

Accuracy and efficiency of electronic energies from systematic molecular fragmentation

Michael A. Collins and Vitali A. Deev

Citation: *The Journal of Chemical Physics* **125**, 104104 (2006); doi: 10.1063/1.2347710

View online: <http://dx.doi.org/10.1063/1.2347710>

View Table of Contents: <http://scitation.aip.org/content/aip/journal/jcp/125/10?ver=pdfcov>

Published by the [AIP Publishing](#)

Articles you may be interested in

[The dissociative chemisorption of methane on Ni\(111\): The effects of molecular vibration and lattice motion](#)
J. Chem. Phys. **138**, 174705 (2013); 10.1063/1.4802008

[Shattering dissociation in high-energy molecular collisions between nitrate esters](#)
J. Chem. Phys. **135**, 114306 (2011); 10.1063/1.3640000

[Molecular acidity: A quantitative conceptual density functional theory description](#)
J. Chem. Phys. **131**, 164107 (2009); 10.1063/1.3251124

[Molecular potential energy surfaces constructed from interpolation of systematic fragment surfaces](#)
J. Chem. Phys. **127**, 024104 (2007); 10.1063/1.2746025

[Addition of water, methanol, and ammonia to Al₃O₃ – clusters: Reaction products, transition states, and electron detachment energies](#)
J. Chem. Phys. **122**, 214309 (2005); 10.1063/1.1926279

 **AIP** | APL Photonics

APL Photonics is pleased to announce
Benjamin Eggleton as its Editor-in-Chief



Accuracy and efficiency of electronic energies from systematic molecular fragmentation

Michael A. Collins^{a)} and Vitali A. Deev^{b)}

Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia

(Received 9 May 2006; accepted 7 August 2006; published online 8 September 2006)

A systematic method for approximating the *ab initio* electronic energy of molecules from the energies of molecular fragments is tested on a large sample of typical organic molecular structures. The detailed methods, including some additional refinements for molecular rings and long range interactions, are described. The accuracy and computational efficiency of the systematic hierarchy of methods are reported. © 2006 American Institute of Physics. [DOI: 10.1063/1.2347710]

I. INTRODUCTION

Ab initio quantum chemistry provides the practical means to calculate the total electronic energy of moderate-sized molecules. There is a hierarchy of techniques ranging in both accuracy and computational cost from the simple Hartree-Fock or density functional theory approaches to those with high level treatments of electron correlation. Thermochemical properties and, in principle, the complete potential energy surface can be evaluated. Hence, chemical reaction dynamics, rate coefficients, and other observables may be evaluated.

However, the computational time required to calculate the total electronic energy increases rapidly with the number of electrons in the molecule, and with the level of *ab initio* theory employed. Accurate treatment of electron correlation is very time consuming. For example, the coupled cluster approximation using single and double excitations and including the triple excitations noniteratively [CCSD(T)] requires a computer time which scales as N^7 , where N is the number of basis functions. Hence, research is directed towards devising new computational algorithms which reduce this “scaling problem” to linearity, N^1 . (see, for example, Ref. 1 and references therein).

One general approach to achieve linear scaling in electronic structure calculations involves breaking the molecule into fragments, evaluating the molecular orbitals or the electron density or the density matrix for the fragments, and combining these to estimate the corresponding quantity for the whole molecule.² These methods are most directly applied to Hartree-Fock (HF) and density functional theory (DFT) calculations. Zhang and co-workers developed^{3,4} and applied⁵ an even simpler idea to estimate the interaction energy between two large molecules from the interactions of fragments of the two molecules. These fragments are not disjoint but overlapping segments of the molecules, similar to the fragments employed in “molecular tailoring.”⁶ Motivated by this work, it soon became clear to a number of

authors that the total energy of a molecule could be estimated very simply from combinations of the energies of such tailored molecular fragments.⁷⁻⁹

In a previous paper (I),⁹ we presented a systematic fragmentation approach to the scaling problem. A molecule is broken up into a set of small molecular fragments. The electronic energy of each fragment is evaluated and the total electronic energy of the whole molecule is approximated by a linear combination of the fragment energies. Importantly, a systematic hierarchy of fragmentation approximations was developed whereby a computer algorithm can automatically and systematically decompose molecules into fragments in a hierarchy of fragment sizes. This provides a sequence of increasingly reliable estimates of the total electronic energy of a molecule. Calculations were presented for the first three members of this hierarchy, denoted as levels 1–3. This automated approach considers the interactions of functional groups within a molecule in terms of “bonded” interactions. For level 1, the interaction of functional groups with α substituents is accounted for; at level 2, α and β substituents are accounted for; and at level 3, α , β , and γ substituents are accounted for. In addition, the total energy of a molecule is affected by “nonbonded” interactions between atoms or functional groups which are close together in space, even though they may be well separated in terms of bonding. Paper I also presented a method for estimating these nonbonded interactions using the fragmentation approach. A number of calculations that indicated the potential utility of this general approach to estimating the energy of organic molecules were presented. However, the range of molecules considered was very small, and so the accuracy and efficiency of the method for general organic compounds were not clearly established.

This paper presents results of tests for the accuracy and computational efficiency of the method for a set of 96 organic molecular structures which were obtained from the Cambridge Structural Database. The computational efficiency is measured as the ratio of the CPU time required to calculate the total energy of a molecule divided by the corresponding time to calculate the energies of all the required molecular fragments. The basic question addressed here is: What accuracy in energy can be obtained for what saving in computational time, for general organic molecules? The pa-

^{a)}Electronic mail: collins@rsc.anu.edu.au

^{b)}Electronic mail: vitali@rsc.anu.edu.au

per also presents some refinements of the original algorithm for fragmenting molecules and for estimating the nonbonded component of the molecular energy.

The paper is set out as follows. Section II briefly reviews the hierarchy of fragmentation methods previously described and presents further refinements. Section III describes the sample of molecular structures used to test the methods. The test results are presented in Sec. IV. Some concluding remarks are contained in the final section. Various details of the method are included in the appendices.

II. FRAGMENTATION

A detailed description of the methods used to fragment a molecule can be found in Appendix A and in Paper I. A very simple description is given here.

A. Review of the basic scheme

We view a molecule as a collection of bonded atoms in the usual chemical sense. These atoms are assembled into functional groups. Basically, atoms connected by multiple bonds are taken to be in the same functional group, as are hydrogen atoms which are bonded to these heavy atoms. Some results will be presented here to illustrate the effect of varying the criteria that define functional groups. For example, we will consider the amide group (HNCO) as one functional group (by arbitrarily defining the CN bond to be a double bond) or as two groups, NH and CO. In any event, a molecule is viewed as a collection of (single) bonded functional groups.

To illustrate the basic idea behind breaking a molecule M into fragments, we review the simple example of a chain molecule of K groups (as described in Paper I):

$$M = G_1 G_2 G_3 \cdots G_K. \quad (2.1)$$

The molecule is broken between groups $n-1$ and n , by stretching the $G_{n-1} \cdots G_n$ single bond to infinity and replacing the $G_{n-1} G_n$ bond by a hydrogen, $H^{(n-1)}$ bonded to G_{n-1} and a hydrogen bonded to G_n $H^{(n)}$. Thus

$$M \rightarrow M_1 + M_2, \quad (2.2)$$

where

$$M_1 = G_1 G_2 G_3 \cdots G_{n-1} H^{(n-1)} \quad (2.3a)$$

$$M_2 = H^{(n)} G_n G_{n+1} \cdots G_K. \quad (2.3b)$$

The geometry of the atoms in both fragments is unchanged from that in M , except for the addition of the hydrogen ‘‘caps.’’ These hydrogens are located along the broken bond vector at an appropriate distance for the particular GH bond (see Appendix A for details).

The total electronic energies of these three molecules are trivially related by

$$E(M) = E(M_1) + E(M_2) + dE_1, \quad (2.4)$$

where dE_1 represents the net energy change. This process is repeated by breaking M at some *other* place, say by stretching the $G_{i-1} G_i$ single bond to infinity. This gives

$$E(M) = E(M_3) + E(M_4) + dE_2, \quad (2.5)$$

where

$$M_2 = G_1 G_2 G_3 \cdots G_{i-1} H^{(i-1)}, \quad (2.6a)$$

$$M_4 = H^{(i)} G_i G_{i+1} \cdots G_K. \quad (2.6b)$$

Now, we break M at *both* places:

$$M \rightarrow G_1 G_2 G_3 \cdots G_{n-1} H^{(n-1)} + H^{(n)} G_n G_{n+1} \cdots G_{i-1} H^{(i-1)} \\ + H^{(i)} G_i G_{i+1} \cdots G_K. \quad (2.7)$$

This gives

$$E(M) = E(M_1) + E(M_5) + E(M_4) + dE_3, \quad (2.8)$$

where

$$M_5 = H^{(n)} G_n G_{n+1} \cdots G_{i-1} H^{(i-1)}. \quad (2.9)$$

Now, if the $G_{n-1} G_n$ bond is far in space from the $G_{i-1} G_i$ bond, then the energy change from the simultaneous fragmentation at $G_{n-1} G_n$ and $G_{i-1} G_i$ will be very nearly equal to the sum of the energy changes from each separate fragmentation (this assumes that no conjugated chain links the two bonds). That is,

$$dE_3 \approx dE_1 + dE_2. \quad (2.10)$$

The more separated the two fragmentation sites are, the closer should be the equality in Eq. (2.10). From the definitions of these energy changes,

$$E(M) - E(M_1) - E(M_5) - E(M_4) \\ \approx E(M) - E(M_1) - E(M_2) + E(M) \\ - E(M_3) - E(M_4)$$

or

$$E(M) \approx E(M_2) + E(M_3) - E(M_5). \quad (2.11)$$

Note that M_5 is the ‘‘overlapping’’ segment, common to M_2 and M_3 , and in this sense

$$M \rightarrow M_2 + M_3 - M_5. \quad (2.12)$$

The fragmentation products are defined by the location of the two bonds which are broken one at a time or both together. For a general bonding structure, as illustrated by the example in Fig. 1, we define levels of fragmentation by the following rules. Level 1: the two bonds are separated by one functional group. Level 2: the two bonds are separated by two functional groups. Level 3: the two bonds are separated by three functional groups. Clearly, one could define additional levels, in which the bond breaks are further and further apart.

1. Exhaustive fragmentation

One application of the fragmentation rule (for some levels) breaks the molecule into, say, three fragments, as in Eq. (2.12). Each fragment has an associated sign (+1 or -1). Each of these three fragments may also be broken into smaller fragments. Fragmentation at level n means that we continue to fragment each successive generation of daughter fragments, until the rule for level n cannot be applied. Many (almost all) fragments generated in this way are ‘‘canceled,’’

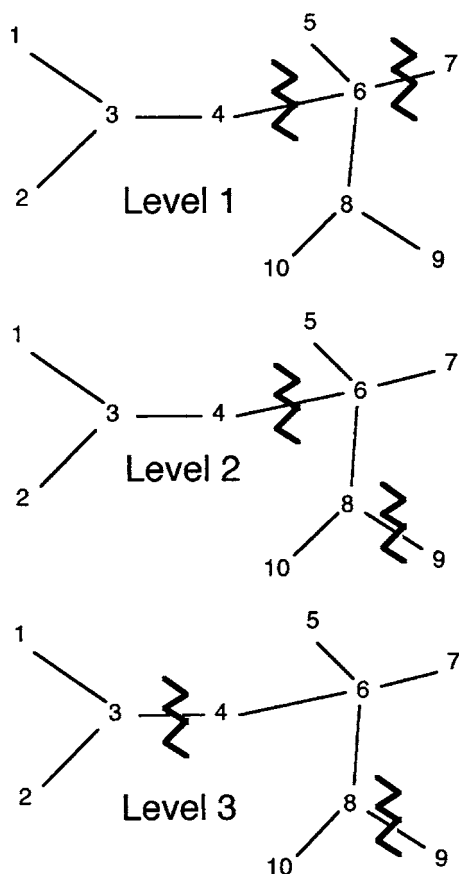


FIG. 1. A schematic representation of the locations of bonds which might be broken in the first three levels of the fragmentation hierarchy.

in the sense that the same fragment is generated twice, once with plus sign and once with a minus sign. The surviving fragments represent the fragmentation of the molecule at level n . The details are included in Appendix A. A simple linear-chain-molecule-like pentane provides the simplest example of fragmentation. The molecule is a bonded chain of five functional groups, numbered 12345. The final fragments are the following: At level 1, $12+23+34+45-2-3-4-5$, at level 2, $123+234+345-23-34$, and at level 3, $1234+2345-234$. In general, the molecule is represented by a sum (and difference) of molecular fragments as

$$M \rightarrow \sum_{n=1}^{N_1} c_n F_n, \quad (2.13)$$

where F_n represents a molecular fragment and c_n are coefficients (typically ± 1).

Moreover, the final estimate of the molecular energy (at least that due to bonding between groups) can be written as

$$E_b(M) = \sum_{n=1}^{N_1} c_n E(F_n). \quad (2.14)$$

The gradients and higher derivatives of the energy with respect to the nuclear positions can be derived from Eq. (2.14) in terms of the corresponding derivatives of the fragment energies (see Paper I for details). Moreover, any molecular property which can be expressed as a derivative of the Hamiltonian (e.g., the electrostatic moments) can be ex-

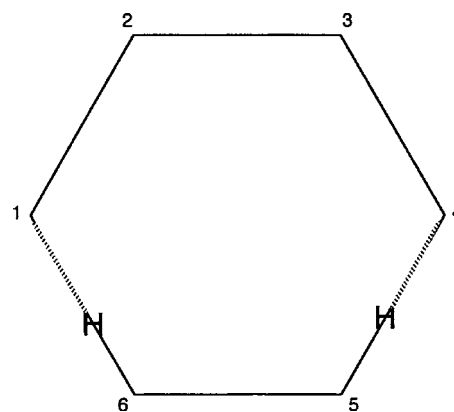


FIG. 2. A schematic representation of the location of two capping hydrogen atoms in a level 3 fragment (containing groups 1–4) in a six member cyclic molecule.

pressed as a sum of the corresponding properties of the fragments.

B. The ring repair rule

When cyclic molecules are fragmented, the capping hydrogens at each end of a fragment may actually be in close proximity. Cyclohexane provides a simple example of this effect, as illustrated in Fig. 2. This implies that there must be a significant hydrogen-hydrogen cap interaction in the fragment that was not present in the original molecule. In fact, the higher the level of fragmentation, the closer together the capping hydrogens may be. This clearly violates the basic premise of the method that the bond breaking and capping at one site should be energetically independent of the bond breaking and capping at another site. Applying the rule of level 2 to cyclohexane (denoted 123456) gives

$$\overline{123456} \rightarrow 123 + 234 + 345 + 456 + 561 + 612 - 23 - 34 - 45 - 56 - 61 - 12. \quad (2.15)$$

Level 3 gives

$$\overline{123456} \rightarrow 1234 + 2345 + 3456 + 4561 + 5612 + 6123 - 234 - 345 - 456 - 561 - 612 - 123. \quad (2.16)$$

As we see in Fig. 2, the capping hydrogens in a level 3 fragment such as 1234 are quite close to one another, much closer than they would be in a level 2 fragment such as 123. Not surprisingly, the level 2 fragmentation gives a more accurate estimate of the energy of cyclohexane than does the level 3 approximation.

Hence, the rules for levels 1–3 which were presented in Paper I are modified here by a *ring repair rule* to preclude significant spurious interactions between capping hydrogen atoms. This new rule has two parts and states.

- (1) If two groups in a fragment were both bonded (in the original molecule) to one group which is not in the fragment, then that group must be included in the fragment, or
- (2) if two groups in a fragment were bonded (in the original molecule) to two different groups which are not in

the fragment, and those two groups are bonded to each other, then those two groups must be included in the fragment.

The details of the algorithm for implementing this new rule are given in Appendix B. The effect of this new rule is to repair some broken rings in fragments to make whole rings. At all three levels, fragments containing two adjacent groups, of what were originally three and four member rings, are replaced by whole rings; at level 2, fragments containing three adjacent groups, of what were originally five member rings, are also replaced by whole rings; and at level 3, fragments containing four adjacent groups of six member rings are replaced by whole rings.

C. Nonbonded interactions

The fragmentation approximation, as in Eq. (2.14), only accounts for the electronic energy associated with functional groups and the bonding between those groups (up to interactions with γ substituents at level 3). However, functional groups that may be well separated in a molecule in terms of bonded connectivity may be close in space, since chain twisting and branching are common motifs in molecular structure. Paper I described how the nonbonded interaction between such groups can be calculated using a modification of the method of Zhang and Zhang.³

1. Review of nonbonded interactions

If M_1 and M_2 represent two distinct molecules, then the nonbonded interaction energy between them would be given by

$$E_{\text{int}}[M_1, M_2] = E[M_1 M_2] - E(M_1) - E(M_2), \quad (2.17)$$

where $E(M_1 M_2)$ represents the *ab initio* energy of both molecules, treated as one supermolecule. Let us denote a level 1 fragmentation of molecule M as

$$M \rightarrow \sum_{n=1}^{N_1} a_n f_n, \quad E(M) = \sum_{n=1}^{N_1} a_n E(f_n), \quad (2.18)$$

and a level 3 fragmentation of the same molecule as

$$M \rightarrow \sum_{m=1}^{N_3} A_m F_m, \quad E(M) = \sum_{m=1}^{N_3} A_m E(F_m). \quad (2.19)$$

The bonded interactions between groups is accounted for by the level 3 fragmentation of Eq. (2.19). We write the nonbonded interactions as

$$E_{\text{nb}}(M) = \sum_{n=1}^{N_1} \sum_{k=n+1}^{N_1} a_n a_k E_{\text{int}}[f_n^*, f_k^*]. \quad (2.20)$$

Here $E_{\text{int}}[f_n^*, f_k^*]$ is defined as in Eq. (2.17), except that the fragments f_n^* and f_k^* , must be “edited” to avoid interactions between groups that have already been accounted for in the level 3 fragmentation of Eq. (2.19) (see Paper I for details).

2. Long range interactions

From Eq. (2.20), we can infer that the number of nonbonded fragment-fragment interactions is proportional to the square of the number of level 1 fragments, N_1 . The number of fragments is linearly proportional to the number of groups in the molecule, and hence, linearly proportional to the number of electrons in the molecule. If Eq. (2.20) is used to calculate the nonbonded contribution to the total electronic energy, the total computational cost would be proportional to the square of the number of electrons in the molecule. However, in a large molecule most fragments are far separated in space. The interaction between two such well separated fragments may be negligible, or may be estimated by some means other than by *ab initio* calculation. Hence, if the separation between the fragments f_n and f_k , $d(f_n, f_k)$ (defined to be the smallest atom-atom distance), exceeds some reasonable tolerance d_{min} , the interaction energy is only evaluated approximately. We denote this approximate interaction energy as $E_{\text{app}}[f_n^*, f_k^*]$. Hence, Eq. (2.20) is amended to

$$E_{\text{nb}}(M) = \sum_{n=1}^{N_1} \sum_{k=n+1}^{N_1} a_n a_k \begin{cases} E_{\text{int}}[f_n^*, f_k^*] & \text{if } d(f_n, f_k) \leq d_{\text{min}} \\ E_{\text{app}}[f_n^*, f_k^*] & \text{if } d(f_n, f_k) > d_{\text{min}} \end{cases}. \quad (2.21)$$

In this way the total *ab initio* computational cost is linear in the number of electrons in the molecule. The total electronic energy of the molecule is then taken to be $E(M)$,

$$E(M) = E_{\text{b}}(M) + E_{\text{nb}}(M). \quad (2.22)$$

3. Electrostatic interactions

In this paper, $E_{\text{app}}[f_n^*, f_k^*]$ is estimated for well separated fragments by a multipole expansion for the electrostatic interaction of the fragments. In the examples studied here, the fragments are not charged, so that dipole-dipole interactions are the “leading” electrostatic interactions. Induction and dispersion contributions have been neglected, a factor which may contribute to the errors presented below.

We will assume that a level 1 fragmentation has been carried out and that the *ab initio* energy of each fragment has been evaluated. As part of these *ab initio* calculations, the electron density is evaluated, and at no significant additional computational cost, all the components of the electrostatic multipole moments of each fragment are calculated. The multipole moments are taken to be Cartesian tensors with components $\alpha, \beta, \gamma, \delta = 1, 2, 3$. Here we assume that, for fragment f_n , the charge, $q(n)$ and all components of the dipole moment $\mu_{\alpha}(n)$, quadrupole moment $\Theta_{\alpha\beta}(n)$, octapole moment $\Omega_{\alpha\beta\gamma}(n)$, and hexadecapole moment $H_{\alpha\beta\gamma\delta}(n)$, have all been evaluated in atomic units. The multipole expansion for f_n is taken to be centered at the center of nuclear charge for the fragment,

$$\mathbf{X}(n) = \frac{\sum_{i=1}^{N_n} c_i \mathbf{x}(i)}{\sum_{i=1}^{N_n} c_i}, \quad (2.23)$$

where c_i is the charge (atomic number) of the nucleus for atom i , $\mathbf{x}(i)$ denotes the corresponding Cartesian coordinate

vector, and the sum is taken over all atoms in the fragment. The relative displacement of fragments f_n and f_k is given by

$$\mathbf{R} = \mathbf{X}(k) - \mathbf{X}(n), \quad (2.24)$$

with magnitude denoted by R . The total electrostatic interaction between neutral fragments is given by¹⁰

$$E_{\text{el}}(n, k) = -\mu_{\alpha}(n)T_{\alpha\beta}\mu_{\beta}(k) - \frac{1}{3}[\mu_{\alpha}(n)T_{\alpha\beta\gamma}\Theta_{\beta\gamma}(k) - \Theta_{\alpha\beta}(n)T_{\alpha\beta\gamma}\mu_{\gamma}(k)] - \frac{1}{15}[\mu_{\alpha}(n)T_{\alpha\beta\gamma\delta}\Omega_{\beta\gamma\delta}(k) + \Omega_{\alpha\beta\gamma}(n)T_{\alpha\beta\gamma\delta}\mu_{\delta}(k) - \frac{5}{3}\Theta_{\alpha\beta}(n)T_{\alpha\beta\gamma\delta}\Theta_{\gamma\delta}(k)] + \frac{1}{105}[\mu_{\alpha}(n)T_{\alpha\beta\gamma\delta\epsilon}H_{\beta\gamma\delta\epsilon}(k) - H_{\alpha\beta\gamma\delta}(k)T_{\alpha\beta\gamma\delta\epsilon}\mu_{\epsilon}(k) - \frac{7}{3}\Theta_{\alpha\beta}(n)T_{\alpha\beta\gamma\delta\epsilon}\Omega_{\gamma\delta\epsilon}(k) + \frac{7}{3}\Omega_{\alpha\beta\gamma}(n)T_{\alpha\beta\gamma\delta\epsilon}\Theta_{\delta\epsilon}(k)], \quad (2.25)$$

where $T=1/R$, and

$$T_{\alpha\beta\dots\nu} = \nabla_{\alpha}\nabla_{\beta}\dots\nabla_{\nu}T. \quad (2.26)$$

Repeated indices are summed in Eq. (2.25). The multipole expansion has been truncated to neglect all terms which decay faster than R^{-6} at large separation.

In Eq. (2.21), we put

$$E_{\text{app}}[f_n^*, f_k^*] = E_{\text{el}}(n, k) \quad \text{if } d(f_n, f_k) > d_{\text{min}}. \quad (2.27)$$

In the results presented here, values of d_{min} between 3.0 and 4.5 Å have been considered. A number of important details for the evaluation of (2.27) are presented in Appendix C.

III. SAMPLE OF MOLECULES

In order to test the accuracy and efficiency of the fragmentation approximation to molecular electronic energies, a sample of “typical” organic molecules has been obtained from the Cambridge Structural Database¹¹ (CSD) as follows. The database was searched for structures containing molecules with the chemical formula $C_{7-30}N_{0-7}O_{0-7}F_{0-3}H_{1-80}$; that is, molecules containing 7–30 carbon atoms and 0–7 nitrogen atoms, and so on. These number ranges were chosen to correspond loosely with the relative abundances of carbon, nitrogen, and oxygen in various classes of organic compounds. The maximum possible number of fluorine atoms was set so that the relative proportion of F atoms might more realistically represent the proportion of all halides in typical organic molecules. Other halides were excluded to restrict the “heavy” atoms to the first row of the periodic table. This was done to restrict the expected *ab initio* quantum chemistry computation time to a manageable limit (for these tests the electronic energy of the whole molecule must be evaluated). Molecules containing the benzene ring were excluded from the search. This search returned 159 structures. Structures were subsequently excluded from the sample if they contained charged species, contained dimers of smaller molecules, rather than monomers, contained solvent molecules, or were missing some hydrogen atoms. After this editing process, the sample contained 96 single molecular structures.

The atomic Cartesian coordinates and bonding assignments of all these molecules were downloaded from the database in an SDF file format, and is available as an EPAPS document.¹²

Benzene rings and charged species were excluded from the sample for this initial extensive test of the fragmentation method. The simple fragmentation method reported in Paper I does not provide an accurate description of benzene in terms of fragments. Bettens and Lee have reported a variation of the fragmentation approach¹³ that allows partial fragmentation of conjugated rings, but this refinement has not yet been incorporated in the code tested herein. Initial investigations indicate, not surprisingly, that charged species represent additional difficulties for molecular fragmentation, and warrant a separate study.

A summary of the chemical composition of all 96 molecules is provided in an EPAPS document.¹² Figure 3 presents histograms of the number of atoms and the number

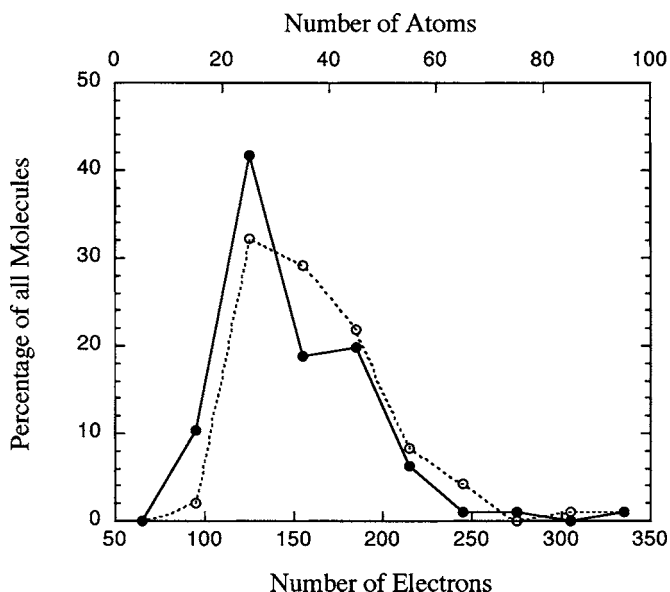


FIG. 3. The number of electrons (●) and of atoms (○) in the sample of 96 molecules are presented as histograms. The bin sizes are 30 electrons and 10 atoms.

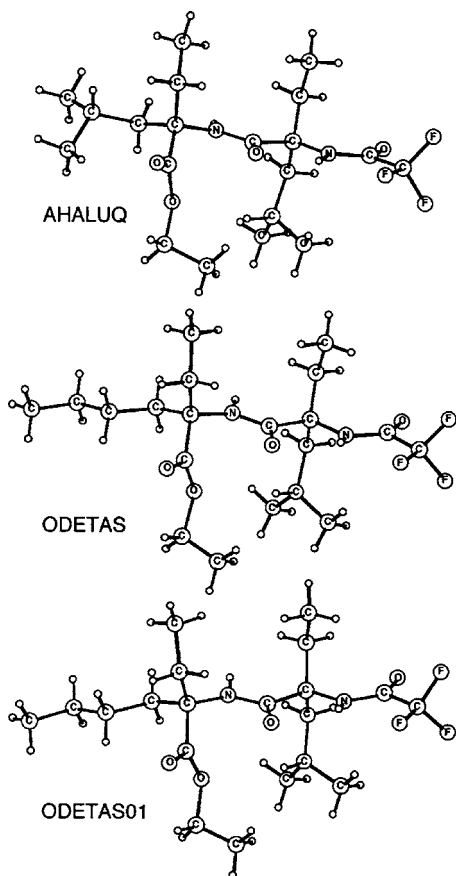


FIG. 4. Schematic representation of the structures of three isomers in the CSD, denoted as AHALUQ, ODETAS, and ODETAS01.

of electrons for the molecules in the sample. In summary, the number of atoms varies from 18 to 91 (with a mean of 37), the number of first row atoms varies from 13 to 43 (with a mean of 20), and the number of electrons varies from 92 to 334 (with a mean of 151).

Most, if not all, common functional groups are found in these molecules, including amines, amides, cyanides, ethenes, esters, ethers, aldehydes, ketones, alcohols, and carbon rings and heterocycles with four or more members. The sample is biased by the fact that all these molecules can be crystallized. Perhaps as a consequence of this fact, 76 of the 96 molecules contain at least one amide group and 39 contain a CF_3 group.

Conveniently, 22 of these molecules are isomers of at least one other molecule in the sample. Figures 4–8 depict the geometries of these 22 molecules in the database. For convenience, the structures are labeled by their CSD code name. There are 13 energy differences between these structures that correspond to isomerization energies. Hence, in addition to a sample of 96 total molecular electronic energies, these molecules provide a sample of 13 energy differences between different geometries of the same molecule. Figures 4–8 also provide some indication of the range of molecules and molecular geometries contained in the sample.

The bonding or connectivity in each molecule was evaluated as described in Eq. (A1). In every case, the bonding calculated is in agreement with that specified in the CSD

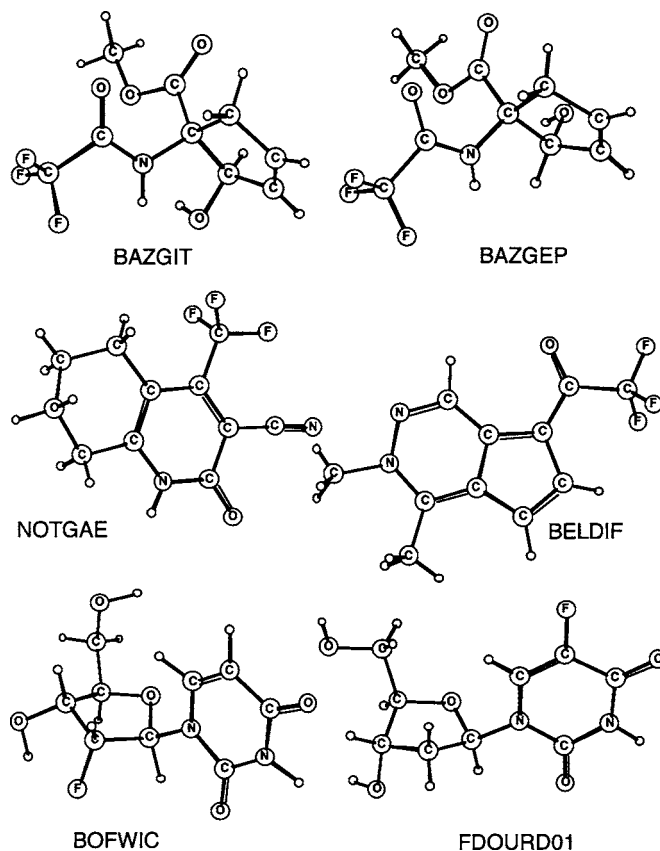


FIG. 5. Schematic representation of the structures of three pairs of isomers in the CSD, denoted as BAZGIT and BAZGEP, NOTGAE and BELDIF, and BOFWIC and FDOURD01.

SDFile for each case in that a bond exists (or does not exist) between two atoms in the molecule. If the CSD file specifies that the bond order is greater than 1, then the CSD value is adopted. Bonds with order greater than 1 are not broken in the fragmentation.

IV. RESULTS

A. Fragmentation

Fragmentation of all 96 sample molecules at levels 1–3 was carried out. The whole calculation required just over 1 s of CPU time per molecule, on average. The average size of the resultant fragments is indicated in Fig. 9, as a function of the number of heavy atoms in the molecule.

Clearly, relatively small molecules tend not to be decomposed into small fragments. The reason for this appears to be that small rings and fused rings constitute a large proportion of the total structure of these smaller systems, and the ring repair rule ensures that such rings are not completely fragmented. In fact, 2 of the 96 molecules cannot be fragmented at all at level 2, and 4 cannot be fragmented at all at level 3. These molecules are shown in Fig. 10. Even if fragmentation is possible, it may not be efficient if the molecule is mostly characterized by small ring motifs. An example of this effect is shown in Fig. 11. There the structure is dominated by a cube of carbon atoms which cannot be broken into fragments at level 2 or 3. Although the molecule can be fragmented at level 3, as shown, three of the resultant fragments contain

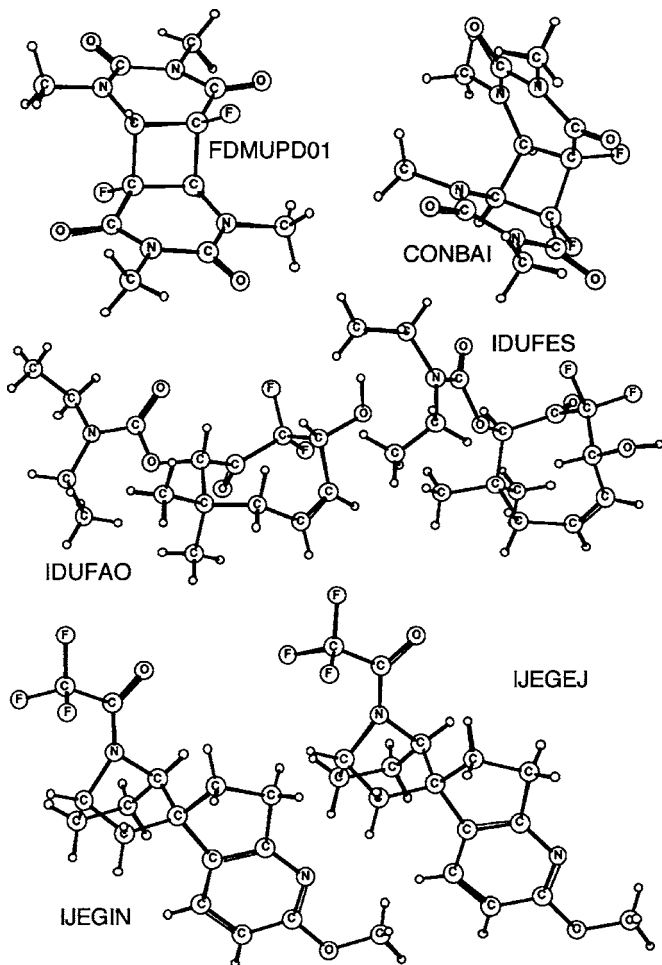


FIG. 6. Schematic representation of the structures of three pairs of isomers in the CSD, denoted as FDMUPD01 and CONBAI, IDUFAO and IDUFES, and IJEGIN and IJEJEJ.

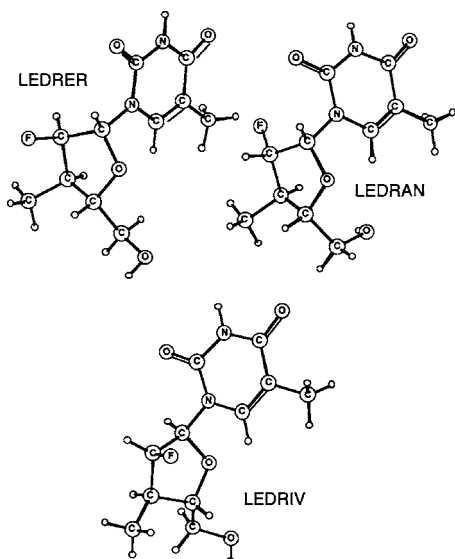


FIG. 7. Schematic representation of the structures of three isomers in the CSD, denoted as LEDRER, LEDRAN, and LEDRIV.

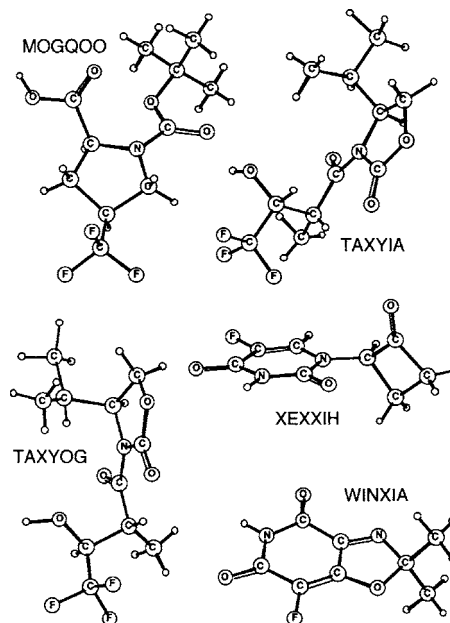


FIG. 8. Schematic representation of the structures of three isomers in the CSD denoted as MOGQOO, TAXYIA, and TAXYOG, and a pair of isomers denoted as XEXXIH and WINXIA.

almost as many heavy atoms as the original molecule. In such cases, fragmentation is not likely to lead to improved efficiency in *ab initio* calculations of the molecular energy.

However, Fig. 9 indicates that as the size of the molecule increases, the average fragment size shows a tendency to decrease. This appears to be due to the fact that small rings make up a smaller proportion of the total structure for larger molecules. It appears that the average fragment size tends towards about four heavy atoms at level 2 and about six heavy atoms at level 3. The simple examples of Paper I [see also Eq. (2.15) above] suggest that smaller fragments would result for linear and simply branched molecules, but one must remember that multiple bonds cannot be broken in the fragmentation scheme. So, for example, every carbonyl (CO)

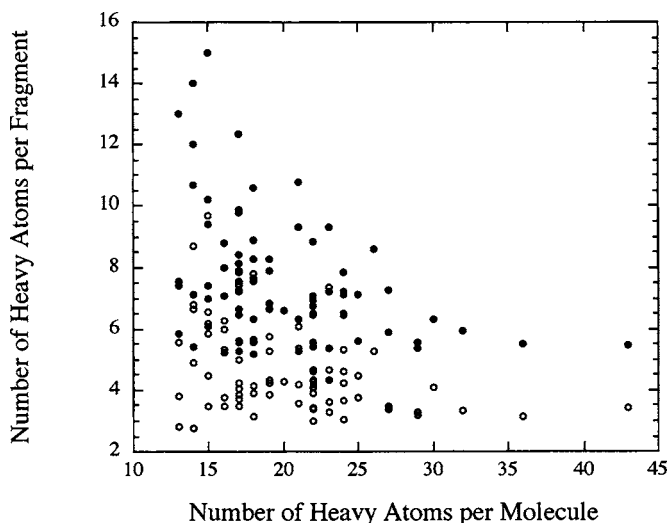


FIG. 9. The average number of nonhydrogen atoms per fragment is shown as a function of the number of nonhydrogen atoms in the molecule, for the sample of 96 molecules, with fragmentation carried out at level 2 (○) and level 3 (●).

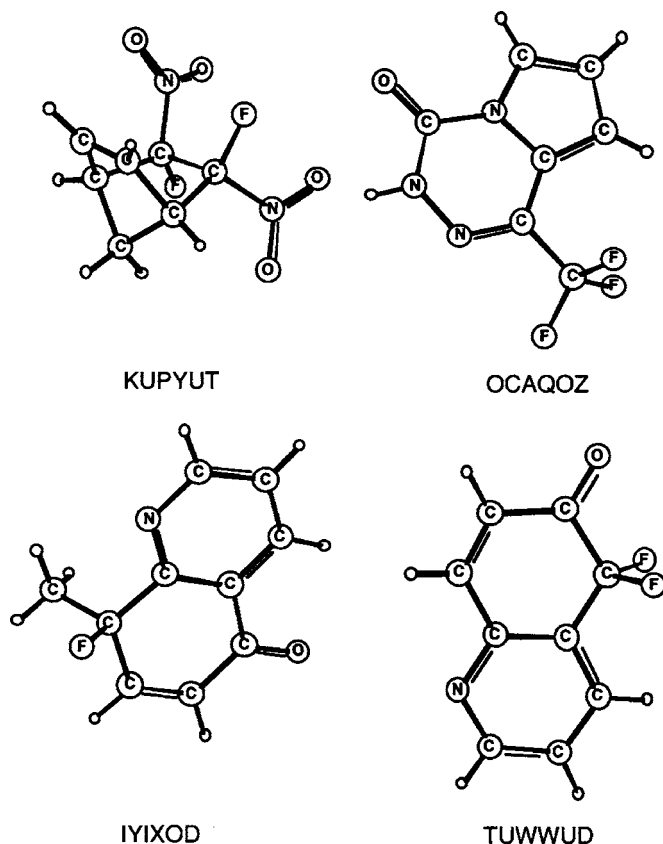


FIG. 10. The structures of the four molecules (denoted by their CSD appellation) which cannot be fragmented at level 3 are shown. At level 2, only IYIXOD and TUWWUD cannot be fragmented.

group, or C=C group, contributes two heavy atoms to each fragment which contains it, and such groups are common in these typical 96 molecules. For all 96 molecules, the average number of heavy atoms per fragment is 4.9 at level 2 and 7.6 at level 3.

In the fragmentation above, only formal multiple bonds have been preserved. The question arises whether other bonds that might be considered to have a “partial double bond” character should be preserved. For example, the amide group, —NH—CO— has been treated as two functional groups, NH and CO. It is well known that the bonding at the N atom is close to planar, consistent with a view of the CN bond as partially double in character. Hence, the fragmentation has been performed with all NC bonds in amide groups declared to be double bonds. As a result, for all 96 molecules, the average number of heavy atoms per fragment is increased to 5.9 at level 2 and 8.8 at level 3. Moreover, one might view a CF_3 group as constituting a single functional group, rather than four functional groups, on the basis that its local effect on the electronic distribution in a molecule is not an additive function of its constituents. To investigate this possibility, the fragmentation has been performed with all CF bonds in CF_3 groups treated as double bonds. As a result, for all 96 molecules, the average number of heavy atoms per fragment is increased to 5.1 at level 2 and 7.8 at level 3. If both amide NC and CF bonds in CF_3 groups are treated as double bonds, the average number of heavy atoms per fragment is increased to 6.2 at level 2 and 9.0 at level 3.

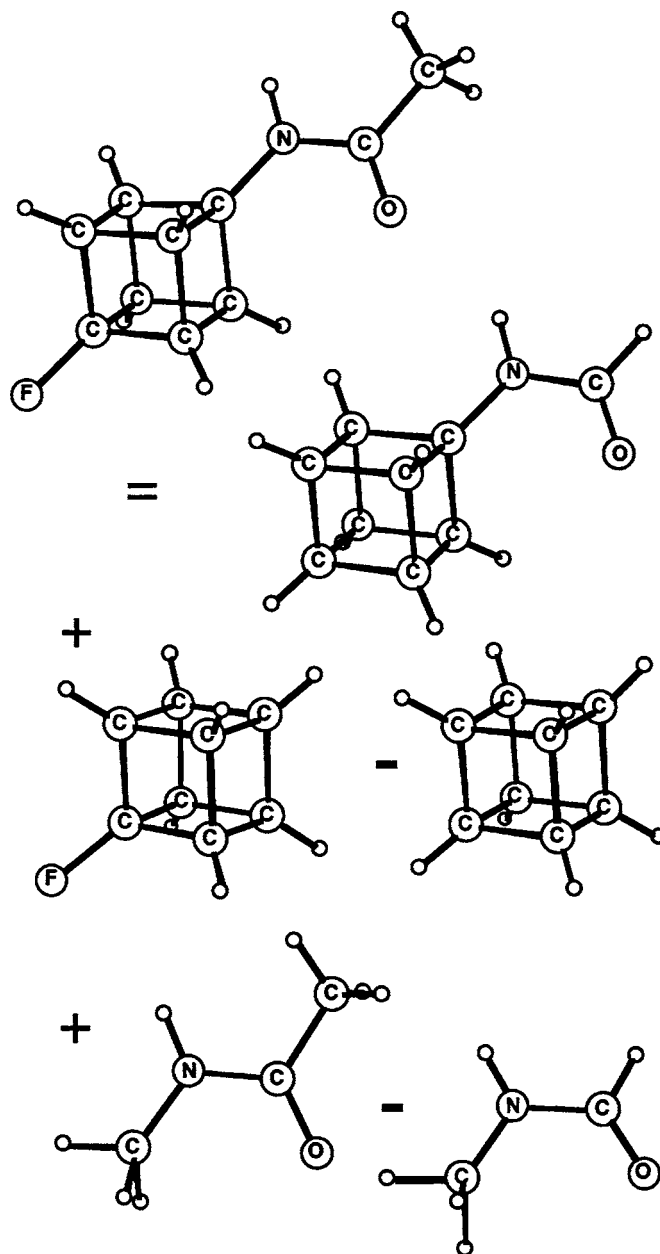


FIG. 11. A schematic representation of a molecule (denoted as GOGXIJ in the CSD) which is fragmented into large fragments at level 3 due to the presence of a cube of atoms.

B. Accuracy of the total energy

The total electronic energy of all 96 molecules has been calculated at levels 1–3 from Eq. (2.14) at the Hartree-Fock and MP2 levels of *ab initio* theory with various basis sets.¹⁴ The nonbonded contribution to the energy has also been calculated using Eqs. (2.21) and (2.26), with a d_{\min} value of 4.5 Å, and combined with the bonding energy at level 3 to give the total energy as in Eq. (2.22). The absolute differences between the fragmentation estimates and the *ab initio* energy of the whole molecule are shown in Table I. For comparison, we note that the average electronic energy for these molecules is close to -1000 hartree.

Table I shows that level 1 fragmentation provides a completely unreliable estimate of the total molecular energy, in error by hundreds of millihartrees. However, the hierarchy of

TABLE I. Absolute errors (millihartree) in the estimate of the total molecular electronic energy, averaged over the sample of 96 molecules, for various *ab initio* methods, using levels 1–3, and level 3 plus the nonbonded energy.

<i>ab initio</i> method	Bonding ^a	Level 1	Level 2	Level 3	Level 3+ nonbonding
HF/STO-3G	Standard	179	6.2	1.6	2.2
HF/STO-3G	C=N	642	4.7	1.4	1.7
HF/6-31G	Standard	152	8.1	3.8	2.8
HF/6-31G	C=N	632	6.5	3.0	1.7
HF/6-31G	CF ₃	129	7.6	3.0	2.5
HF/6-31G	C=N, CF ₃	611	5.8	2.5	1.4
HF/6-311G	Standard	146	9.0	3.9	3.0
MP2/STO-3G	Standard	190	3.2	1.5	2.3
MP2/6-31G	Standard	199	7.7	5.3	3.2

^aThe definition of which bonds are multiple bonds is standard, except that C=N indicates that CN bonds in amide groups have been defined as double bonds, and CF₃ indicates that CF bonds in CF₃ groups have been defined as double bonds for the fragmentation procedure.

fragmentation methods appears to converge very rapidly to quite reliable estimates by levels 2 and 3 and level 3 plus the nonbonded energy. The average of the errors shown in Table I are 6.5 mhartree for level 2, 2.9 mhartree for level 3, and 2.3 mhartree for level 3 plus the nonbonded energy. These last two values represent a relative error in the energy of about two to three parts in 10⁶. More importantly, these errors are consistent with the “chemical accuracy” which *ab initio* quantum chemistry aims to achieve. Comparing the “standard” results for Hartree-Fock (HF) or MP2 calculations, we see that the very small STO-3G basis set gives very small errors. However, the errors for 6-31G and 6-311G bases (on average, 213 and 309 basis functions per molecule, respectively) are very comparable. The contribution of the nonbonded interactions is most apparent for the larger bases where the error is reduced (on average) by about 1.1 mhartree for HF calculations and about 2.1 mhartree, for the MP2 calculations.

Table I does not explicitly show the long range electrostatic contribution to the calculated energies. For HF/6-31G calculations, the electrostatic energy can be as large as about 5 mhartree, for some molecules, but on average contributes only about 0.74 mhartree to the total energy. Using the MP2/6-31G electron density to calculate the multipole moments reduces the largest electrostatic energy to about 4 mhartree and reduces the average electrostatic energy to about 0.58 mhartree. However, the consequent variation of the total error in the energy, using the MP2 rather than the HF density, is found to be negligible on average over all 96 molecules. Clearly, the long range electrostatic energy can be significant for some molecules, but the utility of using the MP2, rather than the HF, electron density is less apparent. The nonbonded interactions in Table I were calculated using $d_{\min}=4.5$ Å in Eq. (2.21). However, it appears that the total nonbonded energy (the sum of the *ab initio* interactions for close fragments and electrostatic interactions for fragments separated by more than d_{\min}) is not very sensitive to the actual value of d_{\min} , so long as $d_{\min} \geq 3$. This is demonstrated in Fig. 12 for MP2/6-31G calculations, where we see that the total interaction energy for each of the 96 molecules for $d_{\min}=4.5$ Å is well correlated with the values obtained for $d_{\min}=3.0$ Å (and

greater), but not with the values obtained with $d_{\min}=2.5$ Å. The total nonbonded energy was similarly seen to be insensitive to the value of d_{\min} for HF/6-31G calculations when $d_{\min} \geq 3$.

Table I does show that treating CN bonds in amides or CF bonds in CF₃ groups, or both, as double bonds does tend to reduce the average calculated error in the energy. However, as indicated in Sec. IV A, this reduction in the error is accompanied by an increase in the average size of the fragments. We would expect an increase in fragment size to result in some reduction in the error.

C. Accuracy of isomerization energies

For most purposes in chemistry, we are concerned with the relative energies of molecular structures, rather than the total energy of a single structure. This is fortunate, since *ab initio* quantum chemistry does not normally provide accurate estimates for total electronic energies, only for relative energies. Hence, in testing whether the fragmentation procedure

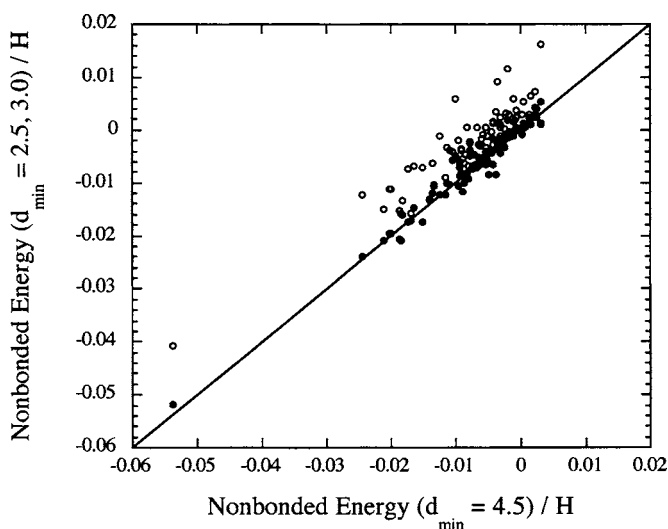


FIG. 12. The total nonbonded energy of Eq. (2.21) (with *ab initio* calculations at the MP2/6-31G level) calculated for $d_{\min}=2.5$ Å (○) and $d_{\min}=3.0$ Å (●) are plotted vs the corresponding values for $d_{\min}=4.5$ Å. The electrostatic energies were calculated using the Hartree-Fock density.

TABLE II. Isomerization energies (kJ mol^{-1}) obtained as the energy difference between structures at various levels of *ab initio* theory.

Isomers	HF STO-3G	HF 6-31G	HF 6-311G	HF 6-31G(<i>d,p</i>)	MP2 6-31G	MP2 6-31G(<i>d,p</i>)
ODETAS-AHALUQ	295.8	213.1	202.3	201.2	282.0	237.5
ODETAS01-AHALUQ	448.5	338.7	322.6	326.5	413.1	366.2
BAZGEP-BAZGIT	320.7	197.0	186.8	209.1	218.4	226.4
BELDIF-NOTGAE	723.4	599.2	577.1	602.7	623.6	574.5
FDOURD01-BOFWIC	704.2	420.1	392.0	410.6	582.0	514.1
CONBAI-FDMUPD10	133.5	120.8	116.3	116.8	105.0	94.1
IDUFES-IDUFAO	91.7	45.7	44.4	48.2	61.4	54.5
IJEJEJ-IJEGIN	16.3	9.0	9.0	6.0	14.4	10.1
LEDNAN-LEDNER	359.6	259.6	254.4	259.4	271.7	264.4
LEDNIV-LEDNER	112.9	70.4	67.0	61.1	95.8	73.8
TAXYIA-MOGQOO	1862.0	1688.7	1637.7	1654.5	1607.8	1583.2
TAXYOG-MOGQOO	891.3	810.4	786.5	794.3	774.9	740.1
WINXIA-XEXXIH	59.8	134.7	126.4	87.2	127.8	88.7

is useful for approximating *ab initio* energies to chemical accuracy, it is the relative energies of different molecular structures which are most relevant. As noted in Sec. II, the sample of 96 molecules contains 22 structures which are isomers of at least one other structure, so that 13 isomerization energies may be evaluated. Table II presents the 13 isomerization energies at various levels of theory. For this set of molecules, HF and MP2 calculations using the larger 6-31G(*d,p*) basis set have also been included. Not surprisingly, the calculated isomerization energy can depend strongly on the level of *ab initio* theory considered, varying by as much as 200 kJ mol^{-1} . Table III presents the absolute errors in these relative energies evaluated using level 3 fragmentation and accounting for the nonbonded interactions. Table IV shows the corresponding errors at level 2. Averaging the results for all *ab initio* methods shows that the isomerization energies are in error by an average of 16.4 kJ mol^{-1} at level 2 and about 3.9 kJ mol^{-1} using level 3 plus nonbonding interactions.

D. Computational efficiency

As a measure of the computational efficiency of the fragmentation approach, we will consider the ratio of the CPU

time for the calculation of the whole molecule energy to the CPU time required for the fragmentation approximation. A value of the ratio greater than 1 implies a net gain by fragmentation.

At levels 1–3, the CPU time for the fragmentation approximation is just the total CPU time to calculate the energy of all the fragments. At level 3 plus nonbonded interactions, the fragmentation CPU time also includes the time required for the *ab initio* calculation of the fragment-fragment interactions. The CPU time for electrostatic interactions of well separated fragments is negligible. However, the total fragmentation CPU time includes the time to perform *ab initio* calculations on all the level 1 fragments. These calculations are required to evaluate the multipole moments needed for the electrostatic calculations.

Figure 13 presents the time ratio for all 96 molecules in the sample, using the MP2/6-31G level of *ab initio* theory, evaluated for level 2 and for level 3 plus nonbonded interactions. In the latter calculations, we have taken $d_{\min}=3.0$. The CPU time for the level 3 calculations is much larger on average than the CPU time for the nonbonded interactions. However, reducing the value of d_{\min} from 4.5 to 3.0 Å reduces the *ab initio* CPU time for the nonbonded interactions

TABLE III. Absolute errors in the isomerization energies of Table II (kJ mol^{-1}) evaluated using level 3 fragmentation plus nonbonded interactions and the standard definition of bond order.

Isomers	HF STO-3G	HF 6-31G	HF 6-311G	HF 6-31G(<i>d,p</i>)	MP2 6-31G	MP2 6-31G(<i>d,p</i>)
ODETAS-AHALUQ	2.3	1.9	2.9	3.2	3.3	4.8
ODETAS01-AHALUQ	3.3	1.1	0.1	0.6	0.8	2.8
BAZGEP-BAZGIT	6.0	5.6	5.5	4.0	5.6	3.4
BELDIF-NOTGAE	5.1	8.1	5.7	5.3	7.3	4.2
FDOURD01-BOFWIC	0.5	0.6	2.7	1.5	0.5	2.5
CONBAI-FDMUPD10	2.8	0.8	1.1	1.6	0.7	1.8
IDUFES-IDUFAO	1.6	6.9	5.2	1.2	9.8	9.8
IJEJEJ-IJEGIN	2.0	4.1	4.8	3.8	4.5	3.3
LEDNAN-LEDNER	0.5	5.4	4.1	7.7	8.6	11.1
LEDNIV-LEDNER	2.3	5.3	6.1	0.7	1.2	4.3
TAXYIA-MOGQOO	14.3	11.8	7.9	3.1	0.7	8.9
TAXYOG-MOGQOO	9.9	1.8	0.6	3.0	1.7	9.6
WINXIA-XEXXIH	0.5	1.4	1.4	1.0	0.6	0.6

TABLE IV. Absolute errors in the isomerization energies of Table II (kJ mol^{-1}) evaluated using level 2 fragmentation and the standard definition of bond order.

Isomers	HF STO-3G	HF 6-31G	HF 6-311G	HF 6-31G(<i>d,p</i>)	MP2 6-31G	MP2 6-31G(<i>d,p</i>)
ODETAS-AHALUQ	6.6	10.9	9.5	7.1	13.0	11.3
ODETAS01-AHALUQ	9.0	11.5	9.8	7.4	13.7	12.3
BAZGEP-BAZGIT	10.6	18.8	18.8	17.1	15.8	16.4
BELDIF-NOTGAE	29.9	86.7	84.9	68.3	59.2	43.6
FDOURD01-BOFWIC	6.2	9.8	12.3	8.0	4.6	0.6
CONBAI-FDMUPD10	18.8	50.8	51.0	46.0	28.5	23.4
IDUFES-IDUFAO	4.7	0.8	0.9	3.3	4.3	3.9
IJEGEJ-IJEGIN	0.1	0.7	0.8	0.1	0.2	0.6
LEDNAN-LEDNER	12.3	20.3	18.3	16.7	21.3	23.1
LEDNIV-LEDNER	5.9	17.8	18.2	15.7	9.3	9.9
TAXYIA-MOGQOO	10.9	9.8	9.4	0.6	15.2	29.1
TAXYOG-MOGQOO	3.0	17.8	17.0	6.0	3.3	14.7
WINXIA-XEXXIH	9.3	20.2	21.1	19.9	5.7	4.5

by a factor of about 5 and thereby reduces the nonbonded CPU time to an insignificant component of the total CPU time. As we have seen above, using $d_{\min}=3.0$ does not significantly affect the accuracy of the approximation.

Figure 13 indicates that the fragmentation approximation is not efficient for the smaller molecules in the sample. This observation is not surprising since Fig. 9 showed that fragmentation of these smaller molecules was ineffective in producing small fragments. However, Fig. 13 shows that as the size of the molecule increases, there is a clear probability of obtaining significant computational efficiency from the fragmentation approach.

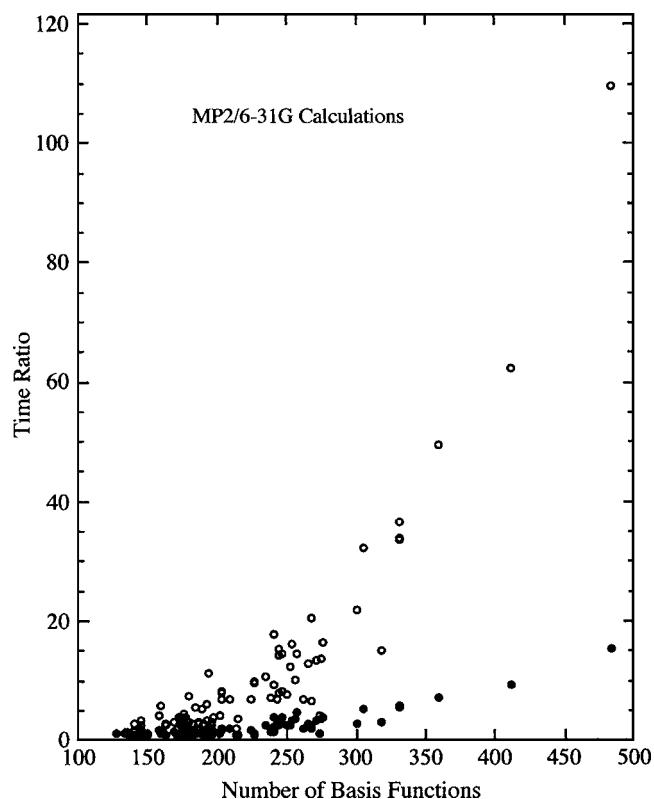


FIG. 13. The ratios of the CPU time for a MP2/6-31G calculation of the whole molecule to the CPU time for the corresponding calculations at level 2 (○) and at level 3 plus nonbonded contributions (●) are shown vs the number of basis functions for the sample of 96 molecules.

V. CONCLUDING REMARKS

We have amended the systematic fragmentation method of Paper I by including a ring repair rule which ensures that cyclic segments of molecules are accurately treated. An electrostatic approximation for long range nonbonded interactions between fragments ensures that the computational time scales only linearly with the size of the molecule.

Long range induction and dispersion effects have been ignored, but might be included via a self-consistent calculation of the induced dipoles (see p. 134 of Ref. 10). Moreover, it may be possible to obtain an approximation to the dispersion energy component of the nonbonded interactions using dispersion coefficients based on the dipole polarizabilities (see p. 56 of Ref. 10.)

The hierarchy of fragmentation approximations have been tested on a large set of typical organic molecules, which contain most of the common types of organic functional groups, except benzene rings and charged species. We have seen that the level 3 approximation, with account for nonbonded interactions, provides an estimate of the total electronic energy, and (most importantly) relative energies of different structures, which is consistent with the chemical accuracy (a few kJ mol^{-1}) which *ab initio* quantum chemistry aims to achieve. Even the level 2 approximation provides a useful estimate of these quantities. As noted above (and in detail in Paper I), energy gradients and higher derivatives of the energy can be obtained from the fragmentation approximation. Hence, geometry optimization has been accurately performed for small molecules in Paper I (Ref. 9) and for a large complex by Chen and Zhang.¹⁵ Vibrational frequencies and other molecular properties can be obtained. Thus far, gradients have been calculated within the fragmentation approximation and geometry optimization performed without inclusion of the nonbonded interactions. Further work is required to include the forces due to these nonbonded interactions in geometry optimization.

Based on the results presented here, it is reasonable to conclude that relatively high levels of *ab initio* quantum chemistry can be accurately and efficiently applied to many

relatively large organic molecules using this fragmentation approach.

ACKNOWLEDGMENTS

The authors thank the Australian Partnership for Advanced Computing for a generous allocation of computer time. One of the authors (M.A.C.) wishes to thank Professor Peter M. Gill for helpful comments and discussions.

APPENDIX A: FRAGMENTATION

Fragmenting a molecule to estimate the electronic energy involves a number of steps.

1. Definition of bonding

The bonding in a molecule is defined as follows.

- (i) The covalent radius of all the elements are read from a table (see EPAPS documents), so that the covalent radius $\text{rad}(i)$ of each atom i is assigned.
- (ii) All atom-atom distances in the molecule, r_{ij} , are calculated. If

$$r_{ij} < \text{rad}(i) + \text{rad}(j) + 0.4 \text{ \AA}, \quad (\text{A1})$$

then a bond exists between atoms i and j , with bond order 1. A record of this bonding is required when the positions of ‘‘capping hydrogen’’ atoms are determined.

- (iii) A bond order greater than 1 is assigned if
 - (a) (in these tests) the CSD file for this molecule assigns a bond order greater than 1 for this pair of atoms, or
 - (b) the bond order is arbitrarily set to a value greater than 1 by choice. For example, in these tests we examined the effect of setting all CN bonds in amide groups as double bonds, and/or all CF bonds in CF_3 groups as double bonds.

Bonds of order greater than 1 are never broken in the fragmentation procedure.

2. Definition of functional groups and group bonding

The atoms in a molecule are assigned to disjoint functional groups as follows.

- (1) Consider first only nonhydrogen atoms. The first such atom is assigned to group 1. This is defined to be the current group.
 - (i) Consider the rest of the nonhydrogen atoms in turn. An unassigned atom which is bonded to an atom in the current group, with bond order greater than one, is assigned to the current group.
 - (ii) The next unassigned atom is assigned to a new group, now denoted as the current group. Go to (i).
- (2) All nonhydrogen atoms have now been assigned to a group. All hydrogen atoms are assigned to be members of the same functional group as the nonhydrogen atom to which they are bonded.

- (3) If a single bond (order 1) exists between two atoms in different groups, a single bond is defined to exist between the two groups. All such atom-atom bonds are considered in order to determine all bonds between groups.

3. The properties of fragments

The molecule is now defined to be a set of ngroup functional groups connected by bonds. The fragmentation of the molecule is carried out by breaking the bonds between these groups. A molecular fragment is a disjoint, compact subset of these groups. *Compact* means that if groups i and j are in a fragment, then there is at least one sequence of bonds, $i \cdots k \cdots m \cdots n \cdots j$, that connects i and j . *Disjoint* means that only such connected groups belong to a fragment.

The connectivity matrix for the molecule is defined by the link array,

$$\begin{aligned} \text{link}(i,j) &= 1, & \text{if } i \text{ is bonded to } j; & \quad i, j = 1, \dots, \text{ngroup} \\ &= 0, & \text{otherwise.} & \end{aligned} \quad (\text{A2})$$

Let

$$\mathbf{M} = \text{link} + \text{link}^2 + \cdots + \text{link}^{\text{ngroup}-1}. \quad (\text{A3})$$

Then $\mathbf{M}(i,j) \neq 0$ if and only if groups i and j are in the same fragment.¹⁶

Similarly, a fragment of nfrag groups, denoted as $F(k)$, is assigned a connectivity matrix, $\mathbf{L}_k(i,j)$, for $i, j = 1, \dots, \text{nfrag}$. For each group in a fragment, the total number of bonds connected to it is easily evaluated from the fragment connectivity matrix. The *connectivity* of group i in fragment k , $\text{con}_k(i)$, is defined to be the number of bonds connected to i in fragment k .

4. Breaking bonds

Breaking the $n \cdot m$ bond in fragment $F(k)$ is equivalent to setting

$$\mathbf{L}_k(n,m) = \mathbf{L}_k(m,n) = 0. \quad (\text{A4})$$

Breaking any bond may result in a fragment being broken into new disjoint fragments. By evaluating

$$\mathbf{M}_k = \mathbf{L}_k + \mathbf{L}_k^2 + \cdots + \mathbf{L}_k^{\text{nfrag}-1}, \quad (\text{A5})$$

the disjoint fragments may be identified as unconnected blocks in \mathbf{M}_k .

5. Iterative fragmentation

To fragment a molecule at a given level of fragmentation, we begin with the whole molecule as fragment 1, denoted as $F(1)$. Fragment 1 is said to be *live* with a sign, $\text{sign}(1)$, of +1.

The iterative ten step procedure is as follows.

- (1) Find the lowest numbered live fragment; denote it as $F(n)$. If no live fragment exists in the current list of fragments, the iteration stops.
- (2) Choose two bonds to break for the level used, as described below. If no pair of allowed bond exists, declare $F(n)$ as *dead* and go to (1).
- (3) Break only the first bond (of two) by setting the two relevant elements of \mathbf{L}_n to zero, and calculate the sets of atoms in all disjoint fragments.
- (4) Create a new live fragment for each disjoint set, with $\text{sign}+1$ times the sign of $F(n)$, and add each to the end of the list of fragments.
- (5) Break only the second bond by setting the two relevant elements of \mathbf{L}_n to zero, and calculate the sets of atoms in all disjoint fragments.
- (6) Create a new live fragment for each disjoint set, with $\text{sign}+1$ times the sign of $F(n)$, and add each to the end of the list of fragments.
- (7) Break both bonds identified in (2) by setting the four relevant elements of \mathbf{L}_n to zero, and calculate the sets of atoms in all disjoint fragments.
- (8) Create a new live fragment for each disjoint set, with $\text{sign}-1$ times the sign of $F(n)$, and add each to the end of the list of fragments.
- (9) Delete fragment $F(n)$ from the list of fragments.
- (10) Go to 1.

Termination: All remaining fragments are dead.

6. Canceling fragments

When this iterative process is complete, any pair of fragments that are composed of the same atoms, but which have opposite signs, are deleted, leaving n_{final} fragments.

7. Capping hydrogens

In each final fragment, the bonds which connected groups in the fragment to groups which are not in the fragment are “missing.” These missing bonds are replaced by bonds to hydrogen atoms. As discussed in Paper I, the total number of hydrogen atoms added to fragments with a sign of $+1$ is exactly equal to the number added to fragments with a sign of -1 . The position of each H atom is taken to lie along the missing bond vector at a distance which is proportional to the expected ratio of bond lengths. That is,

$$\mathbf{x}(\text{H}) = \mathbf{x}(i) + \frac{\text{rad}(i) + \text{rad}(\text{H})}{\text{rad}(i) + \text{rad}(j)} [\mathbf{x}(j) - \mathbf{x}(i)], \quad (\text{A6})$$

where $\mathbf{x}(i)$ denotes the Cartesian position of the atom in the fragment and $\mathbf{x}(j)$ denotes the Cartesian position of the atom *not* in the fragment.

8. *Ab initio* calculations

Now that the positions of all atoms in each fragment k are determined, *ab initio* calculations of the energy, $E(k)$, of each fragment can be carried out. Each fragment is neutral and is taken to be in a singlet spin state.

9. The fragmentation approximation

The molecular energy is approximated at this level by

$$E = \sum_{k=1}^{n_{\text{final}}} \text{sign}(k)E(k). \quad (\text{A7})$$

Similarly, any molecular property which can be evaluated by applying a linear operator to the electronic molecular Hamiltonian can be estimated in a similar fashion. For example, the dipole moment μ of a neutral molecule can be calculated as the expectation value of the derivative of the total Hamiltonian with respect to an electric field \mathbf{F} evaluated at $\mathbf{F}=0$. Thus,

$$\mu = \sum_{k=1}^{n_{\text{final}}} \text{sign}(k)\mu(k) \quad (\text{A8})$$

estimates the total molecular dipole moment as a combination of the dipoles of the fragments. Approximations for the higher multipole moments and other properties can be derived.

10. Choosing bonds to break

We choose which bonds to break in a fragment, $F(n)$, by searching for two bonds which obey the rule for the level of fragmentation considered. The search considers all possibilities in order until either two allowed bonds are found or the fragment cannot be broken, when in step (2) of the iterative fragmentation procedure (see above), $F(n)$ is declared to be dead. The choosing process for level 3 is described as the method is obvious (and simpler) for the lower levels.

At level 3, the searching procedure (a nested process) identifies five atoms we call $n1, n2, n3, n4$, and $n5$ which are bonded in that order: $n1 \cdots n2 \cdots n3 \cdots n4 \cdots n5$. The bonds ($n1 \cdots n2$) and ($n4 \cdots n5$) will be broken in the fragmentation procedure, as they are separated by three groups (the level 3 condition). However, for example, $n1$ cannot be bonded to $n5$, or else the required separation would not be satisfied. The searching process accounts for these bonding and non-bonding requirements.

We begin with the central atom, $n3$. First we declare that all groups are allowed to be any of $n1, \dots, n5$. Then, the following is followed.

- (1) Find the lowest numbered group allowed to be $n3$, which has the highest connectivity and is not a penultimate group (it is bonded to at least two groups which have connectivity greater than 1). Assign this as group $n3$. If no group is possible (satisfies the criteria to be assigned as $n3$), then the search has failed and is finished, or else go on.
- (2) Find the lowest numbered group allowed to be $n2$, which has the highest connectivity and is bonded to $n3$. Assign this group as $n2$. If this is impossible, disallow the current $n3$ as a possible value for $n3$, allow all values for $n1, n2, n4$, and $n5$, and go to (1).
- (3) Find the lowest numbered group allowed to be $n4$, which has the highest connectivity and is bonded to $n3$, and is not $n2$. Assign this as group as $n4$. If this is

impossible, disallow the current $n2$ as a possible value for $n2$, allow all values for $n1$, $n4$, and $n5$, and go to (2).

- (4) Find the lowest numbered group allowed to be $n1$, which has the lowest connectivity and is bonded to $n2$, but is not $n3$ nor bonded to $n4$. Assign this as group as $n1$. If this is impossible, disallow the current $n4$ as a possible value for $n4$, allow all values for $n1$ and $n5$, and go to (3).
- (5) Find the lowest numbered group, which has the lowest connectivity and is bonded to $n4$, but is not $n3$ nor bonded to $n2$ or to $n1$. Assign this group as $n5$. If this is impossible, disallow the current $n1$ as a possible value for $n1$, allow all values for $n5$, and go to (4), or else go on.
- (6) If the search is finished with assigned values for $n1, \dots, n5$, then proceed to break the appropriate bonds in the iterative fragmentation scheme.

APPENDIX B: FRAGMENT REPAIR

The ring repair rule is implemented as follows.

- (i) The fragmentation is completed as detailed in Appendix A.
- (ii) The resultant fragments are sorted in decreasing order of the number of groups.
- (iii) Beginning with the first fragment, we determine
 - (1) if two groups in a fragment were both bonded (in the original molecule) to one group which is not in the fragment, or
 - (2) if two groups in a fragment were bonded (in the original molecule) to two different groups which are not in the fragment, and those two groups are bonded to each other. If neither condition holds, we consider the next fragment, and so on. If the conditions hold for no fragment, we are finished, or else we call the first fragment for which the condition holds, $F(M)$.
- (iv) We construct a new fragment which is composed of $F(M)$ plus the missing (one or two) groups, and call this fragment $F^*(M)$.
- (v) We then consider if conditions (1) and (2) in (iii) hold for $F^*(M)$. If not, we proceed to step (vi). If so, we construct a new fragment which is composed of $F^*(M)$ plus the missing (one or two) groups, and again call this fragment $F^*(M)$. Step (v) is repeated as many times as necessary.
- (vi) We create a second new fragment, denoted as $-F^*(M)$, with the same composition as $F^*(M)$ but with the opposite sign. The fragment $F^*(M)$ is declared to be dead, while the fragment $-F^*(M)$ is defined to be alive. Both $F^*(M)$ and $-F^*(M)$ are added to the list of fragments, and the fragmentation algorithm of Appendix A is repeated.

When the fragmentation is repeated, the live fragment $-F^*(M)$ is decomposed according to the level employed

[$F^*(M)$ is not decomposed, since it is dead]. The process of canceling fragments is carried out. We then return to step (ii) above.

APPENDIX C: ELECTROSTATICS

The charge and multipole moments of each level 1 fragment are calculated by an *ab initio* quantum chemistry program package. It is important to note the following.

- (1) The multipole moments for the fragment charge distribution depend on the level of *ab initio* theory employed to calculate the electron density. It may be, for example, that even though the fragment energy is evaluated at a highly correlated level of theory, the multipole moments are determined by some lower level of theory (for example, at the Hartree-Fock level). Hence, the electrostatic interaction energy between fragments is actually an approximation only to the exact interaction energy at the lower level of theory.
- (2) The values given by the program output are converted to a.u. and the definitions of the multipole moments are made consistent with those of Ref. 10, p. 19.
- (3) The multipole moments may have been evaluated for a geometry which has been translated and rotated into some default orientation preferred by the *ab initio* program. If so, the multipole moments must be rotated to the orientation that the fragment has in the original molecule. The center of the fragment [see Eq. (2.22)] must correspond to its position in the original molecule. The rotation matrix required to effect this reorientation can be evaluated from a knowledge of the original Cartesian coordinates and those used by the *ab initio* program.
- (4) The editing procedure that replaces fragment f_n by f_n^* in Eq. (2.20) may produce a fragment, f_n^* , that is not one of the level 1 fragments, so multipole moments for f_n^* must be evaluated. The only such “nonlevel 1 fragments” that result from the editing process are fragments that contain one terminal group (groups that are only connected to one other group). For simplicity, let us assume that fragment A , containing, say, groups 1 and 2, has been edited to fragment B , containing only group 1. A fragment, call it C , containing just group 2 will exist at level 1. From Eq. (A7), the fragmentation approximation to the molecular electronic energy also supports a fragmentation approximation to other properties, including the electrostatic moments. Hence, we can approximate the unknown dipole moment of fragment B by

$$\mu_\alpha(B) = \mu_\alpha(A) - \mu_\alpha(C). \quad (\text{C1})$$

The higher moments can be estimated as

$$\Theta_{\alpha\beta}(B) = \Theta_{\alpha\beta}(A) - \Theta_{\alpha\beta}(C) - [X_{\alpha}(B) - X_{\alpha}(A)]\mu_{\beta}(C) - [X_{\beta}(B) - X_{\beta}(A)]\mu_{\alpha}(C) - [X_{\alpha}(C) - X_{\alpha}(A)]\mu_{\beta}(B) - [X_{\beta}(C) - X_{\beta}(A)]\mu_{\alpha}(B) \quad (C2)$$

$$\Omega_{\alpha\beta\gamma}(B) = \Omega_{\alpha\beta\gamma}(A) - \Omega_{\alpha\beta\gamma}(C) - [X_{\alpha}(B) - X_{\alpha}(A)]\Theta_{\beta\gamma}(C) - [X_{\beta}(B) - X_{\beta}(A)]\Theta_{\alpha\gamma}(C) - [X_{\gamma}(B) - X_{\gamma}(A)]\Theta_{\alpha\beta}(C) - [X_{\alpha}(C) - X_{\alpha}(A)]\Theta_{\beta\gamma}(B) - [X_{\beta}(C) - X_{\beta}(A)]\Theta_{\alpha\gamma}(B) - [X_{\gamma}(C) - X_{\gamma}(A)]\Theta_{\alpha\beta}(B) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\beta}(B) - X_{\beta}(A)]\mu_{\gamma}(C) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\mu_{\beta}(C) - [X_{\beta}(B) - X_{\beta}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\mu_{\alpha}(C) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\beta}(B) - X_{\beta}(A)]\mu_{\gamma}(B) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\mu_{\beta}(B) - [X_{\beta}(C) - X_{\beta}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\mu_{\alpha}(B), \quad (C3)$$

$$H_{\alpha\beta\gamma\delta}(B) = H_{\alpha\beta\gamma\delta}(A) - H_{\alpha\beta\gamma\delta}(C) - [X_{\alpha}(B) - X_{\alpha}(A)]\Omega_{\beta\gamma\delta}(C) - [X_{\beta}(B) - X_{\beta}(A)]\Omega_{\alpha\gamma\delta}(C) - [X_{\gamma}(B) - X_{\gamma}(A)]\Omega_{\alpha\beta\delta}(C) - [X_{\delta}(B) - X_{\delta}(A)]\Omega_{\alpha\beta\gamma}(C) - [X_{\alpha}(C) - X_{\alpha}(A)]\Omega_{\beta\gamma\delta}(B) - [X_{\beta}(C) - X_{\beta}(A)]\Omega_{\alpha\gamma\delta}(B) - [X_{\gamma}(C) - X_{\gamma}(A)]\Omega_{\alpha\beta\delta}(B) - [X_{\delta}(C) - X_{\delta}(A)]\Omega_{\alpha\beta\gamma}(B) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\beta}(B) - X_{\beta}(A)]\Theta_{\gamma\delta}(C) - [X_{\alpha}(B) - X_{\alpha}(A)]\times[X_{\gamma}(B) - X_{\gamma}(A)]\Theta_{\beta\delta}(C) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\delta}(B) - X_{\delta}(A)]\Theta_{\beta\gamma}(C) - [X_{\beta}(B) - X_{\beta}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\Theta_{\alpha\delta}(C) - [X_{\beta}(B) - X_{\beta}(A)][X_{\delta}(B) - X_{\delta}(A)]\Theta_{\alpha\gamma}(C) - [X_{\gamma}(B) - X_{\gamma}(A)][X_{\delta}(B) - X_{\delta}(A)]\Theta_{\alpha\beta}(C) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\beta}(C) - X_{\beta}(A)]\Theta_{\gamma\delta}(B) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\gamma}(C) - X_{\gamma}(A)]\Theta_{\beta\delta}(B) - [X_{\alpha}(C) - X_{\alpha}(A)]\times[X_{\delta}(C) - X_{\delta}(A)]\Theta_{\beta\gamma}(B) - [X_{\beta}(C) - X_{\beta}(A)][X_{\gamma}(C) - X_{\gamma}(A)]\Theta_{\alpha\delta}(B) - [X_{\beta}(C) - X_{\beta}(A)][X_{\delta}(C) - X_{\delta}(A)]\Theta_{\alpha\gamma}(B) - [X_{\gamma}(C) - X_{\gamma}(A)][X_{\delta}(C) - X_{\delta}(A)]\Theta_{\alpha\beta}(B) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\beta}(B) - X_{\beta}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\mu_{\delta}(C) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\beta}(B) - X_{\beta}(A)][X_{\delta}(B) - X_{\delta}(A)]\mu_{\gamma}(C) - [X_{\alpha}(B) - X_{\alpha}(A)][X_{\gamma}(B) - X_{\gamma}(A)]\times[X_{\delta}(B) - X_{\delta}(A)]\mu_{\beta}(C) - [X_{\beta}(B) - X_{\beta}(A)][X_{\gamma}(B) - X_{\gamma}(A)][X_{\delta}(B) - X_{\delta}(A)]\mu_{\alpha}(C) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\beta}(C) - X_{\beta}(A)][X_{\gamma}(C) - X_{\gamma}(A)]\mu_{\delta}(B) - [X_{\alpha}(C) - X_{\alpha}(A)][X_{\beta}(C) - X_{\beta}(A)][X_{\delta}(C) - X_{\delta}(A)]\mu_{\gamma}(B) - [X_{\alpha}(C) - X_{\alpha}(A)]\times[X_{\gamma}(C) - X_{\gamma}(A)][X_{\delta}(C) - X_{\delta}(A)]\mu_{\beta}(B) - [X_{\beta}(C) - X_{\beta}(A)][X_{\gamma}(C) - X_{\gamma}(A)][X_{\delta}(C) - X_{\delta}(A)]\mu_{\alpha}(B) \quad (C4)$$

¹W. Yang, Phys. Rev. A **44**, 7823 (1991); Phys. Rev. Lett. **66**, 1438 (1991); W. Yang and T. Lee, J. Chem. Phys. **103**, 5674 (1995); S. Liu, J. M. Pérez-Jordá, and W. Yang, *ibid.* **112**, 1634 (2000); N. Flocke and R. J. Bartlett, *ibid.* **121**, 10935 (2004); T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison, and G. Beylkin, *ibid.* **121**, 6680 (2004); C. K. Gan, G. J. Tymczak, and M. Challacombe, *ibid.* **121**, 6608 (2004); X. Li, J. M. Millam, G. E. Scuseria, M. J. Frisch, and H. B. Schlegel, *ibid.* **119**, 7651 (2003); G. E. Scuseria, J. Phys. Chem. A **103**, 4782 (1999); W. Liang, C. Saravanan, Y. Shao, R. Baer, A. T. Bell, and M. Head-Gordon, J. Chem. Phys. **119**, 4117 (2003); H.-J. Werner, F. R. Manby, and P. J. Knowles, *ibid.* **118**, 8149 (2003); P. Ordejon, Comput. Mater. Sci. **12**, 157 (1998); S. Goedecker, Rev. Mod. Phys. **71**, 1087 (1999); S. Dixon and K. Merz, Jr., J. Chem. Phys. **104**, 6643 (1996).

²W. Li and S. Li, J. Chem. Phys. **122**, 194109 (2005); W. Li, T. Fang, and S. Li, *ibid.* **124**, 154102 (2006); D. G. Fedorov and K. Kitaura, *ibid.* **121**, 2483 (2004); Chem. Phys. Lett. **389**, 129 (2004).

³D. W. Zhang and J. Z. H. Zhang, J. Chem. Phys. **119**, 3599 (2003).

⁴Y. Mei, D. W. Zhang, and J. Z. H. Zhang, J. Phys. Chem. A **109**, 2 (2005).

⁵X. H. Chen, D. W. Zhang, and J. Z. H. Zhang, J. Chem. Phys. **120**, 839

(2004); X. H. Chen and J. Z. H. Zhang, *ibid.* **120**, 11386 (2004); D. W. Zhang and J. Z. H. Zhang, J. Theor. Comput. Chem. **3**, 43 (2004).

⁶S. R. Gadre, R. N. Shirsat, and A. C. Limaye, J. Phys. Chem. **98**, 9165 (1994); K. Babu and S. R. Gadre, J. Comput. Chem. **24**, 484 (2003).

⁷X. He and J. Z. H. Zhang, J. Chem. Phys. **122**, 031103 (2005); X. Chen, Y. Zhang, and J. Z. H. Zhang, *ibid.* **122**, 184105 (2005).

⁸S. Li, W. Li, and T. Fang, J. Am. Chem. Soc. **127**, 7215 (2005).

⁹V. Deev and M. A. Collins, J. Chem. Phys. **122**, 154102 (2005).

¹⁰A. J. Stone, *The Theory of Intermolecular Forces* (Clarendon, Oxford, 1996).

¹¹F. H. Allen, Acta Crystallogr., Sect. B: Struct. Sci. **B58**, 380 (2002).

¹²See EPAPS Document No. E-JCPA6-125-305635 for the molecular coordinates and related data file. This document can be reached via a direct link in the online article's HTML reference section or via the EPAPS homepage (<http://www.aip.org/pubservs/epaps.html>).

¹³R. P. A. Bettens and A. M. Lee J. Phys. Chem. A **110**, 8777 (2006).

¹⁴M. J. Frisch, G. W. Trucks, and H. B. Schlegel *et al.*, GAUSSIAN 03, Gaussian, Inc., Pittsburgh, PA, 2003.

¹⁵X. H. Chen and J. Z. H. Zhang, J. Theor. Comput. Chem. **3**, 277 (2004).

¹⁶R. J. Wilson, *Introduction to Graph Theory* (Longman, London, 1972).