Basis Set Dependence of Interaction Energies Computed Using Composite Post-MP2 Methods

James A. Platts,[†],* J. Grant Hill,[‡] Kevin E. Riley,[§] Jan Řezáč,[||] and Pavel Hobza[||]

[†]School of Chemistry, Cardiff University, Park Place, Cardiff CF10 3AT, United Kingdom
[‡]School of Chemistry, University of Glasgow, Joseph Black Building, University Ave, Glasgow, G12 8QQ, United Kingdom
[§]Department of Chemistry, Xavier University of Louisiana, 1 Drexel Drive, Box 22, New Orleans, Louisiana 70125, United States
[||]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, 166 10 Prague, Czech Republic

Ⓢ *Supporting Information*

ABSTRACT: We report the performance of composite post-MP2 *ab initio* methods with small basis sets for description of noncovalent interactions, using the S66 data set as a benchmark. For three representative complexes, it is shown that explicitly correlated coupled cluster (CCSD-F12a) methods yield interaction energies ca. 0.1 kcal/mol from the complete basis set limit with aug-cc-pVDZ. Triple excitations are not explicitly correlated in this approach, but we show that scaling the perturbative triples via the (T*) approximation improves agreement with benchmark values. Across the entire S66 data set, this approach results in a root-mean-square error (RMSE) of 0.13 kcal/mol or 3%, with well-balanced description of all classes of complex. The basis set dependence of traditional CCSD(T) interaction energies is examined, and the small 6-31G*(0.25) basis set is found to give particularly accurate results (RMSE = 0.15 kcal/mol, or 4%). We also employ spin component scaling (SCS) of CCSD-F12a data, which gives slightly better accuracy than CCSD(T*)-F12a if contributions from same- and opposite-spin pairs are optimized for this data set (RMSE = 0.08 kcal/mol, or 2%). Interpolation of local MP2 and MP3 is also shown to accurately reproduce benchmark data with both aug-cc-pVDZ (RMSE = 0.18 kcal/mol or 5%) and 6-31G*(0.25) (RMSE = 0.13 kcal/mol or 4%).
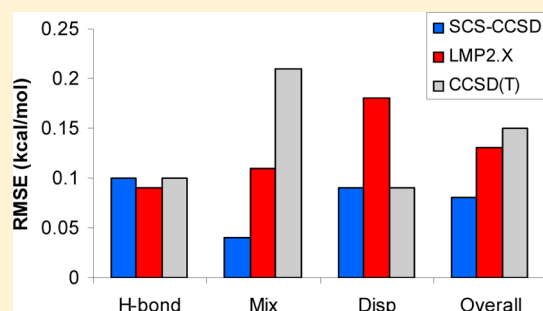
INTRODUCTION

Noncovalent interactions are ubiquitous throughout nature and play important roles in many areas of chemistry, biology, and material science, such as the structure and properties of materials, the bulk properties of liquids, and the structure of biological molecules and their interactions with ligands.[1] An accurate description of noncovalent interactions using computational tools is, in general, not straightforward and requires the use of sophisticated, and computationally expensive, methods.[2] The coupled cluster method including single and double excitations and perturbative triple excitations (CCSD(T)) has been shown to give accurate descriptions for a variety of different interactions and is widely regarded as a benchmark for these.[2] The CCSD(T) method is very resource intensive and requires the use of large basis sets (i.e. aug-cc-pVTZ and larger) to yield accurate results; as such, this technique can only be used for relatively small systems containing up to ~30 atoms.

There are a number of less computationally intensive techniques that can yield reasonably accurate results for certain classes of noncovalent interactions. MP2 has been used to describe a large variety of noncovalent interactions for many years with relatively accurate, but widely varying, results, which

strongly depend on the basis set used.[3,4] When used with large basis sets, or at the extrapolated complete basis set limit (CBS), this method typically produces accurate binding energies for hydrogen bonding complexes and dispersion interactions involving aliphatic species, while strongly overbinding for dispersion interactions involving aromatic groups.[3] In recent years, there have been several attempts to improve the accuracy of MP2 (in terms of noncovalent interactions as well as other molecular properties). Variants of spin component scaled MP2 (SCS-MP2),[5] a technique that scales the contributions from the same-spin (SS) and opposite-spin (OS) MP2 terms, produce improved results for noncovalent complexes. The SCS-MI-MP2 method of Distasio et al.[6] and the SCS-S66-MP2 method of Riley et al.,[3] which were both parametrized specifically for noncovalent interactions, yield binding energies that are significantly more accurate for a variety of interaction types.

MP2/CBS binding energies can be improved by the addition of correction terms based on more accurate correlation methods, computational techniques utilizing these types of corrections are often referred to as composite post-MP2

methods. The most important of these takes advantage of the fact that there is very little basis set dependence in the difference between the CCSD(T) and MP2 binding energies. A correction, $\Delta E(CCSD(T)) - \Delta E(MP2)$, is added to the MP2/CBS binding energy in order to approximate the CCSD(T)/CBS energy, as in eq 1.[7−11]

$$
\begin{aligned}
\Delta E&(CCSD(T)/CBS) \\
&\approx \Delta E(MP2/CBS) + [\Delta E(CCSD(T)/\text{bas-x}) \\
&\quad - \Delta E(MP2/\text{bas-x})]
\end{aligned} \tag{1}
$$

Typically, the MP2/CBS term is computed by extrapolation, using the technique of Helgaker and co-workers,[12] from the correlation consistent basis sets of Dunning (most commonly aug-cc-pVDZ and aug-cc-pVTZ).[13] The size of the basis set, bas-x, used to compute the ΔCCSD(T) term has a large effect on the quality of the binding energy, and usually, a basis set at least as large as aug-cc-pVDZ is used.[2] When used with appropriately large basis sets, this method produces extremely accurate binding energies and is the method most commonly used to compute benchmark-quality data for noncovalent interactions.

Another type of composite post-MP2 correction makes use of the fact that, while MP2 tends to overestimate binding energies, MP3 tends to underestimate them. The magnitudes of the over/underestimation predicted by the MP2 and MP3 methods, respectively, are very similar. This is the basis for the MP2.5[14] and MP2.X[15] methods, which include an MP2.5 correction, the average of binding energies obtained by MP2 and MP3, $[\Delta E(MP3) - \Delta E(MP2)]/2$. The MP2.X binding energy is given as

$$
\begin{aligned}
\Delta E&(MP2.X/CBS) \\
&= \Delta E(MP2/CBS) + X[\Delta E(MP3/\text{bas-x}) \\
&\quad - \Delta E(MP2/\text{bas-x})]
\end{aligned} \tag{2}
$$

Here, the coefficient X of the MP2.X correction may be varied to minimize errors for a given basis set or fixed at exactly 0.5 to give the original MP2.5 method. An advantage of this approach is that the correction is relatively basis set independent. MP2.5 (i.e., $X = 0.5$) is appropriate when bas-x is relatively large (at least as large as aug-cc-pVDZ), while the optimal MP2.X coefficient, X, will tend to deviate from 0.5 when smaller basis sets are used. A recent study showed that suitably chosen small basis sets can yield high accuracy (RMSE $\approx$ 0.15 kcal/mol for the S66 data set) after optimization of the value of X.[15]

In general, the accuracy and computational cost of binding energy computations increases going from MP2 to MP2.X to CCSD(T). The relative accuracy and expense of all of these methods generally depends (sometimes strongly) on the basis sets being employed. The purpose of the current study is to explore the use of relatively small basis sets with composite post-MP2 descriptions of noncovalent interactions, including local and explicit correlation methods. Local correlation methods exploit the "near-sighted" nature of electron correlation by carrying out the calculation in a set of localized orbitals. Only excitations to virtual orbitals that are spatially close to the occupied ones are considered, leading to significant efficiency gains over conventional methods.[16] However, we note that with the small basis sets considered in this work one of the potential advantages of local correlation, that basis set

superposition error (BSSE) is effectively eliminated, cannot be exploited due to residual BSSE in the HF component of the energy, such that counterpoise corrections are required. Explicitly correlated (F12) methods circumvent the slow convergence of correlation energy with respect to basis set size by including terms that depend explicitly on the interelectron distance, r_{12}, and hence yield results close to the CBS limit even with relatively small basis sets.[17]

The S66 data set describes the most important binding motifs found between neutral species in biochemical systems, namely, hydrogen bonding, dispersion interactions, and mixtures of these.[18,19] It was designed to supersede the popular S22 data set, which overemphasized cyclic hydrogen bonds and stacking from nucleic acids. Structures and reference binding energies for this set are available online (http://www.begdb.com/).[20] The accuracy of the S66 benchmark set was discussed in the original paper. We estimate that the average relative error of the best CBS extrapolation scheme used for the S66 set is about 1%. This error was calculated in a set of small systems for which very accurate determination of the CBS limit is possible. The benchmark interaction energies are not relativistic and use the frozen core approximation, and the same is used in the methods tested herein, allowing for valid comparison.

■ COMPUTATIONAL METHODS

Benchmark interaction energy data for the S66 data set, a balanced set of intermolecular complexes containing hydrogen bonded, dispersion bound and "mixed" categories were obtained from the www.begdb.com database, along with atomic coordinates.[20] Following previous work,[3] we report both root-mean-square error (RMSE) and percentage RMSE (%RMSE).

All MP2-F12 calculations used the DF-MP2-F12 method (shortened to MP2-F12 herein),[21] along with density fitted Hartree−Fock (DF-HF).[22] Explicitly correlated coupled cluster calculations were performed using both approximations F12a and F12b,[23] with a conventional (not density fitted) HF reference. All explicitly correlated calculations used the diagonal, fixed amplitude 3C(FIX) Ansatz.[24] The orbital basis sets were either from the cc-pVnZ-F12 (shortened to VnZ-F12)[25] or aug-cc-pVnZ (AVnZ)[13,26] families of correlation consistent basis sets. The OptRI auxiliary basis sets of Yousaf and Peterson were used in the resolution of the identity (RI) that efficiently approximates the many-electron integrals arising in F12 theory. This auxiliary basis is merged with the orbital basis within the complementary auxiliary basis set (CABS) approach,[27] as implemented in Molpro 2010.1.[23a,28] Density fitting of the Fock and exchange matrices (both in F12 calculations and DF-HF) employed the cc-pVnZ/JKFit auxiliary basis sets of Weigend,[29] with fitting of the remaining two-electron integrals performed with the aug-cc-pVnZ/MP2Fit auxiliary sets.[30] The combinations of orbital and auxiliary basis sets matched those detailed elsewhere.[31] The geminal Slater exponent (β) in the F12 methods was also set to values recommended in previous work: 0.9, 1.0, and 1.0 for VDZ-F12, VTZ-F12, and VQZ-F12, respectively, and 1.0, 1.2, 1.4, and 1.5 for AVDZ, AVTZ, AVQZ, and AV5Z, respectively.[31,32]

Extrapolations to estimate the complete basis set (CBS) limit for explicitly correlated methods were carried out using the generalized Schwenke two-point extrapolation of correlation energies:[33]

$$
E_{CBS}^{corr} = (E_{large}^{corr} - E_{small}^{corr})F + E_{small}^{corr}
$$

Table 1. Counterpoise Corrected Interaction Energies (kcal/mol) for Three Complexes[a]

method	basis set	water dimer	uracil...ethene	benzene...ethyne
S66[b]		−5.01	−3.33	−2.86
MP2-F12	VDZ-F12	−4.92 (0.14)	−3.85 (0.22)	−3.34 (0.20)
	VTZ-F12	−4.94 (0.07)	−3.94 (0.11)	−3.42 (0.10)
	VQZ-F12	−4.96 (0.03)	−4.00 (0.04)	−3.46 (0.03)
	CBS	−4.96	−4.03	−3.47
MP2-F12	AVDZ	−4.88 (0.19)	−3.93 (0.44)	−3.36 (0.41)
	AVTZ	−4.94 (0.08)	−3.98 (0.17)	−3.45 (0.18)
	AVQZ	−4.96 (0.04)	−4.00 (0.09)	−3.46 (0.09)
	AV5Z	−4.96 (0.02)	−4.02 (0.03)	
	CBS	−4.96	−4.02	−3.47[c]
CCSD-F12a	AVDZ	−4.67 (0.19)	−2.55 (0.40)	−2.36 (0.37)
	AVTZ	−4.74 (0.09)	−2.51 (0.13)	−2.40 (0.15)
	AVQZ	−4.76 (0.02)		−2.40 (0.03)
	AV5Z	−4.76 (0.00)		
	CBS[d]	−4.75	−2.48	−2.39
CCSD(T)-F12a	AVDZ	−4.83 (0.30)	−3.24 (0.69)	−2.73 (0.69)
	AVTZ	−4.96 (0.12)	−3.29 (0.21)	−2.82 (0.22)
	AVQZ	−4.99 (0.04)		−2.84 (0.06)
	AV5Z	−4.99 (0.00)		
	CBS[d]	−4.99	−3.31	−2.84
CCSD(T*)-F12a	AVDZ	−4.88 (0.33)	−3.45 (0.78)	−2.83 (0.77)
	AVTZ	−4.98 (0.12)	−3.37 (0.22)	−2.86 (0.22)
	AVQZ	−5.00 (0.04)		−2.86 (0.06)
	AV5Z	−5.00 (0.00)		

[a]Values in parentheses indicate the magnitude of the basis set superposition error removed. [b]Approximate CCSD(T)/CBS, taken from ref 18. [c]Using AV5Z resulted in linear dependency errors: the CBS extrapolation was thus performed with AVTZ and AVQZ energies. [d]As extrapolation coefficients for CCSD-F12a are not available, CBS values are estimated using F12b.

with the corresponding HF energy taken as the larger basis set energy including CABS singles corrections. The F coefficients used in the extrapolation have been optimized for specific pairs of basis sets and are taken from ref 31. The interaction energies were corrected for BSSE using the counterpoise (CP) method of Boys and Bernardi.[34] In some cases, the (T) triples contribution was scaled by a simple correction known as (T*),[23b] which employs a scale factor of $E_{corr}^{MP2-F12}/E_{corr}^{MP2}$. Size consistency errors in (T*) were avoided by determining the ratio for the dimer and using it unchanged for the monomers.[35]

For local correlation approaches with correlation consistent orbital basis sets, Hartree−Fock orbitals were localized according to the Pipek−Mezey scheme[36] with contributions from the two most diffuse basis functions of each angular momentum type removed from the localization criterion. Domain selection was carried out using a completeness criterion of 0.985, with the procedure described by Boughton and Pulay.[37] However, the Pipek−Mezey method gave convergence problems when used with the Pople-style basis sets, and in these cases, the natural population analysis (NPA) scheme[38] was used instead, using the domain selection (NPASEL) criterion of 0.07, as recommended by Werner and Schütz.[39] Invariance to unitary transformations of the π-orbitals was achieved through merging the domains of those orbitals.

The default in Molpro for local correlation methods beyond MP2 is to use the specified method only for strongly correlated electron pairs, approximating more distant ones with LMP2.[40] Since strongly correlated pairs are typically located within the same molecule, this would correspond to LMP2 description of intermolecular correlation. We therefore include weakly correlated electron pairs in the specified method through use of keywords "rweak" and "rdist". Basis set superposition error

(BSSE) is non-negligible when local correlation is coupled with the relatively small basis sets used here and was accounted for using counterpoise corrections.[34]

■ RESULTS AND DISCUSSION

We begin by considering the basis set dependence and convergence of interaction energies using explicitly correlated methods. Table 1 reports calculated data for three complexes, taken as representatives of the hydrogen bonded, dispersion bound, and "mixed" classes. It can be seen that the MP2-F12 basis set convergence for the water dimer in particular is very rapid indeed. The VDZ-F12 and AVDZ interaction energies are 0.04 and 0.08 kcal/mol from the CBS estimate, respectively. With both families of basis sets, the CP corrected interaction energy reaches the CBS limit with QZ basis sets, and larger basis sets simply reduce the BSSE. Convergence is slower for the larger complexes, but even here, AVDZ gives interaction energies approximately 0.1 kcal/mol from the MP2/CBS estimate. AVTZ performs very well, with errors (relative to the MP2 limit) of 0.02, 0.04, and 0.02 kcal/mol (i.e. much less than the intrinsic accuracy of MP2).

It is reassuring that for all three systems the MP2-F12 CBS estimates from each family of basis set agree to within 0.01 kcal/mol, indicating that the lack of higher angular momentum diffuse functions in the VnZ-F12 sets does not pose any problems for the extrapolation. However, a comparison between VnZ-F12 and AVnZ sets for the uracil...ethane and benzene...ethyne complexes indicates that with smaller basis sets the AVnZ family produces results closer to the basis set limits. This demonstrates the importance of the higher angular momentum diffuse functions in describing the primarily dispersion bound complexes, although such functions also

Table 2. RMSE and %RMSE for CCSD-F12a/AVDZ and Related Approaches (kcal/mol and %)

	RMSE (%)	H-bond	mixed	dispersion
CCSD(T*)-F12a	0.13 (3.3)	0.13 (1.4)	0.06 (1.8)	0.18 (5.2)
CCSD-F12a	0.84 (20.9)	0.74 (8.2)	0.67 (18.7)	1.04 (29.6)
SCS-CCSD-F12a[a]	0.13 (3.0)	0.19 (2.7)	0.06 (1.7)	0.11 (4.0)
SCS-MI-CCSD-F12a[b]	0.12 (2.7)	0.17 (2.6)	0.06 (1.5)	0.10 (3.5)
SCS-S66-CCSD-F12a[c]	0.08 (2.3)	0.10 (2.3)	0.04 (1.0)	0.09 (2.9)

[a]SS/OS = 1.27/1.13, ref 46. [b]SS/OS = 1.28/1.11, ref 45. [c]SS/OS = 1.36/1.00, this work.

increase the magnitude of the BSSE. As AVnZ for a given element possesses slightly fewer basis functions than VnZ-F12, it seems that AVnZ should remain the basis sets of choice when describing noncovalent interactions. It is worth noting that the VnZ-F12 basis sets still perform well at the MP2-F12 level and might be preferred if CP corrections are not carried out (as previously noted for the carbon dioxide dimer in ref 41).

It is well-known that MP2 overestimates the strength of dispersion interactions; thus, Table 1 shows the basis set convergence of the CCSD-F12a and CCSD(T)-F12a methods, along with estimates of the CCSD and CCSD(T) CBS limits. Due to the significant computational cost of the coupled cluster methods, some larger basis set calculations were beyond the capabilities of the computational resources available. F12a and F12b (see the Supporting Information) approximations give rise to only minor differences, with the F12a approximation producing results that are generally slightly closer to the CBS estimates. However, it is worth noting that CCSD-F12a with larger basis sets exceeds the CCSD/CBS values in several cases, which has been noted in several other investigations.[42,23b] This data illustrates that most of the trends observed for MP2-F12 also apply for the CCSD-F12a methods, although it requires a TZ quality basis to be within 0.1 kcal/mol of the CBS limit, rather than DZ in the MP2 case.

Comparison with literature benchmark data shows the importance of triples terms, as CCSD-F12a values are in several cases worse than corresponding MP2-F12 data. As the (T) component is not explicitly correlated, it should converge more slowly than CCSD-F12, and increase the total BSSE. Scaling the perturbative triples via the (T*) approximation generally improves agreement with benchmark values, especially when used with the smallest basis sets, although the substantial overshooting for uracil...ethene with AVDZ serves as a timely reminder of the inconsistencies that can occur when employing scaling factors. Some degree of caution should be employed when comparing to the CBS limits for uracil...ethene, as our CBS estimates are produced from DZ-TZ extrapolation. It is therefore gratifying to note that not only do CBS estimates agree between the two families of basis sets employed in this work but also they are in excellent agreement with approximate CCSD(T)/CBS in Table 1.

Although a sample size of three complexes is admittedly small, the trends established in Table 1 indicate that for large complexes CCSD(T*)-F12a/AVDZ should represent a good level of theory choice if larger basis sets prove prohibitive. This results in absolute errors relative to the CBS estimates of 0.11, 0.14, and 0.01 kcal/mol for the water dimer, uracil...ethene, and benzene...ethyne, respectively. Increasing the basis set size to AVTZ ensures these errors are below the estimated limit of accuracy of these methods (ca. 0.1 kcal/mol). These findings expand upon previous work,[42−44] in which it was also found that CCSD(T*)-F12a/AVTZ produces excellent results for

slightly smaller complexes bound by intermolecular interactions.

To properly test the suitability of this approach for the description of noncovalent interactions, Table 2 reports data for the entire S66 data set relative to the previously published benchmark data. This shows comparable performance to that found for the three example complexes with AVDZ, yielding overall RMSE of 0.13 kcal/mol, or 3%. The method is well-balanced across the three categories of complex. Within categories, single H-bonds (RMSE = 0.10 kcal/mol) are rather better described than cyclic ones (0.21 kcal/mol), as are aliphatic−π and aliphatic−aliphatic complexes (0.14 and 0.09 kcal/mol, respectively) compared to π−π complexes (0.23 kcal/mol). Mean signed error (MSE) for this method is +0.02 kcal/mol, again indicating a balanced set of results, and the maximum/minimum errors are found for stacked uracil dimer (0.30 kcal/mol, or 3.1%) and acetic acid dimer (−0.27 kcal/mol, or 1.4%).

CCSD-F12a/AVDZ results are also reported in Table 2 and are slightly worse than the CCSD/CBS data reported in ref 18. (RMSE = 0.70 kcal/mol). The importance of triple excitations is evident from the rather large errors for all categories, but this systematic shortcoming can be ameliorated by spin-component scaling. Table 2 reports three variations on this theme: all three approaches perform well, resulting in smaller errors comparable to or better than CCSD(T*)-F12a. The small improvement for SCS-MI[45] (specifically designed for intermolecular interactions) over SCS[46] (originally suggested by Takatani et al. for reaction energies) is rather different from the case reported by Pitonak et al., presumably reflecting the difference between CCSD-F12 used here and CCSD/CBS used in ref 18. Optimizing the scaling factors for this method and data set results in RMSE of 0.08 kcal/mol, identical to that reported previously for SCS-MI-CCSD/CBS.[18] The MSE for this model is exactly 0.00 kcal/mol, and the maximum/minimum errors are +0.24 (or 8.3%, for stacked benzene dimer) and −0.44 kcal/mol (or 10.7%, for methylamine H-bond dimer), further supporting the excellent results from this approach.

Of course, a method that is specifically trained on a certain data set may be biased toward that data set and might not perform well for other data. To test the robustness of our parameters, we randomly split the entire S66 data set into two training and test sets of equal size. Optimizing the SCS parameters for only half of the data allows us to predict interaction energies for the remaining half and hence estimate likely performance for external data sets. This analysis reveals that the SS/OS scaling parameters reported at the foot of Table 2 are almost optimal for either training set and that the largest predicted RMSE is 0.10 kcal/mol that is, only marginally greater than that reported for the full data set. This lends confidence that the excellent performance of this method is not simply a result of fitting to this particular data set. Full details are reported in the Supporting Information.

333

dx.doi.org/10.1021/ct300842d | J. Chem. Theory Comput. 2013, 9, 330−337

While explicitly correlated methods and concomitant use of small basis sets such as AVDZ reduce computational requirements substantially over conventional methods, their application to larger complexes remains a challenge with current computational resources. We therefore continue our search for accurate but expedient approaches to calculation of non-covalent interaction energies using local correlation methods. Table 3 reports the results of counterpoise corrected LMP2.5,

Table 3. RMSE and %RMSE for Local Correlation Methods with AVDZ (kcal/mol and %)[a]

	RMSE (%)	H-bond	mixed	dispersion
LMP2.5	0.22 (7.1)	0.16 (1.8)	0.17 (4.9)	0.29 (11.0)
LMP2.X[b]	0.18 (5.0)	0.18 (2.2)	0.12 (3.7)	0.22 (7.4)
LCCx[c]	0.35 (10.1)	0.25 (2.9)	0.22 (6.7)	0.49 (15.6)
L-CBS(T0)	0.19 (3.0)	0.28 (3.0)	0.09 (2.4)	0.13 (3.4)

[a]Using MP2/CBS data, as in eq 2. [b]X = 0.61. [c]x = 0.14.

generated from local MP2 and MP3 with the AVDZ basis set, used as a correction to MP2/CBS, as in eq 2. Even without optimization of the amount of LMP3 correction, the overall performance is competitive with the best SCS-scaled results from ref 3 and represents a significant improvement over LMP2 (RMSE = 0.73 kcal/mol). However, these results are rather imbalanced in that dispersion-bound complexes have much larger errors than hydrogen bonded or mixed complexes. This is solely due to $\pi-\pi$ complexes, which give RMSE = 0.41 kcal/mol, whereas aliphatic−aliphatic and π−aliphatic complexes are well described, RMSE = 0.12 and 0.18 kcal/mol, respectively.

Optimization of the coefficient of the MP3 contribution improves the performance of this approach, reducing the overall RMSE. Moreover, performance is rather better balanced across categories: errors for hydrogen bonded complexes increase slightly while those for dispersion-bound and mixed complexes fall. All categories of interaction are described with similar accuracy: single and cyclic H-bonds have RMSE = 0.15

and 0.25 kcal/mol, respectively; $\pi-\pi$, aliphatic−aliphatic, and π−aliphatic complexes have RMSE = 0.23, 0.17, and 0.22 kcal/mol, respectively. The mean signed error (MSE) is −0.07 kcal/mol, indicating that the model is relatively unbiased, and the maximum and minimum errors are +0.34 (for stacked pyridine dimer) and −0.41 (for stacked uracil dimer), all of which indicate excellent and balanced performance. The optimized coefficient is noticeably larger than the equivalent from conventional MPx calculations (0.52).

As well as extending the Møller−Plesset series, local correlation approximations have been implemented for coupled cluster theory, of which LCCSD is the simplest available (see ref 47 and references therein). The large basis sets required for description of intermolecular interactions make such calculations demanding for the larger molecules in this set, so in the spirit of MP2.X, we apply a (LCCSD − LMP2) correction to the MP2/CBS binding energy from ref 18; the coefficient of this correction was found by minimizing the overall RMSE. Denoted LCCx in Table 3, the optimal coefficient of 0.14 is much smaller than those found for LMP2.X. This approach yields worse performance than LMP2.5 and LMP2.X. Given the extra computational effort required for LCCSD over LMP3, this approach does not appear to be a promising one.

Instead of relying on interpolation between MP2 and higher-level methods, it is also possible to use local correlation methods to directly approximate the benchmark data. To this end, LMP2 and LCCSD(T0) energies were calculated using AVDZ, and their difference was used as a correction to literature MP2/CBS binding energy: this is denoted L-CBS(T0) in Table 3. T0 differs from standard triples correct by neglecting intertriples couplings that arise from off-diagonal elements in the occupied-occupied block in the Fock matrix.[40] While this approach works reasonably well, the overall RMSE of 0.19 kcal/mol is slightly disappointing. Closer inspection indicates that performance is worst for hydrogen bonded complexes, especially cyclic H-bonds for which RMSE = 0.47 kcal/mol. Performance is worst for the acetic acid dimer, for
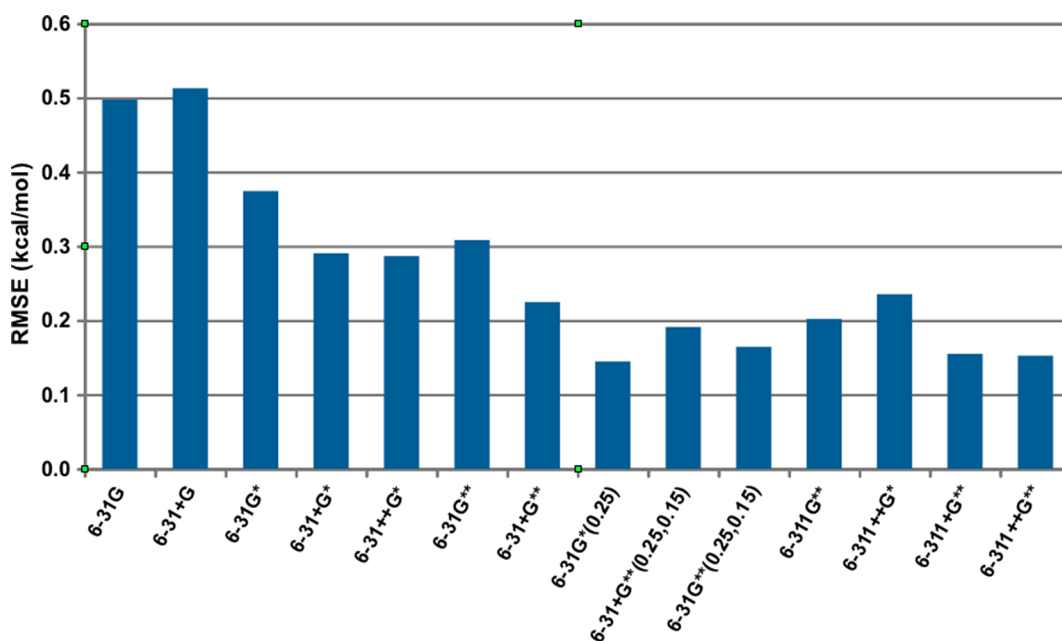

Figure 1. RMSE relative to data from ref 18 as a function of basis set on CCSD(T) corrections (in kcal/mol) in the approximate CCSD(T)/CBS method.

which the error is +0.69 kcal/mol or 3.5%. The reason for this rather poor performance for hydrogen bonded complexes is not clear, and given the more balanced behavior of other methods reported here, it seems that again this is not a particularly useful approach.

AVDZ is already a rather small basis set for studies of noncovalent interactions, but numerous previous studies show that still smaller basis sets can work well if carefully chosen. This is usually assigned to a cancellation of errors between method and basis set. In particular, basis sets belonging to the 6-31G family can perform well if augmented by more diffuse polarization functions than is standard.[44,48] Most such conclusions have been based on MP2 studies, with few checking whether they hold for methods beyond MP2. Here, we have computed approximate CCSD(T)/CBS binding energies for the S66 complexes using a selection of basis sets based on 6-31G and 6-311G for the CCSD(T) correction term: RMSE from counterpoise corrected data for all tested basis sets are shown in Figure 1. As expected, the smallest basis sets without polarization functions perform rather poorly, while conventional basis sets such as 6-31G* and 6-31+G** show improvements to yield RMSE of less than 0.3 kcal/mol. The extra flexibility of the 6-311G family reduces RMSE, especially if augmented with diffuse functions, as in 6-311+G**.

Beyond these standard basis sets, one nonstandard one stands out as offering excellent results for relatively little computational expense, namely, 6-31G*(0.25). This consists of the standard 6-31G* with a smaller exponent (0.25) of the polarization function on heavy atoms and has previously been noted to work well with MP2,[48] but its ability to reproduce CCSD(T) data from much larger basis sets to within 0.15 kcal/mol is striking. Table 4 shows that the CCSD(T)/6-

Table 4. RMSE and %RMSE for Methods with the 6-31G*(0.25) Basis Set (kcal/mol and %)

	RMSE (%)	H-bond	mixed	dispersion
CCSD(T)	0.15 (3.9)	0.10 (1.6)	0.21 (5.9)	0.09 (2.4)
LMP2.5	0.26 (8.8)	0.07 (1.1)	0.21 (6.1)	0.40 (13.8)
LMP2.X[a]	0.13 (4.2)	0.09 (1.5)	0.11 (3.2)	0.18 (6.3)
L-CBS(T0)	0.18 (4.3)	0.20 (2.6)	0.11 (3.0)	0.20 (6.2)

[a]$X = 0.73$.

31G*(0.25) method gives a well balanced description of the interaction types in the S66 set and produces surprisingly small errors for the dispersion interactions (RMSE = 0.09 kcal/mol), which is often the most problematic category. This method has a tendency to overbind $\pi-\pi$ interactions and gives its largest errors for this interaction subcategory, with an RMSE value of 0.30 kcal/mol. However, excellent performance for aliphatic–aliphatic and π–aliphatic complexes mean that the overall RMSE for dispersion bound complexes is small and that the largest errors occur for the mixed category.

The excellent performance of this small basis set prompted us to examine how it fared when used with the most promising of the local approaches discussed above. Table 4 contains data for LMP2.5 and LMP2.X evaluated in this manner and shows that the former is marginally worse than AVDZ. This stems mainly from degradation of the description of dispersion-bound complexes. Optimizing the magnitude of the correction results in a slightly higher value than AVDZ (0.73 cf. 0.61), and significantly improves performance. The overall RMSE falls to 0.13 kcal/mol, and performance for the three main classes of

complex is well balanced. Within these classes, single and cyclic H-bonds are treated similarly (RMSE = 0.10 and 0.06 kcal/mol), as are $\pi-\pi$, aliphatic–π, and aliphatic–aliphatic complexes (0.19, 0.16, and 0.19 kcal/mol). The maximum and minimum errors are +0.23 and −0.35 kcal/mol for stacked pyridine dimer and stacked uracil dimer, respectively, and MSE = −0.04 kcal/mol. As well as these MP3-based estimates, we also evaluated the LCCSD(T0) − LMP2 correction with this small basis set, which, as shown in Table 4, gives marginally better performance than AVDZ. Using the smaller basis set improves performance for cyclic hydrogen bonds (0.26 vs 0.47 kcal with AVDZ) but degrades that for $\pi-\pi$ complexes (0.28 vs 0.17 kcal/mol). The maximum and minimum errors with this approach are +0.35 and −0.40 kcal/mol, for the acetic acid dimer and stacked pyridine...uracil, respectively, and MSE = 0.05 kcal/mol.

As with SCS-S66-CCSD-F12a, splitting the S66 data into training and test sets allows us to examine the robustness of the optimal parameters and of possible predictions of new data. With the 6-31G*(0.25) basis set, values of X vary from 0.70 to 0.76 when optimized against training sets of 33 complexes, and predicted RMSE values vary from 0.14 to 0.15 kcal/mol. With the AVDZ basis set, variations are a little larger, with X values of 0.57 and 0.65 giving predicted RMSE of 0.18 and 0.20 kcal/mol, respectively (see the Supporting Information for full details). Both cases lend confidence that the methods developed here are reasonably robust and not overfitted to the current data set. Application to larger systems is underway and will be reported in future publications.

■ CONCLUSIONS

We have shown that it is possible to generate accurate estimates of the interaction energy of diverse noncovalent interactions using relatively small basis sets and a variety of *ab initio* methods that go beyond the MP2 approximation. The most theoretically rigorous of the methods explored here is the explicitly correlated CCSD-F12 approach, and related methods. For three systems chosen as representatives of hydrogen bonded, dispersion-bound, and "mixed" classes of interaction, we show that interaction energies converge smoothly with both VnZ-F12 and AVnZ basis sets to a complete basis set limit estimated using a Schwenke-like extrapolation. CCSD-F12 interaction energies found in this manner are far from benchmark values, but CCSD(T)-F12 values converge to essentially identical values to those found from ΔCCSD(T) correction of MP2/CBS interaction energy. These studies also demonstrate that the relatively small AVDZ basis set can produce interaction energies approximately 0.1 kcal/mol from the best estimates, especially with the scaled CCSD(T*)-F12a/AVDZ method.

Prompted by these results, we examined the performance of less accurate but more computationally accessible methods with the AVDZ basis set. A local version of the MP2.5 approach, which interpolates between LMP2 and LMP3, is shown to give good performance across the whole of the S66 set of complexes. Errors are further reduced, albeit slightly, if the contribution of the MP3 correction is optimized for the current set of data. Local correlation methods were also used with AVDZ to generate an approximate ΔCCSD(T) correction, which results in similar errors to LMP2.X without use of any scaling parameters. Still better performance stems from spin-component scaling of CCSD-F12a/AVDZ energies; three variants of this approach were tested, including scaling

parameters originally proposed by Takatani et al., another set developed by Pitoňak et al. specifically for noncovalent interactions (SCS-MI), and a third optimized specifically for this method and basis set. All three perform very well, yielding the smallest errors of any method considered here, and similar performance for all classes of complex.

Moving to even smaller basis sets than AVDZ, we showed that a ΔCCSD(T) correction evaluated with 6-31G*(0.25) results in low overall errors and a balanced description of all classes of complex. This basis set also performs well when used in the LMP2.X approach, although a larger contribution of MP3 correction is required than with AVDZ. Similarly, a LCCSD(T0) − LMP2 correction evaluated with 6-31G*(0.25) shows promise. The very small size of this basis set makes it an attractive option for study of large systems.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Results from CCSD-F12b and related methods, details of all basis sets considered for CCSD(T) correction, and training/test data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44-2920-874950. Fax: +44-2920-874030. E-mail: platts@cardiff.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) (a) Brammer, L. *Chem. Soc. Rev.* **2004**, *33*, 476−489. (b) Hobza, P.; Zahradnik, R. *Chem. Rev.* **1988**, *88*, 871−897. (c) Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157−167.

(2) Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023−5063.

(3) Riley, K. E.; Platts, J. A.; Rezac, J.; Hobza, P.; Hill, J. G. *J. Phys. Chem. A* **2012**, *116*, 4159−4169.

(4) Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257−8263.

(5) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095−9102.

(6) Distasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105*, 1073−1083.

(7) Koch, H.; Fernandez, B.; Christianson, O. *J. Chem. Phys.* **1998**, *108*, 2784−2790.

(8) Hobza, P.; Sponer, J.; Reschel, T. *J. Comput. Chem.* **1995**, *16*, 1315−1325.

(9) Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.* **2002**, *124*, 10887−10893.

(10) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104−112.

(11) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *120*, 11586−11599.

(12) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243−252.

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

(14) Pitonak, M.; Neogrady, P.; Cerny, J.; Grimme, S.; Hobza, P. *ChemPhysChem* **2009**, *10*, 282−289.

(15) Riley, K. E.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2011**, *13*, 21121−21125.

(16) (a) Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286−6297. (b) Pulay, P. *Chem. Phys. Lett.* **1983**, *100*, 151−154. (c) Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213−236.

(17) (a) Tew, D.; Hättig, C.; Bachorz, R.; Klopper, W. Efficient explicitly correlated coupled-cluster approximations. In *Recent Progress in Coupled Cluster Methods: Theory and Applications*; Čársky, P., Paldus, J., Pittner, J., Eds.; Springer: Berlin, 2010; pp 535. (b) Werner, H.-J.; Adler, T. B.; Knizia, G.; Manby, F. R. Efficient explicitly correlated coupled-cluster approximations. In *Recent Progress in Coupled Cluster Methods: Theory and Applications*; Čársky, P., Paldus, J., Pittner, J., Eds.; Springer: Berlin, 2010; pp 573.

(18) Rezac, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.

(19) Rezac, J.; Riley, K. E.; Hobza, J. *J. Chem. Theory Comput.* **2011**, *7*, 3466−3470.

(20) Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Rezac, T.; Pitonak, M.; Vondrasek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73*, 1261−1270.

(21) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.

(22) Polly, R.; Werner, H.-J.; Manby, F. R.; Knowles, P. J. *Mol. Phys.* **2004**, *102*, 2311−2321.

(23) (a) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106. (b) Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.

(24) Ten-no, S. *Chem. Phys. Lett.* **2004**, *398*, 56−61.

(25) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.

(26) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.

(27) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190−195.

(28) (a) *Molpro*, version 2010.1, a package of ab initio programs; Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. et al. www.molpro.net (accessed Nov. 9, 2012). (b) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. *WIREs Comput. Mol. Sci.* **2012**, *2*, 242−253.

(29) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285−4291.

(30) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175−3183.

(31) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *131*, 194105.

(32) Marchetti, O.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3400−3409.

(33) Schwenke, D. W. *J. Chem. Phys.* **2005**, *122*, 014107.

(34) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.

(35) Marchetti, O.; Werner, H.-J. *J. Phys. Chem. A* **2009**, *113*, 11580−11585.

(36) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916−4926.

(37) Boughton, J. W.; Pulay, P. *J. Comput. Chem.* **1993**, *14*, 736−740.

(38) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735−746.

(39) Werner, H.-J.; Schütz, M. *J. Chem. Phys.* **2011**, *135*, 144116.

(40) Werner, H.-J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys.* **2003**, *118*, 8149−8160.

(41) McMahon, J. D.; Lane, J. R. *J. Chem. Phys.* **2011**, *135*, 154309.

(42) Werner, H.-J.; Knizia, G.; Adler, T. B.; Marchetti, O. *Z. Phys. Chem.* **2010**, *224*, 493−511.

(43) Marshall, M. S.; Sherrill, C. D. *J. Chem. Theor. Comput.* **2011**, *7*, 3978−3982.

(44) Marshall, M. S.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *135*, 194102.

(45) Pitoňák, M.; Řezáč, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9611−9614.

(46) Takatani, T.; Hohenstein, E. G.; Sherrill, C. D. *J. Chem. Phys.* **2008**, *128*, 124111.

(47) Korona, T.; Kats, D.; Schütz, M.; Adler, T. B.; Liu, Y.; Werner, H.-J. Local approximations for an efficient and accurate treatment of electron correlation and electron excitations in molecules. In *Linear-Scaling Techniques in Computational Chemistry and Physics*; Zaleśny, R., Papadopoulos, M. G., Mezey, P. G., Leszczynski, J., Eds.; Springer: Berlin, 2011; pp 345.

(48) (a) Hobza, P.; Sponer, J. *Chem. Rev.* **1999**, *99*, 3247−3276. (b) Svozil, D.; Hobza, P.; Sponer, J. *J. Phys. Chem. B* **2010**, *114*, 1191−1203.