INCONSISTENCIES IN *AB INITIO* EVALUATIONS OF NON-ADDITIVE CONTRIBUTIONS OF DNA STACKING ENERGIES

BY

QIN KEN

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
SCHOOL OF INFORMATION SCIENCE
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
ACADEMIC YEAR 2020

INCONSISTENCIES IN *AB INITIO* EVALUATIONS OF NON-ADDITIVE CONTRIBUTIONS OF DNA STACKING ENERGIES

A Dissertation Presented

By

QIN KEN
1620406

Submitted to
School of Information Science
Japan Advanced Institute of Science and Technology
In partial fulfillment of the requirement for the degree of
DOCTOR OF PHILOSOPHY

Approved as to style and content by

Supervisor:	Prof. Ryo Maezono, Ph.D.
Second Supervisor:	Prof. Hiroyuki Iida, Ph.D.
Committee Members:	Prof. Yoshifumi Oshima, Ph.D.
	Assoc Prof. Kenta Hongo, Ph.D.
	Prof. Satoshi Tojo, Ph.D.
External Examiner:	Prof. Tamio Oguchi , Ph.D.

MARCH 2020

Abstract

INCONSISTENCIES IN *AB INITIO* EVALUATIONS OF NON-ADDITIVE CONTRIBUTIONS OF DNA STACKING ENERGIES

by

Qin Ken

The main research methods on biomolecules are experiments at present. However, experimental methods are challenging to describe the microscopic details of biomolecules, including molecular-level dynamics and basic functional states. For drug development, such as cancer-targeted drugs, precise matching of biological targets is required. Therefore, computer simulation plays an essential role in understanding and studying the structure and biological function of biomolecules. The description and reproduce of the bonding itself, due to intermolecular forces, is a considerable challenge for *ab initio* methods. The non-additivity in the interactions is expected in intermolecular bindings due to the induced polarization by quantum fluctuations, such as van der Waals (vdW) forces. Non-additivity is a more difficult subject than intermolecular interaction itself, and it has long been far from the mainstream research field and has not been well analyzed yet. We evaluated the non-additive contributions in the intermolecular interactions in B-DNA stacking by using fixed-node diffusion monte carlo (FNDMC) methods. DNA molecules are the basis of biological, genetic variation. In the previous

calculation methods, standard Density Functional Theory (DFT), Hartree-Fock (HF), the sign of the non-additive contribution is positive. While the Self-Consistent Field (SCF) level non-additivity is mostly positive and tiny, the non-additive contributions described by FNDMC are both positive and negative signs. The negative sign is found to be reasonable, which might be supported by a simple model analysis based on the London theory. It would, however, be premature to draw a conclusion that the FNDMC non-additivity reveals the truth. This is because the Watson-Crick base-pair involves the charge transfer caused by the H-bonds. First of all, the evaluation result in dispersion interaction by the standard SCF methods was proved failed due to the lack of dispersion term. And the dispersion correction works well sometimes in the interaction itself, but not in the non-additivity. Second, even the coupled-cluster with singles, doubles, and perturbative triples (CCSD(T)) method still evaluates the non-additive contribution of dispersion interaction as the SCF-level, which will never happened in the binding itself, because of the practical handling of complete basis set correction (CBS) at the feasible level with second order M⊘ller-Plesset perturbation theory (MP2). Finally, although the wavefunction evaluation is trustworthy for the DMC method itself, the FN method cannot be ruled out to cancel the approximation error when the sign problem occurs, because we still divide the system according to the H-bond in the non-additive evaluation.

Keywords: B-DNA, Stacking energy, Non-additivity, Quantum Monte Carlo, *ab initio* methods

Acknowledgments

I would like to thank my supervisor first, Prof. Ryo Maezono and Assoc Prof. Kenta Hongo. Their careful guidance has made me have a very meaningful time in Japan and completed this Ph.D. thesis. Before coming to Japan, I was a person who had no concept of electronic state computing. They led me into this world of computational physics with challenges and opportunities. The two professors are not only experts in quantum chemistry, but also have extensive experience in high-performance computing. Which allowed me to continually improve my computer skills and physics theory knowledge during the years in JAIST. Prof. Ryo Maezono is a very conscientious professor with rigorous attitude towards science and technology, which admires me and guides me to progress.

I would like to thank Dr. Anouar Benali and Dr. Ye Luo for their unpaid support as well, so that I can go further and further in QMC research. Thanks for their help, I have overcome many difficulties in my research. And I give same appreciate for research assistants and the other students in our group. It is they who let our lab fill the atmosphere of learning and let me continue to make progress in constant discussions. I want to thank my parents for their supporting so that I can study at overseas far from home. I am very sorry because I have not been able to accompany them often. So I can only use this Ph.D. thesis as a gift to reward their kindness.

Finally, I want to thank all my friends at JAIST for letting me give me encouragement and comfort in the most difficult times. I have experienced happiness and sorrow at JAIST and have grown considerably in research and life.

Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the continuous development of science, our exploration of micro is getting deeper and deeper. However, nature shows us not simple truths, but the unknown challenges that are constantly emerging. The computer simulation establishes a bridge between the theoretical model and the real world. Then an accurate description of the physical theory is essential. In this chapter, the limitations of traditional computer molecular simulations are analyzed in the background of the discovery of biomolecules. Starting from this background, the research motivation of this thesis is brought forward. The problems solved in this study were defined, and the main findings and contributions of this study were briefly explained. The structure of this thesis is explained at the end of this chapter.

1.1 Background

The discovery of DNA double helix [7] and the determination of its 3D structure have opened the door to life science at the molecular level. Since then, the study of the properties of biological macromolecules has become an important topic in academic research. From the DNA molecules to the transcription of proteins, the microscopic effects of these magical molecules are the "central rule" and key points to uncover the mysteries of life. Life relies on DNA molecules to store genetic information and relies on the basic substances of replicating DNA molecules for the transmission of genetic information. The information

hidden in DNA is transcribed into mRNA by the action of transcriptase, and the corresponding amino acid is transported to the ribosome by tRNA according to the codon-anticodon pairing principle. The basic units of these proteins are arranged in a sequence following the encoding of mRNA, and further proteins are formed. Protein is the basis of all life activities and can be defined as materials with a certain spatial structure formed by the folding of polypeptide chains composed of amino acids in an "dehydration condensation" manner. The complex and subtle geometry of a protein is the structural basis for its realization of different biological functions. Protein has a four-level structure from micro to macro [8,9]:

1. The primary structure refers to the order in which amino acids are arranged on a protein polypeptide chain.
2. A secondary structure refers to a specific spatial structure in which a polypeptide chain is crimped or folded under the action of intermolecular forces.
3. Based on the secondary structure, the polypeptide chain will further form a more complex tertiary structure according to the specific spatial structure.
4. Finally, the quaternary structure of the aggregate is formed in a certain spatial arrangement.

The spatial structure of proteins is an important guarantee for their biological functions. If the protein molecule is in its specific three-dimensional structure, only specific biological activities can be obtained. There is a great significance for drug design and exploration of biochemical principles to study the spatial structure changes of proteins, their functions, and the relationship between protein and ligand action mechanism. In one word, in molecular systems, the Dynamics of structure is determined by intermolecular interactions. Suffice it to say that intermolecular forces explain the stability of important compounds such as DNA and RNA, and also play a vital role in biological behavior such as muscle contraction. [10]

The current main research methods on biomolecules are experiments, including X-ray diffraction (XRD) and nuclear magnetic resonance (NMR). As with other materials science studies, experimental methods are difficult to describe the microscopic details of biomolecules, including molecular-level dynamics and

basic functional states. Especially in the study of the dynamic mechanism of proteins, the folding mechanism is particularly difficult. During protein folding, there are many forces involved, including some structural steric hindrance, van der Waals forces (vdW), hydrogen bonding interactions (H-bond), hydrophobic interactions, ionic interactions, and ruthenium-driven folding resulting from the interaction of the polypeptide and surrounding solvents. For drug development, such as cancer-targeted drugs, precise matching of biological targets is required. Generally, screening of target compounds requires high-throughput experiments, but this is costly. Therefore, computer simulation plays an important role in understanding and studying the structure and biological function of biomolecules compared to traditional experimental methods.

Since 1977, McCammon [11] published the first article on protein molecular dynamics simulations in Nature, opening the era of biomolecular modeling. With the increasing computing power, the rise of large-scale parallel computers [12], and the emerging computing methods, simulation has become one of the important research methods of biomolecules, not just an auxiliary method to explain experiments. With the combination of quantum chemical calculation methods and molecular modeling in recent years, it has become possible to study complex and special biological molecules.

Computer simulation of molecular or molecular systems, also known as molecular modeling, is based on understanding the relationship between molecular structures and properties, the establishment of mathematical models, the inductive laws, and the predictive properties. Molecular simulation can not only study the structures and properties of known molecules but also simulate the structural properties of unknown or unsynthesized compounds. It also could be applied to the study of the structural properties, and guiding the design of necessary molecules with specific structural properties.

Molecular simulations can be regarded as experiments on the computer. Through computer numerical simulation, the information about molecular could be obtained from atomic-level interactions, including their structure, kinetics, thermodynamics, and the relationship between this physical property information and molecular function. Molecular simulation can be used to detect the prop-

erties that are difficult to measure directly by experimental methods and to fill in microscopic details during the experiment. Calculation helps us understand experimental phenomena, provide information that is not available in experiments, and even predict experimental results.

Due to the large system size response of proteins and nucleic acids, current experimental methods often fail to achieve the required accuracy. Computational software can simulate complex, large-scale reaction systems, visually describe complex biochemical processes, which there are great advantages in the analysis process. These advantages are expressed as:

1. Computer simulation method can simulate the actual system through a series of complex systems, thus providing a reference standard that can be used to compare approximation theory;

2. Computer simulation can compare simulated models with experiments and provide a means to assess whether the models are correct;

3. Computer simulation can also strengthen the combination of theory and experiment. Many physical properties are experimentally impossible or difficult to obtain accurate values but can be calculated by simulation.

Molecules are composed of individual atoms, each of which is in an energy state, and what kind of force is received, and formulating these is the so-called molecular force field. As early as 1970, there were already calculation methods based on classical mechanics, namely, force field molecular dynamics. [13] The method calculates various properties of molecules based on the force field of molecules. According to Born-Oppenheimer approximation, the movement of electrons in the calculation is neglected and considered as a force field. The potential energy of the system is considered as a force function of the position of the nucleus. The parameters in the force field of the molecule are obtained by quantum mechanical methods or by experimental methods, that so-called empirical field method. [14]

1.2 Empirical force field

If we introduce quantum mechanics, considering each quantized electron motion, the calculation will be much more complicated, due to the calculations of electron-electron, electron-nucleus interactions and electron kinetic energy will make the computational complexity increase exponentially(If the number of electrons is N, $O\left(N^{5\sim6}\right)$). Compared with the quantum mechanical method, the empirical field method is much simpler in the calculation and can quickly obtain various properties of the molecule. In some cases [15–17], the results are almost identical to those obtained by high-order quantum mechanical methods, but the calculation time is much shorter than the calculation of quantum mechanics. So with the right precision, the advantages of molecular dynamics are obvious:

1. Calculation speed is fast;
2. A system capable of calculating large molecules or containing many molecules;
3. No high-performance computing resources are required.

On the other hand, the molecular force field function is given by the empirical parameter, which causes the molecular force field to be accurate only for the molecules of the structure. Then the force field is applied to other functions, and the force field function cannot be completely Trustworthy. If the chemical bonds in the molecule change, the electron distribution within the molecule will change accordingly, and the energy and structure calculated using such molecular force fields are obviously not correct. Moreover, since the motion of electrons is neglected, the properties related to electron density or molecular orbital cannot be calculated.

However, the force field model constructed by empirical data limitate the simulation results. Because the empirical field does not directly describe the electrons, and different interactions usually form the force field in the real situation, the empirical parameters can only describe the force field in the current state of a target system. When the condition changes, the chemical bond changes due to the change of the state of electrons, and this change are not self-consistent with the geometry. Correspondingly, the excited state and transition state of

electrons affect the accuracy of the force field. The core problem is that the empirical force field is an approximate model from different experimental data, which is also the reason why the simulation of the force field is fast (which could calculate larger systems than electronic state calculations). On the other hand, empirical data is not credible, and any empirical field simulation results require experiments to verify. Namely, the simulations can now be performed on systems containing millions of atoms and in microseconds. Another increasingly popular method is to combine the quantum mechanical potential energy calculation with the molecular force field. Although this semi-empirical hybrid method utilizes quantum mechanical potential energy to solve most of the defects of classical potential energy accurately, these difficulties still exist in the force field part of the calculation. Therefore, many researchers have proposed corresponding improvement methods, such as X-POL force field [18].

Since the molecular force field method, more force fields have been applied to biochemical, molecular systems, polymer compounds, metals, and non-metallic materials. These methods have greatly improved the accuracy of complex computational systems, as well as their thermodynamics and Physical properties. However, particles do not just contain potential energy. When we need to look at the evolution of the system over time from a dynamic perspective, kinetic energy needs to be taken into account. Molecular dynamics simulation (MD) is a computational method developed using these force fields or electron orbit to calculate the potential energy and calculate the ion kinetic energy based on classical motion mechanics. Particle motion in molecular dynamics simulation has the correct physical basis. Therefore, the accuracy is higher, the dynamic and thermodynamic statistics of the system can be obtained, and it can be widely applied to various systems and various characteristics. However, molecular dynamics simulation has certain limitations. Since the calculation requires reference to the mathematical integration method, it is only possible to study the motion of the system in a short time range, and it is impossible to simulate long-term motion problems (such as protein folding).

In macromolecular systems, intermolecular forces within many-body system cannot be ignored. But most force field potential function based on pairwise

function, which means only consider the superposition interactions between two-body. Therefore empirical force field parameters cannot accurately describe the quantum fluctuations in many-body (more than two-body) molecular systems. Due to the interaction between molecules, in a many-body system, electrons will induce polarity with the force field, which will cause orbital deformation, resulting in weakening of the intermolecular interactions. Nevertheless, the interaction between molecules is significant for the properties and actions of molecules, such as protein folding and DNA stacking. Ignoring the electron movement in the empirical field also ignores this incentive, which makes the interpretation of the reaction mechanism at the microscopic level inaccurate.

In many-body systems, the empirical force field cannot consider the non-additivity of force field superposition. Unlike classical mechanics, the force field superposition of quantum systems cannot be added by the instantaneous polarization caused by quantum fluctuations. Therefore, it has a non-additive effect. The non-additivity in the interactions is expected in inter-molecular bindings due to the induced polarization by quantum fluctuations, such as vdW forces.

The description and reproduce of the bonding itself, due to intermolecular forces, is a huge challenge for *ab initio* methods. Non-additivity is a more difficult subject that has long been far from the mainstream research field and has not been well analyzed yet. At present, most molecular force fields are implemented by assuming a superposition of two physical forces, which are widely used in a large number of biomolecules for self-organization simulation [19–26]. This assumption is partially demonstrated, for example, in the report [1], confirming the good agreement between the empirical field, such as AMBER force field [27], and the stacking energy predicted by high-precision quantum chemistry methods. The *ab initio* quantum chemistry theory is a good description of the natural stacking energy, which allows reliable energy to be found on any base structure. Calculations, in any case, need to be done at a sufficient theoretical level. For example, standard DFT, HF, and semi-empirical methods all fail in the description of base stacking because they cannot correctly capture the dispersion effect [28].

Based on the above discussion, we need a high-precision method to simulate the molecular system. Of course, the high-precision method coupled-cluster with

singles, doubles, and perturbative triples (CCSD(T)/CBS) [29] is a widely used method called "gold standard" for quantum chemistry. CCSD(T) is a calculation method based on the HF method and adding electron correlations functions. However, the full electronic correlation calculation makes the calculation cost increase rapidly with the increase of the number of particles $O\left(N^7\right)$. It is impossible to calculates large systems due to its exponential growth computational cost. Recent advances in accurate calculation methods, especially through Diffusion Monte Carlo (DMC) calculations, make it possible to handle larger systems. [28, 30–41] However, some work applied to systems consisting of weakly constrained subsystems shows that non-additiveness is much larger than we expected. [42, 43] Although there is non-additivity in larger molecules, if the non-additive contribution is positive, then there is no research significance. If so, it will only make minor corrections to the C_6 (the coefficient of $1/R^6$deciation interactions) force without any qualitative impact.

1.3 Problem Statement

The subject studied in this thesis is the non-additivity of non-covalent interactions between large molecules such as B-DNA systems. Due to quantum fluctuations, the non-additiveness of the interaction forces between macromolecules is always expected. The interaction between the molecules of a living organism is in the formation of its structure. It plays an important role in biochemical reactions. DNA molecules are the basis of biological, genetic variation. The most basic ten kinds of B-DNA molecular structures are composed of two kinds of purines and pyrimidines. For DNA molecules, the number of atoms is around 10, and the weak force between many-body molecules itself is the challenge in the field of quantum chemistry. The non-additive study of the forces between molecules is at the edge of the research field. In the previous calculation methods, the sign of the non-additive contribution is positive. Recent studies have shown that in the calculation using the fixed node diffusion monte carlo (FNDMC) method, and negative values appear in the results of non-additive contributions. First of all, the appearance of this phenomenon makes the evaluation result of the

SCF methods in dispersion interaction non-additivity doubtful. Second, even the CCSD(T) method, known as the "Gold standard", still evaluates non-additivity as the SCF-level, which will never happen in the dispersion interaction itself. Finally, although the calculation is trustworthy for the DMC method itself, the Fixed-Node method cannot be ruled out in the cancel the approximation error occurs when the symbolic problem occurs, because, in the non-additive evaluation, we still divide the system according to the H-bonds. Therefore, discussing the correctness and rationality of this result is the problem to be solved in this study.

1.4 Contributions

In this study, we performed *ab initio* calculations for the selected many-body molecular systems, B-DNA stacking systems. We investigate non-additivity in B-DNA "systematically" for the first time. A variety of different methods were used to compare stacking energy and non-additivity contribution in different B-DNA combinations. These methods include the simplest HF method, the addition of electronically associated LMP2, CCSD(T), SAPT methods, and several common functional standard DFT methods (LDA, hybrid functionals(B3LYP) and Meta hybrid GGA (M06-2X)), as well as the DFT methods with dispersion correction (B3LYP-D3, CAM-B3LYP-D3, wB97X), and FNDMC, one of the widely used QMC methods. We will explain in detail the theoretical implications, differences, and computational processes of these methods in the theoretical chapters that follow. Detailed data and results, as well as a detailed discussion, will be presented later, summarizing the main findings of this study. Our evaluation of binding energies come up to *common expectations* for the methodologies. However, the expected result of binding energy does not occur equally in the non-additive contribution of binding energy:

1. FNDMC and "CCSD(T)" results are inconsistent with each other in the non-additive contribution (here we use "CCSD(T)" instead of CCSD(T) for some reason described below);
2. "CCSD(T)" gives almost the same results as B3LYP and HF at an SCF-level in the non-additive contribution, which hardly occurs when evaluating

binding energy;

3. In the inconsistent part of the non-additional contribution, we can find that the FNDMC has a significant negative non-additive contribution, and the London theory naturally expects the negative value of this part;

4. In the non-additive contribution of another part of the positive sign, the value of FNDMC is far higher than other methods, and these trends are strictly related to the structural asymmetry of the system. We found the law of this part after analyzing the Hydrogen bonding bridging arrangement (parallel or anti-parallel) in the system.

Only the FNDMC captures the negative values in non-additive contributions, which illustrates the significant advantage of this high-precision calculation method in the calculation of intermolecular interactions. For the quantum many-body system, the approximations based on the two-body superposition cannot achieve an accurate description of the non-additivity of dispersion forces at all, even if the dispersion force correction is increased. Then the reliability of the correction of the dispersion force is questionable, that is, whether the approximations of superposition can describe the induced polar dispersion force with sufficient accuracy. The conclusion that readers most easily misunderstand this study is that the non-additive contribution of CCSD(T) only stays at an SCF-level. These conclusions are not due to our careless choice of calculation specifications or statistical calculation errors. However, more fundamental points – practical approximations used in DMC and CCSD(T), are very effective for evaluating binding energy, but not for the non-additivity contribution results. In the former case, we can see that many previous works are reporting the cancellation of fixed node biases between the entire system and its constituent molecules, which can be well evaluated for binding energy [35–37]. On the other hand, such cancellation has not been investigated yet. In the latter case, we note the fact that "CCSD(T)" applied to B-DNA systems is actually "CCSD(T) with CBS at the MP2 level" [1]. We conclude that this practical approximation can be attributed to the fact that there are the same trends in non-additivity as B3LYP that is not believed to be capable of reproducing vdW interactions [36].

1.5 Outline of thesis

This thesis contains six chapters, where Chapter 1 gives an introduction to the research. The introduction mainly describes the current situation of molecular simulation, as well as the shortcomings of traditional methods in high-precision simulation of large molecular systems, and explains the motivation of the thesis. Moreover, the structure of the article and the main contributions are also described.

Chapter 2 combs the theoretical framework of the method of *ab initio* methods calculation. This chapter starts with the most basic Hartree-Fock (HF) method and goes to Post-HF methods. The DFT methods, which are widely used, and the QMC methods applied in this research, are briefly described. Different methods were compared, including the theoretical and algorithmic processes are provided.

Chapter 3 gives the problem we solved and described the target system. B-DNA systems are typcal samples for non-additivity studies. In order to compare the differences in precision between different methods, we have tried from the widely used DFT methods to the high-precision methods. The details of how to perform the calculation are described. Finally, we also analyze the parallel efficiency of two different algorithms (SCF and MC).

Chapter 4 presents the results of the non-additive nature of stack energy and discusses the results. The categories of molecular action and the mechanism of superposition and non-additive action are elaborated in detail. This study applied the QMC method in a macromolecular 4-body system and a monoatomic molecular four-body system.

Among them, the non-addition result of QMC is found to be much larger than other traditional methods, which is discussed by two simplified models in Chapter 5.

Finally, the conclusion and the recommendations for future studies are discussed in Chapter 6.

The main content of thesis have already been published in Chemical Physics in 2019 as “Inconsistencies in *ab initio* evaluations of non-additive contribu-

tions of DNA stacking energies" [44]. And it also been published in arXiv as https://arxiv.org/abs/1807.04168. Part of the results were presented in the following conferences:

1. APS March Meeting 2019, R31, Boston, USA, Mar/07/2019;
2. QMCPACK users workshop 2019, Oak Ridge, USA, May/14/2019;
3. Workshop on Crystal Structure Prediction: Exploring the Mendeleev Table as a Palette to Design New Materials — (smr 3267), Trieste, Italy, Jan/14/2019;
4. APS March Meeting 2018, A34, Los Angeles, USA, Mar/05/2018.

1.6 Summary

This chapter sets out the current state of molecular simulation from the development of biomolecules as background. It also illustrates the traditional empirical field defects and further gives the importance of high-precision quantum chemical simulation. With the rapid development of the computer industry, especially the computational power of massive-scale parallel computers, molecular simulation has shown a very bright prospect in the fields of revealing biological effects and computer-assisted drug design. However, it is difficult to achieve a balance between high-precision calculations and reasonable computational costs. However, the precise description of the intermolecular interactions is an essential part of molecular modeling. The research demand of larger systems is also a challenge in the conventional computing methods. In recent years, the QMC method based on numerical statistics has achieved good results in the calculation of long-range intermolecular forces, which demonstrates the dawn of high-precision quantum chemical calculations. Although the QMC method is far more costly in computational complexity than conventional *ab initio* methods. Fortunately, parallel computers are better at solving statistics-based calculations such as QMC than the widely used methods based on determinant numerical integration. In this study, the QMC method is used to study the non-additivity of intermolecular interactions. The results show that the QMC method can capture non-additive contribution more than expected which is difficult to capture by other *ab initio*

calculations. This important discovery demonstrates the unparalleled potential of the QMC methods in high-precision large-scale biomolecular simulations. It raises reasonable questions about the approximations methods based on the two-body superposition in the traditional methods, which provides a valuable reference for later research. Finally, this chapter concludes the main contributions and the chapter arrangement structure of this thesis.

Chapter 2

Theoretical framework

ab initio calculation is derived from Latin, meaning "from the beginning", is a description of physicochemical properties of molecules, periodic crystals, etc. using microscopic quantum mechanics theory, which is an important methodology to solve many-electron state problems currently. This chapter begins with the simplest Hartree-Fock (HF) method for solving the Schrödinger equation and expands to the Post-HF methods with the electronic association. The theoretical methods and computational details of the density functional theory (DFT) methods, which are currently widely used, are briefly described in the following sections. However, for these methods, there are some limitations. For large system molecular systems, the Post-HF methods with the addition of electronic correlation are too complex to handle large many-body systems. For DFT methods, the lack of dispersion interactions term makes them seem to be unable to deal with the role of weak intermolecular interactions. Therefore, high-precision methods that can handle large systems are indispensable. Compared with the usage of approximation to improve the calculation speed, it is an important idea to use randomly statistical methods to solve energy integrals. Finally, the main algorithms of the quantum monte carlo (QMC) methods mainly applied in this thesis, including variational monte carlo (VMC) and diffusion monte carlo (DMC), is explained in detail.

2.1 Mean field approximation

2.1.1 Hamiltonian and N-body Schrödinger equation

In condensed matter physics, both atoms and molecules, as well as periodic crystals, consist of elementary particles, electrons, and nucleuses. The description of these particles is based on fundamental theoretical methods in terms of quantum mechanics and statistical mechanics, expressed by the wave function Ψ, and $|\psi|^2$ is the probability of the particle.. For a variety of properties on a system that contain ions and electrons, the relationship can be expressed in terms of Hamiltonian. With this Hamiltonian, all properties of systems can be obtained by solving the time-independent Schrödinger equation in principle. Suppose a system contains n particles, for time-independent N-body Schrödinger equation can be presented as follow:

$$\hat{H}\ \Psi = E \cdot \Psi\left(\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_n\right)\ , \tag{2.1}$$

where $\hat{H}$ is the Hamiltonian operator to obtain properties and E is the eigenstates of $\hat{H}$ and Ψ is the wave function of the particles in position $\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_n$.

In general, not for any value of E, there is a non-zero solution that satisfies the natural condition. In order for such a solution to existing, E can only take certain values. The solution of E is an eigenvalue of Hamiltonian $\hat{H}$, which is a real number that satisfies the eigenvalue equation. If we consider further like particles in a box or harmonic oscillator, there are N electrons and M nuclei in this system, and the electrons in the space vector position $\vec{r}$ use $i, j, \cdots$ to represent. The atomic nucleus is represented by $A, B, \cdots$, and the distance between them is the vector difference of their position. The mass of electrons m is much smaller than the mass of the nucleus M, and Z_A is the atomic number of nucleus. Then the complete Schrödinger description is as follow:

$$\left[-\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 - \frac{\hbar^2}{2M_A}\sum_{A=1}^{M}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A e^2}{r_{iA}} + \sum_{i=1}^{N}\sum_{i<j}^{N}\frac{e^2}{r_{ij}} + \sum_{A=1}^{M}\sum_{A<B}^{M}\frac{Z_A Z_B e^2}{r_{AB}}\right]\Psi = E\cdot\Psi, \tag{2.2}$$

where the first term of Hamiltonian in left is the set of kinetic energy of elec-

trons; the second part represents the kinetic energy of the nuclei, and ∇^2 is the Laplacian operator representing the differential of spatial coordinates; the third term represents the interactions between atomic nucleus A and electron i, and r_{iA} represents the distance between electron and nucleus, $r_{iA} = \left|\vec{r}_i - \vec{r}_A\right|$, this item can be represented by the V function; the fourth term electron and electron correlation, r_{ij} represents the distance between electrons, $r_{ij} = \left|\vec{r}_i - \vec{r}_j\right|$, also expressed as $U(\vec{r}_i, \vec{r}_j)$; the last item is the correlation between different nuclei.

It should be noted that, based on a conservative system, in the case of $\hat{H}$ for total energy, $V(\vec{r}_i)$ is only a function related to spatial coordinates. $\Psi(\vec{r}_1, \cdots, \vec{r}_n)$is the electronic wave function of the n particles on each position coordinate. If each particle is in the ground state, then the eigensolution E of the Hamiltonian is the total system ground state energy under this state. It is unrealistic to solve this equation completely for a large enough system; therefore, how to get a solution within the acceptable accuracy range is being discussed in the next sections. The first is a reasonable approximation, omitting items of a small order of magnitude. The second is to use the idea of division to reduce the time complexity. The electron wave function is the equation on each coordinate, which belongs to all N electrons. The reasonable idea is to map this many-body interacting problem to a set of one-body noninteracting problem (Kohn-Sham equations). This will be explained and described below in detail.

2.1.2 Born-Oppenheimer approximation

The mass of the nucleus is $10^3 \sim 10^5$ times larger than the electron, therefore in the molecular system, the displacement of nuclei is much smaller compare to electrons at a certain time scale. This means that electrons always move around nuclei, and the electrons have an appropriate state of motion under the arrangement of any determined nucleus. In turn, the relative motion between nuclei can also be seen as the average effect of the electronic movement. Hence, to a good approximation, the electrons movement in molecules can be considered as moving in the potential field around fixed nuclei. This simplification of the motion of many-body systems is called the Born-Oppenheimer (BO) approximation. Without considering the interaction between spin and orbital, the interaction between spin and spin, and the

case of relativity and so on. According to the discussion in the previous section, the system's Hamiltonian $\hat{H}$ can be expressed as five items: the kinetic energy of electrons, the kinetic energy of nuclei, the attraction energy of electrons and nuclei, the repulsive energy between electrons, and the repulsive energy between nuclei. In order to separate the electronic motion from the nuclear motion, it is necessary to make an approximation after omitting terms of a small order of magnitude, and to separate the wave function Ψ as $\Psi(\vec{r}_i, \vec{r}_A) = v(\vec{r}_A) \cdot u(\vec{r}_i, \vec{r}_A)$, where u is the wave function of electrons statement and v is wave function of nuclei. The approximated formula after separation is obtained as follow:

$$\begin{aligned} \left[-\frac{\hbar^2}{2m} \sum_{i=1}^{N} \nabla_i^2 + V(\vec{r}_i, \vec{r}_A) \right] u &= E(\vec{r}_A) \cdot u(\vec{r}_i, \vec{r}_A) \\ \left[-\frac{\hbar^2}{2M} \sum_{A=1}^{M} \nabla_A^2 + E(\vec{r}_A) \right] v &= \varepsilon \cdot v(\vec{r}_A). \end{aligned} \tag{2.3}$$

The upper formula is the Schrödinger equation for electrons movement, the lower formula is for nucleus movement. The full version of the formula takes into account the state of thermal motion, including electronic, vibrational, rotational, and translational energy. And when we isolate the movement of the nucleus, we will no longer consider the vibration rotation problem, focusing only on the electronic problem. Therefore, we remove the variables caused by the position of the nucleus, considering only the electronic Hamiltonian and the electron wave function.

2.1.3 Hartree-Fock theory

The Hartree-Fock (HF) method was proposed in the 1930s to calculate the atomic structure and then gradually used to calculate the molecular structure [45], which is the basis of molecular orbital (MO) theory. HF theory is one the simplest approximate theories for solving the many-body Hamiltonian. According BO approximation discussed in the previous section, the Hamiltonian could be given

as follow:

$$\hat{H} = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A e^2}{r_{iA}} + \sum_{i=1}^{N}\sum_{i<j}^{N}\frac{e^2}{r_{ij}}. \tag{2.4}$$

Same as the previous formula, except that this time, we omitted the kinetic energy of the nuclei. Because BO is used and the nucleus coordinates are constant, the repulsive force between the nuclei is constant, so the solution of the wave function of the original equation is not affected, and the difference in the feature value is constant. The electron kinetic energy (the first term) and the potential energy of the electrons attracted by the nuclei (the second term) can be decomposed into the sum of the single-electron Hamiltonian operators. However, the interaction between electrons cannot be ignored. Hartree proposes an approximation method that considers the interaction between electrons. If this repulsion is averaged over all positions of one of the two electrons, the result will be only a function of another electronic coordinate. Thus, the entire system can be decomposed into a single-electron stationary Schrödinger equation:

$$\hat{h}_i\,\psi_i = \varepsilon_i \cdot \psi_i. \tag{2.5}$$

Expand into a form containing non-local potential U and local ionic potential V:

$$-\frac{1}{2}\nabla^2\psi_i(\vec{r}) + V(\vec{r})\psi_i(\vec{r}) + U(\vec{r})\psi_i(\vec{r}) = \varepsilon_i \cdot \psi_i(\vec{r}). \tag{2.6}$$

The many-electron system is the same Fermi sub-system, and the stationary wave functions the Slater should be written as the Slater determinant as follow in order to satisfy the antisymmetric:

$$\Psi(q_1, q_2, \cdots, q_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(q_1), & \psi_1(q_2), & \cdots & ,\psi_1(q_n) \\ \psi_2(q_1), & \psi_2(q_2), & \cdots & ,\psi_2(q_n) \\ \vdots & \vdots & & \vdots \\ \psi_n(q_1), & \psi_n(q_2), & \cdots & ,\psi_n(q_n) \end{vmatrix}, \tag{2.7}$$

where the variables q include the coordinates of space $\vec{r}$ and spin. Since $\hat{h}_i$ only acts on the corresponding wave function, the orthogonal normality of the single

electron wave function ψ_i can be utilized. According to the variational principle, after averaging the average of the Hamiltonian operators, it can be found that the definite integral is independent of the integral variable so that the Fock energy formula can be finally obtained:

$$\begin{aligned} E = & \sum_k \int \left[\psi_k^*(q_i) \left(-\frac{\hbar^2}{2m} \nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A e^2}{r_{iA}} \right) \psi_k(q_i) \right] dq_i \\ & + \frac{\hbar^2}{2m} \sum_{kk'} \int \left[|\psi_k(q_i)|^2 \frac{e^2}{r_{ij}} \left| \psi_{k'}(q_j) \right|^2 \right] dq_i dq_j \\ & - \frac{\hbar^2}{2m} \sum_{kk'} \int \left[\psi_k^*(q_i) \psi_{k'}^*(q_j) \frac{e^2}{r_{ij}} \psi_{k'}(q_i) \psi_k(q_j) \right] dq_i dq_j. \end{aligned} \tag{2.8}$$

The first summation represents the total kinetic energy of attraction between all electrons and electron-nucleus. The second term represents the electrostatic repulsion energy of these two electrons, according to Coulomb's law. Alternatively, called Hartree term, which is the simply electrostatic potential, including an unphysical self-interaction of electrons when $j = i$. The third term is the exchange term, which always appears as a negative sign, which can reduce the interaction energy between parallel spintronics in different orbitals and stabilize the system, also called spin-parallel term.

However, the specific form of $\hat{h}_i$ contains integral terms for electronic wave functions:

$$\sum_j \int \frac{\psi_j^2 e^2 dr_j}{r_{ij}}. \tag{2.9}$$

In other words, it should be included an electronic wave function, and the solution to this problem is as follows: Starting with N wave function ψ_i, the zero-order approximation function is represented by $\psi_i^{(0)}$, and these functions are used to establish the corresponding $\hat{h}_i^{(0)}$. By solving a single electronic equation, a new set of wave functions is obtained, and so on repeat, until the last cycle loop, the obtained wave function can be equal or infinitely close to the exact wave function we need. Such a process is called a self-consistent field (SCF).

2.2 Describing electronic correlation

The HF method considers spin correlation but does not consider the Coulomb correlation. Due to Coulomb repulsion between electrons, two electrons cannot appear in the same position in space, and the probability of being close is also small. This dynamic correlation energy also needs to be corrected in the calculation.

2.2.1 Configuration interaction

Configuration interaction (CI) [46,47] uses the lowest energy n molecular orbitals in the system to form a Slater determinant for describing the ground state Ψ_0:

$$\Psi_0 = |\Psi_1 \Psi_2 \cdots \Psi_e \cdots \Psi_n| , \tag{2.10}$$

where Ψ_e is the lectronic excitation orbitals. These wave functions are all single electronic wave functions that are at occupied orbitals.

If there is an electronic excitation to an empty orbital, each excitation corresponds mapped to a certain Slater determinant, which is referred to as an activated configuration. If only one electron is excited from the occupied orbital to an empty orbital, and $|S\rangle$ is configured for single-electron excitation. Two electrons are excited from the occupied orbitals to the empty orbitals, and $|D\rangle$ is configured for the double-electron excitation. Three electrons are excited from the occupied orbitals to the empty orbitals, and $|T\rangle$ is configured for the triple-electron excitation. Express the exact wave function of the system as a linear combination of Slater determinant wave functions:

$$|\Psi_E\rangle = C_0 |\Psi_0\rangle + C_S |S\rangle + C_D |D\rangle + C_T |T\rangle + \cdots . \tag{2.11}$$

Starting from a complete set of single electron wave functions $\{\psi_i\}$, a complete set of Slater determinant wave function $\{\Psi_i\}$ sets is constructed. The multi-electron wave function Ψ_E of the system is expanded to the complete determinant wave function. In principle, the exact solution of the Schrödinger equation can be

obtained by this way, ie. full CI method.

Where $\{\psi_i\}$ is the space of orbitals, and $\{\Psi_i\}$ is the space of configurations. The CI method calculates the excited states so that each excited state configuration contains many empty orbitals in the original ground state. Thus the probability of distribution of other electrons around each electron becomes smaller as the excitation configuration increases. Thereby the correlation effect becomes smaller and smaller, so the Coulomb correlation can be corrected.

2.2.2 Coupled-cluster method

Coupled cluster (CC) method was first applied in the field of nuclear physics, and Cizek *et al.* [48] used it for the electronic structure calculation of molecules. The most basic equation is:

$$\Psi_E = e^T \Psi_{HF}, \tag{2.12}$$

where Ψ_E is the non-relativistic exact wave function for multi-electron systems; Ψ_{HF} is the HF ground state wave function of the system, as the reference state of CC method; T is a cluster operator, which is actually the total excited state generation operator, which can be expressed as the sum of the generator states T_i for each excited state:

$$T = T_1 + T_2 + T_3 + \cdots = \sum_i T_i. \tag{2.13}$$

The general n-order cluster operator could be shown as :

$$T_n = \frac{1}{(n!)^2} \sum_{i_1,i_2,\ldots,i_n}^{Occupied} \sum_{a_1,a_2,\ldots,a_n}^{Unoccupied} t_{a_1,a_2,\ldots,a_n}^{i_1,i_2,\ldots,i_n} T^{a_1} T^{a_2} \ldots T^{a_n} T_{i_n} \ldots T_{i_2} T_{i_1}, \tag{2.14}$$

where i stands for occupied and a for unoccupied orbitals, T means that the excited electrons from the occupied spin orbitals to the unoccupied spin orbitals, $t_{a_1,a_2,\ldots,a_n}^{i_1,i_2}$ means the corresponding expansion factor is called the cluster amplitude. Solving for the unknown coefficients, cluster amplide, is necessary for finding the approximate solution $|\Psi\rangle$.

The exponential operator may be expanded as a Taylor series. After expanding e^T into a series, we can get the wave function expression obtained by combining the terms of the excited electrons:

$$\Psi_E = \left[1 + \boldsymbol{T}_1 + \left(\frac{1}{2!}\boldsymbol{T}_1^2 + \boldsymbol{T}_2\right) + \left(\frac{1}{3!}\boldsymbol{T}_1^3 + \boldsymbol{T}_1\boldsymbol{T}_2 + \boldsymbol{T}_3\right) + \cdots\right]\Psi_{HF}. \tag{2.15}$$

It can be seen that the common CI method is an approximation of the CC method. For example, in the case of two-electron excitation, the CC method includes $\frac{1}{2}T_1^2$, which is not in the CI method. For multi-electron correlation, it can be divided into connected clusters; that is, electrons are directly related, and disconnected clusters are disconnected, meaning that the clusters of electrons are related in different areas of space. T_n represents n-electron correlation and can be considered as a connected cluster, while the other items T_n^m are unconnected clusters, representing the correlation of electrons in different regions of space. The probability that all electrons will get together at the same time is tiny, but the probability that electrons are made into different sets and then correlated with each other is much greater.

According to the number of items of T, the level of the CC method is expressed as follows:

$$\begin{aligned} &\text{CCD} : T = T_2 \\ &\text{CCSD} : T = T_1 + T_2 \\ &\text{CCSDT} : T = T_1 + T_2 + T_3. \end{aligned} \tag{2.16}$$

CCSD(T) indicates that the cluster amplitude t_{ijk}^{abc} of T_{ijk}^{abc} is not calculated in the iteration, and the coefficients of single excitation and double excitation are used to calculate the triple-electrons term. It is shown that the calculation of the three electrons correlation is an approximation.

2.2.3 Many-body perturbation theory

It is assumed that the Hamilton of the system can be decomposed into a part E before the perturbation and a part W that is subjected to the perturbation.

$$\begin{cases} \hat{H} = E + W \\ |H'\rangle = |0\rangle + |1\rangle + |2\rangle + \cdots \\ H' = E' + a_1 + a_2 + \cdots, \end{cases} \tag{2.17}$$

where the upper formula represents the eigenvector, and the following represents the corresponding eigenvalue. And $|0\rangle, E'$ represents a zero-order approximation, and the following items are first-order corrections, second-order corrections, and so on. This formula can be transformed into a formula that is solved step by step:

$$\begin{aligned} (E' - E)\,|0\rangle &= 0 \\ (E' - E)\,|1\rangle + a_1|0\rangle &= W|0\rangle \\ (E' - E)\,|2\rangle + a_1|1\rangle + a_2|0\rangle &= W|1\rangle. \end{aligned} \tag{2.18}$$

These equations can be solved by a stepwise approximation, that is, the first-order approximation equation is solved, and $|0\rangle, E'$ is obtained, and then it is used to solve the approximation equation of the next stage. In principle, it can be solved indefinitely, but in practice, if W is good selected, the second-order is enough.

M⊘ller-Plesset Perturbation (MP) [49] is a type of many-body perturbation method. When using this method to solve the correlation energy, the HF method is first used as the unperturbed reference of the system. In this case, the sum of the single-electron Fock operators F is taken as the zero-order perturbation Hamiltonian:

$$F_{\psi_r} = \varepsilon_r \psi_r, \tag{2.19}$$

where ψ_r is a single-electron orbital wave function, and ε_r is the energy. The energy eigenvalue at this time, ie the energy of the unperturbed system, is equal

to the sum of the orbital energy:

$$E' = \sum_{r}^{Occupied} \varepsilon_r. \tag{2.20}$$

In this way we can use the single-electron wave function to construct the determinant wave function of the system, Ψ_0. After considering the first-level correction, you can get the energy as:

$$E_1 = E' + a_1 = \langle \Psi_0 | E | \Psi_0 \rangle + \langle \Psi_0 | W_1 | \Psi_0 \rangle . \tag{2.21}$$

It can be noted that this is the total energy in the HF method. Let us consider using this structure to construct a first-order correction of the wave function:

$$\Psi_1 = \sum_{i} \frac{\langle \Psi_{0,i} | \boldsymbol{W}_1 | \Psi_0 \rangle}{E' - a_{1,i}} \Psi_{0,i}. \tag{2.22}$$

By analogy, we can get higher order corrections step by step. The energy sought after this secondary-order correction is called MP2. Obviously, the larger the n of MPn, the more accurate the calculation, but the workload will also increase. MP2 is the most widely used level. In general, it can calculate the correlation energy of 80-90%, which is obviously not accurate enough for the weak interactions.

local electron correlation methods at the second-order perturbation theory level (LMP2) [50] can calculate the intermolecular interactions at the level close to MP2 results without the basis set superposition error (BSSE). That is to say, LMP2 can handle larger systems or use a higher level of the base set due to its computational cost varies linearly with the size of the target system [51]. Another method that has been shown to handle intermolecular forces efficiently is symmetry-adapted perturbation-theory (SAPT) method with energy components up to the second-order in V [52,53]. SAPT is also a method based on many-body perturbation theory to solve non-covalent bonds between molecules directly, and its calculation accuracy increases with the increase of level. These methods are all applied to the superposition energy calculation of the B-DNA system.

2.3 Density functional theory

According to the HF method, the energy E is regarded as an integral function Eq. 2.8 of the wave function ψ, and then the energy E is found to be extreme value according to the variational principle.

$$E = \frac{\int \psi^* \hat{H} \psi \mathrm{d}\tau}{\int \psi^* \psi \mathrm{d}\tau} \tag{2.23}$$

Since ψ itself is a function, E is a function of the function, called a functional, which is a process of finding the extremum of a functional. Density functional theory (DFT) [54, 55], by means of statistical averaging, represents the electron density near a spatial point, using electron density $\rho(\vec{r})$ to represent energy $E(\rho)$. The advantage of the energy density functional $E(\rho)$ instead of $E(\psi)$ is that it greatly reduces the computational complexity. For a multi-electronic system of N electrons, an electron has four coordinates, spatial coordinates, and spin coordinates. The wave function is considered to be able to describe any state of the particle in quantum mechanics, so a system's wave function requires $4N$ coordinates as independent variables. Considering the antisymmetry, the wave function formed as Slater determinant will become very large. If we consider configuring the interaction (CI), multiple Slater determinants will be used. The electron density is only a function of the spatial coordinates, and the magnitude of the variables can be reduced from $4N$ to 3. Besides, the wave function is observable, and the electron density is a considerable measure and can be determined experimentally. This is the basic idea of DFT, and further Hoheberg-Kohn theory lays the theoretical foundation of DFT. The Kohn-Sham equation gives a concrete form of the ground state properties using electron density.

2.3.1 Hohenberg-Kohn theorem

In 1964, Hoheberg and Kohn published a milestone article in the "Physical Review" [56] and began to establish the theoretical foundation of DFT. Hoheberg-Kohn theory includes two principles:

1. The ground-state electron density of the system always mapped with the external potential field. Therefore the properties can be completely determined.
2. For any density function $\rho(\vec{r})$, if the condition $\rho(\vec{r}) \geq 0, \int \rho(\vec{r}) d\tau = N$ is satisfied where N is the number of electrons, then $E[\rho(\vec{r})] \geq E_0$, E_0 is the ground energy of the system.

The first principle of HK indicates that the ground state energy of the system is only a functional of electron density. The second principle states that the variational principle of $E[\rho(\vec{r})]$ is established, and the ground state of the system can be obtained by applying the variational principle. The ground state energy of the system can be obtained by applying the variational principle.

2.3.2 Kohn-Sham Equation

For the total energy of the system, the kinetic energy of the electron, the interaction energy between the electrons, and the energy of the electron in the potential field V can be expressed as a functional of the density function $\rho(\vec{r})$ as follow:

$$E(\rho) = T(\rho) + E_{ee}(\rho) + \int \rho(\vec{r}) V(\vec{r}) \mathrm{d}\vec{r}. \tag{2.24}$$

In order to solve the specific computational problems, Kohn and Sham proposed not to pursue the kinetic energy exact formula of the real system with the electron density ρ as the variable, but to introduce the noninteracting reference frame to establish the expression of the calculated kinetic energy. The use of a non-interactive reference frame means that the electrons are still Fermi, but the electron Coulomb force is not considered when calculating the kinetic energy. Therefore, the universal functional of the real system can be decomposed into the following form:

$$T(\rho) + E_{ee}(\rho) = T_s(\rho) + J(\rho) + E_{xc}(\rho), \tag{2.25}$$

where $T_s(\rho)$ is kinetic energy in a non-interactive reference frame; $J(\rho)$ is classical electron-electron coulomb exclusion energy calculated separately. It is worth noting that the kinetic energy of the real system is not the same as the kinetic energy in the reference frame, and the interaction energy of electrons is also

different from the Coulomb force. These gaps are defined as a functional $E_{xc}(\rho)$ that exchange-correlation (XC) energy. If a good enough electron orbital (KS orbital) was selected, then $E_{xc}(\rho)$ can be small enough to handle the approximation error. DFT is precise in principle, and the key question is how to find "good enough" $E_{xc}(\rho)$ and V.

Then we can get the one-electron equation in DFT, the Kohn-Sham equation [57]:

$$\left[-\frac{1}{2}\nabla^2 - \sum_{p=1}^{\Lambda}\frac{Z_p}{\left|\vec{r}_p - \vec{r}\right|} + \int\frac{\rho\left(\vec{r}'\right)}{\left|\vec{r} - \vec{r}'\right|}\mathrm{d}\vec{r}' + \frac{\partial E_{\mathrm{xc}}}{\partial\rho}\right]\phi_i = \varepsilon_i\phi_i, \tag{2.26}$$

where the first term in the Hamiltonian expression on the left is the electron kinetic energy, the second term represents the attraction potential of the atomic p to the electron, and the third term is the Coulomb potential, and the last term is the exchange-correlation potential energy. If the XC function is determined, then we can solve the SCF solution in the same way as the HF method. ϕ_i can also be written as a linear combination of the basis function set $\{\chi_k\}$:

$$\phi_i = \sum_k c_{ki}\chi_k, \tag{2.27}$$

here $\{\chi_k\}$ can choose either Slater type orbitals (STOs) or Gaussian type orbitals (GTOs). Both electronic orbital descriptions have advantages and disadvantages. The structure of the STO form is more in line with the true wave function description, exponentially decays over long distances, and conforms to Kato's conditions in short distances. However, the form of STO in factorization is more complicated since its calculation is not easy to simplify. Therefore, according to the calculation cost, the application of GTO with easier numerical processing is more extensive. The GTO conforms to the Gaussian Product Theorem, i.e., the product of two GTOs can be expressed as the sum of a finite number of Gaussian functions somewhere in the middle. This greatly simplifies the use of basis functions, which are typically 4-5 orders of magnitude faster than STO. Furthermore, some contracted basis functions, the two forms are mixed in order to balance accuracy and computational efficiency. Therefore, the DFT experiences

three different levels of approximation. [58] The first is the difference between theory and reality, using KS orbital to approximate the physical wave function. The second is a numerical approximation, which involves the choice of methods that come in the actual solution of the differential equation, the main aspect being the choice of the basis function. The third type is the introduction of a non-interactive reference frame, which makes the error all focused on constructing an unknown XC functional expression. How to choose this functional is discussed in the next section. It is worth noting that the choice of XC functionals is like a ladder, and so far, there is no accurate functional expression that can be applied to the general systems. The difference in the processing of different functionals in different problems chooses functionals to contain empirical components.

2.3.3 Exchange and Correlation Functionals

The choice of functionals is the main research problem in the DFT method, and finding the right functional is crucial to the calculation results. Specifically, Local functionals refers to methods for electron density approximation based on harmonious electron gas, including Thomas Fermi (TF) and Local density approximations (LDA). They keep the electrons too close together that the exchange's overall energy was overestimates. At the same time, the correlation energy is underestimated, because only the interaction of local electrons is considered without the long-range parts. The Semilocal or the gradient-expansion approximation (GEA) and generalized gradient approximation (GGA) methods all incorporate density gradient parameters into the density functional, which improve the limitation of LDA. Nonlocal functionals can also be divided into hybrids, orbital Functionals such as meta-GGAs and self-interaction correction (SIC), and Integral-dependent functionals. Among them, SVWN is a kind of LDA, B3LYP is the most widely used hybrids type method, and M06-2X a typical meta-GGA method, while CAM-B3LYP and wB97X are long-range corrected hybrid density functionals. In this study, we used SVWN [59], B3LYP [60], CAM-B3LYP [61], M06-2X [62–64], wB97X for target molecular systems. This includes representatives of local, semilocal and nonlocal functionals.

1. Local Density Approximation (LDA)

The oldest and simplest DFT functional is LDA, which is based on average or uniform electron gas (UEG), which does not apply to chemical problems. The exchange-correlation energy density is assumed to be as uniform as possible in the space of molecules, with the same energy at each location. Therefore, it is considered that electrons can be considered to be uniformly distributed in a small volume. An energy density function approximation based on a uniform electron gas model is established within the spatial volume element, except that the electron density is different at different points.

$$E_{XC}^{LDA}[\rho] = E_{X}^{LDA}[\rho] + E_{C}^{LDA}[\rho]. \tag{2.28}$$

The exchange energy is as the following equation:

$$E_{X}^{LDA}[\rho] = C \int n^{4/3}(\vec{r})d\vec{r}. \tag{2.29}$$

In general, exchange-correlation functionals are represented in the form of unrestricted spins, i.e., two electrons α and β can have different spatial orbitals. Therefore, replacing two variables with one variable improves the applicability of the function. For systems that are not equivalent to spintronics, using two spin density functionals will yield more accurate results. This method is called local spin density approximation (LSDA). The exchange of energy depends only on the electron density at a given location, so the calculation is simple. As a result, LDA calculations are speedy and generally provide good geometry. However, sometimes, the result is a systematic error in energy due to stronger bonding or excessive bonding.

2. Generalized Gradient Approximation (GGA)

   In general, GGA provides enhanced results for LDA. In order to correct errors due to uneven distribution of electron density, and electron density gradient characterizing the inhomogeneity is included in the expression of the energy density functional. This functional is divided into two parts, exchange and correlation functionals, and also derived separately. The exchange energy does not depends on the value of density at a point as

in LDA, but depends on its gradient as follows:

$$E_{XC}^{GGA}[\rho] = \int \rho(\vec{r})\varepsilon_{XC}(\rho(\vec{r}), \mid \nabla\rho(\vec{r}) \mid)d\vec{r}. \tag{2.30}$$

Most of the GGA functionals have been modified from LDA functionals and have the following associations added.

$$\varepsilon_{XC}^{GGA}[\rho] = \varepsilon_{XC}^{LDA}[\rho] + \Delta\varepsilon_{X/C}\left[\frac{\left|\nabla\rho(\vec{r})\right|}{\rho^{4/3}(\vec{r})}\right]. \tag{2.31}$$

If the functionals contain empirical parameters, the values are fitted to reproduce the experimental result, such as exchange B(Becke), CAM, and the correlation B88, LYP. On the other hand, the functionals exclude the empirically determined parameter as the the following: exchange B86, PBE, and the correlation is PW91.

3. Hybrid Exchange Functionals
   These functionalities contain the exact fraction of HF exchange energy, which comes from the molecular orbital functional of KS, which usually contains the following form:

$$E_{XC} = (1-a)\cdot E_{XC}^{DFT} + a\cdot E_{X}^{HF}. \tag{2.32}$$

In general, the exchange energy is significantly larger than the correlation energy. In the HF method, the exchange energy can be accurately calculated, but the correlation can be complicated to calculate. So the Becke process takes advantage of the hybrid approach in 1993, so-called B3LYP 3-parameter functional or Backe3LYP. The calculation of the method includes using the HF method and the DFT method to calculate the exchange energy, and the DFT to calculate the correlation energy. This functional is widely used for molecular calculations, especially for many organic molecule calculations.

$$E_{XC}^{B3LYP} = (1-a)\cdot E_{XC}^{LDA} + a\cdot E_{X}^{HF} + b\cdot\Delta E_{X}^{B} + (1-c)\cdot E_{c}^{LDA} + c\cdot E_{c}^{LYP}, \tag{2.33}$$

where $a = 0.1161$, $b = 0.9262$ and $c = 0.8133$. Basically, there are many hybrid functionals, for instance B3LYP, which is widly used.

4. Hybrid Meta-GGA

   Now that we have the idea of mixing different functionals, we can continue to develop GGA. A meta-GGA DFT functional in its original form includes the second derivative of the electron density. Donald Truhlar at the University of Minnesota published M05 functional family in 2005. The M06 family represents a general improvement over the M05 family and a series of functionals that are constantly being improved. Different types are mixed with different parts. Among them, M06-L is a computational functional based on strong interactions such as metal, and M06 is an improvement, which applies to the corresponding system. And M06-HF is mixing Hartree–Fock and approximate DFT exchange, suitable for systems with non-covalent bonds. M06-2X is its improvement, mixed global hybrid functional with 54% HF exchange. It is worth noting that the M06 functionals are mixed with different levels of empirical optimization parameters, which is one of the reasons why they can only be optimized for different systems. The exchange-correlation functionals are shown as follow:

$$\begin{aligned} E_{\mathrm{X}}^{M06} =& E_{\mathrm{C}}^{PBE}(\rho, \nabla\rho, \nabla^2\rho) + E_{\mathrm{C}}^{LS\,DA} \\ E_{\mathrm{C}}^{M06} =& E_{\mathrm{C}}^{\alpha\beta} + E_{\mathrm{C}}^{\alpha\alpha} + E_{\mathrm{C}}^{\beta\beta} \\ E_{\mathrm{XC}}^{M06} =& pE_{\mathrm{X}}^{HF} + (1-p)\, E_{\mathrm{X}}^{M06} + E_{\mathrm{C}}^{M06}, \end{aligned} \tag{2.34}$$

   Where exchange engery not only used gradient, but also include 2-order gradient, and in the correlation functional, oppsite-spin($\alpha\beta$) and parallel-spin($\alpha\alpha$ and $\beta\beta$) are treated separately, finally, the p is determined by fitting to the data in the training set.

2.3.4 Basis set superposition error

When calculating the weak interaction energy between multibody molecules, such as the stacked molecules system A and B, usually:

$$E_{interaction} \neq E_{\mathrm{AB}} - E(A) - E(B). \tag{2.35}$$

Because the non-additivity of E_{AB} energy relative to $E(A) + E(B)$ does not only contain the interaction energy between the real A and B molecules. On the other hand, the basis functions from the A and B molecules overlap in the complex system. In other words, the basis set of the complex is increased, resulting in a decrease of $E(AB)$ energy. In order to remove this part of the contribution $E_{interaction}$, to prevent overestimation of the interaction energy, it must be corrected using the basis set superposition error (BSSE) [65]. So the interaction energy of the two molecules should be expressed as $E_{interaction} = E_{\mathrm{AB}} - E(A) - E(B) + E_{\mathrm{BSSE}}$. For weak interactions, the ratio of E_{BSSE} to $E_{interaction}$ is often not small or even exceeds it. If not corrected, the symbols may be wrong.

There are several ways to calculate E_{BSSE}, and the counterpoise method is currently the most widely used, developed by Boys and Bernardi. It should be noted that this method only calculates the actual E_{BSSE} approximation, which is not completely strict and precise, and there is no rigorous method to calculate E_{BSSE}. Let E_i be the energy of the i-th molecule under its basis set and E_i' be the energy of the i-th molecule appearing under the basis functions of all n molecules. Then the interaction energy of n molecules is

$$E_{\mathrm{BSSE}} = \sum_{[i]} (E_i - E_i'). \tag{2.36}$$

For the variational method, since the basis set is larger and the energy is lower, E_i' is more negative than E_i, so E_{BSSE} must be positive. Since the composite structure we use to calculate weak interactions is generally optimized. Therefore, generally used counterpoise calculations are generally used to optimize the composite structure.

2.4 Correction of long-range interactions

2.4.1 Intermolecular forces

Chemical bonds refer to the mutual chemical interactions of atoms in a molecule and broadly include interactions between molecules. Two or more atoms or

ions rely on chemical bonds to combine atoms into stable molecules or crystals. Generally, there are three chemical bonds: covalent bonds, ionic bonds, and metal bonds. Among them, the molecules are mainly covalent bonds. Ionic bonds and metal bonds are present in the ionic compound and the metal, respectively. Hydrogen bonds (H-bonds) are sometimes formed between molecules and within molecules, and their strength is between covalent bonds and vdW forces.

The intermolecular force is a general term for the interaction between groups other than covalent bonds, ionic bonds, and metal bonds. It mainly includes ion or radical groups, dipoles, interaction forces between induced dipoles, H-bonding forces, hydrophobic group interaction forces, and non-bonded electron repulsive forces. Most of the intermolecular interactions are below 10 Kj/mol, which is 12 orders of magnitude smaller than the usual covalent bond energy, with a range of 0.3-0.5 nm. Moreover, there is generally no directionality and saturation other than H-bond.

The three forces proportional to $1/r^6$ shown in Table 2.1 are known as vdW forces. It is the role of people in the study of gas behavior, the discovery of the attraction and repulsion between molecules in the gas phase, using the vdW equation to correct the deviation of the actual gas against the ideal gas pair.

Table 2.1: Types of intermolecular-forces. [6]

Type	relation with distance
Charged group electrostatic action	$1/r$
Ion - dipole	$1/r^2$
Ion-induced dipole	$1/r^4$
Dipole - dipole	$1/r^6$
Dipole - induced dipole	$1/r^6$
Induced dipole - induced dipole	$1/r^6$
Non-key repulsion	$1/r^9 - 1/r^{17}$

There are three main sources of vdW forces: electrostatic force, induced force, and dispersion forces.

1. **Electrostatic force**

Polar molecules have a permanent dipole moment, and electrostatic attraction occurs between dipole moments. The average energy is:

$$E = -\frac{2}{3}\frac{\mu_1^2\mu_2^2}{kTr^6} \times \frac{1}{(4\pi\varepsilon_0)^2}, \tag{2.37}$$

where μ is the dipole moments, respectively. r is the distance from the center of mass of the molecule, k is the Boltzmann constant, T is absolute temperature, and negative value represents energy reduction, ε_0 is the vacuum permittivity, also known as vacuum dielectric constant or electrical constant. The electrostatic force increases as the dipole moment increase. For the same type of molecule, since the dipole moment is the same, it is proportional to the fourth power of the dipole moment. When the temperature rises, the orientation of the dipole molecules is destroyed, and the interaction energy is lowered, so it is inversely proportional to the absolute temperature.

2. **Inductive force**

   The permanent dipole moment induces adjacent molecules, causing charge displacement and induced dipole moments. There is an attraction between the permanent dipole moment and the induced dipole moment. The energy of this interaction is called inductive energy. The average induced energy between molecule 1 with a dipole moment of μ_1 and molecule 2 with polarizability of α_2 is:

$$E = -\frac{\alpha_2\mu_1^2}{(4\pi\varepsilon_0)^2 r^6}. \tag{2.38}$$

3. **Dispersion force**

   Non-polar molecules have an instantaneous dipole moment. The instantaneous dipole moment induces a dipole moment in the adjacent molecule, and the interaction between the instantaneous dipole moment and the induced dipole moment is called the dispersive force. This interaction energy is called dispersive energy. London introduced the approximate expression

of the dispersive energy between two molecules as:

$$E = -\frac{\frac{3}{2}I_1 I_2}{I_1 + I_2}\left(\frac{\alpha_1\alpha_2}{r^6}\right)\frac{1}{(4\pi\varepsilon_0)^2}, \tag{2.39}$$

I_1 and I_2 are the ionization energies of two interacting molecules, α is their polarizability, which reflects whether the electron cloud in the molecule is easily deformed. When the number of electrons in the molecule increases, the atom becomes larger, the outer electron is farther away from the core, the polarizability increases, and the dispersion force increases. When there is a π bond in the molecule, the electron cloud is more easily deformed than the σ key; if there is a delocalized π key, α is generally larger.

Electrostatic force and inducing force exist only in polar molecules, and dispersive forces exist in either polar or non-polar molecules. These forces exist not only between different molecules but also in different atoms or groups within the same molecule. Between the regiments. Experiments have shown that these three forces between the general molecules, the dispersion force is dominant.

2.4.2 Polarization interactions

We can use the Rayleigh-Schrödinger perturbation theory to describe the polarization interactions as 2nd-order or higher-order perturbation terms. The approximation solution of Schrödinger equation is considerably simplified to solve the complex systems, according to perturbation theory, when the unperturbed Hamiltonian H_0 add a small perturbation correction as a sum:

$$H = H_0 + V, \tag{2.40}$$

where the solution of H_0 is assumed to be known:

$$H_0\Psi_m^{(0)} = E_m^{(0)}\Psi_m^{(0)}. \tag{2.41}$$

The 0-order undisturbed solution of Schrödinger equation can use different power series of V to perform different forms of expansion. And the expansion of the

perturbation theory is not unique. If we expand E and c_m as series and match them to n-order perturbations:

$$\begin{aligned} E &= E^{(0)} + E^{(1)} + E^{(2)} + \cdots \\ c_m &= c_m^{(0)} + c_m^{(1)} + c_m^{(2)} + \cdots , \end{aligned} \tag{2.42}$$

for the correct state n, where $E^{(n)}$ and $c_m^{(n)}$ are the first-order as V^n, and can get their specific form as:

$$E_n^{(1)} = V_{nn} = \left\langle \Psi_n^{(0)} | V | \Psi_n^{(0)} \right\rangle . \tag{2.43}$$

For higher-order expressions, they are more complex, but can still be derived from undisturbed forms. In the 2nd-order, the energy of interaction expression is given as follow:

$$E_{pol}^{(2)} = -\sum_{n,m}^{q} \frac{\left| \left\langle \Psi_n^A \Psi_m^B | V | \Psi_0^A \Psi_0^B \right\rangle \right|^2}{\left(E_n^A - E_0^A\right) + \left(E_m^B - E_0^B\right)} = E_{ind}^{(2)} + E_{disp}^{(2)}, \tag{2.44}$$

where the summation counts the interactions of each electron pairs, and the corresponding electrons n and m cannot be in the ground state at the same time. The addition of this part can be physically interpreted as the interaction of the induction interactions $E_{ind}^{(2)}$ and the dispersion interaction $E_{disp}^{(2)}$.

2.4.3 London dispersion theory

Dispersion interactions were defined by London in 1930 and can be presented $E_{disp}^{(2)}$ from the previous section. [66–68]

$$E_{disp}^{(2)} = -\sum_{m,n\neq 0} \frac{\left| \left\langle \Psi_n^A \Psi_m^A | V | \Psi_0^A \Psi_0^B \right\rangle \right|^2}{\left(E_n^A - E_0^A\right) + \left(E_m^B - E_0^B\right)} = -\sum_{m,n\neq 0} \frac{\left| V_{nm,00} \right|^2}{\left(E_n^A - E_0^A\right) + \left(E_m^B - E_0^B\right)}. \tag{2.45}$$

The matrix element, appearing in this equation, corresponds to the electrostatic interaction of two mutually induced electron distributions, $\rho_{n0}^A(i)$ and $\rho_{m0}^B(j)$, and may be expressed as follows:

$$V_{nm,00} = \int \rho_{n0}^A(i) \rho_{m0}^B(j) \frac{e^2}{r_{ij}} dV_i dV_j. \tag{2.46}$$

The dispersive force is mainly determined by the quantum fluctuation, and is also affected by the instantaneous induced polarity caused by the change of the electron density.

Due to the electronic movement, the distribution of electrons cannot always be an average distribution. So there will be a momentary charge distribution that produces a momentary dipole moment of the molecule that will induce multipole moments on the other molecule. This energy is always negative in the ground state, indicating that this transient interaction corresponds to an attractive force. London believes that the main terms of the dispersive force can be expressed by the electric field vibration energy generated by the zero-point vibration of the interacting wave dipole moment.

As London's theory, the multipole expansion for dispersion energy is usually written as follow, which is a series with coefficients C_n:

$$E_{disp}^{(2)} = -\sum_{n=6}^{\infty} \frac{C_n}{r^n}. \tag{2.47}$$

It is worth noting that since the interaction is between particles, n can only take an even number. The starting point is $n = 6$, which corresponds to the interaction between dipoles, and when $n = 8$, ($1/r^8$) is a dipole and quadrupole interaction, furthermore, the dipole-octopole and quadrupole-quadrupole interactions in the calculation of the term $1/r^{10}$. Since the quantum mechanical fluctuations are asymmetrical to the ground state electrons, the interaction between the dipoles corresponds to the dispersion multipole, the direct electrostatic interaction is $1/r^7$ term, $1/r^{14}$ term is for rotating molecules and induction interaction, the leading term is $1/r^{10}$.

In the calculation of the dispersion energy in the ground state of two hydrogen atoms [69], the method for accurately evaluate C_6 was verified. Which method was also used to evaluate the dispersion forces in our B-DNA target system in this study. The results show that the non-additive contribution evaluation of a four-body system under the limit state should be in a negative value.

Assuming that there is no polarity of the whole target system, such as a

spherically symmetric system, then we can represent the dispersion coefficient of the dipole-dipole interaction C_6 as the equation:

$$C_6^{AB} = \frac{3}{2} \sum_{n,m\neq 0} \frac{f_{n0}^A f_{m0}^B}{\omega_{n0}^A \omega_{m0}^B \left(\omega_{n0}^A + \omega_{m0}^B\right)}, \tag{2.48}$$

where f_{n0} is the dipole-oscillator strengths of $0 \rightarrow n$ type for one molecule, and it could be expressed in detail about the expectation value of the dipole moment d_{n0}:

$$f_{n0} = \frac{2}{3} \omega_{n0} \left| d_{n0} \right|^2 \tag{2.49}$$

where d_{n0} is from n state to ground state 0. The average dynamic polarizability of the target systems could be defined as function corresponding to the dipole transitions while the static polarizability at $\omega = 0$ and $\omega \neq 0$:

$$\begin{aligned} \overline{\alpha}(0) &= \overline{f}_{k0} / \overline{\omega}_{k0}^2 \\ \overline{\alpha}(\omega) &= \sum_{n\neq 0} \frac{f_{n0}}{\omega_{n0}^2 - \omega^2}. \end{aligned} \tag{2.50}$$

Then, we got the well-known formate London formula by replaced the averaged transition frequencies by the first ionization potentials (empirical parameters):

$$C_6 = \frac{3}{2} \overline{\alpha}_A(0) \overline{\alpha}_B(0) \frac{I_A I_B}{I_A + I_B}. \tag{2.51}$$

2.4.4 DFT+D

Because of the high computational cost of the standard wave function theory (WFT) method, WFT can only handle small-sized systems. For a slightly larger system, even large-scale supercomputers are difficult to calculate, while DFT can easily calculate systems that include more than 200 atoms. However, in the actual calculation of the standard DFT functionals (including local or semilocal), because of its theoretical limitations, the long-range London dispersion energy is not correctly described, which made they unsuitable for the calculation of intermolecular non-covalent interactions. Even when the exchange of electron sets

can be estimated theoretically accurately, it is impossible to accurately estimate the correlation energy [70, 71]. Then, for the physical properties that are mainly affected by the electronic correlation, the results of the standard DFT calculation are not satisfactory, and we need to extend the standard DFT. Therefore dispersion-corrected DFT (DFT+D) provides a practical tool for the investigation and analysis of many-body molecular systems.

According to the previous elaboration of London theory, the electronic shock will cause the charge density to deform, causing an instantaneous dipole moment. Which can distort the charge density of other atoms or molecules, resulting in induced dipole moments at the same time. The existence of two dipole moments forms a total interaction. London gives the relationship between two spherically symmetric atoms with a large mutual distance. The general expression of this interaction is:

$$V^{dispersion} = -C/r^6, \tag{2.52}$$

where C is the compound physical values (including dispersion relationship) in inverse proportion to the 6th power of the distance of the atoms r. A simple practical example that can account for dispersion interactions is the dimer of noble gas atoms such as Ne. It is well known that these atoms are very chemically stable. However, these gases can be liquefied at sufficiently low temperatures, indicating that there is an attraction between the noble gas atoms.

Although the dimer must have minimum potential energy, the minimum value that can be found when using the B3LYP calculation depends on the calculation details. In the calculation without counterpoise corrections, the average absolute error of the equilibrium atomic distance prediction is 0.13 Å, and the average absolute error of the interaction energy is 0.24 kcal/mol. Conceptually, a simple remedy for the DFT's handling of dispersion interactions is to correct the total energy with a dispersion-dependent contribution between each pair of atoms. This idea has been developed in the localized base method, the so-called DFT+D method. In the DFT+D calculation, the atomic set can always be augmented with

the E_{DFT} calculated by the DFT method to the following form:

$$E_{DFT+D} = E_{DFT} + S_6 \sum_{i!=j} \frac{C_{ij}}{r_{ij}^6} f_{damp}(r_{ij}), \tag{2.53}$$

where r_{ij} is the spacing, and C_{ij} is the dispersion coefficient, which can be calculated from the atomic attribute list; $f_{damp}(r_{ij})$ is a damping function that prevents the dispersion term from being unrealistic when the distance is small. The only empirical parameter S_6 in this expression is a scale factor that is uniformly applied to all atom pairs. For each functional, the scale factor should be estimated separately in advance, and the method is to continuously optimize the value of the scale factor for a group of molecular complexes that have an important influence on the dispersion interaction [37].

2.4.5 Non-additivity

Unlike classical mechanics, the force field superposition of quantum systems cannot be added by the instantaneous polarization caused by quantum fluctuations. Therefore, it has a non-additive effect. For example, in a two-body system, there are two molecules A and B. The typical intermolecular interaction is that the polarization fluctuation of molecule A causes polarization fluctuation of molecule B, causing interaction. At this point, the intermolecular force of the total system can be expressed as follows:

1. Coulomb force of A and B.
2. B is subjected to the polarization of A's polarization.
3. A is affected by the polarization fluctuation of B.
4. A is subjected to the polarization fluctuation of B and the polar force of B after the polarization fluctuation of A.

After the addition of the third molecule C, the force analysis of the system is much more complicated. The fluctuation of molecule A also causes the fluctuation of the molecule C, and the polarization fluctuation of molecule C also affects the fluctuation caused by the molecule B. In this case, it is obvious that this is not a simple superposition. Therefore, it is not enough to analyze only the non-additive effects of traditional calculation methods on quantum systems.

Let us further analyze the situation of this non-additive after the introduction of quantum mechanics. In classical mechanics, an object can be described as a point charge or object in a multibody system. In general, in most physical laws (such as Coulomb's law and Newton's law of gravity), the interaction of a point is characterized by its additivity. Because the objects in the system are rigid, additivity can always be expected. However, the situation is different in quantum mechanics. Because the charge cannot be regarded as a rigid point. The internal electronic structures of molecules and atoms are dynamic, which results in constantly changing the electron potential fields. Electrons are also susceptible to other interaction forces that produce energy changes, such as inductive dipoles. Under such quantum fluctuation conditions, additivity will inevitably not work. Non-computational forces are generated, including polar forces and exchange energies. The non-additivity is expected in inter-molecular bindings due to the induced polarizations by the quantum fluctuations, such as vdW forces.

The description and reproduce of the binding itself due to intermolecular forces is a huge challenge for *ab initio* method. Non-additivity is a more difficult subject that has long been far from the mainstream research field and has not been well analyzed yet. The *ab initio* quantum chemistry theory is a good description of the natural stacking energy, which allows reliable energy to be found on any base structure. Calculations, in any case, need to be done at a sufficient theoretical level. For example, standard DFT, HF, and semi-empirical methods all fail in the description of base stacking because they cannot correctly capture the dispersion effect. However, some work applied to systems consisting of weakly constrained subsystems shows that non-additiveness is much larger than we expected [42, 43]. Although there is non-additivity in molecular systems, if the non-additive contribution is positive and tiny, means cohesion reduced than the superposion, then there is no research significance. If so, it will only make minor corrections to the C_6 force without any qualitative impact.

2.5 Quantum monte carlo methods

Although the problem of solving the Schrödinger equation is simplified to an eigenvalue problem (or a generalized eigenvalue problem) based on the HF method and the variational principle. The computational complexity of MO-based methods is very high: when the number of system electrons n rises, the wave function $\psi(r_1, r_2, \cdots, r_n)$ dimension describing the many-body system state also rises linearly, and the number of bases needed to describe the state of the system correctly must rise exponentially. In summary, when the number of electrons rises, the computational complexity becomes very large to solve the Schrödinger equation variational form by the matrix method. At the same time, according to the description of the relevant DFT chapter, the approximation of the density functional makes this complexity greatly reduced. However, the exact XC functional is always unknown, which makes the accuracy of the DFT only stay at the theoretical level. In standard DFT, the lack of dispersion term in XC makes the DFT's credibility greatly reduced. Even if the long-range force correction is added, it can only be considered as a reinstatement. Then we have to look back to the WFT methods to solve the problem accurately for comparing, but this time we use the method of random statistics to reduce the difficulty of calculation. In this way, we can not only approximate the exact solution infinitely in theory, but also, the statistical method has parallel acceleration efficiency in the current parallel computers.

Quantum Monte Carlo (QMC) technology provides a direct and potentially efficient means of solving the many-body Schrödinger equation for quantum mechanics. The simplest QMC, also known as VMC, is based on the direct use of Monte Carlo (MC) integrals to calculate the expected value of multidimensional integration, such as total energy. The MC method is statistical, so the key result is to use the integral value calculated by MC, the convergence speed is faster than the traditional numerical integration method when the problem involves multidimensional relationships. The statistical method thus provides a practical approach to solving the direct integration of the multibody Schrödinger equation and requires only an estimate that is easily controlled. Recent advances in accurate

calculation methods, especially through DMC calculations, make it possible to handle larger systems. [28, 30–41, 43]

2.5.1 Monte Carlo methods

The MC method is a class of methods that use random numbers to implement a certain type of computation. When dealing with high-dimensional integrals, MC integrals have the characteristic that the precision increases linearly with the number of points. Metropolis Monte Carlo (MMC) in an important sampling method that can obtain a random number sequence that conforms to an arbitrary distribution through detailed balance condition.

Since the value of the real wave function in most areas of the Hilbert space is minimal, the sparse problem cannot be avoided when solving the integral method. At this time, if the traditional integration method or MC integration is performed directly using a uniform random number, a large number of computational resources will be wasted. In this case, it is preferable to use the MMC method for integration: Using a large number of particles to perform random walks in the wave function space, the migration rules corresponding to MMC will "push" them to areas with large wave function values to achieve more efficient sampling integration.

2.5.2 Variational principle

As mentioned earlier, since the actual wave function is difficult to derive directly, we need to accurately normalize the eigenstate Hamiltonian. The variational principle can be used in quantum mechanics, which is obtained by extending the normalized trial wave function ψ_T. This principle is also the main theoretical basis of VMC [72, 73]. Suppose we get an infinite wave function ψ_i based on statistical principles, then the trial wave function can be expressed as:

$$\psi_T = \sum_{i=0}^{\infty} c_i \psi_i, \tag{2.54}$$

where c_i is the expansion factor, and $\{c_i\}$ can be normalized to 1 as follow:

$$\sum_{i=0}^{\infty} |c_i|^2 = 1. \tag{2.55}$$

The many-body Hamiltonian can be expanded as follows:

$$\begin{aligned} \langle \psi_T | \hat{H} | \psi_T \rangle &= \left\langle \sum_i c_i \psi_i \middle| \hat{H} \middle| \sum_j c_j \psi_j \right\rangle \\ &= \sum_i \sum_j c_i^* c_j \langle \psi_i | \hat{H} \middle| \psi_j \right\rangle \\ &= \sum_i |c_i|^2 \varepsilon_i, \end{aligned} \tag{2.56}$$

where the energy eigenvalue of a single electron ε_i can be calculated from single wave function as $\varepsilon_i =< \psi_i | \hat{H} | \psi_i >$. Therefore, the expected value of the trial wave function must be greater than or equal to the true ground state energy. The key to variational calculation is to rely on the form of the trial wave function. By selecting the trial wave function on the basis of physical motivation, an accurate wave function can be obtained. Typically, use the wave function obtained from HF method or similar calculations and add additional parameters to construct additional physical properties, such as the known limits and derivatives of the many-body wave function. And then the additional variation degrees of freedom are then used to further optimize the wave function.

2.5.3 Trial wave functions

According to the above, the choice of the wave function determines whether the accuracy and variational principle of the VMC can obtain the exact ground state energy. All observations are related to the probability distribution $\left|\Psi_T(\vec{r})\right|^2$. In order to obtain accurate results of the observations, the trial wave function constituting this probability distribution must be able to obtain a good eigenstate. A good trial wave function also improves the important sampling and reduces the cost of obtaining accurate statistical accuracy. Any wave function can be utilized by QMC whose physical value, gradient, and laplacian can be effectively

calculated. The power of QMC is the flexibility of the trial wave function form. We have to find a trial wave function that is accurate and easy to estimate. In quantum chemistry methods, it is generally desirable to expand a many-body wave function into a linear combination of determinants. However, the convergence of this expansion form is slow because it is difficult to describe cusps, when any two electrons are associated. QMC needs a more reasonable and generalized trial wave function, Slater-Jastrow form: [74]

$$\Psi_T = D\left(\vec{r}_{ij}\right) \exp\left[\sum_{i<j}^{N} J\left(\vec{r}_{ij}\right)\right]. \tag{2.57}$$

It consists of a single Slater determinant multiplied by a Jastrow correlation factor [74, 75], including the cusps [76]. The orbitals in the Slater determinant are usually derived from HF or DFT calculations. The Jastrow factor is chosen by some special functional form and optimized its parameters. The role of the $J\left(\vec{r}_{ij}\right)$ function is to minimize the energy of the entire system, and it is chosen to increase the probability of the particle appearing at the position of the lowest interaction energy. The variation of this method has been successfully applied to various systems, multiplying the determinant wave function by the many-body correlation function. Choosing a good correlation function that contains the relevant effects is more efficient than CI-based methods. A common application and a simple Jastrow factor should be expressed as:

$$J\left(\vec{r}\right) = \frac{A}{\vec{r}}\left(1 - e^{-\frac{\vec{r}}{F}}\right), \tag{2.58}$$

where F is parameterized according to A and chooses to satisfy the electron-electron cusp conditions constraint. The better approximation of the exact many-body wave function and the development of its acquisition methods will remain an important area of research as the results increase the level of accuracy and efficiency.

2.5.4 Variational Monte Carlo

The VMC is a direct application based on MC integration and is used to integrate explicit correlated many-body wave functions. The variational principle in quantum mechanics discussed above indicates that the energy of the trial wave function will always be greater than or equal to the energy of the actual wave function. The exact system energy can be accurately determined by using the optimized form of many-body wave function. According the variational principle, a samples configurations-set of electron positions $|\vec{r}\rangle = \{\vec{r}_i\}$ is constructed from the probability distribution $\left|\Psi_T\left(\vec{r}\right)\right|^2$, where$\Psi_T$ is our trial wave function and satisfy normalized conditions $\int_i \mathrm{d}\vec{r}\left|\vec{r}\right\rangle\left\langle\vec{r}\right| = 1$. We define local energy:

$$E_L(\vec{r}) = \frac{\left\langle\vec{r}\right|\hat{H}\left|\Psi_T\right\rangle}{\left\langle\vec{r}|\Psi_T\right\rangle} = \frac{\hat{H}\Psi_T(\vec{r})}{\Psi_T(\vec{r})}. \tag{2.59}$$

Then, variational enery is obtained by averaging the $E_L(\vec{r})$:

$$E_V = \frac{1}{N}\sum E_L(\vec{r}_i). \tag{2.60}$$

If $\Psi_T(\vec{r})$ is equal to the true wave function, then this local energy is independent of the position $\vec{r}$. Therefore, the energy can be expressed in the following integral form:

$$E_V = \frac{\int \Psi_T^2(\vec{r})E_L(\vec{r})d^3\vec{r}}{\int \Psi_T^2(\vec{r})d^3\vec{r}}. \tag{2.61}$$

Then we can use the distribution:

$$\rho(\vec{r}) = \frac{\Psi_T^2(\vec{r})}{\int \Psi_T^2(\vec{r}')d^3\vec{r}'}, \tag{2.62}$$

to perform the calculation of the MMC sample. The variational QMC method directly calculates the ground state of a multibody system using the ground state energy minimum principle. Its advantage is intuitive and clear. However, the obvious disadvantage is that this method is more dependent on the quality of the constructed and modified wave functions.

2.5.5 Calculation process of VMC

The VMC algorithm contains two distinct phases. In the first phase, a walker containing an initial random set of electronic locations will propagate or multiply according to the Metropolis algorithm. This step is to equilibrate and then start sampling the electronic probability distribution $\pi = |\Psi|^2$. In the second phase, the walker will continue to move, but energy and other observables will also accumulate here, preparing for later average and statistical analysis. To simplify the notation, the wave function $\Psi(\vec{r}')$ is used to indicate that a single electron moves from position $\vec{r}$ to $\vec{r}'$, but no other electrons have a moving wave function. [77, 78]

Algorithm 2.1 The Psudocode of VMC Equilibrium phase

```
//Equilibrium phase to get the distribution π
for i=0; i → m ; equilibrated as r_{i+1} = r'_i do
    Generateing an initial position configuration {r_i} of electrons randomly
    Generateing distribution probability ρ_i
    for each electron on the configuration do
        Propose i th. move from r_i to r'
            with conditional probability T(r'|r_i)
        Update a new distribution ρ(r')
        Evaluate a Metropolis acceptance rate A = min(1, |Ψ(r')/Ψ(r_i)|^2)
        Generate random number x
        if x < A then
            Accept the move r_{i+1} = r';
        else
            Reject the move r_{i+1} = r_i;
        end if
    end for
end for
```

In this algorithm, electrons move independently, rather than all. This is to improve the computational efficiency of the algorithm in large systems, that is to say, coordination movement needs to increase the small movement step to maintain the acceptance rate. Observables also need to be accumulated on each electronic basis and then weighted according to the acceptance and rejection

Algorithm 2.2 The Psudocode of VMC Accumulation phase

```
//Accumulation phase to get the local energy and other observables
for i=0; i → m ; ufficient data accumulated do
    for each electron on the configuration do
        Propose i th. move from r⃗_i to r⃗' with T(r⃗'|r⃗_i)
        Update a new distribution ρ(r⃗')
        Evaluate a Metropolis acceptance rate A = min(1, |Ψ(r⃗')/Ψ(r⃗_i)|^2)
        Accumulate the local energy, and other observables,
            at r_i and r⃗_i, weighted by the A one by one.
        Generate random number x
        if x < A then
            Accept the move r_{i+1} = r⃗';
        else
            Reject the move r_{i+1} = r⃗_i;
        end if
    end for
end for
```

probabilities:

$$< O >= \frac{1}{m} \sum_{i=1}^{m} \left[T_i O_i \left(\mathbf{r}' \right) + \left(1 - T_i \right) O_i(\mathbf{r}) \right], \tag{2.63}$$

where represents the acceptance rate after the m move, where T_i is the acceptance rate of the electron i, and $O_i(\mathbf{r}')$ is the contribution value of the observables. The calculation of this formula improves statistical statistics, which is averaged compared to positions that are only based on movement or not, and can be applied to any observables.

2.5.6 Diffusion Monte Carlo

Obviously, VMC calculations rely heavily on the quality of trial wave functions. [77–80] This limitation can be overcome with the aid by using a projection technique to enhance the ground-state component of a starting trial wave function. The DMC method [72, 73] is based on the Schrödinger equation with imaginary time:

$$\frac{\partial |\Psi\rangle}{\partial \tau} = -\hat{H}|\Psi\rangle, \tag{2.64}$$

Where $\tau = it$ and the state $|\Psi\rangle$ represents the eigenstate of the Hamiltonian quantity:

$$\begin{aligned} |\Psi\rangle &= \textstyle\sum_{i=0}^{\infty} c_i \, |\phi_i\rangle, \\ \hat{H} \, |\phi_i\rangle &= \varepsilon_i \, |\phi_i\rangle. \end{aligned} \tag{2.65}$$

One solution to this form of equation is to use an exponential form of orthogonal projection, given the existence of an unbounded spectrum:

$$|\Psi\left(\tau_1 + \delta\tau\right)\rangle = e^{-\hat{H}\delta\tau} \, |\Psi\left(\tau_1\right)\rangle \, . \tag{2.66}$$

The projection operator in the form of an exponent is represented by the following form of the Green's equation:

$$\mathcal{G}\left(x', x\right) = \left\langle x' \left| e^{-\hat{H}\delta\tau} \right| x \right\rangle . \tag{2.67}$$

This actually contains an exponential form of Hamiltonian that can compute all of the base elements of the base set $|x\rangle$ to $|x'\rangle$. Using this projection operator, we can use the orthogonal projection on the imaginary time to continuously approximate the ground state energy $|\phi_0\rangle$:

$$\lim_{\tau \to \infty} |\Psi(\tau)\rangle = c_0 e^{-\varepsilon_0 \tau} \, |\phi_0\rangle \, . \tag{2.68}$$

This is somewhat similar to the variational principle, but in the virtual time projection, the convergence of this excited state energy to the ground state energy can become very rapid, when we consider the particle's spatial coordinates on $\mathbf{R}$:

$$\lim_{\tau \to \infty} \Psi(\mathbf{R}, \tau) = c_0 e^{-\varepsilon_0 \tau} \phi_0(\mathbf{R}). \tag{2.69}$$

Thanks to the Trotter approximation, which differs in the Green's equation in a small virtual time, the diffusion equation can be used to change the form as:

$$-\frac{\partial \Psi(\mathbf{R}, \tau)}{\partial \tau} = \left[\sum_{i=1}^{N} -\frac{1}{2} \nabla_i^2 \Psi(\mathbf{R}, \tau) \right] + \left(V(\mathbf{R}) - E_T \right) \Psi(\mathbf{R}, \tau). \tag{2.70}$$

We can see that the first item on the right side of the formula is related to the diffusion process, which can be described by the density of the diffusing particles.

The second term is a rate term that can be solved using the branching scheme [81,82]. This process is based on a potential-dependent increase or decrease in the particle density, which is the so-called Birth-death process. The above equation can be transformed into a form suitable for the MC method, but such efficiency is very low. Because the potential energy V is borderless, parts of the second term can diverge and then cause a large error in the particle density and the expected real system. So we can use key sampling to reduce these problems. That is to add a guiding wave function, and this guiding wave function is very close to the estimation of the real system to constrain the diffusion process of the wave function.

2.5.7 Fixed-node approximation

Generally speaking, QMC is a method to solve the integrals of a ground state system by sampling, and then performs weighted averaging. For the fermionic system such as electronics, because of the antisymmetry of the fermionic wave function, as long as it is not half full filling, any wave function in the ground state, we can find the corresponding negative wave function, so that there will be positive and negative weights. Therefore, statistical results will cause cancelling with each other. Such statistical results are meaningless, leading to the so-called "sign problem". This problem is an NP-hardness problem that cannot be completely solved. It can only rely on approximation to improve the result. The most successful attempt to address to this "sign problem" is the fixed-node (FN) approximation [83–85].

The fixed node method is an approximation to the exact Fermi ground state, by translating a fixed nodal so that most of the statistical results are on the same sign side. Its accuracy depends on the established reference nodes, which makes the DMC results more controllable, rather than uncertain empirical parameter corrections. In practice, the nodes of an HF or DFT based wave function are found to be very accurate, giving energies well below those of the best VMC wave functions.

2.5.8 Calculation process of DMC

Just as VMC is divided into equalization phase and accumulation phase, DMC also performs calculations like this. However, the two phases of the DMC are basically the same as the energy. In the calculation of the branching coefficient of the evaluation configuration, these two phases are both required in each movement.

In the algorithm 2.3, step No. 7 is to propose a move as an equation:

$$\mathbf{r}_i' = \mathbf{r}_i + \tau \mathbf{F}(\mathbf{r}_i) + \eta, \tag{2.71}$$

where F is quantum force: $\mathbf{F}(\mathbf{R}) = \frac{\nabla\psi(\mathbf{R})}{\psi(\mathbf{R})}$, and η is a Gaussian random vector with an average sum of 0 and a variance of τ. And the weight for the movement $W(\mathbf{r}', \mathbf{r})$ is calculated as follow in the algorithm 2.3 step No. 14 :

$$W(\mathbf{r}', \mathbf{r}) = \frac{|\Psi_G(\mathbf{r}')|^2 \tilde{G}(\mathbf{r}', \mathbf{r}; \tau)}{|\Psi_G(\mathbf{r})|^2 \tilde{G}(\mathbf{r}, \mathbf{r}'; \tau)}, \tag{2.72}$$

where $\tilde{G}$ is a short-term estimate using the Green's equation. In the step No. 22, the branching factor P_B of the configuration j is calculated as

$$P_B = \exp\left[-\tau\left(\frac{1}{2}\left[E_L(\mathbf{R}') + E_L(\mathbf{R})\right] - E_T\right)\right]. \tag{2.73}$$

2.6 Conclusion

In this chapter, we discussed all electronic state calculation methods applied in the study. Unlike the empirical field, *ab initio* calculations can more accurately describe the electronic state. Correspondingly, higher computational costs are required, which is more pronounced for MO-based methods. The core of solving the many-body electronic state problem is to solve the many-body Schrödinger equation. Based on the BO approximation, we can simplify the equation and use the HF to get the solution of the equation. The addition of electronic correlations makes calculations accurate and expensive, and the advantage of these post-HF methods is that the level of accuracy can be determined. There is no doubt that the accuracy of the third-order CCSDT is much higher than that of the first-order

CCS. Furthermore, the MP4 accuracy is much higher than MP2, with a very high computational cost.

DFT, using the functional of the electronic density as exchange-correlation, opened up the revolution of SCF computing. On the other hand, DFT is a theoretically accurate calculation method, whose key problem is to find the correct functional. Although the precision of functionals can be constantly evolved, in practice, the construction of functionals inevitably incorporates empirical parameters to correct them. Especially in the calculation of long-range intermolecular forces, the instantaneous induced polarity caused by quantum fluctuations cannot be obtained from electron density. Furthermore, the non-covalent bond action itself is a quantum chemical challenge, and the non-additiveness generated by quantum wave action lacks research.

The high-precision MO method cannot handle large systems under the existing computational power, and the DFT can only introduce empirical parameters for correction. Therefore, the statistical method based on the accurate wave function is necessary. With the ability of parallel computing, statistical wave function methods can handle larger systems with an expansive computational cost. The simple QMC method is a VMC method based on the variational principle, but it still relies on the trial wave functions evaluated from other calculations. The introduction of projection operators makes the accuracy of the calculation eliminate the influence of the trial wave function. Although the FN approximation is applied to solve the sign problem, the accuracy of the DMC is still trustworthy. Moreover, with the development of parallel computers, statistical methods make the QMC algorithm more advantageous.

Algorithm 2.3 The Psudocode of DMC

while accumulated enough numbers **do**
  //The overall value of the initial N_c configuration, the total system should be irrelevant, and distributed according to the distribution of the guidance equation $\pi = |\Psi_G|^2$.
  Initialize the trial energy E_T
  for configuration j=0; configuration movement $j \rightarrow N_c$ **do**
    //According to the number of steps, generally $O(100 \sim 1000)$
    for electron i; $i \rightarrow m$; **do**
      Propose a move $\mathbf{r}'_i$
      //Apply the fixed-node approximation
      if $\Psi(\mathbf{r}')$ the same symbol as $\Psi(\mathbf{r})$ **then**
        Accept i move
      else
        rejects i move and consider the next one
      end if
      Calculate the weight of the move $W(\mathbf{r}', \mathbf{r})$.
      Generate random number x
      if $x < \min(1, W(r', r))$ **then**
        Accept the move $r_{i+1} = \vec{r}'$;
      else
        Reject the move $r_{i+1} = \vec{r}_i$;
      end if
    end for
    Calculate the branching factor P_B of the configuration j
    Accumulative local energy and any observables
      weighted by branching coefficients.
    Copy $int(P_B + u)$th configurations,
      where $u \sim$ normal distribution $[0, 1]$.
  end for
  E_T =AverageEnergy($E_T^{(previousblock)}$)
  //Update trial energyBring bring it closer to the current system
  Random birth-death process(walkers) //creat or delete walkers
    Reorganize walkers in this overall system to target number N_c
end while

Chapter 3

Systems and Methods

3.1 Problem Statement

The many-body problem of molecules is an important research topic in quantum chemistry. The complexity of intermolecular forces determines that there are still many mysteries in many-body systems. To explore the non-additivity of intermolecular interactions in the target system is the problem to be solved in this study. Due to quantum fluctuations, the non-additiveness of the interaction forces between macromolecules is always expected, as we described. The interaction between the molecules of a living organism is in the formation of its structure. For DNA molecules and the weak force between many-body molecules itself is the big challenge in the field of quantum chemistry. The non-additive study of the interaction between molecules is at the edge of the research field. In the traditional methods, which are widely used, the non-additive contributions is tiny with positive sign. However, after a simple London model analysis, the non-additive contribution should be negative, which will be discussed in more detail later. Recent studies have shown that in the calculation using the FNDMC method, and negative values appear in the results of non-additive contributions. Even with positive results, FNDMC still captures more non-additivity contributions, resulting in large differences in results than others. First of all, the appearance of this phenomenon makes the evaluation result of the methods based superpositon approximation doubtful. Second, even the CCSD(T) method, one of

high-precision molecular orbital-based electronic correlation quantum computing methods, still evaluates non-additivity as the SCF methods level which will never happened in the binding itself. Finally, although the DMC method evaluate the exact wave functions with imaginary-time evolution, the Fixed-Node (FN) method cannot be ruled out in the cancellation of the error to solve the sign problem, because, we still divide the system according to the H-bonds in the non-additive calculation. Therefore, discussing the correctness and rationality of this result is the problem to be solved in this study.

3.2 Target system

Describe the molecular system of large systems has always been the frontier of research, but how to find a suitable research object is a problem worth exploring. For MO methods such as CCSD(T), too many particle numbers make calculations difficult to achieve at the current stage of computing power. DNA molecules are an typical molecule for researching many-body systems. According to the previous study, we used four-body collections of B-DNA stacking systems as our research objects.

3.2.1 B-DNA molecules

The proposal of a double-helical structure for DNA over 60 years ago provided an eminently satisfying explanation for the heritability of genetic information. DNA not only has the function of information storage in the genetic process, but also plays the role of energy transfer. DNA molecules can have a variety of conformations. In general, B-conformation has a unique role in heredity. Chiral B-DNA consists of A-T and G-C (adenine [A], thymine [T], guanine [G] or cytosine [C]) base pairs and forms a double helix under the action of torsional stress. The nature of the double helix structure depends not only on the base pair, but also on the stacking between base pairs. The power of the stack drives the composition and changes of the chromosomes, ensuring the stable development of the double helix, which is also the key to storing genetic information on the chromosomes [86].

The target systems are shown in Fig. 3.2, ten kinds of Watson-Crick base pairs in B-DNA. The preceding work [1] provides the geometries for all the ten pairs, and we take them to be fixed. Though 'AT:AT' and 'TA:TA' are schematically identical being in a mirror image relation, but not practically identical because of the different geometries in detail. For convenience, we describe the upper and lower layer of four molecules as the schematic geometry of the systems, shown in Fig. 3.1 (b). Base fragments pairs (W,V) and (X,Y) are located within a 'strand' (a box elongating along the stacking direction, shown as a red rectangular), respectively, to form the whole four-body system specified as 'VW:XY' in the convention of the notation. The molecular layers are arranged parallel to each other and have a certain angle, which is following the right-handed helix sense. The distance a between the layers is about 3.25 Å, which is consistent with the experimental data. Moreover, between the base pairs, A-T will form two, and G-C will form three H-bonds, which will further be discussed in the discussion chapter. Because this study explores the non-additivity of stacking energy, not the stacking energy itself, there is no geometry optimization of the structure.

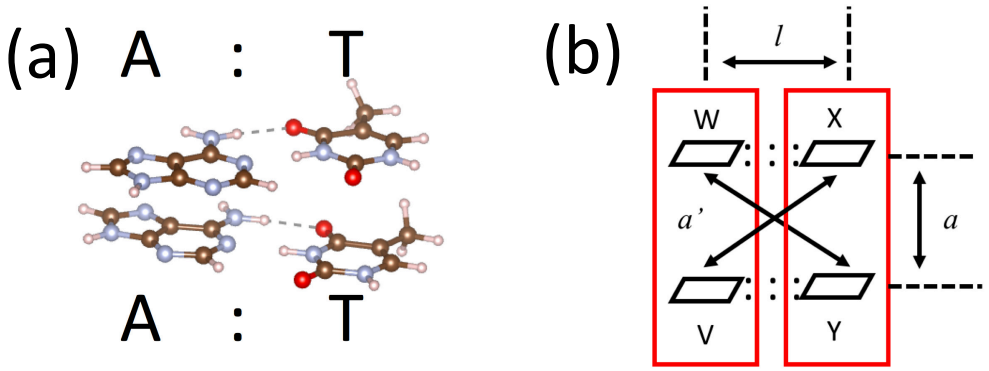

Figure 3.1: Panel (a) shows the example of the geometry for 'AA:TT' pair. The notational convention, 'VW:XY', is according to the standard one [1] in this field, as explained in the panel (b), where the bases V,W,X,Y appear in this order along ∩-shape wise.

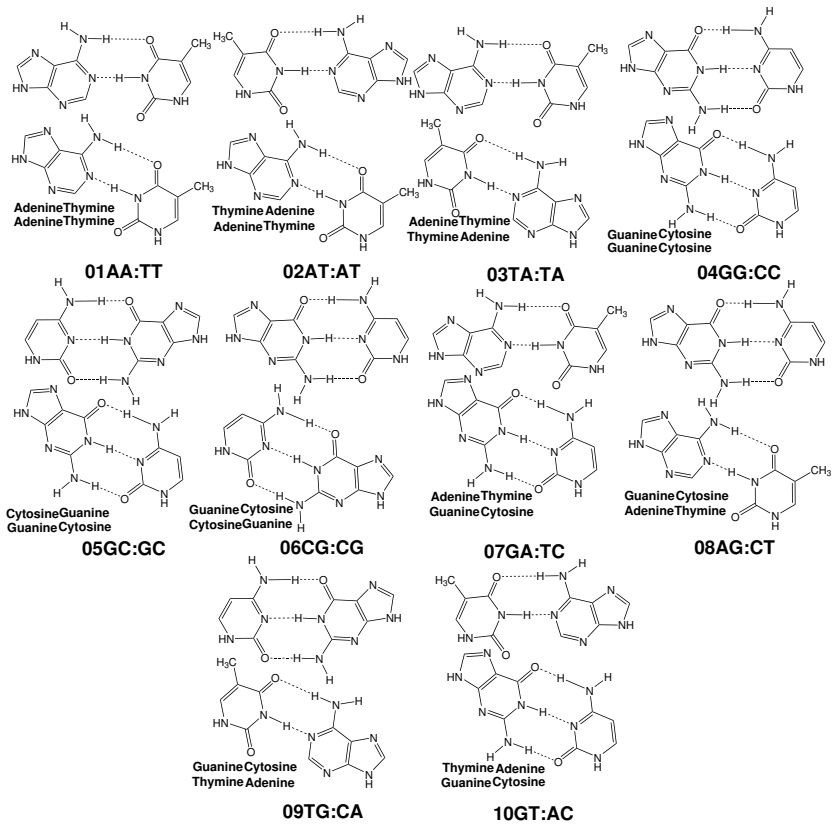

Figure 3.2: Ten kinds of the Watson-Crick base pairs in B-DNA we evaluated. Each system is composed of four kinds of bases, adenine [A], thymine [T], guanine [G], and cytosine [C] molecules.

3.3 Benchmarks of calculation methods

For our target system, this system contains a total of 58 $\sim$ 60 atoms, two layers combined by four molecules. Each layer is linked by 2 $\sim$ 3 H-bonds in the two layers. This system is an ideal system for studying the interactions within many-body systems. What we are concerned with is the stacking energy between layers in the system. Due to their many-body structures, the composition of the stacking energy is not singular. According to the discussion of the theoretical chapter, non-additiveness is always expected. We can verify our conjecture in three levels: First, the mean-field HF, and the standard DFT method, is difficult to express the dispersion interactions, due to their lack of dispersion term. At this stage, we use the HF method and the LDA method, as well as the most widely used B3LYP method in hybrid functionals, and M06-2X in Meta-GGA functional. It can be expected that the description of non-additiveness is not good because these methods completely ignore the instantaneous polarity shift of the electrons.

In the second phase, the conventional XC functionals, we adopted recently developed XC functionals such as ωB97X (B97 functional with long- and short-range corrected exchange), ωB97M-V [87], B3LYP-D3 [88] (B3LYP with an empirical dispersion correction), and CAM-B3LYP-D3 [89] (B3LYP-D3 with the long-range corrected (LC) exchange) in order to investigate not only dispersion effects but also H-bonding. Our HF and DFT simulations were carried out using Gaussian09 [5] with the same basis set and pseudopotential as FNDMC. This level of comparison is to verify empirical corrections at the level of the density functional, which is sufficiently accurate for the many-body's binding energy itself. However, we intentionally adopted them for comparison between correlation- and SCF-level non-additive contributions. So further, we look forward to accurately describing the combined energy of the larger systems, and we need to go back to the WFT method.

There is no doubt that CCSD(T)/CBS is a state-of-the-art or "gold standard" quantum chemistry method – an established protocol of capturing dispersion interactions in non-covalent systems [90]. Its applicability is, however, quite limited to small systems. Unfortunately, the complete basis set is too complicated

for 58 $\sim$ 60 atoms. CCSD(T)/CBS is not able to handle our target systems of B-DNA base-pair steps, even using the best supercomputer. The practically best possible solution [1] was to apply "CCSD(T)/CBS" to all pairwise stacking and add a many-body correction at MP2/VDZ level to the pairwise sum. Here we note that "CCSD(T)/CBS" for the pairs is not a *true* one, but an approximation such that MP2/CBS is combined with an energy difference between CCSD(T) and MP2 obtained using a small basis set. Hereafter we refer to this sort of approximation to CCSD(T)/CBS[MP2]. Very recently, Kraus *et al.* [91] have attempted to avoid the approximation of stacking interaction, such as the sum of four base-base stacking energies. However, their level of theory has not reached CCSD(T)/CBS. Parker *et al.* [3] stated that such an approximate estimate could be used for reference, but not as an absolute standard value. In the above-mentioned context, a *true* non-additivity at CCSD(T)/CBS level of theory remains unknown. We will discuss the estimation of the MP2 level basis set function in the discussion chapter.

To describe the dispersion interactions as the main ingredient in the stacking, such methods going beyond MP2 (Moller-Plesset) level treatment of electronic correlations are required [43, 92–94]. Besides, FNDMC is the most widely used method of statistical WFT methods as our final result. We hence applied the FNDMC method [43, 92–94] using an implementation, CASINO [95]. Due to the statistical approach, the time complexity of FNDMC in handling such systems is within acceptable limits. The feasibility of FNDMC applied to the system size here has well been established, achieving the accuracy to capture electronic correlations at the same level as those by CCSD(T)/CBS (complete basis set limit) [28, 30–41] . Especially for the present B-DNA case, the stacking energies evaluated by DMC are well-calibrated in detail in our preceding work [36]. More detailed information about the computational conditions and numerical results are provided separately later. Nevertheless, we need to note that there is no non-approximation in DMC for evaluating the wave functions part. But, the sign problem is unavoidable in DMC method because anti-symmetry of the fermionic system when inter-changing two particles. The weights of random walkers can be either positive or negative, and their signs are usually *a priori* unknown when we get the mean value of wave functions. In our study, we use fixed-node algorithm

approximate to solving this sign problem. This method allows a refinement of a given guiding wave function by improving its amplitudes. This approximation inevitably leads to errors while eliminating the sign problem. We cannot rule out that this is the cause of the difference in results. The error caused by this part will have a large error on the non-additivity contribution, which we will discuss in detail in the following chapter.

All DMC calculations in this study were performed by CASINO code [95] with Burukatzki-Filippi-Dolg pseudopotentials (BFD-PPs) [96]. We used the conventional N-body Slater-Jastrow form, $\Psi_T(x_1, ..., x_N) = e^{J(x_1,...,x_N)} \cdot \Psi_{AS}(x_1, ..., x_N)$, where x_i are the position of particles. The Jastrow function, $J(x_1, ..., x_N)$, used here consists of the one-body (electron-ion) and two-body (electron-electron) terms. $\Psi_{AS}(x_1, ..., x_N)$, for the many-body wave function used in DMC. Ψ_{AS} is a single Slater determinant formed by the Kohn-Sham orbitals generated by DFT-B3LYP implemented in Gaussian09 [5] code, using VTZ level basis sets. Parameters in the Jastrow functions were optimized in VMC by the variance minimization procedure [97]. In DMC, the pseudo potentials are treated under the locality approximation using T-move scheme [83, 98].

Stacking energies, $\varepsilon^{(4)}$ and $\varepsilon^{(2)}$ are evaluated using *ab initio* methods to evaluate the non-additive contribution,

$$\Delta E^{(4)} = \varepsilon^{(4)}_{\mathrm{VW:XY}} - \left(\varepsilon^{(2)}_{\mathrm{VW}} + \varepsilon^{(2)}_{\mathrm{YX}} + \varepsilon^{(2)}_{\mathrm{VX}} + \varepsilon^{(2)}_{\mathrm{YW}}\right), \tag{3.1}$$

as defined by Sponer *et al.*, [1] which does not include the contributions from H-bonds within each layer, V·Y and W·X, respectively.

3.4 Parallel computing efficiency

A computer with a large function and size is called a "supercomputer". In the past few decades, this has undoubtedly referred to as parallel computers: A computer with multiple CPUs and that can be set to handle the same problem [99]. Multiple processors handle this parallelism with multiple instruction streams, typically explicitly scheduled by the user. In order to parallelize the program

size, it is necessary to design a parallel strategy. Allocate at least a portion of the data and tasks between different N_p processors without introducing a large amount of communication between different CPUs. For different calculations, the parallel strategy usually depends on the way it is calculated, and the specific parallel optimization method is adopted. However, it could be certain that the statistics-based algorithm accelerates in parallel much more than the numerical iterative method, although the latter can also rely to some extent on the power of parallelism.

3.4.1 SCF calculations

Due to the iterative calculation of SCF, it converges to the target energy. This approach means that each iteration calculation relies entirely on the results of the last calculation. This leads to the calculation of the time direction must be linear in this algorithm. But fortunately we can decompose the calculations on the spatial latitude to accelerate the DFT calculations using parallel algorithms. For DFT calculations of different basis sets, we can perform different forms of parallel methods based on the basis set, such as the traditional diagonalisation (TD) scheme or the orbital transformation (OT) method. For example, in the DFT calculation of the plane wave basis set widely used in the periodic operation, we can perform parallel calculation on the wave function in the frequency domain by performing Fourier transform on the plane wave. However, this does not apply to the orbital wave function of the Gaussian basis set function configuration of the center of the nucleus widely used in molecular modeling.

In general, the mapping state of the task decomposition at the processors level can be divided into one-dimensional(1D) or two-dimensional(2D) conditions. Most of the data, including the fragments of molecules, atoms, and their coordinates, the grids and matrices for sampling properties are replicated. These 1D data are distributed in each processor and retrieved according to the spatial order of the atoms. Such parallel level calculations are generally used to calculate the state of the atom itself, such as the charge density, because such variables only have spatial coordinates.

On the other hand, the task can be easily decomposed according to the 1D distribution. The storage space required for a complete 1D distribution is small, so each piece of data can be replicated into a distributed process to reduce the duplication of data transfer and memory allocation. However, it is necessary to establish a corresponding 2D mapping between atoms, such as the interaction between the atoms i and j. So for a 1D atomic distribution, it also needs to locate the positions of the electron pair for i and j. [100]

However, in any case, the parallel basis of SCF is still to decompose the linear matrix, including the diagonalization process and the linear solution. In the process of matrix decomposition, solution, and merging, a large amount of communication overhead is generated for data exchange. According to Amdahl's law [101, 102], SCF is algorithmically impossible to achieve linear scaling. When the processing system is too large, it will generate large-scale parallel communication overhead. For such large-scale parallel computing, the speedup will decrease as the number of cores increases until the acceleration limit is reached.

We tested the parallel efficiency of a widely used quantum computing software gaussian in shared-memory parallel computers. Benchmark uses the DFT calculation of B3LYP-D3. The target system is AA:TT system for parallel computing. The wall times and the CPU times divided by the number of cores are shown as Fig. 3.3, by comparing 6, 12, 24, 36, 48 cores with 256GB shared memory. As shown, the more parallel cores are counted for DFT calculations, the lower the efficiency of parallelism. An important factor that causes this is the need for constant communication of spatial lattice determinants.

3.4.2 Statistical evaluation calcualtions

Since the QMC method is random, a very efficient parallel implementation of the QMC algorithm can be generated. In a VMC, independent runs can be performed by different processors, and then the data from each CPU can be averaged to arrive at a final result. According to the definition of the speedup ratio, the average time for each CPU to complete the operation divided by the maximum single CPU time, the VMC's efficiency can reach ~ 1.

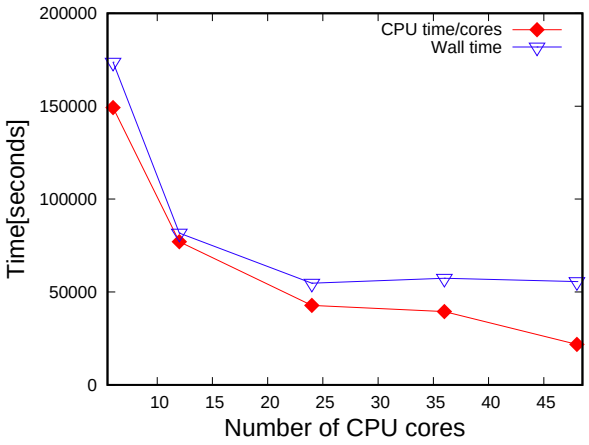

Figure 3.3: The figure shows the total calculation time and CPU time divided by the number of cores. The benchmark's calculation time for Gaussian in the 6, 12, 24, 36, 48 cores states. This calculation selects the AA:TT system in the target system for calculation. In theory, the more parallel cores, the total time should be approximately equal to the CPU time divided by the number of cores N_p. Its curve should be rendered with $\sim O(1/N_p)$. It can be seen from the figure that for DFT calculations, the more parallel cores, the lower the efficiency of parallelism. The important factor leading to this reason is the need for continuous communication of spatial lattice determinants.

The DMC method needs to evolve according to the imaginary-time τ, and each imaginary-time state includes different configurations P. Parallel DMC calculations are performed by allocating a configured initial number P in the available processors, each CPU n performing an imaginary-time evolution of its assigned Pn configuration. According to the description of the DMC algorithm, the number of configurations in the branch DMC is different; therefore, the time τ required for the node n to complete the iteration is proportional to the number of configurations $P_n(\tau)$ on the node n. And the parallel efficiency of iteration τ is approximately $\sum_{n=1}^{N_p} P_n(\tau) / \left(N_p \max\left[P_n(\tau)\right]\right)$. Therefore, if all nodes have the same number of configurations, the operational efficiency can be maximized by ~ 1. Therefore, it is common practice to redistribute configurations between nodes after each iteration. [103]

Notice that a large amount of inter-node communication is still required in QMC calculation, but the communication cost of the statistical method is much lower than that of continuous communication in SCF calculation. Although the easiest way is to separate the branches from the redistribution, it is best to execute them at the same time, transfer the configuration between the nodes, and then copy the configuration on the receiving node. Because copying information between processors is more expensive than copying information within the same processor. In most calculations, this aspect of the redistribution process is not very important, because the branching factor is usually very close to the unit, so multiplicity is also true. However, this can be very important for calculations with large time steps or poor fluctuation functions.

3.5 Conclusion

In this chapter, we discussed different electronic state calculation methods. There are several issues we need to verify for the target system. First, the standard DFT method has been proven to fail in the calculation of dispersive forces. So whether the correction of the dispersion interactions and the long-range force functional can truly describe the dispersion interactions. In addition, high-precision calculation methods such as CCSD(T) can achieve the calculation accuracy of CCSD(T)

when dealing with large-scale systems. These problems all mean that the QMC method is essential to describe the electronic association correctly and to calculate the large system. We designed the calculation of the non-additional contribution to the stacking energy of the target system. For non-additive results, CCSD(T) can only get the same result as DFT. QMC, on the other hand, gets negative results that are difficult to obtain by other methods, and the non-additive contribution of positive values is also significantly higher than other methods. So how to interpret these results is also an important issue. Based on this problem, we have selected 10 B-DNA systems to compare different calculation methods, and detailed target system characteristics are also described. Further, we have also discussed in detail the choice of different calculation methods. Based on our problem, we chose the simplest HF method and the standard DFT methods as a standard for comparison. For the WFT methods, we chose the high-precision CCSD(T)/CBS[MP2] and FNDMC to verify our conjecture. Accordingly, the DFT methods with increased dispersion interactions correction were also tested. Finally, we discuss the computational efficiency of SCF methods and statistical methods. The computational power of the SCF method has unparalleled advantages so that it can handle larger-scale calculations. Nevertheless, by using parallel computers, statistical methods can also handle large-scale calculations, which makes precision and larger system calculations possible.

Chapter 4

Results

4.1 Result of B-DNA systems

The target systems are shown in Fig. 3.1, ten kinds of the Watson-Crick base pairs in B-DNA. The preceding work [1] provides the geometries for all the ten pairs, and we take them to be fixed. The distance between the molecules is around 3.25 Å, which is very close to the experimental data. The molecular layers also have a certain angle between them and conform to the right-handed Helix sense of B-DNA. Though 'AT:AT' and 'TA:TA' are schematically identical being in a mirror image relation, but not practically identical because of the different geometries in detail.

4.1.1 Stacking energies $\varepsilon^{(4)}$

We shall start with the four-body stacking energies ($\varepsilon^{(4)}$) of the B-DNA base-pair steps. Fig. 4.1 (a) and (b) are the $\varepsilon^{(4)}$ values obtained from wave function-based and DFT-based methods, respectively, together with CCSD(T)/CBS[MP2] as reference. Within the wave function methods (Fig. 4.1 (a)), we can see that HF fails to describe the stacking properly because it does not include electron correlations at all by nature. An appropriate description of stacking requires theory more than MP2 level [92, 94]. Overall, the correlated methods agree with each other. Looking closely at the trend, CCSD(T)/CBS[MP2] was found

to overbind compared with the other approaches, LMP2, SAPT, and FNDMC. This can be attributed to two practical approximations adopted in the CCSD(T) reference [1]: Since the B-DNA system is too large to be computed by CCSD(T) with larger basis sets (cc-pVTZ and cc-pVQZ), the *true* CCSD(T)/CBS is not available in any literature so far. Instead, two approximations were applied to estimate "CCSD(T)/CBS" [1]. Firstly, all the pairwise base-base terms ($\varepsilon^{(2)}$ in Eq. (3.1)), were computed by MP2/CBS plus an energy difference between CCSD(T) and MP2 with a common small basis set (6-31G*(0.25)). Secondly, the non-additive contribution ($\Delta\varepsilon^{(4)}$) was evaluated at RI-MP2/aug-cc-pVDZ level. In this sense we refer the CCSD(T) to "CCSD(T)/CBS[MP2]". From the above facts, we can infer that "CCSD(T)/CBS[MP2]" overbinds, similar to MP2. Note that both SAPT and DF-LMP2 are known to correct the overbinding trend in MP2 [2]. Accordingly, 'the overbinding in CCSD(T)/CBS[MP2]' can be also supported by the fact that the magnitudes of stacking energies in SAPT and DF-LMP2 are smaller than those in CCSD(T)/CBS[MP2], as can be seen in Fig. 4.1 (a) and Tab. 4.1. It was found that FNDMC deviates from the other correlated methods depending on base-base pair steps (especially TA:TA and AG:CT), which is closely related to a significant difference in non-additive contributions between the methodologies, as described later.

As for the DFT methods (Fig. 4.1 (b)), we see that only B3LYP cannot properly describe the stacking, which is consistent with its well-known deficiency in describing dispersion interactions [104]. Which are also shown in Tab. 4.2. On the other hand, recently developed dispersion methods within DFT reproduce the stacking for all the steps [88], giving almost the same trend as CCSD(T)/CBS[MP2]. When comparing with CCSD(T)/CBS[MP2], B3LYP-D3 gives the wiggling stacking energies; CAM-B3LYP-D3 slightly overbinds overall; M06-2X and ωB97X both underestimate the stacking energies for all the steps, which has also been demonstrated for other non-covalent systems [37]. Note that the stacking described by LDA is known to be artificial [28, 32, 36–38, 40, 105, 106]. Tab. 4.1 and Tab. 4.2 are only shown the 4-body stacking energies. Moreover, for evaluating the non-additivity contributions, we also need to calculate 2-body stacking energies, including intra- (W,V and X,Y)and interstrand (W,Y

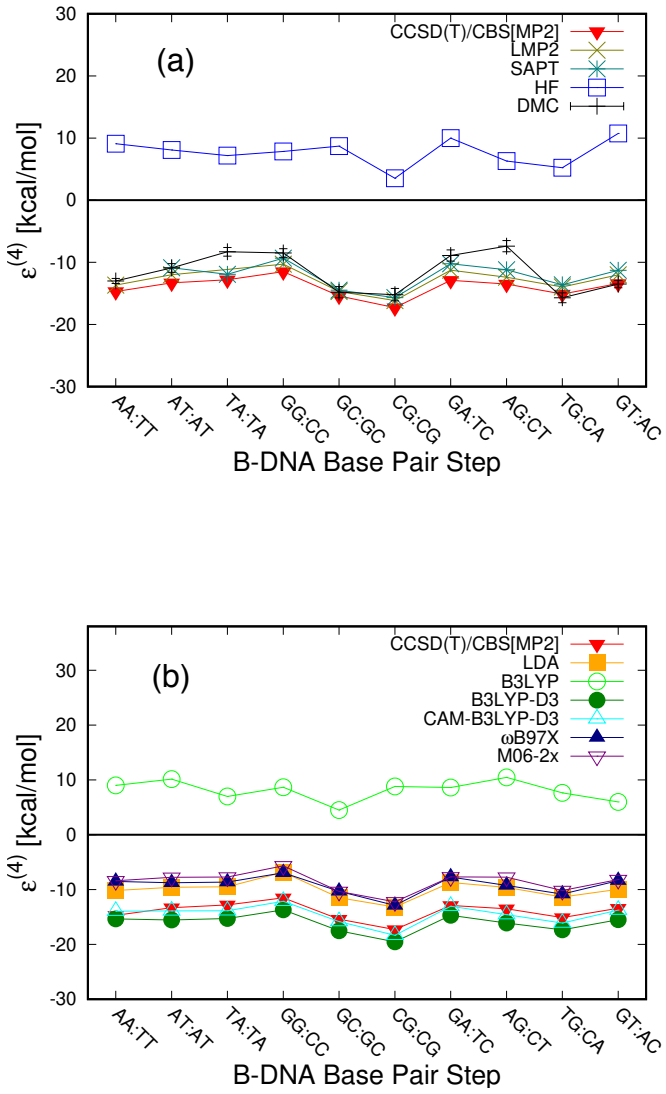

Figure 4.1: Four-body stacking energies, $\varepsilon^{(4)}$ [kcal/mol], for the B-DNA base-pair steps evaluated by various methods. The negative values correspond to the binding, and hence we see that only B3LYP cannot correctly describe the binding. CCSD(T) values were taken from a previous work [1].

and X,V) stacking. The above stacking energy data is put in Appendix I.

Table 4.1: Stacking energies ($\varepsilon^{(4)}$) of B-DNA base-pair steps evaluated from wave function-based methods. All the energies are given in kcal/mol.

Pairs	CCSD(T)/CBS[MP2][1]	LMP2[2]	SAPT[3]	HF[4]	FNDMC[4]
AA:TT	−14.7	−13.66	NA	9.09	−13.0(4)
AT:AT	−13.3	−11.99	−10.87	8.09	−10.9(7)
TA:TA	−12.8	NA	−11.92	7.18	−8.3(7)
GG:CC	−11.5	−10.29	−9.32	7.85	−8.5(7)
GC:GC	−15.4	−14.70	−14.48	8.71	−14.8(9)
CG:CG	−17.3	−16.15	−15.69	3.53	−15.2(0)
GA:TC	−12.9	−11.26	−10.22	10.00	−8.9(9)
AG:CT	−13.5	−12.39	−11.20	6.31	−7.4(9)
TG:CA	−15.1	−13.96	−13.63	5.22	−15.7(8)
GT:AC	−13.4	−12.01	−11.29	10.74	−13.5(6)

Table 4.2: Stacking energies ($\varepsilon^{(4)}$) of B-DNA base-pair steps evaluated from DFT-based methods. All the energies are given in kcal/mol.

Pairs	LDA[1]	B3LYP[1]	B3LYP-D3[1]	CAM-B3LYP-D3[1]	ωB97X[1]	M06-2X[1]
AA:TT	−10.17	9.00	−12.04	−13.94	−8.52	−8.39
AT:AT	−9.60	10.14	−11.19	−13.90	−8.76	−7.76
TA:TA	−9.50	6.97	−12.82	−13.88	−8.63	−7.72
GG:CC	−6.83	8.65	−12.42	−12.14	−6.93	−5.67
GC:GC	−11.48	4.50	−15.97	−15.85	−10.31	−10.33
CG:CG	−13.20	8.80	−13.79	−18.35	−12.89	−12.25
GA:TC	−8.71	8.60	−10.35	−13.05	−7.71	−7.68
AG:CT	−9.62	10.47	−11.99	−14.58	−9.24	−7.74
TG:CA	−11.37	7.64	−12.02	−16.11	−10.83	−10.18
GT:AC	−9.96	5.98	−13.94	−13.76	−8.36	−8.17

[1]Ref. [1]
[2]DF-LMP2+$\Delta(T)$ in Ref. [2]
[3]SAPT0/jaDZl in Ref. [3]
[4]Present study
[1]Present study

4.1.2 Non-additivity contributions $\Delta\varepsilon^{(4)}$

The present study adopts CCSD(T)/CBS[MP2] results due to Šponer *et al* [1] as reference, though they are *not* "true" CCSD(T)/CBS as described in the previous section "System and methods". Here we show not only non-additive contributions $\Delta\varepsilon^{(4)}$, but also stacking energies $\varepsilon^{(4)}$ themselves similar to conventional researches. We will see that FNDMC gives rise to a striking dependence of $\Delta\varepsilon^{(4)}$ on base-pair steps, compared to the other *ab initio* methods including CCSD(T)/CBS[MP2] as well as DFT, while there is not any novel findings in their $\varepsilon^{(4)}$ evaluations.

The non-additivity in the interactions is obviously expected in inter-molecular bindings due to the induced polarizations by the quantum fluctuations, such as vdW forces. Since the binding itself has been a great challenge for ab initio methods to describe and reproduce, the non-additivity as the further difficulty on top of it has been put off from the major interest, being not well analyzed so far. The target systems are ten kinds of the Watson-Crick base pairs in B-DNA. The preceding work [1] provides the geometries for all the ten pairs, and we take them to be fixed. Though 'AT:AT' and 'TA:TA' are schematically identical being in a mirror image relation, but not practically identical because of the different geometries in detail. Stacking energies, $\varepsilon^{(4)}$ and $\varepsilon^{(2)}$ are evaluated using *ab initio* methods to evaluate the non-additive contribution as Eq. 3.1 which does not include the contributions from H-bonds within each layer respectively. Non-additive contributions $\Delta\varepsilon^{(4)}$ evaluated from various methods are shown in Fig. 4.2. We found FNDMC giving a wiggling behavior of $\Delta\varepsilon^{(4)}$, compared to all the other methods. This remarkable sign alternation found in FNDMC is a central issue in the present study.

We first remark that the non-additive contributions do not necessarily arise from the electron correlations but always appear as non-linear processes inside a many-body system. In terms of the perturbation expansion based on SAPT, interaction energy is decomposed into physically meaningful components: electrostatic, induction, exchange, and dispersion terms [92, 94]. Even at the HF level of theory, the 'exchange' and 'induction' parts of the non-additivity occur [92,94], which we refer to 'SCF-level non-additivity' hereafter. The behavior of HF in

Fig. 4.2 can be an appropriate reference to the 'SCF-level non-additive' contribution because HF is incapable of describing the dispersion by nature. Except for FNDMC, CCSD(T)/CBS[MP2] as well as all the other DFT methods exhibit almost the same $\Delta\varepsilon^{(4)}$ as HF; their magnitudes of $\Delta\varepsilon^{(4)}$ are slightly increased or decreased from those of HF owing to their balance between exchange and correlation (as described later), but their differences are quite small (less than 1 kcal/mol). Presumably, the non-additive contributions appearing inn all the DFTs and even CCSD(T)/CBS[MP2] may be regarded as being 'SCF-level' ones or hardly describing 'dispersion-level/correlation-level' ones. On the other hand, we may imply that FNDMC describes more dispersion/correlation contributions than the other methods.

Looking closely at individual methods, it was found from Fig. 4.2 (a) that the behavior of CCSD(T)/CBS[MP2] is the most striking: The electron correlation described by CCSD(T)/CBS[MP2] hardly makes any corrections to the SCF-level non-additivity. It means that the 'correlation-level non-additivity' cannot be well captured by CCSD(T)/CBS[MP2]. Note that the many-body contributions in CCSD(T)/CBS[MP2] are identical to those obtained at RI-MNP2/aug-cc-pVDZ level. The deficiency of correlation-level non-additivity in MP2 is attributed to the fact that MP2 can take into account only the two-electron excitation process. In order to describe the correlation-level non-additivity in a system consisting of four subsystems, MP2 level of theory should be adopted at least, where four one-electron excitations simultaneously appear in each of the subsystems [92,94]. Therefore, the CCSD(T) method with the practical approximation considered in the literature [1] can be regarded as being almost the same as HF and thus, poor at capturing the correlation-level non-additivity. Unfortunately, a *true* behavior of the non-additivity at CCSD(T)/CBS level is still unknown.

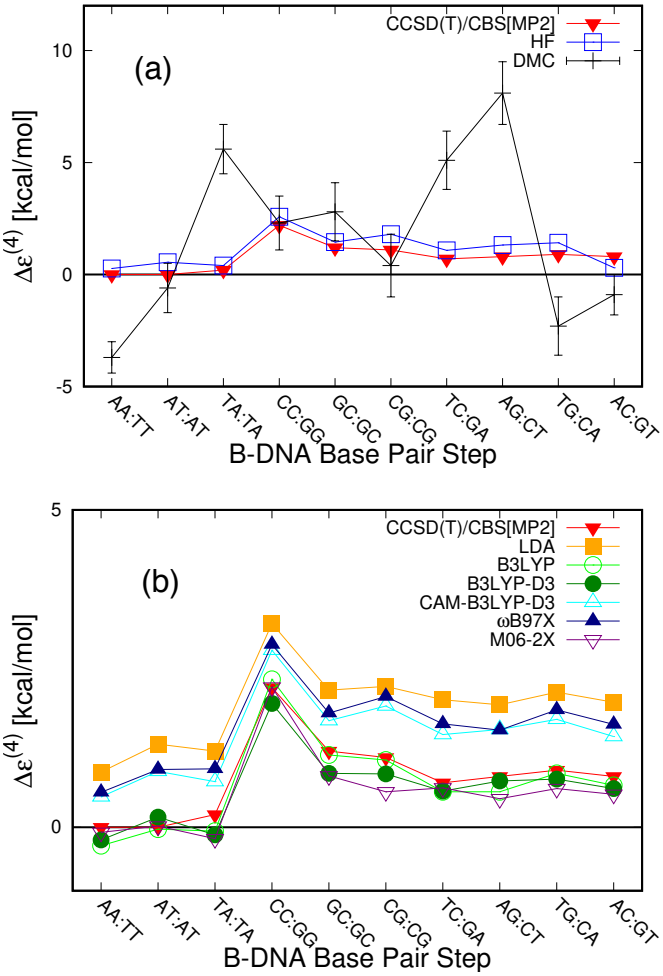

Figure 4.2: Non-additive contribution, $\Delta E^{(4)}$ [kcal/mol], evaluated by various methods. DF-LMP2 [2] and SAPT [3] appearing in Fig. 4.1 are not shown here because their non-additive contributions are not available. For CCSD(T), the data is taken from the preceding work. [1] Unlike stacking energies (Fig. 4.1), CCSD(T) agrees with both HF and B3LYP, while it is far from FNMDC. Plausible discussions for this are given in the text.

Table 4.3: Non-additive contributions ($\Delta\varepsilon^{(4)}$) of B-DNA base-pair steps evaluated from wave function-based methods. The definition of $\Delta\varepsilon^{(4)}$ is given in Eq. 3.1 of the main text. All the energies are given in kcal/mol.

Pairs	CCSD(T)/CBS[MP2][1]	HF[2]	FNDMC[2]
AA:TT	0.0	0.27	−3.7(7)
AT:AT	0.0	0.55	−0.6(11)
TA:TA	+0.2	0.40	+5.6(11)
GG:CC	+2.2	2.58	+2.3(12)
GC:GC	+1.2	1.45	+2.8(13)
CG:CG	+1.1	1.80	+0.4(14)
GA:TC	+0.7	1.08	+5.1(13)
AG:CT	+0.8	1.32	+8.1(14)
TG:CA	+0.9	1.42	−2.3(13)
GT:AC	+0.8	0.30	−0.9(9)

A comparison between B3LYP and B3LYP-D3 gives the most intriguing insight into the non-additivity within the framework of DFT. While the empirical dispersion correction D3 significantly improves the stacking itself (Fig. 4.1 (b)), it hardly modifies its non-additivity from B3LYP (Fig. 4.2(b)). It can be attributed to the fact that dispersion corrections based on D3 or the likes of vdW-XC are additionally made on the original DFT/SCF energies and, thus, never deform their wave functions. From the viewpoint of many-body theory, the deformation of wave functions is the origin of dispersion interactions and hence, essential to the non-additivity. Interestingly enough, as shown in our central results in Fig. 4.2 and Tab. 4.3 this is not true as clarified by the present work, applying DMC to evaluate stacking energies of B-DNA base pairs [1–3, 37, 93, 107, 108]. While the conventionally available techniques, including CCSD(T)/CBS, predict tiny ($\sim$ several kcal/mol), positive definite contributions, the DMC predicts much larger non-additive contributions [even being the same magnitude as those of interactions themselves ($\sim$ 10 kcal/mol)], with those signs alternating from positive to negative depending on the base pairs [it means that the binding of the stacking is reduced (positive) or enhanced (negative) by the non-additivity].

[1] Ref. [1]

[2] Present study

We can also note that the results of FNDMC are more unstable in the combination of A-T base pair appearances compared to systems consisting only of G-C base pair. Especially the difference in the arrangement of the directions, it is easy to associate the relationship between them. We make a conjecture as "whether the difference between these positive and negative signs comes from a change in the polarity direction". Alternatively, we suspect that the whole of the base pairs acts as an inline polar molecule, which leads to a difference in the results of non-additive contributions. The key to the direction of the A-T base and the G-C base is the role of the molecular π ring, and the second is the arrangement of H-bonds. No clear way has been found to quantitatively analyze the increasing effects between π rings, although the dispersive force between π rings can lead to many unique properties. Moreover, there is an angle between the molecules of the B-DNA, which makes the π ring have no upper and lower correspondence. In contrast, the analysis of H-bonds is more explicit. In summary, we will try to explain the result by establishing a model based on H-bond direction later.

Table 4.4: Non-additive contributions ($\Delta\varepsilon^{(4)}$) of B-DNA base-pair steps evaluated from DFT-based methods.

Pairs	LDA[1]	B3LYP[1]	B3LYP-D3[1]	CAM-B3LYP-D3[1]	ωB97X[1]	M06-2X[1]
AA:TT	0.87	−0.21	−0.20	0.48	0.55	−0.08
AT:AT	1.31	0.16	0.16	0.88	0.91	0.02
TA:TA	1.20	−0.12	−0.12	0.72	0.92	−0.18
GG:CC	3.21	1.96	1.95	2.79	2.88	2.21
GC:GC	2.16	0.83	0.85	1.68	1.80	0.81
CG:CG	2.22	0.83	0.84	1.91	2.06	0.56
GA:TC	2.01	0.54	0.57	1.46	1.63	0.62
AG:CT	1.93	0.73	0.73	1.54	1.53	0.45
TG:CA	2.13	0.75	0.76	1.70	1.85	0.61
GT:AC	1.97	0.60	0.61	1.43	1.62	0.52

[1]Present study

4.2 Conclusion

This chapter presents and describes the calculations based on the B-DNA target systems. We calculated the stacking energy in the ten target systems described in the previous chapter. Various methods are employed, including WFT calculations and DFT calculations. The WFT calculation includes the most basic HF calculations, calculations based on perturbation theory, and high-precision CCSD(T) calculations. Moreover, the most important work of this study is FNDMC calculation. In DFT calculations, there is no doubt that the standard DFT calculations, especially the results of widely used hybrid functionals, are used as a negative example for comparison. In comparison, the DFT method with the dispersion-correction is also calculated, and the results are presented in this chapter. Base on the result of the stacking energy, non-additive contributions is also calculated according to the above definition. In this result, we can find that CCSD(T) is at DFT-level. Only the results of the DMC have a negative result, and the overall trend is not at the same level as the other results. We will discuss these results further.

Chapter 5

Discussions

5.1 London model analysis

Although it is difficult to estimate how much non-additivity contributions should be captured by the *ab initio* methods before actual calculations. And none of the conventional *ab initio* methods can describe the negative sign of the non-additive contribution. But we can estimate its sign with a simple model analysis upon London theory, which gives a quick estimation of the contribution being surely negative: Fig. 3.1 (b) shows the schematic geometry of the systems. Base fragments pairs (W,V) and (X,Y) are located within a 'strand', respectively, to form the whole four-body system specified as 'VW:XY' in the convention of the notation. In London theory, a stacking energy between the upper and lower layers scales as $\varepsilon \sim \alpha^{(upper)} \cdot \alpha^{(lower)}$, where $\alpha^{(upper/lower)}$ denotes the polarizability of each layer. In the model defined by London theory, the molecules are attracted to each other by transient dipoles. The instantaneous dipoles caused by the instantaneous relative displacement occurs between the electrons and the nucleus, due to the continuous movement of electrons and the constant vibration of the atomic nucleus. The dispersion force was calculated by the equation 5.1.

$$E_L = -\frac{3}{2}\alpha_A\alpha_B\frac{I_A I_B}{I_A + I_B}{r_{AB}}^{-6}\,, \tag{5.1}$$

For a two layers system, A and B represent two different layer. Where α is polarizability [$Bohr^3$], I is ionization energy [$Hartree$], and r_{AB} is intermolecular distance [Å] The units are given in atomic units (au), and converting to kcal/mol requires multiplication by constant 627.509. Since the polarizability scales to the molecular weight, $\alpha^{(upper/lower)}$ in the total (four-body) system gets doubled from that of the partial (two-body) system, giving a rough estimate of the stacking energy for the whole (four-body) system as $\varepsilon^{(4)} = 2 \times 2 \cdot \varepsilon^{(2)} < 0$, where $\varepsilon^{(2)}$ denotes the stacking energy for a partial (two-body) system (of course, there are other dependence of ε such as on ionic energies, geometries *etc.*, they don't affect so much in the discussion). The estimate then gives the non-additivity being $\Delta E = \varepsilon^{(4)} - 4 \times \varepsilon^{(2)} = 0$. That would be true for the limit, $l \to 0$ [$(a'/a) \to 1$], but for the practical cases, $(a'/a) > 1$, we can ignore the inter-strand interactions [those between 'W and Y' and 'X and V'] due to $(1/a^6) \gg (1/a'^6)$ (the latter is actually in between 2%~9% of the former with $(a'/a) = 1.5 \sim 1.9$), concluding the negative non-additivity, $\Delta E = \varepsilon^{(4)} - 2 \times \varepsilon^{(2)} = 2 \times \varepsilon^{(2)} < 0$. London theory hence supports the non-additivity being negative (the stacking energy *enhanced* by the non-additivity).

5.2 Hydrogen bonds

A-T base pairs are held together by two H-bonds, while G-C base pairs have three H-bonds. The H-bonds in the middle of the base pairs are obviously asymmetrical. This asymmetry causes the bases to not stay on the same plane, but at an angle and bend in practice. This means that the two-layer parallel molecular structure is not an optimized geometric structure, and the power of this distortion tends to change the structure. In contrast, the A-T base pair lacks a H-bond on one side, so this polarity change makes its structure more unstable. This trend is also discussed in the results below to explain the inconsistency results in DMC. The electrostatic potential at the vdW surface of the DNA base pairs is shown as Fig. 5.1.

As described in the result chapter, we presume a main origin of the positive non-additivity coming from the H-bonds bridging between the bases horizontally. Looking back the Fig. 3.1, one notices that all the ten cases include only two kinds

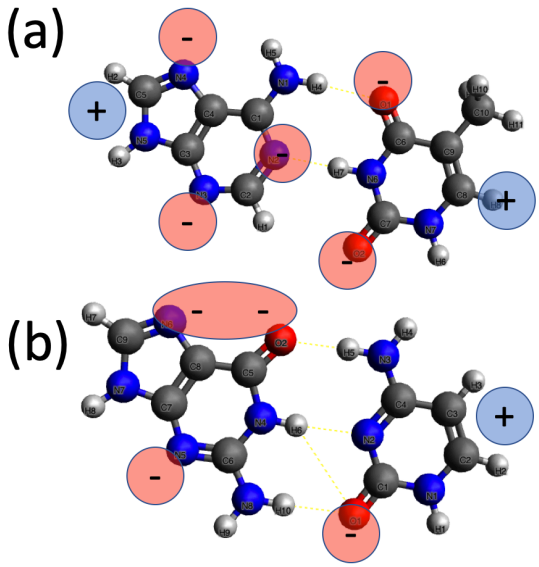

Figure 5.1: The non-covalent interaction between base molecules gives the entire molecule a local polarity, the electrostatic potential at the vdW surface of the DNA base pairs are shown. Regions of positive (blue) and negative (red) charge density are marked. (a) A-T; (b) G-C. [4]

of H-bonds (horizontal), *i.e.*, between 'A-T' and 'G-C'. As shown in Fig. 5.3 (b), these bridging bonds can be sorted into *a* (N-H...O) or *b* (N-H...N), and further labelled such as a^+, b^- *etc.* based on the direction of the charge transfer expected due to the negativity. By writing down the alignments of the 'polarized bonds' for each base pair as given in Table. 5.1, we could extract some interesting trend as follows. Since the '±' denotes the direction of the 'polarization', we can sort the cases into 'P' (parallel) or 'A' (anti-parallel), based on the relation between the 'polarization' in upper and lower layer. Then, we can see that the labelling 'P' or 'A' is fairly in accordance with the sign of the non-additivity observed in Fig. 4.2.

To examine whether such 'dipole' directions could really dominate the trend or not, we evaluated the Mulliken charge analysis on the bridging position using DFT B3LYP-D3, as shown in Fig. 5.3 as well as Fig.7∼16 (in Appendix II). From the analysis, we immediately notice that, 'I. Large negative charges concentrate on Oxygen site', and 'II. On the bond, *b*/N...N-H, little dipole is found'. We may therefore neglect *b*-bond to consider as a dipole contribution, only taking the bond, *a*/O...N-H, into account [even when we consider *b* as well, the consequence doesn't change because when *a* is P(A), *b* is also P(A)].

5.2.1 Sign alternation in FNDMC

Except for B3LYP(-D3) and M06-2X for AA:TT and TA:TA, most of the DFT functionals give positive values of $\Delta\varepsilon^{(4)}$, similar to HF. We then may conclude that the SCF-level non-additivity is mostly positive definite as the previous section. *i.e.*, CCSD(T)/CBS[MP2] gives a negligibly small dispersion-level non-additivity. In contrast, FNDMC values of $\Delta\varepsilon^{(4)}$ alternate their signs depending on the corresponding base-pair steps. This means that the dispersion-level non-additivity in FNDMC for some base-pair steps is large enough to change the signs from the SCF-level non-additivity. According to a simple model analysis based on the London theory, the dispersion contribution to non-additivity was found to be negative definite. Thus, we may conclude that the negative non-additivity in FNDMC for some base-pair steps (AA:TT, TG:CA, and AC:GT) can be attributed to the dispersion contribution. In summary, the dispersion contributions to non-additivity have the potential for changing the sign of non-additivity, depending on

Table 5.1: The bondings located from back to front are shown from left to right in a line. Two lines for each pair corresponds to upper and lower layers of a base step. The sign appearing in the left-most column, *e.g.*, '-/01aatt', means if the non-additivity is negative or positive. For 02atat, we put '-+' because it is 'zero' within the errorbar. 'P/A' appearing in the right-most column means 'parallel' or 'anti-parallel' based on the accordance in the sign ordering in each layer. The pairs, 04∼06, are not considered to be put P/A because these pairs show only the SCF-level non-additivity.

Pairs	upper layer lower layer			'parallel' or 'anti-parallel'
-/01aatt	$a-$	$b+$		(P)
	$a-$	$b+$		
-+/02atat	$a+$	$b-$		(A)
	$a-$	$b+$		
+/03tata	$a-$	$b+$		(A)
	$a+$	$b-$		
/04ggcc	$a+$	$b-$	$a-$	
	$a+$	$b-$	$a-$	
/05gcgc	$a-$	$b+$	$a+$	
	$a+$	$b-$	$a-$	
/06cgcg	$a+$	$b-$	$a-$	
	$a-$	$b+$	$a+$	
+/07gatc	$a-$	$b+$		(A)
	$a+$	$b-$	$a-$	
+/08agct	$a+$	$b-$	$a-$	(A)
	$a-$	$b+$		
-/09tgca	$a+$	$b-$	$a-$	(P)
	$a+$	$b-$		
-/10gtac	$a+$	$b-$		(P)
	$a+$	$b-$	$a-$	

base-pair steps.

If we accept the negative non-additivity in FNDMC in accordance with the London theory, then another doubt arises about why FNDMC also gives more positive non-additivity, depending on the base-pair steps. Comparing FNDMC with HF, FNDMC was found to give almost the same $\Delta\varepsilon^{(4)}$ as HF (within errobar)

for GG:CC, GC:GC, and CG:CG (Fig. 4.2). In contrast, the more positive $\Delta\varepsilon^{(4)}$ deviating from the SCF-level non-additivity appears in TA:TA, GA:TC, and AG:CT, which will be then investigated from the viewpoint of stacking energies.

Fig. 5.2 shows one four-body and four two-body stacking energies given in Eq. (3.1) for ten unique B-DNA base-pair steps. 's' and 'i' appearing in the labels for the horizontal axis in Fig. 5.2 indicate intra- and interstrand stacking. *e.g.*, A//Ai means Adenine molecules are in different strands, and A//As means Adenine molecules are up and down position within strand. It is evident that positive non-additivity correlates with weaker four-body stacking (red bar), which is also noted in Fig. 4.1 that the base-pair steps with the positive non-additivity exhibit the weaker stacking described by FNDMC than by the other *ab initio* methods.

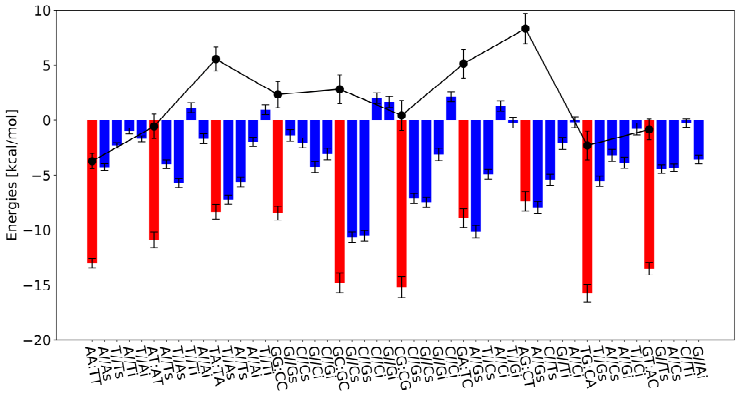

Figure 5.2: Non-additive contributions (black points) decomposed into 4-body (red bars) and 2-body (blue bars)stacking energies evaluated by DMC [kcal/mol]. 's' and 'i' appearing in the labels for the horizontal axis indicate intra- and interstrand stacking.

The weaker four-body staking is identified as being the origin of the positive non-additivity. Furthermore, it was found from Fig. 5.3 (a) and (b) that the weaker four-body stacking can be caused by the bridging bond between the Watson-Crick base pair: Although the bridging formed by the H-bonding partly leads to a

stronger stacking in the horizontal direction [109,110], it simultaneously weakens the staking owing to the vertical repulsion between the bridges at different layers (given in Appendix II in detail). Both the contributions would cancel each other out and thus, giving rise to the overall 'weaker stacking'. This cancellation causes the positive non-additivity depending on the base-pair steps. This factor was not taken into account in our simple London model analysis.

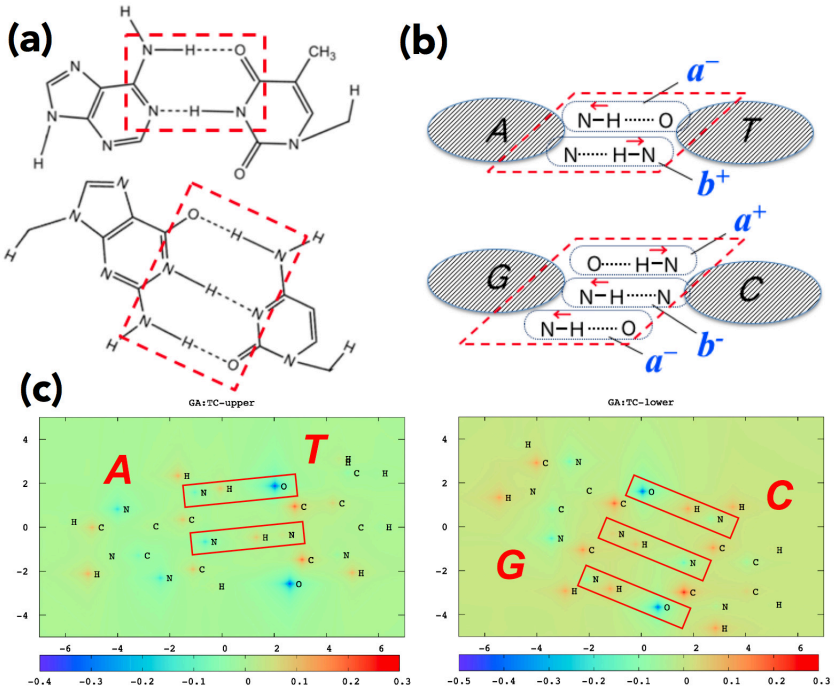

Figure 5.3: H-bonds for GA:TC base pair, shown inside the red broken lines [left panel(a)], and its schematic picture [panel(b)]. Small red arrows put on the N-H bonding in the right panel mean the charge transfer due to the negativity. Bridging bonds can be sorted into *a* (N-H...O) or *b* (N-H...N), and further labelled such as a^+, b^- *etc.*, based on the direction of the charge transfer. Panel(c) shows the Mulliken charge analysis for the upper and lower layers. Blue and red indicate the negative and positive charge values, respectively.

To estimate the contributions quantitatively, we first evaluated the Mulliken charge that appeared in the bridging location, as shown in Fig. 5.3 (c). Based

on the charge, we then evaluated the Madelung repulsion interaction, shown in Fig. 5.4, getting +5∼10 kcal/mol per bond. Since the typical range of the H-bonding energy is known to be less than 5∼6kcal/mol [111], it is likely to result in a positive contribution to non-additivity, thus making the stacking weaker.

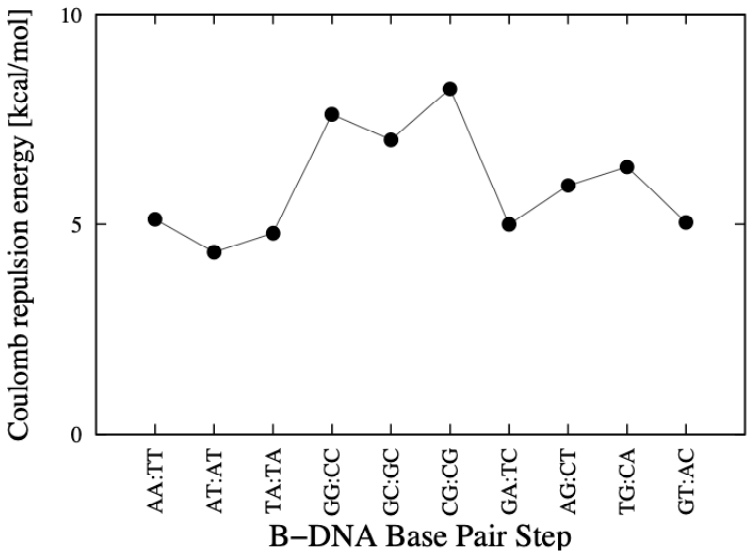

Figure 5.4: Electrostatic interaction energies arising from the Mulliken charges located at atoms involved in H-bonds. The energies are normalized by the number of H-bonds: 4 for a pair of A-T and A-T, 9 for a pair of G-C and G-C, and 6 for a pair of A-T and G-C. Energies are given in kcal/mol.

5.3 CBS[MP2] to CBS[MP4]

The overall coincidence between "CCSD(T)" and SCF (HF/DFT) implies that (1) the present "CCSD(T)" method never describe the dispersion-level non-additivity and thus, losing a large part of *true* non-additivity at CCSD(T) level of theory, or (2) a *true* dispersion-level non-additivity is essentially tiny and thus, well described by the present "CCSD(T)" method. As has been explained in "Systems and methods", the present "CCSD(T)" method relies on the CBS[MP2] approximation, *i.e.*, it is not a *true* "CCSD(T)/CBS". Since the true CCSD(T)/CBS

is believed to well reproduce the non-additivity, it is unlikely for CCSD(T)/CBS to be insufficient to describe the non-additivity. So the second statement is implausible, while the first one is to be studied in more detail. Hereafter we investigate whether or not the CBS[MP2] approximation can be a possible source of damaging a capability inherent in "CCSD(T)/CBS" of capturing the non-additivity.

To address the above issue, we attempted to apply CBS[MP3] and CBS[MP4] (as well as the true CBS) to the B-DNA base-pair steps, but they were too large to compute. Instead, we dealt with a neon tetramer as a simple/model system in which each DNA base is replaced by a Ne atom. We evaluated $\Delta\varepsilon^{(4)}$ values of the Ne tetramer at several distances between the two dimers using CCSD(T)/VTZ (w.o. CBS), CCSD(T)/CBS[MPn] ($n = 2, 3, 4$), and B3LYP-D3. Fig. 5.5 shows how those values differ from each other: CCSD(T)/CBS[MP2] is significantly different from CCSD(T)/CBS[MP3] and CCSD(T)/CBS[MP4], but it is almost same as B3LYP-D3. This indicates that the CBS[MP2] level can be regarded as the SCF-level non-additivity, which is consistent with our finding in the B-DNA case that CCSD(T)/CBS[MP2] is incapable of reproducing the dispersion-level non-additivity properly. In contrast, CCSD(T)/CBS[MP3] almost converges to CCSD(T)/CBS[MP4], while it significantly deviates from both CCSD(T)/CBS[MP2] and B3LYP-D3 – SCF-level non-additivity. From the viewpoint of the perturbation theory, this convergence means that a main contribution to the dispersion-level non-additivity can be described by the CBS[MP3] level. That is, at least CBS[MP3] is essential to describe the dispersion-level non-additivity, while CBS[MP2] is insufficient. Although we could not actually confirm how CBS[MP3/4] differ from SCF/CBS[MP2] for the B-DNA case, we may infer that even in the B-DNA case CBS[MP3/4] dominantly contribute to a description of the dispersion-level non-additivity.

5.4 Dispersion-level non-additivity in FNDMC

Next we move on to FNDMC. Its wiggling dependence of $\Delta\varepsilon^{(4)}$ on the base-pair steps appearing in Fig. 4.2 arouses suspicion whether FNDMC really reproduces

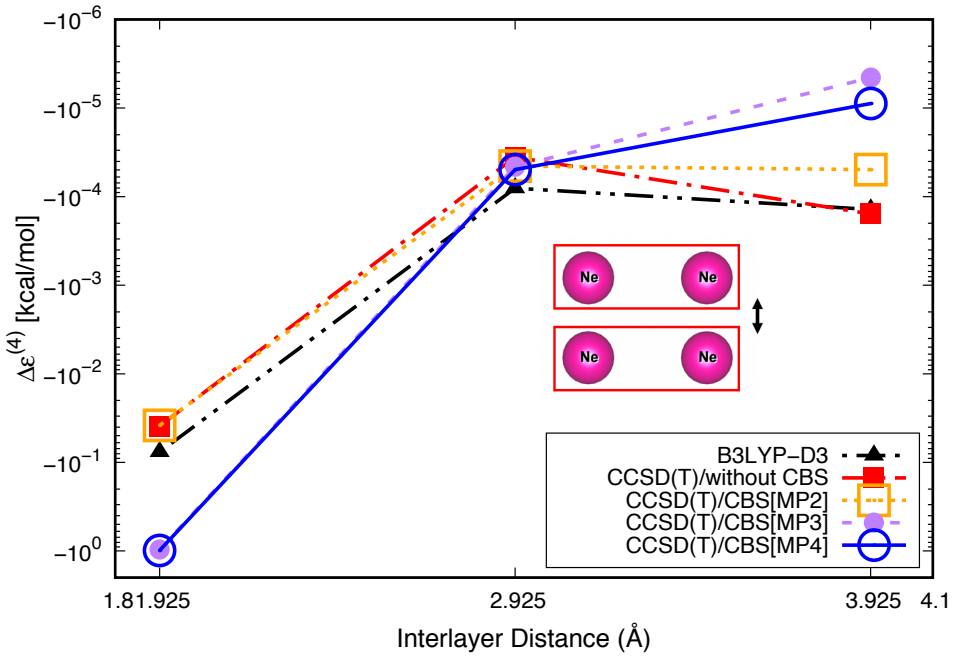

Figure 5.5: The non-additivity $\Delta\varepsilon^{(4)}$ of neon tetramer at several distances between the constituent dimers (described as "Interlayer distance" in the horizontal axis). With a fixed "Interlayer distance", all the Ne atoms located on a plane form a rectangle, where in each dimer its interatomic length is fixed to be 2.925 Å. All the CCSD(T) and DFT calculations were performed using Gaussian09 [5].

the dispersion-level non-additivity appropriately. Here we shall deliberate on two possibilities of causing faults in FNDMC: quality of trial wave functions obtained and reliability of the fixed-node approximation adopted in the present study.

We first note that such a wiggling dependence appears only in Fig. 4.2, but not in Fig. 4.1; both the results were obtained by the wave functions that were optimized at the same level of theory. In the present study we did not optimize the Slater part the trial wave functions, but the Jastrow part only. In FNDMC, the latter changes the statistical error bar only, while the former – related to the fixed-node approximation – changes the final total energy value, unlike VMC. [112] It is well known that the same performance on the Jastrow optimization leads to the same magnitude of error bars for similar system sizes if all the other computational details are assumed to be common to the systems. It is obvious from Figs. 4.1 and 4.2 that this is actually valid for the present B-DNA systems. We also insist that our choice of computational details on FNDMC – basis sets, time step, t-move scheme, as well as Jastrow function – is equivalent to a protocol established in previous studies on non-covalent systems due to Dubecký *et al.* [35]. The reasonable behavior of FNDMC stacking energies in Fig. 4.1 asserts that our choice is valid for the present B-DNA base-pair steps.

5.5 Fixed-node approximation

The fixed-node approximation is the most notorious as the cause of errors in FNDMC. [84] Previous studies on non-covalent systems including B-DNA, however, demonstrated that FNDMC works well for evaluating their complexation energies in general. [35] It is to be noted in the B-DNA stacking that this is valid for the stacking energies, but unknown for the non-additivity. The success in the FNDMC stacking energies relies on the error *cancellation* of the fixed-node approximations between the whole non-covalent system and its constituent sub-systems. This implies that the formation of non-covalent/vdW bonding does not give rise to a significant difference in nodal surface structures between the whole and the sub-systems (tetramer-dimer/dimer-monomer), leading to an accurate complexation energy. The success in the non-additive contribution requires that

two error cancellations of the fixed-node approximations simultaneously occur for $\varepsilon^{(4)}$ and $\varepsilon^{(2)}$. In the case of non-additivity, however, it is possible that the cancellation in $\varepsilon^{(4)}$ would not occur properly. Its possible factor can be attributed to a horizontal bridging between Watson-Crick bases due to H-bonds. The formation of H-bonding accompanied by the charge transfer could deform the fixed-node surface structure of the tetramer more significantly than that of vdW bonding. If this were true, the fixed-node error could not be canceled out more remarkably in $\varepsilon^{(4)}$ than in $\varepsilon^{(2)}$. The less cancellation could arouse the suspicion that the wiggling dependence of FNDMC in Fig. 4.2 is incorrect due to the fixed-node error related to the H-bonding.

In order to prove the conjecture (or anti-conjecture) about the fixed-node errors caused by the H-bonds, the most straightforward way would be to evaluate the nodal surface dependence of the non-additivity. While the stacking energy has been demonstrated to be insensitive to the dependence [36, 38], one would suspect that it is not the case for the non-additivity. Suppose it were true, the non-additivity evaluated with a different trial node would be different from the present FNDMC one in Fig. 4.2. Although we plan to address this issue in our future work, we should mention that such a calculation involves a heavy computational resource, which costs 1.2×10^6 core-hour [1.2×10^5 (core-hour) $\times$ 10 pairs]. In addition to a single reference trial node, we could employ recently developed trial nodes such selected configuration interaction [113] and multipfaffian [114] wave functions, as well as a simple multi-reference one [41, 115]. Beside their feasibility in terms of computation costs, the more sophisticated trial nodes would shed light on the nodal surface dependence of non-additivity in FNDMC, *i.e.*, we could verify whether or not the non-additivity is sensitive to the nodal surface unlike the stacking energies [35, 36].

Although we do not investigate the nodal surface dependence of $\Delta\varepsilon^{(4)}$ further, we alternatively examine if the charge transfer – the key to verifying the issue – could really matter even for the SCF-level non-additivity. Suppose it really matters, one would expect that the charge transfer could somewhat affect the dispersion-level non-additivity. We consider two types of XC functionals in terms of the long-range (LC) exchange corrections: one well describes the charge

transfer with the correction and the other dose not. Their difference in the charge density distribution tells us how significantly the distribution in a B-DNA base-pair change before and after forming the base-pair, thus clarifying an effect of changing their non-additive contributions. For example, it is well known that B3LYP-D3 is not good at capturing the charge transfer because B3LYP also fails for a number of cases relevant to the charge transfer [89] and the D3 correction never improve the B3LYP description of charge density [88]. On the other hand, CAM-B3LYP-D3 [89] remarkably improve the charge transfer, because it enhances the exact exchange for the long-range exchange based on Coulomb-Attenuating Method (CAM). As another choice, ωB97X has been reported to give a better descriptions of properties including the charge transfer than ωB97M-V in some cases. [87] We note that there are further choices of XC for reproducing the charge transfer well, such as 'self-consistent vdW' implemented in a series of 'vdW-DF' [116], but their implementations are unavailable for Gaussian basis set calculations. Fig. 5.6 focuses on a comparison among XC functionals with/without long-range corrections (originally taken from Fig. 3.1). We found that the long-range corrections by CAM-B3LYP(-D3) and ωB97X (positively) enhance the non-additive contributions compared to the counterparts, B3LYP(-D) and ωB97M-V. This implies the importance of the charge transfer caused by the H-bonding when forming a Watson-Crick base pair. Although the above analysis deals with only the SCF-level non-additivity, it would be expected that the charge transfer caused by the H-bonding could significantly deform the nodal surface structures when forming the Watson-Crick base giving rise to a large fixed-node error, and hence the wiggling dependence of $\Delta\varepsilon^{(4)}$ in FNDMC, shown in Fig. 4.2, might be false due to the less error cancellation in the $\Delta\varepsilon^{(4)}$ evaluation.

Lastly, we mention another drawback for FNDMC simulations in a practical sense. According to the previous analysis on the Ne tetramer, it might be expected that CCSD(T)/CBS[MPn] ($n \geq 3$) would get closer to FNDMC predictions as increasing the CBS[MPn] level. For comparison, we attempted to apply FNDMC to evaluate $\Delta\varepsilon^{(4)}$ of the Ne tetramer. Unfortunately, however, we could not obtain numerically/statistically reliable FNDMC results because the magnitude of $\Delta\varepsilon^{(4)}$ itself is an order of/less than the sub-chemical accuracy (0.1 kcal/mol)

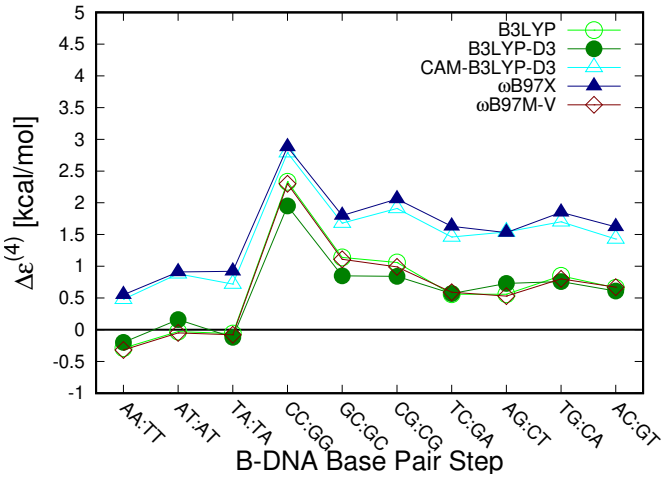

Figure 5.6: Non-additive contributions, $\Delta E^{(4)}$ [kcal/mol], predicted by different XC functionals with/without the long-range exchange corrections. The charge transfer mainly occurs at horizontal H-bonds when forming Watson-Crick bases.

and hence the corresponding error bar is required to be an order of/less than 0.01 kcal/mol. To attain such an error bar, a vast number of statistical samplings must be accumulated even for the smaller system considered here. his is another serious drawback to be noted for FNDMC.

5.6 Conclusion

As we can see in the result chapter, it is found that the FNDMC values of non-additivity alter their sign (*i.e.* they increase or decrease their stacking interactions) depending on the base-pair steps, which is contrary to all the other *ab initio* methods. On the other hand, no significant difference between the methodologies was observed for four-/two-body stacking energies, each of which are used to evaluate the non-additivity. To elucidate this contrast between the stacking and non-additivity, we made two plausible discussions about limitation on practical approximations involved in CCSD(T) and FNDMC:

1. The reason why the unexpected coincidence between CCSD(T)/CBS[MP2] and HF/B3LYP occurs only at the non-additivity level can be attributed

to the imperfect capability for MP2 to reproduce the electron correlation specific to the four-body system. The lack of the correlation never describes the correlation/dispersion-level non-additivity properly. In other words, CCSD(T)/CBS[MP2] mostly describes the SCF-level non-additivity only. In this chapter, we also construct a similar system using Ne atoms to estimate the impact of the MP2 to MP4 level baseset on the calculation results. The results also support our conclusions.

2. FNDMC demonstrates a wiggling dependence of the non-additivity. While the SCF-level non-additivity is mostly positive, the non-additive contributions described by FNDMC are both positive and negative signs. The negative sign is found to be reasonable, which might be supported by a simple model analysis based on the London theory. It would, however, be premature to draw a conclusion that the FNDMC non-additivity reveals the truth. This is because the Watson-Crick base-pair involves the charge transfer caused by the H-bonds, but we could not verify if the error cancellations of the fixed-node errors were successful for the H-bonds, as in the case of complexation energies. However, we can't ignore the FN approximation which is used to solve the QMC symbol problem. We analyzed the possible errors in this approach and discussed this.

In summary, for the results of the previous chapter, we conducted a series of discussions to explain the rationality of the results.

Chapter 6

Summary

After decades of development, quantum chemistry based on computer simulation has become an important means except for experiments. Especially in the biomacromolecules studies, such as stacking, conversion of DNA, folding of protein molecules, etc., the properties between molecules are difficult to measure in experiments. Therefore, how to explain the mechanism of action of large system molecules in biological activities has become an important research topic.

On the other hand, the structure of biomacromolecules is more complex than conventional organic macromolecular. This makes it difficult to simplify the system model using periodic conditions or other methods. The corresponding molecular force field method builds an intermolecular force field based on empirical data, which makes it possible to simulate large-scale molecular systems. The first-principles quantum chemistry is much more complex in calculation than the molecular force field method. Therefore, how to deal with large-scale models has always been a research challenge.

On the other hand, non-covalent effects play an important role in biomolecules such as vdW and H-bond, Since the induced polarity of molecules is involved, a greater cost is required in the calculation of the electronic correlations. The intermolecular interactions in many-body systems are an important research topic in quantum chemistry, but the non-additivity is not well studied. For standard SCF methods have been proven failed to describe dispersion interactions properly. Therefore, hybrid functionals based on long-range force correction and disper-

sion force correction based on empirical data are applied to DFT calculations. However, it is clear that the introduction of empirical parameters makes the DFT uncertain. If the long-range interaction is to be correctly described, calculations above the MP2 level are necessary.

The CCSD(T)/CBS method, known as the "gold standard", is limited to the size of the systems it can handle due to its computational complexity. To deal with many-body molecular systems, CCSD(T) can take a matrix of MP2 levels basis set and perform additive approximation. Then the calculation of such intermolecular interactions is only at the MP2 level, which is at the same level as the DFT method with the dispersion force correction added. Therefore, a quantum chemistry method that can handle macromolecular systems and correctly describe long-range forces (rather than using empirical methods) is particularly important. The QMC method is a wave function theory method based on stochastic statistics, which can truly describe the electron orbit. However, like other wave function theory methods. When the association between the electrons is added, the computational cost will also increase significantly. Fortunately, as the size of side-by-side computers grows larger and larger, and the balance begins to tilt, statistical-based calculations can compress time costs in large-scale parallelism. This makes it possible for the QMC method to handle larger systems.

We performed a non-covalent calculation of the non-covalent binding of the four-molecule system of B-DNA using the highly static QMC method. The results show that using FNDMC, a widely used QMC method, can achieve non-additive contributions that are difficult to capture by other methods. Among them, only the FNDMC method obtains the non-additive contribution of the negative sign in some systems, and according to the analysis of the London model, the negative sign of non-additive is expected. In contrast, the non-additive contributions of other traditional methods are positive symbols. In some B-DNA molecular systems, FNDMC also obtained a larger positive sign non-additive contribution. After a structured analysis, we find that the non-additive fluctuation trend has an important relationship with the polarity direction between molecules. Therefore, we established a model of the H-bond orientation in the base pair to analyze its regularity. We found that when the direction of H-bonds is parallel, the non-

additional contribution will be higher, while in the antiparallel, the non-additional contribution will be lower. It can explain that FNDMC can get more polar energy so that the fluctuation of non-additive contribution is enhanced. However, we cannot rule out the FN approximation, the only approximation in FNDMC, the offsetting of H-bonds generated in separate calculations.

According to the full text, with the deepening of molecular research, more and more precise molecular reaction mechanisms need to be accurately described. The demand for high-precision calculation of macromolecules is also becoming more and more urgent. With the continuous development of large-scale parallel computers, computing power will soon exceed tens of billions of times (as the writing of this article in November 2019) [12], Furthermore, we can predict that larger-scale parallel computing will be more suitable for statistical algorithm-based computing methods than traditional determinant iterative methods. So we can optimistically estimate that in the future, the QMC method will show its unique charm in more directions and fields.

Appendix I Stacking energy

Numerical values for energies

The stacking energies ($\varepsilon^{(4)}$) of ten unique B-DNA base-pair steps evaluated from various *ab initio* methods, are tabulated in Tables I.1, I.2 and I.3 respectively. In order to facilitate comparison and layout, the methods are divided into two parts, one is based on the quantum mechanics-based wave function method (WF), and the density functional theory method (DFT) is placed separately in a subsequent tables.

Stacking Energy of WF methods

Table I.1: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from CCSD(T)/CBS[MP2], HF, and FNDMC. All the energies are given in kcal/mol.

Stacking energy	Pairs	CCSD(T)/CBS[MP2][1]	HF[2]	FNDMC[2]
$\varepsilon^{(4)}$	AA:TT	−14.70	9.09	−13.0(4)
$\varepsilon^{(2)}$	A//As	−6.06	4.56	−4.3(3)
$\varepsilon^{(2)}$	T//Ts	−4.18	5.55	−2.3(3)
$\varepsilon^{(2)}$	A//Ti	−2.34	−0.44	−1.7(3)
$\varepsilon^{(2)}$	T//Ai	−2.16	−0.84	−1.3(3)
$\varepsilon^{(4)}$	AT:AT	−13.32	8.09	−10.9(7)
$\varepsilon^{(2)}$	A//Ts	−6.64	2.47	−4.0(4)
$\varepsilon^{(2)}$	A//Ts	−6.64	2.48	−5.7(4)
$\varepsilon^{(2)}$	T//Ti	0.88	1.72	1.1(5)
$\varepsilon^{(2)}$	A//Ai	−0.92	0.88	−1.7(4)
$\varepsilon^{(4)}$	TA:TA	−12.79	7.18	−7.2(4)
$\varepsilon^{(2)}$	A//Ts	−6.07	0.78	−5.6(4)
$\varepsilon^{(2)}$	A//Ts	−6.07	0.78	−5.6(4)
$\varepsilon^{(2)}$	A//Ai	−1.55	3.83	−2.0(4)
$\varepsilon^{(2)}$	T//Ti	0.70	1.38	0.9(4)
$\varepsilon^{(4)}$	GG:CC	−11.46	7.85	−8.5(7)
$\varepsilon^{(2)}$	G//Gs	−3.54	7.10	−1.4(5)
$\varepsilon^{(2)}$	C//Cs	−1.62	5.46	−2.1(4)
$\varepsilon^{(2)}$	C//Gi	−3.68	−2.63	−3.1(5)
$\varepsilon^{(2)}$	G//Ci	−4.82	−4.67	−4.2(7)
$\varepsilon^{(4)}$	GC:GC	−15.38	8.71	−14.8(9)
$\varepsilon^{(2)}$	G//Cs	−10.80	−1.22	−10.7(5)
$\varepsilon^{(2)}$	G//Cs	−10.80	−1.22	−10.6(5)
$\varepsilon^{(2)}$	C//Ci	3.09	4.16	2.0(5)

Table I.1: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from CCSD(T)/CBS[MP2], HF, and FNDMC. All the energies are given in kcal/mol.

Stacking energy	Pairs	CCSD(T)/CBS[MP2][1]	HF[2]	FNDMC[2]
$\varepsilon^{(2)}$	G//Gi	1.93	5.53	1.6(5)
$\varepsilon^{(4)}$	CG:CG	−17.33	3.53	−15.2(10)
$\varepsilon^{(2)}$	G//Cs	−7.88	−1.40	−7.1(5)
$\varepsilon^{(2)}$	G//Cs	−7.88	−1.40	−7.5(5)
$\varepsilon^{(2)}$	G//Gi	−3.91	2.34	−3.1(4)
$\varepsilon^{(2)}$	C//Ci	1.24	2.17	2.1(4)
$\varepsilon^{(4)}$	GA:TC	−12.86	10.00	−8.9(9)
$\varepsilon^{(2)}$	A//Gs	−9.14	3.14	−10.2(6)
$\varepsilon^{(2)}$	T//Cs	−4.69	2.61	−4.9(5)
$\varepsilon^{(2)}$	A//Ci	−0.31	1.15	1.3(5)
$\varepsilon^{(2)}$	T//Gi	0.58	2.01	−0.2(5)
$\varepsilon^{(4)}$	AG:CT	−13.50	6.31	−7.4(9)
$\varepsilon^{(2)}$	A//Gs	−7.58	1.26	−7.9(6)
$\varepsilon^{(2)}$	T//Cs	−6.07	0.38	−5.4(5)
$\varepsilon^{(2)}$	T//Gi	−0.47	2.22	−2.1(6)
$\varepsilon^{(2)}$	A//Ci	−0.18	1.12	−0.2(5)
$\varepsilon^{(4)}$	TG:CA	−15.20	5.22	−15.7(8)
$\varepsilon^{(2)}$	T//Gs	−5.67	1.09	−5.5(5)
$\varepsilon^{(2)}$	A//Cs	−4.96	2.99	−3.2(5)
$\varepsilon^{(2)}$	A//Gi	−4.22	0.54	−3.9(5)
$\varepsilon^{(2)}$	T//Ci	−1.15	−0.81	−0.8(5)
$\varepsilon^{(4)}$	GT:AC	−13.36	10.74	−13.5(6)
$\varepsilon^{(2)}$	T//Gs	−4.96	6.89	−4.5(4)
$\varepsilon^{(2)}$	A//Cs	−5.44	3.92	−4.3(3)
$\varepsilon^{(2)}$	T//Ci	0.30	0.88	−0.3(4)
$\varepsilon^{(2)}$	A//Gi	−4.06	−1.24	−3.6(4)

Stacking Energy of DFT methods

Table I.2: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from LDA, B3LYP, and B3LYP-D3. All the energies are given in kcal/mol.

Stacking energy	Pairs	LDA[1]	B3LYP[1]	B3LYP-D3[1]
$\varepsilon^{(4)}$	AA:TT	−10.17	9.00	−12.04
$\varepsilon^{(2)}$	A//As	−4.80	4.56	−4.94
$\varepsilon^{(2)}$	T//Ts	−4.02	5.10	−3.84
$\varepsilon^{(2)}$	A//Ti	−1.02	−0.14	−1.36
$\varepsilon^{(2)}$	T//Ai	−1.20	−0.31	−1.70
$\varepsilon^{(4)}$	AT:AT	−9.60	8.65	−12.42
$\varepsilon^{(2)}$	A//Ts	−5.66	2.78	−6.31
$\varepsilon^{(2)}$	A//Ts	−5.66	2.78	−6.31
$\varepsilon^{(2)}$	T//Ti	1.15	1.72	0.96
$\varepsilon^{(2)}$	A//Ai	−0.74	1.21	−0.92
$\varepsilon^{(4)}$	TA:TA	−9.50	7.64	−12.02
$\varepsilon^{(2)}$	A//Ts	−5.24	1.39	−5.96
$\varepsilon^{(2)}$	A//Ts	−5.24	1.39	−5.96
$\varepsilon^{(2)}$	A//Ai	−1.23	3.65	−0.85
$\varepsilon^{(2)}$	T//Ti	1.01	1.33	0.87
$\varepsilon^{(4)}$	GG:CC	−6.83	8.60	−10.35
$\varepsilon^{(2)}$	G//Gs	−1.51	7.03	−2.44
$\varepsilon^{(2)}$	C//Cs	−0.92	5.28	−1.27
$\varepsilon^{(2)}$	C//Gi	−3.22	−1.83	−3.76
$\varepsilon^{(2)}$	G//Ci	−4.39	−3.84	−4.83
$\varepsilon^{(4)}$	GC:GC	−11.48	8.80	−13.79
$\varepsilon^{(2)}$	G//Cs	−9.96	−0.47	−10.19

[1]Ref. [1]

[2]Present works

Table I.2: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from LDA, B3LYP, and B3LYP-D3. All the energies are given in kcal/mol.

Stacking energy	Pairs	LDA[1]	B3LYP[1]	B3LYP-D3[1]
$\varepsilon^{(2)}$	G//Cs	−9.96	−0.47	−10.19
$\varepsilon^{(2)}$	C//Ci	3.38	3.65	3.27
$\varepsilon^{(2)}$	G//Gi	2.90	5.26	2.47
$\varepsilon^{(4)}$	CG:CG	−13.20	4.50	−15.97
$\varepsilon^{(2)}$	G//Cs	−6.48	−0.54	−7.01
$\varepsilon^{(2)}$	G//Cs	−6.48	−0.54	−7.01
$\varepsilon^{(2)}$	G//Gi	−3.85	2.69	−4.03
$\varepsilon^{(2)}$	C//Ci	1.39	2.06	1.24
$\varepsilon^{(4)}$	GA:TC	−8.71	10.14	−11.19
$\varepsilon^{(2)}$	A//Gs	−8.03	3.06	−8.09
$\varepsilon^{(2)}$	T//Cs	−4.22	2.83	−4.67
$\varepsilon^{(2)}$	A//Ci	0.42	1.70	0.06
$\varepsilon^{(2)}$	T//Gi	1.11	2.01	0.94
$\varepsilon^{(4)}$	AG:CT	−9.62	6.97	−12.82
$\varepsilon^{(2)}$	A//Gs	−5.66	1.74	−6.67
$\varepsilon^{(2)}$	T//Cs	−5.36	0.95	−6.12
$\varepsilon^{(2)}$	T//Gi	−0.53	2.20	−0.56
$\varepsilon^{(2)}$	A//Ci	0.00	1.35	−0.20
$\varepsilon^{(4)}$	TG:CA	−11.37	5.98	−13.94
$\varepsilon^{(2)}$	T//Gs	−4.63	1.51	−5.13
$\varepsilon^{(2)}$	A//Cs	−4.05	3.26	−4.18
$\varepsilon^{(2)}$	A//Gi	−3.88	1.06	−4.22
$\varepsilon^{(2)}$	T//Ci	−0.94	−0.60	−1.17
$\varepsilon^{(4)}$	GT:AC	−9.96	10.47	−11.99
$\varepsilon^{(2)}$	T//Gs	−4.28	5.85	−4.45
$\varepsilon^{(2)}$	A//Cs	−4.11	4.27	−4.47

Table I.2: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from LDA, B3LYP, and B3LYP-D3. All the energies are given in kcal/mol.

Stacking energy	Pairs	LDA[1]	B3LYP[1]	B3LYP-D3[1]
$\varepsilon^{(2)}$	T//Ci	0.53	0.82	0.41
$\varepsilon^{(2)}$	A//Gi	−4.07	−1.07	−4.09

Table I.3: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from CAM-B3LYP-D3, ωB97X, and M06-2X. All the energies are given in kcal/mol.

Stacking energy	Pairs	CAM-B3LYP-D3[1]	ωB97X[1]	M06-2X[1]
$\varepsilon^{(4)}$	AA:TT	−13.94	−8.52	−11.84
$\varepsilon^{(2)}$	A//As	−6.08	−4.19	−5.96
$\varepsilon^{(2)}$	T//Ts	−4.62	−2.52	−3.83
$\varepsilon^{(2)}$	A//Ti	−1.68	−1.02	−1.07
$\varepsilon^{(2)}$	T//Ai	−2.04	−1.33	−1.39
$\varepsilon^{(4)}$	AT:AT	−13.90	−8.76	−11.12
$\varepsilon^{(2)}$	A//Ts	−7.04	−5.14	−6.17
$\varepsilon^{(2)}$	A//Ts	−7.04	−5.14	−6.17
$\varepsilon^{(2)}$	T//Ti	0.78	1.25	1.16
$\varepsilon^{(2)}$	A//Ai	−1.47	−0.63	−0.48
$\varepsilon^{(4)}$	TA:TA	−13.88	−8.63	−11.09
$\varepsilon^{(2)}$	A//Ts	−6.90	−5.08	−5.85
$\varepsilon^{(2)}$	A//Ts	−6.90	−5.08	−5.85
$\varepsilon^{(2)}$	A//Ai	−1.58	−0.49	−0.77
$\varepsilon^{(2)}$	T//Ti	0.77	1.10	1.02
$\varepsilon^{(4)}$	GG:CC	−12.14	−6.93	−9.08
$\varepsilon^{(2)}$	G//Gs	−3.49	−1.20	−4.80
$\varepsilon^{(2)}$	C//Cs	−1.90	−0.51	−3.38
$\varepsilon^{(2)}$	C//Gi	−5.21	−3.42	−1.14

[1]Present study

Table I.3: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from CAM-B3LYP-D3, ωB97X, and M06-2X. All the energies are given in kcal/mol.

Stacking energy	Pairs	CAM-B3LYP-D3[1]	ωB97X[1]	M06-2X[1]
$\varepsilon^{(2)}$	G//Ci	−4.33	−4.67	−2.55
$\varepsilon^{(4)}$	GC:GC	−15.85	−10.31	−14.13
$\varepsilon^{(2)}$	G//Cs	−11.42	−9.52	−11.21
$\varepsilon^{(2)}$	G//Cs	−11.42	−9.52	−11.21
$\varepsilon^{(2)}$	C//Ci	3.30	3.65	3.31
$\varepsilon^{(2)}$	G//Gi	2.00	3.28	3.58
$\varepsilon^{(4)}$	CG:CG	−18.35	−12.89	−15.84
$\varepsilon^{(2)}$	G//Cs	−8.20	−6.69	−7.40
$\varepsilon^{(2)}$	G//Cs	−8.20	−6.69	−7.40
$\varepsilon^{(2)}$	G//Gi	−4.97	−3.19	−3.93
$\varepsilon^{(2)}$	C//Ci	1.11	1.62	1.57
$\varepsilon^{(4)}$	GA:TC	−13.05	−7.71	−11.28
$\varepsilon^{(2)}$	A//Gs	−9.37	−7.12	−9.47
$\varepsilon^{(2)}$	T//Cs	−5.39	−3.76	−4.72
$\varepsilon^{(2)}$	A//Ci	−0.41	0.28	0.43
$\varepsilon^{(2)}$	T//Gi	0.66	1.26	1.23
$\varepsilon^{(4)}$	AG:CT	−14.58	−9.24	−11.16
$\varepsilon^{(2)}$	A//Gs	−7.65	−5.74	−6.80
$\varepsilon^{(2)}$	T//Cs	−6.84	−5.23	−5.68
$\varepsilon^{(2)}$	T//Gi	−1.07	0.04	0.07
$\varepsilon^{(2)}$	A//Ci	−0.56	0.16	0.24
$\varepsilon^{(4)}$	TG:CA	−16.11	−10.83	−13.63
$\varepsilon^{(2)}$	T//Gs	−6.10	−4.39	−5.15
$\varepsilon^{(2)}$	A//Cs	−5.37	−3.76	−4.76
$\varepsilon^{(2)}$	A//Gi	−5.02	−3.59	−3.93
$\varepsilon^{(2)}$	T//Ci	−1.32	−0.95	−1.06

Table I.3: Four- and two-body stacking energies ($\varepsilon^{(4)}$ and $\varepsilon^{(2)}$) evaluated from CAM-B3LYP-D3, ωB97X, and M06-2X. All the energies are given in kcal/mol.

Stacking energy	Pairs	CAM-B3LYP-D3[1]	ωB97X[1]	M06-2X[1]
$\varepsilon^{(4)}$	GT:AC	−13.76	−8.36	−11.80
$\varepsilon^{(2)}$	T//Gs	−5.40	−3.16	−4.67
$\varepsilon^{(2)}$	A//Cs	−5.38	−3.73	−5.15
$\varepsilon^{(2)}$	T//Ci	0.32	0.60	0.51
$\varepsilon^{(2)}$	A//Gi	−4.73	−3.69	−3.59

[1]Present study

Appendix II Mulliken charge distribution

Mulliken charge and its electrostatic interaction

As mentioned in the main text, we evaluated the Madelung energy for H-bond as a Coulombic interaction energy between nitrogenous bases in a Watson-Crick pair. The Mulliken charge distributions is shown in Figures II.2, where the Mulliken charge distributions are evaluated at the B3LYP-GD3/VTZ level of theory using the Gaussian09 code.

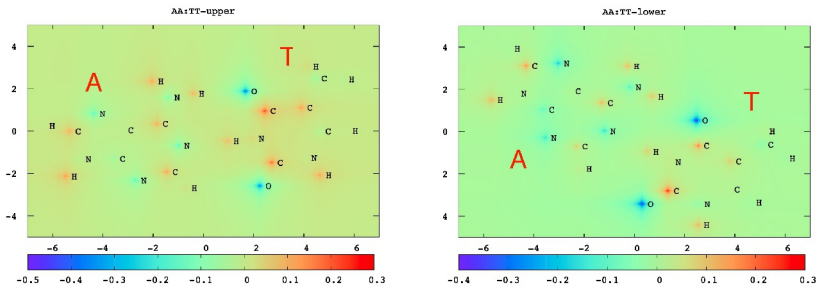

Figure II.1: Mulliken charge distribution on molecular planes of AA:TT. In each row, the left and right panels respectively correspond to upper and lower positions in direction from 5' to 3' carbons.

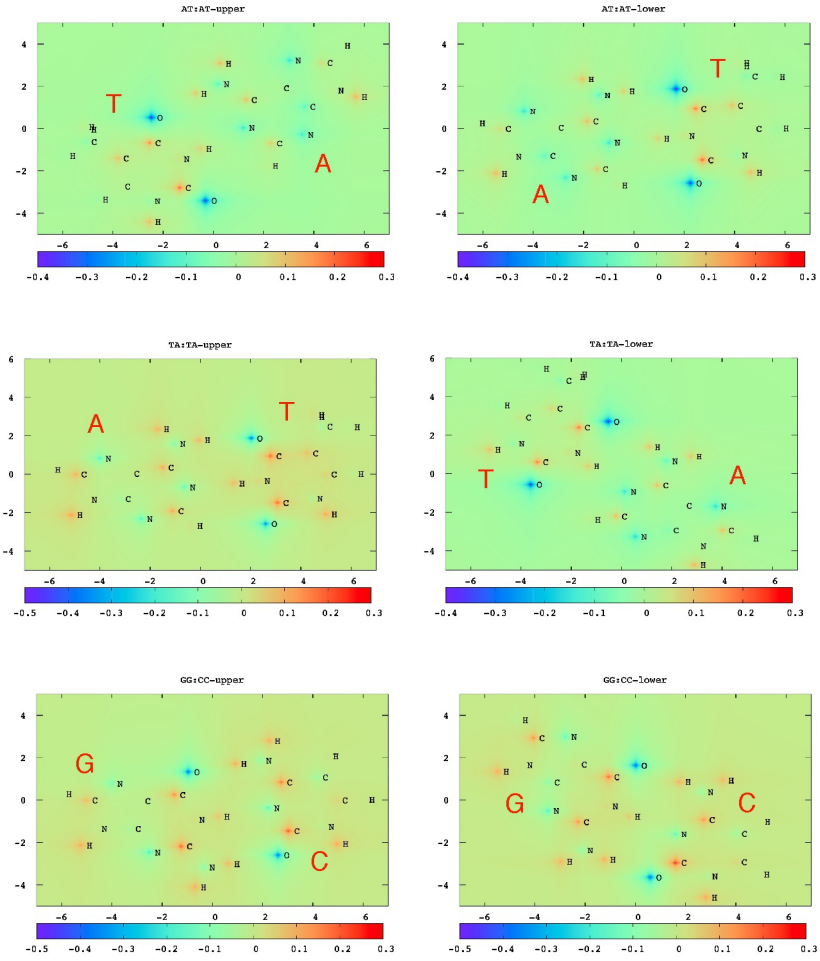

Figure II.2: Mulliken charge distribution on molecular planes of AT:AT (upper panel), TA:TA (middle panel) and GG:CC (lower panel). In each row, the left and right panels respectively correspond to upper and lower positions in direction from 5' to 3' carbons.

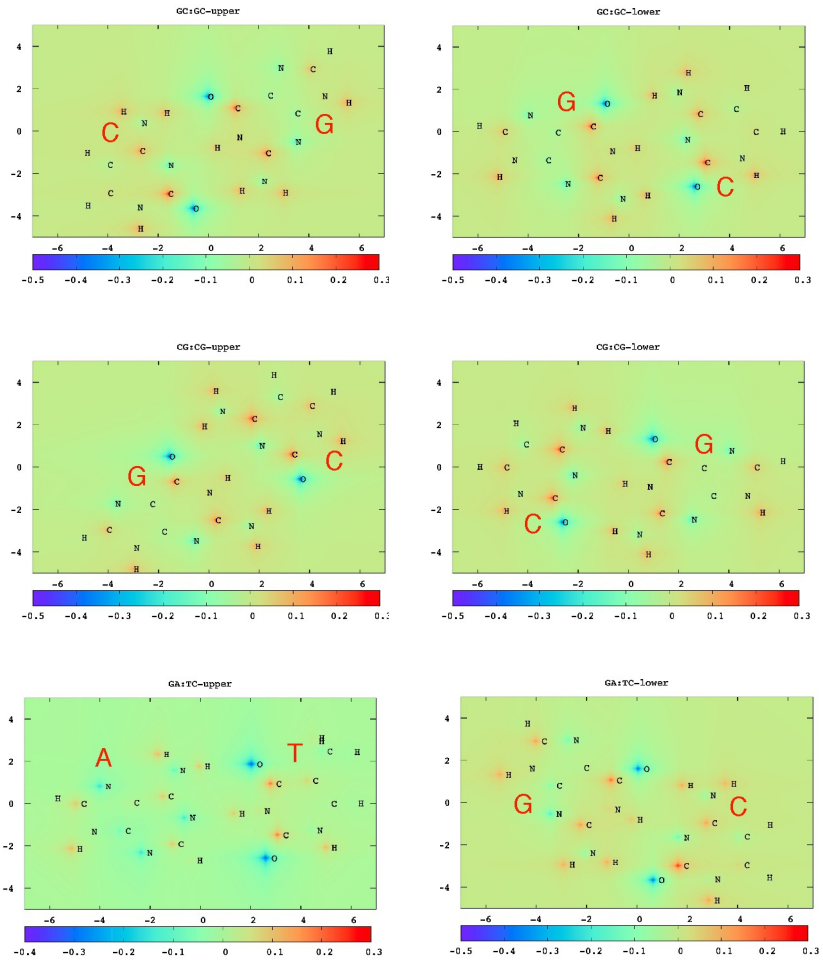

Figure II.3: Mulliken charge distribution on molecular planes of GC:GC (upper panel), CG:CG (middle panel) and GA:TC (lower panel). In each row, the left and right panels respectively correspond to upper and lower positions in direction from 5' to 3' carbons.

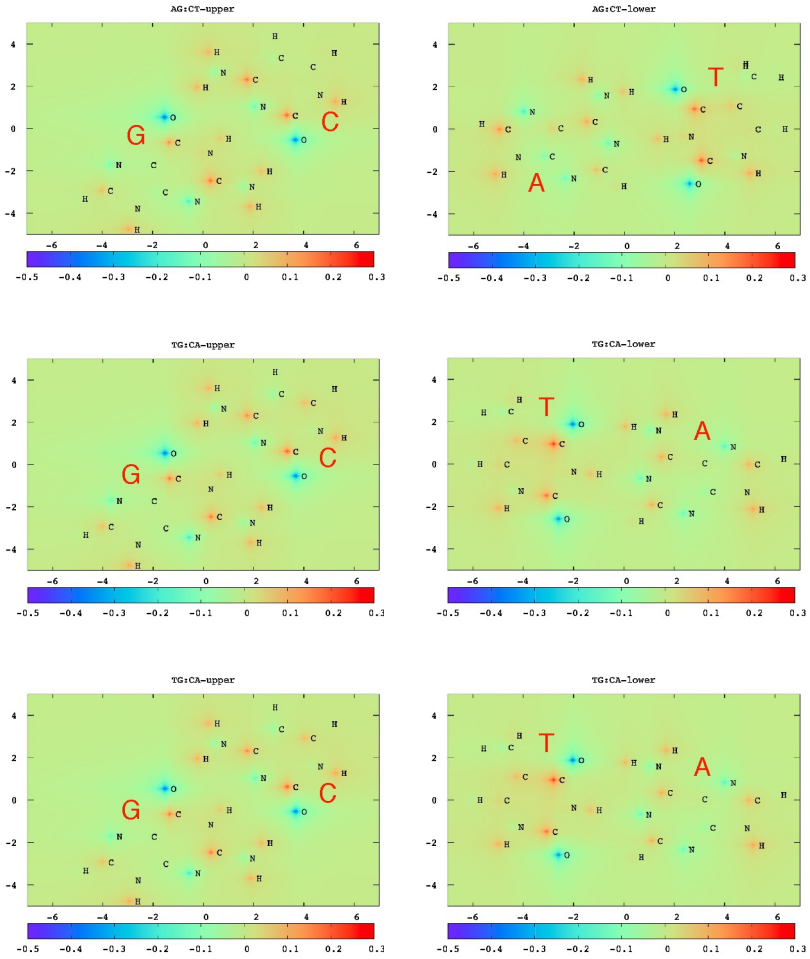

Figure II.4: Mulliken charge distribution on molecular planes of AG:CT (upper panel), TG:CA (middle panel) and GT:AC (lower panel). In each row, the left and right panels respectively correspond to upper and lower positions in direction from 5' to 3' carbons.

References

[1] Jiří Šponer, Petr Jurečka, Ivan Marchan, F. Javier Luque, Modesto Orozco, and Pavel Hobza. Nature of base stacking: Reference quantum-chemical stacking energies in ten unique b-dna base-pair steps. *Chemistry – A European Journal*, 12(10):2854–2865, 2006.

[2] J. Grant Hill and James A. Platts. Calculating stacking interactions in nucleic acid base-pair steps using spin-component scaling and local second order moller-plesset perturbation theory. *Phys. Chem. Chem. Phys.*, 10:2785–2791, 2008.

[3] Trent M. Parker, Edward G. Hohenstein, Robert M. Parrish, Nicholas V. Hud, and C. David Sherrill. Quantum-mechanical analysis of the energetic contributions to π stacking in nucleic acids versus rise, twist, and slide. *Journal of the American Chemical Society*, 135(4):1306–1316, 2013. PMID: 23265256.

[4] Christopher A Hunter. Sequence-dependent dna structure: the role of base stacking interactions. *Journal of molecular biology*, 230(3):1025–1054, 1993.

[5] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi,

J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 revision d.01. Gaussian Inc. Wallingford CT 2009.

[6] Ilya G Kaplan. *Intermolecular interactions: physical picture, computational methods and model potentials*. John Wiley & Sons, 2006.

[7] James D Watson, Francis Crick, et al. A structure for deoxyribose nucleic acid. 1953.

[8] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Science, 2012.

[9] Harvey Lodish, Arnold Berk, Chris A Kaiser, Monty Krieger, Matthew P Scott, Anthony Bretscher, Hidde Ploegh, Paul Matsudaira, et al. *Molecular cell biology*. Macmillan, 2008.

[10] Jinhui Zhan. *Molecular Simulation Study on the interactions between Several Important Proteins and Substrates*. PhD thesis, Jilin University, 2699 Qianjin St, Chaoyang, Changchun, Jilin, China, 5 2009.

[11] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585, 1977.

[12] ACM. Supercomputers top 500. https://www.top500.org. Accessed Nov , 2019.

[13] I Dzidic and Paul Kebarle. Hydration of the alkali ions in the gas phase. enthalpies and entropies of reactions m+ (h2o) n-1+ h2o= m+ (h2o) n. *The Journal of Physical Chemistry*, 74(7):1466–1474, 1970.

[14] Cornelis Altona and Dirk H Faber. Empirical force field calculations. In *Dynamic Chemistry*, pages 1–38. Springer, 1974.

[15] Alexander D MacKerell Jr and Nilesh K Banavali. All-atom empirical force field for nucleic acids: Ii. application to molecular dynamics simulations of dna and rna in solution. *Journal of computational chemistry*, 21(2):105–120, 2000.

[16] Robert B Nachbar Jr, W Douglas Hounshell, Vincent A Naman, Olof Wennerstroem, Alberto Guenzi, and Kurt Mislow. Application of empirical force field calculations to internal dynamics in 9-benzyltriptycenes. *The Journal of Organic Chemistry*, 48(8):1227–1232, 1983.

[17] Viloya S Allured, Christine M Kelly, and Clark R Landis. Shapes empirical force field: new treatment of angular potentials and its application to square-planar transition-metal complexes. *Journal of the American Chemical Society*, 113(1):1–12, 1991.

[18] Wangshen Xie and Jiali Gao. Design of a next generation force field: the x-pol potential. *Journal of chemical theory and computation*, 3(6):1890–1900, 2007.

[19] C David Sherrill. Energy component analysis of π interactions. *Accounts of chemical research*, 46(4):1020–1028, 2012.

[20] Sandro Bottaro, Giovanni Bussi, Scott D. Kennedy, Douglas H. Turner, and Kresten Lindorff-Larsen. Conformational ensembles of rna oligonucleotides from integrating nmr and molecular simulations. *Science Advances*, 4(5), 2018.

[21] Justin A Lemkul and Alexander D MacKerell Jr. Polarizable force field for dna based on the classical drude oscillator: I. refinement using quantum mechanical base stacking and conformational energetics. *Journal of chemical theory and computation*, 13(5):2053–2071, 2017.

[22] Jiri Hostas, Dávid Jakubec, Roman A Laskowski, Ramachandran Gnanasekaran, Jan Rezac, Jiri Vondrasek, and Pavel Hobza. Representative amino acid side-chain interactions in protein–dna complexes: A comparison of highly accurate correlated ab initio quantum mechanical calculations

and efficient approaches for applications to large systems. *Journal of chemical theory and computation*, 11(9):4086–4092, 2015.

[23] Milovanovic Branislav, Marko Kojic, Milena Petkovic, and Mihajlo Etinski. New insight into uracil stacking in water from ab initio molecular dynamics. *Journal of chemical theory and computation*, 14(5):2621–2632, 2018.

[24] Jan Hermann, Dario Alfe, and Alexandre Tkatchenko. Nanoscale π–π stacked molecules are bound by collective charge fluctuations. *Nature communications*, 8:14052, 2017.

[25] Daniel J Cole, Jonah Z Vilseck, Julian Tirado-Rives, Mike C Payne, and William L Jorgensen. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *Journal of chemical theory and computation*, 12(5):2312–2323, 2016.

[26] Sereina Riniker. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *Journal of chemical information and modeling*, 58(3):565–578, 2018.

[27] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[28] Kenta Hongo, Mark A. Watson, Roel S. Sánchez-Carrera, Toshiaki Iitaka, and Alán Aspuru-Guzik. Failure of conventional density functionals for the prediction of molecular crystal polymorphism: A quantum monte carlo study. *J. Phys. Chem. Lett.*, 1(12):1789–1794, 2010.

[29] Mutasem Omar Sinnokrot and C. David Sherrill. Highly accurate coupled cluster potential energy curves for the benzene dimer: Sandwich, t-shaped, and parallel-displaced configurations. *J. Phys. Chem. A*, 108(46):10200–10207, 2004.

[30] Jeffrey C. Grossman. Benchmark quantum monte carlo calculations. *J. Chem. Phys.*, 117(4):1434–1440, 2002.

[31] Martin Korth, Arne Luchow, and Stefan Grimme. Toward the exact solution of the electronic schrodinger equation for noncovalent molecular interactions: Worldwide distributed quantum monte carlo calculations. *J. Phys. Chem. A*, 112(10):2104–2109, 2008.

[32] Mark A. Watson, Kenta Hongo, Toshiaki Iitaka, and Alán Aspuru-Guzik. *A Benchmark Quantum Monte Carlo Study of Molecular Crystal Polymorphism: A Challenging Case for Density-Functional Theory*, chapter 10, pages 101–117. 2012.

[33] Matúš Dubecký, Petr Jurečka, René Derian, Pavel Hobza, Michal Otyepka, and Lubos Mitas. Quantum monte carlo methods describe noncovalent interactions with subchemical accuracy. *J. Chem. Theory Comput.*, 9(10):4287–4292, 2013.

[34] L. Horváthová, M. Dubecký, L. Mitas, and I. Štich. Quantum monte carlo study of π-bonded transition metal organometallics: Neutral and cationic vanadiumbenzene and cobaltbenzene half sandwiches. *J. Chem. Theory Comput.*, 9(1):390–400, 2013.

[35] Matús Dubecký, Rene Derian, Petr Jurečka, Lubos Mitas, Pavel Hobza, and Michal Otyepka. Quantum monte carlo for noncovalent interactions: an efficient protocol attaining benchmark accuracy. *Phys. Chem. Chem. Phys.*, 16:20915–20923, 2014.

[36] Kenta Hongo, Nguyen Thanh Cuong, and Ryo Maezono. The importance of electron correlation on stacking interaction of adenine-thymine base-pair step in b-dna: A quantum monte carlo study. *J. Chem. Theory Comput.*, 9(2):1081–1086, 2013.

[37] Kenta Hongo, Mark A. Watson, Toshiaki Iitaka, Alán Aspuru-Guzik, and Ryo Maezono. Diffusion monte carlo study of para-diiodobenzene polymorphism revisited. *J. Chem. Theory Comput.*, 11(3):907–917, 2015.

[38] Kenta Hongo and Ryo Maezono. *Practical Diffusion Monte Carlo Simulations for Large Noncovalent Systems*, chapter 9, pages 127–143.

[39] Matúš Dubecký, Lubos Mitas, and Petr Jurečka. Noncovalent interactions by quantum monte carlo. *Chemical Reviews*, 116(9):5188–5215, 2016. PMID: 27081724.

[40] Kenta Hongo and Ryo Maezono. A computational scheme to evaluate hamaker constants of molecules with practical size and anisotropy. 2017.

[41] Tom Ichibha, Zhufeng Hou, Kenta Hongo, and Ryo Maezono. New insight into the ground state of fepc: A diffusion monte carlo study. *Sci. Rep.*, 6:29661, Jul 2017.

[42] Anthony M. Reilly and Alexandre Tkatchenko. van der waals dispersion interactions in molecular materials: beyond pairwise additivity. *Chem. Sci.*, 6:3289–3301, 2015.

[43] Alston J. Misquitta, Ryo Maezono, Neil D. Drummond, Anthony J. Stone, and Richard J. Needs. Anomalous nonadditive dispersion interactions in systems of three one-dimensional wires. *Phys. Rev. B*, 89:045140, Jan 2014.

[44] Ken Sinkou Qin, Tom Ichibha, Kenta Hongo, and Ryo Maezono. Inconsistencies in ab initio evaluations of non-additive contributions of dna stacking energies. *Chemical Physics*, 529:110554, 2020.

[45] Charlotte Froese Fischer. General Hartree-Fock program. *Computer Physics Communications*, 43(3):355–365, feb 1987.

[46] RK Nesbet. Configuration interaction in orbital theories. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 230(1182):312–321, 1955.

[47] Stephen R Langhoff and Ernest R Davidson. Configuration interaction calculations on the nitrogen molecule. *International Journal of Quantum Chemistry*, 8(1):61–72, 1974.

[48] J Čížek. J. čížek, adv. chem. phys. 14, 35 (1969). *Adv. Chem. Phys.*, 14:35, 1969.

[49] C Møller and MS Plesset. Mp perturbation theory. *Physical Review*, 46:618, 1934.

[50] S Saebo and Peter Pulay. Local treatment of electron correlation. *Annual Review of Physical Chemistry*, 44(1):213–236, 1993.

[51] S Saebo and P Pulay. Local treatment of electron correlation. *Annual Review of Physical Chemistry*, 44(1):213–236, 1993.

[52] Bogumil Jeziorski, Robert Moszynski, and Krzysztof Szalewicz. Perturbation theory approach to intermolecular potential energy surfaces of van der waals complexes. *Chemical Reviews*, 94(7):1887–1930, 1994.

[53] László Almásy and Attila Bende. Intermolecular interaction in methylene halide (ch2f2, ch2cl2, ch2br2 and ch2i2) dimers. *Molecules*, 24(9):1810, 2019.

[54] Reiner M Dreizler and Eberhard KU Gross. *Density Functional Theory*. Springer, 1990.

[55] W Kohn. Kohn, w., and lj sham, 1965, phys. rev. 140, a1133. *Phys. Rev.*, 140:A1133, 1965.

[56] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, November 1964.

[57] W. Kohn and L. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965.

[58] Klaus Capelle. A bird's-eye view of density-functional theory. *Brazilian Journal of Physics*, 36(4A):1318–1343, 2006.

[59] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.

[60] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[61] Takeshi Yanai, David P Tew, and Nicholas C Handy. A new hybrid exchangecorrelation functional using the coulomb-attenuating method (cam-b3lyp). *Chem. Phys. Lett.*, 393(13):51–57, 2004.

[62] Yan Zhao and DonaldG. Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theor. Chem. Acc.*, 120(1-3):215–241, 2008.

[63] Yan Zhao, Nathan E. Schultz, and D. G. Truhlar. Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions. *J. Chem. Phys.*, 123(16):161103, 2005.

[64] Roberto Peverati and Donald G. Truhlar. Improving the accuracy of hybrid meta-gga density functionals by range separation. *J. Phys. Chem. Lett.*, 2(21):2810–2817, 2011.

[65] Frans B. van Duijneveldt, Jeanne G. C. M. van Duijneveldt-van de Rijdt, and Joop H. van Lenthe. State of the art in counterpoise theory. *Chem. Rev.*, 94(7):1873–1885, 1994.

[66] Fritz London. The general theory of molecular forces. *Transactions of the Faraday Society*, 33:8b–26, 1937.

[67] I.N. Levine. *Physical Chemistry*. McGraw-Hill, New York, 4th ed. edition, 1995.

[68] Jerzy Leszczynski. *Computational Molecular Biology*. Elsevier Science B.V., Jackson, 1st ed. edition, 1999.

[69] W Kołos. Long-range interaction between 1s and 2s or 2p hydrogen atoms. *International Journal of Quantum Chemistry*, 1(2):169–186, 1967.

[70] Stefan Grimme. Density functional theory with london dispersion corrections. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(2):211–228, 2011.

[71] Stephan Ehrlich, Jonas Moellmann, and Stefan Grimme. Dispersion-corrected density functional theory for aromatic interactions in complex systems. *Accounts of Chemical Research*, 46(4):916–926, 2013.

[72] Federico Becca and Sandro Sorella. *Quantum Monte Carlo approaches for correlated systems*. Cambridge University Press, 2017.

[73] WMC Foulkes, L Mitas, RJ Needs, and G Rajagopal. Quantum monte carlo simulations of solids. *Reviews of Modern Physics*, 73(1):33, 2001.

[74] Robert Jastrow. Many-body problem with strong forces. *Phys. Rev.*, 98:1479–1484, Jun 1955.

[75] N. D. Drummond, M. D. Towler, and R. J. Needs. Jastrow correlation factor for atoms, molecules, and solids. *Phys. Rev. B*, 70:235119, Dec 2004.

[76] Tosio Kato. On the eigenfunctions of many-particle systems in quantum mechanics. *Comm. Pure Appl. Math.*, 10(2):151–177, 1957.

[77] Paul Richard Charles Kent. *Techniques and Applications of Quantum Monte Carlo*. PhD thesis, the University of Cambridge, The Old Schools, Trinity Lane, Cambridge, CB2 1TN, United Kingdom, 8 1999.

[78] Ye Luo. *Ab initio molecular dynamics of water by quantum Monte Carlo*. PhD thesis, Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea, 265, 34136 Trieste TS, Italy, 10 2014.

[79] Kousuke Nakano. *Phonon dispersions and Fermi surfaces nesting explaining the variety of charge ordering in titanium-oxypnictides superconductors*. PhD thesis, Japan Advanced Institute of Science and Technology, The Old Schools, Trinity Lane, Cambridge, CB2 1TN, United Kingdom, 12 2017.

[80] Alexander Nikolai Badinski. *Forces in Quantum Monte Carlo*. PhD thesis, the University of Cambridge, The Old Schools, Trinity Lane, Cambridge, CB2 1TN, United Kingdom, 7 2008.

[81] David M Ceperley and Lubos Mitas. Quantum monte carlo methods in chemistry. *New methods in computational quantum mechanics*, pages 1–38, 1995.

[82] Michele Casula. Beyond the locality approximation in the standard diffusion monte carlo method. *Physical Review B*, 74(16):161102, 2006.

[83] Luboš Mitáš, Eric L. Shirley, and David M. Ceperley. Nonlocal pseudopotentials and diffusion monte carlo. *J. Chem. Phys.*, 95(5):3467–3475, 1991.

[84] Peter J. Reynolds, David M. Ceperley, Berni J. Alder, and William A. Lester. Fixed-node quantum monte carlo for molecules. *J. Chem. Phys.*, 77(11):5593–5603, 1982.

[85] Michele Casula. Beyond the locality approximation in the standard diffusion monte carlo method. *Phys. Rev. B*, 74:161102, Oct 2006.

[86] Andrew Travers and Georgi Muskhelishvili. Dna structure and function. *The FEBS journal*, 282(12):2279–2295, 2015.

[87] Gjergji Sini, John S. Sears, and Jean-Luc Brédas. Evaluating the performance of dft functionals in assessing the interaction energy and ground-state charge transfer of donor/acceptor complexes: Tetrathiafulvalene-tetracyanoquinodimethane (ttf-tcnq) as a model case. *Journal of Chemical Theory and Computation*, 7(3):602–609, Mar 2011.

[88] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *J. Chem. Phys.*, 132(15):154104: 1–19, 2010.

[89] Takeshi Yanai, David P Tew, and Nicholas C Handy. A new hybrid exchange correlation functional using the Coulomb-attenu ating method (CAM-B3LYP). *Chemical Physics Letters*, 393(1-3):51–57, jul 2004.

[90] Jan Řezáč and Pavel Hobza. Describing noncovalent interactions beyond the common approximations: How accurate is the "gold standard," ccsd(t)

at the complete basis set limit? *J. Chem. Theory Comput.*, 9(5):2151–2155, 2013.

[91] J. Sponer H. Kruse, P. Banas. Investigations of stacked dna base-pair steps: highly accurate stacking interaction energies, energy decomposition, and many-body stacking effects. *Journal of chemical theory and computation*, 15(1):95–115, 2019.

[92] Grzegorz Chalasinski and Malgorzata M. Szczesniak. Origins of structure and energetics of van der waals clusters from ab initio calculations. *Chemical Reviews*, 94(7):1723–1765, 1994.

[93] J. Sponer, H.A. Gabb, J. Leszczynski, and P. Hobza. Base-base and deoxyribose-base stacking interactions in b-dna and z-dna: a quantum-chemical study. *Biophysical Journal*, 73(1):76 – 87, 1997.

[94] Grzegorz Chałasiński and Małgorzata M. Szczęśniak. State of the art and challenges of the ab initio theory of intermolecular interactions. *Chemical Reviews*, 100(11):4227–4252, 2000. PMID: 11749345.

[95] R J Needs, M D Towler, N D Drummond, and P López Ríos. Continuum variational and diffusion quantum monte carlo calculations. *J. Phys.: Condens. Matter*, 22(2):023201, 2010.

[96] M. Burkatzki, C. Filippi, and M. Dolg. Energy-consistent pseudopotentials for quantum monte carlo calculations. *J. Chem. Phys.*, 126(23):234105: 1–8, 2007.

[97] Peter J. Reynolds, David M. Ceperley, Berni J. Alder, and William A. Lester. Fixed-node quantum monte carlo for molecules. *J. Chem. Phys.*, 77(11):5593–5603, 1982.

[98] Michele Casula. Beyond the locality approximation in the standard diffusion monte carlo method. *Phys. Rev. B*, 74:161102, Oct 2006.

[99] Vipin Kumar. *Introduction to parallel computing*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[100] Joost VandeVondele, Matthias Krack, Fawzi Mohamed, Michele Parrinello, Thomas Chassaing, and Jürg Hutter. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications*, 2(167):103–128, 2005.

[101] John L Gustafson. Reevaluating amdahl's law. *Communications of the ACM*, 31(5):532–533, 1988.

[102] Gene M Amdahl. Computer architecture and amdahl's law. *Computer*, 46(12):38–46, 2013.

[103] Pablo López Ríos. *Backflow and pairing wave function for quantum Monte Carlo methods*. PhD thesis, the University of Cambridge, The Old Schools, Trinity Lane, Cambridge, CB2 1TN, United Kingdom, 9 2006.

[104] Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. *Chem. Rev.*, 112(1):289–320, 2012.

[105] Masayuki Hasegawa and Kazume Nishidate. Semiempirical approach to the energetics of interlayer binding in graphite. *Phys. Rev. B*, 70:205431, Nov 2004.

[106] Alexandre Tkatchenko and O. Anatole von Lilienfeld. Popular kohn-sham density functionals strongly overestimate many-body interactions in van der waals systems. *Phys. Rev. B*, 78:045116, Jul 2008.

[107] Fabian Kilchherr, Christian Wachauf, Benjamin Pelz, Matthias Rief, Martin Zacharias, and Hendrik Dietz. Single-molecule dissection of stacking forces in dna. *Science*, 353(6304):aaf5508, 2016.

[108] Annamaria Fiethen, Georg Jansen, Andreas Hesselmann, and Martin Schütz. Stacking energies for average b-dna structures from the combined density functional theory and symmetry-adapted perturbation theory approach. *Journal of the American Chemical Society*, 130(6):1802–1803, 2008. PMID: 18201088.

[109] George A Jeffrey and Wolfram Saenger. *Hydrogen bonding in biological structures*. Springer Science & Business Media, 2012.

[110] Gastone Gilli, Fabrizio Bellucci, Valeria Ferretti, and Valerio Bertolasi. Evidence for resonance-assisted hydrogen bonding from crystal-structure correlations on the enol form of the. beta.-diketone fragment. *Journal of the American Chemical Society*, 111(3):1023–1028, 1989.

[111] Alan D McNaught and Alan D McNaught. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford, 1997.

[112] Yongkyung Kwon, David M. Ceperley, and Richard McKelvy Martin. Effects of backflow correlation in the three-dimensional electron gas: Quantum monte carlo study. 1998.

[113] Yann Garniron, Anthony Scemama, Emmanuel Giner, Michel Caffarel, and Pierre-François Loos. Selected configuration interaction dressed by perturbation. *The Journal of Chemical Physics*, 149(6):064103, 2018.

[114] M. Bajdich, L. Mitas, L. K. Wagner, and K. E. Schmidt. Pfaffian pairing and backflow wavefunctions for electronic structure quantum monte carlo methods. *Phys. Rev. B*, 77:115112, Mar 2008.

[115] Kenta Hongo and Ryo Maezono. A benchmark quantum monte carlo study of the ground state chromium dimer. *International Journal of Quantum Chemistry*, 112(5):1243–1255, 2012.

[116] Kristian Berland, Valentino R Cooper, Kyuho Lee, Elsebeth Schröder, T Thonhauser, Per Hyldgaard, and Bengt I Lundqvist. van der waals forces in density functional theory: a review of the vdW-DF method. *Reports on Progress in Physics*, 78(6):066501, may 2015.

List of Abbreviations

B3LYP	Becke, 3-parameter, Lee-Yang-Parr
B97	Becke 1997
BO	Born-Oppenheimer
BSSE	Basis Set Superposition Error
CAM-B3LYP	Coulomb-attenuating method-B3LYP
CBS	Complete Basis Set
CC	Coupled Cluster
CCSD(T)	Coupled Cluster with both Single Double and perturbative triples Substitutions
CI	Configuration Interaction
DFT	Density Functional Theory
DFT+D	Dispersion-corrected Density Functional Theory
DMC	Diffusion Monte Carlo
FNDMC	Fixed-Node Diffusion Monte Carlo
GGA	Generalized Gradient Approximation
GTO	Gaussian Type Orbital
HF	Hartree-Fock
KS	Kohn-Sham
LC	Long-range Corrected
LDA	Local Density Approximation

LMP2	Local Electron Correlation Methods at the Second-order Perturbation Theory Level
M06	Minnesota 06
MC	Monte Carlo
MD	Molecular Dynamics
MMC	Metropolis Monte Carlo
MO	Molecular orbital
MPn	n-order Correction M⊘ller-Plesset Perturbation
NMR	Nuclear Magnetic Resonance
QM	Quantum Mechanics
QMC	Quantum Monte Carlo
SAPT	Symmetry-Adapted Perturbation-Theory
SCF	Self-Consistent Field
STO	Slater Type Orbital
UEG	Uniform Electron Gas
vdW	van der Waals
VMC	Variational Monte Carlo
WFT	Wave function Theory
XC	Exchange-Correlation
XRD	X-ray Diffraction
ωB97M-V	B97 functional with VV10 Nonlocal Correlation
ωB97X	B97 functional with Long- and Short-range corrected exchange

Biography

Name	Mr Qin Ken
Date of Birth	12 August 1991
Educational Attainment	Master's degree: Dalian Maritime University, Software Engineering, July, 2016 Bachelor's degree: Dalian Maritime University, Software Engineering (Japanese Enhanced), July, 2013
Publications	Maeda, T. Oshima, T. Ichiba, K. Qin, K. Muraoka, J. J. M. Vequizo, K. Hibino, S. Yamashita, K. Hongo, T. Uchiyama, K. Fujii, R. Kuriki, D. Lu, R. Maezono, A. Yamakata, H. Kato, K. Kimoto, M. Yashima, Y. Uchimoto, M. Kakihana, O. Ishitani, H. Kageyama, "Undoped Layered Perovskite Oxynitride $Li_2LaTa_2O_6N$ for Photocatalytic CO_2 Reduction with Visible Light", Angew. Chem. Int. Ed. 2018, 57(27), 8154-8158 DOI: 10.1002/anie.201803931, Wiley-VCH, 2018. K. Qin, T. Ichiba, K. Hongo, R. Maezono, "Inconsistencies in ab initio evaluations of non-additive contributions of DNA stacking energies", Chemical Physics 2019, 529, 110554, DOI: 10.1016/j.chemphys.2019.110554, Elsevier, 2019.