Spin-component scaled coupled-clusters singles and doubles optimized towards calculation of noncovalent interactions

Michal Pitoňák,[ab] Jan Řezáč[a] and Pavel Hobza*[ac]

Received 8th April 2010, Accepted 29th June 2010
DOI: 10.1039/c0cp00158a

The same- and opposite-spin scaling parameters for the SCS-CCSD method were reparametrized on the basis of benchmark CCSD(T)/CBS set interaction energies from the S22 set. New parameters were close to the original ones but swap between the different spin components, being 1.11 for the opposite- and 1.28 for the same-spin component. The RMSD, and especially, the largest error for the S22 were significantly reduced in comparison with the original parametrization. These statistical factors were only slightly worse when the S22x5 test set, containing not only the equilibrium but also the non-equilibrium geometries, was used. This new method, named the SCS(MI)-CCSD ("MI" stands for "Molecular Interactions") can thus be recommended for highly accurate calculations of interaction energies of various noncovalent interaction types, for which the CCSD(T)/CBS calculations are impractical.

An accurate description of noncovalent interactions is in the literature repeatedly referred to as "the great challenge" for both theory and experiment. From the point of view of quantum chemistry, benchmark quality results for a broad range of noncovalent interaction types can be obtained first on a sophisticated level of theory as coupled-clusters including iterative treatment of single and double excitations with perturbative correction for the triple excitations (CCSD(T)). Steep, O(N^7) (N being the number of AO basis functions), scaling of the CCSD(T) is prohibitive for calculations of complexes with (roughly) more than several tenths of "heavy" (second-row) atoms and hydrogen close to the complete basis set limit (CBS), despite the massive progress in both the CCSD(T) algorithm implementations and the parallel computer architecture capabilities.

The need for an accurate, yet fast method for calculation of noncovalent complexes is strongly motivated by the almost unlimited potential for interesting applications of these methods, such as the calculation of biosystems, nanomaterials, *etc.*, out of the reach of present computational methods. This "need" manifests itself in the large number of new approximate wave-function theory (WFT) and density-functional theory (DFT) methods that have emerged during the past few years. The main direction in development of these methods is focused on the proper description of the dispersion energy, which is typically either completely neglected (as in the Hartree–Fock method or in the most of the DFT methods), or very inaccurate (as in most of the WFT methods, for instance, strongly over-estimated on the MP2 level, or underestimated on MP3 or CCSD levels). The routes which are followed to reach the goal are quite diverse. In the WFT, the most successful attempts employ scaling of spin components of the correlation energy (SCS-MP2,[1] SCSN-MP2,[2] SCS(MI)-MP2,[3] SCS-CCSD),[4] scaling of the higher-order correlation contributions[5] or by use of an "external" source of more accurate dispersion energy itself.[6–8]

In this paper, we would like to report a reparametrization of one of the most promising and the most accurate methods from those mentioned above, the coupled-cluster singles and doubles method with scaled same- and opposite-spin components (SCS-CCSD), developed in the group of Sherrill.[4] In the original work of Takatani *et al.*[4] authors optimized the spin scaling parameters (opposite-spin, c_{os} and same-spin, c_{ss}) to a database of 48 reaction energies, and validated these parameters on atomization energies of several molecules and potential energy curves of benzene and methane dimers. Despite the fact that the parametrization was carried out on reaction energies, the accuracy of noncovalent complexes' interaction energies and geometries, which is our main interest, was excellent. Deviations of the optimal monomer displacements for these two complexes compared to the reference CCSD(T) values were below the displayed accuracy (0.1 Å), while the deviations of the interaction energies were below 0.05 kcal mol^{-1}, *i.e.* less than 1%. Further testing of the SCS-CCSD method on the S22 data set of noncovalent interactions[9] revealed that the accuracy of the SCS-CCSD is still not satisfactory, which motivated us to reinvestigate the optimal nature of the Sherrill's scaling parameters for these applications. Similarly, the accuracy of the original Grimme's SCS-MP2 was successfully improved for calculations of noncovalent interactions by reparametrization (*i.e.* SCS(MI)-MP2) by Distasio and Head-Gordon.[3]

The goal of this paper is to optimize the scaling parameters of the same- and opposite-spin components of the CCSD correlation energy for the set of 22 estimated CCSD(T)/CBS interaction energies of hydrogen-bonded, dispersion dominated and mixed-character noncovalent (closed-shell) complexes of the S22 set. To validate these parameters on the larger data set, we chose the new "S22x5" database[10] consisting of the same species as the S22, both in equilibrium and distorted (one shortened and four elongated intermolecular) geometries, briefly described in the text below. There is a good reason to go beyond the equilibrium distance. It is a fact that moving to the extended complexes, the number of middle- to long-range contacts grows dramatically and even small errors, by

[a] *Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i. and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, 166 10 Praha 6, Czech Republic. E-mail: pavel.hobza@uochb.cas.cz; Fax: (+420) 220 410 320; Tel: (+420) 220 410 311*
[b] *Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava 4, Slovak Republic*
[c] *Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic*

themselves insignificant, at these distances can generate a rather large accumulated error in the total stabilization energy. It should be stressed that the mere fact that a parametrized method works well for the equilibrium distances does not also automatically guarantee its good performance in the middle- or long-range distances. Further validation is done on a set of 24 amino acid side-chain interactions of Berka *et al.*[11] as well as on a collection of databases of noncovalent interactions of Truhlar and coworkers.[12]

Benchmark MP2/CBS and CCSD(T)/CBS values for the S22 set were taken from the ref. 9. Estimated SCS(MI)-CCSD/CBS energies, were calculated analogously to the (estimated) CCSD(T)/CBS energies in the S22 paper, *i.e.* as MP2/CBS augmented with the ΔSCS(MI)-CCSD correction term, $c_{os}CCSD_{os}^{corr} + c_{ss}CCSD_{ss}^{corr.} - MP2^{corr.}$, calculated in small or medium-size basis set (depending on the system size), *e.g.* cc-pVXZ where X = D, T, Q, aug-cc-pVDZ, or, cc-pVTZ-fd, *i.e.* with removed one set of less diffuse f- and d-functions). The only exception was the stacked complex of indole···benzene, for which we have recalculated the CCSD(T) correction in the aug-cc-pVDZ basis set, since the original value calculated in cc-pVDZ was found to be strongly unsaturated.[8,13] Despite the availability of a revised version of the S22 database, so called S22A, published by Takatani *et al.*,[13] we used the CCSD(T)/CBS benchmark data, obtained as described above, for several reasons. Firstly, recalculation of the 'problematic' complex of indole···benzene significantly increased the accuracy of S22 towards that of the S22A. Secondly, it is the availability of valuable S22x5 validation data calculated using the same methodology as the S22. At last, large-scale SCS(MI)-CCSD calculations will most likely be done in rather small basis sets, thus parametrization on higher basis set quality benchmark data would introduce basis set effects to the parametrization, which is not desired. This is related to our motivation, not to use plain, but estimated CBS SCS(MI)-CCSD energies in the optimization, which was to alleviate the basis set effects on the optimal value of spin scaling factors. When SCS(MI)-CCSD/CBS energies are used, the basis set effects are to large extent absorbed in the computationally feasible MP2/CBS. Thus more space is left for the optimization for better incorporation of the higher order correlation effect into the optimized parameters. Having a different scaling coefficients for different basis sets, as used in the SCS(MI)-MP2, is also not too practical in this case, since the CCSD is still rather expensive method and the application to larger complexes will most probably be limited anyway only to aug-cc-pVDZ basis set, or similar, as already mentioned. Scaling parameters were determined in order to minimize Root Mean Square Deviation (RMSD) of the SCS(MI)-CCSD/CBS interaction energies with respect to the CCSD(T)/CBS ones.

Validation set of distorted geometries of complexes from the S22 test set was taken from the work of Gráfová *et al.*[10] For each complex, four additional geometries were prepared (the fifth one was the equilibrium structure from the original S22) by selecting the ''noncovalent bonding coordinate'' and shortening it to 90% of its equilibrium value, R_e, and elongating it to 120, 150 and 200% of the R_e. The selection of the noncovalent bonding coordinate is obviously not well defined, for H-bonded complexes we chose direction roughly corresponding to the hydrogen bond, in stacked complexes it was the perpendicular direction with respect to the stacked ring planes, *etc.* As the character of interaction may vary upon complex dissociation (*e.g.* dispersion dominated complex may turn to electrostatics dominated, if monomer(s) posses permanent electric moment), thus this larger test set should, in our opinion, map all sorts of interaction types fairly well. Validation with respect to the 'Representative Amino Acid Side-Chain Interactions in Proteins' (SCAI) data set of Berka *et al.*,[11] consisting of 24 amimo acid side-chain interaction energies, calculated on the estimated CCSD(T)/CBS level (MP2/CBS from two-point extrapolation from aug-cc-pVDZ and aug-cc-pVTZ basis sets, ΔCCSD(T) correction calculated in the 6-31G**(0.25, 0.15) basis sets)[9] was based on the estimated SCS(MI)-CCSD/CBS interaction energies obtained following the same methodology as used for the CCSD(T)/CBS. Interaction energies of 28 noncovalent complexes from HB6/04, CT7/04, DI6/04 and WI9/04 databases of Truhlar and coworkers[12] were recalculated on the CCSD(T)/aug-cc-pVTZ level for the purpose of this work, to serve as the additional validation data.

Resulting spin scaling parameters from minimizing of the RMSD of SCS(MI)-CCSD/CBS with respect to the CCSD(T)/CBS were 1.11 and 1.28, for opposite-spin and the same-spin scaling parameter. Surprisingly, these values are very close to those of Takatani *et al.*,[3] but swapped with respect to the correlation energy spin components, *i.e.* 1.27 and 1.13 for c_{os} and c_{ss}. Perhaps there is an analogy with the optimal spin scaling parameters swapped from the original SCS-MP2 (1.2 and 0.33, for c_{os} and c_{ss}) to SCS(MI)-MP2 (0.4 and 1.29, for c_{os} and c_{ss}). In addition to the SCS(MI)-CCSD parametrization, we carried out an analogous, but constrained RMSD minimization to obtain the spin scaling parameters, such that $c_{os} + c_{ss} = 2$. The resulting method, named SCS(AC)-CCSD, with the optimized opposite-spin and the same-spin scaling parameters of 0.75 and 1.25, should thus have proper asymptotic scaling, as discussed in ref. 5 and 14, and perform well for the description of noncovalent interactions. The surface of the RMSD of the SCS-CCSD in c_{os} and c_{ss} coordinates for the S22 interaction energies is shown in Fig. 1.

Comparing the performance of SCS-CCSD, SCS(MI)-CCSD and SCS(AC)-CCSD with standard CCSD, see Tables 1–4, in all cases the errors are significantly decreased upon scaling of the correlation energy spin components. Statistical data (RMSD; MUE-Mean Unsigned Error; MSE-Mean Signed Error; MAX-Maximum Unsigned Error) for the S22 test are shown in Table 1. It is obviously not legitimate to compare the performance of the SCS(MI)-CCSD with other methods on its training set, but still some observations can be drawn. A decrease of the maximum error from 1.86 kcal mol^{-1} (indole···benzene stack complex) of CCSD to 0.79 kcal mol^{-1} of SCS-CCSD (for hydrogen-bonded uracil dimer) and finally to only 0.1 kcal mol^{-1} of SCS(MI)-CCSD (for benzene···methane complex) is a step in the right direction. It is worth noting that even the second largest error of the SCS-CCSD for the S22 test set was obtained for the hydrogen-bonded complex, namely the Formic acid dimer (0.75 kcal mol^{-1}). A negative value of MSE for SCS-CCSD indicates certain systematic underestimation of interaction energies by this method, while an

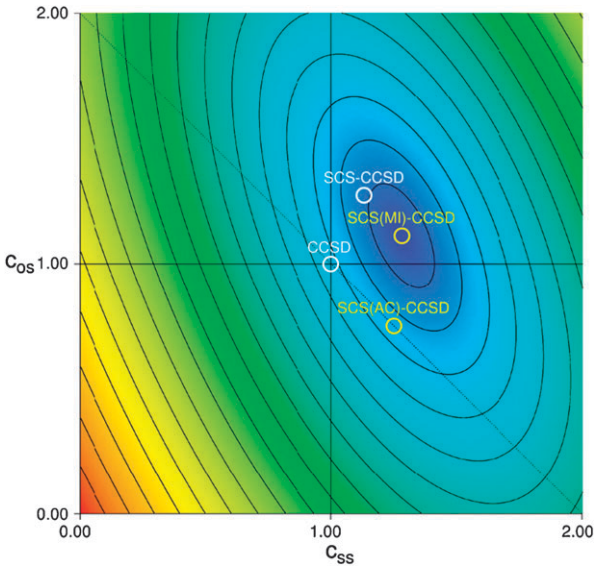

Fig. 1 RMSD surface of the SCS-CCSD interaction energies for the S22 data set in spin-component scaling coefficient coordinates.

Table 1 Selected statistical data for CCSD and SCS variants of the CCSD method for the S22 database. All data in kcal mol^{-1}

	CCSD	SCS(AC)-CCSD	SCS-CCSD	SCS(MI)-CCSD
RMSD	0.82	0.62	0.33	0.06
MUE	0.66	0.43	0.21	0.05
MSE	−0.66	−0.27	−0.20	0.02
MAX	1.86	1.74	0.79	0.10

MSE value of the SCS(MI)-CCSD that is close to zero indicates a more balanced treatment of different interaction types. Performance of the SCS(AC)-CCSD is clearly inferior to that of both the SCS-CCSD and SCS(MI)-CCSD, attributable to the strong constraints applied on the spin-component scaling coefficients in the optimization.

Validation of the new scaling coefficients in the SCS(MI)-CCSD is shown in Tables 2–4. For the S22x5 (Table 2), the RMSD value is practically unchanged compared to the value from the S22 set, still being unprecedentedly small, 0.05 kcal mol^{-1}. Differences in the maximum absolute error from CCSD to SCS(AC)-CCSD, SCS-CCSD and SCS(MI)-CCSD are more pronounced, compared to the S22. The maximum error of the SCS(MI)-CCSD is an order of magnitude lower than that of the CCSD method, and almost 1 kcal mol^{-1} lower than that of SCS-CCSD. Judging by the same MSE value as for the S22, SCS(MI)-CCSD keeps a balanced description of all involved interaction types along the "dissociation" path. Results obtained for the SCAI database (Table 3) and the databases

Table 2 Selected statistical data for CCSD and SCS-variants of the CCSD method for the S22x5 database. All data in kcal mol^{-1}

	CCSD	SCS(AC)-CCSD	SCS-CCSD	SCS(MI)-CCSD
RMSD	0.69	0.53	0.26	0.05
MUE	0.41	0.28	0.14	0.03
MSE	−0.41	−0.16	−0.13	0.02
MAX	2.80	2.64	1.05	0.16

Table 3 Selected statistical data for CCSD and SCS-variants of the CCSD method for the SCAI database. All data in kcal mol^{-1}

	CCSD	SCS(AC)-CCSD	SCS-CCSD	SCS(MI)-CCSD
RMSD	0.52	0.30	0.24	0.10
MUE	0.45	0.24	0.16	0.09
MSE	−0.45	−0.22	−0.07	0.06
MAX	0.99	0.75	0.55	0.18

of Truhlar and coworkers (Table 4) are more balanced, which is most likely due to the missing strong π–π stacked complexes present in the S22 and S22x5. SCS(MI)-CCSD is in both cases clearly the most accurate method. Surprisingly, the SCS(AC)-CCSD performs only slightly worse compared to the SCS-CCSD for the SCAI database and even outperforms the SCS-CCSD method for the collection of Truhlar databases.

The RMSD value of the SCS(MI)-CCSD method for the S22 set was significantly reduced when compared with the original SCS-CCSD parametrization, 0.06 *vs.* 0.33 kcal mol^{-1}. New c_{os} and c_{ss} parameters, 1.11 and 1.28, were close to the original ones, but in swapped order, *i.e.* 1.27, 1.13. When the SCS(MI)-CCSD method was tested for the S22x5 set of equilibrium and stretched geometries, it provided excellent performance, the RMSD value was practically unchanged, 0.06 *vs.* 0.05 kcal mol^{-1}. The high accuracy of the SCS(MI)-CCSD is also retained for the set of interaction energies of representative amino acid side-chain dimers, RMDS of 0.10 kcal mol^{-1}, as well as for the set of databases of nonocovalent interactions of Truhlar and coworkers, RMSD of 0.05 kcal mol^{-1}. This finding indicates that the nature of noncovalent interactions is correctly reflected in the SCS(MI)-CCSD method and that the method is able to provide accurate stabilization energies also at various geometry distortions. This is not so self-evident, several DFT-D procedures that perform fairly well for the equilibrium distances fail at stretched geometries.[15] Being roughly one order of magnitude more computationally efficient than the CCSD(T) method, the SCS(MI)-CCSD can be thus recommended as the method of choice for large complexes, where it is believed to be capable of providing comparably accurate interaction energies.

Suggestions related to the validation data sets, from one of the reviewers of this work, are highly appreciated. This work was a part of the research project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, and was supported by Grants No. LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to PH in 2007 is also acknowledged.

Table 4 Selected statistical data for CCSD and SCS-variants of the CCSD method for the HB6/04, CT7/04, DI6/04 and WI9/04 databases. All data in kcal mol^{-1}

	CCSD	SCS(AC)-CCSD	SCS-CCSD	SCS(MI)-CCSD
RMSD	0.45	0.19	0.25	0.05
MUE	0.35	0.15	0.15	0.03
MSE	−0.35	−0.12	−0.15	−0.01
MAX	1.24	0.41	0.80	0.21

This work was also supported by the Slovak Grant Agency VEGA under Contracts No. 1/0428/09 and No. 1/0520/10 and by the Korea Science and Engineering Foundation (World Class Univ. program: R32-2008-000-10180-0). An important part of the calculations were performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

References

1 S. Grimme, *J. Chem. Phys.*, 2003, **118**, 9095.
2 J. G. Hill and J. A. Platts, *J. Chem. Theory Comput.*, 2007, **3**, 80.
3 R. A. DiStasio and M. Head-Gordon, *Mol. Phys.*, 2007, **105**, 1073.
4 T. Takatani, E. G. Hohenstein and C. D. Sherrill, *J. Chem. Phys.*, 2008, **128**, 124111.
5 M. Pitoňák, P. Neogrády, J. Černý, S. Grimme and P. Hobza, *ChemPhysChem*, 2009, **10**, 282.
6 S. M. Cybulski and M. L. Lytle, *J. Chem. Phys.*, 2007, **127**, 141102.
7 A. Heßelmann, *J. Chem. Phys.*, 2008, **128**, 144112.
8 M. Pitoňák and A. Heßelmann, *J. Chem. Theory Comput.*, 2010, **6**, 168.
9 P. Jurečka, J. Šponer, J. Černý and P. Hobza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1985.
10 L. Gráfová, M. Pitoňák, J. Řezáč and P. Hobza, *J. Chem. Theory Comput.*, 2000, DOI: 10.1021/ct1002253.
11 K. Berka, R. Laskowski, K. E. Riley, P. Hobza and J. Vondrášek, *J. Chem. Theory Comput.*, 2009, **5**, 982.
12 Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2005, **1**, 415.
13 T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall and C. D. Sherrill, *J. Chem. Phys.*, 2010, **132**, 144104.
14 R. C. Lochan, Y. Jung and M. Head-Gordon, *J. Phys. Chem. A*, 2005, **109**, 7598.
15 K. E. Riley, M. Pitoňák, J. Černý and P. Hobza, *J. Chem. Theory Comput.*, 2010, **6**, 66.

