}{\partial r_{i1}^a}=\frac{\partial s^2}{\partial r_{i1}^a}\mathbf{T}\mathbf{H}^{-1}\mathbf{T}^T+s^2\frac{\partial\mathbf{T}}{\partial r_{i1}^a}\mathbf{H}^{-1}\mathbf{T}^T$$

$$-s^2\mathbf{T}\mathbf{H}^{-1}\frac{\partial\mathbf{H}^{-1}}{\partial r_{i1}^a}\mathbf{H}^{-1}\mathbf{T}^T+s^2\mathbf{T}\mathbf{H}^{-1}\left(\frac{\partial\mathbf{T}}{\partial r_{i1}^a}\right)^T$$

$$\forall i\in\mathcal{A}_t,\quad \forall t=1,\ldots,N^{TS},\quad \forall a\in\{x,y,z\} \tag{A6}$$

The derivative of the estimated variance with respect to the satellite charge positions can be expressed as:

$$\frac{\partial s^2}{\partial r_{i1}^a}=\frac{(\mathbf{U}^{\mathrm{QM}}-\mathbf{U})^T}{2(N^P-N^Q+N^C)}\left(\frac{\partial\mathbf{W}}{\partial r_{i1}^a}\mathbf{Q}+\mathbf{W}\frac{\partial\mathbf{Q}}{\partial r_{i1}^a}\right) \tag{A7}$$

where $\mathbf{U}$ is the vector of the model predictions, i.e. $\mathbf{U}=\mathbf{W}\mathbf{Q}$ and $\mathbf{W}$ is the matrix of the sensitivity coefficients.

From Eq. (40), we can compute the derivative of the projection matrix $\mathbf{T}$:

$$\frac{\partial\mathbf{T}}{\partial r_{i1}^a}=\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial r_{i1}^a}\mathbf{H}^{-1}\mathbf{C}^T\boldsymbol{\Omega}^{-1}\mathbf{C}$$

$$-\mathbf{H}^{-1}\left(\frac{\partial\mathbf{C}}{\partial r_{i1}^a}\right)^T\boldsymbol{\Omega}^{-1}\mathbf{C}$$

$$+\mathbf{H}^{-1}\mathbf{C}^T\boldsymbol{\Omega}^{-1}\frac{\partial\boldsymbol{\Omega}}{\partial r_{i1}^a}\boldsymbol{\Omega}^{-1}\mathbf{C}-\mathbf{H}^{-1}\mathbf{C}^T\boldsymbol{\Omega}^{-1}\frac{\partial\mathbf{C}}{\partial r_{i1}^a}$$

$$\forall i\in\mathcal{A}_t,\quad \forall t=1,\ldots,N^{TS},\quad \forall a\in\{x,y,z\} \tag{A8}$$

where $\boldsymbol{\Omega}$ is defined as:

$$\boldsymbol{\Omega}\equiv\mathbf{C}\mathbf{H}^{-1}\mathbf{C}^T \tag{A9}$$

and

$$\frac{\partial\boldsymbol{\Omega}}{\partial r_{i1}^a}=\frac{\partial\mathbf{C}}{\partial r_{i1}^a}\mathbf{H}^{-1}\mathbf{C}^T-\mathbf{C}\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial r_{i1}^a}\mathbf{H}^{-1}\mathbf{C}^T+\mathbf{C}\mathbf{H}^{-1}\left(\frac{\partial\mathbf{C}}{\partial r_{i1}^a}\right)^T \tag{A10}$$

References

[1] Klooster, W.T. and Craven, B.M. (1992) "The electrostatic potential for the phosphodiester group determined from x-ray-diffraction", *Biopolymers* **32**, 1141.

[2] Ghermani, N.E., Bouhmaida, N. and Lecomte, C. (1993) "Modelling electrostatic potential from experimentally determined charge densities. 1. Spherical-atom approximation", *Acta Cryst.* **A49**, 781.

[3] Cox, R.S. and Williams, D.E. (1981) "Representation of the molecular electrostatic potential by a net atomic charge model", *J. Comp. Chem.* **2**, 304.

[4] Singh, U.C. and Kollman, P.A. (1984) "An approach to computing electrostatic charges for molecules", *J. Comp. Chem.* **2**, 129.

[5] Chirlian, L.E. and Francl, M.M. (1987) "Atomic charges derived from electrostatic potentials", *J. Comp. Chem.* **8**, 894.

[6] Breneman, C.M. and Wiberg, K.B. (1990) "Determining atom-centered monopoles from molecular electrostatic potentials - The need for high sampling density in formamide conformational-analysis", *J. Comp. Chem.* **11**, 361.

[7] Besler, B.H., Merz, K.M. and Kollman, P.A. (1990) "Atomic charges derived from semiempirical methods", *J. Comp. Chem.* **11**, 431.

[8] Spackman, M.A. (1996) "Potential derived charges using a geodesic point selection scheme", *J. Comp. Chem.* **17**, 1.

[9] Woods, R.J., Khalil, M., Pell, W., Moffat, S.H. and V.H.S., V.H.S. (1990) "Derivation of net atomic charges from molecular electrostatic potentials", *J. Comp. Chem.* **11**, 297.

[10] Sigfridsson, E. and Ryde, U. (1998) "Comparison of methods for deriving atomic charges from the electrostatic potential and moments", *J. Comp. Chem.* **19**, 377.

[11] Reynolds, C.A., Essex, J.W. and Richards, W.G. (1992) "Atomic charges for variable molecular conformations", *J. Am. Chem. Soc.* **114**, 9075.
[12] Francl, M.M. and Chirlian, L.E. (2000), "The pluses and minuses of mapping atomic charges to electrostatic potentials", chap 1, In: Lipkowitz, K.B. and Boyd, D.B., eds, *Reviews In Computational Chemistry* (John Wiley and Sons, New York), Vol. 14.
[13] Bayly, C.I., Cieplak, P., Cornell, W.D. and Kollman, P.A. (1993) "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges-the RESP model", *J. Phys. Chem.* **97**, 10269.
[14] Woods, R.J. and Chappelle, R. (2000) "Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates", *J. Mol. Struct–Theochem.* **527**, 149.
[15] Hinsen, K. and Roux, B. (1997) "A potential function for computer simulation studies of proton transfer in acetylacetone", *J. Comp. Chem.* **18**, 368.
[16] Lévy, B. and Enescu, M. (1998) "Theoretical study of methylene blue: a new method to determine partial atomic charges; investigation of the interaction with guanine", *J. Mol. Struct–Theochem.* **432**, 235.
[17] Ridard, J. and Lévy, B. (1998) "Effective atomic charges in alanine dipeptide", *J. Comp. Chem.* **20**, 473.
[18] Francl, M.M., Carey, C., Chirlian, L.E. and Gange, D.M. (1996) "Charges fit to electrostatic potentials.2. Can atomic charges be unambiguously fit to electrostatic potentials?", *J. Comp. Chem.* **17**, 367.
[19] Stouch, T.R. and Williams, D.E. (1993) "Conformational dependence of electrostatic potential-derived charges: Studies of the fitting procedure", *J. Comp. Chem.* **14**, 858.
[20] Willock, D.J., Price, S.L., Leslie, M. and Catlow, C.R. (1995) "The relaxation of molecular-crystal structures using a distributed multipole electrostatic model", *Am. J. Comput. Chem.* **16**, 628.
[21] Leach, A.R. (1999) *Molecular Modelling* (Pearson Education, Essex, England).
[22] Williams, D.E. and Abraha, A. (1999) "Site charge models for molecular electrostatic potentials of cycloalkanes and tetrahedrane", *J. Comp. Chem.* **20**, 579.
[23] Dixon, R.W. and Kollman, P.A. (1997) "Advancing beyond the atom-centered model in additive and nonadditive molecular mechanics", *J. Comp. Chem.* **13**, 1632.
[24] Williams, D.E. (1994) "Failure of net atomic charge models to represent the van-der-Waals envelope electric-potential of n-alkanes", *J. Comp. Chem.* **15**, 719.
[25] van Nes, G.J.H. and Vos, A. (1978) "Single-crystal structures and electron density distributions of ethane, ethylene and acetylene. I. Single-crystal X-ray structure determinations of two modifications of ethane", *Acta Cryst.* **B34**, 1947.
[26] Stone, A.J. (1996) *The Theory of Intermolecular Forces* (Clarendon Press, Oxford).
[27] Gelessus, A., Thiel, W. and Weber, W. (1995) "Multipoles and symmetry", *J. Chem. Ed.* **72**, 505.
[28] Bard, Y. (1974) *Nonlinear Parameter Estimation* (Academic Press, New York).
[29] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1984) *Numerical Recipes* (Cambridge University Press, New York).
[30] Sobol', I.M. (1967) "On the distribution of points in a cube and the approximate evaluation of integrals", *Comput. Math. Math. Phys.* **7**, 86.
[31] Kucherenko S., and Sytsko Y (2002) "Application of deterministic low-discrepancy sequences to global optimisation problems", Submitted to J. Glob. Optim.
[32] Rinnooy, K.A.H.G. and Timmer, G.T. (1987) "Stochastic global optimisation methods, part I: clustering methods", *Math. Program.* **39**, 27.
[33] Rinnooy, K.A.H.G. and Timmer, G.T. (1987) "Stochastic global optimisation methods, part II: multilevel methods", *Math. Program.* **39**, 57.
[34] Boender, C.G.E., "The generalized multinomial distribution: A Bayesian analysis and applications", Ph.D. thesis, Erasmus Universiteit Rotterdam, Centrum voor Wiskunde en Informatica (Amsterdam).
[35] Almenningen, A., Bastiansen, O. and Skancke, P.N. (1961) "Preliminary results of an electron diffraction reinvestigation of cyclobutane and cyclopentane", *Acta Chem. Scand.* **15**, 711.
[36] Williams, D.E. (1999) "Improved intermolecular force field for crystalline hydrocarbons containing four-or three-coordinated carbon", *J. Mol. Struct.* **486**, 321.
[37] Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S.J., *et al.* (1993) "General atomic and molecular electronic-structure system", *J. Comp. Chem.* **14**, 1347.
[38] Motherwell, W.D.S., Ammon, H.L., Dunitz, J.D., Dzyabchenko, A., Erk, P., Gavezzotti, A., Hoffmann, D.W.M., Leusen, F.J.J., Lommerse, J.P.M., Mooij, W.T.M., *et al.* (2002) "Crystal structure prediction of small organic molecules: a second blind test", *Acta Cryst.* **B58**, 647.
[39] Lommerse, J.P.M., Motherwall, W.D.S., Ammon, H.L., Dunitz, J.D., Gavezzotti, A., Hoffmann, D.W.M., Leusen, F.J.J., Mooij, W.T.M., Price, S.L., Schweizer, B., *et al.* (2000) "A test of crystal structure prediction of small organic molecules", *Acta Cryst.* **B56**, 697.
[40] Karamertzanis, P.G. (2004) Crystal Structure Prediction of Molecular Solids, PhD Thesis (University of London).
[41] Prusiner, P. and Sundaralingam, M. (1972) "Stereochemistry of nucleic acids and their constituents. XXIX. Crystal and molecular structure of allopurinol, a potent inhibitor of xanthine oxidase", *Acta Cryst.* **B28**, 2148.
[42] Haisa, M., Kashino, S. and Maeda, H. (1974) "The orthorhombic form of p-hydroxyacetanilide", *Acta Cryst.* **B30**, 2510.
[43] Swart, M., Duijnen, P.T.V. and Snijders, J.G. (2001) "A charge analysis derived from an atomic multipole expansion", *J. Comp. Chem.* **22**, 79.
[44] Haisa, M., Kashino, S. and Maeda, H. (1976) "The monoclinic form of p-hydroxyacetanilide", *Acta Cryst.* **B32**, 1283.