BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
FACULTY OF CHEMICAL TECHNOLOGY AND BIOTECHNOLOGY
GEORGE A. OLAH DOCTORAL SCHOOL

Density functionals applied to biomolecules and their non-covalent interactions

Ph. D. thesis summary

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

Department of Inorganic and Analytical Chemistry

2016

2.1. 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.

2.2. Background

The *O*-glycosidic linkage was previously investigated by molecular mechanical (MM) methods with or without constraints from nuclear magnetic resonance (NMR) measurements, and by quantum mechanical (QM) methods. The gas-phase geometry optimizations showed intramolecular hydrogen bonds on the first monosaccharide unit and between its acetamido group and the peptide backbone [1]. In solution, the geometry optimizations with explicit water molecules revealed water pockets on the first monosaccharide unit and water bridges between the acetamido group and the peptide backbone or the neighboring hydroxyl group [2].

The methyl-CpG recognition of methyl-CpG-binding domain proteins was investigated by mutational studies, NMR spectroscopy, as well as by MM and QM methods. Several highly conserved amino acid side chains play key role in the recognition. Most importantly, two arginine side chains bind the two guanines of the palindromic methyl-CpG motif. Also cation-π interactions were reported between the two guanidinium groups and the two cytosines [3]. Nature's preference for the methylated DNA was explained by hydrophobic pockets on the protein surface. Furthermore, important structural water molecules were identified on the interacting surface by X-ray diffraction (XRD) method [4].

For computing accurately and efficiently the non-covalent molecular interactions within the above mentioned biological systems, the possible density functional methods have to be tested first on representative organic and biomolecular benchmark test sets. One interesting problem is the anion-π interactions, which have been recently discovered in biological systems. A test set of 20 binary anion-π and 30 ternary π-anion-π' complexes with Møller-Plesset second-order perturbation theory (MP2) reference energies was proposed in the literature [5] with a small basis set and counterpoise correction to overcome the basis set superposition error. However, the basis set superposition often hastens the convergence, and the counterpoise correction usually overcorrects [6] 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.

The direct random phase approximation (RPA or dRPA) can potentially perform well for non-covalent molecular interactions. It contains all the MP2-like perturbative direct terms summed up to infinite order on the level of double excitations; therefore, it captures the non-pairwise nature of dispersion interactions. However, the dRPA correlation energy has some drawbacks. One problem is that it converges very slowly with the basis set size. To overcome this slow convergence, different complete basis set (CBS) extrapolation techniques have been proposed in the literature. The most popular technique assumes inverse cubic convergence [7], but this assumption is based only on studies of very small molecules. Another well-known problem is that the self-interaction error of the RPA correlation [8], which spoils the RPA interaction energies for charge-transfer complexes.

For testing the self-interaction or delocalization error of semi-local functionals, the DARC test set was suggested in the literature [9]. This test set contains the reference energies and geometries of 14 Diels-Alder reactions with mono-, bi- and tricyclic products. The authors concluded that the self-interaction error in the non-covalent electron density overlaps results in endothermic errors in the reaction energies. The simplest way to correct this error is mixing the semi-local exchange with exact exchange in global hybrid functionals [10].

References

1. Csonka, G. I.; Schubert, G. A; Perczel, A.; Sosa, C. P.; Csizmadia, I. G. *Chemistry* **2002**, *8*, 4718–4733.
2. Corzana, F.; Busto, J. H.; Jiménez-Osés, G.; Asensio, J. L.; Jiménez-Barbero, J.; Peregrina, J. M.; Avenoza, A. *J. Am. Chem. Soc.* **2006**, *128*, 14640–14648.
3. Zou, X.; Ma, W.; Solov'yov, I. A.; Chipot, C.; Schulten, K. *Nucleic Acids Res.* **2012**, *40*, 2747–2758.
4. Ho, K. L.; McNae, I. W.; Schmiedeberg, L.; Klose, R. J.; Bird, A. P.; Walkinshaw, M. D. *Mol. Cell* **2008**, *29*, 525–531.
5. Garau, C.; Frontera, A.; Quiñonero, D.; Russo, N.; Deyà, P. M. *J. Chem. Theory Comput.* **2011**, *7*, 3012–3018.
6. Mentel, Ł.; Baerends, E. *J. Chem. Theory Comput.* **2014**, *10*, 252–267.
7. Eshuis, H.; Furche, F. *J. Chem. Phys.* **2012**, *136*, 084105.
8. Mori-Sánchez, P.; Cohen, A. J.; Yang, W. **2009**, 4.
9. Johnson, E. R.; Mori-Sánchez, P.; Cohen, A. J.; Yang, W. *J. Chem. Phys.* **2008**, *129*, 204112.
10. Perdew, J. P.; Ruzsinszky, A. *J. Chem. Theory Comput.* **2010**, *6*, 3688–3703.

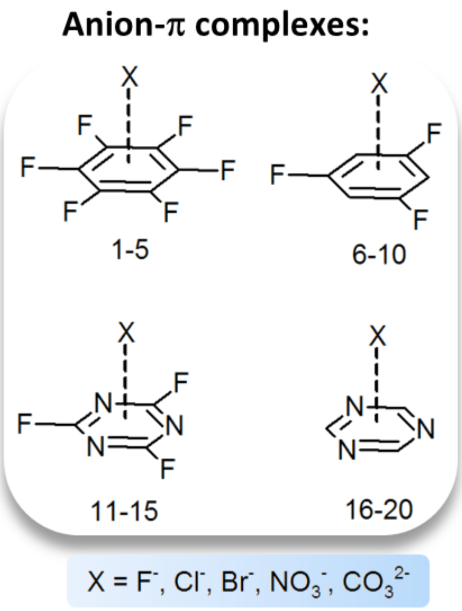
2.3. Computational methods

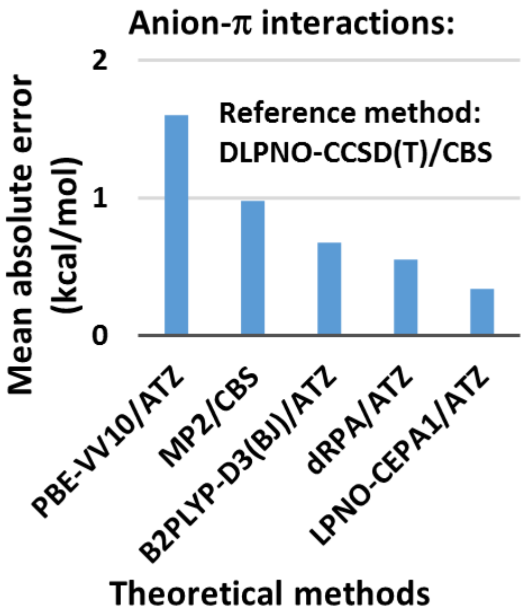
The coupled cluster method with single, double and perturbative triple excitations (CCSD(T)) proved to be a gold standard in main-group thermochemistry. The computations can be accelerated without any significant loss of accuracy in the relative energies by the domain-based local-pair natural orbital (DLPNO) method. A faster but somewhat less accurate approach is the coupled electron-pair approximation (CEPA), which is a size-extensive version of the configuration interaction method with single and double excitations (CISD) and can be accelerated preserving most of the correlation energy by the local-pair natural orbital (LPNO) method. A much more qualitative approach is the symmetry-adapted perturbation theory (SAPT). This theory provides a decomposition of the interaction energy between two molecules into exchange, electrostatics, induction and dispersion terms.

The semi-local density functional methods, *e.g.* PBE, have two well-known errors: the delocalization and dispersion errors. Hybrid functionals, *e.g.* PBE0, B3LYP or ωB97X-D, try to moderate the delocalization error with mixing of exact and semi-local exchange making the electron density more compact. The missing dispersion interactions are usually approximated by empirical dispersion corrections on the level of molecular mechanics with atom-pairwise dispersion corrections, *e.g.* D2, D3 or D3(BJ), or on the level of density functional theory with non-local dispersion corrections, *e.g.* VV10. Other approaches use flexible semi-local or hybrid functional forms with many adjustable parameters, *e.g.* M06L or M06-2X, to consider the midrange correlation effects, or double hybrid form with partial second-order perturbative correlation, *e.g.* B2PLYP or PWPB95, but such functionals also require some dispersion correction. The dRPA is an efficient method for non-covalent molecular interactions. The dRPA energy can be calculated using self-consistent Hartree-Fock (HF) or density functional reference orbitals. The second-order screened exchange (SOSEX) method corrects the self-correlation of RPA considering also the MP2-like perturbative exchange terms; however, it often underestimates the correlation at long range. The RPAX and RPAX2 methods use the exact exchange kernel, but they are considerably more expensive than dRPA.

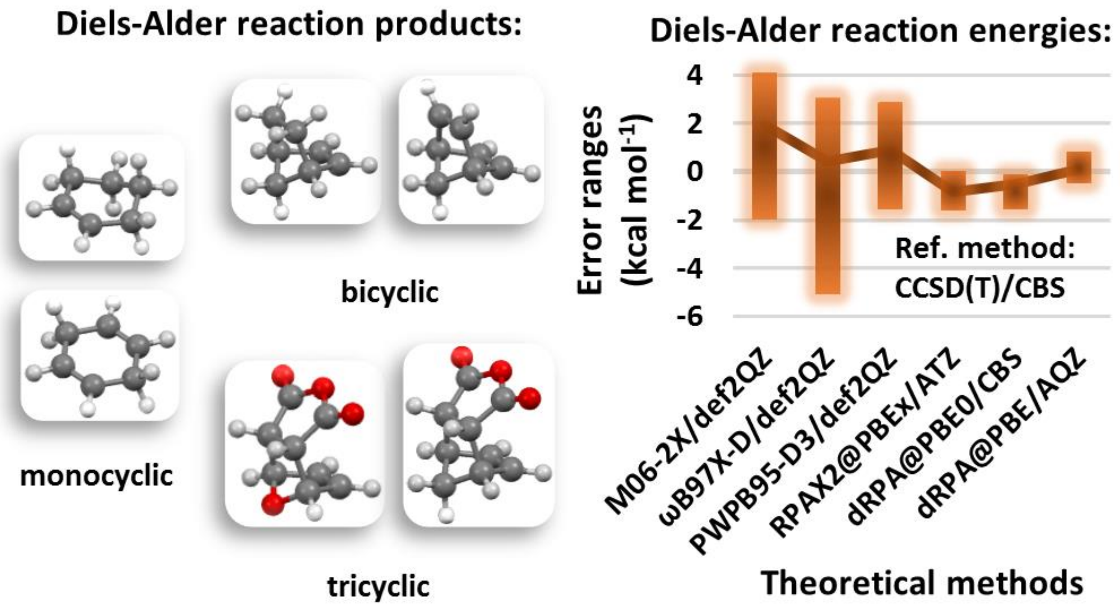
2.4. Results

For the anion-π complexes, we replaced the former MP2 reference with highly accurate DLPNO-CCSD(T) energies extrapolated to the complete basis set (CBS) limit. We confirmed the reliability of the reference method by independent LPNO-CEPA calculations with aug-cc-pVTZ basis set (ATZ). We performed SAPT decomposition on the interaction energy of the binary complexes, and concluded that the leading terms are the exchange repulsion and electrostatic attraction but the induction and dispersion interactions cannot be neglected either. We have shown that the efficient, dispersion corrected methods overestimate the overall interactions, thus they are wrong for anion-π complexes. The dRPA and RPAX methods provide accurate interaction energies. The MP2 method slightly overestimates, the SOSEX method slightly underestimates the interactions. We computed accurate reference interaction energies for the π-anion-π' complexes with the dRPA method. The deviations from the determined non-additivity of the interactions show the PBE-VV10 method significantly overestimates, the B2PLYP-D3(BJ) method slightly overestimates the induction.

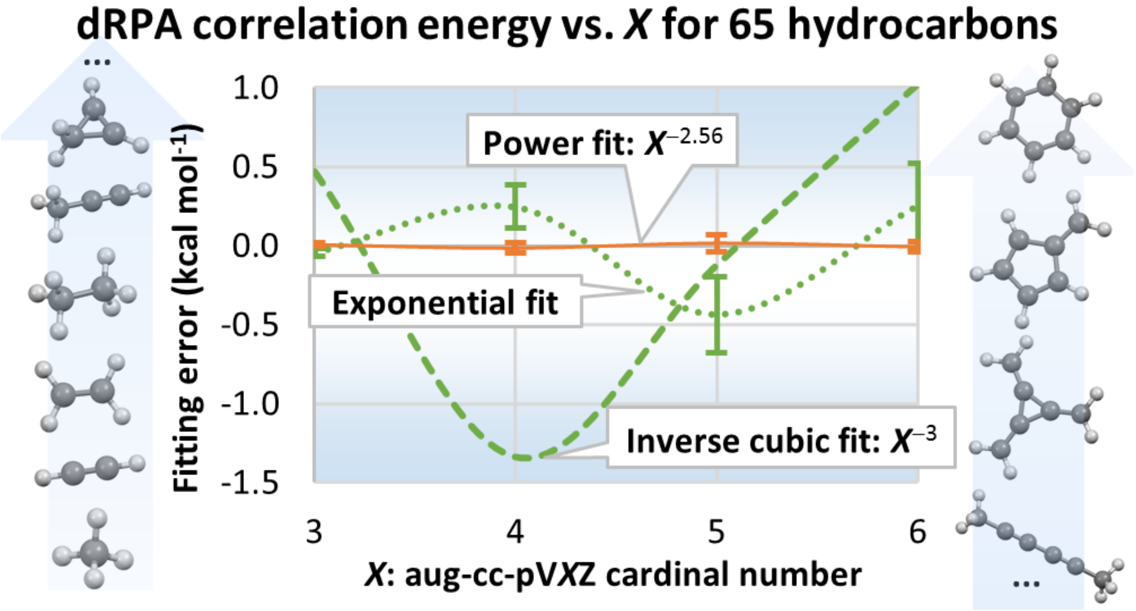




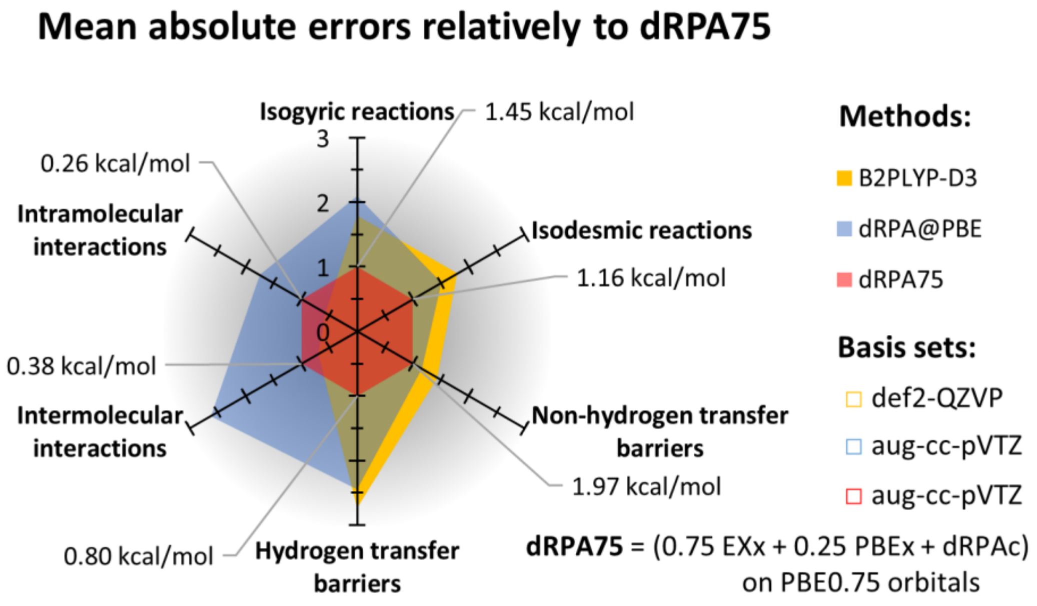
We analyzed the errors on several levels of approximate density functionals on the DARC energy test set. The semi-local functionals showed large endothermic errors especially for the reactions leading to tricyclic products. Originally the DARC test set was suggested to measure the delocalization error of semi-local functionals. To compensate the delocalization error, we tested different exact exchange ratios in the PBE global hybrid form. Our results show that the errors of semi-local functionals cannot be fully compensated by the global hybrid form because the increasing fraction of exact exchange simply shifts the reaction energies towards the exothermic direction by a constant. This suggests that another source of error is present. We identified this error source as the missing intramolecular dispersion interaction. To include also the dispersion interactions, we considered the M06-2X, ωB97X-D and PWPB95-D3 results with def2-QZVP basis set (def2QZ). These methods could not fully correct the dispersion error. The non-self-consistent dRPA methods provided better description of the reaction energies. However, using the more diffuse PBE electron density in dRPA@PBE method resulted in an endothermic error because of the increased exchange repulsion in the non-covalent overlapping regions. Using more compact PBE0 electron density in the dRPA@PBE0 method decreased the overlapping and the exchange repulsion energy but underestimated the dispersion energy. The more expensive RPAX2 method on PBE exchange reference yielded remarkably good results. We noticed that the small endothermic errors of the dRPA@PBE and RPAX2@PBEx methods are systematic enough to be well compensated by an exothermic basis set error.



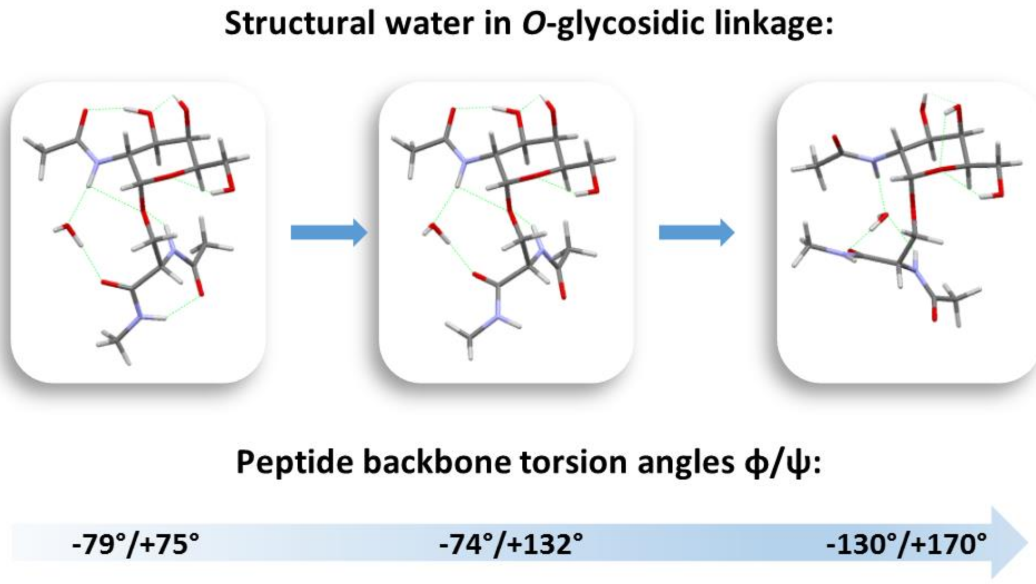
As we have shown, the dRPA correlation energy has excellent properties, but its basis set convergence was not properly addressed in the literature. Therefore, we developed an accurate extrapolation formula for the correlation energy considering 65 hydrocarbons from CH_4 to C_6H_6. The test set contained saturated, unsaturated, linear, branched, cyclic and aromatic isomers, whose geometries were available from the NIST database. We showed for the correlation consistent aug-cc-pV*X*Z (shortly A*X*Z) basis sets that the dRPA correlation energies converge with the increasing basis set cardinal number (*X*) slower than the previously assumed inverse cubic formula. The convergence can be described rather by a power fit with an average exponent of -2.56. Our analysis also revealed that the convergence depends on the atomic composition and structure of the molecules. We showed that the accuracy of the very expensive A6Z/A5Z two-point basis extrapolation scheme can be reached by a less expensive extrapolation scheme using A5Z and AQZ basis sets and corrections based on the chemical formula. To achieve this accuracy, we also developed an even much less expensive two-point extrapolation scheme using AQZ and ATZ basis sets and corrections using also the hybridization states of the carbon atoms in the molecule.



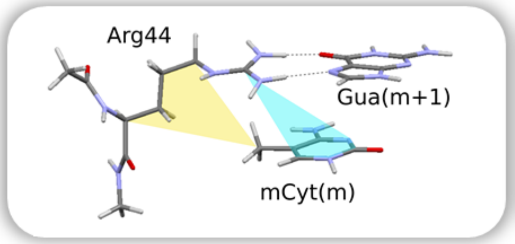
It is very challenging to develop an efficient method which can provide accurate non-covalent interaction energies, reaction energies and barrier heights at the same time. Only the very expensive CCSD(T)/CBS method is known to satisfy those requirements. It has been shown that dRPA can be implemented very efficiently and can correctly describe even the many-body effects needed for non-covalent interactions. But the performance of dRPA should be improved for reaction energies and barrier heights. We discovered that the dRPA correlation works very well on a PBE hybrid orbitals with 75% of exact exchange (PBE0.75). Two small and representative test sets for the hydrocarbon reaction energies and hydrogen-transfer barrier heights showed that the PBE0.75 exchange energy combined with the dRPA@PBE0.75 correlation energy, called dRPA75, might give satisfactory results even with the ATZ basis set. Then we tested the dRPA75 method on many accurately known reaction energies, barrier heights and non-covalent molecular interactions. dRPA75 shows higher accuracy for non-covalent intra- and intermolecular interactions than the non-self-consistent dRPA evaluated on PBE orbitals. Furthermore, it provides more accurate reaction energies and barrier heights than the standard dRPA and dispersion corrected double hybrid methods. The relative errors in the atomization energies shows that the dRPA75 method is very systematic and can be efficiently corrected for example by atomic corrections.

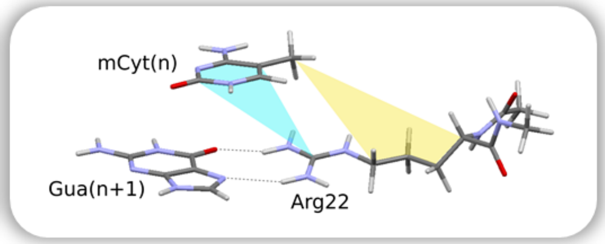


We modelled the *O*-glycosidic linkage with a serine residue closed by two methyl groups and glycosylated on the side chain by 2-acetylamino-2-deoxi-α- or -β-D-galacto- or –mannopiranose. We applied B3LYP method with polarized 6-31G(2df,p) basis set for the geometry optimization and thermal corrections, dRPA75 and dRPA@PBE0 methods with moderately large (aug)-cc-pVTZ(-f,-d) basis set for the single point energy calculation, an explicit structural water molecule for the hydration. The analysis of the conformational space shows that in contrast to the earlier assumptions, the direct hydrogen bonding between the acetamido group and the peptide backbone is less likely and the acetamido group can rotate freely beside an energetically equivalent hydrogen bonding pattern. The glycosidic linkage is stiffened rather by steric effects and water bridges. During the hydration process, a structural water molecule builds in, alters the hydrogen bonding pattern and the conformation. In the hydrated structure, the peptide backbone is preferred to be in extended random loops rather than in the compact γ_L-turn, PPII helix or β-sheet secondary structures.


Structural water in *O*-glycosidic linkage:
Peptide backbone torsion angles ϕ/ψ:
-79°/+75°
-74°/+132°
-130°/+170°

We analyzed the interactions between the methyl-CpG-binding domain proteins and the methylated DNA using two models. The first model contained a methylated or non-methylated cytosine, a guanine and an arginine residue closed by hydrogen atoms. We applied the M06L method with the small 6-31G(d) basis set for the geometry optimizations, and the MP2 method with the medium-sized def2-TZVP(-f) basis set and density fitting for the energy calculations. During the optimization, increasing the flexibility of the protein side chains, the DNA double helix, and peptide backbone corresponded to the strengthening of the hydrogen bonds, cation-π interactions, and dispersion interactions, respectively. The second model contained two-two methylcytosines, guanines, guanidinium groups, and an additional carboxyl group, all units closed by methyl groups. We applied the B3LYP method with the small 6-31G(d) basis set for the geometry optimizations and thermal corrections, the dRPA75 and dRPA@PBE0 methods with the moderately large (aug)-cc-pVTZ(-f,-d) basis set for the single point energy, and different number of explicit water molecules for the hydration. These calculations revealed that the hydrophobic interaction has the largest contribution to the Gibbs interaction energy and turns the arginine side chains into hydrogen bonding position. Our results show that the translation of the protein along the DNA double helix is sterically hindered by the contact of its arginine side chains with the methyl groups of the methyl cytosines. This supports a hopping mechanism for the protein movement along the DNA.





Interaction types:

- hydrogen bonds
- cation-π interaction
- dispersion interaction
- hydrophobic interaction
- steric effects

Interaction energy:

- -42-43 kcal mol^{-1} per triad

Effect of methylation:

- -1-2 kcal mol^{-1} per triad

2.5. 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 *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]

2.6. Applicability

The computed highly accurate DLPNO-CCSD(T)/CBS anion-π interaction energies can be applied as reference for benchmarking density functional methods. The even more efficient but less accurate dRPA/ATZ method can be applied for even larger anion-π systems too.

The new complete basis set extrapolation technique can be directly applied in the field of hydrocarbon chemistry for the efficient extrapolation of the dRPA correlation energy to the basis set limit. Our method can be used for developing dRPA CBS extrapolation schemes for wide range of compounds. The corrected atomic energies method can be directly applied in the field of hydrocarbon chemistry for highly accurate computation of atomization energies.

The new dRPA75 method can be applied in the field of organic and biomolecular chemistry for accurate, efficient and reliable computation of reaction energies, barrier heights, and non-covalent molecular interaction energies. This method has been implemented very efficiently in the MRCC quantum chemistry software, which is available for public use on the internet. Several groups in Hungary, Germany, Thailand, Japan, Australia and US started to use our method.

The new methodology for the glycosidic linkage can be applied also for the accurate computation of the interactions between the first and second monosaccharide subunits in the glycosidic core, and also for the selection of the lowest-energy conformations. The new methodology for the methyl-CpG recognition can be applied also for an extended model of the interacting surface.

2.6. Publications, presentations

Key publications:

S1 Mezei PD, Csonka GI, Ruzsinszky A, Sun J (2015) Accurate, Precise, and Efficient Theoretical Methods To Calculate Anion−π Interaction Energies in Model Structures. J Chem Theory Comput 11(1):360–371. doi: 10.1021/ct5008263 (IF=5.498; ID=5)

S2 Mezei PD, Csonka GI, Kállay M (2015) Accurate Diels–Alder Reaction Energies from Efficient Density Functional Calculations. J Chem Theory Comput 11(6):2879–2888. doi: 10.1021/acs.jctc.5b00223 (IF=5.498; ID=3)

S3 Mezei PD, Csonka GI, Ruzsinszky A (2015) Accurate Complete Basis Set Extrapolation of Direct Random Phase Correlation Energies. J Chem Theory Comput 11(8):3961–3967. doi: 10.1021/acs.jctc.5b00269 (IF=5.498; ID=3)

S4 Mezei PD, Csonka GI, Ruzsinszky A, Kállay M (2015) Construction and Application of a New Dual-Hybrid Random Phase Approximation. J Chem Theory Comput 11(10):4615–4626. doi: 10.1021/acs.jctc.5b00420 (IF=5.498; ID=4)

S5 Mezei PD, Csonka GI (2015) Unified picture for the conformation and stabilization of the O-glycosidic linkage in glycopeptide model structures. Struct Chem 26(5-6):1367–1376. doi: 10.1007/s11224-015-0666-9 (IF=1.837)

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

Other publications:

S7 Perdew JP, Sun J, Ruzsinszky A, Mezei PD, Csonka GI (2016) Why Density Functionals Should Not Be Judged Primarily by Atomization Energies. Period Polytech Chem Eng 60(1):2-7. doi: 10.3311/PPch.8356 (IF=0.296; ID=1)

Presentations:

1. Mezei PD, Csonka GI: Pontos, precíz és hatékony elméleti módszerek az anion-pi kölcsönhatási energiák számítására modell szerkezetekben. Conference of George A Olah Doctoral School, 2015/02/05, Budapest, oral presentation.
2. Mezei P D, Csonka GI: Accurate calculation of Diels-Alder reaction energies: The role of intramolecular interactions. KeMoMo-QSAR Symposium, 2015/05/14-15, Szeged, oral presentation.
3. Mezei PD, Csonka GI: Construction and Application of a New Dual-Hybrid Random Phase Approximation. 16th International Conference on Density Functional Theory and its Applications, 2015/08/31-09/04, Debrecen, poster.