Density functionals applied to biomolecules and their non-covalent interactions

Author: Pál Dániel Mezei
Supervisor: Prof. Gábor István Csonka

Ph. D. thesis
2016
Contents

Abbreviations .................................................................................................................. 3

1. Introduction .................................................................................................................. 5

2. Background .................................................................................................................. 6
   2.1. O-glycosidic linkage in glycoproteins .................................................................... 6
   2.2. Methyl-CpG-binding domain proteins ................................................................... 9
   2.3. Difficult test sets for density functional approximations ......................................... 12

3. Computational methods ............................................................................................. 16
   3.1. Wave function methods ....................................................................................... 16
   3.2. Density functional methods .................................................................................. 18
   3.3. Thermochemical model ....................................................................................... 30

4. Benchmarking density functionals ............................................................................. 31
   4.1. On intermolecular anion-pi interactions [S1] ....................................................... 31
   4.2. On intramolecular interactions in Diels-Alder reactions [S2] ................................. 38

5. Method development ................................................................................................... 43
   5.1. New complete basis set extrapolation for direct random phase approximation [S3] 43
   5.2. New dual-hybrid direct random phase approximation method [S4] ....................... 49

6. Biological applications ............................................................................................... 56
   6.1. Intramolecular interactions in immune recognition [S5] ........................................ 56
   6.2. Intermolecular interactions in epigenetic recognition [S6] ...................................... 63

7. Summary ..................................................................................................................... 70

Related publications ....................................................................................................... 72
Acknowledgement ............................................................................................................ 73
References ....................................................................................................................... 74
Appendix .......................................................................................................................... 79
   A. Main concepts in density functional theory ............................................................. 79
   B. Small test sets for method development ................................................................. 83
   C. Structure and energetics of the O-glycopeptide models ........................................ 84
Abbreviations

**Deviations, errors**
- MD: mean deviation
- MAD: mean absolute deviation
- CSSD: corrected sample standard deviation
- SIE: self-interaction error

**Basis sets**
- AXZ: aug-cc-pVXZ
- aTZ(-f,-d): (aug)-cc-pVTZ(-f,-d)
- def2QZ: def2-QZVP

**Methods**
- MM: molecular mechanics
- MD-tar: molecular dynamics with time-averaged restraints
- HF: Hartree-Fock
- MBPT: many-body perturbation theory
- PT2: second-order perturbation theory
- MP2, RI-MP2: Møller-Plesset second-order perturbation theory without or with density fitting
- SAPT: symmetry-adapted perturbation theory
- CC: coupled cluster
- CEPA: coupled electron pair approximation
- LPNO, DLPNO: local pair natural orbital, domain-based local pair natural orbital
- LDA: local density approximation
- GGA: generalized gradient approximation
- MGGA: meta-generalized gradient approximation
- HGGA: hyper generalized gradient approximation
- RPA, RPAX: random phase approximation without or with exchange kernel
- dRPA: direct RPA
- dRPA75: dual hybrid dRPA with 75% of exact exchange
- SOSEX: second-order screened exchange
- PBE: Perdew-Burke-Ernzerhof
- PBE0: PBE hybrid with 25% of exact exchange
- revPBE: revised PBE
- rPW86: revised Perdew-Wang 1986
- TPSS: Tao-Perdew-Staroverov-Scuseria
- TPSSh: TPSS hybrid with 10% of exact exchange
- TPSS0: TPSS hybrid with 25% of exact exchange
- B3LYP: Becke’s 3-parameter hybrid with Lee-Yang-Parr correlation
- M06L: Minnesota 2006 semi-local
- M06: Minnesota 2006 hybrid
- M06-2X: Minnesota 2006 hybrid with double amount of exact exchange
- oB97X-D: Becke 1997 range-separated hybrid with dispersion
- mPW2PLYP: double hybrid with Perdew-Wang 1991 exchange and Lee-Yang-Parr correlation
- VV10: non-local dispersion correction of Vydrov and van Voorhis
- CBS: complete basis set
- CP: counterpoise
- CAE: corrected atomic energies
- XRD: X-ray diffraction
- NMR: nuclear magnetic resonance
- CD: circular dichroism

**Test sets**
- sHC5: 5 reactions of small hydrocarbons
- BH6: 6 barrier heights
- RCn: n-homodesmotic reaction class
- DBH24: 24 diverse barrier heights
- GMTKN30: 30 test sets in general main-group thermochemistry, kinetics and non-covalent interactions
- ACONF: alkane conformers
- CYCONF: cysteine conformers
- PCONF: tripeptide conformers
- SCONF: sugar conformers
- NCCE31: 31 non-covalent complexation energies

**Compounds**
- DNA: deoxyribonucleic acid
- dsDNA: double-stranded DNA
- RNA: ribonucleic acid
- miRNA: micro RNA
- Gal: galactose
- Man: mannose
- GalNAc: 2-acetylamino-2-deoxy-D-galactose
- ManNAc: 2-acetylamino-2-deoxy-D-mannose
- CpG: 5′-cytidine-phosphate-guanosine-3′
- mCpG: methyl-CpG
- MBD: mCpG-binding domain
- MBP: mCpG-binding protein
- HDAC: histone deacetylase
- DNMT: DNA methyl transferase
- HFB: hexafluorobenzene
- TFB: 1,3,5-trifluorobenzene
- TFZ: 2,4,6-trifluoro-1,3,5-triazine
- TAZ: 1,3,5-triazine

**Conformations**
- PPII: polyproline II
- ac: anticlinal
- ap: antiperiplanar
- sc: synclinal
- sp: synperiplanar
1. Introduction

My doctoral research is following the earlier steps of my supervisor in the direction of accurate, precise and efficient modelling of biologically important molecular interactions by high-level wave function and density functional methods. This topic is significant because the immune recognition can be explained by binding of special glycoproteins. The question is how the binding of the sugar antenna on the protein surface influences the protein conformation? In this thesis, we broaden our research to another biologically very important recognition process in which the methyl-DNA-binding proteins attach to a methylated DNA sequence, and thus they regulate the gene expression. This topic is significant because the alteration of the DNA methylation pattern plays a central role in the initiation of cancer and other epigenetic diseases. The high-level wave function and density functional methods provide deep insight into the mechanisms and interactions which guide these recognition processes.

Firstly, we need a sufficiently accurate and efficient theoretical method for the larger, biologically important systems. This requires smaller model systems with highly accurate and computationally very expensive reference energies and equilibrium geometries. We begin to test much less expensive methods against these references, then select the appropriate methods for computing the molecular interactions. Frequently, we cannot find a sufficiently accurate and efficient method. In these cases, we develop a new methodology. We also develop and modify test sets if it is necessary. We apply the best methods for the computation of the conformational space and potential energy surface of O-glycosylated glycopeptide model structures, and for the analysis of the structure and energetics of the DNA-protein interaction surface in methyl-DNA-binding protein – methyl-DNA complex model structures. Finally, we analyze how O-glycosylation affects the protein structure in immune recognition processes, and how the methyl-DNA recognition works in epigenetic processes.
2. Background

2.1. O-glycosidic linkage in glycoproteins

Glycosylation is among the most frequent post-translational modifications,\(^1\)\(^2\)\(^3\) 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 three-dimensional structure\(^8\)\(^9\)\(^10\) and biological role \(^{11}\)\(^{12}\) of glycoproteins and the synthetic approaches\(^{13}\)\(^{14}\)\(^{15}\)\(^{16}\)\(^{17}\)\(^{18}\) 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-D-galactose (GalNAc) is α-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 \textit{in vivo} profiling of the changes in the O-linked glycosylation.\(^{29}\) Furthermore, glycopeptides are generally applied as antibiotics.\(^{30}\)\(^{31}\)\(^{32}\)\(^{33}\)\(^{34}\)

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 \(^{13}\)C NMR spectroscopy\(^{38}\) 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 modelling 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. The measured $^3J_{HN-H\alpha}$ coupling constants in the internal Thr2 and Thr3 residues limited the associated $\phi$ torsion angles to the region around $-120^\circ$. The nuclear Overhauser effect (NOE) and coupling constant restrained molecular dynamics structural calculations resulted in $\beta$-sheet structure, and the amide proton chemical shift trends independently confirmed the $\beta$-character.

$\phi$: $C_1\text{Ser}-N_1\text{Ser}-C_\alpha-C_2\text{Ser}$

$\psi$: $N_1\text{Ser}-C_\alpha-C_2\text{Ser}-N_2\text{Ser}$

$\chi_1$: $C_1\text{Ser}-N_1\text{Ser}-C_\alpha-C_\beta$

$\chi_2$: $N_1\text{Ser}-C_\alpha-C_\beta-O_1\text{Carb}$

$\chi_3$: $C_\alpha-C_\beta-O_1\text{Carb}-C_1\text{Carb}$

$\tau_2$: $C_1\text{Carb}-C_2\text{Carb}-N_2\text{Carb}-C_7\text{Carb}$

$\tau_3$: $C_4\text{Carb}-C_3\text{Carb}-O_3\text{Carb}-H$

$\tau_4$: $C_5\text{Carb}-C_4\text{Carb}-O_4\text{Carb}-H$

$\tau_5$: $C_4\text{Carb}-C_5\text{Carb}-C_6\text{Carb}-O_6\text{Carb}$

$\tau_6$: $C_5\text{Carb}-C_6\text{Carb}-O_6\text{Carb}-H$

**Figure 1.** Numbering and definition of torsion angles in $\alpha$-GalNAc-Ser. The two subunits are grouped by orange dashed lines; the most important hydrogen bonds are indicated by green dotted lines.

The glycosidic linkage was studied theoretically in vacuum using MM2* and HF/6-31G(d) optimized $O$-(2-acetamino-2-deoxy-$\alpha$- or $\beta$-D-galacto- or -mannopyranosyl)-$N'$-acetyl-$N$-methyl-L-serinamide model structures (**Figure 1**). The $\alpha$- and $\beta$-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 χ₃ and χ₂ torsion angles of the linkage (see Fig. 3b in ref 46). 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)⁴⁷ 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.⁴⁸ The MD-tar studies also showed that β-O-Glycosylation on L-Ser or -Thr increases the probability for the occurrence of this motif in α helical structure.⁴⁹ Similar observations were made for α- or β-GlcNAc-Ser or -Thr from MD-tar simulations.⁵⁰ (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.⁵¹)

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.⁵²

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 were 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 0-glycosidic linkage.53

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.54 (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.55)

2.2. Methyl-CpG-binding domain proteins

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.56 57 58 59 The epigenetic information is carried by the DNA accessibility, DNA methylation and histone modifications, and can be mapped by high-throughput molecular assays.60 61 62 63 64 The DNA methylation plays a central role in embryonic development, cell differentiation, genomic imprinting, X-chromosome inactivation and neoplastic transformation.65 66 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).57 68 In mammalian genomes, ~70% of the CpG sites are methylated.69 Generally, genes can be transcribed from methylation-free promoters, while the noncoding DNA background is permanently silenced.70

The methylated CpG sites (mCpG) are specifically recognized by mCpG binding proteins (MBP), which recruit repressive chromatin modifiers and remodeling complexes.71 72 The methyl-CpG-binding domain (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.73 Furthermore, mutations in the epigenetic marker machinery can also lead to carcinogenesis or Rett syndrome.74 75 76 In cancer, there is a global decrease of CpG methylation, but the promoters of the tumor suppressor genes become hypermethylated.72 77 78 Inter alia the p16INK4a, hMLH1 and BRCA1 genes are silenced in many types of cancer.79 80 Tumors can be classified according to their methylation profile.81 82 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.83 84 85
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. The solution structure of the MBD1-mDNA, 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). 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 Arg22, Asp32, Tyr34 and Arg44. The Arg22 and Arg44 side chains participate in the recognition of the 5-methylcytosine forming two 5mCyt-Arg-Gua triads. The Arg44 side chain shows larger flexibility than the Arg22 side chain. The Arg22 side chain is strongly fixed either by two hydrogen bonds, or by a shorter salt bridge towards Asp32. The Arg44 side chain is weakly fixed either by a single hydrogen bond or by a longer salt bridge towards Glu/Gln/Ser48.

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 2.** (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. For 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(2d{0.8,0.2},p) interaction energy was -41.9 kcal mol\(^{-1}\). (In their modified basis set, the exponents of the \(d\)-type polarization functions are denoted in curly brackets.) Due to the non-optimal 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(d{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\(^2\), which leads to -1.5 kcal mol\(^{-1}\) stabilization due to the presence of the methyl group. Similarly, the free energy perturbation
calculations show -1.2±0.1 kcal mol\(^{-1}\) stabilization by the presence of the methyl group. 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-\(\pi\) interaction. Furthermore, structurally conserved water molecules were suggested to determine further mCpG-binding specificity of MeCP2\(^{105}\). The W1 and W2 water molecules form hydrogen bonds with the amino group of methyl cytosines, as well as with the Tyr34, Arg44 and Asp32 amino acids bridging the DNA and the protein as shown in Figure 2.

2.3. Difficult test sets for density functional approximations

**Anion-\(\pi\) complexes**

The importance of cation–\(\pi\) interactions was recognized some time ago\(^{106,107,108}\), and the accurate description of the preferred arrangements is quite feasible. The analogous anion–\(\pi\) complexes were ignored for a long time because of the assumed electrostatic repulsion\(^{109}\). However, this electrostatic repulsion can be abolished or even turned to electrostatic attraction by electron withdrawing groups\(^{110}\). This relatively unexplored interaction has recently gained attention\(^{111,112}\). Compared to the cation-\(\pi\) interactions, the anion–\(\pi\) complex interaction energy is supposed to have more significant dispersion and induction components because anions have more extended and polarizable electron clouds. Recently direct evidence for the anion-\(\pi\) interactions has been obtained by tandem mass spectrometric experiments with naphthalene-diimide models where only the \(\pi\)-acidic surface is left for anions to interact with\(^{113}\).

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\(^{114,115}\). For complexes with the fluoride ion (e.g. s-triazine ··· F\(^{-}\)), the covalently bonded “Meisenheimer” complex is the most stable in gas phase\(^{116}\). However, adding two water molecules to this complex makes the H-bonded complex more stable, competing with anion–\(\pi\) and displaced anion–\(\pi\) complex forms\(^{116}\). Adding three or four water or acetonitrile molecules makes the displaced anion–\(\pi\) complex form the most stable. Similarly, adding three or four water molecules to the s-triazine ··· Cl\(^{-}\) complex makes the anion–\(\pi\) complex form the most stable\(^{116}\). Since the anion–\(\pi\) and displaced anion–\(\pi\) forms are almost equally stable, these two forms occur in many crystal structures. A survey showed\(^{117}\) that in contrast to the expectations, anion-\(\pi\) interactions are more frequent in the Cambridge Structural Database\(^{118}\) than cation-\(\pi\) interactions.
Figure 3. Structure and the main symmetry axis of the binary anion-π complexes (1-20), and ternary π-anion-π’ complexes (21-50).

A test set of 20 binary anion-π and 30 ternary π-anion-π’ complexes (Figure 3) were proposed in the literature for benchmarking semi-local density functionals. The complexes are constituted by fluoride, chloride, bromide, nitrate or carbonate ions, and by hexafluorobenzene (HFB), 1,3,5-trifluorobenzene (TFB), 2,4,6-trifluoro-1,3,5-triazine (TFZ) or 1,3,5-triazine (TAZ). The reference energies were computed with the MP2 method using small 6-31++G(d,p) basis set and counterpoise correction to overcome the basis set superposition error. However, the basis set superposition error often hastens the convergence, and the counterpoise correction usually overcorrects and results in larger errors than the original basis set error. Furthermore, the MP2 method has a truncated perturbative treatment of the electron correlation, hence it is unable to capture the non-pairwise nature of dispersion interaction (i.e. the third- and higher-order many-body effects).

Notice that calculating anions with common semi-local density functionals requires caution, as a fraction of the electrons leaks to infinity when sufficiently large basis sets are employed. The origin of this is the self-interaction error which leads to incorrect exponential decay of the exchange potential in the asymptotic region (instead of the exact $-1/r$). The effect of the self-interaction error is particularly large for anions and leads to positive HOMO energies. Using a finite basis set stabilizes this unbound state artificially. In the
presence of cations, fractional charge transfer occurs even with finite basis sets, due to the many-electron self-interaction error.\textsuperscript{125} Despite these errors, the approximate density functional methods give useful anion energies and electron affinities.\textsuperscript{126}

\textit{Diels-Alder reactions}

The inexpensive local and semi-local functionals suffer from self-interaction error (SIE).\textsuperscript{127,128,129} This leads to an energetic preference for unrealistically delocalized electron densities. Functionals having SIE show particularly large errors for systems with fractional charges,\textsuperscript{125,130} as too-delocalized charge distribution leads to too-low energies. For the same reason, the SIE overestabilizes the charge transfer complexes, leads to too-low or no reaction barrier heights,\textsuperscript{131} and to qualitatively wrong dissociation energy curves for diatomic cations (e.g. H\textsubscript{2}+). The correct description requires the ground-state total energy to change linearly between two neighboring integer electron numbers.\textsuperscript{132}

All standard local and semi-local functionals give convex electronic energy curves and show no derivative discontinuity at integer electron numbers.\textsuperscript{133,134} Exact exchange has the opposite behavior: it gives concave energy curves between two neighboring integers, due to the missing electron correlation. Mixing the convex and concave energy curves via hybridization of the functionals might help; however, in practice the popular hybrid functionals still show considerable convexity error. These hybrids can be classified as global, local or range-separated. The standard global hybrid functionals with 20-25\% of exact exchange correct only a fraction of the delocalization error and fail seriously for delocalized stretched odd-electron systems. A large fraction (\~{}70\%) of exact exchange in the global hybrid PBE form can eliminate the consequences of the many-electron SIE.\textsuperscript{135} The range-separated MCY3 and rCAM-B3LYP functionals also show improvement in the description of fractional numbers of electrons.\textsuperscript{133}

To quantify the SIE or delocalization error, Johnson et al. proposed\textsuperscript{133} a test set of fourteen representative Diels-Alder reactions (DARC test set). They supposed that there is a highly localized electron density region in the products which is understabilized by most of the semi-local functionals. 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 the SIE.

In the DARC test set (\textit{Table 1}), the diene reactants are butadiene, cyclopentadiene, cyclohexadiene, and furane; the dienophile reactants are ethene, ethyne, maleic anhydride, and maleimide. During the reactions, mono-, bi- and tricyclic products are formed with non-covalent overlapping electron density regions. The reference reaction energies were computed
by the CCSD(T)/CBS method. The DARC test set was included in a larger database called GMTKN30.\textsuperscript{136,137} According to this database, the hybrid and double hybrid functionals provide an improved description of these reaction energies.

\textit{Table 1.} The reagents, products and reference reaction energies in the DARC test set.\textsuperscript{133} The endo or exo orientations of the reagents are noted in parentheses.

<table>
<thead>
<tr>
<th>No</th>
<th>Reagents</th>
<th>Products</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ethene + butadiene</td>
<td>monocyclic</td>
<td>-43.8</td>
</tr>
<tr>
<td>2</td>
<td>ethyne + butadiene</td>
<td></td>
<td>-59.3</td>
</tr>
<tr>
<td>3</td>
<td>ethene + cyclopentadiene</td>
<td></td>
<td>-30.0</td>
</tr>
<tr>
<td>4</td>
<td>ethyne + cyclopentadiene</td>
<td>bicyclic</td>
<td>-33.1</td>
</tr>
<tr>
<td>5</td>
<td>ethene + cyclohexadiene</td>
<td></td>
<td>-36.5</td>
</tr>
<tr>
<td>6</td>
<td>ethyne + cyclohexadiene</td>
<td></td>
<td>-48.2</td>
</tr>
<tr>
<td>7</td>
<td>furan + maleic anhydride (endo)</td>
<td></td>
<td>-14.4</td>
</tr>
<tr>
<td>8</td>
<td>furan + maleic anhydride (exo)</td>
<td></td>
<td>-16.2</td>
</tr>
<tr>
<td>9</td>
<td>furan + maleimide (endo)</td>
<td></td>
<td>-17.2</td>
</tr>
<tr>
<td>10</td>
<td>furan + maleimide (exo)</td>
<td>tricyclic</td>
<td>-19.2</td>
</tr>
<tr>
<td>11</td>
<td>cyclopentadiene + maleic anhydride (endo)</td>
<td></td>
<td>-31.6</td>
</tr>
<tr>
<td>12</td>
<td>cyclopentadiene + maleic anhydride (exo)</td>
<td></td>
<td>-32.1</td>
</tr>
<tr>
<td>13</td>
<td>cyclopentadiene + maleimide (endo)</td>
<td></td>
<td>-34.1</td>
</tr>
<tr>
<td>14</td>
<td>cyclopentadiene + maleimide (exo)</td>
<td></td>
<td>-34.4</td>
</tr>
</tbody>
</table>
3. Computational methods

3.1. Wave function methods

Hartree-Fock method

The Hartree-Fock (HF) method is a variational method based on the one-electron approximation. This method averages the interactions among the electrons. The HF exchange energy ($E_{X}^{HF}$) can be calculated from the $\varphi_i$ occupied orbitals with the $\hat{v}_{ee}$ electron-electron Coulomb operator in a two-electron integral form.

$$E_{X}^{HF} = -\frac{1}{2} \sum_{i,j}^{occ} \langle \varphi_i | \hat{v}_{ee} | \varphi_j \rangle \varphi_i \varphi_j$$  \hspace{1cm} (1)

The Coulomb correlation is defined by the difference between the exact solution of the non-relativistic Schrödinger equation within the Born-Oppenheimer approximation and the HF energy. The missing Coulomb correlation spoils the electronic energy calculation, and makes the binding energies to be seriously underestimated. The self-consistent HF electron densities are too compact, the optimized HF bond lengths are too short. This method has a nominal $O(N^4)$-scaling with the system size.

Coupled pair and coupled cluster methods

The configuration interaction (CI) method superimposes the different electronic states in the wave function $\Psi$. The ground-state determinant $\Phi_0$ and energy $E_0$ is computed using the HF method. The correlation energy can be calculated according to the variational principle. The CI method with all the possible excitations (full-CI) gives the exact solution. In practice, only the CISD method with the singly $\Phi_i^a$ and doubly $\Phi_{ij}^{ab}$ excited states can be applied only for very small systems. Furthermore, the CISD method is not size-extensive. The coupled electron-pair approximation (CEPA) is a size-extensive version of the CI method with single and double excitations (CISD). (The CEPA1 correlation energy, $\hat{H}$ is the Hamiltonian, $\epsilon_{ij}$ are the pair correlation energies.)

$$E_{C}^{CEPA1} = \langle \Phi_0 | \hat{H} - E_0 | \Psi \rangle$$ \hspace{1cm} (2)

$$\langle \Phi_i^a | \hat{H} - E_0 - \sum_k \epsilon_{ik} | \Psi \rangle = 0$$ \hspace{1cm} (3)

$$\langle \Phi_{ij}^{ab} | \hat{H} - E_0 - \frac{1}{2} \sum_k (\epsilon_{ik} + \epsilon_{jk}) | \Psi \rangle = 0$$ \hspace{1cm} (4)
The computations can be accelerated preserving 99.8% of the correlation energy by the local-pair natural orbital (LPNO) method. This enables the calculation of medium-sized systems with at most 2000 basis functions and 100 atoms.

The coupled-cluster (CC) method uses the excitation operator $\hat{T} = \hat{T}_1 + \hat{T}_2 + \cdots$ in an exponential form. This method with all the possible excitations gives the full-CI solution. In practice, the exponential operator is expanded in Taylor series, and the expansion is truncated to only several excitations. The CCSD method contains the single and double excitation operators, and three equations with the $\Phi_0$, $\Phi_i^a$ and $\Phi_{ij}^{ab}$ determinants. The CCSDT method contains also the triple excitation operators, and an additional equation with the $\Phi_{ijk}^{abc}$ determinant. ($E_C^{CCSD..}$ is coupled cluster correlation energy for arbitrary order.)

$$E_C^{CCSD..} = \langle \Phi_0 | e^{-\hat{T}} (\hat{H} - E_0) e^{\hat{T}} | \Phi_0 \rangle$$

$$\langle \Phi_i^a | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Phi_0 \rangle = 0$$

$$\langle \Phi_{ij}^{ab} | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Phi_0 \rangle = 0$$

$$\langle \Phi_{ijk}^{abc} | e^{-\hat{T}} \hat{H} e^{\hat{T}} | \Phi_0 \rangle = 0$$

The CCSD(T) method is a CCSD method with triple excitations approximated non-iteratively from the perturbation theory. Similarly, the CCSDT(Q) method is a CCSDT method with perturbative quadruple excitations. The CCSD(T) method proved to be a gold standard in main-group thermochemistry. This method originally scales iteratively as $O(N^6)$ with one $O(N^7)$-scaling non-iterative step. However, the computations can be accelerated without any significant loss of accuracy in the relative energies by the domain-based local-pair natural orbital (DLPNO) method. The resulted DLPNO-CCSD(T) method has a linear scaling with the system size. Very recently, calculations were reported also with more than 20 000 basis functions and 1000 atoms.

Perturbative methods

The many-body perturbation theory (MBPT) methods are not variational methods, hence the correlation energy of a finite-order MBPT can be lower than the exact solution. In fact, the results usually get better up to sixth-order but oscillate around the exact solution afterwards. The Møller-Plesset second-order perturbation theory (MP2) correlation energy ($E_C^{MP2}$) can be calculated from the $\varphi$ occupied and virtual orbitals and corresponding $\epsilon$ orbital energies, and usually recovers 80-90% of the correlation energy. However, this method only considers the pairwise interactions and diverges for small HOMO-LUMO gap systems. The
original MP2 method scales as \( O(N^5) \), but \( O(N^4) \)-scaling can be achieved with density fitting (RI-MP2).

\[
E_{\text{C}}^{\text{MP2}} = \frac{1}{4} \sum_{\text{occ}} \sum_{\text{virt}} \left| \langle \varphi_i | \varphi_j \rangle \hat{V}_{ee} | \varphi_a \varphi_b \rangle \right|^2 \left( \varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b \right) \quad (9)
\]

The symmetry-adapted perturbation theory (SAPT) provides a direct treatment of the interaction energy between two molecules without computing the energies of the dimer and the monomers.\(^{141}\) The Hamiltonian \( \hat{H} \) of the dimer can be partitioned into contributions from the monomers and the interaction.

\[
\hat{H} = \hat{F}_A + W_A + \hat{F}_B + W_B + V \quad (10)
\]

\( \hat{F} \) and \( W \) are the Fock operator and fluctuation potential of A and B molecules, and \( V \) is the interaction potential. According to this theory, the interaction energy can be decomposed into physically meaningful terms. The simplest truncation of the perturbative expansion is SAPT0, in which the method error is balanced by a truncated aug-cc-pVDZ basis set (denoted by jun-cc-pVDZ).\(^{142,143}\) \( E_{\text{SAPT0}} \) is the SAPT0 energy.

\[
E_{\text{SAPT0}} = E^{(1,0)}_{\text{elst}} + E^{(1,0)}_{\text{exch}} + \left[ E^{(2,0)}_{\text{ind},\text{resp}} + E^{(2,0)}_{\text{exch-ind},\text{resp}} \right] + \left[ E^{(2,0)}_{\text{disp}} + E^{(2,0)}_{\text{exch-disp}} \right] \quad (11)
\]

The first and second number in the superscripts means the order in \( V \) and \( W \), respectively. The resp. tag in the subscripts means the orbital relaxation effects are included. The interaction energy terms are grouped in order with respect to electrostatics, exchange, induction and dispersion. The computations are greatly accelerated by density fitting beside introducing less than 0.01 kcal mol\(^{-1} \) error for small dimers.\(^{144,145}\)

### 3.2. Density functional methods

The main concepts of density functional theory are summarized in Appendix A.

**Local density approximation**

The local density approximation (LDA) was introduced by Kohn and Sham.\(^{146}\) The form of the LDA exchange functional \( E^{\text{LDA}}_X \) can be easily derived from coordinate-scaling the uniform electron gas. The exact exchange energy density per electron \( \varepsilon_{\text{unif}}^X \) of the uniform electron gas depends on the 1/3 power of the electron density \( n \), and thus on the Fermi wave vector \( k_F \).

\[
E^{\text{LDA}}_X[n(r)] \approx \int n(r) \varepsilon_{\text{unif}}^X(n) d^3r \quad (12)
\]
\[\varepsilon^\text{unif}_X(n) = -\frac{3k_F}{4\pi}\]  
(13)

\[k_F = \left(3\pi^2n\right)^\frac{1}{3}\]  
(14)

The spin-polarized exchange energy can be easily derived from the spin-scaling (with the spin densities \(n_\sigma\)). This expression is generally true for all the exchange density functional approximations. Hence we will discuss only the unpolarized exchange functionals \(E^\text{unpol}_X\) in the following sections.

\[E_X[n_\alpha, n_\beta] = \frac{E^\text{unpol}_X[2n_\alpha] + E^\text{unpol}_X[2n_\beta]}{2}\]  
(15)

The exact correlation energy density per electron \(\varepsilon^\text{unif}_C\) of the uniform electron gas depends on the Wigner-Seitz radius \(r_S\), and on the relative spin-polarization \(\zeta\). \(E^\text{LDA}_C\) is the LDA correlation energy.

\[E^\text{LDA}_C[n(r)] \equiv \int n(r)\varepsilon^\text{unif}_C(r_S, \zeta)d^3r\]  
(16)

\[r_S = \left(\frac{3}{4\pi n}\right)^\frac{1}{3}\]  
(17)

\[\zeta = \frac{n_\alpha - n_\beta}{n_\alpha + n_\beta}\]  
(18)

It cannot be written in a closed form; however, analytic representations were fitted to the Green’s function Monte Carlo data of Ceperley and Alder using the Padé interpolation.\(^{147\ 148\ 149}\) These representations also satisfy the small-\(\zeta\), the high-density, and the low-density expansions of the correlation energy per electron.\(^{150\ 151\ 152}\) \((c_0, c_1, c_2, c_3, d_0, \text{and } d_1\) are functions of \(\zeta, \alpha_{\mathcal{C}}\) is the spin stiffness.)

\[\lim_{\zeta \to 0} \varepsilon_C(r_S, \zeta) = \varepsilon_C(r_S, 0) + \frac{1}{2}\alpha_{\mathcal{C}}(r_S)\zeta^2 + \cdots\]  
(19)

\[\lim_{r_S \to 0} \varepsilon_C(r_S, \zeta) = c_0(\zeta) \cdot \ln(r_S) - c_1(\zeta) + c_2(\zeta)r_S \cdot \ln(r_S) - c_3(\zeta)r_S + O(r_S^2)\]  
(20)

\[\lim_{r_S \to \infty} \varepsilon_C(r_S, \zeta) = \frac{d_0(\zeta)}{r_S} + \frac{d_1(\zeta)}{r_S^{3/2}} + O(r_S^{-2})\]  
(21)

The LDA underestimates the exchange and seriously overestimates the correlation, which leads to the overall underestimation of the total energies because the exchange part is much bigger than the correlation part. Hence the binding energies are significantly
overestimated; the bond lengths are slightly underestimated. This method works well for metallic densities but underestimates the HOMO-LUMO gap of semi-conductors. It cannot describe the dispersion interactions because they originate from the correlated density oscillations and not from the density differences. The LDA correlation potential incorrectly vanishes exponentially with the distance.

**Semi-local functionals**

The semi-local functionals take into account some non-locality with the exchange and correlation enhancement factors. The performance of such functionals relies on the error cancellation between the exchange and correlation parts. The semi-local functionals cannot fully describe the dispersion interactions because of their non-local nature. Furthermore, they are loaded with self-interaction error, which results in too diffuse electron densities, too long bond lengths and too low barrier heights. The generalized gradient approximation (GGA) exchange enhancement factor $F_X$ depends on the reduced gradient ($s$).

$$
E_{X}^{GGA}[n(r)] \cong \int n(r)\epsilon_X^{unif}[n](r)F_X^{GGA}(s(r))d^3r
$$

$$
s = \frac{\left|\nabla n\right|}{2k_F n}
$$

The GGA exchange enhancement factor should follow the second-order gradient expansion with the gradient coefficient $\mu_{GE} = 10/81$.

$$
F_X \rightarrow 1 + \mu_{GE} s^2
$$

Since the exact exchange density is not measurable, it is gauge or unitarily variant. It was assumed that a bounded exact exchange density can be constructed by coordinate transformation from the conventional exact exchange density. A lower bound for the exchange-correlation energy (since it is negative) was introduced by Lieb and Oxford (with $C_{LO} = 1.68$). This bound also holds for the exchange energy alone because the correlation energy is always lower or equal to zero.

$$
E_{XC} \geq -C_{LO} \int n^{4/3}(r)d^3r
$$

The Lieb-Oxford bound can be globally satisfied, if the exchange enhancement factor is $F_X(\zeta = 1) < 2.273$ for the fully spin-polarized case.

$$
F_X(\zeta = 1, s) = 2^{1/3}F_X(2^{-1/3}s)
$$

Recalling the definition of the fully spin-polarized enhancement factor, the exchange enhancement factor is $F_X(\zeta = 0) < 1.804$ for the unpolarized case.
\[ F_{X}^{PBE}(s) = 1 + \kappa \left( 1 - \frac{1}{1 + \frac{\mu s^2}{\kappa}} \right) \] 

(27)

The PBE exchange enhancement factor satisfies the Lieb-Oxford bound with \( \kappa = 0.804 \). However, it uses \( \mu = 0.235 \), which is larger than \( \mu_{\text{GE}} \), but works better for molecules due to the increased non-locality.

The non-local part of the GGA correlation energy density per electron \((H)\) depends on \( r_S \) and \( \zeta \) like the LDA correlation energy density per electron. In addition, it depends also on the reduced gradient \((t)\).

\[
E_{c}^{\text{GGA}}[n(r)] \approx \int n(r)\left[\epsilon_{c}^{\text{unif}}[r_S, \zeta](r) + H[r_S, \zeta, t](r)\right]d^3r
\]

(28)

\[
t = \frac{||\nabla n||}{2\phi k_S n}
\]

(29)

The GGA correlation should recover the second-order gradient expansion for the correlation energy at the slowly-varying limit \((H \rightarrow \beta (r_S)\phi^3 t^2)\). Furthermore, it should make the correlation energy vanish at the rapidly varying limit \((H \rightarrow -\epsilon_{c}^{\text{unif}})\), and cancel the logarithmic singularity of \( \epsilon_{c}^{\text{unif}} \).

\[
H^{PBE} = \gamma \phi^3 \ln \left( 1 + \frac{\beta}{\gamma} t^2 + \frac{1 + At^2}{1 + At^2 + A^2 t^4} \right)
\]

(30)

\[
A = \frac{\beta}{\gamma} \left[ \exp \left( \frac{-\epsilon_{c}^{\text{unif}}}{\gamma \phi^3} \right) - 1 \right]^{-1}
\]

(31)

The meta-generalized gradient approximation (MGGA) methods consider also the kinetic energy density \( \tau \) of the Kohn-Sham orbitals \( \varphi_{i\sigma} \). Thus they give a better description of the electronic structure. For example, they provide more accurate atomization energies and bond lengths.
\[
E_X^{MGGA}[n(r)] \equiv \int n(r) \epsilon_X^{unif}[n](r) F_X^{MGGA}(s, \alpha) d^3r
\]

\[
\alpha = \frac{\tau - \tau_W}{\tau_{unif}}
\]

In practice, a dimensionless variable \(\alpha\) is applied, which depends on the kinetic energy density \(\tau\), the von Weizsäcker kinetic energy density \(\tau_W\), and the kinetic energy density of the uniform electron gas \(\tau_{unif}\). This variable can distinguish the single-orbital \(\alpha = 0\), metallic \(\alpha \approx 1\), and non-bonded overlapping \(\alpha \gg 1\) regions.

\[
\tau(r) = \sum_{\sigma} \sum_{i} \frac{1}{2} |\nabla \phi_{i\sigma}(r)|^2
\]

\[
\tau_W(r) = \sum_{\sigma} \frac{1}{8} \frac{|\nabla n_{\sigma}(r)|^2}{n_{\sigma}(r)}
\]

\[
\tau_{unif} = \frac{3}{10} k_F^2 n
\]

The MGGA exchange enhancement factor should follow the fourth-order gradient expansion.

\[
F_X^{GE4} = 1 + \frac{10}{81} s^2 - \frac{1606}{18225} s^4 + \frac{511}{13500} s^2 (1 - \alpha) + \frac{5913}{405000} (1 - \alpha)^2
\]

In fact, the TPSS exchange functional satisfies this constraint.\(^{156}\)

The MGGA correlation energy can be made free from the one-electron self-interaction error. The TPSS correlation functional has this property. Furthermore, the midrange correlation effects can be considered by the MGGA correlation functional. The Minnesota functionals \(e.g.\ M06L^{157}\) do that with many empirical parameters fitted to test sets containing also non-covalent molecular interactions.

**Hybrid functionals**

The hybrid functionals mix the semi-local (sl) exchange with exact (HF) exchange. The dynamic correlation effects are considered by the semi-local correlation. The difference between the hybrid and the pure semi-local exchange contains some static correlation effect. The hybrid methods moderate the self-interaction error of semi-local functionals; therefore, they perform better for barrier heights. However, the error of the static correlation part spoils the energies of the strongly correlated systems.

The global hybrid functionals hybridize the exact \(E_X^{exact}\) and semi-local \(E_X^{sl}\) exchange on level of total energies. The total energy is computed self-consistently.
The global mixing parameter $a$ is $\{0.1, 0.25, 0.25\}$ for TPSSH,\textsuperscript{158} TPSS0 and PBE0,\textsuperscript{159} respectively. Sometimes, we also indicate the mixing factor after the semi-local functional name e.g. TPSS0.1, TPSS0.25 or PBE0.25. The three-parameter global hybrid B3LYP functional contains 20\% of exact exchange.\textsuperscript{160} The M06 and M06-2X functionals are highly empirical functionals from the Minnesota functional family, and contain 27\% and 54\% of exact exchange, respectively.\textsuperscript{157}

The range-separated hybrids mix the exact and semi-local exchange differently in short and long range. The short-ranged and long-ranged Coulomb potentials ($v_{ee}^{sr,\omega}/v_{ee}^{lr,\omega}$) are generally constructed with the error function and complementary error function, and used in the exchange hole model with the $n_{X}^{\text{exact}}(\mathbf{r}, \mathbf{r}')$ exact and $n_{X}^{\text{sl}}(\mathbf{r}, \mathbf{r}')$ semi-local exchange holes instead of the conventional Coulomb potential.

$$
E_{Xc}^{gh} = aE_{X}^{\text{exact}} + (1-a)E_{X}^{\text{sl}} + E_{C}^{\text{sl}}
$$

The ωB97X-D functional contains $a_{sr} = 22.2036\%$ of exact exchange in short range and $a_{lr} = 100\%$ of exact exchange in long range with the range-separation parameter $\omega = 0.2$.\textsuperscript{161}

**Double hybrid methods**

The global double hybrid functionals mix the semi-local exchange with exact exchange, and the semi-local correlation with second-order perturbative correlation (PT2). These methods perform well for barrier heights because of the large fraction of exact exchange. The correlation part also contains some non-locality because of the partial PT2 correlation. But these functionals still miss a large part of the dispersion interactions, hence they also require some dispersion correction. Since the PT2 correlation, as well as the molecular mechanical and non-local corrections contain only pairwise interactions, the double hybrids cannot describe correctly the non-pairwise nature of the dispersion interactions. The computation of the PT2 correlation is also problematic because it diverges for small HOMO-LUMO gap systems.

$$
E_{Xc}^{dth} = a_{X}E_{X}^{\text{exact}} + (1-a_{X})E_{X}^{\text{sl}} + a_{C}E_{C}^{\text{PT2}} + (1-a_{C})E_{C}^{\text{sl}}
$$
The second-order perturbation correlation can be decomposed into singlet (S) and triplet (T) pairs, and the spin components can be scaled to fine tune the functional.

\[
E_C^{PT2} = a_SE_S^{PT2} + a_TE_T^{PT2}
\]

(44)

The parameters are \(\{a_X = 0.53; \ aC = 0.27; \ aS = 1; \ aC = 1\}\) for B2PLYP, \(\{a_X = 0.69; \ aC = 0.54; \ aS = 1; \ aT = 0.8\}\) for DSD-BLYP, \(\{a_X = 0.50; \ aC = 0.269; \ aS = 1; \ aC = 0\}\) for PWPB95.

The spin components are given by the following equations with the pair energies \(e_{ij}\).

\[
E_S^{PT2} = \sum_{iaj\beta} e_{iaj\beta}
\]

(45)

\[
E_T^{PT2} = \frac{1}{2} \sum_{iaj\alpha} e_{iaj\alpha} + \frac{1}{2} \sum_{ipj\beta} e_{ipj\beta}
\]

(46)

The pair energies can be obtained from the amplitudes \(T_{ij}^{ab}\), and two-electron integrals.

\[
e_{iaj\beta} = \sum_{a\beta} T_{iaj\beta}^{a\beta} (i_{a\beta} | a_{\alpha b\beta})
\]

(47)

\[
e_{iaj\alpha} = \sum_{a\beta} \left( T_{iaj\alpha}^{a\beta} - T_{iaj\alpha}^{b\alpha} \right) (i_{a\beta} | a_{\alpha b\alpha})
\]

(48)

\[
e_{ipj\beta} = \sum_{a\beta} \left( T_{ipj\beta}^{a\beta} - T_{ipj\beta}^{b\beta} \right) (i_{p\beta} | a_{\alpha b\beta})
\]

(49)

The amplitudes can be calculated from the spin-orbitals and spin-orbital energies.

\[
T_{iaj\beta}^{a\beta} = \frac{(i_{a\beta} | a_{\alpha b\beta})}{\epsilon_{ia} + \epsilon_{j\beta} - \epsilon_{a\alpha} - \epsilon_{b\beta}}
\]

(50)

The \(i\) and \(j\) indices belong to the occupied, the \(a\) and \(b\) indices belong to the virtual orbitals, \(\alpha\) and \(\beta\) are spin indices.

Dispersion corrections

The dispersion interactions missing from the semi-local functionals can be considered by molecular mechanical and density functional corrections. The atom-pairwise D2 molecular mechanical dispersion correction asymptotically depends on the minus sixth power of the distance \(R_{ij}\) between the atoms \(i\) and \(j\), but incorrectly tends to zero at short range due to a damping function.

\[
E_{disp}^{D2} = -\frac{1}{2} \sum_{i \neq j} s_{ij} f_{damp} (R_{ij}) \frac{C_{ij}^{ij}}{R_{ij}^6}
\]

(51)
The dispersion coefficient $C_{ij}^6$ is the geometric mean of the individual dispersion coefficients. The cutoff radius $R_{0ij}$ is the sum of atomic van der Waals radii. The parameter values are \{s_6 = 0.4; s_r = 1.1; d = 20\} for mPW2PLYP, and \{s_6 = 1; s_r = 1; d = 6\} for ωB97X-D.

The D3 dispersion correction contains terms depending on the minus sixth and eighth power of the interatomic distances. Each term has a damping function similar to the previously introduced damping function. The steepness of the damping is determined by the mean cutoff radius ($R_{0ij}$).

\[
f_{\text{damp}}(R_{ij}) = \frac{1}{1 + e^{-d\left(\frac{R_{ij}}{s_r R_{0ij}} - 1\right)}}
\]

(52)

The universal parameters of the semi-local and hybrid functionals are $s_6 = 1$, $d_6 = 14$, $d_8 = 16$, and $s_{r,8} = 1$. The functional-dependent parameters are $s_8 = \{0.722, 0.928, 0.998, 1.105, 1.219, 1.242\}$, and $s_{r,6} = \{1.217, 1.287, 1.333, 1.166, 1.223, 1.252\}$ for PBE, PBE0.25, PBE0.38, TPSS, TPSS0.1 and TPSS0.25, respectively. The \{s_6, s_{r,8}, s_8\} parameters of the double hybrid functionals 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 D3 correction is also used with the Becke-Johnson (BJ) damping function. In contrast to the D2 and D3 corrections, the D3(BJ) correction tends to a negative constant at short range.

\[
E_{\text{disp}}^{D3(BJ)} = -\frac{1}{2} \sum_{i \neq j} \sum_{n=6,8} s_n f_n(R_{ij}) \frac{C_n^{ij}}{R_{ij}^n}
\]

(53)

\[
f_n(R_{ij}) = \frac{1}{1 + e^{-d_n\left(\frac{R_{ij}}{s_r, n R_{0ij}} - 1\right)}}
\]

(54)

The short-range behavior is determined by $R_{0ij} = \sqrt{\frac{C_8^{ij}/C_6^{ij}}{s_r}}$. The global scaling factors of the energy terms are \{s_6, s_8\} = \{0.418, 0.0\} for PBEP86, \{s_6, s_8\} = \{0.64, 0.9147\} for B2PLYP and \{s_6, s_8\} = \{0.82, 0.29\} for PWPB95. The parameters in the Becke-Johnson damping are
\{a_1, a_2\} = \{0.0, 5.65\} for PBEP86, \{a_1, a_2\} = \{0.3065, 5.057\} for B2PLYP and \{a_1, a_2\} = \{0, 7.3141\} for PWPB95.

The above mentioned molecular mechanical corrections perform well on many datasets, but the large parameter set combined with various functionals might make such methods less transparent. Furthermore, the performance is uncertain for oxidation states not included in the training set as dispersion coefficients strongly depend upon ionization.

The non-local VV10 density functional dispersion correction is based on the local polarizability model. The correction is given by a non-local term minus its uniform density limit ensuring the correction to vanish for the homogeneous electron gas.

\[
E_C^{VV10}[n(r)] = \frac{1}{2} \int \int n(r)\Phi(r, r')n(r') \, dr \, dr' - \int \int c^{nl,unif}_C[n(r)]n(r) \, dr
\]

The non-local term considers the pairwise interactions of two differential volumes determined by a correlation kernel.

\[
\Phi(r, r') = \frac{3}{2} \frac{1}{g(r)g(r')[g(r) + g(r')]} \quad (58)
\]

\[
g(r) = \omega_0(r)R^2 + \kappa(r) \quad (59)
\]

The local function \(g(r)\) depends on the distance \(R = |r - r'|\), behaves as \(\omega_0(r)R^2\) at long range, and converges to \(\kappa(r)\) at short range. This ensures the correct \(-C_6/R^6\) asymptotic and \(-A + B R^2\) short-range behavior of the interaction.

\[
\omega_0(r) = \sqrt{\omega_1^2(r) + \omega_G^2(r)} \quad (60)
\]

The local frequency \(\omega_0(r)\) depends on the local dipole resonance frequency \(\omega_1(r) = \omega_p(r)/\sqrt{3}\) and the local band gap frequency \(\omega_G(r)\). The local plasma frequency \(\omega_p(r) = \sqrt{4\pi n(r)}\) depends on the electron density.

\[
\omega_G(r) = \sqrt{c} \left| \frac{\nabla n(r)}{n(r)} \right|^2 \quad (61)
\]

The local band gap frequency \(\omega_G(r)\) depends on the electron density and its gradient, and contains a universal parameter \(c = 0.0093\).

\[
\kappa(r) = b \frac{v_F^2(r)}{\omega_p(r)} \quad (62)
\]

The local function \(\kappa(r)\) depends on the local plasma frequency \(\omega_p(r)\) and the local Fermi velocity \(v_F(r) = (3\pi^2 n(r))^{1/3}\). The correction is fitted to any density functional method by
parameter $b$. For revPBE, TPSS, TPSSh, rPW86PBE, PBE, PBE0 and B2PLYP, we used $b = \{3.7, 5.0, 5.4, 5.9, 6.2, 6.5, 8.3\}$, respectively.\textsuperscript{173,174}

**Random phase approximation**

The random phase approximation (RPA) is a promising way to obtain improvements upon the standard semi-local density functional results.\textsuperscript{175,176,177} The addition of the exact exchange and the RPA correlation results in an orbital dependent functional, which has many desirable features. The exact exchange is free from self-interaction error, and the fully non-local RPA correlation seamlessly integrates dispersion interactions. RPA is accurate for intra- and intermolecular non-covalent interactions,\textsuperscript{178} for CO adsorption,\textsuperscript{179,180} for interlayer interactions,\textsuperscript{181} and for van der Waals bonded solids.\textsuperscript{182}

However, RPA significantly overestimates the correlation at short-range.\textsuperscript{183,184} It fails seriously in situations which lead to short range rearrangement of the electronic structure (for instance where the number of electron pairs changes). Consequently, RPA overestimates the ionization potentials\textsuperscript{185} and underestimates the atomization energies of molecules and solids.\textsuperscript{186,187} Furthermore, RPA suffers from self-interaction error, which lead to the underestimation of the barrier heights.\textsuperscript{188,189}

Beyond RPA, several approaches were applied to correct these errors. Semi-local density functional corrections in a range separated form can diminish the atomization energy errors.\textsuperscript{190} The SOSEX correction\textsuperscript{191} halves the large atomization error of RPA\textsuperscript{192} and fixes the one-electron self-correlation error.\textsuperscript{193} The renormalized PT2 method improves further the atomization energies, slightly improves the description of hydrogen transfer barrier heights but worsens the non-hydrogen transfer barrier heights.\textsuperscript{194} The approximate exchange kernel (AXK) method means a significant improvement upon RPA for atomization energies and ionization potentials without affecting the barrier heights.\textsuperscript{195}

Another well-documented problem with the RPA calculations is the slow convergence of the RPA correlation energy with the increasing basis set size. The slow convergence can be traced back to the poor description of the electron-electron Coulomb cusps by smooth orbital products.\textsuperscript{196,197}

The complete basis set (CBS) extrapolation method provides a simple way to deal with this problem. The CBS extrapolations are based on the convergence of the energy terms with respect to the increasing completeness of a series of basis sets. Such convergent energy series can be obtained by Gaussian or plane-wave basis sets with systematically increasing quality. A widely used Gaussian-type example is Dunning’s correlation consistent family of the aug-cc-pVXZ split valence atomic basis sets (also noted as AXZ) characterized by the X cardinal
number. The more complete basis sets are characterized with larger cardinal numbers and increasing computational times. The HF energy converges exponentially with respect to the increasing cardinal number. The two-point CBS extrapolation using the A6Z and A5Z basis sets (denoted as CBS(6/5)) reproduces the HF energy limit with route mean square error of 0.01 millihartree. For the so called W1 calculations, the inverse fifth power formula is close to optimal. The MP2 correlation energies converge much slower with the cardinal number. For the coupled cluster correlation energies, exponential, mixed exponential/Gaussian, inverse cubic and shifted quartic extrapolation schemes were suggested. For the RPA correlation energy, the inverse cubic formula, the power function with optimized exponent, and shifted inverse cubic or quartic formulae were proposed. At the CBS(7/6) and CBS(6/5) levels, all these approaches yielded similar results.

The direct random phase approximation (dRPA) can be obtained both in the framework of density functional theory and many-body perturbation theory, and it has a link to the coupled-cluster theory. In density functional theory, the dRPA can be derived from the adiabatic connection fluctuation-dissipation theory.

\[
E_C = \int_0^1 d\lambda \int d^3r \int d^3r' \frac{1}{|r' - r|} \left\{ -\frac{1}{2\pi} \int_0^\infty d\omega \left[ \chi^\lambda(r, r', i\omega) - \chi^{KS}(r, r', i\omega) \right] \right\} \tag{63}
\]

In this expression, the fluctuation potential is the Coulomb potential, and the correlation contribution of the two-particle density matrix (the expression within the curly brackets) comes from the fluctuation-dissipation theorem. The Kohn-Sham (\(\chi^{KS}\)) and \(\lambda\)-dependent (\(\chi^\lambda\)) response functions are given by the following equations (the latter is called the Dyson equation).

\[
\chi^{KS}(r, r', i\omega) = \sum_{\text{occ}} \sum_{\text{unocc}} \frac{\varepsilon_i - \varepsilon_a}{(\varepsilon_i - \varepsilon_a)^2 - (i\omega)^2} \varphi_i(r)\varphi_a(r')\varphi_i(r)\varphi_a(r') \tag{64}
\]

\[
\chi^\lambda(r, r', i\omega) = \chi^{KS}(r, r', i\omega) + \int d^3r_1 \int d^3r_2 \chi^{KS}(r, r_1, i\omega) \left[ \frac{\lambda}{|r_2 - r_1|} + f^\lambda_{XC}(r_1, r_2, i\omega) \right] \chi^\lambda(r_2, r', i\omega) \tag{65}
\]

In the dRPA method, the exchange-correlation kernel is equal to zero (\(f^\text{dRPA}_{XC}(r_1, r_2, i\omega) = 0\)). The above equations can be solved iteratively. (\(\varphi\) and \(\varepsilon\) are the orbitals and the corresponding energies.)
In perturbation theory, the dRPA correlation contains all the MP2-like perturbative direct terms summed up to infinite order on the level of double excitations. In the coupled-cluster theory, the dRPA is equal to the direct ring coupled cluster with double excitations (direct ring-CCD) method. The dRPA method uses full exact exchange, and its correlation part captures the non-locality and non-pairwise nature of dispersion interactions. It gives almost exact correlation at long range but seriously overestimates the correlation at short range.

We use here an efficient implementation for the dRPA correlation in the coupled-cluster framework, which is available in the MRCC quantum chemistry software. The calculations are performed in a postprocessing way, where the single-particle orbitals can be obtained from a self-consistent HF or density functional calculation. The non-self-consistent results are written in the form of method@orbitals (e.g. dRPA@PBE).

\[ E_{\text{C}}^{\text{dRPA}} = \frac{1}{2} \text{tr}[B \ T] \]  

The \( B \) non-antisymmetrized two-electron repulsion integral matrix is defined by the \( B_{ia,jb} = \langle ij|ab \rangle \) four-index matrix elements using occupied \( i,j \) and virtual \( a,b \) spin-orbital indices. The \( T \) double excitation amplitude matrix is given by an iterative procedure initialized with \( T^{(0)} = 0 \). The \( t_{ij}^{ab} \) double excitation amplitudes are equal to the \( T_{ia,jb} \) matrix elements.

\[ T^{(n+1)} = \Delta \circ \left( B + B T^{(n)} + T^{(n)} B + T^{(n)} B^2 T^{(n)} \right) \]  

The \( \Delta \) orbital energy denominator matrix is defined by the \( \Delta_{ia,jb} = \frac{1}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j} \) matrix elements. The circle denotes the Hadamard product. The equations are solved in an \( O(N^4) \)-scaling iterative procedure by density-fitting of the electron repulsion integrals and Cholesky decomposition of the orbital energy denominator matrix. For small HOMO-LUMO gap systems a plasmon-formula-based algorithm is executed in order to prevent unphysical solution or divergence.

The random phase approximation with second-order screened exchange (SOSEX for short) contains also the MP2-like perturbative exchange terms; therefore, it is exact to the second-order perturbation theory. The SOSEX correlation functional uses the \( K \) spin-singlet adapted antisymmetrized two electron repulsion integral matrix with the \( K_{ia,jb} = \langle ij||ab \rangle \) matrix elements.

\[ E_{\text{C}}^{\text{SOSEX}} = \frac{1}{2} \text{tr}[K \ T] \]  

The SOSEX method describes better the correlation at short range, and yields better atomization energies and lattice constants than the dRPA method. However, it often underestimates the correlation at long range.
The RPAX and RPAX2 methods take into account the exchange interactions in the correlation. The RPAX method uses for the correlation energy a similar formula to the dRPA method with a different iterative update equation for the double excitation amplitude matrix.

\[
E_{c}^{RPAX} = \frac{1}{2} \text{tr}[B \, T^{RPAX}]
\]

\[
T^{(n+1)} = -\Delta \circ \left[ (1 + T^{(n)})B(1+T^{(n)}) - (1+T^{(n)})\hat{P}B(1+T^{(n)}) \right] -
\]

The \( \hat{P} \) permutation operator permutes the orbitals. The RPAX (or ring-CCD) method has an unfavorable \( O(N^6) \)-scaling like the CCD method, and the RPAX reaction energies are only slightly better than the MP2 energies.

The RPAX2 correlation energy is written using Cholesky decomposition of the two-electron repulsion integral matrix as \( B = L L^T \).

\[
E_{c}^{RPAX2} = \frac{1}{2} \text{tr}[L^T T^{(\infty)}L]
\]

This method uses also a slightly different iterative update equation for the double excitation amplitude matrix.

\[
T^{(n+1)} = -\Delta \circ \left[ (1 + T^{(n)})B(1+T^{(n)}) - \hat{P}(1+T^{(n)})B(1+T^{(n)}) \right] -
\]

The RPAX2 method can be implemented with an \( O(N^5) \)-scaling, and the RPAX2 reaction and interaction energies agrees well with the CCSD(T) energies.

### 3.3. Thermochemical model

The electronic energy \( (\varepsilon_0)\), the thermal corrections to the energy \( (E_{tot}) \) including the zero-point energy, the Boltzmann contribution \( (k_B T) \) to the enthalpy, and the entropic contribution \( (-TS_{tot}) \) to the Gibbs energy \( (G) \) can be obtained from wave function and density functional methods.

\[
G = \varepsilon_0 + E_{tot} + k_B T - TS_{tot}
\]

However, the computations are so expensive especially for large biomolecules that they can be performed only in gas-phase or in the presence of very few explicit water molecules.
4. Benchmarking density functionals

We always calculate the deviations as: Result – Reference. We use the mean deviation (MD) to quantify the accuracy, the corrected sample standard deviation (CSSD) to quantify the precision, and the mean absolute deviation (MAD) to measure both at the same time.

4.1. On intermolecular anion-pi interactions [S1]

First we discuss the twenty anion-π interaction energies between the F-, Cl-, Br-, NO₃⁻ or CO₃²⁻ ions and the HFB, TFB, TFZ or TAZ π-systems. In the complexes, the anions are bound along the symmetry axis of the aromatic ring as shown in Figure 3. The fully optimized RI-MP2(frozen core)/6-31++G(d,p) coordinates were taken from ref 119. 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⁻, and the longest distance is 3.487 Å for TFB ··· Br⁻.

<table>
<thead>
<tr>
<th>Quadrupole moment (Qzz)</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Hexafluorobenzene (HFB)</td>
<td>Trifluorobenzene (TFB)</td>
</tr>
<tr>
<td>Low</td>
<td>Trifluorotriazine (TFZ)</td>
<td>Triazine (TAZ)</td>
</tr>
</tbody>
</table>

**Figure 4.** Electrostatic potential (ESP) map, and magnitude of axial static dipole polarizability and quadrupole moment of the four anion-acceptor π-systems.

The four aromatic rings can be classified by their Qzz permanent quadrupole moment and αzz static dipole polarizability components along the main symmetry axis of the aromatic rings (Figure 4). The electrostatic potential maps are also shown in Figure 4. The Qzz
components of HFB, TFZ, TFB and TAZ are 9.5, 8.2, 0.9 and 0.9 Buckingham (B), respectively.\textsuperscript{213} For comparison, the $Q_{zz}$ component of benzene is -8.7 B.\textsuperscript{214} The ratios of the calculated $\alpha_{zz}$ components of the HFB, TFB, TFZ and TAZ are 90\%, 93\%, 70\% and 72\% compared to $\alpha_{zz}$ for benzene. The $\alpha_{zz}$ of benzene is around 6.7 Å\textsuperscript{3} calculated with HF or MP2 methods. In the HFB and TFZ molecules, a relatively large partial positive charge concentrated above the center of the ring. In TFB, the partial positive charge is located on the ipso carbon atoms. In TAZ, larger alternating partial charges occur.

We can also characterize the five anions with respect to their $\alpha_{zz}$ components. The larger polarizability might lead to stronger dispersion interaction depending on the anion-\pi distance. The MP2/ATZ polarizabilities of the free F\textsuperscript{-}, Cl\textsuperscript{-}, Br\textsuperscript{-}, NO\textsubscript{3}\textsuperscript{-} and CO\textsubscript{3}\textsuperscript{2-} anions are 1.3, 4.1, 5.9, 5.1 and 8.2 Å\textsuperscript{3}, respectively. Notice that except for fluoride, the polarizability values are comparable to the $\alpha_{zz}$ values (4.7-6.2 Å\textsuperscript{3}) of the aromatic rings discussed above.

![Interaction Energy Components](image)

**Figure 5.** SAPT0 interaction energy decomposition of the binary anion-\pi complexes.

In order to qualitatively analyze the anion-\pi interaction energy components, we have performed SAPT0 analysis (Figure 5). The repulsive exchange terms are very large, 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. Thus conclusions of the earlier analyses 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). The exchange and electrostatic terms are larger for the HFB and TFZ complexes as their $Q_{zz}$ component is larger. This is particularly true with the CO$_3^{2-}$ ion, as the center of the CO$_3^{2-}$ ion is quite close (~2.5 Å) to the center of the aromatic ring. For the complexes with HFB and TFZ, the attractive electrostatic term overcompensates the repulsive exchange term in contrast to for the complexes with TFB and TAZ.

The induction energy components are smaller (on average 60% of the total interaction energy and spanning the 33-96% range). The induction is particularly important in the interactions for the complexes of TFB and TAZ. The induction energy components are particularly large for the complexes with the CO$_3^{2-}$ ion, and smaller but still significant for the complexes with the F$^-$ ion.

The dispersion energy components are indeed smaller but non-negligible in many complexes (on average 50% of the total interaction energy and spanning the 14-107% range). The dispersion interaction energy component is the most important for the complexes with TAZ and TFB, and the least important for the complexes with F$^-$ ion. The dispersion energy components are particularly large for the complexes with the CO$_3^{2-}$ ion. However, the relative importance of the dispersion is smaller because of the large induction component. In the complexes of TAZ or TFB with Br$^-$ or NO$_3^-$ ions, the electrostatic and inductive attraction roughly compensates the exchange repulsion, hence the overall interaction energy is very sensitive to the dispersion interaction.

We computed new benchmark DLPNO-CCSD(T)/CBS interaction energies for the twenty binary anion-π complexes. We extrapolated the HF energy and the MP2 correlation energy separately by the extrapolation parameters $\alpha = 0.269$ and $\beta = 0.712$, respectively. These parameters are slightly different from the earlier parameters optimized for the conformational energies of carbohydrates. Then we used the focal point method to estimate the CCSD(T)/CBS energy.

$$E^{\text{HF}}_{\text{CBS}} = E^{\text{HF}}_{AQZ} + \alpha (E^{\text{HF}}_{AQZ} - E^{\text{HF}}_{ATZ})$$ (74)

$$E^{\text{MP2c}}_{\text{CBS}} = E^{\text{MP2c}}_{AQZ} + \beta (E^{\text{MP2c}}_{AQZ} - E^{\text{MP2c}}_{ATZ})$$ (75)
\[ E_{CBS}^{CCSD(T)} = E_{CBS}^{MP2} + \left( E_{ATZ}^{CCSD(T)} - E_{ATZ}^{MP2} \right) \]

Our benchmark interaction energies (Table 2) show reasonable agreement with the CCSD(T)/CBS halide ion complex binding energies with HFB, TFB, and TFZ but we used the CCSD(T)/ATZ energies for the focal point calculations instead of the CCSD(T)/ADZ energies.\(^{216}\) It has been shown for various benzene and pyrazine anion- and cation-\(\pi\) complexes that the ATZ basis set is suitable to obtain reasonable results combined with RI-MP2 method.\(^{217}\)

**Table 2.** The reference interaction energies (in kcal mol\(^{-1}\)) of the binary anion-\(\pi\) complexes.

<table>
<thead>
<tr>
<th></th>
<th>HFB</th>
<th>TFB</th>
<th>TFZ</th>
<th>TAZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(^-)</td>
<td>-21.08</td>
<td>-10.86</td>
<td>-27.67</td>
<td>-11.9</td>
</tr>
<tr>
<td>Cl(^-)</td>
<td>-14.88</td>
<td>-7.43</td>
<td>-18.03</td>
<td>-7.26</td>
</tr>
<tr>
<td>Br(^-)</td>
<td>-13.81</td>
<td>-6.69</td>
<td>-15.87</td>
<td>-6.37</td>
</tr>
<tr>
<td>NO(_3^+)</td>
<td>-14.89</td>
<td>-8.66</td>
<td>-16.07</td>
<td>-7.32</td>
</tr>
<tr>
<td>CO(_3^{2-})</td>
<td>-38.36</td>
<td>-24.31</td>
<td>-51.07</td>
<td>-22.42</td>
</tr>
</tbody>
</table>

The former MP2-CP reference energies\(^{119}\) have +3.54 kcal mol\(^{-1}\) error on average and +14.13 kcal mol\(^{-1}\) error at most compared to the new DLPNO-CCSD(T)/CBS reference energies. For the complexes of TAZ with F\(^-\), Br\(^-\) and Cl\(^-\) ions, the CP corrected MP2 energies generally have 2-3 kcal mol\(^{-1}\) error compared to MP2/CBS. The CP correction significantly overcompensates the basis set error, and this is the origin of the large and random underbinding error in the MP2-CP interaction energies. Hence the earlier conclusion about the good performance of MPWB1K functional is questionable.\(^{119}\)

Similarly to the SAPT0 analysis, the magnitude of the correlation energy in the DLPNO-CCSD(T)/CBS interaction energies also indicate the importance of the dispersion interaction. This is smaller for the halide ion complexes (2.6-5.5 kcal mol\(^{-1}\)), and considerably larger for the NO\(_3^+\) (8-9 kcal mol\(^{-1}\)) and CO\(_3^{2-}\) complexes (11-13 kcal mol\(^{-1}\)). Despite the short interaction distance in the fluoride complexes, the correlation energy is only a small fraction of the total interaction energy (~10%).

We tested several wave function methods (Figure 6) with the ATZ basis set for the anion-\(\pi\) interaction energies. The MP2 method overestimates the interaction energy because the missing non-pairwise interaction energy terms are positive. The SOSEX method underestimates the interaction energy because it underestimates the long-range correlation. The good dRPA results can be attributed to the small overbinding error of the ATZ correlation. The good dRPA results can be attributed to the small overbinding error of the ATZ correlation. The good dRPA results can be attributed to the small overbinding error of the ATZ correlation. The good dRPA results can be attributed to the small overbinding error of the ATZ correlation. The good dRPA results can be attributed to the small overbinding error of the ATZ correlation.
slightly towards the underbinding direction. The mean absolute deviations from the benchmark increase as LPNO-CEPA1 < RPA = dRPA < SOSEX < MP2.

We have performed a CPU-time based efficiency analysis of the above mentioned wave function methods (the CPU times of the HF calculations are excluded from the following comparisons) with the ATZ basis set. The computational times increase as MP2 << dRPA ≈ SOSEX < LPNO-CEPA1 << DLPNO-CCSD(T) < RPA. The observed scaling with the system size increases as LPNO-CEPA1 < MP2 < dRPA ≈ SOSEX < DLPNO-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. The LPNO-CEPA1 method scales better than the dRPA, so the former method can be more accurate, precise and efficient only for very large systems.

![Figure 6](image)

*Figure 6.* Performance of different theoretical methods on the anion-π interactions relatively to the reference method. Deviation = ΔE(method) – ΔE(ref.). Mean deviations: red line; deviation ranges: yellow bars; ±3 corrected sample standard deviation band: purple lines. The ATZ basis set is applied unless stated otherwise.

Next we used our new benchmark energies to rank various approximate density functionals (with the ATZ basis set). The VV10 corrected PBE and TPSS methods give accurate
but quite imprecise results. The standard hybrid functionals give similar or slightly better results. The double hybrid functionals perform reasonably well but do not reach the precision of the dRPA. The double hybrids without dispersion correction are precise (CSSD ≈ 0.5 kcal mol⁻¹) but inaccurate (MD = 0.9-1.7 kcal mol⁻¹). The D2 and D3(BJ) dispersion corrections efficiently correct this systematic underbinding error, but these corrections make the methods considerably less precise (CSSD ≈ 0.9 kcal mol⁻¹). This shows the somewhat random nature of these a posteriori corrections. The mean absolute deviations from the benchmark increase as dRPA < PWPB95–D3(BJ) < B2PLYP–D3(BJ) < mPW2PLYP–D2 < B2PLYP–VV10 << PBE–VV10 ≈ TPSS–VV10 < revPBE–VV10 << rPWPBE–VV10.

We used the dRPA/ATZ binding energies as a benchmark for the 30 π-anion-π’ sandwich complexes. The former CP corrected RI-MP2(FC)/6-31++G(d,p) reference interaction energies show a serious disagreement with the new reference up to almost 20 kcal mol⁻¹. We computed the interaction energies with several semi-local and double hybrid methods with or without dispersion corrections using the ATZ basis set. The semi-local functionals without dispersion correction show serious underbinding error (8-15 kcal mol⁻¹). The best dispersion corrected semi-local functional among the examined method is PBE–VV10, which provides very accurate (MD = 0.05 kcal mol⁻¹) but considerably less precise (CSSD = 2.63 kcal mol⁻¹) interaction energies. The double hybrid functionals without dispersion correction show 2-3 kcal mol⁻¹ underbinding error. The best dispersion corrected double hybrid is B2PLYP–D3(BJ) among the examined methods, which provides precise (CSSD = 1.31 kcal mol⁻¹) but less accurate (MD = -0.67 kcal mol⁻¹) interaction energies. The mean absolute deviations from the benchmark increase as B2PLYP–D3(BJ) < mPW2PLYP–D2 << PBE–VV10 < TPSS–VV10 < revPBE–VV10. This ranking is consistent with the previous ranking for the binary anion-π complexes, with the exception that PBE–VV10 is clearly better than TPSS–VV10 here. The transferability of the errors from anion-π complexes to ternary π-anion-π’ sandwich complexes suggests that some kind of additivity effects exist within this class of molecules.

The additivity of the interactions can be analyzed in the ternary π-anion-π’ complexes using the highly accurate dRPA/ATZ interaction energies for all possible combinations of the four aromatic compounds and the five anions presented here. The dRPA interaction energies of the ternary complexes are only ~93% of the sum of the two interaction energies of the corresponding binary complexes on average (see the slope of the fitted line in Figure 7). This is only ~89% by the PBE–VV10 method and ~92% by the B2PLYP–D3(BJ) method.
The non-additivity of the interaction energies originates mainly from the induction as the electron density of the anion cannot be induced in the sandwich so much. (This means that the PBE-VV10 method significantly overestimates, the B2PLYP-D3(BJ) method slightly overestimates the induction.) The non-additivity is particularly large in the ternary complexes with HFB and TFZ especially with the CO$_3^{2-}$ ion, where the induction is very large in the corresponding binary complexes. In these cases, the equilibrium distances in the binary and ternary complexes are very similar.

![Figure 7](image)

*Figure 7.* Non-additivity of the anion-π interactions in the 30 ternary π-anion-π complexes.

The additivity is almost perfectly satisfied in the ternary complexes with TFB and TAZ especially with the Br$^-$ ion, where the induction is small, and the dispersion dominates in the corresponding binary complexes. In these cases, the equilibrium distances are slightly longer in the ternary complexes than in the binary complexes. However, the interaction energy is quite insensitive to the distance, as the potential energy surface is very flat around the minima.
4.2. On intramolecular interactions in Diels-Alder reactions [S2]

The accurate description of the Diels-Alder reactions in the DARC test set (Table 1) require a method which can compute correctly the non-covalent intramolecular interactions, and the bond formation and transition at the same time. In this section, we shall analyze the various error sources in the approximate functionals and provide methods to reliably eliminate them.

According to the original suggestion, the semi-local functionals overestimate the exchange repulsion of the non-bonded electron density overlaps because of their self-interaction error. Hence we examine first the role of the exact exchange in global hybrid functionals. We computed the reaction energies with the PBE, PBE0.25, PBE0.32 and PBE0.38 methods and used the (aug)-cc-pVTZ(-f,-d) basis set (abbreviated as aTZ(-f,-d)) for the calculations. The initial PBE method gives quite good reaction energies for the first two reactions with monocyclic products. However, it shows the typical endothermic (underbinding) density functional error for the reactions with cage-like products. This error is particularly large (>8 kcal mol\(^{-1}\)) for reactions with tricyclic products.

![Figure 8](image_url)

**Figure 8.** Deviations from the benchmark reaction energies (\(\Delta E\) (method) – \(\Delta E\) (ref.)) of the DARC test set for semi-local and global hybrid PBE methods (with aTZ(-f,-d) basis set). The positive deviations are endothermic.

The increasing fraction of the exact exchange in the PBE hybrids shifts the reaction energies towards the exothermic direction (Figure 8) but does not considerably improve the
precision (the tendency is similar for the TPSS hybrids in GMTKN30 database).\textsuperscript{136} The average effect of 1\% exact exchange mixing is larger in the reactions of ethane (-0.26 kcal mol\textsuperscript{-1}) than in the similar reactions of ethyne (-0.22 kcal mol\textsuperscript{-1}). This effect is even larger in the reactions, where tricyclic products are formed with more extended electronic structure (-0.29 kcal mol\textsuperscript{-1}). The most accurate PBE hybrid is PBE0.25, but the most precise is PBE0.38. The PBE0.32 hybrid is a good compromise among these functionals (MAD = 2.84 kcal mol\textsuperscript{-1}). Surprisingly, the PBE0.32 reaction energies are considerably better than the MCY3 reaction energies (MAD = 3.4 kcal mol\textsuperscript{-1}) and only slightly worse than the rCAM-B3LYP (2.6 kcal mol\textsuperscript{-1}) results. This clearly shows that one can obtain quite accurate reaction energies using common global hybrid functionals which contain considerable delocalization error.

\textbf{Figure 9.} Deviations from the benchmark reaction energies ($\Delta E($method$) – \Delta E($ref.)) of the DARC test set for the molecular mechanical D3 (with def2QZ basis set, taken from GMTKN30)\textsuperscript{137} and non-local VV10 (with ATZ basis set) dispersion corrected PBE and TPSS methods. The positive deviations are endothermic.

The non-bonding electron density overlaps are accompanied by intramolecular dispersion interactions, which mean medium range correlation. Hence we examine also the dispersion corrected density functionals on the DARC test set (\textbf{Figure 9}). As the attractive intramolecular dispersion interactions stabilize the mono-, bi- and tricyclic products more than the reactants, the calculated dispersion corrected Diels-Alder reaction energies are always shifted in the exothermic direction. We computed the VV10 corrected PBE and TPSS reaction
energies (with the ATZ basis set). For comparison, we also took the D3 corrected PBE and TPSS reaction energies with the def2-QZVP basis set (abbreviated as def2QZ) from the GMTKN30 database.\textsuperscript{136} The VV10 correction has even larger effect on the reaction energies than the D3 correction. Since TPSS usually gives more repulsive potential energy curves than PBE, the corrections are also larger for TPSS than for PBE. The dispersion corrected TPSS reaction energies are more precise than the dispersion corrected PBE reaction energies. These results suggest that the one-electron self-correlation free TPSS has also smaller many-electron self-interaction error (cf. SIE11 subset of GMTKN30 database).\textsuperscript{136}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure10.png}
\caption{Deviations from the benchmark reaction energies ($\Delta E$(method) – $\Delta E$(ref.)) of the DARC test set for hybrid and double hybrid methods which consider the dispersion (with def2QZ basis set, taken from GMTKN30).\textsuperscript{137} The positive deviations are endothermic.}
\end{figure}

Next we assess the performance of the hybrid or double hybrid functionals (e.g. M06-2X, ωB97X-D, PWPB95-D3) which also consider dispersion (\textit{Figure 10}). The semi-local functionals yield different errors for reactions with different diene reactants, and with mono-, bi-, and tricyclic products. Mixing with exact exchange and applying dispersion correction do not help in this respect. In addition, the dispersion corrections always shift the reaction energies towards the exothermic direction, so they worsen the already accurate or even slightly exothermic reaction energies of the hybrid functionals systematically by 2-4 kcal mol\textsuperscript{-1} (e.g. MD = -3.08 kcal mol\textsuperscript{-1} for PBE0-D3). For comparison, the reaction energies by some of the best performing hybrid and double hybrid density functionals are taken from the literature.
which also consider the intramolecular dispersion interactions.\textsuperscript{136,218} The dispersion-corrected range-separated hybrid ωB97X-D with 100\% of exact exchange in long range is accurate (MD = 0.41 kcal mol\textsuperscript{-1}) but not precise (CSSD = 2.49 kcal mol\textsuperscript{-1}). The global hybrid M06-2X functional with 54\% exact exchange is less accurate (MD = 1.92 kcal mol\textsuperscript{-1}) but more precise (CSSD = 1.97 kcal mol\textsuperscript{-1}) than ωB97X-D. The similar M08-SO functional with \sim 57\% exact exchange is even more accurate and precise (MAD = 1.46 kcal mol\textsuperscript{-1}). This performance can be achieved also by the double hybrid PWPB95-D3 functional (MAD = 1.52 kcal mol\textsuperscript{-1}).

\textbf{Figure 11.} Deviations from the benchmark reaction energies (∆E(method) – ∆E(ref.)) of the DARC test set for the complete basis set extrapolated dRPA method (on different self-consistent reference orbitals). The positive deviations are endothermic.

Next we assess the performance of the non-self-consistent dRPA methods (\textit{Figure 11}). dRPA seamlessly integrates dispersion interactions. It can treat chemical reaction energies with good precision, but it contains self-correlation error. It is interesting how dRPA reproduces the quite complicated bond rearrangements, hybridization changes and ring formations in the Diels-Alder reactions. We calculated the dRPA reaction energies on self-consistent HF, PBE and PBE0.25 reference molecular orbitals. Since the dRPA correlation energy converges very slowly with respect to the increasing basis set size, we extrapolated the dRPA energies to the CBS limit (for the details of the extrapolation see the next chapter). The ATZ basis set error might shift the calculated dRPA reaction energies in the exothermic direction by 2-4 kcal mol\textsuperscript{-1}. The basis set effect is larger on PBE orbitals than on HF orbitals because the less polarized
basis sets cannot describe well the more diffuse valence region. The more compact HF electron density also causes an exothermic shift in the dRPA reaction energies because the non-covalent electron density overlaps are smaller, thus the exchange repulsion is smaller. The dRPA@PBE0.25 method shows excellent performance with CBS extrapolation (MD = -0.53, MAD = 0.53 and CSSD = 0.47 kcal mol\(^{-1}\)). The endothermic error of dRPA@PBE is quite efficiently compensated by the AQZ exothermic basis set error leading to very accurate and precise reaction energies (MD = 0.08, MAD = 0.32 and CSSD = 0.42 kcal mol\(^{-1}\)).

Beyond RPA, the SOSEX correction worsens the results (MAD = 2.33 kcal mol\(^{-1}\)) as it gives an unbalanced description of the correlation effects.\(^{219}\) The RPAX2 method using the ATZ basis set and PBE exchange only (PBEx) reference orbitals leads to very precise results but has an exothermic error (MD = -0.83, MAD = 0.83, CSSD = 0.56 kcal mol\(^{-1}\)). The CBS extrapolation of the RPAX2 reaction energies worsens the precision of the reaction energies and leads to an endothermic error (MD = 0.85, MAD = 1.06, CSSD = 0.94 kcal mol\(^{-1}\)).\(^{219}\)

Finally, the all the examined methods showed similar errors on the reactions with furane (7-10) or cyclopentadiene (11-14) because the intramolecular interactions are nearly equal in the cage-like products. Hence there are many redundant reactions in the test set, and the tricyclic reactions are overrepresented. For future benchmarking, we suggest to use a more balanced, smaller database with only six reactions (1-4, 7, 11) called DARC6.
5. Method development

5.1. New complete basis set extrapolation for direct random phase approximation [S3]

We examine the basis set convergence of the dRPA correlation energies (using HF reference orbitals) on a set of 65 hydrocarbon isomers from CH$_4$ to C$_6$H$_6$ (Figure 12). The molecular geometries are optimized with B3LYP/6-31G(2df,p), and taken from NIST Computational Chemistry Comparison and Benchmark Database.\(^{220}\)

![Figure 12. Test set of 65 hydrocarbon isomers from CH$_4$ to C$_6$H$_6$ with B3LYP/6-31G(2df,p) geometries (taken from NIST-CCCBDB).](image)

We use the inverse cubic function suggested in the literature for the extrapolation of the dRPA correlation (dRPAc) energy.\(^{207}\) In addition, we test the exponential decay function, and the generalized power function.\(^{205}\)

\[
E_{CBS}^{dRPAc} = E_{AXZ}^{dRPAc} - AX^{-3} \quad (77)
\]

\[
E_{CBS}^{dRPAc} = E_{AXZ}^{dRPAc} - Ae^{-nX} \quad (78)
\]

\[
E_{CBS}^{dRPAc} = E_{AXZ}^{dRPAc} - A(X + d)^{-\alpha} \quad (79)
\]
Parameters $A$, $n$ and $\alpha$ are fitted, $d$ is a shift parameter, which modifies the cardinal numbers ($d = 0$ for the simple power fit). We define the fitting error as the difference between the fitted and calculated dRPAc energies.

Here, we apply a two-point extrapolation scheme for the dRPAc energy similar to the two-point extrapolations applied in the literature.\textsuperscript{215}

\[
E_{\text{CBS}(X,X-1)}^{\text{dRPAc}} = E_{\text{AXZ}}^{\text{dRPAc}} + C_{X,X-1}^{\text{dRPAc}} \left( E_{\text{AXZ}}^{\text{dRPAc}} - E_{\text{AX}(X-1)Z}^{\text{dRPAc}} \right)
\]  

The $C_{X,X-1}^{\text{dRPAc}}$ extrapolation coefficients can be calculated substituting $(X)$ and $(X - 1)$ into the equations describing the convergence.

**Figure 13.** The average fitting errors of the inverse cubic function, the exponential decay, and the optimized power function with the exponent of -2.56 for the convergence of the dRPAc correlation energies (with HF reference orbitals, with AXZ basis sets, for the 65 hydrocarbons). 95% confidence intervals by error bars; small mean errors in legend.

With the inverse cubic formula, the predicted dRPAc energy values show large and systematic errors (**Figure 13**). This behavior clearly shows that the inverse cubic formula is not suitable for accurate extrapolation of the dRPAc energies. The fitted exponential function formula gives somewhat smaller but still significant, oscillating average fitting errors. The best performance is shown by the fitted power function (with $d = 0$) using $\alpha = 2.56$. (Its shifted versions with $d = \pm 0.5$ perform similarly. The best fitting can be obtained by $d = 0.165$. This is only slightly different from the unshifted version, so we will discuss here only the unshifted
The predicted energies are 100-times more accurate than those of predicted from the inverse cubic formula. The power function form fits particularly well for the ATZ and AQZ energies, so the CBS energy can be estimated from these two energies accurately (with $C_{4,3}^{dRPAc@HF} = 0.918$).

The optimal exponents are 2.59 for the free hydrogen atom, 2.30 for the free carbon atom, 2.52 for ethyne, 2.61 for ethene, 2.60 for ethane, 2.19 for CO, 2.24 for NH$_3$, 2.20 for H$_2$O, and 2.19 for HF. For small molecules, the basis set convergence is sensitive to the molecular structure, but this sensitivity decreases with the increasing molecular size. So the exponents which work well for atoms or small molecules are not well applicable for larger molecules.

Detailed analysis shows that the fitted exponents also slightly differ for each molecule in our test set. The optimal exponent is about 2.70 for hydrogen molecule, 2.66 for methane, 2.53 for the ethyne molecule, 2.50 for 1,3,5-hexatriyne, and 2.35 for the C$_2$ molecule. Since the basis set convergence is faster for the hydrogen atoms than for the carbon atoms, the convergence slows down with the decreasing H:C ratio in the molecule. The origin of the faster convergence on the hydrogen-atom-rich molecules is the smaller electron-electron Coulomb cusp error on the less compact electron densities. In addition, the electronic structure of the carbon atoms significantly varies with the hybridization states. We have found that the exponents can be predicted from the number of carbon and hydrogen atoms in the molecule, or from the hybridization states of the carbon atoms.

$$\alpha \approx m_{\text{atom}} + \frac{N_H}{N_H + N_C} p_{\text{atom}}^H + \frac{N_C}{N_H + N_C} p_{\text{atom}}^C$$  \hspace{1cm} (81)

$$\alpha \approx m_{\text{hybr}} + \frac{N_H}{N_H + N_C} p_{\text{hybr}}^H + \frac{N_C}{N_H + N_C} p_{\text{hybr}}^C + \frac{N_C(p_{sp})}{N_H + N_C} p_{hybr}^{C(sp)} + \frac{N_C(p_{sp^2})}{N_H + N_C} p_{hybr}^{C(sp^2)} + \frac{N_C(p_{sp^3})}{N_H + N_C} p_{hybr}^{C(sp^3)}$$ \hspace{1cm} (82)

$\alpha$ is the exponent; $m$ and $P$ are fitting parameters (Table 3); $N$ is the number of atoms from different types. In the test set of 65 hydrocarbon molecules, we have 450 $H$ and 291 $C$ atoms [48 C(sp), 110 C(sp$^2$) and 133 C(sp$^3$)].
Table 3. The average and atomic parameter values for the exponents of the power formula (in eqs (81) and (82)).

<table>
<thead>
<tr>
<th>Average</th>
<th>Atomic contributions ( (P_{atom}^A) )</th>
<th>Hybridization state contributions ( (P_{hybr}^{A(state)}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
<td>2.5608 ( m_{atom} ) 2.6184 ( P_{H}^{H} ) 0.0268 ( P_{C}^{C(sp)} ) -0.1829 ( P_{C}^{C(sp^2)} ) -0.1852 ( P_{C}^{C(sp^3)} ) -0.2972</td>
<td></td>
</tr>
</tbody>
</table>

Figure 14. Correlation between the predicted and fitted exponents of the power functions describing the basis set convergence of the dRPA correlation energies (with HF reference orbitals, with AXZ basis sets, for the 65 hydrocarbons). Average exponent: yellow line; exponents from atomic contributions: blue crosses; exponents from hybridization states: green plus signs.
The correlation between the calculated and the fitted exponents shows that the consideration of hybridization states leads to better estimations for the exponents (Figure 14). The prediction is the poorest for methane. Two other problematic predictions are for ethyne and ethene. From the hybridization state contributions, we can extrapolate the exponents of larger hydrocarbons. For example, the exponents of C_8H_{18}, C_{10}H_{22}, and C_{12}H_{26} are 2.574, 2.572, and 2.571, respectively. For very large saturated linear hydrocarbons the exponent converges to 2.563, as the C:H ratio approaches to 1:2. This model gives \( \alpha = 2.71 \) for molecules constituted only by hydrogen atoms, \( \alpha = 2.25 \)–2.41 for molecules constituted only by carbon atoms, which are around the fitted exponents for H_2 and C_2.

![Figure 15](image)

*Figure 15.* Mean absolute deviations of the predicted (with average exponent, exponents from the atomic contributions, exponents from the hybridization states) from the fitted complete basis set extrapolated dRPA correlation energies (with HF reference orbitals, with AXZ basis sets, for the 65 hydrocarbons).

Next we analyze the improvements upon the CBS extrapolated dRPAc energies yielded by the application of predicted exponents over the average exponent (Figure 15). The extrapolation error is defined here as the difference between the two-point extrapolated and our best four-point extrapolated dRPAc/CBS energies. (Note that the error of the density fitting is included in these calculations, which also limits the accuracy.) The application of atomic corrections almost halves the CBS(4/3) extrapolation error by the average exponent. This error halves again by considering the hybridization states. The atomic corrections largely decreases
the CBS(5/4) extrapolation. However, the hybridization states corrections do not lead to a significant further improvement. The application of the atomic or hybridization states corrections do not improve significantly the CBS(6/5) extrapolated energies. This means that the energies calculated with smaller ATZ or AQZ basis sets depend considerably on the molecular structure, while the energies calculated with larger A5Z and A6Z basis sets are quite independent of the molecular structure. The main conclusion is that the accuracy and precision of the dRPA/CBS(4/3) extrapolation can be considerably improved by the application of hybridization states corrections, and CBS(6/5) quality results can be reached quickly from ATZ and AQZ calculations, which is a huge speed up.

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 dRPAc energy as the potentially less compact PBE molecular electron density might lead to smaller cusp error. We can estimate quickly the mean extrapolation coefficient for different reference orbitals computing only the dRPAc energies of ethyne and ethane as it is 0.917 for the CBS(4/3) extrapolation, which is close to the above determined \( C_{CBS(4/3)}^{dRPAc@HF} = 0.918 \). The predicted mean coefficients for hydrocarbons are \( C_{CBS(4/3)}^{dRPAc@PBE} = 0.856 \) and \( C_{CBS(4/3)}^{dRPAc@PBE0} = 0.867 \). For the N\(_2\) molecule with more compact electron density, the basis set convergence only slightly depends on the reference orbitals (\( \alpha = 2.20 \) and 2.18 for PBE and HF reference orbitals, respectively). The difference between the extrapolation coefficients is practically negligible.

Finally, we also analyze the basis set convergence for the MP2 correlation energy. We have selected the methane and 1,3,5-hexatriyne molecules for the analysis (with the smallest and largest exponents for the dRPAc energy). The obtained exponents for methane and 1,3,5-hexatriyne are \( \alpha = 2.88 \) and 2.83, respectively. The basis set convergence of the correlation energy is considerably faster for the MP2 correlation energy than for the dRPAc energy, and these exponents are much closer to the inverse cubic convergence. Furthermore, the difference between the exponents for different molecules is also much smaller for the MP2 correlation energy than for the dRPAc energy.
5.2. New dual-hybrid direct random phase approximation method [S4]

**New dual-hybrid method**

Obtaining good results for chemical reaction energies and barriers with the same functional is a challenge for semi-local density functional methods. We have selected two very small but representative test sets (sHC5 and BH6, see Appendix 0) to develop a new method for both reaction energies and barriers. The sHC5 test set contains the reaction energies of five reactions with small hydrocarbons. The BH6 test set contains the barrier heights of six hydrogen transfer reactions. As we develop a model chemistry to be useful for the largest molecules possible in the future, we use the relatively small ATZ basis set, which already gives consistent results. We search for an optimal non-self-consistent method in a dRPA-based double hybrid form (with parameters $a_X$ and $a_C$) using global hybrid reference orbitals (with parameter $a$). For both test sets, the optimization leads to only one accurate and precise dual-hybrid method with $a = 0.75$, $a_X = 0.75$ and $a_C = 1$ (denoted as dRPA75).

$$E_{XC}^\text{dRPA75} = \left(0.75E_X^{\text{exact}} + 0.25E_X^{\text{PBE}} + E_C^{\text{dRPA}}\right)_{\text{PBE}0.75}$$  \hspace{2cm} (83)

![Figure 16](image.png)

**Figure 16.** Optimal $a$ and $a_X$ exact exchange mixing parameter ranges of the dual-hybrid dRPA (with ATZ basis set) for the sHC5 (locally optimal MAD: lower legends; 0.02 kcal mol$^{-1}$ range of the locally optimal MAD: blue band) and BH6 (locally optimal MAD: upper legends; 0.2 kcal mol$^{-1}$ range of the locally optimal MAD: orange band) test sets.
The mixing semi-local PBE correlation with the fully nonlocal dRPA correlation does not lead to improvement, thus $a_C = 1$ is optimal. The overlapping optimal $a_{\lambda}$ parameter regions (Figure 16) are around the dRPA@PBE and dRPA75 methods, but dRPA75 has improved accuracy for both the reaction energies and barriers heights. Notice that the beneficial effects of reference orbitals with such large fraction of exact exchange were also described in the literature.\(^{135}\) Interestingly, the same fraction of exact exchange is optimal for the final energy evaluation. Also notice that the optimal $a_X = 0.75$ exact exchange mixing factor is much larger for dRPA75 than for the popular global hybrids (~25%) or double hybrids (~50%). In the following, we present the performance of this new approach on the test sets introduced above.

**Reaction energies**

First we tested our new method on the $n$-homodesmotic hierarchy of reactions (RC1: isogyric; RC2: isodesmic; RC3: hypohomodesmotic; RC4: homodesmotic; RC5: hyperhomodesmotic).\(^{223}\) The test set contains the reactions of 14 conjugated, 22 nonconjugated, and 2 cyclic hydrocarbons (38 reactions overall).

![Figure 17. Mean absolute deviations relatively to dRPA75, and the dRPA75 mean absolute deviations in legends (in kcal mol\(^{-1}\)) for the RC1-RC5 $n$-homodesmotic hierarchy of reactions (all calculations with ATZ basis set).](image-url)
The double hybrid functionals perform similarly to each other, so we discuss only the performance of the B2PLYP-D3(BJ) method (Figure 17). For the RC1 reaction class, dRPA75 is the best among the examined methods. According to the literature, the dRPA@TPSS/cc-pVTZ results are also considerably less accurate (MAD = 6.7 kcal mol\(^{-1}\)) than our dRPA75 results.\(^{178}\) The commonly used hybrid methods (like M06-2X and B3LYP) also yield large errors (MAD = 9-12 kcal mol\(^{-1}\)). For the RC2 reaction class, dRPA75 is the best again and dRPA@HF has similar performance. For comparison, the dRPA@TPSS and M06-2X methods yield about twice as large errors (MAD ≈ 2 kcal mol\(^{-1}\)). For the RC3 reaction class, the performance of dRPA75 is further improved, and the other examined methods have similar errors within 20%. For the RC4 and RC5 reaction classes, the mean absolute error of dRPA75 is in the insignificant range (<0.3 kcal mol\(^{-1}\)). For comparison, the global hybrid methods yield accurate results with surprisingly low precision (MAD > 2 kcal mol\(^{-1}\)). Note that the error of dRPA75 decreases monotonically from RC1 to RC5. Also note that the basis set error of the dRPA@HF correlation reaction energy is -3.9 kcal mol\(^{-1}\) for the RC1 test set, +1.1 kcal mol\(^{-1}\) for the RC2 test set, and negligible for RC3-5.

**Atomization energies**

Next we analyze the errors of the calculated atomization energies for the 38 hydrocarbons in the RC0 test set.\(^{223}\) The dRPA methods generally underestimates the atomization energies. (Note that part of the underbinding comes from the basis set error of ATZ, which is 14.4 kcal mol\(^{-1}\) on average.) The dRPA75 method is particularly precise for atomization energies (CSSD = 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. This means that dRPA75 atomization energy error is 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 RC\(n\) test sets discussed above.

**Table 4.** The atomic corrections (in kcal mol\(^{-1}\)) for the atomization energy calculation.

<table>
<thead>
<tr>
<th>Method</th>
<th>H</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>dRPA75</td>
<td>4.29</td>
<td>6.79</td>
</tr>
<tr>
<td>dRPA@PBE</td>
<td>1.44</td>
<td>15.12</td>
</tr>
<tr>
<td>dRPA@HF</td>
<td>2.89</td>
<td>13.75</td>
</tr>
<tr>
<td>B2PLYP-D3(BJ)</td>
<td>1.36</td>
<td>-2.56</td>
</tr>
<tr>
<td>DSD-BLYP-D3</td>
<td>1.22</td>
<td>-3.02</td>
</tr>
<tr>
<td>PWPB95-D3(BJ)</td>
<td>1.26</td>
<td>-2.51</td>
</tr>
</tbody>
</table>
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.”

However, due to the systematic nature of the atomization energy errors, an efficient correction is possible (Figure 18) by optimized atomic energies (CAE: corrected atomic energies; the correction are in Table 4). Excellent atomization energies can be obtained from the corrected dRPA energies. Contrarily, the imprecision of the double hybrid methods cannot be corrected as efficiently as the more systematic errors of the dRPA methods.

![Figure 18](image)

**Figure 18.** Corrected sample standard deviations and mean deviations for the RC0 atomization energies with double hybrid, dRPA and dual-hybrid dRPA methods (with ATZ basis set) with or without corrected atomic energies (CAE).

**Barrier heights**

The popular density functional 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. The application of many empirical
parameters might lead to improved results at the cost of a wavy energy surface (e.g. M06-2X functional).

**Figure 19.** Mean absolute deviations for the diverse barrier heights of the DBH24 test set with hybrid and double hybrid (with def2QZ basis set, taken from GMTKN30) as well as with dRPA and dual-hybrid dRPA (with ATZ basis set) methods.

Next we tested our new method for a test set of 24 diverse barrier heights (DBH24) with highly accurate (W3.2, W4 or CCSDT(Q)-full/CBS) reference values. The DBH24 test set is constituted by six hydrogen transfer barrier heights (BH6), six heavy atom transfer barrier heights (HAT6), six nucleophilic substitution barrier heights (NS6), and six unimolecular/recombination barrier heights (UR6). (Since the dRPA barrier heights are quite insensitive to the basis set size, we used the ATZ basis set for the calculations.) The M06-2X, dRPA75 and dRPA@PBE0 methods show an overall good performance for the DBH24 barrier heights (Figure 19). For comparison, the best range-separated RPA shows larger errors (MAD = 2.83 kcal mol$^{-1}$) for this test set. In addition, the considerably more expensive RPAX2@PBE method systematically overestimates the hydrogen transfer barrier heights (by ~1.7 kcal mol$^{-1}$) almost independently of the basis set size, which shows one of the limitations of this method.
**Intermolecular interactions**

Next we tested our new method on various types (charge transfer, $\pi-\pi$ stacking, hydrogen bonding, dipole-dipole, and weak interactions) of intermolecular interactions of the NCCE31 database. The dRPA75 method is much more accurate and precise for charge transfer interaction energies than any other method shown in Figure 20. It is also very accurate for $\pi-\pi$ stacking and hydrogen bonding interactions and quite competitive with the other methods for dipole and weak interactions. These results show that the dual-hybrid dRPA75 method gives well-balanced, accurate and precise interaction energies for all types of complexes in this test set (MAD $<$ 0.2 kcal mol$^{-1}$ for all subsets).

**Figure 20.** Mean absolute deviations for the non-covalent interaction energies of the NCCE31 test set with hybrid (with 6-311+G(3d2f,2df,2p) basis set), dispersion corrected hybrid and double hybrid (with def2QZ basis set), as well as dRPA and dual-hybrid dRPA (with ATZ basis set) methods.

**Intramolecular interactions**

The calculation of the intramolecular interactions is even more challenging than that of the intermolecular interactions because of the non-bonded electron density overlaps as we have pointed out previously by the Diels-Alder reaction energies. To test the performance of the methods for intramolecular interactions, we have selected the broadly-used test set for alkane (ACONF), cysteine (CYCONF), tripeptide (PCONF), and monosaccharide (SCONF).
conformers. The ACONF test set compares the relative energies of two butane, four pentane, and twelve hexane conformers. The CYCONF test set compares the relative energies of eleven cysteine conformer. The PCONF test set compares the relative energies of eleven phenylalanine-glycine-glycine tripeptide conformers. The SCONF test set compares the relative energies of fifteen 3,6-anhydro-4-\(O\)-methyl-D-galactitol and four \(\beta\)-D-glucopyranose conformers. For comparison, we have selected the M06-2X, B3LYP-D3 and B2PLYP-D3 methods. The performance of these functionals is known from the GMTKN30 database. The description of the intramolecular interactions in peptide conformers is challenging for the dRPA methods and not particularly good by M06-2X either (Figure 21). The dRPA75 dual-hybrid gives excellent results for the ACONF and PCONF test sets. However, the double hybrid B2PLYP-D3 performs better for the CYCONF and SCONF test sets. Notice that the 0.2 kcal mol\(^{-1}\) deviations are so small that they can be measurable with the error of the reference energies.

**Figure 21.** Mean absolute deviations for the relative conformational energies of the PCONF, SCONF, CYCONF and ACONF test sets with hybrid and double hybrid (with def2QZ basis set, taken from GMTKN30)\(^{137}\) as well as with dRPA and dual-hybrid dRPA (with ATZ basis set) methods.
6. Biological applications

6.1. Intramolecular interactions in immune recognition [S5]

Gas-phase geometries

First we reoptimized the former HF/6-31G(d) α- and β-Gal- or -ManNAc-Ser model structures (similar to Figure 1). The geometry optimizations (and also frequency calculations later) were performed with the B3LYP/6-31G(2df,p) method suggested in the Gaussian-4 theory with RIJCOSX approximation with automatically generated auxiliary basis set.

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 (cf. Appendix C) are similar to the former MM2* energies. 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 (the increasing numbering shows higher MM2* energies in ref 45) as they were also the most stable rotamers according to the previous HF results. 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 bonding patterns

In the calculated gas-phase geometries, intramolecular hydrogen bonds are formed in the absence of water. The (N2Ser)H...O1Ser hydrogen bond fixes the peptide backbone. The (N2Carb)H...O1Carb hydrogen bond is highly possible in α-Gal and β-Gal, moderately possible in β-Man and never occurs in α-Man. The (N1Ser)H...O1Carb hydrogen bond is frequent in α-Gal and β-Man and less frequent in β-Gal and α-Man. The (N2Carb)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 frequent than previously was considered. This finding agrees with the experimental results that the N2Carb-O2Ser distance fluctuates between ~3.2 and ~4.1 Å. 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.

The intramolecular hydrogen bonding patterns on the carbohydrate unit can be easily followed by torsion angles τ2-τ6. The antiperiplanar (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. The acetamido group is fixed by the (O3Carb)H...N2Carb hydrogen bond in almost all the conformers. 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, it can accept a hydrogen bond from the (N1Ser)H atom with the O7Carb atom.

The (O3Carb)H group can also point towards the O4Carb atom forming a hydrogen bond, or it can be free depending on the existence of the (O4Carb)H...O3Carb hydrogen bond. The (O4Carb)H atom is in this hydrogen bond in almost all the cases, but it can rarely form a hydrogen bond with the O6Carb or the O5Carb atom, or it can be free. If the (O4Carb)H atom binds to the O3Carb atom, there are three possible bonding situations (Figure 22a) for the (O6Carb)H atom with the O4Carb and O5Carb atoms. If the (O4Carb)H atom points towards the O5Carb atom, it reduces to two possibilities (Figure 22b) for the (O6Carb)H atom to bind to the O4Carb and O5Carb atoms. If the (O4Carb)H atom binds to the O6Carb atom, there is only one possibility (Figure 22c) for the (O6Carb)H atom to bind to the O5Carb atom. The (O6Carb)H atom can also form hydrogen bonds with the O1Ser and O2Ser atoms in some cases of the α-Man, β-Gal and β-Man conformers, and also it can be free.

Figure 22. Possible hydrogen bonds on the carbohydrate moiety of the α-GalNAc model with freezing orientation of the acetamido group by the (O3Carb)H...O7Carb hydrogen bond. For clarity, we omitted the (C)H atoms and closed the structure with a carbon atom at the glycosidic position. The two-two possible arrangements for the hydrogen bonds between (O6)H and O5, or between (O6)H and O4 are denoted with or without prime.

Structural flexibility

The flexibility of the glycosidic linkage can be characterized by torsion angles $\chi_1$-$\chi_3$. The torsion angle $\chi_3$ shows that in all the α conformers the peptide linkage is arranged in ap
conformation because of the steric strain in the axial position. In the β conformers, the ap conformation is the most frequent as well, but the (N2_Carb)H...O2_Ser interaction in β-Gal can result in −synclinal (−sc) conformation, and the (N1_Ser)H...O7_Carb interaction in β-Man can stabilize +anticlinal (+ac) conformation.

The torsion angle $\chi_2$ shows anti conformation in almost all the α-Gal conformers but varies more in the other conformers (mostly ap and +/−ac but also +/− sc in some cases). It suggests the structure along the C$\beta$-O1_Carb bond is more rigid in the α-Gal conformers than that in the other conformers. In β-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.

The torsion angle $\chi_1$ shows that the +sc conformation dominates in the conformers, but the −sc and ap conformations are also present. If the conformations are antiperiplanar along the C1_Carb-O1_Carb and O1_Carb-C$\beta$ bonds as it is in most of the α-Gal conformers, then the $\chi_1$ and $\psi$ 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 +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 and −115 degrees, and the torsion angle $\chi_1$ is between 35 and 65 degrees.

The preferred local conformation of the peptide backbone can be characterized by torsion angles $\phi$ and $\psi$ (Figure 23). The torsion angle $\phi$ shows that most of the conformations are +sc or −sc. 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 give 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.
**Figure 23.** Backbone conformation of the α- and β-Gal- or -ManNAc-Ser model structures in the absence of water and the monohydrated α-GalNAc-Ser model structures optimized with B3LYP/6-31G(2df,p) method (The φ/ψ torsion angle ranges for the possible secondary structures are denoted in the figure. The hydration process can be followed by the green arrows.)

**Electronic energy**

In order to determine the starting point of the hydration process, we need to identify the most stable α-Gal conformer. The MM2* method missed some possible hydrogen bonding patterns for the α-Gal conformer, hence we augmented the conformational space with eight new B3LYP gas-phase geometries. The electronic energies for the different hydrogen bonding patterns on the α-Gal unit were calculated with the dRPA75 and dRPA@PBE0 methods and the moderately large aTZ(-f,-d) basis set (*Table 5*). According to our earlier results, these methods can describe the energetics of hydrogen bonds with high accuracy. The relative energies calculated with these two methods agree within 0.5 kcal mol⁻¹. When the (O₃Carb)H atom binds to the O₇Carb atom, and the (O₄Carb)H atom binds to the O₃Carb atom, the (O₆Carb)H atom can bind the O₄Carb atom in two ways and the O₅Carb atom in one way. The most favorable is the (O₆Carb)H...O₅Carb' arrangement, but the (O₆Carb)H...O₄Carb' arrangement is also favorable. When the (O₃Carb)H atom still binds to the O₇Carb atom, but the (O₄Carb)H atom binds either the O₅Carb or O₆Carb atom, the (O₆Carb)H atom can bind the O₄Carb atom in one way and
the O5\textsubscript{Carb} atom in two ways. In the two most favorable situations, the (O6\textsubscript{Carb})H atom is binding to the O5\textsubscript{Carb} atom. The (O6\textsubscript{Carb})H...O5\textsubscript{Carb} is slightly favorable (0.5 kcal mol\textsuperscript{-1}) than the (O6\textsubscript{Carb})H...O5\textsubscript{Carb}' arrangement. This difference becomes more significant (2.0-2.1 kcal mol\textsuperscript{-1}) as the (O3\textsubscript{Carb})H atom releases the O7\textsubscript{Carb} atom and turns towards the O4\textsubscript{Carb} atom.

Table 5. Relative energies (in kcal mol\textsuperscript{-1}) of the α-GalNAcSer rotamers according to the hydrogen bonding pattern on the galactose part calculated with dRPA@PBE0 and dRPA75 methods and aTZ(-f,-d) basis set. (The new geometries are denoted as n1-8. The apostrophe signs an arrangement for the O6\textsubscript{Carb}H group in the ring plane.)

<table>
<thead>
<tr>
<th></th>
<th>(\Delta E)\textsubscript{dRPA@PBE0}</th>
<th>(\Delta E)\textsubscript{dRPA75}</th>
<th>(O3\textsubscript{Carb})H</th>
<th>(O4\textsubscript{Carb})H</th>
<th>(O6\textsubscript{Carb})H</th>
</tr>
</thead>
<tbody>
<tr>
<td>06</td>
<td>1.0</td>
<td>1.1</td>
<td>O7\textsubscript{Carb}</td>
<td>O3\textsubscript{Carb}</td>
<td>O4\textsubscript{Carb}'</td>
</tr>
<tr>
<td>n2</td>
<td>4.6</td>
<td>4.4</td>
<td>O7\textsubscript{Carb}</td>
<td>O3\textsubscript{Carb}</td>
<td>O4\textsubscript{Carb}</td>
</tr>
<tr>
<td>02</td>
<td>0.0</td>
<td>0.0</td>
<td>O7\textsubscript{Carb}</td>
<td>O3\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
</tr>
<tr>
<td>n3</td>
<td>6.9</td>
<td>7.0</td>
<td>O7\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
<td>O4\textsubscript{Carb}'</td>
</tr>
<tr>
<td>n1</td>
<td>4.7</td>
<td>4.8</td>
<td>O7\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}'</td>
</tr>
<tr>
<td>n6</td>
<td>4.2</td>
<td>4.3</td>
<td>O7\textsubscript{Carb}</td>
<td>O6\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
</tr>
<tr>
<td>n7</td>
<td>5.6</td>
<td>5.9</td>
<td>O4\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
<td>O4\textsubscript{Carb}'</td>
</tr>
<tr>
<td>n4</td>
<td>4.6</td>
<td>4.8</td>
<td>O4\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}'</td>
</tr>
<tr>
<td>09</td>
<td>2.6</td>
<td>2.7</td>
<td>O4\textsubscript{Carb}</td>
<td>O6\textsubscript{Carb}</td>
<td>O5\textsubscript{Carb}</td>
</tr>
</tbody>
</table>

Note that in most of the O-glycosidic core structures, the (O3\textsubscript{Carb})H atom is substituted by the second carbohydrate unit. Therefore, in these structures, the (O3\textsubscript{Carb})H...O7\textsubscript{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 counting from the glycosidic linkage.

**Structural water**

Adding a water molecule to the most stable α-GalNAc-Ser vacuum geometry helps the peptide chain to distance from the acetamido group (Figure 24). In the first step, the (N2\textsubscript{Carb})H...O2\textsubscript{Ser} hydrogen bond breaks during the torsion angles \(\chi_1\), \(\chi_2\) and \(\chi_3\) are changing by -12, 36 and -20 degrees, respectively. Then the (N2\textsubscript{Ser})H...O1\textsubscript{Ser} hydrogen bond breaks letting the peptide chain with the O2\textsubscript{Ser} atom to rotate along the C\textsubscript{α}-C2\textsubscript{Ser} bond by 48 degrees. This allows the relaxation of torsion angle \(\chi_1\). In the final step, the (O3\textsubscript{Carb})H and (O4\textsubscript{Carb})H groups rotate towards the O4\textsubscript{Carb} and O5\textsubscript{Carb} atoms forming new hydrogen bonds instead of the (O3\textsubscript{Carb})H...O2\textsubscript{Carb} and the (O4\textsubscript{Carb})H...O3\textsubscript{Carb} hydrogen bonds. Thus the acetamido group rotates along the C2\textsubscript{Carb}-N2\textsubscript{Carb} bond by -52 degrees and the peptide chain with the O2\textsubscript{Ser} atom rotates along the C\textsubscript{α}-C2\textsubscript{Ser} bond by 87 degrees guiding the water molecule towards the (N1\textsubscript{Ser})H atom.
When the water molecule approximates this hydrogen atom, they form a hydrogen bond. Therefore, the (N1Ser)H...O1Carb hydrogen bond breaks, and the peptide chain rotates along the N1Ser-Cα bond towards the water molecule. At the same time, the torsion angle χ₁ increases by 20 degrees, and the torsion angles χ₂ and χ₃ relax. During the three consecutive steps, the distance between the N2Carb and O2Ser atoms changes from 3.189 Å to 4.879, 4.637 and 3.816 Å, respectively.

**Figure 24** Monohydration of the most stable α-GalNAc-Ser model structure in gas phase, optimized with B3LYP/6-31G(2df,p) method (numbering from left to right: 02w1, 02w2, 02w3)

The torsion angles χ₁, χ₂ and χ₃ vary only in the angle range of the initial conformation. The changes in the torsion angles φ and ψ can be followed in **Figure 23**. If the (N2Carb)...O1Carb hydrogen bond exist, the peptide backbone prefers to be in a γ₁-turn. However, a well-defined structural water between the acetamido group and the peptide backbone (and probably also a water bridge between the (O3Carb)H and O7Carb atoms) 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. The varying torsion angle ranges for the peptide backbone and the steric effect of the sugar antenna suggest the O-glycosylated chains prefer to be in random loops, which agrees well with the experimentally observed extended structure of glycoproteins.36 38

Starting from the most stable gas-phase conformer and holding the (O6Carb)H atom in the (O6Carb)H...O5Carb arrangement, the electronic energy deepens by ~9 kcal mol⁻¹ with the addition of a single water molecule to the structure (**Table 6**) because two hydrogen bonds are formed. Although it makes the gas-phase electronic energy more positive (~4 kcal mol⁻¹), the breaking of the (N2Ser)H...O1Ser hydrogen bond is favorable in aqueous phase because the free
(N2_Ser)H and O1_Ser atoms can form one more hydrogen bonds with the solvent than the intramolecular hydrogen bonded ones. Additionally, there is another energetically almost equivalent hydrated structure. 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 further ~2 kcal mol⁻¹ lowering in the electronic energy.

Table 6. Relative energies (in kcal mol⁻¹) of the most stable non- and monohydrated α-GalNAcSer rotamers calculated with the dRPA@PBE0 and dRPA75 methods and aTZ(-f,-d) basis set. (The apostrophe signs an arrangement for the O6_CarbH group in the ring plane.)

<table>
<thead>
<tr>
<th></th>
<th>ΔE_dRPA@PBE0</th>
<th>ΔE_dRPA75</th>
<th>(N1_Ser)H</th>
<th>(N2_Ser)H</th>
<th>(N2_Carb)H</th>
<th>(O3_Carb)H</th>
<th>(O4_Carb)H</th>
<th>(O6_Carb)H</th>
</tr>
</thead>
<tbody>
<tr>
<td>02+w</td>
<td>0.0</td>
<td>0.0</td>
<td>O1_Carb</td>
<td>O1_Ser</td>
<td>O2_Ser</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
<td>O5_Carb'</td>
</tr>
<tr>
<td>02w1</td>
<td>-8.9</td>
<td>-8.6</td>
<td>O1_Carb</td>
<td>O1_Ser</td>
<td>O_Wat</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
<td>O5_Carb'</td>
</tr>
<tr>
<td>02w2</td>
<td>-4.4</td>
<td>-4.0</td>
<td>O1_Carb</td>
<td>-</td>
<td>O_Wat</td>
<td>O7_Carb</td>
<td>O3_Carb</td>
<td>O5_Carb'</td>
</tr>
<tr>
<td>02w3</td>
<td>-3.8</td>
<td>-3.5</td>
<td>O_Wat</td>
<td>-</td>
<td>O_Wat</td>
<td>O4_Carb</td>
<td>O5_Carb</td>
<td>O5_Carb'</td>
</tr>
</tbody>
</table>

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 the first step, the correction does not change so much.
6.2. Intermolecular interactions in epigenetic recognition [S6]

One-triad model structures

First we model the mCyt-Arg-Gua triads separately (cf. Figure 2b). The two smaller one-triad models are built up from the two bases and from the arginine terminated by methyl groups from both sides. All molecules positioned according to the experimental structure of MBD2-mDNA complex are shown in Figure 25 for the mCyt-Arg44-Gua and mCyt-Arg22-Gua triads. We have selected the M06L method (with the 6-31G(d) basis set) for geometry optimizations because this method is efficient and also includes some medium-range correlation effects.

Figure 25 Initial structures of the (a) mCpG-Arg44 and (b) mCpG-Arg22 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.

In the first geometry optimization, we model the interactions between a rigid DNA major groove and a rigid protein side chain and optimize only the positions of the terminal
hydrogen atoms. In the second geometry optimization, we simulate the flexible protein surface by optimizing the arginine atoms except the terminal methyl groups. In the third geometry optimization, simulate a flexible DNA major groove beside rigid sugar-phosphate backbone optimizing the atoms of the bases except for the terminal hydrogen atoms and the hydrogen atoms participating in hydrogen bonding. Finally, we simulate a flexible protein inner structure optimizing also the positions of the arginine terminal methyl groups.

Hydrogen bonding

The strengthening of the hydrogen bonds between the arginine side chain and the guanidine can be followed in Table 7 as the plane of the guanidinium group becomes coplanar with the plane of the guanine. The relaxation of the Arg22 guanidinium group requires only a flexible protein surface. The full relaxation of the Arg44 guanidinium group needs also a flexible protein inner structure. 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 7. Torsion angles (in degree) of the hydrogen bonds ($N\omega_{\text{Arg}}$-$O_{\text{Gua}}$-$N7_{\text{Gua}}$-$N\omega'_{\text{Arg}}$) in the methylated and non-methylated model structures. (Torsion angles between -8° and +8° are highlighted.)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg44</td>
<td>11.7</td>
<td>11.4</td>
<td>11.4</td>
<td>-1.5</td>
</tr>
<tr>
<td>CpG-Arg44</td>
<td>18.5</td>
<td>11.5</td>
<td>10.6</td>
<td>3.8</td>
</tr>
<tr>
<td>mCpG-Arg22</td>
<td>-8.4</td>
<td>-1.2</td>
<td>-6.1</td>
<td>4.6</td>
</tr>
<tr>
<td>CpG-Arg22</td>
<td>-8.4</td>
<td>-3.4</td>
<td>-7.4</td>
<td>-2.2</td>
</tr>
</tbody>
</table>

Cation-pi interaction

The strengthening of the cation-π interactions can be followed in Table 8. The interaction energy depends on the equilibrium distance between the cytosine ring and the guanidinium group, which can be determined also by other influences. The above mentioned distance is shorter in the methylated complex than in the non-methylated one.
Table 8. Distances (in Å) between the center (C<sub>ω</sub>Arg atom) of the positively charged guanidinium group and the center of the aromatic cytosine ring. (Distances under 5 Å are in bold.)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg44</td>
<td>5.21</td>
<td>4.84</td>
<td>4.69</td>
<td>4.42</td>
</tr>
<tr>
<td>CpG-Arg44</td>
<td>5.21</td>
<td>4.90</td>
<td>4.71</td>
<td>4.85</td>
</tr>
<tr>
<td>mCpG-Arg22</td>
<td>4.89</td>
<td>4.71</td>
<td>4.53</td>
<td>4.43</td>
</tr>
<tr>
<td>CpG-Arg22</td>
<td>4.89</td>
<td>4.77</td>
<td>4.57</td>
<td>4.60</td>
</tr>
</tbody>
</table>

Dispersion interaction

The dispersion interactions can be followed by the distance between the arginine side chain and the cytosine methyl group (Table 9). This distance decreases with the increasing flexibility of the protein surface but not with the increasing flexibility of the protein inner structure. Contrarily, the alkyl part of the arginine side chains gets even closer to the cytosine methyl group also with the increasing flexibility of the protein inner structure. Through the optimization, the alkyl part of the Arg44 side chain becomes closer to the corresponding cytosine methyl group than the Arg22 side chain to the other cytosine methyl group.

Table 9. 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></th>
<th>C&lt;sub&gt;ω&lt;/sub&gt;</th>
<th>C&lt;sub&gt;δ&lt;/sub&gt;</th>
<th>C&lt;sub&gt;γ&lt;/sub&gt;</th>
<th>C&lt;sub&gt;β&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>mCpG-Arg44</td>
<td>4.10</td>
<td>5.14</td>
<td>5.68</td>
<td>4.83</td>
</tr>
<tr>
<td>mCpG-Arg22</td>
<td>3.99</td>
<td>5.68</td>
<td>5.94</td>
<td>7.14</td>
</tr>
</tbody>
</table>

Electronic energy

The electronic interaction energy per triad (E(triad) – E(bases) – E(Arg)) is -42-43 kcal mol<sup>−1</sup> calculated with the RI-MP2/def2-TZVP(-f) and M06L/6-31G(d) methods on the final optimized geometries. The methyl group deepens the electronic interaction energy by ~1 kcal mol<sup>−1</sup> on the Arg22 binding site and by ~2 kcal mol<sup>−1</sup> on the Arg44 binding site.
Two-triad model structure

Our larger model contains both triads at the same time. It 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 Arg22 and Arg44 where the methyl groups positioned according to the first atom of the alkyl chains of the arginines, an acetate anion (AcO⁻) representing the Asp32; and two structural water molecules (using here the notations of Figure 2). All components are arranged as shown in Figure 26a.

![Figure 26](image)

**Figure 26** 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)

We optimized the geometry 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 compensate the usual overestimation error of semi-local density functional approximations. The
positions of the terminal methyl groups (closing the bases and the amino acid head groups) were frozen during the geometry optimizations to simulate the sugar-phosphate backbone and the protein surface. The two structural water molecules are fixed by hydrogen bonds in the model geometry (Figure 26a). We consider the hydration of the protein surface by seven explicit water molecules in a similar arrangement (Figure 26b). We also discuss a weakly and strongly hydrated case with six or seven explicit water molecules on the DNA (Figure 26c-d).

In the optimized geometry of the hydrated MeCP2 interacting site model, the water bridges among Arg22, Asp32 and Arg44 stretch two hydrogen bonded pentagons from the heteroatoms sharing one side, which rotates the plane of the guanidinium group of Arg44 towards Arg22 and Asp32. 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-Arg44 triad. The position of the other guanidinium group of Arg22 does not change so much because the Asp32 residue stiffens it. Through a single bridging water molecule, the heteroatoms of Asp32, Arg22, the attached guanine, and its cytosine base pair form again a hydrogen bonded pentagon coplanar with the bases.

**Thermochemical analysis**

We changed the B3LYP/6-31G(d) single-point energies in the B3LYP thermochemistry to significantly more accurate dRPA75/ATZ and dRPA@PBE0/ATZ single-point energies for non-covalent biomolecular interactions (Table 10).

<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>
</tbody>
</table>

In the presence of explicit water molecules, the electronic interaction energy is highly positive. 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 positive. The thermal corrections
to the energy and to the enthalpy slightly increase this value. Considering the entropy results in negative single-point energy corrected Gibbs energy. 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 single-point energy corrected Gibbs energies of the weakly and strongly hydrated models agree within 1 kcal mol\(^{-1}\) again.

**Recognition mechanism**

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 direction, when both CpG stairs are methylated (*Figure 27*).

![Diagram](image)

*Figure 27* Distance between the guanidinium carbon atoms of Arg22 and Arg44 (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.

The Arg22 group would form a larger barrier for sliding towards the 3'→5' direction with respect to the second DNA strand, and the Arg44 group would form a smaller barrier for sliding towards the 3'→5' direction with respect to the first DNA strand. The non-specific interactions
with the sugar-phosphate backbone and the steric hindrance for sliding support together the hopping mechanism for the recognition.

However, the interacting surface has some flexibility, which can modulate our recognition model. The Arg44 side chain is more flexible in comparison to the Arg22 side chain, which is strongly fixed by the Asp32 side chain with hydrogen bonding (locked state). The Arg44 guanidinium group can rotate towards the Arg22 guanidinium group (closed/open states). The smaller barrier is controlled by the hydrophobic effects. When the plane of the Arg44 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.
7. Summary

Our main new results can be summarized in the following thesis statements:

I. We proved that for the anion-π interactions, the MP2 reference interaction energies used in the literature are not suitable for benchmarking; therefore, we replaced the erroneous MP2 reference energies with accurate DLPNO-CCSD(T)/CBS interaction energies for the binary complexes. Based on this knowledge, we also suggested the accurate and efficient dRPA method with TZ basis set to benchmark semi-local density functional methods also on ternary π-anion-π’ sandwich and even larger complexes. We performed SAPT decomposition of the interaction energies, which showed that the exchange repulsion and the electrostatic attraction are the leading terms, but the induction and dispersion components of the interaction energies are not negligible either.[S1]

II. We disproved the claim in the literature that the DARC reaction energy test set is suitable to characterize the self-interaction error of semi-local density functional methods because the origin of the large endothermic error is the missing non-covalent, intramolecular dispersion interaction. We have shown that hybridization with the exact exchange simply shifts the calculated reaction energy errors by a constant in the exothermic direction thus it might improve the accuracy but not the precision. We have shown that hybridization and \textit{a posteriori} VV10, D2, or D3 dispersion corrections cannot reproduce reference reaction energies with the chemical accuracy and sufficient precision while the efficient dRPA can at moderate computational cost.[S2]

III. We have shown for hydrocarbons that the dRPA correlation energy converges considerably slower with the basis set size than the inverse cubic function suggested earlier in the literature. We have optimized the power exponent for different hydrocarbons and observed that it cannot be characterized by a single universal exponent for dRPA. We excluded the possibility of exponential basis set convergence. We developed different formulas which predict the exponent of the basis set convergence depending on the structure of the hydrocarbon in a way that it can be generalized for very large molecules. We developed an efficient QZ/TZ basis set extrapolation procedure, which yields comparable results with the 6Z/5Z extrapolation.[S3]

IV. We developed a unique dual-hybrid dRPA based method which is hybridized at two levels and called dRPA75. It uses the unconventional 75% exact exchange hybridized with PBE functional as orbital reference for dRPA correlation combined with the PBE hybrid
exchange. The dRPA75 method shows balanced performance for non-covalent interactions, and it is much more accurate and precise for charge transfer interaction energies than the other dRPA variants and double hybrid functionals. We obtained accurate and precise results from the dRPA75 method also for the five homodesmotic reaction classes, and for diverse barrier heights. In addition, we provided a correction for the dRPA atomic energies to get highly accurate hydrocarbon atomization energies.[S4]

V. We developed a methodology for calculating accurately the energetics of the O-glycosidic linkage of glycoproteins. We calculated the conformational space of α- and β-Gal- or -ManNAc-Ser model structures. The lowest energy conformers show several possible hydrogen bonding patterns on the first monosaccharide unit, which determine the orientation of the acetamido group, and key hydrogen bonds between the acetamido group and the peptide backbone through the glycosidic oxygen atom, which stiffen the glycopeptide linkage in gas-phase. We also calculated the conformational and energetic changes during the hydration process with an explicit structural water molecule, which shows that the peptide backbone prefers the expanded random coil structure over the more compact secondary structures suggested in the literature.[S5]

VI. We developed a methodology for calculating accurately the energetics of the methyl-CpG recognition by methyl-CpG-binding domain proteins. We calculated the supramolecular structure of methyl DNA – MBD protein model complexes, which revealed the different flexibility of the molecular units, and the steric hindrance for the protein to slide on the DNA double helix. We calculated the thermodynamics of the recognition process with different number of explicit structural water molecules on the interaction surface, which showed the hydrophobic interaction between the DNA cytosine methylation and the MBD arginine side chains largely contributes to the recognition by leading the arginine side chains to the neighboring guanine residues.[S6]
Related publications


S6 **Mezei PD, Csonka GI** (2016) Mechanism of methyl-DNA recognition by methyl-CpG-binding domain proteins. (submitted)
Acknowledgement

I am grateful to Prof. Gábor I. Csonka for supervising my work. I thank to Dr. Mihály Kállay for providing us the MRCC quantum chemical software, and for implementing my newly developed dRPA75 method into his code. This work was supported by the grants TAMOP-4.2.1/B-09/1/KMR-2010-0002 and TAMOP-4.2.2.B-10/1–2010-0009. We also thank to the Hungarian National Information Infrastructure Development Institute for the computer time.
References


Appendix

A. Main concepts in density functional theory

Density functional theory uses the three-dimensional electron density function instead of the multidimensional wave function, which is computationally beneficial. Furthermore, the electron density provides a direct link between theoretical and experimental physics, since it can be measured by X-ray diffraction in contrast to the wave function.

*Hohenberg-Kohn theorems*

The first Hohenberg-Kohn theorem states that the ground state density of an interacting system uniquely determines the potential and thus contains all the information about the system, so the many-body wave function does. To prove this theorem we need to show that the many-body wave function and the potential can be represented by the density. The former statement is called the N-representability and can be shown easier. The latter statement is called the V-representability, but it is not generally true. However, the Levy-Lieb constrained search algorithm avoids this problem for real systems. The second Hohenberg-Kohn theorem states that the exact ground state energy is the minimum of the universal functional, which is essentially the variational principle.

*Kohn-Sham equations*

According to density functional theory, there exists a fictitious non-interacting system which has the same ground-state electron density as the physical interacting system. This idea is very powerful because we can treat our system using simple one-electron eigenvalue equations, the Kohn-Sham equations.

\[
\left( -\frac{1}{2} \Delta + v_S(r) \right) \psi_i(r) = \epsilon_i \psi_i(r) \quad \text{(A-1)}
\]

\[
v_S(r) = v_{\text{ext}}(r) + v_H(r) + v_{\text{XC}}(r) \quad \text{(A-2)}
\]

The first term in the parenthesis is the kinetic energy operator, and the second term is the model potential of the non-interacting system, which is constituted by the external potential, the Hartree potential and the exchange-correlation potential. The electron density can be calculated from the occupied Kohn-Sham orbitals.

\[
n(r) = \sum_{i=1}^{N} |\psi_i(r)|^2 \quad \text{(A-3)}
\]
Energy functional

Generally, the $E[n]$ electronic energy functional can be written in terms of the $T[n]$ exact kinetic energy functional, the $V_{ee}[n]$ electron-electron potential energy functional and the $V_{ext}[n]$ external potential energy functional. The first two terms together are often called the $F[n]$ universal functional. The third term can be expressed as a local functional, since the external potential is a one-electron function.

$$E[n] = T[n] + V_{ee}[n] + V_{ext}[n]$$ (A-4)

$$T[n] = \langle \psi[n] | \hat{T} | \psi[n] \rangle$$ (A-5)

$$V_{ee}[n] = \langle \psi[n] | \hat{V}_{ee} | \psi[n] \rangle$$ (A-6)

$$V_{ext}[n] = \langle \psi[n] | \hat{V}_{ne} | \psi[n] \rangle = \int n(r) v_{ext}(r) d^3r$$ (A-7)

$$v_{ext}(r) = -\sum_{i=1}^{M} \frac{Z_i}{|R_i - r|}$$ (A-8)

In density functional theory, the universal functional is written in terms of the $T_S[n]$ non-interacting kinetic energy functional, the $U[n]$ Hartree energy functional, the $E_X[n]$ exchange functional and the $E_C[n]$ correlation functional. The non-interacting kinetic energy functional is the expected value of the kinetic energy operator on the Kohn-Sham orbitals. The Hartree energy functional is nonlocal, since the Hartree potential is a two-electron function. The sum of the Hartree and exchange energy functionals is the expected value of the electron-electron potential energy operator on the Kohn-Sham orbitals. The exchange-correlation potential can be derived from the exchange-correlation functional with differentiation.


$$T_S[n] = \langle \phi[n] | \hat{T} | \phi[n] \rangle$$ (A-10)

$$U[n] = \int n(r) v_H(r) d^3r$$ (A-11)

$$v_H(r) = \frac{1}{2} \int \frac{n(r')}{|r' - r|} d^3r'$$ (A-12)

$$E_X[n] = \langle \phi[n] | \hat{V}_{ee} | \phi[n] \rangle - U[n]$$ (A-13)

$$E_C[n] = T_C[n] + U_C[n]$$ (A-14)
Hole-model

A neat way to think about density functional theory is the probability theory of the electron-electron interaction. The average electron density function misses the true nature of the electrons that each electron digs a hole around itself in the electron gas because of the Coulomb repulsion (Coulomb hole) and the exchange interaction (Fermi hole). However, the two-particle density matrix expresses the probability of finding an electron in position \( r \) and another electron in position \( r' \). Recalling the Kolmogorov equation, the collective probability of the A and B events is defined by the probability of the A event times the conditional probability of A given B. The conditional probability of finding an electron in position \( r' \) given another electron in position \( r \) can be written as the sum of the electron density in position \( r' \) and the hole density in position \( r' \) of the exchange-correlation hole around position \( r \).

\[
P_2(r, r') = n(r) \cdot n(r') + n_{XC}(r' | r)
\]  

The exchange-correlation energy is the electrostatic interaction between the electron density and the exchange-correlation hole.

\[
E_{XC}[n] = \frac{1}{2} \int \int \frac{n(r) \cdot n_{XC}(r' | r)}{|r - r'|} d^3r' d^3r
\]  

The exact exchange-correlation hole has some formal properties. The hole density in position \( r \) of the exchange hole around position \( r \) (on-top hole) is as large as the electron density in the same position. Therefore, the exchange energy compensates the self-interaction part of the Hartree energy. The exchange hole is negative or zero everywhere. Furthermore, the exchange hole integrates to minus one, and the correlation hole integrates to zero.

Adiabatic connection approach

In density functional theory, the \( \lambda \) coupling constant (do not confuse with the coupling constant in many-body perturbation theory) expresses the strength of the interaction and links the non-interacting Kohn-Sham system (\( \lambda = 0 \)) with the interacting physical system (\( \lambda = 1 \)). The non-interacting Hamiltonian does not contain any potential energy operator, the interacting Hamiltonian contains full potential energy operator.

\[
\hat{H}^\lambda = \hat{T} + \lambda \hat{V}
\]  

The Schrödinger equation can be written for any interaction strength using the coupling constant.
The total energy is given by the Hellman-Feynman theorem. Similarly, the universal functional can be calculated using the electron-electron potential energy operator instead of the whole potential energy operator. The non-interacting part is equal to the non-interacting kinetic energy functional. Finally, the exchange-correlation energy functional can be expressed using the result obtained for the universal functional.

\[ H^\lambda \phi^\lambda = \varepsilon^\lambda \phi^\lambda \]  \hspace{1cm} (A-19)

\[ E[n] = E^{\lambda=0}[n] + \int_0^1 d\lambda \frac{dE^\lambda[n]}{d\lambda} = E^{\lambda=0}[n] + \int_0^1 d\lambda \langle \phi^\lambda[n] | \hat{V} | \phi^\lambda[n] \rangle \]  \hspace{1cm} (A-20)

\[ F[n] = F^{\lambda=0}[n] + \int_0^1 d\lambda \frac{dF^\lambda[n]}{d\lambda} = T_S[n] + \int_0^1 d\lambda \langle \psi^\lambda[n] | \hat{V}_{ee} | \psi^\lambda[n] \rangle \]  \hspace{1cm} (A-21)

\[ E_{XC}[n] = \int_0^1 d\lambda \langle \psi^\lambda[n] | \hat{V}_{ee} | \psi^\lambda[n] \rangle - U[n] = \int_0^1 d\lambda \frac{U^\lambda_{XC}[n]}{\lambda} = \int_0^1 d\lambda U_{XC}[n](\lambda) \]  \hspace{1cm} (A-22)
B. Small test sets for method development

**Table A-1.** Benchmark-quality estimated full CCSDT(Q)/CBS reaction energies (kcal mol\(^{-1}\)) and dRPA75 deviations from benchmark energies \((E - E_{\text{CBS}}^{\text{CCSDT}(Q)})\) of the sHC5 test set. The CBS extrapolations are based on the ATZ and AQZ energies. The dRPA75 calculations were performed with ATZ basis set. All calculations were performed on B3LYP/6-31G(2df,p) optimized geometries.

<table>
<thead>
<tr>
<th>System</th>
<th>(E_{\text{CBS}}^{\text{CCSDT}(Q)})</th>
<th>(\Delta E_{\text{ATZ}}^{\text{dRPA75}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C(_2)H(_4) + 2CH(_4) (\rightarrow) 2C(_2)H(_6)</td>
<td>-21.32</td>
<td>0.76</td>
</tr>
<tr>
<td>2 C(_3)H(_2) + 2CH(_4) (\rightarrow) C(_2)H(_4) + C(_2)H(_6)</td>
<td>-30.82</td>
<td>-0.05</td>
</tr>
<tr>
<td>3 C(_3)H(_8) + CH(_4) (\rightarrow) 2C(_2)H(_6)</td>
<td>2.08</td>
<td>-0.14</td>
</tr>
<tr>
<td>4 C(_3)H(_6) + CH(_4) (\rightarrow) C(_2)H(_4) + C(_2)H(_6)</td>
<td>4.81</td>
<td>-0.10</td>
</tr>
<tr>
<td>5 C(_4)H(_6) + 2CH(_4) (\rightarrow) 2C(_2)H(_4) + C(_2)H(_6)</td>
<td>13.06</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

\(\text{MD}^a\) 0.08  
\(\text{MAD}^b\) 0.23  
\(\text{CSSD}^c\) 0.38  
\(\text{Min}^d\) -0.14  
\(\text{Max}^e\) 0.76

\(^a\) Mean deviation  
\(^b\) Mean absolute deviation  
\(^c\) Corrected sample standard deviation  
\(^d\) Minimum  
\(^e\) Maximum

**Table A-2.** Benchmark quality estimated full CCSDT(Q)/CBS reaction barrier heights (kcal mol\(^{-1}\)) and dRPA75 deviations from benchmark energies \((E - E_{\text{CBS}}^{\text{CCSDT}(Q)})\) of the BH6 test set. The CBS extrapolations are based on the ATZ and AQZ energies. The dRPA75 calculations were performed with ATZ basis set. All calculations were performed on B3LYP/6-31G(2df,p) optimized geometries.

<table>
<thead>
<tr>
<th>System</th>
<th>(E_{\text{CBS}}^{\text{CCSDT}(Q)})</th>
<th>(\Delta E_{\text{ATZ}}^{\text{dRPA75}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 OH + CH(_4) (\rightarrow) [HOHCH(_3)]</td>
<td>6.02</td>
<td>-0.13</td>
</tr>
<tr>
<td>2 H(_2)O + CH(_3) (\rightarrow) [HOHCH(_3)]</td>
<td>19.46</td>
<td>-1.72</td>
</tr>
<tr>
<td>3 H + OH (\rightarrow) [HOH]</td>
<td>10.78</td>
<td>-0.19</td>
</tr>
<tr>
<td>4 H(_2) + O (\rightarrow) [HOH]</td>
<td>12.88</td>
<td>-0.57</td>
</tr>
<tr>
<td>5 H + H(_2)S (\rightarrow) [HSH]</td>
<td>3.78</td>
<td>0.82</td>
</tr>
<tr>
<td>6 H(_2) + HS (\rightarrow) [HSH]</td>
<td>16.97</td>
<td>1.33</td>
</tr>
</tbody>
</table>

\(\text{MD}^a\) -0.07  
\(\text{MAD}^b\) 0.79  
\(\text{CSSD}^c\) 1.07  
\(\text{Min}^d\) -1.72  
\(\text{Max}^e\) 1.33

\(^a\) Mean deviation  
\(^b\) Mean absolute deviation  
\(^c\) Corrected sample standard deviation  
\(^d\) Minimum  
\(^e\) Maximum
C. Structure and energetics of the O-glycopeptide models

**Table A-3.** Equilibrium single point energies, relative energies, the number of hydrogen bonds and torsion angles of the α-GalNAc-Ser model structures (We calculated eight new geometries compared to the reoptimized geometries.)

<table>
<thead>
<tr>
<th></th>
<th>$E_{\text{B3LYP}}^6-31G(2d,p)$ (hartree)</th>
<th>$\Delta E_{\text{B3LYP}}^6-31G(2d,p)$ (kcal mol$^{-1}$)</th>
<th>H-bonds</th>
<th>$\phi$</th>
<th>$\psi$</th>
<th>$\chi_1$</th>
<th>$\chi_2$</th>
<th>$\chi_3$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>-1314.692287</td>
<td>3.8</td>
<td>6</td>
<td>-83.31</td>
<td>67.29</td>
<td>-166.12</td>
<td>-176.42</td>
<td>167.43</td>
<td>-149.71</td>
<td>-161.23</td>
<td>65.95</td>
<td>-42.98</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>-1314.698278</td>
<td>0.0</td>
<td>7</td>
<td>-77.25</td>
<td>83.19</td>
<td>43.36</td>
<td>154.32</td>
<td>-159.16</td>
<td>168.16</td>
<td>-150.33</td>
<td>161.17</td>
<td>176.74</td>
<td>-57.15</td>
</tr>
<tr>
<td>04</td>
<td>-1314.689037</td>
<td>5.8</td>
<td>5</td>
<td>-83.68</td>
<td>66.66</td>
<td>42.22</td>
<td>-168.30</td>
<td>-179.04</td>
<td>168.52</td>
<td>-152.45</td>
<td>-159.74</td>
<td>164.67</td>
<td>-154.71</td>
</tr>
<tr>
<td>06</td>
<td>-1314.697100</td>
<td>0.7</td>
<td>7</td>
<td>-78.95</td>
<td>83.77</td>
<td>44.56</td>
<td>156.60</td>
<td>-162.58</td>
<td>173.28</td>
<td>-155.74</td>
<td>-160.69</td>
<td>-39.94</td>
<td>71.61</td>
</tr>
<tr>
<td>09</td>
<td>-1314.693942</td>
<td>2.7</td>
<td>6</td>
<td>-82.84</td>
<td>69.95</td>
<td>-55.14</td>
<td>153.62</td>
<td>-164.21</td>
<td>24.97</td>
<td>-40.26</td>
<td>19.72</td>
<td>70.06</td>
<td>53.81</td>
</tr>
<tr>
<td>10</td>
<td>-1314.695660</td>
<td>1.6</td>
<td>6</td>
<td>73.08</td>
<td>-44.67</td>
<td>-175.71</td>
<td>89.90</td>
<td>-162.74</td>
<td>160.96</td>
<td>-143.25</td>
<td>-160.98</td>
<td>68.18</td>
<td>-44.08</td>
</tr>
<tr>
<td>12</td>
<td>-1314.683527</td>
<td>9.3</td>
<td>5</td>
<td>56.19</td>
<td>40.87</td>
<td>24.61</td>
<td>157.41</td>
<td>-175.27</td>
<td>167.29</td>
<td>-150.14</td>
<td>-160.13</td>
<td>68.56</td>
<td>-43.77</td>
</tr>
<tr>
<td>13</td>
<td>-1314.685726</td>
<td>7.9</td>
<td>4</td>
<td>70.37</td>
<td>-74.12</td>
<td>179.29</td>
<td>-166.68</td>
<td>-173.51</td>
<td>164.35</td>
<td>-152.87</td>
<td>-161.24</td>
<td>-170.48</td>
<td>-65.16</td>
</tr>
<tr>
<td>15</td>
<td>-1314.691499</td>
<td>4.3</td>
<td>7</td>
<td>-84.88</td>
<td>79.20</td>
<td>47.05</td>
<td>140.40</td>
<td>-164.19</td>
<td>165.02</td>
<td>-135.03</td>
<td>38.65</td>
<td>64.34</td>
<td>53.52</td>
</tr>
<tr>
<td>16</td>
<td>-1314.692679</td>
<td>3.5</td>
<td>6</td>
<td>-77.19</td>
<td>83.52</td>
<td>43.77</td>
<td>157.64</td>
<td>-161.65</td>
<td>170.98</td>
<td>-153.11</td>
<td>-160.64</td>
<td>-171.96</td>
<td>68.30</td>
</tr>
<tr>
<td>n1</td>
<td>-1314.690276</td>
<td>5.0</td>
<td>7</td>
<td>-79.44</td>
<td>73.45</td>
<td>45.77</td>
<td>154.85</td>
<td>-155.46</td>
<td>165.60</td>
<td>-145.66</td>
<td>-42.97</td>
<td>177.04</td>
<td>-55.35</td>
</tr>
<tr>
<td>n2</td>
<td>-1314.692748</td>
<td>3.5</td>
<td>6</td>
<td>-83.32</td>
<td>67.05</td>
<td>42.81</td>
<td>-169.53</td>
<td>-175.22</td>
<td>168.70</td>
<td>-152.97</td>
<td>-160.01</td>
<td>66.17</td>
<td>-42.58</td>
</tr>
<tr>
<td>n3</td>
<td>-1314.688071</td>
<td>6.4</td>
<td>7</td>
<td>-80.27</td>
<td>74.62</td>
<td>44.22</td>
<td>157.46</td>
<td>-156.46</td>
<td>172.06</td>
<td>-153.51</td>
<td>-46.30</td>
<td>-25.61</td>
<td>62.63</td>
</tr>
<tr>
<td>n4</td>
<td>-1314.688585</td>
<td>6.1</td>
<td>6</td>
<td>-83.65</td>
<td>69.81</td>
<td>-55.41</td>
<td>158.87</td>
<td>-165.54</td>
<td>22.55</td>
<td>-39.15</td>
<td>-35.31</td>
<td>-179.28</td>
<td>-55.95</td>
</tr>
<tr>
<td>n5</td>
<td>-1314.687563</td>
<td>6.7</td>
<td>7</td>
<td>-84.51</td>
<td>79.16</td>
<td>46.40</td>
<td>144.51</td>
<td>-160.43</td>
<td>170.76</td>
<td>-152.07</td>
<td>-47.54</td>
<td>-29.80</td>
<td>64.99</td>
</tr>
<tr>
<td>n6</td>
<td>-1314.692085</td>
<td>3.9</td>
<td>7</td>
<td>-85.26</td>
<td>79.29</td>
<td>47.27</td>
<td>140.79</td>
<td>-163.04</td>
<td>164.63</td>
<td>-135.13</td>
<td>38.45</td>
<td>64.42</td>
<td>53.34</td>
</tr>
<tr>
<td>n7</td>
<td>-1314.686345</td>
<td>7.5</td>
<td>5</td>
<td>-83.16</td>
<td>69.63</td>
<td>-56.64</td>
<td>157.29</td>
<td>-163.43</td>
<td>23.29</td>
<td>-37.94</td>
<td>-39.74</td>
<td>-40.88</td>
<td>74.44</td>
</tr>
<tr>
<td>n8</td>
<td>-1314.689462</td>
<td>5.5</td>
<td>7</td>
<td>-80.89</td>
<td>81.84</td>
<td>46.18</td>
<td>145.66</td>
<td>-160.28</td>
<td>163.36</td>
<td>-139.52</td>
<td>-40.64</td>
<td>177.18</td>
<td>-55.62</td>
</tr>
</tbody>
</table>
Table A-4. Equilibrium single point energies, relative energies, the number of hydrogen bonds and torsion angles of the α-ManNAc-Ser model structures.

<table>
<thead>
<tr>
<th></th>
<th>$E_{6-31G(2df,p)}^{\text{B3LYP}}$ (hartree)</th>
<th>$\Delta E_{6-31G(2df,p)}^{\text{B3LYP}}$ (kcal mol$^{-1}$)</th>
<th>H-bonds</th>
<th>$\phi$</th>
<th>$\psi$</th>
<th>$\chi_1$</th>
<th>$\chi_2$</th>
<th>$\chi_3$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>-1314.688082</td>
<td>5.7</td>
<td>5</td>
<td>-82.70</td>
<td>62.13</td>
<td>47.01</td>
<td>130.95</td>
<td>-169.51</td>
<td>-168.00</td>
<td>-108.88</td>
<td>160.89</td>
<td>177.47</td>
<td>-54.70</td>
</tr>
<tr>
<td>02</td>
<td>-1314.685663</td>
<td>7.2</td>
<td>4</td>
<td>-75.00</td>
<td>96.24</td>
<td>-175.71</td>
<td>-167.67</td>
<td>178.82</td>
<td>-154.32</td>
<td>-151.72</td>
<td>165.46</td>
<td>-168.62</td>
<td>52.62</td>
</tr>
<tr>
<td>03</td>
<td>-1314.689308</td>
<td>4.9</td>
<td>6</td>
<td>-83.69</td>
<td>65.74</td>
<td>42.88</td>
<td>113.17</td>
<td>-161.01</td>
<td>-171.25</td>
<td>-103.49</td>
<td>167.43</td>
<td>-72.24</td>
<td>47.33</td>
</tr>
<tr>
<td>04</td>
<td>-1314.680808</td>
<td>10.2</td>
<td>5</td>
<td>74.08</td>
<td>-48.48</td>
<td>-171.73</td>
<td>-152.47</td>
<td>-169.03</td>
<td>-166.40</td>
<td>-110.30</td>
<td>161.02</td>
<td>177.90</td>
<td>-49.95</td>
</tr>
<tr>
<td>05</td>
<td>-1314.683577</td>
<td>8.5</td>
<td>6</td>
<td>-77.10</td>
<td>38.41</td>
<td>73.59</td>
<td>-141.72</td>
<td>-159.19</td>
<td>-167.53</td>
<td>-111.21</td>
<td>161.03</td>
<td>-178.39</td>
<td>-61.75</td>
</tr>
<tr>
<td>06</td>
<td>-1314.686850</td>
<td>6.4</td>
<td>5</td>
<td>-83.18</td>
<td>65.00</td>
<td>48.14</td>
<td>129.81</td>
<td>-178.80</td>
<td>-163.85</td>
<td>-132.47</td>
<td>168.29</td>
<td>64.95</td>
<td>55.99</td>
</tr>
<tr>
<td>07</td>
<td>-1314.681382</td>
<td>9.9</td>
<td>4</td>
<td>-80.60</td>
<td>49.24</td>
<td>71.23</td>
<td>-121.65</td>
<td>-151.62</td>
<td>-165.68</td>
<td>-114.40</td>
<td>161.48</td>
<td>-170.60</td>
<td>-174.77</td>
</tr>
<tr>
<td>08</td>
<td>-1314.687015</td>
<td>6.3</td>
<td>6</td>
<td>-125.20</td>
<td>-7.61</td>
<td>65.21</td>
<td>-86.66</td>
<td>-147.50</td>
<td>-166.85</td>
<td>-110.15</td>
<td>161.58</td>
<td>-171.41</td>
<td>179.68</td>
</tr>
<tr>
<td>09</td>
<td>-1314.682786</td>
<td>9.0</td>
<td>5</td>
<td>71.50</td>
<td>-55.48</td>
<td>-179.79</td>
<td>-164.11</td>
<td>-174.84</td>
<td>-173.67</td>
<td>-100.96</td>
<td>161.12</td>
<td>-74.16</td>
<td>50.36</td>
</tr>
<tr>
<td>10</td>
<td>-1314.677701</td>
<td>12.2</td>
<td>4</td>
<td>74.37</td>
<td>-51.89</td>
<td>138.30</td>
<td>-173.63</td>
<td>177.56</td>
<td>-164.38</td>
<td>-122.24</td>
<td>165.74</td>
<td>-168.97</td>
<td>57.82</td>
</tr>
<tr>
<td>11</td>
<td>-1314.684168</td>
<td>8.1</td>
<td>4</td>
<td>-83.24</td>
<td>63.13</td>
<td>46.33</td>
<td>127.65</td>
<td>-166.53</td>
<td>-169.80</td>
<td>-106.85</td>
<td>159.95</td>
<td>-167.78</td>
<td>179.52</td>
</tr>
<tr>
<td>12</td>
<td>-1314.683297</td>
<td>8.7</td>
<td>5</td>
<td>-84.92</td>
<td>62.84</td>
<td>43.14</td>
<td>121.12</td>
<td>-154.43</td>
<td>-165.09</td>
<td>-108.87</td>
<td>177.67</td>
<td>61.11</td>
<td>-179.34</td>
</tr>
<tr>
<td>13</td>
<td>-1314.686883</td>
<td>6.4</td>
<td>4</td>
<td>-80.63</td>
<td>90.64</td>
<td>-163.44</td>
<td>-145.94</td>
<td>-154.03</td>
<td>-155.04</td>
<td>-137.56</td>
<td>166.17</td>
<td>-160.13</td>
<td>-35.59</td>
</tr>
<tr>
<td>14</td>
<td>-1314.688407</td>
<td>5.5</td>
<td>4</td>
<td>-86.86</td>
<td>64.83</td>
<td>-53.83</td>
<td>-90.72</td>
<td>-160.68</td>
<td>-166.50</td>
<td>-108.93</td>
<td>166.15</td>
<td>-177.17</td>
<td>61.42</td>
</tr>
<tr>
<td>15</td>
<td>-1314.685652</td>
<td>7.2</td>
<td>5</td>
<td>-85.71</td>
<td>70.33</td>
<td>-59.06</td>
<td>-159.40</td>
<td>-174.49</td>
<td>-166.05</td>
<td>-112.76</td>
<td>169.36</td>
<td>-73.92</td>
<td>49.52</td>
</tr>
<tr>
<td>16</td>
<td>-1314.679662</td>
<td>10.9</td>
<td>3</td>
<td>-86.09</td>
<td>-2.96</td>
<td>78.52</td>
<td>-144.85</td>
<td>-163.00</td>
<td>-177.18</td>
<td>-99.24</td>
<td>157.26</td>
<td>-167.58</td>
<td>-174.44</td>
</tr>
<tr>
<td>17</td>
<td>-1314.687599</td>
<td>6.0</td>
<td>4</td>
<td>71.82</td>
<td>-52.53</td>
<td>-58.05</td>
<td>-65.12</td>
<td>-150.58</td>
<td>-178.62</td>
<td>-96.99</td>
<td>157.06</td>
<td>-165.23</td>
<td>-169.03</td>
</tr>
</tbody>
</table>
Table A-5. Equilibrium single point energies, relative energies, the number of hydrogen bonds and torsion angles of the β-GalNAc-Ser model structures

<table>
<thead>
<tr>
<th></th>
<th>$E_{6=31G(2df,p)}^{B3LYP}$ (hartree)</th>
<th>$\Delta E_{6=31G(2df,p)}^{B3LYP}$ (kcal mol$^{-1}$)</th>
<th>H-bonds</th>
<th>$\phi$</th>
<th>$\psi$</th>
<th>$\chi_1$</th>
<th>$\chi_2$</th>
<th>$\chi_3$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>-1314.695693</td>
<td>0.9</td>
<td>6</td>
<td>-81.72</td>
<td>55.39</td>
<td>-57.39</td>
<td>122.78</td>
<td>169.02</td>
<td>176.18</td>
<td>-150.67</td>
<td>-163.35</td>
<td>69.82</td>
<td>-49.68</td>
</tr>
<tr>
<td>02</td>
<td>-1314.693708</td>
<td>2.1</td>
<td>7</td>
<td>-82.83</td>
<td>57.66</td>
<td>56.74</td>
<td>-174.52</td>
<td>-58.65</td>
<td>146.58</td>
<td>179.73</td>
<td>-164.88</td>
<td>64.89</td>
<td>-44.70</td>
</tr>
<tr>
<td>03</td>
<td>-1314.696403</td>
<td>0.4</td>
<td>7</td>
<td>-81.96</td>
<td>56.51</td>
<td>38.19</td>
<td>68.44</td>
<td>162.47</td>
<td>-167.28</td>
<td>-169.07</td>
<td>-161.63</td>
<td>73.47</td>
<td>-49.57</td>
</tr>
<tr>
<td>04</td>
<td>-1314.692794</td>
<td>2.7</td>
<td>7</td>
<td>-82.81</td>
<td>60.40</td>
<td>55.75</td>
<td>-179.48</td>
<td>-57.05</td>
<td>-141.25</td>
<td>174.82</td>
<td>-165.59</td>
<td>-178.61</td>
<td>-51.39</td>
</tr>
<tr>
<td>05</td>
<td>-1314.684087</td>
<td>8.2</td>
<td>6</td>
<td>64.28</td>
<td>-56.16</td>
<td>72.87</td>
<td>-157.85</td>
<td>171.32</td>
<td>174.45</td>
<td>-148.51</td>
<td>-161.22</td>
<td>68.07</td>
<td>-49.41</td>
</tr>
<tr>
<td>06</td>
<td>-1314.693708</td>
<td>2.1</td>
<td>6</td>
<td>-81.73</td>
<td>73.10</td>
<td>177.18</td>
<td>-157.91</td>
<td>157.30</td>
<td>-175.36</td>
<td>-162.73</td>
<td>-160.08</td>
<td>64.51</td>
<td>-40.57</td>
</tr>
<tr>
<td>07</td>
<td>-1314.690259</td>
<td>4.3</td>
<td>5</td>
<td>76.59</td>
<td>-29.03</td>
<td>-61.14</td>
<td>132.80</td>
<td>-178.92</td>
<td>171.49</td>
<td>-159.85</td>
<td>-159.92</td>
<td>-179.29</td>
<td>-73.77</td>
</tr>
<tr>
<td>08</td>
<td>-1314.695236</td>
<td>1.2</td>
<td>7</td>
<td>-86.22</td>
<td>62.78</td>
<td>51.13</td>
<td>177.88</td>
<td>-62.18</td>
<td>-148.29</td>
<td>178.53</td>
<td>-163.83</td>
<td>-178.16</td>
<td>-55.32</td>
</tr>
<tr>
<td>09</td>
<td>-1314.690682</td>
<td>4.0</td>
<td>5</td>
<td>-83.46</td>
<td>56.25</td>
<td>-61.96</td>
<td>133.12</td>
<td>163.89</td>
<td>-155.01</td>
<td>-179.51</td>
<td>-163.14</td>
<td>-176.25</td>
<td>173.11</td>
</tr>
<tr>
<td>10</td>
<td>-1314.691359</td>
<td>3.6</td>
<td>5</td>
<td>-85.35</td>
<td>61.50</td>
<td>-169.19</td>
<td>-139.26</td>
<td>162.33</td>
<td>169.18</td>
<td>-154.95</td>
<td>-159.83</td>
<td>-176.09</td>
<td>-55.66</td>
</tr>
<tr>
<td>11</td>
<td>-1314.679967</td>
<td>10.8</td>
<td>6</td>
<td>61.74</td>
<td>-33.67</td>
<td>66.42</td>
<td>-158.77</td>
<td>169.14</td>
<td>-132.10</td>
<td>168.21</td>
<td>-165.82</td>
<td>-173.61</td>
<td>-58.13</td>
</tr>
<tr>
<td>12</td>
<td>-1314.688281</td>
<td>5.5</td>
<td>6</td>
<td>-84.07</td>
<td>50.79</td>
<td>62.24</td>
<td>172.91</td>
<td>-44.17</td>
<td>169.05</td>
<td>-146.54</td>
<td>-162.03</td>
<td>-164.67</td>
<td>-170.42</td>
</tr>
<tr>
<td>13</td>
<td>-1314.690406</td>
<td>4.2</td>
<td>6</td>
<td>69.45</td>
<td>-53.10</td>
<td>-168.98</td>
<td>-131.87</td>
<td>168.31</td>
<td>-159.71</td>
<td>-175.73</td>
<td>-161.41</td>
<td>65.85</td>
<td>-42.77</td>
</tr>
<tr>
<td>14</td>
<td>-1314.692740</td>
<td>2.7</td>
<td>7</td>
<td>-83.86</td>
<td>50.77</td>
<td>62.57</td>
<td>172.19</td>
<td>-44.69</td>
<td>168.52</td>
<td>-145.01</td>
<td>-162.20</td>
<td>-40.53</td>
<td>69.98</td>
</tr>
<tr>
<td>15</td>
<td>-1314.692734</td>
<td>2.7</td>
<td>5</td>
<td>-114.49</td>
<td>10.74</td>
<td>44.38</td>
<td>107.46</td>
<td>171.20</td>
<td>-170.76</td>
<td>-170.07</td>
<td>-162.95</td>
<td>-165.00</td>
<td>-52.48</td>
</tr>
<tr>
<td>17</td>
<td>-1314.695465</td>
<td>1.0</td>
<td>7</td>
<td>-82.89</td>
<td>51.70</td>
<td>38.42</td>
<td>127.74</td>
<td>-151.02</td>
<td>94.88</td>
<td>-44.24</td>
<td>22.15</td>
<td>64.00</td>
<td>62.43</td>
</tr>
<tr>
<td>18</td>
<td>-1314.688633</td>
<td>5.3</td>
<td>7</td>
<td>-83.76</td>
<td>49.12</td>
<td>59.02</td>
<td>-150.26</td>
<td>176.57</td>
<td>-20.99</td>
<td>-38.87</td>
<td>32.39</td>
<td>65.41</td>
<td>58.36</td>
</tr>
<tr>
<td>19</td>
<td>-1314.684943</td>
<td>7.6</td>
<td>5</td>
<td>64.91</td>
<td>-40.95</td>
<td>80.20</td>
<td>119.29</td>
<td>174.84</td>
<td>-147.03</td>
<td>178.57</td>
<td>-163.88</td>
<td>-179.74</td>
<td>-70.68</td>
</tr>
<tr>
<td>20</td>
<td>-1314.689119</td>
<td>5.0</td>
<td>6</td>
<td>69.43</td>
<td>-53.88</td>
<td>-168.84</td>
<td>-130.92</td>
<td>168.11</td>
<td>-152.95</td>
<td>178.71</td>
<td>-162.98</td>
<td>-176.60</td>
<td>-57.21</td>
</tr>
<tr>
<td>21</td>
<td>-1314.687174</td>
<td>6.2</td>
<td>4</td>
<td>62.62</td>
<td>32.61</td>
<td>-52.66</td>
<td>130.50</td>
<td>-178.33</td>
<td>-150.04</td>
<td>175.76</td>
<td>-163.06</td>
<td>179.86</td>
<td>-74.18</td>
</tr>
</tbody>
</table>
Table A-6. Equilibrium single point energies, relative energies, the number of hydrogen bonds and torsion angles of the β-ManNAc-Ser model structures

<table>
<thead>
<tr>
<th></th>
<th>$E_{B3LYP}^{6-31G(2df,p)}$ (hartree)</th>
<th>$\Delta E_{B3LYP}^{6-31G(2df,p)}$ (kcal mol$^{-1}$)</th>
<th>H-bonds</th>
<th>$\phi$</th>
<th>$\psi$</th>
<th>$\chi_1$</th>
<th>$\chi_2$</th>
<th>$\chi_3$</th>
<th>$\tau_2$</th>
<th>$\tau_3$</th>
<th>$\tau_4$</th>
<th>$\tau_5$</th>
<th>$\tau_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>-1314.685388</td>
<td>7.4</td>
<td>7</td>
<td>58.95</td>
<td>6.50</td>
<td>48.00</td>
<td>118.06</td>
<td>177.73</td>
<td>170.90</td>
<td>-92.82</td>
<td>161.67</td>
<td>-174.49</td>
<td>-50.14</td>
</tr>
<tr>
<td>03</td>
<td>-1314.686496</td>
<td>6.7</td>
<td>7</td>
<td>-78.33</td>
<td>28.84</td>
<td>47.34</td>
<td>136.42</td>
<td>-179.87</td>
<td>-168.40</td>
<td>-113.30</td>
<td>167.32</td>
<td>175.81</td>
<td>-64.81</td>
</tr>
<tr>
<td>04</td>
<td>-1314.691409</td>
<td>3.6</td>
<td>7</td>
<td>-85.03</td>
<td>76.87</td>
<td>50.11</td>
<td>154.07</td>
<td>-158.72</td>
<td>-161.94</td>
<td>-118.80</td>
<td>165.86</td>
<td>64.27</td>
<td>47.93</td>
</tr>
<tr>
<td>05</td>
<td>-1314.681414</td>
<td>9.8</td>
<td>5</td>
<td>-79.28</td>
<td>63.83</td>
<td>41.60</td>
<td>-175.95</td>
<td>144.98</td>
<td>-89.10</td>
<td>50.14</td>
<td>-82.80</td>
<td>61.13</td>
<td>59.84</td>
</tr>
<tr>
<td>06</td>
<td>-1314.687982</td>
<td>5.7</td>
<td>4</td>
<td>-85.35</td>
<td>56.04</td>
<td>-53.55</td>
<td>114.92</td>
<td>157.07</td>
<td>-167.17</td>
<td>-114.53</td>
<td>165.13</td>
<td>179.04</td>
<td>164.82</td>
</tr>
<tr>
<td>07</td>
<td>-1314.686697</td>
<td>6.5</td>
<td>5</td>
<td>74.23</td>
<td>-63.55</td>
<td>-64.20</td>
<td>135.05</td>
<td>176.05</td>
<td>176.49</td>
<td>-93.11</td>
<td>159.48</td>
<td>-172.92</td>
<td>-65.83</td>
</tr>
<tr>
<td>09</td>
<td>-1314.686711</td>
<td>6.5</td>
<td>4</td>
<td>-83.86</td>
<td>56.32</td>
<td>-47.92</td>
<td>119.10</td>
<td>165.49</td>
<td>-169.36</td>
<td>-110.82</td>
<td>166.31</td>
<td>63.73</td>
<td>-160.30</td>
</tr>
<tr>
<td>10</td>
<td>-1314.679162</td>
<td>11.3</td>
<td>5</td>
<td>-75.45</td>
<td>63.18</td>
<td>42.40</td>
<td>-167.39</td>
<td>128.31</td>
<td>-97.17</td>
<td>171.56</td>
<td>177.58</td>
<td>67.17</td>
<td>63.49</td>
</tr>
<tr>
<td>11</td>
<td>-1314.681903</td>
<td>9.5</td>
<td>5</td>
<td>-85.14</td>
<td>59.63</td>
<td>43.69</td>
<td>-144.82</td>
<td>142.89</td>
<td>-92.06</td>
<td>48.47</td>
<td>-49.64</td>
<td>56.66</td>
<td>176.40</td>
</tr>
<tr>
<td>12</td>
<td>-1314.687530</td>
<td>6.0</td>
<td>6</td>
<td>59.17</td>
<td>-19.93</td>
<td>67.13</td>
<td>122.12</td>
<td>-170.39</td>
<td>177.13</td>
<td>-92.80</td>
<td>156.94</td>
<td>70.98</td>
<td>76.98</td>
</tr>
<tr>
<td>13</td>
<td>-1314.680603</td>
<td>10.4</td>
<td>7</td>
<td>-79.61</td>
<td>37.59</td>
<td>38.77</td>
<td>129.14</td>
<td>-169.39</td>
<td>-85.66</td>
<td>49.41</td>
<td>-72.80</td>
<td>170.68</td>
<td>-65.57</td>
</tr>
<tr>
<td>14</td>
<td>-1314.683613</td>
<td>8.5</td>
<td>6</td>
<td>-82.72</td>
<td>62.01</td>
<td>39.29</td>
<td>-153.25</td>
<td>162.01</td>
<td>-177.36</td>
<td>-96.33</td>
<td>163.95</td>
<td>-175.43</td>
<td>-56.01</td>
</tr>
<tr>
<td>15</td>
<td>-1314.690481</td>
<td>4.2</td>
<td>5</td>
<td>-84.13</td>
<td>68.06</td>
<td>-65.99</td>
<td>116.34</td>
<td>172.68</td>
<td>-170.38</td>
<td>-106.37</td>
<td>162.36</td>
<td>172.36</td>
<td>65.73</td>
</tr>
<tr>
<td>16</td>
<td>-1314.678190</td>
<td>11.9</td>
<td>5</td>
<td>-76.87</td>
<td>47.07</td>
<td>56.85</td>
<td>-177.12</td>
<td>101.21</td>
<td>-90.55</td>
<td>171.08</td>
<td>171.62</td>
<td>-177.66</td>
<td>-51.81</td>
</tr>
<tr>
<td>17</td>
<td>-1314.683782</td>
<td>8.4</td>
<td>6</td>
<td>-81.16</td>
<td>70.85</td>
<td>40.39</td>
<td>-143.52</td>
<td>159.80</td>
<td>-176.33</td>
<td>-98.38</td>
<td>164.02</td>
<td>-65.24</td>
<td>42.50</td>
</tr>
<tr>
<td>18</td>
<td>-1314.690900</td>
<td>3.9</td>
<td>6</td>
<td>71.00</td>
<td>-50.93</td>
<td>169.68</td>
<td>87.47</td>
<td>-171.99</td>
<td>-167.66</td>
<td>-109.47</td>
<td>168.00</td>
<td>61.99</td>
<td>54.51</td>
</tr>
<tr>
<td>19</td>
<td>-1314.687337</td>
<td>6.1</td>
<td>5</td>
<td>-102.12</td>
<td>-7.28</td>
<td>58.36</td>
<td>127.34</td>
<td>-168.40</td>
<td>-169.95</td>
<td>-106.52</td>
<td>164.56</td>
<td>70.29</td>
<td>75.24</td>
</tr>
<tr>
<td>20</td>
<td>-1314.693207</td>
<td>2.4</td>
<td>5</td>
<td>-80.42</td>
<td>73.33</td>
<td>49.41</td>
<td>161.72</td>
<td>-158.85</td>
<td>-167.79</td>
<td>-106.46</td>
<td>164.12</td>
<td>-69.20</td>
<td>47.07</td>
</tr>
</tbody>
</table>

87
NYILATKOZAT

Alulírott MEZEI PÁL D ÁNIEL kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Budapest,……………………………...………………………………..