Due to copyright issues this PDF file has all scientific articles removed from Appendixes I-IV

Algorithms and Software Tools for Exploration of Chemical Spaces, Charge Distributions, and Solvation Effects. Applications thereof to DNA Fragments.

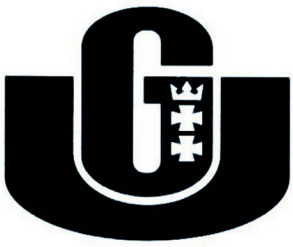
A Study Submitted in Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

at
The Univeristy of Gdańsk

by
Maciej Harańczyk

The Study Prepared under Supervision of

Prof. Maciej Gutowski



Department of Chemistry

Gdańsk, 2008

This Work would not be possible without direct financial support from the following:

- Polish Ministry of Science and Education – Grant N204 127 31/2963
- European Union COST Program P9
  - Short-Term Scientific Mission (STSM) P9-02569 and P9-02841
- British Council and Polish Ministry of Science and Education
  - Young Scientists Programme Grant WAR/342/86

This Work has been also indirectly funded by the following:

- US Department of Energy
- US National Institutes of Health

The following organizations funded fellowships that supported the Author:

- Polish Science Foundation (FNP)
  – START Program
- Foundation of the Development of University of Gdańsk (FRUG)
  – Fellowship for PhD students
- Department of Chemistry, University of Gdańsk
  – Fellowship for PhD students
- European Union Social Funds (EFS)
  - Grant EFS ZPORR/2.22/II/2.6/ARP/U/2/05

The research presented in this Dissertation has been performed in the following computer facilities:

- Molecular Science and Computing Facility (MSCF) located at the Environmental Molecular Sciences Laboratory (EMSL) operated by Pacific Northwest National Laboratory (PNNL) for US Department of Energy
- High Performance Computer Center (KDM) operated by Trójmiejska Akademicka Sieć Komputerowa (TASK)
- National Energy Research Scientific Computing Center (NERSC) operated by Lawrence Berkeley National Laboratory (LBNL) for US Department of Energy
- High Performance Computing and Communication Center (HPCC) of the University of Southern California

Acknowledgements

First I would like to thank my supervisor – Professor Maciej Gutowski for his help, guidance, and encouragement throughout the seven years we worked together. Maciej has not only been my scientific advisor but also wonderful Mentor who always had time and patience for me. Thank you Maciej!

Secondly I would like to thank my hosts and supervisors during my research visits who provided unvalued contribution by the guidance and expertise: Dr. John Holliday and Prof. Peter Willett from University of Sheffield for introducing me to the state-of-the-art chemoinfomatics techniques, Prof. Arieh Warshel from University of Southern California for teaching me secrets of solvation and more importantly, to critically look at my and other people's work. In addition I would like to thank my Master's thesis advisor Prof. Janusz Rak for his support and advice during the time of my graduate school. I would also like to mention people who contributed to my scientific growth: Dr. Karol Krzymiński, Dr. Bogusław Dręczewski, Prof. Jerzy Błażejowski, Dr. Paul Ruttink, Dr. Aleksander Herman, Dr. Marcin Dębowski and Dr. Stanisław Ołdziej.

Next I would like to thank Rafał Bachorz for his friendship that contributed significantly to my research and scientific development. Many thanks go to other my younger collaborators: Sanliang Ling, Giovanni Lupica, Iwona Dąbkowska, Shina Lynn Kamerlin and Tomasz Puzyn.

I would also like to thank people that I met during last few years and who had big influence on my life, and therefore directly or indirectly influenced the Dissertation: Ya-Huei Chin, Sebastian Dejryng, Anita Lagutschenkov and Yogendra Patel. Thank you for inspiration and friendship!

Moreover I would like to mention few friends who supported me during the time of graduate school: Pankaz Sharma (University of Southern California), Dariusz Sobolewski, Monika Kobyłecka, Kamil Mazurkiewicz and Piotr Storoniak (University of Gdańsk), Richard Martin, Iain Mott, George Papadatos, Jerome Hert

and David Wood (University of Sheffield) and Danny Chang (Environmental Protection Agency).

I would also like to give a special thank you to my parents for their support throughout my PhD, and all the years in education.

Maciek

Abstract

The overall aim of this Dissertation is to demonstrate the methodology, tools and approaches we developed to advance research on fragments of DNA. Our contributions include: (i) a combinatorial-computational exploration of chemical spaces in order to identify the "fittest" molecules (e.g., searches for the most stable tautomers of a molecule); (ii) approaches to analyze vast amount of data harvested in quantum chemical calculations of (i); (iii) algorithms and software tools to improve visualization of molecular orbitals and related electron densities; (iv) approaches to improve efficiency of the combinatorial-computational searches for the most stable tautomers by using partial information on the studied chemical space; (v) a methodology to predict accurate solvation free energies of molecules.

These approaches and tools were applied to all nucleic acid bases, but the discussion in this Dissertation will be limited to guanine and uracil. In the case of guanine, we used the approach (i) to perform energy based screening of a library of combinatorially generate 499 tautomers. We discovered 13 adiabatically bound anions of guanine (so far guanine was assumed not to bind an electron) that might be involved in the processes of DNA damage by low-energy electrons and in charge transfer through DNA. These anions correspond to some tautomers that have been ignored thus far.

By using the tools of (ii)-(iv) we concluded that the high stability of adiabatically bound anions originates from the bonding character of π orbitals occupied by the excess electron. This compensates for the antibonding character that usually causes significant buckling of the double-ring structure. Also the excess electron is more homogenously distributed over both rings than in the case of anions of the most stable neutral species. In terms of 2D substructure, the most stable anionic tautomers generally have additional hydrogen atoms at C8 and/or C2 and they do not have hydrogen atoms attached to C4, C5 and C6. They also form an "island of stability" in the tautomeric space of guanine. The latter information may be used to improve the efficiency of future searches for the most stable tautomers using the approach of (i).

In the case of anion of uracil, we calculated the solvation free energies of the most stable anionic tautomers. Our results suggest that the few recently discovered gas-phase tautomers are also the most stable one in water solution.

Preface

This Dissertation summarizes four years of my research. As usually when we try to summarize a longer period of time, we find out that not all of our actions are reasonably connected. In my case, the scientific curiosity drove me towards various projects. If they were presented in chronological order, they might seem loosely connected or just "chaotic". But in fact, the "chaotic" period corresponds to a time of looking for my personal scientific Challenge. With a wonderful guidance from my Advisor, I was able to find my Challenge and address it. This Dissertation is therefore the Story on my scientific Challenge.

In a good story-telling, some events have to be skipped or reordered to achieve an attractive, logical and clear presentation. The same happened to my Story. I decided to select the most interesting parts of all graduate projects and put them into a logical order. Some parts of my initial work ended up being just a background information to the main part of the research presented in this Dissertation. Some other parts of my work did not fit into the main theme of the Dissertation, and have to be moved to the Appendices. The latter is presented in a form of journal articles accessible to the interested Readers.

Eventually, the Dissertation is organized as follows. In the Introduction chapter, I will point out main aims of my work. In the second chapter, I will summarize the background information underlying my research. It will include the biochemical section that will help the Reader to see the presented results in a broader context. It will also include a brief summary of the research methodology used. The third and the most important chapter of this Dissertation is "Methodology developed by Author", in which I describe the most relevant methods, tools and approaches I have developed. The fourth chapter demonstrates their application in the studies of two nucleic acid bases, guanine and uracil. The fifth chapter summarizes the work described in this Dissertation and draws conclusions. In the last chapter I will present few remarks on how my work could be extended in the future into different areas of chemistry and materials science.

Table of Contents

List of Publications

List of publications that result from this dissertation project:

1. **M. Haranczyk**, M. Gutowski – "Valence and dipole-bound anions of the most stable tautomers of guanine" – *The Journal of the American Chemical Society (JACS) 127 (2005) 699-706*.
2. **M. Haranczyk**, J. Rak and M. Gutowski – "Stabilization of very rare tautomers of 1-methylcytosine by an excess electron" – *Journal of Physical Chemistry A* 109 (2005) 11495-11503.
3. **M. Haranczyk**, M. Gutowski - "Quantum Mechanical Energy –Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program" – *Journal of Chemical Information and Modeling 47* (2007) 686-694.
4. **M. Haranczyk**, M. Gutowski – "Finding Adiabatically Bound Anions of Guanine through Combinatorial-Computational Approach" – *Angewandte Chemie Int. Ed.* 44 (2005) 6585-6588.
5. **M. Haranczyk**, M. Gutowski, X. Li, K.H. Bowen – "Adiabatically bound anions of guanine" – *Journal Physical Chemistry* B 111 (2007) 14073-14076.
6. **M. Haranczyk**, J. Holliday, P. Willett, M. Gutowski – "Structure and Singly Occupied Molecular Orbital Analysis of Anionic Tautomers of Guanine" – *Journal of Computational Chemistry* – in press - DOI:10.1002/jcc.20886 .
7. **M. Haranczyk**, M. Gutowski – "Visualization of molecular orbitals and the related electron densities" –*Journal of Chemical Theory and Computation* – in press.
8. E. Rosta, **M. Haranczyk**, Z.T. Chu, A. Warshel – "Accelerating QM/MM Free Energy Calculations: Representing the Surroundings by an Updated Mean Charge Distribution" –*Journal of Physical Chemistry B* – in press.
9. **M. Haranczyk**, M. Gutowski, A. Warshel – " Solvation free energies of molecules. The most stable anionic tautomers of uracil" – submitted to *Physical Chemistry Chemical Physics* **.**

10. **M. Haranczyk**, M. Gutowski, X. Li, K.H. Bowen – "Bound anionic states of adenine. Theoretical and photoelectron spectroscopy study" –*Proceedings of National Academy of Science (PNAS) 104* (2007) 4804-4807.
11. X. Li, K.H. Bowen **M. Haranczyk**, R.A. Bachorz, K. Mazurkiewicz, J. Rak, M. Gutowski –"Photoelectron spectroscopy of adiabatically bound valence anions of rare tautomers of the nucleic acid bases" – *Journal of Chemical Physics* 127 (2007) 174309.
12. **M. Haranczyk**, J. Holliday – "Comparison of Similarity Coefficients for Clustering and Compound Selection" –*Journal of Chemical Information and Modeling* – in press – DOI: 10.1021/ci700413a .
13. **M. Haranczyk**, T. Puzyn, P. Sadowski – "ConGENER – A Tool for Modeling of the Congeneric Sets of Environmental Pollutants" –*QSAR and Combinatorial Science*- in press.
14. **M. Haranczyk**, M. Gutowski – "Differences in Electrostatic Potential Around DNA Fragments Containing Guanine and 8-oxo-Guanine" – *Theoretical Chemistry Accounts 117* (2007) 291–296.
15. **M. Haranczyk**, J.H. Miller, M. Gutowski – "Differences in Electrostatic Potential around DNA Fragments Containing Adenine and 8-oxo-Adenine. An Analysis Based on Regular Cylindrical Projection" – *Journal of Molecular Graphics and Modelling 26* (2007) 282–289.
16. **M. Haranczyk**, G. Lupica, I. Dąbkowska, M. Gutowski – "Cylindrical Projection of Electrostatic Potential and Image Analysis Tools for Damaged DNA. The Substitution of Thymine with Thymine Glycol" – *Journal of Physical Chemistry* B 112 (2008) 2198-2206 **.**

List of Software

List of software tools developed of the course of this dissertation project:

1. **TauTGen** – Tautomer Generator Program, http://tautgen.sourceforge.net
2. **GOT** - Gaussian Output Tools – Scripting tools to extract results from output files of quantum chemical calculations (Gaussian and NWChem), http://gaussot.sourceforge.net
3. **OpenCubMan** – Open-source Cubefile Manipulator - Tools to manipulate cube file with volumetric data containing molecular orbitals, http://opencubman.sourceforge.net
4. **MHcluster** – A Set of Clustering programs
5. **ConGENER** – Congeners Generator - Tools for computational characterization of congeners, http://congener.sourceforge.net
6. **SEPAP** – DNA Shape and Electrostatic Potential Analysis Program, http://sepap.sourceforge.net
7. **Accelerated QM/MM Module** for MOLARIS package

List of Figures

(see Section 1.5 for definition of figures' notation)

List of Tables

(see Section 1.5 for definition of tables' notation)

List of Abbreviations

AC	Adiabatic Charging
AEA	Adiabatic Electron Affinity
BCI	Barnard Chemical Information
BFPT	Barrier Free Proton Transfer
CC	Coupled Cluster Method
CV	Contour Value
DEA	Dissociative Electron Attachment
DFT	Density Functional Theory
DSB	Double Strand Break
EBE	Electron Binding Energy
FEP	Free Energy Perturbation
HF	Hartree-Fock Method
HGAA	Hierarchical Group-Average Agglomerative clustering
LEE	Low Energy Electrons
NAB	Nucleic Acid Base
PCM	Polarizable Continuum Model
PES	Photoelectron Spectroscopy
PT	Proton Transfer
QC	Quantum Chemistry
QM	Quantum Mechanics
SAHN	Sequential Agglomerative Hierarchical Non-overlapping clustering
SCF	Self-Consistent Field
SE	Secondary Electrons
SOMO	Singly Occupied Molecular Orbital
SSB	Single Strand Break
VAE	Vertical Attachment Energy
VDE	Vertical Detachment Energy

1. Introduction

1.1. Motivation of the Dissertation

What was the motivation for writing this Dissertation? What is the motivation for performing research in general? It is the scientific curiosity. It is the challenge that we find in our work. As personalities of researchers differ, so do the personal challenges and interests. Therefore there is no common answer for all researchers. Each one of us has to identify his/her aims, and later on the way to achieve them. So we spend most of our time searching, for questions and answers. Eventually, in a retrospective view we are able to sum up our experiences and present them as a consistent piece of work.

As comes to my dissertation project, I was initially planning to focus on computational characterization of DNA fragments and their interactions with excess electrons. As I performed the first part of the planned research on the tautomers of anionic guanine, I realized that I did not have the appropriate tools to conduct the desired research. Then I had started to develop the tools that I needed. In the meantime, I realized that tools are our allies (often underestimated) in the quest to understand Nature. Any advancement in research requires better tools, more sophisticated, more accurate and more precise. I discovered that working on the development of new tools is my scientific Challenge. My aim became to develop a complete toolbox required to thoroughly characterize the tautomeric space of anionic nucleic acid bases at the level that has never been done before.

1.2. Subject of the Dissertation

This Dissertation presents novel approaches and software tools that have been developed in order to advance the studies on charged nucleic acid bases. I discuss how to generate and characterize combinatorially generated libraries of tautomers. This includes not only the development of a tautomer generation program but also new methods and tools for the analysis of quantum chemical results for numerous though similar molecules.

The application of developed approaches and software tools has been demonstrated in the studies on anionic guaninie and uracil. In the case of guanine,

we performed energy-based screening of a library of 499 tautomers and we identified 14 stable, unknown before, tautomers. These tautomers have been thoroughly characterized by calculating accurate electron binding energies that helped to interpret experimental results (performed in the group of Prof. Kit H Bowen, John Hopkins University, Balitmore), and by identifying the structural and electronic sources of stability. The approaches presented here led to a conclusion that the most stable tautomers form an island of stability in the tautomeric space. In addition, we advanced methodologies for calculations of free energies of solvation and we characterized of the most stable anionic tautomers of hydrated uracil.

1.3. Goals of the PhD Project

The goals of the PhD project can be summarized in the following points:

- develop approaches for systematic, combinatorial explorations of chemical spaces, the tautomeric space in particular. This involves developing software for combinatorial generations of molecular libraries.
- develop an improved visualization of molecular orbitals and the related electron densities. This involves developing software for handling molecular orbitals and the related electron densities represented as volumetric data.
- develop methods for the analysis of quantum chemical results produced when screening the cominatorially generated libraries of molecules.
- develop improved methods for determination of accurate solvation free energies of molecules. This involves developing a quantum mechnical/molecular mechanics methodology.
- demonstrate performance of these approaches and software tools in the studies on tautomers of anionic nucleic acid bases, the anions of guanine and uracil in particular.

1.4 Language Issues

English is without no doubt the language of the scientific community and a common standard for writing graduate thesis in many of the European Union countries. I decided to write this Dissertation in English, because a part of my

research was conducted during research visits in foreign institutions. By presenting this Dissertation in English, I would like to make my work accessible to my Hosts, their students, my collaborators and the institutions that supported the research visits through dedicated grants.

When describing the methodology I developed, or reporting the results I obtained, I am using mostly the active voice form, and the plural "we" instead of singular "I". The latter is reserved to highlight my personal comments and opinions.

As far as selection of English is concerned, I am using American English. However, the only one exception is made for the British word "chemoinformatics" (instead of American "cheminformatics"). Recently David Wild suggested to use 'chem(o)informatics' as a compromise, but for a simplicity I will stay with the British form [1].

1.5 Notation for Figures, Tables and Equations

Due to the size and hierarchical, multi-sectional structure of this Dissertation we use the following, relatively extensive notation to distinguish figures, tables and equations, which are introduced throughout the text. The notation A-x.y-z is defined, where A is T, F or E for tables, figures and equations, respectively. The x and y denotes the chapter and subchapter, where either a table, a figure or an equation is introduced. The z is the numbering within subchapter defined by x and y. For example, T-4.2-3 is the third table introduced within Section 4.2 of the Dissertation.

2. Background Information

2.1 Biochemical Aspect

2.1.1 Introduction

This year we are celebrating the 55th anniversary of decoding the DNA structure by Francis H.C. Crick and James D. Watson [2]. Since then, researchers have significantly enhanced our knowledge of the structure and the function of this biopolymer. We understand now how does the DNA replicate, how does it transcribes on RNA and how it is translated into proteins. All of these processes are of the key importance for life. The cells have to constantly struggle to sustain the "normal" state of living. There are agents of all sorts and sizes that try to interfere with and/or disrupt this "normal" state. These agents include viruses, foreign proteins, biotoxic molecules and a high energy radiation. As will be demonstrated in the next section, the latter interacts with the cell environment to produce low energy electrons. Then, these secondary products might interact with DNA causing single and double strand breaks (SSB and DSB), respectively. We would like to understand these processes. In our work we apply the "from the detail to the wider perspective" approach [3]. As will be shown in the following section, we learned from experimental studies that the SSB and DSB processes run through anionic states localized on the nucleic acid bases. In the framework of this Dissertation we focus our attention on anionic states of selected nucleic acid bases and on tautomerization reactions that might occur upon the excess electron attachment. Such narrow selection has its reasons. The aim of this Dissertation is to describe novel methodologies and illustrate their performance. To avoid a bias towards biochemical applications of the methodology, we will limit the Background Information part to the facts that are absolutely necessary to justify the selection of presented applications.

The Reader interested in interactions of low energy electrons with DNA is directed to our recent review article [4]. This article is an up-to-date summary on the role of anionic nucleic acid basis in the radiation damage of DNA, and therefore it will be widely cited here.

2.1.2 High Energy Radiation, Secondary Electrons and their Interaction with the DNA through Nucleic Acid Bases

Recent years brought advancements in our understanding of the effects of ionizing radiation on biological systems [5,6]. DNA has been of particular interest since it is responsible for storing and processing of genetic information. It also is susceptible to damage by high energy radiation. It was initially assumed that the DNA damage occurs as a result of ionization via direct impact of high-energy quanta. In the theoretical study of Nikjoo et al. [7] the probabilities for the formation of photon-induced single- and double-strand breaks in DNA were examined. They suggested that the minimum photon energy needed to produce SSBs and DSBs is as much as 20 and 50 eV, respectively. However, later on Prise et al. [8] reexamined the estimations of the Nikjoo et al. through experimental studies where samples of dry plasmid DNA were irradiated with photons of energies in the 5-200 eV range. It was demonstrated that damage occurs at photon energies as low as 7-8 eV. The discrepancy between the experimental and calculated threshold energies for strand break formation originated from the deficiencies of Nikjoo's model.

About 20% of the energy deposited by high-energy particles in cellular material leads to the electronically excited species, which may stabilize themselves via hetero- or homolytic dissociation, whereas the remaining energy induces ionization in the cellular material [5]. "As a consequence, ionizing radiation interacts with DNA primarily via products of its interaction with cellular environment [9]. Since water is the most ubiquitous component in all biological systems, most of the high energy radiation absorbed by living matter induces water radiolysis (generation of hydroxyl and hydrogen radicals) and the formation of secondary low-energy electrons (LEEs) [10]. LEEs are formed with the yield of ca. $4x10^4$ per MeV of incident radiation [5,11]. The secondary electron (SE) energy distribution has a maximum around 9-10 eV [12]. It was, however, unclear if such low-energy SEs are able to induce genotoxic damage (SSBs and DSBs) in DNA. To be specific, other secondary species, such as hydroxyl radicals, are known to be highly genotoxic [13,14]. Indeed, abstraction of deoxyribose hydrogen atoms by $OH^{\bullet}$ radicals, formed through water homolysis by ionizing radiation, initiates at least one pathway which ends with the

production of a DNA strand scission [13]. In Figure F-2.1-1 the efficiency of DSB formation (in terms of the percent content of the linear forms of DNA determined with gel electrophoresis) induced by 8.5 eV photons in the water solution of DNA is displayed [15]. Two variants of this experiment were performed – with and without radical (OH•/H• atoms) scavengers – and their results allow one to draw the conclusion that low energy electrons themselves are able to generate DNA strand breaks.

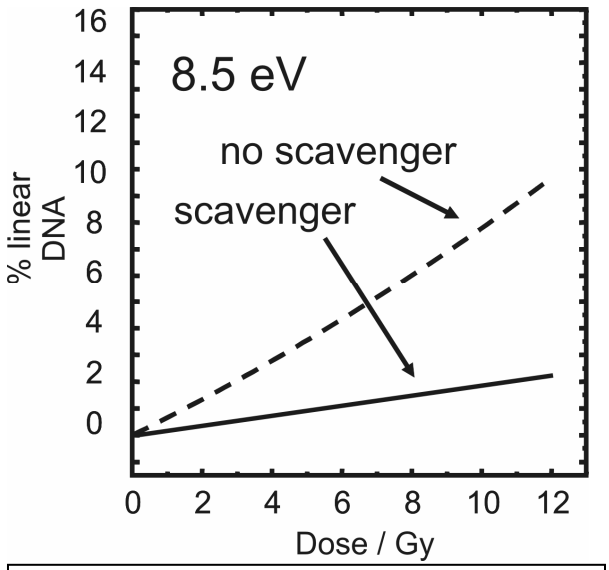


Figure F-2.1-1. The effect of scavenger on the DSBs yields in DNA triggered by photons of 8.5 eV (Source: Ref. 15 and 4).

Plasmid DNA was first bombarded with electrons of energies lower than 100 eV by Folkard et al. [16] who found threshold energies for SSB and DSB at 25 and 50 eV, respectively. Taking into account the fact that the majority of electrons formed within water radiolysis possess energies well below 30 eV, their finding suggested that LEEs are not necessarily an important factor in DNA damage. The paramount role of low energy electrons in the nascent stages of DNA radiolysis was only demonstrated by the pioneering works of Sanche and co-workers [5,6]. In 2000 they published results of their seminal experiments concerning the irradiation of the thin layers of plasmid DNA with electrons of precisely determined energy [17-19]. Using gel electrophoresis to study irradiated samples they demonstrated unequivocally that electrons of sub-ionization energies (i.e. of energies lower than the ionization potential of DNA which are between 7.5 and 10 eV [18]) are capable of producing SSBs and DSBs in DNA (see Figure F-2.1-2). The incident electron energy dependence of damage to DNA was recorded between 3–100 eV in the single-electron regime [19]. The SSB yield threshold was registered near 4–5 eV (due to the cut-off of the electron beam at low energies [6]) whereas the DSB yield begins near 6 eV. Both yield functions possess a strongly structured pattern below 15 eV, have a peak around 10

eV, a pronounced minimum near 14–15 eV, a rapid increase between 15 and 30 eV, and above 30 eV roughly constant yields up to 100 eV.

Above 15 eV the mechanism of chemical bonds dissociation in DNA irradiated with LEEs is probably dominated by direct excitation of dissociative electronically excited states [20]. On the other hand, at lower energies the cleavage process is due to the formation of transient resonance anions [5,6,20-23]. Thus, the SSB and DSB maxima on the yield function observed around 8 and 10 eV (see Figure F-2.1-2), respectively, may be interpreted as originating from resonance anions. The strand break yield as a function of electron impact energy peaks near the threshold for electronic excitation of DNA constituents which suggests that the cleavage process induced by electrons of 8-10 eV is initiated through the short-lived core-excited anion states [21]. The core-excited resonances usually have relatively long lifetimes which promote their dissociation [20]. Therefore, these species should play a key role in the direct dissociative electron attachment (DEA) process. Indeed, electron stimulated desorption (ESD) of anions from the LEE (3-20 eV) irradiated samples of plasmid and synthetic 40-base pair DNA duplex displayed maxima in the yield function of H^-, O^-, and OH^- around 9 eV [22]. The latter value falls in the 8-10 eV range where the main features in the yield functions of strand-break formation in DNA films are located (see Figure F-2.1-2). Thus, the ESD experiments together with the detection of SSBs and DSBs in damaged DNA samples suggest that core-excited resonances might decay in two ways: (i) via the direct DEA process that becomes a

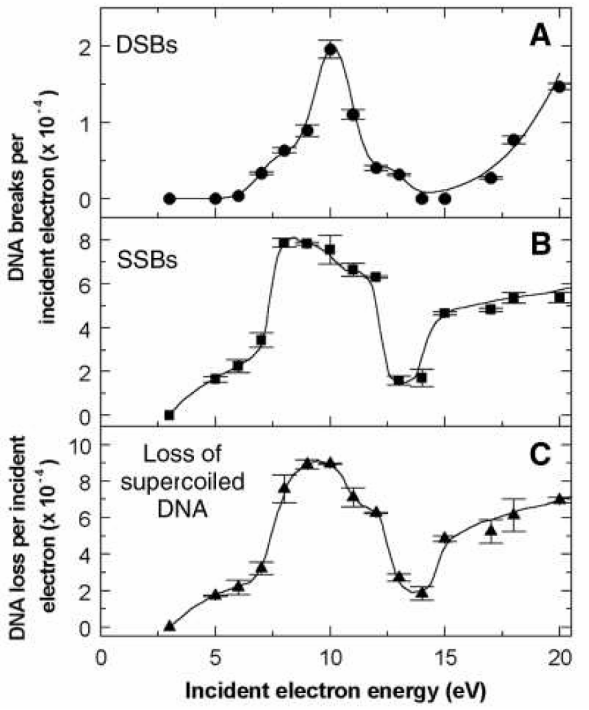


Figure F-2.1-2. Measured quantum yields, per incident electron, for the induction of DSBs (A), SSBs (B), and loss of the supercoiled DNA form (C), in DNA solids by low-energy electron irradiation as a function of incident electron energy (Source: Figure 1 of Ref. 17).

source of small molecular fragments desorbed into the gas phase, and (ii) through electron transfer to the phosphate group which in the next step(s) leads to the formation of SSB. Comparing the yield functions of H^- registered in the ESD experiments on DNA films [22] with that from ESD on films containing nucleobases [24], amorphous ice [25] and deoxyribose analogs [26] it was demonstrated that LEE-induced H^- desorption from DNA below 15 eV occurs mainly via DEA to nucleobases with some contribution from the deoxyribose ring [6]. Hence, in that energy range nucleobases seem to be primary targets for the interaction of LEEs with DNA. "[4].

2.1.3 Mechanisms for Strand Breaks Formation

The mechanism of single and double strand breaking initiated by low-energy electrons has been extensively studied by several experimental and theoretical groups. The possibility of electron transfer from a nucleobase to the phosphate group was indicated in a number of experimental studies. The resonance character of the damage yield function suggests that electron transfer might proceed directly from a resonance anion. This possibility was exploited in a series of papers from the Simons group [27-29]. In their "model the rate of SSB formation has to compete with short lifetimes of resonance states. Thus very low barriers are required to explain the SSB yield observed experimentally [5,6]. " [4].

An alternative for the nonadiabatic mechanism proposed by the Simons group could be a mechanism based on the formation of a stable anionic species localized on the nucleic acid bases [30-33]. As will be shown in the following sections, nucleic acid bases in the condensed phase, or even when solvated by molecules exhibiting proton-donor properties, support bound valence anions. Therefore long-living valence anions rather than metastable resonances are expected in these environments.

"So far, two different mechanisms of single strand break formation based on adiabatically stable anions have been proposed. The first mechanism, suggested by the Leszczynski group [32], assumes the formation of stable anions of 3'- and 5'-phosphates of thymidine and cytidine in which the cleavage of the C-O bond take

place via the S_N2-type process. The second reaction sequence, proposed by Dąbkowska et al. [30], starts from the electron induced barrier-free proton transfer (BFPT) process followed by the second electron attachment to the pyrimidine nucleobase radical, intramolecular proton transfer, and the C-O bond dissociation. In both mechanisms the bottleneck step is associated with very low kinetic barrier which enables the SSB formation to be completed in a time period much shorter than that required for the assay of damage." [4]

2.1.4 Characterization of Anions of Nucleic Acid Bases

2.1.4.1 Nucleic Acid Bases and Quantities of Interest

As shown in the previous sections, the anionic states localized on nucleic acid bases might be involved in the strand break reactions initiated by the low energy electrons. Moreover the charged nucleic acid bases play a key role in the electron and hole transfer phenomena in DNA biopolymer [34-37]. Following the "from the detail to the wider perspective" approach, mentioned in the Section 2.1.1, many research efforts were directed towards characterization of the anions of isolated NABs. The quantities that describe electron affinity are:

- Adiabatic Electron Affinity (AEA), defined as a difference in energies of the anion in its equilibrium geometry and the corresponding neutral in its equilibrium geometry (positive value means a bound anion)
- Vertical Detachment Energy (VDE) for the anion, defined as a difference in energies of the neutral in the geometry of the anion and the anion in its equilibrium geometry (positive value means a bound anion)
- Vertical Attachment Energy (VAE) for the neutral, defined as a difference in energies of the anion in the geometry of the neutral and the neutral in its equilibrium geometry (positive value means a bound anion)

Another issue related to characterization of nucleic acid bases is the possibility of their existence in various tautomeric forms. In the DNA, the canonical tautomers of NABs are favored through the stabilization from the environment. However, the formation of other, so called 'rare', tautomers is still possible and has been suggested as a source of point mutations [38]. In turn, in the case of each isolated NAB, there are a few tautomers with the energy close to the most stable tautomer (which does

not have to the canonical one). Typical low-energy tautomers are amino-oxo, amino-hydroxy, imino-oxo and imino-hydroxy. The existence of the low energy tautomers was reported for uracil [39-44], thymine [42,44-48], cytosine [49], adenine [50] and guanine [51-56]. In conditions where various tautomers coexist, an excess electron can be attached to any tautomer that supports a bound state. Indeed, such an observation has been reported for cytosine [57] and will be considered for guanine in the Results Section.

2.1.4.2 Brief Historical View and the Dipole-Bound Anions

In the past, a large number of theoretical and experimental studies were focused on determination of electron affinity of NABs in the gas- or condensed phases. Anions of hydrated NAB's are believed to support an excess electron on a valence-type molecular orbital as suggested by many experimental [58] as well as theoretical data [59]. However, the existence of stable anions of NAB's in the gas phase has long been a point of discussion. Computational studies conducted as early as in 1960's had predicted negative values of adiabatic electron affinity (AEA) [60-62]. An important development occurred in mid 90's when Adamowicz and co-workers found stable but loosely bound anionic states supported primarily by the large dipole moments of neutral NAB's [63-65].

These insightful theoretical predictions were followed by experimental studies aiming to characterize anions of NAB's in the gas phase. Bowen and co-workers studied uracil and thymine by negative ion photoelectron spectroscopy (PES) [66]. The adiabatic electron affinities were found to be 93±7 meV for uracil and 69±7 meV for thymine and were assigned to dipole-bound states. Desfrancois, Abdoul-Carime, and Schermann produced gas-phase anions of NAB's in charge-exchange collisions with laser-excited Rydberg atoms and reported the following values of AEA: 54±35 meV for uracil, 68±20 meV for thymine and 12±5 meV for adenine [67]. Schlag and co-workers presented photodetachment-photoelectron spectra of the pyrimidine NAB's [57]. They found a dipole-bound state of uracil at 86±8 meV, thymine at 62±8 meV, and cytosine at 85±8 meV.

2.1.4.3 Stabilization of Valence Anions Upon Solvation

An important evolution of anions of NAB's occurs upon solvation, as reported by Bowen and co-workers [68]. A transformation from the dipole-bound to valence anion of uracil was demonstrated upon solvation by a single noble gas atom or water molecule (Figure F-2.1-3). Their photoelectron spectra show that complexes of uracil with argon or krypton support only dipole-bound anions (F-2.1-3 a-c), but a complex of uracil with a more polarizable xenon atom can support both a dipole-bound and valence anionic state that can be distinguished easily by different values of electron vertical detachment energy (VDE) and the shape of the PES feature (F-2.1-3 d). The anionic complex of uracil and water was found to exist only as a valence anion (F-2.1-3 e). This study has convincingly demonstrated that anions of NAB's might exist in dipole-bound states in the gas phase but convert to valence anions upon solvation.

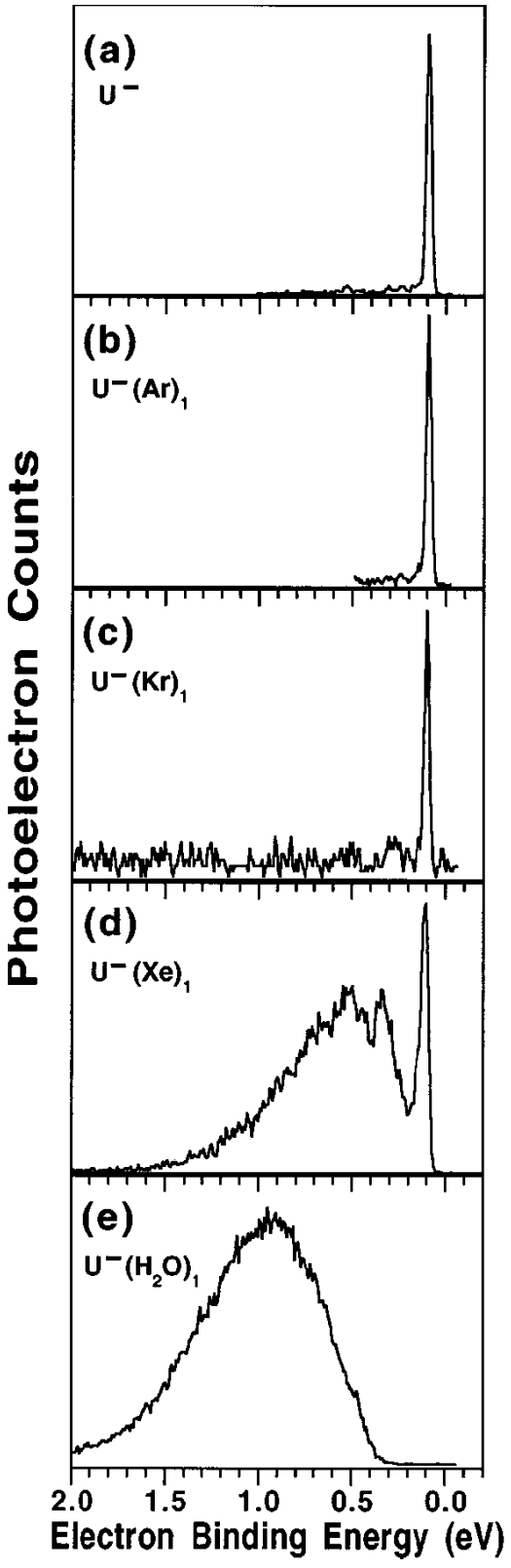


Figure F-2.1-3. Photoelectron spectra recorded using 2.540 eV photons: (a) the photoelectron spectrum of the uracil anion, U^-; (b) the photoelectron spectrum of the uracil anion solvated by an argon atom, $U^-(Ar)_1$; (c) the photoelectron spectrum of the uracil anion solvated by a krypton atom, $U^-(Kr)_1$; (d) the photoelectron spectrum of the uracil anion solvated by a xenon atom, $U^-(Xe)_1$; (e) the photoelectron spectrum of the uracil anion solvated by a water molecule, $U^-(H_2O)_1$.

Solvated NABs anions might be further stabilized by a proton transfer from the solvating molecule. In 2002 Gutowski et al. reported that the excess electron attachment to the complex of uracil with glycine triggers the barrier-free proton transfer reaction (BFPT) [69]. The transferred proton stabilizes the excess electron localized on the π^* orbital of uracil. In general, the reaction may be written as:

$$(NAB...HA) + e \xrightarrow{BFPT} NABH^{\bullet}...A- \qquad \text{(E-2.1-1)}$$

The BFPT reaction as a way of stabilization of anions has been intensively studied in our group. We showed that the BFPT reaction takes places in anionic complexes of thymine, cytosine and adenine [70,71]. Carboxylic acids, inorganic acids and alcohols might be valid proton donors in the BFPT reaction providing that they have appropriate deprotonation energy and ability to form hydrogen bonds [72,73]. A review of the BFPT reactions in the anionic complexes of NABs is presented in Ref 4.

2.1.4.4 Valence Anions of Nucleic Acid Bases

2.1.4.4.1 General Overview of all Nucleic Acid Bases

In this section we will review the research on the anions of all isolated nucleic acid bases, with a particular stress on guanine. In the next two sections we will summarize the most recent studies on the anions of uracil and guanine conducted in our group. The presented most accurate results for these two NABs will be extensively used in the chapter Results Obtained by the Author and Discussion.

From the experimental side, the valence anionic states of NAB's were probed in low-energy electron transmission spectroscopy experiments of Burrow and collaborators [74]. They reported a vertical attachment energy of -0.22, -0.29 and -0.32 for, respectively, uracil, thymine and cytosine. For purine bases, adenine and guanine, they reported VEA of -0.54eV and -0.46 eV, respectively. The latter was assigned to an enol (amino-hydroxy) tautomer of guanine. The negative values mean that the probed anionic state is unbound with respect to the neutral at the optimal geometry of the neutral (see Section 2.1.4.1).

The results on bound anionic states of NABs are very limited in the literature. In the already mentioned Rydberg spectroscopy study of Desfrancois et al., it was suggested that uracil and thymine have positive electron affinities for valence anionic states [67]. Schiedt et al. extrapolated the electron affinity of isolated cytosine from the results obtained for hydrated cytosine with the final prediction being 130 ±120 meV [57]. Many efforts to determine properties of anionic guanine in the PES and RET experiments failed because guanine readily decomposes at elevated temperatures [75]. Guanine and other NAB's also undergo dissociation upon an excess electron attachment. Guanine, however, dissociates into molecular fragments whereas other bases undergo primarily a detachment of a NH hydrogen atom [76].

It should be mentioned here that very recently Bowen and co-workers have made a breakthrough in the experimental studies on valence anions of NABs [77]. They constructed the new ion source that produces anionic states that differ dramatically from those generated in their previous studies. The photoelectron spectra of anionic NABs generated by the new source show them to be exclusively valence-bound, whereas the photoelectron spectra of parent NAB anions formed in the old source revealed them to be dipole-bound. The PES spectrum of the guanine anion obtained by Bowen et al. will be discussed in the Results Section.

With the limited experimental data, the valence anionic states of NABs were mainly characterized in theoretical studies. In the case of guanine, the past computational studies were focused primarily on valence anionic states of the canonical tautomer [59,78,79]. The values of adiabatic electron affinity (AEA) were obtained primarily with the density functional theory (DFT) method using different exchange-correlation functionals and basis sets [78,79]. All but one suggested negative values of the AEA, i.e., the anion was less stable than the neutral. The recent study on electron affinities of NAB's questioned the previously reported values of the AEA of guanine [59]. Sevilla and co-workers suggested that inclusion of diffuse functions in the basis set can result in contamination of the valence state with the dipole-bound state making the results unreliable. In their calculations they used only small basis sets, that do not provide a sufficient extendedness to support a dipole-

bound state, and they reported a negative value of the AEA of -0.75 eV for the canonical tautomer of guanine with the B3LYP exchange-correlation functional.

The remaining NABs have been also characterized extensively. A good summary is contained within the studies conducted in the groups of Sevilla [59] and Schaefer [80]. These authors, not only provided the AEA of each NAB, but also consistently compared the relative AEAs. Schaefer et al. reported the values of AEA for isolated bases (in eVs): A=-0.28, G=-0.07, C=0.03, T=0.20 [80]. Sevilla et al reported results, which provide a similar ordering of the AEA values (in eVs): G=-0.75, A=-0.35, C=-0.05, T=0.22 and U=0.20 [59]. According to both studies, the relative electron affinities favor cytosine and thymine over guanine and adenine. It is consistent with the results of an electron spin resonance (ESR) study of the relative distribution of ion radicals formed in γ-irradiated DNA, which suggested that the anion is divided between the pyrimidine bases and that the excess electron is not localized on purine NABs [81].

In the following two subsections we will summarize recent computational studies on guanine and uracil conducted in our group. The results of these studies provide starting points for the research reported in this Dissertation.

2.1.4.4.2 Guanine*

The most recent and accurate computational results for valence and dipole-bound anions of guanine are provided by us [82]. In our study four most stable neutral tautomers, characterized before in the Hobza group [51] and detected in the gas-phase experiments [83,84], have been considered.

The most important results from the point of view of this Dissertation are summarized in Figure F-2.1-4. The canonical tautomer of guanine (G) is not the most stable neutral tautomer in the gas phase. The GN tautomer is ca. 0.6 kcal/mol more stable than G in terms of Gibbs free energy. Neither of these tautomers can support an adiabatically bound valence anion. The calculated AEAs are -0.459 and -0.503 eV for G and GN, respectively (calculated in respect to the neutral canonical tautomer).

* This section summarizes results obtained by the Author. See Appendix I.

Even though G- and GN- anions are adiabatically unbound, they are characterized by positive values of VDE of 0.585 and 0.212 eV, respectively.

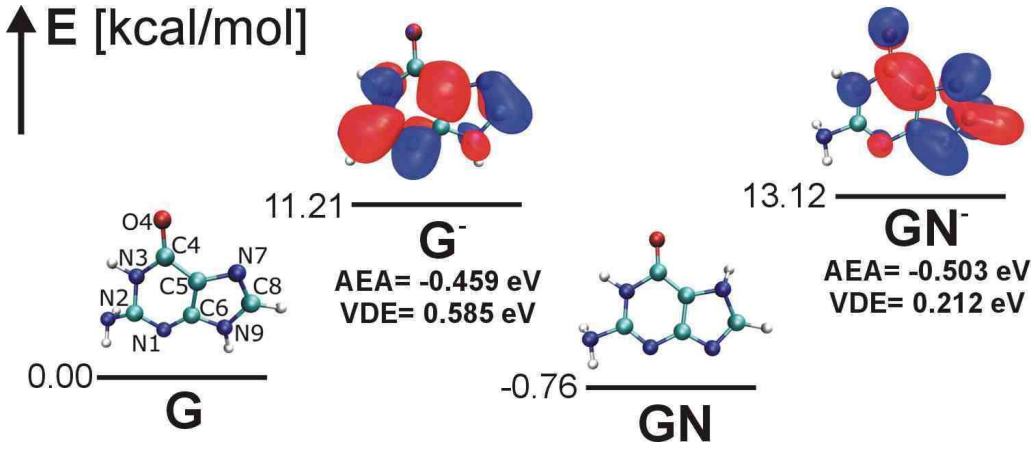


Figure F-2.1-4. Vertically bound valence anions of G and GN.

Our study also characterized the dipole-bound state of canonical guanine, verifying earlier studies of Adamowicz et al. [65]. Since these results are not relevant to the main subject of the Dissertation, we will direct interested Readers to the corresponding article in Appendix I.

2.1.4.4.3 Uracil

The most recent and accurate computational results for valence anions of uracil are contained within a series of recent studies by Bachorz et al. [85-88]. The authors have characterized the anion of canonical tautomer (U0- in Figure F-2.1-5) using the state-of-the-art electronic structure methods, further described in Sections 2.2.2. According to these studies the anion of uracil is adibabatically bound by 40

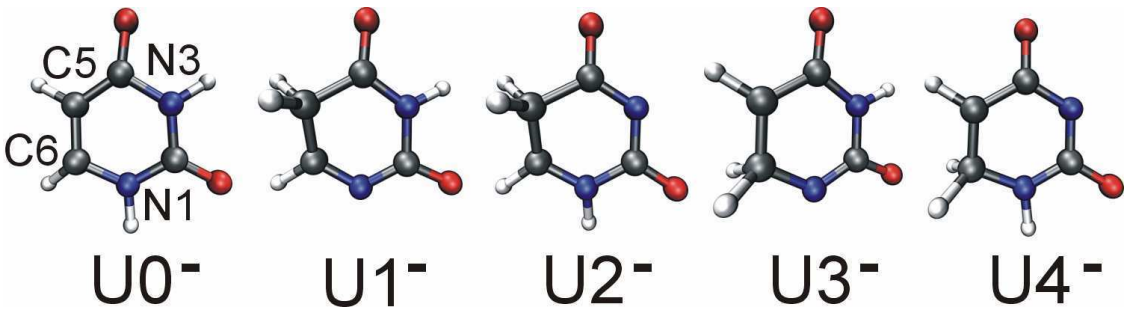


Figure F-2.1-5. The most stable anionic tautomers of uracil identified by Bachorz *et al.* [85,88].

meV and is characterized by a VDE of 0.60 eV [87]. Moreover, based on our experience with proton transfer reactions (E-2.1-1), the authors discovered four important new anionic tautomers, labeled Un- (n=1-4), in addition to the canonical tautomer (Figure F-2.1-5). In comparison with U0-, these new anions have a proton transferred from an NH site to either the C5 or C6 carbon. It is important to note that none of these tautomers belongs to the group of the most stable neutral tautomers mentioned in Section 2.1.4.1. Therefore, we refer to them as "new" tautomers throughout the Dissertation.

The most stable valence anion is $U1^-$, which is more stable than $U0^-$ by ca. 2.5 kcal/mol. The latter is adiabatically bound with respect to the canonical neutral by ca. 0.5 kcal/mol. The valence anions of the remaining tautomers, i.e., $U2^-$, $U3^-$, $U4^-$, are adiabatically unbound with respect to the canonical neutral by 2.0-8.5 kcal/mol, though they are characterized by significant values of VDE of 2.6-3.9 eV. The recently measured photoelectron spectrum of valence anions of U [77] remains in quantitative agreement with the computational predictions.

The results of Bachorz and coworkers will be presented in more detail in Sections 4.2.1-2, when discussing relative free energies of these tautomers in the gas-phase and in water solution.

2.1.4.5 Further Characterization of Anionic NABs and the Missing Tools

In parallel to the studies of Bachorz et al. on the anionic tautomers of uracil [85,88], Mazurkiewicz et al. investigated similar tautomers of anionic thymine [86] and we looked for the most stable anionic tautomers of 1-methylcytosine[†] [89]. For each of pyrimidine bases new anionic tautomers were identified. These finding raised immediate questions on the stability of previously unknown tautomers of anionic purine NABs.

Our initial searches for the most stable anionic tautomers focused on pyrimidine NABs because the number of potentially relevant tautomers was manageable – a few tens of structures. In addition, we had some insights from earlier

[†] See Appendix I.

studies on BFPT (E-2.1-1) which proton donor and proton acceptor sites should be considered. The number of analogous anionic tautomers for purine NABs (guanine and adenine), for which we wanted to perform pre-screening using the density functional level of theory (DFT), was as large as 500-700, as there were no additional suggestions on relevant proton donor and acceptor sites. This is problematic not so much because of the computer time but rather the human time required to prepare, run, and analyze the calculations, which becomes prohibitive. To overcome these limitations, we had to develop a hybrid approach involving both combinatorial and reliable quantum chemical methods. This approach, which is the major development presented in this Dissertation (Sections 3.1 and 3.3), automated the searches for the most stable tautomers of charged purine NABs and facilitated the analysis of generated data.

Another question raised after the discovery of the new tautomers was regarding the stability of these species in condensed phases. Initially the relative stability of hydrated anionic tautomers was estimated at the DFT level with solvent effects approximated through continuum models. These studies exposed the need for approaches that accurately predict the relative stability of important anionic tautomers of NABs in water solution. We addressed this by improving approaches for prediction of free energies of solvation (Sections 3.4). They are based on the microscopic solvent model and quantum mechanical/molecular mechanics (QM/MM) simulations.

2.2 Computational Chemistry

2.2.1 Computational Chemistry Essentials

The fundamental idea behind almost all of the computational chemistry methods is to provide mathematical description of the relation between molecular structure and the energy, or other property. Depending on the studied system, required accuracy of predicted property, and available resources, different models can be selected. In brief, three choices are available‡:

‡ Intentionally, statistical mechanics has not been included

- Molecular mechanics (MM), where interactions between atoms are treated classically with a defined force-field
- Quantum mechanics (QM), where system is defined by a wavefunction or electron density and the energy is obtained by solving the Schrödinger equation
- Mixed models (QM/MM), where two of the above are used to study different fragments of the considered system.

Most of the algorithms and tools being the subject of this Dissertation are based on quantum mechanical methods. Therefore these methods will be briefly summarized in this section. We also use a mixed QM/MM model to determine the free energy of solvation. In this approach solvent molecules are treated classically and a solute molecule is described using QM. A description of this model will be presented in Section 3.4.

The goal of quantum mechanics is to solve the Schrödinger equation:

$$\hat{H}\Psi = E\Psi \qquad \text{(E-2.2-1)}$$

where Ψ is the wavefunction describing the studied system (Ψ is a function of electron and nuclei positions), $\hat{H}$ is the Hamiltonian operator and E is the corresponding energy. In order to obtain a physically relevant solution of the Schrödinger equation, the wave function must be continuous, single-valued, normalizable, and antisymmetric with respect to an interchange of electrons. In principle, solving the Schrödinger equation allows to predict all molecular properties with unlimited precision. However, this is never reached in practice since there are no analytical solutions for most of the systems of interest. A number of approximations have to be introduced in order to obtain the numerical solutions using available computers. The term *ab initio* (Latin for ``from the beginning'') is used to refer to computations that are derived directly from theoretical principles with no inclusion of experimental data.

Solving the Schrödinger equation can be simplified by separating the nuclear and electron motions. This is called the Born-Oppenheimer approximation. When using atomic units, the Hamiltonian for a molecule with stationary nuclei is:

$$\hat{H} = -\sum_{i}^{electrons} \frac{\nabla_i^2}{2} - \sum_{j}^{nuclei} \sum_{i}^{electrons} \frac{Z_j}{r_{ji}} + \sum_{i}^{electrons} \sum_{j<i}^{electrons} \frac{1}{r_{ij}} \qquad \text{(E-2.2-2)}$$

Here, the first term is the kinetic energy of the electrons only. The second term is the attraction of electrons to nuclei. The third term is the repulsion between electrons. The repulsion between nuclei is added onto the energy at the end of calculations. This formulation is the time-independent, nonrelativistic Schrödinger equation. Additional terms can appear in the Hamiltonian when relativity or interactions with electromagnetic radiation or fields are taken into account. The motion of nuclei can be described by considering a potential energy surface determined by the solutions of (E-2.2-2) for various positions of nuclei. The first step is to consider the zero-point motion of nuclei, which is routinely done in harmonic approximation.

Once a wave function has been determined, any property of the individual molecule can be determined. This is done by taking the expectation value of the operator for that property, denoted with angled brackets < >. For example, the energy is frequently given by the expectation value of the Hamiltonian operator

$$\langle E \rangle = \int \Psi^* \hat{H} \Psi d\tau \qquad \text{(E-2.2-3)}$$

The most common type of *ab initio* calculations is performed in the framework of the Hartree-Fock (HF) method, which is based on the central field approximation. This means that the Coulombic electron-electron repulsion is taken into account by integrating the repulsion term. This gives the average effect of the repulsion, but not the explicit repulsion interaction. This is a variational calculation, meaning that the approximate energies calculated are all equal to or greater than the exact energy. One of the advantages of this method is that it breaks the many-electron Schrödinger

equation into many simpler one-electron equations (below an example for an electron "1"):

$$\hat{F}(1)\varphi_i(1) = \varepsilon_i\varphi_i(1) \qquad \text{(E-2.2-4)}$$

Each one-electron equation is solved to yield a single-electron wave function - $\varphi_i(1)$, called an orbital, and an energy - ε_i, called an orbital energy. The orbital describes the behavior of an electron in the average Coulomb-exchange field of all the other electrons expressed in the Fock operator, $\hat{F}$. The Fock operator also contains the kinetic energy operator and the Coulomb interaction with nuclei.

The second approximation in HF calculations is due to the fact that orbitals must be described by some mathematical function, which is known exactly for only a few one-electron systems. In algebraic approximation, one represents orbitals as linear combinations of basis functions. The basis functions, in turn, are represented as linear combinations of Gaussian-type orbitals $\exp(-ar^2)$ centered on the same atom. The orbitals are then combined into a determinant (the Slater determinant). This is done to satisfy a requirement of quantum mechanics, that the wavefunction for electrons must be anisymmetric with respect to any permutation of electrons.

There are variations of the HF procedure depending on the way that orbitals are constructed to reflect paired or unpaired electrons. If the molecule has a singlet spin, then the same orbital spatial function can be used for both the α and β spin electrons in each pair. This is called the restricted Hartree-Fock method (RHF). There are two techniques for constructing HF wave functions of molecules with unpaired electrons. One technique is to use two separate sets of orbitals for the α and β electrons (called spinorbitals). This is called an unrestricted Hartree-Fock wave function (UHF). This means that paired electrons will not have the same spatial distribution. This introduces an error into the calculation, called spin contamination. The amount of spin contamination depends on the chemical system involved. Another way of constructing wave functions for open-shell molecules is the restricted open-shell Hartree-Fock method (ROHF). In this method, the paired electrons share the same spatial orbital; thus, there is no spin contamination.

The major deficiency of the HF method is that it does not include electron correlation (HF only takes into account the average effect of electron repulsion, but not the explicit electron-electron interaction). The concept of correlation energy was introduced by Löwdin [90]: the correlation energy is a difference between the exact nonrelativistic electronic energy and the Hartree-Fock limit (HF energy with a complete basis set). Within this definition, two effects should be distinguished: static correlation and dynamic correlation. Static correlation is present when the system is not well described by one Slater determinant (multi-reference character of the system), while dynamic correlation originates from Coulomb repulsion of electrons (in the HF method the electrons can approach each other closely as they do not "see" each other explicitly).

There are numbers of methods to evaluate the correlation effects, and most of them begin with the HF calculation and then correct it for correlation. Some of these methods are the Møller-Plesset theory [91], configuration interaction (CI) method [91], the generalized valence bond (GVB) method [92], multi-configurational self-consistent field (MCSCF) method [93],and coupled cluster theory (CC) [94]. The first and the last of them will be used within this Study.

In the Møller-Plesset method we improve the Hartree-Fock energy by means of perturbation theory. In the latter we partition the Hamiltonian as:

$$\hat{H} = \hat{H}^{(0)} + \hat{H}^{(1)} \qquad \text{(E-2.2-5)}$$

where $\hat{H}^{(0)}$ is an unperturbed Hamiltonian and $\hat{H}^{(1)}$ is a perturbation. We assume that we know the solution for the unperturbed system:

$$\hat{H}^{(0)}\Psi_0^{(0)} = E_0^{(0)}\Psi_0^{(0)} \qquad \text{(E-2.2-6).}$$

In particular, we assume that in the zeroth-order perturbation, the Hamiltonian is a sum of Fock operators and the zero-th order wavefunction, $\Psi_0^{(0)}$, is the Slater

determinant obtained in the Hartree-Fock calculation ($\Psi_0^{(0)} = \Psi_{HF}$). The zeroth-order energy is then the sum of orbital energies of the occupied spinorbitals:

$$E_0^{(0)} = \sum_i \varepsilon_i \qquad \text{(E-2.2-7)}$$

The sum of the zeroth- and first-order energy is equal to the Hartree-Fock energy:

$$E_0^{(0)} + E_0^{(1)} = E_{HF} = \left(\sum_i \varepsilon_i \right) + \left\langle \Psi_{HF} \middle| \hat{H}^{(1)} \Psi_{HF} \right\rangle \qquad \text{(E-2.2-8).}$$

Electron correlation is added starting from the second-order Møller-Plesset (MP2) level, where the energy is expressed as:

$$E_{MP2} = E_{HF} + \sum_{a<b,p<q}{}' \frac{\left| \langle ab|pq\rangle - \langle ab|qp\rangle \right|^2}{\varepsilon_a + \varepsilon_b - \varepsilon_p - \varepsilon_q} \qquad \text{(E-2.2-9)}$$

where a,b are occupied spinorbitals, p,q are unoccupied spinorbitals and $\varepsilon_a, \varepsilon_b, \varepsilon_p, \varepsilon_q$ are the corresponding orbital energies, respectively. The prime on the summation sign means that the terms that make the denominator equal to zero are excluded.

Third-order (MP3) and fourth-order (MP4) calculations are also encountered in the literature but will not be used in this Study. The perturbation approaches are not variational – the correlation contribution can be overestimated. This fact does not affect their applicability in most of the cases since the method is usually applied to systems where the perturbation (defined as a deviation from the HF description) is relatively small (typically 0.5% in terms of energy). In these cases the second-order perturbation treatment is stable and well behaved.

Another group of methods with growing applications in chemistry are based on the coupled-cluster theory, originally formulated for the problems in nuclear physics. Here, the wavefunction is represented by:

$$\psi = \exp(\hat{T})\Phi_o \qquad \text{(E-2.2-10)}$$

where Φ_o is a reference Slater determinant (it may be obtained by the Hartree-Fock method) and $\exp(\hat{T})$ is an operator that acts on the reference wave function Φ_o and produces the exact waver function. The $\hat{T}$ is a cluster operator, which can be represented as:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \hat{T}_4 + ... + \hat{T}_{n_{\max}} \qquad \text{(E-2.2-11)}$$

where $\hat{T}_k$ (k=1...n_{MAX}) is an excitation operator with k representing the number of excitations. For example, a single excitation operator is defined as:

$$\hat{T}_1 = \sum_{a,r} t_a^r \hat{r}^+ \hat{a} \qquad \text{(E-2.2-12)}$$

and a double excitation operator is defined as:

$$\hat{T}_2 = \frac{1}{4} \sum_{\substack{ab \\ rs}} t_{ab}^{rs} \hat{s}^+ \hat{r}^+ \hat{a}\hat{b} \qquad \text{(E-2.2-13)}$$

where $\hat{x}$ and $\hat{x}^+$ are the annihilation and creation operators, respectively. The t variables, which we try to obtain in the coupled-cluster method (as they define ψ through $\hat{T}$), are called amplitudes. To obtain amplitudes we can consider the Schrödinger equation:

$$\hat{H}\exp(\hat{T})\phi_o = E\exp(\hat{T})\phi_o \qquad \text{(E-2.2-14)}$$

which can be transformed [95] to:

$$\left\{\hat{H}+\left[\hat{H},\hat{T}\right]+\frac{1}{2!}\left[\left[\hat{H},\hat{T}\right],\hat{T}\right]+\frac{1}{3!}\left[\left[\left[\hat{H},\hat{T}\right],\hat{T}\right],\hat{T}\right]+\frac{1}{4!}\left[\left[\left[\left[\hat{H},\hat{T}\right],\hat{T}\right],\hat{T}\right],\hat{T}\right]\right\}\Phi_0 \qquad \text{(E-2.2-15)}$$
$$=E\Phi_0$$

The latter is multiplied by different functions – Slater determinants corresponding to different excitations considered in $\hat{T}$, and then integrated to get a set of nonlinear equations which can be solved iteratively to obtain amplitudes, t. Once amplitudes are obtained we can calculate the energy using:

$$E=\left\langle \exp(-\hat{T}^{+})\Phi_0\middle|\hat{H}\exp(\hat{T})\Phi_0\right\rangle \qquad \text{(E-2.2-16)}$$

In the above formulation the energy is not the mean value of the Hamiltonian and therefore the coupled-cluster method is not variational - it can overestimate the correlation energy. The advantage of coupled-cluster method is that it is size-consistent.

An alternative to the above methods is provided within the density function theory (DFT), which foundations were laid by the work of Hohenberg, Kohn, and Sham [96,97]. The premise behind DFT is that the energy of a molecule can be determined from the electron density instead of a wave function. A practical application of this theory was developed by Kohn and Sham who formulated a method similar in structure to the Hartree-Fock method. In this formulation, the Kohn-Sham orbitals are expressed as linear combinations of basis functions. The occupied orbitals determine the electron. There has been some debate over interpretation of the Kohn-Sham orbitals. It is certain that they are not mathematically equivalent to either HF orbitals or natural orbitals from correlated calculations. However, Kohn-Sham orbitals do describe the behavior of electrons in a molecule, just as the other orbitals mentioned before. DFT orbital energies do not match the energies obtained from photoelectron spectroscopy experiments as well as HF orbital energies do.

Within the orbital expression for the density of the N-electron system, with Kohn-Sham orbitals denoted χ_i, the energy functional can be formulated as:

$$E[\rho(r)] = \sum_{i}^{N}\left(\left\langle \chi_i \left| -\frac{1}{2}\nabla_i^2 \right| \chi_i \right\rangle - \left\langle \chi_i \left| \sum_{k}^{nuclei} \frac{Z_k}{|\mathbf{r}_i - \mathbf{r}_k|} \right| \chi_i \right\rangle\right) + \sum_{i}^{N}\left\langle \chi_i \left| \frac{1}{2}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|} d\mathbf{r}' \right| \chi_i \right\rangle + E_{XC}[\rho(\mathbf{r})] \quad \text{(E-2.2-17)}$$

where the terms on the right hand side refer, respectively, to the kinetic energy of the non-interacting electrons, the nuclear–electron interaction, the classical electron–electron repulsion and the term *Exc*, typically referred to as the exchange-correlation energy. The latter represents the sum of the correction to the kinetic energy deriving from the interacting nature of the electrons, and all non-classical corrections to the electron–electron repulsion energy.

In practice, most modern functionals do not attempt to compute E_{XC} explicitly. Instead, they attempt to construct this function. In many functionals empirical parameters appear. One of the most popular exchange-correlation functionals, and the one being used extensively within this study is the B3LYP functional [98] defined as:

$$E_{XC}^{B3LYP} = (1-a)E_X^{LSDA} + aE_X^{HF} + b\Delta E_X^B + (1-c)E_C^{LSDA} + cE_C^{LYP} \quad \text{(E-2.2-18)}$$

where *a*, *b*, and *c* are 0.20, 0.72, and 0.81, respectively, and the exchange and correlation terms are treated separately. B3LYP is a hybrid functional, which is built from different terms selected in a way to minimize each other deficiencies. The exchange part (marked with x in the subscript) is a sum of terms calculated at different approximations: LSDA, Hartree-Fock exchange terms and Becke's functional, respectively. In local spin density approximation (LSDA) it is assumed that the functional depends on the electron density (Local Density Approximation, LDA) but it is additionally corrected for spin (for open-shell systems). ΔE_X^B represents 'generalized gradient approximation' (GGA), where it is assumed that the functional depends on both the density and the gradient of the density (so it

is 'gradient corrected'). Similarly, the correlation terms, E_C^{LSDA} and E_C^{LYP}, represent the LSDA approximation term and the GGA correlation functional of Lee, Yang, and Parr, respectively.

The main reason for popularity of DFT comes from recovering the correlation effects through a (relatively simple) functional of electron density, which makes their evaluation fast. Another important advantage is that the energies and properties converge to their complete basis set limit relatively fast in comparison with the MPn or CC methods. Thus medium size basis sets provide already very good results.

2.2.2 Selection of Appropriate Methods and Software

The selection of appropriate QM methods to study the desired system is the key to success. What is appropriate is defined by required accuracy and available resources. In the scope of this Dissertation we will practically use only two levels of accuracy depending on application. We will also stick to the popular and well tested methods, for which drawbacks are known and errors can be estimated. These levels are:

- **"Estimate" level,** which is used when the number of systems to be described is large (in the screening of combinatorially generated libraries or sampling in molecular dynamics simulations)
- **"Accurate" level**, which is used to refine the energies of selected system

At our estimate level, the calculations are performed at the DFT level of theory with a B3LYP exchange-correlation functional [98]. The 6-31+G** and 6-31++G** basis sets [99,100] are used. Such a combination is used extensively in the characterization of biomolecules and proved to be a reliable approach [101]. However, in our studies on anionic systems of biomolecules we have observed a tendency of B3LYP to overestimate the excess electron binding energy. For example, for the valence anion of the canonical tautomer of uracil the VDE calculated at the B3LYP/6-31++G** level is overestimated by ca. 0.2 eV in comparison with the coupled-cluster level of theory. This deficiency will be, however, useful in searches for the most stable anionic tautomers (Sections 3.1 and 4.1.1), because it helps to avoid false negatives when screening for adiabatically bound anions.

At the accurate level we use a combination of methods. First, molecular geometries are optimized at the MP2 level of theory with augmented correlation-consistent polarized basis sets of double zeta quality (AVDZ) [102]. The final single-point calculations were performed at the coupled cluster level of theory with single, double, and non-iterative triple excitations (CCSD(T)/AVDZ) [103] at the optimal MP2 geometries. The open-shell CCSD(T) calculations were carried out at the R/UCCSD(T) level. In this approach, a restricted open shell Hartree-Fock calculation was initially performed to generate the set of molecular orbitals and the spin constraint was relaxed in the coupled cluster calculation [104]. The 1s orbitals of carbon, nitrogen and oxygen atoms were excluded from the MP2 and coupled-cluster treatments. The selection of such methods for our accurate level is dictated by the results of the study of Fogarasi, where it was shown that the application of coupled-cluster theory might be required to obtain the correct relative energies of NABs tautomers [49], which is one of the goals of this Dissertation. Moreover, our current experience suggests that the energies (and related VDEs and AEAs) obtained at the R/UCCSD(T)/AVDZ level can be directly compared with the results of photoelectron spectroscopy experiments.

Whenever the energies of zero-point vibrations or thermodynamic corrections to obtain the Gibbs free energy are requited, we calculate them at the MP2/AVDZ level. The thermal corrections as well as the entropy terms, are calculated for T=298 K and p=1 atm in the harmonic oscillator-rigid rotor approximation.

It is also worth to compare the results obtained at our accurate level with the most recent state-of-the-art results. Such a study for the valence anion of uracil has recently been completed by Bachorz and coworkers [87]. They used explicitly-correlated second-order Møller–Plesset perturbation theory (RI-MP2-R12) in conjunction with the conventional coupled cluster method with single, double, and perturbative triple excitations (CCSD(T)) supplemented with basis set extrapolation techniques. The final energies were corrected for zero-point vibration energies, determined in harmonic approximation at the UHF-RI-MP2/aug-cc-pVTZ level of theory. Their best estimate of the VDE is 0.60 eV while 0.51 eV obtained at our

accurate level. A discrepancy of 0.09 eV is not negligible, but not critical for the purpose of this project.

It would be not possible to perform calculation discussed here without appropriate computer programs. We have used NWChem [105], Gaussian03 [106] and Molpro [107]. The latter is used to perform coupled-cluster calculations. The remaining ones are used for all other calculations.

2.3 Chemoinformatics

2.3.1 Chemical Structure Representations

There are various ways to store information about a chemical structure within a computer. The most natural way for computational chemistry methods is to store coordinates of atoms as either Cartesian or internal coordinates [108]. Other ways might include storing as pictures (bitmaps) or text strings (e.g. Simplified Molecular Input Line Entry Specification (SMILES) language [109]). Another way to store a chemical structure, important from the point of applications presented in this Dissertation, is a molecular graph. A graph is an abstract structure that contains nodes connected by edges. In a molecular graph the nodes correspond to the atoms and the edges to bonds. A graph represents the topology of a molecule only, that is, the way the nodes (or atoms) are connected. A subgraph is a subset of the nodes and edges of a graph. A molecular graph can be stored as a connectivity matrix.

Graph theoretic methods can be used to perform substructure searching. A substructure is equivalent to determining whether one graph is entirely contained within another, a problem known as the subgraph isomorphism, also studied in the molecular context [110]. Although efficient algorithms for performing subgraph isomorphism are available, it is often required, when dealing with large molecular databases, to simplify molecular representation even further. It is often done by representing structures by binary vectors, called fingerprints. In the fingerprint vector, each position reflects the presence or absence of a pre-defined substructure or molecular feature. This is often specified using a fragment dictionary. If the i-th fragment from the dictionary is present in the molecule then the relevant (i-th) bit is set to “1” (the fingerprint initially consists of “0”s only). In principle any fragment is

possible, but certain types are frequently encountered. For example, the Chemical Abstracts Service Registry System uses twelve different types of fragments [111]. In this study we will use a format introduced by the Barnard Chemical Information (BCI) company. The substructure fragments used in the BCI fingerprints fall into the following main fragment families:

- Augmented Atom: a central atom and its immediate neighbors, with the connecting bonds
- Atom Sequence: a linear sequence of connected atoms and bonds traced through the molecule
- Atom Pair: two atoms (including details of atom type, number of non-hydrogen connections), and the topological distance between them by the shortest path [112]
- Ring Composition: a sequence of atoms and bonds around a single ring from the Extended Set of Smallest Rings (ESSR) [113]
- Ring Fusion: a sequence of ring-connectivities around a single ring from the ESSR

There has been much work to determine the most effective set of substructures to be used as the fragment dictionary [114,115]. Typically the aim is to select fragments that are independent of each other and that are equifrequent since fragments that are very frequent and occur in most molecules are unlikely to be discriminating and, conversely, fragments that occur infrequently are unlikely to be useful. Thus, frequently occurring fragments are discarded and infrequently occurring fragments are generalized or grouped to increase their frequency. Building of fragment dictionary by detection of fragments in a given structure database, and later creating fingerprints for each molecule is done by application of graphs isomorphism algorithms.

2.3.2 Molecular Similarity

The molecular similarity, and the related dissimilarity, is a very broad topic. For example, comparing the two structures can be done by measuring distances

between corresponding atoms or by analyzing electron density, like in the Carbo method [116]. These methods require to overlay two structures prior to the similarity calculation, which might be a challenge by itself. Many techniques have been introduced to address this problem, including applications of genetic algorithms to find the maximum overlap. This goes, however, beyond the scope of this Dissertation.

In contrast to the mentioned approaches, it is much easier to calculate similarity between molecules represented by fingerprints or other vectors that do not depend on molecular orientation. The most common similarity coefficient, used to measure the similarity between two molecules represented by binary fingerprints, is the Tanimoto coefficient, S_T. S_T is defined:

$$S_T = \frac{c}{(a+b-c)} \qquad \text{(E-2.3-1)}$$

where a and b are the numbers of bits set in the first or the second fingerprint, respectively, and *c* is the number of bits common in two fingerprints. For structures represented by real vectors (such representation will be developed and used in Section 3.1), one can calculate the dissimilarity by using the Euclidean distance, D^{AB}:

$$D^{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2 \right]^{\frac{1}{2}} \qquad \text{(E-2.3-2)}$$

where x_{iA} and x_{iB} are the i-th components of the vectors representing the molecules A and B, respectively. The details of three similarity and dissimilarity coefficients, which are most extensively used in this Dissertation, are summarized in Table T-2.3-1.

For in-depth comparison of the Tanimoto, with 12 other similarity coefficients (as defined in Ref 117, namely: Russell/Rao, Simple Matching, Baroni-Urbani/Buser, Ochiai/cosine, Kulczynski(2), Forbes, Fossum, Simpson, Pearson, Yule, Stiles and

Dennis) in typical applications, we will direct Readers to the article presented in Appendix III.

Table T-2.3-1. Similarity (S) and dissimilarity (D) coefficients in common use. For binary data a is defined as the number of bits set to "1" in molecule A, b as number of bits set to "1" in molecule B and c as the number of bits that are "1" in both A and B.

Name	Formula for continuous variables (real vectors)	Formula for binary variables (fingerprints)
Tanimoto (Jaccard)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA} x_{iB}}$ Range: -0.333 to +1	$S_{AB} = \frac{c}{a+b-c}$ Range: 0 to +1
Euclidean	$D^{AB} = \left[\sum_{i=1}^{N} (x_{iA} - x_{iB})^2 \right]^{\frac{1}{2}}$ Range: 0 to +∞	$D^{AB} = \sqrt{a+b-2c}$ Range: 0 to N
Manhattan (city-block or Hamming)	$D^{AB} = \sum_{i=1}^{N} \lvert x_{iA} - x_{iB} \rvert$ Range: 0 to +∞	$D^{AB} = a+b-2c$ Range: 0 to N

2.3.3 Clustering Methods

Molecules can be grouped according to their similarity by means of clustering methods, providing a similarity measure is defined. There are many clustering methods used in chemical applications (see Refs. 118 and 119 for reviews). The clustering method to be most extensively applied here is a sequential agglomerative hierarchical non-overlapping (SAHN) method [118]. It was implemented by us in our MHcluster program using the stored-matrix algorithm, so named because the starting point is a matrix of all pairwise similarities between molecules in the set to be clustered. Each cluster initially corresponds to an individual structure (singleton). As clustering proceeds, each cluster may contain one or more molecules. Eventually, there evolves one cluster that contains all tautomers. At each iteration, a pair of clusters is merged (agglomerated) and the number of clusters decreases by 1. The various SAHN methods that are available differ in the way in which the similarity between clusters is defined. In the implementation considered in this Dissertation,

we calculate it as the arithmetic average of similarities between all pairs of molecules (which we refer to subsequently as the hierarchical group-average agglomerative clustering, HGAA). The process of clustering may be represented on a dendrogram (for examples see Section 4.1).

The other clustering methods, used in much less extent in this Dissertation, are the non-hierarchical methods: k-means and Jarvis-Patrick, which are described in detail in Ref. 118.

Jarvis-Parick is an example of the nearest neighbor method [120]. Here, the compounds are clustered according to the number of neighbors they have in common (they have to be neighbors of each other, as well). The k-means method is an example of the relocation clustering method [121]. Here, the first step is to choose a set of "seed" compounds. The remaining compounds are assigned to the nearest seed to give an initial set of clusters. The centroids of these clusters are then calculated and the compounds are reassigned to the nearest centroid. The procedure is repeated until convergence is reached.

2.3.4 Virtual-screening, Combinatorial Chemistry and Tautomer Generation Programs

Virtual-screening is a general term describing the process of searching a chemical library for a molecule with the desired structure or property. In the context of this Dissertation we will narrow this definition to screening the libraries of preselected (or generated) compounds. Virtual-screening methods have been primarily applied for drug design. Nowadays the spectrum of possible applications is becoming much broader. For example, they have been used in the design of molecular receptors with binding sites that complement metal ion guests [122]. Extended and diverse libraries of compounds are typically developed using combinatorial chemistry methods. These libraries are built by positioning and connecting molecular fragments or by growing substituents on a core [123].

In the context of this Dissertation, it is important to mention current applications of combinatorial approaches to identify important tautomers. Usually different tautomers are considered in automated docking studies as protein binding

affinity might be very different among tautomers of a molecule being docked. The selection of potentially important tautomers is typically done based on common organic chemistry knowledge. There are also software tools available for generation of tautomers [124-126]. Typically they first identify proton donor (NH or OH groups) and acceptor sites (N or O). Next, a library of compounds is generated with various tautomers resulting from proton transfers between electronegative atoms, such as N or O.

3. Methodology Developed by the Author

3.1 Identification of the Most Stable Tautomers by Screening of Combinatorially Generated Libraries of Tautomers

3.1.1 Introduction

In the Section 2.1.4.4.3 we mentioned the discovery of new stable anionic tautomers of uracil and in Section 2.1.4.5 we highlighted the need for systematic searches of the most stable anionic tautomers of purine bases. The identified obstacle preventing these searches was the human time required to "manually" perform the screening of hundreds of anionic tautomers. Clearly, this problem should be handled using a hybrid approach involving both combinatorial and reliable quantum chemical methods. Such an approach is presented in this and the following three subsections.

The key part of such automated approach is the generation of a diverse library of molecular tautomers. We expect that the most promising tautomers might result from some uncommon transformations of canonical tautomers, i.e., a proton is transferred between N and C atoms. These possibilities are not taken into account in the available software for generation of tautomers (see Section 2.3.4), as these tools have the embedded chemical knowledge based on studies of neutral systems for which such tautomers are highly unstable. To overcome these limitations, we have developed a new program for generation of tautomers, TauTGen. This program builds all possible tautomers from a molecular framework (the core) and a specified number of hydrogen atoms. The hydrogens are attached to the sites specified by a user and a library of tautomers is combinatorally generated within a user-defined list of constraints. The prescreening is performed based on the results of DFT geometry optimizations. We call this approach "energy-based virtual screening" because the most stable tautomers are the target of this screening. The geometries of the top hits identified in the B3LYP energy-based screening were further optimized at the MP2 level of theory and final energies were calculated at the CCSD(T) level.

The details of the developed approach for identification of the most stable anionic tautomers will be discussed in the following three sections.

3.1.2 Combinatorial Generation of Libraries of Tautomers

The TauTGen program was written in the C programming language with the purpose to generate a library of tautomers for a given molecule. TauTGen constructs tautomers from a molecular frame built of heavy atoms and a given number of hydrogens (F-3.1-1).

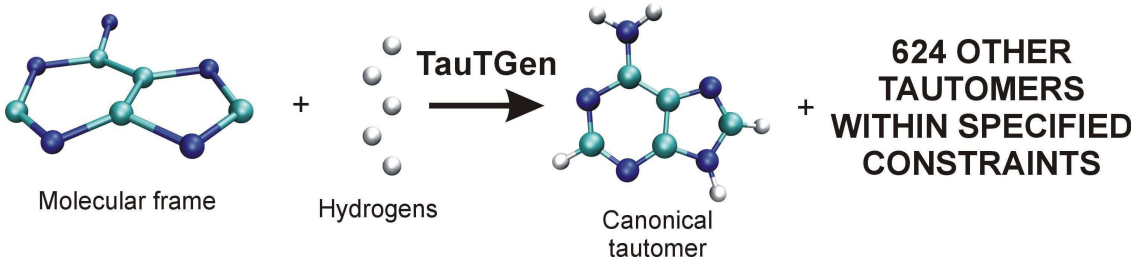


Figure F-3.1-1. TauTGen uses a fixed frame of heavy atoms and a given number of hydrogen atoms to create tautomers of the resulting molecular system (here adenine is used as example)

The user has to provide geometry of the molecular frame and to specify the minimum and maximum number of hydrogen atoms connected to each heavy atom. Sites for placement of hydrogen atoms are also defined by the user. To define a site, the user has to provide the following information:

- Name - a string of characters used to build up a filename for each tautomer
- A point where the hydrogen atom is to be placed. The point is defined relative to the fixed molecular frame
- Information which heavy atom is the holder of this site (connectivity information)
- The required total number of hydrogen atoms assigned to the heavy atom which would make the specific site available for occupation (a site constraint)
- Stereoconfiguration information, which tells the program if occupying a particular site will lead to the R or S configuration of the connected heavy atom.

Special care is taken to precisely name the sites. These names are used to create the names of tautomers that are later used as the filenames. For example, sites A and B (Figure F-3.1-2 a) are named "N4cis" and "N4trans" to distinguish possible rotamers resulting from rotation of the N4H imino group. The connectivity

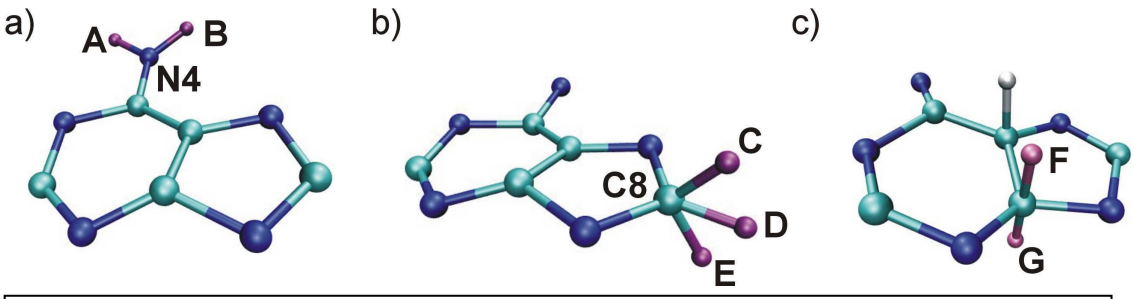


Figure F-3.1-2. Information needed to define sites for hydrogen attachment. The sites are marked with capital letters.

information is used to count the number of hydrogen atoms at each heavy atom, N_s. The number of available sites for hydrogen might be 2 even when N_s=1. Each site has a defined constraint, which tells for which values of N_s the site becomes available for occupation. This is what we mean by the site constraint. This option is used to build proper hybridizations of heavy atoms. For example, the C and E sites (Figure F-3.1-2 b) are occupied only when N_s=2 for C8. Then C8 attains the sp^3 hybridization. On the other hand, the D site is occupied only when N_s=1 for C8 - the sp^2 hybridization is then assigned to C8.

If a user wants to generate stereoisomers, then two sites have to be used for each asymmetric atom in order to describe the R and S configurations. In the case of planar or nearly planar NABs the sites F and G that are "below" and "above" the molecular plane might be distinct (Figure F-3.1-2 c). Each of these sites bears additional information describing the configuration, e.g. 1 or 2 for the "above" or "below" configuration, respectively.

As soon as the framework, available sites, the total number of hydrogen atoms $N_{hydrogens}$, and all constraints are defined, TauTGen generates all possible distributions of $N_{hydrogens}$ hydrogens among N_{sites} sites. For each distribution TauTGen checks whether all applied constraints are respected. The constraints are checked in the following order:

- constraints on the maximum and minimum number of hydrogens connected to each heavy atom
- site constraints; check if the sites are used consistently with the actual values of N_s

- stereoconfiguration; check whether other enantiomer has already been generated (this check is not done by default).

Each new distribution needs to pass all these checks to become an entry in the library of tautomers.

The stereoconfiguration check is done by a separate routine that detects enantiomers of a given distribution. If an enantiomer of the previously generated stereoisomer has been built, the distribution is rejected so the final set of stereoisomers consists of diasteroisomers only. The following steps are parts of the stereoconfiguration check:

- A stereoconfiguration fingerprint is assigned to each new distribution. The fingerprint contains information if hydrogens occupying stereosensitive sites are above or below the molecular plane. In other words, we keep track whether the involved heavy atoms are R or S.
- An inverse stereoconfiguration fingerprint is created for the distribution. It is then compared against the stereoconfiguration fingerprints of all previously generated stereoisomers of the same tautomer.

If there is no match between the fingerprints, the current distribution is a diastereoisomer of the previously generated stereoisomers and it is accepted to the library. If there is a match then the current distribution is an enantiomer and hence it is rejected.

Finally, TauTGen generates filenames and saves atomic coordinates of each member of the library to a separate file. The filename is a string of these site names that were used to build up the molecule. If proper site names are defined, the filename can uniquely name the molecular structure and discriminate various rotamers of the same tautomer. To facilitate files management, we sometimes divide structures among groups and subgroups based on the values of N_s for preselected heavy atoms. If stereoisomers were generated, the tautomer name is supplemented with a stereoconfiguration label: eg. ”Z_nml” where n,m and l are names of these sites that are on the same side of the molecular plane.

The source code of TauTGen is available free of charge and can be downloaded from the Sourceforge Internet archive [127]. The manual of TauTGen is available online and includes examples of input files.

3.1.3 Screening of the combinatorially generated library of tautomers

We used simple UNIX shell scripts to automate the screening procedure. The initial geometries of molecular structures are expressed in Cartesian coordinates and stored in typical .xyz files. They are used to build input files to the Gaussian03 program using a csh shell script. Initial screening is performed at the DFT level of theory with a B3LYP exchange-correlation functional [98] and 6-31++G** basis set for guanine. A tendency of B3LYP to overestimate the excess electron binding energy helps to avoid false negatives when screening for adiabatically bound anions. The 6-31++G** basis set has an advantage that the time required to perform geometry optimization for a NAB is acceptable. This choice of the method and the basis sets was also supported by our earlier experience with calculation of adiabatic electron affinities (AEAs) for some pyrimidine NABs (see Section 2.2.2).

It is known that "buckling" of the ring of a NAB might increase the electronic stability of the anion, because the excess electron occupies a π^* orbital. For this reason, all initial structures of anions were built from buckled molecular frames. In the case of about 15% of generated structures, the initial try of the self-consistent field procedure (SCF) failed to converge. In these cases we applied one, or a combination of up to four approaches: a) start the calculation from orbitals generated with a smaller basis set (3-21G or 6-31G*), b) start the calculation from orbitals generated in water solution simulated with the IEF-PCM method and the cavity built up using the United Atom (UA0) model [128] c) try to converge the SCF procedure using a quadratically converging algorithm, d) start the calculation from a slightly distorted geometry (the distortion was introduced by performing 2 optimization steps, but for the neutral molecule). In consequence, we recorded only a few cases when the SCF procedure failed to converge for the initial structure of the anion. The screening calculations of some guanine tautomers were performed using Gaussian03 on dual Intel Itanium2 nodes. The remaining tautomers of guanine were calculated using NWChem on an SGI Altix computer. For anionic systems, the B3LYP geometry

optimizations were followed by single point calculations for neutral systems at the optimal anionic geometries to determine tautomers' VDEs.

We developed Gaussian Output Tools (GOT) scripts [129] to analyze output files from Gaussian03. The GOT scripts are written in the Practical Extraction and Report (Perl) language and can extract final energies, geometries and forces from the Gaussian03 output files. Analogous scripts were developed for NWChem output files. Other shell scripts were used to identify and restart the calculations of tautomers for which the SCF or geometry optimization failed to converge. The final B3LYP energies for the neutral and anionic species were copied to a Microsoft Office Excel spreadsheet, which was used to calculate the relative energies as well as AEAs and electron vertical detachment energies (VDEs) (anions) and adiabatic and vertical ionization potentials (cations). The spreadsheet was also used to sort the molecular structures according to their relative energies. In some cases we found that two, or more initial structures converged to the same energy. We have analyzed these cases, in addition to the most stable tautomers, using the Molden software package [130]. In all these cases, the same energy resulted from the same converged structure.

3.1.4 Refinement of the energies of the selected tautomers

The anionic tautomers with positive values of AEA determined at the B3LYP level were characterized at our accurate level (see Section 2.2.2). The B3LYP geometries were further optimized at the MP2/AVDZ level of theory [102]. The final single-point calculations were performed at the coupled cluster level of theory with single, double, and non-iterative triple excitations (CCSD(T)/AVDZ) [102] at the optimal MP2 geometries. The relative energies of the anion with respect to the most stable tautomer of the neutral were corrected for the energies of zero-point vibrations to derive the values of AEA. The MP2 geometry optimizations and frequency calculations were performed with Gaussian03 and the CCSD(T) calculations with the MOLPRO package. The codes were run on clusters of dual Intel Itanium2 nodes with and without Quadrics interconnect.

3.2 Manipulation and Visualization of Molecular Orbitals and the Related Electron Densities

3.2.1 Introduction

In the previous sections we demonstrated an approach to identify the most stable anionic tautomers and obtain accurate energies. In this section we discuss approaches and software we developed to improve handling and visualization of molecular orbitals and the related electron densities.

The values of electron density in a molecular fragment and the bonding/antibonding character of the orbital contribute to chemical properties of this fragment. Therefore practically all electronic structure codes give their users an option to access molecular orbital data, either in a form of the coefficients associated with basis functions or as volumetric data with values of the orbital or the related electron density at each point of a predefined grid. Many programs have been developed to visualize orbitals and/or electron density and they are in common use by the community of computational chemists [130-132]. Orbitals and electron densities are typically visualized as finite volumes limited by a boundary defined by a preselected contour value (CV). On occasion 2D maps, which are cross-sections of the finite volumes, are prepared with marked isovalues of the presented quantity.

Interestingly, plotting an orbital or electron density with a pre-defined contour value seems to be the only option implemented in the major visualization software packages. Similarly, when comparing molecular orbitals or electron densities of different systems one usually prepares plots using consistent CVs. This approach works fine when the charge distributions do not differ much in their spatial extension. We found, however, the same approach misleading when the studied charge distributions span a broad range of extension. The problem becomes particularly relevant when dealing with orbitals, which are characterized by very different orbital energies, and therefore different electron binding energies. This results from the long-range asymptotic behavior of bound-state wave functions and orbitals [133, 134] e.g. the occupied Hartree-Fock orbitals decay as $\exp[-(-2\varepsilon_{HOMO})^{1/2} r]$,[133] where ε_{HOMO} is the orbital energy of the highest occupied orbital.

The significant differences in electron binding energies of anions should be reflected in the diffuseness of the singly occupied molecular orbital (SOMO). An opposite relation is often encountered when plotting SOMOs according to the common practice – a consistent CV (in $bohr^{-3/2}$). That is, the SOMO of a strongly bound anion seems to be more diffuse than that of a loosely bound anion. This problem has been demonstrated in detail using an example of the $(ClH...NH_3)^-$, $(Cl^-...NH_4^+)^-$ species, where the former is characterized by an electron binding energy one order of magnitude smaller than the latter (see Appendix II for graphical examples and full discussion). In this study, we concluded that the presented problem is actually an illusion having its origin in the fact that the fractions of electrons (F_e) contained in the volumes determined by the same CV value might be very different.

Our study suggests that an unbiased way to visualize orbitals or electron densities that differ much in the extension of charge distributions would be to assure that a consistent and preselected fraction of the total charge is reproduced in each plot. The same conclusion was reached by Rauk and Armstrong in their studies of dipole-bound and valence anions in clusters involving various hydrogen halides [135-137]. The approach, i.e., plotting different orbitals in such a way that the same fraction of electron charge is reproduced, leads to another question: what are the CVs that lead to the same and preselected F_es? Clearly, these CVs might be different for different orbitals. In the following section we will present an efficient algorithm to determine the desirable CVs. The same algorithm can be used to calculate a fraction of the total charge corresponding to a particular CV. We will also present an algorithm to preselect particular parts of the orbital which can be used for further visualization and analysis. The proposed algorithms are made available to the scientific community by providing appropriate software.

3.2.2 Details of Algorithms and Implementation

The software presented here works with volumetric data containing orbitals or orbital densities. The latter are often referred to as "cube files" [106]. They typically contain the Cartesian coordinates of atoms and a definition of the grid. The

grid is defined by a starting point, three non-parallel vectors and a size of the grid (the numbers of points in each direction defined by the grid vectors). Our software provides the following functionality: (i) identification of a CV that corresponds to a preselected value of F_e, (ii) determination of F_e associated with a given CV, (iii) selection of a particular part of the grid limited by a pre-defined plane. This selection is made by zeroing the to-be-discarded part of the grid. The last functionality can be applied many times, i.e., a few planes can be defined and the grid can be trimmed to the desired slice of the orbital or the related electron density. It is up to the user to define desirable F_es and limiting planes, if any. We believe that instructive plots of orbitals and orbital densities can be generated using the OpenCubMan software [138] in combination with molecular visualization packages, and using "cube files" produced by common quantum chemistry packages. Such examples will be provided in Section 4.1.5.

A CV corresponding to a preselected F_e is determined using an algorithm summarized in Figure F-3.2-1. In this algorithm the charge density is integrated starting from the densest region to the least dense region. The process of density integration is stopped when a preselected fraction of the charge has been recovered. The searched CV is equal to the value of orbital density at the last integrated point (if plotting electron densities) or to the properly signed square root of it (if plotting orbitals).

1. Generate or read-in grid points and the corresponding volumetric data containing orbital or orbital density values
2. If the orbital values are provided in point 1, calculate the corresponding orbital density values
3. Sort grid points according to the orbital density values
4. Loop over sorted grid points and perform numerical integration of the orbital density starting from the point of the highest density value
5. Stop integration when the integrated value exceeds the preselected fraction
6. The searched CV is equal to the value of orbital density at the last integrated point (if plotting electron densities) or to the properly signed square root of it (if plotting orbitals)

Figure F-3.2-1. Algorithm for determination of a contour value corresponding to a preselected fraction of the total orbital charge

Creating a cross-section of an orbital represented on a grid can be achieved by zeroing a part of the volumetric data above or below a predefined plane. A given plane is described with:

$$ax + by + cz + d = 0 \qquad \text{(E-3.2-1)}$$

where a, b and c are components of a vector **v** normal to the plane and d is a parameter which can be calculated by solving Eq. E-3.2-1 for a given point on the plane. A distance D of any point $p_0=(x_0,y_0,z_0)$ from the plane can be calculated using the following equation [139]:

$$D = \frac{ax_0 + by_o + cz_0 + d}{\sqrt{a^2 + b^2 + c^2}} \qquad \text{(E-3.2-2)}$$

Such a definition allows D to have a positive or negative sign. D is positive if p_0 is on the same side of the plane as the vector **v** and negative if it is on the opposite side. When zeroing a part of the grid by using a plane, each point of the grid is tested against Eq. (F-3.2-2), and the value of this point is set to zero or remains unchanged, if appropriate.

All the functions presented above have been implemented in the Open-Source Cubefile Manipulator (OpenCubMan) program, which is provided free of charge under the GNU license [138], and can be downloaded from the SourceForge Internet Archive. OpenCubMan was written in the object oriented C++ programming language and is provided as a C++ object definition. OpenCubMan uses standard C/C++ libraries for all input/output operations, math and sorting (qsort function). This form facilitates incorporating the code into other packages, libraries or scripting languages.

3.3 Analysis of results of multiple quantum mechanical calculations

3.3.1 Introduction

The development of the hybrid quantum mechanical-computational approach presented in Section 3.1 provided an automated way to characterize hundreds of tautomers in the process of identification of the most stable species. At the same time it brought new challenges for us, computational chemists. For example, how to analyze tens of structures characterized at the high level of theory in the last step of the hybrid approach? The natural path forward is to use chemoinformatics techniques, which have been developed to deal with large amount of chemical data. This, however, brings another challenge: how to process quantum chemical (QC) data using existing chemoinformatics tools?

In the followings sections we will present the steps we have taken to meet these challenges. The approaches we have developed combine data from QM calculations (e.g., orbitals, electron density and geometries) with chemoinformatics analysis methods (e.g., similarity calculations and clustering). In this approaches we code QC data into vectors (called *holograms*) and then perform chemoinformatics analysis. For example, for the excess electron distribution represented by holograms derived using Bader's electron density analysis, the similarity can be defined using Euclidean distance and the HGAA clustering can be performed.

3.3.2 Analysis of geometrical parameters

As already mentioned in Section 3.1.3, the important structural feature of anionic tautomers of NABs is the buckling of the molecule. The geometrical parameters related to buckling are the dihedral angles defined among the atoms of molecular frame of non-hydrogen atoms. One can compute the dissimilarity between the buckling modes (D_{BM}^{AB}) of different tautomers using the Euclidean distance (Section 2.3.2):

$$D_{BM}^{AB} = \left[\sum_{i=1}^{N} (\gamma_{iA} - \gamma_{iB})^2 \right]^{\frac{1}{2}} \qquad \text{(E-3.3-1)}$$

where γ_{iA} and γ_{iB} are the i-th dihedral angle related to the buckling of tautomer A and B, respectively. For a given optical isomer of A this optical isomer of B is selected, which provides a smaller value of D_{BM}^{AB} .

Having defined a similarity measure between buckling modes, all pairwise similarities can be calculated, and clustering then performed to group the most similarly buckled tautomers. The clustering method applied to this problem is a HGAA method described in Section 2.3.3.

3.3.3 Analysis of Charge Distributions

The analysis of the excess charge distribution presented in this and the following sections is based on the Hartree-Fock singly occupied molecular orbitals obtained at the optimal MP2/APVDZ geometries in the final step of the hybrid quantum mechanical-computational approach. The major differences in the distribution of the excess electron can be identified by comparing the SOMO plots. However, to get the quantitative information we developed a novel approach, in which the electron density contribution coming from the singly occupied molecular orbital (SOMO) is assigned to heavy atoms using Bader's analysis [140,141]. Bader's analysis defines a unique way of dividing molecules into atoms. The definition of an atom is based purely on the electronic charge density. The atoms are divided by so-called zero flux surfaces, which are 2-D surfaces on which the charge density is a minimum perpendicular to the surface. Having defined the atom limiting surfaces, the charge density is integrated over the volumes occupied by particular atoms.

We propose to define an *orbital density hologram* (in short *an orbital hologram*) as a vector the components of which hold information about population of the excess electron on each heavy atom. We calculate dissimilarity between two orbitals by calculating the Euclidian distance D_{orb}^{AB} between orbital holograms:

$$D_{orb}^{AB} = \left[\sum_{i=1}^{N} \left(x_{iA} - x_{iB} \right)^2 \right]^{\frac{1}{2}} \qquad \text{(E-3.3-2)}$$

where x_{iA} and x_{iB} are the i-th components of the orbital holograms for the tautomers A and B.

Having defined a dissimilarity measure between orbitals, all pairwise dissimilarities can be calculated and then HGAA clustering (as in the previous section) is performed to group the most similar orbitals. The progress of clustering

the SOMO orbitals of the important anionic tautomers can be presented as a dendrogram. The information that is available from orbital clustering contributes to our understanding of the binding modes of the excess electron. The similar shape of the electron density distribution corresponding to the SOMO orbital suggests a similar nature of the corresponding electronic state of the molecule.

During the development of the method for comparing molecular orbitals, we tested several alternative approaches. At first, we employed grid-based similarity calculations. The similarity between two orbitals was defined by the Euclidian distance of two grids, i.e., the sum of the absolute values of the differences in values of wavefunctions at corresponding points on both grids. The obvious disadvantage in this approach is the need to find the maximum overlap between the two wavefunctions before the similarity calculation can be performed. We have also tested the orbital hologram approach with a different similarity measure – the Manhattan distance (as defined in T-2.3-1 in Section 2.3.2). All the orbital comparison methods discussed in this section gave qualitatively the same results.

In addition to the analysis of the excess electron distribution using clustering of orbital holograms we can check how the excess electron is distributed among the main regions of the considered molecule. In the case of guanine, we divided the molecule into three regions corresponding to 6-member, 5-member ring and the part which is common for both rings. More details on this will be provided in the Results, Section 4.1.6.2.

3.3.4 Analysis of Bonding/Antibonding Effects of Singly Occupied Molecular Orbitals

The π^* orbitals occupied by the excess electron in the anionic NABs tautomers have partly a bonding and partly an antibonding character (We will sometimes refer to this orbital as π to stress significant bonding character.). The major differences in the distribution of bonding and antibonding areas of SOMO orbital can be identified by visual inspection. To verify if the bonding and antibonding character of the π orbital correlates with a particular tautomer's stability, we developed an approach that can quantitatively measure the bonding or antibonding character. This

is done by summing the contributions over the chemical bonds present in the molecular framework built from the heavy atoms.

In our approach, the determination of bonding/antibonding character has been designed in the spirit of the Hückel model of π-electron systems [142]. In this model, π orbitals are expressed as linear combination of p_z atomic orbitals (AO) of atoms forming the π-system. It is a minimal basis set for π electrons. Moreover, it is assumed that the AO's are orthonormal and only first immediate neighbors couple through the Hamiltonian. The way to estimate the bonding/antibonding character between two atoms is to look at bond orders resulting from a given orbital. For a given orbital, a contribution to the bond order between atoms X and Y is given by $c_X * c_Y$, where the c's are LCAO coefficients of the contributing p_z functions of atoms X and Y, respectively. Furthermore, the contribution from a given orbital to electronic charge localized on atoms X and Y are c_X^2 and c_Y^2, respectively.

In the spirit of the Hückel method, we introduce a minimal basis set for π electrons. This hypothetical basis does not contain conventional p_z atomic orbitals but rather effective atom-centered basis functions that reproduce an accurate occupied molecular orbital that we want to analyze. This molecular orbital has been obtained with a conventional extended basis set, e.g., AVDZ. We assume that all Hückel model assumptions apply to the new, hypothetical, minimal basis set. Moreover, we assume that bond orders and charges on atoms are calculated in the analogous way. The question remains how to find the LCAO coefficients c_x and c_Y that accompany the hypothetical basis functions centered on the X and Y atom, respectively. For the molecular orbital of interest we determine Bader's charges and we monitor the sign of the orbital in the neighborhood of each heavy atom X. This information is sufficient to determine the c_x coefficients. The details of this procedure will be described below. We will demonstrate in the following section that for the benzene molecule this approach gives practically the same results as the Hückel model.

The detailed procedure to calculate a contribution from a π orbital to the bond order between neighboring atoms X and Y is as follows. The valence anions of NABs typically do not have Cs symmetry and one needs to define an approximate

molecular plane. This plane is selected in a way to minimize the distance of heavy atoms to the plane and it is defined by eigenvectors of inertia tensor. The molecular plane is consistent for all tautomers as they were superimposed before calculating of inertia tensor. This plane can be used to select electron density on either side of the plane by the algorithms implemented in OpenCubMan program. Next, we integrate the electron density associated with the π orbital over the spaces associated with atoms X and Y (where the atomic spaces result from Bader's analysis discussed in the previous section), and the resulting atomic charges are denoted δ^X and δ^Y, respectively. In addition, we focus attention on one side of the approximate molecular plane and we monitor which sign, plus or minus, dominates in the space associated with X and Y. These signs are labeled sign(X) and sign(Y), respectively (Figure F-3.3-1). In the case of tautomers of NABs there was no ambiguity in determining the signs. Finally, the c_X and c_Y coefficients are determined as:

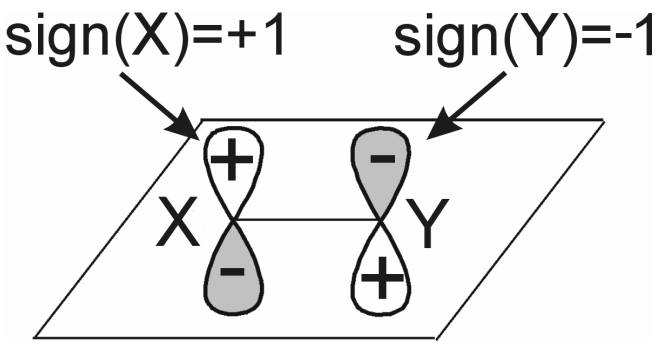


Figure F-3.3-1. Determining the sign of SOMO orbital for the purpose of calculating bonding and antibonding effect on a chemical bond between X and Y.

$$c_Z = sign(Z)\sqrt{\delta^Z}\ , \text{Z=X, Y} \qquad \text{(E-3.3-3)}$$

and a contribution to the bond order between X and Y is given by $c_X c_Y$. The positive and negative sign of $c_X c_Y$ determines whether the interaction is of bonding or antibonding character, respectively. The result does not depend which side of the molecular plane is used to determine *sign(X)* and *sign(Y)*.

Having defined a method to measure the bonding/antibonding character of the SOMO orbital for each bond, we can define a vector, the components of which hold this information for all bonds present in the molecule. We will refer to these vectors as *bonding character holograms*. Similarly to the orbital holograms defined in the previous section, the similarity between bonding character holograms is calculated using the Euclidean or Manhattan distance. Both similarity measures give

qualitatively the same results in this case. The bonding character holograms are clustered using the HGAA method described previously and the corresponding dendrogram can be generated.

The total bonding and antibonding character of the SOMO orbital can be calculated as a sum of, respectively, bonding and antibonding contributions over all the components of a bonding character hologram. These summed values indicate to which extent the SOMO is dominated by bonding and antibonding interactions.

3.3.5 Validation of the approach

To build Readers' confidence in our orbital holograms we discuss a simple case of benzene. The geometry of benzene molecule was obtained at the MP2/APVDZ level. Benzene has 3 doubly-occupied π orbitals presented in Figure F-3.3-2 and the corresponding orbital holograms are presented in the left part of Table T-3.3-1. These holograms seem to reflect what is known about benzene molecule. For the fully symmetric A orbital the electron density is uniformly distributed over six carbon atoms. The orbital holograms for the E_1a and E_1b orbitals properly reflect electron distribution. For the former the largest electron density is on atoms 1 and 4, whereas the latter has no significant electron density on the same atoms (Figure F-3.3-2). Moreover, the partial electron charges associated with heavy atoms (the components of orbital holograms) are very similar for our HF/APVDZ and Hückel orbitals.

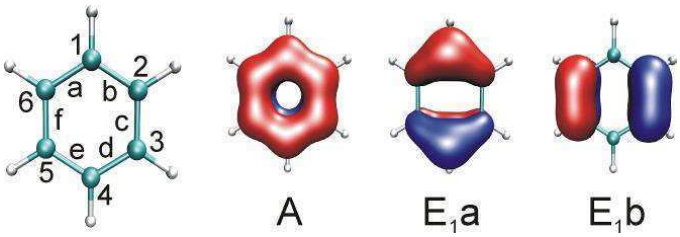


Figure F-3.3-2. Benzene molecule with notation used to discriminate atoms and bonds and three occupied π orbitals.

Can the bonding character holograms provide simple but informative representations of molecular orbitals obtained in extended basis sets? Here we test this concept on occupied π orbitals of benzene (Figure F-3.3-2) and we determine the corresponding bonding character holograms, consisting of contributions from all six

CC bonds (a-f in F-3.3-2). These holograms are presented in the central part of T-3.3-1.

For the full symmetric A orbital, all six bonds have identical bonding character, with the total bonding effect, TOT, equal to 1 and the total nonbonding effect, TOT*, equal to 0. For the E_1a orbital, the bonding interaction is reported between atoms 6-1, 1-2, 3-4 and 4-5 while the antibonding interaction for 2-3 and 5-6. The components to the bonding character hologram related to the a, b, d, and e bonds are positive (bonding character) and equal to 0.17 while the components related to the c and f bonds are negative (antibonding) and equal to -0.09. For this orbital there is a partial cancellation between total bonding (0.67) and anibonding (-0.18) effects. For the E_1b orbital the bonding character is reported for the c and f bonds (bond order 0.25 each). The total bonding and antibonding character (TOT+TOT* in Table T-3.3-1) is the largest (1.0) for the fully symmetric A orbital, and smaller but still positive (0.5) for the degenerated E_1 orbitals. A slight difference of 0.01 between E_1a and E_1b is a manifestation of numerical noise that originates form integration of electron density over cubic grids. Finally we notice that the orbital bond orders derived by us from the HF/APVDZ Bader's electron partial charges are very similar to those obtained from Hückel orbitals (Table T-3.3-1). The above example of occupied orbitals of benzene shows that the bonding character holograms are convenient quantitative representations of molecular orbitals, which are routinely illustrated using contour plots.

Table T-3.3-1. Orbital holograms and bonding character holograms for an electron occupying A, E_1a and E_1b orbitals of benzene obtained using both model presented in this study and the Hückel model. Contributions from atoms 1-6 are presented for orbital hologram. Contributions from bonds a-f, as well as total bonding, TOT, and total antibonding, TOT*, characters are presented for bonding character holograms.

		Orbital hologram						Bonding character hologram						Bonding effect		
Model	Orbital	1	2	3	4	5	6	a	b	c	d	e	f	TOT	TOT*	TOT+TOT*
This study	A	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	1.00	0.00	1.00
	E_1a	0.32	0.09	0.09	0.32	0.09	0.09	0.17	0.17	-0.09	0.17	0.17	-0.09	0.67	-0.18	0.49
	E_1b	0.00	0.25	0.25	0.00	0.25	0.25	0.00	0.00	0.25	0.00	0.00	0.25	0.50	0.00	0.50
Huckel	A	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	1.00	0.00	1.00
	E_1a	0.33	0.08	0.08	0.33	0.08	0.08	0.17	0.17	-0.08	0.17	0.17	-0.08	0.67	-0.17	0.50
	E_1b	0.00	0.25	0.25	0.00	0.25	0.25	0.00	0.00	0.25	0.00	0.00	0.25	0.50	0.00	0.50

3.4 Accurate Free Energies of Solvation

3.4.1 Introduction to QM/MM and the Complex Environments

So far we considered approaches that involve performing only quantum mechanical calculations. Although such calculations may provide accurate results, their real disadvantage is that they cannot be applied to large systems due to the computational cost to characterize the whole system. Moreover, in most of the applications only a subsystem (*e.g.*, where the chemical reaction takes place) requires accurate (=QM) treatment. Because of this, the idea of combining quantum mechanics and molecular mechanics to treat large molecular systems emerged. Applications of such idea are usually referred to as QM/MM. In the QM/MM approach the subsystem of interest is studied at the quantum mechanics level, whereas the surroundings, the environment is modeled using the empirical potentials. There are several implementations of QM/MM differing mainly in the way how QM region interacts with the MM region, to mention IMOMM of Morokuma and Masears [143] or QM-Pot of Sauer and Sierka [144] as examples.

Moreover, when moving our attention to larger systems, especially the ones in the condensed phase (like solvated molecules or proteins) new challenges are encountered. An example of such challenge is the calculation of the thermodynamic properties (like free energies) of the complex system. The mathematical means to calculate the thermodynamic properties of such systems are provided by statistical mechanics theory. The statistical data to calculate condensed-phase properties are obtained using molecular dynamics (MD) or Monte Carlo simulations.

In the following sections we will present an approach to calculate the solvation free energies of molecules. The statistical data required for such calculation are obtained in the MD simulation run on the QM/MM potential energy surface defining a solvated molecule (in water solution or inside a protein). For now on, the QM/MM term will be used to referrer to, generally, all calculations involving QM/MM potential (so it will also include MD and thermodynamic properties obtained from simulations run on the QM/MM potential).

3.4.2 Accelerating QM/MM Free Energy Calculations

QM/MM approaches have provided a general scheme for studies of chemical processes in the condensed phase, like in solutions or in proteins [145-156]. Significant progress has been made with calibrated semiempirical QM/MM approaches [146,151,154,155] that include careful evaluations of the relevant activation free energies by free energy perturbation approaches that date back to the 80's [157]. These studies exploit the rapid evaluation of the semiempirical energies and sample the phase space of the QM atoms and the surrounding MM atoms. However, the current challenge is to move to an *ab initio* representation with a QM/MM treatment, since such representations have been shown to provide "chemical accuracy" in studies of gas phase reactions of small molecules. Here, we used the term *ab initio* mainly to differentiate form semiempirical methods (in the QM part of QM/MM). Therefore, we may also refer to density functional theory, which uses empirical parameters in the common implementations. For simplicity, from now on we limit the meaning of "QM/MM" to *ab initio* QM/MM (unless specified explicitly).

Unfortunately, at present it is extremely challenging to evaluate the free energies of chemical systems using the QM/MM approaches due to the requirement of a very extensive sampling, which results in the extremely computationally expensive repeated evaluation of the QM energies.

The recent realization of the importance of the proper sampling of QM/MM surfaces led to several advances [158-167]. A major direction of these advances have been based on different adaptations [162-167] of the Warshel group idea [158,159] of using a classical potential as a reference for the QM/MM calculations. Other strategies have also been quite promising [161,168]. Nevertheless there is clearly a need for more "mainstream" approaches that can be used in standard implementations and aid in obtaining converging QM/MM free energies.

Here, we consider a simple and powerful treatment that can be viewed as a variable time step approach (from the QM/MM perspective). Within the QM subsystem we represent the average effect of the fluctuating solvent charges by using equivalent charge distributions, which are updated every m steps of MD simulation

of MM subsystem. Since number of required QM evaluations is reduced by the factor of m, we refer to this approach as *accelerated QM/MM*. This approach is formally equivalent to approaches that add the average potential to the solute Hamiltonian and is thus a mean field approximation. Obviously, adding the average potential is an old idea that was implicitly implemented in the QM/Langevin Dipole (QM/LD) model [169,170]. It is also implemented implicitly in continuum models [171-174]. Furthermore, an averaging approach was implemented recently in an instructive work of Yang and co-workers [168]. However, while the addition of the average potential to the semiempirical Hamiltonian is very simple [169], it requires specialized implementation in standard commercial QM/MM codes (that are designed to handle external point charges). Furthermore, the use of the average potential may not reproduce the average energy obtained by using the instantaneous potential in each time step, as in the common case when the solvent fluctuations are significant. In some respects our approach is close to the work of Aguilar and co-workers [175 - 178] who studied different averaging strategies in QM/MM approaches but have not focused the evaluation of free energies.

3.4.3 Accelerated QM/MM to predict solvation free energies of anionic uracil

The developed QM/MM approach has been derived in few variations. Presenting all of them would go beyond the scope of this Dissertation. We will direct the interested Reader to the corresponding article [179], which is also included in the Appendix II. Here we will present only this specific variation that will be used to calculate accurate solvation free energies of anionic tautomers of uracil (presented n the Results, Section 4.2.2). In this model uracil anion is placed in a solvating sphere of explicit water molecules. In our approach, the accurate free energies of solvation are calculated in two steps. The major contribution is calculated using classical MD simulations and the free energy perturbation (FEP) adiabatic charging (AC) approach, where it is assumed that the solvated tautomer has the charge distribution given by the polarizable continuum model (PCM). In the second step the classical free energy of solvation is refined to take into account the real, average charge distribution. This is done using our accelerated QM/MM simulations, where the QM

energy of the solute is calculated in the mean solvent potential averaged over a number of MD steps.

The calculated free energies of solvation can be used to determine the relative free energies of two tautomers, Ux and Uy, in water solution, $(\Delta G_{Ux\text{-}Uy})_s$ by using a simple energy cycle:

$$(\Delta G_{Ux\text{-}Uy})_s = (\Delta G_{Ux\text{-}Uy})_g - \Delta G_{Sol}Ux + \Delta G_{Sol}Uy \qquad \text{(E-3.4-1)}$$

where $(\Delta G_{Ux\text{-}Uy})_g$ is the free energy difference between Ux and Uy in the gas phase and $\Delta G_{Sol}Ux$ and $\Delta G_{Sol}Uy$ are the free energies of solvation of Ux and Uy, respectively.

3.4.4 Detailed description of the approach

When considering the energies of tautomers in solution we will use a QM/MM approach where solute is treated at the quantum mechanical level and solvent is treated classically. The energy of such system is expressed as:

$$\begin{aligned} E_{tot} &= \left\langle \phi_s \psi_S^{pol} \mid H_S^g + V_{Ss}^{el} + V_{Ss}^{vdW} + H_{ss} \mid \phi_s \psi_S^{pol} \right\rangle \cong \\ &\cong E_{QM}^{pol}(\mathbf{R}, \mathbf{Q}(U_s)) + E_{QM/MM}^{el}(\mathbf{R}, \mathbf{r}, \mathbf{Q}(U_s)) + E_{vdW}(\mathbf{R}, \mathbf{r}) + E_{MM}(\mathbf{r}) \end{aligned} \qquad \text{(E-3.4-2)}$$

where Ψ_S and ϕ_s are the wavefunctions of the solute (S) and solvent (s), the H_S^g and H_{ss} are the Hamiltonian operators describing the energy of the solute and solvent, respectively. The V_{Ss}^{el} and V_{Ss}^{vdW} represent, respectively, the electrostatic and van der Waals solute-solvent potentials. The **R** and **r** are the solute and solvent coordinates, respectively, and **Q** is a vector of solute residual atomic charges obtained from Ψ_S^{pol} that depend on the potential U_s of the solvent. Here the first term, $E_{QM}^{pol}(\mathbf{R}, \mathbf{Q}(U_s)) = \left\langle \Psi_S^{pol} \middle| H_S^g \middle| \Psi_S^{pol} \right\rangle$, is the energy of the gas phase Hamiltonian with a solute wavefunction polarized by the solvent. The second term,

($E^{el}_{QM/MM}(\mathbf{R},\mathbf{r},\mathbf{Q}(U_s))$), is the solute-solvent electrostatic interaction, which is approximated with a classical expression:

$$E^{el}_{QM/MM} \equiv 332\sum_{i(S)}\sum_{j(s)}\frac{Q^i q^j}{r_{ij}} \qquad \text{(E-3.4-3)}$$

where q are the solvent residual charges, whereas i and j are indexes of the solute and solvent atoms, respectively. The third term in Eq (E-3.4-2) , $E_{vdW}(\mathbf{R},\mathbf{r})$, is the solute-solvent van der Waals interaction and the last term, $E_{MM}(\mathbf{r})$, is the solvent potential surface.

Such defined E_{tot} can be used to evaluate the solvation free energies using the following approach. The starting point is the free energy perturbation (FEP) adiabatic charging (AC) approach [169,180], where we can use a potential in the form:

$$E_k = (E_{tot} - E_{MM})(1-\lambda_k) + E'\lambda_k + E_{MM} \qquad \text{(E-3.4-4)}$$

where E′ denotes an energy of the system without electrostatic solute-solvent interaction ($E' = \left\langle \Psi^g_S \mid H^g_S \mid \Psi^g_S \right\rangle + E_{vdW}$) and where λ_k changes from zero to one in n+1 steps.

We can use the standard FEP equation [169]:

$$\Delta\Delta G_{sol}(\lambda_k \to \lambda_{k+1}) = \beta^{-1}\ln\left\langle \exp\left\{-(E_{k+1}-E_k)\beta\right\}\right\rangle_{E_k}$$
$$\Delta G_{sol} = \sum_{k=1}^{n+1}\Delta\Delta G_{sol}(\lambda_k \to \lambda_{k+1}) \qquad \text{(E-3.4-5)}$$

where $\beta=1/(k_B T)$; k_B is the Boltzmann constant and T is the absolute temperature. Eq. (E-3.4-5) can be effectively approximated by using the linear response approximation (LRA) treatment [181]:

$$\Delta G_{sol} \cong \left\langle E_{tot} - (E' + E_{MM})\right\rangle_{E_{tot}} + \left\langle E_{tot} - (E' + E_{MM})\right\rangle_{E'} + \Delta G_{cav} \qquad \text{(E-3.4-6)}$$

where ΔG_{cav} is the solvation free energy of the nonpolar neutral form of the solute (all solute residual atomic charges are zero). This term consists of two parts describing

the hydrophobic and van der Waals free energies of cavity, which are not included in the first two terms of Eq. (E-3.4-6). They were described in Ref. 182.

To calculate the solvation free energies of anionic tautomers of uracil we use a cycle illustrated in Figure F-3.4-1, which makes use of both the FEP/AC and LRA

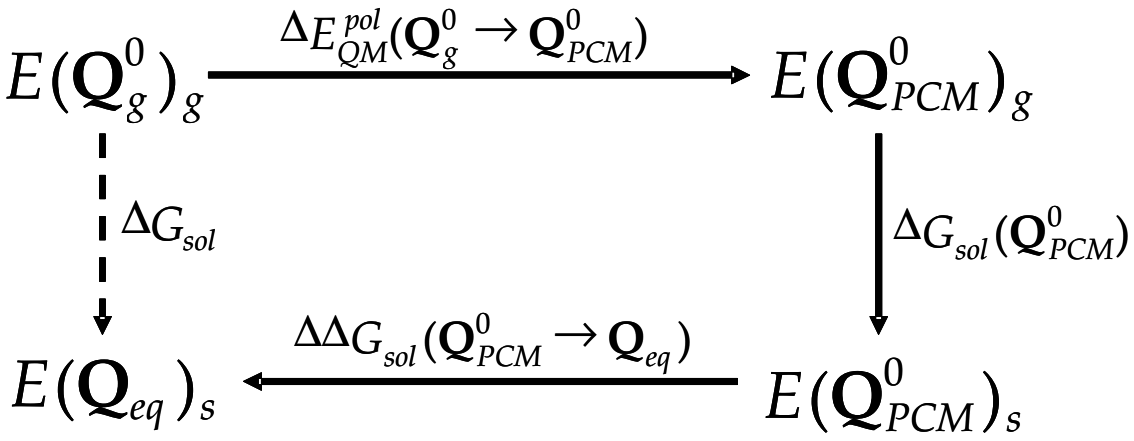


Figure F-3.4-1. An energy scheme used in this study. $\Delta G_{sol}(Q_{PCM}^0)$ is calculated by the classical adiabatic charging approach and $\Delta\Delta G_{sol}(\mathbf{Q}_{PCM}^0 \rightarrow \mathbf{Q}_{eq})$ is calculated using Eq. (E-3.4-9).

approaches presented in Eqs. E-3.4-5 and E-3.4-6, respectively. Here, we first polarize the solute molecule to a given charge distribution (e.g. partial charges obtained with the PCM solvation model ($\mathbf{Q} = \mathbf{Q}_{PCM}^0$, where "0" in a superscript designates a constant value), then we run a classical MM simulations and use the FEP/AC approach [169] to evaluate the free energy of solvation of such polarized solute. Then we evaluate a change in the free energy allowing the solute partial charges to "equilibrate" with the solvent potential. The vector of equilibrated QM/MM residual atomic charges is designed by $\mathbf{Q}_{eq}$. Thus the QM/MM solvation free energy can also be written as:

$$\Delta G_{sol} = \Delta E_{QM}^{pol}(\mathbf{Q}_g^0 \rightarrow \mathbf{Q}_{PCM}^0) + \Delta G_{sol}(\mathbf{Q}_{PCM}^0) + \Delta\Delta G_{sol}(\mathbf{Q}_{PCM}^0 \rightarrow \mathbf{Q}_{eq}) \quad \text{(E-3.4-7)}$$

where the ΔE_{QM}^{pol} term in Eq. (E-3.4-7) is the polarization energy, which is given by:

$$\Delta E_{QM}^{pol} = \left\langle \Psi_S^s \mid H_S^g \mid \Psi_S^s \right\rangle - \left\langle \Psi_S^g \mid H_S^g \mid \Psi_S^g \right\rangle \qquad \text{(E-3.4-8)}$$

where Ψ_S^s and Ψ_S^g are the solute wave function in solution and in the gas-phase, respectively, and H_S^g is the gas-phase Hamiltonian. The $\Delta E_{QM}^{pol}(\mathbf{Q}_g^0 \rightarrow \mathbf{Q}_{PCM}^0)$ represents the polarization energy of a molecule in the PCM solvation model. The $\Delta G_{sol}(\mathbf{Q}_{PCM}^0)$ term is the solvation free energy of the solute, the atomic charges of which have been obtained from the PCM model. The $\Delta G_{sol}(\mathbf{Q}_{PCM}^0)$ can be obtained using the classical adiabatic charging approach based on FEP and therefore can be replaced by $\Delta G_{sol}(\mathbf{Q} = 0 \rightarrow \mathbf{Q}_{PCM}^0) + \Delta G_{cav}$. The last term of Eq. (E-3.4-7), namely $\Delta\Delta G_{sol}(\mathbf{Q}_{PCM}^0 \rightarrow \mathbf{Q}_{eq})$, can be expressed using the LRA approach as:

$$\Delta\Delta G_{sol}(\mathbf{Q}_{PCM}^0 \rightarrow \mathbf{Q}_{eq}) \cong \frac{1}{2}\left[\left\langle E_{tot}(\mathbf{Q}) - E_{tot}(\mathbf{Q}_{PCM}^0)\right\rangle_{E(\mathbf{Q})} + \left\langle E_{tot}(\mathbf{Q}) - E_{tot}(\mathbf{Q}_{PCM}^0)\right\rangle_{E(\mathbf{Q}_{PCM}^0)}\right]$$

(E-3.4-9)

where $E_{tot}(\mathbf{Q})$ is the QM/MM surface with the fluctuating solute partial charges, which respond to changes in solvent configurations.

The main time consuming steps in the evaluation of Eq. (E-3.4-7) are evaluations of the LRA $\langle\ \rangle_{E(\mathbf{Q})}$ and $\langle\ \rangle_{E(\mathbf{Q}_{PCM}^0)}$ terms expressed in Eq. (E-3.4-9). The main problem is the need for a very long computer time to evaluate the QM energy, which in turn makes it extremely challenging to perform proper configurational sampling. To make these calculations possible we employ our accelerated QM/MM approach[179] that introduces the average effect of the fluctuating solvent charges on the QM system by using equivalent charge distributions, which are updated every m steps of a MD simulation.

Our strategy for evaluating the average solvent potential, which is technically similar to the approach of Aguilar and coworkers [175], is demonstrated schematically in Figure F-3.4-2. In this approach, we constrain the QM atoms (the solute atoms), evaluate the QM charges, $\mathbf{Q}^{(1)}$, where (1) designates the first step and run m MM/MD steps allowing the solvent molecules to move in the potential ($E_{QM/MM}^{el}(\mathbf{Q}^{(1)}) + E_{vdW}$). All m snapshots of solvent coordinates from m MD steps are

stored. Then the charge of each solvent atom is scaled by 1/m and all m×N solvent atoms with the scaled solvent charges are sent to the QM program to reproduce an average solvent potential on the solute. The latter is used to obtain the corresponding solute polarization and a new set of solute charges $\mathbf{Q}^{(2)}$. The procedure is repeated until all terms contributing to Eq. (E-3.4-9) converge.

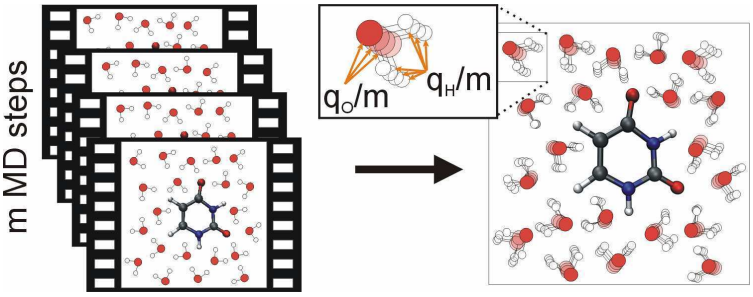


Figure F-3.4-2. A schematic representation of the averaging of the solvent potential over m steps of a MD simulation.

The approach of representing an average solvent potential (Figure F-3.4-2) is simple to implement but unfortunately it generates m×N external charges to be included into the Hamiltonian within the QM program. This can be too expensive and inconvenient as shown in the initial study. Thus, we introduce an approximation described in Figure F-3.4-3. In this treatment we divide the solvent into three regions. In the first region (region I), we convert the N_{regI} solvent atoms to $m \times N_{regI}$ external charges (scaled by 1/m). In the second region (region II) we represent each OH bonds of N_{regII} water molecules with two charges representing average dipole of a moving OH bond, while in region III, we represent the average solvent field coming from $N-N_{regI}-N_{regII}$ solvent molecules, by two point charges (q and -q) using:

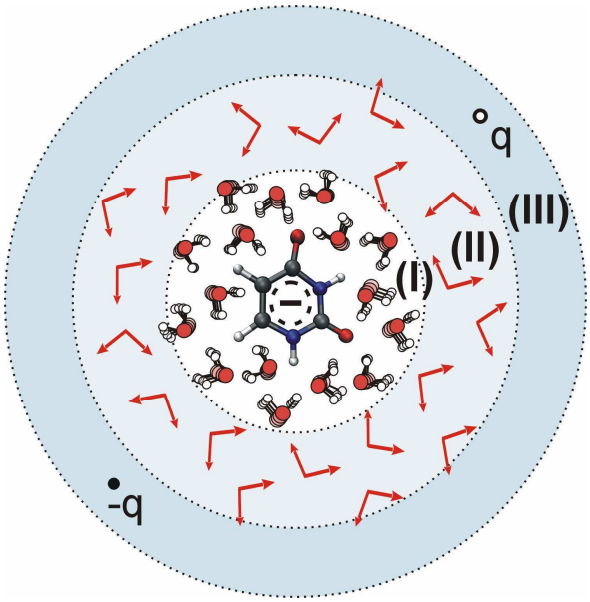


Figure F-3.4-3. Model for the evaluation of the average solvent charges: average the explicit molecules in region (I), while representing the average potential of the molecules in regions (II) and (III) by average dipoles and two charges, respectively.

$$\mathbf{E}_O = \frac{2q}{\left|\mathbf{r}_{OR}\right|^3}\mathbf{r}_{OR} \qquad \text{(E-3.4-10)}$$

where $\mathbf{E}_O$ is an electric field at point O (the geometrical center of the QM system) and $_{OR}$ is pointing along $\mathbf{E}_O$ to the charge q. For the validation of the approach we test a simpler solvent representation, which consisted of only two regions (of I and III type). It will be briefly discussed in the following section.

To perform the required calculation of Eq. (E-3.4-7), every considered anionic tautomer of uracil was hydrated in a sphere with a radius of 16 Å, which contained 558-561 explicit water molecules depending on the tautomer. All solvent molecules are represented by the ENZYMIX force filed [183]. In the simulation model the sphere of explicit water molecules is surrounded by a surface region whose average polarization and radial distribution are determined by the surface constrained all-atom solvent (SCAAS) model [180,184,185]. The surface region is embedded in a bulk continuum region with a dielectric constant of 80. The long range interactions are treated by the local reaction field (LRF) approach [186].

The MD simulations presented here were performed using the MOLARIS package [183]. In every case we first relaxed the system in a 50 ps long simulation of 1 fs time steps. The classical adiabatic charging FEP calculations were preformed in 11 steps of 50ps each for both forward ($\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}^0_{PCM})$) and backward ($\Delta G_{sol}(\mathbf{Q}^0_{PCM} \rightarrow \mathbf{Q}=0)$) processes. The results of both did not differ by more than 0.5 kcal/mol for the whole forward and backward charging processes. Their average is used as $\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}^0_{PCM})$.

The combined QM/MM calculations were preformed to estimate the terms of Eq. E-3.4-9. In each case we ran a 250 ps long simulation, which was sufficient to reach convergence of Eq. E-3.4-9 (appropriated Figure will be presented in Section 4.2). When performing a QM calculation we used the mean solvent potential averaged over 200 MD steps (m=200). Within the QM calculation, the solvent was represented with our three layers model of Figure F-3.4-3, with the radii of regions I, II and III being 10Å, 14Å and 16Å, respectively. All QM calculations were performed using the Gaussian03 package. The B3LYP exchange-correlation potential was used

with 6-31++G** basis sets. The Merz-Kollman scheme [187] with default atom radii was used to determine charges on atoms to be later used in the MD simulations. The hydrophobic and van der Waals contributions to the free energy of solute cavity were calculated using the ChemSol 2.1 program [182].

3.4.5 Validation of the approach

The complete validation of the accelerated QM/MM approach for performing calculations of solvation free energies have been conducted for a water molecule and a formate ion in water solution. The full study is presented elsewhere [179]. However, for the convenience of the Reader, the some representative results obtained for a formate ion solvated by the 16 Å sphere of explicit solvent molecules are presented here. In this example, the solvating water is represented by a two-region model, consisting of region I of 10 Å and region III of 16 Å.

The MD simulations were conducted for different m's, where m is the number of steps in MM simulation over which averaging of the solvent potential takes place. Due to the computational cost, 10 ps simulations were conducted for m = 1 or 10, 50 ps for m = 25 or 50, and finally 100 ps for m≥100. Since representing the effect of the fluctuating solvent charges by effective charges, and then updating the solute polarization by incorporating the effective charges in the solute Hamiltonian, slows down the QM calculation, the corresponding timings are noted.

The results are summarized in Table T-3.4-1 and the convergence of ΔG_{sol} for selected m's is presented in the Figure F-3.4-4. As seen in the Figure F-3.4-4, the ΔG_{sol} converges within the time of the simulation and do not vary more than 1 kcal/mol from the

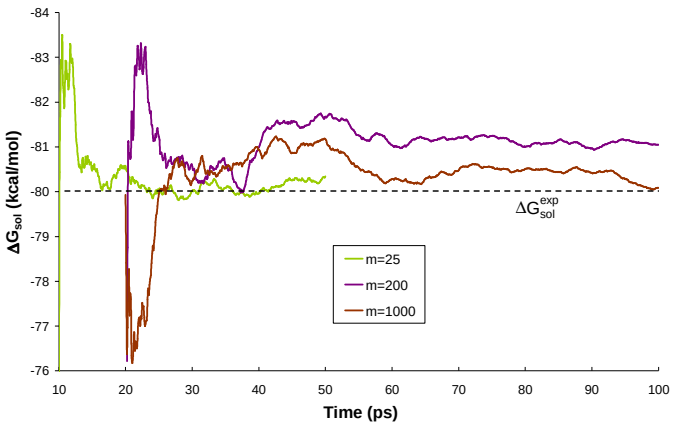


Figure F-3.4-4. Free energy of solvation of the formate ion along 50 and 100 ps simulation. For details, see Ref. 179.

experimental value. The reported timings show how efficient the accelerated approach is. For example, let's consider m=200, which will be used later for uracil anion. In this case time of the single QM calculation is 6.4 times longer then is case on m=1, however the QM is executed only once per 200 MM steps. Therefore the overall speed up is of factor 31 (!), that is, assuming that the time required to run m MD steps for MM subsystem is negligible comparing to QM. As shown in the table, further increasing of m, leads to slowing down the QM calculation as more external charges are introduced into the Hamiltonian. From our experience of the study [179], we conclude that m=200 is a good compromise and this will be used in the case of uracil anion presented in the Section 4.2.2.

Table T-3.4.-1. Free energy of solvation (ΔG_{sol} of Eq. (8) of Ref. 179) obtained during the last 8 ps of 10 ps simulations of $HCOO^-$ anion in water solution. t is the average time required for one QM step. The best estimates of the free energy of solvation (ΔG_{sol}^*) are obtained by taking average over the last 80% of simulation times (available only for m≥25). Average energies and free energies are given in kcal/mol, a time is given in seconds. The free energies of solvation include the ΔG_{cav} term of 1.5 kcal/mol.

m	ΔG_{sol}	ΔG_{sol}^*	t
1	-80.6	-*	27
10	-79.3	-*	41
25	-79.9	-80.1	45
50	-79.8	-79.6	42
100	-78.8	-79.8	73
200	-79.2	-80.9	174
500	-82.2	-80.3	918
1000	-77.8	-80.0	3379

* not converged within the time of calculation

4. Results Obtained by the Author and Discussion

4.1 Guanine

4.1.1 Screening for the Most Stable Tautomers of Anionic Guanine

We have tested our algorithms and the TauTGen program on the generation of guanine tautomers. The constraints are presented in Table T-4.1-1 and Figure F-4.1-1. In this case we created 23 sites available for hydrogen attachment. 17 sites were available for heavy atoms with Ns=1. Within these 17 sites, 4 sites were available to build rotamers of the N2 imino and O4 hydroxy groups and 2 sites were available for each of the C2, C4, C5 and C6 atoms to build stereoisomers with different positions of hydrogens in relation to the molecular plane. Additional 6 sites were available to build tautomers with two hydrogen atoms at N2, O4 and C8.

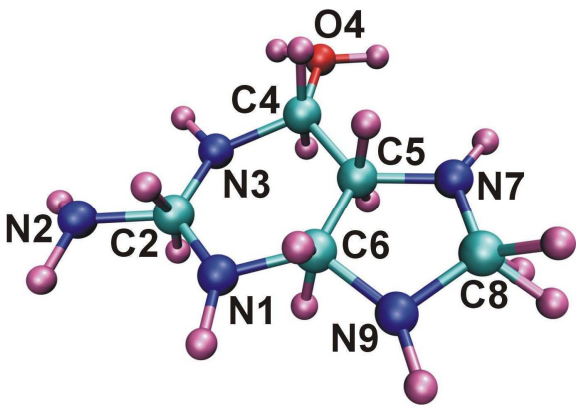


Figure F-4.1-1. Molecular framework of guanine with all sites for hydrogen attachment. The total number of sites differs from the number of sites in Table T-4.1-1 because some sites overlap [184].

Table T-4.1-1. Set of constraints used when searching for the most stable tautomers of anionic guanine.

	Minimum and maximum number of hydrogen atoms at heavy atom		Number of available sites for each number of hydrogens at heavy atom (N_s=1 and 2)		Asymmetric
Atom	Minimum	Maximum	N_s=1	N_s=2	atom
N1	0	1	1		
C2	0	1	2		Yes
N2	1	2	2	2	
N3	0	1	1		
C4	0	1	2		Yes
O4	0	2	2	2	
C5	0	1	2		Yes
C6	0	1	2		Yes
N7	0	1	1		
C8	1	2	1	2	
N9	0	1	1		

The final molecular frame and resulting sites are displayed in Figure F-4.1-1. The preparation time of a TauTGen input file was estimated to be about 15 minutes. The majority of this time is consumed by manually drawing the sites using the Molden software package [130] and naming them. This process could be automated if a larger number of molecules had to be studied.

Within these constraints TauTGen generated 499 unique structures. In the course of generation of tautomers, TauTGen generated initially 33649 distributions of five hydrogen atoms among 23 sites, from which only 9768 tautomers passed a check for the minimum and maximum number of hydrogens at each heavy atom. Only 907 of them passed the site constraint check. This number was later reduced to 499 in the course of the stereoconfiguration check.

The 499 structures of guanine were optimized at the B3LYP/6-31++G** level. An average calculation time for one structure was about 4 hours on a dual Intel Itanium2 node. With this speed of calculations, one needs ca. 2000 node hours to perform screening at the DFT level. A parallel execution of jobs might further shorten the "wall time" required for screening. Indeed, it took us about 2 weeks to screen 499 tautomers of guanine with an unprivileged access to a 128 dual Itanium2 nodes cluster in the TASK academic computer center in Gdańsk [188]. The wall time includes the time when jobs waited in the queuing system. A time scale of 2 weeks is comparable with the time required to perform one MP2/AVDZ calculation of numerical frequencies for anionic guanine on the same dual Itanium2 node.

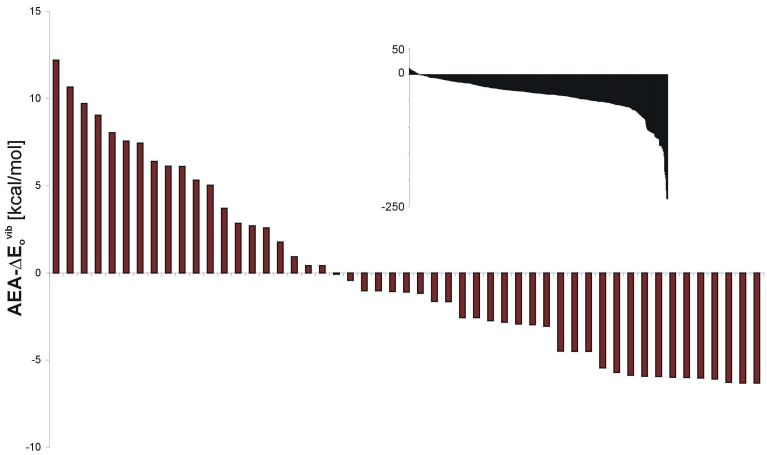


Figure F-4.1-2. Adiabatic electron affinity (AEA) for tautomers of guanine calculated at the B3LYP/6-31++G** level of theory. The values of AEA for 50 most stable tautomers are presented on larger plots. The smaller plots present the AEA of all tautomers. The tautomers are ordered according to the decreasing value of AEA.

With the goal being the determination of adiabatically bound anions of guanine, we compared the final

B3LYP energies for anions with the B3LYP energy of the neutral canonical tautomer at its optimal geometry. The histogram presenting the resulting AEA values for all structures are presented in Figure F-4.1-2. It might be seen that the values of AEA smoothly decrease for about 90% of the structures. A sudden decrease of AEA for the remaining 10% of the structures is related to the fact that some of these structures decompose in the course of geometry optimization.

In case of guanine and the B3LYP/6-31++G** level of theory employed, we found 14 anionic tautomers which were more stable than the canonical neutral. All of them were further studied at the MP2 and CCSD(T) levels with the AVDZ basis set. These calculations revealed that 13 tautomers support adiabatically bound anions. These results will be discussed in the following section.

4.1.2 Accurate Level Characterization of the Adiabatically Bound Anions

The accurate level characterization was performed for 14 adiabatically bound anionic tautomers identified at the screening. These tautomers are labeled Gx (x=1-14) and their adiabatic stability decreases as x increases. In addition, we had studied valence anions of the canonical tautomer (G) and of the most stable neutral tautomer (GN) as presented in Section 2.1.4.4.2. All 16 anionic tautomers are presented in the Figure F-4.1-3. In Table T-4.1-2 we present their energetic characteristics determined at the CCSD(T) level: AEAs (defined with respect to the neutral G) and VDEs [189]. In case of all 16 anions the excess electron occupies a π^* antibonding orbital which is delocalized over both rings. The singly occupied molecular orbitals will be presented in following sections, where they will be characterized in detail using chemoinformatic methods.

Thirteen anionic tautomers, G1-G13, remain bound at the CCSD(T) level. The anions of G1-G7 are adiabatically more strongly bound than any pyrimidine base studied so far., G2-, G3-, and G8- are biologically meaningful, i.e., they have a hydrogen atom at the N9 position, where a sugar unit is attached to guanine in DNA. The more stable G2- and G3- can not form Watson-Crick-type hydrogen bonds with cytosine, thus they may contribute to the instability of DNA. The stability of G8- can be further enhanced in DNA because its hydrogen binding sites are complementary

with those of cytosine. The valence anions based on the most stable neutral tautomers (G and GN) are adiabatically unbound by ca. 0.5 eV. The VDE values for adiabatically bound anions of guanine span a range 1.1-2.5 eV (Table T-4.1-2), and are much larger than VDEs of G and GN (0.5 and 0.2 eV, respectively).

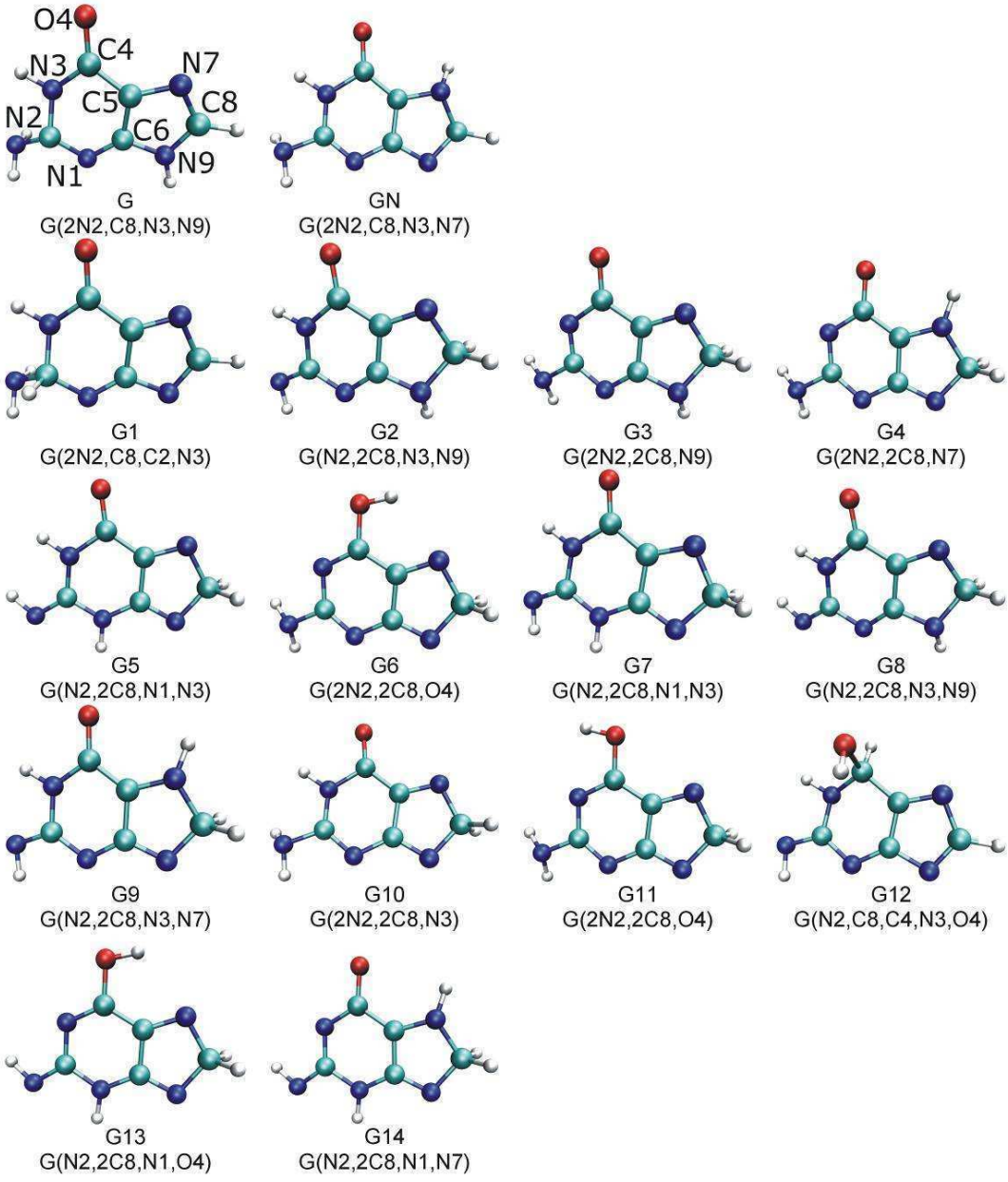


Figure F-4.1-3. The structures of 16 important tautomers of guanine. Two alternative notations are provided.

Table T-4.1-2. AEAs and VDEs (in eVs) for 16 selected anionic tautomers of guanine.

Tautomer	Structure DEA (G/GN)[a]	Energetics			
		Gas phase[b]		Water[c]	
		AEA	VDE	AEA[d]	VDE[e]
G	-/1	-0.459	0.585	1.329	3.381
GN	1/-	-0.503	0.212	1.496	2.643
G1	1/1	0.369	2.426	1.861	5.350
G2	1/2	0.365	1.604	2.026	4.424
G3	1/2	0.304	1.699	2.211	4.625
G4	2/1	0.278	2.205	2.157	5.094
G5	2/2	0.201	1.316	1.861	4.243
G6	2/2	0.174	1.484	1.671	4.343
G7	2/2	0.173	1.289	1.855	4.237
G8	1/2	0.165	1.617	1.953	4.566
G9	2/1	0.116	2.427	1.884	5.234
G10	1/1	0.104	1.137	2.018	4.295
G11	2/2	0.094	1.414	1.702	4.334
G12	2/2	0.078	2.542	1.462	5.434
G13	3/3	0.002	1.450	1.388	4.238
G14	3/2	-0.019	2.318	1.831	5.183

a Number of elementary DEA steps required to form the anionic tautomer from G and GN.
b Calculated at the CCSD(T)/AVDZ//MP2/AVDZ level
c Calculated at the B3LYP/6-31++G** level
d ε=78.
e ε=78 and 2 for the initial and final state, respectively.

4.1.3 Interpretation of the Photoelectron Spectrum of Anionic Guanine. The Formation Pathways

Negative ion photoelectron spectroscopy experiment was conducted in the Bowen group. It was done by crossing a mass-selected beam of negative ions with a fixed frequency photon source and energy analyzing the resultant photodetached electrons. This technique is governed by the energy-conserving relationship $h\nu$=EKE+EBE, where $h\nu$ is the photon energy, EKE is the measured electron kinetic energy, and EBE is the electron binding energy. The photoelectron spectrum of G- was measured with 3.493 eV photons and the result is presented in Figure F-4.1-4. A broad band (or a combination of bands) begins from ~0.5eV and reaches a local maximum at 0.8-1.1 eV. In a broad region of 1.2-2.3 eV the intensity is reduced to a

half of the local maximum intensity. Finally, the intensity steeply increases from 2.4 until 3.2 eV, which is an end of meaningful EBEs obtained with 3.493 eV photons. Because a dipole-bound anion state has a distinctive signature, wherein its spectrum is dominated by a single narrow peak at very low EBE (see Section 2.1.4.3), the spectrum of G- presented here is clearly not that of a dipole-bound state and in fact is due to a valence-bound state or states. The observed broad band spectral congestion is probably due to the simultaneous presence of several tautomers of G-.

The VDE values for adiabatically bound anions of guanine span a range 1.1-2.5 eV (Table T-4.1-2), which overlaps with this range of EBEs, where the PES spectrum has a significant intensity (Figure F-4.1-4). The intensity is negligible for EBEs smaller than 0.6 eV, thus in the region where the VDEs of G- and GN- are. This is consistent with adiabatic instability of the latter anions. Our computational results do not provide an interpretation of a feature in the PES spectrum that develops for EBEs exceeding 2.7 eV. Our library of molecular structures was limited to various tautomers of five-plus-six-member ring structures. The high EBE feature might be related to products with partially or completely broken double-ring structures. Indeed, it is known that guanine undergoes decomposition into molecular fragments in dissociative electron attachment experiments [76]. Further analysis of the high EBE feature would require PES experiments with photon energies larger than 3.493 eV.

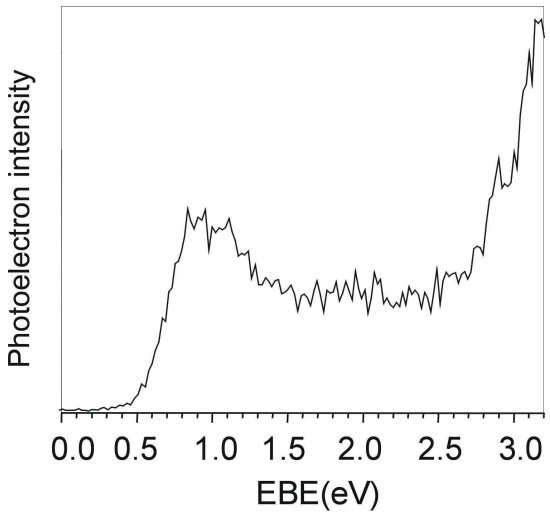


Figure F-4.1-4. Photoelectron spectrum of G- measured with 3.493 eV photons.

What might be formation pathways for the new anionic tautomers? For analogous tautomers of pyrimidine bases the barriers for intramolecular proton transfer are prohibitively large at standard conditions [85,86,89]. We suggest two formation pathways of the new anionic tautomers. First, they might be formed through intermolecular proton transfer. Second, dissociative electron attachment

(DEA) might facilitate their formation. The population of neutral guanine in the gas phase is dominated by two tautomers: GN (70%) and G (30%) (based on relative free energies presented in Section 2.1.4.4.2). Let us consider formation of G2-, which is the most stable among biologically relevant species. Scattering of an excess electron on the neutral target G might lead to:

$$G + e \rightarrow (G)^{*-} \rightarrow (G\text{-}H)^- + H\cdot, \qquad \text{(E-4.1-1)}$$

where $(G)^{*-}$ denotes a scattering state for an excess electron and $(G\text{-}H)^-$ denotes a deprotonated guanine (deprotonation at N2 amino group) in the ground electronic state. The attachment of a hydrogen atom to C8

$$(G\text{-}H)^- + H\cdot \rightarrow G2^- \qquad \text{(E-4.1-2)}$$

is found barrierless. The hydrogen atom attachment reaction is also barrierless for five other tautomers discussed in the following paragraph, whereas the barrier is only 0.03 kcal/mol for $(G10\text{-}H)^-$.

We calculated the overall thermodynamic barrier for the DEA step given by Eq. (E-4.1-1) for all hydrogen sites in G and GN at T= 0 K. Thus only electronic energies and zero-point vibrational corrections were considered. The results obtained at the B3LYP/AVDZ level of theory are: 20.0, 21.5, 22.3, and 60.8 kcal/mol for N9, N2, N3 and C8 of G, respectively, and 20.1, 20.7, 25.0, and 57.0 for N3, N7, N2, and C8 of GN, respectively. Thus the thermodynamic limit for (E-4.1-1) to proceed through the NH sites is approximately 0.87 eV and this is the minimum kinetic energy of electrons required to trigger the above process. Higher kinetic energies, preferably those that match positions of electronic resonances in G and GN, would also be appropriate and the excess energy will be distributed among translational, rotational, and vibrational energy levels of the products. The thermodynamic limit for DEA to proceed through the C8H site is much higher and amounts to ca. 2.47 eV.

The DEA formation pathway would favor formation of these anionic tautomers that are separated from G or GN by one DEA elementary step. The

numbers of DEA elementary steps required to form an anionic tautomer from G and GN are presented in Table T-4.1-2. For the dominating neutral tautomer, GN, only four adiabatically bound anions can be formed by a single DEA step followed by a H atom attachment, namely G1-, G4-, G9-, and G10-. From the second most populated neutral tautomer, G, only five adiabatically bound anions might be formed by a single DEA step: G1--G3-, G8-, and G10-. Notice that G10- and G1- can be formed from both G and GN. There is indeed a local maximum in the PES spectrum at ~1.0 eV, whereas the calculated VDE of G10- is 1.14 eV. A fingerprint of G1- is expected at 2.43 eV, but it would be masked by the unidentified feature that develops for EBEs larger than 2.7 eV. The only anionic tautomers that have VDEs in the 1.2-2.3 eV range and can be associated with a single DEA step are G2-, G3- and G8-. They all can be formed from the less populated G but not from the more populated GN. This might explain why the intensity in the PES spectrum is less intense in the 1.2-2.3 eV range.

4.1.4 Estimation of Stability in Water Solution

How relevant are our findings about new anionic tautomers for solvated species [190,191] ? We considered the effect of electrostatic stabilization by water using the polarizable continuum model and the results are presented in Table T-4.1-2. The new anionic tautomers are again more stable than G- and GN-, both adiabatically and vertically. Moreover, the ordering of Gx- according to their stability is different from that in the gas phase. The biologically relevant tautomers are the first, third, and fifth most stable (G3-, G2-, and G8-), see Table T-4.1-2. Hence the biologically relevant G3- and G2- might dominate in water solutions. The newly identified anionic tautomers have much larger VDEs than G- or GN-, a feature that is amenable to experimental verification. We believe that the new anionic tautomers will dominate not only in the gas phase but also in solvents and we suggest experimental studies in aprotic solvents to verify our predictions.

Application of our accelerated QM/MM method presented in Section 3.4 would provide much better estimation of relative stabilities of anionic tautomers of guanine. However, due to limited computer resources such study has not been performed so far.

4.1.5 Visual Comparison of Extension of the Selected SOMO Orbitals

In this section we demonstrate application of OpenCubMan (see Section 3.2) to visually compare the extension of the selected SOMO orbitals. For this purpose we consider the anion of the canonical tautomer (G) and the most stable anionic tautomer (G1), for which we calculated corresponding electron vertical detachment energies, 0.59 and 2.43 eV respectively. The structures of tautomers are reminded in Figure F-4.1-5a. When the SOMOs of G^- and $G1^-$ are visualized with the same CVs of 0.05 $bohr^{-3/2}$ (Figure F-4.1-5 b), then the corresponding fractions of electron, F_e's, are 0.629 and 0.694 *e*. Clearly, the 6.5% difference is significant and it is a manifestation of different SOMO extension expected from large difference in the values of VDE. We selected consistent values of F_e of 0.95 and 0.99 *e* and the resulting SOMOs are shown in Figure F-4.1-5c-d, respectively. The plots illustrate a larger extension of the SOMO of G- as it is more "bulky" than the SOMO of $G1^-$. The plot also suggests the different bonding/antibonding character of these orbitals, which leads to different values of VDE. This issue will be discussed in detail in the following sections.

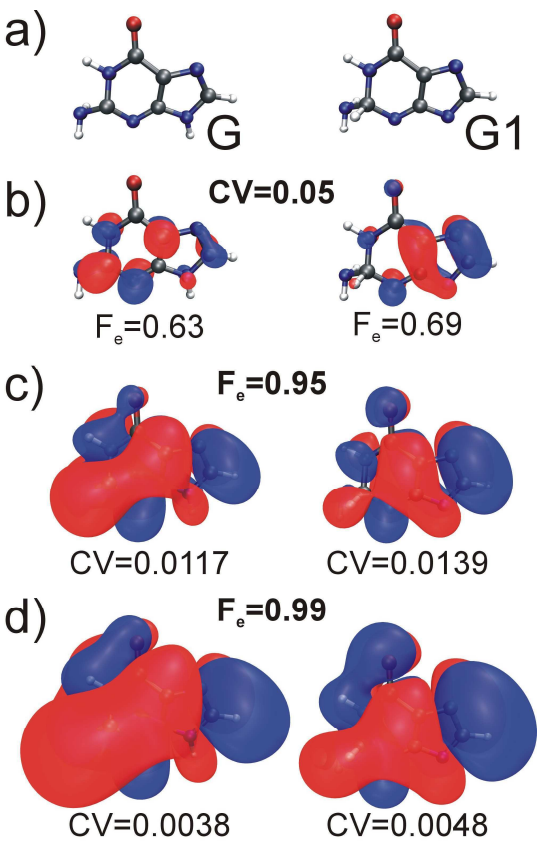


Figure F-4.1-5. (a) Molecular structures of valence anions of the canonical tautomer of guanine (G) and the most stable anionic tautomer (G1). (b) Singly-occupied molecular orbitals plotted using a contour value of 0.05 $bohr^{-3/2}$. (c) and (d) Singly-occupied molecular orbitals plotted F_e equal to 0.95 and 0.99, respectively.

Finally, we show a plot that illustrates differences in the spatial distribution of the excess electron in G^- and $G1^-$ as a demonstration of another capabilities offered by the OpenCubMan program. In Figure F-4.1-6 we superimpose both tautomers and the corresponding electron densities of SOMOs and we focus attention of nine slices,

which are selected by applying specific planes. For G^-, the majority of the excess electron is localized on the six member ring, whereas for $G1^-$ the excess electron is localized on the five member ring.

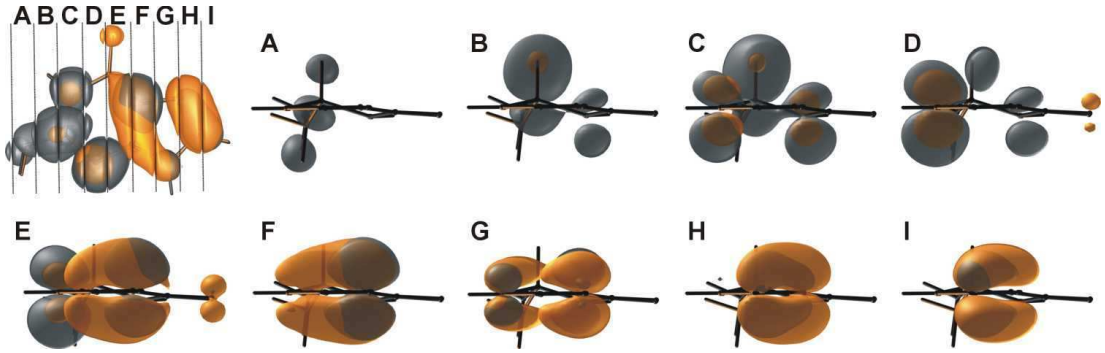


Figure F-4.1-6. Cross-sections of single-occupied molecular orbital densities corresponding to 0.6 *e*. The SOMO densities for G and G1 are superimposed and distinguished with grey and yellow color, respectively.

4.1.6 Chemoinformatics Analysis of QM Results

4.1.6.1 Comparing Buckling Modes of 16 Tautomers

In the case of the 16 important tautomers of anionic guanine, the excess electron occupies the π molecular orbital and might cause buckling of the molecular framework. The latter helps to compensate the antibonding effect of the SOMO orbital [69]. The dihedral angles determining non-planarity are presented in Table T-4.1-3. The buckling is strongest in the case of the anion of the canonical tautomer, which is mainly buckled in the 6-member ring region. The dihedral angle C6N1C2N3 of 28 degrees clearly indicates strong buckling. In the case of the GN tautomer, in turn, buckling is significant in the 5-member ring region confirmed by the C6C5N7C8 angle of 13 degrees. As will be shown in the following section, the buckling modes of G and GN correlate strongly with the excess electron localization on the 6- and the 5-membered ring, respectively.

Interestingly, all the G1-14 tautomers are much less buckled than G and GN, with deviations of dihedral angles from either 0º or 180º rarely exceeding 5º. The G1 and G12 seem to be exceptions: however, the large deviations of dihedral angles here originate from the change in hybridization of the C2 and C4 atom, respectively, from

sp^2 to sp^3. Therefore these cases cannot be compared with buckling coming from the antibonding effect of the occupied π orbital. Reasonably small non-planarity of all adiabatically bound anions implies that the antibonding character of SOMO orbital is much less evident than in the case of G and GN, or, alternatively, that the bonding character of the SOMO orbital is dominant. These hypotheses will be reviewed in the following sections as they would also justify the high stability of G1-G14 species.

The dendrogram in Figure F-4.1-7 presents the clustering of tautomers based on the dihedral angles defining buckling. It can be seen clearly that G, GN, G1 and G12 are most diversely buckled as they are clustered in the last four steps of the clustering process. During this process, at the stage of 7 clusters (marked with a dotted line in Figure F-4.1-7) we can see that besides the four diverse singleton tautomers, there are three more clusters containing similarly buckled tautomers: these clusters are: (G2,G3,G8), (G4,G9,G14) and (G5,G6,G7,G10,G11,G13).

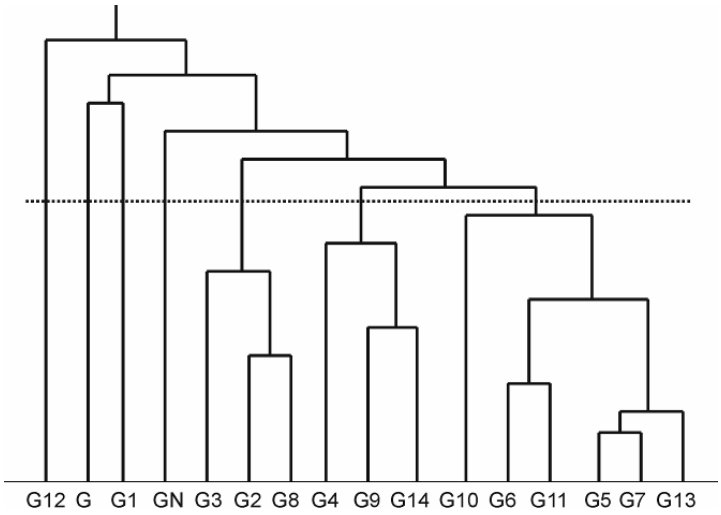


Figure F-4.1-7. Dendrogram presents HGAA clustering of 16 important anionic tautomers of guanine in terms of buckling mode of the molecule. The dotted horizontal line represents the seven-cluster partition.

The tautomers included in these clusters have similar 2D substructural features suggesting that the latter have a major influence on the buckling mode. The tautomers in the (G2,G3,G8) cluster have two hydrogen atoms at C8 and a hydrogen at N9. The G4, G9 and G14 in the next cluster have two hydrogen atoms at C8 and a hydrogen at N7. Finally the tautomers in the (G5,G6,G7,G10,G11,G13) cluster have two hydrogen atoms at C8 and have no hydrogen atoms on either N7 or N9. Interestingly, each of these three clusters contains tautomers with similar values of VDE: ca. 1.6, 2.3 and 1.3 eV, respectively. The big difference between the mean VDE

values for (G4,G9,G14) and the two remaining clusters correlates with the high instability of neutral G4, G9 and G14 reported in Table T-4.1-2.

Table T-4.1-3. Dihedral angles related to the buckling of the guanine molecule. The geometries of the tautomers were optimized at the MP2/APVDZ level of theory.

	Dihedral angles								
Tautomer	C4C5C6N1	C5C6N1C2	C6N1C2N3	C6N1C2N2	C6C5C4N3	C6C5C4O4	N1C6C5N7	N7C5C6N9	C6C5N7C8
G	-7.82	-7.48	27.89	163.44	1.44	184.47	180.51	0.97	-0.84
GN	1.02	-0.78	0.96	183.60	-1.20	178.49	175.53	-5.06	13.05
G1	-5.31	-8.94	27.68	148.78	-0.97	182.75	180.43	-0.51	0.04
G2	3.25	-1.34	-2.05	178.52	-1.44	178.98	183.00	4.42	5.23
G3	2.49	-1.12	-0.36	181.88	-2.26	178.88	182.86	4.50	4.73
G4	-2.58	0.64	1.64	183.12	2.29	182.71	177.85	-2.45	5.02
G5	-0.04	0.11	-0.11	179.90	-0.02	179.97	179.97	-0.01	0.01
G6	-0.20	-0.49	1.08	183.19	0.52	180.84	180.39	-0.01	-0.06
G7	-0.02	0.07	-0.07	179.95	-0.02	180.00	179.99	0.00	0.00
G8	3.90	-1.49	-2.69	177.94	-1.67	178.70	183.61	4.75	5.06
G9	-1.72	0.62	0.98	180.46	0.97	181.17	177.20	-2.85	6.08
G10	1.53	-2.03	1.94	184.04	-0.70	178.87	180.78	-0.08	-0.29
G11	0.14	-0.59	0.82	183.06	0.21	180.33	180.38	-0.01	0.05
G12	2.32	5.52	9.37	185.65	-22.04	101.88	179.48	0.24	0.14
G13	0.00	0.01	-0.01	180.01	0.00	180.00	180.00	0.00	0.00
G14	-2.54	1.27	0.33	179.91	2.11	182.34	176.20	-3.80	7.90

4.1.6.2 Comparing the Electron Density

Figure F-4.1-8 present the SOMO orbitals of 16 tautomers of anion guanine plotted wish consistent CV=0.03 bohr$^{-3/2}$. It shows that the excess electron does not seem to be distributed homogeneously in the various tautomers. The plots prepared for G and GN also suggest that these two tautomers are significantly different from the adiabatically bound anions of guanine. To verify these observations we performed clustering of the excess electron density represented by the orbital holograms (Table T-4.1-3) described in Section 3.3.3, and the resulting dendrogram is shown in Figure F-4.1-9.

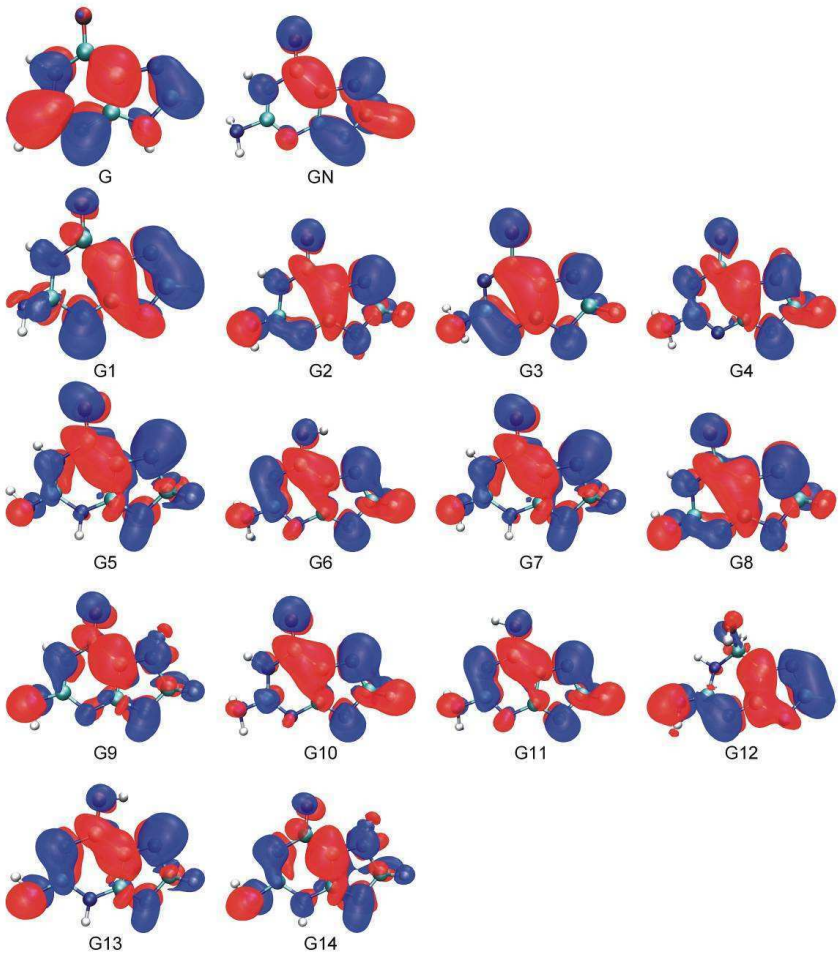


Figure F-4.1-8. Singly occupied molecular orbitals of 16 tautomers of guanine plotted with a spacing of 0.03 bohr$^{-3/2}$.

The clustering confirms the very high similarity between the orbitals of rotamers as they are clustered in the first three steps (G2 and G8, G5 and G7, G6 and G11). Looking down from the top of the dendrogram, it is clear that the orbitals of G and GN are dissimilar from the adiabatically bound anions as their branches join other branches during the last four clustering steps. Looking at the middle of the dendrogram (as denoted by the dotted line in Figure F-4.1-9), we can identify seven clusters containing tautomers with a similar distribution of the excess electron: G, GN and G14 are singletons, and the remaining clusters are (G6,G11,G13), (G4,G5,G7,G10), (G1,G12), (G2,G3,G8,G9). The tautomers of the (G6,G11,G13) and

(G1,G12) clusters have similar values of VDE. In the case of the remaining (G4,G5,G7,G10) and (G2,G3,G8,G9) clusters, the values of VDE vary from 1.1 to 2.4 eV, although significant similarity in the excess electron density is reported. The discrepancies in VDE have their origin in the difference of stability of the neutral counterparts (Table T-4.1-2). Another interesting pattern can be observed in the case of the G1 and G14 pair: here, the difference in stability of both anions and neutrals is almost 0.4 eV while the values of VDE differ by only 0.1 eV. Therefore we expect that these anions bind the excess electron in a "different way" (stronger by 0.3 eV), which is reflected in the dissimilarity of the SOMO density distributions.

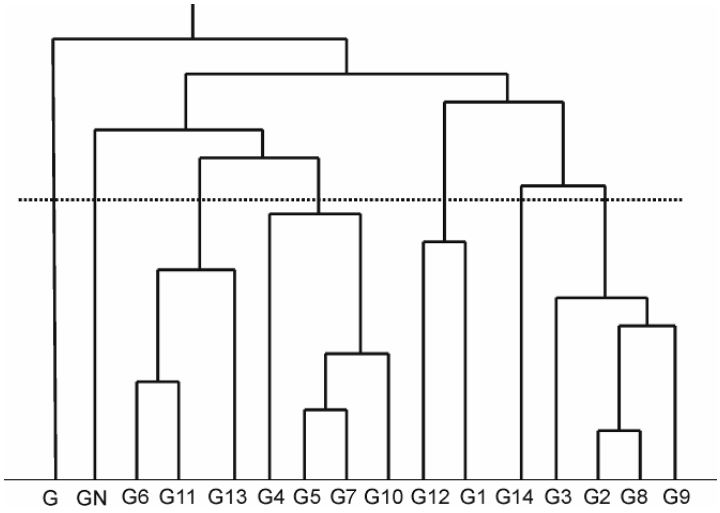


Figure F-4.1-9. Dendrogram presents HGAA clustering of SOMO orbital holograms of 16 important anionic tautomers of guanine. The dotted horizontal line represents the seven-cluster partition.

Table T-4.1-3. SOMO orbital holograms obtained for 16 anionic tautomers of guanine.

	Excess charge on atoms of molecular framework										
Tautomer	N1	C2	N3	C4	C5	C6	N7	C8	N9	N2	O4
G	0.298	0.172	0.121	0.012	0.133	0.009	0.084	0.034	0.041	0.088	0.007
GN	0.026	0.000	0.062	0.086	0.152	0.008	0.153	0.198	0.222	0.002	0.091
G1	0.132	0.016	0.036	0.024	0.174	0.067	0.335	0.062	0.123	0.007	0.025
G2	0.116	0.000	0.045	0.053	0.271	0.058	0.213	0.030	0.045	0.100	0.070
G3	0.143	0.001	0.038	0.039	0.243	0.093	0.190	0.032	0.079	0.047	0.096
G4	0.021	0.001	0.087	0.034	0.307	0.030	0.272	0.067	0.097	0.036	0.049
G5	0.009	0.000	0.069	0.110	0.255	0.028	0.207	0.058	0.115	0.027	0.122
G6	0.045	0.001	0.217	0.114	0.165	0.016	0.147	0.061	0.155	0.033	0.046
G7	0.011	0.000	0.066	0.099	0.261	0.033	0.211	0.059	0.120	0.028	0.112
G8	0.109	0.000	0.047	0.060	0.271	0.051	0.215	0.030	0.040	0.098	0.079
G9	0.109	0.000	0.075	0.042	0.264	0.008	0.217	0.031	0.016	0.192	0.047
G10	0.016	0.000	0.074	0.127	0.222	0.025	0.193	0.063	0.129	0.013	0.139
G11	0.053	0.000	0.231	0.134	0.139	0.013	0.130	0.061	0.158	0.028	0.051
G12	0.190	0.000	0.005	0.009	0.119	0.090	0.291	0.063	0.165	0.054	0.013
G13	0.008	0.000	0.200	0.102	0.216	0.017	0.187	0.049	0.098	0.083	0.040
G14	0.031	0.000	0.160	0.013	0.269	0.028	0.219	0.043	0.059	0.167	0.011

Besides looking at the excess electron distribution among atoms, we calculated the excess electron distribution among three fragments of the guanine molecule. For this purpose we divided the molecule into three regions: I – 6-member ring excluding the C5 and C6 atoms, II – the C5 and C6 atoms common for both 5- and 6-member rings, and III – 5-member ring excluding C5 and C6. For each of the regions I-III, the corresponding components of the orbital hologram were summed, giving information on the excess electron distribution in those regions. The standard deviation of values obtained for regions I-III indicates whether or not the excess electron is homogenously distributed among the regions. The results of this analysis are presented in Table T-4.1-4. The G and GN tautomers distinguish themselves from amongst the 16 tautomers as they have the most heterogeneous distribution of the excess electron, with the majority localized in fragment I and III, respectively. The G1 and G12 pair also exhibits a significant disproportion in distribution of the excess electron mainly localized in the N7C8N9 region. The G6 and G11 is another pair with a heterogeneous distribution of the excess electron, which is mainly localized in the region I. Both pairs are elements of the two clusters described above. In the case of the remaining tautomers, the excess electron is homogeneously distributed among the three regions of the molecule. These homogenous distributions of the excess electron are related to minor molecular buckling, as already reported in the previous section. They also correlate with the bonding and antibonding character of the SOMO orbital, as discussed in the following section.

Table T-4.1-4. Excess electron distribution over fragments I-III of guanine molecule.

	Excess charge on fragments			
Tautomer	I*	II**	III***	Std. div.
G	0.698	0.142	0.159	0.258
GN	0.267	0.160	0.573	0.175
G1	0.239	0.241	0.520	0.132
G2	0.383	0.329	0.288	0.039
G3	0.364	0.336	0.300	0.026
G4	0.227	0.337	0.436	0.086
G5	0.337	0.283	0.380	0.040
G6	0.456	0.181	0.363	0.114
G7	0.317	0.294	0.390	0.041
G8	0.393	0.322	0.284	0.045
G9	0.465	0.272	0.264	0.093
G10	0.369	0.247	0.384	0.062
G11	0.498	0.153	0.349	0.141
G12	0.272	0.209	0.519	0.134
G13	0.433	0.233	0.334	0.082
G14	0.382	0.297	0.320	0.036

*6-member ring (atoms: N1,C2,N2,N3,C4 and O4)

** fragment common for both 6- and 5- member ring (atoms: C5 and C6)

*** 5-member ring (atoms: N7,C8 and N9)

4.1.6.3 Comparing Bonding/Antibonding Character of SOMO

The bonding character holograms of the SOMO orbitals of the 16 important guanine tautomers, see defined in Section 3.3.4, and their summed total bonding and antibonding characters are presented in Tables T-4.1-5 and T-4.1-6, respectively.

In general, the bonding character holograms of SOMO orbitals seem to be good numerical representations of the plots of SOMO orbitals presented in F-4.1-8. For example, in the case of the G tautomer, the bonding character can be observed between C2N2, N7C8, and the antibonding character in the N1C2N3 fragment is clearly reflected in the G bonding character hologram. Moreover, the latter suggests the bonding character of SOMO in the C4C5C6 area, which is not clearly visible with the selected contour spacing used to plot the SOMO in F-4.1-8. All tautomers, with the exception of GN, have a bonding character in the region of C4C5C6. Some bonds, like C5N7 and C8N9, always have an antibonding character in all 16 tautomers.

Table T-4.1-5. Bonding character holograms for 16 tautomers of anionic guanine. The positive number means bonding character, otherwise antibonding character for particular bond between atoms of molecular frame.

	Bond between atoms of molecular frame											
Tautomer	N1C2	C2N3	N3C4	C4C5	C5C6	C6N1	C5N7	N7C8	C8N9	N9C6	C2N2	C4O4
G	-0.204	-0.115	-0.030	0.039	0.033	-0.046	-0.074	0.038	-0.028	0.014	0.071	-0.006
GN	0.000	0.000	-0.052	0.114	-0.021	-0.006	-0.128	-0.145	-0.182	0.022	0.000	-0.068
G1	-0.039	-0.020	-0.025	0.065	0.108	-0.080	-0.211	0.126	-0.060	0.062	0.003	-0.022
G2	0.004	0.002	-0.037	0.119	0.126	-0.076	-0.180	-0.057	-0.028	-0.041	-0.004	-0.054
G3	0.010	-0.001	0.004	0.097	0.150	-0.098	-0.168	-0.059	-0.037	-0.066	-0.005	-0.057
G4	0.002	0.007	-0.051	0.102	0.096	-0.016	-0.243	-0.113	-0.078	-0.053	-0.004	-0.036
G5	0.000	0.000	-0.060	0.167	0.084	-0.002	-0.181	-0.085	-0.080	-0.056	0.000	-0.095
G6	-0.005	0.011	-0.138	0.137	0.051	0.022	-0.128	-0.076	-0.095	-0.049	-0.003	-0.042
G7	0.000	0.000	-0.058	0.161	0.093	-0.004	-0.184	-0.086	-0.082	-0.062	0.000	-0.088
G8	0.004	0.002	-0.039	0.128	0.118	-0.070	-0.180	-0.056	-0.026	-0.036	-0.004	-0.060
G9	0.000	0.000	-0.052	0.105	0.045	-0.029	-0.196	-0.065	-0.021	-0.011	0.000	-0.037
G10	-0.001	0.002	-0.057	0.168	0.074	0.011	-0.168	-0.089	-0.089	-0.056	-0.001	-0.106
G11	-0.005	0.009	-0.151	0.136	0.043	0.023	-0.110	-0.072	-0.097	-0.045	-0.002	-0.047
G12	0.005	0.000	-0.001	0.021	0.103	-0.107	-0.171	0.125	-0.079	0.094	-0.003	0.002
G13	0.000	0.006	-0.128	0.148	0.061	-0.003	-0.154	-0.072	-0.067	-0.041	-0.004	-0.037
G14	0.000	0.000	-0.046	0.060	0.087	-0.027	-0.201	-0.079	-0.048	-0.040	0.000	-0.010

The SOMO orbital character in different tautomers is revealed during clustering of the bonding character holograms, as shown in Figure F-4.1-10. The G, GN, G1, G12, G13 and G14 are different from the remaining tautomers as they are clustered in the last four steps of the agglomeration process. At the seven-clusters level, other important clusters are (G6,G11), (G4,G5,G7,G10), (G2,G3,G8,G9). The formation of clusters similar to these three was also observed when clustering the orbital holograms, verifying the high degree of correlation between the two approaches. However, G13 is clustered here together with G14, whereas it was clustered with G6 and G11 when clustering using orbital holograms. This fact suggests that although G13 has a similar electron distribution to G6 and G11, the bonding and antibonding pattern of its SOMO orbital is different from other tautomers, with G14 being the most similar.

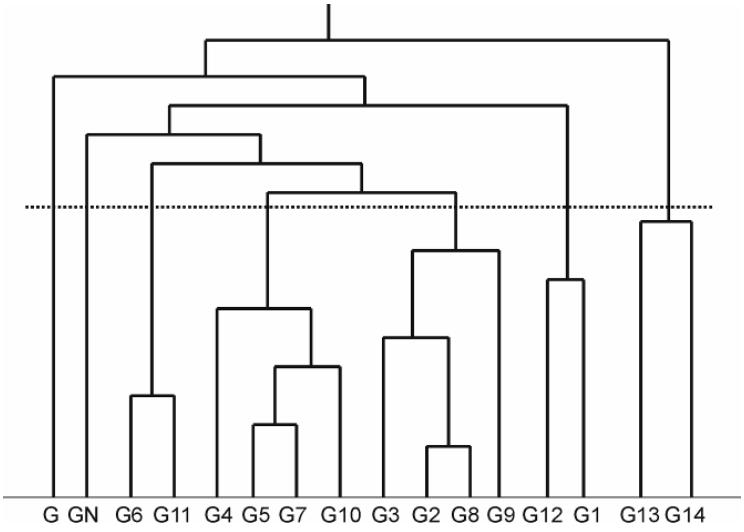


Figure F-4.1-10. Dendrogram presents HGAA clustering of bonding character holograms of 16 important anionic tautomers of guanine. The dotted horizontal line represents the seven-cluster partition.

The total bonding and antibonding character of the SOMO orbital calculated for all 16 tautomers is presented in Table T-4.1-6. The bonding character of SOMO in adiabatically bound anions, G1-G14, is in general larger than in G and GN. The total antibonding character of SOMO for guanine tautomers does not seem to follow any pattern. The G1 and G12 tautomers are distinct in the sense that they have the largest bonding character and quite a small antibonding character. The G1 and G12 tautomers are also distinct from the others because the bonding character of SOMO in the area of C5C6N9 and N7C8 can be observed together with no significant antibonding character in the remaining parts of the molecule (Figure F-4.1-8).

The analysis of the total bonding character provides evidence of correlation between the stability of the adiabatically bound anions and the bonding character of the SOMO orbital. It also supports the conclusion that a high level of planarity of adiabatically bound anions (minor buckling only) originates from the large bonding character of the π orbital occupied by the excess electron. Surprisingly, the antibonding character of SOMO orbital does not seem to correlate with either the stability or the planarity of the 16 anionic tautomers of guanine.

Table T-4.1-6. The total bonding (TOT) and total antibonding (TOT*) character of SOMO orbital derived from bonding character hologram.

Tautomer	TOT	TOT*	TOT+TOT*
G	0.195	-0.503	-0.308
GN	0.136	-0.603	-0.466
G1	0.364	-0.457	-0.093
G2	0.252	-0.477	-0.225
G3	0.262	-0.488	-0.227
G4	0.207	-0.593	-0.386
G5	0.252	-0.560	-0.308
G6	0.221	-0.537	-0.316
G7	0.253	-0.564	-0.311
G8	0.252	-0.472	-0.220
G9	0.150	-0.411	-0.261
G10	0.255	-0.565	-0.310
G11	0.212	-0.529	-0.318
G12	0.350	-0.362	-0.011
G13	0.215	-0.507	-0.291
G14	0.147	-0.450	-0.302

4.1.6.4 Summary

The analysis of quantum chemical data (namely, the excess electron density, the character of the SOMO orbital, and the geometrical parameters related with buckling) allowed us to identify features that either group together or distinguish between the 16 important tautomers. They can be summarized as follows:

- The G1-G14 tautomers are distinguished from the anions of the most stable neutrals, G and GN, as they have a more homogeneous distribution of the excess electron among the fragments of the molecule. The geometries of the G1-G14 are nearly planar due to the greater bonding character of the π orbital occupied by the excess electron. The common structural feature of G1-G14

that distinguishes them from G and GN is an additional hydrogen atom at C2, C4 or C8 when compared with the canonical structure. The most stable anions are also characterized by values of VDE in the range of 1.1 - 2.5 eV whereas the VDEs of G and GN are only 0.6 and 0.2 eV, respectively.

- The G1 and G12 tautomers are significally different from the remaining 12 most stable tautomers. When compared to the canonical tautomer, they have an additional hydrogen atom at C2 or C4, respectively. They are very unstable as neutral species. Therefore both G1 and G12 have large VDE values of ca. 2.4 and 2.5 eV, respectively. The buckling patterns of these molecules also distinguish them from the others. The tautomers are also different in terms of the excess electron distribution. The total bonding character of the SOMO orbital is reported to be the highest for these two tautomers.
- The biologically relevant tautomers, G2, G3 and G8, with hydrogen atom at N9 atom and two hydrogen atoms at C8, seem to be very similar to each other, in terms of 2D structure, the buckling mode, the excess electron distribution, the bonding character of SOMO and values of VDE of ca. 1.6 eV. They are however different from the following groups.
- The G5, G7 and G10 form another group of similar tautomers with two hydrogens at C8 but no hydrogen at either N7 or N9. They are similar to each other in terms of the buckling mode, the excess electron distribution, the bonding character of SOMO and values of VDE of ca. 1.2 eV.
- The G6, G11 and G13 form another group of similar tautomers with two hydrogens at C8, no hydrogen at either N7 or N9, but with a hydroxyl O4H group. They are similar to each other in terms of the buckling mode, the excess electron distribution and values of VDE of ca. 1.4 eV. However, the bonding character of SOMO of G13 is different from the remaining two.
- The G4, G9 and G14 tautomers are similar to each other in terms of 2D substructure (hydrogen at N7, two hydrogens at C8), the buckling mode and large values of VDE of ca. 2.4 eV. The latter correlated with significant instability of neutral counterparts. In terms of the SOMO density and SOMO

bonding/antibonding character, however, they are more similar to other groups than they are to each other.

The possibility to identify subgroups in a small set of 16 tautomers demonstrates a high correlation between the properties that we have considered. Obviously it is expected that the 2D structure of the molecule correlates with the excess electron density. The excess electron density is defined by the SOMO orbital, which defines the buckling mode. However, what is new and unexpected is that the protonation states of particular sites seem to have a larger effect on the excess electron density and related properties than do the others. These sites are C2, C4 and C8 carbons as well as N7 and N9 nitrogens.

4.1.7 Considerations on the Tautomeric Space of Anionic Guanine

4.1.7.1 Introduction

Application of the hybrid quantum mechanical-computational approach proved to be very effective in the identification of the most stable/adiabatically bound anionic tautomers of guanine. So far we concentrated on the characterization of these species, leaving another important output of the hybrid approach – the non-most-stable tautomers - behind. In the following section we demonstrate how the latter might be used to investigate the tautomer structure-stability relationship (SSR) by chemoinformatics techniques like clustering and substructural analysis. In the latter the sets of adiabatically bound and adiabatically unbound tautomers are compared to identify the set of structural features determining the stability. The studies of the tautomeric space using clustering methods can provide suggestions regarding the existence of an "island of stability" in the chemical space of guanine tautomers.

4.1.7.2 Technical Details of Analysis of the Library of Tautomers

The structure-stability relationship analysis was carried out on a reduced set of 165 tautomers (because of software limitations, multiple stereoisomers and rotamers were removed from the set of 499 tautomers as they would become

redundant in the fingerprint representation§ presented in this section, which does not take into account the spatial orientation of the bonded atoms). These tautomers were regenerated with the TauTGen program. We can precisely name the structures using a notation G(a,b…e), where a, b…e are the atoms to which the hydrogen atoms are attached: an example of this notation is presented in Figure F-4.1-2. It might be noted that some structures from the original set, like G6 and G11, are represented by only one structure, G(2N2,O4,2C8), in the reduced set.

All tautomers were represented by fingerprints - Boolean arrays indicating the presence or absence of 2D structural fragments specified in a dictionary of fragments (see Section 2.3.1). The latter was generated from the 2D connection tables (connectivity matrixes) of all tautomers in the library using a 2D descriptor generator program. The updated version (0.04) of TauTGen program was developed to export connection tables, however all bonds were currently assumed to be single. The generation of substructure dictionary as well as fingerprints themselves were performed using the BCI fingerprint package available from Digital Chemistry [192]. This software was not specifically written to work with structures differing only in the position of hydrogen atoms so it does not consider hydrogen positions in substructural fragments. Therefore, to process the library of tautomers we had to use a technical "trick" and substitute all hydrogen atoms with fluorine atoms (which did not exist in the initial dataset). The substructure fragments used in BCI fingerprints fall into the five fragment families presented in Section 2.3.1 (namely: augmented atom, atom sequence, atom pair, ring composition and ring fusion). The substructure dictionary derived from our set of 165 tautomers contained 1143 fragments (meaning that 1143-bit fingerprints were generated). In the latter step, to further extend the fingerprints, we introduced three types of nitrogen atoms (type 1 (amino/imino): N2; type 2 (in 6-member ring): N1 and N3; type 3 (in 5-member ring): N7 and N9). This allowed us to generate 1492 bit fingerprints, which are called extended-fingerprints throughout this Study.

---

§ To be more specific, they become redundant on the level of the subgraph isomorphism algorithm used to detect structural fragments

Having the tautomers represented by fingerprints, one can analyze them by comparing particular bits in the fingerprints, comparing the weighted value of a bit occurring in a subset of structures or comparing bits in modal fingerprints generated for a subset of structures: a weighted fingerprint is generated by summation of the corresponding bits in a set of structures and dividing by the number of structures in the set; and a modal fingerprint [193] can be derived from a weighted fingerprint by setting a bit to 1 if the average value of a bit is higher than a threshold, or 0 otherwise.

A similarity coefficient can be defined between structures coded by the fingerprints. The similarity coefficients were described in the Section 2.3.2. Having defined a similarity measure, molecules can be clustered by their relative similarity with various methods. The main clustering method used to cluster the tautomers was the HGAA method described in the previous section; the other clustering methods that we tried are k-means and Jarvis-Patrick, which are implemented in the BCI fingerprint software package.

4.1.7.3 Substructure Analysis

The structure-stability relationships analysis was started by checking whether any tautomer in the library of 165 tautomers had a unique structural feature (i.e., not shared with any other tautomer). Seven tautomers of this kind were identified (each with one unique bit set) but none of them was an adiabatically bound anion, or was within 20% of the most stable species.

Because we were not able to find any particular structural feature present in all adiabatically bound anions, we looked at the set of tautomers as belonging to two groups: the adiabatically bound tautomers (10 tautomers in the set of 165, which we will call *active group*) and the remaining set of 155 tautomers (*inactive group*). For each group we calculated modal and weighted fingerprints derived from the extended fingerprints. Depending on the threshold used when calculating the modal fingerprints, we are able to identify 184 (threshold 50%), 117 (60%), 100 (70%), 71 (80%) and 66 (90%) substructural features unique to the active group. The most natural approach would be to use a 50% threshold, which means that a particular 2D

substructure feature is considered to be existing (or not existing) when it is found in more than 50% of the tautomers in the library. However, even with the tightest threshold of 90%, 66 substructural features seem to be too many to be analyzed one by one.

When comparing the weighted fingerprints of the active and inactive groups, we looked for substructural features (bits) for which the value differed by 0.59 (as it was the first threshold to identify any feature), i.e., a particular feature is considered unique to a group when it is present 59% more often in one group than the other. For example, if one group has no molecules with the feature, the other group has to have this feature in at least 59% of molecules. We were able to identify 1 non-redundant feature absent in the group of adiabatically bound anions. It is:

1. Absent hydrogen atom separated by 3 bonds from carbon connected with 4 atoms (none of the adiabatically bound anions has this feature)

When we lowered the threshold to 0.58, we could identify six additional non-redundant features absent in the active group (only 1 in 10 adiabatically bound anions has this substructure feature). They are:

2. Absent carbon atom connected to four other atoms (2 carbons, 1 hydrogen and 1 nitrogen)
3. Absent sequence H-C-C, where both carbons are part of a ring
4. Absent sequence H-C-C-N, where C-C-N is part of the ring and N is in a 5-member ring
5. Absent sequence H-C-C-N-C, where C-C-N-C is part of the ring and N is in a 5-member ring
6. Absent sequence H-C-C-N-C-N, where C-C-N-C-N are part of the ring and both Ns are part of a 5-member ring
7. Absent sequence H-C-N-C-C-H, where C-N-C-C are part of a ring and N is in a 5-member ring.

The above substructure features coded in the fingerprints can be translated to more "chemist-friendly" form. For example, the substructural features unique to adiabatically bound anions of guanine are:

1. Absent hydrogen at C2, when there is a hydrogen at C4 or C6 (pt. 1)

2. Absent hydrogen at C4, when there is a hydrogen at C2 or C6 or N7 (pt. 1)
3. Absent hydrogen at C5, when there is a hydrogen at N3 or N1 or C8 or N9 (pt. 1)
4. Absent hydrogen at C6, when there is a hydrogen at C2 or C8 or N7 (pt. 1)
5. Absent hydrogen at C8, when there is a hydrogen at C5 or C6 (pt. 1)
6. Absent hydrogen at C5 (pt. 2)
7. Absent hydrogen at C4, C5 or C6 (pts. 3 and 4 and 5)
8. Absent hydrogen at C5 and C6 (pt. 6)
9. Absent hydrogen at C5 or C6, when there is a hydrogen at C8 (pt. 7)

Caution has to be kept when deriving general organic chemistry rules for predicting the stability of tautomers based on the appearance of substructural features. For example, if the features above were applied as strict rules, only 9 out of 10 adiabatically bound anions would be predicted – the tautomer G(N2,C8,C4,N3,O4) would not fulfill rule 7 (pts. 3-5) - and 55 out of 155 adiabatically unbound anions would be considered bound. Thus, rather than the rules being absolute criteria, they should be regarded as features, that are much more likely to appear in the set of adiabatically bound anions than in the set of remaining tautomers.

When the threshold for comparing weighted fingerprints was lowered further to 0.5, we could identify ten additional substructures unique to each group, but such a long list of features makes them difficult to analyze by eye. It might, however, be possible to apply machine learning methods and use these features to teach the computer to recognize the most stable tautomers: this possibility will be explored in future studies.

4.1.7.4 Clustering

In another approach to structure-stability relationship analysis, we clustered the set of 165 tautomers with the aim of identifying a cluster (or clusters) with a high concentration of adiabatically bound anions. Such cluster(s) would correspond to an island of stability in the chemical space of guanine tautomers. The tautomers were represented with extended fingerprints and the similarities calculated using the

Tanimoto coefficient. The snapshots of HGAA clustering at 165, 80, 60, 50, 30 and 15 clusters are presented in Table T-4.1-7. At each level of clustering we note the number of clusters containing adiabatically bound anions (active clusters) and the composition of these clusters (Table T-4.1-7).

Table T-4.1-7. Progress of HGAA clustering of 165 tautomers of guanine represented with extended fingerprints. The each level of clustering (N_{Cl}) the number of active clusters is noted (N_{ACl}) and composition of these clusters (number of adiabatically bound anions (n_a) and the total number of element in this cluster (n_{tot}). Active clusters with high concentration of adiabatically bound anions are marked with bold font.

N_{Cl}	N_{ACl}	Active cluster composition (n_a/n_{tot})
165	10	1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1 1/1
80	8	1/3 1/3 **2/3** 1/3 1/3 1/1 1/2 **2/2**
60	7	1/3 1/3 **3/6** 1/4 1/1 1/2 **2/2**
50	6	1/3 1/2 **3/6** 1/2 1/1 **3/4**
30	5	1/3 1/10 1/8 1/4 **6/10**
15	4	1/3 **7/24** 1/18 1/8

At the beginning of the clustering, there are 10 active clusters as all tautomers are singletons. As clustering proceeds, larger active clusters are formed, with higher concentrations of adiabatically bound anions. For example, at the level of 60 clusters, two active clusters already contain more than one adiabatically bound tautomer. As clustering proceeds further, a dominant active cluster is formed. For example, at the level of 30 clusters, this dominant cluster contains 6 adiabatically bound tautomers and only 4 less stable ones; and at the final level considered (15 clusters), this dominant cluster has 24 elements and contains 7 out of 10 adiabatically bound anions, including the 5 most stable tautomers (3% of the most stable anions). This cluster contains the following tautomers (adiabatically bound anions marked in bold):

G(N2,N1,C2,N3,C8), G(N2,N1,C2,N7,C8), G(N9,N2,N1,C2,C8), G(2N2,N1,C2,C8), G(N2,N1,C2,2C8), **G(N2,N1,N3,2C8)**, G(N2,N1,N7,2C8), G(N9,N2,N1,2C8), G(2N2,N1,2C8), G(N2,C2,N3,N7,C8), G(N9,N2,C2,N3,C8), **G(2N2,C2,N3,C8)**, G(N2,C2,N3,2C8), G(N9,N2,C2,N7,C8), G(2N2,C2,N7,C8), G(N2,C2,N7,2C8), G(N9,2N2,C2,C8), G(N9,N2,C2,2C8), **G(N2,N3,N7,2C8)**, **G(N9,N2,N3,2C8)**, **G(2N2,N3,2C8)**, G(N9,N2,N7,2C8), **G(2N2,N7,2C8)**, **G(N9,2N2,2C8)**.

The common features of the tautomers in this cluster are:

1. At least two hydrogen atoms distributed among C2 and C8 (0 or 1 atoms at C2, 1 or 2 atoms at C8)
2. No hydrogen atoms at C4, C5 or C6
3. No hydroxyl group

The remaining 3 out of 10 adiabatically bound anions, namely G(2N2,O4,2C8), G(N2,N3,C4,O4,C8) and G(N2,N1,O4,2C8), were found in three other clusters containing 3, 8 and 18 elements. These species are structurally distinct from the tautomers in the dominant active cluster.

The majority of adiabatically bound anions represented by extended fingerprints can be clustered together as they lie close to each other in the chemical space. In other words, there is a set of structural features that make them similar to each other and at the same time dissimilar to the remaining tautomers in the chemical space. These 2D structural features may determine affinity to the excess electron and they will be summarized in the following section.

Besides the HGAA method, we tried two other clustering methods, k-means and Jarvis-Patrick employing the Tanimoto similarity measure. We also tested 13 different similarity coefficients combined with HGAA. These combinations led to clusters with lower concentrations of adiabatically bound anions.

4.1.7.5 Summary and Discussion

The analysis of the tautomeric space of guanine suggests that the most stable tautomers have unique structural features. Some of these features could be identified using the substructural analysis approach. In the presented case the features are the absence of hydrogen atoms at C4, C5 and C6. Substructure analysis is, however, a statistical approach and the results can only be viewed as a suggestion of "more likeness" rather than a definite basis for categorisation. For example, the G12 tautomer has a hydrogen atom at C4, and thus does not follow the rule derived above.

The clustering technique does not provide any structural information directly. It does, however, suggest that some unique features of the most stable anions are

present as the formation of clusters with a high concentration of these species is observed. This suggests the existence of an "island of stability" in the tautomeric space, which may be used in the future to develop faster methods for the identification of the most stable tautomers. For example, one could reduce the number of calculations required to screen the tautomeric space to find the most stable species. Such a reduction could be achieved in the following steps:

- Generate *t* tautomers and the corresponding fingerprints.
- Perform hierarchical agglomerative clustering, stopping at one cluster.
- Choose a level of *P* clusters.
- Select *p* molecules, one molecule from each of *P* clusters.
- Run quantum chemical calculations for the *p* molecules to obtain their relative energy.
- Perform energy-based screening of the *p* molecules to get the *m* most stable molecules representing *M* clusters.
- Analyze the dendrogram representing the clustering. Identify *S* clusters at the level of *F* clusters (*F*<*P*) that contain *M* clusters.
- Run quantum chemical calculations for all molecules (*s*) contained in the *S* clusters.
- Perform energy-based screening of the *s* molecules to get the most stable tautomers.

The efficiency of such a procedure can be estimated using the data collected from the clustering of guanine tautomers presented in Table T-4.1-7. For example, *P*=80 clusters are selected and the energy based screening is performed for 80 representative molecules. In the worst case, we would find only two adiabatically bound anions (as only two clusters have 100% concentration of adiabatically bound anions). We select only one (*m*=1), the most stable molecule in the set of 80, and trace it in the dendrogram up to the level of 15 clusters (*F*=15). In this case only one cluster (*S*=1) is selected. The cluster has 24 elements and we need to characterize 23 of them at the QC level (one is already characterized). This procedure would require us to perform QC calculations for 103 tautomers instead of 165, giving 37.6% of CPU time

saving while retrieving all the five most stable tautomers (and seven adiabatically bound anions total)! If the safer option of *m*=3 is selected, we would end up with 126 calculations (80 at first stage and remaining elements of *S*=3 clusters) – 23.6% of CPU time saving and 8 adiabatically bound anions retrieved. Our ongoing work on anions of adenine and cytosine suggests that such optimized search procedures would successfully identify the most stable tautomers of these molecules: the general application of such procedure would, however, need further investigation. Such a procedure might be valuable for an initial rough exploration of tautomeric spaces of large molecules, or molecules for which little is known about the chemistry (either due to the nature of the molecule or the environment in which it is placed).

4.2 Uracil

4.2.1 Relative Free Energies in the Gas-Phase

Here we consider five tautomers of anionic uracil, that have been identified as the most stable in the gas phase (see Section 2.1.4.4.3). These tautomers are named U0-U4 (Figure F-2.1-5), with U0 being the canonical tautomer, and U1-U4 are the remaining tautomers ordered according to their decreasing stability in the gas-phase. The relative energies of these tautomers were calculated by Bachorz *et al.* using the state-of-the-art methodology (RI-MP2-R12) in conjunction CCSD(T), see Section 2.2.2). These results are summarized in Table T-4.2-1 and were already discussed in Sections 2.1.4.4.3 and 2.2.2.

We supplemented the results of Bachorz and coworkers with thermal corrections to Gibbs free energies of the gas phase anionic tautomers. The geometries were optimized and harmonic frequencies calculated at the MP2/AVDZ level (see Section 2.2.2). Thermal and entropic corrections were calculated at the temperature of 300K. The final relative free energies calculated with respect to the anionic canonical structure are summarized in Table T-4.2-1. These calculations were preformed using the Gaussian03 code.

The calculated relative free energies of five anionic tautomers in the gas phase are compared with the electronic energies corrected for zero-point vibration

energies (Table T-4.2-1). The discrepancies between these two thermodynamic characteristics are smaller than 0.9 kcal/mol. The $U1^-$ is the most stable anionic tautomer on the gas phase free energy surface, and the canonical tautomer, $U0^-$, is less stable by 2.54 kcal/mol. The remaining tautomers, $U2^-$-$U4^-$ are less stable than $U1^-$ by, respectively, 5.35, 9.84 and 10.70 kcal/mol.

Table T-4.2-1. The relative energies (ΔE), energies corrected for zero point vibrations (Δ(E+ZPVE)) and free energies (ΔG) of the most important anionic tautomers of uracil reported in the previous studies and compared with the results obtained in the current study. The energies and free energies (in kcal/mol) are calculated with respect to the anion of the canonical tautomer (U0).

	Δ(E+ZPVE)*	ΔE**	ΔG***	
Tautomer	Gas phase	Water	Gas phase	Water
$U0^-$	0.00	0.00	0.00	0.00
$U1^-$	-2.63	-0.88	-2.54	-3.64
$U2^-$	2.81	-1.09	2.81	-6.49
$U3^-$	7.66	6.12	7.31	1.29
$U4^-$	9.02	1.54	8.16	-5.54

* Ref. 88.
** Ref. 85.
*** This study.

4.2.2 Relative free energies in water solution

The calculation of solvation free energies of five anionic tautomers of uracil was done using Eq. (E-3.4-7), and all the contributing terms are shown in Table T-4.2-2. We started by performing the PCM solvation model calculations. They provided an estimation of residual charges and polarization energies. The polarization energies, $\Delta E_{QM}^{pol}(\mathbf{Q}_g^0 \rightarrow \mathbf{Q}_{PCM}^0)$, amount to 5.62-21.12 kcal/mol, with the largest value for $U4^-$. The polarization of a molecule is typically correlated with the distribution of residual charges. The charges are used in the MD simulations and therefore they influence the values of classical solvation free energies obtained using the free energy perturbation adiabatic charging approach. Indeed, the largest absolute value of $\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}_{PCM}^0)$ of 93.03 kcal/mol is reported for $U4^-$. The classical solvation energy, $\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}_{PCM}^0)$, amounts to -71.59 and -73.29 kcal for $U0^-$ and $U1^-$, respectively. Mid-range values of $\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}_{PCM}^0)$ are reported for $U2^-$ (-84.76

kcal/mol) and U3$^-$ (-79.75 kcal/mol). The ΔG_{cav} terms for all U0$^-$-U4$^-$ tautomers are similar and amount to ca. 2 kcal/mol with the largest deviation of 0.2 kcal/mol reported for U0$^-$.

The most challenging part in calculations of the free energy of solvation is the estimation of the terms contributing to Eq. (E-3.4-9). The sampling required to obtain the $\langle\ \rangle_{E(\mathbf{Q})}$ and $\langle\ \rangle_{E(\mathbf{Q}^0_{PCM})}$ terms of Eq. (E-3.4-9) was performed during a 250 ps MD simulation. We used our accelerated QM/MM approaches, where a calculation of the QM subsystem is performed in the mean field of the solvent averaged over 200 MD steps (Figure F-3.4-2). As demonstrated in Figure F-4.2-1, the convergence of Eq. (E-3.4-9) was reached within ca. 150 ps of the MD simulation. The contributions resulting from Eq. (E-3.4-9) span a range from -0.21 to -3.44 kcal/mol for the U0$^-$-U3$^-$ tautomers. A much larger value of -9.63 kcal/mol is reported for U4$^-$ and reflects a larger polarization of this tautomer. As expected [179], the $\langle\ \rangle_{E(\mathbf{Q})}$ and $\langle\ \rangle_{E(\mathbf{Q}^0_{PCM})}$ terms are quite similar as their difference does not exceed 1 kcal/mol.

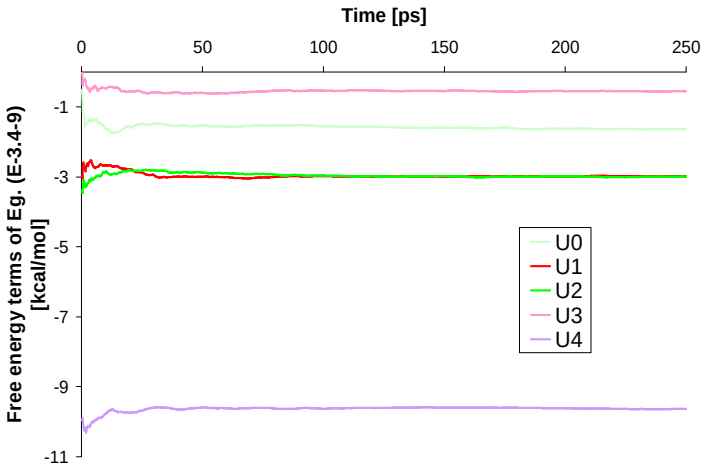


Figure F-4.2-1. Convergence of the results given by Eq. (E-3.4-9) during 250 ps simulations.

The final free energies of solvation calculated using Eq. (E-3.4-7) are -65.80 and -66.90 kcal/mol for U0$^-$ and U1$^-$, respectively. Larger values of ΔG_{sol} are reported for the remaining tautomers, with the largest of -79.50 kcal/mol for U4$^-$. In general, the strongest solvation develops for the tautomers with a proton transferred from N3 to a carbon atom (U2$^-$ and U4$^-$), which is consistent with the largest polarization of these molecules observed at the PCM model, and consistent with previously reported results (summarized in Table T-4.2-2) [86].

The solvation free energies obtained using Eq. (E-3.4-7) differ significantly from the PCM results obtained for the same molecular geometries. The difference is the largest, 7.0 kcal/mol, for the $U4^-$ tautomer and smaller but still significant for $U2^-$ and $U3^-$ of 1.5 and 3.6 kcal/mol, respectively. In the case of $U0^-$ and $U1^-$ the solvation free energies obtained with Eq. (E-3.4-7) agree with the PCM results within 0.7 kcal/mol. The differences between results of Eq. (E-3.4-7) and the PCM solvation energies reflect two facts: (i) in our study we use explicit water molecules; (ii) in our method we allow the solute wavefunction to respond to the field of the explicit solvent molecules.

Big differences in the values of ΔG_{sol} among the considered anionic tautomers are reflected in the relative free energies of anionic tautomers in water. The latter are calculated using Eq. (E-3.4-1) and are summarized in Table T-4.2-1. The most stable anionic tautomer in water solution is $U2^-$, followed by $U4^-$, which is less stable by 0.95 kcal/mol. The $U1^-$ and $U0^-$ are less stable than the most stable tautomer by 2.85 and 6.49 kcal/mol, respectively. The least stable from the set of five tautomers is $U3^-$, being 7.78 kcal/mol less stable than $U2^-$.

The current relative free energies differ significantly from the previously reported both the gas-phase and solution results [85,86]. The most important finding is that $U1^-$, $U2^-$ and $U4^-$ are more stable than $U0^-$. In particular, the $U4^-$ tautomer, which is unstable with respect to $U0^-$ by 8.2 kcal/mol in the gas phase, becomes the second most stable tautomer in solution, more stable then $U0^-$ by 5.5 kcal/mol. Our findings have some important consequences. Firstly, the evaluation of solvation energies for the anionic tautomers of nucleic acid bases should go beyond the initial screening at the PCM level. Secondly, some of the higher energy tautomers, like $U4^-$ for uracil, could become very stable in water solution, even though they are not dominant in the gas phase. Therefore, we should consider repeating the screening for the most stable tautomers of the anionic nucleic acids bases in solution (e.g. using our hybrid combinatorial-computational approach combined with the PCM model), or include in the solvation studies a larger number of promising gas-phase tautomers identified at the level the energy-based screening.

Finally, the fact that $U2^-$ and $U4^-$ are the most stable anionic tautomers in water solution might have important biological consequences. Both of these tautomers have hydrogen at the N1 atom where the sugar unit is connected in RNA. Therefore they can exist in the RNA environment and affect the structure of this nucleic acid. The consequences of these finding have been already discusses extensively in Refs. 85 and 86. It remains to be explored, which anionic tautomers of other NABs dominate in water solutions. These findings might become important for understanding of the effects of high energy radiation on DNA.

Another issue that should be explored is the chemical reactivity of these new anionic tautomers, both with the solvating water molecules and with the most common species in the RNA and DNA environments. For example, it is believed that anions of nucleic acid bases react with water forming hydrogenated nucleic acid bases and OH^- species[194]. The thermodynamics and kinetics of these processes will be studied in our future projects.

Table T-4.2-2. The contributions to the free energy of solvation: polarization energy ($\Delta E_{QM}^{pol}(\mathbf{Q}_g^0 \rightarrow \mathbf{Q}_{PCM}^0)$, hydrophobic (Hdr) and van der Waals (vdW) contributions to ΔG_{cav}, classical AC solvation free energy ($\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}_{PCM}^0)$) and the LRA terms of Eg. (E-3.4-9). The final solvation free energy ΔG_{sol} is calculated using Eq. (E-3.4-7). For comparison, the corresponding ΔG_{sol} values obtained with the PCM model are reported. All energies in kcal/mol.

			ΔG_{cav}		Terms of Eq. (9)		ΔG_{sol}	
Tautomer	$\Delta E_{QM}^{pol}(\mathbf{Q}_g^0 \rightarrow \mathbf{Q}_{PCM}^0)$	$\Delta G_{sol}(\mathbf{Q}=0 \rightarrow \mathbf{Q}_{PCM}^0)$	Hdr	vdW	$\langle\ \rangle_{E(\mathbf{Q})}$	$\langle\ \rangle_{E(\mathbf{Q}_{PCM}^0)}$	This study	PCM
U0$^-$	5.62	-71.59	6.30	-4.49	-2.12	-1.16	-65.80	-65.64
U1$^-$	7.35	-73.29	6.35	-4.31	-3.43	-2.57	-66.90	-66.22
U2$^-$	10.63	-84.76	6.33	-4.29	-3.44	-2.58	-75.10	-73.59
U3$^-$	6.48	-79.75	6.31	-4.30	-0.90	-0.21	-71.81	-68.25
U4$^-$	21.12	-93.03	6.35	-4.32	-10.03	-9.22	-79.50	-86.47

5. Conclusions

5.1 Developed tools and approaches

The aim of this Dissertation was to demonstrate the tools and approaches we developed to advance research on fragments of DNA. They include: (i) approaches for combinatorial-computational exploration of chemical (tautomeric) space in order to identify the most stable tautomers; (ii) approaches to analyze vast amounts of data harvested in quantum chemical calculations of (i); (iii) an algorithm and a software tool to improve visualization of molecular orbitals and related electron densities; (iv) approaches to improve efficiency of the combinatorial-computational searches for the most stable tautomers by using information on the studied chemical space; (v) methodology to predict accurate solvation free energies of molecules. The applications of the above tools and approaches were demonstrated on example studies of the anionic tautomers of selected nucleic acid bases, guanine and uracil, summarized in the following two sections.

5.2 Studies of anionic guanine in the gas-phase

The hybrid quantum mechanical-combinatorial approach turned out to be very efficient in the identification of the adiabatically bound anions of guanine. The set of 499 tautomers (including rotamers and stereoisomers) was pre-screened at the DFT level and 14 adiabatically bound anions were identified. 13 of them were verified as adiabatically stable at the CCSD(T) level of theory (Table T-4.1-2). The most stable anion is characterized by an AEA of 8.5 kcal/mol. The computed values of VDE for the new tautomers, 1.1-2.5 eV, are within the broad range of the dominant feature in the photoelectron spectrum. These new tautomers are obtained from conventional tautomers through N-to-C proton transfers, i.e., a proton is transferred from a nitrogen atom to a carbon atom. Three of adiabatically bound anions are biologically relevant since they have hydrogen atom at N9 position (where sugar unit is connected in the DNA). Seven of these tautomers are adiabatically more

strongly bound than any pyrimidine base studied so far. The results suggest that guanine might be the strongest excess electron acceptor among nucleic acid bases. This property might explain why the probability of capturing of an excess electron increases with the number of guanines in short single and double strands of DNA [195]. We conclude that ignored so far guanine might be critical to radiobiological damage of DNA and it might contribute to those chemical transformations of DNA that proceed through bound anionic states.

In the further study, 16 important anionic tautomers of guanine (Figure F-4.1-2), the 14 mentioned in the paragraph above and 2 anions of the most stable neutral tautomers, were analyzed in terms of molecular geometry and properties of single-occupied molecular orbital, such as the bonding/antibonding character and the distribution of excess electron density. To perform this analysis we developed SOMO orbital holograms and bonding character holograms, vectors containing information about the excess electron distribution related to these properties and derived using Bader's population analysis. By comparing the similarity of the excess electron density represented by orbital holograms, we demonstrated that the anions of the most stable neutral tautomers are significantly different from the most stable anionic tautomers. We observed a more homogeneous distribution of the excess electron among fragments of the molecule in the adiabatically bound anions. The bonding character of the π orbital occupied by the excess electron is greater in the latter, as indicated by the analysis of the total bonding character calculated from the bonding character holograms. As a result of this, the geometries of the most stable anions are nearly planar. The 14 most stable anionic tautomers were compared using the criteria of the buckling mode, the excess electron distribution and the bonding/antibonding character of SOMO. Five groups of similar tautomers could be identified. The correlation is observed between the protonation state of C2, C4, C8, N7 and N9 atoms and the assignment of a tautomer to a particular group under the considered criteria. For example, the most stable tautomer forms its own group as it is different

from others in terms of the considered criteria. All biologically relevant tautomers with hydrogen at N9 are found to be in one group.

Chemoinformatics methods, including substructural analysis and clustering, were used to identify the set of structural features that might determine the stability. The 2D substructure features of a set of 165 tautomers (excluding rotamers and stereoisomers) were coded into Boolean arrays (fingerprints), and then weighted fingerprints generated to represent groups of adiabatically bound and unbound anions. Substructural analysis based on the occurrence of particular substructure features represented in these fingerprints suggested that, in general, there are no hydrogens present at C4, C5 or C6 in the set of adiabatically bound anions.

Additionally, the hierarchical agglomerative clustering of the library of tautomers suggested that most of the adiabatically bound anions are very similar in terms of 2D structure, as represented by the fingerprints. For example, we identified a cluster of 24 tautomers including seven adiabatically bound anions (and all the five most stable tautomers). When compared with the canonical tautomer, the distinct substructural features of these tautomers are additional hydrogen atoms at C8 and/or C2 atoms. Formation of clusters with high concentration of the most stable tautomers, proves the existence of an "island of stability" in the tautomeric space. This information may be used in the future to develop more efficient methods for the identification of the most stable tautomers.

5.3 Studies of anionic uracil in water solutions

We reported the results of our calculations of the solvation free energies of the most stable anionic tautomers of uracil. These free energies were obtained using a two step approach. First the classical MD simulations were performed and the free energy perturbation adiabatic charging approach was employed to obtain classical solvation free energies. In this step it was assumed that the solvated molecules have the charge distributions given by the polarizable continuum model. In the second step the free energy of solvation was refined by taking into account the real, average

solvent charge distribution that reflects polarization caused by explicit water molecules used in the solvation model. This was done using our accelerated QM/MM simulations, where the QM energy of the solute was calculated in the mean solvent potential averaged over 200 MD steps. The results suggest that in water solution three of the recently identified anionic tautomers, namely $U2^-$, $U4^-$ and $U1^-$ (Figure F-2.1-5) are, respectively, 6.5, 5.5 and 3.6 kcal/mol more stable than the anion of the canonical tautomer. We also demonstrated that the solvation free energies of the most stable anionic tautomers obtained using our QM/MM approach are significantly different than the corresponding values obtained using the PCM model. In our opinion the PCM results can be successfully used for initial estimation of the solvation energies. We believe, however, that one would obtain more accurate results by full microscopic QM/MM calculations. Here we would like to emphasize that our conclusion is not completely trivial. It is obvious that microscopic models provide correct specific interactions with the solvent molecule. However, this does not guarantee better results since the use of the energy minimization or other poor sampling approaches would make the microscopic results completely unreliable. Similarly, the use of the semiempirical QM/MM (which would allow proper sampling as they are computational less expensive) might give poor results if the solute charges are not accurate. Only the use of *ab initio* QM/MM approaches with sufficient sampling leads to a stage where the QM/MM results start to be more reliable than those obtained by PCM or related approaches.

6. Closing Remarks

6.1 Inspiration for Future Studies

Here I would like to describe some of the current and future work inspired by the research conducted within this Dissertation project. By doing so, I would like to show the developed methodology, tools and approaches from a much broader perspective. I would like to convince the Reader that my work presented here is applicable to many other areas of research, exceeding the characterization of DNA fragments. The two main directions origin from the hybrid quantum mechanical-combinatorial approach presented in Sections 3.1 and 3.3 and the accelerated QM/MM method, described in Section 3.4. These two directions of ongoing and future development will be presented in the following two sections.

6.2 Development of Combinatorial-Computational-Chemoinformatics (C^3) Approaches

My capability to identify the most stable tautomers of a given molecule was an important accomplishment. The research was very well received by scientific community. Not only the results were published in leading chemical journals,[**] but also my efforts to develop the hybrid quantum chemical-combinatorial chemistry method were awarded with the ACS CINF/Fiz-Chemie Award for Scientific Excellence on the 2006 ACS meeting in San Francisco, and most recently, my work on the same topic was selected for the cover story of the coming issue of the Journal of Computational Chemistry. Stimulated by these facts I made another step forward in the development of my approach and I created software tools to study various Persistent Organic Pollutants (POPs).[††] These potentially dangerous substances exist in the environment as families of halogen substituted congeners. Congeners are

[**] See List of Publications.

[††] See Appendix III, article 3.

molecules based on the same carbon skeleton but differ by a substitution pattern, eg. 1-chloronaphthalene, 1,4-dichloronaphthalene, 1,3,8-trichloronaphthalene etc.. Due to enormous chemical variety of congeners, the identification of potentially toxic ones is practically only possible by the Qualitative Structure-Property Relationships (QSPR) studies. These require accurate molecular descriptors, including those calculated with quantum chemistry methods, to be compared with other well known pollutants like dioxins. With my combinatorial approach I was able to characterize libraries of ca. 100 000 congeners of few common POPs. Being limited by the available computer resources, I performed semiempirical electronic structure calculations for these libraries and the task was completed within a few days. This is a proof that hybrid combinatorial-electronic structure methods are becoming feasible, though they will remain computationally very intensive if accurate descriptors are required.

As the next step, I would like to extend my approach to deal with more complex chemical and materials science problems. Typically a molecular/materials designer has a specific property in mind, such as the emission/absorption wavelength or the particular band structure, and searches for stable molecules/materials that would display the desirable value of the targeted property. My main goal at the current stage is to develop algorithms and software tools that would facilitate combinatorial searches of this type based on the results of quantum mechanical electronic structure calculations. For example, the combinatorial-electronic structure approach could be used in the design of alloys. It could systematically screen various alloy compositions. The scientists would just have to decide on atom types included in the searches, and on ranges of concentrations. The software, with my approach implemented, would generate all possible alloys within the range of requested compositions, run required calculations, analyze the results, and finish up with a short list of alloy compositions that might have the requested property. In a similar manner the searches for novel electronic materials might be performed. For example, different compositions of materials might be screened to

find ones with the requested band gap. The targeted property might be complex, e.g., a specific band gap combined with a specific band offset when interfaced, e.g., with silicon.

The new hybrid methods may have an obvious application in the area of design and development of small-molecules like luminophores or indicators. My rough estimation based on the example presented below, shows that my approach can be already applied to real problems in molecular design. As an example lets consider the design of a novel luminescent molecule based on an efficient light emitter – acridine. Acridine consists of three conjugated rings with eight important substitution sites (although only four are easily experimentally accessible). As might be checked in a patent database, typically about 10-15 chemical groups are considered as facile substituents in the design of luminophores. Based on these observations I conclude that there might be from 40,000 to ca. 2,500,000,000 derivatives of acridine worth considering. The huge numbers that characterize the upper band are often referred to as the combinatorial explosion. The problem of combinatorial explosion can be omitted by "intelligent" scanning of chemical space briefly described in following paragraphs. In my opinion the number of required calculations in this case may be limited to ca. 10,000,000 which can be completed within a few days on a supercomputer.

As discussed, the hybrid quantum chemical/combinatorial approach has the potential to become a powerful tool in rational molecular/materials design. The approach involves three steps: (i) combinatorial generation of libraries of compounds, and (ii) screening of the libraries for the targeted property using electronic structure methods; (iii) analysis of generated data. The steps i-iii correspond to methodology employed, namely combinatorial, computational and chemoinformatics techniques, respectively. Therefore I propose to name this hybrid approach as "Combinatorial*Computational*Chemoinformatics", or just abbreviated as C^3 (or C-cube) approach.

The steps i-iii in C^3 approach do not have to be executed sequentially. For example, results from step (iii) can be feed-backed into (i) to perform navigated searches through chemical spaces. In my initial work on tautomers, I have already proposed one method of accelerating scanning of chemical spaces to avoid a combinatorial explosion (see Section 4.1.7.5). In some approximation, the idea is similar to a strategy in the popular pen and pencil game – Battleships. First, a coarse scan of the (chemical) space is performed. If an opponent's ship (a molecule/material with the targeted property) is identified then all the surrounding area is scanned with a fine grid. This method assumes that similar structure should be reflected in similar properties. In most of the cases this assumption is true at the resolution used in the initial screening of chemical spaces.

In another approach to accelerate scans of combinatorial libraries, I explore artificial intelligence methods, in which training is based on the results of coarse searches. The whole method could be viewed as self-training, in which initial results obtained from electronic structure calculations are accurate enough to guide an intelligent design of materials with superior properties. It is another major advantage in comparison with the current *in-silico* methods, which require costly parameterization and/or validation. I would also like to explore relations between this approach and genetic algorithms for optimizations of desired properties.

6.3 Applications and extensions of accelerated QM/MM approach

In the Section 3.4 I presented the accelerated QM/MM approach and its application in the prediction of the accurate solvation free energies of molecules in water. The same approach is now extended and applied to calculate free energies of molecules inside the proteins. The main advantage of the approach is that it can estimate the solute polarization that reflects the response to the potential from the surrounding protein. Such capability is critical to studies of charged transition states in proteins, which cannot be reliably using the classical approximations.

The QM/MM approach for averaging solvent potential to accelerate calculations currently assumes that the solute coordinates are fixed during the MD run. It can be however extended to the much more challenging (and arguably more important) case where both the solute and solvent are allowed to fluctuate. Such implementation, which is currently under development, could be used for evaluating the potential of mean force (PMF) in the solute-solvent configurational space without fixing the solute coordinate during the free energy calculations.

7. Bibliography

[1] Wild, D., Chemical Structure Association Trust Newsletter **2007**, 16, 5. http://www.csa-trust.org/news07/Issue16.pdf

[2] Watson, J. D.; Crick, F. H. C., *Nature* **1953**, 171, 737.

[3] Dąbkowska, I., PhD Thesis, University of Gdańsk, Gdańsk, **2005**.

[4] Rak, J.; Mazurkiewicz, K.; Kobylecka, M.; Storoniak, P.; Haranczyk, M.; Dabkowska, I.; Bachorz, R.A.; Gutowski, M.; Radisic, D.; Stokes, S.T.; Eustis, S.N.; Wang, D.; Li, X.; Ko, Y.J.; Bowen, K.H. "Stable Valence Anions of Nucleic Acid Bases and DNA Strand Breaks Induced by Low Energy Electrons" in "Radiation Induced Molecular Phenomena in Nucleic Acid: A Comprehensive Theoretical and Experimental Analysis" in the book series "Challenges and Advances in Computational Chemistry and Physics" edited by Jerzy Leszczynski. Springer. The Netherlands – in press.

[5] Sanche, L. *Mass Spectrom. Rev.* 2002, 21, 349-369.

[6] Sanche, L. *Eur. Phys. J. D* **2005**, 35, 367–390.

[7] Nikjoo, H.; Charlton, D.E.; Goodhead, D.T. *Ad. Space Res.* **1994**, 14, 161-180.

[8] Prise, K. M.; Folkard, M.; Michael, B. D.; Vojnovic, B.; Brocklehurst, B.; Hopkirk, A.; Munro, I. H. *Int. J. Radiat. Biol.* **2000**, 76, 881-90.

[9] von Sonntag C., "The chemical basis for radiation biology." London: Taylor and Francis, 1987.

[10] Zheng, Y.; Cloutier, P.; Hunting, D. J.; Sanche L.; Wagner, J. R. *J. Am. Chem. Soc.* **2005**, 127, 16592-98.

[11] Jay, A.; LaVerne, J.A.; Simon, M.; Pimblott, S.A. *J. Phys. Chem.* **1995**, 99, 10540-10548.

[12] Pimblott, S. M.; LaVerne J. A. *Rad. Phys. Chem*. **2007**, 76, 1244-1247.

[13] Pogozelski, W.K.; Tullius, T. D. *Chem. Rev.* **1998**, 98, 1089-1107.

[14] Burrows, C. J.; Muller, J. G. *Chem. Rev.* **1998**, 98, 1109-1152.

[15] Mason, N. private communication.

[16] Folkard, M.K.; Prise, M.; Vojnovic, B.; Davies, S.; Roper, M.J.; Michael, B.D. *Int. J. Radiat. Biol.* **1993**, 64, 651-658.

[17] Boudaiffa, B.; Cloutier, P.; Hunting, D. ; Huels, M.A.; Sanche, L. *Science* **2000**, 287, 1658 – 1660.

[18] Boudaiffa, B. ; Hunting, D.J. ; Cloutier, P. ; Huels, M.A. ; Sanche, L. *Int. J. Radiat. Biol.* **2000**, 76, 1209-1221.

[19] Huels, M.A.; Boudaiffa, B.; Cloutier, P.; Hunting, D.; Sanche, L. *J. Am. Chem. Soc.* **2003**, 125, 4467-4477.

[20] Zheng, Y.; Cloutier, P.; Hunting, D. J.; Wagner, J.R.; Sanche L. *J. Chem. Phys.* **2006**, 124, 064710-064719.

[21] Hotop, H.; Ruf, M.W.; Allan, M.; Fabrikant I.I. *At. Mol. Opt. Phys.* **2003**, 49, 85

[22] Pan, X.; Cloutier, P.; Hunting, D.; Sanche, L. *Phys. Rev. Lett.* **2003**, 90, 208102-208108.

[23] Martin, F.; Burrow, P.D.; Cai, Z.; Cloutier, P.; Hunting, D.J.; Sanche, L. *Phys. Rev. Lett.* **2004**, 93, 068101-4.

[24] Abdoul-Carime, H.; Cloutier, P. ; Sanche, L. *Radiat. Res.* **2001**, 155, 625-633.

[25] Pan, X. ; Abdoul-Carime, H. ; Cloutier, P. ; Bass, A. D. ; Sanche, L. *Radiat. Phys. Chem.* **2005**, 72, 193 - 199.

[26] Antic, D.; Parenteau, L. ; Lepage, M. ; Sanche, L. *J. Phys. Chem. B* **1999**, 103, 6611-6619.

[27] Simons, J. *Acc. Chem. Res.* **2006**, 39, 772 – 779 and the references therein.

[28] Berdys, J.; Skurski, P.; Simons, J. *J. Phys. Chem. B* **2004**, 108, 5800-5805.

[29] Berdys, J.; Anusiewicz, I.; Skurki. P.; Simons, J. *J. Phys. Chem. A* **2004**, 108, 2999-3005.

[30] Dąbkowska, I.; Rak, J.; Gutowski, M. *Eur Phys J D* **2005**, 35, 429 - 435.

[31] Gu, J.; Wang, J.; Rak, J.; Leszczynski, L. *Angew. Chem. Int. Ed.* **2007**, 46, 3479 – 3481.

[32] Bao, X.; Wang, J.; Gu, J.; Leszczynski, J. *Proc. Nat. Acad. Sci. USA* **2006**, 103, 5658 – 5663.

[33] Gu, J.; Wang, J.; Leszczynski, L. *J. Am. Chem. Soc.* **2006**, 128, 9322 – 9323.

[34] (a) Cai, Z.; Sevilla, M.D. *J. Phys. Chem.* B **2000**, 104, 6942-6949; (b) Messer, A.; Carpenter, K.; Forzley, K.; Buchanan, J.; Yang, S.; Razskazovskii, Y.; Cai, Z.; Sevilla, M.D. *J. Phys. Chem.* B **2000**, 104, 1128-1136; (c) Cai, Z.; Gu, Z.; Sevilla, M.D. *J. Phys. Chem.* B **2000**, 104, 10406-10411.

[35] Berlin, Y.A.; Burin, A.L.; Ratner, M.A. *J. Am. Chem. Soc.* **2001**, 123, 260-268 and references cited therein.

[36] Bixon, M.; Jortner, J. *J. Phys. Chem.* A. **2001**, 105, 10322-10328 and references cited therein.

[37] Voityuk, A.A.; Michel-Beyerle, M.E.; Rosch, N. *Chem. Phys. Lett.* **2001**, 342, 231-238.

[38] Löwdin, P.O., Rev. Mod. Phys. 1963, 35, 724.

[39] Tian, S.X.; Zhang, Z.J.; Chen, X.J. ; Xu, K.Z. *Chem. Phys.* **1999**, 242, 217 and references therein.

[40] Beak, P.; White, J. M. *J. Am. Chem. Soc.* **1982**, 104, 7073-7077.

[41] Fujii, M.; Tamura, T.; Mikami, N.; Ito, M. *Chem. Phys. Lett.* **1986**, 126, 583-587.

[42] Tsuchiya, Y.; Tamura, T.; Fujii, M.; Ito, M. *J. Phys. Chem.* **1988**, 92, 1760-1765.

[43] Wolken, J.K.; Turecek, F. *J. Am. Soc. Mass Spectrom.* **2000**, 11, 1065.

[44] Piacenza, M.; Grimme, S. *J. Comp. Chem.* **2004**, 25, 83.

[45] Morsy, M.A.; Al-Somali, A.M.; Suwaiyan, A. *J. Phys. Chem. B* **1999**, *103*, 11205.

[46] Ha, T.-K.; Gunthard, H. H. *J. Am. Chem. Soc.* **1993**, *115*, 11939.

[47] Civcir, P.Ü. *J. Mol. Struct*. (*THEOCHEM)* **2000**, *532*, 157.

[48] Yekeler, H; Özbakir, D. *J. Mol. Model.* **2001**, *7*, 103.

[49] Fogarasi, G., *J. Phys. Chem. A* **2002**, *106,* 1381-1390 and references cited therein.

[50] Hanus, M.; Kabelac, M.; Rejnek, J.; Ryjacek, F.; Hobza, P., *J. Phys. Chem. B* **2004**, *108,* 2087-2097 and references cited therein.

[51] Hanus, M.; Ryjacek, F.; Kabelac, M.; Kubar, T.; Bogdan, T.V.; Trygubenko, S.A.; Hobza, P., *J. Am. Chem. Soc.* **2003**, 125, 7678-7688.

[52] Leszczynski, J. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R., Ed.; John Wiley: Chichester, **1998**, 2951.

[53] Szczepaniak, K.; Szczesniak, M. *J. Mol. Struct.* **1987**, 156, 29-42.

[54] Dolgounitcheva, O.; Zakrzewski, V.G.; Ortiz, J.V. *J. Am. Chem. Soc.* **2000**, 122, 12304-12309.

[55] Leszczynski, J. *J. Phys. Chem.* A **1998**, 102, 2357-2362.

[56] Colominas, C.; Luque, F.J.; Orozco, M. *J. Am. Chem. Soc.* **1996**, 118, 6811-6821.

[57] Schiedt, J.; Weinkauf, R.; Neumark, D.M.; Schlag. E.W. *Chem. Phys.* **1998**, 239, 511-524.

[58] (a) Sevilla, M.D.; Becker, D. Royal Society of Chemistry Special Review on Electron Spin Resonance, **1994**, Chap. 5, Vol. 14, and references therein. (b) C. von Sonntag, Physical and Chemical Mechanism in Molecular Radiation Biology, edited by W.E. Glass and M.N. Varma, Plenum, New York, **1991**, and references therein. (c) Steenken, S. *Chem. Rev.* **1989**, 89, 503-520.

[59] Li, X.; Cai, Z.; Sevilla, M.D. *J. Phys. Chem.* A **2002**, 106, 1596-1603.

[60] Pullman, B.; Pullman, A. Quantum biochemistry (Wiley-Interscience, New York, **1963**).

[61] Younkin, J.M.; Smith, L.J.; Compton, R.N. *Theoret. Chim. Acta* **1976**, 41, 157-176.

[62] Compton, R.N.; Yoshioka, Y.; Jordan, K.D. *Theoret. Chim. Acta* **1980**, 54, 259-259.

[63] Adamowicz, L. *J. Phys. Chem.* **1993**, 97, 11122-11123.

[64] Oyler, N.A.; Adamowicz, L. *Chem. Phys. Lett.* **1994**, 219, 223-227.

[65] Roehrig, G.H.; Oyler, N.A.; Adamowicz, L. *Chem. Phys. Lett.* **1994**, 225, 265-272.

[66] Hendricks, J.H.; Lyapustina, S.A.; de Clercq, H.L.; Snodgrass, J.T.; Bowen, K.H. *J. Chem. Phys.* **1996**, 104, 7788-7791.

[67] Defrancois, C.; Abdoul-Carime, H.; Schermann, J.P. *J. Chem. Phys.* **1996**, 104, 7792-7794.

[68] Hendricks, J. H.; Lyapustina, S. A.; de Clercq, H.L.; Bowen, K.H. *J. Chem. Phys.* **1998**, 108, 8-11.

[69] Gutowski, M.; Dąbkowska, I.; Rak, J.; Xu, S.; Nilles, J.M.; Radisic, D.; Bowen, K.H. *Eur. Phys. J. D* **2002**, 20, 431-439.

[70] Mazurkiewicz, K.; Haranczyk, M.; Gutowski, M.; Rak, J.; Radisic, D.; Eustis, S.N.; Wang, D.; Bowen, K.H. *J. Am. Chem. Soc.* **2007**, 129, 1216–1224.

[71] Haranczyk, M.; Dąbkowska, I.; Rak, J.; Gutowski, M.; Nilles, J.M.; Stokes, S.T.; Radisic, D.; Bowen, K.H. J. Phys. Chem. B **2004**, 108, 6919–6921.

[72] Haranczyk, M.; Bachorz, R.; Rak, J.; Gutowski, M.; Radisic, D.; Stokes, S.T.; Nilles, J.M.; Bowen, K.H. J. Phys. Chem. B **2003**, 107, 7889–7895.

[73] Haranczyk, M.; Rak, J.; Gutowski, M.; Radisic, D.; Stokes, S.T.; Bowen, K.H. J Phys Chem B **2005**, 109, 13383–13391.

[74] Aflatooni, K.; Gallup, G.A.; Burrow, P.D. *J. Phys. Chem.* A **1998**, 102, 6205-6207.

[75] Bowen, K.H. and Shermann, J.P. - private communication.

[76] Illenberger, E.; A presentation during the Workshop on Interaction of Slow Electrons with Molecular Solids and Biomolecules, Harvard-Smithsonian Center for Astrophysics, October 16-18, 2003. http://itamp.harvard.edu/slowelectrons.html

[77] Li, X.; Bowen, K.H.; Haranczyk, M.; Bachorz, R.A.; Mazurkiewicz, K.; Rak, J.; Gutowski, M. *J. Chem. Phys.* **2007**, 127, 174309.

[78] Wetmore, S. D.; Boyd, R.J.; Eriksson, L.A. *Chem. Phys. Lett.* **2000**, 322, 129-135.

[79] Wesolowski, S.S.; Leininger, M.L.; Pentchev, P.N.; Schaefer III, H.F. *J. Am. Chem. Soc.* **2001**, 123, 4023-4028.

[80] Richardson, N.A.; Gu, J.; Wang, S.; Xie, Y.; Schaefer III, H.F. *J. Am. Chem. Soc.* **2004**, 126, 4404.

[81] Sevilla, M.D.; Becker, D.; Yan, M.; Summerfield, S.R., *J. Phys. Chem.* **1991**, 95, 3409-3415.

[82] Haranczyk, M.; Gutowski, M. *J. Am. Chem. Soc.* ***2005****, 127, 699-706*.

[83] Piuzzi, F.; Mons, M.; Dimicoli, I.; Tardivel, B.; Zhao, Q. *Chem. Phys.* **2001**, 270, 205-214.

[84] Mons, M.; Dimicoli, I.; Piuzzi, F.; Tardivel, B.; Elhamine, M. *J. Phys. Chem*. A **2002**, 106, 5088-5094.

[85] Bachorz, R.A.; Rak, J.; Gutowski, M. *Phys. Chem. Chem. Phys.* **2005**, 7, 2116-2125.

[86] Mazurkiewicz, K.; Bachorz, R.A.; Gutowski, M.; Rak, J. *J. Phys. Chem. B* **2006**, 110, 24696-24707.

[87] Bachorz, R.A.; Klopper, W.; Gutowski, M. *J. Chem. Phys.* **2007**, 126, 085101.

[88] Bachorz, R.A.; Klopper, W.; Gutowski, M.; Li, X.; Bowen, K.H. *submitted to J. Chem. Phys*.

[89] Haranczyk, M.; Rak, J.; Gutowski, *J. Phys. Chem. A* **2005**, 109, 11495-11503.

[90] Löwdin, P.-O. *Adv. Chem. Phys.* **1959**, 2, 207.

[91] Shabo, A.; Ostlund, N.S. *Modern Quantum Chemistry*, Dover Publications, **1996**, Chap. 4 and 6.

[92] a) Goddard III, W.A.; Dunning Jr, T.H.; Hunt, W.J.; Hay, P.J. *Acc. Chem. Res.* **1973**, 6, 368; b) Goddard III, W.A.; Harding, L.B. *Ann. Rev. Phys. Chem.* **1978**, 29, 363.

[93] Wahl, A.C.; Das, G. The multiconfiguration self-consistent field method, in *Methods of Electronic Structure Theory*, H. F. Shaefer III (Ed.), Plenum, New York, **1977**, p. 51.

[94] Cizek, J.; Paldus, J. *Physica Scripta*, **1980**, 21, 251.

[95] Piela, L. *Idee Chemii Kwantowej*, PWN, Warszawa, **2003**, p. 564-655.

[96] Hohenberg, P.; Kohn, W. *Phys. Rev. B,* **1964**, 136, 864.

[97] Kohn, W.; Sham, L.J. *Phys. Rev. A* **1965**, 140, 1133.

[98] a) Becke, A.D. *Phys. Rev. A* **1988**, *38*, 3098-3100; b) Becke, A.D. *J. Chem. Phys.* **1993**, *98*, 5648-5652; c) Lee, C.; Yang, W.; Paar, R.G. *Phys. Rev. B* **1988**, *37*, 785-789.

[99] Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*,724.

[100] Hehre, W. J.; Ditchfield, R.; Pople J. A. *J. Chem. Phys.* **1972**, *56*,2257.

[101] a) Rak, J.; Skurski, P.; Simons, J.; Gutowski, M. *J. Am. Chem. Soc.* **2001**, *123*, 11695; b) Skurski, P.; Rak, J.; Simons, J.; Gutowski, M. *J. Am. Chem. Soc.* **2001**, *123*, 11073; c) Kryachko, E. S.; Nguyen, M. T.; Zeegers-Huyskens, T. *J. Phys. Chem. A* **2001**, *105*, 1934; d) Chandra, A. K.; Nguyen, M. T.; Uchimaru, T.; Zeegers-Huyskens, T. *J. Phys. Chem. A* **1999**, *103*, 8853; e) Dkhissi, A.; Adamowicz, L.; Maes, G. *J. Phys. Chem. A* **2000**, *104*, 2112.

[102] Kendall, R.A.; Dunning Jr., T. H.; Harrison, R.J. *J. Chem. Phys.*, **1992**, *96*, 6796-6806.

[103] Taylor, P. R. in *Lecture Notes in Quantum Chemistry II*, (Ed.: B.O. Roos), Springer-Verlag, Berlin, **1994**.

[104] a) Knowles, P.J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1994**, *99*, 5219-5227. b) Deegan, J. J.O.; Knowles, P.J. *Chem. Phys. Lett.* **1994**, *227*, 321-326.

[105] a) Straatsma, T.P.; Aprà, E.; Windus, T.L.; Bylaska, E.J.; de Jong, W.; Hirata, S.; Valiev, M.; Hackler, M.; Pollack, L.; Harrison, R.; Dupuis, M.; Smith, D.M.A; Nieplocha, J.; Tipparaju V.; Krishnan, M.; Auer, A.A.; Brown, E.; Cisneros, G.; Fann, G.; Früchtl, H.; Garza, J.; Hirao, K.; Kendall, R.; Nichols, J.; Tsemekhman, K.; Wolinski, K.; Anchell, J.; Bernholdt, D.; Borowski, P.; Clark, T.; Clerc, D.; Dachsel, H.; Deegan, M.; Dyall, K.; Elwood, D.; Glendening, E.; Gutowski, M.; Hess, A.; Jaffe, J.; Johnson, B.; Ju, J.; Kobayashi, R.; Kutteh, R.; Lin, Z.; Littlefield, R.; Long, X.; Meng, B.; Nakajima, T.; Niu, S.; Rosing, M.; Sandrone, G.; Stave, M.; Taylor, H.; Thomas, G.; van Lenthe, J.; Wong, A.; Zhang, Z.; *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.6* (2004), Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA.

b) Kendall, R.A.; Aprà, E.; Bernholdt, D.E.; Bylaska, E.J.; Dupuis, M.; Fann, G.I.; Harrison, R.J.; Ju, J.; Nichols, J.A.; Nieplocha, J.; Straatsma, T.P.; Windus, T.L.; Wong, A.T. High Performance Computational Chemistry: an Overview of

NWChem a Distributed Parallel Application. *Computer Phys. Comm.* **2000**, *128*, 260-283.

[106] Gaussian 03, Revision C.02, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; Gaussian, Inc., Wallingford CT, 2004.

[107] Amos, R.D.; Bernhardsson, A.; Berning, A.; Celani, P.; Cooper, D.L.; Deegan, M.J.O.; Dobbyn, A.J.; Eckert, F.; Hampel, C.; Hetzer, G.; Knowles, P. J.; Korona, T.; Lindh, R.; Lloyd, A. W.; McNicholas, S.J.; Manby, F.R.; Meyer, W.; Mura, M.E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Rauhut, G.; Schütz, M.; Schumann, U.; Stoll, H.; Stone, A.J.; Tarroni, R.; Thorsteinsson, T.; Werner, H.-J. MOLPRO, a package of *ab initio* programs designed by H.-J. Werner and P. J. Knowles, version 2002.1.

[108]Young, D.C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems*, John Wiley & Sons, **2001**, Chap. 9.

[109] Weininger, D. *J. Chem. Inf. Comp. Sci.* **1988**, 28, 31-36.

[110] Trinajstic, N., *Chemical Graphs Theory*, Boca Raton, CRC Press, **1983**.

[111] Dittmar, P.G.; Farmer, N.A.; Fisanick, W.; Haines, R.C.; Mockus, J. *J. Chem. Inf. Comp. Sci.* **1983**, 23, 93-102.

[112] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. *J. Chem. Inf. Comp. Sci.* **1985**, 25, 64-73.

[113] Downs, G.M.; Gillet, V.J.; Holliday, J.D.; Lynch, M.F. *J. Chem. Inf. Comp. Sci.* **1989**, 29, 207-214.

[114] Adamson, G.W.; Cowell, J.; Lynch, M.F.; McLure, A.H.W.; Town, W.G.; Yapp, A.M. *J. Chem. Doc.* **1973**, 13, 153-157.

[115] Willett, P. *J. Chem. Inf. Comp. Sci.* **1979**, 19, 159-162.

[116] Carbo, R.; Leyda, L.; Arnau, M. *Int. J. Quant. Chem.* **1980**, 17, 1185.

[117] Holliday, J.D.; Hu, C-Y.; Willett, P. *Comb. Chem. High. Throughput Screening* **2002**, 5, 155-166.

[118] Downs, G.M.; Barnard, J.M. Clustering methods and their uses in computational chemistry, Reviews in Computational Chemistry, vol. 18, chapter 1, edited by K.B. Lipkowitz and D.B. Boyd, Wiley-VCH, **2002**.

[119] Willett, P. Similarity and clustering in chemical information systems; Research Studies Press, **1987**.

[120] Jarvis, R.A.; Patrick, E.A. *IEEE Transactions in Computers* **1973**, C-22, 1025-1034.

[121] Forgy, E. Biometrics **1965**, 21, 768-780.

[122] Hay, B.P.; Firman, T.K. *Inorg. Chem.* **2002**, *41*, 5502-5512.

[123] Jorgensen, W.L. *Science* **2004**, *303*, 1813-1818.

[124] Sayle, R.; Delany, J. Canonicalization and Enumeration of Tautomers, in Innovative Computational Applications. Institute for International Research, Sir Francis Drake Hotel, San Francisco, 25-27 October 1999

[125] TAUTOMER, developed and distributed by Molecular Networks GmbH, Erlangen, Germany (http://www.mol-net.com)

[126] Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. *J Recept Signal Transduct.* **2003**, *23*, 361-371.

[127] TauTGen: Tautomer Generator Program. Available at http://tautgen.sf.net.

[128] Tomasi, J.; Persico, M. *Chem. Rev*. **1994**, *94*, 2027-2094.

[129] GOT: Gaussian Output Tools available at http://gaussot.sf.net.

[130] Schaftenaar, G.; Noordik, J.H. *J. Comp.-Aided Mol. Des.* **2000**, 14, 123-34.

[131] Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33-38.

[132] Black, G.; Didier, B.; Bisethagen, T.; Feller, D.; Gracio, D.; Hackler, M.; Havre, S.; Jones, D.; Jurrus, E.; Keller, T.; Lansing, C.; Matsumoto, S.; Palmer, B.; Peterson, M.; Schuchardt, K.; Stephan, E.; Sun, L.; Taylor, H.; Thomas, G.; Vorpagel, E.; Windus, T.,;Winters, C. Ecce, A Problem Solving Environment for Computational Chemistry, Software Version 3.2.5, 2006, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA.

[133] Handy, N.C.; Marron, M.T.; Silverstone, H.J. *Phys. Rev.* **1969**, *180*, 45-48.

[134] Katries, J.; Davidson, E.R. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 4403-4406.

[135] Rauk, A.; Armstrong, D.A. *Int. J. Quant. Chem.* **2003**, *95*, 683-696.

[136] Li, X.; Sanche, L.; Rauk, A.; Armstrong, D.A. *J. Phys. Chem. A* **2005**, *109*, 4591-4600.

[137] Rauk, A.; Armstrong, D.A. *Eur. Phys. J. D* **2005**, *35*, 217-224.

[138] Open-source Cubefile Manipulator Program (OpenCubMan) is available free of charge at SourceForge archive: http://opencubman.sourceforge.net (accessed Feb 27, 2008).

[139] Gellert, W.; Gottwald, S.; Hellwich, M.; Kästner, H.; Künstner, H. (Eds.). *VNR Concise Encyclopedia of Mathematics, 2nd ed.* New York: Van Nostrand Reinhold, 1989.

[140] Henkelman, G.; Arnaldsson, A.; Jónsson, H. *Comput. Mater. Sci.* **2006**, 36, 254-360.

[141] Sanville, E.; Kenny, S.D.; Smith, R.; Henkelman G. *J. Comp. Chem.* **2007**, 28, 899-908.

[142] Atkins, P.; De Paula, J. *Atkins' Physical Chemistry*, p.387, Oxford University Press, 8th Edition, 2006.

[143] Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, 16, 1170.

[144] Sierka, M.; Sauer, J. *J. Chem. Phys.* **2000**, 112, 6983.

[145] Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, 103, 227-249.

[146] Shurki, A.; Warshel, A. *Advances in Protein Chemistry* **2003**, 66, 249-312.

[147] Gao, J. *Acc. Chem. Res.* **1996**, 29, 298-305.

[148] Bakowies, D., Thiel, W. *J. Phys. Chem* **1996**, 100, 10580-10594.

[149] Field, M. J.; Bash, P. A.; Karplus, M. *J. Comp. Chem.* **1990**, 11, 700-733.

[150] Friesner, R.; Beachy, M. D. *Curr. Op. Struct. Biol.* **1998**, 8, 257-262.

[151] Monard, G.; Merz, K. M. *Acc. Chem. Res.* **1999**, 32, 904-911.

[152] Garcia-Viloca, M.; Gonzalez-Lafont, A.; Lluch, J. M. *J. Am. Chem. Soc.* **2001**, 123, 709-721.

[153] Marti, S. A., J.; Moliner, V.; Silla, E.;Tunon, I.; Bertran, J. *Theor. Chem. Acc.* **2001**, 3, 207-212.

[154] Field, M. *J. Comp. Chem.* **2002**, 23, 48-58.

[155] Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, 105, 569-585.

[156] Lyne, P. D.; Mulholland, A. J.; Richards, W. G. *J. Am. Chem. Soc.* **1995**, 117, 11345-11350.

[157] Warshel, A.; Sussman, F.; Hwang, J.-K. *J. Mol. Biol.* **1988**, 201, 139-159.

[158] Muller, R. P.; Warshel, A. *J. Phys. Chem.* **1995**, 99, 17516-17524.

[159] Strajbl, M.; Hong, G.; Warshel, A. *J. Phys. Chem. B* **2002**, 106, 13333-13343.

[160] Olsson, M. H. M.; Hong, G.; Warshel, A. *J. Am. Chem. Soc.* **2003**, 125, 5025-5039.

[161] Zhang, Y.; Liu, H.; Yang, W. *J. Chem. Phys.* **2000**, 112, (8), 3483-3492.

[162] Rod, T. R. *Phys. Rev. Lett.* **2005**, 94, 138302.

[163] Liu, W.W.; Doren, D.J. *J. Phys. Chem. B* **2003**, 107, 9505-9513.

[164] Iftimie, R.S.; Schofield, J. *J. Chem. Phys.* **2003**, 119, 11285-11297.

[165] Crespo, A.M.; Estrin, D.A. *J. Am Chem. Soc.* **2005**, 127, 6940-6941.

[166] Sakane, S.Y.; Liu, W.B. *J. Chem. Phys.* **2000**, 113, 2583-2593.

[167] Pradipta, B. *J. Chem. Phys.* **2005**, 122, 091102.

[168] Hu, H.; Lu, Z. Y.; Yang, W. T. *J. Chem. Theor. Comput.* **2007**, 3, 390-406.

[169] A. Warshel, Computer *Modeling of Chemical Reactions in Enzymes and Solutions.* John Wiley & Sons: New York, 1991.

[170] Luzhkov, V.; Warshel, A. *J. Comp. Chem.* **1992**, 13, 199-213.

[171] Tapia, O.; Goscinski, O. *Mol. Phys.* **1975**, 29, 1653-1661.

[172] Cramer, C. J.; Truhlar, D. G., Continuum Solvation Models: Classical and Quantum Mechanical Implementations. In *Reviews in Computational Chemistry*, Lipkowitz, K. B.; Boyd, D. B., Eds. VCH: New York, 1995; Vol. 6, pp 1-72.

[173] Tomasi, J.; Bonaccorsi, R.; Cammi, R.; Delvalle, F. J. O. *J. Mol. Struct. (Theochem)* **1991**, 80, 401-424.

[174] Rivail, J. L.; Rinaldi, D., *Computational Chemistry: Review of Current Trends*. World Scientific Publishing: Singapore, 1995.

[175] Sanchez, M.L.; Martin, M.E.; Galvan, I.F.; del Valle, F.J.O.; Aguilar, M.A. *J. Phys. Chem. B* **2002**, 106, 4813-4817.

[176] Sanchez, M. L.; Martin, M. E.; Aguilar, M. A.; Del Valle, F. J. O. *J. Comp. Chem.* **2000**, 21, 705-715.

[177] Mendoza, M.L.S.; Aguilar, M.A.; del Valle, F.J.O. *J. Mol. Struct.(Theochem)* **1998**, 426, 181-190.

[178] Sanchez, M.L.; Aguilar, M.A.; delValle, F.J.O. *J. Comp. Chem.* **1997**, 18, 313-322.

[179] Rosta, E.; Haranczyk, M.; Chu, Z.T.; Warshel, A. *J. Phys. Chem. B*, accepted.

[180] Warshel, A.; Sussman, F.; King, G. *Biochemistry*, **1986**, 25, 8368-8372.

[181] Lee, F. S.; Chu, Z.T.; Bolger, M.B.; Warshel, A. *Protein Engineering*, **1992**, 5, 215-228.

[182] Florián, J.; Warshel, A. *J. Phys. Chem. B* **1997**, 101, 5583 - 5595.

[183] Lee, F.S.; Chu, Z.T.; Warshel, A. *J. Comp. Chem.* **1993**, 14, 161-185.

[184] Warshel, A.; Russel, S.T.*Q. Rev. Biophys.* **1984**, 17, 283-422.

[185] King, G.; Warshel, A. *J. Chem. Phys.* **1989**, 91, 3647-3661.

[186] Lee, F.S.; Warshel, A. *J. Chem. Phys.* **1992**, 97, 3100-3107.

[187] Besler, B.H.; Merz, K.M.; Kollman, P.A. *J. Comp. Chem.* **1990**, 11, 431-439.

[188] Academic Computer Center in Gdańsk (TASK), http://www.task.gda.pl.

[189] The relateive stabilities of various anionic tautomers are given by the AEA values. 1eV = 23.0693 kcal/mol = 96.5220 kJ/mol.

[190] Sevilla, M.D.; Becker, D. Special Review on Electron Spin Resonance, (Royal Society of Chemistry 1994), Vol. 14, Chap. 5, and references therein.

[191] Steenken, S. *Chem. Rev*. **1989**, *89*, 503.

[192] http://www.digitalchemistry.co.uk

[193] Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. J. Chem. Inf. Comput. Sci. **2004**, 44, 1177-1185.

[194] M.D. Sevilla, D. Becker, *Royal Society of Chemistry Special Review on Electron Spin Resonance*, Chap. 5, Vol. 14 (1994), and references therein.

[195] Ray, S.G.; Daube, S.S.; Naaman, R. *Proc. Nat. Acad. Sci. USA* **2005**, 102, 15-19.

Appendix I

Research Articles Contributing to the Background Information

Proton transfer in anions of nucleic acid bases

In the following articles we present results of our investigations of the intramolecular proton transfer reaction in the neutral and anionic nucleic acid bases. In the first article on guanine, we consider four tautomers, which are the most stable as neutral species. We characterize their valence and dipole-bound anions by providing the values of AEA and VDE calculated at our standard level of theory (CCSD(T)/AVDZ). A summary of this study was provided in Section 2.1.4.4.2 of the Dissertation.

A similar study is conducted for 1-methylcytosine (Article No. 2). However, in this case we consider also few new anionic tautomers resulting from proton transfer between nitrogen and carbon atoms. These new tautomers are found more stable than the anions of the most stable neutral tautomers.

Appendix II

Research Articles Presenting Methodology and Results of the Dissertation

The articles presented in the Appendix II summarize the methodology, approaches, tools and results that are presented in this Dissertation. Article No. 1 presents the hybrid combinational-quantum mechanical approach, the TauTGen and Gaussian Output Tools software packages and their applications in the identification of the most stable anionic tautomers of nucleic acid bases. Articles No. 2 and 3 present the results of accurate characterization of the most stable anionic tautomers of guanine. Article No. 4 presents the developments of chemoinformatics approaches applied in the characterization of the tautomeric space of anionic guanine (presented in Sections 3.3 and 4.1.6-7).

Article No. 5 presents our findings on the visualization of molecular orbitals and the related densities of systems that significantly differ by the electron density extension. The article describes also OpenCubMan program, and presents its application on the example of anionic HCl-NH_3 complex mentioned in Section 3.2.1.

In the last two articles (Articles No. 6 and 7) we summarized our work on the advancement of the QM/MM methodology. Article No. 6 presents the accelerated QM/MM method (described in the Section 3.4), its various implementations and corresponding validations. Article No. 7 describes an improved QM/MM method used to characterize anionic tautomers of uracil in water solution.

No.	Article reference	Pages
1	**M. Haranczyk**, M. Gutowski - "Quantum Mechanical Energy –Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program" – *Journal of Chemical Information and Modeling 47* (2007) 686-694.	171 – 179
2	**M. Haranczyk**, M. Gutowski – "Finding Adiabatically Bound Anions of Guanine through Combinatorial-Computational Approach" – *Angewandte Chemie Int. Ed.* 44 (2005) 6585-6588.	181 - 183

Appendix III

Research Articles Supplementing and Enhancing the Dissertation

The research articles contributing to the Appendix III supplement and enhance the Dissertation. The first two articles present the results of our searches for the most stable anionic tautomers of adenine and cytosine preformed using our approach described in the Section 3.1. In the case of adenine (Article No. 1) we could identify one adiabatically bound anion and several other stable anionic tautomers. These computational results were positively verified by the photoelectron experiments performed in the Bowen group. Similar computational studies conducted for cytosine has not suggested the existence of adiabatically bound anions. However, some new, stable anionic tautomers have been identified both computationally and experimentally.

The Article No. 3 presents an in-depth comparison of 13 similarity coefficients with applications in clustering and dissimilarity selections. This study was conducted on the 2D fingerprint representations of ca. 20,000 molecules included in the MDL Drug Report Database.

In the Article No. 4 we present a recent extension of the hybrid combinatorial-computational approach. We present the ConGENER package that can be used for generation and quantum mechanical characterization of libraries of congeners. The project summarized in the Article No. 4 was briefly described in Section 6.2.

No.	Article reference	Pages
1	**M. Haranczyk**, M. Gutowski, X. Li, K.H. Bowen – "Bound anionic states of adenine. Theoretical and photoelectron spectroscopy study" –*Proceedings of National Academy of Science (PNAS) 104* (2007) 4804-4807.	251 – 254
2	X. Li, K.H. Bowen **M. Haranczyk**, R.A. Bachorz, K. Mazurkiewicz, J. Rak, M. Gutowski –"Photoelectron spectroscopy of adiabatically bound valence anions of rare tautomers of the nucleic acid bases" – *Journal of Chemical Physics* 127 (2007) 174309.	255 – 260

3	**M. Haranczyk**, J. Holliday – "Comparison of Similarity Coefficients for Clustering and Compound Selection" –*Journal of Chemical Information and Modeling* – in press – DOI: 10.1021/ci700413a.	261 – 271
4	**M. Haranczyk**, T. Puzyn, P. Sadowski – "ConGENER – A Tool for Modeling of the Congeneric Sets of Environmental Pollutants" –*QSAR and Combinatorial Science*- in press.	273 - 280

Appendix IV

Research Articles on Characterization of Larger DNA Fragments

Cylinder Projections of Electrostatic Potential around Intact and Damaged DNA Fragments

The research articles presented in the Appendix IV summarize the development of a method for visualization and analysis of electrostatic potential (EP) around intact and damaged DNA fragments. We perform projection of the electrostatic potential from a complicated 3D surface of the DNA molecule onto the walls of a cylinder, which is a natural approximation to the shape of short fragments of DNA. The resulting 2D EP maps are presented as bitmaps and can be easily analyzed and compared by eye. Moreover, in the most recent article (Article No. 3), we present an extension of the method, where the EP maps are analyzed using image analysis techniques (e.g. automatic feature detection and measurements).

The presented technique is applied to analyze changes of electrostatic potential resulting from the occurrence of the DNA lesions (like 8-oxo-guanine, 8-oxo-adenine or thymine glycol). Our studies suggest that the presence of lesions is reflected in the reorganization of the countercations and phosphate groups neighboring the lesions.

3	**M. Haranczyk**, G. Lupica, I. Dąbkowska, M. Gutowski – "Cylindrical Projection of Electrostatic Potential and Image Analysis Tools for Damaged DNA. The Substitution of Thymine with Thymine Glycol" – *Journal of Physical Chemistry* B 112 (2008) 2198-2206**.**	297 - 305