CHEMICAL REVIEWS

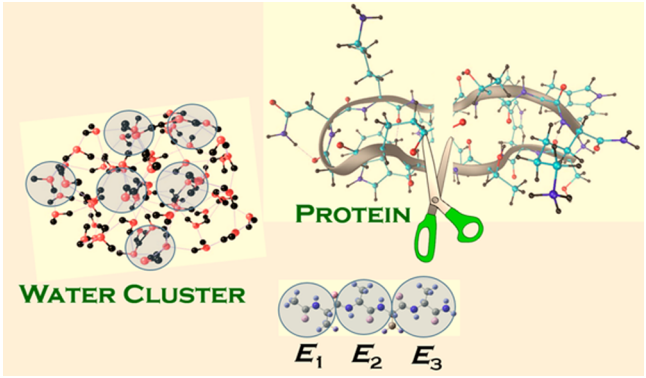
Review

pubs.acs.org/CR

Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules

Krishnan Raghavachari* and Arjun Saha

Department of Chemistry, Indiana University, 800 E. Kirkwood Avenue, Bloomington, Indiana 47405, United States



CONTENTS

1. INTRODUCTION

The last three decades have seen dramatic progress in the development and application of *ab initio* quantum chemical methods. The binding energies and thermodynamic properties of small molecules can now be evaluated to well within chemical accuracy (±1 kcal/mol), frequently rivalling or exceeding the accuracy in many experiments.[1] The ever-increasing power of modern computers has clearly played a role in such applications. However, a much greater contribution has come from a range of new sophisticated computational methods and associated algorithms and program packages that are widely available for use by the larger scientific community. In particular, new developments of highly accurate electron correlation methods (*N*-particle problem) and understanding their convergence behavior with respect to the basis set size (1-particle problem) have both contributed to these developments.

Coupled cluster theory has emerged as the definitive theoretical method for the accurate computation of electron correlation energies of molecules.[2] Nevertheless, the practical realization of chemical accuracy for more than very small molecules is still a formidable computational task. This is a direct result of the steep scaling of the cost of accurate calculations with molecular size. For example, while CCSDTQ (coupled cluster with single, double, triple, and quadruple excitations) is expected to yield results with high accuracy, its evaluation scales as N^{10}, and is impractical except for very small systems or with very small basis sets. The simpler CCSDT calculations still scale as N^8 and are also impractical for most systems. In addition, methods such as CCSDT or CCSDTQ involve iterative processes carried to convergence that make them even more expensive. The era of quantitatively accurate modern *ab initio* quantum chemistry emerged with the development of the perturbative CCSD(T) method by Raghavachari, Trucks, Pople, and Head-Gordon in 1989.[3] While the scaling of CCSD(T) is still steep (N^7), the one-step (noniterative) evaluation of the effects of triple excitations provides the best compromise between accuracy and computa-

Special Issue: Calculations on Large Systems

Received: October 24, 2014
Published: April 7, 2015

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

tional applicability, and the method has come to be known as the "gold standard of quantum chemistry". A quarter century later, it still stands as the method of choice and the focus of most accurate *ab initio* studies.

It is now well-recognized that accurate methods such as CCSD(T) have to be evaluated with large basis sets containing high angular momentum functions to attain convergence in the calculated energies.[4] The most commonly used systematic basis sets in this approach belong to the correlation-consistent family developed by Dunning and co-workers.[4] Ideally, explicit calculations on progressively larger basis sets have to be carried out, and the results have to be extrapolated to the basis set limit. Important effects such as zero-point vibrational corrections, scalar relativistic effects, etc. can then be included to yield accurate results for small molecules. However, a direct evaluation of all the important effects with large basis sets is frequently impractical, and a composite approach involving a series of calculations is almost invariably carried out to incorporate all the key interactions.

The composite quantum chemical approach involves a series of individual calculations that are assembled to extrapolate to the quantity of interest. This can be done at many different levels. The most common composite treatment involves the CCSD(T) extrapolation to the CBS limit. Since CCSD(T) is still expensive (N^7), it is quite common to carry out CCSD(T) calculations with a moderate basis set along with cheaper MP2 calculations (N^5) with a larger basis set, and extrapolate to the CCSD(T) energy with the larger basis set.[5] More generally, a series of Method/Basis combinations using appropriate well-defined protocols can be used to achieve a target accuracy of the expensive method at a reduced computational cost. The key assumption is the additivity approximation that different key energetic quantities can be individually computed, and then summed together to predict their cumulative effects. One of the earliest methods that proposed and used such a composite approach effectively to yield accurate thermochemical information for molecules was the composite Gaussian-1 theory developed by Pople et al. in 1989.[6]

The composite approach can also be used in other dimensions. For example, the different electrons in a molecule are frequently treated by using composite techniques. Almost universally, the contributions of valence electrons and core electrons are computed individually and combined to yield an extrapolated all-electron result. The CASSCF or RASSCF methods[7,8] can treat a subset of valence electrons using a more complete method while the rest of the valence electrons can be treated more approximately to yield results for complex systems where electron correlation effects are expected to be difficult to evaluate fully.

A third key dimension in which a composite approach is used is the fragment-based approach where a large molecule is divided into smaller fragments, and the computed results on the smaller fragments are assembled to extrapolate to the energy of the large molecule that may be difficult to evaluate directly.[9] Many levels of fragmentation may be possible, and the interaction between fragments is frequently restricted to the local vicinity of a given fragment. The computational savings result from the fact that the size of the fragments can be made relatively independent of the size of the parent molecule, while the number of fragments grows slowly (ideally in a linear fashion) with the size of the system. The overall computational cost will then asymptotically approach linear scaling. The ultimate goal is to maintain high accuracy while keeping the cost low. A range of such fragment-based methods will be described in this review.

Using a composite approach mentioned above, a broad range of theoretical methods have been developed to evaluate accurate thermodynamic properties of molecules. While the target of this review is large molecules, we will briefly consider methods that target accurate calculations on small and medium-sized molecules. Some of the key factors that determine the accuracy that can be obtained from theoretical calculations are considered carefully in these methods. For relatively small molecules (*viz.*, up to the size of benzene), accurate composite methods have been developed that can yield highly accurate results without the use of any empirical parameters. Such methods can evaluate the heats of formation of small molecules substantially better than chemical accuracy (approaching kJ/mol). These *ab initio* methods are briefly described in section 2.

An alternative successful strategy that makes it possible to approach chemical accuracy for significantly larger molecules involves a composite multilevel approach using less expensive perturbation theory to perform the larger basis set calculations (often in conjunction with DFT). Some of these methods are used with a small number of empirical parameters to avoid the direct calculations on some of the more difficult components. The most well-known such methods are the Gaussian-*n* theories.[10] Such methods still have chemical accuracy, but can be carried out for larger molecules. Molecules significantly larger than benzene, up to about 20 non-hydrogen atoms, can be treated by such composite techniques applicable for medium-sized molecules. They are briefly described in section 3.

Error cancellation is an important strategy in quantum chemistry. It can be used to achieve accurate results on larger molecules with a modest computational effort. Such strategies are based on using chemical concepts and clever ideas to attempt to cancel the effects of systematic errors in the calculations. For example, it is intuitive that chemical substituent effects can be evaluated using modest levels of theory due to the error cancellation between the parent molecule and its substituted variation. Similarly, isodesmic methods[11] and their variations have been used in quantum chemistry to achieve error cancellation for over 4 decades. Such methods are briefly described in section 4.

The bulk of this review will focus on fragment-based methods in quantum chemistry.[9] Such methods attempt to extend the success already realized for small molecules to the realm of large molecules. They represent one of the most important research areas in quantum chemical method developments. However, unlike the other methods, their focus is somewhat different. Unlike the small molecules where the assessment of performance is carried out by comparison with experiment, the assessment for many fragment-based methods is frequently done by comparison with the energy of a reference calculation. The success of the method is measured by its ability to reproduce the total energy of the large molecule computed directly by the energy obtained using a fragment-based approach. This enables the use of modest levels of theory to evaluate the performance of a given fragment-based method, somewhat akin to using a full CI calculation with a modest basis set to evaluate the performance of electron correlation methods in challenging situations.

There are a broad range of fragment-based methods being pursued by many different research groups. We focus mostly on "energy-based" methods that use different fragmentation

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

strategies to yield accurate predictions using standard quantum chemical methods. Many existing methods are summarized briefly in section 5. While many major contributions in fragment-based quantum chemistry have been considered, we have not analyzed some related approaches such as divide-and-conquer techniques[12] or X-Pol (explicit polarization potential) methods[13] since they have already been considered in a recent review.[9] The goal is not to give all the technical details of the different methods but to focus on the broad concepts of the methods along with the scope of the range of applications possible with the different schemes. An attempt to compare the different methods and to classify them is described in section 6.

A completely different strategy to reduce the computational scaling of CCSD(T) or other correlated methods and to enable accurate calculations on large molecules is the use of local orbital-based methods. The central principle of such methods is that the use of localized molecular orbitals makes the associated pair correlation energies to decay rapidly with the interorbital distance. Starting from the original formalism of Pulay and co-workers,[14,15] later extended by Werner and co-workers,[16,17] Neese and co-workers,[18,19] and others,[20,21] a range of techniques have been developed using orthogonal localized orbitals to represent the occupied space, and projected atomic orbitals or pair natural orbitals to represent the virtual space, to perform accurate calculations on large systems. Some of these methods can also include the interactions between significant pairs using more accurate methods (e.g., using coupled cluster theory), while evaluating other less dominant pair interactions via more approximate techniques (e.g., perturbation theory).[22,23] Other extensions to multiconfiguration and explicitly correlated F12-methods have also been developed.[24,25] While this is an important developing area for performing accurate calculations on large systems, this review will not include local orbital-based methods that have been treated in other reviews.[26,27]

2. ACCURATE MODELS FOR SMALL MOLECULES

The most accurate families of methods strive to predict thermochemical properties of small molecules not just to chemical accuracy (1 kcal/mol) but to significantly exceed chemical accuracy (1 kJ/mol). In such high accuracy *ab initio* methods, the total atomization energy of the molecule of interest is first determined from theoretical calculations, and the experimentally known heats of formation of gas phase atoms are then used to calculate the heat of formation of the molecule. The best nonempirical methods attempt to use the most accurate electron correlation techniques extrapolated to the complete basis set limit. As mentioned earlier, this is typically done using a composite approach where each individual effect is computed using different but appropriate levels of theory and assuming that their contributions are additive. The methods that fit in this category are the W*n* methods by Martin and co-workers;[28−31] the HEAT protocol by Szaley, Gauss, Valeev, Stanton, and co-workers;[32−34] the systematic coupled cluster extrapolation approach by Feller, Peterson, and Dixon;[1,35,36] and the explicitly correlated F12 methods of Klopper, Werner, and co-workers.[37,38]

Dixon, Feller, and Peterson have described the protocol used in most such methods.[36] The starting point is the evaluation of the valence correlation energy of the molecule calculated at the CCSD(T) level extrapolated to the complete basis set limit. A set of additional factors are then calculated and included as additive corrections. Every term that is of kJ/mol magnitude will have to be calculated explicitly without any empirical parameters. The different effects that have to be included are core−valence correlation energy, scalar relativistic corrections, spin−orbit corrections, higher order post-CCSD(T) corrections, zero point energy, and thermal corrections, and corrections for non-Born−Oppenheimer effects.[36] Most such high accuracy methods determine the total atomization energy (TAE) as a sum of all such terms evaluated individually as follows:

$$\mathrm{TAE} = \Delta E_{\mathrm{CCSD(T)}}(\mathrm{CBS}) + \Delta E_{\mathrm{HO}} + \Delta E_{\mathrm{CV}} + \Delta E_{\mathrm{SR}} + \Delta E_{\mathrm{SO}} + \Delta E_{\mathrm{BO}} + \Delta E_{\mathrm{ZPE}} \quad (1)$$

Here, HO = higher order post-CCSD(T) corrections, CV = core−valence correlation, SR = scalar relativistic effects, SO = spin−orbit corrections, BO = diagonal Born−Oppenheimer corrections, and ZPE = zero-point energy.

The foremost effect is still the computation of the valence correlation energy at the complete basis set limit. This is based on the systematic convergence properties of the CCSD(T) correlation energies obtained with the correlation-consistent cc-pV*n*Z or the augmented aug-cc-pV*n*Z basis sets.[39] However, the convergence is usually very slow, due to the poor description of the interelectronic cusp by simple one-particle basis sets. Typically, the results using three (or sometimes two) consecutive *n*Z results can be used with one of several available extrapolation formulas[40] to determine the extrapolated CCSD(T) energy at the CBS limit. More recently, it has been found that the rate of convergence can be accelerated substantially by using explicitly correlated methods that use an exponential correlation factor (e.g., methods that use F12 = $\exp(-\gamma r_{12})$ that offer substantial improvement over the previous generation R12 methods)[41,42] along with the cc-pV*n*Z basis sets.[39]

The core−valence contributions can also be computed at the CCSD(T) level using appropriate large basis sets including core functions. The higher order correlation effects beyond CCSD(T) (if included in the method) have to be computed with smaller basis sets due to their high computational cost. However, this is found to be adequate in many cases since they appear to reach faster convergence (at least for reaction energies) with smaller basis sets.[31,34,43] The first correction is at the CCSDT level (scaling as N^8) where the contributions of triple excitations are calculated with a fully iterative treatment as compared to a perturbative treatment in CCSD(T). The next level is CCSDTQ (scaling as N^{10}) that also includes the full treatment of quadruple excitations. The high cost of these methods as well as the (fortuitous) partial cancellations of their incremental contributions makes it important to choose basis sets appropriately in evaluating such terms.[36] In some molecules whose wave functions have significant multi-configurational character, such post-CCSD(T) methods do make significant contributions to the atomization energy. For example, such effects are as large as 3.2 kcal/mol for the atomization of ozone.[35] If needed, the CCSD, CCSDT, CCSDTQ energies can be used to extrapolate to even higher order correlation contributions[35] (usually found to be very small).

The contributions of zero-point energies and thermal contributions (if 298 K energies are computed) require accurate vibrational frequencies. In most popular methods, the harmonic vibrational frequencies are computed at some appropriate theoretical level (e.g., CCSD), and scaled using an

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

appropriate scale factor to correct for known deficiencies (to correct for anharmonicity effects, for example), and then used in the computation of zero-point corrections. However, if empirical scaling is to be avoided, anharmonic frequencies can also be directly computed and then used for the computation of ZPEs. The remaining terms (scalar relativistic and spin orbit effects, diagonal Born−Oppenheimer corrections for very accurate calculations) can also be computed with appropriate computational models.[36] Such energies have been used by Dixon, Feller, and Peterson for a range of small molecules to obtain heats of formation with kJ/mol accuracy.[36]

As in the Feller, Peterson, Dixon approach, the W*n* methods attempt to evaluate atomization energies to much better than chemical accuracy (kJ/mol) by a careful evaluation of each individual component that contributes to binding energies with high accuracy. Each term is extrapolated to the CBS limit to enable the evaluation of heats of formation to sub-kJ/mol accuracy without the use of empirical parameters. The latest generation of the method, W4 theory, includes post-CCSD(T) corrections for T, Q, and 5 correlations, and achieves an impressive accuracy of about 0.5 kJ/mol for a collection of small molecules. Karton et al.[31] have assembled a test set of 140 total atomization energies of small molecules and radicals based on W4 energies that can be used to assess the performance of more approximate methods. While W4 is limited to small molecules, the earlier generation methods (W1 or W2) can be used in applications to somewhat larger molecules (e.g., amino acids, benzene). The HEAT protocol[34] uses similar strategies and achieves similar sub-kJ/mol accuracy, and has also been applied to systems as large as benzene.[33] Nevertheless, direct applications to large molecules with W*n* or HEAT protocols will be difficult.

3. COMPOSITE MODELS FOR MEDIUM-SIZED MOLECULES

The earliest set of general composite methods that were highly successful are the G*n* methods developed by Pople, Curtiss, and Raghavachari and co-workers. G4, proposed in 2005, is the fourth generation method[44] that follows G1 (1989),[6,45] G2 (1991),[46] and G3 (1998).[47] There are also approximate versions of some of these methods[48−50] that have been suggested for better applicability. As outlined by Pople et al.[6] in the first G1 paper, the objective was to develop a general predictive procedure, applicable to any molecular system in an unambiguous manner, that can reproduce known experimental data to a prescribed accuracy of ±2 kcal/mol (slightly less ambitious than "chemical accuracy"), and can be applied with similar accuracy to other species where the experimental data is unknown or uncertain. The composite approach using a suite of methods with different levels of accuracy in combination with practical basis sets to approach the exact result was also outlined in this important paper. The idea of successfully extrapolating with less expensive perturbation theory (MP2, MP4) or even HF was also demonstrated in this paper. Finally, the idea of using a small number of molecule-independent empirical parameters (i.e., depending only on the number of α and β electrons) to estimate the remaining deficiencies in the calculations was also suggested in this paper. This avoids the use of very large basis sets containing high angular momentum functions and such an approach using "higher level corrections" has led to the most recent methods to achieve an overall accuracy of better than one kcal/mol. Many of the ideas outlined in this paper are now used widely throughout quantum chemistry.

The basic ideas involved in such composite schemes can be illustrated by the energy expression used in G4 theory.[44] The total energy at 0 K ("G4 energy") is obtained by adding several individual energy corrections (at a geometry optimized using DFT methods) in an additive manner.

$$\begin{aligned} E_o[\mathrm{G4}] = {} & \mathrm{E(CCSD(T)/6\text{-}31G(d))} + E(\mathrm{plus}) \\ & + E(\mathrm{2df,\,p}) + E(\Delta\mathrm{G3LXP}) + E(\mathrm{HF/limit}) \\ & + E(\mathrm{SO}) + E(\mathrm{HLC}) + E(\mathrm{ZPE}) \end{aligned} \quad (2)$$

Here the different terms yield additive contributions: E(plus) from diffuse functions, E(2df,p) from the essential polarization functions, E(ΔG3LXP) from larger basis sets, E(HF/limit) from the CBS limit at the HF level, E(SO) from spin−orbit corrections, E(ZPE) from zero-point energy, and E(HLC) from an empirical higher-level correction. The key point is that the accurate CCSD(T) calculations are performed with a very modest basis set (Pople's 6-31G(d) basis set). Since this is usually the most time-consuming part of a G4 calculation (N^7 scaling), it makes the method more applicable for larger molecules (calculations on molecules containing around 20 heavy atoms can be performed). The effects of larger basis sets are obtained at lower levels of theory (MP4, MP2, or HF). For example, the effect of diffuse functions is given as

$$E(\mathrm{plus}) = \mathrm{MP4/6\text{-}31{+}G(d)} - \mathrm{MP4/6\text{-}31G(d)} \quad (3)$$

The effects of larger basis sets containing more polarization functions are obtained using the more cost-efficient MP2 method along with the custom-developed G3LXP basis set. This basis set was developed to have a balanced performance for different elements of the first and second rows of the periodic table. The only term that is extrapolated to the complete basis set (CBS) limit is at the HF level (i.e., HF/limit).[44] Spin−orbit corrections are used only for atoms and taken from the known experimental values.

A "higher level correction", E(HLC), is added to take into account remaining deficiencies in the energy calculations. While empirical, and fitted to reproduce experimental heats of formation, the parameters in G*n* methods are "molecule independent" in the sense that they do not depend on the constituent atoms in the molecule. This is in contrast to the standard semiempirical methods (such as PM6)[51] that have parameters depending on the type of atom and its neighbors. In G*n* theory, however, they only depend on the number of α and β electrons in the molecule. Two parameters were used in G2 theory, 4 in G3 theory, and 6 in G4 theory.

G4 theory was assessed on the full G3/05 test set[52] containing 454 energies including enthalpies of formation of neutral molecules, atomization energies, ionization potentials, electron affinities, proton affinities, and hydrogen bond energies. The mean absolute deviation from experiment at the G4 level was 0.83 kcal/mol, significantly improved over the value of 1.13 kcal/mol obtained for G3 theory. A comparison of the performance of the G*n* methods on the small G2/91 test set[46] as well as the large G3/05 test set is shown in Figure 1.

Petersson and co-workers have developed a related series of methods, referred to as complete basis set (CBS) methods,[53−57] for the evaluation of accurate energies of molecular systems. The central idea in the CBS methods is based on the asymptotic convergence properties of the pair correlation

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

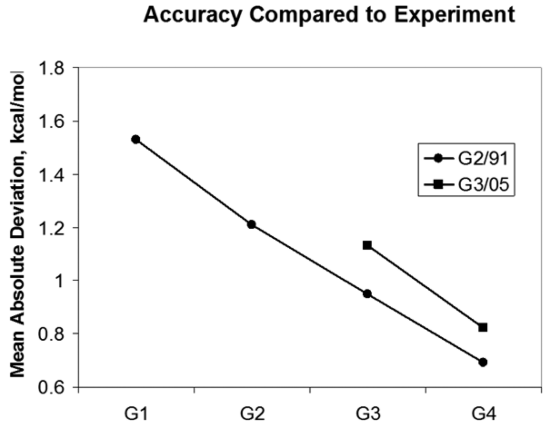


Figure 1. Accuracies of G1−G4 theories for the G2/91 and G3/05 test sets. Reprinted with permission from ref 10. Copyright 2011, John Wiley & Sons, Ltd.

energy, leading to an extrapolation procedure to determine the projected second-order (MP2) energy at the CBS limit.[53] While less empirical than the G*n* methods, a few parameters are nevertheless included in the CBS methods to achieve better performance, similar in spirit to the higher level correction used in the G*n* methods. There are many variations of the CBS methods that are widely used. The CBS-QCI/APNO method[54] is somewhat more expensive than G*n* methods but achieves more accurate results for small molecules. The most popular variants are the CBS-Q, CBS-q, and CBS-4 methods that are progressively applicable for larger molecules using cheaper computational methods and smaller basis sets combined with a small number of empirical parameters.[55] A variation of the CBS-Q method, using B3LYP geometries and ZPEs, termed CBS-QB3 is also widely used.[56,57] Spin-restricted versions of the CBS methods have also been developed to achieve better performance for radical systems.[58]

Wilson, Cundari, and co-workers have developed more recent, broadly applicable and popular, ccCA methods (correlation-consistent composite approach)[59−61] to achieve chemical accuracy without the use of empirical parameters. The basic formalism achieves its efficiency by extrapolating the MP2 energies to the CBS limit using cc-pVnZ or aug-cc-pVnZ (D,T,Q) basis sets using different extrapolation formulas. Higher order correlation corrections are added at CCSD(T) or QCISD(T) levels to attain high accuracy. Relativistic effects are included by Douglas−Kroll calculations at the MP2 level. The performance of ccCA for the G3/05 set of 454 energies is comparable to that of G4 theory, with an overall mean deviation of about 1 kcal/mol.

A range of additional developments on the ccCA method have been introduced by Wilson and co-workers. The method has been extended for heavy p-block elements[62] and s-block metals,[63,64] and achieves excellent performance. An additional variation to extend the method to 3d and 4d transition metals has been developed with a relaxed criterion for accuracy (±3 kcal/mol).[65−67] A variation including multiconfiguration effects (MR-ccCA) permits the exploration of potential energy surfaces for small molecules.[68] Additional efficiency improvements using the RI approximations at the MP2 level as well as computations including local correlation methods enable applications to somewhat larger molecules.[69]

The methods discussed thus far use an additivity approximation to calculating the individual terms in the composite approach. An alternative approach involves scaling of the calculated correlation energy terms using multiplicative parameters determined by fitting to experimental data. This work has been pioneered by Truhlar and co-workers who have derived a series of multicoefficient correlation methods using this approach.[70,71] On the basis of these ideas, multiplicative scaled versions of G3 theory (G3S) have been developed by Curtiss and co-workers.[72] The overall performance of additive and multiplicative versions of G3 theory was found to be similar.

4. ERROR-CANCELLATION STRATEGIES FOR LARGER MOLECULES

The methods considered thus far evaluate the heat of formation of the target molecule by a two-step process. First, the atomization energy of the molecule is directly evaluated using accurate theoretical methods such as coupled cluster theory. The atomization energy is then used in conjunction with the known heats of formation of gas phase atoms to determine the heat of formation for the molecule. Thus, the accuracy of the treatment depends on the accuracy of the calculated atomization energy. The problem with this approach is that the atomization energy is a quantity that is very difficult to calculate accurately from first-principles since every bond in the molecule is broken into the constituent atoms that often contain many unpaired electrons. Highly sophisticated (and expensive) electron correlation calculations using large basis sets are needed to describe the atomization process accurately. While it provides a rigorous test of the inherent accuracy of theoretical methods, this limits the applicability of this approach to small or medium-sized molecules, as described in the earlier sections.

An alternative strategy is to use error-cancellation techniques to derive accurate heats of formation for larger molecules. Since the field of quantum chemistry depends on the evaluation of differences between large quantities (i.e., total energy) to yield small quantities (i.e., reaction energy), error-cancellation strategies are very commonly used. For example, it has long been recognized that substituent effects can be evaluated accurately with modest theoretical methods because they involve energy differences between reactions that are inherently similar. For the calculation of accurate heats of formation, this can be achieved by a strategic choice of chemical reaction schemes that permit such error cancellation. A major advantage of such schemes is that modest computational methods are usually sufficient to achieve high accuracy, enabling calculations on much larger molecules than with the atomization approach. In fact, such an approach is very closely related to the fragment-based methods considered in the next section.

The idea of error-cancellation was recognized to be critical in the early days of *ab initio* quantum chemistry when sophisticated theoretical methods were not available (or too expensive with the available computational resources) to compute the bond energies and thermodynamic properties of molecules. The first key idea was the *isodesmic bond separation* (IBS) scheme proposed by Pople and co-workers[73] in 1970 that enormously improved the accuracy of the theoretical predictions using simple models such as Hartree−Fock theory with moderate basis sets. 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, electronic structure computations on all the reference molecules in the scheme are performed to extract the reaction energy. The available heats of formation for all the reference

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

species (if they are known with sufficient accuracy) can then be used to derive the heat of formation of the parent molecule. Ramarkably, using the IBS scheme with a simple HF/4-31G level of theory, Pople and co-workers were able to achieve a mean absolute deviation from experiment of only 3.5 kcal/mol for the heats of formation for 15 small organic molecules.[73] The IBS scheme provided an excellent demonstration of the significance of appropriately balancing chemical reactions resulting in substantial error-cancellation and yielding improved accuracy.

The use of the IBS scheme can also improve the results for more accurate methods. For example, Raghavachari and co-workers[74] showed that, for a set of 40 closed-shell molecules containing C, H, O, and N, the use of an IBS scheme, in conjunction with composite G2 and G2(MP2) methods, reduced the mean absolute deviation in the calculated heats of formation from 1.49 to 0.54 kcal/mol for G2, and from 1.99 to 0.64 kcal/mol for G2(MP2). This suggests that even sophisticated models can benefit from additional error-cancellation using such ideas to achieve higher accuracy. In a later paper,[75] they also demonstrated substantial performance improvement with DFT methods. However, these methods required information on the experimental heats of formation of the reference molecules.

Bakowies[76] has developed a sophisticated protocol, termed ATOMIC (ab initio thermochemistry using optimal-balance models with isodesmic corrections), that uses isodesmic corrections in conjunction with composite models to derive accurate heats of formation for closed shell organic molecules containing C, H, O, N, and F. Key ingredients of the method include the avoidance of empirical parameters (as in G*n* methods) or the need for experimental information on reference molecules (as in the standard isodesmic schemes). Instead, Bakowies uses the IBS scheme to obtain a set of high level bond increments that are subsequently used to correct the computed atomization energies. The IBS scheme is found to reduce the basis set requirements dramatically for each of the components of the composite model (including sophisticated terms such as complete basis set extrapolations using a sequence of cc-pV*n*Z basis sets, post-CCSD(T) corrections, scalar relativistic effects, diagonal Born−Oppenheimer corrections, etc.). Three different models, termed A, B, and C, that depend on the basis sets used for CBS extrapolations, were proposed that are progressively more applicable for larger molecules.[76] Even the least expensive model C achieved chemical accuracy in the calculated atomization energies (mean absolute deviation = 0.99 kcal/mol) for a set of 173 neutral molecules containing H, C, N, O, and F.[77] The performance was similar to that obtained with the composite G3 method. The more sophisticated models A and B performed even better. The largest calculations in this study contained about 10 non-hydrogen atoms. The ATOMIC protocol was later extended to DFT methods to achieve reasonable results, but outside chemical accuracy.[78] Similar concepts are implicit in other methods such as Fishtik's group additivity schemes.[79]

While the IBS scheme provided significant improvement, it was recognized even in the mid-1970s that more sophisticated error-cancellation techniques may be needed for larger molecules. Furthering the ideas developed by Pople, George, and co-workers[80] in 1975 proposed the hybridization-based *homodesmotic scheme* of reactions applicable for closed shell hydrocarbons. The homodesmotic scheme was designed to offer a superior balance of the bond types and the hybridization of the different carbons and hydrogens in the molecule. Following this, a variety of other similar schemes[81] (e.g., hyperhomodesmotic)[82] that provide higher levels of error cancellation have since been suggested to predict the energies of hydrocarbons.

More recently, Wheeler, Houk, Schleyer, and Allen[83] developed a general and a systematic hybridization-based hierarchy of homodesmotic reactions for closed shell hydrocarbons. By utilizing predefined reactants and products at each level of their hierarchy, they achieve an increased balance in the hybridization and the covalent bonding environment of the carbon atoms within the family of hydrocarbon molecules. In particular, they have developed different classes of homodesmotic reactions, labeled HD1−HD5 (classified according to the increasing levels of bond-type and hybridization matching), and have generalized the concepts as *n*-homodesmotic reactions applicable for all hydrocarbons. They have also pointed out the need for a more general and systematic hierarchy that spans beyond hydrocarbons and is applicable to any organic molecule containing any heteroatom (for instance, O, N, S, etc.). More recently, Wheeler has extended similar ideas to make the methods also applicable to hydrocarbon radicals.[84]

An alternate approach to the construction of the generalized hierarchy, i.e., one based on merely the connectivity of the atoms in an organic molecule instead of utilizing predefined reactants and products in a hybridization-based hierarchy, has been developed by Ramabhadran and Raghavachari.[85] This is an automated hierarchy, termed CBH (connectivity-based hierarchy), applicable to all classes of closed shell organic molecules and does not require manual effort to derive coefficients in the chemical equations to balance the bond types and hybridizations of the different species.

The different levels of CBH can be envisioned as being the rungs of a ladder, such that ascending up the rungs of the hierarchy increasingly preserves the chemical environment (i.e., a better matching of the bond-types and hybridization-types are achieved automatically) of a molecule. They are referred to as CBH-0, CBH-1, CBH-2, CBH-3, etc. The rungs alternate between being atom-centric (CBH-0, CBH-2, etc.), and bond-centric (CBH-1, CBH-3, etc.). For example, the bond-centric CBH-1 is generated by extracting and hydrogen-terminating all the heavy-atom bonds. It turns out that CBH-1 is exactly the same as Pople's IBS scheme. CBH-2 is constructed by extracting all the heavy atoms maintaining their atom connectivities with neighboring heavy atoms, and then hydrogen terminating at the neighboring atoms. Since this rung preserves the immediate chemical environment of an atom, it has been termed as the *isoatomic* reaction *scheme*. 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 atom connectivities with neighboring heavy-atoms, and then hydrogen terminating at the neighboring atoms. Higher rungs such as CBH-4 and beyond can also be defined, but for commonly encountered organic and biomolecules containing about 20 heavy-atoms, CBH-3 was found to be sufficient.

It can readily be seen that, at every rung, certain molecules are overcounted in the construction of CBH. For example, at CBH-1, the simplest valence satisfied hydrides of heavy atoms (ammonia for N, methane for C, etc.) are overcounted. In order to take this aspect of overcounting into account, the overcounted molecules are added to the reactant side of the

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

chemical equations. This was achieved in the CBH hierarchy by noticing a recursive relationship between the products at one rung, and the reactants at the next rung. By taking into consideration some minor differences in the treatment for terminal moieties and branched structures, a general automated procedure was derived that is applicable for all closed shell organic and biomolecules for which a Lewis-type valence structure can be drawn.

The CBH reaction schemes for cyclohexanone are shown below.

CBH-0: $C_6H_{10}O + 8H_2 \rightarrow 6CH_4 + H_2O$

CBH-1:

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

CBH-2:

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

CBH-3: $C_6H_{10}O + 5C_3H_8 + C_3H_6O \rightarrow 4C_4H_{10}$ (n-butane) + $2C_4H_8O$ (2-butanone)

The performance of CBH was initially evaluated for a test set of 30 organic molecules containing the following: different functional groups (alcohol, aldehyde, olefenic, ketone, nitrile, amine, ester), different heteroatoms (N, O, Cl, Br, S), different molecular architectures (linear vs branched, cyclic vs acyclic), 5 molecules possessing ring-strain, three aromatic molecules, and two biomolecules (cysteine and methionine). After the CBH schemes were generated, the performance of a range of DFT methods[85] (BPW91, B3LYP, BMK, M05-2X, M06-2X, TPSSh, and B2PLYP) as well as wave function-based methods[86] (HF, MP2, and CCSD(T)) were assessed using several modest-sized double-ζ and triple-ζ basis sets. The reaction energies at different rungs were first calculated at the levels of theory above, and used with experimental heats of formations for all the reference species in the CBH equations, to derive the enthalpies of formation.

In the case of the 20 simpler molecules in the test set, the performance of CBH was excellent with all levels of theory with all the basis sets considered. Chemical accuracy was achieved at the hierarchies CBH-2 and above. For molecules with ring strain, chemical accuracy was reached with MP2 and CCSD(T), at CBH-2 and CBH-3, while larger errors were encountered with density functionals. For aromatic molecules, only CCSD(T) achieved very good results at the CBH-2 level while deficiencies were found for MP2 as well as DFT methods.[86] The performances for cysteine and methionine were evaluated including the effect of low-lying conformations, and the performance of CBH[87] was in very good agreement with other accurate theoretical studies.[88] These studies were later extended to all 20 amino acids to derive their heats of formation.[89] Other extensions to open shell systems have also been implemented and achieve chemical accuracy using the CBH-2 (isoatomic) scheme.[90]

It turns out that the CBH reaction schemes are closely related to some of the fragment-based methods to be discussed in the next section.[91] Similar ideas in a different context have also been developed by Deev and Collins[92] and by Lee and Bettens.[93]

5. FRAGMENT-BASED METHODS FOR LARGE MOLECULES

In order to dodge the steep computational scaling of *ab initio* quantum chemical calculations, the methods discussed until now use a composite approach with multiple theory/basis combinations to enable accurate calculations on medium-sized molecules. However, all the calculations are performed on the full molecule with lower levels of theory being used with larger basis sets to effectively extrapolate to the results using accurate methods with large basis sets. However, as the molecule gets larger, it may be computationally expensive to treat the entire molecule with large basis sets even at modest levels of theory. An additional strategy that also uses a composite approach on much larger molecules is by means of *fragment-based methods* that aim to break a given molecular system up into computationally efficient units, and piece them back together to approximate the unfragmented energy. This is an extremely active field with a large number of participating groups and a bewildering array of methods and acronyms. While the general idea in all these methods is the same, the details of how the fragments are generated and assembled differ substantially between the different methods. We provide a brief review of a range of methods that focus on deriving accurate energies of large molecules containing both bonded (e.g., peptide) and nonbonded (e.g., water clusters) components.

In the following sections, we first discuss a set of fragment-based methods along with some applications. Later in section 6, we discuss the relationship between the different methods, and classify them in different groups.

5.1. Molecular Tailoring Approach (MTA)

MTA,[94−107] first introduced by Gadre in 1994, is one of the oldest fragment-based methods developed for the *ab initio* study of large clusters. Since the initial proposal, this method has gone through a substantial amount of modifications and improvements over the last 20 years. The formulation of MTA is described below.

In MTA, a large molecular system is divided into smaller *overlapping fragments*. The fragments are selected using an automated procedure that is determined by the desired maximum fragment size (defined in terms of the number of atoms per fragment). The procedure involves creating fragments by centering a sphere on each atom and recursively merging to achieve the desired fragment size. After executing independent electronic structure calculations on all the fragments, the desired molecular property can be predicted by accumulating those fragment results via the *inclusion-exclusion* principle. For a general property P, it can be expressed as

$$P = \sum P^{F_i} - \sum P^{F_i \cap F_j} + \cdots + (-1)^{k-1} \sum P^{F_i \cap F_j \cdots F_k} \cdots \quad (5)$$

Here, P^{F_i} denotes for the molecular property of ith fragment, $P^{F_i \cap F_j}$ stands for the molecular property of binary overlap of ith and jth fragments, and so on.

To achieve the best applicability, the authors have introduced an important parameter that helps to tune the balance between cost and accuracy of MTA. This parameter is termed as *R*-goodness (R_g). R_g is defined as the "minimum distance of an atom A present in a fragment from the atoms not belonging to that fragment". If an atom A occurs in more than one fragment, the maximum of the R_g values is taken for A. The minimum of all the atomic R_g values in the given molecule is a measure of

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

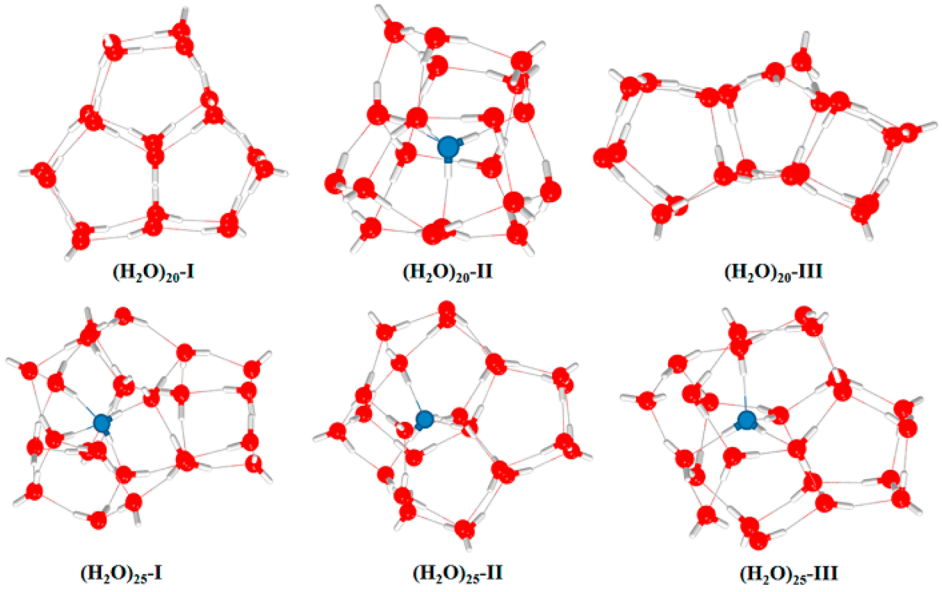


Figure 2. Water clusters studied with MTA method (with grafting correction) by Gadre et al.[105] Reprinted with permission from ref 105. Copyright 2012 American Chemical Society.

the quality of the fragmentation scheme. Typically R_g values in the range of 3−4 Å appear to yield accurate results.

The inclusion of long-range interactions is important in any fragment-based method. As shown by the ONIOM[108−112] and more recently in MIM[113] approaches, this can be achieved effectively by employing a second level of theory to correct for the inherent errors. MTA has adopted a similar idea, termed as a "grafting correction", and has recently been applied on different sizes of water clusters in conjunction with a second layer of theory within MTA.[104,105] Here we provide the sequential steps necessary for the grafting correction within MTA method using the MP2 method for a large molecule as an example. (1) Evaluation of the single point energy of the full system at the HF level using a higher basis set (HB) which is termed as "$E(\mathrm{HF})_{\mathrm{actual/HB}}$". (2) Estimation of the single-point energy through MTA, $E(2)_{\mathrm{MTA/HB}}$, the cardinality-based second-order perturbative correction. (3) At this step, the correction $\delta E(2)_{\mathrm{LB}}$ is determined at the lower basis (LB) by subtracting the MTA-based system energy from the actual system energy at the low level of theory:

$$\delta E(2)_{\mathrm{LB}} = E(2)_{\mathrm{actual/LB}} - E(2)_{\mathrm{MTA/LB}} \tag{6}$$

(4) Finally, MTA-based second order perturbative correction $E(2)^{\mathrm{corr}}_{\mathrm{MTA/HB}}$ and MTA-based energy $E(\mathrm{MP2})_{\mathrm{MTA/HB}}$ are obtained by the following two equations:

$$E(2)^{\mathrm{corr}}_{\mathrm{MTA/HB}} = E(2)_{\mathrm{MTA/HB}} + \delta E(2)_{\mathrm{LB}} \tag{7}$$

$$E(\mathrm{MP2})_{\mathrm{MTA/HB}} = E(\mathrm{HF})_{\mathrm{Actual/HB}} + E(2)^{\mathrm{Corr}}_{\mathrm{MTA/HB}} \tag{8}$$

Since its original proposal, MTA method has undergone several changes and has become more robust for practical applications on large molecules. The current version of MTA method has the capability of performing energy, gradient, and Hessian evaluations along with geometry optimizations and vibrational frequency calculations. The applicability of MTA is well-established in broad ranges of molecular systems including water clusters (small to large range), carbon dioxide clusters, hydrated sodium ion, benzene clusters, solvated ion pairs, N_2O clusters, and so on. A brief overview of some of these applications is discussed below.

Water clusters are simple but important molecules to investigate hydrogen bonding in chemical systems. Treatment of water clusters using fragment-based methods is always challenging due to several factors: (1) large number of hydrogen-bonded interactions, (2) highly compact three-dimensional structures, and (3) high symmetry. Thus, water clusters have attracted substantial attention in fragment-based quantum chemistry, and several research groups around the globe are using a variety of approaches on different sizes of water clusters to either probe or calibrate their methods.

In a recent study,[105] six low-lying structures of $(H_2O)_{20}$ and $(H_2O)_{25}$ clusters have been investigated with MTA method along with the grafting correction. Optimizations of these clusters were carried out with MTA using MP2/aug-cc-pVDZ model. All the previously reported minima were successfully reproduced for these water clusters with MTA (Figure 2).

It is important to notice here that a single level MTA can predict the energetics of these water clusters within an error of 2−20 mH relative to the actual *ab initio* result whereas implementing the grafting correction (i.e., second layer of theory) improves the accuracy remarkably to 0.4 mH for the same clusters. The grafting correction in this particular application used aug-cc-pVDZ and 6-31+G* as high and low level basis sets, respectively. The grafting correction within MTA has further been extended to predict CCSD(T) energies for these clusters. The theoretical model used for this purpose was CCSD(T)/aug-cc-pVDZ (high level method): MP2/aug-cc-pVDZ (low level method). The following extrapolation equation was used to predict the CCSD(T) energies:

$$E(\mathrm{full})_{\mathrm{CCSD(T)/DZ}} \approx E(\mathrm{MTA})_{\mathrm{CCSD(T)/DZ}} + E(\mathrm{full})_{\mathrm{MP2/DZ}} - E(\mathrm{MTA})_{\mathrm{MP2/DZ}} \tag{9}$$

MTA energies at the CCSD(T)/aug-cc-pVDZ level with grafting correction reproduce the full calculation results with remarkable high accuracy matching up to fourth decimal point in energy (hartrees). These results clearly demonstrate that

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

adding a second layer of theory in fragment based methodologies is an effective alternative to the use of embedding charges to account for long-range interactions.

Gadre and co-workers have also tested the applicability of MTA method for different sizes of CO_2 clusters.[114−116] A series of $(CO_2)_n$ clusters (n = 6−13) have been generated by adding CO_2 monomers to smaller CO_2 clusters by an automated cluster-building algorithm. The low-lying isomers of these clusters predicted by MTA resemble the optimized geometries reported by Takeuchi.[117] Although single level MTA reproduces the true optimized geometries with reasonable accuracy, upon grafting correction, the relative energy orderings are in very good agreement with those observed by Takeuchi.[117] Vibrational frequency calculations on CO_2 clusters have also been performed with MTA at MP2/aug-cc-pvDZ level. MTA results show a very good agreement with the results from the direct calculations.

The MTA has been applied to several one- and two-dimensional π-conjugated systems with density functional theory (at M06/6-31+G(d, p) and B3LYP/6-31++G(d,p) levels of theory). In a few cases, MP2 (that includes dispersion interactions that may be important in such systems) has also been used. The π-conjugated systems include β-carotene, retinal, a polyacetylene fragment, heptacene, and a BN nanotube (Figure 3). It has been shown that, for such π-conjugated systems, MTA single point energy calculations can successfully predict the total energy of the system within an error of ∼1 mH. However, the authors have mentioned that larger fragments are necessary to obtain high accuracy for the two-dimensional π-conjugated systems.

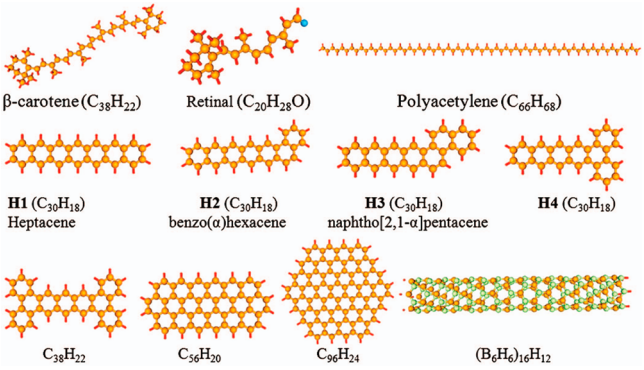


Figure 3. One- and two-dimensional π-conjugated systems studied with MTA method by Gadre et al.[100] Reprinted with permission from ref 100. Copyright 2010 American Institute of Physics.

Gradient calculations[118] and geometry optimizations for these π-conjugated systems (Figure 3) have also been performed with MTA. Subsequent MTA-based single point energies (at B3LYP/6-31+G(d)) have been determined on the MTA-optimized geometry. Detailed comparison of the geometrical parameters show that MTA-optimized structures are in good agreement with the full calculations, leading to mean deviations smaller than 0.005 Å for C−C bond lengths in heptacene and benzo(α)hexacene.

5.2. Electrostatically Embedded Molecular Tailoring Approach (EE-MTA)

In EE-MTA, developed by Truhlar and co-workers,[119] the authors have included long-range electrostatic interactions and higher order polarization effects by augmenting the molecular tailoring approach by embedding each of the fragments in background charges, similar in spirit to the procedure used in QM/MM methods.

The aim of EE-MTA was to increase the accuracy level of MTA. Several important factors were considered in developing the method. (1) Traditional MTA includes only short-range and medium-range interactions within the fragment selected, and does not include the combined effect of electronic polarization and long-range electrostatics, potentially leading to significant errors in highly polar or charged systems. A key factor is to consider interfragment electrostatic interactions by including the background molecular charges from the rest of the fragments. (2) An important consideration is to improve the accuracy of the type of cap atom (i.e., link atom) used when covalent bonds are cut to make the fragments in peptide systems. While hydrogen link atoms are most commonly used, there are other advanced methods available that use a tuned atom in which a pseudo or effective core potential[120−122] containing parameters is adjusted to reproduce specific properties. In EE-MTA, the authors developed general parameters for a fluorine cap atom suitable for all peptide systems. This was accomplished by introducing parameters that minimized the error in the calculated proton affinities of peptide systems. (3) In the traditional MTA, one has to choose relatively larger fragments in order to achieve higher accuracy which eventually contributes to the computational cost of MTA. Since each fragment in EE-MTA is embedded in the background molecular charges of the rest of the fragments, the local chemical environment can be retained without necessarily choosing larger fragments. (4) In normal partitioning of a large molecule into fragments, it is preferable to cut a nonpolar bond (such as the C−C bond). However, it may be desirable to have methods that do not significantly depend on such restrictions. This was examined and found to be valid for EE-MTA.

The total energy of the system in EE-MTA is given for the general case as

$$E^{\text{EE-MTA}} = \sum E_I - \sum E_{I\cap J} + \sum E_{I\cap J\cap K} + \cdots + (-1)^{N-1} \sum E_{I\cap J\ldots\cap N} - E_{\text{OC}} \quad (10)$$

where E_I denotes the electronic energy of fragment I that is coupled with the electrostatic potential of the rest of fragments; $E_{I\cap J}$ denotes the energy of the overlapping section of two fragments I and J; E_{OC} denotes the sum of the Coulomb interactions that are overcounted. In the version of EE-MTA implemented, the authors truncate the previous equation at the 2-body level:

$$E^{\text{EE-MTA}} \approx \sum E_I - \sum E_{I\cap J} - E_{\text{OC}} \quad (11)$$

The overcounted term E_{OC} was analyzed carefully, and two different treatments were considered: a straight implementation of the Coulomb overcounting, and a more sophisticated treatment as considered in the general GEBF[138] method (*vide infra*) in its treatment of embedding. The latter procedure was found to be much more accurate for many systems and was implemented in the general EE-MTA scheme.

The applicability of EE-MTA method was demonstrated by applying it on two conformations (α-helix and β-sheet) of Ace-$(Ala)_{20}$-NMe (Figure 4). In a careful study of the key factors, the authors reached several important conclusions: (1) In general, the tuned fluorine atom performed much better than simple hydrogen link atoms. However, significant deficiencies

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

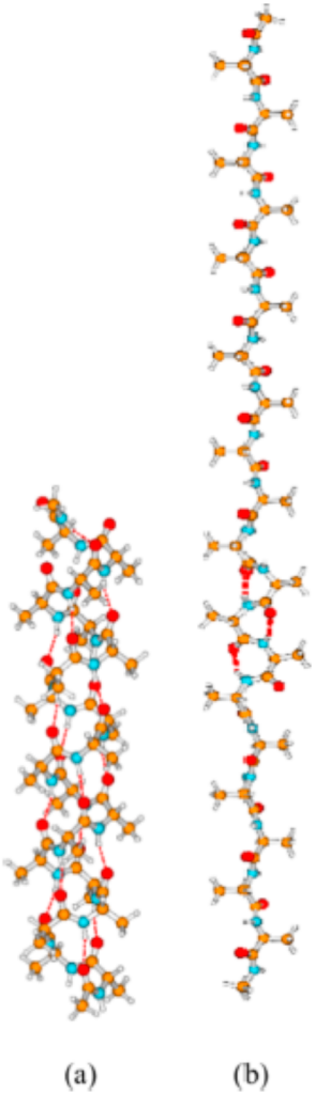


Figure 4. Optimized geometries of Ace-$(Ala)_{20}$-NMe tested with EE-MTA approach.[119] Reprinted with permission from ref 119. Copyright 2013 American Chemical Society.

were still found for some peptide sequences in their protonated forms. (2) The background charges in the EE-MTA method provide a significant improvement in the relative energy between the α-helix and the β-sheet. While larger fragments led to better performance, most of the improvement came from the embedded charges yielding a better energy for the α-helix. While the electronic polarization had only a smaller effect on the β-sheet, the improved energy of the α-helix led to much better relative energies between the two conformations. (3) Much smaller fragment sizes were sufficient in EE-MTA compared to the traditional MTA. For example, EE-MTA, using primary tetramer fragments was more accurate than the MTA method using primary heptamer fragments. (4) The location of the cut bonds had only a small effect. The peptide fragments obtained by cutting the C−C or the N−C bonds performed reasonably well, though the former provided the best results.

Overall, the authors illustrate the better computational cost and high accuracy associated with EE-MTA as compared to the MTA.

5.3. Molecular Fractionation with Conjugate Caps (MFCC)

MFCC is an efficient fragment-based method, initially designed by Zhang et al. for *ab initio* calculations of protein−ligand interaction energies.[123,124] The original scheme in MFCC was to decompose a protein into smaller amino acid fragments to enable linear scaling of the calculated energies. The key feature of MFCC method is that a pair of "conjugate caps" (*vide infra*) is inserted at the location of the cut-bond when a portion is excised from the entire system. Using the fractionation scheme, the interaction energy between the protein and the ligand can then be computed by performing separate calculations on each fragment interacting with the ligand.

In MFCC, the test system (protein) is represented as

$$P = nA_1 - A_2 - A_3 - \cdots - A_N c \tag{12}$$

where A_i (i = 1, ..., N) denote the sequence of amino acid units. nA_1 denotes the first amino acid at the N terminal of the protein, i.e., n = NH_3^+ (or NH_2) for charged (or neutral) terminal species. $A_N c$ denotes the last amino acid at the C terminal of the protein where c = HOO^- (or COOH) for charged (or neutral) terminal species. Since the protein−ligand interaction is assumed to be local, the interaction of protein P with an arbitrary ligand L is represented as the summation of interactions of the individual fragments of protein P with L. In the standard fractionation scheme, the peptide bonds (N−C_α bonds) of the protein are cut. At each location of a cut bond, a pair of caps are added which are conjugate to each other ("concaps"). These caps are denoted as C_{ap} (i.e., CH_2R_iCO-) and C^*_{ap} (i.e., $-NHCH_2R_{i+1}$). The concaps are introduced in such a way that they serve two important functions: (1) serve the purpose of saturating the dangling bonds to preserve the valence requirement of the cut bond and (2) essentially mimic the local chemical and electrostatic environment around the cut-bond location. Hydrogen atoms are added as needed to terminate the dangling bonds. In addition, the pair of concaps is fused to form a proper molecule that can then be used to subtract the doubly counted interaction energies of the caps with the nearest ligands (Figure 5). It can be seen that the use of concaps creates overlapping fragments making the method very similar to several other fragment-based techniques.

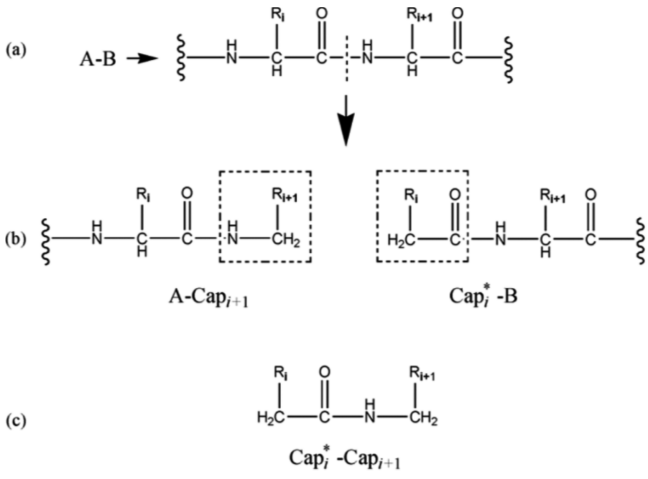


Figure 5. (a) MFCC bond cutting scheme. (b) Cut-bonds with their concaps, C_{ap} and C^*_{ap}. (c) The pair of concaps shown as a fused molecular species.[124] Reprinted with permission from ref 124. Copyright 2014 American Chemical Society.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

The interaction energy between the protein P and the ligand L is denoted as $E(P-L)$. For a protein with N amino acids, there are $N-2$ capped fragments and $N-3$ concaps. Using the fractionation scheme, the interaction energy can be expressed as

$$E_{P-L} = \sum_{k=1}^{N-2} E_{F_k-L} - \sum_{k=1}^{N-3} E_{CC_k-L} - \sum_{k=1}^{N-2} E_{F_k} + \sum_{k=1}^{N-3} E_{CC_k} - E_L \quad (13)$$

The first term is the total energy of the kth capped fragment and ligand. The second term is the total energy of the kth concap and ligand. The third and fourth terms are the self-energy of kth capped fragment and the kth concap, respectively, and the last term is the energy of the ligand. It is important that the geometry of the conjugate caps is preserved in the artificial molecule as in the original protein. This is important to cancel out the artificial (i.e., overcounted) interactions. Thus, prediction of the interaction energy between P and L can be done by summing up several independent energy terms. The method is quite flexible and can be used with any kind of low (HF, DFT) or high level (CCSD(T)) *ab initio* methods.

The applicability of MFCC has been successfully demonstrated on peptide systems.[125−128] In the initial application, the interaction energies of three individual peptide systems with a single water molecule were obtained. The peptides chosen were the following: (1) Gly-Gly-Gly (three glycine residues with charged terminals), (2) Me-His-Ser-Me (two amino acid residues capped with methyl groups on both sides), and (3) Gly-Ser-Ala-Asp-Val (five amino acid residues). For all three systems, the interaction energy of a single water molecule in its rigid orientation approaching the frozen geometry of the peptide at different spherical angles was calculated by MFCC and compared with direct results. Using HF and DFT methods with several relatively small basis sets, MFCC was found to reproduce the results of the direct calculations of the 1D potential energy surfaces with high accuracy.

Initially, MFCC focused only on calculating the protein−ligand interaction energies. Further developments have extended the method for calculation of the total density matrix of the system to evaluate the total energy and other molecular properties of the system. The density matrix elements are constructed in terms of fragment density matrices, and can be used to derive the HF or DFT total energy. Such a variation of the MFCC method to derive the density matrices and total energies has been termed as "MFCC-DM" method.[129] A new version of MFCC called "MFCC-DM-PIC"[130] (molecular fractionation with conjugate caps-density matrix-pairwise interaction) was also proposed to account for electron density polarization due to short-range interactions.

As mentioned earlier, hydrogen link atoms are added in the concaps for termination of the dangling bonds. Since they are not part of the parent molecule, elements of density matrices related to these extra hydrogen atoms have to be treated appropriately. Two types of implementations have been used in MFCC to take this into account. In the simple "MFCC-SDM" approach (SDM stands for small density matrix), the residual density matrix involving the extra hydrogen AOs are completely neglected. While the associated computational cost of this approach is low, it does not conserve the total number of electrons in the system due to imperfect cancellation of the density matrix elements involving extra hydrogen AOs. In the more rigorous MFCC-GDM implementation, these hydrogen atoms are treated as ghost atoms and their AOs are considered explicitly. While the associated computational cost is higher, the total number of electrons is perfectly conserved since residual density matrix elements involving extra hydrogen AOs are explicitly considered. The MFCC-SDM and MFCC-GDM implementations have been tested on extended peptide systems $(GLY)_n$ where n = 3−25 for both HF and DFT methods.

The development of MFCC has been further extended to treat macromolecules with several charged centers. The energy−corrected MFCC method (EC-MFCC) has been proposed to include the long-range interactions between nonbonded fragments (controlled by a preselected truncation distance).

$$E_{\text{total}}^{\text{EC-MFCC}} = \sum E(\text{capped fragments}) - \sum E(\text{conjugated capped}) + \Delta E^{(2)} \quad (14)$$

where $\Delta E^{(2)}$ represents the 2-body through-space interactions between nonadjacent but spatially proximal fragments. Although successful in reproducing the interaction energies of several systems, the EC-MFCC[131] method performs less well in systems with several charged centers where electrostatic interactions fade very slowly over the distance. In order to address the limitations of EC-MFCC, an additional variation, EFA-MFCC (electrostatic field adapted MFCC method), has been presented for the calculation of ground state energies of biological macromolecules and their interactions.

In the EFA-MFCC[132] method, the original formulation of MFCC is used, but the calculation of each fragment and their concaps are executed in a background electrostatic field. This field is simulated by point charges located at the charge centers which then affect the Hamiltonian of the individual fragments and their concaps. The advantage of EFA-MFCC approach over EC-MFCC method is that one can now really avoid the criteria of preselected truncation of long-range electrostatic interactions. By adopting the EFA-MFCC approach, the errors improve moderately over EC-MFCC, but the authors expect the advantage of EFA-MFCC method to emerge more significantly for more complex systems such as DNA fragments.

The applicability of EFA-MFCC has been assessed on the following systems: (1) $(GLY)_n$ (n = 12, 14, 16, and 20), (2) a pentapeptide containing five phenylalanine residues and cysteine dimer connected with disulfide bonds. In EFA-MFCC, the authors also recommend that bonds involving strong electron delocalization should not be separated into two different capped fragments. In order to analyze the effect of different charge schemes, three different models, with charges derived from (1) NBO, (2) ESP, and (3) Mulliken population analyses, were considered with HF and DFT methods. Mean absolute deviations of EFA-MFCC for HF/6-31G(d) and B3LYP/6-31G(d) relative to the direct full system calculations are within a few milihartrees for several test systems, 3.26 (HF) and 5.57 (DFT) mhartrees, respectively, using Mulliken charges.

In order to make MFCC more general, robust, and practical, for applications on large systems (i.e., geometrical optimization of protein complexes, molecular dynamics simulations), Zhang et al. have proposed a more efficient "GMFCC/MM" approach where they implemented a mixed QM/MM scheme for predicting the full QM energy of the protein. In GMFCC/MM,[133] the interaction energies of neighboring residues as well

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

as non-neighboring residues (sequentially not connected) in close spatial contact are treated with QM methods whereas the long-range interactions are treated with molecular mechanics. Thus, a high level QM method can be applied to account for strong short-range interactions while the long-range very weak interactions (intractable with QM methods) are handled with MM. Inspired by the spirit of QM/MM[134−136] in studying chemical reactions in proteins, GMFCC/MM method has been developed by treating each part of the system as a model system (active site).

To include the important effects of sequentially non-neighboring fragments that may be in close contact in the GMFCC/MM method, the authors have introduced the concept of "generalized concap (Gconcap)". If the minimal distance between two non-neighboring residues is within a predefined distance (4 Å), the two residues are considered to be in close contact (defined as Gconcap), and their interactions are calculated at the QM level.

Since in GMFCC/MM the environmental effect was not included, the authors further developed the GMFCC method to present "EE-GMFCC" (electrostatically embedded GMFCC)[137] where the total energy of the system is predicted by linear combination of fragment-based energies of neighboring residues and by considering the interaction energies of non-neighboring fragments with the constrains that each fragment calculation have to be performed in the embedded field of point charges to mimic the local chemical environment of each fragment.

Assessment of applicability of EE-GMFCC method has been carried out on a broad range of protein molecules (18 three-dimensional globular proteins)[133] using HF, DFT, and MP2 methods with the 6-31G* basis set. Two types of charge models have been used in the numerical studies: (1) fixed charge model and (2) polarized point-specific charges (PPCs). The overall mean-unsigned-error of EE-GMFCC method on all these small protein systems is 2.39 kcal/mol[133] with a distance threshold of 4.0 Å at the HF/6-31G* level. Results of EE-GMFCC applied with DFT/6-31G* and MP2/6-31G* theoretical models also have similar errors.

5.4. Generalized Energy Based Fragmentation (GEBF)

GEBF is one of the popular and effective fragment-based methods developed in recent years.[138−143] In GEBF, the ground state energy of a large molecule is assembled from the corresponding energies of overlapping subsystems, each of which is embedded in the background charges of the atoms in other subsystems. The initial development of the GEBF method focused on recovering only the energy, but it has subsequently been extended to calculate energy derivatives and to predict molecular properties (e.g., vibrational frequencies). Since the GEBF method has been applied to a wide range of systems at different theoretical levels, the authors use GEBF-X to denote a GEBF calculation using level X (X = HF, MP2, DFT, etc.).

The total energy of the system in GEBF method is expressed as[143]

$$E_{\text{tot}} = \sum_{m}^{M} C_m \bar{E}_m - [(\sum_{m}^{M} C_m) - 1] \sum_{A} \sum_{B>A} \frac{Q_A Q_B}{R_{AB}} \qquad (15)$$

where $\bar{E}_m$ denotes the total energy of the mth subsystem (including the self-energy of background point charges), C_m denotes the coefficient of the mth subsystem, Q_A is the net atomic charge on atom A, and M is the total number of subsystems.

The overall fragmentation scheme in GEBF can be described as follows: (1) Fragment the entire system into smaller pieces of comparable size by cutting single bonds or hydrogen bonds. An automatic fragmentation procedure with a number of functional groups in its database has been developed[139] to handle large molecules. (2) Each of these fragments is then connected with neighboring fragments (within a given distance threshold (ξ) of 4 Å) to include the interactions most important for this "central fragment". These enlarged fragments are then termed as "primitive subsystems". The coefficients C_n for the primitive fragments are taken as +1. While constructing these "primitive subsystems", the detachment of this section from the rest of the system is accomplished by replacing broken covalent bonds with the bonds to hydrogen. (3) The next step of GEBF determines "derivative subsystems" to take into account the overcounting of some fragments due to the overlapping nature of primitive subsystems. The derivative subsystems usually have a coefficient of −1. (4) The net atomic charges on all the atoms are then computed to take into account long-range electrostatic interactions. This is done sequentially. At the first step, initial atomic charges are calculated by running HF (or DFT) calculations on primitive subsystems independently and using the natural charges. Subsequently, redefined atomic charges for all the atoms can be obtained by calculating the primitive fragments embedded in the electrostatic field of the other subsystems (from separate independent subsystem calculations). While converged atomic charges can be derived by an "iterative procedure", the natural atomic charges from the first iteration appear to be sufficient in most cases. (5) The energy of the entire system is then predicted by proper summation of the embedded subsystem energies (obtained with a standard electronic structure program package) including their coefficients.

Determinations of energy derivatives and different molecular properties have also been implemented within GEBF. Dipole moments and static polarizabilities of large molecules can be approximately evaluated with the GEBF method as[139]

$$\Omega_{\text{tot}} = \sum_{m}^{M} C_m \tilde{\Omega}_m(\Omega = \mu_i, \alpha_{ij,}...) \qquad (16)$$

The fully analytic gradient of the entire system can be achieved within GEBF by the following expression:

$$\frac{\partial E_{\text{tot}}}{\partial q_A} = \sum_{n} C_n \left(\frac{\partial \tilde{E}_n}{\partial q_A} - F_{n,a} Q_a - \sum_{b} f_{ab} \right) + [(\sum_{m}^{M} C_m) - 1] \sum_{b \in \text{all}} f_{ab}$$

Here A denotes a real atom in a given subsystem, a and b denote the point-charge centers, $F_{n,a}$ represents the electric field generated by the nth subsystem on the center a (which can be calculated with any *ab initio* program), and f_{ab} represents the Coulomb force between charge on b and charge on a.

In addition, an approximate gradient evaluation was introduced, and it was suggested that this also performs effectively.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

$$\frac{\partial E_{\text{tot}}}{\partial q_A} \approx \sum_n C_n \left(\frac{\partial \tilde{E}_n}{\partial q_A} \right) \tag{17}$$

The accuracy of the GEBF-gradient method was demonstrated on 10 randomly selected structures[139] of a hydrogelator during its optimization at the HF/6-31G(d) theoretical model. The GEBF method appeared to perform quite well for these systems using either gradient expression. The RMS deviations between conventional HF and GEBF-HF calculations are 0.0005 au/bohr with the exact gradient expression and 0.0006 au/bohr with the approximate gradient expression.

The GEBF gradient method has been applied to locate lowest-energy structures of large water clusters (n = 20−30).[144] A selected set of low energy water cluster isomers chosen from the polarizable AMOEBA force field were optimized at the GEBF-B3LYP/6-311++G(d,p) and GEBF-MP2/6-311++G-(3df,2p) levels. The authors concluded that water clusters undergo a transition from one-centered to two-centered cage structures at n = 26. In this application, each water cluster was considered as a fragment, and the parameter η was (defined as "the maximum number of fragments in any subsystem") set to 6. With the value of η = 6, the maximum absolute error and mean absolute error was 1.23 and 0.74 kcal/mol, respectively, relative to the full system calculations. The maximum absolute error in relative energies of these water clusters was 1.08 kcal/mol (or 0.41 kcal/mol mean error). The vibrational spectra of water clusters (n = 28) as well as peptides (e.g., $(Gly)_{12}$) have also been computed using the GEBF approach. Figure 6 shows a few of the bonded and nonbonded systems investigated with the GEBF method.

Recently GEBF method has been extended to multilayer hybrid models to treat different parts of the system using different theoretical levels, spanning from *ab initio* electronic structure methods to MM methods. Two kinds of multilayer models have been proposed: (1) simultaneous multilayer model (GEBF X:Y method) and (2) sequential multilayer model. In the simultaneous multilayer model, the system is basically partitioned into an active site (treated with high level GEBF-X where X = MP2, for example) and the rest of the system (treated with low level GEBF-Y where Y = HF, for example) like in the popular ONIOM[108] approach. In the second type of multilayer hybrid model, the system is sequentially fragmented as in the ONIOM approach where different layers are treated with different levels of theory. In a recent application,[145] the sequential hybrid GEBF method has been successfully applied to calculate the binding energies of methanol molecules surrounded by water molecules where three different shells or layers (A methanol and its neighboring water molecules within 4 Å (first solvent shell), the water molecules between 4 and 9 Å, and the remaining part (up to 20 Å)) have been treated with GEBF-MP2-F12, GEBF-MP2, and DFTB (density functional theory tight-binding) levels of theory, respectively.

Another multilayer hybrid GEBF model has been proposed to reduce the computational cost of polarizable coarse-gained MM methods.[146] In this model, GEBF-based QM calculations are performed on a predefined set of fragments to estimate the atomic charges and dipole moments which are then further used as input parameters in polarizable MM methods. Such approaches may be useful in studying complex phenomena (e.g., protein−protein interactions) in biologically interesting systems.

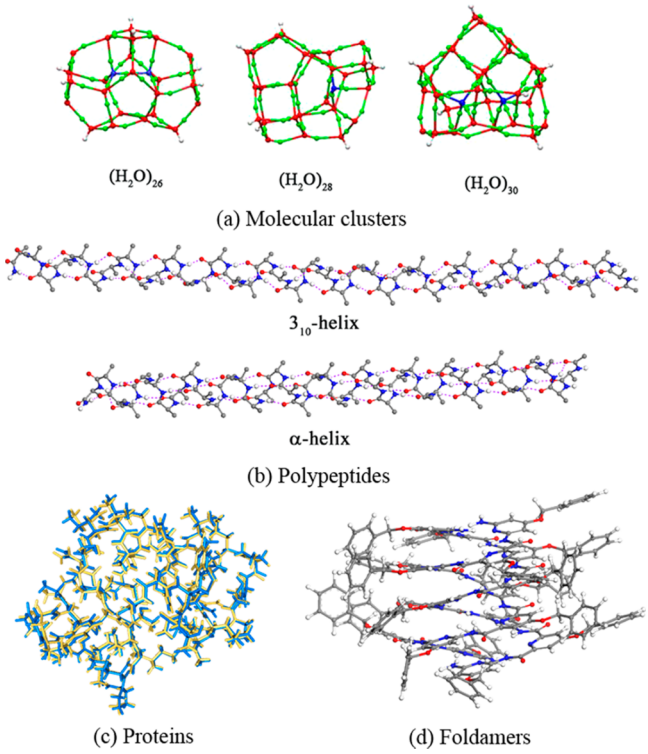


Figure 6. Broad ranges of bonded and nonbonded systems investigated with GEBF.[143] Reprinted with permission from ref 143. Copyright 2014 American Chemical Society.

5.5. Systematic Molecular Fragmentation (SMF)

The SMF method and the updated "systematic molecular fragmentation by annihilation" (SMFA) have been developed by Collins and co-workers, and have been used in a wide range of applications for medium to large molecules. The original SMF method required a "ring repair" procedure to handle some special cases, and has been replaced by the more recent SMFA method that treats such systems automatically and is more easily applicable for very large molecules.[147]

SMFA starts with the molecule M divided into different groups ("groups" are sets of atoms defined by the user or derived following a prescription recommended by the authors). The basic ideas involved in the method can be illustrated for the simplest case involving a chain-like molecule containing N groups connected by single bonds:

$$M = G_1 G_2 ... G_N \tag{18}$$

The target is to derive an accurate value for the total electronic energy:

$$E(M) = E(G_1 G_2 ... G_N) \tag{19}$$

The energy of the molecule M is determined by summation over calculations on fragments (F_n) that are defined in terms of collections of groups. The sizes of the fragments depend on the "Level" of SFM, whereby a given group is assigned to a different fragment if it is separated from others by at least Level groups in the bonded sequence. The fragments can overlap with each other (i.e., a group can belong to multiple fragments). As in many other methods involving overlapping fragments, additional fragments with negative coefficients are generated to cancel the effects of multiple counting.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

The associated "bonded" energy is expressed as a sum over all the N_{frag} fragments

$$E_b = \sum_{n=1}^{N_{frag}} f_n E(F_n) \tag{20}$$

where f_n is an integer coefficient.

For a model system of chain containing five groups, SMFA fragmentation scheme can be expressed as

$$\begin{aligned} G_1G_2G_3G_4G_5 &= G_1G_2 + G_2G_3 + G_3G_4 + G_4G_5 - G_2 - G_3 \\ &\quad - G_4 \text{ (level 1)} \\ &\rightarrow G_1G_2G_3 + G_2G_3G_4 + G_3G_4G_5 - G_2G_3 - G_3G_4 \\ &\quad \text{(level 2)} \\ &\rightarrow G_1G_2G_3G_4 + G_2G_3G_4G_5 - G_2G_3G_4 \text{ (level 3)} \end{aligned} \tag{21}$$

Thus, the fragment sizes increase with the Level used. However, the number of fragments grows linearly with the size of the system. The authors have noted[92] that the different Levels used in SMFA are related to some older concepts used in the field of theoretical thermochemistry. For example, Level 1 reactions are known as "isodesmic reactions", Level 2 is known as homodesmotic reactions, and Level 3 is known as hyperhomodesmotic reactions.

Since the "bonded" energy as defined above only includes nearby interactions, the authors include the "non-bonded" interactions between more distant groups using a composite approach. In the simplest case involving Level 1 fragmentation, the nonbonded interactions are evaluated by the following equation within SMFA

$$E_{nb} = \frac{1}{2} \sum_{n_1=1}^{N_{frag}^{(1)}} \sum_{n_2=1}^{N_{frag}^{(1)}} f_{n_1}^{(1)} f_{n_2}^{(1)} E[F_{n_1}^{(1)} \leftrightarrow F_{n_2}^{(1)}]_{\text{allowed}} \tag{22}$$

where a nonbonded interaction is "allowed" if it is not already included in E_b. For nearby fragments, the fragment−fragment interaction energy is obtained by supermolecule calculations, whereas the interactions between more distant fragments can be approximated using perturbation theory, including contributions from electrostatics, induction, and dispersion. In another variation, Collins, Gordon, and co-workers have explored the efficacy of obtaining the nonbonded interactions with the effective fragment potential (EFP, *vide infra*) method.

As in many other methods, hydrogen link atoms are used in SMFA method when single bonds are broken. The hydrogen atom distance is taken to be proportional to the original heavy-atom distance with a proportionality factor that depends on the ratio of the covalent bond lengths, as in the standard ONIOM method.[108]

The SMFA method has been optimized to perform well for very large molecules by using a "compression" algorithm[150] where the groups are compressed sequentially by factors of 2, followed by an expansion process that reverses the iterative steps. This allows it to perform effectively for large molecules containing thousands of atoms. The authors have illustrated that fragments generated in this way are unique, independent of the ordering of the atoms. Additionally, the method has been adapted to work on periodic systems to derive the electronic structure and energy of nonconducting crystals.

The SMFA method has been applied to calculate the energies of a variety of bonded and nonbonded systems. In addition, the potential energy surfaces for molecules as well as their interaction surfaces with radicals have been derived by SMFA method, enabling them to be used in dynamics simulations. In such applications, the higher Levels of SMFA can be used to derive energy corrections to the potential energy surfaces obtained with lower Levels. Energy derivatives have been used with SMFA to derive properties such as molecular geometries, vibrational frequencies, and NMR chemical shifts.

Since its original formulation, SMF has been systematically applied to a large set of chemical systems as it continues to grow through several theoretical developments. One of the earliest applications[92] of SMF involves substituted alkane molecules. The performance of different Levels of SMF, after the inclusion of nonbonded interactions, reaches "chemical accuracy" for such systems. As mentioned earlier, the early formulation of SMF had some difficulties for ring systems (e.g., piperidine) that were taken care of in subsequent refinements[148] along with improvements for long-range interactions and dispersion effects. The new formulation of SMF has been tested on a set of typical organic molecules containing 96 molecular structures from Cambridge Structural Database having chemical formulas $C_{7-10}N_{0-7}O_{0-7}F_{0-3}H_{1-80}$. Functional groups common to these systems include amines, cyanides, ethenes, esters, ethers, aldehydes, ketones, alcohols, carbon rings, and heterocycles with four or higher members. An important note here is that the fragmentation scheme applied to these systems does not break a bond having a bond order greater than 1. Performance of ring-repaired SMF approach to estimate the total electronic energy and relative energies were consistent with "chemical accuracy". SMF has also been applied systematically to construct a global molecular potential energy surface[149] involving reactions of modest-size molecules (e.g., reaction of hydrogen atom with *n*-pentane) from the corresponding surfaces of small molecular fragments. This is an important example involving the investigation of potential energy surfaces by fragment-based methods.

To demonstrate the applicability of SMFA on very large systems,[150] Collins has carried out a study on a large protein system TM1081,[151] containing 2048 atoms (composed of 615 groups with a net charge of −5). 80 different starting geometries of this protein were initially selected. Energy optimizations of these structures in a "water shell" were carried out with the AMBER force field[152] to identify the 20 lowest energy isomers for study with SFMA. Fragmentation Levels 1−4 generate 1236, 895, 893, and 667 fragments, respectively. The average numbers of atoms in these fragments are 6, 10, 15, and 20, respectively, while the largest fragments contain 19, 22, 30, and 39 atoms. It is important to mention that conjugated rings (if involved in any system) are not broken at any level of fragmentation. The energies of all the structures were determined at the HF/6-31G level, and compared with SMFA results with and without the presence of background charges. A simple embedded charge approach where the energy of each fragment is obtained in the presence of background charges (on all the atoms in the molecule not contained in the fragment) was used to account for the electrostatic and induction effects. An iterative procedure was used to obtain approximate self-consistent charges. Neglecting dispersion effects (because only HF calculations were reported in this particular investigation, the authors neglected the dispersion interaction consistently[150]), the mean absolute deviation in

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

relative energy for these 20 geometries between Level 3 and Level 4 is 42 kJ/mol, relatively small relative to the total bonding energy of over 1000 kJ/mol, but still large on an absolute scale. The size of the system in this particular case (2048 atoms) may have pushed the limit of the SMFA method to its boundary. Better performance may be expected upon proceeding to the higher levels of fragmentation though it may make the computations prohibitively expensive. Moreover, dispersion effects may also have to be included to assess the performance of SMFA in deriving reliable energies for such biomolecules. Finally, applicability of SMFA approach has also been illustrated by applying it to wide range of crystals and crystal surfaces.[153]

5.6. Combined Fragment-Based Method (CFM)

CFM, developed by Bettens and co-workers,[154] bears many similarities to the SMFA method. The total energy expression in CFM is similar though it is written slightly differently as

$$E_{\text{tot}} = \sum_{i=1}^{N_{\text{frag}}} f_i E(\hat{h}F_i) \tag{23}$$

Here f_i is an integer coefficient, and $E(\hat{h}F_i)$ represents the energy of fragment F_i (hydrogen capping is denoted as $\hat{h}$). The initial groups in CFM are typically larger than that in SFMA, and a "precursory fragmentation" of the molecule is carried out, typically at Level 1 of the SMF method. The hydrogen capping procedure in CFM involves standard bond lengths, different from the treatment based on covalent radii in SMFA.

In CFM, after the precursory fragmentation, the fragments interact with each other. The interaction energy between a pair of fragments in CFM is defined as

$$\varepsilon(F_i, F_j) = E(\hat{h}\{F_i \cup F_j\}) - E(\hat{h}F_i) - E(\hat{h}F_j) + E(\hat{h}\{F_i \cap F_j\}) \tag{24}$$

The last term involving a correction for overlapping fragments allows the expression to be simpler without having to determine whether an interaction is "allowed" as in SMFA. The total CFM interaction energy is given as

$$\varepsilon = \sum_{j>i}^{N_{\text{frag}}} f_i^p f_j^p \varepsilon(F_i, F_j) \tag{25}$$

The total combined fragment-based energy in CFM is then given as

$$E_{\text{frag}} = \sum_{i=1}^{N_{\text{frag}}} f_i E(\hat{h}F_i) + \sum_{j>i}^{N_{\text{frag}}} f_i^p f_j^p \varepsilon(F_i, F_j) \tag{26}$$

While the number of such calculations can be large, many of the nonbonded interactions can be evaluated efficiently by perturbation theory. If the groups in the molecule form small cycles (containing 3 or 4 groups), additional modifications are included in CFM to improve accuracy without increasing the computational cost.

Electrostatically embedded CFM method has also been developed to include long-range Coulomb interactions as in SMFA. Again, each fragment is embedded in background molecular charges from the rest of the molecule. In this case, point charges are evaluated by Stone's distributed multipole analysis.[155] The charges are determined iteratively using only the bonded contributions, and the total energy including the nonbonded contributions is then obtained in the field of those charges.

CFM has been applied to determine different molecular properties including central and distributed multipoles and molecular electrostatic potentials.[154] CFM has also been used in a wide range of applications including estimating the density matrix of the entire molecule or self-consistent reaction field (SCRF) energy of a target system. In a recent article,[156] the authors have also carried out applications to calculate the *ab initio* NMR chemical shift via the CFM approach. They observed that, without nonbonded interaction, the RMS errors for ^{1}H, ^{13}C, ^{15}N, ^{17}O, and ^{33}S chemical shifts are 0.340, 0.649, 3.052, 6.928, and 0.122 ppm, respectively, while inclusion of nonbonded interactions results in smaller RMS errors of 0.038, 0.253, 0.681, 3.480, and 0.052 ppm.

Figure 7 shows some of the systems which have been treated with both CFM and SMFA methods. Both methods show very good performance on all these systems. Errors associated with CFM approach on β-carotene, γ-cyxlodextrin, peptide IOAI, and 57-water cluster are −0.592 mE_h (B3LYP/6-311G(d)), 5.638 mE_h (B3LYP/6-311G(d)), −0.901 mE_h (HF/6-311G(d)), and 6.648 mE_h (HF/6-311G(d)), respectively. The corresponding errors associated with SFMA approach are 0.326 mE_h (B3LYP/6-311G(d)), 4.350 mE_h (B3LYP/6-311G(d)), 1.940 mE_h (HF/6-311G(d)), and 1.107 mE_h (HF/6-311G(d)), respectively.

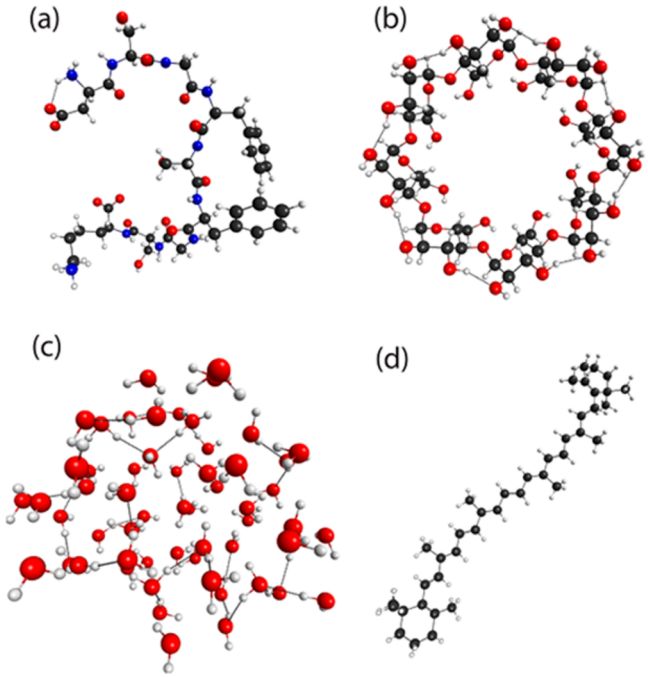


Figure 7. Systems treated with CFM and SMFA methods: (a) peptide IOAI, (b) γ-cyxlodextrin, (c) 57-water cluster, (d) β-carotene.[147] Reprinted with permission from ref 147. Copyright 2014 American Chemical Society.

5.7. Molecules-in-Molecules (MIM)

MIM, a fragment-based method proposed by Mayhall and Raghavachari,[113] uses a multilayer partitioning technique with multiple levels of theory using a generalized hybrid energy expression, similar in spirit to the popular ONIOM methodology. The central point of MIM that makes it distinct from

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

most other fragment-based methods is the use of a multilayer approach using multiple levels of theory.

The starting point in MIM that deals with fragment formation within a single layer is similar in many respects to the approach followed in GEBF. This can be described in four steps.

In the first step, initial fragmentation of the system is carried out. This can be carried out using an automated black-box approach or by a customized approach as defined by the user. In the black-box approach, fragments are generated by breaking each "single" bond between non hydrogen atoms. The fragments play the role of "atoms" in the sense that they are the most fundamental units and are not divided further within the subcalculations.

Each of the fragments from step 1 interacts with nearby fragments to initiate the formation of a primary subsystem in step 2. Formation of subsystems resulting from the local interactions between fragments can be done according to one of several prescribed fragmentation parameters (via connectivity by including one or more directly bonded fragments, or via a distance-based cutoff (r), or via a number-based cutoff (η)). The number of primary subsystems is reduced by considering only unique ones that are not subsets of other primary subsystems. Since link atoms are used to replace broken covalent bonds, steps are taken to ensure that the same center is not replaced by two link atoms (in situations involving rings, for example).

As in many other fragment-based methods, the consequence of using overlapping primary subsystems is that much of the molecule is overcounted. In step 3, derivative subsystems (overlapping regions between primary subsystems that have to be canceled) are formed according to the Inclusion-Exclusion principle. The name of the method (molecules-in-molecules) comes from the fact that the primary and derivative subsystems (after link-atom termination if needed) play the role of "small molecules" composing the large molecule.

$$|A_1. \ldots \cup A_n| = \sum |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| \ldots + (-1)^{n-1} |A_i \cap \ldots \cap A_n| \quad (27)$$

In step 4, energies of all these subsystems are summed up with the appropriate coefficients (derived from the Inclusion-Exclusion principle[113]) to predict the total energy of the system. This yields the MIM energy within a single layer (MIM1).

Most of the hybrid energy methods (e.g., QM/MM) use a single layer partitioning scheme. As mentioned earlier, the key distinctive feature of MIM is the use of multiple layers to handle the long-range interactions instead of the electrostatic embedding approach followed in many other approaches. For example, MIM2 (MIM with 2 layers) indicates a scheme with two fragmentation parameters (e.g., $r < R$). Fragments generated with the smaller parameter are treated with a higher level of theory and the ones with a larger parameter at a lower level of theory. It is also possible to include the entire molecule at the lower level of theory. In such a two layer approach, the long-range interactions are accounted for at the lower level of calculation. Thus, the MIM2 total energy can be written as

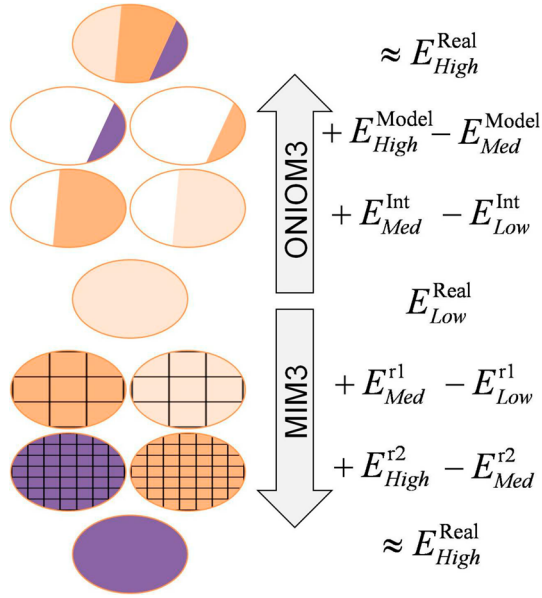


Figure 8. Schematic representation of the MIM methodology. The blue, orange, and pink regions in the different layers represent low, medium, and high levels of theory.[113] Reprinted with permission from ref 113. Copyright 2011 American Chemical Society.

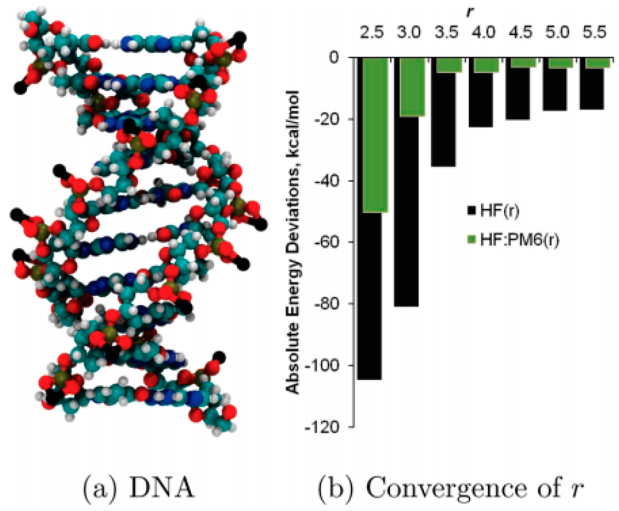


Figure 9. (a) DNA decamer used in the MIM study. (b) Convergence of absolute energy with distance based cutoff (r). Reprinted with permission from ref 113. Copyright 2011 American Chemical Society.

$$E^{\text{MIM2}} = E^{R}_{\text{low}} - E^{r}_{\text{low}} + E^{r}_{\text{high}} \quad (28)$$

where r stands for the threshold parameter used in the first layer and R represents a larger threshold parameter used in the second layer. These are very similar to the energy expression used in the standard ONIOM method where R represents the real system and M stands for the small model system.

$$E^{\text{ONIOM}} = E^{\text{low}}(R) + E^{\text{high}}(M) - E^{\text{low}}(M) \quad (29)$$

In a similar manner, extrapolation of this approximation can be done up to any arbitrary number of theories depending on the complexity of the system and the desired level of accuracy. Following this, the energy expression for MIM3 can be written as

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

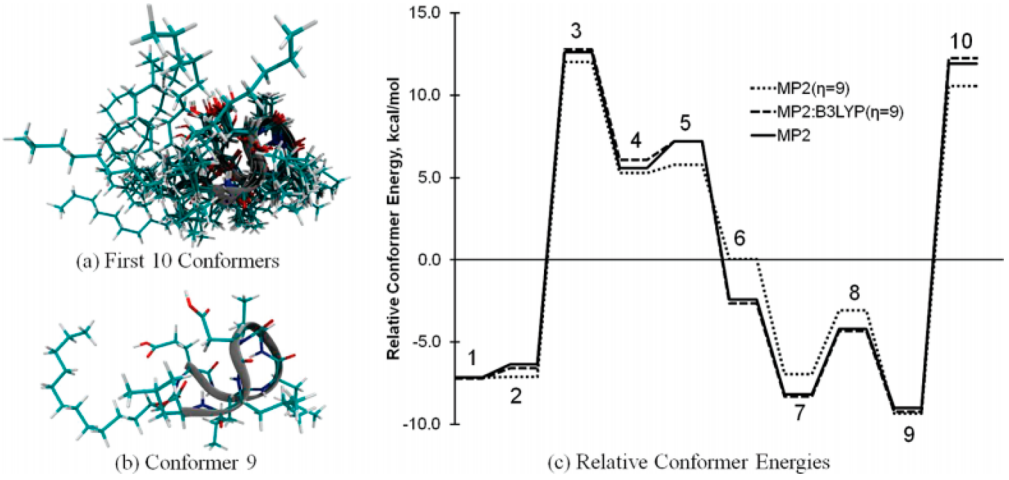


Figure 10. PDB ID 2NPV: (a) Superposition of the first 10 conformers. (b) Structure of conformer 9 given in the PDB file. (c) Relative energies for the first 10 conformers of 2NPV centered at zero. Reprinted with permission from ref 113. Copyright 2011 American Chemical Society.

$$E^{\mathrm{MIM3}} = E^{R}_{\mathrm{low}} - E^{r1}_{\mathrm{low}} + E^{r1}_{\mathrm{medium}} - E^{r2}_{\mathrm{medium}} + E^{r2}_{\mathrm{high}} \quad (30)$$

Here R stands for real system; and $r1$ and $r2$ are the fragmentation parameters ($r1 > r2$); and low, medium, and high indicate the level of theories. Each of the terms in the equation above is the energy of the total system obtained from the summation of individual energy components at the defined fragmentation parameter. The idea of using a second layer to include longer range interactions have been adopted by several different groups. The conceptual relationship between MIM and ONIOM is illustrated in Figure 8.

MIM was assessed on some large molecules. For example, for a DNA (Figure 9) fragment (10 A−T base pairs) containing 656 atoms, a MIM2 calculation using (HF/6-31G:PM6) yielded a total energy with an error of only 2−3 kcal/mol using a distance cutoff parameter of 4−5 Å. The inclusion of long-range effects using PM6 was found to be important since errors larger than 15 kcal/mol were found for a single layer (MIM1) method. For a second example (Figure 10) involving a peptide containing 7 amino acids, the relative energies between the 10 lowest energy conformations were obtained with a mean absolute error of 1.18 kcal/mol for MIM1 and 0.25 kcal/mol for MIM2 using a number-based cutoff (η = 9). Geometry optimizations for large molecules were also demonstrated.

5.8. Fragment Molecular Orbital Method (FMO)

Since its first introduction by Kitaura and co-workers,[157] the FMO method[158] has undergone many important developments and improvements. In particular, during the past decade, persistent attempts have been made to make FMO into a black-box fragmentation method. As pointed out in a recent review article,[9] the following sequential steps are necessary for a successful implementation of FMO method: (1) For each of the monomers, the initial electron density distribution is calculated. (2) Fock operators for each of these monomers are developed using the densities from the previous step, and the energy of each monomer is calculated in the coulomb field of the rest of the system. (3) A converged electrostatic potential is obtained by iterating each of the monomer energies to self-consistency. (4) Fragment dimer calculations (FMO2) are then executed in the converged ESP of rest of the system. Dimer calculations are not performed self-consistently. (5) Depending on the computer resources in hand and the nature of the problem, optionally, fragment trimer calculations (FMO3) are performed in an analogous manner. Each trimer calculation is performed only once.

The total energy in FMO2 and FMO3 are obtained by the following expressions:

$$E^{\mathrm{FMO2}} = \sum_{I}^{N} E_I + \sum_{I>j}^{N} (E_{IJ} - E_I - E_J) \quad (31)$$

$$E^{\mathrm{FMO3}} = E^{\mathrm{FMO2}} + \sum_{I>J>K}^{N} (E_{IJK} - E_I - E_j - E_k) - (E_{IJ} - E_I - E_j) - (E_{IK} - E_I - E_k) - (E_{JK} - E_j - E_k) \quad (32)$$

using the monomer (I), dimer (IJ), and trimer (IJK) energies obtained as described above. E_I, E_{IJ}, E_{IJK} denote the energies of monomer, dimer (energy of species I and J together), and trimer (energy of species I and J and K together), respectively.

The fundamental concepts involved in the sequential steps in a typical FMO calculation are shown graphically in Figure 11.

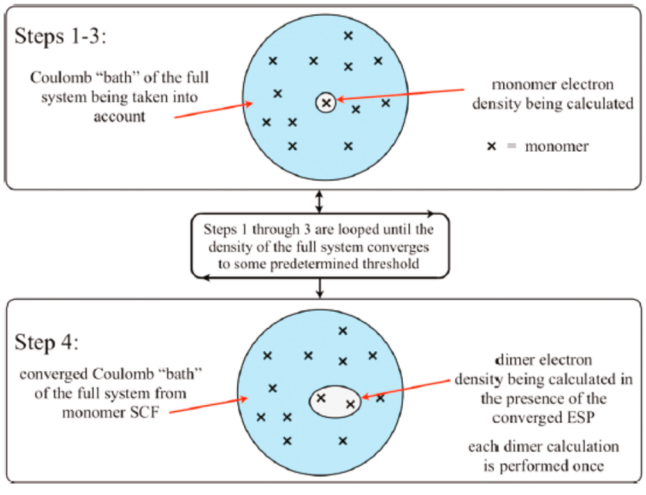


Figure 11. Graphical illustration of the sequential steps involved in the FMO method.[9] Reprinted with permission from ref 9. Copyright 2012 American Chemical Society.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

Since FMO has been reviewed extensively (being one of the earliest fragment-based methods) on numerous occasions, e.g., in a recent review article,[9] we will skip the detailed theoretical developments it has undergone during the past decade. We will briefly focus on a few of the recent impactful applications of FMO in the following paragraphs.

The applicability of FMO method has been demonstrated on a wide range of large molecular systems such as proteins. For example, Ikegami et al.[159] have performed FMO-RHF/6-31G* calculations on a photosynthetic reaction center, *Rhodopseudomonas viridis*, containing 20 581 atoms with 164 442 basis functions. The FMO calculation runtime was 72.5 h on 600 CPUs (2.0 GHz opteron). Recently, geometry optimizations have been performed with the FMO/MM method (a combination of FMO with molecular mechanics) and applied to the complex of a subunit of protein kinase 2 (CK2) with a ligand, involving a broad range of strong to weak interactions including a salt bridge; NH···N, CH···O, and CH···π hydrogen bonds; and $\pi\cdots\pi$ stacking.[160] The hybrid orbital projection operator was used for treating the boundary between FMO and MM regions. Full optimizations on 667 FMO atoms and 1980 MM atoms were performed to yield a structure in good agreement with experiment (rmsd of 0.49 Å). Evaluation of Raman spectra with FMO has been proposed recently.[161] The approach has been applied to calculate Raman and IR spectra of crambin (PDB: 1CNR) and a polystyrene oligomer. In a recent application,[162] FMO-CCSD(T)/6-31G method has been applied to treat HIV-1 protease−lopinavir complex. A dual basis approach[163] within FMO has been implemented to achieve accurate and efficient use of larger basis sets. This approach has been applied to treat water clusters and polypeptides. There are numerous other applications of the FMO method.

Parallel computing is an important technical aspect which has been implemented with many computing environments in various extensions of FMO. As mentioned earlier, Ikegami et al. have demonstrated a massively parallel FMO-HF benchmark study on a photosynthetic reaction center with 600 scalar CPUs. Another example of massive parallelization of FMO method has been illustrated on a variety of large water clusters including 128, 256, 512, 1024, 2048, and 4096 water moleules. Partial or region-limited geometry optimizations[164] based on parallelized FMO-MP2 gradient calculations have been carried out using the BlueGene/P computer (131 072 processors) on TrpCage.[165] The authors reported[166] a calculation time of ∼7 min for atomic forces for a system including 3000 atoms and 44 000 basis functions with MP2/augmented and polarized doubled-ξ basis set. Katouda et al.[167] reported efficient FMO-MP2(RI)/6-31G* calculations for the large influenza protein for a calculation time of 20 min on K-computer by using 12 288 nodes and 86 016 processor cores. Heuristic static load-balancing algorithm has also been developed within FMO method.[168]

EFMO, developed by Jensen et al., is a hybrid method of FMO (fragment molecular orbital) and EFP[169] (effective fragment potential) method. EFMO[170] method was developed to combine the advantages of FMO into EFP method to speed up the EFP runtime. It is based on the FMO fragmentation scheme and many-body energy expression. But for calculating long-range interactions, it uses EFP multipole-based energy expressions.

The total energy of the system in EFMO method is given by

$$E^{\mathrm{EFMO}} = \sum_{I} E_I^0 + \sum_{I<J}^{R_{I,J}\le R_{\mathrm{cut}}} (\Delta E_{IJ}^0 - E_{IJ}^{\mathrm{ind}}) + \sum_{I>J}^{R_{I,J}\le R_{\mathrm{cut}}} E_{IJ}^{\mathrm{es}} + E_{\mathrm{total}}^{\mathrm{ind}} \quad (33)$$

where

$$\Delta E_{IJ}^0 = E_{IJ}^0 - E_I^0 - E_J^0 \quad (34)$$

is the RHF interaction energy of the isolated dimer *IJ*, while E_{IJ} is the intermolecular induction energy of dimer *IJ* computed using induced dipoles. This term is subtracted because E_{total} contains the induction energy of all monomer pairs.

The applicability of the EFMO method has been demonstrated on challenging water clusters. Results of EFMO method on water clusters have been compared with results from FMO2 calculations as well as with full system *ab initio* calculations. It was shown that the error associated with EFMO for these chosen water cluster systems is about 0.5 kcal/mol per hydrogen bond relative to direct *ab initio* results at HF/6-31G(d) theoretical model and 0.1 kcal/mol at HF/6-31G+(d) theoretical model. It was also reported that the EFMO method is 2 times faster than the FMO2 method and 5 times faster when calculating HF gradients.

The EFP method was fully integrated into FMO to develop the fully integrated effective fragment molecular orbital method (FIEFMO)[171] which accounts for all fundamental types of both bonded and nonbonded interactions. FIEFMO has been extensively tested on several different sizes of water clusters (including 8, 16, and 32 water molecules) as well as water−methanol systems. Full *ab initio* calculation of these clusters was carried out at MP2 level using Pople-style basis sets (6-31++G(d,p) and 6-311++G(3df, 2p)). The accuracy of FIEFMO method was tested and compared with standard EFMO, FMO2, and FMO3 results (relative to full MP2 calculations). It was shown that FIEFMO provides superior average errors, relative energies, and binding energies for all these systems. There were many instances where FIEFMO outperformed the accuracy of expensive FMO3 results. In addition, FIEFMO is also capable of providing a detailed energy decomposition analysis (EDA) of a broad range of interactions in a single calculation. In terms of computational efficiency, it was reported that FIEFMO provides a significant reduction in wall-clock time (time savings up to 96%) as compared to standard FMO2 and FMO3 methods through the reduction in the numbers of explicit QM dimer calculations.

5.9. Multicentered QM:QM Method (MC QM:QM)

MC QM:QM, developed by Tschumper et al., makes it possible to extend the idea of a "model system" used in the popular ONIOM method throughout the entire molecule.[172] In ONIOM, the central idea is to treat the chemically interesting part of the molecule (model system) with an accurate (and expensive) quantum mechanical calculation while handing the rest of the molecule (less important part) with a less computationally demanding quantum mechanical method. Integrated MC QM:QM method[172] basically includes multiple fragmentations of the whole system, i.e., treating each fragment of the parent molecule as a "model system". This idea eventually allows the treatment of highly accurate QM methods on systems that involve more than one chemically interesting part. In addition, one can include the effect of surrounding

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

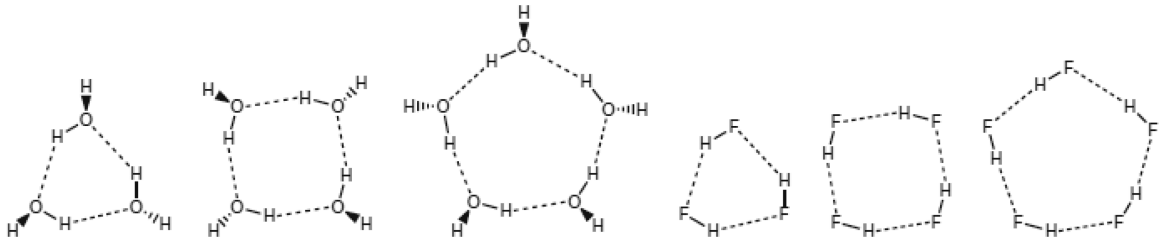


Figure 12. Model systems of $(H_2O)_n$ and $(HF)_n$ clusters (n = 3−5) used for evaluation analytic gradients with MC QM:QM method.[175] Reprinted with permission from ref 175. Copyright 2007 Taylor & Francis.

molecules on each of the model systems to achieve better accuracy.

The MC QM:QM has thus far been applied to molecular clusters bound by weak noncovalent interactions such as hydrogen bonding or dispersion interactions. A typical application of MC QM:QM method on such a weakly bound cluster system allows the estimation of the energy (or any other property) of the entire system by a single low level calculation of the whole system coupled to a series of high and low level calculations on "dimers" and "monomers" within the cluster. This is referred to as the 2-body:many-body QM:QM method. For a system with n monomer fragments, the following expression is used for the 2-body:many-body energy.

$$E_{\mathrm{Hi:Lo}}(\mathrm{mol}) = E_{\mathrm{Lo}}(\mathrm{mol}) + \sum_{i=1}^{n-1}\sum_{j>i}^{n}\{(E_{\mathrm{Hi}}(f_i f_j) - E_{\mathrm{Lo}}(f_i f_j)) - (n-2)(\sum_{i=1}^{n}\{E_{\mathrm{Hi}}(f_i) - E_{\mathrm{Lo}}(f_i)\})\} \quad (35)$$

Here the monomer fragments are denoted as f_i and the dimers as $f_i f_j$. In a similar manner, other related methods such as 3-body:many-body QM:QM can be defined where terms up to 3-body are evaluated with the high level method. The equations are general and can be universally applied to any system as long as the model systems are constructed as unique combinations of monomers within the cluster. For a more complex example involving proton transfer in a water cluster, one subset may include $H_5O_2^+$ while the rest (n−2) subsets can include remaining H_2O molecules.

The applicability 2-body:many-body integrated MC QM:QM method has been successfully demonstrated on π-type prototypes including trimers of cyanogen and diacetylene.[173] The target level of theory used for these model systems is CCSD(T)/aug-cc-pVDZ. The two levels of theory used in MC QM:QM method are CCSD(T)/aug-cc-pVDZ (for each of the dimers) and MP2/aug-cc-pVDZ (for the entire system). This scheme was successful enough to reproduce the interaction energy within an error of ±0.02 kcal/mol on seven different structures of cyanogen. While the performance of MP2 as a low level of theory was outstanding, performance of RHF as low level of theory was also remarkable producing an error of 0.03−0.15 kcal/mol. In the case of diacetylene clusters, with MC QM:QM, the error was no more than 0.04 kcal/mol with the same combination of high and low levels of theory.

In another study,[174] MC QM:QM method has been applied to water, HF, and rare gas clusters ($(H_2O)_n$, $(HF)_n$) and ($(Ne)_n$) to test the applicability of the method in addressing hydrogen bonding and van der Waals interactions. The target level of theory for these clusters was CCSD(T)/aug-cc-pVTZ. The average absolute energy error for six hydrogen bonded clusters was 0.04 kJ/mol per hydrogen bond using MP2 as the low level (0.15 kJ/mol when RHF is used as a low level). For rare gas clusters (helium and neon), with CCSDT:CCSD(T) scheme, the error was not more than 0.52 μE_h (target level of theory was CCSDT). It is clear that adding a second layer of theory can improve the accuracy significantly over a single layer fragmentation.

Evaluation of analytic gradients was also developed within MC QM:QM method and applied on weakly bound clusters [$(HF)_3$, $(HF)_4$, $(HF)_5$, $(H_2O)_4$, and $(CH_3OH)_4$ as shown in Figure 12]. Geometry optimization was carried out on 15 different hydrogen-bonded clusters.[175] As demonstrated by the authors, this 2-body:many body integrated approach not only yields extremely accurate structures and energetics but also reduces the computational cost significantly provided that the selection high level and low level methods is done carefully.

The authors also took a step further to develop 3-body:many-body integrated QM:QM method for weakly bound clusters and successfully applied it to $(H_2O)_{n=3-10,16,17}$.[176] In this application, all the 1-, 2-, and 3-body terms were calculated with CCSD(T) whereas higher-order terms (i.e., the full molecule) were calculated with MP2. The systematic application of 3-body:many-body CCSD-(T):MP2 technique on 40 low-lying $(H2O)_n$ clusters generates maximum absolute deviation (MAD) of 0.07 kcal/mol in total energy or 0.01 kcal/mol per water. For various low-lying structures of $(H_2O)_{16}$ and $(H_2O)_{17}$ clusters, the error was always within 0.13 kcal/mol relative to the CCSD(T)/aug-cc-pVTZ theoretical model. It is also important to note that 3-body:many-body CCSD(T):MP2 technique is efficient since the largest subsystem calculation at CCSD(T) level contains only 3 water molecules, reducing the computational cost significantly while still holding extremely high accuracy.

The applicability of 2-body:many-body and 3-body:many-body QM:QM techniques has also been demonstrated[177] by applying it to obtain CCSD(T) optimized geometries and vibrational frequencies for $(H_2O)_n$ (n = 3−7) and $(HF)_n$ (n = 3−6) clusters. For this particular application, all the 1-body through 2-body (or 3-body) interactions were obtained with CCSD(T) calculation whereas the higher order terms were calculated via a full calculation with MP2. As the authors reported, they were able to obtain virtually identical CCSD(T) optimized geometries for water clusters $(H_2O)_n$ (n = 3−7) by applying 2-body:many-body CCSD(T):MP2 technique. Harmonic vibrational frequencies calculated with 2-body:many-body CCSD(T):MP2 method were also in remarkable agreement with the reference calculations (unfragmented results), with the maximum and average absolute deviation being 6 and 0.8 cm^{-1}, respectively. Furthermore, the authors further reduced these deviations systematically by including more terms from the many-body expansion at the CCSD(T)

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

level. A maximum deviation of few tenth of cm^{-1} was observed when 3-body:Many-body CCSD(T):MP2 method was implemented.

The MC QM:QM method has thus far been applied only to nonbonded cluster systems with a particular focus on high accuracy rather than applicability. The direct extension of such methods on very large clusters has not been demonstrated.

5.10. Electrostatically Embedded Many-Body Method (EE-MB)

EE-MB, developed by Truhlar and co-workers,[178−184] attempts to improve the accuracy of traditional many-body based fragmentation methods. The basic idea is the use of background molecular charges to incorporate the environmental effects in each fragment calculation. In EE-MB, the system is first divided into smaller fragments (monomers). It is then followed by calculations on dimers (and trimers, if needed) of fragments in a field of point charges that represent the electrostatic potential of the rest of the fragments. Since each fragment is calculated in a bath of electrostatic point charges, this method is termed as "electrostatically embedded many-body" method.

EE-MB starts with a system containing N interacting monomers. The total energy of this system can be expressed as a traditional many-body expansion as

$$V = V_1 + V_2 + V_3 + \cdots + V_N \tag{36}$$

Here

$$V_1 = \sum_{i}^{N} E_i \tag{37}$$

$$V_2 = \sum_{i<j}^{N} (E_{ij} - E_i - E_j) \tag{38}$$

$$V_3 = \sum_{i<j<k}^{N} [(E_{ijk} - E_i - E_j - E_k) - (E_{ij} - E_i - E_j) - (E_{ik} - E_i - E_k) - (E_{jk} - E_j - E_k)] \tag{39}$$

All these terms are embedded in a coulomb of point charges. If we retain the equation up to 2-body limit, the total energy of the system will then be given by

$$E_{\text{EE-2B}} = \sum_{i<j}^{N} E_{ij} - (N-2) \sum_{i}^{N} E_i \tag{40}$$

Retaining the equation to 3-body will give rise to the following equation:

$$E_{\text{EE-3B}} = \sum_{i<j<k}^{N} E_{ijk} - (N-3) \sum_{i<j}^{N} E_{ij} + \frac{(N-2)(N-3)}{2} \sum_{i}^{N} E_i \tag{41}$$

As described by the authors, one needs to take the following steps for successful application of EE-MB method: (1) estimation of the partial atomic charges of the system containing N interacting units by a given quantum mechanical method; (2) determination of the energy of the system to include interactions up to 2- or 3-body (each of these calculations is executed in the presence of partial atomic charges of $(N - n)$ molecules determined in step 1); (3) evaluation of the energy by appropriate summation.

There are still several ways in which electrostatic embedding can be carried out. (1) The most rigorous way is via "self-consistent electron charge embedding" as in GEBF or in the FMO method. However, it involves an iterative procedure and can be expensive. (2) In a simpler approach, one can determine the partial charges on all the atoms by performing a full system calculation at a low level of theory. It is then followed by each subsystem calculation embedded in the resulting charges on the nonactive sites. (3) In the simplest approach, the gas phase monomer charges can be determined and used in each separate subsystem calculation to represent the charges on the atoms of nonactive sites.

Using the simplest monomer-based fixed charges, EE-MB has shown reliable performance.

In the initial assessment, EE-MB has been successfully applied on a database of water clusters containing a set of eight trimers, six tetramers, and a pentamer. All 15 clusters have been selected from Monte Carlo and molecular dynamics simulations as well as gas phase optimizations. An extensive combination of different methods, basis sets, and different charge types has been analyzed to test the applicability of the EE-MB method. Six different basis sets (Pople-style or Dunning-style) in combination with four different density functional methods (BLYP, PBE, PBE1W, B3LYP) as well as MP2 have been used. Using fixed charges from several methods, the EE-MB method provides a 10-fold reduction in errors as compared to the conventional pairwise additive method with a mean unsigned error of 0.21 kcal/mol. If the 3-body interaction term is included, the error is 0.05 kcal/mol.

EE-MB has also been tested on water clusters containing up to 21 water molecules (Figure 13). For water 21-mer, EE-2B had an error of 2.97 kcal/mol, whereas with EE-3B the error improved to 0.38 kcal/mol. To compare the performance of EE-MB method with other methods, the authors also provided the results of previously applied FMO method (with B3LYP/6-31+G(d) and B3LYP/6-31++G(d,p) models) on different sized water clusters (n = 16, 32). The errors are consistently lower in the EE-MB method as compared to the FMO method.

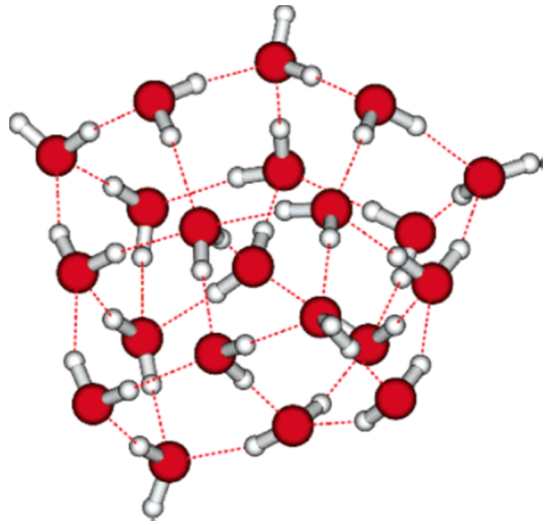


Figure 13. Geometry of 21-mer water cluster tested with EE-MB method by Truhlar et al.[179] Reprinted with permission from ref 179. Copyright 2007 American Chemical Society.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

EE-MB has recently been applied[185] to estimate the binding energy of water 26-mer with CCSD(T)-F12b/CBS theoretical model which shows a mean unsigned error of only 0.009 kcal (per mole of water molecules) for relative energies and 0.015 kcal/mol for absolute binding energies. Other applications to metal-containing systems have also been demonstrated. The EE-MB method has been applied to mixed ammonia−water[186] clusters $NH_3(H_2O)_n$ with n = 3−5. All the calculations were performed with three different DFT functionals (PBE, B3LYP, and M06-2X) and two wave function methods (MP2 and CCSD(T)). The average absolute deviation (in case of tetramers and pentamers) was 0.66 and 0.16 kcal/mol for EE-2B and EE-3B methods, respectively. EE-2B and EE-3B exhibit an average absolute deviation of 0.10 and 0.03 kcal/mol, respectively, in the case of hexamers. The applicability of the EE-MB method was also demonstrated in predicting the binding energies of mixed clusters containing water, ammonia, sulfuric acid, ammonium, and bisulfate ions.[182] The authors illustrated the consistently excellent performance of the EE-MB method in predicting the binding energies of these systems with relative energy errors of less than 1% and average relative absolute error of 0.3%. The authors also investigated and rationalized why the accuracy of the EE-MB method does not depend strongly on the selection of various charges (CheEIPG, ESP-dipole, Merz−Singh−Kollman, NBO, and CM4M) considered in the study. The EE-MB method was further extended to predict the dipole moments, partial atomic charges, and charge transfer for $NH_3(H_2O)_{11}$, $(NH_3)_2(H_2O)_{14}$, $NH_3(H_2O)_{11}$, $(HF)_4$, $(HF)_5$, $(HF)_2(H_2O)$, $(HF)_3H_2O$, $(HF)_3(H_2O)_2$, and $Cl(H_2O)_6^-$ systems.[184] In a recent application,[187] the EE-MB method was applied to predict bond dissociation energies for metal−ligand bonds in positively charged inorganic coordination complexes. This method essentially reproduced the bond-breaking energies for a series of pentacoordinate and hexacoordinate zinc-containing systems with an average absolute deviation of 0.98 kcal/mol. Further extensions of the EE-MB method (EE-MB-CE[180] and EE-MB-NE[188]) were proposed to accurately reproduce the MP2 energies of large water clusters. The applicability of EE-MB method has been illustrated through several other challenging systems.[189−192]

5.11. Kernel Energy Method (KEM)

KEM[193] was developed by Huang et al. to reduce the computational cost of high level *ab initio* methods on large systems such as peptides. The main idea of KEM lies in representing a full biological molecule as a collection of separate kernels. The original formulation and a schematic illustration of KEM is given below.

Figure 14 shows a schematic of an abstract model polymer system. The model polymer here is separated into smaller kernels. The kernels play the role of monomers in a many-body type expansion. Kernels interacting in pairs are termed as "double kernels". After representing the entire system into single and double kernels, individual electronic structure calculations are carried out on each of these kernels. Dangling bonds at the periphery of the kernels are taken care of by adding hydrogen caps at the location of the original truncated atom.

The total energy of the system in KEM is the summation of the contribution of double kernels which are reduced by single kernels to avoid overcounting of their interactions. The summation over double kernels can be restricted to those containing through-bond interactions, or all pairwise interactions can be included via through-space interactions. For the latter case when all double kernels are included, the general energy expression is

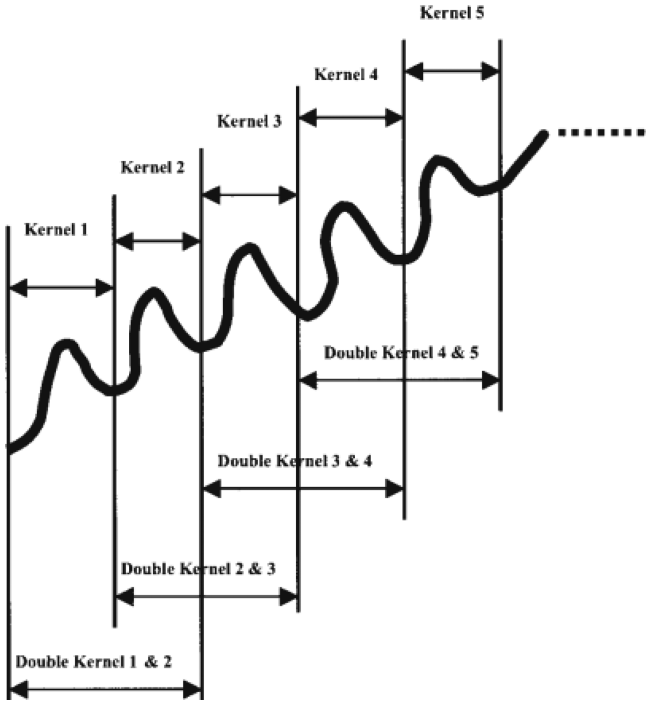


Figure 14. Illustration of kernels and double kernels on an abstract model of polymer.[200] Reprinted with permission from ref 200. Copyright 2005 American Chemical Society.

$$E_{\text{total}} = \sum_{m=1}^{n-1} \Big(\sum_{\substack{i=1 \\ j=i+m}}^{n-m} E_{ij} \Big) - (n-2) \sum_{i=1}^{n} E_i \tag{42}$$

where E_{ij} is the energy of a double kernel ij; E_i is the energy of a single kernel; i, j, m are the running indices; and n is the number of kernels. This is similar to the expressions used in other 2-body methods. The methods have later been expanded to include the three kernel and four kernel interactions.[198]

Applications of KEM have been carried out on a wide range of biological systems[194−200] such as peptides, DNA, and extended systems.[201] A recent application[202] of KEM involves the estimation of interaction energies in a RNA system called "protoribosome" containing 5759 atoms using a B3LYP/3-21G model. KEM has also been used to treat the interactions in aminoglycoside drugs and ribosomal A site RNA targets using a simple HF/STO-3G model.[203]

Huang et al.[201] have reported the treatment of a graphene system with KEM. The principal goal of this application was to show that kernels may be extracted from an extended aromatic molecule such as graphene through fissioning of aromatic bonds. The calculation of the total energy of a graphene flake was carried out at HF and MP2 models using a small 3-21G basis. The graphene flakes ($C_{78}H_{26}$) were composed of a total of 104 atoms arranged in 27 benzenoid rings. The peripheral dangling bonds were saturated with hydrogens. The KEM total energy with HF and MP2 levels was shown to deviate by −1.19 and +0.94 kcal/mol, respectively, from the directly determined total energies, suggesting that "chemical accuracy" in the total energy can be achieved for such systems.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

The formulation of KEM was redefined in a subsequent report[204] to avoid many unnecessary computations of kernels. In the new formulation of KEM, the total energy is no longer estimated as the summation of all possible double, triple, and quadruple kernel interactions. Instead, the total energy is predicted by summation of the interactions of precisely those combinations of kernels that are connected in a mathematical graph (similar to the fragmentation approach put forward by Deev and Collins[205]) that represents the fragmented molecule. In this way, the authors were able to avoid all the negligible kernel energies, particularly for those kernels which are very far apart from each other within the molecule. Applicability of this new formulation of KEM method was demonstrated on the yeast initiator molecule denoted as $ytRN_i^{Met}$ (1YFG in the Protein Data Bank, containing 2565 atoms) where the kernels were calculated with HF/STO-3G. Two basic mathematical models were proposed. The more expensive complete graph model required calculations on 171 double kernels and 19 single kernels, while the less expensive linear chain model required calculations for 18 double kernels and 19 single kernels. The authors showed that both these models were successful enough to reproduce the unfragmented result for the small basis set with reasonable accuracy.

5.12. Multilevel Fragment-Based Approach (MFBA)

In MFBA,[206] Rezac and Salahub have proposed a fragmentation scheme where the energy (and its gradient) is expressed as the sum of the energies of fragments and their pairwise interaction energies.

$$E = \sum_{i=1}^{N} E_i + \sum_{i>j}^{N} \Delta E_{ij} \tag{43}$$

Here, E is the energy of the supersystem. E_i denotes the energy of subsystem i, and ΔE_{ij} is the 2-body interaction term between i and j. While such 2-body expansions are usually carried out for nonbonded systems, MFBA was applied to bonded systems like model peptides. As in QM/MM or ONIOM methods, C−C single bonds were cut and hydrogen link atoms were used as caps. A key feature of MFBA is the implementation using multiple levels of theory within a many-body approach. Distance cutoff criteria were adopted within MFBA to execute different independent fragment calculations with different levels of theory. Fragments within a shorter distance cutoff have been calculated at higher levels of theory while the rest of the pairs were treated at lower levels of theory to capture long-range interactions. Related ideas to incorporate more than two levels of theory have also been suggested.

Gradient calculations have also been implemented within MFBA. Unlike some methods which neglect the gradients from cap atoms, gradients in MFBA are composed of all the atoms including the cap atoms. Since the cap atoms are not present in the full molecule, the gradient on the cap atom r_c is projected on the two original connecting atoms. This is done on the basis of the position of the original atoms and a scaling factor g. This correction is applied for each cap atom, and is exactly analogous to the treatment of the gradients in the popular ONIOM method.

$$r_c = r_f + g(r_o - r_f) \tag{44}$$

$$\frac{\partial E}{\partial r_f} = \frac{\partial E}{\partial r_f} + (1 - g)\frac{\partial E}{\partial r_c} \tag{45}$$

$$\frac{\partial E}{\partial r_o} = \frac{\partial E}{\partial r_o} + g\frac{\partial E}{\partial r_c} \tag{46}$$

Rezac and Salahub have demonstrated the performance of MFBA to four derivative structures of an 18-peptide with the sequence QAAAAKAFGGWISAFAAN with a net charge of +1 from the lysine (K) residue.

Four different model structures were considered to test MFBA: (1) a zwitterionic form (providing a test for strong intramolecular electrostatic interactions), (2) a capped form using acetyl and *N*-methyl caps (providing a test for important interactions between the charged lysine side chain and the other neutral, but polar, residues), (3) a neutral form obtained by deprotonation (providing a test for the attenuation of long-range interactions), and (4) a nonpolar form obtained by replacing all the peptide bonds with nonpolar trans alkene analogs (providing a test for weaker dispersion interactions).

The MFBA calculations on each form were carried out for two conformations (α-helix and an antiparallel β-sheet (Figure 15) combining two different methods: (1) DFT and (2) SCC-DFTB (using TPSS functional and a TZVP basis set). It was shown that MFBA can successfully reproduce the *relative energy* between the two conformations for these systems within an upper limit error of 2.7 kcal/mol. However, the errors in absolute energies were found to be very high. The authors have suggested that the large absolute energy errors come from the following sources: (1) missing important 3-body terms, (2) interactions between caps, (3) difference in the electronic structure of fragments and the full molecule, and (4) missing higher order terms. The dependence of the convergence of the relative energy errors with distance cutoff (fragment-fragment distance) for all four 18-peptide models was also investigated. A cutoff of 5 Å appears to be needed for adequate convergence in the calculated relative energies. In the case of geometry optimizations, a distance cutoff of 3 Å was used which actually implies that all pairs of fragments that are not connected by a covalent or hydrogen bond are calculated at the lower level or neglected. The convergence criteria used for geometry optimization are as follows: (A) energy difference in the optimization step <0.006 kcal/mol, (B) largest element of the gradient vector <1.2 kcal/mol/Å, and (C) absolute value of the gradient vector <1.2 kcal/mol/Å.

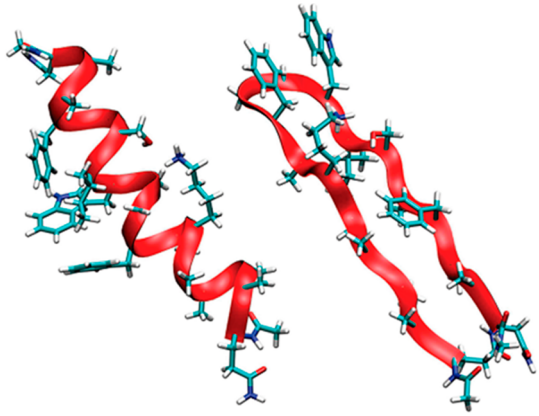


Figure 15. Model 18-peptide in α and β conformations treated using MFBA by Rezac and Salahub.[206] Reprinted with permission from ref 206. Copyright 2010 American Chemical Society.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

5.13. Hybrid Many-Body Interaction (HMBI)

HMBI[207−209] is a striking hybrid many-body fragment-based method proposed by Beran and co-workers to predict organic crystal structures and lattice energies with high chemical accuracy. HMBI uses the idea of modeling systems containing interacting molecules by treating 1-body and short-range 2-body terms with a high level of theory and the other terms with lower levels of theory (Figure 16). Since the calculations are performed on molecular crystals, no link atoms are needed. The distinguishing feature of HMBI distinct from other fragment-based methodologies is the use of classical molecular mechanics polarizable force fields to treat all the non-negligible long-range interactions.

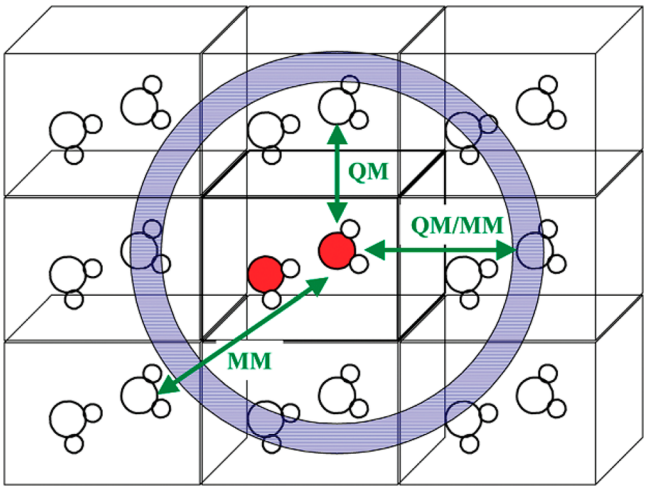


Figure 16. Illustration of the HMBI method. For each monomer in the circle, short-range interactions are treated with quantum mechanical methods, and long-range interactions are treated with molecular mechanics methods. The blue region is treated with a damping function to achieve a smooth transition between short-range and long-range regions.[209] Reprinted with permission from ref 209. Copyright 2010 American Chemical Society.

The working principle of HMBI method can be described as follows: All the central unit cell monomers and short-range 2-body interactions are treated with accurate *ab initio* quantum mechanical methods whereas all long-range 2-body interactions and many-body induction effects are treated with classical polarizable force-fields (Figure 16). For periodic systems such as molecular crystals, the HMBI energy expression[207] is given as

$$E_{\mathrm{PBC}}^{\mathrm{HMBI}} = E_{\mathrm{PBC}}^{\mathrm{MM}} + \sum_{i} (E_i^{\mathrm{QM}} - E_i^{\mathrm{MM}}) + \sum_{ij} d_{ij}^{\mathrm{smooth}} (\Delta^2 E_{ij}^{\mathrm{QM}} - \Delta^2 E_{ij}^{\mathrm{MM}}) + \frac{1}{2} \sum_{i} \sum_{\check{k}}^{\mathrm{images}} d_{i\check{k}}^{\mathrm{smooth}} (\Delta^2 E_{i\check{k}}^{\mathrm{QM}} - \Delta^2 E_{i\check{k}}^{\mathrm{MM}}) \quad (47)$$

Here, the first term is the energy of the periodic system with the force field, the second term includes the corrections to the monomers from the QM treatment in the unit cell, the third term is the correction for short-range 2-body interactions when both molecules reside in the central unit cell, and the last term is the correction when one molecule in the central unit cell and the other is a periodic image. Smooth potential energy surfaces are obtained by using a sigmoid-type damping function $d_{i\check{j}}^{\mathrm{smooth}}$ to transition from the quantum to the classical 2-body interactions over a finite-width buffer region.[207] Polarizable force fields such as Amoeba yield reasonable results while more accurate polarizable force fields can be generated on the fly from monomer *ab initio* calculations.[210,211] Correlated quantum mechanical methods such as MP2 or CCSD(T) are extrapolated to the CBS limit to achieve high accuracy for the short-range part.

The authors point out several features in HMBI that make it an effective hybrid many-body fragment-based method: (1) An inexpensive classical approach is used to replace the expensive traditional quantum mechanical treatment of long-range electrostatics. (2) Within HMBI, it is very easy to apply *nonperiodic* QM methods (MP2, CCSD(T)) to capture the important short-range interactions since long-range lattice interactions have been treated classically. (3) Analytic derivatives are much less expensive within an incremental scheme such as HMBI unlike in other point charge embedded methods (assuming the force field parameters are geometry-independent).

The applicability of periodic HMBI has been demonstrated to predict the lattice energies of five different molecular crystals: (1) hexagonal ice, (2) formamide, (3) acetamide, (4) imidazole, and (5) benzene.[209,211] The first three provide a test for hydrogen bonded systems, and the last two provide a challenge for π-stacking type dispersion interactions. HMBI successfully predicts the experimental lattice energies for all these crystals with an error of 2−4 kJ/mol. DB-RI-MP2/aug-cc-pVnZ (n = D, T, and Q) QM method is used for short-range interactions and Amoeba force-field for long-range 2-body interactions and many-body induction effects.

Another investigation[211] was carried with HMBI where the force fields are constructed on-the-fly from QM calculations on the individual molecules in the unit cell. This *ab initio* force field (AIFF) variation of HMBI reproduces the MP2 results of lattice energies for NH_3 and CO_2 crystals with errors of only 1.8 and 1.3 kJ/mol, respectively, relative to the periodic MP2 results. The strategy of AIFF parametrization within HMBI (with distributed multipoles and atom-centered polarizabilities obtained on-the-fly) has also been tested rigorously on water, formamide, hydrogen fluoride, and mixed glycine−water clusters that show strong many-body effects.[212] The authors have demonstrated that this strategy improves the accuracy of the method significantly and eliminates almost all the empiricism.

Recently, HMBI has shown remarkable success in obtaining an interesting chemical insight of molecular crystals.[213] The authors successfully performed HMBI method on crystalline aspirin (with MP2 level of theory) and probed the experimentally observed "accidental degeneracy" phenomenon (that disagreed with earlier DFT results). The authors were also able to illustrate why the two forms of aspirin (form I and form II) appear to be "virtually isoenergetic" due to the competition between favorable intramolecular relaxation (form I) and improved intermolecular hydrogen bond cooperativity effects (form II). This investigation establishes the credibility of HMBI as an effective computational tool to investigate highly complex and interesting chemical problems.

5.14. Fast Electron Correlation Methods

Hirata et al.[214] have proposed a fast, efficient, and accurate electronic structure method for clusters of weakly interacting molecules (van der Waals or hydrogen-bonded clusters) based

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

(a) (b) (c)

Figure 17. Structure of (a) zwitterionic glycine−water, (b) neutral glycine−water clusters $[(C_2NO_2H_5)(H_2O)_8]$. (c) Structure of hydrated formaldehyde $[(CH_2O)(H_2O)_8]$. Reprinted with permission from ref 214. Copyright 2005 Taylor & Francis.

on the pair-interaction method of Kitaura et al.[157] This methods includes 1-body, 2-body (and 3-body, if necessary) Coulomb, exchange and correlation energies exactly, as well as higher order Coulomb energies in the leading order of multipole expansion. The method is broadly applicable for both ground states (coupled to DFT, MBPT, and CC) and for excited states (coupled to CIS, EOM-CC, and TDDFT). The method can achieve asymptotic linear computational scaling with the cluster sizes for total energies, and constant scaling in the case of excitation energies. Relative to the conventional implementations, the method recovers total energies within 0.001%, binding energies within a few kcal/mol, and excitation energies within a few hundredths of an eV.

The working principles of the fast electron correlation method can be illustrated for the binary interaction method where the total energy of a cluster is approximated by the many-body expansion truncated after the 2-body terms, i.e.

$$\tilde{E}^{\text{binary}} = \sum_{i_1} \tilde{E}_{i_1} + \sum_{i_1<i_2} (\tilde{E}_{i_1 i_2} - \tilde{E}_{i_1} - \tilde{E}_{i_2}) \tag{48}$$

The tildes on the individual terms indicate that the higher-order Coulomb interactions are also included approximately. The individual energy terms are obtained as the eigenvalues of the Schrödinger equation

$$\tilde{H}_{i_1\ldots i_k}\tilde{\varphi}_{i_1\ldots i_k} = \tilde{E}_{i_1\ldots i_k}\tilde{\varphi}_{i_1\ldots i_k} \tag{49}$$

with the effective Hamiltonian

$$\tilde{H}_{i_1\ldots i_k} = H_{i_1\ldots i_k} + \sum_{i_m \notin \{i_1,\ldots,i_k\}} V_{i_m} \tag{50}$$

where $H_{i1\cdots ik}$ is the usual Hamiltonian of the isolated subcluster $i_1\ldots i_k$ in vacuum with eigenvalues $E_{i1\cdots ik}$

$$H_{i_1\ldots i_k}\varphi_{i_1\ldots i_k} = E_{i_1\ldots i_k}\varphi_{i_1\ldots i_k} \tag{51}$$

and V_{im} represents the electrostatic interaction between the subcluster $i_1\ldots i_k$ and subunit i_m (not belonging to the subcluster). V_{im} introduces the n-body ($n > 2$) Coulomb polarization or induction effects into $\tilde{E}_{i_1\ldots i_k}$. Perturbation theory or coupled cluster theory can be used to solve these equations. Similar extensions are also possible for truncating the energy after the inclusion of 3-body terms.

The applicability of this binary interaction model has been demonstrated by predicting the total energies of a variety of molecular clusters including water clusters, and zwitterionic and neutral glycine-water clusters (Figure 17). This method has also been applied to predict the excitation energies of formaldehyde−water clusters. A very large calculation was also performed at an equation-of-motion coupled-cluster singles and doubles level for formaldehyde−$(H_2O)_{81}$ clusters to predict a solvatochromic shift of 1360 cm^{-1} in the lowest transition energy of formaldehyde in water.

A substantial improvement[215] in the performance of this model was later observed by introducing two key extensions. In the first extension, the dipole moments are replaced by atom centered point charges to have an accurate representation of the electrostatic potentials of the clusters.

$$V_{i_m}^{\text{ESP}}(r) = \sum_{a\in i_m} \frac{Z_a}{|r - r_a|} \tag{52}$$

Here Z_a is the partial charge at a nucleus or some arbitrarily chosen position a. In the second extension, the authors accounted for basis set superposition errors (BSSEs) by combining the Valiron−Mayer function counterpoise (VMFC) correction with their binary or ternary interaction method. Reproduction of the VMFC-corrected results within 0.1 kcal/mol was demonstrated with the BSSE-corrected ternary interaction method using atom-centered point charges. These extensions led to impressive improvements in the description of short-range electrostatic potentials for both large and small charge-separated subunits. The overall summary of the computational procedure can be presented as follows for the binary interaction method (with similar extensions for the ternary interaction method): (1) calculation of dipole moment and/or ESP charges for each subunit at the SCF level replacing the rest of the subunits by point charges on ghost atoms (no basis sets), (2) repetition of step 1 until the ESP charges and/or dipole moments of all subunits are converged within a threshold value, (3) calculation of energies of all the dimer subclusters with an appropriate electron correlation method (here all the other subunits are again replaced by point charge), (4) calculation of energies of monomers with dimer subcluster basis sets using ghost atoms (no charges). The BSSE-corrected energy expressions are evaluated as appropriate.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

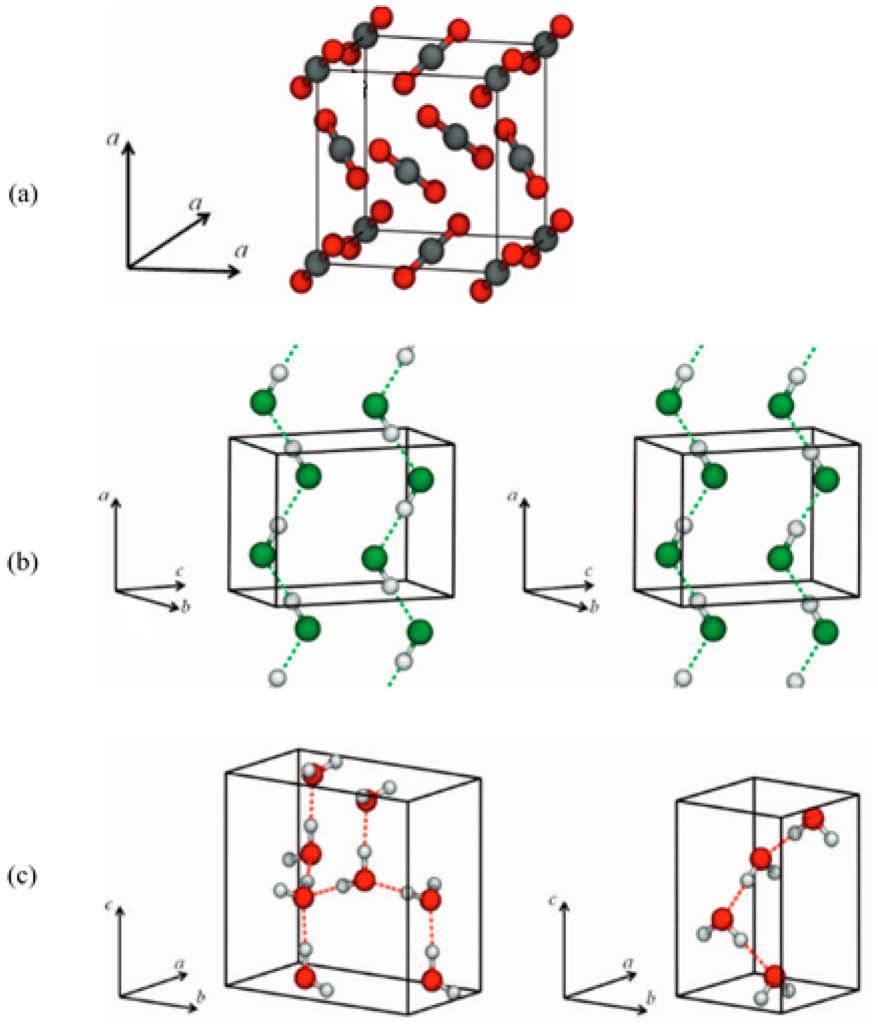


Figure 18. Structures studied by embedded many-body expansion. (a) Structure of the cubic $Pa\bar{3}$ (space group 205) CO_2 unit cell. (b) Structures of the nonpolar (left, space group 62) and polar (right, space group 36) phases of hydrogen fluoride. (c) Structures of two low energy polymorphs, ice XIh (left) and ice XIc (right). Reprinted with permission from ref 216. Copyright 2012 American Institute of Physics.

The performance of these extensions have been assessed on variety of weakly bound clusters including water hexamers, hydrogen fluoride clusters, and glycine−water clusters. The authors have illustrated an improvement by a factor of 2 in the relative energies between zwitterionic and neutral glycine−water clusters by implementing the first extension (use of self-consistent atom-centered ESP charges to describe higher order many-body coulomb interactions). This essentially rationalizes the fact that the short-range electrostatic potential of zwitterionic glycine systems is poorly described by point dipoles. Consistent outperformance by ESP-based models (as compared to corresponding dipole-based models) was also observed in case of water clusters and l-$(FH)_n$, c-$(FH)_n$ clusters. The second extension (counterpoise correction in BSSE) has been applied to remove erroneous energy ordering in the isomers of weakly bound clusters. The counterpoise-ternary + ESP model (which essentially includes explicit 3-body effects, higher-order coulomb effects by self-consistent ESP charges, and BSSE corrections via VMFC) yields relative energies of water hexamers as well as binding energies for l-$(FH)_n$, c-$(FH)_n$ clusters within an impressive accuracy within 0.1 kcal/mol.

Hirata has also extended similar high-accuracy electron correlation methods for the study of energies, structures, and phonons of molecular crystals. The energy per unit cell is given as a sum of monomer and dimer energies in an embedding field of self-consistent, polarizable, atomic charges and dipole moments. The first and second energy derivatives with respect to atomic displacements and lattice constants were also computed efficiently with a long-range electrostatic correction. The method was applied on three polymorphs (β_1, β_2, and α) of solid formic acid (modeled as infinite one-dimensional hydrogen bonded chains) with correlated levels of theory. The stability of the three polymorphs was found to be $\beta_1 > \beta_2 > \alpha$, though the energy difference between the two β forms was small. The observed infrared and Raman spectra of solid formic acid were unambiguously assigned to the calculated normal modes of the β_1 form. The results were found to be very different from previous proposed models on these polymorphs.

5.15. Embedded Many-Body Expansion

Manby et al.[216] have proposed an embedded many-body expansion model[217,218] for reliable prediction of molecular crystal energies (Figure 18) and other properties. The many-body expansion is truncated at the 2-body level (to make it cost-effective) while incorporating the remaining effects through embedding. These methods provide a framework to systematically improve the theories for molecular crystals by increasing the accuracy of the representation of the crystal

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

environment in the embedding potential. In particular, the high accuracy of these methods (with respect to the full periodic MP2 calculations) is obtained by embedding the monomer and dimer calculations in a suitable model of the crystalline environment which includes the two dominant effects: electrostatics and exchange repulsion to capture the important nonadditive terms in the energy. A significant advantage of these methods is the potential extension to coupled-cluster and explicitly correlated F12 methods. The aim of these methods is to reach the accuracy level of a reference molecular electronic structure method (such as CCSD(T)) not by including costly 3-body or 4-body effects, but by improving the quality of the embedding potential. They have also accounted for the effects basis set superposition errors that may be important in such systems.

The applicability of these methods has been demonstrated on several different challenging cluster systems such as carbon dioxide, hydrogen fluoride clusters, and ice XIh and XIc. Implementation of this method on carbon dioxide clusters achieves a cohesive energy of −11.6 mEh which lies within the error bars of the experimental value of −10.6 ± 1.5 mEh, corrected for temperature and zero-point effects. Structural optimizations of molecular clusters have also been accomplished with this model. The optimized lattice parameter of CO_2 clusters are in good agreement with experimental results, being less than 2% shorter while a slight overestimation is observed in the case of bond lengths. Three possible forms of hydrogen fluoride clusters (polar phase, nonpolar phase, and a disordered phase) have been investigated, all having zigzag chains of fluoride atoms. The results on all three phases are in good agreement with earlier benchmark calculations. Ice polymorphs are another challenging systems due to their uncertain nature of boundaries in the phase diagrams. This model addresses the common problem of energy differences between different ice polymorphs, which vary widely in sign and magnitude with the choice of the density functional used. Implementation of this model to ice polymorphs indicates that the ice XIh structure is slightly more stable than XIc by 0.3 kJ/mol. Predictions of structural parameters for ice XIh are within 1% of experimental results while the cohesive energy lies within the error of the zero-point-energy-corrected experimental value.

5.16. Many-Overlapping Body Expansion (MOB or MOBE)

The total energy of a system in the traditional many-body expansion can be expressed as

$$E = \sum_i E_i + \sum_{i>j} \Delta E_{ij} + \sum_{i>j>k} \Delta E_{ijk} + \cdots \quad (53)$$

where

$$\Delta E_{ij} = E_{ij} - (E_i + E_j) \quad (54)$$

$$\Delta E_{ijk} = E_{ijk} - (\Delta E_{ij} + \Delta E_{ik} + \Delta E_{jk}) - (E_i + E_j + E_k) \quad (55)$$

Such an expansion implicitly assumes that the "monomers" do not overlap. If the monomers are allowed to overlap, as seen in many methods described earlier, straightforward application of these terms above would clearly overcount some of the energies of atoms and their interactions. However, the many-body interaction terms can be generalized to take this into account. In the many-overlapping body expansion method[219] (denoted as MOB or MOBE), Mayhall and Raghavachari[219] have suggested that the 2-body and 3-body terms can be redefined as

$$\Delta E_{pq} = E_{pq} - (E_p + E_q - E_{p\cap q}) \quad (56)$$

$$\Delta E_{pqr} = E_{pqr} - (\Delta E_{pq} + \Delta E_{pr} + \Delta E_{qr}) - (E_p + E_q + E_r - E_{p\cap q} - E_{p\cap r} - E_{q\cap r}) - E_{p\cap qqr} \quad (57)$$

Using these redefined many-body terms, they have developed a general procedure for covalently bonded systems as follows (Figure 19): (1) Mayhall and Raghavachari[219] generated the overlapping monomers using ideas from the connectivity-based hierarchy procedure developed by Ramabhadran and Raghavachari.[85,86] Each atom (or bond) along with its nearby atoms as defined by the connectivity in the molecule (up to a depth defined by the level of the hierarchy) generates a primary subsystem, with the restriction that only single bonds are excised. This well-defined procedure yields a "monomer degree" (β) that is closely related to the "level" used in the SMF procedure of Collins and co-workers.[147] However, other definitions of the overlapping monomers are also possible (such as distance-based). (2) After the primary subsystems are generated, the inclusion−exclusion principle is used to determine all the overlap-canceling derivative subsystems. All the primary and derivative subsystems together (with their appropriate coefficients) are collectively known as "monomers". (3) The 2-body interactions are now evaluated to include all the pairwise interactions between the "monomers". While 3-body and higher-body interactions can be evaluated in a similar manner, only 2-body interactions were included in the initial paper. The order of the many-body expansion is denoted as α. (4) Hydrogen link atoms are used to cap severed single bonds. If double bonds or other types of higher order bonds are encountered, the subsystem size is increased until the nearest

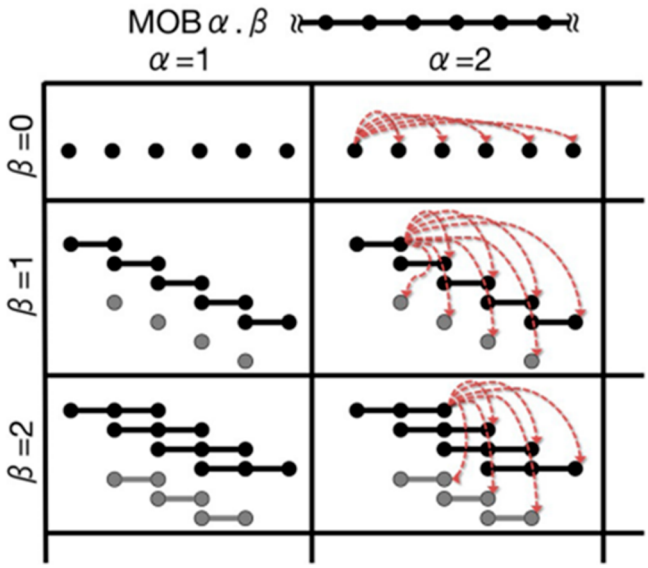


Figure 19. Schematic representation of the MOB methodology. Black monomers represent "positive" monomers. Gray monomers represent "negative" monomers. Red dashed lines denote interaction terms. Reprinted with permission from ref 219. Copyright 2012 American Chemical Society.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

single bond is found. Additional expansion procedures are used to avoid situations where the same center is replaced twice in a subsystem.

Overall, the use of overlapping monomers (denoted by the indices p, q, r, ...) requires a generalization of the many-body expansion expression:

$$E = \sum_{p}^{N} C_p E_p + \sum_{p<q}^{N} C_p C_q \Delta E_{pq} + \sum_{p<q<r}^{N} C_p C_q C_r \Delta E_{pqr} \quad (58)$$

Here C_p is the monomer coefficient obtained from the inclusion−exclusion principle for the pth subsystem. However, since the same intersection may arise from different overlaps, the coefficients may differ from +1 or −1. The MOB procedure is defined as MOB$\alpha.\beta$, where α is the order of the many-body expansion and β is the monomer degree, as defined above. In the absence of monomer overlap, these expressions reduce to the conventional nonoverlapping many-body interaction energies.

The MOB expansion has been used to evaluate the absolute energies of 4 dendritic isomers of $C_{29}H_{60}$ (Figure 20). While

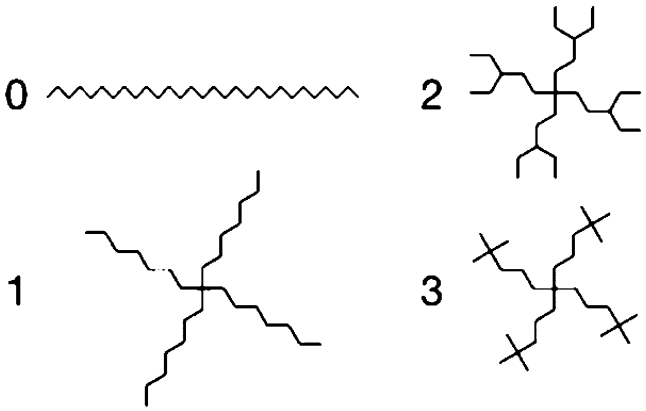


Figure 20. Dendrimers: all four $C_{29}H_{60}$ isomers used in the MOB study are shown. Reprinted with permission from ref 219. Copyright 2012 American Chemical Society.

the errors in the total energy are large when the monomers do not overlap, they are reduced significantly when they begin to overlap.[219] Using a 2-body expansion, the mean absolute deviations in the total energies of the 4 isomers decrease systematically from 52.8 to 6.1 to 1.7 to 0.3 kcal/mol, respectively, as the monomer degree increases from 0 to 3.

In a second example to consider the relative energies of peptide conformers, the 10 lowest energy conformations of a small peptide, 2NPV (163 atoms), were considered.[219] With the progressive increase in the monomer degree from 0 to 3, decreasing RMS deviations in the *total energies* (of the 10 conformations) of 46.01, 12.73, 1.37, and 0.64 kcal/mol, respectively, were obtained (using a 2-body expansion). Substantially better performance was seen in the *relative energies* between the conformations (deviations of 3.35, 1.39, 1.28, and 0.59 kcal/mol, respectively), suggesting some systematic deviations in the total energies of the different conformations.

5.17. Generalized Many-Body Expansion (GMBE)

GMBE, developed independently in 2012 by Richard and Herbert,[220] broadly includes ideas similar to that in MOBE. Again, the authors argue convincingly that the inclusion of intersecting (i.e., overlapping) monomers can improve the convergence of a many-body expansion significantly.

The nomenclature used by Richard and Herbert[220] is slightly different. Figure 21 illustrates the basics of the fragmentation procedure adopted in GMBE. In their nomenclature, they define the entire system as a single "molecule" whether it is a large molecule or a collection of small molecules, i.e., composed of molecular clusters. The "molecule" is then split into groups (groups refer to a collection of atoms that will not be split further). The procedure of selecting the groups is shown in Figure 21a. The fragment-based method is then used to assign groups to fragments, with the possibility that the same groups can be included in multiple fragments. Thus, the method has a collection of intersecting fragments, referred to as "monomers". Thus, "groups" in GMBE correspond to "fragments" in MOBE, and "fragments" in GMBE correspond to "primary subsystems" in MOBE.

The interactions between intersecting "monomers" are then evaluated via a 2-body expansion. Figure 21b,c shows the fragmentation and generation of "dimers" in the case of intersecting and nonintersecting fragments. All 2-body unions of the monomers are generated to form a set, and the inclusion−exclusion principle is then applied to this set to derive the overlap-corrected fragment-based total energy of the molecule. The implementation, as elegantly illustrated in Figure 21, is general and encompasses both intersecting and nonintersecting fragments.

The GMBE has been applied to small and larger water clusters (up to 57-mer) as well as fluoride−water clusters.[220] The latter system is interesting since it includes the effect of an ion participating in solvation. The absolute energies as well as relative energies were considered carefully. Additionally, fixed charge embedding corrections, as well as self-consistent XPOL-type embedding corrections,[221] were also considered. The embedding corrections provide a means of including longer range electrostatic interactions. The GMBE offers substantial improvements in all cases.

In a later paper,[222] Richard and Herbert have performed a careful comparison of MOBE and GMBE for several examples involving cluster systems. While both methods performed quite well in most cases, they observed that, for fluoride−water clusters, GMBE is more robust with respect to the choice of fragments than in MOBE.

6. COMPARISON OF DIFFERENT FRAGMENT-BASED METHODS

While there are numerous fragment-based methods as seen above, there are many common aspects between the methods. This is not surprising since they all share the common goal of fragmenting the molecule, performing the calculations on subsystems, and assembling the results. Thus, while the details differ, they are related to each other since there are only so many distinct ways in which fragmenting and reassembly can be done.

There have been several attempts to classify the fragment-based methods into different classes so that the similarities and differences between them can be examined. Different authors have suggested different classifications. Li et al.[138] have suggested that fragment-based methods can be classified as *energy-based* or *density-based* approaches. While the former class extracts the energy of the molecule directly, the latter class involves a calculation on the density matrix for each fragment

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

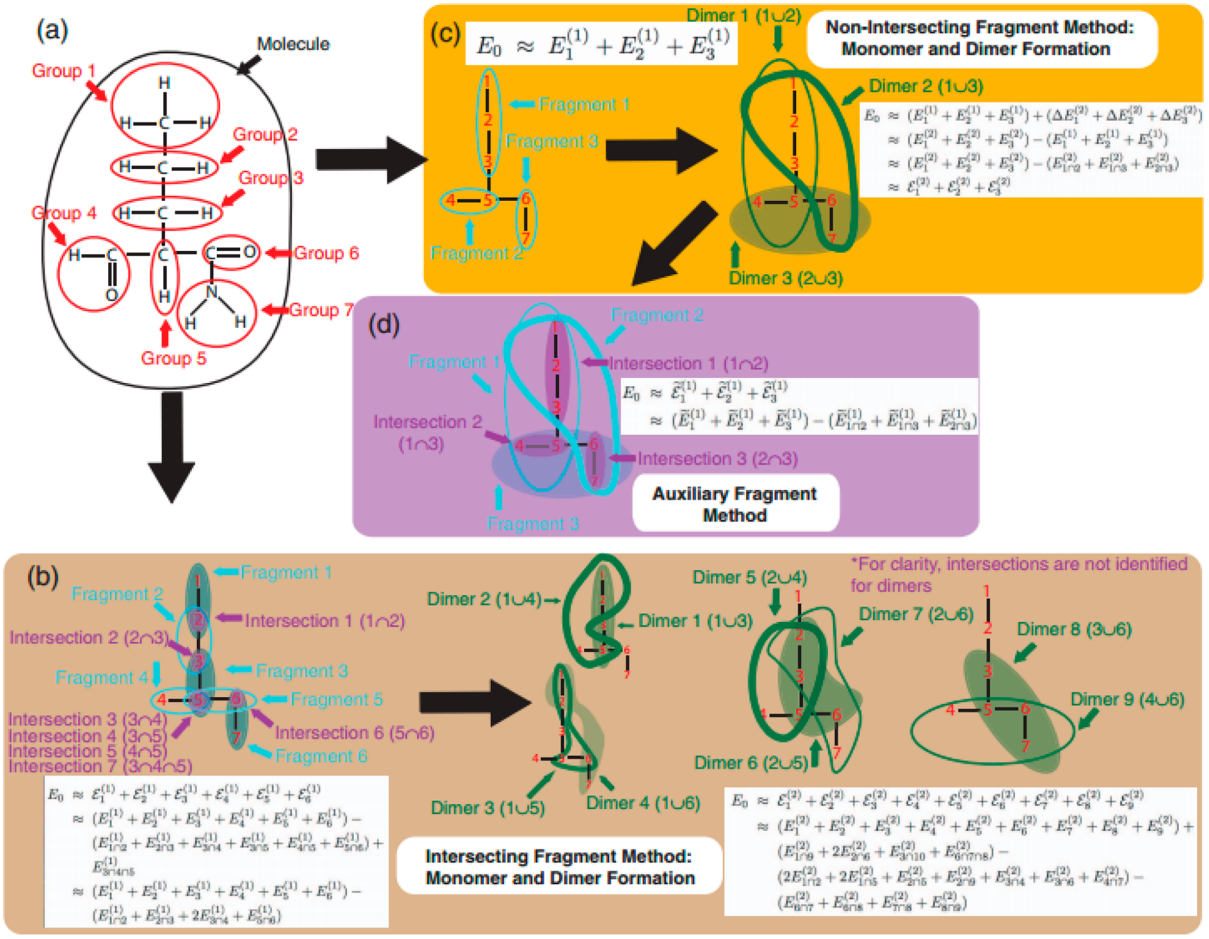


Figure 21. Illustrated example of the nomenclature used in GMBE. Panel (a) defines groups for a particular molecule. Panel (b) illustrates fragmentation of the molecule and dimer formation, using fragments that are allowed to intersect. Panel (c) depicts a particular fragmentation using nonintersecting monomers. Panel (d) illustrates the use of auxiliary monomers. Reprinted with permission from ref 220. Copyright 2012 American Institute of Physics.

first, followed by a treatment to extract the energy of the full molecule. There are some methods such as MFCC which follow both approaches in some of their variants.[124] While density-based methods may appear to be more suitable for property evaluations, the energy-based methods can also evaluate properties effectively as energy derivatives with respect to appropriate parameters such as electric fields or nuclear displacements. Gordon et al.[9] have suggested that the energy-based methods can be described as *one-step* while the density-based methods can be described as *two-step*. However, we do not consider this classification further since almost all the methods that we have considered in this review fall in the class of energy-based (or single step) methods.

Suarez et al.[223] consider all methods as multibody expansion methods that are further classified into those involving *overlapping fragments* and those involving *disjoint fragments*. Mayhall and Raghavachari[219] have suggested that two distinct classes can be identified that differ in the manner in which the fragments are constructed. Broadly, fragmentation methods have been classified as either a *top-down* method, or a *bottom-up* method. Equivalently, they are classified as *inclusion−exclusion principle-based* methods or *many-body-based* methods. Richard and Herbert[220] have classified the different methods in terms of fragments composed of *intersecting nuclei* or *disjoint nuclei*. While the names may seem different from each other, the different classifications are quite similar to each other. The terms, "overlapping fragment methods", "top-down methods", "inclusion−exclusion principle-based methods", and "intersecting fragment methods" are used in a similar context. Similarly, the terms "non-overlapping fragment methods", "bottom-up methods", "many-body based methods", and "disjoint fragment methods" are used in a similar context. With this understanding, we will broadly follow the terms used by Mayhall and Raghavachari.[219]

Apart from the generation of the fragments, there are several additional aspects of fragment-based methods that are also important. Richard and Herbert have suggested that four elements are sufficient to specify different fragment-based methods: (1) the generation of fragments, (2) the capping method, (3) embedding method, (4) number of layers.

While the generation of fragments has already been considered in the previous paragraphs, the other elements have to be specified to completely define the different methods. The capping method is important for techniques that deal with bonded systems where covalent bonds have to be broken during the generation of the fragments. The embedding method needs to be specified for many techniques that attempt to include the long-range electrostatic effects of distant fragments.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

The number of layers included in a method is another important concept. While it has been implicitly used in previous methods, it was introduced explicitly in the MIM method by Mayhall and Raghavachari,[113] in a manner analogous in spirit to the layers used in the ONIOM method.[108] Different layers can be specified by fragmenting the molecule into small fragments, larger fragments, etc. Each layer can have its own theoretical method, capping, and embedding. The *entire molecule* can be treated as a single fragment at a low level of theory, and can be considered as a layer. Many authors use this approach to include all relevant interactions without truncation at a modest level of theory. The important point to note here is that the concept of multiple layers is very general, and additional layers can be used to improve the performance of *any* fragment-based method.

6.1. Top-Down Methods

Many methods take a top-down approach by first creating *overlapping* subsystems, each containing the necessary local interactions of a given fragment (primary subsystems). The inclusion−exclusion principle can then be used to obtain the remaining subsystems (derivative subsystems) required for appropriately canceling out the overcounted units.

$$|A_1 \ldots \cup A_n| = \sum_{i=1}^{n} |A_i| - \sum_{i>j}^{n} |A_i \cap A_j| + \sum_{i>j>k}^{N} |A_i \cap A_j \cap A_k| + \ldots (-1)^{n-1} \sum |A_i \cap \ldots \cap A_n| \quad (59)$$

While the inclusion−exclusion principle is explicitly mentioned in some top-down methods (e.g., MTA,[104] GEBF,[143] and MIM[113]), it is used implicitly in several other methods (MFCC,[124] SMF,[147] CFM[147]). The accuracy of such methods is primarily determined by the size of the primary subsystems (e.g., the distance cutoff parameter).

However, the details are different in the different methods. The MTA has traditionally been applied mostly for nonbonded systems, though bonded systems have also been considered recently.[104] It uses a distance based criteria starting from each of the monomers to determine the overlapping fragments. In addition, it uses an *R*-goodness criterion to control the sizes of the fragments. Since most applications involved nonbonded systems, capping has not been an issue with MTA. While electrostatic embedding is not used, a second layer has been used in recent applications via the "grafting" approach. The EE-MTA[119] uses a difference approach via electrostatic embedding to include longer range interactions.

The MFCC[124] does not explicitly refer to the inclusion−exclusion principle. Here, the molecule (typically a peptide) is cut into individual fragments by cutting the peptide bonds. The overlapping nature of the subsystems on which calculations are carried out comes from the pair of conjugate caps that are added at each cut. The "concaps" provide the environment of each fragment and also result in overlapping subsystems. However, the overlap is based on a bonding criterion along the backbone of the peptide sequence. This is different than the distance criterion used in MTA[104] and GEBF.[143] The inclusion−exclusion principle is not invoked explicitly to take care of the overcounting. However, since the overlap occurs along the backbone of the peptide, the overcounted part is taken care of by subtracting the molecule formed from directly bonding a pair of conjugate caps. This is equivalent to including the first term of the inclusion−exclusion principle if it is assumed that 3 or more fragments do not overlap. Hydrogen capping is used to truncate the outside of the conjugate caps.

The GEBF[143] uses a distance-based criterion to determine the overlapping fragments, similar to MTA. Hydrogen capping is done in the style of ONIOM.[110] It used an embedding method for calculating each fragment in the electrostatic field of the other fragments (charges determined by a natural population analysis). In some variations of GEBF,[143] the embedding charges are determined self-consistently using a "dual-SCF" procedure. Only a single layer is used in GEBF.

The SMF and CFM methods also involve overlapping fragments.[147] The SMF generates fragments using a set of hard and soft guidelines, typically by expanding along the bonded network. As pointed out by the authors,[92] the resulting procedure is similar to methods used in theoretical thermochemistry. The fragments can also be defined by the user. The overlapping nature of the fragments is not determined explicitly via the inclusion−exclusion principle. They are canceled by using a set of rules that are somewhat complicated. Again, since the fragmentation occurs via the bonding backbone, the procedure for canceling overlapping fragments is somewhat simpler. The CFM tends to use larger fragments that are user defined.[147] It follows a more rigorous procedure for canceling the overlapping regions though the inclusion−exclusion principle is not directly invoked.

The MIM[113] defines fragments by cutting single bonds. If a distance criterion is used, the resulting subsystems are very similar to that defined in GEBF[143] (within 1 layer). However, other criteria are also used within the MIM framework. In particular, connectivity-based criteria (defined by coupling fragments separated by a given number of bonds) are also possible (closer to the approach used in SMF).[147] In addition, a number-based criteria that appends a given number of nearby fragments is also used. Such an arrangement is meant to match the sizes of the resulting subsystems, and similar methods have been used previously in GEBF.[143] Once the primary subsystems are defined, the derivative subsystems are determined by explicit consideration of the inclusion−exclusion principle. Hydrogen capping is used as in the ONIOM method. While no electrostatic embedding is used, long-range effects are included by using multiple layers.

6.2. Bottom-Up (or Many-Body) Methods

Many-body based methods take a bottom-up approach, by first obtaining small, *nonoverlapping* (disjoint) "monomers", and then combining them to form larger subsystems which contain the desired interactions.

$$E = \sum_{i=1}^{N} E_i + \sum_{i>j}^{N} \Delta E_{ij} + \sum_{i>j>k}^{N} \Delta E_{ijk} + \cdots \quad (60)$$

The interaction terms are obtained using supersystem calculations.

$$\Delta E_{ij} = E_{ij} - (E_i + E_j) \quad (61)$$

$$\Delta E_{ijk} = E_{ijk} - (\Delta E_{ij} + \Delta E_{ik} + \Delta E_{jk}) - (E_i + E_j + E_k) \quad (62)$$

Here, E_i is the energy of monomer i, and E_{ij} is the energy of combined subsystem $i + j$, etc. It is generally assumed that the monomers do not overlap. Typically, this is truncated at 2-body

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

(or 3-body) interactions. A second approximation involving a distance-based criterion is frequently used to avoid calculating those interaction terms which are spatially well- separated and are likely to be very small. The accuracy of many-body methods is determined by two parameters: the number of interacting monomers (e.g., 2-body vs 3-body) and the distance cutoff (if used).

The most widely applied fragment-based many-body method is FMO.[160] It has been applied in several approximations, 2-body (FMO2), 3-body (FMO3), and even FMO4. Unlike the other methods mentioned thus far, it uses a capping based on localized orbitals obtained from monomers or other related species. The embedding of each fragment is done self-consistently in the field of the electron densities of other fragments, though there are many variations that have been suggested. Only a single layer is used.

The electrostatically embedded many-body (EE-MB) method[178,180,181] includes 1- and 2-body (and 3-body, if needed) terms embedded in the electrostatic field (determined from point charges) of the other fragments. Typically, self-consistent embedding is not carried out, but the authors have indicated that fixed charges determined from the monomers frequently perform very effectively. Since only nonbonded systems have been considered, no capping method is needed. The typical applications focus on accuracy for modest-sized systems, e.g., water 26-mers. Most of the calculations are performed using a single layer since long-range interactions are included by electrostatic embedding. In some applications, an effective second layer is included (e.g., by including a HF calculation on the whole molecule where the fragment-based MP2 energies are computed). The fast electron correlation method of Hirata[214] also uses a truncated 2-body (or 3-body) expansion including the effects of self-consistent electrostatic embedding.

The kernel energy method (KEM)[193,194] is a many-body treatment applied to bonded systems such as biomolecules (e.g., peptides) where the monomers (e.g., amino acids) are defined as kernels. Double kernels involve 2-body interactions where a standard 2-body expansion is used. Extensions to 3- and 4-kernel interactions have been suggested including approaches that restrict the summation to those interaction terms obtained via the connected bond-network.[204] The capping method uses hydrogen atoms at the same location as the truncated atom. The focus has been on applications on very large biomolecules using modest theoretical methods (HF or DFT) with relatively small basis sets. Only a single layer is included.

The hybrid many-body interaction (HMBI) method[207] differs from most of the other methods included in this review since the focus is on molecular crystals. It is a QM/MM method that uses a combination of accurate *ab initio* methods (MP2 or CCSD(T)) for including 1-body and short-range 2-body interactions, while MM methods are used for long-range interactions. No capping is necessary since no covalent bonds are broken. Since MM extends over the whole crystal, an incremental method such as HMBI is an effective 2-layer treatment where the low level of theory is MM. The embedded many-body expansion by Manby[216] also treats molecular crystals by progressively increasing the accuracy of the crystal embedding potential starting from a truncated 2-body treatment.

The multilevel fragment-based approach (MFBA)[206] is another many-body based method involving bonded systems such as biomolecules. Here the 2-body interactions are again split into short-range and long-range. The short-range interactions are computed by a higher (i.e., more accurate) theoretical method while the long-range interactions are computed by a lower level of theory. Since covalent bonds are cut in such biomolecules, capping is done by using hydrogen atoms as in the ONIOM approach. Gradients are obtained by projecting the forces on the link atoms on the two connecting atoms.[206] Only a single layer is included.

6.3. Beyond the Two Classes of Fragmentation

While the different fragment-based methods have been sorted into two classes, relationships exist between them at a high level. There are other considerations that make it impossible to make an unambiguous assignment into one or the other. For example, inclusion−exclusion based methods may be considered as a full N-body expansion within some distance cutoff, determined by the size of the individual subsystems. Clearly, if the distance cutoff is large enough to incorporate the entire molecule, the molecular energy will be exact. Many-body methods also typically employ a distance cutoff to achieve linear scaling.[224] However, contrary to the inclusion−exclusion based methods, truncated many-body methods do not necessarily converge to the exact energy with increasing distance cutoff.

Just as an inclusion−exclusion based method can be thought of as an untruncated many-body method, many-body methods can also be thought of as an inclusion−exclusion based method. For example, if one chooses primary subsystems to be all pairs of monomers, and then obtains the derivative subsystems, the resulting subsystems will be the same as in a many-body method truncated at 2-body terms.

A major benefit of the top-down approach is that since the overlapping primary subsystems are assembled on the basis of local criteria to include the most important interactions, and the fragmentation naturally exhibits a linear scaling with molecule size. However, for systems with polar or charged groups, long-range interactions are critical for deriving accurate molecular energies. Many-body based methods allow the inclusion of longer range interactions without significantly increasing the subsystem size. However, the standard many-body expansion generally assumes that the monomers do not overlap, and thus the two approaches are not compatible.

Two groups have presented new fragmentation schemes that attempt to combine the most advantageous characteristics of both the inclusion−exclusion based methods and the many-body methods.[219,220] Recognizing the ability of inclusion−exclusion based methods to reproduce the absolute energy of large molecules, it is taken as a starting point to do a many-body expansion on top of this. In other words, the many-body energy expression is *generalized* to take into account the overlapping nature of the "monomers". It seems reasonable to expect accelerated convergence of the many-body expansion when the self-energies (or monomer energies) form a larger percentage of the overall binding energy, as this would require less of the higher order terms.

The new methods MOBE[219] and GMBE[220,225,226] combine the advantages of both methods. They carry out a many-body expansion on top of overlapping subsystems. Both sets of authors have argued convincingly and demonstrate that faster convergence of the many-body expansion can be obtained if some of the key interactions are already included in the 1-body self-energy of the system. While the ideas are similar, the implementation is somewhat different.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

Mayhall and Raghavachari[219] carry out the inclusion−exclusion principle on overlapping primary subsystems (obtained via the bonded network via the connectivity-based hierarchy) to obtain derivative subsystems to cancel the overcounting. The primary and derivative subsystems are collectively referred to as monomers. The 2-body interactions between all the monomers are then computed with appropriate coefficients. Capping hydrogen atoms are used using the ONIOM method. A second layer can be included to make further improvements via the MIM2 treatment using a lower level of theory.

Richard and Herbert[220] obtain the fragments representing the primary subsystems first. The 2-body interactions are then included to collect all the dimers. Inclusion−exclusion principle is then applied to eliminate the overcounted terms. Capping atoms are not used since the method has focused thus far on nonbonded systems. Electrostatic embedding can be included by using fixed monomer-based charges, though self-consistent charge embedding has also been considered by the authors.

There are some significant differences between the GMBE and MOBE. The most significant difference is that, in MOBE, the inclusion−exclusion principle is first carried out on the overlapping primary subsystems to generate the set of derivative subsystems that are needed to avoid the overcounting. It is quite efficient to generate all the subsystems since the exponentially scaling inclusion−exclusion principle is only applied to relatively small systems. The primary and derivative subsystems (including the appropriate coefficients) are together defined as "monomers". The 2-body interactions are then computed for the set of monomers. However, in GMBE, the inclusion−exclusion principle is applied to the collection of interacting 2-body (or *n*-body) systems. Since they are significantly larger than the monomers, this can be computationally expensive to use the exponentially scaling inclusion−exclusion principle to generate all the fragments to perform the quantum-chemical calculations. However, Richard and Herbert have shown in a recent paper[222] that the GMBE method is more robust with respect to the choice of fragments.

Finally, we note the relationship between methods such as connectivity-based hierarchy (CBH)[85,86] used in theoretical thermochemistry and fragment-based methods. As mentioned earlier, such methods achieve error-cancellation by using chemically based reaction schemes. The molecules in the right-hand side of the resulting thermochemical equations (*vide supra*) represent the primary subsystems of fragment-based methods, while the reactants added to the parent molecule on the left-hand side represent the derivative subsystems included to correct for the overcounting. Additionally, such methods only include the molecules created by expanding along the connectivity of the bond-network, as pointed out in the SFM method.[92] However, there are also differences since the fragment calculations in such methods are performed at their respective optimized geometries (to be used in conjunction with their experimental thermochemical data), while traditional fragment-based methods use fragments at the geometry found in the parent molecule. Nevertheless, there are clear connections between the two different fields.

The relationship between the different fragment-based methods, resulting in the classifications discussed in the review, is shown pictorially in Figure 22.

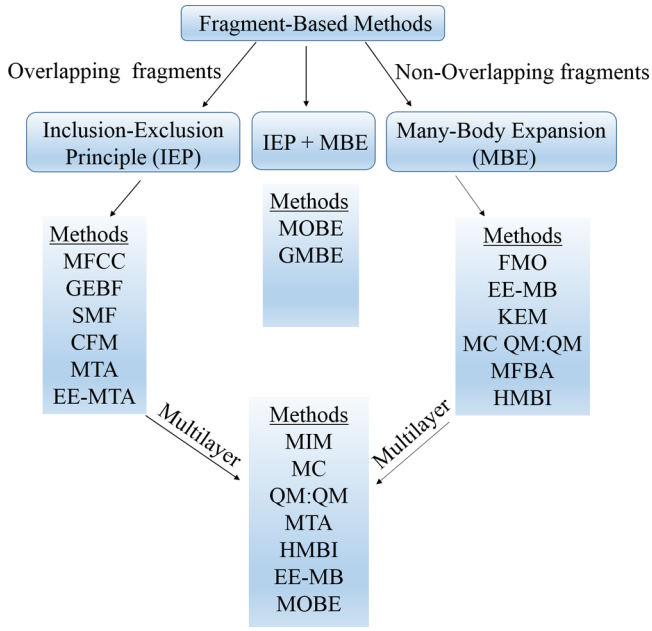


Figure 22. Classification of fragment-based methods.

7. CONCLUSIONS

A range of composite methods have been reviewed with the ultimate goal of obtaining accurate energies on large molecules. Direct calculations are possible on small molecules using extrapolated coupled cluster approaches to obtain results within chemical accuracy. Medium-sized molecules can be treated with composite models such as G*n*. Error-cancellation strategies are discussed for larger molecules using a hierarchy of chemically based ideas. Finally, a variety of fragment-based methods are discussed as important tools to remove the steep computational bottleneck for large molecules. The past decade has observed an enormous amount of activity in this field. In this review, we made a careful effort to capture the essence of all the major fragment-based methods as well as discuss their most impactful applications and interconnections. A generalized view of classification of all the major fragment-based methods has also been provided.

Despite the considerable developmental progress, the real world application of the fragment-based method is still limited due to several factors. First, many methods are still in the calibration stage where the comparison is only with a reference calculation on the supermolecule rather than with experiment. Thus, many assessments are frequently carried out with HF or DFT methods with modest basis sets. Second, the most accurate methods are still being used only on relatively small systems such as a collection of water clusters. Direct experimental information permitting calibration to larger systems is often difficult to obtain. Third, the fragmentation schemes in many methods are user-specific requiring substantial user-intervention, precluding their use as broad black-box methods. Fourth, the lack of implementation of such methods as part of standard quantum chemical packages limits their applications to the groups that develop such methods. Fifth, many important research problems such as the investigation of catalytic behavior of transition metal systems, or photochemistry involving excited states in complex systems, are not easily amenable for fragment-based methods. Finally, many important issues such as solvation effects are only included

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

approximately, precluding direct applications to complex biochemical applications where such fragment-based methods are most applicable.

Nevertheless, the number of studies using fragment-based methods in recent years to perform investigations on large molecular systems is remarkable. Many-body based methods have been mostly used in many successful applications for nonbonded systems such as molecular clusters while methods based on overlapping fragments have been shown to work well for a range of covalently bonded systems such as biomolecules. It is clear that general strategies such as the use of multiple layers[113] can be used to improve the performance of any of the fragment-based methods. The potential of fragment-based methods in understanding complex chemical and physical phenomena in large molecules, like DNA, proteins, clusters, and crystals, can be definitely seen. With the combination of new methods, algorithms, and rapid developments in high performance computing, fragment-based quantum chemistry clearly will have high impact in the next decade.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kraghava@indiana.edu.

Notes

The authors declare no competing financial interest.

Biographies

Krishnan Raghavachari, Distinguished Professor of Chemistry at Indiana University, was born and raised in Madras, India. After completing his B.S. degree in chemistry from Madras University in 1973, he joined IIT Madras to receive his M.S. in chemistry in 1975. He entered the graduate program at Carnegie Mellon University, and received his Ph.D. in 1981 under the guidance of legendary Professor John Pople (1998 Chemistry Nobel Laureate). He joined Bell Laboratories in 1981 to pursue his independent research career, and received the Distinguished Researcher Award at Bell Laboratories in 1987. After a research career of more than two decades at Bell Laboratories, he joined Indiana University as a Professor of Chemistry in 2002. He was appointed as a Distinguished Professor at Indiana University in 2014. Professor Raghavachari is perhaps best known for his work on the development and applications of electron correlation techniques in computational quantum chemistry. He developed the popular and successful CCSD(T) method, widely referred to as the "gold standard of quantum chemistry". He has published extensively on electron correlation methods, surface chemistry, theoretical thermochemistry, cluster science, and electronic structure methods for large molecules. Professor Raghavachari was elected as a Fellow of the American Physical Society in 2001, as a Fellow of the Royal Society of Chemistry in 2008, and as a member of the International Academy of Quantum Molecular Science in 2010. In 2006, he served as the chair of the Theoretical Chemistry subdivision of the American Chemical Society. In 2009, he received the Davisson-Germer Prize in Surface Physics given by the American Physical Society. He was elected as a Distinguished Alumnus of IIT Madras in 2014. He has authored over 340 scientific papers in chemistry, physics, and materials science, and his papers have been cited more than 50 000 times, placing him among the highest cited quantum chemists.

Arjun Saha received his B.S. and M.S. degrees in chemistry from North Bengal University, India. He then spent two years doing research at IIT Bombay, India, and developed an interest in Theoretical and Computational Chemistry. He entered the graduate program in Chemistry at Indiana University Bloomington, in 2011. He is currently working towards his Ph.D. under the guidance of Professor Krishnan Raghavachari. His research interests include study of the structures and catalytic reactions of small transition metal cluster systems, and development of accurate fragment-based electronic structure methods for large molecules.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from NSF Grant CHE-1266154 at Indiana University.

REFERENCES

(1) See, for example: Peterson, K. A.; Feller, D.; Dixon, D. A. *Theor. Chem. Acc.* **2012**, *131*, 1079.
(2) For a comprehensive review of coupled cluster theory, see: Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291.
(3) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
(4) Dunning, T. H., Jr. *J. Phys. Chem. A* **2000**, *104*, 9062.
(5) See for example the papers on the Gn methods by Curtiss and co-workers, the ccCA methods by Wilson and co-workers, and the CBS extrapolation methods by Petersson and co-workers (refs 44−69 in this review).
(6) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys.* **1989**, *90*, 5622.
(7) Andersson, K.; Malmqvist, P.-A.; Roos, B. O. *J. Chem. Phys.* **1992**, *96*, 1218.
(8) Malmqvist, P. A.; Pierloot, K.; Shahi, A. R. M.; Cramer, J. C.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.
(9) For a recent comprehensive review of fragment-based methods, see: Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. *Chem. Rev.* **2012**, *112*, 632.
(10) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 810.
(11) Ditchfield, R.; Hehre, W. J.; Pople, J. A.; Radom, L. *Chem. Phys. Lett.* **1970**, *5*, 13.
(12) Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 1438. Also see ref 9.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

(13) Gao, J. *J. Chem. Phys.* **1998**, *109*, 2346. Also see ref 7.
(14) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151.
(15) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.
(16) Hampel, C.; Werner, H. J. *J. Chem. Phys.* **1996**, *104*, 6286.
(17) Werner, H. J.; Schutz, M. *J. Chem. Phys.* **2011**, *135*, 144116.
(18) Neese, F.; Hansen, A.; Liakos, D. G. *J. Chem. Phys.* **2009**, *131*, 064103.
(19) Riplinger, C.; Neese, F. *J. Chem. Phys.* **2013**, *138*, 034106.
(20) Li, W.; Piecuch, P.; Gour, J. R.; Li, S. *J. Chem. Phys.* **2009**, *131*, 114109.
(21) Kobayashi, M.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 114108.
(22) Li, W.; Piecuch, P. *J. Phys. Chem. A* **2010**, *114*, 6721.
(23) Guo, Y.; Li, W.; Li, S. *J. Phys. Chem. A* **2014**, *118*, 8996.
(24) Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2011**, *135*, 144117.
(25) Shiozaki, T.; Werner, H.-J. *Mol. Phys.* **2013**, *111*, 607.
(26) Sparta, M.; Neese, F. *Chem. Soc. Rev.* **2014**, *43*, 5032.
(27) Korona, T.; Kats, D.; Schütz, M.; Adler, T. B.; Liu, Y.; Werner, H.-J. In *Linear-Scaling Techniques in Computational Chemistry and Physics*; Springer: USA; 2011, pp 345−407.
(28) Martin, J. M. L.; de Oliveira, G. *J. Chem. Phys.* **1999**, *111*, 1843.
(29) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2004**, *120*, 4129.
(30) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *J. Chem. Phys.* **2006**, *125*, 144108.
(31) Karton, A.; Daon, S.; Martin, J. M. L. *Chem. Phys. Lett.* **2011**, *510*, 165.
(32) Tajti, A.; Szalay, P. G.; Csaszar, A. G.; Kallay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vazquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599.
(33) Bomble, Y. J.; Vazquez, J.; Kallay, M.; Michauk, C.; Szalay, P. G.; Csaszar, A. G.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2006**, *125*, 064108.
(34) Harding, M. E.; Vazquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2008**, *128*, 114111.
(35) Feller, D.; Peterson, K. A.; Dixon, D. A. *J. Chem. Phys.* **2008**, *129*, 204105.
(36) Dixon, D. A.; Feller, D.; Peterson, K. A. *Annu. Rep. Comput. Chem.* **2012**, *8*, 1.
(37) Klopper, W.; Adler, T. B.; Ten-no, S.; Valeev, E. F. *Int. Rev. Phys. Chem.* **2006**, *25*, 427.
(38) Werner, H. J.; Knizia, G.; Adler, T. B.; Marchetti, O. *Z. Phys. Chem.* **2010**, *224*, 493.
(39) Feller, D.; Peterson, K. A.; Hill, J. G. *J. Chem. Phys.* **2011**, *135*, 044102.
(40) For an assessment of different extrapolation formulae, see ref 19.
(41) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104 (20 pages).
(42) Feller, D.; Peterson, K. A. *J. Chem. Phys.* **2013**, *139*, 084110.
(43) Feller, D.; Dixon, D. A. *J. Phys. Chem. A* **2000**, *104*, 3048−3056.
(44) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *126*, 084108.
(45) Curtiss, L. A.; Jones, C.; Trucks, G. W.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **1990**, *93*, 2537.
(46) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys.* **1991**, *94*, 7221.
(47) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
(48) Baboul, A.; Curtiss, L.; Redfern, P.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.
(49) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.
(50) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2007**, *127*, 124105.
(51) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(52) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107.
(53) Nyden, M. R.; Petersson, G. A. *J. Chem. Phys.* **1981**, *75*, 1843.
(54) Montgomery, J. A.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1994**, *101*, 5900.
(55) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. *J. Chem. Phys.* **1996**, *104*, 2598.
(56) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822.
(57) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532.
(58) Wood, G. P. F.; Radom, L.; Petersson, G. A.; Barnes, E. C.; Frisch, M. J.; Montgomery, J. A. *J. Chem. Phys.* **2006**, *125*, 094106.
(59) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 114104.
(60) DeYonker, N. J.; Grimes, T.; Yockel, S.; Dinescu, A.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *125*, 104111.
(61) DeYonker, N. J.; Wilson, B. R.; Pierpont, A. W.; Cundari, T. R.; Wilson, A. K. *Mol. Phys.* **2009**, *107*, 1107.
(62) DeYonker, N. J.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Theory Comput.* **2008**, *4*, 328.
(63) DeYonker, N. J.; Ho, D. S.; Wilson, A. K.; Cundari, T. R. *J. Phys. Chem. A* **2007**, *111*, 10776.
(64) Ho, D. S.; DeYonker, N. J.; Wilson, A. K.; Cundari, T. R. *J. Phys. Chem. A* **2006**, *110*, 9767.
(65) Jiang, W.; DeYonker, N. J.; Determan, J. J.; Wilson, A. K. *J. Phys. Chem. A* **2011**, *116*, 870.
(66) DeYonker, N. J.; Williams, T. G.; Imel, A. E.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2009**, *131*, 024106.
(67) Laury, M. L.; DeYonker, N. J.; Jiang, W.; Wilson, A. K. *J. Chem. Phys.* **2011**, *135*, 214103.
(68) Jiang, W.; Wilson, A. K. *Annu. Rep. Comput. Chem.* **2012**, *8*, 29.
(69) Prascher, B. P.; Lai, J. D.; Wilson, A. K. *J. Chem. Phys.* **2009**, *130*, 234104.
(70) Fast, P. L.; Corchado, J. C.; Sanchez, M.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 5129.
(71) Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 6111.
(72) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 1125.
(73) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. J. *Am. Chem. Soc.* **1970**, *92*, 4796.
(74) Raghavachari, K.; Stefanov, B. B.; Curtiss, L. A. *J. Chem. Phys.* **1997**, *106*, 6764.
(75) Raghavachari, K.; Stefanov, B. B.; Curtiss, L. A. *Mol. Phys.* **1997**, *91*, 555.
(76) Bakowies, D. *J. Chem. Phys.* **2009**, *130*, 144113.
(77) Bakowies, D. *J. Phys. Chem. A* **2009**, *113*, 11517.
(78) Bakowies, D. *J. Phys. Chem. A* **2013**, *117*, 228.
(79) Fishtik, I. *J. Phys. Chem. A* **2012**, *116*, 1854.
(80) George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M. *Theor. Chem. Acc.* **1975**, *38*, 121.
(81) For a comprehensive review on homodesmotic reactions in chemistry, see: Wheeler, S. E. *WIREs Comput. Mol. Sci.* **2012**, *2*, 204 and references therein.
(82) Hess, B. A.; Schaad, L. J. *J. Am. Chem. Soc.* **1971**, *93*, 305.
(83) Wheeler, S. E.; Houk, K. N.; Schleyer, P. v. R.; Allen, W. D. *J. Am. Chem. Soc.* **2009**, *131*, 2547.
(84) Wodrich, M. D.; Corminboeuf, C.; Wheeler, S. E. *J. Phys. Chem. A* **2012**, *116*, 3436.
(85) Ramabhadran, R. O.; Raghavachari, K. *J. Chem. Theory Comput.* **2011**, *7*, 2094.
(86) Ramabhadran, R. O.; Raghavachari, K. *J. Phys. Chem. A* **2012**, *116*, 7531.
(87) Ramabhadran, R. O.; Sengupta, A.; Raghavachari, K. *J. Phys. Chem. A* **2013**, *117*, 4973.
(88) Karton, A.; Yu, L.-J.; Kesharwani, M.; Martin, J. M. L. *Theor. Chem. Acc.* **2014**, *133*, 1483.
(89) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. *J. Phys. Chem. B* **2014**, *118*, 9631.
(90) Sengupta, A.; Raghavachari, K. *J. Chem. Theory Comput.* **2014**, *10*, 4342.
(91) Ramabhadran, R. O.; Raghavachari, K. *J. Chem. Theory Comput.* **2013**, *9*, 3986.
(92) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

(93) Bettens, R. P. A; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777.
(94) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165.
(95) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484.
(96) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
(97) Elango, M.; Subramanian, V.; Rahalkar, A. P.; Gadre, S. R.; Sathyamurthy, N. *J. Phys. Chem. A* **2008**, *112*, 7699.
(98) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2008**, *129*, 234101.
(99) Rahalkar, A. P.; Katouda, M.; Gadre, S. R.; Nagase, S. *J. Comput. Chem.* **2010**, *31*, 2405.
(100) Yeole, S. D.; Gadre, S. R. *J. Chem. Phys.* **2010**, *132*, 094102.
(101) Mahadevi, A. S.; Rahalkar, A. P.; Gadre, S. R.; Sastry, G. N. *J. Chem. Phys.* **2010**, *133*, 164308.
(102) Yeole, S. D.; Gadre, S. R. *J. Chem. Phys.* **2011**, *134*, 084111.
(103) Rusinska-Roszak, D.; Sowinski, G. *J. Chem. Inf. Model.* **2014**, *54*, 1963.
(104) Sahu, N.; Gadre, S. R. *Acc. Chem. Res.* **2014**, *47*, 2739.
(105) Furtado, J. P.; Rahalkar, A. P.; Shanker, S.; Bandyopadhyay, P.; Gadre, S. R. *J. Phys. Chem. Lett.* **2012**, *3*, 2253.
(106) Sahu, N.; Yeole, S. D.; Gadre, S. R. *J. Chem. Phys.* **2013**, *138*, 104101.
(107) Sahu, N.; Yeole, S. D.; Gadre, S. R.; Rakshit, A.; Bandyopaddhyay, P.; Miliordos, E.; Xantheas, S. S. *J. Chem. Phys.* **2014**, *141*, 164304.
(108) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.
(109) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959.
(110) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419.
(111) Vreven, T.; Mennucci, B.; da Silva, C.; Morokuma, K.; Tomasi, J. *J. Chem. Phys.* **2001**, *115*, 62.
(112) Vreven, T.; Morokuma, K. *Theor. Chem. Acc.* **2003**, *109*, 125.
(113) Mayhall, N. J.; Raghavachari, K. *J. Chem. Theory Comput.* **2011**, *7*, 1336.
(114) Yeole, S. D.; Sahu, N.; Gadre, S. R. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7718.
(115) Jose, K. V.; Gadre, S. R. *Int. J. Quantum Chem.* **2009**, *109*, 2238.
(116) Gadre, S. R.; Jose, K. V.; Rahalkar, A. P. *J. Chem. Sci.* **2010**, *122*, 47.
(117) Takeuchi, H. *J. Phys. Chem. A* **2008**, *112*, 7492.
(118) Gadre, S. R.; Ganesh, V. *J. Theor. Comput. Chem.* **2006**, *5*, 835.
(119) Isegawa, M.; Wang, B.; Truhlar, D. G. *J. Chem. Theory Comput.* **2013**, *9*, 1381.
(120) Zhang, Y. *J. Chem. Phys.* **2005**, *122*, 24114.
(121) Nasuluzov, V. A.; Ivanova, E. A.; Shor, A. M.; Vayssilov, G. N.; Birkenheuer, U.; Rösch, N. *J. Phys. Chem. B* **2003**, *107*, 2228.
(122) Wang, B.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10556.
(123) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
(124) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. *Acc. Chem. Res.* **2014**, *47*, 2748.
(125) Gao, A. M.; Zhang, Da. W.; Zhang, J. Z. H.; Zhang, Y. *Chem. Phys. Lett.* **2004**, *394*, 293.
(126) He, X.; Zhang, J. H. Z. *J. Chem. Phys.* **2005**, *122*, 031103.
(127) Chen, X. H.; Zhang, D. W.; Zhang, J. H. Z. *J. Chem. Phys.* **2004**, *120*, 839.
(128) Zhang, Da. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3659.
(129) Chen, X.; Zhang, Y.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.
(130) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.
(131) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
(132) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
(133) Xiao, H.; Zhang, Z. H. *J. Chem. Phys.* **2006**, *124*, 184703.
(134) Byun, K.; Mo, Y.; Gao, J. L. *J. Am. Chem. Soc.* **2001**, *123*, 3974.
(135) Alhambra, C.; Corchado, J.; Sanchez, M. L.; Gao, J. L.; Truhlar, D. G. *J. Am. Chem. Soc.* **2000**, *122*, 8197.
(136) Li, G. H.; Cui, Q. *J. Am. Chem. Soc.* **2003**, *125*, 15028.
(137) Wang, X.; Jinfeng, L.; Zhang, J. Z. H.; He, X. *J. Phys. Chem. A* **2013**, *117*, 7149.
(138) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193.
(139) Hua, W.; Fang, T.; Li, W.; Yu, J. G.; Li, S. *J. Phys. Chem. A* **2008**, *112*, 10864.
(140) Hua, S.; Hua, W.; Li, S. *J. Phys. Chem. A* **2010**, *114*, 8126.
(141) Hua, S.; Li, W.; Li, S. *ChemPhysChem* **2013**, *14*, 108.
(142) Wang, K.; Li, W.; Li, S. *J. Chem. Theory Comput.* **2014**, *10*, 1546.
(143) Li, S.; Li, W.; Ma, J. *Acc. Chem. Res.* **2014**, *47*, 2712.
(144) Yang, Z.; Hua, S.; Li, S. *J. Phys. Chem. A* **2010**, *114*, 9253.
(145) Li, W. *J. Chem. Phys.* **2013**, *138*, 014106.
(146) Jiang, N.; Ma, J. *J. Chem. Phys.* **2012**, *136*, 134105.
(147) Collins, M. A.; Cvitkovic, M. W.; Bettens, P. A. *Acc. Chem. Res.* **2014**, *47*, 2776.
(148) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
(149) Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 024104.
(150) Collins, M. A. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7744.
(151) Serrano, P.; Pedrini, B.; Geralt, M.; Jaudzems, K.; Mohanty, B.; Horst, R.; Herrmann, T.; Elsliger, M. A.; Wilson, I. A.; Wüthrich, K. *Acta Crystallogr.* **2010**, *F66*, 1393.
(152) Cornell, D. W.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. M.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(153) Frankcombe, T. J.; Collins, M. A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 8379.
(154) Le, H. A.; Tan, H. J.; Ouyang, J. F.; Bettens, R. P. A. *J. Chem. Theory Comput.* **2012**, *8*, 469.
(155) Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128.
(156) Tan, H. J.; Bettens, R. P. A. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7541.
(157) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.
(158) Pruitt, S. R.; Bertoni, C.; Brorsen, K. R.; Gordon, M. S. *Acc. Chem. Res.* **2014**, *47*, 2786.
(159) Ikegami, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Inadomi, Y.; Umeda, H.; Yokokawa, M.; Sekiguchi, S. *Supercomputing, 2005. Proceedings of the ACM/IEEE SC 2005 Conference*; IEEE Computer Society: Seattle, 2005.
(160) Fedorov, D. G.; Asada, N.; Nakanishi, I.; Kitaura, K. *Acc. Chem. Res.* **2014**, *47*, 2846.
(161) Nakata, H.; Fedorov, D. G.; Yokojima, S.; Kitaura, K.; Nakamura, S. *J. Chem. Theory Comput.* **2014**, *10*, 3689.
(162) Tanaka, S.; Mochizuki, Y.; Komeiji, Y.; Okiyama, Y.; Fukuzawa, K. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10310.
(163) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2014**, *597*, 99.
(164) Tsukamoto, T.; Mochizuki, Y.; Watanabe, N.; Fukuzawa, K.; Nakano, T. *Chem. Phys. Lett.* **2012**, *535*, 157.
(165) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425.
(166) Fletcher, G. D.; Fedorov, D. G.; Pruitt, S. R.; Windus, T. L.; Gordon, M. S. *J. Chem. Theory Comput.* **2012**, *8*, 75.
(167) Katouda, M.; Nakajima, T.; Nagase, S. *Proc. JSST* **2012**, 338.
(168) Alexeev, Y.; Mahajan, A.; Leyffer, S.; Fletcher, G.; Fedorov, D. G. *Proceedings of 2012 Supercomputing Conference*; IEEE Computer Society, Salt Lake City, 2012.
(169) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.
(170) Steinmann, C.; Fedorov, D. G.; Jensen, J. H. *J. Phys. Chem. A* **2010**, *114*, 8705.
(171) Pruitt, S. R.; Steinmann, C.; Jensen, J. H.; Gordon, M. S. *J. Chem. Theory Comput.* **2013**, *9*, 2235.
(172) Hopkins, B. W.; Tschumper, G. S. *J. Comput. Chem.* **2003**, *24*, 1563.
(173) Hopkins, B. W.; Tschumper, G. S. *Chem. Phys. Lett.* **2005**, *407*, 362.
(174) Tschumper, G. S. *Chem. Phys. Lett.* **2006**, *427*, 185.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677

(175) Elsohly, A. M.; Shaw, C. L.; Guice, M. E.; Smith, B. D.; Tschumper, G. S. *Mol. Phys.* **2007**, *105*, 2777.
(176) Bates, D.; Smith, J. R.; Janowski, T.; Tschumper, G. S. *J. Chem. Phys.* **2011**, *135*, 044123.
(177) Howard, C. J.; Tschumper, G. S. *J. Chem. Phys.* **2013**, *139*, 184113.
(178) Wang, B.; Yang, K. R.; Xu, X.; Isegawa, M.; Leverentz, H. R.; Thruhlar, D. G. *Acc. Chem. Res.* **2014**, *47*, 2731.
(179) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.
(180) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.
(181) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
(182) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573.
(183) Miho, I.; Bo, W.; Donald, G.; Truhlar. *J. Chem. Theory Comput.* **2013**, *9*, 1381.
(184) Leverentz, H. R.; Maerzke, K. A.; Keasler, S. J.; Siepmann, J. L.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7669.
(185) Friedrich, J.; Yu, H.; Leverentz, H. R.; Bai, P.; Siepmann, J. L.; Donald, G.; Truhlar. *J. Phys. Chem. Lett.* **2014**, *5*, 666.
(186) Sorkin, A.; Dahlke, E. E.; Thruhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.
(187) Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2012**, *8*, 1.
(188) Qi, H. W.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2013**, *117*, 4486.
(189) Hua, D.; Leverentz, H. R.; Amin, E. A.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 251.
(190) Tempkin, J. O. B.; Leverentz, H. R.; Wang, B.; Truhlar, D. G. *J. Phys. Chem. Lett.* **2011**, *2*, 2141.
(191) Kurbanov, E. K.; Leverentz, H. R.; Truhlar, D. G.; Amin, E. A. *J. Chem. Theory Comput.* **2013**, *9*, 2617.
(192) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1.
(193) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808.
(194) Huang, L.; Massa, L.; Karle, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12690.
(195) Huang, L.; Massa, L.; Karle, J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1233.
(196) Huang, L.; Massa, L.; Karle, J. *J. Chem. Theory Comput.* **2007**, *3*, 1337.
(197) Huang, L. L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 447.
(198) Huang, L.; Massa, L.; Karle, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1849.
(199) Huang, L.; Massa, L. *Int. J. Quantum Chem.* **2011**, *111*, 2180.
(200) Huang, L. L.; Massa, L.; Karle. *J. Biochem.* **2005**, *44*, 16747.
(201) Huang, L.; Bohorquez, H. J.; Matta, C. F.; Massa, L. *Int. J. Quantum Chem.* **2011**, *111*, 4150.
(202) Huang, L.; Krupkinb, M.; Bashanb, A.; Yonathb, A.; Massac, L. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 14900.
(203) Huang, L.; Massa, L.; Karle, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4261.
(204) Weiss, S. N.; Huang, L.; Massa, L. *J. Comput. Chem.* **2010**, *31*, 2889.
(205) Collins, M. A.; Deev, V. *J. Chem. Phys.* **2006**, *125*, 104104.
(206) Rezac, J.; Salahub, D. R. *J. Chem. Theory Comput.* **2010**, *6*, 91.
(207) Wen, S.; Nanda, K.; Huang, Y.; Beran, G. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7578.
(208) Nanda, K.; Beran, G. J. O. *J. Chem. Phys.* **2012**, *137*, 174106.
(209) Beran, G. J. O.; Nanda, K. *J. Phys. Chem. Lett.* **2010**, *1*, 3480.
(210) Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155.
(211) Wen, S.; Beran, G. J. O. *J. Chem. Theory Comput.* **2011**, *7*, 3733.
(212) Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155.
(213) Wen, S.; Beran, G. J. O. *Cryst. Growth Des.* **2012**, *12*, 2169.
(214) Hirata, S.; Valiev, M.; Dupuis, H.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys.* **2005**, *103*, 2255.
(215) Kamiya, M.; Hirata, S.; Manby; Valiev, M. *J. Chem. Phys.* **2008**, *128*, 074103.
(216) Bygrave, P. J.; Allan, N. L.; Manby, F. R. *J. Chem. Phys.* **2012**, *137*, 164102.
(217) Hirata, S.; Sode, O.; Keceli, M.; Shimazaki, T. In *Accurate Condensed Phase Quantum Chemistry*; Manby, F. R., Ed.; Taylor and Francis: New York, 2011; pp 129−162.
(218) Dahlke, E.; Truhlar, D. J. *Phys. Chem. B* **2006**, *110*, 10595−10601.
(219) Mayhall, N. J.; Raghavachari, K. *J. Chem. Theory Comput.* **2012**, *8*, 2669.
(220) Richard, M. R.; Herbert, J. M. *J. Chem. Phys.* **2012**, *137*, 064113.
(221) Gao, J.; Wang, Y. *J. Chem. Phys.* **2012**, *136*, 071101.
(222) Richard, R. M.; Herbert, J. M. *J. Chem. Theory Comput.* **2013**, *9*, 1408.
(223) Suarez, E.; Diaz, N.; Suarez, D. *J. Chem. Theory Comput.* **2009**, *5*, 1667.
(224) Saha, A.; Raghavachari, K. *J. Chem. Theory Comput.* **2014**, *10*, 58.
(225) Richard, M. R.; Herbert, J. M. *J. Chem. Theory Comput.* **2012**, *8*, 4381.
(226) Richard, M. R.; Lao, K. U.; Herbert, J. M. *Acc. Chem. Res.* **2014**, *47*, 2828.

DOI: 10.1021/cr500606e
Chem. Rev. 2015, 115, 5643−5677