ACCOUNTS of chemical research

Article

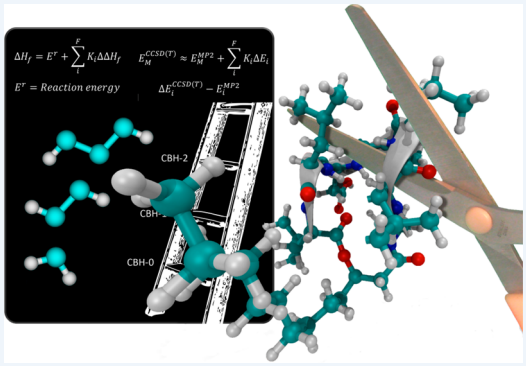
pubs.acs.org/accounts

The Successful Merger of Theoretical Thermochemistry with Fragment-Based Methods in Quantum Chemistry

Raghunath O. Ramabhadran and Krishnan Raghavachari*

Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States

CONSPECTUS: Quantum chemistry and electronic structure theory have proven to be essential tools to the experimental chemist, in terms of both *a priori* predictions that pave the way for designing new experiments and rationalizing experimental observations *a posteriori*. Translating the well-established success of electronic structure theory in obtaining the structures and energies of small chemical systems to increasingly larger molecules is an exciting and ongoing central theme of research in quantum chemistry. However, the prohibitive computational scaling of highly accurate *ab initio* electronic structure methods poses a fundamental challenge to this research endeavor. This scenario necessitates an indirect fragment-based approach wherein a large molecule is divided into small fragments and is subsequently reassembled to compute its energy accurately. In our quest to further reduce the computational expense associated with the fragment-based methods and overall enhance the applicability of electronic structure methods to large molecules, we realized that the broad ideas involved in a different area, theoretical thermochemistry, are transferable to the area of fragment-based methods.



This Account focuses on the effective merger of these two disparate frontiers in quantum chemistry and how new concepts inspired by theoretical thermochemistry significantly reduce the total number of electronic structure calculations needed to be performed as part of a fragment-based method without any appreciable loss of accuracy. Throughout, the generalized connectivity based hierarchy (CBH), which we developed to solve a long-standing problem in theoretical thermochemistry, serves as the linchpin in this merger. The accuracy of our method is based on two strong foundations: (a) the apt utilization of systematic and sophisticated error-canceling schemes via CBH that result in an optimal cutting scheme at any given level of fragmentation and (b) the use of a less expensive second layer of electronic structure method to recover all the missing long-range interactions in the parent large molecule.

Overall, the work featured here dramatically decreases the computational expense and empowers the execution of very accurate *ab initio* calculations (gold-standard CCSD(T)) on large molecules and thereby facilitates sophisticated electronic structure applications to a wide range of important chemical problems.

1. INTRODUCTION

Phenomenal progress in new method developments along with ever-advancing computer technology have made computational quantum chemistry an indispensable tool in chemistry, physics, biology, and material science.[1] In particular, electronic structure theory today rivals experiments for the accurate prediction of geometries, energies, and properties of small molecules.[2] To expand its scope further, a major thrust in contemporary quantum chemistry is on accurately computing the energies and properties of larger molecules by overcoming the inherent steep scaling associated with traditional *ab initio* electronic structure methods.[3,4]

Two distinct areas within quantum chemistry, theoretical thermochemistry[5−12] and fragment-based methods,[13−30] have been at the forefront of enabling applications on larger molecules. While theoretical thermochemistry strives to accurately compute the thermodynamic properties (e.g., enthalpies of formations) of increasingly larger molecules, the objective in fragment-based methods is to calculate the energy of a large molecule by dividing it into smaller fragments.

Herein, we elucidate how the glue of error-cancellation binds together these two areas and results in a key concept in fragment-based methods inspired from theoretical thermochemistry. This merger results in new avenues for performing high-accuracy calculations on large molecules with a significantly reduced computational expense.

2. THEORETICAL THERMOCHEMISTRY

Historically, electronic structure theory has been used for the accurate prediction of thermochemical properties of molecules for over 40 years. Among the prominent highly accurate *ab initio*-based methods are Gaussian-*n* (G*n*),[6] HEAT,[7] Weizmann-*n* (W*n*),[8] coupled cluster-based extrapolation methods,[9] complete basis set methods (CBS),[10] multicoefficient methods,[11] and the correlation-consistent composite approach (ccCA).[12] They have enabled the computation of highly

Received: August 9, 2014
Published: November 13, 2014

   dx.doi.org/10.1021/ar500294s | Acc. Chem. Res. 2014, 47, 3596−3604

accurate (from 1−2 kcal/mol to 1 kJ/mol) enthalpies of formations for small molecules.

However, it is not practical to directly apply these accurate yet computationally demanding methods on large molecules. Thus, when left with the compulsion of using approximate methods (such as DFT) on larger molecules, the use of effective error-cancellation strategies becomes crucial to maintain chemical accuracy (±1−2 kcal/mol).

In this context, Pople and co-workers' 1970 work on the isodesmic bond separation (IBS) scheme is a landmark paper in theoretical thermochemistry.[31] It is among the first publications to explicitly illustrate the role of efficient error-cancellation in accurately computing the thermodynamic properties of organic molecules. The essential idea in an IBS scheme is to "extract all the heavy-atom bonds in a molecule as their simplest valence satisfied molecules". Once the IBS scheme is generated for a molecule, its heat of formation is then calculated from the following two steps: (a) perform electronic structure computations on all the molecules in the scheme to get the reaction energy and (b) use the reaction energy in conjunction with the experimental enthalpies of formations on the reference molecules and apply Hess's law.

Two key aspects learned from the IBS scheme are that (a) it is unique, that is, while several different isodesmic reaction schemes can be written for a molecule, only one IBS scheme can be generated, and (b) it has a simple structure-based definition and hence it does not involve a manual attempt to balance the atom-types, bond types, or hybridizations. These facets, along with the error-cancellation it provides, have afforded the IBS to be a useful scheme even today and is utilized in several modern theoretical thermochemical methods such as in the ATOMIC protocol[32] and in group additivity-based schemes.[33]

As molecules get larger in size, more advanced error-canceling schemes become necessary. Thus, starting with the homodesmotic scheme of George et al.,[34] a considerable amount of effort has been devoted to develop clever reaction schemes that improve upon the IBS scheme. They achieve significant error-cancellation and result in accurate heats of formation for several classes of organic molecules. Yet, despite their success, some of them are specific only to certain organic functional groups, and more importantly, the widely used homodesmotic scheme was found to have definition-based inconsistencies. These points were explicitly brought to light in an important paper in 2009 by Wheeler, Houk, Schleyer, and Allen wherein they recognized the necessity for greater uniformity and generality in such reaction schemes.[35] Consequently, they developed a general hybridization-based hierarchy of homodesmotic reactions[34] for closed shell hydrocarbons. In their hierarchy, they used predefined reactants and products and achieved an increased balance in the hybridization and the covalent bonding environment of the carbon atoms for closed-shell hydrocarbons.[35]

However, in extending their hierarchy beyond hydrocarbons to all classes of organic molecules, they rightly acknowledged the enormous structural variety present in organic chemistry and stated that "The primary challenge for such extensions is the growth in the number of formal bond types and fragments involved in the definitions of the reaction classes". Thus, the development of a general hierarchy applicable to all classes of organic molecules remained an open problem.

In 2011, we developed the fully automated generalized connectivity-based hierarchy (CBH) to address this problem.[36] CBH, as its name implies, is a thermochemical hierarchy based on the connectivity of the atoms in a molecule. It overcomes the problem of complex definitions used in the construction of the reaction schemes, as well as the arduous requirement of balancing hybridization- and bond-types by hand, which had previously detracted the development of a reliable and automated thermochemical hierarchy for all organic molecules. Furthermore, it is very easy to construct the hierarchy, either by hand for smaller molecules or via an automated computer program, thereby making CBH very user-friendly to accurately predict the enthalpies of formations of organic molecules. Finally, CBH, which naturally stems from the chemical structure of a molecule, helped us to identify the reason for the definition-based inconsistency noted by Wheeler et al.[35] in the homodesmotic schemes. Similar ideas in a different context have also been developed by Deev and Collins[26a] and by Lee and Bettens[26b] (*vide infra*).

As previously illustrated in detail in ref 36, it is useful to envision the different levels of CBH as the rungs of a ladder, such that ascending the rungs of the hierarchy increasingly preserves the chemical environment (i.e., a better matching of the bond-types and hybridization-types is automatically achieved) of a molecule (Figure 1). We call these different rungs CBH-0, CBH-1, CBH-2, CBH-3, etc.

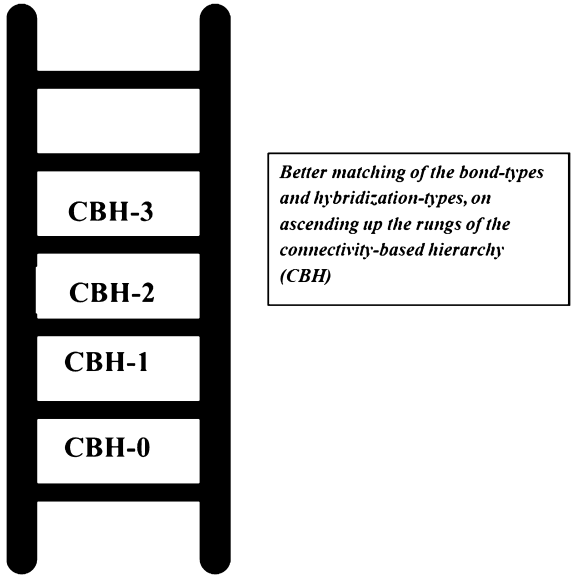


Figure 1. Envisioning the reaction schemes obtained using CBH as the rungs of a ladder.

The rungs alternate between being atom-centric (CBH-0, CBH-2, etc.) and bond-centric (CBH-1, CBH-3, etc.). The atom-centric CBH-0 rung for any molecule is constructed by extracting all the heavy atoms, terminating the open valences with hydrogen atoms (Figure 2a). Similarly, the bond-centric CBH-1 is generated by extracting all the heavy-atom bonds and terminating them with hydrogen atoms (Figure 2b). It turns out that CBH-1 is exactly the same as Pople's IBS scheme.[31] At the next atom-centric rung, CBH-2, we preserve the immediate chemical environment of each heavy atom, that is, it is constructed by extracting all the heavy atoms, maintaining their atom connectivities with neighboring heavy atoms, and then hydrogen terminating them (Figure 2c). Since this rung preserves the immediate chemical environment of an atom, it

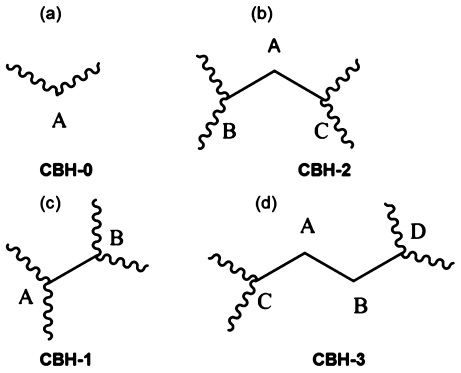


Figure 2. Generic pictorial representation of alternating atom-centric and bond-centric rungs in CBH: (a) CBH-0, (b) CBH-1, (c) CBH-2, and (d) CBH-3. Adapted with permission from ref 37. Copyright 2012 American Chemical Society.

can appropriately be termed as the *isoatomic reaction scheme*. Finally, the bond-centric CBH-3 rung preserves the immediate chemical environment of a heavy-atom bond and is generated by extracting all the heavy-atom bonds, maintaining their connectivities with neighboring heavy atoms, and then hydrogen terminating them (Figure 2d). Higher rungs such as CBH-4 and beyond can also be defined by similar extensions, but we found that for commonly encountered organic molecules and biomonomers containing about 20 heavy-atoms, CBH-2 or CBH-3 usually suffices.[37,38]

It can readily be seen from the discussion thus far that, at every rung, certain regions of the molecule are overcounted in the construction of CBH. This is due to the overlapping nature of the extracted atoms or bonds. In order to take this into account, we need to add back the overcounted molecules to the reactant side of the chemical equations to balance them. For example, at CBH-1, the simplest valence satisfied hydrides of heavy atoms (ammonia for N, methane for C, etc.) are added. This is easily and elegantly extended for all CBH rungs by noticing a recursive relationship between the products at one rung, and the reactants at the next rung (Figure 3), as

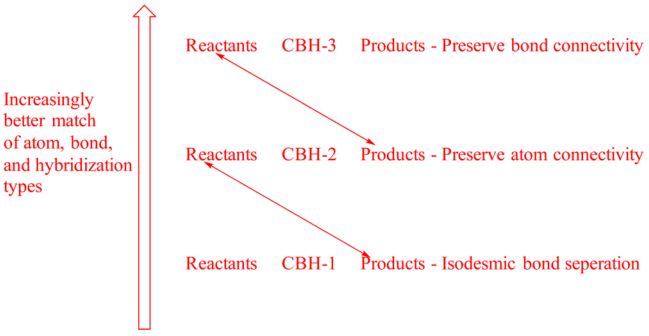


Figure 3. Showing the recursive connection between various rungs of CBH. Products from a lower rung are reactants at a higher rung. Adapted with permission from ref 36. Copyright 2011 American Chemical Society.

documented in our earlier work.[36] The recursive relationship naturally arises since we increasingly preserve the chemical environment of larger parts of a molecule on going up in the hierarchy.

The recursion phenomenon is strictly valid for monocyclic rings without any terminal groups or branching. Only two more facets are needed to generalize CBH to all classes of compounds encountered in organic and bio-organic chemistry. (a) The first facet concerns the cancellation of terminal moieties. A terminal moiety in an organic molecule can be defined as *having fewer than two heavy atom bonds*.[36] At different rungs of CBH, different molecules represent these terminal moieties (which are products at a lower rung), and these molecules do not appear in the next rung of the hierarchy. (b) The second facet is related to branched molecules. As mentioned by us previously,[36] at any branching point in an organic molecule, the atom at the branching point is attached to one (or more) additional heavy atom(s) in comparison to an atom not at the branching point (Figure 4).

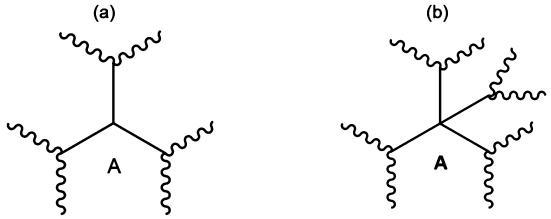


Figure 4. (a) A generic representation of a branch point. Here the branching point is the atom A. The molecule representing this branch point as a reactant is counted twice at the bond centric rungs. (b) A generic representation of two branch points on an atom. Reproduced with permission from ref 36. Copyright 2011 American Chemical Society.

Hence, an additional covalent bond needs to be taken into account for each branching point. To do this, we first identify molecules that represent the branch points and then adjust their coefficients in the bond-centric rungs (by counting them twice) when they occur as reactants. Notice that this feature is applicable only at the bond centric-rungs, that is, at CBH-1, CBH-3, etc., since branching occurs only with bonds. The CBH schemes for cyclohexanone (Figure 5), demonstrate these

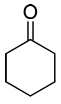


Figure 5. Cyclohexanone. Reproduced with permission from ref 36. Copyright 2011 American Chemical Society.

features (recursion, terminal moiety cancellation, and counting twice at the branch point). More examples explaining all these facets in great detail can be found in ref 36.

CBH-0:

$$C_6H_{10}O + 8H_2 \rightarrow 6CH_4 + H_2O \quad (1)$$

CBH-1: Here we have a combination of terminal moiety cancellation (H_2O representing the terminal moiety) and counting twice at the branch point (the carbonyl carbon is the branch point, with methane representing the corresponding molecule), along with recursion.

$$C_6H_{10}O + 7CH_4 \rightarrow 6C_2H_6 + CH_2O \text{ (formaldehyde)} \quad (2)$$

CBH-2:

$$C_6H_{10}O + 6C_2H_6 \rightarrow 5C_3H_8 + C_3H_6O \text{ (acetone)} \quad (3)$$

dx.doi.org/10.1021/ar500294s | Acc. Chem. Res. 2014, 47, 3596−3604

CBH-3: Here, acetone is the molecule representing both the branch point and the terminal moiety; thus, it gets counted twice and canceled once, eventually yielding,

$$C_6H_{10}O + 5C_3H_8 + C_3H_6O \rightarrow 4C_4H_{10}\ (\text{n-butane}) + 2C_4H_8O\ (\text{2-butanone}) \quad (4)$$

Overall, given a chemical structure of a molecule, the CBH schemes can be readily generated by a computer program based on merely the logic of recursion, cancellation of the terminal moieties, and counting twice at the branch point. Our earlier work[36,37] has shown that CBH increasingly preserves the bond-types and hybridizations for any organic molecule (hydrocarbons and nonhydrocarbons alike) exactly the same way Wheeler et al.'s hierarchy[35] does for hydrocarbons. Additionally, CBH offers a structural basis to recover the IBS scheme of Pople and co-workers[31] and provides the simplest and unique set of isodesmic and homodesmotic reaction schemes that naturally arise from a molecule's chemical structures, even though many different arbitrary isodesmic/homodesmotic reaction schemes can otherwise be written. Lastly, once the CBH schemes were constructed, using experimental heats of formations for the reference species and Hess's law, highly accurate heats of formation for all classes of nonaromatic organic and bio-organic molecules were obtained at CBH-2 and CBH-3 rungs. Since the focus of this Account is the merger of theoretical thermochemistry with fragment-based methods, and not solely thermochemistry, we do not discuss the details of our excellent thermochemical results here, which can be obtained from refs 36−40.

3. A NEW CONCEPT IN FRAGMENT-BASED METHODS AS A RESULT OF ITS MERGER WITH THEORETICAL THERMOCHEMISTRY: FRAGMENTS IN THEIR EQUILIBRIUM GEOMETRIES

The wide applicability of highly accurate *ab initio* electronic structure calculations have traditionally been limited by their prohibitive computational scaling.[3,4] The N^7 scaling of the CCSD(T) method[4] is a nice case in point. While it is termed as the gold-standard in quantum chemistry and is the method of choice for the accurate computation of chemical bond energies, it can become expensive (even with commonly used cc-pV*n*Z type basis functions) for molecules containing 8−10 heavy atoms and can be prohibitive for systems containing more than 12−15 heavy atoms.

The idea of fragmenting a larger molecule into smaller pieces and then reassembling the units to compute its energy has proven to be a very useful strategy to indirectly perform accurate *ab initio* calculations on large systems.[13−30] All such methods share the common central theme of fragmenting a larger molecule and differ in the details of how: (a) the partitioning of the large molecule is carried out, (b) "subsystems" formed from this larger molecule are defined, generally classified as "primary subsystems" (which are the overlapping fragments obtained by chopping the larger molecule) and "derivative subsystems" (which take care of the overcounting due to the overlap between the primary susbsystems), and (c) the energies of the subsystems are assembled.[28]

Interestingly, the construction of error canceling reaction schemes as part of an automated thermochemical hierarchy, such as CBH can be thought as being analogous to fragmenting a larger molecule into smaller entities.[26,41] The reference molecules in the thermochemical hierarchy are the equivalents of the subsystems in a fragment-based method: those on the right-hand-side of the CBH equations represent the "primary subsystems", while those on the left-hand-side represent the "derivative subsystems" that are added to the parent molecule to account for the overcounting.

When CBH is used in thermochemistry, the parent molecule is always present in the equations, and thus a calculation on the whole molecule is carried out at a defined level of theory. The objective therein is to achieve accuracy by error cancellation, and computational efficiency becomes a secondary issue. In contrast, fragment-based methods frequently use a second layer of theory that is computationally expedient on the whole molecule (or larger fragments) to achieve results (by extrapolation) that are more accurate as well as computationally efficient.[28] We realized that combining the best of both the worlds, that of using ground state geometries from theoretical thermochemistry and that of using a second layer from fragment-based methods, can be a very useful new concept.[41] The advantage of this merger is that it can enormously reduce the number of electronic structure calculations needed to compute the energies of larger molecules without any loss of accuracy (*vide infra*). Additionally when using CBH as a fragment-based method, since the largest fragment size at any given CBH rung is independent of the size of the large molecule under consideration, we can achieve an enormous reduction in the computational expense.

While qualitatively similar at a broad level, the CBH formalism for fragment-based methods has some significant advantages over the traditional fragmentation schemes. In particular, at any rung of CBH hierarchy, the reference molecules represent the *optimal* cutting scheme to achieve maximum error cancellation at that level of fragmentation. The higher CBH rungs then represent fragmentation schemes that yield smoothly increasing fragment size while progressively augmenting the efficiency of error cancellation. Related ideas have been previously proposed by Deev and Collins in their systematic fragmentation method,[26a] as well as by Lee and Bettens[26b] in their isodesmic fragmentation method, though there are noteworthy differences (*vide infra*). In addition, the CBH schemes cut across multiple bonds between heavy atoms in a manner exactly analogous to cutting single bonds. This is normally not employed in traditional fragmentation schemes due to the difficulties involved in defining "link atoms" in such situations.

There is another major difference between the CBH formalism and the traditional fragmentation schemes. The geometries of the subsystems in fragment-based methods (including the systematic fragmentation method and the isodesmic fragmentation method[26,27]) are obtained in the same geometry as found in the large parent molecule being fragmented.[13−30] Consequently, the energies of the fragments need to be computed repeatedly every time the energy of a new parent molecule is desired. However, the reference molecules in a thermochemical hierarchy (CBH) are computed in their optimized equilibrium geometries.[36] This is because the experimental enthalpies of formations of these reference species (used to get the enthalpy of formation of the parent molecule) are valid only at their equilibrium geometries. Overall, since many large molecules share the same smaller optimized reference species, repetitive electronic structure computations are avoided in a thermochemical hierarchy such as CBH.

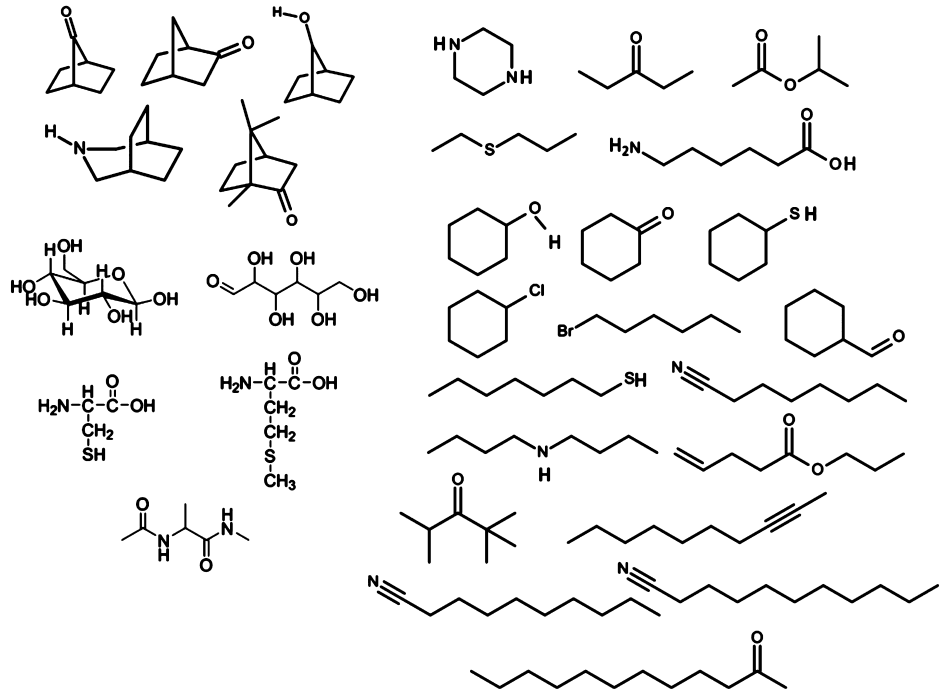


Figure 6. Structures of the organic compounds used in the diverse 30 molecule test set, used in ref 41. Reproduced with permission from ref 41. Copyright 2013 American Chemical Society.

Using these concepts, we recently obtained accurate CCSD(T) energies for large molecules at a significantly reduced computational cost using the CBH method, and in conjunction with MP2 energies.[41] To illustrate the use of CBH as a fragment-based method, we hereafter loosely use the phrase "fragments" to allude to the "reference molecules" in the CBH scheme (i.e., both primary and derivative subsystems).

From the CBH schemes defined in section 2 (and in refs 36 and 37), the reaction energy for any generic molecule at any CBH rung n can be written as

$$\text{reaction energy(CBH-}n) = \sum_{i}^{\text{F}} K_i E_i - E_{\text{M}} \tag{5}$$

The same notations, used in ref 41 are used here, that is, E_{M} is the energy of the parent molecule M at a given level of theory, E_i is the energy of the ith fragment at the same level of theory, and K_i is the signed stoichiometric coefficient of the ith fragment., that is, K_i is a positive integer if i is a product fragment and is a negative integer if i is a reactant fragment. The summation is carried out over all the fragments F generated from M.

For any given basis set and at CBH-2 and higher-rungs, we had earlier demonstrated in ref 37 that the CCSD(T) reaction energies were very similar to MP2 reaction energies for a wide range of organic molecules. The mean absolute deviation between the MP2 and CCSD(T) reaction energies (using 6-31+G(d,p) or aug-cc-pVDZ basis sets) for the 30 molecule test set (*vide infra*) is only ∼0.3 kcal/mol at CBH-2 and ∼0.2 kcal/mol at CBH-3, that is,

$$\text{reaction energy}^{\text{CCSD(T)}}_{\text{CBH-}n} \approx \text{reaction energy}^{\text{MP2}}_{\text{CBH-}n} \tag{6}$$

Hence,

$$E_{\text{M}}^{\text{CCSD(T)}} \approx E_{\text{M}}^{\text{MP2}} + \sum_{i}^{\text{F}} K_i \Delta E_i \tag{7}$$

where

$$\Delta E_i = E_i^{\text{CCSD(T)}} - E_i^{\text{MP2}} \tag{8}$$

Thus, as mentioned by us previously,[41] the use of CBH as a fragment-based method results in the finding that "*the MP2 method is a suitable starting point for approximating to the highly accurate CCSD(T) energies, and the CCSD(T) energy of a larger molecule M can be accurately obtained (vide infra) without having to perform the expensive CCSD(T) calculation on M. The bottleneck CCSD(T) calculation now involves the largest fragment generated in the selected CBH scheme—which is independent of the size of M*". Furthermore, since equilibrium geometries for the fragments are used, once the energy for a fragment has been computed, it can be reused any number of times from a look-up table, whenever the same fragment occurs in another large molecule.

It is straightforward to use CBH to get extrapolated CCSD(T) energies for large organic molecules. At any given CBH-n rung ($n > 1$) and with any basis set, the protocol is as follows (adapted in part from our work in ref 41): (i) Generate the CBH reaction scheme for M and obtain the fragments. (ii) Obtain the equilibrium geometries of M and the fragments at a reasonable level of theory. DFT methods are adequate and computationally inexpensive for this purpose, and we specifically use the B3LYP/6-31G(2df,p) level of theory for optimizing the geometries. Even though we use the fragments in their equilibrium geometries, we find that there is no loss of accuracy (*vide infra*). (iii) Perform CCSD(T) calculations on the fragments to get ΔE_i (defined in eq 8). Note that the necessary MP2 energies are available for free during the CCSD(T) calculation. (iv) Perform an MP2 calculation on the full molecule, M, to obtain $E_{\text{M}}^{\text{MP2}}$. Only a single MP2 calculation

dx.doi.org/10.1021/ar500294s | *Acc. Chem. Res.* 2014, 47, 3596−3604

needs to be performed, making the procedure highly convenient. (v) Finally, use eq 7 to get the extrapolated CCSD(T) energy.

To test the accuracy of CBH to get extrapolated CCSD(T) energies, we included all the varied 27 nonaromatic molecules from our earlier test set for obtaining accurate enthalpies of formation[36,37,39] and added glucose (ring and open forms) and alanine dipeptide.[39] The 30 molecule test we assembled (Figure 6) was further divided into test set A and test set B. Test set A has the 20 common organic molecules, and test set B contains the more challenging systems with ring strain, multiple heteroatoms, or biomolecules with internal hydrogen bonds.

The minimum number of heavy atoms in the test set is 6, and the maximum number is 13. We restricted the test set to containing only about 15 heavy atoms, since direct CCSD(T) computations (which are necessary to assess the errors in the extrapolated CCSD(T) energies) on even larger molecules may be prohibitive. For managing the computational cost and to ensure further applicability to even larger molecules, we tested the method using double-ζ quality 6-31+G(d,p) and aug-cc-pVDZ basis sets. To prove that the method works accurately with larger basis sets, we also used larger triple-ζ basis sets for a subset of the molecules in the test set (see SI of ref 41).

A glance at Table 1 reveals that the mean absolute error (error is defined as the energy obtained with the full calculation − energy obtained using an extrapolated approach) for all 30 molecules is as small as 0.19 kcal/mol at CBH-3 with the aug-cc-pVDZ basis set. Even at CBH-2, which involves smaller fragments, the error is only 0.27 kcal/mol with the aug-cc-pVDZ basis set. Table 1 also tells us that irrespective of the basis set, be it Pople-style or Dunning style, we readily get sub-kilocalorie per mole accuracy at a significantly diminished computational cost. As expected, the errors with the simpler molecules in test set A (Table 1) are excellent, with mean absolute errors throughout being less than 0.20 kcal/mol. With the more interesting molecules in test set B, it is remarkable that the mean absolute errors still fall significantly within chemical accuracy. The individual performances of the molecules are shown in Table 2.

Table 1. Mean Absolute Errors (kcal/mol) between the Full CCSD(T) Energies and CCSD(T) Energies Obtained by Extrapolation in Ref 41 for Test Sets A and B[a]

test set	CBH-2, 6-31+G(d,p)	CBH-3, 6-31+G(d,p)	CBH-2, aug-cc-pVDZ	CBH-3, aug-cc-pVDZ
A(20 molecules)	0.19	0.15	0.17	0.14
B(10 molecules)	0.66	0.51	0.55	0.30

Adapted with permission from ref 41. Copyright 2013 American Chemical Society. [a]Test set A corresponds to the first 20 molecules (piperazine to decyl methyl ketone) in Table 2, and test set B corresponds to the next 10 molecules (camphor to 3-azabicyclo[3.2.2]nonane).

The first 20 molecules listed in Table 2 correspond to test set A, and the next 10 molecules belong to test set B. The largest error from test set A is only 0.56 kcal/mol (for cyclohexanone with CBH-2 using 6-31+G(d,p)). For the molecules in test set B, strained camphor produces the largest errors (1.60 kcal/mol). The performance obtained with both the ring and the open forms of glucose is particularly satisfactory. With as many as six oxygens and ample scope for intramolecular hydrogen bonds, glucose presented the possibility of error-accumulation instead of error-cancellation. Yet, sub-kilocalorie per mole accuracy is achieved with the ring form and the open form (Table 2), thus showcasing the success of our method. More detailed insights on the individual performance of some of the other molecules in the test set, extrapolated CCSD energies, and the poor performance with the CBH-1 rung (thus testifying the need for appropriate error cancellation from CBH-2 and higher rungs) are given in ref 41. It should also be emphasized here that the current formulation of CBH as a fragment-based method is applicable only for energetic minima on the potential energy surface since the fragment energies are constant and do not vary as the geometry of the parent molecule changes.

At first sight, the excellent performance of our method can be thought as being serendipitous, since we use the fragments in their equilibrium geometries. An argument can naively be presented that since the fragments are not maintained in the same chemical environment as found in the parent molecule larger errors should be expected and that any good performance is purely coincidental. However, a careful observation of our approach reveals that the success of our method is based on two strong foundations: (a) the use of a less expensive second layer (MP2) of electronic structure method to recover all the interactions in the parent large molecule[42] and (b) the apt utilization of a systematic error-canceling scheme via CBH (CBH-2 and higher rungs), which ensures that the residual errors in the ΔE_i term in eq 8 are small and an optimal fragmentation scheme is determined. The slightly larger errors (still within 2 kcal/mol) observed with camphor may be partly due to MP2 not being able to take into account the differences in strain energy between the fragments and the parent molecule. This interesting aspect is an object of future investigation.

In some previous applications of fragment-based methods to cyclic molecules,[26b,27] some difficulties have been noted at higher levels (i.e., rungs) due to the intrinsic imbalance created as a result of fragmenting a cyclic molecule into acyclic fragments. This has been ascribed to two reasons by Collins and Bettens:[26,27] (1) from link atoms in a fragment coming too close to each other as the fragment size increases and (2) from some group−group interactions being overcounted in the fragments compared with the parent molecule. *However, our method is not seriously affected by such factors due to two principal reasons.* (1) Spurious interactions between capping hydrogens coming closer does not occur since we use equilibrium geometries for the fragments (inspired by theoretical thermochemistry), and (2) the overcounting of any group−group interactions in the fragments is also effectively canceled since the fragments are treated with CCSD(T) and MP2, and it is only the *difference* between them that counts. Thus, a careful inspection of Table 2 does not show any meaningful deterioration for cyclic molecules on going from CBH-2 to CBH-3.

Having established its accuracy, it is worthwhile to explicitly quantify the huge benefits gained in the computational expense by using our method. The N^7 scaling of the full molecule at the CCSD(T) level is completely avoided. Instead, it is replaced by an MP2 calculation (formal scaling as N^5) of the full molecule. While CCSD(T) calculations are required for each fragment, the number of fragments for any organic or bio-organic molecule grows linearly with the size of the system. Thus, the CCSD(T) part of the calculation on the fragments grows linearly with the size of the molecule. Consider the example of decyl methyl ketone ($C_{12}H_{24}O$). We obtain a speedup of more

Table 2. Listing of the Errors (kcal/mol) between the Full CCSD(T) Energies and CCSD(T) Energies Obtained by Extrapolation in Ref 41

molecular formula	chemical name	CBH-2, 631+G(d,p)	CBH-3, 6-31+G(d,p)	CBH-2, aug-cc-pVDZ	CBH-3, aug-cc-pVDZ
$C_4H_{10}N_2$	piperezine	0.36	0.18	0.40	0.12
$C_5H_{10}O$	3−pentanone	−0.07	−0.01	−0.10	0.00
$C_5H_{10}O_2$	isopropyl acetate	−0.05	−0.01	0.00	−0.01
$C_5H_{12}S$	ethyl propyl sulfide	0.03	−0.05	0.08	−0.07
$C_6H_{13}NO_2$	6−aminohexanoic acid	−0.15	−0.04	−0.24	−0.10
$C_6H_{12}O$	cyclohexanol	0.37	0.30	0.29	0.14
$C_6H_{10}O$	cyclohexanone	0.28	0.38	0.12	0.22
$C_6H_{12}S$	cyclohexanethiol	0.56	0.33	0.46	0.17
$C_6H_{11}Cl$	cyclohexyl chloride	0.49	0.42	0.36	0.24
$C_6H_{13}Br$	n−hexyl bromide	−0.02	−0.05	−0.04	−0.08
$C_7H_{12}O$	cyclohexanal	0.38	0.45	0.19	0.26
$C_7H_{16}S$	1−heptanethiol	−0.08	−0.10	−0.06	−0.15
$C_8H_{15}N$	octanenitrile	−0.04	−0.01	−0.06	−0.07
$C_8H_{19}N$	dibutylamine	−0.11	0.18	−0.03	−0.16
$C_8H_{14}O_2$	propyl pent−4−enoate	−0.26	−0.11	−0.4	−0.23
$C_8H_{16}O$	t−butyl isopropyl ketone	0.08	0.04	−0.06	0.09
$C_{10}H_{18}$	2−decyne	0.07	0.01	0.00	−0.07
$C_{10}H_{19}N$	caprinitrile	−0.11	−0.06	−0.12	−0.13
$C_{11}H_{21}N$	1−cyanodecane	−0.14	−0.08	−0.15	−0.16
$C_{12}H_{24}O$	decyl methyl ketone	−0.31	−0.25	−0.29	−0.25
$C_{10}H_{16}O$	camphor	1.60	1.31	1.21	0.81
$C_6H_{12}O_6$	glucose (ring)	−0.63	−0.04	−0.74	−0.04
$C_6H_{12}N_2O_2$	alanine dipeptide	−0.21	0.06	−0.31	0.09
$C_6H_{12}O_6$	glucose (open)	−0.25	0.14	−0.5	0.06
$C_3H_7NO_2S$	cysteine	0.28	0.30	0.12	0.28
$C_5H_{11}NO_2S$	methionine	0.00	0.10	−0.09	0.06
$C_7H_{12}O$	7-norbornanol	0.98	0.85	0.74	0.47
$C_7H_{10}O$	2-norbornanone	1.16	1.11	0.83	0.76
$C_7H_{10}O$	7-norbornanone	0.87	0.79	0.46	0.41
$C_8H_{15}N$	3-azabicyclo[3.2.2]nonane	0.64	0.39	0.52	0.07
mean absolute error		**0.35**	**0.27**	**0.30**	**0.19**

Reproduced with permission from ref 41. Copyright 2013American Chemical Society.

than a factor of 10 (∼15) using the aug-cc-pVDZ basis set. The speedup will be substantially greater using larger basis sets or for larger molecules. Ultimately, CCSD(T) calculations using our approach are possible for any system where MP2 calculations are feasible.

The convenience of using equilibrium geometries is very clearly demonstrated by both acetone and 2-butanone, common fragments shared by many larger molecules in the test set such as cyclohexanone, 2-norbornanone, 7-nororna-none, camphor, etc. The energies of these fragments are calculated only once, thereby avoiding tedious bookkeeping and repetitive calculations. In fact, as we endeavor to calculate the energies of new and even larger molecules, the energies of most fragments would have been already calculated and can be retrieved from a database of precomputed CCSD(T) energies.

4. CONCLUSIONS AND FUTURE OUTLOOK

In this Account, we have described our work on the successful coming together of theoretical thermochemistry and fragment-based methods in quantum chemistry. This merger, which introduces a new computationally cost-effective concept in the field of fragment-based methods, was enabled by the generalized connectivity-based hierarchy (CBH).

The analogy between error-canceling reaction schemes and fragmenting a larger molecule into smaller pieces permitted the use of CBH as a fragment-based method. In this process, the counterintuitive concept of using the fragments in their equilibrium geometries, an inspiration from theoretical thermochemistry, was introduced to fragment-based methods. The concept proved to be highly useful in massively bringing down the total number of electronic structure calculations that need to be performed as part of a conventional fragment-based method. Moreover, since the largest fragment obtained from CBH at any given rung is independent of the size of the large molecule (whose energy is desired), a prominent gain in the computational savings is achieved without any loss of accuracy.

As a proof of principle, we obtained highly accurate extrapolated CCSD(T) energies of 30 disparate nonaromatic organic and biomolecules, using CBH in conjunction with MP2 energies. The current implementation is applicable for computing accurate energies of large organic and bio-organic molecules at their equilibrium geometries. Since the definition of the different hierarchies is based on the connectivity of the atoms in the molecule, it is not applicable for weak nonbonded interactions such as in water clusters or benzene clusters. In addition, the mismatch between CCSD(T) and MP2 for aromatic systems, as demonstrated in our previous work,[37] makes the current implementation not accurate for aromatic molecules. Various DFT methods are being explored as a low level of theory used in conjunction with CCSD(T) to get accurate ab initio energies of aromatic systems, for molecules away from their equilibrium geometries, and for applications on

even larger molecules. Overall, our method can be very useful to assess the thermodynamic feasibility of organic transformations with the CCSD(T) method at a considerably mitigated computational cost. Some potential application areas are the study of cracking products involved in combustion chemistry (including radical systems), reactions involved in biomolecules such as peptide systems, and a myriad of reactions involving chemical transformations in organic chemistry (structural rearrangements, redox reactions, condensation reactions, addition reactions, carbohydrate chemistry, etc.).

On the whole, our encouraging results bode well for the accurate computation of CCSD(T) energies for substantially larger molecules. Perhaps for very large molecules containing hundreds of atoms, it may be useful to come up with a method that partially uses fragments in their equilibrium geometries and partly incorporates the geometries found in the parent molecule, to accurately and cost-effectively perform otherwise prohibitive high-level *ab initio* calculations. Ultimately it is anticipated that novel chemical applications and a new outlook in the development of fragment-based methods will emerge from using CBH as a fragment-based method.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kraghava@indiana.edu.

Notes

The authors declare no competing financial interest.

Biographies

Raghunath O. Ramabhadran obtained his Ph.D. in Chemistry, working under the guidance of Professor Krishnan Raghavachari at Indiana University, Bloomington. He is currently a postdoctoral fellow in the research group of Professor K. N. Houk, at the University of California, Los Angeles.

Professor Krishnan Raghavachari is a Distinguished Professor of Chemistry at Indiana University, Bloomington. He received his Ph.D. from Carnegie-Mellon University from the research group of Professor John A. Pople in 1981. Previously he was a Distinguished Research Scientist at Bell Laboratories in Murray Hill, NJ. He has published extensively on electron correlation methods, surface chemistry, theoretical thermochemistry, cluster science, and electronic structure methods for large molecules. He is a Fellow of the International Academy of Quantum Molecular Science, Royal Society of Chemistry, and the American Physical Society.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from NSF Grant CHE-1266154 at Indiana University. We thank Ben Gamoke, Arka Sengupta, and Dr. Nicholas Mayhall for the Conspectus figure.

REFERENCES

(1) Friesner, R. A. *Ab Initio* Quantum Chemistry: Methodology and Applications. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6648−6653.

(2) Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E. In *Theory and Applications of Computational Chemistry The First Forty years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds; Elsevier: Amsterdam, The Netherlands, 2005; pp 1−7.

(3) Raghavachari, K.; Anderson, J. B. Electron Correlation Effects in Molecules. *J. Phys. Chem.* **1996**, *100*, 12973.

(4) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.

(5) Raghavachari, K., Curtiss, L. A. In *Quantum-Mechanical Prediction of Thermochemical Data*; Cioslowski, J., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001; pp 67−98.

(6) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. G*n* Theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 810−825.

(7) Harding, M. E.; Vázquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. High-Accuracy Extrapolated *ab initio* Thermochemistry. III. Additional Improvements and Overview. *J. Chem. Phys.* **2008**, *128*, No. 114111.

(8) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/mol Predictions. *J. Chem. Phys.* **2006**, *128*, No. 144108.

(9) Feller, D.; Peterson, K. A.; Dixon, D. A. A Survey of Factors Contributing to Accurate Theoretical Predictions of Atomization Energies and Molecular Structures. *J. Chem. Phys.* **2008**, *129*, No. 204105.

(10) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A Complete Basis Set Model Chemistry. VI. Use of Density Functional Geometries and Frequencies. *J. Chem. Phys.* **1999**, *110*, 2822−2827.

(11) Lynch, B. J.; Truhlar, D. G. What Are the Best Affordable Multi-Coefficient Strategies for Calculating Transition State Geometries and Barrier Heights? *J. Phys. Chem. A* **2002**, *106*, 842−846.

(12) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. The Correlation Consistent Composite Approach (ccCA): An Alternative to the Gaussian-n Methods. *J. Chem. Phys.* **2006**, *125*, No. 104111.

(13) (a) Pruitt, S. R.; Bertoni, C.; Brorson, K. R.; Gordon, M. S. Efficient and Accurate Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2786−2794. (b) Federov, D. G.; Asada, N.; Nakanishi, I.; Kitaura, K. The Use of Many-Body Expansions, and Geometry Optimizations in Fragment-Based Methods. *Acc. Chem. Res.* **2014**, *47*, 2846−2856.

(14) (a) Sahu, N.; Gadre, S. R. Molecular Tailoring Approach: A Route for ab Initio Treatment of Large Clusters. *Acc. Chem. Res.* **2014**, *47*, 2739−2747. (b) Furtado, J. P.; Rahalkar, A. P.; Shanker, S.; Bandyopadhyay, P.; Gadre, S. R. Facilitating Minima Search for Large Water Clusters at the MP2 Level via Molecular Tailoring. *J. Phys. Chem. Lett.* **2012**, *3*, 2253−2258.

(15) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. Fragment Quantum Mechanical Calculation of Proteins and Its Applications. *Acc. Chem. Res.* **2014**, *47*, 2748−2757.

(16) Li, Z.; Li, H.; Suo, B.; Liu, W. Localization of Molecular Orbitals: From Fragments to Molecule. *Acc. Chem. Res.* **2014**, *47*, 2758−2767.

(17) Li, S.; Li, W.; Ma, J. Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates. *Acc. Chem. Res.* **2014**, *47*, 2712−2720.

(18) Wang, B.; Yang, K. R.; Xu, X.; Isegawa, M.; Leverentz, H. R.; Truhlar, D. J. Quantum Mechanical Fragment Methods Based on Partitioning Atoms or Partitioning Coordinates. *Acc. Chem. Res.* **2014**, *47*, 2731−2738.

(19) Mezey, P. Fuzzy Electron Density Fragments in Macromolecular Quantum Chemistry, Combinatorial Quantum Chemistry, Functional Group Analysis, and Shape−Activity Relations. *Acc. Chem. Res.* **2014**, *47*, 2821−2827.

(20) Huang, L.; Massa, L.; Karle, J. The Kernel Energy Method: Application to a tRNA. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1233−1237.

(21) Nanda, K.; Beran, G. J. O. Prediction of Organic Molecular Crystal Geometries from MP2-Level Fragment Quantum Mechanical/Molecular Mechanical Calculations. *J. Chem. Phys.* **2012**, *137*, No. 174106.

(22) Rezac, J.; Salahub, D. R. Multilevel Fragment-Based Approach (MFBA): A Novel Hybrid Computational Method for the Study of Large Molecules. *J. Chem. Theory Comput.* **2010**, *6*, 91−99.

(23) Gao, J.; Truhlar, D. J.; Wang, Y.; Mazack, M. J. M.; Loffler, P.; Provorse, M. R.; Rehak, P. Explicit Polarization: A Quantum

Mechanical Framework for Developing Next Generation Force Fields. *Acc. Chem. Res.* **2014**, *47*, 2837−2845.
(24) Richard, R. M.; Lao, K. U.; Herbert, J. M. Aiming for Benchmark Accuracy with the Many-Body Expansion. *Acc. Chem. Res.* **2014**, *47*, 2828−2836.
(25) Bates, D. M.; Smith, J. R.; Tshumper, G. S. Efficient and Accurate Methods for the Geometry Optimization of Water Clusters: Application of Analytic Gradients for the 2-Body:Many-Body QM:QM Fragmentation Method to $(H_2O)_n$. *J. Chem. Theory Comput.* **2011**, *7*, 2753−2760.
(26) (a) Deev, V.; Collins, M. A. Approximate *ab Initio* Energies by Systematic Molecular Fragmentation. *J. Chem. Phys.* **2005**, *122*, No. 154102. (b) Bettens, R. P. A; Lee, A. M. A New Algorithm for Molecular Fragmentation in Quantum Chemical Calculations. *J. Phys. Chem. A* **2006**, *110*, 8777−8785.
(27) (a) Collins, M. A. Systematic Fragmentation of Large Molecules by Annihilation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7744−7751. (b) Le, H.-A.; Tan, H.-A.; Ouyang, J. F.; Bettens, R. P. A. Combined Fragmentation Method: A Simple Method for Fragmentation of Large Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 469−478.
(28) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials. *J. Chem. Theory Comput.* **2011**, *7*, 1336−1343.
(29) Mayhall, N. J.; Raghavachari, K. Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2669−2675.
(30) Saha, A.; Raghavachari, K. Dimers of Dimers (DOD): A New Fragment-Based Method Applied to Large Water Clusters. *J. Chem. Theory Comput.* **2014**, *10*, 58−67.
(31) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular Orbital Theory of the Electronic Structure of Organic Compounds. V. Molecular Theory of Bond Separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796−4801.
(32) Bakowies, D. *Ab initio* Thermochemistry Using Optimal-Balance Models with Isodesmic Corrections: The ATOMIC Protocol. *J. Chem. Phys.* **2009**, *130*, No. 144113.
(33) Fishtik, I. Unique Stoichiometric Representation for Computational Thermochemistry. *J. Phys. Chem. A* **2012**, *116*, 1854−1863.
(34) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. An Alternative Approach to the Problem of Assessing Stabilization Energies in Cyclic Conjugated Hydrocarbons. *Theor. Chem. Acc.* **1975**, *38*, 121−129.
(35) Wheeler, S. E.; Houk, K. N.; Schlyer, P. v. R.; Allen, W. D. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *J. Am. Chem. Soc.* **2009**, *131*, 2547−2560.
(36) Ramabhadran, R. O.; Raghavachari, K. Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy. *J. Chem. Theory Comput.* **2011**, *7*, 2094−2103.
(37) Ramabhadran, R. O.; Raghavachari, K. Connectivity-Based Hierarchy for Theoretical Thermochemistry: Assessment Using Wave Function-Based Methods. *J. Phys. Chem. A* **2012**, *116*, 7531−7537.
(38) Ramabhadran, R. O.; Sengupta, A.; Raghavachari, K. Application of the Generalized Connectivity-Based Hierarchy to Biomonomers: Enthalpies of Formation of Cysteine and Methionine. *J. Phys. Chem. A* **2013**, *117*, 4973−49801.
(39) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. Accurate and Computationally Efficient Prediction of Thermochemical Properties of Biomolecules Using the Generalized Connectivity-Based Hierarchy. *J. Phys. Chem. B* **2014**, *118*, 9631−9643.
(40) Sengupta, A.; Raghavachari, K. Prediction of Accurate Thermochemistry of Medium and Large Sized Radicals Using Connectivity-Based Hierarchy (CBH). *J. Chem. Theory Comput.* **2014**, *10*, 4342−4350.
(41) Ramabhadran, R. O.; Raghavachari, K. Extrapolation to the Gold-Standard in Quantum Chemistry: Computationally Efficient and Accurate CCSD(T) Energies for Large Molecules Using an Automated Thermochemical Hierarchy. *J. Chem. Theory Comput.* **2013**, *9*, 3986−3994.
(42) Similar ideas using a lower level of theory to improve the results have been used previously. A generalized ONIOM-type approach for fragment-based methods has been introduced by Mayhall and Raghavachari in ref 28. Gadre and co-workers have used very similar ideas in ref 14.