Ab initio thermochemistry using optimal-balance models with isodesmic corrections: The ATOMIC protocol

Dirk Bakowies

Laboratory of Physical Chemistry, ETH Zürich, CH 8093 Zürich, Switzerland

(Received 8 December 2008; accepted 30 January 2009; published online 13 April 2009)

A theoretical composite approach, termed ATOMIC for Ab initio Thermochemistry using Optimal-balance Models with Isodesmic Corrections, is introduced for the calculation of molecular atomization energies and enthalpies of formation. Care is taken to achieve optimal balance in accuracy and cost between the various components contributing to high-level estimates of the fully correlated energy at the infinite-basis-set limit. To this end, the energy at the coupled-cluster level of theory including single, double, and quasiperturbational triple excitations is decomposed into Hartree–Fock, low-order correlation (MP2, CCSD), and connected-triples contributions and into valence-shell and core contributions. Statistical analyses for 73 representative neutral closed-shell molecules containing hydrogen and at least three first-row atoms (CNOF) are used to devise basis-set and extrapolation requirements for each of the eight components to maintain a given level of accuracy. Pople’s concept of bond-separation reactions is implemented in an ab initio framework, providing for a complete set of high-level precomputed isodesmic corrections which can be used for any molecule for which a valence structure can be drawn. Use of these corrections is shown to lower basis-set requirements dramatically for each of the eight components of the composite model. A hierarchy of three levels is suggested for isodesmically corrected composite models which reproduce atomization energies at the reference level of theory to within 0.1 kcal/mol (A), 0.3 kcal/mol (B), and 1 kcal/mol (C). Large-scale statistical analysis shows that corrections beyond the CCSD(T) reference level of theory, including coupled-cluster theory with fully relaxed connected triple and quadruple excitations, first-order relativistic and diagonal Born–Oppenheimer corrections can normally be dealt with using a greatly simplified model that assumes thermonuclear bond-separation reactions and that reduces the estimate of these corrections to the simple task of adding up bond increments. Preliminary validation with experimental enthalpies of formation using the subset of neutral closed-shell (HCNOF) species contained in the G3/99 test set indicates that the ATOMIC protocol performs slightly better than the popular G3 approach. The newly introduced protocol does not require empirical calibration, however, and it is still efficient enough to be applied routinely to molecules with 10 or 20 nonhydrogen atoms. © 2009 American Institute of Physics. [DOI: 10.1063/1.3089241]

I. INTRODUCTION

A substantial number of thermochemical protocols have been suggested in the literature, which are usually judged by their ability to reproduce experimental thermochemical data such as standard enthalpies of formation and atomization energies. In order to achieve chemical accuracy (1–2 kcal/mol) or better, one either needs to resort to very elaborate fully ab initio treatments, severely limiting the range of possible applications to small molecules, or use ab initio methods of reasonably high accuracy whose residual errors are so systematic that they can be absorbed by empirical calibration. Most of the methods suggested so far are compound (or composite) approaches which treat the lower orders of the N-particle problem with larger basis sets than the higher orders, taking advantage of the general observation that energies associated with higher-order excitations usually converge faster to the complete basis-set limit.

Probably the earliest among the fully ab initio schemes is the focal point analysis introduced by East and Allen which systematically probes for (incremental) convergence of both the one- and N-particle descriptions toward the “focal point” of a theoretically exact treatment. Martin and co-workers developed a series of standard protocols termed Weizmann-n [W1,7–9 W2,7–9 W3,10 and W4 (Ref. 11)] which are based on coupled-cluster theory with single and double excitations and quasiperturbational treatments of connected triple excitations [CCSD(T) (Refs. 12 and 13)], extrapolated to the complete basis-set limit and further corrected for effects of even higher-order excitations (up to CCSDTQP in W4), for inner-shell and relativistic contributions (W1–W4) and for non-Born–Oppenheimer (W4) effects. The HEAT protocol is similar in layout to the most advanced variants of Weizmann theory and geared toward benchmark accuracy for small molecules. Less standardized high-accuracy approaches have been presented by Dixon, Feller, and co-workers and used in many applications to first- and second-row chemistry. Klopper and various collaborators re-
ported high-accuracy calculations of atomization energies and enthalpies of formation,\textsuperscript{28-27} often replacing complete basis-set extrapolations with more elaborate explicitly correlated R12 calculations.\textsuperscript{28,29}

The aforementioned approaches achieve high accuracy without application of empirical corrections but are necessarily limited to fairly small molecules. Gaussian-1 theory,\textsuperscript{30} introduced by Pople et al, already in 1989, was the first of a series of general-purpose models applicable to larger molecules. It starts with MP4 (fourth order Møller–Plesset perturbation theory)\textsuperscript{31} calculations using medium-sized basis sets, probes separately for the effects of adding diffuse basis functions and higher polarization functions, adds “post-MP4” corrections estimated at the QCISD(T) level,\textsuperscript{32} and assumes that all these corrections are approximately additive. Additional higher-level corrections (HLCs) for paired and unpaired electrons are added as parameters to correct for the remaining deficiencies of the model. HLC parameters were initially adjusted to reproduce results for the hydrogen atom and molecule but calibrated empirically in the more refined protocols of G2,\textsuperscript{33} G3,\textsuperscript{34} and G4.\textsuperscript{35} While G4 theory is still very new, both G2 and G3 have become very successful and popular, and a large number of alternative formulations have been tested, replacing, e.g., QCISD(T) with CCSD(T) energies,\textsuperscript{36,37} using less expensive MP2 (Ref. 38) or MP3 (Ref. 39) estimates of basis-set extension effects, replacing HLC parameters by multiplicative scaling factors,\textsuperscript{40} or employing more accurate B3LYP geometries and zero-point energies.\textsuperscript{37,41}

The CBS models of Petersson and co-workers\textsuperscript{42} are similar in spirit, but they include complete basis-set extrapolations based on pair natural orbital expansions of the second order correlation energy. Following the proposal of the accurate but very demanding CBS-QCI/APNO procedure,\textsuperscript{33,44} a number of less expensive variants were introduced, including a series of models of increasing accuracy and computational complexity, termed CBS-4, CBS-q, and CBS-Q,\textsuperscript{45} more accurate models based on B3LYP geometries and frequencies (CBS-QB3),\textsuperscript{46} and further refinements ensuring proper orbital localization necessary to enforce size consistency and improve overall robustness.\textsuperscript{47} Extensive validation of the CBS series of models has been reported in the literature.\textsuperscript{48,49}

Based on earlier suggestions to scale either the dynamical\textsuperscript{50} or total\textsuperscript{51,52} correlation energy of more approximate treatments to account for shortcomings in the one- and \textit{N}-particle description, Truhlar and co-workers introduced a number of parametrized multicoefficient methods, mixing scaled energies of several different electron correlation\textsuperscript{53-56} and density functional schemes\textsuperscript{57,58} and aiming at high accuracy for explorations of entire potential energy surfaces. Some of the protocols were designed as multicoefficient versions of G2 (Ref. 59) and G3 theories,\textsuperscript{60} and related parametrized multilevel protocols have been reported by Li et al.\textsuperscript{61,62} Finally, Grimme recently suggested spin-component scaled MP2 theory\textsuperscript{63} with additional calibrated HLC parameters as an economical and fairly accurate alternative for the evaluation of enthalpies of formation.\textsuperscript{64}

Naturally, the use of empirical parameters is unsatisfactory from a theoretical point of view; no matter whether they are scaling factors or HLCs based on counts of paired and unpaired electrons, they establish agreement with experiment in a somewhat uncontrolled fashion. Avoiding calibration is a goal that has been expressed repeatedly,\textsuperscript{65,66} but only recently has a parameter-free approach been introduced by DeYonker et al.\textsuperscript{67,68} that is somewhat intermediate in computational complexity between \textit{Gn} theories\textsuperscript{30,33-35} and the very high-level protocols used by Dixon and co-workers.\textsuperscript{17-24}

Here we introduce a thermochemical protocol that makes consistent use of bond-separation reactions\textsuperscript{69,70} (BSRs) to allow for high-level treatments of bond dissociation embedded in lower-level treatments of the perturbation caused by the chemical environment. High-level calculations are necessary only for the small parent molecules and simple hydrides appearing in the BSRs. Results can then be mapped onto a system of corrective bond increments and stored for later retrieval to improve on low-level treatments for large molecules. The approach is fully \textit{ab initio} in the sense that it does not require empirical calibration or any reference to experiment to guide development or help select from various alternative options. Basis-set requirements are instead carefully assessed for each component contributing to the atomization energy, thus avoiding or at least reducing the risk of uncontrolled, fortuitous error cancellation. The approach has been developed and tested for larger neutral closed-shell molecules containing first-row atoms (C, N, O, F) and hydrogen and is applicable to virtually every type of bonding situation encountered in the chemistry of these elements, thus considerably enhancing the scope of BSRs which was previously limited by the availability of accurate experimental reference data.

The theoretical concept is outlined in Sec. II, and computational details are given in Sec. III. The computational model used to approximate complete basis-set limits of CCSD(T) atomization energies is discussed in Sec. IV and validated in Sec. V. Corrections beyond the CCSD(T) approximation, relativistic, and diagonal Born–Oppenheimer corrections are dealt with in Sec. VI. All analyses focus on equilibrium atomization energies \(E_{\text{A,e}}\), i.e., the negative energy difference between a molecule in its equilibrium geometry and fully dissociated into the constituent atoms. Comparison with experiment requires the additional evaluation of zero-point and thermal energies. While an extensive validation with experimental enthalpies of formation will be published elsewhere,\textsuperscript{71} Sec. VII already provides a preliminary statistical assessment.

II. THEORETICAL CONCEPT

Following established high-level procedures [Weizmann,\textsuperscript{7,10,11} HEAT (Refs. 14–16)], we try to reproduce each contribution to the equilibrium atomization energy \(E_{\text{A,e}}\) as accurately as possible, thereby allowing for simplification wherever possible. The various contributions may be summarized as follows:
(1) the complete basis-set limit of the atomization energy at the CCSD(T) level of theory with all electrons correlated;
(2) higher-order excitations (CCSDT, CCSDTQ);
(3) relativistic and non-Born–Oppenheimer effects.

A reference level of theory is defined based on CCSD(T)(full) energies evaluated with cc-pCVQZ and cc-pCV5Z basis sets and extrapolated to the complete basis-set limit. Further corrections are applied to the dominant contributions (Hartree–Fock, valence-shell CCSD correlation) by replacing (Q5) with (S6) extrapolated energies. Based on previous experience, this level of theory is expected to reproduce the complete basis-set limit to within a few tenths of a millihartree. An energy decomposition analysis (Sec. IV A) is then used to assess in a systematic manner which components of the CCSD(T)(full) energy may be evaluated with smaller basis sets without sacrificing accuracy too much. Such analysis further allows us to balance errors in the various components and to assess systematic (or fortuitous) error cancellation between components. Not surprisingly, basis-set requirements are found to be fairly high, if reasonable accuracy is sought.

We use the concept of isodesmic reactions in their original formulation [bond-separation reactions (BSRs)] to supply very high-level \( \textit{ab initio} \) data for the atomization of each bond and thus improve the accuracy of lower-level quantum-chemical predictions for atomization energies of larger molecules (Sec. IV C). This procedure only requires that a (Lewis) valence structure can be drawn, and we note that the isodesmic reaction is uniquely defined by the valence structure (Sec. IV D). The BSR scheme can easily be mapped onto a system of corrective bond increments (Sec. IV E) which, once evaluated for the reference level of theory, may be used to improve the results of production calculations for larger molecules (Secs. V B–V D).

BSRs are then also used (Sec. VI) to estimate further corrections toward full CI (CCSDT and CCSDTQ corrections), as well as relativistic (mass-velocity and Darwin terms) and non-Born–Oppenheimer corrections (diagonal Born–Oppenheimer terms) to atomization energies. All these contributions are treated under the greatly simplifying assumption that the respective BSRs are thermoneutral, rendering the evaluation of corrections a trivial task of summing up bond increments. The validity of the underlying assumption is examined in Secs. VI A–VI C. Finally one needs to consider spin-orbit corrections, as regular (nonrelativistic) electronic structure calculations refer to weighted averages over states resulting from different couplings between spin and orbital angular momenta. Luckily, no spin-orbit couplings occur for closed-shell molecules, and values for atoms may be obtained from experimental electronic spectra, and they have been listed, e.g., by Curtiss et al.

III. COMPUTATIONAL DETAILS

Geometry optimizations were performed at the RI-MP2/cc-pVTZ level of theory using the resolution of identity (RI) approximation with optimized auxiliary basis sets as implemented in the TURBOMOLE suite of programs (version 5.5, 2002). Although higher-level [CCSD(T)] geometry optimizations would have been preferable, they are impractical for molecules larger than a few atoms. The chosen level is still fairly accurate, typically reproducing bond lengths to within 0.01 Å (or better) and bond angles to within about 1°. All optimized geometries were confirmed to be true minima by standard force-constant analysis. Single-point calculations at the MP2 (Ref. 31) and CCSD(T) levels of theory were performed using the MOLPRO program and the correlation-consistent basis sets cc-pVQZ (Refs. 84–88) ranging from \( X=\text{D} \) to \( X=\text{S} \), for valence-shell treatments (“frozen core”) and the corresponding polarized core-valence basis sets (Ref. 89) cc-pCVXZ (\( X=\text{D} \) to \( X=\text{S} \)) for calculations with all electrons correlated (“full”). Atomization energies are based on regular restricted Hartree-Fock (RHF) wave functions for closed-shell species (all molecules) and on restricted open-shell Hartree-Fock (ROHF) reference wave functions for open-shell species (the atoms C, N, O, and F, see Table S1), for which restricted open-shell MP2 (ROMP2) (Ref. 91) and unrestricted open-shell CCSD(T) (URCCSD(T)) (Refs. 92 and 93) energies are evaluated. Extrapolation procedures are discussed in Sec. IV B, and details on calculations of electron correlation energies beyond CCSD(T), relativistic, and diagonal Born–Oppenheimer corrections are given in Sec. VI.

IV. COMPUTATIONAL MODEL

In this section we discuss details of the CCSD(T) energy decomposition analysis, of procedures to extrapolate the CCSD(T) energy to the complete basis-set limit, and of the way BSRs are used to include high-level \( \textit{ab initio} \) data for refinement of low-level atomization energies.

A. Energy decomposition

We define Hartree-Fock (HF) energies,
\[
E_{\text{HF}}(X) = \text{HF/cc-pVX}_Z, \tag{1}
\]
valence-shell contributions to the correlation energy,
\[
\Delta_{\text{MP2}}(X) = \text{MP2/cc-pVX}_Z - \text{HF/cc-pVX}_Z, \tag{2}
\]
\[
\Delta_{\text{CCSD}}(X) = \text{CCSD/cc-pVX}_Z - \text{MP2/cc-pVX}_Z, \tag{3}
\]
\[
\Delta_{\text{CCSDT}}(X) = \text{CCSD(T)/cc-pVX}_Z - \text{CCSD/cc-pVX}_Z, \tag{4}
\]
and inner-shell (core) contributions,
\[
\Delta_{\text{MP2}}^\text{core}(X) = \text{MP2(full)/cc-pCVX}_Z - \text{HF/cc-pCVX}_Z - \Delta_{\text{MP2}}(X), \tag{5}
\]
\[
\Delta_{\text{CCSD}}^\text{core}(X) = \text{CCSD(full)/cc-pCVX}_Z - \text{MP2(full)/cc-pCVX}_Z - \Delta_{\text{CCSD}}(X), \tag{6}
\]
\[ \Delta_{\text{CCSD(T)}}(X_{8}) = \text{CCSD(T)}(\text{full})/cc-pCVX_{8}Z \]
\[ - \text{CCSD(full)/cc-pCVX}_{8}Z \]
\[ - \Delta_{\text{CCSD(T)}}(X_{8}), \]
such that both the CCSD(T)/cc-pVXZ (frozen core) energy,
\[ \text{CCSD(T)/cc-pVXZ} \]
\[ = E_{\text{HF}}(X) + \Delta_{\text{MP2}}(X) + \Delta_{\text{CCSD}}(X) + \Delta_{\text{CCSD(T)}}(X) \]
and the CCSD(T)(full)/cc-pCVXZ energy with all electrons correlated,
\[ \text{CCSD(T)(full)/cc-pCVXZ} \]
\[ = E_{\text{HF}}(X) + \Delta_{\text{MP2}}(X) + \Delta_{\text{CCSD}}(X) + \Delta_{\text{CCSD(T)}}(X) \]
\[ + \Delta_{\text{core}}(X) + \Delta_{\text{MP2}}(X) + \Delta_{\text{CCSD}}(X) + \Delta_{\text{CCSD(T)}}(X), \]
are recovered if all cardinal numbers \( X_{1}=X_{2}=\cdots=X_{8}=X \) are chosen to be equal. If any of the cardinal numbers is chosen to be smaller than \( X \), one obtains an approximation to the CCSD(T)(full)/cc-pVXZ energy which we generally denote as \([X_{1}X_{2}X_{3}X_{4}|X_{5}X_{6}X_{7}X_{8}]\). For better legibility Arabic numerals are used throughout instead of the common symbols \((D, T, Q)\) for \(X=2, 3, 4\). Boldface numerals with top bar \(X\) energy components obtained through complete basis-set extrapolation based on results for cardinal numbers \(X_{n}1\) and \(X_{n}\). Components omitted completely are indicated by a dash. A MP2/cc-pVTZ energy will thus be written as \([33-\ldots-\ldots]\), and an extrapolated CCSD(full)/cc-pCV(TQ)Z energy as \([444-444]\).

The energy partitioning scheme affords a term \(\Delta_{\text{core}}(X_{8})\) which simply reflects the difference in HF energies obtained with basis sets designed for electron correlation treatments with (cc-pCVXZ) and without (cc-pVXZ) core correlation. \(\Delta_{\text{core}}(X_{8})\) contributions to total energies are negative, as cc-pCVXZ basis sets contain additional basis functions, but they are expected to become smaller as \(X_{8}\) increases and finally vanish for \(X_{8} \rightarrow \infty\).

The quantity of main interest is the atomization energy, i.e., the negative difference in energies between a molecule and its constituent atoms. Atomization energies are decomposed in precisely the same way as total energies, and the same nomenclature is used.

**B. Extrapolation procedures**

Complete basis-set estimates of the HF energy components \(E_{X}=E_{\text{HF}}(X)\) and \(E_{X} = \Delta_{\text{core}}(X)\) are obtained from calculations with basis sets of cardinal numbers \(X-1\) and \(X\) using the exponential extrapolation formula of Halkier et al.,
\[ E_{X} = E_{X} + \frac{1}{\exp(\alpha) - 1} (E_{X} - E_{X-1}), \]
and the calibrated value of \(\alpha=1.63\). Although the authors point out that this extrapolation formula may not always work satisfactorily for large basis sets \((X>4)\), we also note that HF contributions to atomization energy differences \(E_{X}-E_{X-1}\) are small in these cases \((\approx 1.0 \text{ mE}_{h} \text{ for } X=5, \quad \approx 0.3 \text{ mE}_{h} \text{ for } X=6\) using our test set defined in Table S2\(^{90}\)), and that less than optimal extrapolations may be tolerated, particularly for \(X=6\). Quite similar exponents for both two- and three-point extrapolations have recently been reported in a calibration study of Jensen.\(^{95}\) We note in passing that a more refined two-point extrapolation formula has been worked out by Karton and Martin,\(^{96}\) which effectively maps onto Eq. (6) with basis-set dependent exponent \(\alpha\). It was calibrated for large augmented basis sets, however, and is not considered here.

The power-law formula
\[ E_{X} = E_{X} + \frac{X^\beta}{(X-1)^\beta - X^\beta} (E_{X} - E_{X-1}) \]
is used for extrapolations of the correlation energy components \(E_{X} = \Delta_{\text{MP2}}(X)\) (the MP2 correlation energy), \(E_{X} = \Delta_{\text{CCSD}}(X)\) (the CCSD correlation energy), \(E_{X} = \Delta_{\text{CCSD(T)}}(X)\) (the triples contribution), and the corresponding inner-shell contributions \(E_{X} = \Delta_{\text{core}}^{\text{H}}(X), E_{X} = \Delta_{\text{core}}^{\text{H}}(X)\)
\[ + \Delta_{\text{core}}(X), \text{ and } E_{X} = \Delta_{\text{core}}(X) \]
are extrapolated separately. This simplification is motivated by the observation that inner-shell contributions to atomization energies are almost two orders of magnitude smaller than valence-shell contributions, suggesting that it is sufficient to calibrate extrapolation procedures for the dominant terms only. Similar remarks apply to the perturbative triples correction to atomization energies which are about 10–20 times smaller than CCSD correlation contributions. The same exponent as for CCSD \((\beta=3.15)\) is used for the terms \(\Delta_{\text{CCSD(T)}}(X)\) and \(\Delta_{\text{core}}(X)\), in this case, however, also for (DT) extrapolations as limited analysis has indicated that triples contributions to atomization energies show faster convergence for small \(X\) (data not shown).

**C. Isodesmic reactions**

Reactions are called isodesmic if they conserve the number of bonds of each type. Following the introduction of Pople and co-workers,\(^{69,70,73,74}\) isodesmic reactions have often been used in thermochemistry to obtain accurate enthalpies of formation from fairly low-level quantum-chemical approaches. The basic idea is that errors due to the incompleteness of the one- and \(N\)-particle descriptions will be systematic, quite characteristic of a particular bond type, and thus cancel to a large extent, making the evaluation of isodesmic reaction energies significantly more accurate than the evaluation of atomization energies. To use an example, one may obtain the enthalpy of formation for oxirene from the enthalpy of reaction computed for oxirene+ethane.
→ oxirene + ethylene, if one refers to experimentally well-established data for ethane, oxirene, and ethylene.

One problem often discussed is that there are numerous isodesmic reactions relating known enthalpies of formation of auxiliary molecules to the unknown value for the molecule in question, making the reference a matter of choice and the evaluation nonunique. One may, however, refer to BSRs (Ref. 69) [see Fig. 1, reaction C], which map the Lewis valence structure of a molecule in a unique way and generate one parent molecule for each of the nonhydrogen bonds of the molecule in question. Proper stoichiometric balance is achieved by adding the required number of simple hydrides on the reactant side. Still, BSRs have not been considered as often in theoretical work as other types of isodesmic reactions (see Ref. 97 for an example), because experimental data for parent molecules and simple hydrides (collectively referred to as BSR prototypes) are often unavailable or inaccurate, and because larger numbers of auxiliary molecules enter the reaction (six instead of three in the example discussed).

These shortcomings are largely avoided if theoretical rather than experimental reference data are used for BSR prototypes. By construction, these molecules are small, involving only up to two nonhydrogen atoms, and high-level quantum-chemical methods may be used to calculate either their enthalpy of formation or atomization energy very precisely. The approach presented here is based on atomization energies, and for the moment we are only concerned with CCSD(T) atomization energies at the complete basis-set limit. Further corrections (relativistic, non-Born–Oppenheimer, higher-order excitations) are dealt with later (Sec. VI). The following two sections demonstrate in a general way how any given valence structure can be converted uniquely into a BSR and mapped onto a system of corrective bond increments, thus greatly facilitating the computational implementation of our approach.

D. Construction of bond-separation reactions (BSRs)

BSRs may be generated from a sequence of isodesmic reactions, each of them cleaving one of the bonds appearing between the nonhydrogen atoms of a molecule $M$. Let us focus on the $n$-tuple bond $i$-a $X^a_i Y^{k-a}_n$ between nonhydrogen atoms $X$ and $Y$ and denote by superscripts their total numbers of free valencies. The molecule may be written as

$$M = R_{(i-a-n)} H_a (i-a X^a_i Y^{k-a}_n) H_b R_{(k-b-n)}$$

enclosing in curly brackets the number of valencies satisfied by residues $R$ attached to $X$ and $Y$. The formal reaction [Eq. (8), Fig. 1]

$$M + XH_i + YH_k \rightarrow H_{(i-a)} (i-a X^a_i Y^{k-a}_n) H_{(k-b)} + M_1 + M_2,$$

$$M = R_{(i-a-n)} H_a (i-a X^a_i Y^{k-a}_n) H_b R_{(k-b-n)}$$

$$M_1 = R_{(i-a-n)} XH_{i-a-n},$$

$$M_2 = R_{(k-b-n)} YH_{k-b-n}$$

cleaves molecule $M$ at its bond $i$-a $X^a_i Y^{k-a}_n$ into dissociation products $M_1$ and $M_2$ and simultaneously reassembles the bond from the corresponding simple hydrides to form the parent molecule $H_{(i-a)} (i-a X^a_i Y^{k-a}_n) H_{(k-b)}$. The reaction is constructed such that it conserves the number of bonds of each type; hence it is isodesmic. In general two new reaction
products are formed ($M_1$ and $M_2$); however, they form a single chain if the cleaved bond is part of a ring system. If either $i-a=n$ or $k-b=n$, then one of the dissociation products equals the corresponding simple hydride on the reactant side, and if both conditions hold, then the reaction is unproductive as the product and reactant sides become identical.

Following the initial bond cleavage, the same isodesmic scheme, Eq. (8), is applied to arbitrary nonhydrogen bonds appearing in $M_1$ and $M_2$, producing a new set of reaction products $M_1 \cdots M_6$. The iterative process is terminated when only simple hydrides and parent molecules remain, i.e., when the termination condition $i-a=n \land k-b=n$ holds for all dissociation products $M_a$ of the previous cycle. Summing over all individual reactions yields the formal BSR for molecule $M$ which may be written as

$$M + \sum_Z (n_R(Z) - 1) [ZH_2] \rightarrow \sum_{X-Y} [H_{i-a}(i-nX^nY^{k-n})H_{b-n}].$$

(9)

It contains only parent molecules on the product side and only $M$ and a number of simple hydrides on the reactant side, while all other intermediate reaction products cancel out each other. The first sum runs over the nonhydrogen atoms $Z$ of $M$, each of them involved in as many bond-cleavage reactions of type of Eq. (8) as there are residues attached $[n_R(Z)]$. Only the last of these reactions fulfills the above-mentioned termination condition; hence there are $n_R(Z)-1$ simple hydrides $ZH_2$ for a given $l$-valent atom $Z$. In total, a molecule $M$ with $n_R(M)$ nonhydrogen atoms and $n_{cyc}(M)$ ring systems requires $n_R(M) + 2n_{cyc}(M) - 2$ simple hydrides for complete bond separation. This is seen most easily if one processes the iterative procedure, Eq. (8), starting from terminal groups, and moves toward the center of the molecule such that the cleavage of each of the $n_R(M) + n_{cyc}(M) - 1$ nonhydrogen bonds requires precisely one simple hydride, except for the last one (no hydride) and for ring openings (two hydrides).

E. Definition of BSR corrections

BSRs, Eq. (9), may be used to improve on a low-level calculation of the atomization energy of $M$ by considering high-level results for all simple hydrides and parent molecules involved. A high-level ($H$) estimate is given by

$$E_{A,e}^H[M, \text{est.}] = - \Delta E_{\text{rec}}^L + \sum_{X-Y} E_{A,e}^H[H_{i-a}(i-nX^nY^{k-n})H_{b-n}]$$

$$- \sum_Z (n_R(Z) - 1) E_{A,e}^H[ZH_2],$$

(10)

where $\Delta E_{\text{rec}}^L$ is the low-level ($L$) reaction energy which reads

$$\Delta E_{\text{rec}}^L = - E_{A,e}^L[M] + \sum_{X-Y} E_{A,e}^L[H_{i-a}(i-nX^nY^{k-n})H_{b-n}]$$

$$- \sum_Z (n_R(Z) - 1) E_{A,e}^L[ZH_2].$$

(11)

Combining the two equations leads to

$$E_{A,e}^H[M, \text{est.}] = E_{A,e}^L[M] + \sum_{X-Y} \Delta E_{A,e}^L[H_{i-a}(i-nX^nY^{k-n})H_{b-n}]$$

$$- \sum_Z (n_R(Z) - 1) \Delta E_{A,e}^L[ZH_2]$$

(12)

where we have used the shorthand notation $\Delta E_{A,e}^L$ for the difference (high–low level) of atomization energies. Now we decompose the energy difference for a simple hydride $\Delta E_{A,e}^L[ZH_2]$ into $l$ bond contributions $\Delta E_{A,e}^\text{bond}[-1]Z-H$.

$$\Delta E_{A,e}^L[ZH_2] = l \Delta E_{A,e}^\text{bond}[-1]Z-H.$$

(13)

As for bonds between nonhydrogen atoms, the superscript $(l-1)$ indicates the number of free valencies; hence it allows us to specify different atom types referring to the same element. Taking C–H bonds as an example, we distinguish between those originating from methane ($tC-H$), singlet methylene ($tC-H$), and the methyl cation ($tC-H^+$). Next we define the bond contribution $\Delta E_{A,e}^\text{bond}[-nX^nY^{k-n}]$ of parent molecules with two nonhydrogen atoms $X$ and $Y$ as the difference between $\Delta E_{A,e}^L[H_{i-a}(i-nX^nY^{k-n})H_{b-n}]$ and the respective number of $i-X-H$ and $k-Y-H$ bond contributions

$$\Delta E_{A,e}^\text{bond}[-nX^nY^{k-n}] = \Delta E_{A,e}^L[H_{i-a}(i-nX^nY^{k-n})H_{b-n}]$$

$$- (i-n) \Delta E_{A,e}^\text{bond}[-1]X-H$$

$$- (k-n) \Delta E_{A,e}^\text{bond}[-1]Y-H.$$  

(14)

Using this decomposition we can rewrite Eq. (12) more compactly as

$$E_{A,e}^H[M, \text{est.}] = E_{A,e}^L[M] + \sum_{X-Y} \Delta E_{A,e}^\text{bond}[-nX^nY^{k-n}]$$

$$+ \sum_{Z-H} \Delta E_{A,e}^\text{bond}[-1]Z-H.$$  

(15)

Here we have used the simple fact that the BSR, Eq. (9), is isodesmic. Hence the number and type of $Z-H$ bonds appearing in $M$ and in the simple hydrides $ZH_2$ on the reactant side must balance those appearing in the parent molecules on the product side. Subtracting the $Z-H$ bond increments of all simple hydrides from those of all parent molecules will thus leave precisely those $Z-H$ bond terms in Eq. (15) which correspond in type and number to the set of all $Z-H$ terms found in $M$. In essence the definitions in Eqs. (13) and (14) allow a mapping from BSRs, Eq. (12), to bond increments, Eq. (15), as illustrated also in Fig. 1.

The mapping calls for a particularly simple computational implementation, as the identification of molecular topology can easily be automated. Once the optimized geometry of a molecule is known, simple distance criteria can be
used to identify bonds, while neighbor counts and electronic bookkeeping will establish bond types for most standard situations. Little user intervention is thus required unless several resonance structures can be drawn. The mapping also allows for a simple ready-to-use presentation of high-level reference data which only need to be calculated once and can then be used to correct atomization energies of larger molecules treated at lower levels of theory. Reference data and bond increments obtained from high-level calculations will be presented in Sec. V D.

V. RESULTS: ATOMIZATION ENERGIES AT THE CCSD(T) LEVEL

The energy decomposition analysis defined in Sec. IV A is applied to a set of 73 molecules composed of at least three first-row atoms (C, N, O, F) and hydrogen. The set is listed in Table S2.\textsuperscript{90} It represents a large variety of bond types, including more difficult cases with charge-separated valence structures (Fig. 2), thus permitting a proper judgment of the benefits (and weaknesses) of the BSR scheme. The chosen reference level of theory, \{CCSD(T)\}, is expected to reproduce the complete basis-set limit of CCSD(T) to within a few tenths of a kcal/mol at least, is used for all further comparisons and to derive isodesmic corrections from calculations on BSR prototypes (parent molecules and simple hydrides). We start with the component analysis of regular (uncorrected) calculations (Sec. V A), then discuss results for BSR-corrected calculations (Sec. V B), and finally analyze composite approaches assembled from components without (Sec. V C) and with (Sec. V D) application of BSR corrections.

A. Component analysis for uncorrected energies

Figures 3 and 4 present individual analyses for each of the eight components to the atomization energy, showing root mean square (rms) and mean signed errors (MSE) with respect to the reference level of theory, \{CCSD(T)\}, that arise from the choice of smaller basis sets either with (lower panels) or without extrapolation (upper panels). Horizontal lines in Fig. 3 indicate error bars of 0.5 kcal/mol which probably define the maximum individual error one should allow for if overall results of high accuracy are desired. We emphasize that the terms “valence-shell contributions” and “inner-shell contributions.”

FIG. 2. Valence structures of molecules in the test set that show charge separation.

FIG. 3. Energy decomposition analysis: Regular CCSD(T)(full) calculations (rms errors). The eight charts in the top row show rms errors in calculated atomization energies that occur if the corresponding component \(n = 1 \cdots 8\) is evaluated using basis set \(X_n\), while basis sets \(X_k\) (\(k \neq n\)) for all other components remain as defined for the reference level of theory, \{CCSD(T)\}. The bottom row shows corresponding results for the case that component \(n\) is extrapolated to the complete basis-set limit using basis sets of cardinal numbers \((X_{n-1}, X_n)\). In essence, each single chart shows the basis-set incompleteness error observed for a particular energy component as a function of the cardinal number \(X_n\) of the basis set. Components 1–4 and 5–8 have been termed valence-shell contributions and inner-shell contributions, respectively, to denote (a) the types of basis sets involved (cc-pVXZ vs cc-pCVXZ) and (b) the type of MP2, CCSD, or CCSD(T) correlation energy. See Sec. IV A for a complete definition of nomenclature and Table S2 (Ref. 90) for a listing of all 73 molecules used in this analysis. Note that each molecule has at least three nonhydrogen atoms. Some rms errors exceed the range displayed; they are given here in kcal/mol. \(E_{\text{HF}}(X_n)\): \(X_1 = 2\) (15.2), \(X_3 = 0\) (390.5); \(\Delta_{\text{MP2}}(X_n)\): \(X_3 = 5\) (4.1), \(X_4 = 4\) (7.6), \(X_5 = 3\) (16.5), \(X_2 = 2\) (37.6), \(X_4 = 0\) (205.0); \(\Delta_{\text{CCSD}}(X_n)\): \(X_3 = 3\) (3.4), \(X_2 = 2\) (8.6), \(X_1 = 0\) (38.2); \(\Delta_{\text{CCSD(T)}}(X_n)\): \(X_4 = 2\) (5.4), \(X_4 = 0\) (16.5).
shell contributions” apply strictly only to correlation energies while HF energy components have only been classified accordingly to denote the basis-set types used for their evaluation (cc-pVXZ and cc-pCVXZ).

It is immediately evident from Fig. 3 that accurate estimates of atomization energies require complete basis-set extrapolations of valence-shell correlation energies. This is well known and has been considered in all fully *ab initio* thermochemical protocols. Figure 3 also shows that extrapolations from small basis sets (Xn = 2 and 3) are not reliable enough for MP2 and CCSD valence-shell contributions to atomization energies, in agreement with observations discussed previously.52 One may use more accurate extrapolation formulas which capture differences in convergence behavior between molecules, but so far they have only been calibrated for total energy contributions (not for atomization energy contributions), and they are not considered here.

Complete basis-set extrapolation is also necessary for connected-triples contributions unless basis sets of at least quintuple-zeta quality are used, but in this case extrapolations from small basis sets appear to be sufficiently accurate. The situation is somewhat more involved for HF contributions to atomization energies which generally converge faster with basis-set size but less regularly than correlation contributions. The HF energy is essentially converged for Xn = 6 and almost so for Xn = 5, but extrapolations from small basis sets Xn = 2 and Xn = 3 perform quite poorly. On the other hand extrapolations from cc-pVTZ and cc-pVQZ basis sets are quite satisfactory, and those from even larger basis sets, while not really required, appear safe as the energy is already almost converged. These observations largely agree with a more detailed analysis of Halkier et al. on a smaller set of reference data.94

Not unexpectedly, basis-set extrapolation is found to be somewhat erratic for inner-shell contributions. ΔMP2(core)(Xn) [see Eq. (3)] improves slightly upon extrapolation for Xn = 4 but worsens for Xn = 3. Similar remarks apply to ΔCCSD(core)(X7) and ΔCCSD(TD)(X8). Recall that the extrapolation procedure was calibrated for valence-shell contributions but applied to the sum of valence-shell and inner-shell contributions. This appears reasonable as inner-shell correlation contributions to atomization energies are smaller by about two orders of magnitude than valence-shell contributions, but one cannot expect that the small inner-shell contributions are extrapolated accurately, not even at the reference level of theory. The results rather appear to indicate that ΔMP2(core)(Xn) and ΔCCSD(core)(X7) contributions to inner-shell correlation energies need to be calculated at least with cc-pCVQZ and cc-pCVTZ basis sets, respectively, no matter whether extrapolation is applied or not.

In summary, the minimum basis-set requirements for reasonably accurate thermochemical protocols appear to be extrapolations from Xn = 3 and Xn = 4 for HF, MP2, and CCSD valence-shell contributions, extrapolation from Xn = 2 and Xn = 3 for valence-shell triples contributions, and calculations of inner-shell contributions ΔMP2(core)(X6), ΔCCSD(core)(X7), and ΔCCSD(TD)(X8) using Xn = 4, Xn = 3, and Xn = 2, respectively. We note, however, that choices of smaller basis sets may be appropriate if one accepts partial error cancellation (Fig. 4, Sec. V C). Finally, the HF component ΔHFF(core)(Xn) should be omitted as it vanishes for Xn → ∞ and any finite-basis-set calculation necessarily introduces error.

The above analysis shows that basis-set requirements are quite substantial if one aims for reasonable overall accuracy (1–2 kcal/mol) and does not allow for either empirical calibration or fortuitous error cancellation. Unfortunately, expensive CCSD calculations with cc-pVQZ basis sets are required to reach this goal, thus severely limiting the range of possible applications.

### B. Component analysis for BSR-corrected energies

Consistent use of BSRS reduces basis-set requirements dramatically (compare Figs. 3 and 5). Most importantly, CCSD valence-shell energies only need to be evaluated for cc-pVTZ basis sets, thus easily saving a factor of 10 in computational cost, to achieve the same high level of accuracy. Calculations still benefit from complete basis-set extrapolations but not nearly as much as without application of BSR corrections, indicating that the BSR scheme absorbs effects of basis-set incompleteness very efficiently. The MP2 component ΔMP2(core)(Xn) may be evaluated for Xn = 3, either with or without extrapolation, if only medium accuracy is desired, but extrapolation from (Xn − 1, Xn) = (3, 4) is certainly preferable.

Large improvements are observed for inner-shell contributions to the correlation energy. ΔCCSD(core)(X7) and ΔCCSD(TD)(X8) terms are essentially taken care of by BSR corrections alone, not requiring any explicit calculation, and the marginal error observed for ΔCCSD(core)(X8) may be eliminated by small-basis-set calculations (Xn = 3). These results are expected as inner-shell correlation describes a mostly local (atomic) effect that is perturbed only moderately by the process of bond formation with little or no influence from second neighbors. BSRS preserve chemical bonds to immediate neighbors and are thus expected to absorb any effect that is very local in nature. It is noteworthy that BSRS also correct for most of the contribution from the “unphysical” term ΔHFF(core)(Xn) whose nonzero value is entirely due to basis-set incompleteness.

Residual errors are typically less systematic (Fig. 6 versus Fig. 5) than observed for uncorrected calculations (Fig. 4 versus Fig. 3), but some trends are still clearly visible. We note, for example, that unextrapolated HF and valence-shell correlation energies [ΔMP2(core)(Xn), ΔCCSD(core)(Xn)] are often overcorrected through the use of BSRS (Figs. 4 and 6).

### C. Composite models without corrections

Using the experience gathered so far, we may define a sequence of composite methods that balance computational effort and numerical accuracy (Table I). In selecting a computational protocol, we only apply the formal requirements that X1 = X2 ≥ X3 ≥ X4, X6 ≥ X7 ≥ X8, and Xn ≥ Xn+4 (n = 2, 3, 4) to ensure (a) that larger contributions to the atomization energy are always treated more accurately and (b) that energies already available for a particular component are not discarded. It does obviously not make sense, for example, to
perform a MP2/cc-pV5Z calculation and then replace the already available HF component with a HF/cc-pVQZ energy. Let us first consider uncorrected methods.

Starting from the reference level of theory (line 1 in Table I) and discarding the unphysical term $\Delta_{\text{HF}}^\text{core}(X_3)$ (line 2), we define the first level of approximation by replacing (5,6) extrapolations for $\Delta_{\text{CCSD}}(X_3)$ with (4,5) extrapolations (line 3). This introduces a rms error of about 0.1 kcal/mol

![Energy decomposition analysis: Regular CCSD](image1)

![Energy decomposition analysis: CCSD(T)(full)](image2)
which can be maintained if one uses less expensive (4,5) extrapolations also for $E_{\text{HF}}(X_3)$ and $\Delta_{\text{MP2}}(X_2)$ (line 4). The next level of approximation, aiming for about 0.3 kcal/mol rms error, permits further simplifications in $\Delta_{\text{CCSD}}(X_3)$, $\Delta_{\text{CCSD(T)}}(X_4)$, $\Delta_{\text{CCSD(T)}}(X_7)$, and $\Delta_{\text{CCSD(T)}}(X_8)$ (lines 5–8), but valence-shell CCSD contributions need to be evaluated at least with quadruple-zeta basis sets and even inner-shell triples contributions cannot be neglected completely. Any further simplification appears to increase the rms error beyond 0.5 kcal/mol (Fig. 3).

One surprisingly accurate composite method is shown in line 17 of Table I. It only involves complete basis-set estimates of the HF and MP2 valence-shell correlation energies and additional CCSD(T)/cc-pVTZ and MP2(full)/cc-pCVTZ calculations. Inspection of Figs. 3 and 4 shows that the excellent error statistics are mainly due to fortuitous cancellation between quite sizable positive errors for $\Delta_{\text{CCSD}}(X_3)$ (MSE: 3.3 kcal/mol) and negative errors for $\Delta_{\text{CCSD(T)}}(X_4)$, $\Delta_{\text{CCSD(T)}}(X_7)$, and $\Delta_{\text{CCSD(T)}}(X_8)$ (MSE: −1.4, −0.7, −0.2, −0.4 kcal/mol). Table I (line 15) and Fig. 4 further indicate that overall errors increase substantially [mean unsigned error (MUE) from 0.64 to 1.94 kcal/mol] if the valence-shell components $\Delta_{\text{CCSD}}(X_3)$ and $\Delta_{\text{CCSD(T)}}(X_4)$ are treated more appropriately by extrapolation to the complete basis-set limit. We note that a new version of the ccCA approach recently proposed by DeYonker et al. follows a protocol fairly similar to the one shown in line 17 of Table I, although details of the extrapolation procedure vary and augmented basis sets are used for the HF and MP2 components. It is thus quite possible that the good performance of the ccCA approach largely reflects the type of error cancellation discussed above.

D. Composite models with BSR corrections

Turning to BSR-corrected schemes, we note immediately that error statistics vary much less with the precise composition of a compound method, showing that BSR corrections absorb residual errors very efficiently (Table I). Without sacrificing the small rms error of 0.1 kcal/mol observed for $\Delta_{\text{CCSD}}(X_3)$, basis-set cardinal numbers can be lowered by 1 and 2, respectively, for $\Delta_{\text{CCSD}}(X_3)$ and $\Delta_{\text{CCSD(T)}}(X_4)$, and all inner-shell contributions beyond MP2 may be neglected (lines 5–11). Line 11 defines our highest-level BSR-corrected approach, from now on referred to as “model A.” Setting a goal of ±0.3 kcal/mol for the next level of approximation, one may definitely replace all remaining (4,5) extrapolations by (3,4) extrapolations, and, maybe even more importantly, obtain $\Delta_{\text{CCSD}}(X_3)$ from (2,3) rather than from (3,4) extrapolations or even from unextrapolated cc-pVTZ results (lines 13–22). A number of options are available, some of which benefit from very limited error cancellation. Simplifications of $E_{\text{HF}}(X_1)$, $\Delta_{\text{MP2}}(X_2)$ [using (3,4) instead of (4,5) extrapolations], $\Delta_{\text{CCSD}}(X_3)$, $\Delta_{\text{CCSD(T)}}(X_4)$ (abandoning extrapolation), and $\Delta_{\text{CCSD}}(X_7)$ (neglect) all increase rms errors in the corresponding components (Fig. 5) but introduce either positive [$E_{\text{HF}}(X_1)$, $\Delta_{\text{MP2}}(X_2)$] or negative [$\Delta_{\text{CCSD}}(X_3)$, $\Delta_{\text{CCSD(T)}}(X_4)$, $\Delta_{\text{MP2}}(X_2)$] shifts in MSEs (Fig. 6, Table I). Rather than choosing the option with the lowest overall rms error, we suggest two variants of model B, referred to as B$_1$ and B$_2$, which show a small positive MSE and a small negative MSE, respectively. The two options only differ in the term $\Delta_{\text{CCSD}}(X_7)$ which is considered in model B$_1$ ($X_7 = 3$) but neglected in model B$_2$. Evaluation of this term should never be the bottleneck of model B calcu-

<table>
<thead>
<tr>
<th>Valence shell contributions</th>
<th>Inner shell contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{\text{HF}}(X_1)$</td>
<td>$\Delta_{\text{MP2}}(X_2)$</td>
</tr>
<tr>
<td>$\Delta_{\text{CCSD}}(X_3)$</td>
<td>$\Delta_{\text{CCSD(T)}}(X_4)$</td>
</tr>
<tr>
<td>$\Delta_{\text{CCSD}(X_7)}$</td>
<td>$\Delta_{\text{CCSD(T)}}(X_8)$</td>
</tr>
</tbody>
</table>

Fig. 6. Energy decomposition analysis: CCSD(T)(full) calculations with BSR corrections applied (MSEs). See caption of Fig. 4 for details. Some MSE exceed the range displayed; they are given here in kcal/mol. $E_{\text{HF}}(X_1)$: $X_1=0$ (−15.8), $\Delta_{\text{MP2}}(X_2)$: $X_2=0$ (−8.0), $\Delta_{\text{CCSD}}(X_3)$: $X_3=0$ (4.3), $\Delta_{\text{CCSD(T)}}(X_4)$: $X_4=0$ (−1.7).
hydrogen. See Table S2.

valence-shell CCSD component

difficult cases with charge-separated valence structures. Us-
gested which exhausts possible further simplifications that do

ttions lead to individual component errors around 0.5–1.0

Further reduction of computational complexity leads to

with BSR corrections\textsuperscript{ab} and evaluations of

\begin{table}[ht]
\centering
\caption{Statistical analysis of selected compound methods for atomization energies.}
\begin{tabular}{cccccc}
\hline
 & & With BSR corrections\textsuperscript{ab} & & Without BSR corrections\textsuperscript{a} & \\
 & & MSE & MUE & rms & Max. & MSE & MUE & rms & Max. \\
\hline
1 & 6665-5555 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
2 & 6665-555 & 0.00 & 0.00 & 0.00 & -0.02 & -0.01 & 0.01 & 0.02 & -0.03 \\
3 & 6655-5555 & 0.03 & 0.05 & 0.07 & 0.25 & -0.02 & 0.06 & 0.08 & 0.18 \\
4 & 5555-5555 & -0.05 & 0.07 & 0.10 & -0.31 & -0.07 & 0.09 & 0.12 & -0.29 \\
5 & 5554-5533 & -0.01 & 0.05 & 0.06 & -0.17 & -0.06 & 0.12 & 0.15 & -0.47 \\
6 & 5533-5533 & -0.02 & 0.08 & 0.11 & -0.42 & 0.07 & 0.13 & 0.16 & -0.44 \\
7 & 5543-5533 & -0.02 & 0.07 & 0.09 & -0.35 & 0.08 & 0.18 & 0.22 & -0.75 \\
8 & 5543-5532 & -0.02 & 0.07 & 0.09 & -0.33 & -0.01 & 0.23 & 0.29 & -1.09 \\
9 & 5543-53- & -0.03 & 0.07 & 0.10 & -0.36 & -0.39 & 0.40 & 0.51 & -1.62 \\
10 & 5543-5- & -0.02 & 0.07 & 0.10 & -0.33 & -0.59 & 0.59 & 0.63 & -1.20 \\
11 & 5543-3- & -0.04 & 0.09 & 0.13 & -0.49 & -1.28 & 1.28 & 1.41 & -3.03 \\
12 & 5543- - & -0.24 & 0.26 & 0.32 & -0.83 & -2.25 & 2.25 & 2.58 & -5.92 \\
13 & 5533-3- & -0.12 & 0.18 & 0.25 & -0.88 & -1.78 & 1.78 & 1.96 & -4.43 \\
14 & 5533- & -0.32 & 0.34 & 0.44 & -1.20 & -2.75 & 2.75 & 2.92 & -6.01 \\
15 & 4333-3- & 0.11 & 0.24 & 0.31 & 0.90 & -1.94 & 1.94 & 2.19 & -4.74 \\
16 & 4333- & -0.10 & 0.19 & 0.27 & -1.03 & -2.90 & 2.90 & 3.04 & -5.87 \\
17 & 4333-3- & -0.03 & 0.20 & 0.28 & -0.93 & 0.33 & 0.64 & 0.76 & 1.64 \\
18 & 4333- & -0.23 & 0.30 & 0.40 & -1.26 & -0.63 & 0.93 & 1.25 & -3.95 \\
19 & 4332-3- & -0.14 & 0.23 & 0.30 & -0.91 & -3.49 & 3.49 & 3.69 & -7.25 \\
20 & 4332- & -0.34 & 0.36 & 0.46 & -1.33 & -4.46 & 4.46 & 4.67 & -10.15 \\
21 & 4332-3- & -0.12 & 0.21 & 0.27 & -0.78 & -2.88 & 2.92 & 3.29 & -7.25 \\
22 & 4332- & -0.32 & 0.34 & 0.43 & -1.07 & -3.85 & 3.85 & 4.17 & -10.15 \\
23 & 4322-3- & -0.27 & 0.53 & 0.72 & -2.57 & 1.55 & 1.69 & 2.01 & 3.99 \\
24 & 4322- & -0.47 & 0.62 & 0.87 & -2.88 & 0.58 & 1.60 & 1.96 & -4.49 \\
25 & 3322-3- & -0.07 & 0.68 & 0.95 & -3.19 & 1.87 & 2.35 & 2.95 & 7.75 \\
26 & 3322- & -0.27 & 0.76 & 0.99 & -3.28 & 0.90 & 1.88 & 2.51 & 6.28 \\
\hline
\end{tabular}
\end{table}

\textsuperscript{a}Errors are given with respect to \{6665-5555\} (in kcal/mol) and refer to a set of 73 molecules composed of at least three first-row atoms (C, N, O, F) and hydrogen. See Table S2 (Ref. 90) for a complete list of molecules and Sec. IV A for a definition of the nomenclature used. The symbol 3\textsuperscript{a} in methods 21 and 22 refers to a mixed basis set (cc-pVDZ on hydrogen, cc-pVTZ on first-row atoms).

\textsuperscript{b}BSR-corrected methods define our standard approach (ATOMIC).

\textsuperscript{c}Definition of preferred compound methods. See text.

kcal/mol (Fig. 5), but some error cancellation is visible (Fig. 6), so that overall errors still remain below 1 kcal/mol. At this level of accuracy one should aim for computational simplicity, and we suggest a low-cost variant (model C) mainly for explorative studies which requires only two calculations to be performed: MP2/cc-pVTZ and CCSD(T)/cc-pVTZ.

All thermochemical models are defined in Table I. Corresponding atomic energies and BSR corrections are listed in Tables S1 (Ref. 90) and II, respectively. Reference data \{6665-5555\} for BSR prototypes can easily be recovered from Table II, using the definitions provided in Eqs. (13) and (14). To give an example, the atomization energy of fluoromethane \((^{3}\text{C}-\text{F}; \ 3\ ^{3}\text{C}-\text{H})\) equals 107.864+3 \times 105.120 kcal/mol=423.22 kcal/mol. Note that Table II includes a number of data derived from cationic BSR prototypes \([\text{such as methylammonium, } ^{3}\text{C}-\text{N}^{3}(+); \ 3\ ^{3}\text{C}-\text{H}; \ 3\ ^{3}\text{N}-\text{H}(+)]\) needed for molecules with charge-separated valence structures (methyl isocyanide in this case).

Individual results for the entire test set are collected in Table S2.\textsuperscript{90} Without discussing them in great detail, we only note that the BSR model works surprisingly well also for difficult cases with charge-separated valence structures. Us-
ing model A as an example, we observe that BSR corrections reduce the errors for the 11 molecules shown in Fig. 2 from 0.91 to 0.13 kcal/mol on average (MUE), with methyl isocyanide (from –1.30 to –0.12 kcal/mol) and ozone (from –0.48 to –0.14 kcal/mol) showing the largest and smallest improvements, respectively.

VI. BEYOND THE CCSD(T) LEVEL

The consistent use of BSRs has made it possible to find inexpensive but accurate approximations of the complete basis-set limit within the CCSD(T) model. While this is undoubtedly the most important step in calculating atomization energies, it is not sufficient if high accuracy is sought. Other contributions which have little effect on the chemistry of first-row compounds may become important when describing the total atomization of a molecule. Most notable are relativistic effects which are considered in virtually every nonempirical thermochemistry protocol. Apart from spin-orbit coupling (Sec. II), major contributions arise from the spin-free (scalar) mass-velocity (MV) and one-electron Darwin (D) terms99–101 which are evaluated to first order as expectation values of the combined (MVD) operator. Smaller in magnitude but still considered in various thermochemical protocols11,14 are corrections to the Born–Oppenheimer approximation. Ignoring couplings with electronically excited states, the adiabatic (“diagonal”) approximation provides for the leading correction term which is evaluated as expectation value of the nuclear kinetic energy operator102–104 and still retains the familiar picture of potential energy surfaces where each nuclear configuration is associated with one particular value of the energy. Good estimates of diagonal Born–Oppenheimer corrections (DBOCs) are already obtained at the HF level, especially for relative energies,104 so that electron correlation effects on DBOC can be ignored for the expectations of the combined model toward the expectation value of the energy. Good estimates of diagonal Born–Oppenheimer corrections (DBOCs) are already obtained at the HF level, especially for relative energies,104 so that electron correlation effects on DBOC can be ignored for the expectation value of the nuclear kinetic energy operator102–104 and still retains the familiar picture of potential energy surfaces where each nuclear configuration is associated with one particular value of the energy. Good estimates of diagonal Born–Oppenheimer corrections (DBOCs) are already obtained at the HF level, especially for relative energies,104 so that electron correlation effects on DBOC can be ignored for the evaluation of atomization energies.

Finally, corrections of the CCSD(T) model toward the
exact nonrelativistic $N$-particle treatment should be considered as well. While full CI calculations are almost never feasible, coupled-cluster treatments with connected triple, \cite{10,11,14,15,105,106} quadruple, \cite{10,11,14,15,106} and sometimes even pentuple excitations \cite{11,105,106} have been included in the most refined \textit{ab initio} thermochemistry protocols. The large success of the simpler CCSD(T) model is in part due to the fairly consistent overestimation of connected triples contributions to atomization energies which – thanks to error cancellation – makes it agree better with CCSDTQ (Ref. \citenum{107}) than CCSDT (Refs. \citenum{108} and \citenum{109}) itself. \cite{11,105,106} This implies, however, that refinements of the CCSD(T) model need to consider contributions both from the relaxation of connected triple excitations [CCSDT-CCSD(T)] and from connected quadruple excitations (CCSDTQ-CCSDT).

Here, we develop the idea of BSRs one step further and calculate MVD, DBOC, CCSDT-CCSD(T), and CCSDTQ-CCSDT corrections under the assumption that the corresponding energies of bond separation are small and can be neglected. Taking the example of Sec. IV C and Fig. 1 again,

\[ \text{oxirene} + 2\text{CH}_4 + \text{H}_2\text{O} \rightarrow \text{CH}_2 + 2\text{H}_2\text{C} \rightarrow \text{OH}, \]

the energy correction for oxirene is then evaluated as

\[ E_{\text{corr}}[\text{oxirene,est.}] = E_{\text{corr}}[\text{H}_2\text{C} = \text{CH}_3] \]
\[ + 2E_{\text{corr}}[\text{H}_2\text{C} \rightarrow \text{OH}] \]
\[ - 2E_{\text{corr}}[\text{CH}_2] - E_{\text{corr}}[\text{H}_2\text{O}], \]

where we have dropped the superscripts $L$ and $H$ as well as the $\Delta$ of Eq. (10) and following simply because the “low-level energy” is chosen to be zero. According to Eq. (17), the calculation of any of the corrections is reduced to the trivial task of summing up bond increments.

Bond increments are evaluated from explicit calculations for BSR prototypes [Eqs. (13) and (14), Table III], thereby closely following the protocol implemented in the HEAT (Ref. \citenum{14}) procedure: Scalar relativistic corrections have been obtained from UHF-CCSD(T)(full)/aug-cc-pVTZ (Ref. \citenum{110}) one-particle density matrices contracted with one-electron Darwin and mass-velocity terms. DBOCs have initially been computed at the UHF/aug-cc-pVTZ level, but for our purposes virtually identical results are obtained with the smaller cc-pVVDZ basis set (Sec. VI B), an observation that has determined our final choice. CCSDTQ-CCSD(T) corrections (frozen core, UHF) have been evaluated using cc-pVVDZ and cc-pVTZ basis sets, extrapolated to the complete basis-set limit ($X^{-3}$). Tajti \textit{et al.} performed extrapolations based on cc-pVTZ and cc-pVQZ calculations;\cite{14} the latter, however, seemed too taxing for some of the larger BSR prototypes. For a limited set of 25 molecules, we have computed (DT) and (TQ) extrapolated contributions to atomization energies and found that both approaches agree to within 0.04 kcal/mol in all cases, while calculated values vary over a fairly wide range from $-0.1$ kcal/mol (methane) to $-0.8$ kcal/mol (hydrogen cyanide). Additional justification for the use of smaller basis sets may be obtained from related comparisons in the literature \cite{0.11} and from a recent study considering large basis-set extrapolations for selected small systems.\cite{11} Following the HEAT protocol, CCSDTQ-CCSDT corrections are evaluated using a cc-pVVDZ basis set (frozen core, UHF). Calculations with larger basis sets would be intractable for many of the BSR prototypes even though they may become necessary for compounds involving second-row atoms and for “pathological” multireference cases.\cite{11} Fortunately, they appear less mandatory for “normal” first-row chemistry.\cite{14,111} Comparison with data available from a very recent compilation of highly accurate W4 calculations on small molecules\cite{112} indeed shows that the small basis set reproduces CCSDTQ-CCSDT contributions to atomization energies quite well (deviation $\leq 0.1$ kcal/mol) in most cases (18 out of 26) and with acceptable errors ($\leq 0.2$ kcal/mol) in six further cases. Larger errors are only seen for cyanogen (1.92 versus 2.25 kcal/mol) and for ozone (3.43 versus 3.81 kcal/mol), which is known for its multireference character and shows significant further contributions from even higher excitations than considered here.

The \textit{ACESII} (Ref. \citenum{113}) code has been used to evaluate relativistic corrections and DBOCs and coupled with MRCC (Refs. \citenum{114}–\citenum{116}) to obtain molecular CCSDT and CCSDTQ energies. Atomic energies at the CCSDT and CCSDTQ levels of theory were obtained using the \textit{nwchem} program package.\cite{117}–\citenum{119} The designations UHF and UB3LYP used in this section refer to our choice of spin-unrestricted theory for open-shell systems (atoms). Spin-restricted theory has been used throughout for closed-shell systems (all molecules).

The next three sections address the validity of the assumption that BSRs are approximately thermoneutral for MVD, DBOC, CCSDT-CCSD(T), and CCSDTQ-CCSDT corrections to atomization energies. In order to compare direct calculations with sums of precomputed bond increments for a reasonable number of larger molecules, we need to employ somewhat reduced levels of theory for the evaluation of each of the corrections. Results are presented both in graphical (Figs. 7–12) and tabular form (Tables S3–S5\textsuperscript{[60]}).

## A. Relativistic effects

Figure 7 compares UB3LYP (Refs. \citenum{120} and \citenum{121})/cc-pCVTZ [\textit{gaussian} definition\textsuperscript{122,123} using \textit{turbomole} (Ref. \citenum{78})] and UHF-CCSD(T)(full)/aug-cc-pCVTZ estimates of scalar relativistic contributions to atomization energies. The density functional was suggested by Martin and Parthiban\cite{13} as an economical alternative to wave function based methods which require at least some consideration of electron correlation as MVD contributions are severely overestimated at the HF level of theory. The comparison includes 32 small molecules with up to two nonhydrogen atoms, mostly chosen from the set of BSR prototypes (see figure caption for details). The accuracy obtained at the UB3LYP level is quite impressive; a linear regression analysis between UB3LYP and CCSD(T) results affords a slope close to unity.
TABLE III. BSR increments $E_{\text{corr}}$ calculated for relativistic (MVD), non-Born–Oppenheimer (DBOC), and higher excitation corrections to atomization energies [kcal/mol; see Eqs. (13)–(17); values are reported to three digits to avoid accumulation of roundoff error; actual numerical accuracy may be less].

<table>
<thead>
<tr>
<th>Bond</th>
<th>MVD</th>
<th>DBOC</th>
<th>CCSDT-CCSD(T)</th>
<th>CCSDTQ-CCSDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>C—H</td>
<td>-0.049</td>
<td>0.027</td>
<td>-0.021</td>
<td>0.019</td>
</tr>
<tr>
<td>C—H</td>
<td>-0.045</td>
<td>-0.051</td>
<td>0.101</td>
<td>0.048</td>
</tr>
<tr>
<td>N—H</td>
<td>-0.085</td>
<td>0.043</td>
<td>-0.045</td>
<td>0.058</td>
</tr>
<tr>
<td>N—H (+)</td>
<td>-0.074</td>
<td>0.070</td>
<td>-0.032</td>
<td>0.034</td>
</tr>
<tr>
<td>O—H</td>
<td>-0.135</td>
<td>0.055</td>
<td>-0.102</td>
<td>0.119</td>
</tr>
<tr>
<td>O—H (+)</td>
<td>-0.116</td>
<td>0.091</td>
<td>-0.061</td>
<td>0.068</td>
</tr>
<tr>
<td>C—C</td>
<td>-0.101</td>
<td>-0.009</td>
<td>-0.221</td>
<td>0.088</td>
</tr>
<tr>
<td>C—C (+)</td>
<td>-0.133</td>
<td>0.014</td>
<td>-0.375</td>
<td>0.253</td>
</tr>
<tr>
<td>C—C (+)</td>
<td>-0.180</td>
<td>0.064</td>
<td>-0.685</td>
<td>0.512</td>
</tr>
<tr>
<td>C—N²</td>
<td>-0.147</td>
<td>-0.015</td>
<td>-0.252</td>
<td>0.124</td>
</tr>
<tr>
<td>C=N¹</td>
<td>-0.167</td>
<td>-0.015</td>
<td>-0.454</td>
<td>0.373</td>
</tr>
<tr>
<td>C=N³</td>
<td>-0.182</td>
<td>0.047</td>
<td>-0.775</td>
<td>0.749</td>
</tr>
<tr>
<td>C=N¹</td>
<td>-0.187</td>
<td>0.020</td>
<td>-0.541</td>
<td>0.472</td>
</tr>
<tr>
<td>C=N² (+)</td>
<td>-0.171</td>
<td>0.036</td>
<td>-0.311</td>
<td>0.228</td>
</tr>
<tr>
<td>C=N³ (+)</td>
<td>-0.117</td>
<td>0.020</td>
<td>-0.209</td>
<td>0.095</td>
</tr>
<tr>
<td>C≡N (+)</td>
<td>-0.216</td>
<td>0.089</td>
<td>-0.600</td>
<td>0.513</td>
</tr>
<tr>
<td>C≡O (+)</td>
<td>-0.190</td>
<td>-0.015</td>
<td>-0.276</td>
<td>0.172</td>
</tr>
<tr>
<td>C—O¹</td>
<td>-0.159</td>
<td>0.024</td>
<td>-0.300</td>
<td>0.246</td>
</tr>
<tr>
<td>C≡O</td>
<td>-0.249</td>
<td>-0.034</td>
<td>-0.492</td>
<td>0.470</td>
</tr>
<tr>
<td>C—F</td>
<td>-0.242</td>
<td>-0.009</td>
<td>-0.237</td>
<td>0.193</td>
</tr>
<tr>
<td>C—F</td>
<td>-0.187</td>
<td>-0.005</td>
<td>-0.198</td>
<td>0.253</td>
</tr>
<tr>
<td>N—N¹</td>
<td>-0.182</td>
<td>-0.011</td>
<td>-0.270</td>
<td>0.184</td>
</tr>
<tr>
<td>N≡N</td>
<td>-0.161</td>
<td>-0.050</td>
<td>-0.534</td>
<td>0.580</td>
</tr>
<tr>
<td>N≡N (+)</td>
<td>-0.140</td>
<td>0.051</td>
<td>-0.745</td>
<td>0.913</td>
</tr>
<tr>
<td>N≡N</td>
<td>-0.230</td>
<td>-0.123</td>
<td>-0.330</td>
<td>0.439</td>
</tr>
<tr>
<td>N≡N (+)</td>
<td>-0.146</td>
<td>0.018</td>
<td>-0.223</td>
<td>0.154</td>
</tr>
<tr>
<td>N—O¹</td>
<td>-0.166</td>
<td>-0.002</td>
<td>-0.294</td>
<td>0.291</td>
</tr>
<tr>
<td>N≡O</td>
<td>-0.198</td>
<td>-0.100</td>
<td>-0.569</td>
<td>0.829</td>
</tr>
<tr>
<td>N≡O (+)</td>
<td>-0.184</td>
<td>-0.029</td>
<td>-0.588</td>
<td>0.789</td>
</tr>
<tr>
<td>N—O</td>
<td>-0.248</td>
<td>-0.104</td>
<td>-0.383</td>
<td>0.368</td>
</tr>
<tr>
<td>N—O¹ (+)</td>
<td>-0.145</td>
<td>0.011</td>
<td>-0.265</td>
<td>0.241</td>
</tr>
<tr>
<td>N—F</td>
<td>-0.171</td>
<td>-0.001</td>
<td>-0.223</td>
<td>0.330</td>
</tr>
<tr>
<td>N—F (+)</td>
<td>-0.147</td>
<td>0.005</td>
<td>-0.241</td>
<td>0.315</td>
</tr>
<tr>
<td>O—O¹</td>
<td>-0.116</td>
<td>0.004</td>
<td>-0.358</td>
<td>0.473</td>
</tr>
<tr>
<td>O—O</td>
<td>-0.115</td>
<td>-0.054</td>
<td>-0.269</td>
<td>0.369</td>
</tr>
<tr>
<td>O≡O (+)</td>
<td>-0.031</td>
<td>-0.102</td>
<td>-0.573</td>
<td>1.241</td>
</tr>
<tr>
<td>O—F</td>
<td>-0.100</td>
<td>0.007</td>
<td>-0.316</td>
<td>0.564</td>
</tr>
</tbody>
</table>

The assumption of thermoneutral BSRs provides for remarkably accurate estimates of scalar relativistic corrections, with rms and maximum errors of 0.05 and $-0.22 \text{ kcal/mol}$ (tetrafluorohydrazine, BSR: $-0.96 \text{ kcal/mol}$ versus calculated: $-0.74 \text{ kcal/mol}$), respectively. Deviations of 0.15 kcal/mol or more are found for only two further molecules out of the complete set (dioxirane, $-0.65$ versus $-0.48$; tetraazaadamantane, $-2.63$ versus $-2.48$), and we note that the sizable corrections for large organic molecules are reproduced very faithfully by the BSR model (e.g., anthracene, $-2.67$ versus $-2.67$; hexafluorobenzene, $-2.38$ versus $-2.42$; biphenyl, $-2.29$ versus $-2.29$; trans-decalin, $-2.22$ versus $-2.18$). Linear regression analysis confirms the excellent correlation between estimated and calculated values ($b = 0.998, a = -0.012 \text{ kcal/mol, } R^2 = 0.989$). The success of the BSR model infers that scalar relativistic effects on bond formation are very local and hardly influenced by the nature of neighboring bonds. This conclusion is further corroborated by the previous findings that an

(a = 0.012 kcal/mol), and a large correlation coefficient ($R^2 = 0.990$).

Thus it appears perfectly justified to use UB3LYP as reference to assess the accuracy of the simple bond-additivity model that is based on the assumption of thermoneutral BSRs. To ensure a consistent level of treatment, bond increments need to be derived for the same method (UB3LYP) using results obtained for BSR prototypes. For purposes other than this assessment, more appropriate CCSD(T) derived bond increments will be used.

The test set includes 172 molecules (HCNOC) all of which have three nonhydrogen atoms or more, i.e., at least two bonds between nonhydrogen atoms. Figure 8 collects all results in graphical form and highlights data obtained for the subset of 73 molecules discussed in Sec. V. Most of the 99 remaining molecules (Table S390) are larger organic species with substantially larger relativistic corrections to atomization energies.

(b = 1.066), a small value for the intercept ($a = -0.012 \text{ kcal/mol}$), and a large correlation coefficient ($R^2 = 0.990$).
The empirically calibrated bond order model works remarkably well for predictions of scalar relativistic corrections, and that the accuracy of G3 for thermochemical applications does not improve upon explicit calculation of these corrections and successive reoptimization of empirical parameters.

B. Non-Born–Oppenheimer effects

DBOCs to atomization energies have initially been computed at the UHF/aug-cc-pVTZ level, following the scheme implemented in the HEAT model chemistry. Figure 9 shows, however, that evaluations with the smaller cc-pVDZ are absolutely sufficient for our purposes so that we decided to use UHF/cc-pVDZ based DBOCs not only to assess the quality of the BSR model but also to define the standard of our thermochemical approach. Note that the values computed for the set of 32 small molecules cover a range of less than 0.3 kcal/mol and that the largest deviation between UHF/cc-pVDZ and UHF/aug-cc-pVTZ amounts to a negligible 0.02 kcal/mol.

![FIG. 7. Scalar relativistic corrections to atomization energies evaluated at the UHF-CCSD(T)(full)/aug-cc-pVTZ and UB3LYP/cc-pVTZ levels of theory. The test set includes all 32 molecules marked “a” in Table S3 (Ref. 90) all of which have two nonhydrogen atoms or less.](image1)

![FIG. 8. Comparison between estimated (BSR scheme, assuming thermoneutral reactions) and calculated scalar relativistic corrections. All data refer to the UB3LYP/cc-pCVTZ level of theory. The complete set of 172 molecules (filled circles) includes all those not marked “a” in Table S3 (Ref. 90) (three nonhydrogen atoms or more); the smaller set of 73 molecules (crosses) refers to our standard test set (Sec. V).](image2)

![FIG. 9. Diagonal Born-Oppenheimer corrections to atomization energies evaluated at the UHF/aug-cc-pVTZ and UHF/cc-pVDZ levels of theory. The test set includes all 32 molecules marked “a” in Table S3 (Ref. 90) all of which have two nonhydrogen atoms or less.](image3)

![FIG. 10. Comparison between estimated (BSR scheme, assuming thermoneutral reactions) and calculated diagonal Born-Oppenheimer corrections. All data refer to the UHF/cc-pVDZ level of theory. Included are all 172 molecules not marked “a” in Table S3 (Ref. 90) (three nonhydrogen atoms or more). Filled circles denote molecules with two carbon atoms or more and crosses those with zero or one carbon atoms. The two dashed lines enclose an area to the upper right and lower left parts of the graph, for which inclusion of the BSR estimate establishes improvement over neglect of diagonal Born-Oppenheimer corrections.](image4)
Figure 10 compares BSR estimates of DBOCs with calculated values for the set of 172 molecules already studied in Sec. VI A. Obviously, results are somewhat mixed. The assumption of thermoneutral BSRs does not work nearly as well as for relativistic corrections, and both visual inspection and linear regression analysis (\(b=0.826, a=-0.053\) kcal/mol, \(R^2=0.745\)) indicate that DBOCs are almost always underestimated. The largest error is observed for tetraazaadamantane (BSR: 0.15 kcal/mol versus calculated: 0.44 kcal/mol). Simple scaling would improve overall agreement with the reference data, but such an empirical option was not considered here.

On the other hand, DBOCs are fairly small altogether and considered only in the most elaborate \textit{ab initio} thermochemistry protocols. A less than optimal but qualitatively reasonable estimate may thus still be helpful. We note that on average, 0.44 kcal/mol, these cases contain at least three nonhydrogen atoms. Figure 12 analyzes the BSR model for CCSDTQ-CCSDT corrections for the entire set of 45 molecules, all of which were fully relaxed CCSDTQ calculations could be performed. All data refer to total energies evaluated with cc-pVDZ basis sets. A less than optimal but qualitatively reasonable estimate may thus still be helpful. We note that on average, 0.44 kcal/mol, these cases contain at least three nonhydrogen atoms. Figure 12 analyzes the BSR model for CCSDTQ-CCSDT corrections for the entire set of 45 molecules, all of which were fully relaxed CCSDTQ calculations could be performed. All data refer to total energies evaluated with cc-pVDZ basis sets. Using this approximation, (CCSDTQ-CCSDT) contributions to atomization energies were evaluated for a total of 22 molecules, supplementing the set of 23 molecules for which fully relaxed CCSDTQ calculations could be performed. Table S5 lists CCSDTQ-CCSDT and CCSDT-CCSD(T) corrections for the entire set of 45 molecules, all of which contain at least three nonhydrogen atoms. Figure 12 analyzes the quality of the BSR model for the sum of both corrections which is of course the quantity of major interest.

C. CCSDT and CCSDTQ corrections

The assessment of the BSR model for CCSDTQ-CCSDT corrections is seriously restricted by the availability of appropriate reference data. CCSDTQ/cc-pVDZ calculations proved difficult even for some of the larger BSR prototypes such as ethane or the methylammonium cation, mainly reflecting memory and scratch disk limitations. Clearly, only very few systems with three or more nonhydrogen atoms can be studied at this level of theory, making a meaningful comparison between calculated values and BSR estimates almost impossible. A further reduction in basis-set size does not appear appropriate, but both the CCSDT[Q] (Refs. 127 and 128) and the more recently developed CCSDT(Q) model with (quasi)perturbative estimates of quadruple excitations hold promise for relief.

Bomble \textit{et al.} demonstrated for a set of four atoms and 24 small (two to four atomic) radicals and closed-shell molecules that both CCSDT[Q] and CCSDT(Q) provide good estimates of CCSDTQ correlation energies, with absolute (unsigned) values generally being slightly underestimated by CCSDT[Q] and slightly overestimated by CCSDT(Q), i.e., \(|Q|=\text{CCSDT}[Q]-\text{CCSDTQ}>0\) and \(|Q|=\text{CCSDT}(Q)-\text{CCSDTQ}<0\). Using the MRCC code interfaced to ACESII \cite{115}, we have extended this analysis to the subset of 55 molecules of our data set for which CCSDTQ/cc-pVDZ calculations proved feasible (Fig. 11). Indeed, with only two minor exceptions, \(|\Delta|=\text{CCSDT}[Q]/\text{CCSDTQ}(Q)\), singlet methylene, 0.06/0.02 kcal/mol; fluorine, –0.04/ –0.09 kcal/mol], CCSDT[Q] and CCSDT(Q) are found to be upper and lower bounds of the CCSDTQ energy, respectively. Furthermore, CCSDT(Q) generally appears to be the better of the two approximations to CCSDTQ, making it worthwhile to consider scaled [CCSDT(Q)-CCSDT] energy differences as replacement for (CCSDTQ-CCSDT). The lower panel of Fig. 11 shows, with a scaling factor of 0.85 and indicates that CCSDTQ energies are faithful reproduced to within 0.1 kcal/mol if the difference between CCSDT[Q] and CCSDT(Q) is 1 kcal/mol or less. Larger deviations between the two perturbational treatments generally indicate that none of them is very accurate, making predictions based on scaling procedures less reliable. Scaled [CCSDT(Q)-CCSDT] energy differences are thus only used if this criterion is met, and they are reported with an estimated error bar of 0.1 kcal/mol.

Using this approximation, (CCSDTQ-CCSDT) contributions to atomization energies were evaluated for a total of 22 molecules, supplementing the set of 23 molecules for which fully relaxed CCSDTQ calculations could be performed. Table S5 lists CCSDTQ-CCSDT and CCSDT-CCSD(T) corrections for the entire set of 45 molecules, all of which contain at least three nonhydrogen atoms. Figure 12 analyzes the quality of the BSR model for the sum of both corrections which is of course the quantity of major interest.

![Approximations to CCSDTQ](image-url)

**FIG. 11.** Approximations to CCSDTQ. The top graph shows deviations of the (quasi)perturbational estimates CCSDT[Q] (filled circles) and CCSDT(Q) (open circles) from CCSDTQ energies as a function of the difference \(|\Delta|=\text{CCSDT}[Q]-\text{CCSDTQ}\). The lower panel shows deviations of CCSDT(Q)-CCSDT(Q) from CCSDTQ energies as a function of the difference \(|\Delta|=\text{CCSDT}[Q]-\text{CCSDTQ}\) for individual results. All data refer to total energies evaluated with cc-pVDZ basis sets.
such as cyclopropane and propane (C—C / C—C bonds), allene and butatriene (C=C/C=C), propene and cyclopropane (C=C/C=C), propyne and 2-butyne (C=C/C=C), difluoromethane, oxirane, aziridine, and dimethyl ether (C=C / C—X), glyoxal and acetaldehyde (C—C / C=O), and acetonitrile (C—C/C=O)]. For most of these molecules, the BSR predictions are about as accurate as CCSDT(Q) based estimates of CCSDTQ corrections, for which we have assumed error bars of ±0.1 kcal/mol.

Finally, we note that overall results are quite satisfactory also on a statistical basis (ozone and nitrosyl fluoride excluded, see above), with mean absolute and rms deviations of 0.08 and 0.10 kcal/mol, respectively, and absolute values ranging from −0.53 kcal/mol (2-butyne, BSR: −0.45 kcal/mol) to 0.65 kcal/mol (difluorine monoxide, BSR: 0.50 kcal/mol).

VII. ENTHALPIES OF FORMATION

The use of BSR increments for MVD, diagonal Born–Oppenheimer, CCSDT-CCSD(T), and CCSDTQ-CCSDT corrections, evaluated from explicit \textit{ab initio} calculations for BSR prototypes (Table III), concludes the definition of the ATOMIC protocol. Spin-orbit coupling terms still need to be considered, but for closed-shell molecules they contribute to atomization energies only via corrections of the atomic energies which may easily be derived from experimentally observed fine structures.\textsuperscript{34,75} The missing link to experimental enthalpies of formation is the evaluation of zero-point energies, of thermal enthalpies, and of thermochemical data for the atoms. Atomic data (enthalpies of formation and thermal enthalpies) are taken from experiment\textsuperscript{130} as is standard in most \textit{ab initio} thermochemical protocols based on atomization energies. Molecular zero-point energies and thermal enthalpies are evaluated in the rigid-rotor and harmonic-oscillator approximation using scaled harmonic frequencies. A scaling factor of 0.98 has been derived for the RI-MP2/cc-pVTZ level of theory which affords the best agreement between calculated harmonic zero-point energies and experimental (or high-level \textit{ab initio}) reference data which take anharmonicity effects into account. Full details of the protocol and results for a large set of first-row compounds will be given elsewhere.\textsuperscript{71} Here we focus on a statistical evaluation for the subset of neutral, closed-shell HCNOF molecules of the G3/99 standard test set,\textsuperscript{131} excluding molecules with two nonhydrogen atoms or less and discarding also tetrafluoroethylene\textsuperscript{132,133} and carbonyl fluoride\textsuperscript{134} where the results of high-level \textit{ab initio} benchmark studies have cast some doubt on the accuracy of experimental data. The statistical analysis (Table IV) confirms our earlier conclusion that BSR corrections reduce systematic errors in a very effective manner: The MSE is reduced more than tenfold, and the MUE not only becomes much smaller but is also much less model dependent. The ATOMIC protocol achieves an accuracy that is either comparable to (model C) or slightly better (models A and B) than the popular G3 approach, without requiring empirical calibration. The recently introduced G4 method represents a significant improvement over G3 at the expense of somewhat more costly
TABLE IV. Statistical analysis of composite methods A–C for enthalpies of formation [\(\Delta H_f(298.15)\)].

<table>
<thead>
<tr>
<th></th>
<th>Fully corrected (ATOMIC)\textsuperscript{ab}</th>
<th>Partially corrected\textsuperscript{bc}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B\textsubscript{1}</td>
</tr>
<tr>
<td>MUE</td>
<td>0.59\textsuperscript{e}</td>
<td>0.64\textsuperscript{e}</td>
</tr>
<tr>
<td>MSE</td>
<td>0.25</td>
<td>−0.05</td>
</tr>
<tr>
<td>rms</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Max. (+)</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Max. (−)</td>
<td>−1.7</td>
<td>−1.9</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The data set includes all 103 neutral closed-shell HCNOF molecules from the G3/99 set which have three nonhydrogen atoms or more, except tetrafluoroethylene and carbonyl fluoride for which experimental data are suspect. Errors are given with respect to experiment (in kcal/mol). See Table I for a definition of methods.

\textsuperscript{b}Includes all BSR corrections from Tables II and III, defining our standard approach.

\textsuperscript{c}Includes BSR terms only for MVD, DBOC, CCSDT-CCSD(T), and CCSDTQ-CCSDT corrections (Table III). BSR corrections toward the complete basis-set limit CCSD(T) energy (Table II) are omitted.

\textsuperscript{d}Compiled from data reported in the supplementary material to Ref. 35.

\textsuperscript{e}The MUEs are reduced to 0.53 kcal/mol; and only slightly too unstable short 10.7 bonds (4.4 nonhydrogen) per molecule, Table IV, versus 4.8 (2.7, Table I). Somewhat surprisingly, however, not much is gained in using the more accurate model A and not much is lost in using the less accurate model C, indicating that shortcomings in the BSR approach to reproduce CCSD(T) atomization energies may account for only a modest fraction of the total error, and that other sources contribute, leading to either accumulation (model A) or cancellation of error (model C). Note in this context that model A suffers from a much larger MSE when compared to experiment (Table IV) than one would expect from the evaluation of CCSD(T) atomization energies (Table I), while exactly the opposite is true for model C. Potential sources of additional error include inaccuracies in geometries, zero-point energies, and thermal enthalpies, occasional problems with multireference cases such as ozone (see above), and also experimental error.

VIII. DISCUSSION AND CONCLUSIONS

A new \textit{ab initio} thermochemical protocol has been developed, implemented, and tested for a large set of neutral closed-shell molecules composed of first-row atoms (C, N, O, F) and hydrogen. The protocol is geared toward applications in organic chemistry and focuses on accurate calculations of atomization energies and enthalpies of formation. It makes extensive use of the concept of bond-separation reactions (BSRs), thereby supplying very high-level \textit{ab initio} data for the set of parent molecules and simple hydrides (BSR prototypes) involved in order to improve on lower-level treatments for larger molecules. Consistent formulation of BSRs in an \textit{ab initio} framework eliminates implementation restrictions otherwise imposed by the limited availability of accurate experimental reference data. It is no problem, for example, to treat even “exotic” molecules such as dinitrogen trioxide, as the necessary reference data for BSR prototypes (ammonia, ammonia oxide, ammonium and hydrazinium ion, neutral and protonated nitroxyl) can easily be generated by computation. The application of BSR schemes obviously precludes the construction of entire potential energy surfaces, but it allows for accurate and cost-effective complete basis-set estimates of atomization energies at the CCSD(T) level of theory, further corrected for relativistic, diagonal Born–Oppenheimer, and higher-order electron correlation effects. The name ATOMIC has been chosen as an abbreviation of the title of this paper, emphasizing the aspect of optimal balance in accuracy and cost between the various components of the model and the use of isodesmic corrections to reduce systematic error in a controlled fashion. Furthermore, the chosen acronym is meant to underline the focus on optimization energies, and, thus, the use of isolated atoms as absolute reference.

A reference level of theory is defined based on CCSD(T)(full)/cc-pCVXZ (X=Q,S) calculations extrapolated to the complete basis-set limit and further corrected by (56) extrapolations of the valence-shell CCSD contribution. Atomization energies are decomposed into four valence-shell and four inner-shell components of increasing computational complexity [HF, MP2-HF, CCSD-MP2, CCSD(T)-CCSD]. Minimum basis-set requirements are then established for each of the eight components based on comparisons with the complete basis-set reference data. As expected, it turns out that only moderate basis-set reductions are possible if high
accuracy is desired. Use of BSRs reduces basis-set requirements dramatically, and all inner-shell contributions except MP2 may safely be ignored, while triple-zeta basis sets suffice for the (CCSD-MP2) valence-shell component. BSRs can easily be mapped onto a system of corrective bond increments which are evaluated from reference level calculations for BSR prototypes (Table II).

Three levels of accuracy are defined based on rms errors evaluated for a diverse set of 73 molecules with three to six nonhydrogen atoms: A (0.1 kcal/mol), B (0.3 kcal/mol), and C (1.0 kcal/mol). One implementation each is suggested for the highest and lowest levels, and three different implementations labeled B₁, B₂, and B₃ are presented for medium accuracy. The ab initio models are defined by their basis-set and/or extrapolation requirements for HF (X₁), valence-shell MP2-HF (X₂), CCSD-MP2 (X₃), and CCSD(T)-CCSD (X₄) contributions and corresponding inner-shell contributions X₅ to X₈ (see Sec. IV A for notation, Sec. IV B for extrapolation, Table I for definition of models). The best balance of accuracy and computational requirements is probably achieved by model B₃ which uses a mixed basis set (3': cc-pVDZ on hydrogen, cc-pVTZ on first-row atoms) for CCSD calculations. Models B₁ and B₂ are slightly more expensive but may be used to estimate residual errors inherent in the BSR scheme by simply calculating the inner-shell MP2 energy. Model C is an attractive low-cost option as it only requires two calculations to be performed: CCSD(T)/cc-pVQZ and MP2/cc-pVTZ.

Scalar relativistic, diagonal Born–Oppenheimer, and higher-level electron correlation [CCSDTQ-CCSD(T)] corrections are treated under the simplifying assumption that the corresponding BSRs are thermoneutral, rendering the calculation a simple task of adding up bond increments (Table III). With applications to larger (organic) molecules in mind, such simplification is absolutely necessary in the case of higher-order excitation contributions but desirable also for other contributions whose evaluation may otherwise be complicated by implementation restrictions. One should keep in mind that only few quantum-chemical codes allow for the evaluation of DBOCs and relativistic corrections (at the correlated level). Extensive validation with 172 molecules including compounds as large as anthracene shows that the assumption of thermoneutral BSRs leads to remarkably accurate results in the case of scalar relativistic corrections (rms error: 0.05 kcal/mol, range of reference values: −0.17 to −2.67 kcal/mol). Non-Born–Oppenheimer effects are much less local, and the BSR scheme consistently underestimates DBOCs. On the other hand, DBOCs are smaller overall, results are still reasonable on a qualitative level (rms error: 0.11 kcal/mol, range: −0.04 to 0.54 kcal/mol), and no dramatic failures have been observed. More limited validation (45 molecules) was also performed for CCSDTQ-CCSD(T) corrections, although the computational expense of fully relaxed CCSDTQ calculations has forced us to use estimates based on CCSDT(Q) and CCSDT[Q] as reference data for about half of the test set. Not unexpectedly, the BSR model fails for typical multireference cases with higher-order excitations that are nonlocal in character (ozone being the most prominent example), but it proves to be reasonably ac-

accurate for “simpler” molecules mainly consisting of single bonds and isolated multiple bonds, where small effects on a bond-by-bond basis may add up to sizable corrections. Excluding ozone and nitrosyl fluoride, the observed accuracy is quite respectable overall (rms error: 0.10 kcal/mol, range: −0.53 to 0.65 kcal/mol).

The overall accuracy for typical applications in thermochemistry can only be judged from comparisons with experimental enthalpies of formation. Preliminary validation with (HCNFO) data taken from the G3/99 test set shows very promising results and indicates that our protocol achieves a good balance between the various components contributing to enthalpies of formation, although residual errors in contributions other than the CCSD(T) energy seem to foil the superb performance of the most refined model A. On the whole, the ATOMIC protocol performs slightly better than the G3 thermochemistry model, and it does so without the need of empirical calibration. Indeed, one of the most important design principles followed in this work was to avoid reference to experimental data in guiding theoretical development and rather assess theoretical requirements separately for each component of the thermochemical model in order to reduce the risk of fortuitous and uncontrolled error cancellation. The consistent application of BSRs as an integral part of the protocol eliminates the need for dauntingly expensive calculations normally needed if no calibration is permitted and makes the protocol efficient enough to treat systems with about 10–20 nonhydrogen atoms on a standard basis.²¹

So far we have only considered molecules composed of hydrogen and first-row (C, N, O, F) atoms. The ATOMIC approach is not limited to these elements, however, and should in principle be applicable to all molecules for which reasonable valence structures can be drawn, provided that the corresponding BSR prototypes are well-defined stable species. This requirement rarely poses a problem in the chemistry of the elements considered here,²² but demands further investigation for possible extensions of the approach. With an eye toward organic and biochemistry it seems most important to extend coverage to second-row elements and to P, S, and Cl, in particular. Results will be reported in due course.

ACKNOWLEDGMENTS

The author thanks the Swiss National Supercomputing Center (CSCS, Manno TI) for a generous grant of computer time.

EPAPS, see http://www.aip.org/pubservs/epaps.html.


32 M. Källay, mrcc, a string-based quantum-chemical program suite, see also www.mrcc.hu.


53 A notable exception is dinitrogen tetroxide whose canonical valence structure suggests a doubly protonated hydrazine as parent molecule which is of course purely repulsive and does not exist.