NIEEMPIRYCZNE METODY W ANALIZIE I ELEKTROSTATYCZNYM MODELOWANIU ODDZIAŁYWAń W BIOCZąSTECZKACH

NONEMPIRICAL METHODS IN THE ANALYSIS AND ELECTROSTATIC MODELLING OF BIOMOLECULE INTERACTIONS

KAROL M. LANGNER

M.Sc. Solid State Physics, Wrocław University of Technology

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
under the supervision of Prof. W. Andrzej Sokalski

Institute of Physical & Theoretical Chemistry
Wrocław University of Technology

— 2010 —

Wrocław, Poland

Streszczenie

Niniejsza rozprawa przedstawia i ocenia nieempiryczne modele oddziaływań dla biocząsteczek, wykorzystując hybrydową metodę wariacyjno-perturbacyjną do analizy energii stabilizacji kompleksów. Człon elektrostatyczny tej energii jest najbardziej anizotropowym i najmniej kosztownym obliczeniowo, w związku z czym służy jako pierwsze przybliżenie. Badania koncentrują się na problemach dotychczas nie rozwiązanych w literaturze. Należy do nich opis oddziaływań typu π-π pomiędzy aromatycznymi cząsteczkami, zwłaszcza obliczenia dla warstwowych dimerów zasad kwasów nukleinowych. Chociaż oddziaływania tego typu są głównym tematem rozprawy, wykonano także analizy dla układów modelowych innych rodzajów.

Pierwszy rozdział opisuje podstawy metod perturbacyjnych i wariacyjnych używanych do podziału energii oddziaływania, porównując wyniki dla wybranych małych dimerów. Przegląd obejmuje też wielocentrowe rozwinięcia multipolowe, używane do szacowania oddziaływań elektrostatycznych. Omówione są tutaj kwestie zbieżności i problemy związane z zastosowaniem momentów multipolowych w symulacjach dynamiki molekularnej oraz w badaniach reaktywności chemicznej.

Następna część rozprawy wprowadza rangowe miary statystyczne, używane do ilościowej oceny efektów elektrostatycznych, jako predyktora całkowitej energii stabilizacji. Testy dla odległości mniejszych i większych od równowagowych pokazują, że wszędzie oddziaływania elektrostatyczne odtwarzają hierarchię całkowitych energii oddziaływania dla równowagowych geometrii. Takie wyniki mogą mieć praktyczne znaczenie między innymi w ocenie względnej stabilizacji leków zadokowanych przybliżonymi metodami do receptorów. Z kolei w przypadku dimerów zasad nukleinowych w ułożeniu warstwowym, składowa elektrostatyczna jest również w stanie w pewnym stopniu odtworzyć względną stabilność dimerów. Dalsza analiza ujawnia niespodziewaną korelację, pomiędzy wielkościami składowej dyspersyjnej i wymiennej. Jednak jakość tych wniosków statystycznych silnie zależy od jednorodności geometrii dimerów.

Ostatni rozdział poświęcony jest strukturze zawierającej cząsteczkę etydyny interkalującą kwas nukleinowy. Zarówno dokładne obliczenia, jak i multipolowe przybliżenie elektrostatyczne oddziaływań ligandu odtwarzają jego położenie krystalograficzne w płaszczyźnie interkalacji. Szczegółowa analiza weryfikuje podział układu na części i użycie przybliżenia dwuciałowego. Dodatkowo, oceniono wpływ poszczególnych fragmentów kwasu nukleinowego i ładunku na grupach fosforanowych oraz przeciwjonów i cząsteczek rozpuszczalnika.

Abstract

This dissertation explores and evaluates nonempirical interaction models for biomolecules, using an established hybrid variation-perturbation scheme as the underlying method for analyzing stabilization energies. Electrostatic effects are the most anisotropic and easily computable term in a hierarchy of quantum chemical interaction energies, and the first choice for an approximate interaction model. The practical essays here concentrate on molecules that have raised open questions in the literature, for it is in these cases that the physicochemical insight gained from *ab initio* results is most desired. Aromatic π-π interactions particularly attract attention, and special consideration is often given to nucleic acid bases. Although nucleobases are a major topic in this work, the methodologies are also applied to other model systems.

An introductory account summarizes perturbation and variational approaches to interaction energy decomposition including the hybrid method used throughout, and compares results for several small dimers. Atomic multipole expansions are then discussed with regard to estimating electrostatic effects. The disquisition focuses on convergence properties and on the applicability of Cartesian moments in molecular dynamics simulations as well as in studies of chemical reactivity.

The second chapter introduces rank-based statistical measures, used to quantify the predictive power of electrostatic effects. Benchmark calculations at shorter-than-equilibrium distances and in the long range substantiate the idea that electrostatic effects emulate relative equilibrium interaction energies at all distances. A similar analysis shows that the electrostatic component is able to reproduce to a certain degree the relative stability of stacked nucleobase dimers. Additionally, a surprisingly significant statistical relationship is revealed between the dispersion and exchange energies in this case. The quality of this association, however, depends strongly on the homogeneity of the studied structures.

In the last chapter, intercalators bound to nucleic acids are studied in detail. Interaction energy profiles are produced on the intercalation plane for various drug-RNA complexes, reproducing their original crystal binding positions. Further calculations are performed for a single ethidium-RNA complex in order to validate the reconstruction of interactions from pair-wise energies, and to evaluate the influence of nucleic acid backbone fragments. When phosphate groups are included, their charge state becomes an issue, therefore extreme neutral and charged cases are considered, as well as models with a counterion or solvent molecule. Estimates are given for the lower and upper bounds of the interaction energy in the case of ethidium intercalated between UA/AU base pairs.

Preface

> Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry – an aberration which is happily almost impossible – it would occasion a rapid and widespread degeneration of that science.
>
> Auguste Comte
> *Philosophie Positive* **1830**

It is hard to imagine a prediction for the course of chemistry that could have been farther from the truth. Mathematics *is* prominent in chemistry. Statistical thermodynamics and quantum mechanics are routinely used to explain atomic and molecular processes. Comte's warning from almost a century before the quantum revolution has a deeper but simpler interpretation – chemistry is more than the mathematical apparatus used to describe phenomena. Today we might add: more than the numerical methods and algorithms used to execute the mathematics. From his positivistic viewpoint, Comte is saying that chemical observations cannot be wholly expressed by concepts from physics and mathematics.

Fast forward one hundred years and Dirac states no less than the opposite. Chemistry has been reduced to physics and all that remains is to "clean up the details". There is no doubt what he had in mind – the mathematical laws that describe single atoms and their aggregates, molecules, implicitly govern all chemical phenomena and *in principle* can be used to model them to any accuracy. In hindsight, we see nearly another century in which physical theories and their computational implementations have been fruitful, although certainly more than a few details remain to be cleaned up.

These two stances define the conflict between reductionism and approaches that recognize chemical complexity. It is fitting to ask, is what we call chemistry an outdated label or does it offer anything unique? Something that sets it apart conceptually from the other natural sciences? Comte called it *spirit*.

The challenge recurs in recent years, as large parts of physics, chemistry and recently biology morph into a single body of research. Borderline journals and disciplines with names containing "physical chemistry" and "chemical biology" are typical manifestations. Befitting the trend, the chemistry Nobel committee now acknowledges work on green fluorescent protein, eukaryotic transcription and ribosomes, a turn duly noted and discussed by journal editors.[1]

Boundaries are increasingly blurred, but it remains to be seen if that is a green light for reductionism and for trimming chemistry to merely a collection of experimental techniques. Perhaps its practical bent has caused chemistry to be overlooked as an independent discipline

[1] Martens, E. *ACS Chem. Biol.* **2009**, *4*, 885; Mahapatra, A. *ACS Chem. Biol.* **2009**, *4*, 969–970.

in the philosophy of science,[2] which focuses on ontological questions in mathematics and on the interpretation of physical theories.

Meanwhile, chemistry remains a separate, strong field in institutionalized science, in culture and certainly in the popular perception of reality. Transformations of substances and bonding patterns, traditionally at the heart of the discipline, are usually first revealed by linking fundamental physical ideas with biology or materials science. This is especially true now, as pure research tackles life systems and is met halfway by biologists who identify the molecular processes. Does the middle ground lie in chemistry? The question if chemistry can be named "The Central Science" has been discussed at length, for example in a provoking essay by Balaban and Klein who also present relevant scientometric data.[3]

To be sure, these questions have been excogitated in all possible ways. A wary scientist will approach them with caution in fear of venturing onto the wrong, philosophical side of Popper's demarcation line. In fact, inspecting the frontiers of various disciplines and their evolution quickly turns into an exercise in science history, or worse in semantics if one is especially careless. And categorizing abstract ideas according to the theories or fields they are applied in digresses to a task as unfeasible as counting Plato's forms.

Such musings may seem nothing more than byproducts of reductionist or holistic efforts and generalizations in science. One should not be surprised, however, to hear them from an introspective computational *chemist*, who works by applying methods rooted in mechanics and quantum theory. Is it possible, by computations, to ask chemical questions? Do genuinely chemical ideas exist at all in the realm of molecules or can they, following Dirac, be fully reduced to physics? If the computational chemistry we practice is not conceptually redundant, is it a purely speculative science, or is it based on somehow unique observations?

Without further discourse, a point can be made by cherry-picking from computational chemistry's unique methodological memes – these may build upon input stolen from physics, but have little or no legitimate use outside the classical chemistry arena. The transition state is archetypal, because it does not represent one directly observable state and can be studied only at sub-picosecond timescales,[4] being at the same time seminal for the molecular description of reactions.

Less straightforward examples are atomic or molecular orbitals, mathematical constructs so utile and general-purpose that arguments persist about whether they are as real and tangible as the electron density.[5] Aromaticity and π-π interactions, and the various facets of molecular structure[6], these are additional cases that can be called upon to support the cause. Such ideas comprise the spirit of computational chemistry. The author humbly hopes to reinforce these uniquely chemical notions, by relating to them and assessing simplified interaction models based on electrostatic effects.

[2]Connections between chemistry and philosophy are discussed in Scerri, E. R. *J. Chem. Ed.* **2000**, *77*, 522.

[3]Balaban, A., Klein, D. *Scientometrics* **2006**, *69*, 615–637.

[4]Polanyi, J. C., Zewail, A. H. *Acc. Chem. Res.* **1995**, *28*, 119–132.

[5]For a summary of recent arguments, see Scerri, E. R. *J. Chem. Ed.* **2002**, *79*, 310.

[6]Current approaches to molecular structure are reviewed in Sukumar, N. *Found. Chem.* **2009**, *11*, 7–20.

This dissertation simply would not have been completed without the support and motivation provided by my family.

Many, many discussions with my supervisor, W. Andrzej Sokalski, have focused my interests and provoked a wider view of science and research.

For their orgnizational and financial support I am indebted to Wrocław University of Technology, Jackson State University, Basel University and Leiden University.

Wrocław Center for Networking and Supercomputing deserves my special gratitude for their generous allotment of computing time and professional technical assistance.

A number of teachers, colleagues and friends were influential in the past four years, and I should like to mention a few, alphabetically: Alexandra O. Borissova, Hansong Cheng, Agnieszka Dzielendziak, Edyta Dyguda-Kazimierowicz, Robert W. Góra, Tomasz Janowski, Paweł Kędzierski, Ludwik Komorowski, Bartłomiej Krawczyk, Jacek Małecki, Marcus Meuwly, Noel M. O'Boyle, Nuria Plattner, Kevin E. Riley, Szczepan Roszak, Andrzej Sadlej, Krzysztof Strasburger, Adam L. Tenderholt, Elżbieta Walczak.

Table of Contents

1 Introduction

> As I look at a living organism, I see reminders of many questions that need to be answered. Not all these questions are obviously important, nor would their answers be useful — but we want them answered.
>
> Linus Pauling
> *Nature of Forces between Large Molecules of Biological Interest*[7]

1.1 Context

When considering small molecules that are close to each other but not covalently bound, it is rudimentary to recognize the long tradition of experimental, theoretical and computational research already conducted. Much of the past work in the field of supermolecular complexes has been documented in books that deal with the underlying theoretical notions,[8] as well as the experimental applications.[9] Excellent review articles gather the newest facets of research,[10] and no such attempt will be made in this dissertation.

Instead, it is worthwhile to concentrate on one aspect of this tradition. In the course of scientific progress, separate driving forces had been identified that stabilize and repel molecules without forming permanent bonds. Among these, the most notable were named: hydrogen bonding, charge transfer, van der Waals or London dispersion forces. The distinct molecules and bond strengths involved were the first differentiation factors. With time, underlying common quantum mechanical aspects shared by these forces were revealed and used to understand and engineer new non-covalent complexes. Four fundamental contributions – **electrostatic, induction, dispersion and exchange** – can be derived using perturbation theory from the polarization approximation, with exchange effects included by enforcing symmetry.[11] These in turn can be combined in various proportions to compose all known types of intermolecular interactions.

Therefore it is logical, and chemically meaningful, to analyze intermolecular interaction energies as well as interaction-induced properties into terms related to these underlying con-

[7] Pauling, L. *Nature* **1948**, *161*, 707–709.

[8] Chapter 13 in Piela, L. *Ideas of Quantum Chemistry*; Elsevier, 2007.

[9] Ariga, K., Kunitake, T. *Supramolecular Chemistry*; Springer, 2006.

[10] Chałasiński, G., Szczęśniak, M. M. *Chem. Rev.* **1994**, *94*, 1723–1765; Chałasiński, G., Szczęśniak, M. M. *Chem. Rev.* **2000**, *100*, 4227 – 4252; Hobza, P., Zahradnik, R., Muller-Dethlefs, K. *Coll. Czech. Chem. Comm.* **2006**, *71*, 443–531; Schneider, H.-J. *Angew. Chem. Int. Ed.* **2009**, *48*, 3924–3977.

[11] Jeziorski, B., Moszyński, R., Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

tributions, instead of relying on imprecise traditional notions. That is not to say that terms such as 'hydrogen bond', 'van der Waals', 'hydrophobic', 'stacking' or 'charge transfer' should be abandoned altogether. To the contrary, they are valuable in chemistry for classification and quickly referencing typical complexes or interaction scenarios.

There have been substantial efforts to decompose interaction energies using perturbation theory with adapted symmetry (SAPT),[11] or variants of the variational method popularized by Kitaura and Morokuma.[12] Complementary studies on the technical aspects of these procedures have aimed to remove molecular orbital artifacts such as basis set superposition error or at least demonstrate and mitigate them in practice.[13] In particular, a hybrid variation-perturbation scheme has been formulated by Sokalski and coworkers that captures benefits from both approaches, striking a balance between computational cost and the practical utility of extracted terms.[14] All these decomposition schemes applied to small benchmark systems as well as some larger complexes with more complicated features. For examples, the reader is referred to Sections 2.2 and 2.3 and the works referenced there.

Alongside the maturing theoretical field of intermolecular interactions, the last decade has witnessed an unprecedented growth in available numerical methods and computational resources. Large basis sets, increasingly efficient MP2[15] and CCSD(T)[16] algorithms for molecules with more than a few atoms, all these have become commonplace. Accordingly, accurate interaction energies have been achieved for a number of non-covalent complexes, published as reference data sets by Hobza and coworkers[17] and others. Precise data are used as benchmarks to evaluate and eventually improve density functionals[18] or force fields potentials.[19]

Even so, quantum chemical treatment remains problematic for systems with more than about thirty atoms.[19] Since this is essentially below the threshold of any functional biomolecule, a feasible analysis of interactions in larger molecular systems remains both enticing and challenging. In the context of nucleic acids, the last two decades have seen an outburst of reports on the interactions between nucleobases,[20] concentrated more recently on stacking complexes.[21]

A class closely related to stacking are intercalation complexes, where an aromatic ligand enters the space between adjacent Watson-Crick base pairs, causing significant structural modifications to the polymer helix according to Lerman's original deduction.[22] The intercalation of nucleic acids is a still important topic, because it has been characterized thoroughly as a chemical process and has immediate practical significance in medicine.[23] From the computa-

[12]Kitaura, K., Morokuma, K. *Int. J. Quant. Chem.* **1976**, *10*, 325–340.
[13]Duijneveldt, F., Duijneveldt-van de Rijdt, J., Lenthe, J. *Chem. Rev.* **1994**, *94*, 1873–1885.
[14]Sokalski, W. A., Roszak, S., Pecul, K. *Chem. Phys. Lett.* **1988**, *153*, 153–159.
[15]Aikens, C. M., Webb, S. P., Bell, R. L., Fletcher, G. D., Schmidt, M. W., Gordon, M. S. *Theor. Chem. Acc.* **2003**, *110*, 233–253; Ishimura, K., Pulay, P., Nagase, S. *J. Comp. Chem.* **2006**, *27*, 407–413.
[16]Bartlett, R. J., Musiał, M. *Rev. Mod. Phys.* **2007**, *79*, 291–352.
[17]Jurečka, P., Šponer, J., Černý, J., Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
[18]Sherrill, C. D., Takatani, T., Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146–10159.
[19]Stone, A. J., Misquitta, A. J. *Int. Rev. Phys. Chem.* **2007**, *26*, 193–222.
[20]Hobza, P., Šponer, J. *Chem. Rev.* **1999**, *99*, 3247–3276.
[21]Šponer, J., Riley, K. E., Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.
[22]Lerman, L. S. *J. Mol. Biol.* **1961**, *3*, 18–&.
[23]Graves, D. E., Velea, L. M. *Curr. Org. Chem.* **2000**, *4*, 915–929.

tional point of view, well-defined non-covalent interactions are involved, and minimal models for intercalation complexes only recently have come within reach of first-principles methods that adequately include electron correlation.[24]

For hydrogen bonds and systems of charged molecules, the electrostatic component is known to be dominant and structure determining – *electrostatics* understood as the interaction of fragment charge densities unchanged by mutual influence. Many studies concerned with these and other types of interactions also highlight the role of electrostatic interactions; if not as a stabilizing factor then as a key element of molecular recognition at larger distances and as a predictor of relative stability.

Electrostatic interactions between molecules can be approximated asymptotically by expansion into a Taylor series, leading to a description of the electron density in terms of static multipole moments anchored on an expansion center. Such moments provide a compact and mobile representation of molecular electrostatic properties, and should be distributed among atomic centers near van der Waals equilibrium distances. This was first demonstrated in 1983 by Buckingham et al., with relatively precise predictions for angular orientations in non-covalent complexes.[25] It should be kept in mind that direct comparison to experimental electric moments beyond dipoles is limited to very small molecules,[26] and to molecular crystals based on densities obtained from diffraction measurements.[27]

Care should be taken when relating experimental findings to theoretical results, especially those of *ab initio* origin. In many cases, the Coulomb interaction of frozen electron densities does not constitute what is understood by the experimenter as *electrostatic effects*. Often, electronic induction and even dispersion forces are included, as opposed to hydrophobic and other collective effects.

In order to perform computational observations for large systems with finite resources, approximations in the description of interactions seem inevitable. As these resources grow, the smallest complex size for which approximations need to be made grows with them. Nonetheless, it is unimaginable that some day *any* molecular complex of interest, especially to biology, will be tractable using accurate methods derived from first principles.

Following Dirac's recommendation[28], the situation warrants increased interest in "frozen Fragment" electrostatic interactions as a means to modeling biomolecules. To this end, a series of questions are posed here, related to the role and predictive value of electrostatic interactions.

[24] Řeha, D., Kabeláč, M., Ryjáček, F., Šponer, J., Šponer, J. E., Elstner, M., Suhai, S., Hobza, P. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.

[25] Buckingham, A. D., Fowler, P. W. *J. Chem. Phys.* **1983**, *79*, 6426–6428; Buckingham, A. D., Fowler, P. W., Stone, A. J. *Int. Rev. Phys. Chem.* **1986**, *5*, 107–114.

[26] Buckingham, A. D. *J. Chem. Phys.* **1959**, *30*, 1580–1585.

[27] Guillot, B., Muzet, N., Artacho, E., Lecomte, C., Helsch, C. *J. Phys. Chem. B* **2003**, *107*, 9109–9121.

[28] Besides his often cited quantum mechanical reductionist manifesto in *Quantum Mechanics of Many-Electron Systems* (*Proc. Roy. Soc. London A* **1929**, *123*, 714-733), Dirac deemed it "desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation".

1.2 Purpose & overview

The variety of approximations adopted for molecular models can be divided into three broad categories. Parametrizations are employed in force fields and density functionals to describe interactions in a class of systems correctly on average, using accurate results as reference. The most attractive feature in this case is that a method, once parametrized, can be used repeatedly on a range of systems at reduced cost.

Consistent approximations can also be made with nonempirical methods, by dividing systems into fragments and considering inter-fragment interactions separately. This is the basic idea behind methods such as the fragment molecular orbital (FMO) approach,[29] which has been used to handle otherwise unfeasibly large molecules.[30] Lastly, incomplete interaction models that capture decisive energetic parts can be used to reproduce some chosen essential feature. This last possibility is always context-driven, since for various systems the features of interest depend on different effects. Generalization could be achieved by selectively using simplified interaction models based on overlap or distance criteria as proposed recently by Szalewicz and coworkers for the electrostatic component.[31]

The work presented here is concerned with the last two kinds of approximations, and mostly with the second type. With this in mind, it is the purpose of this dissertation to explore the simplest models of intermolecular interactions at the *ab initio* level and to attempt to quantify their practical limits. Electrostatic interactions derived from monomer charge distributions are central to this goal, since they comprise the computationally least demanding component and can be estimated from reusable multipole moments. Furthermore, since electrostatic interactions are dominant at large distances and more so for polar molecules, they are conspicuous in biological complexes.

The word *limits* used in the previous paragraph refers to the extent to which a model reproduces (predicts) interaction energies, relative stability or any other differentiating feature. Estimating such limits is crucial for drug and catalysis design, where it is common to screen a large array of candidate molecules for quantitative relationships.[32] A typical, practical question would be: with what confidence can one molecular complex be said to be more stable than another based on electrostatic interactions? If a number of complexes are analyzed and the energies of each pair compared, the sought confidence can be expressed in statistical terms, for example in the form of a prediction interval.

If an interaction energy is partitioned into physically meaningful terms, the significance of electrostatic effects or other components can be examined and various apposite approximate models can be suggested for the studied system. Here, special focus is given to π-π aromatic stacks, nonetheless complexes involving other types of interactions are also considered. Relevant aspects of the interaction profile not linked directly to electrostatics are also looked at – for example the spatial extent of a model in nucleic acid intercalation, the influence of

[29] Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701–706.
[30] Fedorov, D. G., Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904 – 6914.
[31] Rob, F., Podeszwa, R., Szalewicz, K. *Chem. Phys. Lett.* **2007**, *445*, 315–320.
[32] Karelson, M., Lobanov, V. S., Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027–1043.

solvent and counterions, nonadditive contributions, and statistical relationships between any two interaction energy components.

Logically, this dissertation is organized in three chapters. The first, Section 2, gives an account of the available interaction energy partitioning schemes based on variational and perturbation methods. This is complemented by **a comparison of the most widely used methods for several small dimers in a series of correlation consistent basis sets** in Section 2.3, lending basic observations and a reference point for subsequent discussions.

This chapter also goes into the utility of atomic multipole moment (AMM) expansions for estimating electrostatic interactions between molecules. The discussion concentrates around the Cartesian expansion used, and covers generating and transforming atomic moments, as well as how evaluating interactions. Sections 2.5 and 2.6 present a few **examples of the convergence properties of atomic moments and of the molecular electrostatic potentials and interactions they entail**. Conformational changes are considered, especially in the context of improving the representation of electrostatic interactions in molecular dynamics simulations and studying electron density changes during chemical reactions.

Section 3 starts with an explanation of statistical correlation measures based on ranks, along with several related original concepts. These are adapted to the subject at hand, that is comparing interaction energies between dimers. In recognition of the variable nature of different interaction components, the adopted approach is favored over the traditional Pearson correlation coefficient, which assumes linear relationships.

These statistical concepts are applied in Section 3 to two sets of complexes – one containing a number of small dimers that exhibit interactions typically found in biological systems, the second representing very specific classes of stacked nucleic acid bases. Both cases focus on the ability of electrostatic effects to reproduce the order of dimers with respect to their total interaction energies. The first, discussed in Section 3.2, is the S22 training set published by Hobza and coworkers[17] and extended by Fusti Molnar et al. by varying the separations between molecules.[33] The analysis presented here is based on the same geometries and uses the extrapolated CCSD(T) interaction energies from both studies as references.

Unbalanced interactions in dimers with short artifact contacts are addressed statistically, since these are frequently encountered in force field optimized geometries or structures uncorrected for basis set superposition error (BSSE). Concern for this problem has been expressed recently by Paton and Goodman in a critical review of force field performance for the same S22 training set.[34] A typical example of such error has been explicitly pointed out by Grzywa et al.,[35] who confirm that the geometry generated by a force field can be systematically misguided compared to the optimal MP2 geometry. For inhibitors docked in the urokinase active site with the Tripos force field, they find that the total MP2 interaction fails to correlate with experimental activity. On the other hand, electrostatic interactions manage to recover a large part of that correlation, although the statistical significance and

[33] Fusti Molnar, L., He, X., Wang, B., Merz, K. M. J. *J. Chem. Phys.* **2009**, *131*, 065102.
[34] Paton, R. S., Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944–955.
[35] Grzywa, R., Dyguda-Kazimierowicz, E., Sieńczyk, M., Feliks, M., Sokalski, W. A., Oleksyszyn, J. *J. Mol. Model.* **2007**, *13*, 677–683.

origin of this correlation are uncertain. A confirmation or explanation for this observation was the direct incentive to feature the study in Section 3.2.

At the other side of the scale, namely large distances, interactions are dominated by electrostatic effects, which may or may not have a bearing on the equilibrium stability. A postulated affirmative answer is the basis of hypotheses and methods that describe specific long-range events such as receptor-ligand recognition.[36] **The same section pursues these topics further within the S22 set, comparing interaction energies and their components at various intermolecular separations with the total interaction at equilibrium**.

Other cases exist where the aberrations described in the two previous paragraphs are legitimate problems. For instance, it is common procedure to optimize the geometry of a molecule or complex at a relatively inexpensive or semi-empirical level of theory (using a popular density functional, for instance) and subsequently employ a more accurate method to obtain just the energy. There has been very little discussion, however, how well the final energy relates to its counterpart at the "true" equilibrium of the accurate method, and for some systems these deviations may be significant. A more recognized issue is that of basis set superposition error, which has repeatedly proved its influence on the potential energy surface of intermolecular complexes. BSSE artificially strengthens interactions, which results in shortened intermolecular distances, especially for small or moderate basis sets.[37]

Section 3.3 introduces the computational literature on aromatic π-π stacking and applies the same rank-based statistical approach to stacked nucleobase dimers. For all combinations of B-DNA nucleobase pairs, interactions are analyzed into components and their statistical relationships assessed. **The objective of this study is to quantify the ability of electrostatic interactions to reproduce the total interaction energy**, extending earlier observations made in 2003 by Hill et al.[38]

Another, surprising correlation is obtained for the same B-DNA test set, between the attractive dispersion and repulsive exchange terms, which supports the first result by a partial cancellation of terms. A comparison for different sets of geometries (for example A-DNA versus B-DNA) and their unions reveals, however, that this relationship is highly sensitive to the geometrical homogeneity of the analyzed structures.[39]

The last part of this dissertation employs the methods introduced beforehand in order to answer specific questions related to nucleic acid intercalation motifs. One issue that is addressed is the consequence, in terms of the interaction energy, of breaking down a large molecular model into smaller dimers, with focus on one crystal structure of ethidium bound to RNA, $Eth^{(+1)}$-UA/AU, published by Jain and Sobell.[40] Calculations on extended systems

[36] Kier, L. B. *Pure Appl. Chem.* **1973**, *35*, 509–520; Kier, L. B., Höltje, H.-D. *J. Theor. Biol.* **1975**, *49*, 401–416; Hall, L. H., Kier, L. B. *J. Theor. Biol.* **1976**, *58*, 177–195; Kenny, P. W. *J. Chem. Inf. Model.* **2009**, *49*, 1234–1244.

[37] Simon, S., Duran, M., Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024–11031; Simon, S., Duran, M., Dannenberg, J. J. *J. Phys. Chem. A* **1999**, *103*, 1640–1643; Garden, A. L., Lane, J. R., Kjaergaard, H. G. *J. Chem. Phys.* **2006**, *125*, 144317–7; Shields, A. E., Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.

[38] Hill, G., Forde, G., Hill, N., Lester, W. A., Sokalski, W. A., Leszczyński, J. *Chem. Phys. Lett.* **2003**, *381*, 729–732.

[39] Langner, K. M., Sokalski, W. A., Leszczyński, J. *J. Chem. Phys.* **2007**, *127*, 111102.

[40] Nucleic Acid Databse ID: DRB018, Jain, S. C., Sobell, H. M. *J. Biomol. Struct. Dyn.* **1984**, *1*, 1161–1177.

probe **the influence of the nucleic acid backbone and chemical surroundings on the interaction of the intercalator with its host**, and specifically consider solvent molecules and counterions. Also, interaction energy profiles along an insertion path towards the major groove and on a grid in the intercalation plane show that electrostatic effects can be used to reproduce the crystal structure position.[41]

Two appendices describe selected technical aspects of the presented research. The first (Appendix A) is a description of the algorithms and code used to handle atomic multipole moments and their transformations. Appendix B is a selective overview of open source tools for automating computational chemistry tasks, concentrating on the parsing library cclib.[42]

Problems engaged in this work and its thematic scope were defined by relevant topics from the literature that are currently investigated or unresolved. The main questions can be summarized in five points:

1. Can a systematic study for small, model complexes reinforce the observation that, at artifactually shortened intermolecular separations, electrostatic effects correlate better with experimental results than the total interaction energy?

2. How well do long range interactions reproduce the total interaction strength at equilibrium, and how can this be measured statistically?

3. Do electrostatic interactions have a bearing on the stability of stacked nucleobases – if so, to what accuracy and in what circumstances can electrostatic interactions be used to reproduce or predict total interaction energies?

4. In what way can electrostatic interactions be used to understand the stability and position of an intercalator between the Watson-Crick base pairs of a nucleic acid strand?

5. An intercalator bound between nucleobases interacts strongly with the side chain phosphate groups as well as with surrounding counterions and solvent; how large is the influence of these groups and their various charge states?

In addition, two complementary methodological issues have been addressed:

- The basis set dependence of several interaction energy partitioning schemes, by comparing their particular components for small dimers,

- Convergence of interactions obtained from atomic multipole expansions, and the role of high ranks (above 4) in modeling charge redistribution during chemical reactions.

-

[41] Langner, K. M., Kędzierski, P., Sokalski, W. A., Leszczyński, J. *J. Phys. Chem. B* **2006**, *110*, 9720–9727.
[42] O'Boyle, N. M., Tenderholt, A. L., Langner, K. M. *J. Comp. Chem.* **2008**, *29*, 839–845.

2 First principles analyses of noncovalent interactions

> With courageous simplification, one might assert that the chemistry of the last century was largely the chemistry of covalent bonding, whereas that of the present century is more likely to be the chemistry of noncovalent binding.
>
> Hans-Jörg Schneider
> *Binding Mechanisms in Supramolecular Complexes*[43]

2.1 Introduction

Motivation for the above quote undoubtedly comes from the unprecedented amount of experimental and theoretical findings on noncovalent bonding in the last decades. In his review, Schneider points out that the utilization of intermolecular interactions is increasingly merging with historically dominant synthesis and structure characterization. He also stresses the importance of chelation and similar molecular constructs, where the sum of interactions for a number of centers in reversibly formed complexes leads to strengths comparable to those of single covalent bonds.

Reviewing the subject from a different angle, Hobza et al.[44] remark that *covalent* bonding is one of the most successful concepts of modern science, but in many respects can be considered a closed chapter after almost a hundred years of intense studies. In contrast, the grasp on noncovalent binding is not as firm and disagreements often exist between experimental and theoretical results. For example, although the ubiquitous hydrogen bond was suggested already by 1930, its properties and role in certain contexts are still unclear. The reason for this is arguably that, unlike for covalent bonds, the molecular environment has a strong influence and is not always accounted for.

It is undisputed that intermolecular interactions are an important factor for the properties of many molecular systems and substances. Although their current understanding and description is immature, especially for organic condensed phases,[43] the principle theoretical foundations have long been known.[45] Considerable progress is being made in understanding the nature of these interactions at the fundamental level, and *ab initio* theory has played a central role in this progress.[46]

[43] Schneider, H.-J. *Angew. Chem. Int. Ed.* **2009**, *48*, 3924–3977.
[44] Hobza, P., Zahradnik, R., Muller-Dethlefs, K. *Coll. Czech. Chem. Comm.* **2006**, *71*, 443–531.
[45] Chałasiński, G., Gutowski, M. *Chem. Rev.* **1988**, *88*, 943–962.
[46] Chałasiński, G., Szczęśniak, M. M. *Chem. Rev.* **2000**, *100*, 4227 – 4252.

Similar to other quantum chemical ideas, the accepted perception of intermolecular interactions relies on the Born-Oppenheimer separation of electronic and nuclear motions.[47] The assumption that the electronic part of the wave function relaxes instantaneously for any set of fixed atom positions leads directly to a potential energy surface (PES) as a multidimensional function of nuclear coordinates. Therefore, the quantum chemical interaction between molecules reflects the mutual relaxation of only their electronic degrees of freedom. The interaction energy E_{int} for a system of N components is in fact typically defined as a function of fixed nuclear coordinates $\mathbf{R}_i$ with i spanning all fragments,[48]

$$E_{\text{int}} = E(\mathbf{R}_1, .., \mathbf{R}_N) - \sum_{i=1}^{N} E_i(\mathbf{R}_i). \tag{2.1}$$

The Hamiltonian for the entire system yields the eigenvalue E, and E_i are eigenvalues for the Hamiltonians of isolated components, $\hat{H}_i$. The latter correspond to eigenstates of the time independent Schrödinger equations $\hat{H}_i\psi_i = E_i\psi_i$.

This definition can be extended with nuclear degrees of freedom by including the energies needed to deform components from a chosen dissociation channel $\mathbf{R}_i^0$ to their geometries in the interacting complex. These are the deformation energies, $E_{\text{def}}(\mathbf{R}_i^0 \rightarrow \mathbf{R}_i)$, and together with the interaction energy they comprise the binding energy E_{bind},

$$E_{\text{bind}} = E_{\text{int}} + \sum_{i=1}^{N} E_{\text{def}}(\mathbf{R}_i^0 \rightarrow \mathbf{R}_i) = E(\mathbf{R}_1, .., \mathbf{R}_N) - \sum_{i=1}^{N} E_i(\mathbf{R}_i^0). \tag{2.2}$$

The binding energy understood in this way depends on two different sets of coordinates, $\mathbf{R}_i$ and $\mathbf{R}_i^0$, which are not functionally related. For fixed dissociation channels, the binding energy is equivalent to the total energy, differing by a constant, and for that reason it is often given as a single number relative to a chosen energetic minimum or equilibrium geometry. Furthermore, in order to compare directly with experiment, the binding energy needs to be complemented by the difference in zero vibration energies between the complex and dissociated components. The final sum is often termed the *dissociation energy*, because it corresponds to the experimental energy needed to dissociate the complex.

From a phenomenological viewpoint, the interaction energy defined by (2.1) may seem unreasonable in excluding fragment relaxation. It is physically impossible to keep an isolated fragment in the geometry it assumes in the complex ($\mathbf{R}_i$) without exerting extra force. Conversely, it is even less probable for the electron wave function of a fragment to be in its isolated state (ψ_i^0) when embedded in the complex. In reality, the Born-Oppenheimer approximation dictates that the electronic wave function assumes an eigenstate of the Hamiltonian immediately after the position of any nuclei changes. When the nuclei of a molecule move by infinitesimal distances, the electronic part of its wave function adiabatically goes through a corresponding sequence of eigenstates, much like in a quasistatic thermodynamic process.

[47] Born, M., Oppenheimer, R. *Annalen der Physik* **1927**, *84*, 457–484.
[48] See p.684 in Piela, L. *Ideas of Quantum Chemistry*; Elsevier, 2007.

This idea is depicted in Fig. 2.1 by a generic distance dependence of energy. The thick, solid curve represents the total energy if the entire system is relaxed at various stages of association. The hashed line in turn represents the procedure of deforming fragments at infinite separation to their final geometries, moving them closer and turning on the interaction only at the equilibrium separation. Strictly speaking, the two are equivalent with respect to a constant, since all the energies discussed are potential energy functions associated with a conservative force.

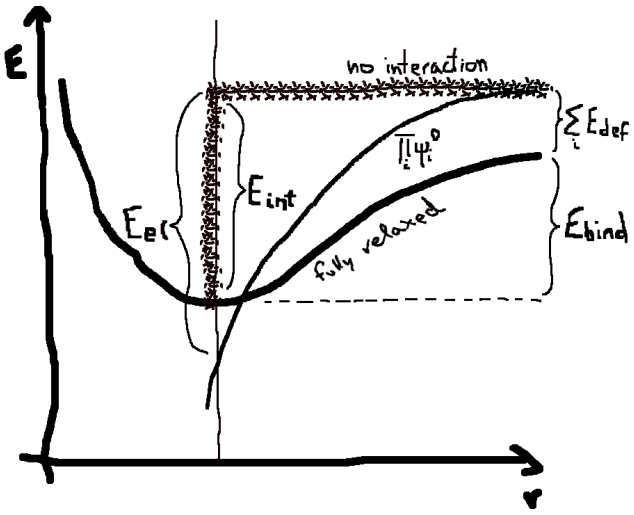


Figure 2.1: Conceptual drawing of binding and interaction energies as defined in (2.1) and (2.2). The solid line represents the total energy during dissociation; the intermediate thin line is the electrostatic component, or how fragments would interact if their electronic wave functions remained unchanged by mutual influence and the polarization approximation $\prod_i \psi_i^0$ were in force. Although the plot does not correspond to any specific system, it is typical for hydrogen bonded dimers.

Another approximation is recognized when more than one molecule is being considered, the adiabatic separation of intramolecular and intermolecular degrees of freedom. If energies determining the internal structure of molecules are significantly larger than noncovalent interactions, then intermolecular and intramolecular vibrations can be legitimately separated.

Although the interaction energy is defined for fixed nuclear coordinates, monomer deformation effects depend on intermolecular separation and in general influence the PES. Calculated potentials normally assume rigid monomers for simplicity, however it is not obvious which intramolecular conformations to use[49] and approaches to obtaining flexible monomer potentials have been discussed in the literature.[50]

The most straightforward way to express the interaction energy is in a supermolecular fashion, by simply following the definition of (2.1). In the case of two molecules A and B, this is the difference between the energy of their dimer and in isolation,

$$\Delta E_{\mathrm{AB}} = E_{\mathrm{AB}} - E_{\mathrm{A}} - E_{\mathrm{B}}. \tag{2.3}$$

When adopting the supermolecular approach, it is important to employ the same size-consistent method in calculating all energies; in practice, this means that at infinite separation E_{AB} should reduce to $E_{\mathrm{A}} + E_{\mathrm{B}}$. In other words, the interaction energy tends to zero. For a larger number of constituents, the total interaction energy can be partitioned into dimer interaction energies and many-body interactions involving three or more monomers,[46]

$$E_{\mathrm{int}} = \Delta E(2-\mathrm{body}) + \Delta E(3-\mathrm{body}) + ... + \Delta E(\mathrm{N}-\mathrm{body}). \tag{2.4}$$

[49] Jeziorska et al. have demonstrated that vibrationally averaged monomer geometries are superior to equilibrium ones for Ar$\cdots$HF, and and that the relaxation energy or change in potential due to monomer deformation is below the accuracy of electronic structure methods; for details, see Jeziorska, M., Jankowski, P., Szalewicz, K., Jeziorski, B. *J. Chem. Phys.* **2000**, *113*, 2957–2968.

[50] Murdachaew, G., Szalewicz, K. *Faraday Discuss.* **2001**, *118*, 121–142; Murdachaew, G., Szalewicz, K., Bukowski, R. *Phys. Rev. Lett.* **2002**, *88*, 123202; Jankowski, P. *J. Chem. Phys.* **2004**, *121*, 1655–1662.

2.1.1 Basis set superposition error

All of the interaction energies and other results presented in this dissertation are based on standard modern quantum chemistry methods – among others the LCAO molecular orbital framework, Hartree-Fock and other self-consistent field procedures,[51] second order Möller-Plesset theory (MP2)[52], coupled cluster approaches[53] and other methods for approximating the wave function of molecules. Without repeating textbook knowledge from applied quantum chemistry,[54] it is perhaps fitting to highlight just one recurring problem for calculations of van der Waals complexes, namely basis set superposition error (BSSE).

BSSE manifests itself in "uncorrected" calculations of the supermolecular interaction energy, for example when different basis sets are used for calculating the dimer energy E_{AB} and monomer energies E_{A} and E_{B} in (2.3). These basis sets are typically *different* in the sense that the dimer calculation contains the functions used in both monomer calculations, but the monomer calculations only use subsets of functions centered on one monomer's nuclei. Since in the dimer, monomers can use the one electron basis set of their partners, the total energy will be artificially lowered compared to that of the monomers. This unmatched extension of the monomer basis set in dimer calculations lowers the energy by virtue of the variational principle and in itself is an improvement as pointed out by Duijneveldt et al.[55] The problem is in the mismatching of basis sets use to generate energies that are subtracted. In many cases this effect has been shown to be around a few kcal/mol, which is comparable to or even larger than the interaction energies of some van der Waals complexes.

The BSSE becomes smaller when larger basis sets are used, as the imbalance between fragments and their complexes diminishes. Obviously, the best solution is to use basis sets as large as possible, in practice available only for the smallest of complexes. Extrapolation to the basis set limit also helps to alleviate the problem somewhat, demonstrated recently for the helium dimer by Varandas.[56]

The error due to unbalanced basis sets can be avoided entirely in calculations of the supermolecule interaction energy by using the function counterpoise version advocated by Boys and Bernardi,[57]

$$\Delta E_{\mathrm{AB}}^{\mathrm{CP}} = E_{\mathrm{AB}}^{\alpha\cup\beta} - E_{\mathrm{A}}^{\alpha\cup\beta} - E_{\mathrm{B}}^{\alpha\cup\beta}. \tag{2.5}$$

where α and β denote the basis sets of the respective monomers A and B, and $\alpha \cup \beta$ is their union. This definition is free of BSSE in the sense explained in the previous paragraphs, because it treats monomers as if they were a sub-case of the entire complex and matches their basis sets to that of the dimer.

Most calculations do not use the interaction energy to optimize the geometries of noncova-

[51] Roothaan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69–89.
[52] Møller, C., Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
[53] Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266.
[54] Cramer, C. J. *Essentials of Computational Chemistry*, 2nd ed.; Wiley, 2004; Lowe, J. P., Peterson, K. A. *Quantum Chemistry*, 3rd ed.; Elsevier, 2006.
[55] Duijneveldt, F., Duijneveldt-van de Rijdt, J., Lenthe, J. *Chem. Rev.* **1994**, *94*, 1873–1885.
[56] Varandas, A. J. C. *Theor. Chem. Acc.* **2008**, *119*, 511–521.
[57] Boys, S. F., Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–&.

lent complexes, but simply follow the total energy to its minimum. In this case a counterpoise (CP) correction equivalent to (2.5) can be added to the system's energy, which is also called the matching error and is the popular definition of basis set superposition error:[55]

$$\delta^{\mathrm{CP}} = \left(E_{\mathrm{A}}^{\alpha} - E_{\mathrm{A}}^{\alpha\cup\beta}\right) + \left(E_{\mathrm{B}}^{\beta} - E_{\mathrm{B}}^{\alpha\cup\beta}\right). \tag{2.6}$$

It is important to stress that in their seminal article, reprinted thirty years later[58], Boys and Bernardi applied the CP correction to two interacting atoms. They expressed concern that "it will still be a moderately difficult matter to put this method into operation for interesting [large] molecules". The extension of their principle to interactions between many-atom molecules, although widely used, has raised controversy in the literature as to how to obtain intermolecular interaction energies that are free of BSSE. Most of these controversies stem from using (2.6) to correct for BSSE instead of the function counterpoise approach of (2.5), and have persisted throughout the last three decades with a steady output of occasionally inconsistent reports.

The first of the controversies concerns whether or not and when the CP correction should be used at all, and what part of the dimer basis set $\alpha\cup\beta$ to use. For example, after performing an extensive study of the HF dimer, with the medium-sized basis sets available in 1985, Schwenke and Truhlar concluded that the CP-corrected interaction energy is not more reliable than its uncorrected counterpart in terms of statistical spread and trends.[59] They also considered using only the virtual or polarization functions of $\alpha \cup \beta$ in calculating δ^{CP}, an approach that Szczęśniak and Scheiner argue against in the case of the water dimer.[60] This particular issue was later resolved by Gutowski and others[61] and reviewed in 1994 by Duijneveldt et al.,[55] who dispel the question whether BSSE overcorrects the interaction energy. They point out that the CP correction in principle does not lead to a *better* result that is closer to the exact interaction energy – BSSE is caused by $\alpha \cup \beta$ being more complete than α or β alone, and by removing BSSE one cannot remove the incompleteness of $\alpha \cup \beta$ or the approximate nature of the computational method chosen.

Additional problems arise when systems beyond dimers are considered. As pointed out by Mierzwicki and Latajka,[62] there is no general consensus on how to apply posterior corrections or the function counterpoise recipe to many-body interactions. Also, it is worth noting that the relative imbalance of basis sets that causes BSSE influences not only the energy. Salvador et al. have investigated the BSSE footprint on the electron density of $(HF)_2$ and other systems van der Waals complexes,[63] and Skwara et al. propose and calculate BSSE effects in terms of interaction-induced properties.[64]

[58] Boys, S. F., Bernardi, F. *Mol. Phys.* **2002**, *100*, 65 – 73.

[59] Schwenke, D. W., Truhlar, D. G. *J. Chem. Phys.* **1985**, *82*, 2418–2426.

[60] Szczęśniak, M. M., Scheiner, S. *J. Chem. Phys.* **1986**, *84*, 6328 – 6335.

[61] Gutowski, M., Duijneveldt-van de Rijdt, J., Lenthe, J. *J. Chem. Phys.* **1993**, *98*, 4728–4727.

[62] Mierzwicki, K., Latajka, Z. *Chem. Phys. Lett.* **2000**, *325*, 465–472; Mierzwicki, K., Latajka, Z. *Chem. Phys. Lett.* **2003**, *380*, 654–664.

[63] Salvador, P., Fradera, X., Duran, M. *J. Chem. Phys.* **2000**, *112*, 10106–10115; Salvador, P., Fradera, X., Duran, M. *Int. J. Quant. Chem.* **2009**, *109*, 2572–2580.

[64] Skwara, B., Bartkowiak, W., Da Silva, D. L. *Theor. Chem. Acc.* **2009**, *122*, 127–136.

Many accounts have been given of how BSSE affects the potential energy surface of noncovalently bound complexes. Dannenberg and coworkers investigated this influence for various hydrogen bonded dimers[65] and for the water dimer at various levels of theory[66], as well as for transition states.[67] More recently, Thar et al. report the magnitude of BSSE in the water dimer along molecular dynamics (MD) trajectories that describe the dissociation of one water molecule, in particular that the error increases for shorter O$\cdots$H distances.[68]

BSSE has also been investigated for a number of hydrated complexes by Garden et al.[69], and Suhai with coworkers have surveyed various basis sets in hydrogen bonded complexes, with the general conclusion that intermolecular distances should be corrected for BSSE when extrapolated to the complete basis set (CBS) limit.[70]

Large, folded molecules have also been considered, for example a tripeptide by Valdés et al..[71] It is not always clear, however, if BSSE actually changes the shape of the PES qualitatively, as in the case of the dipeptide Tyr-Gly studied by Shields and van Mourik.[72] BSSE corrections have been shown to be important for some anion-π complexes[73] and for the energetics of water addition to charged metal ions that exhibit strong dipole-dipole interactions.[74]

The example of benzenium and ethene provides an intriguing, recent history – where the nonexistence of the ionic complex had first been attributed to BSSE[75] and later explained by the overestimation of correlation in the MP2 method.[76] Nonplanarity in benzene and other aromatic molecules has also recently been attributed to BSSE.[77] The sensitivity of these errors and how they are tied with the method and basis set used has also been the topic of recent controversy concerning new density functionals for nucleic acid bases.[78] Such questions are particularly important for studies relying on exact equilibrium geometries or vibrational data.

It is legitimate therefore to ask whether the total energy of a complex ΔE_{AB} should be corrected for the intermolecular BSSE, and the short answer is yes since the interaction energy constitutes a part of it. For example, Simon et al. advance a CP-corrected total energy obtained by adding $E_{\mathrm{A}}^{\alpha}+E_{\mathrm{B}}^{\beta}$ to both sides of (2.6),[65] a formulation that is currently implemented in an automatic CP-corrected geometry optimization option in Gaussian 03.

It is important to point out another controversy concerning BSSE, originating from the

[65]Simon, S., Duran, M., Dannenberg, J. J. *J. Chem. Phys.* **1996**, *105*, 11024–11031.
[66]Simon, S., Duran, M., Dannenberg, J. J. *J. Phys. Chem. A* **1999**, *103*, 1640–1643.
[67]Kobko, N., Dannenberg, J. J. *J. Phys. Chem. A* **2001**, *105*, 1944–1950.
[68]Thar, J., Hovorka, R., Kirchner, B. *J. Chem. Theor. Comp.* **2007**, *3*, 1510–1517.
[69]Garden, A. L., Lane, J. R., Kjaergaard, H. G. *J. Chem. Phys.* **2006**, *125*, 144317–7.
[70]Paizs, B., Salvador, P., Császár, A. G., Duran, M., Suhai, S. *J. Comp. Chem.* **2001**, *22*, 196–207; Salvador, P., Paizs, B., Duran, M., Suhai, S. *J. Comp. Chem.* **2001**, *22*, 765–786.
[71]Valdés, H., Klusák, V., Pitoňák, M., Exner, O., Starý, I., Hobza, P., Rulíšék, L. *J. Comp. Chem.* **2008**, *29*, 861–870.
[72]Shields, A. E., Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272–13277.
[73]Escudero, D., Frontera, A., Quiñonero, D., Deya, P. M. *Chem. Phys. Lett.* **2008**, *455*, 325–330.
[74]Kvamme, B., Wander, M. C. F., Clark, A. E. *Int. J. Quant. Chem.* **2009**, *109*, 2474–2481.
[75]Mourik, T. *J. Phys. Chem. A* **2008**, *112*, 11017–11020.
[76]Schwabe, T., Grimme, S. *J. Phys. Chem. A* **2009**, *113*, 3005–3008.
[77]Asturiol, D., Duran, M., Salvador, P. *J. Chem. Phys.* **2008**, *128*, 10.1063/1.2902974; Asturiol, D., Duran, M., Salvador, P. *J. Chem. Theor. Comp.* **2009**, *5*, 2574–2581.
[78]Mourik, T. *Chem. Phys. Lett.* **2009**, *473*, 206–208; Gu, J. D., Wang, J., Leszczyński, J., Xie, Y., Schaefer III, H. F. *Chem. Phys. Lett.* **2009**, *473*, 209–210.

attempt by Xantheas to relate BSSE to "large fragment relaxation" ($E_{\text{def}} \gg 1$ kcal/mol) and to adopt a unique form of the CP correction that aims to include monomer deformation.[79] This was pointed out to be an unnecessary complication by Szalewicz and Jeziorski two years later,[80] where they also make a number of excellent theoretical and practical arguments for the use of only (2.1) to study interactions energies between molecules.

As discussed a few paragraphs above, adding fragment deformation energies to the interaction energy is equivalent to adding the relaxed monomer energies $E_i(\mathbf{R}^0)$ to the total dimer energy, assuming the same basis set is used throughout. Since the relaxed monomer energies do not depend on intermolecular coordinates, this is strictly a trivial addition of a constant to the total energy. Moreover, as pointed out before, the interaction energy is a potential function and the conceptual route taken from the dissociated monomers to the final complex will not influence the total energy change. This means that the interaction and deformation energies can be attained in entirely separate calculations.

Nonetheless, the viewpoint advanced by Xantheas[79] and the acceptance of (2.2) over (2.1) in practice, as well as complications caused by applying BSSE removal schemes to the binding energy have taken root in the literature.[81] In most cases this is not an important practical issue since the deformation energy of a monomer is by definition small due to the weak nature of intermolecular interactions.[49]

An interesting notion of "intramolecular BSSE" was pointed out by Jensen,[82] where he showed that there is a superposition error associated with the change in the positions of individual basis functions between different geometries of the same molecule. In the context of the above discussion, this effect will also influence the deformation energy. More recently, Galano and Alvarez-Idaboy considered the intramolecular imbalances of basis set functions between individual atoms and their contribution to the conventional counterpoise correction.[83]

As a final note on this topic, the adequacy of the function counterpoise approach has been confirmed by independent Hartree-Fock and correlated interaction energies. In particular, the chemical Hamiltonian approach (CHA) devised by Mayer attempts to prevent basis set mixing in dimer calculations from the onset, by removing terms from the Hamiltonian that contain projections between orbitals of two different monomers.[84] Although the implementation of the CHA method is not publicly available, it has been reported to agree with results obtained using the functional counterpoise approach, with differences between them decaying faster than the superposition error itself.[85]

[79] Xantheas, S. S. *J. Chem. Phys.* **1996**, *104*, 8821–8824.
[80] Szalewicz, K., Jeziorski, B. *J. Chem. Phys.* **1998**, *109*, 1198–1200.
[81] Lendvay, G., Mayer, I. *Chem. Phys. Lett.* **1998**, *297*, 365–373; Rayon, V. M., Sordo, J. A. *Theor. Chem. Acc.* **1998**, *99*, 68–70; Sordo, J. A. *J. Mol. Struct.: THEOCHEM* **2001**, *537*, 245–251.
[82] Jensen, F. *Chem. Phys. Lett.* **1996**, *261*, 633–636.
[83] Galano, A., Alvarez-Idaboy, J. R. *J. Comp. Chem.* **2006**, *27*, 1203 – 1210.
[84] Reviewed at various levels of theory in Mayer, I. *Int. J. Quant. Chem.* **1998**, *70*, 41–63.
[85] Mayer, I., Valiron, P. *J. Chem. Phys.* **1998**, *109*, 3360–3373.

2.2 Methods for analyzing weak interaction energies

Alongside the total intermolecular interaction energy it is helpful to obtain information about its physical origin, especially as a first step in the parametrization of force fields and other empirical methods, which employ different functional forms to model various interaction terms. An overview follows of methods commonly used to analyze intermolecular interactions, with special focus on symmetry-adapted perturbation theory (SAPT) and the hybrid variation-perturbation approach, which is used throughout this work.

In relation to the previous discussion, it should be mentioned that perturbation theories consider the interaction between two monomers as a series of corrections monomer wave function derivable in analytical form, and therefore do not suffer from basis set superposition error in meaning conveyed by (2.6). They are still susceptible, however, to various kinds of basis set extension and truncation effects, and can be performed for both monomer (monomer centered, MCBS) or dimer basis sets (dimer centered, DCBS).

2.2.1 Symmetry-adapted perturbation theory

The formulation of symmetry-adapted perturbation theory for closed-shell monomers typically starts by presenting the supermolecular Hamiltonian $\hat{H}$ as a sum of a Hamiltonian for non-interacting monomers $\hat{H}_0$ and an interaction operator $\hat{V}$[86]. The former is divided further into Fock and fluctuation operators for the isolated monomer as introduced by Möller and Plesset[52]:

$$\hat{H} = \hat{H}_0 + \hat{V} = \hat{F}^A + \hat{F}^B + \hat{W}^A + \hat{W}^B + \hat{V}, \tag{2.7}$$

where $\hat{F}^A$ and $\hat{F}^B$ are the Fock operators for monomer A and B, respectively, and $\hat{W}^A$ and $\hat{W}^B$ are the corresponding intramonomer correlation operators. When not interacting, the monomers are described by two independent Hamiltonians, which means that $H_0 = H_0^A + H_0^B$. Solving the Schrödinger equation $H_0\phi_0 = E_0\phi_0$ therefore yields additive eigenvalues ($E_0 = E_0^A + E_0^B$) and corresponding separate eigenfunctions in the zeroth-order polarization approximation,

$$\phi_0 = \phi_0^A \phi_0^B. \tag{2.8}$$

The interaction energy derived from double perturbation theory is a series of polarization energies and exchange terms originating from symmetry adaptation:[87]

$$E_{int} = \sum_{n=1}^{\infty}\sum_{j=0}^{\infty}\left(E_{pol}^{(nj)} + E_{exch}^{(nj)}\right), \tag{2.9}$$

where n and j indicate the orders of intermolecular interaction and intramolecular electron correlation, connected respectively to the operators $\hat{V}$ and $\hat{W}^A + \hat{W}^B$ in (2.7).

[86] Jeziorski, B., Moszyński, R., Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

[87] The double expansion is discussed on pages 53-56 in Moszyński, R. "Theory of Iintermolecular Forces: an Introductory Account " in: Sokalski, W. A., Leszczynski, J., eds.; Challanges and Advances in Computational Chemistry and Physics, Vol. 4; Springer, 2007; Chapter 1, 1–152.

The first order polarization energy is equivalent to the classical electrostatic interaction,[88] in other words the interaction between unperturbed electron densities of isolated monomers:

$$E_{\mathrm{pol}}^{(10)} = \langle \phi_0^A \phi_0^B | V | \phi_0^A \phi_0^B \rangle. \tag{2.10}$$

The following definition of the first order energy in SAPT (left side) is equivalent to the Heitler-London energy (right side):

$$E^{(1)} = \frac{\langle \phi_0 | \hat{V} | \mathcal{A}\phi_0 \rangle}{\langle \phi_0 | \mathcal{A}\phi_0 \rangle} \simeq \Delta E_{\mathrm{HL}}^{(1)} = \frac{\langle \mathcal{A}\phi_0 | \hat{H} - E_0 | \mathcal{A}\phi_0 \rangle}{\langle \mathcal{A}\phi_0 | \mathcal{A}\phi_0 \rangle}, \tag{2.11}$$

where $\mathcal{A}$ is the antisymmetrizer or antisymmetrizing operator.[89] If ϕ_0 is an exact eigenfunction of H_0, then this becomes an equivalence, but in general a small correction needs to be added to the right hand side, which is called the Murrell delta $\delta_M^{(0)}$.[90]

By breaking down the antisymmetrizer and counting permutations that exchange at least one pair of electrons between monomers as described by Duijneveldt,[91] the first order energy can be written as the sum of the first order electrostatic and exchange contributions:

$$E_{\mathrm{exch}}^{(10)} = \frac{\langle \phi_0 | (V - E_{\mathrm{pol}}^{(10)}) \mathcal{P} \phi_0 \rangle}{1 + \langle \phi_0 | \mathcal{P} \phi_0 \rangle} = E^{(1)} - E_{\mathrm{pol}}^{(10)}, \tag{2.12}$$

where the cumulative transposition operator $\mathcal{P}$ collects all the mentioned permutations with appropriate sign. It is very costly to calculate $E_{\mathrm{exch}}^{(10)}$ directly with no approximations, because the exchange operators included in $\mathcal{P}$ cannot be expressed in terms of monomer properties. For this reason, the approximate equality in (2.11) is sometimes used (as in the hybrid method described in Section 2.2.3) and the first exchange contribution is then computed as the difference between $\Delta E_{\mathrm{HL}}^{(1)}$ and $E_{\mathrm{pol}}^{(10)}$.

Second order polarization contributions are expressed using the reduced resolvent operator $\hat{R}_0$, which is defined by a spectral expansion over the excited eigenstates of $\hat{H}_0$. These second order effects are divided into induction and dispersion components:

$$E_{\mathrm{pol}}^{(2j)} = -\langle \phi_0 | V \bar{R}_0 V | \phi_0 \rangle = E_{\mathrm{ind}}^{(2j)} + E_{\mathrm{disp}}^{(2j)}. \tag{2.13}$$

The induction term $E_{\mathrm{ind}}^{(20)}$ consists of separate contributions arising from both dimers:

$$E_{\mathrm{ind}}^{(20)} = E_{\mathrm{ind}}^{(20)}(A) + E_{\mathrm{ind}}^{(20)}(B) = -(\langle \Phi_A | \Omega_B \hat{R}_0^A \Omega_B | \Phi_A \rangle + \langle \Phi_B | \Omega_A \hat{R}_0^B \Omega_A | \Phi_B \rangle), \tag{2.14}$$

where Ω_A is the electrostatic potential operator for the unperturbed monomer A and the resolvent $\hat{R}_0^B$ contains only those terms in the spectral expansion involving the ground state of A and an exited state of B. Therefore $E_{\mathrm{ind}}^{(20)}(B)$ is the energy correction due to monomer

[88] The interaction terms introduced here and their interpretation are detailed on pages 27-36, ibid.
[89] See p.248 in Dirac, P. A. M. *The Principles of Quantum Mechanics*, 4th ed.; Clarendon, 1958.
[90] Jeziorski, B., Bulski, M., Piela, L. *Int. J. Quant. Chem.* **1976**, *10*, 281–297.
[91] Duijneveldt-van de Rijdt, J., Duijneveldt, F. *Chem. Phys. Lett.* **1972**, *17*, 425–427.

B being polarized by the static electron density around monomer A, induced nominally by a corresponding modification in the wave function of monomer B.

Second order dispersion is defined as the difference between $E_{\mathrm{pol}}^{(20)}$ and $E_{\mathrm{ind}}^{(20)}$, but can be calculated directly:

$$E_{\mathrm{disp}}^{(20)} = -\langle\phi_0|V\hat{R}_0^{AB}V|\phi_0\rangle \equiv E_{\mathrm{pol}}^{(20)} - E_{\mathrm{ind}}^{(20)}. \tag{2.15}$$

In the definition on the left, $\hat{R}_0^{AB}$ contains only excited states from both monomers, which means that the dispersion component is by definition a purely intermolecular effect and represents the interaction between instantaneous electron density fluctuations.

Second order exchange interactions couple polarization and exchange effects. Building on (2.13), they are naturally categorized into contributions pertaining to the induction and dispersion corrections, thus yielding exchange-induction and exchange-dispersion terms:

$$E_{\mathrm{exch}}^{(2j)} = E_{\mathrm{ex-ind}}^{(2j)} + E_{\mathrm{exch-disp}}^{(2j)}, \tag{2.16}$$

which can be understood as resulting from the antisymmetrization of locally excited wave functions.

The third order polarization energy can be analogically written as a sum of appropriate induction, dispersion and a mixed induction-dispersion terms. Involving the polarization of both monomers or quadratic effects, these contributions are harder to interpret and are not easily expressible in terms of monomer properties.

In order to relate perturbation results to supermolecular calculations, it is necessary to determine which SAPT terms compose the interaction energy calculated based on Hartree-Fock, MP2 and other methods. Another issue is whether the supermolecular interaction energy that is compared with should be corrected for basis set superposition error. The latter problem has largely been settled by a consensus to consistently use the functional counterpoise method.[87]

In the case of Hartree-Fock (HF) calculations, the supermolecular interaction energy can be recovered by collecting low order contributions from the the SAPT approach:

$$E_{int}^{HF} = E_{\mathrm{pol}}^{(10)} + E_{\mathrm{exch}}^{(10)} + E_{\mathrm{ind,resp}}^{(20)} + E_{\mathrm{ex-ind,r}}^{(20)} + \delta E_{int,resp}^{HF}. \tag{2.17}$$

where the subscript *resp* means that orbital relaxation effects were considered. The term $\delta E_{int,resp}^{HF}$ includes all third and higher order induction and exchange-induction effects entering the Hartree-Fock interaction energy, and can be interpreted using an exchange-deformation concept formulated by Moszyński et al.[92]

At the MP2 level of theory, the following ansatz was proposed and shown to be reasonable in the case of the helium dimer,[93]

$$\Delta E_{\mathrm{MP2}} \approxeq E_{int}^{HF} + E_{\mathrm{pol,resp}}^{(12)} + E_{\mathrm{ind,resp}}^{(22)} + E_{\mathrm{disp}}^{(20)} + E_{\mathrm{exch}}^{(11)} + E_{\mathrm{exch}}^{(12)} + E_{\mathrm{exch-disp}}^{(20)}. \tag{2.18}$$

[92] Moszyński, R., Heijmen, T. G. A., Jeziorski, B. *Mol. Phys.* **1996**, *88*, 741–758.
[93] Bukowski, R., Jeziorski, B., Szalewicz, K. *J. Chem. Phys.* **1996**, *104*, 3306–3319.

Therefore, the supermolecular MP2 interaction energy contains the leading second order electrostatic and intramonomer induction corrections as well as the major part of the dispersion energy $E_{\mathrm{disp}}^{(20)}$.

While the procedures employed within SAPT provide very accurate values for the dispersion energy $E_{\mathrm{disp}}^{(20)}$ as well as other components, it scales steeply with the number of atoms, roughly as N^7. Recently, a SAPT variant based on density functional theory has been developed that reduces somewhat this severe limitation, although motivated rather by the reported failures of supermolecular density functional theory (DFT) calculations to predict van der Waals interactions where dispersion is of importance. The polarization terms (electrostatic, induction and dispersion) are calculated based on electron densities obtained from DFT computations and time-dependent DFT results. Williams and Chabalowski introduced the first combination of these methods[94], followed by another formulation by Jansen and Heßelmann[95] and a related method for calculating the dispersion energy from TD-DFT monomer response functions.[96]

2.2.2 Analyses based on variational methods

Besides the explicit derivation of contributions to the interaction energy by perturbation methods, there have also been efforts to analyze interactions by modifying the Fock matrix or vectors representing the dimer wave function. One of the first of such attempts can be traced back to Kollman and coworkers,[97] nonetheless the most widely adopted energy decomposition analysis of this type (abbreviated by EDA throughout) was proposed by Kitaura and Morokuma. The scheme, described initially by Morokuma alone in 1971,[98] removes atomic orbital integrals from the Fock matrix and energy expressions if they are not involved in a particular type of interaction. The Hartree-Fock interaction energy is divided into four contributions by following total energy changes after removing the appropriate elements,

$$\Delta E_{\mathrm{RHF}} = \Delta E_{\mathrm{el}} + \Delta E_{\mathrm{ex}} + \Delta E_{\mathrm{pol}} + \Delta E_{\mathrm{CT}}. \tag{2.19}$$

The first of these contributions is obtained from the expectation value of the Hamiltonian using monomer wave functions while neglecting integrals that combine orbitals from different monomers. This is in principle the sum of monomer HF energies and the electrostatic interaction between them. After subtracting the monomer energies, the electrostatic component is obtained, marked by ΔE_{el}. The second contribution arises after including intermolecular overlap effects in the energy computation, thus including the Pauli repulsive exchange effects with ΔE_{ex}. This can be shown to be equal to the Heitler-London interaction energy defined

[94] Williams, H. L., Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 646–659.

[95] Jansen, G., Heßelmann, A. *J. Phys. Chem. A* **2001**, *105*, 11156–11157; Heßelmann, A., Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

[96] Misquitta, A. J., Jeziorski, B., Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201; Misquitta, A. J., Podeszwa, R., Jeziorski, B., Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

[97] Kollman, P. A., Allen, L. C. *Theor. Chim. Acta* **1970**, *18*, 399–403.

[98] Morokuma, K. *J. Chem. Phys.* **1971**, *55*, 1236–1244.

in (2.11) accurately to the order of the Landshoff and Murrell delta terms.[99]

In the third step suggested by Morokuma, the Hartree-Fock equations are iterated until consistency is reached while rejecting the same Fock matrix elements as in the first step. This leads to orbitals for each monomer that are distorted self-consistently by the electrostatic field of the other, which means that the extra contribution introduced in this step contains purely intermolecular induction effects (of the second and all higher orders). Morokuma called this contribution the *polarization* term, ΔE_{pol}.

Finally, in the last step the complete orbitals for the dimer are treated with the Hartree-Fock procedure, which produces the total Hartree-Fock energy. In the original paper by Morokuma,[98] the additional energy introduced after this step is attributed to charge transfer effects (ΔE_{CT}). The charge transfer term defined in this way is often surprisingly large, due to basis set superposition error, and in 1976, with Kitaura, Morokuma introduced modifications to the original scheme.[100] The charge transfer term was redefined, and the remaining portion of the Hartree-Fock interaction energy gathered in a mixing term, $\Delta E_{\mathrm{mix}}^{\mathrm{KM}}$. In all, within the second formulation the total self-consistent Hartree-Fock interaction energy is expressed as the sum,

$$\Delta E_{\mathrm{RHF}} = \Delta E_{\mathrm{el}}^{\mathrm{KM}} + \Delta E_{\mathrm{ex}}^{\mathrm{KM}} + \Delta E_{\mathrm{pol}}^{\mathrm{KM}} + \Delta E_{\mathrm{CT}}^{\mathrm{KM}} + \Delta E_{\mathrm{mix}}^{\mathrm{KM}}. \tag{2.20}$$

The superscript KM will be used throughout this dissertation to mark terms from the second, Kitaura-Morokuma scheme, as implemented in GAMESS. The method has also been extended in GAMESS to many-body systems by Chen and Gordon.[101]

The four components in the 1976 scheme are usually discussed in a similar way to the originally proposed procedure, namely within the framework of electron exchange and promotion between the occupied and unoccupied molecular orbitals of interacting monomers. To shorty recapitulate, the first contribution $\Delta E_{\mathrm{el}}^{\mathrm{KM}}$ is the classical electrostatic interaction between the occupied molecular orbitals of two monomers in their monomer basis sets. Exchange effects between occupied orbitals described by $\Delta E_{\mathrm{ex}}^{\mathrm{KM}}$ cause electron exchange and delocalization between molecules. The polarization contribution $\Delta E_{\mathrm{pol}}^{\mathrm{KM}}$ accounts for the promotion of electrons into vacant orbitals of the same molecule, and the charge transfer term $\Delta E_{\mathrm{CT}}^{\mathrm{KM}}$ results from intermonomer mixing of occupied and unoccupied orbitals. The last term, $\Delta E_{\mathrm{mix}}^{\mathrm{KM}}$, gathers the remaining part of the Hartree-Fock interaction energy.

It should be mentioned that in the original Kitaura-Morokuma formulation and implementation the monomer energies are calculated in their isolated basis set, so that ΔE_{RHF} as well as all of the components obtained in the scheme will be polluted by basis set superposition error. The issues has been addressed by applying the function counterpoise correction (see (2.5) and the related discussion) to the Kitaura-Morokuma analysis, as proposed by Sokalski et al.[102] and later by Cammi et al.[103] The counterpoise-corrected exchange and charge-transfer terms as defined by Cammi will be denoted everywhere by $\Delta E_{\mathrm{ex}}^{\mathrm{EX,CP}}$ and $\Delta E_{\mathrm{CT}}^{\mathrm{KM,CP}}$, respectively,

[99] See page 65 in Moszyński, 2007, in Ref. 87 on page 16.
[100] Kitaura, K., Morokuma, K. *Int. J. Quant. Chem.* **1976**, *10*, 325–340.
[101] Chen, W., Gordon, M. S. *J. Phys. Chem.* **1996**, *100*, 14316–14328.
[102] Sokalski, W. A., Roszak, S., Hariharan, P. C., Kaufman, J. J. *Int. J. Quant. Chem.* **1983**, *23*, 847–854.
[103] Cammi, R., Bonaccorsi, R., Tomasi, J. *Theor. Chim. Acta* **1985**, *68*, 271–283.

following Gordon and coworkers.[101]

Another criticism of the KM scheme was that the polarization and charge-transfer terms it produces do not obey the Pauli exclusion principle. This problem arises more subtly and is connected to the way charge transfer is constructed from exchanges between unoccupied and occupied orbitals of different monomers. In practice, the result is that $\Delta E_{\mathrm{pol}}^{\mathrm{KM}}$ and $\Delta E_{\mathrm{CT}}^{\mathrm{KM}}$ do not converge to an asymptotic limit as the basis set becomes more complete. Two early variants of the basic EDA scheme address this issue – the constrained space orbital variation (CSOV) method proposed by Bagus et al.[104] and reduced variational space (RVS) of Stevens and Fink[105] – both employ a group function approach to ensure that all wave functions used in the procedure do satisfy the Pauli exclusion principle. Functional counterpoise Kitaura-Morokuma as defined by Cammi and RVS methods were implemented in GAMESS and both were extended by Chen and Gordon with a treatment of many-body contributions; they observed that the Kitaura-Morokuma charge transfer term is strongly disrupted by BSSE as well as strong orbital interactions.[101]

In a recent work, Stone and Misquitta have isolated the charge-transfer component of the interaction energy within the framework of symmetry-adapted perturbation theory. Two sources of charge-transfer contributions, the second order induction and exchange-induction energies (see (2.14) and (2.16) below), largely cancel each other. This leaves a true charge-transfer term of a few kJ/mol at equilibrium distances of hydrogen bonds, which is approximately proportional to the exchange component and is even smaller for other charge transfer complexes.

Further modifications to the Kitaura-Morokuma decomposition scheme have been introduced by other researchers. Glendening and Streitwieser combine the supermolecule and fragment wave functions with the natural bond orbital (NBO) method in their natural energy decomposition analysis (NEDA).[106] Mo and coworkers implemented the Kitaura-Morokuma EDA scheme using the block-localized wave function (BLW) method, which inherently corrects for basis set superposition error and exhibits improved basis set stability.[107] Head-Gordon and coworkers on the other hand have recently proposed an EDA scheme based on absolutely localized molecular orbitals.[108]

An more stable EDA scheme has been very recently implemented in GAMESS by Su and Li[109] with the additional advantage of being applicable to open shell wave functions, lending it possible to analyzes covalent bonds. In the restricted Hartree-Fock case at large distances their approach is equivalent to the Kitaura-Morokuma procedure, with the difference that the exchange component $\Delta E_{\mathrm{ex}}^{\mathrm{KM}}$ is separated into attractive and repulsive exchange contributions $\Delta E_{\mathrm{ex}}^{\mathrm{SL}}$ and $\Delta E_{\mathrm{rep}}^{\mathrm{SL}}$. The higher order Morokuma charge transfer and polarization terms on the

[104]Bagus, P. S., Hermann, K., Bauschlicher, J. *J. Chem. Phys.* **1984**, *80*, 4378–4386.
[105]Stevens, W. J., Fink, W. H. *Chem. Phys. Lett.* **1987**, *139*, 15–22.
[106]Glendening, E. D., Streitwieser, A. *J. Chem. Phys.* **1994**, *100*, 2900–2909.
[107]Mo, Y., Gao, J., Peyerimhoff, S. D. *J. Chem. Phys.* **2000**, *112*, 5530–5538.
[108]Khaliullin, R. Z., Cobar, E. A., Lochan, R. C., Bell, A. T., Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 8753–8765.
[109]Su, P., Li, H. *J. Chem. Phys.* **2009**, *131*, 014102–15.

other hand are gathered into a single polarization term,

$$\Delta E_{\mathrm{RHF}}^{SL} = \Delta E_{\mathrm{el}}^{\mathrm{KM}} + \Delta E_{\mathrm{ex}}^{\mathrm{SL}} + \Delta E_{\mathrm{rep}}^{\mathrm{SL}} + \Delta E_{\mathrm{pol}}^{\mathrm{SL}}. \tag{2.21}$$

In their introduction of the method, Su an Li also show how this decomposition scheme can be applied to the Kohn-Sham density functional energy, in which case a dispersion also emerges. Other EDA-type decomposition schemes have also been reported for density functional methods. For example, the extended transition state (ETS) scheme proposed by Ziegler[110] and implemented using Slater orbitals in ADF[111] provides similar contributions attributed to electrostatic, Pauli and orbital interactions.

With the growing computational resources in recent years, many of these energy decomposition methods are routinely complemented by electron correlation terms. This correction to the interaction energy is usually obtained by subtracting the interaction energy at the Hartree-Fock level from the energy attained by Möller-Plesset or coupled cluster methods. Such a correction will be the same for all decomposition schemes based on Hartree-Fock wave function, and will not be discussed in detail here.

2.2.3 Hybrid variation-perturbation theory

Both of the non-empirical methods described above are confined to small or medium-sized systems with in practice up to about 30 atoms[112], although this limit is constantly being raised. The major technical bottleneck in the case of the Morokuma scheme are the disk requirements for storing and sifting through integrals, and in the case of SAPT the main limiting factor are the atomic integral transformations for computing certain perturbation terms. In this dissertation, most interaction energy calculations are performed using an alternate, hybrid variation-perturbation theory (HVPT) approach that avoids both of these bottlenecks, but still provides a reasonable amount of information on physically meaningful interaction components.

The HVPT approach is rooted in the functional counterpoise notion put forward by Boys and Bernardi[57] as discussed in Section 2.1.1, avoiding basis set superposition error by performing calculations consistently in the basis set of the entire complex;[113] the direct SCF procedure[114] is also central to reducing the amount of required disk space, thus increasing the possible number of atomic orbitals that can be considered. The latest implementation of the method was programmed in a customized version of GAMESS and augmented by correlation corrections at the MP2 or coupled cluster levels.

This hybrid decomposition scheme of the total interaction energy can be expressed by an equation, for example at the MP2 level, but it can also be expressed as a sequence of gradually

[110] Ziegler, T., Rauk, A. *Theor. Chem. Acc.* **1977**, *46*, 1–10; Mitoraj, M. P., Michalak, A., Ziegler, T. *J. Chem. Theor. Comp.* **2009**, *5*, 962–975.

[111] Velde, G., Bickelhaupt, F. M., Baerends, E. J., Guerra, C. F., Gisbergen, S., Snijders, J. G., Ziegler, T. *J. Comp. Chem.* **2001**, *22*, 931–967.

[112] Stone, A. J., Misquitta, A. J. *Int. Rev. Phys. Chem.* **2007**, *26*, 193–222.

[113] Sokalski, W. A., Roszak, S., Pecul, K. *Chem. Phys. Lett.* **1988**, *153*, 153–159.

[114] Almlöf, J., Faegri, K., Korsell, K. *J. Comp. Chem.* **1982**, *3*, 385–599; Haser, M., Ahlrichs, R. *J. Comp. Chem.* **1989**, *10*, 104–111.

more accurate components and levels of theory, bound below by advancing braces,

$$\Delta E_{\mathrm{MP2}} = \underbrace{\underbrace{\underbrace{\Delta E_{\mathrm{el,mtp}} + \Delta E_{\mathrm{el,pen}}}_{\Delta E^{(1)}_{\mathrm{el}}} + \Delta E^{(1)}_{\mathrm{ex}}}_{\Delta E^{(1)}_{\mathrm{HL}}} + \Delta E^{(\mathrm{R})}_{\mathrm{del}}}_{\Delta E_{\mathrm{RHF}}} + \underbrace{\Delta E^{(2)}_{\mathrm{disp}} + \Delta E_{\mathrm{corr,intra}}}_{\Delta E_{\mathrm{corr}}}. \tag{2.22}$$

The second order Möller-Plesset interaction energy ΔE_{MP2} is partitioned into a Hartree-Fock energy ΔE_{RHF} and second-order electronic correlation correction ΔE_{corr}. The latter can be further divided into contributions from inter-molecular dispersion $\Delta E^{(2)}_{\mathrm{disp}}$ and intramolecular correlation $\Delta E_{\mathrm{corr,intra}}$. At the Hartree-Fock level, ΔE_{RHF} is decomposed into corresponding electrostatic interactions $\Delta E^{(1)}_{\mathrm{el}}$, Pauli exchange $\Delta E^{(1)}_{\mathrm{ex}}$, and delocalization effects $\Delta E^{(\mathrm{R})}_{\mathrm{del}}$.

The electrostatic component is calculated directly from the perturbation expression (2.10),

$$\Delta E^{(1)}_{\mathrm{el}} \equiv E^{(10)}_{\mathrm{pol}}, \tag{2.23}$$

and can be divided into multipole $\Delta E_{\mathrm{el,mtp}}$ and penetration $\Delta E_{\mathrm{el,pen}}$ contributions by estimating the multipole term with distributed multipole moments,[115] for example on atoms as discussed later in Section 2.5.

The Heitler-London interaction energy $\Delta E^{(1)}_{\mathrm{HL}}$ is computed from the orbitals of isolated monomers obtained after Gram-Schmidt orthogonalization, using the variational expression in (2.11). The result is similar to what Kollman et al. originally called the electrostatic energy[97] and it is close to the first order SAPT interaction energy as mentioned earlier, with the small correction of the Murrell delta:

$$\Delta E^{(1)}_{\mathrm{HL}} = E^{(1)} + \delta^{(0)}_{M}. \tag{2.24}$$

Having obtained $\Delta E^{(1)}_{\mathrm{HL}}$, the exchange component is calculated as the difference between it and the first-order electrostatic interaction term,

$$\Delta E^{(1)}_{\mathrm{ex}} = \Delta E^{(1)}_{\mathrm{HL}} - \Delta E^{(1)}_{\mathrm{el}}. \tag{2.25}$$

It is fitting here to note that other partitioning methods follow this line of thought until this point, a notable example being the open-shell scheme used propsed and used by Cybulski et al. which complements $\Delta E^{(1)}_{\mathrm{HL}}$ with higher order SAPT terms[116]. In HVPT, the problem that there are a multitude of remaining interaction effects at the HF level, which cannot all be interpreted in terms of SAPT components, is dealt with by simply accumulating them in a term coined *delocalization*:

$$\Delta E^{(\mathrm{R})}_{\mathrm{del}} = \Delta E_{\mathrm{RHF}} - \Delta E^{(1)}_{\mathrm{HL}}. \tag{2.26}$$

[115] See Section 2.4.1 for details about the Cartesian multipole expansion and Section 2.5 for a discussion of distributed moments, including the CAMM and DMA methods.

[116] Cybulski, S. M., Burcl, R., Chałasiński, G., Szczęśniak, M. M. *J. Chem. Phys.* **1995**, *103*, 10116.

	SAPT	HVPT	physical interpretation
ΔE_{RHF}	$E_{\mathrm{pol}}^{(10)}$	$\Delta E_{\mathrm{el}}^{(1)}$	electrostatic interaction between Hartree-Fock electron densities
	$E_{\mathrm{exch}}^{(10)}$	$\Delta E_{\mathrm{ex}}^{(1)}$	exchange repulsion between Hartree-Fock monomers
	$E_{\mathrm{ind}}^{(20)}$ $E_{\mathrm{ex-ind}}^{(20)}$ $\delta E_{int,resp}^{HF}$	$\Delta E_{\mathrm{del}}^{(\mathrm{R})}$	other effects complementing the Hartree-Fock energy; $E_{\mathrm{ind}}^{(20)}$ and $E_{\mathrm{ex-ind}}^{(20)}$ are induction and exchange induction components, respectively, while $\delta E_{int,resp}^{HF}$ combines the remaining effects
ΔE_{MP2}	$E_{\mathrm{disp}}^{(20)}$		intermolecular dispersion component
	$E_{\mathrm{pol}}^{(12)}$ $E_{\mathrm{exch}}^{(11)}$ $E_{\mathrm{exch}}^{(12)}$ (others)	ΔE_{corr}	intramolecular second order Möller-Plesset correlation effects; $E_{\mathrm{pol}}^{(12)}$ is the electrostatic correction originating from MP2 densities, and $E_{\mathrm{exch}}^{(11)}$ and $E_{\mathrm{exch}}^{(12)}$ is the excess exchange they cause

Table 2.1: Comparison of selected interaction energy components obtained using the SAPT and HVPT decomposition schemes, along with their nomenclature and accepted physical interpretations. The terms in the SAPT column are defined in (2.10-2.18), while the ones labeled by HVPT are described in (2.23).

Besides analyzing the interaction energy into physically meaningful parts, an important idea conveyed by the HVPT scheme is the building of a hierarchy of theoretical models with increasing completeness and computational cost. Conversely, knowing the interaction energy profile allows an approximate level to be selected based on the most important interaction effects, system size and available resources.

If a coupled cluster calculation is performed for the total interaction energy instead of MP2, then (2.23) can be augmented by an additional correlation term $\delta_{\mathrm{corr}}^{\mathrm{CCSD(T)}}$ that gathers all correlation effects disjunct from the MP2 interaction energy:

$$\Delta E_{\mathrm{CCSD(T)}} = \Delta E_{\mathrm{MP2}} + \delta_{\mathrm{corr}}^{\mathrm{CCSD(T)}}, \tag{2.27}$$

where $\Delta E_{\mathrm{CCSD(T)}}$ is the interaction energy at the appropriate coupled cluster level of theory. Obviously, $\delta_{\mathrm{corr}}^{\mathrm{CCSD(T)}}$ or a variant without triple excitations can be calculated directly as the difference between $\Delta E_{\mathrm{CCSD(T)}}$ and ΔE_{MP2}. The progressive hierarchy of levels of theory implicitly defined by (2.23) and a possible coupled cluster term is then

$$\underset{\mathcal{O}(N_{\mathrm{at}})}{V(\mathbf{r}_i)q_i} \prec \underset{\mathcal{O}(N_{\mathrm{at}}^2)}{\Delta E_{\mathrm{el,mtp}}} \prec \underset{\mathcal{O}(N_{\mathrm{AO}}^4)}{\Delta E_{\mathrm{el}}^{(1)}} \prec \underset{\mathcal{O}(N_{\mathrm{AO}}^4)}{\Delta E_{\mathrm{HL}}^{(1)}} \prec \underset{\mathcal{O}(N_{\mathrm{AO}}^4)}{\Delta E_{\mathrm{RHF}}} \prec \underset{\mathcal{O}(N_{\mathrm{AO}}^5)}{\Delta E_{\mathrm{MP2}}} \prec \underset{\mathcal{O}(N_{\mathrm{AO}}^7)}{\Delta E_{\mathrm{CCSD(T)}}} \tag{2.28}$$

where "$\prec$" is used to indicate a "less than" binary relationship with respect to computational cost and theoretical completeness (N_{at} is the number of atoms and N_{AO} is the atomic orbital count). The most left-hand side term $V(\mathbf{r}_i)q_i$ is the sum of charge-potential products over all atoms, the simplest atomic electrostatic model with practical value.

To date, the HVPT method has been used to investigate a number of noncovalently bound systems, with one of the central goals being to describe the interaction of inhibitors with enzyme active sites. Sokalski and coworkers have repeatedly analyzed such interactions, for

example in the case of leucine analogs in the active site of leucine aminopeptidase (LAP),[117] correlating the interaction energy at various levels of theory with inhibition constants obtained experimentally. A similar study was performed for phenylalanine analogs[118], again highlighting the good correlation with experimental activities achieved by electrostatic interactions and even potential-based estimates. More recent reports by Dyguda et al. also describe the action in the case of phenylalanine ammonia-lyase (PAL)[119] and urokinase.[120]

The nature of catalytic activity has also been addressed, for enzymes in the case of ribonuclease by Kędzierski et al.,[121] in which the interactions of various parts of the active site with the substrates and transition state is analyzed. Combining the HVPT approach with a quantum mechanics/molecular mechanics (QM/MM) treatment, Szefczyk et al. describe the catalytic activity of the chorismate mutase active site,[122] whereas Szarek et al. examine protein phosphotransferase.[123] Catalytic activity has also been examined by Dziekoński et al. using the HVPT approach for the solvated double proton transfer between formamidine and formamide,[124] as well as in prototypical zeolite systems.[125]

All these HVPT studies, through comparison with with experiment, have indicated the dominant role of electrostatic effects in inhibition. More broadly, HVPT has also been applied to study cohesion energies in molecular crystals of urea[126] and various small molecule dimers.[127] A new direction has been taken up lately with increasingly comprehensive reports on nucleobase stacking complexes[128] as well as their complexes with intercalators.[129]

[117]Grembecka, J., Kędzierski, P., Sokalski, W. A. *Chem. Phys. Lett.* **1999**, *313*, 385–392.
[118]Sokalski, W. A., Kędzierski, P., Grembecka, J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 657–663.
[119]Dyguda, E., Grembecka, J., Sokalski, W. A., Leszczyński, J. *J. Am. Chem. Soc.* **2005**, *127*, 1658–1659.
[120]Grzywa, R., Dyguda-Kazimierowicz, E., Sieńczyk, M., Feliks, M., Sokalski, W. A., Oleksyszyn, J. *J. Mol. Model.* **2007**, *13*, 677–683.
[121]Kędzierski, P., Sokalski, W. A., Krauss, M. *J. Comp. Chem.* **2000**, *21*, 432–445.
[122]Szefczyk, B., Mulholland, A. J., Ranaghan, K. E., Sokalski, W. A. *J. Am. Chem. Soc.* **2004**, *126*, 16148–16159.
[123]Szarek, P., Dyguda-Kazimierowicz, E., Tachibana, A., Sokalski, W. A. *J. Phys. Chem. B* **2008**, *112*, 11819–11826.
[124]Dziekoński, P., Sokalski, W. A., Podolyan, Y., Leszczyński, J. *Chem. Phys. Lett.* **2003**, *367*, 367–375.
[125]Dziekoński, P., Sokalski, W. A., Kassab, E., Allavena, M. *Chem. Phys. Lett.* **1998**, *288*, 538–544; Dziekoński, P., Sokalski, W. A., Szyja, B., Leszczyński, J. *Chem. Phys. Lett.* **2002**, *364*, 133–138.
[126]Góra, R. W., Bartkowiak, W., Roszak, S., Leszczyński, J. *J. Chem. Phys.* **2002**, *117*, 1031–1039; Góra, R. W., Sokalski, W. A., Leszczyński, J., Pett, V. B. *J. Phys. Chem. B* **2005**, *109*, 2027–2033.
[127]Góra, R. W., Grabowski, S. J., Leszczyński, J. *J. Phys. Chem. A* **2005**, *109*, 6397–6405.
[128]Hill, G., Forde, G., Hill, N., Lester, W. A., Sokalski, W. A., Leszczyński, J. *Chem. Phys. Lett.* **2003**, *381*, 729–732; Langner, K. M., Sokalski, W. A., Leszczyński, J. *J. Chem. Phys.* **2007**, *127*, 111102; Czyżnikowska, Żaneta *J. Mol. Struct.: THEOCHEM* **2009**, *895*, 161–167; Czyżnikowska, Żaneta *J. Mol. Model.* **2009**, *15*, 615–622; Czyżnikowska, Żaneta, Zaleśny, R. *Biophys. Chem.* **2009**, *139*, 137–143.
[129]Langner, K. M., Kędzierski, P., Sokalski, W. A., Leszczyński, J. *J. Phys. Chem. B* **2006**, *110*, 9720–9727.

2.3 Comparison of interaction energy decomposition schemes

The various perturbation and variational decomposition schemes discussed above have been investigated quite thoroughly in the literature and applied for a number of usually small test cases as well as for some more complicated chemical problems. Not many attempts, however, have been made to systematically compare them for series of increasing basis sets. That is the intent of the present section, and Tables 2.3-2.4 present such a comparison of SAPT, HVPT and several EDA variants available in the current version of GAMESS. Five small dimers are examined at the geometries given in Fig. 2.2, namely He_2,[130] $(H_2)_2$,[131] $(HF)_2$,[132] $(LiH)_2$[113] and $(H_2O)_2$.[133] Correlation consistent basis sets augmented with extra diffuse functions – aug-cc-pVXZ, where X=D,T,Q,5,6 – were retrieved for all the complexes from the Basis Set Exchange. Due to the fact that the current implementation of GAMESS does not support h and i orbitals, the quintuple and sextuple-zeta basis sets were not complete in many cases, as indicated in the table captions.

Since the HVPT and EDA approaches are concerned mainly with effects contained in the Hartree-Fock interaction energy, the comparison is limited to that level. In the case of SAPT, the terms comprising (2.17) and calculated in the dimer basis set (DCBS) are presented. In all cases, the Hartree-Fock energies could be supplemented by adding a correction at any correlated level of theory, and here the MP2 interaction energy was obtained in order to have a full overview of the HVPT scheme as expressed by the hierarchy in (2.23). Table 2.2 shows the interaction energy at the Heitler-London, Hartree-Fock and MP2 levels for each dimer n the aug-cc-pVQZ basis, which was chosen as the largest truly correlation-consistent basis set that GAMESS can use and at the same time the largest available one for the lithium atom).

Obviously, all the terms compared here are additive and complete in the sense that they can be summed up to the total Hartree-Fock interaction energy, and that this energy is necessarily the same for all methods. It is important to point out that this is not true at the Heitler-London level. In the case of HVPT, the Heitler-London interaction energy $\Delta E^{(1)}_{\mathrm{HL}}$ is given exactly by the sum of $\Delta E^{(1)}_{\mathrm{el}}$ and $\Delta E^{(1)}_{\mathrm{ex}}$, and the corresponding SAPT value of $E^{(10)}_{\mathrm{pol}}$ $+E^{(10)}_{\mathrm{exch}}$ according to (2.11) differs by the relatively small value of Murrell delta. The sum of the EDA components $\Delta E^{\mathrm{KM}}_{\mathrm{el}}$ and $\Delta E^{\mathrm{KM}}_{\mathrm{ex}}$ on the other hand can, in general, deviate from $\Delta E^{(1)}_{\mathrm{HL}}$, although it should approach it when the basis set is saturated. The same is true for the counterpoise-corrected, RVS and Su-Li variants of the EDA approach.

Both the helium and molecular hydrogen dimers are dominated by dispersion, hence their Hartree-Fock energies ΔE_{RHF} are destabilizing. It is also interesting to note the sizably stronger reference interaction energy of Korona et al.[130] compared to ΔE_{MP2}, achieved by using explicitly correlated basis set functions and SAPT terms not covered by second-order

[130]Korona, T., Williams, H. L., Bukowski, R., Jeziorski, B., Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 5109 – 5122.
[131]Carmichael, M., Chenoweth, K., Dykstra, C. E. *J. Phys. Chem. A* **2004**, *108*, 3143–3152.
[132]Howard, B. J., Dyke, T. R., Klemperer, W. *J. Chem. Phys.* **1984**, *81*, 5417–5425.
[133]Klopper, W., Duijneveldt-van de Rijdt, J., Duijneveldt, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227–2234.

dimer	best in ref.	geometry	—aug-cc-pVQZ— $\Delta E^{(1)}_{\rm HL}$	$\Delta E_{\rm RHF}$	$\Delta E_{\rm MP2}$
He_2	-35.02 μH[130]	$d_{\rm COM}$ = 5.6 a.u.	30.69μH	29.20μH	-17.53μH
$(H_2)_2$	-135.8 mH[131]	R_{H-H}=0.743 Å, $d_{\rm COM}$ =3.450 Å, θ=90.0°	80.55 μH	60.87 μH	-134.8 μH
$(HF)_2$	-7.39 mH[132]	R_{H-F}=0.917 Å, $d_{\rm COM}$ =2.673 Å, θ=117°	-2.517 mH	-5.934 mH	-6.727 m H
$(LiH)_2$	-14.64 mH[113]	R_{Li-H}=1.633 Å, $d_{\rm COM}$ =5.133 Å, θ=180°	-13.11 mH	-14.45 mH	-14.33 m H
$(H_2O)_2$	-7.987 mH[133]	$R^{\rm don1}_{O-H}$=0.9639 Å, $R^{\rm don2}_{O-H}$=0.9569 Å $R^{\rm acc}_{O-H}$=0.9583, $d_{\rm COM}$ =2.916 Å $\alpha^{\rm don}_{H-O-H}$=104.83°, $\alpha^{\rm acc}_{H-O-H}$=104.87° θ=58.48°	-2.167 mH	-5.801 mH	-7.782 mH

[130] combined SAPT and FCI calculations, with correlated geminals and orbital basis sets

[131] CCD/aug-cc-pVTZ energy

[132] experimental geometry and energy

[113] RHF energy in an uncontracted 20s3p/16s2p basis set

[133] CCSD(T) extrapolated to the basis set limit

Table 2.2: Geometrical parameters for the small dimers used to compare interactions energy decomposition schemes. Results for these dimers are presented in Tables 2.3 and 2.4, while the Heitler-London, Hartree-Fock and MP2 interaction energies are provided here for the aug-cc-pVQZ basis set. The citations listed are the sources for these particular geometries, and the best interaction energy value available for each reference geometry is also given (see text for details). In all cases, $d_{\rm COM}$ denotes the distance between the centers of mass and θ is the angle or dihedral between the monomers, which are all linear or planar. For the water dimer, the superscripts "don" and "acc" respectively mark the proton donor and acceptor, while "don1" denotes the elongated donating bond and "don2" stands for the other O-H bond of the donating molecule.

Möller-Plesset theory. The literature since then has further improved this value with larger calculations,[133] relativistic effects[134] and radiative corrections.[135]

The sum of $E^{(10)}_{\rm pol}$ and $E^{(10)}_{\rm exch}$ in Korona et al. amounts to 33.46μH, which is just a little lower than the best estimate of 33.60μH obtained with similar methods a decade earlier.[136] In any case, the first order effects compare more favorably with the value of $\Delta E^{(1)}_{\rm HL}$ in Table 2.3. Of course, the agreement is excellent with values that were obtained from uncorrelated orbital calculations, in particular the Hartree-Fock limit of 29.2μH published by Gutowski et al.[137] and similar values reported by Sokalski et al.[113]

One more observation needs to be made for the helium dimer, since the multipole expansion of the electrostatic interaction energy should be zero and unexpected residual repulsion is nonetheless observed in the case of the second rank DMA (distributed multipole analysis)[138] term $\Delta E^2_{\rm DMA}$. The DMA results presented here were obtained along with the HVPT decomposition calculations and therefore the interacting monomer densities spanned the basis set of the entire dimer. For limited basis sets this causes nonphysical charge transfer from one atom onto the other, breaking the symmetry of the helium atom, an effect that vanishes quickly as the basis set is increased.

The T-shaped dihydrogen dimer essentially has the same interaction energy profile as He_2,

[133] Cencek, W., Jeziorska, M., Bukowski, R., Jaszuński, M., Jeziorski, B., Szalewicz, K. *J. Phys. Chem. A* **2004**, *108*, 3211 – 3224.

[134] Cencek, W., Komasa, J., Pachucki, K., Szalewicz, K. *Phys. Rev. Lett.* **2005**, *95*, 10.1103/PhysRevLett.95.233004.

[135] Pachucki, K., Komasa, J. *J. Chem. Phys.* **2006**, *124*, 10.1063/1.2166017.

[136] Rybak, S., Szalewicz, K., Jeziorski, B. *J. Chem. Phys.* **1989**, *91*, 4779 – 4784.

[137] Gutowski, M., Duijneveldt, F., Chałasiński, G., Piela, L. *Mol. Phys.* **1987**, *61*, 233.

[138] See Section 2.5.1 for an explanation of DMA, and (2.42) and (2.43) for a definition of expansion rank.

			— electrostatic —			— multipole component —				— exchange —					
	basis	N_{AO}	$E_{pol}^{(10)}$	$\Delta E_{el}^{(1)}$	ΔE_{el}^{KM}	$\Delta E_{DMA}^{\kappa'\leq 2}$	$\Delta E_{CAMM}^{\kappa'\leq 2}$	$\Delta E_{CAMM}^{\kappa'\leq 8}$	$\Delta E_{CAMM}^{MP2,\kappa'\leq 8}$	$E_{exch}^{(10)}$	$\Delta E_{ex}^{(1)}$	ΔE_{ex}^{KM}	$\Delta E_{ex}^{EX,CP}$	ΔE_{ex}^{SL}	ΔE_{rep}^{SL}
He_2	X=D	18	-5.068	-5.068	-3.662	0.0078	0.0000	0.0000	0.0000	36.74	36.74	35.19	35.24	-37.13	72.32
μH	X=T	46	-5.087	-5.087	-4.695	0.0001	0.0000	0.0000	0.0000	35.60	35.56	36.39	36.39	-46.52	82.91
	X=Q	92	-4.955	-4.955	-4.761	0.0000	0.0000	0.0000	0.0000	35.64	35.65	36.04	36.04	-46.57	82.61
	X=5	160	-4.930	-4.930	-4.818	0.0000	0.0000	0.0000	0.0000	35.65	35.65	35.75	35.75	-46.56	82.31
	X=6[a]	210	-4.939	-4.939	N/A	0.0000	0.0000	0.0000	0.0000	35.64	35.62	N/A	N/A	-46.73	82.34
$(H_2)_2$	X=D	36	-91.48	-91.48	-92.19	-54.68	-50.99	-53.13	-46.60	179.2	179.3	186.2	186.6	-333.8	520.0
μH	X=T	92	-97.63	-97.66	-97.41	-62.61	-62.55	-64.74	-58.78	178.7	178.9	177.1	177.2	-295.5	472.6
	X=Q	184	-97.75	-97.75	-97.63	-62.75	-61.54	-63.55	-57.99	178.2	178.3	178.0	178.0	-298.5	476.5
	X=5	320	-97.54	-97.54	-97.42	-61.92	-60.47	-63.23	-57.79	178.1	178.2	178.3	178.3	-299.4	477.7
	X=6[a]	420	-97.56	-97.56	-97.44	-61.90	-60.49	-63.20	-57.79	178.1	178.2	178.3	178.3	-298.9	477.1
$(HF)_2$	X=D	64	-10.78	-10.78	-10.95	-9.24	-9.45	-9.21	-8.155	8.241	8.290	8.196	8.210	-8.998	17.19
mH	X=T	138	-10.70	-10.71	-10.89	-8.842	-10.37	-9.09	-8.086	8.236	8.285	8.278	8.282	-9.20	17.48
	X=Q	252	-10.79	-10.79	-10.80	-9.08	-10.00	-9.09	-8.131	8.226	8.275	8.245	8.246	-9.14	17.39
	X=5[a]	370	-10.80	-10.80	-10.79	-8.992	-9.10	-9.15	-8.199	8.226	8.275	8.276	8.276	-9.18	17.45
	X=6[a]	470	-10.80	-10.80	-10.80	-9.04	-9.68	-9.11	-8.161	8.226	8.275	8.275	8.275	-9.18	17.45
$(LiH)_2$	X=D	64	-13.41	-13.41	-13.09	-13.25	-12.47	-13.01	-12.81	0.2089	0.2089	0.1080	0.1220	-0.2370	0.3450
mH	X=T	138	-13.28	-13.29	-13.25	-13.45	-12.76	-13.30	-13.03	0.1631	0.1642	0.1540	0.1540	-0.1930	0.3470
	X=Q	252	-13.27	-13.27	-13.27	-12.98	-12.63	-13.29	-13.01	0.1639	0.1639	0.1630	–	-0.1840	0.3470
$(H_2O)_2$	X=D	82	-13.17	-13.17	-13.31	-11.02	-11.82	-10.82	-10.11	10.87	10.97	10.84	10.85	-13.46	24.29
mH	X=T	184	-13.09	-13.09	-13.08	-10.56	-11.80	-10.65	-9.98	10.86	10.96	10.84	10.85	-13.35	24.20
	X=Q	344	-13.11	-13.11	-13.07	-10.20	-11.51	-10.66	-10.05	10.84	10.94	10.87	10.88	-13.44	24.31
	X=5[a]	530	N/A	-13.10	-13.10	-10.36	-10.27	-10.66	-10.06	N/A	10.94	10.94	10.94	-13.53	24.47
	X=6[a]	680	N/A	-13.10	-13.11	-10.82	-11.27	-10.68	-10.09	N/A	10.94	10.94	10.94	-13.54	24.48

Table 2.3: Comparison of the electrostatic and exchange components obtained from various interaction energy decomposition schemes, including SAPT in DCBS (2.17), HVPT in DCBS (2.23), Kitaura-Morokuma in MCBS (2.20 and $\Delta E_{ex}^{EX,CP}$ is the counterpoise-corrected counterpart of the exchange term) and Su-Li in MCBS ((2.21)). $\Delta E_{DMA}^{\kappa'\leq 2}$ denotes the interaction between DMA multipole moments for monomers in DCBS, with all interactions summed that are possible with moments up to rank 2 according to the moment truncation described by (2.43). $\Delta E_{CAMM}^{\kappa'\leq 2}$ and ΔE_{CAMM}^{8} are the same, but refer to CAMM moments calculated in MCBS, and $\Delta E_{CAMM}^{MP2,\kappa'\leq 8}$ refers to moments generated from MP2 orbitals.
Note: the current implementation of Morokuma-type decompositions in GAMESS does not support spherical harmonics, therefore in these case the basis set had redundant Cartesian functions.
[a]These basis sets differed from the original by the lack of h and i function, since the corresponding orbital integrals are not implemented in GAMESS.

			— SAPT —			— HVPT —	— Morokuma —					— RVS —		— SL —
	basis	N_{AO}	$E^{(20)}_{\mathrm{ind,resp}}$	$E^{(20)}_{\mathrm{ex-ind,r}}$	$\delta E^{HF}_{int,resp}$	$\Delta E^{(\mathrm{R})}_{\mathrm{del}}$	$\Delta E^{\mathrm{KM}}_{\mathrm{pol}}$	$\Delta E^{\mathrm{KM}}_{\mathrm{CT}}$	$\Delta E^{\mathrm{KM}}_{\mathrm{mix}}$	$\Delta E^{\mathrm{KM,CP}}_{\mathrm{CT}}$	$\Delta E^{\mathrm{KM,CP}}_{\mathrm{mix}}$	$\Delta E^{\mathrm{RVS}}_{\mathrm{pol}}$	$\Delta E^{\mathrm{RVS}}_{\mathrm{CT}}$	$\Delta E^{\mathrm{SL}}_{\mathrm{pol}}$
He_2	X=D	18	-0.7686	0.6800	-1.281	-1.373	-0.0054	-16.64	0.0962	-3.819	2.542	-0.0289	-2.572	-16.55
μH	X=T	46	-0.7986	0.6596	-1.340	-1.439	-0.0137	-4.139	0.6488	-3.282	0.7146	-0.0420	-2.544	-3.504
	X=Q	92	-0.8003	0.6582	-1.349	-1.495	-0.0176	-3.185	0.5417	-2.664	0.5851	-0.0584	-2.038	-2.661
	X=5	160	-0.7993	0.6568	-1.348	-1.494	-0.0245	-2.396	0.4713	-2.202	0.4922	-0.0771	-1.653	-1.949
	X=6[a]	210	-0.7990	0.6563	-1.348	-1.479	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-1.588
$(H_2)_2$	X=D	36	-10.03	4.005	-12.46	-18.61	-2.553	-46.71	1.178	-35.34	12.67	-2.490	-30.97	-48.09
μH	X=T	92	-10.54	3.833	-12.72	-19.61	-3.971	-24.68	8.487	-23.79	9.41	-4.131	-14.75	-20.17
	X=Q	184	-10.59	3.806	-12.78	-19.69	-4.674	-26.07	10.45	-25.60	10.82	-4.402	-14.97	-20.29
	X=5	320	-10.59	3.803	-12.77	-19.69	-5.151	-26.50	11.47	-26.36	11.58	N/A	N/A	-20.18
	X=6[a]	420	-10.59	3.802	-12.77	-19.68	-5.509	-26.46	12.06	-26.43	12.08	N/A	N/A	-19.91
$(HF)_2$	X=D	64	-4.130	1.947	-1.080	-3.312	-2.239	-2.349	1.232	-2.192	1.358	-1.564	-1.481	-3.355
mH	X=T	138	-4.324	2.101	-1.102	-3.388	-3.384	-2.676	2.683	-2.631	2.747	-1.960	-1.240	-3.378
	X=Q	252	-4.342	2.089	-1.116	-3.418	-5.525	-3.046	5.127	-3.025	5.164	-2.116	-1.181	-3.444
	X=5[a]	370	-4.348	2.093	-1.117	-3.421	large[b]	-2.754	large[b]	-2.754	large[b]	N/A	N/A	-3.426
	X=6[a]	470	-4.348	2.093	-1.117	-3.421	large[b]	-3.279	large[b]	-3.279	large[b]	N/A	N/A	-3.422
$(LiH)_2$	X=D	64	-1.419	0.1717	-0.1219	-1.370	-1.122	-1.444	0.2610	-1.284	0.7720	-1.015	-0.9240	-2.305
mH	X=T	138	-1.375	0.1534	-0.1217	-1.343	-1.353	-0.9940	0.9600	-0.9810	0.9840	-1.189	-0.1300	-1.388
	X=Q	252	-1.372	0.1546	-0.1208	-1.339	-1.556	-1.113	1.323	N/A	N/A	N/A	N/A	-1.345
$(H_2O)_2$	X=D	82	-4.429	2.365	-1.388	-3.557	-2.370	-2.783	1.569	-2.630	1.711	-1.636	-1.625	-3.584
mH	X=T	184	-4.647	2.546	-1.412	-3.617	-4.083	-3.219	3.685	-3.183	3.746	-1.965	-1.472	-3.617
	X=Q	344	-4.636	2.526	-1.420	-3.634	-13.95	-3.564	13.85	-3.546	13.90	-2.107	-1.405	-3.671
	X=5[a]	530	N/A	N/A	N/A	-3.635	large[b]	-3.765	large[b]	-3.764	large[b]	N/A	N/A	-3.637
	X=6[a]	680	N/A	N/A	N/A	-3.635	large[b]	-3.871	large[b]	-3.871	large[b]	N/A	N/A	-3.634

Table 2.4: Comparison of higher order contributions to the Hartree-Fock interaction energy obtained from various interaction energy decomposition schemes, including SAPT in DCBS (2.17), HVPT in DCBS (2.23), Kitaura-Morokuma in MCBS (2.20 and the superscript CP denotes a counterpoise correction included) and Su-Li in MCBS ((2.21)). $\Delta E^{\mathrm{RVS}}_{\mathrm{CT}}$ and $\Delta E^{\mathrm{RVS}}_{\mathrm{pol}}$ are obtained from the reduced variational space variants of the Kitaura-Morokuma scheme as discussed in Section 2.2.2.
Note: the current implementation of Morokuma-type decompositions in GAMESS does not support spherical harmonics, therefore in these case the basis set had redundant Cartesian functions.
[a] These basis sets differed from the original by the lack of h and i function, since the corresponding orbital integrals are not implemented in GAMESS.
[b] These results were very large, and in $\Delta E^{\mathrm{KM}}_{\mathrm{pol}}$ approximately cancels out with $\Delta E^{\mathrm{KM}}_{\mathrm{mix}}$ or $\Delta E^{\mathrm{KM,CP}}_{\mathrm{mix}}$.

although the saturation of intramonomer correlation with basis set size is relatively not as important in this equilibrium conformation. Increasing the basis set to aug-cc-pV6Z improves the MP2 interaction by only about 3μH. The best value reported along with the reference geometry by Carmichael et al.,[131] which was obtained from coupled cluster calculations and triple-zeta basis sets, differs by less than 1%.

Moving on to the hydrogen bonded dimers, the experimental geometry used here for the hydrogen fluoride dimer has been refined extensively in the past with computations. Being one of the smallest possible hydrogen-bonded dimers, it is an excellent prototype system for tests and comparisons (in an early study, for example, Latajka and Scheiner incrementally exchange HF molecules with HCl).[139]. The equilibrium structure and energetics of $(HF)_2$ can be described adequately with density functionals,[140] and the potential energy surface has been mapped in detail.[141]

In one of their studies of this surface Klopper et al.[142] give a counterpoise-corrected MP2 estimate of the dissociation energy, -6.79 mH, which is within 1% of the ΔE_{MP2}/aug-cc-pVQZ presented in Table 2.2. This value, obtained by Klopper et al. with an extended quadrupole-zeta basis set, agrees even better with the value obtained here using the aug-cc-pV6Z basis, namely -6.78 mH.

The lithium hydride dimer was arranged linearly, and 5.133 Å between the molecular centers of mass (COM) corresponds to the 3.5 Å intermolecular Li$\cdots$H distance chosen by Sokalski et al. to demonstrate the performance of the HVPT method at large intermolecular distances.[113] Their conclusion, that the electrostatic term exhibits the strongest dependence on basis set size, however, does not carry over to the present case where basis set functions of higher angular momentum are added in the series of correlation-consistent basis sets. In fact, the electrostatic component $\Delta E_{\mathrm{el}}^{(1)}$ for the LiH dimer in Table 2.3 does not change more than 1% and the delocalization term $\Delta E_{\mathrm{del}}^{(\mathrm{R})}$ varies only a bit more.

On the other hand, the exchange component for $(LiH)_2$ changes by over 20% when moving from aug-cc-pVDZ to aug-cc-pVTZ. This apparently takes place only after the addition of f functions to the lithium atoms and d function to hydrogen, since the exchange term does not change any more when moving to aug-cc-pVQZ . Other studies on this linear dimer have usually focused on closer separations, around 1.7 Å for the Li$\cdots$H contact, and report stronger interaction, however the relative importance of correlation remains comparable.[143] Fig. 2.2 shows the changes in $\Delta E_{\mathrm{el}}^{(1)}$, $\Delta E_{\mathrm{ex}}^{(1)}$ and $\Delta E_{\mathrm{del}}^{(\mathrm{R})}$ when moving to larger basis sets, and compares them to corresponding terms obtained in the other decomposition schemes tested.

Finally, for the water dimer geometry a best estimate published in 2000 was used,[133] and the MP2 interaction obtained here is about 3% weaker than the best coupled cluster value calculated therein. The ΔE_{MP2} value in Table 2.2, as well as the -7.83 mH obtained for aug-

[139]Latajka, Z., Scheiner, S. *Chem. Phys.* **1988**, *122*, 413–430.
[140]Latajka, Z., Bouteiller, Y. *J. Chem. Phys.* **1994**, *101*, 9793–9799.
[141]Klopper, W., Quack, M., Suhm, M. A. *Chem. Phys. Lett.* **1996**, *261*, 35–44; Hodges, M. P., Stone, A. J., Lago, E. C. *J. Phys. Chem. A* **1998**, *102*, 2455–2465.
[142]Klopper, W., Quack, M., Suhm, M. A. *J. Chem. Phys.* **1998**, *108*, 10096–10115.
[143]McDowell, S. A. C. *J. Comp. Chem.* **2003**, *24*, 1201–1207; Chen, Y.-L., Huang, C.-H., Hu, W.-P. *J. Phys. Chem. A* **2005**, *109*, 9627–9636.

cc-pV6Z, are between the frozen-core MP2 interaction energies calculated by Klopper et al. for basis sets denoted therein by QQZ (-7.698 mH) and IO249 (-7.846 mH).

A few general remarks can be made for Table 2.3, which compares the first order electrostatic and exchange terms. First of all, the HVPT electrostatic interaction energy $\Delta E_{\rm el}^{(1)}$ is almost always identical to the $E_{\rm pol}^{(10)}$ SAPT term as expected from (2.23). Occasional last-digit deviations most probably arise from differences in the numerical implementation. The exchange terms from SAPT and HVPT are very close to each other, illustrating that the Murrell delta as defined by (2.24) is small – with a largest value in the present comparison of nearly 1% for the hydrogen fluoride and water dimers.

The corresponding components from the Kitaura-Morokuma analysis, $\Delta E_{\rm pol}^{\rm KM}$ and $\Delta E_{\rm ex}^{\rm KM}$, also tend to the SAPT/HVPT values, although this convergence is markedly slower. It becomes worse in the case of the Su-Li exchange (attractive) and repulsive components ($\Delta E_{\rm ex}^{\rm SL}$ and $\Delta E_{\rm rep}^{\rm SL}$, respectively), which otherwise are rather stable with the basis set and added together reproduce $\Delta E_{\rm ex}^{\rm KM}$. Separately, however, they can be a few times larger than their sum. This difference is always largest for the smallest basis set.

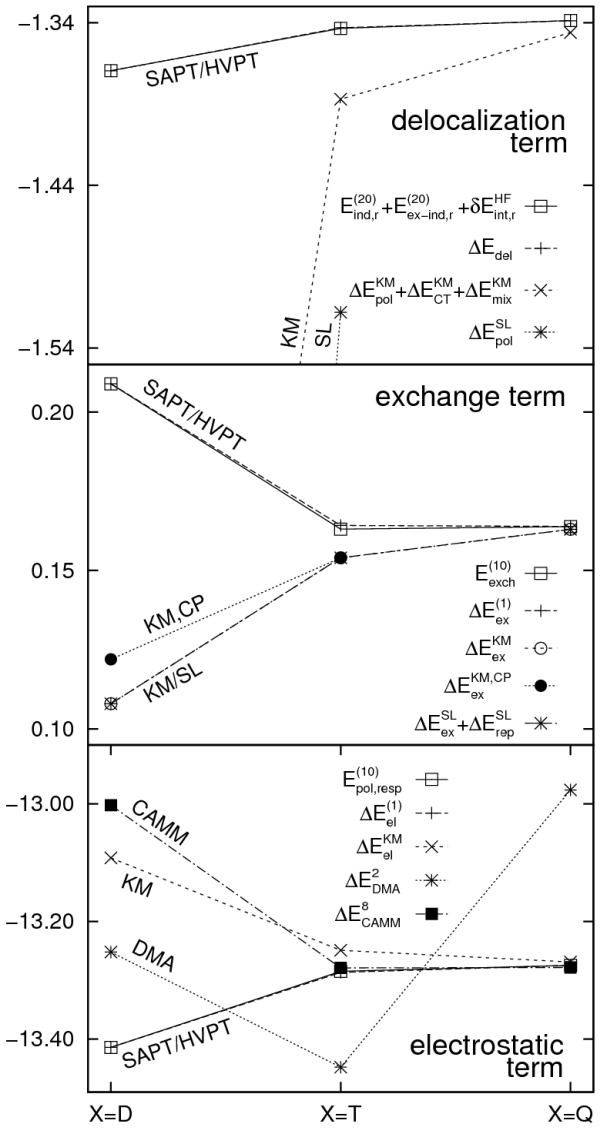


Figure 2.2: Comparison of interaction energy terms for the lithium hydride dimer in three consecutive correlation consistent basis sets (aug-cc-pVXZ, where X=D,T,Q on the axis). Components are compared for the SAPT, HVPT, Kitaura-Morokuma and Su-Li methods, whose numerical values are given in Tables 2.3 and 2.4. All energy values are in millihartree.

Significantly larger basis set dependence is observed for interactions based on atomic multipole moments up to rank two ($\Delta E_{\rm DMA}^{\kappa'\leq 2}$ and $\Delta E_{\rm CAMM}^{\kappa'\leq 2}$), the situation always improves however when moments up to rank 8 are considered ($\Delta E_{\rm CAMM}^{\kappa'\leq 2}$). Furthermore, the multipole electrostatic interaction based on MP2 orbitals ($\Delta E_{\rm CAMM}^{{\rm MP2},\kappa'\leq 2}$) is consistently about 10% weaker than its Hartree-Fock counterpart due to intramolecular correlation.

The stability of the remaining two Kitaura-Morokuma terms $\Delta E_{\rm pol}^{\rm KM}$ and $\Delta E_{\rm CT}^{\rm KM}$ is much worse, as seen in the appropriate columns of Table 2.4. Aside from the charge transfer term oscillating with growing basis set size, which is well-known from the literature, the polarization term $\Delta E_{\rm pol}^{\rm KM}$ as well as it complement $\Delta E_{\rm mix}^{\rm KM}$ clearly diverge beyond aug-cc-pVQZ for the close contacts in the hydrogen bonded HF and water dimers. This divergence possibly originates from linear dependent basis functions within the Cartesian representation used in the EDA

methods, as the spherical coordinate representation currently implemented in GAMESS cannot be used with the decomposition code.

On the other hand, the three terms that comprise ΔE_{RHF}-$\Delta E_{\mathrm{HL}}^{(1)}$ in the SAPT scheme, namely $E_{\mathrm{ind,resp}}^{(20)}$, $E_{\mathrm{ex-ind,r}}^{(20)}$ and $\delta E_{int,resp}^{HF}$, converge systematically as expected with the expanding basis set. It is interesting to notice that while the induction and exchange-induction terms are significant, particularly in the case of hydrogen bonded dimers, the remaining higher order effects gathered in $\delta E_{int,resp}^{HF}$ are also not negligible. Therefore, while providing two additional terms with physical meaning, the SAPT interpretation of the Hartree-Fock interaction energy still does not cover a large portion of it. The remainder is possibly distributed among a number of moderately important polarization terms (since it is always attractive).

As described in Section 2.2.3, The HVPT approach goes further in this direction and collects all of the Hartree-Fock contributions that are not of the first order into the delocalization term $\Delta E_{\mathrm{del}}^{(\mathrm{R})}$. The magnitude of the delocalization component is of the same order as the largest term obtained in the SAPT approach, therefore aside from greatly diminishing the cost of calculations it also gives an idea of the largest contribution that enters the interaction energy at this level. Meanwhile, the relative balance of $E_{\mathrm{ind,resp}}^{(20)}$, $E_{\mathrm{ex-ind,r}}^{(20)}$ and $\delta E_{int,resp}^{HF}$ is typical for different types of interactions, for example the induction part dominates for hydrogen-bonded dimers.

Similar results are obtained, by design, in the Su-Li EDA scheme, where $\Delta E_{\mathrm{pol}}^{\mathrm{SL}}$ or what is called the polarization contribution contains energetic contributions beyond the first order electrostatic and exchange or repulsive terms. For large basis sets its value approaches that of $\Delta E_{\mathrm{del}}^{(\mathrm{R})}$, however for smaller basis sets it again can differ significantly depending on the exchange energy it complements. The stability of this term is commendable, although differences when moving to larger basis sets in most cases are is still an order of magnitude larger than those of $\Delta E_{\mathrm{del}}^{(\mathrm{R})}$ in the HVPT scheme. In this respect, it is important to remember that even the smallest basis set considered here, aug-cc-pVDZ, can be too costly to consider for large complexes in research, in which case this effect will be even more pronounced.

2.4 Electrostatics – a bidirectional force

Ab initio chemical models consist of atomic nuclei and the electron clouds that surround them. They can be constructed to describe a single atom, functional group, or molecule, and the charge distributions associated with them are superpositions of nuclear charges and electronic densities. Furthermore, one point is usually distinguished in each system and denoted by a vector – for example $\mathbf{R}_A$ and $\mathbf{R}_B$ are the centers of systems A and B.

The electrostatic interaction energy for two systems of static charges is understood as the potential energy derived from the Coulomb forces acting between them. For groups of point charges, this is the sum over all products of pairs of charges divided by the distances between them. In the case of non-discrete charge densities ρ^A and ρ^B for systems A and B,

the electrostatic interaction between them can be expressed by a double integral,

$$\Delta E_{\mathrm{el}}^{(1)} = \iint \frac{\rho_A(\mathbf{r}'_A)\rho_B(\mathbf{r}'_B)}{|\mathbf{r}'_A - \mathbf{r}'_B|} d\mathbf{r}'_A d\mathbf{r}'_B, \tag{2.29}$$

where the primes in $\mathbf{r}'_A$ and $\mathbf{r}'_B$ mean these vectors have the same (arbitrary) origin. Although this definition is unambiguous, the accuracy of the interaction energy depends strongly on the quality of the charge densities. Low quality interaction energies are obtained if inadequate *ab initio* methods or very small basis sets are used to describe the charge distribution.

There are electrostatic contributions at each level of intramolecular correlation within the double perturbation expansion used in SAPT, indexed in (2.9) by n. For example, the first order ($n = 1$) electrostatic contribution $E_{\mathrm{pol}}^{(10)}$, equivalent to $\Delta E_{\mathrm{el}}^{(1)}$ within HVPT, is typically attributed to the Hartree-Fock wave function, whereas the second level electrostatic effect $E_{\mathrm{pol}}^{(20)}$ is interpreted as the effect of second order correlation on the densities of isolated monomers. Therefore, using monomer densities from calculations that retrieve electron correlation (MP2, coupled cluster, etc.) allows this part of the electrostatic effect to be retrieved.[144]

Perhaps the most distinguishing feature of electrostatic interactions, understood as in (2.10), is the fact that they are bidirectional. All other terms of the double perturbation expansion (in (2.9) are either attractive (such as induction or dispersive contributions) or repulsive (exchange and its various combined terms). This means that it is not always possible to change the overall balance of forces using just one of those terms. Electrostatic effects, on the other hand, can be favorable or not depending on the distributions and orientation of the participating charge densities. This simple difference makes electrostatic interactions more likely to be the deciding term in the overall interaction energy.

Electrostatic interaction energies are straightforward to calculate using the formula in (2.29). The double integration, however, requires the charge distributions at each point in space to be known or a means to reproduce it, although methods have been proposed to reduce the computational effort required.[145]

In many situations it is necessary to retain the information contained in the charge density in a compact way. High throughput screening, multiple docking trials, building molecules from fragments – these are all scenarios that repeatedly rely on the same charge distributions, and it is impractical to recompute them. Multipole moments, calculated once and stored for later use, can be employed in such cases to approximate the density and its anisotropy; they are discussed in more detail in the following sections.

Numerous case studies have shown that electrostatic effects can in many cases be used to, at least qualitatively, explain stabilization, specific structural features and various other molecular phenomena, especially in biological systems. The role of electrostatic effects and their practical significance differ from system to system, which means that evaluation and testing what can be reproduced needs to be as much a part of research as the final applications.

Sokalski and coworkers have summarized this methodological mindset and many of the

[144] Sokalski, W. A., Sawaryn, A. *J. Chem. Phys.* **1987**, *87*, 526–534.
[145] Cioslowski, J., Liu, G. H. *Chem. Phys. Lett.* **1997**, *277*, 299–305.

early efforts in a book chapter in 1999.[146] Grembecka et al. demonstrate that the electrostatic interactions of phosphonic leucine analogues with the aminopeptidase active site exhibit excellent correlation with experimental activity,[117] and later expand this study to include phosphoanalogues of phenylalanine.[118]

It is prudent to mention that there is always a certain amount of nonphysical charge redistribution onto the orbitals of ghost atoms when interacting molecules are calculated using a dimer basis set, as in the HVPT method.[147] The effect also exists in other methods, such as SAPT, in general whenever a dimer basis set is employed. Utilization of ghost orbitals provides a larger orbital space and more complete wave function, and alleviates basis set superposition error. The resulting charge redistribution, however, can change and in some cases even destabilize electrostatic interactions, which are calculated using (2.10). Polar or charged molecules are particularly prone to this artefact at short separations, and the influence on the multipole expansion is especially strong since charge density contributions from ghost orbitals need to be moved back to the physically present nuclei. Table 2.3 provides a clear illustration for this artefact in the case of the helium dimer, which should exhibit a zero multipole electrostatic interaction.

The following section follows the Cartesian multipole expansion of the electrostatic interaction energy, and continues with a discussion about improving its resolution by distributing the expansion onto atomic nuclei. Issues of convergence with multipole rank for various distances and basis sets are covered, followed by considerations of conformational changes and charge redistribution effects during chemical reactions.

2.4.1 Multipole expansion in Cartesian coordinates

A multipole expansion of the electrostatic interaction energy in (2.29) is usually described as the sum of all terms in a power series of $\frac{1}{|\mathbf{R}_{AB}|}$, with $\mathbf{R}_{AB} = \mathbf{R}_A - \mathbf{R}_B$ being the distance between the centers of charge distributions. The final form, however, always involves products between spatial moments with appropriate factors, which can be abbreviated using contractions between tensors of these moments and an interaction tensor. The connection can be illustrated clearly in Cartesian coordinates by substituting $\mathbf{r}_A = \mathbf{r}'_A - \mathbf{R}_A$ and $\mathbf{r}_B = \mathbf{r}'_B - \mathbf{R}_B$ and expanding $\frac{1}{|\mathbf{R}_{AB}-(\mathbf{r}_A-\mathbf{r}_B)|}$ into a Taylor series. In vector notation this amounts to

$$\Delta E_{\mathrm{el,mtp}} = \iint \varrho_A(\mathbf{r}_A)\varrho_B(\mathbf{r}_B) \sum_{\kappa=0}^{\infty} \frac{(-1)^\kappa}{\kappa!} \left[((\mathbf{r}_A - \mathbf{r}_B)\cdot\nabla_{\mathbf{R}})^\kappa \frac{1}{|\mathbf{R}|} \right]_{\mathbf{R}=\mathbf{R}_{AB}} d\mathbf{r}_A d\mathbf{r}_B, \qquad (2.30)$$

where $\varrho_A(\mathbf{r}_A) = \rho_A(\mathbf{r}_A + \mathbf{R}_A)$ is the charge density function in (2.29) centered at $\mathbf{R}_A$.

The vector $\mathbf{r}_A - \mathbf{r}_B$ cannot be directly separated from the neighboring nabla operators, because they do not commute and the product $(\mathbf{r}_A - \mathbf{r}_B)\cdot\nabla_{\mathbf{R}}$ is not a vector. It is possible,

[146] Sokalski, W. A., Kędzierski, P., Grembecka, J., Dziekoński, P., Strasburger, K. "Theoretical tools for analysis and modelling electrostatic effects in biomolecules ", In *Computational Molecular Biology, 8*; Leszczynski, J., ed.; Elsevier, 1999, 369–396.

[147] Sokalski, W. A. *J. Chem. Phys.* **1982**, *77*, 4529–4541.

however, to expand the operator $((\mathbf{r}_A - \mathbf{r}_B) \cdot \nabla_{\mathbf{R}})^\kappa$ into a multinomial series of scalar operators and subsequently to separate the coordinates in A and B by expanding $(x_A - x_B)^k$ along with the binomials for the other two axes ($\kappa' = k' + l' + m'$ in analogy to κ),

$$\begin{aligned}((\mathbf{r}_A - \mathbf{r}_B) \cdot \nabla_{\mathbf{R}})^\kappa &= \left((x_A - x_B) \frac{\partial}{\partial R_x} + (y_A - y_B) \frac{\partial}{\partial R_y} + (z_A - z_B) \frac{\partial}{\partial R_z} \right)^\kappa \\ &= \sum_{\substack{k,l,m \\ k+l+m=\kappa}} \frac{\kappa!}{k!l!m!} (x_A - x_B)^k (y_A - y_B)^l (z_A - z_B)^m \frac{\partial^k}{\partial R_x^k} \frac{\partial^l}{\partial R_x^l} \frac{\partial^m}{\partial R_z^m} \\ &= \sum_{\substack{k,l,m \\ k+l+m=\kappa}} \frac{\partial^\kappa}{\partial R_x^k \partial R_y^l \partial R_z^m} \sum_{k'=0}^{k} \sum_{l'=0}^{l} \sum_{m'=0}^{m} \kappa! (-1)^{\kappa-\kappa'} \frac{x_A^{k'} y_A^{l'} z_A^{m'} x_B^{(k-k')} y_B^{(l-l')} z_B^{(m-m')}}{k'!(k-k')!l'!(l-l')!m'!(m-m')!}.\end{aligned} \tag{2.31}$$

Inserting this polynomial back into (2.30) and regrouping yields

$$\Delta E_{\mathrm{el,mtp}} = \sum_{\kappa=0}^{\infty} \sum_{\substack{k,l,m \\ k+l+m=\kappa}} \overbrace{\iint \varrho_A(\mathbf{r}_A) \varrho_B(\mathbf{r}_B) (x_A - x_B)^k (y_A - y_B)^l (z_A - z_B)^m \, d\mathbf{r}_A d\mathbf{r}_B}^{\sum_{k'=0}^{k} \sum_{l'=0}^{l} \sum_{m'=0}^{m} (-1)^{\kappa'} M^A_{k'l'm'} M^B_{k-k',l-l',m-m'}} \times \underbrace{(-1)^\kappa \left[\frac{\partial^\kappa}{\partial R_x^k \partial R_y^l \partial R_z^m} \frac{1}{|\mathbf{R}|} \right]_{\mathbf{R}=\mathbf{R}_{AB}}}_{T_{klm}(\mathbf{R}_{AB})}, \tag{2.32}$$

where M_{klm} denotes an unmodified Cartesian multipole moment,

$$M_{klm} = \frac{1}{k!l!m!} \int \varrho(\mathbf{r}) x^k y^l z^m d\mathbf{r}, \tag{2.33}$$

and the interaction tensor element $T_{klm}(\mathbf{R}_{AB})$ contains the partial derivatives of $|\mathbf{R}_{AB}|^{-1}$,

$$T_{klm}(\mathbf{R}_{AB}) = (-1)^\kappa \left[\frac{\partial^\kappa}{\partial R_x^k \partial R_y^l \partial R_z^m} \frac{1}{|\mathbf{R}|} \right]_{\mathbf{R}=\mathbf{R}_{AB}}. \tag{2.34}$$

The Cartesian tensor element in turn can be expressed explicitly in the form of a triple sum, as derived by Cipriani and Silvi[148] and extended by Challacombe et al.[149]

$$T_{klm}(\mathbf{R}) = \frac{(-1)^\kappa k!l!m!}{2^\kappa |\mathbf{R}|^\kappa} \sum_{s=0}^{[k/2]} \sum_{t=0}^{[l/2]} \sum_{u=0}^{[m/2]} \left[\frac{(-1)^\sigma (2\kappa - 2\sigma)!}{s!t!u!(k-2s)!(l-2t)!(m-2u)!(\kappa-\sigma)!} \times \left(\frac{\mathrm{R}_x}{|\mathbf{R}|} \right)^{k-2s} \left(\frac{\mathrm{R}_y}{|\mathbf{R}|} \right)^{l-2t} \left(\frac{\mathrm{R}_z}{|\mathbf{R}|} \right)^{l-2m} \right]. \tag{2.35}$$

[148] Cipriani, J., Silvi, B. *Mol. Phys.* **1982**, *45*, 259–272.
[149] Challacombe, M., Schwegler, E., Almlöf, J. *Chem. Phys. Lett.* **1995**, *241*, 67–72.

Using the symbols M_{klm} and $T_{klm}(\mathbf{R}_{AB})$ in the context suggested by the braces, (2.32) can be rewritten in a simpler form,

$$\Delta E_{\text{el,mtp}} = \sum_{\kappa=0}^{\infty} \sum_{\substack{k,l,m \\ k+l+m=\kappa}} \sum_{k'=0}^{k} \sum_{l'=0}^{l} \sum_{m'=0}^{m} (-1)^{\kappa'} M^{A}_{k'l'm'} T_{klm}(\mathbf{R}_{AB}) M^{B}_{k-k',l-l',m-m'}. \tag{2.36}$$

An even more succinct expression can be presented with tensor notation repalcing the multiple sums, similar to that used by Jansen[150],

$$\Delta E_{\text{el,mtp}} = \sum_{\kappa_a}^{\infty} \sum_{\kappa_b}^{\infty} \mathbf{M}_A^{(\kappa_a)} [\kappa_a] \mathbf{T}^{(\kappa_a+\kappa_b)} [\kappa_b] \mathbf{M}_B^{(\kappa_b)}, \tag{2.37}$$

where κ is called the *rank* or *order* of a multipole ($\mathbf{M}_A^{(\kappa_a)}$ is of rank κ_a) or interaction ($\mathbf{T}^{(\kappa_a+\kappa_b)}$ is of rank $\kappa_a + \kappa_b$) and the operator $[\kappa]$ means that the product of the tensors on both sides is contracted κ times.

The first few terms in this tensor expansion can also be expressed using summations over pairs of indexes, where $T_{\alpha_1 \ldots \alpha_\kappa}$ represents the various elements of $\mathbf{T}^{(\kappa)}$, with $\alpha = x, y, z$,

$$\begin{aligned} \Delta E_{\text{el,mtp}} = \; & T q^A q^B + T_\alpha \left(q^A \mu^B_\alpha - \mu^A_\alpha q^B \right) + T_{\alpha\beta} \left(q^A \Theta^B_{\alpha\beta} + \Theta^B_{\alpha\beta} q^B - \mu^A_\alpha \mu^B_\alpha \right) + \\ & T_{\alpha\beta\gamma} \left(q^A \Omega^B_{\alpha\beta\gamma} - \Omega^A_{\alpha\beta\gamma} q^B - \mu^A_\alpha \Theta^B_{\beta\gamma} + \Theta^A_{\beta\gamma} \mu^B_\alpha \right) + \ldots \end{aligned} \tag{2.38}$$

and the multipole moment symbols used correspond to the multipole moment tensors $\mathbf{M}^{(\kappa)}$:

$$\begin{aligned} q = \mathbf{M}^{(0)} \quad & \int \rho(\mathbf{r}) d\mathbf{r} \\ \mu_\alpha = \mathbf{M}^{(1)} \quad & \int \rho(\mathbf{r}) \alpha d\mathbf{r} \qquad (\alpha = x, y, z) \\ \Theta_{\alpha\beta} = \mathbf{M}^{(2)} \quad & \int \rho(\mathbf{r}) \alpha\beta d\mathbf{r} \qquad (\alpha, \beta = x, y, z) \\ \Omega_{\alpha\beta\gamma} = \mathbf{M}^{(3)} \quad & \int \rho(\mathbf{r}) \alpha\beta\gamma d\mathbf{r} \qquad (\alpha, \beta, \gamma = x, y, z) \\ & \text{etc. ...} \end{aligned} \tag{2.39}$$

In this way, for ranks $\kappa = 0, 1, 2, 3, \ldots$, the subsequent multipole moment tensors $\mathbf{M}^{(\kappa)}$ contain elements M_{klm} such that $k+l+m = \kappa$. These represent the point charge and components of the dipole vector, quadrupole matrix, three-dimensional octupole with 27 moments and so forth in higher dimensions:

$$\begin{aligned} \kappa = 0 \quad & \mathbf{M}^{(0)} \quad q = M_{000} \\ \kappa = 1 \quad & \mathbf{M}^{(1)} \quad \mu = (M_{100}, M_{010}, M_{001}) \\ \kappa = 2 \quad & \mathbf{M}^{(2)} \quad \Omega = \begin{pmatrix} M_{200} & M_{110} & M_{101} \\ M_{110} & M_{020} & M_{011} \\ M_{101} & M_{011} & M_{002} \end{pmatrix}. \end{aligned} \tag{2.40}$$

[150] Jansen, L. *Phys. Rev.* **1958**, *110*, 661–669.

The interaction expression in (2.36) can also be applied to traceless moments as defined by Buckingham,[151]

$$\bar{M}^A_{klm} = (-1)^\kappa \frac{1}{(\kappa)!} \int d\mathbf{r}\rho(\mathbf{r})|\mathbf{r}|^{2\kappa+1} \frac{\partial^\kappa}{\partial x^k \partial y^l \partial z^m} \frac{1}{|\mathbf{r}|}, \tag{2.41}$$

which should replace the unmodified moment along with the extra factor of $2^\kappa\kappa!/(2\kappa)!$, and the interaction tensor elements remain unchanged.

It is expedient at this point to make a note that $\frac{1}{|\mathbf{R}-\mathbf{r}|}$ is not an entire function and its Taylor expansion around $\mathbf{R}$ converges only for $\mathbf{r} < \mathbf{R}$ (in the opposite case an expansion around $\mathbf{r}$ is convergent). In the present context this means that $(\mathbf{r}_A - \mathbf{r}_B)$ should be smaller than $\mathbf{R}_{AB}$. Since integration is to be performed *over all space* and the charge distributions extend to infinity, the series in (2.30) will generally not be convergent everywhere. The complete procedure of using four different Taylor series for the expansion of Coulomb interactions between two charge distributions is imaginable, but tedious and absent from the literature. The complicated boundary conditions are usually implicitly avoided by restricting $\mathbf{R}$ to relatively large values, so that at least one charge distribution ϱ will be small enough to mitigate the divergent character of the integrand at large values of $(\mathbf{r}_A - \mathbf{r}_B)$. If the overlap between ϱ_A and ϱ_B is minimal in this portion of space, the integral over even a divergent expansion will behave asymptotically for small values of κ. In practice the expansion is always truncated at some point, and in the ideal case should behave asymptotically for all of the values of κ used.

Therefore, the order at which the multipole expansion is truncated is usually determined by κ, and only interactions between M^A_{klm} and $M^B_{k'l'm'}$ are included in the interaction energy for which $k + l + m + k' + l' + m'$ is less or equal to some limiting value L. The multipole interaction energy calculated in this sense at order (or rank) L will be denoted by $\Delta E^{\kappa\leqslant L}_{\text{el,mtp}}$ and is given in tensor notation by:

$$\Delta E^{\kappa\leqslant L}_{\text{el,mtp}} = \sum_{\substack{\kappa_a \\ \kappa_a+\kappa_b\leqslant L}} \sum_{\kappa_b} \mathbf{M}^{(\kappa_a)}_A[\kappa_a]\mathbf{T}^{(\kappa_a+\kappa_b)}[\kappa_b]\mathbf{M}^{(\kappa_b)}_B. \tag{2.42}$$

There is another way of truncating the multipole expansion, namely by including all possible interaction terms between available moments. For example, if moments up to order $\kappa = 2$ were produced (quadrupoles), this approach would include an incomplete fourth order interaction due to the terms involving quadrupole moments from both expansions. Throughout this work this moment-based truncated energy will be denoted by $\Delta E^{\kappa'\leqslant L}_{\text{el,mtp}}$,

$$\Delta E^{\kappa'\leqslant L}_{\text{el,mtp}} = \sum_{\kappa_a}^{L} \sum_{\kappa_b}^{L} \mathbf{M}^{(\kappa_a)}_A[\kappa_a]\mathbf{T}^{(\kappa_a+\kappa_b)}[\kappa_b]\mathbf{M}^{(\kappa_b)}_B. \tag{2.43}$$

Equation 2.23 already introduced the multipole expanded part of the electrostatic interaction and its counterpart penetration term $\Delta E_{\text{el,pen}}$, which arises from the non-zero overlap

[151] Pages 2-62 in Buckingham, A. D. *Intermolecular Interactions: From Diatomics to Biopolymers*, Pullman, B., ed.; John Wiley and Sons: New York, 1987.

between charge distributions. Together, at finite distances these constitute the total electrostatic energy:

$$\Delta E_{\mathrm{el}}^{(1)} = \Delta E_{\mathrm{el,mtp}} + \Delta E_{\mathrm{el,pen}}. \tag{2.44}$$

Since for large separations between the interacting charge systems the overlap of charge densities derived from Gaussian functions decays, the total electrostatic interaction $\Delta E_{\mathrm{el}}^{(1)}$ always tends to the multipole moment component described above:

$$\lim_{\mathbf{R}\to\infty} \Delta E_{\mathrm{el}}^{(1)} = \Delta E_{\mathrm{el,mtp}}. \tag{2.45}$$

Almost all interaction scenarios that are of practical interest, and hydrogen bonds in particular, involve contacts for which penetration effects are not entirely negligible. The value of the penetration term is hard to establish accurately, since a converged multipole expansion is needed in order to estimate it from (2.44), although it depends largely on the shape of a molecule and not on the local anisotropy of the electron distribution. Therefore, the assumption that the angular dependence of the electrostatic interaction energy is captured in the multipole term is reasonable.

An interesting approach to modeling overall electric effects has been proposed by Qian and Krimm,[152] who use a multidimensional parametrization of the charge density and response functions to reproduce the interaction of a molecule with a point charge. Such a model includes both the multipole and penetration parts of the electrostatic interaction, as well as other effects arising from polarization and hyperpolarization. The authors also suggest their parametrization could be expanded in a distributed fashion, thus making the description as mobile as the classic multipole approach.

2.5 Enhanced electrostatic resolution with atomic multipoles moments

It is well established that the charge distribution of an entire molecule does not decay fast enough in order for its multipole expansion to converge at van der Waals distances. The charge density is therefore often partitioned spatially and expanded into sets of local multipole moments. It is typical to center these fragment on atoms, or on the bonds between them.

The advantage of using atomic over molecular moments has been demonstrated repeatedly in the literature. Grembecka et al., for example, have modeled the activity of leucine aminopeptidase (LAP) inhibitors using wave function-based as well as potential-derived atomic charges.[153] Various distributed moments have also been used to enhance the accuracy of reaction field models[154] and Coulomb interactions in general.[155]

While the density can be partitioned between atoms or other centers in various ways,

[152] Qian, W., Krimm, S. *J. Mol. Struct.* **2006**, *766*, 93–104.
[153] Grembecka, J., Kędzierski, P., Sokalski, W. A., Leszczyński, J. *Int. J. Quant. Chem.* **2001**, *83*, 180–192.
[154] Rinaldi, D., Bouchy, A., Rivail, J.-L., Dillet, V. *J. Chem. Phys.* **2004**, *120*, 2343–2350.
[155] Popelier, P. L. A., Kosov, D. S. *J. Chem. Phys.* **2001**, *114*, 6539–6547.

some of which are mentioned below, the general use of atomic moments for evaluating the interaction energy remains the same. The multipole electrostatic interaction between two systems A and B that are described by N_A and N_B multipole expansions, respectively, is the sum of interactions between each pair of expansions in the two systems. Extending (2.37) in this way gives

$$\Delta E^{\kappa\leqslant L}_{\mathrm{el,mtp}} \simeq \sum_{i\in A}^{N_A}\sum_{j\in B}^{N_B} \Delta E^{ij,\kappa\leqslant L}_{\mathrm{el,mtp}} = \sum_{\substack{\kappa_a \\ \kappa_a+\kappa_B\leqslant L}}\sum_{\kappa_b} \left(\sum_{i\in A}^{N_A}\sum_{j\in B}^{N_B} \mathbf{M}_i^{(\kappa_a)}[\kappa_a]\mathbf{T}^{(\kappa_a+\kappa_b)}_{|\mathbf{R}_i-\mathbf{R}_j|}[\kappa_b]\mathbf{M}_j^{(\kappa_b)} \right), \tag{2.46}$$

where the interaction tensor needs to be evaluated for each pair of centers separately.

2.5.1 Methods for partitioning the electron density

In the course of describing a charge distribution with multiple expansion centers, an inevitable step is choosing the method for partitioning the charge density between them. This is always an arbitrary choice and influences the usability of the resulting set of multipole expansions at intermediate distances. The literature is abundant with different approaches to this problem, while only a few have been implemented, made public and are currently in use. For example, the distributed multipole analysis (DMA) proposed by Stone and Alderton[156] considers each product of primitive functions that contribute to the charge density and its corresponding multipole expansion. Each such product is centered at a unique point in space, and is ascribed to a final expansion center that is nearest to this point. The closer a product is moved from its original position the less it destabilizes the interaction energy in (2.46). Thus, the solution offered by DMA maximizes the region of convergence for any chosen set of expansion centers.

An alternative approach to partitioning the charge density has been provided by Bader, and is based on topological considerations. His theory of atoms in molecules (AIM) establishes boundaries for individual atoms, defined by surfaces on which the flux of the gradient charge density vector field $\nabla\rho$ vanishes,[157]

$$\nabla\rho(\mathbf{r})\cdot\mathbf{n}(\mathbf{r}) = 0, \tag{2.47}$$

where $\mathbf{n}(\mathbf{r})$ is the normal to that surface at point $\mathbf{r}$.

This approach elegantly removes the arbitrariness of partitioning the electron density by applying a generalized least action principle. Specifically, the electron density is formulated as the expectation value of a quantum mechanical observable (the density operator). The idea, nonetheless, has raised controversy in the literature in the past few years,[158] not without

[156] Stone, A. J., Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.

[157] Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893–928; Bader, R. F. W. *Monatshefte für Chemie* **2005**, *136*, 819–854; Bader, R. F. W. *J. Phys. Chem. A* **2007**, *111*, 7966–7972.

[158] Frenking, G. *Angew. Chem. Int. Ed.* **2003**, *42*, 143–147; Parr, R. G., Ayers, P. W., Nalewajski, R. F. *J. Phys. Chem. A* **2005**, *109*, 3957–3959; Poater, J., Sola, M., Bickelhaupt, F. M. *Chem. Eur. J.* **2006**, *12*, 2902–2905.

response.[159] Bader and Matta have also rebutted various criticisms concerning the charges obtained within AIM, relying invariably on quantum mechanical principles.[160]

AIM theory naturally and uniquely defines bond critical points along bond paths that connect atoms[161] and atomic volumes,[162] which are the basis for integrating atomic charges, polarizabilities[163] or magnetic properties.[164] Multipole integrals can be evaluated over the atomic basins in molecules. For atom i and its three-dimensional basin Ω_i, the Cartesian moment of rank klm will be:

$$M_{klm,i}^{AIM} = \int_{\Omega_i} x^k y^l z^m \rho(\mathbf{r}) d\mathbf{r}. \tag{2.48}$$

While Bader's original method involves identifying surfaces of zero flux, Popelier generates similar basins by following gradient paths from each point in space to an attractor (usually a nucleus).[155] Fuzzy solutions are also possible, such as Hirshfeld's prescription from 1977 that distributes the charge density to atoms based on their contribution to the promolecule density at each point.[165]

Volkov and Coppens have evaluated the performance of AIM and Hirshfeld moments for amino acids[166] and compared them to pixel-by-pixel summation and the electrostatic interaction obtained from a Morokuma-Ziegler decomposition.[110] Their main conclusion was that all these methods could be used, with some reservations, to reproduce the relative strength of bonding.

In the work presented here, atomic moments are generated from densities partitioned the same way as in the Mulliken population analysis. Following Sokalski and others,[167] the starting point is the expectation value of the $x^k y^l z^m$ operator within the LCAO MO approach, written as a sum of products of any two atomic orbitals I and J. For such products $\langle I|x^k y^l z^m|J\rangle$ is a multipole integral over the electron density related to one or two atoms which hold the orbitals I and J. These atoms become the beneficiaries of this one contribution, effectively segregating the molecular moments into atomic contributions:

$$\begin{aligned} \langle x^k y^l z^m \rangle &= \sum_i Z_i x_i^k y_i^l z_i^m - \sum_I^{N_{AO}} \sum_J^{N_{AO}} P_{IJ} \langle I|x^k y^l z^m|J\rangle \\ &\equiv \sum_i \langle x^k y^l z^m \rangle_i = \sum_i \left(Z_i x_i^k y_i^l z_i^m - \sum_{I\in i} \sum_J^{N_{AO}} P_{IJ} \langle I|x^k y^l z^m|J\rangle \right), \end{aligned} \tag{2.49}$$

[159] Bader, R. F. W. *Int. J. Quant. Chem.* **2003**, *94*, 173–177; Bader, R. F. W. *Chem. Eur. J.* **2006**, *12*, 2896–2901; Bader, R. F. W., Matta, C. F. *J. Phys. Chem. A* **2006**, *110*, 6365–6371.

[160] Bader, R. F. W., Matta, C. F. *J. Phys. Chem. A* **2004**, *108*, 8385–8394.

[161] Bader, R. F. W. *J. Phys. Chem. A* **1998**, *102*, 7314–7323.

[162] Bader, R. F. W., Carroll, M. T., Cheeseman, J. R., Chang, C. *J. Am. Chem. Soc.* **1987**, *109*, 7968–7979.

[163] Laidig, K. E., Bader, R. F. W. *J. Chem. Phys.* **1990**, *93*, 7213–7224; Bader, R. F. W., Matta, C. F. *Int. J. Quant. Chem.* **2001**, *85*, 592–607.

[164] Bader, R. F. W., Keith, T. A. *J. Chem. Phys.* **1993**, *99*, 3683–3693.

[165] Hirshfeld, F. *Theor. Chim. Acta* **1977**, *44*, 129–138.

[166] Volkov, A., Coppens, P. *J. Comp. Chem.* **2004**, *25*, 921–934.

[167] Sokalski, W. A., Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86–92; Sokalski; Sawaryn, 1987, in Ref. 144 on page 33; Sawaryn, A., Sokalski, W. A. *Comput. Phys. Commun.* **1989**, *52*, 397–408.

where i spans all atoms and P_{IJ} is an element of the density matrix in which all off-diagonal elements are halved. It should be stressed that $\langle x^k y^l z^m \rangle_i$ is not an expectation value in the sense its sum for the molecule is It is not obtainable as an average using any eigenfunction of the Hamiltonian. Nonetheless, from (2.49) it follows that an atomic multipole moment of rank klm can be defined as

$$M_{klm,i} = \langle x^k y^l z^m \rangle_i = Z_i x_i^k y_i^l z_i^m - \sum_{I \in i} \sum_{J}^{N_{AO}} P_{IJ} \langle I | x^k y^l z^m | J \rangle . \tag{2.50}$$

2.5.2 Cumulative atomic moments

The Cartesian atomic moments defined in (2.50) are all calculated relative to the same origin, and therefore can be directly summed into molecular moments that are also centered at that origin. It is usually beneficial to move the moments to their local atomic coordinate systems, which can be done through coordinate substitution and an iterative recombination of the moments:

$$M_{klm,i}^{\mathrm{CAMM}} = M_{klm,i} - \sum_{\substack{k'\geq 0 \\ k'l'm' \neq klm}}^{k} \sum_{l'\geq 0}^{l} \sum_{m'\geq 0}^{m} \begin{pmatrix} k \\ k' \end{pmatrix} \begin{pmatrix} l \\ l' \end{pmatrix} \begin{pmatrix} m \\ m' \end{pmatrix} \times x_i^{k-k'} y_i^{l-l'} z_i^{m-m'} M_{k'l'm',i}^{CAMM} . \tag{2.51}$$

The resulting moments have been called *cumulative atomic multipole moments* (CAMM) when derived from the wave function projection population analysis in (2.50),[167] although the name can be applied to moments based on any other density partitioning. Besides being invariant with respect to translation, these *cumulative* atomic moments can still be easily rotated while centered on atoms. A moment rotated from an orientation O to a new orientation $\tilde{O}$ is simply the product of the appropriate power of the rotation matrix $R_{O \to \tilde{O}}$ and the original moment tensor, which defines the rotation operator $\hat{O}$:

$$\hat{O} \left(\mathbf{M}^{\kappa} \right) = \left(R_{O \to \tilde{O}} \right)^{\kappa} \times \mathbf{M}^{\kappa} . \tag{2.52}$$

2.5.3 Convergence properties of the atomic multipole expansion

Using molecular multipole moments, for which both N_A and N_B in (2.46) equal one, is inadequate for studying most noncovalently bound systems. There are two main reasons, which were summarized by Stone and Alderton 25 years ago[156] (in an introduction to the DMA method that was reprinted in 2002[168]). Namely, they cannot be used to calculate potentials or interaction energies at van der Waals distances. They also reveal little or no information about the topology of the charge distribution inside or in close proximity to the molecule. The first limitation has been reiterated a number of times, recently by Qian and Krimm in their search of a general charge density approach for hydrogen bonds.[169]

[168] Stone, A. J., Alderton, M. *Mol. Phys.* **2002**, *100*, 221–233.
[169] Qian, W., Krimm, S. *J. Phys. Chem. A* **2005**, *109*, 5608–5618.

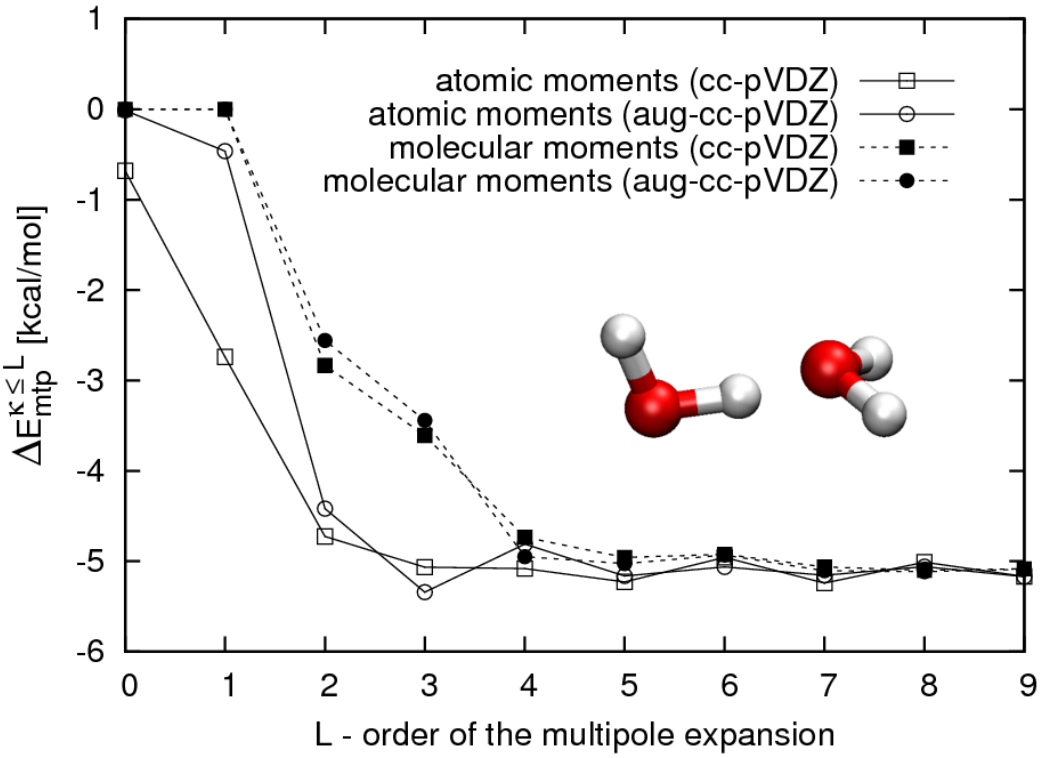


Figure 2.3: Comparison of molecular and atomic multipole expansions of the electrostatic interaction energy for two water molecules. The dimer was set in the equilibrium geometry reported by Szalewicz et al..[a]

[a]Szalewicz, K., Cole, S. J., Kołos, W., Bertlett, R. J. *J. Chem. Phys.* **1988**, *89*, 3662 – 3673.

There have been relatively few systematic attempts to evaluate the convergence behavior of electrostatic multipole interactions, considering the large body of literature that makes use of the methodology. It is customary to show the moments themselves at various orders and present the energy for a single value of L as defined by (2.42) or (2.43). While it is true that there is a correspondence between the magnitude of multipole moments and their interactions, it is not linear. Even if the first ostensibly diverge, that does not imply the divergence of interactions. Moreover, the potentials and interactions entailed by a multipole expansion can by definition vary widely with distance and orientation, so it is important to analyze in detail those configurations that are of interest in a specific system. At the very least it is practical to know at what order the multipole expansion can be expected to converge.

It is worth noting studies that do report on the convergence of multipole interactions. With the CAMM approach used in this work, in their first introductory article Sokalski et al. provide interactions energies obtained from atomic moments of various orders.[167] More recently, in our study on DNA intercalators,[129] the change in atomic and molecular multipole interactions is presented to an order of nine in the supporting information (see Section 4).

Sagui et al., using maximally localized Wannier functions to partition the charge density, demonstrated how the electrostatic potential around water and carbon dioxide converge with the exapnsion rank and distance.[170] Within the DMA method, Stone follows the difference

[170]Sagui, C., Pomorski, P., Darden, T. A., Roland, C. *J. Chem. Phys.* **2004**, *120*, 4530–4544.

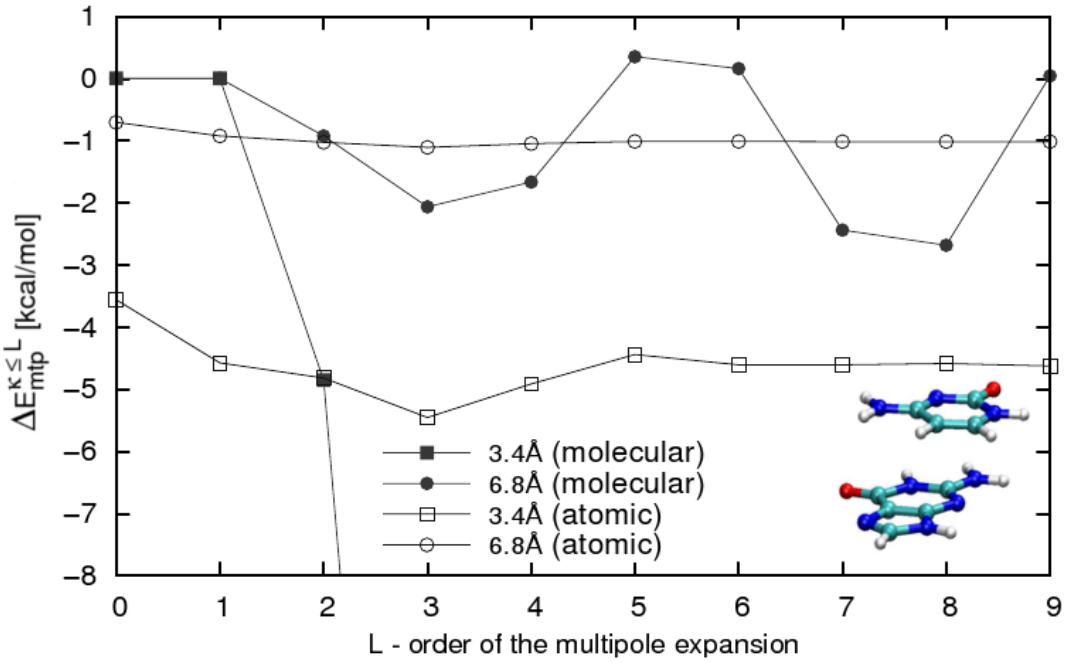


Figure 2.4: Comparison of molecular and atomic multipole expansions of the electrostatic interaction energy for a stacked cytosine-guanine dimer. The geometries at the two presented separations (3.4 Å and 6.8 Å) differed only in the distance between centers of mass.

in electrostatic potential at various orders when revising the method for improved basis set stability.[171] Earlier, Hodges and Wheatley also considered truncation at various orders for two interacting HF molecules.[172]

Much work has been done in this regard by Popelier and coworkers using topological atoms based on AIM and similar approaches, from which it is clear that further issues arise when convergence properties are studied. Using a hard-sphere repulsion potential and multipole moment interactions, they show how multipole interactions change for various small dimers at geometries optimized using multipole expansions of consecutive orders.[155] In another study, Popelier and Rafat illustrate the divergence of the potential obtained by a multipole expansion for orders up to 18, and also consider the use of Bessel function moments to obtain a convergent expansion where the traditional Taylor expansion fails.[173] The same authors have also proposed a generalization that includes the inverse formulation of the multipole expansion to attain a convergent series at all points in space.[174] In another work,[175] they consider displacing atomic moments from their origin on atoms in order to improve the convergence of interactions between nearby atoms. Later they focus on characterizing the convergence properties at large and medium distances.[176]

Here, the point is first illustrated for the water dimer in Fig. 2.3, which contains original results for the Cartesian multipole expansion of the electrostatic interaction energy as described

[171]Stone, A. J. *J. Chem. Theor. Comp.* **2005**, *1*, 1128–1132.
[172]Hodges, M. P., Wheatley, R. J. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1631–1638.
[173]Popelier, P. L. A., Rafat, M. *Chem. Phys. Lett.* **2003**, *376*, 148–153.
[174]Rafat, M., Popelier, P. L. A. *J. Chem. Phys.* **2005**, *123*, 10.1063/1.2126591.
[175]Rafat, M., Popelier, P. L. A. *J. Chem. Phys.* **2006**, *124*, 144102–7.
[176]Rafat, M., Popelier, P. L. A. *J. Comp. Chem.* **2007**, *28*, 832–838.

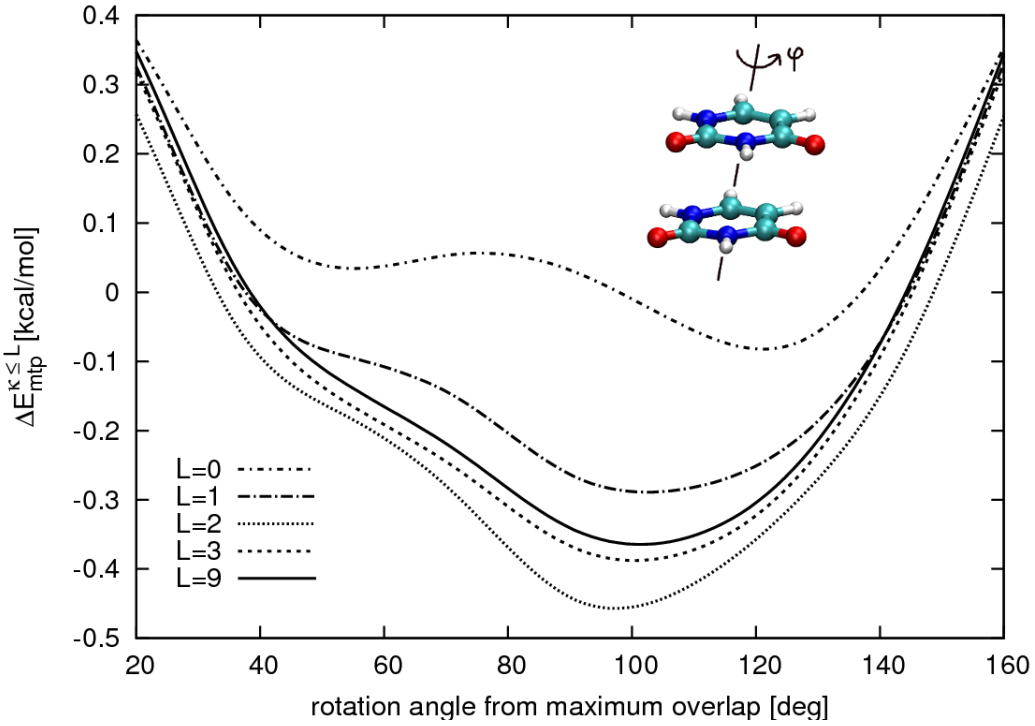


Figure 2.5: Atomic multipole expansions of the electrostatic interaction energy for a stacked uracil dimer with a vertical separation of 6.8Å, plotted against the relative twist of one molecule. Zero corresponds to maximal overlap (ideal stack), and the rotation angle is around the pole connecting centers of mass.

in Section 2.4.1. The molecular expansion, containing one center on each water molecule, stabilizes when moments up to the fourth order (hexadecapoles) are used, while the atomic expansion reaches values near -6 kcal/mol already for $\kappa = 2$ or quadrupoles. Also evident – when comparing the plots for the cc-pVDZ and aug-cc-pVDZ bases – is how little extra diffuse functions in the basis set affect the convergence and final value of the multipole interaction. The multipole and total electrostatic interactions in this case are weaker than for the water dimer in Table 2.3, which is a consequence of using different geometries. The latter included relaxed monomers, unlike the reference geometry pf Fig. 2.3.

A contrasting example is shown in 2.4, namely a stacked dimer containing cytosine and guanine in a conformation typical for B-DNA. The convergence behavior of the CAMM expansion with orders up to $L = 9$ is shown for the biologically relevant intermolecular separation of 3.4 Å and an increased distance of 6.8 Å. In this case, the molecular expansion is strictly divergent for orders larger than one. The molecular charges of these molecules are zero, which means that there are no charge-charge and charge-dipole interactions. It can be noted that the second rank molecular multipole interaction $\Delta E^{\kappa\leqslant 2}_{\mathrm{CAMM}}$, which in this case contains only dipole-dipole interactions, is still moderate, nonetheless the molecular expansion is evidently at best unreliable at biological separations (3.4 Å). The atomic expansion on the other hand reaches a stable value only when moments with orders above five are included. In the second, more distant configuration, the atomic multipole interaction energy is relatively stable already for $L > 1$ and its molecular counterpart does not diverge but oscillates around it.

Fig. 2.5 in turn shows the multipole interaction at various orders for a stacked uracil dimer,

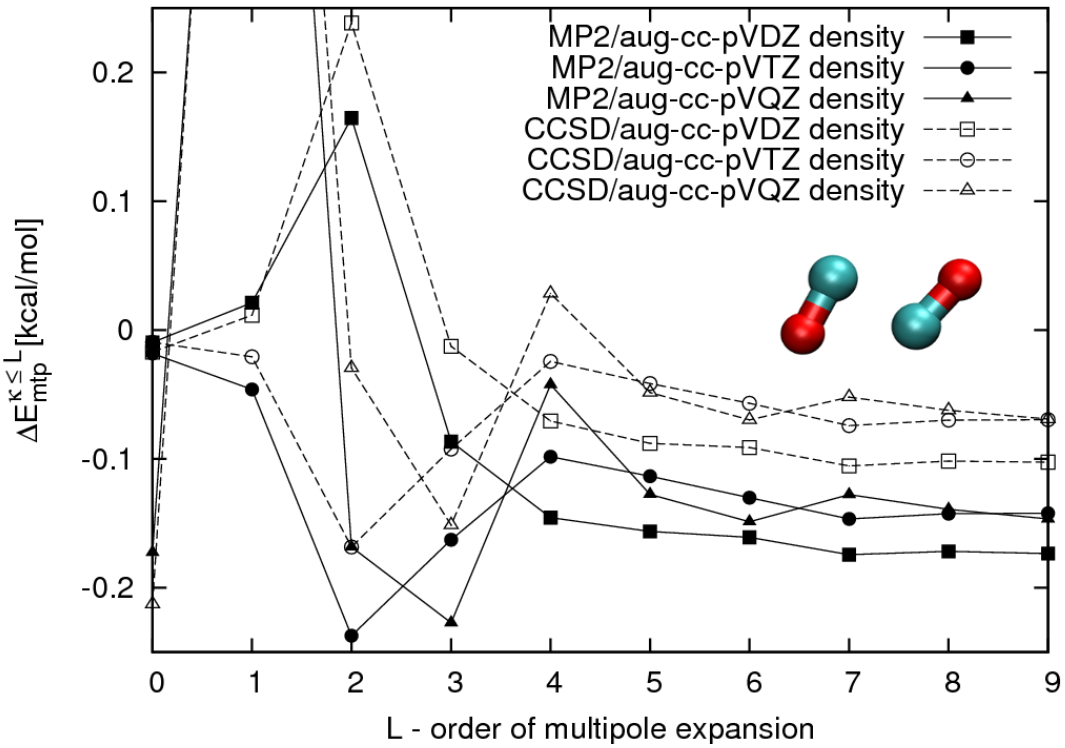


Figure 2.6: Atomic multipole interaction energies in the carbon monoxide dimer evaluated from MP2 and CCSD densities using aug-cc-pVXZ (X=D,T,Q) basis sets.

in which one molecule was rotated around the axis connecting their centers of mass. The initial, maximal overlap configuration is the least favorable one due to the repulsion of like-charges and an optimal rotation is seen for an approximate antiparallel alignment of molecular dipole moments, which is in line with *ab initio* results. The point to be made here is, again, that the atomic multipole expansion converges for orders above L=3. The higher order moments are particularly influential in the attractive region around 180°, which in turn means that the strictly repulsive regions can be eliminated using charges only, if one is performing a crude conformational scan.

The stability of distributed multipole moments with increasing basis set size is also worthy of attention. When the electron density is partitioned in Hilbert space (basis function space) large changes are often observed for particular atomic moments when changing the basis set, while the density around an atom or even the entire molecule does not change substantially. This was the incentive for Stone to recently introduce an integration step when partitioning products of diffuse basis function in the DMA method.[171]

Fig. 2.6 on the other hand illustrates how the multipole interaction converges with increasing expansion order L in the case of the carbon monoxide dimer. Clearly, the energies stabilize only for expansion orders of $L > 3$, and the differences between the interaction of moments generated from MP2 and coupled cluster densities are roughly constant. Possibly this difference corresponds to the correction introduced by Hobza.[177]

[177] Jurečka, P., Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613.

2.5.4 Conformational dependence and fragment transferability

A discussion of multipole moments would be incomplete if it did not touch upon the twin issues of conformational dependence and transferability. The main advantage, in fact, of describing a charge distribution with distributed multipole moments lies in their mobile character and the possibility of using the same moments in systems built from identical or similar molecules. An ideal would be to have a library of reusable fragments, ready to be assembled in any number of combinations. In case where moments for rigid molecules suffice, the procedure for this is straightforward and atomic moments can be taken without modification, aside from translation or rotation into the new environment.

Problems arise when this paradigm is confronted with systems that consist of slightly different molecules, covalently bound fragments, or when molecules change conformations. In the most interesting situations, all these problems need to be dealt with simultaneously, so in general transferability and conformational dependence (or polarizability) are connected. Practical and largely unanswered questions come to mind on this topic, three of which seem to be the most basic:

- how can the conformational dependence of atomic multipole moments be described and how does it influence its interactions with other molecules?
- does the conformational dependence vary with intermolecular distance?
- what is the best way to deal with different residual charges on atoms or fragments when transferring them into another molecule?

Moreover, it is obvious that the answers to these questions can vary between systems, which makes it important to have the possibility to evaluate them repeatedly in an automated way.

In an effort to answer some of these questions, Kędzierski and Sokalski generated a library of uncorrelated and correlated atomic moments for all natural amino acids and tested in detail how well they reproduce the molecular electrostatic potential (MEP) on the solvent accessible surface.[178] They concluded that the transferability of amino acid fragments between molecules is the best in cases with high symmetry, because the main source of non-transferability is the unbalanced residual charge on transferred fragments. An earlier study[179] shows that torsional potential barriers in molecules with elongated bonds can be qualitatively reproduced using atomic multipole moments. Strasburger with Sokalski[180] later extended this treatment further by neglecting inter-fragment density contributions.

Transferability has also been considered for atomic and bond properties within the AIM approach. Lopez et al. have demonstrated good transferability of various properties for a series of *linear* alkanenitriles,[181] in particular the atomic first moment of the charge density or dipole moment. A study by Rafat et al. on the other hand assesses how transferring moments

[178] Kędzierski, P., Sokalski, W. A. *J. Comp. Chem.* **2001**, *22*, 1082–1097.
[179] Sokalski, W. A., Lai, J., Luo, N., Sun, S., Shibata, M., Ornstein, R. L., Rein, R. *Int. J. Quant. Chem.* **1991**, 61–71.
[180] Strasburger, K., Sokalski, W. A. *Chem. Phys. Lett.* **1994**, *221*, 129–135.
[181] Lopez, J. L., Mandado, M., Grana, A. M., Mosquera, R. A. *Int. J. Quant. Chem.* **2002**, *86*, 190–198.

from isolated water molecules into their clusters influences the interaction energy, thus bearing information about the polarization induced by hydrogen bonds.[182] Whitehead et al. have developed a method of reconstructing molecular electrostatic potentials from previously calculated atomic multipole moments.[183] A novel, robust approach to predicting multipole moments for various molecular conformations based on neural nets has also been proposed by Popelier and coworkers.[184] A section in a recent review article by Bushmarinov et al. summarizes some of these studies and other related ideas.[185]

Stone and others have demonstrated that local torsional changes, for example rotations around bonds directly related to an atom, strongly influence multipole moments and propose to interpolate this dependence using short Fourier series.[186] Hodges and Wheatley on the other hand, for hydrogen fluoride, attain accurate results by fitting multipole moments with polynomial functions of the stretching coordinate.[172] Plattner and Meuwly on the other hand have recently evaluate the bond length dependence of multipole moments in carbon monoxide and included their interactions in molecular dynamics simulations.[187]

An interesting comparison of methods for partitioning the charge density was published by Pacios and Gomez,[188] in which they study the values they give for atomic and fragment charges in all the theoretical gaseous conformers of glycine. Heutz et al. in turn point out large variations in atomic charges after conformational changes in dioctadecylamine,[189] and Söderhjelm and Ryde study the conformational dependence of atomic ESP charges in proteins during molecular dynamics simulations.[190]

Overall, these various efforts demonstrate the transferability and reusable character of distributed moments in various contexts, and a few specific sets of atomic moments have been published that could be used in practice.[178] Nonetheless, there is no general framework for recycling atomic moments and a lack of guiding principles to tackle practical problems when conformational changes take place.

A major potential application of atomic multipole moments is enhancing the electrostatic interactions in molecular dynamics simulations. Although only few studies exist, they already show that incorporating multipole moments in the force field model leads to interesting results and improves the quality of simulations. It can be expected that in certain cases improving electrostatic interactions in this way will increase the accuracy of simulations. This is already seen in the advantages gained by using polarizable force field over point charge models, demonstrated for DNA by Sagui and coworkers.[191]

[182]Rafat, M., Shaik, M., Popelier, P. L. A. *J. Phys. Chem. A* **2006**, *110*, 13578–13583.
[183]Whitehead, C. E., Breneman, C. M., Sukumar, N., Ryan, M. D. *J. Comp. Chem.* **2003**, *24*, 512–529.
[184]Darley, M. G., Handley, C. M., Popelier, P. L. A. *J. Chem. Theor. Comp.* **2008**, *4*, 1435–1448.
[185]Bushmarinov, I. S., Lyssenko, K. A., Antipin, M. Y. *Russian Chem. Rev.* **2009**, *78*, 283–302.
[186]Koch, U., Popelier, P. L. A., Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253–260; Koch, U., Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1701–1708.
[187]Plattner, N., Meuwly, M. *Biophys. J.* **2008**, *94*, 2505–2515.
[188]Pacios, L. F., Gomez, P. C. *J. Mol. Struct.: THEOCHEM* **2001**, *544*, 237–251.
[189]Huetz, P., Ramseyer, C., Girardet, C. *Chem. Phys. Lett.* **2003**, *380*, 424–434.
[190]Söderhjelm, P., Ryde, U. *J. Comp. Chem.* **2008**, *30*, 750–760.
[191]Baucom, J., Transue, T., Fuentes-Cabrera, M., Krahn, J. M., Darden, T. A., Sagui, C. *J. Chem. Phys.* **2005**, *121*, 6998–7008; Babin, V., Baucom, J., Darden, T. A., Sagui, C. *J. Phys. Chem. B* **2006**, *110*,

Most recently, Plattner and Meuwly have published a series of articles where they systematically consider multipole interactions for carbon monoxide during simulations.[192] In the one already mentioned,[187] they investigate the dynamics of myoglobin, where the inclusion of a fluctuating quadrupole moment leads to a correct description of the B-state (the location of the CO molecule after the Fe-O bond in myoglobin breaks). In another study,[193] the same authors simulate carbon monoxide with fluctuating multipole moments in various ice models and reproduce the experimental splitting of the CO absorption band, related to two different positions of the CO impurity, at interstitial and substitution sites.

A different approach was adopted by Liem and Popelier based on AIM theory, where they engaged up to quadrupole-quadrupole interactions for improving the electrostatic interaction in simulations of water[194] and liquid hydrogen fluoride.[195] Gresh et al. on the other hand have used distributed multipole moments to improve intramolecular interaction in flexible molecules in their SIBFA potential.[196]

Most of these approaches assume that the simulated molecules are rigid, or represent the conformational dependence of atomic multipoles in linear molecules with an approximate function of the bond length. However, in molecules with more than 4-5 atoms the number of vibrations for which the atomic moments would need to parametrized is overwhelming. So it seems that if full-fledged atomic simulations are to be developed that consider multipole interactions between flexible molecules, new solutions will no doubt be necessary. Appendix A provides a short outlook and discussion of a feasible approach.

2.6 Charge redistribution along reaction paths

Some of the most interesting questions to be asked about bond formation and dissociation concern the changes that take place in the electron distributions around atoms. Since a representation in terms of multipole expanded atomic moments can describe the distribution of charge around molecules, it is natural to ask if they can be used to characterize the changes that occur during reactions.

Two cases are studied here: the alkaline hydrolysis of *O,O*-dimethylphosphorofluoridate (DMPF) and the synthesis of carbonic acid. In both cases, the basic question is how well atomic multipole derived potentials describe the molecular electrostatic potential around reactants and how this representation converges with the expansion rank.

The first reaction – the hydrolysis of DMPF – was studied in greater detail, and was based on a recent report by Dyguda-Kazimierowicz et al.,[197] which describes the hydrolytic degradation of several organophosphorous compounds. All of the compounds studied there

11571–11581.

[192]Plattner, N., Meuwly, M. *J. Mol. Model.* **2009**, *15*, 687–694.

[193]Plattner, N., Meuwly, M. *ChemPhysChem* **2008**, *9*, 1271–1277.

[194]Liem, S. Y., Popelier, P. L. A., Leslie, M. *Int. J. Quant. Chem.* **2004**, *99*, 685–694.

[195]Liem, S. Y., Popelier, P. L. A. *J. Chem. Phys.* **2003**, *119*, 4560–4566; Houlding, S., Liem, S. Y., Popelier, P. L. A. *Int. J. Quant. Chem.* **2007**, *107*, 2817–2827.

[196]Gresh, N., Kafafi, S. A., Truchon, J.-F., Salahub, D. R. *J. Comp. Chem.* **2004**, *25*, 823–834.

[197]Dyguda-Kazimierowicz, E., Sokalski, W. A., Leszczyński, J. *J. Phys. Chem. B* **2008**, *112*, 9982–9991.

are notable in that they are known to be substrates for the enzymatic reactions catalyzed by phosphotriesterase, and the authors acknowledge and carefully study the multistage nature of some of those degradation reactions.

The present analysis is limited to the first stage of the "A" path in DMPF degradation, designated by INT1→TS1a→INT2a,[198] where in the transition state (TS1a) hydroxide is aligned with the phosphoryl oxygen atom. The reaction coordinate in Fig. 2.7 and subsequent plots correspond to this path, with zero denoting the transition state.

Besides the evolution of atomic moments, Fig. 2.7 shows four different measures of the *ab initio* molecular electrostatic potential on the Connolly surface, namely its average, median, minimum and maximum values (the average value is approximately constant, in accordance with Gauss's law). The Connolly or solvent-excluded surface[199] was chosen as it represents a typical region in space at which other molecules could interact.[200]

The first observation to be made here is that these measures do not change in a concerted way. In particular, the maximum or weakest potential on the surface does not exhibit any significant change at all. Meanwhile, the minimum or strongest potential has its largest absolute value at reaction coordinates of approximately -10, where the other measures remain constant. The fact that the minimum (most negative) potential changes the most is not surprising, since the reactants are charged negatively thus emphasizing negative potentials, and these are the most representative for monitoring charge redistribution during the reaction.

Already from these crude characteristics of the MEP it is evident that the largest reorganization take place before and after the transition state, the second region being just before a reaction coordinate value of +5. The circumstances of this second region are different, since along with a rise in the minimum value, there is a visible dip in the median. This means that the negative potential spreads out on the surface and becomes smaller on average, which might be caused by a conformational change in the reactants relative to the surface.

This pattern is reflected in the evolution of atomic charges (Fig. 2.7), where the largest changes also take place after the transition state (reaction coordinate zero), and can also be identified in the evolution of a number of atomic moments (Fig. 2.8). The first region (around reaction coordinate -10) involves local charge redistribution within the hydroxyl ion and nearby methyl groups (electron transfer from C6 to H7). After that, redistribution gradually intensifies with little variation in the minimum value, with even the charge on the central phosphorous (P1) changing by $0.1e$ before the transition state.

Much of the charge transfer takes place in the second region, which can be read from the atomic charge evolution in Fig. 2.7. After the transition state, the approaching oxygen atom is 2.6 Å from the phosphorous situated between the methyl groups, with H7-O14-H11 being

[198]Results presented here are based on the geometries obtained by Dyguda-Kazimierowicz et al. ; potentials and multipole moments were recalculated at the Hartree-Fock level using the 6-311++G(d,p) basis set.

[199]Connolly, M. *J. Appl. Crystall.* **1983**, *16*, 548–558; Connolly, M. *Science* **1983**, *221*, 709–713.

[200]The surface was generated using code published by Connolly and scripts by Paweł Kędzierski. The probing distance was set to the van der Waals radii according to Pauling and Bondi (for carbon and fluorine), extended by the radius of the water molecule (1.4Å). Further details on the implementation can be found in P. Kędzierski, Ph.D. Thesis: Study of the nature of interactions in the active sites of enzymes (2001).

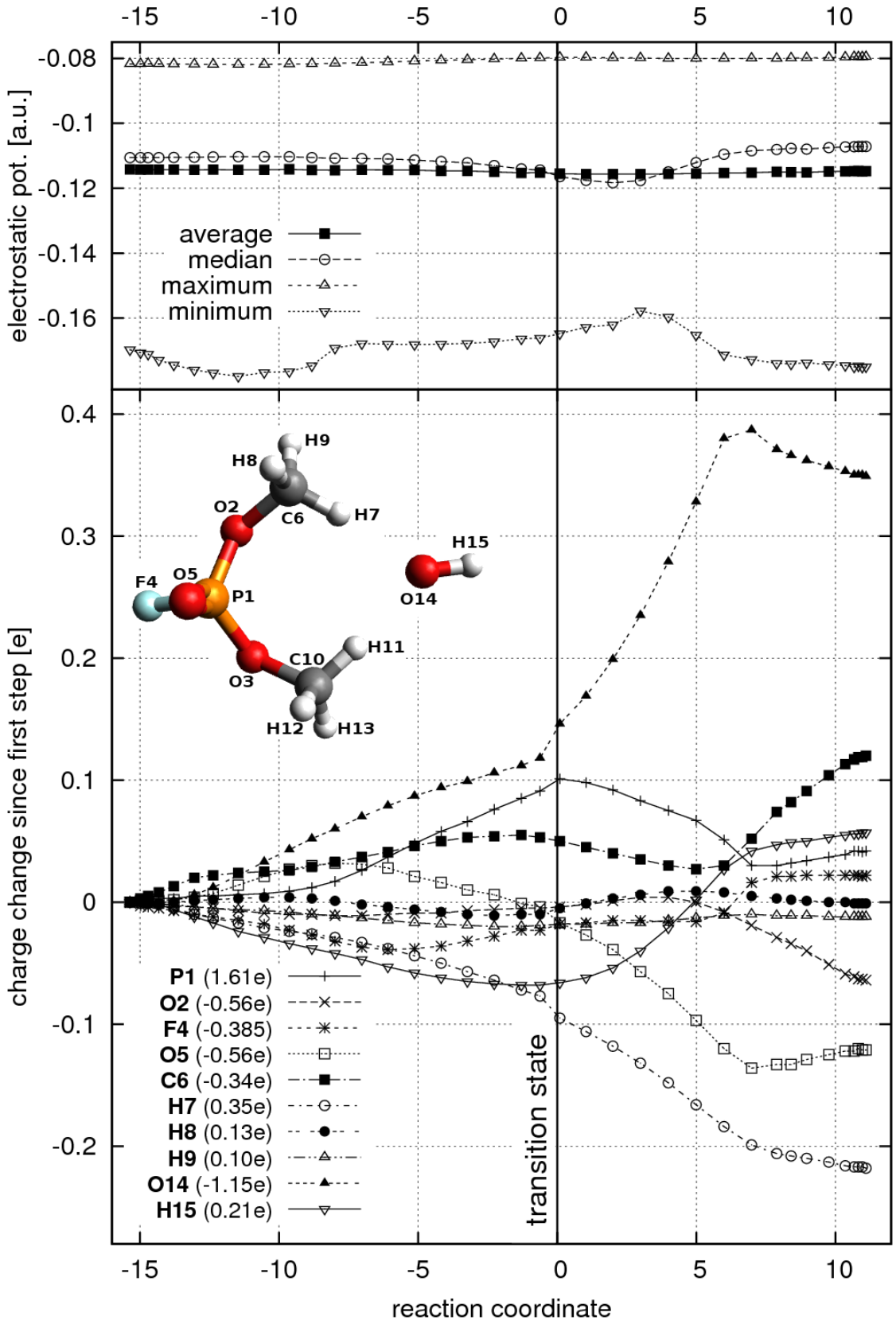


Figure 2.7: Evolution of atomic Mulliken (CAMM) charges and *ab initio* molecular electrostatic potential and on the Connolly surface around reactants during the first stage of the alkaline hydrolysis of DMPF, INT1→TS1a→INT2a. Atomic charges are plotted relative to their value at the first step NT1 (left hand side), with the embedded molecular structure defining the names and numbering of atoms used in the legend. The upper plot shows the corresponding evolution of the average, median and maximum values of molecular electrostatic potential.

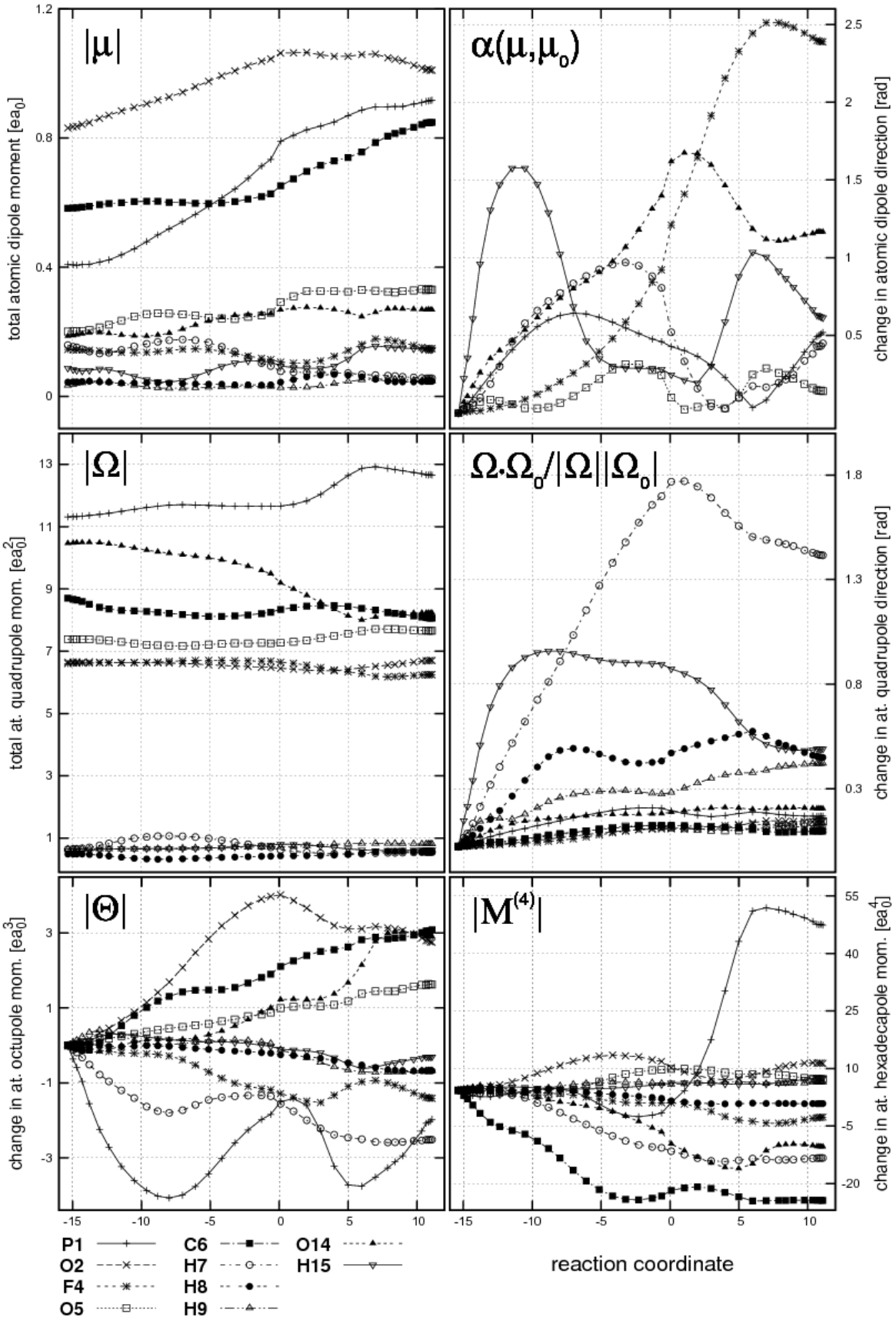


Figure 2.8: Evolution of atomic multipole (CAMM) moments during the first stage of the alkaline hydrolysis of DMPF, INT1→TS1a→INT2a, with the same path and atom definitions as in Fig. 2.9.

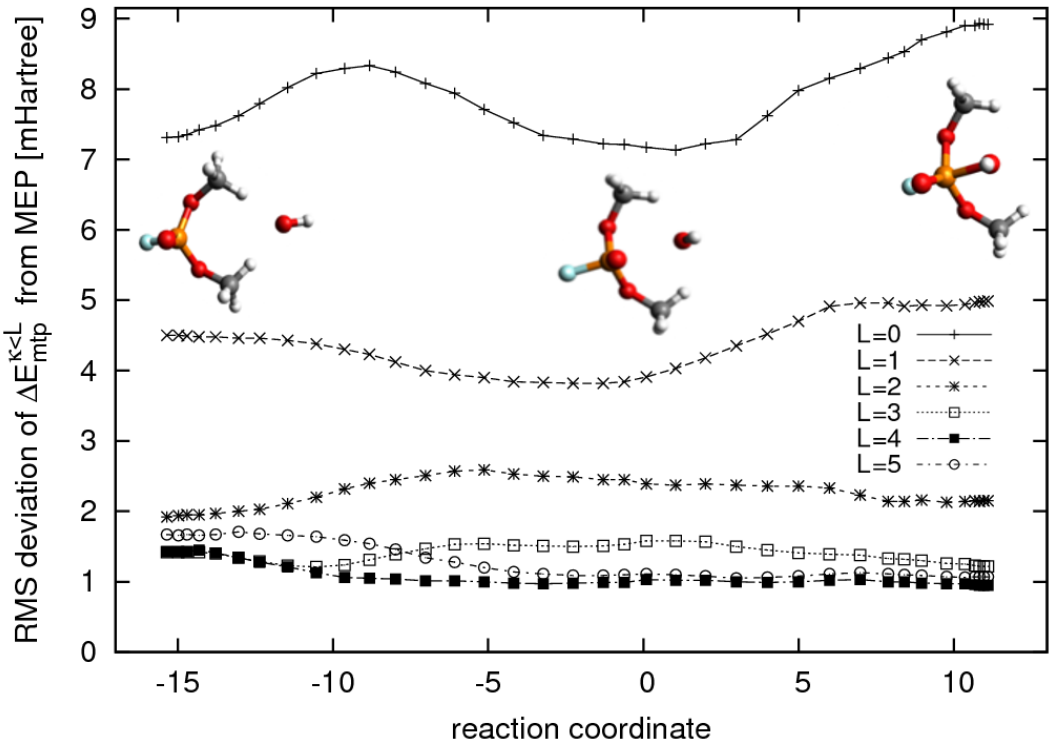


Figure 2.9: Root mean square deviation of the multipole-derived electrostatic potential compared to its *ab initio* value on the Connolly surface around the alkaline hydrolysis reaction of *O,O*-dimethylphosphorofluoridate (DMPF).

roughly linear. The O14 oxygen proceeds to give away almost $0.3e$, and it is not surprising that much of the charge donated by O14 ends up on the other oxygen atoms bonded to the phosphorous atom. However, the second largest change is found for the H7 and H11 atoms (above $0.2e$), a drop that returns their charge to "standard" values, similar to the other hydrogen atoms of the methyl groups.

At first sight, it may seem surprising that much of the charge redistribution takes place after the transition state, nonetheless this agrees with energetic considerations. Since the transition state is essentially a stationary point on the potential energy surface, its derivative there with respect to the reaction coordinates is zero and the energy should not change very much in the vicinity. As the energy is a function of the charge distribution, it follows the latter should also not change significantly.

Mulliken charges and the associated CAMM atomic moments are arbitrary and often strongly basis set dependent,[167] but the plotted *changes* in atomic charges illustrate the magnitude of charge redistribution and should be less sensitive (Löwdin charges were compared in this case). If anything, these changes pinpoint which atoms participate in the reaction and in which direction the flow of charge takes place.

The same is true for atomic moments, some of which are plotted in Fig. 2.8. They describe the finer effects of charge redistribution, also within the bounds of individual atoms. It is hard to draw any final conclusions from them, but some general observations about the role of certain atoms are possible. The largest dipole moments in the system (O2, P1, C6) are all reinforced during the reaction, however their directions do not change significantly. The

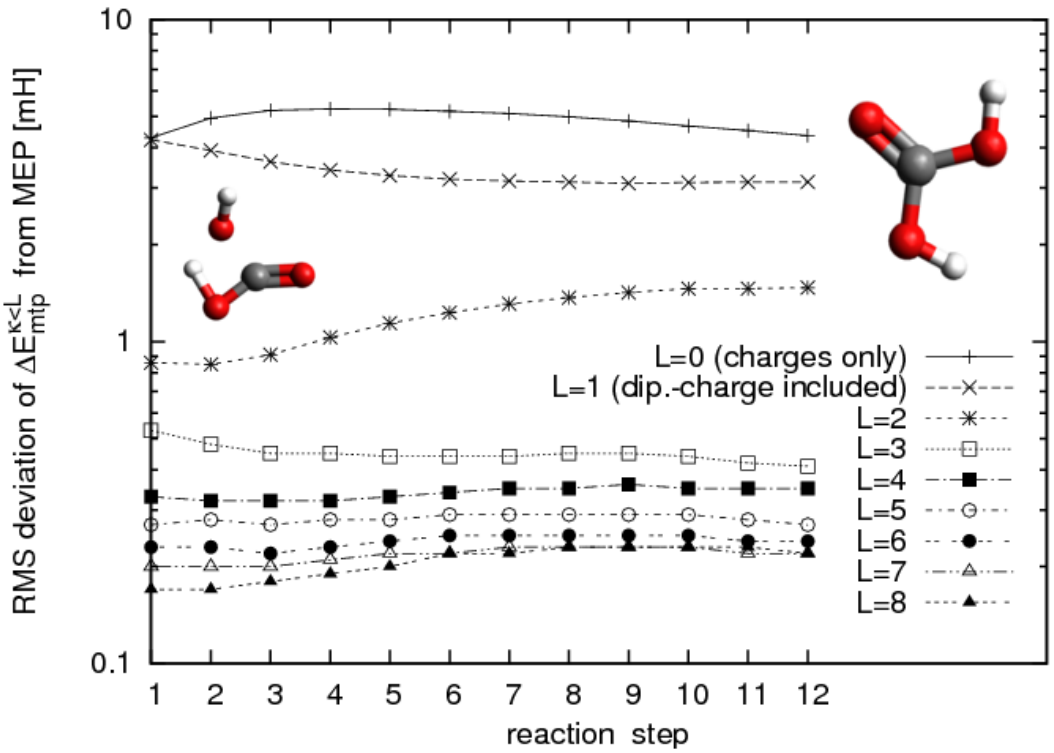


Figure 2.10: Root mean square deviation of the multipole-derived electrostatic potential compared to its *ab initio* value on the Connolly surface along the reaction path of carbon dioxide hydration.

direction of dipoles in the hydroxyl group changes by about a right angle, which is expected when the methyl groups stop interacting and the oxygen atom enters the new bond. Surprisingly, however, the dipole moment on the fluoride atom, although small, changes its direction by almost 180 degrees.

For the higher moments, the central phosphorous atom exhibits by far the largest values and most rapid changes, since the charge distribution around its nucleus is the most anisotropic. Not all of these subtle changes will have an effect on the surroundings, which leads to the question of how the moment-derived molecular potential converges with the increasing rank of the multipole expansion. For example, the hexadecapole (denoted in Fig. 2.8 by $M^{(4)}$) changes drastically after the transition state, while the difference in the multipole interaction on the Connolly surface between ranks L=3 and L=4 (Fig. 2.9) is very small.

Fortunately, the redistribution of molecular charge density can be monitored more directly through the multipole expansion of any well-defined molecular property, such as the MEP or electric fields. In their case, the arbitrary character of a particular atomic charge definition is, to a large degree, eliminated, yielding static or dynamic catalytic fields[122] and aiding *de novo* catalyst design.[201]

The error given in Fig. 2.9 is the root mean square deviation of the MEP estimated from atomic multipole expansions relative to the *ab initio* value on the Connolly surface, using the same reaction and coordinates as in previous figures concerned with DMPF.

[201] Dziekoński; Sokalski; Podolyan; Leszczyński, 2003, in Ref. 124 on page 25, and other references by Dziekoński et al. in Section 2.2.3.

Using only atomic charges (L=0) in this case implies an error of over 7 mH, and adding charge-dipole interactions (L=1) lowers it to 4-5 mH. Moments up to octupoles (L=3) are needed in order to bring the deviation below 2 mH (similar to the octupole-level convergence observed for the transferable atom equivalent method[183]). However, about 1 mH is the lowest root mean square (RMS) that can be achieved on the Connolly surface in this case. The average value on this surface, plotted in Fig. 2.7, is no less that 110 mH, which means that the converged expansion carries an error below 1%.

If the multipole expansion is converged, then this value can be interpreted as an estimate of the average value of penetration effects at this distance according to (2.44). To compare, using atomic charges implies an error of about 6%. It should be mentioned that in this case the multipole expansion starts to diverge for higher ranks, and the RMS deviation starts to increase for $L > 9$.

Fig. 2.10 presents the same RMS deviation of the multipole potential at twelve points along the second phase of the CO_2 reaction with water, namely during the forming of the CO double bond.[202] Using atomic charges only (L=0) in this case implies an error of about 4-5 mH, and adding charge-dipole interactions (L=1) lowers it by only about 1 mH. Moments up to octupoles (L=3) are needed in order to bring the deviation into the sub-millihartree range, and it converges to around 0.2 mH for higher multipole ranks (tested up to L=16).

It is important to emphasize the logarithmic scale, used for clarity, and that the decrease in the error is twenty-fold. Compared to absolute values of the potential, which oscillate around 20 mH for this reaction, this corresponds to improving the deviation from 20% to 1%. Again, using octupoles (L=3) generally brings the MEP estimated from atomic multipole expansions close to the best approximation of the exact potential that is possible – with an estimate of penetration effect of about 1%.

[202]The transition state geometry was obtained here at the RHF/6-31G** level, and potentials and atomic moments were generated using the 6-311++G(d,p) basis set.

2.7 Conclusions

Besides glancing at the current state of research in intermolecular interactions, the present chapter demonstrates the utility of two methods central to this work. The first of these is the hybrid variation-perturbation method (HVPT) described in Section 2.2.3, which partitions the intermolecular interaction energy into components defining a hierarchy of gradually approximate theory levels. HVPT components correspond closely to the values obtained from SAPT, as demonstrated by Tables 2.3 and 2.4 as well as Fig. 2.2. They are also noteworthy for exhibiting the same high degree of basis set stability, although with smaller cost. In particular, compared to other variational schemes of the EDA type, the stability achieved using smaller basis sets (aug-cc-pVDZ was the smallest basis set used in the comparison) is encouraging for applications on large systems.

As a general conclusion, the HVPT scheme is found to be a satisfactory alternative for state-of-the-art, but expensive SAPT calculations. It is adequate for analyzing interaction energies in medium-sized to large complexes, particularly those that are out of reach today for perturbation methods or variational methods that do not employ direct integral evaluation or other technical enhancements. This choice has allowed relatively large systems to be studied in this dissertation, while still retaining the core of physically meaningful interaction energy components.

Section 2.4.1 follows the origin of the Cartesian multipole expansion and the sections that follow discuss its use in various contexts, with several examples of convergence properties for a few select systems. While convergence is not compared to other types of multipole expansions, cases are shown where relatively high multipole moments are necessary in order to obtain a converged approximation to the electrostatic interaction energy. The most relevant example are stacking interactions at distances typical for DNA ($\sim$3.4 Å), which require moments up to rank five or six.

The convergence rate of the multipole expansion was also studied for the electrostatic potential around the reactants of two reaction models, and assessed using its root mean square deviation from the expected value on the Connolly surface. The converged multipole potentials deviated from the expected value in both cases by around 1%, which can be considered an estimate of the average magnitude of penetration effects at such distances.

The first reaction, the alkaline hydrolysis of DMPF, was studied in greater detail. An examination of the median and largest electrostatic potentials on the Connolly surface and of atomic charge changes show that most of the charge redistribution takes place before and after the reaction's transition state (between the reaction coordinates -10 and +5), which is expected energetically. The higher multipole moments on atoms complementing charges give a finer description of the charge redistribution. Magnitudes as well as directions of higher moments may be important in certain cases, especially those of dipoles. These results clearly demonstrate the significant variability of atomic charges along the reaction path as well as the non-trivial role of higher atomic multipole moments during chemical reactions.

3 Statistical relationships between interaction energy terms

> [..] the interaction between a drug and receptor will perturb both molecules. It is our basic hypothesis that at the most remote distance of drug—receptor engagement it is the preferred conformation of the drug or a conformation close in energy which is recognized by the receptor.
>
> Lemont B. Kier
> *The Prediction of Molecular Conformation as a Biologically Significant Property*[203]

3.1 Introduction

Kier called the quoted supposition the *hypothesis of remote recognition of preferred conformation*, which he applied in constructing structure-function relationships for small molecules based on early computational methods such as extended Hückel theory (EHT) or complete neglect of differential overlap (CNDO).[203] Neurotransmitters were of particular interest, and according to the theory a signal molecule makes a "preliminary, weak bond with its receptor while in a minimum energy state".[204], implying that this long-range interaction is favorable and somehow corresponds to the strength of the final bond and to the molecule's activity.

An example of the kind of observations that support this hypothesis were the parallel dual function and two conformations found for histamine. In particular, Kier suggested that two classes of histamine receptors interact selectively with the two distinguished conformations.[205] Since then, however, experiments and calculations have shown histamine to adopt other conformations, especially in solution[206] and its mechanism of action is known to be more complicated, with at least four distinct receptors.

Further studies showed that the remote recognition hypothesis is too simplistic for certain systems and fails for example when interacting lone electron pairs are involved.[207] Nonetheless, the idea and similar ones have permeated molecular studies not only in theoretical pharmacology,appearing in different forms throughout the years. For example, the concept of near attack conformers (NAC) advocated by Bruice[208] is essentially a paraphrase in the context of enzymatic catalysis. A NAC is supposed to be structurally similar to the transition state, while

[203]Kier, L. B. *Pure Appl. Chem.* **1973**, *35*, 509–520.
[204]Kier, L. B., Höltje, H.-D. *J. Theor. Biol.* **1975**, *49*, 401–416.
[205]Kier, L. B. *J. Med. Chem.* **1968**, *11*, 441–445.
[206]Ramirez, F. J., Tunon, I., Collado, J. A., Silla, E. *J. Am. Chem. Soc.* **2003**, *125*, 2328–2340.
[207]Hall, L. H., Kier, L. B. *J. Theor. Biol.* **1976**, *58*, 177–195.
[208]Bruice, T. C. *Acc. Chem. Res.* **2002**, *35*, 139–148.

being in thermal equilibrium with the substrates. The role of the catalyst in this case is understood to also include stabilization of the NAC, thus reorganizing substrates to conformations that are closer to the transition state.

It is the purpose of this work and of general interest to explore such notions with regard to the interaction energy and the physical effects it arises from. Long-range interactions between drug and receptor, as well as between substrate components in their near attack conformations, these must be related to the corresponding interaction energy at equilibrium and in the transition state, respectively. Additionally, long-range interactions will be dominated by electrostatic effects, pointing to a special role of the latter in molecular recognition. In the present chapter, this aspect is approached using statistical analysis, providing measures of the relationship between interaction energies at equilibrium and at large distances.

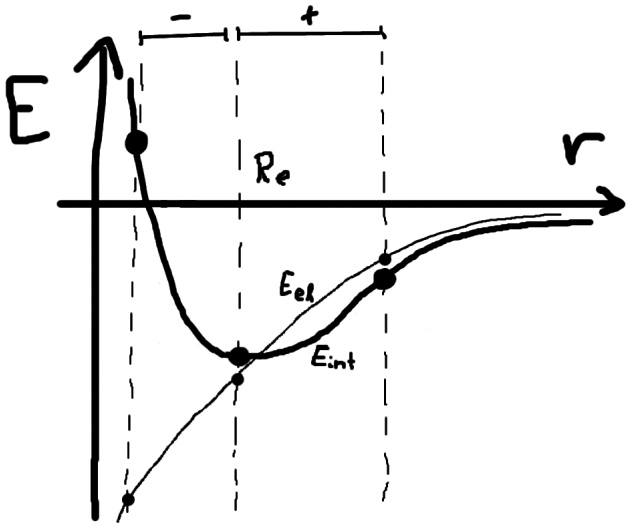


Figure 3.1: Conceptual plot of the total interaction energy E_{int} and its electrostatic component $\Delta E_{\text{el}}^{(1)}$ at misguided intermolecular distances $\mathbf{r}$.

Another motivation for this study, which touches on the subject of electrostatic interactions, lies on the opposite side of the scale, namely at shortened intermolecular distances. It is well known that force fields can generate misguided geometries, with an RMS error that varies from anywhere between 0.2 Å to more than 3 Å, depending on the force field used and type of system tested.[209] A recent study by Grzywa et al. illustrates that an error in the intermolecular distance predicted by a force field can strongly influence a subsequent, theoretically more rigorous analysis.[210] In their case, the nearest contacts between inhibitors and active site residues were shorter than in the more accurate MP2 geometries by up to 1 Å. Due to this, the MP2 interaction energy calculated for geometries obtained using the force field failed to correlate with experimental activities. Surprisingly, its electrostatic component exhibited good correlation. Although the sample size was too small to make objective conclusions, this observation suggests that electrostatic interaction may provide a better prognostic of the actual stability in the equilibrium geometry than the total energy.

Therefore, the present analysis also encompasses interactions at shortened intermolecular distances. The general aim is to test what relationships can be found between interaction energies at various intermolecular distances, analyzed using the HVPT method, and the equilibrium interaction strength. In doing so, the electrostatic component is the major target, for the reasons given above and because it is the least expensive and dominating term at large distances. The final question sought to be answered is: *to what extent can electrostatic*

[209] Paton, R. S., Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 944–955.

[210] Grzywa, R., Dyguda-Kazimierowicz, E., Sieńczyk, M., Feliks, M., Sokalski, W. A., Oleksyszyn, J. *J. Mol. Model.* **2007**, *13*, 677–683.

effects at various distances be used to predict relative equilibrium stability? Before presenting results, however, a short introduction and rationale is provided for the non-parametric statistical measures used.

3.1.1 Rank-based statistics for interaction energies

By far the the most popular way to capture statistical dependence is the Pearson correlation coefficient, which tests for a linear relationship between two samples drawn from (usually normal) probability distributions. When considering any two molecular properties, however, there is no reason to expect a linear relationship. Since the choice of molecules is always biased and usually rooted in a preference for some specific type of connectivity, the properties also cannot be said to depend on a random variable. Furthermore, the Pearson coefficient is also not a robust measure as outliers in the data, which can be frequent when molecules are chosen arbitrarily, can have a strong influence on the correlation coefficient.

In such cases it is natural to turn to non-parametric statistics, which are distribution-free and more robust since they exclude the numerically extreme character of outliers. The interpretation of rank-based statistics is especially straightforward if the tested hypothesis is suitably formulated. For example, the minimum practical information needed when screening a large set of compounds is whether the activity of one compound is larger than another. If this can be decided using energetic criteria, then the main concern becomes how to efficiently reproduce the ascending or descending order of these energies. Their exact values then are not most important, and one may turn towards the analysis of their ranks.

Let two sets of energies with the same number of elements N, $\{A_i\}$ and $\{B_i\}$, be well-ordered. The index i has the same meaning for both A and B, symbolizing a specific molecule or intermolecular complex. The elements of these sets, energies, are treated as raw scores, and can be readily converted into ranks. The rank a_i of an element A_i corresponds to its position when the set is sorted in an ascending or descending order. If descending order is assumed, this is equivalent to the number of elements that are greater or equal to A_i:

$$a_i = |\{j : A_j \geq A_i\}| \,. \tag{3.1}$$

If two elements happen to have the same numerical values, then their rank is taken as the average of what their ranks would otherwise be. To illustrate this case with a simple example, suppose that $A = \{1, 2, 2, 3\}$. The ranks of the middle elements will then be $a_2 = a_3 = 2.5$.

One of the non-parametric measures used in the present study is the Spearman rank correlation coefficient, denoted by ρ_S and often defined as the Pearson correlation coefficient between the two sets of ranks a and b:[211]

$$\rho_S = \frac{\sum_i (a_i - \overline{a}) \left(b_i - \overline{b}\right)}{\sqrt{\sum_i (a_i - \overline{a})^2 \sum_i \left(b_i - \overline{b}\right)^2}}. \tag{3.2}$$

[211] Lehmann, E. L., D'Abrera, H. J. M. *Nonparametrics: Statistical Methods Based on Ranks*, rev.; Prentice-Hall, 1998.

It follows that ρ_S, just as the Pearson correlation coefficient, adopts values between -1 and 1, and is positive when B_i tends to increase as A_i increases. A value of one indicates that there is a perfect monotonic relationship. If there are no ties between elements in the sets, in other no two elements have the same value, then a simpler procedure can be used to calculate the Spearman rank coefficient:

$$\rho_S = 1 - 6 \sum_i \frac{(a_i - b_i)^2}{N(N^2 - 1)}, \tag{3.3}$$

Another way to express the strength of a monotonic relationship between A and B in a non-parametric way is to count the number of concordant and discordant pairs among the sets, N_C and N_D respectively. These two numbers are the cardinalities of sets that contain aligned or misaligned pairs. For example, the number of discordant pairs N_D is equal to $|A_D|$, where

$$A_D = \{(i,j) : i < j, (A_i > B_i \wedge A_j < B_j) \vee (A_i < B_i \wedge A_j > B_j)\}, \tag{3.4}$$

with an analogous definition for $N_\mathrm{C} = |A_C|$.

The numbers N_C and N_D already give a good idea of the monotonic relationship, especially its success rate when expressed as fractions or percentages. Here, however, they will be used as the basis for the Kendall tau coefficient, a non-parametric statistic that measures the correspondence of rankings,

$$\tau_\mathrm{K} = \frac{N_\mathrm{C} - N_\mathrm{D}}{\frac{1}{2}N(N-1)}. \tag{3.5}$$

Considering that the sets of energies dealt with here contain real numbers, this rank correlation coefficient can be equivalently written as

$$\tau_\mathrm{K} = \frac{1}{\frac{1}{2}N(N-1)} \sum_{\substack{i}}^{N_i} \sum_{\substack{j \\ i>j}}^{N_j} \mathrm{sgn}\left((A_i - B_i)\,(B_j - A_j)\right). \tag{3.6}$$

Similar to the Pearson and Spearman coefficients, τ_K assumes values between -1 and 1, extremes that correspond to opposite ideal monotonic relationships. A statistical significance can be assigned to all these correlation coefficients, interpreted as the probability of the observed rank sets a and b assuming the null hypothesis is true. Since a is sought to be used to reproduce b or *vice versa*, the null hypothesis is that there is no monotonic correspondence between or that the order in A and B is random.

As mentioned above, a simple and practical measure of how well two sets of energies are aligned is the number of concordant or discordant pairs. Likewise, the number of errors or misaligned pairs can be counted for either the monotonic or anti-monotonic cases if N_D is taken when $\tau_\mathrm{K} > 0$ and N_C when $\tau_\mathrm{K} < 0$. This will be equal to the minimum of the two numbers, in general. Divided by the total number of pairs, this provides a fractional measure of the amount of mistakes made when reproducing the order of elements in set B from the

order in A, N_{mis}, which can be expressed as a percentage,

$$N_{\text{mis}} = 100\% \cdot \frac{\min(N_D, N_C)}{\frac{1}{2}N(N-1)}. \tag{3.7}$$

In order to estimate the magnitude of the mistakes made, the average difference between the values of mistaken pairs can be calculated. There will be two such averages, one for A_i and another for B_i. The first of these, $\bar{\Delta}^{\text{A}}_{\text{mis}}$, will be

$$\bar{\Delta}^{\text{A}}_{\text{mis}} = \frac{1}{\frac{1}{2}N_{\text{mis}}(N_{\text{mis}}-1)} \sum_{(i,j)\in M}^{N_{\text{mis}}} |A_i - A_j|, \tag{3.8}$$

where M is the set of discordant or concordant pairs, depending on the direction of the monotonic relationship, or whether τ_{K} is positive or negative:

$$M = \begin{cases} A_D & : N_C > N_D \\ A_C & : N_C < N_D \end{cases}. \tag{3.9}$$

The magnitude $\bar{\Delta}^{\text{A}}_{\text{mis}}$, in the context of energies as discussed here, is a measure with clear practical meaning. It is the average difference between two energies in A when their order is opposite than the majority of concordant or discordant pairs with respect to B. The same delta can be defined for B relative to A – $\bar{\Delta}^{\text{B}}_{\text{mis}}$ – and the particular choice of symbols depends on which set is used to predict which.

3.2 Small dimers from the S22 training set

State-of-the-art interaction energies at the CCSD(T) level of theory extrapolated to the complete basis set (CBS) limit were published in 2006 by Hobza and coworkers for a series of van der Waals dimers.[212] This was intended to be a point of reference for future benchmarks and analysis, and a smaller S22 training set was meant to be a testbed for developing new approximate methods. The dimers chosen for the S22 set are listed in Fig. 3.1, along with their symmetries, reference energies and symbols used herein for convenience.

A more recent study by Fusti Molnar et al. is also based on the S22 training set and extends the interaction energy calculations to a range of intermolecular separations, generated by varying the distances between monomer centers of mass.[213] The same fourteen deviations from equilibrium were used for each dimer, namely -0.8, -0.4, -0.2, -0.1, 0.1, 0.2, 0.4, 0.7, 1.0, 1.5, 2.0, 3.0, 5.0 and 10.0 Å.

This case study follows in the footsteps of Fusti Molnar et al. and make use of the geometries published as supplementary material to their article.[213] Mentions of the reference or original CCSD(T) interaction energies, however, will everywhere refer to the first values

[212] Jurečka, P., Šponer, J., Černý, J., Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
[213] Fusti Molnar, L., He, X., Wang, B., Merz, K. M. J. *J. Chem. Phys.* **2009**, *131*, 065102.

published by Jurecka et al.,[212] and denoted by $\Delta E^{\mathrm{CBS}}_{\mathrm{CCSD(T)}}$. Wherever used, the equilibrium or out-of-equilibrium CCSD(T) interaction energies published by Fusti Molnar et al. will be marked additionally with a prime – $\Delta E^{\mathrm{CBS'}}_{\mathrm{CCSD(T)}}$.

Presently, the goal is to study the components of the interaction energy using the HVPT method described in Section 2.2.3, in particular by the hierarchy in (2.23). Curves of the interaction energy at the MP2 and other levels of theory for the geometries borrowed from Fusti Molnar et al. are presented collectively for all of the 22 training dimers in Fig. 3.2. The accompanying plots in Fig. 3.3 correspond to interaction energies at various levels of theory.[214]

		dimer (geometry)	[kcal/mol]
hydrogen-bonded	A	ammonia dimer (C_{2h})	-3.17
	B	water dimer (C_s)	-5.02
	C	formic acid dimer (C_{2h})	-18.61
	D	formamide dimer (C_{2h})	-15.96
	E	uracil dimer (C_{2h})	-20.65
	F	2-pyridoxine·2-aminopyridine (C_1)	-16.71
	G	adenine·thymine (C_1)	-16.37
dispersion-dominated	H	methane dimer (D_{3d})	-0.53
	I	ethene dimer (D_{2d})	-1.51
	J	benzene·methane (C_3)	-1.50
	K	benzene dimer (C_{2h}, stacked)	-2.73
	L	pyrazine dimer (C_s, stacked)	-4.42
	M	uracil dimer (C_2, stacked)	-10.12
	N	indole·benzene (C_1, stacked)	-5.22
	O	adenine·thymine (C_1, stacked)	-12.23
mixed complexes	P	ethene·ethine (C_{2v})	-1.53
	Q	benzene·water (C_s)	-3.28
	R	benzene·ammonia (C_s)	-2.35
	S	benzene·HCN (C_s)	-4.46
	T	benzene dimer (C_{2v}, T-shape)	-2.74
	U	indole·benzene (C_1, T-shape)	-5.73
	V	phenol dimer (C_1, T-shape)	-7.05

Table 3.1: Overview of dimers in the S22 training set published by Jurečka et al..[212] Reference extrapolated CCSD(T) interaction energies and symmetries imposed on the dimers are given, along with conformation types in a few cases.

The first step in any statistical correlation study is often to create a scatter plot that gives an overview of the data. Fig. 3.4 presents such a scatter plot, where the MP2 interaction energy (ΔE_{MP2}) and first order electrostatic component ($\Delta E^{(1)}_{\mathrm{el}}$) are shown against the reference CCSD(T) energy ($\Delta E^{\mathrm{CBS}}_{\mathrm{CCSD(T)}}$), all at the original CCSD(T) equilibrium geometries. The data points are quite well distributed over the whole range of interaction energies (from -20 kcal/mol to almost zero), although a larger part is concentrated in the region above -5 kcal/mol.

Combining the practical issues that were outlined at the start of this chapter with the statistical considerations introduced in Section 3.1.1, a number of specific questions can be raised and addressed by analyzing this collection of benchmark results. Above all, it is interesting whether the statistics introduced, such as τ_{K} and $\bar{\Delta}_{\mathrm{mis}}$, support a remote recognition hypothesis akin to the one proposed by Kier. If the answer is affirmative, guidelines could be given to facilitate computational screening techniques. Another important issue is whether and to what extent electrostatic effects at distances closer than the equilibrium prognose the equilibrium stability of a complex.

Table 3.2 follows the rank of the electrostatic component for all dimers, by showing how the order of dimers changes with distance; the alphabetical symbols used are defined in Table 3.1. It also color-codes the number of misaligned pairs (N_{mis}) with respect to the extrapolated equilibrium CCSD(T) interaction energy, using the intensity of red in the background of each cell to represent the percentage of misaligned pairs that a particular dimer is involved in.

[214] All results presented in this dissertation were based on interaction energies calculated with the aug-cc-pVDZ basis set. The same calculations were als operformed for most of the dimers with larger basis sets, yielding similar conclusions.

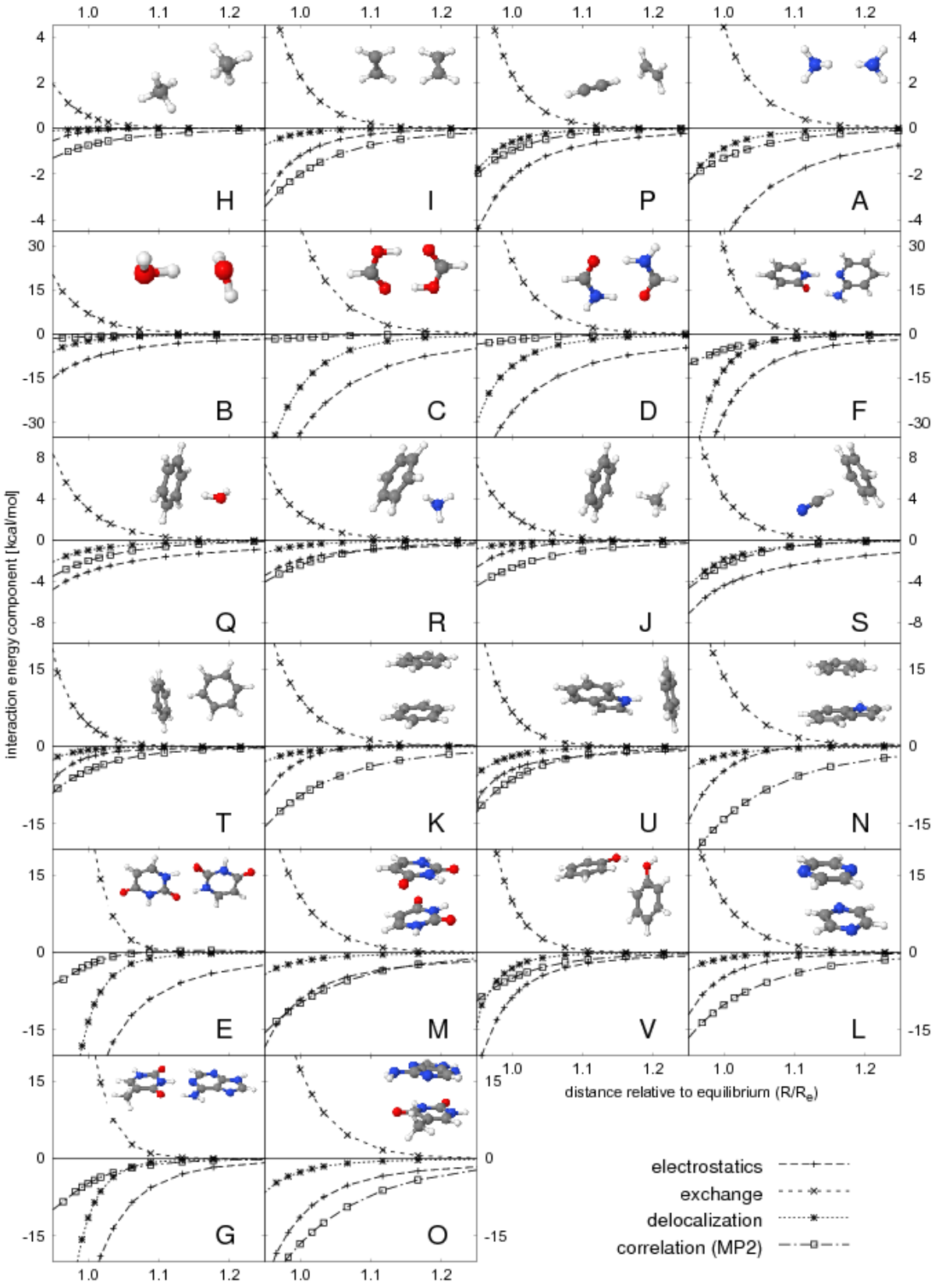


Figure 3.2: HVPT interaction energy components for all molecular dimers in the S22 training set. The letter used to label each dimer corresponds to the list in Fig. 3.1.

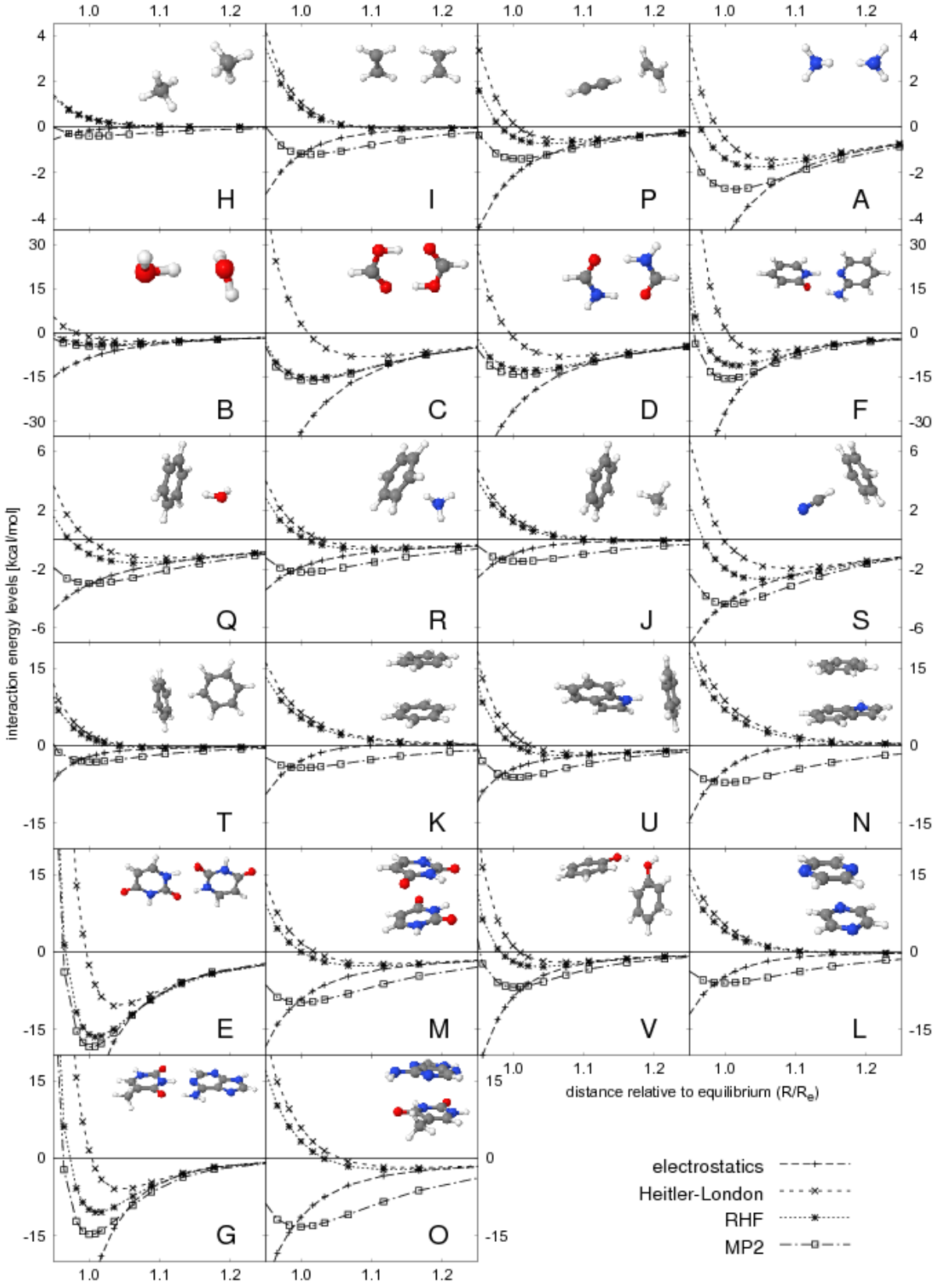


Figure 3.3: The interaction energy at various levels of theory for all molecular dimers in the S22 training set. The letter used to label each dimer corresponds to the list in Fig. 3.1.

It is helpful to clarify the method of presentation with a simple example. Consider the stacked uracil dimer (which is denoted by "M") and the stacked adenine-thymine dimer (O). The ranks for their electrostatic interactions within the S22 set, equivalent to their vertical positions in the table, do not change up to 0.7 Å away from the equilibrium – and the adenine-thymine dimer (O) always shows a stronger interaction energy than dimer M. Therefore, their stability at the equilibrium distance relative to any other S22 dimer can be predicted based on $\Delta E_{\mathrm{el}}^{(1)}$ in this regime. However, when the centers of mass in these dimers are separated by more than 0.7 Å beyond the equilibrium distance, the electrostatic term for the uracil dimer becomes larger. At this point the two dimers switch ranks and rows in Table 3.2, which means that their relative equilibrium stability would not be predicted correctly anymore using $\Delta E_{\mathrm{el}}^{(1)}$.

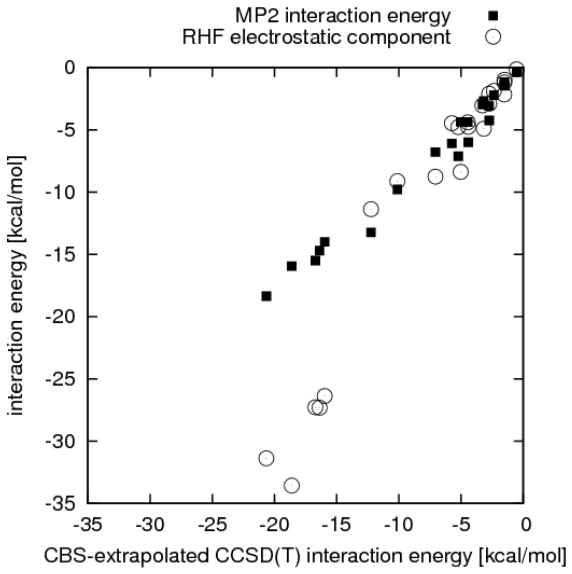


Figure 3.4: Scatter plot of the MP2 and Hartree-Fock electrostatic interaction component against the reference extrapolated CCSD(T) interaction energies.

$\Delta E_{\mathrm{CCSD(T)}}^{\mathrm{CBS}}$ | $\Delta E_{\mathrm{el}}^{(1)}$

	-0.8	-0.4	-0.2	-0.1	0.0	0.1	0.2	0.4	0.7	1.0	1.5	2.0	3.0	5.0	10.0
E	E	C	C	C	C	C	C	E	E	E	E	E	E	E	E
C	G	E	E	E	E	E	E	C	C	C	D	D	D	D	D
F	F	G	G	G	G	F	F	D	D	D	C	C	C	F	F
G	D	F	F	F	F	G	D	F	F	F	F	F	F	C	O
D	C	D	D	D	D	D	G	G	G	G	G	G	M	O	C
O	O	O	O	O	O	O	O	O	O	M	M	M	O	M	M
M	M	M	M	M	M	M	M	M	M	O	O	O	S	S	S
V	N	V	V	V	V	V	V	V	B	B	S	S	G	U	B
U	V	B	B	B	B	B	B	B	V	V	B	U	U	B	U
N	B	N	N	N	A	A	S	S	S	S	U	B	B	V	V
B	L	L	L	L	N	S	A	U	U	U	V	V	V	Q	Q
S	K	A	A	A	L	U	U	A	A	Q	Q	Q	Q	A	A
L	A	K	U	U	U	L	L	Q	Q	A	A	A	A	G	R
Q	U	U	K	S	S	N	Q	L	L	R	R	R	R	R	T
A	S	S	S	K	Q	Q	N	P	P	P	P	P	P	T	P
T	T	T	Q	Q	K	K	P	R	R	L	T	T	T	P	L
K	Q	Q	T	T	P	P	R	T	T	T	L	L	L	L	J
R	P	P	P	P	T	T	T	N	I	I	J	J	J	J	H
P	I	R	R	R	R	R	K	I	J	J	I	I	I	I	I
I	R	I	I	I	I	I	I	K	H	H	H	H	H	H	N
J	J	J	J	J	J	J	J	J	N	K	K	K	K	K	K
H	H	H	H	H	H	H	H	H	K	N	N	N	N	N	G

stronger interaction energies

distance from equilibrium separations [Å]

Table 3.2: Evolution of N_{mis} for the equilibrium CCSD(T) interaction energy and electrostatic component at various distances, for all dimers in the S22 training set. Each cell corresponds to a single dimer from the S22 set (letters the same as in Fig. 3.1), and the intensity of the background red represents the number of misaligned pairs that contain that dimer. Each column is sorted by descending interaction energy from top to bottom. The leftmost, detached column represent the equilibrium CCSD(T) interaction energy extrapolated to the basis set limit. All other columns represent the uncorrelated electrostatic component, at distances marked on the the axis.

Moderate amounts of misalignment are present at all distances. Even for $d_{\mathrm{COM}} = 0$, at which the ammonia dimer (A) is misaligned in over 20% of the possible pairs. There is also a region of exceptionally large misalignment, in the lower right hand corner of the table, where dimers exhibiting the weakest long-range electrostatic interactions are located. In particular, several of the larger stacked complexes incur many misalignments, namely benzene·benzene (K) and indole·benzene (N), as well as the adenine-thymine hydrogen-bonded dimer (G). In the worst case, if one were to base predictions of relative equilibrium stability solely on the electrostatic component in the stacked indole-benzene dimer at distances above ~ 0.5 Å, the choices would be as good as random since the success rate would be roughly 50%.

Table 3.3 in turn shows the summary statistics for such evaluations, testing the electrostatic term as well as other interaction energy components in the role of the prognostic, at the equilibrium distance and for displacements of -0.4 Å and 0.7 Å. Three of these statistics – τ_{K}, N_{mis} and $\bar{\Delta}^{\mathrm{ref}}_{\mathrm{mis}}$– are also plotted as functions of d_{COM} in Fig. 3.5, but for corresponding levels of theory instead of components. It should be stressed that in both cases the distance d_{COM} is given relative to the equilibrium distance between monomers, so that $d_{\mathrm{COM}} = 0.1$ Å refers to the set of all S22 dimers where the distance between centers of mass was extended by 0.1 Å.

The main target of this work, namely the electrostatic component, performs surprisingly well. Especially at shortened distances, none of the three statistics are noticeably worse compared to the correlation coefficient at equilibrium. The Kendall tau in this range is around 0.86-0.87, while the Spearman and Pearson coefficients are well above 0.95. What is the practical value of these correlations? Measures associated with misaligned pairs provide more insight: N_{mis} is below 0.08 in this range, which means that only 8% of all predictions would be false positives. What is more important, the average difference between CCSD(T) energies for these mistakes ($\bar{\Delta}^{\mathrm{ref}}_{\mathrm{mis}}$) is around 1 kcal/mol. This is logical when compared to the scatter plot in Fig. 3.4, in which a few tight groups about 1kcal/mol in diameter can be seen. It is pairs of dimers inside these groups that usually cause the misalignments.

At long range, the statistics for the electrostatic component become gradually worse, but are still significant. Interestingly, the Pearson coefficient remains almost unchanged (with a drop of about 0.01), while both N_{mis} and $\bar{\Delta}^{\mathrm{ref}}_{\mathrm{mis}}$ almost double. This is a typical example where the assumption of a linear relationship can lead to an erroneous conclusion about the quality of a statistical relationship – in this case the quality is clearly overestimated.

It is interesting to compare statistics between interaction components – the exchange ($\Delta E^{(1)}_{\mathrm{ex}}$) and delocalization ($\Delta E^{(\mathrm{R})}_{\mathrm{del}}$) terms both exhibit high correlation coefficients, and their misalignment rates are as good as or better than $\Delta E^{(1)}_{\mathrm{el}}$. On the other hand, the difference in the reference interaction energy that they misjudge on average is always larger than for the electrostatic component. This single difference in the trends of $\bar{\Delta}^{\mathrm{ref}}_{\mathrm{mis}}$ with d_{COM} is probably influenced by the fact that $\Delta E^{(1)}_{\mathrm{el}}$ fades slower other interaction energy terms and dominates at larger values of d_{COM} (above 1.5Å). While the ranks of $\Delta E^{(1)}_{\mathrm{ex}}$ and $\Delta E^{(\mathrm{R})}_{\mathrm{del}}$ may remain similar to that of the equilibrium stability, their ranges will be much narrower, leading to more serious energetic mistakes in the predicted order.

The behavior outlined in the previous paragraph seems logical, since all these terms are

	τ_K/p —	N_{mis} —	$\bar{\Delta}^{ref}_{mis}$ [kcal/mol]	$\bar{\Delta}_{mis}$ [kcal/mol]	ρ_S/p —	ρ_P/p —
$\Delta E_{MP2} \cdot \Delta E^{(1)}_{el}$						
d_{COM} = -0.4	0.844/4e-08	8%	1.13	4.03	0.950/1e-11	0.952/9e-12
d_{COM} = 0.0	0.818/1e-07	9%	1.39	1.54	0.931/3e-10	0.937/1e-10
d_{COM} = 0.7	0.662/2e-05	17%	2.32	0.95	0.798/9e-06	0.911/4e-09
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(1)}_{el}$						
d_{COM} = -0.4	0.861/2e-08	7%	0.99	4.39	0.954/7e-12	0.976/1e-14
d_{COM} = 0.0	0.870/1e-08	6%	0.96	1.00	0.963/8e-13	0.972/4e-14
d_{COM} = 0.7	0.766/6e-07	12%	1.60	0.76	0.879/7e-08	0.960/1e-12
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(1)}_{ex}$						
d_{COM} = -0.4	-0.827/7e-08	9%	1.00	9.30	-0.948/2e-11	-0.965/4e-13
d_{COM} = 0.0	-0.801/2e-07	10%	1.50	2.77	-0.928/5e-10	-0.960/2e-12
d_{COM} = 0.7	-0.758/8e-07	12%	2.25	0.43	-0.895/2e-08	-0.920/1e-09
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(R)}_{del}$						
d_{COM} = -0.4	0.870/1e-08	6%	1.75	2.80	0.970/1e-13	0.917/2e-09
d_{COM} = 0.0	0.879/1e-08	6%	2.07	0.63	0.968/2e-13	0.928/5e-10
d_{COM} = 0.7	0.835/5e-08	8%	2.34	0.10	0.950/1e-11	0.946/3e-11
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(2)}_{disp}$						
d_{COM} = -0.4	0.221/2e-01	39%	7.53	7.76	0.316/2e-01	0.091/7e-01
d_{COM} = 0.0	0.186/2e-01	41%	7.60	4.28	0.269/2e-01	0.085/7e-01
d_{COM} = 0.7	0.048/8e-01	48%	8.20	1.64	0.056/8e-01	0.011/1e+00
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(1)}_{HL}$						
d_{COM} = -0.4	-0.784/3e-07	11%	1.28	6.29	-0.914/3e-09	-0.932/3e-10
d_{COM} = 0.0	-0.022/9e-01	49%	7.19	3.08	-0.074/7e-01	0.102/7e-01
d_{COM} = 0.7	0.654/2e-05	83%	7.99	3.85	0.804/7e-06	0.916/2e-09
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E_{RHF}$						
d_{COM} = -0.4	-0.143/4e-01	57%	6.72	7.36	-0.233/3e-01	-0.072/8e-01
d_{COM} = 0.0	0.455/3e-03	27%	3.65	3.48	0.599/3e-03	0.819/3e-06
d_{COM} = 0.7	0.671/1e-05	16%	2.40	1.16	0.806/6e-06	0.931/3e-10
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E_{MP2}$						
d_{COM} = -0.4	0.411/7e-03	29%	7.43	3.03	0.490/2e-02	0.111/6e-01
d_{COM} = 0.0	0.896/5e-09	5%	0.46	0.79	0.979/3e-15	0.988/1e-17
d_{COM} = 0.7	0.905/4e-09	5%	0.33	0.43	0.982/6e-16	0.990/2e-18
$\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{CBS'}_{CCSD(T)}$						
d_{COM} = -0.4	0.484/3e-03	26%	5.71	1.24	0.544/1e-02	0.709/5e-04
d_{COM} = 0.0	0.989/1e-09	1%	0.20	0.05	0.998/4e-24	1.000/1e-30
d_{COM} = 0.7	0.937/8e-09	3%	0.25	0.07	0.988/5e-16	0.998/2e-22

Table 3.3: Kendall (τ_K), Spearman (ρ_s) and Pearson (ρ_p) correlation coefficients within the S22 training set between a reference equilibrium interaction energy and an interaction component at representative distances. The value of d_{COM} is always relative to the equilibrium separation of monomer centers of mass, and the p in τ_K/p and other coefficients denotes the p-value or statistical significance of the correlation coefficient. The titles on the left of each three-row section define the reference equilibrium energy and the interaction component used as the predictor. For example, $\Delta E^{CBS}_{CCSD(T)} \cdot \Delta E^{(1)}_{el}$ means that the first order electrostatic component $\Delta E^{(1)}_{el}$ was correlated with the reference equilibrium CCSD(T) interaction energy; this particular case is also illustrated in detail in Table 3.2. The superscript "ref" in $\bar{\Delta}^{ref}_{mis}$ means that the average difference was calculated for the reference interaction energy ($\Delta E^{CBS}_{CCSD(T)}$ or ΔE_{MP2}), and $\bar{\Delta}_{mis}$ without a superscript refers to the average difference for the electrostatic or other prognostic component.

actually *part of* the total interaction energy and therefore should hint at the relative equilibrium stability of a complex to some extent. For this reason, it is important to quantify *how well* particular terms perform in this regard, and evaluating $\bar{\Delta}^{\text{ref}}_{\text{mis}}$ is an example of one simple way to do that. The dispersion component $\Delta E^{(2)}_{\text{disp}}$ seems to be an exception to this, as in its case all correlation coefficients at all separations are insignificant, with mistakes in predicting relative stability being made over 40% of the time. Correlation results that consider only stacking complexes in the following section (see Table 3.4, for example) suggest that this is not a consequence of the diversity of interaction types in the S22 set.

Table 3.3 and Fig. 3.5 also show the "auto-correlation" of the CBS-extrapolated CCSD(T) interaction energy – $\Delta E^{\text{CBS}}_{\text{CCSD(T)}} \cdot \Delta E^{\text{CBS}'}_{\text{CCSD(T)}}$, where the original values of Hobza and coworkers ($\Delta E^{\text{CBS}}_{\text{CCSD(T)}}$)[212] are reproduced by those of Fusti Molnar et al. ($\Delta E^{\text{CBS}'}_{\text{CCSD(T)}}$).[213] The percentage of mistakes N_{mis} in this case is not zero due to technical differences methods (note: the latter results also omit two of the largest dimers in the training set). For obvious reasons, this is the best among all correlations shown. The Kendall tau drops below 0.9 only for values of d_{COM} > 3 Å, where the fraction of misaligned pairs exceeds 10% and $\bar{\Delta}^{\text{ref}}_{\text{mis}}$ reaches 1.5 kcal/mol. The statistics associated with the MP2 interaction energy acting as a prognostic are not much worse. For both MP2 and CCSD(T), however, all measures of correlation degrade rapidly at shortened separations, namely for values of d_{COM} below -0.2Å.

Lastly, it is important to notice that statistics for the first order, uncorrelated electrostatic term $\Delta E^{(1)}_{\text{el}}$ do not converge to those of $\Delta E^{\text{CBS}'}_{\text{CCSD(T)}}$ at long range. Since for $d_{\text{COM}} > 1.5$ Å the interaction energy is dominated by non-penetrative, multipole electrostatic effects, intramolecular correlation for long range electrostatic interactions seems to be important. This could be confirmed by repeating the present analysis using multipole-expanded electrostatic interactions based on coupled cluster densities (notice that the delta $\bar{\Delta}^{\text{ref}}_{\text{mis}}$ for $\Delta E^{\text{CBS}}_{\text{CCSD(T)}} \cdot \Delta E_{\text{MP2}}$ also converges to a different limit). If statistical correlation could be obtained with such multipole moments comparable to that found here at large distances for $\Delta E^{\text{CBS}}_{\text{CCSD(T)}} \cdot \Delta E^{\text{CBS}'}_{\text{CCSD(T)}}$, they would provide a valuable and inexpensive tool for predicting stabilization energies, with a success rate of over 90%. More importantly, the average error for the reference interaction energy would probably also be around 1 kcal/mol, comparable to the average deviation reported in a larger portion of the same training set for dispersion corrected density functionals[215].

[215] Antony, J., Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287–5293.

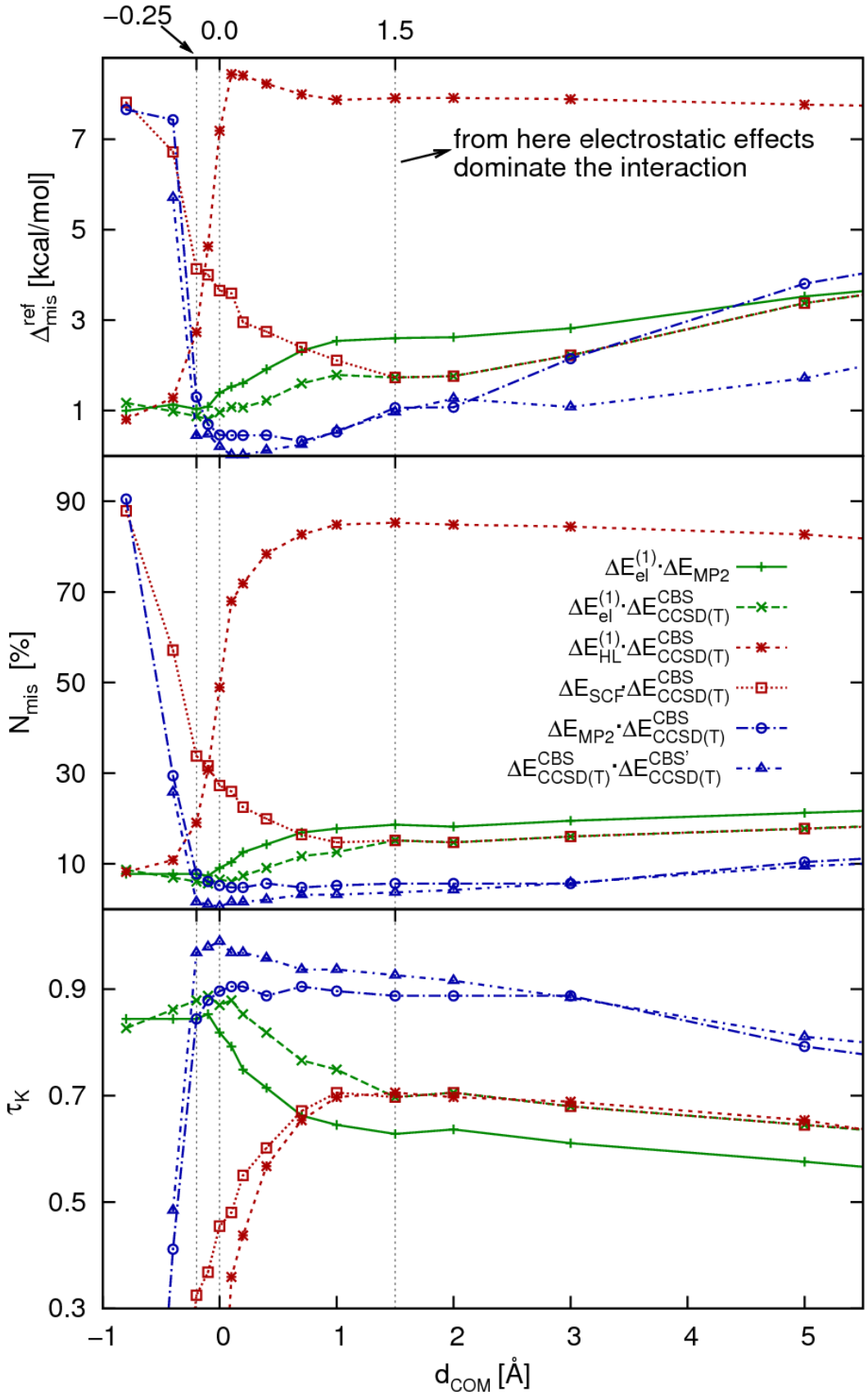


Figure 3.5: Distance dependence of the Kendall tau (τ_K), fraction of misaligned pairs (N_{mis}) and average difference in reference values for misaligned pairs ($\bar{\Delta}_{mis}^{ref}$) in the S22 training set. Green denotes correlations that behave adequately at short separations ($d_{COM} < 0.2$ Å), blue is used for the best correlations in the intermediate and long ranges, and plots involving ΔE_{RHF} and $\Delta E_{HL}^{(1)}$ (with bad correlation around equilibrium) are shown in red.

3.3 Stacked dimers of nucleic acid bases

In this section, the rank-based statistical concepts introduced above are applied to HVPT interaction energy analyses that were performed for stacked nucleic acid base geometries. While a wider, historical context of research on nucleic acids is portrayed in the introduction to Chapter 4, all the topics discussed there depend on recognition and intermolecular interactions, lending more motivation to this study.

Non-covalent interactions of nucleobases in arbitrary intermolecular geometries are complicated by the number of factors that need to be considered simultaneously, including hydrogen bonding and π-π interactions. The first is well characterized in terms of the geometries of donor hydrogens and acceptor atoms and electrostatic forces, the latter however does not seem susceptible to any such straightforward description.[216] Difficulties emerge from multiple intermolecular contacts, variable geometrical parameters and the influence and large number of feasible functional groups. Computational barriers here are also significant. All these complications led researchers to study the smallest representative isolated model systems – such as dimers of benzene-derived molecules and nucleobases. We focus here on stacked nucleobase dimers due the interest they enjoy and the controversy they have caused in the computational literature of the last decade.

As far as the importance of electron correlation effects and choice of methods required to obtain reliable total energies are concerned, the differences between hydrogen-bonded and stacked nucleic acid base dimers are well known. Early studies, summarized already in 1999 by Hobza and Šponer,[217] settled the electrostatic nature of hydrogen bonded base pairs. On the other hand, they revealed the important role of dispersion interactions for the stabilization of π-π stacks.[218] Reliable *ab initio* calculations for nucleic acid bases are themselves relatively new and provide fresh insight into the properties and complexation of these molecules. The general attitude at the turn of the century concerning this topic was concisely described by Sponer et al.:[219] "QM studies of DNA bases have been attempted for more than 30 years. However, before [the] advance of powerful supercomputers in the beginning of the 1990s, no reliable calculations on medium-sized molecular clusters (such as base pairs) were possible. Thus the old results were necessarily highly inaccurate, mutually contradicting, and method dependent. [..] Modern high-level ab initio calculations provide data of great accuracy and reliability, which for nucleobase interactions cannot be presently obtained by any other experimental or computational technique."

Subsequent studies by Hobza and collaborators have chiseled the energies of nucleobases and their dimers to increasingly higher accuracies. The large effort put forward in order to

[216] Hunter, C. A., Lawson, K. R., Perkins, J., Urch, C. J. *J. Chem. Soc. Perkin Trans. 2* **2001**, 651–669; Meyer, E. A., Castellano, R. K., Diederich, F. *Angew. Chem. Int. Ed.* **2003**, *42*, 1210–1250.

[217] Hobza, P., Šponer, J. *Chem. Rev.* **1999**, *99*, 3247–3276.

[218] This is in contrast to isolated nucleobases, for which DFT and MP2 provide similar electric properties, including charges, dipoles and MEPs as shown for methylated nucleobases in Bakalarski, G., Grochowski, P., Kwiatkowski, J. S., Lesyng, B., Leszczyński, J. *Chem. Phys.* **1996**, *204*, 301–311.

[219] Šponer, J., Leszczyński, J., Hobza, P. *Biopolymers* **2002**, *61*, 3–31.

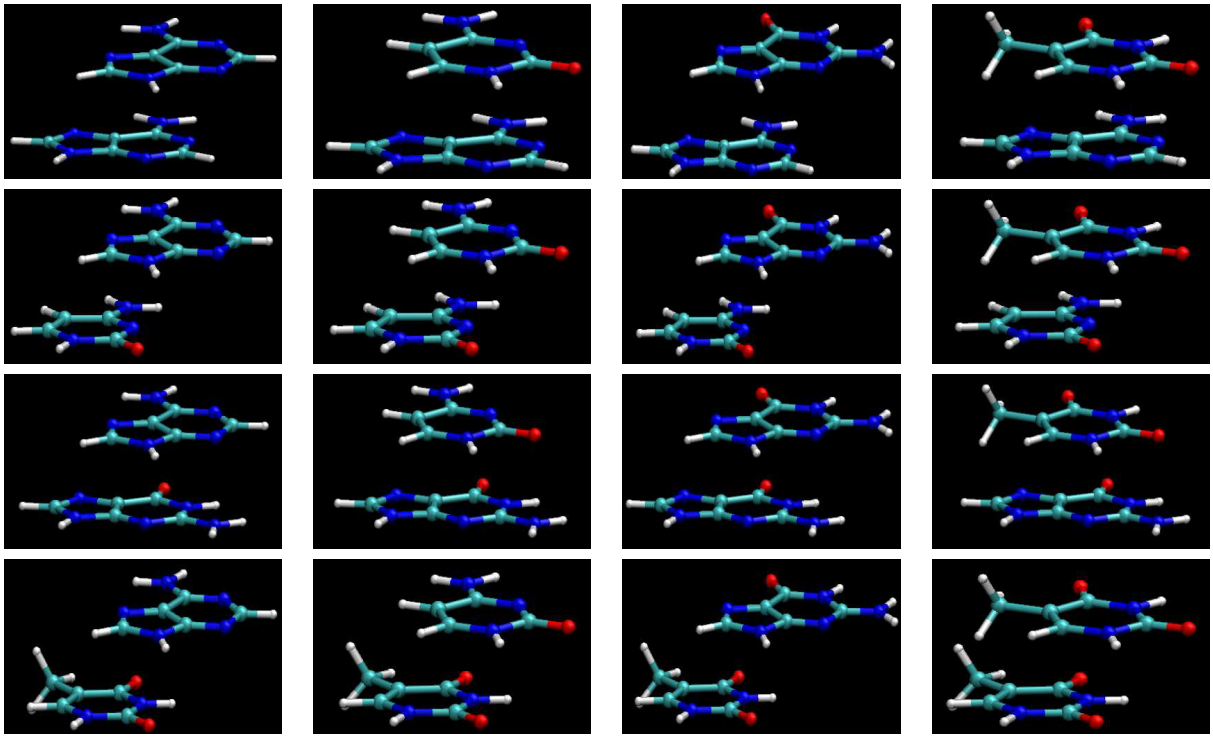


Figure 3.6: All 16 combinations of stacked nucleobase dimers in their typical B-DNA conformations.

apply MP2 and CCSD(T) methods and to extrapolate to complete basis sets[220] gives a clear picture of the gas phase properties of these dimers and the magnitude of the interactions involved.[221] More recent studies have also sought to use these exact references and describe stacked nucleobases with less expensive alternatives, including force fields[222] and approximations such as MP2.5.[223]

These efforts are seconded by studies on simpler stacked aromatic molecules. Tsuzuki et al. showed how important the dispersion interaction is for various conformations of the benzene dimer by comparing interaction energy curves at the Hartree-Fock, MP2 and CCSD(T) levels of theory,[224] followed by a similar inspection of the naphthalene dimer with similar conclusions.[225] Recent, extensive studies of the benzene dimer potential energy surface by Janowski and Pulay have provided more accurate reference data[226] that allow less expensive, but reliable interaction models to be developed.[227] Tschumper and collaborators have demonstrated similar results for

[220]Hobza, P., Šponer, J. *J. Am. Chem. Soc.* **2002**, *124*, 11802–11808.

[221]Jurečka, P., Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608–15613; Šponer, J., Jurečka, P., Marchan, I., Luque, F. J., Orozco, M., Hobza, P. *Chem. Eur. J.* **2006**, *12*, 2854–2865; Šponer, J., Riley, K. E., Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.

[222]Morgado, C. A., Jurečka, P., Svozil, D., Hobza, P., Šponer, J. *J. Chem. Theor. Comp.* **2009**, *5*, 1524–1544.

[223]Pitoňák, M., Janowski, T., Neogrády, P., Pulay, P., Hobza, P. *J. Chem. Theor. Comp.* **2009**, *5*, 1761–1766.

[224]Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M., Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 104–112.

[225]Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. *J. Chem. Phys.* **2004**, *120*, 647.

[226]Janowski, T., Pulay, P. *Chem. Phys. Lett.* **2007**, *447*, 27–32.

[227]Hill, J. G., Platts, J. A., Werner, H. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4072; Rubeš, M., Bludský, O., Nachtigall, P. *ChemPhysChem* **2008**, *9*, 1702–1708; Bludský, O., Rubeš, M., Soldán, P., Nachtigall, P. *J. Chem. Phys.* **2008**, *128*, 114102; Bludský, O., Rubeš, M., Soldán, P. *Phys. Rev. B* **2008**, *77*, 092103; Pitoňák, M., Neogrády, P., Řezáč, J., Jurečka, P., Urban, M., Hobza, P. *J. Chem. Theor. Comp.* **2008**, *4*, 1829–1834.

the diacetylene dimer, an even smaller model for π-π interactions,[228] showing that electron correlation effects must be included in order to retrieve the global minimum on the PES.

With the constant progress of available computer resources, quantum chemical calculations of stacked π-π systems are now branching out to more intricate problems. Rubes et al. addresses stacking interactions in the solid phase,[229] and nucleobases-aromatic amino acid interactions have been studied by others.[230] It should be mentioned, again, that there is a growing body of *ab initio* results on intercalated nucleic acids, which also entail aromatic interactions, as discussed in depth in Section 4.

Yildirim and Turner quite bluntly recap the problems and hopes of the current situation, albeit in the context of RNA folding dynamics, in their review *RNA Challenges for Computational Chemists*[231] – "Some experimental results for [..] folding cannot be explained by simple pairwise hydrogen-bonding models. [..] Presumably, these results can be explained by base stacking effects, which can be partitioned into Coulombic and overlap effects". It is what they call *overlap effects*, and the dispersion interaction in particular, which is so problematic due to the prohibitively expensive computational methods needed to retrieve reliable energies. Although it remains to be seen whether base stacking can in fact satisfactorily explain these experimental results, any information provided via less expensive calculations is valuable if it can make at least qualitative predictions about relative stability or structural properties.

3.3.1 Electrostatic effects in stacked nucleobase dimers

Perhaps the most influential outcome of this branch of research is the widely accepted notion that stacked complexes are stabilized mostly by London dispersion forces, a point repeatedly confirmed by quantum chemistry calculations. This general conclusion is by no means obvious or easy to prove, and in his thoughtful communication *Do Special Noncovalent π-π Stacking Interactions Really Exist?* Grimme recommends a cautious interpretation.[232] He stresses that in stacked aromatic dimers π orbitals do not interact as in conventional overlap-driven covalent bonding, and that the spatial arrangement of fragments is as important as the presence of those π electrons. It is the unique, planar shape of stacked molecules that allows for numerous close atom-atom contacts while remaining outside the extreme Pauli exchange repulsion regime, thus maximizing attractive dispersion forces and leading to overall cooperative π effects. While acknowledging the dominant dispersion component, he also argues that exchange and electrostatic effects push stacked complexes away from maximum overlap into parallel displaced conformations.

The interaction between vertically stacked bases in nucleic acids has long been understood to be an important factor that contributes to their stabilization and recognition in some mech-

[228] Hopkins, B. W., ElSohly, A. M., Tschumper, G. S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1550; ElSohly, A. M., Hopkins, B. W., Copeland, K. L., Tschumper, G. S. *Mol. Phys.* **2009**, *108*, 923–928.

[229] Rubeš, M., Bludský, O. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2611.

[230] Rutledge, L. R., Durst, H. F., Wetmore, S. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2801; Copeland, K. L., Anderson, J. A., Farley, A. R., Cox, J. R., Tschumper, G. S. *J. Phys. Chem. B* **2008**, *112*, 14291–14295.

[231] Yildirim, I., Turner, D. H. *Biochemistry* **2005**, *44*, 13225–13234.

[232] Grimme, S. *Angew. Chem. Int. Ed.* **2008**, *47*, 3430–3434.

anisms, for example those involving enzymatic replication. Early experimental research, such as the NMR studies of 6-methylpurine performed by Chan et al.,[233] succeeded in identifying the average mode of nucleobase association in solution as being partial overlap, in preference to horizontal hydrogen bonding.

A significant, if vague role of electrostatic effects has often been pointed out in the interpretations of experiments involving aromatic stacking complexes. Siegel and coworkers presented elegant experiments to this end. By manipulating the electrostatic potential at the center of the benzene ring face by substituents, they effectively affected the magnitude of interactions between aligned aromatic molecules.[234] A different approach was adopted by Newcomb and Gellman, who compared bis-adenine and bis-naphthyl to control compounds, and found that only the first associates via intramolecular stacking;[235] since the attractive dispersion interaction should be available for both compounds, they concluded that it is not a decisive force for stacking. On the other hand, Guckian et al. have argued that the affinity of nucleobase-derived molecules to form stacking complexes in solution depends mostly on the overlap area of monomers, at the same time indicating the governing role of solvation-driven hydrophobic effects.[236] Interestingly, they also observed that the stacking affinities of natural nucleobases are far from maximal, a fact they ascribe to the necessity of the DNA double helix to unwind.

Another interesting experimental study was performed by Perez-Casas et al.,[237] who determined the association enthalpies of various aromatic π-π complexes by heat capacity measurements in solution. By correlating these with interaction energies from molecular and distributed multipole moments, they succeed in explaining the relative association enthalpies with significant correlation coefficients. Cockroft et al. in different studies argue that substituent effects can also be rationalized by electrostatic effects.[238] Itahawa and Imaizumi also published results of solution experiments that stress electrostatic interactions in the mechanism of aromatic stacking.[239]

From the theoretical side, an early model based on distributed multipole moments employed by Price and Stone[240] demonstrated its utility by demonstrating qualitative agreement with experimental geometries. They also pointed out that simpler models, which employ empirical point charges or molecular multipoles, fail to reproduce these geometries even qualitatively.

More recently, Swart et al. have decomposed the interaction energy in a survey of density functionals for benzene-derived and nucleobase dimers.[241] Their conclusions included an inter-

[233]Chan, S. I., Schweizer, M. P., Ts'o, P. O. P., Helmkamp, G. K. *J. Am. Chem. Soc.* **1964**, *86*, 4182–&.
[234]Cozzi, F., Cinquini, M., Annuziata, R., Siegel, J. S. *J. Am. Chem. Soc.* **1993**, *115*, 5330–5331.
[235]Newcomb, L. F., Gellman, S. H. *J. Am. Chem. Soc.* **1994**, *116*, 4993–4994.
[236]Guckian, K. M., Schweitzer, B. A., Ren, R. X.-F., Sheils, C. J., Paris, P. L., Tahmassebi, D. C., Kool, E. T. *J. Am. Chem. Soc.* **1996**, *118*, 8182–8183; Guckian, K. M., Schweitzer, B. A., Ren, R. X.-F., Sheils, C. J., Tahmassebi, D. C., Kool, E. T. *J. Am. Chem. Soc.* **2000**, *122*, 2213–2222.
[237]Pérez-Casas, S., Hernández-Trujillo, J., Costas, M. *J. Phys. Chem. B* **2003**, *107*, 4167–4174.
[238]Cockroft, S. L., Hunter, C. A., Lawson, K. R., Perkins, J., Urch, C. J. *J. Am. Chem. Soc.* **2005**, *127*, 8594–8595; Cockroft, S. L., Perkins, J., Zonta, C., Adams, H., Spey, S. E., Low, C. M. R., Vinter, J. G., Lawson, K. R., Urch, C. J., Hunter, C. A. *Org. Biomol. Chem.* **2007**, *5*, 1062.
[239]Itahara, T., Imaizumi, K. *J. Phys. Chem. B* **2007**, *111*, 2025–2032.
[240]Price, S. L., Stone, A. J. *J. Chem. Phys.* **1987**, *86*, 2859–2868.
[241]Swart, M., Wijst, T., Guerra, C. F., Bickelhaupt, F. M. *J. Mol. Model.* **2007**, *13*, 1245–1257.

esting point – that the classical electrostatic component is the most important factor shaping the surface and depth of the PES for stacked nucleobases and causes a minimum to occur along the energy profile of two stacked Watson-Crick base pairs at a twist angle of 36°. A similar interpretation was given a decade earlier by Hunter et al.,[242] who surmised the important role of electrostatic interactions in determining the shift and slide of nucleobase stacks. Tsuzuki and coworkers, in one their influential studies of the benzene dimer[224] also conclude that only electrostatic interactions are highly orientation-dependent and that dispersion and electrostatics together determine the dimer's directionality.

3.3.2 Correlations between interaction energy components

The literature reviewed above conjures a vague image of the interaction profile for stacked aromatic complexes; in it, London dispersion effects account for most of the interaction energy. Other effects, supposedly the electrostatic component foremost, determine the geometrical details and relative stability.

We note that although calculations performed for large sets of stacked aromatic molecules are well represented in the recent literature both at the most accurate[212] and more cost-effective[215] levels, **a systematic analysis of interaction components, their interplay and relationship to the stacking geometry is lacking.** It is the intent here to detail this interaction profile from another aspect, by considering the interaction components for stacked nucleobases. Similar to a recent study[243], the interaction energy components were analyzed for each stacked pair of nucleic acid bases. Whereas Heßelmann et al. adopted the DFT-SAPT approach, we use the hybrid variation-perturbation decomposition scheme described in Section 2.2.3.

Methodologically, the published results[244] are a direct extension of a study conducted earlier by Hill et al.,[245] who have shown for a set of 10 stacked DNA bases that electrostatic interactions and their multipole estimates follow ΔE_{MP2} with a reasonable correlation coefficient.

This conclusion was later questioned by Toczyłowski and Cybulski,[246] who point out that the electrostatic penetration contribution ($\Delta E_{\mathrm{el,pen}}$) between two stacked nucleic acids bases cannot be neglected and that basis set dependence may destroy the observed correlations. We aim to better establish the statistical significance of the assertions stated previously by Hill et al. for stacked nucleic acid bases by considering various multipole expansion types and a series of basis sets. Also, we highlight the failure of molecular multipole expansions to estimate electrostatic interactions, and the limitations of multicenter expansions based on atoms due to slow convergence and penetration effects.

[242] Hunter, C. A., Lu, X.-J. *J. Mol. Biol.* **1997**, *265*, 603–619.
[243] Heßelmann, A., Jensen, G., Schütz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730–11731.
[244] Langner, K. M., Sokalski, W. A., Leszczyński, J. *J. Chem. Phys.* **2007**, *127*, 111102.
[245] Hill, G., Forde, G., Hill, N., Lester, W. A., Sokalski, W. A., Leszczyński, J. *Chem. Phys. Lett.* **2003**, *381*, 729–732.
[246] Toczyłowski, R. R., Cybulski, S. M. *J. Chem. Phys.* **2005**, *123*, 154312–12.

		ΔE_{MP2}	ΔE_{corr}	$\Delta E^{(2)}_{\mathrm{disp}}$	ΔE_{RHF}	$\Delta E^{(\mathrm{R})}_{\mathrm{del}}$	$\Delta E^{(1)}_{\mathrm{ex}}$	$\Delta E^{(1)}_{\mathrm{el}}$
$\Delta E^{\mathrm{mol},3}_{\mathrm{el,mtp}}$	cc-pVDZ	0.547	-0.182	-0.041	0.450	0.129	0.071	0.471
	cc-pVTZ	0.559	-0.185	-0.053	0.479	0.156	0.097	0.503
$\Delta E^{\mathrm{mol},8}_{\mathrm{el,mtp}}$	cc-pVDZ	-0.024	-0.103	-0.124	0.185	-0.024	-0.038	-0.009
	cc-pVTZ	-0.047	-0.115	-0.021	0.171	-0.009	-0.062	-0.062
$\Delta E^{1}_{\mathrm{CAMM}}$	cc-pVDZ	0.924	-0.156	0.009	0.688	0.265	-0.044	0.818
	cc-PVTZ	0.659	-0.232	-0.068	0.650	0.156	-0.015	0.776
$\Delta E^{8}_{\mathrm{CAMM}}$	cc-pVDZ	0.818	-0.315	-0.147	0.788	0.156	0.088	0.794
	cc-pVTZ	0.756	-0.321	-0.135	0.821	0.168	0.115	0.812
$\Delta E^{(1)}_{\mathrm{el}}$	cc-pVDZ	**0.874**	0.144	0.329	0.388	0.338	-0.379	
	aug-cc-pVDZ	**0.924**	0.115	0.294	0.447	0.382	-0.315	
	cc-pVTZ	**0.944**	0.144	0.315	0.424	0.353	-0.341	
$\Delta E^{(1)}_{\mathrm{ex}}$	cc-pVDZ	-0.221	-0.921	**-0.950**	0.576	-0.429		
	aug-cc-pVDZ	-0.385	-0.929	**-0.976**	0.582	-0.479		
	cc-pVTZ	-0.365	-0.929	**-0.976**	0.576	-0.429		
$\Delta E^{(\mathrm{R})}_{\mathrm{del}}$	cc-pVDZ	0.147	0.209	0.350	0.032			
	aug-cc-pVDZ	0.315	0.274	0.447	0.012			
	cc-pVTZ	0.238	0.250	0.418	0.032			
ΔE_{RHF}	cc-pVDZ	0.491	-0.753	-0.635				
	aug-cc-pVDZ	0.347	-0.750	-0.600				
	cc-pVTZ	0.371	-0.747	-0.597				
$\Delta E^{(2)}_{\mathrm{disp}}$	cc-pVDZ	0.238	0.965					
	aug-cc-pVDZ	0.400	0.962					
	cc-pVTZ	0.368	0.962					
ΔE_{corr}	cc-pVDZ	0.085						
	aug-cc-pVDZ	0.250						
	cc-pVTZ	0.224						

Table 3.4: Spearman rank correlation coefficients between various interaction components for all the possible pairs of stacked nucleic acid bases in B-form DNA (**set 1**), using the cc-pVDZ, aug-cc-pVDZ and cc-pVTZ basis sets.

However subtle the supposed role of monomer electrostatic interactions is, it may still provide information on the relative stability of the different complexes, and a reliable electrostatic model with a known margin of error is of major interest. That is the practical precedent for this study, in which we survey all 16 B-DNA type stacked nucleobases (shown in Fig. 3.6) and compare the electrostatic contribution to the total interaction energy at the second order Möller-Plesset level ΔE_{MP2}. We go further, however, and address the statistical relationships of any two interaction energy components – statistics for these relationships are summarized in Table 3.4.

The structures chosen for this study fall into one of four categories,

set **1**: all 16 base pairs of stacked model B-form DNA (base step: 3.38 Å, twist: -36^o)
set **2**: all 16 base pairs of stacked model A-form DNA (base step: 2.56 Å, twist: -32.7^o)
set **3**: 6 pairs of stacked bases studied by Hill et al.[245]
set **4**: 18 pairs of stacked bases published by Jurecka et al.[212]

Since there is no reason to assume a normal distribution of the interaction energies being compared, Spearman rank correlation coefficient as defined in (3.3) were used to describe the

relationships between interaction terms, instead of the more popular Pearson product-moment correlation coefficient.

Among all pairs of interaction terms, the correlation coefficients of a few are meaningfully high; among these, $\rho(\Delta E_{\mathrm{el}}^{(1)}, \Delta E_{\mathrm{MP2}})$ and $\rho(\Delta E_{\mathrm{ex}}^{(1)}, \Delta E_{\mathrm{disp}}^{(2)})$ are the most noteworthy. These two pairs of components and a few others are plotted in Fig. 3.7 for the aug-cc-pVDZ and cc-pVTZ basis sets. The correlation coefficients for all pairs of interaction energy components are shown for set **1** in Table 3.4. For the remaining sets of geometries, the two most significant correlations as well as $\rho(\Delta E_{\mathrm{RHF}}, \Delta E_{\mathrm{MP2}})$ and $\rho(\Delta E_{\mathrm{disp}}^{(2)}, \Delta E_{\mathrm{MP2}})$ are listed in Table 3.5.

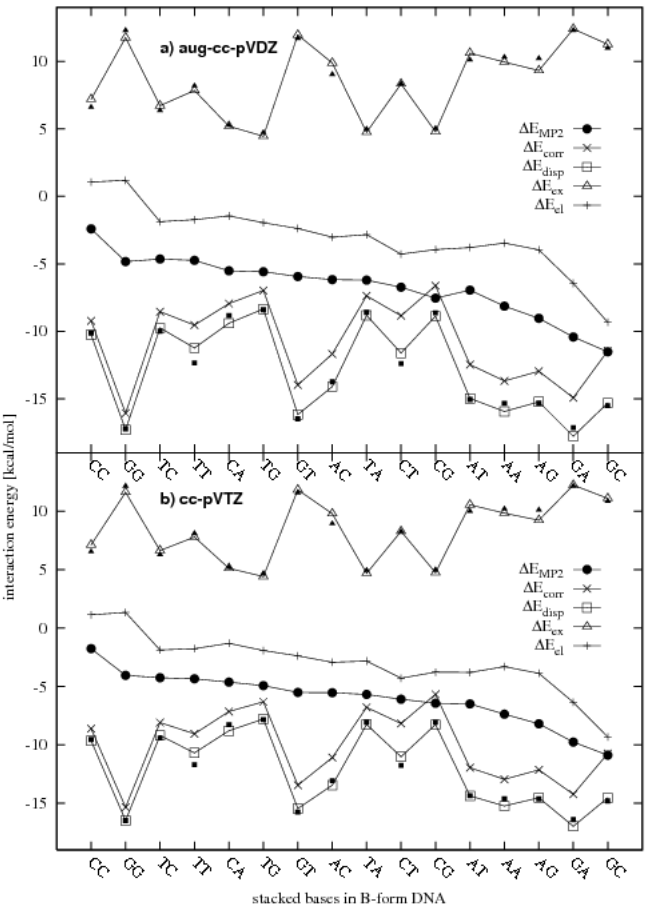


Figure 3.7: Second-order Möller-Plesset interaction energy and its selected components for all 16 pairs of stacked nucleic acid bases in B-form DNA (set **1**): a) aug-cc-pVDZ basis set, b) cc-pVTZ basis set. The structures on the x-axis are ordered by decreasing MP2 interaction energy (ΔE_{MP2}) in the cc-pVTZ basis set. The smaller, filled points (triangles and squares) near the plots of the exchange and dispersion terms ($\Delta E_{\mathrm{ex}}^{(1)}$ and $\Delta E_{\mathrm{disp}}^{(2)}$, respectively) represent values fitted using the function described by 3.10.

Surprisingly, the most pronounced correlation is observed between the exchange $\Delta E_{\mathrm{ex}}^{(1)}$ and dispersion $\Delta E_{\mathrm{disp}}^{(2)}$ terms, with a correlation coefficient below -0.95 in set **1** for all the basis sets studied. Some insight into the practical applicability of such a correlation can be gained by calculating the prediction interval of the difference between $\Delta E_{\mathrm{disp}}^{(2)}$ and its linear regression estimate $a\Delta E_{\mathrm{ex}}^{(1)}+b$. The 95% prediction interval of this kind is below 1.4 kcal/mol for all basis sets, which means that 95% of the values calculated from such a linear regression equation are expected to be within 1.4 kcal/mol of $\Delta E_{\mathrm{disp}}^{(2)}$. This is a crude estimate due to the small population size (16 geometries) and the assumption of a linear relationship; nonetheless, it gives an idea of the minimum difference in energies needed to draw conclusions about the correlated interaction term. The correlation between $\Delta E_{\mathrm{ex}}^{(1)}$ and $\Delta E_{\mathrm{corr}} = \Delta E_{\mathrm{MP2}} - \Delta E_{\mathrm{RHF}}$ is weaker, but still evident - with a correlation coefficient below -0.92 and prediction interval below 1.8 kcal/mol for all basis sets (also within geometry set **1**).

These results may indicate a strong relationship between the dispersion damping and exchange repulsion interactions, both closely related with intermolecular overlap as in the interaction model devised by Tang and Toennies[247]. While it has been acknowledged earlier that the Pauli exchange and dispersion components of the interaction energy in stacked structures

[247]Tang, K.-T., Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

geometries/basis set	$\rho(\Delta E_{\rm el}^{(1)}, \Delta E_{\rm MP2})$	$\rho(\Delta E_{\rm ex}^{(1)}, \Delta E_{\rm disp}^{(2)})$	$\rho(\Delta E_{\rm RHF}, \Delta E_{\rm MP2})$	$\rho(\Delta E_{\rm disp}^{(2)}, \Delta E_{\rm MP2})$
set **1** (B-DNA)/631g(1d,1p)	0.800	-0.962	0.647	0.024
set **1** (B-DNA)/631g(2d,2p)	0.912	-0.965	0.374	0.341
set **1** (B-DNA)/6311g(1d,1p)	0.879	-0.953	0.494	0.253
set **1** (B-DNA)/631g(1d,1p,1f)	0.879	-0.965	0.494	0.203
set **1** (B-DNA)/cc-pVDZ	0.874	-0.950	0.491	0.238
set **1** (B-DNA)/cc-pVTZ	0.944	-0.976	0.371	0.368
set **1** (B-DNA)/aug-cc-pVDZ	0.924	-0.976	0.347	0.400
set **2** (A-DNA)/cc-pVDZ	-0.238	-0.971	0.829	-0.535
set **2** (A-DNA[a])/cc-pVDZ	0.643	-0.930	0.615	0.056
set **3** (Hill et al.[245])/6311g(1d,1p)	0.943	-0.943	0.257	-0.200
set **3** (Hill et al.[245])/cc-pVDZ	0.829	-0.943	0.257	-0.200
set **4** (Jurecka et al.[212])/631g1d1p	0.794	-0.781	0.593	0.480
set **4** (Jurecka et al.[212])/6311g1d1p	0.756	-0.777	0.536	0.540
set **4** (Jurecka et al.[212])/cc-pVDZ	0.771	-0.777	0.546	0.519
all sets combined	0.358	-0.794	-0.115	0.758

Table 3.5: Spearman rank correlation and other coefficients between selected interaction energy components for the four studied sets of stacked nucleic acid geometries and for various basis sets.
[a] Values for set **2** without the AT, CT, GT, and TT stacks (outliers in Fig.2), in which a misplaced methyl group (distance from hydrogen to nearest atom below 2 Å) introduced additional non-stacking exchange interactions.

have opposite signs and large values that partially compensate each other[245], we point out that they correspond closely - although not in a linear fashion - at least within a set of geometrically similar structures (here, for B-DNA type geometries). Such observations complement and may help support approximate methods for calculating dispersion interactions that are presently being developed, for example based on density functional theory[248].

The correlation coefficient between $\Delta E_{\rm el}^{(1)}$ and $\Delta E_{\rm MP2}$ is also significant, above 0.85 for all basis sets, close to values reported earlier by Perez-Casas et al.[237] and Hill et al..[245]. The 95% linear prediction interval in this case is always below 1.5 kcal/mol. A practical use of this result: it would likely be sound to assert based on electrostatic interactions alone that among the B-DNA stacked nucleobases the guanine-cytosine dimer (GC) is more stable than any other. On the other hand, the relative stability of CT, CG, AT, AA and AG should not be discriminated based on the electrostatic component if a confidence level of the order of 95% is needed. In fact, the electrostatic components for this series of dimers are in opposition to the $\Delta E_{\rm MP2}$ energy.

A few observations should be added for the multipole estimates to the electrostatic interaction. Foremost, interactions based on *molecular* electrostatic multipoles do not correlate well with any other terms, even $\Delta E_{\rm el}^{(1)}$, and quickly diverge (the correlation coefficient tends to zero when including higher order moments). Such behavior is reminiscent of the example presented in order to illustrate convergence in Fig. 2.4 of Section 2.5.3. In the case of the atomic multipole expansion used (CAMM), correlation with $\Delta E_{\rm MP2}$ exists but is worse than for the full electrostatic interaction $\Delta E_{\rm el}^{(1)}$ due to penetration effects.

The relationships outlined above refer to the structures in set **1**, namely B-form DNA constructed in ideal geometries. Results for the other geometry sets and basis sets are shown in Table 3.6 and the full list of statistical parameters published as Supporting Information[244].

[248] Misquitta, A. J., Jeziorski, B., Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201; Heßelmann, A., Jansen, G., Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

	$\Delta E^{(2)}_{\rm disp} = a\Delta E^{(1)}_{\rm ex} + b$				$\Delta E_{\rm MP2} = a\Delta E^{(1)}_{\rm el} + b$			
geometries/basis set	a	b	R^2	P	a	b	R^2	P
set **1** (B-DNA)/631g(1d,1p)	-0.85	-1.49	0.94	0.99	0.70	-0.39	0.88	1.09
set **1** (B-DNA)/631g(1d,1p,1f)	-0.92	-1.55	0.94	1.08	0.74	-0.90	0.90	1.01
set **1** (B-DNA)/6-311G(1d,1p)	-0.98	-1.57	0.93	1.18	0.75	-1.48	0.90	1.05
set **1** (B-DNA)/631g(2d,2p)	-1.04	-2.12	0.94	1.16	0.76	-2.42	0.89	1.10
set **1** (B-DNA)/cc-pVDZ	-0.94	-1.32	0.93	1.14	0.73	-0.95	0.90	1.02
set **1** (B-DNA)/cc-pVTZ	-1.15	-2.46	0.94	1.32	0.82	-3.57	0.87	1.35
set **1** (B-DNA)/aug-cc-pVDZ	-1.17	-2.87	0.94	1.36	0.83	-4.14	0.85	1.45
set **2** (A-DNA)/cc-pVDZ	-0.33	-5.60	0.89	3.17	-0.58	-3.18	0.35	7.87
set **3** (Hill et al.[245])/6-311G(1d,1p)	-1.26	-1.61	0.38	0.99	0.52	-3.06	0.81	0.70
set **3** (Hill et al.[245])/cc-pVDZ	-1.22	-1.32	0.37	1.08	0.53	-2.59	0.85	0.59
set **4** (Jurecka et al.[212])/6-31G(1d,1p)	-0.48	-5.49	0.69	1.82	0.48	-2.51	0.86	1.48
set **4** (Jurecka et al.[212])/6-311G(1d,1p)	-0.52	-6.53	0.67	2.11	0.48	-3.99	0.83	1.66
set **4** (Jurecka et al.[212])/cc-pVDZ	-0.50	-6.13	0.67	2.03	0.47	-3.47	0.81	1.67

Table 3.6: Linear regression parameters (a, b, and R^2) and prediction intervals ($P = \sigma_n T_{0.95}\sqrt{1+(1/n)}$, where $T_{0.95}$ is the appropriate percentile of Student's t-distribution) for the pairs $\Delta E^{(1)}_{\rm ex}$ and $\Delta E^{(2)}_{\rm disp}$, and $\Delta E^{(1)}_{\rm el}$ and $\Delta E_{\rm MP2}$. All values in kcal/mol where applicable.

They are similar, provided that the chosen structures are planar and span a significant range of interaction energies. For instance, four dimers in set **2** (A-form DNA) exhibit interatomic distances smaller than 2 Å due to a nonplanar methyl group, whose presence disrupts the relationships between interaction terms. Without these four dimers, the same relationships (between $\Delta E^{(1)}_{\rm ex}$ and $\Delta E^{(2)}_{\rm disp}$, and $\Delta E^{(1)}_{\rm el}$ and $\Delta E_{\rm MP2}$) are observed. Conversely, these relationships are much weaker when dimers with different stacking geometries are evaluated together; set **4** (Jurečka *et al.* contains both experimental and *in vacuo* optimized structures, and in this case $\rho(\Delta E^{(1)}_{\rm el}, \Delta E_{\rm MP2}) = 0.771$ and $\rho(\Delta E^{(1)}_{\rm ex}, \Delta E^{(2)}_{\rm disp}) = -0.777$. Calculated for all the geometry sets together, these coefficients are $\rho(\Delta E^{(1)}_{\rm el}, \Delta E_{\rm MP2}) = 0.358$ and $\rho(\Delta E^{(1)}_{\rm ex}, \Delta E^{(2)}_{\rm disp}) = -0.794$. This also suggests that if a general functional relationship can be drawn between these interaction components, covering various degrees of overlap between stacked molecules, it is certainly not linear.

One more point that needs to be emphasized is that the relationships found hold for all the tested basis sets, even though the energies obtained are certainly not saturated in terms of electronic correlation. The linear regression parameters, however, differ between basis sets.

It is important to keep in mind when studying correlations like those above that they do not not infer anything about the dependence or common origins of interaction energy terms being considered – which would lead to a logical fallacy of the *cum hoc ergo propter hoc* kind. In this context, this means simply that *statistical correlation does not imply causation* or, as would be very useful computationally, *does not imply their compensation*. In the case of $\Delta E^{(1)}_{\rm ex}$ and $\Delta E^{(2)}_{\rm disp}$, both terms possibly depend similarly on the degree to which the stacked planar molecules overlap, despite very different physical origins.

This behavior they have in common can be seen by fitting the exchange and dispersion components with any measure of the overlap. For example, Fig.1 also shows that the *ab initio* exchange and dispersion terms are closely followed by functions that consider the distances

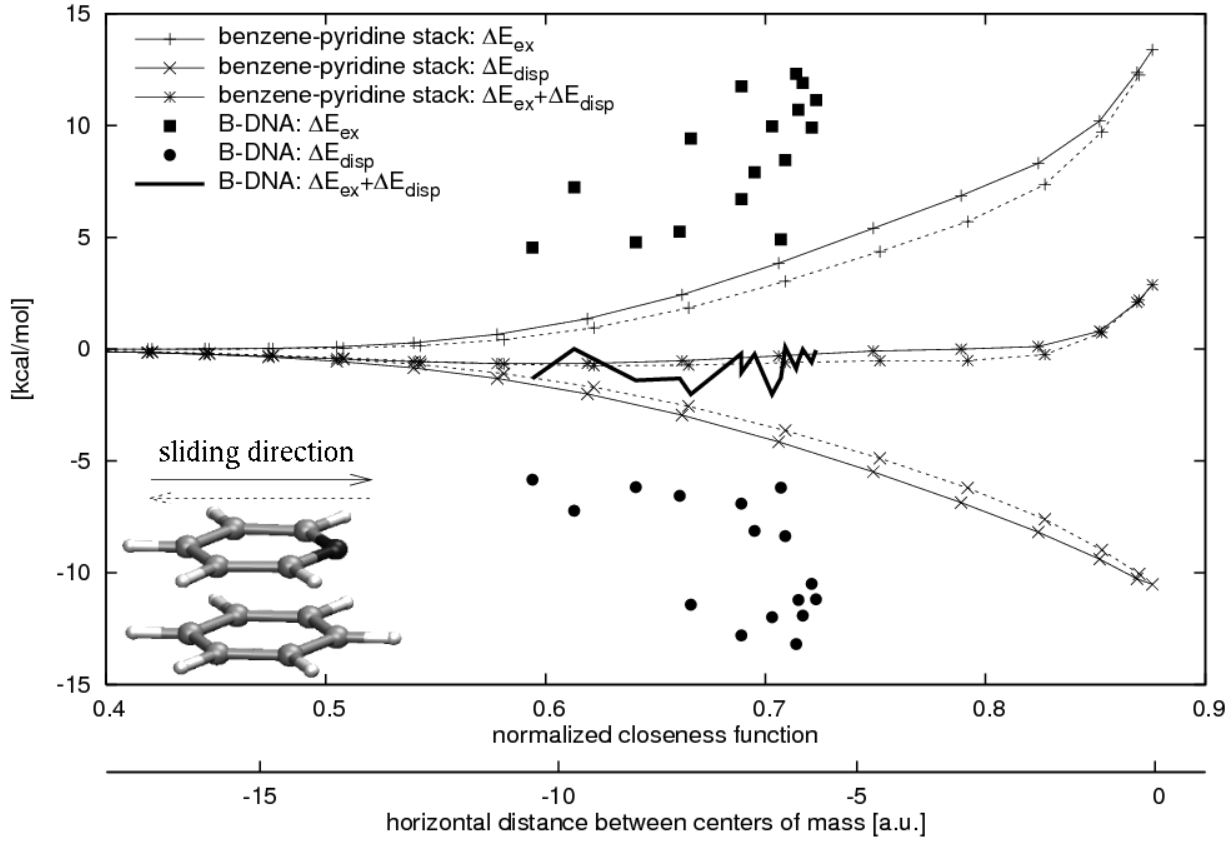


Figure 3.8: The exchange and dispersion interaction components ($\Delta E_{\text{ex}}^{(1)}$ and $\Delta E_{\text{disp}}^{(2)}$, respectively) in a benzene-pyridine stack (parallel with a separation of 3.4Å) and all 16 stacked nucleic acid base pairs of B-DNA (set **1**), plotted against the normalized closeness function in 3.10 with $n = 1$, i.e. $\frac{D}{N_A N_B} \sum_{i,j} \frac{1}{|\vec{r}_i - \vec{r}_j|}$, where D is the distance between the centers of mass and N_A and N_B are the numbers of atoms in each molecule. The range of overlaps for the benzene-pyridine stack were generated by sliding the pyridine along the center of its mass-nitrogen axis, with a normalized closeness of ∼0.8 corresponding to maximum overlap.

between all pairs of atoms in two interacting molecules, in effect a function of their *closeness*:

$$C_n = \sum_i^{N_A} \sum_j^{N_B} \frac{1}{|\vec{r}_i - \vec{r}_j|^n}, \tag{3.10}$$

where i and j span over the heavy atoms in each molecule, in simple analogy to the multipolar dispersion interaction advanced by Hodges and Stone.[249] The exponent in this equation was chosen $n = 12$ for $\Delta E_{\text{ex}}^{(1)}$ and $n = 6$ for $\Delta E_{\text{disp}}^{(2)}$. Other similar, but arbitrary choices also yield satisfactory fits (data not shown).

Looking at the compensation of the exchange and dispersion terms for a wider range of overlaps hints that their relationship is not a linear one, and that the sum $\Delta E_{\text{ex}}^{(1)} + \Delta E_{\text{disp}}^{(2)}$ exhibits a trend for larger relative overlaps. This is illustrated here for a model benzene-pyridine stacked dimer in Fig. 3.8. More importantly, normalized measures of overlap fall into well-defined ranges for sets of related structures. In the case of set **1** (B-DNA), this range is 0.59-0.72 and the sum of $\Delta E_{\text{ex}}^{(1)}$ and $\Delta E_{\text{disp}}^{(2)}$ is not constant due to differences in chemical composition and conformation.

[249] Hodges, M. P., Stone, A. J. *Mol. Phys.* **2000**, *98*, 275–286.

3.4 Conclusions

The surprising observation that at shortened intermolecular separations electrostatic effects correlate stronger with the total equilibrium stabilization energy than either the MP2 or CCSD(T) interaction energies do confirms the motivation behind the first issue outlined in Section 1.2. An abrupt loss of relation between MP2 as well as CCSD(T) interaction energies and the interaction at equilibrium is a significant feature that can impact a number of scenarios met in everyday computational chemistry. The fact that the correlation exhibited by electrostatic interactions degrades marginally for these shorter contacts provides a way to deal with these situations.

On the other hand, these correlations become gradually worse when studied for increasing intermolecular separations. This suggests that the idea of molecular recognition formulated by Kier and brought up in the Introduction can also be applied to the domain of interaction energies. It is in fact possible to reproduce to some extent relative stability based on interactions at larger separations.

Furthermore, the different behavior of the statistics introduced for the electrostatic component and total correlated energies (both ΔE_{MP2} and the reference $\Delta E_{\mathrm{CCSD(T)}}$) point to the possibility that correlated electrostatic effects are important. Repeating the present analyses for electrostatic interaction derived from monomer Coupled-Cluster densities may provide a route to further improve its prognostic value.

The relation observed by Hill et al.[245] is confirmed within sets of similar structures for a selection of basis sets and the predictive strength of this relationship is assessed. Also, the failure of molecular multipole expansions to estimate electrostatic interactions is highlighted, along with the limitations of multicenter expansions based on atoms due to slow convergence and penetration effects.

Major components of the interaction energy that define several approximate levels starting from second order Möller-Plesset theory were studied for 58 stacked nucleic acid dimers. They included typical B-DNA and A-DNA structures, and selected published geometries. A survey of the various terms yields an unexpected correlation between the Pauli exchange and dispersion or correlation terms, which holds for each class of similar planar geometries and for various basis sets. The geometries that exhibit these correlations span a specific range of molecular overlaps when compared to a model benzene-pyridine stacked dimer. Also, the relationship between electrostatic interactions and MP2 stabilization energies reported earlier is confirmed and a prediction interval of practical relevance is estimated.

4 Non-empirical analyses of intercalated nucleic acids

> It is these chromosomes... that contain in some kind of code-script the entire pattern of the individual's future development and of its functioning in the mature state. Every complete set of chromosomes contains the full code...
>
> Erwin Schrödinger
> *What Is Life?* **1944**

4.1 Introduction

Already in the late 19$^{\text{th}}$ century nucleic acid had been isolated and studied as a chemical substance, though it was the identification of its unique function for life on Earth that made it the focal point of molecular biology it is today. A long trail of experimental research in the mid-20$^{\text{th}}$ century inspired the concept: genetic information is embodied and carried by DNA and propagated through processes such as replication and transcription. Several important, less known early studies can be distinguished in this regard, such as those by Avery et al. who already in 1944 induced predictable alterations in the cellular structure of bacterial hosts using isolated salts of deoxyribonucleic acid.[250] Hershey and Chase followed by demonstrating the independent, auxiliary role of protein compared to DNA in bacteriophage growth,[251] with Meselson and Stahl raising important questions about the molecular mechanisms of DNA duplication.[252] These studies and countless others provide the basis for our current understanding of the special, hereditary role of nucleic acids as opposed to proteins, lipids, polysaccharides and other molecules found in cells.

No less important was the recognition of DNA's regular structure in a series of articles in 1953, the first of which placed the phosphate groups on the exterior of the X-ray crystallographic unit[253] and suggested the celebrated double helix.[254]. Franklin et al. found evidence of the same helical structure *in vivo*[255], Wilkins et al. demonstrated *A* and *B* varieties,[256] and Watson and Crick famously discussed the genetic implications of their discovery.[257]

[250] Avery, O. T., MacLeod, C. M., McCarty, M. *J. Exp. Med.* **1944**, *79*, 137–158.
[251] Hershey, A. D., Chase, M. *J. Gen. Physiol.* **1952**, *36*, 39–56.
[252] Meselson, M., Stahl, F. W. *Proc. Natl. Acad. Sci.* **1958**, *44*, 671–682.
[253] Franklin, R. E., Gosling, R. G. *Nature* **1953**, *172*, 156–157.
[254] Watson, J. D., Crick, F. H. C. *Nature* **1953**, *171*, 737 – 738.
[255] Franklin, R. E., Gosling, R. G. *Nature* **1953**, *171*, 740–741.
[256] Wilkins, M. H. F., Stokes, A. R., Wilson, H. R. *Nature* **1953**, *172*, 738–740.
[257] Watson, J. D., Crick, F. H. C. *Nature* **1953**, *171*, 964–968.

Much of the subsequent research – with the basics summarized for example by Lane and Jenkins[258] – has related the macromolecule's structural features to its basic functional properties and recognized the molecular nature of key physicochemical processes such as double helix melting[259] and nucleic acid base dimerization.[260] The place of DNA within the intricate biochemical machinery of the cell has also been mapped, aided by discoveries of specialized molecules dedicated to modifying and assisting the DNA function, such as topoisomerases[261] and telomerases.[262]

A rich body of functional knowledge combined with the unique structural features of nucleic acid strands – their relative planar stacking of aromatic bases along a helix sugar phosphate backbone – are used to tailor new functional materials[263] and sub-micron patterns[264] in an increasingly popular, emerging branch of nucleic acid nanotechnology. Another example of how this knowledge is used are foldamers, which are designed specifically to mimic certain behaviors and conformational patterns of DNA outside the cell environment.[265]

Among the many processes and applications associated with nucleic acids, their complexation with specific small molecules undoubtedly has the largest impact on everyday life. One of the recognized triumphs of medical research in the 20th century was the development and clinical use of biologically active agents for successfully treating cancer. An early major step forward in this regard has been understanding that such chemicals inhibit DNA synthesis by interacting physically so as to distort structure and function.[266]

This medical application of science, which has saved and prolonged the lives of many, directly benefits from fundamental investigations of small active agents. It is not surprising therefore that their interactions, complexes and structural perturbations have consistently provokes interest. Crystallographic and thermodynamic studies played an increasingly central role,[267] relating structure and function with the energetic driving forces behind drug action.[268] Subtleties including selectivity for specific sequences or structural features have been extracted thermodynamically, for example in the competition dialysis experiments reported by Ren and Chaires.[269] Most recently, other less ubiquitous methods have been brought into the arsenal

[258]Lane, A. N., Jenkins, T. C. *Curr. Org. Chem.* **2001**, *5*, 845 – 869.

[259]Zimm, B. H. *J. Chem. Phys.* **1960**, *33*, 1349–1356; Breslauer, K. J., Frank, R., Blöcker, H., Marky, L. A. *Proc. Natl. Acad. Sci.* **1986**, *83*, 3746–3750.

[260]Lamola, A. A., Eisinger, J. *Proc. Natl. Acad. Sci.* **1968**, *59*, 46–51.

[261]Champoux, J. J. *Ann. Rev. Biochem.* **2001**, *70*, 369 – 413; Wang, J. C. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 430 – 440.

[262]Blackburn, E. H. *Nature* **2000**, *408*, 53–56; Blackburn, E. H. *Cell* **2001**, *106*, 661–673.

[263]Katz, E., Willner, I. *Angew. Chem. Int. Ed.* **2004**, *43*, 6042–6108; Liu, X., Diao, H., Nishi, N. *Chem. Soc. Rev.* **2008**, *37*, 2745–2757.

[264]Wengel, J. *Org. Biomol. Chem.* **2004**, *2*, 277–280; Yan, H. *Science* **2004**, *306*, 2048–2049; Condon, A. *Nat. Rev. Genetics* **2006**, *7*, 565–575; Feldkamp, U., Niemeyer, C. M. *Angew. Chem. Int. Ed.* **2006**, *45*, 1856–1876.

[265]Gellman, S. H. *Acc. Chem. Res.* **1998**, *31*, 173–180; Hill, D. J., Mio, M. J., Prince, R. B., Hughes, T. S., Moore, J. S. *Chem. Rev.* **2001**, *101*, 3893 – 4011.

[266]Waring, M. J. *Ann. Rev. Biochem.* **1981**, *50*, 159–192.

[267]Krugh, T. R. *Curr. Op. Struct. Biol.* **1994**, *4*, 351 – 364.

[268]Chaires, J. B. *Biopolymers* **1998**, *44*, 201 – 215; Chaires, J. B. *Curr. Op. Struct. Biol.* **1998**, *8*, 314 – 320; Lane, A. N., Jenkins, T. C. *Q. Rev. Biophys.* **2000**, *33*, 255 – 306; Haq, I. *Arch. Biochem. Biophys.* **2002**, *403*, 1 – 15.

[269]Ren, J. S., Chaires, J. B. *Biochemistry* **1999**, *38*, 16067 – 16075.

to characterize binding modes, such as electrochemical measurements[270] and single molecule atomic force techniques.[271]

The rational design of new active compounds and tailoring them to bind with specific motifs has become a major application of DNA research[272] and is not necessarily limited to double helices anymore.[273] A steadily rising awareness of cellular structure and disease processes continues to add more positions to the list of targeted components, one of the more prominent currently being RNA.[274]

Behind these obvious advances and the practical success, one may easily argue that our understanding of the molecular aspects of nucleic acids is still partial at best. This is especially clear when considering the non-local aspects of DNA and RNA strands in biochemical processes. An excellent example is the formation of secondary quadruplex structures in guanine-rich end regions of the genome. While they have been known to exist for over 40 years[275] as an *in vitro* artefact, G-quartets now attract much debate about their physical and *in vivo* properties,[276] and a number of solution studies have recently revealed new structural details[277] and data on energetic stability.[278]

Quadruplexed guanine-rich strands have also come into the spotlight as specific targets for small molecules, which can function as fluorescent dyes[279] or drugs that inhibit the action of telomerase.[280] One of the possible binding modes for these molecules is intercalation, and the first quadruplex binders were actually derived from duplex DNA intercalators[281]. Whether or not and under what circumstances particular ligands favor an intercalating location, external stacking or another mode is discussed intensively in the literature, for example in the case of porphyrin derivatives.[282]

---

[270]Rauf, S., Gooding, J. J., Akhtar, K., Ghauri, M. A., Rahman, M., Anwar, M. A., Khalid, A. M. *J. Pharm. Biomed. Analysis* **2005**, *37*, 205 – 217.

[271]Krautbauer, R., Pope, L. H., Schrader, T. E., Allen, S., Gaub, H. E. *FEBS Letters* **2002**, *510*, 154 – 158.

[272]Haq, I., Ladbury, J. *J. Mol. Recognit.* **2000**, *13*, 188 – 197.

[273]Jenkins, T. C. *Curr. Med. Chem.* **2000**, *7*, 99 – 115.

[274]Gallego, J., Varani, G. *Acc. Chem. Res.* **2001**, *34*, 836–843; Thomas, J. R., Hernenrother, P. J. *Chem. Rev.* **2007**, *108*, 1171–1224.

[275]Davis, J. T. *Angew. Chem. Int. Ed.* **2004**, *43*, 668–698.

[276]Arthanari, H., Bolton, P. H. *Chem. Biol.* **2001**, *8*, 221 – 230; Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., Neidle, S. *Nucl. Acids Res.* **2006**, *34*, 5402–5415; Huppert, J. L. *Chem. Soc. Rev.* **2008**, *37*, 1375–1384.

[277]Xu, Y., Noguchi, Y., Sugiyama, H. *Bioorg. Med. Chem.* **2006**, *14*, 5584–5591; Dai, J., Punchihewa, C., Ambrus, A., Chen, D., Jones, R. A., Yang, D. *Nucl. Acids Res.* **2007**, *35*, 2440–2450; Phan, A. T., Kuryavyi, V., Burge, S., Neidle, S., Patel, D. J. *J. Am. Chem. Soc.* **2007**, *129*, 4386–4392; Phan, A. T., Kuryavyi, V., Luu, K. N., Patel, D. J. *Nucl. Acids Res.* **2007**, *35*, 6517–6525.

[278]Lane, A. N., Chaires, J. B., Gray, R. D., Trent, J. O. *Nucl. Acids Res.* **2008**, *36*, 5482–5515.

[279]Arthanari, H., Basu, S., Kawano, T. L., Bolton, P. H. *Nucl. Acids Res.* **1998**, *26*, 3724–3728; Koeppel, F., Riou, J., Laoui, A., Mailliet, P., Arimondo, P. B., Labit, D., Petitgenet, O., Helene, C., Mergny, J. *Nucl. Acids Res.* **2001**, *29*, 1087–1096; Rosu, F., Pauw, E. D., Guittat, L., Alberti, P., Lacroix, L., Mailliet, P., Riou, J., Mergny, J. *Biochemistry* **2003**, *42*, 10361–10371.

[280]Oganesian, L., Bryan, T. M. *BioEssays* **2007**, *29*, 155–165; Patel, D. J., Phan, A. T., Kuryavyi, V. *Nucl. Acids Res.* **2007**, *35*, 7429–7455; Cian, A. D., Lacroix, L., Douarre, C., Temime-Smaali, N., Trentesaux, C., Riou, J., Mergny, J. *Biochimie* **2008**, *90*, 131–155; Ou, T., Lu, Y., Tan, J., Huang, Z., Wong, K., Gu, L. *ChemMedChem* **2008**, *3*, 690–713.

[281]Monchaud, D., Teulade-Fichou, M. *Org. Biomol. Chem.* **2008**, *6*, 627–636.

[282]Wei, C., Jia, G., Yuan, J., Feng, Z., Li, C. *Biochemistry* **2006**, *45*, 6681–6691; Parkinson, G. N., Ghosh, R., Neidle, S. *Biochemistry* **2007**, *46*, 2390–2397; Wei, C., Jia, G., Zhou, J., Han, G., Li, C. *Phys. Chem.*

The present chapter deals with intermolecular interactions between intercalators and their hosts, focusing on well studied intercalators in double-stranded nucleic acids. In line with the previous sections, interaction energies are analyzed for specific geometries, using methods based on first principles and highlighting the role electrostatics. In particular, we address the nature of interactions by comparing the different energy components for three popular cationic intercalators, for which crystallographic data is available – ethidium$^{(+1)}$-UA/AU[283], ethidium$^{(+1)}$-CG/GC[284], and proflavine$^{(+1)}$-CG/AU[285]. Discussion is focused on the Eth$^{(+1)}$-UA/AU complex, for which more extensive calculations were performed; a molecular representation of this intercalation site is shown in Fig. 4.3.

4.1.1 Historical review of intercalation research

Nearly half a century ago, Lerman successfully bridged certain drug-DNA complexes and aromatic π-π interactions by proposing a structural intercalation model.[286] Building on observations that the addition of acridine, proflavine, or acridine orange to DNA in solution results in a marked change in the viscosity and sedimentation coefficient, he deduced that these small molecules induce perturbations in the double helix at the intercalation site as illustrated in Fig. 4.1.

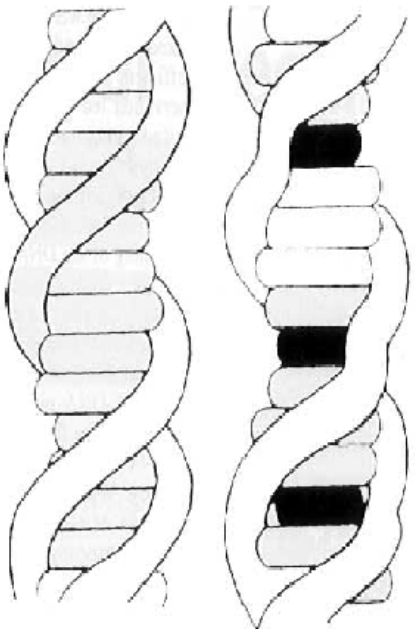


Figure 4.1: DNA intercalation model proposed by Lerman – the unperturbed double helix is shown on the left, and its intercalated counterpart on the right, with the intercalator represented by the dark areas.

Lerman described the net effect of DNA intercalation in terms of three changes:

- an increase in the separation between neighboring base pairs,
- elongation of the nucleic acid strand (hence the changes in sedimentation),
- local unwinding of the double helix.

While the amount of literature concerned with intercalated nucleic acids since Lerman's proposal has been overwhelming and shows no signs of decreasing, it can be roughly divided into five overlapping groups. Starting from the most application-oriented and ending at the most fundamental, the first type is directly related to biological activity. These studies of intercalating drugs and their practical applications in medicine deal first-hand with toxic effects, from the

Chem. Phys. **2009**, *11*, 4025–4032.

[283] Nucleic Acid Database ID: DRB018, Jain, S. C., Sobell, H. M. *J. Biomol. Struct. Dyn.* **1984**, *1*, 1161–1177.

[284] Nucleic Acid Database ID: DRB006, Jain, S. C., Sobell, H. M. *J. Biomol. Struct. Dyn.* **1984**, *1*, 1179–1194.

[285] Nucleic Acid Database ID: DRD004, Aggarwal, A., Islam, S. A., Kuroda, R., Neidle, S. *Biopolymers* **1984**, *23*, 1025–1041.

[286] Lerman, L. S. *J. Mol. Biol.* **1961**, *3*, 18–&.

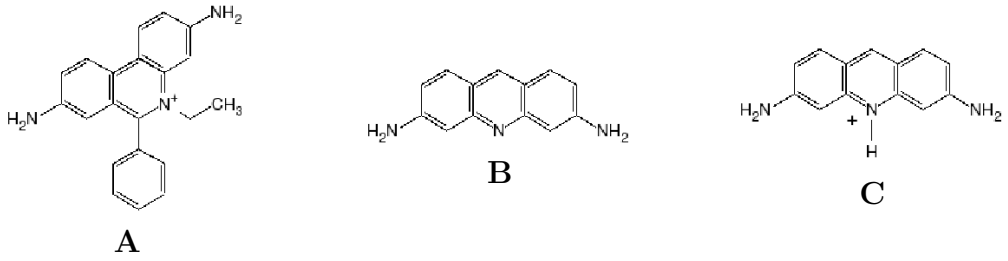


Figure 4.2: Structural formulas of the intercalators studied: ethidium bromide (A), neutral proflavine (B) and cationic proflavine (C).

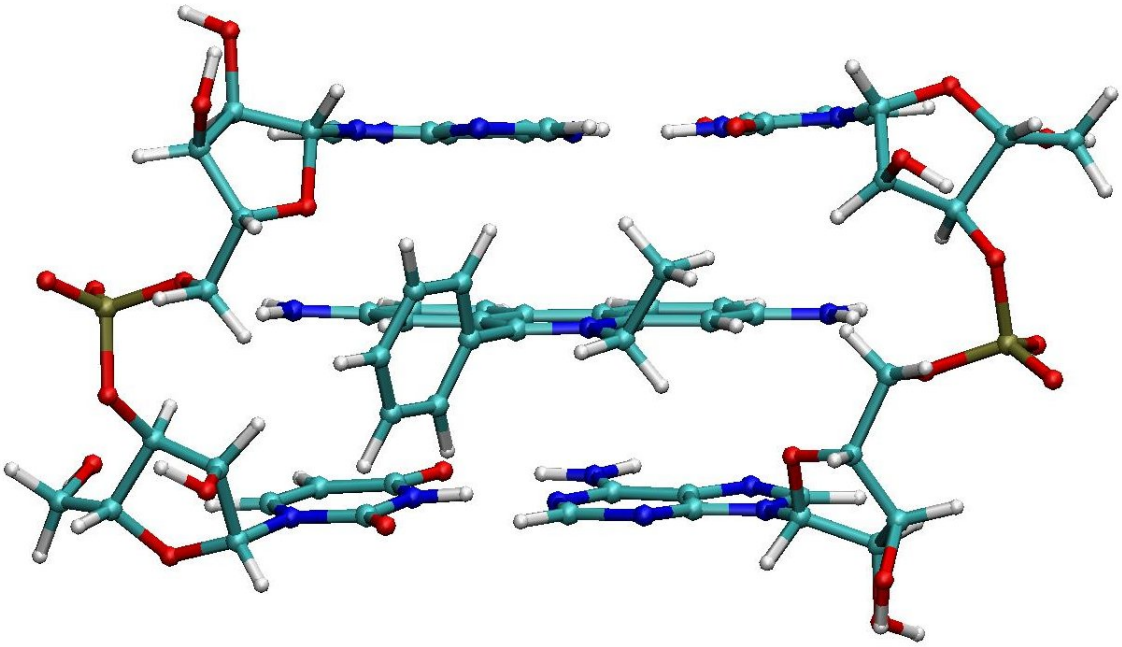


Figure 4.3: Minimal model for the ethidium cation intercalated in RNA between AU/UA base pairs, as used in this study. The geometry was obtained from crystallographic data, and the positions of the added hydrogen atoms optimized as described in the text.

earliest experiments on bacteria growth[287] to current efforts of assessing the inhibition of cell proliferation.[288]

The second, largest group of studies is different in that it considers the intercalation process in solution outside its biological context and focuses on the kinetic and overall thermodynamic characteristics and physicochemical aspects. This approach is represented by a number of important studies following Lerman's seminal experiment and continues to be used extensively in the current literature.

Waring was the first in 1965 to use UV spectroscopy to characterize the interaction of ethidium bromide with DNA as a function of the ionic strength of the solution.[289] In 1967, LePecq and Paoletti showed that the fluorescence of intercalated ethidium is more than 20

[287] Foster, R. A. C. *J. Bacteriol.* **1948**, *56*, 795–809.

[288] Ferguson, L. R., Denny, W. A. *Mutation Research* **2007**, *623*, 14–23; Janovec, L., Sabolová, D., Kožurková, M., Paulíková, H., Kristian, P., Ungvarský, J., Moravčiková, E., Bajdichová, M., Podhradský, D., Imrich, J. *Bioconjugate Chem.* **2007**, *18*, 93–100; Kožurková, M., Sabolová, D., Janovec, L., Mikeš, J., Koval, J., Ungvarský, J., Štefanišinová, M., Fedoročko, P., Kristian, P., Imrich, J. *Bioorg. Med. Chem.* **2008**, *16*, 3976–3984.

[289] Waring, M. J. *J. Mol. Biol.* **1965**, *13*, 269–&.

times stronger than that of the unbound molecule[290] and paved the way to using ethidium bromide as an effective fluorescent marker in various situations,[291] which most prominently has become a standard and accurate tool for conducting electrophoresis assays[292].

Already the case of ethidium appeared to be nontrivial, however, as more detailed experiments showed that its binding exhibits significant sequence specificity.[293] High affinity binding to unique sequences was also later demonstrated and tuned for intercalators based on actinomycin[294], proflavine[295] and other molecules.[296]

A natural further development has been the capability to distinguish intercalation from other binding modes, and consequently a number of polycyclic molecules have been characterized as intercalators.[297] Kinetics, structural changes and other data are now routinely collected using techniques that include spectrophotometry,[298] various other spectroscopies[299], electrochemistry[300] and interferometry.[301] This kind of research has unveiled interesting possibilities, for example the photoinduced switching of an inactive molecule into an intercalator proposed by Starcevic et al.[302]

A related, third branch of experimental research in this field focuses on the energetic requirements and thermodynamic effects of intercalation reactions[303], and emerged from the initial calorimetric differentiation of enthalpic and entropic contributions.[304] Further analytical measurements from physics and biochemistry have revealed much about other aspects

[290]Lepecq, J. B., Paoletti, C. *J. Mol. Biol.* **1967**, *27*, 87–&.

[291]Sharp, P. A., Sugden, B., Sambrook, J. *Biochemistry* **1973**, *12*, 3055–3063; Lai, J.-S., Herr, W. *Proc. Natl. Acad. Sci.* **1992**, *89*, 6958–6962; Chen, W., Turro, N. J., Tomalia, D. A. *Langmuir* **2000**, *16*, 15–19.

[292]Guttman, A., Cooke, N. *Anal. Chem.* **1991**, *63*, 2038–2042.

[293]Krugh, T. R., Reinhardt, C. G. *J. Mol. Biol.* **1975**, *97*, 133–162.

[294]Snyder, J. G., Hartman, N. G., D'Estantoit, B. L., Kennard, O., Remeta, D. P., Breslauer, K. J. *Proc. Natl. Acad. Sci.* **1989**, *86*, 3968–3972; Biver, T., Venturini, M., Jares-Erijman, E. A., Jovin, T. M., Secco, F. *Biochemistry* **2009**, *48*, 173–179.

[295]Bailly, C., Laine, W., Demeunynck, M., Lhomme, J. *Biochem. Biophys. Res. Commun.* **2000**, *273*, 681–685; Baldeyrou, B., Tardy, C., Bailly, C., Colson, P., Houssier, C., Charmantray, F., Demeunynck, M. *Eur. J. Med. Chem.* **2002**, *37*, 315–322.

[296]Nakatani, K., Matsuno, T., Adachi, K., Hagihara, S., Saito, I. *J. Am. Chem. Soc.* **2001**, *123*, 5695–5702; Nakatani, K., Horie, S., Murase, T., Hagihara, S., Saito, I. *Bioorg. Med. Chem.* **2003**, *11*, 2347–2353; Stojković, M. R., Piantanida, I. *Tetrahedron* **2008**, *64*, 7807–7814.

[297]Sartorius, J., Schneider, H.-J. *J. Chem. Soc. Perkin Trans. 2* **1997**, 2319–2327; Ismail, M. A., Sanders, K. J., Fennell, G. C., Latham, H. C., Wormell, P., Rodger, A. *Biopolymers* **1998**, *46*, 127–143.

[298]Ciatto, C., D'Amico, M. L., Natile, G., Secco, F., Venturini, M. *Biophys. J.* **1999**, *77*, 2717–2724; Vardevanyan, P. O., Antonyan, A. P., Manukyan, G. A., Karapetyan, A. T. *Exp. Mol. Med.* **2001**, *33*, 205–208; Vardevanyan, P. O., Antonyan, A. P., Parsadanyan, M. A., Davtyan, H. G., Karapetyan, A. T. *Exp. Mol. Med.* **2003**, *35*, 527–533; Alonso, A., Almendral, M. J., Curto, Y., Criado, J. J., Rodríguez, E., Manzano, J. L. *Anal. Biochem.* **2006**, *355*, 157–164.

[299]Hecht, C., Friedrich, J., Chang, T.-C. *J. Phys. Chem. B* **2004**, *108*, 10241–10244; Chang, T.-C., Yang, Y.-P., Huang, K.-H., Chang, C.-C., Hecht, C. *Optics & Spectroscopy* **2005**, *98*, 655–660; Benevides, J. M., Thomas, G. J. *Biochemistry* **2005**, *44*, 2993–2999.

[300]Tang, T., Huang, H. *Electroanalysis* **1999**, *11*, 1185–1190; Aslanoglu, M. *Anal. Sci.* **2006**, *22*, 439–443; Nowicka, A. M., Zabost, E., Klim, B., Mazerska, Z., Stojek, Z. *Electroanalysis* **2009**, *21*, 52–60.

[301]Wang, J., Xu, X., Zhang, Z., Yang, F., Yang, X. *Anal. Chem.* **2009**, *81*, 4914–4921.

[302]Starcevic, K., Karminski-Zamola, G., Piantanida, I., Zinic, M., Suman, L., Kralji, M. *J. Am. Chem. Soc.* **2005**, *127*, 1074–1075.

[303]Graves, D. E., Velea, L. M. *Curr. Org. Chem.* **2000**, *4*, 915–929.

[304]Breslauer, K. J., Remeta, D. P., Chou, W.-Y., Ferrante, R., Curry, J., Zaunczkowski, D., Snyder, J. G., Marky, L. A. *Proc. Natl. Acad. Sci.* **1987**, *84*, 8922–8926.

of intercalation energetics, such as electrostatic electrolyte contributions[305] and heat capacity changes.[306] Dealing also with the environmental conditions necessary for intercalation, we learn from these studies that the formation of intercalation complexes is dictated by a delicate balance of free energy contributions from the intercalating agent, nucleic acid strand and solution.

Conceptual approaches have been published that compare numerical calculations directly to such experimental results. For example, Kostjukov et al. use a combination of methodologies to determine the electrostatic[307] contribution to intercalation, and more recently the overall profile of the Gibb's free energy[308]. Rocha on the other hand describes a simple model of the mechanical properties of intercalated DNA to reproduce spectroscopic and optical tweezer experiments.[309]

Significant amounts of structural data for intercalation complexes, which provide direct information about the final complex, form another body of literature. This branch has been well represented by X-ray crystallographic reports (for example the geometries of Eth^+-AU/UA[283], Eth^+-CG/GC[284] and Pf-CG/AU[285] used in this work), but also includes solution structures[310] and other forms of binding such as to the minor groove[311].

Lastly, a growing body of research detaches the final intercalation complex from its surroundings and describes it in terms of molecular properties and intermolecular interactions. These studies utilize largely theoretical approaches and often use existing structural data as a starting point. An important experimental counterpart to these various computational studies are single molecule observations of structural perturbations provided by atomic force microscopy. Williams and coworkers have monitored DNA tertiary structure changes with atomic force microscopy (AFM)[312], and other similar reports have followed.[313]

Initially, most computational efforts were limited to analyzing conformational changes, such as the early considerations of flexibility by Berman et al.[314] Numerical methods based on classical electrostatic have also been applied, such as the Poisson-Boltzman approach adapted by Honig and coworkers.[315] Empirical-based methods can also yield useful results – Cashman and Kellogg tune intercalating molecules so they additionally bind to the major or minor

[305]Chaires, J. B., Priebe, W., Graves, D. E., Burke, T. G. *J. Am. Chem. Soc.* **1993**, *115*, 5360–5364; Chaires, J. B., Satyanarayana, S., Suh, D., Fokt, I., Przewloka, T., Priebe, W. *Biochemistry* **1996**, *35*, 2047–2053.

[306]Ren, J. S., Jenkins, T. C., Chaires, J. B. *Biochemistry* **2000**, *39*, 8439–8447.

[307]Kostjukov, V. V., Khomytova, N. M., Davies, D. B., Evstigneev, M. P. *Biopolymers* **2008**, *89*, 680–690.

[308]Kostjukov, V. V., Khomytova, N. M., Evstigneev, M. P. *Biopolymers* **2009**, *91*, 773–790.

[309]Rocha, M. *Phys. Biol.* **2009**, *6*, 036013.

[310]Horowitz, E. D., Lilavivat, S., Holladay, B. W., Germann, M. W., Hud, N. V. *J. Am. Chem. Soc.* **2009**, *131*, 5831–5838.

[311]Neidle, S. *Biopolymers* **1997**, *44*, 105 – 121.

[312]Pope, L. H., Davies, M. C., Laughton, C. A., Roberts, C. J., Tendler, S. J. B., Williams, P. M. *Anal. Chim. Acta* **1999**, *400*, 27–32; Pope, L. H., Davies, M. C., Laughton, C. A., Roberts, C. J., Tendler, S. J. B., Williams, P. M. *J. Microsc.* **2000**, *199*, 68–78.

[313]Berge, T., Jenkins, N. S., Hopkirk, R. B., Waring, M. J., Edwardson, J. M., Henderson, R. M. *Nucl. Acids Res.* **2002**, *30*, 2980–2986; Eckel, R., Ros, R., Ros, A., Wilking, S. D., Sewald, N., Anselmetti, D. *Biophys. J.* **2003**, *85*, 1968–1973; Vladescu, I. D., McCauley, M. J., Rouzina, I., Williams, M. C. *Phys. Rev. Lett.* **2005**, *95*, 158102.

[314]Berman, H. M., Neidle, S., Stodola, R. K. *Proc. Natl. Acad. Sci.* **1978**, *75*, 828–832.

[315]Misra, V. K., Honig, B. *Proc. Natl. Acad. Sci.* **1995**, *92*, 4691–4695.

groove at specific base sequences[316], and Ricci and Netz recently demonstrated that docking protocols can identify the intercalative binding mode of DNA ligands.[317].

By far the most popular methods for conceptually studying nucleic acid intercalation, however, have been molecular dynamics and quantum chemistry. In an interesting early study for example, Elcock et al. manage to reproduce the crystallographic twist angles of DNA intercalated by ellipticine with extensive MD simulations[318]. Trieb et al. provide an interesting analysis of the cooperativity of intercalation events in duplex B-DNA.[319] In a more recent study, Mukherjee et al. study the dynamics of daunomycin in the vicinity of a twelve base pair DNA fragment.[320]

It has only been in the last ten years that *ab initio* calculations for systems representing intercalation sites have been possible. Probably the first to venture in this direction were Bondarev et al., who reported MP2/6-31++G(d,p) interaction energies for the intercalator amiloride with the four DNA bases and compared them to empirical results[321]. Soon afterwords Reha et al. followed with a similar, more extensive investigation of four different intercalating molecules, among them being ethidium[322].

The main conclusion of these first reports was that in such electronic structure calculations it is indispensable to include dispersion (denoted here as $\Delta E^{(2)}_{\mathrm{disp}}$), as it constitutes the largest part of the interaction energy. Other reports of MP2 calculations have followed, including that of Dračinsky and Castano, who study the interaction energy for various distances and twists between ellipticine and base pairs and compare them to force field results.[323] Single point energies were published by Xiao and Cushman for camptothecin in an attempt to correlate them with experimental site selectivity in ternary cleavage complexes.[324]

Due to problems with the proper representation of dispersion interactions, density functional theory has been adopted relatively late for exploring the behavior of intercalators[325]. Recently, Car-Parinello dynamics and time-dependent DFT have also been employed by Fantacci et al. in order to characterize the base pair influence on the excited states of an intercalated ruthenium compound.[326]

Across all these branches of research, the most prominent unifying concept has been that of sequence specificity – where an intercalator binds preferentially to a site depending on its nucleic acid base content. Ironically the last, theoretical branch mentioned above is often the

[316]Cashman, D. J., Kellogg, G. E. *J. Med. Chem.* **2004**, *47*, 1360–1374.
[317]Ricci, C. G., Netz, P. A. *J. Chem. Inf. Model.* **2009**, *49*, 1925–1935.
[318]Elcock, A. H., Rodger, A., Richards, W. G. *Biopolymers* **1996**, *39*, 309–326.
[319]Trieb, M., Rauch, C., Wibowo, F. R., Wellenzohn, B., Liedl, K. R. *Nucl. Acids Res.* **2004**, *32*, 4696–4703.
[320]Mukherjee, A., Lavery, R., Bagchi, B., Hynes, J. T. *J. Am. Chem. Soc.* **2008**, *130*, 9747–9755.
[321]Bondarev, D. A., Skawinski, W. J., Venanzi, C. A. *J. Phys. Chem. B* **2000**, *104*, 815–822.
[322]Řeha, D., Kabeláč, M., Ryjáček, F., Šponer, J., Šponer, J. E., Elstner, M., Suhai, S., Hobza, P. *J. Am. Chem. Soc.* **2002**, *124*, 3366–3376.
[323]Dračinský, M., Castaño, O. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1799–1805.
[324]Xiao, X., Cushman, M. *J. Am. Chem. Soc.* **2005**, *127*, 9960–9961.
[325]Tuttle, T., Kraka, E., Cremer, D. *J. Am. Chem. Soc.* **2005**, *127*, 9469–9484; Barone, G., Guerra, C. F., Gambino, N., Silvestri, A., Lauria, A., Almerico, A. M., Bickelhaupt, F. M. *J. Biomol. Struct. Dyn.* **2008**, *26*, 115–129.
[326]Fantacci, S., Angelis, F. D., Sgamellotti, A., Marrone, A., Re, N. *J. Am. Chem. Soc.* **2005**, *127*, 14144–14145.

starting point for the ultimate goal: rationally designing new, sequence-specific drugs that target a given nucleic acid domain. At the same time, it is the most limited and gives the least information on the final efficacy of a drug molecule.

It is not surprising, therefore, that critical links between the physical and chemical properties of these complexes and their biological effectiveness remain unclear[303]. Several reasons can be formulated. First of all, there are several types of interactions associated with ligands binding to DNA – these include intercalation, non-covalent groove binding, covalent binding with cross linking, cleavage and nucleoside-analog incorporation. All these binding interactions involve changes to both the DNA and drug molecules in order to accommodate complex formation, and all of them can influence DNA function. Further, a single type of ligand may bind in several ways simultaneously, depending on the surrounding conditions.

More importantly, DNA binding affinity in general does not directly correlate with biological activity. This is usually due to the fact that the desired chemotherapeutic effects involve additional components and binding patterns in ternary complexes of the ligand and DNA. A prime example of this issue is the classic case of *m*-AMSA and its structural conformer *o*-AMSA.[327] Both ligands bind to DNA by intercalation – in fact, the binding affinity of *o*-AMSA is about 4 times higher.[328] In contrast, *m*-AMSA stimulates single and double-strand topoisomerase II mediated DNA cleavage,[329] while *o*-AMSA is ineffective in eliciting such an affect.[330] Wadkins and Graves point out that several intermolecular interactions must be considered in this case,

Even when intercalation binding is considered as a separate process in its own right, another problem is encountered when analyzing it, namely the characterization of thermodynamic mechanisms associated with complex formation and therefore kinetics. As already mentioned, thermodynamic studies by Breslauer et al.[304] have shown that conclusions based solely on measured values of the free energy ΔG_{obs} can be misleading. For instance, two complexes may have near identical binding free energies and entirely different thermodynamic profiles – driven mainly by enthalpy or entropy. The energetics of intercalation should therefore be described in terms of separate enthalpy (ΔH_{obs}) and entropy (ΔS_{obs}) changes:

$$\Delta G_{\text{obs}} = \Delta H_{\text{obs}} - T\Delta S_{\text{obs}}. \tag{4.1}$$

A number of separate physical driving forces, each with their own enthalpic and entropic contributions, have been identified for DNA intercalation.[331] The free energy changes induced by these forces are usually implicitly postulated to be independent and additive based on

[327] Cain, B. F., Atwell, G. J., Denny, W. A. *J. Med. Chem.* **1975**, *18*, 1110–1117; Wilson, W. R., Baguley, B. C., Wakelin, L. P. G., Waring, M. J. *Mol. Pharmacol.* **1981**, *20*, 404–414; Pommier, Y., Minford, J. K., Schwartz, R. E., Zwelling, L. A., Kohn, K. W. *Biochemistry* **1985**, *24*, 6410–6416.

[328] Wadkins, R. M., Graves, D. E. *Biochemistry* **1991**, *30*, 4277–4283.

[329] Minford, J. K., Pommier, Y., Filipski, J., Kohn, K. W., Kerrigan, D., Mattern, M., Michaels, S., Schwartz, R. E., Zwelling, L. A. *Biochemistry* **1986**, *25*, 9–16.

[330] Pommier, Y., Covey, J., Kerrigan, D., Mattes, W., Markovits, J., Kohn, K. W. *Biochem. Pharmacol.* **1987**, *36*, 3477–3486.

[331] See section 3.2 in Graves; Velea, 2000, in Ref. 303 on page 86.

general thermodynamic considerations of ligand binding[332],

$$\Delta G_{\text{obs}} = \Delta G_{\text{conf}} + \Delta G_{\text{r+t}} + \Delta G_{\text{pe}} + \Delta G_{\text{hyd}} + \Delta G_{\text{mol}}, \tag{4.2}$$

where the individual contributions are related to the following mechanisms,

ΔG_{conf} – unfavorable conformational changes needed to form the drug-DNA system, including those described by the perturbation model of Lerman (Fig. 4.1),

$\Delta G_{\text{r+t}}$ – loss of translational and rotational degrees of freedom; this is a purely entropic effect, meaning that $\Delta G_{\text{r+t}} = -T\Delta S_{\text{r+t}}$, where typical values are approximately 50 ± 10 entropic units or $+14,9(\pm 3,0)$ kcal/mol at room temperature,

ΔG_{hyd} – favorable entropy changes related to the disruption of the hydration shell around the intercalator,

ΔG_{pe} – polyelectrolyte effect, namely counterions being freed from DNA phosphate groups in the case of cationic intercalators; additional counterions are freed when the electron density along the nucleic acid strand delocalizes due to double helix elongation,

ΔG_{mol} – intermolecular interactions within the binding site, which accounts for local stability and drives the conformational changes that induce ΔG_{conf}.

These contributions are typically much larger than the total observed free energy of binding ΔG_{obs}, making their net thermodynamic values not always easy to identify or quantitate.

That the individual free energy terms are relatively large compared to their sum has also been established by theoretical considerations – Bagiński et al. have argued this point relying on a different kind of partitioning of the the free energy, based on electrostatic and non-electrostatic contributions.[333] They also find that while non-electrostatic interactions are the main driving force in the intercalation of anthracycline derivatives, the electrostatic contributions have greater potential for differentiating between similar intercalators. These results have been used in practice, for example Lesyng and coworkers apply additional assumptions about these free energy contributions and construct a computationally undemanding molecular mechanics model, which is still shown to correspond with experimental data.[334]

Similarly, the binding energy of the final complex ΔG_{mol}, which includes non-covalent molecular interactions, is not easily interpreted when the intercalation process is viewed as a kinetic equilibrium. Obviously, ΔG_{mol} must be favorable in order for the intercalation complex to form and be stable, nonetheless one may imagine a complex where the intermolecular forces would be large but binding is impossible at all relevant environmental conditions due to entropic factors or required unsurmountable conformational changes. Conversely, in situations where the remaining contributions suppress each other, local intermolecular interactions can be the deciding factor even if they are relatively small.

[332] Szwajkajzer, D., Carey, J. *Biopolymers* **1997**, *44*, 181–198.
[333] Bagiński, M., Fogolari, F., Briggs, J. M. *J. Mol. Biol.* **1997**, *274*, 253–267.
[334] Rudnicki, W. R., Kurzepa, M., Szczepanik, T., Priebe, W., Lesyng, B. *Acta Biochim. Pol.* **2000**, 1–9.

4.2 Interaction energy analyses for bound intercalators

It is to the local intermolecular interactions embodied in ΔG_{mol} that we turn to here – interactions which in line with Lerman's model are of the π-π stacking type akin to those studied in Section 3.3. Although there have been a number of theoretical studies on intercalation, works that probe the quantum chemical nature of these complexes are relatively scarce. As mentioned above, these calculations have been performed only in the last decade due to the size of the systems involved and the recognized importance of dispersion effects. Kubar et al. provide a fair discussion of the importance of dispersion and estimates for the free energy terms in (4.2)[335]; they also underline the effective importance of the intermolecular stabilization energy ΔG_{obs} due to the cancellation of other terms.

For most practical purposes, models of intercalation sites are still too large to be treated routinely as entire quantum systems. Therefore, studies have been typically limited to the nearest two or four nucleic acid bases, and interactions are often analyzed pair-wise between the intercalator and each base separately. It is the purpose of our study to confirm the validity of dividing intercalation sites into nucleobases and other fragments within the minimal model proposed by Kubar et al.[335], which consists of the intercalator, four nearest nucleosides and two phosphate groups between them.

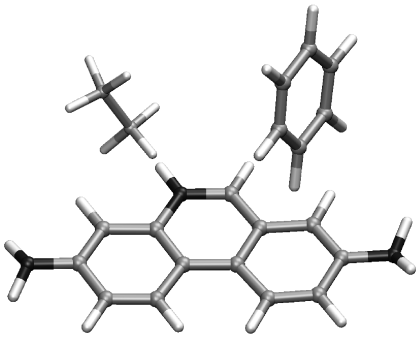


Figure 4.4: A molecular model of the ethidium cation, divided into hydrogen-capped fragments as used in parts of this work.

Starting from a crystallographic structure of ethidium intercalated in a AU/UA base pair step of RNA[284] (shown in Fig. 4.3), we examine models of various extent and evaluate the magnitude of many body interactions.

The models used, which comprise various parts of the intercalation site and divide these parts into smaller fragments in several different ways, are described by the colored schematic representations in Fig. 4.5. Interaction energies for these models, analyzed according to the decomposition scheme given by (2.23), are summarized in Table 4.1. Wherever the number of interacting dimers N_{int} was more than one, interactions were summed in a pair-wise fashion for the intercalator and each RNA fragment. For example, in the case of model **A**2 the total pair interaction energy consists of two parts that correspond to two base pairs:

$$\Delta E(\mathbf{A2}) = \Delta E(\mathrm{Eth..A/U}) + \Delta E(\mathrm{Eth..U/A}), \tag{4.3}$$

where each dimer is calculated in its own basis set.

Model names beginning with **A** were limited to the ethidium molecule and its four nearest base pairs. Model **AC** is a smaller version that additionally accounts for only the chromophore of ethidium, disregarding the side chain and ring. This corresponds to the system studied in

[335]Kubař, T., Hanus, M., Ryjáček, F., Hobza, P. *Chem. Eur. J.* **2006**, *12*, 280–290.

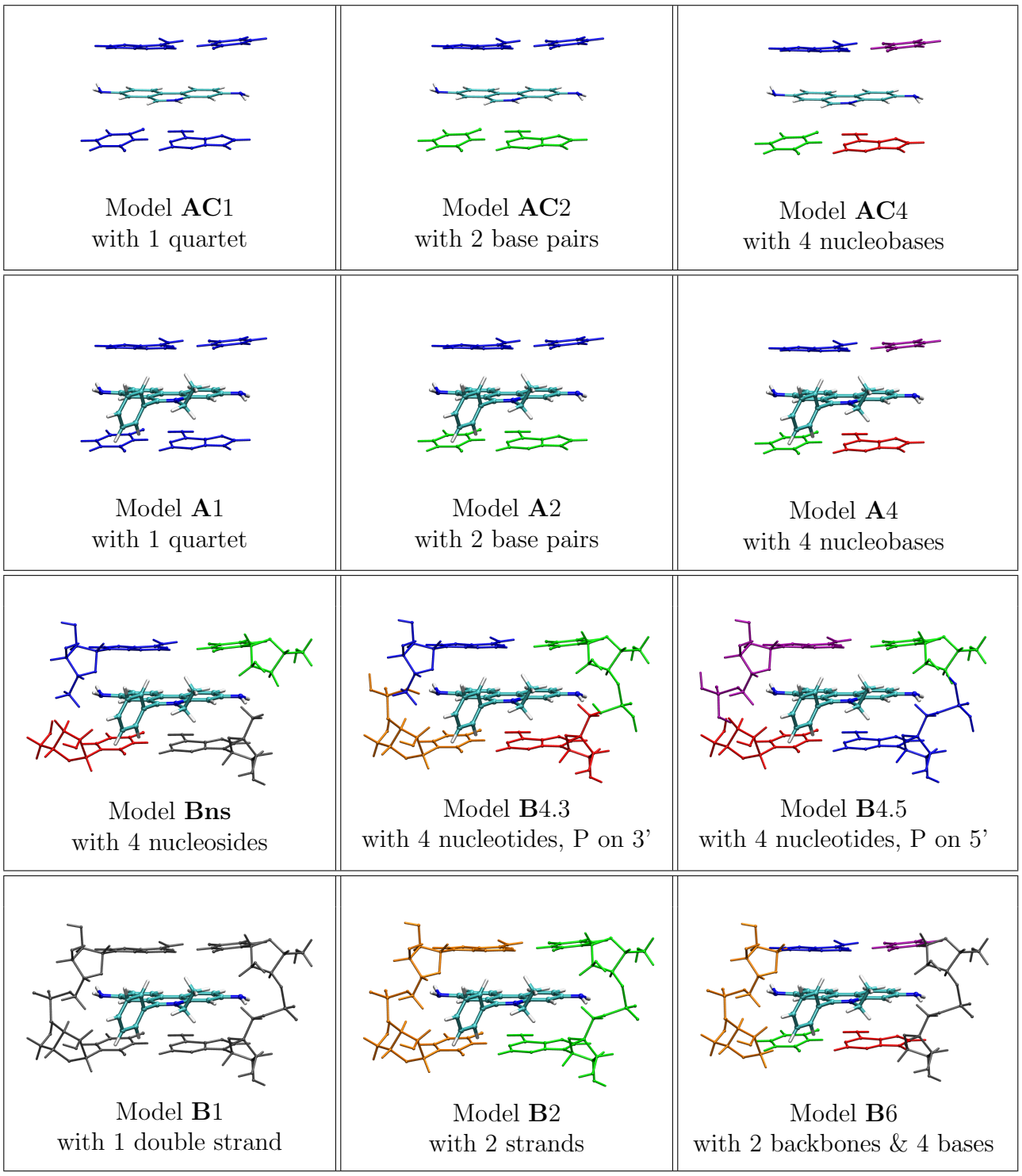


Figure 4.5: Schematic drawings of the models used for the ethidium - AU/UA base pair complex. Each color represents a piece of separately interacting RNA fragment around the binding site. The total interaction was constructed as the sum of all dimer interactions among these pieces with ethidium.

	N_{int}	N_{AO}	$\Delta E_{\mathrm{el}}^{(1)}$	$\Delta E_{\mathrm{ex}}^{(1)}$	$\Delta E_{\mathrm{del}}^{(\mathrm{R})}$	ΔE_{corr}	ΔE_{MP2}
				[kcal/mol]			[kcal/mol]
model **AC**1	1	878	-24.8	31.8	-4.8	-33.4	-31.3
model **AC**2	2	615	-25.3	32.0	-4.7	-33.5	-31.5
model **AC**4	4	0	-26.0	32.3	-4.7	-33.3	-31.7
model **A**1	1	1030	-28.6	40.2	-6.1	-39.7	-34.3
model **A**2	2	733	-28.9	40.2	-6.0	-39.4	-34.2
model **A**4	4	601	-29.8	40.7	-6.2	-39.2	-34.5
model **A**4 (parts)	12	0	-30.1	41.6	-6.4	-39.6	-34.5
model **Bns**	4	753	-35.8	51.7	-10.6	-51.9	-46.6
model **B**1 · neutral (H)	1	1776	-41.9	54.9	-10.9	-54.7	-52.6
model **B**2 · neutral (H)	2	1106	-43.0	55.4	-11.1	-54.4	-53.1
model **B**6 · neutral (H)	6	829	-43.5	56.2	-11.3	-54.4	-52.9
model **B**4.3 · neutral (H)	4	799	-43.6	56.7	-12.1	-54.7	-53.7
model **B**4.5 · neutral (H)	4	832	-40.3	46.2	-10.5	-49.4	-54.0
model **B**1 · neutral (Na)	1	1802	-50.7	55.1	-10.9	-55.0	-61.5
model **B**1 · neutral (K)	1	1830	-52.7	55.1	-10.9	-55.1	-63.6
model **B**1 · charged -2 (H_2O)	1	1814	-117.1	55.2	-11.0	-56.4	-129.3
model **B**1 · charged -2	1	1766	-121.1	55.2	-11.1	-56.5	-133.5
model **B**2 · charged -2	2	1101	-123.5	56.0	-13.9	-56.6	-138.0
model **B**6 · charged -2	6	824	-123.3	57.0	-16.0	-56.8	-139.2
model **B**4.3 · charged -2	4	794	-126.1	56.7	-15.2	-57.0	-141.7
model **B**4.5 · charged -2	4	827	-120.5	46.5	-15.5	-51.8	-141.2

Table 4.1: Components of the interaction energy of ethidium intercalated between AU/UA nucleic acid bases following (2.23); the symbols correspond to various interaction models as illustrated in 4.5. The variant labeled *parts* for model **A4** additionally has the ethidium molecule divided into three parts – its chromophore, side chain and ring[336] – making it a super set of model **AC**4. In the charge variant *neutral (H)* protons were attached to the anionic phosphate groups to simulate counterions, and Na^+ and K^+ were attached in the variants *neutral (Na)* and *neutral (K)*. For *charged -2* the RNA fragment was left charged, and in the case of *charged -2 (H_2O)* both phosphate groups were hydrated. The column N_{int} contains the number of pair-wise calculations comprising the interaction, and N_{AO} denotes the maximum number of atomic orbitals used in any pair-wise calculation in the model.

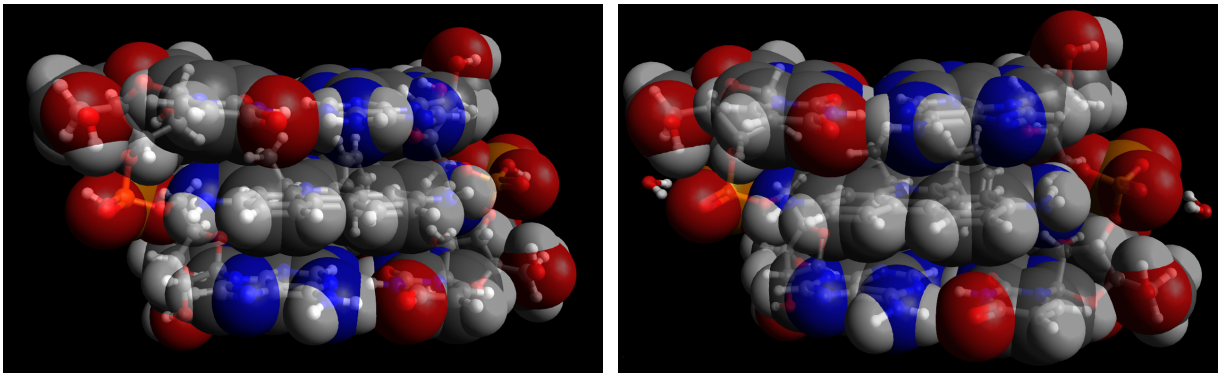
the previously published report[336], restricted to the ethidium chromophore and four nearest nucleobases.

Models designated with **B** include the sugars and phosphate groups connecting these base pairs. The intermediate model name **Bns** disregards the phosphate groups and represents the RNA fragment only by nucleosides. Additional numbers in the model names denote the number of fragments of RNA treated separately when summing interactions from dimer calculations – the different fragments are illustrated with different colors.

In order to estimate the possible magnitude of the influence of dynamic surroundings, the charged and neutralized variants of the largest model (**B**), by optimizing the position of a proton, sodium or potassium ion, or water molecule in the vicinity of the anionic phosphate group oxygen atoms.

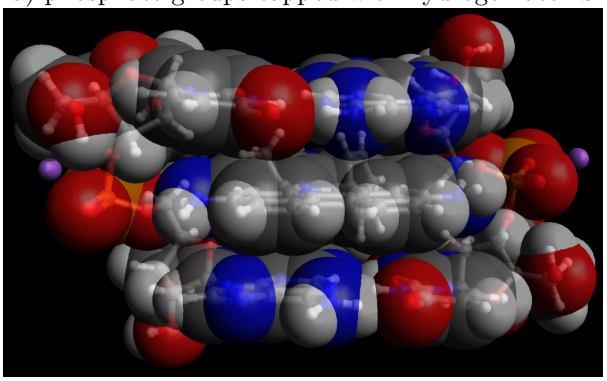
In all the models tested (**AC**, **A**, **B** and its variants), dividing the RNA fragment into nucleobases and backbone strands or into nucleotides is justified from the energetic point of

[336]Langner, K. M., Kędzierski, P., Sokalski, W. A., Leszczyński, J. *J. Phys. Chem. B* **2006**, *110*, 9720–9727.

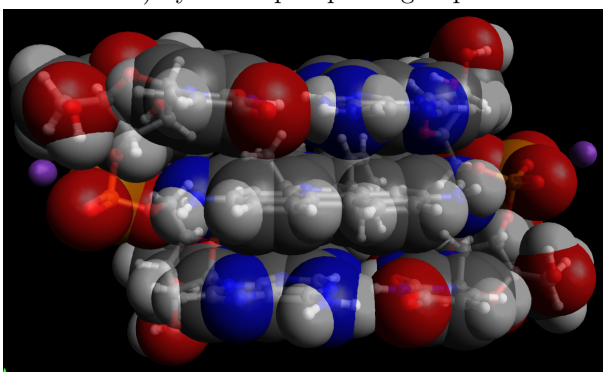


a) phosphate groups capped with hydrogen atoms

b) hydrated phosphate groups



c) two Na^+ counterions added



d) two K^+ counterions added

Figure 4.6: Comparison of the four ways the charged phosphate groups were compensated during calculations of the interactions energy between the nucleic acid strand and intercalated ethidium.

view. For the neutral models, dividing the system changes the total interaction by no more than 2%. In particular, for the smallest model **AC**, the difference between the sum of pair-wise interactions of the ethidium chromophore with each nucleobase (**AC**4) and its interaction calculated with all four nucleobases at once (**AC**1) was 0.4 kcal/mol. For the model including the ethidium side chain and ring (**A**), the analogous difference was only 0.2 kcal/mol. The error associated with approximating the interaction energy pair-wise was largest in the case of the charged model (**B** charged -2), namely up to 6%. This is understandable, since the delocalization of the additional charge is hindered by fragmenting the system.

Additionally, in the extended **B** models, two charge variants were considered. In the first, the anionic phosphate groups were compensated by added extra hydrogen atoms simulating counterions, labeled *neutral*. The other, labeled *charged -2*, left the phosphate groups charged. In the neutral variants, the anionic phosphate oxygen atoms were compensated by either protonation (H), adding one of two counterions (Na or K) or hydration (H_2O).

When considering these results in the context of *in vivo* nucleic acids, one should remember that they are dynamic systems in solution. Also, it has been shown before that the movement of counterions is diffusive around DNA,[337] therefore the type of compensation illustrated by our calculation in models **B** (Na) and **B** (K) is at best a temporary situation. At times when a counterion is not present near the phosphate groups, it is safe to argue that the intercalator will feel a stronger interaction. Meanwhile, in this case hydration effects, as we show, damped this

[337] Varnai, P., Zakrzewska, K. *Nucl. Acids Res.* **2004**, *32*, 4269–4280.

interaction somewhat, and in reality include more than the one water molecule we consider. In this light, our results can be viewed as identifying the range of intermolecular interaction possible in solution. Especially, models **B** neutral (H) and **B** charged -2 can be treated as approximate lower and upper limits of the interaction, respectively.

4.3 Alignment of ligands on the intercalation plane

In a previous study,[336] interaction energy components were also examined for other positions of ethidium intercalated in the Eth$^{(+1)}$-UA/AU complex. Using the nomenclature from the previous section, model **AC**4 was used, which means only the ethidium chromophore and its four nearest nucleic acid bases were considered. The chromophore position was varied in the plane of intercalation – including for example distances of ±0.5 Å and ±2.0 Å towards the major groove – and to a limited extent perpendicular to the intercalation plane.

The interaction energy at various levels of theory and its components at these points are presented in Fig. 4.7. The ΔE_{MP2} energy has a minimum very close to the crystallographic position of the chromophore (zero on the plots) - a parabola fit gives a diagonal offset (originating from two perpendicular directions in the plane of intercalation) of 0.09 Å, and an energy difference of about 0.05 kcal/mol. This supports the notion that factors other than local interactions - that is those between the intercalator's chromophore and its nearest bases - are of minor importance for chromophore alignment, and justifies the fragmentation method adopted for ethidium in this study.

The multipole electrostatic interaction energy was evaluated based on the CAMM moments defined in Section 2.5.2, between each chromophore and its four nearest bases. This energy is presented in the form of a surface plot as a function of the displacement of the chromophore geometric center from the crystal position – in Fig. 4.8 for the Eth$^{(+1)}$-UA/AU system.

Steric constraints caused by the DNA side chain are illustrated in Fig. 4.8 by the shaded region, in which the distance between any pair of atoms is smaller than the sum of their van der Waals radii, scaled by a factor of 0.5. Similar plots were evaluated in the intercalation planes of the other two studied cationic systems, Eth$^{(+1)}$-CG/GC and PF$^{(+1)}$-AU/CG (see Section 3 of Supporting Information in the published report[336]); all three surface plots have a significant central minimum. In contrast, the corresponding multipole interaction surface for the neutral proflavine complex is highly irregular and does not reproduce the crystal binding site in any reasonable way.

The deep central minimum in Fig. 4.8 has a value of -11.0 kcal/mol at the coordinates (-1.3 Å, 0.0 Å), 0.3 kcal/mol below that of the crystal position. This minimum is accompanied by a second one, at the coordinates (1.5 Å, 0.5 Å) with a value of -10.7 kcal/mol. Similar distances from the crystal position were obtained for the $\Delta E^{\kappa\leqslant 9}_{\mathrm{CAMM}}$ surfaces of the other two cationic intercalation complexes: 1.8 Å for Eth$^{(+1)}$-GC/CG and 1.7 Å for PF$^{(+1)}$-AU/GC, which demonstrates the accuracy of this approach in reproducing the alignment of chromophores between base pairs in these types of systems. In all cases, two central minima were located on opposite sides of the crystallographic positions.

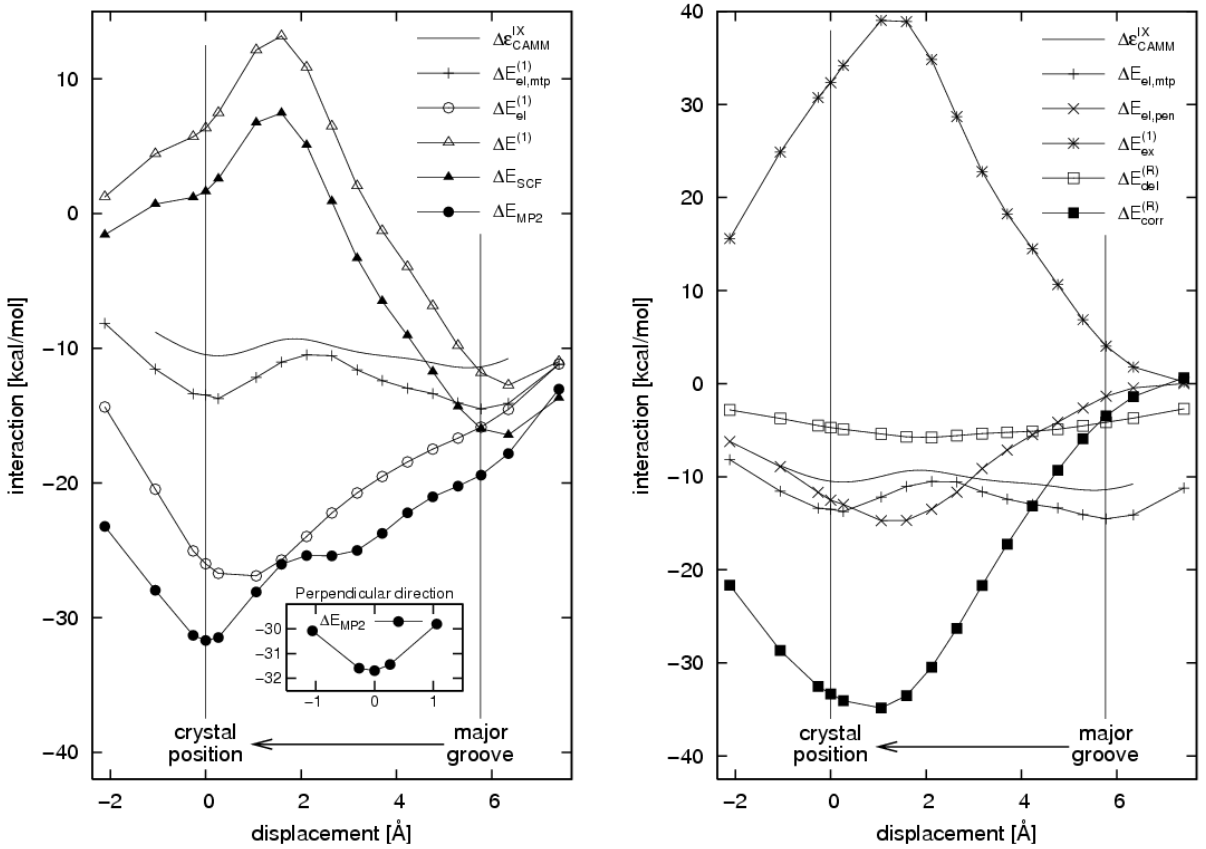


Figure 4.7: Interaction energy profile for the ethidium cation and four nearest nucleobases along a path from the crystal position to the major groove – individual interaction energy terms are shown on the right following (2.23), and the corresponding levels of theory on the left according to (2.28).

An additional minimum is also present in Fig. 4.8 in the direction of the major groove, at a distance of 5.7 Å from the crystal position at -11.5 kcal/mol). This minimum is observed for only the two complexes involving ethidium, and at all interaction ranks. If similar studies were to be performed for model systems without crystallographic data, it would be necessary to study all the obtained minima with a more exact method.

When an intercalator approaches the intercalation site from the major groove, it should experience an increase of electrostatic penetration and dispersion interactions. In order to further investigate the relevance of the off-center minimum in ΔE_{CAMM} for the total interaction energy in the intercalation plane, the components of ΔE_{MP2} were calculated for the ethidium chromophore located along a path connecting the crystallographic binding site and this minimum in the Eth$^{(+1)}$-UA/AU structure, up to 8 Å away. This path simulates, in a very simplified way, the movement of the chromophore when entering the intercalation site (see section 5 in Supporting Information for an animation). Fig. 4.7 presents the components calculated at different levels of theory along this path. The most evident conclusion is that the profiles of the exchange ($\Delta E_{\mathrm{ex}}^{(1)}$) and correlation ($\Delta E_{\mathrm{corr}}$) terms have similar shapes and opposite signs. These two components, along with the electrostatic penetration term, cancel each other out at distances above 5 Å. It is interesting to note that the first order interaction ($\Delta E_{\mathrm{HL}}^{(1)}$) and SCF interaction energy (ΔE_{RHF}) completely fail to reproduce the crystallographic binding site even along this one-dimensional path.

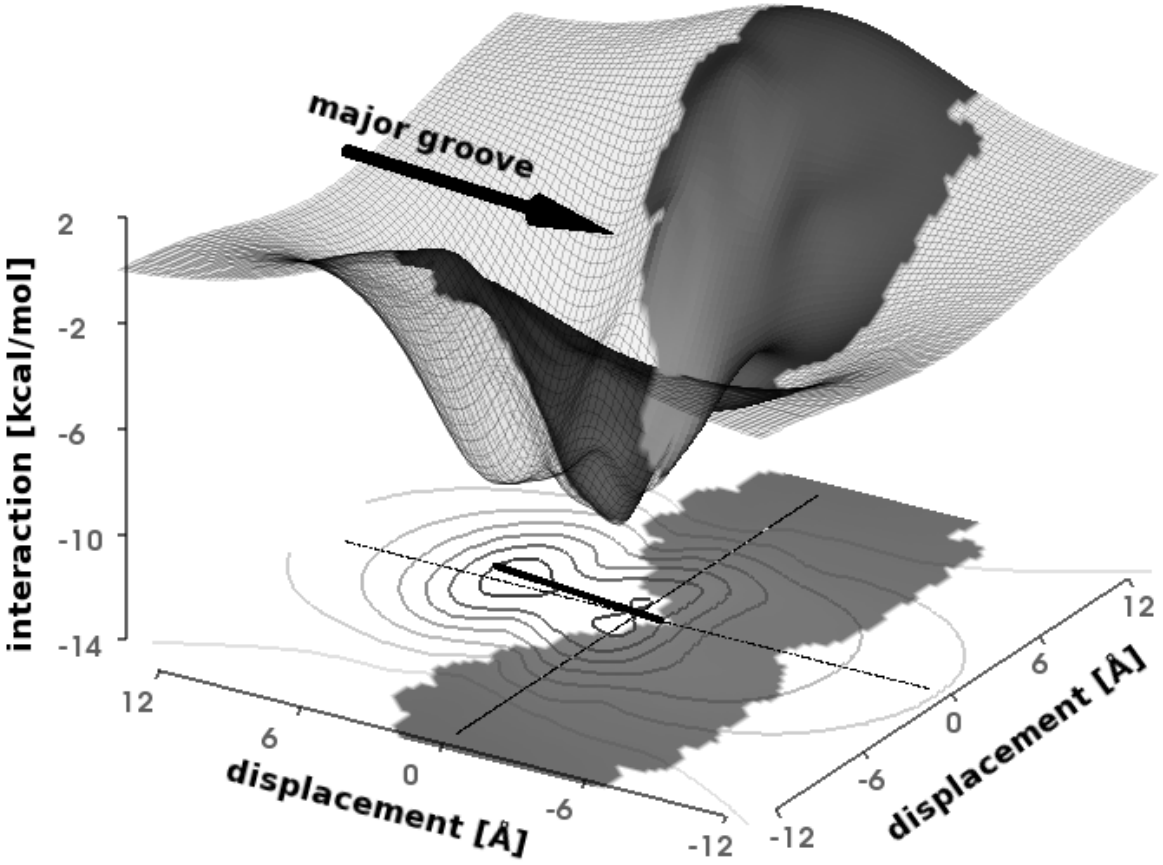


Figure 4.8: Map of the multipole electrostatic interaction between ethidium and the four nearest nucleobases.

Furthermore, besides the full MP2 interaction energy, the multipole components (both the one calculated from CAMM multipoles, ΔE_{CAMM}, and from DMA multipoles, $\Delta E_{\mathrm{el,mtp}}$), best reproduce the crystallographic binding site along this path, more precisely than the entire electrostatic interaction energy $\Delta E_{\mathrm{el}}^{(1)}$ (which includes penetration effects) or the correlation component ΔE_{corr}. On the other hand, the magnitude of the multipole component is significantly smaller that that of ΔE_{MP2}, but becomes more dominant as the chromophore is withdrawn from the intercalation site. Although the MP2 interaction energy does not have a second minimum where $\Delta E_{\mathrm{CAMM}}^{9}$ does, it is not monotonic along the studied path, and exhibits a plateau at around 2.6 Å from the binding site.

4.4 Convergence of multipole electrostatic interactions

In light of the discussions in Section 2.5.3 about the convergence of multipole electrostatic interaction energies based on atomic moments, it is worthwhile to return to this issue in the context of the intercalation complexes studied.

Fig. 4.9 shows a plot similar to the on in Fig. 2.4 for the cytosine-guanine dimer. In this case, the convergence is followed for the three studied intercalation complexes, using CAMM expansions representing the chromophores and base pairs, up to rank nine. At rank nine – $\Delta E_{\mathrm{CAMM}}^{\kappa\leqslant 9}$ – the interaction is converged, but does so only for $\kappa > 5$. The lack of convergence

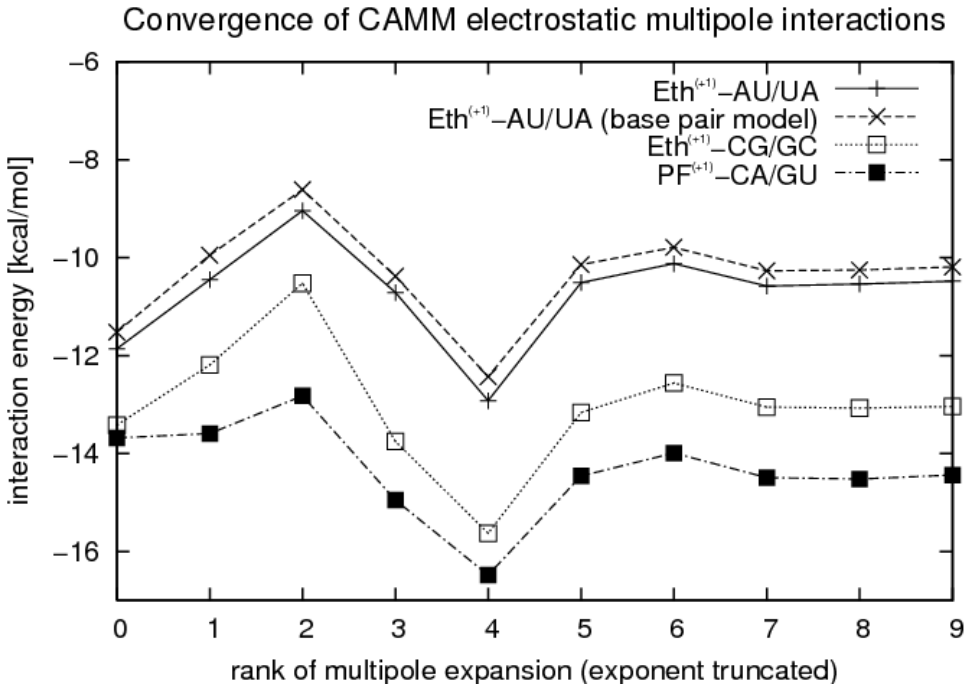


Figure 4.9: Convergence of the multipole approximation to the electrostatic interaction energy $\Delta E^{\mathrm{n}}_{\mathrm{CAMM}}$, where n is the maximum rank.

for lower ranks is due to small intermolecular distances (the normal distance between intercalator and bases is 3.4 Å) and multiple contacts between atoms, characteristic of stacking complexes. It is worthwhile to note that the interactions of the corresponding molecular multipole expansions are divergent (see Section 2 of the published Supporting Information[336]), which is expected considering the previous failures of molecular expansions for relatively simpler systems. The atomic-based multipole interaction energies, on the other hand, exhibit a similar trend already for ranks above 2, which makes them useful in comparative studies, such as dealing with the effect of base pair sequence.

The slow convergence of the electrostatic multipole term is relevant in the context of penetration effects. Since the magnitude of these effects is calculated as the difference between the total electrostatic interaction and its multipole component as in (2.44), it will be inaccurate if the multipole expansion used does not provide a converged value.

Previous studies using atomic expansions have rarely proceeded beyond octupole moments. For instance, Toczyłowski and Cybulski published an exhaustive study on the electrostatics of hydrogen-bonded and stacked DNA bases,[338] in which DMA was expanded through hexadecapoles. The present results suggest that this does not suffice, and that multipole interactions start to converge only above rank five (Fig.4.9). Using DMA moments up to rank 4 results in an overestimate, which leads to an incorrect estimate of the penetration term.

Another interesting observation for Fig.4.9 is that treating base pairs as a whole, as suggested by Řeha et al.[322], does not significantly change the slow convergent trend in the $\mathrm{Eth}^{(+1)}$-AU/UA complex. While there is a constant difference of ~ 0.4 kcal/mol in this case between the base pair (x) and single base (+) models used, the convergence trends behave similarly.

[338]Toczyłowski, R. R., Cybulski, S. M. *J. Chem. Phys.* **2005**, *123*, 154312–12.

	$Eth^{(+1)}$-UA/AU		$Eth^{(+1)}$-GC/CG		$PF^{(+1)}$-AU/CG		$PF^{(0)}$-AU/CG	
	100x100	33x33	100x100	33x33	100x100	33x33	100x100	33x33
Maximal change [kcal/mol]								
$\Delta E^{\kappa\leqslant 0}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 1}_{\mathrm{CAMM}}$	7.20	7.20	6.80	6.80	5.33	4.08	6.00	6.00
$\Delta E^{\kappa\leqslant 1}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 2}_{\mathrm{CAMM}}$	5.46	5.26	5.60	5.60	7.36	7.36	8.38	8.38
$\Delta E^{\kappa\leqslant 2}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 3}_{\mathrm{CAMM}}$	4.83	4.83	4.20	4.20	5.80	5.80	4.89	4.89
$\Delta E^{\kappa\leqslant 3}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 4}_{\mathrm{CAMM}}$	2.27	2.27	1.94	1.94	1.79	1.79	1.99	1.99
$\Delta E^{\kappa\leqslant 4}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 5}_{\mathrm{CAMM}}$	2.44	2.44	2.48	2.48	2.10	2.10	5.81	5.81
$\Delta E^{\kappa\leqslant 5}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 6}_{\mathrm{CAMM}}$	0.62	0.62	0.70	0.70	0.74	0.74	2.01	2.01
$\Delta E^{\kappa\leqslant 6}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 7}_{\mathrm{CAMM}}$	0.52	0.52	0.53	0.53	0.57	0.57	0.98	0.98
$\Delta E^{\kappa\leqslant 7}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 8}_{\mathrm{CAMM}}$	0.43	0.43	0.33	0.33	0.29	0.29	0.47	0.47
$\Delta E^{\kappa\leqslant 8}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 9}_{\mathrm{CAMM}}$	0.16	0.16	0.21	0.21	0.26	0.26	0.41	0.41
Average change [kcal/mol]								
$\Delta E^{\kappa\leqslant 0}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 1}_{\mathrm{CAMM}}$	1.51	3.44	1.50	2.67	1.47	1.90	0.72	2.58
$\Delta E^{\kappa\leqslant 1}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 2}_{\mathrm{CAMM}}$	0.78	2.36	0.84	2.48	0.97	2.69	0.75	3.24
$\Delta E^{\kappa\leqslant 2}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 3}_{\mathrm{CAMM}}$	0.37	1.41	0.36	1.49	0.47	2.20	0.39	1.57
$\Delta E^{\kappa\leqslant 3}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 4}_{\mathrm{CAMM}}$	0.23	0.84	0.21	0.66	0.21	0.77	0.23	0.87
$\Delta E^{\kappa\leqslant 4}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 5}_{\mathrm{CAMM}}$	0.17	0.81	0.14	0.68	0.14	0.72	0.29	1.03
$\Delta E^{\kappa\leqslant 5}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 6}_{\mathrm{CAMM}}$	0.05	0.21	0.05	0.19	0.05	0.19	0.21	0.55
$\Delta E^{\kappa\leqslant 6}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 7}_{\mathrm{CAMM}}$	0.03	0.14	0.03	0.13	0.03	0.12	0.13	0.20
$\Delta E^{\kappa\leqslant 7}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 8}_{\mathrm{CAMM}}$	0.01	0.06	0.01	0.05	0.01	0.06	0.05	0.13
$\Delta E^{\kappa\leqslant 8}_{\mathrm{CAMM}} \rightarrow \Delta E^{\kappa\leqslant 9}_{\mathrm{CAMM}}$	0.01	0.03	0.01	0.04	0.01	0.04	0.04	0.09

Table 4.2: Maximal and average changes in the multipole electrostatic interaction on the intercalation plane when increasing the interaction rank by one. Columns labeled 100x100 consider the entire intercalation plane probed, while 33x33 denotes a central part containing only the vicinity of the binding site. All values are given in kcal/mol.

The fact that these values do not largely differ hints that base pair polarization, while influencing the properties of the bases themselves, does not qualitatively change their interactions with the intercalator.

Finally, an attempt can be made to compare the CAMM interaction energy for the complex $Eth^{(+1)}$-GC/CG to the multipole electrostatic binding energy obtained by Medhi et al.[339], which was based on the same structure and obtained from a distributed multipole analysis (DMA). The quite large discrepancy between the value in that study (around -7 kcal/mol) and our values (-16.8 kcal/mol from DMA analysis and -13.0 kcal/mol for the converged CAMM interaction) probably originates from the fact that Medhi et al. used a differently constructed, idealized model and from the differences in the electron density source (our multipole moments were calculated from RHF density matrices, while the former partly from MP2 wave functions).

To illustrate the convergence of the CAMM interaction on the whole intercalation plane (Fig. 4.9 the convergence at the crystallographic alignment of the chromophore), differences between consecutive $\Delta E^{\kappa\leqslant \mathrm{L}}_{\mathrm{CAMM}}$ surfaces ($L = 0...9$) are presented in Table 4.2. Both maximal and average absolute changes are presented, for the entire 100x100 grid used and for a 33x33 central part. While this gives an overview about how drastically the potential energy surface changes with the interaction rank, the basic conclusion remains the same as with Fig. 4.9 – a significant drop in both the maximal and average values occur for rank $L > 5$.

[339]Medhi, C., Mitchell, J. B. O., Price, S. L., Tabor, A. B. *Biopolymers* **1999**, *52*, 84–93.

4.5 Conclusions

The focus of this chapter was on analyzing *ab initio* interaction energies between cationic intercalators and their hosts, based on the crystallographic structures. Two of the chosen structures involved ethidium, the third contained proflavine.

First, a few important assumptions were confirmed by performing interaction energy decomposition (using the HVPT method discussed in Chapter 2) for the original crystallographic geometry of the $Eth^{(+1)}$-UA/AU for models with different sizes and fragmentation schemes. It is acceptable to dissect the system along chemically intuitive lines – separating nucleobases in the **AC** model and cutting apart the nucleic acid strand at the linkage between the nucleosides and phosphate groups – since this does not introduce relevant errors to the interaction energy and makes the computational task significantly smaller.

The absolute influence of the side chain is shown to be large, increasing the interaction energy from slightly over 30 kcal/mol for model **AB** to almost 60 kcal/mol for model **B**. The charge on the phosphate group here is key, and cannot be left unmatched as it almost doubles the interaction energy again. Compensating with a single hydrating water or single counterions diminishes this difference by around 5 kcal/mol, therefore the actual range of possible interactions can be expected to be narrower in solution.

On the other hand, interactions between the intercalating ethidium and its nearest four nucleobases were found to sufficiently reproduce chromophore alignment in the intercalation plane of the $Eth^{(+1)}$-UA/AU complex. The minimum of the total MP2 interaction energy precisely reproduces the crystallographic position of the ethidium chromophore in the between UA/AU bases.

Furthermore, the electrostatic component constitutes the same fraction of the total energy for all three studied structures. However, the multipole electrostatic interaction energy, calculated from Cumulative Atomic Multipole Moments (CAMMs), was found to converge only after including components above the fifth order. CAMM interaction surfaces, calculated on grids in the intercalation planes of these structures, reasonably reproduce the alignment of intercalators in crystal structures; they exhibit additional minima in the direction of the DNA grooves, however, which also need to be examined at higher theory levels if no crystallographic data are given.

5 Summary & outlook

5.1 Summary

In forty thousand words, this dissertation discusses the key aspects of the current literature on non-empirical electrostatic interactions and provides examples where they can be successfully applied. Two methods are central to the presented work – the first (HVPT) decomposes the interaction energy into physically meaningful components at reduced cost compared to the state-of-the-art SAPT approach, and the second (CAMM) allows for a mobile approximate description of charge density distributions. An interrelated series of studies illustrate how electrostatic effects derived from first principles can be used to reproduce specific structural and energetic features. Stress is placed on pinpointing the limits to which electrostatic interactions can be utilized for such tasks, and non-parametric statistical tools are employed in order to explore these limits.

The majority of conclusions address objectives outlined in *Purpose & overview*, therefore the main points are summarized here in the same sequence.

1. Electrostatic interactions at the Hartree-Fock level are capable of reproducing relative CCSD(T) stabilities of over 90% percent of all pairs of dimers in the S22 training set, which contains various interaction motifs typical for biological systems. This favorable correlation is persistent at shorter intermolecular separations ($d_{\mathrm{COM}} < -0.2$ Å) where, in contrast, MP2 and CCSD(T) interaction energies fail to correlate with the equilibrium reference value. This enforces an earlier observation in the literature, of unexpected correlation between electrostatic interactions and experimental stabilization energies for inhibitors optimized to artifact geometries in an enzyme active site using force fields.

2. The same analysis shows that at larger distances ($d_{\mathrm{COM}} > 1.5$ Å) both the total interaction energy (at the MP2 or CCSD(T) level) and its first-order, uncorrelated electrostatic component successfully reproduce the CCSD(T) equilibrium stabilization energy. Not surprisingly, the CCSD(T) energies do this with the best rating, succeeding over 90% of the time even for distances up to 10 Å from equilibrium, where the interaction is dominated by electrostatic effects. From this and the substantially worse correlation and success rate of first-order electrostatics (around 80%), it follows that the contribution of intramolecular correlation to the electrostatic interaction cannot be neglected. Therefore, using multipole moments generated from correlated densities is recommended.

3. A statistical survey of all 16 pairs of stacked nucleic acid bases in B-DNA conformation shows that the electrostatic component correlates well with the total MP2 interaction. For a series of basis sets, the correlation coefficient is always above 0.85, with a linear prediction interval of about 1.5 kcal/mol. An additional strong correlation was revealed between the exchange and dispersion components, amounting to at least -0.95 for all the basis sets studied, with a prediction interval of 1.4 kcal/mol. Further examination for A-DNA and other sets of structures points to the limitations of these correlations, which are highly dependent on the geometrical homogeneity of structures.

4. The MP2 interaction energy – between the ethidium chromophore and its nearest bases in the Eth(+1)-UA/AU complex – reproduces within 0.1 Å the crystallographic binding site in the intercalation plane. This demonstrates that local interactions alone may decide about the alignment of the chromophore between base pairs, or at least amount to a good model for this system. Two other intercalation complexes were studied, and in all three the electrostatic term comprises the same percentage of the total interaction energy, around 82%. Less than half originates from multipole moments based on Hartree-Fock densities, illustrating the magnitude of electrostatic penetration effects. It should be kept in mind that these values will be smaller for larger basis sets, but can be expected to correlate with the total energy. The binding site on the intercalation plane can also be reasonably reproduced at a lower computational cost than the MP2 level, by examining the interaction energy surface of the chromophore between nucleotides with CAMM multipoles. The accuracy is worse and in the studied systems was no less than about 1.3 Å.

5. An analysis of the position of the ethidium chromophore in the Eth(+1)-UA/AU complex shows that the ethidium side chain and even steric constraints with the nucleic acid backbone are of minor importance. However, the latter almost doubles the absolute interaction energy when the phosphate groups are capped with hydrogen and neutral. Replacing the capping hydrogen with a monovalent ion strengthens the interaction by about 10 kcal/mol. On the other hand, releasing the charge on the phosphate groups boosts the total interaction strength to over 130 kcal/mol, while hydration with a single water on each group decreases it by only 10 kcal/mol. Since the solvent molecules and counterions around nucleic acid are highly mobile, the range of interactions can be expected to be somewhere in between these extremes.

Besides the five major points above, the original results presented in the introductory section have yielded methodological conclusions that are worth mentioning. The first addresses the efficiency and basis set stability of interaction energy components in various decomposition schemes. Using small dimers as examples, a comparison is presented between HVPT and other methods. In particular, the HVPT values correspond very closely to state-of-the-art SAPT numbers, while requiring significantly less computational effort. On the other hand, basis set stability is much more favorable compared to various variational EDA schemes.

In practice, a reasonable approach to sizable molecular problems is to choose the largest model possible and the smallest reasonable basis set set. In this regard, the hybrid method used reduces the entry level and is a strategic choice considering the relatively large molecular systems targeted in this dissertation.

Some conclusive observations are also made while discussing the origins and convergence properties of interactions between sets of atomic multipole moments. Several examples are given – notably π-π stacking complexes of nucleobases and intercalated nucleic acids – where higher multipole moments than normally used (rank five or six) are necessary in order to obtain converged interaction energies. This can be an issue, for instance, if an accurate estimate of electrostatic penetration effects is desired.

Coincidentally, an analysis of the MEP on the Connolly surface around reactants gives an estimate of the penetration part – about 1% of the *ab initio* expected value. Multipole analysis along the reaction path also provides insight into charge redistribution near the transition state. Substantial contributions from higher multipoles indicates that approximate point charge models used in conventional force fields are not adequate for representing such changes in reacting systems.

5.2 Future work

While many directions for prospective research could be based on the work presented here, two stand out as the most appealing. It seems especially valuable to enlarge the S22 test set used to study statistical relationships between interaction energy components at various intermolecular distances. With a larger number of dimers the results will naturally be more significant, but could also be analyzed from additional angles, for example separately for hydrogen bonded dimers and other types of complexes. Another feature already mentioned is that MP2 and coupled cluster interaction energies correlate better with the equilibrium stabilization than Hartree-Fock-based electrostatics, hinting that intramolecular electron correlation can play a significant role. For this reason, it would be worthwhile to reevaluate these statistics using electrostatic interactions based on correlated densities.

Extensions can be formulated from the last chapter, which evaluates the influence of the surroundings of nucleic acid strands on its interaction with intercalating ethidium. The charge state of the phosphate groups predictably has a large effect, changing the interaction energy by about 100% when comparing neutral and charged states. After adding a single counterion or water molecule, this difference is compensated only by about 10%. Sampling additional counterion positions and more hydrated systems could give a clearer picture of the possible dynamic range of interaction strengths.

A Cartesian cumulative atomic multipole moments

Generating atomic moments

The algorithm for generating cumulative atomic multipole moments (CAMM) in Cartesian coordinates is a paraphrase of the equations (2.50) and (2.51), the pseudo-code of which is shown in Algorithm 1. Proceeded by an electronic structure calculation, the computational cost of this algorithm is almost always insignificant.

Algorithm 1 Pseudo-code for generating CAMM atomic moments.

```
for each moment M_kml of rank κ < L do
  for all pairs of atomic orbitals I and J do
    find atoms i and j such that I ∈ i and J ∈ j
    calculate the contribution from this orbital pair, x = P_IJ ⟨I|x^k y^l z^m|J⟩
    add half of x to moment M_kml,i
    add half of x to moment M_kml,j
  end for
end for
for each atom i do
  add contribution from nuclear charge, Z_i x_i^k y_i^l z_i^m
  for all moments with rank lower than κ do
    add cumulative contributions according to (2.51)
  end for
end for
```

The algorithm has been implemented by others in the past. Sokalski and Poirier published the original formulation of CAMM[340] and Sawaryn and Sokalski later provide a description of the implementation[341]. Reimplemented by Strasburger and Sokalski, the approach was ued to study intramolecular interactions for molecules with extended single bonds.[342]

Also, the approach proposed by Stone in the DMA method[343] is conceptually the same, differing in the way the charge density contributions are distributed among atoms (see Section 2.5.1 for details). Therefore, results from CAMM and DMA calculations should yield approximately equivalent results at moderate intermolecular distances, which has been pointed out by Scheiner in his early quantitative hydrogen bonding model.[344]

For this dissertation, Algorithm 1 was coded in a compact form within the quantum chemistry program GAMESS-US, by modifying already existing code for molecular electric mo-

[340] Sokalski, W. A., Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86–92.
[341] Sawaryn, A., Sokalski, W. A. *Comput. Phys. Commun.* **1989**, *52*, 397–408.
[342] Strasburger, K., Sokalski, W. A. *Chem. Phys. Lett.* **1994**, *221*, 129–135.
[343] Stone, A. J., Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.
[344] Spackman, M. A. *J. Chem. Phys.* **1986**, *85*, 6587–6601.

ments. Generalization of subroutines that calculate multipole integrals has additionally allowed for the evaluation of moments up to a rank of $L = 16$. In the context of GAMESS-US, summing orbital product contributions into a two dimensional array `A` of atomic moments can be summarized by the following code fragment,

```
      K=0
      do 100 IA=1,NAT
      do 100 IO=LIMLOW(IA),LIMSUP(IA)
      do 100 JO=1,IO
         IB = ITAB(JO)
         K = K+1
         X = P(K) * AMI(K)
         A(M,IA) = A(M,IA) - X
         if (IO.NE.JO) then
            A(M,IB) = A(M,IB) - X
         endif
  100 enddo
```

where `NAT` is the number of atoms, `LIMLOW` and `LIMSUP` index the lower and upper atomic orbital indexes for an atom, and `ITAB` assigns the atom index an atomic orbital belongs to. Arrays `P` and `AMI` contain the respective density matrix elements and multipole integrals[345] in the order defined by the code (iteration over atoms, atomic orbitals on the atom, and all atomic orbitals with lower indexes). When `I` and `J` belong to different atoms, each contribution in `P` is divided evenly between them; the off-diagonal elements are halved beforehand. The fragment is embedded in an external loop over multipole moments, indexed by `M`.

Another part in the loop iterated by `M` sums the nuclear charge contributions into appropriate elements in `A`. Here, the auxiliary array `INTCOOR` is used to map the multipole moment index `M` into the appropriate powers over all three coordinates, `C` contains atom coordinates, `XP`, `YP` and `ZP` are the molecule's center of mass, and `ZAN` contains nuclear charges,

```
      do IA=1,NAT
         COOR(1) = C(1,IA) - XP
         COOR(2) = C(2,IA) - YP
         COOR(3) = C(3,IA) - ZP
         x=ZAN(IA)
         do I=1,3
            X = X * (COOR(I)**INTCOOR(I,M))
         enddo
         A(M,IA) = A(M,IA)+X
      enddo
```

After these two fragments, the array `A` contains atomic moments as defined by (2.50), which are *additive* in the sense that they sum up to the appropriate molecular moments. A transformation is normally (but optionally) performed that moves the moment to their local atomic coordinates according to the recombination expression in (2.51). Writing it out explicitly, one needs to subtract all products of lower rank moments and appropriate coordinates. Following the notation used by and Sokalski and Poirier[340] and adopted here in

[345] The atomic orbitals are gaussian functions and the multipole integrals held in `X` ($P_{IJ}\left\langle I|x^k y^l z^m|J\right\rangle$ in Algorithm 1) are equivalent to overlap integrals for functions of appropriately higher symmetry.

(2.38), the expressions for the first few moments are,

$$\begin{aligned}
\mu_\alpha^{\mathrm{CAMM}} &= (M_{100}^{\mathrm{CAMM}}, M_{010}^{\mathrm{CAMM}}, M_{001}^{\mathrm{CAMM}}) = \mu_\alpha - q\alpha && (\alpha = x, y, z),\\
\Omega_{\alpha\beta}^{\mathrm{CAMM}} &= \Omega_{\alpha\beta} - q\alpha\beta - \mu_\alpha\beta - \mu_\beta\alpha && (\alpha, \beta = x, y, z),\\
\Theta_{stu,i}^{\mathrm{CAMM}} &= \Theta_{stu,i} - qs_it_iu_i - \mu_{s,i}t_iu_i - \mu_{t,i}u_is_i - \mu_{u,i}s_it_i\\
&\quad - \Omega_{st,i}u_i - \Omega_{tu,i}s_i - \Omega_{us}t_i && (s, t, u = x, y, z),\\
\Psi_{stuv,i}^{\mathrm{CAMM}} &= \Psi_{stuv,i} - qs_it_iu_iv_i - \mu_{s,i}t_iu_iv_i - \mu_{t,i}u_iv_is_i - \mu_{u,i}v_is_it_i - \mu_{v,i}s_it_iu_i\\
&\quad - \Omega_{st,i}u_iv_i - \Omega_{su,i}t_iv_i - \Omega_{sv,i}t_iu_i - \Omega_{tu,i}s_iv_i - \Omega_{tv,i}s_iu_i - \Omega_{uv,i}s_it_i\\
&\quad - \Theta_{stu,i}v_i - \Theta_{tuv,i}s_i - \Theta_{uvs,i}t_i - \Theta_{vst,i}u_i && (s, t, u, v = x, y, z),
\end{aligned}$$

where μ denotes a dipole moment and Ω an octupole moment, and Θ and Ψ are octupoles and hexadecapoles, respectively. This transformation can be performed generically for any rank in-place on the array `A`,

```
      do IA=1,NAT
      X = 0.0
      IK = INTCOOR(1,M)
      IL = INTCOOR(2,M)
      IM = INTCOOR(3,M)
      do 100 IKK = 0,IK
      do 100 ILL = 0,IL
      do 100 IMM = 0,IM
         Y = 1.0
         N = MAPINT(IKK+1,ILL+1,IMM+1)
         if(N.ne.M)then
            Y = Y * A(N,IA)
            Y = Y * IBC(IK,IKK)
            Y = Y * IBC(IL,ILL)
            Y = Y * IBC(IM,IMM)
            Y = Y * COOR(1)**(IK-IKK)
            Y = Y * COOR(2)**(IL-ILL)
            Y = Y * COOR(3)**(IM-IMM)
            X = X - Y
          endif
  100 enddo
      A(M,IA) = A(M,IA) + x
      enddo
```

In the above code fragment, subroutine `IBC` returns the binomial coefficient and `INTCOOR` is the same as before (maps the multipole moment index `M` to the appropriate powers over all three coordinates). Another auxiliary array (`MAPINT`) provides the inverse mapping – iit gives the index of a moment, `M`, given the coordinate ranks klm in $M_{klm,i}^{\mathrm{CAMM}}$. This code need to be iterated over increasing values of the index `M` in order to agree with (2.51). The triple loop performed for each atom in this code corresponds to the triple sum in that equation, while the condition `N.ne.M` allows the case when $k'l'm' = klm$ to be skipped.

Evaluating interaction energies

Operating on the entire traceless multipole tensors in Cartesian coordinates, it is possible to derive explicit symbolic expressions for their interactions based on (2.46). Performing full contractions with the symmetric interaction tensor $\mathbf{T}^{(\kappa_a+\kappa_b)}_{|\mathbf{r}_{ij}|} = -\nabla^{\kappa_a+\kappa_b}\left(\frac{1}{|\mathbf{r}_{ij}|}\right)$ yields the interactions between multipoles of the few first lowest ranks,

$$
\begin{aligned}
\Delta E^{(00)}_{\mathrm{el,mtp}} &= \frac{q_i q_j}{|\mathbf{r}_{ij}|} \\
\Delta E^{(10)}_{\mathrm{el,mtp}} &= \tilde{\mu}_\mathrm{j}[1]\mathbf{T}^{(1)}_{|\mathbf{r}_{ij}|}\mathrm{q_i} = \mathrm{q_i r_{ij}^{-3}}\,(\mathbf{r}_{ij}\cdot\tilde{\mu}_\mathrm{j}) \\
\Delta E^{(11)}_{\mathrm{el,mtp}} &= \tilde{\mu}_\mathrm{j}[1]\mathbf{T}^{(1)}_{|\mathbf{r}_{ij}|}[1]\tilde{\mu}_\mathrm{i} = (\tilde{\mu}_\mathrm{i}\cdot\tilde{\mu}_\mathrm{j})\,\mathrm{r_{ij}^{-3}} - 3\,(\tilde{\mu}_\mathrm{i}\cdot\mathbf{r}_{ij})\,(\mathbf{r}_{ij}\cdot\tilde{\mu}_\mathrm{j})\,\mathrm{r_{ij}^{-5}} \\
\Delta E^{(20)}_{\mathrm{el,mtp}} &= \mathbf{\Omega}_\mathrm{j}[2]\mathbf{T}^{(1)}_{|\mathbf{r}_{ij}|}\mathrm{q_i} = (\mathbf{r}_{ij}\cdot\mathbf{\Omega}_\mathrm{i}\cdot\mathbf{r}_{ij})\,\mathrm{q_j r_{ij}^{-5}} \\
\Delta E^{(21)}_{\mathrm{el,mtp}} &= \mathbf{\Omega}_\mathrm{j}[2]\mathbf{T}^{(1)}_{|\mathbf{r}_{ij}|}[1]\tilde{\mu}_\mathrm{i} = 2\,(\tilde{\mu}_\mathrm{j}\cdot\mathbf{\Omega}_\mathrm{i}\cdot\mathbf{r}_{ij})\,\mathrm{r_{ij}^{-5}} - 5\,(\mathbf{r}_{ij}\cdot\mathbf{\Omega}_\mathrm{i}\cdot\mathbf{r}_{ij})\,(\tilde{\mu}_\mathrm{j}\mathbf{r}_{ij})\,\mathrm{r^{-7}}.
\end{aligned}
$$

One needs only define appropriate scalar products and tensor contractions and these equations can in principle be programmed recursively. It is easier and more efficient, however, to implement the terms of (2.36), using only unique elements of the multipole tensors and different prefactors depending on the frequency of their occurrence as expressed by (2.35).

A straightforward way to do this is the following fragment, although it is not efficient, especially for small values of k, l, m. In practice, it eeds to be optimized by using helper arrays (the one here, `F`, holds factorials) and reorganizing the triple loop.

```
      subroutine T(K,L,M,DR)
      dimension DR(3)
      integer F(11)
      data F /1.0, 1.0, 2.0, 6.0, 24.0, 120.0,
     *        720.0, 5040.0, 40320.0, 362880.0, 3628800.0 /
      R = sqrt(DR(1)*DR(1) + DR(2)*DR(2) + DR(3)*DR(3))
      XR = DR(1) / R
      YR = DR(2) / R
      ZR = DR(3) / R
      N = K + L + M
      T = 0.0
      do 100 KK = 0,int(K/2)
         KKK = K - 2*KK
         do 100 LL = 0,int(L/2)
            LLL = L - 2*LL
            do 100 MM = 0,int(M/2)
               MMM = M - 2*MM
               NN = KK + LL + MM
               X = (-1.0)**NN * F(2*N-2*NN+1) / F(N-NN+1)
               X = X / (F(KK+1)*F(LL+1)*F(MM+1))
               X = X / (F(KKK+1)*F(LLL+1)*F(MMM+1))
               T = T + X * (XR)**(KKK) * (YR)**(LLL) * (ZR)**(MMM)
  100 enddo
      X = (-1.0)**N * F(K+1) * F(L+1) * F(M+1) / (2.0**N)
      T = T * X / (R**(N+1))
      end subroutine T
```

Forces acting between multipole moments

Following (2.36), a derivative of the multipole interaction energy between two molecules with respect to moving the ith atom in molecule A by an infinitesimal distance x_i has the form,

$$\begin{aligned}\frac{\partial}{\partial x_i}\Delta E_{\mathrm{el,mtp}} &= \sum_{j\in B}\sum_{klm}\sum_{k'l'm'}\frac{\partial}{\partial x_i}\left[M_{i,klm}T_{(k+k')(l+l')(m+m')}(|\mathbf{r}_{ij}|)M_{j,k'l'm'}\right]\\ &= \sum_{i'\in A}\sum_{j\in B}\sum_{klm}\sum_{k'l'm'}\frac{\partial}{\partial x_i}\left(M_{i',klm}\right)T_{(k+k')(l+l')(m+m')}(|\mathbf{r}_{ij}|)M_{j,k'm'l'} \qquad \text{(A.1)}\\ &+\sum_{j\in B}\sum_{klm}\sum_{k'l'm'}M_{i,klm}\frac{\partial}{\partial x_i}\left(T_{(k+k')(l+l')(m+m')}(|\mathbf{r}_{ij}|)\right)M_{j,k'm'l'}.\end{aligned}$$

The derivatives of multipole moments – $\frac{\partial}{\partial x_i}\left(M_{i,klm}\right)$ – can be expressed in terms of polarizabilities, that is linear combinations of multipole moments and multipole integrals of lower ranks. Assuming in a first approximation that all moments in the molecule are constant, the first contribution becomes zero and the derivative of the interaction energy is reduced to calculating the derivative of the interaction tensor,

$$\frac{\partial}{\partial x_i}\Delta E_{\mathrm{el,mtp}} \simeq \sum_{j\in B}\sum_{klm}\sum_{k'l'm'}M_{i,klm}\frac{\partial}{\partial x_i}\left[T_{(k+k')(l+l')(m+m')}(|\mathbf{r}_{ij}|)\right]M_{j,k'l'm'}.$$

Any partial derivative of an action tensor element can be easily evaluated using the relations $\partial x_i = \partial r_x$ and $\partial x_j = -\partial r_x$, where $r_x = x_i - x_j$. Using the general definition in (2.34), it can be seen that the derivatives of its elements are equal to the elements of the tensor of one rank higher, with a minus sign depending on the molecule the derivative is calculated for,

$$\frac{\partial}{\partial x_i}T_{klm}(|\mathbf{r}|) = \frac{\partial}{\partial x_i}\frac{\partial^\kappa}{\partial r_x^k\partial r_y^l\partial r_z^m}\frac{1}{|\mathbf{r}|} = \frac{\partial^{\kappa+1}}{\partial r_x^{k+1}\partial r_y^l\partial r_z^m}\frac{1}{|\mathbf{r}|} = T_{(k+1)lm}(|\mathbf{r}|), \qquad \text{(A.2)}$$

$$\frac{\partial}{\partial x_j}T_{klm}(|\mathbf{r}|) = \frac{\partial}{\partial x_j}\frac{\partial^\kappa}{\partial r_x^k\partial r_y^l\partial r_z^m}\frac{1}{|\mathbf{r}|} = -\frac{\partial^{\kappa+1}}{\partial r_x^{k+1}\partial r_y^l\partial r_z^m}\frac{1}{|\mathbf{r}|} = -T_{(k+1)lm}(|\mathbf{r}|), \qquad \text{(A.3)}$$

and analogically for the other two derivatives $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial z}$. These interaction and tensor element derivatives will only be useful in conjunction with the interaction energy itself. Therefore, the tensor elements used to calculate derivatives should be stored between consecutive multipole ranks for each atom pair and used to calculate interaction energies one rank higher.

In practice, this means that the analytical gradient of the electrostatic multipole interaction can be approximated with little extra cost compared to calculating the interactions alone. For example, the gradient of a dipole-dipole interaction in Cartesian form involves the interaction tensor element from quadrupole-dipole interactions. Evaluating derivatives this way introduces only one extra rank of tensor elements that needs to be computed, along with a few auxiliary basic arithmetic operations, as showed in Algorithms 5 and 6.

Prospects for molecular dynamics

The recurrent character of the Cartesian interaction tensor and of relations (A.2) and (A.3) would benefit applications such as molecular dynamics. As discussed in Section 2.5.4 alongside atomic moment transferability, there have been efforts to correct electrostatic interactions during simulations by including multipole interactions. However, most do not change the values of moments during simulations, or rely on a parametrization performed beforehand. Obviously, these are reasonable approaches only for the smallest molecules.

Even the most flexible molecules, however, repeatedly visit the same conformations, therefore some form of library of moments is indispensable. General ideas are presented and discussed in the following paragraphs, and Fig. A.1 presents a schematic protocol for including multipole effects in simulations. Pseudo-code is listed in Algorithms 2-6 for subroutines titled there in upper case, based on CAMM moments and existing code by Plattner and Meuwly that interfaces with CHARMm.[346]

Besides initializing arrays (`MTPINIT` in Algorithm 3) and reading atomic moments from an input file, the implementation needs to be fed the simulation coordinates and influence the next increment by adding corrections to the energy and derivatives. Fig. A.1 illustrates these relations with arrows between nodes.

The additional operations that need to be performed in each time step can be divided into two kinds, the first dedicated to updating atomic moments, the second to evaluating improved electrostatic interactions. These parts are separated into two green clusters in the diagram.

The first green cluster embodies a loop over all the molecules equipped with atomic moments. At this stage a decision is made, by comparing the simulated conformation with those available in the library, whether the deviation from any known conformation is acceptable. If not, then an external program needs to be called that calculates the atomic moments for the new conformation and adds them to the library. If it is acceptable, however, the program can proceed to approximate the atomic moments. This can be done in at least two ways – by treating the conformational changes as internal rotations of atoms as per (2.52), or by interpolating from two or three nearby library conformations. Whatever is chosen, the new atomic moments for each molecule can be finally rotated to their orientation within the simulation.

If the molecules are rigid or the moments are assumed to be constant, then most of these steps can be skipped, and only small internal rotations of atoms (if any) and molecular rotations need to be performed. The most problematic aspect in the described procedure is the reference frame in which conformational changes are evaluated. It is unclear, for example, if it is practical to use natural internal coordinates with constant connectivity definitions or whether some kind of alignment routine would be more useful and efficient. The central question is how an arbitrary change in atomic coordinates should be expressed in terms of the internal molecule conformation. Once this is resolved, the technical means to build up a library and use it to interpolate moments for new conformations remains to be implemented.

The second green cluster shown in Fig. A.1 symbolizes the the evaluation of multipole

[346] Plattner, N., Meuwly, M. *Biophys. J.* **2008**, *94*, 2505–2515; Plattner, N., Meuwly, M. *ChemPhysChem* **2008**, *9*, 1271–1277; Plattner, N., Meuwly, M. *J. Mol. Model.* **2009**, *15*, 687–694.

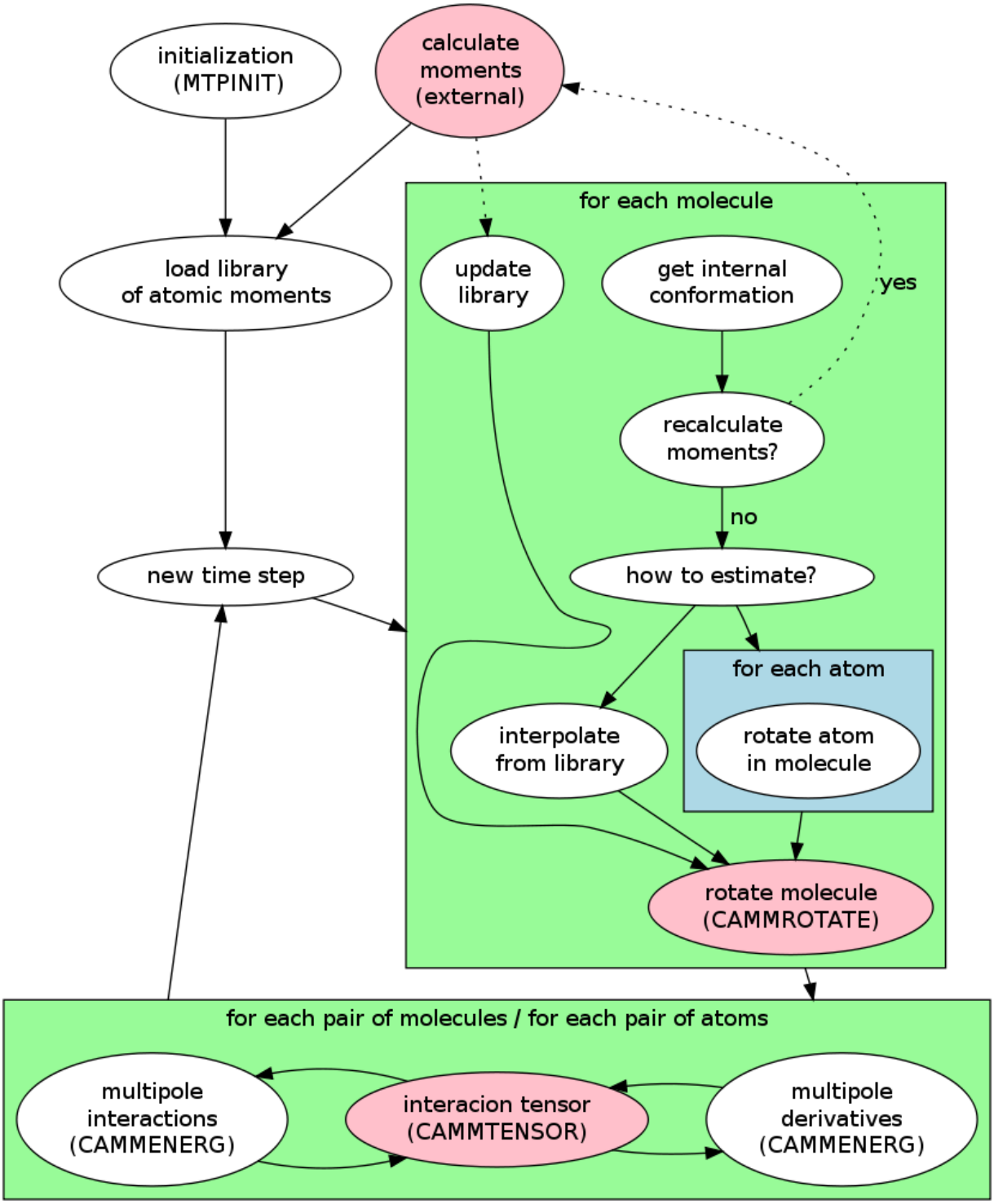


Figure A.1: Schematic for the proposed use of electrostatic atomic multipole moments instead of only charges during atomistic simulations. Arrows denote the sequence of chronological events, while the clusters represent loops and their labels (near top) describe the nature of the loop. Pink is used to highlight the most computationally intensive elements. The upper case titles in several nodes are the names of the procedures listed and discussed in the text.

contributions to the energy and derivatives, given that all the new coordinates and atomic moments are now known. It is a beneficial feature of Cartesian multipole moments and their interactions that they are expressed in the same coordinates as are normally used during simulations. Also, as shown in the previous section, the derivative of (2.37) in any direction can be expressed using an interaction tensor element of a larger rank in that coordinate. That is why the nodes representing both the energy and derivative calculation in the diagram are undersigned with the procedure `CAMMENERG`, which is used in both cases (see Algorithm 5).

A comment should be made about the additional computational cost ensued by including these procedures. While the general equation used for the Cartesian interaction tensor element (2.35) is lengthy, its implementation `CAMMTENSOR` (Algorithm 6) is straightforward and much of the prefactors involved do not change between elements.

By far the most expensive procedure is `CAMMROTATE` (Algorithm 4), which contains trigonometric functions. The rotation matrix $R_{O\to\tilde{O}}$ in (2.52) can be parametrized with Euler angles,

$$\tilde{O}(\gamma_x,\gamma_y,\gamma_z)\mathbf{M}^\kappa = \left(R_{\gamma_x,\gamma_y,\gamma_z}\right)^\kappa \mathbf{M}^\kappa. \tag{A.4}$$

Euler angles γ_x, γ_y, γ_z for the rotation can be obtained from the quaternions of a given rotation axis and angle. The choice of this axis and angle is in fact a choice of the local reference system for each atom, and in principle should be arbitrary for small changes in coordinates or small time steps.

Since CAMM multipoles are defined in the context of an entire molecular fragment, a change in the orientation of an atom i between time steps should be expressed relative to a dynamic center of mass. Namely, the axis angle $\alpha_{i,n}$ for time step n is spanned between the atom vector in the current and previous time steps,

$$\alpha_{i,n} = \arccos\left(\frac{\mathbf{r}_{i,n}\cdot\mathbf{r}_{i,n-1}}{|\mathbf{r}_{i,n}||\mathbf{r}_{i,n-1}|}\right), \tag{A.5}$$

where the the atom vector $\mathbf{r}$ is relative to the center of mass in the current time step,

$$\mathbf{r}_{i,n} = \mathbf{R}_{i,n} - \sum_i m_i\mathbf{R}_{i,n}\cdot\left(\sum_i m_i\right)^{-1}, \tag{A.6}$$

and $\mathbf{R}_{i,n}$ is the position in the global coordinate system and m_i is the atomic mass of atom i.

While the external step, in which multipole moments are calculated, possibly using a very expensive and accurate *ab initio* method, is in itself a bottleneck that will probably block the simulation, it is a rare event. In practice, the moments would be generated mostly at the beginning of the simulation, every time a conformation is found that is judged to be too far from any other conformation in the library. In this case, too, it remains to be seen what criteria for recalculation would be acceptable and how many library instances would be needed to populate a reasonably large portion of a flexible molecule's conformational space.

Algorithm 2 Pseudo-code of a simulation loop that includes CAMM atomic multipole interactions by calling subroutine `MTPX`. Here, N_A denotes the total number of atoms.

Initialize common blocks and multipole moments.	*subroutine `MTPINIT`*
for each time step in simulation (*subroutine `MTPX`*) **do**	
Increment internal `MTP` counter.	*1 sum.*
Rotate Cartesian multipole moments.	*subroutine `CAMMROTATE`*
Update stored coordinates with new values.	N_A *div.*
Evaluate interactions and derivatives.	*subroutine `CAMMENERG`*
Add interactions to simulation energies.	*1 mult., 1 sum.*
Add derivatives to simulation forces.	$3N_A$ *mult.,* $3N_A$ *sum.*
Calculate RMS of interactions and derivative.	*2 sqrt., 2 div.,* $6N_A$ *mult.,* $6N_A$*+2 sum.*
end for	

Algorithm 3 Pseudo-code for `MTPINIT` – subroutine that initializes common blocks related to CAMM atomic moments. Here, N_A denotes the total number of atoms, N_F the number of fragments described by atomic moments and L is the maximum interaction rank considered.

Initialize array of factorials up to 2L+2.	*2L+1 mult., 2L+1 sum.*
Initialize binomial coefficients up to L.	$\sim L^3$ *div.,* $\sim L^3$ *mult.,* $\sim L^3$ *sum.*
Initialize array of (2N!)/N! and powers of -1 and 2 up to L+1.	*L+2 pow., L+2 div., L+2 mult.*
Initialize array of moment indexes.	*L div., L mult., 6L sum.*
Initialize array of fragment masses.	N_F *mult.,* N_A *sum.*
Initialize atom coordinates in bohrs.	N_A *div.*
Initialize centers of mass of all fragments.	$3N_F$ *div.,* $3N_A$ *mult.,* $3N_A$ *sum.*

Algorithm 4 Pseudo-code for `CAMMROTATE` – subroutine that rotates CAMM atomic moments.

for all fragments described by multipole moments **do**	
Update center of mass based on new coordinates	$3N_F$ *div.,* $3N_A$ *mult.,* $3N_A$ *sum.*
for all atoms in the fragment **do**	
Find displacement of atom relative to previous time step	*2 sqrt., 5 div., 12 mult., 12 sum.*
Express the change in direction in quaternions	*1 sqrt., 1 div., 15 mult., 6 sum.*
Express the change in direction in Euler angles	*3 trig., 10 mult., 5 sum.*
Save the angles in array `ANGLES`.	
end for	
for three axes `X`,`Y`,`Z` **do**	
Set indices of axes.	*3 sum.*
for all atoms in the fragment **do**	
Recover angle and evaluate sine and cosine.	*2 trig., 1 sum.*
for all multipole moments on atom **do**	
Operations extracted before loop.	*4 sum.*
for partial exponents over coordinates **do**	
Calculate contribution to new multipole moment.	*2 pow., 4 mult., 13 sum.*
end for	
end for	
for all multipole moments on atom **do**	
Copy moments from temporary array to common block.	*2 sum.*
end for	
end for	
end for	
end for	

Algorithm 5 Pseudo-code for `CAMMENERG` – subroutine that evaluates CAMM atomic moment interactions and derivatives for all atoms in the simulation.

```
Zero interaction and derivative arrays.
for all pairs of atoms IAT and IAT2 do
  if IAT and IAT2 are in different fragments then
    Find distance between the two atoms.                          1 sqrt., 3 mult., 5 sum.
    if The atoms IAT and IAT2 are closer than cut-off distance then
      Operations extracted before loop.                           1 div., 3 mult.
      Zero array of calculated interaction tensor elements.
      for all pairs of ranks L1,L2 for atoms IAT and IAT do
        for all unique moments NM1 of rank L1 do
          Operations extracted before loop.                       4 mult.
          for all unique moments NM2 of rank L2 do
            Intermediate variables.                               1 div., 3 mult., 7 sum.
            if L1+L2 = 0 then
              Calculate interaction tensor element.               subroutine CAMMTENSOR
            else
              Restore interaction tensor element from array TTMP.
            end if
            Sum contribution to total interaction energy.         1 mult., 1 sum.
            for three axes X,Y,Z do
              Find coordinate exponents for this axis.            6 sum.
              if interaction tensor element has already been calculated then
                Restore tensor element from array TTMP.
              else
                Calculate interaction tensor element.             subroutine CAMMTENSOR
                Save tensor element in array TTMP for later use.
              end if
              Add contribution to atomic derivatives of IAT and IAT2.   1 mult., 2 sum.
            end for
          end for
        end for
      end for
    end if
  end if
end for
```

Algorithm 6 Pseudo-code for `CAMMTENSOR` – subroutine that calculates the Cartesian interaction tensor of rank `LMN`.

```
Calculate prefactor of tensor element.                            1 pow., 1 div., 4 mult., 3 sum.
T ← 0.
for partial exponents on three axes do
  Evaluate powers of distances r_x, r_y, r_z.                     3 pow.
  Sum contribution into T.                                        1 div., 12 mult., 7 sum.
end for
Evaluate final value of interaction tensor element.               1 mult.
```

B cclib: interoperability in computational chemistry

In view of the ongoing rapid dissemination of open source software and scripting languages such as Python[347], this outlook aims to give a feeling of their present condition in the field of computational chemistry. In particular, tools facilitating interoperability and the automation of routine tasks are discussed, with primary focus on the parsing library cclib.[348]

Barriers to interoperability

Computational chemists carrying out *ab initio*, density functional or semi-empirical calculations choose from a variety of software packages. Each program is characterized by a range of available methods and theory levels, as well as by how they are implemented.

Due to design differences, the lack of a programming interface (API) in most cases, and the proprietary nature of some codes, most of the packages used in modern computational chemistry are in no way *interoperable*. This often leads to the uncomfortable situation where results obtained from one program may not be reproducible by or even comparable to another within some assumed limits. Furthermore, a particular analysis method may only be available to the users of one program, even though the method is in principle applicable to any.

How do then researchers ensure that analyses apply to output from any program? Typically, they choose several packages that they are interested in, and write routines to extract the necessary information from the text file containing the results (log file) or from a binary file produced during calculations (checkpoint file). The former is often easier since the log file may be readily viewed and it is the output file with which users are most familiar, while the checkpoint files tend to be quite large are are not normally retained.

However, log files can vary, are usually quite free-format, the units may disagree or not be reported, and the same data may be present under different names. On top of this, the specifics of a log file from a particular package may depend on the nature of the calculation, on the version of the software, even on the operating system. As a result, it is unlikely that a parser will be sufficiently robust to deal with the log files of all users. Even if it were, as new versions of the package become available, the author must constantly update the parser or it will become obsolete.

[347] Rossum, G., Drake, F. *Python Reference Manual*; 2001, http://www.python.org.

[348] O'Boyle, N. M., Tenderholt, A. L., Langner, K. M. *J. Comp. Chem.* **2008**, *29*, 839–845; *cclib: A library for package-independent computational chemistry algorithms*; http://cclib.sourceforge.net.

Introducing cclib

The programming library cclib[348] is intended to overcome these difficulties by providing one standard interface to calculation results, independent of the program used. This means that developers do not have to worry about writing a parser, and can concentrate on the analyses or algorithm being implemented. Code based on cclib will work equally well for users of any of the supported computational chemistry packages, and will continue to work with new versions and output file formats as long as cclib is regularly updated to handle new releases.

Two distinct target audiences can be identified. The first consists of chemists who need to repeatedly extract information from log files, who traditionally used a combination of command-line tools such as `grep` and `cut`, or copied text directly to a spreadsheet and edited it by hand or macro. Using cclib, in just a couple of lines of code they can extract information from the log file. The second group are developers of software that, as a necessary first step, parse computational chemistry output files. This could be for molecular visualization software, post-processing code, or an algorithm based on calculation results.

As a collaborative project, cclib is developed using an open source development model[349] and takes full advantage of development resources such as mailing lists, wiki and versioning systems. An open source license, the Lesser GNU Public License (LGPL),[350] was chosen in order to encourage contributions from outside developers and to maximize impact by allowing incorporation into other open and closed source programs.

From the technical point of view, cclib is a library written in Python[347] composed of four modules: `parser`, `bridge`, `method` and `progress`. At the core are the parser classes, of which there are seven in cclib 1.0: for the programs ADF, GAMESS (both GAMESS-US and PC-GAMESS/Firefly), GAMESS-UK, Gaussian, Jaguar, Molpro and ORCA. They can be instantiated directly or by a `ccopen()` function, which detects which package a log file corresponds to. Calling the `parse()` method parses the file and extracts any information found, which is made available through attributes of the resultant data object (`ccData`).

A simple example demonstrates how cclib can be used in practice to calculate a HOMO-LUMO gap (in eV), given the output file of a single point Gaussian energy calculation:

```
>>> from cclib.parser import ccopen
>>> mylogfile = ccopen("methane.log")
>>> print mylogfile
Gaussian log file methane.log
>>> mydata = mylogfile.parse()
>>> homo = mydata.homos[0]
>>> energies = mydata.moenergies[0]
>>> homolumogap = energies[homo+1] - energies[homo]
```

The unified interface to calculation output extends to the data itself. For example, molecular coordinates are always provided in ångström, and vibrational frequencies in cm^{-1}, no matter what units are used in the source log file. In addition to unit conversions, cclib standardizes conventions such as those used for orbital symmetries. For the symmetry labeled BU by GAMESS and Gaussian, ADF uses B.u, GAMESS-UK uses bu and Jaguar uses Bu; cclib

[349] Fogel, K. *Producing Open Source Software*; O'Reilly, 2005.
[350] *GNU Lesser General Public License*; http://www.gnu.org/copyleft/lesser.html.

normalizes all of these to Bu. This issue highlights a general difficultly encountered when parsing log files, namely a lack of detailed documentation. Of these four programs, only ADF provides a manual describing the possible labels. Since many programs do not provide source code, the only way to obtain the full list of labels is to run calculations for all symmetries.

Another notable feature of cclib development is the extensive use of unit tests, short pieces of code designed to test a single functionality. There are two unit test frameworks in the standard Python library, `unittest` and `doctest`, and both are used by cclib. Tests embedded in the documentation strings of modules and functions (*docstrings*) are processed by `doctest`, and are best suited when a correct behavior can be verified by examining a small number of outputs. For example, the symmetry labels produced by ADF are standardized by a function with tests for every possible symmetry that can be found in an ADF log file.

`Unittest` is a more general purpose framework, which cclib uses to ensure consistency between parsers and internal consistency between various data from the same file. For every computational package handled by cclib, a set of standard calculations are performed on the same molecule, including a geometry optimization, single point energy and vibrational frequency calculation. Unit tests ensure that each parser is extracting the correct data – for example, if atom coordinates are in ångström (not bohr), or compare the minimum C-C distance in a molecule to a known value.

All log files used to develop the parsers are stored in source code tree. If, after release, a log file is found that the current parser cannot parse then a test for the bug is added to a *regression* test suite, and the bug is fixed. The regression test suite ensures that a bug, once fixed, will stay fixed. Periodically, a release is made of all log files that have historically broken any parser. In effect, these log files define the behavior and ability of the current parsers, and may be useful to others as a test set for developing similar software.

Several basic computational chemistry algorithms are implemented with cclib, in the `methods` module. After molecular orbital coefficients and the overlap matrix are extracted from a log file, a number of orbital-based population methods can be performed. These include the CSPA method which disregards any overlap between basis set functions and the standard Mulliken population analysis. The density matrix can be calculated, as well as Mayer's bond orders between atoms. The following five lines parse an output file and calculate Mulliken electron populations on atoms (partitioning can be done differently):

```
>>> from cclib.parser import ccopen
>>> from cclib.method import MPA
>>> mydata = ccopen("methane.log").parse()
>>> charges = MPA(mydata)
>>> charges.calculate()
```

Another implemented method is charge decomposition analysis (CDA) developed by Dapprich and Frenking,[351] which describes the interaction between two molecular fragments. Interaction energies are calculated in terms of the mixing between occupied and empty orbitals or occupied orbitals of two fragments. Also included in the methods are functions to calculate the magnitude of the wave function and electron density at grid points in a volume. These

[351] Dapprich, S., Frenking, G. *J. Phys. Chem.* **1995**, *99*, 9352–9362.

functions use the open source quantum chemistry package, PyQuante,[352] to handle the actual calculation, and the resulting volume object can be written to disk in Gaussian cube format or as a Visualization Tool Kit (VTK) file.

Relation to other software

Since one of the goals of cclib is to bring chemical computations and subsequent analysis closer, it promotes interoperability with other open source software, especially written in Python. To this end, the `bridge` module can use the molecular information parsed by cclib to create a Biopython PDB object, OpenBabel OBMol object or a PyQuante Molecule object.

Biopython[353] contains many algorithms of interest in structural biology, such as Superimposer which aligns two molecular conformations by minimizing the root mean squared deviation between the atoms. Of more general interest is the OpenBabel library,[354] a C++ cheminformatics library which provides Python bindings. OpenBabel allows conversion of molecular data between over 60 different file formats, including input file formats for several computational programs. In addition, it contains many algorithms to deal with molecular structures such as ring perception, detection of aromaticity and chirality.

The following shows how to create a PDB file containing the final step of a geometry optimization parsed by cclib,

```
>>> from cclib.parser import ccopen
>>> from cclib.bridge import makeopenbabel
>>> import pybel
>>> data = ccopen("mylogfile.out").parse()
>>> OBmol = makeopenbabel(data.atomcoords[-1], data.atomnos)
>>> pybel.Molecule(OBmol).write("pdb", "finalstep.pdb")
```

PyQuante[352] is a suite of programs for developing quantum chemistry methods. It is written in Python with speed-critical code in C. Since cclib extracts Gaussian basis set in PyQuante format, it is possible to easily create a PyQuante object and carry out an electronic structure calculation with it. The cclib functions for calculating electron density and the wave function magnitude at grid points in a volume interface with PyQuante in this way.

It is prudent to mention other projects that facilitate communication between monolithic computational chemistry codes – they usually rely on converting output, either the binary or associated log file, to a format which can then be used to create an input file for the next program in a workflow. Borini et al.[355] have developed a standardized binary file format based on HDF5 files, a format designed for storing scientific data. Their Q5Cost library provides an API to the contents of the HDF5 file, the scope of which is currently limited to atomic orbital, molecular orbital and wave function data.

[352] Muller, R. P. *PyQuante, Version 1.6.3*; http://pyquante.sourceforge.net/.

[353] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczyński, B., Hoon, M. *Bioinformatics* **2009**, *25*, 1422–1423.

[354] *The Open Babel Package, version 2.0.1*; http://openbabel.sourceforge.net; Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J., Willighagen, E. L. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.

[355] Borini, S., Monari, A., Rossi, E., Tajti, A., Angeli, C., Bendazzoli, G., Cimiraglia, R., Emerson, A., Evangelisti, S., Maynau, D., Sanchez-Marin, J., Szalay, P. *J. Chem. Inf. Model.* **2007**, *47*, 1271–1277.

There are also parallel efforts to convert log files into a more easily parsed format, principally XML, and to manage data from calculations using databases; for details refer to the references given in the article describing cclib.[348]

cclib differs from these projects in that it does not rely on an additional file format or database, and is aimed at users who are developing algorithms or simply need to extract data from log files, although enabling workflows is a possible application. In addition, cclib does not require any changes to be made to the underlying code which, in a field where most computational chemistry codes are proprietary, would be of limited use.

As a final note, it should be stressed that the task of parsing information from output files would be made much easier if information were written using a standard machine-readable format, one that adheres to agreed conventions by means of dictionaries and ontologies so that information and units would be trivial to extract. This is an approach that is gaining momentum in the area of computational materials science, where XML output is in development for packages such as GULP and CRYSTAL. However, until this approach becomes widespread in the computational chemistry community, libraries such as cclib insulate users from the differences between the output files of various packages.

Glossary

Key nomenclature

electrostatic	related to the static distribution of electron charge density around a molecule
interaction	the mutual influence of two molecules, compared to their isolated states
intercalation	insertion of aromatic molecules or their chromophores between nucleic acid bases
multipole moment	approximate point representation of a surrounding charge density distribution
noncovalent	pertaining to a stable complex without the formation of a chemical bond
nonempirical	derived from first principles without assuming arbitrary numerical parameters
stacking	parallel alignment of two aromatic molecules, with significant π orbital overlap
transferability	the process of reusing an atom or fragment in a different molecule or setting

Abbreviations used

AFM	atomic force microscopy
AIM	(quantum theory of) atoms in molecules
AMM	atomic multipole moment
API	application programming interface
BSSE	basis set superposition error
CAMM	cumulative atomic multipole moment
CBS	complete basis set (limit)
CCD	coupled cluster with doubles
CCSD	coupled cluster with singles and doubles
CCSD(T)	coupled cluster with singles, doubles and perturbative triple excitations
CHA	chemical Hamiltonian approach
CNDO	complete neglect of differential overlap
COM	center of mass
CP	counterpoise (correction)
DCBS	dimer centered basis set
DFT	density functional theory
DMA	distributed multipole moment
DMPF	*O,O*-dimethylphosphorofluoridate
DNA	deoxyribonucleic acid
EDA	(variational) energy decomposition analysis
EHT	extended Hückel theory
FCI	full configuration interaction
HF	Hartree-Fock
HVPT	hybrid variation perturbation theory
KM	Kitaura-Morokuma
LCAO	linear combination of atomic orbitals
MCBS	monomer centered basis set
MD	molecular dynamics (simulations)
MEP	molecular electrostatic potential
MP2	second-order Möller-Plesset theory
NAC	near attack conformer
PES	potential energy surface
QM/MM	quantum mechanics/molecular mechanics
RHF	restricted Hartree-Fock
RMS	root mean square (deviation)
RNA	ribonucleic acid
SAPT	symmetry-adapted perturbation theory
SCF	self-consistent field

Software used

ADF	Amsterdam Density Functional[356]
Avogadro	An advanced molecular editor[357]
BSE	Basis Set Exchange[358]
EDS	Energy Decomposition Scheme, an implementation of HVPT by Góra[359]
GAMESS	General Atomic and Molecular Electronic Structure System[360]
Gaussian03	General program for electronic structure modeling[361]
Molden	General purpose program for molecular structure[362]
NumPy	Numerical Python package[363]
Python	A general-purpose high-level programming language[347]
Reduce	Program used for adding missing hydrogen atoms to crystal structures[364]
SAPT	Symmetry-Adapted Perturbation Theory[365]
VMD	Visual Molecular Dynamics[366]

[356]Velde, G., Bickelhaupt, F. M., Baerends, E. J., Guerra, C. F., Gisbergen, S., Snijders, J. G., Ziegler, T. *J. Comp. Chem.* **2001**, *22*, 931–967.

[357]Avogadro*, an advanced molecular editor*; http://avogadro.openmolecules.net.

[358]EMSL Basis Set Exchange, see Schuchardt, K. L., Didier, B. T., Elsethagen, T., Sun, L., Gurumoorthi, V., Chase, J., Li, J., Windus, T. L. *J. Chem. Inf. Model.* **2007**, *47*, 1045–1052.

[359]Góra, R. W. *EDS: Energy Decomposition Scheme*, Wrocław, Poland, Jackson, MS; 1998-2009.

[360]GAMESS-US, see Schmidt, M. W. et al. *J. Comp. Chem.* **1993**, *14*, 1347–1363.

[361]Frisch, M. J. et al. *Gaussian 03, Revisions C.02, D.01 and E.01*; 2004, http://www.gaussian.com.

[362]Schaftenaar, G., Noordik, J. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123–134.

[363]Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9*, 10–20.

[364]Word, J. M., Lovell, S. C., Richardson, J. S., Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1735–1747.

[365]Bukowski, R. et al. *SAPT2008: Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies*; 2008, http://www.physics.udel.edu/~szalewic/SAPT.

[366]Humphrey, W., Dalke, A., Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.

List of Tables

List of Figures

Author index