The effect of midbond functions on interaction energies computed using MP2 and CCSD(T)

Regina Matveeva1 | Merete Falck Erichsen1 | Henrik Koch1,2 | Ida-Marie Høyvik1

Abstract

In this article we use MP2 and CCSD(T) calculations for the A24 and S66 data sets to explore how midbond functions can be used to generate cost effective counterpoise corrected supramolecular interaction energies of noncovalent complexes. We use the A24 data set to show that the primary role of midbond functions is not to approach the complete basis set limit, but rather to ensure a balanced description of the molecules and the interaction region (unrelated to the basis set superposition error). The need for balance is a consequence of using atom centered basis sets. In the complete basis set limit, the error will disappear, but reaching the complete basis set limit is not feasible beyond small systems. For S66 we investigate the need for increasing the number of midbond centers. Results show that adding a second midbond center increases the accuracy, but the effect is secondary to changing the atom centered basis set. Further, by comparing calculations using the 3s3p2d1f1g midbond set with using aug-cc-pVDZ and aug-cc-pVTZ as midbond sets, we see that the requirements for the midbond set to be effective, is not just that it contains diffuse functions, but also that high angular momentum functions are included. By comparing two approaches for placing midbond centers we show that results are not particularly sensitive to placement as long as the placement is reasonable.

KEYWORDS

A24, interaction energies, midbond, noncovalent, S66

1 INTRODUCTION

The role of noncovalent interactions in various areas of natural sciences cannot be underestimated. They play, for example, an important role in the design and construction of supramolecular self-assemblies and materials1,2 as well as in biochemical research.3-5 The computational description of noncovalent interactions has therefore been a topic of research for several decades.6-13 Compared to covalent bonds, noncovalent interactions are characterized by larger intermolecular distances, smaller changes in the electronic structure of the interacting species and energy differences which are orders of magnitude lower.10 Thus, quantifying noncovalent interaction energies through computations has stricter requirements on the level of theory to obtain meaningful results. For quantitative results of interaction energies, correlated \textit{ab initio} electronic structure wave function models in combination with high-quality atomic orbital (AO) basis sets are necessary.14 To this end, the coupled cluster hierarchy of wave function models has been shown to be effective.15 Indeed, the coupled cluster model with singles, doubles and perturbative triple excitations\textsuperscript{16} (CCSD(T)) with a complete basis set (CBS) description...
(also known as “gold standard” of electronic structure theory)\textsuperscript{17} provides quantitative accuracy.\textsuperscript{15,18–20} This method has two apparent challenges, namely the $N^2$ scaling of the CCSD(T) model (with $N$ being a number of correlated electrons/measure of molecular size) and a very slow convergence of the electron correlation energy to the CBS limit. The scaling issue can be avoided by local approximations,\textsuperscript{21–23} while with respect to the latter point, extrapolation techniques\textsuperscript{24–29} towards the CBS limit have proved effective and are therefore widely used.\textsuperscript{30–32} Alternatively, the convergence towards the CBS limit can be improved by including terms into the wave function, which explicitly depend on the intermolecular distance.\textsuperscript{33,34} These so-called explicitly correlated methods are characterized by a much faster convergence to the CBS limit.\textsuperscript{35,36}

A common approach for computing interaction energies is the supramolecular approach, in which the interaction energy of two weakly bound systems is defined as the difference of the energies of the dimer and then energies of the monomers. In addition to scaling and incomplete basis set challenges, the supramolecular approach (in combination with a finite basis set) introduces errors such as the basis set superposition error (BSSE). The BSSE is an artifact of the supermolecular approach, and it appears since the monomers of the complex calculation will have an improved description over the separated monomers.\textsuperscript{37} The BSSE results in an artificially strong computed interaction energy of a complex,\textsuperscript{38} and for noncovalent interactions the BSSE may be on the same order of magnitude as the interaction itself, thus a correction of the error is necessary. The counterpoise (CP) correction method of Boys and Bernardi\textsuperscript{39} has been used in numerous applications and has proven very efficient. At the same time, the CP approach tends sometimes to overcorrect for BSSE.\textsuperscript{40–43} Therefore, Sherill and co-workers\textsuperscript{44} recommended to apply the so-called “half-counterpoise” corrections (average of raw and CP-corrected quantities) for \textit{ab initio} calculations on noncovalent complexes using the basis set of aug-cc-pVQZ and below. This is also valid at explicitly correlated levels of theory, MP2-F12 and CCSD(T)-F12, in combination with intermediate basis sets (e.g., cc-pVQZ-F12), while with small basis sets like cc-pVQZ-F12 incorrect results are closer to CBS limit due to beneficial error compensation between BSSE (over-binding) and intrinsic basis set incompleteness (under-binding).\textsuperscript{45} Other alternatives to the classical CP correction method are known—the so-called virtual CP approach,\textsuperscript{37,46} atom by atom scheme (CPa)\textsuperscript{47} and Same Number Of Optimized Parameters scheme\textsuperscript{48,49} to name few.

There is a second effect on the quality of the computed interaction energies which is much less discussed than the BSSE. In electronic structure theory, we commonly choose atom centered basis sets. Since intermolecular distances are large for noncovalent interactions (relative to distances in covalent bonds) the atom centered basis set functions have limited flexibility in the interaction energy. And hence, the nucleus centered basis set favor the description of the molecule region over the interaction region, thus resulting in an imbalanced description of the noncovalent system. This effect will thus also be present in BSSE corrected calculations since the monomer and dimer energies are still just computed using the atom centered basis set.

This effect is seldom discussed in its own right, and the associated error is attributed to the fact that CBS is not reached. However, the CBS errors will be much larger for the interaction region than for the molecules due to basis sets being centered on atoms. The errors associated with not reaching CBS are commonly addressed by using large atom centered basis sets with diffuse functions, for example, aug-cc-pVXZ basis sets with cardinal number $X = 5$ or higher.\textsuperscript{19} However, this rapidly ends in unfeasible calculations even for small molecular systems and moreover, this is an inefficient way to progress since the approach requires one to go close to CBS limit both in molecule regions and the interaction region, using atom centered basis functions. It is also worth noting that explicitly correlated models alone cannot remove the error introduced by the imbalanced description of molecules and interaction region.

As an alternative, Tao and Pan\textsuperscript{50} introduced the use of midbond functions, a set of functions centered in the interaction region to supplement atom centered basis sets. The 3s3p2d midbond functions set with exponents 0.9, 0.3, 0.1 for s- and p-functions and exponents 0.6 and 0.2 for d-functions was one of the first used midbond sets. This set was subsequently implemented by several research groups in the 1990’s for calculations of interaction energies and potential energy surfaces of rare gas dimers and rare gas-molecule complexes.\textsuperscript{51–59} Extension of this midbond set by one f-type function and further by one g-type function resulted in 3s3p2d1f and 3s3p2d1f1g sets (the added f- and g-functions have both the exponent 0.3), respectively. These sets have been applied in calculations of potential energy surfaces\textsuperscript{60–66} and also dissociation energies.\textsuperscript{67} Some examples of large sets of midbond functions, for example, 6s6p6d3f3g3h are also known.\textsuperscript{68} A prominent example of the computational effectiveness of midbond functions for computing noncovalent interaction energies is the recent extensive studies by the Patkowski group,\textsuperscript{69} demonstrating that at CCSD(T) and CCSD(T)-F12 levels of theory aug-cc-pVDZ basis set supplemented with midbond functions can provide results of high accuracy.\textsuperscript{17,69}

Midbond functions are generally used on ghost atoms located in the intermolecular region, and displacement along the van der Waals bond has been shown to have only a negligible impact on the accuracy of interaction energies of rare gas complexes.\textsuperscript{51,52} The optimal position of midbond functions in small molecular dimers was also investigated in a more recent study by Shaw and Hill\textsuperscript{15} where they optimized the position of the midbond functions. However, since moving the midbond center closer to the larger/heaviest monomer is energetically favorable, such an optimization procedure only maps the effect of placement along the bond. The study by Shaw and Hill\textsuperscript{15} has also challenged the prevalent assumption that the interaction energy is insensitive to the exponents of the midbond functions.\textsuperscript{70} It was demonstrated that significant improvement in both canonical and explicitly correlated calculations of interaction energies can be achieved by optimization of midbond exponents.\textsuperscript{18}

To this date, systematic studies of effects of midbond functions in calculations of interaction energies for medium sized molecular systems (30 atoms) using midbond functions have been performed on the S22 data set.\textsuperscript{11,69} Despite the usefulness and popularity of
the S22 data set, there are some issues associated with it. For example, the S22 data set mainly targets interactions of nucleic acid bases while other interaction categories are under-represented (e.g., single hydrogen bonds) or even missing (aliphatic–aliphatic dispersion interactions). The S66 data set was designed to cover a wider range of interaction motifs. For the S66 data set, Ma and Werner have included midbond functions in local correlation calculations of interaction energies. However, they present no systematic studies on how midbond centers specifically affect the results for the S66 data set.

In this paper we use the A24 and S66 data sets to illustrate how midbond functions affect the computed interaction energies at MP2 and CCSD(T) levels of theory. We show that using midbond functions of the type 3s3p2d1f1g alleviates the requirements on the chosen nucleus centered basis set, and that including a midbond center is highly efficient for small molecular systems, as displayed using the A24 set of molecules. Further, the A24 data set is used to illustrate that midbond functions provide a balanced description of molecules and interaction regions. For the S66 data set we explore the inclusion of two, rather than just one, midbond centers and we show that the accuracy increases upon adding the second midbond center, which, however, is associated with significantly increased computational costs. We also compare computations of interaction energies using the 3s3p2d1f1g midbond set with the use of correlation consistent aug-cc-pVDZ and aug-cc-pVTZ as midbond sets and we see that the midbond set needs to contain functions of high angular momentum (at least f-functions) in order to be effective. Further, by examining two approaches for placing midbond centers we show that results are not particularly sensitive to exact placement as long as it is reasonable.

The paper is organized as follows. Section 2 contains an exposition of the methodology. In Section 3 we use the A24 data set to illustrate the primary objective for using midbond functions. In Section 4 we present results for the S66 data set. In Section 5 we provide a summary and concluding remarks.

2 | METHODOLOGY

2.1 | Data sets

For the presented study we use the A24 and S66 data sets. Whereas the A24 data set consists of small molecular systems, the S66 data set contains larger and more biologically relevant complexes. Both data sets include noncovalent interactions of various kinds, covering hydrogen bonds (referred to as hydrogen group), dispersion (dispersion group) and a mixture of interaction types (others group). The complex geometries are used without further optimization and they can be found in the BEGDB online database.

2.2 | Interaction energy

All interaction energies are calculated using the supramolecular approach, in which the interaction energy $E_{\text{int}}$ of two weakly bound atoms or molecules is defined as the difference of the energies of the dimer $AB$, $E(AB)$, and the energies of the monomers $A$, $E(A)$, and $B$, $E(B)$. CP correction is used to correct for the BSSE, that is, the monomer energies are calculated in the basis of the dimer. The interaction energy is therefore given by,

$$E_{\text{int}}^{CP}(AB) = E(AB) - E(A^{AB}) - E(B^{AB}).$$

The superscript “AB” on monomers indicates that the basis for the dimer is used in the monomer calculations. All results presented in this paper are computed using the CP correction, and for simplicity, the “CP” superscript is made implicit.

2.3 | Midbond sets

In this work, we will investigate the performance of the midbond set $3s3p2d1f1g$ and the correlation consistent aug-cc-pVXZ basis functions ($X = D, T, Q$) for carbon as midbond sets. The exponents of the former set of midbond functions are 0.9, 0.3, 0.1 for $s$-functions and $p$-functions, 0.6, 0.2 for $d$-functions and 0.3 for the $f$- and $g$-functions. For carbon atom the contracted composition of correlation consistent basis functions is: \([3s,2p,1d]^{67}\) for aug-cc-pVDZ, \([4s,3p,2d,1f]^{[68]}\) for aug-cc-pVTZ and \([5s,4p,3d,2f,1g]^{[68]}\) for aug-cc-pVQZ. We will use a shorthand notation to indicate which type of midbond functions we use, that is, \((+33211)\) for $3s3p2d1f1g$ and \((+aXZ)\) for aug-cc-pVXZ, respectively.

The standard procedure of adding midbond functions is to place them on the midpoint between the centers of mass of the interacting monomers. Patkowski and co-workers stress that this method often leads to a midpoint being located closer to one of monomers and thus favoring its description. In order to avoid this, they refer to an algorithm of locating a midbond center, where its location is a $r^{-6}$ weighted average of intermolecular atom–atom midpoints,

$$\overline{r}_b = \frac{\sum_{a \in A, b \in B} o_{ab} \frac{r_{ab}}{r_{a}^6}}{\sum_{a \in A, b \in B} o_{ab}} = \left| \overline{r}_a - \overline{r}_b \right|^{-6},$$

where the summation runs over all atoms $a$ in subsystem $A$ and all atoms $b$ in subsystem $B$. Except where stated explicitly otherwise, we will use Equation (2) to place one midbond center (referred to as systematic). In Section 4.3 we explore the effect of placement of midbond functions, and we include results for where simple chemical intuition is used to place midbond centers (referred to as intuitive, see Figure S1). A detailed description of this approach can be found in Section S3.5 in the Supporting Information. In sections where we use two midbond centers for the S66 data set, we only use the 53 out of 66 complexes which have wide enough interaction regions. The two centers are placed manually. For system names of the 53 complexes with two midbond centers see Table S25. Geometries of the complexes (including midbond centers) of the A24 and S66 data sets are available at https://doi.org/10.18710/2FWECY.
2.4 | Statistical measures

The errors of the interaction energies are calculated in absolute terms

$$\Delta_n = E_{\text{int}}^{\text{ref}}(n) - E_{\text{int}}^{\text{calc}}(n)$$

and in relative (%) terms,

$$\Delta_{n}^{\text{rel}} = \frac{\Delta_n}{E_{\text{int}}^{\text{ref}}(n)} \times 100\%,$$

where $E_{\text{int}}^{\text{ref}}(n)$ is the reference value for the interaction energy of complex $n$, and $E_{\text{int}}^{\text{calc}}(n)$ is the computed value of the interaction energy for complex $n$. Since the magnitude of interaction energies varies significantly with interaction type, using errors in relative terms (%) makes it easier to compare the quality of results for different interaction types.

We will present errors in terms of the mean absolute error, $\overline{\Delta}_{\text{abs}}$, and the relative mean absolute error, $\overline{\Delta}_{\text{rel}}$:

$$\overline{\Delta}_{\text{abs}} = \frac{1}{N} \sum_{n=1}^{N} |\Delta_n|,$$

$$\overline{\Delta}_{\text{rel}} = \frac{1}{N} \sum_{n=1}^{N} \frac{|\Delta_n|}{E_{\text{int}}^{\text{ref}}(n)}.$$

where $N$ is the total number of noncovalent complexes.

2.5 | Computational details

The interaction energies for the A24 (see Section 3) and S66 (see Section 4) data sets are computed at MP2 and CCSD(T) levels of theory using correlation consistent basis sets (aug)-cc-pVXZ with X = D, T, Q for A24 and X = D, T, Q for S66, within the frozen-core approximation and employing CP correction for BSSE. All calculations presented in this work are carried out using the LSDALTON program. The Hartree–Fock convergence threshold was tightened from default to $10^{-7}$ and the integral threshold was tightened from default $10^{-8}$ to $10^{-9}$ for all calculations.

For the timing calculations presented in Sections 4.1.1 and 4.2.1 we use the LSDALTON2020.0 wall times from utilizing a single cluster node with two Intel Xeon E5-2630 v2 CPUs summing up to 20 cores. The timing computations are performed on resources provided by the NTNU IDUN/EPIC computing cluster. All timings include the dimer calculation as well as the CP-corrected calculations for monomers.

The CCSD(T) reference interaction energies for the A24 data set were obtained from aug-cc-pVQZ (+33211) calculations for cc-pVTZ basis set. For CCSD(T) calculations in the S66 set, we used the revised benchmark CCSD(T)/CBS interaction energies presented in the publication of Hobza et al. as reference values. To obtain reference values for both data sets we compute interaction energies at MP2/aug-cc-pVQZ level of theory adding one 3s3p2d1f1g midbond center according to Equation (2) (see Tables S1 and S16).

3 | THE OBJECTIVE OF MIDBOND FUNCTIONS

In this section we use the A24 set to illustrate why the inclusion of a midbond center significantly improves the interaction energy when using small basis sets, and which is unrelated to reaching the CBS limit. We do so by examining computed monomer and dimer energy contributions compared to reference monomer and dimer energies (all within the CP correction). We present CCSD(T) results for the A24 data set using the basis set cc-pVTZ, X = D, T, Q with and without the midbond sets 3s3p2d1f1g (denoted +33211) and carbon aug-cc-pVXZ, X = D, T, Q (denoted +aXZ). In Table 1 we present the energy differences for monomers ($M_A$, $M_B$, and dimers ($D$) for each basis set combination (for monomer and dimer energies see Tables S13–S15) with respect to reference values computed using aug-cc-pVQZ (+33211). $\overline{\Delta}_{\text{abs}}$ and $\overline{\Delta}_{\text{rel}}$ for the interaction energies are presented in Table 2 (interaction energies itself can be found in Tables S7–S9). Note that errors in energy contributions in Table 1 are given in mE, while errors in interaction energies in Table 2 are given in kcal/mol.

We first consider the energy differences for monomers and dimers separately, presented in Table 1. We see that for cc-pVTZ in combination without and with both types of midbond sets (3s3p2d1f1g and aug-cc-pVXZ), $\overline{\Delta}_{\text{abs}}$ in monomer and dimer energies are, as expected, decreasing as the basis set is increased from cc-pVDZ through cc-pVTZ to cc-pVQZ. Further, we note that the errors in $M_A$, $M_B$, and $D$ are similar for results with and without midbond functions but with somewhat larger differences for cc-pVDZ results. However, the errors with and without midbond functions are still on the same order of magnitude even for cc-pVDZ results. In contrast, if we look at $\overline{\Delta}_{\text{abs}}$ and $\overline{\Delta}_{\text{rel}}$ results in Table 2, we see a significant difference between errors for the interaction energies computed with and without midbond sets. For example, for the cc-pVDZ calculations, $\overline{\Delta}_{\text{abs}}$, for cc-pVDZ (+33211), cc-pVDZ (+aDZ), and cc-pVDZ are 11.8%, 23.9% and 68.7%. Increasing the basis set to cc-pVQZ reduces $\overline{\Delta}_{\text{abs}}$ to 6.2%, 5.9%, and 31.5% for cc-pVQZ (+33211), cc-pVQZ (+aDZ), and cc-pVQZ, respectively. Hence, we see that although the errors for $M_A$, $M_B$, and $D$ energies for the cc-pVQZ without a midbond set are much smaller than the errors for cc-pVDZ with midbond sets, the error in the interaction energy itself for the cc-pVQZ calculations without a midbond set ($\overline{\Delta}_{\text{abs}} = 31.5\%$) is larger than the interaction energy error, for example, for cc-pVQZ (+33211) $\overline{\Delta}_{\text{abs}} = 11.8\%$. Thus, we see that more accurate results for $M_A$, $M_B$, and $D$ do not necessarily correlate with more accurate interaction energies. The primary role of a midbond set is therefore not to provide more accurate monomer and dimer energies by more rapidly reaching CBS limit, but rather to ensure that a balanced description of monomers and dimer is obtained. That is, a midbond set corrects imbalances in description of...
4 | RESULTS FOR S66 DATA SET

In this section we present results for the S66 data set. In Section 4.1 we discuss the effect on interaction energies using one and two 3s3p2d1f1g midbond centers. In Section 4.2 we compare the results of using two 3s3p2d1f1g centers with using two carbon-type aug-cc-pVDZ midbond centers. We further comment on relative timings for various combinations of basis sets and midbond functions. In Section 4.3 we discuss the effect of using a systematic method for placing midbond function centers (see Section 2.3) versus using an intuitive (manual) method for placing midbond centers (see Section S3.5 in the Supporting Information).

4.1 | One versus two 3s3p2d1f1g midbond centers

The S66 data set contains complexes with wide interaction regions, and it therefore is reasonable to investigate whether an increased number of midbond centers will result in improved interaction energies. Further, we investigate whether the increased number of midbond centers can compensate for the deficiencies of using a small atom centered basis set (e.g., cc-pVDZ), as was the case for the A24 data set (see Section 3). Note that two midbond centers are only used for 53 out of 66 complexes, where the interaction region is reasonably large. Accordingly, we compare results obtained using two midbond centers (see Tables S17 and S21) with results obtained using one midbond center (see Tables S19 and S23) only for the 53 complexes. Results for Δabs and Δrel are shown in Table 3 and visualized in Figure 1.

For the CCSD(T) calculations we used the CCSD(T)/CBS interaction energies from Hobza et al. as reference values. MP2 reference values are obtained at MP2/aug-cc-pVQZ level of theory and the MP2 reference results can be found in Table S16. Note that the reference values for MP2 are MP2 calculations and the references for the CCSD(T) calculations are CCSD(T) (see details in Table 3). Hence, the errors reflect the basis set error of the method, rather than a combined method and basis set error. Smaller MP2 than CCSD(T) errors therefore do not indicate that MP2 interaction energies are better than corresponding CCSD(T) results.

We first discuss the cc-pVDZ results. The cc-pVDZ(+33211) and cc-pVQZ(+2*33211) results for MP2 show that errors are reduced when including two midbond centers rather than one. The effect is particularly pronounced for dispersion group, where Δabs is reduced from 23.8% to 14.3% upon increasing from one to two midbond centers. The same is seen for CCSD(T) where the Δabs is reduced from 29.1% to 16.3% for dispersion group, but to a less extent for the other two groups. However, the errors for MP2 and CCSD(T) results using cc-pVDZ(+33211) and cc-pVQZ(+2*33211) are quite large. Errors for small complexes when using cc-pVDZ in combination with midbond functions are significantly smaller, because the midbond sets are able to both give an improved description of the molecules and the interaction region. For larger complexes, the midbond functions are not able to compensate for the poor description provided by the cc-pVQZ basis for the molecules.
The quality of results obtained when using one midbond center improves upon increasing the atom centered basis set from cc-pVDZ to cc-pVTZ. The most substantial improvement is observed for the dispersion group. That is, for MP2 $\Delta_{rel}^{abs}$ is reduced from 23.8% (cc-pVDZ) to 8.6% (cc-pVTZ), for CCSD(T) $\Delta_{rel}^{abs}$ is reduced from 29.1% to 12.7%. The inclusion of the second midbond center into the interaction region leads to a further improvement of the results, for example, for MP2 $\Delta_{rel}^{abs}$ is reduced from 8.6% (cc-pVTZ) to 5.5% (cc-pVTZ+2*33211) for dispersion group. CCSD(T)/cc-pVTZ+2*33211 results follow the same trend. However, the positive effect of adding the second midbond center is less significant for cc-pVTZ than for cc-pVDZ.

We now discuss results for the aug-cc-pVDZ basis set in combination with one or two midbond centers. For the MP2 method using aug-cc-pVDZ+33211 we obtain $\Delta_{abs}^{rel}$ of 4.3%, 6.3%, and 5.1% for hydrogen, dispersion, and others interaction groups, whereas aug-cc-pVDZ+2*33211 gives $\Delta_{abs}^{rel}$ of 2.5%, 4.0%, and 3.9%, respectively. The CCSD(T) results display similar errors and reductions in errors upon inclusion of two midbond centers. CCSD(T) results using aug-cc-pVDZ+33211 give $\Delta_{abs}^{rel}$ of 6.0%, 6.8%, and 6.5% for hydrogen, dispersion, and others interaction groups, whereas using aug-cc-pVDZ+2*33211 gives $\Delta_{abs}^{rel}$ of 4.1%, 3.9%, and 4.9%, respectively. The dispersion group is more affected than the other two interaction types, but compared to results for cc-pVDZ and cc-pVTZ, the improvement brought by the second midbond center is reduced in case of aug-cc-pVDZ results. The aug-cc-pVDZ basis set treats individual molecules (and interaction region) better due to the presence of diffuse functions in the basis set itself. Hence, aug-cc-pVDZ basis set supplemented with one midbond function is in general a good choice for the description of weak interactions in S66 complexes, especially to obtain cost efficient description of dispersion. The addition of the second midbond center only has a secondary effect on the accuracy.

The results for the S66 data set demonstrate the importance of both sufficiently describing molecule regions as well as the interaction region, and that the requirement on the molecules is greater for the S66 data set than for the A24 data set. It is likely that for small complexes, the

### Table 3

|            | 33211 | 2*33211 |          |          |          |          |
|------------|-------|---------|----------|----------|----------|----------|
|            | MP2   | CCSD(T) | MP2      | CCSD(T)  |          |          |
|            | $\Delta_{abs}^{rel}$ | $\Delta_{abs}^{rel}$ | $\Delta_{abs}^{rel}$ | $\Delta_{abs}^{rel}$ | $\Delta_{abs}^{rel}$ | $\Delta_{abs}^{rel}$ |
| DZ Hydrogen| 10.1  | 9.58    | 12.0     | 1.189    | 7.2      | 0.595    |
|           | 23.8  | 1.077   | 29.1     | 1.068    | 14.3     | 0.670    |
| Others    | 15.4  | 0.636   | 18.2     | 0.680    | 9.2      | 0.379    |
| TZ Hydrogen| 4.3   | 0.402   | 6.3      | 0.621    | 3.0      | 0.230    |
|           | 8.6   | 0.392   | 12.7     | 0.470    | 5.5      | 0.260    |
| Others    | 6.0   | 0.247   | 9.0      | 0.340    | 3.9      | 0.160    |
| aDZ Hydrogen| 4.3   | 0.489   | 6.0      | 0.697    | 2.5      | 0.258    |
|           | 6.3   | 0.270   | 6.8      | 0.252    | 4.0      | 0.177    |
| Others    | 5.1   | 0.212   | 6.5      | 0.249    | 3.9      | 0.159    |

Note: The reference values for CCSD(T) calculations can be found in Hobza et al., whereas the MP2/aug-cc-pVQZ+33211 reference values are found in Table S16. The errors are sorted according to interaction type.

### Figure 1

Relative mean absolute errors for 53 S66 complexes (see Section 2.3) obtained at MP2 and CCSD(T) levels of theory using (aug)-cc-pVXZ(Y*33211) (X = D, T; Y = 1, 2) basis sets.
FIGURE 2  Timings for MP2/aug-cc-pVDZ(+2*aDZ), MP2/aug-cc-pVDZ(+2*33211) and MP2/cc-pVTZ(+2*aTZ) measured relative to MP2/aug-cc-pVDZ(+33211) timings. Only 53 complexes of the S66 data set are considered as specified in Section 2.3. The assignment of the system numbers is according to the original publication of Hobza and coworkers75 and can be found in Table S25.

TABLE 4  Relative mean absolute errors ($\Delta_{\text{rel}}$, %) and mean absolute errors ($\Delta_{\text{abs}}$, kcal/mol) of interaction energies for 53 complexes of the S66 data set (as specified in Section 2.3) divided into interaction type calculated at MP2 and CCSD(T) levels of theory supplemented with midbond functions 2*3s3p2d1f1g (2*33211) or 2*aug-cc-pVXZ (2*aXZ).

| Interaction Type | cc-pVDZ | cc-pVTZ | aug-cc-pVDZ |
|------------------|---------|---------|-------------|
| Hydrogen         |         |         |             |
| MP2              | 7.2     | 3.0     | 2.5         |
| CCSD(T)          | 0.595   | 0.230   | 0.258       |
| Dispersion       | 14.3    | 5.5     | 4.0         |
| Others           | 9.2     | 9.2     | 3.9         |
| $\Delta_{\text{rel}}$ | 0.750   | 0.414   | 0.435       |
| $\Delta_{\text{abs}}$ | 16.3    | 8.4     | 6.3         |
| Hydrogen         |         |         |             |
| MP2              | 5.2     | 5.2     | 5.2         |
| CCSD(T)          | 0.191   | 0.510   | 0.510       |
| Dispersion       | 22.7    | 5.6     | 4.5         |
| Others           | 18.9    | 4.5     | 3.9         |
| $\Delta_{\text{rel}}$ | 1.018   | 0.261   | 0.177       |
| $\Delta_{\text{abs}}$ | 10.7    | 6.3     | 9.1         |
| Hydrogen         |         |         |             |
| MP2              | 6.1     | 8.0     | 8.3         |
| CCSD(T)          | 0.339   | 0.339   | 0.348       |
| Dispersion       | 21.3    | 4.5     | 3.9         |
| Others           | 12.7    | 7.0     | 10.3        |
| $\Delta_{\text{rel}}$ | 1.200   | 0.972   | 0.394       |
| $\Delta_{\text{abs}}$ | 21.3    | 21.3    | 21.3        |

Note: 2*aXZ means 2*aDZ for cc-pVDZ and aug-cc-pVDZ calculations and 2*aTZ for cc-pVTZ calculations.

FIGURE 3  Relative mean absolute errors for the S66 data set obtained at MP2 and CCSD(T) levels of theory using cc-pVDZ, cc-pVTZ, and aug-cc-pVDZ basis sets supplemented with two 3s3p2d1f1g (2*33211) or aug-cc-pVXZ (2*aXZ) midbond centers. 2*aXZ means 2*aDZ for cc-pVDZ and aug-cc-pVDZ calculations and 2*aTZ for cc-pVTZ calculations.
diffuse midbond functions also improve molecule description enough to obtain a balanced result, whereas this is not the case for larger complexes. As expected, only saturation of the interaction region by an increased number of midbond functions is insufficient for achieving accurate interaction energies for larger complexes, if using small atom centered basis sets, for example, cc-pVdz. Contrary, the combination of an augmented atom-centered basis set and one midbond center is seen to provide sufficient accuracy for describing interaction energies, also for large complexes. The accuracy can be further improved by either adding the second midbond center or increasing the atom-centered basis set (or both). However, this will raise the computational cost.

4.1.1 A comment on timings

Having examined the performance of different basis sets in combination with one and two 3s3p2d1f1g midbond centers, we now turn to a brief examination of their relative computational efficiency. For this purpose, we only consider the MP2 timings. As discussed above, aug-cc-pVdz(±2*33211) provides somewhat better results than aug-cc-pVdz(±33211) but the differences in errors are not dramatic. The largest difference for $\Delta_{\text{abs}}$ is found for the dispersion group where including two midbond centers rather than one reduces the error from 6.3% to 4.0%. The largest $\Delta_{\text{abs}}$ difference is found for the hydrogen group where including two midbond centers rather than one reduces the error from 0.489 to 0.258 kcal/mol. In Figure 2 the computational efficiency of aug-cc-pVdz(±2*33211) relative to aug-cc-pVdz(±33211) (wall time ratios) is visualized for each of the 53 S66 complexes. The results in Figure 2 show that the addition of the second 3s3p2d1f1g center to the interaction region increases the computational time by up to a factor two. It should be noted that calculations for small systems (containing less than 15 atoms), that is, systems numbers 3, 5, 6, 9, 10, and 60 are very fast, resulting in fluctuations in the relative timings. If these systems are not taken into account, an increase in calculation time of up to 1.8 times is observed.

4.2 Comparing the use of 3s3p2d1f1g and aug-cc-pVXZ as midbond functions

In Section 4.1 we see that using two midbond centers for S66 is seen to reduce the error in the interaction energies, in particular for dispersion group. Since the midbond set 3s3p2d1f1g contains diffuse functions of high angular momentum, calculations using several centers with this set increase computational cost (see Section 4.1.1). In this section we explore results for using the correlation consistent aug-cc-pVXZ basis sets for carbon as midbond set (contracted composition [3s,2p,1d] + diffuse (1s,1p,1d) for adZ and [4s,3p,2d,1f] + diffuse (1s,1p,1d,1f) for aTz), to see if standard basis sets can replace the specialized 3s3p2d1f1g. We compare results generated using two centers of 3s3p2d1f1g (denoted +2*33211, see Tables S17 and S21) and two centers of carbon aug-cc-pVXZ (denoted +2*aXZ, see Tables S18 and S22), where X of the midbond basis follows X of the main basis set, as was also the case in Section 3. We show MP2 and CCSD(T) results using cc-pVdz, cc-pVTZ and aug-cc-pVdz atom centered basis sets in combination with 2*33211 and 2*aXZ as midbond functions. Results for $\Delta_{\text{rel}}$ and $\Delta_{\text{abs}}$ for MP2 and CCSD(T) levels of theory using cc-pVXZ basis sets (denoted +2*XZ, see Tables S18 and S22) and two centers of carbon aug-cc-pVXZ are presented in Table 4 and visualized in Figure 3. Although we mainly discuss errors in terms of $\Delta_{\text{rel}}$, we also list $\Delta_{\text{abs}}$ results since equal $\Delta_{\text{rel}}$ values may correspond to different $\Delta_{\text{abs}}$ values (see Equations 5 and 6).

We first discuss the performance of the cc-pVdz and aug-cc-pVdz basis sets in combination with two 3s3p2d1f1g and two aug-cc-pVdz functions as midbond functions. The MP2/cc-pVdz results experience a large decrease in accuracy by replacing 2*3s3p2d1f1g functions with 2*aug-cc-pVdz functions. This effect is especially significant for dispersion and others groups, where $\Delta_{\text{rel}}$ is increased from 14.3% to 22.7% for dispersion and from 9.2% to 18.9% for others group. The same trend is observed for CCSD(T) with a $\Delta_{\text{rel}}$ increasing from 16.3% to 26.4% for dispersion group and from 10.7% to 21.3% for others group. Thus, the cc-pVdz (+2*aD) basis set seems to be an inadequate choice for description of noncovalent interactions. For the aug-cc-pVdz results we see the same trend as for cc-pVdz results, that is, that the accuracy for both MP2 and CCSD(T) results is reduced upon replacing the 3s3p2d1f1g midbond functions with the aug-cc-pVdz functions. In general, the $\Delta_{\text{abs}}$ errors

**Table 5** Relative mean absolute errors ($\Delta_{\text{abs}}$, %) and mean absolute errors ($\Delta_{\text{abs}}$, kcal/mol) of interaction energies for S66 complexes divided into interaction type calculated at MP2 and CCSD(T) levels of theory using cc-pVXZ (+33211), X = D, T, and aug-cc-pVdz (+33211) for systematic and intuitive placement of midbond centers.

| Systematic | Intuitive |
|------------|-----------|
| MP2 | CCSD(T) | MP2 | CCSD(T) |
| $\Delta_{\text{rel}}$ | $\Delta_{\text{abs}}$ | $\Delta_{\text{rel}}$ | $\Delta_{\text{abs}}$ | $\Delta_{\text{rel}}$ | $\Delta_{\text{abs}}$ | $\Delta_{\text{rel}}$ | $\Delta_{\text{abs}}$ |
| cc-pVdz | Hydrogen | 11.1 | 0.918 | 12.9 | 1.107 | 12.9 | 1.023 | 14.7 | 1.215 |
| | Dispersion | 23.8 | 1.077 | 29.1 | 1.068 | 24.8 | 1.125 | 30.3 | 1.119 |
| | Others | 13.0 | 0.526 | 15.7 | 0.576 | 13.6 | 0.552 | 16.5 | 0.604 |
| cc-pVTZ | Hydrogen | 4.6 | 0.379 | 6.5 | 0.564 | 5.0 | 0.403 | 6.9 | 0.588 |
| | Dispersion | 8.6 | 0.392 | 12.7 | 0.470 | 8.9 | 0.407 | 13.2 | 0.488 |
| | Others | 5.1 | 0.206 | 7.9 | 0.291 | 5.3 | 0.214 | 8.2 | 0.300 |
| aug-cc-pVdz | Hydrogen | 3.7 | 0.375 | 5.3 | 0.547 | 3.6 | 0.372 | 5.3 | 0.546 |
| | Dispersion | 6.3 | 0.270 | 6.8 | 0.252 | 6.4 | 0.279 | 7.0 | 0.263 |
| | Others | 4.7 | 0.185 | 6.3 | 0.228 | 4.9 | 0.196 | 6.6 | 0.240 |
are approximately doubled in size, except for CCSD(T) results for the hydrogen group where \( \Delta_{\text{rel}} \) is increased from 4.1% to 7.1%.

We now consider the cc-pVTZ results. From Table 4 and Figure 3 we see that the cc-pVTZ results are much less affected by replacing 3s3p2d1f1g midbond functions with correlation consistent basis sets. For example, the \( \Delta_{\text{rel}} \) which is 4.7% for CCSD(T)/cc-pVTZ\(+2^*33211\) decreases to 4.3% for CCSD(T)/cc-pVTZ\(+2^*aTZ\) in case of hydrogen group, whereas in case of dispersion group \( \Delta_{\text{rel}} \) obtained for CCSD(T) calculations is 8.4% independent of the type of midbond set used. Hence, TZ computations in combination with either of the two types of midbond sets give reasonable results for all interaction types.

From the results above we are in position to discuss what are the criteria for a good midbond set. The carbon aug-cc-pVDZ functions have smaller exponents than 3s3p2d1f1g functions (the smallest exponents for aug-cc-pVDZ are 0.04690, 0.04041, 0.15100 for \( s, p, d \)-functions, respectively, versus 0.100, 0.100, and 0.200 for \( s, p, d \)-functions of the 3s3p2d1f1g set). Hence, they are more diffuse and should cover more of the interaction region. However, 3s3p2d1f1g midbond centers contain functions of higher angular momentum (\( f \) and \( g \) functions). The significant decrease in accuracy when replacing the 3s3p2d1f1g set with an aug-cc-pVDZ for the cc-pVDZ and aug-cc-pVDZ results indicates that high angular momentum of the midbond basis is as important as functions being diffuse. This assumption is supported by the cc-pVTZ results where aug-cc-pVTZ midbond functions were used. The smallest exponents for aug-cc-pVTZ are 0.04402, 0.03569, 0.10000 and 0.26800 for \( s, p, d, f \)-functions, that is, only slightly smaller than for the aug-cc-pVDZ functions. However, carbon aug-cc-pVTZ functions contain an \( f \)-type function, which aug-cc-pVDZ does not. This supports the assumption about the importance of high angular momentum functions in midbond set for the quality of interaction energies made earlier in this section. While it has been proposed that basis sets where the angular momentum of the midbond set is higher than for the atom centered basis set are imbalanced, Patkowski and collaborators argue that this imbalance is not necessarily bad. Our results support the statement of Patkowski and collaborators, because we see that functions of high angular momentum in the midbond set are important to obtain accurate interaction energies.

4.2.1 | A comment on timings

In this section we discuss the relative computational efficiency of aug-cc-pVDZ\(+2^*aDZ\) and cc-pVTZ\(+2^*aTZ\) using the ratio of their wall time to the wall time of the aug-cc-pVDZ\(+33211\), see Figure 2. In Section 4.2 we demonstrate that cc-pVTZ\(+2^*aTZ\) results are of the comparable quality as the aug-cc-pVDZ\(+2^*33211\) results. From Figure 2 we see that the cc-pVTZ\(+2^*aTZ\) calculations are significantly more time consuming than aug-cc-pVDZ\(+33211\) and aug-cc-pVDZ\(+2^*33211\) calculations. In particular, cc-pVTZ\(+2^*aTZ\) calculations take up to 3.6 times longer than corresponding calculations for the systems number 3, 5, 6, 9, 10, 60 containing less than 15 atoms than the corresponding aug-cc-pVDZ\(+33211\) and aug-cc-pVDZ\(+2^*33211\) calculations. Therefore, aug-cc-pVDZ\(+33211\) or \(+2^*33211\) stand out as good choices for computing of interaction energies of noncovalent complexes.

4.3 | Systematic versus intuitive placement

In this section we explore how much the quality of computed interaction energies depend on whether using a systematic approach (see Section 2.3) for placement of midbond centers versus using an intuitive approach where centers are placed manually based on chemical intuition. All input geometries including the location of the midbond centers can be found at https://doi.org/10.18710/2FWECY. The presented calculations are performed at MP2 and CCSD(T) levels of theory using cc-pVXZ\( (X = D, T) \) and aug-cc-pVDZ basis sets (see Tables S19, S20, S23, and S24). The \( \Delta_{\text{rel}} \) errors are given in Table 5, and for illustrative purposes \( \Delta_{\text{rel}} \) are visualized in Figure 4.

For cc-pVDZ\(+33211\) the smallest errors are found for the hydrogen group where \( \Delta_{\text{rel}} \) errors range between 11.1% and 14.7%. For the dispersion group the \( \Delta_{\text{rel}} \) errors range between 23.8% and 30.3%. Hence, the cc-pVDZ\(+33211\) results are generally unreliable, no matter which approach is used for placing the midbond centers. For cc-pVTZ\(+33211\) the results are significantly better than the cc-pVDZ results, but the approach for how to place the midbond
centers only influences the results to a small extent. For example, MP2/cc-pVTZ(3s3p2d1f1g) using the systematic approach gives an \( \Delta \text{abs} \) error of 4.6% for the hydrogen group whereas the intuitive placement gives an error of 5.0%. For CCSD(T) the \( \Delta \text{abs} \) errors are 6.5% and 6.9% for systematic and intuitive approach, respectively, for the hydrogen group. For the dispersion group the MP2/cc-pVTZ(3s3p2d1f1g) \( \Delta \text{abs} \) errors are 8.6% and 8.9% for systematic and intuitive placement, respectively, and the CCSD(T) numbers are 12.7% and 13.2%. For aug-cc-pVQZ(3s3p2d1f1g) we see the same trends as for the other basis sets, namely that there is small dependence on how the placement of midbond centers is determined. In general, although the systematic placement seems to consistently give smaller errors than the intuitive approach, the difference is very small compared to difference introduced by choice of basis set. Overall, the quality of the results attained with both approaches offer distinct advantages. The systematic approach can be beneficial for large data sets of molecules, where the process of locating midbond functions can be automated, whereas the intuitive approach is simple for single systems.

5 | CONCLUSIONS

In this article we use MP2 and CCSD(T) calculations for the A24 and S66 data sets to explore how midbond functions efficiently can be used to generate cost effective CP corrected supramolecular interaction energies of noncovalent complexes. We have used the A24 data set to show that the primary role of midbond centers is not to more rapidly reach the CBS limit for the dimer, but rather in providing a balanced description of the interaction region and the molecules. The need for balance is a consequence of choosing to use atom centered basis sets to describe the electronic structure. The error associated with the imbalanced description is seldom discussed in its own right, but neglected compared to, for example, the BSSE. If CBS limit is reached both errors with respect to an imbalanced description and BSSEs disappear, but for larger complexes it is not feasible to go to the basis set limit.

Further, we have used the S66 data set to explore how midbond functions affect interaction energies of larger molecular systems and whether requirements for large systems are different to those for small systems. We have studied whether it is beneficial to use more than one midbond center in the interaction region of the S66 complexes. However, we show that one midbond center combined with a basis set such as aug-cc-pVQZ yields cost effective results for the S66 data set. Increasing up to two midbond centers improves the results. Results obtained using cc-pVQZ are significantly worse for the S66 data set than for the A24 data set. For small systems such as those in the A24 data set, the midbond functions are able to also improve the description of molecules, and hence reducing the requirement on size of the atom centered basis set. This yields reasonable results even for the cc-pVQZ basis set, as seen both in this paper and in literature. We show that for larger systems, illustrated by using the S66 data set, the role of the midbond functions is primarily to provide the balanced description, since complexes are too large to get their description improved by the midbond functions. Hence, the requirement on choice of atom centered basis set is somewhat stricter for larger complexes. However, including midbond functions in the calculations to improve the flexibility in the interaction region allows for good results to be produced using basis sets, such as aug-cc-pVQZ, which traditionally are deemed insufficient for computing interaction energies.

By comparing calculations using the 3s3p2d1f1g midbond set with the use of aug-cc-pVQZ and aug-cc-pVQZ basis sets as midbond sets, we see that the requirements for the midbond set to be effective, is not just that it contains diffuse functions, but also that high angular momentum functions (at least f-functions) are included. Results for two ways of placing midbond centers show that interaction energies are not sensitive to exact placement as long as it is reasonable. One approach is based on a weighted average of intermolecular atom-atom midpoints (systematic) and the other one is based on the use of chemical intuition (intuitive), which each have their advantages. The systematic approach can be beneficial for large data sets of molecules, whereas the intuitive approach is simple when looking at a particular molecular system.

ACKNOWLEDGMENTS

The authors acknowledge computing resources through UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway—through project numbers nn2962k and nn9409k. I-M. H. and H. K. acknowledges the Research Council of Norway through FRINATEK projects 263110 and 275506.

DATA AVAILABILITY STATEMENT

Geometries of the complexes (including midbond centers) of the A24 and S66 data sets and raw data for presented results are available at https://doi.org/10.18710/2FWECY.

ORCID

Regina Matveeva https://orcid.org/0000-0001-6567-3570
Henrik Koch https://orcid.org/0000-0002-8367-8727
Ida-Marie Høyvik https://orcid.org/0000-0002-1239-7776

ENDNOTE

For system names see Table S25 and for geometries including two midbond centers see https://doi.org/10.18710/2FWECY

REFERENCES

[1] E. Mattia, S. Otto, Nat. Nanotechnol. 2015, 10, 111.
[2] Z. Huang, K. Qin, G. Deng, G. Wu, Y. Bai, J. F. Xu, Z. Wang, Z. Yu, O. A. Scherman, X. Zhang, Langmuir 2016, 32, 12352.
[3] E. Frieden, J. Chem. Educ. 1975, 52, 754.
[4] K. E. Riley, P. Hobza, Wiley Interdiscip. Rev. Comput. Mol. Sci. 2011, 1, 3.
[5] A. Erxleben, Chimia 2017, 71, 102.
[6] P. A. Kollman, Acc. Chem. Res. 1977, 10, 365.
[7] A. Buckingham, P. Fowler, Can. J. Chem. 1985, 63, 2018.
[8] A. Buckingham, P. Fowler, J. M. Hutson, Chem. Rev. 1988, 88, 963.
[9] U. Singh, P. A. Kollman, J. Chem. Phys. 1985, 83, 4033.
[10] K. Müller-Dethlefs, P. Hobza, Chem. Rev. 2000, 100, 143.
[11] P. Jurečka, J. Sponer, J. Černý, P. Hobza, Phys. Chem. Chem. Phys. 2006, 8, 1985.
[12] J. Černý, P. Hobza, Phys. Chem. Chem. Phys. 2007, 9, 5291.
