Molecular forces, geometries, and frequencies by systematic molecular fragmentation including embedded charges

Michael A. Collins

Citation: The Journal of Chemical Physics **141**, 094108 (2014); doi: 10.1063/1.4894185
View online: http://dx.doi.org/10.1063/1.4894185
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/141/9?ver=pdfcov
Published by the AIP Publishing

Molecular forces, geometries, and frequencies by systematic molecular fragmentation including embedded charges

Michael A. Collins
Research School of Chemistry, Australian National University, Canberra, ACT, Australia

(Received 3 April 2014; accepted 18 August 2014; published online 5 September 2014)

The accuracy of energies, energy gradients, and hessians evaluated by systematic molecular fragmentation is examined for a wide range of neutral molecules, zwitterions, and ions. A protocol is established that may employ embedded charges in conjunction with fragmentation to provide accurate evaluation of minimum energy geometries and vibrational frequencies in an automated procedure.
 [http://dx.doi.org/10.1063/1.4894185]

I. INTRODUCTION

A major purpose of *ab initio* quantum chemistry is to calculate the total electronic energy of molecules so that chemical reactivity and other properties can be predicted. There are a hierarchy of quantum chemistry techniques which provide more and more reliable energy estimates at higher and higher computational cost. Unfortunately, the computational time required for the most reliable calculations increases rapidly with the size of the molecule. If N_{basis} is the number of basis functions, MP2 scales as ${N_{basis}}^5$ and CCSD(T) scales as ${N_{basis}}^7$, for example.

These high scaling factors limit the size of the molecules studied. So in recent years, algorithms have been developed that reduce this "scaling problem," ideally to linearity. Based on the idea that chemical functionality is a local phenomenon, one approach breaks the molecule into fragments, performs calculations on the fragments, then reconstitutes the value of the energy or property from the corresponding values for the fragments.[1] Molecular orbitals,[2–5] density matrices,[6,7] and total electronic energies have all been calculated in this way.

When only the total electronic energy is required, a number of methods are currently employed, including QM/MM schemes,[8–10] the effective fragment potential method,[11,12] the X-Pol method,[13–16] the fragment molecular orbital method,[1,17,18] and energy-based fragmentation methods. The later involves breaking the molecule into fragments, evaluating the energy of each fragment, and recombining the fragment energies to estimate the total electronic energy. Several groups have developed such approximations to the energy, and other molecular properties, in recent times.[19–40] The computational time for these fragmentation methods scales only linearly with the size of the molecule. The variation in computation time between different methods is mostly determined by the different sizes of the fragments.

Systematic molecular fragmentation by annihilation (SMFA),[24–26,41] as the name suggests, provides a systematic hierarchy of approximations to the molecular electronic energy, called "Levels." The molecular fragments at each Level are determined by the chemical bonding. Level 1 fragmentation accounts for the energy due to the interaction of functional groups with their α substituent groups, Level 2 accounts for α and β substituents, and so on. That is, the fragmentation depends on the bonded connectivity between functional groups, rather than the spatial distance between groups. The interactions between functional groups that are distant in terms of bonded connectivity but not too far separated in space, so-called non-bonded interactions, are accounted for separately. Perturbation theory is used to describe long range non-bonded interactions. This systematic method has the advantage that convergence of the energy estimate with increasing Level of fragmentation can be reasonably taken as convergence to the correct value.

Previous papers have demonstrated the convergence of the energy estimate by SMFA for a set of 96 organic molecules.[25,26,41] However, this set contained few molecules with extensive hydrogen bonding and no molecules with formal charges. The presence of formally charged functional groups, e.g., $–COO^-$ and $–NH_3^+$ groups, or highly polar groups, introduces a significant many-body induction contribution to the molecular electronic energy. Gao and co-workers,[13–16] Dahlke and Truhlar[42] and Li and co-workers[34] accounted for this effect in the context of cluster expansions and fragmentation methods, by performing the *ab initio* calculations of molecular fragments in the presence of "embedded charges" that represented the rest of the molecule. The purpose of this paper is to present a modification of the established SMFA procedure, involving the use of embedded charges, to account for the presence of charged or highly polar functional groups. The accuracy of the modified procedure is demonstrated using a test set of molecules that includes a much wider range of functional groups than has been considered previously with SMFA, including aromatic groups, sugars, proteins, and organic ions. In particular, the use of embedded charges and the use of perturbation theory for long range non-bonded interactions introduce additional approximations for the energy gradients and hessians. Recently, Gao and Wang[43] introduced a variational procedure into the X-Pol method to account for the mutual polarisation of molecular fragments and to facilitate the evaluation of accurate energy gradients in the X-Pol method. Here, and in related implementations of embedded charges,[34] mutual polarisation is accounted for in an iterative procedure. The accuracy of the

consequently approximate gradients is investigated in terms of the accuracy of optimised geometries. The accuracy of the approximate hessians is investigated in terms of the accuracy of the harmonic vibrational frequencies. An early investigation of the accuracy of optimised geometries and vibrational frequencies in systematic molecular fragmentation only considered very simple molecules,[24] and it is essential[40] to establish the general accuracy of energy gradients and hessians in an approximate method such as SMFA.

The paper is set out as follows. Section II describes the application of SMFA to general molecules, including those containing formal charges. Section III presents the results of numerical tests of accuracy and convergence, while Sec. IV contains some concluding remarks.

II. ENERGY AND ENERGY DERIVATIVES

The SMFA approximation to the molecular energy and other properties has been presented in detail previously.[24–26,41] Here, we simply summarize previous results and provide details for the modifications associated with the calculation of energy gradients and hessians, including the implementation of embedded charges.

In the SMFA approximation, the total molecular electronic energy is given by

$$E = E_b + E_{nb}, \tag{2.1}$$

where E_b denotes the "bonded energy" and E_{nb} denotes the "non-bonded energy."

A. Bonded interactions

A molecule is viewed as a set of functional groups, $\{G_i, i = 1, \ldots, N_G\}$, which are connected by single bonds (see Appendix A). In SMFA, a molecule is decomposed into fragments by removing functional groups, in an automated sequence of steps that preserve the bonding environment of each group to some extent. The method has a systematic set of "Levels" which determine the proximity of eliminated groups, so that with increasing Level, a more extensive bonding environment is retained. The method is automated and applicable to any molecular structure, but a chain-like structure provides an illustrative example. For a chain of five groups, $G_1G_2G_3G_4G_5$, the molecule is decomposed as follows:

$$\begin{aligned} G_1G_2G_3G_4G_5 &\rightarrow G_1G_2 + G_2G_3 + G_3G_4 + G_4G_5 - G_2 - G_3 - G_4, && \text{(Level 1)} \\ &\rightarrow G_1G_2G_3 + G_2G_3G_4 + G_3G_4G_5 - G_2G_3 - G_3G_4, && \text{(Level 2)} \\ &\rightarrow G_1G_2G_3G_4 + G_2G_3G_4G_5 - G_2G_3G_4, && \text{(Level 3)} \end{aligned}$$

and so on for higher Levels. At Level 1, the interaction of each group with its α substituents is included in the fragments. At Level 2, β substituents are included and so on. For any general molecule, we can write the fragmentation of molecule M as

$$M \rightarrow \sum_{n=1}^{N_{frag}} f_n F_n, \tag{2.2}$$

where the F_n represent "overlapping" fragments of the molecule and the f_n are integers. When groups are eliminated in the fragmentation procedure, the remaining groups have unsatisfied valency. The normal valency of each atom is restored by appending hydrogen atoms along the original bond direction, as previously described.[44] These are referred to as "hydrogen caps." Explicitly, the Cartesian coordinates of the hydrogen cap is given by $\mathbf{X}$(H),

$$\mathbf{X}(H) = \mathbf{X}(j) + \frac{r(j) + r(H)}{r(j) + r(m)}[\mathbf{X}(m) - \mathbf{X}(j)], \tag{2.3}$$

where r denotes a standard covalent radius for the element associated with each nucleus; nucleus j is one of the nuclei explicitly contained in fragment n and nucleus m is not contained in fragment n (the $j \ldots . m$ bond was broken in forming the fragment). Previous calculations have indicated that Eq. (2.3) produces chemically sensible bond lengths for the capping $j \ldots .$H bond. The positions of the atoms in each fragment are exactly the corresponding positions in the whole molecule. Hence, a cap defined by Eq. (2.3) has the same position in every fragment in which it occurs.

The "bonded energy" is given by the sum of the Level L fragment energies,

$$E_b = \sum_{n=1}^{N_{frag}} f_n E[F_n; \{\mathbf{Z}(k), Q(k)\}]. \tag{2.4}$$

The term $E[F_n; \{\mathbf{Z}(k), Q(k)\}]$ denotes the electronic energy of fragment F_n, which is evaluated in the presence of a set of charges $Q(k)$ at positions $\mathbf{Z}(k)$, so-called embedded charges.[13–16,34,42] The point charges are taken to represent the charge distribution of functional groups that are not contained in fragment F_n. These point charges are evaluated using a Natural Population Analysis, and the details are presented in Appendix B. The normal (default) approach is that charges are only employed to represent groups that contain formal charges, for example, $RCOO^-$ and RNH_3^+ groups. However, for systems that contain a high density of polar groups, e.g., water clusters or some proteins, highly polar groups may also be represented by a set of charges. The point charges are included in all fragment energy calculations, $E[F_n; \{\mathbf{Z}(k), Q(k)\}]$, *unless* that group is actually contained in the fragment F_n.

The inclusion of point charges in the fragment energy calculations accounts for two significant contributions to the total energy:

(i) the electrostatic interaction of groups that do not share any common fragment F_n; and
(ii) the induction energy produced by the net electric field due to formally charged (or highly polar) groups acting on the rest of the molecule.

It is worth noting that the fragmentation expression of Eq. (2.2) is balanced, in that each group occurs exactly once on the left hand side, and a *net* once on the right hand side (rhs) [see the examples in Eq. (2.1)]. So, counting the interactions of point charges with all groups on the rhs of Eq. (2.4) shows that the point charges representing a some group, A, interact a net once with each other group, B, that does not have any common fragment with A, and (correctly) a net zero times with groups that do share a common fragment with A. As has been noted previously,[34,45] this means that two groups which are represented by point charges and do not share a common fragment interact twice; group A with the charges representing group B and group B with the charges representing group A. This double counting must be corrected (see below). It is also worth noting that since all charges are included in every fragment energy calculation (either within F_n or as a set of point charges), then the many-body induction effect is accounted for, at least approximately. The many-body nature of induction, which depends on the net electric field, is particularly significant for molecules that contain formal charges; which is why embedded charges are implemented herein.

1. Energy gradients

The energy associated with chemical bonding in Eq. (2.4) is a sum of energies of individual molecular fragments,. The Hamiltonian for a molecular fragment, F_n, in the presence of point charges can be written as $H(F_n)$,

$$H(F_n) = H_M(F_n) + \sum_{i=1}^{N_e}\sum_{k=1}^{N_c}\frac{-Q(k)}{\|\mathbf{x}(i)-\mathbf{Z}(k)\|} + \sum_{j=1}^{N_{nuc}}\sum_{k=1}^{N_c}\frac{Q(k)q(j)}{\|\mathbf{X}(j)-\mathbf{Z}(k)\|}, \tag{2.5}$$

where H_M denotes the usual molecular Hamiltonian; $\mathbf{x}(i)$, $i = 1, \ldots, N_e$, denotes the Cartesian coordinate vectors of the electrons; $\mathbf{X}(j)$, $j = 1, \ldots, N_{nuc}$, denotes the Cartesian coordinate vectors of the fragment nuclei, and $q(j)$ is the corresponding nuclear charge; $\mathbf{Z}(k)$, $k = 1, \ldots, N_c$, denotes the Cartesian coordinate vectors of the point charges, and $Q(k)$ is the corresponding charge at that position. Here, we have neglected the mutual Coulomb interaction of the point charges.

We assume that an *ab initio* quantum chemistry program produces an approximate solution for the ground electronic state, subject to this Hamiltonian, which we denote as Ψ_0. The total energy, $E(F_n)$, is given by the expectation value

$$E(F_n) = E[F_n; \{\mathbf{Z}(k), Q(k)\}] = \langle\Psi_0|H(F_n)|\Psi_0\rangle = E_M(F_n) + \sum_{k=1}^{N_c} Q(k)\Phi[\mathbf{Z}(k)], \tag{2.6}$$

where E_M denotes the expectation value of H_M and the second term gives the interaction energy of each charge, $Q(k)$, with the electric potential of the molecule at the position of each point charge, $\Phi[\mathbf{Z}(k)]$. Clearly, the total electronic energy depends on both the positions of the nuclei and the positions of the point charges.

The derivative of this energy with respect to the nuclear positions of atoms in the fragment, $\frac{\partial E(F_n)}{\partial X_\alpha(j)}$, can be obtained from most standard quantum chemistry programs, either by finite differences over nuclear positions or from *so-called* analytic derivatives for some levels of *ab initio* theory. The derivatives of each fragment energy with respect to nuclei in the original molecule have two components: The explicit energy derivative associated with nucleus j in fragment n and a contribution from hydrogen caps.

Suppose that nucleus j contained in fragment F_n is labelled nucleus i in the original molecule and has a capping hydrogen, atom h in F_n, attached to it in place of nucleus m in the original molecule. Then, from Eq. (2.3), there are contributions to the total energy gradient with respect to atoms i and m, given, respectively, by

$$f_n\left\{\frac{\partial E(F_n)}{\partial X_\alpha(j)} + \frac{\partial E(F_n)}{\partial X_\alpha(h)}\left[1 - \frac{r(j)+r(H)}{r(j)+r(m)}\right]\right\}, \quad f_n\left\{\frac{\partial E(F_n)}{\partial X_\alpha(h)}\frac{r(j)+r(H)}{r(j)+r(m)}\right\}. \tag{2.7}$$

The derivatives of the bonding energy are given from Eq. (2.4) as a sum of fragment energy derivatives, including the contributions from the caps,

$$\frac{\partial E_b}{\partial X_\alpha(i)} = \sum_{n=1}^{N_{frag}} f_n \frac{\partial E(F_n)}{\partial X_\alpha(i)}. \tag{2.8}$$

Although Eq. (2.3) is a plausible proposal for locating hydrogen caps and appears to produce sensible associated bond lengths, the consequent contributions to the atomic energy gradients in Eq. (2.7) are nonetheless based on guesswork. The hydrogen caps "cancel" in the fragmentation formula, Eq. (2.2), and as a consequence, "cap contributions" to the gradients appear in Eq. (2.8) with coefficients f_n that sum to zero. The bonding environment of each cap is similar in all fragments in which it occurs. Indeed, as the value of Level increases, the bonding environment of a given cap in fragments with cancelling signs becomes more and more similar, as does the charge distribution in the neighbourhood of the cap. Hence, the energy gradients for a cap are similar in each fragment and, when summed with the cancelling f_n coefficients, result in only small contributions to the net energy gradients for the atomic nuclei in the original molecule. Nonetheless, the capping procedure ensures that the energy gradients cannot be given exactly by SMFA.

Thus far we have partly neglected the role of the point charges in the gradient of the fragment energy. The derivatives of the fragment energy with respect to the positions of any point charges are generally not readily available from *ab initio* quantum chemistry program packages, except by finite difference. To approximate these gradients, we note that *if* Ψ_0 were exact, then the Hellman-Feynman theorem would hold, and

$$\begin{aligned}\frac{\partial E(F_n)}{\partial Z_\alpha(m)} &= \langle\Psi_0|\frac{\partial H(F_n)}{\partial Z_\alpha(m)}|\Psi_0\rangle \\ &= \langle\Psi_0|\frac{\partial}{\partial Z_\alpha(m)}\sum_{n=1}^{N_e}\sum_{k=1}^{N_c}\frac{-Q(k)}{\|\mathbf{x}(n)-\mathbf{Z}(k)\|} \\ &\quad+\sum_{j=1}^{N_{nuc}}\sum_{k=1}^{N_c}\frac{Q(k)q(j)}{\|\mathbf{X}(j)-\mathbf{Z}(k)\|}|\Psi_0\rangle \\ &= \frac{\partial}{\partial Z_\alpha(m)}\sum_{k=1}^{N_c}Q(k)\Phi^{(n)}[\mathbf{Z}(k)] \\ &= Q(m)\frac{\partial\Phi^{(n)}[\mathbf{Z}(m)]}{\partial Z_\alpha(m)} \\ &= -Q(m)\Im_\alpha^{(n)}[\mathbf{Z}(m)], \end{aligned} \tag{2.9}$$

where $\Im_\alpha^{(n)}[\mathbf{Z}(m)]$ is the electric field due to the electrons and nuclei in fragment n, evaluated at the position of the point charge. The electric field at these positions is readily supplied by standard quantum chemistry programs, so Eq. (2.9) is readily evaluated at negligible computational cost. Although the Hellman-Feynman theorem is *not* exact for an approximate electronic wavefunction, nonetheless, test calculations indicate that Eq. (2.9) is quite accurate in the following sense:

$$\sum_{j=1}^{N_{nuc}}\frac{\partial E(F_n)}{\partial X_\alpha(j)}+\sum_{k=1}^{N_c}\frac{\partial E(F_n)}{\partial Z_\alpha(k)}\approx 0,\ \text{for}\,\alpha=1,2,3. \tag{2.10}$$

That is, the sum of the computed forces on the point charges is equal (typically to about 10^{-6} a.u.) and opposite to the sum of the forces on the molecular nuclei.

Now, the positions of these point charges are determined by the positions of atoms in the original molecule,

$$\mathbf{Z}(m)=\sum_{j=1}^{N_{nuc}}c_{mj}\mathbf{X}(j), \tag{2.11}$$

and

$$\frac{\partial Z_\alpha(m)}{\partial X_\alpha(j)}=c_{mj}. \tag{2.12}$$

As discussed in Appendix B for the current protocol, c_{mj} is one or zero, although more complicated schemes have been considered.[46] Combining Eqs. (2.9) and (2.12), by the product rule, gives a contribution to the energy gradient with respect to the coordinates of atoms in groups represented by point charges, associated with the force on the point charges from the *ab initio* fragment charge distributions,

$$\frac{\partial E(F_n)}{\partial X_\alpha(j)}\sim -Q(m)\Im_\alpha^{(n)}[\mathbf{Z}(m)]c_{mj}. \tag{2.13}$$

Such contributions to the total energy gradient rely on merely plausible assumptions about the magnitude of the point charges and their association with the positions of atoms in the molecule, as well as on the applicability of the Hellman-Feynman theorem. Clearly, the use of Eq. (2.13) introduces an error in the estimate of the atomic energy gradients. The magnitude of this error is examined below.

2. Second derivatives of the energy

Similarly, the second derivatives of the energy with respect to the positions of the point charges are approximated by (neglecting contributions from first-order perturbations of the electronic wavefunction)

$$\frac{\partial^2 E(F_n)}{\partial Z_\alpha(m)\partial Z_\beta(m)}=-Q(m)\Im_{\alpha\beta}^{(n)}[\mathbf{Z}(m)], \tag{2.14}$$

where $\Im_{\alpha\beta}^{(n)}[\mathbf{Z}(m)]$ denotes the electric field gradient of fragment n at $\mathbf{Z}(m)$. Again, using Eq. (2.12) and the product rule, Eq. (2.14) leads to contributions to the hessian for atoms represented by point charges. "Off diagonal" second derivatives of fragment energies with respect to the coordinates of atoms in the fragment and atoms represented by charges cannot be easily evaluated and have been neglected.

The major contributions to the hessian for the bonding energy are expected to be given by the corresponding formula for the first derivatives, Eq. (2.8),

$$\frac{\partial^2 E_b}{\partial X_\alpha(j)\partial X_\beta(i)}=\sum_{n=1}^{N_{frag}}f_n\frac{\partial^2 E(F_n)}{\partial X_\alpha(j)\partial X_\beta(i)}, \tag{2.15}$$

where the contributions arising from hydrogen caps are taken into account, using Eq. (2.7).

B. Non-bonded interactions

The non-bonded energy, E_{nb}, contains the interactions between groups in the molecule that are not accounted for in the bonded energy, E_b; that is, interactions between groups that are never contained in a common fragment. These interactions are evaluated from the interaction of one fragment of the molecule with another fragment, as previously described.[24–26]

Briefly, if we fragment the molecule at Levels L_1 and L_2,

$$\begin{aligned}M &\to \sum_{n_1=1}^{N_{frag}^{L_1}} f_{n_1}^{(L_1)}F_{n_1}^{(L_1)}, \\ M &\to \sum_{n_2=1}^{N_{frag}^{L_2}} f_{n_2}^{(L_2)}F_{n_2}^{(L_2)},\end{aligned} \tag{2.16}$$

then the non-bonded energy is given by

$$E_{nb}=\frac{1}{2}\sum_{n_1=1}^{N_{frag}^{L_1}}\sum_{n_2=1}^{N_{frag}^{L_2}}f_{n_1}^{(L_1)}f_{n_2}^{(L_2)}E\left[F_{n_1}^{(L_1)}\leftrightarrow F_{n_2}^{(L_2)}\right]_{allowed}, \tag{2.17}$$

where $E[F_{n_1}^{(L_1)}\leftrightarrow F_{n_2}^{(L_2)}]_{allowed}$ denotes the energy of interaction between two fragments, which is *allowed*. Here, allowed

means that the interaction has not already been accounted for in the bonded energy, E_b. The way in which non-bonded fragment interactions are modified to account for the bonded interactions was discussed in detail in Ref. 26. In principle, the description of these nonbonded interactions can be systematically improved by increasing the Levels L_1 and L_2. In practice, it has been found to be sufficient to take $L_1 = L_2 = 1$.

The non-bonded energy of Eq. (2.17) is actually evaluated in a way that can be written as a sum of distinct terms,

$$E_{nb} = E_{ab} + E_{ele} + E_{ind} + E_{disp}. \tag{2.18}$$

Assuming that an interaction, $F_{n_1}^{(L_1)} \leftrightarrow F_{n_2}^{(L_2)}$ is allowed, then the interaction energy can be evaluated by *ab initio* quantum chemistry calculation,

$$\begin{aligned} E_{ab}\left[F_{n_1}^{(L_1)} \leftrightarrow F_{n_2}^{(L_2)}\right] = E\left[F_{n_1}^{(L_1)} + F_{n_2}^{(L_2)}; \{\mathbf{Z}(k), Q(k)\}\right] \\ - E\left[F_{n_1}^{(L_1)}; \{\mathbf{Z}(k), Q(k)\}\right] \\ - E\left[F_{n_2}^{(L_2)}; \{\mathbf{Z}(k), Q(k)\}\right], \end{aligned} \tag{2.19}$$

as the difference in energy between the combined and separate fragments, calculated in the presence of point charges which represent groups that are not contained in the fragments listed to the left of the ";" in each term in Eq. (2.19). These *ab initio* calculations are only required if the fragments $F_{n_1}^{(L_1)}$ and $F_{n_2}^{(L_2)}$ are separated by a relatively short distance. In practice, all the atom-atom distances between the two fragments are calculated and compared to the sum of the Van der Waals radii for each pair of atoms. An *ab initio* evaluation of this interaction is only performed if the ratio of the atom-atom distance to the sum of the radii is less than a "cut-off" value, denoted d_{tol}, for at least one pair of atoms. The interaction energies evaluated by Eq. (2.19) are summed with coefficients $\frac{1}{2} f_{n_1}^{(L_1)} f_{n_2}^{(L_2)}$ to give E_{ab} in Eq. (2.18).

The contributions to the energy gradient and hessian arising from Eq. (2.19) are evaluated in the same manner as for the *ab initio* energies that contribute to the bonded energy, including the contributions from hydrogen caps and the forces on any point charges. The capping hydrogen atoms formally cancel in the L_1 and L_2 fragmentations, so that the net forces on caps in the non-bonded energy is expected to be small. However, the editing procedure that ensures only "allowed" terms occur in Eq. (2.17) leads to some residual non-vanishing forces on some caps. Hence, the presence of some capping hydrogens in fragments which interact "through space" contributes an additional approximation to the estimation of energy gradients and hessians in SMFA.

For fragments that are separated by larger distances, the interaction can be accurately evaluated using perturbation theory (see below). This distance based criterion is important, because if all non-bonded interactions were evaluated *ab initio*, as in Eq. (2.18), then the computer time for the whole calculation would contain a contribution that scaled as the square of the number of fragments, and hence as the square of the number of atoms in the molecule. Using a distance based criterion for Eq. (2.18) ensures that the computer time for the whole calculation is linear in the number of atoms in the molecule.

When $F_{n_1}^{(L_1)}$ and $F_{n_2}^{(L_2)}$ are separated by more than the cut-off value, the interaction is evaluated using perturbation theory. In first order perturbation theory, the interaction of two fragments is electrostatic. As previously reported,[26] the charge distribution of each fragment is described by a set of distributed electric multipoles (up to the hexadecapole) centred on each atom in the fragment, using Stone's GDMA program.[47] The computational effort required is little more than an evaluation of the energy (and electron density) for each of the (small) Level 1 fragments. The electron density of these Level 1 fragments is calculated in the presence of the point charges (corresponding to all the groups represented by charges, which are not contained in the relevant Level 1 fragment), to account for the mutual polarisation of the fragments.[48] The interaction energy (to first order) is then a sum of multipole-multipole interactions between the fragments

$$\begin{aligned} E\left[F_{n_1}^{(L_1)} \leftrightarrow F_{n_2}^{(L_2)}\right] &= E_{ele}\left[F_{n_1}^{(L_1)} \leftrightarrow F_{n_2}^{(L_2)}\right] \\ &= \sum_{a \in F_{n_1}} \sum_{b \in F_{n_2}} T^{ab} q^a q^b + T_\alpha^{ab} \\ &\quad \times \left(q^a \mu_\alpha^b - \mu_\alpha^a q^b\right) + \ldots\ldots, \end{aligned} \tag{2.20}$$

which is written explicitly in Eq. (26) of Ref. 26. Equation (2.20) employs the tensor convention, wherein repeated subscripts are summed. Here, Eq. (2.20) is intended to indicate that the electrostatic interaction is evaluated as a sum over all the atoms in both fragments of the products of Cartesian multipoles (charges, q^a; dipoles, μ_α^a; etc) with Cartesian tensors, T, of appropriate rank [see page 37 of Ref. 49]. The interaction energies evaluated by Eq. (2.20) are summed with coefficients $\frac{1}{2} f_{n_1}^{(L_1)} f_{n_2}^{(L_2)}$ to give E_{ele} in Eq. (2.18).

Previous studies[26,28] have demonstrated that Eq. (2.20) provides a very accurate description of the long range electrostatic interaction energy between molecular fragments. However, it is not easy to obtain complete derivatives of Eq. (2.20) with respect to the nuclear coordinates. A variation of a nuclear coordinate results in a variation of the Cartesian tensors, which is easily evaluated, and a variation in the Cartesian multipoles, which is not easily evaluated. It would be possible to evaluate Cartesian gradients of the distributed multipole moments by finite difference, but this would be computationally expensive, and is not pursued herein. Here, we approximate the derivatives of the electrostatic energy by

$$\begin{aligned} \frac{\partial E_{ele}\left[F_{n_1}^{(L_1)} \leftrightarrow F_{n_2}^{(L_2)}\right]}{\partial X_\alpha(j)} &= \sum_{a \in F_{n_1}} \sum_{b \in F_{n_2}} q^a q^b \frac{\partial T^{ab}}{\partial X_\alpha(j)} \\ &\quad + \left(q^a \mu_\alpha^b - \mu_\alpha^a q^b\right) \frac{\partial T_\alpha^{ab}}{\partial X_\alpha(j)} + \ldots\ldots \end{aligned} \tag{2.21}$$

The accuracy of this approximation is examined below. The second derivatives of the electrostatic interactions are given by the corresponding form of Eq. (2.21) involving second derivatives of the tensors, only.

In second order perturbation theory, these non-bonded fragment interactions have a dispersion interaction, E_{disp}, and an induction energy, E_{ind}. The dispersion energy is calculated as a sum over the dispersion energy for each pair of groups [unless the groups share a common fragment in Eq. (2.2)]. The details of these dispersion interactions have been presented previously.[26] To summarise,

$$E_{disp} = \sum_{i=1}^{N_G-1} \sum_{\substack{j=i+1 \\ i \leftrightarrow j \ allowed}}^{N_G} U_{disp}^{ij},$$
$$U_{disp}^{ij} = -T_{\alpha\beta}^{ij} T_{\gamma\delta}^{ij} \alpha_{\alpha\gamma}^{i} \alpha_{\beta\delta}^{j} \frac{2P_{ii}P_{jj}}{(P_{ii}+P_{jj})}, \tag{2.22}$$

where i and j denote groups, α^i denotes the zero frequency polarizability of group i, P_{ii} is calculated from the relative magnitude of the imaginary frequency polarisability of group i, and T^{ij} is a Cartesian tensor evaluated from the Cartesian vector connecting the centroids of the two groups.

The exact gradient and hessian of E_{disp} requires evaluation of the Cartesian derivatives of the polarizability and imaginary frequency polarisability of the groups. These are not readily available. Moreover, the T^{ij} are defined in terms of the group centroids, which arbitrarily relates the atomic positions to E_{disp}. Hence, exact evaluation of the nuclear derivatives of E_{disp} is not feasible. Atomic gradients and second derivatives of E_{disp} are estimated from the corresponding first and second derivatives of the T^{ij} tensors alone. Test calculations reported below show that these first and second derivatives are very small. In the case that an estimate of the Hartree-Fock energy of a molecule is required, the dispersion energy is not evaluated.

The static dipole polarizability, α^i, for each group is also used in the calculation of the induction energy. In the case of systems containing formal charges, the induction energy is contained within the bonded energy, E_b, due to the inclusion of embedded charges. Such charges dominate the polarisation of the fragments and, from the results presented below, appear to account satisfactorily for the total induction energy. In most molecules that do not contain formal charges, the induction energy is small, often below 1 mE_h. However, in some systems containing a large number of polar groups the induction energy can be more significant. Hence, for systems without embedded charges, the induction energy is evaluated as follows.

(i) The charge distribution of the Level 1 fragments has been evaluated, in terms of distributed multipoles, in the evaluation of the electrostatic energy.

(ii) The electric field at the centre, $\mathbf{X}(k)$, of each group, $k = 1, \ldots, N_{groups}$, can therefore be readily evaluated as

$$\Im_\gamma[\mathbf{X}(k)] = \sum_{\substack{n=1 \\ allowed}}^{N_{frag}^{L_1}} f_n \sum_{i \in F_n} \Im_\gamma^i[\mathbf{X}(k)], \quad \gamma = 1, 2, 3, \tag{2.23}$$

where the first sum is over all allowed Level 1 fragments, F_n, with coefficients f_n. As usual herein, *allowed* means that the groups in F_n do not share any common fragment with group k in Eq. (2.2). The second sum in Eq. (2.23) is over all atoms in the Level 1 fragment F_n which have distributed multipoles that produce an electric field, $\Im_\gamma^i[\mathbf{X}(k)]$, at the group centre, $\mathbf{X}(k)$,

$$\Im_\gamma^i[\mathbf{X}(k)] = T_\gamma^{ki} q^i - T_{\gamma\varepsilon}^{ki} \mu_\varepsilon^i. \tag{2.24}$$

In the results presented below, only distributed charges and dipoles have been employed at each atomic position.

(iii) The field at each group centre induces a dipole in proportion to the static dipole polarizability tensor, $\boldsymbol{\alpha}(k)$, for each group. These induced dipoles in turn polarise other groups. However, for systems without formal charges, iteration of the field calculation to include the fields due to induced dipoles is generally not essential for reasonable accuracy. Hence, the induction energy is simply estimated as[49]

$$E_{ind} = -\frac{1}{2} \sum_{k=1}^{N_{groups}} \sum_{\gamma=1}^{3} \sum_{\nu=1}^{3} \Im_\gamma[\mathbf{X}(k)] \alpha_{\gamma\nu}(k) \Im_\nu[\mathbf{X}(k)]. \tag{2.25}$$

Again, the exact first and second derivatives of the induction energy cannot be readily calculated since the corresponding derivatives of the multipole moments and group polarizability tensor are not easily obtained. Hence, first and second derivatives of E_{ind} are approximated using constant polarizabilities and distributed multipole moments, and derivatives of the T tensors in Eq. (2.24).

Finally, the total non-bonded energy is given by the sum of these contributions, as in Eq. (2.18), and the gradient and hessian of E_{nb} is given by the corresponding sum of gradients and hessians.

It is important to note that a finite value of the "cut-off" distance, d_{tol}, has the benefit of ensuring that the computational time is linear in the number of functional groups in the molecule. However, a finite value of d_{tol} also results in some discontinuity in the total energy gradient as the molecular geometry changes; as some Level 1 fragment to fragment distance passes through d_{tol}. If d_{tol} is sufficiently large, then when the fragment to fragment distance is near d_{tol}, the gradients calculated by *ab initio* methods should be close to the gradients evaluated by perturbation theory,

$$\frac{\partial E_{ab}}{\partial X_\alpha(j)} \approx \frac{\partial(E_{ele} + E_{ind} + E_{disp})}{\partial X_\alpha(j)}. \tag{2.26}$$

Hence, if d_{tol} is sufficiently large, the discontinuity in the gradients should be small in magnitude.

C. Summary

In summary, the molecular energy is given by a sum of "bonded" energies and non-bonded interactions. Each bonded fragment energy is evaluated in the presence of embedded charges which model the charge distribution of the remainder of the molecule. The non-bonded interaction energies are evaluated *ab initio* for close interactions (in the presence of

embedded charges which model the charge distribution of the remainder of the molecule) and by perturbation theory for long range interactions, in terms of electrostatic interactions, induction and dispersion.

The energy derivatives are composed of the corresponding contributions from bonded energies and non-bonded interactions. The derivatives of the *ab initio* energies with respect to the atomic positions are provided by the *ab initio* program packages, with account taken of the contributions from hydrogen atom caps. An approximate description of the gradients associated with the embedded charges is accounted for. Finally, approximate derivatives of the long range electrostatic, induction and dispersion interactions are evaluated.

III. TEST RESULTS

In this section, the accuracy of the SMFA energy, gradients, and hessians is examined for a set of molecules that vary from relatively small and simple to molecules containing about 250 atoms. This test set contains features which were not common in a set of 96 molecules previously used to examine the accuracy of systematic molecular fragmentation: The molecules considered here contain aromatic rings, extensive hydrogen bonding and formally charged groups. Some of these molecules have been studied using related fragmentation methods, which allow some comparison with alternative approaches.

A. Test molecules

1. Sugars, peptides, and organic ions

The structures for a set of 38 organic molecules have been obtained from the Cambridge Structure DataBase.[50] There are 14 structures that contain formally charged groups, and 24 structures which contain no formal charges. These 24 structures include various sugars, peptides, and miscellaneous functional groups, with between 42 and 169 atoms (the average is 64), and which feature extensive hydrogen bonding. The 14 ions/zwitterions contain between 62 and 180 atoms (the average is 93). The Cartesian coordinates for all structures are included in Table S1 of the supplementary material.[51] Figure 1 illustrates the smallest and largest of these structures.

2. Protein conformers

The structure of TM1081, a *Thermotoga maritima* anti-s factor antagonist, has been examined using NMR spectroscopy.[52] Generally, NMR data may not be sufficient to *completely* determine the structure of a protein. In Ref. 52, various computational methods were used to compile 20 feasible structures, which were deposited in the Protein Data Bank[53] with accession code 2KA5. These 20 structures all contain a terminal chain which is structurally disordered, and these chains, each containing 246 atoms, are used here as a test set of protein conformers. The Cartesian coordinates of all 20 conformers are included in Table S2 of the supplementary material.[51] Two of these structures are illustrated in Figure 2. Each structure contains three formally charged groups, two $-NH_3^+$ groups and one $-COO^-$ group. Deprotonating the $-NH_3^+$ groups and protonating the $-COO^-$ group gives an additional test set of conformers (of 245 atoms) that contain no formal charges. These modified structures are shown in Table S3 of the supplementary material.[51]

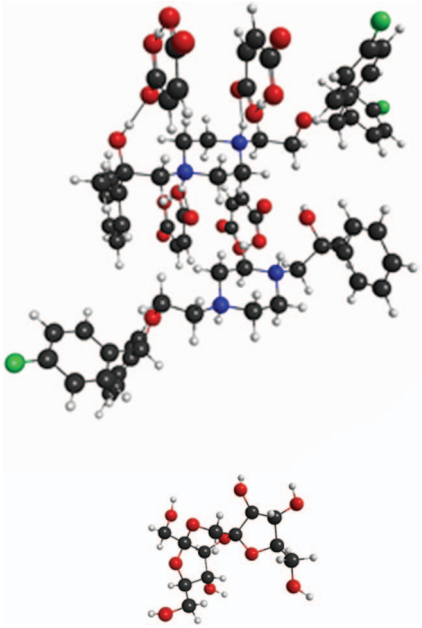

FIG. 1. Ball and stick figures depict the smallest and largest structures in the set of 38 miscellaneous test molecules.

3. Simple zwitterion chains

To illustrate the ability of SMFA to systematically describe zwitterions, a simple set of structures for

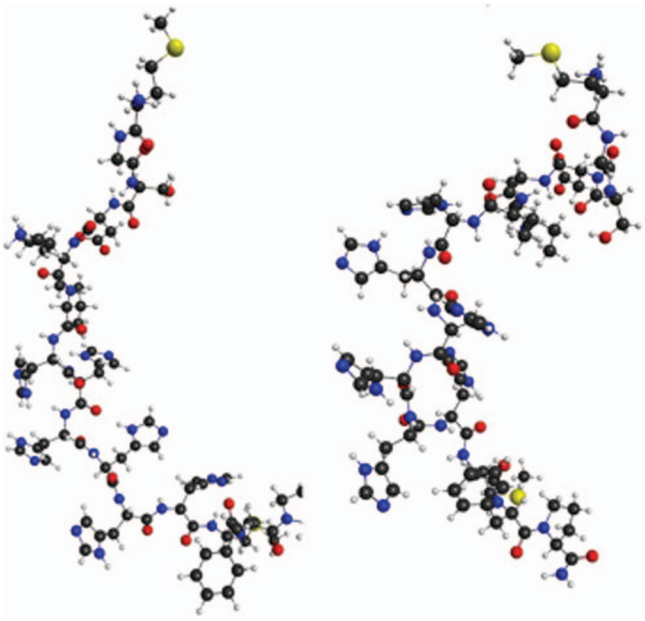

FIG. 2. Ball and stick figures depict two of the 20 protein conformers.

$NH_3^+(CH_2)_nCOO^-$ (for n = 5, …, 9) has been considered. The Cartesian coordinates of these structures are included in the supplementary material:[51] Table S4 gives the structures optimised at HF/6-31G and Table S5 gives the structures optimised at MP2/6-31+G(d,p).

4. Other test cases

To provide some comparison with related work, an additional four test cases have been included that were examined previously by Li and co-workers,[54] denoted $(Gly)_{12}$, $(H_2O)_{28}$, GelA, and GelB in Ref. 54. GelA and GelB are largely alkanes with two charged nitrogen or oxygen containing groups; $(H_2O)_{28}$ is a water cluster; and $(Gly)_{12}$ is a poly-glycine with an α-helical structure. The Cartesian coordinates of these structures [optimised at HF/6-31G(d)] are given in the supplementary material for Ref. 54.

5. Computations

For the miscellaneous test set and protein conformers, *ab initio* calculations were carried out at the HF/6-31G level using the GAUSSIAN09 program package.[55] For the $NH_3^+(CH_2)_nCOO^-$ structures, calculations were carried out at the HF/6-31G and MP2/6-31+G(d,p) levels using GAUSSIAN09. For MP2 calculations, the dispersion energy was evaluated using the DALTON program package.[56] For $(Gly)_{12}$, $(H_2O)_{28}$, GelA, and GelB, HF/6-31G(d) calculations were carried out using GAUSSIAN09.

B. Results

1. Sugars, peptides, and organic ions

The set of miscellaneous organic molecules and ions contain many hydrogen bonds and amide groups. If hydrogen bonds are treated as normal single bonds, these typically produce small ring motifs in the bonding structure. As previously discussed,[41] small rings cannot be fragmented, because partial ring fragments contain capping H atoms in close proximity. Rings of up to 5, 6, 7, 8 groups cannot be fragmented at Levels 2, 3, 4, 5 or above, respectively. Some structures, such as α-helical peptides, may then contain small connected rings which cannot be fragmented. This difficulty can be ameliorated if the C–N bond in an amide group is treated as a single bond, despite being significantly shorter than a normal single C–N bond. Although the default definition of bonding sets the amide C–N bond as multiple, here it is set to be a single bond to allow fragmentation of some α-helical structures.

Figure 3 presents the mean absolute error in the total HF/6-31G electronic energy for the set of miscellaneous organic molecules and ions as a function of the Level of fragmentation. The standard value[26] of d_{tol}, 1.1, is used, except for Level 2 where $d_{tol} = 0$ (in order to avoid close contact of capping hydrogen atoms in the non-bonded interactions). The total energy is not strongly dependent on the value of d_{tol}, for values larger than 1.1. Non-bonded interactions between charged Level 1 fragments are evaluated *ab initio*, as in Eq. (2.19). The graph clearly indicates convergence of the energy to within a few kJ mol^{-1} as the Level of fragmentation increases. Table S8 in the supplementary material[51] presents the individual contributions to the energy [see Eqs. (2.1) and (2.18)] for each of the 38 molecules, at Levels 2–5. For some molecules, the total energy is reasonably well converged by Level 3, but as Figure 3 indicates, generally convergence is established by Level 4. The 38 molecules contain an average of 24.5 groups, while the fragment molecules contain an average of 6.6 groups at Level 3 and 9.3 groups at Level 4. For any arbitrary molecule, a practical approach is to estimate the total energy, using a low (inexpensive) level of *ab initio* theory at Levels 2–5 to observe convergence of the total energy. Earlier work has established that the error due to fragmentation does not depend significantly on the size of the basis set or the method used to treat electron correlation.[24,25] Thus, once the appropriate Level of fragmentation has been determined, convergence of the energy with respect to basis set and treatment of electron correlation can be examined in the usual way. The total calculation time is linearly proportional to the number of groups in the molecule. However, the *ab initio* calculations for all fragments are independent, so if sufficient processors are available the "walltime" is determined by the largest fragment calculation. The practical limit on the size of basis set and treatment of electron correlation is thus determined by the size of the largest fragment.

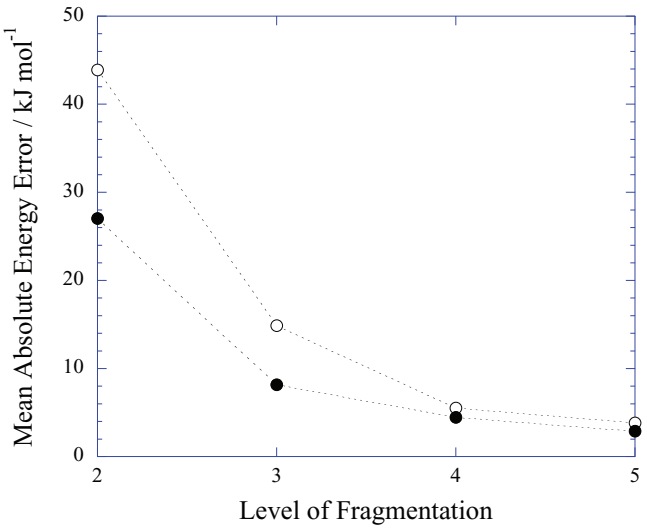

FIG. 3. The mean absolute error in the fragmentation approximation to the total HF/6-31G energy is shown versus the Level of fragmentation. The open symbols correspond to the 24 miscellaneous neutral organic molecules and the filed symbols correspond to the 14 organic ions and zwitterions.

Figure 4 presents the mean absolute error in the Cartesian energy gradients for this test set, for several values of the Level of fragmentation and the tolerance, d_{tol}. As the value of d_{tol} increases, more non-bonded interactions are evaluated *ab initio*, rather than using electrostatic multipole expansions. These 38 molecular configurations, from a crystallography database, are not stationary configurations. The largest gradient at HF/6-31G is of the order of 0.5 (hartree/bohr), and on average a HF/6-31G gradient component is about 0.07 (hartree/bohr) in magnitude. Given the large forces present in these structures, and the different chemical composition of

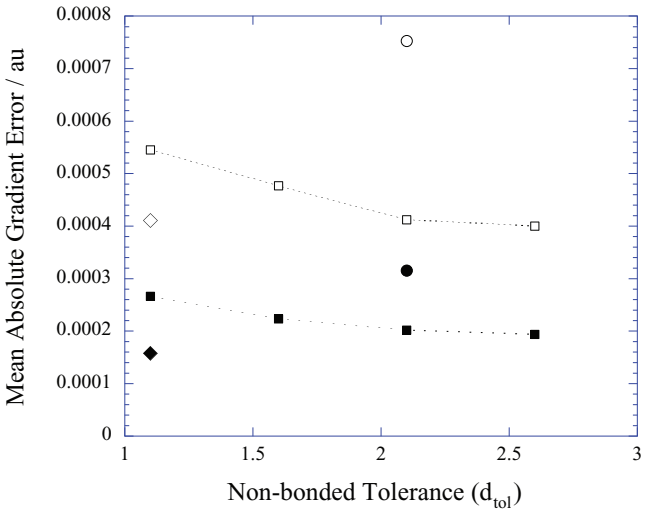

FIG. 4. The mean absolute error in the fragmentation approximation to the Cartesian energy gradients (1 a.u. = 1 hartree/bohr) for the 38 miscellaneous molecules is shown as function of the value of d_{tol} (see text). The open symbols correspond to the 24 miscellaneous neutral organic molecules and the filed symbols correspond to the 14 organic ions and zwitterions; Level 3 (circle), Level 4 (squares), and Level 5 (diamond).

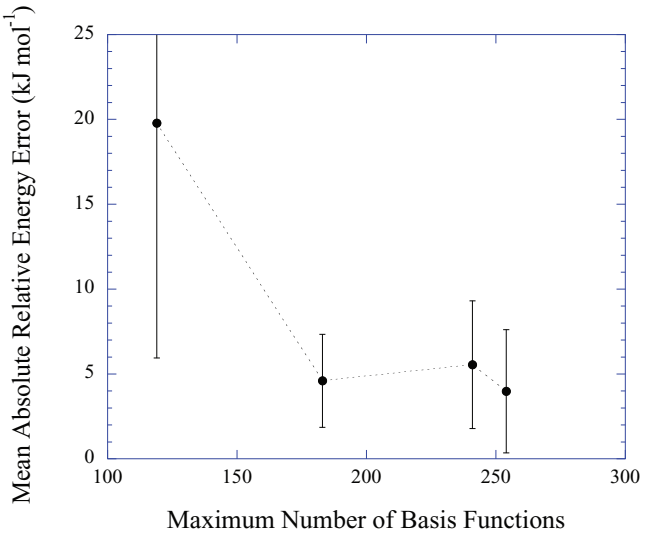

FIG. 5. The mean absolute error in the relative energies of the 20 protein conformations is shown against the maximum number of basis functions for a single fragment calculation for Levels 2–5. The error bars represent one standard deviation for the 20 relative energy errors. For comparison, the whole molecule has 1410 functions for the Pople-style 6-31G basis.

the neutral and charged molecules, the differences in gradient errors between neutral and charged species are not clearly significant. Figure 4 indicates that the error in the gradient is smaller at higher Levels of fragmentation and at larger values of d_{tol}. When considering these gradient errors, it is useful to note that standard *ab initio* program packages consider that a geometry is a stationary point (zero gradient) if the root-mean-square gradient is below about 0.0003 (hartree/bohr), and the largest gradient is below 0.00045 (hartree/bohr). Figure 4 indicates that the error due to the fragmentation approximation is approaching zero on this scale. As noted in Sec. II B, the estimation of gradients arising from E_{ele} is somewhat arbitrary. Hence it is not surprising that the mean gradient error is reduced for larger values of d_{tol}. For practical calculation of gradients for geometry optimization, Figure 4 indicates that Level 4 fragmentation is significantly more accurate than Level 3, and that a value of d_{tol} near 2 would be preferable to the default value of 1.1, which is adequate for energy calculations. At Level 4 and $d_{tol} = 1.1$, the mean absolute gradient error is 5.5×10^{-4} (hartree/bohr) if gradients of E_{ele} are included, but 6.0×10^{-4} if these are neglected. However, for values of $d_{tol} \geq 1.6$, there is no apparent reduction in the gradient error if gradients of E_{ele} are included.

For the ions and zwitterions, the forces on the background charges, as in Eq. (2.13), have been accounted for. However, these appear to be very small contributions in these test cases: For Level 4 and $d_{tol} = 1.1$, the mean absolute gradient error is 2.7×10^{-4} (hartree/bohr), and this increases to only 2.8×10^{-4} (hartree/bohr) if the contributions from Eq. (2.13) are neglected.

2. Protein conformers

The energies of the protein conformers have been determined via whole molecule calculations and via fragmentation using the standard defaults (amide CN bonds are multiple bonds and $d_{tol} = 1.1$). For each conformer the energy is measured relative to the average energy of all the conformers. The error in these relative energies is then obtained by comparing the fragmentation and whole-molecule values. Figure 5 shows the mean and standard deviation of the absolute error in the relative energy of these 20 conformers for Levels 2–5, graphed as a function of the maximum number of basis functions required in a single calculation. Table S9 in the supplementary material[51] shows the energies for each conformer. Figure 5 indicates that the relative energies of these conformers have converged at Level 3. Figure 6 shows the Level 3 relative energies for all 20 conformers versus the exact HF/6-31G value. This plot shows that the 20 conform-

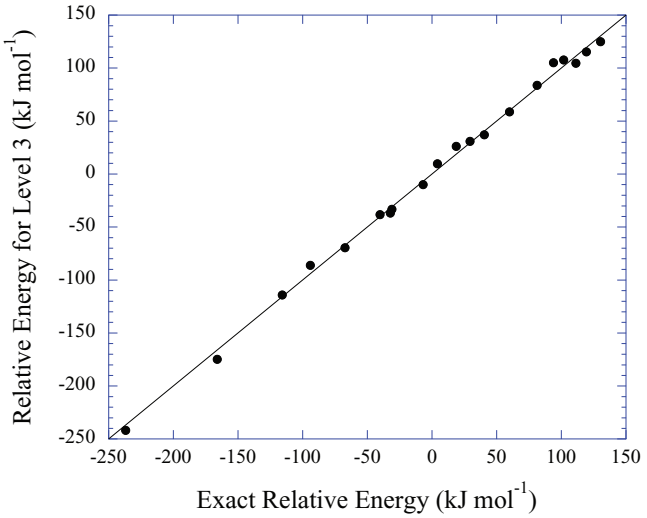

FIG. 6. The relative energy of each protein conformer, evaluated at Level 3 fragmentation, is shown versus the exact HF/6-31G value. The diagonal line indicates perfect agreement.

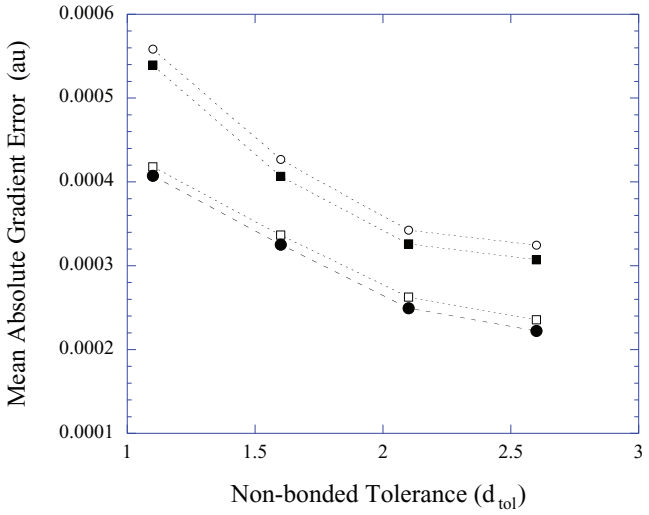

FIG. 7. The mean absolute error in the components of the energy gradient for the 20 protein conformers is shown as a function of the d_{tol} parameter, for Level 3 (■) and Level 4 (●) fragmentation. The results for the corresponding neutral peptides (see text) are shown for Level 3 (o) and Level 4 (□).

ers range in energy over about 367 kJ mol^{-1} and that the relative order of these energies is reproduced well at Level 3 fragmentation.

These 20 structures are not stationary points for HF/6-31G calculations, with gradients as large as about 0.05 a.u. in each case. Figure 7 displays the mean absolute error in the HF/6-31G gradients for the 20 protein structures versus d_{tol} for Level 3 and Level 4 fragmentation. Contributions to the gradient from the electrostatic energy and the forces on the background charges have been included. Figure 7 also presents the corresponding mean absolute errors for the "neutral" proteins, in which the $-NH_3^+$ groups are de-protonated and the $-COO^-$ group is protonated. As for the miscellaneous test set, the contributions from the electrostatic energy, E_{ele}, are small for $d_{tol} \geq 1.6$. Fig. 7 displays the same trend in the gradient error versus d_{tol} as does Fig. 4. At Level 4, the mean absolute error is below 3×10^{-4} for d_{tol} above about 1.8. The figure also clearly indicates that the gradients are evaluated with much the same accuracy for both the charged and neutral proteins.

These proteins contain aromatic rings in the side chains, histidine, and phenylanaline residues, which impose a lower limit on the size of the largest fragments. At Level 4 (Level 3), the largest *ab initio* calculation requires 241 (183) basis functions at 6-31G, compared to 1410 basis functions for the whole molecule.

3. Simple zwitterion chains

Figure 7 indicates that the error in the energy gradient at Level 3 is reasonably converged for $d_{tol} \geq 2$. Hence, geometry optimisation might be carried out using values of d_{tol} in this range. Table I presents the results of geometry optimisation of the molecules $NH_3^+(CH_2)_nCOO^-$, for n = 5, . . . , 9 at the HF/6-31G and MP2/6-31+G(d,p) levels of *ab initio* theory, at Level 3 with $d_{tol} = 2.1$, compared to optimisation of the whole molecules. Tables S6 and S7 in the supplementary material[51] shows the optimised structures for Level 3 and $d_{tol} = 2.1$. Results for $d_{tol} = 1.6$ are very similar to those shown in the tables. The SMFA and whole-molecule optimised Cartesian structures were converted to z-matrix format using the *open babel* program[57] to allow comparison of bond lengths, valance bond angles, and dihedral angles. It is clear from Table I that the optimised geometries are very close to the exact structures in every case, indicating that the energy gradients have been evaluated to sufficient accuracy. The mean errors in the frequencies are also very small, which indicates that the hessians have also been evaluated to sufficient accuracy. The errors for MP2/6-31+G(d,p) appear to be slightly larger than for HF/6-31G, but are still small. There are no significant contributions to the energy derivatives from long range dispersion, in these cases. Figure 8 presents the highest and lowest frequencies in these chains, comparing the exact and Level 3 results. It is interesting to note that the variation of the lowest frequency mode with chain length is well reproduced in the fragmentation approximation, even though this mode is a delocalised cooperative motion of the whole chain.

TABLE I. The geometry, energy, and frequency errors in $NH_3^+CH_2(CH_2)_nCH_2COO^-$ chains, with n = 5, . . . , 9, optimised at the given levels of theory using fragmentation Level = 3, and $d_{tol} = 2.1$. The errors are relative to the corresponding properties for structures optimised without fragmentation. Here, $\langle|\delta r|\rangle$ is the mean absolute error in the bond lengths, $\langle|\delta\theta|\rangle$ is the mean absolute error in the valance bond angles, $\langle|\delta\tau|\rangle$ is the mean absolute error in the dihedral angles, $\langle|\delta\omega|\rangle$ is mean absolute error in the harmonic vibrational frequencies, and Max $|\delta\omega|$ is the maximum absolute error in the frequencies.

System	Energy error (kJ mol^{-1})	$\langle\|\delta r\|\rangle$ (Å)	$\langle\|\delta\theta\|\rangle$ (deg)	$\langle\|\delta\tau\|\rangle$ (deg)	$\langle\|\delta\omega\|\rangle$ (cm^{-1})	Max $\|\delta\omega\|$ (cm^{-1})
HF/6-31G						
5	−0.34	0.00020	0.032	0.15	0.6	3.9
6	−0.74	0.00019	0.024	0.15	0.6	3.1
7	−0.75	0.00018	0.026	0.12	0.7	3.3
8	−0.79	0.00015	0.021	0.028	0.7	2.0
9	−0.95	0.00015	0.023	0.027	0.7	2.9
MP2/6-31+G(d,p)						
5	−1.20	0.00022	0.053	0.49	0.9	4.2
6	−2.59	0.00021	0.063	0.72	1.2	4.7
7	−2.79	0.00022	0.054	0.55	1.1	3.7
8	−2.97	0.00019	0.044	0.53	1.3	8.7
9	−3.38	0.00020	0.045	0.56	1.4	9.4

4. Other test cases

Reference 54 presented measures of the accuracy with which the GEBF fragmentation method produced optimised structures for moderately large molecules, ions, and clusters. Table II presents SMFA results for a selection of these structures, denoted $(Gly)_{12}$, $(H_2O)_{28}$, GelA and GelB in Ref. 54. The SMFA optimised structures are given in Table S10 of the supplementary material[51] for this article, while Table II includes some measures of the deviation of these SMFA structures from the optimised structures.[54] Given, the results above for the accuracy of energy gradients, a value of $d_{tol} = 2.1$ was used for geometry optimisation for $(H_2O)_{28}$, GelA and

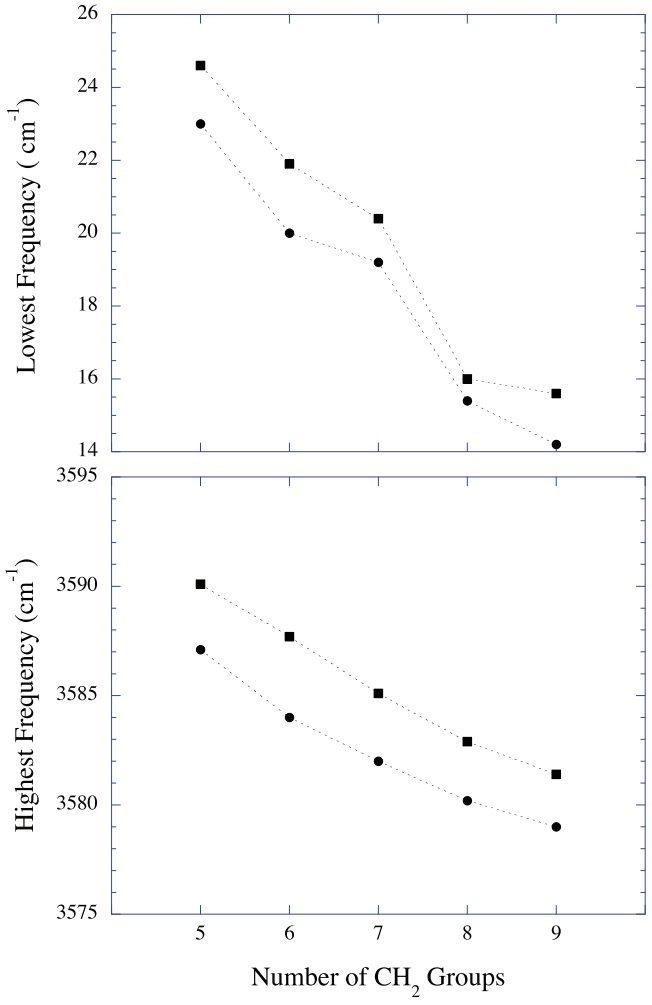

FIG. 8. The highest and lowest frequencies for $NH_3{}^+(CH_2)_nCOO^-$ (for $n = 5, \ldots, 9$) at the MP2/6-31+G(d,p) level are shown for the whole molecule (•) and using Level 3 fragmentation with $d_{tol} = 2.1$ (■).

GelB, while it was found that $d_{tol} = 3.1$ was more accurate for $(Gly)_{12}$ at Level 3. The SMFA and whole-molecule optimised Cartesian structures were converted to z-matrix format using the *open babel* program,[57] so that mean absolute deviations in the valance bond lengths, valance bond angles and dihedral angles could be calculated. Note that for $(H_2O)_{28}$ hydrogen bonds are included in the bond lengths, bond angles, and dihedrals. Harmonic vibrational frequencies were evaluated for all optimised structures, so that the mean absolute deviation in the frequencies and the maximum absolute frequency deviation could be evaluated. For GelA and GelB, embedded charges were employed to represent the formally charged groups only. Table II shows the advantage of using embedded charges in a water cluster, which has a high density of very polar groups. The charges are obtained from a NPA calculation of each isolated water monomer. The $(H_2O)_{28}$ optimised geometry, including the very "soft" dihedral angles, is more accurate when these charges are included. Moreover, the vibrational frequencies are much more accurate. For $(Gly)_{12}$, embedded charges were included to represent only the 13 amide groups (using NPA charges for the isolated groups), and hydrogen bonds were not included. The α-helical structure of $(Gly)_{12}$ was then obtained with the accuracy shown. Similar accuracy was not possible in the absence of embedded charges.

The results in Table II can be compared with corresponding results in Tables I and III of Ref. 54. This comparison indicates that the SMFA method produces approximate energies, geometries, and vibrational frequencies that are comparable or better in accuracy to the GEBF method, and often requires smaller fragments (lower numbers of basis functions).

IV. CONCLUDING REMARKS

This paper has established a protocol which can be used in an automated procedure to estimate the energy, minimum energy geometry, and harmonic vibrational frequencies of general organic and biological molecules using SMFA with embedded charges.

The accuracy of the protocol has been demonstrated for a wide range of molecules. One might ask why are the gradients and hessians in SMFA are so accurate, given the many approximations involved – fragmentation, the use of capping hydrogen atoms, treatment of through space interactions partly by *ab initio* methods and partly by perturbation theory including multipole-multipole interactions, the use of embedded point charges, and a very simple account of the forces on embedded charges. The basic reason why the energy and corresponding derivatives are accurate is because the large forces are described by the Level L fragmentation of Eqs. (2.2)

TABLE II. The structure of the listed molecules has been optimised at the fragmentation Levels shown, at the HF/6-31G(d) level of *ab initio* theory, and compared with the optimised structure of the whole molecule. The notation "(chgs)" denotes that embedded charges have been used (see text for details). "Basis" shows the maximum number of basis functions required for a single fragment calculation in SMFA, compared to the number of basis functions required for the whole molecule, while other notations are defined in Table I.

Molecule	Level	Exact energy (a.u.)	Energy error (a.u.)	Basis	$\langle\|\delta r\|\rangle$ (Å)	$\langle\|\delta\theta\|\rangle$ (deg)	$\langle\|\delta\tau\|\rangle$ (deg)	$\langle\|\delta\omega\|\rangle$ (cm^{-1})	Max $\|\Delta\omega\|$ (cm^{-1})
GelA	3	−1946.05069	−0.00112	216/710	0.00020	0.047	0.32	0.8	6.3
GelB	3	−2335.42596	0.00055	178/706	0.00015	0.0437	0.30	0.7	10.5
$(H_2O)_{28}$	2	−2128.80493	−0.00082	114/532	0.04771	3.308	57.41	5.7	62.2
$(H_2O)_{28}$	3	−2128.80493	0.00329	209/532	0.02230	1.594	27.18	2.5	17.1
$(H_2O)_{28}$	2 (chgs)	−2128.80493	−0.00001	114/532	0.00218	0.263	10.69	1.6	17.3
$(H_2O)_{28}$	3 (chgs)	−2128.80493	0.00030	209/532	0.00065	0.078	0.26	0.7	8.0
$(Gly)_{12}$	3 (chgs)	−2728.84807	−0.00090	140/881	0.00082	0.155	4.82	2.4	8.8

and (2.4). The capping hydrogen atoms formally cancel in Eq. (2.2). If L is sufficiently large, the chemical environment of each cap in a fragment with positive coefficient, f_n, is very nearly the same in other fragments with complementary negative coefficients. Hence, the net forces on the caps are near zero; the caps are not exerting forces on the "real" atoms. As the value of L increases, the net forces of the caps more nearly vanish. Other fragmentation schemes report reasonably accurate energy gradients and hessians, simply by using the *ab initio* derivatives of the "real" atoms in the fragments, with no correction for caps or for long range interactions.[31,54] The correction for caps in Eq. (2.7) is likely very small, any error in this correction is also likely to be very small. Smaller forces arise from the non-bonded interactions. The largest of these are evaluated *ab initio*, for all interactions between fragments closer than d_{tol}. Again, the H atom caps formally cancel in the Level 1 fragmentation, so make little contribution to a Level 1–Level 1 description of the non-bonded interactions; some interactions involving caps arise from the editing process that restricts Eq. (2.17) to *allowed* interactions. Only the quite long range (and likely very small) non-bonded interactions are treated using perturbation theory. Other authors neglect these interactions entirely or account for them via the embedded charges. No doubt the very approximate treatment of long range forces produces a significant *relative* error in the estimates of these forces, but the total error is small because the forces are small. This comment applies to the forces originating from both Eqs. (2.13) and (2.21), as both these forces only make a contribution to interactions between fragments which are separated by more than d_{tol} and/or do not share a Level L fragment in common. The results above show that accurate gradients and hessians are obtained for neutral molecules, ions, and zwitterions, which indicates that the approximations employed for all long range interactions are sufficiently accurate.

The computational efficiency of the SMFA approach resides in the fact that the total computational effort required scales only linearly with the number of functional groups in the molecule. Moreover, given a sufficiently large number of processors, the walltime required is only proportional to the size of the largest single fragment produced by the SMFA process.

A significant advantage of the SMFA procedure is that, as the name suggests, it is systematic. Estimation of molecular properties at Levels 2–5 (say) demonstrates convergence of the property value and establishes the Level of fragmentation required. Systematic convergence of the property values with respect to the level of *ab initio* method and basis set can then be explored in the usual way.

ACKNOWLEDGMENTS

The author thanks Dr. Terry Frankcombe and Mr. David Reid for useful discussions.

APPENDIX A: GROUPS AND BONDING

All atoms in a molecule are assigned to a set of groups, which are generally intended to coincide with the usual chemical definition of a functional group. The definition of groups follows from the definition of "bonding" and is largely the same as that reported previously:[25]

(i) Two atoms are defined to be connected by a single bond if the sum of the covalent radii of the two atoms + 0.4 Å is greater than the distance between the atoms. All single bonds are assigned initially.
(ii) A single bond is replaced by a multiple bond if the sum of the covalent radii of the two atoms +0.08 Å is greater than the distance between the atoms, unless either atom has its normal valence (for example, an oxygen has two single bonds, nitrogen has three single bonds, carbon has four single bonds). For example, one could define the CN bond in an amide group to be a multiple bond (such bonds, at equilibrium, would be sufficiently short in length to be defined as multiple bonds)
(iii) Two atoms connected by a multiple bond are contained in the same group. Hydrogen atoms are defined to be contained in the same group as the "heavy" atom to which they bonded.
(iv) In the case where formal charges are present: If atom A is connected by a single bond to a formally charged atom B (or to an atom which is connected to a formally charged atom B by a multiple bond), then atom A is taken to be in the same group as the formally charged atom.So, for example, in a structure like $–(CH_2)–COO^-$, one O atom is formally charged. Both O atoms may have short bonds to the C atom, taken to be multiple bonds. Thus all three atoms, COO, must be in the same group. Then, since the CH_2 group is connected by a single bond to the C atom, all the atoms, $(CH_2)COO$, are contained in a single group.
(v) Hydrogen bonds can be defined as a separate type of bond. In the current implementation, a hydrogen bond can exist only between two "heavy atoms" that are either oxygen or nitrogen. Such a bond exists if the distance between a hydrogen single bonded to O or N is within 2.4 Å of another O or N atom. Hydrogen bonds are treated like any other single bond by the fragmentation procedure, except that no capping H atom is employed when a hydrogen bond is broken. It is worth noting that if hydrogen bonds are not requested, then such interactions are accounted for as part of the non-bonded interactions. For the case of water clusters, hydrogen bonds are employed, as they are the only bonds connecting groups.

APPENDIX B: POINT CHARGES

The molecule is fragmented into single groups, where hydrogen caps are introduced in place of any bonds broken in the original molecule, as previously described.[25] The ground state electronic wavefunction is evaluated at the chosen level of *ab initio* theory and a Natural Population Analysis[58,59] is used to assign charges to each atomic centre.

Each group may contain one or more hydrogen atom cap, each attached to a heavy atom in the group. The charge on each hydrogen cap is added to that of the heavy atom to which it is attached, at the position of the heavy atom. No charge

is located at the cap positions. This procedure prevents very close contact between atoms in other groups with the point charges in any subsequent *ab initio* calculation. The total (integer) charge of the formally charged group is preserved.

More complicated approaches using distributed multipole representations of the charge distribution[47] were examined but found to be unnecessary.

[1]M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. Slipchenko, Chem. Rev. **112**, 632 (2012).
[2]D. G. Fedorov and K. Kitaura, J. Chem. Phys. **120**, 6832 (2004).
[3]D. G. Fedorov and K. Kitaura, J. Chem. Phys. **122**, 054108 (2005).
[4]W. Li and S. Li, J. Chem. Phys. **122**, 194109 (2005).
[5]Y. Guo, W. Li, and S. Li, Chem. Phys. Lett. **539–540**, 186 (2012).
[6]K. Babu and S. R. Gadre, J. Comput. Chem. **24**, 484 (2003).
[7]W. Yang and T. Lee, J. Chem. Phys. **103**, 5674 (1995).
[8]M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, and K. Morokuma, J. Phys. Chem. **100**, 19357 (1996).
[9]T. Vreven, K. Morokuma, O. Farkas, H. B. Schlegel, and H. B. Frisch, J. Comput. Chem. **24**, 760 (2003).
[10]P. Canfield, M. G. Dahlbom, N. S. Hush, and J. R. Riemers, J. Chem. Phys. **124**, 024301 (2006).
[11]M. S. Gordon, J. M. Mullin, S. R. Pruitt, L. B. Roskop, L. V. Slipchenko, and J. A. Boatz, J. Phys. Chem. B **113**, 9646 (2009).
[12]J. M. Mullin, L. B. Roskop, S. R. Pruitt, M. A. Collins, and M. S. Gordon, J. Phys. Chem. A **113**, 10040 (2009).
[13]J. Gao, J. Phys. Chem. B **101**, 657 (1997).
[14]J. Gao, J. Chem. Phys. **109**, 2346 (1998).
[15]W. Xie and J. Gao, J. Chem. Theory Comput. **3**, 1890 (2007).
[16]W. Xie, M. Orozco, D. G. Truhlar, and J. Gao, J. Chem. Theory Comput. **5**, 459 (2009).
[17]D. G. Fedorov and K. Kitaura, J. Phys. Chem. A **111**, 6904 (2007).
[18]D. G. Fedorov and K. Kitaura, *The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems* (CRC Press, Boca Raton, 2009).
[19]D. W. Zhang and J. Z. H. Zhang, J. Chem. Phys. **119**, 3599 (2003).
[20]Y. Mei, C. Ji, and J. Z. H. Zhang, J. Chem. Phys. **125**, 094906 (2006).
[21]X. H. Chen, D. W. Zhang, and J. Z. H. Zhang, J. Chem. Phys. **120**, 839 (2004).
[22]D. W. Zhang and J. Z. H. Zhang, J. Theor. Comput. Chem. **3**, 43–49 (2004).
[23]L. L. Duan, Y. Mei, Q. G. Zhang, and J. Z. H. Zhang, J. Chem. Phys. **130**, 115102 (2009).
[24]V. Deev and M. A. Collins, J. Chem. Phys. **122**, 154102 (2005).
[25]M. A. Collins and V. A. Deev, J. Chem. Phys. **125**, 104104 (2006).
[26]M. A. Addicoat and M. A. Collins, J. Chem. Phys. **131**, 104103 (2009).
[27]R. P. A. Bettens and A. M. Lee, J. Phys. Chem. A **110**, 8777 (2006).
[28]R. P. A. Bettens and A. M. Lee, Chem. Phys. Lett. **449**, 341 (2007).
[29]H.-A. Le, A. M. Lee, and R. P. A. Bettens, J. Phys. Chem. A **113**, 10527 (2009).
[30]H.-A. Le, H.-J. Tan, J. F. Ouyang, and R. P. A. Bettens, J. Chem. Theory Comput. **8**, 469 (2012).
[31]V. Ganesh, R. K. Dongare, P. Balanarayan, and S. R. Gadre, J. Chem. Phys. **125**, 104109 (2006).
[32]N. Sahu, S. D. Yeole, and S. R. Gadre, J. Chem. Phys. **138**, 104101 (2013).
[33]S. Li, W. Li, and T. Fang, J. Am. Chem. Soc. **127**, 7215 (2005).
[34]W. Li, S. Li, and Y. Jiang, J. Phys. Chem. A **111**, 2193 (2007).
[35]S. Hua, W. Hua, and S. Li, J. Phys. Chem. A **114**, 8126 (2010).
[36]N. Jiang, J. Ma, and Y. Jiang, J. Chem. Phys. **124**, 114112 (2006).
[37]R. O. Ramabhadran and K. Raghavachari, J. Chem. Theory Comput. **7**, 2094 (2011).
[38]N. J. Mayhall and K. Raghavachari, J. Chem. Theory Comput. **7**, 1336 (2011).
[39]N. J. Mayhall and K. Raghavachari, J. Chem. Theory Comput. **8**, 2669 (2012).
[40]R. M. Richard and J. M. Herbert, J. Chem. Phys. **137**, 064113 (2012).
[41]M. A. Collins, Phys. Chem. Chem. Phys. **14**, 7744 (2012).
[42]E. E. Dahlke and D. G. Truhlar, J. Chem. Theory Comput. **3**, 46 (2007).
[43]J. Gao and Y. Wang, J. Chem. Phys. **136**, 071101 (2012).
[44]H. M. Netzloff and M. A. Collins, J. Chem. Phys. **127**, 134113 (2007).
[45]S. R. Pruitt, M. A. Addicoat, M. A. Collins, and M. S. Gordon, Phys. Chem. Chem. Phys. **14**, 7752 (2012).
[46]D. M. Reid and M. A. Collins, J. Chem. Phys. **139**, 184117 (2013).
[47]A. J. Stone, J. Chem. Theory Comput. **1**, 1128 (2005).
[48]J. Gao, J. Comput. Chem. **18**, 1061 (1997).
[49]A. J. Stone, *The Theory of Intermolecular Forces* (Clarendon, Oxford, 1996).
[50]F. H. Allen, Acta Crystallogr. **B58**, 380 (2002).
[51]See supplementary material at http://dx.doi.org/10.1063/1.4894185 for tables of molecular geometries and energies.
[52]P. Serrano, B. Pedrini, M. Geralt, K. Jaudzems, B. Mohanty, R. Horst, T. Herrmann, M.-A. Elsliger, I. A. Wilson, and K. Wüthrich, Acta Cryst. **66**, 1393 (2010).
[53]See http://www.rcsb.org/pdb/home/home.do for Protein Data Bank.
[54]W. Hua, T. Fang, W. Li, J.-G. Yu, and S. Li, J. Phys. Chem. A **112**, 10864 (2008).
[55]M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, Gaussian09, Gaussian, Inc., Wallingford, CT, 2009.
[56]K. Aidas, C. Angeli, K. L. Bak, V. Bakken, R. Bast, L. Boman, O. Christiansen, R. Cimiraglia, S. Coriani, P. Dahle, E. K. Dalskov, U. Ekström, T. Enevoldsen, J. J. Eriksen, P. Ettenhuber, B. Fernández, L. Ferrighi, H. Fliegl, L. Frediani, K. Hald, A. Halkier, C. Hättig, H. Heiberg, T. Helgaker, A. C. Hennum, H. Hettema, E. Hjertenæs, S. Høst, I.-M. Høyvik, M. F. Iozzi, B. Jansik, H. J. A. Jensen, D. Jonsson, P. Jørgensen, J. Kauczor, S. Kirpekar, T. Kjærgaard, W. Klopper, S. Knecht, R. Kobayashi, H. Koch, J. Kongsted, A. Krapp, K. Kristensen, A. Ligabue, O. B. Lutnæs, J. I. Melo, K. V. Mikkelsen, R. H. Myhre, C. Neiss, C. B. Nielsen, P. Norman, J. Olsen, J. M. H. Olsen, A. Osted, M. J. Packer, F. Pawlowski, T. B. Pedersen, P. F. Provasi, S. Reine, Z. Rinkevicius, T. A. Ruden, K. Ruud, V. Rybkin, P. Salek, C. C. M. Samson, A. S. d. Merás, T. Saue, S. P. A. Sauer, B. Schimmelpfennig, K. Sneskov, A. H. Steindal, K. O. Sylvester-Hvid, P. R. Taylor, A. M. Teale, E. I. Tellgren, D. P. Tew, A. J. Thorvaldsen, L. Thøgersen, O. Vahtras, M. A. Watson, D. J. D. Wilson, M. Ziolkowski, and H. Ågren, WIREs Comput. Mol. Sci. **4**, 269, (2014).
[57]See http://openbabel.org/wiki/Main_Page for open babel.
[58]A. E. Reed and F. Weinhold, J. Chem. Phys. **78**, 4066–4073 (1983).
[59]A. E. Reed, R. B. Weinstock, and F. Weinhold, J. Chem. Phys. **83**, 735 (1985).