Accurate, Precise, and Efficient Theoretical Methods To Calculate Anion−π Interaction Energies in Model Structures

Pál D. Mezei, † Gábor I. Csonka,* † Adrienn Ruzsinszky,‡ and Jianwei Sun‡

†Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary
‡Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States

ABSTRACT: A correct description of the anion−π interaction is essential for the design of selective anion receptors and channels and important for advances in the field of supramolecular chemistry. However, it is challenging to do accurate, precise, and efficient calculations of this interaction, which are lacking in the literature. In this article, by testing sets of 20 binary anion−π complexes of fluoride, chloride, bromide, nitrate, or carbonate ions with hexafluorobenzene, 1,3,5-trifluorobenzene, 2,4,6-trifluoro-1,3,5-triazine, or 1,3,5-triazine and 30 ternary π-anion−π sandwich complexes composed from the same monomers, we suggest domain-based local-pair natural orbital coupled cluster energies extrapolated to the complete basis-set limit as reference values. We give a detailed explanation of the origin of anion−π interactions, using the permanent quadrupole moments, static dipole polarizabilities, and electrostatic potential maps. We use symmetry-adapted perturbation theory (SAPT) to calculate the components of the anion−π interaction energies. We examine the performance of the direct random phase approximation (dRPA), the second-order screened exchange (SOSEX), local-pair natural-orbital (LPNO) coupled electron pair approximation (CEPA), and several dispersion-corrected density functionals (including generalized gradient approximation (GGA), meta-GGA, and double hybrid density functional). The LPNO-CEPA/1 results show the best agreement with the reference results. The dRPA method is only slightly less accurate and precise than the LPNO-CEPA/1, but it is considerably more efficient (6−17 times faster) for the binary complexes studied in this paper. For 30 ternary π-anion−π sandwich complexes, we give dRPA interaction energies as reference values. The double hybrid functionals are much more efficient but less accurate and precise than dRPA. The dispersion-corrected double hybrid PWPB95-D3(BJ) and B2PLYP-D3(BJ) functionals perform better than the GGA and meta-GGA functionals for the present test set.

INTRODUCTION

Accurate description of noncovalent interactions is essential for further developments in self-assembly of molecules, catalysis, materials science, medicine, biological functions, and systems biology.1−3 Noncovalent interactions of aromatic rings play a very important role in biology and chemistry,4 for example in protein folding, drug−receptor interactions, crystal engineering, DNA and RNA stacking, protein DNA aromatic interactions, DNA repair, and maintaining DNA stability.5

The importance of cation−π interactions was recognized some time ago,7−9 and an accurate description of preferred arrangements is quite feasible. Cation−π complexes can be described by a single type of minimum energy structure: the cation positions itself close to the ring center. The analogous structure to the cation−π complex, the anion−π complex, was ignored for a long time due to the supposed electrostatic repulsion.10 However, recently this relatively unexplored interaction has gained attention.11,12 A survey showed13 that contrary to general expectations, anion−π interactions are more frequent in the Cambridge Structural Database14 than cation−π interactions. Recently direct evidence for the anion−π interactions is obtained by tandem mass spectrometric experiments with naphthalenediimide models where only the π-acidic surface is left for anions to interact with.15 The correct description of the interaction between anions and aromatic rings can be used for the design of selective anion receptors and channels, and this is important for the advances in the field of supramolecular chemistry.16,17 For complexes with the fluoride ion a covalently bonded structure, called the “Meisenheimer” complex, is the most stable in gas phase.18 A typical example is the 1,3,5-triazine (TAZ)···F− complex. However, adding two water molecules to this complex makes the H-bonded complex more stable, competing with anion−π and displaced anion−π complex forms. Adding three or four water or acetonitrile molecules makes the displaced anion−π complex form the most stable.18 Adding three or four water molecules to the TAZ···Cl− complex makes the anion−π complex form the most
stable.\textsuperscript{18} Since the anion—π and displaced anion—π forms are almost equally stable, these two forms occur in many crystal structures.\textsuperscript{18} Compared to cation—π, the anion—π complex interaction energy is supposed to have a more significant van der Waals (vdW) and induction components, because anions have extended, polarizable electron clouds.\textsuperscript{12}

In this work we focus on a quantitatively accurate and precise description of the anion—π complexes. We give high quality interaction energy values for benchmarks and analyze quantitatively the importance of the exchange, induction, electrostatic, and vdW energy components of the interaction energy. In the following section we provide the computational details. Next we analyze the performance of various dispersion-corrected density functionals up to double hybrids, dRPA and RPA, and give a cost performance evaluation. Finally we select the most efficient best-performing method for further studies.

\section*{METHODOLOGY}

The vdW interaction is a universal noncovalent long-range attraction between chemical species, arising from a correlation between instantaneous electronic charge fluctuations on each. In the large separation limit of two ground-state spherical interacting atoms, the nonretarded potential energy of the vdW interaction between them can be expressed as a power series in the inverse of \(d\). According to second-order perturbation theory at long-range we get a multipolar expansion

\[ E_{vdW} = -\frac{C_6}{d^6} - \frac{C_8}{d^8} - \frac{C_{10}}{d^{10}} + \ldots \]  

where \(d\) is the internuclear separation (\(d\) must be sufficiently large to obtain nonoverlapping electron densities), the system-dependent \(C_6\) describes the dynamic dipole—dipole interaction, \(C_8\) describes the dipole—quadrupole interaction, and \(C_{10}\) describes the quadrupole—quadrupole and the dipole—octupole interactions. Even in the ground state, each sphere has a zero-point motion. The zero-point oscillations of the charge density in one sphere produce a long-range multipolar electric field that interacts with and correlates with the zero-point oscillations of the charge density in the other sphere. The vdW interaction is much weaker in strength than a usual chemical bond, but a correct description of such weak interactions might be critical for biological or material science applications. Long- and medium-ranged weak interactions are largely responsible for many phenomena such as sublimation of molecular solids, high interlayer mobility of graphite, folding of long biomolecular chains such as DNA, RNA, and proteins, folding of polymers, and many others.

The wave function based methods that include electron correlation effects, such as configuration interaction (CI), many-body perturbation theory (MBPT), or coupled cluster (CC) methods, are able to estimate vdW interaction energy quite reasonably. The CC singles and doubles with perturbative triples, CCSD(T), method is currently the reference method for noncovalent interactions, but its applicability is limited to single-reference ground states and to relatively small systems due to its steep \(N^6\) scaling of computational cost with system size \(N\). This method shows slow basis set convergence and expensive basis set extrapolations are necessary for useful precision, but notice the emerging new localized CC methods that efficiently decrease the computational time and scaling with size.\textsuperscript{19} The much less expensive second-order direct Møller—Plesset perturbation theory (MP2) includes long-range interactions but leads to mixed results for anion—π complex interaction energy.

A relatively inexpensive but usefully accurate method can be constructed based on density functional theory (DFT).\textsuperscript{20—24} The asymptotic interaction energy of semilocal DFT approximations decreases exponentially instead of the proper minus sixth power. This failure greatly limits the applicability of pure semilocal DFT to a large class of weakly bound systems such as atom and intermolecular pairs, liquids, and molecular crystals. To overcome or reduce this error, a dispersion correction can be added in a “post-Kohn-Sham manner” by pairwise interactions.\textsuperscript{25} Pair-wise interaction can be constructed in empirical\textsuperscript{26} or nonempirical ways.\textsuperscript{27} In the latter approach, a nonempirical model is proposed for the multipole polarizability or \(C_{ij}\) coefficients based on physical constraints. Grimme et al. introduced the D3 model\textsuperscript{28} that uses precomputed dispersion coefficients, and the environment of an atom in a molecule is modeled by a fractional coordination number. D3(BJ)\textsuperscript{29} is an improved variant of D3 that uses the damping of Becke and Johnson (BJ).\textsuperscript{30}

An alternative strategy is to add approximate dispersion energy via MP2 or Görling-Levy second-order Kohn-Sham perturbation theory.\textsuperscript{31,32} This is the basic idea behind the so-called double hybrid methods that combine semilocal DFT functionals, exact exchange, and second order perturbation theory reviewed in ref 26. Notice that double hybrid functionals miss a considerable part of long-range correlation effects and require a \textit{a posteriori} dispersion correction for better results.\textsuperscript{29}

Interesting alternatives are the nonlocal vdW correlation functionals that model the multipole polarizability. Unlike the pairwise approaches, which usually take the static polarizability as an input from experiment or higher-level calculation, the nonlocal correlation approximations are functionals of the density, making a seamless approximation for the nonlocal correlation energy. Long-range van der Waals functionals of the Langreth\textsuperscript{33—35} or Vydrov-van Voorhis-type (VV10)\textsuperscript{36} start from a semilocal exchange-correlation functional and then add a fully nonlocal correction. Thus, these methods strongly rely upon the underlying semilocal approximation.

Most approximations, including the pairwise models and the nonlocal van der Waals functionals as well, miss the many-body terms arising from the coupling of the oscillators. The number of atoms is additive, but the exact multipole polarizability is not. The exact multipole polarizability is also nonlocal. The direct Random Phase Approximation (dRPA) using the Kohn—Sham orbitals and orbital energies\textsuperscript{37—39} provides a “seamless” treatment of all forces including vdW at any separation. dRPA neglects antisymmetrization effects on the correlation energy that are included in full RPA. The correlation energy in the dRPA can be understood as the zero-point vibrational energy of virtual electronic transitions. dRPA contains the zero-point energy of plasmons coupled by the long-ranged Coulomb interaction between subsystems. Its long-range correlation is almost exact, and it properly captures the nonpairwise-additive nature of long-range interactions.\textsuperscript{40—42} Notice that dRPA is exact for the exchange energy but not for the correlation energy. It can also be derived from the “ring-coupled cluster” theory by removal of the exchange terms from the two-particle Hamiltonian.\textsuperscript{43} The absence of these terms is responsible for a self-correlation error and for the expected failure of the dRPA for stretched \(H_2^+,\text{He}_2^+,\text{and Ne}_2^+\).\textsuperscript{44—46} We have argued that the RPA correlation hole, even after semilocal or short-range correction, is not correct at midrange in a molecule.\textsuperscript{47,48} This is
an interesting complementarity to semilocal DFT, where the short- and midrange energy might be correct but not the long-range component of the energy.

## COMPUTATIONAL DETAILS

We investigate 20 binary anion–π complexes of fluoride, chloride, bromide, nitrate, or carbonate ions with hexafluorobenzene (HFB), 1,3,5-trifluorobenzene (TFB), 2,4,6-trifluoro-1,3,5-triazine (TFZ), or 1,3,5-triazine (TAZ) as shown in Figure 1 and 30 ternary π–anion–π′ sandwich complexes as shown in Figure 2.1

Figure 1. Structure and the main symmetry axis of the anion–π binary complexes 1–20 taken from ref 1.

Figure 2. Structure and the main symmetry axis of the π–anion–π′ ternary complexes 21–50 taken from ref 1.

The resolution of the identity MP2 (RI-MP2) method is considerably faster than the MP2, and it was shown that the interaction energies and equilibrium distances are almost identical for both methods for HFB with anions, TFB with cations and anions, and TAZ with cations and anions.49 Consequently the RI-MP2 method can replace the MP2 for the calculations of ion–π interactions.49 The RI-MP2/6-31++G(d,p) equilibrium molecular geometries of the monomers and the 20 anion–π and 30 sandwich complexes were taken from the literature.1 We did not use the alternatively available MPWB1K/6-31++G(d,p) optimized geometries1 as those yield higher CCSD(T) energies and are thus less optimal for our purposes. The application of different geometries results in a 1–2 kcal mol⁻¹ energy change.

The DFT calculations were performed by using Orca 3.0.160 on Intel Core i7-2630QM Processor. At the generalized gradient approximation (GGA) level we used the Perdew–Burke–Ernzerhof functional (PBE),51 its revised version with more repulsive exchange (revPBE),52 and the reftitted Perdew–Wang 1986 exchange functional (rPW86) with the PBE correlation (rPWBPBE).53 At the meta-GGA level we used the Tao-Perdew-Staroverov-Scuseria functional (TPSS).54 We also used three double hybrid functionals: 1) the B2PLYP 55 containing Becke 1988 exchange (B88),56 exact exchange, Lee–Yang–Parr correlation (LYP),57 and second order perturbation theory correlation (PT2); 2) the mPW2PLYP58 containing modified Perdew–Wang 1991 exchange (mPW91),59 exact exchange, and LYP and PT2 correlation terms; and 3) the PWBP9560 containing the parametrized version of PW91 exchange and B9562 correlation functionals and also exact exchange and PT2 correlation. To accelerate the calculations, we applied the RI63 approximation for the Coulomb, for the exact exchange and for the PT2 correlation terms (RI-J, RI-JK, and RI-C algorithms, respectively). The DFT-BLYP and B3LYP results are poor for these complexes, and the interested readers might find those results in ref 49.

We used the MRCC64 (06/20/2014 release) program for the density-fitted MP2, the random phase approximation with exchange terms (RPAX), direct RPA (dRPA) and dRPA + second order screened exchange (SOSEX) calculations.65 The dRPA correlation energy is obtained as

\[ E^{dRPA} = \frac{1}{2} \text{tr} [BT] \]

where \( B_{ijkl} = (ijkl) \) are nonantisymmetrized two electron repulsion integrals using occupied \( ij \) and virtual \( ab \) spin–orbital indices, and \( T_{ijkl} = t_{ij}^a t_{kl}^b \) are the double excitation amplitude matrix elements. The amplitude matrix can be obtained from the solution of the Riccati equation as described in refs 43 and 66. This algorithm scales as \( (N_{occ} N_{virt})^4 \), where \( N_{occ} \) and \( N_{virt} \) are the number of occupied and unoccupied orbitals, respectively, leading to overall scaling of \( N^4 \) with the molecular size \( N \). However, this unfavorable scaling is reduced by applying a Cholesky-decomposition of energy denominators as proposed by Hesselman.67 The dRPA + SOSEX (we shall use SOSEX for short) correlation energy is obtained as

\[ E^{SOSEX} = \frac{1}{2} \text{tr} [KT] \]

where \( K_{ijkl} = (ij||ab) \) are spin-singlet adapted antisymmetrized two electron repulsion integrals.

The RPAX (RPA or ring-CCD or rCCD) correlation energy is obtained according to eqs 10 and 13 in ref 65. We shall use RPA in this paper for RPAX.

We used Orca 3.0.1 for local pair natural orbital (LPNO)68 coupled electron pair approximation (CEPA/1)69 and domain-based DLPNO-CCSD(T)70 calculations. The post-HF calculations were performed on AMD Opteron 6174 and Intel Xeon X5680 processors of the National Information Infrastructure Development Institute (NIIF).

For all calculations we used Dunning’s correlation consistent aug-cc-pVXZ (AXZ) basis sets and the corresponding auxiliary density fitting basis sets (aug-cc-pVXZ/J, aug-cc-pVXZ/JK, and aug-cc-pVXZ/C). The extrapolated complete basis set (CBS) HF and MP2 correlation energies were calculated as shown in eqs 4 and 5.
\[ E_{\text{CSR}}^{\text{HF}} = E_{\text{CSR}}^{\text{HF}} + \alpha(E_{\text{CSR}}^{\text{HF}} - E_{\text{CSR}}^{\text{CCSD(T)})}) \]

(4)

and

\[ E_{\text{CSR}}^{\text{MP2}} = E_{\text{CSR}}^{\text{MP2}} + \beta(E_{\text{CSR}}^{\text{MP2}} - E_{\text{CSR}}^{\text{CCSD(T)})}) \]

(5)

where \( E_{\text{CSR}}^{\text{HF}} \) is the estimated HF/CSR, \( E_{\text{CSR}}^{\text{HF}} \) is the HF/AQZ energy, etc. The extrapolation parameters are \( \alpha = 0.269 \) and \( \beta = 0.712 \). These parameters are slightly different from the earlier parameters optimized for the conformational energies of carbohydrates.\(^\text{7,1}\)

The DLPNO-CCSD(T)/CBS energy is estimated according to eqs 6 and 7:

\[ E_{\text{CSR}}^{\text{MP2}} = E_{\text{CSR}}^{\text{HF}} + E_{\text{CSR}}^{\text{MP2}} \]

(6)

\[ E_{\text{CSR}}^{\text{CCSD(T)})} = E_{\text{CSR}}^{\text{MP2}} + (E_{\text{CSR}}^{\text{CCSD(T)})} - E_{\text{CSR}}^{\text{MP2}}) \]

(7)

In this paper we use DLPNO-CCSD(T)/CBS electronic energies for reference.

The semilocal density functional methods cannot describe the purely nonlocal dispersion interaction. To account for the dispersion we applied the nonself-consistent VV10 nonlocal correction.\(^\text{36}\) The electron density is only slightly altered by the VV10 correction, and thus the nonself-consistency results in a negligible error, as our analysis shows for these complexes. The VV10 correlation energy is given by the nonlocal term minus its negligible error, as our analysis shows for these complexes. The VV10 correction, and thus the nonself-consistency results in a negligible error, as our analysis shows for these complexes. The VV10 correlation energy is given by the nonlocal term minus its negligible error, as our analysis shows for these complexes. The VV10 correction, and thus the nonself-consistency results in a negligible error, as our analysis shows for these complexes.

\[ \Phi(\mathbf{r}, \mathbf{r'}) = \frac{1}{2} \int \int n(\mathbf{r}) \Phi(\mathbf{r}, \mathbf{r'}) n(\mathbf{r'}) d^3r d^3r' \]

(9)

\[ \int \int \int_C n(\mathbf{r}) d^3r d^3r' \]

(8)

According to the local polarizability model, the nonlocal term can be written in a general form of the average interaction between two differential electron gas volumes, where the quality of the interaction is determined by a correlation kernel (eq 9)

\[ \Phi(\mathbf{r}, \mathbf{r'}) = \frac{3}{2} g(\mathbf{r}) g(\mathbf{r'}) (g(\mathbf{r}) + g(\mathbf{r'})) \]

(10)

where \( g(\mathbf{r}) = \omega(\mathbf{r}) R^2 + \kappa(\mathbf{r}) \) with \( R = |\mathbf{r} - \mathbf{r'}| \) distance and \( \omega(\mathbf{r}) = (\omega_c(\mathbf{r}) + \omega_r(\mathbf{r}))^{1/2} \) local frequency, where \( \omega_c(\mathbf{r}) = 3^{-1/2} \omega_0(\mathbf{r}) \) is the local dipole resonance frequency computed from the local plasma frequency: \( \omega_0(\mathbf{r}) = (4 \pi n(\mathbf{r}))^{1/2}. \omega_0(\mathbf{r}) \) is the local band gap, which might be approximated as shown in eq 10

\[ \omega_0(\mathbf{r}) = \sqrt{C} \left| \nabla n(\mathbf{r}) \right|^2 \]

\[ \int n(\mathbf{r}) d^3r \]

where \( C = 0.0093. \)

The second or \( \kappa(\mathbf{r}) \) term of \( g(\mathbf{r}) \) can be expressed as follows

\[ \kappa(\mathbf{r}) = B \frac{V_r^2(\mathbf{r})}{\omega_0(\mathbf{r})^2} = 3B \frac{\omega_0(\mathbf{r})}{k_0^2} \]

(11)

where \( V_r(\mathbf{r}) = 3(\pi n(\mathbf{r}))^{1/2} \) is the local Fermi velocity, and \( k_0 \) is the Thomas-Fermi screening wave vector. The correction can be fit to any density functional method by choosing an appropriate value for parameter \( B \) in eq 8. For revPBE, TPSS, rPWB6PBE, PBE, and B2PLYP we used \( B = 3.7, 5.0, 5.9, 6.2, \) and 8.3, respectively.\(^\text{2,7,3}\)

For double hybrid functionals we applied D2\(^\text{24} \) and D3(BJ)\(^\text{29} \) molecular mechanics dispersion corrections proposed by Grimme et al. The D2 correction contains a \( C_6 \) term (eq 12) and Tang-Toennies damping (eq 13)

\[ E_{\text{disp}}^2 = -\frac{1}{2} \sum_{i \neq j} \epsilon_{ij}(R_{ij}) \]

(12)

\[ f_{ij}(R_{ij}) = \left( \frac{1}{1 + e^{-\frac{R_{ij}}{\alpha}} - 1} \right) \]

(13)

where the global scaling factor of the energy is \( \kappa_0 = 0.4 \) for mPW2PLYP, \( R_0 \) is the sum of the atomic van der Waals radii, the scaling factor of the distance is \( \kappa_0 = 1.10 \), and the parameter in the Tang-Toennies damping is \( d = 20 \). The D3(BJ) correction contains \( C_6 \) and \( C_8 \) terms (eq 14) and Becke-Johnson damping (eq 15)

\[ E_{\text{disp}}^{\text{D3(BJ)}} = -\frac{1}{2} \sum_{i \neq j} \sum_{n=0}^{\infty} \epsilon_{ij}(R_{ij}) \]

(14)

\[ f_{ij}(R_{ij}) = a_1 R_{ij}^6 + a_2 \]

(15)

where \( R_0 = (\alpha / C_0)^{1/2} \), the global scaling factors of the energy terms are \( \{s_0, s_0\} = \{0.64, 0.9147\} \) for B2PLYP and \( \{s_0, s_0\} = \{0.82, 0.29\} \) for mPW2PLYP, while the parameters in the Becke-Johnson damping are \( \{a_1, a_2\} = \{0.3065, 5.057\} \) for B2PLYP and \( \{a_1, a_2\} = \{0, 7.3141\} \) for mPW2PLYP.

For statistical analysis we calculated the mean deviation (MD), the mean absolute deviation (MAD), the minimum and maximum deviations from the reference, and the corrected sample standard deviation (CSD) or \( s \) from the mean. In the computational chemistry literature, the MAD and the root-mean-square deviation are frequently used for method ranking. However, these are not measures for the precision of a method and consequently they might hide problems with precision. It is possible that a systematic error (e.g., systematic underbinding) is corrected with a random factor that shifts all energies toward overbinding. The MD and MAD might show considerable improvement, while the precision of the method is deteriorated. This is why we suggest CSD to estimate the unbiased sample variance.

We used Gaussian 09 Revision C.01\(^\text{75} \) to calculate the electrostatic potential maps and the static dipole polarizability tensors. We carried out basic-level density-fitted symmetry-adapted perturbation theory DF-SAPT0/jun-cc-pVDZ calculations (a modified aug-cc-pVDZ basis, with the diffuse functions on hydrogen and the diffuse \( d \) functions on heavy atoms removed)\(^\text{76-78} \) for an energy decomposition analysis. SAPT analyses for several anion–\( \pi \) (HF and TAZ) and cation–\( \pi \)–anion interaction energies were published in refs 12 and 9, respectively.

## RESULTS AND DISCUSSION

First we discuss the 20 anion–\( \pi \) interaction energies between fluoride, chloride, bromide, nitrate, or carbonate ions and HFB, TFB, TFZ, or TAZ. The anions are bound along the symmetry axis of the aromatic ring as shown in Figure 1. The fully optimized RI-MP2(frozen core)/6-31++G(d,p) coordinates were taken from ref 1. Good agreement between RI-MP2 and MP2 equilibrium distances was observed as mentioned above. The frozen core and full MP2 calculations agree well except for the \( Br^- \) complexes, where the MP2(full) optimum is shorter by around 0.1 Å.\(^\text{1,79} \) Notice that MP2(full) calculations with standard valence basis sets lead to controversial results,
and the application of the core–valence basis sets (e.g., cc-pCVTZ) is advised.\textsuperscript{79} The equilibrium distances of the center of the anions from the plane of the π system in increasing order are the following: fluoride < carbonate < nitrate < chloride < bromide and TFZ < HFB < TAZ < TFB. The shortest distance is 2.385 Å for TFZ···F\textsuperscript{−}, and the longest distance is 3.487 Å for TFB···Br\textsuperscript{−} as shown in Table S1. The distances related to carbonate, nitrate, chloride, and bromide complexes are 6%, 15%, 24%, and 30% longer on average than those of the fluoride complexes, respectively. The distances related to HFB, TAZ, and TFB complexes are 6%, 8%, and 12% longer on average than those of the TFZ complexes, respectively.

The electrostatic potential (ESP) maps of HFB, TFB, TFZ, and TAZ are shown in Figure 3. Notice the relatively large positive charge on the TFZ surface that explains partly the relatively short distances for the TFZ complexes mentioned above. Qualitatively similar ESP maps can be found for HFB in Figure 1 of ref 80 and for TFB and TAZ in Figure 3 of ref 10. The quite uneven charge distribution of TAZ is well visible in Figure 3. The central area of the TFZ surface has a higher positive potential than of the TAZ surface due to the attached三氟甲基三氮杂环己烷 (TFZ) calculated at a PBE/ATZ electron isodensity value of 0.005 au. Electrostatic potential surface energies range from –0.05 (red) to +0.05 (blue) au.

Figure 3. Electrostatic potential surfaces of hexafluorobenzene (HFB), 1,3,5-trifluorobenzene (TFB), 2,4,6-trifluoro-1,3,5-triazine (TFZ), and 1,3,5-triazine (TAZ) calculated at a PBE/ATZ electron isodensity value of 0.005 au. Electrostatic potential surface energies range from –0.05 (red) to +0.05 (blue) au.

In order to analyze the magnitude of the anion−π interaction energy components, we have performed a density-fitted symmetry-adapted perturbation theory (SAPT) analysis.\textsuperscript{83} Figure 4 shows the calculated SAPT0/jun-cc-pVDZ compo-

Table 1. The repulsive exchange terms in Figure 4 are very large, see also ref 12, and variable (on average 131% of the total interaction energy and spanning the 74−200% range). The origin of the large exchange repulsion is the close orbital overlap.\textsuperscript{12} Thus, conclusions of the earlier analyses\textsuperscript{4} that neglect

Scheme 1. Classification of the Four Aromatic Compounds with Respect to the Magnitude of $Q_{zz}$ and $\alpha_{zz}$

![Image](https://via.placeholder.com/150)

Components of the Interaction Energies

<table>
<thead>
<tr>
<th>Anion-π complexes</th>
<th>Exchange</th>
<th>Electrostatics</th>
<th>Induction</th>
<th>Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAZ</td>
<td>0.1</td>
<td>9.5</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>TFZ</td>
<td>0.1</td>
<td>9.5</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>TFB</td>
<td>0.1</td>
<td>9.5</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>HFB</td>
<td>0.1</td>
<td>9.5</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

In order to analyze the magnitude of the anion−π interaction energy components, we have performed a density-fitted symmetry-adapted perturbation theory (SAPT) analysis.\textsuperscript{83} Figure 4 shows the calculated SAPT0/jun-cc-pVDZ compo-
the exchange repulsion based exclusively on inductive and electrostatic terms might be questionable. The attractive electrostatic terms are also very large (on average 122% of the total interaction energy and spanning the 83−150% range) in agreement with the literature. Figure 4 shows that the exchange and electrostatic terms are particularly large for $5$ (HFB···CO$_3^{2−}$) and $15$ (TFZ···CO$_3^{2−}$) complexes, as the centers of the CO$_3^{2−}$ ions are quite close (at about 2.5 Å) to the aromatic ring center and the $Q_{zz}$s are large (cf. Scheme 1). For HFB and TFZ (large $Q_{zz}$s), the attractive electrostatic terms overcompensate the repulsive exchange terms, while for TFB and TAZ (small $Q_{zz}$s) the repulsive exchange terms overcompensate the attractive electrostatic terms as shown in Figure 4.

The induction energy components are smaller (on average 60% of the total interaction energy and spanning the 33−96% range; see also ref 12 for HFB and TAZ complexes). The induction is particularly important in the interactions of $6$ (TFB···Cl$^−$), $94%$ of the total interaction energy) and $20$ (TAZ···CO$_3^{2−}$, 96% of the total interaction energy). Notice that for TFB and TAZ the $Q_{zz}$s components are small (cf. Scheme 1). Figure 4 shows that the induction energy components are particularly large for CO$_3^{2−}$ containing complexes.

Several earlier studies have also suggested that dispersion forces, which are generally important in weak interactions involving aromatic rings, play only a minor role in anion−π bonding. The dispersion energy component is particularly interesting as HF calculations miss this component completely, and the semilocal DFT calculations miss the long-range part of it. In the SAPT0 energy partition the dispersion component is the difference of the SAPT0 and the HF interaction energy (the HF interaction energy is composed of exchange, electrostatic and induction components). Larger polarizability might lead to stronger dispersion interaction (larger electron correlation energy), depending on the anion−π distance. We have calculated the MP2/ATZ polarizabilities of the free fluoride, chloride, bromide, nitrate, and carbonate ions that are 1.3, 4.1, 5.9, 5.1, and 8.2 Å$_3$, respectively. Notice that, except for fluoride, the polarization values are comparable to the $\alpha_{zz}$ values (4.7−6.2 Å$_3$) of the aromatic rings discussed above. The dispersion energy components are smaller indeed but non-negligible in many complexes (on average 50% of the total interaction energy and spanning the 14−107% range, see also ref 12). The dispersion energy component of the interaction is the least important for fluoride complexes (due to the small polarizability of fluoride ions) and most important for bromide and nitrate complexes of TAZ and TFB characterized by small $Q_{zz}$ (up to 70−107%). So in these complexes the binding arises completely or almost completely from the attractive dispersion interaction. Large dispersion attraction is observed for carbonate complexes (cf. large polarizability), but the relative importance of dispersion is smaller due to the large electrostatic components arising from the two negative charges.

Table 1. Benchmark Quality DLPNO-CCSD(T)/CBS Binding Interaction Energies, ($\text{kcal mol}^{-1}$), $E_{\text{CBS}}^{\text{CCSD(T)}}$, and MP2, HF, SOSEX, RPA, dRPA, and CEPA/1 Deviations from Benchmark Energies ($E − E_{\text{CBS}}^{\text{CCSD(T)}}$) of 20 Anion−π Complexes.

<table>
<thead>
<tr>
<th>complex</th>
<th>$E_{\text{CBS}}^{\text{CCSD(T)}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{CCSD(T)}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{MP2}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{RPA}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{dRPA}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{SOSEX}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{RPA/ATZ}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{dRPA/ATZ}}$</th>
<th>$\Delta E_{\text{CBS}}^{\text{CEPA/1}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HFB + F$^−$</td>
<td>−21.08</td>
<td>2.77</td>
<td>2.59</td>
<td>0.32</td>
<td>−0.20</td>
<td>−0.63</td>
<td>−0.47</td>
<td>−0.20</td>
</tr>
<tr>
<td>2</td>
<td>HFB + Cl$^−$</td>
<td>−14.88</td>
<td>2.00</td>
<td>4.54</td>
<td>−1.31</td>
<td>0.38</td>
<td>−0.54</td>
<td>−0.14</td>
<td>−0.17</td>
</tr>
<tr>
<td>3</td>
<td>HFB + Br$^−$</td>
<td>−13.81</td>
<td>1.70</td>
<td>5.46</td>
<td>−1.20</td>
<td>0.79</td>
<td>−0.19</td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>HFB + NO$_3^−$</td>
<td>−14.89</td>
<td>2.24</td>
<td>8.94</td>
<td>−1.37</td>
<td>0.76</td>
<td>−0.90</td>
<td>−0.43</td>
<td>−0.77</td>
</tr>
<tr>
<td>5</td>
<td>HFB + CO$_3^{2−}$</td>
<td>−38.36</td>
<td>5.29</td>
<td>10.87</td>
<td>−1.77</td>
<td>0.33</td>
<td>−1.88</td>
<td>−0.95</td>
<td>−0.42</td>
</tr>
<tr>
<td>6</td>
<td>TFB + F$^−$</td>
<td>−10.86</td>
<td>3.09</td>
<td>3.58</td>
<td>0.79</td>
<td>0.98</td>
<td>0.23</td>
<td>0.55</td>
<td>0.83</td>
</tr>
<tr>
<td>7</td>
<td>TFB + Cl$^−$</td>
<td>−7.43</td>
<td>2.64</td>
<td>4.69</td>
<td>−0.50</td>
<td>1.18</td>
<td>0.18</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td>8</td>
<td>TFB + Br$^−$</td>
<td>−6.69</td>
<td>2.30</td>
<td>5.08</td>
<td>−0.72</td>
<td>1.19</td>
<td>0.16</td>
<td>0.59</td>
<td>0.04</td>
</tr>
<tr>
<td>9</td>
<td>TFB + NO$_3^−$</td>
<td>−8.66</td>
<td>3.04</td>
<td>8.50</td>
<td>−0.59</td>
<td>1.62</td>
<td>0.00</td>
<td>0.54</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>TFB + CO$_3^{2−}$</td>
<td>−24.31</td>
<td>6.98</td>
<td>10.66</td>
<td>−0.82</td>
<td>1.96</td>
<td>−0.66</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td>11</td>
<td>TFZ + F$^−$</td>
<td>−27.67</td>
<td>3.44</td>
<td>2.68</td>
<td>0.33</td>
<td>−0.37</td>
<td>−0.51</td>
<td>−0.37</td>
<td>−0.38</td>
</tr>
<tr>
<td>12</td>
<td>TFZ + Cl$^−$</td>
<td>−18.03</td>
<td>3.14</td>
<td>5.71</td>
<td>−0.94</td>
<td>1.18</td>
<td>0.34</td>
<td>0.74</td>
<td>0.32</td>
</tr>
<tr>
<td>13</td>
<td>TFZ + Br$^−$</td>
<td>−15.87</td>
<td>1.87</td>
<td>6.32</td>
<td>−1.12</td>
<td>1.36</td>
<td>0.39</td>
<td>0.82</td>
<td>0.37</td>
</tr>
<tr>
<td>14</td>
<td>TFZ + NO$_3^−$</td>
<td>−16.07</td>
<td>3.06</td>
<td>8.67</td>
<td>−1.50</td>
<td>0.70</td>
<td>−0.75</td>
<td>−0.25</td>
<td>−0.60</td>
</tr>
<tr>
<td>15</td>
<td>TFZ + CO$_3^{2−}$</td>
<td>−51.07</td>
<td>14.13</td>
<td>13.07</td>
<td>−1.76</td>
<td>0.32</td>
<td>−0.95</td>
<td>−0.50</td>
<td>−0.27</td>
</tr>
<tr>
<td>16</td>
<td>TAZ + F$^−$</td>
<td>−11.90</td>
<td>2.20</td>
<td>4.37</td>
<td>−0.01</td>
<td>0.97</td>
<td>0.06</td>
<td>0.38</td>
<td>−0.16</td>
</tr>
<tr>
<td>17</td>
<td>TAZ + Cl$^−$</td>
<td>−7.26</td>
<td>2.04</td>
<td>5.42</td>
<td>−0.96</td>
<td>1.43</td>
<td>0.33</td>
<td>0.75</td>
<td>0.11</td>
</tr>
<tr>
<td>18</td>
<td>TAZ + Br$^−$</td>
<td>−6.37</td>
<td>1.36</td>
<td>5.57</td>
<td>−0.97</td>
<td>1.48</td>
<td>0.37</td>
<td>0.79</td>
<td>0.17</td>
</tr>
<tr>
<td>19</td>
<td>TAZ + NO$_3^−$</td>
<td>−7.32</td>
<td>1.98</td>
<td>7.54</td>
<td>−1.07</td>
<td>1.38</td>
<td>−0.06</td>
<td>0.41</td>
<td>−0.23</td>
</tr>
<tr>
<td>20</td>
<td>TAZ + CO$_3^{2−}$</td>
<td>−22.42</td>
<td>5.57</td>
<td>11.58</td>
<td>−1.63</td>
<td>2.10</td>
<td>−0.09</td>
<td>0.73</td>
<td>−0.47</td>
</tr>
</tbody>
</table>

$^a$HFB = hexafluorobenzene, TFB = trifluorobenzene, TFZ = trifluorotriazine, TAZ = triazine. $^b$Domain-based local pair natural orbital (DLPNO) method was applied for approximating the CCSD(T) energies at the aug-cc-pVTZ level, and focal point calculation was performed based on the MP2/CBS results. $^c$The counterpoise correction (CP) is based on the 6-31+G(dp) level (ref 1). $^d$The complete basis set extrapolation (CBS) uses aug-cc-pVTZ and aug-cc-pVQZ results. $^e$The sum of the complete basis set extrapolated MP2 correlation and HF energies. $^f$Local pair natural orbital (LPNO) method was applied for approximating the CEPA/1 energies. $^g$MD = mean deviation from benchmark. $^h$MAD = mean absolute deviation from benchmark. $^i$CSSD = corrected sample standard deviation from the mean. $^j$Min = minimal deviation from the benchmark. $^k$Max = maximal deviation from the benchmark. $^l$The MP2 geometries used for these calculations were taken from ref 1.
Table 1 reports 20 new benchmark DLPNO-CCSD(T)/CBS binding interaction energies. Our CCSD(T) benchmark energies show reasonable agreement with nine HFB, TFB, and TFZ halide ion complex binding energies published in ref 84, but we used CCSD(T)/ATZ energies for CBS energy estimation, as shown by eq 7, instead of the less accurate and precise CCSD(T)/ADZ energies published in ref 84. Results in ref 85 show that the ATZ basis set is suitable to obtain reasonable results combined with the RI-MP2 method for various benzene and pyrazine anion- and cation-π complexes.

In Table 1 we also show the energy deviations, \( \Delta E_{MP2}^{CBS} = (E_{MP2}^{CBS} - E_{CBS}^{D(Z)}), \) where \( E_{CBS}^{D(Z)} \) is a basis set superposition error (BSSE) corrected counterpoise (CP) MP2/6-31++G(d,p) energy taken from ref 1. Earlier these energies were used as benchmark binding energies; however, inspection of the large deviations, up to 14 kcal mol\(^{-1}\), between DLPNO-CCSD(T)/CBS and CP-MP2 binding energies shown in Table 1 yields that these benchmark energies are inaccurate and imprecise, thus an earlier conclusion about the good performance of MPWB1K functional\(^{16}\) is questionable. The BSSE correction to binding energies of MP2/6-31++G(d,p), MP2/6-311++G(d,p), MP2/ADZ, and MP2/ATZ model chemistries is analyzed in ref 79, and the energies were compared to MP2/CBS(D,T) binding energies for TAZ...F\(^{-}\), Cl\(^{-}\), and Br\(^{-}\) complexes. In this work we have calculated the considerably more accurate MP2/CBS(T,Q) energies according to eq 5 as shown in Table 1. This makes possible to re-evaluate the results of ref 79. For TAZ...F\(^{-}\) the binding energy converges quickly with the basis set, the MP2/ADZ binding energy in ref 79 has a small basis set error around 0.5 kcal mol\(^{-1}\), and the MP2/AQZ binding energy in Table S2 has a negligible basis set error of 0.15 kcal mol\(^{-1}\). Applying the CP correction worsens these good and almost converged results by almost 2 kcal mol\(^{-1}\) for the ADZ basis set rendering this BSSE correction useless for this complex. Applying CP to MP2/6-31++G(d,p) binding energy that has 0.6 kcal mol\(^{-1}\) error results in inaccurate energies with errors more than 2.4 kcal mol\(^{-1}\). For TAZ...Cl\(^{-}\) the binding energy is sensitive to the basis set, and it is an example where MP2/CBS(D,T) extrapolation fails compared to MP2/CBS(T,Q) extrapolation. Due to the unreliable DZ basis set error here the D,T extrapolation goes in the opposite way (\(-8.4, -8.56,\) and \(-8.62\) kcal mol\(^{-1}\), D, T, and CBS binding energies in ref 79) than the T,Q extrapolation (\(-8.49, -8.33,\) and \(-8.22\) kcal mol\(^{-1}\), T, Q, and CBS binding energies in Table S2), thus the MP2/CBS(D,T) energy in ref 79 has 0.4 kcal mol\(^{-1}\) error. For this complex the CP correction to ADZ and ATZ basis sets is smaller, and the BSSE corrected CP-MP2/ATZ result in ref 79 has only 0.4 kcal mol\(^{-1}\) deviation. Interestingly MP2/6-31+ +G(d,p) and MP2/6-311++G(d,p) binding energies have relatively small overbinding errors (0.6 and 1.3 kcal mol\(^{-1}\), respectively), but the 3.6 and 4 kcal mol\(^{-1}\) CP corrections yield around 2 kcal mol\(^{-1}\) underbinding error making these energies unsuitable for benchmark. For TAZ...Br\(^{-}\) the comparison is more difficult as a slightly different equilibrium distance is used in refs 79 and 1, but for this complex again the CBS(D,Z)\(^{79}\) and CBS(T,Q) extrapolation (cf. Table S2) goes in the opposite way. For this complex again the BSSE corrected MP2/6-31++G(d,p), MP2/6-311++G(d,p), and MP2/ATZ binding energies show around 2 kcal mol\(^{-1}\) errors. These inaccurate and imprecise BSSE overcompensations are the origins of the large and random \( \Delta E_{CP}^{MP2} \) underbinding deviations in Table 1.

The HF/CBS deviations, \( \Delta E_{HF}^{CBS} = (E_{HF}^{CBS} - E_{CBS}^{D(Z)}), \) in Table 1, show the importance of the missing correlation energy and the vdW interaction as discussed above. The absolute value of the correlation energy is smaller for the halogen anion complexes (2.6–5.5 kcal mol\(^{-1}\)), and it is considerably larger for NO\(_3^−\) (8–9 kcal mol\(^{-1}\)) and CO\(_2^2−\) complexes (11–13 kcal mol\(^{-1}\)). Despite the short interaction distance in the fluoride complexes (cf. 2.39–2.75 Å given in ref 1), the correlation energy is only a small fraction of the total interaction energy (cf. around 2.6 kcal mol\(^{-1}\), 10–12% errors of the HF/CBS interaction energies for HFB and TFZ...F\(^−\) complexes in Table 1). These results agree with the SAPT0 results in Figure 4, and they can be explained by the small polarizability of the fluoride ion.

In agreement with the SAPT0 analysis results for TFZ...Br\(^−\), NO\(_3^−\), TAZ...Cl\(^−\), Br\(^−\), and NO\(_3^−\) complexes, the origin of the binding energy can be attributed in 75–100% to electron correlation (dispersion) effects. This can be explained by the small quadrupole moments for TFZ and TAZ and the large polarizabilities of Cl\(^−\), Br\(^−\), and NO\(_3^−\) discussed above. This fact contradicts an earlier supposition in the literature that the dispersion is not important for anion–π interactions. For HFB and TFZ...NO\(_3^−\) complexes, the correlation energy is about 55–60% of the total interaction energy, in perfect agreement with the SAPT0 dispersion energy component. The two negative charges of the carbonate ion makes the electrostatic interaction much larger than the dispersion interaction. Notice the particularly stable complexes of HFB and TFZ...CO\(_3^2−\) (at about 2.5 Å equilibrium distances) where the large positive \( Q_{ez} \) interacts with the two negative charges result in \(-38.36\) and \(-51.07\) kcal mol\(^{-1}\) interaction energies, respectively (cf. Table 1 and see also Figure 4 for SAPT0 analysis). The correlation energy vdW contributions to the interaction energies are also large, \(-10.87\) and \(-13.07\) kcal mol\(^{-1}\), respectively (cf. Table 1).

Statistical analysis of the deviations in Table 1 shows that LNO-CEPA/1/ATZ interaction energies agree best with the DLPNO-CCSD(T)/CBS energies. RPA and dRPA/ATZ perform almost equally well, and they provide only slightly worse agreement with the benchmark energies than the LNO-CEPA/1. The SOSEX performs slightly worse than the dRPA. The statistical analysis shows that the MP2/ATZ (cf. MAD = 1.64 kcal mol\(^{-1}\) and CSSD = 1.17 kcal mol\(^{-1}\) in Table S2) or the more expensive MP2/AQZ (cf. MAD = 1.22 kcal mol\(^{-1}\) and CSSD = 0.85 kcal mol\(^{-1}\) in Table S2) perform worse than the dRPA/ATZ (cf. MAD = 0.53 kcal mol\(^{-1}\) and CSSD = 0.55 kcal mol\(^{-1}\) in Table 1). Even after the CBS extrapolation the MP2 results remain worse (cf. MAD = 0.98 kcal mol\(^{-1}\) and CSSD = 0.72 kcal mol\(^{-1}\) in Table 1) than the dRPA/ATZ results. The five correlation energy calculation methods analyzed here can be classified according to the increasing deviations (MD, MAD, and CSSD) from the reference as follows: LNO-CEPA/1 < RPA \( \approx \) dRPA < SOSEX < MP2 (cf. Table 1 and Table S1 results with the ATZ basis set).

We have performed a CPU-time based efficiency analysis of the above-mentioned correlation energy calculations, but notice that such analysis is always limited by the uncertainties arising from the actual computational environment (hardware, operating system, compiler optimizations, run time environment, etc.). We have selected 3 complexes from Table 1: the smallest, complex 16, TAZ...F\(^−\), a medium sized, complex 1, HFB...F\(^−\), and one of the largest, complex 5, HFB...CO\(_2^2−\). These complexes require 391, 598, and 736 Gaussian ATZ basis functions for calculations and contain 38, 66, and 90
The correlation energy calculation methods can be classified according to the scaling with the system size as follows: MP2 ≈ dRPA ≈ SOSEX < LPNO-CCSD(T) < RPA.

We can conclude that the dRPA/ATZ is quite accurate, precise, and efficient, and it can be suggested for benchmarking anion–π interactions on moderately large systems. LPNO-CEPA/1 scales better than the dRPA, so for very large systems the former method can be more accurate, precise, and efficient.

In order to study basis set dependence, we calculated the dRPA energies of 65 hydrocarbons from CH₄ to C₆H₆ with ATZ basis sets with increasing cardinal numbers: X = {T, Q, 5, 6}. The optimized B3LYP/6-31G(2df,p) geometries were taken from Computational Chemistry Comparison and Benchmark Database. The dRPA correlation energy converges to the basis set limit as follows:

\[
E_{\text{RPA}}^\text{CBS} = E_{\text{RPA}}^\text{CBS} + AX^{-n}
\]

where the statistical average \( n = 2.566 \). We suggest a two-point CBS(T, Q) extrapolation scheme, as we did for the MP2c in eq 5 but with larger \( \beta = 0.916 \). (A larger value of \( \beta \) shows that dRPA converges to the basis set limit slower than MP2c.)

With this parameter we obtained MD = 0.016 kcal mol⁻¹ and CSSD = 0.16 kcal mol⁻¹ compared to our most precise four point \( \{T, Q, 5, 6\} \) fitted extrapolated values. For 63 hydrocarbons the value of \( \beta \) is in the range of 0.89–0.94. This relatively small uncertainty in the extrapolation might lead to small error in the extrapolated correlation energy as shown above.

Using our new extrapolation scheme, we determined dRPA/CBS interaction energies for four TAZ anion complexes from dRPA/ATZ and AQZ interaction energies (Table S3). The

Table 2. VV10 Corrected rPWPBE, TPSS, revPBE, PBE, B2PLYP, D3(BJ) Corrected PWPB95, B2PLYP, and D2 Corrected mPW2PLYP Deviations from Benchmark \((E - E^{\text{CCSD(T)}}_{\text{CBS}})\) of 20 Anion–π Complexes (kcal mol⁻¹)^[a]

<table>
<thead>
<tr>
<th>complex[^a]</th>
<th>(E^{\text{rPWPBE}}_{\text{VV}})</th>
<th>(E^{\text{TPSS}}_{\text{VV}})</th>
<th>(E^{\text{revPBE}}_{\text{VV}})</th>
<th>(E^{\text{PBE}}_{\text{VV}})</th>
<th>(E^{\text{B2PLYP}}_{\text{VV}})</th>
<th>(E^{\text{D3(BJ)}}_{\text{VV}})</th>
<th>(E^{\text{PWPB95}}_{\text{D2}})</th>
<th>(E^{\text{B2PLYP}}_{\text{D3}})</th>
<th>(E^{\text{D2}}_{\text{CBS}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFB + F⁻</td>
<td>0.41</td>
<td>0.81</td>
<td>2.09</td>
<td>1.24</td>
<td>0.03</td>
<td>0.36</td>
<td>0.83</td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td>HFB + Cl⁻</td>
<td>-0.22</td>
<td>0.19</td>
<td>0.72</td>
<td>0.42</td>
<td>-0.47</td>
<td>-0.32</td>
<td>-0.15</td>
<td>-0.38</td>
<td></td>
</tr>
<tr>
<td>HFB + Br⁻</td>
<td>1.23</td>
<td>1.58</td>
<td>1.98</td>
<td>1.79</td>
<td>-0.81</td>
<td>-0.57</td>
<td>-0.63</td>
<td>-0.79</td>
<td></td>
</tr>
<tr>
<td>HFB + NO₃⁻</td>
<td>0.46</td>
<td>0.87</td>
<td>1.29</td>
<td>1.09</td>
<td>-1.07</td>
<td>0.24</td>
<td>-0.22</td>
<td>-0.91</td>
<td></td>
</tr>
<tr>
<td>HFB + CO₃²⁻</td>
<td>-2.66</td>
<td>-3.01</td>
<td>-1.62</td>
<td>-1.90</td>
<td>-2.75</td>
<td>-1.15</td>
<td>-1.44</td>
<td>-2.35</td>
<td></td>
</tr>
<tr>
<td>TFB + F⁻</td>
<td>-0.17</td>
<td>0.45</td>
<td>1.07</td>
<td>0.50</td>
<td>0.38</td>
<td>0.64</td>
<td>1.01</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>TFB + Cl⁻</td>
<td>-0.06</td>
<td>0.49</td>
<td>0.67</td>
<td>0.49</td>
<td>0.12</td>
<td>0.47</td>
<td>0.34</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>TFB + Br⁻</td>
<td>1.20</td>
<td>1.67</td>
<td>1.80</td>
<td>1.71</td>
<td>-0.40</td>
<td>0.05</td>
<td>-0.32</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td>TFB + NO₃⁻</td>
<td>0.44</td>
<td>1.12</td>
<td>1.13</td>
<td>1.00</td>
<td>-0.20</td>
<td>1.04</td>
<td>0.40</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>TFB + CO₃²⁻</td>
<td>-9.74</td>
<td>-3.18</td>
<td>-2.82</td>
<td>-2.78</td>
<td>-1.65</td>
<td>-0.64</td>
<td>-0.68</td>
<td>-1.28</td>
<td></td>
</tr>
<tr>
<td>TFZ + F⁻</td>
<td>0.02</td>
<td>-0.07</td>
<td>1.54</td>
<td>0.55</td>
<td>-0.20</td>
<td>0.15</td>
<td>0.78</td>
<td>-0.33</td>
<td></td>
</tr>
<tr>
<td>TFZ + Cl⁻</td>
<td>0.52</td>
<td>0.71</td>
<td>1.35</td>
<td>0.98</td>
<td>0.08</td>
<td>0.58</td>
<td>0.47</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>TFZ + Br⁻</td>
<td>1.65</td>
<td>1.65</td>
<td>2.23</td>
<td>1.99</td>
<td>-0.64</td>
<td>-0.07</td>
<td>-0.40</td>
<td>-0.60</td>
<td></td>
</tr>
<tr>
<td>TFZ + NO₃⁻</td>
<td>-0.05</td>
<td>0.14</td>
<td>0.68</td>
<td>0.36</td>
<td>-1.18</td>
<td>0.03</td>
<td>-0.22</td>
<td>-1.05</td>
<td></td>
</tr>
<tr>
<td>TFZ + CO₃²⁻</td>
<td>-4.16</td>
<td>-6.93</td>
<td>-4.44</td>
<td>-4.89</td>
<td>-3.72</td>
<td>-2.83</td>
<td>-2.01</td>
<td>-2.67</td>
<td></td>
</tr>
<tr>
<td>TAZ + F⁻</td>
<td>-1.55</td>
<td>-0.98</td>
<td>-0.42</td>
<td>-1.11</td>
<td>-0.31</td>
<td>0.08</td>
<td>0.39</td>
<td>-0.41</td>
<td></td>
</tr>
<tr>
<td>TAZ + Cl⁻</td>
<td>-0.73</td>
<td>-0.20</td>
<td>-0.11</td>
<td>-0.36</td>
<td>-0.15</td>
<td>0.08</td>
<td>0.05</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>TAZ + Br⁻</td>
<td>0.46</td>
<td>0.91</td>
<td>0.96</td>
<td>0.79</td>
<td>-0.60</td>
<td>-0.31</td>
<td>-0.54</td>
<td>-0.47</td>
<td></td>
</tr>
<tr>
<td>TAZ + NO₃⁻</td>
<td>-0.88</td>
<td>-0.18</td>
<td>-0.31</td>
<td>-0.53</td>
<td>-0.77</td>
<td>0.31</td>
<td>-0.26</td>
<td>-0.61</td>
<td></td>
</tr>
<tr>
<td>MD[^b]</td>
<td>-1.32</td>
<td>-0.58</td>
<td>0.02</td>
<td>-0.31</td>
<td>-0.87</td>
<td>-0.20</td>
<td>-0.24</td>
<td>-0.71</td>
<td></td>
</tr>
<tr>
<td>MAD[^c]</td>
<td>1.96</td>
<td>1.64</td>
<td>1.73</td>
<td>1.60</td>
<td>0.93</td>
<td>0.58</td>
<td>0.67</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>CSSD[^d]</td>
<td>3.66</td>
<td>2.65</td>
<td>2.43</td>
<td>2.40</td>
<td>1.13</td>
<td>0.93</td>
<td>0.88</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>Max[^f]</td>
<td>1.65</td>
<td>1.67</td>
<td>2.23</td>
<td>1.99</td>
<td>0.93</td>
<td>1.04</td>
<td>1.01</td>
<td>0.78</td>
<td></td>
</tr>
</tbody>
</table>

[^a]: HFB = hexafluorobenzene, TFB = trifluorobenzene, TFZ = trifluorotrizine, TAZ = trizine. [^b]: MD = mean deviation from benchmark. [^c]: MAD = mean absolute deviation from benchmark. [^d]: CSSD = corrected sample standard deviation from the mean. [^e]: Min = minimal deviation from the benchmark. [^f]: Max = maximal deviation from the benchmark. [^g]: The MP2 geometries used for these calculations were taken from ref 1.
Table 3. Benchmark dRPA Interaction Energies and VV10 Corrected TPSS, revPBE, PBE, D3(BJ) Corrected B2PLYP and D2 Corrected mPW2PLYP Deviations from the Benchmark (E – E\text{dRPA}^{\text{ATZ}}) of 30 π − π′ Complexes (kcal mol\(^{-1}\))^a

<table>
<thead>
<tr>
<th>complex^a</th>
<th>(E_{\text{dRPA}}^{\text{ATZ}})</th>
<th>(\Delta E_{\text{VV}^{\text{ATZ}}},\pi − \pi')</th>
<th>(\Delta E_{\text{revPBE}-\text{VV}^{\text{ATZ}}},\pi − \pi')</th>
<th>(\Delta E_{\text{PBE}-\text{VV}^{\text{ATZ}}},\pi − \pi')</th>
<th>(\Delta E_{\text{D3(BJ)-VV}^{\text{ATZ}}},\pi − \pi')</th>
<th>(\Delta E_{\text{mPW2PLYP-D2}^{\text{ATZ}}},\pi − \pi')</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>HFB + F + TFB</td>
<td>−30.72</td>
<td>1.62</td>
<td>3.57</td>
<td>2.24</td>
<td>1.58</td>
</tr>
<tr>
<td>22</td>
<td>HFB + Cl + TFB</td>
<td>−21.47</td>
<td>0.46</td>
<td>1.14</td>
<td>0.72</td>
<td>−0.27</td>
</tr>
<tr>
<td>23</td>
<td>HFB + Br + TFB</td>
<td>−19.30</td>
<td>2.00</td>
<td>2.54</td>
<td>2.32</td>
<td>−1.83</td>
</tr>
<tr>
<td>24</td>
<td>HFB + NO(_3^-) + TFB</td>
<td>−23.08</td>
<td>2.31</td>
<td>2.69</td>
<td>2.50</td>
<td>0.04</td>
</tr>
<tr>
<td>25</td>
<td>HFB + CO(_3^{2-}) + TFB</td>
<td>−59.66</td>
<td>−1.62</td>
<td>0.04</td>
<td>−0.16</td>
<td>−0.32</td>
</tr>
<tr>
<td>26</td>
<td>HFB + F + TFZ</td>
<td>−46.02</td>
<td>2.15</td>
<td>4.75</td>
<td>3.19</td>
<td>2.08</td>
</tr>
<tr>
<td>27</td>
<td>HFB + Cl + TFZ</td>
<td>−30.65</td>
<td>0.61</td>
<td>1.79</td>
<td>1.22</td>
<td>−0.31</td>
</tr>
<tr>
<td>28</td>
<td>HFB + Br + TFZ</td>
<td>−27.34</td>
<td>2.17</td>
<td>3.05</td>
<td>2.72</td>
<td>−2.00</td>
</tr>
<tr>
<td>29</td>
<td>HFB + NO(_3^-) + TFZ</td>
<td>−29.94</td>
<td>2.16</td>
<td>2.93</td>
<td>2.60</td>
<td>0.19</td>
</tr>
<tr>
<td>30</td>
<td>HFB + CO(_3^{2-}) + TFZ</td>
<td>−79.81</td>
<td>−3.72</td>
<td>−0.68</td>
<td>−1.12</td>
<td>−0.47</td>
</tr>
<tr>
<td>31</td>
<td>HFB + F + TAZ</td>
<td>−31.72</td>
<td>0.59</td>
<td>2.44</td>
<td>1.02</td>
<td>1.11</td>
</tr>
<tr>
<td>32</td>
<td>HFB + Cl + TAZ</td>
<td>−21.15</td>
<td>−0.27</td>
<td>0.33</td>
<td>−0.15</td>
<td>−0.67</td>
</tr>
<tr>
<td>33</td>
<td>HFB + Br + TAZ</td>
<td>−18.81</td>
<td>1.22</td>
<td>1.68</td>
<td>1.38</td>
<td>−2.14</td>
</tr>
<tr>
<td>34</td>
<td>HFB + NO(_3^-) + TAZ</td>
<td>−21.73</td>
<td>1.25</td>
<td>1.48</td>
<td>1.18</td>
<td>−0.43</td>
</tr>
<tr>
<td>35</td>
<td>HFB + CO(_3^{2-}) + TAZ</td>
<td>−57.09</td>
<td>−5.64</td>
<td>−4.19</td>
<td>−4.57</td>
<td>−1.88</td>
</tr>
<tr>
<td>36</td>
<td>TFB + F + TFZ</td>
<td>−36.57</td>
<td>0.73</td>
<td>2.90</td>
<td>1.48</td>
<td>1.42</td>
</tr>
<tr>
<td>37</td>
<td>TFB + Cl + TFZ</td>
<td>−23.43</td>
<td>0.05</td>
<td>0.82</td>
<td>0.33</td>
<td>−0.55</td>
</tr>
<tr>
<td>38</td>
<td>TFB + Br + TFZ</td>
<td>−20.70</td>
<td>1.51</td>
<td>2.13</td>
<td>1.84</td>
<td>−2.00</td>
</tr>
<tr>
<td>39</td>
<td>TFB + NO(_3^-) + TFZ</td>
<td>−23.77</td>
<td>1.28</td>
<td>1.76</td>
<td>1.45</td>
<td>−0.16</td>
</tr>
<tr>
<td>40</td>
<td>TFB + CO(_3^{2-}) + TFZ</td>
<td>−68.20</td>
<td>−5.70</td>
<td>−3.40</td>
<td>−3.55</td>
<td>−1.46</td>
</tr>
<tr>
<td>41</td>
<td>TFB + F + TAZ</td>
<td>−21.21</td>
<td>−0.88</td>
<td>0.29</td>
<td>−0.90</td>
<td>0.35</td>
</tr>
<tr>
<td>42</td>
<td>TFB + Cl + TAZ</td>
<td>−13.23</td>
<td>−0.86</td>
<td>−0.61</td>
<td>−1.01</td>
<td>−0.93</td>
</tr>
<tr>
<td>43</td>
<td>TFB + Br + TAZ</td>
<td>−11.65</td>
<td>0.64</td>
<td>0.70</td>
<td>0.47</td>
<td>−2.07</td>
</tr>
<tr>
<td>44</td>
<td>TFB + NO(_3^-) + TAZ</td>
<td>−14.98</td>
<td>0.29</td>
<td>0.15</td>
<td>−0.13</td>
<td>−0.88</td>
</tr>
<tr>
<td>45</td>
<td>TFB + CO(_3^{2-}) + TAZ</td>
<td>−42.98</td>
<td>−7.87</td>
<td>−9.68</td>
<td>−7.41</td>
<td>−2.92</td>
</tr>
<tr>
<td>46</td>
<td>TFZ + F + TAZ</td>
<td>−37.49</td>
<td>−0.18</td>
<td>1.93</td>
<td>0.41</td>
<td>0.98</td>
</tr>
<tr>
<td>47</td>
<td>TFZ + Cl + TAZ</td>
<td>−23.14</td>
<td>−0.65</td>
<td>0.03</td>
<td>−0.53</td>
<td>−0.92</td>
</tr>
<tr>
<td>48</td>
<td>TFZ + Br + TAZ</td>
<td>−20.18</td>
<td>0.77</td>
<td>1.31</td>
<td>0.94</td>
<td>−2.32</td>
</tr>
<tr>
<td>49</td>
<td>TFZ + NO(_3^-) + TAZ</td>
<td>−22.37</td>
<td>0.29</td>
<td>0.64</td>
<td>0.21</td>
<td>−0.65</td>
</tr>
<tr>
<td>50</td>
<td>TFZ + CO(_3^{2-}) + TAZ</td>
<td>−65.38</td>
<td>−9.03</td>
<td>−6.76</td>
<td>−7.28</td>
<td>−2.78</td>
</tr>
</tbody>
</table>

\(^a\text{HFB = hexafluorobenzene, TFB = trifluorobenzene, TFZ = trifluorotrizine, TAZ = triazine.}\)\(^b\text{MD = mean deviation from benchmark.}\)\(^c\text{MAD = mean absolute deviation from benchmark.}\)\(^d\text{CSSD = corrected sample standard deviation from the mean.}\)\(^e\text{Min = minimal deviation from the benchmark.}\)\(^f\text{Max = maximal deviation from the benchmark.}\)\(^g\text{The MP2 geometries used for these calculations were taken from ref 1.}\)
presence of the second aromatic ring on the other side of the anions. For example the $-11.65$ kcal mol$^{-1}$ dRPA/ATZ interaction energy of the TFB···Br$^-$···TAZ ternary complex (cf. Table 3) is reasonably approximated by the sum of the TFB···Br$^-$ and the Br$^-$···TAZ interaction energies, $-6.11$ and $-5.59$ kcal mol$^{-1}$, respectively. Notice that the interaction energies are quite insensitive to the distance, as the energy surface is very flat around the minima.

\section*{CONCLUSIONS}

The correct description of the interaction between anions and aromatic rings might be used for the design of selective anion receptors and channels, and this is important for advances in the field of supramolecular chemistry. In this paper we focus on the so-called anion--π interaction. The anions are bound along the symmetry axis of the aromatic ring, as shown in Figure 1. The accurate, precise, and efficient calculation of this interaction is challenging, and inaccurate results are available in the literature. Here we propose benchmark quality solutions and detailed explanation of the origin of such interactions. The insight given in this paper might be used for future design and prediction of the properties of anion--π sandwiches and larger complexes.

We have calculated the $\alpha_{zz}$ static polarizability components parallel to the main symmetry axis of the aromatic rings (cf. Figure 1) for hexafluorobenzene (HFB), 1,3,5-trifluorobenzene (TFB), 2,4,6-trifluoro-1,3,5-triazine (TFZ), or 1,3,5-triazine (TAZ). These results and the experimental $Q_{zz}$ permanent quadrupole moments show that the four molecules fall into four different classes with respect to the magnitude of $Q_{zz}$ and $\alpha_{zz}$. TAZ(small, small), HFB(large, large), TFZ(large, small), and TFB(small, large), respectively.

We have also calculated the electrostatic potential maps that show the uneven charge distribution on the surface of the aromatic rings and the electron withdrawing effects of the various fluorine substitutions. The off center positive charge distribution on the ESP maps of TFB, TFZ, and TAZ show that displaced anions might be stabilized. This explains the displaced anions observed in several crystal structures.

We have performed a symmetry-adapted perturbation theory analysis (SAPT0/jun-cc-pVDZ), DLPNO-CCSD(T)/CBS, MP2, HF, SOSEX, RPA, dRPA, LPH-CEPA-1/1, VV10 corrected rPWPE, TPSS, revPBE, PBE, B2PLYP, D3(BJ) corrected PWPB95, B2LYP, and D2 corrected mPW2PLYP calculations for 20 fluoride, chloride, bromide, nitrate, or carbonate complexes of HFB, TFB, TFZ, or TAZ.

Our SAPT results show that the repulsive exchange and the attractive electrostatic, induction, and dispersion energy components are on average $-131\%$, $122\%$, $60\%$, and $50\%$ of the total interaction energy, respectively. The repulsive exchange, the attractive electrostatic, induction, and dispersion energy components vary from $-74\%$ to $-200\%$, $83\%$ to $150\%$, $33\%$ to $96\%$, and $14\%$ to $107\%$ of the total interaction energy, respectively. The exchange and electrostatic terms are particularly large for HFB···CO$_3^{2-}$ and TFZ···CO$_3^{2-}$ complexes because the center of the CO$_3^{2-}$ ion is relatively close to the aromatic ring center and HFB and TFZ Q$_{zz}$s are large. The induction is particularly important in the interactions of TFB···CI$^-$ and TAZ···CO$_3^{2-}$, as the Q$_{zz}$ components of TFB and TAZ are small. The vDW dispersion energy components are relatively small in the fluoride complexes, and they are particularly important for bromide and nitrate complexes of TFB and TAZ. Differences of DLPNO-CCSD(T)/CBS and
HF/CBS energies show good agreement with SAPT0 dispersion energy components.

Our results show that the investigated correlation energy calculation methods can be classified according to increasing deviations from the reference DLPNO-CCSD(T)/CBS interaction energies as LPNO-CEPA/1 < dRPA ≈ RPA < SOSEX < MP2. These density-fitted methods can be classified according to the increasing CPU times as MP2 ≪ dRPA ≈ SOSEX < LPNO-CEPA/1 ≪ DLPNO-CCSD(T) < RPA and according to the scaling with the system size as LPNO-CEPA/1 < MP2 ≪ dRPA ≈ SOSEX < DLPNO-CCSD(T) < RPA, all with augmented triple-ζ (ATZ) basis set.

We can conclude that the dRPA/ATZ is quite accurate, precise, and efficient and it can be suggested for benchmarking anion−π interactions on moderately large systems where DLPNO-CCSD(T) is too expensive. We show that the slight underbinding error of dRPA/CBS (0.6−1.3 kcal mol−1) is efficiently compensated by the comparable overbinding error of the ATZ basis set. LPNO-CEPA/1 scales better than the dRPA, so for very large systems it can be more accurate, precise, and efficient. According to our results in agreement with ref 84, the counterpoise corrected (CP) MP2/6-31++G(d,p) energies suggested in ref 1 are not suitable for benchmarking.

We used our new benchmark energies to rank various approximate density functionals. Comparison of mean absolute deviations (MAD) and precisions (CSSD) shows the following ranking from the best to worst: PWPB95−D3(BJ) < B2PLYP−D3(BJ) < mPW2PLYP−D2 < B2PLYP−V10 ≈ MP2−V10 ≈ TPSS−V10 < revPBE−V10 ≈ rPWB1PBE−V10. All calculations used the ATZ basis set. The double hybrid functionals with dispersion correction perform best, almost reaching the accuracy of the dRPA/ATZ binding energies but not the precision. Without dispersion correction all functionals underbind and show serious errors. We have also shown that the precision (CSSD) of the investigated D3(BJ) dispersion corrected double hybrid functionals deteriorate for the current set of complexes.

For 30 π−anion−π′ sandwich complexes we have calculated dRPA/ATZ binding energies for benchmarks. For the investigated density functionals, we obtained the following ranking from the best to worst: B2PLYP−D3(BJ) < mPW2PLYP−D2 < B2PLYP−V10 < TPSS−V10 < revPBE−V10. This ranking for π−anion−π′ sandwich complexes is very similar to the previous ranking for anion−π complexes, and this supports some kind of additivity effects.

Systematic comparison of our dRPA/ATZ energies shows that the ternary interaction energy can be calculated from the two corresponding binary interaction energies with negligible error. Consequently the energies are practically additive. The origin of this is the near equality of the equilibrium distances in the binary π−anion and the ternary π−anion−π′ as observed earlier. Moreover the interaction energies are quite insensitive to the small variations of the π−anion−π′ distances as the energy surface is sufficiently flat around the minima.

For future benchmarking of various efficient methods, we suggest using our new interaction energies. Among the investigated methods, the dRPA/ATZ model chemistry is quite accurate, precise, and efficient so we suggest using it for solving chemical and biological problems of similar size as those studied in this paper. If accuracy is more important than efficiency, then the more accurate LPNO-CEPA/1/ATZ is suggested, and because of the excellent scaling with size it might be more efficient for sufficiently large systems than dRPA. The most efficient but somewhat less precise methods are among the double hybrid functionals; for the present test set the PWPB95−D3(BJ) and B2PLYP−D3(BJ) are the best performers.

© 2015 American Chemical Society. All rights reserved.
Accurate Diels–Alder Reaction Energies from Efficient Density Functional Calculations

Pál D. Mezei,† Gábor I. Csonka,*† and Mihály Kállay‡

†Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary
‡MTA-BME Lendület Quantum Chemistry Research Group, Department of Physical Chemistry and Materials Science, Budapest University of Technology and Economics, H-1521 Budapest, P.O. Box 91, Hungary

ABSTRACT: We assess the performance of the semilocal PBE functional; its global hybrid variants; the highly parametrized empirical M06-2X and M08-SO; the range separated rCAM-B3LYP and MCY3; the atom-pairwise or nonlocal dispersion corrected semilocal PBE and TPSS; the dispersion corrected range-separated ωB97X-D; the dispersion corrected double hybrids such as PWPB95-D3; the direct random phase approximation, dRPA, with Hartree–Fock, Perdew–Burke–Ernzerhof, and Perdew–Burke–Ernzerhof hybrid reference orbitals and the RPAX2 method based on a Perdew–Burke–Ernzerhof exchange reference orbitals for the Diels–Alder, DARC; and self-interaction error sensitive, SIE11, reaction energy test sets with large, augmented correlation consistent valence basis sets. The dRPA energies for the DARC test set are extrapolated to the complete basis set limit. CCSD(T)/CBS energies were used as a reference. The standard global hybrid functionals show general improvements over the typical endothermic energy error of semilocal functionals, but despite the increased accuracy the precision of the methods increases only slightly, and thus all reaction energies are simply shifted into the exothermic direction. Dispersion corrections give mixed results for the DARC test set. Vydrov–Van Voorhis 10 correction to the reaction energies gives superior quality results compared to the too-small D3 correction. Functionals parametrized for energies of noncovalent interactions like M08-SO give reasonable results without any dispersion correction. The dRPA method that seamlessly and theoretically correctly includes noncovalent interaction energies gives excellent results with properly chosen reference orbitals. As the results for the SIE11 test set and H2+ dissociation error show that the dRPA methods suffer from delocalization error, good reaction energies for the DARC test set from a given method do not prove that the method is free from delocalization error. The RPAX2 method shows good performance for the DARC, the SIE11 test sets, and for the H2+ and H2 potential energy curves showing no one-electron self-interaction error and reduced static correlation errors at the same time. We also suggest simplified DARC6 and SIE9 test sets for future benchmarking.

INTRODUCTION

The inexpensive local spin density (LSD) approximation,1 the semilocal generalized gradient approximation (GGA) density functionals, e.g., PW86,2 PW91,3 PBE,4 and PBEsol;5 and meta-GGA functionals, e.g., TPSS,6 revTPSS,7 and regTPSS,8 suffer from many-electron self-interaction error (SIE)9 leading to energetic preference for unrealistically delocalized electron densities. Functionals having SIE show particularly large errors for systems with fractional charges,12,13 as too-delocalized charge distribution leads to too-low energies. For the same reason, SIE overestimates the charge transfer complexes, leads to too-low or no reaction energy barriers14 of chemical reactions, and leads to seriously wrong dissociation energy curves for diatomic cations (e.g., H2+).9 The correct description requires that the ground state total energy E(N) versus number of electrons should be a linkage of straight-line segments connecting the energy values (e.g., E(N − 1), E(N), E(N + 1)) at integer N’s.15 All standard local and semilocal functionals give erroneously smooth convex, almost parabolic energy curves and show no derivative discontinuity at integer electron numbers.16 Exact exchange has the opposite behavior: it gives the opposite concave energy curve between two integers, due to the missing electron correlation.16 Mixing the convex and the concave energy curves via hybridization of density functionals might help; however, in practice the popular hybrid functionals still show considerable convexity error.16 These hybrids can be classified as global (e.g., B3PW91 or B3LYP,17 PBE0,18 and TPSSH19 and variants of BPW91, PBE, and PBEsol hybrids20), local (e.g., B05,21 MCY1,22 MCY2,23 and PSTS24), or range-separated (e.g., HSE03,25 oB97 or oB97X,26 MCY3,27 CAM-B3LYP,28 or rCAM-B3LYP27). The standard global hybrid functionals that mix 20–25% of exact exchange correct only a fraction of the delocalization error (Figure 4 in ref 16) and fail seriously for delocalized stretched odd-electron...
systems (e.g., dissociating H₂⁺). It was found that only MCY3 and rCAM-B3LYP showed signs of improvement in the description of fractional numbers of electrons.

To illustrate the above-discussed SIE or delocalization error, Johnson et al. proposed a test set of 14 representative Diels–Alder reactions. These are the reactions of butadiene, cyclopentadiene, cyclohexadiene, and furane with typical dienophiles as ethene, ethyne, maleic anhydride, and maleimide, as shown in Table 1.

Table 1. Reagents in the 14 Diels–Alder Reactions in the DARC Test Set

<table>
<thead>
<tr>
<th>no.</th>
<th>reagents</th>
<th>products</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ethene + butadiene</td>
<td>monocyclic</td>
</tr>
<tr>
<td>2</td>
<td>ethyne + butadiene</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ethene + cyclopentadiene</td>
<td>bicyclic</td>
</tr>
<tr>
<td>4</td>
<td>ethyne + cyclopentadiene</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ethene + cyclohexadiene</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ethyne + cyclohexadiene</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>furan + maleic anhydride</td>
<td>(endo) tricyclic</td>
</tr>
<tr>
<td>8</td>
<td>furan + maleic anhydride</td>
<td>(exo)</td>
</tr>
<tr>
<td>9</td>
<td>furan + maleimide</td>
<td>(endo)</td>
</tr>
<tr>
<td>10</td>
<td>furan + maleimide</td>
<td>(exo)</td>
</tr>
<tr>
<td>11</td>
<td>cyclopentadiene + maleic anhydride</td>
<td>(endo)</td>
</tr>
<tr>
<td>12</td>
<td>cyclopentadiene + maleic anhydride</td>
<td>(exo)</td>
</tr>
<tr>
<td>13</td>
<td>cyclopentadiene + maleimide</td>
<td>(endo)</td>
</tr>
<tr>
<td>14</td>
<td>cyclopentadiene + maleimide</td>
<td>(exo)</td>
</tr>
</tbody>
</table>

*Reactions 7–14 lead to tricyclic products with two different endo or exo conformations as noted in parentheses.*

This test set is called DARC, and it was included in a larger test set called GMTKN30. Inspection of the reactions reveals that intramolecular noncovalent interaction occurs in the larger bicyclic and tricyclic products. Semilocal functionals miss completely the long-range part of the correlation energy, the origin of noncovalent dispersion bonding. And this error cannot be alleviated by the mixing of the exact exchange into the functional as the exact exchange misses completely the electron correlation effects. To include the missing intramolecular attractive dispersion effects, a variety of atom-pairwise dispersion corrections were applied (e.g., DFT-D3, DFT-D2, DFT-D3, DFT-D3(BJ), D3, dD10, dDsC, and dDXDM). A different approach is applied by the nonlocal van der Waals density functionals (e.g., vDW-DF, vDW-DF2, vDW-DF3, VV09, and VV10). The Minnesota functionals (e.g., M06-L and M11-L) have a flexible meta-GGA form with many parameters to describe main group thermochemistry, kinetics, and noncovalent interactions. Minnesota functionals might improve the medium-range interaction energies; however, they miss the long-range dispersion effects due to exponential decay of the calculated interaction energy with the distance. The Minnesota functionals have several global hybrid variants (e.g., M05, M05-2X, M06, M06-2X, M08-HX, and M08-SO). The global double hybrid functionals (e.g., B2PLYP, DSD-BLYP, and PWPP95) yield improved results for reaction energies. But even these elaborate functionals that include second order perturbation theory require dispersion correction for long-range dispersion effects as in B2PLYP-D3, PWPP95-D3, and DSD-BLYP-D3.

RPA is accurate for noncovalent intra- and intermolecular interactions, and in contrast with the dispersion corrected functionals it captures the nonpairwise-additive nature of the dispersion interactions. It performs moderately for covalent bond rearrangement reactions and gives erroneous short-range correlation and therefore fails seriously for atomization and ionization processes. It also suffers from self-interaction and static correlation error, which lead to the underestimation of barrier heights and to the overestimation of the energy of breaking bonds, respectively. Furthermore, it converges very slowly to the complete basis set limit because the smooth orbital products cannot describe well the electron−electron Coulomb cusps. We also include in this study the SIE11 self-interaction error related test set which exclusively deals with the SIE. It contains 11 reaction energies that are especially sensitive to SIE, including the dissociation energies of five cationic reactions (M⁺ → M + M⁺, where M = He, H₂O, and NH₃, C₂H₄⁺ → C₂H₂ + C₂H₅⁺, and (CH₃CO⁺ → CH₃ + CH₃CO⁻) and six neutral reactions (CIFCI → CICIF, C₂H₄⋯F₂ → C₂H₄ + F₂, benzene⋯Li → benzene + Li, NaOMg → Na + MgO, and PdLiF → Li + F₂). The reference reaction energies were obtained by CCSD(T)/CBS extrapolation. Earlier studies show that many functionals perform relatively poorly for this test set. One of the worst result is given by PBE (cf., Figure 1 in ref S8). The meta-GGA functionals improve over PBE noticeably. Despite its optimization of the parameters, oTPSS gives only a negligible improvement over TPSS, as the many-electron self-interaction error cannot be corrected at the level of semilocal approximation. Among the semilocal functionals, MGGA_MS gives the smallest error. The hybrid functionals perform better, as mixing of exact exchange decreases the electron self-interaction error for exchange, but do not deliver good results. The best performance for the SIE11 test set was observed for the DSD-BLYP-D3 dispersion corrected doubly hybrid functional.

### COMPUTATIONAL DETAILS

The exact exchange dependence of the PBE global hybrid functionals was tested using Gaussian. The calculations, we used an efficient triple-ζ basis set called (aug)-cc-pVTZ(-f,-d). It uses the cc-pVTZ(-d) basis set for the hydrogen atoms and the aug-cc-pVTZ(-f) basis set for the heavy atoms. Our PBE reaction energies with this efficient basis are lower approximately by 0.3 kcal mol⁻¹ than those with aug-cc-pVTZ (ATZ), and the calculations are considerably faster. The VV10 dispersion corrected density functional results (with ATZ basis set) were obtained with Orca 3.0.1. The DSD-PBE86 energies (with aug-cc-pVQZ basis set noted as AQZ) were calculated using the MRCC program, with the D3(BJ) dispersion corrections obtained separately from Orca. The dD10 dispersion corrections (applied on the de2-QZVP energies taken from ref 30) were calculated using our own program. The dRPA, SOSEX, and RPAX2 correlation energies are calculated in the MRCC program code using the efficient algorithm of Heilman. In the Appendix, we give a detailed presentation of these methods. These calculations are very efficient with the ATZ basis set; for example, for one of the largest DARC reaction products (P10), the PBE and TPSS computation times were 5.5 and 7 min, and those of the PBE0, M06-2X, and PWPP95 were 21, 22, and 24 min. The calculation of dispersion corrections took only seconds. The calculation of the dRPA and RPAX2 correlation energies took 10 and 22 min. More details about the efficiency of dRPA can be found in our earlier work. There are no practical self-consistent dRPA implementations in the density functional framework for single-particle reference orbitals (RO). These calculations are commonly performed in a postprocessing way, where single-particle ROs from a self-consistent...
DFT or HF calculation are used to evaluate both the EX and cRPA terms. Such noniterative density functional calculations are called dRPA@RO or (EX + cRPA)@RO in the literature, where EX + cRPA is an orbital-dependent functional of exact exchange and RPA correlation.65 In order to distinguish our results yielded by the Ricatti equation with density fitting and Cholesky decomposition, we adopt the nomenclature of dRPA@RO.

We extrapolate the exact exchange and the dRPA correlation energies (see the equations in the Appendix) to the complete basis set (CBS) according to eqs 1 and 2:

\[
E_{\text{exact}}^{\text{CBS}} = E_{\text{exact}}^{\text{(AQZ)}} + C_{4,3}^{\text{exact}}(E_{\text{exact}}^{\text{(AQZ)}} - E_{\text{exact}}^{\text{(ATZ)}})
\]

(1)

\[
E_{\text{dRPA@RO}}^{\text{CBS}} = E_{\text{dRPA@RO}}^{\text{(AQZ)}} + C_{4,3}^{\text{dRPA@RO}}(E_{\text{dRPA@RO}}^{\text{(AQZ)}} - E_{\text{dRPA@RO}}^{\text{(ATZ)}})
\]

(2)

where \(C_{4,3}^{\text{exactHF}} = 0.274^{66}\) and \(C_{4,3}^{\text{dRPA@RO}}\) depends on the choice of the self-consistent reference orbitals. For PBE, PBE0.25, and HF reference orbitals, \(C_{4,3}^{\text{dRPA@RO}(PBE)} = 0.856, C_{4,3}^{\text{dRPA@RO(PBE0.25)}} = 0.867,\) and \(C_{4,3}^{\text{dRPA@RO(HF)}} = 0.917\), respectively (for details see ref 67). These values were obtained from a basis set convergence study of the dRPA correlation energies of ethyne and ethene for A\(_{\alpha}\) values were obtained from a basis set convergence study of the DARC Diels–Alder reaction test set using the (aug)-cc-pVTZ(-f,-d) basis set with cardinal numbers \(X = 3–6\) by finding an optimal \(\alpha\):

\[
C_{\text{dRPA@RO}}^{\alpha} = \frac{1}{\left(\frac{X}{X-1}\right)^\alpha - 1}
\]

(3)

This corresponds to an inverse power basis set convergence of the \(E_{\text{dRPA@RO}}\) controlled by \(\alpha^{66}\). The CBS(S/4) extrapolations by the \(H_2^+\) and \(H_2\) reference potential energy curves were performed according to eqs 4 and 6 of ref 66 for the HF and CISD correlation energies, respectively.

\section*{RESULTS AND DISCUSSION}

Inspection of the 14 Diels–Alder reactions in the DARC database reveals that, in the first six reactions, double (ethene) and triple (ethyne) bonds transformed into single and double bonds. The electron delocalization in the dienes is destroyed; moreover, in reactions 3–6, bicyclic products are formed. Thus, intramolecular dispersion effects occur compared to the first two reactions where the product is simple monocyclic (cf. Table 1). In the last eight reactions, furane or cyclopentadiene reacts with maleine, and maleimide yields tricyclic products with endo and exo conformations (cf. Table 1). All these features of the products suggest that these reactions require a method that describes the noncovalent intramolecular interactions, the bond multiplicity change, and new bond formation energies correctly. The energetic consequences of the delocalization error (SIE) can be overcompensated by many different factors listed above. In this section, we shall analyze the various error sources and give methods to reliably eliminate those errors.

First, we examine the role of the exact exchange in hybrid functionals. Figure 1 shows the deviations from the reference reaction energies based on CCSD(T)/CBS calculations for PBE and global hybrid PBE0.25, PBE0.32, and PBE0.38 with increasing weight of exact exchange. Table S1 shows the PBE hybrid (aug)-cc-pVTZ(-f,-d) deviations from the benchmark reaction energies. The def2-QZVP results are also presented in Table S1 for comparison. The results in Figure 1 show that increasing the weight of the exact exchange shifts the reaction energies in a negative, exothermic direction but does not improve the precision (CSSD) considerably (the tendency is similar for TPSS hybrids in GMTKN30 \textsuperscript{29} database). Figure 1 shows that PBE gives quite good reaction energies for the first two reactions yielding monocyclic products (cyclohexene and 1,4-cyclohexadiene). However, for reactions leading to bicyclic or tricyclic products, PBE shows a typical DFT error: the reaction energies are erroneously shifted to an endothermic (underbinding) direction. This error is particularly large (more than 8 kcal mol\textsuperscript{-1}) for the tricyclic products (see reactions 7–14 in Figure 1). In these products, the interactions of the bridgehead atoms are intramolecular noncovalent interactions at relatively short distances. The energy errors are related to the tendency of the functionals to overestimate noncovalent repulsion. Johnson et al.\textsuperscript{16} supposed that, in the products, there is a region of highly localized electron density that is understabilized by most of the semilocal functionals like PBE in our example.\textsuperscript{18} They observed improvement in the calculated reaction energies for functionals with improved treatment of fractionally charged systems. Because of this, they supposed that the errors are related to SIE, the electron delocalization error in DFT, as this error causes semilocal functionals to give too-high energy for localized states or too-low energy for delocalized states. The rCAM-B3LYP and MCY3 functionals, designed to minimize delocalization error, give good results for these reactions (MD = −2.6 and −3.4 kcal mol\textsuperscript{-1} and MAD = 2.6 and 3.4 kcal mol\textsuperscript{-1}, respectively). This logic supposes that the electrons are more localized in the products of the Diels–Alder reactions than in the reactants. Notice that the Diels–Alder reactions are exothermic; thus the large product molecules with an extended electronic system are more stable than the reactants.

Our analysis shows that the self-consistent HF exact exchange electron densities are more compact (localized) than the PBE electron densities in the same molecules with the same nuclear arrangement. Thus, as we increase the weight of the exact exchange in a hybrid functional, we increase the localization of the electron density, and that shifts the reaction energies in the exothermic directions. Thus, the products stabilize more than the reactants. Our results show that we can obtain quite accurate reaction energies using common global PBE hybrid functionals that contain considerable delocalization error. The most accurate is the PBE0.25 hybrid, but the most precise is the PBE0.38 (compare the MD values for accuracy and the CSSD values for precision). The PBE0.32 hybrid yields the best MAD (2.84 kcal mol\textsuperscript{-1}) among these functionals with relatively good accuracy and precision. PBE0.32 hybrid reaction energies are considerably better than those obtained from MCY3 and only slightly worse...
than the rCAM-B3LYP results. In the reactions of ethene, the effect of 25% exact exchange is larger than in the similar reactions of ethyne. The effect of exact exchange is even larger in the reactions of maleic anhydride and maleimide (reaction 7–14 in Figure 1), where the reacting carbon–carbon double bond is a part of a larger conjugated system. Figure 1 shows that the reactions leading to mono-, bi-, and tricyclic products behave quite differently, and the increasing fraction of the exact exchange simply shifts the errors in the exothermic direction.

Next, we discuss the dispersion corrected DFT results. As attractive intramolecular dispersion interactions stabilize the bicyclic and tricyclic products more than the reactants, the calculated dispersion corrected Diels–Alder reaction energy is always shifted in the exothermic direction. Thus, this kind of dispersion correction, the exothermic shift of accurate (small MD) or already exothermic DFT reaction energies, worsens the results (vide infra PBE0.3220 or 38 results). In Figure 2, we show

![Figure 2](image)

Figure 2. The D3 dispersion corrected (taken from GMTKN30 database29) as well as the nonlocal VV10 corrected (this work, using ATZ basis set) PBE and TPSS deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set.

the deviations of D3 and VV10 corrected PBE and TPSS/ATZ reaction energies from the reference. The energies were taken from the GMTKN30 database.29 Furthermore, the PBE- and TPSS-D2, -D3, -dD10, and -VV10 deviations from the benchmark reaction energies are presented in Tables S2 and S3 for comparison. The simpler D2 corrected methods seem to be superior to the D3 corrected ones for Diels–Alder reaction energies. The dD10 correction performs similarly to the D2 correction for PBE. The VV10 correction has an even larger effect. Since TPSS is more repulsive than PBE, the corrections are also larger for TPSS than PBE. The dispersion corrected TPSS results are also more precise. These results show that the one-electron self-correlation free TPSS has also a smaller many-electron self-interaction error (cf. SIE11 subset of GMTKN30 database29) and this leads to improvements over the dispersion corrected PBE for DARC test set. The dispersion corrected TPSS-dD10 results are even better as MAD = 2.17 kcal mol⁻¹, and the results are quite precise (CSSD = 1.86 kcal mol⁻¹). The VV10 correction for TPSS increases the accuracy of the calculations more than the atom-pairwise corrections.

Figures 1 and 2 show clearly that semilocal functionals yield different errors for reactions with ethene and ethyne, and mixing with exact exchange and applying dispersion correction does not help in this respect. The deviations for the reactions 7–10 or 11–14 are similar because the intramolecular interactions are nearly equal in the cage-like products. In the database, there are many redundant reactions, and the tricyclic reactions are over-represented. Figures 1 and 2 also show that the methods have very similar errors for reactions 7–10 and 11–14. We suggest using a more balanced, smaller database with only six reactions (reactions 1–4, 7, and 11 in Table 1), called DARC6.

Next, we assess the performance of the hybrid functionals that also consider dispersion (Table S4). The dispersion correction shifts the reaction energies in the exothermic direction by stabilizing the products more than the reactants. Consequently, for all methods that reproduce accurately the reaction energies (MD value is around zero) or yield exothermic error (MD value is negative), the accuracy of the dispersion-corrected reaction energies is worsened systematically by 2−4 kcal mol⁻¹. For example, the PBE0.25-D3 results in MD = −3.08 and MAD = 3.13 kcal mol⁻¹, a worsening compared to the PBE0.25 results above. Figure 3 shows the deviations from reference energies for

![Figure 3](image)

Figure 3. The global hybrid M06-2X, the dispersion corrected range separated hybrid oB97X-D, and the D3 dispersion corrected global double hybrid PWPB95 (taken from GMTKN30 database29) deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set.

oB97X-D, M06-2X, and PWPB95-D3 (taken from GMTKN30 database29). The dispersion-corrected range-separated hybrid oB97X-D with 100% of exact exchange in the long range is accurate (MD = 0.41 kcal mol⁻¹) but not precise (CSSD = 2.49 kcal mol⁻¹), leading to MAD = 1.98 kcal mol⁻¹. The M06-2X functional (with 54% of exact exchange) is less accurate (MD = 1.92 kcal mol⁻¹), but it is significantly more precise (CSSD = 1.97 kcal mol⁻¹) than oB97X-D. Overall, the M06-2X is somewhat worse (MAD = 2.45 kcal mol⁻¹) than oB97X-D. Somewhat better results can be obtained with the M08-SO functional yielding MD, MAD, and CSSD equal to 0.41, 1.46, and 1.69 kcal mol⁻¹, respectively. The dispersion corrected global double hybrid PWPB95-D3 gives MD, MAD, and CSSD equal to 0.88, 1.52, and 1.59 kcal mol⁻¹, respectively. These latter methods are among the best DFT methods in the literature for the DARC test set.

Finally, we assess the performance of the RPA methods (Tables S5 and S6). dRPA seamlessly integrates dispersion interactions. It can treat chemical reaction energies with good precision, but it contains the SIE. It is interesting how dRPA reproduces these quite complicated bond rearrangements, hybridization changes, and ring formations characteristic to Diels–Alder reactions. It is expected that the intramolecular dispersion effects are well described. A posteriori dispersion corrections like D2 or D3 are practical, but several features of such corrections are undesirable, as such corrections are parametrized empirically and might lead to random corrections as it was experienced for anion−π interactions.21 Even the VV10 correction is pairwise, thus inherently limited. dRPA does not suffer from these problems, and it is very efficiently implemented
The complete basis set extrapolated dRPA (calculated on self-consistent HF, PBE, PBE0.25 orbitals) deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set.

Figure 4. The complete basis set extrapolated dRPA energies (calculated on self-consistent HF, PBE, PBE0.25 orbitals) deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set.

Table 2. Reference Reaction Energies Obtained from CCSD(T)/CBS (kcal mol$^{-1}$) Deviations for Various Calculated Reaction Energies (calculated − reference) and Statistical Data As Mean Deviation, Mean Absolute Deviations, Corrected Sample Standard Deviations, and Minimal and Maximal Deviations for 11 Reactions of SIE11 Test Set$^{a}$

<table>
<thead>
<tr>
<th>react.</th>
<th>ref. energy</th>
<th>PBE0.25</th>
<th>DSD-BLYP-D3</th>
<th>dRPA: HF</th>
<th>SOSEX: HF</th>
<th>dRPA: PBE</th>
<th>dRPA: PBE0.25</th>
<th>RPAX2: PBE$^{x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57.44</td>
<td>12.58</td>
<td>5.68</td>
<td>2.29</td>
<td>−5.82</td>
<td>13.15</td>
<td>9.40</td>
<td>−3.43</td>
</tr>
<tr>
<td>2</td>
<td>35.34</td>
<td>6.25</td>
<td>2.24</td>
<td>−2.47</td>
<td>−6.22</td>
<td>4.56</td>
<td>2.57</td>
<td>−1.79</td>
</tr>
<tr>
<td>3</td>
<td>37.25</td>
<td>10.99</td>
<td>5.45</td>
<td>−0.95</td>
<td>−7.52</td>
<td>9.81</td>
<td>6.85</td>
<td>−1.07</td>
</tr>
<tr>
<td>4</td>
<td>35.28</td>
<td>3.75</td>
<td>0.24</td>
<td>−2.12</td>
<td>−5.68</td>
<td>3.52</td>
<td>2.27</td>
<td>−1.57</td>
</tr>
<tr>
<td>5</td>
<td>22.57</td>
<td>7.74</td>
<td>−2.47</td>
<td>−0.87</td>
<td>1.43</td>
<td>0.99</td>
<td>1.43</td>
<td>0.06</td>
</tr>
<tr>
<td>6</td>
<td>−1.01</td>
<td>11.48</td>
<td>−1.22</td>
<td>−16.36</td>
<td>−14.89</td>
<td>4.29</td>
<td>1.10</td>
<td>0.80</td>
</tr>
<tr>
<td>7</td>
<td>1.08</td>
<td>0.03</td>
<td>−0.07</td>
<td>−0.26</td>
<td>−0.54</td>
<td>−2.86</td>
<td>−0.60</td>
<td>−2.81</td>
</tr>
<tr>
<td>8</td>
<td>9.50</td>
<td>−4.09</td>
<td>−7.52</td>
<td>−11.36</td>
<td>−12.61</td>
<td>−10.36</td>
<td>−9.11</td>
<td>−7.24</td>
</tr>
<tr>
<td>9</td>
<td>10.50</td>
<td>3.08</td>
<td>1.13</td>
<td>−2.20</td>
<td>−2.43</td>
<td>−2.13</td>
<td>−1.15</td>
<td>−1.54</td>
</tr>
<tr>
<td>10</td>
<td>69.56</td>
<td>4.73</td>
<td>−1.36</td>
<td>8.63</td>
<td>13.23</td>
<td>−9.26</td>
<td>−0.76</td>
<td>−6.36</td>
</tr>
<tr>
<td>11</td>
<td>94.36</td>
<td>13.31</td>
<td>5.75</td>
<td>11.62</td>
<td>−4.68</td>
<td>42.75</td>
<td>26.46</td>
<td>−6.18</td>
</tr>
<tr>
<td>MD</td>
<td>6.35</td>
<td>0.71</td>
<td>−1.28</td>
<td>−4.16</td>
<td>4.95</td>
<td>3.50</td>
<td>−2.83</td>
<td></td>
</tr>
<tr>
<td>MAD</td>
<td>7.09</td>
<td>3.01</td>
<td>5.38</td>
<td>6.82</td>
<td>9.43</td>
<td>5.61</td>
<td>2.98</td>
<td></td>
</tr>
<tr>
<td>CSSD</td>
<td>5.53</td>
<td>4.02</td>
<td>7.82</td>
<td>7.47</td>
<td>14.44</td>
<td>8.95</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>−4.09</td>
<td>−7.52</td>
<td>−16.36</td>
<td>−14.89</td>
<td>−10.36</td>
<td>−9.11</td>
<td>−7.24</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>13.31</td>
<td>5.75</td>
<td>11.62</td>
<td>13.23</td>
<td>42.75</td>
<td>26.46</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

$^{a}$ All RPA calculations were performed with the ATZ basis set. PBE0.25 and DSD-BLYP-D3 results are taken from ref 30.
are considerably smaller for the rest of the reactions. Even using the PBE0.25 hybrid reference yields quite a large (about 26.5 kcal mol\(^{-1}\)) error for this reaction, while the results for the other 10 reactions are reasonable, as shown in Table 2. Standard hybrid PBE0.25 results in Table 2 are less accurate but more precise than the dRPA:PBE0.25 results. The RPAX2:PBE\(_{\text{Ex}}\) method shows errors above 6 Å are due to the fractional charge problem leading to the direction of dRPA:HF < dRPA:PBE0.25 < dRPA:PBE. The huge underbinding error of the dRPA:PBE (cf. Table 4) is particularly large errors for reactions 1, 3, 8, and 11. The correct descriptions of reactions 8 and 11 are particularly difficult for the methods presented here. The deviations in Table 2 show that SIE does not manifest well in reaction 7. Notice also that dRPA performs particularly well for reaction 5; consequently the remaining nine reactions (SIE9) would be sufficient to test SIE. We also suggest changing the direction of reaction 6 in order to make all reactions endothermic, this would facilitate the discussion of the endo- or exothermic errors.

Figure 5 shows the dissociation curve of the \(\text{H}_2^+\) molecular ion that is the simplest way to present the delocalization error, and it shows very clearly that all dRPA results suffer from serious one electron SIE, but the magnitude depends on the choice of the reference orbitals. The HF reference leads to considerably better results than the PBE reference, and the PBE0.25 hybrid reference is between them. Thus, the delocalization error increases in the direction of dRPA:HF < dRPA:PBE0.25 < dRPA:PBE. The huge errors above 6 Å are due to the fractional charge problem leading to delocalization error. The dRPA:PBE0.25/ATZ method shows almost 60 kcal mol\(^{-1}\) delocalization error at a 6 Å internuclear distance, while it yields excellent performance for the DARC test set. This supports the view that the DARC test set does not really test the delocalization error. The results in Table 3 show that the SIE leads to exothermic error at the equilibrium distance, and the dRPA:HF performs better than dRPA:PBE for the equilibrium distance and bond energy. As expected, the RPAX2 is one-electron self-interaction-free as shown in Table 3 and Figure 5 (the small error of RPAX2 is the consequence of the PBE reference and the ATZ basis set).

Figure 6 shows the dissociation curve for the \(\text{H}_2\) molecule that is the simplest way to present the static correlation error. Notice the large (74 kcal mol\(^{-1}\)) error for HF reference orbitals, and the relatively good performance (around 7 kcal mol\(^{-1}\)) error for PBE reference orbitals. This also shows that the dRPA error can be diminished substantially by the proper choice of the reference orbitals. This large improvement can be observed for the equilibrium distance and bond energy calculated with dRPA:PBE shown in Table 4. However, notice the unphysical barrier in Table 3. Equilibrium Distance (R) and Bond Energy (E) of \(\text{H}_2\) Molecule Calculated with dRPA:PBE, dRPA:PBE0.25, dRPA:HF, and RPAX2:PBE\(_{\text{Ex}}\) Methods Using the ATZ Basis Set

<table>
<thead>
<tr>
<th>method</th>
<th>R (Å)</th>
<th>E (kcal mol(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>dRPA:PBE</td>
<td>1.11</td>
<td>−68.35</td>
</tr>
<tr>
<td>dRPA:PBE0.25</td>
<td>1.09</td>
<td>−67.60</td>
</tr>
<tr>
<td>dRPA:HF</td>
<td>1.07</td>
<td>−66.36</td>
</tr>
<tr>
<td>RPAX2:PBE(_{\text{Ex}})</td>
<td>1.05</td>
<td>−63.90</td>
</tr>
<tr>
<td>HF/CBS(5/4)</td>
<td>1.06</td>
<td>−64.40</td>
</tr>
</tbody>
</table>

The reference distance and bond energy is calculated from HF/CBS(5/4) with complete basis set extrapolation, which uses AQZ and ASZ basis sets.

Table 4. Equilibrium Distance (R) and Bond Energy (E) of \(\text{H}_2\) Molecule Calculated with dRPA:HF, dRPA:PBE0.25, dRPA:PBE, and RPAX2:PBE\(_{\text{Ex}}\) Methods Using the ATZ Basis Set

<table>
<thead>
<tr>
<th>method</th>
<th>R (Å)</th>
<th>E (kcal mol(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>dRPA:HF</td>
<td>0.74</td>
<td>−101.38</td>
</tr>
<tr>
<td>dRPA:PBE0.25</td>
<td>0.74</td>
<td>−105.64</td>
</tr>
<tr>
<td>dRPA:PBE</td>
<td>0.75</td>
<td>−107.61</td>
</tr>
<tr>
<td>RPAX2:PBE(_{\text{Ex}})</td>
<td>0.74</td>
<td>−110.42</td>
</tr>
<tr>
<td>CISD/CBS(5/4)</td>
<td>0.74</td>
<td>−109.52</td>
</tr>
</tbody>
</table>

The reference distance and bond energy are calculated from CISD/CBS(5/4) with complete basis set extrapolation, which uses AQZ and ASZ basis sets.

Figure 6 for the dRPA:PBE curve. A 25% mixing of exact exchange considerably improves the results compared to the HF reference orbitals leading to 41 kcal mol\(^{-1}\) static correlation error. The results in Table 4 show that dRPA:PBE0.25 yields excellent equilibrium geometry and 4 kcal mol\(^{-1}\) underbinding error in the dissociation energy. This is an improvement over the 8 kcal mol\(^{-1}\) error of the dRPA:HF but worse than the 2 kcal mol\(^{-1}\) underbinding error of the dRPA:PBE (cf. Table 4). Inspection of Figure 6 and the results in Table 4 shows the excellent performance of the RPAX2:PBE\(_{\text{Ex}}\) method. Notice that
CONCLUSIONS

In this paper, we assessed the performance of semilocal PBE; global hybrid PBE0.25; PBE0.32; PBE0.38; range-separated hybrid rCAM-B3LYP; MCV3; the D2, D3, D10, and VV10 dispersion corrected PBE and TPSS; and the M06-2X, M08-SO, M08-1X-D, and PWPB95-D3 as well as the dRPA:HF, dRPA:PBE, dRPA:PBE0.25, and RPAX2:PBEx functionals for the DARC Diels–Alder reaction test set with various basis sets. For PBE and TPSS calculations, we have found that a reduced triple-ζ basis set called (aug)-cc-pVTZ(-f,d) speeds up the calculations without a significant loss of accuracy. We have found that the energies calculated with semilocal functionals show rapid basis set convergence, and our and literature reaction energies calculated with ATZ or AQT are practically converged with respect to the basis set extension. This is not true for the dRPA results; thus we performed a complete basis set extrapolation based on the ATZ and AQT energies. For reference, we used CCSD(T)/CBS reaction energies.

Although the DARC test set was constructed to test the delocalization errors spoiling semilocal DFT results, it is not satisfactorily suitable for that purpose. We have found that it is possible to obtain quite good results for this DARC test set with functionals that contain large delocalization error. This is quite understandable as inspection of the 14 Diels–Alder reactions reveals that in the first six reactions double (ethene) and triple (ethyne) bonds transformed into single and double bonds, the electron delocalization in the dienes is destroyed; moreover bicyclic products are formed with possible intramolecular dispersion effects. In eight reactions, furane or cyclopentadiene reacts with maleine, and maleimide yielding tricyclic products with endo and exo conformations. Correct energies for these reactions require a method that describes the intramolecular noncovalent bonding, the bond multiplicity change, and formation of new bonds correctly, or systematic multiple error compensation might also lead good results.

We have shown that systematic error compensation is present in PBE hybrids, and the best results can be obtained with mixing 32% exact exchange with 68% PBE exchange (PBE0.32) for the DARC test set. In the PBE0.32 hybrid, the endothermic PBE energy error for reactions leading to bicyclic and tricyclic products is compensated by the exothermic energy error of the exact exchange. This leads to increased average accuracy, but the precision (the corrected sample standard deviation) of the PBE0.32 method does not improve considerably, because the reaction energies are simply shifted in the negative direction. These PBE0.32 results are about the same quality (MAD = 2.8 kcal mol⁻¹) as the results obtained with rCAM-B3LYP and MCV3 with minimal delocalization error (MAD = 2.6 and 3.4 kcal mol⁻¹, respectively), despite the large delocalization error in PBE0.32.

We have shown that a posteriori empirical dispersion corrections improve the PBE and TPSS results, by shifting the too-endothermic reaction energies into the exothermic direction. This is because the attractive intramolecular dispersion interactions stabilize the bicyclic and tricyclic products more than the reactants. Consequently, such corrections worsen the already good or too-exothermic reaction energies yielded by PBE hybrids. Our results show that the D2 corrections are better than the D3 corrections, and the D3 corrections are too small. The VV10 corrections are larger and give better results when applicable. We obtained reasonable results with the one-electron self-correlation free TPSS-VV10 functional (MAD = 2.34 kcal mol⁻¹). Minnesota functionals are parametrized to include medium range noncovalent interactions. The M06-2X functional does not show good performance, but the M05-2X and M08-SO functionals give reasonably good agreement with the reference energies (MAD = 1.69 and 1.46 kcal mol⁻¹, respectively). The more expensive dispersion corrected global double hybrid PWPB95-D3 gives good agreement with the reference energies (MAD = 1.52 kcal mol⁻¹) with the best precision for this class of functionals.

Direct random phase approximation, dRPA, has many good properties as it seamlessly describes the many-body noncovalent interactions, but it suffers from SIE, thus the delocalization error. We used CBS extrapolated dRPA with HF, PBE, and hybrid PBE0.25 reference orbitals. The too-compact HF electron density leads to more than 2 kcal mol⁻¹ exothermic error. The dRPA:PBE and dRPA:PBE0.25 yield excellent results (MAD = 1.55 and 0.53 kcal mol⁻¹, respectively) with much better precision than the best methods discussed above. We have found that dRPA:PBE shows a systematic endothermic error, and that can be compensated by the systematic exothermic error of the AQZ basis set (MAD = 0.32 kcal mol⁻¹). Excellent results can be obtained from RPAX2:PBEx with the ATZ basis set (MAD = 0.83 kcal mol⁻¹) at a considerably higher computational cost compared to dRPA.

Our results show that the DARC test set is not particularly suitable to identify the delocalization error caused by the self-interaction error, but it is an excellent test set for intramolecular interactions. We observed that many reactions in the test sets give the same results; thus by leaving out these calculations, considerable computational effort and time can be saved without the loss of information. We suggest using a reduced DARC6 test set composed of reactions 1, 2, 3, 4, 7, and 11:

<table>
<thead>
<tr>
<th>no.</th>
<th>reactants</th>
<th>products</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ethene + butadiene</td>
<td>monocyclic</td>
</tr>
<tr>
<td>2</td>
<td>ethyne + butadiene</td>
<td>bicyclic</td>
</tr>
<tr>
<td>3</td>
<td>ethene + cyclopentadiene</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ethyne + cyclopentadiene</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>furan + maleic anhydride</td>
<td>tricyclic (endo)</td>
</tr>
<tr>
<td>11</td>
<td>cyclopentadiene + maleic anhydride</td>
<td></td>
</tr>
</tbody>
</table>

The SIE11 test set is more suitable to test the delocalization error than the DARC test set; however, we identified two reaction energies in these sets that are not sensitive to SIE. We observed that it is particularly difficult to obtain good reaction energy for reaction 11 (FLiF → Li + F₂). Only SOSEX:HF and RPAX2:PBEx give reasonable energy for this reaction, while dRPA:PBE gives a very large (around 43 kcal mol⁻¹) endothermic error. Overall, the best results for SIE11 were obtained from the RPAX2:PBEx/ATZ method (MAD = 3 kcal mol⁻¹). We suggest reducing SIE11 to SIE9 by omitting reaction numbers 5 and 7 and changing the direction of reaction 6.

Inspection of H₂ and H₂ dissociation curves reveals that all dRPA methods fail for H₂, independent of reference orbitals, and dRPA:PBE performs quite well for H₂ while dRPA:HF fails seriously.

Although the dRPA results presented here are very promising for the DARC test set, the dRPA suffers from serious self-interaction error; thus it yields erroneous dissociation curves and bad atomization energies. A good but more expensive alternative might be the one-electron self-correlation free RPAX2 method.
with lower many-electron self-interaction and static correlation error.

### APPENDIX

The general form of the exchange-correlation energy of a global hybrid functional is given by eqs A1 and A2:

\[
E_{XC}^{\text{gh}} = a E_X^{\text{exact}} + (1 - a) E_X^a + E_C^d
\]  

(A1)

\[
E_X^{\text{exact}} = -\frac{1}{2} \sum_{i<j} \langle \hat{\psi}_i \hat{\psi}_j \rangle
\]  

(A2)

where \(E_X^{\text{exact}}\) is the exact exchange energy calculated from the \(i\) and \(j\) occupied orbitals (using Dirac’s notation). The mixing parameter, \(a\), is equal to 0.25, 0.32, and 0.38 for PBE0.25 (PBE0.25 also called as PBE0 or PBEh in the literature), PBE0.32, and PBE0.38 and 0.10 and 0.25 for TPSS0.1 and TPSS0.25, respectively (TPSS0.1 is called TPSSH and TPSS0.25s called TPSS in the literature). \(E_C^d\) and \(E_X^a\) are the semilocal exchange and correlation energies, respectively.

The atom-pairwise DFT-D2 dispersion correction uses terms depending on the minus sixth and eighth power of the interatomic distances (eq A3).

\[
E_{d2}^{\text{D2}} = -\frac{1}{2} \sum_{i \neq j} \sum_{n=0,8} s_n \phi_{\text{damp}}(R_{ij}) \frac{C_n^{\text{vdW}}}{R_{ij}^n}
\]

(A3)

\[
\phi_{\text{damp}}(R_{ij}) = \frac{1}{1 + e^{-d(R_{ij} - R^0_{\text{vdW}})}}
\]

(A4)

where \(C_n^{\text{vdW}} = (\langle \tilde{C}_n \rangle_{\text{vdW}})^{1/2}\), \(s_6 = 1\), \(s_8 = 0.722, 0.928, 0.998\) for PBE, PBE0.25, and PBE0.38 and 1.105, 1.219, and 1.242 for TPSS, TPSS0.1, and TPSS0.25. Each term has a damping function (eq A6), which has a steepness determined by the mean cutoff radius \(R_0\).

\[
f_n(R_{ij}) = \frac{1}{1 + e^{-d(R_{ij} - R^0_{\text{vdW}})}}
\]

(A5)

The parameters in eq A4 are \(d_6 = 14\), \(d_8 = 16\), \(s_{6,8} = 1\); \(s_{6,8}\) is 1.217, 1.287, 1.333 for PBE, PBE0.25, and PBE0.38 and 1.166, 1.223, and 1.252 for TPSS, TPSS0.1, and TPSS0.25.

With the Becke–Johnson damping function, the correction can be written according to eqs A7 and A8.

\[
E_{d3}^{\text{D3}} = -\frac{1}{2} \sum_{i \neq j} \sum_{n=0,8} s_n \phi_{\text{damp}}(R_{ij}) \frac{C_n^{\text{vdW}}}{R_{ij}^n}
\]

(A6)

\[
f(R_{ij}) = a_n R_{ij}^{n} + a_1
\]

(A7)

where \(R_0 = (\langle C_0^{\text{vdW}} \rangle)^{1/2}\), the global scaling factors of the energy terms are \(\{s_6, s_8\} = \{0.418, 0.0\}\) for PBE, PBE0.25, while the parameters in the Becke–Johnson damping are \(\{a_0, a_1\} = \{0.0, 5.65\}\) for PBE0.25.

The dD10 dispersion correction is also an atom-pairwise correction but contains also terms depending on the tenth power of the interatomic distances and double damping.

\[
E_{d10}^{\text{D10}} = -\frac{1}{2} \sum_{i \neq j} \sum_{n=0,8} f_n(b,R_{ij}) \frac{C_n^{\text{vdW}}}{R_{ij}^{n+2}}
\]

(A8)

The atomic van der Waals radii are \(R_{\text{vdW}} = 1.2\) Å, \(R_{\text{vdW}} = 1.7\) Å, \(R_{\text{vdW}} = 1.55\) Å, and \(R_{\text{vdW}} = 1.52\) Å. The parameter of the Fermi damping function is a global constant \(a = 1.45\). The individual damping function of each term is given by eq A11.

\[
f_{2n}(b,R_{ij}) = 1 - e^{-b R_{ij}^{2n}}
\]

(A9)

The parameter of the individual damping functions depends on the choice of density functional \((b = 1.11 \text{ for PBE and } b = 1.8 \text{ for TPSS})\) fitted to the S22 database. We slightly modified the original method using the \(C_6\) coefficient for any type of \(i\)-\(j\) atom pairs given by

\[
f_{2n}(b,R_{ij}) = 1 - e^{-b R_{ij}^{2n}}
\]

(A10)

The local band gap is approximated by eq A12.

\[
E_C^{\text{LMO}}[n(\vec{r})] = \frac{1}{2} \iint n(\vec{r}) \Phi(\vec{r}, \vec{r'}) n(\vec{r'}) d^3\vec{r} d^3\vec{r'} - \sum_{\lambda} E_{\text{unif}}^{\text{LMO}}[n(\vec{r})] n(\vec{r}) d^3\vec{r}
\]

(A11)

The first term is a two electron integral with the \(\Phi(\vec{r}, \vec{r'})\) interaction kernel (eq A14), the second term assures that the correction vanishes at uniform electron density.

\[
\Phi(\vec{r}, \vec{r'}) = -\frac{3}{2} \frac{1}{g(\vec{r}) g(\vec{r'})}(g(\vec{r}) + g(\vec{r'}))
\]

(A12)

where \(g(\vec{r}) = \omega_0(\vec{r}) R^2 + \kappa(\vec{r})\) with \(R = \vec{r} - \vec{r}'\) distance and \(\omega_0(\vec{r}) = (\omega_1(\vec{r}) + \omega_2(\vec{r})^{1/2})\) local frequency. The local dipole resonance frequency \(\omega_1(\vec{r}) = 3 + 1/2 \omega_0(\vec{r})\) is computed from the local plasma frequency \(\omega_2(\vec{r}) = (4 \pi n_0(\vec{r})\overline{C})^{1/2}\). The local band gap is approximated by eq A11.

\[
E_C^{\text{hyb}}[\omega_0(\vec{r})] = (3 \pi n_0(\vec{r})\overline{C})^{1/2}
\]

(A13)

The general form of the local double hybrid exchange-correlation functionals is given by eq A15.

\[
E_C^{\text{hyb}} = E_C^{\text{exact}} + (1 - a_C) E_C^{\text{ed}} + a_C E_C^{\text{PT2}} + (1 - a_C) E_C^{\text{bb}}
\]

(A14)
where $E_{PT2}^{C}$ is the second-order perturbation correlation given by eq A16. The parameters in eq A14 are $a_\alpha = 0.53$ and $a_c = 0.27$ for B2PLYP ($a_\alpha = 1, a_T = 1$ in eq A16). For DSD-BLYP, the parameters are $a_\alpha = 0.69$ and $a_c = 0.54$ with scaling down the same-spin PT2 component to 80% ($a_\alpha = 1, a_T = 0.8$ in eq A16). For PWPB95, the parameters are $a_\alpha = 0.50$ and $a_c = 0.269$ with only the opposite-spin PT2 component ($a_\alpha = 1, a_T = 0$ in eq A16).

$$E_{PT2}^{C} = a_e E_{PT2}^{E} + a_T E_{PT2}^{T}$$  \hspace{1cm} (A16)

where the spin components are given by eqs A17 and A18 with the pair energies (eqs A19, A20, and A21) and the amplitudes (eq A22) using Dirac’s notation for the integrals.

$$E_{PT2}^{S} = \sum_{\alpha \beta} e_{\alpha \beta}$$  \hspace{1cm} (A17)

$$E_{PT2}^{T} = \frac{1}{2} \sum_{\alpha \beta} e_{\alpha \beta} + \frac{1}{2} \sum_{\alpha \beta} e_{\beta \alpha}$$  \hspace{1cm} (A18)

$$e_{\alpha \beta} = \sum_{\alpha \beta} T^{\alpha \beta}_{\alpha \beta}$$  \hspace{1cm} (A19)

$$e_{\alpha \beta} = \sum_{\alpha \beta} \sum_{\alpha \beta} T^{\alpha \beta}_{\alpha \beta} - T^{\alpha \beta}_{\alpha \beta}$$  \hspace{1cm} (A20)

$$T^{\alpha \beta}_{\alpha \beta} = \frac{\langle \alpha \beta \mid \alpha \beta \rangle}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\alpha} - \epsilon_{\beta}}$$  \hspace{1cm} (A21)

and

$$T^{\alpha \beta}_{\alpha \beta} = \frac{\langle \alpha \beta \mid \alpha \beta \rangle}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\alpha} - \epsilon_{\beta}}$$  \hspace{1cm} (A22)

The $i$ and $j$ indices belong to the occupied orbitals. The $\alpha$ and $\beta$ indices belong to the virtual orbitals. $\alpha$ and $\beta$ are spin indices. The energy of the $ith$ orbital is $\epsilon_{\alpha}$, etc. $T^{\alpha \beta}_{\alpha \beta}$ are the double excitation amplitude matrix elements. The DFT-D3 parameters ($t_{ij}, s_{ij}, g_{ij}$) are $(0.64, 1.427, 1.022)$ for B2PLYP, $(0.50, 1.569, 0.705)$ for DSD-BLYP, and $(0.82, 1.557, 0.705)$ for PWPB95.

The direct random phase approximation (dRPA) exchange-correlation energy is given by eq A23.

$$E_{dRPA}^{XC} = E_{dRPA}^{T} + E_{dRPA}^{C}$$  \hspace{1cm} (A23)

where the dRPA correlation is given by eq A24. The exact exchange (eq 2) and dRPA correlation energies are generally calculated on self-consistent HF, PBE, or PBE0.25 orbitals.

$$E_{dRPA}^{C} = \frac{1}{2} \text{tr} \left[ K_{T} \right]$$  \hspace{1cm} (A24)

$$B_{\alpha \beta} = \langle \alpha \beta \mid \alpha \beta \rangle$$  \hspace{1cm} (A25)

where $B$ is the non-antisymmetrized two-electron repulsion integral matrix (eq A25) and $T$ is double excitation amplitude matrix given by the iterative update (the iteration is initialized with $T^{(0)} = 0$) according to eq A26 with the Hadamard product of $\Delta$ defined by eq A27.

$$T^{(n+1)} = -\Delta \cdot (B + BT^{(n)}) + T^{(\alpha \beta)B} + T^{(\alpha \beta)BT^{(\alpha \beta)}}$$  \hspace{1cm} (A26)

$$\Delta_{\alpha \beta} = \frac{1}{\epsilon_{\alpha} + \epsilon_{\beta} - \epsilon_{\alpha} - \epsilon_{\beta}}$$  \hspace{1cm} (A27)

The equations are solved in an $O(n^2)$-scaling iterative procedure with the density-fitted form of electron repulsion integrals and the Cholesky decomposition of matrix $\Delta$. $\Delta$

The equations are solved in an $O(n^2)$-scaling iterative procedure with the density-fitted form of electron repulsion integrals and the Cholesky decomposition of matrix $\Delta$. $\Delta$

$$E_{dRPA}^{C} = \frac{1}{2} \text{tr} \left[ K_{T} \right]$$  \hspace{1cm} (A28)

where $K_{\alpha \beta} = \langle \alpha \beta \mid \alpha \beta \rangle$ are spin-singlet adapted antisymmetrized electron repulsion integrals.

The RPAX2 correlation energy is defined by eq A29 with the density-fitted form of matrix $B$ according to eq A30.

$$E_{RPAX2}^{C} = \frac{1}{2} \text{tr} \left[ LT^{(\alpha \beta)}L^T \right]$$  \hspace{1cm} (A29)

$$B = LL^T$$  \hspace{1cm} (A30)

The iteration is initialized with $T^{(0)} = 0$; then the amplitudes of the iteration cycles are given by

$$T^{(n+1)} = -\Delta \cdot (B + BT^{(n)}) + T^{(\alpha \beta)B} + T^{(\alpha \beta)BT^{(\alpha \beta)}}$$  \hspace{1cm} (A31)

where $\hat{P}$ is a permutation operator that permutes the orbitals. $\Delta$ is decomposed using the Cholesky decomposition. The equations are solved in an $O(n^2)$-scaling iterative procedure.

**ASSOCIATED CONTENT**

**Supporting Information**

Six tables show the deviations from the reference Diels–Alder reaction energies and statistics for several methods discussed in this paper. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00223.

**AUTHOR INFORMATION**

*E-mail: csonkagi@gmail.com.*

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

G.I.C. acknowledges the support by the New Hungary Development Plan (Project ID: TAMOP-4.2.1/B-09/1/KMR-2010-0002) at the BME project. The authors thank for the Hungarian National higher education and research network for the computer time.

**REFERENCES**

Accurate Complete Basis Set Extrapolation of Direct Random Phase Correlation Energies

Pál D. Mezei,*† Gábor I. Csonka,*‡ and Adrienn Ruzsinszky†

†Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary
‡Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States

ABSTRACT: The direct random phase approximation (dRPA) is a promising way to obtain improvements upon the standard semilocal density functional results in many aspects of computational chemistry. In this paper, we address the slow convergence of the calculated dRPA correlation energy with the increase of the quality and size of the popular Gaussian-type Dunning’s correlation consistent aug-cc-pVXZ split valence atomic basis set family. The cardinal number $X$ controls the size of the basis set, and we use $X = 3−6$ in this study. It is known that even the very expensive $X = 6$ basis sets lead to large errors for the dRPA correlation energy, and thus complete basis set extrapolation is necessary. We study the basis set convergence of the dRPA correlation energies on a set of 65 hydrocarbon isomers from CH$_4$ to C$_6$H$_6$. We calculate the iterative density fitted dRPA correlation energies using an efficient algorithm based on the CC-like form of the equations using the self-consistent HF orbitals. We test the popular inverse cubic, the optimized exponential, and inverse power formulas for complete basis set extrapolation. We have found that the optimized inverse power basis extrapolation delivers the best energies. Further analysis showed that the optimal exponent depends on the molecular structure, and the most efficient two-point energy extrapolations that use $X = 3$ and 4 can be improved considerably by considering the atomic composition and hybridization states of the atoms in the molecules. Our results also show that the optimized exponents that yield accurate $X = 3$ and 4 extrapolated dRPA energies for atoms or small molecules might be inaccurate for larger molecules.

INTRODUCTION

The direct random phase approximation (noted as RPA or dRPA)$^{1−3}$ is a promising way to obtain improvements upon the standard semilocal density functional theory or in the coupled cluster formalism. RPA is accurate for intra- and intermolecular noncovalent interactions,$^5$ for adsorption,$^6,7$ for interlayer interactions,$^8$ and for van der Waals bonded solids.$^9$ RPA captures the nonpairwise-additive feature of dispersion interactions missed by many pairwise $a$ posteriori molecular mechanics-like dispersion corrections. For covalent bond breaking reaction energies, RPA performs moderately.$^4$ However, RPA describes the short-range correlation poorly. It fails seriously in situations that lead to short-range rearrangement of the electronic structure. It also fails for ionization and atomization processes, where the number of electron pairs are changing. RPA overestimates the ionization potentials$^{10}$ and underestimates the atomization energies of molecules$^{11}$ and solids.$^{12}$ Furthermore, RPA suffers from self-interaction and static correlation error, which lead to the underestimation of the barrier heights$^{13}$ and to the overestimation of the energy of breaking bonds.$^{14}$ These errors can be understood from the perspective of fractional charges and fractional spins.$^{15}$ Beyond RPA methods, like RPA$^+,$$^{16}$ RPA$^{++}$,$^{17}$ RPA plus second-order screened exchange (SOSEX),$^{18}$ RPA plus renormalized single excitations (rSE),$^{19}$ and several other approaches, are applied to correct these errors.$^{19−21}$

Another well-documented problem with the RPA calculations is the slow convergence of the RPA correlation energy with the increase of the basis set size. The slow convergence can be traced back to the poor description of the electron–electron Coulomb cusps by smooth orbital products.$^{22,23}$ The complete basis set (CBS) extrapolation method provides a simple way to deal with this problem. There exist more sophisticated, but very expensive methods, like the explicitly correlated wave function,$^{24}$ transcorrelated,$^{28,29}$ diffusion quantum Monte Carlo,$^{30}$ and other methods that treat this error.$^{33−35}$ The CBS extrapolations are based on the convergence of the energy terms with respect to the increasing completeness of a series of basis sets. The CBS energy limit of the infinite basis set can be extrapolated from convergent energy series calculated with systematically increasing quality Gaussian-type$^{36}$ or plane-wave$^{37}$ basis sets. One such widely
used Gaussian-type basis set is Dunning’s correlation consistent aug-cc-pVXZ split valence atomic basis set family (noted here as AXZ) characterized by the X cardinal number. More complete basis sets are characterized with larger cardinal numbers, increasing computational times, and convergent series of energies or other properties.38,39

The Hartree–Fock (HF) energy converges exponentially with respect to the increase of the cardinal number X.40,41 A two point extrapolation form of eq 5 in ref 41 yields HF energy limits with an RMS error of 0.01 millihartree using the AXZ and A6Z basis sets (noted as CBS(6/5)). It was also found that the inverse fifth power form is close to optimal for the so-called W1 calculations.41

The Møller–Plesset second order perturbation (MP2) correlation energies converge much slower with the cardinal number.42 Feller, Peterson, and Hill compared the exponential, mixed exponential/Gaussian, inverse cubic,42 and shifted quartic CBS schemes for the coupled cluster correlation energies.44 For the random-phase approximation (RPA) correlation energy, Eshuis and Furche used the inverse cubic formula.45 Fabiano and Della Sala examined shifted inverse cubic or quartic formulas and inverse power functions for the basis set extrapolation of RPA correlation. At the CBS(7/6) and CBS(6/5) levels, all these approaches yield similar results.46 For intermolecular correlation energy in dimers, the counterpoise (CP) correction might improve the basis set convergence (but its application for systems of more than two entities is problematic).47 For a detailed explanation about the errors and for further information about the development of CP corrections, see refs 48 and 49. Recently, it was shown that CP corrections yield poor results for anion-π interaction energies calculated with MP2.50

■ COMPUTATIONAL DETAILS

We examine the basis set convergence of the dRPA correlation energies on a set of 65 hydrocarbon isomers (Figure 1) from CH₄ to C₆H₆. The molecular geometries are optimized with B3LYP/6-31G(2df,p) and taken from the NIST Computational Chemistry Comparison and Benchmark Database.51

The dRPA correlation energy (dRPAc) is given by eq 1

\[ E_{dRPA} = \frac{1}{2} \text{tr}[BT] \]

where B is the nonantisymmetrized two electron repulsion matrix, and T is the double excitation amplitude matrix. The dRPAc is calculated using the efficient algorithm of Heßelmann which is based on the CC-like form (by using antisymmetrized two-electron integrals in the Riccati equation) of the dRPA equations. The equations are solved in an $O(n^4)$-scaling iterative procedure by the density-fitted (DF) form of electron repulsion integrals and Cholesky decomposition of the orbital energy denominators. Recently an efficient linear-scaling implementation of the dRPA is applied for molecules with more than 1000 atoms and 10 000 basis functions.54

We calculate the iterative DF-dRPAc energies on self-consistent HF orbitals, called as dRPAc:HF, using the MRCC (30/12/2014 release) program. We apply Dunning’s correlation consistent aug-cc-pVXZ basis sets (noted here as AXZ, where X = 3, 4, 5, 6) with aug-cc-pVXZ-RI-JK and aug-cc-pVXZ-RI auxiliary basis sets for the self-consistent field (SCF) and correlation calculations, respectively. We call these
calculated dRPAc:HF/AXZ or for short dRPAc/AXZ. (We used RI approximation instead of RI-JK for A6Z and consider these results as $X = 5.99$ based on the difference between the RI and RI-JK approximations for small molecules such as methane, ethylene, ethene, and ethane.) We note that we excluded double-$\zeta$ basis sets from this study because the random errors of such popular basis sets are disturbingly large, and extrapolations that include such basis sets are considerably less reliable.\(^{43}\)

We use the inverse cubic function (eq 2)\(^{43}\) suggested in the literature for dRPAc energy extrapolation.\(^{49}\) In addition, we test the exponential decay (eq 3) function and the generalized inverse power function (eq 4)\(^{43}\)

$$E_{CBS}^{dRPAc} = E_{CBS}^{dRPAc,AXZ} - AX^{-3}$$ \hspace{1cm} (2)

$$E_{CBS}^{dRPAc} = E_{CBS}^{dRPAc,AXZ} - Ae^{-nX}$$ \hspace{1cm} (3)

$$E_{CBS}^{dRPAc} = E_{CBS}^{dRPAc,AXZ} - A(X + d)^{-\alpha}$$ \hspace{1cm} (4)

where $A$, $n$, and $\alpha$ are optimized parameters, and $d$ is a shift parameter that modifies the cardinal numbers. For the simple inverse power fit, $d = 0$. We define the average fitting error as the average of the errors yielded by eqs 2–4 using optimized parameters $A$, $n$, $d$, $\alpha$, and $E_{CBS}^{dRPAc}$ (e.g., for eq 2 the error $= E_{CBS}^{dRPAc} - E_{CBS}^{dRPAc,AXZ}$), compared to the calculated $E_{CBS}^{dRPAc}$ energy values for every molecule in the test set, where $X = 3–6$. Notice that optimal $\alpha$ in eq 4 linearizes the $E_{CBS}^{dRPAc}$ vs $X^{-\alpha}$ function and yields the $E_{CBS}^{dRPAc}$ as the intercept and $A$ as the slope (vide infra).

Eqs 2–4 are one-point extrapolation formulas. Here, we apply a two-point extrapolation formula for the dRPAc energy extrapolation similar to the two-point extrapolations applied in refs 56 and 57:

$$E_{CBS}^{dRPAc,X,X-1} = E_{CBS}^{dRPAc,AXZ} + E_{CBS}^{dRPAc} - E_{CBS}^{dRPAc,AXZ}(X, X-1)$$ \hspace{1cm} (5)

The $C_{dRPAc}^{X,X-1}$ extrapolation coefficients can be calculated from eq 4 ($d = 0$) as follows:\(^{56,57}\)

$$C_{dRPAc}^{X,X-1} = \frac{1}{\left(\frac{X}{X-1}\right)^{\alpha} - 1}$$ \hspace{1cm} (6)

\section*{RESULTS AND DISCUSSION}

The average fitting errors for dRPAc energies for 65 hydrocarbons using cardinal numbers $X = 3, 4, 5, 6$ (see Table S1 for the energies) and eqs 4–6 are shown in Figure 2. As the $E_{CBS}^{dRPAc}$ energy values predicted by the inverse cubic formula do not fit well to the calculated $E_{CBS}^{dRPAc}$ energy values, for $X = 3–6$, systematic errors occur. The predicted and calculated energies agree relatively well for $X = 3$ and $5$ but not for $X = 4$ (large, systematic negative deviations) and $X = 6$ (large, systematic positive deviations). This behavior clearly shows that the inverse cubic form is not suitable for accurate extrapolation of dRPAc energies. Thus, the inverse cubic function yields very large and oscillating average fitting errors (over 1 kcal mol$^{-1}$) and large individual fitting errors (up to almost 2 kcal mol$^{-1}$) for cardinal numbers 4 and 6, respectively (cf. Figure 2). Figure 2 shows that the fitted exponential function form, eq 3, is considerably better. Clearly the best performance is shown by the optimized inverse power functions, eq 4 with average $\alpha = 2.5608$, as the predicted energies are 100-times more accurate than those of calculated from the inverse cubic formulas, eq 2 as shown in Figure 2. Notice that the small average fitting errors for eq 6 are around 0.01 kcal mol$^{-1}$, thus they are practically invisible in Figure 2; this is why we show the numerical values.

The inverse power function ($d = 0$ in eq 4) performs similarly to its shifted versions ($d = \pm 0.5$ in eq 4), in agreement with the results in Figure 4 of ref 43. The average fitting errors for shifted versions compared to the inverse power fitting errors are shown in Figure S1 of the Supporting Information. The best fitting can be obtained by $d = 0.165$ in eq 4. We show these results in Table S2 and Figures S3–S5. The inverse power function form fits particularly well for the ATZ and AQQ energies (cf. Figure 2, $X = 3$ and 4), so the CBS energy can be estimated from these two energies with $C_{dRPAc}^{dRPAc,AXZ} = 0.9183$ from eq 6.

Detailed analysis shows that the extrapolation coefficient in eq 5 and the $\alpha$ exponent in eq 6 are slightly different for every molecule in the test set (cf. Supporting Information Figures S2 and S3). Figure S4 shows the calculated dRPAc:HF correlation energies for ethyne as a function of $X^{-\alpha}$ and as a function $X^{-2.529}$ for $X = 3, 4, 5, 6$. The nonlinearity shown for the inverse cubic formula (eq 2) leads to poorly extrapolated CBS energy. Approximating the parabolic function by a linear fit for the inverse cubic formula causes systematic negative deviations for $X = 4$ and positive deviations for $X = 6$ as shown in Figure 2. The optimal $\alpha = -2.529$ for ethyne molecule linearizes the correlation energy function vs $X^{-\alpha}$ and yields more consistently and accurately extrapolated CBS energy (cf. Figures 2 and S4). Using the inverse cubic formula will yield different CBS(4/3), CBS(5/4), and CBS(6/5) extrapolated dRPA correlation energies, while the optimal $\alpha$ yields practically equal extrapolated correlation energies independent of the cardinal numbers used for the extrapolation.

Notice that the dRPAc:HF energy of free C atoms converges with the $X$ as $X = 2.303$ in eq 6. We observed that the optimal exponent for methane and $H_2$, $\alpha = 2.66$ and 2.70, respectively, are considerably larger than the optimal exponent for 1,3,5-hexatriyne ($\alpha = 2.495$) or C atom. Notice that the H:C ratio for...
methane is 4:1 and for 1,3,5-hexatriyne is 1:3. While the dRPAc
energy for 1,3,5-hexatriyne converges significantly slower than
that of the methane, it converges significantly faster than that of
the free C atom. The origin of this behavior might be the
considerable sensitivity of the $E_{\text{dRPAc}}$ on the cusp error. For less
compact electron densities in the hydrogen rich molecules, the
basis set convergence might be faster as the cusp error might be
smaller for these compounds compared to the heavy atom rich
compounds. This observation also shows that the basis set
convergence for atoms is not transferable to molecules without
significant loss of accuracy. Analysis of Figures S2 and S3 yields
that increasing the number of C atoms in the molecule slows
down the basis set convergence by decreasing the value of $\alpha$.
We have found that the exponents for eq 4 ($d = 0$) can be
estimated from the number of carbon and hydrogen atoms in
the molecule (eq 7 with the parameters of Table 1) or from the
hybridization states of the carbon atoms (eq 8 with the
parameters of Table 1)

$$\alpha \approx m_{\text{atom}} + \frac{N^H}{N^H + N^C} m_{\text{hybr}} + \frac{N^C}{N^H + N^C} m_{\text{atom}} \tag{7}$$

$$\alpha \approx m_{\text{hybr}} + \frac{N^H}{N^H + N^C} p_{\text{hybr}} + \frac{N^C}{N^H + N^C} P_{\text{hybr}} + \frac{N^C(p^{sp})}{N^H + N^C} P_{\text{hybr}}(p^{sp}) + \frac{N^C(p^{sp})}{N^H + N^C} P_{\text{hybr}}(p^{sp}) \tag{8}$$

where $\alpha$ is the exponent in eq 6, $m$ and $P$ are fitting parameters,
and $N$ is the number of the specific atoms (e.g., C atoms in a
specific hybrid state). In the test set of 65 hydrocarbon
molecules we have 450 H atoms, 48 C(sp), 110 C(sp$^2$), and
133 C(sp$^3$) atoms, altogether 291 C atoms. Notice that the
classification of the C atoms according to their hybrid states
is not always straightforward. Inspection of the parameter values
in Table 1 suggests that it is enough to distinguish the C(sp$^3$)
atoms from all the rest of the C atoms (eq 9), as the parameters
for C(sp) and C(sp$^2$) agree quite well. This leads to the
following simpler equation

$$\alpha \approx m_{\text{hybr}} + \frac{N^H}{N^H + N^C} p_{\text{hybr}} + \frac{N^C}{N^H + N^C} P_{\text{hybr}} + \frac{N^C(p^{sp})}{N^H + N^C} P_{\text{hybr}}(p^{sp}) + \frac{N^C(p^{sp})}{N^H + N^C} P_{\text{hybr}}(p^{sp}) \tag{9}$$

where $m_{\text{hybr}}$ and $P_{\text{hybr}}$ can be found in Table 1; the weighted
average of the C(sp) and C(sp$^2$) parameters can be used for the
carbon atoms, $P_{\text{hybr}}^{C} = (48P_{\text{hybr}}(p^{sp}) + 110P_{\text{hybr}}(p^{sp}))/158 = -0.1712$, with the C(sp$^3$) correction of $\Delta P_{\text{hybr}}(p^{sp}) = P_{\text{hybr}}(p^{sp}) - P_{\text{hybr}}^{C} = -0.1260$.

Inspection of the results shows that increasing the ratio of H
atoms increases the basis set convergence, increasing the ratio
of C atoms slows that down, and this effect is particularly large
for C(sp$^3$) atoms in the molecules. The exponent calculated
for 1,3,5-hexatriyne that contains 6 C(sp) and only two H atoms is
the smallest, eqs 7 and 8 give 2.488 in good agreement with the
optimized $\alpha = 2.495$ given above. For methane eqs 7–9 do not
work that well as the optimal exponent is $\alpha = 2.653$, but the
predicted exponents are 2.603 and 2.624 for eqs 7 and 8,
respectively (see also Figure 3). From the parameter values in
Table 1, we can predict the $\alpha$ values for larger saturated
hydrocarbons. For C$_8$H$_{18}$, C$_{10}$H$_{22}$, and C$_{12}$H$_{26}$, we obtain $\alpha = 2.574$, 2.572, and 2.571, respectively, from hybrid state
contributions, and we obtain very similar values from atomic
contributions. For large saturated linear hydrocarbons, as the
C:H ratio converges to 1:2, the $\alpha$ converges to 2.563.

The correlation between the calculated and the fitted
exponents is shown in Figure 3. Inspection of the figure shows
that consideration of the hybridization states (cf. eq 8
and Table 1) leads to better agreement (see also Figures S2 and S3)
between the calculated and fitted exponents. The
agreement is poorest for methane as discussed before, and
this error is well visible in Figure 3. The two other problematic
predictions are for ethyne and ethene.

Our present results show that the exponents working well for
atoms or small molecules are not well applicable for dRPAc of
larger molecules. In our examples the basis set convergence of
large hydrocarbons is slower than that for methane, thus the
dRPAc/CBS(4/3) energies that use $\alpha = 2.654$ for such large
molecules might be in serious endothermic error. For CO and
C$_2$ molecules we obtained $\alpha = 2.188$ and 2.348, respectively.
Notice that this latter value is not far from the value of $\alpha = 2.303$ for the C atom shown above. For NH$_3$, H$_2$O, and HF

![Figure 3. The correlation between the calculated and the fitted exponents of the power functions describing the basis set convergence of dRPAc:HF energy for the 65 hydrocarbon molecules shown in Figure 1. Legend: Average means that we use the average $\alpha = 2.5608$ exponent in eq 6. Atomic contribution ($x$) means that we calculate the value of $\alpha$ for every molecule according to eq 7 using the parameters from Table 1. Hybridization states (+) means that we calculate the value of $\alpha$ for every molecule according to eq 8 using the parameters from Table 1.](image-url)
molecules we obtained similar α = 2.235, 2.203, and 2.189, respectively. These exponents are much smaller than the exponents that control the basis set convergence for larger hydrocarbons (see above α around 2.563). Application of such small inverse power exponents for large hydrocarbons would lead to very large exothermic CBS extrapolation error. These examples clearly show that the basis set convergence is sensitive to the molecular structure, but this sensitivity decreases with the increase of the size of the molecules. While for small molecules, like methane, water, ammonia, HF, N₂, CO, and C₂, the optimal α values are very different (vary from 2.18 to 2.65), for the larger molecules the optimal α values are similar and predictable (cf. eqs 7–9 and see also Figures S2a, S2b, and S3). Notice also that all the dRPAc energies in this study converge considerably slower with the cardinal number than the inverse cubic formula suggested in the literature.

Next we analyze how the basis set convergence is influenced by the reference orbitals for the dRPA correlation energy calculation. Replacing the HF orbitals by PBE orbitals might improve the basis set convergence of the ECBS, as the potentially less compact PBE molecular electron density might lead to smaller cusp error. For methane we obtained α = 2.80 vs 2.65 for PBE vs HF reference orbitals, respectively (ECBS(PBE/HF) = 0.807 and ECBS(PBE) = 0.876). In our previous paper for PBE, PBE0.25, and HF reference orbitals applied for ethyne and ethene we obtained ECBS(PBE/HF) = 0.856, ECBS(PBE0.25) = 0.867, and ECBS(HF) = 0.917. For the more compact electron density of N₂, the basis set convergence is slower but less dependent on the reference orbitals, leading to α = 2.20 vs 2.18 for PBE vs HF reference orbitals, respectively. For this latter molecule the difference between the basis set convergences is practically negligible (cf. ECBS(PBE) = 1.144 and ECBS(HF) = 1.132). Using the published PBE energies calculated with SCF PBE orbitals for the N₂ molecule (cf. Table 1 in ref 11), we obtained α = 2.08, leading to a slightly slower basis set convergence.

We have also analyzed the basis set convergence for the MP2 correlation energy. We have selected methane and 1,3,5-hexatriyne molecules for the analysis and obtained α = 2.88 and 2.83, respectively. Our results show that the basis set convergence of the correlation energy is considerably faster for the MP2 correlation energy than for the dRPA energy, and these power values are much closer to the α = 3 used extensively in the literature. It can be noticed that the difference between the negative power values for different molecules is also much smaller for MP2 than for dRPA.

Next we analyze the improvements yielded by applying eqs 7 and 8 over the simple average α for CBS extrapolation of the dRPA energies. The individual deviations for the simple average α = 2.5608 are calculated as a difference: 
\[ \alpha_{dRPA(CCBS(X/-1))} - \alpha_{CBS(X/-1)} \] 
for every molecule according to eq 7 using the \( \alpha \) from Table 1 and use this exponent in eq 6. Atomic correction means that we calculate the value of \( \alpha \) for every molecule according to eq 7 with the parameters from Table 1 and use this exponent in eq 6. The CBS(4/3) extrapolation coefficients obtained this way are used in eq 5 to obtain dRPAc:HF/CBS(X/(X-1)) energies.

Figure 4. The mean absolute deviations of the predicted from the fitted complete basis set extrapolated dRPAc:HF/CBS(X/(X-1)) energies, where \( X = 4, 5, 6 \) for the 65 hydrocarbon molecules shown in Figure 1. Legend: Average means that we use the average \( \alpha = 2.5608 \) exponent in eq 6. Atomic contribution means that we calculate the value of \( \alpha \) for every molecule according to eq 7 using the parameters from Table 1 and use this exponent in eq 6. The CMS extrapolation coefficients obtained this way are used in eq 5 to obtain dRPAc:HF/CBS(X/(X-1)) energies.
Figure 5. The minimum, maximum, mean (MD), and corrected sample standard deviations multiplied by 3 (3CSSD) of the two-point CBS(4/3), CBS(5/4), and CBS(6/5) extrapolated dRPAc/HF energies for the test set of 65 hydrocarbons. Legend: Avg. means that we use the average $\alpha = 2.5608$ exponent in eq 6. Atom means that we calculate the value of $\alpha$ for every molecule according to eq 7 using the parameters from Table 1 and use this exponent in eq 6. Hybr. means that we calculate the value of $\alpha$ for every molecule according to eq 8 using the parameters from Table 1 and use this exponent in eq 6. The $\alpha = 2.5608$ extrapolation coefficients obtained this way are used in eq 5 to obtain dRPAc/HF/CBS(X/(X-1)) energies.

(down from 1.44 to 0.44 kcal mol$^{-1}$). Figure 5 shows that application of the hybridization corrected exponents via eq 8 improves the dRPAc/CBS(4/3) energy extrapolation in every statistical aspect and makes it practically equivalent to the very expensive CBS(6/5) energy extrapolation for the present test set (see also Figures S5 and S6). Figure 5 also shows that for dRPAc/CBS(5/4) extrapolation the atomic correction of the exponents is sufficient (eq 7), and no further improvements can be obtained from hybridization corrections (eqs 8 or 9). For dRPAc/CBS(6/5) extrapolation, the atomic and hybridization corrections of the exponents yield only small but well appreciable improvements as shown in Figure 5.

For practical CBS(6/5) extrapolation of the dRPAc energy we suggest using the average $\alpha = 2.5608$ exponent in eq 6 for larger hydrocarbons. For CBS(5/4) extrapolations we suggest using the $\alpha$ values calculated from eq 7 in eq 6. For CBS(4/3) extrapolations we suggest using the $\alpha$ values calculated from the eqs 8 or 9 in eq 6.

**CONCLUSION**

Accurate and precise calculation of the dRPAc energy requires very large basis sets, as observed earlier in the literature. In this paper, we have selected a test set of 65 hydrocarbon molecules in order to provide insight into the basis set convergence problems. The DF-dRPA correlation (dRPAc) energy is calculated using an efficient algorithm which is based on the CC-like form by using antisymmetrized two-electron integrals in the Riccati equations. These equations are solved in an iterative procedure by the density-fitted form of electron repulsion integrals and Cholesky decomposition of the orbital energy denominators ($O(n^4)$-scaling).

The accurate CBS extrapolation of the HF energy is given in the literature. In this paper, we focus on the CBS extrapolation of dRPAc energies using Dunning's correlation consistent aug-cc-pVXZ basis sets (noted here as AXZ) with aug-cc-pVXZ-R1 auxiliary basis sets for the correlation calculations ($X = \{3, 4, 5, 6\}$), respectively, where $X$ is the cardinal number of the basis set. We use the self-consistent HF wave function for dRPAc energy calculation noted as dRPAc/HF.

The dRPAc energy converges slowly, and even the dRPAc/A6Z energies are far from being converged. We have tested 3 functions for the complete basis set (CBS) energy extrapolation of dRPAc energy with respect to the X cardinal number of the AXZ basis sets. We have found that the popular inverse cubic extrapolation formula fits poorly to dRPAc/AXZ energies, giving well over 1 kcal mol$^{-1}$ error for the CBS extrapolated energies for the tested molecules. The optimized exponential extrapolation gives a considerably better fit. However, the best fit was obtained for an inverse power formula with optimized exponent. We tested the accuracy and the precision of the dRPAc/CBS energies for an averaged inverse power exponent $\alpha = 2.5608$.

We observed that the C atom and the hydrocarbon molecules show different convergence for dRPAc, and a simple atomic correction of the inverse power exponent leads to considerable improvement for the dRPAc/CBS(4/3) or CBS(5/4) energy extrapolations. The dRPAc/CBS(4/3) energies improved further by considering the hybrid states of the C atoms in the molecules: the mean absolute deviation (MAD) obtained by the averaged inverse power exponent is reduced to one-third (MAD = 0.13 kcal mol$^{-1}$). The other statistical parameters, like minimum and maximum deviations from the mean deviation and the corrected sample standard deviations, improve greatly by considering atomic hybridization states in the molecules. dRPAc/CBS(4/3) energies that account for the hybridization information in the molecule via eq 8 reach the accuracy and precision of the very expensive dRPAc/CBS(6/5) energies that use the averaged inverse power exponent $\alpha = 2.5608$.

Our results show that basis set convergence of the dRPAc energy also depends on the reference orbitals. In general, using PBE determinant instead of the HF determinant speeds up the basis set convergence for molecules containing a large fraction of hydrogen atoms. A large fraction of carbon or nitrogen atoms slows down the basis set convergence and also diminishes the difference between the application of the HF or PBE determinants. The MP2 correlation energy converges considerably faster with the basis set quality increase, and the difference between the molecules is considerably smaller in this respect. For methane and 1,3,5-hexatriyne molecules, we obtained exponents that are close to the widely used inverse cubic.

Our results suggest that A3Z and A4Z basis sets have some atom specific convergence properties in the dRPAc/CBS energy extrapolations and that can be corrected by the method proposed here. Possibly redesigned AZX basis sets also could solve this problem. We noticed that atomic or hybrid corrections of the inverse power exponent do not improve considerably the dRPAc/CBS(6/5) energies, as these ASZ and A6Z basis sets are considerably more complete than A3Z. Our results also show that the optimized exponents that yield accurate results for atoms or small molecules might be inaccurate for larger molecules, as the basis set convergence depends on the size and structure of the molecules for dRPAc/CBS(4/3) extrapolations. For practical CBS(6/5) extrapolation of the dRPAc energy of hydrocarbons we suggest using the average $\alpha = 2.5608$ exponent in eq 6. For CBS(5/4) extrapolations we suggest using the $\alpha$ values calculated from eq 7 in eq 6. For CBS(4/3) extrapolations we suggest using the $\alpha$ values calculated from eqs 8 or 9 in eq 6.
Figures are provided for the average fitting errors calculated with aug-cc-pVXZ basis sets ($X = 3–6$) for 65 hydrocarbon isomers, the actual and the predicted exponents of the inverse power function, for the correlation between the predicted and the fitted exponents of the slightly shifted ($d = 0.165$) power function, and for the mean absolute deviations of the predicted from the fitted complete basis set extrapolated dRPA correlation energies using the slightly shifted ($d = 0.165$) inverse power function formula. Tables are provided for the atomic compositions, the hybrid states, and for the dRPA energies calculated with HF SCF wave function using AXZ basis set, where $X = 3–6$, and the extrapolated CBS energies, and for the average and atomic parameters values for estimated slightly shifted ($d = 0.165$) inverse power function exponents. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00269.
Construction and Application of a New Dual-Hybrid Random Phase Approximation

Pál D. Mezei,*,† Gábor I. Csonka,*,† Adrienn Ruzsinszky,‡ and Mihály Kállay§

†Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, H-1521 Budapest, Hungary
‡Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States
§MTA-BME Lendület Quantum Chemistry Research Group, Department of Physical Chemistry and Materials Science, Budapest University of Technology and Economics, P.O. Box 91, H-1521 Budapest, Hungary

ABSTRACT: The direct random phase approximation (dRPA) combined with Kohn–Sham reference orbitals is among the most promising tools in computational chemistry and applicable in many areas of chemistry and physics. The reason for this is that it scales as \(N^4\) with the system size, which is a considerable advantage over the accurate ab initio wave function methods like standard coupled-cluster. dRPA also yields a considerably more accurate description of thermodynamic and electronic properties than standard density-functional theory methods. It is also able to describe strong static electron correlation effects even in large systems with a small or vanishing band gap missed by common single-reference methods. However, dRPA has several flaws due to its self-correlation error. In order to obtain accurate and precise reaction energies, barriers and noncovalent intra- and intermolecular interactions, we construct a new dual-hybrid dRPA (hybridization of exact and semilocal exchange in both the energy and the orbitals) and test the performance of this new functional on isogyric, isodesmic, hypohomodesmotic, homodesmotic, and hyperhomodesmotic reaction classes. We also use a test set of 14 Diels–Alder reactions, six atomization energies (AE6), 38 hydrocarbon atomization energies, and 100 reaction barrier heights (DBH24, HT-BH38, and NHT-BH38). For noncovalent complexes, we use the NCCE31 and S22 test sets. To test the intramolecular interactions, we use a set of alkane, cysteine, phenylalanine-glycine-glycine tripeptide, and monosaccharide conformers. We also discuss the delocalization and static correlation errors. We show that a universally accurate description of chemical properties can be provided by a large, 75% exact exchange mixing both in the calculation of the reference orbitals and the final energy.

INTRODUCTION

Within the direct Random Phase Approximation (dRPA) framework, the addition of the exact exchange (EX) and the dRPA correlation (dRPAc) results in an orbital dependent functional, which has many desirable features. dRPAc can be constructed either in the framework of density functional theory or in the ring coupled cluster formalism. The fully nonlocal dRPAc seamlessly integrates dispersion interactions, the EX is free of self-interaction error (SIE), and dRPA is applicable to small-gap or metallic systems because the “ring” diagrams are summed to finite order. dRPA is accurate for intra- and intermolecular noncovalent interactions, for adsorption, for interlayer interactions, and for van der Waals bonded solids. However, the method has several serious disadvantages because the dRPAc energy is spoiled by the self-correlation error, the total energies are systematically and significantly overestimated in magnitude as the dRPAc overestimates the correlation at short-range. Consequently, dRPA fails seriously in the description of changes which lead to short-range rearrangement of the electronic structure. RPA performs moderately for covalent bond breaking reaction energies, and it underestimates the binding energies. RPA overestimates the ionization potentials and underestimates the atomization energies of molecules and solids.

There is also no practical self-consistent dRPA implementation in the adiabatic-connection fluctuation–dissipation theorem (ACDFT) density functional approach. The nonsel-

Received: May 7, 2015
Published: August 31, 2015
the mean absolute deviation from the reference energies is abbreviated as MAD. Paier et al. reported that (EX + dRPA)@PBE provides accurate results for the barrier heights of the hydrogen transfer reactions in the HT-BH38 database\(^\text{2,5}\) (\(\text{MD} = -0.2 \ \text{kcal mol}^{-1}\), \(\text{MAD} = 1.7 \ \text{kcal mol}^{-1}\)) but significantly underestimates the barrier heights of heavy atom, non-hydrogen transfer (NHT) reactions in the NHT-BH38 database (MD = -2.51 kcal mol\(^{-1}\), MAD = 2.89 kcal mol\(^{-1}\)) compared to the moderately accurate W1 level reference energies. Only modest improvements using the corrected (EX + dRPAc + SOSEX + rSE)@PBE were achieved for HT reaction barriers, where SOSEX stands for Second-Order Screened Exchange,\(^\text{22}\) and rSE refers to renormalized single excitations within the Rayleigh–Schrödinger perturbation theory.\(^\text{19}\) The NHT reaction barriers are even worsened due to the SOSEX and rSE corrections.\(^\text{20}\) The SOSEX correction to dRPA about halves the large underbinding atomization error of dRPA.\(^\text{23}\) Adding rSE correction to dRPA + SOSEX further improves the atomization energies.\(^\text{20}\) Notice also that dRPA + SOSEX is one-electron self-correlation free.\(^\text{24}\) Ren et al. suggested\(^\text{25}\) replacing the EX@PBE reference energy with the HF total energy. This HF + (dRPAc)@PBE hybrid method (dRPAh) improves binding energies of molecules and solids by almost 50% compared to (EX + dRPAc)@PBE.

Sai et al. observed\(^\text{26}\) that a large fraction of exact exchange in the PBE hybrid can eliminate the consequences of the many-electron SIE. To achieve many-electron SIE-freedom, the ground state total energy \(E(N)\) should be linear in the electron number \(N\) between integers and should have a realistic slope.\(^\text{26–29}\) This condition was achieved for benzene, naphthalene, and anthracene by mixing 0.7 fraction of exact exchange with 0.3 fraction of PBE exchange in a PBE hybrid.\(^\text{25}\) This hybrid gives an accurate description for ionization and dimer dissociation.\(^\text{25,27}\) The predicted formation energies of both localized and delocalized holes in perfect molecular crystals due to ionization are consistent with experimental values.\(^\text{25}\)

Atalla et al. proposed\(^\text{30}\) a scheme to obtain an unambiguous system-dependent fraction of exact exchange for hybrid DFT that is consistent with the G0W0 approach, where G0 is the noninteracting Green’s function of the system and W0 is the screened Coulomb interaction. To achieve this the exact condition of DFT was used: the energy of the highest occupied molecular orbital corresponds to the ionization potential.\(^\text{31}\) The approach which uses 0.8 fraction of exact exchange is essential for describing electron transfer in the TTF/TCNQ dimer and yields the vertical ionization potentials of the G2 test set with a mean absolute percentage error of only \(\approx 3\%\).

Macher et al. compared\(^\text{32}\) the performance of the (EX + dRPAc)@PBE and the hybrid HF + (dRPAc)@PBE for ice. The (EX + dRPAc)@PBE underestimates the binding energies and overestimates all predicted volumes. The HF + (dRPAc)@PBE overestimates the binding energy and underestimates the equilibrium volumes. A simple solution for this problem suggested by Macher et al. is to determine the exact exchange energy from orbitals of a functional in which half of the HF exchange is mixed with half of the DFT exchange.

Hefele\textsuperscript{mann} proposed the RPAX2 method based on PBE exchange (PBEx) reference orbitals.\(^\text{33}\) RPAX2 is obtained by a simple modification of the ring coupled-cluster doubles amplitude equation and by a slightly modified update equation for the amplitudes. Numerical tests for chemical reaction energies and intermolecular interaction energies have shown that the RPAX2 method, if it is based on a PBEx Kohn–Sham reference determinant, yields results which are very close to coupled-cluster with single, double, and perturbative triple excitations (CCSD(T)) reference results.

Using the renormalized perturbation theory, Bates and Furche developed a promising beyond-RPA exchange kernel (AXK) method.\(^\text{34}\) AXK yields a significant improvement for atomization energies and ionization potentials compared to dRPA without affecting reaction barriers. AXK improves the binding energy of \(H_2\) compared to SOSEX but reintroduces some self-interaction error in the binding energy curve of \(H_2^+\).

In this paper, the exact exchange and dRPA correlation energies are calculated based on self-consistent HF, PBE, or PBE hybrid orbitals. We develop a new dual-hybrid dRPA functional and test it on test sets for chemical reaction energies, atomization energies, barrier heights (not involving the zero-point vibration energies), intermolecular weak interactions, and intramolecular noncovalent interactions. We use high level reference energies up to CCSDT(Q)/CBS and compare the performance of the new dual-hybrid functional to the performance of dRPA based on HF and DFT determinants, RPAX2 and double hybrid methods. We use MD as a measure of the accuracy and the corrected sample standard deviation \(\text{CSSD} = \frac{\left(\sum (E_i - \overline{E})^2/(N-1)\right)^{1/2}}{N}\), where \(N\) is the number of elements in the sample) as a measure of the precision. Notice that a small MAD signals that a method is accurate and precise at the same time, but larger MAD above 1–2 kcal mol\(^{-1}\) does not always characterize sufficiently the problem with a method (e.g., underbinding or overbinding, etc.).

#### METHODOLOGY

In the semilocal (sl) global hybrid (gh) density functional theory, the exchange-correlation energy, \(E^\text{gh}_{\text{XC}}\), is given by mixing the sl exchange, \(E^\text{sl}_{\text{XC}}\) with the exact exchange energy, \(E^\text{ex}_{\text{XC}}\), and adding a sl correlation energy, \(E^\text{sl}_{\text{C}}\).

\[
E^\text{gh}_{\text{XC}} = a E^\text{ex}_{\text{XC}} + (1-a) E^\text{sl}_{\text{XC}} + E^\text{sl}_{\text{C}}
\]

where \(a\) is the mixing factor.

In a global double hybrid (dhh) functional, the correlation energy is expressed by mixing the semilocal correlation energy with the second-order perturbation theory correlation energy, \(E^\text{CT}2\), calculated on the Kohn–Sham (KS) orbitals.

\[
E^\text{dhh}_{\text{C}} = a_k E^\text{ex}_{\text{XC}} + (1-a_k) E^\text{d}_{\text{C}} + a_c E^\text{PT}2_{\text{C}} + (1-a_c) E^\text{sl}_{\text{C}}
\]

These functionals perform well for molecules but suffer from problems like MP2 as they overestimate the binding energy of systems with small HOMO–LUMO energy gaps. Here we examine three popular functionals based on PT2 (B2PLYP),\(^\text{35}\) spin-component scaled PT2 (DSD-BLYP),\(^\text{36}\) and scaled opposite spin PT2 (PWBP95)\(^\text{37}\).

The direct random phase approximation (dRPA) exchange-correlation energy is given by eq 2

\[
E^\text{RPA}_{\text{XC}} = E^\text{ex}_{\text{XC}} + E^\text{RPA}_{\text{C}}
\]

where the dRPA correlation energy is given by the following equation

\[
E^\text{RPA}_{\text{C}} = \frac{1}{2} \text{tr} [B T]
\]

where \(B\) is the nonantisymmetrized two-electron repulsion integral matrix defined by the \(B_{ijkl} = \langle i|j|l|k\rangle\) four-index matrix.
elements, and $T$ is the double excitation amplitude matrix given by the iterative procedure (initialized with $T^{(0)} = 0$) as a Hadamard product with $\Delta$:

$$T^{(n+1)} = -\Delta \phi (B + BT^{(n)} + T^{(n)}B + \Delta \phi BT^{(n)})$$  \hspace{1cm} (5)

The $\Delta$ matrix is defined by $\Delta_{ji} = 1/(\epsilon_i + \epsilon_j - \epsilon_i - \epsilon_j)$ matrix elements. The equations are solved in an $O(N^4)$-scaling iterative procedure by density-fitted form of electron repulsion integrals and Cholesky decomposition of the orbital energy denominators $\Delta$.\(^{38}\) For small HOMO–LUMO gap (e.g., metallic) systems a plasmon-formula-based algorithm is executed in order to prevent unphysical solution or divergence as described in ref $38$.

The RPAX2 correlation energy is defined by eq 6, where we suppose that $B$ is written in density-fitted form as $B = LL^T$.

$$E_{\text{RPAX}2}^C = \frac{1}{2} tr[L^T T^{(oo)} L]$$  \hspace{1cm} (6)

The iteration is initialized with $T^{(0)} = 0$, then the amplitudes of the iteration cycles are given by

$$T^{(n+1)} = -\Delta \phi [(1 + T^{(n)})B(1 + T^{(n)}) - \hat{P}(1 + T^{(n)})B(1 + T^{(n)})]$$  \hspace{1cm} (7)

where $\hat{P}$ is a permutation operator that permutes the orbitals, $\Delta$ is decomposed using the Cholesky decomposition. The equations are solved in an $O(N^4)$-scaling iterative procedure.

RPA correlation energy can be calculated with self-consistent DFT or HF calculations, and we also use reference orbitals from self-consistent hybrid DFT calculations. In this work, we propose a scheme of hybrid (eq 8) and double hybrid (eq 9) dRPA energies evaluated using reference orbitals from self-consistent hybrid calculations.

$$E_{\text{XC}dRPA}^{\text{gh}} = a_X E_{\text{XC}}^{\text{exact}} + (1 - a_x)E_{\text{XC}}^{\text{dRPA}} + E_{\text{XC}}^C$$  \hspace{1cm} (8)

$$E_{\text{XC}dRPA}^{\text{ghh}} = a_X E_{\text{XC}}^{\text{exact}} + (1 - a_x)E_{\text{XC}}^{\text{dRPA}} + a_{\mu} E_{\text{XC}}^{\text{dRPA}} + (1 - a_{\mu})E_{\text{XC}}^C$$  \hspace{1cm} (9)

All energy components are evaluated using self-consistent PBE hybrid orbitals characterized by the fixed $\alpha$ mixing parameter from eq 1.

### COMPUTATIONAL DETAILS

The single-point energy calculations were performed using the aug-cc-pVQZ basis sets (denoted here as AZX), where $X$ was T, Q, 5, or 6. To accelerate the calculations we used the corresponding AXZ-RJ-JK and AXZ-RJ auxiliary basis sets for the SCF and correlation calculations, respectively. To consider the basis set error we extrapolated the exact exchange and the dRPA correlation parts to the CBS limit according to eqs 10 and 11 similarly to the two-point extrapolations applied in refs 39 and 40.

$$E_{\text{XC \text{CBS} (X,x-1)}}^{\text{EX}} = E_{\text{AXZ}}^{\text{EX}} + C_{\text{AXB}X}^{\text{EX}} (E_{\text{AXZ}}^{\text{EX}} - E_{\text{AXB}X}^{\text{EX}})$$  \hspace{1cm} (10)

$$E_{\text{RPA-CBS (X,x-1)}}^{\text{RPA-CBS (AXZ)}} = E_{\text{RPA-CBS (AXZ)}}^{\text{RPA-CBS (AXZ)}} + C_{\text{RPA-CBS (AXZ)}}^{\text{RPA-CBS (AXZ)}} (E_{\text{RPA-CBS (AXZ)}}^{\text{RPA-CBS (AXZ)}} - E_{\text{RPA-CBS (AXZ)}}^{\text{RPA-CBS (AXZ)}})$$  \hspace{1cm} (11)

where $C_{\text{EX}}^{\text{X}} = 0.274$, as well as $C_{\text{RPA}}^{\text{X}}$ is equal to 0.917, 0.856, and 0.867 for (dRPA)@H, (dRPA)@PBE, and (dRPA)@PBE@2S, respectively.

The dRPA and RPAX2 single-point energies were calculated using the MRCC program code.\(^{41}\) For comparison, the dispersion corrected double hybrid energies were partly taken from the literature and partly calculated with the Orca 3.0.1\(^{42}\) quantum chemistry program using the ATZ basis set with the auxiliary ATZ-RJ-JK and ATZ-RJ (denoted as ATZ/JK and ATZ/C in Orca) basis sets.

### TEST SETS

For quick testing, we suggest a very small but representative test set ($\text{shC5}$) from five reactions of ethylene, acetylene, propane, propene, and butadiene taken from ref 43. The reactions represent $\pi \rightarrow \sigma$ bond transformations of sp$^2$ or sp$^3$ carbon atoms in the molecules, protobranching reactions, and $\sigma$- or $\pi$-conjugation breaking reactions. The optimized B3LYP/6-31G(2df,p) geometries of these molecules were taken from the NIST Computational Chemistry Comparison and Benchmark Database.\(^{44}\) For reference energies, we calculated high-quality extrapolated CCSD(T)-full/CBS energies. (This test set is used in Figures 1, S1 and Tables 1, S1, S2.)

![Figure 1](image-url)


Table 1. Benchmark-Quality Estimated Full CCSDT(Q)/CBS Reaction Energies (kcal mol\(^{-1}\)) and Self-Consistent dRPA@HF, dRPA@PBE0.25, dRPA@PBE, dRPAh, and dRPA75 (Eq 12) Deviations from Benchmark Energies (\(E - \Delta E_{CCSDT(Q)}\)) of the sHC5 Set\(^{c}\)

<table>
<thead>
<tr>
<th>system</th>
<th>(\Delta E_{CCSDT(Q)})</th>
<th>(\Delta \Delta E_{dRPA@HF}^{a})</th>
<th>(\Delta \Delta E_{dRPA@PBE0.25}^{a})</th>
<th>(\Delta \Delta E_{dRPA@PBE}^{a})</th>
<th>(\Delta \Delta E_{dRPAh}^{a})</th>
<th>(\Delta \Delta E_{dRPA75}^{a})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-21.32</td>
<td>-2.02</td>
<td>-0.16</td>
<td>1.63</td>
<td>1.23</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>-30.82</td>
<td>-2.85</td>
<td>-1.07</td>
<td>0.48</td>
<td>-1.27</td>
<td>-0.05</td>
</tr>
<tr>
<td>3</td>
<td>2.08</td>
<td>-0.39</td>
<td>-0.24</td>
<td>-0.18</td>
<td>0.04</td>
<td>-0.14</td>
</tr>
<tr>
<td>4</td>
<td>4.81</td>
<td>-0.34</td>
<td>-0.51</td>
<td>-0.80</td>
<td>-0.35</td>
<td>-0.10</td>
</tr>
<tr>
<td>5</td>
<td>13.06</td>
<td>-0.98</td>
<td>-1.27</td>
<td>-1.88</td>
<td>-0.44</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

\(^{a}\)Mean deviation. \(^{b}\)Mean absolute deviation. \(^{c}\)Corrected sample standard deviation. \(^{d}\)Minimum. \(^{e}\)Maximum. dRPAh denotes the HF + (dRPAc)@ PBE hybrid. \(^{f}\)The CBS extrapolations are based on the ATZ and AQZ energies. The dRPA75 calculations were performed with the ATZ basis set. All calculations were performed on B3LYP/6-31G(2df,p) optimized geometries.

C\(_2\)H\(_4\) + C\(_2\)H\(_6\) 4.81
2C\(_2\)H\(_4\) + C\(_2\)H\(_6\) 13.06

C\(_2\)H\(_4\) + C\(_2\)H\(_6\) 4.81
2C\(_2\)H\(_4\) + C\(_2\)H\(_6\) 13.06

The dRPA75, dRPA@PBE0.25, dRPA@PBE, and dRPAh deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set are shown in Figure 3. The reaction energies of 38 hydrocarbons (the RC0 test set) were used in this study. The test set consists of 14 Diels–Alder reactions, abbreviated as DARC. These are condensation reactions of dienes like butadiene, cyclopentadiene, cyclo-

hexadiene, and furane with dienophiles like ethene, ethyne, maleine, and maleimide. The majority of the DFT functionals underestimate the noncovalent intramolecular interactions, which occur in the larger bicyclic and tricyclic products leading to endothermic reaction energy error. Johnson et al. observed a systematic improvement in the calculated reaction energies for functionals which treat fractionally charged systems more correctly (rCAM-B3LYP and MCY3). Based on this observation, it was supposed that the DARC test set is suitable for testing the electron delocalization error in DFT. Our recent study\(^{37}\) has shown that the DARC test set is not particularly suitable for identifying the delocalization error caused by the self-interaction error, but it is an excellent test set for intramolecular interactions and bond reorganizations. This is because the reactions in the DARC test set are influenced by many effects, e.g., double (ethene) and triple (ethyne) bonds are transformed into single and double bonds, the electron delocalization in the dienes is destroyed, double and triple ring formations, etc.\(^{37}\) (This test set is used in Figures 3, S5, S6.)

Figure 2. MAD(method)/MAD(dRPA75) ratios of B2PLYP-D3(BJ), dRPA@HF, and dRPA@PBE methods and the dRPA75 MAD values in kcal mol\(^{-1}\) for RCI-RC5 \(n\)-homodesmotic hierarchy of reactions. Smaller ratio means better performance. For the RCI the dRPA@HF yields a very large, out of chart MAD (11.6 kcal mol\(^{-1}\)). Abbreviations: RCI, isogyric; RC2, isodesmic; RC3, hypohomodesmotic; RC4, homodesmotic; RC5, hyperhomodesmotic.

Figure 3. The dRPA75, dRPA@PBE, dRPA@PBE0.25, and dRPA@HF deviations from the benchmark reaction energies of the DARC Diels–Alder reaction test set. Abbreviations: ATZ, aug-cc-pVTZ basis set; CBS, complete basis set.

\[4618\] DOI: 10.1021/acs.jctc.0c00420
For testing the performance on reaction barriers, we use the accurately known DBH24 test set, which is composed of four subsets: BH6 for hydrogen transfer; HAT6 for heavy atom transfer; NS6 for nucleophilic substitution; UR6 for unimolecular and recombination reactions. For BH6, we use the estimated CCSDT(Q)-full/CBS reference values,55 and for the other three test sets, we use the energies obtained from the W3.2 and W4 ab initio computational thermochemistry protocols, which have a 95% confidence interval well below 0.2 kcal mol⁻¹. We also use the less accurately known but considerably larger HT- and NHT-BH38 reaction barrier test sets.51 (These test sets are used in Tables 3, 4, S6, S7 and Figures 5, S8.)

For testing the performance on intermolecular interactions, we used the noncovalent complexes of the NCCE31 test set (with MC-QCISD/3 reference energy values), which contains six hydrogen-bonded, seven charge transfer, six dipole–dipole, seven weak interacting, and five π–π stacking complexes. We also included the biologically important complexes of the S22 database54 (with estimated CCSD(T)/CBS reference values), which contains seven hydrogen bonded complexes, eight dispersion dominated complexes, and seven mixed complexes. (These test sets are used in Figures 6, 7, S9 and Table S8.)

Furthermore, we used four databases to test the performance for intramolecular interactions. The aliphatic dispersion interaction is tested using a set of all alkane conformers (ACONF)36 which compares the relative energies among two butane, three pentane, and 11 hexane conformers with WH-val reference values. A set of cysteine conformers (CYCONF),57 a set of phenylalanine-glycine-glycine tripeptide conformers (PCONF),58 and a set of sugar derivatives (SCONF)59 test various kinds of intramolecular hydrogen bonds, all with estimated CCSD(T)/CBS reference values. (These test sets are used in Figure 8 and Table S9.)

Common semilocal density functional approximations underestimate the energy of systems with fractional charge.60 Thus, the delocalization of electrons is preferred, and this error is called delocalization error. Fractional charge is present e.g. in transition states and leads to the underestimation of the barrier heights. The prototype indicator of the delocalization error is the potential energy curve of the dissociating H₂ molecule.61 Similarly, the energy of systems with fractional spin is overestimated by the static correlation error. Fractional spin can occur, e.g. by the breaking of chemical bonds. The prototype indicator of this error is the potential energy curve of the dissociating H₂ molecule.61 (These systems are used in Figure S7 and Table S3.)

### RESULTS AND DISCUSSION

#### New Dual-Hybrid Method

Obtaining good results for chemical reaction energies and barriers with the same functional is a challenge for semilocal DFT methods. We have selected two very small but representative test sets (sHC5 and BH6) to develop a new method for both reaction energies and barriers. As we develop a model chemistry which is useful for the largest molecules possible, we use the relatively small ATZ basis set, which gives consistent results. We present all the details of the parameter optimization in Tables S1a-d, S6a-d.
and Figure S1 of the Supporting Information. For either the sHCS or the BH6 test set, the global optimum in the $a$, $a_X$, and $a_C$ parameter space of eqs 1 and 9 leads to only one accurate and precise dual-hybrid method (shortened as dRPA75).

$$E_{XC}^{\text{dRPA75}} = (0.75E_X^{\text{exact}} + 0.25E_X^{\text{PBE}} + E_C^{\text{dRPA}})@PBE0.75/ATZ$$

(12)

For this hybrid, we use a PBE0.75 global hybrid orbital reference with $a = 0.75$ in eq 1. The beneficial effects of such large 0.7–0.8 fraction of exact exchange were also described in refs 25 and 30. The same 0.75 fraction is optimal for $a_X$ in eq 8 combined with full dRPA correlation ($a_C = 1$ in eq 9).

Notice that mixing semilocal PBE correlation with the fully nonlocal dRPA correlation does not lead to improvement thus $a_C = 1$ is optimal in eq 9. Figure 1 shows the distribution of the optimal values (within 0.2 kcal mol$^{-1}$ deviation from the locally minimal MAD) for the sHCS and BH6 test sets in the $a$, $a_X$, and $a_C$ parameter space. The overlapping regions in Figure 1 are around the dRPA@PBE and dRPA75 methods, but the latter has improved accuracy for both the reaction energies and the barrier heights. This finding underlines the good properties of PBE hybrid orbitals with a large fraction of exact exchange discussed in the literature.25,30

Also notice that the optimal $(1 - a_C) = 0.25$ weight factor of the nondynamic correlation correction ($E_X^{\text{corr}} - E_X^{\text{PBE}}$) to the exchange-correlation energy is much smaller for dRPA75 than
for the popular global hybrids (cf. 0.80 for B3LYP, 0.75 for PBE0.25, 0.46 for M06-2X) or double hybrids (0.5 for PWBP95, 0.47 for B2PLYP, 0.45 for mPW2PLYP). Table S1a and Figure S1 show a local optimum about 53% dRPA + 47% PBE correlation, which agrees well with our earlier suggestion of dRPA + α(PBE) (α = 0, α = 0.5, α = 0.5 in eq 9), but this functional shows considerably poorer performance for the HCS test set than the new dRPA75 hybrid. In the following, we present the performance of this new approach on the test sets introduced above.

Reaction Energies. Table 1 shows the estimated CCSDT(Q)-full/CBS reference energies and dRPA energy deviations for the HCS test set. The results show a particularly good performance for dRPA75/ATZ. For the HCS reaction energies, dRPA shows quite rapid basis set convergence as shown in Tables S2a and S2b. We note also that the RPAX2@PBEX/ATZ, AQZ, and ASZ energies are very accurate and precise (cf. Table S2c).

Figure 2 shows the radar or polar chart for the relative performance represented by the MAD(method)/MAD(dRPA75) ratios of B2PLYP-D3(BJ), dRPA@HF, dRPA@PBE methods and the dRPA75 MAD values in kcal mol\(^{-1}\) for RC1-RC5 n-homodesmotic hierarchy of reactions. The double hybrid functionals perform similarly to each other, so we show only the typical performance of the B2PLYP-D3(BJ) method. For the RC1 reaction class, dRPA75 is the best among the methods shown in Figure 2, while the MAD(dRPA@HF)/MAD(dRPA75) = 8 is out of the chart. (The error distributions are shown in Figure S2.) The dRPA@TPSS/cc-pVTZ results (MAD = 6.7 kcal mol\(^{-1}\)) are also considerably poorer than our dRPA75 results. The commonly used DFT methods like M06-2X and B3LYP yield large MADs (9.1 and 11.9 kcal mol\(^{-1}\), respectively). For the RC2 reaction class, dRPA75 is the best again and dRPA@HF has similar performance, while the dRPA@PBE and B2PLYP-D3(BJ) yield larger deviations. dRPA@TPSS/cc-pVTZ, M06-2X, and B3LYP yield MAD = 2.1, 2.4, and 4.2 kcal mol\(^{-1}\), respectively. For the RC3 reaction class, the performance of dRPA75 improves further, and the other methods discussed here have similar MADs within 20%. For RC1-RC5, the MAD of dRPA75 decreases monotonically. Other methods discussed here have similar MADs within 20%.

We show detailed results for the A, B, and C subsets of the five reaction classes in Figures S3a-f for dRPA@HF, dRPA@PBE, dRPA75, and dRPA@TPSS/cc-pVTZ and the D3 dispersion corrected B2PLYP, PWBP95, and DSD-BLYP (taken from ref 8). For the isogric RC1A test set, the D3 dispersion corrected double hybrid methods perform about the same as dRPA75, but dRPA75 shows much better performance for the RC1B and RC1C test sets. For the RC4 and RC5 reaction classes, the MADs of dRPA75 are in the insignificant range, 0.1–0.3 kcal mol\(^{-1}\).

Comparison of Figures S4b and S4c shows that the basis set convergence of the RC1 reaction energies is considerably slower than that of the RC2 reaction energies. For detailed analysis of the basis set convergence of the dRPA correlation energies, see ref 40.

Next we test the performance of the dRPA method with various reference orbitals on the DARC reaction energy test set. Figure 3 shows that our new dRPA75 method has about 2 kcal mol\(^{-1}\) endothermic reaction energy error with a good precision (CSSD = 0.7 kcal mol\(^{-1}\)). dRPA@PBE0.75 shows an overbinding error very similar (the average difference is around 0.1 kcal mol\(^{-1}\)) to dRPA@HF (cf. Figure S5a). Consequently, the 25% of PBE exchange in PBE0.75 leads to the above-mentioned underbinding error. The more diffuse PBE electron densities result in more endothermic reaction energies than the compact HF electron densities (cf. Figures S5a and S5b). The AQZ or ATZ basis set errors shift the DARC reaction energies in the exothermic direction (cf. Figures S5a-c).

Our recent results show that the systematic endothermic error of the dRPA@PBE/CBS reaction energies can be efficiently compensated by the systematic exothermic error of the AQZ basis set (MD = 0.08 kcal mol\(^{-1}\), MAD = 0.32 kcal mol\(^{-1}\) and CSSD = 0.42 kcal mol\(^{-1}\); see also Figure S5b), because on more diffuse reference orbitals the self-correlation error of the dRPA correlation overestimates the chemical delocalization in the reactants (see ref 47), and the cusp error stabilizes the diffuse electron densities in the products (see ref 40). Mixing PBE exchange with 0.25 fraction of exact exchange can also compensate the systematic endothermic error of dRPA@PBE/CBS reaction energies (cf. MD = −0.53 and MAD = 0.53 kcal mol\(^{-1}\) in ref 47). The more expensive RPAX2@PBEX/ATZ calculations lead to very precise reaction energies with a small exothermic error (MD = −0.83, MAD = 0.83, in ref 47). Standard functionals like B3LYP, PBE0, TPSS, or TPSSH yield inaccurate and imprecise results for the DARC reaction energies characterized by MAD = 15 kcal mol\(^{-1}\) and CSSD = 5 kcal mol\(^{-1}\). The highly empirical M06-2X functional is relatively accurate (MD = 1.92 kcal mol\(^{-1}\)) but considerably less precise (CSSD = 2 kcal mol\(^{-1}\)) than the new dRPA75. The best performing double hybrid functionals show similar performance to dRPA methods presented here as shown in Figure S6, but they are far from the accuracy and precision of the dRPA@PBE/AQZ for the DARC test set.

Delocalization and Static Correlation Errors. The delocalization error of the semilocal density functionals can be well represented by the dissociation of homonuclear diatomic radical cations (i.e., H\(^+_2\)) as the semilocal DFT incorrectly stabilizes the fractionally charged product over the integer charged products. In Figure S7a and Table S3a, we show the dRPA@HF, dRPA@PBE0.25, dRPA@PBE, and dRPA75 errors. Our earlier results show that the one-electron self-interaction error for RPAX2@PBEX yields perfect binding curve for H\(^+_2\). Another serious error of semilocal DFT is the so-called static correlation error represented by the H\(_2^+\) binding energy curve. Figure S7b and Table S3b show the dRPA@HF, dRPA@PBE0.25, dRPA@PBE, and dRPA75 errors. All dRPA methods discussed here underbind the H\(_2^+\) molecule. The underbinding increases from 2 kcal mol\(^{-1}\) up to 8 kcal mol\(^{-1}\) in the order of dRPA@PBE < dRPA75 < dRPA@HF (Table S3b). Again the RPAX2@PBEX gives the best agreement with the reference curve.

Atomization Energies. The underbinding error of dRPA/CBS is very large, around 20 kcal mol\(^{-1}\) as shown in Table 2 and Table S4a-b for the AE6 test set. The underbinding error of the ATZ basis set increases this error further by 11 kcal mol\(^{-1}\) on average. It was shown in the literature that range separation can diminish the atomization energy errors. Another possibility is the AXK correction, but still it has similar
atomization energy errors to TPSS. We have also calculated the RPAX2@PBE@ATZ, AQZ, and ASZ atomization energies (Table S4c) and found that the CBS extrapolated atomization energies are quite accurate and precise for the AE6 test set. The RPAX2@PBE@ atomization energies also converge well with the basis set quality. The underbinding error of the ASZ basis set efficiently compensates the small overbinding error of the RPAX2@PBE method.

Next we analyze the errors of the calculated atomization energies for the 38 hydrocarbons in the RC0 test set. Table S5 shows the statistical data of the deviations and the relative deviations of the calculated atomic energies from the reference energies for dRPA@HF, dRPA@PBE, dRPA75, and B2PLYP-D3(BJ). The dRPA75 method is particularly precise for atomization energies (CSDS = 0.07%) but underbinds the molecules by slightly more than 5%. The B2PLYP-D3(BJ) method is very accurate but much less precise than the dRPA75. The dRPA@PBE and dRPA@HF methods are considerably less precise. This means that dRPA75 atomization energy error is systematic, even more systematic than that of the accurate B2PLYP-D3(BJ), thus the relative energies of the molecules are quite correct leading to good chemical reaction energies as it was the case for the RCn test sets analyzed above.

As we stated previously: “The atomization energy errors of a density functional can strongly magnify minor errors in its spin-polarization dependence which are of little importance for other properties of many molecules and solids. Thus, density functionals should not be judged primarily by their atomization energy errors, but by a wider spectrum of tests.”

Due to the systematic nature of the atomization energy errors an efficient correction is possible by optimized atomic energies (CAE = corrected atomic energies). Figure 4 shows these corrected atomization energies for dRPA@HF, dRPA@PBE, dRPA75, and D3 dispersion corrected B2PLYP methods. (Notice the much smaller corrections for double hybrids compared to dRPA.) These a posteriori fitted atomic energies considerably reduce the atomization energy errors for the RC0 test set, but the imprecision of the double hybrid methods cannot be corrected as efficiently as the more systematic errors of the dRPA methods. Excellent results can be obtained from dRPA75-CAE, dRPA@PBE-CAE, and dRPA@HF-CAE.

As we noted for the AE6 test set, the dRPA atomization energies are very sensitive to the basis set error. In Figure S4a, we show the very large basis set errors of (dRPAc)@HF with the ATZ basis sets for nine instances and the convergence of the atomization energies as a function of the increasing cardinal numbers up to six. Even the A6Z basis set shows considerable dRPAc@HF atomization energy error for larger hydrocarbons like pentane or other pentane derivatives.

**Barrier Heights.** Popular DFT methods seriously underestimate the reaction barrier heights, while the HF method seriously overestimates them. Some kind of uncertain error compensation might lead to somewhat improved results for global hybrid functionals, but the improvement is unreliable. Also the large fraction of exact exchange needed for good reaction barriers worsens the results for reaction energies. Application of a very large number of empirical parameters might lead to somewhat improved results at the cost of a wavy energy surface (e.g., M05-2X and M06-2X functionals). Table 3 shows the estimated CCSD(T)(Q)-full/CBS reference energies and deviations for various dRPA methods for the BH6-DHB24 test set. These results show the typical dRPA@HF errors (in agreement with the values reported earlier). The hybrid dRPAh fails to deliver good results for reaction barriers. This is because the average HF/CBS exchange energy error is 12.5 kcal mol⁻¹ in agreement with ref 66. This error is decreased by the (dRPAc)@HF correlation energy for the reaction barriers by about 7 kcal mol⁻¹. Replacing the HF orbitals by PBE orbitals increases the average exact exchange energy error to 19.1 kcal mol⁻¹. This error is almost perfectly compensated by the average (dRPAc)@PBE correlation energy for the reaction barriers leading to very accurate dRPA@PBE energies. However, correcting the too small average HF/CBS exchange energy error with the same average (dRPAc)@PBE correlation energy for the reaction barriers leads to serious underestimation of the barrier heights, and this makes the dRPAh method unsuitable for reaction barrier height calculations. Comparison of the barrier heights in Tables 3, S7a and S7b shows that the calculated dRPA barrier heights are quite insensitive to the basis set errors, and the ATZ and AQZ results are practically converged with respect to the basis set size. Inspection of Table 3 reveals that despite the excellent accuracy, the dRPA@PBE/CBS results are quite imprecise. The dRPA@PBE0.25/CBS results are less accurate but more precise than the dRPA@PBE/CBS results.

The most accurate and precise barrier heights by far are obtained again with dRPA75/ATZ. These results are very close to the CBS results, as good as the CBS-Q, MC-QCISD/3 results, much better than the MP2, MP3, MP4, CCD, or QCISD(T) results and also considerably better than the results obtained with the MPWB1K method empirically optimized for reaction barriers (MAD = 1.32 kcal mol⁻¹). We have also calculated the considerably more expensive RPAX2@PBEx/ATZ, AQZ, and ASZ barrier heights (see Table S7c) and obtained a systematic overestimation independently of the basis set quality. This shows one of the limitations of the RPAX2@PBE method.

The overall good performance of the dRPA75/ATZ method for reaction barriers can be followed in Tables S7d-f for the HAT6-, NS6-, and U6-DHB24 test sets. The overall performances for the 24 barrier heights are summarized in Table 4. The best method is dRPA75/ATZ followed closely by dRPA@PBE0.25/ATZ. The best range-separated RPA shows larger MAD (2.83 kcal mol⁻¹) for the DHB24 test set.

We have also included larger but less accurately known (for comparison see Table S7g) reaction barrier test sets into our study: the NHT-BH38 (Table S7h) and HT-BH38 (Table S7i) test sets of 76 barrier heights. The error distributions of the dRPA methods for the BH76 test set is shown in Figure 5. The

**Table 4. Statistics of the DBH24 Database for dRPA@HF, dRPA@PBE0.25, dRPA@PBE, and dRPA75**

<table>
<thead>
<tr>
<th>Method</th>
<th>ΔE_{dRPA@HF}^{DBH24}</th>
<th>ΔE_{dRPA@PBE0.25}^{DBH24}</th>
<th>ΔE_{dRPA@PBE}^{DBH24}</th>
<th>ΔE_{dRPA75}^{DBH24}</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD‡</td>
<td>5.87</td>
<td>6.65</td>
<td>2.11</td>
<td>1.42</td>
</tr>
<tr>
<td>CSSD</td>
<td>5.46</td>
<td>1.90</td>
<td>2.37</td>
<td>1.72</td>
</tr>
<tr>
<td>MinΔ</td>
<td>1.00</td>
<td>-2.72</td>
<td>-5.17</td>
<td>-2.24</td>
</tr>
<tr>
<td>MaxΔ</td>
<td>18.87</td>
<td>3.90</td>
<td>3.95</td>
<td>3.14</td>
</tr>
</tbody>
</table>

*Mean deviation. ‡Mean absolute deviation. †Corrected sample standard deviation. ‡Minimum. †Maximum. With estimated CCSDT(Q)-full/CBS reference values for the BH6 subset, W3.2 and W4 reference values otherwise. All calculations were performed with the ATZ basis set.*
new dual-hybrid dRPA75 is among the best performing methods (MAD = 1.38 kcal mol$^{-1}$), and it is also better than the AXK results (MAD = 1.63 kcal mol$^{-1}$). Note that although dRPA@PBE0.25 looks slightly better (MAD = 1.37 kcal mol$^{-1}$) than dRPA75 in Figure 5, the reference values have 0.3–0.4 kcal mol$^{-1}$ error, which makes this comparison uncertain (cf. the DBH24 and BH76 results in Figure S8a). (For the sake of completeness, the deviations from the benchmark reaction energies of the BH76 test set are compared in Figure S8b.)

Inter- and Intramolecular Interactions. First we tested our new method (dRPA75) on various types (charge transfer, π–π stacking, hydrogen bond, dipole interaction, and weak interaction) of intermolecular interactions of the NCCE31 database. Figure 6 shows the double hybrid GGA with nonlocal correction (taken from ref 67); dRPA@HF, dRPA@PBE, dRPA@PBE0.25, and dual-hybrid dRPA75 mean absolute deviations from the benchmark noncovalent complexation energies of the NCCE31 database. dRPA75 is much more accurate and precise for charge transfer interaction energies than any other method shown in Figure 6. The situation is similar for π–π stacking and hydrogen bond energies. The dispersion corrected double hybrid B2PLYP-NL or revPBE0-DH-NL functionals show considerably poorer performance than dRPA75. For dipole and weak interactions, dRPA75 is quite competitive with the other methods in Figure 6. The dual-hybrid dRPA75 gives well-balanced, accurate, and precise interaction energies for all types of complexes in this test set (all MAD values are below 0.2 kcal mol$^{-1}$). The detailed results for the components of the NCCE31 test set can be found in Tables S8a-f.

Then we also examined the performance of our new method on the S22 test set that contains several biologically important weak interactions. Figure 7 shows the MADs of the D3 dispersion corrected double hybrid GGA, dRPA, and dual-hybrid dRPA75 from the benchmark interaction energies of the complexes of S22 set for D3 dispersion corrected PWPB95, DSD-BLYP, and B2PLYP (taken from the GMTKN30 database), as well as dRPA@HF, dRPA@PBE, dRPA@PBE0.25, and dRPA75 (calculated with the ATZ basis set).

Figure 7. The mean absolute deviations (MAD) from the benchmark interaction energies of the complexes of S22 set for D3 dispersion corrected PWPB95, DSD-BLYP, and B2PLYP (taken from the GMTKN30 database), as well as dRPA@HF, dRPA@PBE, dRPA@PBE0.25, and dRPA75 (calculated with the ATZ basis set).

The calculation of the intramolecular interactions is even more challenging than that of the intermolecular interactions because of the nonbonded electron density overlaps as we have pointed out previously for the Diels–Alder reaction energies. To test the performance of the methods for intramolecular interactions, we have selected the well tested ACONF, CYCONF, PCONF, and SCONF test sets from the GMTKN30 database. For comparison, we have selected three well-performing double hybrid functionals, the D3 dispersion corrected B2PLYP, PWPB95, and DSD-BLYP. The performance of these functionals is known in the GMTKN30 database. Figure 8 shows that the performances of dRPA@HF, dRPA@PBE, and dRPA@PBE0.25 are not particularly good for these energies. However, dRPA75 dual-hybrid gives excellent results for ACONF and PCONF test sets and performs reasonably well for the other two test sets. Several double hybrids like B2PLYP-D3 and DSD-BLYP-D3 perform better for the CYCONF and SCONF test sets. Notice that the 0.2 kcal mol$^{-1}$ deviation is so small that it can be smaller than the errors in the reference energies. Further details about the performance of the new dual-hybrid dRPA75 can be found in Tables S9a-d.

Figure 8. The mean absolute deviations (MAD) from the benchmark relative conformational energies of the ACONF, CYCONF, PCONF, and SCONF test sets for D3 dispersion corrected PWPB95, DSD-BLYP, and B2PLYP (taken from the GMTKN30 database), as well as dRPA@HF, dRPA@PBE, dRPA@PBE0.25, and dRPA75 (calculated with the ATZ basis set).

In Table 5, we summarize the MAD values (in kcal mol$^{-1}$) for 27 test sets. We selected the B2PLYP-D3(BJ), dRPA@HF, dRPA@PBE, dRPA@PBE0.25, and dRPA75 methods for comparison. The B2PLYP-D3(BJ) results are taken from the literature. Inspection of the results shows that the dRPA75 method shows excellent overall performance (the smallest MAD values are in bold), and it is applicable for a wide range of chemical problems.
In this paper, we have developed a dual-hybrid method which is hybridized at two levels. First we calculate the self-consistent PBE0.75 reference orbitals yielded by a 0.75 fraction exact exchange containing PBE hybrid. Then we mix 0.75 fraction of exact exchange energy with 0.25 fraction of PBE exchange energy and add direct Random Phase Approximation, dRPA, correlation energy to obtain the dRPA75 energy. All energy components are calculated using the PBE0.75 hybrid reference orbitals. dRPA75 requires no extra computational effort compared to dRPA.

We used five reaction energies of small hydrocarbons, sHCS, and the six reaction barrier heights of the BH6 test set to analyze the dRPA hybridization errors. We have obtained extremely accurate and precise dRPA75 results even with the moderate augmented triple-ζ (ATZ) basis set for the reaction energies and barriers. Comparison to the best available methods for reaction energies (five homodesmotic reaction classes) and barriers (DBH24 and BH76 test sets) shows that dRPA75 is one of them, and it systematically improves the dRPA results. For Diels–Alder reactions (DARC) the dRPA75/ATZ leads to an average 2 kcal mol$^{-1}$ endothermic reaction energy error with a good precision. The origin of this error is that the 0.25 fraction of the PBE exchange energy overcorrects the large exothermic error of the dRPA@PBE0.75/ATZ. Conventional DFT methods perform very poorly for the DARC test set.

The analysis of the binding energy curves of H$_2^+$ and H$_2$ yields that dRPA75 gives typical dRPA results for equilibrium bond distance and energy but shows large one-electron self-correlation error and a moderate static correlation error. The one-electron self-correlation error free RPAX2@PBE6 yields perfect binding curve for H$_2^+$ and H$_2$. The beyond-RPA AXK removes most of the self-correlation of dRPA, but not all, and yields better description of static correlation than SOSEX.

### Table 5. Summary of Mean Absolute Deviations (MAD in kcal mol$^{-1}$) from Reference Energies for Dispersion Corrected B2PLYP-D3(BJ), dRPA@HF, dRPA@PBE0.25, dRPA@PBE, and dRPA75$^c$

<table>
<thead>
<tr>
<th></th>
<th>B2LYP-D3</th>
<th>dRPA@HF</th>
<th>dRPA@PBE0</th>
<th>dRPA@PBE</th>
<th>dRPA75</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reaction Energies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sHCS</td>
<td>1.32$^c$</td>
<td>0.65$^c$</td>
<td>0.99$^c$</td>
<td>0.23$^c$</td>
<td></td>
</tr>
<tr>
<td>RC1</td>
<td>2.62$^c$</td>
<td>11.60$^b$</td>
<td>3.05$^b$</td>
<td>1.45$^b$</td>
<td></td>
</tr>
<tr>
<td>RC2</td>
<td>2.09$^c$</td>
<td>1.13$^b$</td>
<td>1.75$^b$</td>
<td>1.16$^b$</td>
<td></td>
</tr>
<tr>
<td>RC3</td>
<td>0.46$^c$</td>
<td>0.32$^b$</td>
<td>0.48$^b$</td>
<td>0.42$^b$</td>
<td></td>
</tr>
<tr>
<td>RC4</td>
<td>0.15$^c$</td>
<td>0.26$^b$</td>
<td>0.45$^b$</td>
<td>0.27$^b$</td>
<td></td>
</tr>
<tr>
<td>RC5</td>
<td>0.27$^c$</td>
<td>0.21$^b$</td>
<td>0.09$^b$</td>
<td>0.19$^b$</td>
<td></td>
</tr>
<tr>
<td>DARC</td>
<td>5.03$^c$</td>
<td>2.22$^b$</td>
<td>0.53$^b$</td>
<td>1.55$^b$</td>
<td>2.23$^b$</td>
</tr>
<tr>
<td><strong>Atomization Energies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AE6</td>
<td>1.90$^c$</td>
<td>34.20$^a$</td>
<td>21.70$^a$</td>
<td>26.90$^a$</td>
<td>30.70$^a$</td>
</tr>
<tr>
<td>RC0</td>
<td>3.29$^c$</td>
<td>102.30$^a$</td>
<td>97.51$^a$</td>
<td>75.04$^a$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.96)$^c$</td>
<td>(0.35)$^c$</td>
<td>(0.40)$^c$</td>
<td>(0.55)$^c$</td>
<td></td>
</tr>
<tr>
<td><strong>Barrier Heights</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BH6</td>
<td>2.28$^d$</td>
<td>5.04$^c$</td>
<td>1.73$^b$</td>
<td>2.24$^b$</td>
<td>0.79$^b$</td>
</tr>
<tr>
<td>HAT6</td>
<td>2.86$^d$</td>
<td>11.57$^b$</td>
<td>2.17$^b$</td>
<td>2.99$^b$</td>
<td>0.20$^b$</td>
</tr>
<tr>
<td>NS6</td>
<td>3.63$^d$</td>
<td>2.72$^b$</td>
<td>1.56$^b$</td>
<td>2.93$^b$</td>
<td>1.75$^b$</td>
</tr>
<tr>
<td>Ur6</td>
<td>0.73$^c$</td>
<td>4.15$^c$</td>
<td>1.16$^b$</td>
<td>0.97$^b$</td>
<td>1.16$^b$</td>
</tr>
<tr>
<td>DBH24 total</td>
<td>2.38$^d$</td>
<td>5.87$^b$</td>
<td>1.66$^b$</td>
<td>2.11$^b$</td>
<td>1.42$^b$</td>
</tr>
<tr>
<td>NT-BH38</td>
<td>2.83$^d$</td>
<td>5.49$^b$</td>
<td>1.30$^b$</td>
<td>2.82$^b$</td>
<td>1.97$^b$</td>
</tr>
<tr>
<td>HT-BH38</td>
<td>2.20$^d$</td>
<td>5.20$^b$</td>
<td>1.45$^b$</td>
<td>1.96$^b$</td>
<td>0.80$^b$</td>
</tr>
<tr>
<td>BH76 total</td>
<td>2.51$^d$</td>
<td>3.35$^b$</td>
<td>1.37$^b$</td>
<td>2.12$^b$</td>
<td>1.38$^b$</td>
</tr>
<tr>
<td><strong>Intramolecular Interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HB6</td>
<td>0.49$^d$</td>
<td>0.19$^b$</td>
<td>0.21$^b$</td>
<td>0.22$^b$</td>
<td>0.09$^b$</td>
</tr>
<tr>
<td>CT7</td>
<td>1.09$^d$</td>
<td>0.88$^b$</td>
<td>0.86$^b$</td>
<td>2.37$^b$</td>
<td>0.13$^b$</td>
</tr>
<tr>
<td>D16</td>
<td>0.26$^d$</td>
<td>0.19$^b$</td>
<td>0.16$^b$</td>
<td>0.22$^b$</td>
<td>0.18$^b$</td>
</tr>
<tr>
<td>W7</td>
<td>0.05$^d$</td>
<td>0.04$^b$</td>
<td>0.06$^b$</td>
<td>0.07$^b$</td>
<td>0.06$^b$</td>
</tr>
<tr>
<td>PP5</td>
<td>0.10$^d$</td>
<td>0.59$^b$</td>
<td>0.51$^b$</td>
<td>0.33$^b$</td>
<td>0.11$^b$</td>
</tr>
<tr>
<td>NCCE31 total</td>
<td>0.42$^d$</td>
<td>0.38$^b$</td>
<td>0.36$^b$</td>
<td>0.69$^b$</td>
<td>0.11$^b$</td>
</tr>
<tr>
<td>H-bond (S22)</td>
<td>0.42$^d$</td>
<td>0.33$^b$</td>
<td>0.58$^b$</td>
<td>0.43$^b$</td>
<td>0.25$^b$</td>
</tr>
<tr>
<td>Dispersion (S22)</td>
<td>0.16$^d$</td>
<td>0.28$^b$</td>
<td>1.30$^b$</td>
<td>1.50$^b$</td>
<td>0.40$^b$</td>
</tr>
<tr>
<td>Mixed (S22)</td>
<td>0.23$^d$</td>
<td>0.61$^b$</td>
<td>1.01$^b$</td>
<td>1.01$^b$</td>
<td>0.50$^b$</td>
</tr>
<tr>
<td>S22 total</td>
<td>0.27$^d$</td>
<td>0.40$^b$</td>
<td>0.98$^b$</td>
<td>1.00$^b$</td>
<td>0.38$^b$</td>
</tr>
<tr>
<td><strong>Intermolecular Interactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACONF</td>
<td>0.07$^d$</td>
<td>0.19$^b$</td>
<td>0.12$^b$</td>
<td>0.23$^b$</td>
<td>0.07$^b$</td>
</tr>
<tr>
<td>CYCONF</td>
<td>0.12$^d$</td>
<td>0.36$^b$</td>
<td>0.68$^b$</td>
<td>0.61$^b$</td>
<td>0.42$^b$</td>
</tr>
<tr>
<td>PCONF</td>
<td>0.17$^d$</td>
<td>0.34$^b$</td>
<td>0.69$^b$</td>
<td>0.82$^b$</td>
<td>0.21$^b$</td>
</tr>
<tr>
<td>SCONF</td>
<td>0.23$^d$</td>
<td>0.89$^b$</td>
<td>0.32$^b$</td>
<td>0.37$^b$</td>
<td>0.38$^b$</td>
</tr>
</tbody>
</table>

$^c$CBS(4/3) extrapolated values. $^b$Using the ATZ basis set. $^a$Using the ATZ basis set and DFT-D3(BJ). $^d$Taken from GMTKN30 database, using the def2-QZVP basis set and DFT-D3. $^e$Using corrected atomic energies. $^f$Taken from ref. $^6$9 using NL correction ($b = 7.8$). $^g$The smallest MADs within 0.2 kcal mol$^{-1}$ are shown in bold.
The dRPA methods show typical serious underbinding errors, and the calculated atomization energies are very sensitive to the basis set. Even the A6Z basis set yields large but very systematic errors. The systematic underbinding and the basis set error can be corrected easily via corrected atomic energies. Contrarily, the imprecisions of the double hybrid methods (e.g., B2PLYP-D3) cannot be corrected as efficiently as the imprecisions of the dRPA methods.

For weak interactions in the NCCE31 test set, our results show that dRPA75 is much more accurate and precise for charge transfer interaction energies than the other dRPA methods and double hybrid functionals. dRPA75 also shows a very balanced performance for all categories of the test set, and it is considerably better than any of the methods discussed in this paper. For the S22 test set, dRPA75 performs quite well with slight overestimation of the interaction energies due to the basis set error. For intramolecular interactions in alkanes and peptides (ACONF, PCONF), dRPA75 shows excellent performance. It performs better than the other dRPA methods for cysteine and monosaccharide conformers (CYCONF, SCONF). These results are quite promising. dRPA75 performs well in many areas of molecular physical chemistry and integrates seamlessly the noncovalent interactions. It shows all the strengths of the dRPA methods but performs better than dRPA@PBE or dRPA@PBE0.25. However, problems arise due to the self-interaction error of PBEx and dRPa functional. The elimination of these errors will be the topic of our future work.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00420.

Further information about the parameter optimization, the complete basis set extrapolations, and detailed results with statistics (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: mezei.pal@mail.bme.hu.
*E-mail: gcsonka@mail.bme.hu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Department of Energy under Grant No. DE-SC0010499. The authors thank the Hungarian National higher education and research network for the computer time.

REFERENCES

Unified picture for the conformation and stabilization of the \(O\)-glycosidic linkage in glycopeptide model structures

Pál D. Mezei \(^1\) · Gábor I. Csonka \(^1\)

Received: 10 August 2015 / Accepted: 14 August 2015 / Published online: 4 September 2015 © Springer Science+Business Media New York 2015

Abstract Glycoproteins play a central role in the immune response. In this study, we focus on the core of the common \(O\)-linked mucin-type glycopeptides. It has been observed that glycosylation stabilizes the protein in a stiffened, extended structure. We provide a unified picture for the conformation and stabilization of the \(O\)-glycosidic linkage using \(O\)-(2-acetylamino-2-deoxy-\(\alpha\)- or \(-\beta\)-D-galacto- or mannopyranosyl)-N-acetyl-L-serinamide model structures. We have calculated equilibrium geometries of the model structures with the B3LYP/6-31G(2\(df\),\(p\)) method suggested in the Gaussian-4 theory. According to the relative energies, we confirm that the GalNAc-Ser linkage is more stable than its mannose analogues. The natural preference for the \(\alpha\)-GalNAc-Ser over the \(\beta\)-GalNAc-Ser anomers can be explained by entropic effects. We explored the hydrogen bonding patterns on the carbohydrate unit calculating highly accurate dRPA@PBE0.25 and dRPA75 energies and found that in some cases, the acetamido group can be fixed by hydrogen bonding from the \((O3_{\text{carb}})H\) atom, but in most of the cases, it can rotate more freely. The torsion angles in the glycosidic linkage show that the linkage is stiffened more in the \(\alpha\)-anomers and the most in the \(\alpha\)-GalNAc-Ser structure because of the steric strains in the axial position and by two or three intramolecular hydrogen bonds. We also found that although, in our gas-phase model geometries, the peptide backbone prefers to be in a \(\gamma\_i\)-turn, a structural water molecule can stabilize a polyproline II helix of a proline-rich sequence, a \(\beta\)-sheet, or more likely random coils.

Keywords Glycoproteins · Glycopeptides · Glycosidic linkage · Density functional calculations

Introduction

Glycosylation [1, 2] is among the most frequent post-translational modifications [3]. In biology, glycans play structural or modulatory roles (i.e., protein folding, stability [4]) or participate in intrinsic or extrinsic recognition [5]. The additional information conveyed by the oligosaccharide structure represents the glycocode [6, 7]. (The 3D structure [8–10] and biological role [11, 12] of glycoproteins and the synthetic approaches [13–18] are reviewed in the given references.) The two most common glycosylation types are the \(N\)-linked [19, 20] and the \(O\)-linked glycosylations [21, 22]. We focus here on the core of \(O\)-linked mucin-type glycans, in which a 2-acetylamino-2-deoxy-\(\alpha\)-galactose (GalNAc) is \(\alpha\)-glycosidically linked to a serine or a threonine residue [23, 24].

As almost all the key molecules in the immune response are glycoproteins [25], mucin-related \(O\)-linked glycopeptides have been highlighted in cancer treatment because immunization against cancer cells might be obtained with carbohydrate vaccines [26], which mimic posttranslational modification [27, 28]. The disease states also can be monitored by the in vivo profiling of the changes in the...
O-linked glycosylation [29]. Furthermore, glycopeptides are generally applied as antibiotics [30–34].

According to hydrodynamic studies, mucins have random coil structure [35]. NMR and CD spectroscopic studies also indicate that mucins relatively lack α- and β-secondary structures [36]. This is because glycosylation stiffens the peptide chains and thus leads to an extended structure as it was reported by transmission electron microscopy on leukosialin (CD43) [37], by light-scattering [36] and 13C NMR spectroscopies [38] on ovine submaxillary mucin (OSM) and by rotary shadowing electron microscopy on P-selectin glycoprotein ligand-1 (PSGL-1) [39]. The chain-stiffening effect is caused by the steric interaction between the GalNAc residue and the peptide backbone. The rigidity is increased by the second carbohydrate unit, but the further carbohydrate units have very little effect [40, 41].

An NMR, CD, and molecular modeling analysis of human salivary mucin (MUC7)-derived O-linked model glycopeptides revealed that an intramolecular hydrogen bond between the amide proton of GalNAc and the carbonyl oxygen of threonine stabilize the structure [42]. However, the serine analogues lacked such intramolecular hydrogen bonding. Note that in this particular case, the proline-rich core was found to be in polyproline II (PPII) helix.

The NMR analysis of the mucin glycopeptide motif derived from the N-terminal fragment SSTTAV of the cell surface glycoprotein CD43 also provided an elongated structure and the interaction of the peptide and the first surface glycoprotein CD43 also provided an elongated derived from the N-terminal fragment SSTTAV of the cell helix. The proline-rich core was found to be in polyproline II (PPII) secondary structures [36]. This is because glycosylation stiffens the peptide chains and thus leads to an extended structure as it was reported by transmission electron microscopy on leukosialin (CD43) [37], by light-scattering [36] and 13C NMR spectroscopies [38] on ovine submaxillary mucin (OSM) and by rotary shadowing electron microscopy on P-selectin glycoprotein ligand-1 (PSGL-1) [39]. The chain-stiffening effect is caused by the steric interaction between the GalNAc residue and the peptide backbone. The rigidity is increased by the second carbohydrate unit, but the further carbohydrate units have very little effect [40, 41].

An NMR, CD, and molecular modeling analysis of human salivary mucin (MUC7)-derived O-linked model glycopeptides revealed that an intramolecular hydrogen bond between the amide proton of GalNAc and the carbonyl oxygen of threonine stabilize the structure [42]. However, the serine analogues lacked such intramolecular hydrogen bonding. Note that in this particular case, the proline-rich core was found to be in polyproline II (PPII) helix.

The NMR analysis of the mucin glycopeptide motif derived from the N-terminal fragment SSTTAV of the cell surface glycoprotein CD43 also provided an elongated structure and the interaction of the peptide and the first N-acetylgalactosamine residue [43, 44]. The measured 3 JHN-H2 coupling constants in the internal T2 and T3 residues limited the associated ϕ torsion angles to the region around −120°. The NOE and coupling constant restrained MD structural calculations resulted in β-sheet structure, and the amide proton chemical shift trends independently confirmed the β-character.

The glycosidic linkage was studied theoretically in vacuum using MM2* and HF/6-31G(d) optimized O-(2-acetylamino-2-deoxy-α- or -β-D-galacto- or -mannopyranosyl)-N-acetyl-l-serinamide model structures (Fig. 1) [45]. The α- and β-GalNAc-Ser anomers were found to be the most stable due to the hydrogen bond between the acetamido group and the peptide backbone. The acetamido group and the peptide backbone were also constrained by two intraresidual hydrogen bonds. The entropy terms stabilize the α-anomers relatively to the β-anomers, and the mannose derivatives relatively to the galactose derivatives. According to the relative Gibbs energies, an α-GalNAc-Ser conformer was found to be the most stable geometry, which explains the natural preference for the α-GalNAc-Ser linkage. The calculated geometries agreed well with the experimentally derived χ1 and χ2 torsion angles of the linkage. The preferred secondary structure of these gas-phase models was reported to be γL-turn.

The solvent interactions were considered by molecular dynamics simulations with time-averaged restraints (MD-tar) from NMR measurements and by B3LYP/6-31G(d) optimized geometries with water pockets and bridges [46]. The β-sheet and the PPII helix were found to be more likely than the α-helical structure. The comparison of serine and threonine derivatives using MD-tar simulations showed antiperiplanar/antiperiplanar (ap/ap) or synclinal/antiperiplanar (sc/ap) and antiperiplanar/anticlinal (ap/+ac) or synclinal/+anticlinal (−sc/+ac) conformations for the χ3/χ2 torsions of the α- or β-GalNAc-Ser and α- or β-GalNAc-Thr structures, respectively. Furthermore, the +synclinal (+sc) conformation was found to be the most probable for the χ1 torsion, more probable in α-GalNAc derivatives and in the threonine derivatives, and the only possibility in α-GalNAc-Thr [47]. The MD-tar studies also showed that β-O-Glycosylation on L-Ser or L-Thr increases the probability for the occurrence of this motif in α-helical structure [48]. Similar observations were made for α- or β-GlcNAc-Ser or -Thr from MD-tar simulations [49]. (The β-GlcNAc modification is also significant because it aberrantly alters the behavior of proteins, and it is related to diabetes, tumorigenesis, and Alzheimer’s disease. Furthermore, the β-Glc-Ser glycoform of
brain-derived peptides, in contrast to the unglycosylated form, can penetrate the blood–brain barrier, which can lead to the therapeutic use of these peptides [50]).

The hydration process was also studied in gas phase comparing the measured infrared ion depletion spectra to the calculated IR spectra on B3LYP/6-31++G* optimized mono-, bi-, and trihydrated 1-phenyl-2-acetylamino-2-deoxy-β-D-glucose complexes. In the two most stable monohydrated complexes, the acetamido group was either fixed by the (O3 Carb)H···O7 Carb hydrogen bond and hydrated from the (N2 Carb)H moiety or rotated and bridged by a water molecule between the (O3 Carb)H and O7 Carb atoms [51].

In another study in the absence of NMR restraints using CHARMM carbohydrate force field with Hamiltonian replica exchange (HREX) sampling, the extended β-sheet or the PPII helix was found to be more favorable than the compact α-helix. (Note that by NMR, it is difficult to distinguish between PPII helix and the random coils.) According to these data, it was concluded that there is a synergistic interplay between the intramolecular H bonds and the water bridges, which determines the stability of the O-linkage [52].

It was also shown that methylation or other modification in the α-position can cause the break up of the (N2 Pept)H···O1 Pept hydrogen bond and force the peptide to adopt α-helix conformation [53]. (Although it does not affect much the stability of a short glycopeptide receptor complex, it is believed that it may modulate the binding properties of larger glycopeptides to appropriate receptors [54]).

Computational details

We reoptimized the former HF/6-31G(d)-optimized α- and β-Gal- or -ManNAc-Ser model structures [45] and calculated eight new gas-phase geometries. The geometry optimizations and frequency calculations were performed with B3LYP/6-31G(2df,p) method (suggested in the Gaussian-4 theory) [55] with RIJCOSX approximation (with automatically generated auxiliary basis set) using Orca 3.0.2 quantum chemistry software. The effect of hydration was estimated after the optimization on the equilibrium geometries using the same method with COSMO solvation model with the relative permittivity of 78.3553 and refractive index of 1.333.

We used the dRPA@PBE0 (25 % exact exchange, 75 % PBE exchange, and 100 % PBE correlation for the self-consistent Kohn–Sham orbital calculations; 100 % exact exchange and 100 % dRPA correlation for the non-self-consistent final energy evaluation [56–59]) and dRPA75 (75 % exact exchange, 25 % PBE exchange, and 100 % PBE correlation for self-consistent the Kohn–Sham orbital calculations; 75 % exact exchange, 25 % PBE exchange, and 100 % dRPA correlation for the non-self-consistent final energy evaluation) methods for the energy calculations because these methods can describe the energetics of hydrogen bonds with high accuracy [60]. The dRPA@PBE0 and dRPA75 calculations were performed with (aug)-cc-pVTZ(-f,-d) basis set (aug-cc-pVTZ basis set with diffuse functions only on the non-hydrogen atoms, without f functions on the non-hydrogen atoms, and without d functions on the hydrogen atoms; this basis set approximates well the aug-cc-pVTZ energies by density functional calculations [61]), density fitting with cc-pVTZ-RJ-JK for SCF calculations, cc-pVTZ-RI for correlation calculations, and natural auxiliary functions (with very tight threshold: naf_cor = 1e−3) using MRCC quantum chemistry software (released on 05/28/2015).

Results and discussion

Relative energies

The relative energies of the most stable α- and β-Gal- or -ManNAc-Ser model structures calculated with the B3LYP/6-31G(2df,p) method (Fig. 2) are similar to the former MM2* energies [45]. This method gives smaller energy difference for the most stable α-Gal/β-Gal derivatives than the HF/6-31G(d) method. However, the relative energy ranges of the B3LYP/6-31G(2df,p) rotamers are on the same scale than that of the HF/6-31G(d) rotamers. The two most stable α-Gal rotamers are the α-Gal02 and α-Gal06 as they were also the most stable rotamers according to the previous HF results (see also the numbering of the conformers in Ref. [45]). The other most stable rotamers are the α-Man03 > α-Man14 > α-Man01, the β-Gal03 > β-Gal01 > β-Gal17 > β-Gal08 and β-Man20.

Hydrogen bonds

In the calculated vacuum geometries in the absence of water, intramolecular hydrogen bonds are formed (Table 1). The (N2 Ser)H···O1 Ser hydrogen bond fixes the peptide backbone. The (O3 Carb)H···O7 Carb and the (O4 Carb)H···O3 Carb hydrogen bonds fix the position of the acetamido group. These three hydrogen bonds are present in almost all the conformers. The (O6 Carb)H atom forms a hydrogen bond with either the O4 Carb or the O5 Carb atom. In the latter case, the (O4 Carb)H atom can form a hydrogen bond with the O6 Carb atom with low probability.

The (N2 Carb)H···O1 Carb hydrogen bond is highly possible in α-Gal and β-Gal, moderately possible in β-Man, and never occurs in α-Man. The (N1 Ser)H···O1 Carb hydrogen bond is frequent in α-Gal and β-Man and less frequent in β-Gal and α-Man. The (N2 Carb)H atom can form a hydrogen bond with
the O₂Ser atom in α-Gal and β-Gal, but it is not common or present in β-Man or α-Man. The (N₂Carb)H···O₁Carb, (N₁Ser)H···O₁Carb, and (N₂Carb)H···O₂Ser hydrogen bonds can freeze the sugar–peptide linkage, but according to the B3LYP/6-31G(2df,p) calculations, the latter hydrogen bond is less probable than previously was considered.

This finding agrees with the experimental results that the H₂Carb-O₂Ser distance fluctuates between ~3.2 and ~4.1 Å [44]. We calculated this distance to be 3.189 and 3.266 Å in the two most stable α-Gal conformers without water molecule, and it was calculated to be 4.144 Å in the presence of one water molecule (on the level of B3LYP/6-31G(d)) accepting two hydrogen bonds from (N₂Carb)H and (N₁Ser)H and donating one to O₂Ser [46].

Torsion angles

The ap conformation along the C₂Carb–N₂Carb bond gives the opportunity for the (N₂Carb)H···O₁Carb hydrogen bond to appear in α-Gal, β-Gal, and β-Man, but this bonding is impossible in α-Man because both of the neighboring groups in positions 1 and 2 are in axial position. Inspection of the torsion angles τ₂ and τ₃ (Tables S1a-d) shows that the acetamido group is fixed by the (O₃Carb)H···N₂Carb hydrogen bond in almost all the conformers (and also by the (N₂Carb)H···O₂Ser hydrogen bond in half of the cases). Otherwise, the acetamido group can rotate freely with the (N₂Carb)H atom forming hydrogen bond with the O₂Ser, O₁Ser, or O₃Carb atom in α-Gal, β-Gal, and β-Man, or with the O₅Carb atom in α-Man and β-Man. Furthermore, with the O₇Carb atom, it can accept a hydrogen bond from the (N₁Ser)H atom. The (O₃Carb)H group can also point toward

The newly identified hydrogen bonds are in italic.

<table>
<thead>
<tr>
<th>Hydrogen bond</th>
<th>α-Gal (%)</th>
<th>α-Man (%)</th>
<th>β-Gal (%)</th>
<th>β-Man (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O₃Carb)H···O₄Carb</td>
<td>10.0</td>
<td>0.0</td>
<td>10.0</td>
<td>16.7</td>
</tr>
<tr>
<td>(O₅Carb)H···O₇Carb</td>
<td>90.0</td>
<td>100.0</td>
<td>90.0</td>
<td>72.2</td>
</tr>
<tr>
<td>(O₄Carb)H···O₃Carb</td>
<td>80.0</td>
<td>100.0</td>
<td>90.0</td>
<td>83.5</td>
</tr>
<tr>
<td>(O₄Carb)H···O₆Carb</td>
<td>20.0</td>
<td>0.0</td>
<td>10.0</td>
<td>5.6</td>
</tr>
<tr>
<td>(O₆Carb)H···O₄Carb</td>
<td>40.0</td>
<td>17.6</td>
<td>35.0</td>
<td>11.1</td>
</tr>
<tr>
<td>(O₆Carb)H···O₅Carb</td>
<td>30.0</td>
<td>23.5</td>
<td>35.0</td>
<td>50.0</td>
</tr>
<tr>
<td>(N₂Carb)H···O₁Carb</td>
<td>90.0</td>
<td>0.0</td>
<td>85.0</td>
<td>50.0</td>
</tr>
<tr>
<td>(N₂Carb)H···O₃Carb</td>
<td>10.0</td>
<td>0.0</td>
<td>5.0</td>
<td>5.6</td>
</tr>
<tr>
<td>(N₂Carb)H···O₅Carb</td>
<td>0.0</td>
<td>41.2</td>
<td>0.0</td>
<td>5.6</td>
</tr>
<tr>
<td>(N₁Ser)H···O₅Carb</td>
<td>60.0</td>
<td>29.4</td>
<td>45.0</td>
<td>61.1</td>
</tr>
<tr>
<td>(N₁Ser)H···O₅Carb</td>
<td>0.0</td>
<td>5.9</td>
<td>15.0</td>
<td>16.7</td>
</tr>
<tr>
<td>(N₁Ser)H···O₆Carb</td>
<td>0.0</td>
<td>29.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(N₁Ser)H···O₇Carb</td>
<td>10.0</td>
<td>0.0</td>
<td>10.0</td>
<td>27.8</td>
</tr>
<tr>
<td>(N₂Ser)H···O₆Carb</td>
<td>0.0</td>
<td>5.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>(N₂Ser)H···O₇Carb</td>
<td>10.0</td>
<td>0.0</td>
<td>10.0</td>
<td>11.1</td>
</tr>
<tr>
<td>(N₂Carb)H···O₁Ser</td>
<td>50.0</td>
<td>0.0</td>
<td>45.0</td>
<td>16.7</td>
</tr>
<tr>
<td>(O₆Carb)H···O₁Ser</td>
<td>0.0</td>
<td>11.8</td>
<td>10.0</td>
<td>5.6</td>
</tr>
<tr>
<td>(O₆Carb)H···O₂Ser</td>
<td>0.0</td>
<td>11.8</td>
<td>15.0</td>
<td>33.3</td>
</tr>
<tr>
<td>(N₂Ser)H···O₁Ser</td>
<td>90.0</td>
<td>88.2</td>
<td>90.0</td>
<td>94.4</td>
</tr>
</tbody>
</table>

The newly identified hydrogen bonds are in italic.
the O4 Carb atom forming a hydrogen bond, or it can be free depending on the existence of the (O4 Carb)H···O3 Carb hydrogen bond, which can be followed by the torsion angle $\tau_4$. The (O4 Carb)H atom is in this hydrogen bond in almost all the cases, but it can rarely form a hydrogen bond with the O6 Carb or the O5 Carb atom, or it can be free. If the (O4 Carb)H atom binds to the O3 Carb atom, the ranges of torsion angles $\tau_5$ and $\tau_6$ show for the (O6 Carb)H atom three possible bonding situations with the O4 Carb and O5 Carb atoms (Fig. 3). If the (O4 Carb)H atom points toward the O5 Carb atom, it reduces to two possibilities for the (O6 Carb)H atom to bind to the O4 Carb and O5 Carb atoms, and if the (O4 Carb)H atom binds to the O6 Carb atom, there is only one possibility for the (O6 Carb)H atom to bind to the O5 Carb atom. The (O6 Carb)H atom can also form hydrogen bonds with the O1 Ser and O2 Ser atoms in some cases of the $\alpha$-Man, $\beta$-Gal, and $\beta$-Man conformers, and also it can be free.

The torsion angle $\chi_3$ shows that in all the $\alpha$-conformers, the peptide linkage is arranged in ap conformation because of the steric strain in the axial position (Fig. 4). In the $\beta$-conformers, the ap conformation is the most frequent as well, but the (N2 Carb)H···O2 Ser interaction in $\beta$-Gal can result in $\psi$-synclinal ($\psi$-sc) conformation, and the (N1 Ser) H···O7 Carb interaction in $\beta$-Man can stabilize $\psi$-anticlinal ($\psi$-ac) conformation.

The torsion angle $\chi_2$ shows anti-conformation in almost all the $\alpha$-Gal conformers but varies more in the other conformers (mostly ap and $\pm$ac but also $\pm$sc in some cases). It suggests the structure along the C$\beta$–O1 Carb bond is more rigid in the $\alpha$-Gal conformers than that in the other conformers. In $\beta$-Man, since the neighboring acetamido and peptide groups are in the axial–axial arrangement, the peptide chain can rotate freely under the plain of the sugar molecule along the torsion angles $\chi_1$ and $\chi_2$ forming the (N1 Ser)H···O6 Carb or (N2 Ser)H···O6 Carb hydrogen bonds.

Fig. 3 Possible hydrogen bonds on the carbohydrate moiety of the $\alpha$-GalNAc model with freezing orientation of the acetamido group by the (O3 Carb)H···O7 Carb hydrogen bond (for clarity, we omitted the (C)H atoms and closed the structure with a carbon atom at the glycosidic position. There are two–two possible arrangements for the hydrogen bond between the (O6)H and O5 atoms as well as for the hydrogen bond between the (O6)H and O4 atoms as the hydroxyl-methyl group can be above the ring plane (without apostrophe sign) or in the ring plane (with apostrophe sign).)

Fig. 4 Side-chain rotamers of the $\alpha$- and $\beta$-Gal- or -ManNAc-Ser model structures in the absence of water and the monohydrated $\alpha$-GalNAc-Ser model structures optimized with B3LYP/6-31G(2df,p) method (The angle ranges of the anti-/antiperiplanar conformations for the $\chi_2/\chi_3$ torsions are denoted in the figure).

The torsion angle $\chi_1$ shows that the $\psi$-sc conformation dominates in the conformers, but the $\psi$-ap and $\psi$-conformations are also present (Fig. 5). If the conformations are antiperiplanar along the C1 Carb–O1 Carb and O1 Carb–C$\beta$ bonds as it is in most of the $\alpha$-Gal conformers, then the $\psi$ and $\tau$ torsion angles determine the position of the O2 Ser atom, which is also related to the existence of the (N2 Carb)H···O2 Ser hydrogen bond beside the torsion angle $\tau_2$. In one case, the (O6 Carb)H···O1 Ser hydrogen bond can fix the peptide chain in $\psi$-ac conformation. In another case, the (N2 Carb)H···O1 Ser hydrogen bond can fix the peptide chain in synperiplanar (sp) conformation. If the (N1 Ser) H···O1 Carb hydrogen bond exists, then the torsion angle $\phi$ is between $-75^\circ$ and $-115^\circ$, and the torsion angle $\chi_1$ is between $35^\circ$ and $65^\circ$. 


\[123\]
The torsion angle \( \phi \) shows that most of the conformations are +sc or −sc (Fig. 6). In few cases, there can be also −ac conformation because the (N1_Ser)H···O5_Carb or (O6_Carb)H···O1_Ser hydrogen bonds can fix the peptide chain. The torsion angle \( \psi \) shows that most of the conformations are +sc or −sc. In some cases, the (O6_Carb)H···O2_Ser hydrogen bond can distort the +sc and −sc conformations to +ac or sp. The torsion angles \( \phi \) and \( \psi \) indicate together the existence of (N2_Ser)H···O1_Ser hydrogen bond. In other words, the gauche− arrangement along the N1_ser−C\( \alpha \) bond with gauche+ arrangement along the C\( \alpha \)−C1_ser bond (g−g+) or the opposite g+g− conformation gives the possible ranges for the (N2_ser)H···O1_ser hydrogen bond. Note that the g−g+ conformation (\( \gamma_L \)) is possible by L amino acids in contrast to the g+g− conformation.

**Electronic energy**

In order to determine the starting point of the hydration process, we identified the most stable \( \alpha \)-Gal conformer. The electronic energies for the different hydrogen bonding patterns on the \( \alpha \)-Gal unit were calculated with dRPA75 and dRPA@PBE0 methods (Table 2). The relative energies calculated with these two methods agree within 0.5 kcal mol\(^{-1}\). When the (O3_Carb)H atom binds to the O7_Carb atom, and the (O4_Carb)H atom binds to the O3_Carb atom, the (O6_Carb)H atom can bind the O4_Carb atom in two ways and the O5_Carb atom in one way. The most favorable is the (O6_Carb)H···O5_Carb' arrangement, but the (O6_Carb)H···O4_Carb' arrangement is also favorable. When the (O3_Carb)H atom still binds to the O7_Carb atom, but the (O4_Carb)H atom binds either the O5_Carb or O6_Carb atom, the (O6_Carb)H atom can bind the O4_Carb atom in one way and the O5_Carb atom in two ways. In the two most favorable situations, the (O6_Carb)H atom is binding to the O5_Carb atom. The (O6_Carb)H···O5_Carb is slightly favorable (0.5 kcal mol\(^{-1}\)) than the (O6_Carb)H···O5_Carb' arrangement. This difference becomes more significant (2.0–2.1 kcal mol\(^{-1}\)) as the (O3_Carb)H atom releases the O7_Carb atom and turns toward the O4_Carb atom.

**Table 2** Relative energies (in kcal mol\(^{-1}\)) of the \( \alpha \)-GalACSer rotamers according to the hydrogen bonding pattern on the galactose part calculated with dRPA@PBE0 and dRPA75 methods and (aug)-cc-pVTZ(-f,-d) basis set on B3LYP/6-31G(2df,p) reoptimized geometries.

<table>
<thead>
<tr>
<th>( \Delta E ) dRPA@PBE0</th>
<th>( \Delta E ) dRPA75</th>
<th>(O3_Carb)H</th>
<th>(O4_Carb)H</th>
<th>(O6_Carb)H</th>
</tr>
</thead>
<tbody>
<tr>
<td>06</td>
<td>1.0</td>
<td>1.1</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
</tr>
<tr>
<td>n2</td>
<td>4.6</td>
<td>4.4</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
</tr>
<tr>
<td>02</td>
<td>0.0</td>
<td>0.0</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
</tr>
<tr>
<td>n3</td>
<td>6.9</td>
<td>7.0</td>
<td>O7_Carb</td>
<td>O5_Carb</td>
</tr>
<tr>
<td>n1</td>
<td>4.7</td>
<td>4.8</td>
<td>O7_Carb</td>
<td>O5_Carb</td>
</tr>
<tr>
<td>n6</td>
<td>4.2</td>
<td>4.3</td>
<td>O7_Carb</td>
<td>O6_Carb</td>
</tr>
<tr>
<td>n7</td>
<td>5.6</td>
<td>5.9</td>
<td>O4_Carb</td>
<td>O5_Carb</td>
</tr>
<tr>
<td>n4</td>
<td>4.6</td>
<td>4.8</td>
<td>O4_Carb</td>
<td>O5_Carb</td>
</tr>
<tr>
<td>09</td>
<td>2.6</td>
<td>2.7</td>
<td>O4_Carb</td>
<td>O6_Carb</td>
</tr>
</tbody>
</table>

The new geometries are denoted by n1–8. The hydrogen bonding arrangements of the (O6_Carb)H atom are denoted. The apostrophe signals an arrangement in the ring plane; all other arrangements are above the ring plane.
Note that in most of the O-glycosidic core structures [45], the (O3_Carb)H atom is substituted by the second carbohydrate unit. Therefore, in these structures, the (O3_Carb)H···O7_Carb hydrogen bonding, which could fix the acetamido group, is not possible. However, the hydrogen bonds on the carbohydrate moiety might be important for fixing the orientation of the second carbohydrate unit.

**Structural water**

Adding a water molecule to the most stable α-GalNAc-Ser vacuum geometry needs the peptide chain to distance from the acetamido group (Fig. 7). In the first step, the (N2_Ser)H···O1_Ser hydrogen bond breaks during the torsion angles $\chi_1$, $\chi_2$, and $\chi_3$ (Table S2) are changing by $-12^\circ$, $36^\circ$, and $-20^\circ$, respectively. Then, the (N2_Ser)H···O1_Ser hydrogen bond breaks letting the peptide chain with the O2_Ser atom to rotate along the C$_\alpha$–C2_Ser bond by $48^\circ$. This allows the relaxation of torsion angle $\chi_1$. In the final step, the (O3_Carb)H and (O4_Carb)H groups rotate toward the O4_Carb and O5_Carb atoms forming new hydrogen bonds instead of the (O3_Carb)H···O2_Carb and (O4_Carb)H···O3_Carb hydrogen bonds. Thus, the acetamido group rotates along the C2_Carb–N2_Carb bond by $-52^\circ$ and the peptide chain with the O2_Ser atom rotates along the C$_\alpha$–C2_Ser bond by $87^\circ$ guiding the water molecule toward the (N1_Ser)H atom. When the water molecule approximates this hydrogen atom, they form a hydrogen bond. Therefore, the (N1_Ser)H···O1_Carb hydrogen bond breaks, and the peptide chain rotates along the N1_Ser–C$_\alpha$ bond toward the water molecule. At the same time, the torsion angle $\chi_1$ increases by $20^\circ$, and the torsion angles $\chi_2$ and $\chi_3$ relax. During the three consecutive steps, the distance between the N2_Carb and O2_Ser atoms changes from 3.189 to 4.879, 4.637, and 3.816 Å, respectively.

The torsion angles $\chi_1$, $\chi_2$, and $\chi_3$ are also presented in Figs. 4, 5. These torsion angles vary only in the angle range of the initial conformation. The changes in the torsion angles $\phi$ and $\psi$ can be followed in Fig. 6. If the (N2_Carb)···O1_Carb hydrogen bond exist, the peptide backbone prefers to be in a $\gamma_1$-turn. However, a well-defined structural water between the acetamido group and the peptide backbone (and probably also a water bridge between the (O3_Carb)H and O7_Carb atoms) [51] can stabilize a PPII helix or a β-sheet. Note that the PPII helix requires a proline-rich sequence; otherwise, it can be thought only as a protein folding intermediate. Also note that random coils can be in similar torsion angle ranges to PPII helices or β-sheets.

Starting from the most stable conformer and holding the (O6_Carb)H atom in the (O6_Carb)H···O5_Carb' arrangement, the electronic energy largely deepens ($\sim 9$ kcal mol$^{-1}$) with the addition of a single water molecule to the structure (Table 3). Breaking the (N2_Ser)H···O1_Ser hydrogen bond makes the electronic energy more positive ($\sim 4$ kcal mol$^{-1}$). Breaking the (N1_Ser)H···O1_Carb bond and forming the (N1_Ser)H···O_Wat bond accompanied by the rotation of the acetamido group and the relaxation of the peptide chain yield only small positive change (0.5–0.6 kcal mol$^{-1}$) in the electronic energy. Note that releasing the (O6_Carb)H atom from the (O6_Carb)H···O5_Carb' arrangement and letting to be in the most stable pattern related to the (O6_Carb)H···O5_Carb arrangement can result in a $\sim 2$ kcal mol$^{-1}$ lowering in the electronic energy.

The corrections to the Gibbs energy in standard circumstances mean about $+11$ kcal mol$^{-1}$ change mostly because of the losses in the entropy of the water molecule. After that, the correction does not change so much.

The COSMO solvation model predicts that building in the structural water molecule leads to more and more hydrophilic conformation. The corrections to the energy
The hydrogen bonding arrangements of the (O_{6Carb})H atom are denoted. The lack of the apostrophe sign belongs to an arrangement above the ring plane; the presence of the apostrophe sign belongs to an arrangement in the ring plane.

are in order $-19.3$, $-20.8$, $-23.9$, and $-26.6$ kcal mol$^{-1}$. (For comparison, the correction to the energy of a single water molecule is $-6.7$ kcal mol$^{-1}$.)

**Conclusion**

Comparison of the calculated energies of $\alpha$- and $\beta$-Gal- or -ManNAc-Ser model structures yields that the $\alpha$-GalNAc-Ser linkage is significantly more stable than its mannose analogue. This can be explained by the hydrogen bonds between the acetalamido group and the peptide backbone. The energy of the $\beta$-GalNAc-Ser linkage is relatively close to that of the $\alpha$-GalNAc-Ser linkage. The preference for the latter can be explained rather by the more negative entropic term in the Gibbs energy of the $\alpha$-anomers [45].

The relative frequencies of the hydrogen bonds show that the acetalamido group can be fixed by an intraresidual hydrogen bonding pattern in all the carbohydrate model structures. However, there exists an energetically almost equally possible intraresidual hydrogen bonding pattern, which allows the acetalamido group to rotate. Furthermore, in most of the $O$-glycosidic core structures, the hydrogen atom of the C3 hydroxyl group is substituted by the second carbohydrate unit; therefore, the acetalamido group cannot accept any hydrogen bond from this side.

The glycopeptide linkage can be stiffened by three hydrogen bonds. The glycosidic oxygen can accept one hydrogen bond from the acetalamido group and the peptide backbone, and there can be a hydrogen bond between the acetalamido group and the peptide backbone. These hydrogen bonds are more frequent in the $\alpha$-GalNAc-Ser linkage, less frequent in the $\beta$-linkages, and very rare in the $\beta$-GalNAc-Ser linkage. Our data also show that the hydrogen bond between the acetalamido group and the peptide backbone is less possible than it was previously considered. The vacuum geometries also show an intraresidual hydrogen bond in the peptide backbone, but this can be broken in aqueous phase.

The torsion angles of the side-chain rotamers show that the glycosidic linkage is frozen in $\alpha$-anomers and especially in the $\alpha$-GalNAc-Ser structure. The axial arrangement of the glycosidic bond affects the torsion along the $C_{1Carb}$-$O_{1Carb}$ bond. The equatorial acetalamido group in the $\alpha$-GalNAc-Ser affects more the torsions along the $O_{1Carb}$-$C_{\beta}$ and the $C_{\rho}$-$C_{\alpha}$ bonds. In gas phase, the peptide backbone prefers to be in a $\gamma_{1}$-turn. However, in aqueous phase, a structural water between the acetalamido group and the peptide backbone can stabilize a polyproline II helix of a proline-rich sequence or more generally a $\beta$-sheet or random coils. But the random coils agree more with the experiments and can be explained by the glycosylation caused rigidity, which prevents the peptide chains to fold.

Finally, the identified hydrogen bonding patterns on the carbohydrate moiety suggest that intramolecular hydrogen bonding or water bridges can fix the linkage between the first and second carbohydrate. Future works can be about to analyze the hydrogen bonding pattern of this linkage and expand our picture from the glycopeptide linkage to the glycosidic core.

**Acknowledgments** This project is supported by the Grants TÁMOP-4.2.1/B-09/1/KMR-2010-0002 and TÁMOP-4.2.2.B-10/1–2010-0009. The authors thank the Hungarian National Higher Education and Research Network for the computer time.

**References**

Why Density Functionals Should Not Be Judged Primarily by Atomization Energies

John P. Perdew¹²*, Jianwei Sun¹, Adrienn Ruzsinszky¹, Pál D. Mezei³, Gábor I. Csonka³**

Received 25 June 2015; accepted after revision 16 July 2015

Abstract

While most molecules and solids are spin-unpolarized, most chemically-active atoms are partly spin-polarized. As a result, the errors of the spin-dependence of a density functional are much more troublesome for atomization energies than they are for typical reaction or formation energies. This observation explains why the atomization energy errors of approximate functionals do not correlate with their other errors, and why the errors of atomization energies for a given functional can be radically reduced by fitting the energies of the atoms. We present an illustrative example from the recent nonempirical construction of the SCAN meta-generalized gradient approximation.

Keywords

density functional, atomization energy, electronic structure theory

1 Introduction

This year is the 50th anniversary of Kohn-Sham density functional theory [1], which from a humble beginning has grown up to be the most widely-used method of electronic structure calculation in condensed matter physics, quantum chemistry, and materials engineering. The popularity of this theory arises first because it is computationally efficient, with an orbital structure like that of Hartree or Hartree-Fock theory, and second because it is usefully accurate. There is an underlying exact theory for the ground-state energy and electron density, which inspires the continuing and successful (if slow) search for more accurate but still computationally tractable approximations to the needed density functional for the exchange-correlation energy.

This is also the tenth anniversary of collaboration among the authors of this paper. In 2005, three of us (GIC, AR, JPP) published the first [2] of 35 collaborative research papers in density functional theory. (One year later, another of us (JS) started research in this theory as a graduate student with JPP.) That first collaborative paper raised questions about the atomization energies of molecules, which up to that time had been widely regarded as a “gold standard” of accuracy for density functional approximations. In 2015, when three of us (JS, AR, and JPP) have proposed what we regard as an optimally accurate functional that preserves computational efficiency [3], we return to the issue in the title of this article and present an explanation which is supported by numerical evidence at this high level of approximation.

2 Density and Spin-Density Functional Theories

The original Kohn-Sham theory [1] assumed that the electrons (with mutual Coulomb repulsion) see a static, spin-independent, and multiplicative external potential \( v(\vec{r}) \) (typically but not necessarily the Coulomb attraction to the nuclei). Under this condition, the exact exchange-correlation energy \( E_{xc} \) is a functional \( E_{xc}[n] \) of the electron density distribution \( n(\vec{r}) \); i.e., there is a rule (in practice an uncomputable one) that assigns one energy \( E_{xc} \) to each possible electron density distribution \( n(\vec{r}) \). Apart from \( E_{xc} \), the rest of the total energy is computable exactly.

¹ Department of Physics, College of Science and Technology, Temple University, Philadelphia, PA 19122 USA
² Department of Chemistry, College of Science and Technology, Temple University, Philadelphia, PA 19122 USA
³ Department of Inorganic and Analytical Chemistry, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, H-1521 Budapest, Hungary

* First corresponding author, e-mail: perdew@temple.edu
** Second corresponding author, e-mail: gcsonka@mail.bme.hu
Later this theorem was generalized [4] to electrons that see spin-dependent external potentials \( v_\uparrow (\mathbf{r}) \) and \( v_\downarrow (\mathbf{r}) \). In this case, \( E_{xc}[n_\uparrow, n_\downarrow] \) is a functional of the electron spin-density distributions \( n_\uparrow (\mathbf{r}) \) and \( n_\downarrow (\mathbf{r}) \). In nature, non-relativistic electrons in the absence of a magnetic field always see an external potential \( v(\mathbf{r}) = v_\uparrow (\mathbf{r}) = v_\downarrow (\mathbf{r}) \), so these two exact theories should yield the same exact total density \( n = n_\uparrow + n_\downarrow \) and the same exact total energy (which includes \( E_0 \) as one term).

But approximations to these two exact theories need not agree. We can see this already at the simplest level, the local spin-density approximation [5]

\[
E_{xc}^{\text{LDA}}[n_\uparrow, n_\downarrow] = \int d^3 r n_{\text{ex} \uparrow \downarrow}^{\text{LDA}}(n_\uparrow, n_\downarrow).
\]

(1)

Here \( n_{\text{ex} \uparrow \downarrow}^{\text{LDA}}(n_\uparrow, n_\downarrow) \) is the exchange-correlation energy per particle of an electron gas with uniform or \( \mathbf{r} \)-independent spin densities \( n_\uparrow \) and \( n_\downarrow \). Eq. (1) is exact for spin densities that are uniform or that vary slowly-enough over space. The local density approximation [1]

\[
E_{xc}^{\text{LDA}}[n] = \int d^3 r n_{\text{ex} \uparrow \downarrow}^{\text{LDA}}(n)
\]

(2)

will give the same total energy as Eq. (1) only when the system is fully spin-unpolarized, with \( n_\uparrow (\mathbf{r}) = n_\downarrow (\mathbf{r}) = n(\mathbf{r})/2 \). This is the case for the He atom ground state, but not for the fully spin-polarized H atom ground-state with \( n_\uparrow (\mathbf{r}) = n(\mathbf{r}) \) and \( n_\downarrow (\mathbf{r}) = 0 \). Yet both atoms see a spin-independent external potential.

Tong [5] made an early LDA calculation of the cohesive energy of metallic sodium, i.e., the energy per atom to convert the solid to a vapor of atoms. He found a value 23% larger than the experimental result. Kohn [6] argued that LDA should be accurate for solid sodium (in which the spin-unpolarized valence electron density resembles a uniform gas), and that the error must be in the energy of the atom. Gunnarsson, Lundqvist, and Wilkins [7] applied LSDA to the partly spin-polarized sodium atom, and found a cohesive-energy error of only 4%. It is understandable that approximate spin-density functionals can be more accurate than the corresponding approximate density functionals. The spin-density functionals get more input information about the electrons, and don’t have to be as “smart” as the density functionals would have to be to achieve the same level of accuracy.

The cohesive energy of a solid was never of great interest to condensed matter physicists, who were typically more interested in crystal structures, lattice constants, bulk moduli, phonon frequencies, band structures, etc. However, around 1993 chemists also started to be interested in density functional theory. To a chemist, the atom is rightly the building block of all ordinary matter. As a result, an era began in which the atomization energies of molecules were a kind of “gold standard” for density functionals. (In theoretical chemistry, atomization energies were traditionally used for the calculation of the standard enthalpies of formation, because early wavefunction theory was unable to yield reliable energies for elements in the solid state. In physical chemistry, atomization energies are not required, as the enthalpies of formations require only the standard-state enthalpies of the reagents and the product.) In those earlier days, empirical functionals were fitted to large data sets of atomization energies, and all functionals were judged by their error statistics for atomization energies. That era is now ending, with the availability of much larger and more diverse molecular data sets and direct calculations of the energies of solid state elements.

Perdew and Schmidt [8] have proposed a “Jacob’s Ladder” of spin-density functional approximations:

\[
E_{xc}^{\text{approx}}[n_\uparrow, n_\downarrow] = \int d^3 r n_{\text{ex} \uparrow \downarrow}^{\text{approx}}(n_\uparrow, n_\downarrow, \nabla n_\uparrow, \nabla n_\downarrow, \tau_\uparrow, \tau_\downarrow, \ldots).
\]

(3)

The inputs at each point \( \mathbf{r} \) are the local spin densities (the only inputs in LSDA), their gradients (added in a generalized gradient approximation or GGA), the orbital kinetic energy densities (further added in a meta-GGA), etc. The inputs displayed explicitly in Eq. (3) are all available at each point \( \mathbf{r} \) in a Kohn-Sham calculation, making the first three rungs (LSDA, GGA, and meta-GGA) semilocal and computationally efficient in this hierarchy. The ingredients not displayed explicitly in Eq. (3) are fully nonlocal, so the higher rungs (hybrid-like and random phase approximation (RPA)-like approximations) are computationally more demanding.

Given that the exact total energies for atoms, molecules and solids in the absence of magnetic fields must be the same, we can ask whether for example MAE(DFT)-MAE(SDFT) gets smaller as we climb up the ladder and approach the exact theory. Here MAE is the mean absolute error for some property over a data set, applying the same functional in density functional (DFT) and spin-density functional theory (SDFT). We hope to answer this question numerically in future work. The answer is currently unknown, since nearly all studies are based on SDFT.

3 Atomization Energy Puzzle

As we climb the Jacob’s Ladder of approximate spin-density functionals from LSDA (which is more or less unique) to various flavors of GGA and meta-GGA, there is a clear decrease in the MAE of the atomization energies of molecules and solids, as one would expect.

Following suggestions from Dewar [9] and Delley [10], as further explored by Delley [11], and Grimme [12], Csonka, Ruzsinszky, Tao and Perdew [2] found that the MAE’s of the atomization energies for functionals on all rungs of the ladder...
could be reduced, sometimes drastically, by using atomic energies shifted by the addition of empirical constants, with one set of atomic constants for each functional. [Very recently, optimized atomic correction energies were successfully applied to direct RPA (dRPA) and dual-hybrid dRPA75 [13].] Large atomic shifts indicated an inconsistency between a functional’s description of atoms and its description of molecules built up from those atoms. (The shifts were almost negligibly small for the TPSS [14] meta-GGA.) The fact that the empirically-corrected MAE’s were so small indicated that most functionals were more consistent for molecules than for their constituent atoms. Thus the functionals were expected to be (and are) much more accurate for reaction molecules than for their constituent atoms. Thus the functionals that make for atomization energies are almost unrelated with the errors that it makes for other molecular properties.

Those are the facts, but what is the explanation?

Atoms obviously differ from molecules and solids in several ways: (1) The density of an atom is more confined, and further from the slowly-varying limit, than is the density of a molecule or solid. (2) Open-subshell atoms like carbon might be expected to have smaller gaps between the highest-occupied and lowest unoccupied orbital energies than molecules or semiconducting solids do, leading to near-degeneracy correlation, although SCAN calculations (Table 1) do not support this expectation. (3) Open-subshell atoms like carbon have ground-state densities that are non-spherical, with shapes very different from the shape of the external potential. (4) Open-subshell atoms are partly spin-polarized, while most molecules and solids are spin-unpolarized, so spin-polarization errors can be more troublesome for atomization energies than for reaction or formation energies that involve no free atom. So which of these differences explains the atomization energy puzzle?

The early work [7] on the cohesive energy of sodium metal, described in section 2 of this paper, suggests that the spin-dependence of the approximate functionals provides the right explanation. But that work (largely unknown to chemists) was done on the lowest and least-accurate rung of the ladder. Complicating the answer is the fact that the exchange-correlation energy can be written as

\[ E_x = E_x + E_c, \]

the sum of a large exchange energy and a small correlation energy. Nearly every approximate density functional for exchange satisfies the exact spin-scaling relation [16]

\[ E_x[n_\uparrow, n_\downarrow] = E_x[2n_\uparrow]/2 + E_x[2n_\downarrow]/2. \]

Thus, given a sufficiently accurate \( E_x [n] \), errors in the spin dependence must arise from the relatively small correlation energy \( E_c [n_\uparrow, n_\downarrow] \). To show that these are responsible for the atomization energy puzzle on the higher rungs of the ladder, we need a very accurate spin-density functional. As discussed in the next section, we have one now.

### 4 Exact Constraints and the SCAN Meta-GGA

Although the exact \( E_x [n_\uparrow, n_\downarrow] \) is incomputable, there are exact formal expressions for it which can be used to derive its mathematical features. These features can be regarded as exact constraints, and used to construct approximations that are fully or partly nonempirical. Since the exact constraints are universal (independent of the external potential), constraint-based approximations can be reliable even for systems unlike those on which they have been tested and benchmarked.

The early successes of LSDA, even for metal surfaces, were surprising and demanded an explanation. The explanation [17,18] was not only that LSDA was appropriately normed on the only density for which it could be exact (the electron gas of uniform spin densities), but also that LSDA satisfied a certain subset of exact constraints that it inherited from that appropriate norm. Generalized gradient approximations (GGA’s) like the PBE of Ref. [19] and meta-GGA’s like the TPSS of Ref. [14] were then constructed to satisfy additional exact constraints. The added ingredients on higher rungs of Jacob’s Ladder can be used to satisfy exact constraints that cannot be satisfied on lower rungs.

Recently three of us have proposed SCAN [2], a strongly constrained and appropriately normed meta-GGA. We believe that SCAN is nearly as accurate as a computationally-efficient semilocal exchange-correlation functional can be. SCAN is the first functional to satisfy all 17 known exact constraints that a semilocal functional can, including a tight new lower bound [20] on the exchange energy. But there are still infinitely many ways to satisfy these 17 exact constraints. Thus, to the previous meta-GGA appropriate norms (the electron gas of uniform

| Table 1: Energy gap (eV) between the lowest-unoccupied and highest-occupied atomic orbital energies, from SCAN |
|---|---|---|
| atom | \( \Lambda \) | atom | \( \Lambda \) |
| H | 8.6 | Ne | 19.0 |
| He | 23.5 | Na | 1.7 |
| Li | 2.1 | Mg | 3.6 |
| Be | 4.0 | Al | 0.7 |
| B | 1.4 | Si | 0.4 |
| C | 1.4 | P | 3.3 |
| N | 5.6 | S | 0.7 |
| O | 2.4 | Cl | 1.3 |
| F | 27 | Ar | 12.6 |
or slowly-varying spin densities, and the one-electron atom), SCAN adds the exchange energies and correlation energies of rare-gas atoms and other nonbonded systems in which the exact exchange-correlation hole remains close to its electron (as it must if a semilocal functional is to be very accurate for a system). It is important that, unlike empirical functionals, SCAN is not fitted to any bonded system. Then the SCAN predictions for bonded systems are genuine predictions and not fits.

SCAN is constructed as an interpolation/extrapolation on the dimensionless meta-GGA ingredient

$$\alpha = \left( \frac{\tau - \tau_{\omega}}{\tau_{\text{unif}}} \right),$$

(7)

where \( \tau = \tau_\uparrow + \tau_\downarrow \) is the exact positive orbital kinetic energy density, \( \tau_\omega = |\nabla n|^2 / (8n) \) is the von Weizsäcker kinetic energy density (exact for one- and spin-unpolarized twoelectron ground states), and \( \tau_{\text{unif}} \) is the kinetic energy density of a uniform electron gas. We have argued [21] that \( \alpha \) is the right meta-GGA ingredient because it can recognize and give an appropriate description to three kinds of bonds: covalent single (\( \alpha = 0 \)), metallic (\( \alpha \approx 1 \)), and weak (\( \alpha >> 1 \)). Since many of the exact constraints are for \( \alpha = 0 \) or \( \alpha \approx 1 \), we need an interpolation/extrapolation guided by the remaining constraints and the appropriate norms.

We have tested SCAN and other semilocal functionals on the G3 set (223 atomization energies of molecules), the BH76 set (76 barrier heights to chemical reactions), the S22 set (22 binding energies for weakly-bound molecular complexes, with hydrogen bonds or van der Waals bonds), and the LC20 set (20 lattice constants of solids). Table 2 shows the MAE’s in kcal/mol (first three sets) or Å (fourth set). LSDA is the local spin-density approximation, PBE [19] is a GGA, while TPSS [14] is our earlier nonempirical meta-GGA and SCAN [2] is our current nonempirical meta-GGA. pSCAN (proto-SCAN) is identical to SCAN except in the way it interpolates the \( \alpha = 0 \) correlation energy between fully spin-unpolarized and fully spin-polarized limits. pSCAN is in fact the version of SCAN that we first submitted for publication, while SCAN is the version finally accepted for publication.

Table 2 Mean absolute errors of semilocal spin-density functionals for three molecular (G3, BH76, S22, in kcal/mol) and one solid-state (LC20, in Å) test sets, as described in the text. References for the test sets are given in Ref. [2].

<table>
<thead>
<tr>
<th></th>
<th>G3</th>
<th>BH76</th>
<th>S22</th>
<th>LC20</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSDA</td>
<td>83.7</td>
<td>15.4</td>
<td>2.3</td>
<td>0.081</td>
</tr>
<tr>
<td>PBE</td>
<td>22.2</td>
<td>9.2</td>
<td>2.8</td>
<td>0.059</td>
</tr>
<tr>
<td>TPSS</td>
<td>5.8</td>
<td>8.7</td>
<td>3.7</td>
<td>0.043</td>
</tr>
<tr>
<td>pSCAN</td>
<td>12.0</td>
<td>7.5</td>
<td>0.9</td>
<td>0.016</td>
</tr>
<tr>
<td>SCAN</td>
<td>5.7</td>
<td>7.7</td>
<td>0.9</td>
<td>0.016</td>
</tr>
</tbody>
</table>

A first observation is that most calculated properties (the atomization energies of G3, the barrier heights of BH76, and the lattice constants of LC20) improve as we climb from the first rung (LSDA) to the second (GGA PBE) and then to the third (meta-GGA’s TPSS, pSCAN, and SCAN). An exception is the weak-interaction binding energies of S22, which actually worsen from LSDA to PBE to TPSS but then improve dramatically in pSCAN and SCAN. The improvement in the pSCAN and SCAN lattice constants is also dramatic. The added constraints and norms beyond those in TPSS are doing what they should do.

A second observation, and one more relevant to the conclusions of this article, is that proto-SCAN and SCAN are identical for S22 and LC20 (where all systems are fully spin-unpolarized). They make very similar errors for BH76, even though the transition states of chemical reaction are typically partly spin-polarized. But SCAN is much better than proto-SCAN for the G3 atomization energies, where the open-shell free atoms are typically partly spin-polarized.

The difference between proto-SCAN and SCAN arises only from the way in which we interpolate the \( \alpha = 0 \) correlation energy between fixed fully spin-unpolarized and fully spin-polarized limits. Let us define the relative spin polarization

$$\zeta = \left( n_\uparrow - n_\downarrow \right) / \left( n_\uparrow + n_\downarrow \right),$$

(8)

The function that exactly interpolates the exchange energy per particle of a uniform electron gas between its \( \zeta = 0 \) and \( |\zeta| = 1 \) limits is

$$d_\zeta (\zeta) = \left[ \left( 1 + \zeta \right)^{\frac{2}{3}} + \left( 1 - \zeta \right)^{\frac{4}{3}} \right] / 2,$$

(9)

which varies from 1 at \( \zeta = 0 \) to \( 2^{\frac{1}{3}} = 1.26 \) at \( |\zeta| = 1 \). For \( \alpha = 0 \) correlation, proto-SCAN uses the interpolation

$$G_\alpha^p (\zeta) = \left[ 2^{\frac{1}{3}} - d_\zeta (\zeta) \right] / \left[ 2^{\frac{1}{3}} - 1 \right],$$

(10)

which varies from 1 at \( \zeta = 0 \) to 0 at \( |\zeta| = 1 \). The latter limit is needed to make the correlation energy per particle vanish for all one-electron (\( \alpha = 0 \) and \( |\zeta| = 1 \) ) densities. But SCAN uses a different interpolation between the same two limits,

$$G_\alpha (\zeta) = \left[ 1 + 2.363 \left[ 1 - d_\zeta (\zeta) \right] \right] \left[ 1 - \zeta^2 \right],$$

(11)

designed to make the low-density limit of the SCAN exchange-correlation energy nearly independent of \( \zeta \) over the range \( 0 \leq |\zeta| \leq 0.7 \), as in TPSS. In this way, SCAN and TPSS satisfy, as best they can, the exact constraint that the exchange-correlation energy per particle in the low-density limit should be independent of \( \zeta \). Thus both SCAN and TPSS yield accurate atomization energies, while proto-SCAN is less accurate.
The two interpolation functions are compared in Table 3. For a given $|\varsigma|$ between 0 and 1, $G_\varsigma > G^c_\varsigma$, so SCAN provides a little more negative correlation energy in a partly spin-polarized atom than proto-SCAN does. This slightly lowers the energy of the atom and reduces the too-high atomization energies found in proto-SCAN.

<table>
<thead>
<tr>
<th>$\varsigma$</th>
<th>$d_\varsigma$</th>
<th>$G^c_\varsigma$</th>
<th>$G_\varsigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.2</td>
<td>1.009</td>
<td>0.966</td>
<td>0.979</td>
</tr>
<tr>
<td>0.4</td>
<td>1.036</td>
<td>0.861</td>
<td>0.891</td>
</tr>
<tr>
<td>0.6</td>
<td>1.083</td>
<td>0.680</td>
<td>0.802</td>
</tr>
<tr>
<td>0.8</td>
<td>1.153</td>
<td>0.410</td>
<td>0.594</td>
</tr>
<tr>
<td>1.0</td>
<td>1.260</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

While some exact constraints can only be satisfied on a sufficiently high rung of Jacob’s Ladder, others (including the absence of spin-polarization dependence in the low-density limit) are already satisfied on the first or LSDA rung and get harder to satisfy on higher rungs.

The effect on the atomization energies of stricter enforcement of the low-density-limit constraint is rather subtle. It is magnified in the G3 set by the presence of numerous large molecules that contain many C atoms that “compound the error”. While the cohesive energy of a solid is an intensive (per atom) quantity, the atomization energy of a molecule is an extensive one; it and its density functional errors grow with the size of the molecule.

5 Conclusions

The atomization energy errors of a density functional can strongly magnify minor errors in its spin-polarization dependence which are of little importance for other properties of many molecules and solids. Thus density functionals should not be judged primarily by their atomization energy errors, but by a wider spectrum of tests. Although SCAN atomization energies are not significantly better than TPSS atomization energies, we believe that SCAN is significantly better than TPSS. We already have evidence [2] that SCAN can reach a new level of accuracy for the energy differences between molecules or between solids at fixed atomic composition. In future work, we will test SCAN and other functionals for reaction energies, isomerization energies, structural energy differences in solids, and formation energies (and thus for relative stabilities).

Acknowledgement

JPP was honored by the conferral of a Doctor Honoris Causa degree from the Budapest University of Technology and Economics in May 2015. This work was supported in part by the U.S. National Science Foundation under Grant No. DMR-1305135 (JP, JS, and AR).

References


Mechanism of methyl-DNA recognition by methyl-CpG-binding domain proteins

Abstract:
Recognition of the methylated regions of the DNA plays an important role in the epigenetic processes. We analyze the interactions between the methylated DNA and the methyl-CpG-binding proteins using two models. The first model was built from a methylated or non-methylated cytosine, a guanine and an arginine residue in the experimental arrangement. We applied the M06L density functional method with a small, polarized double-ζ basis set for the geometry optimizations, and the MP2 method with polarized triple-ζ basis set for the energy calculations. The second model was built from two methylcytosines, guanines, guanidinium groups plus an additional carboxyl group in the experimental arrangement. We applied the B3LYP method with a small, polarized double-ζ basis set for the geometry optimizations and thermal corrections. The single point energies were obtained from dual-hybrid dRPA75 and dRPA@PBE0 calculations supplemented by a moderately large polarized triple-ζ basis set. The hydration effects were modeled by adding explicit water molecules combined with the COSMO solvation model. These calculations revealed that the hydrophobic interaction has the largest contribution to the Gibbs interaction energy and turns the arginine side chains into hydrogen bonding position. Our results show that the translation of the protein along the DNA double helix is sterically hindered by the contact of its arginine side chains with the methyl groups of the methyl cytosines. This supports a hopping mechanism for the searching movement of the protein along the DNA.
Mechanism of methyl-DNA recognition by methyl-CpG-binding domain proteins

Pál D. Mezei, Gábor I. Csonka

2016-01-25

Abstract

Recognition of the methylated regions of the DNA plays an important role in the epigenetic processes. We analyze the interactions between the methylated DNA and the methyl-CpG-binding proteins using two models. The first model was built from a methylated or non-methylated cytosine, a guanine and an arginine residue in the experimental arrangement. We applied the M06L density functional method with a small, polarized double-ζ basis set for the geometry optimizations, and the MP2 method with polarized triple-ζ basis set for the energy calculations. The second model was built from two methylcytosines, guanines, guanidinium groups plus an additional carboxyl group in the experimental arrangement. We applied the B3LYP method with a small, polarized double-ζ basis set for the geometry optimizations and thermal corrections. The single point energies were obtained from dual-hybrid dRPA75 and dRPA@PBE0 calculations supplemented by a moderately large polarized triple-ζ basis set. The hydration effects were modeled by adding explicit water molecules combined with the COSMO solvation model. These calculations revealed that the hydrophobic interaction has the largest contribution to the Gibbs interaction energy and turns the arginine side chains into hydrogen bonding position. Our results show that the translation of the protein along the DNA double helix is sterically hindered by the contact of its arginine side chains with the methyl groups of the methyl cytosines. This supports a hopping mechanism for the searching movement of the protein along the DNA.

Introduction

Over the genome, epigenetics describes changes in the regulation of gene expression that can be conserved during cell division but without changing the nucleotide sequence of the DNA. The epigenetic information is carried by the DNA accessibility, DNA methylation and histone modifications, and can be mapped by high-throughput molecular assays. The DNA methylation plays a central role in embryonic development, cell differentiation, genomic imprinting, X-chromosome inactivation and neoplastic transformation. The methylation occurs at the C5 position of the cytosine residues, mostly in the context of palindromic CpG dinucleotides in body cells (and predominantly in non-CpG dinucleotides in neurons). In mammalian genomes, ~70% of the CpG sites are methylated. Generally, genes can be transcribed from methylation-free promoters, while the noncoding DNA background is permanently silenced. The CpG islands (CGI) are short (<200 bp) non-methylated CpG-rich regions mostly in the promoter region of genes or in the first exon. In the less frequent non-CGI promoters, the level of DNA methylation in the promoter region inversely correlates with the transcriptional activity of the gene. Contrarily, the more frequent CGI promoters are kept free from methylation even in their silenced state.
The methylation markers are de novo established by DNA methyltransferase (DNMT) proteins such as DNMT3A, DNMT3B, and the catalytically inactive DNMT3L member, which modulates the activity of the other two.\textsuperscript{20, 21, 22, 23} The patterns are likely formed due to the exclusion of the DNMT proteins from the CpG islands, and maintained by the DNA methyltransferase DNMT1 and the ubiquitin-like plant homeodomain and RING-finger domain containing UHRF1 proteins.\textsuperscript{24, 25, 26} The methylation is usually copied after DNA synthesis, thus the epigenetic states can be conserved during cell division.\textsuperscript{27} Between generations of organisms, there are two sensitive periods, in which epigenetic reprogramming happens. In the first sensitive period during the early embryonic development, the paternal genome is enzymatically demethylated (active demethylation) either by oxidation to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylycitosine catalyzed by the ten-eleven translocation TET1-3 proteins, or by deamination to thymine followed by base excision repair (BER) catalyzed by the thymine-DNA glycosylase (TDG) enzyme.\textsuperscript{28-34} In addition, the maternal genome becomes less methylated (passive demethylation) by dilution due to the lacking DNMT1 enzyme. At later steps of the embryonic development, methylation is reestablished in cell-specific manner. In the second sensitive period in primordial germ cells, the somatic marks are replaced by sex-dependent marks; furthermore, imprinted genes are reprogrammed only in this period.

The cytosine methylation alters the appearance of the major groove and the interactions between DNA and DNA binding proteins, which results in an epigenetic marker. The methylation mark can directly prevent the binding of transcriptional factors when it is present at their target sites.\textsuperscript{10} In other cases, the methylated CpG sites (mCpG) are specifically recognized by mCpG binding proteins (MBP), which recruit repressive chromatin modifiers and remodeling complexes.\textsuperscript{35-38} MBPs have three families: the methyl-CpG-binding domain (MBD) proteins, the UHRF proteins, and the Kaiso-like proteins.\textsuperscript{20} The corresponding three classes of methylated cytosine recognition domains are the MBDs, the SET and RING-finger-associated (SRA) domains, and the C2H2 zinc finger (ZnF) DNA binding domains, respectively. Generally, MBDs recognize symmetrically methylated CpG dinucleotides, SRA domains bind hemi-methylated CpG sites, and ZnF domains prefer the methylated CpGs within longer specific sequences. MBD and ZnF proteins approach the B-DNA from the major groove, and form hydrogen bonds and specific interactions with the bases constituting one or two 5mCyt-Arg-Gua triads.\textsuperscript{37} Contrarily, the SRA domain proteins flip the methylated cytosine into a hydrophobic binding cage.\textsuperscript{38, 39, 40}

MBD proteins merit particular interest among the nuclear factors, which establish the connection between DNA methylation and the histone modifications. Several experimental results suggest that MBD proteins are bound to the aberrantly methylated sequences triggering transcriptional misregulation.\textsuperscript{41} Furthermore, mutations in the epigenetic marker machinery can also lead to carcinogenesis or Rett syndrome.\textsuperscript{32, 43, 44} In cancer, there is a global decrease of CpG methylation, but the CGI promoters of the tumor suppressor genes become hypermethylated.\textsuperscript{36, 45, 46} Inter alia the p16\textsuperscript{INK4a}, hMLH1 and BRCA1 genes are silenced in many types of cancer.\textsuperscript{47, 48} Tumors can be classified according to their methylation profile.\textsuperscript{49, 50} In addition, the aberrant promoter hypermethylation is involved in silencing of the transcriptional inhibitor micro RNAs, which leads to a tumor-specific miRNA expression profile.\textsuperscript{51, 52, 53}

MBD proteins (MeCP2, MBD1, MBD2) specifically bind mCpG steps and recruit corepressor complexes with histone deacetylase (HDAC) and nucleosome remodeling activities (such as MeCP1 and NuRD) inhibiting the transcription.\textsuperscript{41, 54} (Of course MBD proteins can have also a methylation-independent DNA binding site beside the MBD domain.\textsuperscript{55, 56, 57, 58}) The solution structure of the MBD1-mDNA,\textsuperscript{59, 60} MBD2-mDNA, MBD4-mDNA, and the XRD structure of MeCP2-mDNA complexes can be found under the PDB entry codes 1IG4, 2KY8, 2MOE, and 3C2I, respectively. It was reported for the MBD1 protein that during the recognition process, the MBD domain reverses its orientation on the DNA double helix without its complete dissociation from the DNA (flipping).\textsuperscript{61} In this domain, there
are a four-stranded beta sheet (β1-4) and an alpha helix (α1), which orient the double-stranded DNA (dsDNA). Furthermore, one of the phosphate backbones of the dsDNA is partially surrounded by a loop (L1) between the β2 and β3 strands. The ability of the MBD proteins to bind the phosphate backbone of the DNA double helix reflects non-specific interactions. It was suggested for MBD4 that the recognition of the fully methylated CpG dinucleotides goes by facilitated diffusion (likely by hopping). There are several highly conserved residues on the interacting surface: among the others Arg'22', Asp'32', Tyr'34' and Arg'44'. (In this paper, we use the human MBD1 numbering for residues (denoted by primes), although the actual numbering differs for each protein.) The Arg'22' and Arg'44' side chains participate in the recognition of the 5-methylcytosine forming two 5mCyt-Arg-Gua triads. The Arg'44' side chain shows larger flexibility than the Arg'22' side chain. The Arg'22' side chain is strongly fixed either by two hydrogen bonds, or by a shorter salt bridge towards Asp'32'. The Arg'44' side chain is weakly fixed either by a single hydrogen bond or by a longer salt bridge towards Glu/Gln/Ser'48'.

The function of MBD proteins is reported to depend on the surrounding sequence and methylation pattern. The MeCP2 protein recognizes mCpG steps with adjacent A/T bases likely because the Asx-ST motif stabilizes the narrowed minor groove of the AT run. It binds to a single methylated CpG and ensures the long-term silencing of the methylated DNA. The MBD1 protein prefers TmCGCA and TGmCGCA sequences, and it can bind also to hypomethylated promoters. The MBD2 protein likely binds to the mCGG sequence. It has low affinity to the DNA and requires densely methylated sequences. The mammalian MBD3 protein cannot bind to methylated DNA, because it has a phenylalanine in position ‘34’ instead of tyrosine. The MBD4 protein weakly binds to mCpG dinucleotides and rapidly exchanges between successive mCpG sites, meanwhile its glycosylase domain repairs the mCpG/TpG mismatches. Furthermore, the MBD2-MBD3 heterodimer binds to hemimethylated DNA and maintains the silenced state of the chromatin by recruiting HDACs and DNMT1.

Figure 1. (a) Secondary structure (α-helix: red; β-strand: yellow; loop: green; DNA: blue; methylation: sphere) and (b) molecular elements (foreground: lines; background: sticks; nucleobase: grey; amino acid carbon atom: green; oxygen atom: red; nitrogen atom: blue; water molecule: red sphere; hydrogen atoms: hidden; hydrogen bridge: dashed line) in the recognition region according to the crystal structure of the MeCP2-mDNA complex (PDB: 3C2I).

In an earlier theoretical protein-DNA interface study the interaction was modeled on a Gua-Arg-Gua stair motif and three types of pairwise interactions were observed: aromatic base stacking, hydrogen
bonding, and cation-π interactions. The DNA-binding domain of Tc3 transposase from *Caenorhabditis elegans* (1TC3) the gas phase interaction energy between Arg C236 and the two successive Gua A7 and A8 was calculated using the experimental geometry. The calculated counterpoise corrected (CP) MP2/6-31G(0.8,0.2,p) interaction energy was -41.9 kcal mol\(^{-1}\). Due to the nonoptimal geometries taken from the crystal structures this energy might be too high. In order to show the geometry optimization effects and mimic the DNA-protein environment the structure was optimized. The resulting MP2/6-31G(1.2,p)+CP/MP2/6-31G(d,p) gas-phase interaction energy was -49.7 kcal mol\(^{-1}\). Finally, it was shown that the solvent plays a crucial role in the interaction energy.

The recognition of methylated DNA through methyl-CpG binding domain proteins was also studied by quantum chemical calculations. It was found that that methylation increases the buried hydrophobic surface between MBD and DNA by about 100 Å\(^2\) and this leads to -1.5 kcal mol\(^{-1}\) stabilization due to the presence of the methyl group, which is close to the value (-1.2±0.1 kcal mol\(^{-1}\)) obtained through free energy perturbation calculations. It was concluded that the effects of the 5-methyl group arise from a change of binding affinity due to an increased contact area and cation-π interaction.

However, the suggested cation-π interactions have not been found universal in 5mCyt-Arg-Gua triads, since in contrast to the MBD proteins, ZnF proteins cannot distinguish TpA pairs from mCpG pairs. The recognition might be explained rather by hydrophobic interactions because the more oxidized is the group at the fifth position of the cytosine (*i.e.* 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylycytosine), the weaker the ZnF proteins bind it. Furthermore, structurally conserved water molecules were suggested to determine further mCpG-binding specificity of MeCP2. The W1 and W2 water molecules form tetrahedral orientation of hydrogen bonds with the amino group of methyl cytosines, as well as with the Tyr‘34’, Arg‘44’ and Asp‘32’ amino acids bridging the DNA and the protein as shown in the Figure 1. As it was mentioned earlier the increase of the hydrophobic surface introduced by the methyl groups was estimated to decrease the Gibbs energy by 1.5 kcal mol\(^{-1}\), but this hardly explains the natural preference for the mCpG dinucleotides and does not capture the effect of structural water molecules on the interacting surface.

In this paper, we start from solution and crystal MBD-mDNA structures given in the literature. We apply a simple model to qualitatively discover the interactions within the mCyt-Arg-Gua triads. Next we construct MeCP2-mDNA model complex with two structural water molecules, the hydrated MeCP2 model, and the six or seven water hydrated mDNA models. Finally, based on the experimental and computational results, we suggest a mechanism for the methyl DNA recognition by the MBD proteins.

**Computational details**

The model for the mCyt-Arg-Gua triads is built up from the two bases and from the arginine terminated by methyl groups from both sides. All molecules positioned according to the experimental PDB structure of MBD2-mDNA complex are shown in Figure 2 for the mCyt-Arg‘44’-Gua and mCyt-Arg‘22’-Gua triads. We have selected the M06L density functional theory (DFT) for geometry optimizations because this method is efficient and it also includes some middle range correlation effects, which are missing from many standard DFT methods and such correlation effects might be important for more realistic non-covalent interaction energies. The polarized double zeta valence 6-31G(d) basis set was selected because of the size of the system and the complexity of the time consuming geometry optimizations. First we model the interaction between a rigid DNA major groove and a rigid protein side chain and optimize only the positions of the terminal hydrogen atoms. Next in a less constrained geometry optimization the arginine atoms were also allowed to change their positions except the terminal methyl groups. In the third geometry optimization, the relaxation of the bases was allowed except for the directly hydrogen bonded oxygen and hydrogen atoms participating in the base pair formation. The positions of the terminal hydrogen atoms were also fixed to simulate the backbone rigidity. This simulates a more
flexible DNA structure, which can adopt to the interaction with the protein. Finally, we also optimized the positions of the arginine terminal methyl groups in order to model the flexibility of the peptide chain.

The single point energies of the equilibrium geometries were computed more accurately with RI-MP2/def2-TZVP(-f) model chemistry combined with def2-TZVPP/J auxiliary basis set using RJCOSX approximation for the HF energy calculation. All these calculations were made in gas-phase with the Orca 3.0 quantum chemistry software. The interaction energy is calculated as the energy of the 5-MeCyt-Arg-Gua complex minus the energy of the Arg, Gua and mCyt shown in Figure 2.

![Figure 2. Initial structures of the (a) mCpG-Arg’22’ and (b) mCpG-Arg’44’ models (carbon atom: dark grey; hydrogen atom: light grey; oxygen atom: red; nitrogen atom: blue; hydrogen bond: dashed line; cation-π interaction: light blue beam; dispersion interaction: yellow beam). The molecules are positioned according to the solution NMR structure of the MBD2-mDNA complex (PDB: 2KY8). However, the positions of the terminal hydrogen atoms were optimized. (c) The numbering of the mCyt-Arg-Gua triads is indicated on a schematic arrangement.](image)

The larger model geometry is based on the crystal structure of the MeCP2-mDNA complex with two structural water molecules (PDB: 3C2I). The initial model geometry for the mCpG recognition site is built up from two 1,5-dimethyl cytosines (1,5-Me2Cyt), two 9-methyl guanines (9-MeGua), two methyl guanidinium cations (MeGdm+) representing of Arg’22’ and Arg’44’ where the methyl groups positioned according to the first atom of the alkyl chains of the arginines, an acetate anion (AcO−) representing the Asp’32’; and two structural water molecules. All components are arranged as shown in Figure 3a.

The geometry optimization and thermochemical analysis were performed with the Gaussian 09 quantum chemistry software using the B3LYP/6-31(d) method. This method is known to give reasonable equilibrium bond lengths as it contains 20% of exact exchange, which compensates the usual overestimation error of semi-local density functional approximations. The positions of the terminal
methyl groups (9-Me of guanines and 1-Me of cytosines) were frozen during the geometry optimizations to simulate the backbone. The DNA double helix is in the B-DNA conformation. The two structural water molecules are fixed by hydrogen bonds in the model geometry. Figure 3b-d shows a possible further hydration with six and seven explicit water molecules.

As the B3LYP/6-31G(d) single point energies of the equilibrium geometry might be quite inaccurate we also calculated considerably more accurate dRPA75 and dRPA@PBE0 single point energies with the (aug)-cc-pVTZ(-f,-d) basis set (abbreviated here as aTZ(-f,-d)) using the MRCC (09/07/2015) quantum chemistry program. These latter methods quite accurately and seamlessly yield the non-covalent interaction energies. Furthermore, if the dRPA75 and dRPA@PBE0 results agree well, we can exclude the density driven and self-interaction errors as these methods use two different Kohn-Sham determinants (PBE determinants with 0.75 and 0.25 fraction of exact exchange, respectively). These errors could make the non-self-consistent RPA calculations highly uncertain as dRPA correlation energy is sensitive to the KS determinants. Then we also apply the COSMO solvation model (using Orca) which includes further solvent effects. We used the $\varepsilon = 78.3553$ relative permittivity of water on 25 °C to calculate the screened charges.

![Figure 3. Optimized structures of (a) the MeCP2-mDNA model geometry with two structural water molecules, (b) the hydrated MeCP2 model, as well as (c) the six and (d) seven water hydrated mDNA models. (nucleotide part: capped sticks; amino acid part: balls and sticks; water molecules: balls and sticks; carbon atom: dark grey; hydrogen atom: light grey; oxygen atom: red; nitrogen atom: blue)](image-url)
Results and discussion

Hydrogen bonding

Firstly, we discuss the gas-phase models for the mCyt-Arg-Gua triads on different geometry optimization levels. The strengthening of the hydrogen bonds between the arginine side chain and the guanidine can be followed in Table 1. On the first geometry optimization level, as we optimize only the positions of the terminal hydrogen atoms, the position of the arginine side chain comes from the original solution structure. The plane of the guanidinium group is not coplanar with the plane of the guanine. The torsion of the Arg’44’ guanidinium group is larger than that of the Arg’22’ guanidinium group. On the second geometry optimization level, the Arg’22’ side chain relaxes on the flexible protein surface. The full relaxation of the Arg’44’ guanidinium group needs also a flexible protein inner structure, because it happens only on the fourth geometry optimization level. The hydrogen bond lengths change from 3.0-3.3 Å to 2.7-3.0 Å. The methylation has a small effect on the relaxation of the guanidinium group.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg’44’</td>
<td>11.7</td>
<td>11.4</td>
<td>11.4</td>
<td>-1.5</td>
</tr>
<tr>
<td>CpG-Arg’44’</td>
<td>18.5</td>
<td>11.5</td>
<td>10.6</td>
<td>3.8</td>
</tr>
<tr>
<td>mCpG-Arg’22’</td>
<td>-8.4</td>
<td>-1.2</td>
<td>-6.1</td>
<td>4.6</td>
</tr>
<tr>
<td>CpG-Arg’22’</td>
<td>-8.4</td>
<td>-3.4</td>
<td>-7.4</td>
<td>-2.2</td>
</tr>
</tbody>
</table>

Table 1. Torsion angles (in degree) of the hydrogen bonds (NωArg-OGua-N7Gua-Nω’Arg) in the methylated and non-methylated model structures. (Torsion angles between -8° and +8° are highlighted.)

Cation-π interaction

The strengthening of the cation-π interactions can be followed in Table 2. It might be expected that the inductive effect of the methyl group on the aromatic cytosine ring strengthens the cation-π interaction. However, the interaction energy rather depends on the equilibrium distance between the center of the cytosine ring and the central carbon atom of the guanidinium group, which can be determined also by other influences. On the second geometry optimization level, the guanine attracts the flexible arginine side chain by hydrogen bonding, and the cation-π interaction strengthens at the same time. On the third geometry optimization level, the flexible DNA major groove allows the arginine side chain to get even closer to the aromatic cytosine ring, and the importance of the cation-π interaction further increases. The above mentioned distance is shorter in the methylated complex than in the non-methylated one. Interestingly, the guanidinium-cytosine distance increases in the non-methylated structures between the third and fourth geometry optimization levels.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg’44’</td>
<td>5.21</td>
<td>4.84</td>
<td>4.69</td>
<td>4.42</td>
</tr>
<tr>
<td>CpG-Arg’44’</td>
<td>5.21</td>
<td>4.90</td>
<td>4.71</td>
<td>4.85</td>
</tr>
<tr>
<td>mCpG-Arg’22’</td>
<td>4.89</td>
<td>4.71</td>
<td>4.53</td>
<td>4.43</td>
</tr>
<tr>
<td>CpG-Arg’22’</td>
<td>4.89</td>
<td>4.77</td>
<td>4.57</td>
<td>4.60</td>
</tr>
</tbody>
</table>

Table 2. Distances (in Å) between the center (CωArg atom) of the positively charged guanidinium group and the center of the aromatic cytosine ring. (Distances under 5 Å are in bold.)
Dispersion interaction

The dispersion interactions can be followed by the distance between the arginine side chain and the cytosine methyl group (Table 3). The average distance from the guanidinium group is represented by the distance from the C\textsubscript{\textomega} atom. This distance decreases on the second and third geometry optimization levels but not on the fourth level. Contrarily, the alkyl part of the arginine side chains gets even closer to the cytosine methyl group also on the fourth geometry optimization level. Through the optimization, the alkyl part of the Arg’44’ side chain becomes closer to the cytosine methyl group than that of the Arg’22’ side chain.

Table 3. Distances (in Å) between the methyl carbon atom of the methyl cytosine and the arginine side chain carbon atoms. (Distances shorter than 5 Å are in bold.)

<table>
<thead>
<tr>
<th>Method</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg’44’</td>
<td>4.10</td>
<td>3.97</td>
<td>3.62</td>
<td>3.60</td>
</tr>
<tr>
<td>mCpG-Arg’22’</td>
<td>3.99</td>
<td>3.75</td>
<td>3.72</td>
<td>3.60</td>
</tr>
</tbody>
</table>

Thermochemical analysis

The changes in the gas-phase interaction energies can be followed in Table 4. The interaction energy differences between the methylated and non-methylated model structures indicate that the effect of methylation is smaller for Arg’22’ than for Arg’44’ because the alkyl moiety of the Arg’22’ side chain cannot get into the proximity of the cytosine methyl group so much.

Table 4. Gas-phase interaction energies (in kcal mol\(^{-1}\)) for the methylated and non-methylated CpG-Arg model structures. The geometries were optimized with the M06L/6-31G(d) method. The interaction energies were also calculated with the RI-MP2/def2-TZVP(-f) method. The interaction energy difference between the methylated and non-methylated complexes are in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>M06L</td>
<td>mCpG-Arg’44’</td>
<td>-32.1</td>
<td>-40.9</td>
<td>-42.1</td>
</tr>
<tr>
<td>CpG-Arg’44’</td>
<td>-31.2</td>
<td>-39.8</td>
<td>-40.5</td>
<td>-41.6</td>
</tr>
<tr>
<td>(\Delta E_{\text{Me}}-\Delta E_{\text{H}})</td>
<td>-0.9</td>
<td>-1.1</td>
<td>-1.6</td>
<td>-1.7</td>
</tr>
<tr>
<td>mCpG-Arg’22’</td>
<td>-33.6</td>
<td>-41.3</td>
<td>-43.3</td>
<td>-43.7</td>
</tr>
<tr>
<td>CpG-Arg’22’</td>
<td>-33.4</td>
<td>-41.1</td>
<td>-42.4</td>
<td>-42.6</td>
</tr>
<tr>
<td>(\Delta E_{\text{Me}}-\Delta E_{\text{H}})</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-0.9</td>
<td>-1.1</td>
</tr>
</tbody>
</table>

RI-MP2

<table>
<thead>
<tr>
<th>Method</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg’44’</td>
<td>-31.8</td>
<td>-39.8</td>
<td>-40.6</td>
<td>-42.0</td>
</tr>
<tr>
<td>CpG-Arg’44’</td>
<td>-30.9</td>
<td>-38.7</td>
<td>-39.3</td>
<td>-39.8</td>
</tr>
<tr>
<td>(\Delta E_{\text{Me}}-\Delta E_{\text{H}})</td>
<td>-0.9</td>
<td>-1.1</td>
<td>-1.3</td>
<td>-2.2</td>
</tr>
<tr>
<td>mCpG-Arg’22’</td>
<td>-32.7</td>
<td>-40.2</td>
<td>-41.9</td>
<td>-42.0</td>
</tr>
<tr>
<td>CpG-Arg’22’</td>
<td>-32.6</td>
<td>-39.6</td>
<td>-40.8</td>
<td>-41.1</td>
</tr>
<tr>
<td>(\Delta E_{\text{Me}}-\Delta E_{\text{H}})</td>
<td>-0.1</td>
<td>-0.6</td>
<td>-1.1</td>
<td>-0.9</td>
</tr>
</tbody>
</table>
In the following part, we discuss the solvent effects using two larger models with different number of explicit water molecules on the interacting surface. In the weakly hydrated mDNA model, two structural water molecules are arranged in the center somewhat shifted from the axial orientation towards the guanines, and two-two water molecules are singly bonded to the guanines. These six water molecules are also bonded to each other forming a chain. In the strongly hydrated mDNA model, one water molecule is added to the position of the Asp’32’ binding site, forming a hydrogen bonded pentagon coplanar to the bases. This causes the rearrangement of the two central water molecules parallel to the DNA axis. In the optimized geometry of the hydrated MeCP2 interacting site model, the water bridges among Arg’22’, Asp’32’ and Arg’44’ stretch two hydrogen bonded pentagons from the heteroatoms sharing one side, which rotates the plane of the guanidinium group of Arg’44’ towards Arg’22’ and Asp’32’. This group rotates back into the plane of the guanine because of the formation of the two hydrogen bonds as it was observed for the relaxation of the individual mCpG-Arg’44’ triad. The position of the other guanidinium group of Arg’22’ does not change so much because the Asp’32’ residue stiffens it. Through a single bridging water molecule, the heteroatoms of Asp’32’, Arg’22’, the attached guanine, and its cytosine base pair form again a hydrogen bonded pentagon coplanar with the bases.

In the B3LYP/6-31G(d) thermochemistry in the presence of the explicit water molecules, the electronic interaction energy is highly positive (Table 5). The zero-point vibrational energy decreases this value (using the harmonic frequencies instead of the anharmonic frequencies probably causes a 0.2-0.3 kcal mol\(^{-1}\) overestimation of the zero-point vibrational energies in these cases), but it still remains highly positive. The thermal corrections to the energy and to the enthalpy slightly increase this value. Considering the entropy results in a moderately large positive Gibbs energy. Only the solvation correction from the COSMO model can result in a negative Gibbs energy in this model chemistry. Changing the B3LYP single point energies to highly accurate dRPA75 and dRPA@PBE0 single point energies has large effect on the thermochemistry of biomolecules.\(^{75}\)\(^ {76}\) These methods can describe well the non-covalent interactions which are missing from the B3LYP interaction energies. Therefore, the Gibbs energies even without the COSMO correction become negative. The interaction energy differences between the dRPA75 and dRPA@PBE0 interaction energies are within 1 kcal mol\(^{-1}\), which suggests that the electronic energies are not loaded by density driven or self-interaction errors. The uncorrected Gibbs energies of the weakly and strongly hydrated models agree within 1 kcal mol\(^{-1}\) again. The comparison between the final COSMO corrected Gibbs energies shows 4 kcal mol\(^{-1}\) difference. This can be explained as the addition of a single water molecule makes the hydrated mDNA model structure more compact, and thus reduces the solvent accessible polar surface for the COSMO model.

<table>
<thead>
<tr>
<th>Energy</th>
<th>Six water molecules</th>
<th>Seven water molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B3LYP</td>
<td>dRPA75</td>
</tr>
<tr>
<td>(E_e)</td>
<td>114.5</td>
<td>59.9</td>
</tr>
<tr>
<td>(E_e + ZPVE)</td>
<td>99.4</td>
<td>44.8</td>
</tr>
<tr>
<td>(E(298K))</td>
<td>102.8</td>
<td>48.2</td>
</tr>
<tr>
<td>(H(298K))</td>
<td>108.7</td>
<td>54.1</td>
</tr>
<tr>
<td>(G(298K))</td>
<td>13.1</td>
<td>-41.5</td>
</tr>
<tr>
<td>(G(298K) + COSMO)</td>
<td>-28.6</td>
<td>-83.2</td>
</tr>
</tbody>
</table>

Table 5. Thermochemistry (in kcal mol\(^{-1}\)) for the protein-DNA interactions using six and seven explicit structural water molecules and with approximated additional COSMO solvent effects. B3LYP/6-31G(d), dRPA75/aTZ(-f,-d) and dRPA@PBE0/aTZ(-f,-d) single point energies are shown. (The more accurate dRPA standard Gibbs energies are in bold.)
Mechanism for methyl DNA recognition

The arginine side chains recognize the guanine bases from the major groove by reading their hydrogen bonding donor-acceptor pattern. On a fully rigid interacting surface, the coplanar orientation of the guanidinium groups to the planes of the base pairs would geometrically mean an additional steric hindrance for the protein from sliding on the DNA double helix towards both directions, when both CpG stairs are methylated (Figure 4). The Arg’22’ group would form a larger barrier for sliding towards the 3’→5’ direction with respect to the second DNA strand, and the Arg’44’ group would form a smaller barrier for sliding towards the 3’→5’ direction with respect to the first DNA strand.

However, the interacting surface has some flexibility. Furthermore, the Arg’44’ side chain is more flexible in comparison to the Arg’22’ side chain, which is strongly fixed by the Asp’32’ side chain with hydrogen bonding (locked state). The Arg’44’ guanidinium group can rotate towards the Arg’22’ guanidinium group (closed/open states). The smaller barrier is controlled by the hydrophobic effects. When the plane of the Arg’44’ guanidinium group is out of the plane of the bases, the steric hindrance tends to vanish towards one direction. Although the CpG dinucleotides have similar hydrogen bonding donor-acceptor pattern to the mCpG pairs, the MBD proteins do not recognize them because of the missing steric hindrance and hydrophobic interactions.

Figure 4. Distance between the guanidinium carbon atoms of Arg’22’ and Arg’44’ (in locked and open/closed states) and the methyl carbon atoms of the 5-methylcytosine residues with respect to the rigid helical rotation of the amino acid side chains of the MeCP2 model along the axis of the DNA double helix.

Conclusion

The constrained geometry optimizations of the mCyt-Arg-Gua triad models revealed the flexibility of the interacting structure, the possible interaction modes in the recognition and the effect of the methylation on the interactions. (For the residues in various proteins, we use the human MBD1 numbering as reference mapped on the actual protein.) The geometries were optimized by the M06L method, which was empirically parametrized to include mid-range correlation effects. In the initial structures taken from experiment only the guanidinium group of the Arg’22’ forms a strong hydrogen
bridge (shorter than 3 Å) with a neighboring guanine(n+1) while the guanidinium group of the Arg’44’ does not. During the next stage of the geometry optimization we allowed to change the positions of the arginine side chain atoms in the rigid DNA and protein environment. This results in the formation of strong hydrogen bonds in both triads. The plane of the guanidinium group of the Arg’22’ and the guanine becomes coplanar. The Arg’44’ guanidinium group is initially farther from the coplanar orientation, and the fully coplanar orientation requires some flexibility from the protein environment. The effect of methylation is small on the orientation of the guanidinium group and marginal on the length of the hydrogen bonds. The cation-pi interactions require the flexibility of the DNA double helix, and the methylation has a small effect on the Arg’22’ binding site and a moderate effect on the Arg’44’ binding site assuming a flexible protein environment. The effect of dispersion interactions between the arginine side chain and the methyl group of the methylcytosine are more important on the Arg’44’ binding site than on the Arg’22’ binding site as the Cβ(Arg)-C(mCyt) distances indicate. The strength of these interactions increases on each geometry optimization level.

The electronic interaction energy in gas-phase originates mostly from the hydrogen bonding as it changes largely on the second geometry optimization level, during the strengthening of the hydrogen bonds. The cation-pi and dispersion interactions have smaller contributions to the electronic interaction energy according to the third geometry optimization level, in which mostly the cation-pi interactions strengthen; and according to the fourth geometry optimization level, in which mostly the dispersion interactions strengthen. The electronic interaction energy per binding site is -42-43 kcal mol⁻¹ calculated with the RI-MP2 and M06L methods on the optimized geometries from the final level. The methyl group deepens the electronic interaction energy by ~1 kcal mol⁻¹ on the Arg’22’ binding site and by ~2 kcal mol⁻¹ on the Arg’44’ binding site. The thermochemical analysis shows the hydrophobic interactions are crucial in the recognition. Beside different number of explicit water molecules, the highly accurate dRPA75 and dRPA@PBE0 electronic energy corrected B3LYP interaction energies and enthalpies are highly positive. The overall Gibbs energy is about -42 kcal mol⁻¹ for the mDNA-MBD model complex. The applied COSMO solvation model lowers the Gibbs interaction energy to -83-84 kcal mol⁻¹ and -87-88 kcal mol⁻¹ in the weakly and strongly hydrated cases. In the optimized model complex, the rigid helical rotation of the amino acid side chains around the axis of the DNA double helix shows that the Arg’22’ guanidinium group forms a larger steric hindrance for sliding towards one direction (locked state). The Arg’44’ guanidinium group forms a smaller steric hindrance for sliding towards the other direction; however, it rotates towards the Asp’32’ side chain without the presence of the hydrophobic interaction between its alkyl chain and the methyl group of the neighboring methylcytosine, which controls the orientation of the guanidinium group (open-close mechanism). The steric hindrance of the arginine side chains for the protein for sliding on the DNA double helix agrees with the earlier suggested model of hopping to search the target site. The open-close mechanism for the Arg’44’ side chain corresponds to the experimental observation that the Arg’44’ side chain is more flexible than the Arg’22’ side chain.

The developed model gives a possible explanation for the experimentally observed methyl DNA recognition process by MBD proteins, and it will hopefully prove to be successful to predict more realistic thermodynamics and kinetics of epigenetic recognition processes, which can be useful for designing new therapeutic anti-cancer drugs targeting directly the interaction between the MBD proteins and the promoter region of the silenced tumor suppressor genes avoiding the chromosome instability caused by demethylating drugs.