Structure and spectroscopy of bio- and nano-materials from first-principles simulations

Weijie Hua

(花伟杰)

Doctoral Thesis in Theoretical Chemistry and Biology
School of Biotechnology
Royal Institute of Technology
Stockholm, Sweden 2011
Structure and spectroscopy of bio- and nano-materials from first-principles simulations
Thesis for Philosophy Doctor degree
Department of Theoretical Chemistry and Biology
School of Biotechnology
Royal Institute of Technology
Stockholm, Sweden 2011

© Weijie Hua, 2011
pp i-xvi, 1-74
ISSN 1654-2312
TRITA-BIO Report 2011:5
Printed by Universitetsservice US-AB
Stockholm, Sweden 2011
To Juan, my parents, and grandparents

獻給我的祖父母、
父母、和妻子李娟
Abstract

This thesis is devoted to first-principles simulations of bio- and nano-materials, focusing on various soft x-ray spectra, ground-state energies and structures of isolated large molecules, bulk materials, and small molecules in ambient solutions.

*K*-edge near-edge x-ray absorption fine structure (NEXAFS) spectra, x-ray emission spectra, and resonant inelastic x-ray scattering spectra of DNA duplexes have been studied by means of theoretical calculations at the density functional theory level. By comparing a sequence of DNA duplexes with increasing length, we have found that the stacking effect of base pairs has very small influence on all kinds of spectra, and suggested that the spectra of a general DNA can be well reproduced by linear combinations of composed base pairs weighted by their ratio.

The NEXAFS spectra study has been extended to other realistic systems. We have used cluster models with increasing sizes to represent the infinite crystals of nucleobases and nucleosides, infinite graphene sheet, as well as a short peptide in water solution. And the equivalent core hole approximation has been extensively adopted, which provides an efficient access to these large systems. We have investigated the influence of external perturbations on the nitrogen NEXAFS spectra of guanine, cytosine, and guanosine crystals, and clarified early discrepancies between experimental and calculated spectra. The effects of size, stacking, edge, and defects to the absorption spectra of graphene have been systematically analyzed, and the debate on the interpretation of the new feature has been resolved. We have illustrated the influence of water solvent to a blocked alanine molecule by using the snapshots generated from molecular dynamics.

Multi-scale computational study on four short peptides in a self-assembled cage is presented. It is shown that the conformation of a peptide within the cage does not correspond to its lowest-energy conformation in vacuum, due to the Zn-O bond formed between the peptide and the cage, and the confinement effect of the cage.

Special emphasis has been paid on a linear-scaling method, the generalized energy based fragmentation energy (GEBF) approach. We have derived the GEBF energy equation at the Hartree-Fock level with the Born approximation of the electrostatic potential. Numerical calculations for a model system have explained the accuracy of the GEBF equation and provides a starting point for further refinements. We have also presented an automatic and efficient implementation of the GEBF approach which is applicable for general large molecules.
Preface

The work presented in this thesis was conducted during a period of about three and half years, 2007.12-2011.03, at the Department of Theoretical Chemistry and Biology, School of Biotechnology, Royal Institute of Technology, Stockholm (Sweden) and Institute of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing (P. R. China).

List of papers included in the thesis

Paper 1. Systematic study of soft-X-ray spectra of poly(dG)-poly(dC) and poly(dA) -poly(dT) DNA duplexes,
Weijie Hua, Hiroyuki Yamane, Bin Gao, Jun Jiang, Shuhua Li, Hiroyuki S. Kato, Maki Kawai, Takaki Hatsui, Yi Luo, Nobuhiro Kosugi, and Hans Ågren,

Paper 2. Refinement of DNA structures through NEXAFS analysis: applications on guanine and cytosine nucleobases, nucleosides and nucleotides,
Weijie Hua, Bin Gao, Shuhua Li, Hans Ågren, and Yi Luo,

Paper 3. X-ray Absorption Spectra of graphenes from first-principles simulations,
Weijie Hua, Bin Gao, Shuhua Li, Hans Ågren, and Yi Luo.

Paper 4. NEXAFS spectroscopies of blocked alanine in water solution from first-principles simulations,
Weijie Hua, Yue-Jie Ai, Bin Gao, Hans Ågren, and Yi Luo,
Manuscript.

Paper 5. Understanding the influence of the guest-host interaction on the conformations of short peptides in the hydrophobic cavity: a computational study,
Weijie Hua, Lina Xu, Yi Luo, and Shuhua Li,
Paper 6. Understanding the accuracy of the generalized energy-based fragmentation approach at the Hartree-Fock level by multipole expansion analysis.
Weijie Hua,
*In preparation.*

Paper 7. An efficient implementation of the generalized energy-based fragmentation approach for general large molecules.
Shugui Hua, Weijie Hua, and Shuhua Li,

**List of related papers not included in the thesis**

Paper 1. Geometry optimizations and vibrational spectra of large molecules from a generalized energy-based fragmentation approach,
Weijie Hua, Tao Fang, Wei Li, Jian-Guo Yu, and Shuhua Li,

Paper 2. The energy-based fragmentation approach for computing total energies, structures, and molecular properties of large systems at the ab initio levels, in Jeffrey R. Reimers (ed.), *Computational Methods for Large Systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*,
Wei Li, Weijie Hua, Tao Fang, and Shuhua Li,

Paper 3. Identification of two $I_h$-symmetry-breaking $C_{60}$ isomers by means of X-ray spectroscopy,
Jia-Yuan Qi, Weijie Hua, Bin Gao, Wei Wu, and Yi Luo.
*Manuscript.*

**Comments on my contributions to the papers included**

I have taken the major responsibility in the papers where I am the first author, that is, Papers 1, 2, 3, 4, 5, and 6. I have assisted in the preparation of the manuscript and part of the writing for Paper 7.
Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Yi Luo for his guidance, support, continuous trust and encouragement. I thank him for introducing me to a new and interesting research field, and for his valuable and insightful suggestions both on the academic problems and the research itself.

I am grateful to my supervisor at Nanjing University in China, Prof. Shuhua Li, for his instructions and considerable care. I appreciate the extensive scientific trainings that I have received and everything I have learnt from him.

Many thanks to head of the Department of Theoretical Chemistry and Biology, Prof. Hans Ågren, for his cooperations and for providing a stimulating atmosphere here. I am thankful to Prof. Faris Gel’mukhanov, Dr. Pawel Salek, Dr. Yaoquan Tu, Dr. Ying Fu and Dr. Olav Vahtras for helps in my study and laboratory life. I am also indebted to Dr. Bin Gao who helped me on using the x-ray spectral programs and had many in-depth discussions. I want to thank my colleges Guangjun Tian, Xin Li, Dr. Abdelsalam Mohammed, Xing Chen, Dr. Hui Cao, Yu-Ping Sun, Fuming Ying, Dr. Jun Jiang, Sai Duan, Dr. Hao Ren, Qiang Fu, Ke-Yan Lian, Xiao-Fei Li, Igor Ying Zhang, Yue-Jie Ai, Sathya S. R. R. Perumal, Dr. Arul Murugan, Jia-Yuan Qi, Yong Ma, Lili Lin and Xiao Cheng for frequent discussions, members of the support group for their helps in computers, and every former and present member in the department for friendship.

The work presented in the thesis has also benefited from the suggestions, collaborations from Prof. Jing Ma, Dr. Wei Li, Dr. Shugui Hua, Lina Xu, Yang Guo, and Dr. Zhen Yang. And I would also like to thank Dr. Hiroyuki Yamane for providing us with very exciting experimental results and cooperations.

Special thanks to Yue-Jie Ai, LiLi Lin, Bonaman Xin Li, and Yong Ma for reading the preprint and providing valuable suggestions.

I want to acknowledge the overseas study program and financial support (2007-2009) from the China Scholarship Council. The computer time via the Swedish National Infrastructure for Computing is acknowledged.

Finally, I wish to thank my wife Juan, my parents, and grandparents for their love, understanding and support.

Hope this work could evoke your pleasant memories.

Weijie Hua
Spring 2011, Stockholm
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO</td>
<td>atomic orbital</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribose nucleic acid</td>
</tr>
<tr>
<td>BE</td>
<td>binding energy</td>
</tr>
<tr>
<td>BO</td>
<td>Born-Oppenheimier</td>
</tr>
<tr>
<td>BP</td>
<td>base pair</td>
</tr>
<tr>
<td>DFT</td>
<td>density functional theory</td>
</tr>
<tr>
<td>DM</td>
<td>density matrix</td>
</tr>
<tr>
<td>ECH</td>
<td>equivalent core-hole</td>
</tr>
<tr>
<td>ESCA</td>
<td>electron spectroscopy for chemical analysis</td>
</tr>
<tr>
<td>FCH</td>
<td>full core-hole</td>
</tr>
<tr>
<td>FWHM</td>
<td>full-width at half-maximum</td>
</tr>
<tr>
<td>(G)EBF</td>
<td>(generalized) energy-based fragmentation</td>
</tr>
<tr>
<td>GNR</td>
<td>graphene nanoribbon</td>
</tr>
<tr>
<td>HB</td>
<td>hydrogen bond</td>
</tr>
<tr>
<td>HCH</td>
<td>half core-hole</td>
</tr>
<tr>
<td>HF</td>
<td>Hartree-Fock</td>
</tr>
<tr>
<td>HOMO</td>
<td>highest occupied molecular orbital</td>
</tr>
<tr>
<td>HWHM</td>
<td>half-width at half-maximum</td>
</tr>
<tr>
<td>IGLO-III</td>
<td>triple-ζ quality individual gauge for localized orbital</td>
</tr>
<tr>
<td>IP</td>
<td>ionic potential</td>
</tr>
<tr>
<td>KS</td>
<td>Kohn-Sham</td>
</tr>
<tr>
<td>LJ</td>
<td>Lennard-Jones</td>
</tr>
<tr>
<td>LUMO</td>
<td>lowest unoccupied molecular orbital</td>
</tr>
<tr>
<td>mH</td>
<td>milli-Hartree</td>
</tr>
<tr>
<td>MAE</td>
<td>maximum absolute error</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MD</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ME</td>
<td>multipole expansion</td>
</tr>
<tr>
<td>MM</td>
<td>molecular mechanical</td>
</tr>
<tr>
<td>MO</td>
<td>molecular orbital</td>
</tr>
<tr>
<td>NEXAFS</td>
<td>near-edge x-ray absorption fine-structure</td>
</tr>
<tr>
<td>ON</td>
<td>occupation number</td>
</tr>
<tr>
<td>ONIOM</td>
<td>our own n-layer integrated molecular orbital molecular mechanics</td>
</tr>
<tr>
<td>PBC</td>
<td>periodic boundary condition</td>
</tr>
<tr>
<td>PCM</td>
<td>polarizable continuum model</td>
</tr>
<tr>
<td>ps</td>
<td>pico second</td>
</tr>
<tr>
<td>QM</td>
<td>quantum mechanical</td>
</tr>
<tr>
<td>RHF</td>
<td>restricted Hartree-Fock</td>
</tr>
<tr>
<td>RIXS</td>
<td>resonant inelastic x-ray scattering</td>
</tr>
<tr>
<td>RMS</td>
<td>root mean square</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>RMSE</td>
<td>root mean square error</td>
</tr>
<tr>
<td>r.h.s.</td>
<td>right hand side</td>
</tr>
<tr>
<td>SCF</td>
<td>self-consistent-field</td>
</tr>
<tr>
<td>SE</td>
<td>Schrödinger equation</td>
</tr>
<tr>
<td>TDPT</td>
<td>time-dependent perturbation theory</td>
</tr>
<tr>
<td>TP</td>
<td>transition potential</td>
</tr>
<tr>
<td>TS</td>
<td>transition state</td>
</tr>
<tr>
<td>UHF</td>
<td>unrestricted Hartree-Fock</td>
</tr>
<tr>
<td>vdW</td>
<td>van der Waals</td>
</tr>
<tr>
<td>XANES</td>
<td>x-ray absorption near-edge structure</td>
</tr>
<tr>
<td>XAS</td>
<td>x-ray absorption spectroscopy</td>
</tr>
<tr>
<td>XCH</td>
<td>eXcited core-hole</td>
</tr>
<tr>
<td>XPS</td>
<td>x-ray photoelectron spectroscopy</td>
</tr>
<tr>
<td>XES</td>
<td>x-ray emission spectroscopy</td>
</tr>
<tr>
<td>ZPVE</td>
<td>zero-point vibrational energy</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$A$</td>
<td>vector potential</td>
</tr>
<tr>
<td>$C_m, \tilde{E}_m, E_m$</td>
<td>coefficient, total energy with and without including the electrostatic energy of background point charges of subsystem $m$</td>
</tr>
<tr>
<td>$e$</td>
<td>elementary charge</td>
</tr>
<tr>
<td>$e$</td>
<td>transcendental number</td>
</tr>
<tr>
<td>$f$</td>
<td>oscillator strength</td>
</tr>
<tr>
<td>$F$</td>
<td>scattering amplitude</td>
</tr>
<tr>
<td>$i$</td>
<td>imaginary unit</td>
</tr>
<tr>
<td>$</td>
<td>i\rangle,</td>
</tr>
<tr>
<td>$p$</td>
<td>total electronic angular momentum</td>
</tr>
<tr>
<td>$P$</td>
<td>transition probability</td>
</tr>
<tr>
<td>$Q_A$</td>
<td>net charge distribution on atom $A$</td>
</tr>
<tr>
<td>$\mathbf{R}, \mathbf{R}_A$</td>
<td>Cartesian coordinates, sum of all nuclei or of one nucleolus $A$</td>
</tr>
<tr>
<td>$\mathbf{r}, \mathbf{r}_j$</td>
<td>Cartesian coordinates, sum of all electrons or of one electron $j$</td>
</tr>
<tr>
<td>$W$</td>
<td>transition rate</td>
</tr>
<tr>
<td>$\mathbf{X} \subset \mathbf{1}$</td>
<td>inclusion complex of guest $\mathbf{X}$ and host $\mathbf{1}$</td>
</tr>
<tr>
<td>$\Gamma, \gamma$</td>
<td>half-width at half-maximum</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>spectral broadening function</td>
</tr>
<tr>
<td>$\delta$</td>
<td>energy shift or Dirac delta function</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>electronic polarization vector</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>molecular orbital energy</td>
</tr>
<tr>
<td>$\xi$</td>
<td>distance threshold</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>cross section</td>
</tr>
<tr>
<td>$\Psi, \Phi$</td>
<td>wavefunction</td>
</tr>
<tr>
<td>$\psi$</td>
<td>molecular spatial orbital</td>
</tr>
<tr>
<td>$\omega$</td>
<td>angular frequency</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>solid angle</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Overview of x-ray absorption and emission processes and corresponding x-ray spectroscopies (listed below the final state of each process). The final state of XAS (XPS) can be considered as the intermediate state for RXES (XES). Solid (dot) lines stand for occupied (unoccupied) orbitals. Electrons are colored for clear trace. .......................... 5

2.2 Schematic illustration of (a) absorption and (b) induced emission of photon. 8

2.3 Schematic illustration of linear polarized x-ray and graphene. x, y, z denote molecular orientations, is the electronic polarization vector and can be represented as (1, θ, ϕ) in spherical coordinates. .......................... 11

2.4 Energy level occupation of various core-hole states, with the nuclear charge of the excited atom shown below each state. Solid (dot) lines stand for occupied (unoccupied) orbitals. Electrons from the core and valence levels of the ground state are colored by red and blue. See text for detail. .......................... 13

2.5 Schematic illustration of spontaneous emission of photon. .......................... 18

2.6 Depiction of RIXS process by (a) electronic wavefunctions; and (b) molecular orbitals of the final state. .......................... 20

3.1 Two strategies to treat large systems: (a) using a hybrid way; (b) using linear scaling technique (DFT methods taken as an example). .......................... 23

4.1 Structural illustration of DNA: (a) Side and (b) front views of a B-DNA duplex fragment; (c) deoxyribonucleotide unit; (d-e) two base-pairing patterns with imine (=N-) and amine (-NH-) nitrogens labeled. .......................... 30

4.2 Comparison of N 1s NEXAFS spectra (contribution from one nitrogen atom) in (gc)2 calculated from two models: (a) The spectra; models (b) with and (c) without the phosphate groups. Hydrogens without forming HBs are hid for clarity, and arrows denote the excited atoms. .......................... 30

4.3 N1s x-ray spectra of poly(dG)-poly(dC) DNA duplexes: (a) NEXAFS (b) normal XES, and (c) RIXS spectra calculated from different sizes of (gc)n models compared with experiments; (d) experimental XES spectra of GC-DNA film. Calculated NEXAFS spectra are shifted by +0.2 eV for better comparison with experiment[1]; calculated XES spectra are shifted by +10 eV to match with experiment (incident photon energy hν=410 eV); calculated RIXS spectra are shifted by +10 eV to match with experiment. Selected from Paper 1, reprinted with permission from American Chemical Society. .......................... 32

4.4 Selected oligomer configurations of guanines: (a-c) dimers; (d) tetramer. 34
LIST OF FIGURES  

4.5 N1s NEXAFS spectra and structures of guanine G7H isomer. (a) Cluster model used for calculation as generated from the crystal structure. Spectra of anhydrous guanine crystal calculated by ECH method at the crystal structure (0%) and at geometries with the distances between the central and all neighboring molecules scaled (±1%, ±3%), compared with experiment. (c) Gas phase spectrum calculated by FCH method (shifted by +0.57 eV for comparison with experiment). Selected from Paper 2, reprinted with permission from American Chemical Society.

4.6 Calculated Δ values (Δ denotes the energy difference of the amine and imine 1s → π* peaks) of crystals of (a) cytosine, (b) guanine, and (c) guanosine with different models vs different scaled geometries. Experimental values are also shown as horizontal dashed lines. Selected from Paper 2, reprinted with permission from American Chemical Society.

4.7 Comparison of N1s NEXAFS spectra of cytosine monohydrate calculated with different functionals. (a) Cluster model used for calculation. Spectra of cytosine monohydrate crystal calculated by ECH method using (b) B3LYP and (c) M062X functionals at the crystal structure (0%) and at geometries with the distances between the central and all neighboring molecules scaled (±1%, ±3%), compared with experiment. Selected from Paper 2, reprinted with permission from American Chemical Society.

4.8 (a) Illustration of Ice-Ih crystal structure. (b) O K-edge NEXAFS spectra of bulk ice-Ih with one particular hydrogen order with respect to system size (characterized by ξ, and number of waters are included in parenthesis). Inset: cluster model at ξ = 10.0 Å. (c) Comparison of our converged result with experimental result by Wernet et al. and calculation results by Chen et al., Prendergast et al., Wernet et al., and Leetmaa et al.

4.9 O K-edge NEXAFS spectra of Ice-Ih with 16 kinds of hydrogen order compared with experimental result by Wernet et al. Calculations are based on cluster models constructed with ξ = 7.0 Å.

4.10 Depiction of ideal infinite graphene sheet with supercell labeled.

4.11 [Reprinted with permission from American Physical Society.] (a) Experimental C K-edge NEXAFS spectra of (a) graphene and pristine graphite; (b) FLG around 283.7 eV; and (c) FLG around 288 eV. Characteristic peaks are labeled by arrows and numbers in eV. Selected from Paper 3, reprinted with permission from American Physical Society.

4.12 C K-edge NEXAFS spectra of graphene: (a) three graphene nanoribbons (GNRs) of different size: Size A, Size B and Size C; (b) 1-, 2-, and 3-layers (1L, 2L, 3L) GNRs based on Size A cluster; (c) atom-specific spectra of 1L GNR with Size B as excited from central, outer, and edge regions; (d) Stone-Wales defected GNR with Size A. Left and middle panels illustrate 1s→π* and →σ* transitions, respectively. Panels a3 [with FWHM=0.2 eV (lower) and FWHM=0.8 eV (higher)] and b3 (with FWHM=0.8 eV only) provide fine structures in the first π* peak. Panels c3 and d3 denote the structure of GNR used for calculations with the excited centers labeled. Selected from Paper 3, reprinted with permission from American Physical Society.
List of Figures

4.13 Ground-state DOS of ideal one-layer graphene from different density functionals. (a) B3LYP for Size C cluster with (bottom continuous curve) and without (bottom dash curve) edge states, LDA (LDA-1\cite{13} and LDA-2\cite{14}), and GGA (GGA-1\cite{15} and GGA-2\cite{16}) results. Energies are relative to \(E_F\) (absolute value -3.7 eV) and Lorentzian with FWHW=0.2 eV is used for convoluting B3LYP results. (b) Enlarged B3LYP results above \(E_F\). Dominant spectral peaks are labeled by star or daggers. Selected from Paper 3, reprinted with permission from American Physical Society.

4.14 Spin-specific contributions to the C 1s\(\rightarrow\)\(\pi^*\) NEXAFS spectra of GNRs with different sizes: (a) 2.4 nm \(\times\) 1.5 nm; (b) 5.0 nm \(\times\) 4.4 nm. Calculations are run at the UB3LYP level.

4.15 Size dependence of N1s NEXAFS spectra of snapshot 1 calculated with the (a) ECH and (b) FCH methods. Each size is labeled by the distance threshold \(\xi\) in Å, and number of water is included in parenthesis. Structures of selected models are shown around the spectra. Bars denote the IP values.

4.16 Atom-specific N1s NEXAFS spectra calculated with ECH method using models constructed with thresholds: (a) 0.0 Å; (b) 2.6 Å; (c) 10.0 Å. Selected bars are labeled by numbers and corresponding final-state MOs are visualized in panel (d).

4.17 Statistics of N1s NEXAFS spectra calculated with the ECH method based on models constructed at \(\xi=7.0\) Å: (a) 10 randomly chosen geometries; (b) averages over all 100 snapshots (solid), snapshots 1-50 (dash-dot), and snapshots 51-100 (dot). (c) N1s NEXAFS spectra calculated with ECH-PCM method with different size of models (each characterized by \(\xi\) in Å, and number of waters is included in paracentesis) based on snapshot 1. Spectra have been shifted for better comparison with the statistical result.

4.18 a) Structural illustration of peptides 2-5. b) Full view of the crystal structure of 2cryc1. Among six Pd atoms and three Zn atoms in the cage 1, only one Zn atom forms a Zn-O bond with the peptide in the crystal structure. Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

4.19 B3LYP and MP2//B3LYP energies at B3LYP-optimized structures for several low-lying conformers of each peptide in the gas phase. Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

4.20 B3LYP-optimized structure of 2oe and ONIOM(B3LYP:PM6)-optimized structure of 2oe2\subset1. Only the Zn atom is shown for the cage and motif turn types are labeled. Distances are labeled in Å, and total number of HBs is listed in parentheses after the name of each species. Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

4.21 Schematic illustration for (a): a two-fragment model system H\(_3\)O\(^+\)\(\cdots\)OH\(^-\); and (b-c) its two subsystems. The fragments containing real atoms are labeled as I or J, and those containing point charges (denoted by “+” and “−”) are labeled by I’ or J’.
## Contents

Abbreviations ix
Symbols xi
List of Figures xii

1 Introduction 1

2 Soft x-ray spectroscopies 4
  2.1 NEXAFS 6
    2.1.1 Theory 6
    2.1.2 Basic concepts and formulas 9
    2.1.3 Approximations 11
    2.1.4 Spectral calibration and broadening 15
  2.2 XPS 17
  2.3 XES 18
  2.4 RIXS 19

3 Ground-state energies and structures for large systems 22
  3.1 The hybrid approaches 22
  3.2 Linear-scaling methods 24
    3.2.1 Standard linear-scaling methods 24
    3.2.2 Fragment-based methods 25

4 Complex systems: challenges, efforts, and achievements 28
  4.1 DNA 29
  4.2 Graphene 41
CONTENTS

4.3 Short peptides ........................................... 47
4.4 Multipole expansion .................................... 57

5 Summary of included papers .......................... 60
  5.1 Soft x-ray spectroscopies ............................ 60
      5.1.1 DNA and related molecules, Papers 1-2 .... 60
      5.1.2 Graphene, Papers 3 ............................... 61
      5.1.3 Solvent effect of peptide, Paper 4 .......... 61
  5.2 Ground state energies and structures ............. 62
      5.2.1 Host-guest system, Paper 5 ................. 62
      5.2.2 Multipole expansion, Paper 6 ............... 62
      5.2.3 Automatic fragmentation implementation, Paper 7 62

References ................................................. 64
Chapter 1

Introduction

Bio- and nano-materials represent two categories of systems that have attracted much attention from researchers due to their ongoing and potential medical and industrial applications. Experimentally, with the development of biotechnology and nanoscience, exceptional opportunity in characterization and design of new materials is provided. Theory and modelling are another indispensable part that can give information inaccessible to the experiments. Since its birth in the 1920s, various quantum mechanical (QM) methods have been widely used to probe the energies, structures, spectroscopies and various properties of small- and medium-sized systems. Especially, the development of first-principles methods since the 1960s provides an accurate solution without losing efficiency. However, as the bio- and nano-materials usually have large size, the extension of first-principles simulations to them is not a trivial task.

The electronic structure is the most fundamental property of molecules or materials. One can however not “see” it directly but can only “reflect” it through the external electric, optical, magnetic interactions. X-ray techniques have been effective ways for such an investigation since its production and detection by Wilhelm C. Röntgen in 1895. X-ray can be generated by synchrotron radiation or x-ray tubes. Its usage generally falls into two categories: x-ray diffraction and x-ray spectroscopy. Diffraction technique is first introduced by William L. Bragg and William H. Bragg, which is based on the elastic scattering of x-ray photons from the electrons of the target system. And it has been widely used in determining the crystal structures of inorganic or biological solids. While the early spectral usage started from the work by Manne Siegbahn, which is based on the electronic transitions.

Soft (long-wavelength) x-ray spectroscopies have become efficient ways to
probe the electronic and chemical structure of molecules, surfaces, and bulk materials due to its element-selectivity and environment-sensitivity. The physics behind is that the interaction of light with electrons causes excitation or de-excitation of the system; and different electronic structure corresponds to a characteristic profile of absorption, emission, or scattering. Not only covalent bonding, but also weak interaction, can have fingerprints in soft-x-ray spectra. Especially, the weak interaction including the hydrogen bonding and $\pi-\pi$ stacking, plays a significant role in large molecules or materials due to the cumulative effect.

Theoretically, the interaction of light adds a perturbation to the molecule, and these phenomena can be described by time-dependent perturbation theory when the electromagnetic field is week. Approximations must be made for computations, which is even crucial for large molecules. Two important approximations which simplify the calculations are the final state rule and the dipole approximation. The former one states that the problem can be well treated by using only the final state, without knowing information of the initial one. This turns a multi-electron problem into a single-electron one, thus the transitions among states are represented by those among molecular orbitals of the final state. While the latter approximation avoids an exponential term in the transition strength by only the leading term in its Taylor-expanded series, which is specially applicable for light with long wavelengths.

On the other hand, it is yet a challenging area even for the ground-state of large systems. The high scaling of the conventional QM method with respect to the number of atoms in the system is the bottleneck of the problem. One simple and direct solution is the hybrid way that employs the accurate QM method only for the important region of system under study, and an economic method, usually semi-empirical or force-field method, is used for the other parts. Such a treatment is suitable for chemical problems which is restricted in a small region (e.g. enzyme-catalyzed reactions), but it is not applicable for problems involving global geometrical changes (e.g. folding of proteins).

Another solution involves in the development of linear-scaling methods, which can be used for more general problems. Linear-scaling means that the computational time scales only linearly with the system size. The underlying physical principle is the “nearsightedness” of the density matrix and/or fast-decay of the electron correlation. In the last few decades, these methods have been well developed and various methods emerge.

One branch of the linear-scaling methods consists of various molecular-fragment-based methods. In these methods, a target large molecule is cut into pieces, and
the energies can be obtained from the density matrices and/or energies of various subsystems (each constructed from a few neighboring pieces). There is a great need to extend these methods for more general use, including to a more general molecule with complex structure and to more properties.

This thesis is devoted to the first-principles simulations of bio- and nano-systems, focused on the soft x-ray spectroscopies as well as ground state energies and structures. And representative biomolecules (DNA, nucleobases, peptides) and nano-materials (graphene, self-assembled cages) will be our target molecules. In the following chapters, the theory and computational methods for soft-x-ray spectra (Chapter 2) and ground-state energies/structures (Chapter 3) are first outlined. Then, case studies are provided in Chapter 4. Finally, a summary of included papers is provided in Chapter 5.
Chapter 2

Soft x-ray spectroscopies

X-ray is electromagnetic wave with wavelength $\lambda \sim 0.01-10$ nm or photon energy $\hbar \omega \sim 10^2-10^5$ eV that associates with the inner-shell excitation of electrons. A variety of x-ray techniques have been developed and applied to probe the electronic structure and chemical composition of molecules, surfaces, and bulk materials. The low-energy part (0.1-10 nm or $10^2-10^4$ eV) is usually called soft x-ray, and single core-electron is believed to be excited to a high level or knocked out from an atom; while the high-energy part (0.01-0.1 nm or $10^4-10^5$ eV) is termed as hard x-ray, which can excite multi electrons or even knock all the electrons out of an atom. This chapter deals only with soft x-ray that is more related to the chemical structures of molecular and bio-systems. Various absorption, emission, and scattering processes as well as corresponding spectral techniques will be discussed.

Figure 2.1 illustrates an overview of selected x-ray processes. When a photon with frequency $\omega$ is absorbed by a molecule in the ground state, a core electron can be excited to a virtual orbital or ejected to vacuum. These processes can be used to measure the x-ray absorption spectroscopy (XAS) or x-ray photoelectron spectroscopy (XPS). These states with the existence of a core hole usually have short life time. An electron from the valence level can fall to the core level and lower the energy. Simultaneously, an photon with frequency $\omega'$ can be emitted if radioactive emission happens. One could accordingly measure the resonant x-ray emission spectroscopy (RXES) or non-resonant x-ray emission spectroscopy (XES). Note that in the former case, the absorption and emission is actually a concerted process, and the spectra are classified as resonant elastic ($\omega' = \omega$) and inelastic ($\omega' \neq \omega$) scattering (RIXS) spectra according to the energies of the in- and out-photons.
In Sections 2.1-2.4, we will summarize the theory and calculation methods for XAS, XPS, XES, and RIXS spectra, respectively.

A few terminologies need to be mentioned. For XAS, we only study the near-edge part, and it is considered the same as the near-edge x-ray fine structure (NEXAFS) or the x-ray absorption near-edge structure (XANES) spectrum. XPS is also known as electron spectroscopy for chemical analysis (ESCA) invented by Nobel Laureate Kai Siegbahn. RIXS spectroscopy sometimes is also termed as resonant Raman x-ray scattering spectroscopy or resonance x-ray fluoresces spectroscopy.

Figure 2.1  Overview of x-ray absorption and emission processes and corresponding x-ray spectroscopies (listed below the final state of each process). The final state of XAS (XPS) can be considered as the intermediate state for RXES (XES). Solid (dot) lines stand for occupied (unoccupied) orbitals. Electrons are colored for clear trace.
2.1 NEXAFS

NEXAFS technique was developed several decades ago, aiming to probe the structure of molecules in surface, especially for the light elements.\textsuperscript{[17]} Nowadays, it has been widely used in surfaces, bulk solids, liquids, or gas-phase molecules to map out the electronic structure, especially the character of the virtual orbitals, as well as the local chemical structure. It is necessary to mention that the absorption spectroscopy is actually measured by the decay process. Because of the high energy due to the existence of core-hole, the system tends to lower its energy. A valence-shell electron (i.e., Auger electron) can be emitted to vacuum; and/or a valence electron will fall down to the core level and emit fluorescence. Then experiments can measure the Auger electron or the fluorescence yields, respectively.

The purpose of this section is to present the calculation methods of absorption spectra from theory to practice. Theoretically, the absorption of x-ray photon can be solved in the framework of time-dependent perturbation theory (TDPT), if a weak electromagnetic field is satisfied. We will first give a short summary of TDPT in general and its application for soft x-ray absorption in particular. Then in Section 2.1.2, the basic concepts and formulae are given. Next, various practical approximation methods are discussed in Section 2.1.3. Finally, post treatments of raw spectra are discussed in Section 2.1.4.

2.1.1 Theory

Time-dependent perturbation theory

TDPT provides a way to solve the time-dependent Schrödinger equation (SE) with a total Hamiltonian composed of two parts,

\[ \mathcal{H} = \mathcal{H}_0 + \mathcal{V}(t); \quad \mathcal{H}_0|j\rangle = E_j|j\rangle, \quad (2.1) \]

where \( \mathcal{H}_0 \) is the zero-order Hamiltonian (with the eigenvectors \( |j\rangle \)'s and eigenvalues \( E_j \)'s already solved) and \( \mathcal{V}(t) \) denotes the time-dependent perturbation. This solution is often convenient to be done in the interaction picture, and the SE is given by,

\[ \hbar \frac{\partial}{\partial t}[\alpha, t_0; t]_I = \mathcal{V}_I[\alpha, t_0; t]_I, \quad (2.2) \]
where $|\alpha, t_0; t\rangle_f$ denotes a ket in the interaction picture which is $|\alpha\rangle$ at $t_0$.[18] Further expanding it in the basis of $\{|j\rangle\}$,

$$|\alpha, t_0; t\rangle_f = \sum_j c_j(t)|j\rangle,$$

the problem that leaves is to find the expansion coefficients $\{c_j\}$, which can be achieved by using the method of Dyson series.[18] Considering that initially only one state $|i\rangle$ is populated, the transition probability for $|i\rangle \rightarrow |f\rangle$ ($f \neq i$), is given by,

$$P_{fi} = \left| \sum_{k=1}^{\infty} c_j^{(k)}(t) \right|^2.$$

For practical applications, a truncated solution at low order is often used.

**Fermi’s golden rule**

Particularly when describing the interaction between the light and the molecule, $\mathcal{H}_0$ represents the molecular Hamiltonian, and $\mathcal{V}(t)$ is the time-dependent interaction between light (plane electromagnetic wave assumed) and electrons of the molecule given by

$$\mathcal{V}(t) = \frac{e}{mc} A(t) \cdot p,$$

$$A(t) = \epsilon A_0 \cos (k \cdot x - \omega t) = \epsilon \frac{A_0}{2} \left( e^{i(k \cdot x - \omega t)} + e^{-i(k \cdot x - \omega t)} \right).$$

Here $A(t)$ stands for the vector potential of light and Coulomb gauge, $\Delta \cdot A = 0$, is chosen; $\epsilon$, $k$, and $\omega$ denote the polarization vector, the wave vector, and the angular frequency, respectively; and $p = \sum_{j=1}^{N} p_j$ is the total linear momentum of all the electrons (each with mass $m$ and charge $-e$) in the molecule. If only up till the first-order perturbation is considered, i.e., $P_{fi} = |c_f^{(1)}(t)|^2$, Fermi’s golden rule gives an estimate of the transition rate (transition probability per time) from initial state $|i\rangle$ to final state $|f\rangle$:

$$W_{fi}(\omega) = \frac{d}{dt} P_{fi} = \frac{2\pi}{\hbar} |V_{fi}|^2 \delta(E_f - E_i \pm \hbar \omega),$$

where $V_{fi} = \langle \Phi_f | \mathcal{V} | \Phi_i \rangle$. 

Dipole approximation

For soft x-ray, \( k \cdot x \ll 1 \) can be satisfied, one can expand the \( e^{ikx} \) term in Taylor series as \( e^{ikx} = 1 + ik \cdot x + \cdots \), and choose the leading term which makes the major contribution to the transition strength. This is often termed as electric dipole approximation and correspondingly the approximated perturbation (known as the electric dipole Hamiltonian) and its matrix elements are given by

\[
\mathcal{V}(t) = e \cdot p \frac{e A_0}{2mc} (e^{-i\omega t} + e^{i\omega t}),
\]

(2.8)

\[
V_{fi} = e \cdot (\Phi_f | p | \Phi_i) \frac{e A_0}{2mc} (e^{-i\omega t} + e^{i\omega t}).
\]

(2.9)

Using the relation \([18,19]\)

\[
\langle \Phi_f | p | \Phi_i \rangle = im\omega \langle \Psi_f | r | \Psi_i \rangle,
\]

(2.10)

where \( r = \sum_{j=1}^{N} r_j \) is the summation of all electronic coordinates, eq 2.9 can be rewritten within the position representation as

\[
V_{fi} = e \cdot (\Psi_f | r | \Psi_i) \frac{ie A_0}{2c} \omega (e^{-i\omega t} + e^{i\omega t}).
\]

(2.11)

Inserting eq 2.11 to eq 2.7, the transition rate is given by

\[
W_{fi}(\omega) = \frac{\pi e^2 A_0^2 \omega^2}{2\hbar c^2} |e \cdot (\Psi_f | r | \Psi_i)|^2 \delta(E_f - E_i \pm \hbar\omega).
\]

(2.12)

The negative and positive signs in the above equation correspond to the photon absorption and induced (i.e., stimulated) emission, respectively, as schematically illustrated in Figure 2.2. In the following, we only consider the absorption case.

![Figure 2.2](image_url) Schematic illustration of (a) absorption and (b) induced emission of photon.
2.1 NEXAFS

2.1.2 Basic concepts and formulas

Absorption cross section and oscillator strength

Two important concepts to describe the absorption are cross section and oscillator strength. The absorption cross section $\sigma_{\text{abs}}$ for transition from initial state $|\Psi_i\rangle$ to final state $|\Psi_f\rangle$ is defined as the transition rate per unit photon flux:\[17]\[2.13]

$$\sigma_{\text{abs}} = \frac{W_{fi}}{F_{\text{ph}}}, \quad F_{\text{ph}} = \frac{A_0^2 \omega}{8\pi\hbar c}.$$ 

Here $F_{\text{ph}}$ is the photon flux (number of incident photons per unit time per unit area), and thus $\sigma$ has the unit of area. Within the dipole approximation, eq 2.13 is expressed as

$$\sigma_{\text{abs}}(\omega) = 4\pi^2 \alpha \omega |\mathbf{e} \cdot \langle \Psi_f | r | \Psi_i \rangle|^2 \delta(\omega_{fi} - \omega),$$ \[2.14\]

where $\omega_{fi} \equiv (E_f - E_i)/\hbar$, and $\alpha \equiv e^2/(\hbar c) \approx 1/137$ is the fine-structure constant.

The integration of eq 2.14 over all frequency range leads to

$$\int d\omega \sigma_{\text{abs}}(\omega) = 4\pi^2 \alpha \omega_{fi} |\mathbf{e} \cdot \langle \Psi_f | r | \Psi_i \rangle|^2,$$

$$= \frac{2\pi^2 e^2}{mc} f_{fi},$$ \[2.15\]

where $f_{fi}$ is the optical oscillator strength defined as

$$f_{fi} \equiv \frac{2m\omega_{fi}}{\hbar} |\mathbf{e} \cdot \langle \Psi_f | r | \Psi_i \rangle|^2.$$ \[2.16\]

The optical oscillator strength is a dimensionless quantity which is convenient for practical use. And the sum of oscillator strength from state $|\Psi_i\rangle$ to all possible final states is equal to the number of electrons,

$$\sum_f f_{fi} = N,$$ \[2.17\]

which is known as the Thomas-Reiche-Kuhn sum rule. Note that when considering continuum states, the summation is changed to integration for the region above the ionic potential.
Final-state rule and single-particle approximation

According to the final-state rule\textsuperscript{[20-23]}, the accurate absorption spectra of finite molecule systems could be obtained by only the final-state wavefunction instead of knowing both. This is because the 1s orbital is well separated from the other ones, and they behave similar in both the initial and final states. The electron transition process can be approximated by a single-electron picture, when the process is sudden compared to the relaxation time of the other passive electrons. Thus, for single-determinants \( |\Psi_{i,f}\rangle \), we have

\[
\langle \Psi_f | r | \Psi_i \rangle \approx S \langle \psi_f(1) | r_1 | \psi_i(1) \rangle ,
\]

(2.18)

where \( |\psi_{i,f}\rangle \) are a pair of final-state molecular orbitals involved in the transition. As the co-determinantal factor \( S \) does not change much for different transitions (\( S \approx 0.7-0.9 \textsuperscript{[24]} \)), the oscillator strength can be simply calculated as

\[
f_{fi} = \frac{2m\omega_{fi}}{\hbar} |\epsilon \cdot \langle \psi_f | r_1 | \psi_i \rangle|^2 .
\]

(2.19)

Now \( \omega_{fi} \) has the meaning of orbital energy difference, \( \omega_{fi} = (\epsilon_f - \epsilon_i)/\hbar \).

Polarization

For molecules in the gas, solution, polycrystal, or amorphous phases, random orientation can be assumed. The interaction of linear-polarized light with the ensemble of molecules in different configurations should be considered in a statistical way. Averaging over the \( x, y, \) and \( z \) directions provides a simple estimation of the averaged absorption oscillator strength:

\[
f_{fi} = \frac{2m\omega_{fi}}{3\hbar} \left( |\langle \psi_f | x | \psi_i \rangle|^2 + |\langle \psi_f | y | \psi_i \rangle|^2 + |\langle \psi_f | z | \psi_i \rangle|^2 \right) .
\]

(2.20)

While for molecules that are orderly aligned, for example, layered materials, mono-crystals, there will be evident angular dependence of the oscillator strength. A good example is a two-dimensional material graphene as shown in Figure 2.3. This system (placed in the \( x - y \) plane) has evident anisotropy: the \( \pi^* \) orbitals are parallel to the \( z \) direction, while the \( \sigma^* \) orbitals spread in the molecular plane. Thus, for a given \( \theta \) (angle between \( z \) and \( \epsilon \)), the \( \sigma^* \) and \( \pi^* \) transitions can be activated with different strengths. Experimentalists can adjust the incident angle to strengthen a specific transition. While theoretically we can obtain the pure \( \sigma \)}
2.1 NEXAFS

and π transitions by using the following equations:

\[ f_{fi}^\pi = \frac{2m\omega_{fi}}{3\hbar}|\langle \psi_f | z | \psi_i \rangle|^2, \]  

\[ f_{fi}^\sigma = \frac{2m\omega_{fi}}{3\hbar} \frac{1}{2} \left( |\langle \psi_f | x | \psi_i \rangle|^2 + |\langle \psi_f | y | \psi_i \rangle|^2 \right). \]  

Here the σ transition is simply approximated as an average over \( x \) and \( y \).

2.1.3 Approximations

As most NEXAFS calculations nowadays are performed with Kohn-Sham (KS)\(^{[25]}\) density functional theory (DFT)\(^{[26]}\), and the concept of occupation number (ON), especially fractional ON, is involved in different approximations, we first summarize the KS-DFT as well as Janak’s theorem\(^{[27]}\), and then discuss various approximations for NEXAFS calculations.

Density functional theory and Janak’s theorem

In KS-DFT, the total energy is given by

\[ E[\rho(r)] = T_s[\rho] + U[\rho] + E_{XC}[\rho], \]  

\[ \rho(r) = \sum_{i=1}^{N} |\psi_i(r)|^2. \]
where $T_s$ denotes the kinetic energy of the reference system, $U$ and $E_{\text{XC}}$ represent the Coulomb and exchange-correlation energies, respectively, $\rho(\mathbf{r})$ is the electronic density, and $\psi_i$'s are KS orbitals, which are obtained from solving a set of single-electron KS equations,

$$\hat{f}_{\text{KS}}|i\rangle = \varepsilon_i|i\rangle.$$  \hspace{1cm} (2.25)

Janak’s theorem extends the density expression by introducing occupation numbers $\{n_i\}$ as parameters, i.e.,

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} n_i |\psi_i(\mathbf{r})|^2,$$  \hspace{1cm} (2.26)

and proves that the KS orbital energy of orbital $i$ can be written as the derivative of a generalized KS energy $\tilde{E}$ with respect to $n_i$:

$$\varepsilon_i = \frac{\partial \tilde{E}}{\partial n_i},$$  \hspace{1cm} (2.27)

$$\tilde{E}[\rho(\mathbf{r})] = \tilde{T}_s[\rho] + U[\rho] + E_{\text{XC}}[\rho].$$  \hspace{1cm} (2.28)

Here in $\tilde{E}$, a set of $\{n_i\}$ is also introduced in the kinetic energy part, i.e., $\tilde{T} = \sum_i^{N} n_i \langle i \mid -\frac{1}{2} \nabla_i^2 \mid i \rangle$. In many places, Janak’s theorem is also named as Slater-Janak’s theorem. Because historically Slater first introduced this concept in the $X\alpha$ method,\cite{28,29} while Janak first brought this into the DFT level.

**Approximation methods in NEXAFS calculations**

From above we have known that one only needs the electronic structure of the core-excited state to calculate a specific transition. For all possible transitions in the whole energy range, a state-by-state calculation over all possible pair of initial and final states is of course right. However, for large systems, this becomes not practicable as there are a large number of orbitals and consequently even more possible transitions are involved. An economic way is to find a reference electronic configuration which could better represent the characters of all transitions, so that they can be generated from single electronic-structure calculation. Various approximations have been developed to define the reference state. Here we summarize the full core-hole (FCH), excited core-hole (XCH), equivalent core-hole (ECH) and half core-hole (HCH) approximations, as schematically depicted in Figure 2.4.

1. Full core-hole approximation
2.1 NEXAFS

![Energy level occupation of various core-hole states](image)

Figure 2.4  Energy level occupation of various core-hole states, with the nuclear charge of the excited atom shown below each state. Solid (dot) lines stand for occupied (unoccupied) orbitals. Electrons from the core and valence levels of the ground state are colored by red and blue. See text for detail.

The FCH approximation takes the core-ionized state as the reference state for spectra calculations. This is because the excited electron (which now appears in the outer shell) has relatively small influence compared to the core-hole, and thus can be neglected. The electronic structure is obtained by optimizing the MO coefficients with the constraint $n_i = 0$ ($i$ is the core orbital involved). The final spectra could be obtained by the static-exchange (STEX) method$^{30,31}$. For practical calculations with a Gaussian basis set, usually the triple-$\zeta$ quality individual gauge for localized orbital (IGLO-III) basis set$^{32}$ is chosen for the excited atom. This method has been implemented in the StoBe$^{33}$ package. While for calculations employing plane-wave basis set with pseudo-potentials (usually for periodic system), this is realized by using a FCH potential.$^{34}$

2. Excited core-hole approximation

This approximation is first introduced by Prendergast and Galli$^{35}$ in 2006 when studying the spectra of ice and liquid water. Compared with the FCH approximation, an electron in the LUMO is also considered$^{36}$. This method is developed based on the planewave basis sets and the core-excited state is represented by modifying the pseudopotential of the excited atom. It is also straightforward to be used for a Gaussian basis set. It has been applied to study the spectra of peptides in solution$^{37}$ and molecular crystal$^{38}$, as well as the influence of nuclear vibrational effects$^{39,40}$.

3. Equivalent core-hole or $Z + 1$ approximation
The ECH approximation\textsuperscript{[\textit{41--44}] is based on the intuition that the effect with a disappeared core-electron can be modulated by increasing the nuclear charge by one ($Z + 1$). This can be simply understood by the concept of effective nuclear charge of a valence electron $i$:

$$Z_{\text{eff}} = Z - s.$$  (2.29)

Here $Z$ is the nuclear charge and $s$ denotes the shielding from the electrons between the nucleolus and electron $i$. The decrease of $s$ and increase of $Z$ has a similar effective charge, as seen by electron $i$. A more rigorous explanation is that the increased nuclear charge models a similar electronic potential for the core-hole state.\textsuperscript{[\textit{44}]}

The implementation of the ECH concept for NEXAFS and XPS calculations are first introduced by Ågren and coworkers.\textsuperscript{[\textit{44}] Practically, the core hole effect is approximately represented by replacing the excited atom with the next element in the periodic table, while the number of electrons remains unchanged (thus, for a common molecule which is neutral, closed-shell, in the ground state, the total charge is +1 for the new species). For example, the ground state of $\text{C}_{59}\text{N}^+$ can be used to model the core-hole state of $\text{C}_{60}$ while studying its $\text{C}1s$ NEXAFS spectrum. Note that for the excited atom (now as N), the IGLO-III basis set\textsuperscript{[\textit{32}] is usually used.

As the electronic structure is solved in the ground-state of the new species within the ECH method,\textsuperscript{[\textit{44}] it is efficient and valid for practical usage in large systems. This method has been successfully used for various isolated systems, such as phthalocyanine\textsuperscript{[\textit{45]}, DNA duplexes\textsuperscript{[\textit{46]}, fullerene\textsuperscript{[\textit{47]}, single-walled carbon nanotubes\textsuperscript{[\textit{48]}, graphene\textsuperscript{[\textit{49}, as well as small molecules in ambient conditions.\textsuperscript{[\textit{50]}} This method has been implemented in the Bio-Nano LEGO package\textsuperscript{[\textit{51}] which calculates the spectra based on the MO coefficients and multipole integrals read from the GAUSSIAN program\textsuperscript{[\textit{52].}

4. Half core-hole approximation

The HCH (or transition potential, TP) approximation by Triguero \textit{et al.}\textsuperscript{[\textit{53}] is a simplification of Slater’s transition state (TS) method.\textsuperscript{[\textit{28,29]} Both method tries to model an intermediate state (half occupation of electron in the core level) which represents the characters of both the initial and final states.

Especially, the TS method also considers a half occupation of electron in the virtual level. The energies of both the initial and final states are considered
as functions of \( n_i \) and \( n_f \), and Taylor expanded at \((n_i, n_f) = (\frac{1}{2}, \frac{1}{2})\) over the two variables. The transition energy, difference of the initial and final-state energies, has cancelation of the second-order terms. Thus, truncation at this order gives only simple mathematical forms \([9]\) and is often used.

While the HCH method avoids the state-by-state requirement within the TS method by only considering the half-electron in the core state. \([9]\) The Taylor expansion is now executed over only one variable \( n_i \) at \( n_i = \frac{1}{2} \), and the transition energy between states \( |i\rangle \) and \( |f\rangle \) can be calculated as \( \varepsilon_f(n_i)|_{n_i=\frac{1}{2}} - \varepsilon_i(n_i)|_{n_i=\frac{1}{2}} \), which still considers to relaxation up till the second order.

Besides, there are also other methods to model the spectra, for example, the multiple scattering approach \([54,55]\).

### 2.1.4 Spectral calibration and broadening

Post treatments are necessary in order to account for the approximations made in theoretical calculations and for better comparisons with experimental data as well. Calibration of transition energy and broadening of spectral lines are often made.

**Energy calibration**

The \( \Delta \)SCF (or \( \Delta \)KS for Korn-Sham DFT)\([56,57]\) which fully considers the relaxation is often used as a standard procedure for accurate transition-energy calculations. Although a state-by-state approach is computationally expensive, we can however do this for only one transition, and obtain a shift value \( \delta \) which is assumed to be constant for all other transitions. Usually, the transition to the LUMO is used for the \( \Delta \)SCF calculation. For example, for a \( K \)-edge transition, the accurate transition energy is given by

\[
\varepsilon_{1s-\text{LUMO}}^{\Delta \text{SCF}} = N E_{1s-\text{LUMO}} - N E_{\text{GS}}.
\]  

(2.30)

By comparing with the calculated \( 1s \rightarrow \text{LUMO} \) transition energy using the previous mentioned approximations, one can get the shift value as \( \delta = \varepsilon_{1s-\text{LUMO}}^{\Delta \text{SCF}} - \varepsilon_{1s-\text{LUMO}}^{\text{calc}} \). Then the whole raw spectra is shifted by adding this \( \delta \) for calibration.

Our previous discussions are all within the non-relativistic framework. It is however necessary to account for the relativistic effect associated with the removal of one electron from the core level. Global shift of the spectrum by a small constant
value avoids solving the equations within the relativistic framework and provides
the simplest treatment. For example, for N K-edge XAS, a shift value of +0.3 eV
is often used.

Calibration to experiment peak directly is also another option, which is usu-
ally used when an accurate energy is not applicable. In practice, one can choose
the first clear peak in both the calculated (and also broadened, see below) and
experimental profiles for the alignment. Such a peak usually exists for the low-Z
elements.

Underestimation of bandgap is common in some DFT methods, which can
narrow the calculated spectrum compared with experimental one.\(^8\) In such cases,
a scaling of energy might also provides a calibration for the spectra profiles.

**Line broadening**

The calculated bar spectrum needs to be broadened for many thermal, mechanical,
or optical reasons, and the lack of vibronic coupling in the calculated spectrum is
the major reason. The vibronic coupling is neglected when the Born-Oppenheimer
approximation is used. A rigorous solution is to include the Franck-Condon (FC)
factor \(|\langle \nu' | \nu \rangle|^2\) to weight the transition between two electronic states. It is known
that in many cases, the vibration modes with frequency around 0.2-0.25 eV con-
tribute the most to the FC factor. However, for large system, one is not able to
calculate this at this stage. A broadening by distribution of each transition is
often used to represent many effects.

Usually, the Gaussian (\(\Delta^G\)), Lorentzian (\(\Delta^L\)), and Voigt (\(\Delta^V\)) distribution
functions are widely used in literatures:

\[
\Delta^G(\omega; \omega_{fi}, \Gamma) = \frac{1}{\sigma_{sd}\sqrt{2\pi}} \exp \left[-\frac{(\omega - \omega_{fi})^2}{2\sigma^2}\right], \quad \sigma_{sd} = \frac{\Gamma}{\sqrt{2\ln 2}}, \quad (2.31)
\]

\[
\Delta^L(\omega; \omega_{fi}, \Gamma) = \frac{1}{\pi} \frac{\Gamma}{(\omega - \omega_{fi})^2 + \Gamma^2}, \quad (2.32)
\]

\[
\Delta^V(\omega; \omega_{fi}, \Gamma) = \Delta^G(\omega; \omega_{fi}, \Gamma) \otimes \Delta^L(\omega; \omega_{fi}, \Gamma)
= \int_{-\infty}^{+\infty} d\tau \Delta^G(\tau; \omega_{fi}, \Gamma) \Delta^L(\omega - \tau; \omega_{fi}, \Gamma). \quad (2.33)
\]

In the above expressions, each bar spectrum is broadened centering at the transi-
tion energy \(\omega_{fi}\), and \(\Gamma\) denotes the half-width at half-maximum (HWHM). Each
bar intensity is multiplied by the distribution function, and the weighted sum-
mation at the same \(\omega\) adds to the final broadened spectral profile. The spectra
obtained from different distribution functions usually look similar with each other. But to the best of our knowledge, the Gaussian distribution is the most widely-used broadening function for the NEXAFS spectra in literatures.

For the continuum part of the spectrum (region higher than IP), a large HWHW value (fixed or gradually increasing) with conventional broadening functions mentioned above can be used. A better solution is to use the Stieltjes imaging procedure.\[58\]

\section{2.2 XPS}

The photoelectric effect—electrons are ejected from matter due to the electromagnetic radiation—was first discovered by Heinrich Hertz in the 19th century. Albert Einstein explained this effect by describing light as discrete photons, and he received the Nobel Prize in Physics in 1921.\[59\] A practical use of photoemission as an analytical tool was done by Kai Siegbahn and coworkers since 1950s, and he was awarded the Nobel Prize in 1981.\[59\]

X-ray photoelectron spectroscopy has been one of the most widely-used techniques for surface analysis. For an atom or molecule $A$, the process can be expressed by $h\omega + A \rightarrow A^+ + e^-$, where the electron $e^-$ is termed as a photoelectron. According to the rule of energy conservation, the binding energy (BE) of the emitted electron is given by

\[ E_b \equiv E_{A^+} - E_A = h\omega - E_{\text{kin}}, \]  

where only the kinetic energy $E_{\text{kin}}$ is considered for the emitted electron. While for solids, an electron transferring from bulk to surface also costs energy, and a work function term $\phi$ needs to be introduced to account for this effort, i.e.,

\[ E_b = h\omega - (E_{\text{kin}} + \phi). \]  

As experimentally the photon energy of the x-ray source is nearly a constant value (e.g. for Mg $K\alpha$ radiation, $h\omega = 1253.6$ eV), so the XPS profile is often given as the intensity with respect to either $E_b$ or $E_{\text{kin}}$.

Koopmans’ theorem provides a simple way to calculate XPS. After the Roothaan or KS equation is solved, one obtains a series of occupied and virtual orbitals with energies $\varepsilon_i$’s. Koopmans’ theorem employs a “frozen orbital” approximation, i.e., when electron $i$ is removed, the passive orbitals keep the same as the original
N-electron state. Within this approximation, the ionic potential (IP, or BE) is given by,

\[
\text{IP}_i \equiv N^{-1}E_i - NE = -\epsilon_i. \tag{2.36}
\]

More accurate IP can be calculated with a \(\Delta\text{SCF}\) philosophy which fully considers the relaxations:

\[
\text{IP}_i = N^{-1}E_{\text{FCH}}(n_i = 0) - N\text{E}_{\text{GS}}. \tag{2.37}
\]

Here FCH approximation is used to represent the core-ionized state, and \(N\text{E}_{\text{GS}}\) stands for the ground-state energy. The relativistic due to the removal of the core electron must be considered, and can be accounted for by a global shift as we do in the NEXAFS case. After the IP is obtained, the XPS spectra are obtained by broadening the IP bars, each weighted with the portion of equivalent centers among all centers of the studied element.

It is necessary to mention that in some cases, the photoemission can leads to the excitation of valence electron(s) to unoccupied level, and cause so-called shake-up peaks. These peaks are usually weak, and as more energies are needed, they appear several eV above the main peak.

### 2.3 XES

The non-resonant (or normal) XES is a second-order optical process which aims to map out the information of the valence levels. The fluorescence generated when an electron falls into a lower-energy state is used for experimental detections.

\[
E_f = E_i - \hbar\omega
\]

Figure 2.5 Schematic illustration of spontaneous emission of photon.

Theoretically, one can not calculate the probability of spontaneous emission using the TDPT with the electromagnetic field treated as a classical field. This can be done in a statistical way considering the detailed balance of spontaneous and induced emissions. And the Einstein coefficient for spontaneous emission from
\(|\Psi_i\rangle\) to \(\Psi_f\rangle\) (see Figure 2.5) is given by\(^{[60]}\)

\[
A_{if} = \frac{4e^2}{3\hbar c^3} \omega_{if}^3 \langle \Psi_i | \mathbf{r} | \Psi_f \rangle^2,
\]

(2.38)

where \(\omega_{if} = (E_i - E_f)/\hbar\). Note that there is no \(\epsilon\) in the above equation as different orientation has already been included. The transition probability can be given by,\(^{[61]}\)

\[
W_{if} = P_i \frac{A_{if}}{\sum_j A_{ij}},
\]

(2.39)

where \(P_i\) is the probability of the initial state and the summation is over all possible decay ways. By further assuming that the Auger emission dominates the decay process over the fluoresces channel (which is reasonable for light elements) and has a constant probability for all initial states, the transition rate of normal XES spectroscopy is given by\(^{[61,62]}\)

\[
W_{if}(\omega_{if}) \propto \omega_{if}^3 |\langle \Psi_i | \mathbf{r} | \Psi_f \rangle|^2,
\]

(2.40)

Similarly as previous section, the integral over wave functions can be simplified by single-electron MOs \(|\psi_{i,f}\rangle\) of the final state wavefunction \(|\Psi_f\rangle\). For practical calculation, the final state usually refers to the ground state (ignoring the effect of the valence hole). The emission cross section is finally expressed as

\[
\sigma_{\text{XES}}(\varepsilon_{if}) \propto W_{if}(\varepsilon_{if}) \propto \varepsilon_{if}^3 |\langle \psi_i | x | \psi_f \rangle|^2 + |\langle \psi_i | y | \psi_f \rangle|^2 + |\langle \psi_i | z | \psi_f \rangle|^2
\]

(2.41)

where \(\varepsilon_{if} = \varepsilon_i - \varepsilon_f\) is orbital energy difference.

2.4 RIXS

The resonant inelastic x-ray scattering is also a second-order optical process. As more orbitals (valence and virtual) are involved, it can reflect more information of the studied system. However, compared to other x-ray spectral techniques, the RIXS signal is much weaker. In the 1970s, the first resonant inelastic x-ray scattering experiment was carried out by Eisenberger et al.\(^{[63,64]}\) by using synchrotron radiation. Now this technique has been developed due to the use of high-brightness synchrotron radiation sources.
Theoretically, the RIXS spectrum can simply be obtained from normal XES with tuned incident photon energy. That is, the selectivity is added to the normal XES (which is non-selective to different transitions) to obtain the RIXS spectrum corresponding to specific excited atoms. For example, in order to study the nitrogen K-edge RIXS spectra for DNA (see Paper 1), we selected the XES transitions to the amine and imine 1s orbitals and obtained their respective RIXS spectra.

![Figure 2.6 Depiction of RIXS process by (a) electronic wavefunctions; and (b) molecular orbitals of the final state.](image)

While a rigorous calculation needs to treat the photon-in/photon-out processes together as one step. As depicted in Figure 2.4a, $|i\rangle$, $|n\rangle$, and $|f\rangle$ stand for the initial, intermediate, and final states, respectively; $\omega$ and $\omega'$ represent the frequencies of the absorbed and emitted photons. Let $\Omega$ be the solid angle, the double differential cross section of RIXS spectrum is given by

$$
\frac{d^2\sigma(\omega,\omega')}{d\Omega d\omega'} = \frac{\omega'}{\omega} \sum_{f_i} |F_{fi}(\omega,\omega')|^2 \Delta^L(\omega - \omega'; \omega_{fi}, \Gamma_f),
$$

(2.42)

where small positive number $\Gamma_n$ or $\Gamma_f$ stands for the lifetime broadening of corresponding state, and $F_{fi}(\omega,\omega')$ denotes the resonant inelastic part of the total (including elastic, resonant inelastic, and non-resonant inelastic parts) scattering amplitude. Each term in the r.h.s. of eq 2.43 is obtained from the Kramers-Heisenberg formula (also known as Kramers-Heisenberg-Dirac formula) within the dipole approximation. As the lifetime of the final state, $\frac{1}{\Gamma_f}$, is much longer than that of the intermediate state, $\frac{1}{\Gamma_n}$, (i.e. $\Gamma_f \ll \Gamma_n$), it is a good approximation to express the Lorentzian distribution function in eq 2.42 as Dirac-$\delta$ function $\delta(\omega - \omega' - \omega_{fi})$.

Similarly to previous sections, one can approximately use the a single-electron picture in spectral calculations when the electronic correlation is relatively weak.
The x-ray process has been shown in the left part of Figure 2.1 (and Figure 2.4 recaptures the final state). Let \( c, v, \) and \( u \) denote the core, valence, and unoccupied orbitals involved in the transition, we have the relations

\[
|n\rangle = a_u^\dagger a_c |i\rangle; \quad |f\rangle = a_v^\dagger a_v |n\rangle = a_v^\dagger a_v |i\rangle,
\]

(2.44)

where \( a^\dagger \) and \( a \) are the creation and annihilation operators, respectively. By using the Slater-Condon rule,\(^{[69]} \) the differential cross section can be rewritten as,

\[
\frac{d^2\sigma(\omega, \omega')}{d\Omega d\omega'} = \frac{\omega'}{\omega} \sum_{uv} |F_{uv}(\omega, \omega')|^2 \delta(\omega - \omega' - \omega_{uv}), \quad (2.45)
\]

\[
F_{uv}(\omega, \omega') = \alpha \sum_c \omega_{vc} \omega_{uc} \frac{\langle u|\mathbf{r}_1 \cdot \mathbf{e}|c\rangle \langle c|\mathbf{r}_1 \cdot \mathbf{e}'|v\rangle}{\omega - \omega_{uc} + i\Gamma_u}. \quad (2.46)
\]

To consider practical experimental instrumentation situation, the incident photon energy has a distribution which can be specified by a Gaussian distribution \( \Delta G(\omega; \omega_0, \gamma).^{[70]} \) Thus, integration operator \( \int d\omega \Delta G(\omega; \omega_0, \gamma) \) executed on both sides of eq 2.45 leads to

\[
\int d\omega \Delta G(\omega; \omega_0, \gamma) \frac{d^2\sigma(\omega, \omega')}{d\Omega d\omega'} = \sum_{uv} \Delta G(\omega' + \omega_{uv}; \omega_0, \gamma) \frac{\omega'}{\omega' + \omega_{uv}} |F_{uv}(\omega' + \omega_{uv}, \omega')|^2. \quad (2.47)
\]

Further, to account for the orientational average of molecules, one need to integrate with \( \int d\Omega \) to the above equation, and an efficient solution is provided by Luo et al.\(^{[71,72]} \) More details on RIXS can be found in reviews by Gel’Mukhanov and Ågren\(^{[65]} \), Kotani et al.\(^{[73]} \), and book by Sakurai\(^{[74]} \).
Chapter 3

Ground-state energies and structures for large systems

The application of quantum mechanical methods for large systems is hindered by the high scaling of the conventional computational approaches with respect to the system size (characterized by the number of atoms, $N$). The widely-used Hartree-Fock (HF) and density functional theory (DFT) methods have quartic and cubic scaling, respectively [often denoted as $O(N^4)$ or $O(N^3)$]. Various post-HF methods have even higher computational scaling, for example, many-body perturbation theory truncated at the second-order (MP2) scales as $O(N^5)$, and coupled-cluster singlets and doublets (CCSD) scales as $O(N^6)$. Therefore, it is demanding to have more efficient algorithms to solve the corresponding equations within these methods. Generally, the solutions to this problem falls into two categories, the hybrid approaches and linear scaling approaches. They will be presented in Section 3.1 and Section 3.2, respectively, and the discussions in this chapter are limited only in the ground states.

3.1 The hybrid approaches

Hybrid way means mixing different theoretical levels in the study of a target system. That is, the important, usually also small, region of the system (model system) is treated with accurate method (high level), while an economic method (low level) is employed for the rest part, as schematically shown in Figure 3.1a. For instance, for an enzyme-catalyzed reaction problem, one can use a DFT method for the reaction center, and a force-field method for the rest part. Such a combination of quantum mechanical (QM) and molecular mechanics (MM) methods is termed
as a QM/MM approach. In many places, one can also see a mixed usage of two QM methods with different accuracy, usually a DFT together with a semi-empirical methods, which is denoted as a QM/QM’ approach.

The total Hamiltonian within a QM/MM method is given by,

\[ \hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}}, \]  

(3.1)

where the three terms denote the QM, MM, and the interaction parts. The performance of a QM/MM approach depends largely on a good description of the interaction term. Usually, the mechanical embedding and the electronic embedding are used to describe the van der Waals (vdW) and electrostatic interactions between the two parts, respectively. For practical usage, the boundary between the QM and MM regions may be specially considered if, for example, a covalent bond is broken when defining the two regions. A so-called “link atom”, usually a hydrogen, is added for valence saturation. Various QM/MM implementations have been suggested\[^{75-78}\], which differ in the ways to handle the link atoms, to define the theoretical levels, and to treat the solvent.\[^{79}\]

One widely-used hybrid method is the ONIOM (our own n-layer integrated molecular orbital molecular mechanics) method, which is developed by Morokuma et al.\[^{80-82}\] and has been implemented in the GAUSSIAN package\[^{83}\]. This approach defines a different theoretical level as a “layer”, and calculate the target system at the high level as a hierarchy of underlying calculations. A two-layer (2L) model is the most often used, and the ONIOM total energy of the target full
Ground-state energies and structures for large systems

The model system is part of the full system with particular interest. For more complicated systems, a three-layer (3L) can be defined.各项工作

Various hybrid methods have been widely employed in theoretical investigation of reactions, catalysis, conformational transitions, and so on. Acceptable accuracy is guaranteed if the studied problem is localized in a small part; while high accuracy (∼1 kcal/mol) usually is not easy to be obtained. For instance, the predicted relative energies by choosing different theoretical levels can differ by about 10 kcal/mol, as revealed in a recent case study by Morokuma and coworkers. The hybrid method is appropriate to cases which involve some local regions during the chemical transformation. However, it can not be employed in studying chemical problems where the geometrical changes are global, for example, the folding of proteins. Limitations exist for the hybrid methods and actually the QM or MM region is not easy to be defined. For such problems, the hybrid methods are not appropriate, while the following linear-scaling methods are more preferred.

3.2 Linear-scaling methods

The linear-scaling [O(N)] methods are usually based on the “near-sightedness” of the density matrix and/or the locality of electron correlation. Figure 3.1b provides an illustration for the DFT methods, where the cubic scaling of conventional DFT method can be lowered to linear when O(N) algorithm is employed. In the region beyond the crossover point, usually around several hundred atoms, the computational time is saved.

3.2.1 Standard linear-scaling methods

Standard linear-scaling algorithms have been developed at various levels. For HF/DFT methods, the strategy is to lower the computation scaling of the key steps in the calculations, that is, the Fock matrix formulation and the diagonalization of the transformed Fock matrix. The linear scaling of the former step can be achieved by using the Order-N exchange (ONX) or fast multipole methods, while the latter one can be avoided by applying the density matrix search methods. The recombination of these methods by using localized

\[ E^\text{ONIOM(high:low)}_{\text{full}} = E^\text{high}_{\text{model}} - E^\text{low}_{\text{model}} + E^\text{low}_{\text{full}}, \]  

where the model system is part of the full system with particular interest. For more complicated systems, a three-layer (3L) can be defined.各项工作
molecular or atomic orbitals can also guarantee linear scaling for post-HF methods. Besides, linear-scaling are also available for property calculations.\textsuperscript{[91,92,102–107]} However, these linear scaling algorithms have not been established as practical tools for routine uses as the crossover between conventional algorithms and them occurs usually at several hundred atoms.

### 3.2.2 Fragment-based methods

Alternatively to the conventional linear-scaling methods, fragment-based methods which are based on chemical intuitions, have been proposed.\textsuperscript{[108–132]} In these methods, a target molecule is cut into small pieces (named “fragments”), and the energy or properties of the target molecule are obtained by assembling corresponding values from a series of “subsystems” (each formed from a few neighboring fragments). Various fragment-based approaches can roughly be classified as density matrix (DM)-based, energy-based, or combination of both. In a DM-based approach, the total DM of the target system is first constructed by assembling DMs of a series of subsystems, and then the ground-state energy and properties can be calculated from it. While in an energy-based approach, the total energy or properties of a target system are directly assembled from corresponding quantity of each subsystem, which makes it the simplest and the most convenient solution. Mixed-DM/energy method has also been introduced, for example, the fragment molecular orbital (FMO) method\textsuperscript{[121,122,133–135]} by Kitaura, Fedorov and coworkers.

**Energy-based fragmentation method**

The energy-based fragmentation (EBF) method for the ground-state total energy and energy derivatives is first proposed by Li et al.\textsuperscript{[109]} and Collins et al.\textsuperscript{[110]} independently in 2005. The basic underlying principle is the fast decay and good transferability (in structurally similar molecules) of the $n$-fragment interaction energies.\textsuperscript{[109]} In the EBF approach, the total energy is expressed as,

$$E_{\text{tot}}^{\text{EBF}} = \sum_{m} C_{m} E_{m},$$

where $E_{m}$ denotes the total energy of the $m$th subsystem, $C_{m}$ is the coefficient, and the summation is over all constructed subsystems. The set of $\{C_{m}\}$ is generated subject to that all one- and two-fragment terms, and all considered multi-fragment terms appear only once. In practice, the flowchart of the EBF approach includes
(1) dividing a target molecule into fragments with similar sizes; (2) constructing subsystems from neighboring fragments subjecting to some rules; (3) performing conventional calculation for each subsystem; (4) calculating the total energy using eq 3.3. The EBF method are applicable to closed-shell systems without localized charges or highly-polarized groups, and well reproduces total energy, dipole moment, forces, Hessians, and various properties of a large molecule. However, it does not have a good performance for systems with many polar groups or charged groups.

Generalized energy-based fragmentation method

The extension of EBF method to more general closed-shell molecules, especially for those containing many polar groups or charged groups, is proposed in 2007 under the name of the generalized energy-based fragmentation (GEBF) approach.\textsuperscript{108} The procedure of GEBF approach is similar to the EBF one except that each subsystem is embedded in the background charges at the positions of the distant atoms, which approximately includes the electrostatic influence from them. As the charges are obtained iteratively from atomic population in a self-consistent way,\textsuperscript{108} the polarization effect is also approximately considered. The underlying idea is that the interaction energy between two fragments $I$ and $J$ can be approximately obtained from

$$E_{I,J} \approx E_{I',J} + E_{I,J'} - E_{I',J'},$$  \hspace{1cm} (3.4)

where $I'$ ($J'$) represents that each QM atom in fragment $I$ ($J$) is replaced by a point charge at the same position. The GEBF total energy is given by

$$E_{\text{GEBF}}^{\text{tot}} = \sum_{\text{subsys}} C_m \tilde{E}_m - \left( \sum_{\text{subsys}} C_m - 1 \right) \sum_{A<B} \frac{Q_A Q_B}{R_{AB}},$$  \hspace{1cm} (3.5)

where $\tilde{E}_m$ denotes the total energy of the embedded subsystem $m$ (with the inclusion of the electrostatic energy of all background point charges), and $Q_A$ represents the point charge on atom $A$.

Various properties are (or as functions of) energies derivatives with respect to nuclear coordinates, electric field, magnetic field, and so on. The advantage of the GEBF method is that the energy derivatives can also be obtained in a similar
way as the total energy. Thus, we have the force and Hessian given by:\cite{136,137}

\[
\frac{\partial F_{\text{tot}}^{\text{GEBF}}}{\partial q_{Aa}} \approx \sum_n C_n \frac{\partial \tilde{E}_n}{\partial q_{Aa}}, \tag{3.6}
\]

\[
\frac{\partial^2 F_{\text{tot}}^{\text{GEBF}}}{\partial q_{Aa} \partial q_{Bb}} \approx \sum_n C_n \frac{\partial^2 \tilde{E}_n}{\partial q_{Aa} \partial q_{Bb}}. \tag{3.7}
\]

In the above expressions, \( a, b = \{x, y, z\} \), \( q_{Aa} \) or \( q_{Bb} \) denotes the Cartesian displacement, and the summation over \( n \) is limited to subsystems containing atom \( A \) (atoms \( A \) and \( B \)) as a real atom (real atoms). Thus this method can be employed for geometrical optimization and frequency calculations. Similarly, the derivatives over electric field or combined derivatives are given by:\cite{108,136}

\[
\Omega \approx \sum_{\text{subsys}} C_m \Omega_m (\Omega = \mu_a, \alpha_{ab}), \tag{3.8}
\]

\[
\frac{\partial \Omega}{\partial q_{Aa}} \approx \sum_n C_n \frac{\partial \Omega_n}{\partial q_{Aa}}. \tag{3.9}
\]

Thus, this method can be used for dipole moment (\( \mu_a \)), polarizability (\( \alpha_{ab} \)), as well as infrared and Raman spectra\cite{138} calculations.
Chapter 4

Complex systems: challenges, efforts, and achievements

The application of quantum mechanical theory to large and complex systems is challenging. For isolated large molecules, the accuracy depends on a good description of the intramolecular interactions. While for molecules in ambient conditions or bulk materials, it also depends on the reproduction of important intermolecular interactions and considerations of statistical effects and environmental influences. In other words, theory needs to treat a more “realistic” system instead of just calculation on a “model”. However, the challenge lies in the competition between an accurate theoretical description of the system and the increasing computational cost. One possible strategy is to construct a series of models with increasing sizes, and thus the real system can be achieved by asymptotic or convergent measurements. Another way is to make use of the localized properties of a system, and obtain energy and properties from a series of subsystems, each of which is part of the target system.

In this chapter, we show our efforts in studies of spectroscopy, energy, and structure of large systems. The case studies include both isolated large molecules and small molecules in ambient conditions, and both strategies above have been used. Special focus is paid on the influences of weak interaction (hydrogen bonding, stacking effect) and covalent bonding to the molecular or electronic structures. The story behind each project, and planned future outlook are also mentioned in the text. In the following, the DNA duplexes, nucleobases, and related molecules are first discussed in Section 4.1. Then in Section 4.2, mono- and few- layered graphenes are studied, and the influences of edge and defects are considered. Next in Section 4.3, we present the studies of short peptides in water solution and in a
self-assembled cage. Finally in Section 4.4, we provide the progress of multipole expansion for approximate Hartree-Fock calculations.

4.1 DNA

Why DNA?

Our interest first falls in DNA (deoxyribose nucleic acid) which is known as the molecule of heredity. It has an amazing duplex structure, which was first suggested by Watson and Crick\textsuperscript{139} in 1953. Various helical forms exist, say, A-DNA, B-DNA, and Z-DNA, which differ in pitch, twist angle, and other structural parameters. An example of the most common double helical structure, B-DNA duplex (pitch distance 3.375 Å, twist angle 36°), is illustrated in Figure 4.1a-b. Structurally, DNA is a polymer made of deoxyribonucleotide units, each of which contains a base, pentose, and a phosphate group, as shown in Figure 4.1c. There are altogether four types of bases in DNA, viz., guanine (G), cytosine (C), adenine (A), and thymine (T); and they normally have definite pairing patterns of G·C and A·T (Figure 4.1d-e). Thus, the hydrogen bonding within a base pair (BP) and the \(\pi-\pi\) stacking between different base pairs are the major weak intramolecular interactions in a DNA.

How the base pair components and the \(\pi-\pi\) stacking between base pairs influence the electronic structure, and other properties, of DNA has been a fundamental problem in biochemistry. This can however not be directly observed, but can only be reflected by measurements in an indirect way. Various soft x-ray spectra which measure the electronic transitions between different states, are efficient tools to examine their electronic structures. So it would be appealing to study the soft x-ray spectra of DNA. It is noticed that early soft x-ray studies on DNA are all by experiments.\textsuperscript{1,140–143} Theoretical works are only done for the small building-block molecules\textsuperscript{142,144–147} (nucleobases, nucleotides, etc.). This is based on the “building block principle”\textsuperscript{17,148–151} which states that the spectrum of a complex molecule can be approximately obtained by linear combination of the composed small units. These calculations provide a useful starting point for spectral interpretation of DNA, however, obtaining the spectra of a real DNA would be important and necessary. So it comes to the work of Paper 1.
Figure 4.1 Structural illustration of DNA: (a) Side and (b) front views of a B-DNA duplex fragment; (c) deoxyribonucleotide unit; (d-e) two base-paring patterns with imine (=N-) and amine (-NH-) nitrogens labeled.

Figure 4.2 Comparison of N 1s NEXAFS spectra (contribution from one nitrogen atom) in (gc)₂ calculated from two models: (a) The spectra; models (b) with and (c) without the phosphate groups. Hydrogens without forming HBs are hid for clarity, and arrows denote the excited atoms.
Phosphate groups and model simplification

It is natural to study the spectra of nitrogen, because it is the element that appears only in the nucleobases and thus gives the clearest fingerprint of DNA. In order to study the component-dependent features, we define two pure G·C or A·T stacked B-DNA duplexes, \((dGdC)_n·(dCdG)_n\) and \((dAdT)_n·(dTdA)_n\) [denoted as \((gc)_n\) and \((at)_n\) for simplicity]. The number \(n\) is increased \((n = 1, 2, 4, 6, \ldots)\) in order to get a series approaching a real DNA with a large number of base pairs. The NEXAFS, XES, and RIXS spectra are calculated for the \((gc)_n\) and \((at)_n\) sequences, the comparison of spectra at different \(n\) reflects the influence of stacking effect. We employ the mostly-used DFT method with the B3LYP functional\(^{[152]}\) to do the electronic structure calculations. Because these are large molecules containing some hundreds of atoms, we use the ECH approximation to calculate the NEXAFS spectra.

Our first plan is ambitious to include every atom in our models. It works fine for small models up till \(n=2\), but for larger molecules, we meet the SCF convergence problem. We think the problem may come from the phosphate groups. Since they locate relatively far from the nitrogens, they are then proved to have no influence to the spectra and thus can be neglected. By choosing a medium-sized molecule \((gc)_2\), test calculations are performed for models with (Figure 4.2b) and without (Figure 4.2c) the phosphate groups. As shown in Figure 4.2a, the atom-specific spectra generated from two models look nearly the same. Tests on a few excited atoms give the same conclusion. So we are confident to neglect them and the calculations run smoothly.

Insensitive stacking

Figure 4.3a-c display the NEXAFS, XES, and RIXS spectra calculated from different models, compared also with corresponding experimental spectra of poly(dG)·poly(dC) film, respectively. To our surprise, the calculated spectra converge so quickly. One can see that \((gc)_n\) with \(n = 1, 2, 4, 6\) give nearly the same spectra, which are even quite similar to the profile of the GC base pair. The convergence of spectra convincingly shows that our calculations can represent a real DNA with many base pairs. The stacking effect is found to be insensitive to the N1s NEXAFS spectra of DNA. The agreement with the experimental spectra is also good (see Figure 4.3), with all qualitative features accurately reproduced. The same conclusion applies to the \((at)_n\) sequence. One could deduce that for a general DNA
Figure 4.3  N1s x-ray spectra of poly(dG)-poly(dC) DNA duplexes: (a) NEXAFS (b) normal XES, and (c) RIXS spectra calculated from different sizes of \((gc)_n\) models compared with experiments; (d) experimental XES spectra of GC-DNA film. Calculated NEXAFS spectra are shifted by +0.2 eV for better comparison with experiment\(^{[1]}\); calculated XES spectra are shifted by +10 eV to match with experiment (incident photon energy \(h\nu_i=410\) eV); calculated RIXS spectra are shifted by +10 eV to match with experiment. Selected from Paper 1, reprinted with permission from American Chemical Society.
with mixed base pairs, its spectra can be well reproduced by linear combinations of those of GC and AT, weighted by their ratio.

**Discrepancy in the nucleobase XAS**

Table 4.1  Peak positions (eV) of the main $\pi^*$ features in the N1s NEXAFS spectra of nucleobases, nucleosides and nucleotides from literatures. $\Delta$ denotes the difference of the $\pi_1^*$ and $\pi_2^*$ bands. $^a$ Selected from Paper 2, reprinted with permission from American Chemical Society.

<table>
<thead>
<tr>
<th>mol.</th>
<th>$\pi_1^*$</th>
<th>$\pi_2^*$</th>
<th>$\Delta$</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>401.2</td>
<td>402.3</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400.6$^b$</td>
<td>401.7$^b$</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>399.9</td>
<td>400.7</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>399.6</td>
<td>400.8</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>399.4$^b$</td>
<td>401.9$^b$</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>399.6$^b$</td>
<td>402.7$^b$</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400.8$^b$</td>
<td>402.7$^b$</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>G-dS</td>
<td>399.9</td>
<td>401.6</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>G-dS-P</td>
<td>401.3</td>
<td>403.4$^b$</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>399.8</td>
<td>401.0</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>400.7</td>
<td>402.0</td>
<td>1.3</td>
<td>exp.$^{[140]}$</td>
</tr>
<tr>
<td></td>
<td>399.9$^b$</td>
<td>401.1$^b$</td>
<td>1.2</td>
<td>exp.$^{[144]}$</td>
</tr>
<tr>
<td></td>
<td>399.0</td>
<td>400.2</td>
<td>1.2</td>
<td>exp.$^{[4]}$</td>
</tr>
<tr>
<td></td>
<td>399.1</td>
<td>400.4</td>
<td>1.3</td>
<td>exp.$^{[3]}$</td>
</tr>
<tr>
<td></td>
<td>400.1$^b$</td>
<td>402.8$^b$</td>
<td>2.7</td>
<td>calc./StoBe$^{[144]}$</td>
</tr>
<tr>
<td></td>
<td>399.9$^b$</td>
<td>402.9$^b$</td>
<td>3.0</td>
<td>calc./GSCF3$^{[144]}$</td>
</tr>
<tr>
<td></td>
<td>400.7$^b$</td>
<td>402.9$^b$</td>
<td>2.2</td>
<td>calc./ECA$^{[147]}$</td>
</tr>
</tbody>
</table>

$^a$ G=guanine, C=cytosine, d=deoxyribose, S=sugar, P=phosphate group.

$^b$ Values recaptured from the figures. (Others are explicitly given in text.)

Above we have shown the x-ray spectra of DNA that demonstrate a good agreement between the experiment and the theory. However, for the absorption spectrum of the smallest building block, a nucleobase, discrepancy exists not only between the theory and the experiment, but also among different experimental results. This is the most evident for guanine, and next for cytosine. Table 4.1 summarizes earlier experimental and theoretical results. $\pi_1^*$ and $\pi_2^*$ represent the $\pi^*$ band energies of imine and amine nitrogens, respectively, and $\Delta$ denotes their difference. For example, for G, experiments predict a $\Delta$ of about 0.8-1.2 eV, while theoretical works give a much larger $\Delta$ of 1.9-3.1 eV. We think this is because that all calculations were done in the gas phase, while the experiments were carried out in the condensed phase, usually as powders. The same proposal was also mentioned by Macnaughton et al.$^{[144]}$, but nobody had ever convincingly shown that. Besides, there is discrepancy among the spectra of a base and its nucleoside or nucleotide. For example, guanine nucleoside has a much larger $\Delta$ (1.7 eV) than the corresponding nucleotide (0.8 eV) and the nucleobase (1.2 eV).$^{[4]}$ Since our previous study has already shown that phosphate has nearly no effect (and
sugar is later shown to have very little effect), we believe that this is also due to
different condensed phases of the bases. **Paper 2** mainly deals with this problem,
and provides results for the x-ray absorption spectra of solid samples.

**Initial attempts for oligomers**

Generally, the difference between a molecule in condensed and isolated phases
is mainly the intramolecular interactions. One simple way to study the solid-
state spectra is to enumerate possible low-lying oligomer structures to consider
the local interactions. This is especially useful when the crystal structure does
not exist (e.g. a noncrystalline sample) or not available. As an early attempt,
we have ever tried this approach for guanine, and several oligomer structures are
given in Figure 4.4. The resulted spectra (not shown) have evident difference,
which all exhibit significant difference to the gas-phase spectra. Although there
is still discrepancy to the experimental spectra, these initial attempts provide an
important starting point for the solid sample studies. Note that as there are
many candidate positions to form hydrogen bonds in guanine, the enumerations
above are incomplete. A better solution for amorphous guanine powder would
be to include more configurations by using Monte Carlo or molecular dynamics
techniques.

![Figure 4.4 Selected oligomer configurations of guanines: (a-c) dimers; (d) tetramer.](image)

**Cluster models for infinite crystal structure**

Why not try crystal structures if they exist? The early attempts have shown that
neighboring molecules have great influence to the spectra. This motivates us to
look for possible crystal structures. Then, we find five crystal structures, say, anhydrous guanine\cite{2}, guanine monohydrate\cite{153}, anhydrous cytosine\cite{154}, cytosine monohydrate\cite{5}, guanosine dihydrate\cite{155}, which are used in our calculations.

There are generally two ways to study the x-ray spectra for a molecular crystal. One way is to use the periodic boundary condition (PBC), which is often associated also with planewave basis sets and pseudopotentials.\cite{38} A modified pseudopotential is specially used for the core-excited atom, and a large supercell is necessary to consider more interactions in the real space, and to get convergent spectra. Another simpler way is to construct cluster model which represents the infinite crystal at considerable large size. In this case, the near-range intermolecular interactions are all considered in the real space.

In this thesis, we only use the latter way which is combined with the ECH method. In practice, we choose an ordinary residue (labeled as central residue), and construct cluster models by capping it with neighboring residues. A distance threshold $\xi$ is used to define the cutoff. By adjusting $\xi$, spectra calculated from models with different sizes can approach convergence, which is considered as the spectrum of the infinite crystal. In practice, either the center-of-mass distance or minimum atom-atom distance can be used, without causing much difference.

Figure 4.5a schematically shows the construction of models for the guanine G7H crystal structure. In panel b, the calculated spectrum for crystal structure of anhydrous guanine (curve labeled by 0%) gets very close to the experimental spectrum. It also shows an obvious difference to that of an isolated molecule (panel c). The good agreement with the experiment show the reliability of our method, and major intermolecular interactions have been reproduced.

**Consideration of external perturbations**

A experimental spectral measurement is carried out in a specific laboratorial condition (temperature, pressure, etc.), which could affect the measured results. It would be interesting to see the response of spectra under the external perturbations. This also helps to see the range of spectra obtained by different experimental groups. To generate such external perturbations in theoretical calculations, we assume the influence does not change the internal geometry of each single residue, but only the residue-residue distances. In other words, we introduce “expansion” and “compression” to the system.

The expansion or compression can change the intermolecular interactions, mainly the hydrogen bonding. In the cases of cytosine monohydrate and the an-
Figure 4.5  N1s NEXAFS spectra and structures of guanine G7H isomer. (a) Cluster model used for calculation as generated from the crystal structure.[2] (b) Spectra of anhydrous guanine crystal calculated by ECH method at the crystal structure[2] (0%) and at geometries with the distances between the central and all neighboring molecules scaled (±1%, ±3%), compared with experiment.[3] (c) Gas phase spectrum calculated by FCH method (shifted by +0.57 eV for comparison with experiment[3]). Selected from Paper 2, reprinted with permission from American Chemical Society.

Figure 4.6  Calculated Δ values (Δ denotes the energy difference of the amine and imine 1s → π* peaks) of crystals of (a) cytosine, (b) guanine, and (c) guanosine with different models vs different scaled geometries. Experimental values[3,4] are also shown as horizontal dashed lines. Selected from Paper 2, reprinted with permission from American Chemical Society.
hydrous guanine, some residues appear to be in parallel, the stacking effect is also especially considered in two different ways. In the first way, the spectra predicted by two models with and without the stacking bases are compared. Another way is more explicit. We selectively scale only the distances of stacking molecules, with those of the other molecules unchanged.

The performance of such a strategy is viewed in Figure 4.5b. The compression or expansion by up to 3% (with respect to the original distances in each crystal structure) causes a decrease or increase of $\Delta$ by around 0.3 eV (compared to that predicted from the crystal structure). This is because the compression make the hydrogen bonds much stronger. As a consequence, the electrons become more delocalized, and the amine and imine nitrogens become less distinguishable.

Furthermore, Figure 4.1 summarizes the change of $\Delta$ with respect to the change of hydrogen bond of all crystal structures and models. In a word, one can find that the change of hydrogen bond of around 0.2-0.3 Å (denoted by ± 3%) can cause a change of the $\Delta$ around 0.2-0.8 eV. Our calculations finally confirm that the stacking effect nearly makes no influence on the spectra, which is consistent with the conclusion obtained for DNA duplexes.

To summarize, we have solved the problem of the guanine and cytosine absorption spectra, successfully illustrated the discrepancies at different experimental conditions, and clarified the spectral difference of a base, nucleoside and nucleotide. From the above discussions, one can also find that the influence of hydrogen bonds in the condensed-phase is obvious. The cluster model plus ECH approach is also proved to be applicable for the absorption spectra calculation of molecular crystals. We then extend this approach to more solid systems, for example, crystalline ice.

**Functionals**

One might comment on the B3LYP functional used in the study of stacking effect, as it is generally accepted that it lacks long-range dispersion. However, one should keep in mind that what we are dealing is excited state with a core-hole. The MOs become quite extended, and thus one can get a good description even without the long-range correction.

To verify this point, we compare the B3LYP-predicted spectra with those obtained by using the newly-developed M062X functional, which is validated for long-range dispersions. Figure 4.1 illustrates the N1s NEXAFS spectra of cytosine monohydrate calculated from the two functionals. One can find that both
functionals can produce nearly the same $\pi_1^*$ and $\pi_2^*$ peaks at each scaled geometry. Small difference of course exists in the $\sigma^*$ regions; the pre-shoulder structure predicted by M062X is simply because the contribution from non-equivalent N atoms have slightly different transition energies in this region. The agreement verifies that our B3LYP results are reliable.

![Figure 4.7](image)

**Application of the same strategy to crystalline ice**

Ice exists in many different forms under different temperature and pressure conditions. Until now, 16 crystalline and 3 non-crystalline (amorphous) phases have been found. For the crystalline forms, the oxygens are in fixed positions while the hydrogens can be disordered or ordered. Ordinary ice or natural snow on Earth has the hexagonal crystal structure (space group 194), which is denoted as ice-Ih using the Bridgman nomenclature\(^{157}\), as depicted in Figure 4.8a. In the region around 130-220 K of the phase diagram, ice can also take a face-centered-cubic crystal structure with half of the tetrahedral holes filled (space group 227), which is denoted as ice-Ic. As the other forms of ice exist in conditions far from room temperature and atmosphere pressure, ice-Ih and ice-Ic are the two most discussed forms in literatures.
There are altogether eight waters in one unit cell. Based on the coordinates of unit cell\textsuperscript{[10]}, cluster models are constructed by including waters surrounding a chosen central water molecule subject to increasing distance threshold $\xi$. Due to the high symmetry of the ice I$\mathrm{h}$ structure, currently only one oxygen is chosen as the excited atom.

We first study the size-dependence of spectra for bulk ice I$\mathrm{h}$, and one specific hydrogen order is used. The spectra are illustrated in Figure 4.8b, where the convergence is achieved since $\xi = 7.0$ Å. The agreement between the computed and experimental spectrum is fairly good (Figure 4.8c). Compared with the spectra calculated by other methods, our spectrum well reproduces the energy difference 4 eV between the main-edge (537 eV) and the post-edge (541 eV). This shows that our approach can give reasonable results for crystal structures. None of the theoretical spectra has predicted the pre-edge structure, which might comes from the transitions from surface ice. The computed spectrum at $\xi = 10.0$ Å has been aligned with the post-edge peak of the experimental spectrum\textsuperscript{[6]}, and the shift value is also used for models with other sizes. Each spectrum is broadened using Lorentzian line shape with FWHM of 1.5 eV and 2.0 eV in regions lower than
537 eV and higher than 541 eV, respectively. In the in-between area, a linear-increasing FWHM is adopted.

Figure 4.9 shows the NEXAFS spectra of 16 hydrogen-order structures of Ice Ih. Calculations are performed using cluster models with $\xi = 7.0$ Å, constructed from the theory-predicted crystal coordinates\textsuperscript{[10]}. One can find that each spectrum exhibits similar main-edge structure (differ by around 0.3 eV in energy position), while large difference is found in post-edge structure. Thus we can see that different hydrogen order can influence the NEXAFS spectra. The average spectrum also matches well with the experimental one.

Figure 4.9  O K-edge NEXAFS spectra of Ice-Ih with 16 kinds of hydrogen order\textsuperscript{[10]} compared with experimental result by Wernet et al.\textsuperscript{[6]}. Calculations are based on cluster models constructed with $\xi = 7.0$ Å.
4.2 Graphene

Why graphene?

Graphene is a two-dimensional (2D) infinite hexagonal sheet of carbon with only one-atom thick, as shown in Figure 4.10. It was first experimentally obtained in 2004 by Konstantin Novoselov and Andre Geim et al.\textsuperscript{[158]}, and they were awarded the Nobel Prize in Physics in 2010.\textsuperscript{[59]} It is however many years before the discovery that theoretical studies have been performed for this amazing, “assumed” structure. Graphene has peculiar electronic structure (e.g. zero band-gap) and properties (e.g. strong mechanical strength\textsuperscript{[159]}, quantum spin Hall effect\textsuperscript{[160]} and half-quantized quantum Hall effect\textsuperscript{[161]}, high optical transparency\textsuperscript{[162]} and electrical conductivity\textsuperscript{[158]}), which establishes importance in potential industrial applications. As it is usually considered as the bridge between molecules and materials, the graphene studies are also helpful for conceptual understanding. By rolling or stacking of graphenes, other systems with various dimensions can be obtained: fullerenes (0D), carbon nanotubes (1D) and graphites (3D). Therefore, one usually views graphene as the origin of the other carbon-based systems. Due to the importance of this material, many works have been done to obtain its electronic structure, including the soft-x-ray spectra techniques.

Two characteristic features and debate on assignments

The first NEXAFS experiment on graphene was performed by Pacilé et al. who found two distinct features different from bulk graphite: one in 283.7 eV which is termed as the sign of the splitting of the \(\pi^*\) peak, and the other in 288 eV which is assigned as an interlayer state.\textsuperscript{[11]} Jeong et al. however disagree with the assignment at 288 eV. They think that it might comes from COOH group and/or C-H contamination because graphene is very easy to be oxidated.\textsuperscript{[12]} While Pacilé et al. claimed the purity of their samples.\textsuperscript{[163]} The main features of the experimental spectra is recaptured in Figure 4.11. It would be important for theoretical studies to give an explanation, and our Paper 3 provides such a solution. Furthermore, a systematic analysis of the effects of size, stacking, edge, and defects for the graphene XAS is conducted.
Figure 4.10  Depiction of ideal infinite graphene sheet with supercell labeled.

Figure 4.11  [Reprinted with permission from American Physical Society] (a) Experimental C K-edge NEXAFS spectra of (a) graphene\cite{11} and pristine graphite\cite{12}; (b) FLG around 283.7 eV\cite{11}; and (c) FLG around 288 eV\cite{11}. Characteristic peaks are labeled by arrows and numbers in eV. Selected from Paper 3, reprinted with permission from American Physical Society.
Figure 4.12  C K-edge NEXAFS spectra of graphene: (a) three graphene nanoribbons (GNRs) of different size: Size A, Size B and Size C; (b) 1-, 2-, and 3-layers (1L, 2L, 3L) GNRs based on Size A cluster; (c) atom-specific spectra of 1L GNR with Size B as excited from central, outer, and edge regions; (d) Stone-Wales defected GNR with Size A. Left and middle panels illustrate 1s→π* and →σ* transitions, respectively. Panels a3 [with FWHM=0.2 eV (lower) and FWHM=0.8 eV (higher)] and b3 (with FWHM=0.8 eV only) provide fine structures in the first π* peak. Panels c3 and d3 denote the structure of GNR used for calculations with the excited centers labeled. Selected from Paper 3, reprinted with permission from American Physical Society.
Convergence verified by different sizes

We use a graphene nanoribbon (GNR) saturated with hydrogens in the edges as our models. In the semiconductor field, GNR is designed by cutting graphene in order to open the zero-band-gap of graphene. The GNRs with increasing sizes are used to modulate the features of the infinite graphene. The unrestricted B3LYP method is used for electronic structure calculations, and the ECH method is employed due to the large size of this system.

The convergence is however much slower than the DNA case because the electrons are much more correlated due to the delocalized $\pi$-bond. Figure 4.12a illustrates the size dependence and convergence of NEXAFS spectra of GNRs with three sizes: 2.9 nm $\times$ 2.7 nm (size A), 5.0 nm $\times$ 4.0 nm (size B), and 5.0 nm $\times$ 4.4 nm (size C). In the $\pi$ spectra (panel a1), a pre-edge peak 0 locating at 285 eV appears, which is around 0.5 eV lower than the main peak. In 288 eV, a stronger peak appears in a small-sized than a large-sized GNR, as clearly viewed in panel a3. Similarly, in the $\sigma$ spectra (panel a2), the features appearing at 295.5 and 298 eV have all smeared out as a result of the increase in the DOS. Our calculations thus show that the experimental feature 288 eV can only appear in a small-sized graphene.

Convergence as verified by ground-state DOS

The density of states (DOS) in ground state and the NEXAFS spectrum in core-excited state are related. After the NEXAFS calculations, we check it with the ground state DOS of graphene. Actually, there are discrepancies in literatures even for this ground-state problem. Figure 4.13a depicts the DOS results calculated from local density approximation (LDA) and gradient corrected functional approximation (GGA). Especially, the calculation of Trickey et al.\cite{13} (LDA-1) shows a new peak, which was used as a verification for the experimental feature 283.7 eV in the XAS experiment.\cite{11} Our previous study\cite{164} has revealed that nonlocal exchange interactions may be important in describing the graphene electronic structure. Since nobody has ever calculated the DOS with a hybrid functional, we decide to perform a DOS calculation using the B3LYP functional by ourself.

Our calculated DOS is consistent with LDA-2\cite{14}, GGA-1\cite{15}, and GGA-2\cite{16} results, without predicting a new featured peak in LDA-1.\cite{13} The ground state calculations verify our conclusions in the NEXAFS spectra. As other calculations are based on periodic boundary conditions, the similarity of DOS also verifies that our models are large enough to represent an infinite graphene sheet.
4.2 Graphene

Figure 4.13  Ground-state DOS of ideal one-layer graphene from different density functionals. (a) B3LYP for Size C cluster with (bottom continuous curve) and without (bottom dash curve) edge states, LDA (LDA-1\cite{13} and LDA-2\cite{14}), and GGA (GGA-1\cite{15} and GGA-2\cite{16}) results. Energies are relative to $E_F$ (absolute value -3.7 eV) and Lorentzian with FWHW=0.2 eV is used for convoluting B3LYP results. (b) Enlarged B3LYP results above $E_F$. Dominant spectral peaks are labeled by star or daggers. Selected from Paper 3, reprinted with permission from American Physical Society.

Influences of stacking, edge, and defect effects

The XAS of few-layer graphene (FLG) was also measured by Pacilé et al.\cite{11}, especially the peak 288 eV was explained as an interlayer state. While our calculations for one-layer (1L), two layer (2L), and three-layer (3L) GNRs based on size A model does not support such an assignment as the presence of the multilayers can also smear out the spectral profile around this energy (see Figure 4.12b). The stacking effect can have evident influence of spectra in peak 4 in the $\pi$ spectra, while it has nearly no influence in the $\sigma$ spectra. There is also no evidence for the 283.7 eV peak.

We begin to consider the gap between theory and experiment. Experiments treat real samples which have edges and also possible defects; while our calculations deals with an the ideal model. A consideration of the edge effect is realized by performing ECH calculations for carbons with different distances to the center of the model (see Figure 4.12c3, labeled as $a - h$). Our calculations show a red-shift of the main peak in the $\pi$ spectra and a new peak at 288 eV in the $\sigma$ spectra, both contributed by the outmost carbons. However, as the experiments with polarized x-rays\cite{11} were biased to strengthen the $\pi$ transitions, the weak
feature in the $\sigma$ spectra may not be observed in the experiments.

As too many possible defects exist, we choose only three representative defects, namely mono-vacancy (1V), di-vacancy (2V) Stone-Wales (SW)\cite{165} defects, and perform ECH calculations for selected non-equivalent carbons. For example, in the SW-defected model, we find strong evidence of the peaks both at 283.7 and 288 eV (Figure 4.12d). Our calculations show that defect can have the largest influence to the electronic structure of graphene, and we can explain that the two experimental features are most probably due to the defects, and they do not exist in an ideal graphene.

Alpha and beta electrons

![Graph showing Alpha and Beta Electrons](image)

Figure 4.14 Spin-specific contributions to the C 1s$\rightarrow$ $\pi^*$ NEXAFS spectra of GNRs with different sizes: (a) 2.4 nm $\times$ 1.5 nm; (b) 5.0 nm $\times$ 4.4 nm. Calculations are run at the UB3LYP level.

Interestingly, during the study, we also find that the spectra calculated from $\alpha$ and $\beta$ electrons are different in a small-sized GNR, and they approach nearly the same when the size increases. Figure 4.14 shows the 1s$\rightarrow$ $\pi^*$ NEXAFS spectra of graphene with two sizes: 2.4 nm $\times$ 1.5 nm and 5.0 nm $\times$ 4.4 nm. In panel a, the $\beta$ spectrum shows a clear splitting of the main peak around 284-286 eV, which makes it different from the $\alpha$ one. Such a difference disappears when the size increases, as shown in panel b. This observation might be related to the fact that small hydrogenated GNR could have peculiar spin property\cite{164}.
4.3 Short peptides

Why short peptides?

A peptide is a short polymer of amino acids connected with peptide bonds (O=\text{C-NH}). One or more polypeptides can fold to protein with biological function. Short peptides with a few amino acids have attracted attention from researchers. Because this could provide a starting point for the understanding of more complex proteins. Besides, sometimes they are considered as modulators of protein terminals.

Solvent effect: two categories of approaches

Our interest lies in the solvent effect of a peptide in water solution. Solvents could greatly influence the local chemical and electronic structures of the solute molecule by hydrogen bonds, electrostatic and polarization effects, etc. The theoretical investigation in solution phase is more complex than in gas phase due to larger size of the models and more possible configurations. The overall scope is to develop a method to probe the NEXAFS spectra of peptides in dilute solution, which could effectively reproduce the influences of solvent molecules from both short-range and long-range.

Two categories of approaches to treat the solvent effect are implicit- and explicit-solvent methods. In the former one, the structure of the solvent is ignored, the solvent molecules are replaced by a continuous dielectric with dielectric constant $\varepsilon_r$ (also called relative permittivity) that surrounds a cavity containing the solute molecule. This is a computationally economic way to treat the interaction between a solute and the solvent molecules. The interaction with the continuous dielectric is represented by a term that is added to the molecular Hamiltonian of the solute. An electric field (called reaction field) at the solute molecule is generated by the polarized solvent, and the electronic wavefunction of the solute molecule changes to respond to the interaction. The problem can be solved iteratively until self-consistency between the charge distribution of the solute and the solvent reaction field is achieved. Such a solution is thus often known as the self-consistent reaction-field (SCRF) model. Kinds of SCRF models exist, which differ in the way to define the cavity and the reaction field. In the widely-used polarizable continuum model (PCM), the cavity is defined as a set of spheres centered at each atom. Many models are developed based on the PCM model, for example, the isodensity PCM (IPCM) model, the self-consistent IPCM (SCI-PCM) model,
and so on. In the latter category, many individual solvent molecules are explicitly included. This can reproduce more accurate interactions but more computational cost is also needed. Besides, as the solvent can adopt different geometries around the solute, a statistical average over different configurations is also necessary.

We employ the explicit-solvent models by using the ECH method. Statistical average is accounted from classical molecular dynamics simulations, and different size of models is constructed to achieve converged NEXAFS spectral profile. Meanwhile, we also test the performance of the PCM model\cite{166,167} implemented in the GAUSSIAN 09 package.\cite{52}

**Influence of solvent shells on the absorption spectra**

The blocked alanine molecule (Ac-ALA-NHMe, denoted as bALA), which is often considered as a modulator of the peptide backbones, is chosen as our system. It contains two nitrogens N\(_1\) and N\(_2\) which are easy to form hydrogen bonds with water molecules, as visualized in Figure 4.15. It is solvated in about 1000 waters, and classical molecular dynamics with the OPLS-AA force field\cite{168,169} is employed to generate the structures in isothermal-isobaric ensemble with temperature at 300 K and pressure at 1 atm by using the fDYNAMO package\cite{170,171}. We choose 100 snapshots from the 1 ns production phase (one per 10 ps) for further spectral calculations. As we are interested in the N\(_1\)s NEXAFS spectra, the distance \(r\) between the water oxygen (O\(_w\)) and the hydrogen linked to nitrogen (H\(_n\)) is used as a criterion, and waters with \(r\) no less than a given distance threshold \(\xi\) are selected to construct cluster models.

Figure 4.15a illustrates the size dependence of the N\(_1\)s spectra calculated with the ECH method for snapshot 1, wherein \(\xi = 2.6\) and 7.0 Å include the first and second solvation shells of H\(_n\), respectively. The spectra almost converge at \(\xi = 7.0\) Å, exhibiting a sharp peak at 400.6 eV (peak \(i\)) and a broad peak at 403.7 eV (peak \(ii\)). However, a larger \(\xi\) is still necessary to be employed in order to get more accurate spectra, since \(\xi = 10.0\) Å gives a peak \(i\) with 0.2 eV lower energy and a slightly wider peak \(ii\). In panel b, the spectra calculated with the FCH method are visualized. As more computational cost is needed for a FCH calculation, models with \(\xi\) up till 5.0 Å are considered. At each model, the FCH and ECH methods reproduce roughly similar spectra except for the isolated peptide (\(\xi = 0.0\) Å). This is because the replacement of O\(^+\) for a core-excited N gives a much extended core-excited state; while the existence of a few neighboring waters stabilizes the core hole and makes ECH a better approximation for these cases.
4.3 Short peptides

Figure 4.15  Size dependence of N1s NEXAFS spectra of snapshot 1 calculated with the (a) ECH and (b) FCH methods. Each size is labeled by the distance threshold $\xi$ in Å, and number of water is included in parenthesis. Structures of selected models are shown around the spectra. Bars denote the IP values.

Figure 4.16 presents atom-specific spectra of models with $\xi = 0.0$, 2.6, and 10.0 Å for snapshot 1, as well as final-state MOs of selected transitions. Peak 1 denotes the transition to the $\pi^*$ orbital of the peptide bond around N$_1$, which always appears as the lowest transition when increasing the number of waters. The MO features of the three models keep nearly the same and show little correspondence with waters. Peak position has small red shift (from 401.3 eV to 400.4 eV) as the slight distribution of electrons to neighboring waters lowers the energy of the core-excited state. Besides, it becomes well separated from the other transitions ever since the inclusion of the waters in the first solvation shell ($\xi = 2.6$ Å). Peak 2 denotes the first transition of N$_2$, and shows significant difference in MOs between the isolated and solvated peptide. The former case is mainly a Rydberg state of the excited nitrogen as the MOs localize mostly in its outer $s$-orbitals; considerable electronic distribution however also exists in the $\pi^*$ orbital of the peptide bond. While in the latter case, the electrons mostly localize in the peptide bond. Peak 3 also shows a large influence of solvent: it stays both on the $\pi^*$ orbital of the peptide bond and the outer $p$ orbital of N$_2$ at $\xi = 0.0$ and 2.6 Å; while it spreads a lot to the solvents at $\xi = 10.0$ Å. The strong intensity has also smeared out when more solvent molecules are included. Peak 4 is well delocalized in the whole peptide in all the three models, and the inclusion of waters attracts part of the electron distribution from the peptide. Besides, the increase of size makes a dense DOS in the largest model, and transitions with similar energies constitute
Figure 4.16 Atom-specific N1s NEXAFS spectra calculated with ECH method using models constructed with thresholds: (a) 0.0 Å; (b) 2.6 Å; (c) 10.0 Å. Selected bars are labeled by numbers and corresponding final-state MOs are visualized in panel (d).
4.3 Short peptides

![Graphs showing N1s NEXAFS spectra](image)

Figure 4.17  Statistics of N1s NEXAFS spectra calculated with the ECH method based on models constructed at $\xi=7.0\,\text{Å}$: (a) 10 randomly chosen geometries; (b) averages over all 100 snapshots (solid), snapshots 1-50 (dash-dot), and snapshots 51-100 (dot). (c) N1s NEXAFS spectra calculated with ECH-PCM method with different size of models (each characterized by $\xi$ in Å, and number of waters is included in parentheses) based on snapshot 1. Spectra have been shifted for better comparison with the statistical result.

to a strong peak at around 404 eV.

**Statistical effects**

From above discussions, we then employ $\xi = 7.0\,\text{Å}$ to construct models and perform ECH calculations for the rest snapshots, and statistical results are illustrated in Figure 4.17. Panel a depicts the spectra of 10 randomly-selected snapshots, and significant impact of configurations to the spectra is observed. Each geometry exhibits a $\pi^*$ peak with similar shape, and the energy covers a range of about 1 eV (400.4 eV–401.3 eV). While large difference appears in the high-energy region. This is because the first peak comes from localized transition to the peptide bonds, while the high-energy region depends more on the geometries of both the peptides and the waters. In panel b, the averaged spectra over the 100 snapshots are presented, and the mean values over the first and last 50 snapshots are also given for comparison. Similarity among the three curves verifies that these geometries can represent the configurational influence with acceptable accuracy, exhibiting a $\pi^*$ peak at around 400.5–400.7 eV and a broad peak at 404.0 eV. However, more configurations are still necessary so as to get a refined statistical spectra.
On the other hand, ECH-PCM results with small or medium models based on snapshot 1 are also plotted in Figure 4.17c. The convergence of peak $i$ can be reproduced at very small models, as this is less related to the solvents. While for the high-energy part, a small model ($\xi = 2.6, 3.5 \text{ Å}$) or a bare peptide (not shown) can not well predict a statistical effect of the solvents; while a moderate model ($\xi = 4.0, 5.0 \text{ Å}$) can roughly account for the statistical effects, as compared with the spectra in panel b. Further increasing $\xi$ neglects more statistical effects of the near-coordinated waters as well as adds more dependence on the given snapshot geometry. It just gives a similar spectrum to the ECH one and thus is not recommended to be used.

**Cavity-protection of short peptides**

![Figure 4.18](image)

Figure 4.18  
(a) Structural illustration of peptides 2-5.  
(b) Full view of the crystal structure of 2cry<sub>⊂</sub>1. Among six Pd atoms and three Zn atoms in the cage 1, only one Zn atom forms a Zn-O bond with the peptide in the crystal structure. Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

Understanding factors that control the folding of peptides or proteins is still a fundamental problem in biochemistry. However, the existence of solvents can greatly influence the folding through hydrogen bonds, electrostatic, or polarization effects, which adds the complexity of problem. Thus, gas-phase studies which better reflect the intrinsic propensities of peptides, have attracted the interests of researchers. Experimentally, one can use the infrared-ultraviolet spectroscopy double resonance technique\textsuperscript{[172–176]} as well as mass spectrometry techniques for gas-phase peptide ions.\textsuperscript{[177–179]} Another way is to encapsulate short peptides into
4.3 Short peptides

Figure 4.19  B3LYP and MP2//B3LYP energies at B3LYP-optimized structures for several low-lying conformers of each peptide in the gas phase. Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

the cavity of self-assembled cages to study their folding behaviors, since these cages provide a local nearly-solvent-free environment.\[180\text{–186}\] Such a solution constructs so-called host-guest systems, which cover a large class of complex systems in chemistry and biology.

Our work is inspired by the experimental work by Fujita and coworkers\[183\]. They have successfully encapsulated four short peptides (shown in Figure 4.18a, denoted as peptides 2 – 5) into a prism-like self-assembled cage (Figure 4.18b, denoted as host 1). The cage has 357 atoms, including 3 Zn(II) and 6 Pd(II) atoms, with a total charge of +12. They suggested that the conformations of these peptides in this cage corresponds to the lowest-energy geometries of these peptides in the gas phase, rather than restricted conformations resulted from the confinement of 1. Clearly, computational study can provide valuable information on the guest-host interaction and its influence on the conformations of peptides, which are not directly accessible from experiments. It is interesting to investigate whether the cavity-protected peptides can represent their conformational propensities in the gas phase.
Isolated peptides

Our strategy is to study both the isolated peptides and peptide-cage inclusion complexes by comparing energies of low-lying conformers with those of the optimized crystal structures. We first study the peptides in the gas phase. Classical MD using the Amber ff03 force field\textsuperscript{[187,188]} is first used to generate candidate geometries followed by relaxation at the same level. Then, 10-20 low-lying isomers are selected, and DFT with the B3LYP functional is used for further geometry optimizations. Finally, MP2 single-point calculations at B3LYP-optimized structures are used for verification. Figure 4.19 gives the B3LYP and MP2/B3LYP energies of five low-lying isomers (denoted as X\textsubscript{a}, X\textsubscript{b}, \cdots, X\textsubscript{e} with X=2-5) compared with the optimized crystal structure (denoted as X\textsubscript{oe}) for each peptide. One can clearly see that crystal structure does not correspond to the lowest-energy isomer in the gas phase. The agreement of B3LYP and MP2 energies verifies that our calculations are reliable.

Hybrid approaches for peptide-cage complexes

On the other hand, for the inclusion complex, we first dock the gas-phase optimized peptides to the cage by using the Monte Carlo approach, and then ONIOM\textsuperscript{[189]} (B3LYP:PM6\textsuperscript{[190]}) method is used for the geometry relaxation, followed by a full B3LYP verification. The energies are also compared to that of the optimized crystal structure of the inclusion complex. Table 4.2 lists the relative energies of the ONIOM-optimized conformers of peptide-cage inclusion complexes. One can see that the ONIOM relative energies exhibit significant differences to the B3LYP relative energies. The relative energy order of different inclusion complex conformer is quite different from that of isolated peptide conformers in the gas phase, which shows the influences of the cages. And for each peptide, the optimized crystal structure of the inclusion complex, X\textsubscript{oe}⊂1, is among two or three lowest-energy conformers. For peptide 2, 2\textsubscript{oe}⊂1 is predicted to have the lowest energy; while for other peptides, they appear only 3-6 kcal/mol above the lowest energy conformer. It shows that our results are qualitatively right. Quantitatively accuracy is however not achieved due to the approximations used in our methods. Figure 4.20 illustrates the optimized structures of peptide 2 in both the gas phase and the inclusion complex. One can find the geometrical change from gas phase to the cage locates mostly around the O(Ac), the strong Zn-O bond (with bond length around 2.3 Å) is mainly accounted for the host-guest interaction. The underlying reason for such local structural changes is mainly that when the Zn-O bond forms
Table 4.2 Relative energies (in kcal/mol) of several low-lying conformers at ONIOM-optimized geometries for the inclusion complexes. The energy of the optimized crystal structure for the inclusion complex (Xoe⊂1) is taken as zero (X = 2, 3, 4, 5). Selected from Paper 5, reprinted with permission from John Wiley & Sons, Inc.

<table>
<thead>
<tr>
<th>Conformer</th>
<th>E(ONIOM)</th>
<th>E(B3LYP//ONIOM)</th>
<th>E₀(B3LYP//ONIOM)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>2⊂1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a⊂1</td>
<td>0.56</td>
<td>8.08</td>
<td>8.51</td>
</tr>
<tr>
<td>2b⊂1</td>
<td>6.53</td>
<td>17.85</td>
<td>19.09</td>
</tr>
<tr>
<td>2c⊂1</td>
<td>6.80</td>
<td>12.74</td>
<td>12.81</td>
</tr>
<tr>
<td>2d⊂1</td>
<td>-0.55</td>
<td>4.39</td>
<td>4.42</td>
</tr>
<tr>
<td>2oe⊂1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2e⊂1</td>
<td>3.97</td>
<td>8.73</td>
<td>9.09</td>
</tr>
<tr>
<td>3⊂1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a⊂1</td>
<td>-1.08</td>
<td>-1.63</td>
<td>-1.31</td>
</tr>
<tr>
<td>3b⊂1</td>
<td>6.63</td>
<td>14.42</td>
<td>16.52</td>
</tr>
<tr>
<td>3c⊂1</td>
<td>-5.74</td>
<td>-5.10</td>
<td>-4.96</td>
</tr>
<tr>
<td>3oe⊂1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3d⊂1</td>
<td>0.14</td>
<td>6.23</td>
<td>6.48</td>
</tr>
<tr>
<td>3e⊂1</td>
<td>1.34</td>
<td>4.25</td>
<td>4.99</td>
</tr>
<tr>
<td>4⊂1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4a⊂1</td>
<td>7.06</td>
<td>20.65</td>
<td>22.28</td>
</tr>
<tr>
<td>4b⊂1</td>
<td>5.21</td>
<td>22.23</td>
<td>23.64</td>
</tr>
<tr>
<td>4c⊂1</td>
<td>2.37</td>
<td>10.13</td>
<td>11.74</td>
</tr>
<tr>
<td>4d⊂1</td>
<td>1.17</td>
<td>7.44</td>
<td>8.23</td>
</tr>
<tr>
<td>4e⊂1</td>
<td>-4.10</td>
<td>-5.70</td>
<td>-5.89</td>
</tr>
<tr>
<td>4oe⊂1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5⊂1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a⊂1</td>
<td>15.30</td>
<td>32.15</td>
<td>35.40</td>
</tr>
<tr>
<td>5b⊂1</td>
<td>3.99</td>
<td>15.88</td>
<td>17.95</td>
</tr>
<tr>
<td>5c⊂1</td>
<td>11.67</td>
<td>30.31</td>
<td>32.61</td>
</tr>
<tr>
<td>5d⊂1</td>
<td>-3.75</td>
<td>-3.04</td>
<td>-3.20</td>
</tr>
<tr>
<td>5oe⊂1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5e⊂1</td>
<td>-4.38</td>
<td>2.13</td>
<td>1.26</td>
</tr>
</tbody>
</table>

a The B3LYP//ONIOM energy plus ZPVE calculated at the ONIOM level.
between the cage and the peptide, the Zn-bound oxygen turns less favorable as a HB acceptor.

\[ \Delta E_b \equiv E^{\text{inclusion complex}} - (E^{\text{peptide}} + E^{\text{cage}}) = \Delta E^{\text{def}} + \Delta E^{\text{int}}. \] (4.1)

We then analyze the sum of Mulliken charge distribution on the peptide atoms in the inclusion complexes ($\rho^{\text{pep}}$). As the sum of Mulliken charges on the isolated peptide is zero, $\rho^{\text{pep}}$ represents the strength of charge transfer between the cage and the peptide during the encapsulation process. We find that $\rho^{\text{pep}}$ is in the range from -0.2 to -0.8, indicating that in all inclusion complexes the cage acts as an electron donor and the peptide as an electron acceptor. These inclusion complexes may be classified as metal-to-ligand charge-transfer complexes.\[191\] We also find that stronger peptide-cage interaction energy $\Delta E^{\text{int}}$ usually corresponds to more charge transfer. Thus, we have provided a better understanding of the host-guest interaction.
4.4 Multipole expansion

Why multipole expansion?

The GEBF method\cite{108} has shown its advantage in reproducing the energies, and low-order energy derivatives like dipole moments, polarizabilities, forces, force constants, infrared (IR) and Raman intensities, as compared with those obtained from conventional QM methods.\cite{136,137,192–196} The main approximation of the GEBF method is to use point charges to represent the distant atoms. The interaction energy between two fragments $I$ and $J$ are approximately obtained from underline calculations with one or two fragments represented as point charges (see eq 3.4). The electrostatic interaction of distant atoms has been well considered; and the polarization effect is partly reproduced from the generation of atomic point charges in a self-consistent way.

During our further extension of this method to higher-order properties, we find that the current GEBF method provides strange first hyperpolarizability ($\beta$) results. We think that the problem with $\beta$ is because the point charge is insufficient to represent the high-order interactions. The usage of point charges is related to multipole expansion (ME) for the Coulomb potential, with each term coming from charges, dipoles, quadruples, and so on.

A two-fragment system at RHF level

![Figure 4.21 Schematic illustration for (a): a two-fragment model system H$_3$O$^+$···OH$^-$; and (b-c) its two subsystems. The fragments containing real atoms are labeled as $I$ or $J$, and those containing point charges (denoted by “+“ and “-“) are labeled by $I'$ or $J'$.

Our strategy is to derive the GEBF total energy from the multipole expansion first, then refinement is expected to be achieved by including higher-order terms. We start with the simplest case, i.e. a two-fragment system at RHF level. The conclusion from this model system can be easily extended to a general large molecule. Figure 4.21 illustrates a representative two-fragment (with fragments
I and J) system, H$_3$O$^+$···OH$^-$, as well as two subsystems within the GEBF method. The GEBF total energy (eq 3.5) can be explicitly written as,

$$E_{IJ}^{\text{GEBF}} = \tilde{E}_{\text{sub1}} + \tilde{E}_{\text{sub2}} - \sum_{A<B} \frac{Q_A Q_B}{R_{AB}}.$$  \hspace{1cm} (4.2)

We need to prove the approximate equivalence between the GEBF-HF ($E_{IJ}^{\text{GEBF}}$) and conventional HF ($E_{IJ}$) energies, i.e.,

$$\Delta E \equiv E_{IJ}^{\text{GEBF}} - E_{IJ} \approx 0,$$  \hspace{1cm} (4.3)

where the total HF energy of the target system $E_{IJ}$ is given by,

$$E_{IJ} = \frac{1}{2} \sum_{\mu} \sum_{\nu} P_{\nu \mu} (H_{\mu \nu}^{\text{core}} + F_{\mu \nu}) + \frac{1}{2} \sum_{A=1}^{M} \sum_{B=1}^{M} Z_A Z_B R_{AB},$$  \hspace{1cm} (4.4)

with $P_{\nu \mu}$, $H_{\mu \nu}^{\text{core}}$ and $F_{\mu \nu}$ being the density, core Hamiltonian and Fock matrices, respectively. The energy can further be decomposed as

$$E_{IJ} = E_I + E_J + E_{IJ},$$  \hspace{1cm} (4.5)

where $E_I$ and $E_J$ are one-body energies and $E_{IJ}$ denotes the two-body energy.

If we assume full transferability of density matrix from the target system to the subsystems, the subsystems can provide exact one-body energies, while the two-body energy needs to be examined in an approximate way.

The two-body interaction energy is actually weighted summation of a series of one-electron (1e) and two-electron (2e) integrals, which involves basis functions and/or nucleolus in different fragments. The $\frac{1}{r_{1A}}$ and $\frac{1}{r_{12}}$ terms (1, 2 denote electrons, and A labels nucleolus) in $E_{IJ}$ can be expressed by ME within the Cartesian coordinates$^{[197]}$:

$$\frac{1}{r_{1A}} = \frac{1}{R_{AB}} + \frac{\mathbf{R}_{AB} \cdot \mathbf{r}_{1B}}{R_{AB}^3} + \cdots,$$  \hspace{1cm} (4.6)

$$\frac{1}{r_{12}} = \frac{1}{r_{1A}} + \frac{\mathbf{r}_{1A} \cdot \mathbf{r}_{2A}}{r_{1A}^3} + \cdots.$$  \hspace{1cm} (4.7)

The above expressions can be cut at the leading terms (Born approximation), when the distances of electron-nucleolus or two electrons are far away (i.e. the distance between the two fragments is large). And finally we have proved that
\[ \Delta E = \delta_1 + \delta_2 + \delta_3 - T_4 \approx 0, \]  

(4.8)

where the terms \( \delta_1, \delta_2, \delta_3, \) and \(-T_4\) all have small values at large fragment-fragment distance. Thus we have proved that the GEBF method can be derived from the ME at the zeroth-order.

Numerical tests at HF/6-31G** level for the model system is listed in Table 4.3. One can find the error terms decrease with the fragment-fragment distance (minimum atom-atom distance, denoted as \( R_{frg} \)), and each term at 4.0 Å has the accuracy of several milli-Hartrees (mHs). Besides, one can see that the error components are canceled, which leads to a very small \( \Delta E \). It reveals that the high accuracy of the GEBF equation lies in a balanced treatment in approximating the 1e and 2e integrals. Besides, one can see that the distance threshold \( \xi = 3.5–4.0 \) Å chosen in our previous works is reasonable for small or medium-sized basis sets. We will extend it to the first order in the near future, which is expected to improve the accuracy for some ground-state property calculations.

<table>
<thead>
<tr>
<th>( R_{frg}/\text{Å} )</th>
<th>( \delta_1 )</th>
<th>( \delta_2 )</th>
<th>( \delta_3 )</th>
<th>(-T_4)</th>
<th>( \Delta E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>-18.14</td>
<td>-3.76</td>
<td>21.81</td>
<td>-141.07</td>
<td>-141.16</td>
</tr>
<tr>
<td>4.0</td>
<td>-3.94</td>
<td>4.30</td>
<td>-0.37</td>
<td>-2.72</td>
<td>-2.72</td>
</tr>
<tr>
<td>5.0</td>
<td>-2.10</td>
<td>3.21</td>
<td>-1.11</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>6.0</td>
<td>-1.26</td>
<td>2.47</td>
<td>-1.20</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7.0</td>
<td>-0.82</td>
<td>1.97</td>
<td>-1.15</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Chapter 5

Summary of included papers

5.1 Soft x-ray spectroscopies

The C1s or N1s soft x-ray spectra of some large systems in the gas, solution or solid phases have been studied at the DFT levels. Early theoretical spectra are mainly on small or medium-sized systems. Aiming at better describing the intermolecular interaction for a real system, spectral convergence has been obtained based on a series of cluster models with increasing sizes. Aiming at better modelling the real experimental situations, the effect of external perturbations have also been considered.

5.1.1 DNA and related molecules, Papers 1-2

N1s x-ray spectra were studied for DNA and related molecules. For NEXAFS spectra calculations, ECH approximation were used for large models, while FCH approximation were employed for small molecules. Ground state wavefunction were used for XES calculation, and the transitions to imine (-N=) and amine (-NH-) nitrogens were selected to constitute the corresponding RIXS spectra. The amine and imine nitrogens exhibit noticeable contributions in all studied spectra (peak position difference denoted as $\Delta$).

In Paper 1, NEXAFS, XES, and RIXS spectra were studied for poly(dG)-poly(dC) and poly(dA)-poly(dT) DNA duplexes by both experiments and calculations. Starting from GC and AT base pairs, we created models of DNA duplex with up to 12 base pairs. We found that the stacking effect of base pairs has very small influence on all kinds of spectra studied. And for a general DNA, its spectra
can be well reproduced by linear combinations of GC and AT base pairs weighted by their ratio.

In Paper 2, NEXAFS spectra of G, C, and related molecules in the gas and solid phases were calculated and influence of intramolecular and intermolecular interactions were analyzed. In the gas phase, hydrogen tautomerism has the largest influence to $\Delta$ (\sim 0.5 eV), while sugar and phosphate groups have less or nearly no influence. For solid-state spectra, cluster models were created from crystal structures, and were further “expanded” or “compressed” to reveal the environmental perturbations. We found that change of HBs of 0.2-0.3 Å can leads to 0.2-0.8 eV change of $\Delta$ depending on different crystal structures; while the stacking effect has nearly no influence to the spectra. Early discrepancies between experimental and calculated spectra have been overcome and clarified. The potential of NEXAFS spectra in refining HB structures in DNA is highlighted.

5.1.2 Graphene, Papers 3

We calculated the C1s-$\pi^*$ and C1s-$\sigma^*$ NEXAFS spectra of graphenes by the ECH method at the UB3LYP level and systematically analyzed the influence of size, stacking, edges and defects. Hydrogen-terminated graphene nanoribbons of different sizes were used to represent the infinite graphene sheet. We found the size-dependence and convergence of the spectra. Stacking effects were found to smear out the double-peaks feature of the first main $\pi^*$ peak (around 285 eV) and blue shift of the second $\pi^*$ structure by about 2 eV in the $\pi$ spectra. The excitation from edge atoms have a red-shift in the $\pi$ spectrum and can introduce a new weak in the $\sigma^*$ peak around 288 eV. New spectral features in both $\pi$ and $\sigma$ regions can be brought in by defects. The debate on the interpretation of the new feature has been resolved.

5.1.3 Solvent effect of peptide, Paper 4

The N1s NEXAFS spectra of a blocked alanine molecule in water solution have been investigated using cluster models constructed from molecular dynamics snapshots. With one ordinary snapshot, we first show that the convergent spectra with respect to system size require at least the inclusion of the second solvation shell. The statistics over different configurations have been realized by averaging the spectra of 100 snapshots. Comparison has also been made with those obtained
from the PCM model, and we have demonstrated that the calculation based on a medium-sized peptide-water cluster can roughly reflect the configurational effects.

5.2 Ground state energies and structures

5.2.1 Host-guest system, Paper 5

We studied the conformational preferences of four short peptides in a self-assembled cage. We used classical MD to generate the candidate geometries, then optimized selected low-lying geometries at B3LYP level, and verified at single-point MP2 calculations. We then docked the gas-phase optimized conformers into the cage by Monte Carlo simulations, followed by ONIOM optimization and verification by single-point B3LYP calculations. We revealed that the conformation of a peptide in the cage does not correspond to its lowest-energy conformation in vacuum, caused by the Zn-O bond formed between the peptide and the cage, and the confinement effect of the cage.

5.2.2 Multipole expansion, Paper 6

We derived the GEBF energy equation at the Hartree-Fock level by keeping the leading term of the multipole expansion of the electrostatic potential. Our numerical calculations for a model system showed that the balanced treatment in approximating one-electron and two-electron integrals is mainly responsible for the high accuracy of the GEBF equation. The numerical results suggested that for small or medium-sized basis sets the distance threshold (3.5–4.0 Å) chosen in our previous works is reasonable, but for large basis sets a larger distance threshold should be used. This work also suggests that the GEBF approach may be further refined by introducing higher-order terms of the electrostatic potential. Such refinement is expected to improve the accuracy of the GEBF approach for ground-state energy and some property calculations.

5.2.3 Automatic fragmentation implementation, Paper 7

We presented an automatic and efficient implementation of the GEBF approach. In this scheme, one functional group can be defined as a fragment. Correspondingly, as the fragment-size decreases and number of fragments increases, we provided a fast scheme to derive the coefficients by using compound index. The
scheme is now applicable for a general large molecule, including structurally-
difficult systems, for example, the supermolecules with fused aromatic rings. Test
calculations at the HF and DFT levels validated that the GEBF approach can
produce reasonably accurate ground-state energies and optimized structures com-
pared to conventional HF or DFT calculations.
References


[3] Zubavichus, Y.; Shaporenko, A.; Korolkov, V.; Grunze, M.; Zharnikov, M. J. Phys. Chem. B 2008, 112, 13711–13716, and Figure 4 therein, with the experimental spectrum made on pristine powder samples of guanine, is embedded to our Figure 4.5b for comparison.


[7] Chen, W.; Wu, X.; Car, R. Phys. Rev. Lett. 2010, 105, 017802, and Figure 1a therein, with the calculated spectra of Ice-Ih, is recaptured to our Figure 4.8 for comparsion.

[8] Prendergast, D.; Galli, G. Phys. Rev. Lett. 2006, 96, 215502, and Figure 1 therein, with the calculated spectra of Ice-Ih, is recaptured to our Figure 4.8 for comparsion.


[13] Trickey, S. B.; Müller-Plathe, F.; Diercks, G. H. F.; Boettger, J. C. Phys. Rev. B 1992, 45, 4460, and Figure 1 therein, with the PBC-LDA predicted DOS of single-layer graphene, is recaptured to our Figure 4.13a.

[14] Klintenberg, M.; Lebègue, S.; Ortiz, C.; Sanyal, B.; Fransson, J.; Eriksson, O. J. Phys.: Condens. Matter 2009, 21, 335502, and Figure 2 therein, with the PBC-LDA predicted DOS of single-layer graphene, is recaptured to our Figure 4.13a.

[15] Duplock, E. J.; Scheffler, M.; Lindan, P. J. D. Phys. Rev. Lett. 2004, 92, 225502, and Figure 3a therein, with the PBC-GGA (revised PBE) predicted DOS of single-layer graphene, is recaptured to our Figure 4.13a.


REFERENCES


REFERENCES


