Approaching Coupled Cluster Accuracy with Density Functional Theory using the Generalized Connectivity-Based Hierarchy

Krishnan Raghavachari, Sarah Maier, Eric M. Collins, Sibali Debnath,[†] Arkajyoti Sengupta[‡]*

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

Abstract

This perspective reviews Connectivity-Based Hierarchy (CBH), a systematic hierarchy of error-cancellation schemes developed in our group with the goal of achieving chemical accuracy using inexpensive computational techniques ("coupled cluster accuracy with DFT"). The hierarchy is a generalization of Pople's isodesmic bond separation scheme that is based only on the structure and connectivity and is applicable to any organic and biomolecule consisting of covalent bonds. It is formulated as a series of rungs involving increasing levels of error cancellation on progressively larger fragments of the parent molecule. The description of the method and our implementation are discussed briefly. Examples are given for the applications of CBH involving (1) energies of complex organic rearrangement reactions, (2) bond energies of biofuel molecules, (3) redox potentials in solution, (4) pK_a predictions in aqueous medium, and (5) theoretical thermochemistry combining CBH with machine learning. They clearly show that near-chemical accuracy (1-2 kcal/mol) is achieved for a variety of applications with DFT methods, *irrespective of the underlying density functional used*. They demonstrate conclusively that seemingly disparate results, often seen with different density functionals in many chemical applications, are due to an *accumulation of systematic errors* in the smaller local molecular fragments that can be easily corrected with higher-level calculations on those small units. This enables the method to achieve the accuracy of the high level of theory (e.g., coupled cluster) while the cost remains that of DFT. The advantages and limitations of the method are discussed along with areas of ongoing developments.

[†] Current address: Department of Chemistry, Columbia University, New York, NY 10027.

[‡] Current address: Department of Chemistry, University of California Los Angeles, Los Angeles, CA 90095.

1. Introduction

There is a constant battle between "accuracy" and "applicability" in computational quantum chemistry. While methods based on sophisticated electron correlation techniques such as coupled cluster theory,[1] e.g., the "gold-standard" CCSD(T),[2] can yield relative energies to within chemical accuracy (±1 kcal/mol),[3-9] the steep computational scaling of such methods (N^7) makes them intractable for medium to large molecules. On the other hand, widely popular, cheaper methods such as DFT (density functional theory),[10] allow studies on substantially larger systems, though often with significant errors (5 kcal/mol or more for many problems) and show disconcerting variations with choice of underlying exchange-correlation functional.[11] While developments of newer functionals in DFT have yielded better performance for some chemical applications,[10] the consistent achievement of CCSD(T)-level accuracy with DFT remains unattained. There is a critical need for methods that are chemically accurate as well as broadly applicable. This can potentially be achieved in two ways: by making accurate (coupled cluster) methods more applicable or making applicable (DFT) methods more accurate.

Various groups are in pursuit of two major strategies to make coupled-cluster methods faster and more applicable. The first is based on local-orbital treatments such as PNO- or DLPNO-based CCSD(T),[12,13,14] and the second is based on fragmentation-based methods.[15-23] Both approaches can potentially achieve asymptotic linear scaling and are in different stages of development. In this perspective, we describe an alternate strategy that we have developed where systematic error-cancellation is used to correct for the deficiencies of DFT to approach chemical accuracy.[24] An important side benefit is that our strategy renders the results largely *independent* of the underlying density functional used.[25] The overarching goal of our approach labelled "Connectivity-Based Hierarchy" can be summarized as "coupled cluster accuracy at DFT cost". In this perspective, we outline our approach and demonstrate its performance for several different chemical applications and discuss its successes as well as limitations and discuss areas for future developments.

This perspective is organized as follows. **Sections 2** and **3** discuss the method and our implementation. **Section 4** discusses different applications using this approach from our group and a demonstration of its performance in different areas of chemistry. Conclusions and outlook are discussed in **Section 5**.

2. Connectivity-Based Hierarchy (CBH)

Error-cancellation has remained an inherently intuitive and ubiquitous concept in computational quantum chemistry. Techniques in error-cancellation were particularly important before the advent of modern era computers, when accurate calculation of thermochemical data even for modest systems with just a few heavy (non-hydrogen) atoms was nearly impossible. While various ideas were used for individual systems under investigation, the first *systematic* approach to error-cancellation, the *isodesmic bond separation scheme*, was proposed by John Pople and coworkers in 1970.[26] In this widely used scheme, a select "large" molecule is "separated" into small fragments, each consisting of individual valence-satisfied (i.e., hydrogen-terminated) heavy-atom bonds, preserving formal bond types. A chemical reaction, analogous to that shown in **Figure 1**, is then created using the full molecule and its fragments.

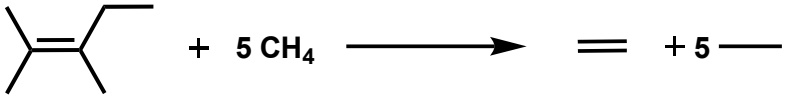


Figure 1. Isodesmic bond separation reaction scheme

The product fragments contain the individual heavy-atom bonds while the small reactant fragments represent the molecules needed to balance the reactions to account for any overcounting. The heat of such reactions may be calculated with reasonable accuracy using relatively inexpensive levels of theory (such as Hartree-Fock theory with a modest basis set in the early 1970s), since errors particular to local chemical units (heavy-atom bonds) are well-balanced in the products and reactants and are thus cancelled to a large extent. The calculated reaction energy could then be used along with the experimentally known heats of formation of the smaller fragments to derive the heat of formation of the parent molecule with reasonable accuracy. Overall, the isodesmic scheme capitalizes on fundamental ideas of error cancellation to determine corrections to low-level methods.

While the isodesmic protocol provided reasonable error-cancellation, it was clear that more sophisticated approaches would be needed for chemical accuracy. Since the early 1970s, several groups have worked to apply similar principles more generally using larger structural units to achieve better error-cancellation. One notable early contribution was the hybridization-based homodesmotic method by George and co-workers.[27] Over the years, other assorted schemes based on matching bond-types and hybridizations of larger units were brought forward to achieve better error cancellation, for example the hyperhomodesmotic, semihomodesmotic, quasihomodesmotic, and homomolecular homodesmotic methods.[28] However, many of these schemes involved explicit

tabulations of the bond types and hybridizations of all the component molecules in the reaction schemes and were formulated to be applicable mostly for hydrocarbons.

In 2011, the *Connectivity-Based Hierarchy* (CBH) protocol was developed in our group as a generalized way to bring order to the various hybridization-based error cancellation methods detailed above.[24] CBH, as its name implies, is a thermochemical hierarchy based entirely on the connectivity of the atoms in a molecule and the underlying valence bond structure (*vide infra*). The CBH scheme is systematic, well-defined, and is applicable for any organic or biomolecule without complex notations. It provides an intuitive and meaningful approach to correcting deficiencies in low-level methods, based entirely on connectivity and chemical bonding principles. A careful analysis shows that the first three rungs of the hierarchy (CBH-1, CBH-2, CBH-3) can be associated with the isodesmic, hypohomodesmotic and hyperhomodesmotic schemes considered in previous literature, though CBH reaction schemes are generated much more readily for a general, larger, organic molecule. For the sake of completeness, we also include the CBH-0 reaction scheme, often called the isogyric scheme, that separates the large molecule into the corresponding isolated single heavy-atom molecules. Overall, CBH provides a chemically sensible hierarchy of correction schemes ("rungs" of the hierarchy), employing the most basic components of molecular structure. The protocol has been used to calculate various thermochemical properties with high accuracy, including heats of formation,[29,30] bond dissociation energies,[31] acid dissociation constants (pK_a),[32] and redox potentials.[33]

At the center of CBH is a series of chemical reactions, whose successive levels include larger molecular fragments, affording better error cancellation and thus higher accuracy.[24,34] An example of CBH-0 to CBH-3 reaction schemes is shown below **(Scheme 1)** for the amino acid, methionine (**Figure 2**), containing the heteroatoms N, O and S, to illustrate the generality of the schemes.[35]

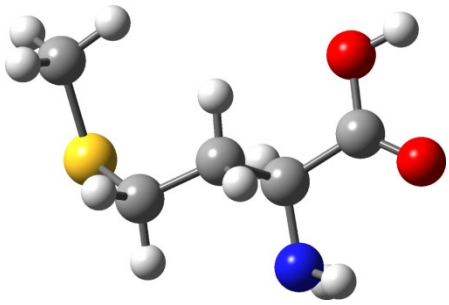
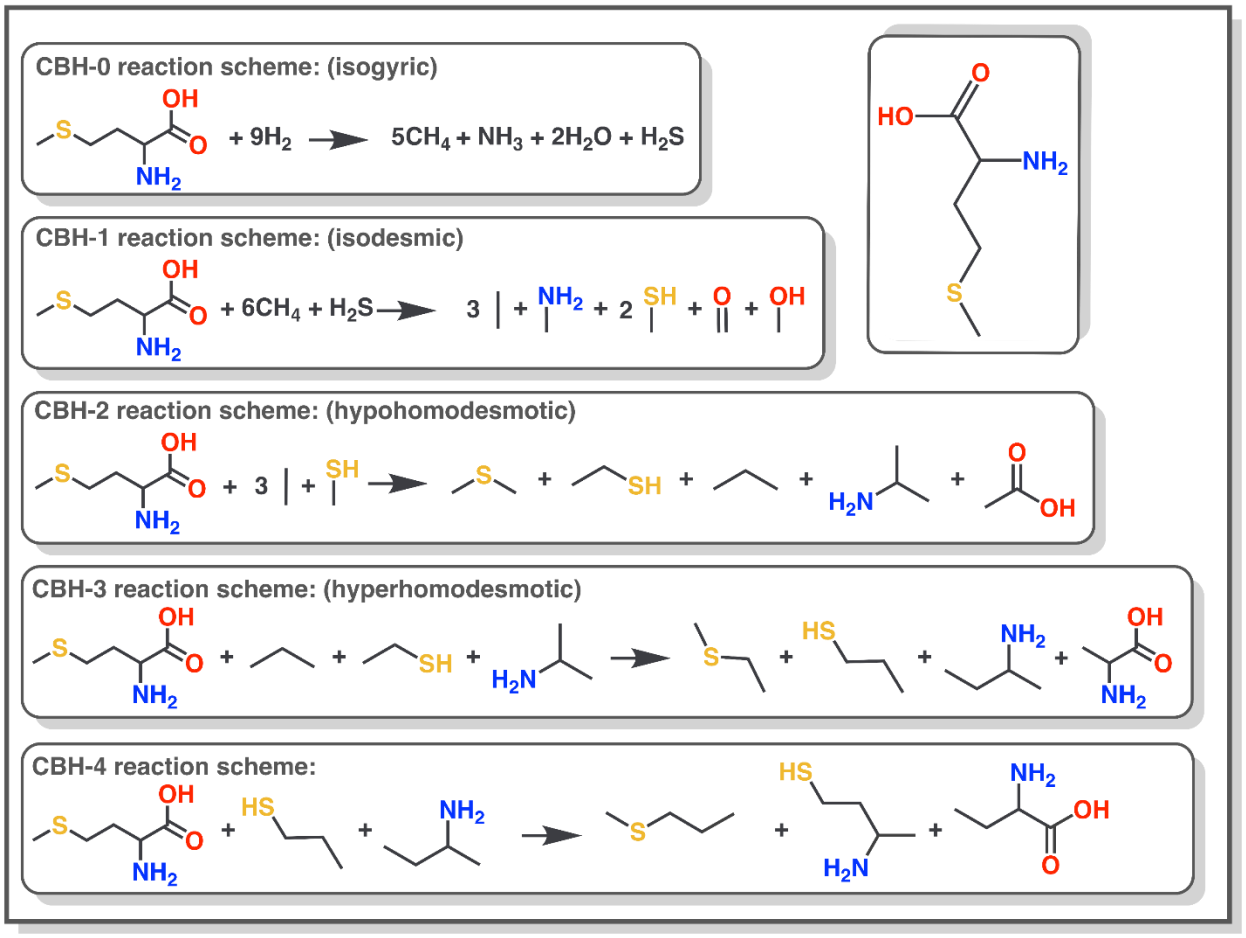


Figure 2: Ball and stick representation of methionine, used to illustrate the construction of CBH. Reproduced with permission from ref 35. Copyright 2013, American Chemical Society.



Scheme 1 – CBH reaction schemes for amino acid methionine, shown in **Figure 2**.

The following observations can be made about the reaction schemes shown above. Each reaction has a set of product fragments and a set of overlapping smaller reactant fragments to balance the reactions. Fragment size grows systematically while progressing through the hierarchy, with CBH-0 products formed by a single heavy atom, CBH-1 products formed by one heavy-atom bond, CBH-2 products formed by one heavy atom along with all heavy atoms in its immediate bonding environment, etc. Because they capture greater portions of the molecular environment, larger fragments grant better error cancellation between products and reactants. While the CBH-1 scheme is identical to Pople's isodesmic bond separation scheme, we have clearly demonstrated over the years that much better performance is achieved at the CBH-2 level (*vide infra*). Since CBH-2 maintains the environment of each heavy atom in the parent molecule, we have labeled it "isoatomic". It may also be noted that many of the products on one rung appear as reaction fragments on the next rung. Such a recursive relationship arises naturally from the systematic

nature of the growth of the fragments. Similar ideas in a different context have also been developed by Deev and Collins[36] and by Lee and Bettens.[37]

The rungs alternate being between "atom-centric" and "bond-centric". Thus CBH-1 and CBH-3 are bond-centric where the latter involves fragments for each heavy-atom-bond along with their bonded heavy atoms. CBH-2 and CBH-4 are atom-centric where the latter (not shown) involves fragments for each heavy atom along with its first and second neighbor heavy atoms. Furthermore, it is very easy to construct the hierarchy – either by hand for smaller molecules, or via an automated computer program (*vide infra*), thereby making CBH very user-friendly to accurately predict the enthalpies of formations of organic molecules. Most of our studies have been carried out using CBH-2 or CBH-3, and they appear adequate to achieve chemical accuracy.

In the original isodesmic formalism, experimental heats of formation on the smaller fragments were needed to get accurate results on the parent molecule. However, we have modified the procedure using two levels of theory (low and high) so that no experimental data are necessary to get the CBH-corrected results.[35] Instead, accurate high-level computations (typically G4) on the relatively small fragment units are used to generate a correction term to the low-level total energy (typically DFT). The CBH correction and approximate high-level energy is calculated as:

$$\mathrm{E_{high}(full)} - \mathrm{E_{low}(full)} \approx \sum_{\mathrm{i}} \mathrm{E_{high}(i)} - \sum_{\mathrm{i}} \mathrm{E_{low}(i)} = \Delta\mathrm{CBH_{correction}} \qquad (5)$$

$$\mathrm{E_{High}(full)} \approx \mathrm{E_{Low}(full)} + \Delta\mathrm{CBH_{correction}} = \mathrm{E^{CBH}} \qquad (6)$$

where $\mathrm{E_{high}(full)}$ is the (extrapolated) energy of the full molecule calculated at the high-level of theory, $\mathrm{E_{low}(full)}$ is the energy of the full molecule calculated at the low-level of theory, $\mathrm{E_{high}(i)}$ is the energy of the *i*th fragment calculated at the high-level of theory, $\mathrm{E_{low}(i)}$ is the energy of the *i*th fragment calculated at the low-level of theory, and $\Delta\mathrm{CBH_{correction}}$ is the total CBH correction to the full low-level energy. The summation in Eq. (5) is performed over all product fragments (with a positive coefficient) and the reaction fragments (with a negative coefficient). Hydrogens are added as appropriate to maintain the original hybridization of each atom. Similar ideas, viz. using two levels of theory to improve the accuracy, are commonly used in fragmentation-based methods such as Molecules-in-Molecules (MIM),[38] and the relationship between CBH and MIM have been discussed in one of our previous publications.[39]

Overall, the method requires a full calculation at the low level of theory (DFT) and fragment calculations at the low and high levels of theory. While high-level calculations are required on the fragments, it should be noted that the size of the fragments (at any rung of CBH) is *independent of the size of the parent molecule*. Thus, as the parent molecule gets larger, the computational cost of the high-level calculations grows only linearly with the size of the system. In addition, all the fragment calculations are carried out at their equilibrium geometries (*vide infra*). Since the same fragments frequently appear for many different parent molecules, their energies can often be obtained from a lookup table where the energies can be stored. Overall, the cost of the extrapolated E_{High}(full) is dominated by the cost of E_{Low}(full). In the applications shown below, we will show that the performance of CBH approaches that of the high level of theory while the cost of CBH is that of the low level of theory, hence the term "coupled cluster accuracy at DFT cost". For the remainder of this perspective, the term "coupled cluster accuracy" will be defined to be 1-2 kcal/mol. We also use the terms "mean absolute deviation" (MAD) and "mean absolute error" (MAE) interchangeably (as used in the original publications), obtained as the mean absolute difference between a calculated quantity and its reference value (coupled cluster or experiment).

3. Generalized graph-theoretic approach for CBH generation and implementation

Molecular systems are defined by their constituent atoms along with the bonds between them. Likewise, graphs are defined by a set of nodes connected by edges representing relationships between nodes. In chemical graph theory, regions of a molecule are coarse-grained into nodes and connected to form edges based on a defined interaction threshold. Standard CBH fragmentation utilizes the most basic coarse-grained graph, which is formed by taking into consideration only the *heavy* (non-hydrogen) atoms as nodes, treating the hydrogens implicitly, and connecting nodes to form edges. This procedure provides a *hydrogen-suppressed* chemical graph, which is a simplified version of the typical Kekulé style molecular structure drawings universally used in organic chemistry.[40] A set of nodes and edges define a graph *G* with a well-defined structure and connectivity.

Since the fragments of the *Connectivity-Based* Hierarchy are constructed from connectivity information alone, each rung of CBH can be defined as a graph neighborhood from Graph Theory—more specifically based on the geodesic distance.[40] In graph theory, the geodesic or graph distance *d(u,v)* between two nodes *u* and *v* is defined to be the length of the shortest path along

edges connecting the nodes, where each edge has a length of 1. In this work, we define a graph neighborhood N_k as a subgraph of G containing all nodes within a distance k of a certain point. The smallest neighborhood N_0 corresponds to the base entity with no neighbors. Within this definition, graph neighborhoods can be centered on either a node or edge, denoted as $N_k(n)$ or $N_k(e)$ respectively. Node-centered graph neighborhoods $N_k(n)$ include the base node along with all other nodes in which $d \leq k$. On the other hand, edge-centered graph neighborhoods $N_k(e)$ measure the graph distance from the center of an edge, with the nearest nodes at a graph distance of 0.5, and all nodes in which $d < k$ are collected as $N_k(e)$. As an illustration, the molecular graph of methyl 3-butenoate is shown in **Figure 3a** with the first three node- and edge-centric neighborhoods (**Figure 3b**), where neighborhoods are centered on the entity closest to the dot and subgraphs are highlighted in black.

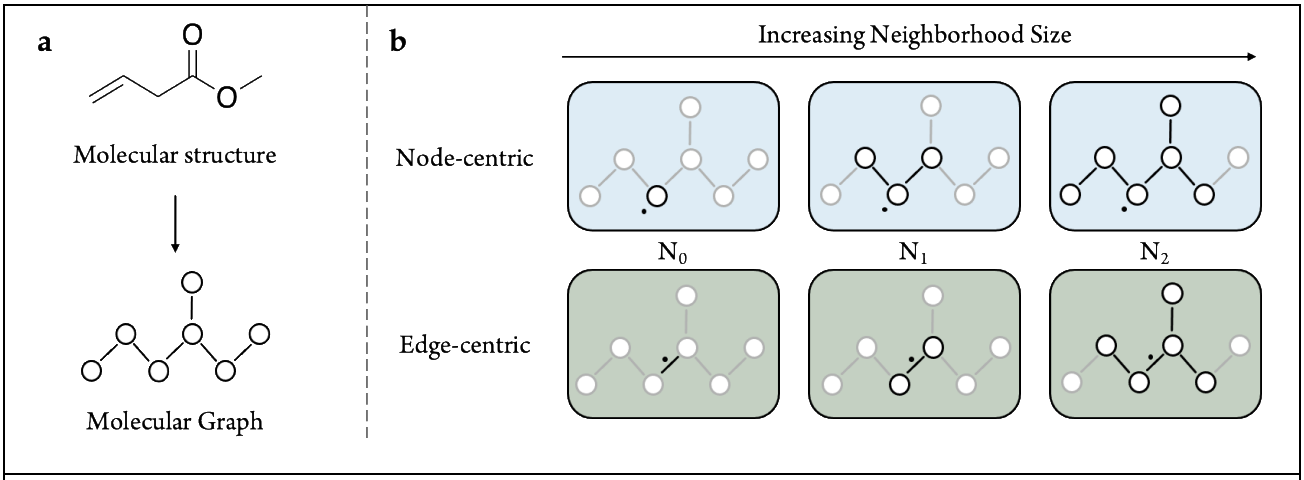


Figure 3. Graph neighborhoods in chemical graph theory, (a) molecular graph representation from the skeletal formula of a molecule with heavy (non-hydrogen) atoms as nodes and bonds as edges, (b) node- (blue) and edge-centric (green) graph neighborhoods (N_k) of increasing size, where k is the maximum graph distance included in the subgraph

The Connectivity-Based Hierarchy of reaction schemes can be visualized as the rungs of a ladder (**Figure 4**), such that ascending the rungs of the hierarchy increasingly preserves the local chemical environments of the parent molecule, achieving a better matching of the bond and hybridization types. CBH-n rungs alternate between atom- and bond-centric reactions, with even numbered rungs preserving the chemical environments of atoms and odd numbered rungs preserving the environments of bonds. The fundamental definition of CBH allows for the

automated generation of the reaction schemes since the reactant side fragments of a given reaction can be derived from the product side fragments of the previous rung.

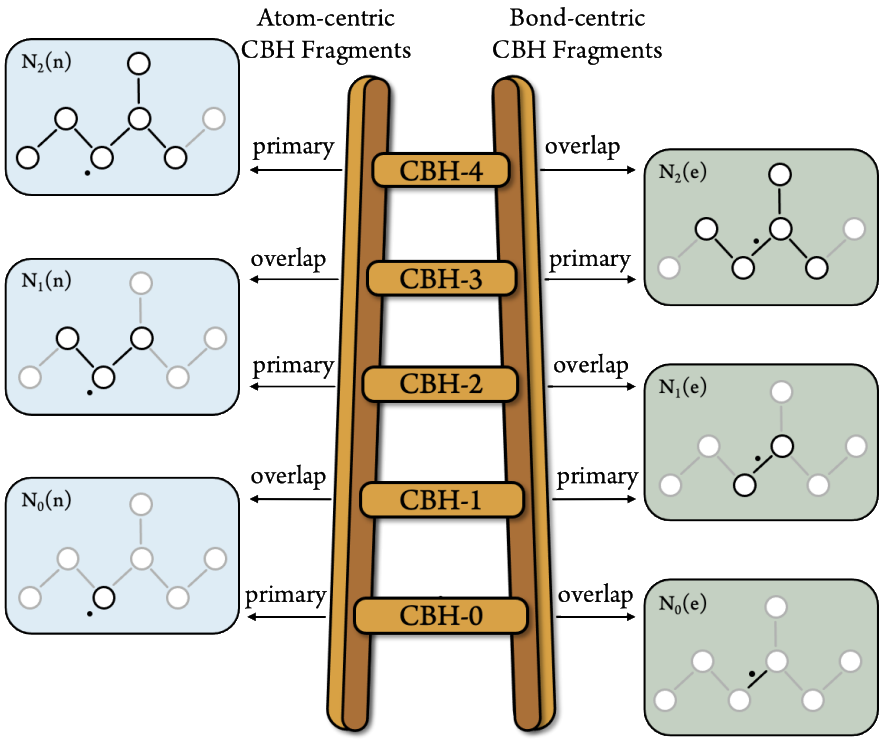


Figure 4. Primary and overlap fragment generation of various rungs of the Connectivity-Based Hierarchy ladder as they relate to atom- and bond-centric graph neighborhoods

Here, we outline the generalization of the CBH fragmentation protocol in the graph neighborhood definition (**Figure 5**). First, the full molecular graph is formed and divided into primary subgraphs to form the product side fragments of the selected (CBH-n) reaction scheme using neighborhood N_k(n) for atom-centric rungs or N_k(e) for edge-centric rungs (see **Figure 4**). Second, to cancel the overcounting of atoms in the overall reaction, the overlapping regions are calculated from the corresponding graph neighborhood of the previous CBH-(n−1) rung. Finally, each subgraph is expanded back into their full molecular form along with sufficient hydrogens to account for the atomic features of the full molecule. To form the full CBH reaction, identical fragments are collected to give the reaction coefficients where each primary-type subgraph has a coefficient of +1 and each overlap-type subgraph has a coefficient of −1.

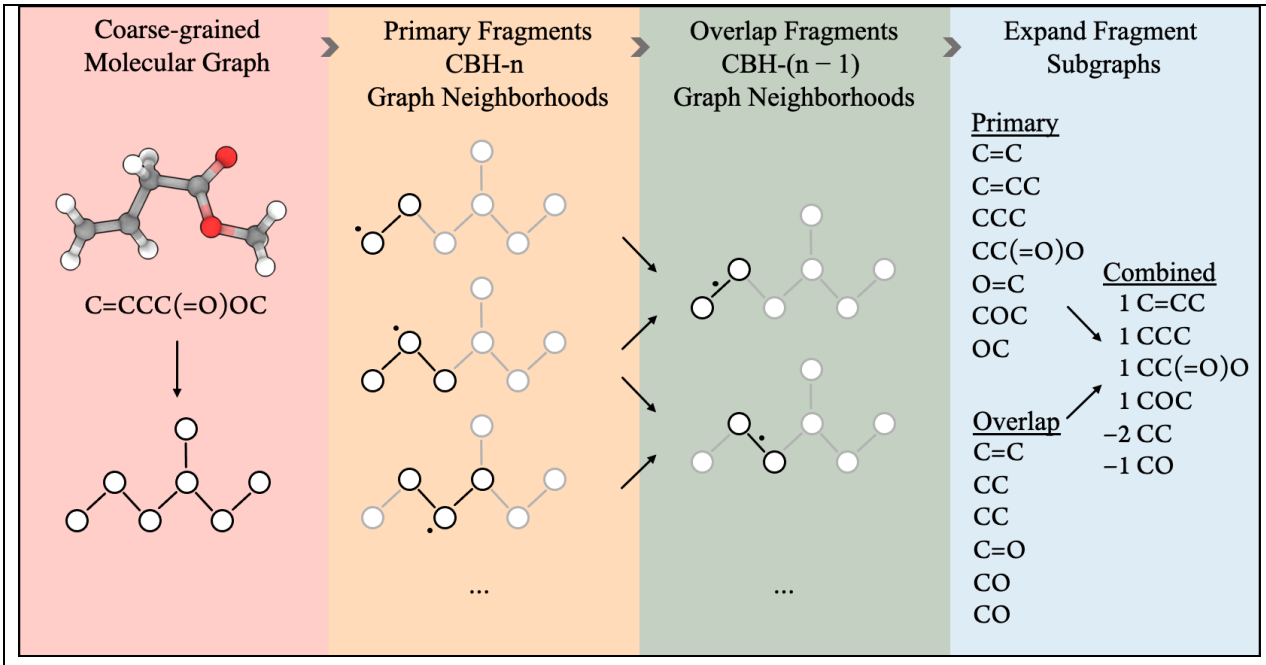


Figure 5. Graph-based construction of CBH-2 for methyl 3-butenoate ($C_5H_8O_2$)

Using the ideas mentioned above, an automated python package to generate the CBH reaction schemes (pyCBH) has been developed in our group by one of the authors (EMC). The pyCBH package is open-source and freely available on Github at https://github.com/colliner/pyCBH. The development of pyCBH was motivated by the inherent systematic structure of the rungs of fragmentation in CBH as well as the need to quickly calculate thousands of CBH corrections in an automated manner. pyCBH follows the exact procedure outlined in the previous section and **Figure 3**. Fragments are formed from a parent molecule given in either cartesian coordinates or from the SMILES representation for any user-requested CBH rung. Typically, in other fragmentation-based methods, the calculation of all two- to *n*-body overlaps between fragments is required in order to satisfy the inclusion-exclusion principle.[19-23] For CBH, however, the total number of fragments (including overlaps) is equal to the number of nodes and edges in the molecular graph, and *no higher-order overlaps need to be calculated*, making the algorithm highly efficient. In this context, we note that a few other groups have also taken notice of CBH's ease of automation, usually employing a graph-based algorithm in their implementation of the protocol.[41,42]

Included with pyCBH is a lookup table of many of the common fragments formed with CBH-0 to CBH-3 along with a database of energies calculated at various levels of theory. If all

fragments of a generated CBH reaction are present in the database, the ΔCBH correction (*vide infra*) can be computed automatically from the lookup table without the need for further electronic structure calculations.

4. Chemical Applications with CBH

As mentioned earlier, the initial applications with CBH were carried out using experimental heats of formation for the reference molecules to obtain accurate enthalpies of formation for neutral organic molecules,[24] organic radicals,[29] carbocations,[30] and biomolecules[43] such as amino acids.[44] The focus in these early studies was on applications using CBH-2 and CBH-3 to demonstrate significant improvement over the conventional isodesmic (CBH-1) formalism. In most of these studies, CBH-2 as well as CBH-3 yielded results within 0-2 kcal/mol of experiment while the CBH-1 errors were much larger. For example, in the initial paper on the study of 20 neutral organic molecules containing 6-13 heavy atoms, the mean absolute deviations from experiment, averaged between seven different density functionals, were 5.2, 1.4 and 1.1 kcal/mol for CBH-1, CBH-2 and CBH-3, respectively.[24]

The strategy of combining two levels of theory (high and low) to derive the CBH-corrections without the need for any experimental data on the component systems was developed in 2013 and used to extrapolate to CCSD(T) energies using MP2 as the low level.[35] For a slightly larger test set of 30 neutral organic molecules containing 6-13 heavy atoms, CBH-2 and CBH-3 showed remarkably low mean absolute deviations of 0.3 and 0.2 kcal/mol from the directly evaluated CCSD(T)/aug-cc-pVDZ energies. For a smaller subset of 14 molecules containing 6-8 heavy atoms, the deviations were slightly larger 0.8 and 0.6 kcal/mol with the 6-311++G(3df,2p) basis set, but well within chemical accuracy. For these test sets, the corresponding errors for CBH-1 were well outside the chemical accuracy range. Overall, these results showed an excellent compatibility between MP2 as the low level and CCSD(T) as the high level of theory.[35]

In all our more recent studies, we have focused our attention on our stated goal at the beginning of this manuscript – assessing the performance of DFT methods and trying to approach coupled cluster accuracy starting from DFT. We list a few examples below to illustrate the success of this approach.

4.1. Enthalpies of Complex Organic Reactions.

An illustrative study demonstrating the power of CBH, carried out in 2017, involves a careful comparison of the performance of more than 15 different density functionals (along with some MP2 variants) on a carefully assembled set of complex organic rearrangement reactions.[25] Reaction energies were computed with different DFT-based methods for a set of 25 organic reactions (named as CBH-R25 test set) to assess the systematic error-cancellation for the different methods using the CBH reaction schemes. CBH-R25 set contains molecules with 6-26 heavy atoms, including a variety of common organic reactions like Diels-Alder (R1–R6), aldol condensation (R7–R8), Pausson-Khand reaction (R13), aminoxylation (R14), and isomerization reactions (R17–R25). It includes a broad range of functional groups to provide a rigorous calibration of CBH performance. A few illustrative examples are shown in Figure 6, and the full set of reactions can be found in the original publication.[25] G4 energies for the reactions were used as the reference "experimental" values to assess the performance of DFT methods.

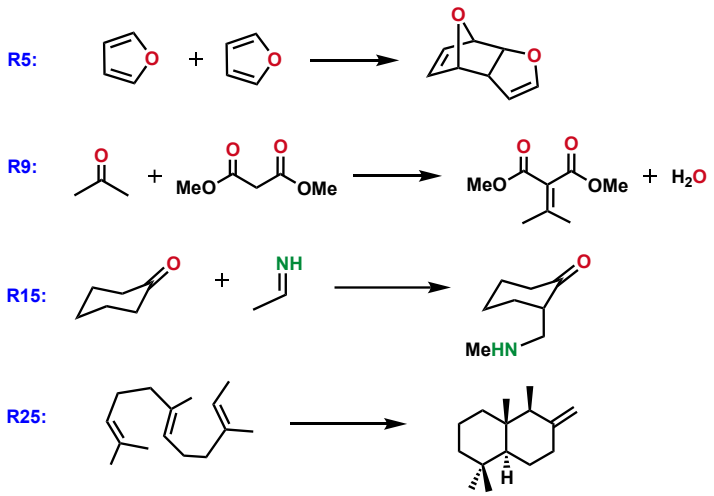


Figure 6. A few illustrative examples from the CBH-R25 set of 25 organic reactions. Adapted with permission from reference 25. Copyright 2017, American Chemical Society.

This work incorporates an additional important idea to improve the efficiency of CBH computations. This can be seen via an example (**Figure 7**) from the original publication.[25] It illustrates how CBH fragments (on which higher level calculations are needed to obtain the corrections) are well-balanced on the product and reactant sides while obtaining reaction energies, even for a seemingly complicated reaction like the ring opening rearrangement shown below. For each rung of CBH-*n* (*n*=1,2 in this study), the CBH reaction schemes are set up for the reactant

and the product to identify the *net change* in the elementary model reactions (**Figure 7**). Fragment molecules common to both the reactant and product side cancel each other, and the *resultant* CBH-1 and CBH-2 schemes (**ΔCBH-1** and **ΔCBH-2**) are then used to provide error correction. As seen in the illustration, the resultant (or net) CBH-1 and CBH-2 reactions involve only a modest number of small molecules, making it very easy to obtain the corrections. The energies of the resultant reactions, labeled as ΔCBH-1 and ΔCBH-2, are calculated using high and low levels of theory, yielding the corrections to the DFT reaction energies.

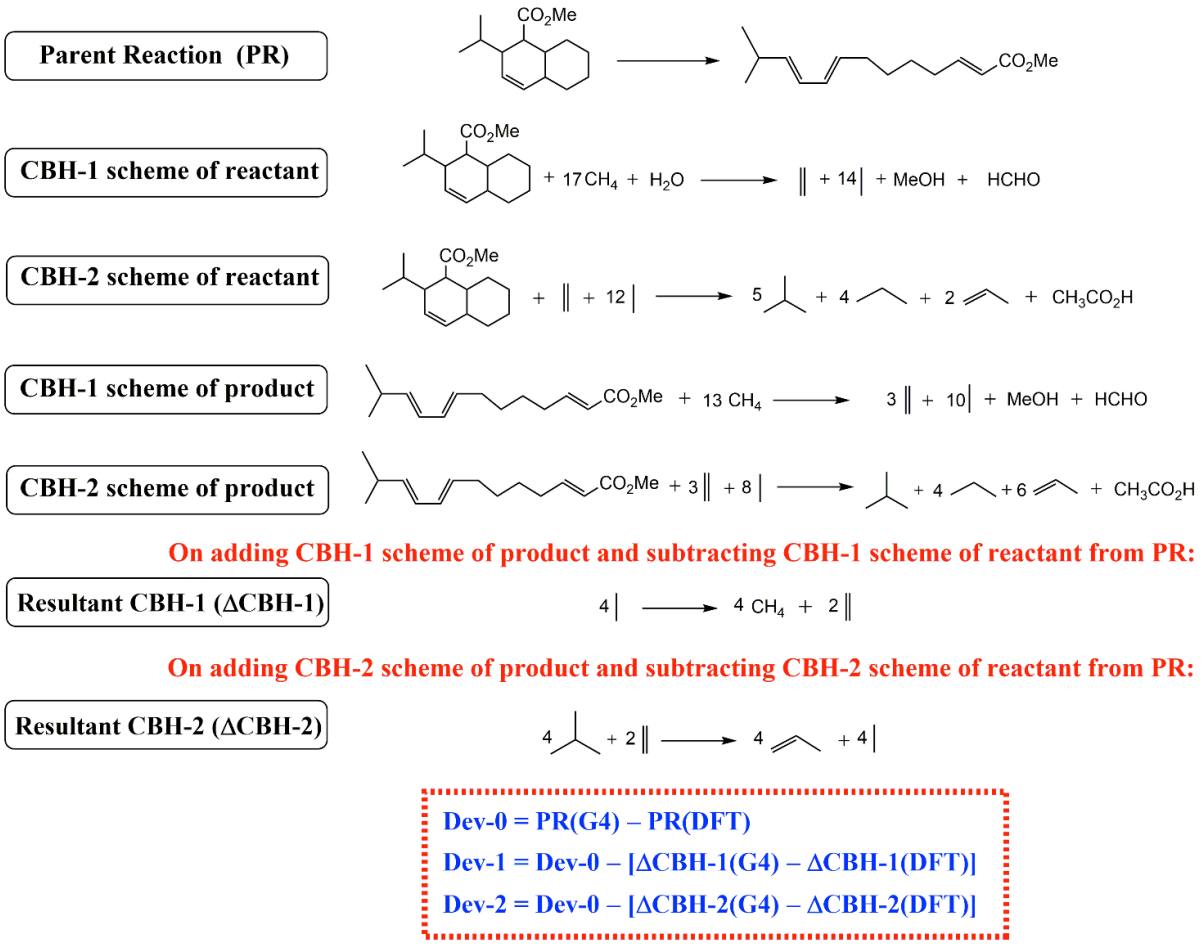


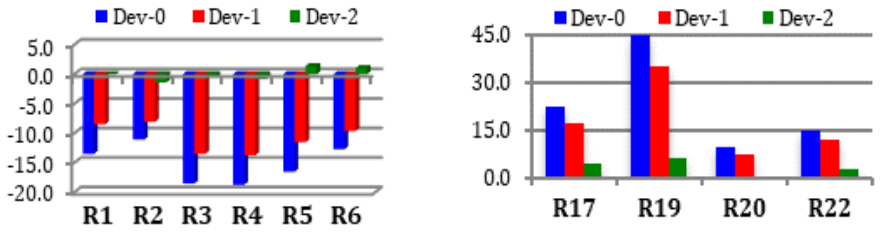
Figure 7. Derivation of ΔCBH-1 and ΔCBH-2 schemes with an illustrative example to define Dev-0, Dev-1, Dev-2. Reproduced with permission from reference 25. Copyright 2017, American Chemical Society.

The following systematic procedure can be used to derive the CBH-corrected energies.

(1) Compute the reaction energy with standard DFT method and the performance is evaluated *vs.* the accurate energies from G4 theory, and the deviation is denoted as "Dev-0". Dev-0 varies across a wide range of 0-45 kcal/mol, depending on the method used and the reaction.

(2) For each rung of CBH-*n* (*n* = 1, 2 illustrated above), calculations are carried out on the "resultant" reactions with both the current method (DFT) and the reference method (G4 theory) to calculate the associated corrections. The energy deviations at rung 1 (isodesmic) and rung 2 (isoatomic) are denoted as Dev-1 and Dev-2.

Averaged over the 25 reactions, the raw DFT mean absolute deviations range from a value as low as 2-3 kcal/mol (ωB97X-D[45] and M06-2X[46]) to as high as 12.9 kcal/mol (B3LYP[47,48]). In this context, it is interesting to note that the most popular density functional, B3LYP, shows the largest errors for this test set. Thus, to illustrate its usefulness, we first examine the performance of the ΔCBH schemes in conjunction with the B3LYP functional. As pointed out in previous studies,[49] B3LYP underestimates the reaction enthalpies of the six Diels-Alder reactions in the test set due to an inadequate description of σ → π bond transformations (delocalization), hyperconjugation, and dispersion interactions present in the cyclic and bicyclic products. Application of the ΔCBH-1 (isodesmic) scheme give marginal improvement, but dramatic improvement is observed with ΔCBH-2 (isoatomic). The mean absolute deviations of Dev-0, Dev-1 and Dev-2 for these 6 reactions are 15.2, 10.8 and 0.9 kcal/mol, respectively. The small value of Dev-2 suggests that similar 1,3 alkyl-alkyl interactions and hyperconjugation effects are present in both the CBH-2 fragments and the parent molecules.[50] Similar substantial improvements are seen for most other reactions in the test set. Considering the full test set, the large B3LYP mean absolute deviations (MAD-0 = 12.9 kcal/mol) in reaction enthalpies decrease only slightly with the popular isodesmic (ΔCBH-1) schemes (MAD-1= 9.6 kcal/mol) but improve dramatically using ΔCBH-2 schemes (MAD-2 = 1.7 kcal/mol).



(a) (b)

Figure 8. B3LYP calculated Dev-0 and CBH corrected deviations (Dev-1 and Dev-2 at ΔCBH-1 and ΔCBH-2, respectively) in the reaction energies of (a) R1–R6 and (b) R17, R19, R20 and R22. Reproduced with permission from reference 25. Copyright 2017, American Chemical Society.

The substantial improvement at CBH-2 (isoatomic) over CBH-1 (isodesmic) for the B3LYP functional is illustrated in **Figure 8**. **8(a)** shows the deviations for six Diels-Alder reactions while **8(b)** shows the deviations for four larger isomerization reactions with significant deviations. The raw B3LYP deviations are shown in blue, the CBH-1-corrected values in red, and the CBH-2-corrected values in green. In all cases, the isodesmic correction is modest while the isoatomic correction is dramatic.

Similar improvements are seen for all DFT (and MP2) methods.[25] **Figure 9** represents the mean absolute deviations (MAD-0, red) in the reaction energies of reactions **R1**–**R25,** and corrected deviations through ΔCBH-2 schemes (MAD-2, green). The dramatic decrease of the deviations from the MAD-0 to MAD-2 across the various DFT and WFT based methods demonstrates the consistently excellent performance of the ΔCBH-2 schemes. Only the local density functional (SVWN5) has a MAD-2 of greater than 3 kcal/mol (3.6 kcal/mol) after error-cancellation. Interestingly, even Hartree-Fock theory shows a deviation of under 3 kcal/mol.

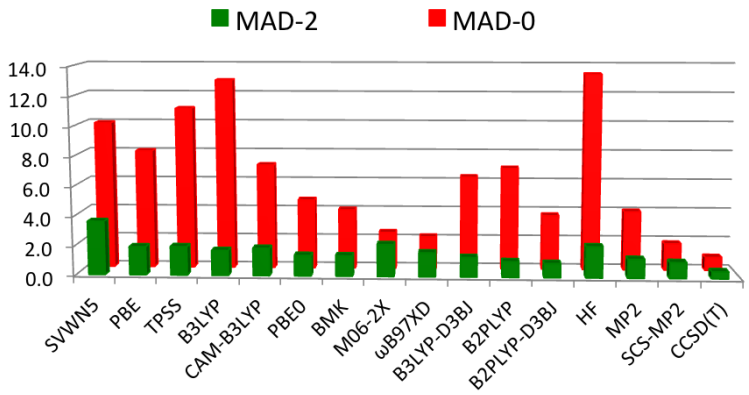


Figure 9. Calculated mean absolute deviations (MAD-0) and CBH corrected MAD-2 in the reaction energies of R1–R25 reactions at selected levels of theory. Reproduced with permission from reference 13. Copyright 2017, American Chemical Society.

The conclusions from this study are extremely important for understanding the deficiencies of DFT. Our results clearly demonstrate that *the disparate results from different functionals stem from the systematic errors in the underlying elementary reactions that represent the changes in the bonding environment between reactants and products*. Our rigorous CBH-protocol corrects for the systematic errors of DFT methods to yield accurate enthalpies of complex organic reactions. Most notably, the performance differences between different density functionals decrease dramatically. In conjunction with ΔCBH-2 schemes, most functionals yield deviations of 1-2

kcal/mol, and the best functionals such as the double-hybrid B2PLYP-D3BJ yield a MAD of only 1.0 kcal/mol.

A careful analysis of the relative performance of all the DFT methods reveals that many of the previously known performance trends[10] for *families of DFT functionals* still hold after ΔCBH-2 corrections, but *the range of errors is compressed*. For example, hybrid functionals work better than gradient corrected (GGA) functionals. Thus, PBE has a MAD-2 of 1.9 kcal/mol while the hybrid PBE-0[51] functional performs better (1.4 kcal/mol). Inclusion of D3 dispersion corrections[52] improves the performance for these organic reactions. Thus, B3LYP has a MAD-2 of 1.7 kcal/mol while B3LYP-D3BJ has a smaller error of 1.3 kcal/mol. Double hybrid DFT functionals include a component of MP2 electron correlation and are known to perform better than hybrid functionals. B2PLYP,[53] a double-hybrid derivative of B3LYP resulted in significant improvement over B3LYP. Thus, after the CBH-2 corrections, B2PLYP and B2PLYP-D3BJ yield MAD-2 of 1.1 and 1.0, respectively. With only 2 deviations greater than 2 kcal/mol and a MAD of 1.0 kcal/mol, B2PLYP-D3BJ shows consistent performance irrespective of the size of the molecules in the reactions, making it the most accurate functional tested.

Our broad conclusions from this study[25] are that traditional isodesmic corrections, though useful, are far from achieving chemical accuracy. Most importantly, the simplest next level correction from the CBH hierarchy (CBH-2) can achieve dramatically improved results, reaching near chemical accuracy (1-2 kcal/mol). An even more striking observation is that the performance differences between the different density functionals mostly evaporate after the application of the CBH-2 corrections. Thus, any functional can be used with CBH-2 corrections to achieve high accuracy. In the next sections, we demonstrate that similar results are likewise obtained in other examples, though CBH-3 corrections may be needed in some cases to obtain further error cancellation.

4.2 Bond dissociation energies in biofuel molecules

As illustrated thus far, the CBH protocol offers a route to derive accurate thermochemical properties of organic molecules using computationally inexpensive methods such as density functional theory. However, all the organic reactions considered in the previous section were closed shell systems. In a 2019 study, we explored the performance of CBH to obtain accurate bond-dissociation energies (BDEs) of various biodiesel esters,[31] exploring two new aspects for

CBH. First, bond dissociation leads to radicals, and thus the performance of CBH for open shell systems can be explored. Second, since the unpaired electron is contained in one of the product fragments, the stronger tendency of radical systems to delocalize can be explored at higher rungs of CBH. To this end, we explored CBH-2 and CBH-3 schemes in this work, and the excellent performance at CBH-3 suggested that further rungs are not needed for this system.

BDEs of several C–C, C–O, and C–H bonds, comprising a total of 21 reactions involving smaller to medium-sized biodiesel esters were chosen for initial calibration. The ΔCBH procedures adopted are illustrated in **Figure 10**.

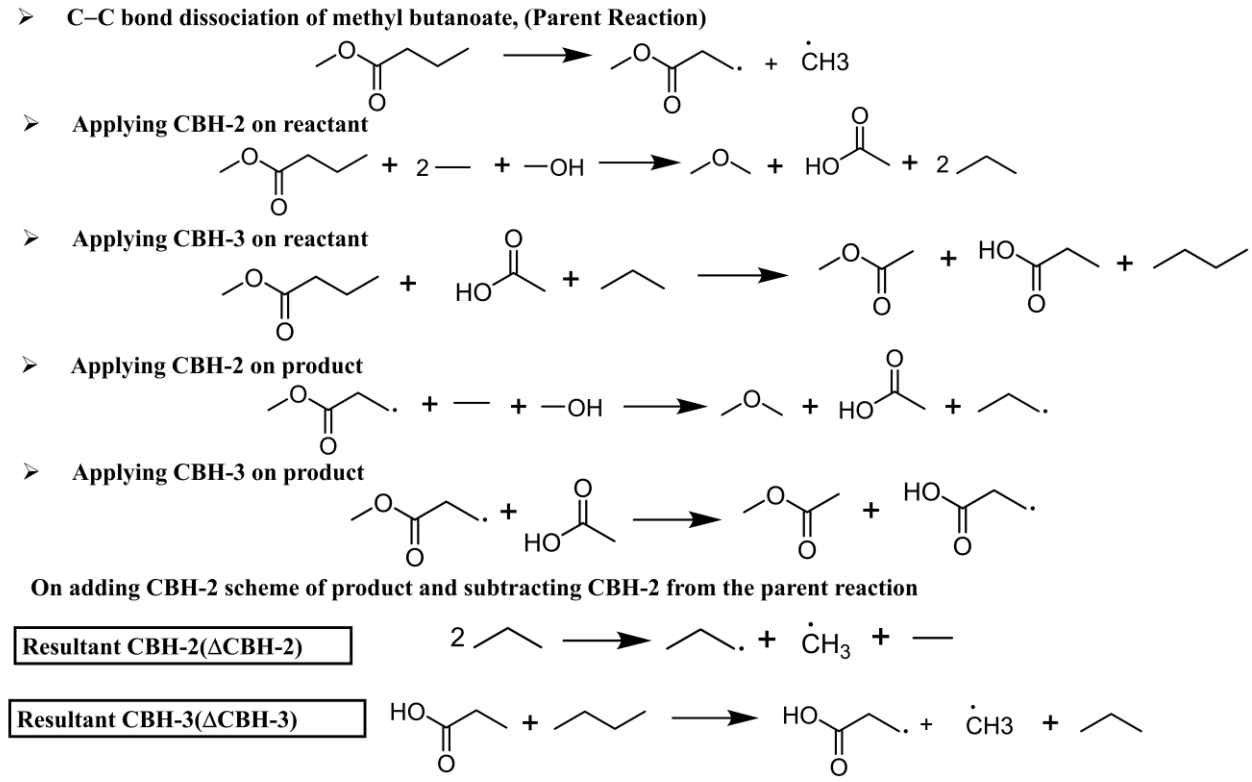


Figure 10. Derivation of ΔCBH-2 and ΔCBH-3 schemes for the bond dissociation reaction for methyl butanoate. Reproduced with permission from Reference 31. Copyright 2019, American Chemical Society.

The performances of five different popular DFT methods (B3LYP, B97, M06-2X, ωB97X-D, and B2PLYP), with and without empirical dispersion correction, in conjunction with CBH protocol, were compared with MRACPF2 (a multireference treatment using a modified coupled pair functional approach) values reported by Carter and coworkers.[54] Overall, DFT results after application of ΔCBH corrections are comparable with those from the multireference methods. Accuracy improves for all DFT functionals, yielding similar overall deviations. In particular,

MADs, especially for dispersion corrected functionals, fall within a narrow range of 0.2 kcal/mol. Among the different density functionals, ωB97X-D and B97-D3 show the best performance with a MAD of 1.3 kcal/mol from MRACPF2. Moreover, further improvement is achieved by applying ΔCBH-3 corrections, yielding a MAD within 1 kcal/mol (0.9 kcal/mol).

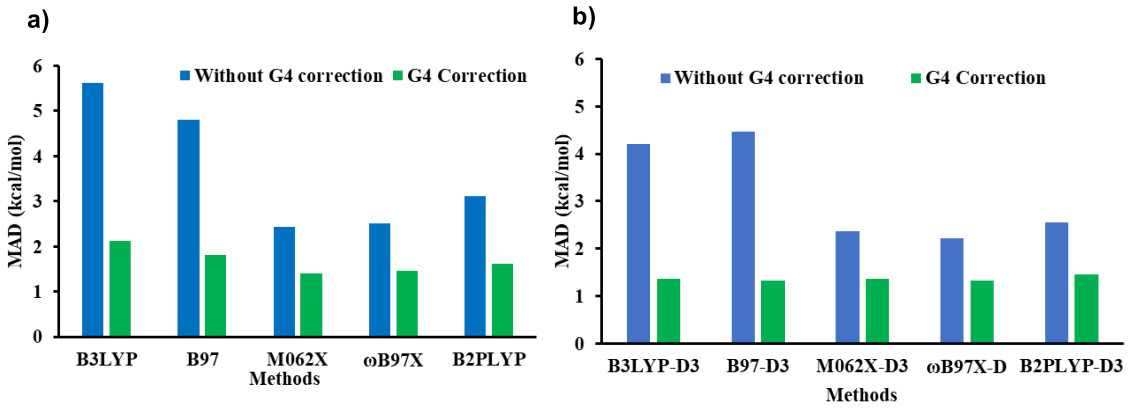


Figure 11. Graphical representations of the calculated mean absolute deviations (MAD) in BDEs of all the reactions with and without including G4 correction (ΔCBH-2 scheme) using (a) DFT and (b) DFT-D methods. Reproduced with permission from Reference 31. Copyright 2019, American Chemical Society.

The ΔCBH-2 and ΔCBH-3 correction schemes have also been applied to a larger biofuel component, methyl linolenate (**Figure 12**).

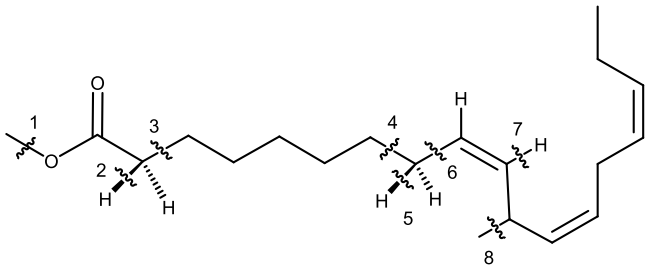


Figure 12. BDEs of eight different bonds in methyl linolenate calculated using ωB97X-D including ΔCBH-2 and ΔCBH-3 corrections. Reproduced with permission from Reference 31. Copyright 2019, American Chemical Society.

For a set of eight different bond dissociation reactions in methyl linolenate, the MAD for both CBH-2 and CBH-3 schemes are within 1-2 kcal/mol of the MRACPF2 results. The computed ωB97X-D BDEs with ΔCBH-2 and ΔCBH-3 corrections yielded a MAD of 1.8 kcal/mol and 1.1 kcal/mol, respectively, again illustrating the excellent performance of CBH.

4.3 Calculation of Redox Potentials with CBH

The redox potential gives the free energy cost of electron loss/gain and is a useful thermodynamic and kinetic tool. While extremely important, obtaining accurate and reliable redox potentials remains a steep challenge.[55-57] In particular, solid computational protocols that focus on high accuracy and reproducibility are critical for cases where consistent experimental measurement is difficult. We have applied our ΔCBH protocol for the calculation of accurate redox properties of organic molecules.[33] Redox calculations consider, in addition to the presence of open shell radicals in both oxidized and reduced species, *effects from solvation*. Thus, solvation models are an additional component for redox potential evaluations. A dependable protocol for redox property prediction that eliminates systematic errors in DFT, while remaining computationally feasible, holds tremendous value.

In 2020, our group introduced such a protocol, called CBH-Redox, a method for calculating accurate redox potentials using *implicit solvent* models.[33] This protocol is an extension of CBH and is appropriate for chemical processes involving an electron transfer. An example of the CBH-Redox fragmentation scheme is given in **Figure 13**. As in the case of the previous two applications, for a chemical reaction involving the loss or gain of a single electron, similarities in reactant and product structures results in a cancellation of fragments on either side of the reaction, and only a few high-level calculations of fragment molecules must be performed. Thus, the CBH protocol provides substantial computational speedups for reaction energies. It must be noted that to attain effective error cancellation, the main atomic site of oxidation should be known with some certainty. Incorrect identification of the oxidation site can lead to insufficient error cancellation. Nonetheless, many of the effects of electron delocalization are captured by the low level of theory, and low-level population calculations for the reduced and oxidized species may be useful to determine the most likely site of oxidation.

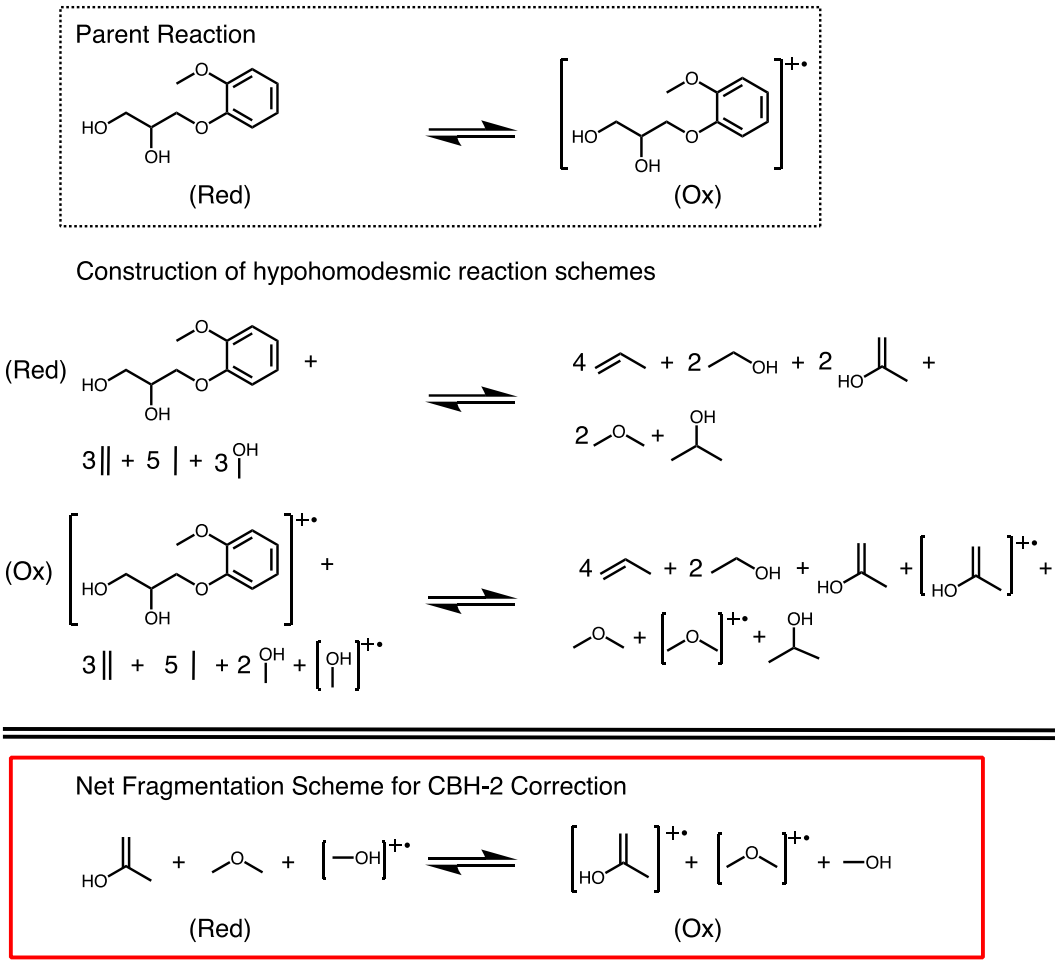


Figure 13. CBH-2 fragmentation scheme for 3-(2-methoxyphenolxy)-1,2-propanediol redox couple. Reproduced with permission from Reference 33. Copyright 2020, Royal Society of Chemistry.

We applied CBH-Redox to a test set of 46 C, O, N, Cl, F, and S-containing molecules in SMD implicit solvation and achieved impressive accuracy.[33] The test set for CBH-Redox features a range of functional groups, namely alcohols, aldehydes, alkyl-halides, amines, ethers, ketones, nitriles, nitro compounds, phenyls, thioethers. A proton-coupled electron transfer (PCET) formalism was used to compute the redox potentials for the phenolic compounds and the tyrosine derivatives. To test the strength and robustness of the protocol, CBH-Redox calculated potentials were evaluated with four popular density functionals. B3LYP, CAM-B3LYP, ωB97X and M06-2X were tested, with and without dispersion corrections. **Figure 14** shows the comparison of the calculated results with the corresponding G4 values. Comparisons with experimental values are very similar.

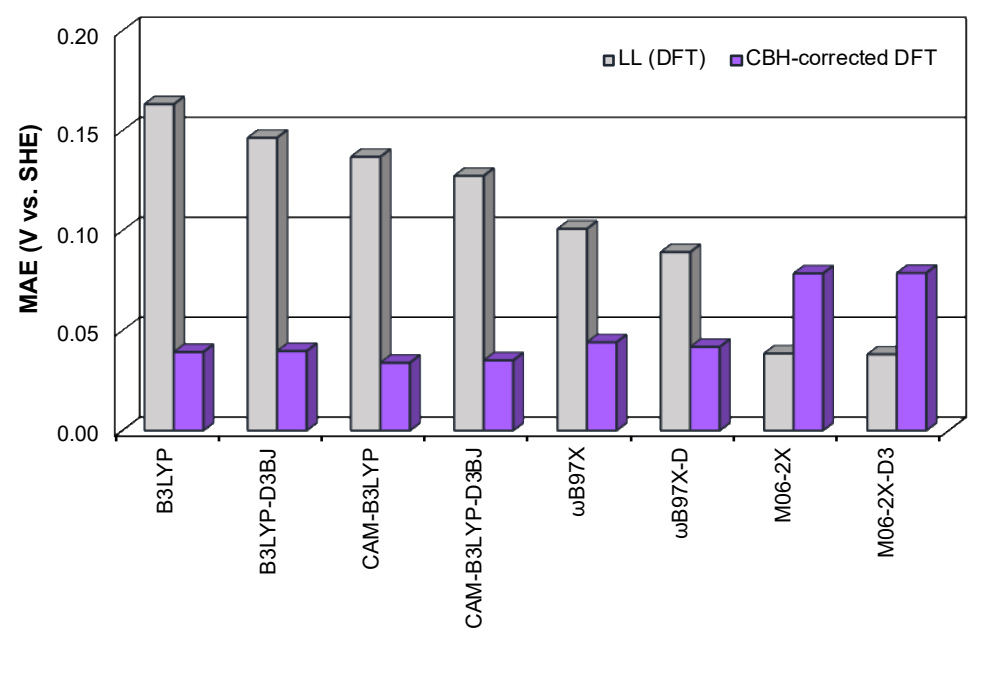


Figure 14. MAE of low level and CBH-2 redox potentials versus G4. Reproduced with permission from Reference 33. Copyright 2020, Royal Society of Chemistry.

We note that the raw performance of DFT methods in this case is quite reasonable with the MAEs ranging from 0.04-0.16 eV. Overall, the CBH-Redox protocol achieves a significant improvement in accuracy, yielding a MAE of 0.05 V or below versus G4 for six of the eight density functionals tested (B3LYP, B3LYP-D3BJ, CAM-B3LYP, CAM-B3LYP-D3BJ, ωB97X, and ωB97X-D). It is interesting to note that the M06-2X functional, achieves an overall MAE on par with G4 even before application of the CBH correction for this test set. Therefore, trying to improve upon this method using G4 fragments perhaps is inappropriate. That may explain the seemingly worse performance of M06-2X with the CBH correction. Nevertheless, considering the entire test set, the protocol's MAE falls well within the benchmark threshold for CBH.

4.4 Calculation of Accurate Acid Dissociation Constants (p*K*a) with CBH

We have also developed a standard protocol for accurately calculating pK_a's of a wide range of bio-organic molecules in the aqueous medium. This involves the evaluation of the free energy changes for protonation/deprotonation reactions.[58,59] While the spin state of the system does not change upon protonation, the *solvation requirements for the calculation of accurate* pK_a*'s are more stringent*.[60] In particular, inclusion of *a few explicit water molecules* directly hydrogen-bonded to the functional group of interest is key to the determination of accurate values.[61] Thus, we have used an explicit-implicit solvation model (also called a microsolvation model) by including a few (1-3) explicit solvent molecules along with implicit solvation effects from the SMD model.[62] For a calibration set of 224 small bio-organic molecules containing a variety of functional groups, by using the explicit-implicit solvation model at the CBS-QB3 level (a variant of the complete basis set extrapolation model CBS-Q using B3LYP geometries),[63] an impressive accuracy of MAE = 0.45 pK_a units was achieved compared to experimental pK_a values in the range of −1 to 20. For the larger molecules, where CBS-QB3-based approach is computationally unaffordable, we have developed an efficient pK_a calculation protocol based on the CBH error-cancelation scheme. Full details can be found in the original publication, but two new factors from the protocol used in this work should be noted. The full molecule calculations are done with DFT with implicit solvation while ΔCBH-2 corrections are determined for the CBH fragments using the CBS-QB3 method with an explicit-implicit solvation model. If the group undergoing deprotonation is directly bonded to an aromatic ring, the full aromatic ring was considered as a single group to maintain the delocalization across the aromatic ring.

The CBH protocol was assessed on a set of 28 relatively complex drug molecules (**Figure 15**) and the results are shown in **Figure 16**. This is a challenging set of molecules with some of them containing multiple ionizable groups or tautomeric forms (e.g., structures **2, 3, 9** in **Figure 15**), and provides a critical test of the performance of computational models for pK_a predictions.

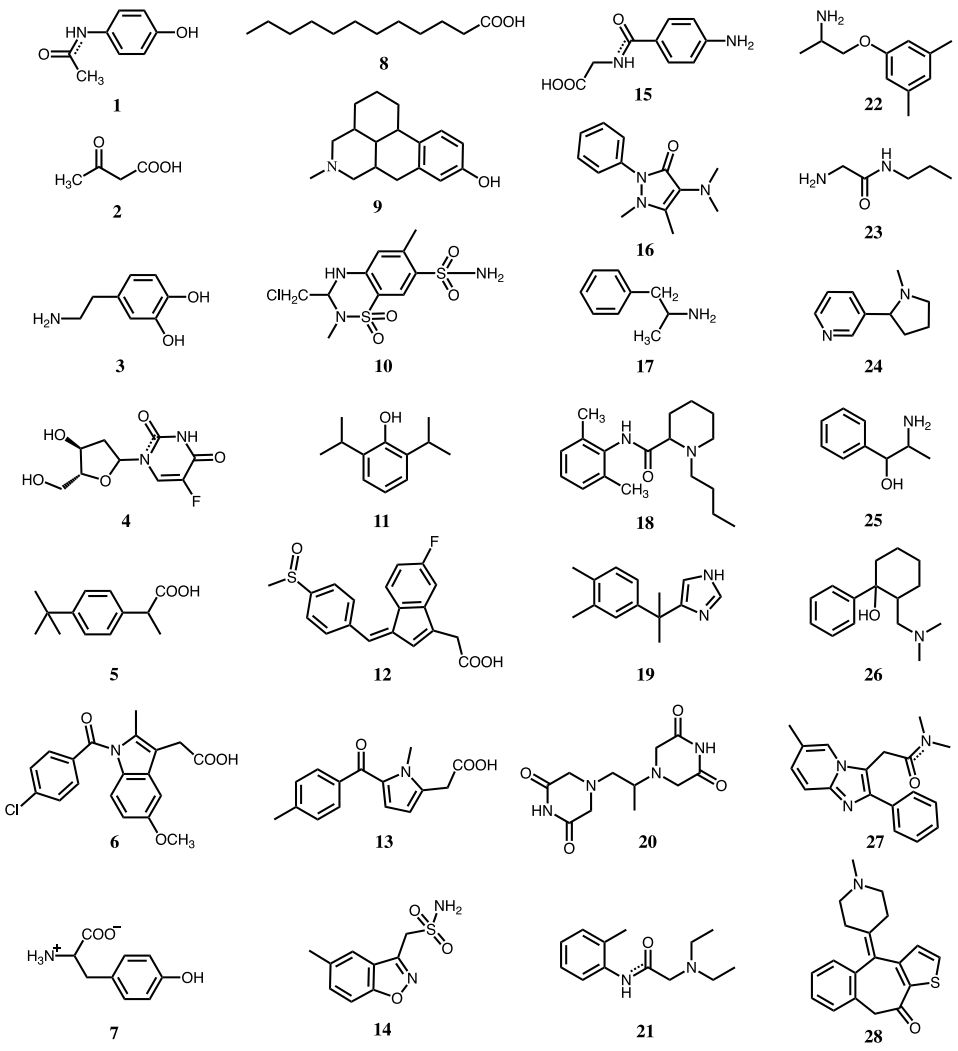


Figure 15. Structures of 28 drug molecules for testing the CBH-p*K*a protocol. Reproduced with permission from reference 32. Copyright 2019, American Chemical Society.

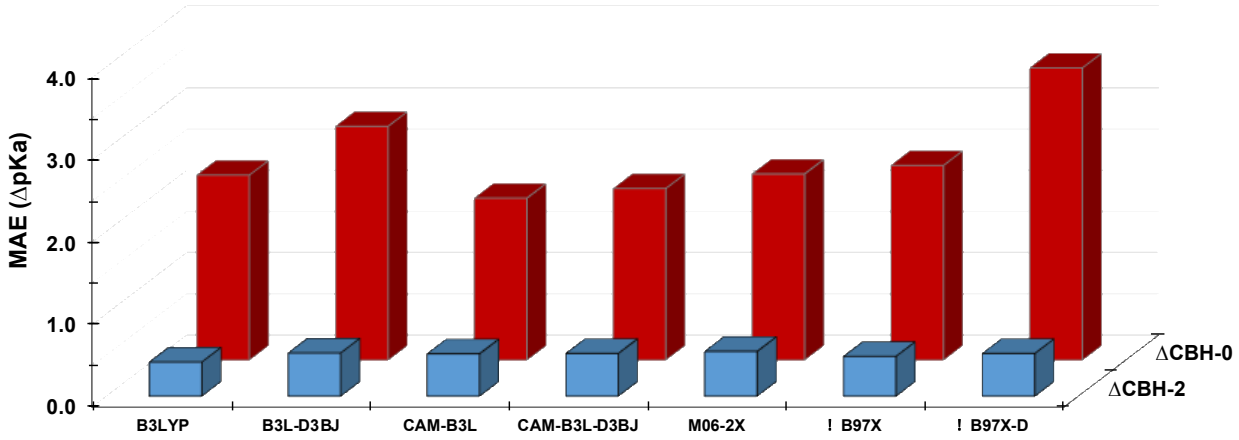


Figure 16. Calculated mean absolute error in pK_a (ΔpK_a -0) and ΔCBH-2 corrected pK_a errors (ΔpK_a-2) for various DFT methods for the 28 molecule test set of drug molecules from **Figure 15**. Reproduced with permission from reference 32. Copyright 2019, American Chemical Society.

The raw errors are shown in red and the ΔCBH-2 corrected values are shown in blue for seven different density functionals. Using the ΔCBH-2 scheme, our protocol eliminates the systematic errors in different DFT methods to yield accurate pK_a values (MAE of 0.40-0.54 pK_a units) for these relatively complex molecular systems. In particular, our results show that by treating the elementary deprotonation reactions at the CBS-QB3 level with explicit-implicit solvation, the calculated pK_a's are nearly independent of the underlying DFT method used. The pK_a calculation protocol based on CBH scheme also works if a molecule possesses multiple ionizable groups or tautomeric forms. For such molecules, separate CBH schemes can be constructed for each of the deprotonating functional groups to derive the separate elementary deprotonation reactions. In such cases, the *explicit water molecules are placed only near the deprotonating functional group under consideration*. In this way, each of the functional groups can be microsolvated locally and separately, without having to include explicit solvent molecules around all of the functional groups at once. This also avoids the complication that may arise while placing explicit water molecules around all of the functional groups at the same time. For example, for the molecule **III.3** which has two phenolic OH groups and one aliphatic amine group, three separate CBH reactions are constructed for each of the functional groups, and the corresponding pK_a values are then obtained.

Overall, the ΔCBH-2 model yields pK_a values with an impressive accuracy of ~ 0.5 pK_a units. In more complicated cases, the accuracy can potentially be further improved by using higher rungs of the CBH schemes (e.g., ΔCBH-3). Nonetheless, we note that the current protocol covers most of the common functional groups present in organic and biomolecular systems and should be useful for widespread application.

4.5. Thermochemistry for a Large Dataset: Combining CBH with Machine Learning

Finally, we show some results using our latest models combining the CBH approach with machine learning (ML). Since CBH is a strategy for error cancellation, this is a natural extension since ML deals with automated pattern recognition that could be used for further error cancellation.[64-67] In our work, we have focused explicitly on Δ(ML) models[68,69] that learn the difference between DFT and CCSD(T) for theoretical thermochemistry.[70] Full details are beyond the scope of the current review. Briefly, we used ideas based on CBH-type fragmentation to introduce a new family of molecular descriptors for machine learning. CBH naturally offers a hierarchy of simple, chemically intuitive grouping of atoms, tuned for progressive error-cancellation across the rungs. In the simplest model, we used CBH to enumerate the substructures in a given rung, and both product and reactant fragment coefficients were encoded to provide structure-based fingerprints.[70] This has two advantages relative to other structure-based fingerprints. First, since hydrogens are implicitly included in CBH, the shorter resulting input vector leads to more efficient encoding. Second, the use of both product and reactant coefficients provides some balance, leading to better performance.[70] We have labeled this model as DFT+ΔML(CBH) and have assessed its performance for the first three rungs (CBH-0, CBH-1 and CBH-2) on a test set of G4 calculations on over 1000 molecules containing H, C, N, O, Cl and S atoms ("1k-G4-C9" test set consisting of 1051 molecules with 9 or fewer carbon atoms).[70] Just like in the traditional ΔCBH corrections, the ML(CBH) molecular descriptors provide information about local structures. The ΔML(CBH) models are not based on any fragment energy calculations, but the trends in systematic errors can be directly learned instead. Our results are shown in **Figure 17**.

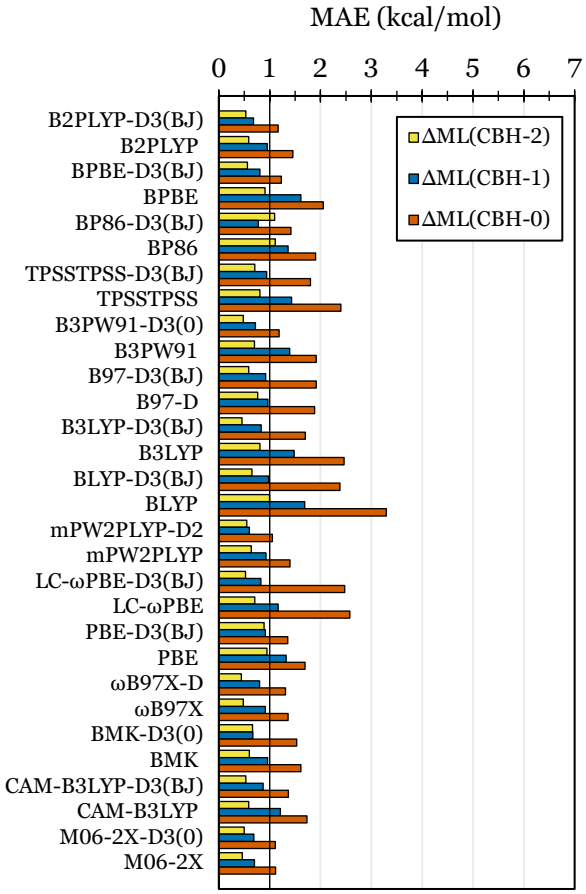


Figure 17. Final out-of-sample performance for all DFT+ΔML(CBH) models across 30 DFT baselines. Reproduced with permission from reference 70. Copyright 2020, American Chemical Society.

The results in **Figure 17** are very impressive. For a wide variety of density functionals, DFT+ΔML(CBH-2) models, trained on a set of small to medium-sized organic H, C, N, O, S, and Cl-containing molecules, achieve an out-of-sample MAE within 0.5 kcal/mol (and 2σ (95%) confidence interval of <1.5 kcal/mol) compared to accurate G4 reference values at DFT cost. All functionals tested, aside from BP86 and BP86-D3(BJ), achieve average errors within chemical accuracy using ΔML(CBH-2), with six functionals achieving less than 0.5 kcal/mol. The best performing functionals are ωB97XD, B3LYP-D3(BJ), and M06-2X ranging from 0.44 to 0.46 kcal/mol. In general, more sophisticated families of density functionals, i.e., double-hybrid and long-range corrected hybrid density functionals, outperform GGA functionals. B2PLYP-D3(BJ) and CAM-B3LYP-D3(BJ), for example, have mean absolute errors around 0.53 kcal/mol, while

some of the GGA functionals, such as TPSS, and BLYP, range from 0.89 to 1.00 kcal/mol. As expected, systematic errors are automatically cancelled out. Indeed, the MSEs of all three ΔML models are close to zero.

More advanced techniques in molecular machine learning have appeared in more recent times which utilize some of the deep learning models used in other fields.[71-73] To this end, we expanded our fragmentation-based ML approach to the FragGraph graph-network model[74] in which a molecular graph is constructed and local information about each CBH-2 fragment is embedded on the nodes of a graph.[75] Then, a graph network uses message-passing[76] to learn both from the structure of the fragments in a molecule as well as their relationship between one another through the graph structure. These methods have taken our ideas much further and achieved outstanding performance,[74] well within benchmark accuracy (kJ/mol), in predicting G4(MP2) energies for the ~130k molecules in the GDB9 test set.[77] However, our ML models have only been tested thus far on relatively small neutral molecules (nine heavy atoms or less) and it is not clear if their excellent performance can be extended for systems beyond the class of molecules included in the training set. In general, machine learning models are very good in interpolation but can fail in extrapolation to new systems (or predicting new properties). Much more work is clearly needed in this area, and machine learning is an active area of ongoing and future research in our group.

5. Other Applications and Future Prospects

In this perspective, we have only focused on works of our own. But more research groups are now adopting concepts associated with CBH and carrying out new applications. While a comprehensive discussion is beyond the scope of this perspective, we point out a few select papers. In one particular study, CBH-type reactions were used to order the relative thermochemical energies of 24 C_{84} isomers.[78] In another noteworthy study, CBH was used in combination with fast low-level computational methods (PBEh-3c, HF-3c, and HF/STO-3G), tight-binding DFT methods (GFN-xTB, DFTB, and DFTB-D3), and semiempirical methods (AM1, PM3, PM6, PM6-DH+, PM6-D2, PM6-D3H+, PM6-D3H4X, PM7, and OM2) on the set of 25 organic reactions first studied by us, showing the value of CBH, even when coupled with less accurate theoretical methods.[41]

One particularly thoughtful study involving CBH introduces CBH-ANL, an approach developed by Elliot and coworkers.[79] The method combines ANL1 energies for CBH-1 reference

fragments with ANL0 energies for CBH-2 reference fragments in a laddered scheme to improve the energy predictions. In this way, reliable values for the heats of formation for CBH-2 reference fragments may be achieved. The study also quantifies uncertainties of each reference fragment species, as well as their propagation to the full species, the largest uncertainty being 0.28 kcal/mol. The laddering approach in this study, which allows for the extension of CBH to larger molecules, follows naturally from CBH's systematic hierarchy and is a topic of research that we have systematically explored within our own group.[80]

Finally, we highlight some of the advantages and limitations of CBH and point out ongoing and future research directions that may be beneficial. As mentioned earlier, CBH has its foundation based on error-cancellation in theoretical thermochemistry. This has advantages and disadvantages. The biggest advantage is that the fragments (reactants or products) in CBH are calculated at their optimized equilibrium geometries.[39] This is because the experimental enthalpies of formations of these reference species (used to calculate the enthalpy of formation of the parent molecule) are valid only at their equilibrium geometries. Overall, since many large molecules share the same smaller optimized reference species, repetitive electronic structure computations are avoided in a thermochemical hierarchy such as CBH. As mentioned earlier, the energies of the recurring fragment species can easily be stored in a look-up table to avoid calculating them altogether. Thus, the overall computational cost is determined by the cost of the underlying DFT calculation, supporting the premise of "coupled cluster accuracy at DFT cost".

The traditional approach to CBH discussed thus far is applicable only for equilibrium structures. At any rung of CBH hierarchy, the reference molecules represent the *optimal* cutting scheme to achieve maximum error cancellation at that level of fragmentation. The higher CBH rungs then represent fragmentation schemes that yield smoothly increasing fragment size while progressively augmenting the efficiency of error cancellation. The application of CBH to nonequilibrium structures would unlock a vast domain of unexplored chemistry. In the original scheme utilizing experimental values of the fragments, such an extension would not be possible. But using our more recent efforts involving a second higher level of theory instead of experiment,[35] these restrictions can be relaxed. Such ideas are regularly used in the generalized implementation of fragmentation-based methods. A balanced approach merging the ideas from CBH with fragmentation may lead to more powerful and more broadly applicable computational techniques.

An important aspect of CBH is that the reaction schemes depend on the underlying valence bond structure that is used for generating valence-satisfied fragments. Hence, CBH, as described in this work, is not valid for delocalized systems (such as metals) or when hydrogen-terminated reference fragments are not easily generated (such as dative bonded systems), though the method could be adapted to make it applicable for the latter. In addition, all the applications thus far have been on the first- and second-row main group molecules, though extensions to heavier main group systems should be straightforward. CBH has not been applied to any transition metal systems.

A more serious concern for CBH is that if multiple resonance structures are possible for a given species, more than one CBH reaction may be obtained at a given rung, making the scheme non-unique. This would be true for some aromatic structures, at least for the higher CBH rungs. Similar ambiguities may be present in some radical structures or charged species. Since CBH restricts the unpaired electron or the charge to one of the fragment species, there may be multiple CBH reactions if the spin (or charge) is delocalized. However, non-uniqueness does not necessarily lead to poor performance. We have explored this to a limited extent in our study on carbocations[30] and shown that the results are quite insensitive to the choice of the CBH fragments. Nevertheless, non-uniqueness in such cases is not a satisfactory situation, and our group is currently considering strategies for tackling this issue in a rigorous manner.

It is well understood that the accuracy of a CBH reaction depends on the extent of error cancellation between reactants and products. Error cancellation should, in principle, increase with fragment size, since larger fragments capture greater portions of the molecular environment. In some cases, CBH fragments generated via lower rungs may prove insufficient in capturing the true molecular environment and may thus compromise accuracy. Conversely, higher rungs of CBH may compromise computational efficiency. To address this problem, we have developed coarse-grained models of CBH, which we have only briefly investigated thus far. For example, we pointed out earlier that the aromatic units (e.g., phenyl groups) were left intact in our pK_a studies.[32] In a more recent study, we have obtained slightly better performance from coarse-graining other functional groups such as nitro groups, sulfoxides, nitriles, etc.[81] This is an active topic of ongoing research.

As mentioned above, if there is a large mismatch between a substructure and the parent molecule, there could potentially be significant errors in the CBH approach. However, in many cases, the starting DFT does reasonably well for strained structures or crowded structures and the

issue is not major. However, when the mismatch is between electronic structures, e.g., delocalized aromatic structure vs. a localized small CBH fragment, there may be more significant problems and more caution is required in such cases. A signature of such a mismatch is that such systems will show a much stronger dependence between different density functionals while such differences disappear when there is a good match. To avoid such mismatches, coarse-grained CBH could be used such that highly strained substructures are not broken during fragmentation.

The formulation that we have discussed in this manuscript does not address the application of CBH schemes to conformers. If the same fragment conformations are used starting from two different parent conformations, there is no higher order correction and the performance remains the same as the low-level theory (i.e., DFT). This could partially be avoided by using fragment conformations that most closely resemble the structure of that unit in the parent molecule. For example, if a long-alkane chain in the fully extended conformation is compared with that of its folded form, the fragment butane units from the former will be in the trans conformation while some of them will have the gauche conformation in the latter, yielding a CBH contribution to the energy difference.

In principle, CBH methods are applicable for much larger molecules than illustrated in this work. However, as mentioned above, the CBH fragments are small and the resulting large number of fragments for larger molecules will potentially lead to accumulation of errors,[82] growing linearly with the number of fragments. Thus, error accumulation will be less pronounced if larger fragments are used. While coarse-grained CBH, briefly discussed in the final section, is a possible strategy, a general fragmentation approach gives much more flexibility in the generation of fragments of different sizes to optimize the accuracy and applicability of the calculations. Thus, we have carried out calculations on biological systems containing well over a thousand atoms[83,84] using our MIM[38] fragmentation method.

Finally, as mentioned briefly in the last section of this perspective, we stress that CBH-based descriptors can serve as useful candidates for the development of machine-learning strategies for chemical discovery. While we have given preliminary insight into this topic, we plan to investigate this avenue of chemical research more extensively to assess the performance of such models for chemical investigations in a broader context.

Acknowledgement

We acknowledge financial support from the National Science Foundation Grant CHE-2102583 at Indiana University. Scientific discussions on this topic over the years with Dr. Bishnu Thapa and Prof. Raghunath Ramabhadran are gratefully acknowledged.

Author Information

Corresponding Author:

E-mail: kraghava@indiana.edu

ORCID

Krishnan Raghavachari: 0000-0003-3275-1426

Sarah Maier: 0000-0002-3817-1476

Arkajyoti Sengupta: 0000-0002-9917-2472

Eric Collins: 0000-0002-9113-1705

Notes

The authors declare no competing financial interest.

References

[1] Bartlett, R. J.; Musial, M., Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys*. **2008**, *79*, 291-352.

[2] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M., A 5th-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989,** *157*, 479-483.

[3] Karton, A., A computational chemist's guide to accurate thermochemistry for organic molecules. *WIRES Comput. Mol. Sci.* **2016,** *6*, 292-310.

[4] Karton, A.; Daon, S.; Martin, J. M., W4−11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett*. **2011**, *510*, 165−178.

[5] Harding, M. E.; Vázquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F., High-accuracy extrapolated ab initio thermochemistry. III. Additional improvements and overview. *J. Chem. Phys.* **2008**, *128*, 114111

[6] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K., Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.

[7] Petersson, G. A., Complete Basis Set models for chemical reactivity: from the helium atom to enzyme kinetics. In Cioslowski J (Ed.), *Quantum-mechanical prediction of thermochemical data*. New York: Kluwer Academic, **2001**, pp 99–130.

[8] Deyonker, N. J.; Cundari, T. R.; Wilson, A. K., The correlation consistent composite approach (ccCA): an alternative to the Gaussian-*n* methods. *J Chem Phys*. **2006**, *124*, 114104-1–17.

[9] Chan, B.; Collins, E. M.; Raghavachari, K., Applications of isodesmic-type reactions for computational thermochemistry, *WIRES Comput Mol. Sci*. 2020, 11, e1501.

[10] Mardirossian, N.; Head-Gordon, M., Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017,** *115*, 2315-2372.

[11] See for example: Pieniazek, S. N.; Clemente, F. R.; Houk, K. N., Sources of error in DFT computations of C-C bond formation thermochemistries: pi → sigma transformations and error cancellation by DFT methods. *Angew. Chem. Intl. Ed.* **2008**, *47*, 7746.

[12] Ma, Q.; Werner, H.-J., Scalable Electron Correlation Methods. 7. Local open-shell coupled-cluster methods using pair natural orbitals: PNO-RCCSD and PNO-UCCSD. *J. Chem. Theory Comput*. **2020**, *16*, 3135−3151.

[13] Ma, Q.; Werner, H.-J., Scalable Electron Correlation Methods. 8. Explicitly Correlated Open-Shell Coupled-Cluster with Pair Natural Orbitals PNO-RCCSD(T)-F12 and PNO-UCCSD(T)-F12. *J. Chem. Theory Comput*. **2021**, *17*, 902–926.

[14] Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An Improved Linear Scaling Perturbative Triples Correction for the Domain Based Local PairNatural Orbital Based Singles and Doubles Coupled Cluster Method [DLPNO-CCSD(T)]. J. Chem. Phys. 2018, 148, 011101.

[15] For a comprehensive review of fragmentation methods, see: Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V., Fragmentation methods: A route to accurate calculations on large systems *Chem. Rev.* **2011**, *112*, 632-672.

[16] Raghavachari, K.; Saha, A., Accurate composite and fragment-based quantum chemical models for large molecules, *Chem. Rev*., **2015**, *115*, 5643–5677.

[17]Sahu, N.; Gadre, S. R., Molecular tailoring approach: A route for ab initio treatment of large clusters. *Acc. Chem. Res.* **2014,** *47* (9), 2739-2747.

[18] Collins, M. A.; Bettens, R. P. A., Energy-based molecular fragmentation methods. *Chem. Rev.* **2015,** *115* (12), 5607-5642.

[19] He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z., Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.* **2014,** *47* (9), 2748-2757.

[20] Li, S.; Li, W.; Ma, J., Generalized energy-based fragmentation approach and its applications to macromolecules and molecular aggregates. *Acc. Chem. Res.* **2014,** *47* (9), 2712-2720.

[21] Richard, R. M.; Herbert, J. M., A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012,** *137* (6), 064113.

[22] Dahlke, E. E.; Truhlar, D. G., Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput.* **2007,** *3* (1), 46-53.

[23] Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J., Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* **2012,** *14* (21), 7578-7590.

[24] Ramabhadran, R. O.; Raghavachari, K., Theoretical thermochemistry for organic molecules: Development of the generalized Connectivity-Based Hierarchy. *J. Chem. Theory Comput.* **2011**, *7*, 2094-2103.

[25] Sengupta, A. and Raghavachari, K., Solving the density functional conundrum: Elimination of systematic errors to derive accurate reaction enthalpies of complex organic reactions. *Org. Lett.* **2017**, *19*, 2576–2579.

[26] Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A., Molecular orbital theory of electronic structure of organic compounds .5. Molecular theory of bond separation. *J. Am. Chem. Soc.* **1970,** *92*, 4796-4801.

[27] George, P.; Trachtman, M.; Bock, C. W.; Brett, A. M., An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. *Theor. Chim. Acta* **1975,** *38*, 121-129.

[28] Wheeler, S. E.; Houk, K. N.; Schleyer, P. V. R.; Allen, W. D., A hierarchy of homodesmotic reactions for thermochemistry. *J. Am. Chem. Soc.* **2009,** *131*, 2547-2560.

[29] Sengupta, A.; Raghavachari, K., Prediction of accurate thermochemistry of medium and large sized radicals using Connectivity-Based Hierarchy (CBH). *J. Chem. Theory Comput.* **2014,** *10*, 4342-4350.

[30] Collins, E. M.; Sengupta, A.; AbuSalim, D. I.; Raghavachari, K., Accurate thermochemistry for organic cations via error cancellation using Connectivity-Based Hierarchy. *J. Phys. Chem. A* **2018,** *122*, 1807-1812.

[31] Debnath, S.; Sengupta, A.; Raghavachari, K., Eliminating systematic errors in DFT via Connectivity-Based Hierarchy: Accurate bond dissociation energies of biodiesel methyl esters. *J. Phys. Chem. A* **2019,** *123*, 3543-3550.

[32] Thapa, B.; Raghavachari, K., Accurate pK_a evaluations for complex bio-organic molecules in aqueous media. *J. Chem. Theory. Comput.* **2019,** *15*, 6025-6035.

[33] Maier, S.; Thapa, B.; Raghavachari, K., G4 accuracy at DFT cost: Unlocking accurate redox potentials for organic molecules using systematic error cancellation. *Phys. Chem. Chem. Phys.* **2020,** *22*, 4439-4452.

[34] Ramachandran, R. O.; Raghavachari, K. Connectivity-Based Hierarchy for theoretical thermochemistry: Assessment using wave function-based methods. *J. Phys. Chem. A*, **2012**, *116*, 7531-7537.

[35] Ramabhadran, R. O.; Raghavachari, K., Extrapolation to the Gold-Standard in quantum chemistry: Computationally efficient and accurate CCSD(T) energies for large molecules using an automated thermochemical hierarchy. *J. Chem. Theory Comput.* **2013**, *9*, 3986-3994.

[36] Deev, V.; Collins, M. A., Approximate *ab initio* energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

[37] Bettens, R. P. A; Lee, A. M., A New Algorithm for Molecular Fragmentation in Quantum Chemical Calculations. *J. Phys. Chem. A* **2006**, *110*, 8777-8785.

[38] Mayhall, N. J.; Raghavachari, K., Molecules-in-molecules: An extrapolated fragment-based approach for accurate calculations on large molecules and materials. *J Chem Theory Comput* **2011,** *7*, 1336-1343.

[39] Ramabhadran, R. O.; Raghavachari, K., The successful merger of theoretical thermochemistry with fragment-based methods in quantum chemistry. *Acc. Chem. Res*. **2014**, *47*, 3596-3604.

[40] Trinajstić, N., *Chemical graph theory*. 2nd ed.; CRC Press: Boca Raton, 1992; p 322 p.

[41] Kromann, J. C.; Welford, A.; Christensen, A. S.; Jensen, J. H., Random versus systematic errors in reaction enthalpies computed using semiempirical and minimal basis set methods. *ACS Omega* **2018,** *3*, 4372-4377.

[42] Liu, J.; Wang, R. W.; Tian, J.; Zhong, K.; Nie, F. D.; Zhang, C. Y., Calculation of gas-phase standard formation enthalpy via ring-preserved Connectivity-Based Hierarchy and automatic bond separation reaction platform. *Fuel* **2022,** *327*.

[43] Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K., Accurate and computationally efficient prediction of thermochemical properties of biomolecules using the generalized Connectivity-Based Hierarchy. *J. Phys. Chem. B*, **2014**, *118*, 9631-9643.

[44] Ramabhadran, R. O, Sengupta, A.; Raghavachari, K., Application of the generalized Connectivity-Based Hierarchy to biomonomers: Enthalpies of formation of cysteine and methionine. *J. Phys. Chem. A*, **2013**, *117*, 4973-49801.

[45] Chai, J.-D.; Head-Gordon, M., Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. *Phys. Chem. Chem. Phys*. **2008**, *10*, 6615.

[46] Zhao, Y.; Truhlar, D.G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor Chem Acc.* **2008**, *120,* 215–241.

[47] Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange.*J. Chem. Phys*. **1993**, *98*, 5648.

[48] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem*. **1994**, *98*, 11623.

[49] Mezei, P. D.; Csonka, G. I.; Kállay, M., Accurate Diels-Alder reaction energies from efficient density functional calculations. *J. Chem. Theory Comp.* **2015**, *11*, 2879.

[50] McKee, W. C.; Schleyer, P. v. R., Correlation effects on the relative stabilities of alkanes. *J. Am. Chem. Soc.* **2013**, *135*, 13008.

[51] Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys*. **1996**, *105*, 9982.

[52] Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C., Dispersion-Corrected Mean-Field Electronic Structure Methods, *Chem. Rev*. **2016**, *116*, 5105.

[53] Grimme, S.; Neese, F., Double-hybrid density functional theory for excited electronic states of molecules, *J. Chem. Phys.* **2007**, 127(15):154116.

[54] Oyeyemi, V. B.; Dieterich, J. M.; Krisiloff, D. B.; Tan, T.; Carter, E. A., Bond dissociation energies of C_{10} and C_{18} methyl esters from local multireference Average-Coupled Pair Functional theory. *J. Phys. Chem. A* **2016**, *120*, 4025-4036.

[55] Schüring, J.; Schulz, H. D.; Fischer, W. R.; Böttcher, J.; Duijnisveld, W. H., *Redox: Fundamentals, processes and applications*. Springer Science & Business Media: 2013.

[56] Marenich, A. V.; Ho, J. M.; Coote, M. L.; Cramer, C. J.; Truhlar, D. G., Computational electrochemistry: Prediction of liquid-phase reduction potentials. *Phys. Chem. Chem. Phys.* **2014,** *16*, 15068-15106.

[57] Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R., Calculation of one-electron redox potentials revisited. Is it possible to calculate accurate potentials with density functional methods? *J. Phys. Chem. A* **2009,** *113*, 6745-6750.

[58] Alongi, K. S.; Shields, G. C., Theoretical calculations of acid dissociation constants: A review article. *Annu. Rep. Comput. Chem.* **2010,** *6*, 113-138.

[59] Seybold, P. G.; Shields, G. C., Computational estimation of pKa values. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015,** *5*, 290-297.

[60] Thapa, B.; Schlegel, H. B., Theoretical calculation of pKa's of selenols in aqueous solution using an implicit solvation model and explicit water molecules. *J. Phys. Chem. A* **2016,** *120*, 8916-8922.

[61] Thapa, B.; Schlegel, H. B., Improved pKa prediction of substituted alcohols, phenols, and hydroperoxides in aqueous medium using density functional theory and a cluster-continuum solvation model. *J. Phys. Chem. A* **2017,** *121*, 4698-4706.

[62] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009,** *113*, 6378-6396.

[63] Montgomery Jr., J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A., A complete basis set model chemistry. VI. Use of density functional geometries and frequencies, *J. Chem. Phys.*, **1999**, *110*, 2822-27.

[64] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J Phys Chem Lett* **2015,** *6* (12), 2326-2331.

[65] Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys Rev Lett* **2012,** *108* (5).

[66] Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J., Constant size descriptors for accurate machine learning models of molecular properties. *J Chem Phys* **2018,** *148* (24).

[67] Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J., Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J Phys Chem Lett* **2017,** *8* (12), 2689-2694.

[68] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J Chem Theory Comput* **2015,** *11*, 2087-2096.

[69] Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A., Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J Chem Theory Comput* **2019,** *15*, 1546-1559.

[70] Collins, E. M.; Raghavachari, K., Effective molecular descriptors for chemical accuracy at DFT cost: Fragmentation, error-cancellation, and machine learning. *J. Chem. Theory Comput*. **2020**, *16*, 4938-4950.

[71] Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R., SchNet – A deep learning architecture for molecules and materials. *J Chem Phys* **2018,** *148*, 241722

[72] Chen, H. M.; Engkvist, O.; Wang, Y. H.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. *Drug Discov Today* **2018,** *23*, 1241-1250.

[73] Xu, J., Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences* **2019,** *116*, 16856.

[74] Collins, E. M.; Raghavachari, K., Fragmentation-based graph embedding framework for QM/ML, *J. Phys. Chem. A*. **2021**, *125*, 6872-6880.

[75] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials* **2019,** *31*, 3564-3572.

[76] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P., Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **2016,** *30*, 595-608.

[77] Ward, L.; Blaiszik, B.; Foster, I.; Assary, R. S.; Narayanan, B.; Curtiss, L., Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations. *MRS Commun* **2019,** *9*, 891-899.

[78] Waite, S. L.; Chan, B.; Karton, A.; Page, A. J., Accurate thermochemical and kinetic stabilities of C_{84} isomers. *J Phys Chem A* **2018,** *122*, 4768-4777.

[79] Elliott, S. N.; Keceli, M.; Ghosh, M. K.; Somers, K. P.; Curran, H. J.; Klippenstein, S. J., High-accuracy heats of formation for alkane oxidation: From small to large via the automated CBH-anl method. *J Phys Chem A* **2023**, *127*, 1512-1531.

[80] Collins, E. M.; Raghavachari, K., Stepping-Stone CBH: Benchmark and Application of a Multilayered Isodesmic-based Correction Scheme. **2023**, to be published.

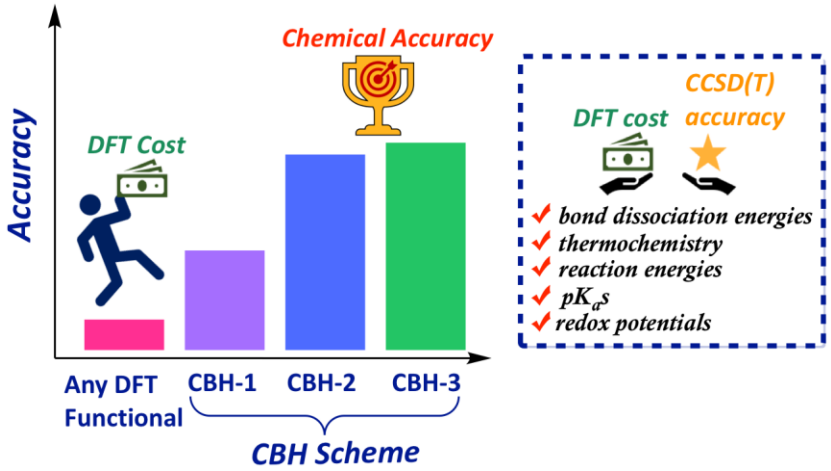
[81] Maier, S.; Collins, E. M.; Raghavachari, K., Quantitative prediction of vertical ionization potentials from DFT via a graph network-based Delta machine learning model incorporating electronic descriptors. *J. Phys. Chem. A*. (**2023**), in press.

[82] Chan, B.; Karton, A., Polycyclic aromatic hydrocarbons: from small molecules through nano-sized species towards bulk graphene. Phys. Chem. Chem. Phys. **2021**, *23*, 17713.

[83] Thapa, B.; Erickson, J.; Raghavachari, K., Quantum Mechanical Investigation of Three-Dimensional Activity Cliffs using the Molecules-in-Molecules Fragmentation-Based Method, *J. Chem. Inf. Mod*. **2020**, *55*, 2932.

[84] Chandy, S. K.; Raghavachari, K., Accurate and Cost-Effective NMR Chemical Shift Predictions for Nucleic Acids Using a Molecules-in-Molecules Fragmentation-Based Method, *J. Chem. Theory Comput*. **2023**, *19*, 544.

TOC Figure


Chemical Accuracy
Accuracy
DFT Cost
CCSD(T)
DFT cost
accuracy
bond dissociation energies
thermochemistry
reaction energies
pKas
redox potentials
Any DFT Functional
CBH-1
CBH-2
CBH-3
CBH Scheme