STRUCTURES AND QUANTUM EFFECTS OF SHORT HYDROGEN BONDS IN
CONDENSED PHASE SYSTEMS

By

SHENGMIN ZHOU

A dissertation for phd submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology

Written under the direction of

Lu Wang

And approved by

__________________________________

__________________________________

__________________________________

__________________________________

New Brunswick, New Jersey
January 2022
ABSTRACT OF THE DISSERTATION FOR PHD

Structures and Quantum Effects of Short Hydrogen Bonds in Condensed phase Systems

by SHENGMIN ZHOU

Dissertation for PhD Director: Lu Wang

Short hydrogen bonds (SHBs), which have the donor and acceptor separations below 2.7 Å, occur widely in condensed phase system and exhibit prominent quantum mechanical characters. SHBs have been associated with crucial functions in biological macromolecules, and hence I first conduct a statistical analysis on atomic-resolution biomolecular structures from the Protein Data Bank and demonstrate the structural and chemical features of biological SHBs. The exact determination of the geometry and functional roles of SHBs is often subject to the limited sources of high-resolution protein structures. As such, I collaborate with the group of Professor Sijian Wang to develop boosting based machine learning models to predict the presence of biological SHBs in a protein structure with moderate or low resolution. Using electronic structure calculations, I further elucidate how the interplay of the structural and chemical features determines the proton potential energy surfaces and the proton sharing conditions, and how the competition of intermolecular interactions results in different preference to form SHBs between amino acid side chains. As one of the spectral signatures of SHBs is their highly downfield (> 14 ppm) $^1$H NMR chemical shift, I carry out first principles simulations on a set of model molecules to assess how quantum effects determine the symmetry and chemical shift of their SHBs. From these simulations, I reveal a universal relation between the instantaneous chemical shift and the position of the proton in a SHB and develop a metric that allows one to determine
the proton position directly from its $^1$H chemical shift. Besides biological systems, SHBs have been observed in protic ionic liquids and aqueous solutions. I apply first principles simulations and IR spectra calculations to the mixture of 1-methylimidazole and acetic acid and the aqueous solutions of bifluoride ions to demonstrate how quantum effects promote the delocalization of the hydrogen atom between acid and base in protic ionic liquids and how SHBs disturb the hydrogen bond structures of liquid water.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Lu Wang. Without her guidance and encouragement over the last five years, this dissertation would not have been possible. As the second graduate student in her group, I always get opportunities to learn new things from her and discuss research with her. Besides research, I have learned a lot about how to write a descent scientific article and how to make a great presentation. Her help on my academic career will always be the most valuable asset in my whole life. Also, her efforts and enthusiasm on science make me understand what a great scientist is.

I would also like to thank my committee members Professor David Case, Professor Darrin York and Professor Sijian Wang for their help all the way. With their guidance and helpful suggestions on my qualify exams and annual meetings, I could make sure I am on the right path to purse my degree. In addition, I really appreciate that they are always supportive on my job searching. With their letters of recommendations, I successfully get an offer for a postdoctoral position from the Argonne National Laboratory.

I would like to thank my collaborators for doing many amazing works together. In particular, I want to thank Professor Sijian Wang and Yuanhao Liu from the Department of Statistics of Rutgers University. I appreciate how much they have taught me about machine learning methods, and our close collaboration makes it possible to finally develop a web server to predict biological short hydrogen bonds and contribute to the scientific community. I also want to thank Professor Daniel Kuroda, Dr. Xiaoliu Zhang and Dr. Fedra Leonik from the Department of Chemistry of Louisiana State University. I am grateful for their help in the experimental measurements, which has brought me into the field of ionic solutions. Finally, I want to thank Professor Andrei Tokmakoff and Dr. Bogdan Dereka from the Department of Chemistry of the University of Chicago. We have had a lot of fruitful discussions about how to study the structures and spectra of aqueous solutions.

I would like to thank all the members in Professor Lu Wang’s group, Dr. Yaoyukun
Jiang, Yuxuan Wu, Cheng Qian, Wenting Meng and Zelin Wang, for their help and support. It’s my pleasure to work with a group of amazing people, and I will always remember the life when we work together and support each other.

I also would like to thank the Department of Chemistry and Chemical Biology and the Institute for Quantitative Biomedicine of Rutgers University to provide such a wonderful environment for me to study and perform research. In particular, I want to thank the Office of Advanced Research Computing to offer the computing sources in the powerful servers Amarel and Caliburn. Without their support, I would not be able to conduct my research in the field of computational chemistry.

Last but not least, I would like to thank my whole family and friends for supporting every decision in my life. I appreciate my parents to give me life and make it possible for me to experience different things in the world. Their cultivation on me makes me who I am today. I am also grateful to all my friends and their care and support help me go through every hard time in my life.
TABLE OF CONTENTS

Abstract ................................................................. ii

Acknowledgments ...................................................... iv

List of Tables ........................................................ x

List of Figures ........................................................ xii

Chapter 1: Introduction and Background ......................... 1

Chapter 2: Structural and chemical features of biological short hydrogen bonds in proteins ........................................ 8
  2.1 Introduction ......................................................... 8
  2.2 Methods .......................................................... 9
  2.3 Results and discussion ........................................... 11
    2.3.1 Validation of the biomolecular structures ................. 11
    2.3.2 Short hydrogen bonds in biological macromolecules .... 11
    2.3.3 Structural and chemical features of short hydrogen bonds in proteins ........................................ 14
    2.3.4 Proton potential energy surfaces for side chain-side chain short hydrogen bonds .................................. 17
    2.3.5 Hybrid hydrogen bond networks in proteins ............... 22
  2.4 Conclusion ......................................................... 27
Chapter 3: Effective prediction of short hydrogen bonds in proteins via machine learning method

3.1 Introduction ................................................................. 29
3.2 Methods ................................................................. 30
  3.2.1 Preparation of the hydrogen bond dataset ................. 30
  3.2.2 Development of the machine learning model .......... 31
  3.2.3 Electronic structure calculations ......................... 31
3.3 Results and discussion .............................................. 32
  3.3.1 Dataset for biological hydrogen bonds .......... 32
  3.3.2 Machine Learning model for the prediction of SHBs in proteins ........ 33
  3.3.3 Side-chain Tyr-Asp and Tyr-Glu pairs are most likely to form SHBs 39
3.4 Conclusions .............................................................. 48

Chapter 4: Structural and chemical features of biological short hydrogen bonds in protein-ligand complexes

4.1 Introduction ................................................................. 50
4.2 Methods ................................................................. 50
4.3 Results and discussion .............................................. 51
  4.3.1 Short hydrogen bonds in protein-ligand complexes ........ 51
  4.3.2 Structural and chemical features of amino acid-ligand short hydrogen bonds ........ 54
  4.3.3 Machine learning model to predict amino acid-ligand SHBs in proteins ........ 57
  4.3.4 Quantum chemistry calculations of the amino acid-ligand SHBs ........ 60
4.4 Conclusions .............................................................. 63
Chapter 5: Symmetry and proton NMR chemical shifts of short hydrogen bonds  65

5.1 Introduction ........................................... 65
5.2 Methods .................................................. 67
5.3 Results and discussion .................................. 69
  5.3.1 Geometry and symmetry of the intramolecular SHBs ............... 69
  5.3.2 $^1$H NMR chemical shifts of the SHBs ........................... 77
5.4 Conclusions ............................................. 87

Chapter 6: Geometry and proton NMR chemical shifts of bifurcated short hydrogen bonds  90

6.1 Introduction ........................................... 90
6.2 Methods .................................................. 92
6.3 Results and discussion .................................. 94
  6.3.1 Structure of the bifurcated SHB .............................. 94
  6.3.2 Competing quantum effects in linear and bifurcated SHBs ........ 99
  6.3.3 Downfield $^1$H NMR chemical shifts of TAPH .................. 102
6.4 Conclusions ............................................. 108

Chapter 7: Short hydrogen bonds in acid-base mixtures  110

7.1 Introduction ........................................... 110
7.2 Methods .................................................. 112
7.3 Results and discussions .................................. 115

Chapter 8: Short hydrogen bonds in the aqueous solution of bifluoride ions  125

8.1 Introduction ........................................... 125
8.2 Methods ................................................................. 126

8.3 Results and discussion .................................................. 128
  8.3.1 Geometry of the F-H-F hydrogen bond ......................... 128
  8.3.2 Solvation structures of the bifluoride and fluoride ions ........ 129
  8.3.3 Hydrogen bonds in the solvation shells of the ions ............ 133
  8.3.4 IR spectra of the O-H stretch modes in liquid water and the ionic solutions ......................... 135

8.4 Conclusion ............................................................... 138

Chapter 9: Summary ......................................................... 139

Publication List ............................................................... 142

References ................................................................. 143
List of Tables

3.1 Summary statistics of the training and test datasets when the donor (D) and acceptor (A) residues of a hydrogen bond in the location combination (DA) is in the backbone (B) or side chain (S) of a peptide or protein. 33

3.2 The precision and recall of the MAPSHB model with different probability thresholds. The recommended threshold and its relevant metrics are highlighted in bold. 37

3.3 Occurrence and probability of surface exposure of the side chain-side chain hydrogen bonds when Tyr/Ser/Thr/His/Arg/Lys is the donor and Asp/Glu is the acceptor in the entire data set. 40

3.4 Decomposition of interaction energies of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds at R = 2.55 Å. The energies are in the unit of kcal/mol. 45

4.1 Chemical IDs in the PDB and the categories of the ligands in this study. 54

4.2 The precision and recall of the machine learning model with a probability threshold of 0.5 for 10 splits of training and test datasets. 58

4.3 Occurrence of the side chain of amino acids and functional group of ligands in hydrogen bonds when the number of SHBs is larger than 20 from our dataset. “D” and “A” denote that the ligand serves as the donor or acceptor, respectively, in a hydrogen bond. The functional groups are denoted using chemical formula: hydroxyl (-OH), phosphate (-H₃PO₄⁻²) and carboxylic acid (-COOH and -C(OH) with the underlined atom representing the specific atom in the hydrogen bonds. 61

5.1 $\nu_0$ and $\langle \delta_H \rangle$ from the AIMD and AI-PIMD simulations of the model molecules. The predicted and the experimental $^1$H NMR chemical shifts are included for comparison. 77
6.1 Average geometric parameters of the bifurcated SHB as obtained from the AIMD and AI-PIMD simulations of TAPH in the acetonitrile solution.

6.2 Chemical shifts and isotropic magnetic shielding constants of the hydrogen bonded proton in two configurations of TAPH and DMANH. $r_1$ represent the distance between the $N_1$ and $H$ atom in the hydrogen bonds. For TAPH, $S_{HB,12}$, $S_{HB,13}$ and $S_{HB,23}$ are the shielding constants from the dimeric $N_1$–$H$····$N_2$, $N_1$–$H$····$N_3$ and $N_2$–$H$····$N_3$ hydrogen bonds, respectively.

7.1 The number of molecules and the box size in the systems with different concentrations of the acetic acid ($X_{HA}$).

8.1 The number of hydrogen bonds per water in liquid water and the aqueous ionic solutions. The first and second solvation shells of the $F^-$ and HF$_2^-$ ions have $r_{FO} \leq 3.4$ Å and $r_{FO} \leq 5.4$ Å, respectively. The O–H–F hydrogen bonds can be further divided into two types, O–H–F$_r$ and O–H–F$_o$ in the solvation shells of a F atom. $F_r$ denotes the reference F atom in the ions to define the solvation shell, and $F_o$ denotes all the other F atoms except for the reference one in the solution.

8.2 Averaged values of $R$, $\nu$ and $r_{OH}$ of the O–H–O and O–H–F hydrogen bonds in liquid water and the $F^-$ and HF$_2^-$ solutions.
LIST OF FIGURES

1.1 Potential energy surface diagrams for transferring a proton in a hydrogen bond: (a) symmetric double-well potential, (b) asymmetric double-well potential, (c) low-barrier hydrogen bond potential, (d) single-well potential. The upper solid and the lower dashed red horizontal lines represent the zero-point energy levels for the donor-hydrogen or donor-deuterium bonds, respectively. ......................................................... 2

1.2 Examples of SHBs that form between amino acids in proteins and between amino acids and ligands in protein-ligand complexes. (a) A SHB between Asp32 and His64 in Bacillus lentus subtilisin (PDB ID 1GCI).33 (b) A SHB between Ser70 and Lys73 in CTX-M-14 class A β-lactamase (PDB ID 4UAA).85 (c) SHBs between the inhibitor and Asp32 and Asp215 in endothiapepsin (PDB ID 2JJJ).91 (d) SHBs between equilenin and Tyr14 and Asp99 in ketosteroid isomerase (PDB ID 1OH0).92 The red and blue dashed lines indicate SHBs and NHBs, respectively. ......................................................... 6

2.1 Properties of the potential energy surfaces predicted using the B3LYP-D3 method and the 6-31+G(d), 6-31+G(d,p) and aug-cc-pVDZ basis sets: (a) $\nu_{eq}$ and (b) $\Delta E_{\nu=0}$ as a function of R. .............................................. 10

2.2 (a) Percentage distribution of R in all the 15968 biological SHBs. (b) Distribution of 11814 enzyme-related SHBs in 7 enzyme classes. ...................................................... 13

2.3 Distribution of SHBs in (a) the backbone and side chains of proteins at different hydrogen bond lengths, and (b) in different secondary structures when the donor or acceptor groups are in the protein backbone. .............................................. 14

2.4 Chemical features of BB-SC and SC-SC SHBs. (a) Occurrence of 11 proteinogenic amino acids as acceptors or donors in SHBs. (b) Distribution of charged and neutral SHBs at different hydrogen bond lengths. ...................... 16
2.5 Three types of proton potential energy surfaces in biological SHBs. (a) A double-well potential, calculated from the Arg331–Glu328 hydrogen bond in a glucose isomerase (PDB ID 4A8I). (b) A single-well potential with a shoulder, calculated from the Asp35–Tyr109 hydrogen bond in a cellulbiohydrolase (PDB ID 2V3I). (c) A single-well potential, calculated from the Arg947–Glu972 hydrogen bond in a mineralocorticoid receptor (PDB ID 4PF3). The $\nu_{eq}$ and $\Delta E_{\nu=0}$ are highlighted for each system.  

2.6 Probability of different shapes of the potential energy surfaces from analyzing 3665 side chain-side chain short hydrogen bonds. The three types are single-well potential (SWP), single-well potential with a shoulder (SWP with a shoulder) and double-well potential (DWP).  

2.7 Conditional probabilities (a) $P_{cp}(R, \nu_{eq})$ and (b) $P_{cp}(R, \Delta E_{\nu=0})$ from electronic structure calculations of 3665 SC-SC SHBs. In each graph, the probabilities are normalized by their maximum value.  

2.8 (a) Average $\nu_{eq}$ for each R, calculated from DFT methods ($\nu_{QM}$) and the Amber14SB force field ($\nu_{cl}$) with the dotted lines as the linear fitting of the data. (b) Average $\Delta E_{\nu=0}$ for each R, calculated from DFT methods on the side chain-side chain hydrogen bonds with the dotted lines as the linear fitting of the data.  

2.9 Patterns of hybrid hydrogen bond networks. The top panels are schematic representations of the networks, in which nodes and lines represent atoms and hydrogen bonds, respectively. The bottom panels show example structures in proteins. The structural patterns include (a) the chain geometry of hydrogen bonded trimers (PDB ID 2BCH), the (b) chain and (c) branched geometries of tetramers (PDB IDs 2CI1 and 2EVW), and the (d) chain and (e) branched geometries of the pentamers (PDB IDs 5A0Y and 3RWN). Silver, red, blue and white represent C, O, N and H, respectively, and the hydrogen bonds are represented by dotted lines.  

2.10 (a) Occurrence of the protein backbone and side chains in the center or terminal of hybrid hydrogen bond networks. The amino acids are donors or acceptors in SHBs. (b) Correlation between $\Delta \Delta E_{\nu=0}$ and the proton positions in the reference state, $\nu_{eq}^{\text{single}}$. Insets shows the most probable configurations of the hydrogen bonded trimers in each quadrant.  

2.11 Example proton potential surfaces of SHBs when they are treated as single hydrogen bonds or in hydrogen bond networks. Potential energy surface for the hydrogen bonded trimer (a) Glu575-Lys366-Gln372 and (b) Glu128-Tyr195-His310.  

xiii
3.1 Workflow for the development of the MAPSHB model. In the first step, we randomly sample the NHBs in the training dataset and create 10 subsets that contain 6181 NHBs. In the second step, we form 10 balanced datasets by combining each subset of NHBs with all the SHBs. Next, we develop a boosting model from each of the 10 balanced training set and obtain the final MAPSHB model from their average.

3.2 Normalized importance scores for the 21 input features of the MAPSHB model. The features that contribute ≤1% to the predictions are grouped as the “other” type. These include the atom type, charge and secondary structure of the donor and acceptor groups and the location of the acceptor residue.

3.3 The normalized distribution of $P(R, \theta)$ for the (a) Tyr-Asp, (b) Ser-Asp, (c) Arg-Asp, (d) Tyr-Glu, (e) Ser-Glu and (f) Arg-Glu side chain-side chain hydrogen bonds.

3.4 Structures of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds that represent their most probable geometry arrangements in proteins. (a) A SHB between Tyr1147 and Asp1156 (R = 2.65 Å) in a CBP bromodomain (PDB ID 5I86). (b) A SHB between Ser261 and Asp11 (R = 2.65 Å) in an amidase mutant (PDB ID 4LF0). (c) Two regular hydrogen bonds between Arg141 and Asp135 (R = 2.79 Å and 2.85 Å) in the majastridin protein (PDB ID 2NXV).

3.5 Hydrogen bond energies for different configurations and solvation environments of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds when R is between 2.3 Å and 3.2 Å. For each pair, the red and blue lines represent the the PCM model with the $\varepsilon = 10.0$ and 78.4, respectively. The solid and the dashed lines represent the configurations from the first and the second maximum of the probability distributions, respectively.

3.6 Interaction energies of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds. For each type of the hydrogen bond, the red dots represent the interaction energies of 150 randomly chosen configurations. We further calculate how the hydrogen bond energies of the configurations from the first and second maxima of the geometric probability distributions vary with R, which are shown as the solid and dashed lines, respectively, in each plot. The vertical line represents the potential energy minimum of the Tyr-Asp pair at R = 2.55 Å.
3.7 Decomposition of the interaction energies for the Tyr-Asp, Ser-Asp and Arg-Asp hydrogen bonds at R = 2.55 Å. The total interaction energy \( \Delta E_{TOT}^{(s)} \) is partitioned into the frozen interaction \( \Delta E_{FRZ}^{(s)} \), polarization \( \Delta E_{POL}^{(s)} \) and charge transfer \( \Delta E_{CT}^{(s)} \) energies.  

4.1 (a) Percentage distribution of SHBs formed between two amino acids and between amino acids and ligands. (b) Distribution of the different categories of ligands in the 1393 amino acid-ligand SHBs. All the other ligands are grouped into the “Other” category. 

4.2 Examples of SHBs formed between amino acids and (a) NADP\(^+\) (PDB ID 5FI3)\(^{204}\) and (b) FAD (PDB ID 2VFR)\(^{207}\) in the active-site of proteins. Silver, red, blue, white and tan represent C, O, N, H and P, respectively. The SHBs are represented by dotted lines. 

4.3 Chemical features of amino acid-ligand SHBs. (a) Occurrence of the side chain of 11 proteinogenic amino acids and 11 most commonly observed functional groups in ligands in SHBs. (b) Distribution of charged and neutral amino acid-ligand SHBs at different hydrogen bond lengths. 

4.4 Examples of protein-ligand interactions in the active-sites of proteins: (a) SHBs of fucose to Asp96 and Asp99 in the calcium and monosaccharide binding loop (96-104) of lectin PA-III (PDB ID 1UZV),\(^{217}\) (b) SHBs of amprenavir to Asp25 in the catalytic triplet Asp25-Thr26-Gly27 of HIV-1 protease (PDB ID 3NU3),\(^{90}\) (c) SHBs of heme to Asp70 and Lys125 in the proximal side of nitrophorin 4 (PDB ID 1X8O).\(^{209}\) All figures are made using the LigPlot\(^+\) software.\(^{229,230}\) 

4.5 Normalized importance scores for the 16 input features of the machine learning model. The features that contribute \( \leq 2\% \) to the predictions are grouped as the “Other” type. These include the charges of amino acid and ligand, the atom type and secondary structure of the amino acid and the location, \( pK_a \), \( pK_b \) and logP of the ligand. 

4.6 (a) Hydrogen bond network composes of Asp96, Asp99 and Fuc in lectin PA-III (PDB ID 1UZV).\(^{217}\) (b) Interaction energies of the Fuc-Asp96 (red) and Fuc-Asp99 (blue) hydrogen bonds with (solid line) and without (dashed line) the SHB network.
5.1 Chemical structures of the model molecules that contain intramolecular SHBs. 66

5.2 Free energy profiles for DMANH and HM in organic solvents and aqueous solutions. In the top panels, the free energy surfaces are calculated from (a) AIMD and (b) AI-PIMD simulations of DMANH in acetonitrile, and from (c) AIMD and (d) AI-PIMD simulations of DMANH in water. In the bottom panels, the curves are calculated from (e) AIMD and (f) AI-PIMD simulations of HM in acetone, and from (g) AIMD and (h) AI-PIMD simulations of HM in water. 71

5.3 Representative configurations of HM with the hydrogen bonded water molecules. The conformations have R of 2.5 Å and ν of 0.2 Å. Silver, red and white represent C, O and H, respectively. The orange spheres are the ring polymer beads of the hydrogen bonded proton in HM. 74

5.4 Free energy profiles for DHND in water from (a) AIMD and (b) AI-PIMD simulations, and for CUA in DMSO from (c) AIMD and (d) AI-PIMD simulations. 75

5.5 (a) Correlation plot for δH and ν. Each data point is calculated from a solute-solvent cluster as sampled from the AIMD and AI-PIMD simulations of the model molecules in different solvents. (b) Correlation of δH and ν from the model molecules, KSI and previous computational studies. 79

5.6 Decomposition of the total shielding constants into the contributions from covalent bonding (Covalent) and hydrogen bonding (HB) for DMANH. 80

5.7 The natural bond orbitals of DMANH. (a) The σNH orbital and (b) the overlap of the np and σ∗NH orbitals at ν = −1.0 Å. (c) The np−s∗H orbital overlap diagram at ν = 0. Silver, blue and white represent C, N and H atoms, respectively. The red and green spheres are the positive and negative molecular orbitals calculated from the NBO analysis. 81

5.8 The probability distribution of ν from the (a) AIMD and (b) AI-PIMD simulations of the model molecules in different solvents (symbols) and the least squares fitting of them using Equation 5.4 (lines). The R2 values for the fitting range between 0.971 and 0.999 for AIMD simulations and between 0.986 and 0.998 for AI-PIMD simulations. 83

5.9 The np−σ∗ orbital overlap diagrams for CUA at (a) ν=−0.9 Å and (b) ν=1.0 Å. Silver, red and white represent C, O and H atoms, respectively. The red and green spheres are the positive and negative molecular orbitals. 86
5.10 Comparison between the predicted and experimental \(^{234,236,267}\) (a) chemical shifts and (b) proton positions of a series of inorganic and organic molecules with SHBs. The dashed diagonal lines represent perfect correlation between the variables. 87

6.1 Chemical structures of (a) TAPH and (b) DMANH with the hydrogen bonded N atoms highlighted in blue. Due to the symmetry of TAPH and DMANH, the N atoms in the hydrogen bonds are equivalent and they are numbered for discussion purposes. 91

6.2 Geometric parameters for the bifurcated SHB in TAPH. From AIMD and AI-PIMD simulations, \(N_1, N_2\) and \(N_3\) form an approximately equilateral triangle. 95

6.3 Representative structures of TAPH from (a) AIMD and (b) AI-PIMD simulations. Silver, blue and white represent C, N and H atoms, respectively. The orange spheres represent the ring polymer beads of the hydrogen bonded proton. 96

6.4 Instantaneous positions of the proton in the triangle formed by the N atoms (a and b) and the free energy surfaces for proton movement (c and d) in the bifurcated SHB in TAPH. (a) and (c) are from AIMD simulations, and (b) and (d) are from AI-PIMD simulations. The free energy profiles are obtained by subtracting their lowest values such that the minimal free energy in each plot is 0. 97

6.5 The average QKEs of the hydrogen bonded proton in TAPH and DMANH as obtained from the AI-PIMD simulations. For each system, the total QKE is decomposed into three internal coordinates: the vector along the N–H bond (N–H), the perpendicular vector in the hydrogen bond plane (Plane), and the vector orthogonal to the hydrogen bond plane (Orth). The internal coordinate system for TAPH is illustrated in the inset picture, and the orange spheres and springs represent the ring polymer of the hydrogen bonded proton. 100

6.6 Correlation between the averaged instantaneous \(^1\)H NMR chemical shift and the geometric parameters of the bifurcated SHB. The data from the AIMD and AI-PIMD simulations of TAPH are combined as they follow the same trend. 103

6.7 Chemical structures of (a) the protonated N-(4-methylphenyl)-2-pyridinamine and (b) the protonated \(N^2-(\text{pyridin-2-yl})-N^2,N^6\)-di-\(p\)-tolylpyridine-2,6-diamine. These are the monomers and dimers used in the decomposition of the magnetic shielding constants of the bifurcated SHB in TAPH. 105
6.8 The overlap of the \( n_p \) and \( s_H^* \) orbitals in the bifurcated SHB with (a) \( r_1 = 1.1 \text{ Å} \) and \( \theta_1 = 30.0^\circ \) and (b) \( r_1 = 1.5 \text{ Å} \) and \( \theta_1 = 30.0^\circ \). Silver, blue and white represent C, N and H atoms, respectively. The red and cyan spheres are the positive and negative molecular orbitals obtained from the NBO analysis.

7.1 Potential energy surfaces along the O-H stretch coordinate in an acetic acid dimer and an acid-base complex. In the top panels, silver, red, blue and white represent C, O, N and H atoms, respectively. The dotted line shows a conventional hydrogen bond, and the orange spheres represent the ring polymer beads of the delocalized proton. In the bottom panels, \(|0\rangle\) and \(|1\rangle\) represent the ground and first excited vibrational states of the system, respectively.

7.2 Spectroscopic and theoretical features of the complex formed from acetic acid and 1-methylimidazole. (a) ATR-FTIR spectra of the mixture of acetic acid and 1-methylimidazole for different molar ratios of acetic acid (\( X_{HA} \)). (b) Radial distribution functions between the oxygen atom of the hydroxyl group in acetic acid and the unsubstituted nitrogen atom of the imidazole ring (red highlighted atoms of the cartoon) from AIMD simulations at different concentrations. (c) Distribution of the rotation angles between the acid and the base from AIMD simulations. (d) IR spectrum in the 2000 – 3500 cm\(^{-1}\) region derived from AIMD simulations (\( X_{HA} = 0.67 \)) and the contributions of the different components: acetic acid (HAc) and 1-methylimidazole (MIm).

7.3 \textit{Ab initio} frequency calculations of acid-base representative complexes. (a) Representative structures of the complexes composed of acetic acid hydrogen bonded to 1-methylimidazole and their IR vibrational modes from DFT calculations. Left and right panels correspond to two different angular conformations of the complex as depicted on the top. (b) Atomic displacement for the proton shuttling mode of the complex. The left, center and right modes have an associated frequency of ca. 1585 cm\(^{-1}\), 1696 cm\(^{-1}\) and 2211 cm\(^{-1}\), respectively.

7.4 Energy profiles for shuttling the acidic hydrogen for different hydrogen bonded complexes. (a), (b) and (c) show the potential energy surfaces for the three hydrogen bonded pairs derived from \textit{ab initio} computations, and (d) and (e) display the free energy surfaces for the different pairs calculated from the AIMD and AI-PIMD simulations of the pure acid and the solution with \( X_{HA} \) of 0.67, respectively. The free energy surfaces are calculated along the proton sharing coordinate \( \nu \) and the hydrogen bond length \( R \), and are shifted so that their minimal values are 0.
7.5 Experimental and computational $^1$H NMR chemical shifts of the hydrogen bonded proton as a function of the concentration of the acid in the mixtures of acetic acid and 1-methylimidazole. The experimental $^2$H chemical shifts of the acetic acid-1d and 1-methylimidazole are also included for comparison.

8.1 IR spectra of (a) water (blue) and the potassium bifluoride solution (red).

8.2 Geometry of the F-H-F hydrogen bond in the 3.61 M potassium bifluoride solution from the AIMD simulations. (a) Free energy surface along R and $\nu$. (b) Normalized probability distribution of the proton XY-plane coordinates.

8.3 Radical distribution functions between the O atom in water and the F atoms in the fluoride and bifluoride ions in the aqueous solutions of different concentrations.

8.4 (a) Potential energy surface of proton shuttling in the bifluoride ion in the gas phase (red) and in the solute-solvent cluster (blue). (b) Autocorrelation function of the hydrogen bonds between water molecules and the fluoride and bifluoride ions with high and low concentrations.

8.5 (a) Geometric parameters to describe the water-ion interactions between the bifluoride ions and the water molecules. (b) Normalized probability distribution of r and $\theta$. (c) SDFs of the H atoms (white) and O atoms (red) in the water molecules surrounding the bifluoride ion. The isovalue is set to be 50 nm$^{-3}$.

8.6 (a) IR spectra of liquid water and the F$^−$ and HF$^2_2$ solutions in the 3000 to 3750 cm$^{-1}$ region derived from AIMD simulations. Contributions of water molecules in the first and second solvation shells of the ions in the 3.61 M (b) potassium fluoride and (c) potassium bifluoride solutions.
CHAPTER 1
INTRODUCTION AND BACKGROUND

The hydrogen bond has attracted wide interest since its introduction by Pauling in 1935. The hydrogen bonding interaction plays fundamental roles in condensed phase systems ranging from small molecules to biological polymers, and has been widely used in explaining the structure and dynamics of water clusters, liquid ammonia, hydrogen fluoride, DNA and proteins. While hydrogen bonds are conventionally described as classical dipole-dipole interactions, experimental measurements and quantum chemistry calculations have suggested their partially covalent nature. As such, the International Union of Pure and Applied Chemistry (IUPAC) has recommended the definition of a hydrogen bond as “an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation”. According to this definition, a series of different atoms or molecular fragments with high electronegativity, such as F, O, N, S and halides, are able to participate in the formation of hydrogen bonds. Here, given the frequency of occurrence, I will only focus on the common types of hydrogen bonds with O, N and F as the donor and acceptor heavy atoms.

The geometry, strength and proton behavior of a hydrogen bond D–H–·–A are strongly dependent on the distance between the donor and acceptor atoms, R. A typical hydrogen bond has R over 2.8 Å, and its potential energy surface of proton sharing contains two minimal energy configurations. As illustrated in Figure 1.1a and b, it has a double-well potential and the proton is covalently bonded to the donor atom. It takes a symmetric double-well potential when the \( pK_a \) values of the donor and acceptor atoms are identical (Figure 1.1a), and an asymmetric double-well potential otherwise (Figure 1.1b). When R is shortened to around 2.55 Å, the hydrogen bond becomes a low-barrier hydrogen bond.
(LBHB) and the energy barrier for proton sharing is comparable to the zero-point energy of a typical O-H or N-H stretch mode, which is about 5 kcal/mol (Figure 1.1c). R is considerably shorter than the sum of the van der Waals radius (1.52 Å for O, 1.55 Å for N and 1.47 Å for F)\(^{17}\) of the donor and acceptor atoms in LBHBs, which promotes quantum proton delocalization.\(^{14–16,18–22}\) When R is further shortened to around 2.3 Å, the energy barrier for proton sharing vanishes and a single-well potential appears (Figure 1.1d).\(^{18,19,23–25}\) As the potential energy surface has no barrier, the proton is delocalized across a wide region and is able to be equidistant or nearly equidistant between the donor and acceptor atoms. For example, single-well hydrogen bonds have been observed in potassium hydrogen malonate, KH(CF\(_3\)COO)\(_2\) and methylammonium hydrogen succinate monohydrate crystals.\(^{26–29}\) Given the similarity of the apparent elongation of the D-H bond and the covalent characters, we combine LBHBs and single-well potential hydrogen bonds as short hydrogen bonds (SHBs).\(^{30,31}\)

![Figure 1.1: Potential energy surface diagrams for transferring a proton in a hydrogen bond: (a) symmetric double-well potential, (b) asymmetric double-well potential, (c) low-barrier hydrogen bond potential, (d) single-well potential. The upper solid and the lower dashed red horizontal lines represent the zero-point energy levels for the donor-hydrogen or donor-deuterium bonds, respectively.](image)

In our definition, a SHB satisfies the following criteria: (1) the donor and acceptor atoms are O, N or F; (2) R ≤ 2.7 Å; (3) \(\angle\) DHA ≥ 135°. In comparison, we also define a normal hydrogen bond (NHB) as 2.8 Å ≤ R ≤ 3.2 Å. A value over 2.6 Å for R is longer
than usual distance (\( \sim 2.55 \) \( \text{\AA} \)) for LBHBs in both small molecules and the active sites of proteins. However, we use a relatively long cutoff of 2.7 \( \text{\AA} \) for R based on the following two reasons: (1) the coordinate uncertainty is about 0.1 \( \text{\AA} \) even in atomic-resolution (<1.2 \( \text{\AA} \)) crystal structures of biological macromolecules;\(^{32}\) (2) some LBHBs have R over 2.6 \( \text{\AA} \), such as the hydrogen bond between Asp32 and His64 (R = 2.64 \( \text{\AA} \)) in serine protease\(^{33}\) and that between imidazolium cation and carboxylate anion in the crystal phase (R = 2.65 \( \text{\AA} \)).\(^{22}\) Due to the close proximity between the donor and acceptor atoms, SHBs often exhibit prominent covalent characters and quantum mechanical features.\(^{19,31,34–37}\) For example, with the shortening of R, the proton tends to be more shared in the hydrogen bond, in which the donor-hydrogen bond can elongate from 0.95 \( \text{\AA} \) to 1.25 \( \text{\AA} \) and the acceptor-hydrogen distance would shorten from over 2.0 \( \text{\AA} \) to 1.25 \( \text{\AA} \). Notably, the proton can be located equidistantly between the donor and acceptor atoms when R is around 2.45 \( \text{\AA} \), for example, in the crystal structure of NH\(_4\)H(ClCH\(_2\)COO)\(_2\).\(^{31,38–40}\) Besides electrostatic interactions, SHBs exhibit prominent electronic and nuclear quantum effects. For example, electronic quantum effects such as charge transfer and dispersion play an important role in stabilizing these compact structures.\(^{12,31,41}\) As a result of the comparable proton sharing energy barriers and zero-point energies in SHBs, nuclear quantum effects can promote the delocalization of the proton.\(^{42–46}\) In addition, when hydrogen (H) is replaced by deuterium (D) in a SHB, the zero-point energy of the donor-deuterium bond is reduced, which prevents the proton from overcoming the barrier and leads to an unusual low H/D isotopic fractionation in protein systems.\(^{8,14–16,19,21,22,31,47}\)

Apart from the structural features, SHBs have distinctive spectral properties. They provide an indirect way to detect the presence of SHBs and a great opportunity to better understand their quantum nature. One of the spectral signatures of SHBs is their highly downfield \(^1\)H nuclear magnetic resonance (NMR) chemical shift, \( \delta_H \). \( \delta_H \) is below 10 parts per million (ppm) for a hydrogen atom in a typical functional group, while the proton delocalization in a SHB promotes the lengthening of the donor-hydrogen bond and shifts the
$\delta_H$ value to over 14 ppm.\textsuperscript{48-60} For example, serine proteases utilize a highly conserved Asp-His-Ser triad in the active site to catalyze the hydrolysis of peptide bonds\textsuperscript{61} and the length of the Asp-His hydrogen bond is often below 2.7 Å in the presence of an inhibitor.\textsuperscript{16,21,22,62,63} The $\delta_H$ values of this SHB vary from 18.6 to 19.0 ppm in chymotrypsin, a prototypical serine protease, depending on the structure of the inhibitor in the active site.\textsuperscript{64} Similarly, the Asp-His SHB shows a chemical shift of 17.4 ppm in subtilisin E and 19.9 ppm in a serine protease from the Dengue type II virus.\textsuperscript{63,65} Another spectral signature of SHBs is the red-shifted donor-hydrogen stretching frequency in the infrared (IR) spectra. With the elongation of the donor-hydrogen bond, the frequency of O-H stretching shifts from about $3500$ cm\(^{-1}\) to below $1000$ cm\(^{-1}\) in the IR spectra.\textsuperscript{66,67} For example, the distance between two O atoms in the potassium hydrogen maleate crystal is found to be 2.44 Å, indicating the formation of a SHB. The proton stretching vibrational frequencies for the O-H stretches are $711$ cm\(^{-1}\) and $563$ cm\(^{-1}\), respectively.\textsuperscript{67,68} Furthermore, two-dimensional infrared spectroscopy has been used to distinguish between SHBs and NHBs by measuring the ratio of the $|0\rangle \rightarrow |1\rangle$ transition frequency of proton shuttling, $\omega_{10}$, and the $|1\rangle \rightarrow |2\rangle$ transition frequency, $\omega_{21}$, $\gamma = \omega_{21}/\omega_{10}$.\textsuperscript{69} NHBs follow the conventional rule of positive anharmonicity and have $\gamma$ smaller than 1, while SHBs can show negative anharmonicity (superharmonicity) with $\gamma$ greater than 1. For example, the $\gamma$ value of the SHB in the aqueous solutions of potassium bifluoride is found to be 1.17.\textsuperscript{69}

SHBs have been widely observed in condensed phase systems. We will divide the systems containing SHBs into two categories: non-biological and biological systems. The non-biological systems include inorganic and organic small molecules in the solid or solution phases. For example, the bifluoride ion is the very first inorganic ion that is observed to contain a short, strong hydrogen bond.\textsuperscript{14,70} In the crystal structure of $p$-toluidine bifluoride, the intramolecular SHBs in the bifluoride ion have an R of 2.28 Å and the estimated bond strength is greater than 35 kcal/mol.\textsuperscript{14,70} In addition, bifluoride ions demonstrate different features when solvated in the aqueous solution, as the surrounding sol-
vation environment has a huge impact on the behavior of the SHB that is exposed to the water molecules. Compared to the solid state, the frequency of the F-H-F bending mode shifts from 1233 cm$^{-1}$ to 1206 cm$^{-1}$ and the frequency of the F-H-F asymmetric stretching mode has a blue shift from 1473 cm$^{-1}$ to 1536 cm$^{-1}$ in the aqueous solution of potassium bifluoride.$^{71}$ Another famous example is the SHB in the Proton Sponge (1,8-bis(dimethylamino)naphthalene).$^{72–74}$ Proton Sponge is an exceptionally strong base with $pK_a = 12.3$ and has a weak nucleophilic character due to its highly strained structure.$^{72,73}$ The protonated cation of Proton Sponge contains a stable intramolecular N-H-N hydrogen bond with $R$ of 2.55–2.63 Å, and shows a highly downfield $^1$H NMR chemical shift over 17 ppm.$^{75,76}$ Besides pure small molecules and ions, their mixtures such as protic ionic liquids (PILs) are also sources of SHBs. For example, the N–H···O hydrogen bond was found to contain short $R$ and linear bond angle in imidazolium based PILs.$^{77}$ One way to detect the presence of SHBs in PILs is to compare the IR spectra between the pure precursors and the PILs. For example, in N,N-dimethylethanolammonium acetate PILs, the stretching mode of the C=O bond (1710 cm$^{-1}$) in neutral acetic acid became weaker and the asymmetric stretching mode of the COO$^-$ (1570 cm$^{-1}$) in the acetate anion appear, assigned to the formation of hydrogen bonds between two components.$^{78–80}$

SHBs in the second category, biological systems, have been associated with essential functions in enzyme catalysis, including promoting protein structural stability of intermediate states, mediating antibiotic resistance of bacteria in acetyltransferases, and synchronizing long-range communication in the signalling cooperativity pathways of multimeric enzymes.$^{16,19,21,22,63,81–89}$ Among them, two types of SHBs are of particular interest due to their common appearance in biological macromolecules. In the first type, SHBs form between two proteinogenic amino acids in proteins. For example, a SHB is found between Asp32 and His64 in the catalytic triad Asp-His-Ser of the serine protease *Bacillus lentus* subtilisin, in which the proton is located 1.2 Å from the side chain O atom of Asp32 and 1.5 Å from the side chain N of His64 with an $R$ of 2.64 Å (Figure 1.2a). This SHB is pro-
posed to enhance the ability of His64 to deprotonate the catalytic Ser221. In another case, a SHB is found between Lys73 and Ser70 in the active site of the CTX-M-14 class A β-lactamase, in which R is 2.53 Å and the proton is shared equidistantly between the heteroatoms (Figure 1.2b). This SHB participates in a hydrogen bond network including Ser70, Lys73, Glu166, Asn170 and water molecules, and contributes to the stabilization of the transition state prior to general acid-base catalysis. In the second type, SHBs form between proteinogenic amino acids and ligands in protein-ligand complexes. For example, two SHBs are observed between a gem-diol inhibitor PD-135,040 and the catalytic residues Asp32 and Asp215 in an aspartic proteinase, endothiapepsin, for which the values of R are 2.54 Å and 2.64 Å, respectively (Figure 1.2c). The SHB with Asp32 is proposed to stabilize the transition state when a peptide binds to the active site of endothiapepsin. In another example, two SHBs form between equilenin and the catalytic Tyr14 and Asp99 residues in the bacterial Δ5-3-ketosteroid isomerase, for which the values of R are 2.54 Å and 2.48 Å, respectively (Figure 1.2d). From mutational analyses, the SHB associated with Tyr14 has been shown to significantly enhance the enzymatic catalysis rate and stabilize the transition state.

Figure 1.2: Examples of SHBs that form between amino acids in proteins and between amino acids and ligands in protein-ligand complexes. (a) A SHB between Asp32 and His64 in Bacillus lentus subtilisin (PDB ID 1GCI). (b) A SHB between Ser70 and Lys73 in CTX-M-14 class A β-lactamase (PDB ID 4UAA). (c) SHBs between the inhibitor and Asp32 and Asp215 in endothiapepsin (PDB ID 2JJJ). (d) SHBs between equilenin and Tyr14 and Asp99 in ketosteroid isomerase (PDB ID 1OH0). The red and blue dashed lines indicate SHBs and NHBs, respectively.
Obtaining the accurate geometry is always the first concern to investigate SHBs. With the improvement of experimental technologies such as X-ray crystallography, electron microscopy and neutron diffraction, more and more high-resolution structures of small molecules and biological macromolecules have been deposited in databases such as the Protein Data Bank (PDB) and the Cambridge Structural Database, which provides an opportunity to investigate the relation between R and proton positions in SHBs.\textsuperscript{93,94} High-resolution neutron diffraction has already enabled unambiguous determination of the proton positions of SHBs in the active sites of many biological systems, including photoactive yellow proteins, elastases and GCN5-related N-acetyltransferase in the presence of inhibitors.\textsuperscript{83,87,95,96} However, the number of atomic-resolution biomolecular structures is still limited by the long-standing challenge due to the experimental difficulty to map the electron density of hydrogen atoms using X-ray crystallography and the small number of high-flux neutron sources globally. Computational modelling provides a powerful alternative approach. Note that classical molecular dynamics (MD) simulations treat hydrogen bonding interaction mainly as electrostatic and van der Waals interactions based on the force fields, and cannot properly deal with SHBs due to their lack of quantum descriptions.\textsuperscript{30,97,98} This is especially problematic for hydrogen with its inherent quantum nature as the lightest element. A suitable method is to take electronic quantum effects into account and use first principles simulations. The \textit{ab initio} molecular dynamics (AIMD) simulations evolve the nuclear motion from on-the-fly electronic structure calculations and allows the chemical bonds to dynamically form and break as the conformations of the compounds fluctuate.\textsuperscript{12,99–104} In addition, \textit{ab initio} path integral molecular dynamics (AI-PIMD) simulations effectively incorporate the quantum nature of the nuclei using the path integral formalism of quantum mechanics, which exploits the isomorphism between the partition function of a quantum mechanical system and a classical system of ring polymers.\textsuperscript{105–107} As such, I will use these computational methods to study the structure, function and dynamics of SHBs in different condensed phase systems in this dissertation.
CHAPTER 2

STRUCTURAL AND CHEMICAL FEATURES OF BIOLOGICAL SHORT HYDROGEN BONDS IN PROTEINS

This chapter is reprinted with permission from Ref. [P1]

2.1 Introduction

SHBs have been widely observed in proteins,\textsuperscript{108–111} possibly because the three-dimensional folds of these biological macromolecules can help position the hydrogen bonded groups in close proximity. Despite the importance of biological SHBs, their structural features, energetics and the protein environment suitable for their formation are still under debate.\textsuperscript{112–115} The PDB contains over 184,000 biological macromolecular structures\textsuperscript{116,117} and offers a unique opportunity to dissect the features of SHBs. For example, previous analysis of the database has provided valuable insight into the geometries and locations of SHBs in proteins and on protein-ligand interfaces.\textsuperscript{108,109,111,118,119} In this chapter, we systematically examine the top 1% highest-quality structures in the PDB to unravel the structural and chemical factors that promote the formation of SHBs. For this purpose, we evaluate biomolecules that are refined to a resolution better than 1.1 Å from X-ray or neutron diffraction measurements, and reveal that SHBs and their networks are prevalent in proteins, protein-ligand complexes and nucleic acids. Combining statistical analysis and electronic structure calculations, we further uncover their preferred patterns in connectivity and amino acid composition and evaluate the impact of quantum effects on the proton behavior.
2.2 Methods

We conduct a search of the PDB for biomolecular structures that are determined by X-ray or neutron diffraction and have resolution of 1.1 Å or higher. This search yields 1663 structures, which include 103 nucleic acids and 1564 proteins. There are 4 protein-nucleic acid complexes among them, which we treat as both proteins and nucleic acids.

Except for the potential energy surfaces, all the calculations and analyses are performed using the Amber 2016 software package. The biomolecules and ligands are modeled using the Amber14SB force field and the generalized Amber force field, respectively. For each structure, we remove the crystallographic waters and add the H atoms using Amber 2016, and optimize the geometry with all non-hydrogen atoms maintained at their positions in the crystal structures. When both the hydrogen bond donor and acceptor atoms are in the backbone of a protein, we determine the corresponding secondary structures using the DSSP algorithm as implemented in Amber 2016.

We use electronic structure methods to obtain the optimized geometries and proton energy surfaces of the SHBs that formed from the side chains of Tyr, Lys, Arg, His, Asp and Glu. If the SHBs were involved in hydrogen bond networks, we further carry out electronic structure calculations in the presence of the networks. All calculations have been performed with the non-hydrogen atoms fixed at their positions in the crystal structures, using the TeraChem software package. The electronic structures are described with the B3LYP density functional, the D3 dispersion correction and the 6-31+G(d) basis set. To represent a side chain of an amino acid, we include all side chain atoms and the α-C atom, which is capped with hydrogens to saturate the bonds. In each SHB or hydrogen bond trimer, we compute the potential energy surface by scanning the acceptor–H or donor–H bond length and optimizing the position of all the protons at each step. This procedure is taken because the H atoms that are added using Amber 2016 might not be at their optimal positions in the electronic structure calculations. In addition, the protons can have con-
certed movements when the SHBs or their networks involve the side chains of Lys or Arg, which contain multiple N–H bonds. To assess the performance of the basis set, we repeat the calculations on 101 randomly chosen SHBs using the 6-31+G(d,p) and aug-cc-pVDZ basis sets and find that the equilibrium proton position and the barrier for proton sharing predicted from the three basis sets agreed well with each other, as shown in Figure 2.1. On average, the equilibrium proton position calculated from the 6-31+G(d,p) and aug-cc-pVDZ basis sets differs from that of the 6-31+G(d) basis set by 0.0039 and 0.0046 Å, respectively. Similarly, the average barrier differs from the value obtained from 6-31+G(d) by 0.54 and 0.45 kcal/mol, respectively. These results verifies that the 6-31+G(d) basis set is sufficient to capture the correct proton potential energy surfaces in the SHBs. We carry out all electronic structure calculations in the gas phase. To validate this approach, we consider 648 single SHBs and repeated the geometry optimization by representing the protein environment as point charges, as described using the Amber14SB force field. The resulting proton positions are in quantitative agreement with the gas-phase results with an average error of 0.03 Å.

Figure 2.1: Properties of the potential energy surfaces predicted using the B3LYP-D3 method and the 6-31+G(d), 6-31+G(d,p) and aug-cc-pVDZ basis sets: (a) $\nu_{eq}$ and (b) $\Delta E_{\nu=0}$ as a function of $R$. 
2.3 Results and discussion

2.3.1 Validation of the biomolecular structures

In this chapter, we consider the top 1% highest quality structures from the PDB that have resolution equal to or above 1.1 Å. Before conducting the analysis, we first examine their R-factor and R-free values to validate the 1663 crystal structures. 98.2% of the biomolecules have the R-factor $\leq 0.20$ and the difference between the R-free and R-factor $\leq 7\%$, demonstrating that they are reliable structures. The rest of them have slightly larger R-factor between 0.21 and 0.28.

As we define a SHB based on its heavy atom distance, the statistical analysis strongly depends on the accuracy of the atom position and R in the biomolecular structures. In our dataset, all biomolecules are at atomic resolution and the coordinate errors are expected to be around 0.03 Å. To verify this rule on our dataset, we find that 946 structures contain the estimated overall coordinate error calculated by the maximum likelihood method, $\Delta x$, in their PDB files. In each biomolecule, the $\Delta x$ value measures the coordinate error of all the non-hydrogen atoms and is expected to give an upper limit to the error in specific SHBs. In the 946 structures, $\Delta x$ values vary from 0.004 to 0.3 Å with an average of 0.04 Å, confirming the accuracy of the atom positions. The average $\Delta x$ gives rise to an error of $\sqrt{0.04^2 + 0.04^2} = 0.057$ Å in the heavy atom distance, R. Given that the coordinate error can extend beyond the average value, we find that 94% of the structures have $\Delta x \leq 0.1$ Å, which corresponds to an error up to 0.14 Å in R. Therefore, by focusing on biomolecular structure that are at atomic resolution, we can reliably analyze the SHBs as the errors in atomic position and R are relatively small.

2.3.2 Short hydrogen bonds in biological macromolecules

After examining 1663 high-quality crystal structures, we have found that 1504 biomolecules contain at least one SHB. These include 1475 proteins and protein-ligand complexes as
well as 30 nucleic acids, among which there is 1 protein-nucleic acid complex. We have identified a total of 15968 SHBs, which gives an average of 11 SHBs in each structure. Moreover, when considering both short and regular hydrogen bonds in the 1663 structures, we find a total of 258753 cases with $2.3 \leq R \leq 3.0 \text{ Å}$. This suggests that one can observe 1 SHB in every 16 hydrogen bonds, highlighting the prevalence of these special structural elements in biological macromolecules.

A small amount of 57 SHBs are present in nucleic acids, which form in Watson-Crick base pairs, guanine-uracil wobble base pairs and between the backbone ribose and phosphate groups of adjacent nucleotides. From our analysis, 99.6% of the observed SHBs are present in protein and protein-ligand complexes, with the number varying from 1 to 215 in each structure. Given the rarity of SHBs in nucleic acids, we will focus on these systems and characterize SHBs and hydrogen bond networks that form between amino acids, and show how the interplay of their geometric and chemical features determines the proton potential surfaces in this chapter. We will then identify the types of amino acids and ligands that commonly participate in the formation of SHBs in protein-ligand complexes in Chapter 4. As shown in Figure 2.2a, 50.6% of these biological SHBs have $R$ between 2.65 and 2.7 Å. However, there are 3314 very short hydrogen bonds with $R < 2.6 \text{ Å}$. Considering that the van der Waals radii for the N and O atoms are 1.55 and 1.52 Å, respectively, these SHBs are conformationally highly compact with the donor and acceptor groups in much closer proximity than those typically observed in the condensed phase. Chemically, 98.8% of the SHBs have O as the acceptor atom, and $\text{O} \cdots \text{O}$ is the most commonly observed. This is followed by $\text{N} \cdots \text{O}$ hydrogen bonds, which are more likely to occur when $R$ is shorter than 2.55 Å.

Given the observation that SHBs are extensively distributed in biological systems, they might play a role in enhancing the functions of proteins and nucleic acids. While it is not the main focus of this chapter, we will use two categories of proteins to demonstrate the possible functional importance of SHBs. In the first category, we have identified 226 SHBs
from the analysis of 37 proteins that are crucial for cellular signal transduction. These include Ras the RAF proteins, which are pivotal components in the Ras-RAF-MARK pathway to mediate mammalian gene expression, and response regulatory proteins for bacterial photo- and chemotaxis. As an example, the light-sensing chromophore in photoactive yellow protein, a photoreceptor that controls the negative phototaxis of purple sulfur bacteria, forms a network of SHBs with residues Tyr42 and Glu46 with R of 2.49 and 2.58 Å, respectively. The SHB network is proposed to stabilize the deprotonated chromophore in the hydrophobic protein interior and maintain the ground receptor state of the protein in its signal transduction pathway. In the second category, we have found a total of 11814 SHBs in 900 enzymes. As shown in Figure 2.2b, SHBs exist in all 7 classes of enzymes, which include 484 hydrolases, 208 oxidoreductases, 86 lyases, 59 transferases, 57 isomerases, 5 ligases and 1 translocase. SHBs are most abundant in hydrolases, followed by oxidoreductases, lyases and transferases, in accordance with the fractions of these enzymes in our dataset. On average, we find that each hydrolase and lyase contain 12 SHBs, whereas each oxidoreductase, transferase and ligase contain 17 SHBs. In addition, we find an average of 8 SHBs in each isomerase, and there are 5 SHBs formed in the...
only translocase structure. For example, as one of the largest groups of hydrolases, serine proteases utilize a highly conserved Asp–His–Ser catalytic triad to facilitate the hydrolytic cleavage of peptide bonds.\cite{61,142,143} From the statistical analysis, we have identified SHBs in serine proteases ranging from trypsin to proteinase K and elastase,\cite{144-146} and these SHBs in the catalytic triad have been proposed to aid the initiation of the enzymatic reactions and stabilize the reaction intermediates.\cite{16,21,63}

### 2.3.3 Structural and chemical features of short hydrogen bonds in proteins

A total of 13724 SHBs occur between amino acids in proteins. As shown in Figure 2.3a, 5281 SHBs are backbone-backbone (BB-BB) and backbone-side chain (BB-SC) hydrogen bonds. 82.2% of these backbone-involving SHBs have the peptide bond C=O as acceptor and the side chain O–H or N–H groups as donors, and they are the predominant types across all hydrogen bond lengths. The rest have the main chain N–H groups as donors and the backbone or side chains as acceptors, which are more frequently observed when R is around 2.7 Å.

![Figure 2.3: Distribution of SHBs in (a) the backbone and side chains of proteins at different hydrogen bond lengths, and (b) in different secondary structures when the donor or acceptor groups are in the protein backbone.](image)

\[\text{(Figure 2.3)}\] Distribution of SHBs in (a) the backbone and side chains of proteins at different hydrogen bond lengths, and (b) in different secondary structures when the donor or acceptor groups are in the protein backbone.
From Figure 2.3a, 90.5% of the acceptors in the BB-BB and BB-SC hydrogen bonds are the amide bond C=O groups, consistent with the finding that O is the most common acceptor in biological SHBs. As shown in Figure 2.3b, these backbone acceptors are distributed among all types of secondary structures. 40.1% of them are in ordered protein configurations, including α- and 310-helices and β-sheets. In BB-BB hydrogen bonds, this ratio increases to 63.9%, indicating that regular protein structural patterns can facilitate the formation of SHBs. In contrast, in BB-SC hydrogen bonds, the majority of the backbone carbonyl acceptors reside in more disordered regions of the proteins such as coils, bends and turns, in agreement with a previous study of the PDB. Similarly, when the backbone N–H groups serve as donors in the SHBs, their preferred locations are in disordered secondary structure motifs. Therefore, Figure 2.3b suggests that proteins can not only use regular secondary structures to position backbone amide groups in close proximity, but also take advantage of flexible structural elements to bring the backbone and side chain groups together and facilitate the formation of SHBs.

In Figure 2.3a, the side chains of amino acids are present in 13284 SHBs, and they account for over 95% of SHBs at each R. Among them, there are 4841 BB-SC SHBs and 8443 side chain-side chain (SC-SC) SHBs. To elucidate their chemical features, we have examined the occurrence of 11 proteinogenic amino acids with polar side chains that are capable of forming hydrogen bonds. These amino acids include Ser, Thr and Tyr with side chain –OH groups, Asp and Glu with –COO\(^-\) groups, Asn and Gln with –CONH\(_2\) groups, Lys with the –NH\(_2\) group, Trp with the indole group, Arg with the guanidinium group, and His with the imidazole group. Figure 2.4a shows that except Trp, all other 10 amino acids are frequently involved in the formation of SHBs. In all BB-SC and SC-SC hydrogen bonds, 80.0% have the negatively charged Asp and Glu as acceptor residues while 9.5% have the neutral Asn and Gln as acceptors. In contrast, the donor residues in these SHBs are predominantly amino acids with neutral side chains. For example, Ser and Thr have aliphatic side chains with hydroxyl groups and serve as donors in 52.8% of
SHBs. Tyr contains the aromatic phenol side chain and acts as donors in 26.9% of SHBs. The remaining 20.3% of SHBs mainly have positively charged Lys, His and Arg as donor groups. From Figure 2.4a, the most favorable acceptor and donor residues in the BB-SC and SC-SC hydrogen bonds contain carboxyl and hydroxyl groups, respectively, which contribute to the observation that O–H⋯O is the most common type of biological SHBs. In addition, many N–H⋯O hydrogen bonds form when the side chains of Lys, His and Arg are the donor groups. Here, the observations that amino acid side chains are present in the majority of SHBs and that the charged Lys, His, Arg, Asp and Glu as well as the neutral Tyr, Ser and Thr are enriched in SHBs are consistent with a recent study by Qi and Kulik on close contacts in the crystal structures of proteins.\(^{119}\)

Figure 2.4: Chemical features of BB-SC and SC-SC SHBs. (a) Occurrence of 11 proteinogenic amino acids as acceptors or donors in SHBs. (b) Distribution of charged and neutral SHBs at different hydrogen bond lengths.

Figure 2.4a indicates that the charge and aromaticity of the amino acids are important chemical factors in the formation of BB-SC and SC-SC SHBs. To further elucidate the role of side chain charges, we have computed the distribution of “charged” and “neutral” SHBs at different hydrogen bond lengths. While the residues involved in SHBs might have considerably disturbed acidity, it is computationally demanding to accurately calculate their \(pK_a\) in the protein interior. Therefore, we use the solution \(pK_a\) value as a reference to
determine the ionization states of the amino acid side chains. A SHB is defined as charged if at least one hydrogen bond participant bears a charge, and as neutral if both the donor and acceptor groups are neutral. As shown in Figure 2.4b, both types of SHBs are abundant at all hydrogen bond lengths. The majority (71.7%) of neutral SHBs are BB-SC hydrogen bonds in which the peptide bond C=O groups are acceptors. In contrast, 89.2% of charged SHBs are SC-SC hydrogen bonds. Consistent with the findings in Figure 2.4a, the most favorable acceptor residues in the charged SHBs are Asp and Glu, whereas the most common donors are the neutral Tyr, Ser and Thr as well as the positively charged Arg, Lys and His. As there are almost twice as many SC-SC hydrogen bonds as BB-SC hydrogen bonds, it is more likely to find charged SHBs when R is between 2.35 and 2.65 Å. Accordingly, Figure 2.4b demonstrates that the possession of charges in the donor or acceptor groups facilitates the formation of SC-SC SHBs. From recent symmetry-adapted perturbation theory calculations by Qi and Kulik, this phenomenon arises because the electrostatic and induction interactions are significantly enhanced when a charged residue is present, providing stabilization to the SHBs.\(^{119}\)

### 2.3.4 Proton potential energy surfaces for side chain-side chain short hydrogen bonds

Shortening R in a hydrogen bond often results in a larger degree of proton sharing between the donor and acceptor groups.\(^{19,31,41,44,57}\) As such, compared to hydrogen bonds that are typically observed in the condensed phase, SHBs can have distinct electronic energy surfaces when the proton is moved between the donor and acceptor atoms. To uncover how the structural and chemical features impact the SHBs in proteins, we have used electronic structure methods to compute the proton potential energy curves for 3665 SC-SC hydrogen bonds that are composed of Tyr, Lys, Arg, His, Asp and Glu. Here we only consider SC-SC SHBs that contain specific amino acids because the backbone amide groups and the side chains of Trp, Ser, Thr, Asn and Gln are protonated under neutral pH conditions, and hence are energetically unfavorable to participate in the sharing or transferring of protons.
Figure 2.5: Three types of proton potential energy surfaces in biological SHBs. (a) A double-well potential, calculated from the Arg331–Glu328 hydrogen bond in a glucose isomerase (PDB ID 4A8I). (b) A single-well potential with a shoulder, calculated from the Asp35–Tyr109 hydrogen bond in a cellobiohydrolase (PDB ID 2V3I). (c) A single-well potential, calculated from the Arg947–Glu972 hydrogen bond in a mineralocorticoid receptor (PDB ID 4PF3). \( \nu_{eq} \) and \( \Delta E_{v=0} \) are highlighted for each system.

To characterize a SC-SC hydrogen bond A–H···B, we have determined the donor and acceptor atoms from its optimized geometry and defined the proton sharing coordinate as

\[
\nu = d_{AH} - d_{BH},
\]

where \( d_{AH} \) and \( d_{BH} \) are the distance from the H atom to the donor and acceptor, respectively. From this definition, the equilibrium proton positions, \( \nu_{eq} \), in all of the 3665 SHBs are negative. As shown in Figure 2.5 and Figure 2.6, the proton potential energy curves fall into 3 categories, and their fractions depend heavily on R. For relatively long hydrogen bonds with \( R > 2.55 \) Å, the potential energy surface can take the form of a symmetric or asymmetric double well curve (Figure 2.5a). In addition to the negative \( \nu_{eq} \), they have a second minimum at \( \nu > 0 \), suggesting that the proton can form a stable B–H bond after being transferred to the acceptor group. However, these SHBs are more likely to adopt a single-well potential curve with a small shoulder (Figure 2.5b). Here the proton transferred configuration is not thermodynamically stable, as evident from the presence of a shoulder rather than a second minimum at \( \nu > 0 \). When \( R < 2.55 \) Å, over 70% of the SHBs have a single-well potential energy surface, and this ratio increases to 100% when R
becomes shorter than 2.4 Å. As shown in Figure 2.5c, \( \nu_{eq} \) in the single-well potentials are closer to 0 than those in other types of surfaces, indicating that protons are more shared in the hydrogen bonds as their lengths shorten. Figure 2.5 hence demonstrates the well-known phenomenon that as \( R \) of hydrogen bonds shortens, the proton energy surfaces change from double-well to single-well potentials,\(^ {16,19,31,41,64,148} \) and it has been extensively shown that these differences in the shape of the potential energy curves lead to unique residual entropy and spectroscopic properties in small molecule crystals such as ice and bifluoride ions.\(^ {1,31,41,149,150} \)

![Figure 2.6: Probability of different shapes of the potential energy surfaces from analyzing 3665 side chain-side chain short hydrogen bonds. The three types are single-well potential (SWP), single-well potential with a shoulder (SWP with a shoulder) and double-well potential (DWP).](image)

The compact structures of SHBs strongly impact the extent to which quantum effects modulate the potential energy surfaces and the proton behavior. From the electronic structure calculations, we have examined the optimized geometry of the SHBs and calculated the conditional probability of finding a hydrogen bond with length \( R \) and the proton at \( \nu_{eq} \),

\[
P_{cp}(R, \nu_{eq}) = \frac{P(R, \nu_{eq})}{P(R)},
\]

where \( P(\alpha) \) represents the probability distribution of the property \( \alpha \). As shown in Figure 2.7a, while the 3665 SC-SC hydrogen bonds have differ-
ent donor and acceptor residues, their equilibrium proton positions follow the same trend with the change in R. At R of 2.7 Å, $\nu_{eq}$ distributes between -0.4 and -0.9 Å with an average value of -0.7 Å. As R shortens, the average $\nu_{eq}$ increases almost linearly with a slope of -1.2 (Figure 2.8a). When R < 2.4 Å, the average $\nu_{eq}$ becomes larger than -0.3 Å and noticeable amount of the SHBs has $\nu_{eq}$ close to 0, where the proton resides equidistantly between the donor and acceptor atoms. We observe the strong correlation between $\nu_{eq}$ and R, demonstrating that the classical force field is capable of providing a qualitatively correct description of the proton behavior in SHBs. However, the interplay of R and electronic quantum effects results in two distinct features. First, the explicit inclusion of the quantum nature of the electrons promotes proton sharing in the SHBs, because the average $\nu_{eq}$ is larger at any given R and moves more rapidly towards 0 as R shortens compared to the classical results (Figure 2.8a). Second, electronic quantum effects significantly increase the fluctuations of $\nu_{eq}$ around their average values, hence capturing the sensitivity of the proton positions to the surrounding chemical environment.

Figure 2.7: Conditional probabilities (a) $P_{cp}(R, \nu_{eq})$ and (b) $P_{cp}(R, \Delta E_{\nu=0})$ from electronic structure calculations of 3665 SC-SC SHBs. In each graph, the probabilities are normalized by their maximum value.

To further delineate the potential energy surfaces, we define the barrier for proton sharing in the SHBs as the energy required to move the proton from its equilibrium state to the
equally shared position, $\Delta E_{\nu=0}$, as illustrated in Figure 2.5. Similar to the case of $\nu_{eq}$, we have examined the 3665 SC-SC hydrogen bonds and computed the conditional probability $P_{cp}(R, \Delta E_{\nu=0}) = P(R, \Delta E_{\nu=0})/P(R)$. As shown in Figure 2.7b, $\Delta E_{\nu=0}$ of the SHBs exhibit a strong positive correlation with $R$. When $R$ is at 2.7 Å, $\Delta E_{\nu=0}$ of the SHBs can go up to 34.6 kcal/mol and have a large average value of 10.3 kcal/mol (Figure 2.8b). Due to the high barrier, the protons in these relatively long hydrogen bonds are covalently linked to the donor atoms with highly negative $\nu_{eq}$ values, as observed from Figure 2.7a. When $2.4 \ A \leq R \leq 2.6 \ A$, the average barrier decreases to 2.6 – 6.7 kcal/mol, which makes the proton more shared in the SHBs with the average $\nu_{eq}$ between -0.3 and -0.6 Å. These SHBs are also in the low-barrier hydrogen bond regime, where $\Delta E_{\nu=0}$ is comparable to the zero-point energy of the O–H or N–H vibration ($\sim$ 5 kcal/mol). The zero-point energy hence promotes the quantum delocalization of the proton in the SHBs, as demonstrated in previous simulation studies of a hydrogen bond network in the active site of an enzyme.\textsuperscript{30,151}

When $R$ further shortens to below 2.4 Å, the potential energy curves becomes a single-well potential (Figure 2.5c) with the average $\Delta E_{\nu=0}$ smaller than 3 kcal/mol. Accordingly, both
electronic and nuclear quantum effects will facilitate the sharing of protons in these very short hydrogen bonds. Note that while nuclear quantum effects allow the proton to be delocalized between the donor and acceptor groups and strengthen a SHB, they also enhance the motion of the proton in other directions that act to distort and weaken the hydrogen bond. Therefore, the net impact results from a delicate balance between two competing effects, with their relative importance depending strongly on \( R \). From a series of recent simulations on hydrogen bonded systems, nuclear quantum effects strengthen shorter hydrogen bonds and weaken longer ones.\(^{37,151–156}\)

### 2.3.5 Hybrid hydrogen bond networks in proteins

Properties of a SHB can be significantly changed when it is involved in a hydrogen bond network. From all the proteins, we have identified a total of 4967 networks that contain at least 1 SHB. We refer to these structures as hybrid hydrogen bond networks because 96.2\% of them are formed from both SHBs and regular hydrogen bonds. As schematically represented in the top panels of Figure 2.9, the hybrid networks exhibit 5 characteristic connectivity patterns. 76.4\% of them are hydrogen bonded trimers, which take a V-shaped chain geometry (Figure 2.9a). The second largest population have 4 hydrogen bond participants, among which 792 systems adopt a chain structure that provides 2 hydrogen bonds to each of the central residues and 1 hydrogen bond for the terminal groups (Figure 2.9b). 95 of these tetramers take a branched geometry, in which the central residue forms 3 hydrogen bonds with the surrounding terminal groups, as shown in Figure 2.9c. In addition, 168 networks are hydrogen bonded pentamers that are in either a chain or branched configuration, as demonstrated in Figure 2.9d and Figure 2.9e, respectively.

The protein backbone amide groups and the polar side chains, except that in tryptophan, have the capacity to form multiple hydrogen bonds. From Figure 2.9, the two amino acids in a SHB can reside in the center or terminal of a hybrid hydrogen bond network. We hence examine their preferred locations in hybrid networks and plot the distributions
Figure 2.9: Patterns of hybrid hydrogen bond networks. The top panels are schematic representations of the networks, in which nodes and lines represent atoms and hydrogen bonds, respectively. The bottom panels show example structures in proteins. The structural patterns include (a) the chain geometry of hydrogen bonded trimers (PDB ID 2BCH), the (b) chain and (c) branched geometries of tetramers (PDB IDs 2CI1 and 2EVW), and the (d) chain and (e) branched geometries of the pentamers (PDB IDs 5A0Y and 3RWN). Silver, red, blue and white represent C, O, N and H, respectively, and the hydrogen bonds are represented by dotted lines.

in Figure 2.10a. 44.3% of hybrid networks have negatively charged Asp and Glu as central residues, possibly because multiple hydrogen bonds can act to stabilize the negatively charged carboxylate groups in the protein interior. The neutral side chains in Ser, Thr and Tyr are commonly observed both in the center and terminal of hybrid networks, demonstrating that the –OH functional group is highly favored in the hybrid networks. Furthermore, the protein backbone amide groups frequently occur in the centers of hybrid networks and are the most favored terminal residues, highlighting their prevalence in hydrogen bond networks that involve SHBs.
Figure 2.10: (a) Occurrence of the protein backbone and side chains in the center or terminal of hybrid hydrogen bond networks. The amino acids are donors or acceptors in SHBs. (b) Correlation between $\Delta\Delta E_{\nu=0}$ and the proton positions in the reference state, $\nu_{eq}^{single}$. Insets shows the most probable configurations of the hydrogen bonded trimers in each quadrant.

Next, we investigate how the presence of a hydrogen bond network alters the proton energy surface of a SHB. Here we only consider hydrogen bonded trimers because the hybrid networks predominantly take a trimer structure and that the most prominent influence on a SHB comes from its closest hydrogen bond partner. To directly compare the properties of SHBs in the absence and presence of the network, we have carried out electronic structure calculations on 947 trimers in which the SHBs are formed from the side chains of Tyr, Lys, Arg, His, Asp and Glu. Their structures are schematically presented in the insets of Figure 2.10b: the terminal residue $T_1$ forms a SHB with the central residue $C$, which is further linked to another terminal residue $T_2$ to form a hydrogen bond network. In the reference state, the pair of $T_1$ and $C$ is treated as an isolated single SHB and its proton energy curve is characterized by the equilibrium proton position, $\nu_{eq}^{single} = d_{T_1H} - d_{CH}$, and the barrier for proton sharing, $\Delta E_{\nu=0}^{single}$. When the SHB is involved in a network, its barrier becomes $\Delta E_{\nu=0}^{network}$. As shown in Figure 2.10b, the impact of the hydrogen bond network on the barrier for proton sharing, $\Delta\Delta E_{\nu=0} = \Delta E_{\nu=0}^{network} - \Delta E_{\nu=0}^{single}$, depends heavily on $\nu_{eq}^{single}$ in the reference state.
In the reference state that residues T_1 and C forms a single SHB, 77.8% of the systems have the protons reside closer to T_1 and ν_{eq}^{\text{single}} < 0 and hence belong to Quadrants I and II in Figure 2.10b. In the presence of residue T_2, 650 of them have increased barrier (Quadrant I). In these cases, residue C are almost exclusively Asp or Glu that accept hydrogen bonds from both T_1 and T_2, as shown in the inset picture. Because of this connectivity, the electronic induction effects from T_2 result in a slight decrease in ν_{eq} in the SHBs and an increase in their barriers (ΔΔE_{ν=0} > 0) compared to the reference state. In contrast, 87 SHBs are in Quadrant II and have reduced barriers upon forming the hybrid networks. Over 50% of these systems have ΔΔE_{ν=0} < -1 kcal/mol and lysine as the central residue, which accepts a hydrogen bond from T_1 and donates a hydrogen bond to T_2. As such, T_2 electronically induces the proton to be more shared in the SHBs and lowers the barrier for proton sharing. The reduced barriers lead to proton transfer from residues T_1 to C in a few systems. As an example, the proton potential energy surfaces of a Glu–Lys SHB are shown in Figure 2.11a. The hydrogen bonded trimer Glu575-Lys366-Gln372 contains a SHB formed from Glu575 and Lys366 with R of 2.69 Å in a pyruvate oxidase (PDB ID 4FEG).\textsuperscript{161} When the side chain of a Gln residue is hydrogen bonded to Lys, a proton transfer occurs and the shape of the energy curves qualitatively changes as the barrier decreases by 3.7 kcal/mol and ν_{eq} shifts from -0.6 to 0.5 Å.

In the reference state, a total of 210 SHBs have residue C as the hydrogen bond donor and ν_{eq}^{\text{single}} > 0. When involved in hydrogen bond networks, the majority of them have decreased barriers and are in Quadrant IV of Figure 2.10b. In these systems, the most common central residue is Lys, which is followed by Asp and Glu. As illustrated in the inset picture, residue C donates a hydrogen bond to T_1 and accepts one from T_2. From this connectivity, the presence of T_2 stabilizes residue C, facilitates the sharing of the proton in the SHB and reduces the potential energy barrier. For example, we have observed 3 cases where ΔΔE_{ν=0} < -17 kcal/mol, all of which have a Tyr–Tyr SHB connected to a Glu residue as T_2. Due to the barrier reduction, proton transfer occurs in 32% of the
Figure 2.11: Example proton potential surfaces of SHBs when they are treated as single hydrogen bonds or in hydrogen bond networks. Potential energy surface for the hydrogen bonded trimer (a) Glu575-Lys366-Gln372 and (b) Glu128-Tyr195-His310.

SHBs in Quadrant IV, particularly when T_2 are the side chains of Arg, Lys or His as their positive charges provide stronger induction effects. This is demonstrated in Figure 2.11b using a Glu–Tyr SHB. The hydrogen bonded trimer Glu128-Tyr195-His310 contains a SHB formed from Glu128 and Tyr195 with R of 2.58 Å in a methionine aminopeptidase (PDB ID 2B3H). In the presence of a third His residue, the barrier for proton sharing decreases by 5.8 kcal/mol, leading to a proton transfer and a shift in ν_{eq} from 0.6 to -0.5 Å. Finally, a small number of 34 SHBs are in Quadrant III, which have increased barrier when hydrogen bond networks are formed. When ΔΔE_{ν=0} > 2 kcal/mol, Arg is the predominant residue C as it contains more than one hydrogen atoms in the side chain and can serve as dual donors in the hydrogen bond network. In these cases, residue T_2 are Asp or Glu and their strong electrostatic interactions with residue C increase the barrier for proton sharing in the SHBs (ΔΔE_{ν=0} > 0). Therefore, Figure 2.10b demonstrates that the potential energy curves, and hence the proton behavior in the SHBs are significantly influenced by the geometries and chemical features of the hydrogen bond networks.
2.4 Conclusion

In this chapter, we statistically analyze the PDB and find that on average, each of the 1504 high-resolution biomolecular structures contains 11 SHBs. This observation demonstrates that SHBs are ubiquitous in proteins, protein-ligand complexes and nucleic acids, and indicates the importance to incorporate these special structural elements in X-ray or NMR structure refinement as conventional methods tend to avoid the formation of very close contacts between atoms. Structurally, these SHBs all have $R \leq 2.7$ Å and are frequently involved in the formation of hydrogen bond networks. Chemically, they often contain the charged side chains of Asp, Glu, Arg, Lys and His as well as the neutral side chains of Ser, Thr and Tyr. SHBs can also be functionally important as they are widely distributed in signaling proteins and enzymes, which shows their biological importance.

The interplay of the structural and chemical features results in characteristic proton potential energy surfaces that are universal for all biological SHBs. In particular, as $R$ shortens, the potential energy barrier decreases and the proton is more shared in the hydrogen bond, and the influence of quantum effects becomes prominent. For example, our calculations have shown that the classical Amber14SB force field can only provide a qualitative description of this relation and explicit inclusion of electronic quantum effects is required to accurately predict the equilibrium proton positions and the barrier for proton sharing in the SHBs. Note that we have carried out all calculations with the non-hydrogen atoms fixed at their positions in the crystal structures, and one can further investigate the impact of conformational fluctuations using molecular simulations that obtain forces from instantaneous quantum mechanical calculations.\textsuperscript{163–167} Moreover, our results confirm that when $R$ is between 2.4 and 2.6 Å, one enters the low-barrier hydrogen bond regime as the barrier for sharing the proton between the donor and acceptor groups is comparable to the zero-point energies of typical O–H and N–H vibrations. To elucidate how quantum effects facilitate the sharing and transferring of the protons in these SHBs and unravel their
functional importance, one can exploit simulations that incorporate the quantum mechanical nature of both electrons and nuclei, which have offered crucial insight into hydrogen bonded systems in proteins.\textsuperscript{30,45,86,156,168–171} These simulations will also provide benchmark data for the development of new force fields that accurately and efficiently describe the conformations and proton sharing conditions in biological SHBs.
CHAPTER 3
EFFECTIVE PREDICTION OF SHORT HYDROGEN BONDS IN PROTEINS
VIA MACHINE LEARNING METHOD

This chapter is reprinted with permission from Ref. [P4]

3.1 Introduction

Despite the importance of SHBs, pinpointing them often requires a protein structure at atomic resolution (< 1.2 Å), which poses a challenge to structure determination techniques. For example, the coordinate error in a well-ordered protein structure with a 2 Å resolution is about 0.2 Å, which gives a significant uncertainty of 0.3 Å in R. While force field based minimization and computational structure predictions have greatly improved the crystallographic refinement process, the relevant classical force fields cannot fully account for the quantum nature of the SHBs and their van der Waals parameters would push the length of a hydrogen bond to ∼3 Å. Due to the coordinate uncertainty in the crystal structures and the inaccuracy of classical force fields, it is often difficult to determine the presence of SHBs in a biological macromolecule. As such, it is highly desirable to develop a method that allows for an effective prediction of SHBs based on a protein structure with moderate or low resolution.

In Chapter 2, we have analyzed the top 1% highest-quality macromolecular structures in the PDB and characterized the preferred location, geometry and amino acid composition of SHBs. The interplay of the geometric and chemical features helps position the hydrogen bonded atoms in close proximity in the interior or on the surface of proteins, making it more likely to observe SHBs in these biological molecules than in small molecules and simple liquids. Combining these features and the sequence information, in this chapter, we consider 1260 high-resolution peptide and protein structures from the PDB and develop
a machine learning model that effectively predicts the formation of SHBs between amino acids. This model is based on the boosting method and utilizes the undersampling strategy. We then reveal the key factors that facilitate the formation of SHBs and unravel why the phenol side chain of Tyr and the carboxylate side chains of Asp and Glu have a significant probability of forming such close contacts. This work is in collaboration with Professor Sijian Wang and Yuanhao Liu in the Department of Statistics of Rutgers University, and they are in charge of the machine learning model development.

3.2 Methods

3.2.1 Preparation of the hydrogen bond dataset

The structural preparation is carried out using the same method discussed in Chapter 2. Then we obtain a series of structural, chemical and sequence features for each hydrogen bond, including the residue and atom types, charge and location of the hydrogen bond donor and acceptor atoms and the relevant secondary structures and sequence information using AmberTools16.175 We gather the sequence within 3 amino acids of the donor and acceptor residues of a hydrogen bond, which we refer to as their +3/-3 residues. We find that some hydrogen bonds are near the C- or N-terminus of a peptide or protein and missed one or more of the +3/-3 residues around their donor and acceptor groups. In these cases, we mark the missing residues as XXX. In addition, some proteins have local structural uncertainty with multiple possible amino acids in one position, and we mark these uncertain residues as +++ in the sequence analysis. The probabilities of observing the XXX and +++ residues are 5.1% and 0.1%, respectively, in the entire dataset. Given their rare occurrence, we remove all hydrogen bonds with the XXX or +++ residues in their +3/-3 sequences in both the training and test datasets.
3.2.2 Development of the machine learning model

This is a collaborative work, and the machine learning model is developed by Professor Sijian Wang and Yuanhao Liu. They use the generalized boosted regression modeling (gbm) function in the R programming language to develop a boosting model for each balanced dataset (step 3 in the procedure and Figure 3.1). They set the shrinkage as 0.01 and the number of decision tree models as 5000. The interaction depth is treated as a tuning parameter, for which they use a 10-fold cross validation to choose the optimal value. Here they randomly split a balanced dataset into 10 groups with equal sizes, and for each group, they train the model on the remaining 9 groups and record the validation error on this group. They then calculate the average validation error of all groups for each interaction depth. By varying the interaction depth from 1 to 15, they determine the optimal value that minimizes the average validation error. Using the optimal interaction depth, they refit the boosting model on the whole balanced dataset and save it as the final boosting model.

For each boosting model, they use the varImp function in the caret package to calculate the importance score of each feature. The final importance scores for the whole machine learning model are calculated from averaging over the 10 boosting models.

3.2.3 Electronic structure calculations

The interaction energy of the hydrogen bond pair is obtained from electronic structure calculations of the energy difference between the hydrogen bonded dimer and the corresponding monomers, \( \Delta E = E_{Hbond} - (E_{Donor} + E_{Acceptor}) \). The hydrogen bond energy surfaces are computed by keeping the relative orientation of the hydrogen bond donor and acceptor groups and changing \( R \) of the hydrogen bond. In all calculations, the side chains of Tyr, Ser, Arg and Asp are represented using the compounds 4-ethylphenol, ethanol, protonated 2-butylguanidine and propanoate, respectively, and the interaction energies are computed after optimizing the geometry of each structure with non-hydrogen atoms fixed in space. The electronic structure is described using the B3LYP density functional, the D3(op)
dispersion correction\textsuperscript{178} and the aug-cc-pVDZ basis set.\textsuperscript{179} Here density functional theory, rather than post-Hartree-Fock methods such as MP2 and coupled cluster, is used to enable the energy decomposition analysis with the continuum solvation model. The conductor-like polarizable continuum model (PCM) method is used to mimic the protein and solvent environment using the ALMO-EDA(solv) method\textsuperscript{180} as implemented in the Q-Chem 5.3 software.\textsuperscript{181}

### 3.3 Results and discussion

#### 3.3.1 Dataset for biological hydrogen bonds

From the PDB, we obtain 2171 peptide and protein structures that are determined from X-ray or neutron scattering experiments and have a resolution equal to or higher than 1.1 Å. 35\% of these structures are small proteins with fewer than 150 amino acids, and 12\% of them are peptides containing less than 50 amino acids. We note that some proteins have multiple PDB entries for their wild-type, mutated and ligand-bound structures. Considering their structural similarity, we remove the “redundant” structures with the same protein name and only include one structure with the highest resolution for each peptide and protein in the training dataset. Finally, we include 782 structures with unique protein names in the training set. To further assess the redundancy of the peptide and protein sequences, we use the CD-HIT web server\textsuperscript{182} to analyze the resulting 782 structures in the training set. After setting the sequence identity cutoff to 0.9, we find that the 853 chains reduce to 741 independent chains, suggesting that there is still a small degree of redundancy in the training data. However, we expect this small sequence redundancy to have minor impact on the development of the machine learning model. In comparison, we keep all the 478 remaining structures of peptides and proteins in the test set so that the model predictions cover all available data.

We find that the location of the donor and acceptor residues can play an important role in determining the class of a hydrogen bond. As shown in Table 3.1, the probability of finding
a SHB is 24% and 13% for a side chain-side chain and side chain-backbone hydrogen bond, respectively. In contrast, this probability reduces to 0.4% when the donor group is in the protein backbone, and hence we preclude these cases to avoid highly imbalanced data. From the 1260 protein structures in the training and test datasets, we collect a total of 10161 SHBs and 44871 NHBs that form between amino acids and have the side chain of an amino acid as the donor group. Here we consider a NHB as $2.8 \leq R \leq 3.2$, and we use a separation of 0.1 Å in R to better distinguish a SHB and NHB. This analysis yields 6181 SHBs and 26929 NHBs in the training set, and 3980 SHBs and 17942 NHBs in the test set. Therefore, the probability of observing SHBs in the training and test sets are 18.7% and 18.2%, respectively.

Table 3.1: Summary statistics of the training and test datasets when the donor (D) and acceptor (A) residues of a hydrogen bond in the location combination (DA) is in the backbone (B) or side chain (S) of a peptide or protein.

<table>
<thead>
<tr>
<th>Location Combination</th>
<th>Number of SHBs</th>
<th>Number of NHBs</th>
<th>Percentage of SHBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>6711</td>
<td>21619</td>
<td>23.7%</td>
</tr>
<tr>
<td>SB</td>
<td>3450</td>
<td>23252</td>
<td>12.9%</td>
</tr>
<tr>
<td>BS</td>
<td>347</td>
<td>25095</td>
<td>1.4%</td>
</tr>
<tr>
<td>BB</td>
<td>375</td>
<td>149991</td>
<td>0.2%</td>
</tr>
<tr>
<td>SS + SB</td>
<td>10161</td>
<td>44871</td>
<td>18.5%</td>
</tr>
</tbody>
</table>

3.3.2 Machine Learning model for the prediction of SHBs in proteins

Based on our work in Chapter 2, we use 21 variables that cover the structural, chemical and sequence information of SHBs as the input features of the machine learning model. These include the residues and heteroatoms involved in the formation of hydrogen bonds, the charge, location (backbone or side chain) and secondary structure of the donor and acceptor groups, and the sequence of amino acids that are within 3 residues before and after the hydrogen bonded residues. We only consider N and O as the type of the heteroatoms in the hydrogen bonds, and do not further separate their types in amino acids such as Asn and Gln with complex side chains since these amino acids have been found to mostly form
The location of the donor residues is fixed at the protein side chain. As the output, the model is expected to classify a given hydrogen bond as a SHB or NHB.

We face two challenges in the development of the machine learning model. First, it is difficult to predict SHBs correctly as they occur much less frequently than NHBs in proteins. For example, SHBs take only 18.7% of our training data and a standard machine learning model would be more likely to categorize a given hydrogen bond as a NHB. Second, the model involves a considerable number of input features, each of which can take multiple values, and there are strong interaction effects among them when predicting the formation of SHBs. Taking the type of the donor residue as an example, its effect depends heavily on the type of the acceptor residue in the predictions. If Gln is the acceptor residue, the probability of observing a SHB is 32% when the donor is Thr and 11% when the donor is Lys, giving an odds ratio of 3.8 (\(= \frac{\left(\frac{32\%}{1-32\%}\right)}{\left(\frac{11\%}{1-11\%}\right)}\)). In comparison, if Glu acts as the hydrogen bond acceptor, the probabilities become 65% and 23% for the donor residues Thr and Lys, respectively. In the latter case, the odds ratio of 6.2 is almost two fold of that in the former case. Therefore, it is crucial to account for the interaction effects among the 21 input features to achieve a good prediction performance of our model.

To address the first challenge, Professor Wang and Yuanhao use the undersampling strategy\(^{183}\) to develop the machine learning model. The main idea is to create multiple balanced datasets, which contain equal numbers of SHBs and NHBs, by sampling subsets of NHBs from all the training data. As SHBs and NHBs appear equally frequently in these datasets, the resulting models have enhanced performance in predicting the occurrence of SHBs and they will ensemble them to obtain the final model. To address the second challenge, they invoke the boosting model\(^{184}\) that fits a series of decision tree models sequentially and adaptively. From the tree structure of each model, they can automatically and efficiently incorporate the interaction effects among the 21 structural, chemical and sequence input features. They then combine them to form the final model, which has a significant boosted performance in the prediction of SHBs compared to the individual de-
cision tree models. They have also tested other traditional models such as random forests, support-vector machines and multilayer feedforward neural networks and have found that these models achieve similar level results. They have chosen the boosting based model in this chapter as it can provide a good interpretation to the SHB predictions. They hence take the following steps to develop the machine learning model and the workflow is schematically represented in Figure 3.1.

**Step 1.** Generate a subset of NHBs by randomly selecting 6181 NHBs from the entire training dataset;

**Step 2.** Combine this subset of NHBs with the SHBs to form a balanced training set that contains equal numbers of SHBs and NHBs;

**Step 3.** Train a boosting model on this balanced training set;

**Step 4.** Repeat steps 1 – 3 to obtain 10 boosting models;

**Step 5.** Average over 10 models to construct the final machine learning model.

We name the resulting model as the machine learning assisted prediction of short hydrogen bonds (MAPSHB) model. We have further designed a web server for it on https://www.sas.rutgers.edu/cms/wanggroup/mapshb-model/the-mapshb-model, which allows a user to upload a protein structure with moderate or low resolution and obtain the probability of each hydrogen bond as a SHB. By defining a probability threshold, we then predict a hydrogen bond to be a SHB if its probability is greater or equal to this value, or a NHB if the probability is below it. To assess the performance of the MAPSHB model, we apply it to the test dataset and examine two metrics of its predictions. One metric is precision, which is the proportion of true SHBs among all of the predicted SHBs; the other is recall, which is the proportion of the predicted SHBs among all the SHBs in the test dataset. For both metrics, larger values correspond to higher model performances.

As shown in Table 3.2, one can use the probability threshold to tune the precision and recall of the MAPSHB model and modulate the predictions for a specific research need. A larger threshold value yields a prediction of fewer hydrogen bonds as SHBs, and results
Figure 3.1: Workflow for the development of the MAPSHB model. In the first step, we randomly sample the NHBs in the training dataset and create 10 subsets that contain 6181 NHBs. In the second step, we form 10 balanced datasets by combining each subset of NHBs with all the SHBs. Next, we develop a boosting model from each of the 10 balanced training set and obtain the final MAPSHB model from their average.

In a higher precision but lower recall of the model. In contrast, a smaller threshold leads to a prediction that most of the hydrogen bonds are SHBs and hence a lower precision but higher recall. From Table 3.2, if one requires the model to have highly precise predictions of SHBs in their protein structures, a stringent threshold of 0.996 can be used to control the precision to be 95%. However, the recall is relatively low and only 20% of SHBs in the test set are identified. If one instead wants to explore all the plausible SHBs in a protein, a small threshold of 0.062 can be chosen to reach an excellent recall of 94%. Therefore, one can use the data in Table 3.2 as a guidance and adjust the balance between the precision and recall of the MAPSHB model for their systems. Our recommended probability threshold is 0.870, with which the model predictions can achieve a precision of 80% while maintaining a relatively high recall of 75%. Here, the best score of prediction is about 80% because we are dealing with a strongly unbalanced dataset and it is difficult to reach both high precision and high recall for the model prediction.

To reveal the key factors that promote the formation of SHBs between amino acids, we use the MAPSHB model and calculate the relative importance of its 21 input features.
Table 3.2: The precision and recall of the MAPSHB model with different probability thresholds. The recommended threshold and its relevant metrics are highlighted in bold.

<table>
<thead>
<tr>
<th>Probability Threshold</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.996</td>
<td>95%</td>
<td>20%</td>
</tr>
<tr>
<td>0.979</td>
<td>90%</td>
<td>50%</td>
</tr>
<tr>
<td>0.943</td>
<td>85%</td>
<td>66%</td>
</tr>
<tr>
<td><strong>0.870</strong></td>
<td><strong>80%</strong></td>
<td><strong>75%</strong></td>
</tr>
<tr>
<td>0.740</td>
<td>70%</td>
<td>83%</td>
</tr>
<tr>
<td>0.555</td>
<td>60%</td>
<td>87%</td>
</tr>
<tr>
<td>0.062</td>
<td>40%</td>
<td>94%</td>
</tr>
</tbody>
</table>

demonstrated in Figure 3.2, the type of the donor residue plays a major role in determining the class of a hydrogen bond and possesses the most significant importance score of 28.2% in our predictions. From the training dataset, the donor amino acids can be divided into three groups that give vastly different probability of forming SHBs. The first group is Tyr, which has a remarkable 86% probability to form a SHB when its phenol side chain serves as the donor of a hydrogen bond. The second group includes Arg, Lys, Asn, Gln and Trp, as the probability of observing SHBs is below 15% when the N-containing functional groups of their side chains are donors. Given their distinct preferences in forming SHBs or NHBs, the MAPSHB model can determine the class of a hydrogen bond almost solely from the donor residue when it belongs to these two groups. The third group comprises Ser, Thr and His, and their SHB-forming probabilities are 53%, 40% and 29%, respectively. This suggests that when a hydrogen bond contains the hydroxyl or imidazole side chain of these amino acids as the donor, it is feasible to form either a SHB or a NHB. In this case, the MAPSHB model cannot judge the class of the hydrogen bond only from the donor residue, and it is essential to consider the acceptor properties such as its charge, location and residue and atom types. For example, when Thr is the donor, the probability of observing a SHB is 73.8% if the anionic Asp or Glu is the acceptor residue, and is only 29.2% if a neutral amino acid is the acceptor. Similarly, when Ser and His are the hydrogen bond donors, the presence of the anionic Asp and Glu as acceptors significantly enhances the likelihood of forming SHBs.
Figure 3.2: Normalized importance scores for the 21 input features of the MAPSHB model. The features that contribute ≤1% to the predictions are grouped as the “other” type. These include the atom type, charge and secondary structure of the donor and acceptor groups and the location of the acceptor residue.

From Figure 3.2, the amino acids next to the hydrogen bond donor and acceptor residues in the protein sequence are important features that facilitate the formation of SHBs, and each of them contributes ~5% to the prediction of the MAPSHB model. Interestingly, we observe an enhanced SHB-forming probability when the donor and acceptor residues are separated by only one amino acid in the sequence, possibly because the secondary structure of the protein backbone can help position the hydrogen bonded groups in close proximity. This is particularly the case when Ser or Thr acts as the donor and Asp is the acceptor of a hydrogen bond. For instance, we observe the Ser-Xxx-Asp and Asp-Xxx-Ser (Xxx represents any amino acid) sequence patterns in 36.4% of the hydrogen bonds formed between the side chains of Ser and Asp, and they show a considerable 82.5% probability to form SHBs. As an example, we observe an Asp-Met-Ser sequence in the active site of a cyclopropanase, in which the Asp and Ser side chains form a SHB in the presence of a bound NADP$^+$ cofactor (PDB ID 5DP2). In Figure 3.2, apart from the hydrogen bond donor and acceptor residues and their relevant sequence information, the other features such
as the charge, atom type and secondary structure of the hydrogen bonded groups appear to be less important and contribute $\leq 1\%$ to the overall prediction. It is because they partially overlap with the residue information and only play essential roles when the hydrogen bond donor belongs to the third group of amino acids, which are associated with 25\% of the data.

3.3.3 Side-chain Tyr-Asp and Tyr-Glu pairs are most likely to form SHBs

Using a data-driven approach, the MAPSHB model provides useful guidelines to pinpoint SHBs in proteins. In its top 1623 predictions, the amino acid pairs have a significant (> 99\%) probability of forming SHBs and 71\% of them have the anionic side chains of Asp or Glu as the acceptor. Consistent with Figure 3.2, the most common donors are Tyr in the first group (937 SHBs) and Ser in the third group (419 SHBs). Residues in the second group, such as Arg and Lys, are predicted to have a low probability of forming SHBs. To confirm these predictions, we further analyze all side chain-side chain hydrogen bonds in the dataset. As shown in Table 3.3, the Tyr-Asp pair has the highest probability (96\%) to form a SHB and this is followed by the Tyr-Glu pair (93\%). The Ser-Asp and Ser-Glu pairs have a lower probability of $\sim 79\%$ to form SHBs, but they constitute the largest portion (20\%) of all the side chain-side chain SHBs. The charged Arg-Asp and Arg-Glu pairs possess strong electrostatic stabilization and form the largest number (8671) of hydrogen bonds, but the probabilities of observing them in SHBs are below 6\%.

To uncover the origin of the observed trend, we choose Tyr, Ser and Arg to represent the three groups of donor amino acids and take Asp as an example acceptor residue given the structural similarities between the side chains of Asp and Glu. We have randomly selected 150 configurations for each of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side hydrogen bonds to ensure that their $R$ covers the whole range between 2.3 Å and 3.2 Å and to compute their interaction energies from electronic structure calculations. Among them, we include two representative configurations for each pair of the first and second maxima of their geometric probability distributions (Figure 3.3) and evaluate how their interaction
Table 3.3: Occurrence and probability of surface exposure of the side chain-side chain hydrogen bonds when Tyr/Ser/Thr/His/Arg/Lys is the donor and Asp/Glu is the acceptor in the entire data set.

<table>
<thead>
<tr>
<th>Donor</th>
<th>Acceptor</th>
<th>n(HBs)</th>
<th>n(SHBs)</th>
<th>n(NHBs)</th>
<th>$P_{SHBs}$</th>
<th>$P_{s,donor}$</th>
<th>$P_{s,acceptor}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>Asp</td>
<td>4654</td>
<td>179</td>
<td>4475</td>
<td>3.85%</td>
<td>84.62%</td>
<td>75.87%</td>
</tr>
<tr>
<td>Arg</td>
<td>Glu</td>
<td>4017</td>
<td>230</td>
<td>3787</td>
<td>5.73%</td>
<td>82.72%</td>
<td>76.75%</td>
</tr>
<tr>
<td>Lys</td>
<td>Asp</td>
<td>1191</td>
<td>198</td>
<td>993</td>
<td>16.62%</td>
<td>88.50%</td>
<td>84.38%</td>
</tr>
<tr>
<td>Ser</td>
<td>Asp</td>
<td>1159</td>
<td>920</td>
<td>239</td>
<td>79.38%</td>
<td>68.33%</td>
<td>71.01%</td>
</tr>
<tr>
<td>Lys</td>
<td>Glu</td>
<td>1122</td>
<td>251</td>
<td>871</td>
<td>22.37%</td>
<td>93.67%</td>
<td>90.64%</td>
</tr>
<tr>
<td>Thr</td>
<td>Asp</td>
<td>1113</td>
<td>785</td>
<td>328</td>
<td>70.53%</td>
<td>69.81%</td>
<td>65.23%</td>
</tr>
<tr>
<td>Tyr</td>
<td>Asp</td>
<td>821</td>
<td>786</td>
<td>35</td>
<td>95.74%</td>
<td>48.48%</td>
<td>63.70%</td>
</tr>
<tr>
<td>Tyr</td>
<td>Glu</td>
<td>710</td>
<td>661</td>
<td>49</td>
<td>93.10%</td>
<td>52.11%</td>
<td>63.66%</td>
</tr>
<tr>
<td>Ser</td>
<td>Glu</td>
<td>572</td>
<td>443</td>
<td>129</td>
<td>77.45%</td>
<td>55.07%</td>
<td>59.44%</td>
</tr>
<tr>
<td>His</td>
<td>Asp</td>
<td>549</td>
<td>274</td>
<td>275</td>
<td>49.91%</td>
<td>67.94%</td>
<td>58.29%</td>
</tr>
<tr>
<td>Thr</td>
<td>Glu</td>
<td>424</td>
<td>349</td>
<td>75</td>
<td>82.31%</td>
<td>59.43%</td>
<td>56.13%</td>
</tr>
<tr>
<td>His</td>
<td>Glu</td>
<td>356</td>
<td>172</td>
<td>184</td>
<td>48.31%</td>
<td>65.45%</td>
<td>65.45%</td>
</tr>
</tbody>
</table>

energies vary with R. The first maxima of the Tyr-Asp, Ser-Asp and Arg-Asp hydrogen bonds occur at (2.65 Å, 150°), (2.65 Å, 90°) and (2.85 Å, 90°), respectively. Their second maxima are at (2.60 Å, 100°), (2.70 Å, 150°) and (2.85 Å, 150°), respectively. Similarly, the first maxima of the Tyr-Glu, Ser-Glu and Arg-Glu hydrogen bonds occur at (2.60 Å, 150°), (2.65 Å, 90°) and (2.85 Å, 90°), respectively. Their second maxima are at (2.60 Å, 90°), (2.65 Å, 150°) and (2.85 Å, 140°), respectively. In Figure 3.3a-f, the ratio between the second and first maxima are 56%, 42%, 14%, 32%, 93% and 23%, respectively. As shown in Figure 3.4, the Tyr-Asp and Ser-Asp pairs are highly likely to form SHBs while the Arg-Asp pair is most frequently observed as a NHB with R above 2.8 Å.

In the calculation of the hydrogen bond energies, we find it crucial to account for the distinctive electrostatic environment around the donor and acceptor residues. As the 20 common amino acids have different polarity and hydrophobicity, the hydrogen bonds formed from them can be located at different places in proteins. For example, amino acids with polar or charged side chains are more likely to be on the protein surface and exposed to the aqueous environment, while those with non-polar side chains are often buried inside the protein. To estimate the solvation condition of these biological hydrogen bonds, we
first calculate the solvent-accessible surface area (SASA) of the relevant amino acids in the peptides and proteins, which represents the surface areas of these residues accessible to the solvent water molecules. The calculations have been carried out with the FreeSASA software package using Lee and Richards’ algorithm and a probe radius of 1.4 Å. Next, we compute the relative solvent accessibility (RSA) values, which accounts for the size differences of the amino acids and allows us to compare the solvent accessibility among different amino acids. RSA is defined as the SASA of a residue normalized by a maximum possible value of the residue,

\[
RSA = \frac{SASA}{SASA_{\text{max}}}. \tag{3.1}
\]

The theoretical maximum SASA values from the study of Tien and coworkers are taken as \(SASA_{\text{max}}\) values. For each type of the side chain-side chain hydrogen bonds, we calculate the average RSA values of its donor and acceptor residues from the entire dataset. Considering that a residue is usually defined as surface exposed if its RSA is \(\geq 5\%\),
we also include the probability of finding a hydrogen bonded residue on the surface of the peptide or protein, $P_s$, in Table 3.3. Here, for a specific type of hydrogen bond, if its donor or acceptor appears $N$ times in the hydrogen bond dataset with $n$ cases where the corresponding RSA value is $\geq 5\%$, we compute $P_s$ as

$$P_s = \frac{n}{N}. \quad (3.2)$$

From these calculations, we find the Tyr-Asp and Ser-Asp hydrogen bonds are often found in the protein interior (Table 3.3), and we implement the PCM with a dielectric constant of 10.0 to mimic the protein environment. In contrast, the Arg-Asp hydrogen bond is mostly observed on the protein surface as the charged side chains of Arg and Asp have a significant probability of 85% and 76% to be solvent exposed, respectively (Table 3.3), and we use a dielectric constant of 78.4 to represent a water environment around this pair. To examine the impact of our choice of $\varepsilon$ in the solvent model, we repeat the calculations of the hydrogen bond energies using $\varepsilon = 78.4$ for the Tyr-Asp and Ser-Asp hydrogen bonds, and $\varepsilon = 10.0$ for the Arg-Asp hydrogen bonds. As demonstrated in Figure 3.5, the poten-
tial energy curves for the Tyr-Asp and Ser-Asp pairs resemble each other when different dielectric constants are used, with an energy difference < 2 kcal/mol. This means that the hydrogen bond energies are nearly invariant to the $\varepsilon$ values used in the PCM calculations. In contrast, the energies of the Arg-Asp hydrogen bond are very sensitive to the solvent environment as the energy can differ by 10.5 kcal/mol when different $\varepsilon$ values are used.

Figure 3.5: Hydrogen bond energies for different configurations and solvation environments of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds when $R$ is between 2.3 Å and 3.2 Å. For each pair, the red and blue lines represent the the PCM model with the $\varepsilon = 10.0$ and 78.4, respectively. The solid and the dashed lines represent the configurations from the first and the second maximum of the probability distributions, respectively.

As shown in Figure 3.6, for each type of the hydrogen bond, the interaction energies of the 150 geometry arrangements closely follow the potential energy surfaces that are computed from their most probable configurations. For example, the minima of the interaction energy occur at an O–O distance, $R_{\text{min}}$, of 2.55 Å for the Tyr-Asp hydrogen bond, and 2.65 Å for the Ser-Asp hydrogen bond. Consistent with the MAPSHB predictions, the $R_{\text{min}}$ values of both pairs are within the range of SHBs, indicating that it is energetically favorable for their donor and acceptor residues to stay in close proximity. In contrast, $R_{\text{min}}$ of the Arg-Asp hydrogen bond are at 2.75 Å in its potential energy surface and the interaction energies rise sharply when $R$ shortens, confirming that this combination of amino acid side chains predominantly forms NHBs. In the following, we will analyze the most frequently observed configuration of each hydrogen bond since their potential energy curves
well represent the corresponding hydrogen bond energies in proteins and are lower by 0.3-6.4 kcal/mol than those of the second most probable configurations at all hydrogen bond lengths.

Figure 3.6: Interaction energies of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds. For each type of the hydrogen bond, the red dots represent the interaction energies of 150 randomly chosen configurations. We further calculate how the hydrogen bond energies of the configurations from the first and second maxima of the geometric probability distributions vary with R, which are shown as the solid and dashed lines, respectively, in each plot. The vertical line represents the potential energy minimum of the Tyr-Asp pair at R = 2.55 Å.

From Figure 3.6, we choose an $R_{\text{min}}$ of 2.55 Å in the potential energy curves of the Tyr-Asp, Ser-Asp and Arg-Asp pairs, and decompose their interaction energies using the ALMO-EDA(solv) method\textsuperscript{180} to elucidate why the Tyr-Asp combination shows the highest probability of forming SHBs in proteins. As demonstrated in Figure 3.7, the total interaction energy of a hydrogen bond is partitioned into the frozen interaction, polarization and charge transfer energies,\textsuperscript{180}

$$\Delta E_{TOT}^{(s)} = \Delta E_{FRZ}^{(s)} + \Delta E_{POL}^{(s)} + \Delta E_{CT}^{(s)}. \quad (3.3)$$

Here the superscript \textsuperscript{(s)} indicates that the decomposition is carried out using a solvent model. The frozen interaction energy, $\Delta E_{FRZ}^{(s)}$, describes the energy difference between a solvated hydrogen bond pair and its individually solvated, non-interacting donor and acceptor residues without any orbital relaxation. It can be further decomposed into the
contributions from the permanent electrostatics, Pauli repulsion and dispersion interactions in vacuum and a solvation term (Table 3.4),

\[ \Delta E_{TOT}^{(s)} = (\Delta E_{ELEC}^{(v)} + \Delta E_{PAULI}^{(v)} + \Delta E_{DISP}^{(v)} + \Delta E_{SOL}) + \Delta E_{POL}^{(s)} + \Delta E_{CT}^{(s)} \]  

(3.4)

The effects of orbital interactions and electron redistribution are incorporated in the \( \Delta E_{POL}^{(s)} \) and \( \Delta E_{CT}^{(s)} \) energies.\(^{180}\)

![Figure 3.7: Decomposition of the interaction energies for the Tyr-Asp, Ser-Asp and Arg-Asp hydrogen bonds at R = 2.55 Å. The total interaction energy (\( \Delta E_{TOT}^{(s)} \)) is partitioned into the frozen interaction (\( \Delta E_{FRZ}^{(s)} \)), polarization (\( \Delta E_{POL}^{(s)} \)) and charge transfer (\( \Delta E_{CT}^{(s)} \)) energies.]

![Table 3.4: Decomposition of interaction energies of the Tyr-Asp, Ser-Asp and Arg-Asp side chain-side chain hydrogen bonds at R = 2.55 Å. The energies are in the unit of kcal/mol.]

<table>
<thead>
<tr>
<th></th>
<th>Tyr-Asp</th>
<th>Ser-Asp</th>
<th>Arg-Asp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrostatics</td>
<td>-63.6</td>
<td>-53.4</td>
<td>-187.2</td>
</tr>
<tr>
<td>Pauli repulsion</td>
<td>64.3</td>
<td>55.6</td>
<td>121.1</td>
</tr>
<tr>
<td>Dispersion</td>
<td>-6.3</td>
<td>-5.0</td>
<td>-10.3</td>
</tr>
<tr>
<td>Solvation</td>
<td>10.8</td>
<td>8.7</td>
<td>101.2</td>
</tr>
<tr>
<td>Polarization</td>
<td>-7.0</td>
<td>-5.3</td>
<td>-14.5</td>
</tr>
<tr>
<td>Charge transfer</td>
<td>-11.1</td>
<td>-9.1</td>
<td>-22.4</td>
</tr>
<tr>
<td>Total interaction energy</td>
<td>-12.9</td>
<td>-8.5</td>
<td>-12.1</td>
</tr>
</tbody>
</table>
From Figure 3.7, the total interaction energies of the Tyr-Asp, Ser-Asp and Arg-Asp pairs are distributed among all three terms, with the frozen interaction acting to weaken the hydrogen bonds and the polarization and charge transfer interactions strengthening them. As shown in Table 3.4, the positive frozen interaction energies come from the large Pauli repulsion (over 55 kcal/mol), which is due to the short contact between the donor and acceptor groups, and the solvation energies (8–102 kcal/mol) as the solute-solvent interactions damp the Coulomb attraction in the hydrogen bond. In the meantime, the attractive electrostatic and dispersion interactions cancel about 90% of these unfavorable interactions, leading to a $\Delta E_{FRZ}^{(s)}$ between 5 and 25 kcal/mol for the three types of hydrogen bonds. Comparing the Tyr-Asp and Ser-Asp pairs, they share similar $\Delta E_{FRZ}^{(s)}$ values as the donor residues both contain hydroxyl groups. However, the former has a longer O–H bond and a more shared proton in the hydrogen bond because the phenol side chain of Tyr is more acidic (with a $pK_a$ value of 10.1) than the hydroxymethyl side chain of Ser. As a result, the Tyr-Asp hydrogen bond has more prominent contributions from the polarization and charge transfer terms, which lower the interaction energy by 3.7 kcal/mol more than that of the Ser-Asp hydrogen bond. Similar behavior is observed over the whole range of R (Figure 3.8), and a delicate balance of the frozen interaction, polarization and charge transfer makes the Tyr-Asp pair more stable at a short R of 2.55 Å than the Ser-Asp pair. Comparing the Tyr-Asp and Arg-Asp hydrogen bonds, the latter has significantly larger components of the intermolecular interactions because both of its donor and acceptor groups are charged (Figure 3.7). For example, the polarization and charge transfer interactions in the Arg-Asp hydrogen bond give rise to a stabilization energy of -36.9 kcal/mol, which is over twice of that in the Tyr-Asp case. However, its frozen interaction is 5 times larger than that of the Tyr-Asp pair and destabilizes the overall interaction by 24.8 kcal/mol. From Table 3.4, this arises from a Pauli repulsion of 121.1 kcal/mol in this N-H· · ·O type hydrogen bond, and a solvation energy of 101.2 kcal/mol that smears out the Coulomb attraction between the cationic Arg and anionic Asp residues. From Figure 3.8, the repulsive frozen interaction of
the Arg-Asp pair decays rapidly when R lengthens, making it more stable when the donor
and acceptor residues are separated in the NHB range.

Figure 3.8: Decomposition of the interaction energies for the Tyr-Asp (red), Ser-Asp (blue)
and Arg-Asp (brown) side chain-side chain hydrogen bonds when R is between 2.3 Å and
3.2 Å. The hydrogen bonds take their most likely configurations (Figure 3.4).

Given their prevalence in proteins, we have considered the hydrogen bonds that form
between the hydroxyl side chains of Tyr, Ser or the guanidinium side chain of Arg and
the carboxylate side chains of Asp or Glu, and identified that their preferred lengths come
from a balance of several competing effects. As demonstrated in Figure 3.7, the frozen
interaction tends to weaken a SHB mainly because the large Pauli repulsion prevents the
donor and acceptor residues from staying in close proximity. The solute-water interactions
also considerably destabilize hydrogen bonds that contain charged residues, and are partic-
ularly repulsive when a hydrogen bond forms between cationic and anionic residues, as in
the cases of the Arg-Asp and Lys-Asp pairs. While electrostatic interactions play an impor-
tant role in counteracting these effects, polarization and other purely quantum mechanical
effects such as dispersion and charge transfer also contribute 20% – 28% to the overall
stabilization energy (Table 3.4). Among the neutral amino acids, Tyr has the most ioniz-
able chain from its conjugated ring structure and hence is likely to share its proton with
the Asp/Glu acceptor residue. As a result, the stabilization effects from polarization and
charge transfer dominate in the Tyr-Asp hydrogen bond and make it the most energetically
favorable at a short heteroatom separation of 2.55 Å. Furthermore, the interaction energy
of the Tyr-Asp pair is mostly invariant with the solvation condition, whereas the aqueous environment considerably weakens the charged Arg-Asp hydrogen bond (Figure 3.5) and makes it most stable with $R$ above 2.7 Å. Due to the interplay of the frozen interaction, polarization and charge transfer interactions, the Tyr-Asp/Glu pair has the highest probability of forming SHBs among the commonly observed combinations of amino acid side chains.

3.4 Conclusions

In this chapter, we have developed a MAPSHB model that effectively predicts SHBs in proteins that form between amino acids and have the side chain of an amino acid as the donor residue. By tuning the probability threshold, we successfully control the precision and recall of its predictions to both around 80%. We have designed a web server for the MAPSHB model on https://www.sas.rutgers.edu/cms/wanggroup/mapshb-model/the-mapshb-model, which allows a user to input a protein structure with low to moderate resolution and identify probable SHBs in the biomolecule. It will provide additional restraints for the experimental and computational refinement of protein structures and facilitate the determination of the structure and functional roles of SHBs in proteins.

From the MAPSHB model, we uncover three main features that promote the formation of SHBs in proteins. First, the donor amino acid plays a major role in determining the class of a hydrogen bond. This is particularly the case for Tyr, which exhibits a strong preference for the formation of SHBs. Second, the carboxylate side chains of Asp and Glu are the most frequently observed acceptors of SHBs in proteins. They strongly modulate the likelihood of observing these compact structures when the donor residue belongs to the third group of amino acids, namely Ser, Thr and His. Finally, the sequence of amino acids next to the hydrogen bond groups can also facilitate the formation of these close contacts. For example, there is an enhanced probability of observing a close contact in a Ser-Asp or Thr-Asp hydrogen bond when the donor and acceptor residues are separated by only one amino acid in the sequence. Following the first two rules, we find that the Tyr-Asp and
Tyr-Glu side chain-side chain hydrogen bonds have the highest probability of forming a SHB in proteins. Combining electronic structure calculations and energy decomposition analysis, we compare the Tyr-Asp pair with the Ser-Asp and Arg-Asp hydrogen bonds and reveal that this trend comes from a competition of intermolecular interactions. While the frozen interaction tends to push the heteroatoms away in space, Coulomb attraction, polarization and other purely quantum mechanical effects such as dispersion and charge transfer play key roles in stabilizing a SHB. A delicate balance of these effects makes the Tyr-Asp and Ser-Asp pairs energetically favorable at an O-O distance below 2.7 Å. In comparison, while the cationic Arg and anionic Asp tend to form the largest amount of hydrogen bonds, their interactions with the aqueous solution make the Arg-Asp pair most stable as a NHB. The predictions of the MAPSHB model, in conjunction with our elucidation of the origin of these hydrogen bonding interactions, will guide the design of novel protein systems that take advantage of SHBs to enhance their functions in biochemical and materials applications.
CHAPTER 4
STRUCTURAL AND CHEMICAL FEATURES OF BIOLOGICAL SHORT HYDROGEN BONDS IN PROTEIN-LIGAND COMPLEXES

4.1 Introduction

Despite the prevalence of SHBs in protein-ligand complexes, understanding their structures, functions and dynamics is challenging due to the limited resources of high-resolution structures. The PDB not only contains over 184,000 biological macromolecular structures, but also archives over 35,000 ligands, providing a data source for investigating the features of protein-ligand interactions. In this chapter, we systematically analyze the atomic-resolution structures in the PDB to uncover the structural and chemical features of SHBs that form between amino acids and ligands. Similar to our MAPSHB model in Chapter 3, we collaborate with Professor Sijian Wang and Yuanhao Liu to develop a machine learning model for the SHBs between one amino acid and one ligand based on 1070 atomic-resolution protein structures. We then take the carbohydrate-ligand SHBs as an example and use electronic structure calculations to uncover the factors that promote the formation of these SHBs.

4.2 Methods

The structural preparation is carried out using the same method discussed in Chapter 2. We obtain the residue and atom types, charge and location of the amino acids and the relevant secondary structures and sequence information using AmberTools16. For ligands, we determine the type of atoms and functional groups in the hydrogen bonds, and use the Molecular Operating Environment (MOE) software to estimate their $pK_a$, $pK_b$ and octanol-water partition coefficients (logP). Many of the ligands involved in the SHBs are
inorganic anions, small inorganic molecules or ions and polyols, such as \( \text{SO}_4^{2-} \), \( \text{PO}_4^{3-} \), ethylene glycol and glycerol. We will not consider them since they are mainly used in the solvation of biomolecules for experimental measurements. A list of chemical IDs for the excluded ligands is shown in Table 4.1.

We obtain the interaction energy of the hydrogen bond pair from the energy difference between the hydrogen bonded dimer and the corresponding monomers, \( \Delta E = E_{\text{Hbond}} - (E_{\text{Donor}} + E_{\text{Acceptor}}) \). The potential energy surfaces for the hydrogen bond interactions are computed by keeping the relative orientations of the amino acids and ligands and changing R of the hydrogen bond. The representations of the side chains of amino acids and the computational details are the same as discussed in Chapter 3. We also apply a conductor-like PCM method with a dielectric constant of 10 to mimic the protein environment, and perform energy decomposition analysis using the ALMO-EDA(solv) method as implemented in the Q-Chem 5.3 software.

4.3 Results and discussion

4.3.1 Short hydrogen bonds in protein-ligand complexes

From 1070 protein-ligand complexes in the PDB, we find a total of 2799 amino acid-ligand SHBs. After removing small inorganic particles and polyols from the ligand categories, we finally identify 1393 SHBs that are distributed among 634 protein-ligand complexes. Structurally, hydrogen bonds that involve ligands are more likely to have shorter R when compared to those between amino acids, as demonstrated in Figure 4.1a. For example, 29.1% of these SHBs have \( R < 2.6 \ \text{Å} \), whereas the ratio in protein-protein SHBs is only 18.2%. From Figure 2.7, we expect to observe more prominent proton sharing in the ligand-containing SHBs, which arise from an interplay of R and quantum effects in both the electronic and nuclear degrees of freedom.

Many of the SHBs are formed between active-site residues and ligands, and they possibly participate in crucial biological functions. Given the diversity of the ligands, we
Figure 4.1: (a) Percentage distribution of SHBs formed between two amino acids and between amino acids and ligands. (b) Distribution of the different categories of ligands in the 1393 amino acid-ligand SHBs. All the other ligands are grouped into the “Other” category.

combine some of them with similar molecular structures and biological functions to form a few categories, which include carbohydrates, nucleotides, acids or anions, heme and non-proteinogenic amino acids. The remaining ligands include alcohol, drugs and metal-containing ligands, and we combine them into the “Other” category (Table 4.1). As shown in Figure 4.1b, the most abundant ligands in the amino acid-ligand SHBs are carbohydrates. A typical structure of a carbohydrate contains a closed ring with various hydroxyl groups and it is capable of forming multiple hydrogen bonds with other molecules. In particular, we find that $\alpha$-L-fucose (FUC), $\beta$-D-glucose (BGC) and $\alpha$-D-mannose (MAN) are the most common carbohydrates to form SHBs with amino acids and they mostly exist in carbohydrate-binding proteins. Nucleotides also actively participate in the formation of SHBs in proteins. This is especially the case for pyridine nucleotides and flavin nucleotides, which contribute to 72% SHBs in this category. Pyridine nucleotides include nicotinamide adenine dinucleotide (NAD$^+$), nicotinamide adenine dinucleotide phosphate (NADP$^+$) and their reduced forms NADH and NADPH, while the flavin nucleotides include flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN) and their proto-
nated or hydroxylated forms. These enzyme cofactors are composed of two nucleotides joined through ribose and phosphate groups, in which the O-containing functional groups are good sources to form SHBs. The pyridine nucleotides, in particular NADP$^+$, are electron carriers for important redox reactions in metabolism.\textsuperscript{202–204} As an example, Figure 4.2a shows the active-site cavity of a tetrahydroalstonine synthase that plays a key role in the biosynthesis of heteroyohimbine.\textsuperscript{204} NADP$^+$ is anchored by residue Glu59 through bidentate hydrogen bonds, one of which is a SHB with R of 2.49 Å. Furthermore, NADP$^+$ accepts a hydrogen bond from Ser211 with an R of 2.59 Å, and these SHBs hold the ligand in place for enzymatic redox reactions.\textsuperscript{204} Besides, the flavin nucleotides are widely observed in flavoproteins, in which they serve as cofactors to catalyze cellular redox reactions.\textsuperscript{205–207} As an example, the FAD-binding domain of aldotol oxidase, a flavoprotein that selectively oxidizes the terminal hydroxyl groups of sugar alcohols, is shown in Figure 4.2b.\textsuperscript{207} The pyrophosphate group of FAD forms two SHBs with residues Ser44 and Ser47 with R of 2.60 and 2.54 Å, respectively. These SHBs likely act to position the FAD cofactor in the FAD-binding domain of the enzyme to facilitate catalysis. From Figure 4.1b, 12\% of the ligands in the amino acid-ligand SHBs belong to acids or anions, and most of them are fatty acids and carboxylates such as citric acid (CIT) and malonate ion (MLI). Due to the presence of the carboxylic or carboxylate groups, these acids or anions act like Asp or Glu and their neutral protonated forms in forming SHBs. As hemes contain carboxylic groups connected to porphyrins, they also contribute to 6.4\% amino acid-ligand SHBs. These heme-containing SHBs are distributed in a variety of proteins ranging from nitrophorin 4, myoglobin to cytochrome c and dehaloperoxidase hemoglobin.\textsuperscript{208–212} Finally, non-proteinogenic amino acids, such as S-adenosyl-L-homocysteine (SAH) and D-glutamic acid (DGL), are also occasionally observed to form SHBs.
Table 4.1: Chemical IDs in the PDB and the categories of the ligands in this study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Chemical IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded ligands</td>
<td>HOH WAT DOD OH NH3 NH4 CYN NO3 CO2 CO3 BCT SO4</td>
</tr>
<tr>
<td></td>
<td>PO4 PI WO4 8AR CAD SE4 PO3 NO2 NO POP EOH GOL EDO</td>
</tr>
<tr>
<td></td>
<td>PEG MOH PGE PGO PG4 1PE NH2 OXY F CAC AZI DMS SCN P33 XPE HEZ IPA BU3 15P MXE DIO PE4 P6G ME2 ETF DXE</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>GLC BMA MAN BG6 ADA RAM CBI E4P F6R BGC NAG GAL</td>
</tr>
<tr>
<td></td>
<td>LMT MMA FRU FUC NDG 5RP XYS GN1 XYP MGL MFU G7P</td>
</tr>
<tr>
<td></td>
<td>MA4 AHR PRP LAT CE5 BOG GTR GLA IDC 4IP SGC SSG A2G</td>
</tr>
<tr>
<td>Nucleotides</td>
<td>FAD FMN FAE BYN NAP NAD NAJ NDP ODP 8GD URI</td>
</tr>
<tr>
<td></td>
<td>60G 7CI UDP 8DG AP5 B4P AMP ATP KBD UMC GDP ADP</td>
</tr>
<tr>
<td></td>
<td>ACP AMZ G3D NEU 3AM U1P APC U5P GNP ACO JBT 5GP</td>
</tr>
<tr>
<td></td>
<td>SAM DCM COA 6U4 MCN ADN</td>
</tr>
<tr>
<td>Acids or anions</td>
<td>FMT CIT SAL ACT FLC MHA OCA TLA SIN ACA HLP IAC</td>
</tr>
<tr>
<td></td>
<td>FQZ 3DS XZA ZPO HHT HJK HGH HHB HHW SIA PLM DKA</td>
</tr>
<tr>
<td></td>
<td>ACY AKG S3P TFA ZGB GRI ISC HCA NIO MLT TAR 01F 3EB</td>
</tr>
<tr>
<td></td>
<td>BEZ 173 OXM FOL PYR DAO MLI 0V5 K12 PPF EDT HC4 PHB FER MCO 1DF</td>
</tr>
<tr>
<td>Hemes</td>
<td>HEM HEC FDD FDE 6HE</td>
</tr>
<tr>
<td>Amino acids</td>
<td>DLY DTH DCY ALO OCS OGA GHP OMY AIB DLE DAS DSN</td>
</tr>
<tr>
<td></td>
<td>DGL DAL DTR TPO CCS MEA 9KK B2H SAH DA2 PCA MSE</td>
</tr>
<tr>
<td></td>
<td>HYP CSO OHI ACE GYS CRO PHL CSD CXM 23F VDL VLL</td>
</tr>
<tr>
<td></td>
<td>D3P NLE LYZ TRQ KOR SNC CGU OSE M3L CSX ZZU BB9</td>
</tr>
<tr>
<td></td>
<td>SLB ETA FVA HM8 MLY SEP AGD C6L FME MDO CSS MIR DCL OLD IYR TPQ QIL C4L FP9 GSF GTS PIA DIV AGM MGN</td>
</tr>
<tr>
<td></td>
<td>DY A MHS SMC MHO PTR MLZ HIC VOL SU1 DBB</td>
</tr>
</tbody>
</table>

4.3.2 Structural and chemical features of amino acid-ligand short hydrogen bonds

Among the 1393 amino acid-ligand SHBs, we find that ligands can be both donor and acceptor in a SHB and 63.6% of them act as the donor groups. Similar to the observation in the protein-protein SHBs, 82.5% of the amino acids and 97.3% of the ligands have O atoms in the amino acid-ligand SHBs. This makes O–H···O is the most commonly observed type of amino acid-ligand SHBs. Moreover, we find 91.3% of the SHBs are between a side chain of an amino acid and a ligand, which results in a total of 1272 SHBs.

Unlike the limited number of proteinogenic amino acids, there is a large diversity in the molecular structures of ligands. However, the classification of ligands in Figure 4.1
does not specify the type of atoms participating in the formation of SHBs. Given that ligands usually have large molecular size, not all the atoms play important roles in forming SHBs. We find that atoms of a specific functional group, such as the hydroxyl group in carbohydrates or nucleotides, have similar propensity to form SHBs. We thus focus on the functional groups in ligands and find 11 types that are most likely to form SHBs with amino acids in proteins. As shown in Figure 4.3a, these functional groups include amide, carboxyl, carboxylate, ester, alkyl and aromatic hydroxyl, phosphate, sulphate, alkyl amine and the N-containing aromatic rings. Among them, the alkyl hydroxyl groups are the predominant donor groups and contribute to over half of the SHBs. There are also many SHBs formed with carboxyl and phosphate groups, and they can be either donor or acceptor depending on their protonation states. In Figure 4.3a, we have also examined the occurrence of 11 proteinogenic amino acids with polar side chains, and observe that the negatively charged Asp and Glu are strongly favorable as the acceptor in the amino acid-ligand SHBs, especially for the formation of shorter hydrogen bonds with $R < 2.6$ Å. Similar to the case of SHBs between two amino acids, the neutral amino acids Ser, Thr and Tyr are frequently observed as both donors and acceptors, and histidine can be the donor or

Figure 4.2: Examples of SHBs formed between amino acids and (a) NADP$^+$ (PDB ID 5F13) and (b) FAD (PDB ID 2VFR) in the active-site of proteins. Silver, red, blue, white and tan represent C, O, N, H and P, respectively. The SHBs are represented by dotted lines.
acceptor of a SHB depending on its protonation state. We expect to observe both charged and neutral amino acid-ligand SHBs due to the large population of neutral alkyl hydroxyl groups and the negatively charged Asp, Glu, phosphate and carboxylate groups. As shown in Figure 4.3b, these SHBs are indeed abundant at all hydrogen bond lengths and it is more likely to find charged ones when \( R \) is larger than 2.35 Å.

Figure 4.3: Chemical features of amino acid-ligand SHBs. (a) Occurrence of the side chain of 11 proteinogenic amino acids and 11 most commonly observed functional groups in ligands in SHBs. (b) Distribution of charged and neutral amino acid-ligand SHBs at different hydrogen bond lengths.

As amino acid-ligand SHBs often appear in the active-site of proteins, the sequence of the active-site residues might encode the significant features to form a SHB. We find some common sequences for the amino acid-ligand SHBs. For example, a SHB network between one carbohydrate and two aspartates in the bacterial lectin PA-IIL of *Pseudomonas aeruginosa* is found to facilitate protein-ligand binding.\(^{213-217}\) A conserved calcium and monosaccharide binding loop with the sequence 96-104 and two calcium cations contributes to the high affinity binding to the sugar. As shown in Figure 4.4a, the residues Asp96 and Asp99 form two SHBs to the hydroxyl groups in \( \alpha \)-L-fucopyranose, in which the O-O distance are 2.62 Å and 2.55 Å, respectively.\(^{217}\) Another example is the SHB between the protease inhibitor drugs (Indinavir, Darunavir, Saquinavir, Ritonavir and Amprenavir) and the catalytic Asp25 in human immunodeficiency virus 1 (HIV-1) proteases.\(^{90,218-224}\) The catalytic
triplets Asp25-Thr26-Gly27 from both subunits are found to contribute to the formation of the enzyme active site. As shown in Figure 4.4b, there is an asymmetrical hydrogen bond network between the hydroxyl group of the drugs and the side chains of the catalytic Asp25 and Asp25’, in which the four O-O distance are 2.56 Å, 2.95 Å, 2.72 Å and 3.12 Å, respectively.90 In addition, there are SHBs between heme and Asp70 and Lys125 in the nitric oxide (NO)-transport protein nitrophorin 4 (NP4) from the blood-sucking insect Rhodnius prolixus.209,210,225–228 The hydrogen bond network at the proximal side of NP4 includes the His59, Phe68, Asp70, Lys125 residues and one water molecule. The Asp70 side chain is buried and has an altered \( pK_a \), which plays an important role in restraining His59 through water-mediated hydrogen bonds.209 As shown in Figure 4.4c, the residues Asp70 and Lys125 form two SHBs with one of the heme propionate, for which the O-O distances are both 2.50 Å.

4.3.3 Machine learning model to predict amino acid-ligand SHBs in proteins

From the 1070 protein-ligand complex structures, we have collected a total of 1272 SHBs and 2733 NHBs that form between the side chain of an amino acid and a ligand, resulting in
a probability of 31.8% to observe a SHB. They account for over 90% of the SHBs formed between amino acids and ligands. The criteria of $2.3 \, \text{Å} \leq R \leq 2.7 \, \text{Å}$ for a SHB and $2.8 \, \text{Å} \leq R \leq 3.2 \, \text{Å}$ for a NHB are used to distinguish one from the other. We use 16 variables to represent the structural and chemical features of both the amino acids and ligands. These include the charge of the amino acid and ligand, the residue type, heteroatom, secondary structure and the sequence that are within 3 residues before and after the hydrogen bonded amino acid, and the functional group, position (donor or acceptor), $pK_a$, $pK_b$ and logP of the ligand. We collaborate with Professor Sijian Wang and Yuanhao Liu to develop a machine learning model to predict the amino acid-ligand SHBs in proteins. Similar to the MAPSHB model in chapter 3, they use a boosting method and the undersampling strategy to construct the model. In order to evaluate the performance of the model, they randomly split the whole dataset into training and test datasets with a split ratio of 80:20 for 10 times and average over the results of 10 splits. A probability threshold of 0.5 is applied to predict a hydrogen bond to be a SHB or a NHB. From Table 4.2, the model prediction achieves a precision of 79% and a recall of 91%. We are developing a web server that allows one to upload a protein-ligand complex structure and obtain the probability of each amino acid-ligand hydrogen bond as a SHB.

Table 4.2: The precision and recall of the machine learning model with a probability threshold of 0.5 for 10 splits of training and test datasets.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>79%</td>
<td>82%</td>
<td>74%</td>
<td>2%</td>
</tr>
<tr>
<td>Recall</td>
<td>91%</td>
<td>93%</td>
<td>87%</td>
<td>2%</td>
</tr>
</tbody>
</table>

To discover the key parameters that promote the formation of amino acid-ligand SHBs, we have calculated the relative importance of the 16 input features. As shown in Figure 4.5, the residue type of the amino acids and the functional group of the ligands play a major role in determining the class of an amino acid-ligand hydrogen bond, as they possess the highest importance score of 20.0% and 13.0%, respectively. The amino acids can be divided into three types that differ in the probability to form SHBs. The first type includes
Tyr, Asp and Glu, in which Tyr has 74% probability to form a SHB with its phenol side chain and Asp and Glu have the probabilities of 71% and 64%, respectively, to form a SHB when their carboxylate side chain are the acceptor groups. The second type includes Ser, Thr and His, where the probability becomes 45%, 50% and 41%, respectively. The third type includes Arg, Lys, Asn, Gln and Trp, whose probability to form amino acid-ligand SHBs is below 16%. The functional groups of ligands can also be divided into four groups. The first group is the phenol group as it has a high probability of 73% to form a SHB. The second group is the alkyl hydroxyl group, with the probability decreasing to 51%. Although its probability is lower that the first group, the alkyl hydroxyl groups contribute to the largest population of amino acid-ligand hydrogen bonds. The third group includes the anion and acid forms of carboxylate, phosphate and sulphate groups and carbonyl groups of ketones and aldehydes, as the probability ranges from 22% to 40%. The fourth group includes all the other functional groups as their probability to form a SHB is below 14%. In the second group of amino acids and the second and the third groups of ligand functional groups, the model cannot determine the class of a hydrogen bond solely from these two features, and it is necessary to consider other features like the atom type of amino acids and the position of the ligands. For example, when the ligand contains an alkyl hydroxyl group, the probability of observing a SHB is 78.5% if it serves as the donor, while the probability is only 11.1% when it is the acceptor. In addition, the sequences within 3 residues before and after the hydrogen bonded amino acids each contribute \(\sim 10\%\) to the prediction of SHBs. For example, we can observe that many SHBs contain the sequences of “Gly-Ser-Glu-Asp-Gly-Thr-Asp” and “Asp-Gly-Thr-Asp-Asn-Asp-Tyr” in our dataset, and they partially coincide with the active-site sequence in the conserved calcium and monosaccharide binding loop of lectin.\(^{213–217}\) Finally, the other input features are less important and only contribute \(\leq 2\%\) to the overall prediction (Figure 4.5).
Figure 4.5: Normalized importance scores for the 16 input features of the machine learning model. The features that contribute $\leq 2\%$ to the predictions are grouped as the “Other” type. These include the charges of amino acid and ligand, the atom type and secondary structure of the amino acid and the location, $pK_a$, $pK_b$ and logP of the ligand.

4.3.4 Quantum chemistry calculations of the amino acid-ligand SHBs

Figure 4.5 suggests the essential parameters for the formation of SHBs between amino acids and ligands. We collect a total of 1758 hydrogen bond pairs from the top predictions of the machine learning model, which has $> 99\%$ probability to form SHBs. 53\% of them contain the alkyl hydroxyl group of the ligands as the hydrogen bond donor and the side chain of Asp or Glu as the acceptor. Although the other pairs occur much less frequently, we observe that some amino acids and ligand functional groups are favored in the formation of SHBs. For example, while only 11\% of the top predictions have a phosphate group in the ligands, 58\% of these cases contain the side chain of Thr or Ser as the hydrogen bond donor or acceptor. We list the probability to form SHBs for different combinations of amino acids and ligand functional groups in Table 4.3. Here we only list the cases that are observed in at least 20 SHBs to remove the rare cases of the amino acid-ligand pairs. From Table 4.3, we find that the alkyl hydroxyl-Asp/Glu, phosphate-Thr and carboxylic acid-Tyr/His pairs are particularly interesting, as they are commonly found in proteins and have
over 85% probability to form SHBs.

Table 4.3: Occurrence of the side chain of amino acids and functional group of ligands in hydrogen bonds when the number of SHBs is larger than 20 from our dataset. “D” and “A” denote that the ligand serves as the donor or acceptor, respectively, in a hydrogen bond. The functional groups are denoted using chemical formula: hydroxyl (-OH), phosphate (-H$_n$PO$_n$$^−_2$) and carboxylic acid (-COOH and -COO$^-$H) with the underlined atom representing the specific atom in the hydrogen bonds.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Ligand</th>
<th>Ligand location</th>
<th>n(SHBs)</th>
<th>n(NHBs)</th>
<th>$P_{SHB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASP</td>
<td>-OH</td>
<td>D</td>
<td>328</td>
<td>47</td>
<td>87.47%</td>
</tr>
<tr>
<td>GLU</td>
<td>-OH</td>
<td>D</td>
<td>202</td>
<td>35</td>
<td>85.23%</td>
</tr>
<tr>
<td>SER</td>
<td>-H$_n$PO$_n$$^−_2$</td>
<td>A</td>
<td>37</td>
<td>28</td>
<td>56.92%</td>
</tr>
<tr>
<td>THR</td>
<td>-H$_n$PO$_n$$^−_2$</td>
<td>D</td>
<td>29</td>
<td>0</td>
<td>100.00%</td>
</tr>
<tr>
<td>ASP</td>
<td>-COOH</td>
<td>D</td>
<td>28</td>
<td>12</td>
<td>70.00%</td>
</tr>
<tr>
<td>HIS</td>
<td>-COOH</td>
<td>A</td>
<td>27</td>
<td>10</td>
<td>72.97%</td>
</tr>
<tr>
<td>TYR</td>
<td>-OH</td>
<td>A</td>
<td>24</td>
<td>6</td>
<td>80.00%</td>
</tr>
<tr>
<td>TYR</td>
<td>-COOH</td>
<td>A</td>
<td>24</td>
<td>2</td>
<td>92.31%</td>
</tr>
<tr>
<td>THR</td>
<td>-H$_n$PO$_n$$^−_2$</td>
<td>A</td>
<td>21</td>
<td>2</td>
<td>91.30%</td>
</tr>
<tr>
<td>SER</td>
<td>-COOH</td>
<td>A</td>
<td>21</td>
<td>5</td>
<td>80.77%</td>
</tr>
<tr>
<td>HIS</td>
<td>-COOH</td>
<td>A</td>
<td>21</td>
<td>3</td>
<td>87.50%</td>
</tr>
<tr>
<td>LYS</td>
<td>-H$_n$PO$_n$$^−_2$</td>
<td>A</td>
<td>20</td>
<td>78</td>
<td>20.41%</td>
</tr>
</tbody>
</table>

From Table 4.3, the SHBs between the alkyl hydroxyl group of ligands and the side chain of Asp or Glu constitute the largest portion (42%) of all amino acid-ligand SHBs, and their probability to form SHBs is as high as 87%. We find that 41% of them have the side chain of Asp or Glu as the hydrogen bond acceptor and the alkyl hydroxyl group of carbohydrates as the donor. An example is shown in Figure 4.4a, in the active site of lectin PA-IIL, the carbohydrate is capable of forming two SHBs with the aspartate residues in the conserved calcium and monosaccharide binding loop.\textsuperscript{217} We further find that 57% of the carbohydrate-Asp SHBs in our dataset involve the interaction between one carbohydrate and two aspartates, suggesting its structural importance in forming SHBs. We will take Asp96, Asp99 and $\alpha$-L-fucopyranose (Fuc) from lectin PA-IIL (PDB ID 1UZV) (Figure 4.6a) and perform electronic structure calculations to evaluate how the interaction energies with and without the SHB network vary with R.\textsuperscript{217} As shown in Figure 4.6b, the minima of the interaction energy of the Fuc-Asp96 and Fuc-Asp99 pairs occur at an O–O
distance, \( R_{\text{min}} \), of 2.65 Å and 2.75 Å, respectively. In the presence of the hydrogen bond network, we find a shortening of \( R_{\text{min}} \) for both pairs: the \( R_{\text{min}} \) values become 2.60 Å and 2.65 Å, respectively. Consistent with the model predictions, the \( R_{\text{min}} \) values of both pairs are within the range of SHBs, indicating that the formation of the hydrogen bond network makes it favorable for the Asp residues and the Fuc ligand to stay in close proximity. Besides the changing of \( R_{\text{min}} \), we find that the hydrogen bond energies of both pairs decrease by 5.5-7.6 kcal/mol in the SHB regions and increase by 2.2-14.8 kcal/mol when \( R \geq 3.1 \) Å in the presence of the hydrogen bond network. In particular, a SHB network can lower the hydrogen bond energies by 6.2 kcal/mol and 7.6 kcal/mol for the Fuc-Asp96 and Fuc-Asp99 SHBs when \( R \) is 2.6 Å, demonstrating the energetic benefits of forming a network of SHBs.

![Figure 4.6](image_url)

**Figure 4.6:** (a) Hydrogen bond network composes of Asp96, Asp99 and Fuc in lectin PA-II (PDB ID 1UZV). (b) Interaction energies of the Fuc-Asp96 (red) and Fuc-Asp99 (blue) hydrogen bonds with (solid line) and without (dashed line) the SHB network.

We are still in the process of evaluating the phosphate-Thr and carboxylic acid-Tyr/His pairs to identify the origin of their preference in forming SHBs (Table 4.3). In particular, 82% of the phosphate-Thr SHBs occur in the binding of NADP\(^+\) in proteins, and 63% of
the carboxylic/carboxylate-Tyr SHBs occur in protein binding to small molecule acids. These electronic structure calculations will possibly provide insights into their biological functions.

### 4.4 Conclusions

In this chapter, we have conducted a statistical analysis of amino acid-ligand SHBs in the protein-ligand complexes from the atomic-resolution structures in the PDB. Structurally, SHBs formed between the side chains of amino acids and ligands usually have shorter $R$ than the ones between two amino acids. Chemically, ligands such as carbohydrates and nucleotides are frequently observed to form SHBs with amino acids from their alkyl hydroxyl, carboxyl, carboxylate and phosphate functional groups. From the amino acid side, the negatively charged side chains of Asp and Glu are most frequently observed as the acceptor, while the side chains of Tyr, Ser, Thr and His are observed as donor or acceptor in the hydrogen bonds. As amino acid-ligand SHBs are often found in the active site of proteins, the protein sequences in the active-site region can be important to promote the formation of these SHBs.

In collaboration with Professor Sijian Wang and Yuanhao Liu, we have constructed a machine model to predict SHBs between side chains of amino acids and ligands. Using a probability threshold of 0.5, our model successfully gives a precision of 79% and a recall of 91% in the prediction. Our model will facilitate the experimental and computational refinement of protein-ligand complex structures, and the design of novel proteins and bio-inspired materials that incorporate these compact structural elements to achieve enhanced functions. From our model, we find that the type of amino acids and the functional groups of ligands play major roles in determining whether a hydrogen bond is a SHB or a NHB. In particular, the alkyl hydroxyl-Asp/Glu pairs exhibit a strong preference to form SHBs. Besides, the sequences of amino acids are also important to the formation of SHBs. Using electronic structure calculations, we demonstrate how a network of SHBs promotes the for-
mation of alkyl hydroxyl-Asp/Glu SHBs in carbohydrate-binding proteins. We will carry out more electronic structure calculations to further explain the formation of phosphate-Thr and carboxylic acid-Tyr/His SHBs in protein-ligand complexes.
CHAPTER 5
SYMMETRY AND PROTON NMR CHEMICAL SHIFTS OF SHORT HYDROGEN BONDS

This chapter is reprinted with permission from Ref. [P2]

5.1 Introduction

One of the most distinctive features of SHBs is their far downfield $^1$H chemical shifts, $\delta_H$, in the nuclear NMR spectra. However, dissecting the geometry and $^1$H chemical shifts of a SHB is highly challenging due to the experimental difficulty to probe specific protons in a large biomolecule. To tackle this problem, researchers have examined a series of organic and inorganic small molecules that contain a single SHB to unravel the common properties of these compact structures. A famous example is the compound 1,8-bis(dimethylamino)naphthalene (DMAN), also known by its trade name Proton Sponge.$^{72,73}$ DMAN is an exceptionally strong base with $pK_a = 12.3$ and has a weak nucleophilic character due to its highly strained structure.$^{72,73}$ Protonation releases this steric strain and the resulting cation, DMANH, contains a stable intramolecular N–H···N hydrogen bond with $R$ of 2.55–2.63 Å and $\delta_H > 17$ ppm.$^{75,76}$ Other examples include the salts of hydrogen maleate (HM)$^{21,64,231}$ and 4,5-dihydroxynaphthalene-2,7-disulfonate (DHND),$^{53,232}$ which serve as model systems for the O–H···O type SHBs, and cis-urocanic acid (CUA), which has been designed to mimic the Asp–His SHB in serine proteases.$^{21}$

The chemical structures of the 4 model molecules are depicted in Figure 5.1. By measuring the X-ray diffraction patterns and NMR spectra of these small molecules, researchers have identified a strong correlation between $\delta_H$ and $R$ and utilized this relation to obtain the hydrogen bond lengths in proteins.$^{22,48,55,232-234}$ However, as the lengths of the O–H or N–H bonds are only available for a few systems, how $\delta_H$ depends on the proton positions in the
SHBs has not been established. While electronic structure calculations have provided crucial insights into this problem, they are carried out on the optimized configurations of the small molecules or proteins and hence neglecting the critical influences from structural fluctuations and nuclear quantum effects.

Figure 5.1: Chemical structures of the model molecules that contain intramolecular SHBs.

In this chapter, we consider DMANH, HM, DHND and CUA in organic solvents and aqueous solutions and utilize AIMD and AI-PIMD simulations to investigate how quantum effects determine the symmetry and chemical shifts of their SHBs. As shown in Figure 5.1, these molecules form stable intramolecular hydrogen bonds with $R < 2.6$ Å and exhibit large downfield $^1$H chemical shifts, providing excellent models for N–H· · ·N, O–H· · ·O and O–H· · ·N SHBs in proteins. We perform both AIMD and AI-PIMD simulations of these solutions and directly compare the trajectories from them to unravel how the interplay of electronic and nuclear quantum effects impacts the structure, proton sharing conditions and $^1$H chemical shifts of the SHBs. By extracting representative configurations from the simulations, we further establish a relation between $\delta_H$ and the position of the proton and elucidate the origin of the highly downfield chemical shifts for the SHBs.
5.2 Methods

We obtain the structures of the 4 model molecules from the Cambridge Structural Database\textsuperscript{94} and solvat them using the Amber 2016 software package\textsuperscript{120} to form the following systems: DMANH in acetonitrile and in water, HM in acetone and in water, DHND in water and CUA in dimethyl sulfoxide (DMSO). The solvation is carried out by placing 1 DMANH, HM and CUA in 86 acetonitrile, 62 acetone and 66 DMSO molecules, respectively. The aqueous solutions of DMANH, HM and DHND contain 1 solute molecule in 151, 185 and 160 waters, respectively. As DMANH, HM and DHND possess net charges, sodium or chloride ions are added to neutralize the system. For each solvated structure, the simulations are performed in a cubic box with periodic boundary conditions, and the box length varies between 17 and 25 Å to ensure that the system have the correct solvent density at the temperature of the corresponding experimental NMR measurements. The simulation temperatures are 300 K for DMANH in acetonitrile and CUA in DMSO, 298 K for DMANH in water, 268 K for DHND in water, and 308 K for HM in acetone and water.\textsuperscript{21,52,53,231,238–240}

As a first step, we equilibrate the systems with classical molecular dynamics (MD) simulations using the Amber 2016 package.\textsuperscript{120} The potential energies of the solute and organic solvent are modeled using the general AMBER force field,\textsuperscript{123} the water molecules are described with the TIP3P model,\textsuperscript{241} and the ions are modeled using Joung/Cheatham ion parameters.\textsuperscript{242} The SHAKE algorithm is used to constrain all bonds that involved hydrogen atoms,\textsuperscript{243} and the particle-mesh Ewald method is implemented to treat long-range electrostatic interactions.\textsuperscript{244} After energy minimization, each system is equilibrated with a constant pressure at 1 bar and a constant temperature using the Langevin thermostat\textsuperscript{245} and Berendsen barostat.\textsuperscript{246} The equilibration has been carried out for 4 ns with a time step of 2 fs. We take the final configuration of each system from the MD simulations and use them for the first principles simulations.

AIMD and AI-PIMD simulations are performed using the Quickstep module in the
CP2K package\textsuperscript{247,248} for the on-the-fly evaluation of the electronic structures and the i-PI software\textsuperscript{249} for the propagation of the nuclear motion. For each system, the electronic structures are described using the BLYP density functional\textsuperscript{250,251} with the D3 dispersion correction.\textsuperscript{128} The Goedecker-Teter-Hutter pseudopotentials\textsuperscript{252} are chosen to describe the core electrons and the DZVP-GTH plane-wave basis set with a cutoff of 300 Ry is used for the valence charge density. The simulations are performed in the canonical ensemble at a time step of 0.5 fs. In AIMD simulations, the stochastic velocity rescaling method\textsuperscript{253} is applied to keep the temperature at the desired values. In Al-PIMD simulations, each atom is represented by 6 ring polymer beads using the path integral generalized Langevin equation method.\textsuperscript{254} Each system is equilibrated for 10 ps, followed by AIMD simulations for 50 ps or Al-PIMD simulations for 25 ps.

From the AIMD and Al-PIMD simulations of DMANH, HM and DHND in aqueous solutions, we analyze the hydrogen bonds between the solute molecules and water. We consider a pair of A–H⋯B to be hydrogen bonded if its R < 3.5 Å and the A–H–B angle $\theta_{AHB} > 135^\circ$. For all systems, we also calculate the average residence time of a proton in a SHB. Here we define the residence time as the duration that the proton is closer to the donor or acceptor side before hopping happened, and the report values are averaged over the donor and acceptor atoms and over the simulation trajectories.

To compute the $^1$H NMR chemical shifts of the model molecules, we extract configurations of the solute-solvent clusters every 100 fs from the AIMD simulations, and every 50 fs from the Al-PIMD simulations by randomly selecting one ring polymer bead. We include all solvent molecules that are within 5.5 Å of any atom of the solute and the resulting clusters contained 143–330 atoms. We then compute the $^1$H NMR chemical shift of the solute-solvent clusters using the Gauge-Independent Atomic Orbital (GIAO) method,\textsuperscript{255–257} as implemented in the Gaussian 16 software package.\textsuperscript{258} The calculations are carried out using the B3LYP density functional,\textsuperscript{127} the D3 dispersion correction\textsuperscript{128} and the 6-31+G(d,p) basis set. The $\delta_H$ value of tetramethylsilane (TMS) at the same level of theory is 31.6
ppm and is subtracted from the value of the molecules to produce the chemical shifts. To elucidate the electronic quantum effects in the SHBs and how they lead to the downfield chemical shifts, we carry out the Natural Bond Orbital (NBO) analysis of DMANH with the NBO 6.0 software. Here we use the optimized structure of DMANH and scan the lengths of the A–H or B–H bonds in the gas phase. For each scanned configuration, we optimize the position of all the H atoms while maintaining the positions of the heavy atoms, and then conduct NMR and NBO calculations. To decompose the total chemical shift into contributions from covalent bonds and hydrogen bonds, we repeat the NMR calculations on the compound N,N-dimethylaniline, which is chosen to model the donor and acceptor groups in DMANH. For this purpose, we remove the atoms from the other half of the DMANH molecule and add two H atoms to saturate the bonds of the C atoms in N,N-dimethylaniline. The coordinates of these H atoms are optimized while the positions of all the other atoms are fixed. We then take the same proton positions in the hydrogen bond of DMANH in the scan process and carry out NMR calculations. The overall shielding constant from covalent bonding is computed as the sum of the chemical shifts from the donor and acceptor groups. The shielding constant of hydrogen bonding is calculated as the difference between the total shielding and that from covalent bonding.

5.3 Results and discussion

5.3.1 Geometry and symmetry of the intramolecular SHBs

In a symmetric hydrogen bond, the proton is centered between the donor and acceptor groups. It has been proposed that the formation of a symmetric hydrogen bond requires 1) a short distance between the hydrogen bond partners (R ≤ 2.5 Å), 2) a matched proton affinity of the donor and acceptor atoms, and 3) a non-aqueous environment. By examining the 4 model molecules, we will elucidate how these criteria arise from the quantum nature of the hydrogen bonds.

As a first step, we evaluate the impact of R on the geometry and symmetry of the
SHBs and use a coordinate \( \nu \) to characterize the proton positions in these compounds. For DMANH, HM and DHND, we define \( \nu = d_{X_1H} - d_{X_2H} \), where X is N or O and \( d_{X_1H} \) and \( d_{X_2H} \) are the distances from the hydrogen atom to the two equivalent heavy atoms in the SHBs. For CUA, we take this coordinate as \( \nu = d_{OH} - d_{NH} \). From its definition, \( \nu = 0 \) means that the proton equally bridges the donor and acceptor groups, leading to a symmetric hydrogen bond, whereas a negative or positive \( \nu \) shows that the proton is closer to the donor or acceptor groups, respectively. From the AIMD and AI-PIMD simulations, we compute the joint probability for finding the hydrogen bond at length \( R \) and the proton at position \( \nu \), \( P(R, \nu) \). At a given temperature \( T \), the free energies are calculated as

\[
F = -k_B T \ln \frac{P(R, \nu)}{P_{\max}},
\]

where \( k_B \) is the Boltzmann constant. \( P_{\max} \) is the probability of the corresponding system at the most likely \( R \) and \( \nu \), and is included to ensure that the minimal free energy is 0. At thermal equilibrium, the molecules dynamically switch between different conformations and form an ensemble of hydrogen bonding geometries. When these interconversions happen faster than the time scale of the NMR measurements, one would observe a symmetric hydrogen bond if the free energy surface is symmetric and the average \( \nu \) is 0.

As Figure 5.2a demonstrates, DMANH forms a relatively symmetric \( N_1-H \cdots N_2 \) hydrogen bond in acentonitrile from AIMD simulations. \( R \) of the hydrogen bond fluctuates between 2.4 and 2.9 Å and the energetically most favorable configuration has \( R = 2.7 \) Å and \( \nu = -0.5 \) Å. The second minimum occurs symmetrically at \( R = 2.7 \) Å and \( \nu = +0.5 \) Å, which is only 0.2 kcal/mol higher in energy compared to the most stable hydrogen bonding conformation. In the AIMD simulations, we observe frequent hopping of the proton in the SHB, consistent with the observations in previous Car-Parinello molecular dynamics simulations of DMANH in other environment. However, there is a barrier of 2.4 kcal/mol connecting the 2 proton transferred configurations and hence the proton spends more time...
around the N\(_1\) or N\(_2\) atoms, rather than the middle of the SHB. From Figure 5.2a, the absolute values of \(\nu\) decrease almost linearly when R shortens from 2.9 Å to 2.4 Å, regardless of whether the proton is closer to N\(_1\) (\(\nu \leq 0\)) or N\(_2\) (\(\nu > 0\)). In other words, the hydrogen bond becomes more symmetric as the separation between N\(_1\) and N\(_2\) decreases. We have recently shown that electronic quantum effects give rise to this pronounced proton sharing in the SHBs, and calculations based on a classical force field can only provide a qualitative trend in lengthening the Donor–H bond.\(^6\)

Figure 5.2: Free energy profiles for DMANH and HM in organic solvents and aqueous solutions. In the top panels, the free energy surfaces are calculated from (a) AIMD and (b) AI-PIMD simulations of DMANH in acetonitrile, and from (c) AIMD and (d) AI-PIMD simulations of DMANH in water. In the bottom panels, the curves are calculated from (e) AIMD and (f) AI-PIMD simulations of HM in acetone, and from (g) AIMD and (h) AI-PIMD simulations of HM in water.

From Figure 5.2b, inclusion of nuclear quantum effects significantly increases the region that the proton can move and washes out the double-well feature of the free energy profile. From AI-PIMD simulations, the most stable configuration of the SHB in DMANH has R of 2.6 Å and \(\nu\) of 0.1 Å, where the proton resides nearly in the middle of the SHB. While shortening R again facilitates proton sharing in Figure 5.2b, nuclear quantum effects allow the proton to be delocalized between the N\(_1\) and N\(_2\) atoms as it takes less than
0.7 kcal/mol to move the proton between \( \nu \) of -0.5 and +0.5 Å. In the mean time, nuclear quantum effects also make the \( \text{N}_1-\text{H}-\text{N}_2 \) angle vary over a wide range of 114–180° in AI-PIMD simulations, as compared to 131–176° in AIMD simulations. This demonstrates the well-known competition of nuclear quantum effects in facilitating the \( \text{N}–\text{H} \) stretch motion, which strengthens the hydrogen bond, and enhancing the bending and rotational of the bonds, which weaken the hydrogen bonds. For DMANH, the strengthening effect dominates as the average residence time of the proton on the two N atoms decreases from 0.15 ps in AIMD simulations to 0.04 ps in AI-PIMD simulations. The average residence time is inversely related to the proton exchange rate in the SHB, and the values from both simulations are much shorter than the millisecond time scale in a typical NMR experiment. Considering that the free energy surfaces in Figs. Figure 5.2a and b have symmetric features and that the average \( \nu \) is -0.05 and 0.01 Å when the nuclei are treated classically and quantum mechanically, respectively, both simulations predict that the SHB in DMANH is considerably symmetric in acetonitrile.

The chemical structure of DMANH has reflection symmetry about a plane that bisects the naphthalene ring (Figure 5.1). Therefore, it is not surprising that in the organic solvent acetonitrile, the donor and acceptor atoms of the \( \text{N}–\text{H} \cdot \cdot \cdot \text{N} \) hydrogen bond have identical proton affinities and hence their \( pK_a \) mismatch, \( \Delta pK_a \), is 0. One can perturb this \( pK_a \) matching condition by putting DMANH in aqueous solution. As shown in Figs. Figure 5.2c and d, water molecules induce slight asymmetry in the distributions of the proton position. Comparing DMANH in acetonitrile and water, the solvent molecules do not form direct interactions with the \( \text{N}_1 \) and \( \text{N}_2 \) atoms due to the steric hindrance of the bulky methyl groups and the \( \text{Cl}^- \) counterions are 6–8 Å away from the N atoms, leading to minor influence on the proton positions. However, unlike acetonitrile, water forms slightly different solvation shells around the methyl groups of DMANH. From both the AIMD and AI-PIMD simulations, we find that the first solvation shells occur at a distance of 4.9 Å between the C atom in the methyl group and the O atom in water, and there are on average 15 and 16
water molecules in the first shells of the methyl groups surrounding the N1 and N2 atoms, respectively, leading to the slight asymmetry in the free energy surfaces.

Similar to DMANH, HM possesses reflection symmetry in its structure (Figure 5.1) and its O1–H⋯O2 hydrogen bond is highly symmetric in the organic solvent acetone. As shown in Figure 5.2e, accompanying the structural fluctuations in the AIMD simulations, electronic quantum effects make the proton more shared in the SHB when R decreases from 2.8 to 2.2 Å. The free energy surface exhibits two minima at R = 2.5 Å and ν = ±0.2 Å. As the average R in HM is shorter than that in DMANH, the proton transfer barrier is only 0.4 kcal/mol, and as such, the zero-point energy associated with the O–H vibrations facilitates the quantum delocalization of the proton in the hydrogen bond. This is reflected in the free energy profiles in Figure 5.2f, which has a single minimum with an average ν of 0.

In contrast, we observe prominent transient asymmetry in the SHB of HM in water, as demonstrated in Figs. Figure 5.2g and h. From AIMD simulations of 50 ps, the most favorable hydrogen bonding conformation has R = 2.5 Å and ν = 0.4 Å and the average residence time of the proton on O1 or O2 is 0.09 ps. In the 25-ps AI-PIMD simulations, the interplay of nuclear and electronic quantum effects moves the minimum to ν = 0.2 Å and reduces the average residence time by 56%. In both cases, the most likely position of the proton is around the O2 atom, rather than at the center of the hydrogen bond. Comparing HM in acetone and water (Figs. Figure 5.2f and h), we expect the changes in the proton sharing conditions to arise from the distinct solute-solvent interactions. After examining the minimal-energy configuration of HM in aqueous solutions, we find that the O1 and O2 atoms have different solvation environment and form an average of 1.1 and 0.7 hydrogen bonds with the surrounding water molecules, respectively. 26% of these conformers accept 2 hydrogen bonds from water, 1 for each O atom, with a representative snapshot shown in Figure 5.3a. 20% of them form only 1 hydrogen bond between O1 and the solvent, as depicted in Figure 5.3b. Apart from these commonly observed conformations, the SHB in HM can also take other solvation structures with the O1 and O2 atoms forming up to 3 hy-
drogen bonds with the water molecules. Therefore, although the $\Delta pK_a$ in the SHB of HM is 0 in a nonpolar environment, the heterogeneous hydrogen bonding patterns disturb its $pK_a$ matching condition and introduce asymmetry in its free energy surface. In particular, hydrogen bonding with water effectively reduces the proton affinity of the O atom, as our results and previous simulations both show that the proton is more likely to stay closer to the O that is less solvated.\textsuperscript{263} It is important to note that the asymmetry in Figs. Figure 5.2g and h is transient in nature as it reflects the instantaneous solvation environment around the O\textsubscript{1} and O\textsubscript{2} atoms. Although our simulations are not sufficiently long to fully sample the rearrangements of the water molecules, we expect the free energy surfaces to become symmetric in the time scale of the NMR measurements.

![Figure 5.3: Representative configurations of HM with the hydrogen bonded water molecules. The conformations have R of 2.5 Å and $\nu$ of 0.2 Å. Silver, red and white represent C, O and H, respectively. The orange spheres are the ring polymer beads of the hydrogen bonded proton in HM.](image)

Similar to HM, the O\textsubscript{1}–H···O\textsubscript{2} hydrogen bond of DHND exhibit instantaneous asymmetry in water. From AIMD simulations, the O\textsubscript{1} and O\textsubscript{2} atoms form an average of 2.0 and 1.6 hydrogen bonds with the solvent molecules, respectively, which effectively introduces a nonzero $\Delta pK_a$ in its SHB. This is reflected in its free energy surface in Figure 5.4a, which has a global minimum at R = 2.5 Å and $\nu = 0.5$ Å with the proton closer to the O\textsubscript{2} atom. A second minimum that is 0.4 kcal/mol higher in energy occurs at $\nu$ of -0.5 Å. Although the average R of the intramolecular hydrogen bond is only 2.5 Å, DHND has a large barrier of
3.7 kcal/mol between the two minima and the average residence time of the proton is 0.56 ps. As shown in Figure 5.4b, nuclear quantum effects enhance the transient asymmetry of the free energy surface. From AI-PIMD simulations, the ratio of solute-solvent hydrogen bonds for the O₁ and O₂ atom become 2:1 and accordingly, the proton spends 82% of the time closer to the less solvated O₂ atom ($\nu \geq 0$). In the mean time, nuclear quantum effects also promote proton sharing in the SHB as it takes only 1.2 kcal/mol to move the proton between -0.5 and +0.5 Å and the average residence time of the proton decreases to 0.11 ps.

To further evaluate how $pK_a$ mismatch influences the properties of a SHB, we consider CUA, which possesses an intrinsic $\Delta pK_a$ of 3.7 between the imidazole and carboxylic groups,²³⁹,²⁶⁴ and find that it has an asymmetric hydrogen bond. From AIMD simulations, its minimal energy configuration occurs at $R = 2.6$ Å and $\nu = -0.6$ Å and electronic quantum effects give rise to a strong inverse correlation between $\nu$ and $R$, as demonstrated in Figure 5.4c. The proton is almost entirely attached to the O atom ($\nu < 0$) in the simulations, resulting in a long residence time of 1.16 ps. As shown in Figure 5.4d, the proton can sample a much broader region in the hydrogen bond and the free energy minimum shifts to $\nu = -0.5$ Å in AI-PIMD simulations. Upon the addition of nuclear quantum effects, the proton becomes more shared in the SHB with a 13% probability of being transferred to the acceptor N atom ($\nu \geq 0$) and the average residence time is reduced to 0.09 ps, an order of magnitude shorter than that from the AIMD simulations.

From the first principles simulations, we now assess the proposed criteria for the for-
mation of symmetric SHBs. First, electronic quantum effects facilitate proton sharing in a hydrogen bond when the distance between the donor and acceptor atoms decreases, as demonstrated by the free energy surfaces from the AIMD simulations. In the model molecules, the SHBs have $R \leq 2.7$ Å and hence there are considerable probability of finding the proton around $\nu = 0$. Due to the short $R$, we observe that the protons frequently hop between the donor and acceptor groups and their average residence time is less than 1 ps, much shorter than the typical NMR time scale. As a result, what is measured from the NMR experiments is the ensemble average of the proton positions. Second, the $pK_a$ matched condition, which depends on the intrinsic proton affinity of the donor and acceptor groups and the instantaneous solvation environment, is required for a symmetric hydrogen bond. For example, as DMANH, HM and DHND have reflection symmetry in their chemical structures, their $\Delta pK_a$ in the SHBs are 0 with no external perturbations. When these molecules are placed in aqueous solutions, the solute-solvent hydrogen bonds significantly modulate the $\Delta pK_a$ value of the SHBs and the protons are preferentially attached to the atom that is less solvated instantly, leading to transient asymmetry in their proton distributions. In comparison, the SHB in CUA has an intrinsic $\Delta pK_a$ of 3.7 and a prominently asymmetric SHB, in which the proton spends 87% of the time around the donor O atom. Finally, AI-PIMD simulations closely mimic the experimental conditions as they include both electronic and nuclear quantum effects, and the interplay of these quantum effects plays a fundamental role in determining the symmetry of a SHB. From Figs. 5.2 and Figure 5.4, nuclear quantum effects significantly enhance the region that the protons can move and shorten their average residence time on the N or O atoms by 60–92%. Notably, the influences of nuclear quantum effects are comparable to, if not greater than, those from $R$ and $\Delta pK_a$ on the free energy surfaces and the proton behavior. Instead of experiencing electrostatic attraction by the donor and acceptor groups, the protons are quantum mechanically delocalized in the SHBs and the symmetry of the hydrogen bonds are enhanced.
5.3.2 $^1$H NMR chemical shifts of the SHBs

From the AIMD and AI-PIMD simulations, we extract 6012 solute-solvent clusters of the model molecules and calculate their $\delta_H$ values using the B3LYP density functional, the D3 dispersion correction and the 6-31+G(d,p) basis set. As shown in Table 5.1, the average chemical shifts, $\langle \delta_H \rangle$, from the simulations are in good agreement with the experimental measurements. Considering that AI-PIMD simulations closely resemble the experimental conditions, the excellent comparison between their predicted chemical shifts and the experimental values validates our approach of combining these simulations and DFT calculations to obtain $\delta_H$. From Table 5.1, we find that the chemical shifts from our calculations and experiments are all highly downfield (> 16 ppm). In addition, the $\langle \delta_H \rangle$ values from the AI-PIMD simulations are systematically larger than those from the AIMD simulations by 0.3–2.1 ppm.

Table 5.1: $\nu_0$ and $\langle \delta_H \rangle$ from the AIMD and AI-PIMD simulations of the model molecules. The predicted and the experimental $^1$H NMR chemical shifts are included for comparison.

<table>
<thead>
<tr>
<th>Model molecule</th>
<th>Solvent</th>
<th>DMANH</th>
<th>HM</th>
<th>DHND</th>
<th>CUA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acetonitrile</td>
<td>Water</td>
<td>Acetone</td>
<td>Water</td>
<td>Water</td>
</tr>
<tr>
<td>$\nu_0,\text{AIMD}$ (Å)</td>
<td>0.42</td>
<td>0.42</td>
<td>0.27</td>
<td>0.34</td>
<td>0.51</td>
</tr>
<tr>
<td>$\nu_0,\text{AI−PIMD}$ (Å)</td>
<td>0.32</td>
<td>0.32</td>
<td>0.23</td>
<td>0.28</td>
<td>0.42</td>
</tr>
<tr>
<td>$\langle \delta_H \rangle,\text{AIMD}$ (ppm)</td>
<td>18.9</td>
<td>19.0</td>
<td>20.2</td>
<td>18.9</td>
<td>16.4</td>
</tr>
<tr>
<td>$\langle \delta_H \rangle,\text{AI−PIMD}$ (ppm)</td>
<td>20.1</td>
<td>20.2</td>
<td>20.3</td>
<td>19.6</td>
<td>17.5</td>
</tr>
<tr>
<td>$\langle \delta_H \rangle,\text{predict}$ (ppm)</td>
<td>18.9</td>
<td>18.9</td>
<td>19.6</td>
<td>19.2</td>
<td>17.7</td>
</tr>
<tr>
<td>$\delta_{H,\text{exp}}$ (ppm)</td>
<td>18.7\textsuperscript{238}</td>
<td>18.5\textsuperscript{52}</td>
<td>20.7\textsuperscript{231}</td>
<td>20.2\textsuperscript{240}</td>
<td>17.7\textsuperscript{53}</td>
</tr>
</tbody>
</table>

Pioneering first principles simulations have shown that one can use $\nu$ to effectively represent the electronic state of a hydrogen bond and link its structure to the chemical shielding on the protons.\textsuperscript{44,266} To uncover the molecular origin of the observed trends in Table 5.1, we use $\nu$ as a collective coordinate and decompose $\langle \delta_H \rangle$ into two components,

$$\langle \delta_H \rangle = \int_{-\infty}^{\infty} \delta_H(\nu) P(\nu) d\nu.$$  \hspace{1cm} (5.2)

For a model molecule, $\delta_H(\nu)$ describes how its chemical shift changes with the proton
position $\nu$ as its structure fluctuates, and $P(\nu)$ represents the probability distribution of $\nu$ in the SHBs. In the following, we will examine the two properties individually.

**Universal relation between the chemical shifts and the proton position**

As a first step, we compute the proton position and chemical shift of each solute-solvent configuration to evaluate $\delta_H(\nu)$ in Equation 5.2. As shown in Figure 5.5, when the proton moves from $\nu$ of $\pm 1.1$ Å to the center of the hydrogen bond ($\nu=0$), its chemical shift changes from a regular value of 6 ppm to the downfield 23 ppm, demonstrating the sensitivity of $\delta_H$ to the fluctuations of the proton positions. Furthermore, while the model molecules have distinct donor and acceptor groups and are in different solvents, their chemical shifts follow the same relation with $\nu$ in their SHBs. From a least squares fitting to the data, we obtain a quadratic function for $\delta_H(\nu)$:

$$\delta_H(\nu) = 21.9 - 16.1\nu^2.$$  \hspace{1cm} (5.3)

This observation of a universal trend is consistent with previous experimental and computational studies.\textsuperscript{44,236,237,266,267} In particular, Ceriotti and coworkers have used AI-PIMD simulations to show that the $^1$H NMR chemical shifts of water in 3 distinct thermodynamic state points fall on the same curve when $\nu$ is used to represent the proton position in the hydrogen bonds.\textsuperscript{44} Compared to their findings that $\delta_H$ follows an almost linear relation with the proton position when $\nu$ is between -1 and 0 Å, here we identify a strong non-linearity for systems containing SHBs.

The $^1$H NMR chemical shift describes the electronic shielding effect on a proton by the surrounding atoms, which arises from a competition between the covalent bonding and hydrogen bonding interactions in a SHB. We will take the $\text{N}_1\text{H} \cdots \text{N}_2$ hydrogen bond in DMANH as an example to elucidate how this competition leads to the trend in Figure 5.5. For this purpose, we rewrite the chemical shift as $\delta_H = S_{TMS} - S_H$, where $S_H$ is the
Figure 5.5: (a) Correlation plot for $\delta_H$ and $\nu$. Each data point is calculated from a solute-solvent cluster as sampled from the AIMD and AI-PIMD simulations of the model molecules in different solvents. (b) Correlation of $\delta_H$ and $\nu$ from the model molecules, KSI and previous computational studies.

isotropic magnetic shielding constant for the H atom and $S_{TMS}$ is the shielding constant in the reference compound TMS ($S_{TMS} = 31.6$ ppm from our DFT calculations). Therefore, a downfield chemical shift corresponds to a small shielding constant and vice versa. We then decompose $S_H$ as

$$S_H = S_{N_1H} + S_{N_2H} - S_{HB} = S_{Covalent} - S_{HB}.$$ 

In the SHB of DMANH, $S_{N_1H}$ and $S_{N_2H}$ are the chemical shielding from the molecular bond between the proton and the N$_1$ and N$_2$ atoms, respectively. Their sum gives the overall impact of the covalent bonds, $S_{Covalent}$, which is counteracted by the influence of hydrogen bonding, $S_{HB}$. To obtain $S_{N_1H}$ and $S_{N_2H}$, we use N,N-dimethylaniline to mimic the donor and acceptor groups in DMANH and scan the N–H bond length in its protonated form. The value of $S_{HB}$ is then computed by taking the difference between the total shielding constant and $S_{Covalent}$. All of the shielding constants vary with the proton position, and their relations with $\nu$ are shown in Figure 5.6.

When $\nu < -0.8$ Å, the proton is covalently attached to the N$_1$ atom in the SHB of
Figure 5.6: Decomposition of the total shielding constants into the contributions from covalent bonding (Covalent) and hydrogen bonding (HB) for DMANH.

DMANH. For example, at $\nu = -1.0$ Å, the $N_1$–H bond length is 0.9 Å and $S_{N_1H}$ has a large value of 32.7 ppm (Figure 5.6). To characterize this $\sigma_{NH}$ bond, we further carry out the NBO analysis\textsuperscript{268} to obtain its molecular orbitals. As Figure 5.7a demonstrates, there is a significant amount of electron density between the $N_1$ and H atoms, which shields the proton from the external magnetic field. Due to the compact structure of this SHB ($R = 2.6$ Å), the electron density around the $N_2$ atom also leads to a $\sigma_{N_2H}$ of 15.7 ppm, giving an overall $S_{Covalent}$ of 48.4 ppm. In the mean time, the presence of the hydrogen bond induces an opposite effect. As shown in Figure 5.7b, while the antibonding $\sigma_{NH}^*$ orbital contains a nodal plane between the $N_1$ and H atoms, it has a lobe that points towards the $N_2$ atom and overlaps with its $p$-type lone pair orbital, $n_p$. This leads to a $n_p \rightarrow \sigma_{NH}^*$ charge transfer in the hydrogen bonding interaction and weakens the covalent $N_1$–H bond. From Figure 5.6, The hydrogen bonding interaction deshields the proton by 22.5 ppm and renders a 46% cancellation to the impact of the covalent bonds, giving an overall $\delta_H$ of 5.7 ppm. We observe similar competition of these electronic quantum effects in all model molecules, and as a result, their $\delta_H$ are in the regular range of 5–10 ppm when $\nu < -0.8$ Å in the SHBs (Figure 5.5).

As the proton migrates towards the center of the SHB in DMANH, the covalent bond to the $N_1$ atom weakens while its new bond with $N_2$ begins to form. As Figure 5.6 shows, this
results in a decrease in $S_{N_1 H}$ and an increase in $S_{N_2 H}$, with the overall $S_{Covalent}$ reaching a minimum of 39.5 ppm at $\nu = 0$. In this symmetric position, the charge transfer in the hydrogen bonding interaction occurs from the lone pairs of the N$_1$ and N$_2$ atoms to the valence $s_H^*$ orbital of the proton, as represented in Figure 5.7c. Accordingly, $S_{H H}$ reaches a maximal deshielding of 30.7 ppm (Figure 5.6) and the amount of cancellation between the two electronic quantum effects increases to 78%. As shown in Figure 5.5, the chemical shift reaches its maximal value of 22.8 ppm. As the chemical structure of DMANH is symmetric, we expect to observe a similar $n_p \rightarrow \sigma_{N H}^*$ charge transfer when the proton moves closer to the N$_2$ atom in the SHB ($\nu > 0$). Therefore, the changes of $\delta_H$ and $S_{H H}$ with the proton position in Figs. Figure 5.5 and Figure 5.6 exhibit symmetric features with respect to $\nu = 0$.

Due to the competition of the covalent and hydrogen bonding interactions, $\nu$ makes a good coordinate to characterize $\delta_H$ of a SHB. To further examine this observation, we compute the $^1$H chemical shifts of a network of SHBs in the active site of an enzyme ketosteroid isomerase (KSI) and compare the results with those from the model small molecules in Figure 5.5b. KSI and its Asp40Asn mutant KSI$^{D40 N}$ contain a network of SHBs in the active site, which is composed of residues Tyr16, Tyr32 and Tyr57. The NMR chemical shifts of the hydrogen bonds between Tyr16/Tyr32 and Tyr57 from the AI-
PIMD simulations in the QM/MM setup are computed using the B3LYP density functional and the 6-31+G(d,p) basis set. In addition, we include the data from previous computational studies that implement DFT methods to calculate the chemical shifts of a variety of organic compounds with intra and intermolecular hydrogen bonds. In the data set T06, $\delta_H$ is for a proton in the N–H···N hydrogen bond in a pyrrole derivative. In G13, NMR calculations are performed for the hydrogen bond between phenol and different solvents. In G15, NMR chemical shifts are calculated for a set of organic compounds with intermolecular or intramolecular hydrogen bonds using the B3LYP and the M06-2X density functionals and the 6-31+G(d) basis set. In G17, the NMR chemical shifts of the enol-enol tautomers of $\beta$-dicarbonyl compounds are computed using the B3LYP density functional and the 6-31+G(d), 6-311G(d,p) and def2TZVP basis set. As shown in Figure 5.5b, the $\delta_H$ values follow the same quadratic relation with $\nu$ for all systems. While $\delta_H$ is strongly correlated with the elongation of the Donor–H covalent bond, we find that the universal trend in Figure 5.5 does not hold if we invoke the O–H or N–H bond length, rather than $\nu$, as a coordinate.

Symmetry of the SHBs determines the average chemical shift

From Equation 5.2, the average $^1$H chemical shift of a SHB depends on $\delta_H(\nu)$ and $P(\nu)$. Figure 5.5 demonstrates that $\delta_H(\nu)$ of the model molecules follow a universal trend, which can be well described using Equation 5.3, despite the differences in their chemical compositions and solvation conditions. Therefore, the variations in $\langle \delta_H \rangle$ observed in Table 5.1 arise from the probability distribution of $\nu$ in the molecules.

From the free energy surfaces in Figs. Figure 5.2 and Figure 5.4, we assume that $P(\nu)$ can be written as a linear combination of two Gaussian functions,

$$P(\nu) = \frac{A_1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\nu-\nu_0)^2}{2\sigma^2}} + \frac{A_2}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\nu+\nu_0)^2}{2\sigma^2}}$$

(5.4)
Here the constants $\nu_0$ and $\sigma$ define the center and width of the Gaussian functions, which are normalized, $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\nu-\nu_0)^2}{2\sigma^2}} d\nu = 1$. The coefficients $A_1$ and $A_2$ represent the symmetry of a SHB and they satisfy $A_1 + A_2 = 1$ to ensure that $P(\nu)$ is normalized, i.e., $\int_{-\infty}^{\infty} P(\nu) d\nu = 1$. As shown in Figure 5.8, the probability distribution of $\nu$ from the AIMD and AI-PIMD simulations of the model molecules can be well represented using Equation 5.4.

![Figure 5.8: The probability distribution of $\nu$ from the (a) AIMD and (b) AI-PIMD simulations of the model molecules in different solvents (symbols) and the least squares fitting of them using Equation 5.4 (lines). The R2 values for the fitting range between 0.971 and 0.999 for AIMD simulations and between 0.986 and 0.998 for AI-PIMD simulations.](image)

Given Equation 5.3, the average $^1$H chemical shift can be calculated as,

$$\langle \delta_H \rangle = \int_{-\infty}^{\infty} \delta_H(\nu) P(\nu) d\nu$$

$$= \int_{-\infty}^{\infty} (21.9 - 16.1\nu^2) \left[ \frac{A_1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\nu-\nu_0)^2}{2\sigma^2}} + \frac{A_2}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\nu+\nu_0)^2}{2\sigma^2}} \right] d\nu$$

$$= 21.9(A_1 + A_2) - 16.1(\nu_0^2 + \sigma^2)(A_1 + A_2)$$

$$= 21.9 - 16.1(\nu_0^2 + \sigma^2)$$

(5.5)

From the first principles simulations, the proton hopping frequencies between the donor and acceptor groups of a SHB is much faster than the time scale of a typical NMR measurement. Therefore, the experimental $^1$H NMR chemical shift of a system containing a
SHB corresponds to $\langle \delta_H \rangle$ from the simulations. We consider $\nu_0$ as the average proton position in a SHB, which is equivalent to the proton position observed in experiments. When a SHB is highly asymmetric, like in the case of CUA, $\nu_0$ is the average position of the proton from the simulations. When a SHB is symmetric, one cannot distinguish between its donor and acceptor atoms, and so we take the absolute value of $\nu$ from the simulations and use its average as $\nu_0$.

Compared to AIMD simulations, AI-PIMD simulations include both electronic and nuclear quantum effects and mimic the experimental conditions more closely. Therefore, we will use the $\nu_0$ values from the AI-PIMD simulations of the model molecules, as listed in Table 5.1. As the BLYP density functional is known to overstructure hydrogen bonded systems, it overestimates the $^1$H chemical shift in the AI-PIMD simulations. To alleviate the problem, we determine $\sigma$ by globally minimizing the differences between the predicted chemical shift, $\langle \delta_H \rangle_{\text{predict}}$, and the experimental chemical shift of all the systems. The resulting $\sigma$ is 0.3 Å and the root-mean-square deviation is 0.8 ppm. The $\langle \delta_H \rangle_{\text{predict}}$ values are also listed in Equation 5.3. Therefore, Equation 5.5 becomes

$$\langle \delta_H \rangle = 20.5 - 16.1 \nu_0^2.$$  \hspace{1cm} (5.6)

Here $\nu_0$ is the average proton position in a SHB, which can be calculated from the AIMD and AI-PIMD simulations. Note that the chemical structures of DMANH, HM and DHND are symmetric, and hence one cannot distinguish between the donor and acceptor atoms in their SHBs. In these cases, $\nu_0$ should be taken as the average of the absolute values of $\nu$ from the simulations. $\nu_0$ from the AIMD and AI-PIMD simulations of the model molecules are listed in Table 5.1, and we will use them to assess how the $^1$H NMR chemical shifts are influenced by the symmetry and quantum nature of the SHBs.

In the model molecules, $R$ of the SHBs are all below 2.7 Å. As shown in Table 5.1, $\nu_0$ of these systems are mostly below 0.5 Å from the AIMD simulations. Explicit inclu-
sion of nuclear quantum effects allows the proton to move further towards the center of the SHBs and reduces the value of \( \nu_0 \) by 15-37\%. As a result, \( \langle \delta_H \rangle \) from the first principles simulations are all above 16 ppm, and the chemical shifts from AI-PIMD simulations are larger by 4–12\% than those from AIMD simulations. Therefore, the interplay of electronic and nuclear quantum effects gives rise to the highly downfield \(^1\)H NMR chemical shift, a spectral signature of SHBs. Note that in Table 5.1, in a few cases the \( \langle \delta_H \rangle \) values predicted from AIMD and AI-PIMD simulations are larger than those from experimental measurements, as it is well-known that GGA functionals tend to overstructure the hydrogen bonds and overestimate the amount of proton delocalization.\(^{155,276,277}\) To alleviate the problem, one can use hybrid functionals such as B3LYP\(^{127}\) in the first principles simulations to better estimate \( \nu_0 \) and \( \langle \delta_H \rangle \).

In the SHBs of DMANH, HM and DHND, the intrinsic \( \Delta pK_a \) is 0. As Table 5.1 shows, when DMANH and HM are placed in organic solvents, AI-PIMD simulations predict that \( \nu_0 \leq 0.32 \) Å and \( \langle \delta_H \rangle \geq 20 \) ppm. In contrast, \( \langle \delta_H \rangle \) of HM and DHND decrease in aqueous solutions because the heterogeneous solvation environment perturbs their \( pK_a \) matched condition and induces asymmetry in their SHBs. Compared to the other 3 model molecules, CUA lacks the reflection symmetry in its chemical structure (Figure 5.1) and possesses a \( \Delta pK_a \) of 3.7 in the SHB. Its \( \nu_0 = -0.34 \) Å and \( \langle \delta_H \rangle = 19.6 \) ppm from the AI-PIMD simulations. To elucidate how a non-zero \( \Delta pK_a \) leads to the asymmetry of a SHB, we carry out the NBO analysis of CUA and evaluate the charge transfer characters when the proton moves between the donor O atom and the acceptor N atom. As shown in Figure 5.9a, when \( \nu \) is -0.9 Å, the charge transfer goes from the \( n_p \) orbital of the N atom to \( \sigma^*_{OH} \) with substantial orbital overlap. From second-order perturbation theory calculations,\(^{268}\) this charge delocalization stabilizes the SHB by 24.5 kcal/mol. When the proton is transferred to the N atom, the charge transfer becomes \( n_p \rightarrow \sigma^*_{NH} \), as demonstrated in Figure 5.9b. However, due to the geometry constraint of the imidazole ring, the \( \sigma^*_{NH} \) orbital points towards the \( n_p \) orbital of the O atom with an angle, resulting in a smaller amount of overlap and a
reduced stabilization energy of 12.9 kcal/mol. Therefore, $\Delta pK_a$ impacts how the electrons are quantum mechanically distributed between the donor and acceptor atoms, which in turn determines the symmetry of the SHB and its $^1H$ NMR chemical shift.

Figure 5.9: The $n-\sigma^*$ orbital overlap diagrams for CUA at (a) $\nu=-0.9$ Å and (b) $\nu=1.0$ Å. Silver, red and white represent C, O and H atoms, respectively. The red and green spheres are the positive and negative molecular orbitals.

In Equation 5.6, one can take $\nu_0$ as the experimentally measured proton position in a SHB. To evaluate its validity, we use the experimental data on a series of inorganic and organic compounds with intra or intermolecular O–H···O SHBs, and we refer to the data sets as S88,234 B04267 and D09.236 For these small molecule crystals, the proton positions, $\nu_{0,exp}$, are determined from neutron diffraction and the $^1H$ chemical shifts, $\delta_{H,exp}$, are measured by solid-state NMR spectroscopy. As shown in Figure 5.10a, we combine $\nu_{0,exp}$ and Equation 5.6 to predict the chemical shifts and find them to be in good agreement with the experimental measurements. As Equation 5.6 is developed from the AI-PIMD simulations with the electronic surface described by the BLYP functional, which tends to overly delocalize the protons in the SHBs, it slightly overestimates $\langle \delta_H \rangle$ for a few cases. However, this equation is capable of quantitatively predict the chemical shifts with a root-mean-square deviation of 1.7 ppm as compared to the experimental values. In turn, one can invoke Equation 5.6 to derive the proton position in a SHB from the experimental NMR chemical shift. From the 3 data sets, we calculate the $\nu_0$ values and compare them with $\nu_{0,exp}$ in Figure 5.10b. The predicted values follow a strong positive correlation with $\nu_{0,exp}$,
giving a root-mean-square deviation of 0.1 Å. Therefore, Equation 5.6 provides an efficient and accurate way to obtain the location of a proton based on its downfield chemical shift.

Figure 5.10: Comparison between the predicted and experimental \(^{1234,236,267}\) (a) chemical shifts and (b) proton positions of a series of inorganic and organic molecules with SHBs. The dashed diagonal lines represent perfect correlation between the variables.

5.4 Conclusions

In this chapter, we exploit first principles simulations to examine the symmetry and \(^1H\) NMR chemical shifts of SHBs, which incorporate the impact of quantum effects and structural and environmental fluctuations on the proton behavior in these compact structures. By performing AIMD and AI-PIMD simulations on a set of model molecules in organic solvents and aqueous solutions, we reveal how the geometrical and chemical criteria required for the formation of symmetric hydrogen bonds reflect the quantum nature of the SHBs. First, the interplay of short R and electronic quantum effects promotes the lengthening of the Donor–H bond and the sharing of the proton in a hydrogen bond. Second, a matched \(pK_a\) in a SHB ensures a substantial overlap of the molecular orbitals whether the proton is closer to the donor group or transferred to the acceptor atom, leading to a charge delocalization and stabilization of the hydrogen bonding interaction. Compared to the less polar
organic solvents, water molecules have non-uniform interactions with the donor and acceptor atoms in the SHB, which modulates the $\Delta pK_a$ value of the hydrogen bond and results in an instantaneous asymmetric distribution of the proton positions. Finally, nuclear quantum effects can qualitatively and quantitatively change the properties of a SHB and hence must be included when considering its symmetry. In particular, the zero-point energy associated with a typical O–H or N–H stretch, which is $\sim 5$ kcal/mol at room temperature, allows the proton to be quantum mechanically delocalized and more likely to reside near the center of a hydrogen bond.

While the model molecules have different chemical compositions and solvation environment, Figure 5.5 shows that their $^1H$ NMR chemical shifts follow a universal trend with the proton position, which arises from a competition between the covalent bonding and hydrogen bonding interactions. We expect the universal relation in Equation 5.3 to hold in SHBs not only in small molecules, but also in biological macromolecules. From the first principles simulations, the proton frequently hops between the donor and acceptor groups with an average residence time less than 1 ps. As such, the $^1H$ chemical shift measured from NMR experiments reflects an ensemble average from a probability distribution of the proton positions, and hence it provides a highly sensitive probe on the symmetry of a SHB. Based on the AI-PIMD simulations and the experimental $\delta_H$ of the model molecules, we develop Equation 5.6 and uncover why the highly downfield $^1H$ chemical shift is a spectral signature of SHBs.

SHBs occur extensively in proteins and have been proposed to play essential roles in biological functions.$^{16,21,83,85}$ However, the geometric features and energetics of these compact structures are still not well understood mainly because of the experimental challenge to determine the proton positions in a large protein.$^{112-114,278}$ While neutron diffraction has enabled unambiguous assignment of the proton positions,$^{83,87,96}$ its application to macromolecules is limited by the small number of high-flux neutron sources globally.$^{279}$ As a result, NMR spectroscopy becomes one of the most commonly used technique in detecting
biological SHBs, and the relation between the $^1$H chemical shift and the hydrogen bond length has been well established.\textsuperscript{22,49,233,273} In this chapter, we demonstrate in Figure 5.10b that Equation 5.6 allows one to accurately and efficiently calculate the proton position, $\nu_0$, from the chemical shift. Our metric, combined with the previous developed relations, will allow researchers to determine the length and proton position in a SHB directly from NMR measurements, and hence facilitate the investigation of the physical properties and biological functions of these specialized structural elements with unprecedented detail.
CHAPTER 6
GEOMETRY AND PROTON NMR CHEMICAL SHIFTS OF BIFURCATED SHORT HYDROGEN BONDS

This chapter is reprinted with permission from Ref. [P3]

6.1 Introduction

Bifurcated three-center hydrogen bonds, which involve one hydrogen bond donor and two acceptors, are widely observed in crystalline structures ranging from organic small molecules to amino acids and biological macromolecules.\textsuperscript{280–284} Notably, a series of enzymes contain bifurcated hydrogen bonds with relatively short distances between the active-site residues and intermediate analogs, indicating their functional importance in enzyme catalysis.\textsuperscript{285–287} Due to their compact structures, bifurcated short hydrogen bonds (SHBs) can have covalent characters that deviate significantly from classical electrostatic interactions. However, while the quantum nature of linear SHBs has been extensively investigated,\textsuperscript{12,30,35,37,44,119,155,168,288–290} much less attention has been paid to the structure and properties of bifurcated SHBs.

The macrocyclic compound N,N',N''-tris(p-tolyl)azacalix[3](2,6)pyridine (TAP) provides an excellent model for examining bifurcated SHBs in the condensed phases. As illustrated in Figure 6.1a, it takes a triangular shape and the rigid ring structure forms a cavity, in which three pyridine N atoms are in close proximity with R varying between 2.60 and 2.67 Å.\textsuperscript{291} As a means to release the steric restraint, TAP acts as a superbase with a $pK_{BH^+}$ of 23.1 in acetonitrile, and readily captures a proton.\textsuperscript{292} The resulting cation, TAPH, adopts an approximate $S_3$ symmetry and contains a bifurcated SHB in its cavity (Figure 6.1a). This protonation slightly distorts the macrocyclic framework with R in the N$_1$–N$_2$ and N$_2$–N$_3$ pairs reduced to 2.53 Å and 2.57 Å, respectively. The PF$_6$ salt of TAPH
shows a prominently downfield $^1\text{H}$ NMR chemical shift, $\delta_H$, of 22.1 ppm, indicating that the proton is shared among the pyridine N atoms and the bifurcated SHB is partially covalent in nature. This is consistent with the X-ray crystal structure of the molecule, in which the $\text{N}_1$–H covalent bond is elongated to 1.18 Å and the distances from the H atom to the $\text{N}_2$ and $\text{N}_3$ atoms are shortened to 1.63 Å.

TAP behaves like DMAN\textsuperscript{72} as both of them have exceptionally high proton affinity and act as base catalysts in a variety of organic reactions.\textsuperscript{73,293,294} In Chapter 5, we show the quantum nature and large $\delta_H$ values of DMANH (Figure 6.1b).\textsuperscript{290} As TAPH and DMANH share common chemical features, including ring structures that constrain the distances between the N atoms in their intramolecular SHBs and the $^1\text{H}$ NMR chemical shifts that are much further downfield than those of common hydrogen-containing functional groups (1 – 10 ppm), we expect the bifurcated SHB in TAPH to exhibit prominent quantum mechanical features.

In this chapter, we invoke AIMD and AI-PIMD simulations to determine the geometry, quantum fluctuations and chemical shifts of TAPH in the acetonitrile solution. Comparing the AIMD and AI-PIMD simulations, we will unravel how the interplay of electronic and nuclear quantum effects changes the proton sharing conditions in the bifurcated SHB.
and results in the large deshielding effects as observed in the NMR experiments. Using DMANH as a reference, we will further elucidate how the arrangement of donor and acceptor groups in a SHB influences its strength and $^1$H chemical shift.

### 6.2 Methods

The initial structure of TAPH is obtained from the Cambridge Structural Database\textsuperscript{94} with the refcode of XEKQUA. The TAPH molecule is solvated in 167 acetonitrile molecules to form a cubic box with a length of 25 Å, and a chloride ion is added to neutralize the positive charge of the system. Using the same method discussed in Chapter 5, we perform molecular dynamics (MD) simulations to get the initial structure for the first principle simulations. After equilibrating the TAPH solutions for 5 ps, we carry out AIMD simulations for 60 ps and AI-PIMD simulations for 35 ps with a time step of 0.5 fs in the canonical ensemble at a temperature of 300 K by applying the same set of simulations in Chapter 5.

From the first principles simulations, we calculate the positions of the proton inside the triangle formed by the three N atoms. At each time step, we define an internal two-dimensional coordinate system by placing the N$_2$ and N$_3$ atoms in the x axis and the N$_1$ atom along the positive y direction, and compute the proton position as a projection onto the coordinate system. The results are shown in Figure 6.4a and b, in which N$_1$, N$_2$ and N$_3$ are placed at their average positions from the AIMD and AI-PIMD simulations, respectively.

We also extract TAPH-solvent clusters every 100 fs from the AIMD simulations and every 10 fs from the AI-PIMD simulations, randomly selecting a ring polymer bead each time, and calculate the $^1$H NMR chemical shift of the hydrogen bonded proton in TAPH. We further set a grid of 0.05 Å in $r$ and 2.0° in $\theta$ within the triangle formed by the three N atoms to build the relation between the chemical shift and the geometry of the bifurcated SHB. In the rare case that a grid is not sampled, we randomly select 5 configurations in this region from the AI-PIMD simulations and compute their chemical shifts. We then average the $\delta_H$ values within each grid and plotted the results in Figure 6.6. In these
calculations, we include the acetonitrile molecules that are within 5.0 Å of any atom in TAPH, and the resulting TAPH-solvent clusters containing 103-169 atoms. We use the same method in Chapter 5 to calculate the NMR chemical shifts of TAPH clusters. In addition, we extract two TAPH configurations, one at the free energy minimum ($r_1 = 1.1$ Å and $\theta_1 = 30.0^\circ$) and the other at a shared position ($r_1 = 1.5$ Å and $\theta_1 = 30.0^\circ$), from the AI-PIMD simulations and decomposed their shielding constants. In vacuum, the shielding constants of the two configurations are 20.2 ppm and 27.6 ppm, respectively. When solvents within 5 Å of the TAPH molecule are included, the shielding constants are 20.0 ppm and 27.7 ppm. Given the small differences, we take the gas-phase values and further performed the following decomposition in vacuum. We use the protonated compound N-(4-methylphenyl)-2-pyridinamine to mimic the donor and acceptor groups in the bifurcated SHB, and use the protonated $N^2$-(pyridin-2-yl)$-N^6$-$N^6$-di-$p$-tolylpyridine-2,6-diamine to mimic the dimeric hydrogen bonded system. To divide TAPH into these monomeric and dimeric structures, we remove the corresponding C–N covalent bonds, add H atoms to saturate the bonds, and optimize the position of these capping atoms with all the other atoms fixed. We then carry out the NMR calculations by placing the atoms, except the capped H atoms, in their positions in the corresponding TAPH configuration. In addition, we also perform the Natural Bond Orbital (NBO) analysis on the two TAPH structures using the NBO 6.0 program.

As the atoms in molecules (AIM) method provides an effective way to examine intramolecular hydrogen bonds from the topological properties of the electron density, I use it to estimate the energies of the intramolecular SHBs of DMANH and TAPH. Specifically, we optimize the geometry of DMANH and TAPH in vacuum and performe the AIM analysis with the B3LYP-D3 method and the 6-311++G(2d,p) basis set using the AIMPAC program. After locating the $(3,-1)$ bond critical point along the acceptor-proton
path of the molecule, \( r_{CP} \), we estimate the hydrogen bond energy using the relation\(^{296,297}\)

\[
E_{HB} = -\frac{a_0^3}{2} V(r_{CP}).
\]  

(6.1)

Here \( a_0 \) is the Bohr radius and \( V(r_{CP}) \) is the local potential energy density at the bond critical point of the hydrogen bond.

### 6.3 Results and discussion

#### 6.3.1 Structure of the bifurcated SHB

The bifurcated hydrogen bond in TAPH is formed by three N atoms interacting with a proton. Constrained by the macrocyclic framework of the molecule, the separations between \( \text{N}_1, \text{N}_2 \) and \( \text{N}_3 \) in the cavity sample a relatively small range of 2.3 – 3.0 Å from the AIMD and AI-PIMD simulations. The average \( R \) between each pair of the N atoms is 2.6 Å, and hence these heavy atoms in the bifurcated SHB adopt an approximately equilateral triangular shape. To further characterize the geometry of the hydrogen bond, we define the distances from the proton to the \( \text{N}_1, \text{N}_2 \) and \( \text{N}_3 \) atoms as \( r_1, r_2 \) and \( r_3 \), respectively, and the angles between the \( \overrightarrow{\text{NH}} \) vectors as \( \alpha_1, \alpha_2 \) and \( \alpha_3 \), as illustrated in Figure 6.2. We obtain the average values of these parameters from the first principles simulations and list them in Table 6.1. In both AIMD and AI-PIMD simulations, the sum of \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) is close to 360°, demonstrating that the proton lies mostly in the plane formed from the three N atoms. The average \( r_1, r_2 \) and \( r_3 \) are between 1.4 and 1.6 Å, much longer than the length of 1.0 Å for a regular N–H covalent bond. This suggests that the proton either resides near the center of the N triangle, or frequently transfers to different donor atoms in the bifurcated hydrogen bond.

To distinguish between the two scenarios, we examine the instantaneous positions of the proton from the first principles simulations. In the AIMD simulations, we observe no cases where the proton is close to the center of the bifurcated SHB. As demonstrated in
Figure 6.2: Geometric parameters for the bifurcated SHB in TAPH. From AIMD and AI-PIMD simulations, $N_1$, $N_2$ and $N_3$ form an approximately equilateral triangle.

Table 6.1: Average geometric parameters of the bifurcated SHB as obtained from the AIMD and AI-PIMD simulations of TAPH in the acetonitrile solution.

<table>
<thead>
<tr>
<th></th>
<th>$r_1$ (Å)</th>
<th>$r_2$ (Å)</th>
<th>$r_3$ (Å)</th>
<th>$\alpha_1$ (°)</th>
<th>$\alpha_2$ (°)</th>
<th>$\alpha_3$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMD</td>
<td>1.6</td>
<td>1.5</td>
<td>1.6</td>
<td>121.4</td>
<td>116.9</td>
<td>119.5</td>
</tr>
<tr>
<td>AI-PIMD</td>
<td>1.4</td>
<td>1.6</td>
<td>1.6</td>
<td>120.2</td>
<td>123.4</td>
<td>113.2</td>
</tr>
</tbody>
</table>

Figure 6.3a, at any instant, the proton resides asymmetrically in the bifurcated SHB as it stays covalently bonded to a donor N atom and has different distances to the two acceptors. After collecting the proton positions throughout the AIMD simulations in Figure 6.4a, we find that the proton moves around a N atom with an average residence time of 4.6 ps. Electronic quantum effects enable it to hop occasionally towards another N in the bifurcated hydrogen bond, forming three distinct clusters in the triangle constructed by the N atoms. Due to the proton transfer events, the average $r_1$, $r_2$ and $r_3$ are above 1.3 Å. Note that we cannot directly compare the average $r_1$, $r_2$ and $r_3$ from the AIMD simulations to the N–H bond length measured in the X-ray crystallography experiments. Due to the symmetry of TAPH and its SHB, $N_1$, $N_2$ and $N_3$ are equivalent and the crystal structure shows a distance of 1.18 Å between the proton and the donor N atom in the hydrogen bond. If we assign the N atom that the proton is closest to as the donor atom at each simulation step, we find the average Donor–H length to be 1.08 Å, which is consistent with the experimental value.
In AI-PIMD simulations, each atom is represented by 6 ring polymer beads and the proton is capable of simultaneously bonded to both the donor and acceptor atoms in the hydrogen bond, as demonstrated in Figure 6.3b. As such, the interplay of electronic and nuclear quantum effects allows both scenarios of the proton arrangement to occur and leads to the observed large $r_1$, $r_2$ and $r_3$ values. From Figure 6.4b, the proton’s trajectory covers most of the available region in the N triangle and hence the proton becomes shared in the bifurcated SHB. In addition, Figure 6.4b shows that the proton can move around the center of the bifurcated SHB, smearing out the clustering characters in Figure 6.4a and contributing to the elongation of the N–H bonds. In both cases, the properties of the bifurcated SHB are qualitatively and quantitatively different from the standard description of a hydrogen bond as classical dipole-dipole interactions.

To further reveal how quantum fluctuations impact the structure of TAPH, we calculate the free energies for proton movement in the bifurcated SHB from the AIMD and AI-PIMD simulations. As illustrated in Figure 6.2, we use the three N atoms to define a coordinate plane. We first place $N_1$ at the origin and represent the position of the proton using the polar coordinates $r_1$ and $\theta_1$, and repeat this process by treating $N_2$ ($N_3$) as the origin and considering the corresponding coordinates $r_2$ ($r_3$) and $\theta_2$ ($\theta_3$). After collecting all the data,
Figure 6.4: Instantaneous positions of the proton in the triangle formed by the N atoms (a and b) and the free energy surfaces for proton movement (c and d) in the bifurcated SHB in TAPH. (a) and (c) are from AIMD simulations, and (b) and (d) are from AI-PIMD simulations. The free energy profiles are obtained by subtracting their lowest values such that the minimal free energy in each plot is 0.

we compute the probability of finding the proton at a distance $r$ and angle $\theta$ from a N atom, $P(r, \theta)$. The free energy is then calculated as

$$ F = -k_B T \ln P(r, \theta), $$

(6.2)

where $k_B$ is the Boltzmann constant and $T$ is the simulation temperature of 300 K. From AIMD simulations of TAPH, we observe three minima in the free energy surface of Figure 6.4c, consistent with the three clusters in Figure 6.4a. Since the three N atoms are
equivalent in the bifurcated SHB, we will use N₁ as an example to explain the features in Figure 6.4c. Minimum 1 in the free energy profile occurs at \( r = 1.1 \) Å and \( \theta = 30.0^\circ \), in which the proton is covalently bonded to the N₁ atom. When \( \theta \) becomes smaller than 30°, the hydrogen bond is mainly between the N₁ and N₂ atoms, whereas a \( \theta > 30^\circ \) leads to a hydrogen bond mostly between the N₁ and N₃ atoms. Minima 2 and 3 have \( r \) of 1.75 Å and are symmetric about a horizontal line of \( \theta = 30^\circ \). Their \( \theta \) values are 18.0° and 42.0°, corresponding to the proton transferred configurations in which the H atom is covalently bonded to the N₂ and N₃ atoms, respectively. Accompanying the conformational dynamics of TAPH and its solvation structure, the proton fluctuates around a N atom and occasionally moves to another donor atom with a barrier of 3.4 kcal/mol.

Compared to the AIMD results, the locations of the free energy minima are largely unchanged in the AI-PIMD simulations, as demonstrated in Figure 6.4d. However, inclusion of nuclear quantum effects results in two prominent features in the free energy surface. First, the proton samples a much wider phase space that covers \( r \) of 0.8 – 2.5 Å and \( \theta \) of 1.1° – 64.3°. Second, the proton becomes quantum mechanically delocalized in the bifurcated hydrogen bond. In particular, the proton can reside almost equidistantly between two N atoms (\(|r_i - r_j| \leq 0.1 \) Å for \( i, j = 1, 2, 3 \)) and it takes less than 1.4 kcal/mol for the proton to move from a minimum to such a shared position. As a result, the probability of finding the proton shared between two N atoms increases by over 30 times from 0.1% in AIMD simulations to 3.3% in AI-PIMD simulations. In addition, we observe a relatively low barrier of 2.3 kcal/mol to transfer the proton from a free energy minimum to the center of the N triangle (\( r_1 \approx r_2 \approx r_3 \)), which further promotes the delocalization of the proton in the bifurcated SHB. Here, both the AIMD and AI-PIMD simulations are performed using the BLYP density functional\(^{250,251}\) and the D3 dispersion correction.\(^{128}\) As the GGA functionals are known to over-delocalize the electron density in hydrogen bonded systems,\(^ {155,276,302}\) one can further improve the description of the proton positions using a hybrid functional such as B3LYP\(^ {127}\) in the simulations. However, we expect the resulting
free energy surfaces and the impact of nuclear quantum effects to exhibit similar features as those observed in Figure 6.4.

6.3.2 Competing quantum effects in linear and bifurcated SHBs

In conventional two-center hydrogen bonds, one often observes a nearly linear arrangement of the hydrogen bonded atoms. For example, DMANH contains a SHB (Figure 6.1b), and its average N1-H-N2 angle is 155.1° in the acetonitrile solution from our AI-PIMD simulations in Chapter 5.290 The bifurcated SHB in TAPH involves one donor and two acceptor N atoms, which interchange with one another when proton transfer events occur. Accordingly, the relevant Donor-H-Acceptor angles in this system are α1, α2 and α3 (Figure 6.2) and their average values are about 120° from the first principles simulations (Table 6.1). While DMANH and TAPH share similar chemical features and the average R of their intramolecular hydrogen bonds are both 2.6 Å from the AI-PIMD simulations, this 35° difference in their hydrogen bond angle suggests that the linear and bifurcated SHBs have different quantum mechanical characters. As a comparison between the two systems, we calculate the quantum kinetic energy (QKE) of the protons in the intramolecular SHBs, which provides a sensitive probe to their local chemical environment. From the AI-PIMD simulations, we compute the average QKE of the proton using the centroid-virial estimator,303,304

$$\langle \text{QKE} \rangle = \frac{3}{2} k_B T + \frac{1}{2P} \sum_{j=1}^{P} (r^{(j)} - \bar{r}) \cdot \frac{\partial V}{\partial r^{(j)}}.$$  \hspace{1cm} (6.3)

Here \(P\) is the number of ring polymer beads, \(r^{(j)}\) is the position of the \(j^{th}\) ring polymer bead that represents the proton and \(\bar{r} = \sum_{j=1}^{P} r^{(j)}/P\) is the centroid position. \(V\) is the potential energy of the system and \(-\frac{\partial V}{\partial r^{(j)}}\) is the force acting on the \(j^{th}\) ring polymer bead of the proton. If the proton were a classical particle, the last term in Equation 6.3 would vanish and its average QKE would be 38.7 meV \(\left(\frac{3}{2} k_B T\right)\) in both the linear and bifurcated hydrogen bonds. As shown in Figure 6.5, the average QKE of the proton is 123.6 meV in
TAPH and 131.9 meV in DMANH, suggesting that the proton experiences a confinement force in both systems and this confinement is slightly smaller in the bifurcated SHB of TAPH.

Figure 6.5: The average QKEs of the hydrogen bonded proton in TAPH and DMANH as obtained from the AI-PIMD simulations. For each system, the total QKE is decomposed into three internal coordinates: the vector along the N–H bond (N–H), the perpendicular vector in the hydrogen bond plane (Plane), and the vector orthogonal to the hydrogen bond plane (Orth). The internal coordinate system for TAPH is illustrated in the inset picture, and the orange spheres and springs represent the ring polymer of the hydrogen bonded proton.

Comparing Figs. Figure 6.4c and Figure 6.4d, for the bifurcated SHB, nuclear quantum effects allow the proton to get as close as 0.8 Å and extend as far away as 2.5 Å from a N atom. In the meantime, the θ angle varies over a much wider range in AI-PIMD simulations as compared to AIMD simulations. As such, incorporating the quantum nature of the nuclei not only enhances the N–H stretch vibration, but also increases the motion of the proton in the orthogonal directions, demonstrating the well-known phenomenon of competing quantum effects in hydrogen bonded systems.\textsuperscript{37,152,154,155,171} To elucidate how these geometric changes influence the bifurcated SHB in TAPH, we decompose the total QKE of the proton into three internal coordinates: the N–H bond vector, the perpendicular vector in the hydrogen bond plane formed by N, H and the N atom on the right side of the
N–H bond, and the vector orthogonal to the hydrogen bond plane. The internal coordinate system is illustrated in the inset of Figure 6.5. As a reference, we also perform the QKE decomposition on the hydrogen bonded proton of DMANH. At each simulation step of TAPH or DMANH, we define the N–H bond as the covalent bond between the proton and the N atom that it is closest to in the hydrogen bond.

As shown in Figure 6.5, all of the QKE components are larger than the classical contribution of 12.9 meV ($\frac{1}{2}k_B T$), and the confining forces mainly come from the N–H covalent bonds in the molecules. The total QKEs result from two competing quantum effects, which act differently for the linear and bifurcated SHBs due to their distinct geometries. For the linear SHB of DMANH, the N–H stretch motion lengthens the covalent bond and facilitates the sharing of the proton between the donor and acceptor atoms, whereas the bending and rotation of the N–H bond in the molecular plane disrupt the hydrogen bonding interaction. In TAPH, the N atoms adopt a triangular arrangement in its bifurcated SHB, which lacks linearity in the Donor-H-Acceptor angle. While the stretch vibration still strengthens the bifurcated SHB, proton delocalization along the N–H direction is less favored in this system compared to the linear SHB, especially when the $\theta$ angles are large. On the other hand, the in-plane bending and libration of the N–H bond enable the proton to be delocalized between a pair of N atoms and enhances the hydrogen bonding interaction. Therefore, the zero-point energy associated with the N–H stretch vibration strengthens both the linear and bifurcated SHBs, although its effect is more prominent for the former case and makes its proton much less confined in the covalent bond. Accordingly, the QKE component of the proton along the N–H direction is smaller by 18.5 meV in DMANH than that in TAPH (Figure 6.5). In contrast, the zero-point energies in the in-plane bending and librational motion distort the linear SHB but enhance the bifurcated SHB, leading to a larger in-plane component of the proton QKE for DMANH (Figure 6.5). The zero-point energies in the bending and rotational motion also allow the proton to move out of the molecular plane and break the hydrogen bonds in both DMANH and TAPH. Compared to the relatively planar
cavity of TAPH, the proton in DMANH moves less freely in this orthogonal direction due to the large steric hindrance created by the methyl groups attached to the donor and acceptor N atoms, and its corresponding QKE component is higher by 15.3 meV. Compared to the linear SHB of DMANH, the hydrogen bonded proton in TAPH has an overall smaller QKE, which arises from a 69% cancellation between these competing effects. The net influence of nuclear quantum effects is to make the proton move more freely in the bifurcated SHB as compared to its linear counterpart.

6.3.3 Downfield $^1$H NMR chemical shifts of TAPH

One of the most prominent features of SHBs is their far downfield $^1$H NMR chemical shifts. We collect 4100 TAPH-solvent clusters from the first principles simulations and find that the average $\delta_H$ for the hydrogen bonded proton are 20.7 ppm and 21.5 ppm from the AIMD and AI-PIMD simulations, respectively. The predicted chemical shifts are in good agreement with the experimental value of 22.1 ppm, and are 2-4 times larger than those typically observed for functional groups involving the N–H bonds. This quantitative agreement between the calculated and experimental $\delta_H$ values suggests that the BLYP-D3 approach correctly describes the proton sharing conditions of the bifurcated SHB in the first principles simulations and validates our method of obtaining the isotropic shielding constants on the protons. Furthermore, we observe a universal correlation between the $\delta_H$ value of a TAPH-solvent configuration and the geometry of the bifurcated SHB, in particular the N–H length $r$ and the angle $\theta$ (Figure 6.2), regardless of whether the configuration is obtained from AIMD or AI-PIMD simulations. This is demonstrated in Figure 6.6 using the averaged instantaneous $^1$H chemical shifts. Considering that the instantaneous $\delta_H$ fluctuates with the conformation of TAPH and its solvation environment in each snapshot of the first principles simulations, here we use a grid of 0.05 Å in $r$ and 2.0° in $\theta$ and average the instantaneous $\delta_H$ values within each grid to smooth the data. As shown in Figure 6.6, the distribution of the averaged instantaneous chemical shifts resembles the free energy profiles
in Figs. Figure 6.4c and Figure 6.4d, and can be divided into three regions. Around the free energy minima, the chemical shifts are less downfield. For example, when the proton is covalently bonded to the N\textsubscript{1} atom with \(r_1\) of 1.1 Å and \(\theta_1\) of 30.0°, the averaged \(\delta_H\) is 20.9 ppm and it decreases to a minimal value of 10.7 ppm when the proton further approaches the N\textsubscript{1} atom. As the N–H bond lengthens to over 1.2 Å, the chemical shifts become larger than 23.5 ppm. This is particularly the case when the proton is shared between two N atoms. For example, the proton is equidistant between the N\textsubscript{1} and N\textsubscript{2} atoms when \(r_1\) and \(\theta_1\) take the values of 1.3 Å and 28.1°, respectively, and its \(\delta_H\) value is 25.4 ppm. Finally, in the rare event that the proton moves to the center of the bifurcated hydrogen bond, the chemical shift reaches a maximal value of 27.7 ppm. In general, the averaged instantaneous chemical shift increases as the proton becomes more quantum mechanically delocalized in the bifurcated SHB of TAPH.

![Figure 6.6: Correlation between the averaged instantaneous \(^1\text{H} \) NMR chemical shift and the geometric parameters of the bifurcated SHB. The data from the AIMD and AI-PIMD simulations of TAPH are combined as they follow the same trend.](image)

The \(^1\text{H} \) NMR chemical shift stems from the electronic shielding effects that the surrounding atoms exert on the proton. For TAPH, the chemical shift is related to the isotropic magnetic shielding constant, \(S_{tot}\), by \(\delta_H = S_{TMS} - S_{tot}\). Here, \(S_{TMS}\) is the shielding constant of the protons in the reference molecule TMS and has a value of 31.6 ppm from our DFT calculations. In Chapter 5, we have shown that \(S_{tot}\) in linear SHBs arises from
a competition between covalent and hydrogen bonding interactions. We hence choose two TAPH configurations from the AI-PIMD simulations and decompose their isotropic magnetic shielding constants as

\[ S_{\text{tot}} = S_{\text{cov}} - S_{\text{HB}} = S_{N_1H} + S_{N_2H} + S_{N_3H} - S_{\text{HB}}. \] 

(6.4)

Here \( S_{\text{cov}} \) is the overall shielding constant on the proton due to its covalent bonding to the \( N_1, N_2 \) and \( N_3 \) atoms, which we represent as \( S_{N_1H}, S_{N_2H} \) and \( S_{N_3H} \), respectively. To calculate them, we approximately divide the TAPH molecule into three identical monomers, each containing one N atom and the proton from the bifurcated SHB. This involves breaking two C–N covalent bonds, and we add capping H atoms to saturate the dangling bonds. The resulting monomer is the protonated N-(4-methylphenyl)-2-pyridinamine and its chemical structure is depicted in Figure 6.7a. For each TAPH configuration, the \( S_{N_1H}, S_{N_2H} \) and \( S_{N_3H} \) values are computed by putting the C, N and H atoms, except the capping H’s, of the monomers at the same position as in the TAPH molecule. By taking the difference between these covalent shielding constants and \( S_{\text{tot}} \), we obtain \( S_{\text{HB}} \), which characterizes the deshielding of the proton due to hydrogen bonding interactions.

We choose the first TAPH configuration from a minimum of the free energy profile and it has \( r_1 = 1.1 \, \text{Å}, r_2 = r_3 = 1.8 \, \text{Å} \) and \( \theta_1 = 30.0^\circ \). The chemical shift and shielding constants of the hydrogen bonded proton are listed in Table 6.2. As the proton is close to the \( N_1 \) atom, \( S_{N_1H} \) has a large value of 21.5 ppm, while \( S_{N_2H} \) and \( S_{N_3H} \) are 15.7 ppm and 14.8 ppm, respectively. Accordingly, the proton experiences a significant shielding of 52.0 ppm due to the presence of the N atoms. This is counteracted by the hydrogen bonding interaction. As shown in Figure 6.8a, there is a noticeable overlap between the p-type lone-pair orbital of the N atom, \( n_p \), and the antibonding 1s orbital of the H atom, \( s_H^* \), from the NBO analysis. Hence the bifurcated hydrogen bond involves a \( n_p \rightarrow s_H^* \) charge transfer, which deshields the proton by 40.6 ppm. The 78% cancellation between the covalent and
Figure 6.7: Chemical structures of (a) the protonated N-(4-methylphenyl)-2-pyridinamine and (b) the protonated $N^2$-(pyridin-2-yl)-$N^2,N^6$-di-$p$-tolylpyridine-2,6-diamine. These are the monomers and dimers used in the decomposition of the magnetic shielding constants of the bifurcated SHB in TAPH.

hydrogen bonding contributions leads to an overall shielding constant of 11.4 ppm.

Given the ability to identify the deshielding effects due to the hydrogen bonding interactions, we will use them as an indicator for the strength of the intramolecular SHBs of TAPH and DMANH. For TAPH, we remove a donor or acceptor group in its bifurcated SHB and use the protonated $N^2$-(pyridin-2-yl)-$N^2,N^6$-di-$p$-tolylpyridine-2,6-diamine as a model system to mimic a hydrogen bonded dimer of the system, as demonstrated in Figure 6.7b. The resulting shielding constants of the hydrogen bonded proton are again calculated by keeping the positions of the relevant atoms identical for the dimer and TAPH. As shown in Table 6.2, $S_{HB}$ of these dimeric hydrogen bonds in the geometry of a bifurcated hydrogen bond are about 22 ppm. Note that due to our decomposition scheme, the sum of the dimeric $S_{HB}$ does not give the total $S_{HB}$ of the bifurcated SHB. As a comparison, we choose a DMANH structure with an N$_1$–H distance of 1.1 Å from the AI-PIMD simulations and compute its overall and decomposed shielding constants. As shown in Table 6.2, $S_{cov}$ of the linear SHB is much smaller than that of the bifurcated SHB because
Figure 6.8: The overlap of the $n_p$ and $s_H^*$ orbitals in the bifurcated SHB with (a) $r_1 = 1.1$ Å and $\theta_1 = 30.0^\circ$ and (b) $r_1 = 1.5$ Å and $\theta_1 = 30.0^\circ$. Silver, blue and white represent C, N and H atoms, respectively. The red and cyan spheres are the positive and negative molecular orbitals obtained from the NBO analysis.

DMANH contains only two N atoms, and its total shielding constant results from a 69% cancellation between the covalent and hydrogen bonding contributions. Comparing to a $S_{HB}$ of 28.2 ppm for DMANH, the shielding constants in the dimeric hydrogen bonds of TAPH are smaller by about 6 ppm. This suggests that the N–H···N interaction is weaker in the bifurcated hydrogen bond of TAPH than in the two-center hydrogen bond of DMANH because the former lacks linearity in its geometry. However, the overall $S_{HB}$ of TAPH is 12.4 ppm larger than that of DMANH, indicating that the bifurcated SHB becomes stronger than the linear SHB when its two N–H···N interactions are taken together.

To characterize how the proton sharing condition influences the $^1$H chemical shift and strength of a SHB, we pick a TAPH configuration that has $r_1 = r_2 = r_3 = 1.5$ Å and $\theta_1 = 30.0^\circ$ from the AI-PIMD simulations. From Table 6.2, as the proton goes from the free energy minimum to the center of the bifurcated SHB, its $S_{tot}$ decreases by 65% and the cancellation between the covalent and hydrogen bonding contributions increases to 91%. From Figure 6.8b, the $n_p \rightarrow s_H^*$ charge transfer also occurs when the proton is symmetri-
Table 6.2: Chemical shifts and isotropic magnetic shielding constants of the hydrogen bonded proton in two configurations of TAPH and DMANH. $r_1$ represent the distance between the N$_1$ and H atom in the hydrogen bonds. For TAPH, $S_{HB,12}$, $S_{HB,13}$ and $S_{HB,23}$ are the shielding constants from the dimeric N$_1$–H···N$_2$, N$_1$–H···N$_3$ and N$_2$–H···N$_3$ hydrogen bonds, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$r_1$ (Å)</th>
<th>$\delta_H$ (ppm)</th>
<th>$S_{tot}$ (ppm)</th>
<th>$S_{cov}$ (ppm)</th>
<th>$S_{HB}$ (ppm)</th>
<th>$S_{HB,12}$ (ppm)</th>
<th>$S_{HB,13}$ (ppm)</th>
<th>$S_{HB,23}$ (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAPH</td>
<td>1.1</td>
<td>20.2</td>
<td>11.4</td>
<td>52.0</td>
<td>40.6</td>
<td>22.4</td>
<td>23.1</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>27.6</td>
<td>4.0</td>
<td>47.0</td>
<td>43.0</td>
<td>24.4</td>
<td>23.6</td>
<td>24.6</td>
</tr>
<tr>
<td>DMANH</td>
<td>1.1</td>
<td>18.8</td>
<td>12.8</td>
<td>41.0</td>
<td>28.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>23.2</td>
<td>8.4</td>
<td>40.5</td>
<td>32.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

cally shared in the bifurcated hydrogen bond. Compared to the TAPH configuration at the free energy minimum, the overlap between the antibonding $s^*_H$ of the H atom and $n_p$ of the N atoms becomes more significant and the deshielding effects arising from the bifurcated and dimeric hydrogen bonds both increase. As a reference, we find a configuration of DMANH from the AI-PIMD simulations, which has $r_1 = r_2 = 1.3$ Å and hence the proton equally shared between the two N atoms. As shown in Table 6.2, we again find its $S_{HB}$ of 32.1 ppm to be larger than the dimeric $S_{HB}$ of TAPH (about 24 ppm), but smaller than the total $S_{HB}$ (43.0 ppm) of the bifurcated SHB.

Therefore, like the linear SHBs, the downfield $^1$H NMR chemical shifts of the bifurcated SHB of TAPH come from a balance between the covalent and hydrogen bonding interactions, which counteract each other. The magnetic shielding effects of these interactions depend strongly on the geometry of the bifurcated hydrogen bond, in particular the proton position relative to the three N atoms, which can be well characterized by the $r$ and $\theta$ parameters. This has led to the universal relation as observed in Figure 6.6. By analyzing two representative configurations of TAPH and DMANH from the AI-PIMD simulations, our calculated $S_{HB}$ values suggest that the N–H···N interaction is weaker in the bifurcated SHB due to its lack of linearity in the geometric arrangement. However, combining these N–H···N pairs, the overall interaction is stronger in the bifurcated SHB of TAPH than the linear SHB of DMANH. To further validate this assessment, we perform the AIM analy-
sis on the optimized structures of TAPH and DMANH in vacuum, and find one (3,-1) bond critical point for the \( N_1-H \cdots N_2 \) hydrogen bond in DMANH and two critical points for the \( N_1-H \cdots N_2 \) and \( N_1-H \cdots N_3 \) interactions in TAPH. Using Equation 6.1, we estimate the hydrogen bond energy to be 27.6 kcal/mol for DMANH, and 15.3 kcal/mol for each of the \( N_1-H \cdots N_2 \) and \( N_1-H \cdots N_3 \) interactions in TAPH. While this calculation only includes electronic quantum effects and is performed without the solvent environment, the estimated hydrogen bond energies show that the \( N-H \cdots N \) interaction in the bifurcated SHB of TAPH is weaker than that in the linear SHB in DMANH. However, if one assumes these energies are additive, the bifurcated SHB becomes more stable than its linear counterpart, in good agreement with our findings from the NMR calculations.

### 6.4 Conclusions

In this chapter, we carry out first principles simulations to uncover the geometry, competing quantum effects and \( ^1H \) NMR chemical shifts of a bifurcated SHB in the macrocyclic compound TAPH. In AIMD simulations, the proton remains bonded to a N atom for most of the time and electronic quantum effects enable occasionally proton transfer to occur. The resulting distribution of the proton positions exhibits a three-fold symmetry, consistent with the triangular arrangements of the N atoms in the bifurcated SHB. Explicit inclusion of nuclear quantum effects in AI-PIMD simulations makes the proton delocalized in the bifurcated SHB, and it takes less than 2.3 kcal/mol to move the proton from a free energy minimum to a shared position between two or three N atoms. In both cases, the behavior of the bifurcated SHB deviates significantly from a standard description of hydrogen bonds as classical electrostatic interactions. By analyzing the average QKE of the hydrogen bonded proton, we further elucidate the distinct features of competing nuclear quantum effects in the bifurcated SHB as compared to its linear counterpart. The quantum nature of the SHB manifests as a \( ^1H \) NMR chemical shift that is larger than 20 ppm. This remarkably downfield chemical shift arises from the opposite shielding effects of the covalent and hydrogen
bonding interactions, which give rise to a universal correlation between the $\delta_H$ values and the geometric parameters of the bifurcated SHB.

Compared to a two-center SHB, the bifurcated SHB in TAPH adopts a triangular shape and possesses a Donor-H-Acceptor angle that is far from linearity. These geometric features suggest that TAPH and DMANH might have distinct hydrogen bond energies. Combining DFT calculations and the AIM analysis in the gas phase, we show that each N-H···N interaction is weaker in the bifurcated SHB, but their sum gives a more stable hydrogen bond in TAPH as compared to the linear SHB in DMANH. However, as the donor and acceptor groups are in the same molecule, there are more than one possible definitions for the energy of an intramolecular hydrogen bond.$^{298,305,306}$ In addition, due to the close contact between the N atoms, one has to incorporate both electronic and nuclear quantum effects and properly account for the condensed phase environment, making it difficult to accurately compute the energy of the SHBs using quantum chemistry methods. As an alternative approach, we take the deshielding effects in the $^1$H NMR chemical shifts as a measure of the hydrogen bond strength and divide TAPH into monomeric and dimeric structures to examine the degree of deshielding due to the two-center and three-center hydrogen bonding interactions. After taking two representative configurations of TAPH and DMANH, one at a free energy minimum and the other with the proton shared in the SHB, from their corresponding AI-PIMD simulations, we reach the same conclusion that the bifurcated SHB becomes stronger than its linear counterpart when both of its N-H···N interactions are taken together. Given that bifurcated SHBs are frequently observed in condensed phase systems, our results provide useful benchmark values for future experimental and computational studies of the structure, energetics and $^1$H NMR chemical shifts of these compact structures.
CHAPTER 7
SHORT HYDROGEN BONDS IN ACID-BASE MIXTURES

This chapter is a reprint of Ref. [P5]

7.1 Introduction

The Brønsted-Lowry theory, proposed in 1923, establishes that an acid and a base react by exchanging a proton to form their conjugated base and acid.\textsuperscript{307,308} This theory has been widely used in science to predict and explain the chemistry when mixing an acid and a base. The acid-base reaction is typically represented by the following chemical equation,

\[
HA + B \rightleftharpoons A^- + HB^+ \quad (7.1)
\]

where HA and B are the acid and base, respectively, and A\(^-\) and HB\(^+\) their conjugated forms. The double arrow in Equation 7.1 denotes the presence of chemical equilibrium between the species, which is strictly defined by the thermodynamics of the reaction.\textsuperscript{309,310} Hence, according to the Brønsted-Lowry theory acid-base theory, a solution containing an acid and a base can only exist as a mixture of the four components, where the ratio between the reactants and products is defined by the equilibrium constant. It is possible to argue that the Brønsted-Lowry theory should only be used in the context of aqueous solutions, and instead, for mixtures of other solvents the reaction should be expressed in terms of the autoprotolysis of the components,

\[
A^- + HB^+ \rightleftharpoons HA + B \quad (7.2)
\]
where HA and B are the conjugate acid and base of the ionic species, respectively. For a fully ionized solution (ionic liquid), the equilibrium constant is expressed as \( K_S = [HA][B] \).\(^{311}\)

While this formalism can be used to predict the ionic conductivity of many molecular liquids and solutions, it fails to explain the cases such as the mixture of acetic acid and 1-methylimidazole.\(^ {312}\) Acetic acid is a weak acid with a pKa of 4.75, and 1-methylimidazole is a weak base with a pKb of 6.99\(^ {313}\) and their mixtures present high ionic conductivity comparable to strong acids.\(^ {312,314}\) However, no ionization reaction is observed (left direction in Equation 7.2), since neither imidazolium nor acetate ions are present in their mixture.\(^ {312}\)

In this chapter, we collaborate with the group of Professor Daniel Kuroda at the Department of Chemistry, Louisiana State University, and report the formation of a new chemical species when mixing acetic acid and 1-methylimidazole. In this new chemical species, the hydroxyl group of the acetic acid directly interacts with the unsubstituted nitrogen atom of the imidazole ring, forming an acid-base hydrogen bonded complex (top right panel of Figure 7.1). Notably, the hydrogen bond in the acid-base pair is not conventional because it presents a significantly reduced distance between the donor and acceptor atoms, and hence, a different potential energy surface in which the acidic hydrogen atom participating in the hydrogen bond is delocalized (Figure 7.1). While hydrogen atom delocalization has often been observed in short and strong hydrogen bonds in the active site of enzymes\(^ {20,88,123,315}\) and excess proton and conformationally rigid molecules in solutions,\(^ {62,155,316–318}\) the hydrogen bond complex observed in this chapter is substantially different from those systems because the donor and acceptor are both molecules in solution where entropy plays a major role in determining the energetics of the system. Moreover, the presence of a molecular species having a delocalized hydrogen atom challenges the conventional knowledge of acid base chemistry.
Figure 7.1: Potential energy surfaces along the O-H stretch coordinate in an acetic acid dimer and an acid-base complex. In the top panels, silver, red, blue and white represent C, O, N and H atoms, respectively. The dotted line shows a conventional hydrogen bond, and the orange spheres represent the ring polymer beads of the delocalized proton. In the bottom panels, $|0\rangle$ and $|1\rangle$ represent the ground and first excited vibrational states of the system, respectively.

7.2 Methods

This is a collaborative work, and all the experimental measurements and AIMD simulations have been done by the group of Professor Daniel Kuroda, and I contribute to the work by performing AI-PIMD simulations, computing free energy surface and calculating spectral properties. The geometry optimization, vibrational frequency calculations and potential energy surface calculations are performed in vacuum at the PBE level\cite{319} with the 6-311++G basis set using the Gaussian 16 software.\cite{320} Initial configurations are built by
Avogadro with the classical force field MMFF94. AIMD simulations are performed for the mixtures of acetic acid and 1-methylimidazole at five concentrations using the CP2K package. AI-PI-MD simulations are performed for the solution with $X_{HA}$ of 0.67 using CP2K for the electronic structure evaluation and the i-PI software for propagating the nuclear motion. The electronic structure is described using the PBE density functional and the D2 dispersion correction. The Goedecker-Teter-Hutter pseudopotentials are used for the core electrons, and the TZV2P-GTH plane-wave basis set with a cut-off of 400 Ry is used to describe the valence charge density. Periodic boundary conditions are applied for all systems. The simulations are performed with a time step of 0.5 fs in the canonical ensemble at 321 K. In the AIMD simulations, the Nosé-Hoover thermostat is used to control the temperature. In the AI-PI-MD simulations, each atom is represented by 6 ring polymer beads using the path integral generalized Langevin equation method. The AIMD simulations are performed for a total of 300 ps for the mixtures with $X_{HA}$ of 0.02, 0.67 and 0.86. The simulation lengths are 140 ps and 150 ps for the systems with $X_{HA}$ of 0.99 and 1.00, respectively. The AI-PI-MD simulations are performed for 50 ps. The two-dimensional free energy surfaces for transferring the proton in a hydrogen bond are computed from the first principles simulations. For a hydrogen bond between two acetic acids, $R$ is calculated as the O-O distance and $\nu$ as the difference between the two O-H distances, $\nu = d_{O1H} - d_{O2H}$. For a hydrogen bond between acetic acid and 1-methylimidazole, $R$ is the O-N distance and $\nu = d_{OH} - d_{NH}$. Here, a pair is considered as a hydrogen bond if $R \leq 3.5 \text{ Å}$ and the O-H-O or O-H-N angle is equal to or above 135°. The free energies are computed as $F = -k_B T \ln P(\nu, R)$, where $k_B$ is the Boltzmann constant, $T$ is the simulation temperature, and $P(\nu, R)$ is the probability of observing a hydrogen bond with the hydrogen atom at $\nu$ and the O-O or O-N distance at $R$.

AIMD simulations are performed for the mixtures of acetic acid and 1-methylimidazole at five concentrations of acetic acid ($X_{HA}$). The systems are placed in cubic boxes with the application of periodic boundary conditions, as shown in Table 7.1.
Table 7.1: The number of molecules and the box size in the systems with different concentrations of the acetic acid ($X_{HA}$).

<table>
<thead>
<tr>
<th>$X_{HA}$</th>
<th>0.02</th>
<th>0.67</th>
<th>0.86</th>
<th>0.99</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of acetic acid</td>
<td>1</td>
<td>42</td>
<td>60</td>
<td>70</td>
<td>72</td>
</tr>
<tr>
<td>Number of 1-methylimidazole</td>
<td>51</td>
<td>21</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Box length (Å)</td>
<td>19.0</td>
<td>19.0</td>
<td>19.3</td>
<td>18.9</td>
<td>19.2</td>
</tr>
</tbody>
</table>

The IR spectrum is calculated from the AIMD simulations of the acid-base mixture with $X_{HA}$ of 0.67. The IR spectrum is obtained from the Fourier transform of the molecular dipole time correlation function,

$$ I(\omega) = Q \int_{-\infty}^{\infty} e^{-i\omega t} \langle \vec{\mu}(0) \cdot \vec{\mu}(t) \rangle dt $$

(7.3)

Here $\vec{\mu}(t)$ is the molecular dipole moment at time $t$, and $Q$ is a quantum correction factor to the formula since the AIMD simulations treated the nuclei classically. The molecular dipole moments are obtained using the maximally localized Wannier function scheme,\textsuperscript{326} and we compute the Wannier centers every 2.5 fs from the AIMD trajectories using the CP2K software.\textsuperscript{247,248} This scheme also allows us to decompose the overall IR spectrum into the contributions from the acetic acid and 1-methylimidazole molecules. The IR spectra of acetic acid, 1-methylimidazole and their mixture are calculated using the TRAVIS software.\textsuperscript{326–328} A harmonic quantum correction factor, $Q = \frac{\beta \hbar \omega}{1 - e^{-\beta \hbar \omega}}$, where $\beta$ is the inverse temperature, is then applied to obtain the final spectra.\textsuperscript{329} The average $^1$H NMR chemical shifts are also calculated from the AIMD simulations according to the proton position-chemicallshift relation in Chapter 5, where $\langle \nu \rangle$ is the average proton position in a short hydrogen bond.\textsuperscript{290} From the AIMD simulations, we collect $\nu$ of all the hydrogen bonds every 0.5 fs and use the values between -1.0 Å and 1.0 Å, which belongs to shared hydrogen atoms in short hydrogen bonds, to obtain $\langle \nu \rangle$. The average chemical shifts are then computed for each concentration of the acid-base mixture.
7.3 Results and discussions

The lack of a chemical reaction between acetic acid and 1-methylimidazole is readily seen in the IR spectra of the acid-base mixture (Figure 7.2a). Even at very high concentrations of the base ($X_{HA} = 0.1$), the band corresponding to the asymmetric stretch of acetate, usually located at around $1560 \text{ cm}^{-1}$, is not observed, in agreement with previous studies on the system. In contrast, the addition of acetic acid to 1-methylimidazole leads to the appearance of at least one very broad band between $2200 \text{ cm}^{-1}$ and $3500 \text{ cm}^{-1}$ (blue and red zones in Figure 7.2a). This broad band is very visible in the IR spectrum at any molar fractions of the acid or the base, but is absent in the pure acid or base. The intensity of the new IR peak presents a strong temperature dependence (i.e., they decrease when the temperature increases) and a linear dependence with the concentration of the acid. In addition, a substitution of the acidic hydrogen in acetic acid with deuterium results in a broad band at $1900 \text{ cm}^{-1}$, showcasing the isotopic dependence of the transition frequency of the bands. The dependence of the IR features with temperature, acid concentration and the isotopic nature of acidic hydrogen demonstrates that the broad IR peaks observed in the acid-base mixtures are not mixed states as previously proposed, but the vibrational signature of a one-to-one acid-base complex formed between the acetic acid and 1-methylimidazole (Figure 7.1). This acid-base complex is not only a new chemical species with specific IR signatures, but also is thermodynamically more stable at room temperature than the acid, the base, and their conjugated pairs as demonstrated by the lack of chemical reaction. These observations are in agreement with previous studies on the same system, which identify that there is no acid-base reaction and that the mixture has an excess enthalpy of around $-7 \text{ kJ/mol}$.

The hypothesis that the experimental IR features originate from the formation of an acid-base complex is further validated by ab initio molecular dynamics (AIMD) simulations. These simulations reveal the presence of many different hydrogen bonded structures
Figure 7.2: Spectroscopic and theoretical features of the complex formed from acetic acid and 1-methylimidazole. (a) ATR-FTIR spectra of the mixture of acetic acid and 1-methylimidazole for different molar ratios of acetic acid ($X_{HA}$). (b) Radial distribution functions between the oxygen atom of the hydroxyl group in acetic acid and the unsubstituted nitrogen atom of the imidazole ring (red highlighted atoms of the cartoon) from AIMD simulations at different concentrations. (c) Distribution of the rotation angles between the acid and the base from AIMD simulations. (d) IR spectrum in the 2000 – 3500 cm$^{-1}$ region derived from AIMD simulations ($X_{HA} = 0.67$) and the contributions of the different components: acetic acid (HAc) and 1-methylimidazole (MIm).

in the acid-base mixtures. However, in all investigated concentrations, 1-methylimidazole always forms a one-to-one hydrogen bonded complex with acetic acid (Figure 7.1) even when the acid is in very low or high concentrations. As expected, the simulations confirm that the complex is formed by the direct interaction of the hydroxyl group of the acid with the unsubstituted nitrogen of the imidazole ring via a hydrogen bond (Figure 7.2b). Notably, it is observed that the acid-base pair forms a hydrogen bond with a distance shorter than 2.7 Å, irrespective of the molar fraction of the acid (Figure 7.2b), showcasing that the
acid-base complex has a short hydrogen bond at all concentrations. While this hydrogen bond keeps the participating molecules at short distances, the complex appears to have a large number of possible conformations stemming from the almost free rotation of the acid and the base around the hydrogen bond (Figure 7.2c). In addition, the hydrogen bonded complexes are not isolated, but generally participate in extended hydrogen bond networks with up to three other acid or base molecules at very high or low concentrations of the acid, respectively.\textsuperscript{336}

The assignment of the IR bands to the formation of the acid-base complex is confirmed by the IR spectrum computed from the AIMD simulations of the mixture with $X_{HA} = 0.67$. Consistent with the experimental results (Figure 7.2a), the theoretical IR spectrum presents at least one band in the 2200-3500 cm\textsuperscript{-1} region (Figure 7.2d). The decomposition of the simulated IR spectrum shows that the spectral components below 2750 cm\textsuperscript{-1} originate from 1-methylimidazole and the high frequency component at 3100 cm\textsuperscript{-1} arises from acetic acid (Figure 7.2d). This assignment validates our hypothesis that the formation of the 1-methylimidazole-acetic acid complex gives rise to the IR bands seen experimentally in the 2200-3500 cm\textsuperscript{-1} region of the IR spectra. It is important to note that IR spectrum in the AIMD does not contain mixed states, which strengthens the hypothesis that the IR signatures in the 2200-3500 cm\textsuperscript{-1} region arise from the complex. Furthermore, the strong resemblance between the simulated and the experimental IR spectra corroborates the molecular structure of the complexes captured by AIMD simulations because the vibrational modes are very sensitive to their local molecular environment.\textsuperscript{337} Thus far, the experimental and theoretical data evidence the formation of short and strong hydrogen bonds as a result of the formation of the complex between acetic acid and 1-methylimidazole, which are observed through their spectral signatures in the IR region.

Recently, the link between short hydrogen bonds and the appearance of broad vibrational bands at lower frequencies (1000-2000 cm\textsuperscript{-1}) than the hydroxyl stretch (3500 cm\textsuperscript{-1}) is demonstrated for the excess proton in solution.\textsuperscript{338–342} For example, a study on acid aque-
ous solutions reveals that molecular species, such as the Zundel cation, present broad IR peaks in the same spectral regions arising from vibrational modes directly linked to the shuttling of the excess proton.\textsuperscript{343,344} Hence, the broad bands in the IR spectra (Figure 7.2a) are attributed to the acidic hydrogen shuttling modes of the complex upon the formation of a short hydrogen bond between acetic acid and 1-methylimidazole. This relation is further confirmed from \textit{ab initio} frequency calculations of the vibrational modes for 6 representative configurations (Figure 7.3a) deduced from the AIMD simulations. The computed vibrational modes show that all of the complexes exhibit strong vibrational absorptions in the 2200-3000 cm\textsuperscript{-1} region (Figure 7.3a), in agreement with the experimental and simulated IR spectra (Figure 7.2a and c). Moreover, all these vibrational modes involve atomic motions directly relate to the shuttling of the acidic hydrogen atom in the acetic acid molecule (Figure 7.3b). The vibrational mode computations also reveal the presence of two distinct groups. The first group corresponds to the hydrogen shuttling modes of the acidic hydroxyl group directly interacting with 1-methylimidazole and has transition frequencies below 2500 cm\textsuperscript{-1} (blue lines in Figure 7.3a). In contrast, the second group has transition frequencies between 2500 cm\textsuperscript{-1} and 3100 cm\textsuperscript{-1} (red lines in Figure 7.3a) and is associated to hydrogen shuttling modes that involve exclusively acetic acid molecules further along the chains of the hydrogen bond network (Figure 7.3a). Remarkably, the vibrational modes with displacements of the hydrogen atom parallel to the hydrogen bond (Figure 7.3b) have very large transition dipoles (greater than 2000 km/mol), while the shuttling modes with perpendicular displacements of the hydrogen atom have more typical transition dipoles (\textasciitilde 500 km/mol).\textsuperscript{337} Normal mode computations also reveal that the broadness of the hydrogen shuttling bands (Figure 7.2a) arises from the dependence of the vibrational transition frequency with the geometry of the complex. More specifically, the hydrogen shuttling frequency is strongly dependent on the angle between the planes of the hydrogen bond donor and acceptor (Figure 7.2) due to a concomitant shortening of the complex hydrogen bond. As a result, the large number of angular conformations adopted by the acid-base pair...
causes the hydrogen shuttling mode to span in a 250 cm$^{-1}$ region. Hence, the transition dipole and frequency of the hydrogen shuttling modes in the acid-base complex make these modes easily observed via IR spectroscopy, and the strong dependence of the acidic hydrogen shuttling transition frequencies on the geometrical arrangement of the complex along with its thermally allowed distribution of angles give rise to the large bandwidth of these shuttling modes.

![Figure 7.3](image)

Figure 7.3: Ab initio frequency calculations of acid-base representative complexes. (a) Representative structures of the complexes composed of acetic acid hydrogen bonded to 1-methylimidazole and their IR vibrational modes from DFT calculations. Left and right panels correspond to two different angular conformations of the complex as depicted on the top. (b) Atomic displacement for the proton shuttling mode of the complex. The left, center and right modes have an associated frequency of ca. 1585 cm$^{-1}$, 1696 cm$^{-1}$ and 2211 cm$^{-1}$, respectively.

Previous studies on the excess proton in aqueous solutions demonstrate that the broad vibrational bands in the IR spectra are directly associated with the changes in the potential energy surface of the proton shuttling coordinate.$^{69,341}$ The same effect is observed here for
the potential energy surfaces of the hydrogen shuttling coordinate in both the acid dimer and the acid-base complex (Figure 7.4a and b). In the case of the acetic acid dimer, the potential energy surface has a single minimum with a curvature of 504 kcal/(Å²·mol). In contrast, when acetic acid forms a hydrogen bond with 1-methylimidazole, the potential energy surface shows a second minimum with a low barrier of 5 kcal/mol for transferring the proton between the donor and acceptor atoms, and the curvature in its global minimum reduces to 192 kcal/(Å²·mol). While the associated frequency for the hydrogen shuttling changes from 3550 cm⁻¹ in the acid dimer to 2190 cm⁻¹ in the acid-base complex, or to 1590 cm⁻¹ upon deuteration, and evidences the variations in the curvature of the potential energy surfaces, the shape of the potential surfaces explain why the acid-base complex has a special kind of hydrogen bond. Unlike the case of the acetic acid dimer (Figure 7.4a), the acidbase pair (Figure 7.4b) presents a hydrogen transfer barrier (5 kcal/mol) comparable to the zero point energy of the O-H stretch mode (4 kcal/mol), which allows the delocalization of the hydrogen atom in the potential well. Moreover, the hydrogen delocalization is heightened by hydrogen bonding to an additional acetic acid, as seen in the almost barrier less (1 kcal/mol) potential energy surface of the hydrogen shuttling for such complex (Figure 7.4c).

The special characteristics of the hydrogen bond in the acid-base pair are also evaluated from the free energy surfaces of proton shuttling from the AIMD simulations. In the absence of a hydrogen bonded 1-methylimidazole, the shuttling of acidic hydrogen in an acid dimer (HAc·HAc, Figure 7.4d) has two local minima with a barrier of 3.6 kcal/mol, indicating that the proton is positioned close to one of the two acid molecules forming a normal hydrogen bond. In contrast, the free energy landscape for the shuttling of acidic hydrogen is drastically altered when 1-methylimidazole is part of the hydrogen bond network. In this case, the barrier for proton shuttling (MIm·H·Ac of Figure 7.4e) is three times lower (1.1 kcal/mol) than in the pure acid. Consequently, the acidic hydrogen in the acid-base complex has a much larger probability of being delocalized within the hy-
Figure 7.4: Energy profiles for shuttling the acidic hydrogen for different hydrogen bonded complexes. (a), (b) and (c) show the potential energy surfaces for the three hydrogen bonded pairs derived from ab initio computations, and (d) and (e) display the free energy surfaces for the different pairs calculated from the AIMD and AI-PIMD simulations of the pure acid and the solution with $X_{HA}$ of 0.67, respectively. The free energy surfaces are calculated along the proton sharing coordinate $\nu$ and the hydrogen bond length $R$, and are shifted so that their minimal values are 0.

The hydrogen shuttling potential in the acid-base complex is almost barrier-less (Figure 7.4e). Although a similar decrease in the barrier is observed for the other hydrogen bond pairs, the hydrogen shuttling between two acetic acid molecules still maintains its double well shape even in
the mixture. These results demonstrate that quantum mechanical effects play a significant role in the energetics of the acid-base complex as they lower the hydrogen shuttling barrier, and consequently, favor the delocalization of the hydrogen atom within the hydrogen bond. The energetic stabilization of chemical species via nuclear quantum effects is not new and has been proposed to play a major role in the formation of Watson-Crick base pairs.\(^{168}\) The presence of a significant non-classical effect in the acid-base complex arises from the quantum kinetic energy penalty as a result of localizing a quantum particle, such as the hydrogen atom.\(^ {30}\) Hence, a decrease of the energy barrier in the hydrogen shuttling potential not only results in a reduction of the energy penalty in the system, but also leads to a substantial expansion of the region where the hydrogen atom can be positioned; i.e., delocalization.

The delocalization of the acid hydrogen in the acetic acid and 1-methylimidazole mixtures is also deduced from the \(^1\)H NMR spectra of the different solutions, since one of the most distinctive features of delocalized hydrogen atoms in short hydrogen bonds is their far downfield \(^1\)H NMR chemical shifts.\(^ {30}\) Typically, the \(^1\)H chemical shift for the hydroxyl group of an acid is around 12 ppm.\(^ {345}\) However, when an acidic hydrogen atom is delocalized within a strong hydrogen bond, the \(^1\)H nucleus is further deshielded and results in a \(^1\)H chemical shift well above 14 ppm or even close to 20 ppm.\(^ {21,76}\) This extreme behavior in the chemical shift of the delocalized hydrogen is also found in the acetic acid and 1-methylimidazole samples, where the addition of 1-methylimidazole shifts the chemical shift of the acid proton from 12 ppm in the pure acid to 15-16 ppm in the mixture (Figure 7.5). It is important to note that the observed chemical shift is an average of the chemical shift of the possible species, in this case the delocalized and non-delocalized acidic hydrogen atom, due to the ultrafast times scale of the chemical exchange of the process compared to the NMR time resolution.\(^ {346}\) The observed chemical shifts are in good agreement with the formation of species where the hydrogen atom is delocalized, or equivalently, deshielded. Mixtures containing deuterated acetic acid also displays a large \(^2\)H chemical
shifts of the deuterium atom, though they are slightly lower than the corresponding \(^1\)H samples (Figure 7.5). The large chemical shift of the \(^2\)H nucleus confirms the delocalization of the acidic deuterium atom, where their lower chemical shift as compared to \(^1\)H evidences the quantum mechanical effect driven exclusively by a change in the zero point energy rather than changes in the interaction potential between the deuterium atom and the carboxylate group\(^5\). Computation of the NMR chemical shift for the acidic hydrogen atom in different mixtures confirm the large downfield shift in the \(^1\)H NMR region of the acid hydrogen seen in the experimental data (Figure 7.5). Hence, the NMR data further validates our hypothesis that the hydrogen delocalization occurs in the mixture of acetic acid and 1-methylimidazole because of the presence of a strong hydrogen bond in the acid-base complex.

Figure 7.5: Experimental and computational \(^1\)H NMR chemical shifts of the hydrogen bonded proton as a function of the concentration of the acid in the mixtures of acetic acid and 1-methylimidazole. The experimental \(^2\)H chemical shifts of the acetic acid-1d and 1-methylimidazole are also included for comparison.

In summary, in this chapter, we showcase a new acid-base chemistry, not predicted by the classical theories, as a result of the direct and strong interaction between acetic acid and 1-methylimidazole. Unlike typical hydrogen bonded complexes, the interaction between these acid-base pair gives rise to the formation of a short hydrogen bond, in which
the shared hydrogen atom is delocalized. This new species has distinct broad vibrational signatures in the 2250-3000 cm$^{-1}$ region arising from the shuttling vibrational mode involving the hydroxyl hydrogen atom. In particular, the IR features evidence substantial changes in the potential energy surface associated with the formation of a strong hydrogen in the acid-base complex. The presence of different potential energy surfaces associated with the new species is confirmed using AIMD and AI-PIMD simulations. More importantly, it is established from the simulations that quantum mechanical effects play a significant role in defining the structure and energetic of the complex formed from acetic acid and 1-methylimidazole. Evidence of the quantum mechanical states and hydrogen delocalization of the complex is directly deduced from the NMR chemical shifts of the mixtures and corroborated via computational chemistry. Overall, the present study shows that quantum mechanical effects play significant roles in the chemistry of conventional substances, such as acid and bases.
CHAPTER 8
SHORT HYDROGEN BONDS IN THE AQUEOUS SOLUTION OF BIFLUORIDE IONS

8.1 Introduction

The bifluoride ion, [F···H···F]−, has a linear and symmetric structure and is likely the shortest and strongest hydrogen bond. Its F-F length is as low as 2.28 Å in the crystal structure of p-toluidine bifluoride and is associated with an extremely high bond strength. Both experimental and theoretical studies have suggested that the strength of the bifluoride hydrogen bond ranges from 30 to 60 kcal/mol, which is much larger than that of typical hydrogen bonds and closer to that of covalent bonds. From the perspective of molecular orbital theory, the bifluoride hydrogen bond has been considered as a three-center four-electron bond, rather than a usual hydrogen bond. Given the proximity of the two fluoride atoms, the bifluoride ion has a single minimum potential for transferring the proton between the two F atoms, which results in many interesting spectral signatures.

The normal F-H stretch frequency is around 4000 cm⁻¹, for example, 3962 cm⁻¹ in liquid hydrogen fluoride. However, the symmetric stretch, antisymmetric stretch and bending modes of the bifluoride ion have the frequencies of ca. 600 cm⁻¹, ca. 1425 cm⁻¹ and ca. 1230 cm⁻¹, respectively in the IR and Raman spectra of the potassium bifluoride crystal. When potassium bifluoride is solvated in aqueous solutions, the frequency of the F-H-F bending mode shifts from 1233 cm⁻¹ to 1206 cm⁻¹ and the frequency of the F-H-F asymmetric stretch mode shifts from 1473 cm⁻¹ to 1536 cm⁻¹. The red-shift is often considered a spectral signature of SHBs. All the geometric features and spectral properties make the bifluoride ion a good model for studying SHBs.

Recently, the research group of Professor Andrei Tokmakoff in the Department of
Chemistry of the University of Chicago has measured the IR spectrum of potassium bifluoride in aqueous solutions. As shown in Figure 8.1, the IR spectrum shows a broad O-H stretch absorption around 3500 cm$^{-1}$, which is accompanied by both red-shifted and blue-shifted features as compared to that of liquid water. The frequency shifts are proposed to be related to the O–H–F hydrogen bonds between the bifluoride ion and water and the disturbed hydrogen bond network of liquid water. As a comparison, the IR spectra of the fluoride ions in aqueous solutions often show a red-shifted O-H stretch frequency when compared to liquid water, indicating a strong water-ion interaction. For example, the O-H stretch frequency of water shifts from 3421 cm$^{-1}$ in the liquid phase to 3412 cm$^{-1}$ in the solution of sodium fluoride with a mole fraction of 0.015.\textsuperscript{368} In this Chapter, we use AIMD simulations and electronic structure calculations to study the solvation structures and spectral properties of the solutions of potassium bifluoride and potassium fluoride.

![Figure 8.1: IR spectra of (a) water (blue) and the potassium bifluoride solution (red).](image)

8.2 Methods

We carry out computational modeling on five systems: liquid water, low-concentration (1.43 M) and high-concentration (3.61 M) KF solutions, low-concentration (1.43 M) and
high-concentration (3.61 M) KHF$_2$ solutions. We place 3 and 10 solute molecules in cubic water boxes to generate the low and high concentration solutions, respectively. Each solution contains 111 to 139 water molecules and the length of the simulation box ranges from 15 to 17 Å. We solvate the solute molecules using the Amber 2016 software package, with the potential energies described using the TIP3P water model and the Joung/Cheatham ion parameters.$^{120,241,242}$

We perform AIMD simulations of liquid water and the aqueous solutions using the CP2K package.$^{247}$ The electronic structure is described using the revised PBE density functional (revPBE)$^{369}$ with the D3 dispersion correction.$^{128}$ The core electrons for all the atoms are described using the Goedecker-Teter-Hutter pseudopotentials.$^{252}$ The valence charge density is modeled by the DZVP-GTH plane-wave basis set for the H, O, F atoms and the DZVP-MOLOPT-SR-GTH plane-wave basis set for the K atoms with a cutoff of 400 Ry.$^{323}$ All the simulations are carried out in the canonical ensemble with a time step of 0.5 fs, and the Nosé-Hoover thermostat chain is applied to control the temperature at 294.15 K.$^{324}$ After equilibrating each systems for 10-50 ps, we carry out 80 ps of AIMD simulations.

The potential energy surface for proton sharing in the bifluoride ion is calculated by scanning the F-H bond length and optimizing the position of the hydrogen atoms with the F and O fixed at each step. The electronic structure calculations are performed using the TeraChem software package$^{125,126}$ with the B3LYP density functional,$^{127}$ the D3 dispersion correction$^{128}$ and the aug-cc-pVDZ basis set.$^{179}$

We use the TRA VIS software to analyze the trajectories from the AIMD simulations and calculate the IR spectra of liquid water and the aqueous solutions.$^{326–328}$ Here, we define a hydrogen bond when R ≤ 3.5 Å and H–D–A angle ≤ 30°.$^{370}$ From the simulations, we compute the autocorrelation functions of the water-ion hydrogen bonds and calculate the spatial number density of water molecules surrounding the bifluoride ions to get the spatial distribution functions. In addition, we compute the molecular dipole moments using the maximally localized Wannier function scheme,$^{326}$ and calculate the Wannier centers every
2.5 fs from the AIMD trajectories using the CP2K software. A harmonic quantum correction factor, $Q = \frac{\beta \hbar \omega}{1 - e^{-\beta \hbar \omega}}$ is applied to correct for nuclear quantum effects in the IR spectrum calculation. A polynomial function in the range of 3000 cm$^{-1}$ to 3750 cm$^{-1}$ is applied to smooth the curve for the final IR spectra. To decompose the overall IR spectrum into the contributions from water molecules in the first and second solvation shells of the F atoms, we collect a series of trajectories with time period longer than 10 ps from the AIMD simulations, in which one water molecule is confirmed to stay in the same solvation shell during that period. We calculate the decomposed IR spectrum of that water molecule and average over all the trajectories to obtain the final decomposition.

8.3 Results and discussion

8.3.1 Geometry of the F-H-F hydrogen bond

It has been observed that the proton is centered between two F atoms in the crystal structure of the bifluoride ion. In the AIMD simulations of the KHF$_2$ solutions, we find that the averaged R between the two F atoms is 2.36 Å and 2.35 Å in the low- and high-concentration solutions, respectively. The hydrogen bond is nearly linear, with an average F–H–F angle of 168.2° and 168.4° in the low- and high-concentration solutions, respectively. Besides R, we will use the collective variables $\nu$ and the proton XY-plane coordinates X and Y to describe the geometry of the bifluoride hydrogen bond. We define $\nu = d_{F_1H} - d_{F_2H}$ to characterize the proton position in the hydrogen bond, where $d_{F_1H}$ and $d_{F_2H}$ are the distances from the proton to the two equivalent F atoms. Therefore, $\nu = 0$ means that the proton is equidistant between the two F atoms and a symmetric F–H–F hydrogen bond. We compute the probability, $P(R, \nu)$, for finding the hydrogen bond at length R and the proton at position $\nu$ and calculate the free energy for proton movement as

$$ F = -k_B T \ln \frac{P(R, \nu)}{P_{\text{max}}} , $$

(8.1)
Here $k_B$ is the Boltzmann constant and $T$ is the simulation time. $P_{max}$ is the maximal probability of $P(R, \nu)$ and is included to ensure that the minimal free energy is 0. Given the similarity of the F–H–F hydrogen bonds in the low- and high-concentration solutions, we will take the high-concentration KHF$_2$ solution to demonstrate the results. Figure 8.2a shows that the proton moves in an apparent single-well potential when it shuttles between the two F atoms. Although the proton occasionally moves to be closer to one F atom, it is more likely to stay in the middle, which is consistent with the crystallographic result of the solid state (Figure 8.2a).

In order to quantify the linearity of the F–H–F hydrogen bond, we redefine an internal coordinate system with the middle point of the two F atoms as the origin and the unit vector along the $\overrightarrow{F_1F_2}$ direction as the Z-axis. The proton XY-plane coordinates X and Y are the X- and Y-axis coordinates of the proton in this internal coordinate system. When both X and Y are equal to 0, the F–H–F hydrogen bond is perfectly linear. We then calculate the normalized probability distribution, $P(X,Y)/P(X,Y)_{max}$, as its two-dimensional distribution of allows us to visualize how the proton movements deviate from linearity. As shown in Figure 8.2b, the probability decreases to below 10% when X or Y only deviates more than 0.18 Å from linearity, indicating that the F–H–F hydrogen bond is mostly linear in the KHF$_2$ solution.

### 8.3.2 Solvation structures of the bifluoride and fluoride ions

Adding ions to liquid water will lead to structural rearrangement of the water molecules and heterogeneity in different solvation shells. Both the F$^-$ and HF$_2^-$ ions are negatively charged and capable of forming O–H–F hydrogen bonds with the solvent molecules. To examine the microscopic solvation structures, we calculate the radical distribution functions (RDFs) between the O atoms in water and the F atoms in the ions, $g(r_{FO})$. As shown in Figure 8.3, the RDF curves have multiple peaks, and those of the low- and high-concentration solutions have similar features. For both the F$^-$ and HF$_2^-$ ions, the F-O distance ($r_{FO}$)
of the first solvation shell is around 3.4 Å, although the first peak of the F⁻ ion is much higher in intensity due to the stronger hydrogen bonding interaction between the ion and water. Accordingly, the number of water molecules within the first solvation ($r_{FO} \leq 3.4$ Å) is 5.5 and 3.7 for the low-concentration F⁻ and HF₂⁻ solutions, and 5.3 and 3.1 for the high-concentration solutions, respectively. The $r_{FO}$ values for the second solvation shells is around 5.4 Å for both the F⁻ and HF₂⁻ ions.

While the presence of the HF₂⁻ ions changes the structure of surrounding water molecules, the solvents also influence the properties of in return. We pick one configuration of the solvated HF₂⁻ ion from the AIMD simulations, which has the averaged values of $R = 2.35$ Å and F–H–F angle = 168.4° for the bifluoride ion to demonstrate the influence of the solvation environment. In this solute-solvent cluster, we include all the water molecules within the first solvation shell of each F atom in the HF₂⁻ ion, resulting in 3 and 4 water molecules around the F₁ and F₂ atoms, respectively. We intentionally include unequal numbers of water molecules around the two F atoms to represent the instantaneous asymmetric solvation structures of the bifluoride ion. As shown in Figure 8.4a, in the gas phase, the potential energy surface of proton movement in the HF₂⁻ ion shows a symmetric single-well feature and the energy minimum occurs when the proton is equidistant between the two F atoms.
However, the potential energy surface becomes asymmetric in the solute-solvent cluster, and the proton is closer to the F\textsubscript{1} atom which has fewer solvated water molecules (\(\nu = -0.26\ \text{Å}\)). The dynamics of the solvation shells creates the instantaneous difference in the number of water molecules hydrogen bonded to the two F atoms and promotes the shuttling of proton in the HF\textsubscript{2}\textsuperscript{−} ion. In addition, the reorientation and diffusion of water molecules lead to the forming and breaking of hydrogen bonds between the water and the ions. To quantify the time scale of this dynamical change, we calculate the autocorrelation function of the O–H–F hydrogen bond and measure its lifetime. Based on Luzar and Chandler’s definition for intermittent hydrogen bonds,\textsuperscript{370–372} the hydrogen bond lifetime depends on the population operator \(h(t)\), which is equal to 1 when a given pair is hydrogen bonded and 0 otherwise. The autocorrelation function, \(C_{HB}(t)\), is defined as

\[
C_{HB}(t) = \frac{\langle h(0)h(t) \rangle}{\langle h(0) \rangle}, \quad h(t) = \begin{cases} 1, & \text{if a pair is hydrogen bonded} \\ 0, & \text{if a pair is not hydrogen bonded} \end{cases}
\]

(8.2)

where \(h(t)\) is the population operator at time \(t\). We then calculate the integral of the auto-
correlation function, \( \tau_{HB} \), to estimate the lifetime of hydrogen bonding using the TRAVIS software.\(^{327,328,373,374} \) As shown in Figure 8.4b, the autocorrelation functions of the \( \text{HF}_2^- \) ions at both the low and high concentration decay faster then those of the \( \text{F}^- \) ions, giving a shorter lifetime of the water-ion hydrogen bonds. Compared to the partially charged F atoms in the \( \text{F}^- \) ions, the attractive interactions between the negatively charged \( \text{F}^- \) ions and water are stronger, and thus it takes longer time to break these hydrogen bonds. The \( \tau_{HB} \) values are 6.15 ps and 5.66 ps for the \( \text{HF}_2^- \) ions in the low- and high-concentration solutions, respectively. In comparison, \( \tau_{HB} \) are 19.50 ps and 17.41 ps for the \( \text{F}^- \) ions at the low and high concentrations, respectively. A slightly shorter lifetime is found in the high-concentration solutions because larger number of ions disturbs the water structures more strongly.

Figure 8.4: (a) Potential energy surface of proton shuttling in the bifluoride ion in the gas phase (red) and in the solute-solvent cluster (blue). (b) Autocorrelation function of the hydrogen bonds between water molecules and the fluoride and bifluoride ions with high and low concentrations.
8.3.3 Hydrogen bonds in the solvation shells of the ions

The presence of the F\textsuperscript{−} or HF\textsubscript{2}− ions alters the hydrogen bond network of liquid water, particularly the molecules in their inner solvation shells. As shown in Table 8.1, from the AIMD simulations, each water molecule forms an average of 3.57 hydrogen bonds with other water molecules in the pure liquid state. This number decreases considerably when the F\textsuperscript{−} or HF\textsubscript{2}− ions are added to the solution. This is especially the case for the high-concentration solutions, as the number of hydrogen bonds per water molecule decreases to 3.22 and 3.20 in the presence of the F\textsuperscript{−} and HF\textsubscript{2}− ions, respectively. Interestingly, the F\textsuperscript{−} and HF\textsubscript{2}− ions have opposite effects on their first and second solvation shells. For the aqueous solutions of F\textsuperscript{−}, the number of hydrogen bonds per water molecule is slightly larger in the first two solvation shells of the solute than in the bulk, which means that a water molecule tends to form more hydrogen bonds when it is closer to a F\textsuperscript{−} ion. In contrast, in the aqueous solutions of the HF\textsubscript{2}− ions, the number of hydrogen bonds per water molecule is smaller in the first two solvation shells of the solute than the bulk, indicating that the HF\textsubscript{2}− ions act as a defect in the hydrogen bond network of water. The average number of O–H–O hydrogen bonds in the first solvation shell of HF\textsubscript{2}− ions is slightly larger than that of the F\textsuperscript{−} ions, but there is a 20%-30% reduction in the number of O–H–F hydrogen bonds in the HF\textsubscript{2}− solution. Compared with the F\textsuperscript{−} ions, which enhance hydrogen bonding ability of the surrounding water molecules, the HF\textsubscript{2}− ions disrupt the hydrogen bond network of the water molecules possibly because of its larger size and weaker electronegativity of the F atoms.

Compared to liquid water, the geometry of the hydrogen bonds that form between water and other molecules also changes in the F\textsuperscript{−} and HF\textsubscript{2}− solutions. Here, we use the distance between the donor and acceptor R, the proton sharing coordinate \(\nu\), and the donor-hydrogen distance \(r_{OH}\) to describe the geometry of the hydrogen bonds. As shown in Table 8.2, the geometry of the O–H–O hydrogen bonds stays almost unchanged in aqueous solutions. Given the different electronegativity of the F and O atoms, the O–H–F hydrogen bonds have shorter R than the O–H–O hydrogen bonds. In the F\textsuperscript{−} solutions, we observe an over
Table 8.1: The number of hydrogen bonds per water in liquid water and the aqueous ionic solutions. The first and second solvation shells of the F$^-$ and HF$_2^-$ ions have $r_{FO} \leq 3.4$ Å and $r_{FO} \leq 5.4$ Å, respectively. The O–H–F hydrogen bonds can be further divided into two types, O–H–F$_r$ and O–H–F$_o$ in the solvation shells of a F atom. F$_r$ denotes the reference F atom in the ions to define the solvation shell, and F$_r$ denotes all the other F atoms except for the reference one in the solution.

<table>
<thead>
<tr>
<th>Solution</th>
<th>Solvation shell</th>
<th>Number of HBs per water molecule</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total O–H–O</td>
</tr>
<tr>
<td>KF (1.43 M)</td>
<td>All</td>
<td>3.43</td>
</tr>
<tr>
<td></td>
<td>First shell</td>
<td>3.46</td>
</tr>
<tr>
<td></td>
<td>Second shell</td>
<td>3.46</td>
</tr>
<tr>
<td>KF (3.61 M)</td>
<td>All</td>
<td>3.22</td>
</tr>
<tr>
<td></td>
<td>First shell</td>
<td>3.30</td>
</tr>
<tr>
<td></td>
<td>Second shell</td>
<td>3.24</td>
</tr>
<tr>
<td>KHF$_2$ (1.43 M)</td>
<td>All</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>First shell</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>Second shell</td>
<td>3.33</td>
</tr>
<tr>
<td>KHF$_2$ (3.61 M)</td>
<td>All</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>First shell</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>Second shell</td>
<td>3.13</td>
</tr>
<tr>
<td>Pure water</td>
<td>All</td>
<td>3.57</td>
</tr>
</tbody>
</table>

0.1 Å shortening in R for the O–H–F hydrogen bonds as compared to the O–H–O hydrogen bonds. In comparison, R only shortens by 0.02-0.03 Å in the aqueous solutions of the HF$_2^-$ ions. The F$^-$ ions are capable to form shorter and stronger O–H–F hydrogen bonds than the HF$_2^-$ ions because the former has higher electronegativity and thus higher proton affinity. The $r_{OH}$ value in the O–H–F hydrogen bonds stays almost the same in the F$^-$ and HF$_2^-$ solutions. In general, adding the ions disturbs the hydrogen bond structures of water due to the formation of the O–H–F hydrogen bonds. The F$^-$ ions form much shorter water-ion hydrogen bonds while the HF$_2^-$ ions tend to form slightly shorter water-ion hydrogen bonds.

We then define a set of parameters to describe the spatial distribution of the water molecules around the HF$_2^-$ ions. As illustrated in Figure 8.5a, these parameters are r, the distance between the H atom in water and the F atom in the HF$_2^-$ ions, and $\theta$, the H–F–F angle with the center F atom denoting the acceptor of the O–H–F hydrogen bonds. As
Table 8.2: Averaged values of $R$, $\nu$ and $r_{OH}$ of the O–H–O and O–H–F hydrogen bonds in liquid water and the F$^-$ and HF$_2^-$ solutions.

<table>
<thead>
<tr>
<th>Solution</th>
<th>$R$ / Å</th>
<th>$\nu$ / Å</th>
<th>$r_{OH}$ / Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>KF (1.43 M)</td>
<td>2.89</td>
<td>2.77</td>
<td>-0.94</td>
</tr>
<tr>
<td>KF (3.61 M)</td>
<td>2.89</td>
<td>2.77</td>
<td>-0.94</td>
</tr>
<tr>
<td>KHF$_2$ (1.43 M)</td>
<td>2.89</td>
<td>2.85</td>
<td>-0.94</td>
</tr>
<tr>
<td>KHF$_2$ (3.61 M)</td>
<td>2.89</td>
<td>2.86</td>
<td>-0.95</td>
</tr>
<tr>
<td>Pure water</td>
<td>2.88</td>
<td>/</td>
<td>-0.94</td>
</tr>
</tbody>
</table>

shown in Figure 8.5b, the H atoms in water occasionally approach the H atom in a HF$_2^-$ ion ($\theta \leq 90^\circ$). Instead, they mostly stay further away from the H atom in the HF$_2^-$ ion with $\theta$ larger than 90$^\circ$. The H–F–F angle rarely approaches 180$^\circ$, as the $\theta$ values are usually smaller than 120$^\circ$, consistent with the tetrahedral structure of the lone pairs in a F atom. The most probable configuration appears when $r = 1.82$ Å and $\theta = 102^\circ$. Here the F–H distance is shorter than the usual hydrogen bond length in water (1.97 Å). To further probe the solvation structure, we calculate the spatial distribution functions (SDF) of both the H and O atoms in water around the HF$_2^-$ ions using the TRAVIS program.$^{327,328}$ Figure 8.5c shows that both the H and O atoms distribute as a ring-shape structure around the F atoms in the HF$_2^-$ ion, which is consistent with the most probable configurations in Figure 8.5b. Even for the O atoms in water, there is a small probability to distribute around the H atom in the HF$_2^-$ ion, which confirms that it is not likely for water to form a hydrogen bond in the HF$_2^-$ ions.

### 8.3.4 IR spectra of the O-H stretch modes in liquid water and the ionic solutions

From the AIMD simulations of liquid water and the F$^-$ and HF$_2^-$ solutions, we calculate the IR spectra from the Fourier transform of the molecular dipole moment time correlation functions. Here the dipole moments are computed using the maximally localized Wannier function scheme.$^{326}$ As shown in Figure 8.6a, the computed O-H stretch vibrations of water have similar spectral features in the 3000 – 3750 cm$^{-1}$ region in all the systems. Also, the
peak frequencies range from $3368 \text{ cm}^{-1}$ to $3427 \text{ cm}^{-1}$, which are close to the experimental spectra, $3410 \text{ cm}^{-1}$, of liquid water and potassium bifluoride (Figure 8.1). Compared to the IR spectrum of liquid water, the absorption line shapes in the KF solutions are broader. When we consider the water molecules in the first solvation shell of the $F^-$ ions in the 3.61 M KF solution, we find a 50 cm$^{-1}$ red shift in the absorption peak ($3325 \text{ cm}^{-1}$) as compared to liquid water (Figure 8.6b). The red-shifted frequency is assigned to the strong O–H–F hydrogen bonds formed between water and the $F^-$ ions. In comparison, the water molecules in the second solvation shell show a 10 cm$^{-1}$ blue shift in the absorption peak, which corresponds to the weaker O–H–O hydrogen bonds among the water molecules with a slight elongation of $R$ (Table 8.2). For the HF$_2^-$ solutions, we currently are able to reproduce the blue shift of the absorption peak (Figure 8.6a). In the 3.61 M KHF$_2$ solution, the IR spectrum of the water molecules in the first solvation shell of the ion clearly shows a 50 cm$^{-1}$ blue shift in peak frequency to $3425 \text{ cm}^{-1}$ as compared to liquid water (Figure 8.6c). Since shorter $R$ strengthens a hydrogen bond while a shorter $r_{OH}$ weakens it, the blue shift indicates a weaker O–H–F hydrogen bonds and the shortening on $r_{OH}$ might play a role in determining the O-H stretch frequency (Table 8.2). However, there is a 0.01 Å shortening, which can just be the error bar.

Currently, we cannot reproduce the experimentally observed red shift in the O-H stretch
Figure 8.6: (a) IR spectra of liquid water and the $\text{F}^-$ and $\text{HF}_2^-$ solutions in the 3000 to 3750 cm$^{-1}$ region derived from AIMD simulations. Contributions of water molecules in the first and second solvation shells of the ions in the 3.61 M (b) potassium fluoride and (c) potassium bifluoride solutions.

frequency in the $\text{HF}_2^-$ solutions (Figure 8.1). There are a few possible reasons. Firstly, AIMD simulations ignore nuclear quantum effects, and thus they would underestimate the extent of proton delocalization in the O–H–F hydrogen bonds. Since proton delocalization promotes the elongation of the O-H bond, we would expect to observe a red-shifted stretch frequency of the water molecules in the $\text{HF}_2^-$ solutions. Another possible reason is that we do not observe the breaking of H-F bonds in the $\text{HF}_2^-$ ions in the relatively short AIMD simulations. If one $\text{HF}_2^-$ ion dissociates into one hydrogen fluoride and one $\text{F}^-$ ion, these particles might influence the hydrogen bond structure of water and its IR absorption spectra.
8.4 Conclusion

In this chapter, we carry out AIMD simulations to reveal the solvation structures, hydrogen bond geometries and IR spectra of liquid water, and the 1.43 M and 3.61 M potassium fluoride and potassium bifluoride solutions. From the simulations, we find that the SHBs in the HF\text{$_2$} ions are likely to be linear with shared protons. The instantaneous asymmetric distribution of the water molecules around the two F atoms in the HF\text{$_2$} ions leads to transient asymmetry of the F–H–F SHBs. The hydrogen bond lifetime shows that the HF\text{$_2$} ions have frequent changes of the water-ion hydrogen bonding interactions. We find that the F$^-$ ions increase the number of hydrogen bonds that water molecules can form in their first two solvation shells as compared to the bulk, and the O–H–F hydrogen bonds are 0.1 Å shorter than the O–H–O hydrogen bonds in liquid water, suggesting that the F$^-$ ions are better acceptors for hydrogen bonding than the water molecules. In contrast, water molecules in the first two solvation shells of the HF\text{$_2$} ions tend to form fewer hydrogen bonds than in the bulk, indicating that the HF\text{$_2$} ions are structural breakers for the water hydrogen bond network. The SDFs shows that the O–H–F hydrogen bonds in the HF\text{$_2$} solutions are most likely to adopt a F–H distance of 1.82 Å and a H–F–F angle of 102°, similar to the tetrahedral structure of the hydrogen bond networks in liquid water. Finally, our calculated IR spectra show a broad absorption for the O-H stretch vibration in the KF solutions, which is due to the formation of the strong O–H–F hydrogen bonds and weaker O–H–O hydrogen bonds. Our current IR spectra for the KHF\text{$_2$} solutions can only reproduce the blue shift in the O-H stretch frequency in the experimental measurement, and it mainly arises from the weaker water-ion hydrogen bonds in the first solvation shells of the HF\text{$_2$} ions. We are conducting more calculations to explain the experimentally observed red shifts.
CHAPTER 9
SUMMARY

In this dissertation, I have studied the structural, chemical and spectral features of SHBs in various condensed phase systems using electronic structure calculations and first principle simulations that include both electronic and nuclear quantum effects. Our computational results demonstrate the importance to take quantum effects into consideration when dealing with SHBs in condensed phase systems.

In Chapters 2 and 4, I focus on the SHBs in biological systems, and conduct statistical analyses on macromolecules with atomic resolution from the PDB. Based on the types of hydrogen bonds, I have divided my discussions into two parts, SHBs between two amino acids in proteins and SHBs between one amino acid and one ligand in protein-ligand complexes. The first type of SHBs often contain the negatively charged Asp and Glu residues as the acceptor, and the neutral residues of Tyr, Ser, Thr as the donor. The side chains of these amino acids also frequently participate in the formation of the second type of SHBs. From the ligand side of amino acid-ligand SHBs, alkyl hydroxyl is the dominant functional group in the formation of SHBs, which mainly originate from carbohydrates and ribose in nucleotides. Other favorable functional groups are phosphates from nucleotides and carboxyl groups from fatty acids and heme.

The property of a SHB highly depends on its geometric structure, so understanding the geometries and functions of biological SHBs often requires accurate determination of the protein structures. However, only about 1% of the available PDB structures are of atomic resolution. In Chapter 3, I collaborate with Professor Sijian Wang and Yuanhao Liu to develop a machine learning model that effectively predicts the presence of biological SHBs between two amino acids in proteins. Our model considers the residue, heteroatom, location, charge, secondary structure and sequence information of the amino acids, and we
successfully control the precision and recall of the model to both around 80% by selecting an appropriate value of the probability threshold. We have further designed a web server for this model on https://www.sas.rutgers.edu/cms/wanggroup/mapshb-model/the-mapshb-model. Then in Chapter 4, we are currently developing a new machine learning model for the prediction of amino acid-ligand SHBs by considering more features of ligands, such as the functional group, $pK_a$, $pK_b$ and logP. Both two models will provide additional restraints for both experimental and computational refinement of biomolecular structures, and facilitate the design of novel bio-inspired materials to achieve enhanced functions.

Using electronic structure calculations, I have uncovered the structural and energetic properties of biological SHBs. In Chapter 2, a series of R-dependent properties are identified, demonstrating the importance of quantum effects in SHBs. When R shortens, I observe more single-well potentials instead of double-well potentials in the proton potential energy surfaces of SHBs. At the same time, the proton becomes more shared in the SHBs and the energy barrier for proton transfer decreases. In Chapters 3 and 4, the machine learning models reveal the major factors that contribute to the formation of biological SHBs. For amino acid-amino acid SHBs, the donor residue type is of the most importance. There is a decreasing preference to form SHBs for Tyr, Ser and Arg when considering Asp as the acceptor. Based on the energy decomposition analysis, I find that this trend comes from a competition of intermolecular interactions. The Pauli repulsion and solvation effect tend to push the heteroatoms away in space, while electrostatics, polarization, dispersion and charge transfer interactions play key roles in stabilizing a SHB. The balance of these effects makes different types of donor residues energetically favorable at different R. In the machine learning model for amino acid-ligand SHBs, both the amino acid residue type and the functional group of ligands contribute significantly to the predictions. In particular, the alkyl hydroxyl-Asp/Glu pairs frequently appear as SHBs in carbohydrate-binding proteins. By calculating the hydrogen bond energy, I have demonstrated the function of SHB networks in promoting the formation of SHBs. I am currently carrying out more calculations
for other favorable pairs, such as the phosphate-Thr and carboxylic acid-Tyr/His SHBs.

Highly downfield $^1$H NMR chemical shifts are a spectral signature of SHBs, although it is difficult to investigate the structure-property relation of biological SHBs due to the large size of the biomolecules. In Chapters 5 and 6, I have selected a set of model molecules in organic solvents and aqueous solutions to mimic the behavior of biological SHBs and have studied them using first principles simulations. More frequent proton delocalization is observed when both electronic and nuclear quantum effects are included. After computing the $^1$H NMR chemical shifts of solute-solvent clusters, the downfield chemical shifts are confirmed to arise from a competition between the covalent bonding and hydrogen bonding interactions. I have further discovered a universal relation between the instantaneous proton position and the chemical shift of SHBs, and have proposed a quadratic metric that allows one to efficiently detect the proton position of a SHB in a large biomolecule directly from $^1$H NMR measurement. I have also considered the structure, energetics and spectrum of a bifurcated SHB, and have concluded that each hydrogen bond interaction is weaker than a normal dimer SHB due to its nonlinearity, while their sum makes it more stable as a bifurcated SHB.

In Chapters 7 and 8, I focus on the SHBs in non-biological condensed phase systems, which include ionic solution and aqueous solutions of potassium bifluoride. Using first principles simulations, I have showcased the formation of new hydrogen bonded species between the acid and base in the mixture composing of acetic acid and 1-methylimidazole. The shared proton is delocalized in this SHB and contributes to the broad vibrational signatures in the 2250-3000 cm$^{-1}$ region. In addition, I find that the hydrogen bond structures of liquid water are strongly disturbed when bifluoride ions are added due to their interactions with the solvation shells. Our calculations on the IR spectra have explained that the blue shift of the O-H stretching frequency in the water molecules comes from the first solvation shell. I are currently working on explaining the red shift as observed in the experimental spectra.
ACKNOWLEDGMENT OF PREVIOUS PUBLICATIONS


[P7] Zhou, S.*; Liu, Y.*; Wang, S.; Wang, L. Characteristics and machine learning prediction of biological short hydrogen bonds in protein-ligand complexes. *In preparation*. (*These authors contributed equally to this work.*)

REFERENCES


(26) Macdonald, A. L.; Speakman, J. C.; Hadži, D. Crystal structures of the acid salts of some monobasic acids. Part XIV. Neutron-diffraction studies of potassium hydrogen bis(trifluoroacetate) and potassium deuterium bis(trifluoroacetate): crystals


(120) Case, D. et al., *AMBER 2016*; University of California, San Francisco.


(170) Zuehlsdorff, T. J.; Napoli, J. A.; Milanese, J. M.; Markland, T. E.; Isborn, C. M. Unraveling electronic absorption spectra using nuclear quantum effects: Photoac-


(343) Daly, C. A.; Streacker, L. M.; Sun, Y.; Pattenaude, S. R.; Hassanali, A. A.; Petersen, P. B.; Corcelli, S. A.; Ben-Amotz, D. Decomposition of the Experimental Raman


