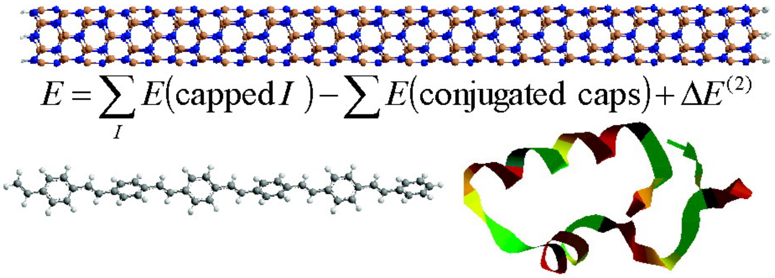
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

An Efficient Fragment-Based Approach for Predicting the Ground-State Energies and Structures of Large Molecules

Shuhua Li, Wei Li, and Tao Fang

J. Am. Chem. Soc., **2005**, 127 (19), 7215-7226• DOI: 10.1021/ja0427247 • Publication Date (Web): 23 April 2005



Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

JACS
ARTICLES

An Efficient Fragment-Based Approach for Predicting the Ground-State Energies and Structures of Large Molecules

Shuhua Li,* Wei Li, and Tao Fang

Contribution from the Department of Chemistry, Institute of Theoretical and Computational Chemistry, Lab of Mesoscopic Chemistry, Nanjing University, Nanjing, 210093, People's Republic of China

Received December 3, 2004; E-mail: shuhua@nju.edu.cn

Abstract: An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules at the Hartree−Fock (HF) and post-HF levels is described. The physical foundation of this approach is attributed to the "quantum locality" of the electron correlation energy and the HF total energy, which is revealed by a new energy decomposition analysis of the HF total energy proposed in this work. This approach is based on the molecular fractionation with conjugated caps (MFCC) scheme (Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, 119, 3599), by which a macromolecule is partitioned into various capped fragments and conjugated caps formed by two adjacent caps. We find that the MFCC scheme, if corrected by the interaction between non-neighboring fragments, can be used to predict the total energy of large molecules only from energy calculations on a series of small subsystems. The approach, named as energy-corrected MFCC (EC-MFCC), computationally achieves linear scaling with the molecular size. Our test calculations on a broad range of medium- and large molecules demonstrate that this approach is able to reproduce the conventional HF and second-order Møller−Plesset perturbation theory (MP2) energies within a few millihartree in most cases. With the EC-MFCC optimization algorithm described in this work, we have obtained the optimized structures of long oligomers of *trans*-polyacetylene and BN nanotubes with up to about 400 atoms, which are beyond the reach of traditional computational methods. In addition, the EC-MFCC approach is also applied to estimate the heats of formation for a series of organic compounds. This approach provides an appealing approach alternative to the traditional additivity rules based on either bond or group contributions for the estimation of thermochemical properties.

---

1. Introduction

The development of linear-scaling algorithms for electronic structure calculations has been an active field in the past decade. Such developments would make ab initio quantum chemistry calculations become possible for molecules with thousands of atoms, greatly expanding the applications of computational chemistry. At the Hartree−Fock (HF) and density functional theory (DFT) levels, the conventional high power scaling behavior (O(N^3) or worse) of several key steps with the system size has been reduced to near-linear scaling in the large molecule limit. The fast multipole methods reduce the calculation of the Coulombic matrix from O(N^2) to O(N),[1−4] and both the "order-*N* exchange"[5] and "near-field-exchange"[6] methods allow linear scaling to be achieved in the calculation of the exchange matrix for large molecules.[7] For very large molecules, the O(N^3) Fock matrix diagonalization will become the dominant time-consuming step. Among approaches replacing the diagonalization of the Fock matrix, density matrix search methods have been proven to be an efficient linear-scaling approach.[8] A combination of these methods has enabled single-point HF calculations of systems containing several thousands of basis functions to become possible. In addition to this category of linear-scaling algorithms that are aimed to calculate the whole system at once, there also exists a category of fragment-based approaches,[9−23] which are capable of reproducing ab initio HF or post-HF results of large molecules quite accurately but with much fewer computational costs. The basic ideas of these fragment-based

---

(1) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325.
(2) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.
(3) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253*, 268.
(4) White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593.
(5) Schwegler, E.; Challacombe, M. *J. Chem. Phys.* **1996**, *105*, 2726.
(6) Burant, J. C.; Scuseria, G. E.; Frisch, M. J. *J. Chem. Phys.* **1996**, *105*, 8969.
(7) Scuseria, G. E. *J. Phys. Chem. A* **1999**, *103*, 4782.
(8) Li, X.; Millam, J. M.; Scuseria, G. E.; Frisch, M. J.; Schlegel, H. B. *J. Chem. Phys.* **2003**, *119*, 7651.
(9) Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 1438.
(10) Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.
(11) Lee, T.-S.; York, D. M.; Yang, W. *J. Chem. Phys.* **1996**, *105*, 2744.
(12) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791.
(13) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980.
(14) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, 4301.
(15) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
(16) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039.
(17) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. *Chem. Phys. Lett.* **2004**, *394*, 293.
(18) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2.
(19) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103.
(20) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.
(21) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.
(22) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832.
(23) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129.

10.1021/ja0427247 CCC: $30.25

approaches are to divide a large molecule into a series of fragments, then perform conventional quantum chemical calculations on fragments, and finally construct the properties, such as electron density and the HF energy, of the whole molecule in some way. Depending on different ways of treating covalent bonds that join fragments, many fragment-based approaches have been proposed in the literature in recent years. For instance, Yang and co-workers developed the divide-and-conquer (DC) method,[9−11] Mezey and Exner developed the adjustable density matrix assembler (ADMA) approach,[12−14] Zhang and co-workers developed the molecular fractionation with conjugated caps (MFCC) approach,[15,16] and Kitaura et al. proposed the fragment molecular orbital (FMO) method.[20−23] The comparisons of these approaches and other developments[3,7,24−31] with conventional HF or DFT calculations for some kinds of macromolecules have demonstrated the accuracy and computational linear-scaling behavior of fragment-based approaches.

On the other hand, it is well-known that the computational cost of the traditional post-HF (or electron correlation) methods with the molecular size (at least $O(N^5)$) is even higher than that of the HF method. To extend correlation calculations to large systems, some lower-order even linear-scaling local correlation methods within the coupled cluster (CC) theory[32−47] and Møller−Plesset perturbation theory (MPPT)[44−46,48−66] have been developed.

The central idea of these local correlation (LC) methods is to recast the traditional MPPT and CC equations in the localized molecular orbital (LMO)[32−38,59−65] or atomic orbital (AO)[41−46,48−56] basis so that the "quantum locality" of the electron correlation problem can be exploited. It has been demonstrated that linear scaling of computational time with molecular size can be achieved in local Møller−Plesset (MP) and coupled cluster (CC) calculations of large molecules. Moreover, these local correlation methods are shown to be capable of recovering more than 99.0% of the correlation energy obtained in the corresponding conventional CC and MP calculation by using typical thresholds. Nevertheless, at present the computational cost of these LC methods is still demanding since they achieve linear scaling only in the large molecule asymptote.

A few linear-scaling fragment-based methods have also been developed for correlation calculations on large molecules.[47,66−69] Within these fragment-based methods, the total electron correlation energy of a large system is estimated as the sum of the intrafragment and interfragment correlation contributions, which can be approximately obtained from correlation calculations on fragments or combined fragments. The implicit assumption exploited in these methods is that the intrafragment and interfragment correlation energies are transferable from these fragments to the whole system due to the similarity between LMOs in fragments and those in the whole molecule. These methods include, for example, the FMO-based second-order Møller−Plesset perturbation theory (FMO-MP2) proposed by Fedorov and Kitaura,[66] the incremental method proposed by Stoll,[67−69] and a MFCC-based approach proposed by us.[47] Among these fragment-based methods, the MFCC-based approach is the simplest one but its accuracy is competitive with other more sophisticated local correlation methods.[59−65]

In this article, our aim is to introduce an efficient fragment-based approach that can predict the total energy of a large molecule at the HF and post-HF levels with good accuracy with low computational costs. This approach can be formally derived from a new energy decomposition analysis of the HF total energy proposed in this work. With this approach, an optimization algorithm is proposed and implemented that will make geometry optimizations of large molecules with hundreds of atoms become feasible at the present time. In addition, the present fragment-based approach will also be applied to estimate the heats of formation for gas-phase compounds from structurally similar small species. Different from the well-known bond or group additivity approaches,[70,71] the present approach provides an alternative appealing approach for chemists to understand and predict the physical and chemical properties of a huge amount of chemical compounds.

This article is organized as follows. In section II, we first introduce a new energy decomposition analysis of the HF total energy. Next, on the basis of this energy decomposition analysis, the energy-corrected MFCC (EC-MFCC) approach is proposed for calculating the total energy of large molecules at the HF and post-HF levels. Then, an optimization algorithm for

---

(24) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165.
(25) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484.
(26) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E. *Theor. Chem. Acc.* **2004**, *111*, 255.
(27) Strout, D. L.; Scuseria, G. E. *J. Chem. Phys.* **1995**, *102*, 8448.
(28) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1996**, *104*, 6643.
(29) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879.
(30) Couty, M.; Bayse, C. A.; Hall, M. B. *Theor. Chem. Acc.* **1997**, *97*, 96.
(31) Ochsenfeld, C.; Kussmann, J.; Koziol, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 4485.
(32) Laidig, W. D.; Purvis, G. D.; Bartlett, R. J. *J. Phys. Chem.* **1985**, *89*, 2161.
(33) Schutz, M.; Werner, H.-J. *Chem. Phys. Lett.* **2000**, *318*, 370.
(34) Rauhut, G.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 4853.
(35) Schutz, M.; Werner, H.-J. *J. Chem. Phys.* **2001**, *114*, 661.
(36) Schutz, M. *J. Chem. Phys.* **2002**, *116*, 8772.
(37) Li, S.; Ma, J.; Jiang, Y. *J. Comput. Chem.* **2002**, *23*, 237.
(38) Li, S.; Li, W.; Ma, J. *Chin. J. Chem.* **2003**, *21*, 1422.
(39) Schutz, M.; Manby, F. R. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3349.
(40) Flocke, N.; Bartlett, R. J. *J. Chem. Phys.* **2003**, *118*, 5326.
(41) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286.
(42) Constans, P.; Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **2000**, *113*, 10451.
(43) Schutz, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 3941.
(44) Scuseria, G. E.; Ayala, P. Y. *J. Chem. Phys.* **1999**, *111*, 8330.
(45) Head-Gordon, M.; Maslen, P. E.; White, C. A. *J. Chem. Phys.* **1998**, *108*, 616.
(46) Maslen, P. E.; Lee, M. S.; Head-Gordon, M. *Chem. Phys. Lett.* **2000**, *319*, 205.
(47) Li, W.; Li, S. *J. Chem. Phys.* **2004**, *121*, 6649.
(48) Saebo, S.; Pulay, P. *J. Chem. Phys.* **1987**, *86*, 914.
(49) Haser, M.; Almlof, J. *J. Chem. Phys.* **1992**, *96*, 489.
(50) Azhary, A. E.; Rauhut, G.; Pulay, P.; Werner, H.-J. *J. Chem. Phys.* **1998**, *108*, 5185.
(51) Hetzer, G.; Pulay, P.; Werner, H.-J. *Chem. Phys. Lett.* **1998**, *290*, 143.
(52) Maslen, P. E.; Head-Gordon, M. *Chem. Phys. Lett.* **1998**, *283*, 102.
(53) Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **1998**, *109*, 7093.
(54) Ayala, P. Y.; Scuseria, G. E. *J. Chem. Phys.* **1999**, *110*, 3660.
(55) Lee, M. S.; Maslen, P. E.; Head-Gordon, M. *J. Chem. Phys.* **2000**, *112*, 3592.
(56) Pulay, P.; Saebo, S.; Wolinski, K. *Chem. Phys. Lett.* **2001**, *344*, 543.
(57) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. *J. Chem. Phys.* **1995**, *103*, 1481.
(58) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908.
(59) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. *J. Phys. Chem. A* **1999**, *103*, 1913.
(60) Rauhut, G.; Azhary, A. E.; Eckert, F.; Schumann, U.; Werner, H.-J. *Spectrochim. Acta, Part A* **1999**, *55*, 647.
(61) Schutz, M.; Hetzer, G.; Werner, H.-J. *J. Chem. Phys.* **1999**, *111*, 5691.
(62) Finley, J. P.; Hirao, K. *Chem. Phys. Lett.* **2000**, *328*, 51.
(63) Hetzer, G.; Schutz, M.; Stoll, H.; Werner, H.-J. *J. Chem. Phys.* **2000**, *113*, 9443.
(64) Saebo, S.; Pulay, P. *J. Chem. Phys.* **2001**, *115*, 3975.
(65) Nakao, Y.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 6375.
(66) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.
(67) Stoll, H. *J. Chem. Phys.* **1992**, *97*, 8449.
(68) Stoll, H. *Phys. Rev. B* **1992**, *46*, 6700.
(69) Stoll, H. *Chem. Phys. Lett.* **1992**, *191*, 548.
(70) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. *Chem. Rev.* **1969**, *69*, 279.
(71) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. *J. Phys. Chem.* **1973**, *77*, 1687.

geometry optimizations of large molecules is proposed. In section III, we first validate the accuracy of the EC-MFCC approach by applying it to calculate the total energies for a broad range of chemical systems. Then, the EC-MFCC optimization algorithm is applied to fully optimize the structures of long oligomers of *trans*-polyacetylene (*trans*-PA) and BN nanotubes, in which the largest system contains about 400 atoms. In addition, the heats of formation of some organic compounds are estimated to further illustrate the use of the present approach. Finally, a brief summary is given in section IV.

2. Methodology

2.1. The Energy Decomposition Analysis at the HF Level. It is well-known that the total energy of a molecule at the HF level can be expressed as:

$$E_{\text{total}} = \frac{1}{2}\sum_{u}\sum_{v}\mathbf{P}_{vu}(\mathbf{H}^{\text{core}}_{uv} + \mathbf{F}_{uv}) + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{1}$$

where $\mathbf{P}$ is the density matrix, $\mathbf{H}^{\mathbf{core}}$ is the core-Hamiltonian matrix, and $\mathbf{F}$ denotes the Fock matrix that contains a one-electron part $\mathbf{H}^{\mathbf{core}}$ and a two-electron part that depends on the density matrix $\mathbf{P}$ and a set of two-electron integrals over basis functions. According to eq 1, one can see that the total energy can be decomposed into the sum of the single-atom, two-atom, three-atom, and four-atom terms if the basis functions are centered at the atoms. More generally, if several given atoms are combined to form a fragment, the total energy can be similarly divided into the sum of n-fragment terms ($n = 1-4$). For instance, if a system is divided into N fragments, the total energy can be decomposed as below

$$E_{\text{total}} = \sum_{I=1}^{N}E_I + \sum_{I=1}^{N}\sum_{J>I}^{N}E_{I,J} + \sum_{I=1}^{N}\sum_{J>I}^{N}\sum_{K>J}^{N}E_{I,J,K} + \sum_{I=1}^{N}\sum_{J>I}^{N}\sum_{K>J}^{N}\sum_{L>K}^{N}E_{I,J,K,L} \tag{2}$$

$$E_I = \sum_{u,v\in I}\mathbf{P}_{vu}\left\langle u\left| -\frac{1}{2}\nabla^2\right| v\right\rangle + \sum_{u,v\in I}\sum_{A\in I}\mathbf{P}_{vu}\left\langle u\left| -\frac{Z_A}{|r-R_A|}\right| v\right\rangle + \sum_{A=1}\sum_{\substack{B>A\\A,B\in I}}\frac{Z_A Z_B}{R_{AB}} + \frac{1}{2}\sum_{u,v\in I}\sum_{\lambda,\sigma\in I}\mathbf{P}_{vu}\mathbf{P}_{\lambda\sigma}\left[(uv|\sigma\lambda) - \frac{1}{2}(u\lambda|\sigma v)\right] \tag{3}$$

$$E_{I,J} = E_{IJ} - E_I - E_J \tag{4}$$

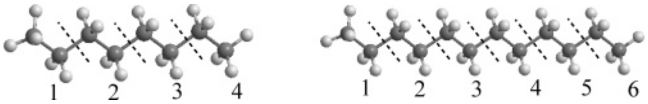
$$E_{I,J,K} = E_{IJK} - E_I - E_J - E_K - E_{I,J} - E_{I,K} - E_{J,K} \tag{5}$$

$$E_{I,J,K,L} = E_{IJKL} - E_I - E_J - E_K - E_L - E_{I,J} - E_{I,K} - E_{I,L} - E_{J,K} - E_{J,L} - E_{K,L} - E_{I,J,K} - E_{I,J,L} - E_{I,K,L} - E_{J,K,L} \tag{6}$$

where E_I represents the intrafragment energy within the fragment I, E_{IJ} is the energy contribution within a combined fragment IJ composed of fragments I and J, and $E_{I,J}$ stands for the two-body interaction energy between fragments I and J. In a similar manner, $E_{I,J,K}$ and $E_{I,J,K,L}$ denote the three-body and four-body terms, respectively. It should be pointed out that the expansion of the total energy into the sum of n-body ($n = 1-4$) terms as in eq 2 is complete for arbitrary molecular systems, since all the one-electron or two-electron integrals involve at most four atoms or four fragments. A similar energy decomposition scheme was proposed in the fragment molecular orbital method,[20] but different formulas were defined for the n-body contributions.

Now we want to demonstrate the features of the energy decomposition analysis (EDA) described above. First, we take dodecane ($C_{12}H_{26}$) as a model system. The geometry of this molecule is optimized at the HF/6-31G level with the Gaussian 03 package.[72] By partitioning this molecule into six fragments (each fragment consists of two carbon atoms), we have calculated all the n-body contributions at the HF/6-31G level according to the present EDA scheme, with part of the results collected in Table 1. As seen from Table 1, the sum of the intrafragment energies accounts for more than 99% of the total energy. Among the remaining terms, the total two-body terms make the most important contribution. Clearly, the magnitude of a given two-body term decreases rapidly with increasing distance between two fragments. The total three-body terms still contribute significantly, especially when three fragments are spatially close to each other. For example, $E_{1,2,3}$ is even larger than $E_{1,3}$ in magnitude. For a trimer IJK, the data shown in Table 1 show that its contribution depends strongly on its spatial “compactness”. A similar trend also exists for the four-body terms, but the total four-body interaction energies are about 3 orders of magnitude smaller than the total three-body contributions. From the discussions above, one can clearly see that the order of the total n-body contributions to the total energy obtained by the present EDA scheme is in accord with the intuition of chemists. On the other hand, it is interesting to know whether the calculated intrafragment and interfragment contributions are approximately transferable in a series of structurally similar compounds. For this purpose, we have presented all the one-body energies and some n-body ($n \geq 2$) interaction energies of octane (C_8H_{18}) in Table 1, with each fragment being the same as that in dodecane. As one can see, the intrafragment energies are almost identical for the terminal fragments in both molecules and are within the 0.30 milliHartree (mH) for the next-terminal fragments. Among the two-body terms, $E_{1,2}$ differs by about only 0.2 mH from octane to dodecane, and the difference of $E_{1,3}$ in both molecules is even smaller. Similarly, the term $E_{1,2,3}$ in both molecules varies by only 0.03 mH. The near-equivalence of these terms can be attributed to the fact that the fragments 1, 2, and 3 (shown in Table 1) have analogous neighboring groups in these two molecules. Thus, both the intrafragment and interfragment contributions are found to be highly transferable if the fragments involved have similar local environments in different systems. To conclude from the discussions above, the present EDA

Table 1. Intrafragment and Interfragment Energies (au) of C_8H_{18} and $C_{12}H_{26}$ Calculated at the HF/6-31G Level



C_8H_{18} components	C_8H_{18} energies	$C_{12}H_{26}$ components	$C_{12}H_{26}$ energies
One-Body Terms		One-Body Terms	
E_1	−78.58259	E_1	−78.58261
E_2	−77.96443	E_2	−77.96472
E_3	−77.96443	E_3	−77.96480
E_4	−78.58259	E_4	−77.96474
		E_5	−77.96459
		E_6	−78.58257
Two-Body Terms		Two-Body Terms	
$E_{1,2}$	−6.03330E-2	$E_{1,2}$	−6.0134E-2
$E_{1,3}$	−8.65634E-3	$E_{1,3}$	−8.6692E-3
$E_{1,4}$	−6.96346E-5	$E_{1,4}$	7.3123E-5
		$E_{1,5}$	1.4560E-5
		$E_{1,6}$	7.5777E-6
Three-Body Terms		Three-Body Terms	
$E_{1,2,3}$	−1.16498E-2	$E_{1,2,3}$	−1.1625E-2
$E_{1,2,4}$	1.00788E-4	$E_{1,2,4}$	3.1972E-5
		$E_{1,2,5}$	−6.4807E-5
		$E_{1,2,6}$	−2.1953E-5
Four-Body Term		Four-Body Terms	
$E_{1,2,3,4}$	−2.65493E-5	$E_{1,2,3,4}$	−4.5985E-5
		$E_{1,2,3,5}$	−3.2165E-5
		$E_{1,2,3,6}$	−5.9300E-9
total energy	−313.30844	total energy	−469.38214

---

(72) Frisch, M. J. et al. *Gaussian 03*, revision B.04; Gaussian, Inc.: Wallingford, CT, 2004.

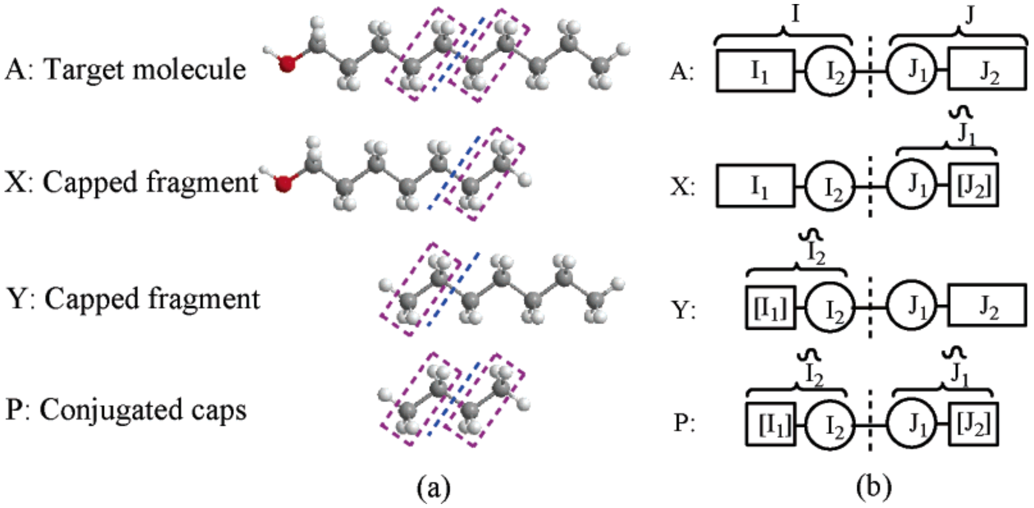


Figure 1. (a) Capped fragments and conjugated cap for 1-decanol (the target molecule). (b) An illustration of fragments, caps, and conjugated caps for a general molecule.

procedure has two important features: one is the fast decay of the calculated n-body ($n \geq 2$) interaction energies with the largest distance between any two fragments (of n fragments involved), and the other is the good transferability of the n-body contributions within the structurally analogous systems.

It should be mentioned that the correlation energy can also be decomposed into the sum of n-atom (or body) ($n = 1-4$) terms or the sum of the atom pair interactions (after partial summation of three- and four-atom terms), as discussed by Ayala and Scuseria.[73] Their calculations also showed the fast decay of the atom pair interactions with the distance between atoms. Thus, the total energy of a molecule, at the HF or post-HF level, can always be partitioned into the sum of the one-, two-, three-, and four-atom contributions. Due to the transferability of the n-body contributions within the structurally analogous systems, such an energy decomposition scheme could provide a theoretical foundation for the following fragment-based approach or the well-known bond or group additivity rules,[70,71] which have been widely used to understand and predict the physical and chemical properties of a huge amount of chemical compounds.

2.2. The Energy-Corrected MFCC Approach for Estimating the Total Energy of Macromolecules. Due to the features of the n-body ($n \geq 2$) interaction energies, as described above, eq 2 can be used to devise an approximate approach for estimating the total energy in large molecules at the HF level. Since the n-body ($n \geq 2$) interaction energies decrease rapidly with the largest separation between any two fragments of n fragments involved, only those significant terms with fragments spatially close to each other are required to be included in eq 2 for a given accuracy. Due to the good transferability of the n-body contributions in a homologous series of compounds, the n-body contributions can be derived from the EDA calculations on small systems. Thus, one could approximately determine the total energy of large molecules at the HF level from HF calculations on a series of small systems.

Now we will introduce a very simple procedure to implement these ideas. First, a large molecule is partitioned into various fragments. Normally, neighboring fragments are connected by a single covalent bond. Second, for each fragment we cap it with its local environments (environmental groups are normally no smaller than the selected fragment) and add hydrogen atoms (if necessary) to those terminal non-hydrogen atoms in environmental groups to form complete caps. For each cut bond, a right cap is constructed to mimic the rest of the molecule to the right, and similarly a left cap is to mimic the rest of the molecule to the left. For simplicity, these two caps are said to form conjugated caps. Take 1-decanol as an example; we illustrate this fragmentation-and-capping scheme in Figure 1a. This scheme, termed by MFCC, was first proposed by Zhang and Zhang.[15] This approach has been employed to calculate interaction energies between small molecules and biological molecules,[15,16] the electron density,[17,18] and kinetic energy[19] of large molecules. However, this approach has not been used to calculate the total energy of large molecules in the simple way described below. For a general molecule, a schematic picture of the MFCC scheme is shown in Figure 1b. A large system A is partitioned into two large fragments I and J. Similarly, fragment I is further divided into two subunits I_1 and I_2, and fragment J is further divided into two subunits J_1 and J_2. For convenience, a small part of J_2 capped with hydrogen atoms from the right is denoted as $[J_2]$, and the substructure that consists of J_1 and $[J_2]$ is denoted as a cap $\tilde{J}_1$. Assume that fragment I capped by $\tilde{J}_1$ forms a subsystem X, fragment J capped by $\tilde{I}_2$ forms a subsystem Y, and two conjugated caps $\tilde{I}_2$ and $\tilde{J}_1$ form a subsystem P. According to eq 2, the total energies of the whole system and these subsystems can be approximately calculated as below:

$$E_{\mathrm{HF}}(A) \approx E_{I1} + E_{I2} + E_{J1} + E_{J2} + E_{I1,I2} + E_{I2,J1} + E_{J1,J2} + E_{I1,J1} + E_{I2,J2} + E_{I1,I2,J1} + E_{I2,J1,J2} \tag{7}$$

$$E_{\mathrm{HF}}(X) \approx E_{I1} + E_{I2} + E_{J1} + E_{[J2]} + E_{I1,I2} + E_{I2,J1} + E_{J1,[J2]} + E_{I1,J1} + E_{I2,[J2]} + E_{I1,I2,J1} + E_{I2,J1,[J2]} \tag{8}$$

$$E_{\mathrm{HF}}(Y) \approx E_{[I1]} + E_{I2} + E_{J1} + E_{J2} + E_{[I1],I2} + E_{I2,J1} + E_{J1,J2} + E_{[I1],J1} + E_{I2,J2} + E_{[I1],I2,J1} + E_{I2,J1,J2} \tag{9}$$

$$E_{\mathrm{HF}}(P) \approx E_{[I1]} + E_{I2} + E_{J1} + E_{[J2]} + E_{[I1],I2} + E_{I2,J1} + E_{J1,[J2]} + E_{[I1],J1} + E_{I2,[J2]} + E_{[I1],I2,J1} + E_{I2,J1,[J2]} \tag{10}$$

In the equations above, some implicit assumptions are made. For example, the four-body term and some three-body terms such as $E_{I1,J1,J2}$ are totally ignored in all these equations due to their negligible contributions. The one-body, two-body, and three-body energies are assumed to be transferable for those fragments with analogous neighboring groups in different systems. Combining these four equations together, one can obtain a simple relationship:

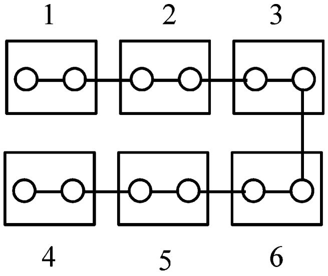
$$E_{\mathrm{HF}}(X) + E_{\mathrm{HF}}(Y) - E_{\mathrm{HF}}(P) \approx E_{\mathrm{HF}}(A) \tag{11}$$

This relationship clearly shows that the total HF energy of the whole large system A can be approximately determined from HF calculations on subsystems X, Y, and P. In the following, for convenience we call subsystems constructed from a fragment capped by its local environ-

---

(73) Ayala, P. Y.; Scuseria, G. E. *Chem. Phys. Lett.* **2000**, *322*, 213.

Scheme 1



mental groups as capped fragments. The above procedure can be easily extended to larger systems, which may be decomposed into three or more fragments. In general, the total energy of a large molecule at the HF level can be approximately calculated by using the following expression

$$E = \sum_I E(\text{capped } I) - \sum E(\text{conjugated caps}) \qquad (12)$$

Here the summation in the first term runs over all fragments. One should note that for the central fragments, they are capped from both sides, while the terminal fragments are capped from the left or the right.

In our previous work,[47] we have derived the same equation as in 12 for approximately evaluating the correlation energies of large molecules, based on the fast decay of pair correlation energies with the separation between localized molecular orbitals (MO) and the transferability of pair correlation energies within the structurally similar molecules. Of course, this can also be derived from the fast decay property of the atom pair interactions, if the correlation energy is partitioned into atom pairs.[73] As a result, eq 12 can be applied to approximately determine the total energy of a large molecule at the HF and any post-HF levels.

It is important to know how to construct the geometries of capped fragments and conjugated caps. To ensure the transferability of the *n*-body interactions in the parent system and its subsystems, the positions of all atoms in caps and fragments, except those added hydrogen atoms, are kept to be the same as those in the parent molecule. To construct a cap for a given fragment, we first choose its neighboring subunit. If this subunit is connected to the rest of the system by a covalent single bond X−Y (this holds true in most cases), we then replace the X−Y bond with the X−H bond (X belongs to this subunit) to form a complete cap. This added hydrogen atom is placed somewhere between atoms X and Y. In this work, we use 1.07 Å for C−H, 1.00 for N−H, 0.96 for O−H, 1.18 for B−H, and 1.31 for S−H as defaults. Thus, the position of this added hydrogen atom is unambiguously determined. Of course, other atoms rather than the hydrogen atom may also be chosen as "link" atoms. Here the hydrogen atom is chosen for simplicity, as in combined quantum-mechanical/molecular-mechanics approaches.[74] If the X−Y bond is not a single bond, a more complicated procedure for capping the X atom was suggested previously.[47]

It should be pointed out that eq 12 is applicable only for one-dimensional or quasi-one-dimensional systems, in which the interactions between nonbonded or distant fragments can be ignored. For two- or three-dimensional macromolecules, two fragments separated by two or more fragments in the sequence may also be adjacent to each other in space (as shown in Scheme 1), thus the *n*-body ($n \geq 2$) interaction energies involving these spatially close fragments, which primarily originate from "through-space" interactions (e.g., van der Waals or electrostatic interactions), may be still significant. Clearly, these "through-space" *n*-body ($n \geq 2$) interactions, which are called the *n*-body corrections in the following, are not taken into account in eq 12. In general, the three- and four-body corrections are far less important than the two-body correction if the fragments are reasonably chosen and thus can be neglected in most cases. For a given molecule, the two-body correction may be approximately evaluated by the formula below:

$$\Delta E^{(2)} = \sum_I \sum_J [E(\text{capped } I + \text{capped } J) - E(\text{capped } I) - E(\text{capped } J)] \qquad (13)$$

Here the first term in the summation stands for the total energy of the dimer composed of the capped fragments *I* and *J*. To avoid overcounting the two-body correction, each fragment here is capped with only hydrogen atoms. If large caps are used for fragments, more complicated procedures, other than eq 13, are required to obtain a reasonable estimate to the "through-space" two-body interactions. Thus, one should be aware that in deriving eqs 12 and 13 two different capping schemes are used, respectively. To simplify the following discussions, only capping schemes used in deriving eq 12 are described if necessary. To conclude, we propose that the total energy of a general macromolecule at the HF and any post-HF levels can be approximately determined using the following equation:

$$E = \sum_I E(\text{capped } I) - \sum E(\text{conjugated caps}) + \Delta E^{(2)} \qquad (14)$$

For convenience, this approach is called as the energy-corrected MFCC (or EC-MFCC) approach. Similarly, this formula should also be applicable within the density functional theory (DFT) framework. Thus, by performing calculations on capped fragments, conjugated caps, and subsystems composed of two capped fragments, one should be able to obtain a good approximation to the total energies of any sufficiently large molecules. Since the two-body interaction energy will become negligible when the separation between the two fragments is larger than a given threshold, the number of the subsystems containing two capped fragments to be calculated will increase only linearly with the number of fragments. As a result, the overall computational cost of the present method should scale linearly with the molecular size in a sufficiently large molecule.

It would be useful to have some ideas on the applicability and limitations of the present EC-MFCC method. Clearly, the accuracy of this approach noticeably depends on whether the sizes of fragments and caps are reasonably chosen or not. For saturated close-shell molecules, satisfactory predictions of their total energies can usually be obtained with relatively small fragments and caps. But for some aromatic systems, where electrons are highly delocalized, significantly larger fragments and caps may be required to achieve the same accuracy as in their analogous saturated systems. However, for those systems with very small gaps between highest occupied and lowest unoccupied MOs, the present method may even break down. A feasible method to validate the present approach is to check the dependence of the calculated total energy on the size of fragments and caps for a given system. If the calculated energy is convergent within a given accuracy, it should be a good approximation to the total energy.

2.3. Geometry Optimization with the EC-MFCC Approach. With various linear-scaling methods, ab initio single-point calculations on molecules containing hundreds of atoms are now feasible. However, ab intio geometry optimization of such large molecules is not yet routinely possible, especially at the electron correlation level. With the simple EC-MFCC approach, we will show that ab initio geometry optimizations for very large systems can now be performed at the HF (or DFT) and post-HF levels with fairly low computational resources, which is likely to have significant impacts on many areas outside traditional computational chemistry.

For small- and medium-sized molecules, geometry optimization techniques such as the quasi-Newton[75−77] and conjugated gradient

---

(74) Singh, U. C.; Kollman, P. *J. Comput. Chem.* **1986**, *7*, 718.

(75) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214.

(76) Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.

(77) Farkas, O.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 10806.

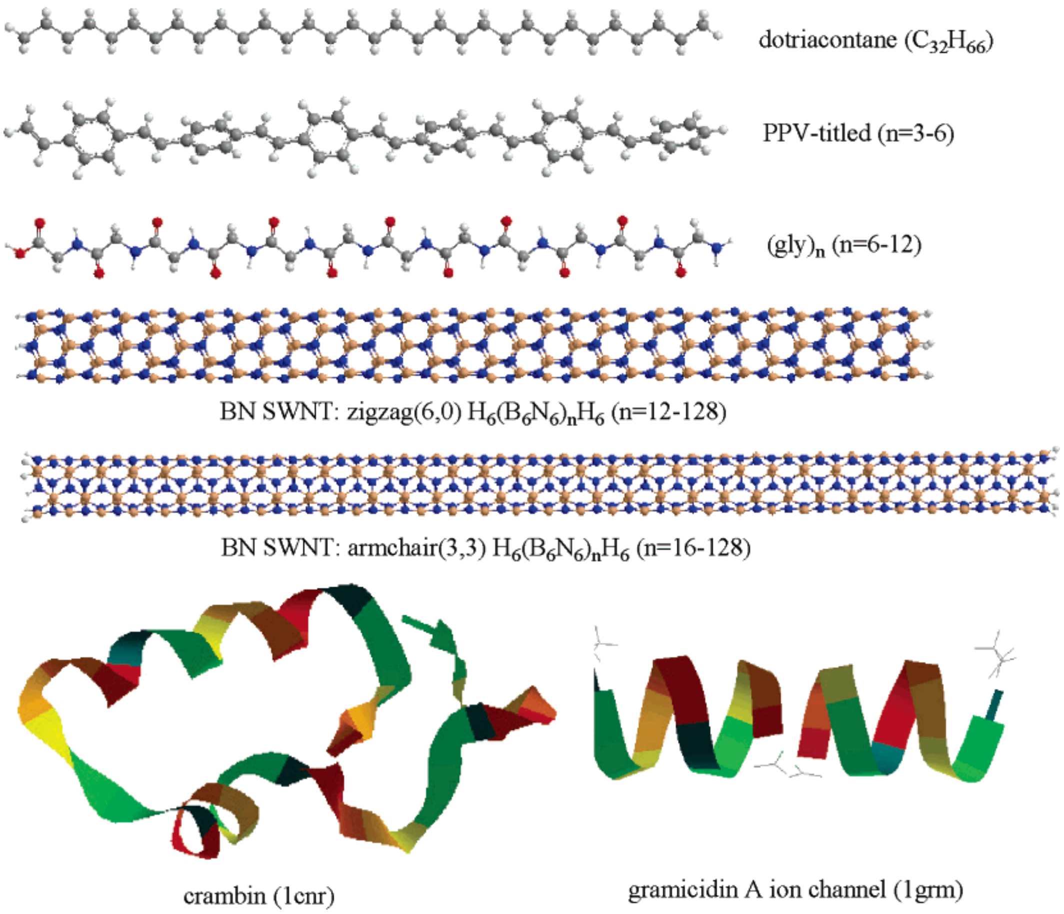


Figure 2. Some selected systems.

methods[78] are now well-established at the HF and post-HF levels, based on the availability of analytical gradients. In fact, these optimization algorithms can be directly employed for our purpose, with only the energy of the system calculated with eq 14 and the gradient of the system calculated with the equation below:

$$\frac{\partial E}{\partial q_i} = \sum_{I} \frac{\partial E(\text{capped } I)}{\partial q_i} - \sum \frac{\partial E(\text{conjugated caps})}{\partial q_i} + \frac{\partial \Delta E^{(2)}}{\partial q_i} \quad (15)$$

It should be pointed out that only the coordinates of the real system are involved in eq 15, and the coordinates of the “link” atoms (hydrogen) in capped fragments or conjugated caps are not considered (because the relative positions of the “link” atoms are unambiguously determined by the positions of the related atoms in the real system, as described earlier). Since all the gradient components of various subsystems can be obtained analytically, the gradient of the whole system can be computed readily. In the present work, we choose the quasi-Newton method[75−77] to be the optimization algorithm. The BFGS procedure is used to update the Hessian matrix,[78] and a line search procedure is performed at each optimization step except the first one. Both Cartesian coordinates and internal coordinates can be used in our optimization program.

3. Results and Discussions

In this section, we will first test the accuracy and applicability of the EC-MFCC approach by applying it to obtain approximate total energies for a broad range of medium and large-sized molecules within the HF and MP2 theories. The energy calculations are carried out directly with the Gaussian 03[72] program. Next, we will employ the EC-MFCC optimization program to obtain the optimized structures for long oligomers of *trans*-PA and long BN nanotubes at the MP2 and HF levels, respectively. Finally, we will demonstrate that the EC-MFCC approach can be used to estimate the heats of formation for some organic systems with good accuracy.

3.1. Total Energies Predicted by the EC-MFCC Approach. In this subsection, the EC-MFCC approach is applied to a variety of molecules to obtain their total energies at the HF and MP2 levels. The selected systems, some of which are shown in Figure 2, include saturated and conjugated molecules and also exhibit sufficient structural complexity. Hereafter, the HF (MP2) energies without two-body corrections are denoted as MFCC-HF (MP2) energies for simplicity. If the closest atom−atom distance between the two fragments is less than a given threshold (R Å), the two-body correction should be considered for this pair of fragments. Obviously, a short distance threshold (R_1 Å) should be used for the van der Waals interaction, and a much long distance threshold (R_2 Å) should be used for the long-range electrostatic interaction. Thus, for those molecules without charged functional groups, the HF (MP2) energies with two-body corrections are denoted as EC-MFCC(R_1)-HF (MP2) energies, since only the van der Waals distance threshold is required to be defined, while for those molecules containing charged groups, two distance thresholds are defined. In this case, the HF (MP2) energies are abbreviated as EC-MFCC(R_1,R_2)-HF (MP2) energies. In all MP2 calculations, the core orbitals are frozen and the basis set is the same as that used in the

---

(78) Leach, A. R. *Molecular Modelling: Principles and Applications*; Addison Wesley Longman: London, 1996.

Table 2. Energy Differences (ΔE) between the Conventional HF (MP2) and the Corresponding EC-MFCC Energies for Selected One-Dimensional Systems

molecule	basis set[a]	ΔE (mH)[b] MFCC	EC-MFCC(8)[c]	basis functions of the largest subsystem
dotriacontane	6-31G* (612)	−0.01 (0.23)	−0.04 (−0.13)	160
β-carotene	6-31G (472)	0.05 (0.83)	0.06 (0.74)	209
		PPV-Planar		
$n = 3$	cc-pVDZ (436)	0.05 (0.26)	0.08 (0.24)	304
$n = 4$	cc-pVDZ (578)	0.11 (0.53)	0.17 (0.50)	304
$n = 5$	cc-pVDZ (720)	0.16	0.26	304
$n = 6$	cc-pVDZ (862)	0.22	0.34	304
		PPV-Titled		
$n = 3$	cc-pVDZ (436)	0.04 (0.22)	0.07 (0.20)	304
$n = 4$	cc-pVDZ (578)	0.08 (0.44)	0.14 (0.41)	304
$n = 5$	cc-pVDZ (720)	0.12	0.21	304
$n = 6$	cc-pVDZ (862)	0.16	0.27	304
		$H(NHCH_2CO)_nOH$		
$n = 6$	6-31G (265)	0.15 (0.24)	−0.09 (−0.03)	196
$n = 8$	6-31G (349)	0.46 (0.71)	−0.09 (0.06)	200
$n = 10$	6-31G (433)	0.84 (1.27)	−0.04 (0.24)	200
$n = 12$	6-31G (517)	1.26 (1.87)	0.05 (0.46)	200
		$^+H_2(NHCH_2CO)_nO^-$		
$n = 6$	6-31G (265)	32.65 (34.16)	1.53 (2.38)	198
$n = 8$	6-31G (349)	27.32 (29.12)	1.39 (2.48)	202
$n = 10$	6-31G (433)	24.36 (26.39)	1.23 (2.50)	202
$n = 12$	6-31G (517)	22.58 (24.85)	1.22 (2.66)	202

[a] The total number of basis functions included in parentheses. [b] The relative energies with respect to corresponding conventional energies, and the relative energies from EC-MFCC-MP2 calculations included in parentheses. [c] For zwitterions of glycine peptides, EC-MFCC(8,30)-HF (MP2) energies are given instead.

corresponding HF calculations. The Cartesian coordinates and conventional HF or MP2 energies of all compounds studied are presented in the Supporting Information.

Since in the EC-MFCC treatment a fragmentation scheme is required for each system before actual calculations can be done, we have displayed fragmentation schemes for most of the studied systems in Figure S1 of the Supporting Information. For other systems studied, a brief description on how to fragment them will be mentioned elsewhere. A general guide for partitioning a large molecule we adopt here is that fragments should be of roughly equal size and should contain at least two non-hydrogen atoms. For each fragment, each cap is chosen as its neighboring fragment saturated by necessary hydrogen atoms in this work for simplicity. In a homologous series of molecules, the sizes of fragments are chosen to be the same for easy comparison.

3.1.1. One-Dimensional and Quasi-One-Dimensional Systems. Among this category of systems, the selected molecules include: dotriacontane ($C_{32}H_{66}$), β-carotene, planar and titled poly(p-phenylene vinylene) (PPV) ($n = 3-6$), glycine peptides $(gly)_n$ ($n = 6, 8, 10, 12$) and their zwitterions, and BN single-walled nanotubes (SWNTs) of zigzag (6,0) and armchair (3,3). For most of these compounds, their geometries are optimized by using the conventional HF method with various basis sets.

For selected one-dimensional systems, the energy differences between the conventional HF (MP2) energies and the corresponding EC-MFCC values are summarized in Table 2. Both MFCC and EC-MFCC(8) or EC-MFCC(8,30) values at the HF or MP2 level are presented for comparison. In the last column of Table 2, the number of basis functions included in the largest subsystem is given for each species. One can see that the deviations of the EC-MFCC(8) energies relative to those of the conventional HF or MP2 energies are less than 4 mH for all these apolar species. Actually, except for the zwitterionic form of $(gly)_n$, the calculated MFCC energies are almost identical to the EC-MFCC(8) values at both the HF and MP2 levels for all other systems, suggesting that the two-body corrections are insignificant in these systems. For saturated dotriacontane and glycine peptides without charged residues, the EC-MFCC(8) energies are consistent with the conventional values within 0.1 mH at the HF level and within 0.5 mH at the MP2 level. For glycine peptides zwitterions with one positive charge in its N-terminal residue and one negative charge in its C-terminal residue, the accuracy of the MFCC-HF and MFCC-MP2 energies is significantly lower than that for their uncharged counterparts. This is because the electrostatic interaction between the N-terminal residue and the C-terminal residue is a long-range effect, and thus the two-body correction between the two terminal fragments must be considered. By setting two distance thresholds R_1 and R_2 to be 8 and 30 Å, respectively, the calculated EC-MFCC(8,30) energies are found to deviate from the conventional HF and MP2 energies by at most 1.6 and 2.6 mH, respectively, for all glycine peptides zwitterions under study. These results demonstrate that the EC-MFCC scheme could predict quite accurate ground-state total energies for one-dimensional systems without or with charged groups.

To investigate how the accuracy of the present treatment varies with the size of fragments, the difference between the conventional HF (MP2) energy and the corresponding EC-MFCC value for $(gly)_{12}$ zwitterion with the size of each fragment is shown in Figure 3. The calculations are performed with the 6-31G basis set. In the EC-MFCC calculations, we have considered the size of each fragment to be of about one, two, and three residues. From Figure 3, one can see that the accuracy of the EC-MFCC approach increases rapidly when the fragment size is enlarged, and for $(gly)_{12}$ zwitterion the fragment

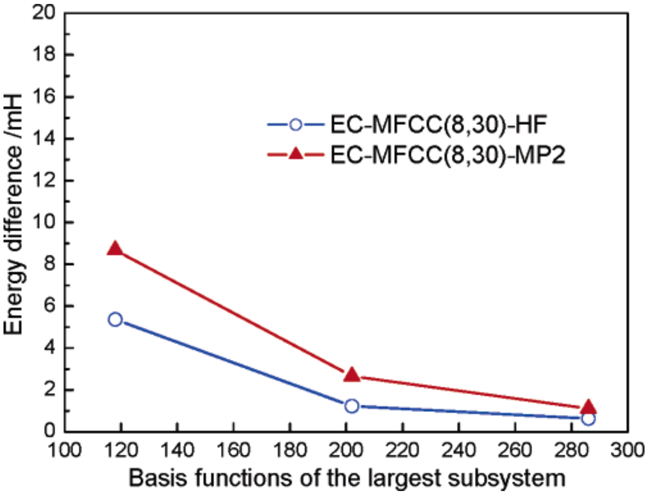


Figure 3. Difference between the conventional HF (MP2) energy and the EC-MFCC(8,30)-HF(MP2) values for $(gly)_{12}$ zwitterion for fragments with different sizes.

containing about two residues can achieve satisfactory results with errors of less than 3 mH.

For selected quasi-one-dimensional systems, zigzag (6,0) and armchair (3,3) BN SWNTs, the calculated EC-MFCC energies and corresponding conventional results (if available) are listed in Table 3. Both ends of these two kinds of BN SWNTs are saturated by hydrogen atoms. In our EC-MFCC calculations, each fragment in all the tubes is chosen to contain four repeating units, that is, $(B_6N_6)_4$. As seen from Table 3, the calculated MFCC-HF energies deviate from the conventional HF values by less than 1.2 mH for two zigzag (6,0) tubes with $n = 12$ and 16, and by 0.2 mH for the $n = 16$ armchair (3,3) tube. The MFCC-MP2 energies are found to have an accuracy similar to the corresponding MFCC-HF values for the smallest tube of both types. As expected, the accuracy of the EC-MFCC(10) energies at the HF or MP2 level is even higher than that of the MFCC energies. Here $R_1 = 10$ Å implies that the two-body corrections between each fragment and its next-nearest neighboring fragments are included. It should be mentioned that the size of fragments for nanotubes should be significantly larger than that for the chainlike oligomers described previously if a similar accuracy is required for these two different categories of systems. For longer zigzag (6,0) and armchair (3,3) tubes with $n > 32$, their total energies are evaluated directly from energies of those subsystems, which already occur in the $n = 32$ tube. If the energy per unit (BN) can be defined as the energy difference between successive nanotubes divided by 6, it is predicted from the EC-MFCC(10) calculations at the HF/3-21G level to be −78.82586 au and −78.81911 au for the infinite zigzag (6,0) and armchair (3,3) tubes, respectively. These values are almost equal to the extrapolated values, −78.82589 au and −78.81910 au, from conventional HF calculations on smaller tubes ($n = 6-17$).[79] Furthermore, for the infinite zigzag (6,0) tube the energy per unit at the MP2 level is −78.97234 au from the corresponding EC-MFCC calculations.

3.1.2. Biological Molecules. The selected biological molecules include indinavir (an HIV-protease inhibitor), two base pairs (adenine−thymine (A−T) and cytosine−guanine (C−G)), and six protein structures from the protein data bank (PDB).[80] The geometries of six proteins are the same as those in Exner and Mezey's recent work.[14] The HF (or MP2) energies of all these molecules obtained from the conventional approach and the EC-MFCC approach are summarized in Table 4. For indinavir, base pairs A−T and G−C, the EC-MFCC(8) energies at the HF and MP2 level are quite satisfactory, differing from their conventional counterparts by less than 1.0 mH. By comparing MFCC values with corresponding EC-MFCC results, however, one can see that the two-body correction, which is mainly caused by the van der Waals interaction for indinavir and the hydrogen-bond interaction for base pairs, is significant for these molecules and has to be taken into account. For six proteins, the EC-MFCC(8,30) approach is still very satisfactory in reproducing the conventional HF energies of the first five with the mean absolute error of a few millihartree, but is less satisfactory for the last one, EETI II. The reason the EC-MFCC approach gives rise to relatively larger errors for EETI II may be attributed to the fact that in this system there are seven positive or negative charges on different residues. In this case, the effect of those remote charged fragments on a given fragment may not be well-mimicked by only summing the two-body interactions between this fragment and each of those charged fragments. The overall performance of the present approach for selected biological molecules is competitive with some existing fragment-based methods such as the field-adapted ADMA approach[14] and the FMO method.[20−23] The difference between MFCC and EC-MFCC(8,30) results strongly indicates that the inclusion of the two-body correction, which is caused by the electrostatic and van der Waals interactions, is very important for biological molecules such as proteins, especially those polar and highly charged molecules.

3.2. Optimized Structures for Long Oligomers of *trans*-Polyacetylene and BN Nanotubes. To illustrate the application of the EC-MFCC optimization method, we have performed full geometry optimizations for two interesting one-dimensional systems, oligomers of *trans*-polyacetylene (or *trans*-oligoenes) and BN nanotubes of varying lengths, with the optimization algorithm described in the preceding section. As discussed above, these two kinds of systems can be well treated with the EC-MFCC approach. Furthermore, it is well-known that the electronic structure of *trans*-PA is closely related to its C−C bond length alternation. A number of experimental and theoretical studies have been conducted to resolve this issue.[81−86] Theoretically, one usually obtains by extrapolation the structure of an infinite chain (*trans*-PA) from those of *trans*-oligoenes. Since the correlation effects significantly reduce the bond length alternation in short *trans*-oligoenes,[84−86] the structures of long *trans*-oligoenes must be optimized at the post-HF level, which limits the application of standard quantum chemistry calculations to small- and medium-sized oligomers. For BN nanotubes, the zigzag (6,0) BN SWNT with a diameter of 0.5 nm has been experimentally observed by Bengu and Marks.[87] A recent theoretical study[79] has provided the fully optimized geometries at the HF/3-21G level for nanotubes with 6−17 repeating units,

---

(79) Xu, H.; Ma, J.; Chen, X.; Hu, Z.; Huo, K.; Chen, Y. *J. Phys. Chem. B* **2004**, *108*, 4024.
(80) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
(81) Fincher, C. R., Jr.; Chen, C.-E.; Heeger, A. J.; MacDiarmid, A. G. *Phys. Rev. Lett.* **1982**, *48*, 100.
(82) Yannoni, C. S.; Clarke, T. C. *Phys. Rev. Lett.* **1983**, *51*, 1191.
(83) Kahlert, H.; Leitner, O.; Leising, G. *Synth. Met.* **1992**, *83*, 179.
(84) Fogarasi, G.; Liu, R.; Pulay, P. *J. Phys. Chem.* **1993**, *97*, 4036.
(85) Hirata, S.; Torii, H.; Tasumi, M. *J. Chem. Phys.* **1995**, *103*, 8964.
(86) Choi, C. H.; Kertesz, M. *J. Chem. Phys.* **1997**, *107*, 6712.
(87) Bengu, E.; Marks, L. D. *Phys. Rev. Lett.* **2001**, *86*, 2385.

Table 3. Energy Differences (ΔE) between the Conventional HF (MP2) and Corresponding EC-MFCC Energies for Two Types of BN SWNTs with the 3-21G Basis Set

molecule	basis functions	ΔE (or total energy)[a] MFCC	EC-MFCC(10)	basis functions of the largest subsystem
		Zigzag (6,0)		
$n = 12$	1320	0.43 (0.92)	−0.12 (0.03)	912
$n = 16$	1752	1.11 (−7587.96552)	0.16 (−7587.96714)	912
$n = 32$	3480	−15141.14118 (−15169.30735)	−15141.14372 (−15169.31190)	912
$n = 64$	6936	−30275.70261 (−30331.99105)	−30275.70834 (−30332.00145)	912
$n = 128$	13848	−60544.82547 (−60657.35844)	−60544.83757 (−60657.38056)	912
		Armchair (3,3)		
$n = 16$	1752	−0.16	0.01	912
$n = 32$	3480	−15139.94237	−15139.94178	912
$n = 64$	6936	−30273.21193	−30273.21048	912
$n = 128$	13848	−60539.75106	−60539.74789	912

[a] For tubes ($n = 32$, 64, 128) without conventional energies and zigzag (6,0) tube ($n = 16$) without conventional MP2 energy, the total energies from MFCC and EC-MFCC calculations are presented. The energy differences (or total energies) from EC-MFCC-MP2 calculations are included in parentheses.

Table 4. Energy Differences (ΔE) between the Conventional HF (MP2) and Corresponding EC-MFCC Energies for Selected Biological Molecules

molecule[a]	basis set[b]	ΔE (mH)[c] MFCC	EC-MFCC (8,30)	basis functions of the largest subsystem
indinavir	6-31G (499)	5.20 (9.84)	0.72 (1.07)	215
adenine−thymine (A−T)	6-31G** (660)	19.42 (27.36)	−0.30 (−0.16)	345
cytosine−guanine (G−C)	6-31G** (655)	44.62 (51.67)	−0.23 (−0.05)	360
crambin (1cnr)	3-21G (3597)	670.02	4.93	594
gramicidin A ion channel (1grm)	3-21G (3000)	277.96	0.68	711
α-conotoxin pni1 (1pen)	3-21G (1192)	−1.31	4.43	555
α-conotoxin pnib from *Conus pennaceus* (1akg)	3-21G (1211)	7.88	6.61	555
R-conotoxin mii (1m2c)	3-21G (1268)	51.193	−10.34	547
trypsin inhibitor II (EETI II) (2eti)	3-21G (2164)	486.19	−38.09	604

[a] The fragmenting scheme for all these six proteins is similar to that adopted for glycine peptides (see Figure S1). The size of each fragment is about three residues. [b] The total number of basis functions included in parentheses. [c] The relative energies with respect to conventional HF energies, and the relative energies from EC-MFCC-MP2 calculations included in parentheses.

but optimized structures for longer nanotubes are required to extrapolate the geometry of an infinite nanotube.

For the *trans*-oligoenes $H(C_2H_2)_nH$ ($n = 16, 32$) and the zigzag (6,0) BN nanotubes, $H_6(B_6N_6)_nH_6$ ($n = 16, 32$), their structures optimized by the EC-MFCC approach at the MP2/6-31G* and HF/3-21G levels, respectively, are displayed in Figure 4. For *trans*-oligoenes, each fragment contains two C═C bonds, and each cap contains three C═C bonds. The fragmentation and capping scheme for BN nanotubes are the same as described earlier. For comparison, we also display in Figure 4 the geometry data of the $n = 16$ *trans*-oligoene and the $n = 16$ nanotube optimized by the conventional MP2 and HF methods, respectively. As one can see, for these two medium-sized molecules, the geometrical parameters derived from the EC-MFCC approach are remarkably close to those from the conventional methods. The average deviation is about ±0.001 Å for the bond lengths, ±1° for the bond angles, and ±2° for the dihedral angles. Thus, the EC-MFCC approach is ideally suitable for these two systems. As seen from the optimized structure of the *trans*-oligoene $H(C_2H_2)_{32}H$, the C═C and C−C bond lengths converge quickly from 1.347 to 1.366 Å and from 1.448 to 1.431 Å, respectively, as their positions go inner. Since the calculated C═C and C−C bond lengths in the central part of the $n = 32$ *trans*-oligoene are equal to those in the $n = 16$ *trans*-oligoene, the C═C and C−C bond lengths in the *trans*-PA can be predicted to be 1.366 and 1.431 Å, respectively, which is in good agreement with the observed bond lengths (1.36 and 1.44 Å). In contrast, a previous study using the extrapolation technique predicted the C═C and C−C bond lengths in the *trans*-PA to be 1.373 and 1.423 Å at the same theoretical level. For the $n = 32$ BN nanotube, one can see that the geometrical parameters of the hexagon converge rapidly to constant values as it moves toward the middle of the tube, and both the terminal and central hexagons have almost the same geometry data as those in the short $n = 16$ tube. Thus, on the basis of the optimized structure for the $n = 32$ BN nanotube the geometries of longer BN nanotubes can be reasonably constructed.

3.3. The Estimation of the Standard Heats of Formation for Some Organic Compounds. It has been known that most molecular properties of large molecules, such as the standard heats of formation $\Delta H^0_{f,T}$, can be derived from the additivity of bond properties or group properties.[70,71] In the group additivity scheme, a group is generally defined as a polyvalent atom in a molecule together with all of its ligands. The values for the contributions of groups to molecular properties are obtained from multilinear regression analyses of the thermochemical data of a large number of known compounds. By summing the contributions of all the groups in a given molecule, the molecular property of an unknown compound can be estimated. For general compounds, the simple group additivity scheme is often modified by introducing several types of corrections, which are

---

(88) Lide, D. R.; Kehiaian, H. V. *Handbook of Thermophysical and Thermochemical Data*; CRC Press: Boca Raton, FL, 1994.

(89) Dean, J. A. *Lange's Handbook of Chemistry*; McGraw-Hill: New York, 1999.

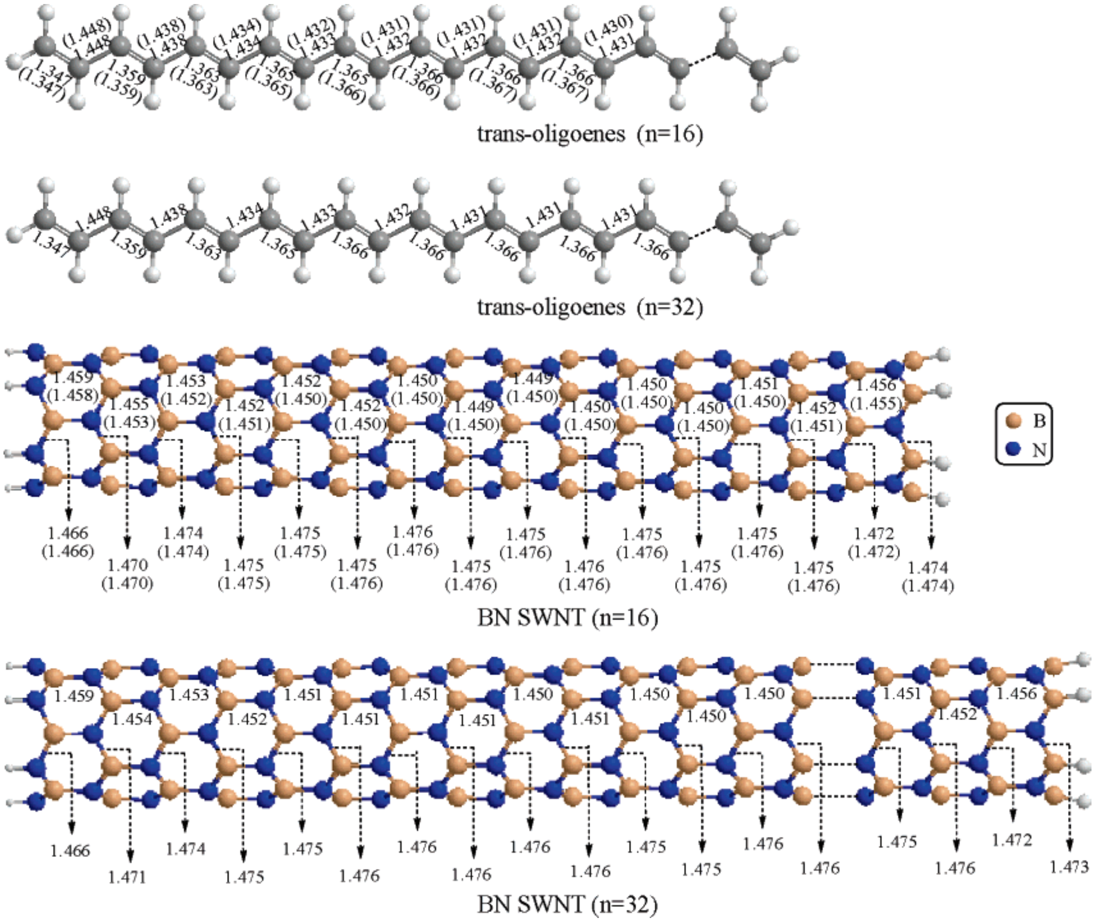


Figure 4. Optimized structure of *trans*-oligoenes ($n = 16, 32$) and zigzag (6,0) BN SWNTs ($n = 16, 32$). The *trans*-oligoenes are optimized at the MP2/6-31G* level, and BN SWNTs are optimized at the HF/3-21G level. The bond lengths (in Å) obtained by the conventional optimization methods at the same theoretical level are given in parentheses (the data of the $n = 16$ BN SWNT are from ref 79).

from the interactions of nonbonded, next-nearest neighboring groups. This group additivity scheme can provide satisfactory $\Delta H^0_{\mathrm{f},T}$ (and molar heat capacity, etc.) for structurally relatively simple organic compounds, with the average deviation of ±2.0 kJ/mol.[70,71] But for structurally complex systems, values of estimated $\Delta H^0_{\mathrm{f},T}$ deviate by ±12 kJ/mol from the experimentally observed values.[70,71] The success of eq 14 in predicting the total energy suggests an alternative approach for estimating the standard heats of formation of molecules. If the two-body correction can be neglected (this approximately holds for medium-sized molecules), the standard heats of formation of gas-phase species can be approximately predicted from the following equation:

$$\Delta H^0_{\mathrm{f},T} = \sum_I \Delta H^0_{\mathrm{f},T}(\text{capped } I) - \sum \Delta H^0_{\mathrm{f},T}(\text{conjugated caps}) \tag{16}$$

Clearly, this MFCC approach differs from the group additivity scheme in two ways. First, $\Delta H^0_{\mathrm{f},T}$ of an unknown compound can be directly estimated from related small molecules with experimental data. Second, since capped fragments or conjugated caps often contain several groups defined above, most of next-nearest neighboring interactions of groups are automatically incorporated. To illustrate the application of the MFCC approach, we have estimated values of $\Delta H^0_{\mathrm{f},T}$ at $T = 298\ K$ for 18 organic molecules shown in Figure 5, with the results listed in Table 5. The fragmentation and capping schemes for all these molecules are also shown in Figure 5, so that readers may easily do the calculations by themselves. For example, $\Delta H^0_{\mathrm{f},T}$ of 1-decanol can be estimated from those of 1-heptanol, heptane, and butane:

$$\Delta H^0_{\mathrm{f},T}(1-\text{decanol}) = \Delta H^0_{\mathrm{f},T}(1-\text{heptanol}) + \Delta H^0_{\mathrm{f},T}(\text{heptane}) - \Delta H^0_{\mathrm{f},T}(\text{butane}) = (-336.4) + (-187.7) - (-125.6) = -398.5\ \text{kJ/mol} \tag{17}$$

From Table 5, one can see that for all studied systems the mean deviation of estimated $\Delta H^0_{\mathrm{f},298}$ is about ±2.3 kJ/mol. In general, for structurally simple compounds the accuracy of the MFCC approach is comparable to that of the group additivity approach, but for more complex systems the MFCC approach may show better performance. For instance, for bibenzyl the deviation of $\Delta H^0_{\mathrm{f},298}$ predicted from the present approach is 5.8 kJ/mol, while that from the group additivity approach is 7.9 kJ/mol.[70] Furthermore, the MFCC approach can give correct predictions on the relative order of $\Delta H^0_{\mathrm{f},T}$ for some isomeric species. For example, the relative magnitude of $\Delta H^0_{\mathrm{f},T}$ between 2-methylnonane and 5-methylnonane is correctly predicted with the MFCC approach. On the other hand, it should be pointed out that the MFCC approach is only applicable for molecules with four or more polyvalent atoms (or groups). In some cases, where the experimental thermochemical data of subsystems are not available, the MFCC approach could not be used, too. The combination of the present approach with the traditional group

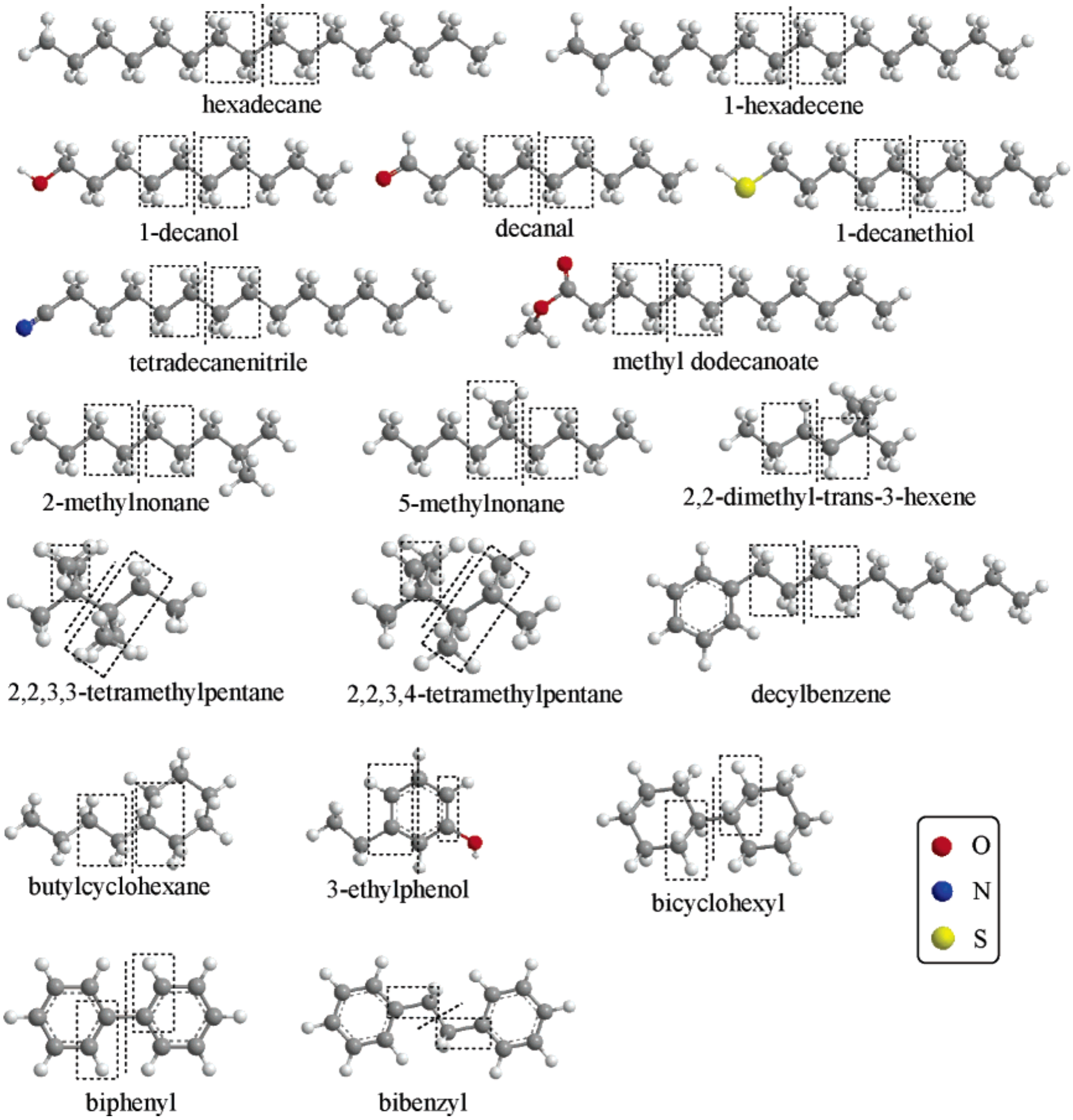


Figure 5. Selected organic compounds for the estimation of $\Delta H^0_{f,T}$. The fragmentation and capping schemes are also displayed. The cut bond is denoted by a broken line, and a cap is denoted by a frame.

Table 5. Standard Heats of Formation $\Delta H^0_{f,T}$ (kJ/mol) for Some Organic Molecules at $T = 298$ K

molecule	exptl[a]	estimated[b]	molecule	exptl[a]	estimated[b]
hexadecane	−374.8	−373.4 (1.4)	2,2-dimethyl-*trans*-3-hexene	−109.5	−110.3 (−0.8)
1-hexadecene	−248.5	−247.3 (1.2)	2,2,3,3-tetramethylpentane	−237.1	−237.4 (−0.3)
1-decanol	−396.4	−398.5 (−2.1)	2,2,3,4-tetramethylpentane	−236.9	−238.4 (−1.5)
decanal	−330.9	−326.1 (4.8)	decylbenzene	−138.6	−137.0 (1.6)
1-decanethiol	−211.5	−212.1 (−0.6)	butylcyclohexane	−213.3	−212.6 (0.7)
tetradecanenitrile	−174.9	−174.4 (0.5)	3-ethylphenol	−146.1	−149.1 (−3.0)
methyl dodecanoate	−614.8	−616.5 (−1.7)	bicyclohexyl	−215.7	−217.8 (−2.1)
2-methylnonane	−259.9	−256.9 (3.0)	biphenyl	181.4	185.8 (4.4)
5-methylnonane	−258.6	−253.4 (5.2)	bibenzyl	135.6	141.4 (5.8)

[a] Values from refs 88 and 89. [b] The relative standard heats of formation with respect to the corresponding experimental value included in parentheses.

additivity approach may be used to predict the thermochemical data for a large number of structurally complex compounds with good accuracy.

4. Conclusions

In this work, we present an efficient fragment-based approach for predicting the total energies of macromolecules at the HF and post-HF levels. The physical foundation of this approach is revealed by a new energy decomposition analysis of the total HF energy proposed in this work. Within this approach, the total energy of a macromolecule can be directly obtained from energy calculations on a series of subsystems, which are capped fragments, conjugated caps, and subsystems composed of two capped fragments. Since the computational cost of the present method scales linearly with the molecular size in large molecules, the total energy of systems with hundreds or thousands of atoms can now be approximately computed at the ab initio HF and MP2 levels with existing quantum chemistry programs. Our test calculations for a variety of medium-sized and even quite large-sized systems show that the present approach is able to reproduce the conventional HF and MP2 energies within a few millihartree in most cases. The accuracy of the present approach is competitive with other sophisticated linear-scaling methods.[3,7,12−14,20−66] Furthermore, with the EC-MFCC optimization technique described in this work, now we can perform geometry optimizations for large systems with hundreds of atoms at the HF or post-HF level. The optimized structures of very long oligomers of *trans*-PA and BN nanotubes have been obtained with this optimization technique, which are beyond the reach of traditional computational methods. It is expected that the use of the present optimization technique will greatly expand the application ranges of computational chemistry. In addition, the EC-MFCC approach is also applied to estimate the heats of formation for a number of organic systems. Satisfactory results can be obtained, if the molecule under study is properly fragmented and capped. Therefore, the EC-MFCC method provides an alternative approach to the traditional additivity rules based on either bond or group contributions for the estimation of thermochemical properties.

Compared to other fragment-based methods, the present approach has several advantages. First, the approach may be applicable to any existing quantum chemistry methods including HF, MP2, or CCSD, and even DFT methods. Second, the approach is computationally the simplest one among existing fragment-based methods or alternative linear-scaling methods. For example, the approach does not really construct the total density matrix of the whole macromolecule from the density matrixes of the capped fragments, like the ADMA[12−14] approach and other methods. In addition, the parallel implementation of the present approach could be highly efficient because calculations on constructed subsystems can be carried out independently. We would also like to address some inherent limitations of the EC-MFCC approach. This approach is not a “black-box” method like some linear-scaling methods,[32−38,59−65] since the user needs to construct fragments and caps manually. Of course, for a group of molecules made up of limited building blocks such as proteins, polymers, and nanotubes, an automated procedure for fragmenting these molecules and capping the fragments is feasible. The accuracy of this approach noticeably depends on the size of fragments and caps selected. Hence, one should calibrate the accuracy of this approach before applying the present approach to unknown large systems. In addition, the present approach may be less satisfactory for some polar and highly charged molecules, as discussed in subsection 3.1.2. Furthermore, the present approach may even break down for some conjugated systems and radicals, in which electrons are highly delocalized in the whole system. It should be noted that other linear-scaling approaches[3,7,12−14,20−66] might also face this problem.

To summarize, the EC-MFCC approach is demonstrated to be capable of predicting total energies and structures of very large molecules with good accuracy at the ab initio HF and post-HF levels. Furthermore, the present approach provides an appealing approach for chemists to understand and predict the physical and chemical properties of a huge amount of chemical compounds. Future developments would make this approach a promising tool for performing quantum chemistry calculations on very large molecules.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Grant Nos. 20373022, 20233020, and 20433020) and the National Basic Research Program (Grant No. 2004CB719901). We are grateful to the reviewers for their comments and good suggestions on the original manuscript.

Supporting Information Available: Fragmentation and capping schemes for some studied molecules, complete ref 72, Cartesian coordinates, and convention HF or MP2 energies of all compounds studied. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0427247