



AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4

Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

http://www.cfcopies.com/V2/leg/leg_droi.php

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>

THÈSE

présentée en vue de l'obtention du titre de

Docteur de l'Université de Lorraine
en Chimie

par

Antoine MARION

**Dynamiques moléculaires utilisant un champ de force quantique
semiempirique : développement et applications à des systèmes
d'intérêt biologique**

**Molecular dynamics using a semiempirical quantum force field:
development and applications to systems of biological interest**

Soutenue publiquement le 8 décembre 2014
à l'Université de Lorraine, Nancy, France

Membres du jury

Président :	Prof. Claude Millot	UMR SRSMC, Université de Lorraine, Vandœuvre-lès-Nancy, France
Rapporteurs :	Prof. Walter Thiel	Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Allemagne
	Dr. Daniel Borgis	UMR P.A.S.T.E.U.R., Département de Chimie, École Normale Supérieure, Paris, France (D.R. CNRS)
Examineur :	Prof. Abdenacer Idrissi	UMR LASIR, Université de Lille 1, Villeneuve d'Ascq, France
Directeur :	Prof. Gérald Monard	UMR SRSMC, Université de Lorraine, Vandœuvre-lès-Nancy, France
Co-directrice :	Dr. Francesca Ingrosso	UMR SRSMC, Université de Lorraine, Vandœuvre-lès-Nancy, France



Remerciements / Acknowledgments

I first would like to express my gratitude to the jury members who accepted to judge this work. In particular, I am grateful to Prof. Walter Thiel and Dr. Daniel Borgis for having reviewed this manuscript.

I had the chance to be guided by two brilliant and complementary thesis directors, Dr. Francesca Ingrosso and Prof. Gérald Monard. They taught me so much during the past three years and the present work is the result of their constant availability to answer my numerous questions. Thanks to both of you for your time, patience, help and for the nice moments that we shared together.

I am also particularly grateful to my former Master thesis advisor, Prof. Mironel Enescu. You brought me to the world of computational sciences and initiated my passion for investigating and understanding (bio-)chemistry by sharing yours.

I had the invaluable opportunity to evolve in a wonderful working environment composed by very talented and welcoming people. I would like to thank all the permanent and non permanent staff of the SRSMC laboratory. It is not possible to be exhaustive but I want to thank personally a few members of the TMS research team and collaborators:

- Prof. Jean-Louis Rivail for the fascinating discussions.
- Dr. Manuel F. Ruiz-López for his precious contribution to this work and for helping me to remain optimistic during the difficult moments.
- Prof. Xavier Assfeld for being a thoughtful doctoral school director and a great friend.
- Dr. Mounir Tarek for the valuable discussions as well as for his constant and communicative happiness.
- Prof. Margarita I. Bernal-Uruchurtu for her contribution to this work and for her warm welcome in Mexico.

I am also grateful to Dr. François Dehez, Dr. Alessandro Genoni, Dr. Alexandrine Lambert, Dr. Marilia Martins-Costa, Prof. Claude Millot and Dr. Antonio Monari for their welcome and for sharing their knowledge. I feel very lucky to have met so many friendly and talented young researchers from all around the world: Andraž, Agisilaos, Benjamin, Burcu, Daniel, Gülşah, Hatice, Hugo, İlke, Julian, Lucie, Magda, Marina, Maura, Oleksandr, Thibaud, Thibaut, Nicola, ... Thanks to all of you for the nice moments, discussions and friendship that we shared. I am also particularly grateful to Séverine, the kind and thoughtful secretary,

which was always here to remind me what I had to do and to fix everything.

This accomplishment does not only result from the last three years that I spent in Nancy. It indeed also stems from the constant support of my family and close friends. I will hereafter express my gratitude to them in French.

Tout d'abord, merci à mes parents, Véronique et Pierre. Merci Maman de m'avoir transmis ta joie de vivre et ton optimisme à toutes épreuves. Merci également pour ton soutien sans faille dans les bons comme dans les mauvais moments. Merci enfin de m'avoir tant donné et appris pour réussir dans le monde professionnel et surtout en dehors, mes succès sont les tiens. Merci Papa de m'avoir transmis ta curiosité scientifique et technologique. Merci de m'avoir initié si tôt à l'informatique et de m'avoir donné ton goût du travail bien fait.

Je tiens également à remercier mes sœurs, Charlotte et Virginie. Merci d'avoir supporté votre petit frère pendant toutes ces années et de lui avoir transmis votre expérience. C'est une chance et un moteur exceptionnels que de tenter de suivre l'exemple de deux grandes sœurs talentueuses, qui se sont battues sans relâche pour mériter les grands succès qu'elles ont obtenus. Merci à vous deux pour votre aide, pour votre soutien permanent et pour nos innombrables fous rires.

Le long chemin pour arriver jusqu'ici aurait été beaucoup plus compliqué sans le soutien de mes amis. J'ai la chance exceptionnelle de faire partie d'un groupe d'amis soudés depuis de longues années, composé de femmes et d'hommes brillants au cœur énorme. Difficile d'être exhaustif, mais je tiens à remercier particulièrement : Adrien, Alban, Alice, Anne, Antoine, Baptiste, Benjamin, Benoit, Boris, Céline, Clémence, Cyrielle, Gaëtan, Hugues, Loïc, Lucie M, Lucie N, Manon, Marion, Mathieu, Morgane, Meddy, Nicolas, Paul, Perrine, Romain, Samira, Sarah, Stéphanie et Yohan. Merci à vous et à tous les autres pour tous ces bons moments et tout simplement pour être vous-mêmes.

Finally, spending three years in Nancy brought me much more than a starter for my professional life. I had the chance to meet İlke, wonderful woman and talented scientist. Thank you for your help in this work but most of all, thank you for your support, your happiness and for sharing my life. *Seni çok seviyorum bi'tanem.*



Contents

Remerciements / Acknowledgments	iii
Contents	vii
Introduction générale	1
General introduction	5
1 Methodology	11
Résumé	11
1.1 Quantum chemistry	13
1.1.1 The Born-Oppenheimer approximation	14
1.1.2 Orbitals, secular equation and Slater determinant	15
1.1.3 The Hartree-Fock method	16
1.1.4 The MP2 method	18
1.1.5 The density functional theory	19
1.1.6 Summary	20
1.2 Molecular mechanics	21
1.3 Molecular dynamics	22
1.3.1 Principle	23
1.3.2 Ergodic hypothesis	23
1.3.3 Integration and time step	24
1.3.4 Periodic boundary conditions	26
1.3.5 Short and long range electrostatic interactions: the Ewald summation	27
1.3.6 Thermodynamical ensembles and temperature coupling	29
1.3.7 General remarks	30
1.4 Free energy calculation techniques	31
1.4.1 Umbrella sampling	32
1.4.2 Metadynamics	34
1.5 Trajectory analysis tools	36
1.5.1 Radial distribution functions	36

1.5.2	Spatial distribution functions	38
1.5.3	Infrared spectroscopy	39
1.6	Concluding remarks	43
2	Semiempirical methods	47
	Résumé	47
2.1	Historical overview	49
2.2	NDDO-based methods: MNDO, AM1 and PM3	52
2.2.1	The NDDO approximation	52
2.2.2	Core-core repulsion function in MNDO, AM1 and PM3	55
2.3	Corrections of existing methods	56
2.3.1	Parameterizable interaction function: PIF	57
2.3.2	Method adapted for intermolecular studies: MAIS	59
2.4	Concluding remarks	60
3	SemiEmpirical Born-Oppenheimer Molecular Dynamics (SEBOMD): description and related developments	63
	Résumé	63
3.1	Amber14 and AmberTools14	65
3.1.1	General comments	65
3.1.2	The sander program	66
3.2	SEBOMD in Amber14	67
3.2.1	Implementation	67
3.2.2	SEBOMD features	68
3.3	Improvements and developments of seboimd in sander	70
3.3.1	Setup and communication with sander	70
3.3.2	Rewriting of the routine for corrections to semiempirical methods . . .	70
3.4	Other related developments around sander and seboimd	72
3.4.1	Vibrational normal mode analysis: calculation and visualization	72
3.4.2	Spline definition of smoothing functions for metadynamics simulations	76
3.5	Liquid water as a test case	80
3.5.1	Structure of water: the choice of the Hamiltonian	80
3.5.2	Long range interactions: the choice of the charge model	82
3.6	Concluding remarks	87
4	PIF3: improvement of semiempirical methods for the interaction of water with hydrophobic groups	91
	Résumé	91
4.1	Using semiempirical methods to describe intermolecular interactions	93
4.2	Hydrocarbon-Water Interactions	95

4.2.1	Test of NDDO Methods using GCFs in the core-core terms	95
4.2.2	Test of NDDO Methods Using a Core-Core Parametrizable Interaction Function (PIF)	98
4.3	Recalibration of the PM3-PIF2 Method: PM3-PIF3	99
4.3.1	The PIF3 strategy	100
4.3.2	Parametrization procedure and test calculations on the methane-water complex	100
4.4	Test computations	103
4.5	Concluding remarks	109
5	Solvent effects on small biological compounds	113
	Résumé	113
5.1	Computational details	115
5.2	Hydrophobic compounds	119
5.2.1	Solvent structure	119
5.2.2	Electronic properties	128
5.2.3	Vibrational properties	130
5.2.4	Summary	141
5.3	Hydrophilic compounds	142
5.3.1	Solvent structure and solute geometrical properties	142
5.3.2	Electronic properties	155
5.3.3	Vibrational properties	157
5.3.4	Summary	162
5.4	Peptide model	163
5.4.1	General aspects on the alanine dipeptide conformation	163
5.4.2	Dynamics of structural and electronic properties	166
5.4.3	Solvent structure	169
5.4.4	Vibrational properties	171
5.4.5	Summary	173
5.5	Concluding remarks	175
6	Water self-dissociation in confined systems	179
	Résumé	179
6.1	Background on water clusters	182
6.2	Proton transfer in water: performance of semiempirical Hamiltonians	186
6.2.1	The water dimer	187
6.2.2	Ionic model: the Eigen cation	194
6.2.3	Zwitterionic minima of (H ₂ O) ₂₁ with PM3-MAIS	196
6.3	Relation between proton transfer free energy and cluster topology	199

6.3.1	Umbrella sampling simulations	200
6.3.2	Reaction descriptors	202
6.3.3	Hydrogen bonds network analysis	204
6.3.4	Environment descriptors	207
6.3.5	Data correlation	210
6.3.6	Prediction of favorable situations	212
6.4	Concluding remarks	215
Conclusions générales		217
General conclusions		221
Supplementary material		I
List of Figures		III
List of Tables		XIII
Bibliography		XV



Introduction générale

Modéliser et comprendre les processus biochimiques et biophysiques qui ont lieu dans la nature est l'un des plus grands défis des sciences chimiques modernes. Des petites biomolécules modèles aux larges et complexes systèmes macromoléculaires, la gamme des études possibles est sans limite. Afin de traiter des problèmes biochimiques à un niveau moléculaire, le chimiste théoricien a accès, aujourd'hui, à une large gamme d'outils, allant des approches simplifiées de mécanique moléculaire (MM) aux modèles plus élaborés de chimie quantique. Dans la majeure partie des cas, il est nécessaire de faire le lien entre les résultats tirés de simulations moléculaires et ceux issus d'observations expérimentales : afin de vérifier la validité d'un modèle, de l'améliorer, d'aider la compréhension de données expérimentales ou encore de suggérer la mise en place de nouvelles expériences. Dans tous les cas, la méthode de calcul doit être capable de reproduire au mieux les conditions expérimentales et doit prendre en compte les propriétés pertinentes du système afin d'espérer atteindre une représentation raisonnable du phénomène d'intérêt.

L'étude des propriétés vibrationnelles dans les systèmes biologiques est un exemple typique de domaine de recherche qui illustre la relation proche qui lie la théorie et l'expérience. Certains mouvements vibrationnels sont sensibles à leur environnement, et cette sensibilité provient d'une polarisation mutuelle entre la molécule d'intérêt et son entourage. Les informations structurales et dynamiques obtenues au niveau expérimental sont souvent quelque peu complexes à interpréter. C'est ici que la modélisation peut jouer un rôle : d'une part pour proposer une interprétation détaillée au niveau moléculaire, et d'autre part, pour assister la mise en place de nouvelles expériences. Cependant, les résultats expérimentaux obtenus par spectroscopie infrarouge contiennent des informations à la fois électroniques et dynamiques. La façon dont les mouvements vibrationnels d'une molécule ou d'une macromolécule se couplent avec l'environnement peut être particulièrement subtile et dépend de la façon qu'a le milieu de polariser le système, de la possibilité d'un transfert de charges et de la possibilité de donner lieu à des réactions chimiques. Ainsi, afin de modéliser de tels systèmes, il est nécessaire d'avoir un modèle théorique qui tient compte d'une description de la structure électronique des molécules impliquées, ainsi que de leurs propriétés dynamiques. Cela requiert donc une description quantique des électrons ainsi qu'un échantillonnage conformationnel permettant d'obtenir de bonnes statistiques (corrélations, événements rares, *etc.*).

Parce que plus les systèmes moléculaires sont grands, plus leur nombre de degrés de libertés est important, il est primordial d'échantillonner correctement leur espace conformationnel afin d'obtenir une description satisfaisante de leurs propriétés. Ceci peut être obtenu, par exemple, par des simulations de dynamique moléculaire qui décrivent les variations géométriques des systèmes moléculaires au cours du temps. Pour permettre la convergence des statistiques thermodynamiques de ces simulations, il est important de les mener jusqu'à des temps longs, et par conséquent, d'en calculer l'énergie pour un grand nombre de géométries successives. Dans le même temps, afin de décrire les propriétés électroniques, l'utilisation de calculs de chimie quantique est nécessaire à chaque pas de temps de la dynamique moléculaire. Cependant, le coût d'un calcul de chimie quantique peut être très important en termes de ressources informatiques. En utilisant des algorithmes standards, ce coût de calculs augmente de manière non linéaire avec la taille du système. Ainsi, il peut donc devenir impossible d'obtenir la structure électronique complète d'une grande molécule et de son environnement, en particulier si ce calcul doit être répété à chaque pas de temps d'une simulation de dynamique moléculaire. L'un des défis du chimiste théoricien est ainsi d'être capable de mener de longues simulations de dynamique moléculaire pour des systèmes de grande taille tout en gardant une description précise de leurs propriétés électroniques grâce aux outils de la chimie quantique.

Un moyen de résoudre ce problème est d'utiliser la méthode de dynamique moléculaire Car-Parrinello, dans laquelle un système moléculaire est décrit, dans son entier, par la théorie de la fonctionnelle de la densité. Cependant, cette méthode requiert des temps de calculs longs, et est, de nos jours, limitée à des systèmes contenant au mieux quelques centaines d'atomes durant quelques picosecondes en utilisant plusieurs centaines de processeurs. Une autre approche, proposée à l'origine par les lauréats du Prix Nobel Martin Karplus, Michael Levitt et Arie Warshel, consiste à partitionner un système moléculaire en deux parties : une partie active, décrite par la mécanique quantique (QM), et une partie inactive, décrite au moyen de la mécanique moléculaire (MM). Le second perturbe les propriétés électroniques du premier au travers d'interactions électrostatiques et de van der Waals. Cette méthode est connue sous le nom de Mécanique Quantique/Mécanique Moléculaire (QM/MM) et est largement utilisée depuis les années 1990, moment à partir duquel les ressources informatiques sont devenues suffisamment importantes pour permettre la description de systèmes contenant plusieurs milliers d'atomes dont seulement quelques uns sont décrits à un niveau de chimie quantique.

Bien que très puissante et largement utilisée, en particulier dans le domaine de la biochimie, l'approche QM/MM peut soulever quelques questions fondamentales. Comment les interactions électrostatiques et de van der Waals doivent-elles être traitées à l'interface entre la partie QM et la partie MM ? Si une liaison covalente existe entre deux descriptions distinctes, comment la frontière doit-elle être définie et comment en tenir compte ? Lors

d'une simulation de dynamique moléculaire, comment les molécules passant d'une région à l'autre doivent-elles être traitées ? Jusqu'ici, différentes solutions techniques ont été proposées dans la littérature, au sein de l'approche QM/MM. Cependant, une description quantique de l'Hamiltonien électronique du système dans son entier permettrait de passer outre toutes ces questions.

Dans notre groupe, nous avons récemment décidé d'initier une troisième manière de coupler une description électronique quantique de systèmes de grande taille avec la dynamique moléculaire : la méthode SEBOMD (*SemiEmpirical Born-Oppenheimer Molecular Dynamics*). Dans cette méthode, le système complet est décrit à un niveau quantique, bien qu'approché, grâce à l'utilisation de méthodes quantiques semiempiriques rapides. L'avantage de ce type d'approche réside dans le fait que les calculs semiempiriques sont suffisamment rapides pour permettre leur répétition le long d'une simulation de dynamique moléculaire. Ainsi, par la méthode SEBOMD, il est possible de traiter des systèmes moléculaires de grande taille (jusque plusieurs centaines d'atomes) sur des échelles de temps permettant la convergence de nombreuses propriétés statistiques (de l'ordre de la nanoseconde).

Notre approche SEBOMD a été récemment implémentée dans la suite de programmes Amber et a été rendue publique dans sa dernière version (Amber14, sortie en avril 2014). Cette implémentation consiste à relier un module de dynamique moléculaire d'Amber (le module sander) à un code de calcul quantique semiempirique (originellement, le programme DivCon99) qui permet de mener des calculs quantiques standards ou à croissance linéaire en utilisant une variété d'Hamiltoniens semiempiriques. Alors que l'implémentation du code est relativement simple, plusieurs questions surgissent quant à l'interprétation des résultats issus de ce type de calculs, en particulier concernant la précision chimique et la capacité des méthodes semiempiriques à caractériser les interactions intermoléculaires ayant lieu au sein des systèmes moléculaires de grande taille.

L'objectif de ce travail est de fournir les outils de chimie quantique et de développement du code permettant la simulation de systèmes en phase condensée offrant une description fiable des propriétés inter- et intra-moléculaires. La nouvelle méthode SEBOMD a été testée et, durant ce travail, nous avons adressé plusieurs questions quant à la validité de cette méthodologie : les méthodes semiempiriques sont-elles suffisamment fiables pour fournir des résultats qualitatifs et quantitatifs au regard de valeurs expérimentales et de modèles basés sur un niveau de théorie plus élevé ? La méthode SEBOMD est-elle suffisamment rapide pour permettre de longues dynamiques moléculaires, nécessaires à la convergence de nombreuses propriétés statistiques, notamment pour l'étude des propriétés vibrationnelles de bio-molécules ? Est-il possible de traiter des systèmes réactifs avec la méthode SEBOMD ? Les simulations SEBOMD apportent-elles plus de réponses que leurs équivalents au niveau MM ou QM/MM ?

Ce manuscrit est destiné à rassembler nos avancées pour répondre à ces questions et est organisé comme suit :

- dans le premier Chapitre, nous décrivons les outils de chimie théorique que nous avons utilisés dans ce travail (à l'exception des méthodes semiempiriques, qui sont détaillées dans le Chapitre suivant) : les méthodes *ab initio*, la théorie de la fonctionnelle de la densité, la dynamique moléculaire, les calculs d'énergie libre et les outils permettant l'analyse des trajectoires.
- dans le deuxième Chapitre, nous nous attardons sur la description des méthodes semiempiriques qui sont au cœur de ce travail. Après une brève introduction historique, nous détaillons les méthodes semiempiriques basées sur l'approximation NDDO (*Neglect of Diatomic Differential Overlap*).
- dans le troisième Chapitre, après avoir introduit l'implémentation de SEBOMD dans Amber14, nous décrivons les différents développements et améliorations apportés au cours de ce travail. Nous discutons également une simulation test d'eau liquide.
- dans le quatrième Chapitre, le test systématique des méthodes semiempiriques de type NDDO quant à la description d'interactions moléculaires est suivi par l'introduction de l'Hamiltonien PM3-PIF3. Ce dernier est une nouvelle méthode semiempirique développée originellement dans ce travail, qui tient compte de manière explicite des interactions entre certains groupes fonctionnels important et leur environnement aqueux.
- dans le cinquième Chapitre, nous appliquons l'Hamiltonien PM3-PIF3 pour mener des simulations SEBOMD de composés organiques dans l'eau. Nous considérons une série de petits modèles contenant des groupements hydrophobes et/ou hydrophiles ainsi qu'un modèle de peptide, le dipeptide alanine. Nous analysons la structure du solvant autour de ces molécules ainsi que leurs propriétés électroniques et vibrationnelles en présence du solvant.
- dans le sixième Chapitre, nous étudions le processus d'autoprotolyse de l'eau en milieux confinés. Après avoir discuté du choix de l'Hamiltonien semiempirique à utiliser pour cette étude, nous caractérisons le transfert de proton dans l'agrégat d'eau (H₂O)₂₁ en établissant une corrélation entre l'énergie libre associée à la première étape de ce transfert et certaines propriétés physiques collectives.



General introduction

Modeling and understanding the biochemical and biophysical processes that take place in life is one of the greatest challenges of modern computational sciences. From small model biomolecules to large complex macromolecular systems, the range of investigations is limitless. To tackle biochemical problems at the molecular level, the computational chemist has access, today, to a wide range of tools, going from simplified classical point charge molecular mechanics to elaborate quantum chemistry models. In most cases, it is necessary to link the results of molecular simulations to experimental observations: either to verify the validity of a model, to improve it, to help the interpretation of these experimental data, or to suggest new experiments. In any case, the computational technique should be capable of reproducing at best the experimental conditions and should take into account the relevant properties of the studied system in order to expect reaching a reasonable representation of the phenomenon of interest.

The study of vibrational properties in biological systems is a typical example of a research field that illustrates the close relationship between theory and experience. Some vibrational motions are sensible to their environment and this sensitivity comes from a mutual polarization between the molecule of interest and its surrounding. The structural and dynamical information obtained at the experimental level is often quite complex to interpret. This is where molecular modeling can play a key role: on one hand it can propose a detailed interpretation at the molecular level, while on the other hand it can assist in the design of new experiments. However, experimental results obtained from infrared spectroscopy contain both electronic and dynamical information. The way in which the vibrational motions of a molecule or a macromolecule couple with the environment can be very subtle and depends on the way in which the medium polarizes the system, on the possibility of charge transfer and on the possibility of giving rise to chemical reactions. Therefore, to model such systems, it is necessary to have a theoretical model that tackles both the description of the electronic structure of the molecules involved and of their dynamical properties. This means that a quantum description of the electrons is required as well as extensive conformational sampling to allow for good statistics (correlations, low probability events, *etc.*).

Because large molecular systems have a large number of degrees of freedom, it is important to properly sample their conformational space to obtain a correct description of their properties. This is, for instance, obtained by running molecular dynamics simulations

that describe the geometrical variations of molecular systems along time. To converge the thermodynamical statistics of these simulations, it is important to run them for a long time, therefore to compute their energies for many different successive geometries. At the same time, to describe electronic properties, the use of quantum chemistry is required for each step of the molecular dynamics. However, the computational cost of a quantum chemical calculation is high. Using standard algorithms, it is a non linear function of the size of the system. Therefore, it can become prohibitive to obtain the complete electronic structure of a large molecule and its environment, especially when the calculation is repeated at each time step of a molecular dynamics simulation. One of the challenges of the computational chemist is indeed to be able to run long molecular dynamics simulations of large systems together with the accurate description of their electronic properties through the use of quantum chemistry tools.

One way to solve this problem is to use the Car-Parinello Molecular Dynamics method, in which an entire molecular system is described using the density functional theory approach. However, this method requires quite long calculation times and it is still limited, today, to the simulations of systems containing at most a few hundred of atoms for a few picoseconds using thousands of cores in supercomputers. Another approach originally proposed by the Nobel Prize laureates Martin Karplus, Michael Levitt and Arie Warshel, consists in partitioning a molecular system into two parts: an active part, described using quantum mechanics (QM), and a nonactive part, described by means of molecular mechanics (MM). The latter perturbs the electronic properties of the former through electrostatic and van der Waals interactions. This method is referred to as Quantum Mechanics/Molecular Mechanics (QM/MM) and it became extensively used starting from the 1990s, when the computational power became high enough to allow the description of systems containing thousands of atoms with only a few being described by quantum chemistry.

Although powerful and widely applied, especially in the biochemistry field, the QM/MM approach may arouse some fundamental questions. How should the electrostatics and van der Waals interactions be treated between the QM and the MM part? If a covalent bond exists between the two distinct descriptions, how should this frontier be defined and taken care of? In a molecular dynamics simulation, how should one treat the molecules moving from one region to the other? So far, different technical solutions have been suggested in the literature within the QM/MM approach. However, a quantum mechanical description of the electronic Hamiltonian of the entire system would allow overcoming these issues.

In our group, we have recently decided to initiate a third way of merging quantum electronic description of large systems with molecular dynamics: the use of SEBOMD, which stands for SemiEmpirical Born-Oppenheimer Molecular Dynamics. In this method, the entire system is described at the quantum, albeit approximate, level through the use of a fast semiempirical quantum approach. The advantage of such approach is that semiempirical

calculations are fast enough to enable their repetitions along a molecular dynamics simulation. Therefore, in SEBOMD, it is possible to tackle large molecular systems (up to several hundreds of atoms) along time scales that allow for many statistical properties to converge (up to the nanosecond scale).

Our SEBOMD approach has been recently implemented in the Amber suite of programs and it has been made public in its last version (Amber14, released in April 2014). It consists of linking a molecular dynamics module of Amber (sander) to a semiempirical quantum code (originally, the DivCon99 program) that is able to perform either standard or linear scaling quantum calculations using various semiempirical Hamiltonians. While the code implementation is straightforward, many questions can be raised on the results obtained from such calculations, especially concerning the chemical accuracy and the ability of semiempirical methods to characterize the intermolecular interactions taking place in large molecular systems.

The objective of the present work resides in providing both the quantum chemistry and the code development tools to run simulations of condensed phase systems with a reliable description of molecular properties. The new SEBOMD approach has been tested and, during our study, we have addressed some questions that are critical to assess the validity of this methodology. Are semiempirical quantum methods reliable enough to provide qualitative and quantitative results as compared to experimental values and to higher levels of theory? Is the SEBOMD method efficient enough to allow for long molecular dynamics that are needed to make statistical properties converge, especially in the study of the vibrational properties of biomolecules? Can reactive systems be treated with SEBOMD? Do SEBOMD simulations bring more information than their MM or QM/MM counterparts?

This manuscript is devoted to reporting our scientific advances in answering these questions and is organized as follows:

- in the first Chapter, we describe the computational chemistry tools that we have used during this work (except for semiempirical methods, which are detailed in the following Chapter): *ab initio* methods, density functional theory, molecular dynamics, free energy calculations and trajectory analysis tools.
- in the second Chapter, we focus on the description of the semiempirical methods that are at the core of our work. After a brief historical introduction, we focus on the semiempirical methods based on the NDDO (Neglect of Diatomic Differential Overlap) approximation.
- in the third Chapter, after introducing the implementation of SEBOMD in Amber14, we describe the developments and improvements brought during this work. We also discuss, as a test case, a SEBOMD simulation of liquid water.
- in the fourth Chapter, a systematic test of the performance of NDDO semiempirical methods in the description of intermolecular interactions is followed by the intro-

duction of the PM3-PIF3 Hamiltonian. It is a new semiempirical method originally developed in this work, which explicitly accounts for the interaction between some important functional groups and their aqueous environment.

- in the fifth Chapter, we apply the PM3-PIF3 Hamiltonian to perform SEBOMD simulations of organic compounds in water. We consider a series of small models having hydrophobic and/or hydrophilic groups as well as a model of polypeptide, the alanine dipeptide. We analyze the solvent structure around these compounds as well as their electronic and vibrational properties in the presence of the solvent.
- in the sixth Chapter, we investigate the water self-dissociation process in confined environments. After discussing the choice of the semiempirical Hamiltonian to be used for this purpose, we characterize the proton transfer in the $(\text{H}_2\text{O})_{21}$ water cluster by establishing a correlation between the free energy of the first step of this process and some collective physical properties.

Chapter

1

Methodology

Résumé

Ce chapitre est destiné à exposer les fondements théoriques des méthodes utilisées pour ce travail. Nous présentons tout d'abord les approches permettant d'évaluer l'énergie potentielle d'un système moléculaire : la chimie quantique et la mécanique moléculaire. Puis nous abordons les principes de la dynamique moléculaire. Enfin, nous détaillons certaines techniques permettant d'évaluer l'énergie libre associée à un processus physico-chimique ainsi que les outils d'analyse donnant accès aux diverses propriétés dynamiques du système étudié.

La chimie quantique est destinée à décrire le comportement quantique de la structure électronique d'un système moléculaire donné. Elle se base sur les principes de la mécanique quantique et cherche à résoudre l'équation de Schrödinger indépendante du temps. La résolution analytique de l'équation de Schrödinger n'étant pas possible pour des systèmes polyélectroniques et/ou polyatomiques, différentes approximations sont nécessaires afin d'approcher la solution exacte. Notamment, l'approximation de Born-Oppenheimer constitue la base des méthodes de chimie quantique et permet de découpler le mouvement de noyaux de celui des électrons en remarquant la différence importante de masse entre ces deux particules. Il existe principalement trois types d'approches en chimie quantique : les méthodes *ab initio* cherchant à évaluer la fonction d'onde électronique du système (par exemple, les méthodes Hartree-Fock et post-Hartree-Fock), la théorie de la fonctionnelle de la densité (*density functional theory*, DFT) et les méthodes approximées de type semiempirique. Les approches *ab initio* et la DFT sont détaillées dans ce chapitre alors que les approches semiempiriques seront traitées séparément dans un chapitre dédié.

La mécanique moléculaire se base sur une représentation classique des systèmes moléculaires. Elle nécessite la définition d'un champ de forces : une équation paramétrée permettant de décrire les interactions atomiques et moléculaires (liaisons, interactions électrostatiques, interactions de van der Waals, ...). Les paramètres d'un champ de forces sont habituellement optimisés pour reproduire certaines propriétés clés du système, en référence à des résultats obtenus expérimentalement ou par le biais de calculs réalisés à un plus haut niveau de théorie. La mécanique moléculaire offre un moyen rapide et efficace de traiter des

systèmes de grande taille au détriment des propriétés quantiques de ces derniers.

Parmi les approches capables de décrire le comportement dynamique d'un système moléculaire, la dynamique moléculaire est celle qui a été retenue au cours de ce travail. Connaissant la surface d'énergie potentielle du système d'intérêt, l'intégration de la seconde loi de Newton conduit à une trajectoire classique des atomes représentant l'évolution temporelle de ce système. Afin d'atteindre des temps de simulation suffisamment longs pour décrire le phénomène d'intérêt, la méthode de calcul d'énergie sélectionnée ne doit requérir qu'un temps de calcul suffisamment bas. Il est possible de simuler la phase condensée en appliquant des conditions périodiques aux limites, rendant ainsi compte des interactions du système avec ces répliques virtuelles dans toutes les directions de l'espace. Dans ce contexte, les interactions électrostatiques peuvent être traitées efficacement grâce à la méthode des sommes d'Ewald. Différents algorithmes permettent de produire une simulation dans divers ensembles thermodynamiques en assurant, par exemple, la conservation du nombre de particules, du volume total et de la température (ensemble canonique, NVT).

Différentes approches permettent d'évaluer l'énergie libre associée à un phénomène donné. Dans le cadre d'une réaction chimique, les méthodes dites *umbrella sampling* ou métadynamique font partie des plus utilisées. Ces deux méthodologies permettent d'une part de biaiser la dynamique moléculaire du système afin de favoriser l'exploration de l'espace des phases. D'autre part, à partir de l'analyse de la force de biais appliquée, il est possible de rendre compte de l'énergie libre nécessaire au système pour passer d'un état à un autre.

Enfin, de nombreuses propriétés dynamiques des systèmes moléculaires peuvent être calculées à partir de la trajectoire obtenue. Notamment, le calcul des fonctions de distributions radiales dans un système en phase condensée permet de rendre compte de la structure du solvant autour d'un atome ou d'un groupe fonctionnel donné. Les fonctions de distributions spatiales offrent une analyse similaire à celle obtenue grâce aux fonctions de distributions radiales en conservant les informations relatives à l'arrangement spatial du système. Les propriétés vibrationnelles des molécules étudiées peuvent aussi être calculées à partir de la trajectoire. Le spectre infrarouge d'une molécule peut notamment être obtenu par transformée de Fourier de la fonction d'auto-corrélation temporelle de son moment dipolaire.

The present Chapter is intended to give an overview of the methods used in this work. We first detail the methods of computing the energy of a molecular system using *quantum mechanical*[1–3] (QM) and *molecular mechanical*[4] (MM) approaches. Then, we introduce the *molecular dynamics* technique[4, 5] as well as the related tools used to perform free energy calculations[4] and to retrieve macroscopic properties from such computer simulations. In the first part about quantum mechanical approaches, semiempirical methods are not treated since Chapter 2 will be dedicated to this topic. For most of the following Sections, the general theoretical framework of the method is first presented before we discuss some practical details related to the applications performed in the present work.

1.1 Quantum chemistry

The study of the *electronic structure* of a molecular system requires to solve the time-independent Schrödinger equation:[6]

$$\hat{H}\Psi = E\Psi \quad (1.1)$$

which expresses the energy (E) and the wave function (Ψ) of a given stationary system as the eigenvalues and the eigenvectors of its Hamiltonian (\hat{H}), respectively. According to the Born postulate,[7, 8] the behavior of a molecular system is completely characterized by its wave function (Ψ).

The non-relativistic molecular Hamiltonian operator represents all the interactions between the particles of the system and is given by:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{ee} + \hat{V}_{Ne} \quad (1.2)$$

where, in atomic units:

- \hat{T}_N is the kinetic energy operator applied to all the nuclei of mass M_K written as,

$$\hat{T}_N = -\frac{1}{2} \sum_K \frac{1}{M_K} \Delta_K \quad (1.3)$$

where Δ_K is the Laplacian operator.

- \hat{T}_e defines the kinetic energy of the electrons:

$$\hat{T}_e = -\frac{1}{2} \sum_i \Delta_i \quad (1.4)$$

- \hat{V}_{NN} represents the repulsion between nuclei:

$$\hat{V}_{NN} = \sum_K \sum_{L>K} \frac{Z_K Z_L}{R_{KL}} \quad (1.5)$$

with Z_K the charge of a nuclei and R_{KL} the distance between two nuclei K and L.

- \hat{V}_{ee} gives the repulsion between electrons:

$$\hat{V}_{ee} = \sum_i \sum_{i>j} \frac{1}{r_{ij}} \quad (1.6)$$

where r_{ij} is the distance between the electrons i and j .

- \hat{V}_{Ne} accounts for the attraction between nuclei and electrons:

$$\hat{V}_{Ne} = - \sum_K \sum_i \frac{Z_K}{r_{Ki}} \quad (1.7)$$

where r_{Kj} is the distance between a nuclei of charge Z_K and a negatively charged electron.

For polyelectronic and polyatomic systems, the Schrödinger equation cannot be solved analytically. To overcome this issue, it is necessary to introduce approximations. Among all of them, one is central to most of the quantum chemistry methods and we shall discuss it in the following subsection.

1.1.1 The Born-Oppenheimer approximation

Considering the large difference of mass between nuclei and electrons, the first approximation that can be made is to consider the position of the nuclei to be fixed with respect to the electrons. This approximation, formulated by Born and Oppenheimer in 1927,[9] has a direct impact on Eq. 1.2. The term \hat{V}_{NN} becomes constant while the kinetic energy of the nuclei (\hat{T}_N) is zeroed. The molecular Hamiltonian (\hat{H}) can thus be rewritten as:

$$\hat{H} = \hat{H}_{el} + \hat{V}_{NN} \quad (1.8)$$

where \hat{H}_{el} defines the electronic Hamiltonian of the system which is given by:

$$\hat{H}_{el} = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} \quad (1.9)$$

Under the Born-Oppenheimer approximation, the wave function only depends on the electrons and is obtained for a given position of the nuclei.

This approximation, which allows to consider the motions of the electrons as independent of the one of the nuclei, is the base of most of the electronic structures methods. It is also of great importance in the present study since the SEBOMD methodology (see Chapter 3) is based on this approximation.

Despite the simplification brought by the Born-Oppenheimer approximation, the Schrö-

dinger equation of a complex polyatomic and polyelectronic system still does not admit any analytic solution. To overcome this issue, other considerations have to be made and we shall describe some them in what follows.

1.1.2 Orbitals, secular equation and Slater determinant

Each electron of a molecular system can be described by a monoelectronic function. Under the *orbital approximation*, such a function is a *molecular orbital* (ϕ_i) and it is common to expand each ϕ_i as a *linear combination of atomic orbitals* (LCAO). If we note the N atomic orbitals $\varphi_1, \varphi_2, \dots, \varphi_N$, the i^{th} molecular orbital is written as:

$$\phi_i = c_{1i}\varphi_1 + c_{2i}\varphi_2 + \dots + c_{Ni}\varphi_N \quad (1.10)$$

where the c_{vi} are the coefficients of each atomic orbital. According to the *variational principle*, the best set of c_{vi} 's is obtained by minimizing the electronic energy of the system. This leads to a set of N equations that has a non trivial solution if the following *secular equation* is fulfilled:

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} & \cdots & H_{1N} - ES_{1N} \\ H_{21} - ES_{21} & H_{22} - ES_{22} & \cdots & H_{2N} - ES_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \cdots & H_{NN} - ES_{NN} \end{vmatrix} = 0 \quad (1.11)$$

where the elements of the Hamiltonian are given, using Dirac's notation, by:

$$H_{12} = \langle \phi_1 | \hat{H} | \phi_2 \rangle \quad (1.12)$$

and the elements of the overlap matrix:

$$S_{12} = \langle \phi_1 | \phi_2 \rangle \quad (1.13)$$

To completely describe an electron, its spin should also be taken into account. To this end, two functions are introduced: α and β corresponding to a spin *up* and *down*, respectively. The product between a molecular orbital and a spin function gives rise to the definition of a *spin-orbital* (χ_i) which, for the coordinates x_1 of an electron is written as:

$$\begin{aligned} \chi_i(x_1) &= \phi_i(x_1)\alpha \\ \text{or} \\ \chi_i(x_1) &= \phi_i(x_1)\beta \end{aligned} \quad (1.14)$$

In the *orbital approximation*, the electronic Hamiltonian (Eq. 1.9) becomes the sum of the energy of each electron in the system. The wave function can thus be written as a product of spin-orbitals. Such a product is called a *Hartree product* (Φ^{HP}) and leads, for the wave function of a system composed of n electrons, to:

$$\Psi^{\text{HP}}(x_1, x_2, \dots, x_n) = \chi_1(x_1)\chi_2(x_2)\dots\chi_N(x_n) \quad (1.15)$$

However, such a Hartree product does not satisfy the antisymmetry principle, which is a requirement of the *Pauli's principle*.^[10] A way to satisfy the latter is to express the wave function as a *Slater determinant* of Hartree products.^[11] Thus, the wave function becomes:

$$\Psi(x_1, x_2, \dots, x_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \dots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \dots & \chi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_n) & \chi_2(x_n) & \dots & \chi_N(x_n) \end{vmatrix} \quad (1.16)$$

1.1.3 The Hartree-Fock method

From Eq. 1.9, we define for each electron, a *core Hamiltonian* ($\hat{H}^c(x_i)$) that contains the term of the electron kinetic energy and the potential energy related to the nuclei-electron interactions. The electronic Hamiltonian appears thus as:

$$\hat{H}_{el} = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} = \sum_k \hat{H}^c(x_k) + \sum_k \sum_{l>k} \frac{1}{r_{kl}} \quad (1.17)$$

$$\begin{aligned} E_{el} &= \langle \Psi | \hat{H}_{el} | \Psi \rangle \\ &= \sum_i \langle \chi_i(x_1) | \hat{H}^c(x_1) | \chi_i(x_1) \rangle \\ &+ \sum_i \sum_{j>i} \left[\langle \chi_i(x_1)\chi_j(x_2) | \frac{1}{r_{12}} | \chi_i(x_1)\chi_j(x_2) \rangle - \langle \chi_i(x_1)\chi_j(x_2) | \frac{1}{r_{12}} | \chi_i(x_2)\chi_j(x_1) \rangle \right] \end{aligned} \quad (1.18)$$

From this, one defines the *Hartree-Fock operator* ($\hat{F}(x_1)$) for one electron as:

$$\hat{F}(x_1) = \hat{H}^c(x_1) + \sum_j [\hat{J}_j(x_1) - \hat{K}_j(x_1)] \quad (1.19)$$

where $\hat{J}_v(x_1)$ is the *Coulombian operator* given by:

$$\hat{J}_j(x_1) = \int_{x_2} \chi_j^*(x_2) \frac{1}{r_{12}} \chi_j(x_2) d\tau_2 \quad (1.20)$$

with $d\tau_2$, a volume element in the space of x_2 . $\hat{K}_j(x_1)$ is the *exchange operator* defined by its application to a function ($\chi_j(x_1)$) as:

$$\hat{K}_j(x_1)\chi_i(x_1) = \chi_j(x_1) \int_{x_2} \chi_j^*(x_2) \frac{1}{r_{12}} \chi_i(x_2) d\tau_2 \quad (1.21)$$

For each molecular spin-orbital, the *Hartree-Fock* equations are derived as:

$$\hat{F}\chi_i = \epsilon_i \chi_i \quad (1.22)$$

and relate to every spin-orbital (χ_i) its respective eigenvalue (ϵ_i), the orbital energy.

To solve Eq. 1.22, and thus obtain the set of molecular orbitals that defines the wave function as a Slater determinant, one needs to know the Hartree-Fock operator (\hat{F}). However, this operator is directly defined by the molecular orbital as this can be seen from Eqs. 1.20 and 1.21. The resolution of such a problem requires the use of an iterative procedure called a *Self-Consistent Field* (SCF).

In the LCAO approximation (Eq. 1.10) and for a closed shell system, the Hartree-Fock equation (Eq. 1.22) leads to the *Roothaan-Hall equation*[12, 13] given in its matrix form by:

$$\mathbf{FC} = \mathbf{SCE} \quad (1.23)$$

where \mathbf{F} is the Fock matrix, \mathbf{C} is the matrix of the atomic orbital coefficients (*i.e.*, the matrix formed by the $c_{v\mu}$ elements in Eq. 1.10), \mathbf{S} is the overlap matrix between atomic orbitals and \mathbf{E} is the diagonal matrix containing the different orbital energies.

In the LCAO approximation, the elements of the Fock matrix for a wave function developed in a base of atomic orbitals (φ_μ , φ_ν , φ_λ , φ_σ) is obtained by:

$$F_{\mu\nu} = H_{\mu\nu}^c + \sum_{\lambda,\sigma} P_{\lambda\sigma} \left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\sigma|\nu\lambda) \right] \quad (1.24)$$

where $(\mu\nu|\lambda\sigma)$ is a contracted notation of:

$$(\mu\nu|\lambda\sigma) = \int_{x_1} \int_{x_2} \varphi_\mu(x_1) \varphi_\nu(x_1) \frac{1}{r_{12}} \varphi_\lambda(x_2) \varphi_\sigma(x_2) d\tau_1 d\tau_2 \quad (1.25)$$

The electronic energy of the system becomes:

$$E_{el} = \sum_{\mu,\nu} P_{\mu\nu} H_{\mu\nu}^c + \frac{1}{2} \sum_{\mu,\nu} \sum_{\lambda,\sigma} P_{\mu\nu} P_{\lambda\sigma} \left[(\mu\nu|\lambda\sigma) - \frac{1}{2} (\mu\sigma|\nu\lambda) \right] \quad (1.26)$$

From Eq. 1.26 arises the definition of the $P_{\mu\nu}$ elements of the *density matrix* (\mathbf{P}). These elements are expressed from the atomic orbital coefficients as:

$$P_{\mu\nu} = 2 \sum_i^{occ} c_{\mu i} c_{\nu i} \quad (1.27)$$

where the sum runs over all the occupied molecular spin orbitals. To solve these equations one uses the SCF procedure and thus obtains the coefficients that define the total wave function of the system.

The Hartree-Fock method gives an approximated solution to the Shrödinger equation for polyelectronic systems. However, because of the orbital approximation, it neglects part of the correlation between electrons. Many methods have been dedicated to overcome this issue by correcting the Hartree-Fock solution. Such methods are called *post Hartree-Fock* and we shall briefly describe one of them, MP2.

1.1.4 The MP2 method

As we discussed above, in the Hartree-Fock method the wave function is described by a single Slater determinant. Such an approximation leads to an only partial description of the electronic correlation (*i.e.*, only the correlation between electron of same spin is taken into account because of the antisymmetric form of the Slater determinant). Thus, the “exact” energy of the system can be expressed as:

$$E_{exact} = E_{Hartree-Fock} + E_{correlation} \quad (1.28)$$

Considering that $E_{correlation}$ is a perturbation (*i.e.*, a small variation compared to the value of $E_{Hartree-Fock}$) leads to one of the mainly used post Hartree-Fock method, the *Møller-Plesset* perturbation theory.[14] In this approximation, the Hamiltonian of the system is written as:

$$\hat{H} = \hat{F} + \hat{V} \quad (1.29)$$

where \hat{F} is the Fock Hamiltonian as defined by Eq. 1.19 and \hat{V} is the perturbation operator. The energy of the system is then found as the Hartree-Fock energy plus the perturbation. Usually, the perturbation is taken into account up to the second or fourth order, leading to the methods known as MP2[15] and MP4,[16] respectively. Such approaches allow to account for the dynamical correlation of the electrons while other methods, based on a

multideterminantal definition of the wave function, can account for the static correlation contribution to the energy.

Post Hartree-Fock methods strongly depend on the quality of the initial wave function and are particularly expensive in terms of computational cost. Other quantum mechanical methods exist. Some are also based on the computation of the system wave function but bearing stronger approximations than the techniques detailed above. This is the case of semiempirical methods for which we shall give a short review later in this manuscript. Another type of QM approaches is based on the *Density Functional Theory* (DFT) and we will now briefly described the major points of this theory.

1.1.5 The density functional theory

The density functional theory (DFT) is based on a theorem which was first formulated by Hohenberg and Kohn in 1964: the ground state of a polyelectronic system is completely described by its electronic density ($\rho(r)$).^[17] Thus, all the properties of this system, including its energy, are defined as a functional of $\rho(r)$. This theorem also states that any other density $\rho'(r)$ leads necessarily to a higher energy.

The formulation of the energy as a functional of $\rho(r)$ arose from the work of Kohn and Sham,^[18] and is expressed as:

$$E[\rho(r)] = U[\rho(r)] + T[\rho(r)] + E_{xc}[\rho(r)] \quad (1.30)$$

where $U[\rho(r)]$ represents the potential energy related to the interacting particles of the system (*i.e.*, electron-electron and nuclei-electron). $T[\rho(r)]$ is the sum of the kinetic energy of each electron with the assumption that those particles do not interact with each others. The last term of Eq. 1.30, $E_{xc}[\rho(r)]$, is introduced to account for all the phenomena that are neglected by the above assumptions and is called *exchange-correlation* functional. If the exact form of this functional was completely know, $E[\rho(r)]$ would be the “exact” energy of the considered system. However, this term is only known for a free electron gas and applications to molecular systems require the introduction of some approximations.

Many attempts to find a good exchange-correlation functional have been made during the past decades. We shall not give a complete review of these methods here but we will briefly discuss some of the main models.

The first level approach is to consider that E_{xc} only depends on the value of the density at the space coordinates where it is evaluated. This leads to the Local Density Approximation (LDA). In such methods, E_{xc} is separated into two contributions: E_x and E_c , the exchange and the correlation energy, respectively. E_x is known under this approximation and several forms and parameterizations of the correlation energy functional have been developed (*e.g.*,

Volko-Wilkes-Nusair (VWN)[19], Perdew-Zunger (PZ)[20], Perdew-Wang (PW)[21]).

The LDA approach hypothesizes a uniform distribution of electrons, which is quite far from the case of atomic and molecular systems. The next level of approximation expresses E_{xc} as a functional of the density and of the gradient of the density: the Generalized Gradient Approximation (GGA). As in the case of LDA methods, E_{xc} is separated into E_x and E_c , but here, E_x is unknown and has to be approximated too. There exist several exchange functionals such as PW91[22], PBE[23] or B[24] and one can use either an LDA correlation functional (*e.g.*, WNV, PZ) or a correlation functional developed in the GGA formalism such as LYP[25], leading for example to the well known BLYP combination. There exist extensions of the GGA approach which include the second derivative of the density in E_{xc} and such methods are called meta-GGA.

Another approach is to consider the exchange part of E_{xc} as a combination between the Hartree-Fock exchange energy and an exchange density functional.[26] Such a hybrid functional is usually made of a parameterized combination of different exchange and correlation functionals with a given amount of Hartree-Fock exchange energy. The parameterization of such a method is made by fitting experimental data such as ionization potentials or proton affinity. We can cite as an example of hybrid functionals the well known B3LYP[24–26], which intends to correct the wrong $1/r$ asymptotic behavior of most of the GGA functionals, as well as the more recent M06[27] and derivatives that have been developed to improve the description of, *e.g.*, non covalent interactions or thermodynamical properties.

The development of accurate density functionals is an active field of theoretical chemistry and other approaches are available for particular applications. For example, Grimme *et al.* introduced a dispersion term to account for weak intermolecular interactions in their B97D[28, 29] and B97D3[30] DFT methods. Finally, corrections have been developed to correctly account for long range phenomena, through the use of a *range separated* functional such as CAM-B3LYP[31]. They are for example particularly applied to the field of photochemistry to study excited states using *time dependent*-DFT calculations.[32, 33]

1.1.6 Summary

The methods cited above are based on quantum mechanics (QM). They give a description of the electrons of the system using different types of approximations. Some intend to find the wave function and are based on an *ab initio* definition of the Hamiltonian (Hartree-Fock and post-Hartree-Fock) and others consider the properties of the system as a functional of the electron density (DFT). The latter can be either *ab initio* or partially parameterized (*e.g.*, B3LYP).

Despite the efforts made to lower the computational cost of QM approaches, it is still difficult to use such methods to describe very large systems. Nevertheless, approximations

of the Hartree-Fock and DFT methods (*e.g.*, semiempirical methods and density functional tight binding, respectively) can be considered in order to reduce the computational expense of such methods (see Chapter 2).

Another way to compute the energy of larger systems at a molecular level for a low computational cost is based on a classical representation of the electron interactions. Such an approach is known as molecular mechanics and is a popular tool to study the dynamics and thermodynamics of very large systems (*e.g.*, proteins, membranes) in their environment.

1.2 Molecular mechanics

Molecular mechanics (MM) is a method for computing the energy of a given molecular system, based on classical mechanics. In such approaches, a covalent bond between two atoms can be seen as a classical spring and its related energy is given by a simple harmonic potential. Similar assumptions are made for bond angles and dihedral angles leading to a definition of the energy as an additive two body potential. Such a potential is called a *force field* and various definitions exist in the literature, such as CHARMM[34] (Chemistry at HARvard Molecular Mechanics), OPLS[35] (Optimized Potentials for Liquid Simulations) or GROMOS.[36]. As an illustration, we shall focus on the Amber[37] (Assisted Model Building with Energy Refinement) force field, which is defined as:

$$\begin{aligned}
 E = & \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
 & + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_i \sum_{j>i} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
 \end{aligned} \tag{1.31}$$

where the three first terms account for bounded interactions (*i.e.*, bonds, bond angles and dihedral angles, respectively) while the last double sum gives a description of non bonded interactions between pairs, namely van der Waals and Coulombic interactions (the terms in Eq. 1.31 are given in internal Amber units). Such a force field implies the use of many parameters:

- K_b and b_0 are, respectively, the force constant and the center of the harmonic potential used to describe covalent bonds (b).
- K_θ and θ_0 have the same definition as K_b and b_0 , respectively, for bond angles (θ)
- V_n is the force constant of the potential applied to the dihedral angle ϕ and γ is the phase angle (usually 0° or 180°).
- A_{ij} and B_{ij} are the parameters of the Lennard-Jones potential.
- q_i is the atomic point charge associated with the atom i .
- ϵ is the dielectric constant of the medium if the solvent is not explicitly taken into account (*i.e.*, equal to 1.0 otherwise).

Except ϵ , all the parameters described above are optimized for a set of predefined atom types (e.g., the bond between an sp^3 carbon atom and the oxygen atom of an alcohol group). The parameterization is based on experimental and/or quantum mechanical calculations. Several sets of parameters for the Amber force field have been derived. In this work, we will mainly use the ff03[38, 39] version of these parameters.

The force field described in Eq. 1.31 involves the calculation of the energy through simple formulas which can be easily and efficiently parallelized in a computational program. This last point is one of the main advantages of molecular mechanics over QM approaches and makes such a method able to compute the energy of very large systems for a reasonable computational expense. The large number of parameters is both an advantage and an inconvenient of this methodology. The advantage is the flexibility of the method since it is relatively easy to improve. The main shortcoming is that such a method is not general and one needs to parameterize all the interactions when dealing with systems that were not included in the original parameterization procedure, unlike most of the QM approaches.

More sophisticated schemes can be used and their development represents an active field of computational sciences. One can imagine other potentials to describe for example bonds by taking into account anharmonic potentials.[40] Other models overcome the point charge approximation and their development is based on a distribution of multipoles rather than point charges. Such methods are named *polarizable force field* and among those, one can cite AMOEBA[41] (Atomic Multipole Optimized Energetics for Biomolecular Applications) or SIBFA[42] (Sum of Interactions Between Fragments Ab initio computed).

Compared to QM methods, MM force fields have a clear advantage because they allow fast and efficient calculations on very large systems. However, the price to pay is the loss of quantum mechanical properties. Thus, there is a choice to be made in the model used to describe a molecular system and this choice depends on the properties that need to be evaluated.

1.3 Molecular dynamics

As it has been shown in the previous sections, there are different ways to compute the energy (E) of a molecular system for a given position of the atoms. Within the Born-Oppenheimer approximation, the position of the atoms can be optimized in order to minimize E , leading to stationary structures (*i.e.*, minima, transition states...). This gives a *static* representation of molecular systems that can be sometimes far from experimental observations, especially when the number of degrees of freedom in the system increases.

Considering only stationary structures results in the neglect of temperature effects. In biological sciences, most of the experiments are performed at room temperature (*i.e.*, $\sim 300\text{K}$) and the associated extra amount of energy allows molecular systems to explore different con-

formational states. In order to reach a more realistic model, the temperature must be taken into account within a methodology that allows to sample the phase space (*i.e.*, to explore all the possible states of the system).

Various methodologies exist to reach this goal, such as Monte-Carlo (MC) or molecular dynamics (MD) simulations. An advantage of MD over MC simulations is that the former allows to compute time related quantities. The advantages and disadvantages of those methods have been discussed in Ref. [43]. Here, we shall focus only on the molecular dynamics technique for which we will give an overview in what follows.

1.3.1 Principle

Molecular dynamics is a technique that aims at computing successive configurations of a given molecular system along time. It is based on the integration of the classical equations of motion and yields a *trajectory* which describes the time evolution of the atomic positions. Such a trajectory is used *a posteriori* to derive statistical and thermodynamical data related to the studied system.

The trajectory of a given system is obtained by integrating over time (t) and for each particle the equation related to the Newton's second law of motion, which is given by:

$$M_K \frac{d^2 \vec{R}_K(t)}{dt^2} = \vec{f}_K \quad (1.32)$$

This equation relates the time evolution of the position \vec{R}_K of a given particle (*i.e.*, an atom in the present case) having the mass M_K , to the force (\vec{f}_K) applied on this particle.

The forces \vec{f}_K appearing in Eq. 1.32 are derived from the energy of the system. One can choose any of the models described in Sections 1.1 and 1.2, as long as the first derivatives of this energy with respect to the atomic positions can be obtained. However, as we shall see in what follows, an MD simulation requires a large number of energy calculations and thus, the computational cost of the chosen method should be taken into account.

1.3.2 Ergodic hypothesis

In statistical mechanics, the average of an observable is defined by an *ensemble average*. Such a quantity is obtained by averaging a property over all the possible configurations of the system. A simulation is usually intended to treat large systems containing a large number of atoms and thus, a large phase space (*i.e.*, all the possible states/configurations of the system). To compute an average property, one should thus make sure that all the configurations of the system have been sampled along the simulation.

In molecular dynamics simulations, the phase space is explored sequentially along time and thus, *temporal averages* can be accessible. The value of an observable measured from a

laboratory experiment arises from an ensemble average. Temporal and ensemble averages are not strictly equal and this observation gives rise to a fundamental hypothesis in statistical dynamics: the *ergodic hypothesis*. This hypothesis states that if a system is simulated during an infinite time, then the phase space will be completely explored and the temporal average will become equal to the ensemble average.

The main purpose of molecular dynamics is thus to generate a sufficient and representative amount of configurations for a given system, allowing to assume that the ergodic principle is valid. Other methodologies than MD can assess statistical quantities but the main advantage of MD is that the simulation keeps track of the “history” of the system. This allows the computation of time dependent quantities, as we shall discuss in Section 1.5.

1.3.3 Integration and time step

The energy of a molecular system depends on the coordinates of all its atoms. Considering this large number of variables, the resolution of the Newton's equations of motion (Eq. 1.32) requires a numerical integration. Several methods exist such as the *Verlet*,^[44] the *velocity Verlet*^[45] or the *leapfrog* algorithm.^[46]

These algorithms are based on a Taylor series of the position (\vec{R}), the velocity (\vec{v}) and the acceleration (\vec{a}) of each particle and the evaluation of these quantities at a time $t + \delta t$ is given by,

$$\begin{aligned}\vec{R}(t + \delta t) &= \vec{R}(t) + \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 + \mathcal{O}(\delta t^3) \dots \\ \vec{v}(t + \delta t) &= \vec{v}(t) + \vec{a}(t)\delta t + \frac{1}{2}\vec{b}(t)\delta t^2 + \mathcal{O}(\delta t^3) \dots \\ \vec{a}(t + \delta t) &= \vec{a}(t) + \vec{b}(t)\delta t + \mathcal{O}(\delta t^2) \dots\end{aligned}\tag{1.33}$$

where δt is the integration *time step* of the simulation, \vec{b} is the jerk (*i.e.*, third derivative of the position) and \mathcal{O} represents higher orders of the development.

The limitation of Eq. 1.33 to the second order leads for the position to:

$$\begin{aligned}\vec{R}(t + \delta t) &= \vec{R}(t) + \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 \\ \vec{R}(t - \delta t) &= \vec{R}(t) - \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2\end{aligned}\tag{1.34}$$

By summing those two equations and by replacing \vec{a} with its expression (according to Eq. 1.32), one obtains:

$$\vec{R}(t + \delta t) = 2\vec{R}(t) - \vec{R}(t - \delta t) + \frac{1}{M_K}\vec{f}(t)\delta t^2\tag{1.35}$$

known as the Verlet integration method. Notice that in this method, the velocity is not ex-

plicitly computed and can be obtained by:

$$\vec{v}(t) = \frac{\vec{R}(t + \delta t) - \vec{R}(t - \delta t)}{2\delta t} \quad (1.36)$$

The Verlet algorithm bears several inaccuracies and is not commonly used. The velocity Verlet integrator is usually preferred and is defined as:

$$\begin{aligned} \vec{R}(t + \delta t) &= \vec{R}(t) + \vec{v}(t)\delta t + \frac{1}{M_K} \vec{f}(t)\delta t^2 \\ \vec{v}(t + \delta t) &= \vec{v}(t) + \frac{1}{2M_K} \left[\vec{f}(t) + \vec{f}(t + \delta t) \right] \delta t \end{aligned} \quad (1.37)$$

Finally, in the leapfrog algorithm, the velocities are estimated at the mid point between t and $t + \delta t$ leading to:

$$\vec{v}(t + 1/2\delta t) = \frac{\vec{R}(t + \delta t) - \vec{R}(t)}{\delta t} \quad (1.38)$$

and thus:

$$\vec{R}(t + \delta t) = \vec{R}(t) + \vec{v}(t + 1/2\delta t)\delta t \quad (1.39)$$

From Eq. 1.38, one can similarly define $\vec{v}(t - 1/2\delta t)$ to express the acceleration as:

$$\vec{a}(t) = \frac{1}{M_K} \vec{f}(t) = \frac{\vec{v}(t + 1/2\delta t) - \vec{v}(t - 1/2\delta t)}{\delta t} \quad (1.40)$$

and eventually obtain:

$$\vec{v}(t + 1/2\delta t) = \vec{v}(t - 1/2\delta t) + \frac{1}{M_K} \vec{f}(t)\delta t \quad (1.41)$$

The velocity Verlet and the leapfrog algorithms explicitly express the velocity, unlike the Verlet integrator. The advantage of the velocity Verlet method over the leapfrog algorithm is that in the former, the positions and velocities are synchronized but the latter is easier to implement in an MD program.

One can notice from Eqs. 1.39 and 1.41 that, if the positions and velocities of the atoms are known at given time steps t and $t - 1/2\delta t$, respectively, the next step leading to $\vec{R}(t + \delta t)$ and $\vec{v}(t + 1/2\delta t)$ can be obtained in a straightforward way. Only $\vec{f}(t)$ has to be evaluated at each time step to propagate the simulation. This procedure can be repeated as much as it is needed in order to reach a simulation time in agreement with the ergodic theory. It appears clearly now that the choice of the method used to compute the energy of the system (and thus \vec{f}) will limit the simulation time that can be reached.

The choice of the integrator and the Hamiltonian is of great importance to carry out a molecular dynamics simulation as much as the choice of the time step is. A too small δt will increase the computational cost while a too large δt will yield instabilities in the numerical integration procedure. The time step is usually defined by the smallest oscillation period in

the system. For example, in a molecular system, the smallest period is related to the vibrations of the hydrogen atoms and the time step will be typically set to 1 fs or less. In order to increase the value of the time step, algorithms exist to freeze the bonds involving a hydrogen atom (*e.g.*, the SHAKE algorithm[47]) allowing to set for example a δt of 2 fs and thus to double the accessible simulation time for the same amount of energy calculations.

1.3.4 Periodic boundary conditions

Molecular dynamics simulations can be performed for isolated or bulk systems (*i.e.*, gases, liquids and solids). At the atomic scale, bulk systems can be considered as infinite. However, a simulation has a finite size, though the largest simulations can contain millions of atoms. To avoid border effects, one approximation is to consider the system as periodic. A unit cell is chosen (usually cubic but not necessarily) and replicated in each direction. This method is known as *periodic boundary conditions (PBC)*.

Let us consider a system made of four particles in two dimensions (see Figure 1.1). The unit cell (box) having edges length of a_x and a_y is replicated in all directions leading to 8 neighbored replicas of the box. From this consideration it results that, when one particle leaves the unit cell during an MD simulation, it is automatically replaced by its replica com-

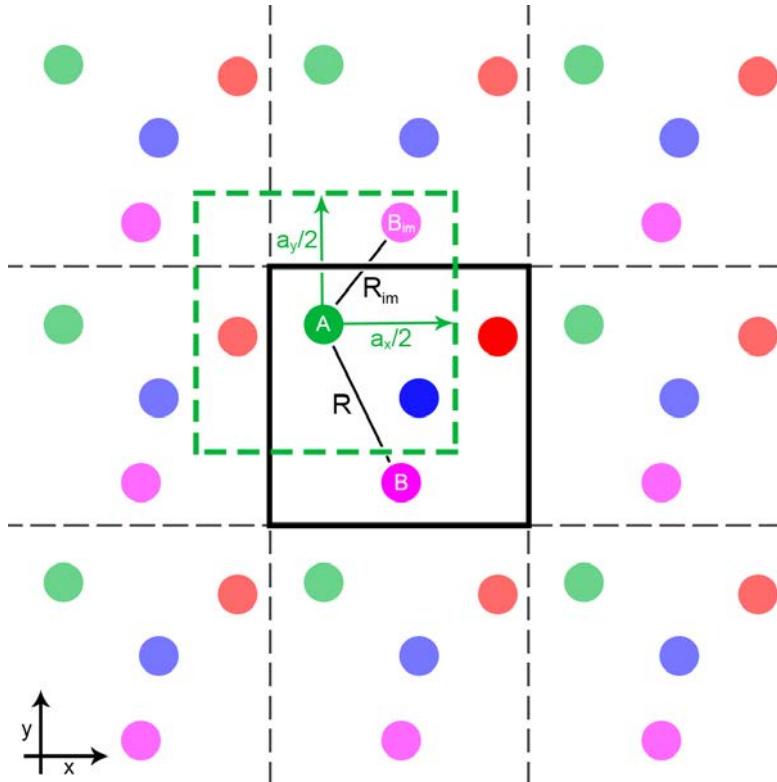


Figure 1.1: Schematic representation of a simulation box in two dimensions (x,y), using periodic boundary conditions (PBC). The unit cell having edges of length a_x and a_y (in this example: $a_x = a_y$), is represented in the middle by a black square. The green dashed square is the representation of the box around the particle A. R is the distance between A and the particle B inside the unit cell while R_{im} is the same distance after applying the minimum image convention (see text).

ing from one of the neighboring cells. This procedure allows to keep a constant number of particles in the unit cell and to simulate a finite number of particles as if they were actually placed in an infinite volume.

In practice, only the unit cell is simulated and the PBCs are applied using the *minimum image convention*. The idea behind this convention is to find the minimum distance between two particles. Keeping as an example the system depicted in Figure 1.1, let us focus on the distance between two particles, namely A and B (represented in green and purple in the Figure, respectively). The distance R between A and B in the unit cell is not the shortest distance that can be found in the system. Indeed, considering B_{im} the image of B, it appears that R_{im} , the distance between A and B_{im} is shorter than R . This can be easily seen by centering the unit cell around A (green dashed square in Figure 1.1). In other words, the minimum image convention “rebuilds” a box centered around the particle of interest to ensure that the other particles are located at the shortest possible distance.

For a molecular system, there are several ways to apply the minimum image convention. It can be based on atoms or residues (*i.e.*, predefined pieces of molecules). The former case is the one explained above while the latter considers the distances between the center of mass of each residues in the system. For example, a residue based minimum image method applied to a water molecule will translate the two hydrogen and the oxygen atoms as soon as the molecule center of mass crosses the border with another neighboring cell.

1.3.5 Short and long range electrostatic interactions: the Ewald summation

In a molecular system, electrostatic Coulombic interactions between pairs of particles are proportional to R_{ij}^{-1} (Eqs. 1.20 and 1.31). Such *long range* interactions between particles and all their infinite periodic images are given by the following Coulombic potential (V_C):

$$V_C = \frac{1}{2} \sum_{\vec{n}} \sum_i^N \sum_j^N{}' \frac{q_i q_j}{\epsilon R_{ij, \vec{n}}} \quad (1.42)$$

where \vec{n} is the translation vector of the unit cell: $\vec{n} = u\vec{a}_x + v\vec{a}_y + w\vec{a}_z$ with $(\vec{a}_x, \vec{a}_y, \vec{a}_z)$ the vectors defining the unit cell and $u, v, w \in \mathbb{Z}$, $(u, v, w) = (0, 0, 0)$ representing the original unit cell. The prime symbol in the last sum notifies that j cannot take the same value as i only if $\vec{n} = \vec{0}$. The convergence of the sum V_C is very slow because of its dependence in $R_{ij, \vec{n}}^{-1}$ and can represent a large part of the computational expense in MD simulations.

To overcome this issue, different schemes can be considered. The use of a *cutoff* will consider only the interactions within a given distance range. Outside this range, the interactions are simply neglected. Such a brute approach leads to strong discontinuities in the energy of the system and induces instabilities in MD simulations. An elegant approach is given by the

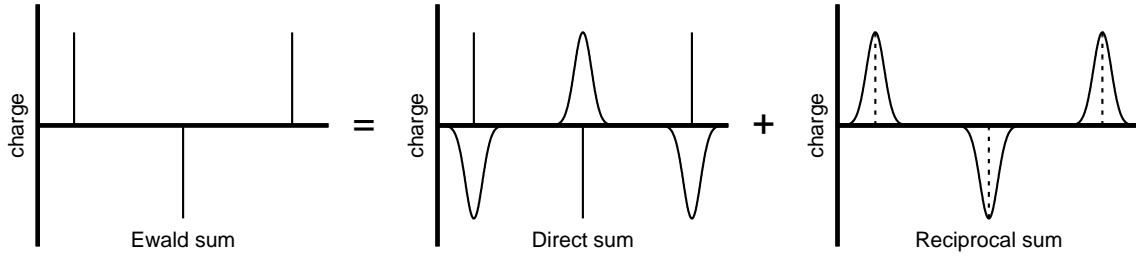


Figure 1.2: Schematic representation of the Ewald summation for a distribution of point charges along a one dimensional axis. Plain lines represent the positive and negative point charges while dotted lines indicate the position of the Gaussian function in the reciprocal space.[48]

use of *Ewald summation* (or *Ewald sum*) and we shall briefly summarize the concepts of this methodology. A complete survey of Ewald summation schemes can be found in Ref. [48].

The idea of Ewald methods is to replace V_C by a potential which consists in the sum of two series converging rapidly in the direct and in the reciprocal space:

$$\sum_{\vec{n}} \frac{1}{|\vec{n}|} F(\vec{n}) + \sum_{\vec{m}} \frac{1}{|\vec{m}|} (1 - F(\vec{m})) \quad (1.43)$$

\vec{n} and \vec{m} being the translation vectors of the direct and of the reciprocal space, respectively. $F(\vec{n})$ is a function that decays rapidly and thus makes the first sum converge quickly in the direct space. The second sum involves a smooth function for which the Fourier transform rapidly decays in the reciprocal space.

A schematic representation of the Ewald summation is represented in Figure 1.2. For the direct sum, each point charge q_i is surrounded by a Gaussian distribution ($\rho_i(\mathbf{r})$) of equal magnitude and opposite sign as,

$$\rho_i(\mathbf{r}) = -q_i \left(\frac{\beta}{\sqrt{\pi}} \right)^3 e^{-\beta^2 \mathbf{r}^2} \quad (1.44)$$

where \mathbf{r} is the distance with respect to the center of the distribution and β is a parameter that controls the width of the distribution. As it can be seen from the middle panel in Figure 1.2, this procedure screens the interaction between the charges and limits the sum in the direct space to a short range.

For the reciprocal sum, a second Gaussian distribution is added with the same magnitude and sign as the original point charge distribution to cancel out the Gaussian distribution introduced in the direct space (right hand side in Figure 1.2). This sum is performed in the reciprocal space *via* Fourier transforms to solve the resulting Poisson's equation.[49]

In practice, three parameters need to be defined to perform an Ewald summation: n_{max} and m_{max} , the maximum number of vectors used in the direct and in the reciprocal sum, respectively, and β . In large systems, the direct sum is usually limited to the first unit cell

by using a cutoff: $R_{cut} < a/2$ with a the edge length of the box. Large values of n_{max} , m_{max} and R_{cut} will yield more accurate results but might be inefficient. Finally, larger values of β will make the direct sum converge more rapidly while the reciprocal sum will converge slowly. However, efficient algorithms exist[49] to perform the latter summation and thus, a large value of β usually represents the preferred choice.

1.3.6 Thermodynamical ensembles and temperature coupling

The computation of statistical data from a simulation requires the definition of a statistical ensemble, which gives a representative distribution of the different states of the system. Molecular dynamics simulations can be performed in thermodynamic ensembles such as the *microcanonical* and the *canonical* ensemble. The former ensemble arises from the Newton's equations (Eq. 1.32) that yield systems for which the number of particles (N), the volume (V) and the total energy (E) remain constant along the simulation (NVE ensemble). In the canonical ensemble, the Newton's equations of motion are biased in order to keep the temperature (T) constant (NVT ensemble). Thus, in an NVT simulation, the total energy of the system is no longer conserved. It is also possible to perform simulations with constant pressure (P) and temperature. The resulting *isothermal-isobaric* ensemble is derived from the canonical ensemble and is referred as the NPT ensemble. Most of the simulations performed in this work use the NVT ensemble and we shall highlight the main aspects of this methodology in what follows.

Different approaches exist to maintain the temperature of a given system constant along the simulation. These methods are based on the following relationship between the kinetic energy (E_{kin}), the atomic velocities (v) and the instantaneous temperature (T):

$$E_{kin} = \frac{N_f}{2} k_B T = \frac{1}{2} \sum_K \frac{p_K^2}{M_K} \quad (1.45)$$

where N_f is the number of degrees of freedom (*i.e.*, $N_f = 3N - 6$ if no constraint are applied to a non linear system) k_B is the Boltzmann's constant and $p_K = M_K v_K$ is the *linear momentum* of the particle K. A way to keep the temperature constant is thus to adjust the velocity of each particle according to the target temperature, by coupling the system to a *heat bath via a thermostat*.

The Berendsen's thermostat[50] scales the velocities of the particles at each time step of the simulation and yield a new velocity $v_{new} = \lambda v(t)$. The corresponding variation of temperature is:

$$\Delta T = T_{new} - T(t) \quad (1.46)$$

where $T(t)$ and T_{new} are the instantaneous temperature before and after the velocity scale. From Eq. 1.45, the temperature can be expressed as a function of v and yield the following

expression of Eq. 1.46:

$$\Delta T = (\lambda^2 - 1)T(t) \quad (1.47)$$

Finally, the velocity scaling factor (λ) is obtained by combining Eqs. 1.46 and 1.47 as:

$$\lambda = \sqrt{1 + \frac{\delta t}{\tau} \left[\frac{T_{new}}{T(t)} - 1 \right]} \quad (1.48)$$

where δt is the simulation time step and τ is a parameter. This method yields a simulation in the NVT ensemble. However, such a thermostat does not conserve the linear momentum of the particles and thus loses the correlation of the successive steps of the simulation.

The Andersen's thermostat[51] uses another methodology. Rather than scaling the velocities, it uses a stochastic collision methodology. Every n_r steps, the velocity of each particle is randomly reassigned using a Maxwell-Boltzmann distribution that reflects the target temperature. This method also yield an NVT simulation and its main advantage compared to the Berendsen's thermostat is that the $n_r - 1$ steps between two velocities randomization follow a dynamics in the microcanonical ensemble (*i.e.*, NVE). Time correlated quantities are thus accessible with such an approach. Finally, the choice of n_r has to be regarded with attention. A too small value will slow down the phase space exploration while a too large value will imply a slow convergence of the temperature. In practice, small values of n_r are used during the equilibration steps of the dynamics while n_r is typically set to 1000 for production runs.

Finally, other methods to control the temperature along an MD simulation exists in the literature. One can cite as an example, the Nosé-Hoover thermostat[52, 53] or the Langevin's dynamics.[5]

1.3.7 General remarks

Molecular dynamics is a powerful tool to study the dynamical properties of a molecular system. It requires the choice of a Hamiltonian to compute the energy of the system at each time step, and this can be done at a quantum mechanics, a molecular mechanics or a mixed QM/MM level of theory (not detailed in the present manuscript, see Refs. [54, 55] for reviews). The choice of the Hamiltonian will define the type and the accuracy of the properties that can be computed as well as the length of the simulation that can be reached.

As we discussed earlier in the present Section, the value of a given observable computed from an MD simulation is related to the convergence of the phase space exploration. According to the ergodic hypothesis, if the simulation time is long enough, it can be considered that all the conformations of the system have been sampled. However, the methodology to be used in order to reach this goal remains controversial. It is indeed still not clear rather a single (very) long MD simulation or multiple shorter independent simulations should be pre-

formed to yield the convergence of the phases space sampling, especially when the number of atoms increases.[56, 57]

Finally, interesting properties related to the dynamical behavior of the system can be computed from MD simulations using various biasing techniques and trajectory analysis methodologies. We shall detail some of them in the following section.

1.4 Free energy calculation techniques

The free energy surface of a molecular system usually bears many minima. Those minima can be separated by energetically high barriers, thus preventing the system to explore other states during the simulation. Figure 1.3 shows the schematic representation of the free energy surface of a given system. Going to a state B for a system initially in a state A can be thermodynamically favorable (*i.e.*, $\Delta G < 0$) but, as it is shown in the Figure, kinetically prevented (*i.e.*, high activation barrier, ΔG^a). According to the ergodic hypothesis, a long enough molecular dynamics simulation should explore the two states. However, the length of an MD simulation is limited by the available computational power, making the exploration of the whole phase space and thus the computation of thermodynamical data impossible.

To tackle this issue, several methodologies have been developed to bias a molecular dynamics simulation in order to sample different states of a molecular system within a reasonable amount of computational time. We shall detail in what follows, the two methods used in the present work: the umbrella sampling and the metadynamics techniques. Such biased simulations allow to compute the free energy of a transformation relative to the different states of interest, by analyzing the biasing force introduced in the simulation.

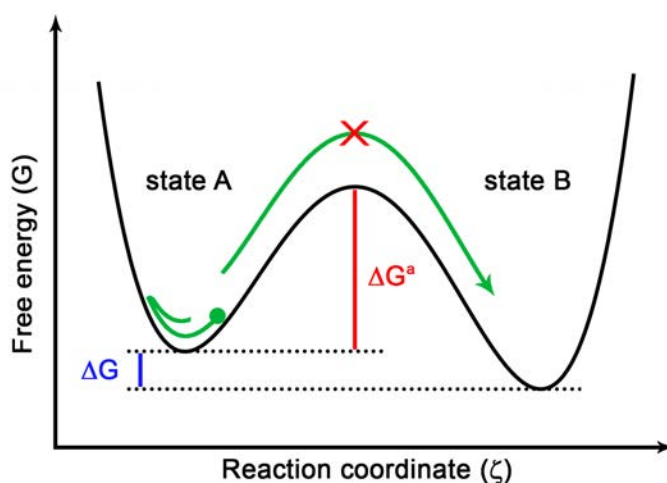


Figure 1.3: Schematic representation of a free energy (G) surface as a function of a given reaction coordinate (ζ). The activation free energy barrier is represented in red and labeled as ΔG^a while the free energy difference between the states A and B (ΔG) is displayed in blue.

1.4.1 Umbrella sampling

The free energy ($G(\zeta)$) of a molecular system as a function of a given reaction coordinate (ζ) is related to the probability ($P(\zeta)$) of finding the system in a state corresponding to ζ :

$$G(\zeta) = -k_B T \ln(P(\zeta)) \quad (1.49)$$

where k_B is the Boltzmann constant and T the temperature.

The umbrella sampling[58] technique is based on a discretization of the space into *windows*. We shall skip the details about statistical mechanics that led to the development of umbrella sampling, and focus only on the application of this technique. A schematic representation of the method is given in Figure 1.4. To enforce the system to remain in a given state, a potential V_i (usually harmonic but not necessarily) centered on each window (i) is added to the unbiased energy (E^0), leading to the biased energy (E^*):

$$E^*(\zeta) = E^0(\zeta) + \sum_{\text{windows}} V_i(\zeta) \quad (1.50)$$

An MD simulation performed using $E^*(\zeta)$ yields the biased probability P_i^* in each i window. The biased free energy is expressed from Eq. 1.49 with P_i^* and the unbiased free energy is obtained as:

$$G_i(\zeta) = -k_B T \ln(P_i^*(\zeta)) - V_i(\zeta) - F_i \quad (1.51)$$

where F_i is a constant associated to a given simulation (*i.e.*, it does not depend on ζ). The main issue that arises from Eq. 1.51 is to find the constant F_i for each window in order to rebuild the full free energy surface of the studied transformation as a *potential of mean force* (PMF).

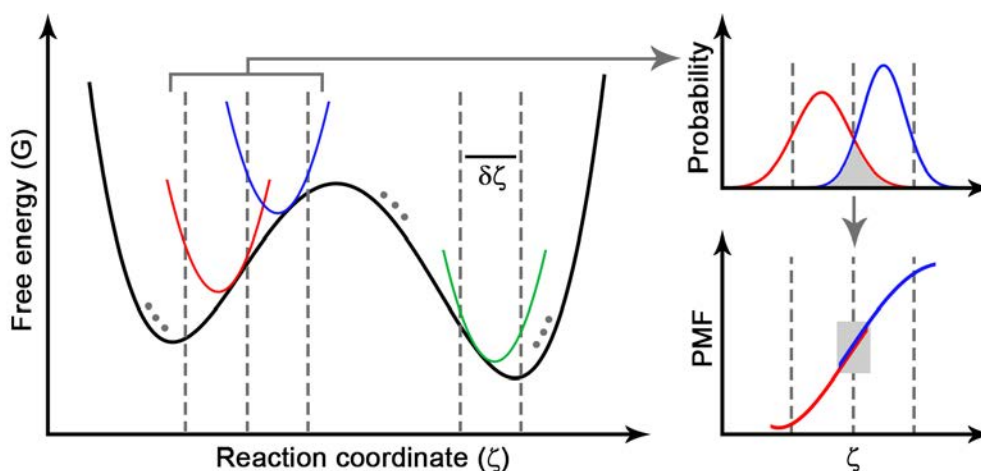


Figure 1.4: Schematic representation of the umbrella sampling technique applied to obtain the potential of mean force (PMF, bottom right panel) of a given transformation from the biased free energy profile (G , left panel) *via* the computation of the biased probability in each window (top right panel).

One of the most popular technique intended to overcome this issue is the *weighted histogram analysis method*[59, 60] (WHAM). In a series of umbrella sampling simulations that sample a given reaction coordinate, if the probabilities associated with two successive windows overlap (*i.e.*, if the same state can be found in two neighboring windows), then the free energy between those two windows should be equal. If a sufficient overlap between each successive windows is ensured, then the F_i 's can be optimized in order to minimize the difference in free energy between each neighboring window in the overlap region (see Figure 1.4 top and bottom right panels).

Practical details

The success of an umbrella sampling simulation to compute the PMF of a given transformation requires several prerequisites:

- a reaction coordinate (ζ) should be chosen.
- the width ($\delta\zeta$) of each window as well as the parameter(s) defining the bias potential must be set to have a reasonable overlap between the probability of two neighboring windows. A too small overlap would compromise the connection between two windows while a too large overlap will slow down the probability convergence in the windows.
- the maximum of the biased probability distribution must be centered in the middle of the corresponding window.

The first point is actually the most complicated when setting an umbrella sampling simulation. The reaction coordinate should reflect the reaction pathway in order to yield the lowest free energy profile. Also, the dimension of the reaction coordinate should not be too high (*i.e.*, usually not more than the combination of two/three variables) in order to ensure that the exact state of the system is controlled for each value of ζ . The choice of the reaction coordinate is system dependent and can change the shape of the resulting PMF. For the two last points, we will give some simple considerations which help to set the corresponding variables.

Let us consider Eq. 1.51 with $G_i(\zeta) = 0$ and $F_i = 0$ (*i.e.*, a flat free energy profile and the constant F_i set to zero). If the biasing potential is given by, $1/2k(\zeta - \zeta^0)^2$ with ζ^0 the center of the harmonic potential and k the force constant, then the related biased probability ($P_i^*(\zeta)$) is a Gaussian function such that:

$$P_i^*(\zeta) = e^{-\frac{k(\zeta - \zeta^0)^2}{2k_b T}} \quad (1.52)$$

A good assumption is to set the window size ($\delta\zeta$) equal to the *full width half maximum* (FWHM) of the Gaussian distribution:

$$\delta\zeta = \text{FWHM} = 2\sqrt{\frac{2\ln(2)k_b T}{k}} \quad (1.53)$$

$\delta\zeta$ can be obtained for a given force constant (k) and *vice versa*, ensuring a sufficient overlap between the successive windows. Notice that the force constant in Amber (k_A) and the one in the WHAM (k_W) program are not defined in the same way; $2k_A = k_W = k$. This last observation must be taken into account in order to avoid any misinterpretation of the resulting PMF.

The parameter that controls the position of the biased distribution maximum in each window is ζ^0 . For a given window, i , the biased free energy is given by:

$$G_i^*(\zeta) = G^0(\zeta) + V_i(\zeta) \quad (1.54)$$

If we admit that we already have a correct guess of the PMF, we can assume that,

$$G_i^*(\zeta) = \text{PMF}(\zeta) + \frac{1}{2}k(\zeta - \zeta_i^0)^2 \quad (1.55)$$

To ensure that the maximum of the biased probability in the window i will be located at the center of the window (ζ_i^c), the first derivative of G_i^* in ζ_i^c should be equal to zero. Thus,

$$\begin{aligned} \left. \frac{dG_i^*(\zeta)}{d\zeta} \right|_{\zeta=\zeta_i^c} &= \left. \frac{d\text{PMF}(\zeta)}{d\zeta} \right|_{\zeta=\zeta_i^c} + k(\zeta_i^c - \zeta_i^0) \\ &= 0 \end{aligned} \quad (1.56)$$

Then, ζ_i^0 is simply obtained by,

$$\zeta_i^0 = \zeta_i^c + \frac{1}{k} \left. \frac{d\text{PMF}(\zeta)}{d\zeta} \right|_{\zeta=\zeta_i^c} \quad (1.57)$$

In practice, the PMF can be obtained iteratively. As a first guess, the umbrella sampling procedure can be performed by setting $\zeta_i^0 = \zeta_i^c$ for each window. The obtained PMF is then interpolated (*e.g.*, by a polynomial function or a spline) and used to find a better set of ζ_i^0 's according to Eq. 1.57. This procedure is then repeated until a satisfactory convergence of the PMF is achieved.

1.4.2 Metadynamics

Metadynamics[61, 62] is another way to force a system to explore different states of the free energy surface (FES), which would have been barely or not accessible otherwise. The method focuses on a small part of the coordinates space defined by a few collective coordinates (*i.e.*, a subset of the $3N-6$ coordinates of a system containing N particles). Those coordinates are commonly named *collective variables* (σ) and are chosen to define the process of interest (*e.g.*, distances, angles, dihedral angles, coordination number as well as combinations of

those coordinates).

In this method, a standard MD simulation is biased by the addition of a *history-dependent* potential.[61, 63, 64] This potential applies a penalty function to the visited “places”, forcing the system to explore other states of the collective variables subspace. A schematic representation of the method applied to a one dimensional subspace is given in Figure 1.5. Every τ_G steps of the dynamics, a Gaussian function is added, centered at the instantaneous value of σ . The expression of the resulting biasing potential (V_G) is:

$$V_G(\sigma, t) = W \sum_{\substack{t' = \tau_G, 2\tau_G, \dots \\ t' < t}} e^{-\frac{(\sigma - \sigma_{t'})^2}{2\delta s^2}} \quad (1.58)$$

where t' is the collection of the time steps at which a Gaussian function has been added and W and δs are the height and the width of the Gaussian functions. Figure 1.5 presents the shape of the biased FES at three successive moments of the simulation, separated by $n\delta t$ and by $(n+m)\delta t$ (with $n, m \in \mathbb{N}$ and δt the time step of the simulation). At $t = t_0$, the well is not completely filled and the system remains “trapped” in its initial state. After n time steps, n/τ_G Gaussian functions have been added to the biasing potential, thus allowing the system to escape the local minimum and to begin exploring the next one. The metadynamics methodology can be seen as a flooding of the FES wells by Gaussian functions.

It is interesting to notice that, using this methodology, the system will go from one state to another by using the lowest energy pathway in the collective variables subspace. In the example presented in Figure 1.5, when the two wells are completely filled, the system diffuses between the two states without any barrier. The simulation can thus be stopped and the relative free energy surface (ΔG) is recovered by analyzing the total biasing potential (V_G). This potential represents a molding of the FES and the envelop of this molding gives approximately the relative FES (*i.e.*, $-V_G(\sigma) \simeq \Delta G(\sigma)$).

Compared to the umbrella sampling technique, metadynamics has the advantage that the transformation from a state A to a state B can be performed without any assumption of the pathway. However, as it is the case with umbrella sampling, the present method requires

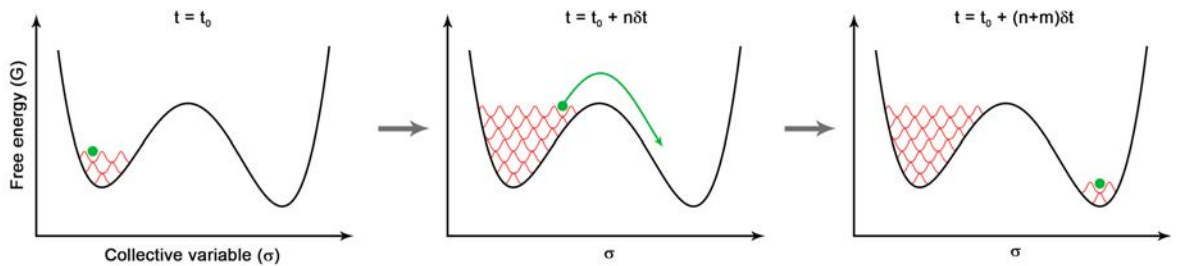


Figure 1.5: Schematic representation of the metadynamics technique. The successive steps separated by $n\delta t$ and $(n+m)\delta t$ (with $n, m \in \mathbb{N}$ and δt the time step of the simulation) show the flooding of the free energy surface with Gaussian functions. The green dot represents an instantaneous state of the system.

the setting of several parameters which directly impact the shape of the resulting FES. The definition and the number of collective variables must be chosen carefully to ensure that the obtained FES reflects the lowest energy pathway. As for the umbrella sampling technique, the dimension of the collective variables has to be regarded with attention. Also the height and the width of the added Gaussian functions will impact the quality of the FES. Finally, τ_G has to be chosen large enough to allow the system to equilibrate after the addition of a Gaussian but a too large value will increase the required computational time.

Many different types of collective variables can be envisaged. However, the use of a new collective variable in an MD program requires the explicit implementation of its first derivative. A large collection of collective variables has been developed as a plugin (PLUMED[65]) interfaced with most of the popular MD codes. Nevertheless, some collective variable require the use of a *smoothing function* (*i.e.*, a function that goes from one to zero continuously) in their definition (*e.g.*, coordination number) and many of those functions exist. In Chapter 3, we shall introduce an elegant way to implement a general definition of the smoothing function.

1.5 Trajectory analysis tools

Macroscopic properties can be retrieved from a molecular dynamics simulation. While some quantities are not “history dependent” and only require a large collection of configurations, such as *radial* or *spherical distribution functions*, some others are related to the *time-correlation function* of a given property (*e.g.*, the computation of infrared spectra from the dipole moment time-correlation function). In what follows, we shall detail some of the tools used to analyze a molecular dynamics simulation.

1.5.1 Radial distribution functions

The radial pair distribution function (RDF or $g(r)$) can be obtained from computational simulations and (in some cases) experimentally.[66] It gives a representation of the condensed matter structure, describing the arrangement of given particles around a specific site compared with an ideal distribution.

We shall give here a general definition of a site-site RDF,[5, 67] *i.e.*, for a homogeneous atomic distribution (*e.g.*, argon gas) or molecular distribution (*e.g.*, liquid water) as well as for binary, tertiary,... mixtures (*e.g.*, one solute in a box of water molecules). We define two types of sites (*typ1* and *typ2*) and N_1, N_2 the number of sites per type. A site can be an atom or the center of mass of a molecule/residue. The RDF is given by,

$$g(r) = \lim_{dr \rightarrow 0} \frac{n(r)/N_p}{4\pi r^2 dr/V} \quad (1.59)$$

with r and dr the distance between the two sites and an infinitesimal interval, respectively. The function $g(r)$ is normalized with respect to an ideal distribution of particles where V is the total volume of the system, $n(r)$ the average number of site pairs within a range of distance between r and $r + dr$, and N_p is the total number of site pairs. The definition of N_p differs from one calculation to another. If the two sites belong to the same type (*i.e.*, $typ1 = typ2$), then

$$N_p = N_1(N_1 - 1) \quad (1.60)$$

If $typ1 \neq typ2$, thus

$$N_p = N_1 N_2 \quad (1.61)$$

For example, considering a system made of N_w water molecules, the number of hydrogen (H) atom pairs is $N_p^{HH} = 4N_w^2 - 2N_w$ (*i.e.*, Eq. 1.60 with $N_1 = 2N_w$). If we consider one methane molecule in a box of N_w water molecules, $N_p^{CO} = N_w$ with N_p^{CO} the number of carbon-oxygen (CO) pairs.

In a molecular dynamics simulation, $n(r)$ is computed from a temporal average:

$$n(r) = \frac{1}{N_s} \sum_i^{N_s} \sum_j \sum_k \delta(r - r_{ijk}) \quad \left| \begin{array}{l} j \in typ1 \\ k \in typ2; \quad k \neq j \end{array} \right. \quad (1.62)$$

with N_s the number of simulation steps, δ the Dirac delta function and r_{ijk} the distance between the two sites j and k in the frame i . In practice, the continuous function $n(r)$ is discretized into an histogram with the step Δr .

An RDF usually presents a succession of maxima and minima. Those oscillations are smoothly attenuated when r increases to finally converge to 1, representing a non structured fluid (*i.e.*, an ideal distribution). The maxima represent a high local density of particles around the site of interest while a minimum is related to a depletion of this density. For liquid systems, the successive peaks give information about the local structure in the first, second, ... solvation shells, and about their size. It can be interesting in some cases to compute the average number of sites present in the first solvation shell (*e.g.*, to know the number of hydrogen bonds formed by the oxygen atom of a given water molecule in liquid water). This can be done by expressing $n(r)$ from Eq. 1.59 and by integrating the resulting expression between the beginning and the end of the first peak (r_0 and r_1 , respectively). This gives rise to the definition of the *coordination number* (n^c) as,

$$n^c = \frac{N_p}{V} \int_{r_0}^{r_1} g(r) 4\pi r^2 dr \quad (1.63)$$

As any temporal average, computing RDF's requires a sufficiently large sampling. If the RDF of two sites belonging to the same type (*e.g.*, water-water RDF) converges rapidly because of the important number of sites (ensemble average), it converges much slowly in the

case of a single site of reference (*e.g.*, one solute molecule in liquid water). From our experience, the former case will typically require only tens of picoseconds of MD simulations while hundreds will be needed in the latter case.

For a solute molecule in a given solvent, the radial distribution function gives a one dimensional representation of the solvent structure resulting from an average in all the space directions. Depending on the symmetry of the system, this can sometimes be insufficient since the distribution might be locally different from one direction to another. The use of *spatial distribution functions* can overcome this issue and we shall detail this methodology in what follows.

1.5.2 Spatial distribution functions

A spatial distribution function (SDF) [68] shares similarities with the definition of RDF. It is used to compute the density of all the particles (or sites, S_{mob}) of a given type around a site of reference (S_{ref}). However, unlike the RDF, which only contains information about the radial component of the distribution, the SDF is defined in three dimensions and thus also takes into account the angular coordinates.

To compute an SDF, the reference site of interest is fixed in a local frame and a spacial grid of unitary volumes (δV) is constructed around it. For all the different configurations of the system, the number of occurrences (hits) of S_{mob} in each δV are counted. Because the position of each unitary volume is defined in space with respect to the origin of the grid (*i.e.*, S_{ref}), the spatial distribution of S_{mob} around S_{ref} is obtained.

The number of hits per δV is usually normalized in a similar way to what it is done for RDFs, in order to help the interpretation of the results. The normalization factor (*fact*) used in this work was adapted from Ref. [69] and is given by,

$$fact = \left[N_{sym} N_s \frac{N_p}{V} \delta V \right]^{-1} \quad (1.64)$$

where N_p , N_s and V have the same definition as in the previous Subsection, and N_{sym} is the number of symmetry operations considered to perform the analysis. Using such a definition of *fact* yields a density in each unitary volume normalized with respect to an ideal distribution of particles (atoms or molecules). A value larger than one shows a locally stable structure around the site of interest while values smaller than one indicate a depletion of the particles density.

Compared to an RDF, the computation of an SDF requires an even larger sampling. This comes from the definition of the unitary volume used in both cases. For an RDF, $\delta V = 4\pi r^2 dr$ while for an SDF, δV is fixed by the definition of the grid. The former is usually much larger than the latter, making the average in each bin of the histogram converge faster in the case of

an RDF. However, the symmetry of the system can be taken into account in order to increase the sampling when computing an SDF.[69]

Finally, an SDF can be represented in three dimension around the site of interest by plotting surfaces of iso-density. It can also be projected in a relevant plane allowing to visualize *volume slices* in three dimensions, the third dimension being the density.

1.5.3 Infrared spectroscopy

As we discussed in Section 1.3, a molecular dynamics simulation performed in the NVE ensemble keeps track of the history of the trajectory. This allows to compute time correlated quantities such as the vibrational properties of a given system. Infrared (IR) spectroscopy is widely used experimentally to probe the structure of complex systems and this analysis can be performed based on an MD simulation.

The IR absorption line shape $I(\nu)$, where ν is the absorption frequency, is related to the time correlation function (TCF) of the quantum mechanical dipole moment ($\hat{\mu}$) through a Fourier transform operation as:[70, 71]

$$I(\nu) \sim \int_{-\infty}^{+\infty} dt e^{-i2\pi\nu t} \frac{\hat{\text{Tr}} \left[e^{-\hat{H}/k_B T} \hat{\mu}(0) \cdot \hat{\mu}(t) \right]}{\hat{\text{Tr}} \left[e^{-\hat{H}/k_B T} \right]} \quad (1.65)$$

The expression of $I(\nu)$ in Eq. 1.65 can be approximated by using the classical analog of the TCF and by replacing the QM dipole moment operator by the classical dipole moment $\vec{\mu}$ leading to:[72]

$$I(\nu) \sim Q(\nu) \int_{-\infty}^{+\infty} dt e^{-i2\pi\nu t} \langle \vec{\mu}(t) \cdot \vec{\mu}(0) \rangle \quad (1.66)$$

where $\langle \vec{\mu}(t) \cdot \vec{\mu}(0) \rangle$ is the time correlation function. To compute this quantity from an MD simulation, the following discrete definition is used:[5]

$$\langle \vec{\mu}(t) \cdot \vec{\mu}(0) \rangle = \frac{1}{t_{max}} \sum_{t_0=1}^{t_{max}} [\mu_x(t_0)\mu_x(t_0+t) + \mu_y(t_0)\mu_y(t_0+t) + \mu_z(t_0)\mu_z(t_0+t)] \quad (1.67)$$

with t_{max} the maximum number of time steps and $(\mu_x(t), \mu_y(t), \mu_z(t))$ the components of the instantaneous dipole moment. Computing this TCF numerically can represent a large amount of computational time, especially for long simulations. However, several methods exist to estimate the Fourier transform of a TCF for a much lower computational expense.[73] Over all those methods, it has been shown[74] that the *maximum entropy method* gives satisfactory results for the application to IR spectra calculations.

In Eq. 1.66 Q is a correction factor introduced to compensate the approximation introduced by the classical definition of the different operators. Because of this approximation, the intensity $I(\nu)$ does not match the experimental prediction. However, the position of the

resulting absorption bands is usually correct. Different versions of Q have been proposed in the literature (see Ref. [75] for a survey). Though the used of this factor does not significantly affect $I(\nu)$, it usually yields a better ratio between the intensities related to the low and to the high frequencies. In this work, the IR spectra will be calculated using the so-called “harmonic” factor[75–77] defined as:

$$Q(\nu) = \frac{\beta h \nu}{1 - e^{\beta h \nu}} \quad (1.68)$$

with β being the invert of the product of the Boltzman’s constant with the temperature.

Computation of IR spectra from NVT simulations

The computation of time correlated quantities from an MD simulation requires this simulation to be performed in the NVE ensemble. As we discussed in Section 1.3.6, the use of a thermostat during an NVT simulation implies the loss of correlation each time the velocities are modified. However, we have shown in the same Section that some thermostats (*e.g.*, the Andersen’s thermostat) do not change the velocities at each time step of the simulation but only every n_r time steps. Thus, the $n_r - 1$ steps between two velocities modification follow a Newtonian dynamics (*i.e.*, in the NVE ensemble). We shall describe here a methodology which allows to compute an IR spectrum from an NVT simulation performed using the Andersen’s thermostat.[74]

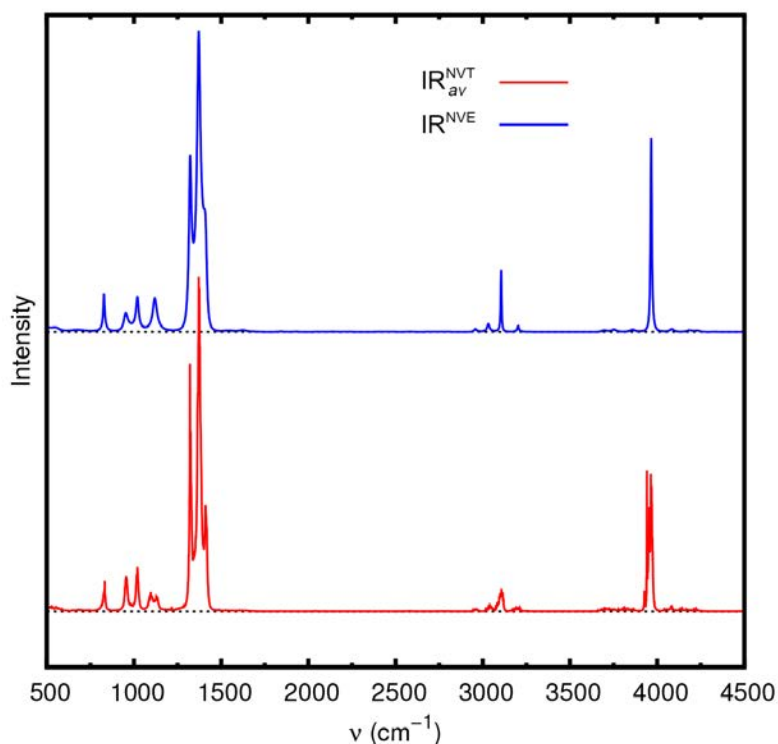


Figure 1.6: Infrared spectra computed from a 500 ps gas phase SEBOMD simulation of ethanol in the NVE ensemble (IR^{NVE}) and in the NVT (IR_{av}^{NVT}).

From an NVT simulation made of n_T time steps, n_T/n_r independent simulations of $n_r - 1$ steps can be considered as performed in the NVE ensemble. Thus, n_T/n_r IR spectra (IR_i^{NVT}) can be computed. Compared to the IR spectrum (IR^{NVE}) obtained from an n_T steps long NVE simulation, the IR_i^{NVT} 's are far less converged because of the loss of statistics in the latter case. However, it has been shown in our group from a simulation of the N-methylacetamide molecule in the gas phase,[74] that the average of all the n_T/n_r independent spectra computed from an NVT simulation yields a spectrum $\text{IR}_{av}^{\text{NVT}}$ equivalent to IR^{NVE} obtained from the NVE counterpart simulation (with $n_r = 1000$ and $n_T = 300000$).[74] The loss of statistics discussed above is thus compensated by the large amount of IR_i^{NVT} 's spectra that are taken into account.

We give here another example for two gas phase SEBOMD simulations of the ethanol molecule in the gas phase performed in the NVE and in the NVT ensemble with $n_r = 1000$ and $n_T = 500000$ and the PM3[78] semiempirical Hamiltonian (see Chapters 3 and 5 for more details about the SEBOMD method and the simulation protocol, respectively). Figure 1.6 shows the infrared spectra obtained from the NVE and NVT simulations computed using the methodology described above (*i.e.*, IR^{NVE} and $\text{IR}_{av}^{\text{NVT}}$, respectively). As we can see from the Figure, though $\text{IR}_{av}^{\text{NVT}}$ is more noisy than IR^{NVE} , the line shape of both of the two spectra presents the same features.

Convergence of the IR spectrum

A common issue when facing the computation of properties arising from an ensemble or a temporal average is to make sure that the simulation performed is converged. We propose here to address this question in the case of the computation of the IR spectra in the condensed phase.

As we shall discuss in Chapter 3, only NVT simulations can be performed in the SEBOMD framework with PBC in the current state of the method development. Thus, the convergence of the spectrum ($\text{IR}_{av}^{\text{NVT}}$) will be tested against the number of simulation time steps and not with respect to IR^{NVE} . In what follows, we shall focus on the IR spectrum of a given solute in water. Such a calculation rigorously requires to take into account the solute-solvent dipole moment *cross-correlation* and the solute-solute correlation TCF.[79] However, this would also require a much better statistics and we will restrict our calculations to the self-correlation of the solute dipole moment as it has been suggested in the literature.[74, 80]

The IR spectrum of a single solute was computed from the SEBOMD simulation of this molecule in a box of 128 water molecules with PBC in the NVT ensemble using the PM3-PIF3 Hamiltonian (see Chapters 5 and 4 for details about the protocol and the Hamiltonian). We consider two different solutes, *i.e.*, ethanol and N-methylacetamide (NMA).

Figure 1.7 presents the IR spectra of ethanol and NMA as a function of the simulation length. We shall not describe the specific features of those spectra since Chapter 5 is ded-

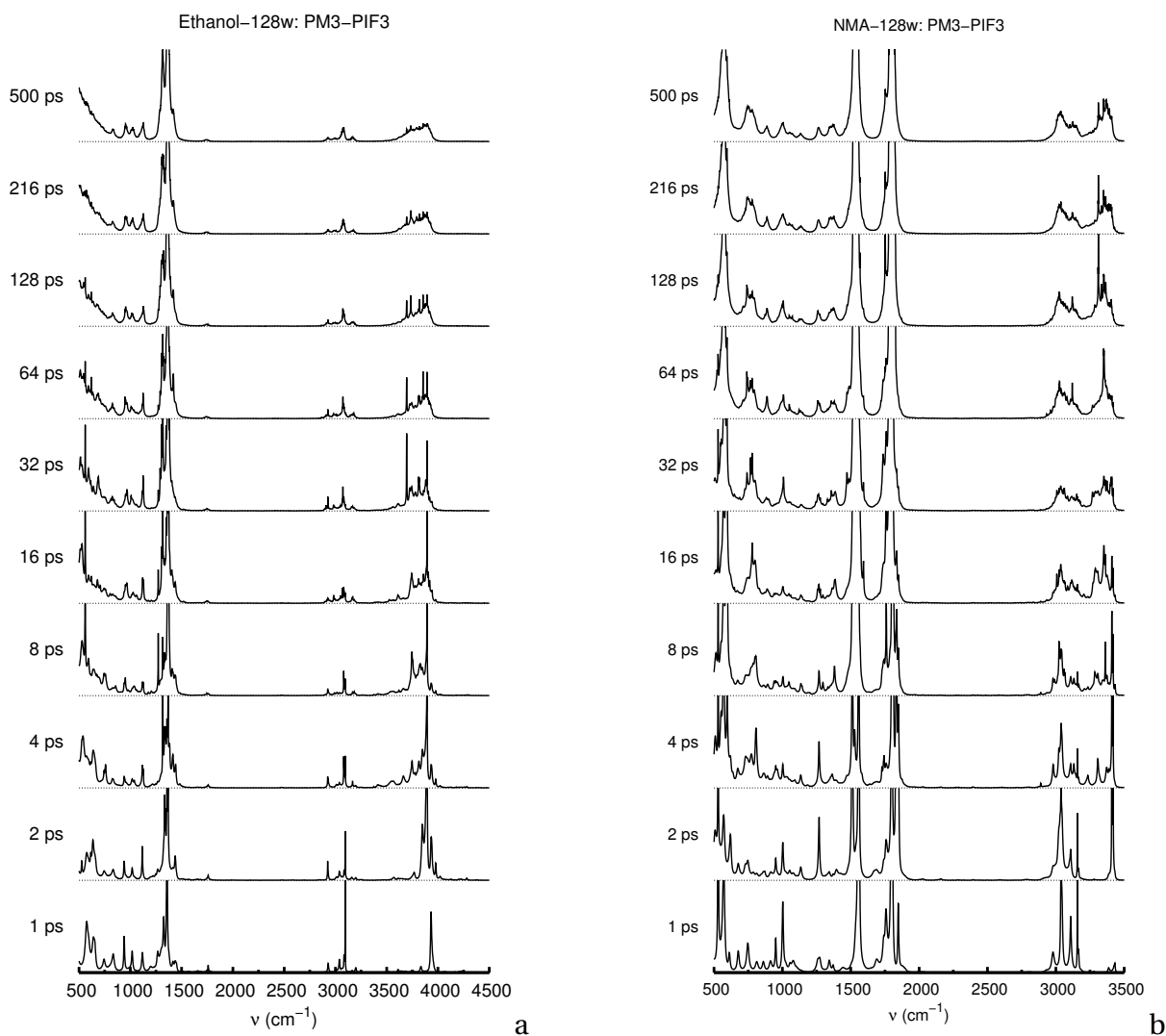


Figure 1.7: Convergence of the IR spectrum as a function of the simulation time. The simulation were performed using the SEBOMD method and the PM3-PIF3 Hamiltonian in the NVT ensemble. In each case, one solute is solvated in a box of 128 water molecules. a: ethanol. b: N-methylacetamide (NMA).

icated to this purpose. Let us focus first on the IR spectrum of ethanol (Figure 1.7a). If we concentrate on the high frequencies region (*i.e.*, between 3500 and 4000 cm^{-1}), we observe that after one picosecond a single peak appears. After four picosecond of simulation, this peak starts to broaden. It is only after 16-32 ps that the line shape starts to converge and it appears completely converged after 216 ps (since the spectrum does not change if we double the simulation time).

Similar observations can be drawn from the IR spectrum of NMA (Figure 1.7b). However, in this case, not only the width of the bands requires time to converge, but we also observe a peak appearing (at about 3400 cm^{-1}) after two picoseconds, which was not present considering only one picosecond of simulation.

This example shows that the convergence of the IR spectrum requires at least about 30 ps of simulation for the system considered here. It also shows that a simulation of 500 ps is sufficient to compute the IR spectra of simple molecules in water.

1.6 Concluding remarks

We have presented in this Chapter the general methodology related to the present work. This work combines different computational techniques, from the quantum description of the electronic structure to the modeling of dynamical properties.

We have discussed here an overview of quantum chemistry techniques starting from the Born-Oppenheimer approximation and the Hartree-Fock method as well as more elaborated schemes. The molecular dynamics technique plays a key role in the following Chapters and we gave here the basic concepts of this method. In addition, we detailed some of the relevant calculation and analysis tools used in this work. For each of them, we gave practical details and examples to illustrate the methodology.

Chapter

2

Semiempirical methods

Résumé

Les méthodes semiempiriques en chimie quantique sont au cœur de ce travail et le chapitre présenté ici leur est dédié. Nous parcourons tout d'abord l'historique de ces méthodes en soulignant les développements clés. Puis, étant donné que ce travail de thèse s'articule autour des méthodes dérivées de MNDO, nous en présentons les détails théoriques. Enfin, nous expliquons les motivations et les équations ayant conduit à la mise en œuvre de corrections spécifiques des méthodes semiempiriques, notamment pour améliorer le traitement des interactions intermoléculaires. Nous nous attardons sur deux méthodologies en particulier, PIF et MAIS, qui ont été et sont encore développées au sein de notre groupe.

Depuis les premiers travaux de Hückel en 1931, jusqu'au développement des méthodes approximées les plus élaborées, nous retraçons ici les grandes lignes du développement des méthodes semiempiriques. Ce développement a notamment été marqué par la série d'approximations énoncées par Pople dans les années 1960. La base de ces approches est de négliger complètement (CNDO, *complet neglect of differential overlap*) ou en partie (NDDO et INDO, *neglect of diatomic differential overlap* et *intermediate neglect of differential overlap*, respectivement) le recouvrement entre deux orbitales atomiques différentes. Parmi les méthodes dérivées de ces approximations, celles basées sur l'approche MNDO sont les plus courantes. D'autres méthodes impliquant une reformulation des approximations de Pople ont vu le jour à la fin des années 1990, telles que les modèles OMx. On peut également noter l'utilisation toujours importante aujourd'hui dans le domaine de la spectroscopie de la méthode ZINDO dérivée de l'approximation INDO.

La méthode MNDO développée par Dewar à la fin des années 1970 a rapidement montré certaines limitations quant au traitement des interactions intermoléculaires. Les approches dérivées de MNDO, telles que AM1 et PM3, ont cherché à corriger ce défaut en introduisant une somme de fonctions Gaussiennes permettant d'améliorer la description du terme de répulsion entre les noyaux. Cependant, il a été démontré que l'introduction de ce type de fonctions entraîne la présence d'artéfacts sur la surface d'énergie potentielle d'interaction de nombreux complexes moléculaires.

Enfin, différentes approches, destinées à résoudre le problème des interactions inter-

moléculaires en semiempirique, ont été proposées dans la littérature. Deux d'entre-elles ont notamment été développées dans notre groupe. Ces approches (PIF et MAIS) sont basées sur deux principes : remplacer l'utilisation de fonctions Gaussiennes et effectuer une paramétrisation plus complète de la surface d'énergie d'interaction entre molécules d'eau et composés organiques à partir de données quantiques *ab initio*. Elles permettent ainsi de corriger des erreurs importantes, commises sur des structures moléculaires hors équilibre et/ou en interaction, obtenues avec la plupart des méthodes semiempiriques (NDDO) actuelles qui sont paramétrées principalement à partir de géométries correspondant à des minima d'énergie.

We described, in Section 1.1, *ab initio* electronic structure methods to solve the Schrödinger equation. To evaluate the electronic energy of a given molecular system, these methodologies involve the calculation of a large number of *mono* and *bielectronic integrals*. This represents the most time consuming part of such a calculation beside requiring a self-consistent field (SCF) procedure. Since the early times of computational chemistry, approximations of *ab initio* theories (sometimes severe) have been necessary to predict the properties of “large” molecular systems (the term “large” should be intended as related to the available computational power). The idea behind semiempirical (SE) methods is to formulate approximations and to avoid the computation of several terms by replacing them with *parameters*, either derived from experiments or optimized to reproduce empirical/*ab initio* data.

To understand the state of current methodologies and to locate the present work in the field of the SE method development, we shall present in what follows a brief overview of the main breakthroughs in this area. For more complete references about this topic, the reader can refer to the reviews in Refs. [81–85]. This Chapter is not expected to be exhaustive but to present the issues faced and the answers brought by the developers along time, in order to reach a good balance between accuracy and efficiency in order to treat systems of larger and larger size. A time line is presented in Figure 2.1 in order to illustrate the following discussion.

2.1 Historical overview

In 1931, Hückel formulated[86] one of the first approximation of the secular equation (Eq. 1.11). This method was devoted to the study of unsaturated and aromatic hydrocarbons and was proven to give valuable predictions of the structure and spectroscopy of such compounds. In this approach, a basis set made of parallel $2p$ atomic orbitals (one per carbon atom) is used and the overlap matrix is set to unity. The key points of this methodology are that most of the resonance integral are neglected and that the remaining ones are not explicitly computed but parameterized from experimental data. Hückel’s theory has been extended later by Hoffmann[87] to a minimal basis set of valence orbitals and was still limited to applications on π -conjugated systems.

In the mean time (1953), Pariser, Parr[88] and Pople[89] developed a method (PPP) based on the *zero-differential overlap approximation* (ZDO), which was first formulated by Parr[90] in 1952. The PPP method was also mainly applied to π -electrons systems. Unlike Hückel’s theory and derivatives, the PPP methodology, as well as those following it, explicitly accounts for some of the two-electron integrals in the SCF procedure.

Based on this work, Pople *et al.* formulated, in 1965, two ways of rationalizing the ZDO approximation in order to keep the invariant property of SCF molecular orbital methods with respect to transformations in the atomic orbital basis set (*e.g.*, rotation of the axis, hy-



The first parametrization of the methodologies introduced by Pople *et al.* were done to

reproduce Hartree-Fock (HF) *ab initio* calculations and gave rise to the CNDO/2[98] and INDO methods. Thus, such approaches could, at best, reach the (limited) level of accuracy of HF predictions.

Between 1969 and 1977, Dewar *et al.* performed the parameterization of INDO- and NDDO-based methods with respect to empirical results. The MINDO/1-3[102–104] and the MNDO[114] methods arose from this strategy, based on the INDO and on the NDDO approximation, respectively. MINDO and MNDO were intended to reproduce and predict the ground state properties of organic molecules. The MNDO method was shown rapidly to fail at describing intermolecular interactions (*e.g.*, hydrogen bonds) and two later corrections and reparameterizations have been derived from MNDO to overcome this issue: AM1[115] by Dewar *et al.* and PM3[78] by Stewart. AM1 and PM3 have been some of the most popular ones for applications to biomolecular and to solid state systems until the early 2000's.

For spectroscopic applications, two CNDO- and INDO-based Hamiltonians were developed between 1968 and 1992: CNDO/S[99, 100] and INDO/S[105] (also referred as ZINDO for “Zerner’s INDO”), respectively. The latter, developed by Zerner *et al.*, which was extended to the parameterization of transition metals and lanthanides,[106, 107] is still in use nowadays for excited states applications (CIS) as recently reviewed in Ref. [84].

Over the methods cited above, MNDO-based approaches are those that experienced and still experience the largest improvements since 1990's. The basis set has been increased to *spd* atomic orbitals leading for example to MNDO/d[145–147], AM1/d[118] and later to the last versions of the PMx model, PM6[131] and PM7.[141] Further extensions or reparameterization of these methods have been developed, such as AM1/d-PhoT[128] and AM1*.[119–126] Other approaches have induced some changes in the *core-core repulsion function*, associated with a full reparameterization (*e.g.*, PDDG/MNDO and PDDG/PM3,[133–135] PM3-CARB1[136]) or only a straight reparameterization of an existing methodology (*e.g.*, RM1,[127] which was derived from AM1 and mainly dedicated to biological applications).

Although the AM1 and PM3 models did improve the wrong prediction of intermolecular interactions of MNDO, those methods were still not accurate enough to be applied to the study of the condensed phase. Notably, it has been shown that PM3 predicts artifacts in the intermolecular *potential energy surface* (PES), such as a spurious minimum for the interaction between two hydrogen atoms at about 1.7 Å.[168] We will show later in this manuscript that other methods present similar artifacts. To overcome this issue, different schemes have been proposed in the literature, based on the correction of the core-core repulsion function while keeping the electronic part of the Hamiltonian intact. The PIF[132, 138–140] (Parameterizable Interaction Function) and MAIS[137, 139] (Model Adapted for Intermolecular Studies) strategies (detailed in Section 2.3), which were partly developed in our group, are some of those and have been proven to yield good results for water-water and water-organic compounds intermolecular interactions.[74, 140, 169] Martin and Clark[161] adapted the disper-

sion term developed for DFT methods to AM1 and PM3 (-D correction). Later, Hobza's group developed additional terms to the -D correction in order to treat hydrogen bonded system: -DH,[163] -DH2(X),[164, 165] -DH+[166] and -D3H4(X).[167]

Other approaches, developed by Thiel *et al.*, have gone “beyond the MNDO model” by introducing orthogonalization corrections to the NDDO approximation. This gave rise to the OM1,[149] OM2,[150, 151] and OM3[152] models.

Finally, other schemes of semiempirical approaches have been developed, in parallel to the methods discussed above, based on the early work of Slater: the *tight binding* (TB) theory. These methodologies have been mainly dedicated to study solid states until the 1990's. In 1995, Porezag *et al.* adapted such an approximation to the density functional theory, leading to the *density functional tight binding* (DFTB) method.[154] DFTB has been improved later by Elstner *et al.* through a *self-consistent charge* (SCC) model: SCC-DFTB.[155] This methodology has been shown to provide good results for biological and material science applications[170–172] and it represents an active field in approximate method developments (*e.g.*, DFTB3,[157] DFTB-CPE[158]).

The present work is restricted to the test and applications of NDDO-based methods to model the dynamics of biological systems in water. We shall now detail the NDDO approximation and highlight the treatment of the electronic and core-core parts of the Hamiltonian for the MNDO, AM1 and PM3 methods.

2.2 NDDO-based methods: MNDO, AM1 and PM3

MNDO, AM1 and PM3 share the same methodological framework. All three methods are derived from the neglect of diatomic differential overlap (NDDO) approximation. Only the parameters (as well as the way they have been derived) and the expression of the core-core energy change from one method to another. We shall detail in what follows the NDDO approximation of the electronic energy and the way it has been adapted to MNDO, AM1 and PM3. Thus, we will briefly discuss the formulation of the core-core repulsion term for those three methods.

2.2.1 The NDDO approximation

As we commented earlier, the *neglect of diatomic differential overlap* (NDDO) is one of the approximations formulated by Pople *et al.*[91] It is based on the concept of *zero differential overlap* (ZDO) developed by Parr,[90] in which the product of two different atomic orbitals (ϕ_μ and ϕ_ν) is neglected:

$$\phi_\mu \phi_\nu = 0 \quad \text{if } \mu \neq \nu \quad (2.1)$$

Consequently, the overlap integrals ($S_{\mu\nu}$, the elements of the overlap matrix) are given by:

$$S_{\mu\nu} = \delta_{\mu\nu} \quad (2.2)$$

Thus, the overlap matrix \mathbf{S} becomes unitary and the N Roothaan-Hall equations (Eq. 1.23) for the N molecular orbitals are simplified to:

$$\sum_{\nu} (F_{\mu\nu} - E_i \delta_{\mu\nu}) c_{\mu i} = 0 \quad \forall i \in N \quad (2.3)$$

where E_i is the energy of the molecular orbital ϕ_i , that is defined as a linear combination of atomic orbitals (ϕ_{μ}) and $c_{\mu i}$ the coefficients of this combination (Eq. 1.10). The electronic energy is given by Eq. 1.26 and can also be written as:

$$E_{el} = \frac{1}{2} \sum_{\mu} \sum_{\nu} P_{\mu\nu} (H_{\mu\nu}^c + F_{\mu\nu}) \quad (2.4)$$

where $H_{\mu\nu}^c$ are the elements of the core Hamiltonian and $P_{\mu\nu}$, the elements of the density matrix. The elements of the Fock matrix ($F_{\mu\nu}$, defined in Eq. 1.24), involve a large number of electron repulsion integrals, $(\mu\nu|\lambda\sigma)$ (Eq. 1.25), and most of them are neglected under the ZDO approximation.

The NDDO approximation remains invariant under transformations of the reference axis or of the basis set. NDDO-based methods use a minimal basis-set of *Slater-type atomic orbitals* (STO). The nuclei are replaced by *cores* that include the core electrons.

In what follows, we will consider that the set of atomic orbitals ϕ_{μ} and ϕ_{ν} belongs to an atom A while the set ϕ_{λ} and ϕ_{σ} belongs to the atom B (with $A \neq B$). Under the NDDO approximation, all the repulsion integrals $(\mu\nu|\lambda\sigma)$ involving the product of two atomic orbitals located on two different atoms are neglected as:

$$\phi_{\mu}\phi_{\lambda} = 0 \quad \text{if } \mu \in A \text{ and } \lambda \in B, \text{ with } A \neq B \quad (2.5)$$

Thus the only remaining repulsion integrals are of type: $(\mu\mu|\nu\nu)$, $(\mu\nu|\mu\nu)$ and $(\mu\nu|\lambda\sigma)$.

The elements of the Fock matrix simplify as follows:

- the diagonal elements:

$$F_{\mu\mu} = U_{\mu\mu} + \sum_B V_{\mu\mu}^B + \sum_{\nu}^A P_{\nu\nu} \left[(\mu\mu|\nu\nu) - \frac{1}{2} (\mu\nu|\nu\mu) \right] + \sum_B \sum_{\lambda,\sigma}^B P_{\lambda\sigma} (\mu\mu|\lambda\sigma) \quad (2.6)$$

where \sum_{ν}^A denotes a sum over all the atomic orbitals ν located on the atom A.

- the block-diagonal elements (*i.e.*, for two different atomic orbitals, μ and ν , located

on the same atom, A):

$$F_{\mu\nu} = \sum_B V_{\mu\nu}^B + \frac{1}{2} P_{\mu\nu} [3(\mu\nu|\mu\nu) - (\mu\mu|\nu\nu)] + \sum_B \sum_{\lambda,\sigma} P_{\lambda\sigma} (\mu\nu|\lambda\sigma) \quad (2.7)$$

- the off-diagonal elements (*i.e.*, for two different atomic orbitals located on two different atoms):

$$F_{\mu\lambda} = H_{\mu\lambda}^c - \frac{1}{2} \sum_{\nu}^A \sum_{\sigma}^B P_{\nu\sigma} (\mu\nu|\lambda\sigma) \quad (2.8)$$

In Eq. 2.6, $U_{\mu\mu}$ is the sum of the kinetic energy of an electron in an atomic orbital centered on the atom A and the electron-core attraction potential energy due to the core of atom A. In Eqs. 2.6 and 2.7, $V_{\mu\nu}^B$ represents the two-center one-electron attraction potential due to the core of a B atom on the electron represented by the distribution $\varphi_{\mu}\varphi_{\nu}$ and is given by:

$$V_{\mu\nu}^B = \left\langle \varphi_{\mu} \left| \frac{-Z_B^*}{r_B} \right| \varphi_{\nu} \right\rangle \quad (2.9)$$

with Z_B^* the core charge of B.

In the first part of this Section, we have detailed the exact NDDO formalism for the electronic energy. In what follows, we shall discuss the modifications brought by Dewar *et al.*, [114] which led to the MNDO method and to its derivatives, AM1 and PM3.

Those modifications imply the parameterization of most of the quantities present in Eqs. 2.6, 2.7 and 2.8. Their objective is to simplify and thus fasten the evaluation of most of the one- and two-electron terms. The MNDO method can be summarized as follows:

- The one-center terms, namely $U_{\mu\mu}$, $(\mu\mu|\nu\nu)$ (the Coulomb integrals, denoted $g_{\mu\nu}$) and $(\mu\nu|\mu\nu)$ (the exchange integrals, denoted $h_{\mu\nu}$), are treated as atomic parameters, which are obtained for each atom from experimental spectroscopic data using Oleari's technique.[173]
- The $V_{\mu\nu}^B$ terms are approximated to be the interaction between an electron represented by the distribution $\varphi_{\mu}\varphi_{\nu}$ centered on the atom A with an atomic core of charge Z_B^* having the size of the valence shell s orbital of the atom B (s^B):

$$V_{\mu\nu}^B = -Z_B^* (\mu\nu|s^B s^B) \quad (2.10)$$

- The off-diagonal elements of the core Hamiltonian ($H_{\mu\nu}^c$) are proportional to the overlap matrix elements $S_{\mu\nu}$ and are replaced by:

$$H_{\mu\lambda}^c = S_{\mu\lambda} \frac{\beta_{\mu} + \beta_{\lambda}}{2} \quad (2.11)$$

where β_{μ} and β_{λ} are atomic parameters optimized to reproduce experimental results

and $S_{\mu\nu}$ are evaluated analytically with the ζ exponent of the STO treated as an adjustable parameter.

- The two-electron two-center integrals $(\mu\nu|\lambda\sigma)$ are functions of ζ and of the two-electrons one-center integrals. Thus, their evaluation makes use of the $g_{\mu\nu}$, the $h_{\mu\nu}$ and the ζ parameters.[174]

In MNDO and AM1, beside the one-center terms, all the parameters are optimized to reproduce a set of experimental data for isolated molecules, such as heats of formation, ionization potentials or dipole moments. The parameterization of PM3 is similar but all the parameters are optimized, including the one-center terms.

To complete the total energy of the system, the core-core repulsion term needs to be evaluated. We shall now discuss the treatment of this term that differs from MNDO to AM1 or PM3.

2.2.2 Core-core repulsion function in MNDO, AM1 and PM3

The core-core repulsion energy is a term similar to the nuclei-nuclei repulsion energy described in Eq. 1.5. However, to remain consistent with the approximation of a valence orbital basis-set, Dewar *et al.* designed a function that represents the interaction between two cores of effective charge Z_A^* and Z_B^* ($E_{cc}^{MNDO}(A,B)$). In a similar manner as in the case of $V_{\mu\nu}^B$ in Eq. 2.10, the core repulsion function in MNDO is defined as the interaction between two cores having the size of an s valence orbital centered around A and B (s^A and s^B , respectively):

$$E_{cc}^{MNDO}(A,B) = Z_A^* Z_B^* (s^A s^A | s^B s^B) [1 + f_0(R_{AB})] \quad (2.12)$$

where $f_0(R_{AB})$ is a function introduced to avoid $E_{cc}^{MNDO}(A,B)$ to (almost) cancel out with $V_{s^A s^A}^B$ when the distance between the two cores (R_{AB}) decreases (see Eq. 2.10). $f_0(R_{AB})$ has the following expression:

$$f_0(R_{AB}) = \begin{cases} \tilde{R}_{AB} e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} & (\text{if } A,B = \text{N,H or O,H}) \\ e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} & (\text{otherwise}) \end{cases} \quad (2.13)$$

where α is an atomic parameter expressed in \AA^{-1} and $\tilde{R}_{AB} = R_{AB}[\text{\AA}]/1[\text{\AA}]$.

The formulation and parameterization of MNDO happened to yield good results for the prediction of the properties of isolated molecules but failed to reproduce weak intermolecular interactions (*e.g.*, hydrogen bonds). At least two reasons are responsible for this deficiency: i) the set of experimental data used in the parameterization procedure did not include such interactions and ii) the current formulation of the method does not allow to reproduce minima at typical intermolecular distances.

Since hydrogen bonds are fundamental interactions in biological systems, Dewar *et al.* attempted to overcome this issue by tackling the second point described above, leading to the development of AM1.[115] The idea was to add corrective terms to the core-core repulsion function, in order to correct the overestimate repulsion of intermolecular interactions predicted by MNDO. Several types of functions can be envisaged but Dewar *et al.* “decided to use a brute force approach”[115] by adding a sum of Gaussian terms in the definition of $E_{cc}(A,B)$. The core-core repulsion function in AM1 was originally defined as:[115]

$$E_{cc}^{AM1}(A,B) = Z_A^* Z_B^* (s^A s^A | s^B s^B) [1 + f_0(R_{AB}) + f_1^A(R_{AB}) + f_1^B(R_{AB})] \quad (2.14)$$

with:

$$\begin{aligned} f_1^A(R_{AB}) &= \sum_{i=1}^K a_{A,i} e^{-b_{A,i}(R_{AB}-c_{A,i})^2} \\ f_1^B(R_{AB}) &= \sum_{i=1}^K a_{B,i} e^{-b_{B,i}(R_{AB}-c_{B,i})^2} \end{aligned} \quad (2.15)$$

where a , b and c are adjustable atomic parameters.

In the development of PM3, Stewart used a similar expression for $E_{cc}^{PM3}(A,B)$:

$$E_{cc}^{PM3}(A,B) = E_{cc}^{MNDO} + g^{PM3}(A,B) \quad (2.16)$$

where

$$g^{PM3}(A,B) = \frac{Z_A^* Z_B^*}{R_{AB}} [f_1^A(R_{AB}) + f_1^B(R_{AB})] \quad (2.17)$$

Notice that further references of AM1 do not define $E_{cc}^{AM1}(A,B)$ as it is in Eq. 2.14.[175] Instead, the same definition as $E_{cc}^{PM3}(A,B)$ is used, where $g^{AM1}(A,B) = g^{PM3}(A,B)$. This definition assumes that $(s^A s^A | s^B s^B) \simeq 1/R_{AB}$. The only difference in the expression of $g^{AM1}(A,B)$ and $g^{PM3}(A,B)$ resides in the number (K) of Gaussian functions involved in $f_1^A(R_{AB})$ and $f_1^B(R_{AB})$ (Eq. 2.15). In AM1, K is set to four while it is equal to two in PM3.

AM1 and PM3 did correct the overestimated repulsion predicted by MNDO for intermolecular interactions. However, as we shall discuss in what follows, the introduction of a sum of Gaussian correction terms has no real physical meaning and can yield artifacts on the intermolecular potential energy surface. The fact that the parameterization data set used for AM1 and PM3 did not include any information about intermolecular interactions, is also a source of critical mistakes for applications in the condensed phase. We will discuss some corrections of these methods in the following Section.

2.3 Corrections of existing methods

The development of AM1 and PM3 has allowed to overcome some of the deficiencies of MNDO. However, there are still important issues that prevent the use of such methods to

be applied to condensed phase studies or biological systems. Intermolecular interactions are usually poorly described (*e.g.*, hydrogen bonds) and the electronic properties of some elements are not well predicted (*e.g.*, the electronic environment of a nitrogen atom involved in an amide group is not planar with PM3[176]).

Several methodologies have been developed to correct the misbehaviors of semiempirical methods with various degrees of success.[161–167] Corrections including both dispersion and hydrogen bond terms to treat intermolecular interactions were developed by Hobza *et al.* The first generation correction ('-DH'[163]) was developed for the PM6 method and extended to RM1 and the OMX methods, however a further reparametrization was introduced shortly after through the second generation correction ('-DH2'[164, 165]) for PM6 (see Ref. [177] for a review). These methods were used to analyze docking in protein-ligand complexes[178] by means of a linear scaling algorithm (MOZYME[179]). A different correction scheme ('-DH+'), including a damping function to improve both the short- and long-range behavior, was proposed for PM6, AM1 and OM3.[166] Finally, the PM6-D3H4X was developed by treating dispersion, hydrogen bonds and halogen bonds through a post-SCF correction.[167] Nonetheless, in a very recent paper by Hobza's group[180], in which the authors analyzed the performance of most of the corrected methods described above on a large selection of non-covalent complexes, it has been pointed out that the wrong complex structures and large errors on the interaction energies can be found. Quite a different strategy to improve NDDO semiempirical approaches was proposed in our group a few years ago and we shall briefly detail it in what follows.

2.3.1 Parameterizable interaction function: PIF

It has been shown by Csonka and Ángyán,[168] that the use of Gaussian functions in the definition of $g(A,B)$ induces artifacts on the interaction potential energy surface (IPES). We present in Figure 2.2, a representation of $g^{\text{PM3}}(A,B)$ as a function of the interatomic distance between a couple of hydrogen-oxygen atoms (H,O) and a couple of two hydrogen atoms (H,H). The two profiles show spurious minima and maxima. The $g(H,H)$ function, corresponding to the interaction between two hydrogen atoms, presents a minimum at about 1.8 Å. This minimum is enhanced in the case of $g(H,O)$. Such unphysical behavior has a dramatic impact on the IPES of fundamental intermolecular interactions, such as hydrogen bonds (see Chapters 4, 5 and 6).

An elegant way to overcome this issue has been proposed in our group,[132] through the use of a *parameterizable interaction function* (PIF). The PIF strategy has been developed originally to correct the PM3 Hamiltonian but can be adapted to other methods in a straightforward way.

The idea of the PIF correction is to overcome the spurious artifacts introduced by the

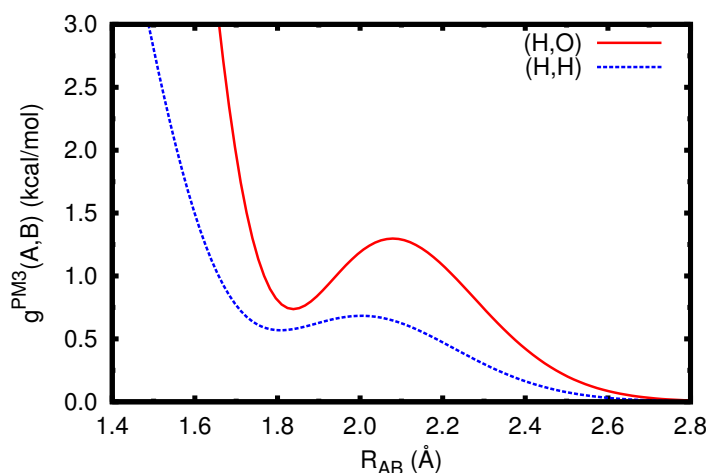


Figure 2.2: Representation of $g^{\text{PM3}}(\text{A,B})$ as a function of the interatomic distance for two couples of atoms: hydrogen-oxygen (H,O) and hydrogen hydrogen (H,H).

Gaussian functions, by using a different $g(\text{A,B})$ function for a pair of atoms belonging to different molecules, according to the following expression:

$$g^{\text{PIF}}(\text{A,B}) = \begin{cases} g^{\text{PM3}}(\text{A,B}) & \text{if A,B intra} \\ \alpha_{\text{AB}} e^{-\beta_{\text{AB}} R_{\text{AB}}} + \frac{\chi_{\text{AB}}}{R_{\text{AB}}^6} + \frac{\delta_{\text{AB}}}{R_{\text{AB}}^8} + \frac{\epsilon_{\text{AB}}}{R_{\text{AB}}^{10}} & \text{if A,B inter} \end{cases} \quad (2.18)$$

where α , β , χ , δ , and ϵ are diatomic parameters optimized to reproduce the interaction potential energy surface of a set of dimers obtained using high level *ab initio* calculations. g^{PIF} intends to bring a more physical meaning to the intermolecular interactions by introducing an exponential term (which corresponds to the expected behavior for the Pauli repulsion at short distance) and a series of terms inversely proportional to the sixth, eighth, and tenth powers of the interatomic distance (which are expected to correct for inaccuracies of the electronic contribution in electrostatic, induction, and dispersion energies).

The first version of the PIF method focused only on water-water interactions. We shall refer to this set of parameters as PM3-PIF1. Subsequently, the set of PM3-PIF1 parameters was extended to molecules containing H, C, N, O and Cl atoms interacting with water.[138, 139, 181] This parametrization will be named PM3-PIF2 in the following discussion.

The PIF2 correction uses a different set of parameters to describe the interaction between water molecules or between a water molecules and an organic solute. The strength of the PIF strategy is both based on the physically meaningful formulation of g^{PIF} and on the parameterization set of data used in its development. Instead of considering only minimum structures, as it is the case in most of the SE methods, intermolecular potential energy surfaces were introduced for a set of chosen complexes involving water: *i.e.*, water-water, water-amonia, water-methanal, water-methanol, water-formic acid, water-formonitril and

water-methane. A total number of 1485 configurations were used in the parameterization procedure.

One of the purpose that motivated the development of the PM3-PIF1 and the PM3-PIF2 SE Hamiltonians was to study the dynamics of biological compounds in water. Those two Hamiltonians have been shown to yield good results when applied to simulate the dynamics of pure liquid water and of the N-methylacetamide molecule in water,[74, 169] respectively. We shall give other examples of the application of the PIF methodology in Chapters 3, 4, 5 and 6.

2.3.2 Method adapted for intermolecular studies: MAIS

The strategy used in the PIF correction (*i.e.*, replacing only the core-core repulsion function in a given semiempirical Hamiltonian) was shown to yield a better description of intermolecular interactions. However, the resulting PM3-PIF methods is by definition not applicable to intermolecular reactivity studies. In a similar manner, a second correction was developed in our group, in collaboration with Prof. Margarita I. Bernal-Uruchurtu from the University of Morelos in Cuernavaca, Mexico: the *method adapted for intermolecular studies* (MAIS).[137]

The MAIS correction was developed originally for PM3. However, as it is the case for PIF, a straightforward adaptation of this correction can be performed for other methods. Unlike PIF, the $g(A,B)$ is systematically replaced in MAIS by:

$$g^{\text{MAIS}}(A,B) = \sum_{i=1}^3 \alpha_{iAB} e^{-\beta_{iAB}(\gamma_{iAB}-R_{AB})^2} \quad (2.19)$$

where α , β and γ are diatomic parameters. This definition of the $g(A,B)$ is continuous and can thus be used to model intermolecular reactions. $g^{\text{MAIS}}(A,B)$ involves a sum of three Gaussian functions. This could seem surprising considering the above discussion about AM1 and PM3. However, because the parameterization procedure include IPES and not only stationary structures, the artifacts caused by the Gaussian terms are avoided.

The first set of parameters was dedicated to water-water interactions. We shall refer to this set as PM3-MAIS1 in what follows. Subsequently, a second set version was developed to treat the HCl dissociation in water clusters, where the whole set of interactions parameters was reoptimized, leading to the PM3-MAIS2 Hamiltonian.[139] Both PM3-MAIS1 and PM3-MAIS2 have been shown to yield good results to study the stability of protonated water complexes[182] as well as the dynamics of pure liquid water.[169] In the present work, only PM3-MAIS1 has been used. Thus, if not otherwise stated, PM3-MAIS will always refer to this set of parameters.

2.4 Concluding remarks

We have presented in this Chapter an overview of semiempirical (SE) methods. Among the numerous available Hamiltonians, the present work mainly focuses on those derived from MNDO. We detailed this Hamiltonian as well as some of its derivative approaches.

Despite the great efforts that have been made during the past decades to improve the description of intermolecular interactions using SE methods, some deficiencies still prevent the use of such approaches to study molecular systems in aqueous environments. In particular, most of the existing methods show artifacts on the potential energy surface, due to the presence of Gaussian correction functions in the core-core repulsion term of the total energy. Two corrections of the PM3 Hamiltonian have been developed in our group: PM3-PIFx and PM3-MAISx. These approaches replace the Gaussian correction functions by a pairwise physically meaningful function. The parameters of the PIF and MAIS functions have been optimized to reproduce MP2/aug-cc-pVTZ interaction energies of 1:1 complexes, by considering representative configurations of the related potential energy surface (PES). This strategy yields a better description of the PES, which is mandatory to perform molecular dynamics simulations.

Chapter

3

**SemiEmpirical Born-Oppenheimer
Molecular Dynamics (SEBOMD):
description and related developments**

Résumé

Une nouvelle méthode de dynamique moléculaire, décrivant la totalité d'un système donné au niveau semiempirique, a récemment été développée au sein de notre groupe. La méthode SEBOMD (*SemiEmpirical Born-Oppenheimer Molecular Dynamics*) a été implémentée et est officiellement distribuée dans la dernière version du logiciel de dynamique moléculaire Amber (version 14). Nous parcourons brièvement ici les caractéristiques d'Amber14 relatives au développement de SEBOMD et décrivons l'implémentation de la méthode ainsi que ces principales options. Une partie du projet présenté ici a été d'aider l'amélioration de cette implémentation et nous en discutons ici les points les plus importants. Étant donné le caractère récent de la méthode SEBOMD, les outils nécessaires à l'analyse des calculs effectués par cette approche ont dû être développés également. Nous en présentons ici quelques uns. Enfin, nous traitons un système test, l'eau liquide, pour discuter du choix de paramètres tels que celui de l'Hamiltonien décrivant le système, et celui du modèle de charge à utiliser pour prendre en compte les interactions à longue portée.

Amber est une suite de programmes destinés à l'étude des propriétés dynamiques des systèmes d'intérêt biologique. Elle contient divers outils, contenant notamment le module `sander`, un moteur de dynamique moléculaire. Le module `sander` est distribué gratuitement sous une licence GPL et c'est dans ce programme que nous avons implémenté la méthode SEBOMD.

La méthode SEBOMD a été développée dans Amber14 en couplant `sander` avec un programme de calcul quantique semiempirique, DivCon99. Cependant, considérant le nombre important de modifications ayant été apportées à DivCon99, le nouveau module permettant d'effectuer les calculs d'énergie au niveau semiempirique à été renommé `sebomd`. L'implémentation de `sebomd` suit la philosophie des précédents développements effectués dans `sander`. Le programme maître reste `sander`, qui conduit la dynamique moléculaire en appelant `sebomd` pour le calcul de l'énergie et des forces à chaque pas de temps de la dynamique. `sebomd` tire avantage de la plupart des options de calculs présentes originellement dans `sander` telles que les différents thermostats, les outils de calcul d'énergie libre ou encore les algorithmes d'exploration efficace de l'espace des phases.

Afin d'aider l'implémentation de la méthode SEBOMD dans Amber14, nous avons tout d'abord travaillé à l'amélioration de la communication entre sander et sebomd. Un module est maintenant présent dans sander afin d'assurer le passage des données nécessaires aux calculs dans sebomd. Au cours de nos différents travaux traitant des corrections apportées aux méthodes semiempiriques, il nous a été nécessaire de ré-écrire la partie de sebomd relative à ces corrections. Le nouveau module développé à cet effet a été écrit de manière à rendre facile l'implémentation de nouvelles corrections dans l'avenir.

Divers outils ont été développés afin de permettre l'analyse des résultats issus de simulations SEBOMD. Notamment, nous avons écrit un programme permettant l'analyse des modes normaux de vibration du système simulé, dans l'approximation harmonique. Nous avons également ajouté à ce programme un outil permettant de produire une série de trajectoires traduisant les mouvements du système selon toutes les coordonnées normales de ce dernier. Un module compatible avec le programme VMD a également été écrit pour faciliter la visualisation de ces données. Nous présentons également ici une manière pratique de définir les fonctions d'atténuation pour des applications à la métadynamique. La méthode imaginée ici décrit ce type de fonctions par une spline cubique, facilitant ainsi le test de différents paramètres par l'utilisateur, pour la mise en œuvre d'une étude par métadynamique.

Enfin, nous montrons par l'intermédiaire de simulations tests d'eau liquide, que le résultat d'un calcul SEBOMD dépend fortement du choix de différents paramètres. Notamment, le choix de l'Hamiltonien semiempirique s'avère être un point critique. Nous montrons que parmi les Hamiltoniens implémentés dans la version courante de sebomd, seuls PM3-PIF2 et PM3-MAIS permettent une description satisfaisante de la structure de l'eau liquide. Dans une moindre mesure, l'Hamiltonien AM1/d-PhoT apparaît comme un candidat raisonnable pour réaliser ce type d'étude. Nous avons également montré ici un problème quant à l'utilisation des méthodes d'Ewald pour la prise en compte des interactions à longue portée. Ce type d'approche décrit de manière classique les interactions à l'extérieur de la boîte centrale et le modèle de charges ponctuelles utilisé à cet effet doit être capable de reproduire les propriétés électrostatique de la boîte principale. Nous montrons ici que si le modèle de charges surestime le moment dipolaire comparé à celui issu de la fonction d'onde quantique, alors un champ électrique externe est créé, modifiant de manière significative la structure de l'eau. Nous proposons ici un moyen efficace de palier à ce problème, en définissant un nouveau modèle de charge cohérent avec la fonction d'onde, dérivé des charges de Mulliken.

Modeling biomolecular systems by explicitly taking into account the quantum mechanical behavior of the electrons represents one of the greatest challenges for theoretical chemistry studies. Moreover, the complexity of such studies increases when the number of degrees of freedom is large and when meaningful statistics are necessary to model the phenomenon of interest. Although outstanding progress has been made in the past decades in performing molecular dynamics (MD) simulations with density functional theory based methods (*e.g.*, Car-Parrinello molecular dynamics, CPMD[183]) to include the quantum nature of the electrons, long time scales and/or systems containing a large number of atoms still demand very high computational costs. A reasonable compromise is represented by using a more approximate level of quantum chemistry to model the electronic Hamiltonian. In particular, NDDO(Neglect of Diatomic Differential Overlap)-based semiempirical (SE) methods are particularly appealing. Such approaches demand a low computational cost and can be improved through the optimization of the parameters that they contain. Moreover, they are particularly well adapted to efficient parallelization schemes allowing the use of linear scaling algorithms.

A molecular dynamics technique using SE methods as Hamiltonians has been recently developed in our group: SEBOMD, standing for SemiEmpirical Born-Oppenheimer Molecular Dynamics.[169] We shall detail in what follows the general features of SEBOMD as well as its recent implementation in Amber14.[184] Part of the present work was devoted to test and to improve this implementation, and we will detail the modifications that have been done. SEBOMD is a new MD method and we had to develop the related analysis tools. We will also detail here a code developed to compute and visualize the normal modes of a given molecular system from a SEBOMD calculation. In addition, we tested the implementation of a new way of defining smoothing functions in Amber14 for applications to metadynamics simulations. Finally, as a test case, we shall discuss the ability of the SEBOMD methodology to reproduce the structure of liquid water with respect to the choice of the SE Hamiltonian and of the method used to take into account long range interactions.

3.1 Amber14 and AmberTools14

Before moving on, a few comments about the Amber14[184] suite of programs are necessary. The previous version of this molecular simulation program package (Amber12) has been recently reviewed by Salomon-Ferrer *et al.*[185] We shall briefly detail some relevant aspects of Amber14 with respect to the implementation of SEBOMD.

3.1.1 General comments

Amber is a molecular modeling suite of programs. It is distributed in two parts: Amber and AmberTools. Amber is a license restricted program while AmberTools is distributed under

a GNU General Public License (GPL). Two main programs are part of those packages and are intended to carry efficient molecular dynamics (MD) simulations: `sander` and `pmemd`. The `sander` program is capable of performing MD simulations in different thermodynamics ensembles, with or without periodic boundary conditions and allows the application of restraints on various types of geometrical coordinates. The MD simulations in `sander` can be carried out using different types of molecular mechanics (MM) force fields as well as hybrid quantum mechanical/MM (QM/MM) approaches. The `pmemd` program is an improved version of `sander` that contains the most used and tested features of the latter. Its implementation is optimized to perform faster than `sander` and to scale more efficiently for parallel MM calculations. It also includes an NVIDIA GPU acceleration version of the code (`pmemd.cuda`), allowing highly efficient MM-MD simulations of very large molecular systems.

In the previous releases of Amber and AmberTools, both `sander` and `pmemd` were distributed within Amber. From the last version, `pmemd` is included in the license restricted Amber14 while `sander` is included in the GNU AmberTools14. AmberTools14 also includes a variety of independent programs that allow to perform a complete molecular modeling study:

- Setting of *ad hoc* force field parameters (`antechamber`).
- Preparation of the inputs (`leap` and `tLeap`).
- MD simulations (`sander`).
- Analysis of the simulation(s) (`ptraj` or `cpptraj`).

In the present work, we did not use any of the features of Amber14. All the calculations, analysis, input preparations and developments have been performed using the AmberTools14 program package only. However, for the sake of simplicity, we might refer to the former as Amber14 or simply Amber in what follows.

Finally, we mention that the acronym Amber is also used to name a collection of molecular mechanics force fields such as ff99SB,[186] ff03[38] or lipid14.[187]

3.1.2 The `sander` program

The `sander` program of AmberTools14 is dedicated to perform MD simulations of biomolecular systems. At each time step of the simulation, `sander` calls a subprogram to compute the energy and forces of the system in order to predict the next geometry by integrating the Newton's second law of motion (see Section 1.3). The simulations can be performed with or without periodic boundary conditions, in different thermodynamical ensembles and using various types of Hamiltonians:

- MM force fields.
- QM/MM schemes using the `sqm` module of AmberTools14 (for DFTB or semiempirical Hamiltonians) as well as external QM programs (*e.g.*, Gaussian09[188]).

- Full SE QM approaches using the SEBOMD methodology (see below).

In such an implementation, `sander` is considered as a *master*, which controls and calls the relevant *slaves* to perform the required energy calculations. For more details about the features of `sander`, one can refer to the Amber14 user manual and to the numerous tutorials available on the Amber Web site.[189] We shall now detail the basic idea of the implementation of SEBOMD in `sander`.

3.2 SEBOMD in Amber14

To perform a molecular dynamics (MD) simulation in Amber14, at least two program modules are required: one that integrates the Newton's second law of motion as well as one that computes the energy and forces of an instantaneous geometry of the system (see Section 1.3). SEBOMD was developed by coupling one of the MD driver programs of Amber14, *i.e.*, `sander`, with a semiempirical (SE) quantum chemistry package, *DivCon99*, [190] for the energy calculation. Many changes have been made from the original version of *DivCon99* in SEBOMD. We shall therefore refer to this program as a new module simply called `sebond`. From now on, SEBOMD will refer to the methodology while `sebond` is the related program module.

3.2.1 Implementation

The implementation of `sebond` within `sander` follows the standard development protocol of the Amber developers community. The `sander` program is kept as the *master*, which manages the input/output and directs the calculation by calling the `sebond` module. This choice allows a general implementation of SEBOMD in Amber14, which is now distributed within the official version of the program package. Moreover, the new `sebond` module can take advantage of nearly all the available features of `sander`, such as:

- Molecular dynamics
- Energy minimization
- Simulations in different thermodynamical ensembles:
 - NVE
 - NVT
 - NPT (only the Monte-Carlo barostat has been tested up to now)
- Replica exchange molecular dynamics
- Biased molecular dynamics for free energy calculations:
 - Umbrella sampling
 - Metadynamics

- Path integrals molecular dynamics

However, specific tests must be performed prior to any new application of the SEBOMD methodology with an available feature of `sander`, in order to ensure the compatibility between the different techniques.

3.2.2 SEBOMD features

Beside the general features brought by `sander`, `sebond` has some specific characteristics. Various semiempirical Hamiltonians are available as well as different SCF algorithms. Condensed phase calculations are available using periodic boundary conditions and Ewald summation schemes to account or not for long range interactions. A parallel version of the code is also implemented in `sander .MPI` through the use of a linear scaling algorithm.

Self-consistent field

`Sebond` performs energy and forces calculations at each time step of the dynamics. The MNDO, AM1, PM3, PDDG/PM3, RM1 and AM1/d-PhoT Hamiltonians are available. The accessible elements are limited to H, C, N, O, S, F, Cl, Br, I and P for all these Hamiltonians. However, because the current version of `sebond` does not include d-orbitals, phosphorus parameters are not available for AM1/d-PhoT.

The convergence criterion for an SCF calculation is determined by the relative energy obtained between two successive iterations. Although the threshold can be set by the user, it has been shown favorable to set it to 10^4 times the machine precision in order to ensure the conservation of the energy during an NVE simulation.

Two different routines are available for the diagonalization of the Fock matrix. A standard full diagonalization and a pseudo-diagonalization scheme. The latter significantly improves the calculation speed.

In order to facilitate the SCF convergence, two options are available in `sebond` to generate an initial (guess) density matrix. One can simply use the density matrix of the previous time step of the dynamics or use a polynomial interpolation of the three last steps density matrices. During an MD simulation, the latter significantly reduces the number of SCF iterations and thus accelerates the overall calculation.

Condensed phase simulations

Residue based periodic conditions are implemented in `sebond`. This allows to simulate condensed phase systems by keeping a constant number of particles/molecules in the unit cell. As it is the case for classical molecular mechanics molecular dynamics simulations, the use of residue based PBC has almost no effect on the conservation of the total energy. However, when a molecule crosses the unit cell frontier, a discontinuity is introduced in the forces

calculations. Basically, this discontinuity arises from the change of orientation of the forces between the particle in the center of the box and those located at the edges. For a large enough box size, this effect can be neglected. However, as the unit cell size is limited by the available computational power, this should be taken into account. In the current version of `sebomd`, no strategy is implemented to tackle this issue. Therefore, it is preferable to perform simulations in the NVT ensemble when PBCs are used. Different schemes are currently envisaged to lower the impact of this phenomenon, in a similar way as it is done in most of the MM-MD simulation programs. Most of those approaches are based on the definition of a buffer region at the edges of the unit cell, in which the position vector of the particle of interest is scaled by an attenuating factor, defined by a smoothing function.

Long range Coulombic interactions can be taken into account or not in `sebomd` when using PBC. Two Ewald schemes are implemented: a standard Ewald summation scheme and an approximate particle mesh Ewald (PME) approach. The former is explicitly included in the SCF procedure, while the latter consists of an external classical correction that uses a fixed set of atomic partial charges (usually extracted from a force field). In the case of the standard Ewald method, two atomic partial charge models are currently available: Mulliken and CM1 (the CM1 model is only available for AM1 and PM3). Nevertheless, as we shall discuss in Section 3.5, our test calculations have shown that it is preferable to use the Mulliken charge model in the present implementation.

Linear scaling and parallel calculations

Semiempirical QM approaches are particularly well adapted for linear scaling approaches. In `sebomd`, the Divide&Conquer algorithm is available to fasten the Fock matrix diagonalization procedure. Such an approach is based on the splitting of the Fock matrix into small regions. Each region is diagonalized independently and the total diagonal matrix is rebuild at the end of the procedure. The power of such an approach is that, though the number of diagonalization procedures is increased, each diagonalization can be treated on a single core. A parallel version of `sebomd` is implemented in `sander`. For this version, only the SCF calculation using the Divide&Conquer algorithm is available.

Available properties

Various quantities are computed and outputted at each time step of a SEBOMD simulations:

- Mulliken atomic partial charges as well as CM1 and CM2 when available.
- Total dipole moment (extracted from the wave function).
- Partial numerical Hessian matrix and gradient vector.
- Atomic Cartesian positions and velocities (as provided by `sander`).

The Hessian matrix of a molecular system containing N atoms is a $3N \times 3N$ array. For large systems, the size of this array can rapidly produce large output files. Thus, the maximum number of Hessian matrix elements that can be outputted is predefined and hard coded in the memory allocation routine of `sander`.

3.3 Improvements and developments of `sebomd` in `sander`

Along the present work, the different tests that have been carried out on the SEBOMD methodology and on its implementation in `sander` have helped to improve the method as well as to identify and to fix minor bugs. We shall not detail all of the modifications that have been performed but only the most relevant. Mainly, the communication between `sebomd` and `sander` has been improved and simplified, in order to facilitate further developments and a major modification of the way that semiempirical corrections are handled has been performed.

3.3.1 Setup and communication with `sander`

The original version of `sebomd` was written in Fortran77 while the last version of `sander` in AmberTools14 is in Fortran90. To facilitate the further developments of `sebomd`, we first improved the communication between `sebomd` and `sander`.

The early implementation of `sebomd` in `sander` used a DivCon99 formatted input file generated by `sander` to pass all the required information to `sebomd` (e.g., the Hamiltonian, the use of PBC, the residues and atomic numbers, *etc.*). From now on, all the semiempirical calculation options are kept in `sander` and a Fortran90 module is used to bridge the two programs. This strategy allows an easier modification of the information to be passed from one program to another.

All the tables and matrices required for the setup of the SE calculations are also stored in `sander`. A `sebomd_arrays.f90` module was created for this purpose, allowing a simple *put/get* interface between `sander` and `sebomd`. Such an implementation allows to modify the way in which data are stored or calculated without affecting the rest of the code. In addition, the use of arrays of variable size is facilitated by this mean, since the master can manage the allocation and the cleaning of the required memory parts.

3.3.2 Rewriting of the routine for corrections to semiempirical methods

In the original version of `sebomd`, the implementation of the semiempirical corrections was part of the standard SE calculation. The limitation of such an implementation arises from the numerous tests that have to be made to classify each interaction in the system. Also, the

modification and implementation of new methodology was made inefficient and we had to rationalize this part of the code.

We wrote a new Fortran90 module dedicated to apply modification of the standard SE methods. This procedure is independent of the rest of the calculation which is now performed as follows:

1. SCF calculation for the electronic part of the Hamiltonian.
2. Calculation of the core-core term of the energy for all atom-atom interactions.

And when needed:

- 3a. Removal of the original interatomic interactions intended to be modified.
- 3b. Addition of the modified interaction with respect to the selected scheme (*e.g.*, PIF3, MAIS1).
4. Application of the peptidic correction, implemented in a similar manner as described by Ludwig *et al.*[176]

The corrections that have been implemented in the new `se_correction` modules up to now are for PM3: PIF2, PIF3, MAIS1 and MAIS2. For all available Hamiltonians, we also implemented the peptidic correction. However, the former requires the setting of a force constant and those are only available for MNDO, AM1 and PM3.[176] Nevertheless, the user can set its own definition of this force constant in the `sander` input file. Moreover, the `se_correction` modules have been developed in a way that any type of SE correction can be added in a straightforward way (*e.g.*, -DHx corrections).

Another reason that has motivated the rewriting of the semiempirical corrections in the `sebond` module was the implementation of the PIF3 correction of PM3 developed in the present work.[140] As we will see in Chapter 4 (where we detail the development and tests of PM3-PIF3), this correction treats differently the interaction between water and a “hydrophobic” or a “hydrophilic” hydrogen atom. The implementation of such a method does not only request the atomic number of each atom (as it is the case in most of the quantum chemistry programs) but also requires the definition of the atom type of each hydrogen atom present in the molecular system. The `sander` program was originally developed to perform MM calculations and takes as an input a topology file containing the relevant information about the force field chosen by the user. Beside other data, such a topology file attributes a type to all the atoms of the system using a nomenclature specific to the Amber force field. For example, an aliphatic sp^3 carbon atom is of type CT and a hydrogen atom bonded to the latter is of type HC while the hydrogen atom of a hydroxyl group is referenced as HO. The knowledge allowing to differentiate between hydrophobic and hydrophilic hydrogen atoms when applying the PM3-PIF3 method is thus already present in `sander` and we used the bridge defined previously to send this information to `sebond`.

In the `se_correction` modules, we assigned a hydrophobic or hydrophilic character

to all the available Amber hydrogen atom types. At the first step of a molecular dynamics simulation or energy minimization, each interatomic interaction is identified, classified and stored in the `sebomd_array.f90` module of `sander`. This procedure has to be performed only once, during the calculation setup, and the application of the chosen interaction specific semiempirical correction is efficiently handled.

3.4 Other related developments around `sander` and `sebomd`

SEBOMD is a new molecular modeling method. Therefore, most of the analysis tools that are necessary to perform the different studies detailed in the following Chapters had to be built and this has been an important part of the present work. We shall not detail all of them but we will focus on the development of a harmonic vibrational normal mode calculation and visualization tool. In addition, we have developed an elegant way of defining the smoothing function for applications to metadynamics free energy calculations in `sander`.

3.4.1 Vibrational normal mode analysis: calculation and visualization

The `sebomd` module is compatible with the energy minimization routines of `sander`. To study stationary molecular structures, it is often necessary to characterize the obtained geometry (*e.g.*, minimum, transition state, *etc.*). This analysis can be done by computing the harmonic vibrational frequencies associated with each normal coordinate of the system. Such an analysis tool is not available in the current version of AmberTools14. We present here the methodology that we used to build a program able to process the outputs of `sebomd` in order to perform this analysis. We first discuss the computation of the vibrational frequencies and then the visualization of each harmonic vibrational normal modes.

Frequencies calculation

Under the harmonic approximation, the vibrational frequencies of a given molecular system composed by N atoms are obtained from the diagonalization of the mass weighted Hessian matrix (\mathbf{U}^m). The $3N \times 3N$ elements of \mathbf{U}^m (u_{ij}^m) are computed from the elements of the Cartesian Hessian matrix (u_{ij}) as:

$$u_{ij}^m = \frac{u_{ij}}{\sqrt{M_i M_j}} \quad (3.1)$$

where M_i is the mass of the atom related to the coordinate i . The diagonalization of \mathbf{U}^m leads to the matrix of orthogonal eigenvectors (\mathbf{L}) and to the diagonal matrix of eigenvalues ($\mathbf{\Lambda}$) as:

$$\mathbf{U}^m \mathbf{L} = \mathbf{\Lambda} \mathbf{L} \quad (3.2)$$

A vibrational frequency (ν_i) is assigned to each eigenvalue (λ_i) of $\mathbf{\Lambda}$ as:

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi} \quad (3.3)$$

We note that, for a geometry that does not correspond to a minimum of the system potential energy surface, some elements of $\mathbf{\Lambda}$ can have a negative value. To avoid computing imaginary frequencies, it is common to use the absolute value of λ_i in Eq. 3.3 and to attribute *a posteriori* the sign of λ_i to ν_i .

The Cartesian Hessian matrix can be outputted from a SEBOMD calculation by setting the `ntwh` option to 1 in the input of `sander`. This option is still in the beta stage and thus it does not appear in the current version of the AmberTools14 documentation. We wrote a Fortran90 program (`ENMcalc`, which is accessible in Supplementary Material) to process the SEBOMD Hessian matrix and compute the vibrational frequencies of a given molecular system.

It is sometimes useful to visualize the vibrational normal mode associated to each frequency. We added this functionality to our code and we detail in the following the methodology that we used to this end.

Computation and visualization of the normal modes

To analyze the vibrational properties of a given minimized molecular system composed by N atoms of Cartesian coordinates $(x_1^0, x_2^0, \dots, x_{3N}^0)$, one first defines the $3N$ mass weighted Cartesian displacement coordinates (a_i) as:

$$a_i = \sqrt{M_i} (x_i - x_i^0) \quad i = 1, \dots, 3N \quad (3.4)$$

where $(x_1, x_2, \dots, x_{3N})$ defines a given geometry of the system. From this, the $3N$ vibrational normal displacement coordinates (Q_j), are defined as a linear combination of the mass weighted Cartesian displacement coordinates by:

$$Q_j = \sum_{i=1}^{3N} l_{ij} a_i \quad j = 1, \dots, 3N \quad (3.5)$$

where l_{ij} are the elements of the eigenvectors matrix \mathbf{L} obtained from Eq. 3.2. Since the l_{ij} eigenvectors are orthogonal, the backward transformation is obtained from Eq. 3.5 by:

$$a_i = \sum_{j=1}^{3N} l_{ij} Q_j \quad i = 1, \dots, 3N \quad (3.6)$$

In the following, the i and j indexes will always refer to the Cartesian and to the normal coordinate system, respectively. If the geometry of interest is the same as the geometry for which the Hessian matrix has been computed, all the a_i are zeroed and so are the Q_j .

The vibrational potential energy of the system (V_T^{vib}) can be expressed as the sum of the potential energies (V_j^{vib}) associated with each of the $3N$ normal coordinates as:

$$V_T^{\text{vib}} = \sum_{j=1}^{3N} V_j^{\text{vib}} \quad (3.7)$$

with:

$$V_j^{\text{vib}} = \frac{1}{2} \lambda_j Q_j^2 \quad (3.8)$$

and conversely:

$$Q_j = \pm \sqrt{\frac{2V_j^{\text{vib}}}{\lambda_j}} \quad (3.9)$$

From Eq. 3.9, we can see that an increase of V_j^{vib} will induce a shift of the corresponding normal coordinate and *vice versa*.

In our `ENMcalc` code, we add a functionality that allows to analyze and output the Cartesian motion of the system atoms with respect to a displacement along a given normal coordinate (j). For each normal coordinate, `ENMcalc` outputs a short looped trajectory (in the `sander .crd` format), with a normal displacement step ΔQ_j that can be visualized with any compatible program.

The displacement step ΔQ_j is defined in normal coordinates as:

$$\Delta Q_j = \frac{|Q_{j,\text{max}}|}{k_{\text{max}}} \quad (3.10)$$

where $Q_{j,\text{max}}$ is the maximum normal displacement and k_{max} is the maximum number of steps for the trajectory to output. The latter is a constant given by the user, while the former is obtained from Eq. 3.9 as:

$$Q_{j,\text{max}} = \pm \sqrt{\frac{2V_{i,\text{max}}^{\text{vib}}}{\lambda_j}} \quad (3.11)$$

where $V_{i,\text{max}}^{\text{vib}}$ defines a maximum vibrational potential energy associated with the displacement along the normal coordinate of interest. In order to keep the characteristic amplitude of each vibrational motion, it is convenient to express $V_{i,\text{max}}^{\text{vib}}$ as:

$$V_{i,\text{max}}^{\text{vib}} = n_{\text{max}} h \nu_i \quad (3.12)$$

where h is the Planck's constant, ν_i is the vibrational frequency computed from Eq. 3.3 and n_{max} is the maximum level of excitation that one wants to consider under the harmonic

approximation. Fixing the value n_{\max} to be the same for each normal vibration allows to quantitatively compare the amplitude of each motion with respect to the value of its associated vibrational frequency. For example, the stretching motion of a CH bond will have a smaller amplitude than the motion of a protein backbone, since the former is associated with a higher vibrational frequency than the latter.

The normal displacement ($Q_{j,k}$) at each step (k) of the trajectory is obtained by:

$$Q_{j,k} = k\Delta Q_j \quad k = -k_{\max}, \dots, 0, \dots, k_{\max} \quad (3.13)$$

From Eq. 3.6, we obtain the Cartesian displacement at each step of the trajectory as:

$$a_{i,k} = \sum_{j=1}^{3N} l_{ij} Q_{j,k} \quad \begin{array}{l} i = 1, \dots, 3N \\ k = -k_{\max}, \dots, 0, \dots, k_{\max} \end{array} \quad (3.14)$$

If we apply a displacement along a single normal coordinate j , all the $Q_{l \neq j} = 0$. Thus, Eq. 3.14 simplifies in:

$$a_{i,k} = l_{ij} Q_{j,k} \quad \begin{array}{l} i = 1, \dots, 3N \\ k = -k_{\max}, \dots, 0, \dots, k_{\max} \end{array} \quad (3.15)$$

We obtain the set of Cartesian coordinates ($x_{i,k}$) at each step k by:

$$x_{i,k} = \frac{a_{i,k}}{\sqrt{M_i}} + x_i^0 \quad \begin{array}{l} i = 1, \dots, 3N \\ k = -k_{\max}, \dots, 0, \dots, k_{\max} \end{array} \quad (3.16)$$

Finally, by combining Eqs. 3.10, 3.11, 3.12, 3.13 and 3.15, Eq. 3.16 becomes:

$$x_{i,k} = \frac{l_{ij}}{k_{\max}} \sqrt{\frac{2n_{\max} v_j}{\lambda_j M_i}} + x_i^0 \quad \begin{array}{l} i = 1, \dots, 3N \\ k = -k_{\max}, \dots, 0, \dots, k_{\max} \end{array} \quad (3.17)$$

Using such a definition, the user only has to define two parameters, *i.e.*, the maximum level of excitation (n_{\max}) and the maximum number of points to output in the trajectory (k_{\max}).

To help the visualization of the trajectories resulting from this analysis, we wrote a short VMD plugin. This script is written in Tcl and is available as Supplementary Material.

3.4.2 Spline definition of smoothing functions for metadynamics simulations

We introduced in Section 1.4.2 the metadynamics simulation method to assess the free energy related to a given process. Such a technique requires the definition of one or more collective variable(s) that will be sampled along the dynamics and that will reflect the process of interest.

Many definitions of a collective variable can be adopted and one of those is the coordination number of an atom with respect to a subset of the N other atoms of the system (N_{sub}). The coordination number of an atom i (η_i) with respect to all the N_{sub} atoms is usually given by:

$$\eta_i = \sum_j^{N_{sub}} S(R_{ij}) \quad (3.18)$$

where R_{ij} is the interatomic distance and S is a *smoothing* function that continuously goes from one to zero as R_{ij} increases. This smoothing function is used to determine if a bond exists or not between two atoms considering their interatomic distance. At each step of the metadynamics simulation, η_i and its first derivative with respect to the atomic positions need to be evaluated in order to apply the corresponding biasing force to the system. As a consequence, the value and the first derivative of the smoothing function need to be known for any value of R_{ij} .

Mainly, two families of smoothing functions exist: *shifting* and *switching* functions.[191] A shifting function is equal to one only for $R_{ij} = 0$ Å and either is equal to 0 for a given value of R_{ij} or tends to 0 when R_{ij} tends to infinity. An example of shifting function can be a piecewise-defined function as:[191]

$$S_a(R_{ij}) = \begin{cases} 1 - \frac{2R_{ij}^2}{R_{cut}^2} + \frac{R_{ij}^4}{R_{cut}^4} & R_{ij} \leq R_{cut} \\ 0 & R_{ij} > R_{cut} \end{cases} \quad (3.19)$$

where R_{cut} is a parameter that defines the region in which S_a smoothly goes from 1 to 0. Another example is the function implemented in the sander program of AmberTool14:[184]

$$S_b(R_{ij}) = \frac{1 - \left(\frac{R_{ij}}{R_0}\right)^6}{1 - \left(\frac{R_{ij}}{R_0}\right)^{12}} \quad (3.20)$$

where R_0 defines the point at which $S_b = 0.5$.

Switching functions have a definition similar to that of shifting functions but contain a parameter that allows to define the region in which the smoothing function is applied. A switching function is equal to 1 up to a given value of R_{ij} and smoothly goes or asymptotically

cally tends to 0 when R_{ij} increases. Such a function can be a piecewise defined polynomial as:[191]

$$S_c(R_{ij}) = \begin{cases} 1 & R_{ij} \leq R_{on} \\ \frac{(R_{off}^2 - R_{ij}^2)^2 (R_{off}^2 + 2R_{ij}^2 - 3R_{on}^2)}{(R_{off}^2 - R_{on}^2)^3} & R_{on} < R_{ij} \leq R_{off} \\ 0 & R_{ij} > R_{off} \end{cases} \quad (3.21)$$

were R_{on} and R_{off} define the range of application of S_c . A switching version of S_b (Eq. 3.20) is implemented in the PLUMED program as:[65]

$$S_d(R_{ij}) = \begin{cases} 1 & R_{ij} \leq d_0 \\ \frac{1 - \left(\frac{R_{ij} - d_0}{\rho_0}\right)^n}{1 - \left(\frac{R_{ij} - d_0}{\rho_0}\right)^m} & R_{ij} > d_0 \end{cases} \quad (3.22)$$

This definition is more flexible than the one of S_b . The parameters n and m allow to tune the shape of the function, d_0 defines the range of applicability of S_d and the value of $d_0 + \rho_0$ sets the position at which the function equals a half.

A different type of smoothing function has been proposed by Sprik for a metadynamics study of the water self-dissociation process in liquid water.[192] This smoothing function is given by:

$$S_e(R_{ij}) = \left[e^{\kappa(R_{ij} - R_c)} + 1 \right]^{-1} \quad (3.23)$$

where R_c defines the point at which $S_e = 0.5$ and κ controls the slope of the function. This function is different from the other examples given above (Eqs. 3.19, 3.20, 3.21 and 3.22), because here S_e never equals 1 nor 0 for any value of R_{ij} . It only tends to 1 and to 0 for small and large values of R_{ij} , respectively.

The smoothing function to be used in a metadynamics simulation needs to be carefully chosen in order to correctly reflect the process of interest. To illustrate this issue, let us consider that we are looking for a smoothing function that fits at best the following set of points: $P = [P_1(1.0, 1.0), P_2(1.5, 0.9), P_3(2.0, 0.5), P_4(2.5, 0.1), P_5(3.0, 0.0)]$. For an application to a coordination number, this set of points would define a bond between two atoms for an R_{ij} distance shorter than 1.0 Å and no bond for R_{ij} greater than 3.0 Å. We optimized the parameters of each of the five smoothing functions introduced previously (Eqs. 3.19, 3.20, 3.21, 3.22 and 3.23) with respect to the set of points P . We report the resulting functions in Figure 3.1. For each plot of the Figure, the obtained parameters as well as the set of points P are also displayed.

As one can see in Figure 3.1, not all the selected smoothing functions are flexible enough to fit the set of points P . Moreover, among those that give the most satisfactory results (*i.e.*, S_c ,

S_d and S_e), the shape of the function strongly varies from one definition to another. For an application intended to study a chemical reaction using the metadynamics technique, each of the presented smoothing function will yield a different result and the best choice of such a function will vary from one application to another. As a result, if one wants to test which smoothing function is the more adapted to study a given system, one will have to perform a specific implementation for each trial.

To overcome this issue, we imagined a different way of implementing smoothing functions in a molecular dynamics program. We use a *cubic spline* definition of the smoothing function. A cubic spline is a piecewise polynomial function. For a given set of points (knots), it defines one polynomial function in each interval of the set and ensures that the total resulting function is continuous and derivable within the range of knots. More details about cubic splines can be found in Ref. [73].

A spline is only defined in the given range of initial points. In the case of the previous example (*i.e.*, the set of points P), the spline function is only defined for values of R_{ij} between 1.0 and 3.0 Å. Outside this range, we want the smoothing function to be either 1 or 0 and to remain constant. To ensure a continuous definition of the function, we need to fix the condition for the limits of the spline definition interval. The most intuitive way to do so is to force the first derivative of the spline to be equal to zero at $R_{ij} = 1.0$ and 3.0 Å (*i.e.*, $y'(P_1) = 0$ and $y'(P_5) = 0$). In this way, we defined a continuous and derivable function in all the range of R_{ij} values, which fits the initial set of points that we targeted to reproduce. A representation of the resulting function is reported in the bottom right panel of Figure 3.1.

We implemented the definition of the smoothing function as a cubic spline in *sander* by adapting the example code reported in Ref. [73]. The user can set an option in the standard input file of *sander* to specify the file containing the given number of points used to define the smoothing function. For the example developed above, this file is given by:

```
5 ! number of points
1.0  1.0
1.5  0.9
2.0  0.5
2.5  0.1
3.0  0.0
```

At the first step of the dynamics, the cubic spline is computed and stored. Thus, at each time step, the value of the spline is evaluated as well as its first derivative.

The power of this implementation resides in the fact that it does not require to write a new piece of code to test different forms of smoothing functions. We believe that this methodology will help the user to efficiently test and chose the most relevant function to perform a metadynamics simulation depending on the chemical process of interest. However, this implementation still requires some further tests and thus, it is not yet officially distributed within the AmberTools14 program package.

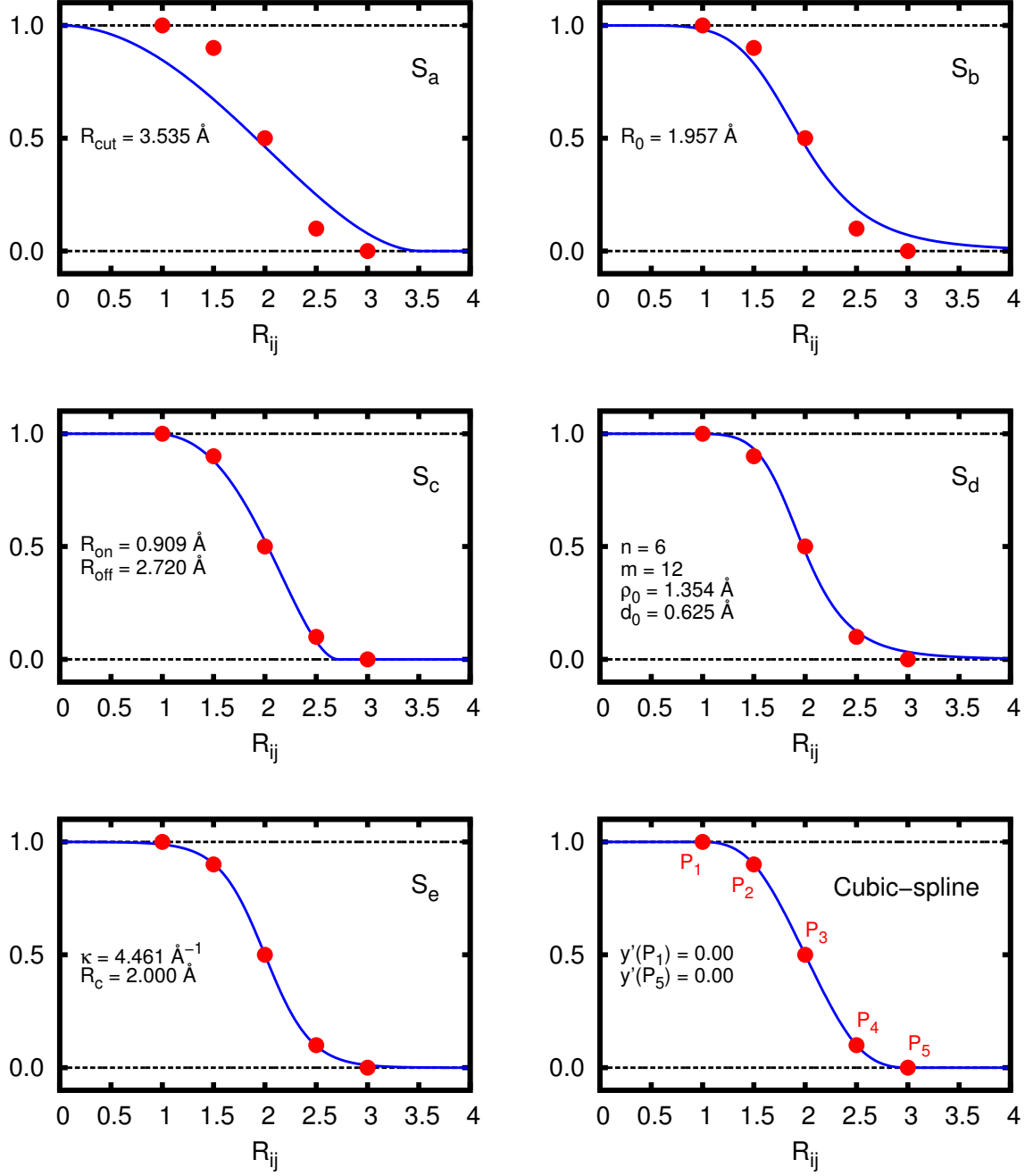


Figure 3.1: Representation of various smoothing functions (see text). The parameters of each function have been optimized to fit the initial set of points (red dots): P_1, \dots, P_5 . These parameters are displayed on the corresponding plot.

3.5 Liquid water as a test case

The SEBOMD methodology has already been successfully applied to simulate liquid water [169] as well as the N-methylacetamide molecule in water.[74] We show here an update of the former test case, while the simulations of various organic solutes in water will be investigated more deeply in Chapter 5. We also discuss the impact of the choice of the charge model on the molecular dynamics when long range interactions are taken into account through an Ewald summation scheme.

3.5.1 Structure of water: the choice of the Hamiltonian

As we will discuss intensively in the rest of this work, the results of a molecular dynamics simulation do not only depend on the computational methodology that is used. The choice of the force field, or of the SE Hamiltonian in the present case, is indeed of major importance. Here, we performed short SEBOMD simulations of liquid water using the Hamiltonians available in the current implementation of `sebomd` and compared the results obtained for the structure of water against recent experimental measurements.[66]

The system was composed of 125 water molecules in a cubic box with an edge length of 15.52 Å, corresponding to a density of 1.00 g/cm⁻³. No long range interaction scheme was used in this case and only the minimum image procedure was applied through the use of periodic boundary conditions. For each Hamiltonian, the simulation box was first equilibrated during 10 ps and the productions run was performed in the NVT ensemble for 100 ps with a time step of 1 fs. The site-site pair radial distribution functions of water were computed using `cpptraj` from the AmberTools14 program package for each of the simulations.

In Figure 3.2, we present the three site-site radial distribution functions of water (*i.e.*, $g(\text{OO})$, $g(\text{OH})$ and $g(\text{HH})$) obtained using the MNDO, AM1, PM3, RM1, PDDG/PM3 and AM1/d-PhoT SE Hamiltonians. In the case of PM3, we also present the results obtained by applying the corrections described in Section 2.3, *i.e.*, PIF2 and MAIS1. For comparison purpose, the recent experimental results of Soper for the corresponding RDF are also reported in the Figure.[66]

As it has been discussed in Chapter 2, the MNDO Hamiltonian fails to describe intermolecular interactions, especially hydrogen bonds. The RDFs obtained with this Hamiltonian confirm the latter observation, the expected peak around 1.8 Å corresponding to a hydrogen bond on the OH RDF being absent on the corresponding profile. The deficiencies of AM1 and PM3 to reproduce the structure of water have already been discussed by Monard *et al.*[169] and similar observations can be drawn for RM1. The PDDG/PM3 Hamiltonian yields results similar to those of PM3, with the peaks intensity enhanced. AM1/d-PhoT seems to yield satisfactory results for all of the three studied interactions, though the position and intensities of the peaks are slightly different compared to the experimental observations. Fi-

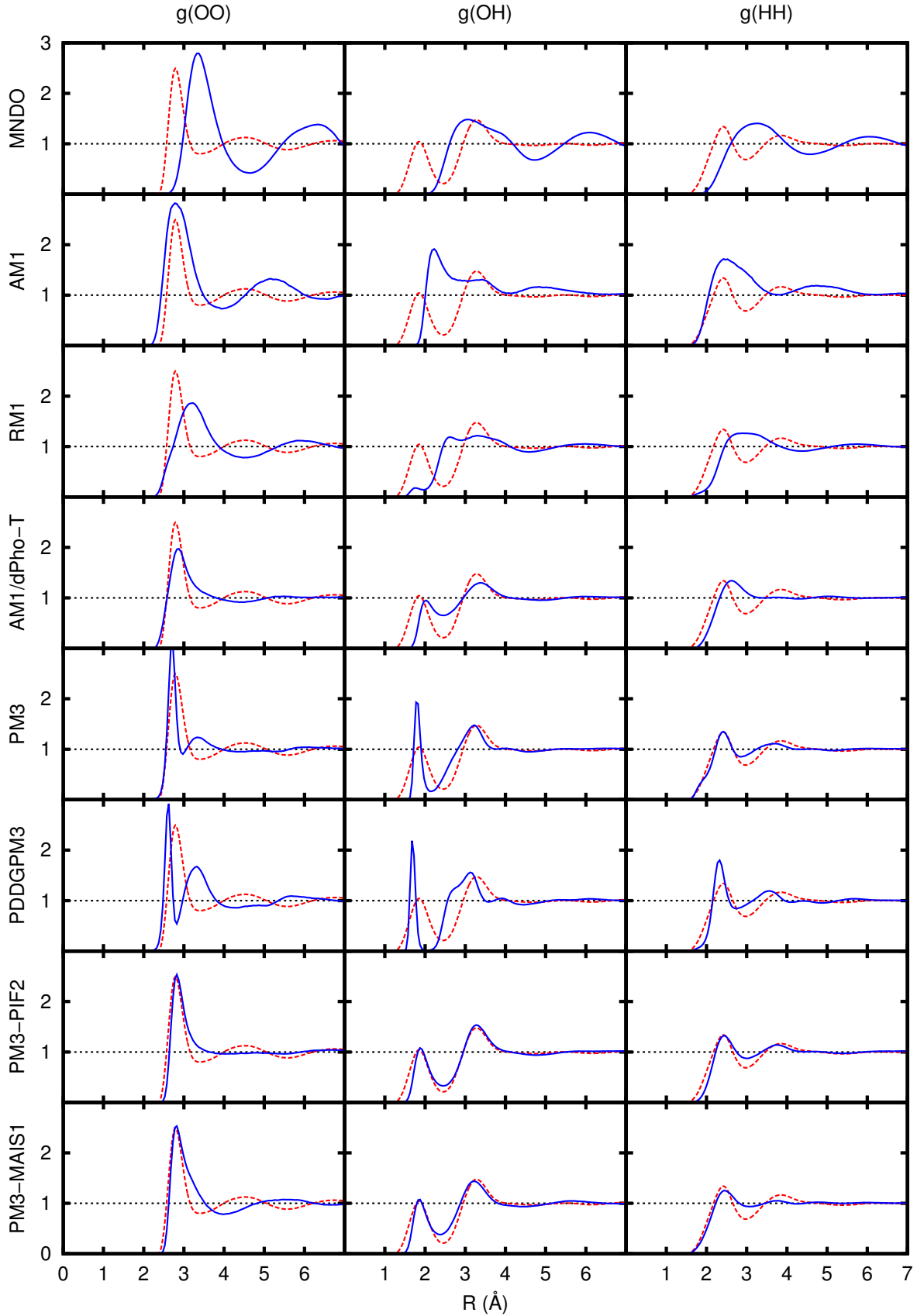


Figure 3.2: Pair radial distribution functions of water from the SEBOMD simulations using various semi-empirical Hamiltonians. Left panel: oxygen-oxygen RDF. Middle panel: oxygen-hydrogen RDF. Right panel: hydrogen-hydrogen RDF. For each plot, the experimental result by Soper[66] is represented using red dashed lines and the corresponding RDFs obtained from SEBOMD simulations are given using plain blue lines.

nally, as it has been shown by Monard *et al.*, PM3-PIF2 and PM3-MAIS1 both predict a structure of water in good agreement with the results by Soper,[66] though the shape of the second solvation shell is only fairly well reproduced in the case of the oxygen-oxygen distribution.

3.5.2 Long range interactions: the choice of the charge model

The sebomd program can perform calculations with periodic boundary conditions and, as we discussed above, long range interactions can be taken into account using a particle mesh Ewald (PME) or a standard Ewald summation scheme. During our test calculations to simulate liquid water, we pointed out that the choice of the atomic charges that are used for long range interactions can dramatically affect the result of the simulation. We shall illustrate this issue and discuss the possible ways to overcome it in what follows.

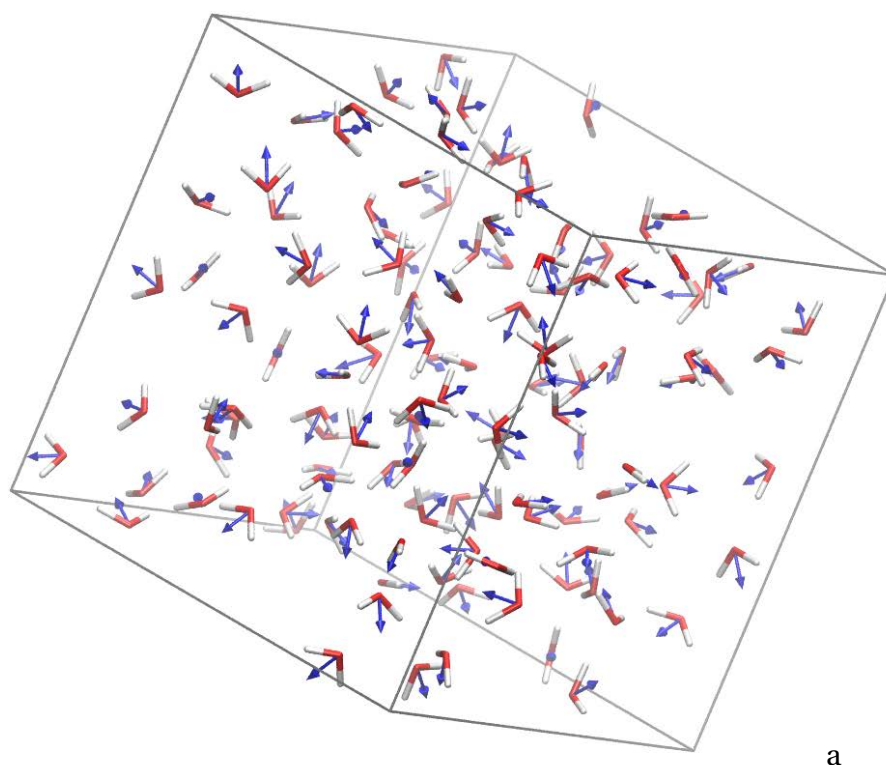
We performed SEBOMD simulations of the system presented above, containing 125 water molecules with periodic boundary conditions and the PM3-PIF2 Hamiltonian, in the same conditions as it has been discussed in the previous subsection. For long range interactions, we used a standard Ewald summation scheme included in the SCF procedure. We tested two atomic charge models for the Ewald summation, Mulliken and CM1 (referred to as Ewald Mulliken and Ewald CM1 in the following, respectively). An instantaneous equilibrated geometry of each of the two simulations is presented in Figure 3.3. While the simulation performed using Mulliken charges yields, qualitatively, the expected structure for a liquid system (Figure 3.3a), it appears clearly in the second case (*i.e.*, using CM1 charges) that all the water molecules are oriented and directed along the same axis (Figure 3.3b). To help the visualization of this phenomenon, we displayed the dipole moment of each water molecule computed from the Mulliken atomic partial charges and from the position of the atoms in the frame of the molecule.

To better quantify this phenomenon, we computed at each time step of the simulation, the averaged cosine (C_θ) of the angle between the dipole moments of each couple of water molecules. For a given time step, $C_\theta(t)$ is given by:

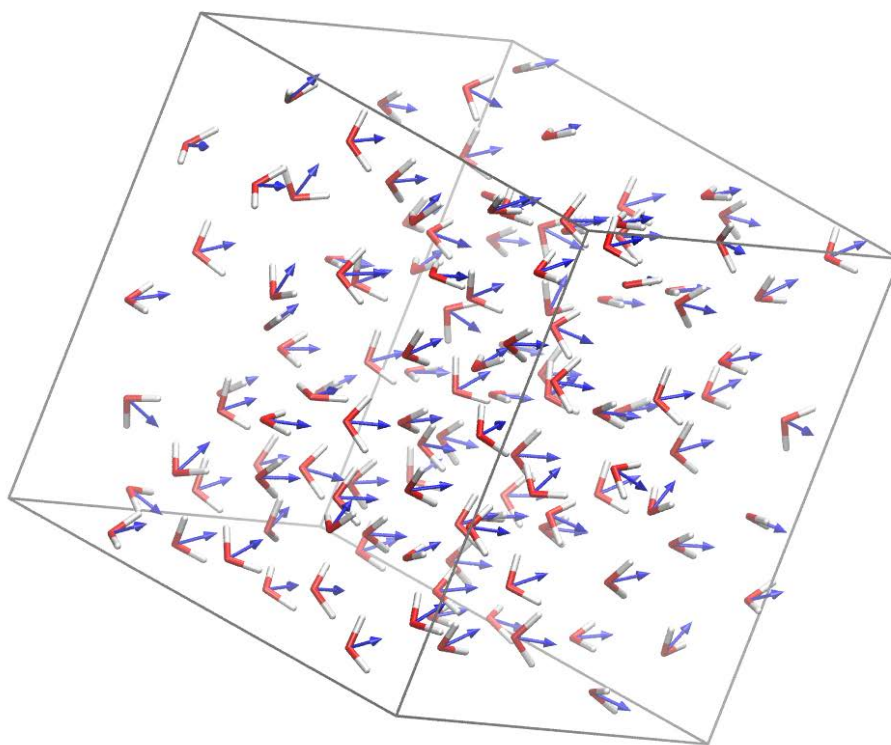
$$C_\theta(t) = \frac{1}{N_w(N_w - 1)} \sum_{i=1}^{N_w} \sum_{j>i}^{N_w} \frac{\vec{\mu}_i(t) \cdot \vec{\mu}_j(t)}{\|\vec{\mu}_i(t)\| \|\vec{\mu}_j(t)\|} \quad (3.24)$$

where N_w is the number of water molecules and $\|\vec{\mu}_i(t)\|$ defines the norm of $\vec{\mu}_i(t)$. C_θ ranges from -1 to 1 giving information on the relative orientation of the water molecules.

We performed an initial SEBOMD simulation without taking into account the long range interactions (*i.e.*, only using the minimum image convention). Starting from the last configuration of this simulation, we continued the minimum image simulation and we performed two other molecular dynamics by turning on the long range interactions, using either Mulliken or CM1 charges in the Ewald summation. For each simulation, we computed the value



a



b

Figure 3.3: Instantaneous geometries from SEBOMD PM3-PIF2 simulations of liquid water. a: using Mulliken charges in the Ewald summation. b: using CM1 charges in the Ewald summation. The dipole moment of each water molecule is represented by a blue arrow.

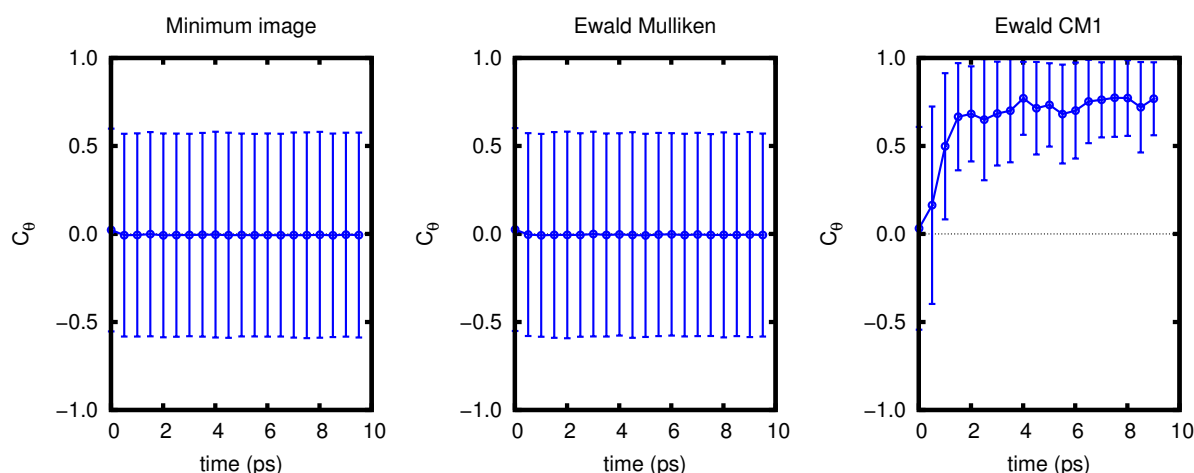


Figure 3.4: Time evolution of the C_θ descriptor (see text) and of its standard deviation along PM3-PIF2 SEBOMD simulations with PBC. Left panel: no long range interactions, only the minimum convention is applied. Middle panel: long range interactions are taken into account using a standard Ewald scheme and Mulliken atomic partial charges. Right panel: same as for the middle panel using CM1 atomic partial charges for the Ewald summation.

of C_θ at each time step and we present these results in Figure 3.4. It is clear from the Figure that during the simulation with Ewald Mulliken, the orientation of the water molecules is not much affected compared to the simulation without long range interactions. To the opposite, for Ewald CM1, the value of C_θ rapidly rises up to 0.75 and remains constant during the rest of the simulation. Moreover, the small value of the standard deviation indicates that the orientation of the molecules with respect to each others is almost fixed. In the two other cases, the standard deviation of C_θ is large, indicating that many different relative orientation of the water molecules are represented in the average.

In Table 3.1, we present the average dipole moment of each water molecule computed along a 10 ps SEBOMD simulation. The dipole moment has been computed from the atomic components of the dipole moment extracted from the wave function (μ_Ψ), or using different model of atomic partial charges (*i.e.*, μ_{Mul} , μ_{CM1} and μ_{CM2} for Mulliken, CM1 and CM2 charge schemes, respectively). We also computed this dipole moment using a fixed set of charges obtained from the TIP3P MM force field (μ_{TIP3P}). Four different simulations are considered using PBC:

- without long range interactions (minimum image).
- using the Ewald summation scheme in the SCF with:
 - Mulliken charges.
 - CM1 charges.
- using the particle mesh Ewald method with a fixed set of atomic charges obtained from the TIP3P force field.

The time average value of C_θ ($\langle C_\theta \rangle_t$) is also reported for each simulation. Gas phase results obtained for the PM3 optimized geometry of one water molecule are reported for compari-

Table 3.1: Average dipole moment computed from the wave function or using different charge models (see text) along PBC simulations with or without including the long range interactions. The values are given in Debye. The indexes n and t refer to an ensemble and time average, respectively.

	$\langle\mu_\Psi\rangle_{n,t}$	$\langle\mu_{\text{Mul}}\rangle_{n,t}$	$\langle\mu_{\text{CM1}}\rangle_{n,t}$	$\langle\mu_{\text{CM2}}\rangle_{n,t}$	$\langle\mu_{\text{TIP3P}}\rangle_{n,t}$	$\langle C_\theta\rangle_t$
Gas Phase	1.74	0.97	1.92	1.90	2.25	–
Minimum image	1.95 ± 0.08	1.25 ± 0.09	2.19 ± 0.13	2.11 ± 0.12	2.31 ± 0.14	0.00 ± 0.00
Ewald Mulliken	1.94 ± 0.08	1.23 ± 0.09	2.18 ± 0.13	2.10 ± 0.12	2.31 ± 0.14	0.01 ± 0.00
Ewald CM1	1.96 ± 0.08	1.27 ± 0.09	2.23 ± 0.13	2.15 ± 0.12	2.36 ± 0.14	0.73 ± 0.02
PME (TIP3P)	1.84 ± 0.07	1.12 ± 0.08	2.12 ± 0.12	2.05 ± 0.10	2.40 ± 0.13	0.84 ± 0.01

son.

From Table 3.1, we can observe that none of the considered charge models can reproduce μ_Ψ for an individual water molecule, neither in the gas phase nor in the condensed phase simulations. Mulliken charges systematically underestimate the dipole moment, while CM1, CM2 and TIP3P overestimate it. For the condensed phase simulations, as we observed in Figure 3.4, neither the minimum image nor the Ewald Mulliken method yields the alignment of the water molecule dipole moments since the averaged value of C_θ is null. To the opposite, the use of CM1 charges in the Ewald summation or the use of a particle mesh Ewald method with TIP3P charges both result in a simulation in which the water molecules are aligned.

The above observations tend to show that, if the charge model used to model long range electrostatic interactions overestimates the dipole moment of the water molecules with respect to value predicted using the wave function, then a non physical electric field is created outside the unit simulation cell, forcing the alignment of the molecules. To the opposite, if the charge model underestimates the dipole moment, then the long range interactions will be underestimated and no additional electric field will be formed. Another observation that arises from this analysis is that this phenomenon does not depend on the Ewald scheme, since the same effect is observed using a standard Ewald summation in the SCF with CM1 charges or with an external PME using a fixed set of MM charges.

To further test the impact of the set of charges on the system, we performed a series of 10 ps SEBOMD simulations using PBC and PME for long range interactions. The charge of each hydrogen atom was set to q_H while the charge of the oxygen atoms was $-2q_H$, and we changed the value of q_H from one simulation to another. The results are presented in Figure 3.5.

We can see from Figure 3.5 that the value of q_H has almost no effect on the structure of water up to a critical value. For values of q_H greater than ~ 0.335 e, the charges used in the Ewald summation start to gradually induce an unphysical structuration of liquid water.

In order to fix this issue, it is necessary to obtain a set of atomic partial charges that

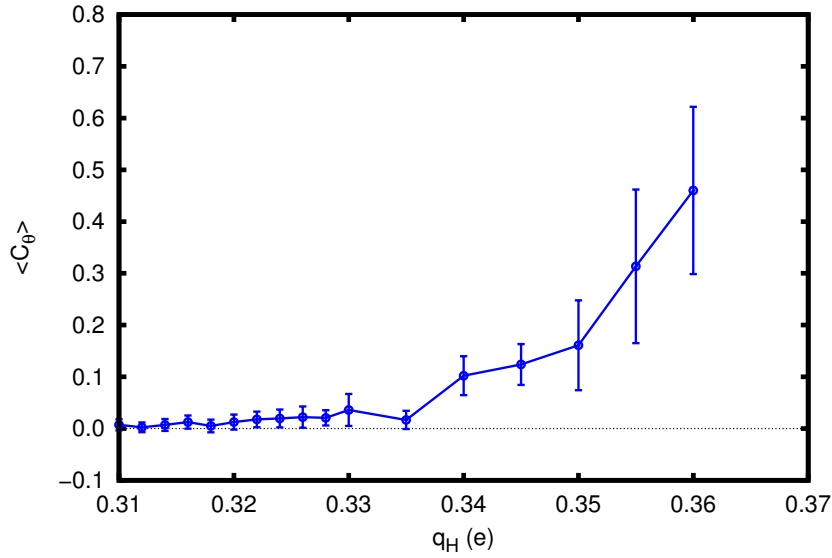


Figure 3.5: Variation of the C_θ descriptor for SEBOMD simulations performed using a particle mesh Ewald scheme for long range interactions with different sets of atomic charges (q_H).

reproduces at best the dipole moment of the wave function. To this end, we have envisaged different approaches. i) One can find a correct value of q_H and use only a particle mesh Ewald method for long range interactions. ii) In order to use a standard Ewald scheme included in the SCF procedure, one should derive a new charge model intended to reproduce the wave function dipole moment for each considered semiempirical Hamiltonian.

In addition, a different strategy to be applied to a standard Ewald method included in the SCF procedure is currently under way. To ensure that the set of atomic charges (q_A^*) reproduces the total QM dipole moment (μ_Ψ^T), we require the following equation to be true:

$$\vec{\mu}_\Psi^T = \sum_A q_A^* \vec{R}_A \quad (3.25)$$

where q_A^* and \vec{R}_A are the atomic charge and the position vector of the atom A, respectively. We have seen above that Mulliken charges fail at reproducing the wave function dipole moment. We note $\vec{\mu}_{Mul}^T$, the total dipole moment computed using Mulliken atomic charges. By assuming that $\vec{\mu}_{Mul}^T$ and $\vec{\mu}_\Psi^T$ are collinear, we can apply a scaling factor (γ) to $\vec{\mu}_{Mul}^T$ so that:

$$\vec{\mu}_\Psi^T = \gamma \vec{\mu}_{Mul}^T \quad (3.26)$$

and thus:

$$\vec{\mu}_\Psi^T = \sum_A \gamma q_A^{Mul} \vec{R}_A \quad (3.27)$$

where q_A^{Mul} is the Mulliken charge of the atom A. Multiplying the two parts of Eq. 3.27 by $\vec{\mu}_\Psi^T$ yields:

$$||\vec{\mu}_\Psi^T||^2 = \sum_A \gamma q_A^{Mul} \vec{R}_A \cdot \vec{\mu}_\Psi^T \quad (3.28)$$

and consequently:

$$\gamma = \frac{||\vec{\mu}_\Psi^T||^2}{\sum_A q_A^{\text{Mul}} \vec{R}_A \cdot \vec{\mu}_\Psi^T} \quad (3.29)$$

The set of q_A^* charges that would satisfy Eq. 3.25 is thus:

$$\begin{aligned} q_A^* &= \gamma q_A^{\text{Mul}} \\ q_A^* &= \frac{||\vec{\mu}_\Psi^T||^2}{\sum_B q_B^{\text{Mul}} \vec{R}_B \cdot \vec{\mu}_\Psi^T} q_A^{\text{Mul}} \end{aligned} \quad (3.30)$$

Some preliminary calculations on single conformations of 125 water molecules have shown encouraging results. The total wave function dipole moment of the simulation box is reproduced by the charges obtained from Eq. 3.30, as well as the molecular dipole moment of each water molecule.

The methodology described above can be implemented in an Ewald summation algorithm in a straightforward way. The great advantage of such an approach is that, at each step of an SCF cycle, the set of charges used for the long range interactions will be consistent with the current electronic density matrix.

The next step would be the implementation of this protocol in *sebond* and its validation for condensed phase studies. In the present version of the *sebond* code, the Ewald scheme can be used only by selecting Mulliken charges. This approach is the choice made in the literature for QM/MM studies including long range interactions.[193]

3.6 Concluding remarks

The SEBOMD methodology has been recently implemented in the GNU AmberTools14 and is now officially distributed with this program package. We have worked on improving and simplifying the interface between *sander* and *sebond* in order to allow a straightforward implementation of further considered methods.

Because SEBOMD is a new methodology, the development of various tools has been necessary to conduce the present work, such as a normal mode analysis tool. Furthermore, we took advantage of the fact the *sebond* is now part of *sander* to help the development of an original definition of the smoothing functions used to perform metadynamics simulations.

The test case on liquid water has shown that not only the molecular dynamics methodology plays a significant role in the quality of the results but also the choice of the semiempirical Hamiltonian is of dramatic importance. It also shows that the PIF and MAIS strategies give satisfactory results for the structure of liquid water.

In this project, we have developed most of the tools that are necessary for an extensive application of the SEBOMD methodology, revealed some issues and proposed a few paths

for solving them. In the next Chapters, we shall present further method developments and some applications intended, on one hand, to optimize the semiempirical Hamiltonian chosen for the simulation of organic solutes in the condensed phase and, on the other hand, to provide some results that can be compared with experimental and theoretical studies in the literature. Indeed, the developments and the applications carried out in this project open the way to further tests and improvements toward the study of large biological systems and of their environment.

chapter

4

**PIF3: improvement of semiempirical methods
for the interaction of water with
hydrophobic groups**

Résumé

L'étude présentée dans ce chapitre vise dans un premier temps à évaluer la capacité des méthodes semiempiriques courantes à traiter les interactions intermoléculaires en solutions aqueuses. Peu d'études dans la littérature ont été consacrées au test des méthodes semiempiriques vis-à-vis des interactions entre molécules d'eau et groupements ou molécules hydrophobes. Ce travail cherche notamment à combler ce manque et nous nous focalisons sur un système modèle composé d'une molécule d'eau et d'une molécule de méthane. La deuxième partie de cette étude est dédiée à l'amélioration de l'approche PM3-PIF2 et mène à la définition d'une nouvelle correction, PM3-PIF3. Enfin, la dernière partie vise à évaluer la généralisation de PM3-PIF3 à l'étude des interactions entre une molécule d'eau et des composés présentant à la fois un caractère hydrophobe et un caractère hydrophile.

Le système modèle choisi pour cette étude est le complexe 1:1 méthane-eau. Nous avons exploré la surface d'énergie potentielle de ce complexe au niveau MP2/aug-cc-pVTZ, selon différentes coordonnées choisies pour refléter les interactions clés entre ces deux molécules. Nous nous attardons notamment sur l'interaction entre un atome d'hydrogène de l'eau (H_W) et un atome d'hydrogène du méthane (H_C), ainsi que sur l'interaction entre un atome d'oxygène de l'eau (O_W) et un H_C . L'analyse des résultats obtenus en utilisant les méthodes semiempiriques courantes montre qu'aucune d'entre elles ne fournit une description satisfaisante de ces interactions. Nous remarquons que la plupart des méthodes choisies présente des artefacts sur la surface d'énergie potentielle principalement dus à l'utilisation de fonctions de corrections Gaussiennes dans leur développement. Le test de la méthode PM3-PIF2, développée dans notre groupe, fournit une surface dépourvue d'artefact. En revanche, nous notons une forte surestimation de la répulsion entre H_W et H_C , faussant ainsi la description de l'interaction entre l'eau et le méthane.

Nous proposons une amélioration de la méthode PM3-PIF2, visant à donner une description correcte de l'interaction d'une molécule d'eau avec des composés à caractère hydrophobe et/ou hydrophile. La nouvelle méthode PM3-PIF3 se base sur une différenciation des atomes d'hydrogène présents dans une molécule organique en fonction de leur environnement direct. Dans cette approche, un atome d'hydrogène lié à un atome de car-

bone est considéré comme hydrophobe (H_C) alors qu'un atome d'hydrogène porté par un hétéroatome est défini comme hydrophile (H_X). Nous avons optimisé les paramètres décrivant les interactions $H_W H_C$ et $O_W H_C$ à partir de configurations représentatives de la surface d'énergie potentielle du complexe méthane-eau. Pour toutes les autres interactions, les paramètres originaux de la méthode PM3-PIF2 sont conservés.

La méthode PM3-PIF3 est tout d'abord testée sur un système plus large et plus concret, un modèle d'hydrate de méthane. Ce type de structure est généralement formé d'une molécule de méthane emprisonnée dans une cage de 20 molécules d'eau. L'analyse de la structure et de l'énergie de stabilisation de ce système montre un très bon accord entre les résultats obtenus avec PM3-PIF3 et ceux reportés dans la littérature à de plus hauts niveaux de théorie. L'effet de la cage d'eau sur les propriétés vibrationnelles du méthane est également bien reproduit par PM3-PIF3 au regard des observations expérimentales. Nous soulignons que certaines méthodes récentes, telles que PM6 et PM7, sous-estiment très largement la valeur des fréquences associées aux vibrations OH, caractéristiques de la molécule d'eau, d'environ 1000 cm^{-1} par rapport aux résultats expérimentaux. Par opposition, l'erreur commise par la méthode PM3-PIF3 pour le calcul de cette quantité est plus acceptable, soit environ 200 cm^{-1} .

Pour tester la transférabilité de la méthode PM3-PIF3 à l'étude de composés organiques divers en interaction avec l'eau, nous avons mené une série de calculs sur une large variété de molécules. Notamment, nous étudions l'interaction d'une molécule d'eau formant une liaison hydrogène avec une molécule d'éthanol. Nous comparons les résultats MP2/aug-cc-pVTZ avec ceux obtenus en utilisant diverses méthodes semiempiriques, incluant PM3-PIF2 et PM3-PIF3. Parmi les Hamiltoniens testés, nous observons que la récente méthode PM7, présente un artefact important sur la surface d'énergie potentielle de cette interaction fondamentale. PM3-PIF2 fournit un résultat qualitativement satisfaisant par rapport aux calculs de références mais surestime cependant l'énergie d'interaction entre les deux molécules. Cette erreur est due à la surestimation de la répulsion entre la molécule d'eau et le groupement méthyl de l'éthanol. La méthode PM3-PIF3 corrige cette erreur, conduisant à un accord remarquable avec MP2.

Enfin, l'Hamiltonien PM3-PIF3 apparaît comme un très bon candidat pour l'étude de la solvation de molécules bio-organiques dans l'eau. Ceci constitue la prochaine étape de ce travail et sera détaillé dans le chapitre suivant.

Semiempirical molecular orbitals based methods have been developed and used for almost 50 years. They represent a valid alternative to *ab initio* approaches since they allow keeping a quantum mechanical description of large systems with a much reduced cost. The interest for these methods has not only persisted along the last few decades but it has even increased in recent years due to (at least) two main reasons. On the one hand, semiempirical methods are particularly well adapted to linear scaling calculations and parallel computing technologies.[194–196] In systems of biological interest, for instance, it becomes possible to treat a macromolecule (or a large part of it) and its environment by using a quantum electronic Hamiltonian. Various efficient programs have been recently developed for this purpose, such as SEBOMD (see Chapter 3), LocalSCF[197] or EMPIRE.[198] On the other hand, the use of semiempirical electronic Hamiltonians is extremely appealing in the perspective of treating systems of large size along relatively long statistical simulations (see for instance Refs. [169, 199] and references therein). However, attention must be paid when treating intermolecular interactions, since originally these methods were developed by introducing a parametrization of the Hamiltonian intended to fit experimental measurements (or in some cases *ab initio* results) of physical-chemical properties of isolated molecules.

4.1 Using semiempirical methods to describe intermolecular interactions

Most of the recent semiempirical Hamiltonians are based on the NDDO approximation (Neglect of Diatomic Differential Overlap)[91] introduced by MNDO[114] and its related methods AM1[115] and PM3[78] (see Chapter 3). Despite considerable improvements provided by the latter in the description of intermolecular interactions with respect to MNDO, major problems remain. It was pointed out, for instance, that the AM1 and PM3 methods give a poor description of the water dimer energy surface [132, 168, 200–202] leading to erroneous pair distribution functions of liquid water, when the corresponding Hamiltonians are used to model the condensed phase (see Chapter 3). Moreover, both methods lead to unphysical intermolecular bonds. For instance, PM3 predicts quite stable geometries for A-H...H-B interactions with an intermolecular H...H bond distance of about 1.7-1.8 Å[138, 168] (the interaction energy is about -2 kcal/mol). This feature is responsible for many deficiencies of the method and in particular the wrong prediction of stereochemical molecular properties.[203, 204]

As we discussed in Chapter 2, quite a few efforts have been made in the past years to improve the potential of standard semiempirical Hamiltonians to be employed in the description of clusters and of condensed phase systems. The focus has been put on hydrogen bonds because they represent an essential interaction in aqueous systems and in molecules of biological interest. A thorough reparameterization of the PM3 method has been provided

by J. J. P. Stewart. The most recent versions of the method, PM6[131] and PM7[141], have allowed eliminating some of the artifacts of PM3 for crystal structures and for the description of hydrogen bonds. Thiel and collaborators have carried out extensive work to revise standard NDDO schemes leading to the so called orthonormalization corrected models (OM1-OM3, see Ref. [85] for an updated review), which were recently improved in order to achieve a realistic description of proton transfer in bulk water.[153]

Extensive work has been devoted in recent years to the introduction of different types of *ad hoc* corrections to the interaction energy, in particular for the dispersion term, starting with the work by Martin and Clark.[161] The use of an empirical dispersion term in Hartree-Fock [205, 206] and Density Functional Theory [28, 207] methods was generalized to semiempirical methods in a straightforward way ('D' correction).[162, 208] Terms for hydrogen bonds description have been introduced by Hobza's and Korth's groups through a series of corrections: -DH, DH2, -DH+, -D3H4.[163–167]

We introduced in Chapter 2 a different type of correction that have been developed in our group: the PIF and MAIS approaches. The main concern of the former was to avoid the use of Gaussian correction functions (GCFs) in the core-core term of the PM3 Hamiltonian. The last version of the PIF parameters for PM3, PM3-PIF2, has been proven to yield good results for the study of the solvent effects experienced by the N-methylacetamide molecule in aqueous solution.[74]

Although the most up to date versions of semiempirical methods significantly improve the description of intermolecular interactions for hydrogen-bonded systems, the reliability of these methods to describe hydrophobic interactions has been less investigated. This is however an important topic because, as said above, one of the main shortcomings of AM1 and PM3 was the prediction of stable unphysical H...H bonds, which artificially form between non-polar hydrocarbon groups and water molecules, thus preventing the correct hydrophobic/hydrophilic balance description. Checking whether or not recent methods correctly describe the hydration of hydrophobic compounds is obviously essential to investigate the chemistry of biological systems in aqueous media.

In this work, one first aim is to provide a detailed study of the performance of semiempirical methods applied to hydrophobic molecule-water interactions. For the sake of simplicity, the calculations will be illustrated with a simple model system, the 1:1 methane-water complex. Different methods will be tested against *ab initio* references (MP2[15]/aug-cc-pVTZ,[209] CCSD(T)[210]/aug-cc-pVTZ). As it will be explained below, none of the methods tested is able to give satisfactory results. Our second aim is, consequently, to optimize a set of PIF parameters in such a way that they correctly describe solvation phenomena in water not only of polar molecules (which was the main target of previous investigations) but also of non-polar ones. Starting from the PIF2 parameters, we introduce the idea of hydrogen type-specific parameters for treating the core-core intermolecular interaction terms and the

new set of parameters is named PM3-PIF3. PM3-PIF3 can therefore be seen as an elaborated quantum mechanical force-field that is able to describe the solvation of organic molecules in aqueous solutions to a much lower cost than *ab initio* or Density Functional Theory methods. The results presented in this Chapter have been recently published in Ref. [140].

4.2 Hydrocarbon-Water Interactions

The first step of our analysis was to test the behavior of some of the most employed semiempirical NDDO methods to describe the potential energy surface in the case of hydrophobic molecules interacting with water. We chose, as a model system, the 1:1 methane-water complex. As previously noted, our reference calculations were run at the MP2 level of theory with an aug-cc-pVTZ basis set. Corrections to the basis set superposition error (BSSE) were not included in the calculation of the interaction energies. However, the PESs obtained at the MP2 level were tested against the corresponding at the CCSD(T) level of theory with the same basis set (see the Supplementary Material for a comparison between the two methods). The *ab initio* calculations were performed using Gaussian09,[188] whereas the calculations using semiempirical Hamiltonians were performed by means of the Amber14 suite of programs [184] (using the sqm and seboimd modules, see Chapter 3) and MOPAC2012.[211]

4.2.1 Test of NDDO Methods using GCFs in the core-core terms

We first tested the following methods: MNDO,[114] MNDO/d,[145–147] AM1,[115] AM1/d-PhoT, [128] RM1,[127] PM3,[78] PM6,[131] PM7.[141] Then, among the methods corrected for hydrogen bonds and/or dispersion described above, we considered AM1-D, PM6-D, AM1-DH+, PM6-DH+, and PM6-DH2. We also analysed the results obtained with the PM3-CARB1 [136] model (developed to improve the description of flexible and polar carbohydrates), the PDDG/PM3 as well as the PDDG/MNDO models,[133–135] (developed to improve the relative stability of isomers for hydrocarbons and organic molecules containing the heteroatoms N and O).

In Figure 4.1 we report the interaction energy of the methane-water complex obtained for a choice of semiempirical methods (AM1, AM1-DH+, PM3, PDDG/PM3, PM6, PM6-DH+, PM7) and for the MP2 level. The integrality of our results including more semiempirical Hamiltonians is available as Supplementary Material. On the left hand side of Figure 4.1, we analyze the behavior of the interaction energy of the complex as a function of the distance between a methane hydrogen atom and a hydrogen atom on the water molecule (r_{HH}) in order to check for any possible unphysical stabilization. On the right hand side of the Figure we report the corresponding results along the coordinate defined by the distance between a methane H atom and the O atom of water (r_{OH}) in order to inspect the behavior of these

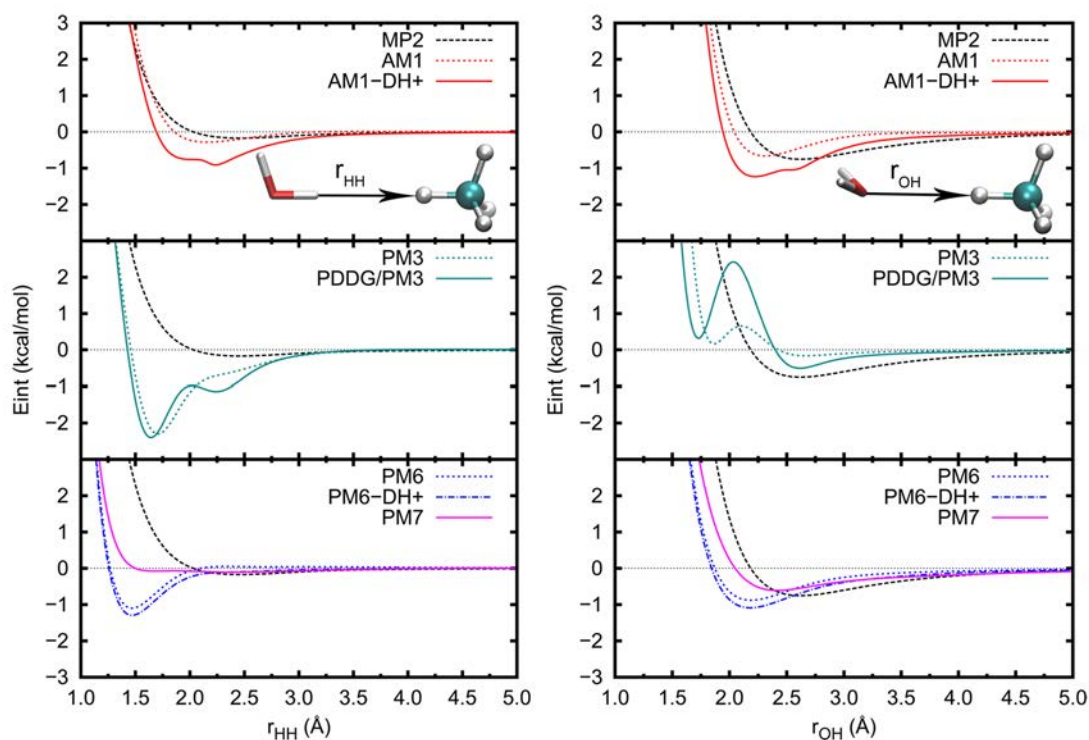


Figure 4.1: Performance of a choice of semiempirical methods to reproduce the MP2/aug-cc-pVTZ curves (black dashed curves) describing the interaction between the H atom of water and the H atom of methane (left plots) and between the O atom of water and the H atom of methane (right plots).

methods under the formation of a weak C-H...O hydrogen bond. In the two cases, a scheme of the respective orientation of the two molecules is presented as an inset in the top plots.

While AM1 well reproduces the H...H interaction curve compared to the MP2 reference, the corresponding Hamiltonian reparametrized by including a dispersion and a hydrogen bond (DH+) correction overestimates the interaction and predicts two minima around 2.0 and 2.3 Å, about -1.0 kcal/mol deep. For the same interaction, both PM3 and PM6 predict a minimum (about -2.2 kcal/mol and -1.0 kcal/mol, respectively) at a much shorter distance (at about 1.7 and 1.5 Å). The DH+ correction of PM6 does not change the shape of the interaction energy profile but slightly increases the depth of the energy minimum. On the other hand, the use of the PDDG correction for PM3 induces a second, less deep minimum on this profile. The interaction energy profile obtained with PM7 is qualitatively correct though flatter compared to the MP2 reference and it does not reproduce the repulsive part of the interaction, which starts at shorter intermolecular distance (roughly at an H...H distance equal to 1.5 Å, *i.e.*, about 0.5 Å shorter than in the case of MP2).

Concerning the O...H interaction, when comparing with the MP2 curve the first thing that we observe is that most of the methods considered predict a minimum of the interaction energy at a shorter distance, comparable to that of a weak hydrogen bond. The case of PM3 is quite surprising. Without correction, the PM3 profile presents two minima at about 2.7 and 1.9 Å. The minimum at 1.9 Å corresponds to a metastable structure that lies about 0.3 kcal/mol above the energy of the dissociated complex and is separated by about 0.5 kcal/mol

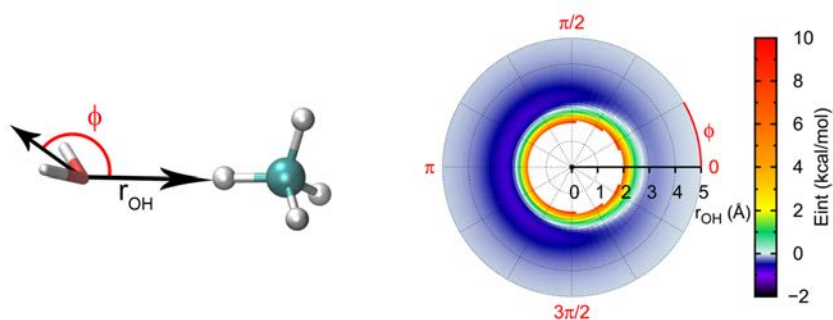


Figure 4.2: Definition of the ϕ angle used to describe the orientational dependence of the interaction energy, and the corresponding plot of the surface obtained at the MP2/aug-cc-pVTZ level.

from the second (very shallow) energy minimum. This behavior is enhanced when applying the PDDG correction: the position and energy of the two minima are little affected but the barrier that separates them appears to be higher, reaching 2.0-3.0 kcal/mol. The PM7 curve, on the other hand, is in reasonably good agreement with the MP2 reference.

To further explore the PES of the complex, we extended our investigation to the angular dependence of the interaction energy. We introduced a two dimensional rigid scan of the interaction energy of our complex in terms of the r_{OH} distance and of the ϕ angle between the water bisector and the O...H vector, as shown on the top of Figure 4.2. On the right hand side of this Figure we report the interaction energy surface computed at the MP2 level using cylindrical coordinates. We observe that the interaction energy has a minimum basin centered around $r_{OH} = 2.7$ Å, in a region where $\phi \in [2\pi/3, 4\pi/3]$. The same surface was calculated using different semiempirical Hamiltonians and the corresponding results are reported in Figure 4.3 and in the Supplementary Material. In the case of the PM3 Hamiltonian (as well as PDDG/PM3), we qualitatively observe a mirror surface compared to the MP2 reference: in this case, the interaction is dominated by the presence of a minimum in the H...H interaction, *i.e.*, for values of $\phi \sim 0$. Regarding the PM6 method (as well as PM6-DH+), the two

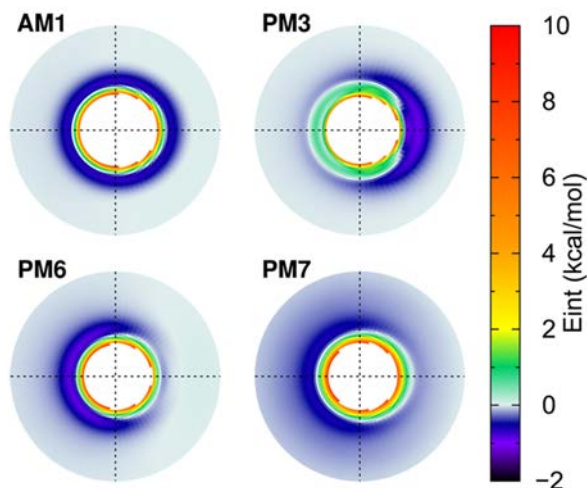


Figure 4.3: Orientational dependence of the methane-water interaction energy, using the ϕ angle, for a choice of semiempirical methods.

dimensional plots display a much better agreement with the MP2 reference compared with PM3 but, as we have already shown, the H...H curve has a well defined minimum. In addition, by inspecting the results obtained for the AM1 (as well as the AM1-DH+) method we can conclude that this Hamiltonian does not reproduce correctly the orientational dependence of the interaction energy, since all orientations are equally favorable for a given value of the O...H distance. Finally, PM7 displays the best agreement with the MP2 PES and, among all the methods tested above, it is the only one to provide a qualitatively correct picture of the methane-water PES, in spite of a significant underestimation of the intermolecular repulsion for H...H distances in the range around 1.5-2.0 Å. We would like to mention that, though this can be considered as a minor issue for the overall behavior of the PM7 method, it could lead to non-negligible errors in the description of the solvation structure around hydrophobic groups in MD simulations in aqueous systems.

4.2.2 Test of NDDO Methods Using a Core-Core Parametrizable Interaction Function (PIF)

It has been shown that significant improvements to the description of intermolecular interactions by NDDO methods can be achieved by modifying the analytical form of the core-core interactions, and by providing its reparametrization based on *ab initio* calculations.[132, 137]

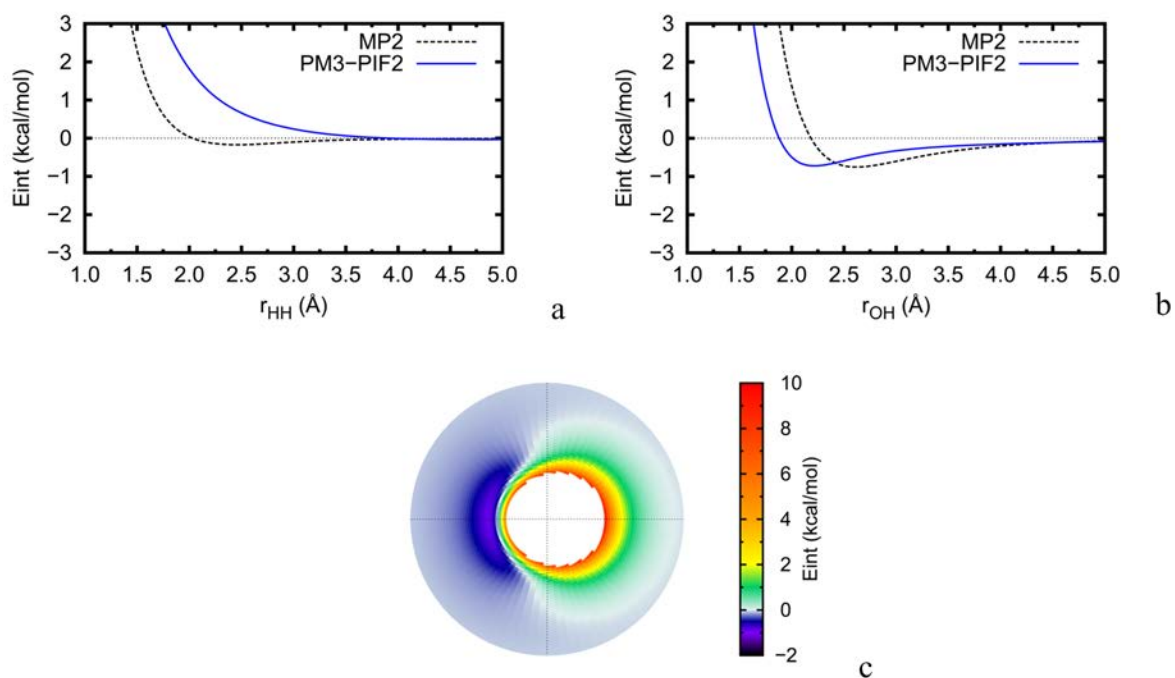


Figure 4.4: Interaction between the H atom of water and the H atom (top plot) and the O atom (middle plots) of a methane molecule: comparison between the MP2 curve (black dashed line) and the curve obtained with the PM3-PIF2 Hamiltonian (blue line). In the bottom plot (c) we report the angular dependence of the PM3-PIF2 interaction energy.

In this work, we have tested the PM3-PIF2 (described in Chapter 2) to the methane-water interactions, according to the analysis presented in the previous Subsection for other semi-empirical Hamiltonians. The results are presented in Figure 4.4. As a first remark, we note that, as already observed for the other semiempirical models tested, the minimum for the O...H hydrogen bond interaction is located at a shorter distance compared to the MP2 minimum. Most importantly, the PES for the methane-water interaction calculated along the H...H coordinate is too repulsive for H...H distances shorter than 3 Å, in contrast with PM7, which predicts too small a repulsion. Despite the qualitative agreement between the PIF2 angular distribution of the interaction energy and the corresponding MP2 results (Figure 4.4), we can conclude that the PM3-PIF2 method is not sufficiently general to describe the interaction of hydrophobic groups in water. In the next Section, the method is recalibrated from an *ad hoc* reparametrization using the MP2 methane-water dimer PES, and a new set of PIF parameters (PIF3) is derived.

4.3 Recalibration of the PM3-PIF2 Method: PM3-PIF3

Before discussing the approach proposed here to improve hydrocarbon-water interactions, some comments on PM3-PIF2 are necessary. This set of parameters was developed to describe hydrated systems. The training set included different complexes of water with polar molecules but also with methane. As we discussed in Chapter 2, the total number of configurations was 1485, from which about 25% (377 to be precise) corresponded to the methane-water complex PES. The PM3-PIF2 parameters, therefore, should be close to the most favorable solution that can be found within the PIF approximation to describe the interactions of water with both hydrophilic and hydrophobic compounds.

Our attempts to improve the PM3-PIF2 results in this work keeping the same theoretical framework proved to be rather unfruitful. We considered tuning only the parameters for the H...H interaction, since they are mostly responsible for the exaggerated methane-water repulsion obtained along the corresponding coordinate. Starting from the original PM3-PIF2 parametrization, we increased the weight of the methane-water PES, while keeping the other parameters fixed to their PM3-PIF2 value (see the Supplementary Material for details about the parametrization protocol). After the optimization of H...H interactions, the repulsive part of the PES along the HO-H...H-CH₃ coordinate was ameliorated but the H₂O...H-CH₃ interaction became too attractive, with a minimum about 2 kcal/mol deep (see the Supplementary Material for the PES obtained in this case). The conclusion that can be derived from this analysis is that the development of NDDO methods capable of describing intermolecular PESs with chemical accuracy does probably require a profound change in the parametrization strategies that have been considered up to now. However, focusing on the core-core terms is still useful, because one preserves the simplicity of pairwise additive potentials.

4.3.1 The PIF3 strategy

A more elaborated description of the methane-water complex in such a case may require taking into account the chemical environment of the atoms, and not just their effective core charges. This points to the development of valence-dependent functions, in which some information about the molecular wave function is included into the core-core terms, which can be done at different theoretical levels. A simple approach consists in using the effective charge of the atom (obtained by a Mulliken population analysis for instance), and this method has indeed been assumed by previous works on dispersion corrections cited in Chapter 2. Even simpler is the definition of different PIF corrections for different atom-types, and not only for different atomic numbers, somehow imitating the guidelines of classical force-fields. This choice has also been made to develop hydrogen bonds corrected semi-empirical methods in the past.[129] The main drawback of previous approaches that contemplated this kind of developments is that Gaussian terms were preserved, leading to the unphysical artifacts that have been illustrated in Figure 4.1. The proposal here is to replace the Gaussian functions by the physically meaningful, atom-type dependent PIFs.

Specifically, in this work, we have explored the improvement that can be gained through the categorization of PIF contributions by H-atom type, basically by distinguishing between H atoms linked to a carbon atom and those linked to a heteroatom. Obviously, more elaborated schemes can be imagined, for instance by differentiating H-atoms linked to O or N, to saturated, unsaturated or aromatic carbon atoms, *etc.* However, in the present work we focused on evaluating the capabilities of current semiempirical methods to describe the interactions of water with hydrophobic molecules and to propose some directions for future research.

In the following, a hydrogen atom connected to a heteroatom will be considered as “hydrophilic” (H_X), while a hydrogen atom bonded to a carbon atom will be considered as “hydrophobic” (H_C) and correspondingly two different sets of parameters will be used. In the first case, the PM3-PIF2 parameters previously developed in Refs. [138, 204] are employed, whereas a new parametrization will be provided in this work in the latter case. Since we are interested on hydrated systems, only $H_X...H_C$ and $O...H_C$ functions are relevant and will be optimized here.

4.3.2 Parametrization procedure and test calculations on the methane-water complex

We used four different training sets of configurations on which we parametrize the PM3-PIF3 Hamiltonian. Each considered geometry consists of one water molecule and one methane molecule. The two first sets correspond to the methane-water rigid scans presented in Section 4.2 (Figure 4.1), to include explicitly the $H...H$ and $O...H$ interactions in the parameter-

ization procedure. For the other two sets, the methane-water complex configurations were extracted from molecular dynamics simulations of a system that was composed by 128 explicit water molecules and one methane molecule in a cubic box of 15.681 Å (leading to a density of 1.00 kg/m³). The simulations were performed with the sander program of Amber14. One was carried using the MM force field ff03[38] for methane and the SPC/E model of water,[212] while the second simulation was performed in the SEBOMD methodological framework (as implemented in sander, see Chapter 3) with the PM3 Hamiltonian to describe the whole system. More details about the simulation protocol (equilibration, simulation runs, *etc.*) are given in Chapter 5.

From these two simulations, we extracted snapshots every 50 fs, which led to 10,000 and 6,400 snapshots, respectively for the MM-MD and for the SEBOMD simulation. The same procedure (described in the following) was applied to select methane-water geometries in the two cases. From the extracted snapshots, we analyzed and ranked each couple of methane-water molecules as a function of the C(methane)...O(water) distance by applying a cut-off of 5.5 Å. The geometries were then sorted in 0.1 Å wide windows (C...O distance). For each window containing more than 100 structures, we randomly selected 100 geometries of the methane-water complex. This procedure led to 2600 configurations having a C...O distance in the range [2.9-5.5] Å for the MM-MD simulation, and 2700 configurations having a C...O distance in the range [2.8-5.5] Å for the SEBOMD simulation.

The interaction energy between water and methane has been computed at the MP2/aug-cc-pVTZ level of theory for the whole set of configurations. The basis set superposition error (BSSE) was not included in the calculation, in order to remain consistent with the previ-

Table 4.1: PIF3 parameters for organic molecules interacting with water. The parameters reported with an asterisk were obtained in this work, whereas the others were obtained in Refs. [138, 204]. All parameters are in atomic units.

(A,B)	α_{AB}	β_{AB}	χ_{AB}	δ_{AB}	ϵ_{AB}
(H _X ,H _X)	2.47949	2.896120	47.262	-505.73	1,549.81
(H _C ,H _X)*	370.07613	3.585786	29.896	-315.75	513.54
(H _X ,C)	0.53126	0.830561	-353.470	4643.40	-7083.63
(H _X ,N)	33.11381	1.782993	-172.487	158.43	2423.97
(H _X ,O)	29.32517	2.092648	-55.779	44.53	313.44
(H _C ,O)*	18.55763	1.825201	-78.110	277.11	-305.52
(C,N)	0.94697	0.981346	2.085	6.12	102.30
(C,O)	0.87884	0.886032	1.879	33.95	843.67
(N,N)	1.43533	0.841558	1.941	35.45	887.74
(N,O)	1.94702	1.016194	1.941	35.45	887.74
(O,O)	75.15466	1.512063	338.656	-32,185.68	274,634.10

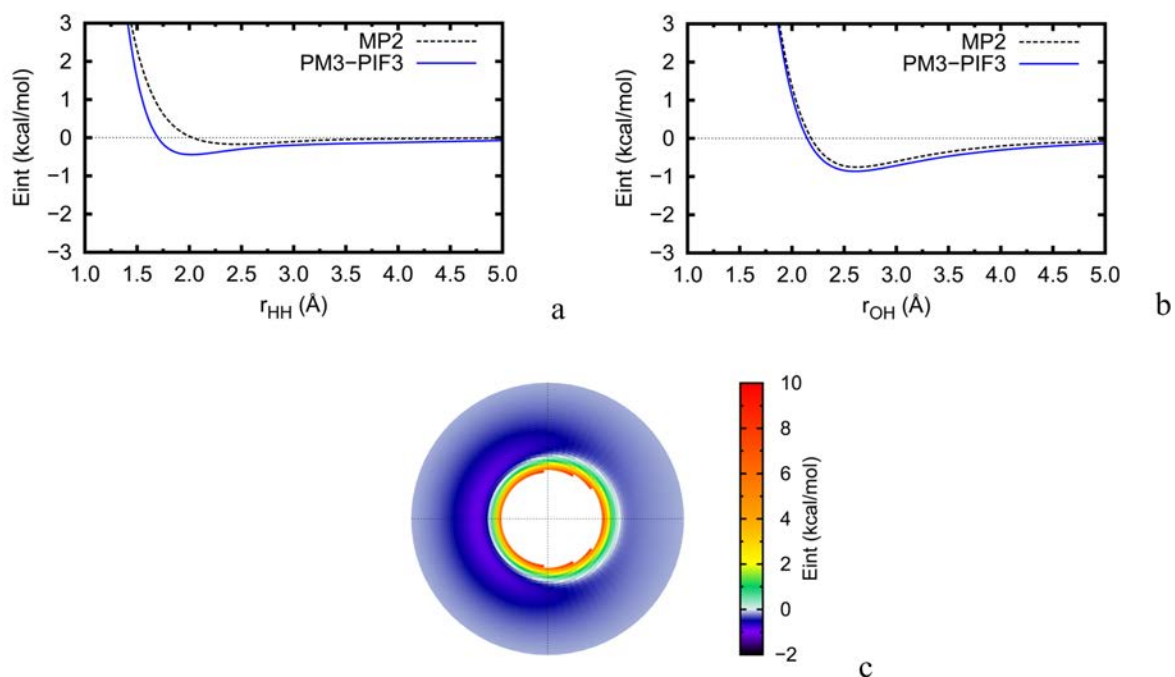


Figure 4.5: Same as in Figure 4.4, for the PM3-PIF3 Hamiltonian.

ous development of the PIF2 parameters. These energies have been used as a reference to perform the parameters optimization, carried out using the so called Levenberg-Marquardt algorithm[213, 214] as implemented in the `scipy.optimize.leastsq` python package.

Introducing a new atom type results in 10 parameters to be fitted, stemming from the PIF function, which is defined for atom (cores) pairs, to be compared to the corresponding 12 atomic parameters to be fitted for the Gaussian functions in the PM3 formalism. Beside the time required to perform the reference MP2 calculations, the optimization of parameters to fit such a large number of points can induce quite a large computational time, especially if an SCF procedure is required at each step of the optimization. However, some consideration about the PIF strategy can be made here to considerably fasten this process. As we detailed in Chapter 3, the PM3-PIF_x Hamiltonian differs from PM3 only for the core-core repulsion part. Thus, for a given geometry, the electronic part of the two Hamiltonians are strictly the same. In the present case, a PM3 calculation has been performed for all the structures present in the training set in order to evaluate the electronic part of the interaction energy of each complex. Finally, only the core-core repulsion part was included in the parameterization procedure. The python script developed for this purpose is available as Supplementary Material.

As we mentioned, the new parametrization is referred to as PM3-PIF3, and the resulting full set of parameters is summarized in Table 4.1. We shall now test this new methodology on the methane-water complex.

Figure 4.5 presents two unidimensional rigid scans of the interaction between methane and water using the new set of parameters. First of all, we note the excellent agreement ob-

tained for the $H_C \dots O$ interaction compared with the MP2 curve. A good agreement is also obtained in the case of the $H_C \dots H_X$ interaction. In the latter case, the position of the minimum is located at a slightly shorter distance but as far as this minimum is quite shallow, this difference should not dramatically affect the description of hydrophobic groups interacting with water molecules. Finally, Figure 4.5c displays the interaction energy surface as a function of the position and orientation of the water molecule. By comparison with Figure 4.2, it appears clear that the results obtained with this new set of parameters are convincingly close to the MP2 prediction.

4.4 Test computations

An interesting and more complex system for testing the interactions between methane and water is represented by clathrate hydrates, in which methane is a guest molecule incapsulated in a cage of water molecules, organized in a highly regular shape.[215, 216] One of the most abundant basic units in methane clathrate hydrates are formed by a pentagonal dodecahedron (referred to as 5¹²) of twenty water molecules surrounding one methane molecule located at the center of the cavity. We chose one of the isomers of this cage and optimized the clathrate complex by using the PM3-PIF3 Hamiltonian. The structure that we obtained is represented in Figure 4.6. This structure was characterized as a minimum by a vibrational normal mode analysis (using the program developed in this work, see Section 3.4). Some information on the effect of the water cage on the methane properties can be retrieved by comparing the shift induced on the C-H asymmetric stretch mode inside the cage compared to the value obtained for the isolated molecule. The calculated shift amounts to -31 cm^{-1} , to be compared with the experimental shift of about -20 cm^{-1} measured in a clathrate hydrate at 7K.[217] It is not possible to make a similar comparison for the vibrational properties of the water molecules in the cluster but a reasonably upper limit to the frequencies shift can be obtained by comparing experimental data for water in gas and liquid phases. In the gas phase, the water bend frequency is measured at 1594.7 cm^{-1} , whereas the symmetric and asymmetric stretch are at 3657.1 and 3755.9 cm^{-1} , respectively.[218] The experimental value of the shift for the water bend vibration is $+48 \text{ cm}^{-1}$. [218, 219] In the case of the stretch modes, in the liquid phase only one large band is observed between 3000 and 3800 cm^{-1} , the maximum of which is at 3404 cm^{-1} . The calculated shifts in the clathrate cluster with the PM3-PIF3 method fall in the range $-20 / +20 \text{ cm}^{-1}$ for the bend and about $-350 / -470 \text{ cm}^{-1}$ for the stretch, which seems consistent with the experimental data above. This is not the case for other methods such as PM6 or PM7, which lead to much larger shifts, such as for instance roughly 150 cm^{-1} for the bend frequency. It must be emphasized here, in addition, that PM6 and PM7 display extremely large errors in the calculation of absolute vibrational frequencies. The frequencies for the symmetric and asymmetric OH stretching modes of the water molecule in gas phase are, respectively, 3989.3 and 3868.6 cm^{-1} with PM3, 2613.4

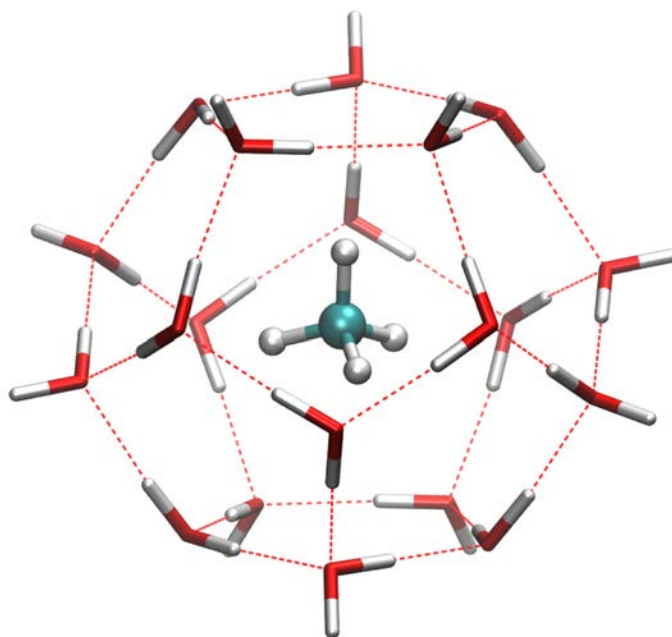


Figure 4.6: Optimized geometry of the cluster formed by methane in the 5^{12} cage of water molecules obtained using PM3-PIF3.

and 2526.5 cm^{-1} with PM6, 2856.5 and 2810.1 cm^{-1} with PM7, to be compared with the experimental values above. We note here that all the methods predict an inversion in the ordering between asymmetric and symmetric stretch. However, errors with PM6 and PM7 can therefore be larger than 1000 cm^{-1} , which precludes the use of these methods in vibrational spectroscopy studies.

Similar structures for methane in the 5^{12} cage have been treated in the literature[29, 220], in particular calculations at the DFT level using a B97-D functional[28] and a TZV(2d,2p) basis set[221] were performed and used to evaluate the binding energy of methane in the cage (ΔE) and the energy spent to deform the cage when the methane molecule is included, (ΔE_{def}). [29] Following the definition given in the latter work, we computed ΔE as the difference between the energy of the whole optimized system and the energy of the relaxed fragments. ΔE_{def} is defined as the difference between the energy of the water cage at the geometry of the complex (after removing CH_4) and the energy of the optimized 5^{12} cage. Regarding ΔE , we obtained a value of -6.22 kcal/mol for PM3-PIF3, to be compared with -6.90 kcal/mol for B97-D/TZV(2d,2p)[29] and -7.00 kcal/mol for MP2/6-31G*[220]. The value obtained for ΔE_{def} by using PM3-PIF3, 0.07 kcal/mol , is identical to the one stemming from calculations in Ref. [29]. We can safely conclude that the new parametrization leads to quite good results in the prediction of the stability of model clusters for the study of clathrate hydrates.

The next step is to verify the transferability of the PM3-PIF3 parameters, for which we carried out a series of tests. First, we performed calculations on 1:1 complexes between water and the following molecules: water, methane, methanol, ethanol, methanal, ethanal, butanone, formic acid, ethanoic acid, methylamine, dimethylamine, trimethylamine, for-

Table 4.2: Mean unsigned errors (MUE, in kcal/mol) for the interaction between water and small organic molecules at the MP2 minimum, and between water and methane throughout the PES (see text). The maximum error is also shown in parentheses for the calculations based on the set of 33 complexes.

Method	Water-molecule 1:1 complexes	Methane-water complex		
		<i>a</i>	<i>b</i>	<i>c</i>
PM3-PIF2	1.85 (4.27)	3.39	1.93	2.58
PM3-PIF3	1.35 (4.13)	0.47	0.15	0.29
AM1	4.21 (9.85)	0.37	0.49	0.44
AM1-DH+	2.57 (4.94)	0.64	0.80	0.73
PM3	4.18 (7.59)	1.79	1.62	1.69
PDDG/PM3	5.83 (11.90)	1.93	2.15	2.05
PM6	2.10 (4.57)	1.82	1.86	1.84
PM6-DH+	0.64 (1.58)	1.91	1.94	1.92
PM7	0.73 (1.87)	1.32	1.32	1.32

a H...H interaction

b O...H interaction.

c Both H...H and O...H interactions.

mamide and N-methylacetamide. Following the protocol described by Rablen *et al.*, [222] we performed a non exhaustive search of stable conformations for each molecule of our set in interaction with one water molecule at the MP2/6-311+G(d,p) [223] level. This strategy resulted in 33 fully optimized complexes (frequency calculations were performed at the same QM level). The geometries thus obtained are reported in the Supplementary Material. For each complex, the interaction energy between the two fragments was performed at the MP2 level with an aug-cc-pVTZ basis set, considering the geometry for the isolated fragments as the one obtained in the corresponding optimized complex. Corrections for the basis set superposition error (BSSE) were not included, to be consistent with the protocol adopted to parametrize the PM3-PIFx (see Ref. [132] for more details). For each of the 33 MP2-optimized geometries, we performed single point calculations to estimate the interaction energy using the semiempirical Hamiltonians considered in this work. As a first estimate of the performance of the different methods chosen to reproduce the MP2 standard, we calculated the Mean Unsigned Error (MUE) for the interaction energy, computed at the geometries corresponding to the respective MP2 minima. In the first column of Table 4.2, we report the results obtained for a choice of methods, the full set being presented in the Supplementary Material. For the sake of completeness, the other columns of Table 4.2 summarize the MUE obtained for the methane-water interactions, calculated as explained below.

Let us consider first the MUE of the 1:1 complexes interaction energies. Interestingly, PM3-PIF3 performs better compared to PM3-PIF2 (the MUE is about 0.5 kcal/mol smaller),

even though it was not explicitly reparametrized with the aim of improving hydrogen bond interactions. We note that the PM3-PIF3 value is about 0.6-0.7 kcal/mol larger than the values obtained for PM7 and PM6-DH+. Overall, the results confirm the improvement compared with the AM1 and PM3 methods. Similar conclusions can be drawn by considering the maximum error.

The structures of the complexes considered so far bring information on energy minima for hydrogen bonds but do not allow to assess the ability of methods to describe the interaction between water and H atoms belonging to hydrophobic saturated carbon atoms. To this aim, we shall first analyze the results obtained for the potential energy surface (PES) of the methane water complex. Columns 2-4 in Table 4.2 (see also Supplementary Material) contain the MUE from the different calculations. The calculated MUE values are based on the difference between the interaction energy curves obtained at the semiempirical level with respect to the MP2 standard (those reported in Figure 4.1 and in the Supplementary Material), by defining a relevant interaction window for each case. Considering the methane-water complex, in the case of the H...H curves we chose distances in the range between 1.2 and 2.5 Å, whereas for the O...H curves we considered distances between 1.5 and 3.5 Å. In the last column we report the total MUE for the H...H and O...H interactions.

The best result is now obtained with the PM3-PIF3 method, displaying a total MUE as small as 0.29 kcal/mol, representing a strong improvement over PM3-PIF2 (2.6 kcal/mol) and PM7 (1.32 kcal/mol). This has to be expected, since the new parameters were developed based on this specific complex. The MUE obtained with AM1 is also small, as expected from the discussion in Section 4.2.

To extend our tests to systems containing different types of C-H bonds displaying a hydrophobic character, we analyzed the interaction of water with ethylene and with benzene. In addition to the H...H and O...H interactions that were considered in the case of methane, here we also examine the interactions of one H atom of water with the π electron cloud of the C-C bonds in ethylene and benzene. Selected results for the MUE are reported in Table 4.3, while the full results and the different interaction energy curves are reported as Supplementary Material. The configurations and relative orientations of the two molecules in the complexes for the three different interactions are displayed in the Supplementary Material as well. The ranges chosen to calculate MUE from the PESs were the following: between 1.3 and 3.0 Å for H...H interactions, between 1.7 and 3.5 Å for H...O interactions and between 2.8 and 4.5 Å for the H... π interactions.

As a general remark, and as already observed for the O...H and H...H interactions in the methane-water complex, small errors are given by the AM1 method for the two systems considered, but the description of π interactions is very poor in this case. In the case of ethylene, the errors obtained for the PM3-PIF3 method are comparable to those found by using PM6 and PM7, with PM3-PIF3 giving a better description of the interaction of water with the π

Table 4.3: Mean unsigned errors (MUE, in kcal/mol) for the interaction between water and ethylene, benzene throughout the PES (see text).

Method	Ethylene		Benzene	
	a	b	a	b
PM3-PIF2	0.98	2.52	0.73	0.48
PM3-PIF3	0.82	0.99	1.15	0.76
AM1	0.35	2.02	0.46	3.04
AM1-DH+	0.51	1.00	0.44	1.25
PM3	1.30	1.55	1.21	2.48
PDDG/PM3	1.44	1.57	1.34	2.27
PM6	0.98	1.62	0.91	1.68
PM6-DH+	1.04	1.32	0.93	0.94
PM7	0.70	1.36	0.62	1.07

a Both H...H and O...H interactions.

b H... π interaction.

electron density of the C-C bond. As in the case of methane, PM3-PIF3 provides an important improvement over PM3-PIF2. In the case of benzene-water interactions, the comparison between PM3-PIF3 and PM7 methods shows that the former performs slightly better for H- π interactions while the latter performs slightly better for the “in-plane” (H...H and O...H) interactions. It can be also noted that in this particular case, PM3-PIF2 leads to smaller errors compared with PM3-PIF3.

Finally, tests and comparisons were done for the PES of the 1:1 water complexes with ethanol and trimethylamine. These complexes are particularly interesting because both ethanol and trimethylamine can form a hydrogen-bond with water and both bear one or more methyl groups. Starting from the MP2 minimum we define the axis joining the water H atom and the O atom of ethanol, in the first case, and the N atom of the amine, in the second case. The PESs are obtained by moving the two molecules along these axes. The results are reported in Figures 4.7a and b, respectively. As shown in Figure 4.7a, the PM3-PIF3 results for the ethanol-water complex are in remarkably good agreement with the MP2 data, even though no further *ad hoc* parametrization was introduced to obtain this curve. We also show that this agreement is not obtained when we do not use different parameters for hydrophobic (H_C) and hydrophilic (H_X) H atoms (PM3-PIF3*). Considering the minimum of the PM3-PIF3 PES, located at 1.88 Å, the hydrogen bond energy is predicted to be -6.13 kcal/mol, to compare with the MP2 value, -6.43 kcal/mol, corresponding to a distance of 1.90 Å. The PM7 curve presents a characteristic artifact of the GCFs close to the MP2 energy minimum. The PM7 interaction energy minimum is predicted to be at a shorter distance (1.72 Å) and 0.92

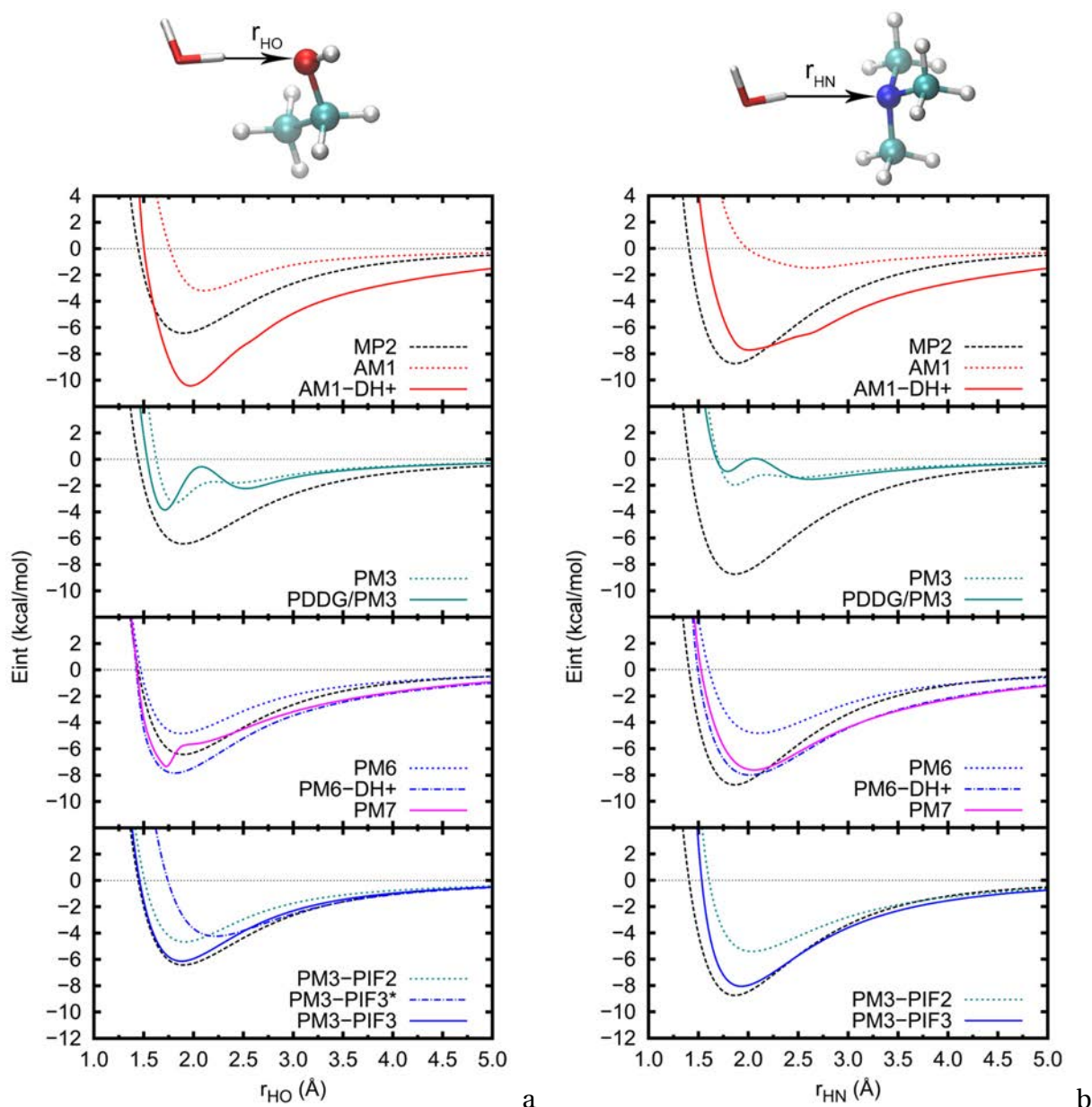


Figure 4.7: Interaction between the H atom of water and the O atom of the hydroxyl group of ethanol (a) as well as the N atom of tri-methylamine (b): performance of a choice of semiempirical methods compared to the MP2 reference (black dashed line, reported in each plot). In the case of the PM3-PIF3* curve, the parameters for all hydrogen atoms are those obtained for 'hydrophobic' H_C's (see text).

kcal/mol lower in energy compared to MP2. In the case of trimethylamine (see Figure 4.7b), PM3-PIF3 is also in excellent agreement with MP2 and it provides slightly better results than PM7. The PES minimum is located at 1.86 Å in the case of MP2, 1.94 Å for PM3-PIF and at 2.06 Å for PM7. The corresponding values for the interaction energy are -8.75 kcal/mol, -8.06 kcal/mol and -7.63 kcal/mol, respectively.

We can then conclude that the parameterization strategy used in the PM3-PIF3 method, *i.e.*, using PIF instead of GCFs in the core-core functions and using different parameters for H atoms linked to carbon or heteroatoms, is very promising. It provides significantly better results than any other semiempirical method in the case of water interactions with methyl groups, and good results for the interaction with methylene groups, as in the case of ethy-

lene. Further analysis concerning aromatic systems interacting with water, which were not included in the PIF3 parametrization, is necessary to assess whether the introduction of atom types should also be extended to different types of carbon atoms. By analyzing the shape of the interaction energy along the PESs, we showed that PM3-PIF3 avoids the presence of GCF artifacts, still present in PM7, though to a lower extent than in the preceding versions of the method PM6 and PM3.

4.5 Concluding remarks

In this Chapter we presented a study of how different semiempirical methods behave with respect to the description of the intermolecular interactions that have to be taken into account to be able to model water solutions. Different model systems were considered, in particular 1:1 complexes between water and small organic molecules. In these complexes, we analyzed the potential energy surfaces along two relevant coordinates, which we chose to describe attractive and repulsive interactions between water and molecules containing polar and/or apolar groups. The most recent semiempirical Hamiltonians show remarkable improvements compared to the description given by the originally developed methods, optimized to reproduce mostly gas phase properties. However, none of the available methods is able to give a correct physical representation of the interactions of water with both hydrophilic and hydrophobic groups.

Most of the methods that we analyzed treat the core-core interactions by using Gaussian correction functions. On the other hand, it was shown in our group that significant improvement in the description of intermolecular interactions can be achieved by modifying the analytical form of the functions to be used, and by reparametrizing them based on *ab initio* PESs. Notwithstanding these progresses, the parametrization provided so far (*i.e.*, PM3-PIF2) still gave inaccurate results for the interaction between water and hydrophobic methyl groups. We therefore went forward in a recalibration of this method that allows to increase the weight of such hydrophobic interactions in the optimization procedure. As in previous works, the reference PESs were obtained at the MP2/aug-cc-pVTZ level of theory. While working at the parametrization, we realized that considering the chemical environment of H atoms is a key point to introduce in order to be able to treat intermolecular interactions of water with polar and apolar groups in a balanced way. Hence, in this work we propose to introduce H...H and O...H core-core interaction parameters that are different based on the nature of the atom to which H is attached (carbon or a heteroatom) and the PM3-PIF3 parameters were developed according to this hypothesis. They were tested on a set of 1:1 complexes, for which we show that the results are very encouraging. However, it is worth mentioning that distinguishing between H atoms in a quantum chemical method introduces some practical limitations in its potential applications. In particular, at-

tention must be paid in the case of reactive systems. If a covalent bond involving an H atom is broken and a new bond is formed with an atom of a different type (*e.g.*, in the case of a keto-enol equilibrium) the parameters used for the reactants might not be consistent with the description of the products. We note that this issue also affects reactivity in several semiempirical methods described above, as well as in the general QM/MM schemes, since the van der Waals interactions between the QM and the MM regions depend on the atom type. We may also note that in the case of semiempirical methods, the optimized QM/MM van der Waals parameters are usually quite different compared to those of the MM force-field and in addition they depend on the method used to calculate the QM/MM electrostatic interactions.[224]

Our work provides a new strategy to improve the description of intermolecular interactions by semiempirical methods by keeping simple protocols to reparametrize existing Hamiltonians. Our preliminary results point to the possibility of achieving a good description of the interactions in systems immersed in aqueous solutions, to be exploited in condensed phase studies and for the molecular dynamics of biomolecules in their environment. Further extensions of this methodology will require including a larger number of heavy atoms in the parametrization set as well as more extensive analysis on water interactions with hydrophobic compounds. Such analysis should include condensed phase simulations to test the accuracy of the PM3-PIF3 method in predicting the solvation pattern of organic and bioorganic molecules in aqueous solution. This will be the focus of the next Chapter.

Chapter

5

Solvent effects on small biological compounds

Résumé

L'étude présentée dans ce chapitre constitue la première application de l'Hamiltonien PM3-PIF3 à la dynamique de molécules (bio-)organiques en phase aqueuse condensée. Nous cherchons ici à valider l'utilisation de cet Hamiltonien et de la méthode SEBOMD pour des études futures de larges systèmes biologiques. Dans ce but, nous discutons ici les résultats des simulations de onze systèmes moléculaires, présentant des groupes fonctionnels caractéristiques des systèmes biologiques. Pour chaque molécule, nous étudions de manière systématique la structure de sa sphère de solvation ainsi que l'effet du solvant sur ses propriétés électroniques et vibrationnelles. Chaque système en phase condensée est composé d'un soluté plongé dans une boîte de 128 molécules d'eau. Tout d'abord, nous nous focalisons sur une série de molécules à caractère hydrophobe. Puis, dans une deuxième partie, nous traitons de petites molécules présentant un caractère mixte, à la fois hydrophile et hydrophobe. Enfin, nous étudions une molécule comportant un plus grand nombre de degrés de liberté et se rapprochant d'un composé biologique, le dipeptide alanine.

Parmi les molécules hydrophobes sélectionnées ici, deux sont purement aliphatiques (le méthane et l'isobutane) et deux sont des composés aromatiques (le benzène et le toluène). Notamment, l'interaction entre la molécule de méthane et l'eau a servi de modèle pour le développement de la méthode PM3-PIF3 et l'analyse de la dynamique menée en phase condensée confirme les bons résultats de cet Hamiltonien. La position de la première sphère de solvation autour du méthane, prédite par PM3-PIF3, est en très bon accord avec d'autres modèles reportés dans la littérature. PM3-PIF3 corrige ainsi le problème de surestimation de la répulsion entre méthane et eau obtenu avec la méthode PM3-PIF2. Nous montrons également un effet du solvant sur la bande de vibration CH du méthane, cohérent avec les observations expérimentales, qui ne peut pas être reproduit avec un modèle de champs de force non-polarisable. Les observations relatives aux autres molécules hydrophobes sont similaires et en bon accord avec les quelques résultats rapportés dans la littérature.

La série de molécules hydrophiles a été choisie pour refléter certains groupements clés présent dans les acides aminés : deux molécules présentant un groupement hydroxyl (l'éthanol et le para-éthylphénol), une amine tertiaire (la tri-méthylamine) et trois amides (le for-

mamide, la propanamide et la N-méthylacetamide). Ces molécules présentent un caractère mixte, hydrophile/hydrophobe, et nous permettent de tester si l'Hamiltonien PM3-PIF3 est capable de traiter correctement ces deux types d'interactions. Nous montrons dans la majeure partie des cas étudiés ici, un bon accord avec d'autres travaux théoriques quant à la structure du solvant autour de ces molécules. L'étude des propriétés électroniques et vibrationnelles montre l'intérêt de traiter ces systèmes à un niveau de théorie quantique. Par exemple, l'élargissement et le décalage de la bande de vibration OH d'un alcool dus à la présence du solvant sont bien reproduits par la méthode PM3-PIF3, alors que cet effet n'apparaît pas en mécanique moléculaire. Les effets de solvant sur les bandes de vibrations caractéristiques des amides sont également bien reproduits par nos simulations.

Le dipeptide alanine est le modèle de polypeptide le plus simple. L'étude de ce système en phase condensée constitue le lien entre l'analyse de petites molécules (bio-)organiques et le traitement de larges systèmes biologiques. Ici encore, la structure des différentes couches de solvation autour du dipeptide alanine est en bon accord avec d'autres résultats théoriques. Bien qu'une analyse conformationnelle quantitative de ce peptide requiert des temps de simulation plus longs que ceux présentés ici (500 ps), les tendances obtenues montrent un effet clair de la conformation sur les propriétés électroniques de la molécule. L'analyse du spectre infrarouge du dipeptide en fonction de sa conformation tend vers des observations similaires à celles fournies par d'autres études théoriques. Cependant, le temps de simulation doit être augmenté afin de confirmer ces résultats préliminaires.

Enfin, l'étude présentée ici fournit des résultats très encourageants quant à l'application de la méthode SEBOMD et de l'Hamiltonien PM3-PIF3 pour étudier la dynamique de systèmes biologiques en solution. L'étape suivante consistera à traiter des systèmes de plus grande taille, tels que de petites protéines ou enzymes, dans leur environnement.

The semiempirical Born-Oppenheimer molecular dynamics (SEBOMD, see Chapter 3) methodology is an efficient tool that allows a good compromise between system size (*i.e.*, number of atoms and simulation time) and level of accuracy in the description of electronic effects.[169] Because of the low computational cost offered by semiempirical (SE) methods, the SEBOMD technique is suitable to simulate large biomolecular systems including their aqueous environment (*i.e.*, about 1000 atoms) along large simulation time scales (*i.e.*, up to a few tenths of a nanosecond). SEBOMD is based on NDDO SE approaches, therefore its accuracy is directly related to the choice of the Hamiltonian used to evaluate the potential energy surface of the target system. In this Chapter, we propose to validate the SEBOMD methodology and the PM3-PIF3 Hamiltonian by focusing on the simulation of model hydrated systems.

We have chosen ten organic molecules intended to reproduce some of the representative building groups encountered within biological systems (*i.e.*, aminoacid sidechains such as valine, serine or asparagine) and their interactions with water. To model the interactions of water with hydrophobic groups, we have studied methane, isobutane, benzene and toluene. Six additional compounds were chosen to model hydrophilic interactions in proteins: ethanol, p-ethylphenol, tri-methylamine, formamide, propanamide and N-methylacetamide. The last three molecules contain an amide site and will be used as models of biomolecules containing peptide bonds. Finally, to make a step further toward the simulation of polypeptides and proteins in water, we shall focus on the alanine dipeptide, which is the simplest and most studied model for conformational analysis.[225–227]

The following discussion will be based on the analysis of SEBOMD and molecular mechanics based molecular dynamics (MM-MD) simulations of the eleven compounds cited above dissolved in water and on the comparison of our results with existing theoretical and experimental data. To validate the use of the PM3-PIF3 Hamiltonian, we shall first focus on the solvent structure around the different solutes. We will also investigate the solvent effects on the dynamical features of each system. To this end, we shall compare the electronic and vibrational properties of the molecules when going from the gas phase to aqueous solutions. In the case of flexible molecules, we will also discuss the effect of water on the structural characteristics of the solute.

5.1 Computational details

The same protocol has been followed to prepare the topology and coordinate files of each system. For MM simulations, we used the ff03[38] force field for the solute and the SPC/E model to represent water molecules.[212] For the molecules that were not defined in the standard ff03 library, atomic charges were derived using the RESP procedure applied to calculations performed at the B3LYP/cc-pVTZ//HF/6-31+G** level with implicit solvent to re-

main consistent with the original development of the ff03 parameters.[38] A minimal Gaussian03 input is given as,

```
%chk=fname.chk
#P HF/6-31G** opt=Tight

resp calculation

0 1
. . .

--Link1--
%chk=fname.chk
#P B3LYP/cc-pVTZ SCRF(IEFPCM,Solvent=Ether)\
SCF=Tight Geom=AllCheck Guess=Read Pop=MK IOp(6/33=2)
```


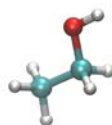
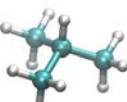
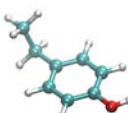
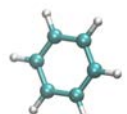
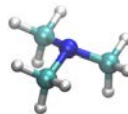
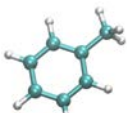
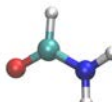
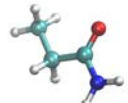
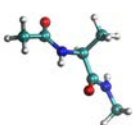
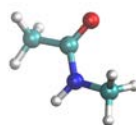
Condensed phase simulations were performed using periodic boundary conditions (PBC) and the Ewald[228–233] summation scheme to account for long range interactions. For each system, one solute molecule was placed in a cubic box of 128 water molecules. Molecular dynamics simulations were performed for each system using the same setups. In all cases, the time step was 1 fs and the Andersen thermostat was used to keep the system at 300K. Each box was first equilibrated at the MM level during 700 ps in the NPT ensemble with a strong temperature coupling every 50 fs (vrand=50) and the Berendsen barostat[50] to keep a constant pressure of 1 bar. During these NPT simulations, the bond lengths of each molecule (*i.e.*, including the solute) were frozen using the SHAKE algorithm.[47] The size of the box was determined from the average volume of the last half of the simulation after the system was equilibrated. This procedure led to the box sizes and densities presented Table 5.1. An input file to perform such a simulation is given by:

```
NPT MM simulation
&cntrl
  imin=0, ntx=1, irest=0
  ntc=3, dt=0.001, nstlim=700000,
  cut=4.2, ntb=2, ntp=1
  ntt=2, vrand=50, temp0=300.
  ntp=100, ntwx=100, ntwv=100, ntwe=100
/
&ewald
  ew_type=1
/
```

where `ntp = 1` and `ntc = 3` respectively turn on the Berendsen barostat and the SHAKE algorithm. The `ew_type=1` keyword ensures the use of the standard Ewald calculation rather than the approximated particle mesh Ewald (PME) method.

From this pre-equilibrated trajectory, the same procedure, for both equilibration and production runs, was used for MM-MD and SEBOMD simulations. Each of the selected solutes was simulated in the gas and in the condensed phase within the NVT ensemble. The same equilibration protocol was followed in all cases to smoothly increase the temperature

Table 5.1: Description of the 11 studied systems. Each cubic box is composed by one solute and 128 water molecules. For molecular mechanics simulations, the water molecules are treated using the SPCE/E force field. The box size of each system was determined after an ff03-SPC/E NPT simulation (see text). The origin of the charges used for the solute in the MM simulation is also provided.

Hydrophobic			Hydrophilic		
Solute name	Box size (Å)	Charges	Solute name	Box size (Å)	Charges
Structure	Density (kg/m ³)		Structure	Density (kg/m ³)	
Methane			Ethanol		
	15.69 1.00	RESP ^a		15.70 1.01	RESP ^a
Isobutane			p-ethylphenol		
	15.77 1.00	RESP ^a		15.83 1.02	RESP ^a
Benzene			Tri-methylamine		
	15.75 1.01	RESP ^a		15.74 1.01	RESP ^a
Toluene			Formamide		
	15.80 1.01	RESP ^a		15.66 1.02	RESP ^a
Peptide model			Propanamide		
				15.73 1.02	RESP ^a
Alanine dipeptide			N-methylacetamide		
	15.83 1.03	ff03 ^b		15.74 1.01	ff03 ^b

^a RESP charge derivation from B3LYP/cc-pVTZ//HF/6-31+G**

^b charges from Amber ff03[38]

of the system up to 300K:

1. 5 ps at 10K, with temperature coupling every 10 fs
2. 5 ps at 100K, with temperature coupling every 10 fs
3. 10 ps at 300K, with temperature coupling every 50 fs
4. 10 ps at 300K, with temperature coupling every 1 ps

Then, the production runs were performed for 500 ps at 300K with a velocity randomization every 1ps (default in Amber14). An example of production run input for an MM simulation

in condensed phase is given as:

```
NVT MM simulation
&cntrl
  imin=0, ntx=5, irest=1,
  ntc=1, dt=0.001, nstlim=10000,
  cut=4.2,
  ntb=1, ntp=0,
  ntt=2, vrand=1000, temp0=300., tempi=300.,
  iwrap = 1,
  ntp=1, ntwx=1, ntwv=1, ntwe=1,
/
&ewald
  ew_type=1,
/
```

In SEBOMD simulations, the PM3-PIF3 Hamiltonian was used for each system, though in some test cases other Hamiltonians such as PM3-PIF2 and PM3 were used to provide a comparison with the newly developed PM3-PIF3 method. The MM correction to the PM3 Hamiltonian for peptidic bonds was applied when necessary (*i.e.*, for formamide, propanamide, N-methylacetamide and alanine dipeptide). In Chapter 3, we provided more details about the implementation of this correction available in Amber14. A typical input for a PM3-PIF3 SEBOMD simulation in Amber14 is written as follows:

```
PM3-PIF3 SEBOMD
&cntrl
  . . .
  ifqnt=1
/
&qmmm
  qm_theory=SEBOMD
/
&sebomd
  hamiltonian=PM3
  modif=PIF3
  peptcorr=1
  method=2, ipolyn=1
  screen=1, ntwc=1
  longrange=2
  solute=1
  nsol=1
/
```

where the `longrange` keyword is set to 2 in order to use Mulliken charges in the Ewald[193] summation according to the discussion in Section 3.5.

A systematic analysis on each SEBOMD and MM-MD simulation was performed by first focusing on the solvent structure during the condensed phase simulations. In addition, for both the gas phase and the condensed phase simulations, we analyzed the instantaneous electronic and the vibrational properties.

5.2 Hydrophobic compounds

SEBOMD simulations are expected to be carried out on large biological systems in which the solvent plays a major role (*e.g.*, protein folding and reactivity).[234] Thus, the first investigation to be done is to verify the ability of our methodology to predict the solvation of biologically relevant solutes. Molecules containing hydrophobic groups are part of those key compounds. As we discussed in Chapter 1, radial pair distribution functions (RDFs) are widely used in molecular dynamics simulation to investigate the structure of the solvent around a given atom. In what follows, we will first discuss the RDFs obtained from our SEBOMD and MM-MD simulations.

5.2.1 Solvent structure

For each of the hydrophobic molecules selected in this work, we will focus only on specific and relevant pair distribution functions between one atom of the solute and one atom of the solvent. However, all RDFs are available as Supplementary Material. To clarify the discussion, we will add a “w” after each solvent atom. Thus the hydrogen atoms and oxygen atoms of water will be referred as “Hw” and “Ow”, respectively. A specific nomenclature for solute atoms will be introduced if needed.

Methane. As we have shown in Chapter 4, PM3 and PM3-PIF2 present deficiencies to reproduce the interaction between methane and water hydrogen atoms (HHw). We recall that PM3 predicts a minimum for this interaction of about ~ 2.2 kcal/mol for a HHw distance of ~ 1.7 Å. The use of PIF2 corrects this issue but overestimates the HHw repulsion. This has motivated the development the PIF3 correction and a comparison of the methane solvation shell obtained from PM3, PM3-PIF2 and PM3-PIF3 SEBOMD simulations appears to be particularly interesting. Indeed, it is not obvious that the problems pointed out from static calculations will also be present in dynamics. To analyze this issue, the COw and HHw RDFs are particularly relevant. The former will give insights about the position of the solvation shell around methane while the latter will highlight possible HHw interaction artifacts. Figures 5.1a and b present respectively the COw and HHw RDFs obtained from SEBOMD and MM-MD simulations. The integration of each RDF is also reported on those plots. In the case of the COw RDFs, we show in Table 5.2 the values of the position and height of the first maximum and minimum given by each method, as well as the coordination number (N_w) integrated up to the first minimum (see Section 1.5 for more details). Since no experimental results exist for the hydration of methane in liquid water, we also report data from similar theoretical studies[235, 236] performed with different methods as a comparison.

Let us first discuss the results coming from the literature. Rossato *et al.* performed Car-Parinello MD (CPMD) simulations of methane in a box of 64 water molecules using the PBE DFT functional.[236] We remind that the temperature used in CPMD is higher than the room

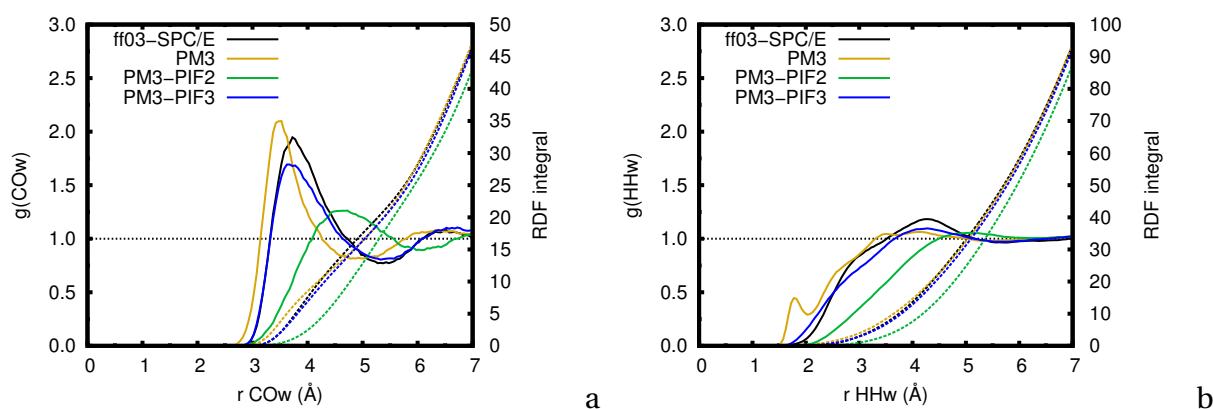


Figure 5.1: Comparison of radial pair distribution functions of methane in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (Ow) with respect to the methane carbon atom (C). b: Pair distribution of water hydrogen atoms (Hw) with respect to the methane hydrogen atoms (H).

temperature, to compensate some deficiencies of the method and obtain the correct behavior of liquid water. Depending on the temperature, the authors report a position of the first COW RDF peak at about ~ 3.76 Å and an averaged coordination number (N_w) of 21.0 water molecules. MD simulations using a classical force field have been carried out with a larger water box by Mateus *et al.* leading to similar results,[235] *i.e.*, a first maximum at 3.65 Å and $N_w=19.7$ water molecules. The two CPMD simulations of Rossato *et al.* and the MM-MD of Mateus and coworkers also agree in having the position of the first minimum consistent with the results of N_w . The MM simulation performed in the present work leads to very similar results. The position of the first peak is found at 3.73 Å and 21.0 water molecules compose the first solvation shell around methane.

The three SEBOMD simulations give different results depending on the Hamiltonian used. A qualitative look at the COW RDFs Figure 5.1a clearly shows that the position and shape of the first solvation shell are predicted differently by the three SE methods. The first maximum in PM3 is found at a slightly shorter distance compared to the MM simulation using the ff03 force field (-0.25 Å) and the peak appears to be sharper, leading to a smaller coordination number ($N_w=15.7$). To the opposite, the PM3-PIF2 Hamiltonian leads to a large displacement of the first solvation shell ($+0.92$ Å compared to ff03). Consequently, the width of the first peak is broadened and a larger number of molecules is found in the first solvation shell ($N_w=27.5$). The PIF3 correction leads to a COW RDF very close to ff03. The position of the first peak and the coordination number (3.71 Å and 20.3 molecules, respectively) confirm this agreement and is also comparable with other results in the literature.

An inspection of the HHw RDFs on Figure 5.1b leads to similar observations concerning PM3-PIF2 and PM3-PIF3 results. PM3-PIF2 predicts a too long HHw distance compared to PM3-PIF3 and the other theoretical methods. The RDF obtained with PM3 shows an additional well defined peak centered at 1.80 Å. This peak is related with a direct interaction between methane and water hydrogen atoms. As predicted by the static calculations (see

Table 5.2: First maximum and first minimum positions in Å (height in parenthesis) of the COw RDF of methane and the corresponding coordination number (N_w). Comparison of different Hamiltonians used in this work and some results from the literature.

System	first max	first min	N_w
128w, ff03 SPC/E (300K) ^a	3.73 (1.94)	5.38 (0.78)	21.0
128w, PM3 (300K) ^a	3.48 (2.10)	4.94 (0.82)	15.7
128w, PM3-PIF2 (300K) ^a	4.65 (1.26)	6.11 (0.89)	27.5
128w, PM3-PIF3 (300K) ^a	3.71 (1.69)	5.39 (0.81)	20.3
64w, CPMD/PBE (400K) ^b	3.79 (2.59)	5.28 (0.52)	19.8
64w, CPMD/PBE (460K) ^b	3.72 (1.91)	5.55 (0.75)	22.2
256w, OPLS-AA TIP3P (298K) ^c	3.65 (1.84)	5.35 (0.82)	19.7

^a This work, ^b Reference [236], ^c Reference [235]

Chapter 4), the artifact present on the PM3 interaction energy surface appears to play an important role in condensed phase. Even though this interaction is weak compared to water water interactions, it modifies the shape of the first solvation shell of methane.

The results obtained for the methane water system show the importance of avoiding strong artifacts on the interaction energy surface. Indeed, the problem pointed out for PM3 (see Chapter 4) also exists in other SE methods such as PM6 and its derivatives, making them bad candidates for solution studies involving similar interactions. This analysis also confirms the great improvement of PIF2 provided by PIF3 for this particular system. Such results are very promising for further applications to alkanes and compounds containing alkyl groups.

Isobutane. The isobutane water system implies the same type of interactions as in the case of methane. We can thus expect some similarities in the water organization around those two molecules, though isobutane is larger than methane. Figure 5.2a and b present the RDF between the isobutane central carbon atom (Cc) and Ow as well as the one between the methyl carbon atom (Cme) and Ow, obtained with the PM3-PIF3 and ff03-SPC/E methods. Ashbaugh *et al.* developed an *ad hoc* united atom potential to simulate the hydration of alkanes, HH-Alkane.[237] The parameters of this potential have been fitted to reproduce experimental data. The authors applied it on several alkane molecules in a box of 400 TIP4P/2005 water molecules and showed a very good agreement with experimental results for thermodynamical properties such as solute partial molar volume (PMV). Their results for isobutane are also reported Figure 5.2.

At first glance, one can see that all the three models give qualitatively similar results for both pair distribution functions presented Figure 5.2. The position of the first peak in the CcOw RDF is found at 4.64, 4.79 and 4.81 Å for HH-Alkane, ff03 and PM3-PIF3, respectively.

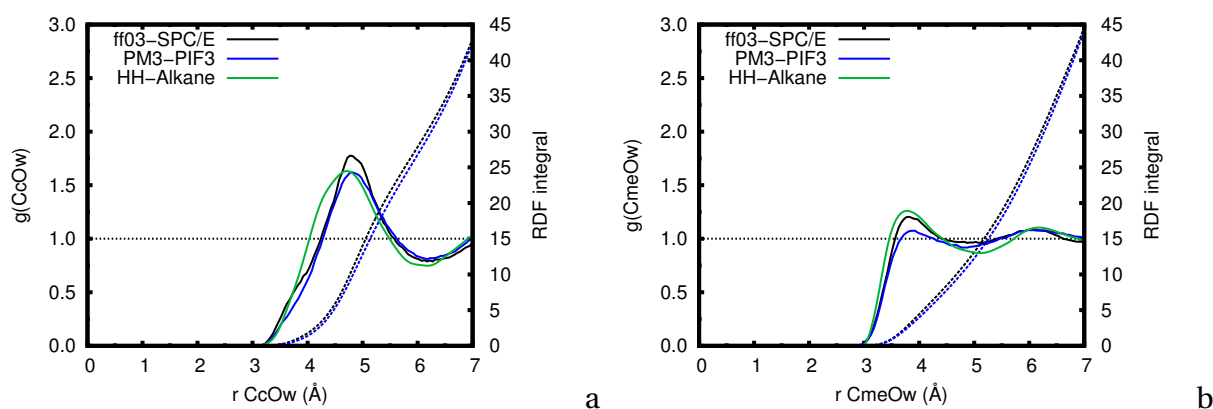


Figure 5.2: Comparison of radial pair distribution functions of isobutane simulated with ff03-SPC/E, PM3-PIF3 and the HH-Alkane united atoms potential.[237] Plain lines: RDF. Dashed lines: RDF integral. a: Isobutane center carbon atom (Cc)-water oxygen RDF b: Isobutane methyl carbon atom (Cme) Ow RDF.

The positions of the first peak for the CmeOw RDF are also very close with all three models (*i.e.*, 3.78, 3.81 and 3.89 Å for HH-Alkane, ff03 and PM3-PIF3, respectively). The height of the peaks slightly varies from one model to another. Concerning the CcOw RDF, we notice that HH-Alkane predicts a larger width for the first peak compared to ff03 and PM3-PIF3. However, HH-Alkane is a united atom model and thus, the precise shape of the RDF peaks can be affected by this approximation. Since it has been shown that the PMV and RDF are closely related, according to the Kirkwood-Buff theory,[238] the good agreement found here with HH-Alkane allows us to believe that the structure of water around isobutane is correctly predicted by the PM3-PIF3 Hamiltonian.

Benzene. Understanding the structure of water around benzene and in general aromatic compounds has been the focus of many theoretical and experimental works during the past decades.[69, 239–245] Although it is now clear and commonly accepted that a hydrogen bond exists between the water hydrogen atoms and the π -electrons cloud of benzene,[239, 240] interpreting the shape of the solvation shell around benzene is still controversial. A few studies addressed this issue by using molecular dynamics or Monte-Carlo simulations.[69, 242–245] The specific π Hw interaction is reproduced by all these works, but the shape of the related solvation shell strongly depends on the model. In the following, we shall discuss those differences while presenting our own results.

Figures 5.3 a and b show, respectively, the RDFs for the π Hw and π Ow pairs in the simulations of benzene solvated by 128 water molecules simulations. Despite the difference in the position of the second peak for the three models considered here, it is noteworthy that ff03-SPC/E, PM3-PIF2 and PM3-PIF3 predict a non zero density of hydrogen and oxygen atoms at about 2 and 3 Å from the benzene center of mass, respectively. While ff03-SPC/E shows a well structured peak, the QM models predict a more broadened peak, and in the case of PM3-PIF3, the peak is so broad that the minimum before the second peak is not very pronounced.

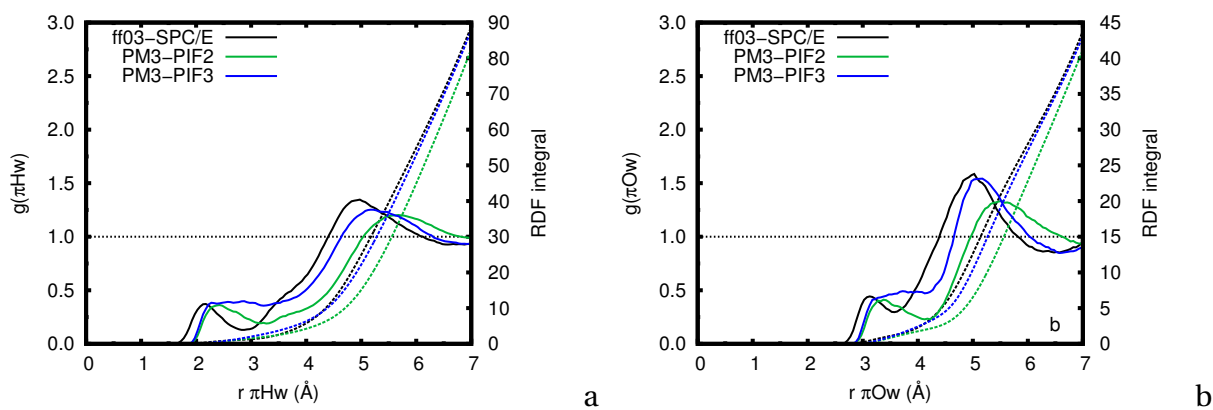


Figure 5.3: Comparison of radial pair distribution functions of benzene simulated with ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Plain lines: RDF Dashed lines: RDF integral. a: Benzene center of mass (π)-water hydrogen RDF. b: Benzene center of mass-water oxygen RDF.

Similar observations have been reported in the literature. Mateus *et al.* performed MD simulation of benzene in a box of 256 water molecules with PBC, using either a point charges or polarizable force field, *i.e.*, OPLS-AA/TIP3P or AMOEBA.[244] The authors compared the structure of the first peak representing the π Hw interaction using those two models. They found that the point charges model predicts a well structured peak (as ff03-SPC/E in the present work) while the polarizable force field shows a shoulder, similarly to the PM3-PIF3 results.

Allesch and coworkers also focused on this issue by means of Car-Parrinello MD simulations (CPMD).[69, 243] Simulations were performed for a system composed by one benzene molecule in a box of 73 water molecules with periodic boundary conditions. They considered two levels of approximations by using a flexible or a rigid description of water molecules, the latter allowing a larger sampling by increasing the simulation time step. However, the use of a rigid water model prevents the solvent molecules geometry to respond to the polarization induced by the solute. Their two models (*i.e.*, either with flexible or rigid water) both show a well defined first peak for the π Hw interaction but the intensity depends on the approximation used to describe water molecules. Indeed, compared to the rigid water model, the intensity of this peak is divided by two when the internal coordinates of the water molecules are relaxed. It should be noticed that the temperature used in these simulations was about 300K. It is commonly admitted that the liquid behavior of water using CPMD is obtained with higher temperature (*i.e.*, approximately 400K). This fact might affect the water dynamics and can lead to an over-structuration of the solvation shell. It appears from the work of Mateus *et al.* and Allesch and coworkers that the polarizability of water plays a key role in predicting the solvation structure around benzene.

To get more insights into the hydration shell, it is worth looking at the spatial distribution functions (SDFs, see Chapter 1), since RDFs do not contain any straightforward information about the orientational properties. Although it is possible to split the RDF of benzene into different regions (*i.e.*, axial and equatorial),[69, 243] SDFs give a global understanding of the

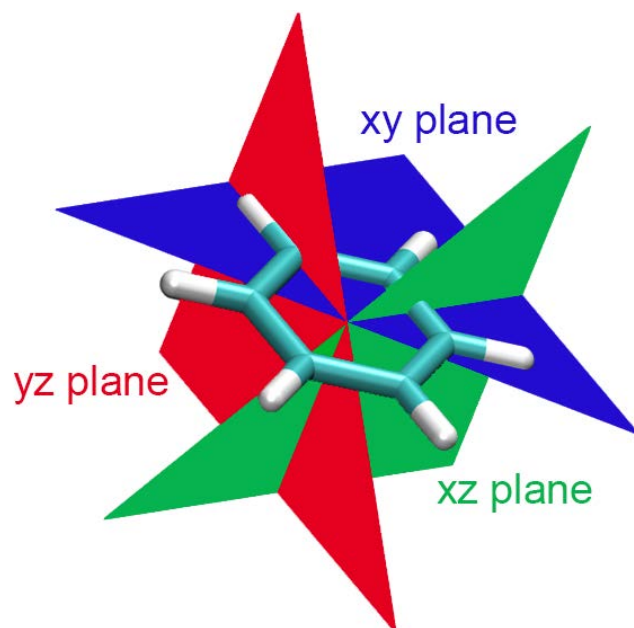


Figure 5.4: Definition of the three planes used to produce volume slices of the spatial distribution functions (SDFs) of water around benzene.

solvent structure, by computing the probability over the simulation time of finding a given atom in a unit volume of the MD box. SDFs have been computed over the three 500 ps MD simulations (*i.e.*, ff03-SPC/E, PM3-PIF2 and PM3-PIF3). To increase the sampling, we took advantage of some chosen symmetry elements of benzene. In Figure 5.4, we define three planes of interest for this system. The xy plane is the plane of the benzene ring, while the xz and yz planes are orthogonal to the aromatic ring. The xz plane intersects two opposite CC aromatic bonds and the yz one contains two opposite carbon atoms. We report SDFs slices according to those three planes in Figures 5.5 and 5.6 for the distribution of hydrogen and oxygen water atoms, respectively. To help the analysis, we also report on the SDFs, the position of hydrogen and carbon atoms for the xy and yz planes as well as the CC aromatic bonds for the xz plane. The SDFs presented here have been normalized with respect to an ideal distribution of water molecules with a density of 1.0 g/cm^3 . The regions corresponding to this ideal distribution are represented in white (*i.e.*, around 1 on the color scale). Gray regions correspond to a depletion in the water density and black ones to the solvent inaccessible part of space. Around 2 on the color scale, the blue regions are related to the different solvation shells and the rest of the scale, from 3 to 10 (*i.e.*, from green to red), shows the increasing strength of the intermolecular interaction.

We shall discuss first the results for the Hw and Ow distributions in the xy plane. The Hw distribution varies from one model to another (see the first column of Figure 5.5). ff03-SPC/E shows a hexagonal first solvation shell (blue region) which follows the benzene geometry. This shell is located between 4.5 and 5.5 Å, which corresponds to the second peak in Figure 5.3. Similar results have been observed by Allesch *et al.* from CPMD simulations.[69] The

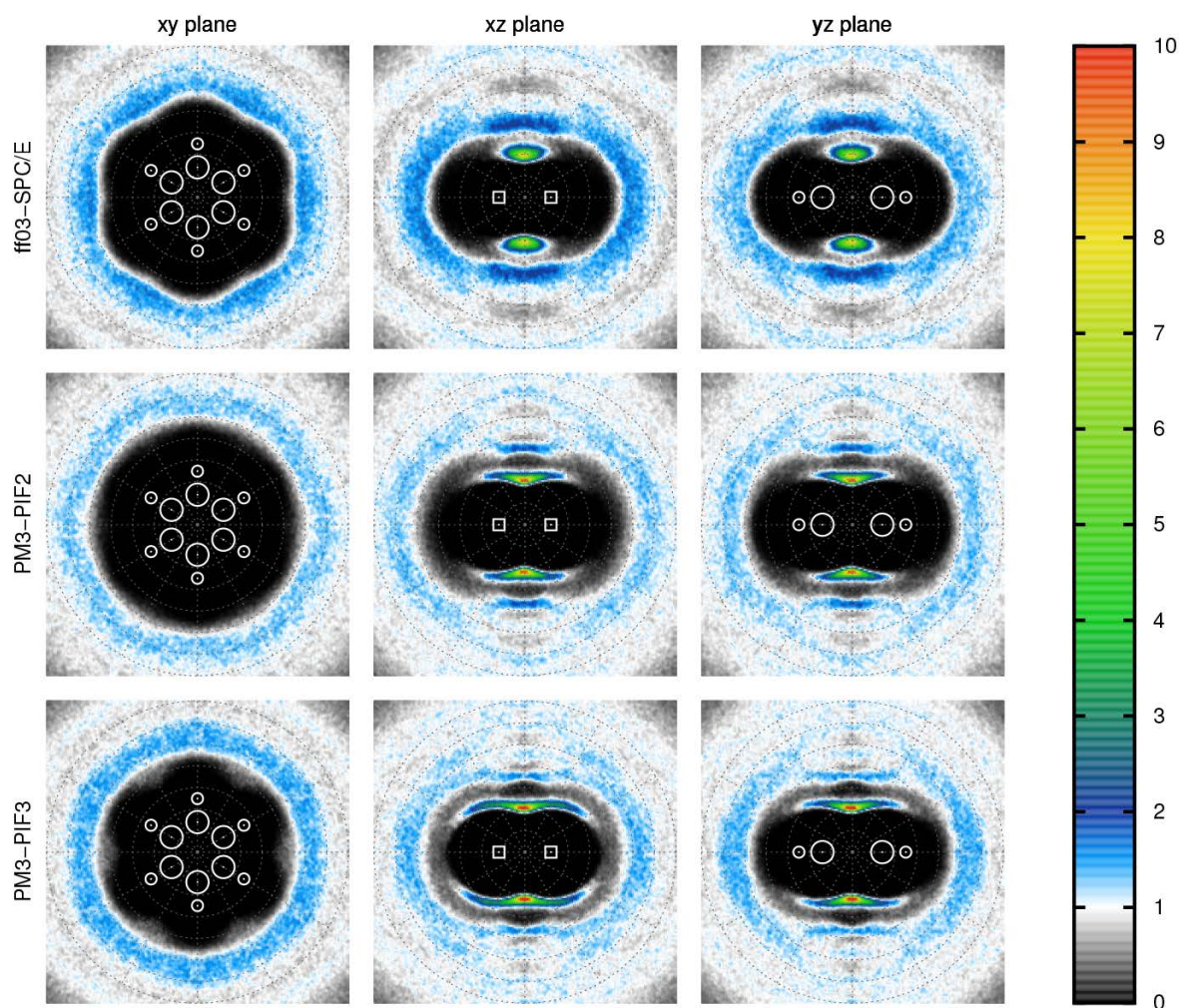


Figure 5.5: Volume slices of water hydrogen atoms (Hw) spatial distribution function (SDF) around benzene for the condensed phases dynamics using ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Large and small white circles are respectively the projections of benzene carbon and hydrogen atoms in the xy and yz planes. Squares correspond to the projection of the CC aromatic bonds in the xz plane.

PM3-PIF2 and PM3-PIF3 shells are more circular and their position is also in good agreement with the representation given by the RDFs in Figure 5.3. It is noteworthy that, using the PM3-PIF3 Hamiltonian, the water hydrogen atoms have a non zero probability to be found at shorter distances when facing the CC aromatic bonds, giving a general shape of the hydration structure similar to the one of ff03-SPC/E. The same observations can be drawn for the Ow distributions (Figure 5.6). ff03-SPC/E and PM3-PIF3 Ow SDFs are similar and show two characteristic distances of the first solvation shell to the center of the aromatic ring. This distance is shorter when the value of the angle is $\pi/3$ modulo $\pi/3$ (*i.e.*, facing a CC bond) than it is for angles equal to $\pi/6$ modulo $\pi/3$ (*i.e.*, facing a hydrogen atom). However, this difference is less strong in the case of PM3-PIF3. The results obtained with the PM3-PIF2 Hamiltonian give a different picture, with a minimum distance between the benzene hydrogen atoms and Ow of about 2 Å. This indicates that, in the case of PM3-PIF2, the interaction is dominated by the attraction between the hydrogen atom of benzene and Ow. This last observation is in agreement with the results discussed Chapter 4.

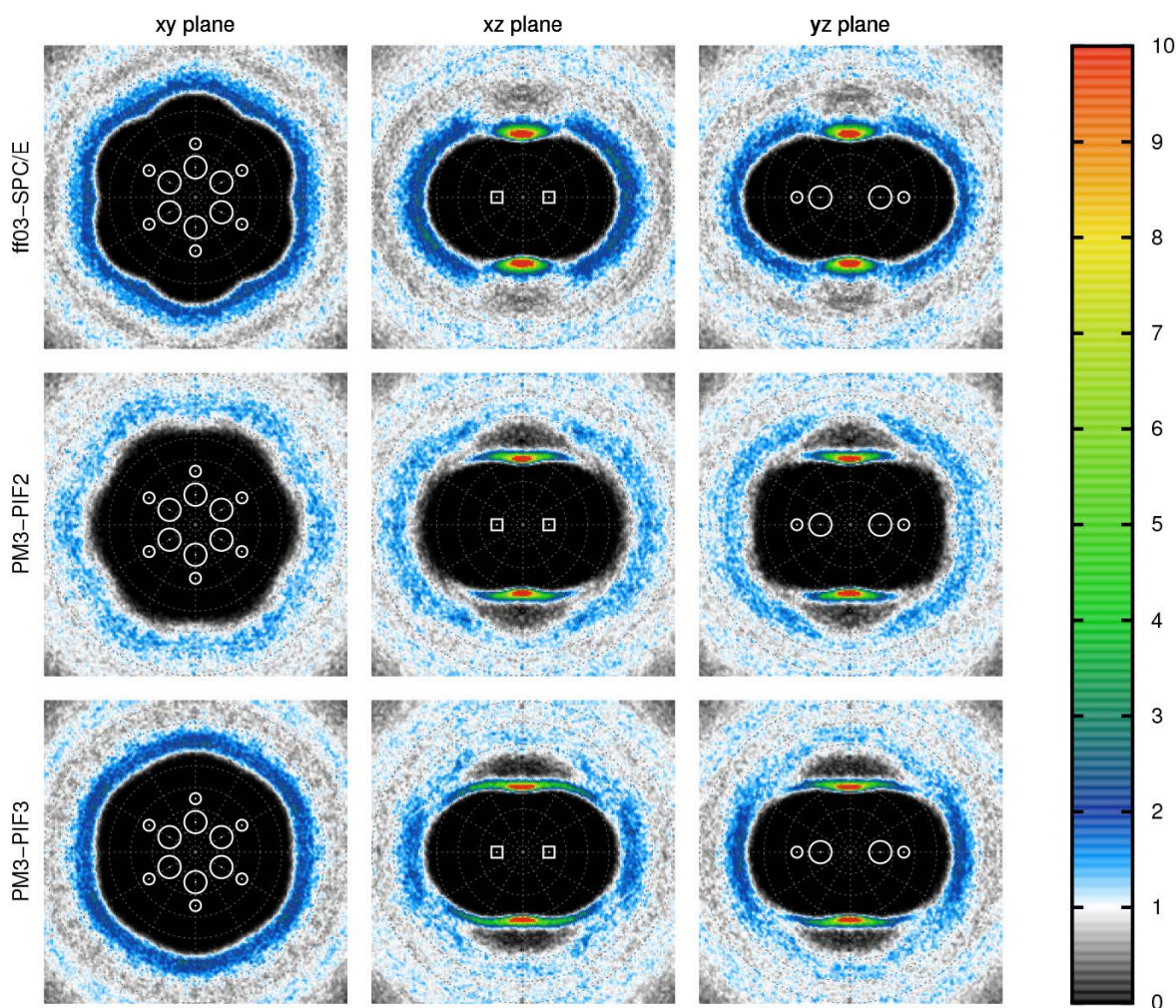


Figure 5.6: Volume slices of water oxygen atoms (Ow) spatial distribution function (SDF) around benzene for the condensed phases dynamics using ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Large and small white circles are respectively the projections of benzene carbon and hydrogen atoms in the xy and yz planes. Squares correspond to the projection of the CC aromatic bonds in the xz plane.

The results for the xz and yz planes can be discussed together. From Figures 5.5 and 5.6, one can see that all three methods considered here lead to a high probability density of both Hw and Ow atoms on the top of the aromatic ring (axial region). While the area occupied by Hw and Ow is quite localized and does not extend to an angle larger than $\pi/3$ with ff03-SPC/E, it is much more delocalized using the two QM models. Indeed, this region occupies an angle slightly larger than $\pi/3$ with PM3-PIF2 and reaches almost $2\pi/3$ with PM3-PIF3. This last observation explains the differences in the shape of the first peaks appearing on Figure 5.3. It is also interesting to notice that with the QM approaches, the π Hw interaction induces a structure of the next solvation shells in the axial region. Indeed, one can see four “layers” of Hw and two of Ow in this region. This phenomenon is not observed with the point charges force field (ff03-SPC/E). This last comment confirms the importance of using a polarizable model to describe such a system. What is observed for the Hw and the Ow distributions in the equatorial region is consistent with the discussion about the xy plane. In the case of ff03-SPC/E and PM3-PIF3, the averaged positions of Hw and Ow are closer

to the benzene molecules in the xz planes than they are in the yz plane for steric reasons. The Hw distribution in the equatorial region obtained using PM3-PIF2 is almost identical for the xz and yz planes due to the circular distribution in the xy plane. In the latter case, since the Ow distribution in the xy plane is tilted compared to the other methods, water oxygen atoms appear to be closer to benzene in the yz plane than in the xz plane, which is the opposite behavior compared to ff03-SPC/E and PM3-PIF3. Finally, it is interesting to notice the rectangular shape of the Ow distribution in the yz plane when using the PM3-PIF2 Hamiltonian. This shape is very different from the one obtained by the other methods and the results reported in the literature, which appears to be more spheroidal.[69, 240]

The description of the benzene hydration shell by our SEBOMD simulations appears to be in relatively good agreement with other works. However, the results for the Ow distribution in the equatorial region obtained with PM3-PIF2 are in disagreement with all other reported results. On the other hand, the use of PM3-PIF3 seems thus to be reasonable for such system. Nevertheless, the slight overestimation of the CHw interaction with this Hamiltonian (see discussion in Chapter 4) can be seen on the SDFs and might require further tests in future works.

Toluene. This compound shares similarities with methane, isobutane and benzene and thus implies the same type of interactions with water. To our knowledge, no MD simulation of toluene in aqueous solution devoted to the solute hydration structure has been reported in the literature. Thus, only a qualitative comparison between ff03-SPC/E force field and the PM3-PIF3 Hamiltonian will be discussed here.

We present in Figure 5.7a and b the RDFs of toluene for the pairs π Hw and CmeOw respectively. The labels are defined as for the previous compounds. In Figure 5.7a, we observe the same qualitative trend as in the case of benzene. The ff03-SPC/E force field predicts a well structured first peak for the π Hw interaction while the results obtained using PM3-PIF3 show a broadened shoulder for the same interaction. In a similar manner, the RDF for the

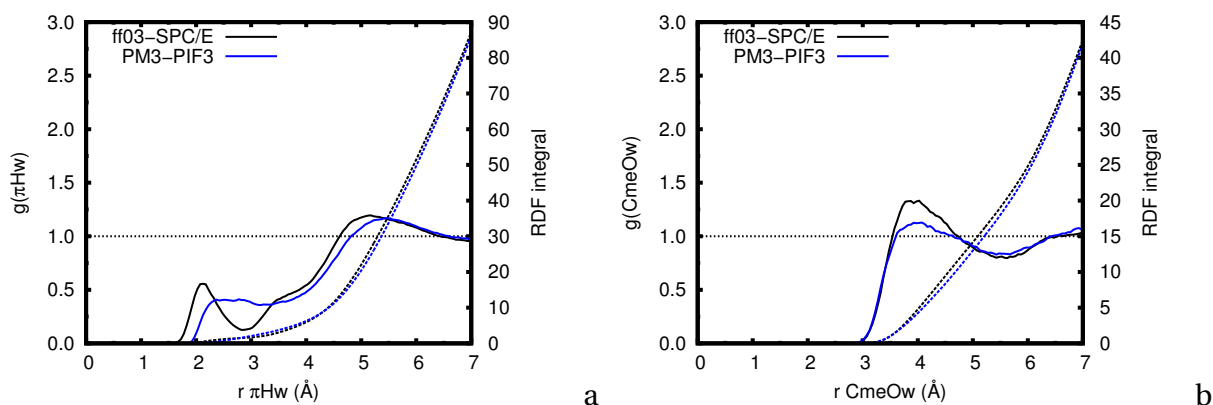


Figure 5.7: Comparison of radial pair distribution functions of toluene simulated with ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Plain lines: RDF, dashed lines: RDF integral. a: Toluene geometric center of the phenyl ring (π)-water hydrogen RDF. b: Toluene methyl carbon atom (Cme)-water oxygen RDF.

CmeOw pair is close to those obtained for methane and isobutane. In this case, the MM and QM methods both predict the same position for the peak corresponding to the first solvation shell.

Finally, we conclude that, in this first analysis of the solvent structure around hydrophobic compounds, SEBOMD simulations with the PM3-PIF3 Hamiltonian provide results in good agreement with other works. We shall now discuss the effect of the solvent on the solutes electronic properties along the dynamics.

5.2.2 Electronic properties

From our SEBOMD simulations, we can obtain the atomic partial charges with three different models, *i.e.*, Mulliken, CM1 and CM2. We note that these three models give different results for hydrophilic molecules but, in the case of hydrophobic compounds, the atomic charges are almost identical. This stems from the definition of these models that apply a correction to Mulliken charges, which is either null or very small for carbon and hydrogen atoms in CM1 and CM2, respectively. However, in order to remain consistent with the discussion about hydrophilic molecules in the next Section, only the results using CM1 will be considered in what follows. The total CM1 charge of each solute has been followed along the condensed phase SEBOMD simulations. The solute dipole moment of each solute has been computed along both gas phase and condensed phase simulations from atomic partial charges and atomic positions. The definition of dipole moment is not valid for charged compounds since it varies depending on the choice of the frame. Thus, all dipole moments have been computed with respect to the center of mass of the targeted molecule.

Figures 5.8a shows the distribution of the solute CM1 total charge along the PM3-PIF3 dynamics in percent of elementary charge (%e). One can first notice that the charge of the solute is not a fixed property and varies during the dynamics. The total charge of each solute

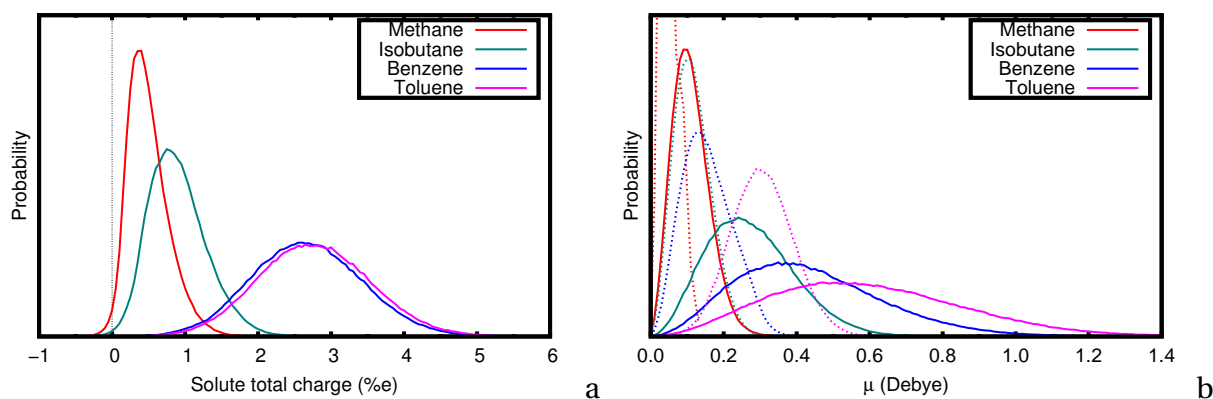


Figure 5.8: Distribution of instantaneous electronic properties of the four hydrophobic molecules along the PM3-PIF3 SEBOMD simulations. a: CM1 total charge of the solute in condensed phase. b: Solute dipole moment computed from CM1 atomic partial charges. Comparison between gas phase (dotted line) and condensed phase (plain line).

Table 5.3: Averages of the instantaneous total charge and dipole moment (in parenthesis, the difference with respect to gas phase) of the hydrophobic solutes during the PM3-PIF3 condensed phase simulation. The CM1 model was used to derive the charges.

Molecule	Total charge (%e)	Dipole moment (Debye)
Methane	+0.5	0.11 (+0.05)
Isobutane	+0.9	0.28 (+0.16)
Benzene	+2.7	0.44 (+0.28)
Toluene	+2.8	0.60 (+0.29)

averaged over the simulation time is reported in Table 5.3. For the two alkane molecules (*i.e.*, methane and isobutane), the charge transfer between solute and solvent is very small leading to a charge of 0.5×10^{-2} e and 0.9×10^{-2} e for methane and isobutane, respectively. The charge transfer is more important in the case of aromatic compounds. Benzene and toluene bear a similar total charge in solution (*i.e.*, 2.7×10^{-2} e and 2.8×10^{-2} e, respectively). As we discussed previously, aromatic compounds experience a specific hydrogen bond with a water molecule *via* the interaction of a water hydrogen atom with the π electrons of the aromatic ring. Such an interaction leads to a partial electron transfer from the aromatic ring to the water molecule, which is consistent with the observed positive total charge of benzene and toluene in water.

The distribution of the instantaneous dipole moment of the different solutes is plotted in Figure 5.8b. This Figure shows a comparison between the gas phase and the condensed phase results (dotted and plain line, respectively). Here again, one can see that the spread of these distributions is quite large, reflecting the variations of this property along the dynamics. Table 5.3 presents the average of the condensed phase dipole moment of the four hydrophobic compounds and the difference with the respective gas phase simulation. It is noteworthy that, in the gas phase, methane and benzene bear a small instantaneous dipole moment (*i.e.*, 0.06 ± 0.03 and 0.16 ± 0.07 Debye, respectively) even though those molecules are non-polar on average. This instantaneous dipole moment arises from asymmetric vibrational modes that break the symmetry of the molecule during the dynamics. We shall discuss the vibrational properties of these systems later in this Section. Isobutane and toluene are polar molecules and bear a dipole moment of 0.12 ± 0.05 and 0.31 ± 0.08 Debye in the gas phase respectively. All the four molecules discussed here are subject to a polarization induced by the solvent when moving from the gas phase to the condensed phase. The dipole moment of methane and isobutane are roughly doubled, reaching 0.11 ± 0.05 and 0.28 ± 0.12 Debye, respectively. The dipole moment of benzene and toluene experience a similar change induced by the solvent (*i.e.*, $\sim +0.3$ Debye), leading to a dipole moment in solution of 0.44 ± 0.20 and 0.60 ± 0.25 Debye, respectively. As suggested by the large values of standard deviations, the dipole moment of benzene and toluene fluctuate with a large amplitude during the dynamics. Similar results have been reported by Mateus *et al.* from

BHandHLYP/aug-cc-pVTZ DFT single point calculations on geometries extracted from an OPLS-AA TIP3P simulation of benzene in water.[244]

The case of methane and isobutane is particularly interesting. Indeed, we showed in the previous section that the first solvation shell is located at a relatively far distance from the solute in those two systems (see Figures 5.1 and 5.2). Therefore, our results suggest that a long range polarization exists. Mateus *et al.* carried out an extensive study of the electronic properties of methane in water by means of MM-MD and *ab initio* static calculations.[235] In particular, the authors have investigated the variation of the methane dipole moment as a function of the number of water molecules present in the first solvation shell. To this end, they extracted 500 geometries from their MM-MD simulation and performed BHandHLYP/aug-cc-pVTZ single point calculations by including sequentially the first 18 water molecules. They computed the methane dipole moment based atomic partial charges derived from the CHelpG charge fitting scheme. They predicted that the dipole moment of methane increases from ~ 0.2 Debye (with no water molecules) to ~ 0.5 Debye (with 18 surrounding water molecules). Our SEBOMD results show a qualitatively good agreement with the observations of Mateus *et al.*, though the dipole moments obtained in that work are much larger than those described in the present work. We performed a similar analysis on 500 geometries extracted from our SEBOMD PM3-PIF3 simulation of methane in water. For each frame, single points calculations were performed at the B97D/aug-cc-pVTZ level on the methane molecule including the 20 closest water molecules. The dipole moment of methane was computed from the atomic partial charges derived from four different models, *i.e.*, Mulliken,[246] Löwdin,[247] Merz-Kollman[248, 249] charge fitting of the electrostatic potential (ESP-MK) and the natural population analysis[250] (NPA). When we evaluate the dipole moment shift with respect to the gas phase model, we find 2.05 (Mulliken), 0.08 (Löwdin), 0.83 (ESP-MK) and 0.06 (NPA) Debye. The spread of these values shows that such a property is strongly related to the charge model used to derive it.

5.2.3 Vibrational properties

Vibrational properties are influenced by the variation of the electronic structure of the solute. Thus, considering the results presented in the previous Subsection, we should expect modifications of the infrared (IR) spectrum of our solutes as a consequence of the change in dipole moment induced by the solvent. The IR spectra have been computed following the protocol described in Chapter 1. Since the intensity of the different vibrational bands cannot be compared with experimental measurements using such an approach, it will be reported in arbitrary units in all the related plots.

By definition, hydrophobic compounds are poorly soluble in water, thus, to our knowledge, no experimental IR data are available in the literature for neither of the four hydrophobic selected molecules in liquid water. However, as discussed previously in Chapter 4, a par-

ticular state of water can encapsulate gas molecules such as molecular hydrogen or methane. Such structure of water is named clathrate hydrate and can be found under particular conditions of pressure and temperature. More emphasis will be made on methane and benzene systems since, respectively, isobutane and toluene share the same type of properties.

Methane. The methane molecule bears 5 atoms and a tetrahedral symmetry (T_d point group). Such molecule is characterized by 9 vibrational frequencies ($3N-6$, N being the number of atoms). The symmetry of the system induces several of those vibrational modes to be degenerate. Thus, only four distinct modes (represented in Figure 5.9) are present in the methane molecule and are described as follows:[251]

- ν_1 : symmetric C-H stretch
- ν_2 : rocking (two times degenerate)
- ν_3 : asymmetric C-H stretch (three times degenerate)
- ν_4 : bend (three times degenerate)

Because of symmetry reasons, ν_1 and ν_2 do not affect the dipole moment of the molecule and thus are not active in IR spectroscopy (*i.e.*, zero intensity).

In Table 5.4 (static section), we report the harmonic equilibrium frequencies of gas phase methane computed by Lee *et al.* at the CCSD(T) level of theory with a cc-pVQZ basis set.[252] In this work, the authors proved a very good agreement with experimental results.[253] We also computed the same property with various SE Hamiltonian and the ff03 force field as a comparison. SE frequencies were obtained with the MOPAC2012 program while the ff03 results have been computed from the diagonalization of the numerical Hessian matrix (see Section 3.4).

The static calculations presented in Table 5.4 show that the results obtained with ff03 are in good agreement with the work of Lee *et al.*[252] While none of the SE methods selected here reproduces the correct order between the two C-H stretch vibrations, PM7 also inverts the position of ν_2 and ν_4 resulting in a wrong ordering of all the frequencies. PM6 and PM7 strongly underestimate the value of ν_3 compared to CCSD(T) predictions. Finally, despite the overestimation of the frequency of ν_1 , AM1 and PM3 show a reasonable agreement with CCSD(T) for the other frequencies.

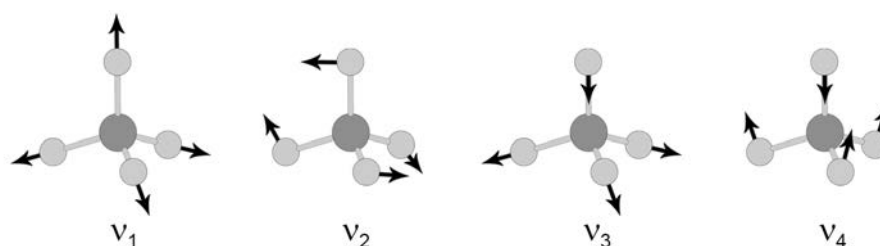


Figure 5.9: Schematic representation of the four vibrational normal modes of the methane molecule.

Table 5.4: Methane characteristic vibrational frequencies in gas phase (GP) and aqueous solution (Aq). Static SE and ff03 results are harmonic force constant computed on the minimized geometry using the related Hamiltonian. Two results are given for molecular dynamics simulations (MD). IR: IR spectrum from methane CM1 dipole moment time correlation function. VDOS: vibrational density of states. For the IR spectra, only ν_3 and ν_4 are active.

		ν_1	ν_2	ν_3	ν_4
		sym.	rocking	asym.	bend
		C-H stretch		C-H stretch	
Harmonic (GP)	Exp ^a	3025.5	1582.7	3156.8	1367.4
	CCSD(T)/cc-pVQZ ^b	3036.2	1570.4	3157.1	1345.3
	ff03	2820.5	1463.7	2981.9	1330.6
	AM1	3214.9	1411.6	3103.1	1379.5
	PM3	3310.9	1450.9	3207.5	1361.9
	PM6	2827.5	1261.6	2721.0	1253.7
	PM7	2813.6	1291.5	2689.8	1306.6
MD (GP)	ff03				
	IR	–	–	3022.9	1285.9
	VDOS	2851.8	1444.6	3023.9	1287.2
	PM3				
	IR	–	–	3229.0	1362.0
	VDOS	3335.8	1449.6	3229.9	1361.8
MD (Aq)	ff03 SPC/E				
	IR	–	–	3035.4	1286.7
	VDOS	2867.6	1454.3	3057.0	1288.6
	PM3-PIF3				
	IR	–	–	3182.1	1352.7
	VDOS	3294.3	1435.8	3189.1	1352.7

^a Ref. [253], ^b Ref. [252]

From our MD simulations, we computed the infrared spectrum of methane, both in the gas and in the condensed phase. Details about the methodology used to compute those spectra are given in Section 1.5. Figure 5.10a shows the IR spectra obtained with PM3-PIF3 and ff03 and gives a comparison between gas phase and condensed phase results. As expected, the methane IR spectrum shows only two active vibrations, ν_3 and ν_4 . For both PM3-PIF3 and ff03, the difference between gas and condensed phase is very slight. However, when zooming on the high frequencies region of the spectrum (see Figure 5.10b), a shift of the C-H stretch band (ν_3) occurs, both in MM-MD and SEBOMD simulations. The position of the two bands ν_3 and ν_4 from those IR spectra are reported in Table 5.4. In particular,

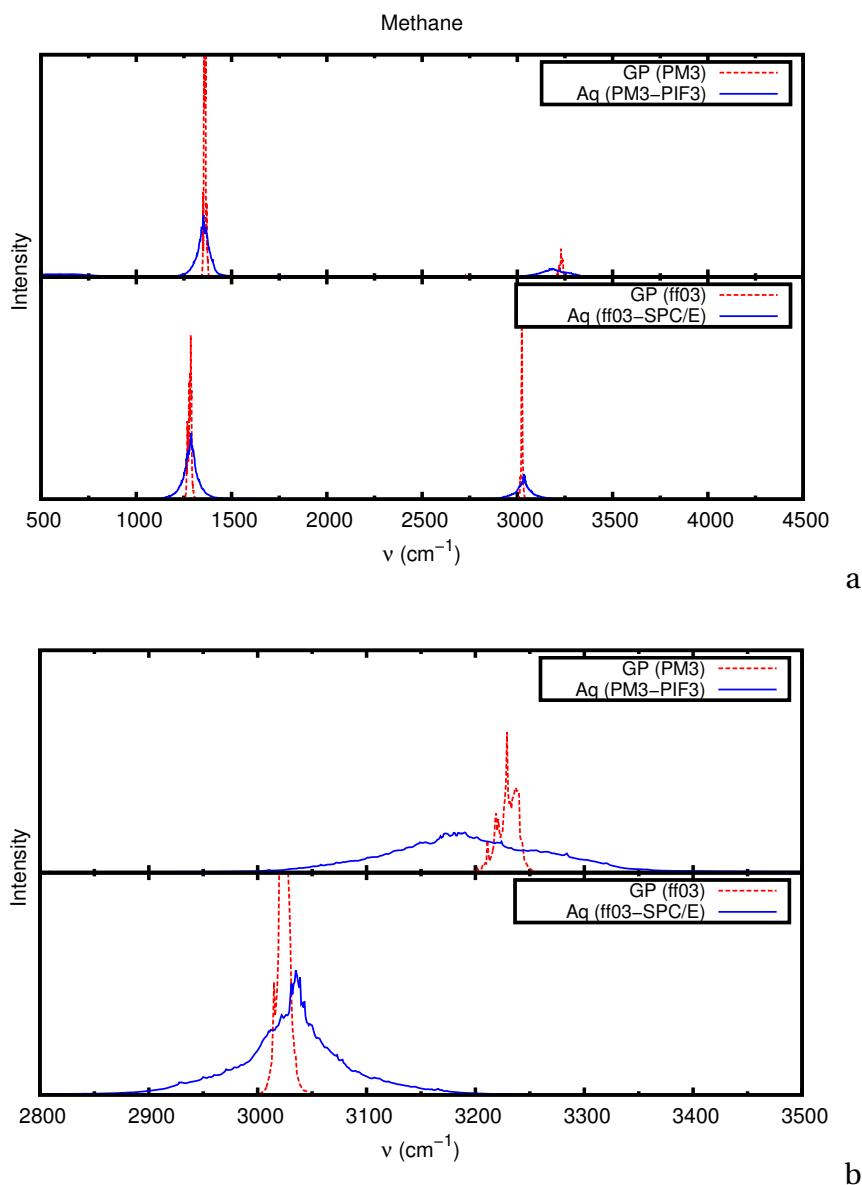


Figure 5.10: Comparison of infrared spectra of methane obtained with PM3-PIF3 and ff03. a: full frequency range. b: zoom on the high frequency region. In both a and b, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

Dartois *et al.* focused on the measurement of methane IR spectrum both in gas phase and encapsulated within clathrate hydrates for astronomic applications.[217] They have shown that the water cage induces a red shift (*i.e.*, displacement to the low frequencies) of the C-H asymmetric stretch band (ν_3). A theoretical study of this phenomenon has been carried out by Greathouse *et al.* using an MM point charges force field.[254] The sign of the shifts observed with this MM force field is the opposite of the experimental predictions (*i.e.*, blue instead of red shift). A blue shift of 11.5 cm^{-1} is also obtained here using the ff03-SPC/E force field. The latter result is also consistent with the observations of Greathouse *et al.* in their study of methane encapsulated in clathrate hydrate with a classical force field.[254] On the other hand, simulations with the PM3-PIF3 Hamiltonian predict a red shift of ν_3 in solution of about 46.9 cm^{-1} . This result is in reasonable agreement with those from Dartois *et al.*

who measured a red shift of about 20 cm^{-1} in a clathrate hydrate at 7K.

As discussed above, ν_1 and ν_2 are not active in the IR spectrum. A way to obtain such vibrational bands is to compute the vibrational density of states (VDOS).[74, 80] This method yields a vibrational spectrum from the time correlation function of the atomic velocities. This technique is also helpful to identify vibrational modes since the spectrum can be decomposed into atomic contributions. However, the IR spectrum of methane is simple enough and does not require such analysis. Here, we use VDOS to get the position of ν_1 and ν_2 . These results are reported Table 5.4, as well as ν_3 and ν_4 for a comparison with IR. We note that the values obtained with VDOS for the latter are consistent with IR results. As noticed from the static calculations, the SEBOMD results show an inversion of the two C-H stretch bands (*i.e.*, ν_1 and ν_3), while MM predicts the correct order. The symmetric C-H stretch experiences the same effect as ν_3 , being red shifted with PM3-PIF3 and blue shifted with ff03. Finally, the same observations can be drawn for ν_2 , though a much smaller shift is obtained.

Isobutane. Figure 5.11 shows the IR spectra of isobutane calculated from SEBOMD and MM-MD simulations in the gas and in the condensed phase. The low frequencies part of the isobutane spectrum does not seem to suffer any modification induced by the presence of the solvent. Semiempirical and MM results show some differences. However, since this region of the spectrum contains many modes that are coupled with each other, attempts to identify all the bands are out of the scope of this study. We shall focus on the $2800\text{--}3300\text{ cm}^{-1}$ region of the spectrum, which contains all types of C-H stretch. Schachtschneider *et al.* reported a complete analysis of the spectra of a large selection of hydrocarbon compounds.[255] For isobutane, they attributed the six highest frequencies of the spectrum to the different asymmetric C-H stretching modes of the methyl groups ($\nu_{\text{CH}_3}^{\text{asym}}$), lying in a range of $2958\text{--}2962$

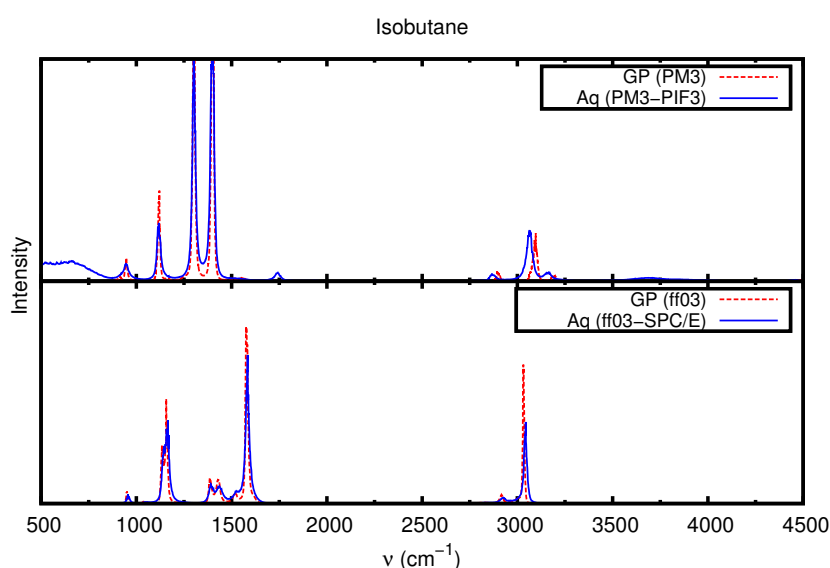


Figure 5.11: Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

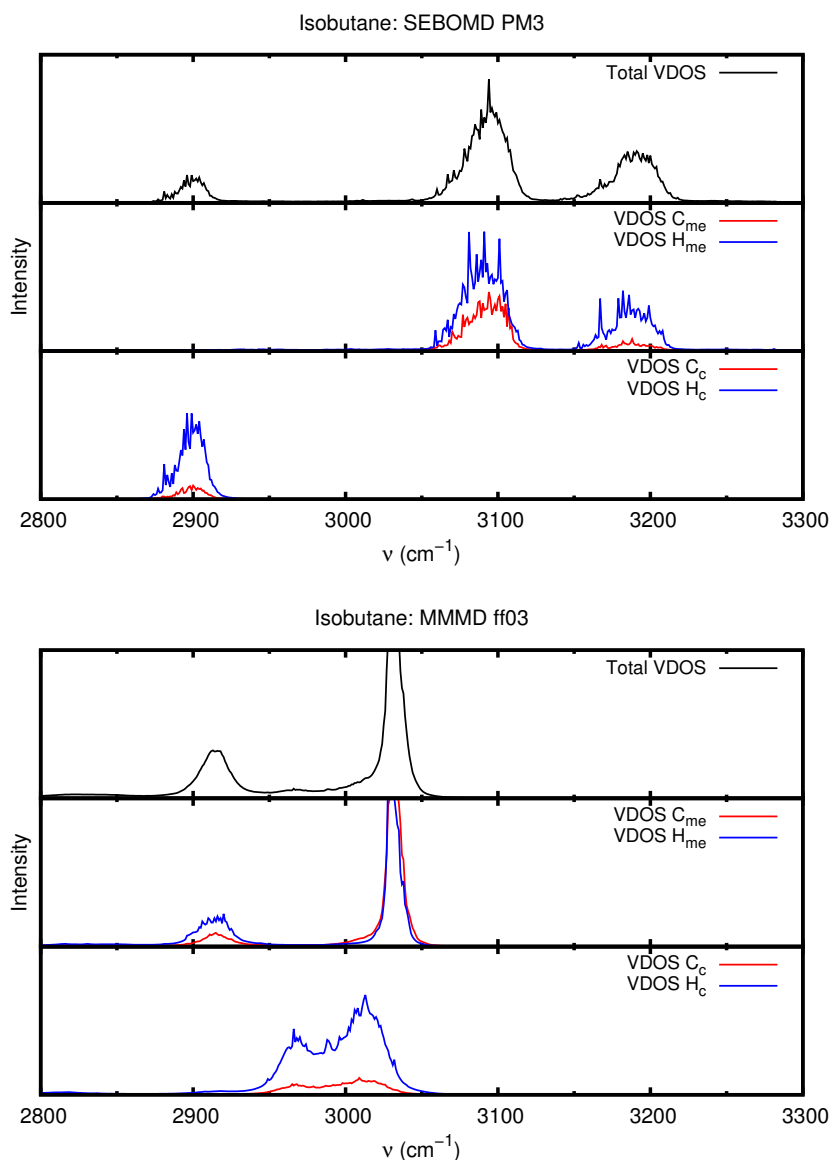


Figure 5.12: Decomposition of the VDOS of isobutane into atomic contributions simulated in the gas phase with PM3 (top panel) and ff03 (bottom panel). Intensities are reported in arbitrary units.

cm^{-1} . The C-H stretch of the central CH group (ν_{CH}) is located at 2904 cm^{-1} , and finally, the three symmetric C-H stretching modes of the methyl groups ($\nu_{\text{CH}_3}^{\text{sym}}$) are found between 2880 and 2894 cm^{-1} .

We used the decomposition of the VDOS into atomic contributions to identify the vibrational bands on the isobutane IR spectra. We shall use the same nomenclature as introduced earlier to differentiate the carbon and hydrogen atoms. Thus, the central carbon and hydrogen atoms will be named C_c and H_c , respectively, while the carbon and hydrogen atoms of the methyl groups will be referred as C_{me} and H_{me} , respectively. Finally, since gas phase and condensed phase results are very close, only the gas phase VDOS spectra will be analyzed here.

Figure 5.12 shows the VDOS decomposition into atomic contributions for the gas phase simulations. Let us start by analyzing the SEBOMD simulation (top panel of Figure 5.12). The

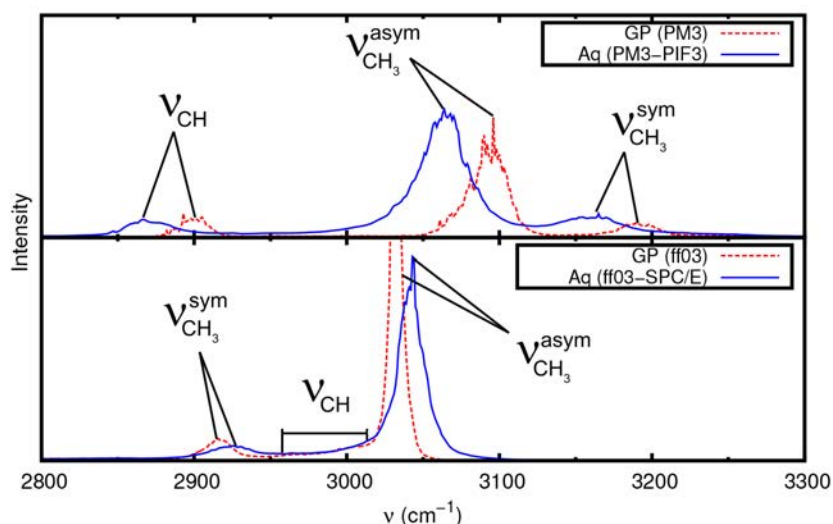


Figure 5.13: Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03 in the frequency range between 2800 and 3300 cm^{-1} . The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

contribution of C_c and H_c are both located in one peak around 2900 cm^{-1} . This peak can thus be unequivocally attributed to ν_{CH} . In the case of C_{me} and H_{me} , two peaks appear around 3095 and 3190 cm^{-1} . To attribute the peaks to $\nu_{\text{CH}_3}^{\text{asym}}$ and $\nu_{\text{CH}_3}^{\text{sym}}$, one should analyze the contribution of C_{me} . Indeed, this atom is almost not involved in the symmetric C-H stretch of the methyl groups. From this observation, it comes that $\nu_{\text{CH}_3}^{\text{asym}}$ and $\nu_{\text{CH}_3}^{\text{sym}}$ are attributed to the frequencies 3190 and 3035 cm^{-1} respectively. For the MM-MD simulation (bottom panel of Figure 5.12), similar analysis can be done. ν_{CH} appears to yield a broadened peak centered approximately around 3000 cm^{-1} . From the contribution of C_{me} , we can identify the peak of $\nu_{\text{CH}_3}^{\text{sym}}$ at 2915 cm^{-1} and the peak of $\nu_{\text{CH}_3}^{\text{asym}}$ at 3030 cm^{-1} .

In Figure 5.13, we report the IR spectra of isobutane between 2800 and 3300 cm^{-1} in the gas and in the condensed phase for both SEBOMD and MM-MD simulations. The peak assignment proposed above is also depicted on these plots. As it has been observed previously in the case of methane, the PM3 Hamiltonian appears to invert the position of the C-H stretch vibrational bands. To the opposite, ff03 gives the correct frequency order and the position of the peaks in good agreement with experimental results.[255] Despite the wrong position of the peaks in SEBOMD simulations (which is inherent to the PM3 Hamiltonian), one can observe a red shift of all the C-H stretch modes comparable to the one discussed above in the case of methane. In a similar manner, ff03 predicts the wrong solvent effect, with a blue shift of the C-H stretch bands.

Benzene. Benzene is a cyclic aromatic molecule bearing a D_{6h} symmetry. Because of this high symmetry, only four normal vibrations are active in the IR spectrum. Those are reported in the book of Varsányi and Szöke,[256] and are given as follows in gas phase: out-of-plane

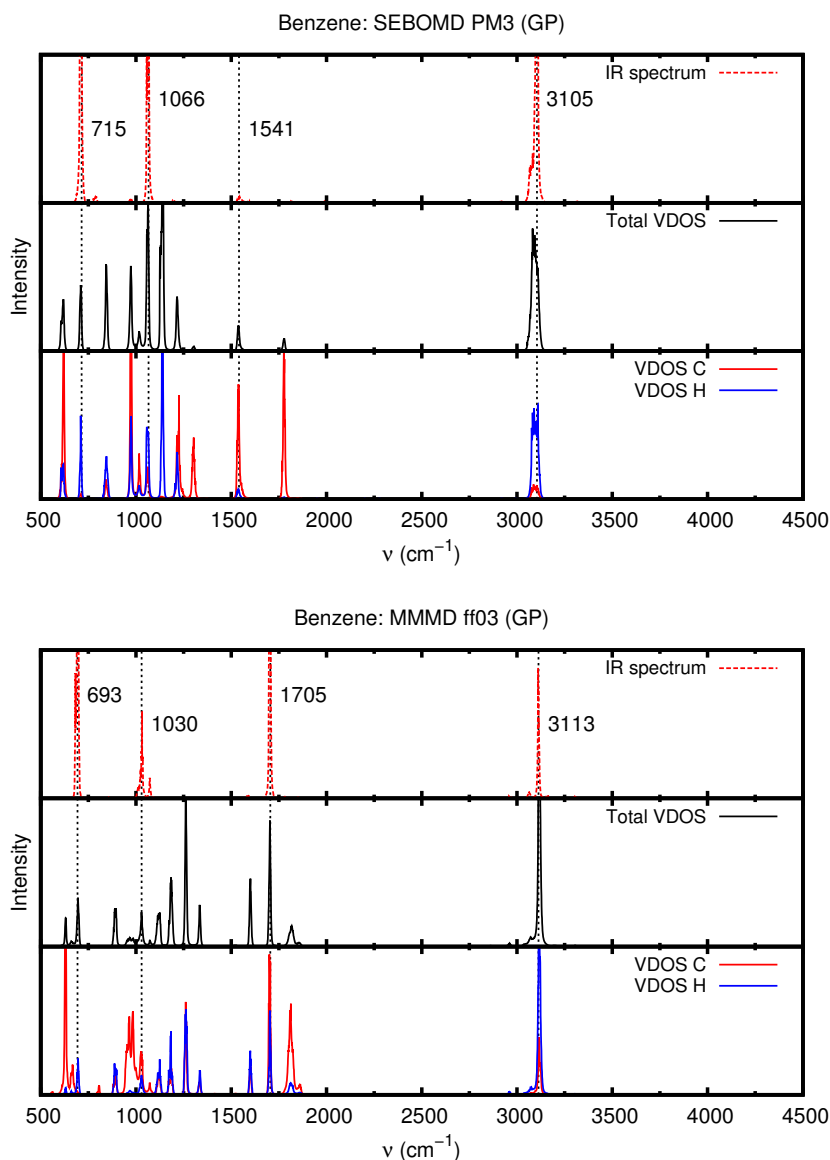


Figure 5.14: Decomposition of the VDOS of benzene simulated in the gas phase with PM3 (top panel) and ff03 (bottom panel). Intensities are reported in arbitrary units. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.

C-H bending ($\nu_{\text{CH}}^{\text{oop}} = 673 \text{ cm}^{-1}$), in-plane C-H bending ($\nu_{\text{CH}}^{\text{ip}} = 1037 \text{ cm}^{-1}$), C-C stretching ($\nu_{\text{CC}} = 1482 \text{ cm}^{-1}$) and C-H stretching ($\nu_{\text{CH}} = 3064 \text{ cm}^{-1}$).

In Figure 5.14, we report the gas phase infrared and VDOS spectra of benzene obtained from SEBOMD and MM simulations. As expected, only four peaks appear in the IR spectra. The VDOS contains all the vibrational modes. The first and the last peak are quite simple to assign. Indeed, both have a larger contribution of the hydrogen component of the VDOS. The first one is $\nu_{\text{CH}}^{\text{oop}}$ and is located at 715 and 693 cm^{-1} for PM3 and ff03 respectively. The last peak is assigned to ν_{CH} : the frequencies using PM3 and ff03 are 3105 and 3113 cm^{-1} , respectively. The two peaks in the middle are slightly more complicated. In SEBOMD simulation, the peak at 1066 cm^{-1} shows a mixed contribution of C and H atoms to the total VDOS. This peak can be assigned to $\nu_{\text{CH}}^{\text{ip}}$. The one at 1541 cm^{-1} has a very small intensity,

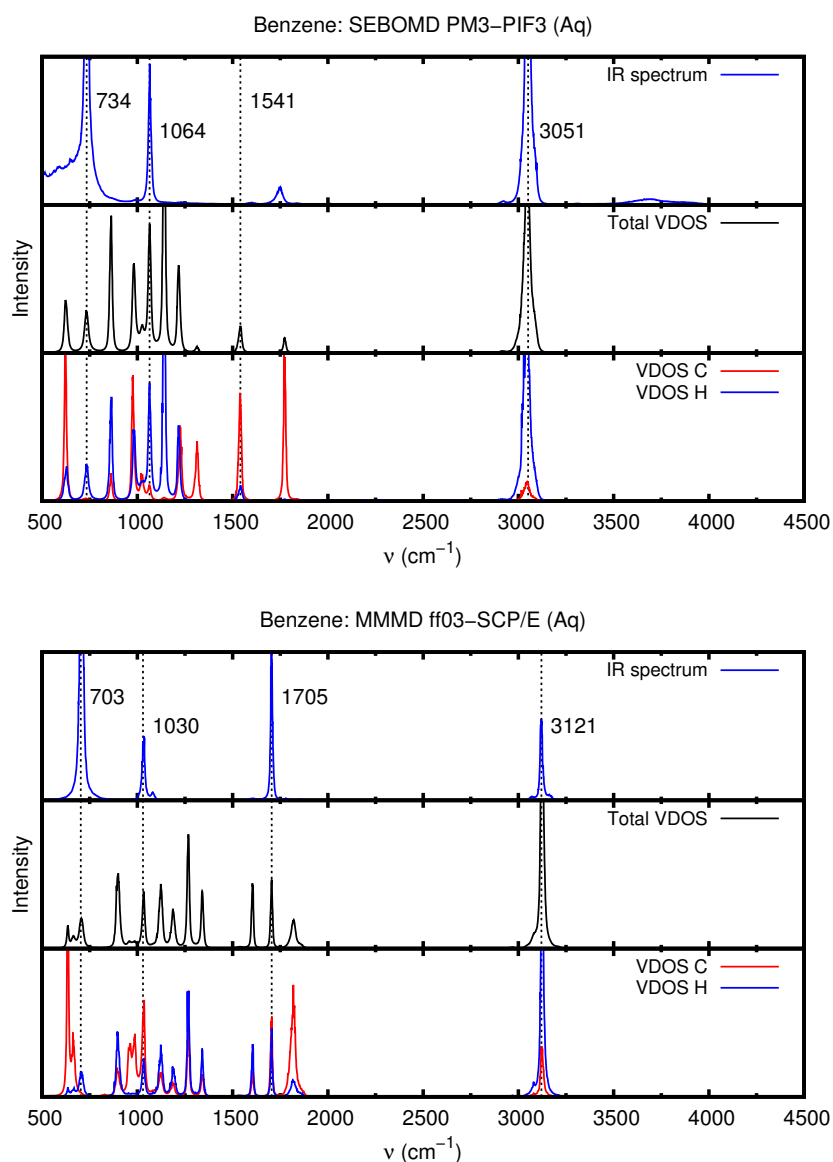


Figure 5.15: Decomposition of the VDOS of benzene simulated in the condensed phase with PM3-PIF3 (top panel) and ff03-SPC/E (bottom panel). Intensities are reported in arbitrary units. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.

but corresponds to a peak in the VDOS. Notice that the experimental intensity is also quite small. This peak seems to involve the motion of C atoms only, which can be attributed to the C-C stretching (ν_{CC}). The assignment is made easier by performing an equilibrium normal modes analysis using the tools presented in Section 3.4 (the normal mode analysis of each molecule presented in this work is available as Supplementary Material). This analysis is of great help to attribute ν_{CH}^{ip} and ν_{CC} in MM-MD. Indeed, the VDOS shows a mixed contribution of C and H atoms to the total VDOS of these peaks (1030 and 1705 cm^{-1}). The visualization of these normal modes shows a strong coupling between those and thus, no unequivocal assignment can be made. Finally, despite the mixed character of ν_{CH}^{ip} and ν_{CC} with ff03, both Hamiltonians give the correct frequency order in gas phase and PM3 shows results in very good agreement with experimental measurements.

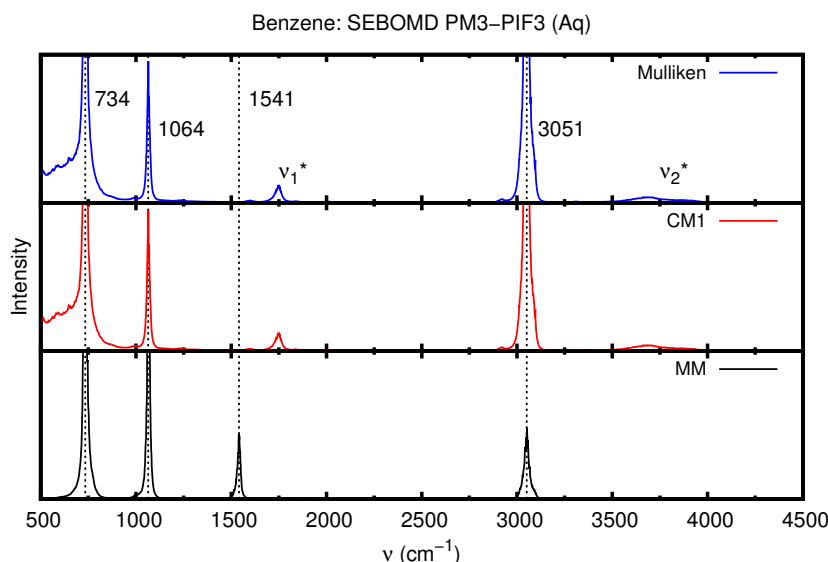


Figure 5.16: PM3-PIF3 IR spectra of benzene in water computed with different charge models: Mulliken, CM1 and MM point charges. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.

Figure 5.15 presents the corresponding IR and VDOS spectra in aqueous solution. First, it appears clearly that the ff03 spectra are not modified by the presence of the solvent. Only the ν_{CH} (3121 cm^{-1}) stretch and the peak at 703 cm^{-1} are slightly blue shifted compared to the gas phase. In SEBOMD simulation, the solvent has a clear impact on the IR spectrum. $\nu_{\text{CH}}^{\text{oop}}$ (734 cm^{-1}) is blue shifted compared to the gas phase and the most important effect is seen on ν_{CH} , which is red shifted by about 54 cm^{-1} . This results is consistent with what we observed in the case of methane. However, to our knowledge, this effect is not observed experimentally. What could appear at first glance like a strong blue shift of ν_{CC} (*i.e.*, a peak at 1748 cm^{-1}) does actually not correspond to any peak on the VDOS, the closest being placed at 1774 cm^{-1} . The origin of this peak will be discussed later in this Section. ν_{CC} is actually not affected by the solvent. Only its intensity is strongly lowered and thus, it does not appear on the IR spectrum.

To better understand the origin of the peak appearing at 1748 cm^{-1} (ν_1^*), one should focus on the methodology used to compute the dipole moment of the solute molecule. Indeed, the dipole moment is only defined for a neutral system since its value depends on the chosen referential for a charged one. Here, since the total charge of benzene is quite small (*i.e.*, 2.7 %e, see Table 5.3) we can use the dipole moment computed from CM1 atomic partial charges in the solute molecular frame. However, another assumption could be made by using fixed point charges coming from the MM force field. The later would thus be only dependent on the geometrical variations and not on the charge fluctuations. We present in Figure 5.16 the benzene IR spectrum in solution coming from SEBOMD simulation using different charge models (*i.e.*, Mulliken, CM1 and MM). The first observation that we can draw from these plots is that the charge model does not affect the general shape of the spectrum.

Indeed, $\nu_{\text{CH}}^{\text{oop}}$, $\nu_{\text{CH}}^{\text{ip}}$ and ν_{CH} appear at the same position in all the cases. Both QM models (*i.e.*, Mulliken and CM1) show the same bands in the $1500\text{--}1750\text{ cm}^{-1}$ region of the spectrum (ν_1^*). The observed phenomenon is thus not related to the CM1 charge model. In addition, a broad band appears with QM charges between 3500 and 4000 cm^{-1} (ν_2^*). To the opposite, the spectrum computed with MM charges only shows the four characteristic vibrational bands of benzene. In this case, ν_1^* and ν_2^* do not appear and ν_{CC} is placed at the expected position. ν_1^* and ν_2^* are thus due to the fluctuations of QM charges. It is noteworthy that ν_1^* has a frequency of 1748 cm^{-1} , which is particularly close to the value obtained with PM3 for the bend of water in gas phase (*i.e.*, 1742 cm^{-1}). The same observation can be made for ν_2^* which corresponds to the two O-H stretch of water, *i.e.*, 3868 and 3989 cm^{-1} with PM3. It is reasonable to conjecture that those vibrational bands may thus result from a coupling between the instantaneous charges of benzene and water due to the charge transfer. The fluctuations of the benzene electronic density appear to be directly related with the vibrational properties of water. Moreover, if one looks carefully at all the IR spectra presented in this work, these bands appear only for those systems for which no other vibrational band with stronger intensity is present in the spectrum in the same region. However, no further investigation will be pursued here since the analysis of such a phenomenon would require additional developments in the methodology used for the calculation of IR spectra.

Toluene. Toluene is a mono-substituted derivative of benzene. It thus bears similarities with both aromatic and alkane compounds. A comparison of gas and condensed phase IR spectra of this molecule is presented Figure 5.17. The discussion of these plots will be more concise since most of the solvent effects have been discussed previously. Here again, only the

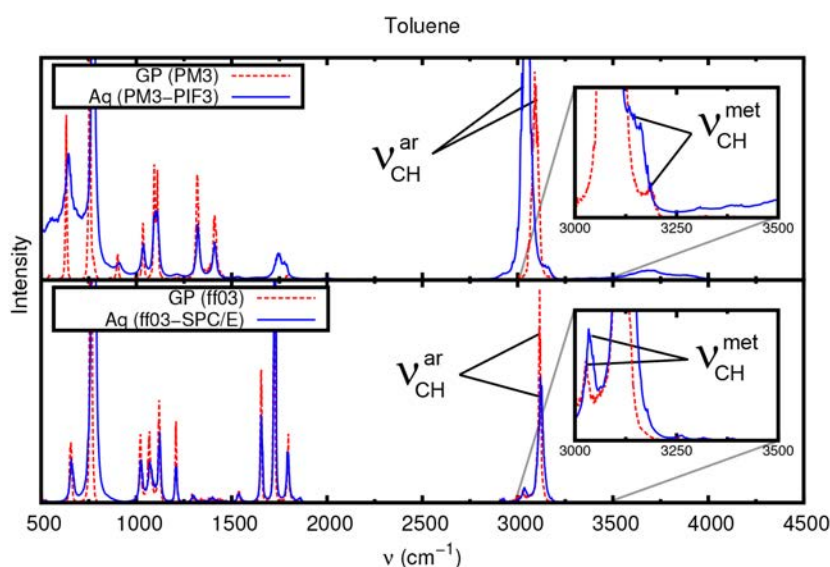


Figure 5.17: Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line. In both plots, a zoom on the $3000\text{--}3500\text{ cm}^{-1}$ region is presented as an inset.

high frequency region (*i.e.*, larger than 2000 cm^{-1}) appears to be affected by the presence of solvent. Two types of C-H bonds exist in toluene, on the aromatic ring and the methyl group and we will refer the related vibrational bands as $\nu_{\text{CH}}^{\text{ar}}$ and $\nu_{\text{CH}}^{\text{met}}$ respectively. Using the same methodology as previously introduced, it is possible to decompose the VDOS into contributions related to these two different domains. The VDOS spectra are not presented here but are available as Supplementary Material.

The peaks assignment depicted Figure 5.17 has been obtained using such a methodology. As previously observed, $\nu_{\text{CH}}^{\text{ar}}$ and $\nu_{\text{CH}}^{\text{met}}$ are red shifted in SEBOMD simulations while the corresponding bands are slightly blue shifted in MM-MD. It is however interesting to notice that the frequency order of those two peaks is correctly predicted by ff03 while it is inverted by PM3-PIF3. Nevertheless, the PM3-PIF3 results are consistent with the observations derived from the three other hydrophobic compounds studied in the present work. Finally, we notice that the two vibrational bands ν_1^* and ν_2^* are still present on the PM3-PIF3 spectrum in aqueous solution.

5.2.4 Summary

This study shows that the SEBOMD methodology is well suited for the simulation of hydrophobic compounds in water. We have performed an analysis of the solvent structure around methane, isobutane, benzene and toluene and studied the effect of the solvent on key properties of such compounds, *i.e.*, electronic and vibrational.

Even though the solvent effects are relatively small, they are reproduced by this methodology and are in quite good agreement with experimental measurements and other theoretical works. We also show in this Section that the quality of the results obtained with SEBOMD strongly depends on the choice of the semiempirical methods. The PM3-PIF3 Hamiltonian appears to be a good candidate to study hydrophobic hydration, confirming the good results obtained from static calculation in Chapter 4. Finally, an interesting phenomenon has been pointed out concerning the IR spectra of hydrated compounds, *i.e.*, the presence of bands having a low intensity, which could be related to a coupling between the electronic structure of the target molecule with the vibrational modes of water. However, further investigations are necessary to find out about the relevance of such an observation.

5.3 Hydrophilic compounds

In biological systems, the balance between hydrophobic and hydrophilic interactions dictates the way in which proteins interact with water and directs their folding. We propose here to verify if the PM3-PIF3 Hamiltonian qualitatively reproduces this balance for the solvation of small hydrophilic compounds along SEBOMD simulations. To this end, we shall first discuss the solvent structure around key sites of each molecule and the effect of water molecules on the internal structural properties of the solutes. We will also focus on the changes in electronic properties that the molecules experience when they are immersed in water. Finally, we will give an analysis of the infrared spectra of the selected hydrophilic molecules and compare the results obtained in the gas phase and in aqueous solution.

5.3.1 Solvent structure and solute geometrical properties

The hydrophilic molecules that we chose for this study bear two interesting features. First, they are more flexible than the hydrophobic compounds that we discussed in the previous Section. This will lead our discussion to elucidate the solvent effects on the conformational properties of these compounds. Second, most of the six selected molecules (see Table 5.1) were chosen to model amino acid side chains and bear both a hydrophilic and a hydrophobic group. This will allow us reach a step further to our objective, *i.e.*, to simulate polypeptides in water.

Ethanol. The ethanol molecule is known to exist in two conformations, *i.e.*, *trans* and *gauche* as illustrated in Figure 5.18.[257] They are characterized by different values of the ω dihedral angle defined as $C_1 - C_2 - O_a - H_a$ (labels are depicted in the Figure). Understanding the relative stability of both conformers as well as the stability of the related ethanol-water 1:1 complexes has been the focus of several works in the past decades. However, all those studies were performed using different level of theory and thus no correct global picture was available. In 2004, Fileti *et al.* carried an exhaustive study of this system at the MP2/aug-cc-pVDZ level of theory.[257] In this work, the authors pointed out that, for the isolated ethanol molecule, the *trans* conformer (E_t) is predicted to be more stable than the *gauche* one (E_g) by about 0.25 kcal/mol.

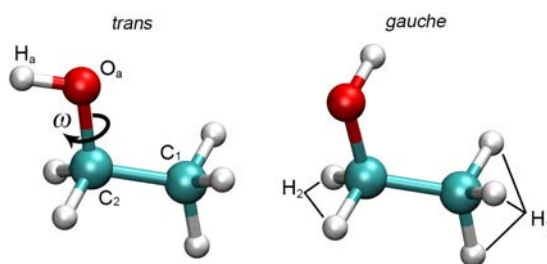


Figure 5.18: Illustration of *trans* and *gauche* conformers of ethanol and atom label definition. The ω dihedral angle is defined as $C_1 - C_2 - O_a - H_a$.

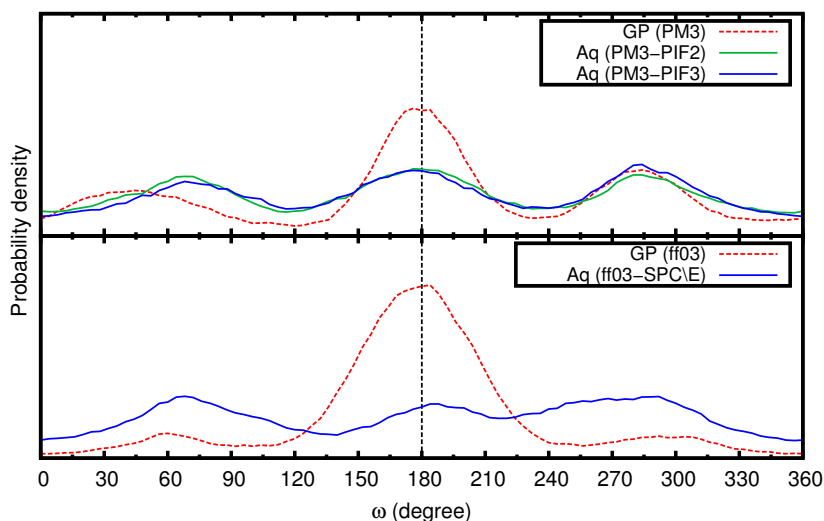


Figure 5.19: Comparison of the ethanol ω dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.

Ethanol forms hydrogen bonds and can act either as a proton donor (HD) or as a proton acceptor (HA). The difference in binding energy between HA and HD has been computed by the same group to be 1.22 and 1.25 kcal/mol in favor of the HA form of the E_t -water and of the E_g -water complexes, respectively. They also showed that, for the ethanol-water complex, the relative stability of the two conformers is inverted compared to gas phase. The *gauche* conformer appears to be more stable by about 0.26 and 0.29 kcal/mol than the *trans* one in the HD and HA forms, respectively.

We present in Figure 5.19 the distribution of the ω dihedral angle from our SEBOMD and MM-MD simulations. Both gas and condensed phase results are presented on these plots. For the SEBOMD condensed phase simulations, we report the results obtained with the PM3-PIF2 and PM3-PIF3 Hamiltonians. The gas phase results show that both PM3 and ff03 MD simulations predict the E_t as the most probable conformer (*i.e.*, $\omega = 180^\circ$). While the PM3 simulation predicts a qualitatively reasonable ratio between E_t and E_g compared to MP2 calculations,[257] the difference between the *trans* and *gauche* conformers is largely overestimated with the MM force field in the gas phase. When the ethanol molecule is dissolved in water, one can see that all three models (*i.e.*, ff03-SPC/E, PM3-PIF2 and PM3-PIF3) predict the *gauche* form to be more probable. We do not observe any difference concerning the conformational behavior of ethanol in water when using the PIF2 or the PIF3 correction of PM3. This first analysis allows us to be quite confident about the structural properties of ethanol along our simulations.

Ethanol is one of the simplest bifunctional organic molecule. It bears both a hydrophilic and a hydrophobic group and thus constitutes the first step to model any larger biological compound. Its hydration structure is the result of a well balanced combination of hydrophilic and hydrophobic interactions. Fidler *et al.* studied the ethanol hydration shell

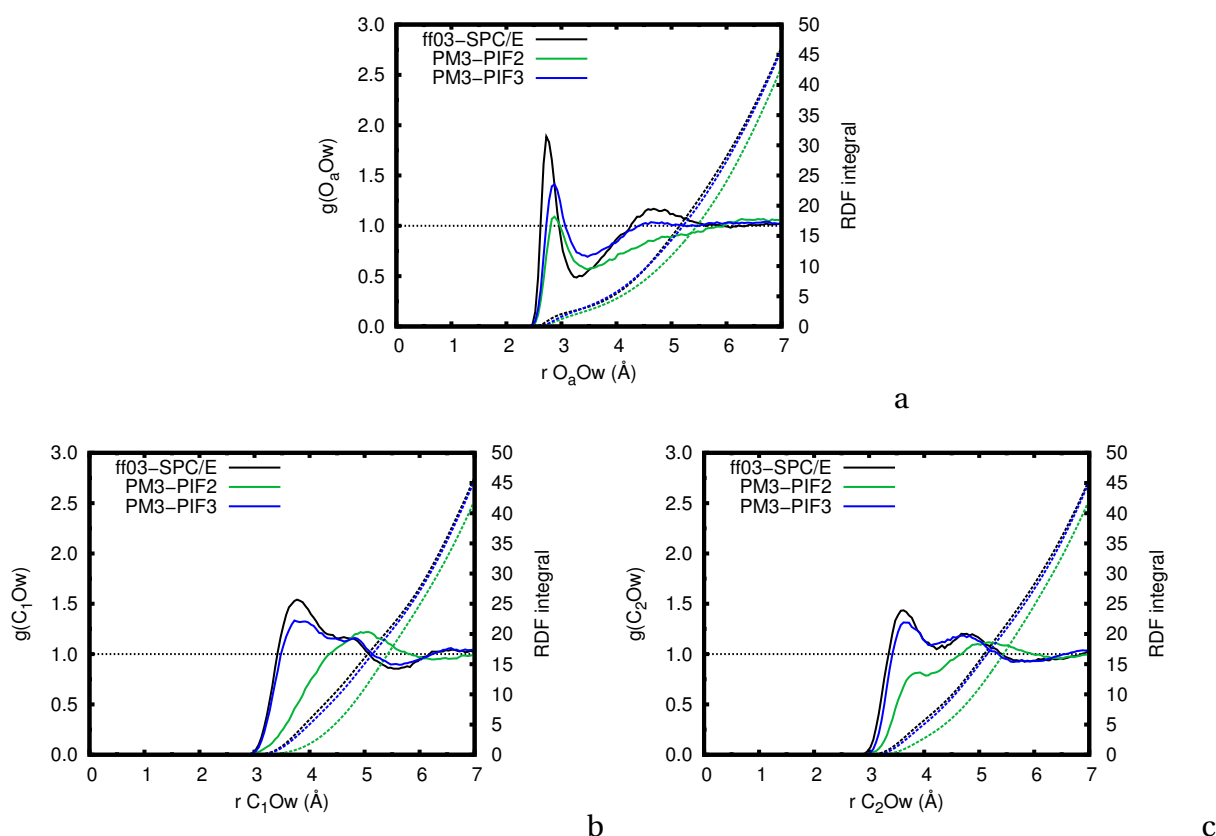


Figure 5.20: Comparison of radial pair distribution functions of ethanol in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF. Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (Ow) with respect to the ethanol oxygen atom (O_a). b: C₁Ow. c: C₂Ow

from MD simulations of one ethanol molecule in a box of 256 TIP4P water molecules.[258] They used the OPLS force field for the solute with a united atom description of the methyl and methylene groups. Later, van Erp and Meijer performed similar simulations using the CPMD technique with the BLYP functional, though they used smaller simulation boxes (*i.e.*, 31 and 63 water molecules).[259] Finally, Noskov and coworkers investigated the hydration of ethanol in water at different concentrations using a dedicated polarizable force field.[260] We shall now discuss our results and compare them with those reported by the three groups above cited.

We present in Figure 5.20a, b and c the ethanol-water RDFs for the O_aOw, C₁Ow and C₂Ow pairs, respectively. Let us start with the O_aOw RDF (Figure 5.20a). The ff03-SPC/E results obtained from the present work give a global shape of the RDF in good agreement with other results in the literature.[258–260] Also the position of the first peak (*i.e.*, 2.7 Å) is close to the results of other groups. The PM3-PIF2 and PM3-PIF3 simulations both predict a peak located at a slightly larger distance (*i.e.*, 2.9 Å). This value is similar to the minimum OO distance predicted by Fileti *et al.* at the MP2 level in the ethanol-water 1:1 complex,[257] which was found to be between 2.83 and 2.92 Å, depending on the configuration. The integration of the peaks for the three methods up to their respective first minimum leads to 2.8, 2.8 and 3.0 water molecules for ff03-SPC/E, PM3-PIF2 and PM3-PIF3, respectively. Both ff03-SPC/E

and PM3-PIF3 results predict a second broad peak at approximately 4.5 Å, in good agreement with other works. Such a broad peak is representative of the solvation shell around the methyl group of ethanol. Considering the results obtained for the hydration of methane using the PM3-PIF2 Hamiltonian, it is not surprising that this peak is not present on the RDF obtained with this method.

For the distribution of water oxygen atoms around the carbon atom of the ethanol methyl group (Figure 5.20b), Fidler *et al.* and Noskov *et al.* found a first peak at 3.8 Å and a broad shoulder placed around 4.8 Å. The CPMD calculation of van Erp *et al.* also showed a first peak at 3.8 Å, however, the limited simulation time used in this work is not sufficient to yield a well converged second solvation shell. In the present calculations, both ff03-SPC/E and PM3-PIF3 show similar results with a first peak at 3.8 Å and a shoulder at 4.7 and 4.8 Å respectively. To the opposite, the PM3-PIF2 Hamiltonian presents a completely different picture with a single broad peak centered around 5.0 Å. Also in this case, this observation is related to the overestimation of the repulsion between water and alkyl fragments.

Finally, the C_1Ow RDFs depicted Figure 5.20c show a good agreement between ff03-SPC/E and PM3-PIF3. Both methods predict a first peak at 3.7 Å and a second one at 4.8 Å having a lower intensity. These observation are also in good agreement with the literature. [258–260] Here again, the PM3-PIF2 gives different results. The position of the two peaks is shifted by about +0.2 Å and the relative intensities are inverted.

All the cited methods bear different deficiencies because of the type of approximations that they involve. Nevertheless, they all lead to qualitatively similar results for the structure of the ethanol solvation shell. Only PM3-PIF2 yields a different picture, however, these results are consistent with the observed problems of this Hamiltonian. This analysis also shows that only the prediction of hydrophobic interactions was problematic with PM3-PIF2 and that the strategy used to develop PM3-PIF3 corrected it, without affecting the hydrophilic interactions part.

p-ethylphenol. The p-ethylphenol molecule bears the three types of molecular groups that we studied up to now: an hydrophobic alkyl tail, an aromatic ring and a hydrophilic hydroxyl group. The flexibility of this molecule is characterized by two dihedral angles ω_1 and

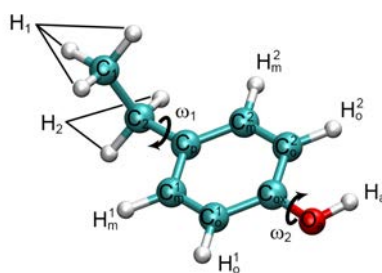


Figure 5.21: p-ethylphenol labels definition. ω_1 and ω_2 dihedral angles are defined as $(C_m^2 - C_p - C_2 - C_1)$ and $(C_o^1 - C_{ox} - O_a - H_a)$, respectively.

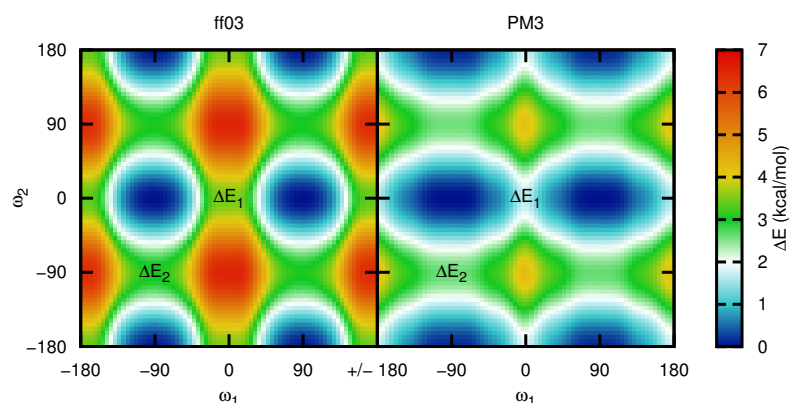


Figure 5.22: p-ethylphenol potential energy surface in the gas phase as a function of the two dihedral angles ω_1 and ω_2 . ΔE_1 and ΔE_2 are positioned at $(0^\circ, 0^\circ)$ and $(-90^\circ, -90^\circ)$, respectively.

ω_2 (depicted Figure 5.21), which represent the rotation of the ethyl and hydroxyl substituents of the phenyl ring, respectively.

Figure 5.22 presents a rigid scan of the p-ethylphenol potential energy (ΔE) surface (PES) as a function of the two dihedral angles ω_1 and ω_2 in the gas phase. ff03 and PM3 give similar results for the variation of the energy as a function of ω_2 . The rotational barrier of the OH group (ΔE_2) is estimated at 3.0 and 2.6 kcal/mol for ff03 and PM3, respectively. This internal rotation has been studied experimentally by Larsen *et al.* as well as Berden and coworkers for phenol.[261, 262] Both groups measured a similar rotational barrier about 1215 cm^{-1} (*i.e.*, ~ 3.5 kcal/mol). For the variation of ΔE as a function of ω_1 , no experimental nor theoretical results have been reported in the literature. ff03 and PM3 give two different pictures. In the former case, the rotation barrier of the ethyl group (ΔE_1) is about 3.5 kcal/mol while the latter predicts it to be 1.8 kcal/mol lower. The shape of the two wells also differs for the two methods. Indeed, the well is much sharper with ff03 than it is with PM3. As we have shown in Chapter 4, the PM3 Hamiltonian presents an artifact in the interaction energy surface of water with methane. This method predicts a minimum for the HH interaction of about 2 kcal/mol located at 1.7 Å. Along our rigid scan, short HH distances (*i.e.*, between 2.0 and 2.2 Å) can occur for values of ω_2 between -40° and 40° . The HH interaction artifact of PM3 might be the cause of the lower ΔE_1 barrier obtained with this Hamiltonian compared to ff03. This observation suggests that also the intramolecular interactions might have to be reconsidered to improve the PM3 method.

We analyzed the value of ω_1 along our molecular dynamics simulations in the gas and in the condensed phase. Figure 5.23 presents the results obtained for the simulations with PM3 and ff03 for the gas phase and PM3-PIF2, PM3-PIF3 and ff03-SPC/E for the condensed phase. In both MM-MD and SEBOMD gas phase simulations, the distribution of ω_1 is not symmetrical and shows two maxima about 90° and 270° , meaning that the ethyl group is oriented perpendicularly to the aromatic ring plane. This asymmetry can be due to the limited simulation time considered in this study (*i.e.*, 500 ps). However, such results give us insights about

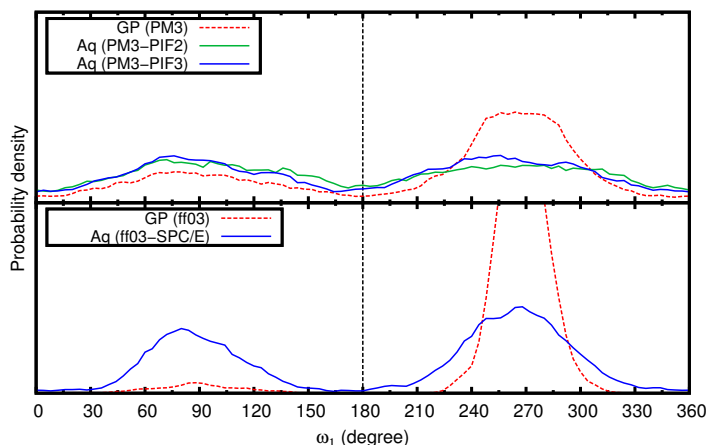


Figure 5.23: Comparison of the p-ethylphenol ω_1 dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.

the relative rigidity of the system depending on the chosen Hamiltonian. The difference in intensity between PM3 and ff03 results is consistent with the results obtained from the rigid scans above. In the condensed phase, the intensities of the peaks at 90° and 270° are equivalent. PM3-PIF2 and PM3-PIF3 both give a broad distribution with two maxima centered at 90° and 270° . It is worth noting that, the probability of finding configurations corresponding to a value of ω_1 equal to 0° is not null with those Hamiltonians. ff03-SPC/E predicts a similar distribution but much more structured than it is with the QM Hamiltonians. In the former case, the ethyl group is almost fixed during the simulation and just switched from one side of the plane to the other. This result is also in good agreement with the observed sharpness of the well in the ff03 PES (Figure 5.22). It is worth noting that, unlike PM3-PIF2 and PM3-PIF3, ff03 predicts the conformations corresponding to a value of ω_1 equal to 0° to be almost null.

We performed a similar analysis of ω_2 . The distribution of this dihedral angle is pre-

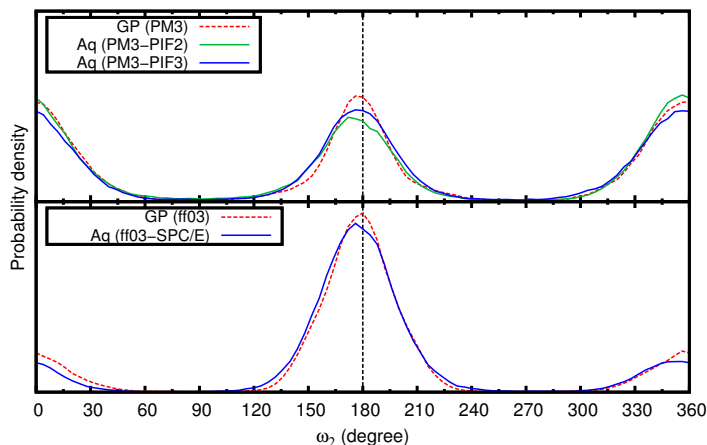


Figure 5.24: Comparison of the p-ethylphenol ω_2 dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.

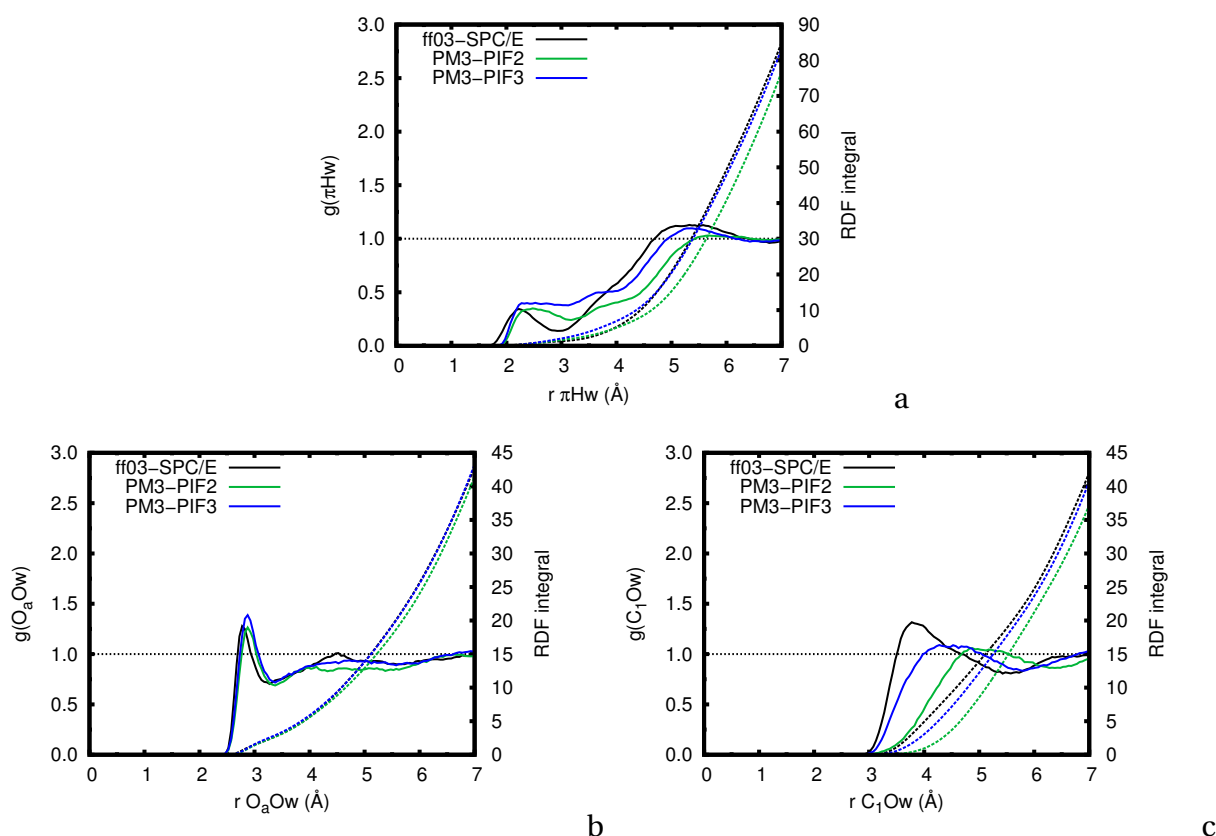


Figure 5.25: Comparison of radial pair distribution functions of p-ethylphenol in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF. Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (Ow) with respect to the ethanol oxygen atom (O_a). b: πHw . c: C_1Ow

sented Figure 5.24. The PM3 distribution appears to be more symmetrical than the ff03 one. Here again, the asymmetry is due to the relatively short sampling of the simulation. Nevertheless, this difference transcribes the lower rotational barrier predicted by PM3 compared to ff03. Unlike ω_1 , ω_2 does not seem to be affected by the solvent with any of the Hamiltonians chosen here.

Figure 5.25 shows the p-ethylphenol water RDFs for three sites, πHw , O_aOw and C_1Ow . In Figure 5.25a, the πHw pair distributions obtained with ff03, PM3-PIF2 and PM3-PIF3 are similar to the results discussed for the toluene molecule. Only the intensity of the first peak is slightly smaller, which is consistent with a larger steric effect, due to the ethyl substituent.

The O_aOw distributions (Figure 5.25b) obtained from the ff03-SPC/E, PM3-PIF2 and PM3-PIF3 simulations are close to those coming from the simulations of ethanol in water. Only the shape of the second solvation shell is affected. In the MM simulation, the intensity of the first peak with the MM force field is strongly lowered compared to the case of ethanol, which can be due to the larger size of the aromatic ring compared to the alkyl tail of ethanol. The integration of the first peak leads to 2.3, 2.6 and 2.9 water molecules for ff03-SPC/E, PM3-PIF2 and PM3-PIF3 respectively.

The C_1Ow RDF (Figure 5.25c) is the first example for the hydration of alkyl groups in which ff03-SPC/E and PM3-PIF3 do not predict the same result. ff03-SPC/E gives a similar

hydration sphere to the one of the methyl groups of isobutane, with a first peak located at about 3.8 Å. Compared to this value, PM3-PIF2 and PM3-PIF3 predict a first peak shifted by about 1.2 and 0.7 Å respectively. The difference between ff03 and PM3-PIF2 has already been discussed. The shift observed in the case of PM3-PIF3 is related to the flexibility of the ethyl group. Indeed, as we have seen earlier, this group is more flexible with the PM3-PIF3 Hamiltonian than it is with ff03-SPC/E. We have seen that the probability of finding $\omega_1 = 0^\circ$ or 180° is not null during the PM3-PIF3 simulation (see Figure 5.23). Such a conformation yields a shorter C_1C_m distance than in the case where $\omega_1 = 90^\circ$ or 270° (*i.e.*, about 2.8 and 3.3 Å, respectively). In the case of the PM3-PIF3 simulation, the aromatic ring repels the water molecules when $\omega_1 = 0^\circ$ or 180° leading to an averaged position of the peak located at a larger distance than it is during the MM simulation.

Trimethylamine. Trimethylamine bears three hydrophobic methyl groups, which surround one nitrogen atom that is able to form a hydrogen bond with one water molecule. Considering the above results about hydrophobic hydration, we shall not consider the PM3-PIF2 Hamiltonian for this analysis. The strength of the Hbond has been computed and found to be between 6 and 7 kcal/mol depending on the level of theory.[222, 263] Such a strong interaction is expected to be present in the dynamics of the condensed phase. Considering the geometry of the trimethylamine molecule, we can also expect that the interaction of the methyl groups with water will be equivalent to that in the case of the isobutane molecule. In Figure 5.26a and b we report the RDFs of the trimethylamine in water from our SEBOMD and MM-MD simulations for the NHw and COw pairs, respectively. Rozzio *et al.* have performed a reparametrization of the OPLS-AA force field, based on both theoretical and experimental results, dedicated to the hydration of amines, from amonia to trimethylamine showing good agreement with experimental results of hydration free energy.[264] Their results for the NHw pair along a simulation with TIP4P water molecules are also reported Figure 5.26a.

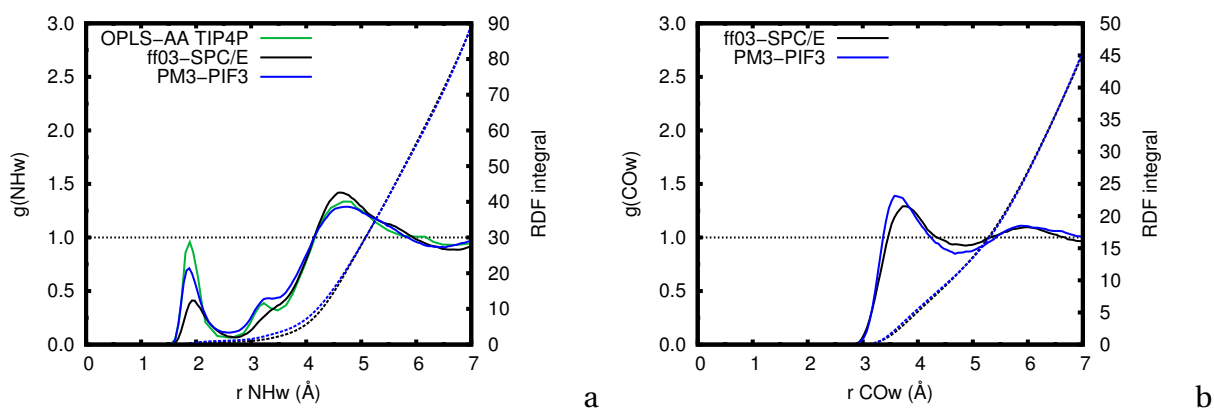


Figure 5.26: Comparison of radial pair distribution functions of trimethylamine in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF Dashed lines: RDF integral. a: pair distribution of water hydrogen atoms (Hw) with respect to the trimethylamine nitrogen atom. The OPLS-AA TIP4P results have been obtained from Ref. [264]. b: pair distribution of water oxygen atoms (Ow) with respect to the carbon atoms of the trimethylamine methyl groups.

From Figure 5.26a, we can see that all three selected models predict a first peak for the NHw interaction at 1.88, 1.94 and 1.84 Å with OPLS-AA, ff03 and PM3-PIF3, respectively. The intensity of this peak is slightly higher with OPLS-AA than it is with PM3-PIF3. However, the integration of the PM3-PIF3 peak leads to 1.09 Hbonds which is identical to the value reported by Rozzio *et al.*[264] The intensity of this peak and its integration are much lower with ff03, leading to 0.83 Hbonds. Similar results for the position of this first peak (*i.e.*, 1.7 Å) have been reported in the literature for methylamine in a box of 62 water molecules using CPMD with periodic boundary conditions.[265]

We can also identify in Figure 5.26a a second peak at around 3.1 Å for both OPLS-AA and PM3-PIF3 results, which is absent in the case of ff03 (*i.e.*, only a shallow shoulder appears at this distance). This peak can be assigned to the second hydrogen atom of the water molecule interacting with the nitrogen atom. This is consistent with a strong hydrogen bond as predicted from *ab initio* calculations,[222, 263] resulting in a water molecule almost fixed during the simulation. To conclude with the NHw interaction, we can also notice that the second solvation shell is predicted to be at the same distance by all the three models. This peak is defined by the hydrophobic part of the molecule which involves alkyl-water interactions. This observation is confirmed by the analysis of the COW RDF in Figure 5.26b, which shows that ff03-SPC/E and PM3-PIF3 predict a similar shape of the solvation shell around the methyl groups. We note that the shape and the position of this solvation shell is consistent with that of the isobutane molecule, as discussed in Subsection 5.2.1.

Amides. We selected three molecules containing an amide group: formamide (FA), propanamide (PA) and N-methylacetamide (NMA). In the case of NMA, only the *trans* conformer was considered, since it is the most stable in the gas and in the condensed phase.[74] AM1 and PM3 are known to present deficiencies in predicting the correct mesomery of peptide bonds.[74, 176, 266] Indeed, using those Hamiltonians, the geometry of the nitrogen atom is predicted to be pyramidal instead of planar. Such a misbehavior is of dramatic importance when dealing with biological compounds, since the geometry of the peptide bond dictates the secondary structure of any polypeptide. To overcome this issue, it is common

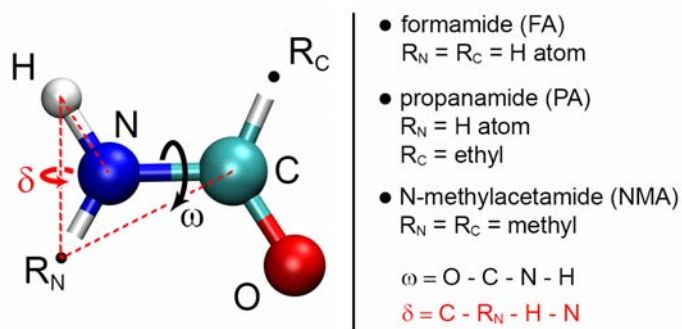


Figure 5.27: Labels definition for the three amide molecules.

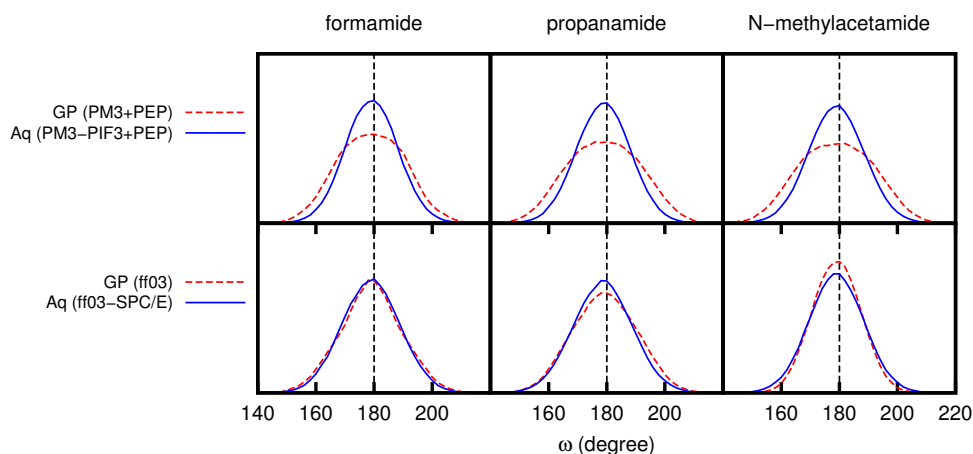


Figure 5.28: Distribution of the ω dihedral angle along the gas and the condensed phase dynamics. Comparison between formamide, propanamide and N-methylacetamide. For PM3 and PM3-PIF3, the peptidic correction (PEP) was applied.

to use an additional PEPTidic correction (PEP) to enforce the planar geometry of the peptide bond.[176] If not otherwise stated, the PEP correction was applied to each system containing at least one amide group.

We first examine the solvent effects on the geometry. To this end, we define in Figure 5.27 one dihedral angle (ω) and one improper angle (δ). The former is related to the *cis/trans* conformation of the peptide bond while the latter gives us information about the out of plane angle of the N-H bond. The value of the ω dihedral is not expected to change along the simulation, considering the large value of the *cis/trans* rotational barrier. Indeed, it has been reported for NMA in the literature to be about 15-20 kcal/mol in the gas phase and enhanced by about 2-3 kcal/mol in aqueous solution (see Ref. [74] for a recent review).

In Figure 5.28, we present the distribution of the ω dihedral along the gas and the condensed phase simulations for each of the three amides. We also show in this Figure a comparison between the MM force field and the QM approach. The first observation that can be drawn from these plots is that, in all cases, the value of ω is distributed around 180° . As expected, the conformation of the peptide bond does not change during the simulation. Another observation is that, using the ff03 force field, the presence of the solvent does not affect the shape of the distribution. In other words, the height of the rotational barrier is not changed, unlike it is suggested by experimental results. To the opposite, when using the PM3-PIF3 Hamiltonian, the distribution of ω is much sharper in solution than it is in the gas phase. The height of the barrier is thus modified by the solvent in a way which is consistent with experimental observations.

The same analysis was performed for the δ improper angle. Ingrosso *et al.* performed a SEBOMD simulations of NMA in the gas and in the condensed phase using the PM3-PIF2 Hamiltonian without the PEP correction.[74] They reported distributions of δ presenting two separated maxima in the gas phase (*i.e.*, -40° and $+40^\circ$) and a much broader distribution in

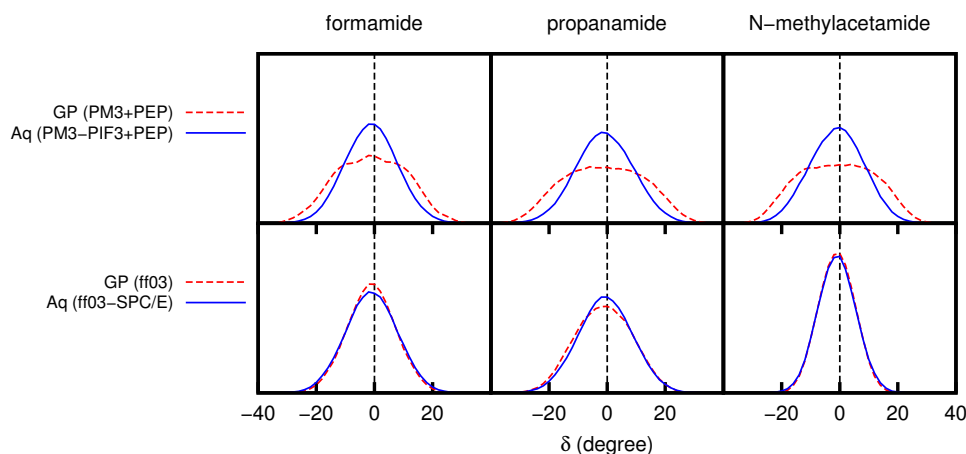


Figure 5.29: Distribution of the δ improper angle along the gas and the condensed phase dynamics. Comparison between formamide, propanamide and N-methylacetamide. For PM3 and PM3-PIF3, the peptidic correction (PEP) was applied.

the condensed phase, however still presenting two maxima (*i.e.*, -25° and $+25^\circ$). In a polar environment, the zwitterionic form of the peptide bond is favored. This corroborates the fact that the geometry of the nitrogen atom environment tends to be planar in solution.

We present the distributions of δ coming from our simulations in Figure 5.29. At first glance, one can see that the PEP correction, applied to PM3 and PM3-PIF3, actually leads to a distribution of δ presenting solely one maximum centered around 0° for each of the three amides considered here. It is worth noting that again, in the case of ff03, the solvent does not affect the distribution of δ . In solution, the zwitterionic form of the amide group is favored and the planarity of the nitrogen environment is enforced. This phenomenon appears to be well reproduced by the PM3-PIF3 Hamiltonian, since in each case, the distribution of δ is much sharper in solution than it is in the gas phase.

The variation of other geometrical features of the three amide molecules was also followed along the dynamics. As it has been observed by Ingrosso *et al.*, the bond lengths are not modified in the condensed phase compared to the gas phase when using a classical force field.[74] For our SEBOMD simulations of NMA, we observed slight modifications of the C-O, C-N and N-H bonds (*i.e.*, $+0.02$, -0.02 and $+0.01$ Å, respectively) compared to the corresponding values in the gas phase (*i.e.*, 1.23, 1.41 and 1.00 Å, respectively). This observation is due to the zwitterionic character of the peptide bond that is enhanced in aqueous solutions. Similar results were obtained for the two other amide molecules. These results are in good agreement with those reported by Ingrosso *et al.* as well as those from Gaigeot *et al.*[80] The latter performed a similar study of NMA in box of 50 water molecules with PBC, using DFT-based Car-Parrinello molecular dynamics.

We shall now analyze the structure of the solvation shell around specific sites of the three amides selected here. Since the solvation shell around alkyl groups has been extensively discussed in the present work, we will only focus on three atoms shared by all FA, PA and NMA

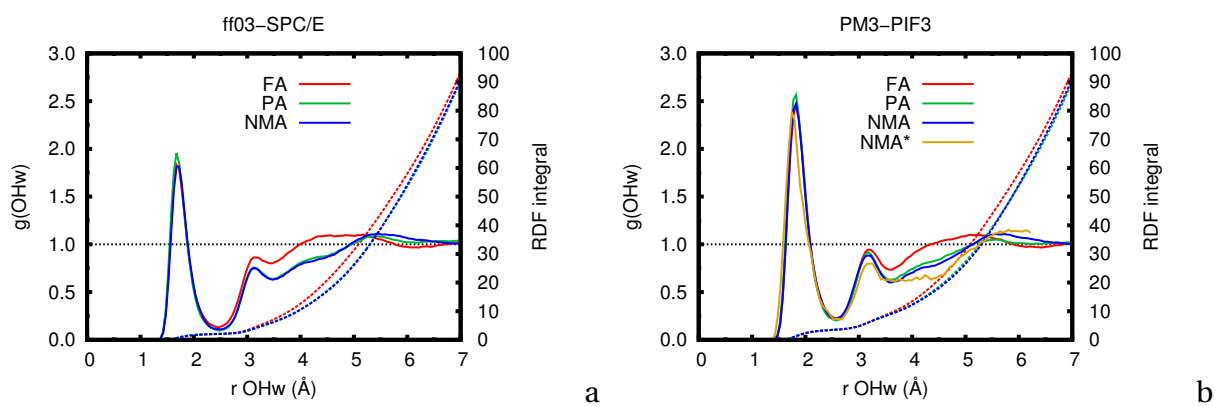


Figure 5.30: Comparison of OHw RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.

(i.e., the O, N and H atoms). On Figure 5.30a and b, we present a comparison of the OHw RDF of each compound obtained from MM-MD and SEBOMD simulations, respectively. We also report in these plots the unpublished RDF of NMA obtained with PM3-PIF2, without PEP correction, by Ingrosso *et al.* as a comparison.[74] As a first observation, the results obtained from MM-MD and SEBOMD simulations are consistent. For all of the three amides, ff03-SPC/E and PM3-PIF3 predict a peak of large intensity centered around 1.70 and 1.80 Å, respectively. The result obtained with the PM3-PIF2 Hamiltonian shows a slight shift of this peak for NMA about -0.05 Å compared to PM3-PIF3. The number of hydrogen bonds around the carbonyl group obtained by integrating the RDFs is about 2.06 and 3.20 for ff03-SPC/E and PM3-PIF3, respectively.

Chalmet *et al.* performed a hybrid QM/MM study of the formamide molecule in a box of 216 TIP3P water molecules with PBC.[267] In this work, the authors considered the solute using two DFT methods. They found a position of this first peak between 1.77 and 1.83 Å, and a number of Hbonds lying between 2.0 and 3.0, depending on the QM level. More recently, Gaigeot and coworkers have performed a deep analysis of the dynamical properties of NMA

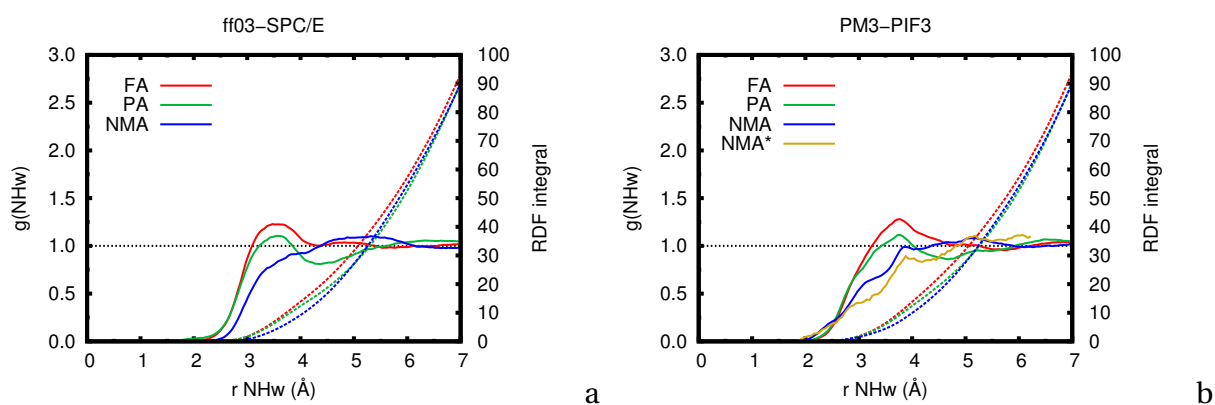


Figure 5.31: Comparison of NHw RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.

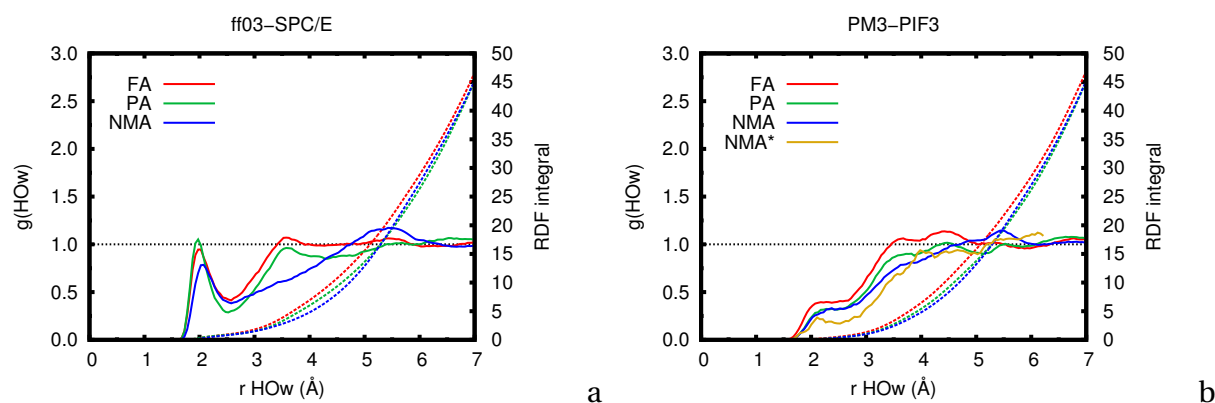


Figure 5.32: Comparison of HOw RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.

in a box of 50 water molecules with PBC with CPMD.[80] The authors reported the position of the first OHw peak to be at 1.80 Å, and 2.0 Hbonds formed by the carbonyl group of the *trans*-NMA molecule.

Because SEBOMD simulations allow a larger sampling compared with CPMD, it is possible to discuss the shape of the second solvation shell as well. In Figure 5.30, we observe that both MM-MD and SEBOMD simulations predict a similar shape of the RDF after 3 Å, for PA and NMA. Only the RDF of FA differs in this region with a slightly higher intensity, consistently with a smaller R_C group bounded to the carbonyl carbon atom.

Figure 5.31 displays the NHw RDF obtained from MM-MD and SEBOMD. Because the lone electron pair of the nitrogen atom is delocalized along the peptide bond, a strong Hbond between the N atoms and the Hw atoms of water is not expected. The RDFs obtained from MM-MD and SEBOMD simulations do not show any peak at a distance characteristic of a hydrogen bond. Instead, we observe a broad peak located between 3 and 4 Å, for ff03-SPC/E and PM3-PIF3 in the case of FA and PA. In the RDF of NMA no peak is present, which is consistent with the presence of a methyl group place at the R_N position (See Figure 5.27). For the FA molecule, our results are in good agreement with the observations of Chalmet *et al.*

In Figure 5.32a and b, we report the HOw RDFs obtained from MM-MD and SEBOMD simulations, respectively. Unlike the other interactions discussed above, ff03-SPC/E and PM3-PIF3 (as well as PM3-PIF2) predict here a different picture of the solvation shell around the H atom of the three amides. While ff03-SPC/E predicts a well structured peak at approximately 2.0 Å for each of the three molecules, the corresponding RDF obtained with the QM approaches only shows a shallow shoulder. In similar studies, about FA and NMA, Chalmet *et al.* as well as Gageot *et al.* remarked a peak in good agreement with the ff03-SPC/E results.[80, 267] A deeper investigation of the interaction between the nitrogen atom of an amide and the hydrogen atom of water is indeed required to confirm these observations and different strategies could be considered. A specific parameterization of the core-core repul-

sion function could be proposed, following the PIF3 methodology (see Chapter 4). However, considering the issues related with the electronic environment of an N atom involved in a peptide bond using PM3, a modification of the electronic parameters could also be envisaged.

The results obtained for the geometrical features and solvent structure of amide compounds in water are in good agreement with other theoretical works. The use of the PEP correction yields a correct geometry of the nitrogen atom and is very promising for further applications on larger polypeptides. Finally, most of the interactions with water are qualitatively well reproduced by the PM3-PIF3 Hamiltonian, compared with other theoretical results.

5.3.2 Electronic properties

It has been shown that the CM1 model leads to a good prediction of the N-methylacetamide (NMA) dipole moment with respect to gas phase experimental results.[74] To allow a consistent comparison with all the other molecules considered in this work, we shall discuss here only the results obtained with CM1. The same protocol as the one used for hydrophobic compounds (see Subsection 5.2.2) was followed to obtain the total charge and the instantaneous dipole moment of each hydrophilic solute.

Figure 5.33 presents the distribution of the total charge and dipole moment of the solute using CM1 charges. Concerning the dipole moment, a comparison with gas phase results is also plotted in Figure 5.33b. The average of those electronic properties are summarized in Table 5.5.

From Figure 5.33a, we can observe that all the six hydrophilic molecules experience a charge transfer from the solute to the solvent that fluctuates along the simulation time. The charge transfer for all the molecules is consistent with the strength of the different hydrogen

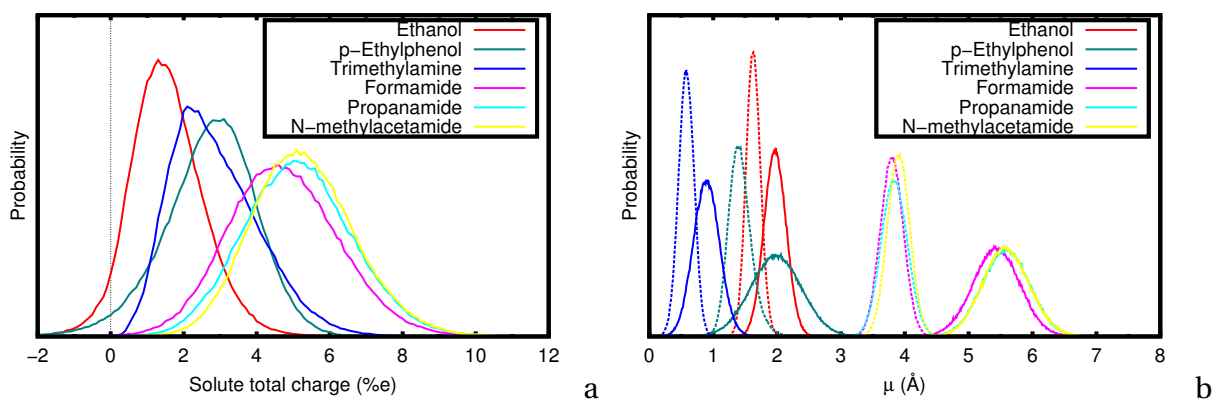


Figure 5.33: Distribution of instantaneous electronic properties of the six hydrophilic molecules along the PM3-PIF3 SEBOMD simulations. a: CM1 total charge of the solute in condensed phase. b: Solute dipole moment computed from CM1 atomic partial charges. Comparison between gas phase (dotted line) and condensed phase (plain line).

Table 5.5: Averages of the instantaneous total charge and dipole moment (in parenthesis, the difference with respect to gas phase) of the hydrophilic solutes during the PM3-PIF3 condensed phase simulation. The CM1 model was used to derive the charges.

Molecule	Total charge (%e)	Dipole moment (Debye)
Ethanol	+1.6	1.98 (+0.35)
p-Ethylphenol	+2.8	1.99 (+0.56)
Trimethylamine	+2.9	0.90 (+0.32)
Formamide	+4.8	5.44 (+1.64)
Propanamide	+5.2	5.59 (+1.75)
N-methylacetamide	+5.3	5.59 (+1.68)

bonds involved in the solute water interaction. The p-ethylphenol molecule has a charge in solution comparable to that of benzene and toluene (see Tables 5.5 and 5.3 for a comparison). Notice that the distribution of the trimethylamine total charge is not symmetrical. Thus, a relatively large difference exists between the maximum of this distribution and the average presented in Table 5.5. The three amides experience the largest charge transfer among all the molecules treated up to now (*i.e.*, between 4.8 and 5.3 %e). The value obtained for NMA is slightly lower than the one obtained by Ingrosso and coworkers using the PM3-PIF2 Hamiltonian (*i.e.*, 6.0 %e).

The dipole moment of the different molecules is also affected by the solvent (see Figure 5.33b and Table 5.5). The ethanol dipole moment increases from 1.63 to 1.98 Debye going from the gas phase to solution, which is consistent with the work of van Erp *et al.*[259] However, the intensity of the shift was predicted to be about 1.2 Debye in the latter study. Comparable results are obtained for the p-ethylphenol while the trimethylamine dipole moment in solution is quite small. Nevertheless, the latter molecule experiences a relatively large augmentation of its dipole moment compared to the gas phase (*i.e.*, it is roughly doubled in solution). Finally, the dipole moment of the three amide molecules are the largest among the solutes considered here (*i.e.*, between 5.44 and 5.59 Debye in solution). In the case of NMA, we obtained a dipole moment in solution of 5.59 Debye. This value is consistent with the HF/6-31G(d)-RISM results by Du *et al.* (5.79 Debye). However, Ingrosso *et al.* and Gaigeot *et al.* found a larger value from a SEBOMD PM3-PIF2 (6.63 Debye) and a CPMD (6.96 Debye) simulation, respectively.

As we discussed in Section 5.2.2, a direct comparison of the absolute values of electronic properties obtained from different methods and charge models is not straightforward. However, the trend observed from our SEBOMD simulations is consistent with other methods and reflects the expected properties of the selected hydrophilic molecules in water.

5.3.3 Vibrational properties

The infrared spectrum of hydrophilic compounds is particularly sensitive to direct solute-solvent interactions (*e.g.*, hydrogen bonds, π - π interactions ...). Compared to hydrophobic compounds, more extensive experimental and theoretical work has been devoted to understand the effect of those interactions on the vibrational properties of small hydrophilic molecules. However, presenting a review of these studies is out of the scope of the present work. We shall refer to relevant results of the literature while discussing the infrared spectra computed from our simulations.

Ethanol and p-ethylphenol. Alcohol and phenol molecules contain a hydroxyl group (OH) giving rise to hydrogen bonds with surrounding water molecules. This interaction affects the vibrational behavior of the OH group and can be seen on the infrared spectrum of the targeted compound.

We present the IR spectra computed from our simulations in Figures 5.34a and b for the ethanol and the p-ethylphenol molecules, respectively. We recall that the peak assignment follows a procedure based on the decomposition of the VDOS (see Subsection 5.2.3). In the IR spectra obtained from SEBOMD simulations, we can observe that, for both molecules, only the vibrational modes in the frequencies higher than 2500 cm^{-1} are affected by the solvent. At about 3000 cm^{-1} , the bands of the different C-H stretches appear to be red shifted in a way that is consistent with what has been discussed for isobutane and toluene in the previous section. In the gas phase, the peak corresponding to the O-H stretch is positioned around 3960 cm^{-1} for the alcohol and the phenol molecules while it is broadened and red shifted by about $150 \pm 50\text{ cm}^{-1}$ and $200 \pm 100\text{ cm}^{-1}$ in the condensed phase for the ethanol and the p-ethylphenol molecule, respectively. The broadening experienced by this band in both cases is typical of a hydrogen bonded hydroxyl group. These results are consistent with experimental and theoretical observations. Although the absolute position of the O-H peak is overestimated by about 300 cm^{-1} in the present work,[259, 268–271] the solvent effect on

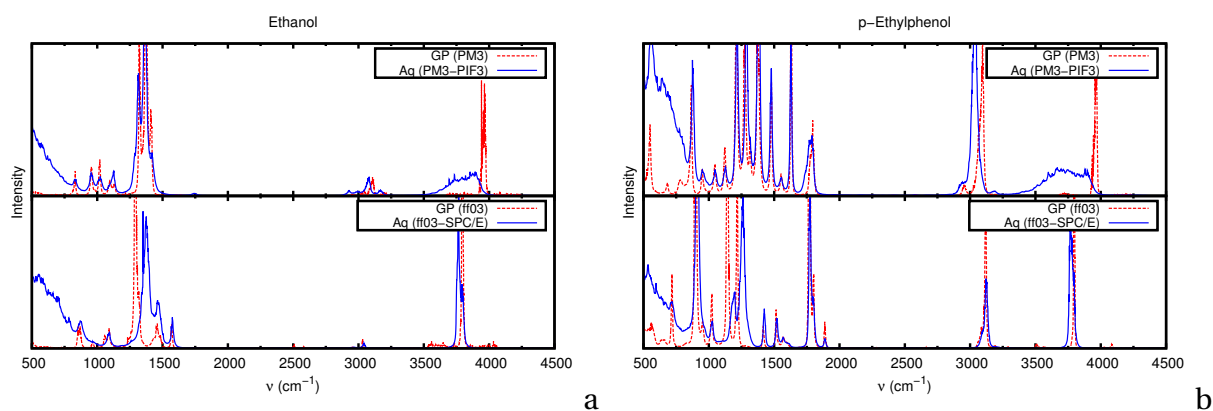


Figure 5.34: Comparison of infrared spectra of ethanol (a) and p-ethylphenol (b) obtained with PM3-PIF3 and ff03. For each molecule, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

the IR spectrum is well reproduced by our SEBOMD simulations.

From the MM simulations, we observe a slight red shift of the O-H peak from the gas to the condensed phase and no broadening. The solvent has clearly no effect on this part of the spectrum when using a classical force field. It is worth noting that, unlike the case of SEBOMD simulations, the MM force field predicts a difference in the low frequencies region of the spectrum (*i.e.*, between 500 and 2500 cm^{-1}) between the gas and the condensed phase. For the ethanol molecule, one peak positioned at 1290 cm^{-1} is blue shifted by 85 cm^{-1} , and correspond to the COH bend vibrational mode. In the case of the p-ethylphenol, two peaks are affected by the solvent in this region. Those peaks, positioned at 1140 and 1215 cm^{-1} , involve a combination of the COH bending and the CCH bending at the meta position of the hydroxyl group. Both peaks are blue shifted by 50 cm^{-1} . This phenomenon has been observed by van Erp and coworkers in their CPMD simulations of methanol in water (*i.e.*, a shift about 50-100 cm^{-1} of the COH bending vibrational mode).[259] To the opposite, the experimental work by Max *et al.* on the 1-propanol molecule in water does not show any modification in the position of this peak.[269]

Trimethylamine. To our knowledge, there is no experimental nor theoretical study that has been devoted to the effect of solvent on the IR spectrum of amines allowing a direct comparison with our present results. The results reported in the literature mainly deal with primary or secondary amines as well as their protonated form. In Figure 5.35 we show the IR spectra obtained from our simulations of the trimethylamine molecule. The spectrum obtained from the ff03-SPC/E simulation in the gas and in the condensed phase show a similar line shape. No effect of the solvent on the vibrational properties of the trimethylamine molecule is predicted by this MM force field. By analyzing the IR spectra obtained from the

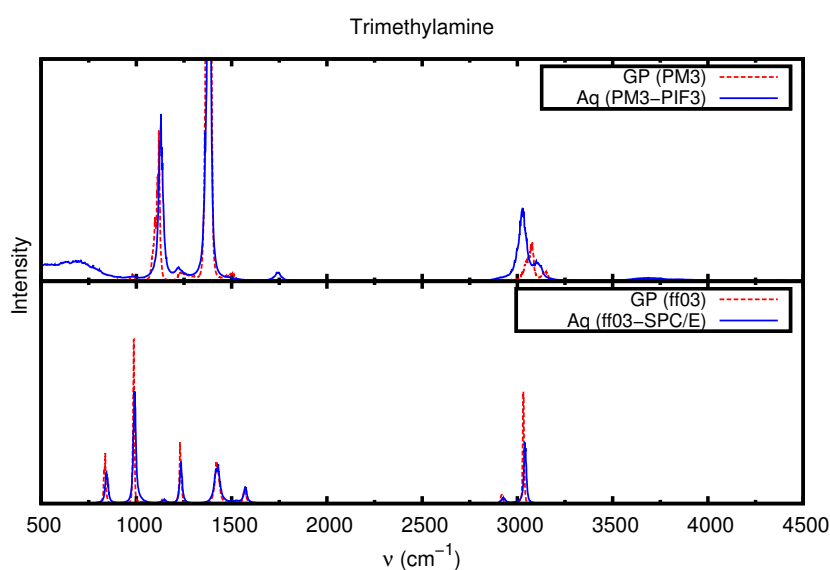


Figure 5.35: Comparison of infrared spectra of thrimethylamine obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

SEBOMD simulations, we observe that the C-H stretching modes of the methyl groups are affected when moving from the gas phase to the condensed phase, in a similar way than it has been discussed for the isobutane molecule (*i.e.*, slightly red shifted in solution). In addition, we notice that the peaks identified in the previous sections (*i.e.*, ν_1^* and ν_2^* at about 1700 and 3500 cm^{-1} , respectively) still appear on the IR spectrum computed from the SEBOMD simulation in solution. Finally, it is worth noting that, despite the hydrophilic character of the nitrogen atom of the trimethylamine, which forms a hydrogen bond with water (see the corresponding RDF in Figure 5.26), no strong effect of the solvent can be observed here, compared to the case of the isobutane molecule (see Subsection 5.2.3).

Amides. The N-methylacetamide molecule is the simplest model of a peptide bond, and thus, it has been widely studied both from an experimental and from a theoretical point of view.[74, 80, 227, 272, 273] In particular, Gaigeot *et al.* as well as Ingrosso *et al.* recently reviewed this topic by means of CPMD and SEBOMD simulations, receptively.[74, 80, 227] Four vibrational modes (illustrated for the *trans*-NMA molecule in Figure 5.36) are of great interest and characteristic of the amide functional group:[272]

- The amide I (AmI) mode is mainly composed by the C-O carbonyl stretch and found experimentally at 1714-1731 cm^{-1} in the gas phase.
- The amide II (AmII) vibration involves a large contribution of the CNH bend, as well as a smaller contribution of the CN stretch. This band has a frequency of about 1500 cm^{-1} in the gas phase.
- The amide III (AmIII) mode is similar to AmII but with inverted contributions (*i.e.*, the main contribution is due to the CN stretch). The associated frequency in the gas phase is about 1300 cm^{-1} .
- The NH stretch (AmA) is found in the high frequencies region, *i.e.*, around 3476 cm^{-1} in the gas phase.

All those four vibrational modes are affected by the presence of water as follows:

- AmI and AmA are known to experience a red shift of about 90 and 130 cm^{-1} , respectively.
- AmII and AmIII are both blue shifted in solution by about 80 and 60 cm^{-1} , respectively.

The formamide and propanamide molecules share similar vibrational properties with NMA, whereas the AmII and AmIII modes are modified because of the lack of substituent on the nitrogen atom.

In the case of NMA, the work by Ingrosso *et al.* showed that classical force fields that do not include polarization are not able to reproduce the experimentally observed shifts of the four bands discussed above.[74] In that work, the authors also showed that, despite an error in the absolute position of the bands, SEBOMD simulations using the PM3-PIF2 Hamil-

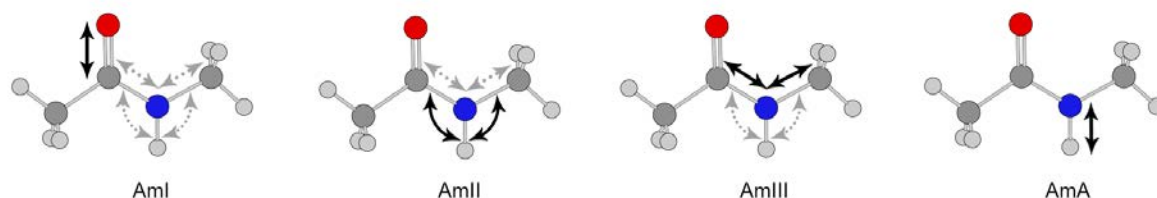


Figure 5.36: Schematic representation of the four amide vibrational modes in the *trans*-N-methylacetamide (NMA) molecule. The plain black arrows and dotted gray arrows show the main and secondary contributions, respectively.

tonian reproduce very well the solvent effects on the vibrational properties of NMA. Since these effects have been extensively discussed in the latter paper, especially for the low frequencies region, it will not be detailed here. Only a comparison between the three selected amide molecules will be made, with more emphasis on the different C-H and N-H stretching modes.

In Figure 5.37, we present the IR spectra of the three amide molecules computed from our SEBOMD and MM-MD simulations. As expected, when comparing the gas and the condensed phase results, it appears that the MM force field cannot reproduce any of the solvent induced IR shifts discussed above. In Ref. [74], the authors showed that the ff03 force field does not reproduce the correct frequency order between AmI and AmII, and this observations is also true in the present work for all the three amides. In the case of NMA, the PM3-PIF3 results presented here show a very good agreement with the work by Ingrosso *et al.* A slight difference compared to their results can be seen for the position of AmII and AmA which are respectively shifted by +36 and +25 cm^{-1} in our work. This difference can be attributed to the use of the PEP correction in the present study, which was not included in the simulations of Ref. [74].

For the three amide molecules, the solvent effects on the low frequencies region are well reproduced by our PM3-PIF3 simulations. In the high frequencies region (*i.e.*, above 2750cm^{-1}), we can observe the effect of water on the C-H and N-H stretching modes. A zoom on this particular region of the spectrum is presented in Figure 5.38. In the gas phase IR spectrum of FA and PA, two peaks are present at 3520 and 3600 cm^{-1} corresponding to the asymmetric (ν_{NH}^a) and symmetric (ν_{NH}^s) N-H stretching modes, respectively. The assignment of these peaks is possible by using the VDOS decomposition into atomic contributions, the nitrogen atom participating more to the asymmetric mode than it does to the symmetric one. In both cases, these peaks are red shifted in solution and form a single broad band centered around 3450 cm^{-1} . This last result is consistent with experimental observations.[274] In the case of NMA, as expected, only one N-H peak is observed. This peak also suffers a red shift in solution, consistent with other theoretical and experimental results.[74, 80, 227, 272, 273]

In the C-H stretch region, we can observe the effect of water on the different substituents (*i.e.*, R_N and R_C in Figure 5.27). In the case of FA, the C-H stretch appears at 2940 cm^{-1} and

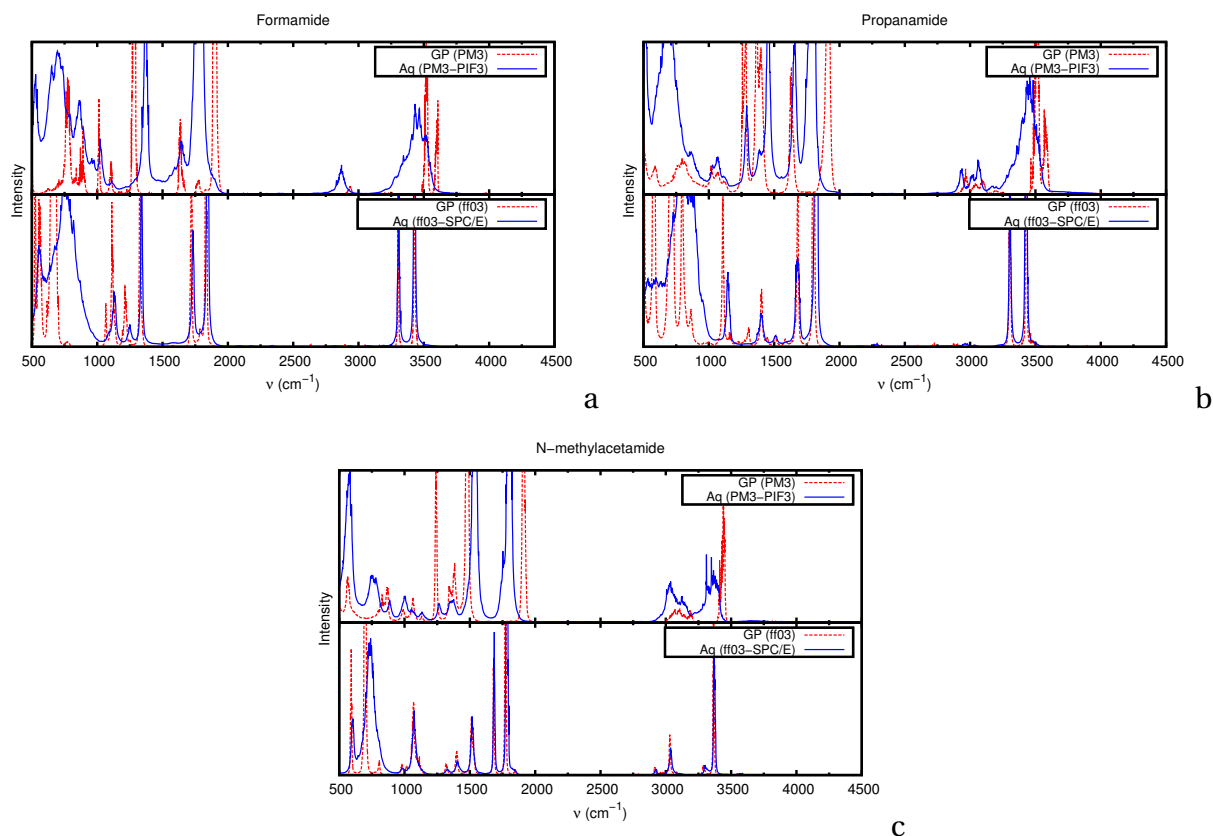


Figure 5.37: Comparison of infrared spectra of formamide (a) propanamide (b) and N-methylacetamide (c) obtained with PM3-PIF3 and ff03. For each molecule, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.

is red shifted in solution by about 80 cm^{-1} . Notice that this shift is twice as large than in the case of alkane compounds. For PA, two different C-H stretching modes are present between 2900 and 3100 cm^{-1} , with their respective symmetric and asymmetric contributions: ν_{CH}^1 and ν_{CH}^2 corresponding respectively to the C-H stretching in the methyl and methylene groups of the propanamide molecule. Although it is possible, from the VDOS decomposition, to see that ν_{CH}^2 has a lower frequency than ν_{CH}^1 , one cannot assign unequivocally the respective symmetric and asymmetric modes. Nevertheless, one can see that those vibrational bands experience a small red shift in solution. Finally, the spectra of NMA show a broad peak between 3000 and 3200 cm^{-1} which is also red shifted in the condensed phase. This band is due to the C-H symmetric and asymmetric stretching modes of the two methyl groups in NMA (*i.e.*, $\nu_{\text{CH}}^{\text{N}}$ and $\nu_{\text{CH}}^{\text{C}}$ corresponding to the methyl group of the nitrogen and to the one of the carbon atom, respectively). Here again, a detailed assignment of this broad band cannot be performed in a straightforward way. Nevertheless, we identified the asymmetric contributions of $\nu_{\text{CH}}^{\text{N}}$ and $\nu_{\text{CH}}^{\text{C}}$ to correspond to a frequency of 3060 and 3100 cm^{-1} in the gas phase, respectively. Even though the absolute values obtained in the present work are slightly higher than experimental results, the ordering and the frequency difference between the two vibrational modes are in good agreement with the observations made by Ataka *et al.*[272]

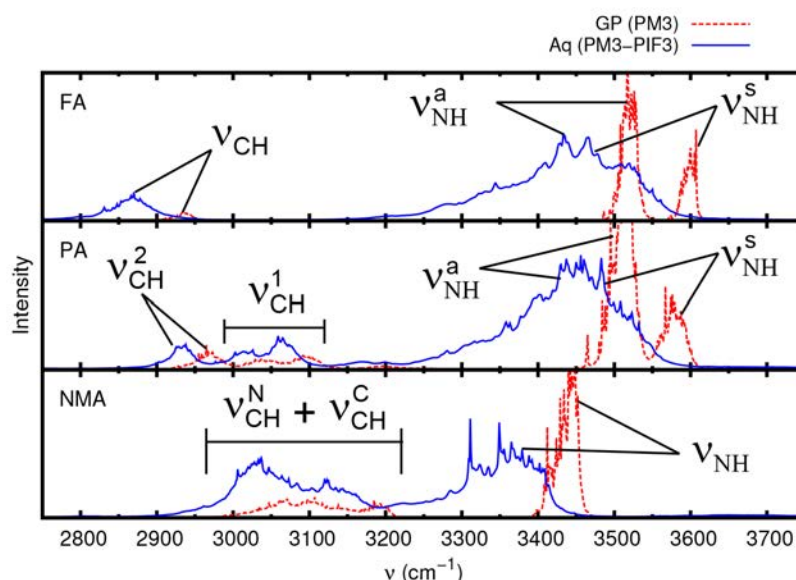


Figure 5.38: Comparison of infrared spectra of formamide (top panel) propanamide (middle panel) and N-methylacetamide (bottom panel) obtained with PM3-PIF3. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line. Peak assignment was made from the VDOS decomposition into atomic contributions.

The effect of water on the vibrational properties of amide compounds is qualitatively well reproduced by the PM3-PIF3 Hamiltonian compared to other works. We also show here that the different substituents are also affected by the solvent in a consistent way with the discussion about hydrophobic compounds.

5.3.4 Summary

We have shown in this section that SEBOMD simulations using the PM3-PIF3 Hamiltonian lead to results in good agreement with other experimental and theoretical works for hydrophilic molecules. We observed that the PM3-PIF3 Hamiltonians corrects the misbehavior of PM3-PIF2 while keeping the previously well defined interactions unchanged. This makes the PM3-PIF3 method a very good candidate for the simulation of biological compounds in aqueous solution. However, attention must be paid on some particular interactions such as the hydrogen bond between H atoms bonded to the nitrogen atom of an amide group and the oxygen atom of water.

5.4 Peptide model

Proteins dynamical properties play a key role in biological functions. Such macromolecules are built from an assembly of amino-acids connected to each other *via* a peptide bond (*i.e.*, an amide function). This sequence of peptide bonds represents the backbone of the protein and its flexibility directs the folding into a particular secondary structure of the polypeptide. The conformational properties of a protein are due to a balance between stabilizing and repulsive inter/intra-molecular interactions. Of course, the interaction of the side chains and backbone of the protein with water plays a key role in this balance.[234]

In the previous Section, we have studied the N-methyl acetamide molecule as the simplest peptide bond model. In order to validate the SEBOMD methodology and to test the application of the PM3-PIF3 Hamiltonian to biological systems, the next step is represented by the study of more realistic systems, for which conformational equilibria can be analyzed. The alanine dipeptide (denoted as 2Ala in the following) represents the most commonly used model of polypeptide containing two peptide bonds.[225] We propose in what follows to investigate the dynamical properties of this molecule in the gas and in the condensed phase, by means of SEBOMD simulations. We shall first discuss the conformational characteristics of 2Ala in gas phase as described by the PM3 Hamiltonian. Thus, we will analyze our molecular dynamics simulations of 2Ala by focusing on its structural and electronic features, on the structure of its hydration shell and finally, on its vibrational properties.

5.4.1 General aspects on the alanine dipeptide conformation

The alanine dipeptide (represented in the top panel of Figure 5.39) is composed by three residues that can be identified as: an acetyl group (ACE), an alanine aminoacid (ALA) and an N-methyl group (NME). Disregarding the rotation of the three methyl groups (*i.e.*, one borne by each of the three residues), the molecule is quite rigid, with the exception of two particularly important degrees of freedom. The two dihedral angles ϕ and ψ , depicted on the top molecule in Figure 5.39, constitute some of the most important structural characteristic of polypeptides and proteins in general. In 1963, Ramachandran *et al.* have rationalized the study of polypeptides conformational properties by defining those angles and by drawing the well known Ramachandran map.[275] This map is a representation of the ϕ and ψ dihedral angles of each amino acid in a polypeptide. From simple steric considerations and an extensive comparison of polypeptides structural data, Ramachandran *et al.* defined “allowed” and “disallowed” regions (basins) of the map. The authors also identified in the allowed regions some particular structures that affect the shape of the protein secondary structure (*i.e.*, α -helix, β -strands *etc.*).

The configurational preferences of 2Ala have been extensively studied experimentally and theoretically. We will not give a review of these works here and shall refer to some recent

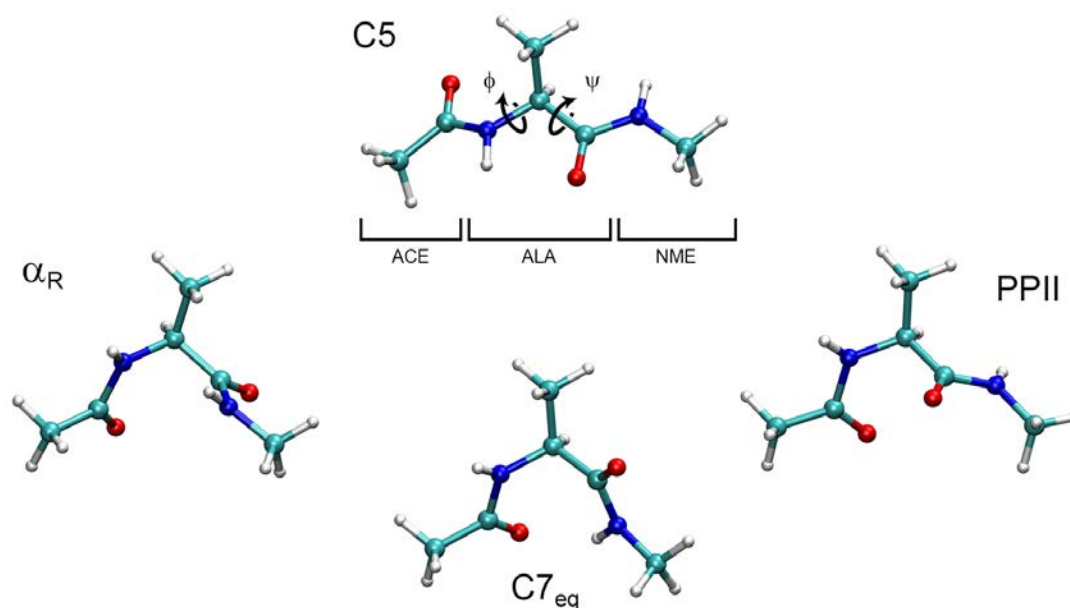


Figure 5.39: Schematic representation of the four main configurations of the alanine dipeptide and definition of the two dihedral angles (ϕ and ψ). $\phi = \text{C}^{\text{ACE}} - \text{N}^{\text{ALA}} - \alpha\text{C}^{\text{ALA}} - \text{C}^{\text{ALA}}$ and $\psi = \text{N}^{\text{ALA}} - \alpha\text{C}^{\text{ALA}} - \text{C}^{\text{ALA}} - \text{N}^{\text{NME}}$.

reviews.[225–227, 276] The most representative structures of 2Ala are depicted in Figure 5.39. In the gas phase, it is commonly admitted that the C7_{eq} ($\phi = -75^\circ$ and $\psi = 75^\circ$) configuration is the most stable and involves a seven membered ring closed by a hydrogen bond between the carbonyl O atom of ACE and the H atom bonded to the nitrogen atom of NME. Two other configurations are also favorable in the gas phase and belong to the β basin of the Ramachandran map: C5 (involving the formation of a five membered ring) and the poly-L-proline type II of helix (PPII, see Ref. [277] for more details) for $\phi = -170^\circ$ to -60° , and $\phi = 120^\circ$ to 170° . The α_{R} configuration ($\phi = -60^\circ$ and $\psi = -50^\circ$) belongs to the α basin and can also be found for the dipeptide in the gas phase. Other configurations are possible, such as α_{L} and C7_{ax} but less probable. In the condensed phase, although it is known that the preferred configurations of 2Ala are dramatically affected by the presence of water molecules, the precise picture of the Ramachandran map is still not well known and controversial.[276] However, it is commonly accepted that the PPII configuration, and in general the β basin, should be favored due to intermolecular hydrogen bonds with a neighbor water molecule.

Before starting the discussion about the effect of water on the conformational properties of our system, it is worth to investigate the gas phase Ramachandran map of 2Ala with the methods used in this work. To this end, we performed a semi-rigid scan of the potential energy surface (PES) of 2Ala as a function of ϕ and ψ . In the current implementation of Amber (Amber14), it is not possible to keep a given internal coordinate (*i.e.*, bond, angle or dihedral angle) constant. However, the “belly” methodology (*i.e.*, `ibelly=1` in the Amber input) allows the user to freeze the Cartesian coordinates of a given set of atoms and this is the procedure that we adopted here. For each couple of ϕ and ψ coordinates, the belly

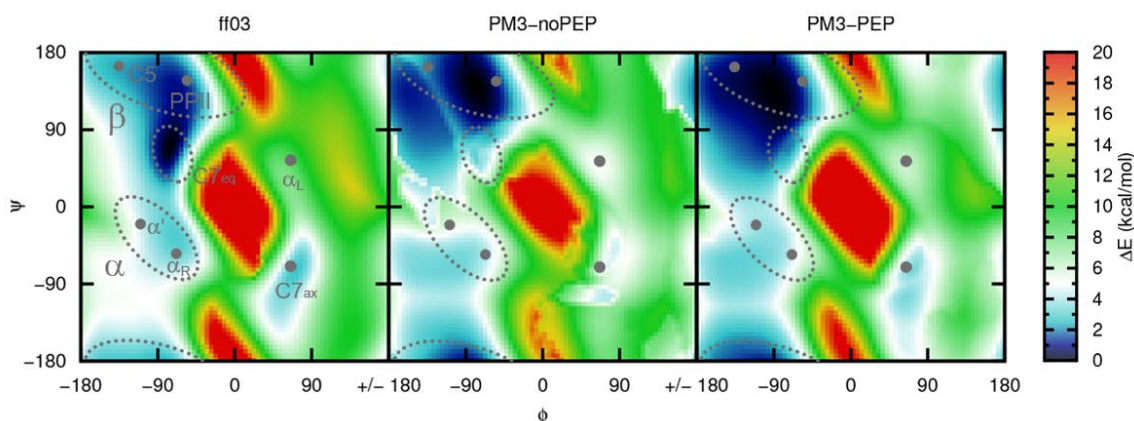


Figure 5.40: Potential energy surfaces of the isolated alanine dipeptide computed with ff03 (left panel) and PM3 without (middle panel) and with (right panel) peptidic correction. A representation of the different conformational regions as well as indicators of the corresponding structures are also represented.

procedure was applied to the five atoms which define those two angles: C^{ACE} , N^{ALA} , αC^{ALA} , C^{ALA} and N^{NME} . The reset of the peptide geometry was relaxed and minimized.

We present in Figure 5.40 the PES obtained using the ff03 force field and the PM3 Hamiltonian in the gas phase. We recall that the PM3 Hamiltonian is known to predict a pyramidal geometry of the nitrogen atom environment in amide groups. We shall discuss the effect of the peptidic correction (PEP) by comparing the results given by PM3 with and without applying this additional potential (*i.e.*, PM3-PEP and PM3-noPEP, respectively).

The first observation that can be derived from Figure 5.40 is that the PEP correction does change the shape of the PES predicted by PM3. Two main effects can be seen. First, considering the PM3-noPEP results, the β basin contains two separated wells, while a single broader well appears when applying the PEP correction. The second observation is that the PM3-PEP surface is much smoother than the one predicted by PM3-noPEP. In the latter case, many discontinuities appear on the map, that are directly related with the umbrella effect experienced by the nitrogen atoms (*i.e.*, the inversion of the improper angle ω , as defined for the amide molecules in Subsection 5.3.1). When comparing the QM surfaces with the ff03 one, it appears clearly that PM3-PEP and ff03 give a qualitatively similar picture of the map. In the region of positive ϕ values, the wells corresponding to the α_L and $C7_{ax}$ conformations are correctly reproduced by PM3-PEP and ff03 compared to other works,[225, 226, 276, 278] while it is not well defined in the case of PM3-noPEP. To our knowledge, the barrier predicted by PM3-noPEP in the β region has not been reported by other methods in the literature. In this β region, ff03 and PM3-PEP predict similar results, though the latter shows a larger well than the former. While ff03 correctly predicts the $C7_{eq}$ conformation to be the most stable in the gas phase, PM3 with and without the PEP correction barely present a minimum corresponding to this structure. Finally, PM3-PEP and ff03 predict a similar barrier between the β and the α basin, while, in this area, the surface computed using PM3-noPEP is located at a higher energy and presents some unexpected minima.

The Ramachandran map of 2Ala has been extensively investigated and several works compared the results obtained by different methods.[226, 279] In particular, Seabra *et al.* performed an extensive exploration of the configurational space of 2Ala in the condensed phase using the replica exchange sampling methodology.[226] The authors used some of the most common classical force fields devoted to biological studies and compared their performances with some semiempirical methods in a QM/MM framework. For the classical force fields, the authors showed that, even though such methods result from an *ad hoc* parametrization of biological systems, not all of them yield a good agreement with experimental data such as dipolar coupling constant and conformational distribution. Over the tested MM methods, the ff99SB and ff03 Amber force fields were the closest to the experience, though the latter appears to overestimate the population of the α basin. Among the QM approaches, the RM1 Hamiltonian appeared to yield the best results, while PM3 with peptidic correction predicted a correct picture of the Ramachandran map. Finally, in a recent joint experimental and theoretical study, Parchaňský *et al.* have shown a good agreement between the free energy surface of 2Ala obtained from Raman optical activity measurements fitted with theoretical curves and the potential of mean force computed using the ff03 force field.[280]

It seems that the PEP correction actually improves the PM3 Hamiltonian for the conformational properties of 2Ala, though this correction does not directly affect the ϕ and ψ dihedral angles. It is also important to notice that this correction does not introduce any artifact on the PES. For those reasons, the use of the PM3 Hamiltonian without PEP correction will be avoided in the rest of the present work and, from now on, the PM3 acronym will always refer to PM3-PEP if not otherwise stated. This first analysis was performed from static gas phase calculation and we shall now take into account dynamical and solvation effects.

5.4.2 Dynamics of structural and electronic properties

The investigation of conformational properties in biological systems requires the use of dedicated tools. Even for a small polypeptide such as 2Ala, one needs to use highly efficient sampling techniques (*e.g.*, replica exchange MD, metadynamics, *etc.*) to expect exploring a large part of the conformational space.[226, 281] Since here we performed a single 500 ps MD trajectory of our system, we do not expect to obtain a representative distribution of the different conformations of 2Ala. Nevertheless, we will show in what follows that the conformation of the dipeptides directly affects the electronic properties of this molecule.

The structural properties of the dipeptide were followed along the simulation. In particular, we present in Figure 5.41 the ϕ and ψ values explored by the molecule during the MM-MD and SEBOMD simulations and the related probability map in a Ramachandran plot. To help the discussion, we reported on the Figure a delimitation of the Ramachandran map into three main regions as proposed by Seabra *et al.*[226]: C5, PPII and α . As suggested by other works, the conformational preferences of 2Ala are modified when going from the gas to the

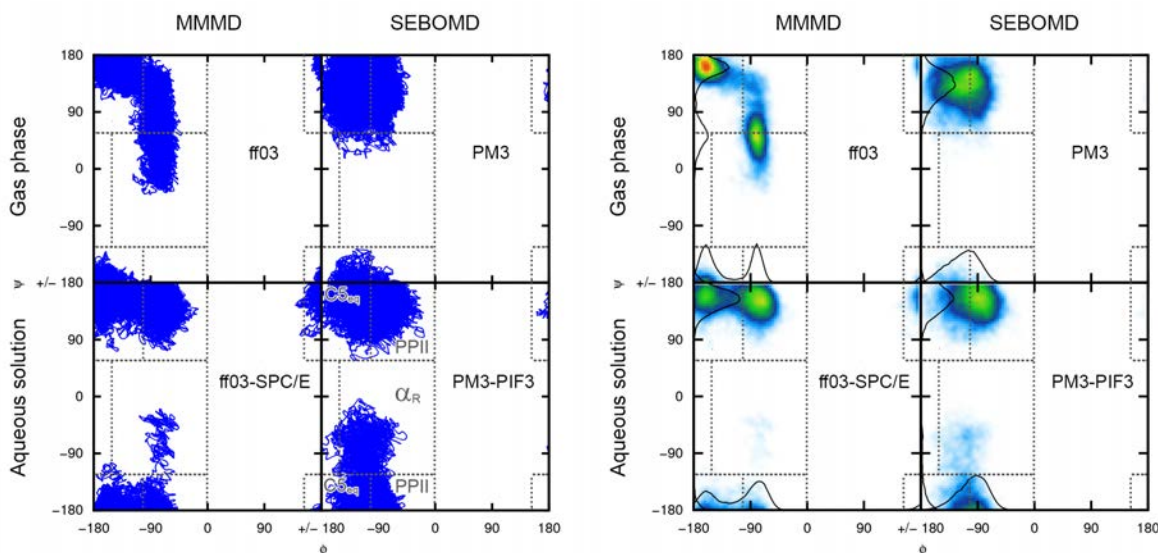


Figure 5.41: Values of the ϕ and ψ angles along the gas and the condensed phase MM-MD and SEBOMD simulations (left panel) and the related distribution of probability (right panel). On the left panel, we show the space delimitation of the three basins as defined in Ref. [226] In the left panel, the plain lines correspond to the projection of the distributions on the ϕ and ψ axes.

condensed phase. In the gas phase, both ff03 and PM3 results follow the PES discussed above and presented in Figure 5.40. This can be seen from the trajectories (left panel of Figure 5.41) and is confirmed by the associated probability densities (see the right panel of Figure 5.41). Here, mostly the β basin is explored in MM-MD and SEBOMD simulations. The ff03 results show two well defined maxima of the ϕ/ψ distribution corresponding to the C5 and C7_{eq} conformations, while the PM3 Hamiltonian predicts one maximum extending through the C5 and PPII conformations. This is in agreement with the fact that the C7_{eq} conformation was not observed on the PM3 PES. In the condensed phase, for PM3-PIF3, the distribution of ϕ/ψ angles is close to the one obtained in the gas phase. In the former case, the PPII conformation seems to be favored and the α basin is explored, though to a much lower extent. Similar observations can be drawn from the ff03-SPC/E results. In the condensed phase, the C7_{eq} is not explored and a small amount of configurations is found in the α basin.

We also followed the variation of the CM1 dipeptide dipole moment along our SEBOMD simulations. The corresponding distributions are represented in Figure 5.42 for the gas and the condensed phase. As a first observation, one can see that not only the distribution in the condensed phase is shifted compared to the gas phase, but also its shape is strongly affected. While in the gas phase, only one peak centered at about 1.5 Debye is observed, with a small tail extending up to 6 Debye, the distribution in the condensed phase is much wider and presents a different structure. A main maximum is present at around 6.0 Debye, a shallow shoulder can be distinguished at 4.4 Debye and a second maximum of low intensity is located at about 10.0 Debye. From our trajectories, we sorted each combination of ϕ/ψ angles into three main families as suggested by Seabra *et al.*[226] The averaged dipole moment of each basin is reported in Figure 5.42 as well as the related probability of each conformation

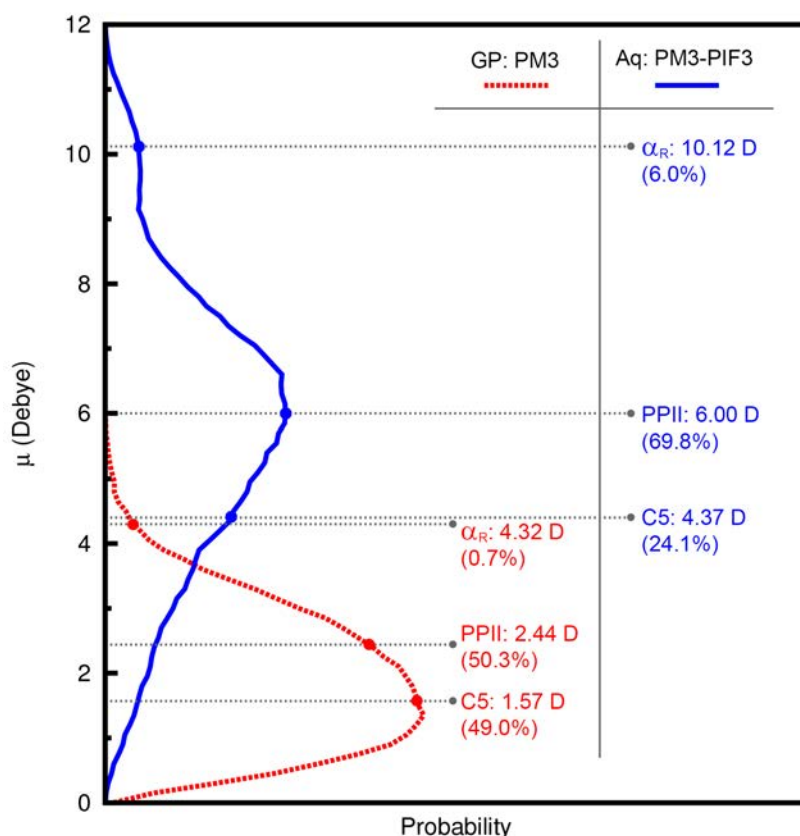


Figure 5.42: Comparison of the CM1 alanine dipeptide dipole moment distribution along the SEBOMD simulations in the gas and in the condensed phase. The average dipole moment within the different basins is given in red for the gas phase (GP) and in blue for the condensed phase (Aq). In parenthesis, we report the population of each basin along the simulation.

to be found along the dynamics. Although our sampling is far from being sufficient to extract statistical data about the relative probability of the different configurations of 2Ala, a trend appears, showing different values of the dipole moment as a function of the conformation. In the gas phase, the C5, PPII and α conformations are related with a dipole moment of 1.57, 2.44 and 4.32 Debye, respectively. In aqueous solution, the dipole moment of 2Ala in each of the three basins increases considerably, reaching 4.37, 6.00 and 10.12 Debye. We note that, in this case, the average dipole moment of each conformation corresponds to the different features of the band structure discussed above. Kwac *et al.* performed a similar analysis from long QM/MM simulations (up to 10 ns) of 2Ala in a box of 1461 TIP4P water molecules with PBC, using various semiempirical methods (including PM3) to treat the solute.[279] The authors have shown results in very good agreement with the present work in the condensed phase. They found an average dipole moment for the C5, PPII and α conformations of about 2.0, 3.8 and 8.2 Debye, respectively. The authors performed the analysis based on Mulliken atomic partial charges, while in this work we used the CM1 model. Mulliken charges are known to underestimate the magnitude of the dipole moment of peptides compared with CM1 or CM2.[74] We repeated our analysis with Mulliken charges for a comparison and found absolute values in reasonable agreement with the work by Kwac *et al.* (*i.e.*, 3.6, 4.8 and 8.1 Debye in the condensed phase for the C5, PPII and α conformations, respectively).

Despite the low statistics, our results are consistent with a close correlation between conformational and electronic properties of the alanine dipeptide. Because SEBOMD simulations are much less time consuming than other QM based MD methods (*e.g.*, CPMD), the use of high performance sampling methods combined with a QM description of the whole system represent a realistic further development of the work performed so far.

5.4.3 Solvent structure

Radial distribution functions require significant statistics in order to obtain a converged result. For this reason and considering the discussion above, we did not split our trajectory into conformational basins for this analysis. However, the distributions of ϕ and ψ dihedral angles presented in the previous Subsection tend to show that the main contribution to our results comes from the β basin for both PM3-PIF3 and ff03-SPC/E simulations. We shall discuss and compare with other works the RDFs of water around 2Ala by keeping this observation in mind.

Figure 5.43 presents the RDFs of the carbonyl oxygen atom in the ACE and the ALA residue with respect to the Hw atoms of water. The ff03-SPC/E and the PM3-PIF3 simulation provide similar results. This observation is in good agreement with the discussion about the solvation of NMA (see Section 5.3.1). The RDF of both carbonyl groups are identical in the first solvation shell and the peak height is slightly higher in the second shell in the case of the ACE residue compared to ALA. Compared to ff03-SPC/E, the first peak is much higher in the case of PM3-PIF3, which is consistent with the results obtained for amide molecules. The position of this first peak is found at 1.7 and 1.8 Å for ff03-SPC/E and PM3-PIF3, respectively. From a CPMD simulation of 2Ala in a box of 118 water molecules, Gaigeot *et al.* reported a position at about 1.7 Å in the β basin.[276] The authors noticed a difference between the ACE and ALA residue in the α basin, but not for the β conformations. Similar values were obtained by Kwac *et al.* from their QM/MM study of this system using the PM3 Hamiltonian to describe the alanine dipeptide.[279]

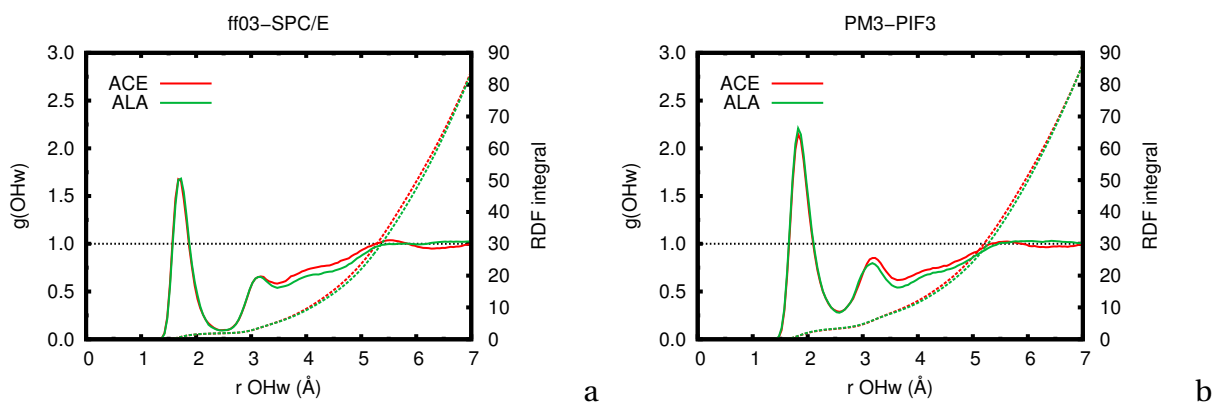


Figure 5.43: Comparison of OHw RDFs from MM-MD (a) and SEBOMD (b) simulations for the ACE and the ALA carbonyl group of 2Ala in solution.

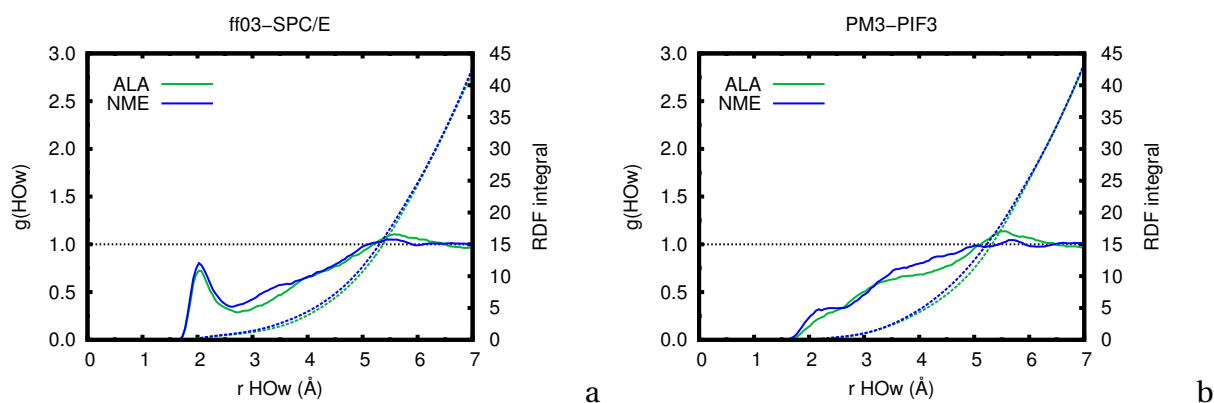


Figure 5.44: Comparison of H-Ow RDFs from MM-MD (a) and SEBOMD (b) simulations for the H atom bonded to the nitrogen in the ALA and the NME residue of 2Ala in solution.

In Figure 5.44 we show the RDF between the water Ow atoms and the H atom bonded to the nitrogen atom of the ALA and NME residues. No difference is observed between the RDFs coming from the two residues. The ff03-SPC/E method predicts a peak at about 2 Å, in good agreement with the results of Refs. [276] and [279]. As we observed in the case of amides, the PM3-PIF3 Hamiltonian does not reproduce the H-Ow interaction when the H atom belongs to the nitrogen atom of a peptide bond, as predicted by MM and CPMD studies, but shows a shoulder in the RDF structure at the correct position. Further investigation are required to state if this difference could lead to a change in the conformational preferences of 2Ala in aqueous solution.

Finally, we would like to discuss the interaction between water and the three methyl groups of 2Ala. The hydration structure of those sites has not been discussed in the literature. However a qualitative comparison with the previous results about smaller solutes can be carried out. Figure 5.45 shows the RDF between the methyl C atoms and the oxygen atoms of water. As expected, the global shape and position of the peaks are consistent with the results discussed in the previous Sections of this Chapter. ff03-SPC/E and PM3-PIF3 predict a similar solvation shell around the methyl group of the two extremities of the peptide, while

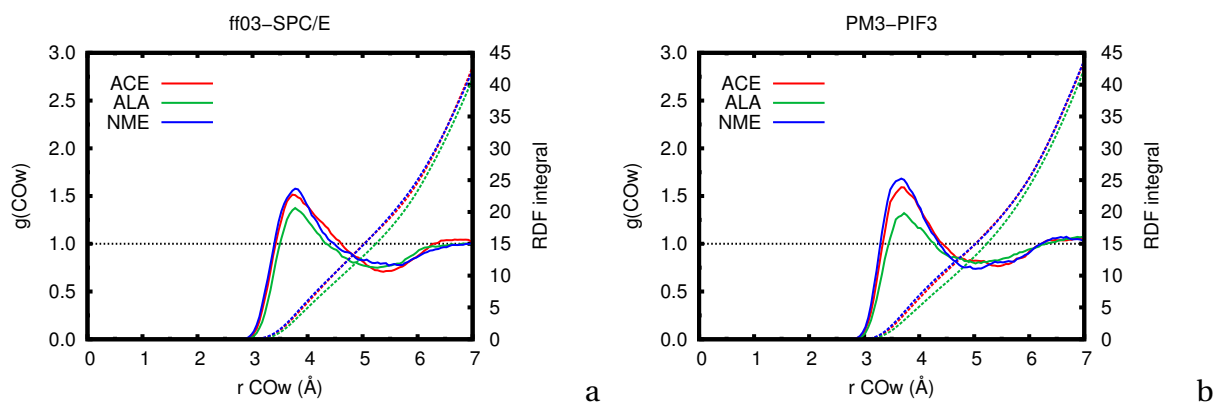


Figure 5.45: Comparison of C-Ow RDFs from MM-MD (a) and SEBOMD (b) simulations for the three methyl C atoms of 2Ala in solution.

the intensity of the peak in the RDF of the ALA residue methyl group is slightly lower.

The structure of water around 2Ala obtained with the PM3-PIF3 Hamiltonian is in very good agreement with other works and consistent with the present results about smaller molecules. However, the weak H-Ow interaction (for H bonded to the nitrogen of a peptide bond) predicted by PM3-PIF3 compared to other works suggests that a deeper analysis is required to assess the impact of such difference on the description of the conformational features of 2Ala.

5.4.4 Vibrational properties

As we discussed in the previous Section, four vibrational bands are characteristic of amide and polyamide compounds: AmI, AmII, AmIII and AmA. The position and the shape of these bands in the infrared spectrum of a polypeptide are strongly and specifically affected by intra-/inter-molecular interactions. Because the conformation of a peptide involves different interactions with the solvent and the rest of the molecule, infrared spectroscopy is used experimentally as a fingerprint of peptides conformations.[282–285] Recently, Gaigeot *et al.* investigated the IR spectrum of 2Ala in solution by means of CPMD simulations.[227, 276, 286] The authors performed two simulations starting from geometries belonging to the β and to the α basin of the Ramachandran map. From those two simulations of approximately 60 ps each, the authors computed the IR spectrum related to each of the two basins. They have been able to interpret the effect of the peptide conformation on the AmI vibrational band of 2Ala.

Let us start with the solvent effect on the global infrared spectrum of 2Ala. In Figure 5.46, we present the comparison between the gas and the condensed phase IR spectra of 2Ala obtained from SEBOMD and MM-MD simulations. We note that a non polarizable force field (*i.e.*, ff03 and ff03-SPC/E in the present study) predicts an almost negligible solvent effect on the vibrational properties of 2Ala. From the SEBOMD simulations, the effect is clear and consistent with the previous discussions about the model molecules detailed earlier in this Chapter. We assigned some characteristic bands from the vibrational density of state (VDOS) decomposition into atomic contributions and those are reported in Figure 5.46. In a similar manner compared with NMA and the other amide molecules treated in the present work, the AmI and the AmA vibrational bands are red shifted in solution. Similarly, AmII and AmIII appear to be blue shifted due to the interaction with the solvent. It is worth to notice that, in the present case, the AmIII band is composed by two peaks while only one peak was observed for the molecules containing only one amide group. The C-H stretching mode of the three methyl groups ($\nu_{\text{CH}}^{\text{met}}$) gives rise to two broad peaks between 3000 and 3500 cm^{-1} , which are red shifted in solution, in a consistent way compared to the results obtained for alkane compounds. Finally, $\nu_{\text{CH}}^{\text{C}}$, the band corresponding to the stretching mode of the H atom bonded to the α carbon atom is found at a lower frequency than $\nu_{\text{CH}}^{\text{met}}$ and is also slightly red shifted

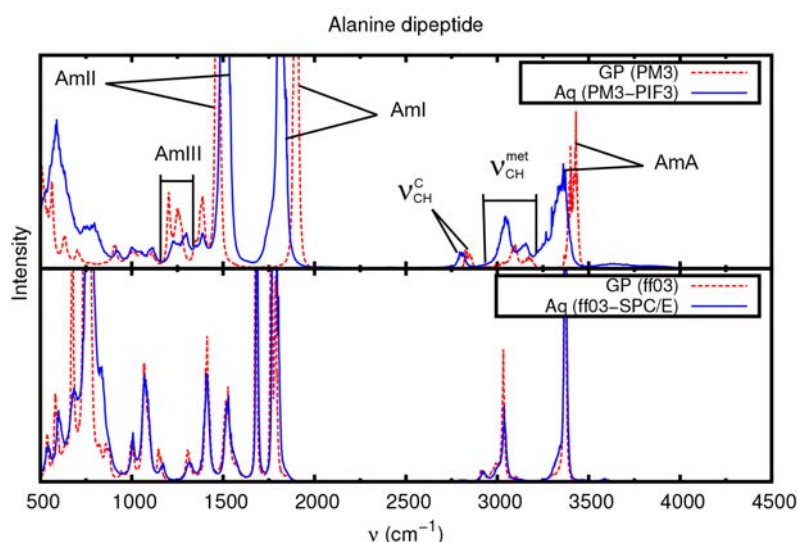


Figure 5.46: Comparison of the infrared spectra of the alanine dipeptide obtained from SEBOMD (top panel) and MM-MD (bottom panel) simulations. Gas phase results are shown as dashed red lines while condensed phase ones are shown as plain blue lines.

in solution. This last observation is consistent with the discussion about the isobutane molecule in Subsection 5.2.3.

We shall now investigate the effect induced by the peptide conformation on AmI and AmA. As we have shown in the present Section, our PM3-PIF3 trajectory explores three basins of the Ramachandran map during the simulation time (*i.e.*, 500 ps). As it has been described in Section 1.5, the IR spectra are computed from the Fourier transform of the dipole moment time correlation function. We also showed that the use of the Andersen thermostat requires to split the trajectory into windows of 1 ps. We analyzed our trajectory and we sorted each of the 1 ps windows in the different basins represented in the Ramachandran map. The windows in which the conformation of 2Ala changes from one basin to another during the 1 ps trajectory were discarded. This analysis led to 151, 16 and 20 windows corresponding to the PP2, the C5 and the α basin, respectively. We notice that, considering the discussion in Section 1.5, the statistics in the C5 and the α basin will be too low to allow a quantitative analysis of the results.

Figure 5.47 presents the comparison of IR spectra obtained in solution from the PP2, the C5 and the α basin along the PM3-PIF3 SEBOMD simulation. The total IR spectrum is also reported in this Figure. Figure 5.47a focuses on the AmI frequency region. Although the spectra coming from the C5 and α basins are not fully converged with respect to the width of the bands, the position of the latter is already representative. The total spectrum as well as the PPII and the C5 spectra show a peak around 1810 cm^{-1} . For the α basin, it is interesting to observe that the AmI band presents two peaks at about 1800 and 1840 cm^{-1} , the latter showing the highest intensity. Compared to PPII and C5, this last peak is blue shifted by about 30 cm^{-1} . In their analysis of the α/β IR spectra of 2Ala, Gaigeot *et al.* reported results in very

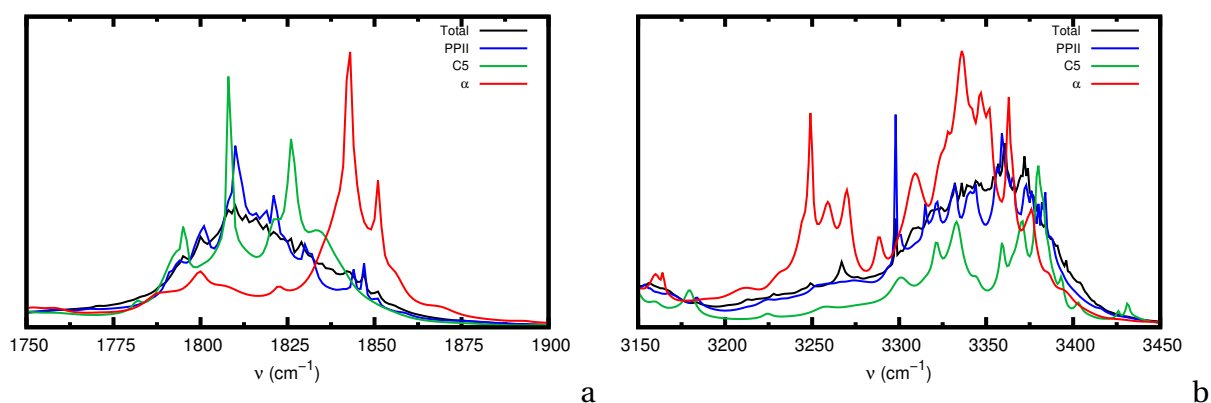


Figure 5.47: Comparison of the infrared spectra of the alanine dipeptide obtained from the different basins of the Ramachandran map along the SEBOMD simulations in solution using the PM3-PIF3 Hamiltonian. a: AmI frequency region. b: AmA frequency region.

good agreement with the present observations, *i.e.*, an α AmI peak blue shifted compared to the β basin and presenting two peaks.[286] It is also worth noting that, considering the large difference of probability in finding the PPII or the α conformation along the present dynamics, it seems reasonable that the main contribution to the total IR spectrum arises from the PPII basin.

Gaigeot *et al.* only focused on the AmI region of the spectrum. However, from our simulations it appears that the AmA band is also affected by the conformation of the peptide. Indeed, when looking at Figure 5.47b, one can see that both PPII and C5 conformations show a broad peak centered at around 3360 cm^{-1} , while the α basin predicts another structure of this band. Here again, two peaks are present, and the one presenting the largest intensity is slightly red shifted compared to what is observed in the other basins. However, considering the low statistics used to compute the IR spectrum in the α basin, this second peak is more likely to be an artifact.

Although the sampling of the different conformations is relatively poor in the present simulation, some features related with the 2Ala structure can be identified on the IR spectrum of the dipeptide. In particular, the difference in the shape and position of the AmI vibrational band between the PPII and the α basin, obtained from our PM3-PIF3 SEBOMD simulation, is in good agreement with other theoretical studies.

5.4.5 Summary

We have shown here that the PM3 Hamiltonian needs to be corrected in order to describe the dynamics of polypeptides in water. First, the peptidic correction (PEP) is necessary to overcome the issues stemming from the non planarity of the nitrogen environment in an amide group. Second, as we have shown along the present Chapter, the PIF3 corrections leads to a correct description of intermolecular interactions, avoiding the well known deficiencies of PM3.

This work also shows that SEBOMD simulations of the alanine dipeptide using the PM3-PIF3 semiempirical Hamiltonian can reproduce most of the features related to the conformational equilibria in the gas and in the condensed phase. The structural properties compare fairly well to other experimental and theoretical works, as well as the electronic properties. We could also show the effect of the conformation of 2Ala on the dipole moment and on the infrared spectrum of this compound. However, further work to increase the sampling of the α basin should be done in order to obtain a better picture of the contribution of the different conformations to the total IR spectrum.

5.5 Concluding remarks

We have reported in the present Chapter an extensive investigation of solvent effects on the dynamical properties of molecules chosen to reproduce chemically significant fragments of biological compounds. By means of SEBOMD simulations combined with the PM3-PIF3 Hamiltonian developed in the present work, we have compared the solvent structure around small hydrophilic and hydrophobic compounds, as well as the effect of water on the electronic and vibrational properties of those molecules. We finally presented an application to the solvation of a larger model, *i.e.*, the alanine dipeptide.

The SEBOMD technique has been shown here to be a powerful tool to include polarization into molecular dynamics simulations, while keeping a relatively large sampling compared with similar methods. It appears clearly from this work that the quality of the SEBOMD predictions strongly depends on the quality of the chosen semiempirical method, as it is the case for any MD technique (*e.g.*, MM-MD, *ab initio*-MD, CPMD, *etc.*).

This work also represents the first application of the PM3-PIF3 Hamiltonian in the condensed phase. We have shown that this method is particularly well suited to study the hydration of hydrophobic compounds as well as molecules bearing both hydrophobic and hydrophilic groups. However, some features are still to be improved in order to reach more general predictions of solute-water interactions in biological systems.

Finally, the application to the alanine dipeptide gave very promising results. Although the sampling of the present simulations should be extended, we have been able to show solvent and conformational effects on the electronic and vibrational properties of the molecule in good agreement with other works.

Further applications will focus on the simulation of larger polypeptides in water. This work has shown that the SEBOMD technique and the PM3-PIF3 Hamiltonian are suitable to perform such a study. The main issue when dealing with systems bearing a large amount of degrees of freedom is to reach a sufficient sampling of the phase space. Because of the low computational cost of SEBOMD compared to other similar methods, the use of efficient sampling techniques (such as replica exchange molecular dynamics) becomes realistic. However, an investigation of the compatibility of such methodologies with a full QM description of the system is required to state about the feasibility of such an application of SEBOMD.

Chapter

6

Water self-dissociation in confined systems

Résumé

Ce chapitre est dédié à l'étude de la réaction d'autoprotolyse de l'eau en milieu confiné à l'aide de calculs quantiques semiempiriques réalisés sur un système modèle, $(\text{H}_2\text{O})_{21}$ (un agrégat de 21 molécules d'eau en phase gazeuse). Nous expliquons dans un premier temps le choix de ce système modèle en rapport avec les études expérimentales et théoriques menées sur les micelles inverses au cours des dernières années. Nous présentons également un état de l'art des calculs théoriques liés à l'étude des agrégats d'eau neutres ou protonés ainsi qu'aux transferts de protons dans ces systèmes. La deuxième partie de cette étude permet de justifier du choix de l'Hamiltonien semiempirique à utiliser, en se basant sur l'analyse des performances de ces méthodes quant au calcul des propriétés de systèmes allant du dimère d'eau à $(\text{H}_2\text{O})_{21}$. Enfin, nous réalisons une étude approfondie des caractéristiques favorisant ou inhibant le transfert de proton entre deux molécules d'eau dans $(\text{H}_2\text{O})_{21}$ à l'aide de dynamiques moléculaires couplées à la méthode d'*umbrella sampling*.

La réaction d'autoprotolyse de l'eau est certainement l'un des processus les plus importants dans la nature. Elle contrôle les équilibres acido-basiques et par conséquent la réactivité des systèmes moléculaires en milieux aqueux. Cette réaction est désormais bien connue dans l'eau liquide. En revanche, la compréhension de ce phénomène dans des systèmes confinés reste un défi, d'un point de vue expérimental comme théorique. Un exemple de système représentatif de l'eau en milieu confiné est donné par un objet moléculaire appelé micelle inverse. Ce type d'objet est formé par l'agrégation de molécules de tensioactifs qui forment une cavité dans laquelle un nombre limité de molécules d'eau est emprisonné. Des études expérimentales ont montré que la dynamique et la réactivité de l'eau varient en fonction de la taille de la cavité d'eau. Certaines études théoriques ont confirmé les effets du confinement sur la dynamique de l'eau. Quant au transfert de protons, beaucoup d'études existent sur des agrégats d'eau protonés de taille variable mais peu de travaux portant sur des systèmes neutres ont été reportés. Nous proposons ici d'apporter quelques éléments de réponse quant aux transferts de protons dans un système neutre, l'agrégat $(\text{H}_2\text{O})_{21}$.

Le nombre important de degrés de liberté dans le système $(\text{H}_2\text{O})_{21}$ nécessite de prendre en compte les propriétés dynamiques de ce dernier. Cependant, considérant la taille de

$(\text{H}_2\text{O})_{21}$, des calculs *ab initio* utilisant un haut niveau de théorie nécessiteraient des temps de calcul trop importants. Les méthodes semiempiriques semblent être des candidats intéressants pour mener ce type d'étude. Nous réalisons ici un test approfondi de la capacité des méthodes semiempiriques courantes à traiter de la structure et de la réactivité des systèmes aqueux. Nous montrons tout d'abord les résultats obtenus par ces méthodes sur un système modèle simple, le dimère d'eau. De cette étude, seules les méthodes explicitement développées pour traiter l'eau semblent capables de donner une bonne description de la géométrie du dimère d'eau. Notamment, nous obtenons de meilleurs résultats avec l'approche PM3-MAIS. Ces bons résultats sont également confirmés par l'étude de la structure de $(\text{H}_2\text{O})_{21}$ et de la stabilité d'une paire d'ions (H_3O^+ et HO^-) dans cet agrégat comparé à des calculs MP2/aug-cc-pVTZ. Ces observations sont particulièrement encourageantes quant à l'utilisation de l'Hamiltonien PM3-MAIS pour la suite de cette étude.

Enfin, nous cherchons à expliquer en quoi les caractéristiques de certaines géométries données de $(\text{H}_2\text{O})_{21}$ favorisent ou inhibent le transfert de proton entre deux molécules d'eau voisines. À partir de deux configurations de l'agrégat, nous étudions chaque transfert de proton possible au sein de $(\text{H}_2\text{O})_{21}$ et en calculons l'énergie libre associée par le biais de simulations d'*umbrella sampling*. En nous basant sur des travaux présents dans la littérature, nous mettons au point différents descripteurs, traduisant du réseau de liaisons hydrogènes ainsi que de la géométrie et de l'environnement direct des deux molécules d'eau impliquées dans un transfert de proton. Deux de ces descripteurs semblent montrer une corrélation avec l'énergie libre nécessaire à ce transfert : la distance entre les deux atomes d'oxygène impliqués ainsi que la direction et l'intensité du champ électrique créé par les 19 autres molécules sur le proton à transférer. Nous testons l'utilisation de ces deux descripteurs afin de prédire les configurations favorisant le transfert à partir d'une dynamique SEBOMD de $(\text{H}_2\text{O})_{21}$. Les résultats obtenus sont particulièrement encourageants pour guider une étude future de la réaction d'autoprotolyse dans des systèmes de plus grande taille.

Water self-dissociation is certainly one of the most important processes in Nature. It controls the acid/base equilibrium and thus dictates the reactivity in aqueous environment. This reaction is related to the well known equilibrium,



with the associated rate constant $K_W = 10^{-14}$ ($T = 298$ K and $P = 1$ atm). As suggested by the very low value of the rate constant (K_W), such a process is a rare event. Indeed, the production of a stable ionic pair from any randomly chosen couple of water molecules takes about ~ 10 hours in liquid water.[287] From Eq. 6.1 comes the definition of pH as $\text{pH} \simeq -\log([\text{H}_3\text{O}^+])$ proposed by Sørensen in 1909[288] and commonly adopted by the chemistry community since then.[289] However, as the notion of concentration, this definition is valid only from a macroscopic point of view in a system containing a statistically significant amount of particles (*i.e.*, bulk water). As the number of particles decreases because of the limited size of the system, those concepts are no longer valid.[290] This is the case of confined systems such as water nanodroplets and reverse micelles (RMs), in which the water molecules are constrained to occupy a limited space.

A reverse micelle consists of a cavity created in non polar environments by surfactant molecules. Such compounds bear an amphiphilic character and are composed by a small polar head and a hydrophobic tail of variable length. The polar heads of the surfactant molecules aggregate and give rise to the formation of cavities that contain a limited amount of water molecules. Those cavities are usually found to be spherical, since such a shape minimizes the interaction area at the interface between the polar water pool and the exterior hydrophobic environment. The size of RMs containing water is commonly characterized by a single parameter, $w_0 = [\text{H}_2\text{O}]/[\text{surfactant}]$. For spherical RMs, this parameter is directly proportional to the radius of the cavity.[291] If the physical properties and the self-dissociation of water are now well understood in the bulk, measuring and modeling such phenomena represents a challenge for both experimental and theoretical studies in confined environments.[290]

During the past decades, extensive experimental and theoretical studies have been devoted to understand the dynamical properties of water, encapsulated in RMs, as a function of the w_0 parameter.[291–298] In the late 90s, Riter *et al.* already remarked that the water mobility decreases with the value of w_0 and that the molecules tend to be almost frozen for low values of w_0 (*i.e.*, $w_0 \simeq 1$ corresponding to about 21 water molecules).[292] This trend was also confirmed by molecular dynamics (MD) simulations carried out by the Ladanyi's group.[293–295] Later, Harpham *et al.* explained from quasielastic neutron scattering experiments and from MD simulations, that the water mobility varies from the interface to the center of the RM.[296] Baruah *et al.* showed the same trend for the structure of the environ-

ment of a molecular probe depending on its location within the RM.[297] All those findings point out that the motion of water is strongly affected by its confinement in RMs and that the water pool bears an anisotropic character that differentiates the behavior of the molecules in the center of the RM from those at the interface. These observations yield the definition of a core-shell model of nanoconfined water, which was partially confirmed by Piletic and coworkers.[298] In this work, the authors have shown that this anisotropy is only true for dynamical properties that do not imply a rearrangements of the global hydrogen bonds network. The authors suggested that this network links the dynamics of the core and that of the shell region. Finally, in 2009, Levinger *et al.* reviewed the recent findings brought by ultrafast spectroscopy on the physical properties of RMs water pools, which corroborate the statements discussed above.[291]

The observations mentioned here point out several modifications of the water physical properties induced by its confinement. If such effect exists, one can wonder whether the chemical properties of water in those systems differ from a bulk environment as well. We propose here to assess this question by means of theoretical chemistry tools, using a model system of confined water: *i.e.*, the $(\text{H}_2\text{O})_{21}$ water cluster in gas phase. This system has a similar size compared to water encapsulated in small RMs with values of w_0 close to 1.[293–295] The structure and reactivity of gas phase water clusters has been widely studied from a theoretical point of view and we shall discuss the main findings related to the present work in what follows.

6.1 Background on water clusters

Despite the apparent simplicity of a water molecule, understanding and predicting the structural organization of aqueous systems is still a challenge from a theoretical point of view. In what follows, we shall first describe the structure of water aggregates, from the water dimer to large water clusters. A comparison with the condensed phase will be considered, though the attention of the following discussion is mainly drawn on gas phase water clusters.

Water molecules interact with each others *via* hydrogen bond (Hbond) formation. Predicting such a fundamental intermolecular property represents a required test case for any computational model intended to treat aqueous systems.[299–301] The water dimer is the smallest system that illustrates this interaction and the geometry of this dimer is frequently updated in the literature, as the computational power allows to increase the precision.[302–304] To date, the most precise calculations on this system have been recently reported by Lane up to the CCSDTQ level of theory.[304] Furthermore, understanding the arrangement and the n-body interactions among water molecules in small $(\text{H}_2\text{O})_n$ clusters is of key interest and has been the focus of many theoretical studies in the literature, from trimer to hexamer[305–309] and up to 22 water molecules.[310–313] In particular, Xantheas focused

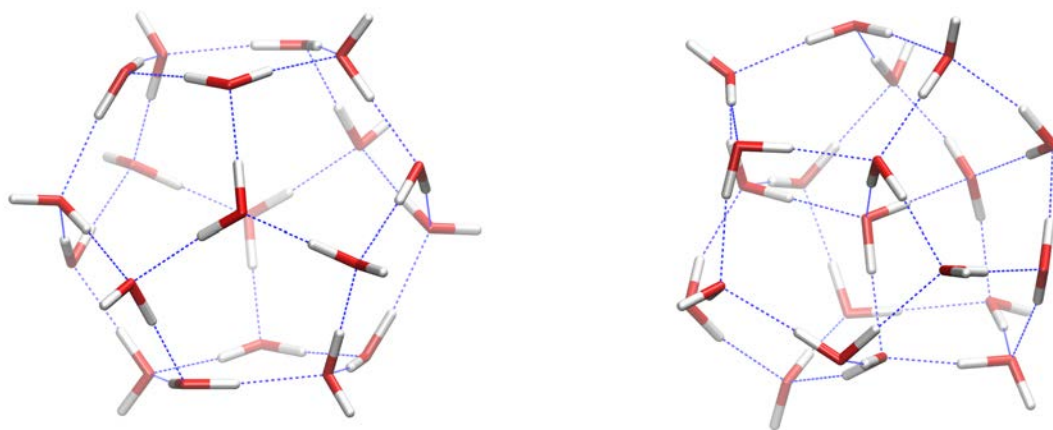


Figure 6.1: 5^{12} structures of the $(\text{H}_2\text{O})_{20}$ (left) and the $(\text{H}_2\text{O})_{21}$ (right) clusters, optimized at the PM3-MAIS level of theory.

on predicting the global minimum for the pentagonal dodecahedron (5^{12}) form of the $(\text{H}_2\text{O})_{20}$ water cluster.[313] One can refer to this paper for a complete review and good references of such structures.

As mentioned above, we focus here on $(\text{H}_2\text{O})_{21}$, denoted in the literature as the “magic number” water cluster for its remarkable stability. This cluster corresponds to the number of water molecules contained in an RM bearing a w_0 value of 1.[293–295] As $(\text{H}_2\text{O})_{20}$, $(\text{H}_2\text{O})_{21}$ bears a 5^{12} geometry. The main difference between $(\text{H}_2\text{O})_{20}$ and $(\text{H}_2\text{O})_{21}$ is that the extra molecule of the latter is located in the interior of the 20 molecules cage (see Figure 6.1). This difference is of critical importance since it implies a heterogeneity in the environment surrounding different water molecules. Indeed, in $(\text{H}_2\text{O})_{20}$, all the water molecules are located at the surface of the cage and participate to only 3 Hbonds. Each water molecule is either a donor of two Hbonds (2D) and an acceptor of one Hbond (1A), or it is a donor of one Hbond (1D) and an acceptor of two (2A). In the case of $(\text{H}_2\text{O})_{21}$, 5 molecules are both 2D and 2A, as reported by Lenz *et al.* in a complete study of the hydrogen bond topology in various water clusters.[311, 312] Compared to periodic systems (*e.g.*, liquid or ice water), the interface plays a dominant role in gas phase water clusters. It implies the presence of dangling protons (*i.e.*, hydrogen atoms that do not participate to any Hbond) and thus, most of the surface molecules do not follow the Bernal-Fowler model, which states that each water molecule should be a donor of two Hbonds and, implicitly, an acceptor of two Hbonds.[314] Such structural differences with respect to bulk water impacts the dynamical behavior of water as well as its chemical properties.

The water self-dissociation process results from a collective phenomenon.[315] It has been widely studied in liquid water by means of free energy calculations *via* Car-Parrinello molecular dynamics (CPMD) simulations.[192, 316, 317] In particular, Geissler *et al.* have carried out an extensive study on the particular phenomena that enhance the separation of the ions (*i.e.*, H_3O^+ and HO^-) resulting from the self-dissociation in liquid water.[317] The au-

thors pointed out the importance of the rare combination of collective properties: the proper hydrogen bonds network and the correct orientation of the environment electric field. Those criteria appear to be essential to favor a proton transfer between two neighboring molecules and thus to initiate the reaction. The authors also emphasized that this first proton transfer is the rate limiting step of the water self-dissociation, the next step being a Grotthuß like mechanism, which separates each of the two resulting ions from each other. Experimental studies showed that the separation of the two charges should be at least of 6 Å for this ionic pair to be stable.[318] Moreover, as explained by Geissler *et al.*,[317] this charge separation is not only a matter of distance but also requires a certain number of water molecules between the two ions. To quantify this, the authors defined a parameter (l), being the number of Hbonds connecting the two ions (*e.g.*, $l = 2$ if only one molecule separates the two ions). They suggested that this parameter should be $l \geq 3$ to stabilize the ionic pair, even though, in some rare cases, they observed stable ionic pairs in liquid water with $l = 2$.

When moving from liquid periodic systems to gas phase water clusters, several of the aspects discussed above change. Indeed, if a proton transfer occurs in such system, the finite size of the cluster limits the charge separation after the initial transfer. In the case of 5^{12} (H_2O)₂₀ or (H_2O)₂₁, the largest possible charge separation is slightly smaller than 6 Å. The range of values for the l parameter is also restricted by the size and by the structure of (H_2O)₂₀ and (H_2O)₂₁. If in (H_2O)₂₀ they are several possibilities for the parameter l to have values greater than 3, this is not the case in (H_2O)₂₁: because of the presence of the central water molecule, only a few organizations of the ionic pair lead a value of $l \geq 3$.

Another question arises from the presence of an interface: which location of the two ions is the most favorable? In other words, are those ions more stable in the shell or in the core region of the cluster? This question has been the focus of many studies carried out on the structure and reactivity of protonated water clusters.[313, 319–325] All those studies suggested that the extra proton is more stable at the surface of the structure for large clusters. This can be illustrated by few considerations about the solvation structure of H_3O^+ .[182] Two geometries of this cation have been reported in the literature:

- The *Eigen cation*[326] is made of a protonated water molecule (H_3O^+) acting as a pure Hbond donor and surrounded by three acceptor water molecules.
- The *Zundel cation*[327] where the extra proton is shared between two water molecules (H_5O_2^+) and that is usually solvated by four acceptor water molecules.

Either Eigen or Zundel cations are better accommodated at the surface of (H_2O)₂₁ rather than in its interior for geometrical reasons. In the latter case, a twist in the shape of the cluster would be observed and the energy would thus increase.[325]

Despite the apparent less favorable situation for water self-dissociation within (H_2O)₂₀ and (H_2O)₂₁ compared to liquid water, stable structures containing a separated ion pair have been reported in the literature.[310, 328–331] We will refer to such zwitterionic structures as

Z^\pm in what follows. From a *graph theoretical analysis*, McDonald *et al.* revealed 30026 stable arrangements of the Hbond network in the 5^{12} form of $(\text{H}_2\text{O})_{20}$. [320] Later, Kuo *et al.* characterized energetically those isomers [310] with both the OSS2 [332] empirical potential and the B3LYP [25, 26] DFT functional. The authors highlighted some particular structures, relatively high in energy with respect to the global minimum, which showed a very short oxygen-oxygen (OO) distance (*i.e.*, 2.425 Å). By forcing the proton between those two oxygen atoms to be transferred, they found a Z^\pm minimum lying 14.3 kcal/mol below the starting geometry. This Z^\pm minimum was found to be only about ~ 10 kcal/mol higher in energy than the global minimum. The authors investigated the factors that favor such situation and their findings will be of key interest in the following Sections. Later, Mrázek *et al.* examined the proton transfer free energy in small water clusters at the MP2 [14] level of theory with a 6-31++G(d,p) basis set and found a linear relationship between this free energy and the OO distance between the two oxygen atoms involved in the transfer. [329] However, this study was performed only for a few structures and the authors emphasized the fact that, depending on the configuration of the other molecules, strong deviations could be found. More recently, Torrent-Sucarrat *et al.* performed *ab initio* calculations to understand the protonation of $(\text{H}_2\text{O})_{20}$ and $(\text{H}_2\text{O})_{21}$ induced by a hydroperoxyl radical. [331] In the last part of this study, the authors focused on the finding of stationary points related to the water self-dissociation within $(\text{H}_2\text{O})_{21}$. They obtained two stable Z^\pm geometries of $(\text{H}_2\text{O})_{21}$ and the related transition states. For the most stable of those two Z^\pm minima (denoted as C10 in Ref. [331]), they found a free energy barrier of 17.4 kcal/mol and a reaction free energy of only 10.1 kcal/mol. This result is particularly surprising when compared to the experimental value in bulk water. Indeed, the activation and reaction free energies for this process in bulk water are of 23.8 and 21.4 kcal/mol, respectively. Not only the forward process seems more favorable in $(\text{H}_2\text{O})_{21}$, but also the product appears to be more stable considering the higher backward barrier compared to bulk results.

The findings discussed above suggest that water self-dissociation can be energetically more favorable in water clusters than it is in bulk water. Several possible reasons for this observation have been discussed in the literature, pointing to a complex cooperative process involving the Hbond network. We propose in what follows to reach a better understanding of the particular features of $(\text{H}_2\text{O})_{21}$ that enhance water self-dissociation.

We first discuss the QM approach to be used in the rest of the study. Considering the large number of degrees of freedom in $(\text{H}_2\text{O})_{21}$, a sufficient amount of statistical data will be required here. High level *ab initio* calculations are thus too costly from a computational point of view. Parallel to the present work, two recent studies by Wu *et al.* [153] and by Wang *et al.* [333] have shown that semiempirical methods can be suitable to investigate neutral or protonated aqueous systems, through a specific reparameterization and/or correction of these QM approaches against high level *ab initio* reference calculations. Wu *et al.* developed

specific reaction path parameters using a training set composed by optimized geometries of neutral and protonated water clusters as well as a series of geometries reflecting the proton transfer within the Zundel cation. Among the methods tested by the authors, the reparameterized version of the OM3 model (*i.e.*, OM3n) has shown promising results through an application to the study of the hydration of a proton in water using a QM/MM approach. Wang *et al.* adopted a similar strategy to that by Wu *et al.* in order to reparameterize the AM1 Hamiltonian, though the former authors only considered minimum structures in their training set. Wang *et al.* performed a new parameterization of AM1 following two approaches: by keeping the AM1 Hamiltonian as originally formulated (*i.e.*, AM1-W) and by adding a pairwise Gaussian correction function to the core-core repulsion term (*i.e.*, AM1PG-W), which somehow resembles the idea of the MAIS correction developed in our group (see Chapter 2). Also in this case, the specific semiempirical methods show a good agreement with higher levels of theory to model proton transfer in water.

The PM3-MAIS Hamiltonian (see Section 2.3.2) has been tailored to study proton transfers in water and thus appears as a good candidate for this study. However, a comparison with other semiempirical methods against *ab initio* calculations is necessary in order to validate the use of PM3-MAIS and to get a better picture of the quality of this Hamiltonian. To this end, we will perform an analysis of the water dimer using various SE methods as well as a research of stable Z^{\pm} structures of $(H_2O)_{21}$ using the PM3-MAIS Hamiltonian.

We use umbrella sampling simulations to perform an extensive analysis of all the possible proton transfers within two different initial structures of $(H_2O)_{21}$. We focus only on the initial proton transfer between two neighboring molecules. The molecular dynamics simulations are performed with SEBOMD (see Chapter 3), which can access the different tools required for this study. From those results, we propose some criteria to explain the system requirements that favor a proton transfer in confined systems. Finally, we summarize a few guidelines for further studies.

6.2 Proton transfer in water: performance of semiempirical Hamiltonians

Before starting any further investigation, the choice of the Hamiltonian to be used needs to be discussed. In Chapters 4 and 5, we have shown that most of the available semiempirical methods fail to describe intermolecular interactions. We shall discuss here their ability to treat the structure and the reactivity of aqueous systems. We first investigate the potential energy surface of the water dimer with various SE methods and the proton transfer within this system. Then, we focus on the proton transfer in an ionic model: the Eigen cation. We show that, among the methods tested here, only the PM3-MAIS Hamiltonian gives satisfactory results for the purpose cited above. Finally, we study the ability of this SE method to

locate zwitterionic (Z^\pm) minima of the $(\text{H}_2\text{O})_{21}$ water cluster with respect to the findings of Torrent-Sucarrat *et al.*[331]

6.2.1 The water dimer

As we discussed in the previous Section, the correct prediction of the water dimer structure is a compulsory test for any model intended to treat aqueous systems. We recall that the most up to date calculations on this structure have been reported recently by Lane at the CCSDTQ level of theory.[304] However, even if this represents the most accurate results on this system, strong deviations in the geometry of the water dimer are not obtained compared to MP2 calculations. Thus, we use here as a reference MP2/aug-cc-pVTZ results.

Interaction energy profiles

Smith *et al.* reported a complete study of the stationary structures of the water dimer.[334] Among those geometries, we chose three known structures of the dimer (A, B and C in Figure 6.2). The structure A is the global minimum while B and C are transition states. In addition, considering the numerous artifacts present in most of the SE methods discussed in Chapters 4 and 5, we constructed two extra structures that are not supposed to yield any minimum (D and E on Figure 6.2). The structures D and E are expected to reflect the impact of the OO and HH interactions on the potential energy surface, respectively. For each structure, we performed a rigid scan of the interaction energy as a function of the distance between the two molecules. The atoms selected to perform each scan are colored in blue in Figure 6.2. The non relaxed interaction energy was computed using MP2/aug-cc-pVTZ calculations as a reference, as well as several semiempirical Hamiltonians. AM1, RM1, PM3, PM6, PM6-DH2, PM6-DH+ and PM7 calculations were performed using the MOPAC program.[211] We used the `sqm` module of Amber14[184] for the calculations with PDDG/PM3 and PM3-CARB1, and the `sebond` module of Amber14 for PM3-PIF1 and PM3-MAIS. The Gaussian09[188] program was used for the MP2 calculations.

Figure 6.3 presents the results obtained for a selection of SE methods (*i.e.*, AM1, RM1, PM3, PM3-PIF1, PM3-MAIS, PM6, PM6-DH+ and PM7). The profiles obtained with other Hamiltonians are available as Supplementary Material. For each SE method, the profiles corresponding to the five structures (see Figure 6.2) are plotted in plain lines and the corresponding MP2 reference is reported in dashed lines. Let us first discuss the MP2 results. As expected, the structures D and E do not show any minimum on the related profiles. Both correspond to unstable geometries and thus repulsive profiles. Structure A is the global minimum geometry of the water dimer, followed in increasing interaction energy by structures C and B. The minimum of the profiles A, B and C is located at 1.94 Å, 3.01 Å and 2.30 Å corresponding to an interaction energy of -5.22 kcal/mol, -3.37 kcal/mol and -4.30 kcal/mol, respectively.

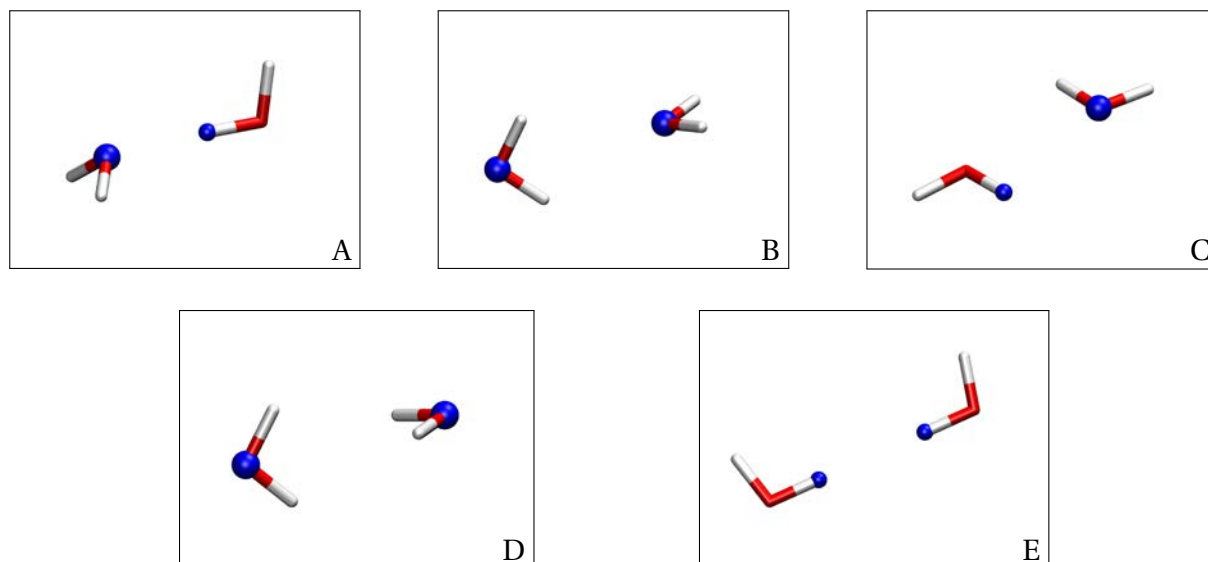


Figure 6.2: Chosen structures of the water dimer. Blue atoms are those chosen to perform the rigid scans.

Among the SE methods presented in Figure 6.2, PM3, PM3-PIF1, PM3-MAIS, PM6-DH+ and PM7 predict profile A to display the lowest energy minimum. However, PM3 and PM7 present typical artifacts of SE Hamiltonians containing Gaussian correction functions in the core-core energy term (see Chapter 4). In the case of PM7, it is noteworthy that the minimum structure of the water dimer was explicitly included in the parameterization set. PM7 predicts indeed a minimum of the profile A located at the correct distance and energy (*i.e.*, 2.00 Å and -4.86 kcal/mol, respectively), but this is not the global minimum. Another minimum is found at a shorter OH distance (*i.e.*, 1.73 Å) and corresponds to an energy of -6.39 kcal/mol. Hostaš *et al.* already pointed out that PM7 predicts a too short Hbond for the water dimer minimum but no further discussion was devoted to the interaction energy surface.[180] This observation is also consistent with the discussion of the interaction energy profile between water and ethanol in Chapter 4. Such a deficiency shows a limitation of PM7, which might arise from a parameterization that only takes into account minimum structures, thus disregarding the whole potential energy surface. PM3 shows artifacts on each of the considered interaction energy profiles.

Only PM3-PIF1 and PM3-MAIS correctly reproduce the profile corresponding to structure B. Nevertheless, the profile obtained with PM3-PIF1 shows that the energy tends to minus infinity when the OO distance is reduced. However, the barrier required to reach this “infinitely deep well” is about ~110 kcal/mol. This does not represent a strong limitation for simulations at room temperature and below, but this observation should be taken into account if one wants to use the PM3-PIF1 Hamiltonian with high temperatures (*e.g.*, replica exchange MD simulations). RM1 predicts a profile close to MP2 for the structure B but presenting an artifact in the repulsive wall.

Finally, most of the SE methods fail at reproducing the correct repulsive profile for the structures D and E. For structure D, AM1 and PM3 show a minimum on the interaction en-

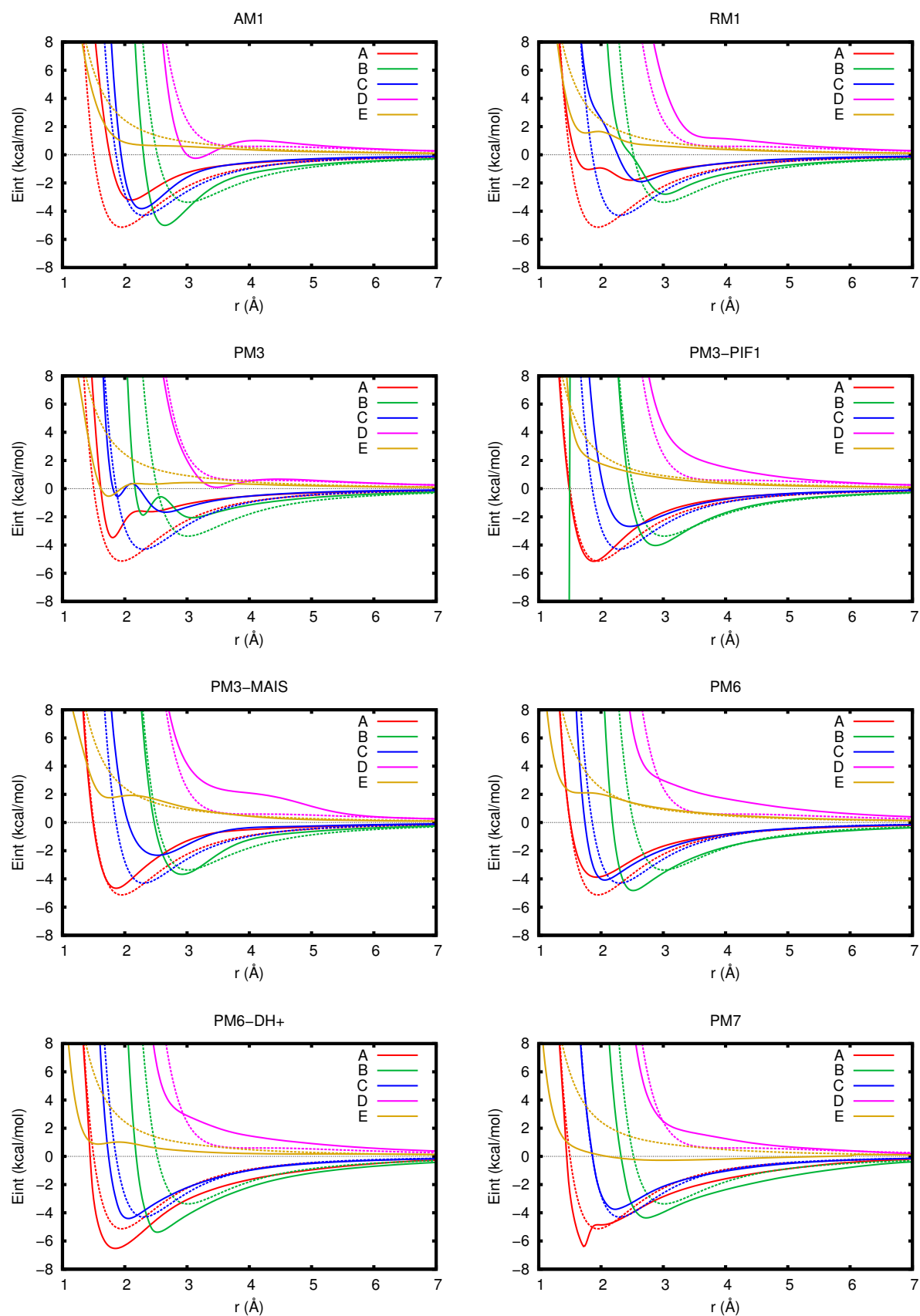


Figure 6.3: Water dimer rigid scan using a chosen set of SE methods. Dashed lines correspond to MP2 results. The schematic representation of the A, B, C, D and E geometries of the water dimer are displayed in Figure 6.2.

ergy profile and the other methods overestimate the repulsion. For structure E, PM3-PIF1 gives a profile in good agreement with MP2 while PM7 underestimates the position of the repulsive wall and the other approaches present a spurious minimum on this profile.

Minimum structure


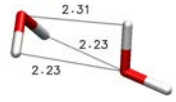
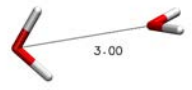
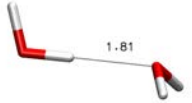
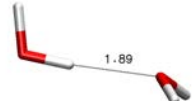
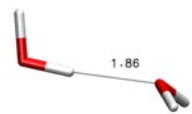
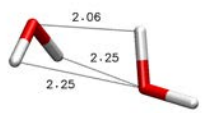
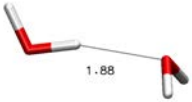
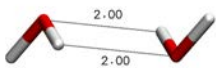



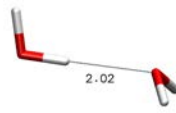
We also performed an extensive research of the possible minima predicted by some semi-empirical methods for the water dimer. From the profiles presented in Figure 6.3 and those available as Supplementary Material, we extracted all the minimum structures of the corresponding method. This led to a set of initial geometries. Each of those geometries was first optimized with all the SE methods considered here, then all the stationary structures were used to build a second set of geometries. Each of the new configurations was optimized by using all of the SE methods and we kept all the structures corresponding to a minimum (*i.e.*, no imaginary frequency) for each method.

This procedure yield several fully optimized minima for each of the considered SE methods. As an example, we found 21 different stable structures of the water dimer using the PM3 Hamiltonian, which is consistent with the numerous artifacts present on the interaction energy surface of this method (see Figure 6.3). We summarized our results in Table 6.1 for a choice of SE methods. We report in this Table the minimum structure optimized at the MP2/aug-cc-pVTZ level and, for each SE method, we give the global minimum that we found with the procedure described above, as well as the structure closest to the MP2 minimum. The quality of each structure is characterized by a root mean square deviation (RMSD) with respect to the MP2 result. We also report for each structure the interaction energy difference with respect to MP2 (*i.e.*, $\Delta E_{int} = E_{int}^{SE} - E_{int}^{MP2}$).

Among the tested methods, only PM3, PM3-PIF1, PM3-MAIS and PM6-DH+ predict a structure close to MP2 to be the global minimum of the water dimer. AM1, RM1, PM6 and PM6-DH2 yield a global minimum with an RMSD higher than 0.5 Å. While we could not find a minimum close to MP2 with AM1 and RM1, we obtained a correct structure with PM6 and PM6-DH2 lying in a range of 0.5-2.0 kcal/mol higher than the global minimum of the corresponding method. The two minima identified on the profile A with PM7 (see Figure 6.3) have been found here. A minimum very close to MP2, both for the energy and the structure (*i.e.*, low RMSD) as well as a global minimum lying at about -1.2 kcal/mol below and presenting a short intermolecular OH distance (*i.e.*, 1.73 Å). Finally, the closest prediction in energy is obtained with PM3-PIF1 and PM3-MAIS: +0.1 kcal/mol and +0.5 kcal/mol with respect to MP2, respectively.

From the analysis of the interaction energy profiles and of the minimum structures of the water dimer, only PM3-PIF1, PM3-MAIS and PM6-DH+ would be recommended to study the dynamics of water. The other methods either fail at reproducing the fundamental water dimer interaction energy profile (*i.e.*, profile A) and/or present nonphysical artifacts. One

Table 6.1: Comparison of the minimum structures found with different semiempirical methods with respect to MP2/aug-cc-pVTZ. $\Delta E_{int} = E_{int}^{SE} - E_{int}^{MP2}$ where $E_{int}^{MP2} = -5.22$ kcal/mol (interaction energy computed without applying the BSSE).

Method	Global minimum [†]		Minimum closest to MP2	
	Structure	ΔE_{int} RMSD//MP2	Structure	ΔE_{int} RMSD//MP2
MP2/aug-cc-pVTZ		0.00 0.000	—	—
AM1		-0.26 0.623	not found	—
RM1		+2.43 0.560	not found	—
PM3		+1.67 0.085	same as global minimum	—
PM3-PIF1		+0.10 0.100	same as global minimum	—
PM3-MAIS		+0.50 0.106	same as global minimum	—
PM6		-1.52 0.584		+1.18 0.090
PM6-DH2		-0.26 0.574		+0.25 0.025
PM6-DH+		-1.52 0.072	same as global minimum	—
PM7		-1.42 0.132		+0.26 0.065

[†] Found in this work.

aim of the present study is to investigate the dynamical properties of water. The use of Hamiltonian presenting artifacts would have a dramatic impact on the stability of a molecular dynamics simulation.

Proton transfer between two water molecules

One motivation of the present work is to model the proton transfer between two neighboring water molecules in $(\text{H}_2\text{O})_{21}$. To assess the feasibility of such a reaction with SE methods, we performed a rigid scan from the MP2/aug-cc-pVTZ minimum structure presented in Figure 6.4. Along this scan, all the atoms were fixed with the exception of the hydrogen atom to be transferred (H_T). We defined the reaction coordinate ζ as:

$$\zeta = R_{\text{O}_\text{D}\text{H}_\text{T}} - R_{\text{O}_\text{A}\text{H}_\text{T}} \quad (6.2)$$

where $R_{\text{O}_\text{D}\text{H}_\text{T}}$ is the distance between H_T and the donor oxygen atom (O_D), and $R_{\text{O}_\text{A}\text{H}_\text{T}}$ is the distance between H_T and the acceptor oxygen atom (O_A).

We present in Figure 6.5a the relative potential energy (ΔE) profile as a function of ζ , obtained at the MP2 and at the PM3-MAIS level. Not surprisingly, both methods predict that ΔE increases with ζ , without reaching any minimum for the dissociated product ($\zeta \simeq 1$). The two profiles seem qualitatively similar, presenting two inflection points (*i.e.*, zero second derivative) at about ~ -0.5 and $\sim +0.5$. However, PM3-MAIS overestimates ΔE when ζ increases.

We analyzed in Figure 6.5b the correlation between the PM3-MAIS and MP2 results. A linear regression without intercept (*i.e.*, $\Delta E_{\text{PM3-MAIS}} = a\Delta E_{\text{MP2}}$) is also reported in the Figure. As suggested by the value of the coefficient of determination $R^2 = 0.999$, a linear relationship exists between the results of the two methods. Thus, we scaled $\Delta E_{\text{PM3-MAIS}}$ by the coefficient obtained from the linear regression ($1/1.216$) and reported the results on the profiles in Figure 6.5a. The qualitative correspondence between the scaled PM3-MAIS results and MP2 is remarkable. This observation is very promising for the rest of the present study. Although the absolute energy values predicted by PM3-MAIS do not fit exactly the MP2 results, the former Hamiltonian reproduces the correct physical features of the proton transfer between two water molecules along the selected reaction coordinate (ζ).

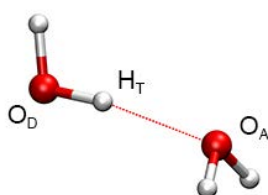


Figure 6.4: Optimized geometry of the water dimer obtained at the MP2/aug-cc-pVTZ level of theory. The label of the atoms used to define the reaction coordinate of the proton transfer is given for the donor oxygen atom (O_D), the acceptor oxygen atom (O_A) and the hydrogen atom to be transferred (H_T).

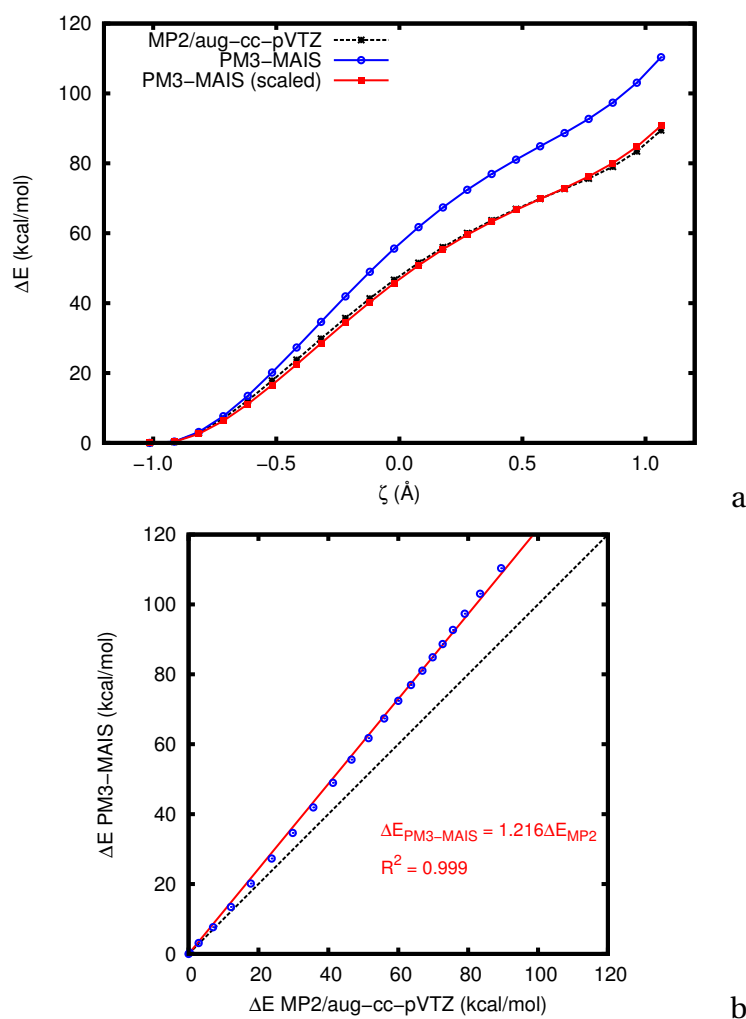


Figure 6.5: a: Relative potential energy (ΔE) profile for the proton transfer between two water molecules using MP2/aug-cc-pVTZ and PM3-MAIS. The profile labeled PM3-MAIS(scaled) was obtained by applying a scaling factor of $1/1.216$ to $\Delta E_{\text{PM3-MAIS}}$. The curves corresponding to MP2 and to PM3-MAIS(scaled) are superimposed. b: Correlation between MP2 and PM3-MAIS and linear regression of the results.

We also performed the same analysis using other SE Hamiltonians and the results are reported as Supplementary Material. Among the tested methods, we note the good performance of AM1, AM1-D and PM3. However, these three Hamiltonians are not capable of reproducing the correct structure of the water dimer, which makes such approaches inadequate for the present study. We also note that the profile obtained with PM7 is not perfectly smooth, which tends to show that this Hamiltonian may bear some discontinuities.

A similar study has been performed by Bernal-Uruchurtu *et al.* for the protonated water dimer.[137] The authors showed a good agreement of the PM3-MAIS predictions with MP2 calculations, for the proton transfer and the interaction energy between H_3O^+ and H_2O . We performed here a similar study for the deprotonated water dimer ($\text{HO}^-/\text{H}_2\text{O}$), which is reported in Supplementary Material. We observed a good agreement of PM3-MAIS with MP2 for the proton transfer. However, it appeared that PM3-MAIS tends to overestimate the interaction energy between water and the hydroxy anion.

Summary

From this preliminary study on the water dimer, only one semiempirical method (among those tested here) appears to be suitable to investigate the structure and reactivity of water: PM3-MAIS. As we have shown, most of the tested methods fail at reproducing the fundamental interaction between two water molecules. Some of them present severe artifacts on the intermolecular potential energy surface of the water dimer and some do not predict the correct minimum for this interaction.

The PM3-PIF and PM6-DH+ methods also give good results for the interaction of two water molecules. However, the former, by definition, cannot be applied to study intermolecular reactivity and the latter is not recommended either since its Hamiltonian is not continuous, as described in Ref. [167].

6.2.2 Ionic model: the Eigen cation

The reaction that we aim to model involves the solvation of ionic species in the relaxation process. The stability of those ions is of key importance to expect finding local minima of the $(\text{H}_2\text{O})_{21}$ water cluster that present a zwitterionic character. To assess the feasibility of such a study with semiempirical methods, we chose to model, as a test case, the proton transfer between H_3O^+ and one neighboring water molecule in the Eigen cation.

The MP2/aug-cc-pVTZ optimized geometry of the Eigen cation is presented in Figure 6.6. This structure is particularly stable since it maximizes the solvation of the H_3O^+ cation. [182, 326] In a similar manner to what has been done for the water dimer above, we performed a rigid scan of the proton transfer. The reaction coordinate was defined by ζ in Eq. 6.2 with the same definition of the two distances involved in the transfer. Due to the constraint applied to the system, this reaction does not yield any local minimum for the product

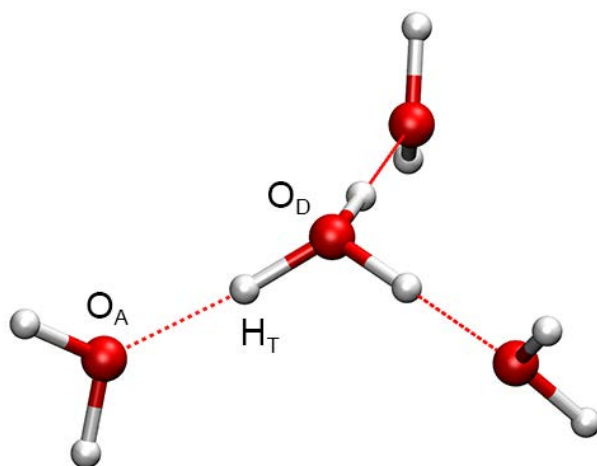


Figure 6.6: MP2.aug-cc-pVTZ optimized geometry of the Eigen cation. The atomic labels used for the proton transfer are also given.

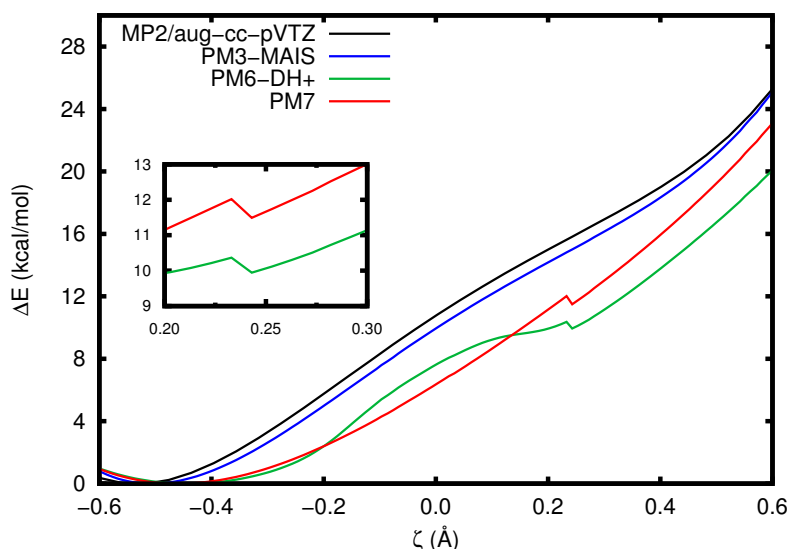


Figure 6.7: Relative energy (ΔE) profiles of the proton transfer between H_3O^+ and H_2O in the Eigen cation using MP2/aug-cc-pVTZ calculations as a reference and some semiempirical Hamiltonians.

as it can be seen on the MP2/aug-cc-pVTZ relative energy profile presented in Figure 6.7.

We performed the same calculation using different semiempirical methods. We chose PM3-MAIS, PM6-DH+ and PM7. The -DH+ co correction of the PM6 Hamiltonian is dedicated to treat hydrogen bonded systems and a similar formalism has been used by Stewart to develop PM7.[141] The results of the three selected semiempirical methods are displayed in Figure 6.7, while those obtained using a wider set of methods are reported in Supplementary Material.

At first glance, it appears clearly that the PM3-MAIS Hamiltonian predicts a profile in very good agreement with the MP2 results.

As it can be seen in the inset of the plot in Figure 6.7, the profile obtained with PM6-DH+ bears a discontinuity between $\zeta = 0.23 \text{ \AA}$ and $\zeta = 0.24 \text{ \AA}$. This is in agreement with the observations of Řezáč *et al.*[167] A similar discontinuity in the PM7 profile, which appears in the same range of ζ values than it is for PM6-DH+, shows the close relationship in the treatment of hydrogen bonds in those two methods. This discontinuity also appears on the profile predicted by AM1-DH+ (see Supplementary Material). Among the other methods reported in Supplementary Material, we note that RM1 predicts a profile in very good agreement with MP2. However, as it has been discussed above, this method cannot be used for the present study, considering its wrong description of the water dimer minimum.

Finally, it is clear from this study that the PM3-MAIS Hamiltonian is well suited to study proton transfer in water. Thus, we shall only comment the results obtained with this method in the rest of this study. We also showed that -DHx type of corrections cannot be applied to such process since the potential energy surface of those methods is not continuous. During the development PM7, the author did not explicitly advance nor deter the applicability of this method to the study of reactive and dynamical systems. Nevertheless, we emphasize

here that the use of the PM7 Hamiltonian to study the structure and reactivity of hydrogen bonded systems is not recommended since this methods bears artifacts and discontinuities on the related potential energy surface.

6.2.3 Zwitterionic minima of $(\text{H}_2\text{O})_{21}$ with PM3-MAIS

As a last test case, we investigated the ability of the PM3-MAIS Hamiltonian to predict local minima of the $(\text{H}_2\text{O})_{21}$ water cluster that present a zwitterionic (Z^\pm) character. Torrent-Sucarrat *et al.* identified two stable Z^\pm geometries of $(\text{H}_2\text{O})_{21}$ optimized at the B3LYP/6-311+G(d) level (labeled as C9 and C10 by the authors).[331] The authors also reported the optimized transition states related to the formation of C9 and C10 from a common “neutral” local minimum of $(\text{H}_2\text{O})_{21}$ (referred here as w21-0): TS6 and TS7, respectively. They characterized energetically those structures from single point calculations at the MP2/6-311+G(2df,2d)//B3LYP/6-311+G(d) and we report their results in Table 6.2.

The relative potential energy of TS6, TS7, C9 and C10 are given with respect to the energy of w21-0. We can see from those results that C10 is the most stable Z^\pm structure of $(\text{H}_2\text{O})_{21}$ found by the authors and that its related transition state bears a relatively high energy with respect to w21-0. C9 appears to be a metastable minium, since the backward energy barrier is almost zero (this cannot be seen from the degree of precision of the results reported in Ref. [331]). We performed single point calculations on the same structure using the PM3-MAIS method and the results are reported in Table 6.2. The energy values calculated with PM3-MAIS are in qualitatively good agreement with the MP2 calculations (*i.e.*, the same trend is obtained for the relative stability of the structures), which is consistent with the discussion in Subsections 6.2.1 and 6.2.2.

We optimized the C9 and C10 structures using PM3-MAIS but all our attempts systematically yield a non zwitterionic water cluster. Considering the high energy values obtained with PM3-MAIS compared to MP2, we hypothesized that the stabilization of the ionic pair with PM3-MAIS would require a reorganization of the hydrogen bond network in the cluster. To test this hypothesis, we performed a 500 ps SEBOMD simulation of the C10 cluster at 75K, by applying a harmonic potential on each OH bond in order to avoid any proton transfer and thus to keep the Z^\pm character of the system. The potential energy of the system is repre-

Table 6.2: Relative potential energy (in kcal/mol) of the structure reported in Ref. [331] calculated at the MP2/6-311+G(2df,2d) and at the PM3-MAIS level.

Method	w21-0	TS6	C9	TS7	C10
MP2/6-311+G(2df,2d)//B3LYP/6-311+G(d) [†]	0.0	17.1	17.1	20.0	9.3
PM3-MAIS//B3LYP/6-311+G(d)	0.0	25.2	25.0	29.7	24.4

[†] Results from Ref. [331]

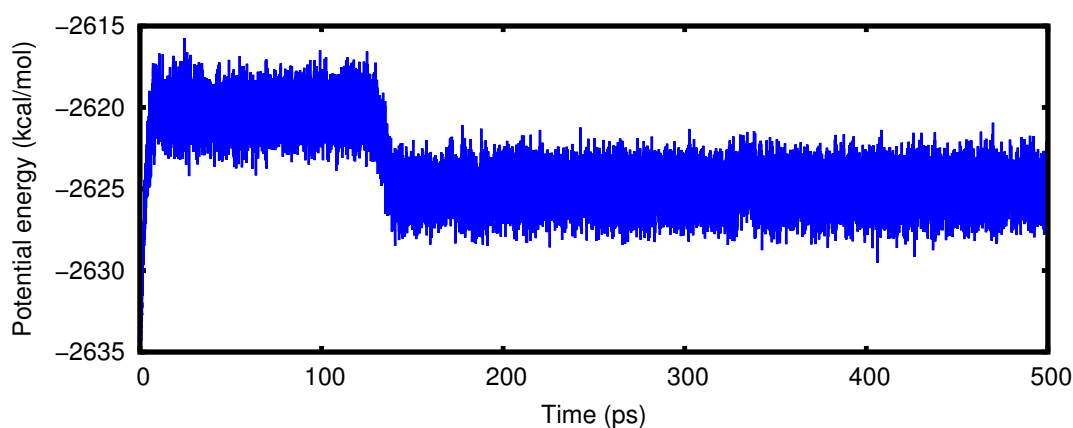


Figure 6.8: Relaxation dynamics of the C10 structure.

sented in Figure 6.8 as a function of the simulation time. Disregarding the heating part of this simulation, we can see that the system remains in a stable basin during the first 100 ps of the dynamics. In the following ps, a reorganization the Hbond network occurs and the system reaches another basin, lower in energy by about ~ 5 kcal/mol. The energy remains constant until the end of the simulation. We used the last point of this dynamics and optimized it by keeping the constraint on the OH bonds. After this optimization step completed, we optimized the structure by relaxing all the constraints. This procedure led to the Z^{\pm} minimum of (H₂O)₂₁ (w21-A- Z^{\pm}) represented in Figure 6.9a. Following the same procedure, we identified two other Z^{\pm} minima, w21-B- Z^{\pm} and w21-C- Z^{\pm} (reported as Supplementary Material) as well as the corresponding neutral forms, w21-B-N and w21-C-N, respectively.

In Table 6.3 we report single point calculations of the relative energy of the three PM3-MAIS Z^{\pm} structures with respect to their corresponding neutral form. We compare in this Table the results obtained using the PM3-MAIS Hamiltonian by applying or not the scaling factor discussed in Subsection 6.2.1 against MP2-aug-cc-pVTZ calculation performed on the

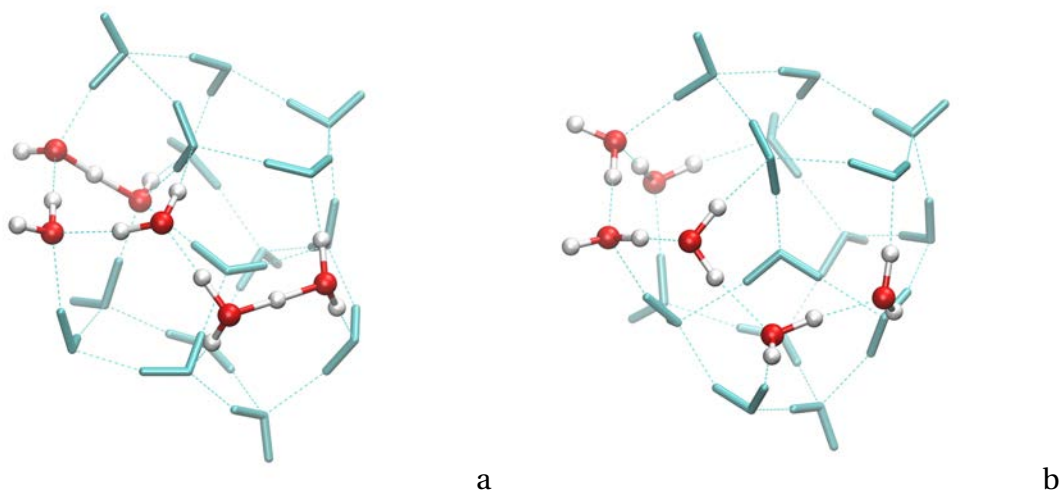


Figure 6.9: Optimized geometries of the (H₂O)₂₁ water cluster at the PM3-MAIS level of theory. a: The zwitterionic minimum w21-A- Z^{\pm} . b: The corresponding “neutral” form w21-A-N. The highlighted water molecules are those involved in the proton transfer.

Table 6.3: Relative potential energy (in kcal/mol) of the three Z^\pm structures (*i.e.*, w21-A- Z^\pm , w21-B- Z^\pm , w21-C- Z^\pm) with respect to their corresponding neutral form. The scaling factor applied to PM3-MAIS is 1/1.216, as discussed in Subsection 6.2.1.

Method	w21-A- Z^\pm	w21-B- Z^\pm	w21-C- Z^\pm
PM3-MAIS//PM3-MAIS	19.01	17.19	20.75
PM3-MAIS(scaled)//PM3-MAIS	15.63	14.14	17.07
MP2/aug-cc-pVTZ//PM3-MAIS	15.34	14.56	18.69

PM3-MAIS minima. The unscaled PM3-MAIS results are systematically higher than the corresponding MP2 energy, which is in good agreement with the discussion of self-dissociation in the water dimer (see Figure 6.5). When applying the scaling factor of 1/1.216 to the PM3-MAIS results, the agreement with MP2 becomes remarkable. While the error for w21-C- Z^\pm is about +1.6 kcal/mol, the relative energy of w21-A- Z^\pm and w21-B- Z^\pm are found in a range of ± 0.5 kcal/mol with respect to the MP2-aug-cc-pVTZ predictions. This result tends to confirm that the MAIS procedure to correct PM3 is well adapted to study such a phenomenon. Although a systematic error is made on the absolute energy, the physics of the process is well reproduced by the method compared to a higher level of theory.

We investigated the stability of w21-A- Z^\pm by means of a short SEBOMD simulation at 75K. The structure was first equilibrated with restraints on the OH bonds and then, the dynamics was continued by relaxing the system. After a few femtoseconds, a proton transfer occurs and the cluster reaches a neutral form. We optimized this structure and the corresponding minimum (w21-A-N) is presented in Figure 6.9b. From the proton transfer dynamics simulation, we observed the mechanism of the transfer (see the movie in Supplementary Material). This reaction involves six water molecules (highlighted in Figure 6.9) through a

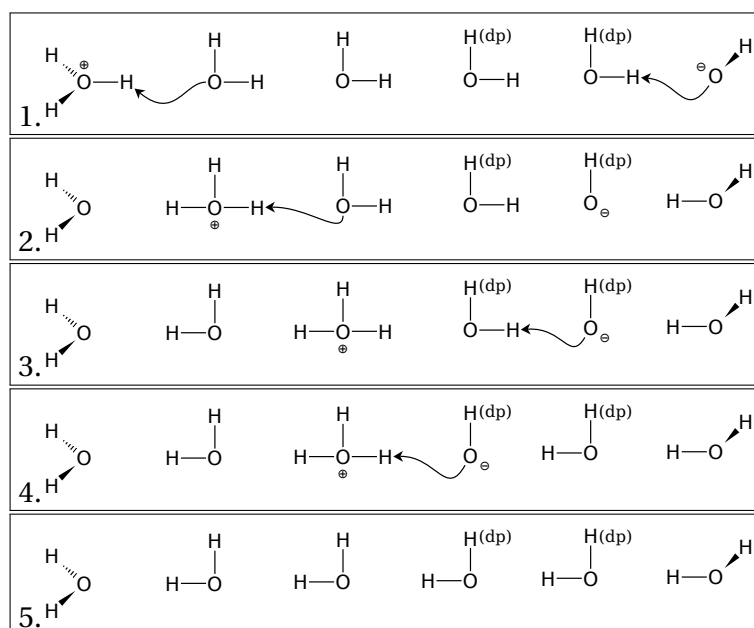


Figure 6.10: Neutralization mechanism from w21-A- Z^\pm to w21-A-N.

complex mechanism that we summarized in Figure 6.10. Our observations on this test case are consistent with a Grothuß-like mechanism. We note that the neutralization process is initiated from both sides of the water chain. Thus, the plus and minus charges are successively moved from one molecule to another, until the neutralization occurs in the last step.

Two main observations arise from this test case:

- The PM3-MAIS Hamiltonian can predict stable Z^\pm structures of the $(\text{H}_2\text{O})_{21}$ water cluster. Moreover, the PM3-MAIS PES is in qualitatively good agreement with MP2/aug-cc-pVTZ results and in quantitative agreement up to a scaling factor.
- The proton transfer reaction in $(\text{H}_2\text{O})_{21}$ implies a complex mechanism involving many water molecules.

As we have seen here, the PM3-MAIS Hamiltonian is well suited to study proton transfer in water and particularly in confined systems. We also showed that the proton transfer is a dynamical process that involves many degrees of freedoms in $(\text{H}_2\text{O})_{21}$. In what follows, we shall perform an extensive analysis of the required characteristics of the water cluster that enhance or prevent the water self-dissociation reaction, using the PM3-MAIS SE method and umbrella sampling simulations.

6.3 Relation between proton transfer free energy and cluster topology

The water self-dissociation reaction is a rare event that results from the specific combination of various collective properties.[317] It has been shown in the literature that during this autoionization process, the first proton transfer between two neighboring water molecules is the rate determining step,[317, 329] the rest of the reaction being a relaxation process *via* a Grothuß-like mechanism to stabilize the resulting $\text{H}_3\text{O}^+/\text{HO}^-$ ionic pair. We propose here to focus on this first step in order to get insights about the required features of the two water molecules involved in the reaction that favor the dissociation reaction, as well as the characteristics of their direct environment. To this end, we performed a series of umbrella sampling calculations in $(\text{H}_2\text{O})_{21}$ and defined several descriptors that we shall discuss in what follows.

As we discussed in the previous Section, the water self-dissociation process involves many degrees of freedoms in $(\text{H}_2\text{O})_{21}$. However, the umbrella sampling technique, as well as most of the free energy calculation methods (see Section 6.1) bears deficiencies to model such complex mechanisms. To apply this method to the present analysis, several approximations are required. We based those approximations on four main considerations:

- The first proton transfer between two neighboring water molecules is the rate determining step.
- The proton transfer between two water molecules is enhanced by the instantaneous

combination of several descriptors.

- A proton transfer occurs in a few femtoseconds.
- The mobility of water molecules is slowed down in confined systems compared to bulk water.

From these observations, it seems reasonable to assume that, during the proton transfer between two neighboring molecules, the global structure of the cluster should not be strongly affected. We thus decided to study this process for fixed instantaneous configurations of the $(\text{H}_2\text{O})_{21}$ water cluster. In what follows, we will consider two initial geometries of $(\text{H}_2\text{O})_{21}$ and freeze the position of all oxygen atoms during the simulations while the hydrogen atoms will remain free. This strategy has two main advantages. First, it reduces the number of degrees of freedom within the cluster, allowing the use of one umbrella sampling reaction coordinate, which we will detail below. Second, we can get insights about the impact of a given Hbond network on the reaction free energy.

6.3.1 Umbrella sampling simulations

We choose two initial geometries of the cluster: a PM3-MAIS minimum (w21-0K) and a snapshot extracted from a SEBOMD/PM3-MAIS simulation at 150K (w21-150K). The same protocol was applied to both of those two initial structures. We first defined all the possible proton transfer between two neighboring water molecules. This procedure was based on the identification of all hydrogen bonds in the water cluster and was performed by applying a cutoff of 2.2 Å on the oxygen hydrogen intermolecular distance. Combining w21-0K and w21-150K, this analysis yielded a total of 65 different proton transfers. In Figure 6.11, we present the initial structure w21-0K, in which we highlighted two identified proton transfers between two neighboring water molecules. For each couple of water molecules, the proton donor and the proton acceptor molecule will be labeled as molD(x) and molA(x), respectively, where x denotes the index of the molecule. In the following, we will mainly illustrate our results on those two examples (the entirety of our calculations is available as Supplementary Material).

For each identified proton transfer, we performed an umbrella sampling simulation using the reaction coordinate (ζ) described by Eq. 6.2 in Subsection 6.2.1. The reaction coordinate was split into 39 windows separated by 0.05 Å. The range of ζ was [-0.95:0.95] going from reactants $\text{HO}_\text{D}-\text{H}_\text{T}\dots\text{O}_\text{A}\text{H}_2$ to products $\text{HO}_\text{D}^-\dots\text{H}_\text{T}-^+\text{O}_\text{A}\text{H}_2$, where O_D and O_A are the proton donor and proton acceptor oxygen atoms, respectively and H_T is the proton to be transferred.

For each window, we performed a constrained simulation to sample a given range of the ζ reaction coordinate in the NVT ensemble with a time step of 0.2 fs. The Andersen thermostat was used to maintain the temperature at 300K. After 2 ps of equilibration, the data production was run for 15 ps. As discussed above, the position of all oxygen atoms was frozen using a belly type MD framework (ibelly=1 in Amber input) while the hydrogen

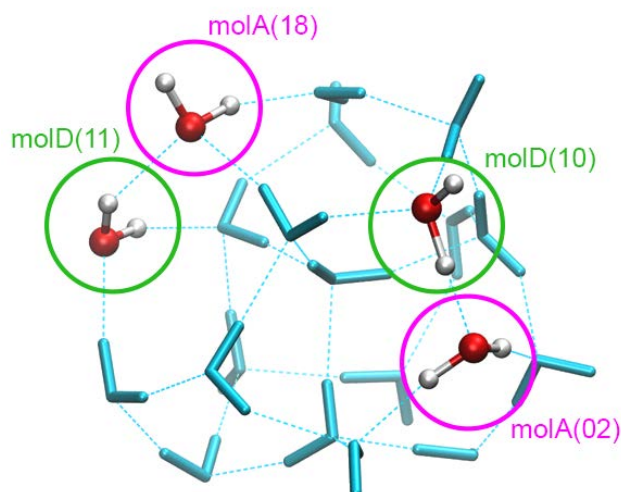


Figure 6.11: Representation of the w21-0K geometry of $(\text{H}_2\text{O})_{21}$ showing to example of identified proton transfer between two neighboring water molecules. For each couple of water molecules, the donor molecule (molD) is circled in green and the acceptor (molA) is circled in purple. The numbers are the indexes of the water molecules.

atoms were free. We set the force constant of the harmonic potential to $1322 \text{ kcal}\cdot\text{\AA}^{-2}\cdot\text{mol}^{-1}$ (661 in the Amber input) following Eq. 1.53 detailed in Subsection 1.4.1. An example of Amber input for the production run in one window is given by:

```
UMB run
&cntrl
  imin=0, ntx=5, irest=1
  cut=12.
  ntc=1, dt=0.0002, nstlim=75000
  ntbf=0, ntp=0
  ntt=2, vrand=5000, ig=-1, tempi=300., temp0=300.
  ntp=25000, ntwx=10, ntwv=25000, ntwe=25000
  ibelly=1, bellymask='%HW'
  nmropt=1
  ifqnt=1
/
&qmmm
  qm_theory='SEBOMD'
&sebomd
  hamiltonian='PM3'
  modif='MAIS'
/
&wt type='DUMPFREQ', istep1=1/
&wt type='END'/
DISANG='rc_0001.RST'
DUMPAVE='rc_0001.dat'
```

with the corresponding restraint file:

```
&rst
  restraint="coordinate(distance(1,2),1.0,distance(52,2),-1.0)"
  r1=-1.45, r2=-0.95, r3=-0.95, r4=-0.45
  rk2=661.0, rk3=661.0
/
```

Each umbrella sampling simulation has been performed several times (up to 5) in order to find the best set of r_0 value for the applied harmonic potentials as discussed in Subsection 1.4.1. The potential of mean force (PMF) of each proton transfer was computed using the WHAM procedure.

6.3.2 Reaction descriptors

The reaction that we intend to model does not lead to any stationary structure (see Section 6.2.1). It results in the formation of an ionic pair that is not stable because of the proximity of the two ions and because of the constraint applied to the position of the oxygen atoms. We first need to define a descriptor that identifies the value of ζ for which the proton transfer can be estimated as completed, as well as the corresponding relative free energy (ΔG).

We report in Figure 6.12 the PMF of the two proton transfers highlighted in Figure 6.11, *i.e.*, from molD(10) to molA(02) and from molD(11) to molA(18) (Figure 6.12a and 6.12b, respectively). As expected, those PMFs do not show any minimum corresponding to the ionic product. We also report on those plots two reaction descriptors:

- The bond order (n_H) of the two oxygen atoms involved in the proton transfer, that are calculated following the definition of Sprik *et al.*[192]. For any oxygen atom O^* , n_H is computed as follows:

$$n_H = \sum_{i=1}^{N_H} S(R_{H_i O^*}) \quad (6.3)$$

where $R_{H_i O^*}$ is the distance between the hydrogen and oxygen atoms of interest, the sum runs around all the N_H hydrogen atoms and $S(R)$ is a switch function that continuously defines if a bond exists between the two particles. Many definitions of $S(R)$ exist in the literature. Here we used the definition of Sprik *et al.*:

$$S(r) = \frac{1}{e^{\kappa(R-p_0)} + 1}. \quad (6.4)$$

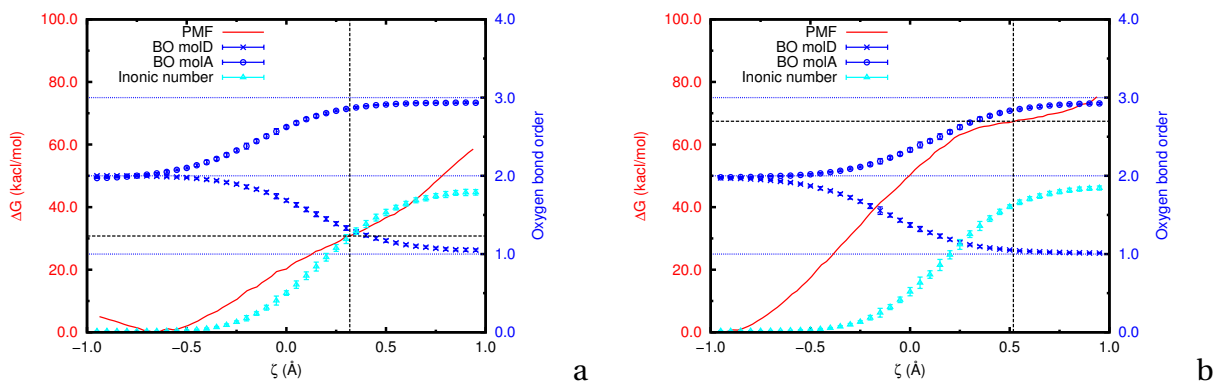


Figure 6.12: Potential of mean force (PMF) of two proton transfers in w21-0K as a function of the reaction coordinate ζ . a: Transfer from molD(10) to molA(02). b: Transfer from molD(11) to molA(18). The bond order of the two oxygen atoms involved in the reaction as well as the ionic number of the cluster are also reported (see text).

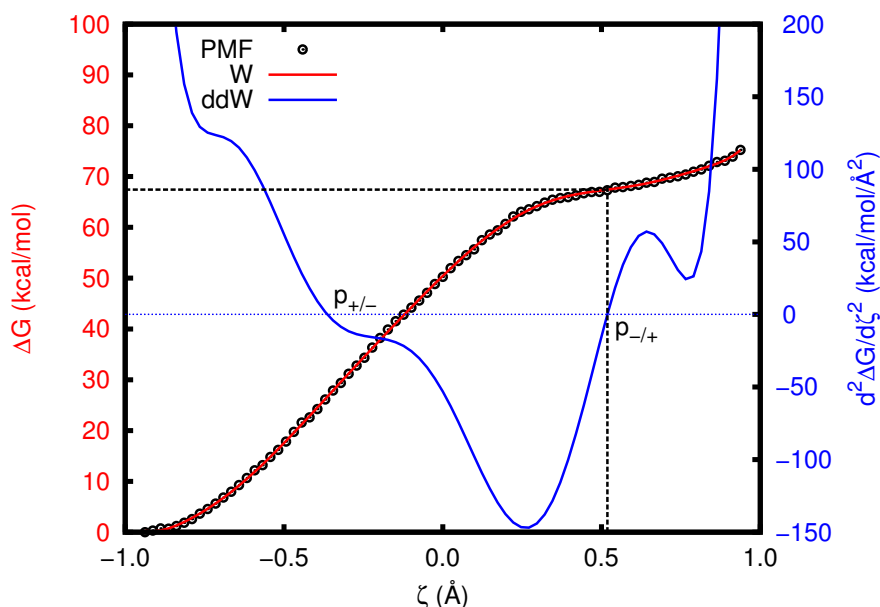


Figure 6.13: Illustration of the procedure used to determine the end of the proton transfer using the identification of the inflection points ($p_{+/-}$ and $p_{-/+}$) of the PMF.

κ^{-1} and ρ_0 are parameters that define the position and the shape of the switch. The authors derived those parameters from the analysis of the OH radial distribution function of water simulated using CPMD: 0.1 and 1.38 Å, respectively.

- The *ionic number* (I^\pm) that characterizes the zwitterionic state of the cluster:

$$I^\pm = \sum_{i=1}^{N_O} (n_H - 2)^2 \quad (6.5)$$

where the sum runs around all the oxygen atoms (N_O). By definition, I^\pm is 0 when all the molecules are neutral while it tends to 2 when an ionic pair is created.

The reaction descriptors (*i.e.*, n_H and I^\pm) allow us to follow the evolution of the reaction as a function of ζ and ensure that the umbrella sampling simulation followed the correct pathway. As it can be seen in Figure 6.12, both descriptors tend to show that the proton transfer has been completed, *i.e.*, that the reaction resulted in the formation of an ionic pair located on the expected molecules. Such descriptors could be used to identify the “end” of the reaction. Nevertheless, a cutoff should be applied and there is no straightforward nor objective way to do so. We also contemplated other geometric descriptors but any definition implies the choice of a cutoff.

To avoid incurring in the use of a cutoff, we propose here another strategy to determine, *a posteriori* the end of the proton transfer. We will define this strategy by describing one proton transfer example. We present in Figure 6.13 the PMF of the transfer from molD(11) to molA(18) fitted using a 10th order polynomial function (W in the Figure). We also report in this plot the second derivative of W with respect to ζ , ddW. We can see that ddW has two

roots in the considered range of ζ values, which correspond to two inflection points in W : $p_{+/-}$ and $p_{-/+}$, where the sign of ddW changes from positive to negative and from negative to positive, respectively. The second inflection point (*i.e.*, $p_{-/+}$) corresponds to the point at which the variation of the free energy rises due to the fixed geometry imposed in the cluster. By identifying the $p_{-/+}$ of each of the 65 profiles, we define the value of ζ that corresponds to the end of the reaction as well as the corresponding value of ΔG . We represented this point in Figure 6.12 by the intersection of the horizontal and vertical dashed lines. We can thus estimate the proton transfer free energy from the plots in Figure 6.12a and 6.12b to be of 30.77 kcal/mol and of 67.43 kcal/mol, respectively.

We used this analysis to define the window corresponding to the end of each proton transfer, the starting one being taken as the minimum of the corresponding PMF. This allows us to perform systematic analysis of the water cluster topology in order to characterize each proton transfer. We shall now introduce the descriptors used to this purpose.

6.3.3 Hydrogen bonds network analysis

We characterized each proton transfer by analyzing the hydrogen bond network in the cluster at the beginning and at the end of the reaction. The corresponding windows were identified following the procedure described in the previous Subsection. For each frame, we built an *instantaneous hydrogen bond connectivity matrix* (IHBCM) between each of the 21 molecules in the cluster. We first assigned each hydrogen atom to one oxygen atom by applying a distance cutoff of 1.2 Å, thus defining a molecule (either a neutral water molecule or an ionic compound, depending on the number of hydrogen atoms). Then, we tested the existence of a hydrogen bond between each couple of molecules (by applying a cutoff of 2.2 Å for the OH intermolecular distances) and stored the results in a matrix by scoring 1 or 0 if the hydrogen bond is found or not, respectively. Finally, all the IHBCM were average per window leading to an *averaged hydrogen bond connectivity matrix* (AHBCM).

In Figures 6.14a and 6.14b, we present the AHBCM corresponding, respectively, to the beginning and to the end of the proton transfer between molD(10) and molA(02). For each matrix, rows and columns give, respectively, the Hbond donor and the Hbond acceptor character of one molecule with respect to another. To help the reading of these matrices, we show in dashed lines the position of the donor and acceptor molecules. In addition, a histogram is given for each molecule, showing the total number of Hbonds given or accepted. The averaged number of Hbond between two molecules is given by a color scale from white to red (0 to 1) for the two first plots.

This representation contains all the information about the Hbond network within the cluster and one can follow the connectivity between all the water molecules. For example, let us consider the AHBCM represented in Figure 6.14a. In the row corresponding to

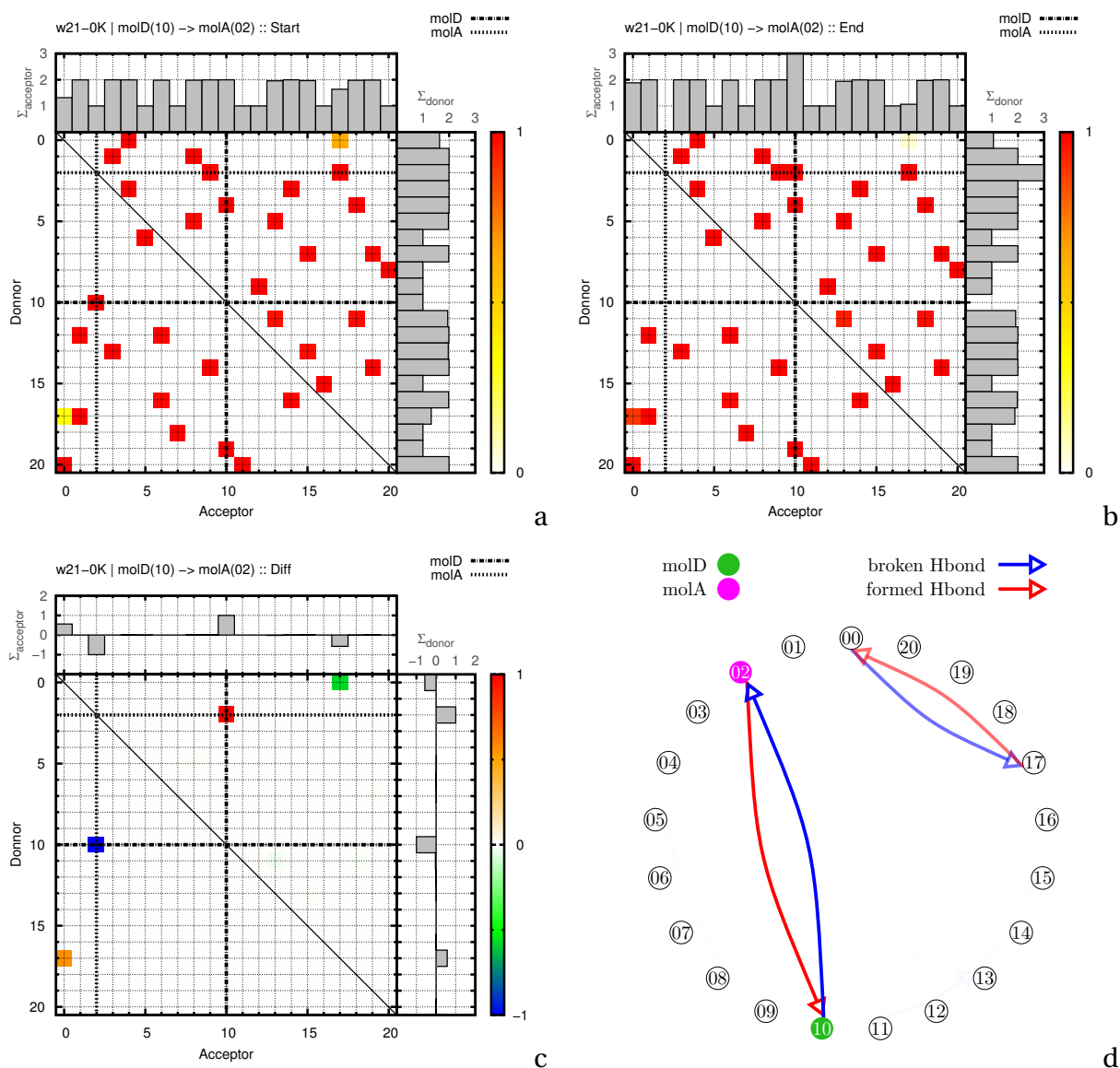


Figure 6.14: Hydrogen bond network analysis of the structure w21-0K for the proton transfer between molD(10) to molA(02). a: Averaged hydrogen bond connectivity matrix (AHBCM) at the beginning of the reaction. b: AHBCM at the end of the reaction. c: Reaction hydrogen bond connectivity matrix (RHBCM) obtained by the difference between the AHBCM's at the end and at the beginning of the reaction. d: Connectivity tree of RHBCM.

molD(10), one can identify only one red square at the position (row,column)=(10,2). This means that the molecule labeled 10 acts as an Hbond donor with only one neighboring molecule: molA(02). This is in agreement with the histogram facing this row that shows a total number of one Hbond formed by molD(10) as a donor. Analyzing the column corresponding to molD(10), we can see two red squares showing that molD(10) receives two Hbonds: one from molecule 4 and another one from molecule 19. Moving to the right from the position (4,10) makes us cross a red square at position (4,18) showing the Hbond donor character of molecule 4 with respect to molecule 18. We can continue this analysis and thus follow all the routes connecting the 21 water molecules of the cluster.

To analyze the modifications induced by the proton transfer on the Hbond network, we computed the difference between the AHBCM's at the end and at the beginning of the reac-

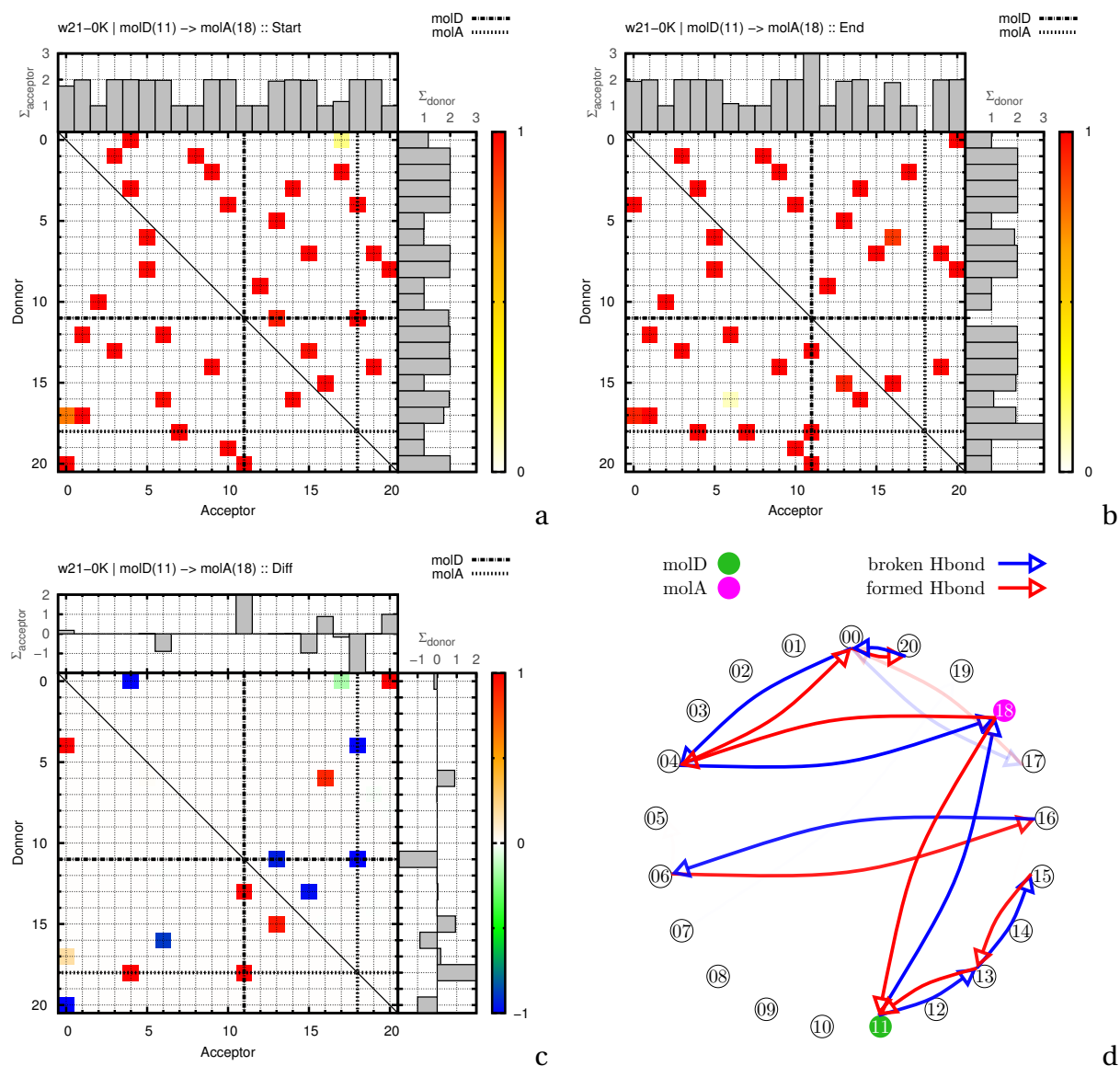


Figure 6.15: Same as Figure 6.15 for the proton transfer between molD(11) and molA(18).

tion (Figures 6.14a and 6.14b, respectively). The resulting *reaction hydrogen bond connectivity matrix* (RHBCM) is represented in Figure 6.14c for the proton transfer between molD(10) and molA(02). The RHBCM shows the Hbonds that have been broken and formed along the reaction. The scale goes from blue to red, corresponding to a range of values between -1 and 1 (0 being represented in white). A value of -1 represents a broken Hbond while 1 corresponds to a formed Hbond. Another representation of the RHBCM is presented in Figure 6.14d as a connectivity tree. Here, we represent by an arrow the hydrogen bond directed from the Hbond donor to the Hbond acceptor molecule. A blue arrow corresponds to a broken Hbond while the formed Hbonds are represented with a red arrow. In addition, an opacity factor is applied to the color of the arrow, equal to the absolute value of the corresponding matrix element. Finally, we identify on this plot the two molecules involved in the proton transfer by a green and by a purple circle, corresponding to molD and molA, respectively.

The analysis of the connectivity tree in Figure 6.14d shows a double arrow between the

molecules labeled as molD(10) and molA(02), which is related to the proton transfer: the molecule molD(10) lost its Hbond donor character by loosing one proton while the molecule molA(02) won this Hbond donor character by gaining this proton. When two water molecules are connected by a hydrogen bond, if the Hbond acceptor molecule has a dangling proton they can exchange their character: the Hbond acceptor becomes donor and the Hbond donor becomes an acceptor with a dangling proton. This situation is typically represented by the second double arrow between the molecules 0 and 17. Finally, this analysis of the Hbond network during the proton transfer between molD(10) and molA(02) shows us that the cluster topology is not much affected by this reaction.

We performed the same analysis for each of the 65 identified proton transfers in w21-0K and w21-150K. As another example, we present this analysis for the transfer between molD(11) and molA(18) in w21-0K in Figure 6.15. The connectivity tree (Figure 6.15d) appears to be more complicated than it was for the previous case (Figure 6.14d). Here, the proton transfer induces many changes in the Hbond network as it can be seen from the numerous double arrows in this plot. One of the most interesting observation that arises from this analysis is that most of the water molecules that experience a change in their Hbond connectivity are directly linked to the molecules involved in the proton transfer. Moreover, when comparing the free energy required for the proton transfer between molD(10) and molA(02) with the one from molD(11) and molA(18) (*i.e.*, 30.77 kcal/mol and of 67.43 kcal/mol, respectively. See Subsection 6.3.2), it seems that the reaction free energy is related to the number of changes in the cluster topology. A small number of Hbond rearrangements appears to favor the proton transfer.

The characterization of the Hbond network within the cluster represents a promising tool to reach a better understanding of the required features that enhance or prevent the proton transfer. However, this is still a visual tool and to reach our goal, we need to extract more quantitative results from this analysis. This is the scope of the next Subsection in which we use the results described above to analyze the environment of each proton transfer.

6.3.4 Environment descriptors

The last family of descriptors that we will consider in this work is intended to give information about the environment of the reaction. Kuo *et al.* [310] already carried out such an analysis on $(\text{H}_2\text{O})_{20}$ and found several orientations of the Hbond network that are energetically unstable and lead to an ionic pair separation with a small activation barrier. The authors characterized such structures in terms of several geometric descriptors detailed below. The main difference between the work by Kuo *et al.* and the case in which we are interested in here, is the extra water molecule that $(\text{H}_2\text{O})_{21}$ bears in the middle of the cage. This observation has an importance in the definition of several parameters which we shall discuss in what follows.

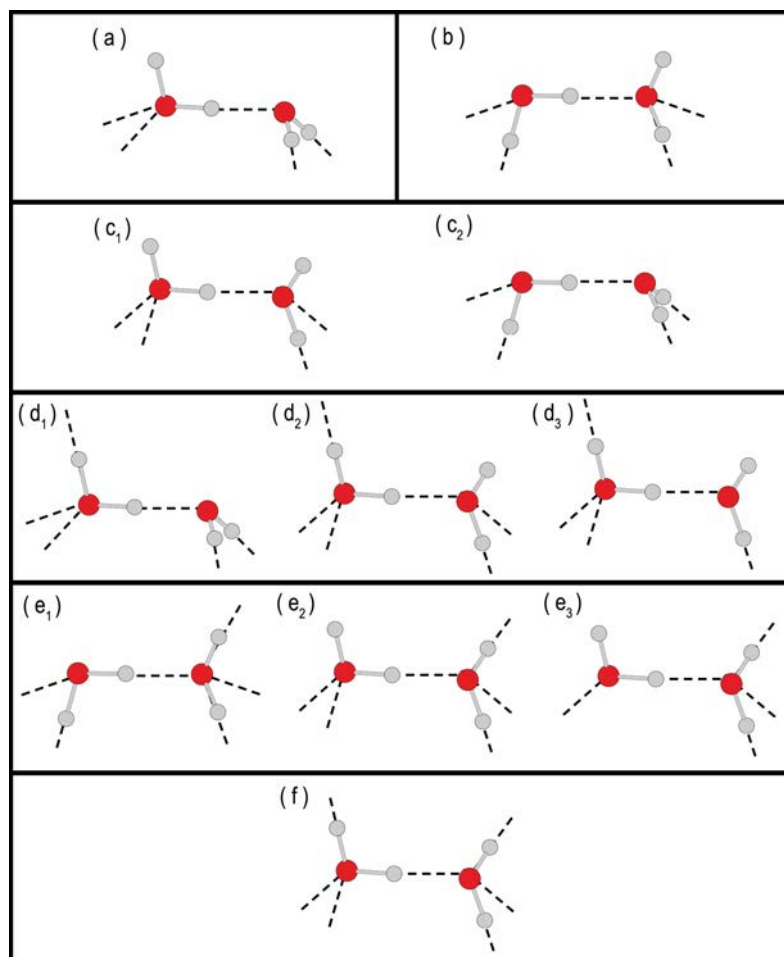


Figure 6.16: Definition of the groups of water dimer geometries. The groups from a to c were defined in Ref. [310].

Before moving on, few words about the nomenclature is necessary. We already introduced some definitions that we will recall here to avoid confusion. In the reaction that we model here, a proton is transferred from one water molecule to another. The molecule that loses the proton is referred as molD (*i.e.*, proton donor molecule) and the one which receives this proton is named as molA (*i.e.*, proton acceptor molecule). We will discuss the character of each water molecule as a function of its direct Hbond environment. 1A and 2A will refer to a water molecule that accepts one or two Hbond(s), respectively. In a similar manner, we define as 1D and 2D, a molecule that is a donor of one or two Hbond(s), respectively.

Kuo *et al.* identified four different possible configurations of two neighboring water molecule within the $(\text{H}_2\text{O})_{20}$ cluster (reported Figure 6.16). They pointed out that the “best” configuration to promote a proton transfer should be of (a) kind. This means that molD should bear a dangling proton (1D,2A) while molA should give two Hbond to its neighbors (2D,1A). This observation is consistent with the idea that the two ions resulting from this transfer should be well solvated. To the opposite, it appears quit clearly that the (b) type of configuration should be the least favorable. (c₁) and (c₂) are intermediate configurations in which one of the two water molecules is well positioned while the other one is not. We will

refer to this groups as (c). Understanding that the (a) type is the most favorable, the authors also investigated the effect of the direct neighbors of the two molecules involved in the transfer. The authors emphasized that the best configuration for the neighbors of molD would be to have two 2A water molecules while it seemed more favorable for molA to be connected to two 2D neighbors. Consequently, the worst configuration would be: molD with two 2D neighbors and molA being connected to two 2A molecules. To quantify the differences between the configurations of (a) type, Kuo *et al.* defined a descriptor (ξ) as follows:

$$\xi = (\text{number of 2D neighbors of molD}) + (\text{number of 2A neighbors of molA}) \quad (6.6)$$

By definition, ξ can take all integer values between 0 and 4, 0 corresponding to the best case and 4 to the worst one. The authors showed a strong correlation between the interatomic distance of molD and molA (R_{ODOA}) and the value of ξ , where ξ takes small values when the $ODOA$ distance is short. This suggests that the probability of having the correct configuration to transfer a proton from one molecule to another increases when the interatomic distance R_{ODOA} decreases.

We will use here the idea and observations of Kuo *et al.* to analyze our proton transfers. As introduced above, $(H_2O)_{21}$ bears one water molecule in its interior. Compared to the work by Kuo *et al.*, we need to define at least three additional families of water couples.

- (d) : molD accepts and gives 2 Hbonds
 - (d_1) : molA accepts 1 and gives 2 Hbonds
 - (d_2) : molA accepts 2 and gives 1 Hbonds
 - (d_3) : molA accepts 1 and gives 1 Hbonds
- (e) : molA accepts and gives 2 Hbonds
 - (e_1) : molD accepts 1 and gives 2 Hbonds
 - (e_2) : molD accepts 2 and gives 1 Hbonds
 - (e_3) : molD accepts 1 and gives 1 Hbonds
- (f) : both molD and molA accepts and gives 2 Hbonds

We reported a sketch representing all the possible configurations discussed above in Figure 6.16. We performed this analysis from the average hydrogen bond connectivity matrices, presented in the previous Subsection, to characterize the initial configuration of the cluster (*i.e.*, using the AHBCM at the beginning of the reaction).

Another descriptor that can quantify the role of the “non-reactive” water molecules of the cluster (*i.e.*, the 19 water molecules not directly involved in the proton transfer) is the electric field that they produce on the proton to be transferred (H_T). To focus only on the impact of the orientation of each water molecule on this electric field, we considered the point

charges defined by the SPC/E force field. This avoids any problematic definition related to the QM charge transfer between the water molecules. We computed this electric field (\vec{E}_{HT}^e) for each proton transfer as:

$$\vec{E}_{\text{HT}}^e = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{Q_i \vec{R}_{\text{HT}i}}{R_{\text{HT}i}^3} \quad i \notin [\text{molD}, \text{molA}] \quad (6.7)$$

where the sum runs over all the atoms that are not included in molD and molA, Q_i is the SPC/E point charge of the atom i and ϵ_0 is the vacuum permittivity. In order to quantify the impact of \vec{E}_{HT}^e on the proton transfer free energy, we focus on the projection of this vector on the $\vec{O}_{\text{D}}\vec{O}_{\text{A}}$ vector, referred as $P_{\text{OO}}(\mathbf{E})$ in what follows, where the bold characters denote the vectorial quantities.

We shall now discuss the correlation between the environments descriptors defined above and the proton transfer free energy.

6.3.5 Data correlation

For each of the 65 proton transfers identified in w21-0K and w21-150K, we computed the transfer free energy (ΔG), the distance between the two oxygen atoms (R_{ODOA}), the Hbond network and the projection of the environment electric field ($P_{\text{OO}}(\mathbf{E})$) as described in the previous Subsections. The entirety of these results is available as Supplementary Material.

Table 6.4 summarizes those results per conformation family. We can see that the (a) type of configurations is the most favorable to transfer a proton. The averaged free energy of transfer of all the other groups is found to be in a 10 kcal/mol range while the (a) group lies about 15 kcal/mol below the most favorable one, (f). The free energy rises when R_{ODOA} increases and when $P_{\text{OO}}(\mathbf{E})$ decreases. Again, quite an important gap exists between the (a)

Table 6.4: Averages obtained for each type of water couple configurations: proton transfer relative free energy $\overline{\Delta G}$, the $\text{O}_{\text{D}}\text{O}_{\text{A}}$ distance $\overline{R_{\text{ODOA}}}$ and the environment electric field projection $\overline{P_{\text{OO}}(\mathbf{E})}$. Results are ordered by increasing free energy of transfer. The # symbol gives the number of configuration taken into account in the average. Standard deviation are given for all the averaged quantities.

Conf. type	#	$\overline{\Delta G}$ (kcal/mol)			$\overline{R_{\text{ODOA}}}$ (Å)			$\overline{P_{\text{OO}}(\mathbf{E})}$ (V/nm)		
a	6	36	±	8	2.63	±	0.05	+10	±	1
f	7	51	±	4	2.75	±	0.03	+6	±	1
c	11	51	±	12	2.76	±	0.08	+5	±	2
d	11	56	±	9	2.78	±	0.09	+5	±	1
e	12	56	±	10	2.80	±	0.12	+4	±	2
b	18	63	±	10	2.83	±	0.09	+3	±	2

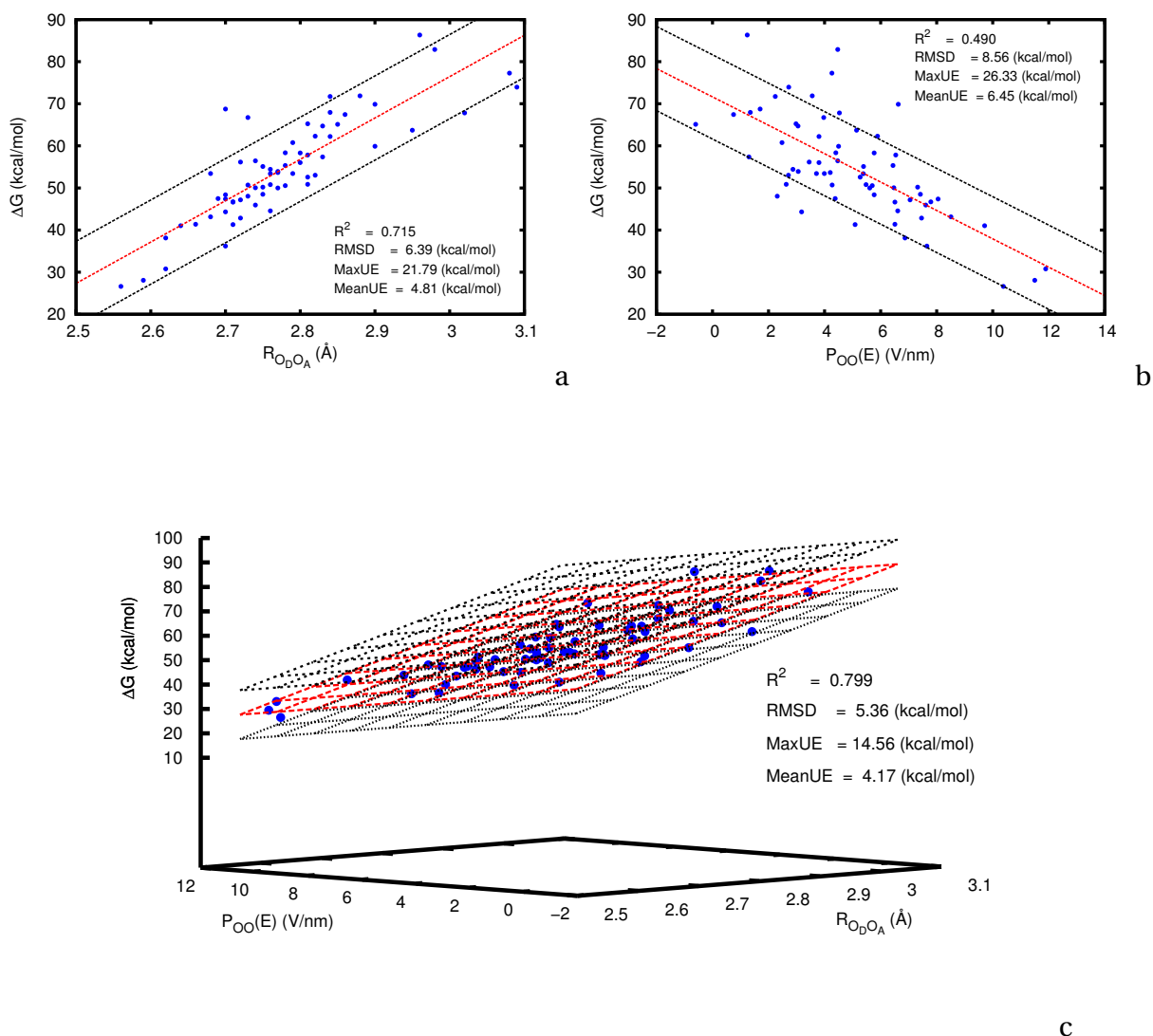


Figure 6.17: Correlation between the proton transfer free energy and the two descriptors detailed in the text. a: ΔG versus R_{ODOA} . b: ΔG versus $P_{OO}(E)$. c: Three dimensional correlation between ΔG , R_{ODOA} and $P_{OO}(E)$. The coefficient of determination (R^2), the root mean square deviation (RMSD), the maximum unsigned error (UE) and the mean UE are displayed for each fit in the corresponding plot.

group and the others for those two descriptors. From the definition of the (a) group of configurations (Figure 6.16), it appears that the proton transfer is more favorable at the surface than in the inner part of the cluster. This observation is in good agreement with the hypothesis that the resulting ion pair is stabilized at the surface of a water cluster.[313, 319–325]

From Table 6.4 (as well as from the complete set of data presented as Supplementary Material), it appears that a correlation between ΔG and R_{ODOA} exists as well as a correlation between ΔG and $P_{OO}(E)$. The correlation between ΔG and R_{ODOA} was already suggested by Mrázek *et al.* but from a much smaller set of configurations.[329] Concerning the projection of the environment electric field, a positive value of $P_{OO}(E)$ goes in the direction of a transfer of the proton along the \overrightarrow{ODOA} axis. It seems reasonable to think that a large value of $P_{OO}(E)$ would favor this process. It is also in good agreement with the findings of Kuo *et al.* about the type of configuration since $P_{OO}(E)$ bear a larger value in the (a) configuration type.

We present in Figures 6.17a and 6.17b the correlation between ΔG and R_{ODOA} as well as the correlation between ΔG and $P_{\text{OO}}(\mathbf{E})$, respectively. Even though the observations made above seem valid, no linear correlation exists here when looking at the very low values of determination coefficient (R^2) for the linear regression in both cases. In Figure 6.17c, we plotted ΔG as a function of both R_{ODOA} and $P_{\text{OO}}(\mathbf{E})$, and fit our results with a plan using the following equation:

$$\Delta G = \alpha R_{\text{ODOA}} + \beta P_{\text{OO}}(\mathbf{E}) + \delta \quad (6.8)$$

where $\alpha = 76.68$, $\beta = -1.66$ and $\delta = -150.04$. Here, the correlation appears to be more clear. The determination coefficient reaches almost 0.8 and the root mean square deviation is about 5.4 kcal/mol. Considering the large range of free energy values in our set of data (*i.e.*, between 27 and 86 kcal/mol) such a deviation is small enough to validate the three dimensional correlation of ΔG with R_{ODOA} and $P_{\text{OO}}(\mathbf{E})$.

The present study confirms the hypothesis of Mrázek *et al.*, who showed a correlation between the proton transfer free energy and a short intermolecular distance. It also confirms the observation of Kuo *et al.*, who stated that the proton transfer between two neighboring water molecule is enhanced by their direct environment topology. The descriptor introduced here to characterize the electric field caused by the environment of those two water molecules shows a similar trend to the geometrical descriptor of Kuo *et al.* However, $P_{\text{OO}}(\mathbf{E})$ is a more quantitative and objective descriptor than ξ , since no assumption needs to be made to determine the hydrogen bond network of the water cluster.

We shall now use these two descriptors (*i.e.*, R_{ODOA} and $P_{\text{OO}}(\mathbf{E})$) to analyze the instantaneous configurations of $(\text{H}_2\text{O})_{21}$ along a SEBOMD simulation. We expect to predict which structure is about to favor a proton transfer using the above considerations.

6.3.6 Prediction of favorable situations

To generate a sufficient amount of configurations, we performed a 500 ps SEBOMD simulation. We first considered a simulation at low temperature (*i.e.*, 150K) but this prevents the system to reorganize in a reasonable time scale. We thus increased the temperature up to 300K. To avoid the system to evaporate, a spherical potential was applied on the cluster. This potential consists in a soft half harmonic potential that prevents the atoms to move further than a given distance from the geometrical center of the cluster. After several tests, we found preferable to place this harmonic potential at 3.8 Å from the center of the system. We computed the accessible surface of the system using the *cpptraj* module of the AmberTools14 (with the default settings) and used this surface to compute the radius of the cluster, by assuming a spherical shape. This led to an averaged radius of 5.09 Å and a corresponding density of 1.135 g/cm³ to be compared with the experimental value measured by Amararene, 1.136 g/cm³ (corresponding to an apparent specific volume of 0.88.10⁻³ m³/kg in a non charged reverse micelles, with $\omega_0 = 1$).[335]

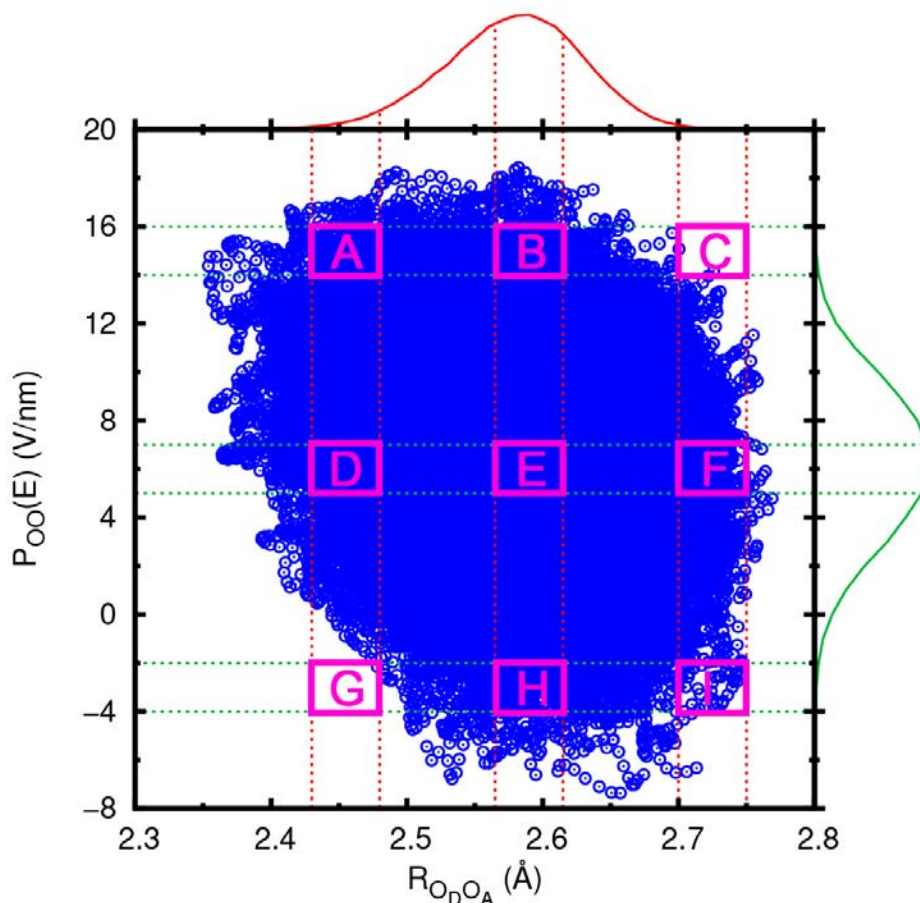


Figure 6.18: Representation of the “best” combination of $R_{O_D O_A}$ and $P_{OO}(E)$ (see text) for each frame of the SEBOMD simulation. The probability of $R_{O_D O_A}$ and $P_{OO}(E)$ are projected on the corresponding axis. The nine selected zones (from A to I) described in the text are also represented.

For each frame of the simulation, we analyzed the possible proton transfers and selected the one presenting the shortest $O_D O_A$ distance. For this particular proton transfer, we also computed the projection of the environment electric field experienced by the proton (H_T) as described above. A representation of $P_{OO}(E)$ with respect to $R_{O_D O_A}$ at each time step of the simulation is plotted in Figure 6.18. The distribution of points is projected on the two axes of this plot, given the probability of $R_{O_D O_A}$ and $P_{OO}(E)$ along the simulation.

From Figure 6.18, we can see that the values of the two descriptors are spread in a relatively large range: $R_{O_D O_A} \in [2.35 : 2.80]$ Å and $P_{OO}(E) \in [-8 : 20]$ V/nm. However, the probability projection of the two descriptors shows that the most probable combination is found for $R_{O_D O_A} = 2.59$ Å and $P_{OO}(E) = 6.0$ V/nm and that this probability decreases rapidly around those values.

In order to assess the accuracy of the couple of descriptors developed in the previous Subsection, we defined nine zones (labeled from A to I) of this distribution as reported in Figure 6.18 and in the three first columns of Table 6.5. From each of these zones, we extracted randomly one representative structure, also reported in the Table. We used the parameterized equation obtained in Eq. 6.8 to compute the expected value of the proton transfer free

Table 6.5: Definition of the regions reported in Figure 6.18 and details about the selected structures (see text). The expected and calculated free energy of each proton transfer (ΔG^\ddagger and ΔG , receptively) are reported in kcal/mol. The range of ΔG^\ddagger is given with respect to the mean unsigned error reported in Figure 6.17c (*i.e.*, ± 4 kcal/mol).

Zone	R_{ODOA}	$P_{OO}(E)$	Number of matches	Selected structure		ΔG^\ddagger	ΔG
	range (Å)	range (V/nm)		R_{ODOA}	$P_{OO}(E)$		
A	2.430:2.480	14.0:16.0	349	2.430	15.0	[7:15]	17.2
B	2.565:2.615	14.0:16.0	826	2.591	14.8	[20:28]	30.0
C	2.700:2.750	14.0:16.0	2	2.726	14.2	[32:40]	41.6
D	2.430:2.480	5.0:7.0	2217	2.475	5.7	[26:34]	19.2
E	2.565:2.615	5.0:7.0	46380	2.594	7.0	[33:41]	35.3
F	2.700:2.750	5.0:7.0	685	2.722	6.9	[43:51]	48.0
G	2.430:2.480	-3.0:-1.0	9	2.474	-1.8	[39:47]	23.0
H	2.565:2.615	-3.0:-1.0	1230	2.609	-1.5	[49:57]	44.4
I	2.700:2.750	-3.0:-1.0	78	2.722	-2.0	[58:66]	52.3

energy (ΔG^\ddagger). Then, we performed the umbrella sampling simulation for each of the selected structures to model the proton transfer, by means of the same protocol used previously. The resulting proton transfer free energy (ΔG) as well as ΔG^\ddagger are reported in Table 6.5.

As expected from the probability displayed in Figure 6.18, only the region E is well represented by a large amount of structures, while the zones C and G contain almost no matching configurations. The value of ΔG is outside the range of predicted values for most of the chosen configurations. Only for the configurations in the zones E and F the prediction matches the effective value. Nevertheless, as expected, the structure coming from zone A bears the lowest free energy. Moreover, the trend of the variation of free energy as a function of R_{ODOA} and $P_{OO}(E)$ obtained from ΔG^\ddagger and ΔG show a similar behavior, only the variation in the diagonal (G,E,C) is inverted.

The main factor that affects the feasibility of a proton transfer between two neighboring molecules appears to be the distance between the two oxygen atoms. To a lower extent, the electric field produced by the environment of the two considered water molecules also plays a role to enhance or to prevent the reaction.

Although the present results have been obtained from a limited number of configurations of the $(H_2O)_{21}$ water cluster, they give a clear picture of the factors influencing the first step of the proton transfer. The use of the two descriptors proposed here results to be very promising and can be used in other studies of larger water clusters.

6.4 Concluding remarks

The first part of the present Chapter was intended to give a global picture of the applicability of semiempirical NDDO methods to study proton transfers in water. We have shown that most of the currently available SE Hamiltonians fail at describing this process as well as the fundamental structure of the water dimer. From our test calculation, only the PM3-MAIS Hamiltonian seems to be suitable to perform such a study.

We applied the PM3-MAIS Hamiltonian to study the water self-dissociation reaction in confined systems. We chose as a model system, the $(\text{H}_2\text{O})_{21}$ water cluster in gas phase. By focusing on the first step of this complex reaction, we designed a protocol to elucidate the particular combinations of features that enhance or prevent a proton transfer between two neighboring molecules. We showed that the distance between the two oxygen atoms involved in the reaction is, as already suggested in the literature, the main factor. We also showed that the direction and intensity of the electric field caused by the environment of the two reacting molecules plays a significant role to lower the proton transfer free energy.

To assess the evaluation of the free energy associated with water self-dissociation in a more general way, one needs to take into account the relaxation of the cluster and the stabilization of the ionic pair that result from an initial proton transfer. Among the currently available methods to perform free energy calculations, two of them are widely used to model chemical reactions: the umbrella sampling and the metadynamics technique. In a preliminary study, we have tested the applicability of these two methods.

As we discussed in Section 6.2, the water self-dissociation reaction is related with a complex mechanism that involves many degrees of freedom in the system. The reaction coordinate to consider in order to model this reaction should thus contain many dimensions. The umbrella sampling technique is not well adapted to treat such a multi-dimensional reaction coordinate. This method requires to control the state of the system for each value of the reaction coordinate and this is not possible in such a case. The model system must then be simplified to apply umbrella sampling. The metadynamics technique appears to be better adapted to such a study. In particular, the use of one (or two) collective variable(s) that control the coordination number of one (or two) oxygen atom(s) seems quite appealing. We used the implementation of such a collective variable performed in the present work (see Section 3.4) and metadynamics simulations based on it are currently under way.

The results obtained in this study will constitute a solid base of knowledge for further studies of the water self-dissociation reactions. The next step of this work would be to enlarge the size of the model water cluster and to use the criteria defined here to select relevant structures in order to evaluate the free energy of reaction and thus estimate the related rate constant in confined systems.



Conclusions générales

Modéliser à la fois la structure électronique des systèmes bio-moléculaires et leurs propriétés dynamiques est l'un des plus grands défis de la chimie théorique moderne. Au cours de ces dernières années, différentes approches ont été développées pour répondre à cette question, telles que les approches multi-échelles et la dynamique moléculaire Car-Parrinello. Alors que la première permet de traiter des systèmes de grande taille durant de longues échelles de temps, en utilisant une description hybride de mécanique quantique/mécanique moléculaire, la seconde étend le traitement quantique à la totalité du système mais ne peut atteindre que des temps de simulation limités. Dans notre travail, nous cherchons à étudier les propriétés dynamiques de molécules biologiques solvatées en gardant une description quantique du système dans son entier.

Certains compromis sont nécessaires afin d'accélérer les calculs le long d'une simulation de dynamique moléculaire. Au travers de l'utilisation de SEBOMD, nous avons fait le choix d'utiliser des méthodes quantiques approchées afin de modéliser l'Hamiltonien électronique, et en particulier, nous nous sommes focalisés sur les méthodes semiempiriques de type NDDO. La méthode SEBOMD a été développée récemment dans notre groupe et, bien que quelques études préliminaires [74, 169] ont montré SEBOMD comme une méthodologie prometteuse, il était nécessaire d'apporter des réponses à certaines questions afin de permettre une utilisation plus poussée de la méthode.

Dans ce travail, nous nous sommes intéressés à quelques questions fondamentales, en contribuant à la fois au développement théorique de méthodes quantiques approchées et aux applications à l'étude des propriétés électroniques et vibrationnelles de molécules solvatées. Ces questions sont résumées dans ce qui suit.

Les méthodes semiempiriques sont-elles suffisamment fiables pour fournir des résultats qualitatifs et quantitatifs au regard de valeurs expérimentales et de modèles basés sur un niveau de théorie plus élevé ? À l'origine, les méthodes semiempiriques (SE) ont été développées pour traiter des systèmes isolés. Malgré les efforts importants qui ont été réalisés au cours des dernières années, nous montrons ici que la plupart des méthodes SE disponibles aujourd'hui présentent des difficultés à traiter certaines interactions intermoléculaires fondamentales.

Dans le contexte de simulations de la phase condensée, il est primordial de détenir un

modèle capable de reproduire ces interactions. Il a été montré dans notre groupe que l'utilisation d'une fonction d'interaction paramétrisable (PIF) peut fournir des améliorations significatives dans le cas de systèmes aqueux. Cependant, nous avons montré que la dernière version des paramètres (PM3-PIF2) ne permet pas d'obtenir une description satisfaisante de l'interaction de l'eau, à la fois avec des groupements hydrophobes et hydrophiles. De plus, les autres méthodes SE disponibles ne permettent pas d'apporter une alternative valide à l'approche PIF. Pour résoudre ce problème, nous avons développé et paramétrisé une nouvelle correction des méthodes SE, qui prend en compte de manière explicite les interactions entre certains groupes fonctionnels importants et leur environnement. Cet Hamiltonien SE dépendant du type d'atome, que nous avons appelé PM3-PIF3, a été appliqué à la simulation de molécules organiques dans l'eau. Les résultats des simulations SEBOMD PM3-PIF3 fournissent un accord remarquable avec d'autres travaux expérimentaux et théoriques pour la prédiction des effets du solvant sur des composés organiques.

Envisager des développements futurs requiert d'apporter une réponse à la question suivante : faut-il continuer de corriger les méthodes existantes en modifiant la partie cœur-cœur du terme d'interaction ou est-il nécessaire de reprendre à zéro le développement des Hamiltoniens semiempiriques ? Une extension de la méthode PM3-PIF3 nécessiterait d'introduire d'autres types d'atomes, conduisant à une méthode de mécanique quantique assimilable à un champ de forces. D'autre part, travailler au développement de nouveaux Hamiltoniens nécessiterait probablement une paramétrisation tenant compte de données provenant de la phase condensée et de traiter séparément les paramètres électroniques et les paramètres de cœur durant le processus de paramétrisation.

La méthode SEBOMD est-elle suffisamment efficace pour permettre de longues dynamiques moléculaires, nécessaires à la convergence de nombreuses propriétés statistiques, notamment pour l'étude des propriétés vibrationnelles de bio-molécules ? Dans ce travail, nous avons réalisé les simulations d'une série de 11 molécules organiques dans l'eau. Chaque système était composé d'environ 400 atomes et a été simulé durant 500 ps. De plus, nous avons mené une étude approfondie du transfert de proton dans l'agrégat d'eau $(\text{H}_2\text{O})_{21}$, qui a nécessité de réaliser 65 simulations par *umbrella sampling*.

La méthode SEBOMD nous a permis de réaliser ce nombre important de simulations dans un temps relativement court et en utilisant une quantité moindre de moyens informatiques comparé à d'autres approches. De plus, nous avons montré dans le cadre des simulations de composés organiques que l'échelle de temps de nos simulations était, en général, suffisant pour permettre la convergence des propriétés dynamiques, telles que les spectres infrarouges de ces molécules. Néanmoins, dans le cas d'un modèle biologique plus complexe (le dipeptide alanine), il apparaît qu'une simulation de 500 ps est toujours trop courte pour permettre un échantillonnage correct de l'espace conformationnel.

Cependant, le faible coût de calculs de notre méthode rend possible l'accès à des échelles

de temps plus grandes. L'extension de la simulation du dipeptide alanine est en cours. De plus, des tests préliminaires ont montré que la méthode SEBOMD pourrait être couplée à des algorithmes d'échantillonnages efficaces tels que la méthode de dynamique moléculaire *replica exchange*. Ces calculs prometteurs seront au cœur de prochaines études.

La méthode SEBOMD peut-elle traiter des systèmes réactifs ? Nous avons tiré avantage d'une extension du modèle PIF menant à une fonction de correction permettant de traiter des systèmes aqueux réactifs, l'Hamiltonien PM3-MAIS. Nous avons testé son application à l'étude du transfert de proton dans $(\text{H}_2\text{O})_{21}$.

Nous avons tout d'abord montré que les prédictions de PM3-MAIS sont en très bon accord avec les résultats obtenus au niveau MP2/aug-cc-pVTZ pour le calcul de la surface d'énergie potentielle de systèmes modèles pertinents. De plus, nous avons montré que la plupart des méthodes SE courantes ne peut pas être appliquée à l'étude des systèmes réactifs en milieu aqueux. Finalement, nous avons caractérisé la première étape du processus d'autoprotolyse de l'eau dans l'agrégat $(\text{H}_2\text{O})_{21}$, aux moyens d'un modèle basé sur des simulations SEBOMD par *umbrella sampling*, en trouvant quelques facteurs clés intervenant dans le mécanisme réactionnel.

La possibilité de modéliser une réaction dans un environnement entièrement quantique conduit également à de nouvelles problématiques techniques. Beaucoup d'approches théoriques adaptées à l'étude de la réactivité de systèmes dynamiques sont basées sur l'hypothèse que la partie réactive du système est relativement localisée. Cette approximation n'est pas valide, par exemple, dans le cas de la déprotonation d'un acide aminé dans l'eau. Un modèle plus réaliste devrait prendre en compte le fait que la réaction n'est pas contrainte à une région donnée de l'espace et que le système entier y participe, particulièrement lorsque le solvant est impliqué.

Les simulations SEBOMD apportent-elles plus de réponses que leurs équivalents au niveau MM ou QM/MM ? Comparé à des études MM, nous avons montré qu'il est nécessaire de prendre en compte le comportement quantique des propriétés électroniques du système afin de reproduire les observations expérimentales. Un exemple typique est donné par le décalage en fréquence et l'élargissement de la bande correspondant au mode de vibration OH sur le spectre infrarouge de l'éthanol, qui est caractéristique de l'effet du solvant sur ces types de composés. L'utilisation d'une description quantique du système dans son entier permet de prendre en compte les effets de polarisation mutuelle de la molécule d'intérêt et de son environnement lors du calcul de ces propriétés. Nous avons montré, en revanche, qu'une description MM ne permet pas de rendre compte de cet effet.

Comparé à un modèle QM/MM, une description quantique du système complet apparaît comme nécessaire pour rendre compte du caractère non local de la fonction d'onde lors de la modélisation d'une réaction. Lors de notre étude SEBOMD du transfert de proton dans $(\text{H}_2\text{O})_{21}$ nous avons montré qu'un arrangement collectif spatial et électronique du système

complet est nécessaire pour modéliser ce type de réaction.

Finalement, ce travail ouvre la porte vers une utilisation plus poussée de la méthode SE-BOMD. Il montre aussi la nécessité d'analyser en profondeur les performances des méthodes semiempiriques quant à la prédiction des interactions intermoléculaires. Bien que de futurs développements méthodologiques soient nécessaires, l'utilisation d'un modèle capable de traiter efficacement à la fois la structure électronique et les propriétés moléculaires dépendantes du temps fournira une description plus réaliste de la structure, de la dynamique et de la réactivité des systèmes biologiques.



General conclusions

Modeling together the electronic structure of biomolecular systems and their dynamical properties is one of the greatest challenges of modern theoretical chemistry. During the past decades, different approaches have been developed to tackle this question, such as multi-scale methods and Car-Parrinello molecular dynamics. While the former allows to treat large systems during long time scale simulations by using a hybrid quantum mechanics/molecular mechanics description, the latter extends the quantum chemistry treatment to the whole system but can only be run on shorter time scales. In our work, we are interested in the dynamical properties of solvated biomolecules by keeping a quantum description of the whole system.

Indeed, some compromises need to be made to attain fast calculations along a molecular dynamics simulation. Through the use of SEBOMD, we made the choice of using approximate quantum methods to model the electronic Hamiltonian, in particular we focused on NDDO semiempirical methods. The SEBOMD methodology has been developed very recently in our group and, though some preliminary studies[74, 169] had showed it to be very promising, some questions needed to be answered to make a more extensive application possible.

In this work we have focused on a few fundamental questions, by contributing both to the theoretical development of approximated quantum methods and to applications to the study of electronic and vibrational properties of solvated molecules. Such questions are summarized in the following.

How do semiempirical quantum methods perform compared to experimental measurements and to higher levels of theory to obtain electronic and vibrational molecular properties? Originally, semiempirical (SE) methods were developed to treat isolated systems. Despite the great effort that have been made during the last decades, we show here that most of the currently available SE methods present deficiencies to treat some fundamental intermolecular interactions.

In the context of condensed phase simulations, it is crucial to have a model that is able to reproduce these interactions. It has been shown in our group that the use of a parameterizable interaction function (PIF) can provide some significant improvements in the case of aqueous systems. However, we have found that the existing PIF2 parametrization (i.e., PM3-

PIF2) is not sufficient to reach a qualitative and quantitative description of the interaction of water with both hydrophilic and hydrophobic groups. In addition, other currently available SE methods do not provide a valid alternative to the PIF approach. To overcome this issue, we developed and parameterized a new correction of SE methods, which explicitly accounts for the interaction between some important functional groups and their environment. This atom type dependent SE Hamiltonian, which we named PM3-PIF3, was applied to the simulation of organic molecules in water. The results of the PM3-PIF3 SEBOMD simulation showed a remarkable agreement with other experimental and theoretical works for the prediction of the solvation effects on organic compounds.

Envisaging further developments of this work is related to a basic question: should we continue to correct existing methods by modifying the core-core interaction term or is it necessary to start over the development of semiempirical Hamiltonians? An extension of the PM3-PIF3 method would require introducing other atom types, leading to the definition of a force field-like QM method. On the other hand, working at new SE Hamiltonians would probably require a new parametrization by explicitly taking into account condensed phase data and by separating the treatment of the electronic and of the core parameters during the parameterization procedure.

Is the SEBOMD method efficient enough for running molecular dynamics simulations on a time scale allowing the convergence of statistical properties, especially in the study of the vibrational properties of biomolecules? In this work, we have performed simulations of a series of 11 organic molecules in water. The systems were composed by about 400 atoms and were simulated for 500 ps. In addition, we conducted an extensive study of the proton transfer in the $(\text{H}_2\text{O})_{21}$ water cluster, which required to perform 65 umbrella sampling simulations.

The SEBOMD methodology has allowed us to perform this large amount of simulations in a relatively short time and with a reasonable use of computational resources compared to other approaches. Moreover, we have shown in the case of the simulation of organic compounds that the time scale of our simulations is, in general, sufficient to provide the convergence of dynamical properties, such as the infrared spectra of the molecules. However, in the case of a more complex biological model (i.e., the alanine dipeptide), a 500 ps simulation is still too short to expect a good sampling of the conformational space.

Nevertheless, considering the modest computational cost of our method, larger time scales are indeed accessible. The extension of the alanine dipeptide simulation is currently under way. In addition, some preliminary tests show that SEBOMD can be coupled with efficient sampling algorithms such as replica exchange molecular dynamics. These promising calculations will be the focus of further studies.

Can reactive systems be treated with SEBOMD? We have taken advantage of an extension of the PIF model to a correction function allowing to treat reactive aqueous systems,

the PM3-MAIS Hamiltonian. We have thus tested its application to study proton transfer in $(\text{H}_2\text{O})_{21}$.

First, we have shown that the PM3-MAIS predictions are in good agreement with the results obtained at the MP2/aug-cc-pVTZ level for the potential energy surface of relevant model systems. In addition, we have found that most of the currently available SE methods cannot be applied to study the reactivity of aqueous systems. Finally, we have characterized the first step of the water self-dissociation in the $(\text{H}_2\text{O})_{21}$ cluster by means of a model based on umbrella sampling SEBOMD simulations, finding a few key factors intervening in the reaction mechanism.

The possibility to model a reaction in a fully QM environment also yields new technical issues. Many theoretical techniques adapted to study the reactivity of dynamical systems are based on the assumption that the reactive part of the system is relatively localized. This approximation fails, for instance, in treating the deprotonation of an aminoacid in water. A realistic model should take into account the fact that the reaction is not constrained to a given spacial region and that the entire system participates, especially when the solvent is involved.

Do SEBOMD simulations bring more information than their MM or QM/MM counterparts? Compared to MM studies, we have shown by analyzing the solvent effects on biomolecules that it is necessary to account for the quantum behavior of the system electronic properties in order to reproduce experimental observations. A typical example is given by the shift and broadening of the band corresponding to the OH stretching vibrational mode in the infrared spectrum of the ethanol molecule in solution, which is characteristic of the effect of the solvent on such compounds. The use of a full quantum description of the system allows to account for the mutual polarization of the molecule and of its surrounding in the computation of such properties. We have shown that an MM description is not capable of reproducing this effect.

Compared to a hybrid QM/MM model, the quantum description of the entire system appears to be required to model reactivity, accounting for the non local character of the total wave function. In our SEBOMD study of proton transfer in $(\text{H}_2\text{O})_{21}$ we have shown that a cooperative spatial and electronic arrangement of the entire system is necessary to model such reaction.

Finally, this work opens the door toward a more extensive use of the SEBOMD methodology. It also points out the necessity of a deep analysis of the performance of semiempirical methods with respect to the prediction of intermolecular interactions. Although further methodological developments will be needed, the use of a model capable of efficiently treating the electronic structure together with the time dependent molecular properties will provide a more realistic description of the structure, the dynamics and the reactivity of biological systems.



Supplementary material

Additional contents are available as an external medium. One can access it by running the `SupplementaryMaterial.html` script contained in the attached CD-ROM using any Web Browser.

List of Figures

1.1	Schematic representation of a simulation box in two dimensions (x,y), using periodic boundary conditions (PBC). The unit cell having edges of length a_x and a_y (in this example: $a_x = a_y$), is represented in the middle by a black square. The green dashed square is the representation of the box around the particle A. R is the distance between A and the particle B inside the unit cell while R_{im} is the same distance after applying the minimum image convention (see text).	26
1.2	Schematic representation of the Ewald summation for a distribution of point charges along a one dimensional axis. Plain lines represent the positive and negative point charges while dotted lines indicate the position of the Gaussian function in the reciprocal space.[48]	28
1.3	Schematic representation of a free energy (G) surface as a function of a given reaction coordinate (ζ). The activation free energy barrier is represented in red and labeled as ΔG^a while the free energy difference between the states A and B (ΔG) is displayed in blue.	31
1.4	Schematic representation of the umbrella sampling technique applied to obtain the potential of mean force (PMF, bottom right panel) of a given transformation from the biased free energy profile (G, left panel) <i>via</i> the computation of the biased probability in each window (top right panel).	32
1.5	Schematic representation of the metadynamics technique. The successive steps separated by $n\delta t$ and $(n+m)\delta t$ (with $n, m \in \mathbb{N}$ and δt the time step of the simulation) show the flooding of the free energy surface with Gaussian functions. The green dot represents an instantaneous state of the system.	35
1.6	Infrared spectra computed from a 500 ps gas phase SEBOMD simulation of ethanol in the NVE ensemble (IR^{NVE}) and in the NVT (IR_{av}^{NVT}).	40
1.7	Convergence of the IR spectrum as a function of the simulation time. The simulation were performed using the SEBOMD method and the PM3-PIF3 Hamiltonian in the NVT ensemble. In each case, one solute is solvated in a box of 128 water molecules. a: ethanol. b: N-methylacetamide (NMA).	42

2.1	Time line of semiempirical method development with the related reference publications.	50
2.2	Representation of $g^{\text{PM3}}(\text{A,B})$ as a function of the interatomic distance for two couples of atoms: hydrogen-oxygen (H,O) and hydrogen hydrogen (H,H).	58
3.1	Representation of various smoothing functions (see text). The parameters of each function have been optimized to fit the initial set of points (red dots): P_1, \dots, P_5 . These parameters are displayed on the corresponding plot.	79
3.2	Pair radial distribution functions of water from the SEBOMD simulations using various semiempirical Hamiltonians. Left panel: oxygen-oxygen RDF. Middle panel: oxygen-hydrogen RDF. Right panel: hydrogen-hydrogen RDF. For each plot, the experimental result by Soper[66] is represented using red dashed lines and the corresponding RDFs obtained from SEBOMD simulations are given using plain blue lines.	81
3.3	Instantaneous geometries from SEBOMD PM3-PIF2 simulations of liquid water. a: using Mulliken charges in the Ewald summation. b: using CM1 charges in the Ewald summation. The dipole moment of each water molecule is represented by a blue arrow.	83
3.4	Time evolution of the C_θ descriptor (see text) and of its standard deviation along PM3-PIF2 SEBOMD simulations with PBC. Left panel: no long range interactions, only the minimum convention is applied. Middle panel: long range interactions are taken into account using a standard Ewald scheme and Mulliken atomic partial charges. Right panel: same as for the middle panel using CM1 atomic partial charges for the Ewald summation.	84
3.5	Variation of the C_θ descriptor for SEBOMD simulations performed using a particle mesh Ewald scheme for long range interactions with different sets of atomic charges (q_H).	86
4.1	Performance of a choice of semiempirical methods to reproduce the MP2/aug-cc-pVTZ curves (black dashed curves) describing the interaction between the H atom of water and the H atom of methane (left plots) and between the O atom of water and the H atom of methane (right plots).	96
4.2	Definition of the ϕ angle used to describe the orientational dependence of the interaction energy, and the corresponding plot of the surface obtained at the MP2/aug-cc-pVTZ level.	97
4.3	Orientational dependence of the methane-water interaction energy, using the ϕ angle, for a choice of semiempirical methods.	97

4.4	Interaction between the H atom of water and the H atom (top plot) and the O atom (middle plots) of a methane molecule: comparison between the MP2 curve (black dashed line) and the curve obtained with the PM3-PIF2 Hamiltonian (blue line). In the bottom plot (c) we report the angular dependence of the PM3-PIF2 interaction energy.	98
4.5	Same as in Figure 4.4, for the PM3-PIF3 Hamiltonian.	102
4.6	Optimized geometry of the cluster formed by methane in the 5 ¹² cage of water molecules obtained using PM3-PIF3.	104
4.7	Interaction between the H atom of water and the O atom of the hydroxyl group of ethanol (a) as well as the N atom of tri-methylamine (b): performance of a choice of semiempirical methods compared to the MP2 reference (black dashed line, reported in each plot). In the case of the PM3-PIF3* curve, the parameters for all hydrogen atoms are those obtained for 'hydrophobic' H _C 's (see text).	108
5.1	Comparison of radial pair distribution functions of methane in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF. Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (Ow) with respect to the methane carbon atom (C). b: Pair distribution of water hydrogen atoms (Hw) with respect to the methane hydrogen atoms (H).	120
5.2	Comparison of radial pair distribution functions of isobutane simulated with ff03-SPC/E, PM3-PIF3 and the HH-Alkane united atoms potential.[237] Plain lines: RDF. Dashed lines: RDF integral. a: Isobutane center carbon atom (Cc)-water oxygen RDF. b: Isobutane methyl carbon atom (Cme) Ow RDF	122
5.3	Comparison of radial pair distribution functions of benzene simulated with ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Plain lines: RDF. Dashed lines: RDF integral. a: Benzene center of mass (π)-water hydrogen RDF. b: Benzene center of mass-water oxygen RDF	123
5.4	Definition of the three planes used to produce volume slices of the spatial distribution functions (SDFs) of water around benzene.	124
5.5	Volume slices of water hydrogen atoms (Hw) spatial distribution function (SDF) around benzene for the condensed phases dynamics using ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Large and small white circles are respectively the projections of benzene carbon and hydrogen atoms in the xy and yz planes. Squares correspond to the projection of the CC aromatic bonds in the xz plane.	125
5.6	Volume slices of water oxygen atoms (Ow) spatial distribution function (SDF) around benzene for the condensed phases dynamics using ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Large and small white circles are respectively the projections of benzene carbon and hydrogen atoms in the xy and yz planes. Squares correspond to the projection of the CC aromatic bonds in the xz plane.	126

5.7	Comparison of radial pair distribution functions of toluene simulated with ff03-SPC/E, PM3-PIF2 and PM3-PIF3. Plain lines: RDF, dashed lines: RDF integral. a: Toluene geometric center of the phenyl ring (π)-water hydrogen RDF. b: Toluene methyl carbon atom (Cme)-water oxygen RDF.	127
5.8	Distribution of instantaneous electronic properties of the four hydrophobic molecules along the PM3-PIF3 SEBOMD simulations. a: CM1 total charge of the solute in condensed phase. b: Solute dipole moment computed from CM1 atomic partial charges. Comparison between gas phase (dotted line) and condensed phase (plain line).	128
5.9	Schematic representation of the four vibrational normal modes of the methane molecule.	131
5.10	Comparison of infrared spectra of methane obtained with PM3-PIF3 and ff03. a: full frequency range. b: zoom on the high frequency region. In both a and b, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.	133
5.11	Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.	134
5.12	Decomposition of the VDOS of isobutane into atomic contributions simulated in the gas phase with PM3 (top panel) and ff03 (bottom panel). Intensities are reported in arbitrary units.	135
5.13	Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03 in the frequency range between 2800 and 3300 cm^{-1} . The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line. . . .	136
5.14	Decomposition of the VDOS of benzene simulated in the gas phase with PM3 (top panel) and ff03 (bottom panel). Intensities are reported in arbitrary units. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.	137
5.15	Decomposition of the VDOS of benzene simulated in the condensed phase with PM3-PIF3 (top panel) and ff03-SPC/E (bottom panel). Intensities are reported in arbitrary units. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.	138
5.16	PM3-PIF3 IR spectra of benzene in water computed with different charge models: Mulliken, CM1 and MM point charges. The position (vertical dashed lines) and the frequency value of the four IR vibrational bands are also displayed.	139

5.17	Comparison of infrared spectra of isobutane obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line. In both plots, a zoom on the 3000-3500 cm^{-1} region is presented as an inset.	140
5.18	Illustration of <i>trans</i> and <i>gauche</i> conformers of ethanol and atom label definition. The ω dihedral angle is defined as $C_1 - C_2 - O_a - H_a$	142
5.19	Comparison of the ethanol ω dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.	143
5.20	Comparison of radial pair distribution functions of ethanol in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (O_w) with respect to the ethanol oxygen atom (O_a). b: C_1O_w . c: C_2O_w	144
5.21	p-ethylphenol labels definition. ω_1 and ω_2 dihedral angles are defined as $(C_m^2 - C_p - C_2 - C_1)$ and $(C_o^1 - C_{ox} - O_a - H_a)$, respectively.	145
5.22	p-ethylphenol potential energy surface in the gas phase as a function of the two dihedral angles ω_1 and ω_2 . ΔE_1 and ΔE_2 are positioned at $(0^\circ, 0^\circ)$ and $(-90^\circ, -90^\circ)$, respectively.	146
5.23	Comparison of the p-ethylphenol ω_1 dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.	147
5.24	Comparison of the p-ethylphenol ω_2 dihedral angle distribution in the gas phase and in aqueous solution. Top panel: SEBOMD simulations. Bottom panel: MM-MD simulations.	147
5.25	Comparison of radial pair distribution functions of p-ethylphenol in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF Dashed lines: RDF integral. a: Pair distribution of water oxygen atoms (O_w) with respect to the ethanol oxygen atom (O_a). b: πH_w . c: C_1O_w	148
5.26	Comparison of radial pair distribution functions of trimethylamine in a box of 128 water molecules using various Hamiltonians. Plain lines: RDF Dashed lines: RDF integral. a: pair distribution of water hydrogen atoms (H_w) with respect to the trimethylamine nitrogen atom. The OPLS-AA TIP4P results have been obtained from Ref. [264]. b: pair distribution of water oxygen atoms (O_w) with respect to the carbon atoms of the trimethylamine methyl groups.	149
5.27	Labels definition for the three amide molecules.	150
5.28	Distribution of the ω dihedral angle along the gas and the condensed phase dynamics. Comparison between formamide, propanamide and N-methylacetamide. For PM3 and PM3-PIF3, the peptidic correction (PEP) was applied.	151

5.29	Distribution of the δ improper angle along the gas and the condensed phase dynamics. Comparison between formamide, propanamide and N-methylacetamide. For PM3 and PM3-PIF3, the peptidic correction (PEP) was applied.	152
5.30	Comparison of OHw RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.	153
5.31	Comparison of NHw RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.	153
5.32	Comparison of HOW RDFs from MM-MD (a) and SEBOMD (b) simulations of formamide (FA), propanamide (PA) and N-methylacetamide (NMA) in solution. * Unpublished results from Ref. [74] obtained using PM3-PIF2 without PEP correction.	154
5.33	Distribution of instantaneous electronic properties of the six hydrophilic molecules along the PM3-PIF3 SEBOMD simulations. a: CM1 total charge of the solute in condensed phase. b: Solute dipole moment computed from CM1 atomic partial charges. Comparison between gas phase (dotted line) and condensed phase (plain line).	155
5.34	Comparison of infrared spectra of ethanol (a) and p-ethylphenol (b) obtained with PM3-PIF3 and ff03. For each molecule, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.	157
5.35	Comparison of infrared spectra of thrimethylamine obtained with PM3-PIF3 and ff03. The top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.	158
5.36	Schematic representation of the four amide vibrational modes in the <i>trans</i> -N-methylacetamide (NMA) molecule. The plain black arrows and dotted gray arrows show the main and secondary contributions, respectively.	160
5.37	Comparison of infrared spectra of formamide (a) propanamide (b) and N-methylacetamide (c) obtained with PM3-PIF3 and ff03. For each molecule, the top panel is related to SEBOMD simulations and the bottom one to MM-MD. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line.	161

5.38	Comparison of infrared spectra of formamide (top panel) propanamide (middle panel) and N-methylacetamide (bottom panel) obtained with PM3-PIF3. Gas phase results are given in dashed red line while condensed phase ones are shown in plain blue line. Peak assignment was made from the VDOS decomposition into atomic contributions.	162
5.39	Schematic representation of the four main configurations of the alanine dipeptide and definition of the two dihedral angles (ϕ and ψ). $\phi = C^{ACE} - N^{ALA} - \alpha C^{ALA} - C^{ALA}$ and $\psi = N^{ALA} - \alpha C^{ALA} - C^{ALA} - N^{NME}$	164
5.40	Potential energy surfaces of the isolated alanine dipeptide computed with ff03 (left panel) and PM3 without (middle panel) and with (right panel) peptidic correction. A representation of the different conformational regions as well as indicators of the corresponding structures are also represented.	165
5.41	Values of the ϕ and ψ angles along the gas and the condensed phase MM-MD and SEBOMD simulations (left panel) and the related distribution of probability (right panel). On the left panel, we show the space delimitation of the three basins as defined in Ref. [226] In the left panel, the plain lines correspond to the projection of the distributions on the ϕ and ψ axes.	167
5.42	Comparison of the CM1 alanine dipeptide dipole moment distribution along the SEBOMD simulations in the gas and in the condensed phase. The average dipole moment within the different basins is given in red for the gas phase (GP) and in blue for the condensed phase (Aq). In parenthesis, we report the population of each basin along the simulation.	168
5.43	Comparison of OHw RDFs from MM-MD (a) and SEBOMD (b) simulations for the ACE and the ALA carbonyl group of 2Ala in solution.	169
5.44	Comparison of HOw RDFs from MM-MD (a) and SEBOMD (b) simulations for the H atom bonded to the nitrogen in the ALA and the NME residue of 2Ala in solution.	170
5.45	Comparison of COw RDFs from MM-MD (a) and SEBOMD (b) simulations for the three methyl C atoms of 2Ala in solution.	170
5.46	Comparison of the infrared spectra of the alanine dipeptide obtained from SEBOMD (top panel) and MM-MD (bottom panel) simulations. Gas phase results are shown as dashed red lines while condensed phase ones are shown as plain blue lines.	172
5.47	Comparison of the infrared spectra of the alanine dipeptide obtained from the different basins of the Ramachandran map along the SEBOMD simulations in solution using the PM3-PIF3 Hamiltonian. a: AmI frequency region. b: AmA frequency region.	173
6.1	5^{12} structures of the $(H_2O)_{20}$ (left) and the $(H_2O)_{21}$ (right) clusters, optimized at the PM3-MAIS level of theory.	183

6.2	Chosen structures of the water dimer. Blue atoms are those chosen to perform the rigid scans.	188
6.3	Water dimer rigid scan using a chosen set of SE methods. Dashed lines correspond to MP2 results. The schematic representation of the A, B, C, D and E geometries of the water dimer are displayed in Figure 6.2.	189
6.4	Optimized geometry of the water dimer obtained at the MP2/aug-cc-pVTZ level of theory. The label of the atoms used to define the reaction coordinate of the proton transfer is given for the donor oxygen atom (O_D), the acceptor oxygen atom (O_A) and the hydrogen atom to be transferred (H_T).	192
6.5	a: Relative potential energy (ΔE) profile for the proton transfer between two water molecules using MP2/aug-cc-pVTZ and PM3-MAIS. The profile labeled PM3-MAIS(scaled) was obtained by applying a scaling factor of $1/1.216$ to $\Delta E_{PM3-MAIS}$. The curves corresponding to MP2 and to PM3-MAIS(scaled) are superimposed. b: Correlation between MP2 and PM3-MAIS and linear regression of the results. .	193
6.6	MP2.aug-cc-pVTZ optimized geometry of the Eigen cation. The atomic labels used for the proton transfer are also given.	194
6.7	Relative energy (ΔE) profiles of the proton transfer between H_3O^+ and H_2O in the Eigen cation using MP2/aug-cc-pVTZ calculations as a reference and some semiempirical Hamiltonians.	195
6.8	Relaxation dynamics of the C10 structure.	197
6.9	Optimized geometries of the $(H_2O)_{21}$ water cluster at the PM3-MAIS level of theory. a: The zwitterionic minimum w21-A- Z^\pm . b: The corresponding “neutral” form w21-A-N. The highlighted water molecules are those involved in the proton transfer.	197
6.10	Neutralization mechanism from w21-A- Z^\pm to w21-A-N.	198
6.11	Representation of the w21-0K geometry of $(H_2O)_{21}$ showing to example of identified proton transfer between two neighboring water molecules. For each couple of water molecules, the donor molecule (molD) is circled in green and the acceptor (molA) is circled in purple. The numbers are the indexes of the water molecules.	201
6.12	Potential of mean force (PMF) of two proton transfers in w21-0K as a function of the reaction coordinate ζ . a: Transfer from molD(10) to molA(02). b: Transfer from molD(11) to molA(18). The bond order of the two oxygen atoms involved in the reaction as well as the ionic number of the cluster are also reported (see text).	202
6.13	Illustration of the procedure used to determine the end of the proton transfer using the identification of the inflection points ($p_{+/-}$ and $p_{-/+}$) of the PMF.	203

6.14	Hydrogen bond network analysis of the structure w21-0K for the proton transfer between molD(10) to molA(02). a: Averaged hydrogen bond connectivity matrix (AHBCM) at the beginning of the reaction. b: AHBCM at the end of the reaction. c: Reaction hydrogen bond connectivity matrix (RHBCM) obtained by the difference between the AHBCM's at the end and at the beginning of the reaction. d: Connectivity tree of RHBCM.	205
6.15	Same as Figure 6.15 for the proton transfer between molD(11) and molA(18). . . .	206
6.16	Definition of the groups of water dimer geometries. The groups from a to c were defined in Ref. [310].	208
6.17	Correlation between the proton transfer free energy and the two descriptors detailed in the text. a: ΔG versus R_{ODOA} . b: ΔG versus $P_{OO}(E)$. c: Three dimensional correlation between ΔG , R_{ODOA} and $P_{OO}(E)$. The coefficient of determination (R^2), the root mean square deviation (RMSD), the maximum unsigned error (UE) and the mean UE are displayed for each fit in the corresponding plot.	211
6.18	Representation of the "best" combination of R_{ODOA} and $P_{OO}(E)$ (see text) for each frame of the SEBOMD simulation. The probability of R_{ODOA} and $P_{OO}(E)$ are projected on the corresponding axis. The nine selected zones (from A to I) described in the text are also represented.	213



List of Tables

3.1	Average dipole moment computed from the wave function or using different charge models (see text) along PBC simulations with or without including the long range interactions. The values are given in Debye. The indexes n and t refer to an ensemble and time average, respectively.	85
4.1	PIF3 parameters for organic molecules interacting with water. The parameters reported with an asterisk were obtained in this work, whereas the others were obtained in Refs. [138, 204]. All parameters are in atomic units.	101
4.2	Mean unsigned errors (MUE, in kcal/mol) for the interaction between water and small organic molecules at the MP2 minimum, and between water and methane throughout the PES (see text). The maximum error is also shown in parentheses for the calculations based on the set of 33 complexes.	105
4.3	Mean unsigned errors (MUE, in kcal/mol) for the interaction between water and ethylene, benzene throughout the PES (see text).	107
5.1	Description of the 11 studied systems. Each cubic box is composed by one solute and 128 water molecules. For molecular mechanics simulations, the water molecules are treated using the SPCE/E force field. The box size of each system was determined after an ff03-SPC/E NPT simulation (see text). The origin of the charges used for the solute in the MM simulation is also provided.	117
5.2	First maximum and first minimum positions in Å (height in parenthesis) of the COw RDF of methane and the corresponding coordination number (N_w). Comparison of different Hamiltonians used in this work and some results from the literature.	121
5.3	Averages of the instantaneous total charge and dipole moment (in parenthesis, the difference with respect to gas phase) of the hydrophobic solutes during the PM3-PIF3 condensed phase simulation. The CM1 model was used to derive the charges.	129

5.4	Methane characteristic vibrational frequencies in gas phase (GP) and aqueous solution (Aq). Static SE and ff03 results are harmonic force constant computed on the minimized geometry using the related Hamiltonian. Two results are given for molecular dynamics simulations (MD). IR: IR spectrum from methane CM1 dipole moment time correlation function. VDOS: vibrational density of states. For the IR spectra, only ν_3 and ν_4 are active.	132
5.5	Averages of the instantaneous total charge and dipole moment (in parenthesis, the difference with respect to gas phase) of the hydrophilic solutes during the PM3-PIF3 condensed phase simulation. The CM1 model was used to derive the charges.	156
6.1	Comparison of the minimum structures found with different semiempirical methods with respect to MP2/aug-cc-pVTZ. $\Delta E_{int} = E_{int}^{SE} - E_{int}^{MP2}$ where $E_{int}^{MP2} = -5.22$ kcal/mol (interaction energy computed without applying the BSSE).	191
6.2	Relative potential energy (in kcal/mol) of the structure reported in Ref. [331] calculated at the MP2/6-311+G(2df,2d) and at the PM3-MAIS level.	196
6.3	Relative potential energy (in kcal/mol) of the three Z^\pm structures (<i>i.e.</i> , w21-A- Z^\pm , w21-B- Z^\pm , w21-C- Z^\pm) with respect to their corresponding neutral form. The scaling factor applied to PM3-MAIS is 1/1.216, as discussed in Subsection 6.2.1.	198
6.4	Averages obtained for each type of water couple configurations: proton transfer relative free energy $\overline{\Delta G}$, the $O_D O_A$ distance $\overline{R_{O_D O_A}}$ and the environment electric field projection $\overline{P_{00}(\mathbf{E})}$. Results are ordered by increasing free energy of transfer. The # symbol gives the number of configuration taken into account in the average. Standard deviation are given for all the averaged quantities.	210
6.5	Definition of the regions reported in Figure 6.18 and details about the selected structures (see text). The expected and calculated free energy of each proton transfer (ΔG^\ddagger and ΔG , receptively) are reported in kcal/mol. The range of ΔG^\ddagger is given with respect to the mean unsigned error reported in Figure 6.17c (<i>i.e.</i> , ± 4 kcal/mol).	214



Bibliography

- [1] Szabo, A. *Modern Quantum Chemistry*; MacMillan: New York, US, 1982.
- [2] Rivail, J.-L. *Éléments de chimie quantique à l'usage des chimistes*; Savoirs actuels Inter Édition/Éditions du CNRS: Paris, FR, 1994.
- [3] Cramer, C. J. In *Essentials of Computational Chemistry Theories and Models Second Edition*; John Wiley & Sons: Chichester, U., Ed.; 2004.
- [4] Leach, A. R. *Molecular modelling: principles and applications. Second edition.*; Pearson Education: UK, 2001.
- [5] Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press; Oxford, UK, 1989.
- [6] Schrödinger, E. *Phys. Rev.* **1926**, 28, 1049.
- [7] Born, M. *Zeitschrift für Phys.* **1926**, 37, 863.
- [8] Born, M. *Zeitschrift für Phys.* **1926**, 38, 803.
- [9] Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, 84, 457.
- [10] Pauli, W. *Zeitschrift für Phys.* **1925**, 31, 765.
- [11] Slater, J. C. *Phys. Rev.* **1929**, 34, 1293.
- [12] Roothaan, C. *Rev. Mod. Phys.* **1951**, 23, 69.
- [13] Hall, G. G. *Proc. R. Soc. Lond. A* **1951**, 205, 541.
- [14] Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, 46, 618.
- [15] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1988**, 153, 503.
- [16] Krishnan, R.; Pople, J. *Int. J. Quantum Chem.* **1978**, XIV, 91.
- [17] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, 136, 864.
- [18] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, 140, 1133.
- [19] Vosko, S.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, 58, 1200.

- [20] Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, 23, 5048.
- [21] Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, 45, 244.
- [22] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, 46, 6671.
- [23] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, 77, 3865.
- [24] Becke, A. D. *Phys. Rev. A* **1988**, 38, 3098.
- [25] Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785.
- [26] Becke, A. D. *J. Chem. Phys.* **1993**, 98, 5648.
- [27] Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, 120, 215.
- [28] Grimme, S. *J. Comput. Chem.* **2006**, 27, 1787.
- [29] Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, 5, 741.
- [30] Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, 32, 1456.
- [31] Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, 393, 51.
- [32] Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, 105, 4009.
- [33] Etienne, T.; Assfeld, X.; Monari, A. *J. Chem. Theory Comput.* **2014**, 10, 3896.
- [34] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, 4, 187.
- [35] Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, 110, 1657.
- [36] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. *Eur. Biophys. J.* **2011**, 40, 843.
- [37] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, 117, 5179.
- [38] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, 24, 1999.
- [39] Lee, M. C.; Duan, Y. *Proteins* **2004**, 55, 620.
- [40] Császár, A. G. *WIREs Comput. Mol. Sci.* **2012**, 2, 273.

-
- [41] Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.
- [42] Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960.
- [43] Jorgensen, W. L.; Tirado-Rives, J. *J. Phys. Chem.* **1996**, *100*, 14508.
- [44] Verlet, L. *Phys. Rev.* **1967**, *159*, 98.
- [45] Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.
- [46] Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136.
- [47] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- [48] Toukmaji, A. Y.; Broad Jr., J. A. *Comput. Physics Commun.* **1996**, *95*, 73.
- [49] York, D.; Yang, W. *J. Chem. Phys.* **1994**, *101*, 3298.
- [50] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
- [51] Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384.
- [52] Nosé, S. *Mol. Phys.* **1984**, *52*, 255.
- [53] Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.
- [54] Monard, G.; Merz Jr., K. M. *Acc. Chem. Res.* **1999**, *32*, 904.
- [55] Monard, G.; Prat-Resina, X.; González-Lafont, A.; Lluch, J. M. *Int. J. Quantum Chem.* **2003**, *93*, 229.
- [56] Genheden, S.; Ryde, U. *J. Comput. Chem.* **2010**, *31*, 837.
- [57] Genheden, S.; Ryde, U. *Phys. Chem. Chem. Phys.* **2012**, *14*, 8662.
- [58] Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.
- [59] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Senden, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011.
- [60] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Senden, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.
- [61] Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.
- [62] Laio, A.; Gervasio, F. L. *Reports Prog. Phys.* **2008**, *71*, 126601.
- [63] Huber, T.; Torda, A. E.; van Gunsteren, W. F. *J. Comput. Aided. Mol. Des.* **1994**, *8*, 695.
-

- [64] Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.
- [65] Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. a.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961.
- [66] Soper, A. K. *ISNR Phys. Chem.* **2013**, *2013*, 1.
- [67] Levine, B. G.; Stone, J. E.; Kohlmeyer, A. *J. Comput. Phys.* **2011**, *230*, 3556.
- [68] Svishchev, I. M.; Zassetsky, A. Y.; Kusalik, P. G. *Chem. Phys.* **2000**, *258*, 181.
- [69] Allesch, M.; Schwegler, E.; Galli, G. *J. Phys. Chem. B* **2007**, *111*, 1081.
- [70] Berens, P.; Wilson, K. *J. Chem. Phys.* **1981**, *74*, 4872.
- [71] Gordon, R. G. *J. Chem. Phys.* **1965**, *43*, 1307.
- [72] Schmidt, J. R.; Corcelli, S. A. *J. Chem. Phys.* **2008**, *128*, 184504.
- [73] Press, W. H.; Teukolosky, S. A.; Vetterling, W. R.; Flannery, B. P. *Numerical Recipies in Fortran*; Cambridge University Press; Cambridge, UK, 1992.
- [74] Ingrosso, F.; Monard, G.; Hamdi Farag, M.; Bastida, A.; Ruiz-López, M. F. *J. Chem. Theory Comput.* **2011**, *7*, 1840.
- [75] Egorov, S. a.; Everitt, K. F.; Skinner, J. L. *J. Phys. Chem. A* **1999**, *103*, 9494.
- [76] Berens, P. H.; White, S. R.; Wilson, K. R. *J. Chem. Phys.* **1981**, *75*, 515.
- [77] Bader, J. S.; Berne, B. J. *J. Chem. Phys.* **1994**, *100*, 8359.
- [78] Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
- [79] Iuchi, S.; Morita, A.; Kato, S. *J. Phys. Chem. B* **2002**, *106*, 3466.
- [80] Gageot, M.; Vuilleumier, R.; Sprik, M.; Borgis, D. *J. Chem. Theory Comput.* **2005**, *1*, 772.
- [81] Thiel, W. *Tetrahedron* **1988**, *44*, 7393.
- [82] Zerner, M. C. *Rev. Comput. Chem.*; John Wiley & Sons: Chichester, UK, 1991; Vol. 2; pp 313–365.
- [83] Clark, T. *J. Mol. Struct. THEOCHEM* **2000**, *530*, 1.
- [84] Voityuk, A. A. *WIREs Comput. Mol. Sci.* **2013**, *3*, 515.
- [85] Thiel, W. *WIREs Comput. Mol. Sci.* **2014**, *4*, 145.
- [86] Hückel, E. *Zeitschrift für Phys.* **1931**, *70*, 204.

-
- [87] Hoffmann, R. *J. Chem. Phys.* **1963**, 39, 1397.
- [88] Pariser, R.; Parr, R. G. *J. Chem. Phys.* **1953**, 21, 466.
- [89] Pople, J. A. *Trans. Faraday Soc.* **1953**, 49, 1375.
- [90] Parr, R. G. *J. Chem. Phys.* **1952**, 20, 1499.
- [91] Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, 43, S129.
- [92] Pople, J. A.; Beveridge, D. L.; Dobosh, P. A. *J. Chem. Phys.* **1967**, 47, 2026.
- [93] Slater, J. C.; Koster, G. F. *Phys. Rev.* **1954**, 94, 1498.
- [94] Slater, J. C.; Johnson, K. H. *Phys. Rev. B* **1972**, 5, 844.
- [95] Johnson, K. H.; Smith Jr., F. C. *Phys. Rev. B* **1972**, 5, 831.
- [96] Johnson, K. H. *Adv. Quantum Chem.* **1973**, 7, 143.
- [97] Warshel, A.; Weiss, R. M. *J. Am. Chem. Soc.* **1980**, 102, 6218.
- [98] Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1966**, 44, 3289.
- [99] Del Bene, J.; Jaffé, H. H. *J. Chem. Phys.* **1968**, 48, 1807.
- [100] Kuehnlenz, G.; Jaffé, H. H. *J. Chem. Phys.* **1973**, 58, 2238.
- [101] Noor Mohammad, S.; Hopfinger, A. J. *Int. J. Quantum Chem.* **1982**, 22, 1189.
- [102] Baird, N. C.; Dewar, M. J. S. *J. Chem. Phys.* **1969**, 50, 1262.
- [103] Dewar, M. J. S.; Haselbach, E. *J. Am. Chem. Soc.* **1969**, 92, 590.
- [104] Bingham, R. C.; Dewar, M. J. S.; Lo, D. H. *J. Am. Chem. Soc.* **1975**, 97, 1285.
- [105] Ridley, J.; Zerner, M. *Theor. Chim. Acta* **1973**, 32, 111.
- [106] Bacon, A. D.; Zerner, M. C. *Theor. Chim. Acta* **1979**, 53, 21.
- [107] Kotzian, M.; Rösch, N.; Zerner, M. C. *Theor. Chim. Acta* **1992**, 81, 201.
- [108] Nanda, D. N.; Jug, K. *Theor. Chim. Acta* **1980**, 57, 95.
- [109] Jug, K.; Iffert, R.; Schulz, J. *Int. J. Quantum Chem.* **1987**, 32, 265.
- [110] Zhanpeisov, N. U.; Pel'menshchikov, A. G.; Zhidomirov, G. M. *J. Struct. Chem.* **1987**, 28, 1.
- [111] Bliznyuk, A. A.; Voityuk, A. A. *J. Mol. Struct. THEOCHEM* **1988**, 164, 343.
-

- [112] Lipiński, J. *Int. J. Quantum Chem.* **1988**, 34, 423.
- [113] Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, 14, 639.
- [114] Dewar, M.; Thiel, W. *J. Am. Chem. Soc.* **1977**, 99, 4899.
- [115] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, 107, 3902.
- [116] Dewar, M. J. S.; Jie, C.; Yu, J. *Tetrahedron* **1993**, 49, 5003.
- [117] Holder, A. J.; Dennington II, R. D.; Jie, C. *Tetrahedron* **1994**, 50, 627.
- [118] Voityuk, A. A.; Rösch, N. *J. Phys. Chem. A* **2000**, 104, 4089.
- [119] Winget, P.; Horn, A. H. C.; Selçuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, 9, 408.
- [120] Winget, P.; Clark, T. *J. Mol. Model.* **2005**, 11, 439.
- [121] Kayi, H.; Clark, T. *J. Mol. Model.* **2007**, 13, 965.
- [122] Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, 15, 1253.
- [123] Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, 15, 295.
- [124] Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, 16, 1109.
- [125] Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, 16, 29.
- [126] Kayi, H.; Clark, T. *J. Mol. Model.* **2011**, 17, 2585.
- [127] Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, 27, 1101.
- [128] Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, 3, 486.
- [129] Burstein, K. Y.; Isaev, A. N. *Theor. Chim. Acta* **1984**, 64, 397.
- [130] Suck Salk, S. H.; Chen, T. S.; Hagen, D. E.; Lutrus, C. K. *Theor. Chim. Acta* **1986**, 70, 3.
- [131] Stewart, J. J. P. *J. Mol. Model.* **2007**, 13, 1173.
- [132] Bernal-Uruchurtu, M. I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-López, M. F. *J. Comput. Chem.* **2000**, 21, 572.
- [133] Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, 23, 1601.
- [134] Tubert-Brohman, I.; Guimarães, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, 25, 138.
- [135] Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, 1, 817.

- [136] McNamara, J. P.; Muslim, A.-M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, 394, 429.
- [137] Bernal-Uruchurtu, M. I.; Ruiz-López, M. F. *Chem. Phys. Lett.* **2000**, 330, 118.
- [138] Harb, W.; Bernal-Uruchurtu, M. I.; Ruiz-López, M. F. *Theor. Chem. Acc.* **2004**, 112, 204.
- [139] Arillo-Flores, O. I.; Ruiz-López, M. F.; Bernal-Uruchurtu, M. I. *Theor. Chem. Acc.* **2007**, 118, 425.
- [140] Marion, A.; Monard, G.; Ruiz-López, M. F.; Ingrosso, F. J. *Chem. Phys.* **2014**, 141, 034106.
- [141] Stewart, J. J. P. *J. Mol. Model.* **2013**, 19, 1.
- [142] Ignatov, S. K.; Razuvaev, A. G.; Kokorev, V. N.; Alexandrov, Y. A. *J. Phys. Chem.* **1996**, 100, 6354.
- [143] Hehre, W. J.; Yu, J.; Adei, E. *Abstr. Pap. ACS* **1996**, 212, COMP 092.
- [144] <http://www.wavefun.com>.
- [145] Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, 81, 391.
- [146] Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, 93, 315.
- [147] Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, 100, 616.
- [148] Ahlswede, B.; Jug, K. *J. Comput. Chem.* **1999**, 20, 563.
- [149] Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, 14, 775.
- [150] Weber, W. *Ph.D. thesis, Univ. Zürich, Switz.* **1996**,
- [151] Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, 103, 495.
- [152] Scholten, M. *Ph.D. thesis, Univ. Düsseldorf, Düsseldorf, Ger.* **2003**,
- [153] Wu, X.; Thiel, W.; Pezeshki, S.; Lin, H. *J. Chem. Theory Comput.* **2013**, 9, 2672.
- [154] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, 51, 947.
- [155] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, 58, 7260.
- [156] Kumar, A.; Elstner, M.; Suhai, S. *Int. J. Quantum Chem.* **2003**, 95, 44.
- [157] Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, 7, 931.

- [158] Kaminski, S.; Giese, T. J.; Gaus, M.; York, D. M.; Elstner, M. *J. Phys. Chem. A* **2012**, *116*, 9131.
- [159] Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618.
- [160] Chang, D. T.; Schenter, G. K.; Garrett, B. C. *J. Chem. Phys.* **2008**, *128*, 164111.
- [161] Martin, B.; Clark, T. *Int. J. Quantum Chem.* **2006**, *106*, 1208.
- [162] McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
- [163] Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.
- [164] Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.
- [165] Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2011**, *506*, 286.
- [166] Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808.
- [167] Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141.
- [168] Csonka, G. I.; Ángyán, J. G. *J. Mol. Struct. THEOCHEM* **1997**, *393*, 31.
- [169] Monard, G.; Bernal-Uruchurtu, M. I.; van der Vaart, A.; Merz Jr., K. M.; Ruiz-López, M. F. *J. Phys. Chem. A* **2005**, *109*, 3425.
- [170] Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.
- [171] Seifert, G.; Joswig, J.-O. *WIREs Comput. Mol. Sci.* **2012**, *2*, 456.
- [172] Gaus, M.; Cui, Q.; Elstner, M. *WIREs Comput. Mol. Sci.* **2014**, *4*, 49.
- [173] Oleari, L.; De Michelis, G.; Di Sipio, L. *Mol. Phys.* **1966**, *10*, 97.
- [174] Dewar, M. J. S.; Thiel, W. *Theor. Chim. Acta* **1977**, *46*, 89.
- [175] Stewart, J. J. *J. Comput. Aided. Mol. Des.* **1990**, *4*, 1.
- [176] Ludwig, O.; Schinke, H.; Brandt, W. *J. Mol. Model.* **1996**, *2*, 341.
- [177] Riley, K. E.; Pitoňák, M.; Jurecka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023.
- [178] Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Prenosil, O.; Konvalinka, J.; Hobza, P. *J. Phys. Chem. B* **2010**, *114*, 12666.
- [179] Stewart, J. J. P. *J. Mol. Model.* **2009**, *15*, 765.
- [180] Hostaš, J.; Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2013**, *568-569*, 161.
- [181] Thiriot, E.; Monard, G. *J. Mol. Struct. THEOCHEM* **2009**, *898*, 31.

- [182] Bernal-Uruchurtu, M. I.; Ruiz-López, M. F. *Beyond standard quantum chemistry: Applications from gas to condensed phases*; Transworld Research Network, Kerala, India, 2007; Chapter Eigen and Zundel ions in aqueous environments. A theoretical study using semi-empirical force fields, pp 65–86.
- [183] Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- [184] Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; T.E. Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *AMBER14*; University of California, San Francisco, 2014.
- [185] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198.
- [186] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *725*, 712.
- [187] Dickson, C. J.; Madej, B. D.; Skjevik, A. a.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. *J. Chem. Theory Comput.* **2014**, *10*, 865.
- [188] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, M. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian09*; Gaussian, Inc., Wallingford CT, 2009.
- [189] <http://www.ambermd.org>.
- [190] Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, J. J.; Brothers, E. N.; Westermhoff, L. M.; Merz Jr., K. M. *DivCon99*; The Pennsylvania State University, 1999.
- [191] Tasaki, K.; McDonald, S.; Brady, J. W. *J. Comput. Chem.* **1993**, *14*, 278.
- [192] Sprik, M. *Chem. Phys.* **2000**, *258*, 139.

- [193] Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2.
- [194] VandeVondele, J.; Borštnik, U.; Hutter, J. *J. Chem. Theory Comput.* **2012**, *8*, 3565.
- [195] Maia, J. D. C.; Urquiza Carvalho, G. A.; Manguiera, C. P.; Santana, S. R.; Cabral, L. A. F.; Rocha, G. B. *J. Chem. Theory Comput.* **2012**, *8*, 3072.
- [196] Wu, X.; Koslowski, A.; Thiel, W. *J. Chem. Theory Comput.* **2012**, *8*, 2272.
- [197] Anisimov, V. M.; Bliznyuk, A. A. *J. Phys. Chem. B* **2012**, *116*, 6261.
- [198] Hennemann, M.; Clark, T. *J. Mol. Model.* **2014**, *20*, 2331.
- [199] Murdachaew, G.; Mundy, C. J.; Schenter, G. K.; Laino, T.; Hutter, J. *J. Phys. Chem. A* **2011**, *115*, 6046.
- [200] Ventura, O. N.; Coitiño, E. L.; Lledós, A.; Berteán, J. *J. Mol. ...* **1989**, *187*, 55.
- [201] Csonka, G. *J. Comput. Chem.* **1993**, *14*, 895.
- [202] Csonka, G.; Éliás, K.; Csizmadia, I. G. *J. Comput. Chem.* **1997**, *18*, 330.
- [203] Salvatella, L.; Mokrane, A.; Cartier, A.; Ruiz-López, M. F. *Chem. Phys. Lett.* **1998**, *296*, 239.
- [204] Harb, W. *Ph.D. thesis, Univ. Henri Poincaré, Nancy, Fr.* **2003**,
- [205] Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119.
- [206] Clementi, E.; Corongiu, G. *J. Phys. Chem. A* **2001**, *105*, 10379.
- [207] Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- [208] Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.
- [209] Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.
- [210] Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
- [211] Stewart, J. J. P. *MOPAC2012*; Stewart Computational Chemistry, Colorado Springs, CO, 2012.
- [212] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- [213] Levenberg, K. *Quart Appl. Math.* **1944**, *2*, 164.
- [214] Marquardt, D. *SIAM J. Appl. Math.* **1963**, *11*, 431.
- [215] Struzhkin, V. V.; Militzer, B.; Mao, W. L.; Mao, H.-K.; Hemley, R. J. *Chem. Rev.* **2007**, *107*, 4133.

- [216] Koh, C. A. *Chem. Soc. Rev.* **2002**, 31, 157.
- [217] Dartois, E.; Deboffle, D. *Astron. Astrophys.* **2008**, 22, 19.
- [218] Tennyson, J.; Bernath, P. F.; Brown, L. R.; Campargue, A.; Császr, A. G.; Daumont, L.; Gamache, R. R.; Hodges, J. T.; Naumenko, O. V.; Polyansky, O. L.; Rothman, L. S.; Vandaele, A. C.; Zobov, N. F. *Pure Appl. Chem.* **2014**, 86, 71.
- [219] Max, J.-J.; Chapados, C. *J. Chem. Phys.* **2009**, 131, 184505.
- [220] Khan, A. *J. Chem. Phys.* **1999**, 110, 11884.
- [221] Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, 100, 5829.
- [222] Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, 102, 3782.
- [223] Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, 72, 650.
- [224] Luque, F. J.; Reuter, N.; Cartier, A.; Ruiz-López, M. F. *J. Phys. Chem. A* **2000**, 104, 10923.
- [225] Feig, M. *J. Chem. Theory Comput.* **2008**, 4, 1555.
- [226] Seabra, G. D. M.; Walker, R. C.; Roitberg, A. E. *J. Phys. Chem. A* **2009**, 113, 11938.
- [227] Gaigeot, M.-p. *Phys. Chem. Chem. Phys.* **2010**, 12, 3336.
- [228] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, 98, 10089.
- [229] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, 103, 8577.
- [230] Crowley, M. F.; Darden, T. A.; Cheatham III, T. E.; Deerfield II, D. W. *J. Supercomput.* **1997**, 11, 255.
- [231] Sagui, C.; Darden, T. A. In *Simulation and Theory of Electrostatic Interactions in Solution*, american institute of physics ed.; Pratt, L. R., Hummer, G., Eds.; 1999; pp 104–113.
- [232] Toukmaji, A.; Sagui, C.; Board, J.; Darden, T. *J. Chem. Phys.* **2000**, 113, 10913.
- [233] Sagui, C.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2004**, 120, 73.
- [234] Ball, P. *Chem. Rev.* **2008**, 108, 74.
- [235] Mateus, M. P.; Galamba, N.; Costa Cabral, B. J.; Coutinho, K.; Canuto, S. *Chem. Phys. Lett.* **2011**, 506, 183.
- [236] Rossato, L.; Rossetto, F.; Silvestrelli, P. L. *J. Phys. Chem. B* **2012**, 116, 4552.
- [237] Ashbaugh, H. S.; Liu, L.; Surampudi, L. N. *J. Chem. Phys.* **2011**, 135, 054510.

- [238] Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1951**, *19*, 774.
- [239] Nathaniel Pribble, R.; Zwier, T. S. *Faraday Discuss.* **1994**, *97*, 229.
- [240] Laaksonen, A.; Stilbs, P.; Wasylishen, R. E. *J. Chem. Phys.* **1998**, *108*, 455.
- [241] Tarakeshwar, P.; Choi, H. S.; Lee, S. J.; Lee, J. Y.; Kim, K. S.; Ha, T.-K.; Jang, J. H.; Lee, J. G.; Lee, H. *J. Chem. Phys.* **1999**, *111*, 5838.
- [242] Dang, L.; Feller, D. *J. Phys. Chem. B* **2000**, *104*, 4403.
- [243] Allesch, M.; Lightstone, F. C.; Schwegler, E.; Galli, G. *J. Chem. Phys.* **2008**, *128*, 014501.
- [244] Mateus, M. P. S.; Galamba, N.; Costa Cabral, B. J. *J. Chem. Phys.* **2012**, *136*, 014507.
- [245] Junqueira, G. M. A.; Dos Santos, H. F. *J. Mol. Model.* **2014**, *20*, 2152.
- [246] Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.
- [247] Löwdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365.
- [248] Chandra Singh, U.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129.
- [249] Besler, B. H.; Merz Jr., K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- [250] Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735.
- [251] Stuart, B. *INFRARED SPECTROSCOPY: Fundamentals and Applications*; John Wiley & Sons: Chichester, UK, 2004.
- [252] Lee, T. J.; Martin, J. M. L.; Taylor, P. R. *J. Chem. Phys.* **1995**, *102*, 254.
- [253] Gray, D. L.; Robiette, A. G. *Mol. Phys.* **1979**, *37*, 1901.
- [254] Greathouse, J. a.; Cygan, R. T.; Simmons, B. a. *J. Phys. Chem. B* **2006**, *110*, 6428.
- [255] Schachtschneider, J. H.; Snyder, R. G. *Spectrochim. Acta* **1963**, *19*, 117.
- [256] Varsányi, G. *Vibrational spectra of benzene derivatives*; Academic Press: New York, 1969.
- [257] Fileti, E. E.; Chaudhuri, P.; Canuto, S. *Chem. Phys. Lett.* **2004**, *400*, 494.
- [258] Fidler, J.; Rodger, P. M. *J. Phys. Chem. B* **1999**, *103*, 7695.
- [259] van Erp, T. S.; Meijer, E. J. *J. Chem. Phys.* **2003**, *118*, 8831.
- [260] Noskov, S. Y.; Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2005**, *109*, 6705.
- [261] Larsen, N. W.; Nicolaisen, F. M. *J. Mol. Struct.* **1974**, *22*, 29.

- [262] Berden, G.; Leo Meerts, W.; Schmitt, M.; Kleinermanns, K. *J. Chem. Phys.* **1996**, *104*, 972.
- [263] Mmereki, B. T.; Donaldson, D. J. *J. Phys. Chem. A* **2002**, *106*, 3185.
- [264] Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827.
- [265] Hesske, H.; Gloe, K. *J. Phys. Chem. A* **2007**, *111*, 9848.
- [266] Feigel, M.; Strassner, T. *J. Mol. Struct. THEOCHEM* **1993**, *283*, 33.
- [267] Chalmet, S.; Ruiz-López, M. F. *J. Chem. Phys.* **1999**, *111*, 1117.
- [268] Benoit, D. M.; Clary, D. C. *J. Phys. Chem. A* **2000**, *104*, 5590.
- [269] Max, J.-J.; Daneault, S.; Chapdos, C. *Can. J. Chem.* **2002**, *80*, 113.
- [270] Yamashita, T.; Takatsuka, K. *J. Chem. Phys.* **2007**, *126*, 074304.
- [271] Petković, M. *J. Phys. Chem. A* **2012**, *116*, 364.
- [272] Ataka, S.; Takeuchi, H.; Tasumi, M. *J. Mol. Struct.* **1984**, *113*, 147.
- [273] Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001**, *105*, 10922.
- [274] Lucas, B.; Lecomte, F.; Reimann, B.; Barth, H.-D.; Grégoire, G.; Bouteiller, Y.; Schermann, J.-P.; Desfrancois, C. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2600.
- [275] Ramachandran, G.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95.
- [276] Gaigeot, M.-P. *J. Phys. Chem. B* **2009**, *113*, 10059.
- [277] Adzhubei, A. a.; Sternberg, M. J. E.; Makarov, A. a. *J. Mol. Biol.* **2013**, *425*, 2100.
- [278] Drozdov, A. N.; Grossfield, A.; Pappu, R. V. *J. Am. Chem. Soc.* **2004**, *126*, 2574.
- [279] Kwac, K.; Lee, K.-K.; Han, J. B.; Oh, K.-I.; Cho, M. *J. Chem. Phys.* **2008**, *128*, 105106.
- [280] Parchaňský, V.; Kapitán, J.; Kaminský, J.; Šebestík, J.; Bour, P. *J. Phys. Chem. Lett.* **2013**, *4*, 2763.
- [281] Vymetal, J.; Vondrásek, J. *J. Phys. Chem. B* **2010**, *114*, 5632.
- [282] Chin, W.; Piuzzi, E.; Dognon, J.-P.; Dimicoli, I.; Mons, M. *J. Chem. Phys.* **2005**, *123*, 084301.
- [283] Chin, W.; Piuzzi, E.; Dimicoli, I.; Mons, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1033.
- [284] Gloaguen, E.; Pagliarulo, F.; Brenner, V.; Chin, W.; Piuzzi, E.; Tardivel, B.; Mons, M. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4491.

- [285] Biswal, H. S.; Loquais, Y.; Tardivel, B.; Gloaguen, E.; Mons, M. *J. Am. Chem. Soc.* **2011**, *133*, 3931.
- [286] Gageot, M.-P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10198.
- [287] Eigen, M.; De Maeyer, L. *Zeitschrift für Elektrochemie, Berichte der Bunsengesellschaft für Phys. Chemie* **1955**, *59*, 986.
- [288] Sørensen, S. P. L. *Biochem. Z.* **1909**, *21*, 131.
- [289] Baucke, F. G. K.; Brett, C. M. A.; Milton, M. J. T.; Mussini, T.; Naumann, R.; Pratt, K. W.; Spitzer, P.; Wilson, G. S. *Pure Appl. Chem.* **2002**, *74*, 2169.
- [290] Crans, D. C.; Levinger, N. E. *Acc. Chem. Res.* **2012**, *45*, 1637.
- [291] Levinger, N. E.; Swafford, L. A. *Annu. Rev. Phys. Chem.* **2009**, *60*, 385.
- [292] Riter, R. E.; Willard, D. M.; Levinger, N. E. *J. Phys. Chem. B* **1998**, *102*, 2705.
- [293] Faeder, J.; Ladanyi, B. M. *J. Phys. Chem. B* **2000**, *104*, 1033.
- [294] Faeder, J.; Ladanyi, B. M. *J. Phys. Chem. B* **2001**, *105*, 11148.
- [295] Faeder, J.; Albert, M. V.; Ladanyi, B. M. *Langmuir* **2003**, *19*, 2514.
- [296] Harpham, M. R.; Ladanyi, B. M.; Levinger, N. E.; Herwig, K. W. *J. Chem. Phys.* **2004**, *121*, 7855.
- [297] Baruah, B.; Roden, J. M.; Sedgwick, M.; Mariano Correa, N.; Crans, D. C.; Levinger, N. E. *J. Am. Chem. Soc.* **2006**, *128*, 12758.
- [298] Piletic, I. R.; Moilanen, D. E.; Spry, D. B.; Levinger, N. E.; Fayer, M. D. *J. Phys. Chem. A* **2006**, *110*, 4985.
- [299] Millot, C.; Soetens, J.-c.; Martins Costa, M. T. C.; Hodges, M. P.; Stone, A. J. *J. Phys. Chem. A* **1998**, *30*, 754.
- [300] Guillot, B.; Guissani, Y. *J. Chem. Phys.* **2001**, *114*, 6720.
- [301] van Duijneveldt-van de Rijdt, J. G. C. M.; Mooij, W. T. M.; van Duijneveldt, F. B. *Phys. Chem. Chem. Phys.* **2003**, *5*, 1169.
- [302] Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 2993.
- [303] Kloppe, W.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt, F. B. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *2*, 2227.
- [304] Lane, J. R. *J. Chem. Theory Comput.* **2013**, *9*, 316.

-
- [305] Pedulla, J. M.; Kim, K.; Jordan, K. D. *Chem. Phys. Lett.* **1998**, 291, 78.
- [306] Xantheas, S. S. *Chem. Phys.* **2000**, 258, 225.
- [307] Keutsch, F. N.; Saykally, R. J. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, 98, 10533.
- [308] Keutsch, F. N.; Cruzan, J. D.; Saykally, R. J. *Chem. Rev.* **2003**, 103, 2533.
- [309] Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. *J. Phys. Chem. A* **2008**, 112, 3976.
- [310] Kuo, J.-L.; Ciobanu, C. V.; Ojamäe, L.; Shavitt, I.; Singer, S. J. *J. Chem. Phys.* **2003**, 118, 3583.
- [311] Lenz, A.; Ojamäe, L. *Phys. Chem. Chem. Phys.* **2005**, 7, 1905.
- [312] Lenz, A.; Ojamäe, L. *J. Phys. Chem. A* **2006**, 110, 13388.
- [313] Xantheas, S. S. *Can. J. Chem. Eng.* **2012**, 90, 843.
- [314] Bernal, J. D.; Fowler, R. H. *J. Chem. Phys.* **1933**, 1, 515.
- [315] Bernal-Uruchurtu, M. I.; Ortega-Blake, I. *J. Phys. Chem. A* **1999**, 103, 884–892.
- [316] Trout, B. L.; Parrinello, M. *Chem. Phys. Lett.* **1998**, 288, 343.
- [317] Geissler, P. L.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, 291, 2121.
- [318] Natzle, W. C.; Moore, C. B. *J. Phys. Chem.* **1985**, 89, 2605.
- [319] Chang, H.-C.; Wu, C.-C.; Kuo, J.-L. *Int. Rev. Phys. Chem.* **2005**, 24, 553.
- [320] McDonald, S.; Ojama, L.; Singer, S. J. *J. Phys. Chem. A* **1998**, 102, 2824.
- [321] Vuilleumier, R.; Borgis, D. *Chem. Phys. Lett.* **1998**, 284, 71.
- [322] Hodges, M. P.; Wales, D. J. *Chem. Phys. Lett.* **2000**, 324, 279.
- [323] Ku, T.; Lotrich, V. F.; Perera, A.; Bartlett, R. J. *J. Chem. Phys.* **2009**, 131, 104313.
- [324] Wróblewski, T.; Karwasz, G. P. *Eur. Phys. J. Spec. Top.* **2013**, 222, 2217.
- [325] Parkkinen, P.; Riikonen, S.; Halonen, L. *J. Phys. Chem. A* **2012**, 116, 10826.
- [326] Eigen, M. *Angew. Chemie Int. Ed. English* **1964**, 3, 1.
- [327] Zundel, G.; Metzger, H. *Zeitschrift für Phys. Chemie* **1968**, 58, 225.
- [328] Anick, D. J. *J. Mol. Struct. THEOCHEM* **2001**, 574, 109.
-

- [329] Mrázek, J.; Burda, J. V. *J. Chem. Phys.* **2006**, *125*, 194518.
- [330] Karthikeyan, S.; Singh, N. J.; Kim, K. S. *J. Phys. Chem. A* **2008**, *112*, 6527–32.
- [331] Torrent-Sucarrat, M.; Ruiz-López, M. F.; Martins-Costa, M.; Francisco, J. S.; Anglada, J. M. *Chem. Eur. J.* **2011**, *17*, 5076.
- [332] Ojamäe, L.; Shavitt, I.; Singer, S. J. *J. Chem. Phys.* **1998**, *109*, 5547.
- [333] Wang, S.; MacKay, L.; Lamoureux, G. *J. Chem. Theory Comput.* **2014**, *10*, 2881.
- [334] Smith, B. J.; Swanton, D. J.; Pople, J. A.; Schaefer III, H. F.; Radom, L. *J. Chem. Phys.* **1990**, *92*, 1240.
- [335] Amararene, A.; Gindre, M.; Le Huérou, J.-Y.; Nicot, C.; Urbach, W.; Waks, M. *J. Phys. Chem. B* **1997**, *101*, 10751.

Résumé :

Ce travail est destiné au développement de méthodes approchées de chimie quantique capables de traiter des systèmes biologiques de grande taille. En particulier, nous réalisons des simulations de dynamique moléculaire dans l'approximation de Born-Oppenheimer, permettant une description quantique de l'Hamiltonien électronique du système dans son entier. Nous utilisons un code développé dans notre groupe et disponible au sein de la suite de programmes AmberTools14 : SEBOMD (*SemiEmpirical Born-Oppenheimer Molecular Dynamics*).

Notre approche se base sur un Hamiltonien électronique semiempirique (SE). Nous avons remarqué que l'une des principales difficultés rencontrées lors d'une simulation SEBOMD de la phase condensée est représentée par le choix de la méthode SE. Nous avons montré que la plupart des méthodes courantes ne permet pas une description convenable de certaines interactions fondamentales. Poursuivant les développements méthodologiques initiés dans notre groupe, nous avons élaboré et paramétrisé une nouvelle correction pour les Hamiltoniens SE qui prend en compte de manière explicite l'interaction de certains groupes fonctionnels importants avec leur environnement. Cette méthode, dénommée PM3-PIF3, a été appliquée à l'étude par dynamique moléculaire de molécules organiques dans l'eau. Les résultats que nous avons obtenus montrent que notre méthode est appropriée pour le traitement de molécules comportant des groupements hydrophobes et/ou hydrophiles en milieu aqueux.

Nous avons travaillé sur un choix *ad hoc* de molécules organiques, représentatives de systèmes biologiques modèles, ainsi que sur un modèle de peptide, le dipeptide alanine. Nous avons analysé la structure de l'eau autour de ces composés ainsi que leurs propriétés électroniques et vibrationnelles en présence du solvant. Nos résultats montrent un très bon accord avec les études expérimentales et théoriques présentes dans la littérature.

Finalement, nous nous sommes intéressés au processus d'autoprotolyse de l'eau en milieux confinés. Après avoir discuté du choix de l'Hamiltonien SE à utiliser pour cette étude, nous avons caractérisé le transfert de proton dans l'agrégat d'eau (H₂O)₂₁. Nous avons établi une corrélation entre l'énergie libre associée à la première étape de ce transfert et certaines propriétés physiques collectives.

Mots-clés :

dynamique moléculaire Born-Oppenheimer; méthodes quantiques semiempiriques; simulations en phase condensée; spectroscopie infrarouge; transfert de proton; agrégats d'eau.

Abstract:

The present work is devoted to the development of approximate quantum chemistry methods that are suitable to treat biological systems of large size. In particular, we run molecular dynamics under the Born-Oppenheimer approximation, allowing a quantum mechanical description of the electronic Hamiltonian of the full system. We use a code developed in our group and available in the AmberTools14 suite of programs: SEBOMD (*SemiEmpirical Born-Oppenheimer Molecular Dynamics*).

Our method is based on a semiempirical (SE) electronic Hamiltonian. We pointed out that one of the key issues arising in a condensed phase SEBOMD simulation is represented by the choice of the SE method. We showed that most of the currently available approaches fail in describing some relevant intermolecular interactions. Following a methodology formulated in our group, we developed and parameterized a new correction of SE Hamiltonians, which explicitly accounts for the interactions between some important functional groups and their environment. This method, which we named PM3-PIF3, was applied to study the molecular dynamics of organic molecules in water. The results that we obtained showed that our technique is suitable to treat molecules having hydrophobic and/or hydrophilic groups in an aqueous medium.

We worked on an *ad hoc* choice of organic molecules, representative of biological model systems as well as on a model of polypeptide, the alanine dipeptide. We analyzed the structure of water around these compounds as well as their electronic and vibrational properties in the presence of the solvent. Our results show a very good agreement with experimental and theoretical studies in the literature.

Finally, we investigated the water self-dissociation process in confined environments. After discussing the choice of the SE Hamiltonian to be used for this purpose, we characterized the proton transfer in the (H₂O)₂₁ water cluster. We established a correlation between the free energy of the first step of this process and some collective physical properties.

Keywords:

Born-Oppenheimer molecular dynamics; quantum semiempirical methods; condensed phase simulations; infrared spectroscopy; proton transfer; water clusters.

Publication associée/Related publication:

A. Marion, G. Monard, M. F. Ruiz-López and F. Ingrosso, *J. Chem. Phys.*, **2014**, 141, 034106.