Stepwise Basis Set Selection

M. W. Li, P. M. Zimmerman*

Department of Chemistry, University of Michigan, Ann Arbor, MI 48104.

*paulzim@umich.edu

Abstract

The computational cost of quantum chemical methods grows rapidly with increasing level of theory and basis set size. At increasing costs, higher accuracies can be reached, forcing a compromise between cost and accuracy for most molecular systems. Heats of reaction, however, are mostly determined by a subset of atoms that experience significant bonding and/or electronic changes. To exploit this fact, the Stepwise Basis Builder (SBB) algorithm selectively adds basis functions to reactive atoms and maintains small basis sets on spectator atoms. This article introduces the SBB algorithm and how it chooses a basis for each atom, predicts calculation errors, and uses these predicted errors to reach target levels of accuracy. Benchmarks show SBB heats of reaction and activation barriers converge to values consistent with higher-quality calculations using a greatly reduced number of basis functions.

Keywords: Basis set selection, basis set extrapolation, computational cost, computational scaling, stepwise selection

Introduction

Computational quantum chemistry has become a widely used, powerful research tool because it allows many physically relevant properties to be predicted from first principles. While the Schrödinger equation provides a formally exact representation for any chemistry, standard quantum chemical studies must make approximations to avoid intractable computational costs. In general, increasing levels of theory and basis set size can improve the quality of electronic structure computations, with massive costs at the highest accuracy levels. Usually some compromise must be made by creating simpler molecular models, using lower levels of theory, and/or reducing the basis set quality. This study focuses in on the third component, and presents a new means of selecting a small basis set while achieving a target degree of accuracy.

The algorithmic scaling of quantum chemical methods is central to understanding their high cost and the size of basis that may be used in practice. At the second order Møller-Plesset (MP2) level of perturbation theory, molecular energies can be calculated at $O(n^5)$,[1] where n is the number of atoms in the system. A more accurate method, coupled cluster theory with single and double excitations (CCSD), scales as $O(n^6)$.[2-3] In MP2, the scaling with respect to number of basis functions is $O(N^3)$, and in CCSD, $O(N^4)$. In each case, linear scaling approaches (with respect to number of atoms in the system) have been developed to reduce the computational burden, at the cost of introduction of systematic errors.[1,4-22] Despite the linear scaling, these methods still grow superlinearly in cost with increasing basis set size, though some progress has been made to mitigate these costs using explicit correlation methods like R12.[23-27] Generally speaking, to achieve the full accuracy of MP or CC levels of theory, large basis sets must be used.

The complete basis set (CBS) limit represents the high accuracy, high computational cost limit for basis set selection. Extrapolation to the CBS limit[3,28-29] takes advantage of convergence with

successively larger basis sets, and aims to eliminate any errors from choice of basis set.[4-5,30-64] Basis set truncation,[65] dual basis approaches,[66] and related extrapolations[3] have also been used to approach the CBS limit. The quality of CBS extrapolation using various basis sets, however, is dependent on the specific chemical systems and methodologies used,[33-34] as has been shown by various benchmarks.[58-64,67-68] Basis sets of at least triple-zeta quality—and often larger—are needed to approach this limit.

Approximate wave function methods do not reach arbitrarily high levels of accuracy compared to the exact electronic energy, even in the CBS limit. Additionally, since the CBS limit is not always attainable due to high costs, we pose a different question than prior studies in this area. In our view, we ask: in order to reach a specified level of accuracy, what basis set is required? The goal of the present article is to build a methodology to answer this question. In short, the proposed method will tailor basis sets for molecular systems to achieve a desired accuracy level, while at the same time keeping those basis sets small so the computations will remain tractable. The electronic structure theory will remain fixed, and therefore the accuracy represents the distance from the large basis limit for that level of theory.

This study is further motivated by three key observations: 1) judicious choice of basis set may entail considerable savings with arbitrarily small loss in accuracy, 2) the effectiveness of additional basis functions is dependent on which atoms they are applied to, and 3) increasing level of theory and basis set quality allows for convergence of calculated energies to more accurate values. These observations suggest that iterative addition of select basis functions provides a route to reach convergence of computed energies. This study specifically will consider reactive systems, where in particular, reacting atoms require larger basis sets than nonreactive atoms. By examining the response of the energy to changes in individual atomic basis functions, estimates of the size of basis required for each atom can be obtained.

Herein is described a *Stepwise Basis Builder* (SBB) algorithm that assigns basis functions for each atom individually, with energetic errors limited by a specified threshold. To control the error from truncating the basis, the relative error with respect to a change in basis on each atom is measured. A subset of the atoms which dictate significant changes in energy are given a larger basis, and the process repeated until convergence. This strategy will be shown to be particularly advantageous where the target quantities are energies of reaction, and cancellation of errors can lead to dramatic reductions in computational cost.

Theory

SBB will be presented after an introduction of standard energy calculations, where basis sets are fixed (e.g. the same Pople or Dunning type basis set is used for each atom). The energy, E, of molecular structure M is a function of the geometry and the basis:

$$E = f(M, g)$$

with g denoting the basis and f being a function specific to the level a theory (e.g. Hartree-Fock, MP2, or CCSD). Using the Dunning cc-pVXZ type basis sets (with X ranging from 2 to ~8 for double zeta and larger basis sets), the basis size |G| scales as:

$$|G| \approx 2^X$$

Complete convergence (i.e. μHa) of the total energy E typically occurs for $X > 5$, which is impractical for most systems of interest.[29,69-70] Absolute energy calculations, however, are not the most practical goal in quantum chemistry. Often relative energies (e.g. thermochemical energies and activation energies) are the key quantities,

$$\Delta E = f(M_2, g) - f(M_1, g) = f'(M_1, M_2, g)$$

with M_1 representing the structure of the reference molecule and M_2 being the second (i.e. the product or the transition state). The relative energy is a new function, f', of the two structures and the basis used. If M_1 and M_2 are fixed (i.e. already known), then the relative energy can be further simplified to

$$\Delta E = f''(g)$$

where the basis is the only important quantity. ΔE usually is converged to within 1 kcal/mol for $X = 3$ or 4, allowing $X > 4$ computations to be avoided. For example, considering a molecule with n atoms, applying RI-MP2, and taking |G| to be 2^X,

$$\text{Cost} \sim n^5 \sim O^2 V^2 N_{aux} \sim (n)^2 (n \cdot 2^X)^2 (n \cdot 2^X) \sim (2^X)^3$$

with O being the number of occupied orbitals, V being the number of virtual orbitals, and N_{aux} being the number of auxiliary functions used in the RI approximation. Under this scaling law, reduction of the basis size from quintuple zeta to quadruple zeta results in a ~10-fold reduction in computational cost.

So far the basis set was assumed to be the same for each atom. Now, consider the following expression for the basis:

$$g = \bigcup_i g_i \; ; \; i \in A$$

where A is the set of atoms in the molecule in question, i indexes the atoms in A, and g_i is the basis set used for each atom. Each atom's basis is flexible, giving

$$g_i = \text{cc-pV}X_i\text{Z}$$

To motivate this assumption, see the example shown in Figure 1, where it is obvious that ΔE depends more strongly on basis functions in the central region of the molecule.

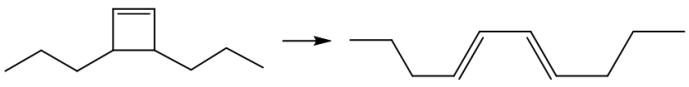

Figure 1. 1,4 dipropyl cyclobutene ring opening.

The problem of interest is determining how to reduce the total number of basis functions with minimal loss in accuracy. The ring opening reaction depicted in Figure 1 has reactive atoms only in the ring and the propyl chains are spectators. ΔE is highly dependent on basis functions on the four carbons of the ring, with gradually weakening dependencies as the distance increases from this center. The basis functions for each atom can be selected by the iterative method outlined in Figure 2, which will be described in detail by the following derivation of the algorithm.

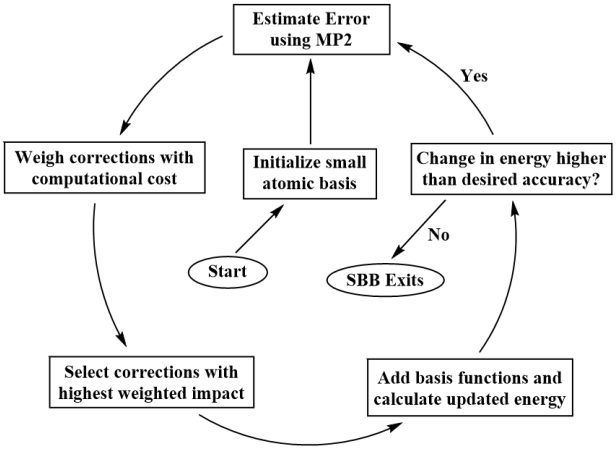

Figure 2. Overview of SBB algorithm.

The problem of choosing which size of basis to be used for each atom can be formulated,

$$\text{minimize} \sum_i |g_i|$$

where $|g_i|$ refers to the size of the basis on each atom i, with constraint

$$|\Delta E - \Delta E^{exact}| < \varepsilon$$

where ε is the desired accuracy level. Selecting the optimal basis without knowing ΔE^{exact}, however, is not a tractable problem. Assuming this quantity is not known, it is instead approximated using

$$\left|\Delta E - \Delta E^{ref}\right| < \varepsilon$$

The constraint can be understood in the context of forward-stepwise iteration.[71-73] Starting with a minimal basis set size, and iteratively adding functions to each atomic basis, g_i, fulfilling this constraint is the core action of the proposed SBB algorithm. ΔE^{ref} is the current best estimate of ΔE^{exact} and can be approximated as:

$$\Delta E^{ref} \approx \Delta E + \sum_i \frac{\Delta E^c}{\Delta g_i} \Delta g_i$$

with ΔE^c being the energy correction due to addition of basis functions (Δg_i) on atom i. Under this prescription, ε will be used to control the number of basis functions added on each iteration. $\left|\Delta E - \Delta E^{ref}\right|$ can thus be thought of as an estimate for the residual error, and ε is used to ensure that this error remains small. This error is approximate, as it depends on the current (small) basis set, and because adding atomic basis functions is not additive in the energy.

ΔE^{ref} measures the approximate effect of adding basis functions to *all* atoms. However, the algorithm must choose specific, reactive atoms for basis function addition. In order to select these atoms, the following objective function was chosen:

$$s_i = \left|\frac{\Delta E}{\Delta g_i}\right| \Delta g_i - \lambda X(g_i)^2$$

such that s_i is a measure of the impact of adding basis functions to atom i, weighted against the computational cost of adding those functions. λ is a regularization factor, and will be discussed in more detail shortly.

The largest values of s_i correspond to atoms i where additional basis functions are predicted to be most useful. The ordered list of s_i values may be expressed as:

$$B = \{s_m, s_n, s_p \ldots\} \text{ with } s_m > s_n > s_p, \ldots$$

Based on set B, the algorithm chooses m atoms corresponding to the first m s_i values for the calculation of the updated value of ΔE, denoted $\Delta E'$, such that

$$|\Delta E' - \Delta E| = \left|\sum_{j=1}^{m} \frac{\Delta E^c}{\left|\Delta g_{B_j}\right|} \Delta g_{B_j}\right| < \varepsilon$$

where the term Δg_{B_j} refers to the basis added to the atom corresponding to the jth element of B. All that remains is to choose λ, which up until now appears as an unknown regularization parameter. λ is chosen such that $|\Delta E' - \Delta E| < \varepsilon$ and m is minimized. This is achieved by scanning over a large range of λ values, and choosing the λ where m is smallest and the constraint is satisfied. The efficacy of this procedure will be demonstrated by numerical application in the results section.

Computational Details

Geometries were optimized using the B3LYP density functional[74-75] in a spin restricted formalism with the double-zeta, polarized 6-31G** basis set.[76] Electronic energies were computed for reactant-product pairs using Q-Chem 4.0.[77] These calculations utilize second order Møller-Plesset perturbation theory under the resolution-of-the-identity approximation (RI-MP2) with basis sets from the set of unpolarized double zeta 6-31G,[76] double-zeta cc-pVDZ, triple-zeta cc-pVTZ, and quadruple-zeta cc-pVQZ,[78] selected based on the method outlined in the **Theory** section. The RI basis set was chosen to be RI-MP2-cc-pVQZ throughout to mitigate errors coming from this approximation (this choice is not necessary, and RI basis could be chosen on an atomic level to match the primary basis). Energies are reported as gas phase electronic energies. All geometric structures can be found in the Supporting Information.

Results and Discussion

Benchmarks of the SSB method are presented using the basis sequence 6-31G, cc-pVDZ, cc-pVTZ, and cc-pVQZ. The smallest basis, 6-31G, is composed of 5 functions as [3s2p], and the largest, cc-pVQZ with 55 functions, [5s4p3d2f1g]. In SBB, we label these $X = 1$ through $X = 4$, which means we have introduced 6-31G at the beginning of the cc-pVXZ sequence. Since relative energies—not absolute energies—are of greatest interest, computations using the cc-pVQZ basis on all atoms are assumed to be reasonably converged, and therefore provide ΔE^{exact}. The ΔE^{exact} values from cc-pVQZ will be used to determine convergence of ε, providing a baseline for determining the SBB savings and basis set transferability between related reactions.

Tautomerism in Methyl Acetamide

Figure 3 shows H transfer from N to O in methyl acetamide, an example of a simple tautomerism reaction. At a threshold level (ε) of 1 mHa (0.63 kcal/mol), which has greater precision than the target 1 kcal/mol accuracy, SBB assigns the basis set depicted in Figure 3. Most of the basis functions are concentrated on the three reactive atoms of O, N, and H, where the cc-pVQZ basis is placed on N and O, cc-pVTZ on the reactive H, and cc-pVDZ on all other atoms. For this compact reaction, the smallest, unpolarized basis set (6-31G) introduces significant errors in the energy of reaction even for nominally spectator atoms. On the other hand, the largest basis sets are only needed on the reactive atoms to reach the specified level of convergence.

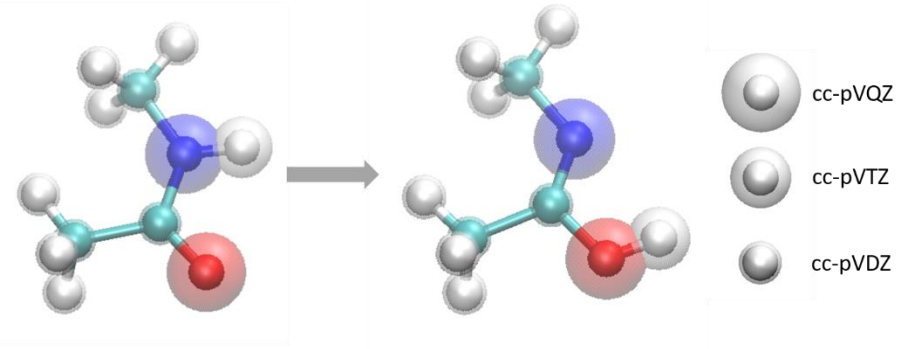

Figure 3. Amine tautomerization. Atom colors are as follows: blue – N, red – O, teal – C, grey – H. Basis size is shown for ε = 1 mHa.

A more rigorous analysis was performed using a sweep over ε to compare ΔE values to the full cc-pVQZ basis benchmark. At $\varepsilon = 10$ mHa, the computed ΔE is too high by almost 5 kcal/mol (8 mHa). Similarly, at $\varepsilon = 5$ mHa, the error is 1.3 kcal/mol (2 mHa). Errors of less than 1 kcal/mol are achieved at $\varepsilon \leq 2$ mHa, where ΔE = 11.45 kcal/mol, compared to the cc-pVQZ value of 11.55 kcal/mol. Errors fluctuate somewhat, but remain within 0.2 kcal/mol of each other for thresholds lower than 1 mHa. These errors therefore remain within chemical accuracy (1 kcal/mol) of the RIMP2/cc-pVQZ limit. At $\varepsilon = 1$ mHa, 196 basis functions were selected, compared to 485 total for the full cc-pVQZ computation. This reduction (Figure 4) corresponds to 60% fewer basis functions than the benchmark computation (incidentally, using the cc-pVTZ basis across all atoms corresponds to 248 basis functions, 27% more than SBB).

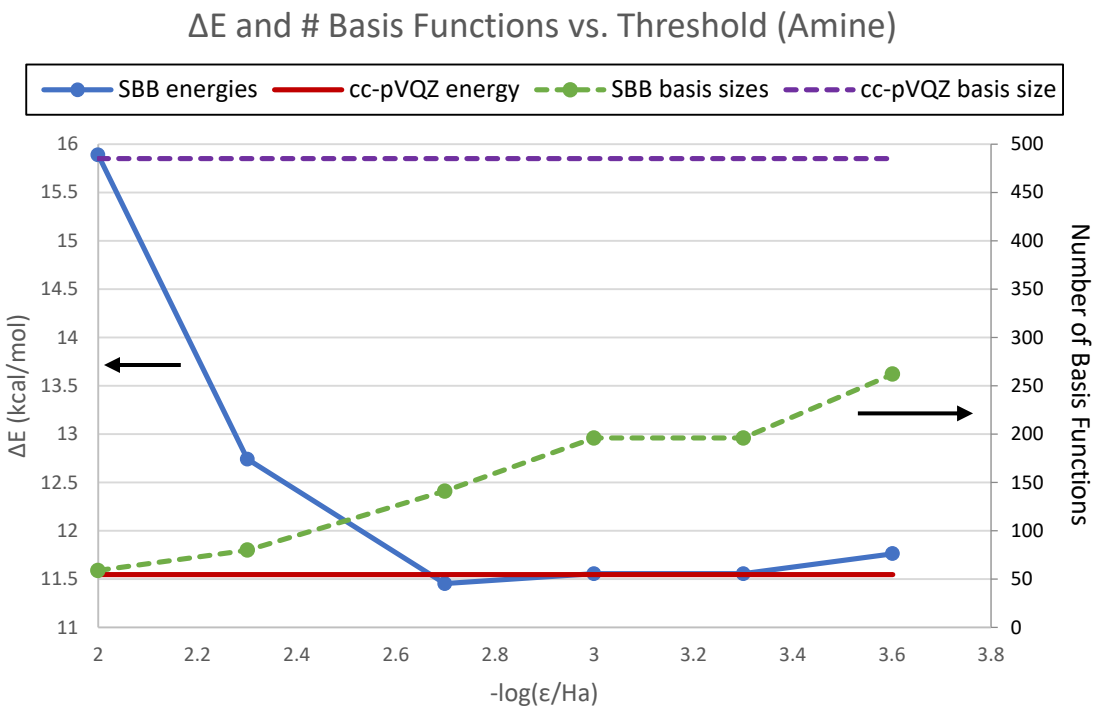

Figure 4. ΔE and number of basis functions versus -log ε in Ha.

Ene Reaction

The Ene reaction between ethylene and *cis*-butadiene is shown in Figure 5, which consists of a concerted hydrogen shift with C-C bond formation. Applying SBB with $\varepsilon = 1$ mHa, cc-pVQZ was placed on carbons 1 and 5, while cc-pVTZ was placed on carbons 2-4, between which π bonding orbitals were swapped (1,3 butene → 2,5 hexene). Carbon 6 received cc-pVDZ, as did all hydrogens except the reactive H, where SBB applied cc-pVTZ.

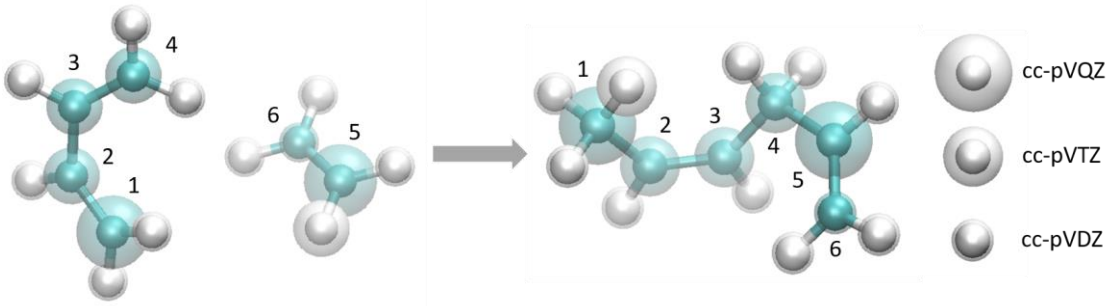

Figure 5. 1,5 hydride shift from ethylene to butene. Atom colors are as follows: teal – C, grey – H. Basis function distribution is shown at $\varepsilon = 1$ mHa.

Analysis of the number of basis functions and computed relative energies with respect to ε are provided in Figure 6. Convergence to within 1 kcal/mol of the full cc-pVQZ value (-27.21 kcal/mol) was obtained at a threshold of $\varepsilon = 2$ mHa (-26.78 kcal/mol). The error in reaction energy is decreased further, to 0.13 kcal/mol, at $\varepsilon = 0.25$ mHa. Analysis of basis function savings (Figure 6) shows that 273

basis functions were needed at $\varepsilon = 1$ mHa, compared to 630 for the full cc-pVQZ calculation (57% savings) and 320 basis functions for full cc-pVTZ (15% reduction). The smallest tested threshold, 0.25 mHa, resulted in 420 basis functions, or 33% savings.

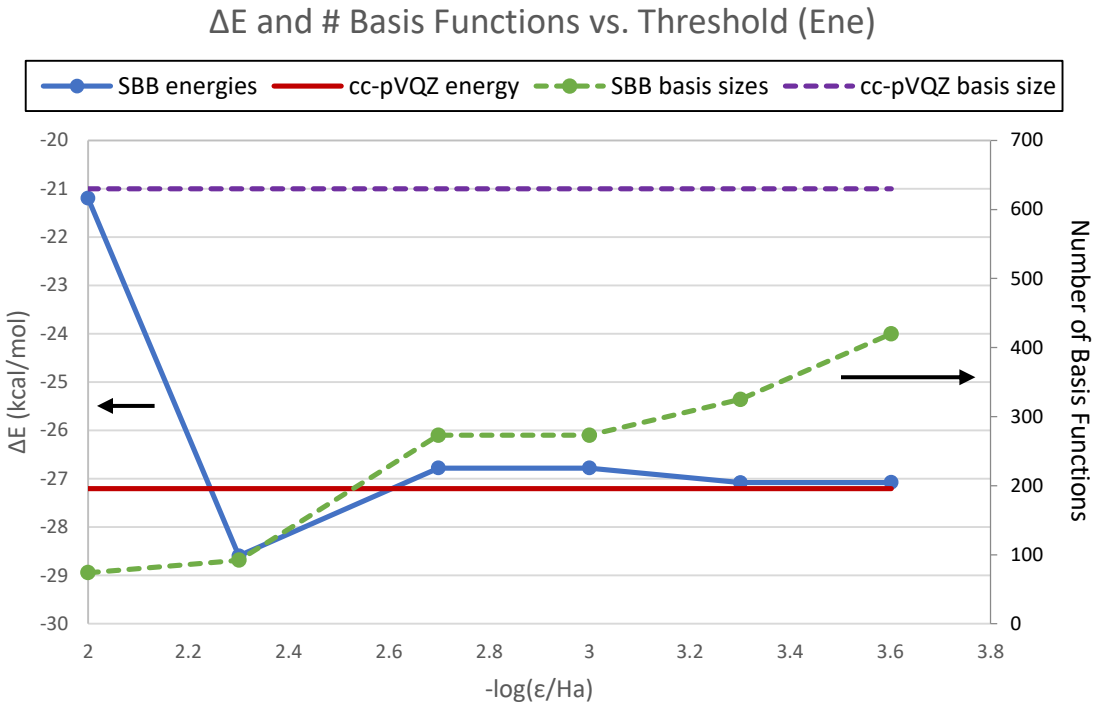

Figure 6. ΔE and number of basis functions versus -log ε for the *Ene* reaction.

Naphthalene Ring Expansion/Contraction

As a third example, naphthalene's cyclorearrangement to azulene is shown in Figure 7. As with the two prior examples, the cc-pVQZ basis is placed on atoms involved in the most significant changes of bonding, and cc-pVTZ is placed on neighboring carbons in the 5-membered ring of the product. 6-31G or cc-pVDZ is placed on remaining atoms, making this the first example where atoms were far enough from the reaction center to receive the small 6-31G basis.

ε analysis for the ring exchange shows that absolute errors are within 1 kcal/mol of the cc-pVQZ ΔE value (34.65 kcal/mol) at $\varepsilon = 2$ mHa (34.60 kcal/mol), and at the useful threshold of $\varepsilon = 1$ mHa, 296 basis functions were needed (compared to 790 for full cc-pVQZ, 412 for full cc-pVTZ; see Figure S1). With this third example in hand, the SBB method is showing considerable promise for basis set selection, saving 50% or more functions compared to full cc-pVQZ.

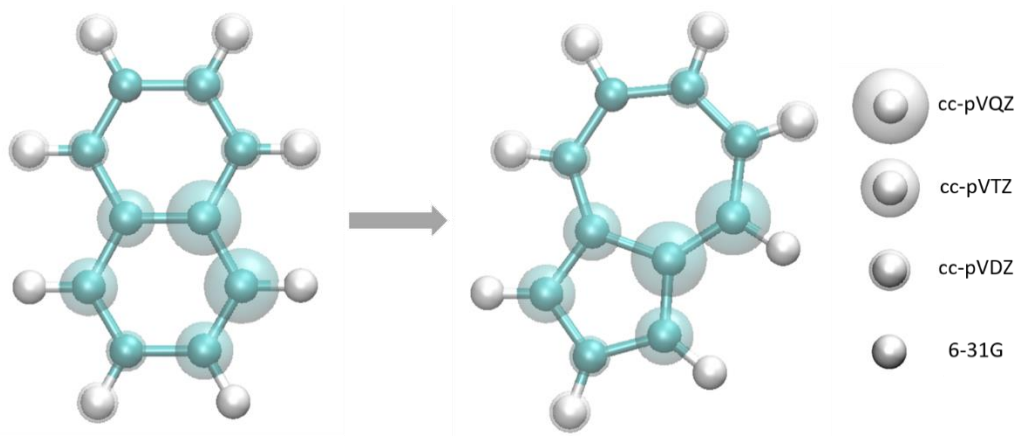

Figure 7. Naphthalene cyclorearrangement. Atom colors are as follows: teal – C, grey – H. Outer rings correspond to basis sets as shown in the legend. Basis function distribution is shown at ε = 1 mHa.

Ammonia-Borane (AB) and CO_2

The following five reactions (Figure 8) are related to carbon dioxide reduction by ammonia borane, a system of interest to carbon-neutral chemistry and materials development.[79] The polarized B-N and B=N bonds have been noted as leading to a wide variety of reactivity,[79-91] which becomes even more complex as C-O bonds are introduced as reactive partners.[79] Compared to the previous SBB examples, significant polarizations during reaction means more basis functions will be needed to capture this reactivity.

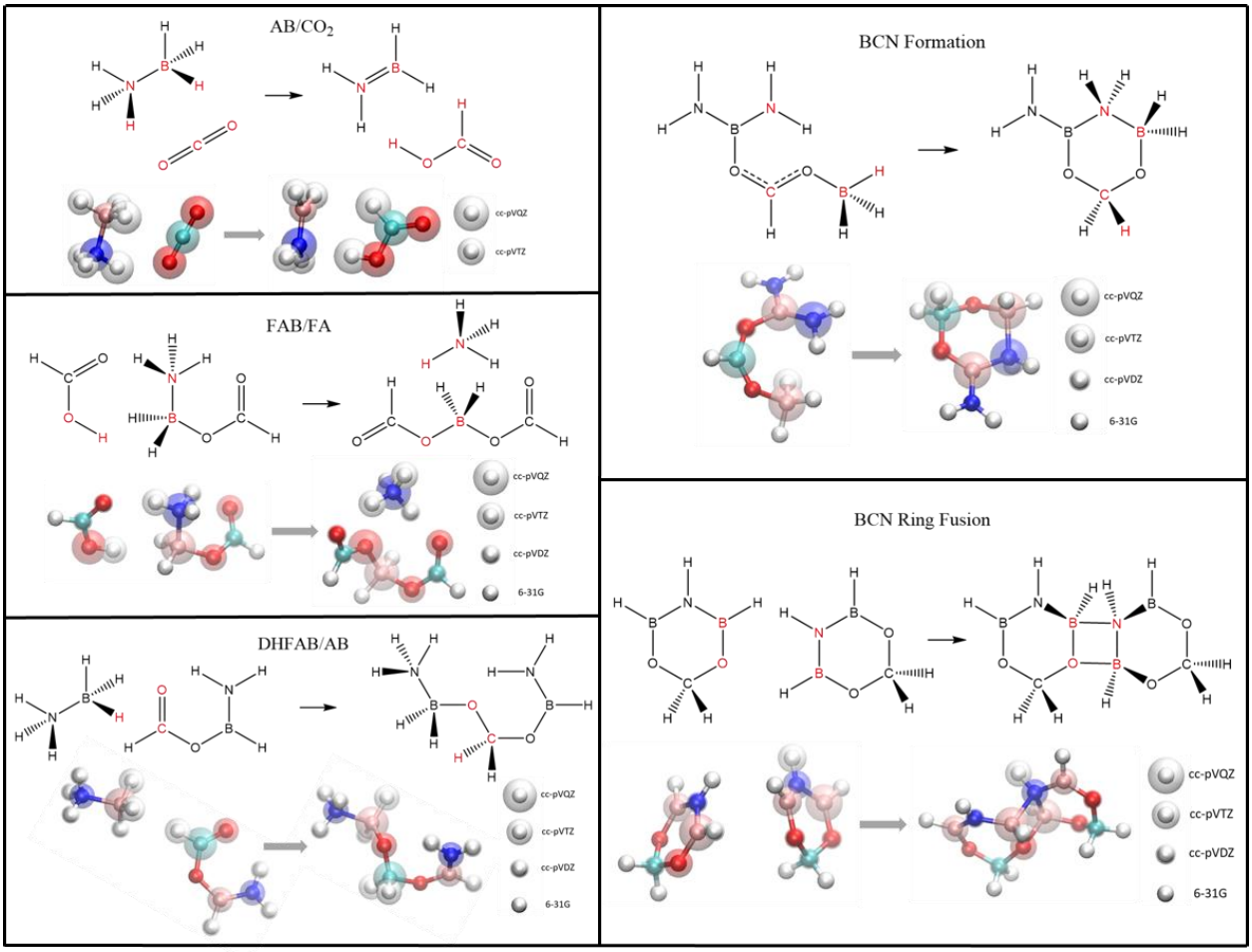

Figure 8. AB/CO_2 reactions and SBB basis for $\varepsilon = 1$ mHa. Key atoms are highlighted in red in 2D reaction mechanisms, where basis functions were expected to be focused on based on previous studies. Atom colors are as follows: blue – N, pink – B, red – O, teal – C, grey – H.

Figure 9 plots the relative error (compared to cc-pVQZ) in energies of reaction for the five AB/CO_2 test examples. At $\varepsilon = 1$ mHa, the largest error is for the ring fusion reaction (bottom right of Figure 8), at just under 1 kcal/mol. Other examples have errors under 0.5 kcal/mol, and all errors decrease with decreasing ε. This fine error control comes with varying cost in terms of number of required basis functions (Figure 10). At $\varepsilon = 1$ mHa, the 2H transfer from AB to CO_2 (the smallest system of the 5 examples) requires about 80% of the full cc-pVQZ basis. This is somewhat unsurprising, as each spectator atom is connected to at least one reactive atom. The other four cases, however, require 40 to 60% fewer basis functions, representing a significant reduction in number compared to the benchmark basis.

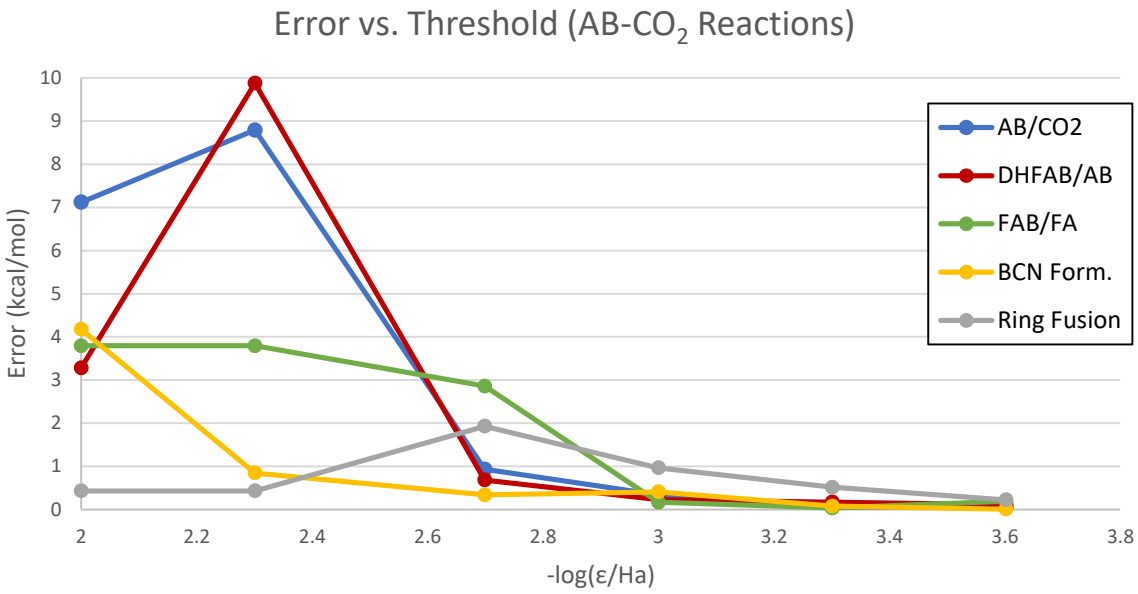

Figure 9. Error in ΔE versus -log ε for AB-CO_2 reactions.

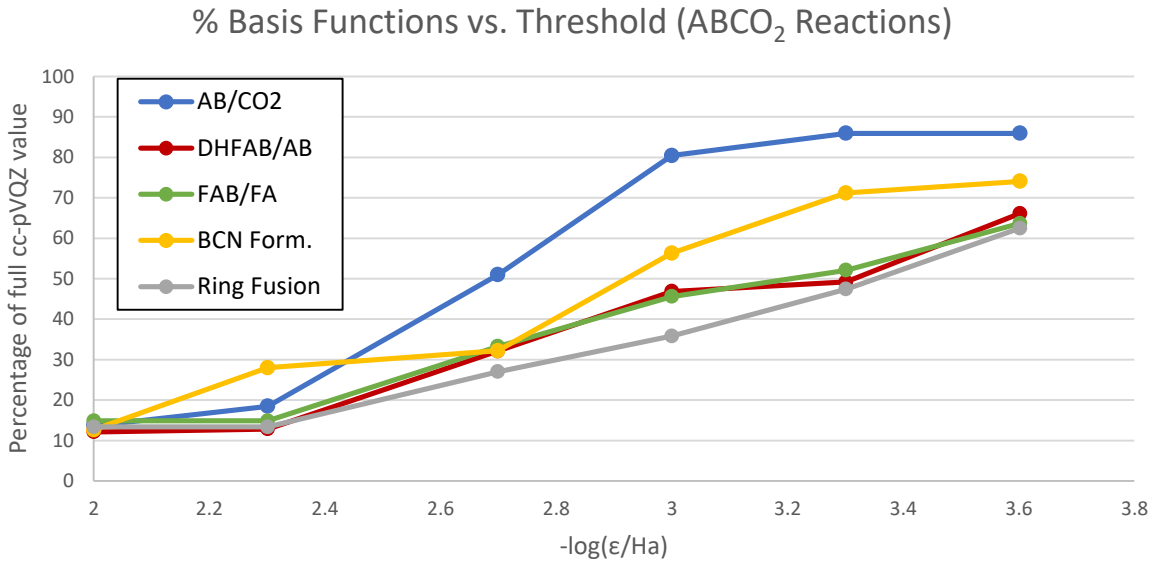

Figure 10. Percentage of basis functions (compared to cc-pVQZ calculation) versus -log ε in Ha.

Discussion of First Eight Examples

The above examples lead to three main observations. First, basis functions are concentrated on reactive atoms and those involved in important changes in electronic interactions. Second, the threshold ε is a reasonable estimate of the true basis truncation error, and reactions converge to within 1 kcal/mol error by $\varepsilon = 1 - 2$ mHa; ε is within a factor of two of the calculation error (Table S1). Third, basis function savings with $\varepsilon < 1$ mHa are reduced with minimal gains in accuracy compared to $\varepsilon = 1$ mHa (Table S2). In general, reactions with a greater fraction of atoms involved (whether through connectivity

or electronic interaction) have reduced savings compared to those with fewer electronically active atoms.

The results so far have examined thresholds for convergence with respect to ε. Given the nature of SBB as an algorithm based on *predicted* error (i.e. $|\Delta E - \Delta E^{ref}|$), it is also of interest to analyze how closely SBB's internal error metric matches the actual error (i.e. $|\Delta E - \Delta E^{exact}|$). To estimate this difference, maximum estimated error (MEE)—$\sum_i \left|\frac{\Delta E_c}{\Delta g_i}\right| \Delta g_i$—for the first eight reaction examples was calculated for each value of ε (Figure S2). Figure 11 shows the ratio of actual error to MEE, which is less than 1 for all reactions at all thresholds, indicating that MEE overestimates the true error.

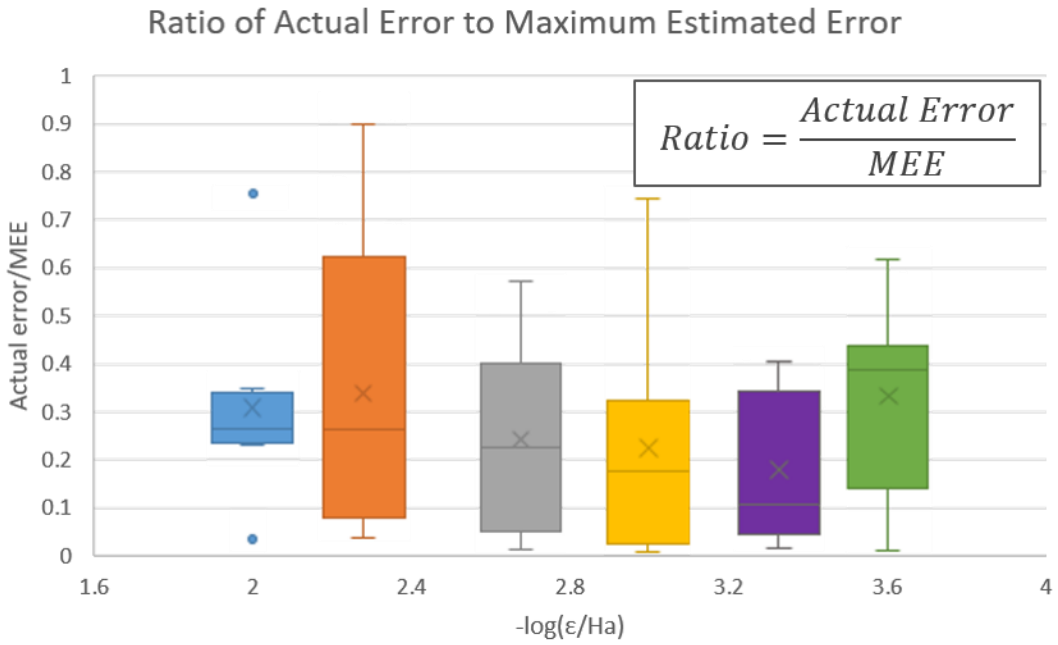

Figure 11. Ratio of actual error to maximum estimated error for first eight reaction examples plotted as whisker plots for each threshold value tested (0.25 – 10 mHa). The ratio is always less than one, showing that MEE > actual error for all reactions at all thresholds.

The combination of Figure S2 and Figure 11 suggests that the actual error, MEE, and ε are all similar in value. Fortunately, this means that SBB for ε < 2 mHa is expected to have at least kcal/mol accuracy, and ε can be taken as a reasonable estimate of true error in reaction energy. The above indications of basis set size required to achieve this convergence suggest that SBB is reaching target accuracies with significant reduction in number of atomic orbitals.

Transferability of basis sets to transition states

Having demonstrated convergence of the SBB algorithm for reaction intermediates, it is now of interest to examine whether these basis function distributions are transferable to different geometries. The transition states for the AB/CO_2 reactions,[79] shown in Figure 12, will serve as test cases. Transferability of SBB basis sets was tested by taking the $\varepsilon = 1$ mHa SBB bases and using these to

calculate the corresponding activation energies. Table 1 compares these results to the E_a calculated with cc-pVQZ on all atoms.

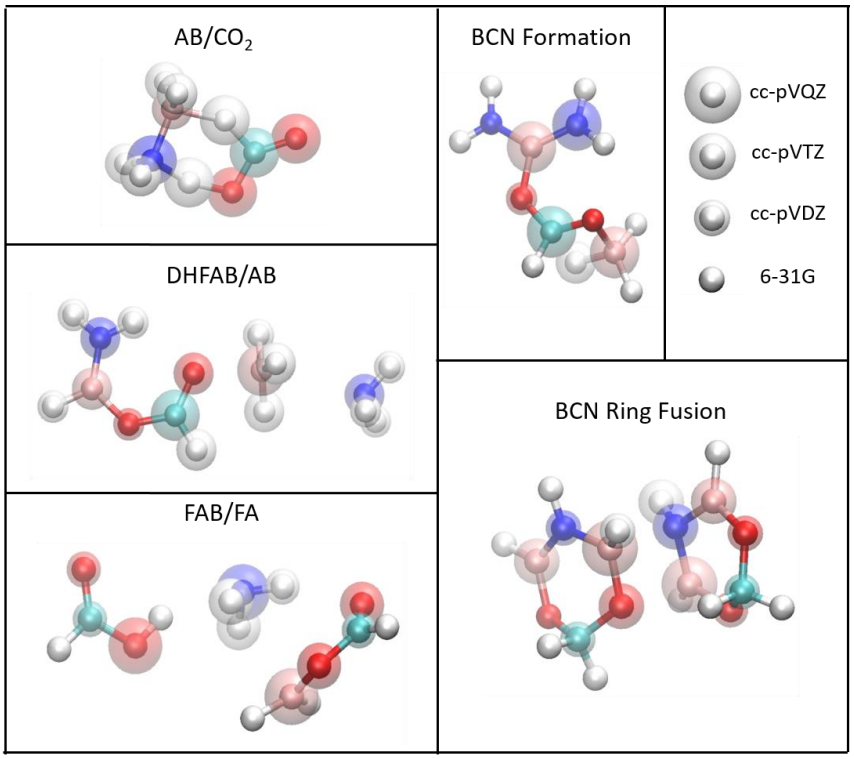

Figure 12. AB/CO_2 transition states. Basis function distributions shown here are the same as for the heats of reaction computation with ε = 1 mHa (Figure 8).

Table 1. Activation energies and error for AB/CO_2 transition states using SBB basis generated for reactant/product pair.

Reaction	E_a (ε = 1 mHa)	E_a (cc-pVQZ)	Error
AB/CO_2	28.59 kcal/mol	28.63 kcal/mol	0.04 kcal/mol
DHFAB/AB	23.81 kcal/mol	22.35 kcal/mol	1.46 kcal/mol
FAB/FA	36.11 kcal/mol	36.14 kcal/mol	0.03 kcal/mol
BCN formation	11.93 kcal/mol	9.95 kcal/mol	1.98 kcal/mol
BCN ring fusion	7.42 kcal/mol	8.57 kcal/mol	1.15 kcal/mol

Errors in E_a vary from nearly zero to almost 2 kcal/mol compared to full cc-pVQZ. These increased errors likely stem from two sources: 1. Transition states are more challenging than intermediate structures, and 2. Not all subtleties of electronic variation at the transition state are captured using SBB in the stable geometries. For example, B atoms in the DHFAB/AB and BCN formation reactions appear in non-optimal configurations at the transition state, but these geometries are relaxed at the endpoints of the reaction path. This degree of accuracy, however, is excellent given that the basis was not selected for the transition state, but for the reactant product pair.

Performing SBB for the DHFAB reactant/transition state pair adds additional functions to reactive N, both O, and four H, increasing the number of basis functions from 321 to 423 and reducing

the error in E_a to 0.08 kcal/mol. SBB performed for the BCN formation reactant/transition state pair increases the number of basis functions from 280 to 326 by adding additional functions to both O, and reduces error to 1.06 kcal/mol. SBB therefore works well even for transition states, with comparable accuracy to stable intermediate computations.

Large system sizes: Lever reactions

As molecules become larger, it becomes more and more infeasible to apply large basis sets uniformly across all atoms. SBB, however, may be able to overcome such difficulties and provide results similar in quality to full cc-pVQZ computations. Two reactions of mechanochemical interest[92-94] are examined to test SBB in this limit: *Lever 1* depicts gem-dichloro-cyclopropane mechanophore in a polybutadiene polymer, while *Lever 2* includes the same mechanophore in a polynorbornene chain (Figure 13).

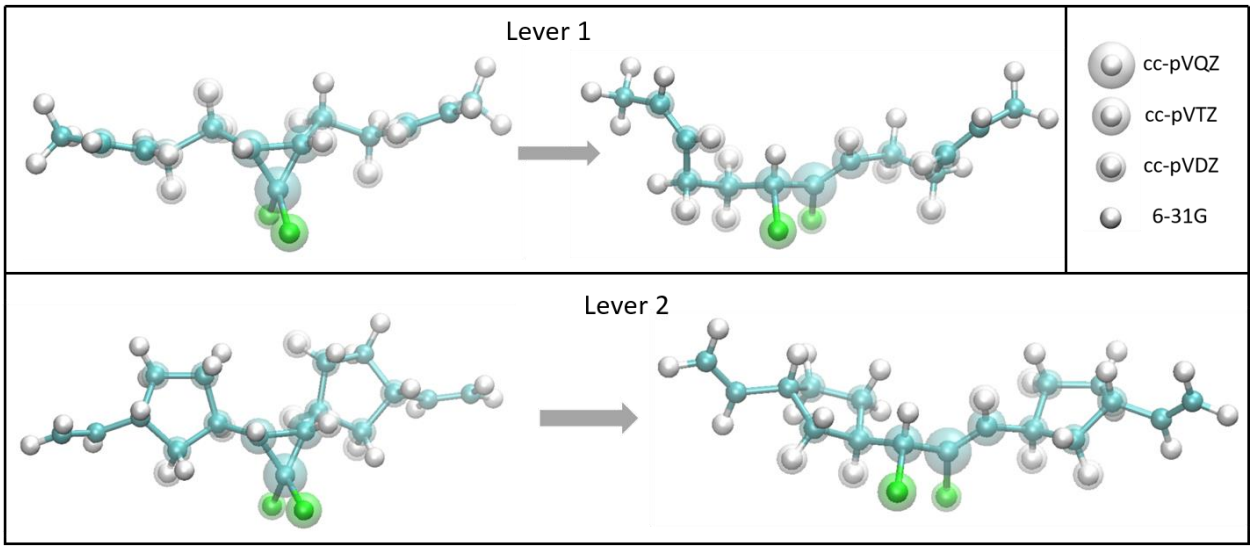

Figure 13. Lever reactions. Atom colors are as follows: teal – C, green – Cl, grey – H. Basis function distribution is shown at ε = 1 mHa.

In this example, the full cc-pVQZ basis would result in 1433 and 1773 basis functions total for *Levers 1* and *2*, respectively. Instead of calculating this limit, the $\varepsilon = 0.25$ mHa result is considered a benchmark, as its MEE is only 0.7 kcal/mol for both reactions (Figure S3). The SBB basis size for *Lever 1* is 61% smaller at $\varepsilon = 2$ mHa compared to $\varepsilon = 0.25$ mHa (Figure S4), and an analogous 44% smaller basis size is observed for *Lever 2*. As usual for SBB, cc-pVQZ is *not* assigned to all atoms, where basis functions are focused around the gem-dichloro-cyclopropane core, where the primary reactivity of the ring opening and chloride transfer occurs (Figure 13). Atoms farther from the reactive core are less affected by electrostatic field of the Cl, and thus are assigned fewer basis functions (Figure S5).

Discussion and Conclusions

A new basis set selection technique—SBB—was herein evaluated in detail for its accuracy and efficiency. Using a target accuracy of 1 kcal/mol error for RI-MP2 heats of reaction, the benchmark studies provide significant evidence that SBB can systematically reach this limit, and it does so with $\varepsilon = 1$ mHa. Even lower errors are achievable with tighter convergence, with $0.25 < \varepsilon < 0.5$ mHa

being a useful lower range. These results are independently evaluated using SBB's internal MEE metric, which shows that MEE is larger than the actual error, for all thresholds ε. This strongly suggests the error control strategy—built by design into SBB—is effective.

This error control can be achieved alongside reduction in basis set size by more than 50% at $\varepsilon = 1$ mHa. Savings are generally higher for reactions with proportionally fewer reactive atoms, while more polarizing reactions require higher quality basis sets (e.g. Figure 9). Tests also indicate that SBB basis sets can be transferred to geometries that are distinct from the structure pair where they were generated, at a small reduction in accuracy. This includes transition state geometries, which are harder to describe than stable intermediates.

Applications of SBB are likely in two areas. The first is in *ab initio* molecular dynamics simulations, where the reduced basis sets will allow longer trajectories at much reduced computational costs. Second, preliminary tests indicate that SBB basis sets chosen for one level of theory (e.g. RI-MP2) are transferable to a second level of theory (e.g. CCSD). Basis set selection at the lower level of theory therefore will allow meaningfully accurate electronic structure simulations at higher levels of theory. This use of SBB will be examined in combination with localized correlation methods (e.g. incremental full configuration interaction[22,95-96]) in the near future.

In summary, the herein introduced SBB algorithm is derived and tested thoroughly on a variety of molecular reactions. Error analysis, both internal and external, indicates the method consistently performs as designed. SBB will likely see continued development and be useful for reaching high accuracy in correlated computations.

Acknowledgments

The authors thank Dow Chemical for support of this project. David Braun is thanked for continued computational support.

References

(1) P. Y. Ayala, G. E. Scuseria, *J. Chem. Phys.* **1999**, *110* (3660).

(2) R. Sedlak, K. E. Riley, J. Řezáč, M. Pitoňák, P. Hobza, *ChemPhysChem* **2013**, *14* (4), 698–707.

(3) A. Karton, E. Rabinovich, J. M. L. Martin, B. Ruscic, *J. Chem. Phys.* **2006**, *125* (14).

(4) H. Fliegl, W. Klopper, C. Hättig, *J. Chem. Phys.* **2005**, *122*.

(5) F. R. Manby, *J. Chem. Phys.* **2003**, *119*, 4607–4613.

(6) R. J. Azar, M. Head-Gordon, *J. Chem. Phys.* **2015**, *142* (20).

(7) A. Venkatnathan, A. B. Szilva, D. Walter, R. J. Gdanitz, E. A. Carter, *J. Chem. Phys.* **2004**, *120* (4), 1693–1704.

(8) D. G. Fedorov, K. Kitaura, *J. Chem. Phys.* **2004**, *121* (6), 2483.

(9) M. Schütz, H. J. Werner, *J. Chem. Phys.* **2001**, *114* (2), 661–681.

(10) M. Schütz, *J. Chem. Phys.* **2000**, *113* (22), 9986–10001.

(11) G. Hetzer, M. Schütz, H. Stoll, H. J. Werner, *J. Chem. Phys.* **2000**, *113* (21), 9443.

(12) M. Schütz, G., Hetzer, H. J. Werner, *J. Chem. Phys.* **1999**, *111* (13), 5691–5705.

(13) A. Hansen, D. G. Liakos, F. Neese, *J. Chem. Phys.* **2011**, *135* (21).

(14) J. E. Subotnik, M. Head-Gordon, *J. Phys. Condens. Matter* **2008**, *20* (29), 294211.

(15) J. E. Subotnik, A. Sodt, M. Head-Gordon, *J. Chem. Phys.* **2006**, *125* (7).

(16) T. S. Chwee, A. B. Szilva, R. Lindh, E. A. Carter, *J. Chem. Phys.* **2008**, *128* (22).

(17) M. Schütz, F. R. Manby, *Phys. Chem. Chem. Phys.* **2003**, *5* (16), 3349–3358.

(18) M. Schwilk, D. Usvyat, H. J. Werner, *J. Chem. Phys.* **2015**, *142* (12).

(19) H. J. Werner, F. R. Manby, P. J. Knowles, *J. Chem. Phys.* **2003**, *118* (18), 8149–8160.

(20) H. Larsen, J. Olsen, P. Jørgensen, *J. Chem. Phys.* **2001**, *115* (21), 9685–9697.

(21) M. Kobayashi, Y. Imamura, H. Nakai, *J. Chem. Phys.* **2007**, *127* (7).

(22) P. M. Zimmerman, *J. Chem. Phys.* **2017**, *146* (10).

(23) E. F. Valeev, *Chem. Phys. Lett.* **2004**, *395* (4–6), 190–195.

(24) W. Klopper, F. R. Manby, S. Ten-No, E. F. Valeev, *Int. Rev. Phys. Chem.* **2006**, *25* (3), 427–468.

(25) E. F. Valeev, C. L. Janssen, *J. Chem. Phys.* **2004**, *121* (3), 1214–1227.

(26) L. Kong, F. A. Bischoff, E. F. Valeev, *Chem. Rev.* **2012**, *112* (1), 75–107.

(27) E. F. Valeev, T. Daniel Crawford, *J. Chem. Phys.* **2008**, *128* (24).

(28) D. Moncrieff and S. Wilson *J. Phys. Chem. B* **1996**, *29*, 6009

(29) T. Helgaker, W. Klopper, H. Koch, J. Noga, *J. Chem. Phys.* **1997**, *106* (23), 9639–9646.

(30) D. G. Truhlar, *Chem. Phys. Lett.* **1998**, *294*, 45–48.

(31) W. Klopper, *Mol. Phys.* **2001**, *99*, 481–507.

(32) E. Fabiano, F. della Sala, *Theor. Chem. Acc.* **2012**, *131*, 1–10.

(33) M. Okoshi, T. Atsumi, H. Nakai, *J. Comput. Chem.* **2015**, *36*, 1075–1082.

(34) W. Klopper, K. L. Bak, P. Jørgensen, J. Olsen, *J. Phys. Chem. B* **1999**, *32*, R103-R130.

(35) M. Jeziorska, W. Cencek, K. Patkowski, B. Jeziorski, K. Szalewicz, *Int. J. Quantum Chem.* **2008**, *108*, 2053–2075.

(36) P. Jurečka, J. Šponer, J. Černý, P. Hobza, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(37) J. W. Ochterski, G. A. Petersson, J. A. Montgomery, *J. Chem. Phys.* **1996**, *104*, 2598–2619.

(38) K. Szalewicz, B. Jeziorski, *J. Chem. Phys.* **1998**, *109*, 1198–1200.

(39) G. A. Petersson, M. A. Al-Laham, *J. Chem. Phys.* **1991**, *94*, 6081–6090.

(40) W. Kutzelnigg, J. D. Morgan, *J. Chem. Phys.* **1992**, *96*, 4484–4508.

(41) D. Feller, K. A. Peterson, J. Grant Hill, *J. Chem. Phys.* **2011**, *135*.

(42) D. Feller, *J. Chem. Phys.* **2013**, *138*.

(43) E. Papajak, D. G., Truhlar, *J. Chem. Phys.* **2012**, *137*, 0–8.

(44) K. A. Peterson, T. B. Adler, H. J. Werner, *J. Chem. Phys.* **2008**, *128*.

(45) W. Klopper, C. C. M. Samson, *J. Chem. Phys.* **2002**, *116*, 6397–6410.

(46) H. J. Werner, T. B. Adler, F. R. Manby, *J. Chem. Phys.* **2007**, *126*.

(47) F. R. Manby, H. J. Werner, T. B. Adler, A. J. May, *J. Chem. Phys.* **2006**, *124*.

(48) P. L. Fast, J. C. Corchado, M. L. Sánchez, D. G. Truhlar, *J. Phys. Chem. A* **1999**, *103*, 5129–5136.

(49) B. J. Lynch, Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 1643–1649.

(50) B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2003**, *107*, 3898–3906.

(51) S. Boys, F. Bernardi, *Mol. Phys.* **1970**, *19*, 553–566.

(52) L. Turi, J. J. Dannenberg, *J. Phys. Chem.* **1993**, *97*, 2488–2490.

(53) P. Valiron, I. Mayer, *Chem. Phys. Lett.* **1997**, *275*, 46–55.

(54) S. Simon, M. Duran, J. J. Dannenberg, *J. Chem. Phys.* **1996**, *105*, 11024–11031.

(55) L. A. Burns, M. S. Marshall, C. D. Sherrill, *J. Chem. Theory Comput.* **2014**, *10*, 49–57.

(56) R. M. Richard, K. U. Lao, J. M. Herbert, *J. Phys. Chem. Lett.* **2013**, *4*, 2674–2680.

(57) Y. Li, J. Yuan, M. Chen, F. Ma, M. Sun, *J. Comput. Chem.* **2013**, *34*, 1686–1696.

(58) P. Jurecka, J. Sponer, J. Cerný, P. Hobza, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(59) D. Feller, *J. Chem. Phys.* **2013**, *138*.

(60) P. Hobza, J. Šponer, *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

(61) D. G. Liakos, R. Izsák, E. F. Valeev, F. Neese, *Mol. Phys.* **2013**, *111*, 2653–2662.

(62) P. Jurečka, P. Hobza, *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

(63) R. M. Richard, M. S. Marshall, O. Dolgounitcheva, J. V. Ortiz, J. L. Bredas, N. Marom, *J. Chem. Theory Comput.* **2016**, *12*, 595–604.

(64) A. Karton, A. Tarnopolsky, J.-F. Lamère, G. C. Schatz, J. M. L. Martin, *J. Phys. Chem. A* **2008**, *112*, 12868.

(65) R. P. Steele, R. A. DiStasio, Y. Shao, J. Kong, M. Head-Gordon, *J. Chem. Phys.* **2006**, *125* (7).

(66) W. Liang, M. Head-Gordon, *J. Phys. Chem. A* **2004**, *108*, 3206.

(67) B. Chan, L. Radom, *J. Chem. Theory Comput.* **2011**, *7* (9), 2852–2863.

(68) K. B. Wiberg, *J. Comput. Chem.* **2004**, *25* (11), 1342–1346.

(69) A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, J. Olsen, *Chem. Phys. Lett.* **1999**, *302*, 437–446.

(70) A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, A. K. Wilson, *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(71) J. Elith, J. R. Leathwick, T. J. Hastie, *J. Anim. Ecol.* **2008**, *77* (4), 802–813.

(72) B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *Ann. Stat.* **2004**, *32* (2), 407–499.

(73) T. Hastie, J. Taylor, R. Tibshirani, G. Walther, *Electron. J. Stat.* **2007**, *1*, 1–29.

(74) A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648-5652

(75) C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785-789

(76) R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.* **1980**, *72*, 650-654.

(77) Y. Shao, L. F. Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S. T. Brown, A. T. B. Gilbert, L. V. Slipchenko, S. V. Levchenko, D. P. O'Neill, R. A. DiStasio Jr, R. C. Lochan, T. Wang, G. J. O. Beran, N. A. Besley, J. M. Herbert, C. Yeh Lin, T. Van Voorhis, S. Hung Chien, A. Sodt, R. P. Steele, V. A. Rassolov, P. E. Maslen, P. P. Korambath, R. D. Adamson, B. Austin, J. Baker, E. F. C. Byrd, H. Dachsel, R. J. Doerksen, A. Dreuw, B. D. Dunietz, A. D. Dutoi, T. R. Furlani, S. R. Gwaltney, A. Heyden, S. Hirata, C.-P. Hsu, G. Kedziora, R. Z. Khalliulin, P. Klunzinger, A. M. Lee, M. S. Lee, W. Liang, I. Lotan, N. Nair, B. Peters, E. I. Proynov, P. A. Pieniazek, Y. Min Rhee, J. Ritchie, E. Rosta, C. David Sherrill, A. C. Simmonett, J. E. Subotnik, H. Lee Woodcock III, W. Zhang, A. T. Bell, A. K. Chakraborty, D. M. Chipman, F. J. Keil, A. Warshel, W. J. Hehre, H. F. Schaefer III, J. Kong, A. I. Krylov, P. M. W. Gill, M. Head-Gordon, *Phys. Chem. Chem. Phys.*, **2006**, *8*, 3172-3191.

(78) T. H. Dunning Jr., *J. Chem. Phys.* **1989**, *90*, 1007-1023.

(79) M. W. Li, I. M. Pendleton, A. J. Nett, P. M. Zimmerman, *J. Phys. Chem. A* **2016**, *120* (8), 1135–1144.

(80) P. M. Zimmerman, Z. Zhang, C. B. Musgrave, *J. Phys. Chem. Lett.* **2011**, *2*, 276–281.

(81) A. Al-Kukhun, H. T. Hwang, A. Varma, *Int. J. Hydrogen Energy* **2013**, *38*, 169–179.

(82) C. Lim, A. Holder, J. T. Hynes, C. B. Musgrave, *Inorg. Chem.* **2013**, 1–5.

(83) P. M. Zimmerman, Z. Zhang, C. B. Musgrave, *Inorg. Chem.* **2010**, *49*, 8724-8728

(84) L. Roy, P. M. Zimmerman, A. Paul, *Chem. Eur. J.* **2011**, *17*, 435–439.

(85) D. Ai, Y. Guo, W. Liu, Y. Wang, *J. Phys. Org. Chem.* **2014**, *27*, 597–603.

(86) A. Pal, K. Vanka, *Chem. Commun. (Camb).* **2011**, *47*, 11417–11419.

(87) T. Malakar, L. Roy, A. Paul, *Chem. Eur. J.* **2013**, *19*, 5812-5817

(88) H. A. Kalviri, F. Gärtner, G. Ye, I. Korobkov, R. T. Baker, *Chem. Sci.* **2014**, *6*, 618–624.

(89) T. Malakar, S. Bhunya, A. Paul, *Chem. Eur. J.* **2015**, *21*, 6340–6345.

(90) S. Bhunya, P. M. Zimmerman, A. Paul, *ACS Catal.* **2015**, *5*, 3478–3493.

(91) P. M. Zimmerman, A. Paul, Z. Zhang, C. B. Musgrave, *Inorg. Chem.* **2009**, *48,* 1069-1081

(92) H. M. Klukovich, T. B. Kouznetsova, Z. S. Kean, J. M. Lenhardt, S. L. Craig, *Nat. Chem.* **2013**, *5* (2), 110–114.

(93) J. Wang, T. B. Kouznetsova, Z. S. Kean, L. Fan, B. D. Mar, T. J. Martínez, S. L. Craig, *J. Am. Chem. Soc.* **2014**, *136* (43), 15162–15165.

(94) C. L. Brown, S. L. Craig, *Chem. Sci.* **2015**, *6*, 2158–2165.

(95) P. M. Zimmerman, *J. Chem. Phys.* **2017**, *146* (22).

(96) P. M. Zimmerman, *J. Phys. Chem. A* **2017**, *121* (24), 4712–4720.