JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

# Off-Center Gaussian Functions, an Alternative Atomic Orbital Basis Set for Accurate Noncovalent Interaction Calculations of Large Systems

Miroslav Melicherčík,[†] Michal Pitoňák,*[,‡,§] Vladimír Kellö,[‡] Pavel Hobza,[⊥,∥] and Pavel Neogrády*[,‡]

[†]Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 974 01 Banská Bystrica, Slovakia

[‡]Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University, Mlynská Dolina, 842 15 Bratislava, Slovakia

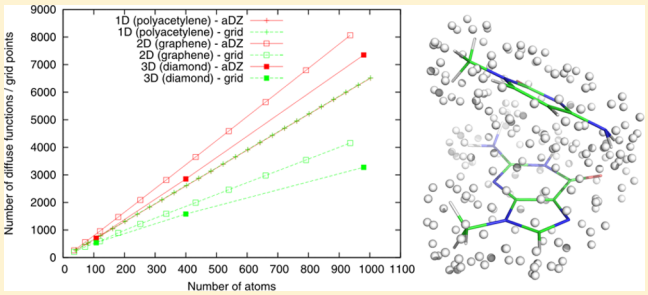
[§]Computing Center of the Slovak Academy of Sciences, Dúbravská cesta č. 9, 845 35 Bratislava, Slovakia

[⊥]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, v. v. i., Flemingovo nám. 2, 166 10 Praha 6, Czech Republic

[∥]Department of Physical Chemistry, Palacký University, 771 46 Olomouc, Czech Republic

Supporting Information

**ABSTRACT:** Proper description of noncovalent interactions requires, among other things, the use of diffuse atomic orbital (AO) basis sets. However, the presence of diffuse functions, especially in extended molecular systems, can lead to linear dependent AO basis sets. This in turn results, for example, in molecular orbital optimization problems or, when dependencies are removed in unpredictable and possibly geometry-dependent accuracy fluctuations. In this work, an alternative approach is proposed which suffers no linear dependence problems and delivers comparably accurate noncovalent interaction energies. An algorithm is proposed and implemented to construct a grid of off-center s-type Gaussian functions surrounding the molecule; substituting the presence of atom-centered diffuse basis functions. While the number of basis functions in the grid is comparable to the number of diffuse basis functions in aug-cc-pVXZ (for each cardinality number "X") basis sets for small molecular systems, the ratio becomes more favorable with increasing system size. The grid is constructed in a way that it is unique for a molecule (monomer) and, thus, independent of noncovalent complex/cluster geometry. The grid parameters, such as the density of grid points and s-function exponents, are obtained via optimization toward the S22 data set on the MP2 level. The quality, transferability, and versatility of the grid is tested on the S66 data set as well as on several cuts through the potential energy surface for noncovalent complexes, such as methyl-guanine···methyl-cytosine conversion from stacked to hydrogen-bonded structure.




## 1. INTRODUCTION

Accurate theoretical description of noncovalent interactions is currently recognized as crucially important in many fields of science, including those focused on the structure and function of biomolecules, drug design, and catalysis.[1] This presents a challenging task for theoretical chemistry, as often reported in many scientific works and clearly demonstrated in intense research activity in this field. Despite noteworthy progress in the development of density functional theory (DFT) methods suited for description of noncovalent interactions, reliable and "benchmark-quality" methods of choice are still bound to wave function theory (WFT).[2] Particularly, the coupled clusters methods with iterative treatment of single and double excitations corrected with the perturbative inclusion of connected triple excitations (CCSD(T)) is recognized as the best trade-off between computational demands and accuracy toward the full configuration interaction (FCI) limit; often referred to as the "gold standard". This is no doubt true for most single-reference noncovalent complexes in ground state, but an important condition for accuracy in completeness of the atomic orbital (AO) basis set expansion has been concealed yet.

In the early days of ab initio noncovalent interaction calculations, the so-called basis set superposition error (BSSE), as one of the artifacts resulting from an incomplete AO basis set expansion was identified and several ways of alleviating its impact on accuracy were proposed.[3] The impact of BSSE is almost impossible to separate from other companion errors arising from insufficient basis set saturation. These cover proper description of relevant monomer electric properties,

**Received:** August 5, 2013
**Published:** November 5, 2013

   dx.doi.org/10.1021/ct400692b | J. Chem. Theory Comput. 2013, 9, 5296−5304

including dipole moment and polarizabilities. The conceptually simplest, yet not necessarily the most computationally efficient solution to this complex problem, is to operate close to the complete basis set (CBS) limit; either by using extended basis sets or exploiting some basis set extrapolation schemes. Currently, this approach is often pursued for molecular systems with a size of approximately tens of atoms due to the availability of high-performance computer hardware and efficient software. Moreover, the computational costs of this approach are efficiently alleviated using so-called focal point analysis[4,5] combined with the hierarchy of many-body perturbation theory (MBPT) and/or CC energies calculated in series of AO basis sets of different size. Nevertheless, at certain stages, all approaches applied to noncovalent interaction investigations require energies to be calculated in a fairly large, preferably diffuse, AO basis set.

The popularity of the focal-point analysis is also facilitated by the knowledge that large basis set MP2 calculations used in routine applications no longer produce a bottleneck when "small basis set" post-MP2 calculations follow. Nevertheless, as the size of the molecular systems of interest increases, various approximations to computational AO basis sets are being introduced. These include the so-called heavy-augmented basis sets advocated by Sherrill and co-workers,[6,7] "midbond" functions,[8,9] and the use of diffuse basis functions on only a subset of atoms in the investigated molecular system.[10] The use of these approximations essentially proceeds from a twofold motivation. First, the use of uniform, large diffuse basis sets does not deliver the expected improvements on nondiffuse ones, and second, the presence of diffuse basis functions on all atoms in the molecule can cause convergence problems or other issues resulting from numerical instability.

The primary goal of this work is to address the second motivation; problems caused by the linear dependence of basis functions. In addition, the more economic basis set expansion of the first motivation is implicitly addressed, as a byproduct of this proposed methodology; see section 3. A computational setup which eventually leads to a linear dependent AO basis set is quite common and often occurs in routine calculations. This is observed where a small basis set with diffuse functions or a medium size basis set without diffuse functions is used for extended, compact molecular systems such as polycyclic aromatic hydrocarbons[10,11] or compounds where atoms or molecular fragments are drawn closer than their equilibrium distances.

Although it is quite straightforward to control the linear dependence of an AO basis set by elimination of problematic basis functions (identified by AO overlap matrix eigenvalues larger than a predefined threshold), insufficient accent has attended the impact of this elimination on the resulting energy and the properties derived from it. We are not aware of a method capable of estimating these errors, especially for interaction energy as a function of the linear dependent AO function elimination threshold. Furthermore, the overlap of AO basis functions changes in mutual rotation or displacement of intramolecular fragments, thus potentially causing the error resulting from elimination of the linear dependent basis function to be geometry-dependent. Although a similar problem can arise in intersystem AO basis function overlap, we are not aware of a particular case where this eventuates. In addition, the presence of diffuse basis functions often causes problems in orbital localization which must be properly converged. Local correlation treatments provide an example where this is necessary.

The computational setup presented in this work is inspired by two widely accepted concepts. The first of these is midbond or floating functions;[8,9] where off-center Gaussian functions spanning the space between interacting species were found to profoundly increase convergence toward the CBS limit in noncovalent interaction energy calculations. This approach is not widely used in routine applications, because its drawbacks include ambiguity in the midbond function position, dependence on super-system geometry, and an unpredictable convergence pattern toward the CBS limit. Floating Gaussian functions found use in much diverse range of studies, such as the investigations of the excess electron binding in water clusters,[12,13] and many others. The second concept is related to the so-called Gaussian-lobe functions.[14,15] The aim of this approach is to mimic the presence of higher-angular momentum basis functions in the AO basis set by properly placing a manifold of lower-angular momentum basis functions between atoms in the molecule. The approach presented in this work is based on the concept of a grid of off-center Gaussian s-functions, unique for a molecule (monomer, in noncovalent interaction calculations). This grid combined with the standard Dunnig-type cc-pVXZ[16] (XZ) basis sets should perform comparably well with the corresponding atom-centered basis set augmented with diffuse basis functions, aug-cc-pVXZ (aXZ). Two key features are expected from the basis set combined with a Gaussian s-function grid: (1) linear dependence of basis functions should not exceed that of the underlying XZ basis set and (2) the total number of basis functions do not considerably exceed that of aXZ basis set for the same cardinality "X". The second feature is clearly ambiguous, especially in regard to the size of the investigated molecular system. The results presented in section 3 show that using grids with DZ basis set it is difficult to achieve simultaneously comparable accuracy with aDZ and the total number of basis functions also; particularly in "small" molecular systems. However, the number of grid basis functions scales more favorably with the size of the molecule compared to the number of diffuse basis functions, as in aXZ basis sets. This aspect will be discussed further, but it is most important to reiterate here that the proposed methodology is primarily suited for large-scale applications of approximately fifty atoms or more, depending on molecular structure, where the second condition is more easily satisfied.

## 2. METHODOLOGY

Grid parametrization and the validation of its quality and versatility is limited to the MP2 level. Although it is clear that treatment at MP2 correlation level is substandard for benchmark quality description of noncovalent interactions, it aptly describes typical interaction motifs. As the AO basis set exponents and contractions for a particular atom are expected to be transferable across various chemical environments and theory levels, so are the grid parameters obtained in this work. Grid parameters transferability naturally has its limitations, as demonstrated in section 3. However, we are convinced that errors resulting therein are greater than those resulting from the level of correlation treatment applied in its parametrization. Needless to say, grid parameter optimization at post-MP2 level would be computationally impractical.

In this study, we deal exclusively with grids of basis functions added to the DZ basis set. Therefore, the goal of this grid optimization is to approach the accuracy and efficiency of

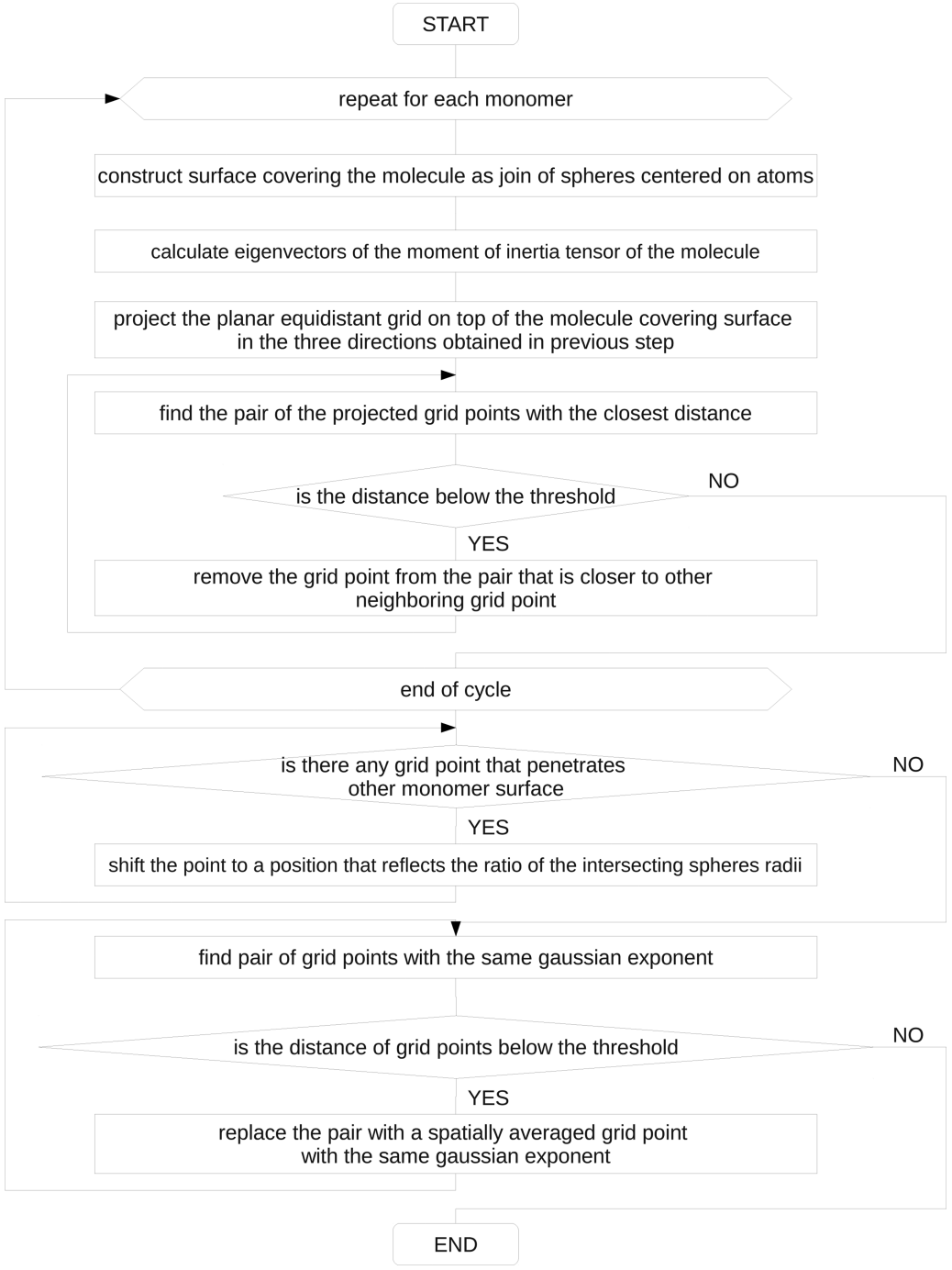



**Figure 1.** Flowchart of the grid generation algorithm.

description delivered by the aDZ basis set. This choice was made in light of the following considerations. The largest molecular systems feasibly treated at the CCSD(T) level are currently tens (and closely approximate a hundred) first and second-row atoms. Taking to account more rapid convergence of post-MP2 correlation contribution with the basis set size compare to MP2 itself, CCSD(T)/aDZ (or even smaller, CCSD(T)/6-31G*(0.25)) combined with MP2/CBS, delivers fairly accurate and reliable interaction energies for molecular systems of this size. However, the MP2/CBS is often difficult to obtain using diffuse basis sets at this size, so even having an alternative for aDZ basis set is an appealing choice at the MP2 level. There is no principal limitation in building the grid on top of any other basis set, such as cc-pVTZ which is currently in progress in our laboratory; however, this is beyond the scope of our paper.

**2.1. Grid Generation Algorithm.** We have proposed and implemented the grid generation algorithm to produce a grid of points specific for the particular geometry of a molecule (monomer). It is invariant to "initial" molecular orientation. The algorithm consists of three essential steps: construction of a surface surrounding the molecule, projection of a grid of points on the surface, and elimination of the grid points with mutual distance below a predefined threshold. The grid generation algorithm flowchart is depicted in Figure 1.

The first estimate of the surface surrounding a molecule is constructed by placing spheres centered on each atom in the molecule. The "exposed parts" of spheres, which are not penetrated by other spheres, are merged to create a coherent surface with a guaranteed minimum distance from atoms. The distance of the grid surface from atoms in a molecule varies for different atom types, and in the current implementation, this sphere radius is related to the van der Waals radius of the closest atom. The overall surface distance is then scaled by a single parameter obtained from grid parameter optimization (see further).

The grid must be invariant with respect to the initial orientation of the molecule, and therefore, it must be projected on the molecule's surrounding surface from unambiguously defined directions in three-dimensional space. Hence, the center of mass of the molecule and the moment of inertia tensor are calculated, thus providing eigenvectors. In the current approach, these eigenvectors define the directions from which the planar, equidistant square grid of points is projected on the surrounding molecular surface. In the actual implementation of the algorithm, the surface covering the molecule is never "physically" constructed. The intersection of spheres surrounding atoms is evaluated "on-the-fly" in the step of projection of the planar, equidistant grid, individually for each grid point.

The mutual distance of grid points can be arbitrarily set, and this forms another parameter in grid optimization. As expected, an inhomogeneous density of points on the molecule's surrounding surface is obtained by projection of an equidistant, planar square grid, and some points can appear "too close" to each other. Therefore, the pairs of grid points with mutual distance below a predefined threshold are identified, and the point closer to another grid point is removed. This procedure is repeated until no more pairs with mutual distance under the threshold are found. After this step, the grid generation for a molecule (monomer) is completed.

However, an additional step should be introduced into intermolecular interaction calculations. In situations where the interacting molecules are mutually too close, their individual grids may penetrate each other. In this case, the penetrating grid points (if any) are shifted to a position that reflects the ratio of the intersecting spheres radii. Finally, grid points (of the same Gaussian exponent) with a distance from atoms less than the predefined threshold are then averaged. The finalized grid presents the centers of the Gaussian basis s-functions. The exponents of the Gaussian functions are optimized "per atom type", where one exponent is used for hydrogen atom grid points and different exponents are employed for the C, N, and O atoms considered in this work. An example of a grid for the methyl-guanine···methyl-cytosine complex is depicted in Figure 2.

**2.2. Grid Parametrization.** Grid parameters, such as the radius of spheres, distance of grid points in the planar grid, and exponents of Gaussian s-functions are subjects of optimization. This delivers grid accuracy comparable to that of aDZ basis set which has a comparable number of grid points to the number of diffuse functions in aDZ. The most accurate and efficient grid methodology is achieved when each parameter is individually optimized for grid points on spheres of each type of atom. Because this is computationally impractical, we decided to optimize two sets of parameters: one for hydrogen atoms and another for the C, N, and O atoms (six parameters in total, $x_1 \cdots x_6$). We selected representative systems from the S22 data set, stacked and hydrogen-bonded uracil dimer and stacked benzene···indole complex, with both in equilibrium, and elongated geometries for use in optimization at the MP2 level with the cc-pVDZ basis set. The optimization problem is formulated as the search for the minimum of the following function; presuming the interaction energies in aDZ being negative:

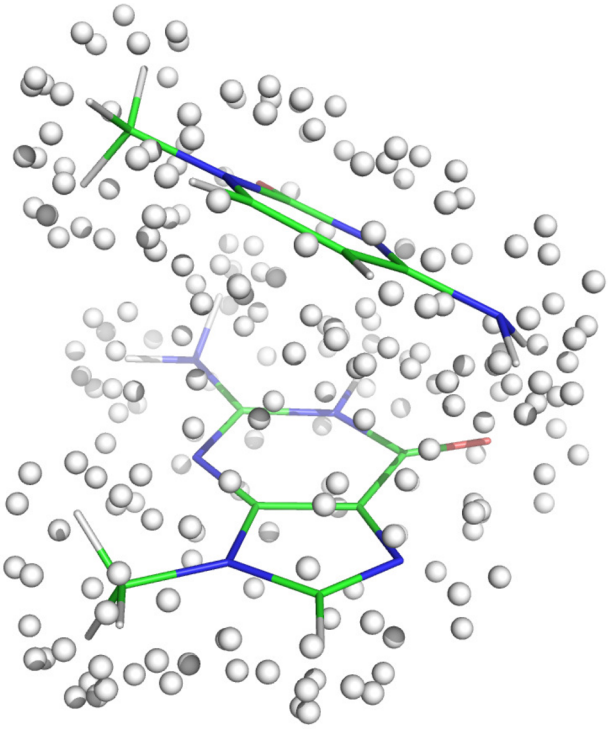


**Figure 2.** Illustration of the methyl-guanine···methyl-cytosine complex covered by a grid. Spheres represent centers of the s-type Gaussian functions.

$$f(x_1, \cdots, x_6) = \max(E_{\text{grid}} - E_{\text{aDZ}} + \text{penalty} + \text{del}) \quad (1)$$

where $E$ refers to pertinent interaction energies and del stands for a penalty proportional to the number of deleted linear dependent basis functions using the default threshold of $1 \times 10^{-5}$ a.u.; see eqs (2) and (3). Maximum in eq 1 is taken from the six values obtained for each noncovalent complex in the optimization, in each step. Two grids are optimized; the "standard", where *penalty* is for excess basis functions evaluated against aDZ (if any) and a "+10%", where *penalty* is for excess basis functions evaluated against aDZ + 10% tolerance (10% extra functions allowed, if any).

$$\text{penalty} = \text{abs}\left(\frac{\text{nbf}_{\text{grid}} - \text{nbf}_{\text{aDZ}}}{\text{nbf}_{\text{aDZ}}} E_{\text{aDZ}}\right) \quad (2)$$

where nbf refers to pertinent numbers of basis functions. The penalty for the deleted basis functions is calculated as

$$\text{del} = \text{abs}\left(\frac{10\text{nbf}_{\text{del}}}{\text{nbf}_{\text{aDZ}}} E_{\text{aDZ}}\right) \quad (3)$$

Minimization is a difficult process in itself, because the optimization function in eq 1 is not continuous. Parallel implementation of a modified simplex method is used to find the global minimum. The basic premise of the global simplex method is to generate a large set of vertices in $n$-dimensional

space where the minimum is expected to be found. A random subset with the minimum required number is then selected to form a simplex from this set of vertices. Parallel implementation was inevitable because the number of simplexes calculated during this optimization is huge.

**2.3. Computational Details.** The calculations were carried out using the Cholesky-decomposed (CD) two-electron integrals based MP2 method implemented in the MOLCAS package.[17] The threshold of $1 \times 10^{-6}$ a.u. for CD integrals was used because it delivers interaction energies with accuracy within a few hundredths of kilocalories per mole, yet still increases calculation speed significantly. Geometries were taken from other studies and are referenced in the relevant parts of the paper.

The grid parameters obtained via optimization toward the standard and +10% grid (in parentheses) are the following: hydrogen, 0.4273060304 (0.3993125136), and non-hydrogen atom grid s-function exponent, 0.1979525804 (0.1613579315); hydrogen, 0.6824029146 (0.7777672163), and non-hydrogen atom sphere radius scaling factor, 0.8533025490 (0.91457259450); hydrogen, 0.6249311576 (0.60884790960), and non-hydrogen atom planar grid edge length, 0.902745557195 (0.835508865780). Numbers are purposely presented with maximum number of valid digits, so that the resulting grids are identical to those used in this work.

## 3. RESULTS

One of the benefits expected from the proposed grid methodology is to provide more economical basis set expansion with increased investigated molecular system size compared to the corresponding basis set with diffuse functions. To prove this concept we show scaling of the number of grid points and the number of diffuse basis functions in aDZ basis set for three model systems representing three different molecular system classes: a pseudo one-dimensional system, polyacetylene (20 to 500 C-H units), as a model for molecular chains; a pseudo two-dimensional system, graphene ("2" coronene, "3" circumcoronene, up to "13" (1014 carbon and 78 hydrogen atoms)) as a model for planar molecules and, finally, hydrogenated fragments of diamond ("3 × 3 × 3" 54 carbon and hydrogen atoms, "5 × 5 × 5" 250 carbon and 150 hydrogen atoms, and "7 × 7 × 7" 686 carbon and 294 hydrogen atoms) as a model for a compact, three-dimensional molecule; see Figure 3.

The number of grid points approximately scales the number of diffuse basis functions for polyacetylene, and therefore, nothing is gained in this respect for 1D molecules. For 2D molecules, such as graphene, the benefit of using grid methodology is apparent. For the largest 2D system of 1092 atoms, the number of diffuse basis functions is more than twice the number of grid points.

The scaling of the number of grid points for the hydrogenated diamond fragments in the 3D molecule is the modest of all the model systems shown. For the largest fragment of 980 atoms, the factor between the number of diffuse basis functions and the grid points is approximately 2.24. The reason for steeper scaling of the number of diffuse basis functions for graphene-like molecules compared to diamond-like ones is in the ratio of atom count; where hydrogen has four diffuse basis functions per atom and carbon has nine. The ratio between hydrogen and carbon atom counts is higher in diamond-like systems at 1.0 for 3 × 3 × 3 to 0.42 for 7 × 7 × 7 compared to graphene-like ones where it is 0.5 for 2 units to 0.08 for 13 units. Nevertheless, as expected, the number of grid points scales more favorably for the 3D molecular system size than for 2D.

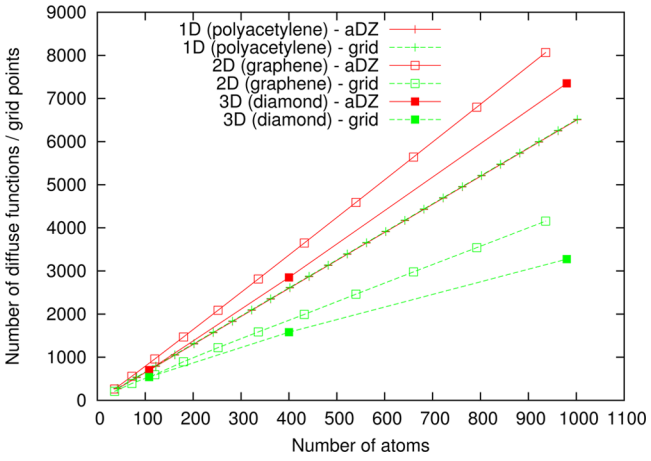



**Figure 3.** Scaling of the number of diffuse basis functions ((1s,1p,1d), as added in aDZ basis set) and grid points with the number of atoms in pseudo one-dimensional ("1D") molecule, polyacetylene, pseudo two-dimensional ("2D") molecule, graphene and thee-dimensional ("3D") molecule, hydrogenated diamond fragment.

The quality of the grid parameters obtained from the optimization restricted to the number of total basis functions lower than the corresponding aDZ basis set count is demonstrated in Tables 1 and 2. The reason for showing S22 interaction energy errors is that S22 is not only the parametrization but also a partial test data set because this parametrization was performed for only the small subset of 3 of 22 complexes. This was done in both equilibrium and elongated geometries with twice the optimal mutual monomer displacement along the interaction coordinate. The average error for equilibrium structures is negative, as seen in Table 1, and this shows that, on average, both grids "outperform" the aDZ basis set. However, this is not a strict analysis for the following two reasons. First, monotonous convergence toward the CBS limit with the basis set size is not guaranteed for interaction energies, and second, the MP2 method is not variational. However, according to quite numerous publications on the S22 data set, particularly related to the basis set convergence pattern, we know that the MP2 interaction energies in (a)XZ basis set series converge monotonously for at least the noncovalent complexes present in the data set. The total number of basis function using grids according to Table 1, again, is slightly larger compared to aDZ; on average by 9 to 38 basis functions, for the standard and +10% grids, respectively. The maximum error for the standard grid is 0.13 kcal/mol in the adenine··· thymine stacked complex, but this decreases to −0.03 kcal/mol in the larger, +10% grid. The largest error noted for the +10% grid is 0.02 kcal/mol in the water dimer. This dimer is simply too small a complex/molecule for efficient grid treatment, and therefore, it does not represent a true measure of grid methodology accuracy.

The situation is quite different in elongated complex structures, taken from the work of Gráfová et al.,[18] shown in Table 2. The errors for both grids are positive, highlighting that grid performance is inferior to aDZ. The most likely explanation for this stems from the fact that the interaction in stretched geometries is dominated by electrostatics, which is sensitive to the quality of the monomer electric properties description. The properties, such as dipole and quadrupole


dx.doi.org/10.1021/ct400692b | J. Chem. Theory Comput. 2013, 9, 5296−5304

**Table 1. MP2/aug-cc-pVDZ (aDZ) Interaction Energies [kcal/mol] for the S22 Data Set in Equilibrium Complex Geometries and the Respective Errors Obtained Using Grids (standard and +10%; See Section 2.2)[a]**

| | | | Δ(DZ + Grids) | | | |
|---|---|---|---|---|---|---|
| complex[b] | aDZ | | standard | | +10% | |
| A···T (H) | −14.71 | (536) | −0.19[c] | (−17)[d] | −0.24 | (35) |
| A···T (S) | −13.24 | (536) | 0.13 | (−14) | −0.03 | (36) |
| $(NH_3)_2$ | −2.68 | (100) | −0.08 | (30) | −0.11 | (45) |
| $(H_2O)_2$ | −4.37 | (82) | 0.08 | (20) | 0.02 | (40) |
| $(CH_4)_2$ | −0.39 | (118) | −0.05 | (28) | −0.05 | (44) |
| $(C_2H_4)_2$ | −1.18 | (164) | −0.12 | (32) | −0.18 | (40) |
| $C_2H_2\cdots C_2H_4$ | −1.39 | (146) | −0.10 | (28) | −0.07 | (48) |
| $(\text{formic acid})_2$ | −15.99 | (174) | −0.24 | (16) | −0.09 | (44) |
| $\text{formamide}_2$ | −13.95 | (192) | 0.04 | (20) | −0.04 | (44) |
| B···$H_2O$ | −2.98 | (233) | −0.13 | (15) | −0.16 | (38) |
| B···$NH_3$ | −2.21 | (242) | −0.10 | (17) | −0.12 | (42) |
| B···$CH_4$ | −1.47 | (251) | −0.13 | (15) | −0.10 | (41) |
| $B_2$ (T) | −3.10 | (384) | −0.11 | (10) | −0.11 | (40) |
| $B_2$ (PD) | −4.25 | (384) | 0.03 | (4) | −0.02 | (38) |
| B···I (T) | −6.10 | (462) | −0.21 | (0) | −0.20 | (39) |
| B···I (S) | −7.13 | (462) | 0.09 | (−2) | 0.01 | (29) |
| $\text{pyridine}_2$ | −6.00 | (348) | 0.05 | (−8) | −0.06 | (32) |
| 2-A···2-P | −15.55 | (421) | −0.18 | (3) | −0.19 | (34) |
| $\text{phenol}_2$ | −6.79 | (430) | −0.03 | (5) | 0.01 | (34) |
| $U_2$ (S) | −9.80 | (440) | 0.01 | (−6) | −0.03 | (34) |
| $U_2$ (H) | −18.41 | (440) | −0.27 | (−14) | −0.21 | (28) |
| B···HCN | −4.38 | (247) | −0.28 | (18) | −0.23 | (40) |
| MAX | | | 0.13 | (32) | 0.02 | (48) |
| MD | | | −0.08 | (9) | −0.10 | (38) |
| RMS | | | 0.14 | (17) | 0.13 | (39) |

[a]The total number of basis functions and the difference with respect to aDZ are shown in parentheses. "MAX" stands for maximum signed deviation, "MD", for mean deviation, and "RMS", for root mean squared deviation of errors. [b]"H" and "S" stand for hydrogen-bonded and stacked; "T" and "PD" stand for T-shaped and parallel-displaced; other symbols: "A···T"—adenine···thymine, "B"—benzene, "I"—indole, "2-A"—2-aminopyridine, "2-P"—2-pyridoxine, "U"—uracil. [c]Negative value indicates "superior" IE compared to aDZ. [d]Negative value indicates less total number of basis functions compared to aDZ.

moments and polarizabilities, are naturally described more accurately using diffuse basis functions on all atoms compared to surface basis function grids. Nevertheless, the basis set saturation for a particular basis set type improves with the size of the calculated molecular system (at least for "compact" ones), thus the discrepancy is expected to diminish in larger systems. Furthermore, the discrepancy is expected to become lower with increase in the underlaying grid basis set size, for example TZ vs DZ, because basis set saturation is dependent on system size, and is therefore expected to converge even more rapidly. When the grid is combined with a basis set capable of delivering more accurate monomer electric properties, this discrepancy is expected to decrease even further. This issue, however, is still not completely understood and we assume that the mutual overlap of diffuse basis functions in each monomer, as in aDZ for example, has smaller impact on the resulting interaction energies compared to deficiency in the electrostatics description.

The grid accuracy for the S66 data set[19] is summarized in Table 3. Results for the complete data set are shown for equilibrium and elongated complex structures. Analogous to S22, all errors are positive in elongated complex geometries for both standard and +10% grids. The maximum error obtained for equilibrium structures for S66 is slightly larger than that for S22, 0.15, and 0.06 kcal/mol, for the standard and +10% grid, respectively; however, the average errors are about the same, −0.05 and −0.11 kcal/mol. These results confirm again the versatility and transferability of the optimized grid parameters.

The results for stretched S22 and S66 geometries evaluated in Tables 2 and 3 demonstrate the grid's performance in the asymptotic region rather than for distorted geometries within "contact", close to the equilibrium region. Therefore, we decided to perform accuracy assessment for potential energy curves (PECs) corresponding to the transition between various important structures on the potential energy surface (PES) of the benzene dimer and methyl-guanine···methyl-cytosine (mG···mC). Two PECs, one for the parallel-displaced (PD) to in-plane (IP) in Figure 4, and the other for the T-shaped benzene dimer (T) to sandwich (S) structure conversion illustrated in Figure 5. Geometries of both PECs are taken from the work of Grimme et al.[20] and were calculated using DZ and aDZ basis sets and grids. The mG···mC PEC, calculated using the same methodology based on structures taken from the work of Černý and Hobza,[21] is highlighted in Figure 6.

Let us first analyze the results obtained for PD-to-IP PEC in Figure 4. The MP2/aDZ interaction energies range from −4.01 kcal/mol (PD structure, number "1" in the plot) to −0.57 kcal/mol (IP structure, number "8" in the plot). The importance of diffuse basis functions is obvious. The span of MP2/DZ interaction energies is 1.38 kcal/mol compared to the 3.44 kcal/mol obtained using the aDZ basis set. The maximum error for the standard grid is 0.07 kcal/mol, for structure 3, and decreases to 0.03 kcal/mol in the larger +10% grid for

**Table 2. MP2/aug-cc-pVDZ (aDZ) Interaction Energies [kcal/mol] for the S22 Data Set in Elongated Complex Geometries and the Respective Errors Obtained Using Grids (standard and +10%; See Section 2.2)[a]**

| | | | Δ(DZ + grids) | | | |
|---|---|---|---|---|---|---|
| complex[b] | aDZ | | standard | | +10% | |
| A···T (H) | −2.55 | (536) | 0.02[c] | (−17)[d] | 0.02 | (35) |
| A···T (S) | −0.99 | (536) | 0.04 | (−16) | 0.03 | (34) |
| $(NH_3)_2$ | −0.36 | (100) | 0.03 | (30) | 0.01 | (46) |
| $(H_2O)_2$ | −0.94 | (82) | 0.01 | (22) | 0.01 | (40) |
| $(CH_4)_2$ | −0.01 | (118) | 0.00 | (28) | 0.00 | (46) |
| $(C_2H_4)_2$ | −0.03 | (164) | 0.00 | (32) | 0.00 | (42) |
| $C_2H_2\cdots C_2H_4$ | −0.15 | (146) | 0.01 | (28) | 0.01 | (44) |
| $(\text{formic acid})_2$ | −3.53 | (174) | 0.05 | (16) | 0.03 | (45) |
| $\text{Formamide}_2$ | −3.44 | (192) | 0.07 | (22) | 0.04 | (46) |
| $B\cdots H_2O$ | −0.51 | (233) | 0.04 | (13) | 0.03 | (41) |
| $B\cdots NH_3$ | −0.29 | (242) | 0.02 | (11) | 0.02 | (47) |
| $B\cdots CH_4$ | −0.14 | (251) | 0.01 | (15) | 0.01 | (40) |
| $B_2$ (T) | −0.41 | (384) | 0.03 | (4) | 0.02 | (40) |
| $B_2$ (PD) | −0.11 | (384) | 0.00 | (4) | 0.00 | (36) |
| B···I (T) | −1.21 | (462) | 0.07 | (0) | 0.05 | (37) |
| B···I (S) | −0.18 | (462) | 0.00 | (−2) | 0.00 | (27) |
| $\text{pyridine}_2$ | −0.24 | (348) | 0.01 | (−5) | 0.01 | (36) |
| 2-A···2-P | −3.28 | (421) | 0.03 | (3) | 0.01 | (33) |
| $\text{Phenol}_2$ | −1.43 | (430) | 0.03 | (5) | 0.02 | (35) |
| $U_2$ (S) | −0.71 | (440) | 0.02 | (−5) | 0.02 | (35) |
| $U_2$ (H) | −4.43 | (440) | 0.03 | (−14) | 0.03 | (29) |
| B···HCN | −0.89 | (247) | 0.05 | (18) | 0.05 | (38) |
| MAX | | | 0.07 | (32) | 0.05 | (47) |
| MD | | | 0.03 | (9) | 0.02 | (39) |
| RMS | | | 0.03 | (17) | 0.02 | (39) |

[a]Elongated geometries correspond to twice the optimal displacement along the interaction coordinate, as defined in S22 × 5.[18] The total number of basis functions and the difference with respect to aDZ are shown in parentheses. [b]Complex name codes and statistical functions are explained in the footnote of Table 1. [c]Positive value indicates "inferior" IE compared to aDZ. [d]Negative value indicates less total number of basis functions compared to aDZ.

**Table 3. S66 Data Set:[19] Statistical Evaluation of Interaction Energies (IE) [kcal/mol] and Basis Function Excess (Δnbf) of Grids with Respect to MP2/aug-cc-pVDZ (aDZ) Reference Data in Equilibrium and Elongated (Twice the Equilibrium Distance) Geometries**

| | MAX | MD | RMS |
|---|---|---|---|
| S66—equilibrium geometries | | | |
| standard grid | | | |
| IE | 0.15 | −0.05[a] | 0.11 |
| Δnbf[b] | 31 | 14 | 18 |
| +10% grid | | | |
| IE | 0.04 | −0.11 | 0.14 |
| Δnbf | 94 | 49 | 50 |
| S66—elongated geometries | | | |
| standard grid | | | |
| IE | 0.06 | 0.02 | 0.03 |
| Δnbf | 31 | 14 | 18 |
| +10% grid | | | |
| IE | 0.04 | 0.01 | 0.02 |
| Δnbf | 95 | 48 | 50 |

[a]Negative value indicates "superior" IE compared to aDZ. [b]Positive value indicates more total number of basis functions compared to aDZ.

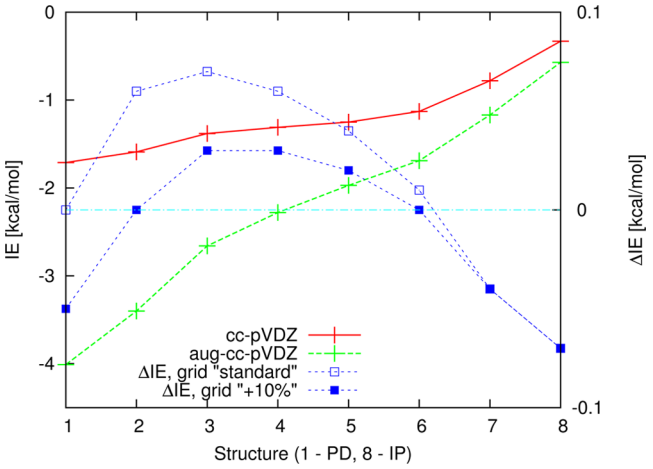



**Figure 4.** Interaction energy (IE) change along the transition from parallel-displaced (PD) to in-plane (IP) benzene dimer structure.[20] MP2 IEs in cc-pVDZ (DZ, 228 basis functions (b.f.)) and aug-cc-pVDZ (aDZ, 384 b.f.) basis sets are shown (left *y*-axis) as well as errors of MP2 DZ + grids (390−391 b.f. for standard and 416−421 for +10%, respectively) IEs with respect to aDZ basis set (right *y*-axis).

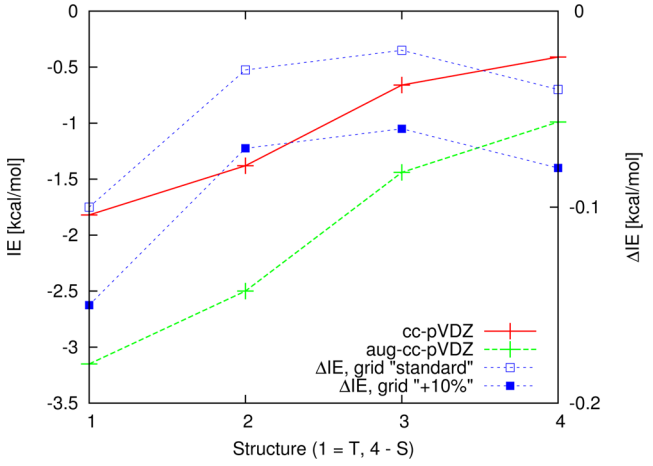



**Figure 5.** Interaction energy (IE) change along the transition from T-shaped (T) to sandwich (S) benzene dimer structure.[20] MP2 IEs in cc-pVDZ (DZ, 228 basis functions (b.f.)) and aug-cc-pVDZ (aDZ, 384 b.f.) basis sets are shown (left *y*-axis) as well as errors of MP2 DZ + grids (392 b.f. for standard and 430−436 for +10%, respectively) IEs with respect to aDZ basis set (right *y*-axis).

structures 3 and 4. Errors along the PEC span the intervals of −0.07 to 0.07 (2.6%) kcal/mol for the standard, and −0.07 to 0.03 (1.3%) kcal/mol for +10% grid, respectively. This indicates geometry dependence within approximately 3−4%. Slightly more favorable results are obtained for the T-to-S PEC in Figure 5. The interaction energy span is smaller than that PD-to-IP at 2.16 kcal/mol for aDZ and 1.41 kcal/mol for DZ, as also is the deviation between the DZ and aDZ PECs which are more parallel. The errors for both grids are negative values which indicates that the grid results are superior to those from aDZ. The geometry dependence of errors varies within approximately 4% for both grids.

Description of the mG···mC PEC in Figure 6 is more challenging than that for benzene dimer PECs, because the grid is required to substitute the presence of diffuse basis functions equally well for different interaction motifs, the $\pi-\pi$ stacking and hydrogen bonding. Indeed, the span of "errors" is notably


dx.doi.org/10.1021/ct400692b | J. Chem. Theory Comput. 2013, 9, 5296−5304

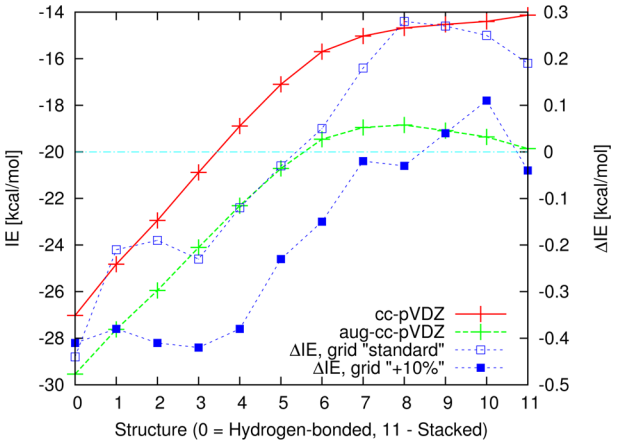



**Figure 6.** Interaction energy (IE) change along the transition from hydrogen-bonded methyl-guanine···methyl-thymine (mG···mC) to stacked structure. MP2 IEs in cc-pVDZ (DZ, 364 basis functions (b.f.)) and aug-cc-pVDZ (aDZ, 609 b.f.) basis sets are shown (left *y*-axis) as well as errors of MP2 DZ + grids (591−602 b.f. for standard and 651−663 for +10%, respectively) IEs with respect to aDZ basis set (right *y*-axis).

larger here at −0.44 to 0.28 (1.5%) kcal/mol for the standard grid and −0.42 to 0.11 (0.6%) kcal/mol for the +10% grid. It should be noted here that the absolute interaction energies and their span is an order of magnitude larger than the PD-to-IP and T-to-S PECs although the errors in percentile scale are comparable.

## 4. CONCLUSIONS

An alternative concept for assembling an AO basis set for molecules suited to noncovalent interaction calculations is proposed in this work. Diffuse basis functions uniformly used for all atoms in a molecule can cause linear dependencies in the basis set, especially for extended and compact molecular species. The presence of linear dependent basis functions leads to numerical instabilities in quantum chemical calculations. These are manifested, for example, in orbital optimization convergence problems. In the proposed methodology, the presence of diffuse basis functions is mimicked by a grid of off-center Gaussian s-functions covering the surface of the molecule. Exponents of the s-functions and the density of their distribution are optimized for a subset of noncovalent complexes from the S22 data set in order to reproduce the MP2/aug-cc-pVDZ (aDZ) results; both in terms of accuracy and the total number of basis functions. Because the extra Gaussian basis functions are placed on the surface of the molecule, the probability of linear dependence occurrence does not exceed that in the underlying basis set without diffuse basis functions. The side benefit gained from using such grids is in the scaling of the total number of basis functions with the size of the molecule. This proves more favorable than using basis sets with diffuse basis functions on all atoms.

Two sets of optimized grid parameters are presented in this work. The first is the so-called standard grid, aimed at the most faithful reproduction of the number of diffuse basis functions of aDZ. The second set is the so-called +10% grid which can exceed the number of diffuse basis functions of aDZ by approximately 10%. The performance of the grids is tested on equilibrium and elongated (twice the optimal mutual monomer displacement along the interaction coordinate) geometries of complexes in the S22 and S66 data sets. The average errors obtained are negative, proving that the grids outperform the aDZ basis set. The total number of basis functions is, however, slightly larger compared to aDZ. Maximum errors for the standard and +10% grids calculated for S22 and S66 are 0.13 and 0.02 kcal/mol and 0.15 and 0.04 kcal/mol, respectively.

The versatility of the grid methodology is tested on three potential energy curves (PECs) corresponding to (1) parallel-displaced benzene dimer conversion to in-plane structure, (2) T-shaped benzene dimer conversion to sandwich structure, and (3) the most challenging test on hydrogen-bonded methyl-guanine···methyl-cytosine (mG···mC) conversion to the stacked structure. The errors with respect to the reference aDZ interaction energies change with the calculated complex structures, but the deviation is quite small, on the order of a few percentile. The maximum error obtained along the potential energy curves using the standard and +10% grid are 0.28 and 0.11 kcal/mol, both for mG···mC PECs, which corresponds to 1.5 and 0.6% deviation.

## ■ ASSOCIATED CONTENT

### Supporting Information

Cartesian coordinates for the investigated noncovalent complexes (with and without grids) and also the program for grid generation, written in C-language, together with the documentation and case studies. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Authors

*E-mail: pitonak@fns.uniba.sk.
*E-mail: neogrady@fns.uniba.sk.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the Slovak Research and Development Agency, contract No. APVV-0059-10, and it also forms part of Research project No. Z40550506 of the Institute of Organic Chemistry and Biochemistry, Academy of Science of the Czech Republic. Some calculations were performed in the Computing Centre of the Slovak Academy of Sciences and the High Performance Computing Center of the Matej Bel University in Banská Bystrica using the supercomputing infrastructure acquired in project ITMS 26230120002 and 26210120002 (Slovak infrastructure for high-performance computing) and supported by the Research & Development Operational Programme funded by the ERDF.

## ■ REFERENCES

(1) Lee, E. C.; Kim, D.; Jurečka, P.; Tarakeshwar, P.; Hobza, P.; Kim, K. S. *J. Phys. Chem. A* **2007**, *111*, 3446−3457.
(2) Riley, K. E.; Pitoňák, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023−5063.
(3) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553−566.
(4) Jankowski, K.; Peterson, K. A. *Phys. Rev. A* **2012**, *86*, 022526.
(5) Jurečka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365*, 89−94.
(6) Marshall, M. S.; Sherrill, C. D. *J. Chem. Theory Comput.* **2011**, *7*, 3978−3982.
(7) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *135*, 194102.
(8) Tao, F. M.; Pan, Y. K. *J. Chem. Phys.* **1992**, *97*, 4989−4995.
(9) Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374−7391.

(10) Janowski, T.; Ford, A. R.; Pulay, P. *Mol. Phys.* **2010**, *108*, 249−257.
(11) Sedlák, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2013**, *9*, 3364−3374.
(12) Nicolas, C.; Boutin, A.; Levy, B.; Borgis, D. *J. Chem. Phys.* **2003**, *118*, 9689−9696.
(13) Choi, T. H.; Jordan, K. D. *Chem. Phys. Lett.* **2008**, *464*, 139−143.
(14) Whitten, J. L. *J. Chem. Phys.* **1966**, *44*, 359−364.
(15) Harrison, J. F. *J. Chem. Phys.* **1967**, *46*, 1115−1118.
(16) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(17) Aquilante, F.; De Vico, L.; Ferre, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitoňák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010**, *31*, 224−247.
(18) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 2365−2376.
(19) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.
(20) Grimme, S.; Muck-Lichtenfeld, C.; Antony. *J. Phys. Chem. Chem. Phys.* **2008**, *10*, 3327−3334.
(21) Černý, J.; Hobza, P. *Chem. Commun.* **2010**, *46*, 383−385.