Hydration of drug-like molecules with molecular density functional theory and the hybrid-4th-dimension Monte Carlo approach

Sohvi Luukkonen

université
PARIS-SACLAY

Hydration of drug-like molecules with molecular density functional theory and the hybrid-4th-dimension Monte Carlo approach

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 571, Sciences Chimiques: Molécules, Matériaux, Instrumentation et Biosystemes (2MIB)
Spécialité de doctorat: Chimie
Unité de recherche: Université Paris-Saclay, UVSQ, Inria, CNRS, CEA, Maison de la Simulation, 91191, Gif-sur-Yvette, France
Référent: : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Maison de la Simulation, CEA Saclay, le 24 Novembre 2020, par

Sohvi LUUKKONEN

Composition du jury:

Tâp HA-DUONG Professeur, Université Paris-Saclay	Président du jury
Francesca INGROSSO Maîtresse de Conférence, HDR, Université de Lorraine	Rapporteur & Examinatrice
Agilio PADUA Professeur, École Normale Supérieur Lyon	Rapporteur & Examinateur
Hélène BERTHOUMIEUX Chargée de Recherche, CNRS, Sorbonne Université	Examinatrice
Riccardo SPEZIA Directeur de Recherche, CNRS, Sorbonne Université	Emaninateur
Daniel BORGIS Directeur de Recherche, CNRS	Directeur
Maximilien LEVESQUE Chargé de Recherche, CNRS	Co-encadrant
Luc BELLONI Directeur de Recherche, CEA Saclay	Invité
Guillaume JEANMAIRET Chargé de Recherche, CNRS	Invité

Thèse de doctorat

NNT: 2020UPASF030

DISSERTATION BY

SOHVI LUUKKONEN

HYDRATION OF DRUG-LIKE MOLECULES WITH MOLECULAR DENSITY FUNCTIONAL THEORY AND THE HYBRID-4TH-DIMENSION MONTE CARLO APPROACH

MAISON DE LA SIMULATION, CEA SACLAY, UNIVERSITÉ PARIS-SACLAY

UNDER THE SUPERVISION OF

DANIEL BORGIS & MAXIMILIEN LEVESQUE

"The least important things,
sometimes, my dear boy,
lead to the greatest discoveries."

— The First Doctor

ACKNOWLEDGMENTS

First of all, I would like to thank Edouard Audit, director of Maison de la Simulation at CEA Saclay, and Rodolph Vuilleumier, director of PASTEUR laboratory at ENS, for welcoming me into your laboratories. I would like to acknowledge all the organizations and the staff at both institutions support they gave me during these three years.

I wish to acknowledge the members of the jury: Francesca Ingrosso, Agilio Padua, Hélène Berthoumieux, Ricardo Spezia and Tâp Ha-Duong, for their willingness to evaluate and taking interest in my work.

Most of all, I would like to express my most respectful gratitude to my thesis supervisors Maximilien Levesque and Daniel Borgis. I would like to thank you for the opportunity that you gave me and for all the help and advise I got from you in these three years. I would also like to give a special thank you to Luc Belloni with whom I collaborated all along this thesis. I have learned so much from you three and it was a great pleasure to work with you.

I would like to warmly thank everybody at the Maison de la Simulation and the theoretical chemistry pole at the ENS for scientific and non-scientific lunchtime discussion. Especially my fellow younglings: Hugo, Anton, Benoît, Sascha, and Elsa(s); and Marie-Laure Bocquet. Hugo, it was a pleasure the share my office with you during these three years and Anton, my fellow Zerg, it was a great pleasure to work with you. I appreciate these discussions and working with you even more after not seeing you for these past few months.

I would also like to acknowledge some people without whom I would not be here or at least I would not have had such a great time in the past few years:

My high school chemistry and physics teachers, Kaija Kauppi, Titta Linderborg and Timo Taskinen, who guided me to find my passion for science.

Cheers to Anni and Kasimir, who got me through my first few years in Paris and made it a great time. You made the Saturday library sessions and the three-day study camps before the exams bearable, and I fondly remember our evenings at the 'Café Sophie'.

Cheers also to all my Monday night Pub Quiz team members from the last six years, from the Polarization team to the Polarized Squirrels with Jad and Andrew, and to our quiz master Kahina. It has been a weekly highlight to play with you!

And cheers, to all physics friends who adopted me "just a mere" chemist as a friend and with whom we made through these three years together. Especially to Jad, Flo and most of all Elena, I'm not sure I would have survived this without our afternoon coffee breaks. We did it!

Last but to least, I would thank my family and Maxime.

Thanks 'äiti' and 'isi' for everything and always supporting me in my studies here in Paris

Thanks Maxime for the support and encouragement you have given me and for everything else. And a huge extra nod to you for supporting my face 24/7 for the last few months and for the courage of correcting this thesis.

RÉSUMÉ EN FRANÇAIS

Le développement d'un médicament prend en moyenne plus de 10 ans et coûte 1 milliard de dollars. Pour accélérer le processus, et diminuer le coût, on utilise des méthodes in silico lors de l'étape de la découverte du médicament. Cela consiste à faire du criblage et de l'optimisation de ligands à partir de bases de données de $\sim$100 000 molécules de type médicament pour proposer quelques candidats à l'étape préclinique. Le critère majeur de la sélection est l'affinité entre le potentiel médicament et la cible biologique.

L'interaction se passant dans notre corps, cette affinité, i.e. l'énergie libre de liaison, doit être prédite dans l'eau. De plus, le médicament doit être soluble dans l'eau et parfois être capable de traverser la membrane cellulaire modélisée par le coefficient de partage eau/n-octanol pour avoir accès à la cible. Globalement, les propriétés de solvatation jouent un rôle important dans la conception de médicaments avec une grandeur importante au cœur des processus : l'énergie libre de solvatation et particulièrement l'énergie libre d'hydratation (ELH).

Numériquement, la solvatation peut être étudiée soit par (i) des simulations exactes mais coûteuses avec plusieurs centaines d'heures CPU par ELH, soit par (ii) des modèles de continuum rapides mais qui ne tiennent pas compte de la nature moléculaire du solvant et donc manque de précision, soit par (iii) des théories des liquides approximées qui gardent l'information moléculaire du solvant pour une diminution en temps de calcul. L'objectif de cette thèse est de proposer un outil numérique précis mais rapide pour la prédiction des énergies libres d'hydratation de molécules d'intérêt pharmaceutique.

La théorie de la fonctionnelle de densité moléculaire (MDFT) est une approche de théorie de liquide qui permet l'étude des propriétés thermodynamiques d'équilibre de n'importe quelle soluté rigide. L'avantage de cette théorie et du code de haute performance associé est de prédire les énergies libres de solvatation et la structure d'équilibre moléculaire de solvatation de petites molécules type médicament en quelques minutes CPU. Dans son état actuel, la théorie est une approximation au niveau "hyper netted-chain" (HNC) et a deux désavantages : (i) elle traite des molécules rigides donc des solutés à conformère unique figés dans l'espace et (ii) l'approximation HNC introduit une forte surestimation de la pression du système qui en conséquence conduit à la surestimation de l'énergie libre de la formation de la cavité et doit être compensée par une correction de pression (PC) a posteriori.

Pour bien développer et évaluer la performance de MDFT, on a besoin des données de référence à conformère unique et d'évaluer l'importance de la flexibilité du soluté quand on prédit des EHL de molécules type médicament. Pour cela, on s'est tourné vers l'approche Monte-Carlo hybride à 4e dimension et son code maison développée par Luc Belloni car la majorité des codes de simulation bien développés sont écrits pour des molécules flexibles.

H4D-MC est une méthode originale de simulations exactes pour calculer les ELH à partir du principe de Jarzynski avec simulations courtes hors-équilibre pendant lesquelles on introduit le soluté dans le solvant ou on le retire du solvant avec un paramètre de couplage dans une 4ème dimension qui dépend du temps. En combinant les données des insertions et des destructions avec le théorème de Crooks et après une analyse paramétrique, on a pu montrer que H4D-MC permet de calculer les ELH des molécules de type médicament cinq fois plus rapidement que l'approche classique de perturbation de l'énergie libre pour une précision statistique de moins de 0.1 kcal/mol.

Nous avons utilisé l'approche H4D-MC pour calculer les ELH à conformère unique des 642 petites molécules type médicament de la base de données FreeSolv. En comparant les résultats H4D-MC aux ELH obtenues par des calculs MD+FEP etat de l'art avec un soluté flexible fourni

par la base de donnée, on trouve que les deux méthodes donnent le même ELH pour la majorité des molécules mais des déviations importantes sont présentes pour certaines solutés.

Pour vérifier que ces déviations viennent du manque de flexibilité dans les calculs H4D-MC et pas de problèmes dans le code, nous avons implémenté la flexibilité dans le code H4D-MC. On a fait cela à deux niveaux : (i) on peut propager les conformères de soluté en même temps que les configurations du solvant pendant le MC et (ii) les conformères du soluté ont la possibilité de se relaxer pendant l'insertion et la destruction. En recalculant les ELH de la base de données FreeSolv avec des solutés flexibles, nous retrouvons les ELH flexible de la base de données FreeSolv avec un gain de temps de fois quatre par rapport à l'approche MD+FEP.

En faisant une analyse chemoinformatique et multiconformationnelle, nous avons pu identifier un sous-ensemble de FreeSolv, nommée FreeSolv-rigide, de 214 molécules (un tiers de la base de données originale) pour lesquelles la flexibilité du soluté n'affecte pas du tout leur ELH et 80% des déviations dues à la flexibilité sont plus petites que la précision expérimentale de 0.6 kcal/mol. Par ailleurs, nous avons pu identifier la caractéristique primaire pour prédire si la flexibilité joue un rôle important : la capacité de former des liaisons hydrogènes avec l'eau et surtout la possibilité de former des liaisons hydrogènes intramoléculaires. Nous avons aussi pu montrer que, pour la majorité de ces petites molécules type médicament, on réussit à récupérer l'ELH flexible avec quelques calculs à conformère unique.

Donc nous avons utilisé l'approche H4D-MC pour calculer des données de références à conformère unique et nous avons montré que pour une grande partie des molécules type médicament de la base de données FreeSolv, la flexibilité du soluté n'a pas un rôle important pour prédire leurs ELH. Nous pouvons donc appliquer l'approche rapide mais approximée sur ces molécules.

Nous montrons que les ELH brutes prédites par MDFT-HNC sont très éloignées des valeurs de référence pour la base de données FreeSolv comme attendu à cause de la surestimation de la pression du système. Si nous appliquons une première correction de pression simple, proposée par le groupe en 2014, on prédit les ELH de référence avec une erreur moyenne de 2 kcal/mol. Cette correction dépend de la pression HNC du système ($\sim$10 000 atm à la place de l'atm attendue expérimentalement) et du volume du soluté définie originalement comme le volume molaire partielle (PMV) du soluté (un output direct de MDFT).

Nous visons une précision d'un demi kcal/mol ($\sim$1 kT $\approx$ précision expérimentale) pour MDFT. Pour cela nous avons proposé trois améliorations de la correction de précision : (i) en définissant le volume soluté comme une union de volumes atomiques de van der Walls (vdW) optimisés, (ii) en ajoutant une correction machine learning avec des réseaux de neurones entraînés à minimiser la différence entre les résultats MDFT-HNC et H4D-MC, ou (iii) en ajoutant un terme de surface inspiré de "scale particle theory" et un terme de correction de chargement empirique à la correction de pression originale. Tous les trois développements mènent à une précision de 0.5 kcal/mol. La dernière est la correction la plus rigoureuse et est applicable à n'importe quel type de soluté, mais la correction électrostatique dépend de la variation du volume du soluté pendant le chargement et donc nécessite deux minimisations MDFT. Donc, pour appliquer MDFT dans le cadre des applications pharmaceutiques, où la vitesse est clef, nous conseillons d'utiliser la correction pression vdW pour éviter la deuxième minimisation.

Pour évaluer rigoureusement la performance de MDFT-HNC à prédire des ELH et des structures de solvatation, nous avons fait un benchmark sur une variété de systèmes : solutés sphériques neutres ou chargés et des solutés moléculaire de la base de données FreeSolv et la molécule d'eau comme soluté elle-même. Pour les ELH, nous trouvons que MDFT-HNC couplé à la correction de pression et de surface prédit celles des systèmes neutres avec une précision de moins d'un demi kcal/mol et celles des ions avec une précision de moins de 5 kcal/mol ($\sim$5% d'erreur relative).

Pour les structures de solvatation moléculaires, nous montrons que MDFT-HNC les prédit généralement plutôt bien mais a quelques défauts systématiques : elle surestime les densités autour des sites neutres ou peu chargés et sous-estime les densités autours des sites anioniques. Elle ne

réussit pas non plus à bien prédire la structure tétraédrique des liaisons hydrogènes autour de la molécule d'eau.

Le but final est de réussir à prédire avec précision mais efficacement des ELH expérimentales de molécules d'intérêt pharmaceutique. Pour cela, nous avons comparé la capacité de MDFT-HNC avec la correction de pression vdW et des simulations MD+FEP à prédire les ELH expérimentales de solutés rigides de la base de données FreeSolv. Nous avons trouvé que les deux approches prédisent les ELH expérimentales avec une erreur moyenne de 1 kcal/mol mais MDFT permet un gain de temps de 3 à 4 ordres de magnitude par rapport aux calculs de référence de MD+FEP. Avec une analyse chemoinformatique, nous avons pu montrer qu'une grande partie de l'erreur de MDFT-HNC vient de la paramétrisation du champ de force et pas de la théorie approximée en elle-même.

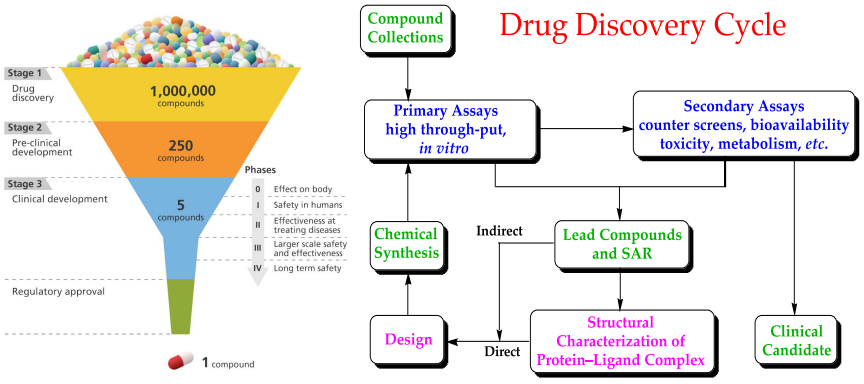
A drug molecule is a small (few hundreds of Daltons) organic molecule that will bind to biological target (a protein such as enzymes, receptors or ion channels) related to the disease/condition, and modify, increase (agonist receptor) or block (antagonist receptor), the activity of the biological target.

The development of a new drug, from drug discovery, after the therapeutic target is selected, to commercialisation, through (pre-)clinical trials and approval, takes on average over 10 years and the median cost is $985 million, counting expenditures on failed trials [1] (fig. 0.1a). Starting from a database of millions of already existing molecules, *eg.* the ZINC database of ~ 750 million purchasable compounds [2], the first stage is to do *in vitro* high-throughput screening (HTP) to select lead compounds which can be optimised, with *in silico* methods (fig. 0.1b). The main criterion of selection is the affinity, *i.e.* the binding free energy, between the solute target's active site $\Delta G_{\text{bind}}^{\text{ligand}-\text{target}}$. This first stage can take up to three years and is the least expensive stage of the drug development process. Its objective is to propose few hundred potential drug leads to the *in vivo* pre-clinical trials via efficient screening with hopefully not missing good candidates (false negatives) and not letting through false positives. Currently, only 0.02 and 10 % of starting molecules in pre-clinical and clinical trials lead to an approved drug, respectively. Earlier the 'bad' ligands are rejected, more time and money is saved.

As the drug-receptor interaction happens in our body, this affinity needs to be computed in to be predicted in water where the binding free energy is defined as a combination of binding free energy in vacuum and the hydration free energies of the ligand, the target and the complex:

$$\Delta G_{\text{bind},H_2O}^{\text{ligand}-\text{target}} = \Delta G_{\text{bind.vac}}^{\text{ligand}-\text{target}} + \Delta G_{\text{hydr}}^{\text{ligand}-\text{target}} - \left(\Delta G_{\text{hydr}}^{\text{ligand}} + \Delta G_{\text{hydr}}^{\text{target}}\right).$$

Other criteria include target selectivity, the ligands solubility as it needs to disolve into the blood ($\log S = -(\Delta G_{\text{hydr}} + \Delta G_{\text{sub}})/2.303RT$ where ΔG_{sub} is the sublimation free energy), and potentially the capacity to petrate the cell membrane to access the receptor modelled by the water/n-octanol partition coefficient ($\log P_{ow} = (\Delta G_{\text{octanol}} - \Delta G_{\text{hydr}})/2.303RT$ where $\Delta G_{\text{octanol}}$ is the octanol solvation free energy). Solvation and especially hydration free energies are omnipresent in these quantities. Thus it is important to have accurate but fast computational tools to predict solvation free energies.



(a) (b)

Figure 0.1: Scheme of (a) the drug design process and (b) the drug discovery and pre-clinical phases.

The most rigorous way to do the *in silico* optimisation would be done with explicit solvent simulations, but as a computation of a single HFE takes hundreds of cpu.h, it is impossible to do virtual screening on hundreds of thousands of molecules with this approach. Therefore, the pharmaceutical industry and academics use docking and scoring functions to evaluate the affinity [3, 4, 5]. Docking is a method for predicting the preferred orientation and position of the ligand, *i.e.* potential drug, in the targets active site and scoring functions are fast approximate numerical methods used to estimate the intensity of affinity between the ligand and the target when docked. They calculate a score, representing the binding free energy, between the two molecules, in few seconds from a set of descriptors characterising the complex with a numerical function. It has been shown that these approaches work sometimes very well and sometimes not at all [6]. Hence they are not very predictive and can lead to multiple false positives or negatives. In 2016, considering the difficulty but nevertheless necessity of evaluating precisely SFEs in the drug design process, important actors of the pharmaceutical industry publicly called the academic world for alternatives, pointing out the lack of accuracy or speed of current methods [7].

Beyond predicting the SFE, and its derivatives, of potential drug molecules, studying solvation of the biological target site can be very fruitful in the structure-based drug design processes. As subtle structural variations of the ligand can have profound impact on the binding with the target's active site, mapping the location and orientation, and the binding affinity of water molecules in the target site can offer ample information into the properties of the active site. They can describe the hydrophobic forces and potential hydrogen bonds driving the binding of the potential ligand.

CONTENTS

1 INTRODUCTION 1
- 1.1 Solvation and hydration 1
  - 1.1.1 Energetic informations 2
  - 1.1.2 Structural informations 3
- 1.2 Modelling solvation effects 4
- 1.3 Other applications 5
  - 1.3.1 Synthesis/production 5
  - 1.3.2 Aquatic toxicity 6
- 1.4 Scope of this thesis 6

I STATE-OF-THE-ART: SOLVATION MODELS AND METHODS

2 CONTINUUM MODELS 8
- 2.1 Non-polar term 8
- 2.2 Charging term 10
- 2.3 Limitations 11

3 SOLVATION FREE ENERGIES WITH SIMULATIONS 12
- 3.1 Free energy perturbation 12
  - 3.1.1 Widom test particle 13
  - 3.1.2 Bennett acceptance ratio 15
- 3.2 Alchemical intermediates 15
  - 3.2.1 Stratification: discretised coupling parameter 16
  - 3.2.2 Slow growth: time-dependent coupling parameter 17
- 3.3 Alternative methods and improving sampling 19

4 LIQUID STATE THEORIES 21
- 4.1 Molecular integral equations and RISM 22
- 4.2 Molecular density functional theory 24
  - 4.2.1 MDFT functional 24
  - 4.2.2 MDFT algorithm and code 25
  - 4.2.3 Free energy corrections 27
  - 4.2.4 Solvation structure 28
- 4.3 Correspondence between MIET and MDFT 29

II HYDRATION WITH H4D-MC

5 RIGID SOLUTES 32
- 5.1 Analysis of insertion/destruction parameters 32
  - 5.1.1 Computation time 37
- 5.2 Spherical solutes 38
  - 5.2.1 Hydrophobic solutes 38
  - 5.2.2 Monovalent ions 39
  - 5.2.3 $\{q, \sigma, \epsilon\}-$spheres 40
- 5.3 Molecular solutes - FreeSolv 41

6 SOLUTE FLEXIBILITY IN H4D-MC 43
- 6.1 Solute flexibility in H4D-MC 43
- 6.2 Simulation parameters related to flexibility 44

6.3 Comparison of single rigid conformers vs. fully flexible solute 47

7 FLEXIBILITY IN THE FREESOLV DATABASE 49
7.1 Single conformer vs. flexible solute 49
7.2 Multiple conformers analysis 51
7.3 Focus on flexible solutes 53
7.3.1 Effect of mass, number of bonds and rings 53
7.3.2 Effect of H-bond donors and acceptors 53
7.3.3 Effect of functional groups 54

III HYDRATION WITH MDFT-HNC

8 PRESSURE CORRECTION 60
8.1 Original pressure correction 61
8.1.1 Compressible fluids 62
8.2 Optimized van der Waals volume 62
8.3 Machine learning fitted correction 64
8.3.1 Neural network 64
8.3.2 Input data and hidden layers 65
8.3.3 Cross-validation 66
8.3.4 ML corrected MDFT results 67
8.4 Surface term 68
8.4.1 Hydrophobic spheres 68
8.4.2 Molecular solutes 70
8.5 Recapitulation 71

9 MDFT-HNC BENCHMARK 74
9.1 Spherical solutes 74
9.1.1 Hydrophobic solutes 74
9.1.2 Monovalent ions 76
9.2 Molecular solutes 78
9.2.1 Water as solute 78
9.2.2 FreeSolv database 81

10 MDFT FOR DRUG–LIKE MOLECULES 84
10.1 Effect of solute's mass, charges and solvation structure 85
10.2 Effect of functional groups 88

IV CONCLUSION AND PERSPECTIVES

11 CONCLUSION 94

12 PERSPECTIVES 96
12.1 For H4D-MC 96
12.2 For MDFT 96
12.2.1 Going beyond the HNC approximations 97
12.2.2 Reducing the memory footprint 97
12.2.3 Solute flexibility / MM-MDFT 98
12.2.4 Coupling MDFT 99
12.2.5 Effect of temperature 99
12.2.6 Partition coefficients 100

V APPENDIX

A GRADIENTS OF THE MDFT FUNCTIONAL 102

B STATISTICAL MEASURES 104
C FREESOLV DATABASE 106
D H4D-MC SIMULATION DETAILS 107
E 'FREESOLV-RIGID' 109
F NEURAL NETWORK DETAILS 110
G PROBLEMATIC SOLUTES FOR MDFT 111
H MD+FEP AND RISM ERROR ANALYSIS 113
I MDFT-HNC ERROR BARS 114

BIBLIOGRAPHY 121

PUBLICATIONS

The ensemble of published or accepted articles on the subject of this thesis at the time of the redaction :

S. Luukkonen, M. Levesque, L. Belloni, and D. Borgis. Hydration free energies and solvation structures with molecular density functional theory in the hypernetted chain approximation. In: J. Chem. Phys. 152 (2020), p. 064110. doi: 10.1063/1.5142651

A. Robert, **S. Luukkonen**, and M. Levesque. Pressure correction for solvation theories. In: J. Chem. Phys. 152.19 (2020), p. 191103. doi: 10.1063/5.0002029

S. Luukkonen, L. Belloni, D. Borgis, and M. Levesque. Predicting hydration free energies of the FreeSolv database of druglike molecules with molecular density functional theory. In: J. Chem. Info. Model 60 (2020), p. 3558. doi: 10.1021/acs.jcim.0c00526

Two other artciles are in preparation at the time of the redaction.

Additionally, the author of thesis co-authored the following paper not discussed in this thesis :

D. Borgis, **S. Luukkonen**, L. Belloni, and G. Jeanmairet. Simple parameterfree bridge functionals for molecular density functional theory. Application to hydrophobic solvation. In: J. Chem. Phys. B 124 (2020), p. 6885. doi: 10.1021/acs.jpcb.0c04496

NOTATIONS

$\Delta G_{solv}, \mu_{exc}$	solvation free energy $\equiv$ excess chemical potential [$\mathrm{kcal \cdot mol^{-1}}$]
$\rho(\boldsymbol{r}, \omega)$	molecular solvent density [$\mathrm{Å^{-3}}$]
ρ_{bulk}	molecular solvent bulk density [$\mathrm{Å^{-3}}$]
$\boldsymbol{r}$	cartesian position vector
ω	Euler angles: $\omega \equiv (\theta, \phi, \psi)$
n_{max}	order of quadrature of angle discretisation
$n(\boldsymbol{r})$	number solvent density [$\mathrm{Å^{-3}}$]: $n(\boldsymbol{r}) \equiv \int \rho(\boldsymbol{r}, \omega) \mathrm{d}\omega$
n_{bulk}	number solvent bulk density [$\mathrm{Å^{-3}}$]: $n_{bulk}^{H_2O} = 0.033$ $\mathrm{Å^{-3}} \simeq 1$ kg/l at 300 K and 1 atm
T	temperature [K]
k_B	Boltzmann constant: $k_B = 1.3806488 \times 10^{-23}$ [$\mathrm{J \cdot K^{-1}}$]
β	reciprocal of thermal energy: $\beta = (k_B T)^{-1}$ [$\mathrm{J^{-1}}$]
σ, ϵ	Lennard-Jones force field parameters [Å,$\mathrm{kJ \cdot mol^{-1}}$]

e	electron charge: $e = 1.60217662 \times 10^{-19}$ C
q	point charge [e]
ε_0	vaccuum permittivity: $\varepsilon_0 = 8.854187817 \times 10^{-12}$ [$\mathrm{C}^2 \cdot \mathrm{J}^{-1} \cdot \mathrm{m}^{-1}$]
ε_r	dielectric constant (relative permittivity) of the solvent
P	pressure [atm]
γ_S	surface tension [$\mathrm{J} \cdot \mathrm{m}^{-2}$]
V_{PM}	partial molar volume [Å^3]
V_{vdW}	solute's van der Waals volume [Å^3]
F	Helmholtz free energy [$\mathrm{kcal} \cdot \mathrm{mol}^{-1}$]
N	number of solvent molecules: $N = \int n(\boldsymbol{r}) \mathrm{d}\boldsymbol{r}$
$\boldsymbol{q}$	phase space coordinates (spatial coordinates and momenta)
$\lambda, \lambda(t)$	(time-dependent) alchemical coupling parameter
W	work [$\mathrm{kcal} \cdot \mathrm{mol}^{-1}$]
$p_{\mathrm{ins}}, p_{\mathrm{des}}$	insertion and destruction distributions in H4D-MC
w_{max}	maximum altitude in the 4$^{\mathrm{th}}$ dimension in H4D-MC [Å]
M	solvent's molar mass: $M = 18$ g/mol for water
v	insertion/destruction speed in H4D-MC [$\sqrt{k_{\mathrm{B}}T \cdot M^{-1}}$]
Δt	time-step in H4D-MC [$\sqrt{\beta \cdot M}$Å]
$g(\boldsymbol{r}, \omega)$	pair correlation function: $g = \rho / \rho_{\mathrm{bulk}}$
h	total correlation function: $h = g - 1$
c	direct correlation function
b	bridge function
$g(r)$	radial distribution function
$\mathcal{F}[\rho]$	MDFT functional
U_{ext}	solute-solvent interaction potential
γ	indirect solute-solvent correlation function: $\gamma = c * \Delta\rho$
$\Delta\rho$	solvent excess density: $\Delta\rho = \rho - \rho_{\mathrm{bulk}}$

ACRONYMS

BAR	Bennett acceptance ration
FEP	free energy perturbation
GB	generalized Born (equation)
H4D-MC	hydrid 4$^{\mathrm{th}}$ dimension Monte Carlo
HFE	hydration free energy
HNC	hypernetted-chain (approximation)
KH	Kovalenko-Hirata (approximation)

LJ	Lennard-Jones
LST	liquid state theory
MAE	mean absolute error
MC	Monte Carlo
MD	molecular dynamics
MDFT	molecular density functional theory
ME	mean (signed) error
MIET	molecular integral equations theory
ML	machine learning
MM	molecular mechanics
MOZ	molecular Ornstein-Zernike (equation)
NN	neural network
PB	Poisson-Boltzmann (equation)
PC	pressure correction
PCM	polarizable continuum model
PMV	partial molar volume
QM	quantum mechanics
RISM	reference interaction site model
RMSE	root-mean-squared error
SAS	surface accesisble area
SFE	solvation free energy
SPT	scale particle theory
vdW	van der Waals

1 INTRODUCTION

Water covers 71% of the earth's surface and makes $\sim 65\%$ of the human body weight. It is the essence of life and surrounds us everywhere and thus making it the environment of most chemical reaction. Therefore, predicting if a substance, an ion, a molecule or a complex, likes water or not, *i.e.* is it hydrophilic or -phobic, is of first importance in *in silico* physical chemistry. Furthermore, the hydration of substance can greatly affect its structure and activity. Thus making important of studying compounds in their natural environment, which is commonly water.

1.1 SOLVATION AND HYDRATION

In chemistry, a solution is a homogeneous mixture of two or more components defined as

> *'A liquid or solid phase containing more than one substance, when for convenience one (or more) substance, which is called the solvent, is treated differently from the other substances, which are called solutes. When, as is often but not necessarily the case, the sum of the mole fractions of solutes is small compared with unity, the solution is called a dilute solution. A superscript attached to the ∞ symbol for a property of a solution denotes the property in the limit of infinite dilution.'*
>
> - *IUPAC* [8]

and solvation is a fundamental phenomenon in chemistry defined as

> *'Any stabilizing interaction of a solute (or solute moiety) and the solvent or a similar interaction of solvent with groups of an insoluble material (i.e. the ionic groups of an ion-exchange resin). Such interactions generally involve electrostatic forces and van der Waals forces, as well as chemically more specific effects such as hydrogen bond formation.'*
>
> - *IUPAC* [8]

This thesis is solely focused on liquid solutions and even more specifically on hydration, *i.e.* the solvation process with water as the solvent. Theoretically, the solvation process can be defined as bringing/submerging a solute, whether it is a solid, a liquid or a gas compound, from a fixed position in an ideal gas to fixed position in a solution [9]. However, in practice, the solvation process happens experimentally by the transfer of a molecule from a gas phase to the solution or by the dissolution of a solid solute. Whereas, the most common theoretical pathway is by the creation of an artificial solute cavity in the solution where the solute is inserted (fig. 1.1). In both cases, the solution (and solute) structure can relax leading to a chemical/physical equilibrium.

Once solvated the compound (molecule, complex, salt pair etc.), the stability, the structure or the activity of the solute compound can be strongly altered by the influence of the solvent molecules, *eg.* salt molecules can dissolve to single ions, proteins can change conformations drastically or proton exchanges can occur with a protonic solvent. The prediction of these solvation effects has been for a long time a goal of many physical-chemists and there are two main aspects that one can consider when studying solvation: energetics and structural profiles.

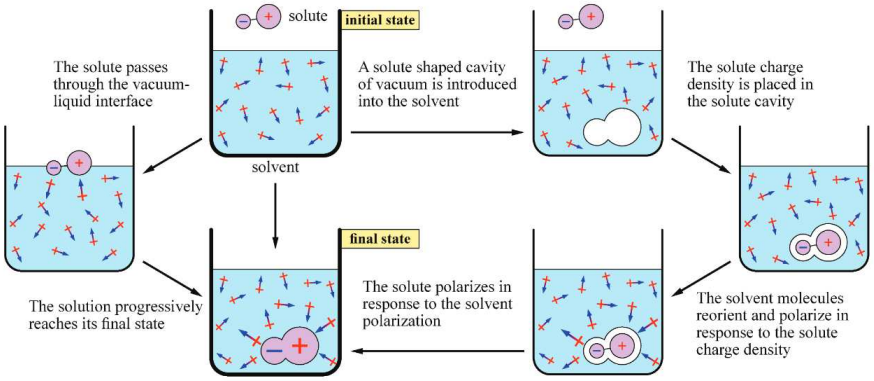


Figure 1.1: Thermodynamic cycles of the solvation process via (left) the experimental path and (right) theoretical path decomposed into an alchemical (artificial non-physical) process.

1.1.1 ENERGETIC INFORMATIONS

At the heart of all energetic information connected to solvation lies the solvation free energy (SFE) ΔG_{solv}, *i.e.* the excess chemical potential μ_{exc} of the solute molecule. A *free energy* is defined as the chemical potential that measures the reversible work during a transformation of a thermodynamic quantity (*eg.* volume, pressure, number of particles) or a change of the interaction potential along a physical or a non-physical reaction coordinate. For the physical-chemists, the most commonly used free energy is the Gibbs free energy, *i.e.* the work in a thermodynamic system at constant temperature T and pressure P.

The solvation free energy is the necessary work to bring a solute from vacuum to the solvent, *i.e.* work due to the change of the number of particles, which is equivalent to 'turning on' the interaction between the solute and the solvent molecules or to the excess chemical potential supplied to the system by the solute (fig. 1.2). It can be considered as the Gibbs free energy difference between the final state, solvated system $(N+m)$, and the initial state, bulk solvent (N) and solute (m) in the vacuum:

$$\Delta G_{\text{solv}} = G_{N+m} - (G_N + G_m). \tag{1.1}$$

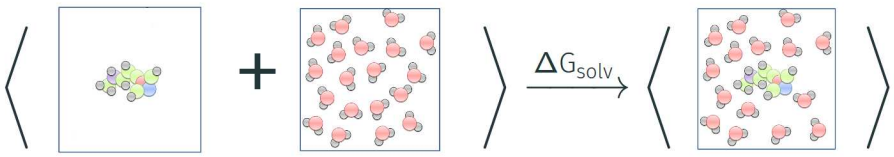


Figure 1.2: Illustration of the solvation free energy.

The solvation free energy can be decomposed into an enthalpic and an entropic term,

$$\Delta G = \Delta H - T\Delta S \tag{1.2}$$

where ΔH is the solvation enthalpy resulting from the broken solute-solute and solvent-solvent and formed solute-solvent bonds; and ΔS the solvation entropy resulting from the increase of disorder due to the dispersion of the solute in the solvent and the increase of order due to the structuration of the solvent around the solute.

Experimentally, the solvation free energy is defined as the transfer free energy of 1 mol of the solute from the gas phase to a solution at a concentration of 1 mol/l. These solvation free energies, and enthalpies, can be measured by isothermal titration (micro)calorimetry or gas phase

chromatography. The computation of solvation free energies, as for any free energy, is non-trivial as it requires the sampling of all possible states that can be visited during the transformation. Computational methods for predicting solvation free energies are discussed in detail in part i.

Nevertheless, the ability to predict solvation free energies, possibly combined with gas-phase calculations, unlocks the access to a multitude of other important physical quantities such as partition ($\log P$) coefficients [10], activities (γ), solubilities ($\log S$) [11], binding free energies ($\Delta G_{\text{bind}}^{\text{AB,solv}}$) [12, 13] or potentials of mean force (PMF) defined as

$$\log P_{\alpha\beta} = (\Delta G_{\text{solv}}^{\alpha} - \Delta G_{\text{solv}}^{\beta})/2.303RT \tag{1.3}$$

$$\log \gamma = \Delta G_{\text{solv}}/2.303RT \tag{1.4}$$

$$\log S = -(\Delta G_{\text{solv}} + \Delta G_{\text{sub}})/2.303RT \tag{1.5}$$

$$\Delta G_{\text{bind}}^{\text{AB,solv}} = \Delta G_{\text{bind}}^{\text{AB,vac}} + \Delta G_{\text{solv}}^{\text{AB}} - (\Delta G_{\text{solv}}^{\text{A}} + \Delta G_{\text{solv}}^{\text{B}}) \tag{1.6}$$

$$\text{PMF}(r) = \Delta G_{\text{solv}}^{\text{AB}}(r) - \Delta G_{\text{solv}}^{\text{AB}}(\infty) + U^{\text{AB}} \tag{1.7}$$

where α and β are two non-miscible solvents, ΔG_{sub} the sublimation free energy of a substance, $\Delta G_{\text{bind}}^{\text{AB},x}$ the binding free energy between molecules A and B in solution or vacuum; $\Delta G_{\text{solv}}^{\text{AB}}(r)$ the solvation free energy of the AB pair separated by r and U^{AB} the direct interaction between the pair, and R the gas constant.

1.1.2 STRUCTURAL INFORMATIONS

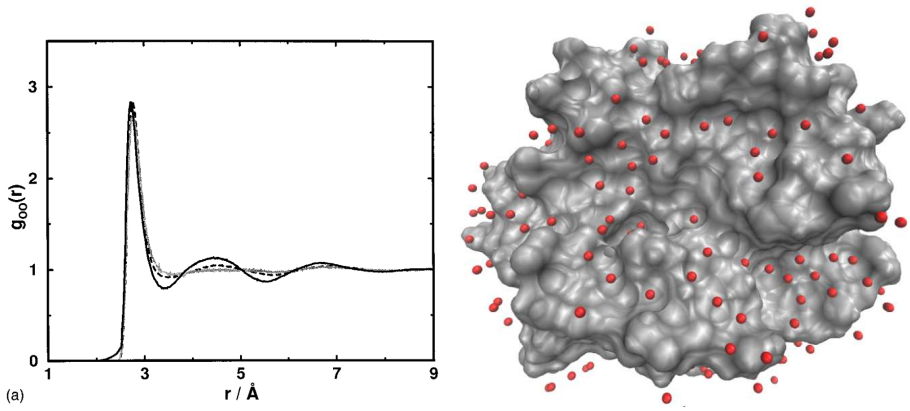
The most exhaustive information that one can have on the structure is to the knowledge of all molecular positions at all instants. However, as the positions vary non-stop it renders the information on the instantaneous positions very difficult to read and impossible to measure experimentally. Therefore, one needs to introduce quantities that measure the average structure of the solution. The most complete average quantity is the molecular equilibrium solvation structure $\rho_{\text{eq}}(\boldsymbol{r}, \omega)$ that is the average static microscopic structure of the solvent around a solute which is a function of the solvent position ($\boldsymbol{r}$) and orientation (ω). From $\rho_{\text{eq}}(\boldsymbol{r}, \omega)$ one can deduce a multitude of information. One of the most commonly used features in the liquid state physics is the radial site-site distribution function $g(r) = \rho(r)/\rho_{\text{bulk}}$ (fig. 1.3a).

Another piece of information that can be extracted from the solvation structure can be for example the identification of (i) hydrophilic or -phobic regions of a solute, important information for protein folding or self-assembly; (ii) highly bounded water molecules in a protein's active site (fig. 1.3b), an important piece of information when selecting/optimizing a potential drug molecule in structure-based drug design; or (iii) polarization fields around charged solutes.

Experimentally this fully molecular (spatial and orientational) structure of the liquids is impossible to measure. However, one can measure the structure factor S with X-ray or neutron diffraction experiments [14]. Note that, contrary to solids, liquids are isotropic. Hence, the structure factor only depends on the norm of the diffraction vector $k = \|\boldsymbol{k}\|$ and not on its orientation. The structure factor can be linked to the fully molecular equilibrium structure *via* the radial distribution function

$$S(k) = 1 + \rho \iiint_{\mathbb{R}^3} e^{-i\boldsymbol{k}\boldsymbol{r}} g(r) \mathrm{d}\boldsymbol{r} \tag{1.8}$$

Numerically the equilibrium solvation structure can be obtained either by explicit solvent simulation or liquid state theories and cannot be obtained with implicit solvent methods (see following sections and chapters for more information on these methods). The molecular solvation structure, that depends on the position and the orientation of the solvent molecules is a direct output of liquid state theories whereas one needs to accumulate lots of data from long trajectories to obtain it from explicit solvent simulations.



(a) (b)

Figure 1.3: Illustration of structural information: (a) radial distribution function between water oxygen atoms (black line: experimental results, grey line: TIP3P simulations, dashed line: SPC simulations) [15], (b) crystallographic solvent molecules on a protein.

1.2 MODELLING SOLVATION EFFECTS

As mentioned above, computing solvation free energies and even solvation profiles is not trivial. To do this there are 3 families of *in silico* methods to study solvation, recapitulated in table 1.1 and presented in more detail in the following chapters [16]. The first two approaches, implicit calculation, also called polarizable continuum models (PCM), and explicit solvent simulations have been wildly used for years, the former for its cheap computational cost and the latter for its precision. However, in the last few decades, the third approach, liquid state theories (LST), is gaining momentum because of their good balance between precision, simplicity, and speed.

Method	Speed	Structure	Energetics
Implicit solvent calculation	Fast	No	Yes
Explicit solvent simulations	Slow	Yes	Yes
Liquid state theories	Fast	Yes	Yes

Table 1.1: Summary of computational methods to study solvation

Some approaches that combine multiple methods, *eg.* treating the first few solvation layers with explicit solvents and treating the long-range interactions with either an implicit solvent model or LST [17, 18]; or using explicit solvent simulations to compute the solvation structure from trajectories and plug it to a functional of solvent density to extract local thermodynamic information, such as solvation free energies. The family of the latter approaches is called inhomogeneous solvation theories (IST) [19, 20, 21] that include methods like WaterMap [22, 23], solvation thermodynamics of ordered water (STOW) [24] and grid IST (GIST) [25].

Beyond choosing the computational theory, one needs to choose how to model the solute, either with a quantum (QM) or a classical force field (MM) representation, and, in the case of explicit solvent, how to model the solvent also either with a QM or a MM representation. Table 1.2 summarises the different computational approaches to study solvation as a function of the solute and solvent representations. In the force field representation, the atomic charges can be treated as either fixed point charges or fluctuating with a polarizable force field model. There are also some coarse-grain models, intermediates between implicit and explicit solvent/atom models, which gather a group of atoms into a single interaction site [26].

Solute	Solvent	Methods
Quantum	Quantum	*Ab initio* simulations
Quantum	Force field	QM/MM and QM/LST
Quantum	Implicit	QM/PCM
Force field	Force field	Classical simulations and LST
Force field	Implicit	MM/PBSA or MM/GBSA

Table 1.2: Summary of computational methods to study solvation depending on the solute and solvent models.

As the focus of this work is MDFT, a classical liquid state theory, and classical explicit solvent simulations are used as a reference, we used force field representation for the solutes and the solvent and restrict ourselves to fixed point charge models for simplicity. The most common force field representation of non-bonded interactions is a pair potential composed of a Lennard-Jones term and a Coulomb potential,

$$U_{\text{non-bonded}}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] - \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{1.9}$$

where r_{ij} is the distance between atomic sites i and j, $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ are the mixed Lennard-Jones parameters, and ε_0 the vacuum permittivity. The bonded intramolecular potentials are given in equation 6.2.

For the solvent, as we studied hydration, we used either the rigid SPC/E or TIP3P water models. Their force field parameters are detailed in table 1.3. Both models are simple 3-sites point charge models with a single Lennard-Jones sphere sitting the oxygen atom and partial charges situated on the oxygen and the hydrogen sites (fig. 1.4). Both models are widely used by the community.

	σ [Å]	ϵ [kJ/mol]	r_{OH}[Å]	q_{O} [e]	q_{H} [e]	θ [°]	ϵ_r
SPC/E[a]	3.11600	0.6500	1.0000	-0.820	+0.410	109.47	71
TIP3P[b]	3.15061	0.6364	0.9572	-0.834	+0.417	104.52	99
Exp.[c]			0.9910			105.5	78[d]

Table 1.3: Force field parameters and relative permittivity of SPC/E and TIP3P water models. Refs. [a][27],[b][28],[c] [29] and [d][30]

In general, water cannot be perfectly described with a simple pair-potential force field due to multi-body effects, hydrogen bonding, quantum effects etc. Multiple models containing different numbers of sites have been developed, with the principle of that increasing the number of sites should improve the quality of the model. However, for simple 3-sites models, SPC/E produces relatively well the solvation structures and TIP3P the solvation free energies. For further information on water models, Martin Chaplin has collected a great review of the most commonly used water models [31].

1.3 OTHER APPLICATIONS

1.3.1 SYNTHESIS/PRODUCTION

As mentioned before, the largest majority of natural or laboratory chemical reactions happens in organic or aqueous solutions. At the end of a chemical synthesis, in laboratory or industry, the aim

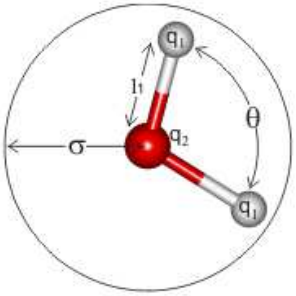


Figure 1.4: Illustration of a three-point water model.

is to separate and collect the final molecule/product of interest. One of the easiest ways to do it is to have selective precipitation of the final product. To do this optimally one needs to know the relative solubilities $\log S$ of the different compounds in the solution. Furthermore, most organic chemistry reactions are done in organic solvents and one of the most common ways to recover the final product is via liquid-liquid extraction between the non-miscible organic phase, commonly modelled with cyclohexane, and aqueous phase. To do this efficiently, one should know the partition coefficient $\log P$ of the molecule of interest between the solvents. Additionally, the last few years have seen a large push towards 'green chemistry' [32]. One of the principles is to 'use safer solvents and reaction conditions' and which means to substitute more toxic and environmentally hazardous organic solvents, like toluene, to aqueous solutions if possible. One possible way to do this is to modify the substances so that they will be more soluble in water, i.e. try to increase the hydration free energy of the compound.

1.3.2 AQUATIC TOXICITY

Another important use of solubilities and partition coefficients is for the predictions of aquatic toxicity [33]. All chemicals authorised by the European Chemical Agency (ECHA) require the prediction of the water solubility and the n-octanol/water partition coefficient $\log P_{ow}$ which can be done in some cases with QSAR models [34, 35]. These are important information for the protection of the environment. Furthermore, water-soluble substances gain access to humans and other living organisms and the $\log P_{ow}$ models the capacity of a compound to pass the cell membrane. These are properties that we want to avoid in most compounds, *eg.* used in food packaging.

1.4 SCOPE OF THIS THESIS

In this thesis, we aim to develop and apply two 'newish' solvation free energy calculation methods: hybrid 4th dimension Monte-Carlo (H4D-MC), originally developed for grand canonical simulation [36], and molecular density functional theory (MDFT) [37, 38, 39, 40].

Part I reviews a selection of models and methods to study solvation with a focus on the computation of solvation free energies. It presents briefly the three main families of SFE calculation methods: implicit solvent calculations, explicit solvent simulation and liquid state theories.

Part II presents the developments and results of the H4D-MC approach for simple spherical solutes and small organic molecules. A special focus is given on including solute flexibility in H4D-MC and to an analysis of the effect of solute flexibility on the hydration free energies of small drug-like molecules.

Part III focus on the MDFT in the hyper-netted chain approximation (HNC). It includes a presentation of the development made to MDFT-HNC, a rigorous benchmarking of MDFT-HNC from simple hydrophobic spheres to molecular solutes, through ions, and a chemoinformatics analysis on the performance of MDFT-HNC for small drug-like molecules.

Part I

STATE-OF-THE-ART: SOLVATION MODELS AND METHODS

The aim of this thesis is to develop a method to compute solvation free energies *fast*. Today the reference to that is polarizable continuum models. It's worth starting by a few pages to make clear what PCMs are and what are their limitations which are related to their accuracy.

In chapter 3, I describe the free energy perturbation approach, the reference for computing SFEs accurately. There are possible enhancements of the FEP approach with the use of alchemical intermediates with either stratification or slow growth. They are also described in this chapter. The inconvenience of FEP and its enhancements is that they are very time-consuming.

Liquid state theories make today the promise of fast computation of accurate SFEs. Chapter 4 gives an introduction to these liquid state theories with a focus on the classical density functional theory approach.

2

CONTINUUM MODELS

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The refence in *fast* computation of solvation free energies is continuum solvent approaches.

How to compute solvation free energies with implicit solvent? What are their limitations ?

Implicit solvent models consider the solvent as a polarizable continuous isotropic medium characterized by a dielectric constant ε_r with the solute M placed in a cavity within this medium (figure 2.1). These mean-field approaches referred to as polarizable continuum models (PCM), are numerically very cheap compared to explicit solvent calculations presented in the following chapter. Hence they are popular in QM calculation or classical simulations of large biomolecular systems. This chapter is mainly based on the reviews by Roux and Simonson [41] and Skyner et al. [16]. More in-depth discussion, especially for QM calculations, are available in the reviews by Cramer and Truhlar [42] and by Tomasi et al. [43, 44].

In the continuum models, the solvation free energy is decomposed into a non-polar and a polar part,

$$\Delta G_{\text{solv}} = \Delta G_{\text{np}} + \Delta G_{\text{elec}} \tag{2.1}$$

where ΔG_{np} is usually referred to as the 'cavity formation free energy' and ΔG_{elec} as the 'charging free energy'. The former is modelled by solvent accessible area term and the latter by a continuum electrostatic reaction field.

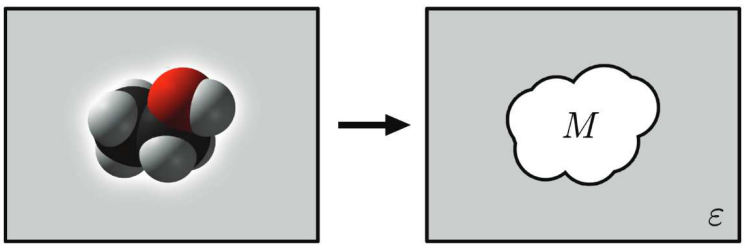


Figure 2.1: Illustartion of continuum solvent model: a solute formed cavity in mean field characterized by the solvent dielectric constant.

2.1 NON-POLAR TERM

In scale particle theory (SPT) [45, 46, 47], which describe the free energy of inserting a non-polar repulsive sphere into a solvent, the reversible work to produce a spherical cavity of radius R can

be rigorously calculated for a hard-sphere liquid of bulk density n_{bulk} as long as $2R \leq a$, the hard-sphere diameter (for water with a non-polar solute $a = 2.75$ Å [46]) and reads,

$$W(R) = -k_B T \ln\left(1 - \frac{4}{3}\pi R^3 n_{\text{bulk}}\right) \tag{2.2}$$

In the limit of a large cavity, or solute particle, thermodynamic considerations lead to

$$W(R) = \frac{4}{3}\pi R^3 P + 4\pi R^2 \gamma \left(1 - \frac{\delta}{R}\right) + \ldots \tag{2.3}$$

where P is the isotropic pressure, γ_s the solvent surface tension and δ a molecular length scale related to the so-called curvature correction of the surface tension ($\delta \sim 0.5$ Å for water [46]). The remaining terms are assumed to be negligible. In practice, (i) the PV-like (in R^3) term is expected to be negligible, an atmospheric pressure corresponds to an energy of 1.5×10^{-5} kcal/mol per Å^3, and the free energy is dominated by the surface term; and (ii) the length scale is such that the curvature dependence is only significant when the radius R is very small.

SPT provides an important concept of relating the non-polar free energy contribution to the surface area of the solute. If one neglects pressure and the curvature effect, the non-polar term of solvation free energy can be approximated as

$$\Delta G_{\text{np}} = \gamma_s A(M) \tag{2.4}$$

where is $A(M)$ is the solute surface of the solute in the M conformation which is most commonly defined as solvent-accessible surface area (SAS) A_{SAS} (figure 2.2).

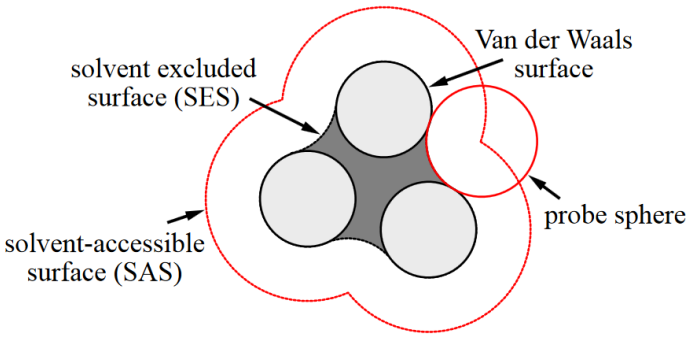


Figure 2.2: Illustration of the solute surfaces. The solvent accessible surface (SAS) is traced out by the centre of the probe sphere, representing a solvent molecule, when rolling on the atoms defined by their van der Waals radii. The solvent excluded surface (SES) is the topological boundary of the volume where the probe sphere can not penetrate.

Another approach is to assume that the non-polar term of solvation free energy can be represented as a linear sum of atomic contributions weighted by their solvent-exposed area,

$$\Delta G_{\text{np}} = \sum_{i}^{\text{atoms}} \xi_i A_i(M) \tag{2.5}$$

where ξ_i is a parametrized atomic free energy per unit are constant for each atom type and A_i the solvent-exposed area of the atom i which depends on the solute configuration M.

The deficiencies of these simple surface area approaches have been recognized and further decomposition of the non-polar term into cavity ΔG_{cav} and van der Waals dispersion ΔG_{vdW} has been proposed $\Delta G_{\text{np}}(M) = \Delta G_{\text{cav}} + \Delta G_{\text{vdW}}$ and have shown to improve results [48].

2.2 CHARGING TERM

For an isotropic solvent with random thermal motion, the average electric field is null at any given point. When introducing a charged solute, or a solute with partial charges, it creates a net change in orientation, introducing an overall change in the charge distribution represented by a continuous electric field $E(\boldsymbol{r})$, known as the 'reaction field'. The charging free energy in a continuum model, *i.e.* the work required to create the charge distribution is determined by

$$\Delta G_{\text{elec}} = \frac{1}{2}\int \mathrm{d}\boldsymbol{r}\rho_q(\boldsymbol{r})V_{\text{elec}}(\boldsymbol{r}) \tag{2.6}$$

where ρ_q is the interaction of solute charge density and $V_{\text{elec}}(\boldsymbol{r})$ electrostatic potential.

The Maxwell-Gauss equation reads

$$\nabla \cdot D(\boldsymbol{r}) = \frac{\rho_q(\boldsymbol{r})}{\varepsilon_0} \tag{2.7}$$

where $D(\boldsymbol{r}) = \varepsilon_0 E(\boldsymbol{r}) + P(\boldsymbol{r}) = \varepsilon_r(\boldsymbol{r})E(\boldsymbol{r})$ is the electric displacement field, $P(\boldsymbol{r})$ the system's polarization field, $\varepsilon_r(\boldsymbol{r})$ the solvent's position-dependent dielectric constant and ε_0 the vacuum permittivity. As the electric field at any given point is the gradient of the electrostatic potential $V_{\text{elec}}(\boldsymbol{r})$, the Maxwell-Gauss equation can be rewritten as a second-order differential equation, called the Poisson equation :

$$\nabla \cdot [\varepsilon_r(\boldsymbol{r})\nabla V_{\text{elec}}(\boldsymbol{r})] = -\frac{\rho_q(\boldsymbol{r})}{\varepsilon_0}. \tag{2.8}$$

This equation can not be solved analytically for complex geometries such as large molecules and hence it needs to be solved numerically with appropriate methods [49, 41, 50, 51, 52]. The Poisson equation is valid only for systems under non-ionic condition, whereas in a real solution, dissolving a solute produces a mobile electrolyte. This effect is taken into account by coupling the Poisson equation with the thermal Boltzmann distribution, yielding the following Poisson-Boltzmann (PB) equation

$$\nabla \cdot [\varepsilon_r(\boldsymbol{r})\nabla V_{\text{elec}}(\boldsymbol{r})] - \frac{2zn_{\text{ion}}}{\varepsilon_0}\sinh\left(\beta q V_{\text{elec}}(\boldsymbol{r})\right) = -\frac{\rho_q(\boldsymbol{r})}{\varepsilon_0} \tag{2.9}$$

where $z = |q|$ is the valence of the ions and n_{ion} the ion density. The slow convergence of the PB equation makes the prediction of the electrostatic potential for the charging free energy computationally expensive and often inaccurate. It works best for systems where the solute cavity is near-spherical or ellipsoidal, but for systems with more complex geometries, it is cumbersome. Therefore most continuum models use derivations approximating the Poisson equation. In the classical cases, the most common methods are the generalized Born (GB) equation and self-consistent reaction field models (SCRF).

For a simple net charge q in a spherical cavity of radius a the charging free energy is given by the Born formula:

$$\Delta G_{\text{elec}} = -\frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2}{2a}. \tag{2.10}$$

For more complex geometries the empirical GB model gives the charging free energy as a superposition of several net charges in spherical cavities described by the Born formula :

$$\Delta G_{\text{elec}} = -\frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\sum_i\sum_j\frac{q_i q_j}{f_{ij}} \tag{2.11}$$

with

$$f_{ij} = \sqrt{r_{ij}^2 - a_i a_j \exp\left(\frac{r_{ij}^2}{4 a_i a_j}\right)} \tag{2.12}$$

where r_{ij} is the distance between the centre of atoms i and j; and a_i the (empirical) effective Born radius of the atom i. GB provides a very fast method with similar accuracy to PB for computing the charging free energy and hence has large success in molecular modelling. PB/GB methods coupled with the cavity term and molecular mechanics for the possible solute movements are called MM/PBSA and MM/GBSA approaches [53].

The other approaches are self-consistent reaction field models, such as the Onsager model which models a dipole(s), characterized by a (total) dipole moment μ, in a spherical cavity [54, 55] or the Kirkwood-Westheimer model with a general multipole model in a spherical or ellipsoidal cavity [56, 57]. These simple models are not fully capable of predicting solvent behaviour in many realistic cases.

All of the previous models use (point) charges inside the cavity, whereas PCM models for quantum calculations use the apparent surface charge of the solute to compute the charging free energy. These methods can be split into two categories [58]: dielectric (DPCM) where the solvent is treated as a polarizable medium [59], and conductor-like (CPCM) polarizable continuum models, where the solvent is treated as a conductor-like medium.

The DPCM uses the exact dielectric boundary condition for the calculation of the polarization charge densities σ on the surface segments of the solute cavity M. Currently the most commonly used DPCM methods are integral equation formalism PMC (IEFPCM) [60, 61] which has shown to perform well compared to other DPCM models [44].

The CPMC uses a simpler boundary condition of a conductor and takes into account the reduction of the polarization charge densities occurring at finite permittivity by a slightly empirical scaling. The most common CPCMs are "conductor-like screening model" (COSMO) [62] and its derivative for the computation of chemical potentials "conductor-like screening model for real solvent" (COSMO-RS) [63, 64].

2.3 LIMITATIONS

By definition, continuum solvent models are mean-field approaches where the solvent is represented by only its approximate electrostatic properties and lack all molecular information of the solvation structure. The solvation structure can have an important effect on the SFE as the solvation enthalpy depends on the change of the number of solvent-solvent bonds and the entropy on order in the solvent. These effects are especially important with protonic polar solvents like water where the bulk solvent is ordered by hydrogen bond networks and where the solvation structure can by guided by hydrogen bonding. Dismissing these structural effects leads to important inaccuracies in the implicit solvent models.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The refence in *fast* computation of solvation free energies is continuum solvent approaches.

Here, we briefly described emprical way of computing solvation free energies with continuum models. These approaches are very fast but lack precision as they miss all molecular information of the solvent. In the next chapter, we will describe how to calculate them exactly.

3

SOLVATION FREE ENERGIES WITH SIMULATIONS

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The *exact* approach is to compute them from explicit solvent simulations with free energy perturbation.

This chapter briefly presents the free energy perturbation methods to compute solvation free energies and approaches the improve the statistics, mainly with the use of alchemical intermediates either by stratification or by a time-dependent coupling parameter.

The free energy methods presented in this chapter are based on classical molecular dynamics or Monte Carlo simulations with a classical force field representation of the solute and solvent molecules. As mentioned before the computation of a solvation *free* energy is not trivial and cannot be obtained from a single simulation of the solvated system or from a simple difference between simulations of the solvated system and the bulk solvent, but require the sampling of all possible states that can be visited during the transformation. The majority of this section is based on information from AlchemistryWiki edited by M.R. Shirts, L. Nade, D.L. Mobley and J.D. Chodera [65] and lecture notes of M. S. Shell [66].

The basic idea behind all free energy calculations and methods yields from a core equation derived from statistical mechanics: the free energy difference in a canonical (NVT) ensemble is

$$\Delta F_{ij} = -k_B T \ln \frac{Q_j}{Q_i} \tag{3.1}$$

where ΔF_{ij} is the Helmholtz free energy difference between states i and j, k_B the Boltzmann constant, T the temperature and Q the canonical partition function.

3.1 FREE ENERGY PERTURBATION

One of the earliest free energy methods consists of computing the free energy difference between a reference state and some perturbed state via exponential averaging. Perturbation techniques have a long history in statistical mechanics and were pioneered by Born and Kirkwood in the 1920s and '30s. Robert Zwanzig introduced the free energy perturbation (FEP) method in the context of MD and MC simulations in 1954 [67]. Starting from equation 3.1 the Zwanzig relationship[1] reads,

$$\begin{aligned}
\beta\Delta F_{01} &= -\ln \frac{Q_1}{Q_0} = -\ln \frac{\int e^{-\beta U_1(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N}{\int e^{-\beta U_0(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N} \\
&= -\ln \frac{\int e^{-\beta \Delta U(\boldsymbol{q}^N) - \beta U_0(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N}{\int e^{-\beta U_0(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N} = -\ln \int p_0(\boldsymbol{q}^N) e^{-\beta \Delta U(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N \\
&= -\ln \langle e^{-\beta \Delta U(\boldsymbol{q}^N)} \rangle_0
\end{aligned} \tag{3.2}$$

1 Free energy perturbation ≡ exponential averaging ≡ Zwanzig relationship/equation

where $\beta = \frac{1}{k_B T}$ is the inverse of the thermal energy, $U_0(\boldsymbol{q}^N)$ and $U_1(\boldsymbol{q}^N)$ are the potential energy of the initial state 0 and the perturbed state 1, $\boldsymbol{q}^N$ the spatial coordinates and the momenta of the N particles system, $\Delta U = U_1 - U_0$ the potential energy difference between states initial and perturbed state, $p_0 = \frac{e^{-\beta U_0}}{\int e^{-\beta U_0}}$ the probability of having the system in state $U_0(\boldsymbol{q}^N)$ and $\langle \cdots \rangle_0$ the ensemble average over the state 0. The free energy difference is obtained exponentially averaging the potential energy difference between the initial and perturbed state for an ensemble of configuration obtained by the propagation of an equilibrium simulation in the initial state 0. The Zwanzig equation is asymmetric and the free energy difference can be obtained from a simulation propagated in state 1 via

$$\beta \Delta F_{01} = \ln \frac{Q_0}{Q_1} = \ln \langle e^{\beta \Delta U} \rangle_1. \tag{3.3}$$

In principle the Zwanzig equation is straightforward. However, calculating a correct free energy difference is hard due to important statistical errors because of the exponential averaging. The Zwanzing equation can be reformulated as

$$\beta \Delta F_{01} = -\ln \langle e^{-\beta \Delta U} \rangle_0 = -\ln \int e^{-\beta \Delta U} p_0(\Delta U) \mathrm{d}\Delta U \tag{3.4}$$

with $p_0(\Delta U) = \frac{\int \delta[U_1(\boldsymbol{q}^N) - U_0(\boldsymbol{q}^N) - \Delta U] e^{-\beta U_0(\boldsymbol{q}^N)} d\boldsymbol{q}^N}{\int e^{-\beta U_0(\boldsymbol{q}^N)} d\boldsymbol{q}^N}$. As illustrated in figure 3.1, $e^{-\beta \Delta U}$ grows very large for negative values of ΔU whereas $p_0(\Delta U)$ typically peaks at some intermediate value of ΔU and tends to zero away from it. An important part of $e^{-\beta \Delta U} p_0(\Delta U)$ is in a region where $p_0(\Delta U)$ is quasi-zero. Hence some configurations that have a large contribution to the integral due to the exponential averaging are sampled very rarely in the simulation, *i.e.* a substantial portion of the average depends on the rare events of the simulation in state 0. Therefore the simple FEP can only be used for very small perturbations where $\Delta U \approx 0$ as larger perturbations result in very large statistical errors.

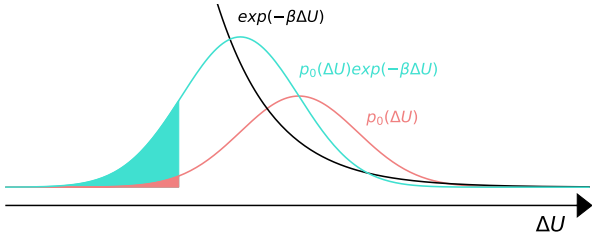


Figure 3.1: Problem of computing free energy differences. Representation of equation 3.4 terms.

3.1.1 WIDOM TEST PARTICLE

The Zwanzig equation and its derivatives (see below) are highly general and can be applied to any free energy calculation. They are based on the direct calculation of the overall partition functions $Q_0(N)$ and $Q_1(N)$ yielding the free energy difference of the system ΔF can be applied to any free energy calculations with a variation of U. Benjamin Widom proposed a method in 1963 for computing a free energy difference in the specific case of changing the number of particles in the system, *eg.* computing solvation free energy of the new particle [68]. The Widom insertion method is an application of the Jarzynski equality [69] since it measures the free energy difference via the average work to change the system from a state with N molecules (with $Q(N)$) to a state with $N+1$ molecules (with $Q(N+1)$). It yields the excess chemical potential of one component rather

than the system free energy as $\mu_{\text{exc}} = \frac{\Delta F}{\Delta N}$ with $\Delta N = 1$. The excess chemical potential, *i.e.* the difference between the chemical potential of a given species and that of an ideal gas under the same conditions reads as

$$\beta\mu_{\text{exc}} = \ln \frac{\int e^{-\beta U(\boldsymbol{q}^N)} \mathrm{d}\boldsymbol{q}^N}{\int e^{-\beta U(\boldsymbol{q}^{N+1})} \mathrm{d}\boldsymbol{q}^{N+1}}. \tag{3.5}$$

This problem can be reformulated in the formalism of FEP, where the perturbation is to 'turn on' the interaction of $(N+1)^{\text{th}}$ molecule of the system with all of the other molecules, *i.e.* the $(N+1)^{\text{th}}$ molecule is converted from a non-interacting ideal gas particle to an interacting one. The potential energies of the initial and perturbed states are $U_0(\boldsymbol{q}^{N+1}) = U(\boldsymbol{q}^N)$ and $U_1(\boldsymbol{q}^{N+1}) = U(\boldsymbol{q}^{N+1})$ respectively and the energy difference is $\Delta U = U(\boldsymbol{q}^{N+1}) - U(\boldsymbol{q}^N) = u_{cross}$ with $u_{cross}$the interaction energy between the inserted particle with all other particles in the system. In this case, the Zwanzig equation can be written as

$$\beta\mu_{\text{exc}} = -\ln\langle e^{-\beta u_{cross}}\rangle_N. \tag{3.6}$$

The asymmetric Zwanzig equation can be applied to particle destruction and the excess chemical potential can be obtained with $\beta\mu_{\text{exc}} = \ln\langle e^{\beta u_{cross}}\rangle_{N+1}$. The practical details for test particle insertions and destructions are detailed in algorithms 3.1 and 3.2 respectively.

Algorithm 3.1 Widom test particle insertion

1. perform an equilibrium simulation of N solvent molecules
2. periodically pause the simulation and insert a test particle, *i.e.* the solute molecule, at a random position and orientation in the simulation box, *i.e.* convert the $(N+1)^{\text{th}}$ ideal gas particle into the interactive test particle
3. compute u_{cross} $(\equiv \Delta U)$
4. remove the test particle and continue the simulation as if the test never occurred
5. compute the average $-\ln\langle e^{-\beta u_{cross}}\rangle_N$ over the simulation.

Algorithm 3.2 Widom test particle destruction

1. perform an equilibrium simulation of $N+1$ particles, *i.e.* N solvent molecules and the solute molecule
2. periodically pause the simulation and remove a particle, *i.e.* the solute molecule, from the simulation box, *i.e.* convert the $(N+1)^{\text{th}}$ interacting test particle to an ideal gas particle
3. compute $-u_{cross}$ $(\equiv -\Delta U)$
4. replace the test particle and continue the simulation as if the test never occurred
5. compute the average $\ln\langle e^{\beta u_{cross}}\rangle_{N+1}$ over the simulation.

Note that, Widom insertion and destruction are instantaneous, *i.e.* there is no relaxation of the test molecules or the environment. This leads to large statistical errors as (i) the particle insertion fails for systems with a large density as there is almost always a core overlap resulting in large ΔU and (ii) as the removal of 'large' solutes is very unfavourable due to the creation of non-physical cavities in the solvent.

3.1.2 BENNETT ACCEPTANCE RATIO

In 1976 Charles Bennett proposed an approach, later called the Bennett acceptance ration (BAR) [70], to optimize the Zwanzig equation as to minimize the statistical error via error propagation. He rewrote the FEP equation as,

$$\begin{aligned}\beta\Delta F_{01} &= -\ln\frac{\int e^{-\beta U_1(q^N)}\mathrm{d}q^N}{\int e^{-\beta U_0(q^N)}\mathrm{d}q^N}\\ &= -\ln\left(\frac{\int e^{-\beta U_1(q^N)}\mathrm{d}q^N}{\int w(r^N)e^{-\beta U_0(q^N)-\beta U_1(q^N)}\mathrm{d}q^N}-\frac{\int w(r^N)e^{-\beta U(q^N)-\beta U_1(q^N)}\mathrm{d}q^N}{\int e^{-\beta U_0(q^N)}\mathrm{d}q^N}\right)\\ &= -\ln\frac{\langle we^{-\beta U_1}\rangle_0}{\langle we^{-\beta U_0}\rangle_1}\end{aligned} \tag{3.7}$$

where $w(q^N)$ is a weight function and the two averages imply two simulations: one simulation in state 0 averaging values of $we^{-\beta U_1}$ and another simulation in state 1 averaging values of $we^{-\beta U_0}$. Note that this expression combines information from the transformations in both ways and it is symmetric, *i.e.* both states 0 and 1 appear in equal roles. The idea is to find an optimal value of w that minimizes the expected statistical error of the free energy difference. Using standard error propagation rules to determine $\sigma^2_{\beta\Delta F}$ and minimizing it variationally with respect to w, he found the optimal value of weight function as,

$$w(r^N) \propto \left(\frac{e^{-\beta F_0-\beta U_1(q^N)}}{n_0}+\frac{e^{-\beta F_1-\beta U_0(q^N)}}{n_1}\right)^{-1} \tag{3.8}$$

where n_0 and n_1 are the numbers of configurations used from states 0 and 1. Details of the derivation are given by Frenkel and Smit [71] and they showed that the proportionality constant does not affect the statistical error. If we assume $n_0 = n_1$ and plug the equation 3.8 into equation 3.7, we obtain the following BAR equation

$$\beta\Delta F_{01} = \ln\langle\frac{1}{1+e^{-\beta\Delta U+\beta\Delta F_{01}}}\rangle_0 / \langle\frac{1}{1+e^{\beta\Delta U+\beta\Delta F_{01}}}\rangle_1. \tag{3.9}$$

Note that, one can not perform a simple average during the simulations as ΔF_{01} appears on both sides of the equation. Instead, we must save a list of energies during the simulation in order to evaluate the averages by solving the BAR equation self-consistently at the end. The BAR equation can be shown to be identical to the Ferrenberg-Swendsen multiple histogram reweighting technique in the limit that the histogram bin size goes to zero, another free energy calculation method not presented here [72].

BAR method provides a statistically optimal estimator of a free energy difference between two states. However even with BAR one obtains large statistical errors if the two states are substantially different, which is the case for most solvation free energy calculations, as the there is little to no overlap between the phase spaces explored by the two states, *i.e.* ΔU is too large.

3.2 ALCHEMICAL INTERMEDIATES

To overcome the problem of the small, or non-existent, overlap between the initial and final state, one can divide the total FE difference into a series of small steps in which the system hops gradually from state 0 to state 1. This is done with the use of 'partial' solutes, *i.e.* alchemical intermediates, for which the solute-solvent interaction u^{λ}_{cross}increases gradually from 0 (non-interacting ideal particle) to u_{cross} (fully interacting solute molecule). For this, we introduce a coupling parameter λ that links the alchemical intermediate states to the initial and final states typically via

$$U_\lambda = (1-\lambda)U_0 + \lambda U_1. \tag{3.10}$$

There are two ways to set the coupling parameter λ:

1. Stratification: λs are set at fixed values defined *a priori* and an equilibrium simulation is propagated at each U_λ potential
2. Slow/fast growth: a time-dependant $\lambda(t)$ that evolves *on the fly* during an out-of-equilibrium simulation
3. Extra degree of freedom: λ dynamical variable that varies during a simulation and the potential of mean force is constructed from the measured λ-distribution (not discussed here for simplicity).

3.2.1 STRATIFICATION: DISCRETISED COUPLING PARAMETER

The idea of stratification is to decrease the ΔU between two states, and in consequence improve the statistics, by simulating m intermediates states between the initial and final state, with fixed values of λ and computing the free energy difference between two following states $\Delta F_{\lambda_i\lambda_{i+1}}$. The total free energy difference is the sum of the $m+1$ individual free energy differences :

$$\begin{aligned}\beta(F(\lambda=1)-F(\lambda=0)) &= \ln\frac{Q(\lambda=0)}{Q(\lambda=1)}\\ &= \ln\left(\frac{Q(\lambda_0=0)}{Q(\lambda_1)}\frac{Q(\lambda_1)}{Q(\lambda_2)}\cdots\frac{Q(\lambda_m)}{Q(\lambda_{m+1}=1)}\right)\\ &= \beta\sum_{i=0}^{m}\Delta F_{\lambda_i\lambda_{i+1}}. \end{aligned} \tag{3.11}$$

Stratification can strongly reduce the statistical error and it permits the computation of free energy differences for systems with $\Delta U \gg 0$ such as solvation free energies. However $K = m+2$ equilibrium simulations must be performed in order to compute the total free energy difference. One uses typically ~ 20 intermediate states for a solvation free energy calculation. Hence, the statistical accuracy comes at the expense of more simulations.

In the case where the intermediates states are non-physical, *eg.* in the solvation process, the process is called an alchemical transformation but stratification can also be used for physical intermediates where for example λ fixes the position of a molecule in a potential of mean force calculations. For alchemical transformations, the simple linear scaling presented in equation 3.10 raises some problems as it gives very unequal phase space overlaps. This linear mixing rule is sufficient for some changes, such as for the change of atomic partial charge. But for cases where the number of interacting particles changes, the free energy differences for low values of λ diverge [65]. This 'end-point catastrophe' is due to the singularity at $r=0$ because of the r^{-12} term in the LJ potential. This singularity causes mostly problems to thermodynamic integration (see 3.3) but also to FEP calculation.

Hence, in the case of alchemical transformation, the intermediate states are divided into 2 categories (figure 3.2) :

- the van der Waals part is treated with 'soft core potential' to get around numerical instabilities [73, 74]. In the soft core potential, the alchemical variable λ is coupled with the configuration variable r to smoothing out the singularity. This is corresponds to adding a extra dimension to r and it reads,

$$U(\lambda,r) = 4\epsilon\lambda^n\left[\left(\alpha(1-\lambda)^m+\left(\frac{r}{\sigma}\right)^6\right)^{-2}-\left(\alpha(1-\lambda)^m+\left(\frac{r}{\sigma}\right)^6\right)^{-1}\right] \tag{3.12}$$

where α is positive constant (usually 0.5), m and n are positive integers (usually 1) and, σ and ϵ the Lennard-Jones parameters of the solute site. Typically one uses $\sim 15-20$ values of λ_{LJ} for the van der Waals (vdW) part.

- the electrostatic part is treated with linear scaling of the partial charges. Typically one uses ~ 4 values of λ_{elec} for the electrostatic part.

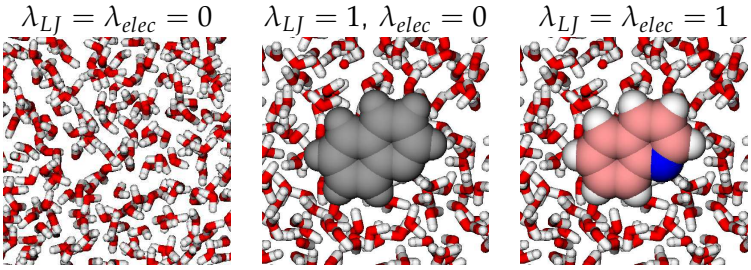


Figure 3.2: Alchemical transformation: staring from the bulk solvent with solute as a non-interacting particle, the solute is created by first turning on the non-polar interactions and then turning on the charges.

The statistics of stratification can be improved by applying (i) BAR to all $\lambda_i\lambda_{i+1}$ couples or (ii) using multistate Bennett acceptance ratio (MBAR) developed by Shirts and Chodera in 2008 [75]. MBAR is a generalisation of BAR, which uses data from all K states simultaneously. MBAR equation is identical to the Weighted Histogram Analysis Method (WHAM) [76], an extension of Ferrenberg-Swendsen multiple histogram reweighting technique, with zero-width bins.

3.2.2 SLOW GROWTH: TIME-DEPENDENT COUPLING PARAMETER

The idea of slow growth is to improve the statistics of a Widom test particle method by introducing the solute into the bulk with a finite speed, *i.e.* not to create the test particle suddenly into the bulk but introducing it slowly by imposing a time-dependant coupling parameter $\lambda(t)$ during an out-of-equilibrium simulation joining the states 0 and 1, which gives time to relax to the solvent environment. Solvation free energy is constructed from the external work W due to the introduction of the solute with the Jarzynski equality. The Jarzynski equality is an equation in statistical mechanics that relates the free energy difference between two states and the irreversible work along an ensemble of trajectories joining the same states and reads as

$$\beta\Delta F = -\ln\langle e^{-\beta W}\rangle. \tag{3.13}$$

It is valid no matter how fast the process happens and it has experimental (RNA folding [77]) and theoretical (Widom [68]) applications.

Recently, Luc Belloni developed a theory, and a code, 'hybrid 4th dimension'-MC (H4D-MC) for grand canonical simulations (μVT) based on the idea of adding or removing molecules via a 4th dimension with a time-dependent coupling parameter [36] to keep the chemical potential constant μ. This methodology can also be used for measuring the excess chemical potential of a solute during a canonical (NVT) or isobaric (NPT) simulation. The algorithms for measuring the solvation free energy via solute insertion or destruction are detailed in algorithms 3.3 and 3.4. The slow insertion or destruction gives time for the solvent (and the solute molecule) to relax during the transformation improving the statistics drastically compared to simple Widom.

In the case of propagation in the NPT ensemble, as it will be the case in this thesis, the Jarzynski equality is modified to

$$\beta\mu_{\text{exc}} = -\ln\langle Ve^{-\beta\Delta H}\rangle + \ln\langle\frac{1}{V}\rangle \tag{3.14}$$

where V is the simulation cell volume of the system at the time of the insertion/destruction.

As for simple FEP, the statistics of H4D-MC can be improved with coupling information from the forward and the backward transformations with the Crooks fluctuation theorem [78], a non-equilibrium version of BAR, which reads

$$p_{des}(\Delta H) = p_{ins}(\Delta H)e^{-\beta(\Delta H - \mu_{exc})}. \tag{3.15}$$

where $p_{ins}$and p_{des} are the insertion and destruction distributions of $\Delta H = \Delta H_{N+1} - \Delta H_N$ respectively.

This in-house code by Luc Belloni was originally written for rigid solutes whereas most of the well known MD codes are written for flexible molecules and one needs to use constraints to have a rigid solute. As MDFT calculations are done on rigid solutes, we need reference data for rigid solutes. Hence H4D-MC is used as the reference method in this thesis for MDFT. Implementing solute flexibility in H4D-MC is discussed in the chapter 6.

Algorithm 3.3 H4D-MC insertion

1. perform an equilibrium simulation of N solvent molecules with intermolecular interactions given by $U_{ij}(r_{ij})$ pair potential (with $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$)
2. periodically pause the simulation and slowly insert a solute molecule, at a random position and orientation in the simulation box, via 4th dimension with a short MD simulation. The interaction potential between the solute and the solvent molecules is given by $V_{ik}\left(\sqrt{r_{ik}^2 + (w_i - w_k)^2}\right)$ where w is the 'altitude' of a molecule in the 4th dimension that governs the 3D motion of all solvent molecules. $w_i = 0$ for all solvent molecules and the solutes altitude w_k evolves during the simulation from w_{max} (input parameter) to 0 with a given speed v (input parameter).
3. compute $\Delta H = H_{N+1} - H_N$ ($\approx W$) of the insertion process
4. return the simulation to the state before the insertion and continue the simulation as the test never occurred
5. compute the average $-\ln\langle e^{-\beta\Delta H}\rangle_N$ over the simulation.

Algorithm 3.4 H4D-MC destruction

1. perform an equilibrium simulation of N solvent molecules and the solute molecule with intermolecular interactions given by $U_{ij}(r_{ij})$
2. periodically pause the simulation and slowly delete the solute molecule *via* 4th dimension with a short MD simulation where the interaction potential between the solute and the solvent molecules is given by $V_{ik}\left(\sqrt{r_{ij}^2 + (w_i - w_k)^2}\right)$. The solutes altitude w_k evolves during the simulation from 0 to w_{max} with a given speed v.
3. compute $-\Delta H = H_N - H_{N+1}$ ($\approx -W$) of destruction process
4. return the simulation to the state before the deletion and continue the simulation as the test never occurred
5. compute the average $\ln\langle e^{\beta\Delta H}\rangle_{N+1}$ over the simulation.

A similar approach with multiple short non-equilibrium MD simulations measuring the annihilation/creation work was proposed by Procacci and collaborators in 2014 [79, 80, 81]. They found that these out-of-equilibrium approaches have wall clock times that are at least one order of magnitude smaller than for stratified MD+FEP approaches. These computation times are discussed in section 5.1.1.

In practice, solvation free energies can not be done with simple FEP between initial and final states, *i.e.* the Widom test particle method, and need alchemical transformation done either by

stratification with $K \sim 25$ ergodic equilibrium simulations or by a time-dependent coupling with two ergodic equilibrium simulations coupled with multiple (over 1000) short out-of-equilibrium simulations. Both methods are exact, in the force field approximation, if the sampling is done correctly. However, these methods are time-consuming, requiring 10s to 100s of CPU-hours for a single solvation free energy.

3.3 ALTERNATIVE METHODS AND IMPROVING SAMPLING

In addition to the FEP and the approaches to improve sampling presented above, there are few other methods/approaches to either (i) compute solvation free energies or (ii) improve the sampling in a FEP calculation. They include the thermodynamic integration, weighted histogram methods, Hamiltonian replica exchange, Umbrella sampling, λ-dynamics or ensemble dynamics.

Thermodynamic integration (TI) [82] is one of the most common and intuitive methods to compute free energy differences. The idea is to compute free energy differences by taking the derivative of the free energy difference with respect to λ over m intermediate states. Starting with the identity of the free energy $\beta F = -\ln Q$, its derivative with respect to λ yields

$$\begin{aligned}
\beta\frac{dF}{d\lambda} &= -\frac{d}{d\lambda}\ln Q = -\frac{d}{d\lambda}\ln\int e^{-\beta U_\lambda(\boldsymbol{q}^N)}d\boldsymbol{q}^N \\
&= -\frac{1}{Q}\frac{d}{d\lambda}\int e^{-\beta U_\lambda(\boldsymbol{q}^N)}d\boldsymbol{q}^N \\
&= \frac{\beta}{Q}\int \frac{dU_\lambda(\boldsymbol{q}^N)}{d\lambda}e^{-\beta U_\lambda(\boldsymbol{q}^N)}d\boldsymbol{q}^N \\
&= \beta\left\langle\frac{dU_\lambda(\boldsymbol{q}^N)}{d\lambda}\right\rangle_\lambda .
\end{aligned} \tag{3.16}$$

If one does an integration over the whole λ range we get the following thermodynamic integration equation

$$\Delta F_{01} = \int_0^1 \left\langle\frac{dU_\lambda(\boldsymbol{q}^N)}{d\lambda}\right\rangle_\lambda d\lambda. \tag{3.17}$$

TI is not quite as general as FEP based techniques, since it requires the computation of Hamiltonian's derivatives and since only a finite number states are used a numerical integration scheme is required for the computation of equation 3.17. Nonetheless, if the method is used correctly it is one of the most accurate free energy computation methods. Note that the infinite potential singularity of linear alchemical potentials causes problems for TI as derivate diverges at the singularity. Therefore a soft-core potential should be used when doing alchemical transformations with TI.

Weighted Histogram Analysis Method (WHAM) [76] is the earliest method taking into account information from all intermediate states. WHAM was developed for alchemical transformations from its precursor, the first version of multiple histograms relighting techniques proposed by Ferrenberg and Swendsen [72]. It is based on the principle that if you have a finite number of states, one can create a histogram with discrete bins that provide the relative probability of observing the states of interest, assuming that the bins were created along the selected reaction path, here along the alchemical variable λ. From these probabilities, one can compute free energies among other things. In the limiting case where the bin width is chosen to be zero, the WHAM equation is equivalent to the MBAR equation.

Umbrella sampling [83] is used to improve the sampling of configurations by adding bias terms to constrain the simulation in some way. This method can lower energetic barriers and accelerate the sampling of slow degrees of freedom in the system.

Hamiltonian replica exchange (HREX) [84, 85] can be used to improve the sampling of configurations. In 'classic' stratified FEP calculations, all K states are run independently, however, these simulations can be run in parallel with each state λ_i allowed to swap atoms/molecules, under certain conditions, from another state λ_j with different energy barriers. This accelerates the sampling of configuration that are rare in state λ_i.

Expanded Ensemble dynamics (EED) [86] samples both the spatial coordinates and the alchemical variable during a single simulation. The probability of any given state is given by $P(\boldsymbol{r},\lambda) \propto \exp\left(-\beta U_\lambda + g_\lambda\right)$ with g_λ an user-specified weight factor of the λ^{th} state. EED simulations are a serialized version of the REX formalism, allowing the exploration of multiple intermediate states in a single simulation.

λ-dynamics [87, 88, 89] treats the alchemical coupling parameter λ as a dynamical variable, *i.e.* as an extra degree of freedom with a fictitious mass, and construct a potential of mean force from the measured λ-distribution.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The *exact* approach is to compute them from explicit solvent simulations with free energy perturbation.

Here, we presented the exact approach to compute solvation free energies with free energy perturbation and two approaches to improve the statistical pression with alchemical intermediates:

- to most widely used approach of stratification
- a novel approach with a time-dependent coupling parameter.

Both approaches lead to exact results, in the force field approximation, but are time-consuming: the former requires hundreads and the latter tens of CPU-hours.

4

LIQUID STATE THEORIES

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The reference of speed are emprical continuum models that lack accuracy and the exact free energy perturbation approach is very time-consuming. The liquid state approach offers a compromise between speed and accuracy.

This chapter briefly presents the use of liquid state theories to study solvation with the introduction of two LST approaches:

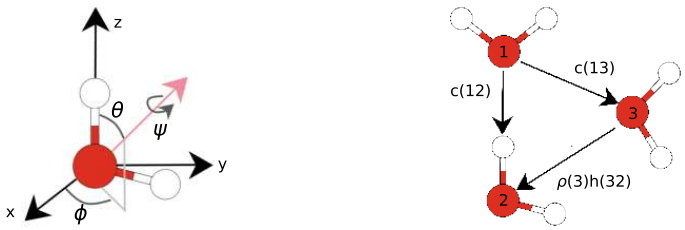
- the reference interaction site model
- the molecular density functional theory (the main theory of this thesis).

In the previous chapters, we briefly presented the widely used implicit solvent calculations and explicit solvent simulation to study solvation. Another approach to study liquids and solvation is the use of statistical mechanics methods to deduce thermodynamic quantities directly from the Hamiltonian of any given system without sampling. This chapter includes a short presentation of basic concepts of liquid state theories and the integral equation approach before a more detailed presentation of the molecular density functional approach. This chapter, based in majority on the books by Hansen and McDonald [90] and Hirata [91]; and the thesis of Guillaume Jeanmairet [92] and Lu Ding [93].

First, in statistical mechanics, a rigid molecule can be fully described by a six-dimensional vector with three positional degrees of freedom $\boldsymbol{r} = (x, y, z)$ and three orientational degrees of freedom $\omega = (\theta, \phi, \psi)$ corresponding to the Euler angles illustrated in figure 4.1a.

Secondly, liquids are characterized by their density fluctuations in time and space due to thermal fluctuations. Unlike in solids, molecules in liquids have a diffusive motion changing their position and orientation continuously, hence basic concepts when describing liquids are the local densities $n(\boldsymbol{r}) = \sum_i \delta(\boldsymbol{r} - \boldsymbol{r}_i)$ and their fluctuations $\delta n(\boldsymbol{r}) = n(\boldsymbol{r}) - n_{\text{bulk}}$. In a uniform liquid, the first moment of the density field is constant, $n_{\text{bulk}} \equiv \langle n(\boldsymbol{r}) \rangle \equiv N/V$ with N the number of particles in a system and V the volume of the system. It does not carry microscopic information of the liquid structure. However, the second moment of the density fluctuations $\langle \delta n(\boldsymbol{r}) \delta n(\boldsymbol{r}') \rangle$ contains ample information about the structure of liquids.

A natural way to represent the microstructure in liquids is the use of a pair correlation function (PCF) $g(\boldsymbol{r}_1 - \boldsymbol{r}_2, \omega_1, \omega_2) = \rho(\boldsymbol{r}_1 - \boldsymbol{r}_2, \omega_1, \omega_2)/\rho_0$ where ρ and ρ_0 are the six-dimensional density field ($n(\boldsymbol{r}) = \int \rho d\omega$) that represents the density at $\{\boldsymbol{r}_2, \omega_2\}$ when another molecule is at $\{\boldsymbol{r}_1, \omega_1\}$ and bulk density,$\rho_{\text{bulk}} = n_{\text{bulk}} i/8\pi^2$, respectively that depend on the position and the orientation of the molecules. n_{bulk} is the spatial homogeneous bulk density, typically 0.033 molecules per Å^3 for water at room conditions ($\equiv 1$ kg/L), and $i/8\pi^2$ the angular normalization constant with i the order of the main symmetry axis of the solvent molecule ($i = 2$ for water which has an axial C_{2v} symmetry along the molecules dipole with all the integrals of ψ calculated implicitly between 0 and π). It translates the probability of having two molecules at positions $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ with orientations ω_1 and ω_2, respectively, in the environment created by the surrounding molecules. The starting point



(a) (b)

Figure 4.1: Illustration on a water molecule of (a) the three Euler angles in the ZXZ representation and (b) the direct and indirect contributions to the total correlation function.

to use and determine these functions in statistical mechanics is the molecular Ornstein-Zernike equation (MOZ) [94]. It is an integral equation defining the direct correlation function $c(12)$, with the notation $(12) = (\boldsymbol{r}_1 - \boldsymbol{r}_2, \omega_1, \omega_2)$ in terms of the total correlation function $h(12) = g(12) - 1$ and it reads

$$h(12) = c(12) + \int_V c(13)\rho(3)h(32)\mathrm{d}3 \tag{4.1}$$

There are two contributing effects for the total correlation between molecules 1 and 2: (i) the direct correlation between 1 and 2, and (ii) the indirect correlation *via* a third body 3. Figure 4.1b illustrates these contributions.

As the equation contains two unknown functions h and c, another equation, a closure relation, is required to solve the MOZ equation. The exact closure relation reads as

$$g(\boldsymbol{r}, \omega) = e^{-\beta U(\boldsymbol{r},\omega)+\gamma(\boldsymbol{r},\omega)+b(\boldsymbol{r},\omega)} \tag{4.2}$$

where U is the interaction potential, typically a Hamiltonian with a sum of Lennard-Jones and Coulombic potential, $\gamma = h - c$ the indirect correlation function and b the bridge function coming from the graph theory. The bridge function is known formally as an infinite diagrammatic re-summation of virial diagrams, but is not numerically tractable and hence the closure relation needs to be approximated. The 'original' closure relation approximate is the hyper netted-chain approximation where the bridge term is neglected completely, $b = 0$. Until recently, MOZ equation in the HNC approximation was numerically very challenging to solve for molecular systems. Therefore, other approximations of the closure relations have been developed such as the mean-spherical approximation (MSA) [95], the Percus-Yevick approximation (PYs) [96], particularly adapted for hard spheres or the Kovalenko-Hirata approximation (KH) [97] a hybrid of MSA and HNC developed for the RISM approach presented in the following section.

There are two families of methods to solve the MOZ : (i) directly by solving the integral equations or their approximates or (ii) by using the classical density functional theory framework. A brief presentation of the first approach is presented in the following section before presenting the second approach in more detail as it is the core method of this thesis.

4.1 MOLECULAR INTEGRAL EQUATIONS AND RISM

The direct resolution of the fully molecular OZ equations with spatial and angular dependencies is numerically not tractable. Blum proposed to expand the angular-dependant correlation function onto rotational invariants so that the MOZ equation can be reduced to (much) smaller number of

function evaluations [98, 99, 100]. In the '80, Fries and Patey adopted this formalism to propose a numerical solution for the full HNC approximation [101]. However, the numerical resolution of the fully molecular MOZ-HNC, for shapes deviating from spheres, is cumbersome and therefore the use of molecular integral equations theory (MIET) has not gained a large success in the community.

To overcome the numerical cost, an approximate approach, called the reference interaction site model (RISM), was proposed by Chandler and Andersen in 1972 [102]. It averages the orientations of the interacting molecules and approximates the molecular solvents as combinations of spherical atomic sites with strong intramolecular correlations representing chemical bonds. It can be seen as an expansion of the generalized (non-molecular) OZ of n atomic components (eq. 4.20) with c_{ij} the typical intermolecular correlation function if i and j are not sites of the same molecule and a very strong intramolecular correlation function if i and j are sites of the same molecule. The original RISM theory accounts approximately for one of the two important chemical aspects of molecules: the geometry, in terms of the intramolecular correlation. However, it does not handle the other chemical aspect of molecules, electrostatics, in its original form. The complete characterisation of the chemical specificity of the molecular liquids became possible with the so-called extended RISM theory (XRISM) [102, 103, 104, 105]. The numerical resolution of the 'molecular' OZ equation is drastically simplified as now we have N one-dimensional problem, with N the number sites in the molecule, depending only on r each, instead of one fully molecular six-dimensional one.

Note that the original (X)RISM theory is a one-dimensional theory producing site-site radial distributions and cannot produce 3D profiles. A multitude of development and approaches have been made to obtain a three-dimensional reduction of the fully molecular solute-solvent MOZ equation [106, 107, 108, 109] with the most successful approach being the 3D-RISM [110, 97] which computes the three-dimensional solvent density $\rho(\boldsymbol{r})$ around a single fixed solute at the origin recovering the N densties of each site $\rho_i(\boldsymbol{r})$.

Beyond the solvation profile, Chandler and Singer [111], and later Kovalenko and Hirata, have shown that one can obtain analytically the solvation free energy of a solute from the RISM+HNC and 3D-RISM+KH formalism.

In summary, the RISM approaches enable the fast computation of solvation profiles and solvation free energies by approximating the MOZ equation with site-site OZ equation (SSOZ) that drastically simplified the numerical resolution as the original six-dimensional problem has been transformed into a N times a one- or three-dimensional problem. In principle, this approximation is permitted. However, there are no guarantees the MOZ closure relations, like PY or HNC, work for the SSOZ as the diagrammatic development is not proper. RISM+HNC does not have exact virial coefficients, it predicts wrong dielectric constants, and for a lot of cases the molecule 3D structures are not respected and it even does not have a solution for a lot of cases. To overcome some of these problems, multiple phenomenological modifications/corrections have been proposed without any base in liquid state theories: RISM specific closure relations (*eg.* KH [97]), additional Lennard-Jones site on water's hydrogen sites (not typically there in most water force field models) or proposal of bridge functions in r^{-1}.

Nevertheless, as the RISM approaches are numerically cheap, few tens of minutes compared to hundreds of hours with MD+FEP, and for a long time there were no numerically tractable LST alternatives, these approaches are gaining in momentum in last few years in the computation of SFEs, their derivatives and solvation profiles [91, 112, 113, 114, 115, 116]. Also, the RISM approach can be coupled with (i) *ab initio* quantum calculations, RISM-SCF/MCSCF [117] or EC-RISM (embedded cluster RISM) [118], to do QM/LST calculations where the solute(s) is treated with a quantum model and the solvent with RISM; (ii) with classical simulation (MM/RISM) to provide solvent effects for example in the conformational sampling of large biomolecules [119]; (iii) with the generalized Langevin equation (RISM-GLE) to study dynamical processes in solution [120].

4.2 MOLECULAR DENSITY FUNCTIONAL THEORY

The density functional theory of classical molecular fluids, in its atomic or molecular version [37, 38, 39], is the cousin of the well-known electronic Kohn-Sham density functional theory [121, 122], extended to finite temperature in the grand canonical ensemble by D. Mermin [123] and further developed for classical fluids by R. Evans [124, 125]. In the grand canonical ensemble, the Hohenberg-Kohn theorems can be rewritten as

1. Theorem: For a given fluid, potentially subject to an outside potential, a unique free energy functional $\mathcal{E}[\rho]$ of the fluid can be written.
2. Theorem: This functional is minimal for the fluid density corresponding to the thermodynamic equilibrium and $\Omega = \min_{\rho \to \rho_{eq}} \{\mathcal{E}[\rho]\} = \mathcal{E}[\rho_{\text{eq}}]$ is the grand potential of the system.

In consequence, if the exact expression of $\mathcal{E}[\rho]$ is known, the grand potential Ω and the equilibrium fluid density, *i.e.* the structural and energetic equilibrium properties, can be obtained by the variational principle. Even though the discovery of the electronic and classical DFT (cDFT) was quasi-simultaneous, the cDFT did not have the same success as its electronic counter-part. This was due to the prior existence of methods, such as molecular dynamics (MD) and Monte Carlo (MC) simulations, for studying systems with N-body in the fields of classical mechanics, that always stayed 'reasonably solvable numerically'.

Nevertheless, the solvation free energy of a solute can be defined as the difference of the grand potential of the solvated system Ω and the grand potential of the bulk solvent. In the cDFT framework, this difference can be expressed in a functional form :

$$\Delta G_{\text{solv}} = \Omega - \Omega_{\text{bulk}} = \min_{\rho \to \rho_{eq}} \{\mathcal{F}[\rho]\} = \mathcal{F}[\rho_{\text{eq}}] \tag{4.3}$$

where $\mathcal{F} = \mathcal{E} - \mathcal{E}_{\text{bulk}}$ is the free energy functional to be minimized, $\rho \equiv \rho(\boldsymbol{r}, \omega)$ the molecular 6-dimensional solvent density characterizing the position and the orientation of the rigid solvent molecule relative to the rigid solute, and ρ_{eq} the equilibrium solvent density. Note, that here we solve MDFT in its 3D version: $\boldsymbol{r}$ represents the absolute position of a molecule and ω represents its orientation with respect to a fixed reference and with respect to the solute as in MIET and MDFT in 1D. In the absence of solute, the equilibrium density is the homogeneous angular and spatial bulk density ρ_{bulk} .

4.2.1 MDFT FUNCTIONAL

Without approximations, the MDFT functional can be split as follow :

$$\mathcal{F} = \mathcal{F}_{\text{int}} + \mathcal{F}_{\text{ext}} = \mathcal{F}_{\text{id}} + \mathcal{F}_{\text{exc}} + \mathcal{F}_{\text{ext}} \tag{4.4}$$

where $\mathcal{F}_{\text{int}} = \mathcal{F}_{\text{id}} + \mathcal{F}_{\text{exc}}$ is the intrinsic term composed of the ideal term $\mathcal{F}_{\text{id}}$ for a fluid of non-interacting particles and the excess term $\mathcal{F}_{\text{exc}}$ that introduces structural correlations between solvent molecules; and $\mathcal{F}_{\text{ext}}$ the external term is the direct cost of the interaction of the solute and the solvent density.

The ideal term reads

$$\mathcal{F}_{\text{id}} = k_B T \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega \left[\rho(\boldsymbol{r}, \omega) \ln \left(\frac{\rho(\boldsymbol{r}, \omega)}{\rho_{\text{bulk}}} \right) - \Delta\rho(\boldsymbol{r}, \omega) \right] \tag{4.5}$$

where $k_B T$ is thermal energy (~ 0.59 kcal/mol or ~ 2.479 kJ/mol at 300K) and $\Delta\rho(\boldsymbol{r}, \omega) \equiv \rho(\boldsymbol{r}, \omega) - \rho_{\text{bulk}}$ the excess density over the bulk density.

The external contribution comes from the interaction potential U_{ext} between the solute molecule and a solvent molecule (the molecule, protein, ligand or their complex, ... embedded in water). It reads

$$\mathcal{F}_{\text{ext}} = \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho(\boldsymbol{r},\omega)U_{\text{ext}}(\boldsymbol{r},\omega). \tag{4.6}$$

Here we study classical solutes so the interaction potential uses the same non-bonded force field parameters as in a molecular dynamics simulation, typically made of Lennard-Jones (LJ) potentials and electrostatic interactions. In principle, the solute could be quantic [126].

The excess term describes the effective solvent-solvent interactions, formally known to be an infinite diagrammatic resummation of virial diagrams, but is not numerically tractable. It may be written as a density expansion around the homogeneous bulk density ρ_{bulk}:

$$\begin{aligned}\mathcal{F}_{\text{exc}} = -\frac{k_B T}{2}\int \mathrm{d}\boldsymbol{r}_1\mathrm{d}\omega_1 \int \mathrm{d}\boldsymbol{r}_2\mathrm{d}\omega_2\Delta\rho(\boldsymbol{r}_1,\omega_1)& \\ \times c^{(2)}(r_{12},\omega_1,\omega_2)\Delta\rho(\boldsymbol{r}_2,\omega_2) + \mathcal{O}(\Delta\rho^3)& \qquad (4.7)\\ = -\frac{k_B T}{2}\int \mathrm{d}\boldsymbol{r}_1\mathrm{d}\omega_1\Delta\rho(\boldsymbol{r}_1,\omega_1)\gamma(\boldsymbol{r}_2,\omega_2) + \mathcal{O}(\Delta\rho^3)& \\ = \mathcal{F}_{\text{HNC}} + \mathcal{F}_{\text{B}}& \qquad (4.8)\end{aligned}$$

where $c^{(2)}(r_{12},\omega_1,\omega_2)$ is the homogeneous solvent-solvent molecular direct correlation function with $r_{12} = |\boldsymbol{r}_1 - \boldsymbol{r}_2|$, $\mathcal{F}_{\text{B}}$ the bridge functional containing all the (unknown) terms of a higher order, $\Delta\rho^3$ and beyond, and $\gamma = c^{(2)} * \Delta\rho$ the indirect solute-solvent correlation function defined as the spatial and angular convolution of the excess density with the direct correlation function $c^{(2)}$, see eq. 4.1. The second-order direct correlation function of the bulk solvent for a given thermodynamic conditions $c^{(2)}$ is an input of the MDFT theory and is provided by previous Monte Carlo simulations coupled to integral equations calculations [17, 18], carefully corrected for finite-size effects [127] performed for the neat liquid.

If one cuts the expansion at order two in excess density, that is, if one cancels the bridge functional, one finds that the MDFT functional produces at its variational minimum the HNC equation for the solute-solvent distribution:

$$\ln g = \ln\left(\frac{\rho}{\rho_{\text{bulk}}}\right) = -\beta U_{\text{ext}} + \gamma. \tag{4.9}$$

See appendix A for the derivation. The rest of the excess term, the so-called bridge functional can be approximated empirically [128, 129, 130, 131, 132] or rigorously through higher-order direct correlation functions that are not numerically tractable as of today. The work in this thesis is performed with MDFT at its lowest level of accuracy: MDFT-HNC, *i.e.* with a vanishing bridge functional (The HNC approximation was initially qualified as the homogeneous reference fluid approximation (HRF) [133, 134, 38, 93]). This HNC level can only be improved by adding subsequent, well-funded, bridge functionals. MDFT-HNC can be considered as a rigorous basis that one can only improve.

4.2.2 MDFT ALGORITHM AND CODE

As defined by equation 4.3, once the MDFT functional is defined, to obtain the solvation free energy of the solute and its equilibrium solvation structure, one needs to minimise it for :

- a given rigid solute, composed of 'atomic' sites described by an xyz-position and a partial charge q and Lennard-Jones parameter σ, ϵ triplet (or a quantum solute, not discussed here for simplicity)

- a given rigid solvent (or a mixture of solvents, not discussed here for simplicity) described by 'atomic' sites with xyz-position and the $\{q, \sigma, \epsilon\}$-triplet, its homogeneous direct correlation function $c^{(2)}$ and its bulk density ρ_{bulk}

This minimisation is done by an in-house MDFT high-performance Fortan95 code developed by Maxilimilien Levesque, Daniel Borgis *et al.* which implements the MDFT theory. Figure 4.2 represents the main structure of the MDFT code.

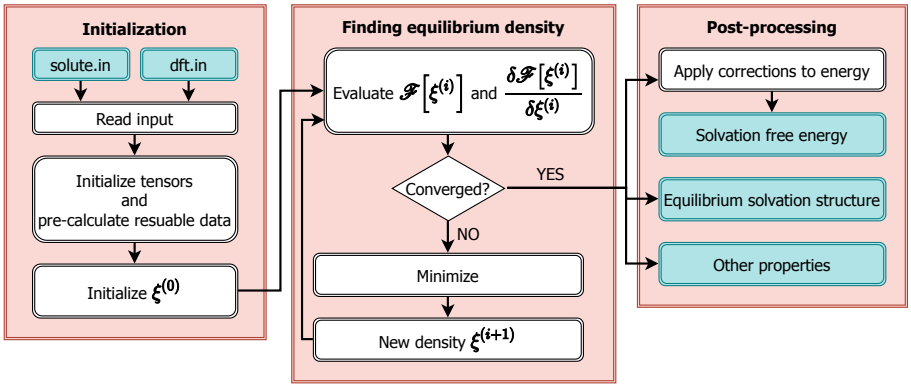


Figure 4.2: Main structure of the MDFT code.

The following four technical details are briefly discussed below :

- Supercell discretisation
- Minimiser
- Variable substitution to avoid unphysical densities
- Convolutions

Supercell discretisation: the six-dimensional tensors of $(\boldsymbol{r}, \omega)$ are discretised on a spatial and an angular grids. The Cartesian $L_x \times L_y \times L_z$ [Å^3] space is discretised on a homogeneous grid of $n_x \times n_y \times n_z$ nodes and the angular grid is discretised with the Gauss-Legendre quadrature for $\theta \in [0, \pi]$ and trapezoidal quadratures for $\phi \in [0, 2\pi]$ and $\psi \in [0, \frac{2\pi}{i}]$ with i the main symmetry axes of the solvent molecule ($i = 2$ for water). Furthermore, the angular discretization can be expressed in terms of the projections onto generalized spherical harmonics $R^m_{\mu\nu}$ [98, 99, 93]. The number of each angular dimension is linked to the order of the quadrature, $n_{\max}$. Table 4.1 shows the equivalence between the value of $n_{\max} = m_{\max}$ and the number of projections and angles for a solute with C_{2V}-symmetry like water.

	C_{2V} symmetry		no symmetry	
$n_{\max}$	$n_{\text{independent projections}}$	n_ω	$n_{\text{independent projections}}$	n_ω
1	4	6	10	6
2	14	45	35	75
3	28	84	84	196
4	55	225	165	405
5	88	330	286	726

Table 4.1: Correspondence between value of $n_{\max}$ and number of projections and the number of orientations n_ω for a molecule with C_{2V} symmetry and without one.

Minimiser: the MDFT code proposes two minimisers for the functional minimisation: (i) simple steepest-descent or (ii) limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [135, 136]

which uses the densities ρ_i and the gradients $\nabla\mathcal{F}[\rho_i]$ (see appendix A) of the past m iterations to speed up the minimisation.

Variable substitution to avoid unphysical densities: during minimisation, the density variable $\rho(\boldsymbol{r},\omega)$ can have unphysical negative values, which can also cause the divergence of the minimisation. To avoid this phenomenon, the density variable is substituted during the minimisation with the variable ξ, defined by

$$\rho(\boldsymbol{r},\omega) = \rho_{\text{bulk}}\xi^2(\boldsymbol{r},\omega) \tag{4.10}$$

which can take negative values while the density ρ always stays positive.

Convolutions: the main advantage of MDFT compared to other methods is its speed. The ideal and external terms are local, therefore their computation time scales linearly with $n_{xyz}n_\omega$ with $n_{xyz} = n_x \times n_y \times n_z$ and $n_\omega = n_\theta \times n_\phi \times n_\psi$. However, the excess term is non-local and involves the spatial and angular convolution $\gamma = c^{(2)} * \Delta\rho$. The spatial convolution can be computed efficiently with fast Fourier transformations (FFT) due to the following property of convolutions

$$f * g = \text{FT}^{-1}[\text{FT}(f)\text{FT}(g)], \tag{4.11}$$

i.e. the convolution of real space functions can be computed as the inverse Fourier transformation (FT) of a simple product of functions in the reciprocal space. In an equivalent manner, the orientational convolution is replaced by an algebric product between projections. Recently, our group proposed a similar computation of the angular convolution with the fast generalised spherical harmonic transformation (FGSHT) [40]. The coupling of these two methods decreases drastically the computation time of the excess term, which is now in the same order of magnitude as for the ideal and external terms. This was a major breakthrough which enabled the computation of the MOZ in the HNC approximation for fully molecular solvents.

4.2.3 FREE ENERGY CORRECTIONS

To obtain a final solvation free energy with MDFT-HNC two types of *a posteriori* corrections need to be added to the final value of the MDFT functional $\mathcal{F}_{\text{min}}$ obtained by variational minimization :

- standard free energy corrections briefly discussed here
- and pressure correction (PC) due to the HNC approximation discussed in detail in Chapter 8.

Similarly to most MD or MC simulations, MDFT uses periodic boundary conditions (PBC), treats the van der Waals part of the potential with a cut-off scheme and the electrostatics with Ewald scheme. Hence, usual free energy correction should be applied to MDFT results to (i) include the effect of long-range van der Waals interactions ($\Delta G_{\text{vdW-LR}}$); and for charged systems (ii) a Madelung-like correction incorporating the contribution of all the periodic images ($\Delta G_{\text{elec-B}}$) and (iii) taking into account the choice of the summing up convention of the solvent charges ($\Delta G_{\text{elec-C}}$) . The corrected MDFT solvation free energy reads,

$$\Delta G_{\text{solv}} = \mathcal{F}_{\text{min}} + \Delta G_{\text{vdW-LR}} + \Delta G_{\text{elec-B}} + \Delta G_{\text{elec-C}} + \text{PC}. \tag{4.12}$$

The long-range correction for the Lennard-Jones potential [137] reads

$$\begin{aligned}\Delta G_{\text{vdW-LR}} &= n_{\text{bulk}} \int_{r_c}^{\infty} U^{\text{LJ}}(r) 4\pi r^2 \text{d}r \\ &= \sum_{i}^{n_{solute}} \sum_{j}^{n_{solvent}} 16\pi n_{\text{bulk}} \epsilon_{ij} \sigma_{ij}^3 \left[\frac{1}{9}\left(\frac{\sigma_{ij}}{r_c}\right)^9 - \frac{1}{3}\left(\frac{\sigma_{ij}}{r_c}\right)^3\right] \\ &\approx \sum_{i}^{n_{solute}} \sum_{j}^{n_{solvent}} -\frac{16}{3}\pi n_{\text{bulk}} \epsilon_{ij} \frac{\sigma_{ij}^6}{r_c^3}\end{aligned} \tag{4.13}$$

where $U^{LJ}(r)$ is the Lennard-Jones potential, r_c the cut-off distance, n_x the number of solute or solvent sites, $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ are the geometric and arithmetic averages of the Lennard-Jones parameters between the solute and the solvent sites, according to the Lorentz-Berthelot mixing rules.

The electrostatic corrections for charged systems [138, 139] read

$$\Delta G_{\text{elec-B}} = \frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2}{L}\left[\xi + \frac{4\pi}{3}\left(\frac{R_I}{L}\right)^2 - \frac{16\pi}{45}\left(\frac{R_I}{L}\right)^5\right] \tag{4.14}$$

$$\approx \frac{\xi}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2}{L} \tag{4.15}$$

$$\Delta G_{\text{elec-C}} = -\frac{1}{6\varepsilon_0} n_{\text{bulk}} \gamma q \left[1 - \frac{4\pi}{3}\left(\frac{R_I}{L}\right)^3\right] \tag{4.16}$$

$$\approx -\frac{1}{6\varepsilon_0} n_{\text{bulk}} \gamma q \tag{4.17}$$

where ε_0 is the vacuum permittivity, ε_r the relative permittivity of the solvent, q the charge of the solute, L the box length, R_I the ion radius, $\xi \approx -2.837297$ the energy per particle in a simple cubic lattice [140] and $\gamma = \mathrm{Tr}(\mathcal{Q})$ the solvent's spheropole moment with $\mathcal{Q}$ the quadrupole moment of the solvent molecule. As R_I is significantly smaller than the size of the computational supercell, *i.e.* $R_I \ll L$, its quadratic and higher-order values of (R_I/L) are considered negligible.

4.2.4 SOLVATION STRUCTURE

Equation 4.3 states that at the same time as MDFT produces the solvation free energy of an arbitrarily complex molecule (the value of the functional at its minimum), it produces the equilibrium solvent structure around this solute (the density that minimizes the functional) in its full molecular description; the molecular (density reduced) solvent distribution function is given by $g(\boldsymbol{r},\omega) = \frac{\rho(\boldsymbol{r},\omega)}{\rho_{\text{bulk}}}$. From this full molecular distribution, one can extract more readable information. For instance, the first moment of $g(\boldsymbol{r},\omega)$ is the three-dimensional scalar field

$$g(\boldsymbol{r}) = \frac{1}{8\pi^2}\int g(\boldsymbol{r},\omega)\mathrm{d}\omega \tag{4.18}$$

from which one can derive the usual spherically symmetric radial distribution function $g_i(r)$ between solute and solvent sites or the number density $n(\boldsymbol{r}) = n_{\text{bulk}} g(\boldsymbol{r})$.

Another important quantity embedded in $g(\boldsymbol{r},\omega)$ is the polarization field

$$P(\boldsymbol{r}) = \frac{\mu n_{\text{bulk}}}{8\pi^2}\int \hat{\omega} g(\boldsymbol{r},\omega)\mathrm{d}\omega \tag{4.19}$$

where μ is the value of solvent molecule dipole moment, $\hat{\omega}$ the unitary vector along the dipole axis depending on (θ,ϕ) only. One can also obtain so-called water maps from $g(\boldsymbol{r},\omega)$, catching the most probable water molecules position and orientation around the solute.

It should be noted that the equilibrium molecular solvent density $g(\boldsymbol{r},\omega)$ is a direct output of MDFT. This information cannot be obtained with implicit solvent methods, or even with RISM, and in the case of molecular simulations, one would have to accumulate such data during a long trajectory, averaging in spatial voxels of typical size 0.1–0.5 Å for a series of orientations. For just the spatial density profile, this can be tackled nowadays, especially with the recent approach using an estimator based on forces rather than simple binning to decrease the variance of the estimate of g [141, 142, 143]. Nevertheless, accumulating data in the full six-dimensional orientation and position space is a daunting task, even more difficult than computing SFE's. MDFT produces this six-dimensional map in the same few minutes as needed to predict the SFE.

4.3 CORRESPONDENCE BETWEEN MIET AND MDFT

The direct connection between MIET, and its approximate RISM, and MDFT might seem non-existent. However, starting from the generalized MOZ of multicomponent systems,

$$h_{\mu\nu}(12) = c_{\mu\nu}(12) + \rho \sum_{\lambda} x_\lambda \int h_{\lambda\nu}(13) c_{\mu\lambda}(23) \mathrm{d}3 \tag{4.20}$$

where $x_\lambda = N_\lambda / N$ is the fraction of species $\lambda \in [1, n]$; the coupled MOZ relations for a homogeneous two-component solute-solvent mixture, where the solute M is infinitely diluted ($x_M \to 0$) in the solvent S ($x_S \to 1$), can be written as

$$h_{SS}(12) = c_{SS}(12) + \rho \int h_{SS}(13) c_{SS}(23) \mathrm{d}3 \tag{4.21}$$

$$h_{SM}(12) = c_{SM}(12) + \rho \int h_{SS}(13) c_{SM}(23) \mathrm{d}3 \tag{4.22}$$

$$h_{MS}(12) = c_{MS}(12) + \rho \int h_{MS}(13) c_{SS}(23) \mathrm{d}3 \tag{4.23}$$

$$h_{MM}(12) = c_{MM}(12) + \rho \int h_{MS}(13) c_{SM}(23) \mathrm{d}3. \tag{4.24}$$

Eq. 4.21 is the MOZ equation for bulk solvent. Eqs. 4.22 and 4.23 are equivalent and describe the correlation between the solute and the solvents molecules. Eq. 4.24 is the MOZ equation for the solute-solute correlation which is rarely used in (M)IET as HNC closure relation is insufficient for describing the solute-solute correlations.

The MDFT excess functional, eq. 4.8, can be deduced from eq. 4.23 if one imposes the HNC approximation, *i.e.* $\mathcal{F}_\mathrm{B} = 0$.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The reference of speed are emprical continuum models that lack accuracy and the exact free energy perturbation approach is very time-consuming. The liquid state approach offers a compromise between speed and accuracy.

Here, we presented two ways to study solvation by the resolution of the molecular Ornstein-Zernike integral equation by either approximating it by site-site correlations (RISM) or by resolving it rigorously at the hyper netted-chain approximation level in the density functional formalism (MDFT). They are approximate theories that give a good compromise between speed and accuracy with a main disadvantage: the solute is rigid.

To recapitulate the presentations of three families of methods to compute solvation free energies:

- Polarizable continuum models (PCM) are the reference for speed (seconds) but as they do not include any structural informations of the solvation they are prone to be inaccurate
- Free energy perturbation simulations (MD+FEP) are the reference for accuracy as they are exact in the force field approximation but their use is limited by their large computation times (tens/hundreds of hours)
- Liquid state theories (LST) are approximate theories that compute solvation free energies of rigid solutes with a compromise between speed (minutes) and accuracy.

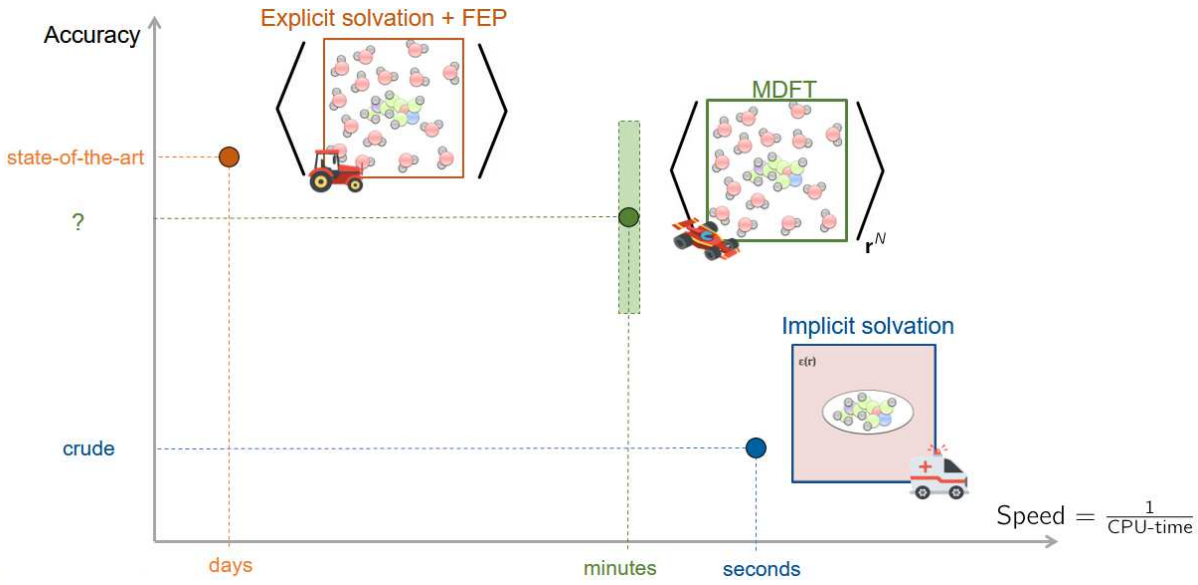


Figure 4.3: Illustration of accuracy-computation time trade of the solvation free energy methods. Creation by Anton Robert.

A main disadvantage of LST is that predict the SFE for a single conformer, *i.e.* a rigid solute. From this arises three important points:

- Is solute flexibility important when computing SFEs?
- If yes, how to take the effect of into account in LSTs?
- To correctly evaluate and develop the approximations in LSTs, one need rigid solute reference data.

These points are addressed in the following part.

Part II

HYDRATION WITH H4D-MC

This part presents how to compute solvation free energies of rigid and flexible solutes with H4D/MC.

Chapter 5 focuses on rigid single conformer solutes. It includes a parameter analysis to optimise the statistical efficiency of the hydration free energy calculation and presents the HFE results from simple (charged) spherical solutes to a set of small organic molecules of the FreeSolv database.

In chapter 6, the H4D-MC method is extended to flexible solutes. This chapter focuses on methodological developments and technical aspects for computing efficiently HFEs of flexibles solute with H4D-MC.

Chapter 7 presents a chemoinformatics analysis of the solute flexibility of small drug-like molecules. We quantify at which point solute flexibility is important in HFE calculations of the FreeSolv database and define a sub-set of 'rigid' molecules, i.e. solutes for which flexibility does not affect the HFEs that can be used as a reference database for single conformer SFE methods. Additionally, we try to identify solute features for which solute flexibility is needed.

5

RIGID SOLUTES

Why this chapter?

In order to develop single conformer SFE methods, one needs to have rigorous reference data for rigid solutes. However, most MD(+FEP) codes are written for flexible solutes and thus one needs to apply constraints in order to compute SFEs of rigid molecules. Is there a way to compute exact SFEs of rigid solutes more efficiently?

In this chapter, we show how to compute efficiently HFEs of simple neutral and charged spherical solutes and small rigid molecules with the H4D-MC approach. The first part of the chapter includes a parameter analysis of the H4D-MC method and comparison of computation times to the classic MD+FEP approach, before showing the results for the spherical solutes and small organic molecules.

As mentioned in chapter 4.2, MDFT calculations are done with a single conformer rigid solute, and we need accurate and precise rigid solute single conformer reference data to correctly evaluate and develop the MDFT approach. To produce this reference data we developed and used an original approach: the 'hybrid 4$^{\text{th}}$ dimension Monte Carlo' (H4D-MC) [36] and its in-house code. We originally choose to use H4D-MC as it was written for rigid molecules whereas most of the well-known commercial or open-source MD codes are written for flexible solutes and needs to use constraints to obtain a rigid solute. But as shown in the following chapters, we discovered H4D-MC to be a very efficient alternative to classic MD+FEP simulations when computing HFE of rigid and flexible solutes.

The principle of using H4D-MC approach for solvation free energy calculations was presented in sec. 3.2.2, and in algorithms 3.3 and 3.4. Briefly, as a reminder, (i) it propagates two equilibrium simulation with MC, one of bulk solvent and one with the solvated solute, (ii) periodically introduces or removes 'slowly' the solute from the simulation box with a short non-equilibrium MD simulation in a 4$^{\text{th}}$dimension where the 'distance' to the simulation box is imposed by time-dependent coupling parameter $w(t)$, and (iii) computes the excess chemical potential of the solute, *i.e.* the solvation free energy, work needed to insert or destruct the solute with Jarzynski principle.

This thesis project was the first time that the H4D-MC approach was used to systematically compute solvation free energies as it was originally written for grand canonical μVT simulation. Therefore, the first section of this chapter is a systematic analysis of the H4D insertion/destruction parameters. The second and the third sections, present HFE calculation results for *(i)* neutral and charged spherical solutes, and *(ii)* molecular solutes, respectively.

5.1 ANALYSIS OF INSERTION/DESTRUCTION PARAMETERS

A systematic analysis of MD simulation parameters of the H4D insertion/destruction (ins/des) process was done to determine optimal values to minimize the statistical error with the minimal simulation time. This analysis was done on the amitriptyline molecule (FreeSolvID: 5282042 [144], sec. C for more information of the FreeSolv database) represented in fig. 5.1. It is the largest solute,

by site number (44), of the FreeSolv database with a partial molar volume $V_{\mathrm{PM}} \sim 335$ Å^3 (this value is evaluated from a fast $n_{\max} = 1$ MDFT calculation, see eq. 8.3) and largest intramolecular distance $d_{\max}$ equal to 11.7 Å.

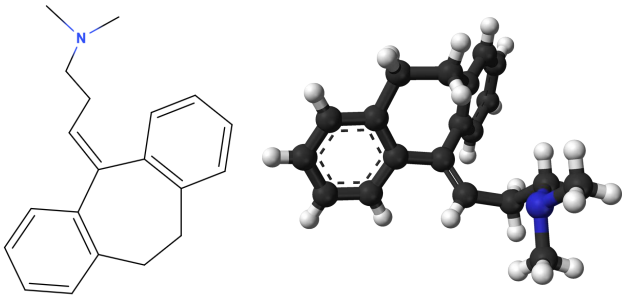


Figure 5.1: Amitriptyline 2D- and 3D-structure.

The five simulation parameters were :

- simulation box size with four different number of the solvent molecules: $N = 50, \mathbf{100}, 200$ and 400, which correspond to $L \approx 11.5, 14.5, 18.2$ and 23.00 Å and $V \approx 1500, 3000, 6000$ and 12000 Å^3 for the bulk solvent
- relaxing the simulation box volume with the solute's partial molar volume during insertion/destruction or not (*)
- maximal "altitude" in the 4^{th} dimension for the solute: $w_{\max} = 2, \mathbf{3}$ or 5 Å
- insertion/destruction speed: $v = 0.1, \mathbf{0.05}$ or $0.10\ \sqrt{\frac{kT}{M}}$, *i.e.* $3.7 \times 10^{-5}, \mathbf{1.9 \times 10^{-4}}$ or 3.7×10^{-4} Å/fs
- MD time step of ins/des process: $\Delta t = 0.01, \mathbf{0.02}$ or $0.04\sqrt{\frac{M}{kT}}\mathring{A}$, *i.e.* $2.7, \mathbf{5.4}$ or 10.8 fs

(*) The propagation is done with a NPT MC simulation but in the basic case the out-of-equilibrium MD trajectory is done without a thermostat and with a fixed volume. An additional option changes the volume of the box with $V \pm V_0 \frac{t}{t_{max}}$ during the insertion $(+)$ and the destruction $(-)$ to help the solvent relaxation. V_0 is a arbitrarily volume chosen at the beginning. Closer it is to the partial molar volume of solute better the statistics will be. Note that this changes the definition of the work in the Jarzynski equation (eq. 3.14) from ΔH to $\Delta H + PV_0 + N \ln\left(\frac{V+V_0}{V}\right)$ with P the system pressure.

Table 5.1 recapitulates the evolution of the amitriptyline's HFE, its error bars, and the computation time as a function of these simulation parameters. Figures 5.2 and 5.3 plot the evolution of the HFE and the evolution of the insertion and destruction distributions $p_{\mathrm{ins}}(\Delta H)$ and $p_{\mathrm{des}}(\Delta H)$, respectively. Narrower and more gaussian-like the distribution and larger the overlap between them is, better the statistics will be.

First of all, as the H4D-MC uses Ewald method for both the electrostatics and the Lennard-Jones (LJ) part, the simulation supercell can be smaller than in most MD and MC codes where the minimum box size is $L \geq 2r_{\mathrm{LJ-cutoff}} \gtrsim 20$ Å with $r_{\mathrm{LJ-cutoff}}$ is LJ cut-off radius, typically ~ 10 Å. Also, most codes use tin-foil conditions for the electrostatic Ewald summation, *i.e.* the relative permittivity of the imaginary surface ε' around the infinity system created by the periodic supercells is set to ∞ , which sets the dipole correction of the Ewald summation to zero. In H4D-MC this quantity ε' set to the closest possible to the systems dielectric constant, *eg.* $\varepsilon' = 99$ the dielectric constant of TIP3P bulk water. This improves the quality of the Ewald summation even at small cut-off distances at least for neutral solutes.

Nonetheless, the insertion of a "large" solute like amitriptyline, with a V_{PM} that is over 10% of the initial bulk box of 100 water molecules, the destruction and especially the insertion statistics with a

N	$w_{\max}$ $[\AA]$	v $\left[\sqrt{\frac{kT}{M}}\right]$	Δt $\left[\sqrt{\frac{M}{kT}}\AA\right]$	V_0 $[\AA^3]$	ΔG_{100}	ΔG_{1000}	ΔG_{3000}	ΔG_{5000}	$t_{100\,\mathrm{MC}+1\mathrm{MD}}$ [CPU.s]
100	**3**	**0.05**	**0.02**	**335**	$\mathbf{-10.21 \pm 0.88}$	$\mathbf{-8.13 \pm 0.32}$	$\mathbf{-8.13 \pm 0.21}$	$\mathbf{-8.00 \pm 0.17}$	**2+20=22**
100	3	0.05	0.02	**0**	-5.16 ± 10^3	-8.27 ± 16	-7.23 ± 5.93	-6.93 ± 4.44	2+20=22
100	**5**	0.05	0.02	335	-7.59 ± 0.66	-8.06 ± 0.30	-8.23 ± 0.18	-8.34 ± 0.14	2+30=32
100	3	**0.01**	0.02	335	-7.98 ± 0.25	-8.10 ± 0.08	-8.19 ± 0.05	-8.15 ± 0.04	2+73=75
100	3	**0.10**	0.02	335	-7.91 ± 40	-10.70 ± 1.05	-8.07 ± 0.53	-8.36 ± 0.43	2+6=8
100	3	0.05	**0.01**	335	-7.64 ± 2.07	-8.07 ± 0.44	-8.37 ± 0.22	-8.24 ± 0.18	2+37=39
100	3	0.05	**0.04**	335	-8.85 ± 10^6	-8.27 ± 10^4	-6.38 ± 10^3	-6.33 ± 10^3	2+9=11
50	3	0.05	0.02	335	-9.25 ± 4.99	-7.40 ± 0.53	-7.34 ± 0.31	-7.50 ± 0.22	1+10=11
200	3	0.05	0.02	335	-8.96 ± 1.85	-8.06 ± 0.36	-8.10 ± 0.18	-8.13 ± 0.17	5+46=51
400	3	0.05	0.02	335	-10.59 ± 0.79	-8.74 ± 0.50	-7.98 ± 0.29	-7.75 ± 0.20	15+112=127

Table 5.1: Amitriptyline: evolution of H4D-MC hydration free energy and computation times as a function of the simulation parameters and number of accumulations. ΔG_x corresponds to the HFE obtained after x iterations of 100 MC cycles of simulation box propagation and one MD insertion/destruction process. Hydration free energies in kcal/mol and CPU-times as a sum of 100 propagation MC cycles and 1 MD insertion/destruction.

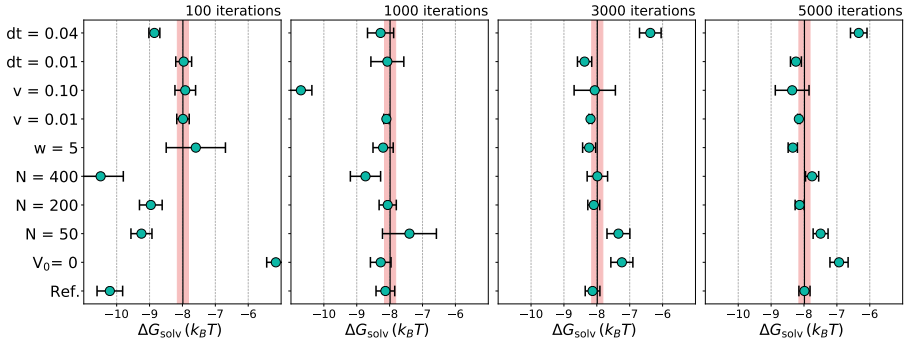


Figure 5.2: Evolution of the HFE of amitriptyline as a function of simulation parameters and the number of accumulations.

very wide $p_{\text{ins}}(\Delta H)$ and little to none overlap between the insertion and destruction distributions as it can be seen in fig. 5.3b. This leads to very large statistical errors of the HFE even for long accumulations as the solvent molecules do not have the place to correctly relax in fixed volume. This problem was overcome by the addition of a new feature, that imposes the increase/decrease (linear in time) of the supercell volume by V_0, a value close to the solute volume, during the ins/des process, thus helping the simulation box relaxation and improving drastically the statistics (see figs. 5.2 and 5.3a.).

Three other simulation box sizes, $N = 50, 200$ and 400, were tested to verify that there are no finite-size effects. As can be seen in figs. 5.3a,d,e and f, all four box sizes give similar $p_{\text{ins}}(\Delta H)$ and $p_{\text{des}}(\Delta H)$ profiles. In fig. 5.2 it can be seen that HFEs for simulation boxes with $N \geq 100$ converge to the same value, hence confirming that a small box of 100 water molecules is large enough for a neutral molecule of 44 atomic sites. This is a great advantage of the method as the computation time increases with N^2. As a comparison, the FreeSolv MD+FEP calculation of the amitriptyline was done in a supercell of 2366 water molecules ($L \approx 41$ Å).

Secondly, the starting/ending 'altitude' in the 4th dimension during the ins/des process, should be large enough so that there is no important solute-solvent overlap at $w_{\max}$, at least larger than $\max\{\sigma_i/2\}$ with σ_i LJ diameter of the site i ($\max\{\sigma_i/2\} = 1.7$ Å for amitriptyline). On the other hand, large values of $w_{\max}$ increase the computation time, as it increases linearly with $w_{\max}$, for little to none improvement in the statistics. Three values were tested. The overlap between solute and solvent molecules were too big for the smallest one, $w_{\max} = 2$ Å, leading to unstable simulations. When comparing results for $w_{\max} = 3$ and 5 Å, we can see a slightly better overlap between $p_{\text{ins}}(\Delta H)$ and $p_{\text{des}}(\Delta H)$ for 5 Å than 3 Å (figs. 5.3a and c), leading to slightly better statistics (fig. 5.2 and table 5.1). However, the improvement is so slight that it does warrant the increase of computation time by a factor 1.6.

The third parameter to be determined was the insertion speed v, with three values at $0.01, 0.05$ and 0.10 in the code's intrinsic units of $\sqrt{k_{\text{B}}T/M}$ with $M = 18$ g/mol the molar mass of water molecules. These values correspond to $v = 0.37, 1.9$ and 3.7×10^{-4} Å/fs, in more classical units. This parameter seems to have the most important effect on the statistics. In figs. 5.3a,g and h, we see that the width of the distributions decreases and the overlap increases significantly, with the decrease of the ins/des speed, and thus leads to better statistics as seen on the error bars in fig. 5.2. The very slow speed of ins/des, $v = 0.01$, gives clearly the best statistics. However, this comes with a great computational cost as there is a factor five between $v = 0.01$ and 0.05 $\sqrt{k_{\text{B}}T/M}$. Hence we chose to use $v = 0.05$ as a compromise between precision and speed, as $v = 0.10$ was too fast producing no overlap at all between the distributions and thus producing bad statistics. It should be noted that the amitriptyline was one of the few solutes for which good statistics were hard to obtain. The average statistical error for the FreeSolv database, with $v = 0.05\sqrt{k_{\text{B}}T/M}$ and 3 000 accumulations was 0.03 kcal/mol. We also preferred to use $v = 0.05$ to be consistent with the

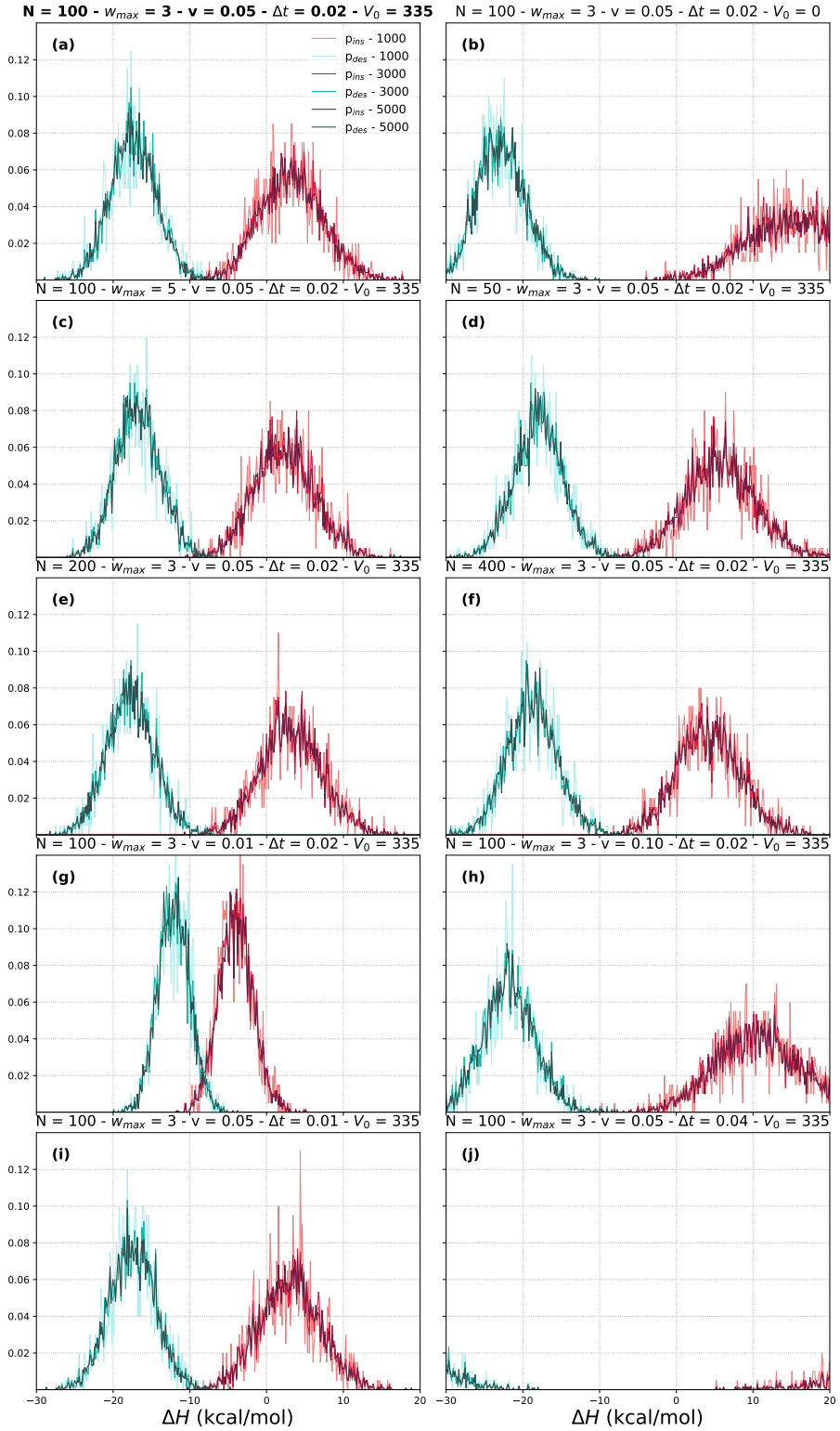


Figure 5.3: Evolution of the $p_{\text{ins}}(\Delta H)$ (red) and $p_{\text{des}}(\Delta H)$ (turquoise) as a function of simulation parameters and the number of accumulations.

flexible solute calculations of the following chapter, where a larger number of accumulations is favoured to better sample the solute conformers.

For the MD time step, the two smallest values $\Delta t = 0.01$ and $0.02\ \sqrt{M/k_{\mathrm{B}}T}\mathring{A}$ ($\Delta t = 2.7$ and 5.4 fs) give similar distributions (figs. 5.3a and i) and thus similar statistics (fig. 5.2). Note that both values are already larger than the typical MD time step of $\Delta t = 2$ fs $= 0.007\sqrt{M/k_{\mathrm{B}}T}\mathring{A}$. Using larger time step, $0.04\ \sqrt{M/k_{\mathrm{B}}T}\mathring{A}$ ($= 10.8$ fs), gives the widest and most separated $p_{\mathrm{ins}}(\Delta H)$ and $p_{\mathrm{des}}(\Delta H)$ distributions, leading to huge error bars. Hence, we chose $\Delta t = 0.02\sqrt{M/k_{\mathrm{B}}T}\mathring{A}$ as default time step for optimal statistics and minimal computation time.

Note that all the simulations converge, with 3000 sampling points, to the same value with exception of simulations with $V_0 = 0$ and $\Delta t = 0.04\sqrt{M/k_{\mathrm{B}}T}\mathring{A}$ for which the error bars are too large to conclude anything, and for the smallest simulation box, $N = 50$, for which the bulk box length was smaller than the largest intramolecular distance of amitriptyline, so problems should be expected. Even in those cases, the predicted HFEs are within 2 kcal/mol of the other predictions.

All following H4D-MC simulations are done with the reference parameters of $N = 100$, $w_{\max} = 3$ Å, $v = 0.05\ \sqrt{k_{\mathrm{B}}T/M}$, $\Delta t = \sqrt{M/k_{\mathrm{B}}T}$Å and $V_0 \sim V_{\mathrm{PMV}}$ for molecular solutes and $V_0 = 0$ for small spherical solutes. There is a small improvement of the error bars when increasing the accumulation from $2 \times 3\,000$ to $2 \times 5\,000$ but does not warrant the increase of computation time from $\sim$ 36h to $\sim$ 53h. Moreover, as mentioned before, the mean statistical error with $3\,000$ accumulations is 0.03 kcal/mol for the FreeSolv database. Full simulation details are given in appendix D.

5.1.1 COMPUTATION TIME

With these reference parameters and 3 000 accumulations, the computation of an SFE takes ~ 18 hours (wall-clock time) if the insertion and the destruction simulations are done in parallel on a single core each (~ 36 CPU.h). The FreeSolv's MD+FEP calculation done with Gromacs (v.2018.3) [145, 146, 147, 148, 149] of the amitriptyline was done in a simulation box of length 41 Å. It leads to a wall-clock time of ~ 11 hours (~ 220 CPU.h) if the $K = 20$ states are simulated in parallel, 5 ns propagation per state, on a single core each. This leads to a speed-up of $\times 6$ when comparing H4D/MC and MD+FEP CPU-computation times. Note that the MD+FEP calculation was done with a flexible solute, for a rigid solute the sampling could be potentially decreased thus reducing the computation time.

Moreover, contrary to molecular dynamics based MD+FEP calculations, H4D-MC can be, in principle, infinitely parallelised. For the Gromacs MD+FEP calculation, each λ can be run as a separate simulation and the simulation box of $L = 41$ Å can be efficiently parallelised of 16 cores. Thus a single MD+FEP SFE calculation can be parallelised efficiently on a maximum of 320 cores, leading to a wall-clock time of $\sim$ 1h45 (~ 545 CPU.h). On a typical laboratory cluster of 64 cores, one can run four separates λs in parallel with 16 cores for each simulation leading to a wall-clock time of $\sim$ 8h30.

Currently, the H4D-MC code is not parallelised and for each SFE calculation, it uses two cores (one for the insertions and another for the destructions). However, each ins/des out-of-equilibrium simulation is independent and could be performed on a separate core. Moreover, as the propagation is done with MC and we are not interested in dynamical quantities, like correlation times, even the propagation could be done with several parallel simulations of the bulk solvent and the solvated system. In fig. 5.4, we show the evolution of the wall time as a function of the total number used and the number of cores used for propagation. The figure is plotted for the reference parameters defined in the previous section for $2 \times 3\,000$ accumulations and takes into account the 3 minutes used for initialisations and equilibration (10 000 MC cycles).

On a large cluster of 1024 cores, without any parallelisation of the propagation the wall-clock time is $\sim$ 1h45, and if one parallelises the propagation to 2×32 cores the wall-clock time should

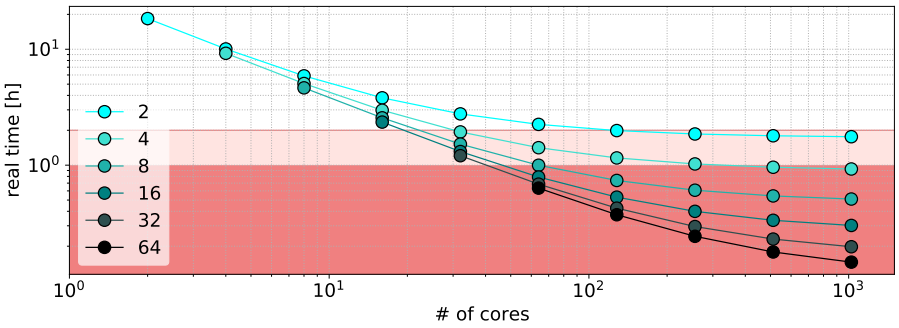


Figure 5.4: Evolution of H4D-MC wall-clock time as a function of the total number of cores. Each line corresponds to a different number of cores used for propagation. The light and dark rose indicates wall times under 1 and 2 hours respectively.

be only ~ 10 minutes. On a typical laboratory cluster of 64 cores, the wall-clock time should be $\sim$ 2h15 and ~ 40 minutes without and with the parallelisation of the propagation on 2×32 cores, respectively. This leads to a speedup of ~ 4 and ~ 13, without and with the parallelisation of the propagation, of the wall-clock time when computing SFEs with H4D-MC compared to computing them with MD+FEP on a typical laboratory cluster.

All computation times were estimated on a cluster of *Intel Xeon CPU E5-2690 v4 @ 2.60GHz* CPUs.

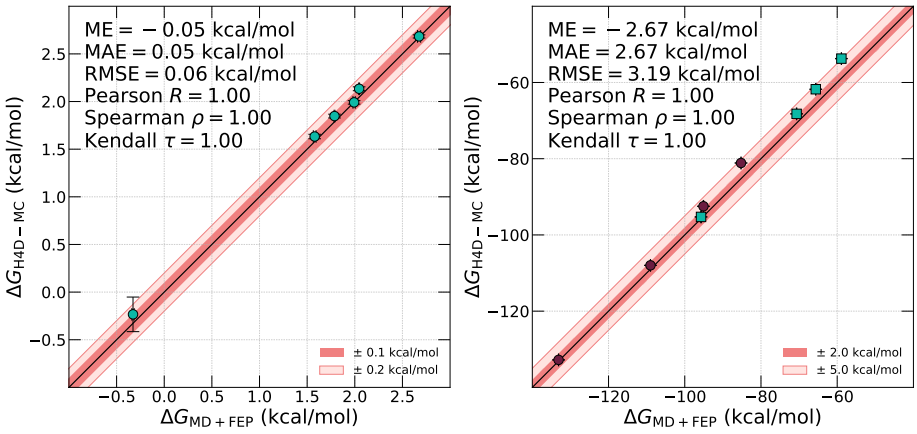
5.2 SPHERICAL SOLUTES

5.2.1 HYDROPHOBIC SOLUTES

First, we H4D-MC computed the HFEs of simple hydrophobic spherical solutes, noble gases and unified-atom methane and neopentane, to test the H4D-MC code by comparing them to MD+FEP calculation done with Gromacs. The H4D-MC calculations were done with the parameters defined in the previous section and the Gromacs calculations with the parameter files given by FreeSolv with 884 solute molecules ($L \approx 30$ Å). V_0 was set to 0 for all solutes except the neopentane for which it was set to 120 Å. The hydration free energies are computed in TIP3P water [28]. Table 5.2 recapitulates the force field parameters and calculated HFEs of the hydrophobic solutes. As the fig. 5.5a shows, H4D-MC reproduces well the HFEs obtained with the classic MD+FEP approach.

			Hydration free energy [kcal/mol]		
Solute	σ [Å]	ϵ [kJ/mol]	ΔG_{Exp} [150]	$\Delta G_{\text{MD+FEP}}$	$\Delta G_{\text{H4D-MC}}$
Neon	3.035	0.15432	2.48	2.68 ± 0.02	2.66 ± 0.04
Argon	3.415	1.03931	1.99	1.99 ± 0.01	2.05 ± 0.04
Krypton	3.675	1.40510	1.66	1.79 ± 0.01	1.81 ± 0.04
Xenon	3.975	1.78510	1.45	1.59 ± 0.01	1.62 ± 0.05
Methane	3.730	1.23000		2.04 ± 0.01	2.14 ± 0.04
Neopentane	6.150	3.49000		-0.33 ± 0.01	-0.22 ± 0.09

Table 5.2: Lennard-Jones force field parameters and hydration free energies of rare gases and unified methane and neopentane molecules.



(a) (b)

Figure 5.5: Correlation between HFEs obtained with H4D-MC and MD+FEP of (a) hydrophobic spheres and (b) monovalent ions (cations in purple and anions in turquoise).

5.2.2 MONOVALENT IONS

Next, we computed the HFEs for a series of spherical monovalent ions with H4D-MC. The force field parameters of these ions are given in table 5.3 [151]. Reference calculations were done with Gromacs with the parameter files given by FreeSolv with 2164 solute molecules. To study finite-size effects of charged systems we performed the HFE calculations for the sodium and chloride ions with four different box sizes: $N = 50, 100, 256$ and 400 ($L \approx 11.5, 14.5, 20$ and 23 Å). As we simulate charged systems with periodic boundary conditions, we should apply Hünenberger's finite-size corrections to the results [139, 138] (cf. 4.2.3). Note that, in the Ewald implementation of H4D-MC, the self-energy correction of the first term of eq. 4.14 ($q^2\xi/8\pi\varepsilon_0$) is already accounted for. In fig. 5.6, we plot these 'brute' HFEs, with the self-correction, and with the rest of the 'type B' correction with ionic radii that vary from 0 Å to 3.0 Å. Firstly, the brute results seem to converge for relatively small boxes, $N = 256$, as the deviation to the largest box are below 0.1 kcal/mol and even with only 100 water molecules, HFEs are close to the converged value.

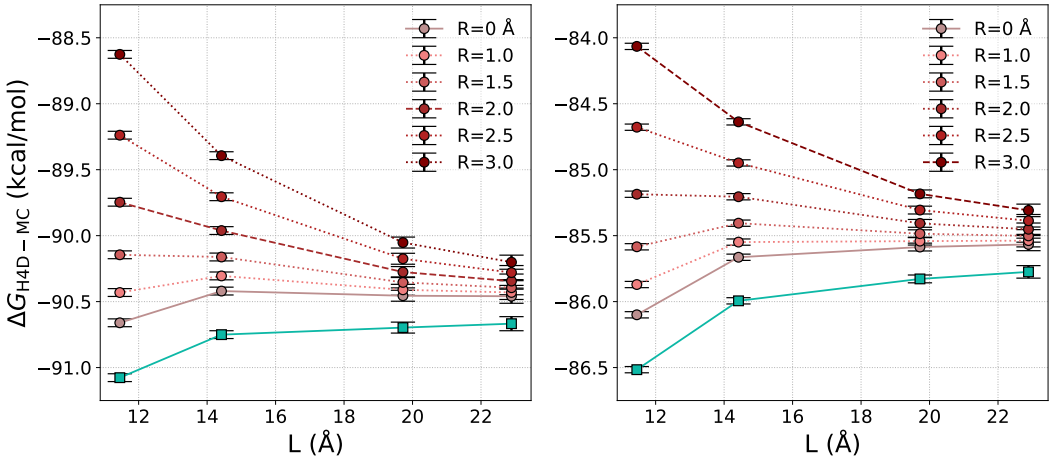


Figure 5.6: Evolution of H4D-MC HFE of sodium (left) and chloride (right) ions as a function of simulation box length. Turquoise data correspond to the 'brute' results and other to 'typeB' corrected HFEs with different ionic radius.

Contrary to what Hünenberger *et al.* claim, we found large variations to the HFE depending on the ionic radii used in the 'type B' correction. Moreover, the ionic radius that produces a plateau and therefore corrects for the finite-size effects, $R \approx 1.5$ Å for both ions, does not correspond to the usual way to define an ionic radius from the first peak of the radial distribution function, $R_{\mathrm{Na}} \sim 2$ Å and $R_{\mathrm{Cl}} \sim 3$ Å. For both cases, if we use the ionic radius defined from the g_{ion-O} (dashed line in fig. 5.6) the quality of size convergence decreases significantly. As we are not convinced by the pertinence of radius dependant terms in the 'type B' correction and our brute results seem to converge for relatively small box sizes, we decide not to apply the radius dependant 'type B' correction terms to our future results. We found, similar problems with the radius dependence of 'type C' correction, *i.e.* convergence improved for radii close to 1 Å instead of the 2 or 3 Å expected for which the convergence is degraded compared to brute results. Thus, only the non-radius-dependent part of the 'type C' correction is applied, $-17.8q$ kcal/mol for TIP3P water.

Figure 5.5b plots the correlation between HFEs obtained with the MD+FEP and H4D-MC approach, and table summarises 5.3 the HFEs for the full series monovalent ions. The H4D-MC calculations were done with $N = 400$ to be sure.

Ion	σ [Å]	ϵ [kJ/mol]	R [Å]	$\Delta G_{\mathrm{MD+FEP}}$	$\Delta G_{\mathrm{H4D-MC}}$
F^-	3.434	4.654×10^{-1}	2.7	-95.9 ± 0.1	-95.6 ± 0.1
Cl^-	4.394	4.160×10^{-1}	3.1	-70.8 ± 0.1	-68.6 ± 0.1
Br^-	4.834	2.106×10^{-1}	3.2	-65.7 ± 0.1	-62.2 ± 0.1
I^-	5.334	1.575×10^{-1}	3.4	-59.0 ± 0.1	-54.0 ± 0.1
Li^+	2.874	6.154×10^{-4}	1.9	-133.3 ± 0.2	-133.2 ± 0.2
Na^+	3.874	*idem*	2.2	-109.0 ± 0.1	-108.6 ± 0.1
K^+	4.543	*idem*	2.4	-95.2 ± 0.1	-92.8 ± 0.1
Cs^+	5.173	*idem*	2.6	-85.4 ± 0.1	-81.6 ± 0.1

Table 5.3: Force field parameters and hydration free energies (in kcal/mol) of monovalent ions.

5.2.3 $\{q, \sigma, \epsilon\}-$SPHERES

In the previous two sections, we tested and confirmed the functioning of H4D-MC to compute SFE with an exact comparison of H4D-MC results to classical MD+FEP results. These results will be used as reference data for an MDFT-HNC benchmark in chapter 9. This analysis will show that MDFT in the HNC approximation somewhat struggles to predict the HFEs of anions especially and does not capture correctly the solvation profile around hydrophobic and hydrophilic solutes. To correct these, the next step will be the development of a solute-independent bridge functional. A first step in the goal of developing a bridge functional is to compute the bridge function which is defined from

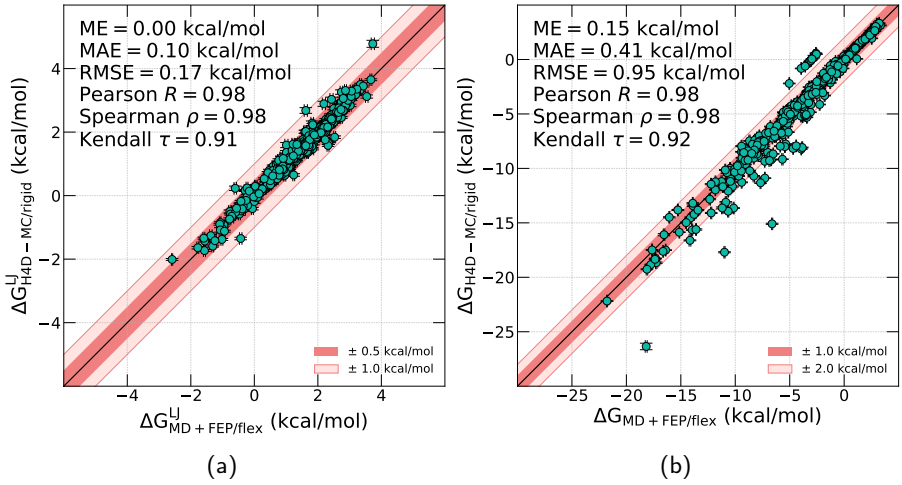
$$g_{\mathrm{sim}} = e^{-\beta U + h_{\mathrm{sim}} * c + b} \tag{5.1}$$

In this aim, we computed reference HFEs and molecular solvation profiles for a large range of spherical solute with σ varying from 1.0 to 4.0 Å with steps of 0.5 Å, ϵ varying from 0.1 to 2.1 kJ/mol with steps of 0.2 kJ/mol and charges varying from -1.0 to +1.0 e with steps of 0.2 e. The σ and ϵ ranges correspond to values present in the GAFF force field. The HFEs of these $\{q, \sigma, \epsilon\}-$triplets are given a file at github.com/sohviluukkonen/Thesis.

5.3 MOLECULAR SOLUTES - FREESOLV

The FreeSolv database, a widely used reference database produced by Mobley et co-workers [152, 144], contains HFEs obtained by experiments and state-of-the-art MD+FEP calculations for 642 small neutral organic molecules. However, as mentioned before, MDFT calculations are done with a single conformer rigid solute, and we need accurate and precise rigid solute single conformer reference data to correctly evaluate and develop the MDFT approach. Therefore we recomputed the HFEs of the FreeSolv database with rigid solute, with and without partial charges, for the initial configuration given in ref. [144]. Same non-bonded force fields parameters, GAFF (v1.7) [153] with AM1-BCC [154, 155] partial charges for the solutes and TIP3P [28] for the water.

Figures 5.7a and 5.7b show the correlations between HFEs obtained with rigid solute H4D-MC and flexible solute MD+FEP methods as given in ref. [144] for the whole FreeSolv database without and with partial charges. For most of the solutes, flexibility is not a very important factor in the HFE calculation as the correlation between the rigid and flexible solute calculations are high with small discrepancies.



(a) (b)

Figure 5.7: Correlations between HFEs obtained with rigid solute H4D-MC and flexible solute MD+FEP calculations for the FreeSolv database (a) without and (b) with partial charges.

For the hydrophobic solutes (fig. 5.7a) , the deviations between rigid and flexible results are especially with an MAE of 0.1 kcal/mol, no deviations above 1 kcal/mol, and 66% of the solutes (421) with deviations below 0.1 kcal/mol. The criterion of 0.1 kcal/mol was chosen as (i) it corresponds to the statistical error of these methods and (ii) we want to develop a theory with a precision of 0.1 kcal/mol. For most of the solutes with partial charges solute (fig. 5.7b), flexibility is not important as 46% of the solutes (295) have deviations lower than 0.1 kcal/mol. These molecules will be considered completely rigid concerning the HFE, *i.e.* solute flexibility does no affect at all on HFE prediction. However, the MAE of 0.41 kcal/mol, even if it is smaller than the experimental precision of ~ 0.5 kcal, is significantly higher than the statistical errors of the rigid H4D-MC and flexible MD+FEP methods. The effect of solute flexibility and how to treat in H4D-MC are discussed in details in the following chapters.

To remember

In order to develop single conformer SFE methods, one needs to have rigorous reference data for rigid solutes. However, most MD(+FEP) codes are written for flexible solutes and thus one needs to apply constraints in order to compute SFEs of rigid molecules. Is there a way to compute exact SFEs of rigid solutes more efficiently?

Here, We showed that the HFEs of spherical model solutes and small rigid molecules can be rigorously calculated with the H4D-MC approach leading to a speed-up of 6 times when compared to the classic MD+FEP approach. As a result, we now have produced rigorous reference data for single conformer SFE approaches. In the next chapter, we implement solute flexibility to H4D-MC to verify that we recover the flexible solute results obtained with MD+FEP.

6

SOLUTE FLEXIBILITY IN H4D-MC

Why this chapter?

There are some divergences between rigid solute H4D-MC and flexible solute MD+FEP hydration free energies. Are these deviations due to solute flexibility or due to problems in H4D-MC?

In this chapter, we extend the H4D-MC approach to flexible solutes. It presents the theoretical development made to H4D-MC to enable the use of flexible solutes and the tricks to compute solvation free energies of flexible solutes efficiently.

The H4D-MC theory and code were originally written for grand canonical μVT simulations with a fluctuating number of rigid particles/molecules [36]. Then it was modified to measure the excess chemical potential due to the insertion/destruction of a rigid single conformer solute. During this thesis, the theory and code were extended to more realistic, flexible solutes. The first section of this chapter introduces the theoretical methodology to perform flexible solute simulations with H4D-MC, the second part presents some techniques to optimise the statistics and the last section presents the comparison of the two methods presented in the first section.

6.1 SOLUTE FLEXIBILITY IN H4D-MC

In the most general case, the total energy (potential and kinetic) difference between the final (solute in solvent) and initial (bulk solvent + solute in vacuum) states of a solvation free energy process reads

$$\begin{aligned}\Delta H &= H(w=0) - H(w=\infty) \\ &= H_N(0) - H_N(\infty) + H_M(0) - H_M(\infty) + v_{cross}(0) - v_{cross}(\infty) \\ &= H_N(0) - H_N(\infty) + H_M(0) - H_M(\infty) + v_{cross}\end{aligned} \tag{6.1}$$

where $H_N(\infty)$ and $H_N(0)$ are the internal energies of the system when the solute does not interact with the solvent ($w=\infty$) and is fully interacting with the solvent ($w=0$). These energies can be decomposed into three terms : the total energy of the N solvent molecules and their kinetic energy, the total energy of the solute molecule M and the cross interaction between the N solvent molecules and the solute M. By definition, the cross term at infinite limit is null, $v_{cross}(\infty) = 0$.

In the case of the simple Widom test particle method [68], the passage from ∞ to 0 is instantaneous, *i.e.* the solute or the solvent does not have time to relax ($H_N(\infty) = H_N(0)$ and $H_M(\infty) = H(0)$), thus $\Delta H_{\text{Widom}} = v_{cross}$. In the case of H4D-MC with a rigid solute considered up to now, the solvent relaxation is permitted during the insertion/destruction, but the solute is kept rigid during the ins/des process (during propagation the solute may be flexible) yields $\Delta H_{\text{rigid}} = H_N(0) - H_N(\infty) + v_{cross}$. In the case where the solute can relax during the insertion/destruction process the energy difference ΔH_{flex} is given by equation 6.1.

There are two ways to take solute flexibility into account during an H4D-MC hydration free energy calculation:

1. as a combination of multiple single conformer calculations, *i.e.* the solute conformers are propagated in vacuum and solvent during the MC cycles but kept rigid during the ins/des process (ΔH_{rigid})
2. as a fully flexible simulation, *i.e.* the solute conformers are propagated in vacuum and solvent during the MC cycles and subject to relaxation during the ins/des process as well (ΔH_{flex}).

Both approaches should give the same results but we hope that the latter approach would improve the statistics compared to the former one.

The solute internal energy is defined with the same bonded and non-bonded force field parameters and equations as in any classical MD or MC simulations: Lennard-Jones and Coulombic potential for the non-bonded part, with the exclusion of "1-4" interactions, *i.e.* interactions between solute sites separated by less than 3 bonds are not accounted for; and the bonds, angles and dihedrals (proper and improper) are defined with the following harmonic potentials

$$\begin{aligned} U_{\text{bond}}(r_{ij}) &= k_r(r - r_{\text{eq}})^2 \\ U_{\text{angles}}(\theta_{ijk}) &= k_\theta(\theta - \theta_{\text{eq}})^2 \\ U_{\text{dihedrals}}(\phi_{ijkl}) &= k_\phi(1 + \cos(n\phi - \phi_s))^2 \end{aligned} \tag{6.2}$$

where k_x is the force constant of harmonic potential, r_{eq} and θ_{eq} the equilibrium bond distance and angle, and n and ϕ_s the multiplicity and phase of the dihedral. In principle, this flexible molecule approach could be applied to the solvent too, but as we study hydration and most of the water force fields are for a rigid molecule there is no need (for now) to do it.

6.2 SIMULATION PARAMETERS RELATED TO FLEXIBILITY

For flexible solutes, the MC propagation works the same as before except that, now, the shifting of solute sites are also considered, with force-bias as before, with a maximum displacement Δr_{max}. As the intramolecular forces vary faster than the intermolecular one, one needs to choose a smaller value for the solute site displacement $\Delta r_{\text{max}}^{\text{solute sites}}$ than for solvent molecule translation $\Delta r_{\text{max}}^{\text{H}_2\text{O}} = 0.3$ Å. We found that $\Delta r_{\text{max}}^{\text{solute sites}} = 0.1$ Å leads to acceptation probabilities close to $\sim 40\%$ and was chosen as the default value. Setting $\Delta r_{\text{max}}^{\text{solute sites}}$ to 0, leads to a rigid solute.

By default, the probability to move a solvent molecule or a solute site is equivalent. However, for some solutes, we observe very long propagation times to sample some rare conformers and conformers with high energy barriers. For example, the diflunisal molecule (FreeSolvID: 6055410, the molecule with the largest deviation between a rigid H4D-MC and flexible MD+FEP calculation) have two main conformers: one with an intramolecular H-bond and one without. They are illustrated in figs. 6.1a and 6.1b. Figure 6.1c shows the internal energy of a flexible diflunisal molecule in water as a function of MC cycles for a total of 10^6 MC cycles. As it can be seen, even for a relative long simulation time, the conformers with an intramolecular H-bond are almost exclusively sampled even though one could expect the conformer without the intramolecular H-bond to be favourable in water as the H-bonds between the solute and the solvent stabilise the solvation.

To improve the conformational sampling in solution, an option for preferential displacement $p_{\text{H}_2\text{O/solute-site}}$ was implemented, *i.e.* the possibility to change the relative probabilities to move a solvent molecule or a solutes site. Figure 6.1d shows that the sampling of solute conformers in solution is drastically improved by increasing the probability to move a solute site by five compared to a solvent molecule ($p_{\text{H}_2\text{O/solute-site}} = 1/5$). Both states are sampled with slightly more weight on the states with an intramolecular H-bond, for no increase in computation time. The value $p_{\text{H}_2\text{O/solute-site}} = 1/5$ is set to be the default value for flexible solute simulations as a compromise of sampling flexible solute conformers but also letting the solvent structure propagate between solute destructions.

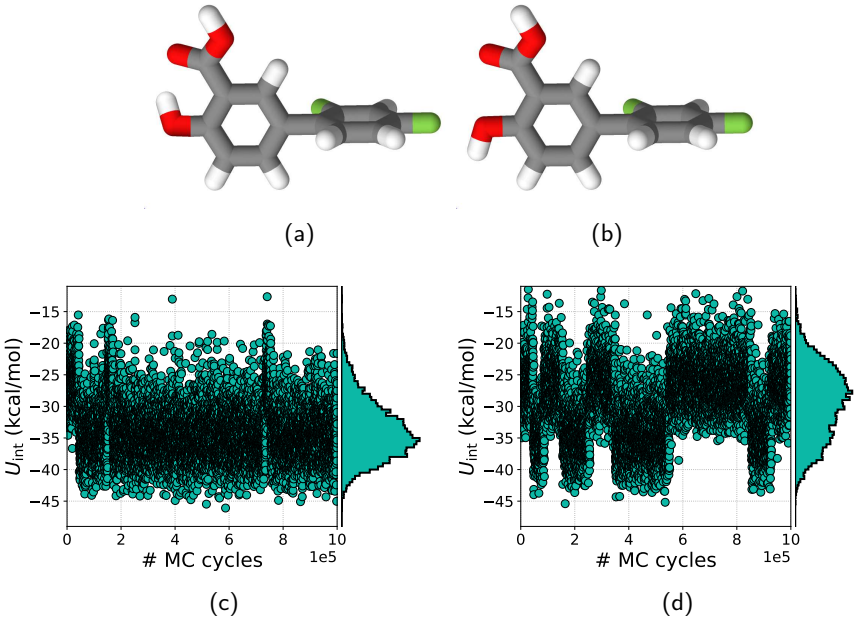


Figure 6.1: Illustration of diflunisal conformers (a) with ($U_{\text{int}} \sim -25$ kcal/mol) and (b) without an intramolecular H-bond ($U_{\text{int}} \sim -35$ kcal/mol) . Evolution of the diflunisal internal energy in solution as a function of the MC cycles for (c) $p_{\text{H}_2\text{O/solute-site}} = 1$ and (d) $p_{\text{H}_2\text{O/solute-site}} = 1/5$.

Still, accumulating 3 000 ins/des process (same the number of accumulations used for the rigid solute calculations in the previous chapter) with an interval of $n_{\text{MC}}^{\text{H}_2\text{O}} = 100$ MC cycles leads to 3×10^5 MC cycles in total, for which the sampling of diflunisal conformers is rather poor as illustrated in fig. 6.1d. Therefore, for the case of destruction, the accumulation interval is extended to $n_{\text{MC}}^{\text{H}_2\text{O}} = 1\,000$ MC cycles, to improve sampling, which increases the destruction simulation computation time by two (~ 33 cpu.h). For the insertions, the interval for the bulk solvent is kept at 100 MC and the vacuum conformers, for which the propagation is computationally very cheap, are obtained with a separate MC simulation with an accumulation interval of $n_{\text{MC}}^{\text{vacuum}} = 10^4$, *i.e.* each vacuum conformer is obtained with $n_{\text{MC}}^{\text{vacuum}} \times n_{solute\,sites}$ elementary displacement attempts.

Figures 6.2a and 6.2b show the evolution of the internal energy of the diflunisal in solution ($n_{\text{MC}}^{\text{H}_2\text{O}} = 1\,000$) and in vacuum ($n_{\text{MC}}^{\text{vacuum}} = 10^4$), respectively. For the diflunisal, in solution, both conformers are present, whereas, in vacuum, the molecule stays in the more stable conformer with the intramolecular H-bond. These results are expected as the two states are separated by ~ 10 kcal/mol ($\sim 16\,k_{\text{B}}T$) which (i) is a relatively high energy difference to overcome in vacuum and (ii) makes the intramolecular H-bond conformer much more stable, whereas in water (i) the thermal fluctuations enable the crossing of larger energy gaps and (ii) the addition of two H-bonds with the solvent, obtained by breaking the intramolecular H-bond, stabilises the solvated molecule.

The last parameter related to solute flexibility is the solute sites mass during the ins/des process. $m_{solute\,sites} = \infty$ (numerically 10^{20}) makes the solute rigid during the ins/des process, thus leading to the case of 'mixture of single conformer calculations', whereas $m_{\text{solute sites}} = 1$ makes the solute fully flexible and gives the solute's sites the same mass as the solvent molecules during the MD ins/des process. In order to obtain good statistics we expect the solvent relaxation to be more important than the solute relaxation during the ins/des process. Therefore, it is interesting to have $m_{\text{solute sites}} > 1$ to prioritise the movement of solvent molecules. We expect that the overlap between insertion and destruction distributions increases with a finite solute mass compared to the rigid solute and we expect a better overlap for $m_{\text{solute sites}} > 1$ than $m_{\text{solute sites}} = 1$.

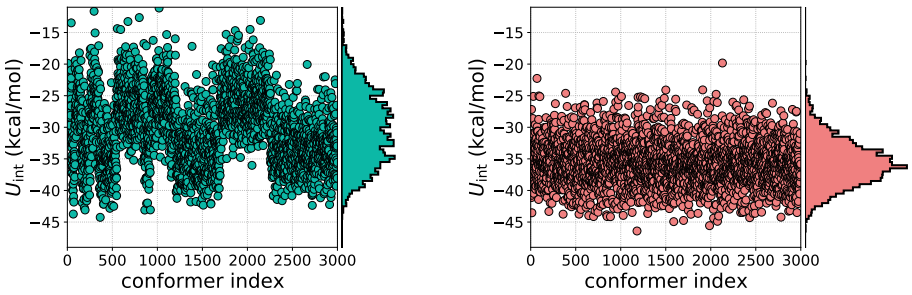


Figure 6.2: Evolution of the diflunisal internal energy in (a) solution and (b) vacuum.

Figure 6.3a plots the insertion and destruction distributions for diflunisal with $m_{\text{solute sites}} = 1, 3, 5, \infty$ after 3 000 accumulations. The plots are smoothed by a Gaussian filter ($\sigma = 1$) to make them more readable. Firstly, the insertion distributions p_{ins} are narrower than the destruction distributions p_{des}. This is not surprising as 'all' the insertions are started with the 'same' conformer with the intramolecular H-bond, whereas the destructions are started both in the conformations with and without the intramolecular H-bond. This effect can be seen the most clearly on $p_{\text{des}}^{m=\infty}$ which seems to be a combination of two 'Gaussians' centred at ~ 20 and ~ 14 kcal/mol. In the case of a finite mass, this double peak seems to disappear forming a single large blob, as the solute structure can evolve between the two forms during the destruction process. Secondly, looking at the position of the distribution, one sees that the $p_{\text{des}}^{m=3}$ and $p_{\text{des}}^{m=5}$ are similarly positioned, whereas $p_{\text{des}}^{m=1}$ and $p_{\text{des}}^{m=\infty}$ are somewhat shifted/stretched to lower values of ΔH, *i.e.* away from the p_{ins} distributions. For the insertion distributions, they are all very similar with maybe $p_{\text{ins}}^{m=3}$ and $p_{\text{ins}}^{m=5}$ slightly narrower than $p_{\text{ins}}^{m=1}$ and $p_{\text{ins}}^{m=\infty}$.

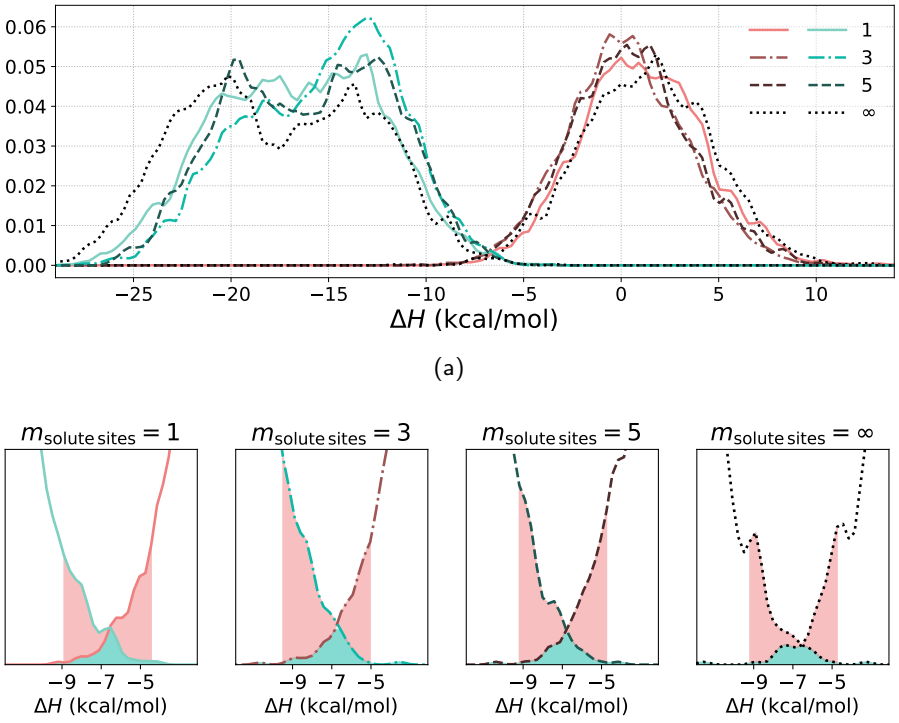


Figure 6.3: Diflunisal's (a) insertion (red) and destruction (turquoise) distributions with $m_{\text{solute sites}} = 1, 3, 5, \infty$ and a zooms of overlap region in (b).

Overall all the distributions are very similar and have similar overlaps of p_{ins} and p_{des} (fig. 6.3b). If one integrates the overlap area, one finds that is slightly larger areas for $m_{\text{solute sites}} = 3$ and 5 that $m_{\text{solute sites}} = 1$and ∞ as we expected. There is no significant difference between $m_{\text{solute sites}} = 3$ and 5, so we arbitrary chose $m_{\text{solute sites}} = 3$ as the default value for fully flexible H4D-MC calculations.

To summarize, the default parameters for flexible solute in H4D-MC are

- maximum solute site displacement $\Delta r_{\text{max}}^{\text{solute sites}} = 0.1$ Å
- solute mass during ins/des : $m_{\text{solute sites}} = 10^{20}$ (without solute relaxation) or $m_{\text{solute sites}} = 3$ (with solute relaxation)
- for insertion
  - 10^2 MC cycle propagation of the bulk solvent + 10^4 MC propagation of the solute in vacuum
- for destruction
  - 10^3 MC cycle propagation of the solvated system
  - preferential displacement of solute's sites $p_{\text{H}_2\text{O/solute-site}} = 1/5$

6.3 COMPARISON OF SINGLE RIGID CONFORMERS VS. FULLY FLEXIBLE SOLUTE

In this section, we compare the performance of the two flexible solute H4D-MC methods: the solute conformers are sampled in vacuum and solution during the MC propagation and (i) the solute is kept rigid during the ins/des processes (= combination of multiple single conformer ins/des) or (ii) the solute can be relaxed during ins/des process. Figure 6.4 shows the correlation between the HFEs obtained by the two approaches after $2 \times 3\,000$ accumulations. With excellent correlations and very small deviations, it shows that both methods produce similar results. As expected, the average statistical error of each solute, shown in fig. 6.5 (triangles) is smaller if the solute is relaxed during ins/des than when not.

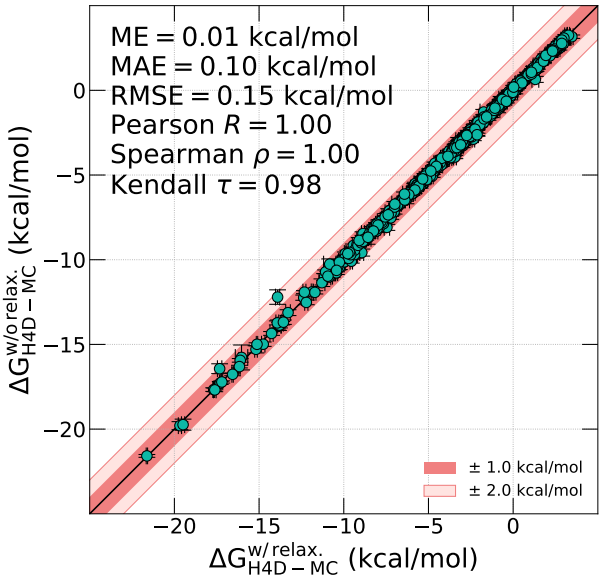


Figure 6.4: Correlation between HFEs obtained with flexible solute H4D-MC with and without solute relaxation during the ins/des process for the FreeSolv database.

Figure 6.5 also shows the evolution of the ME, MAE and RMSE between the flexible MD+FEP and flexible H4D-MC with and without solute relaxation during the ins/des process. First of all, deviations of both approaches to the classic flexible solute MD+FEP calculations are very small.

With relaxation and 3 000 accumulations, the ME is quasi-zero, the MAE is 0.12 kcal/mol and the RMSE 0.21 kcal/mol. As we recover the MD+FEP results, this is a confirmation that H4D-MC approach works for the computation of HFEs of flexible solutes and the errors of single conformer calculations are due to solute flexibility and not bugs in the H4D-MC code. Now, we can be confident in our rigid solute reference data produced in the previous chapter.

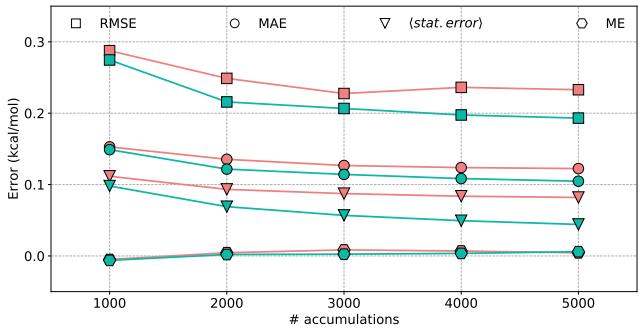


Figure 6.5: The MAE (circles) and RMSE (squares) between flexible solute MD+FEP and H4D-MC with (turquoise) and without (pink) solute relaxation, and the average statistical error (triangles) as a function of H4D-MC accumulations.

Similarly to the statistical error of each solute, the MAE and RMSE are lower with the fully flexible solute approach than with the mixture of single conformers approach. However, the effect of solute relaxation is very small. The small difference between the fully flexible solute and the mixture of single conformers approach demonstrates than the flexible solute HFEs can be recovered with single conformer methods, like MDFT. However, we used few thousands of conformers which is not feasible with MDFT, as the computation of the HFEs eg. 6 000 conformers would lead to ~ 200 cpu.h, *i.e.* computation times similar to MD+FEP. The use of only a few conformers is discussed in the following chapter.

Moreover, with the flexible solute calculation with H4D-MC, with $2 \times 3\,000$ accumulations, leads to a computation time of ~ 51 cpu.h compared to ~ 220 cpu.h with Gromacs. Thus, we have proposed a new method to compute flexible solute HFEs with a speedup of four times when compared to classic MD+FEP approaches.

Why this chapter?

There are some divergences between rigid solute H4D-MC and flexible solute MD+FEP hydration free energies. Are these deviations due to solute flexibility or due to problems in H4D-MC?

We showed that H4D-MC can be used for the efficient computation of the flexible solute HFEs of small molecules. First of all, these results confirmed the legitimacy of our single conformer reference data. Moreover, showed that flexible solute SFEs can be recovered from single conformer SFE calculations and H4D-MC allows a speedup of four times compared to classical MD+FEP when computing HFEs.

7

FLEXIBILITY IN THE FREESOLV DATABASE

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The liquid state approach offers a compromise between speed and accuracy but is a single conformer approach. Is a rigid solute free energy method usefull ?

In this chapter we analyse if solute flexibility is important for the computation of HFEs of small drug-like molecules. If it's the case can we predict for which solute features is important. Additionally, we identify a set of solutes suited to benchmarking rigid solute methods like our MDFT and (ii) solute features that can predict if solut

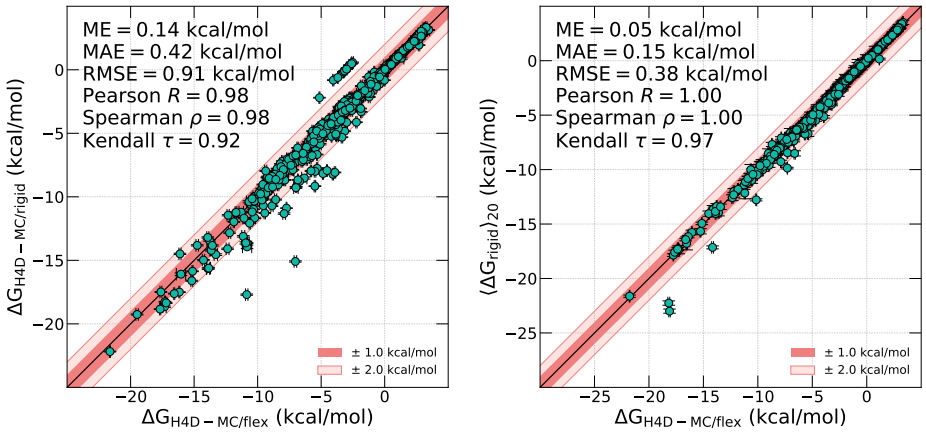
In this chapter, we try to evaluate whether it is necessary to have a flexible solute molecule or is a single conformer calculation enough when predicting hydration free energies of small drug-like molecules. This analysis is done by comparing rigorous H4D-MC simulation results obtained either with a single conformer solute as calculated in chapter 5 and a fully flexible solute as computed in chapter 6 for the FreeSolv database of small drug-like molecules [144]. Moreover, we try to identify solute features which can predict if solute flexibility is necessary for the HFE calculation. Most of the results presented in this chapter are going to be published in [156].

7.1 SINGLE CONFORMER VS. FLEXIBLE SOLUTE

Figure 7.1a, shows the correlation between HFEs obtained with a single conformer and flexible solute simulations. Table 7.1 summarizes the statistical measures characterizing this correlation. In general, the agreement between the single conformer and flexible solute results is good with high correlation coefficients and mean absolute error (MAE) of 0.41 kcal/mol. The MAE is smaller than a typical experimental error of ∼0.5 kcal/mol of modern-day calorimetry, but significantly larger than the statistical errors of the H4D-MC method with error bars below 0.1 kcal/mol.

Figure 7.2a and table 7.2 quantify at which point a single conformer calculation is sufficient for the FreeSolv database. For almost half of the database, the deviations are below 0.1 kcal/mol, *i.e.* similar to the statistical errors. Hence, for these molecules, the solute flexibility does not have any effect on the hydration free energy. Moreover, for 80% of the database, the effect of flexibility is not critical as they have deviations smaller than the experimental error. However, there are some important outliers with 10% of the database having deviations of more than 1 kcal/mol between the rigid and flexible solute results and there are even five solutes with a deviation larger than 4 kcal/mol (illustrated in fig. 7.2b).

In the following, a solute is considered 'rigid' in respect to the hydration free energy, *i.e.* the hydration free energy of a molecule is not affected by the lack of solute flexibility, if the deviation is between the flexible and single conformer calculation is below 0.1 kcal/mol (295 solutes) for development purposes in chapter 8 or below the average experimental error of 0.6 kcal/mol for analysis in chapter 10.



(a) (b)

Figure 7.1: Comparison of (a) single conformer and (b) average of 20 single conformers and fully flexible solute hydration free energies obtained with H4D-MC for the FreeSolv database.

	Full dataset (642)	
	1 conformer	20 conformers
MAE [kcal/mol]	0.41 ± 0.07	0.15 ± 0.03
RMSE [kcal/mol]	0.94 ± 0.19	0.38 ± 0.13
ME [kcal/mol]	0.15 ± 0.07	0.05 ± 0.03
Max. error [kcal/mol]	8.47	4.89
Pearson's R	0.98 ± 0.01	1.00 ± 0.01
Spearman's ρ	0.98 ± 0.01	1.00 ± 0.01
Kendall's τ	0.92 ± 0.01	0.97 ± 0.01

Table 7.1: .Summary of the statistical measures characterizing the correlations between HFEs obtained with flexible solute and with (i) one single conformer or (ii) a combination of 20 single conformers H4D/MC simulations calculations for the full FreeSolv database.

AE [kcal/mol]	< 0.1	< 0.2	< 0.5	< 1.0	< 2.0	< 4.0
Single conformer						
#solutes (%)	295 (46)	409 (64)	515 (80)	576 (90)	607 (95)	637 (99)
MAE [kcal/mol]	0.05	0.07	0.12	0.18	0.25	0.37
20 conformers						
#solutes (%)	399 (62)	524 (82)	613 (95)	628 (98)	635 (99)	640 (100)
MAE [kcal/mol]	0.04	0.07	0.10	0.11	0.13	0.14

Table 7.2: Number of solutes and the cumulative mean absolute error (MAE) as a function of the absolute error (AE).

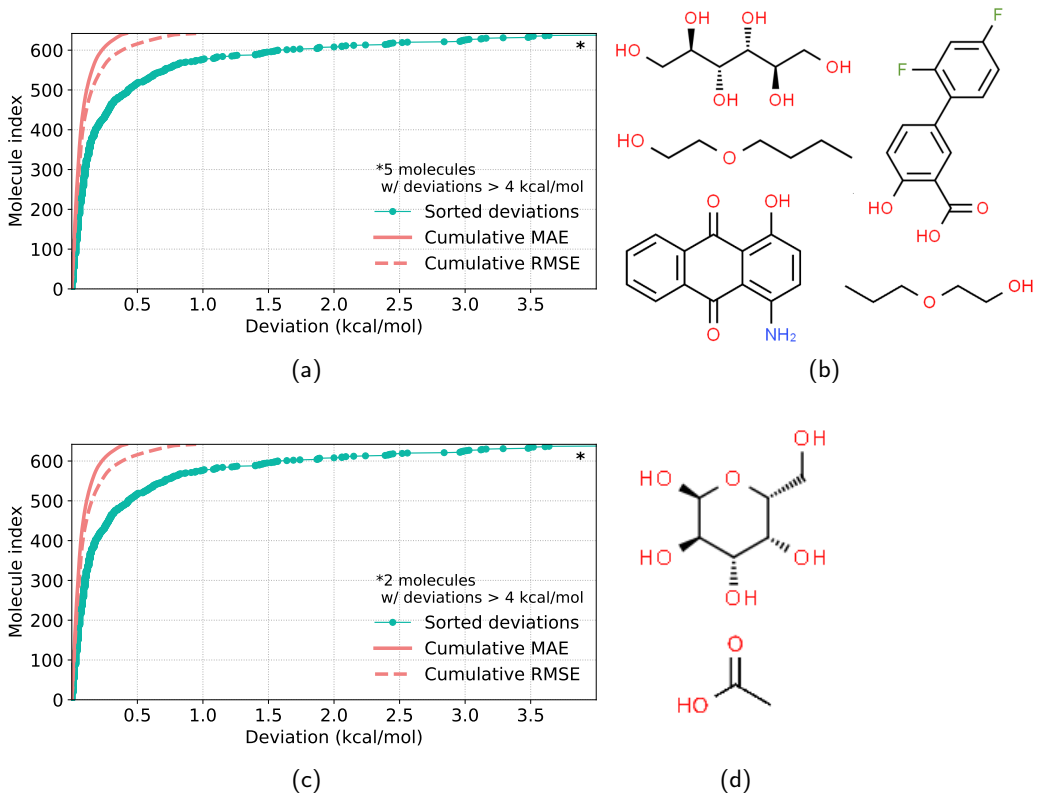


Figure 7.2: Sorted deviations between flexible solute and (a) single conformer and (c) 20-conformer average hydration free energies with cumulative MAE (solid pink line) and RMSE (dashed pink line). (b) and (d) Chemical structure of the solutes with deviations larger than 4 kcal/mol in each case.

7.2 MULTIPLE CONFORMERS ANALYSIS

The single conformer calculations were done with the initial conformer given in the FreeSolv database. The origin of these conformers is not discussed by Mobley and co-workers [152, 144] as, in their flexible solute simulations, the solute structures will be equilibrated and propagated. They seem to correspond to a (local) minima in vacuum. However, from the previous comparison, we cannot be sure that a solute really is 'rigid', concerning its HFE, or were we just '(un)lucky' and the single conformer simulation with initial conformer gives the right average HFE of a flexible solute whereas another conformer would not give it.

To confirm that the 'rigid' solutes, as defined above, are really rigid we performed 2×10 shorter rigid solute simulations for $N_{conf} = 2 \times 10$ conformers obtained after 10^5 MC cycle propagation in vacuum or solution for all the FreeSolv database molecules. Two short 500 accumulations ins/des simulations were run for each conformer to obtain a rough estimate of the HFE of each conformer (the average statistical error is 0.14 kcal/mol). The length of the simulations was limited, as even with 500 accumulations, the computation of the analysis reaches ~ 120 cpu.h for each molecule bringing the computation time for the analysis of the whole database to $\sim 77\,000$ cpu.h.

For each solute, the Boltzmann average of the HFEs $\langle \Delta G_{\text{rigid}} \rangle = -k_{\text{B}} T \ln \langle e^{-\beta \Delta G_{\text{rigid}}} \rangle$, their standard error $\sigma(\Delta G_{\text{rigid}})$ and the average statistical error $\langle \text{ste}(\Delta G_{\text{rigid}}) \rangle$ was computed from these 20 conformer calculations. Fig. 7.1b shows the correlation between HFE obtained with the fully flexible solute HD4-MC calculation and the $\langle \Delta G_{\text{rigid}} \rangle$. Using 20 conformers per solute instead of only one improves all the statistical measures (see table 7.1: the correlation coefficients are now $R = \rho = 1.00$ instead of $R = \rho = 0.98$ and the MAE and the RMSE are 0.15 and 0.38 kcal/mol instead of 0.41 and 0.94 kcal/mol before. The use of 20 conformers instead of only one decreases the error of 67 %.

First of all, this good news for single conformer SFE methods, like MDFT, as for most small flexible solutes of the FreeSolv database, the flexible solute HFE can be recovered from rigid solute HFE calculations of a few conformers. However, there are still some solutes with large deviations to the flexible result. Therefore, this approach should be applied with caution to single conformer calculations and more reliable sampling methods should be considered or developed when trying to include solute flexibility to single conformer methods.

However, this analysis does not give us the confirmation of whether the solutes defined as 'rigid' in the previous section, *i.e.* $|\Delta G_{\text{flex}} - \Delta G_{\text{rigid}}^{\text{initial}}| < 0.1$ kcal/mol, are really rigid within respect to the HFE or were we just '(un)lucky' with initial conformer. Therefore we define a solute really to be rigid if it fulfils the following two criteria

$$|\Delta G_{\text{rigid}}^{\text{initial}} - \langle \Delta G_{\text{rigid}} \rangle| < 0.1\,\text{kcal/mol}$$
$$\text{ste}(\Delta G_{\text{rigid}}) = \frac{\sigma(\Delta G_{\text{rigid}})}{\sqrt{N_{conf}}} < \langle \text{ste}(\Delta G_{\text{rigid}}) \rangle. \tag{7.1}$$

where $\text{ste}(\Delta G_{\text{rigid}})$ the standard error of the multi-conformer HFE. Figures 7.3a and 7.3b illustrate these two new rigidity criteria: they plot the correlation between the original rigidity criterion, *i.e.* the deviation between the flexible solute and single conformer calculation, and the deviation between the original single conformer and the multi-conformer average of short single conformer simulations (7.3a) and the standard deviation of the short simulations (7.3b). It seems that the majority of the molecules defined as rigid previously are really rigid in the sense that the solute flexibility does not affect the hydration free energy as defined by the new criteria.

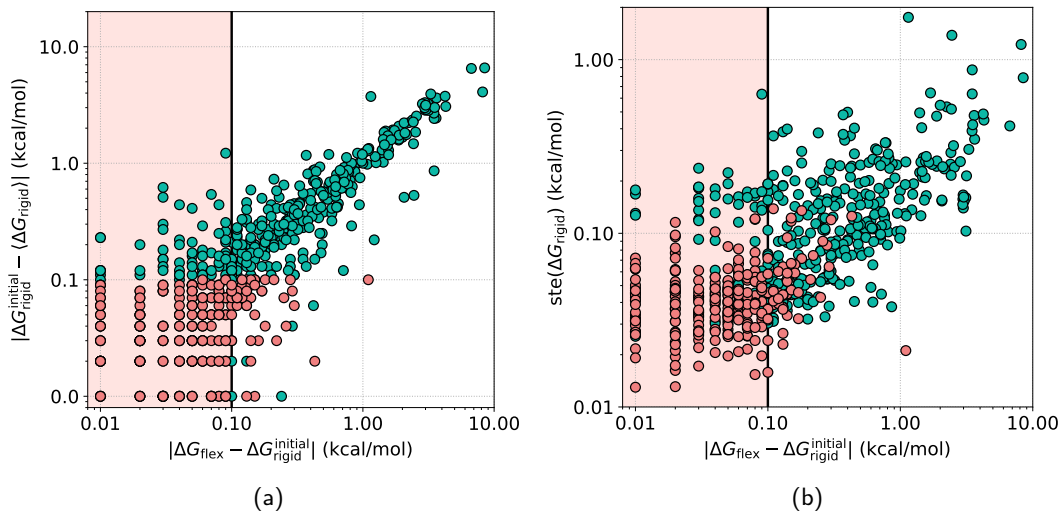


Figure 7.3: The deviation between the initial single conformer HFE and (a) the mean HFE of short single conformer simulations and (b) the standard deviation of the short single conformer HFEs as a function of the original flexible solute-single conformer deviation for the FreeSolv database. In pink solutes that fill the rigidity criteria of equation 7.1 and the flexible ones in turquoise.

However, there are some 'outliers'. Some solutes, that were defined as rigid previously are in reality flexible, *i.e.* we were '(un)lucky' to find the correct HFE with a single conformer. There are also molecules defined as 'flexible' previously that seems to be rigid in the reality. To quantify this, table 7.3 shows the confusion matrix, *i.e.* the error matrix, between solutes defined as rigid or flexible from the initial conformer calculation and the multi-conformer average. In general, both approaches define a solute flexible or rigid for 80 % of the molecules. However, of the 295 solutes originally defined as 'rigid' only 214 (72 %) are really rigid as defined by the multi-conformer analysis.

Initial conf. \ Multi-conf.	R	F	
R	214	81	295
F	46	301	347
	259	267	

Table 7.3: Single conformer and multi-conformer confusion matrix for solute flexibility of the FreeSolv database.

We can now, with confidence, identify these 214 solutes defined as rigid with the original and the new rigidity criteria as a rigid solute sub-set of the FreeSolv database, called 'FreeSolv-rigid' that can be used for the development of single conformer SFE methods like liquid state theories and continuum model approaches. This sub-set is given in appendix E.

7.3 FOCUS ON FLEXIBLE SOLUTES

7.3.1 EFFECT OF MASS, NUMBER OF BONDS AND RINGS

In this section, we analyse the error distributions along a few selected solute that could be related to solute flexibility. The aim is to be able to predict a priori for which types of solutes, or solute features, solute flexibility is necessary. Fig. 7.4 plots the error distribution in function of the solute size, represented by the solute mass, and the number of rotational bonds and cycles in the solute molecules. One could expect solute flexibility to increase with the solute size and the number of rotational bonds, and decrease with the number of cycles. As it can be seen on the fig. 7.4 there is no significant correlation between these features and deviation between the single conformer and flexible solute.

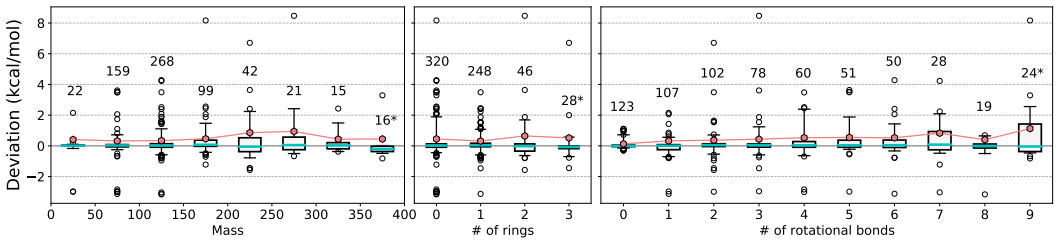


Figure 7.4: Distribution of the deviations between HFEs obtained with flexible solute and single conformer H4D/MC simulation as a function of three features (a) solute's molar mass (with a bin size of 50 Da), (b) number of rings in a solute and (c) the number of rotational bonds in a solute with turquoise lines corresponding to the median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. The pink hexagons and the numbers above each bin correspond to the MAE and the population of the bin. (*) The last represented bins regroup the solutes with a mass = 350-493 Da, number of rings = 3-5 and number of rotational bonds = 9-16 to be statistically significant.

7.3.2 EFFECT OF H-BOND DONORS AND ACCEPTORS

In figure 7.5, we show the same plot as fig. 7.4, but for the number of H-bond donors (HBD) and acceptors (HBA). There is an important effect of the number of H-bond donors on the deviation with a factor 4 between the MAE of 0.89 kcal/mol, twice as much as for the whole database, for the solutes with an HBD (29% of the database, fig. 7.6b) compared to solutes without one at 0.22

kcal/mol (fig. 71% of the database fig. 7.6a), which is half of the MAE of the whole database. The increased number of HBDs in a solute introduces a somewhat systematic bias to the single conformer calculations as the mean (signed) error (ME) increases with the number of HBD in a solute. There seems to be the same effect of HBAs, with a factor 5 between the MAE of 0.56 kcal/mol for the solutes with HBAs (67% of the database, 7.6d) and the MAE of 0.11 kcal/mol for solutes without them (33% of the database, 7.6c). Note that, almost all the HBD groups, like hydroxyls and amines, are also HBAs.

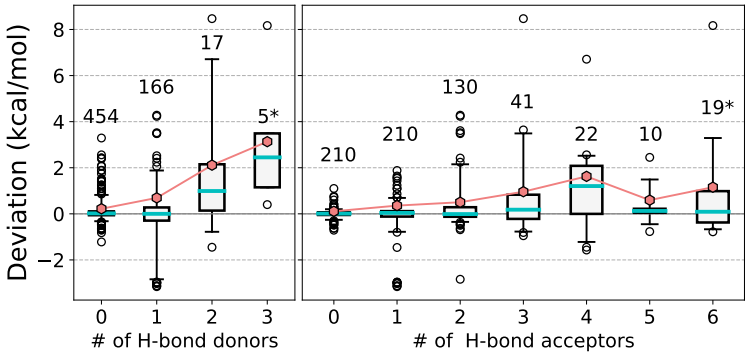


Figure 7.5: Distribution of the deviations between flexible solute and single conformer HFEs as a function of the number of H-bond (a) donors and (b) acceptors in a solute with turquoise lines corresponding to the median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. The pink hexagons and the numbers above each bin correspond to the MAE and the population of the bin. (*****) The last represented bins regroup the solutes with 3-6 H-bond donors and 6-8 H-bond acceptors to be statistically significant.

We identified 47 molecules in the database, including the five solutes with the largest deviation illustrated in fig. 7.2b, that have at least HBD-HBA couple positioned in a way that they can form an intramolecular H-bond. In fig. 7.6f, we can see that the largest deviations are found for these solutes with a potential intramolecular H-bond with an MAE of 1.80 kcal/mol, over four times higher than for the full database. The single conformer calculations systematically underestimate the hydration free energies of these solutes with a ME of 1.29 kcal/mol.

7.3.3 EFFECT OF FUNCTIONAL GROUPS

In figure 7.7, we plot the deviation distribution between the single conformer and the flexible solute calculation as a function of the chemical groups present in the database. For statistical significance, only groups contained at least in five solutes are represented. This analysis confirms the dependence of solute flexibility on the presence of H-bond donors. Nine out of the ten chemical functions with the smallest deviations (MAE $<$ 0.3 kcal/mol) do not contain H-bond donors whereas seven out of the ten groups with the largest deviations (MAE $>$ 1.0 kcal/mol) are H-bond donors. Moreover, the largest deviations are distinctly found for 1,2-diols, *i.e.* solutes with two hydroxyl groups next to each other in a way that they can form an intramolecular H-bond. We note also, the less there is steric encumberment around the H-bond donor the larger the effect of flexibility is: primary alcohols (MAE = 1.51 kcal/mol) have larger deviations that secondary alcohols (1.08 kcal/mol) and primary amines (1.63 kcal/mol) have larger errors than secondary amines (0.54 kcal/mol).

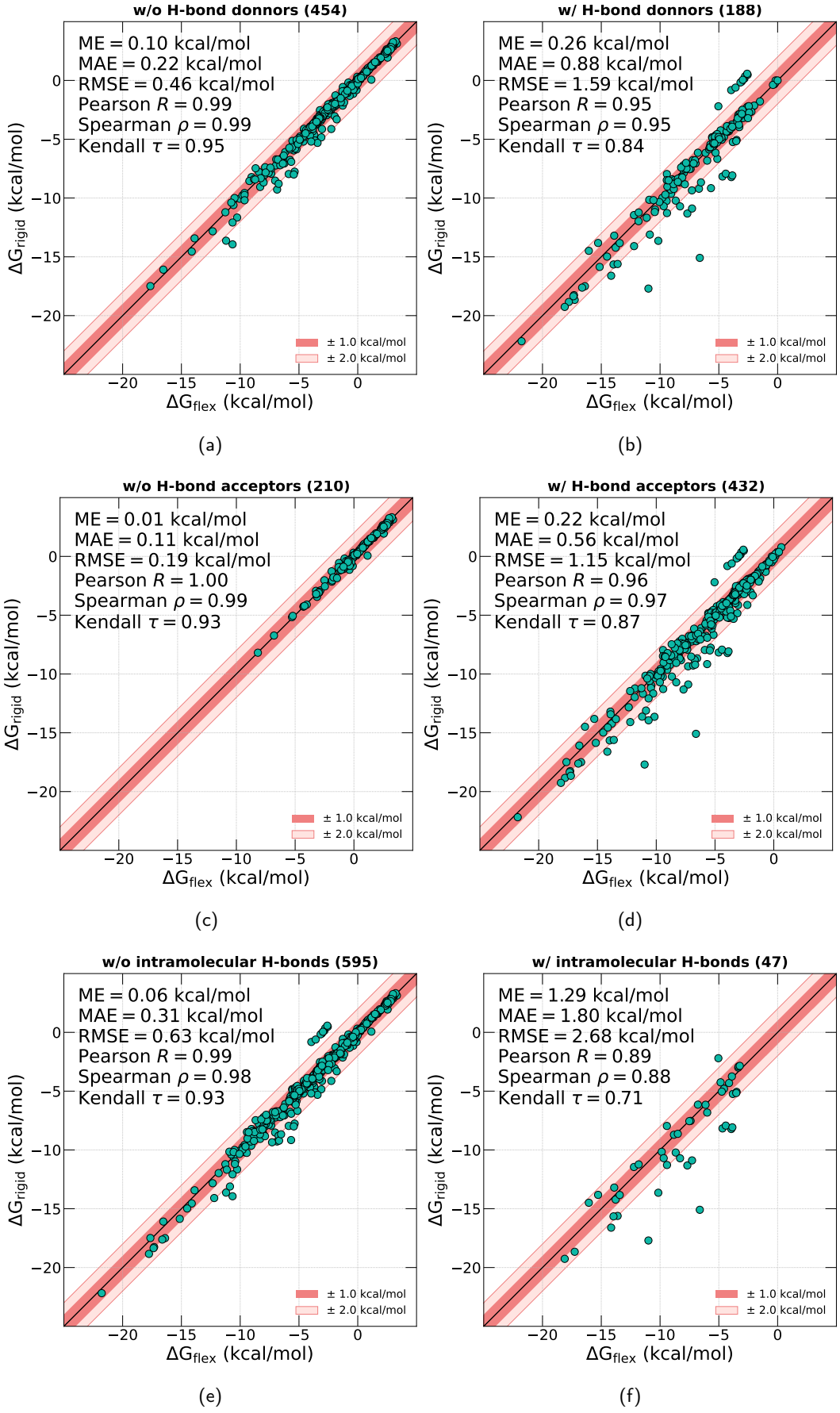


Figure 7.6: Correlations between hydration free energies predicted by flexible solute MD+FEP and single conformer H4D/MC for the sub-sets of solutes with or without H-bond donors or possible intramolecular H-bonds.

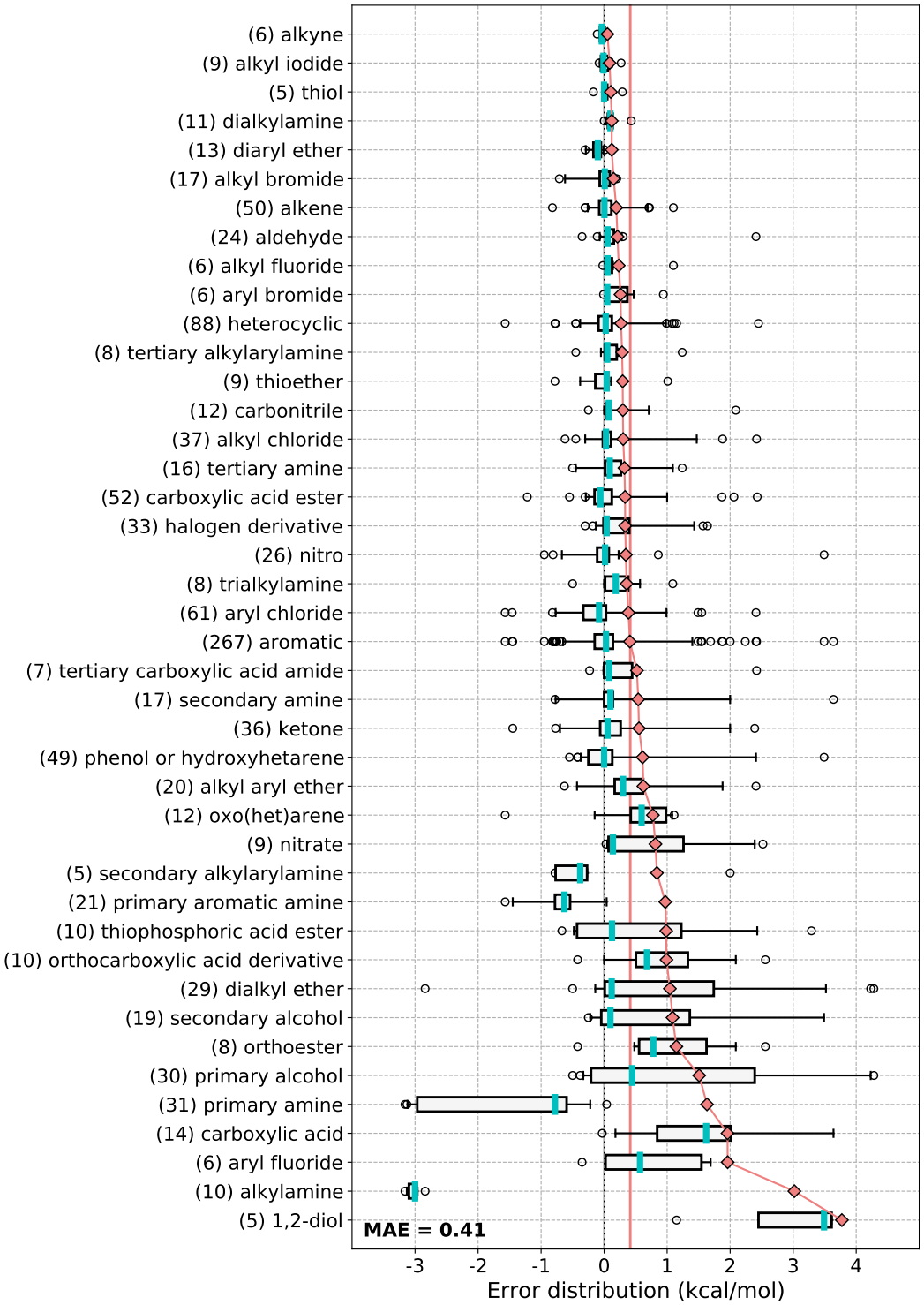


Figure 7.7: Distribution of the deviations between flexible solute and single conformer HFEs for the chemical groups with more than 5 solutes present in the database. The number of molecules in each group is written within parenthesis. Turquoise lines correspond to median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. Pink diamonds correspond to the MAE of each functional group and the vertical pink line to total MAE of the rigid subset.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. The liquid state approach offers a compromise between speed and accuracy but is a single conformer approach. Is a rigid solute free energy method usefull ?

We showed by comparison of single conformer HFEs to flexible solute HFEs and by multi-conformer analysis that solute flexibility does not have any effect on the HFEs of 34 % of FreeSolv database and can be considered as a reference "FreeSolv-rigid" for the developement of single conformer SFE methods. The multi-conformer analysis showed also that the flexible solute HFE can be recovered from a single conformer calculation of few solutes.

Additionally, we identified that the main features determining if solute flexibility is important in HFE calculations are potential hydrogen bond donors and acceptors. For molecules without H-bond acceptors, solute flexibility does not have any importance with an MAE of 0.11 kcal/mol between single conformer and flexible solute predictions. Solute flexibility is almost as negligible for solutes without H-bond donors with an MAE of 0.22 kcal/mol, whereas solute flexibility is very important for solutes with potential intramolecular H-bonds with an MAE of 1.81 kcal/mol.

Hybrid 4th dimension Monte Carlo: the novel method that compute hydration free energies via short out-of-equilibrium simulations where the solute is inserted or removed from the simulation box enables the efficient computation of hydration free energies of small drug-like molecules with a speed up 6 (rigid solute) or 4 (flexible solute) times in the computation time when compared to classic stratified free energy perturbation calculations.

FreeSolv database: of the 642 molecules of the database 213 (33%) are rigid with respect to the hydration free energy, i.e. solute flexibility does not affect the computed HFE of the solute. For the rest of the molecules, solute flexible affects their HFE. However, for 520 (81%) molecules the deviation between the rigid and flexible solute calculations is smaller than the typical experimental error of 0.6 kcal/mol. Solute flexibility is important for molecules with H-bond donors and/or acceptors and especially for molecules with potential intramolecular H-bonds.

Part III

HYDRATION WITH MDFT-HNC

This part presents the developments made to MDFT-HNC and the main results obtained with MDFT-HNC.

Chapter 7 briefly introduces an indispensable *a posteriori* pressure correction of the hydration free energies predicted by MDFT-HNC. This pressure correction compensates of the large overestimation of the cavitation formation energy for the HNC approximation. The chapter also includes the presentation of four developments made to improve this correction by (i) taking into account solvent compressibility, (ii) optimization of the solute volume, (iii) addition of machine learning terms or (iv) by addition of a surface term to the correction.

Chapter 8 presents a rigorous benchmarking of MDFT-HNC with the new surface term corrected pressure correction on a multitude of systems: starting from simple hydrophobic spheres to molecular solutes like water itself and small organic molecules, *via* spherical ions. It aims to assess carefully the accuracy of MDFT at the HNC level, acknowledge its successes, and more importantly enlighten where it fails, in order to pinpoint on which aspects the efforts for proper bridge functionals should be put. To this end, the MDFT results will be compared systematically throughout this paper to 'exact' results generated by ourselves by Monte-Carlo.

Chapter 9 makes more through chemo-informatics analysis of MDFT-HNC results of the small drug-like molecules of the FreeSolv database. The aim is the assess the performance of MDFT-HNC, with the faster optimized solute volume pressure correction, to predict solvation free energies in the scope of drug design and identifying features of solutes for which MDFT performs well or not.

8

PRESSURE CORRECTION

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. MDFT could offer a compromise between speed and accuracy. However, the bare MDFT-HNC hydration free energy predictions are in the woods. We know that most of this error is due to a bad estimation of the system's pressure. Can we correct this pressure estimation?

An original pressure correction was proposed by the group in 2014. In this chapter we try to improve it with three approaches :

- optimisation of the solute volume
- fitting by machine learning
- addition of a surface term

One of the characteristics that should-be accounted for when describing water is the presence of a liquid-gas coexistence at normal conditions, $T = 300$ K and $P = 1$ atm. Due to this, experimentally, the creation of a microscopic gas bubble, or cavity, in bulk water does not cost anything energetically. This coexistence of liquid and gas phases transcribes as a double-well in the homogeneous free energy as a function of the density (fig. 8.1) with minima at gas density $n_{\text{gas}} \approx 0$ and at liquid bulk density $n_{\text{bulk}} = 1$ kg/l or 0.033 molecules/Å^3.

The major weakness of the HNC approximation has been known for a long time [157], as its original sin is to be a quadratic theory around the liquid bulk density n_{bulk} and thus to only have a single well at n_{bulk}. Therefore, it can not accommodate for the liquid-gas transition as it largely overestimates the pressure of the system with $P_{\text{HNC}} \sim 10\,000$ atm instead of the experimental 1 atm. This consequently leads to large overestimations of the cavity creation energy in HNC.

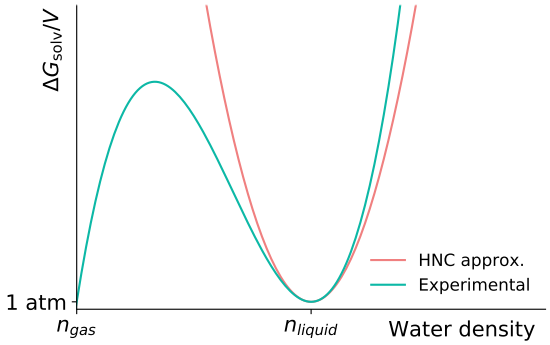


Figure 8.1: Homogenous free energy as a function of water density.

There are two options to overcome this problem : (i) by pushing the theory beyond the second-order and introducing the so-called bridge functional which, by definition, starts at the cubic order in perturbation [128, 129, 130, 131, 132]; (ii) by adding a so-called pressure correction (PC),

an *a posteriori* correction to HNC results to correct for the overestimation of the bulk pressure. This chapter focuses on the new developments of the pressure correction with a short description of the original pressure correction proposed in refs. [158, 159] before a presentation, of first a fundamental addition to the PC, and then of the developments made to improve it with three different approaches.

8.1 ORIGINAL PRESSURE CORRECTION

For large spheres or any general cavity, the solvation free energy should tend to PV, where P is the experimental or simulated pressure, thus virtually zero at a normal pressure of 1 atm unless micrometric sizes are reached. On the other hand, the HNC approximation has non-zero values of $P_{\mathrm{HNC}}V$ as the pressure is given by

$$P_{\mathrm{HNC}} = \frac{\mathcal{F}[0]}{V_{\mathrm{cell}}} = k_B T n_{\mathrm{bulk}} \left(1 - \frac{1}{2} n_{\mathrm{bulk}} \hat{c}^{000}(0)\right) \tag{8.1}$$

where V_{cell} is the volume of the supercell and $\hat{c}^{000}(0)$ is the $q = 0$ Fourier component of the spherically averaged direct correlation function in eq. 4.8. P_{HNC} has largely overestimated values compared to experiment or simulations ones : $P_{\mathrm{HNC}}^{\mathrm{SPC/E}} = 11\,260$ atm and $P_{\mathrm{HNC}}^{\mathrm{TIP3P}} = 9\,400$ atm. This leads to a large overestimation of the cavity formation free energy and the solvation free energy as illustrated in fig. 8.2a on the FreeSolv database. Hence, a first pressure correction was proposed by Sergiivskyi et al. [158, 159] to correct the overestimation of the bulk pressure and reads

$$\mathrm{PC} = -(P_{\mathrm{HNC}} - P_{\mathrm{Exp}})V \simeq -P_{\mathrm{HNC}}V. \tag{8.2}$$

Note two remarks on the PC: (i) even if this correction is justified in the macroscopic limit, it is not at the molecular level; and (ii) how should one define the solute volume V.

Originally the solute volume was defined by the unambiguous partial molar volume (PMV or V_{PM}). The PMV can be derived rigorously in LSTs from the variation of ΔN of the number of solvent molecules in the supercell while inserting the solute at a constant temperature, volume and solvent chemical potential. The PMV pressure correction reads

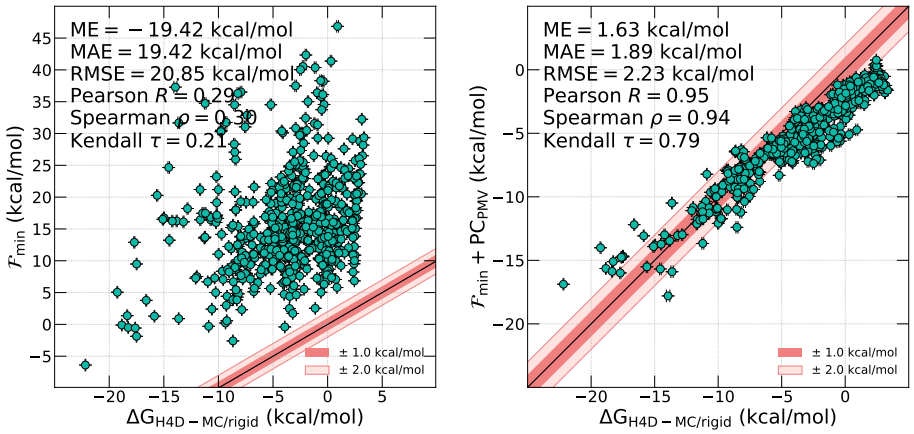
$$\begin{aligned} \mathrm{PC}_{\mathrm{PMV}} &= -P_{\mathrm{HNC}} V_{\mathrm{PM}} \\ &= -P_{\mathrm{HNC}} \frac{\Delta N}{n_{\mathrm{bulk}}} \end{aligned} \tag{8.3}$$

This correction improves drastically the solvation free energies predicted by MDFT (fig. 8.2b). For the FreeSolv database, MDFT$^{\text{HNC-PMV}}$ yields an MAE of 2.14 kcal/mol and $R = 0.95$ compared to 19.42 kcal/mol and $R = 0.29$ for the uncorrected MDFT-HNC results. A closely related pressure correction of type $aV_{\mathrm{PM}}+b$ with empirically adjusted constants a and b has been proposed and is wildly used for 3D-RISM calculations [160, 161, 162].

At the same time, an additional empirical pressure correction was proposed that improved significantly predictions [158, 159], and read

$$\begin{aligned} \mathrm{PC}^{+}_{\mathrm{PMV}} &= -(P_{\mathrm{HNC}} - \mathrm{P_{id}}) V_{\mathrm{PM}} \\ &= -(P_{\mathrm{HNC}} - k_{\mathrm{B}} T n_{\mathrm{bulk}}) V_{\mathrm{PM}}. \end{aligned} \tag{8.4}$$

Although this correction is now wildly used in the RISM community [162], we find that it does not improve the results with our current ($n_{\max} \geq 3$) MDFT-HNC version. Besides we could never finally justify it theoretically. A field theory approach [163] eventually leads to $-(P_{\mathrm{HNC}} - P_{\mathrm{id}}/2)V_{\mathrm{PM}}$ but not to eq. 8.4. We thus prefer to stick to the well-justified original pressure correction of eq. 8.3.



(a) (b)

Figure 8.2: Comparison between the hydration free energies of the FreeSolv database obtained by reference simulations and MDFT-HNC (a) without correction and (b) with the $\mathrm{PC}_{\mathrm{PMV}}$ correction.

8.1.1 COMPRESSIBLE FLUIDS

In the original $\mathrm{PC}_{\mathrm{PMV}}$, the PMV was defined as $V_{\mathrm{PM}} = \Delta N / n_{\mathrm{bulk}}$ which is the correct definition for incompressible fluids. This is not the case of the water and the correct partial molar volume reads,

$$V_{\mathrm{PMV}} = (\Delta N - n_{\mathrm{bulk}} k_{\mathrm{B}} T \chi_T) n_{\mathrm{bulk}}^{-1} \tag{8.5}$$

where χ_T is the isothermal compressibility, $n_{\mathrm{bulk}} k_{\mathrm{B}} T \chi_T = 0.0773$ and 0.0630 for TIP3P and SPC/E at normal conditions. Now, the $\mathrm{PC}_{\mathrm{PMV}}$ reads

$$\begin{aligned} \mathrm{PC}'_{\mathrm{PMV}} &= -P_{\mathrm{HNC}} V_{\mathrm{PM}} \\ &= -\frac{P_{\mathrm{HNC}}}{n_{\mathrm{bulk}}} (\Delta N - n_{\mathrm{bulk}} k_{\mathrm{B}} T \chi_T). \end{aligned} \tag{8.6}$$

Note that, the new second term is constant for a given solvent: 0.32 kcal/mol for TIP3P and 0.33 kcal/mol for SPC/E. As it can be seen in table 8.1, this modification does not improve the results for the FreeSolv database, it even decreases the quality. Nevertheless, this correction will be applied in future results as it uses fundamentally the correct definition of the PMV in a compressible fluid.

The choice of the PMV as the solute volume is not unambiguous and the PC is only exact for macroscopic volumes. In the following section, we propose three approaches to improve the PC by (i) using an optimized volume for the solute, (ii) fitting the MDFT/simulations (or experiment) difference with machine learning or (iii) adding a surface term to the PC correction.

8.2 OPTIMIZED VAN DER WAALS VOLUME

First, we proposed [164] an empirical approach to improve the pressure correction: define an optimized van der Waals volume pressure correction that reads

$$\mathrm{PC}_{\mathrm{vdW}} = -P_{\mathrm{HNC}} V_{\mathrm{vdW}} \tag{8.7}$$

	$\mathcal{F}_{\text{min}}$	$\mathcal{F}_{\text{min}} + \text{PC}_{\text{PMV}}$	$\mathcal{F}_{\text{min}} + \text{PC}'_{\text{PMV}}$
MAE	19.42 ± 0.61	1.88 ± 0.09	2.14 ± 0.10
RMSE	20.84 ± 0.71	2.23 ± 0.10	2.47 ± 0.10
ME	-19.42 ± 0.61	1.63 ± 0.12	1.94 ± 0.12
Pearson's R	0.29 ± 0.09	0.95 ± 0.01	0.95 ± 0.01
Spearman's ρ	0.30 ± 0.08	0.94 ± 0.02	0.94 ± 0.02
Kendall's τ	0.21 ± 0.06	0.79 ± 0.02	0.79 ± 0.03

Table 8.1: Summary of the statistical measures characterizing the correlations between single conformer for HFEs obtained from H4D-MC and MDFT-HNC calculations (i) without any PC, (ii) with the original PC_{PMV} and (iii) with the compressibility corrected PC'_{PMV} for the FreeSolv database.

where the solute volume is defined as the sum of the volume of the voxels (of width 0.05 Å) within R_i^{vdW} of the solute atom i. This corresponds to the volume inside the solvent excluding surface in fig. 2.2. The radii depend upon the chemical nature of each atom and were initially taken from Bondi's paper [165] that gathers multiple experimental estimations and are gathered in table 8.2.

vdW radius (Å)	C	N	O	H	F
Initial values	1.700	1.550	1.520	1.200	1.470
Optimized/Sim.	1.711	1.734	1.588	1.588	1.318
Optimized/Exp.	1.682	1.893	1.430	1.353	1.510
	Cl	Br	I	P	S
Initial values	1.750	1.850	1.980	1.800	1.800
Optimized/Sim.	1.590	1.815	1.872	1.458	1.721
Optimized/Exp.	1.887	1.984	1.960	1.426	1.804

Table 8.2: van de Waals radii used for PC_{vdW}. First row: initial values as taken from experiments [165]. Second and third rows: optimized values within respect to rigid H4D-MC results and experimental results.

Since those experimental values are not defined unambiguously and they are subject to large incertitude, we optimized them so that PC_{vdW} minimizes the RMSE of MDFT compared to reference simulations or experimental values. The vdW volumes, and thus the vdW radii, were iteratively calculated and optimized via the Nelder-Mead algorithm [166, 167] using a bootstrap technique on sub-set of 288 molecules of the FreeSolv database (45% of the database) determined to be rigid in chapter 7. The radii were modified by 6% on average and are reported in table 8.2.

Firstly, we optimized the vdW radii with reference to rigid H4D/MC results in order to discard all error compensation effects due to the force field or solute flexibility. The optimization was done on a sub-set of 288 molecules to test and ensure the transferability of the new pressure correction. Figure 8.3a shows the final correlation between hydration free energies obtained with $\text{MDFT}^{\text{HNC-vdW}}$ and reference simulations for the full FreeSolv database. The PC_{vdW} divides the error by almost a factor five with respect to H4D-MC simulations: the MAE is now 0.46 kcal/mol compared to 2.14 kcal/mol with PC_{PMV}. The correlations are also improved with $R = 0.99$ and $\tau = 0.93$ compared with $R = 0.95$ and $\tau = 0.79$ before. Note that even though the vdW radii were optimised on less than half of the database, these results are obtained on the whole database showing a high transferability to the other molecules.

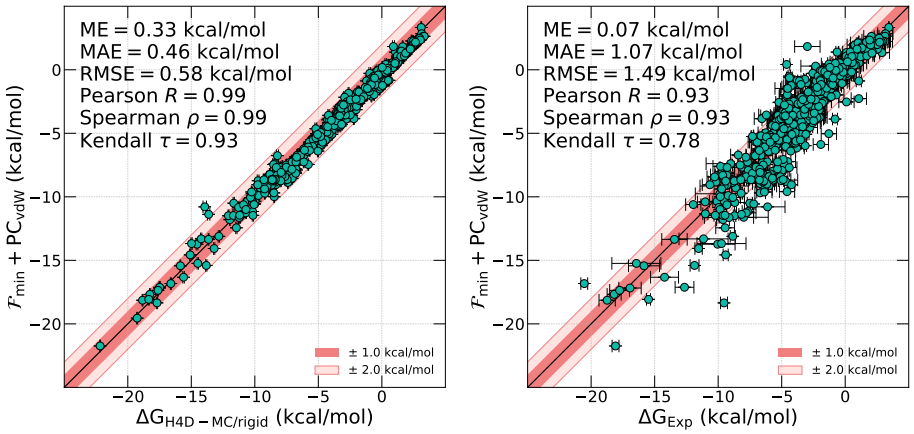


Figure 8.3: Comparison between the hydration free energies of the FreeSolv database of MDFT-HNC with the $\mathrm{PC_{vdW}}$ correction and (a) the reference simulations and (b) experimental values.

Secondly, we turned to optimise the vdW radii on experimental solvation free energies. Since MDFT computes the SFE of rigid solutes, we restrict ourselves to the sub-set of 288 rigid molecules. Figure 8.3b shows the comparison between $\mathrm{PC_{vdW}}$-corrected MDFT results and experimental SFEs. We find an MAE of 1.07 kcal/mol, thus reaching the same accuracy as reference simulation at 1.06 kcal/mol.

In this approach, the radii optimisation was done once on the 288 rigid solutes and took a few tens of CPU.min only. If the radii are optimized on a larger pool of molecules, a slight improvement can be seen but it comes from over-fitting, especially for flexible molecules. Note that the calculation of V_{vdW} is negligible compared to the MDFT minimisation and it only needs one MDFT calculation compared to the more rigorous approach presented in the section 8.4 that needs two MDFT minimisations. This is a simple, versatile and efficient pressure correction that can be applied beyond MDFT to other HNC-level liquid state theories.

8.3 MACHINE LEARNING FITTED CORRECTION

Another approach is to directly fit the free energy difference between the MDFT result and the reference calculation or experimental values instead of fitting the solute volume to have an optimal pressure correction. This fitting can be done efficiently with machine learning approaches. We propose to do it with neural networks (NNs). Here we used a feed-forward neural network (FNN), also called multilayer perceptron (MLP), regression as our ML algorithm.

8.3.1 NEURAL NETWORK

The FNNs are the simplest artificial neural networks as the information only flows in one direction, *i.e.* forward, from the input nodes, through the hidden nodes, to the output node(s) without any cycle or loop in the network. In general NNs are algorithmic structures consisting of multiple layers, each layer containing nodes, *i.e.* the neurons (fig. 8.4). Every node in a layer i is connected to all nodes of the layer $i+1$ with weights $W_{i,j\to i+1,k}$ with j and k nodes of layers i and $i+1$ respectively. The value of neuron k is given by

$$x_{i+1,k} = f_a\left(\sum_j W_{i,j\to i+1,k} x_{i,j} + b_k\right) , \tag{8.8}$$

where x is the value of a node, f_a an activation function which makes the system non-linear and b_k an additional bias. The weights W and biases b are the parameters to be adjusted by the computer during the training via back-propagation. Normally, the activation function is not applied for the transition between the last hidden layer and the output layer.

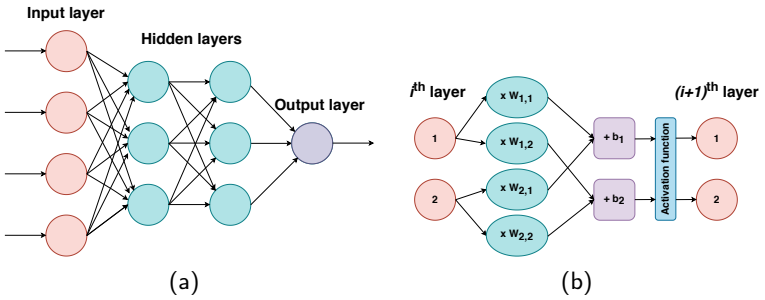


Figure 8.4: Illustration of (a) a two-layer feed-forward neural network and (b) scheme of operations between two layers.

We want to fit a single value correction to the MDFT results to minimise the deviation to simulation (or experimental) values. Therefore, the ML task is to do a regression with n_{in} input nodes corresponding each to a parameter/descriptor describing the solute molecule and a single output node, $n_{\text{out}} = 1$, corresponding the final energy correction. The learning is supervised as we try to minimise the deviation to known experimental values: at each iteration, the training outputs are calculated from the input data with the current weights and biases. Then, the difference between the calculated output and target values are evaluated with a loss function, *e.g.* cross-entropy. The training process consists of minimising this loss function *via* a back-propagation process which consists of correcting the weight and bias values, with a gradient calculation, layer by layer in reverse order starting from the output layer.

8.3.2 INPUT DATA AND HIDDEN LAYERS

How to represent molecules, by digital encoding, in a way that they can be used by a NN model as input data while it captures the essential structural and chemical information of the molecule? Two important properties that are desirable, but not required, for representations are uniqueness, *i.e.* the molecular structure is associated with a single representation, and invertibility, *i.e.* each representation is associated with a single molecule, in other words, there is a one-to-one mapping in both ways [168].

The most 'complete' way to represent a molecule it is three-dimensional chemical structure, but the direct implementation of the nuclear coordinates as ML input creates several issues. The major one is that these coordinates are not invariant to molecular translation and rotation, and the permutation of atomic indexes. Recently, multiple schemes have been proposed the essential information of the 3D structure to a more ML appropriate format using the molecule's internal coordinates or atomic densities [169]. These 3D structure-based representations are needed for ML methods applied to quantum systems but are quite cumbersome.

Therefore, many molecular representations are based on two-dimensional graphs of the molecule [168, 170]. The most well-known 2D representation of molecules is the (canonical) 'simplified molecular-input line-entry system' (SMILES) that are unique and invertible. Their problem is they are not fixed in length and thus are not optimal input for NN where the number of input nodes does not vary. The most common way is to use chemical fingerprints, which is a list, of fixed length, of binary values (0 or 1) characterising a molecule (*eg.* is there any halogen atoms present? is there more than 3 oxygen atoms? *etc.*). We chose the widely used MACCS keys [171], 166 bits long 2D fragment-based keyed fingerprints to represent the solute plus an additional input node

corresponding to the HFE predicted with MDFT, $n_{\text{in}} = 166\text{MACCs} + 1\Delta G_{\text{MDFT}} = 167$. The MACCS were generated with python's rdkit package [172] from the SMILES given in FreeSolv.

Now that, we have chosen our input data format and thus the number of input nodes, we need to choose hidden layers' structure. The optimal structure depends on the complexity of the task and the level of abstraction, *i.e.* the number of hidden layers, needed to resolve it. There is no way of knowing a priori how many layers and how many nodes in a hidden layer one should use to get an optimal result. For the large majority of ML task, only a few levels of abstraction is needed (1,2 or 3 layers) and a thumb-rule for the number of hidden nodes is given as

$$n_{\text{hidden}} \approx \frac{n_{\text{in}} + n_{\text{out}}}{2} = \frac{167+1}{2} = 84. \tag{8.9}$$

Building on trials and errors, we found that a neural network with two hidden layers with 84 hidden nodes with the rectified linear unit (ReLU) activation function gave the optimal results. The neural network was done with the python open-source library scikit-learn [173] and the model details are given in appendix F.

8.3.3 CROSS-VALIDATION

An important part of building and validating an ML model is to split the data, here the FreeSolv database, into a training set with whom the NN optimization is done, and a test set not used during the training with whom the final optimized model is validated. The imperative obligation of having a test set, *i.e.* data to validate the model not seen during training, is not a problem for large datasets ($N > 10\,000$) where a part of the original dataset, can be 'wasted' as the test set and is not used during the training. Here, our dataset is small ($N = 620$) so either we have to (i) 'waste' a large part of the dataset as the test set leading to a very small training (larger the training set better the model will be), or (ii) to use only a very small test set to validate the model leading to large uncertainty on the model's real performance. This means either having a model (i) that is not very well optimised and probably over-fitted to a small set of molecules or (ii) not be very confident in the model as it validated on a very limited number of data points.

To overcome this, we used cross-validation: the dataset is split into K sub-sets and K parallel neural networks are trained separately with each one using a different sub-set $X_{\text{test}} = X^k$ as the test set and the rest as the training set, $X_{\text{train}} = \sum_i^K X^{i \neq k}$ (fig. 8.5). We chose to split the initial dataset with $K = 62$ leading to 62 separate NN optimisations with $N_{\text{test}} = 10$ and $N_{\text{train}} = 610$. To limit over-fitting, 5% of the training set (30 molecules) were used as a validation set during training and some additional measures were taken (see appendix F for more information). The optimisation of the 62 NNs took one and a half minutes with 60 minimisation iteration on average per NN (∼1.5 s/NN and ∼ 0.02 s/minimisation iteration).

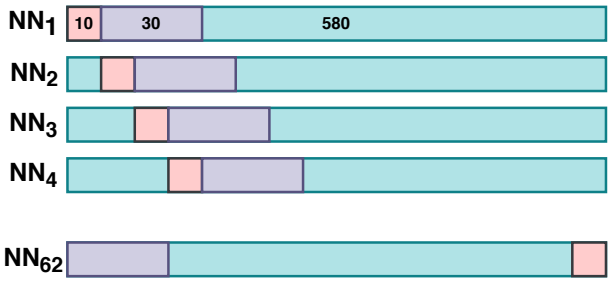


Figure 8.5: Illustration of the test (pink), training (turquoise) and validation (purple) set splitting.

Figure 8.6 shows the MAE, RMSE and R distributions of the optimized 62 NNs for the training and test sets. As expected, the training set errors are lower and correlation is higher than for the test set with $\langle \text{MAE}_{\text{train}} \rangle = 0.36 < \langle \text{MAE}_{\text{test}} \rangle = 0.42$ kcal/mol, $\langle \text{RMSE}_{\text{train}} \rangle = 0.57 <$

$\langle \mathrm{RMSE_{test}} \rangle = 0.65$ kcal/mol and $\langle \mathrm{R_{train}} \rangle = 0.990 > \langle \mathrm{R_{test}} \rangle = 0.987$. The optimisations were run several times with different random variable initiations, all the optimisations gave similar results with less $< 10\%$ of the variation in average statistical measures between different optimisations. The test set distributions are quite asymmetric with long sparse tails to high errors and small correlations, i.e. few NNs are not well optimized for their test sets. Is this due to bad luck in training/test set split, *eg.* all the nine thioesters are in a single test set without any representation in the training set? This seems not to be the case, as we tried multiple different training/test set splits and in all cases a few 'outlier' NNs emerged. For example, the HFE of the 1-amino-4-hydroxy-9,10-anthracenedione (FreeSolvID: 4371692) was systematically not well corrected by the different NNs.

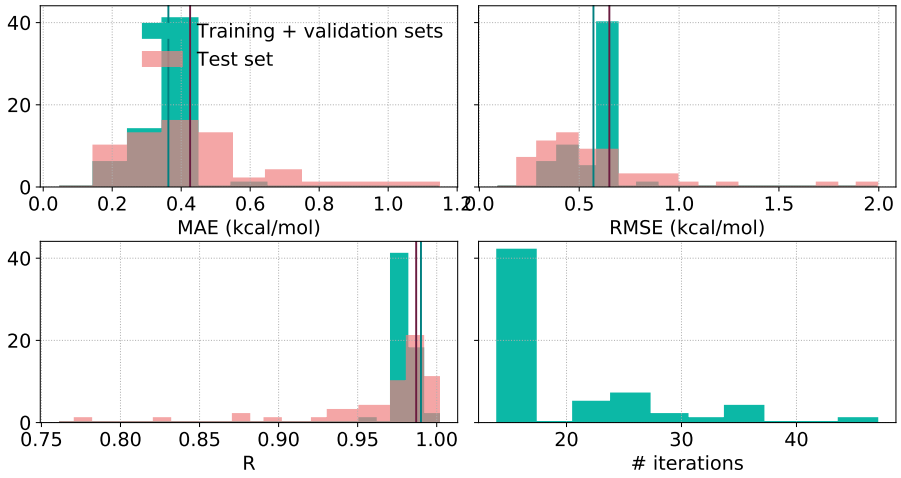


Figure 8.6: Statistical measure (MAE, RMSE and R) distributions for training and test sets and the number minimization steps of the 61 NNs. Vertical lines correspond to the averages over all the NNs (training set in teal and test set in purple).

8.3.4 ML CORRECTED MDFT RESULTS

Figure 8.7, plots the correlation between reference HFEs and those obtained with the ML-corrected MDFT. The final MDFT results for the whole database were obtained by regrouping all the 62 test set's results. The correlation between reference values and ML-corrected MDFT results is high, with $R = \rho = 0.97$ and $\tau = 0.87$. The deviations are almost twice smaller than with $\mathrm{MDFT^{HNC+vdwW}}$, with an MAE of 0.59 kcal/mol ($\sim$ experimental precision) and RMSE of 0.90 kcal/mol compared to 1.07 and 1.49 kcal/mol before. We can conclude, that this simple and fast (the training takes a few minutes and the prediction is instantaneous) ML correction works very well for the FreeSolv database.

However, as any ML model, especially the NNs, are 'black boxes' tasked to do interpolation between specific data of the training set: this model should correctly predict a correction to MDFT HFEs for other small neutral organic molecules but if used for example on small charged organic molecules there is no guaranty at all that it will work. Moreover, as this model uses MACCS fingerprints as input data it can not be used for example inorganic molecules as the MACCS representation was developed for (small) organic compounds. To develop a more 'universal' ML correction the input data format, *i.e.* the solute representation, should be a more general one and the training should be done on the 'full' chemical space of the MDFT calculations.

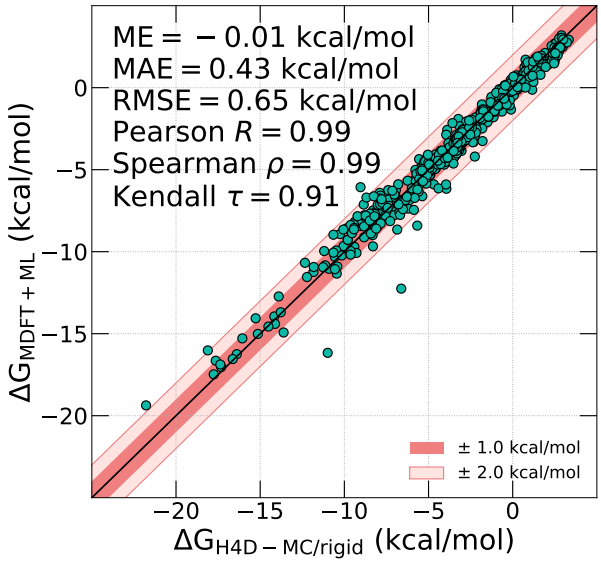


Figure 8.7: Comparison between the reference hydration free energies and those obtained with MDFT-HNC with an ML correction for the FreeSolv database.

8.4 SURFACE TERM

Finally, we propose a third, more rigorous, approach to correct the PC term. As mentioned before the original PC in eq. 8.2 is justified as the dominant term in the macroscopic limit is the pressure term. But at the molecular level as for small cavities or solutes, this correction should also depend on the solute's surface area.

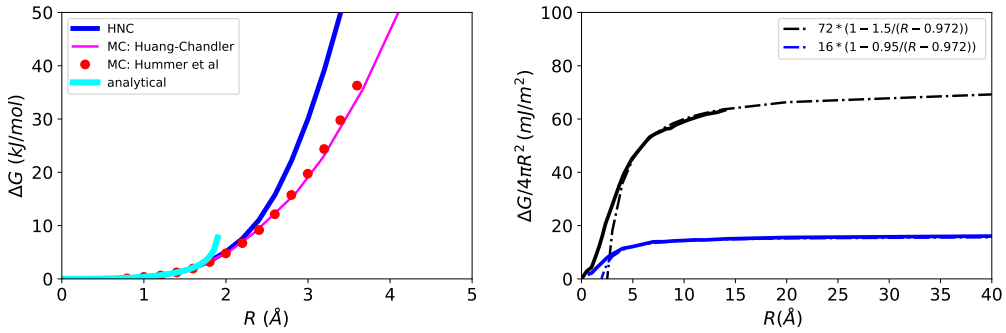
8.4.1 HYDROPHOBIC SPHERES

To study this aspect we started by examining the solvation of hydrophobic spheres, which is the paradigmatic problem for either the standard scaled particle theory (SPT) [174, 175] or more recent advances in the theory of hydrophobicity and hydrophobic interactions [176, 177]. Figure 8.8a shows the hydration free energy of a hard-sphere of increasing radius R computed by MDFT-HNC and compared to the simulation results given by Hummer et al. [178] and Huang and Chandler [176] with the SPC/E water model. Figure 8.8a also includes the analytical limit for cavity volumes that can only accommodate 0 or 1 water molecules, namely [178],

$$\Delta G = -k_{\mathrm{B}}T \ln(1 - n_{\mathrm{bulk}}V), \tag{8.10}$$

where $V = 4\pi R^3/3$ is the hard-sphere volume. It can be seen in fig. 8.8a that MDFT-HNC and simulations fulfil this exact small-radii limit for $R < 1.8$ Å; MDFT-HNC even matches the simulation results slightly beyond that radius and diverges from them afterwards. In fig. 8.8b, we compare the solvation free energy per surface area, computed either by MC by Huang-Chandler as $\Delta G(R)/4\pi R^2$, or by MDFT after pressure correction as $(\mathcal{F}_{\mathrm{min}}(R) - \mathrm{PC}_{\mathrm{PMV}})/4\pi R^2$. Both curves present a horizontal asymptote pointing to the surface tension γ. Simulations yield $\gamma_{\mathrm{sim}} = 72$ mJ/m^2, a value close to the experimental one but somewhat larger than the reported gas-liquid surface tension of SPC/E [179] whereas MDFT-HNC yields the much smaller value $\gamma_{\mathrm{HNC}} = 16$ mJ/m^2: thus not only the HNC pressure has to be corrected but also the surface tension.

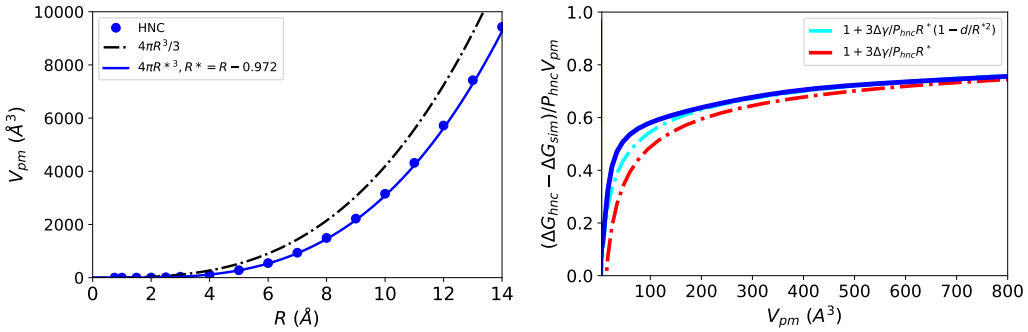
An important question while using SPT is the definition of solute volume and surface to be considered, usually derived from either the solute van der Waals surface (vdW) or the solvent-accessible surface (SAS) illustrated in fig. 2.2. The two of them differ by the extension of the probe sphere, *i.e.* solvent (water) molecule, radius R_w. Here, the question is rather the relationship



(a) (b)

Figure 8.8: (a) Hydration free energy of a hard-sphere (HS) of radius R obtained by MDFT-HNC (blue line) or by MC simulations by Hummer et al. [178] or Huang and Chandler [177] (red bullets and red line, respectively). The line in cyan is the analytical result of eq. 8.10. (b) Hydration free energy per unit area, $\Delta G/4\pi R^2$, as a function of HS radius computed by MC or by MDFT with $\mathrm{PC_{PMV}}$.

between the measured V_{PM} and the hard-sphere volume V. Figure 8.9a, clarifies that relationship. We observe that V_{PM} is optimally fitted by $V_{\mathrm{PM}} = 4\pi R^{*3}/3$ with a shifted radius $R^* = R - R_w$ and $R_w = 0.972$. R^* can be identified to the vdW radius generating a vdW surface of area $S = 4\pi R^{*2}$ rather than the solvent-accessible surface of area $S = 4\pi R^2$.



(a) (b)

Figure 8.9: (a) Partial molar volume V_{PM} vs hard-sphere radius R obtained by MDFT-HNC. It is best fitted by $V_{\mathrm{PM}} = 4\pi R^{*3}$ with $R^* = R - 0.972$. (b) Correction to the HNC solvation free energy normalised by the first-order pressure correction [thus, the quantity $a(R^*)$ defined in eq. 8.12] as a function of the PMV: exact computation (solid blue line), or estimation using eq. 8.13, with the value of $\delta = 0.32$ Å (cyan dashed-dotted curve) or with $\delta = 0$ (red dashed-dotted curve).

Supposing a truncated SPT expression (eq. 2.3) for the hydration free energy with the same R_w for both simulations and MDFT-HNC

$$\Delta G = P\frac{4}{3}\pi R^{*3} + \gamma 4\pi R^{*2}\left(1 - \frac{\delta}{R^*}\right) \tag{8.11}$$

with $R^* = \sqrt[3]{3V_{\mathrm{PM}}/4\pi}$ and accounting for the fact that $P_{\mathrm{sim}} \simeq 0$, one can write a correction to the HNC approximation as

$$\mathrm{PC_{PMV-surf}} = \Delta G_{\mathrm{HNC}} - \Delta G_{\mathrm{sim}} = a(R^*)P_{\mathrm{HNC}}V_{\mathrm{PM}} \tag{8.12}$$

with

$$a(R^*) = 1 + \frac{3\Delta\gamma}{P_{\mathrm{HNC}}R^*}\left(1 - \frac{\delta}{R*}\right) \tag{8.13}$$

and $\Delta\gamma = \gamma_{\mathrm{HNC}} - \gamma_{\mathrm{sim}} = -56$ mJ/m^2. The first term of $a(R^*)$ yields the pure PMV pressure correction $\mathrm{PC}_{\mathrm{PMV}}$ and the second one a surface correction to it; the length parameter δ relates to the so-called curvature correction to the surface tension or, in this case, to the surface tension difference. It can be determined by imposing the condition that $\Delta G_{\mathrm{HNC}} = \Delta G_{\mathrm{sim}}$ for small radii, e.g. for $R^* = 1$ Å ($\Longleftrightarrow$ $R \simeq 2$ Å, see fig. 8.8a). This condition yields $\delta = 1 + P_{\mathrm{HNC}}/3\Delta\gamma = 0.32$ Å. The approximation of eq. 8.13 is compared to the simulation results in fig. 8.9b and fits quite well. Note that it is a parameter-free expression and only $\Delta\gamma$ enters. The simpler approximation with $\delta = 0$ applies only above ~ 500 Å^3.

As shown in this section, the pressure correction proportional to the PMV as we proposed previously is strictly valid for very large solutes of micrometric size. For microscopic to nanoscale solutes, at least a surface correction $\Delta\gamma$, preferably the next correction term in a scaled-particle theory parametrization, should be accounted for. On this simple paradigmatic example, one observes that there is no way that a simple correction strictly proportional to the PMV can be applied unless limited to a small range of PMV.

8.4.2 MOLECULAR SOLUTES

As shown in the previous section, MDFT-HNC fails for what seems to be the simplest case, *i.e.*, estimating the free energy cost of creating cavities. This failure extends to non-polar solutes composed of LJ sites with no partial charges. Fig. 8.11a plots the correlation between the MDFT and H4D-MC results for the hydration free energies ΔG^{LJ} of the solutes without partial charges, when the simple pressure correction $\mathrm{PC}_{\mathrm{PMV}}$ is applied. As for the fully charged solutes (fig. 8.2a), this correction improves greatly the bare results, but it cannot be considered as satisfactory yet, with an MAE at 3.05 kcal/mol and $R = 0.55$.

Similarly to fig. 8.9a for hard spheres, fig. 8.10 plots the "exact" correction factor of eq. 8.13, $\left(\Delta G_{\mathrm{HNC}} - \Delta G_{\mathrm{sim}}\right)/\mathrm{PC}_{\mathrm{PMV}}$, and a fit using the analytical form of eq. 8.13, with $R^* = \sqrt[3]{3V_{\mathrm{PM}}/4\pi}$, thus as for an hypothetical, equivalent spherical solute. Only the parameter $\Delta\gamma$ has to be adjusted since it stands here for an effective value accounting for a mean Lennard–Jones attraction that was not present in the derivation for hard spheres. We find an optimal value $\Delta\gamma = -6.9$ mJ/m^2 For a purely repulsive hard-sphere in TIP3P water, one would expect $\Delta\gamma = -39.2$ mJ/m^2 with $\gamma_{\mathrm{HNC}} = 13.1$ mJ/m^2 and $\gamma_{\mathrm{sim}} = 53.2$ mJ/m^2 from Ref. [179]; the corresponding curve is also presented in fig. 8.10. The figure also represents the horizontal line corresponding to a simpler correction of the form $a * \mathrm{PC}_{\mathrm{PMV}}$, compatible with previous suggestions [161] (eq. 8.4), with an optimal value $a = 0.86$. This type of correction only applies because the range of PMV values that are spanned is relatively small. In contrast to the formula in eq. 8.13, this correction gives an incorrect limit when PMV becomes larger.

Figures 8.11b and 8.11c, show the new correlations between the MDFT-HNC and simulation HFEs with just a pressure correction renormalized by the constant factor 0.86, or applying the more elaborated analytical form of eq. 8.13, which gives a better description of the surface effects and yields the correct large volume limit. From (a)–(c), one goes initially from an MAE of 2.92 kcal/mol and an R of 0.56, to 0.41 kcal/mol and 0.83, and finally to 0.38 kcal/mol and 0.88. Note that this agreement is obtained with a very rude, spherical approximation for the vdW surface area, which could certainly be improved. In particular, the slope of the correlation should be corrected. Note that, previously, an empirical $\mathrm{PC}^{+}_{\mathrm{PMV}}$ was proposed in eq. 8.4. For TIP3P $\left(P_{\mathrm{HNC}} - P_{\mathrm{id}}\right)/P_{\mathrm{HNC}} = 0.86$, which is exactly what we now suggest by introducing surface contributions. With the current understanding, we consider this agreement as satisfying, but fortuitous.

Figure 8.11d shows the correlation between MDFT-HNC and simulation results for the electrostatic contribution of the HFE, *i.e.* $\Delta G^{\mathrm{elec}}_{\mathrm{H4D-MC}} = \Delta G_{\mathrm{H4D-MC}} - \Delta G^{\mathrm{LJ}}_{\mathrm{H4D-MC}}$ and $\mathcal{F}^{\mathrm{elec}}_{\mathrm{min}} = \mathcal{F}_{\mathrm{min}} - \mathcal{F}^{\mathrm{LJ}}_{\mathrm{min}}$, where the first terms are computed for the fully charged solutes. Without any correction, we observe already a very good agreement with an MAE of 0.53 kcal/mol and a correlation of 0.99,

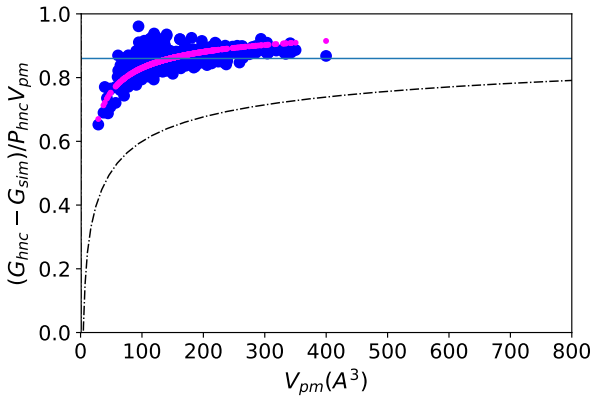


Figure 8.10: Exact correction to HNC normalised by the first order-pressure correction [the quantity $a(R^*)$ in eq. 8.12] for the Lennard-Jones contribution to the SFE as a function of the solutes PMV. Each blue dot represents a molecule in FreeSolv dataset. The dots in magenta correspond to the analytical fit of eq. 8.13, using an effective spherical radius $R^* = \sqrt[3]{3V_{\mathrm{PM}}/4\pi}$. The parameter-free result for a purely repulsive sphere in TIP3P water is given by the dashed line.

but nevertheless, a mean slope of 0.88 instead of 1. We find that the agreement can even be improved to an MAE of 0.26 kcal/mol and a slope of nearly 1 by adding a pressure-like correction $+0.6\Delta\mathrm{PC}_{\mathrm{PMV}} = 0.6P_{\mathrm{HNC}}\Delta V_{\mathrm{PM}}$, involving the difference of the PMV with and without charges $\Delta V_{\mathrm{PM}} = V_{\mathrm{PM}} - V_{\mathrm{PM}}^{\mathrm{LJ}}$; see fig. 8.11e. ΔV_{PM} is always negative, and this new correction goes with an opposite sign with respect to the standard one. It means that the regular pressure/surface tension correction $\mathrm{PC}_{\mathrm{PMV-surf}}$ above, roughly $-0.86\mathrm{PC}_{\mathrm{PMV}}$, is overcompensated by electrostatic effects that we do not yet fully understand; this correction remains empirical at this stage.

Overall, fig. 8.11f displays the final correlation results adding both the Lennard–Jones and electrostatic contributions. For each solute, this requires two independent minimizations, with and without the solute partial charges. Reducing the parametrization to its minimum to capture the correct physics, *i.e.*, a single parameter $\Delta\gamma$ correcting the pressure correction by surface effects, and no empirical correction of the electrostatics, yields an MAE of 0.66 kcal/mol and an R of 0.99 with a mean slope of 0.85. This is significantly better than with the initial values obtained with an original pressure correction with an MAE of 2.14 kcal/mol and R of 0.95 with a slope of 0.72. Accounting for the full story reported above, *i.e.* incorporating in addition to the well-justified one-parameter correction for the LJ contribution, another one-parameter correction for the electrostatic contribution yields an MAE of 0.44 kcal/mol, a correlation of 0.99, and a correlation slope close to 1.

8.5 RECAPITULATION

Table 8.3 shows that all three developments of the PC give similar results and improves the quality MDFT-HNC results compared to the simple PMV correction incompressible fluids $\mathrm{PC}'_{\mathrm{PMV}}$, and table 8.4 summarises the advantages and disadvantages of each approach.

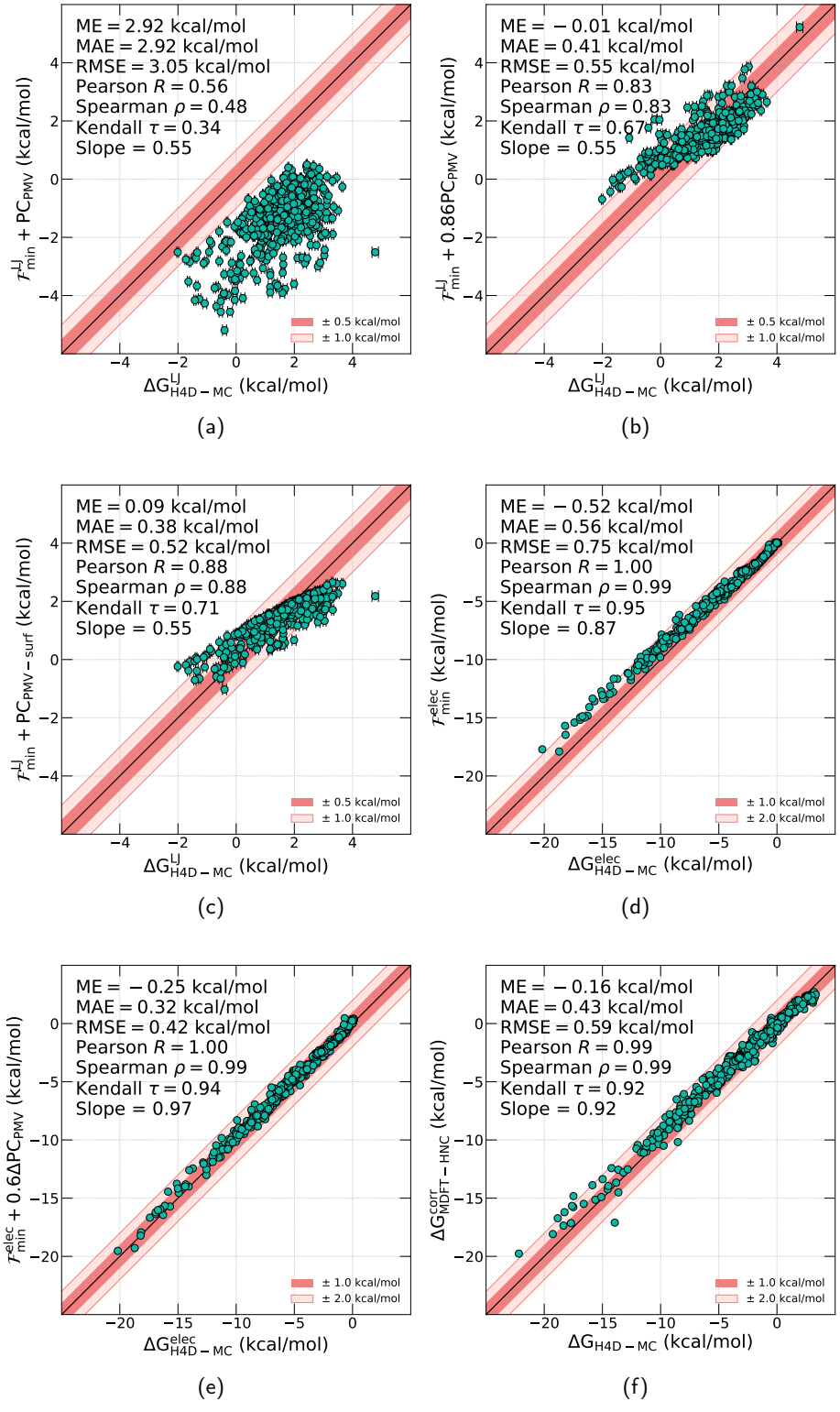


Figure 8.11: Comparison between HFEs obtained with MDFT-HNC and H4D-MC simulations for the FreeSolv database : (a,b,c) correspond to the MDFT-HNC results for non-polar part (LJ) with (a) the original $\mathrm{PC}_{\mathrm{PMV}}$, (b) a renormalized pressure correction $0.86\mathrm{PC}_{\mathrm{PMV}}$, and (c) the surface corrected $\mathrm{PC}_{\mathrm{PMV-surf}}$ of eq. 8.13. (d) corresponds to the MDFT-HNC result for electrostatic SFE, (e) to the same plus an empirical pressure-like correction $0.6\Delta\mathrm{PC}_{\mathrm{PMV}}$, and (f) to the total HFE as the sum of (c) and (e).

	$\mathcal{F}_{\text{min}} + \text{PC}'_{\text{PMV}}$	$\mathcal{F}_{\text{min}} + \text{PC}_{\text{vdW}}$	$\mathcal{F}_{\text{min}} + \text{PC}_{\text{ML}}$	$\mathcal{F}_{\text{min}} + \text{PC}_{\text{PMV}-\text{surf}}$
MAE	2.14 ± 0.10	0.46 ± 0.03	0.43 ± 0.04	0.45 ± 0.03
RMSE	2.47 ± 0.10	0.58 ± 0.04	0.65 ± 0.07	0.61 ± 0.06
ME	1.94 ± 0.12	0.33 ± 0.04	-0.01 ± 0.04	-0.18 ± 0.04
Pearson's R	0.95 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
Spearman's ρ	0.94 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.02
Kendall's τ	0.79 ± 0.03	0.92 ± 0.01	0.91 ± 0.03	0.92 ± 0.03

Table 8.3: Summary of the statistical measures characterizing the correlations between HFEs obtained from H4D-MC and MDFT-HNC calculations (i) with the original pressure correction for compressible fluids PC'_{PMV} , (ii) with the van der Waals volume pressure correction PC_{vdW}, (iii) with the machine learning pressure correctionPC_{ML} and (iv) with the surface corrected pressure correction $\text{PC}_{\text{PMV}-\text{surf}}$ for the FreeSolv database. All energies in kcal/mol.

	Pros	Cons
PC_{vdW}	No extra cost	Only for organic molecules
PC_{ML}	No extra cost	Only for neutral organic molecules
$\text{PC}_{\text{PMV}-\text{surf}}$	Applicable to all solutes (see next chapter)	Needs two MDFT minimisations

Table 8.4: Summary of the pros and cons of three new pressure corrections.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. MDFT could offer a compromise between speed and accuracy. However, the bare MDFT-HNC hydration free energy predictions are in the woods. We know that most of this error is due to a bad estimation of the system's pressure. Can we correct this pressure estimation?

Here, we proposed three approaches to improve the original pressure correction proposed by the group. All three approaches improved the MDFT results for the FreeSolv database. Of the three approaches, the addition of a surface term leads to best results and can be applied to any type of solute. However, as it demands two MDFT calculations instead of a singles one, we recommend the use of the vdW volume-based correction for organic solutes as it improves the original results without increasing the computation time.

9

MDFT-HNC BENCHMARK

Why this chapter?

The aim of this thesis is to compute hydration free energies accurately but very efficiently. MDFT could offer a compromise between speed and accuracy. Is MDFT-HNC well-suited for any type of solute? Is the HNC approximation with an appropriate pressure correction good enough?

In this chapter, we compare MDFT$^{\text{HNC+PMV-surf}}$ results rigorously to exact single conformer simulation results for four types of solutes :

- hydrophobic spheres
- monovalent ions
- water as solute
- small organic molecules

This chapter benchmarks the performance of MDFT-HNC with new pressure-surface correction $\text{PC}_{\text{PMV}-\text{surf}}$, the best correction to date. This is done by a rigorous comparison of MDFT results, solvation free energies and solvation profiles, to state-of-the-art H4D-MC simulation results for a multitude of systems, with the same force field parameters and same fixed solute geometries: from simple spheric hydrophobic solutes and spherical ions to small organic molecules.

All computations were done with the TIP3P water model for the solvent. For water in MDFT, we find in general that $n_{\max} = 3$, corresponding to 84 orientations per grid point, gives sufficient accuracy compared to higher-order expansions, *e.g.*, $n_{\max} = 5$ corresponding to 330 orientations. Most of the calculations presented below were performed with $n_{\max} = 5$ just for safety as this is completely affordable for the relatively small solutes that were considered. Most of the results presented in this chapter were published in Luukkonen *et al.* [180].

9.1 SPHERICAL SOLUTES

9.1.1 HYDROPHOBIC SOLUTES

We started with simple one-site hydrophobic solutes: rare gases and united-atom representation of methane and neopentane. H4D-MC simulations were performed with a box of 100 water molecules and with the same simulation parameters determined in 5.1. MDFT calculations are done with a solute embedded in a cubic supercell of length 24 Å, with periodic boundary conditions, a spatial resolution of 0.25 Å (= 96x96x96 grid nodes) and an angular resolution of 330 ($n_{\max} = 5$) orientations per spacial grid node. The force field parameters of these hydrophobic solutes are described in table 9.1.

Table 9.1 also includes their experimental HFEs and those obtained with H4D-MC simulations and MDFT-HNC. The correlation between the computed HFEs is plotted in fig. 9.1a. The agreement between MDFT and simulations is good with a small MAE between the methods of 0.23 kcal/mol

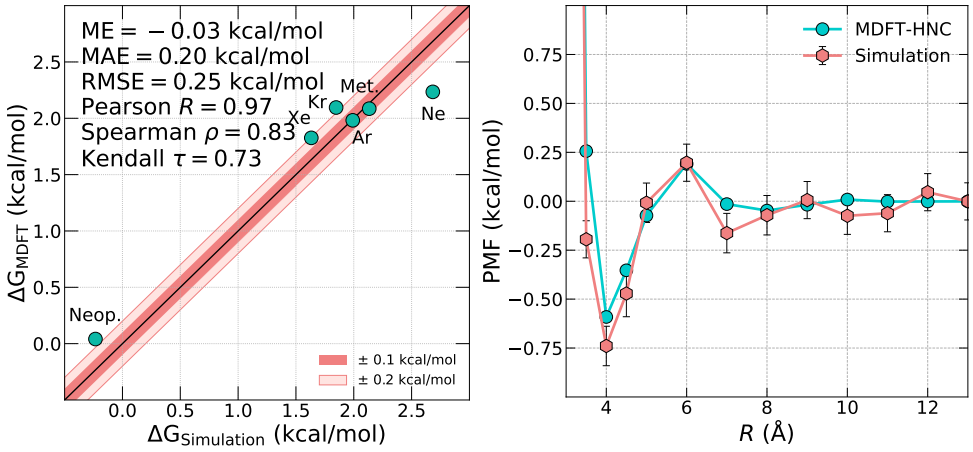
			Hydration free energy [kcal/mol]		
Solute	σ [Å]	ϵ [kJ/mol]	ΔG_{Exp} [150]	ΔG_{sim}	ΔG_{MDFT}
Neon	3.035	0.15432	2.48	2.68 ± 0.02	2.23
Argon	3.415	1.03931	1.99	1.99 ± 0.01	1.98
Krypton	3.675	1.40510	1.66	1.79 ± 0.01	2.06
Xenon	3.975	1.78510	1.45	1.59 ± 0.01	1.83
Methane	3.730	1.23000		2.04 ± 0.01	1.96
Neopentane	6.150	3.49000		-0.33 ± 0.01	0.09

Table 9.1: Lennard-Jones force field parameters and hydration free energies of rare gases and unified methane and neopentane molecules.

and a high linear correlation with $R = 0.97$. We can also go beyond the HFEs and compute, for example, potentials of mean forces (PMFs) between two solutes via

$$\text{PMF}(R) = \Delta G^{\text{AB}}_{\text{solv}}(R) - \Delta G^{\text{AB}}_{\text{solv}}(\infty) + U^{\text{AB}}(R) \tag{9.1}$$

where $\Delta G^{\text{AB}}_{\text{solv}}(R)$ is the SFE and $U^{\text{AB}}(R)$ the direct interaction of the AB pair separated by a distance R. Figure 9.1b shows the PMF obtained with MDFT ($n_{\max} = 5$, $L = 43.75$ Å and $dx = 0.25$ Å) and H4D-MC (400 water molecules) between two unified-atom methane solutes. MDFT captures correctly the first minimum and maximum but not quite the subsequent fluctuations in the PMF.



(a) (b)

Figure 9.1: Comparison of (a) HFEs of spherical hydrophobic solutes and (a) the PMF between two methane atoms obtained with MDFT and simulation.

Figure 9.2 compares the solute-solvent water site-site distribution profiles of the united-atom methane and neopentane obtained with MDFT and simulations. As already evidenced in the past [38], the HNC approximation predicts correctly the cavity volume and the rising of the first peak. Its characteristic feature for small hydrophobic solutes is to slightly displace the first peak and overestimate its height. The peak location is better for the larger solute but the height overestimation remains. Overall, however, the approximation is also doing fine on the structure.

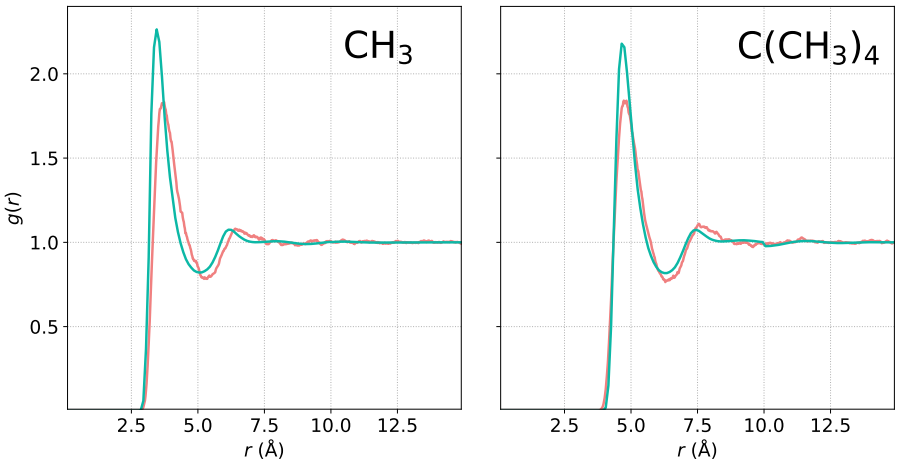


Figure 9.2: Radial solvation profiles obtained with MDFT-HNC (turquoise) and simulations (purple) for two hydrophobic Lennard-Jones solutes.

9.1.2 MONOVALENT IONS

Here, we study the solvation of simple monovalent ions: four anions, F^-, Cl^-, Br^-, I^-, and four cations, Li^+, Na^+, K^+, Cs^+, described with force field parameters given by Horinek *et al.* [151]. Those are recapitulated in Table 9.2. H4D-MC simulations were performed with a box of 400 water molecules and with the same simulation parameters determined in 5.1. MDFT results were obtained from two calculations with the solute, one with partial charges and one without, embedded in a cubic supercell of length 32 Å, with periodic boundary conditions, a spatial resolution of 0.25 Å (= 128x128x128 grid nodes) and an angular resolution of 330 ($n_{\max} = 5$) orientations per spacial grid node. In both cases, since the calculations are done for a periodic system, two types of correction have been applied, of the so-called B and C types (see sec. 4.2.3, [138, 139, 40])

Ion	σ [Å]	ϵ [kJ/mol]	ΔG_{sim}	ΔG_{MDFT}	$\Delta\Delta G_{\text{sim}}$	$\Delta\Delta G_{\text{MDFT}}$
F^-	3.434	4.654×10^{-1}	-95.6 ± 0.1	-95.0	-27.13	-24.6
Cl^-	4.394	4.160×10^{-1}	-68.6 ± 0.1	-70.3	- -	- -
Br^-	4.834	2.106×10^{-1}	-62.2 ± 0.1	-65.4	6.48	4.9
I^-	5.334	1.575×10^{-1}	-54.0 ± 0.1	-58.9	14.49	11.4
Li^+	2.874	6.154×10^{-4}	-133.2 ± 0.2	-132.5	-201.65	-202.9
Na^+	3.874	*idem*	-108.6 ± 0.1	-104.8	-176.89	-175.1
K^+	4.543	*idem*	-92.8 ± 0.1	-91.3	-161.31	-161.6
Cs^+	5.173	*idem*	-81.6 ± 0.1	-81.9	-145.99	-152.2

Table 9.2: Force field parameters and hydration free energies (in kcal/mol) of monovalent ions.

Table 9.2 reports HFEs obtained with MDFT and reference simulations and figures 9.3a and 9.3b shows the correlation between MDFT and simulation for the absolute solvation free energies ΔG and the relative solvation free energies as defined in Ref. [151], $\Delta\Delta G = \Delta G + z\Delta G(\text{Cl}^-)$, with $z = \pm 1$ according to the ion valence and $\Delta G(\text{Cl}^-)$ the value obtained for Cl^-. The latter was the reference free energies that Horinek *et al.* used to fit their ions force field parameters since it cancels the somehow uncontrolled surface charge corrections that should be added when comparing to experimental values. These figures include the MDFT results with the original $\text{PC}_{\text{PMV}}$to illustrate

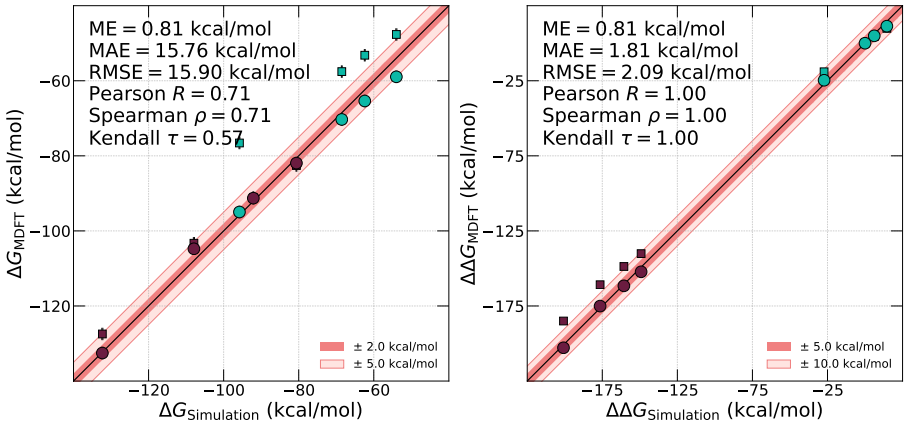


Figure 9.3: Comparison of (a) absolute and (b) relative HFEs of monovalent ions obtained with MDFT-HNC and simulations. Cations in purple and anions in turquoise. Small squares correspond to results with $\mathrm{PC}_{\mathrm{PMV}}$ and larger spheres to results with $\mathrm{PC}_{\mathrm{PMV-surf}}$.

that the new $\mathrm{PC}_{\mathrm{PMV-surf}}$, originally developed for neutral solutes, improve the MDFT-HNC results for charged solutes too. Overall, MDFT-HNC results are quite good for small cations with an MAE of 1.66 kcal/mol (and of 4.00 without the surface correction) but not so well for the anions with an MAE of 2.58 kcal/mol (and of 11.30 without the surface correction).

A primary output of MDFT is the full molecular equilibrium solvent structure $g(\boldsymbol{r},\omega)$ from which one can derive easily radial $g(r)$ and polarization $P(r)$ distribution functions or any other angular-dependent density distribution. Figures 9.4a and 9.4b present the radial density and polarization distributions for all the ions. Concerning the $g(r)$'s, MDFT-HNC clearly performs better for the cations than for the anions. For the cations, MDFT correctly predicts the position of the first two maxima and first minimum. For the smallest cation, Li+, MDFT sightly overestimates the intensities of the maxima, and for the larger cations, MDFT slightly underestimates the relative intensities of the maxima and the minimum. This effect increases with the cation size. In the case of the anions, the $g(r)$ predicted with MDFT deviates much more from the simulation results. The position of the first peak and its width are correct. As for the second peak, it is displaced to larger distances. Since the position of the second peak in water is a sign of tetrahedral order, the cation here taking the place of one water molecule, we concluded before that the HNC approximation is missing here some tetrahedral order. For the polarization radial distributions $P(r)$, the correspondence between simulations and MDFT is much better both for cations and anions, with some differences in the intensities of the minima and maxima, but globally an excellent agreement.

Beyond the traditional computation of the atomic pair distribution functions, MDFT has the great advantage of providing, in addition, the complete information on the orientations of the water molecules around the solute. Here, in spherical symmetry, this translates to the knowledge of the angular-dependent density maps $g(r,\cos\theta',\psi')$, where θ' is the angle between the dipole direction of one water molecule at position r from the ion and $\boldsymbol{r}$ itself, ψ' the rotation angle around the dipole direction. $g(r,\cos\theta',\psi')$ is easily deduced from the full distribution in laboratory frame, $g(\boldsymbol{r},\omega)$, by a spherical average over all r-orientations.

In fig. 9.5a and 9.5b, we have concatenated all this information into the 2D-plots of $g(r,\cos\theta') = \frac{1}{2\pi}\int \mathrm{d}\psi' g(r,\cos\theta',\psi')$, indicating the preferred orientation of the solvent dipoles as a function of the radial distance and $g(\cos\theta',\psi') \equiv g(r_{\max},\cos\theta',\psi')$, where $r_{\max}$ is the distance corresponding to the maximum of the radial density distribution $g(r)$; this indicates the preferred orientation of the hydrogen atoms as a function of the dipole orientation. The plots are given for both Cl^- and Na^+ and compare MDFT to simulations. As can be seen, the agreement is again excellent for the

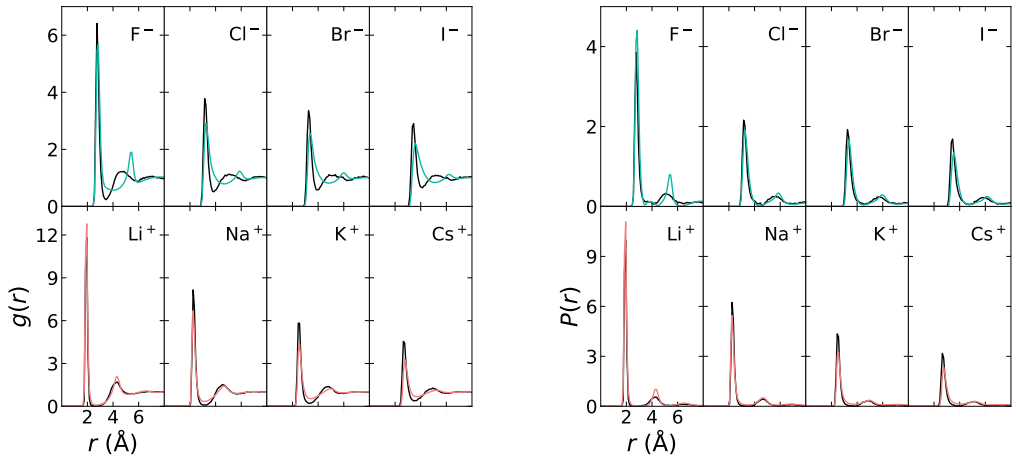


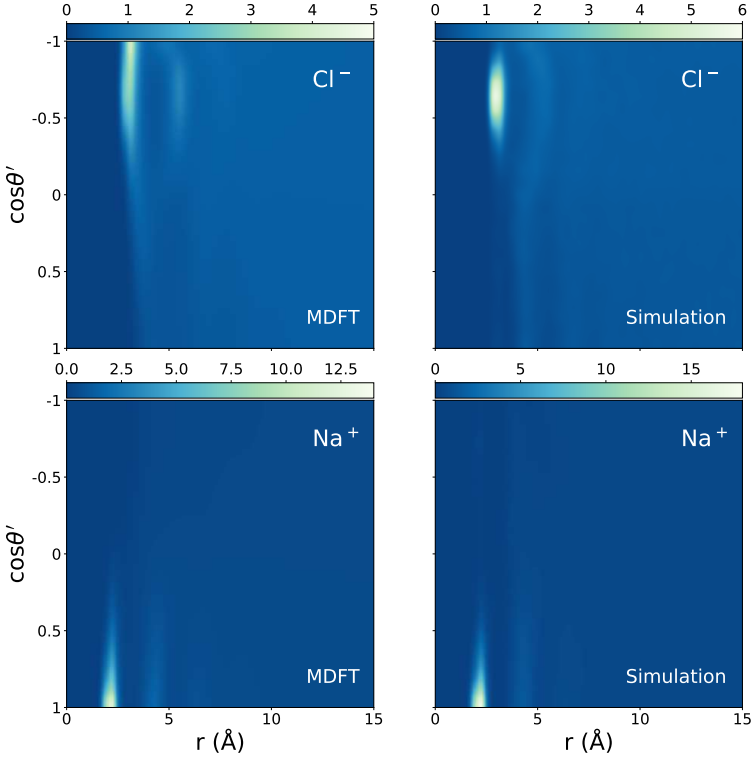
Figure 9.4: Radial (a) density and (b) polarization distribution around ions. Simulation results in black and MDFT results in turquoise (anions) and pink (cations).

cation. At the peak of $g(r)$, the water dipole is directed radially away from the cation ($\cos\theta' = 1$), *i.e.*, with the oxygen closest to the cation and the hydrogen atoms pointing away symmetrically, with no angular dependency in ψ' close to $\cos\theta' = 1$: the hydrogen sites rotate freely around the dipole axis. For values of $\cos\theta'$ departing from 1, a distribution in ψ' appears around $\psi' = \pi/2$. In our conventions, $\psi' = 0$ or π, corresponds to the water molecule in the plane formed by the dipole direction and ion-oxygen direction. The value $\pi/2$ corresponds to a configuration in which the two hydrogen atoms become equidistant from the cation, thus maximizing the sum of the two distances. As for the anion, the $g(\cos\theta', \psi')$ maps look quite similar for MDFT and simulations and display a peak centred around $\cos\theta' \simeq -0.58 = \cos(\pi - \theta_0/2)$, where θ_0 is the HOH angle of the TIP3P model, and $\psi' = 0$ or π in order to have the optimal H-bond to the anion. So far so good, but as seen before, it cannot be perfect: a difference does appear in the $g(r, \cos\theta')$ map in which the peak in simulation appears consistently at cos $\theta\prime = -0.58$ as before and extends roughly between -0.4 and-0.8, whereas in MDFT, it extends more floppily from -0.4 and -1, with its maximum at -1. The strength and directionality of the O–H–X$^-$ bond are clearly underestimated. The second peak is displaced and is somewhat narrower in angle and more pronounced.

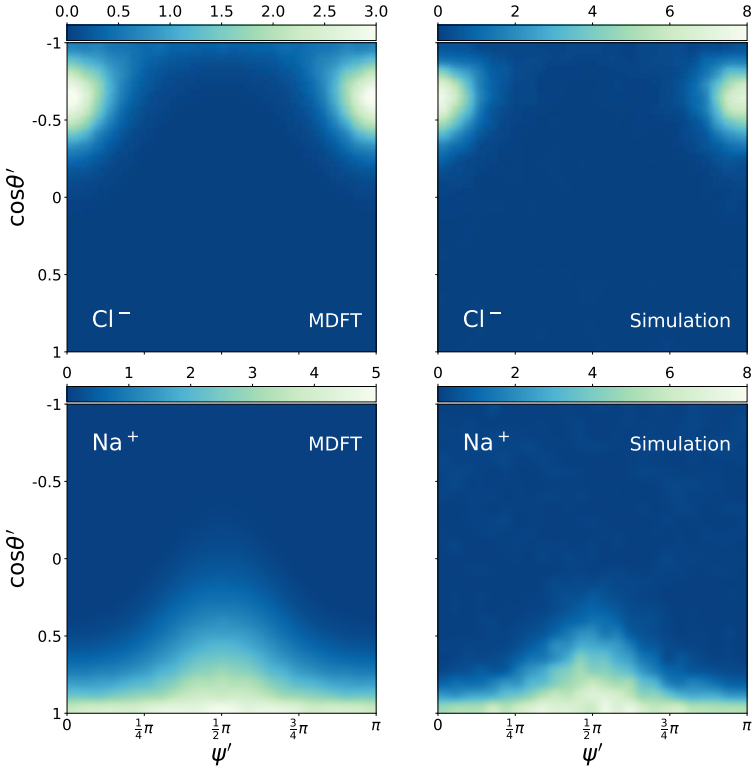
9.2 MOLECULAR SOLUTES

9.2.1 WATER AS SOLUTE

For the molecular solutes, we begin by the case of a TIP3P water molecule in TIP3P, a paradigm for both an H-bond acceptor and donor. The MDFT-HNC result was obtained from two calculations, one with the solute partial charges and one without, within a cubic supercell of side 24 Å, a spatial resolution of 0.25 Å (=96x96x96 grid nodes), and $n_{\max} = 5$. H4D-MC results were obtained with a box of 100 solvent water molecules and with the simulation parameters determined in 5.1. For the hydration energy of an additional water molecule in water (namely, the chemical potential of TIP3P water), MDFT-HNC predicts a value of -6.3 kcal/mol (-4.69 without the surface correction, a large part of the correction comes from the ΔV_{PM}) which is in excellent accord with the value given by simulation at -6.04 ± 0.07 kcal/mol.



(a)



(b)

Figure 9.5: Two-dimensional maps of $g(r, \cos\theta')$ and $g(\psi', \cos\theta')$ computed by simulations or by MDFT for Cl− (top) or Na+ (bottom).

Figure 9.6a shows the radial site-site pair distribution functions between the solute oxygen and hydrogen sites and the solvent oxygen and hydrogen sites obtained from the MC simulation and MDFT-HNC. Here, we recover the equivalent results obtained already 20 years ago by Richardi *et al.* [181] and Lombardero *et al.* [182] using 1D-MOZ-HNC integral equations for both TIP3P and SPC/E waters. Indeed, the same deficiencies of HNC appear: it does miss some of the (subtle) tetrahedral symmetry in water. The the first O–O peak is correctly placed but too wide on its right side; the second peak is misplaced and appears at a position pertinent to the second neighbour in a general dipolar fluid, and not at the 4.6 Å value imposed by the tetrahedral symmetry. The first O–H or H–O, peak is also at the correct position but underestimated. The H–H pair distribution function appears almost structureless in MDFT-HNC.

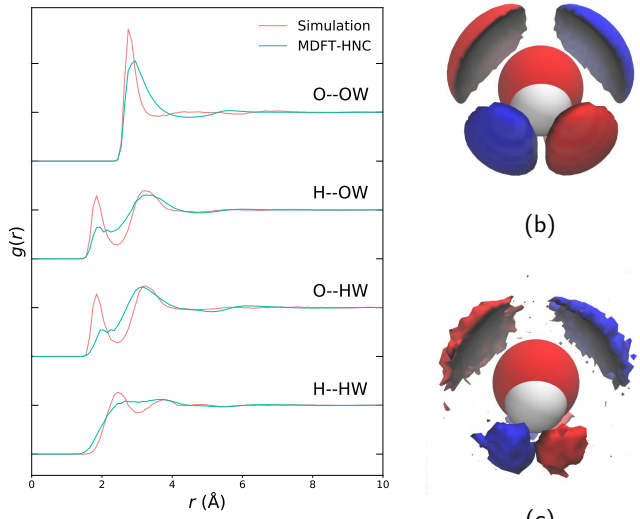
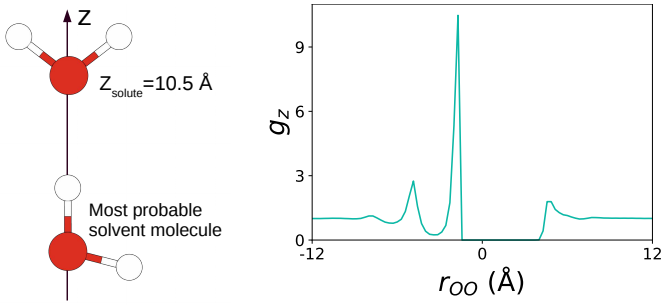


Figure 9.6: (a) TIP3P site-site radial distribution function. Isosurfaces for the polarization density $P_y(\boldsymbol{r})$ (red and blue for positive and negative, respectively) computed (b) by MDFT with a voxel size of 0.25 Å and (c) from a 50 ns long MD simulation with identical voxel size.

One can extract also the three-dimensional solvent charge densities, easily with MDFT, more painfully by simulation since one needs to explore three-dimensional space with sufficient statistics. This is illustrated in figs. 9.6b and 9.6c. They represent the isosurfaces of the 3D-polarization density $P_y(\boldsymbol{r})$, where y is the axis perpendicular to the molecular plane. 9.6b shows the isosurfaces $P_y(\boldsymbol{r}) = \pm 0.035$ D/Å^{3}obtained by MDFT with a grid size of 0.25 Å, whereas 9.6c shows the same quantity obtained by collecting histograms of identical voxel size along a 50-ns-long MD trajectory (25 000 independent configurations). These 3D plots look familiar compared to previous simulations [143] with a change of sign when crossing the symmetry plane. The two rather loose, upper caps correspond to the solvent donor molecules presenting their hydrogen atoms to the solute oxygen negative partial charge. The two lowest ones represent the solvent water molecule presenting its oxygen to the hydrogen site pointing in the figure, and whose orientation can depart from the average, symmetric one with the two hydrogen atoms pointing away and a vanishing P_y. Beyond the satisfactory agreement between MDFT and simulation, the noise appearing in the MD results illustrates the statistical difficulty of accumulating 3D-densities by simulation, not to speak of position and orientation densities, which are the direct output of MDFT.

Finally, we present in fig. 9.7b a feature that is not accessible to RISM-based approaches and would require intense statistics in simulations: we plot the probability of finding a water molecule in a fixed orientation at distance z from another, here the most probable orientation for a O–H–O bond on the positive side, $z < 0$, which becomes an H-bond mismatch on the other side of the donating molecule, for $z > 0$.



(a) (b)

Figure 9.7: (a) Representation of a solute water fixed in the centre of the box and a second solvent one at its most probable location ($r_{OO} = 2.45$ Å) and in its optimal angular configuration for H-bonding. (b) The probability distribution for a water molecule keeping the same fixed orientation as in d. and gliding along the z-axis.

9.2.2 FREESOLV DATABASE

As described already in part in chapter 8, for a more systematic study of molecular solutes, we assessed the performance of MDFT-HNC on the FreeSolv database of small drug-like molecules. As reference, we use rigid solute H4D-MC simulation results presented in 5.1. MDFT results were obtained from two calculations of solute, one with and one without partial charges, in the same initial FreeSolv configurations within a supercell of side 32 Å with a spatial resolution of 0.33 Å (=96x96x96 grid nodes) and angular resolution of 84 orientations per spatial grid node ($n_{\max} = 3$). The MDFT calculations did not converge for 22 molecules (3% of the database, see appendix G.1 for the solutes that did not converge). All results presented below are for the 620 molecules that led to convergence. The computational cost or average computation time per molecule on a single CPU was 8 min—we usually use 8 CPU-threads, so ~1 min per solute in real-time. Note that we could have done the calculation in a box of 21 Å, and hence, have had only 64 nodes in each direction, for the vast majority of the molecules in the database, decreasing the simulation time to under two minutes on average.

Figure 9.8 shows the comparison between simulation and MDFT results for the FreeSolv database with the whole $\mathrm{PC}_{\mathrm{PMV-surf}}$ included. The correlation between the two methods is almost perfect with correlation coefficients close to 1, with an average deviation below half a kcal/mol (MAE=0.44 kcal/mol) and negligible bias (ME=0.12 kcal/mol). MDFT-HNC with the $\mathrm{PC}_{\mathrm{PMV-surf}}$ predicts HFEs of small neutral molecular solutes at the same accuracy as reference simulations for a speedup of at least 3 orders of magnitude in computation time.

Additionally, we illustrate the capacity of MDFT-HNC to predict solvation profiles around a molecule. It is illusory to span the whole database. We have chosen for illustration the case of quinoline [FreeSolv ID: mobley_5857, a typical molecule of the FreeSolv database, whose chemical structure presented in fig. 9.10]. The 3D-solvation structure obtained by MDFT is represented in fig. 9.9: 9.9a displays the number density in the plane of the molecule, with the associated alternation of maxima and minima. 9.9b concerns another important quantity embedded in $g(\boldsymbol{r}, \omega)$ that is, the polarization field $P(\boldsymbol{r})$. It displays the norm of the polarization field, *i.e.*, $||P(\boldsymbol{r})||$, in the plane of the quinoline molecule obtained with MDFT-HNC. As expected, we find high polarization close to the sites wearing localized charges, and the expected polarization with OH pointing toward N. 9.9c illustrates the solvation isosurface of $n = 3n_{\mathrm{bulk}}$ around the solute.

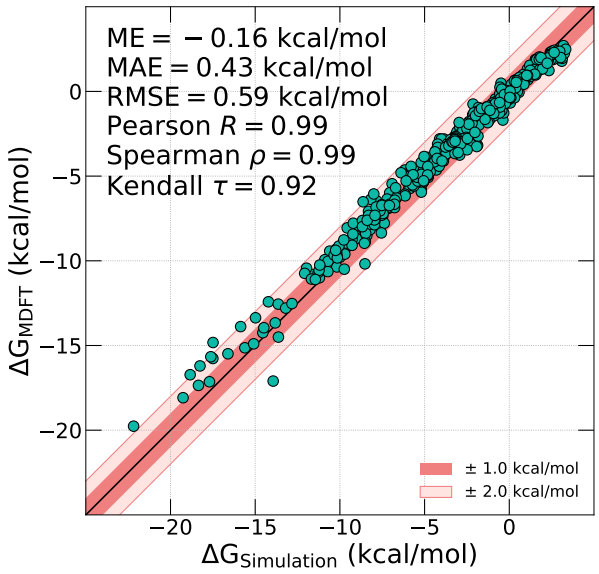
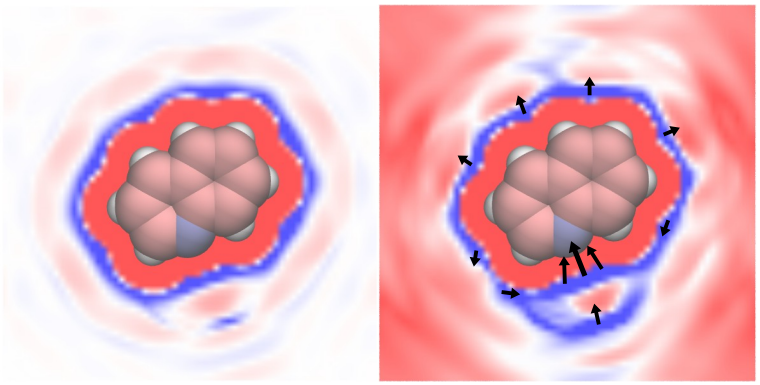
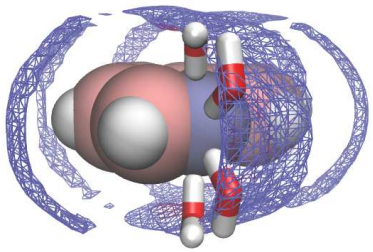


Figure 9.8: Correlation between HFEs obtained with single conformer H4D-MC and MDFT-HNC with $\mathrm{PC}_{\mathrm{PMV-surf}}$ for the FreeSolv database.



(a) (b)



(c)

Figure 9.9: (a) Water density map (red: $n < n_{\mathrm{bulk}}$, white: $n = n_{\mathrm{bulk}}$ and blue: $n > n_{\mathrm{bulk}}$) and (b) norm of the polarization vector field (blue: high polarization, black arrows representing the orientation) in the plane of the molecule obtained with MDFT-HNC. (c) Representation of the quinoline molecule with four most probable water molecules and the water isosurface at $n = 3n_{\mathrm{bulk}}$.

A direct comparison to simulation results is made for the site radial distribution functions in fig. 9.10. The agreement appears very reasonable. For all solute atoms, the rise of the first peak follows exactly that of the simulation: the shape of the cavity is perfectly reproduced. The maximum of the first peak, if any, is correctly located, meaning that the first solvation shell lies where it should. For

the carbon sites exposed to the solvent (e.g., C1, C6, C8, C9), one does recover the overestimation of the height that was found for hydrophobic solutes in fig. 9.2. The $g(r)$ for the nitrogen site misses the important H-bond first peak. Since the nitrogen atom wears a relatively high partial charge of $-0.65e$, one is back to the problem encountered before for strong negative charges, *e.g.*, for anions or the oxygen of the water molecule.

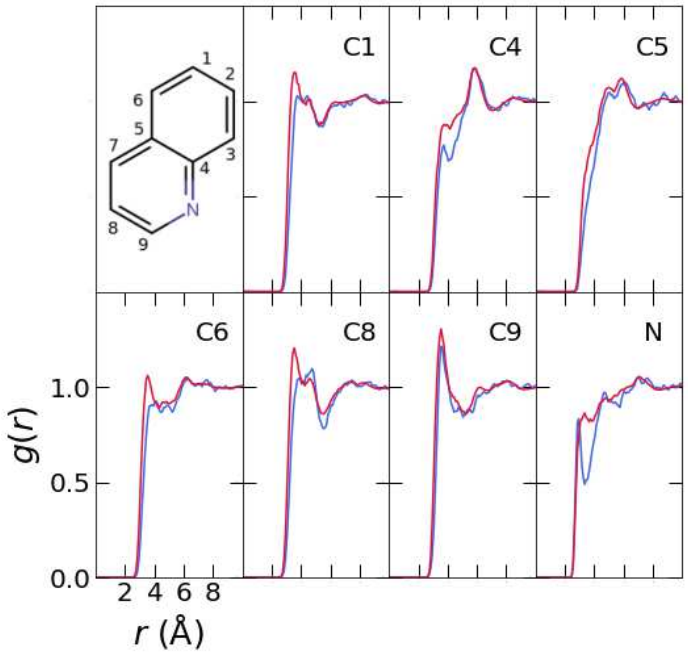


Figure 9.10: Chemical structure of the quinoline and the radial distribution function $g(r)$ between the heavy atoms of the quinoline and water oxygens. Blue lines correspond to MDFT-HNC as obtained in a few minutes, and red lines to MD simulations as obtained in a few hours.

Finally, from $g(\boldsymbol{r},\omega)$ one can also obtain so-called water maps catching the most probable water molecule positions and orientations around the solute. Figure 9.9c shows the four most probable position and the orientation of water molecules around the quinoline. Expectedly, they are found close to the nitrogen atom.

To remember

The aim of this thesis is to compute hydration free energies accurately but very efficiently. Is MDFT-HNC well-suited for any type of solute? Is the HNC approximation with an appropriate pressure correction good enough?

When compared to single conformer simulations, MDFT-HNC with predicts HFEs of neutral solutes within 1 kcal/mol and monovalent ions within 5 kcal/mol. For the solvation profiles, MDFT-HNC finds the general form but typically overestimates the first peak around neutral (or low charged), and underestimates and spreads the first peak around a (highly) charged site. MDFT-HNC also misses part of the hydrogen bonding around a water molecule and strongly negative charges.

10

MDFT FOR DRUG–LIKE MOLECULES

Why this chapter?

The aim of this thesis is to compute hydration free energies of *drug-like molecules* accurately but very efficiently. We know that MDFT-HNC when corrected with an appropriate pressure correction predict hydration free energies of drug-like molecules within 1 kcal/mol. But for which types of molecules does it work especially well and for which does it struggle more?

To get the insight we

- compare MDFT-HNC results with $\mathrm{PC}^{\mathrm{vdW}}$ to experimental data and state-of-the-art simulations (the reference in HFE calculations) as well as to 3D-RISM, a competing liquid-state theory
- identify solute features for which MDFT performance well and for which it struggles

This chapter focuses on a deeper chemoinformatics analysis of the MDFT results for the FreeSolv database to assess MDFT-HNC's capacity for drug design. As discussed in the introduction there is a need for fast but precise prediction of solvation free energies for the process of drug design. MDFT could be an answer for this demand as MDFT-HNC with rigorous $\mathrm{PC}_{\mathrm{PMV-surf}}$ *a posteriori* correction has similar accuracy as time-consuming reference simulations, as shown in the previous chapter. However, $\mathrm{PC}_{\mathrm{PMV-surf}}$ requires two MDFT calculations per solute and the 8 cpu.min computation time per solute in the previous section is still somewhat expensive for the evaluation of large databases. Therefore, we re-performed the MDFT-HNC calculations for FreeSolv database with the $\mathrm{PC}_{\mathrm{vdW}}$ as it requires only one MDFT minimization per solute, within a supercell of length 21 Å, a spatial resolution of 0.33 Å (=64x64x64 grid nodes) and an angular resolution of 84 orientations per spatial grid node ($n_{\max} = 3$). In this case, the average computation time on a single CPU-thread was 1 min 53 sec. The MDFT minimization process did not converge for 23 solutes (4% database, see appendix G for more information). In chapter 6, we found that there are 520 molecules for which the difference between a rigorous single conformer and flexible solute H4D-MC simulation is below 0.6 kcal/mol, the average experimental precision of the database. These molecules are defined as rigid, in respect of their hydration free energy, and as MDFT does a single conformer calculation, the following analysis is performed on this sub-set of 520 quasi-rigid solutes. Most of the content of this chapter will be published in Luukkonen *et al.* [183].

Firstly, MDFT results will be compared to state-of-the-art MD+FEP simulations and 3D-RISM, another LST approach gaining success in the last few years. The following sections include an error analysis on selected features of the drug-like molecules in order to identify MDFT's strengths and weaknesses and to be able to infer error bars on the method.

Figure 10.1a shows the correlation between experimental HFEs and those obtained with $\mathrm{MDFT}^{\mathrm{HNC+vdW}}$. The MAE is 0.92±0.07 kcal/mol and the Pearson's correlation coefficient R is 0.93±0.01. MDFT results also have a small mean (signed) error of -0.07±0.11 kcal/mol which indicates that MDFT does not have a systematic bias: it is lower in amplitude than the statistical error bars. All the statistical measures characterizing this correlation are summarized in table 10.1.

The error bars on the measures correspond to the 95% confidence interval [1]. Note that, as we are comparing MDFT$^{\text{HNC+vdW}}$, an approached theory, to experimental data, the deviations could be the results of incorrect approximations in MDFT$^{\text{HNC+vdW}}$ or due to bad force field parametrisation.

Table 10.1 contains also the statistical measures characterizing the correlations between experimental values and those obtained with MD+FEP (fig. 10.1b) given by Duarte Ramos Matos *et al.* [144] and those obtained with 3D-RISM-KH (fig. 10.1c) by Roy and Kovalenko [114]. Overall the three methods perform at the same accuracy level with similar errors and correlation coefficients. However, MDFT's computation time is on average less than 2 CPU.min compared to hundreds of CPU.h or tens gpu.h with MD+FEP and few tens cpu.min with 3D-RISM[2]. Hence for the same accuracy, MDFT has a speed-up of 1-2 and 3-4 orders of magnitude, when compared to 3D-RISM and MD+FEP respectively. Compared to 3D-RISM, MDFT does not have the consequences from approximating MOZ [185, 186, 187, 90].

	MD+FEP$^{(a)}$	3D-RISM$^{(b)}$	MDFT$^{\text{HNC+vdW}}$
MAE	0.98 ± 0.07	1.04 ± 0.09	0.92 ± 0.07
RMSE	1.29 ± 0.11	1.45 ± 0.11	1.25 ± 0.11
ME	-0.40 ± 0.11	-0.19 ± 0.11	-0.07 ± 0.11
Max. error	4.57	7.11	4.82
Pearson's R	0.94 ± 0.02	0.91 ± 0.02	0.93 ± 0.01
Spearman's ρ	0.94 ± 0.01	0.89 ± 0.03	0.93 ± 0.02
Kendall's τ	0.79 ± 0.02	0.73 ± 0.03	0.78 ± 0.03
cpu.h per solute	$\sim 10^{2}$	$\sim 10^{-1}$	$\sim 10^{-2}$

Table 10.1: Summary of the statistical measures characterizing the correlations between experimental HFEs and those obtained with simulation-based free energy techniques, 3D-RISM-KH and MDFT-HNC calculations for a sub-set of rigid molecules (520). All error measures are given in kcal/mol and all error bars correspond to the 95% confidence interval. *(a)* Duarte Ramos Matos *et al.* [144]. *(b)* Roy and Kovalenko [114].

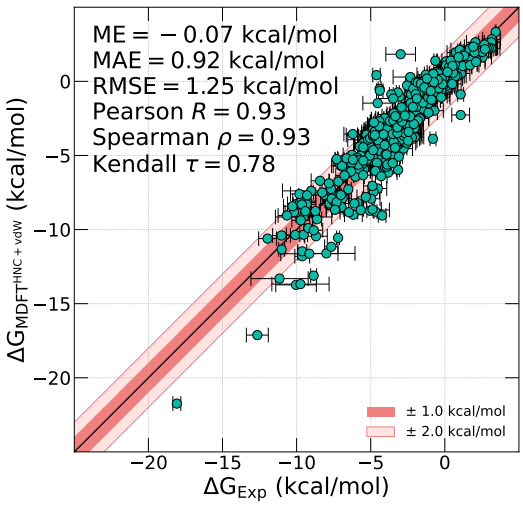
10.1 EFFECT OF SOLUTE'S MASS, CHARGES AND SOLVATION STRUCTURE

In order to give an optimal set of requirements and confidence intervals to MDFT-HNC predictions, we now focus on finding sources of errors or correlations between errors. Figure 10.2 shows the error distribution as a function of the solute's *(i)* molar mass, *(ii)* largest partial charge $\max\{q_i\}$ and *(iii)* highest value of the 3D solvation structure $\max\{g(\boldsymbol{r})\}$. As shown in figure 10.2a, the heaviest molecules have the largest deviations to experimental values: the MAE increases with the solute's mass. For solutes with a molar mass larger than 200 Da, the MAE is 1.75 kcal/mol, *i.e.* almost the double than for the whole database. However, these molecules present only 12% of the rigid subset so their effect on the total MAE is not significant as seen on the cumulative MAE. Similar trends are present also for the MD+FEP and RISM results with an MAE of 1.78 and 2.21 kcal/mol respectively for these molecules heavier than 200 Da (see fig. H.1)

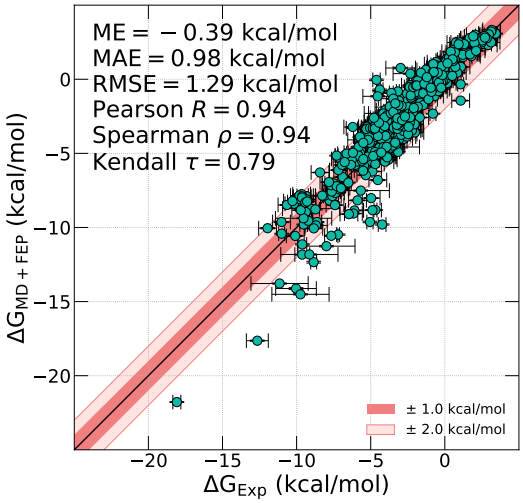
Similarly to the molar mass, the deviation to experimental values increases with the magnitude of the largest partial charge of the drug-like molecule, positive or negative, (see fig. 10.2b) with an

1 For each statistical measure X characterizing a dataset of N points (eg. $N = 619$ for the full set), the measure X' was computed 10 000 times on N values chosen at random each iteration from the dataset. The error bars of X correspond to two standard deviations of the X' distribution.

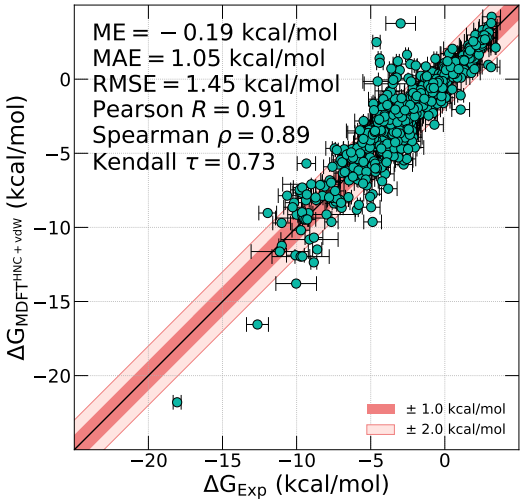
2 3D-RISM computation times were recovered from a 2010 paper [184] as the 2019 paper [114] did not discuss computation times



(a)



(b)



(c)

Figure 10.1: Comparison of experimental HFEs and those obtained with (a) MDFT-HNC with $\mathrm{PC}_{\mathrm{vdW}}$, (b) MD+FEP simulations and (c) 3D-RISM for a sub-set of rigid solutes (520).

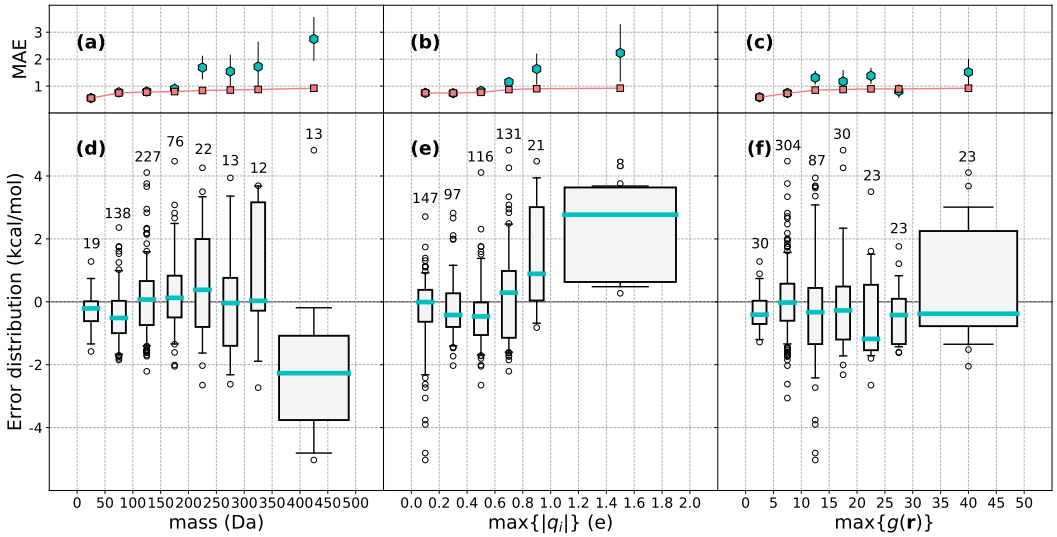


Figure 10.2: Distribution of mean absolute error between predicted and experimental hydration free energies as a function of three features (a) solute molar mass (with a bin size of 50 Da), (b) solute's largest local charge (0.2 e) and (c) the maximum of the 3D $g(\boldsymbol{r})$ around the solute (5) with turquoise hexagons (error bars correspond to the 95% confidence interval) presenting the MAE of each bin and pink squares the cumulative MAE. The corresponding signed error distributions are presented in (d), (e) and (f) with turquoise lines corresponding to the median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. The number above each bin is the population of the bin for the FreeSolv database and for each distribution the last three (350-500 Da), five (1-2 e) or four (30-50) bins are gathered into one to be statistically significant.

MAE of 1.84 kcal/mol for solutes with $\max\{q_i\} > 0.8e$ (6% of the rigid sub-set). The effect is less pronounced for MD+FEP and RISM with MAEs at 1.55 and 1.50 kcal/mol respectively for these molecules. This is expected for MDFT at the HNC approximation: the second-order density expansion of the functional around $\rho = \rho_{\text{bulk}}$, or $g = 1$, misses higher order repulsion terms. This leads to problems for cases with densities getting away from ρ_{bulk}: either high densities typically found next to high (partial) charges or large solutes with large volumes where $g = 0$.

Besides the solute's molar mass and partial charges, solute features known *a priori*, we can also look at the output of an MDFT calculation, that is, the solvation profile, to predict, on this dataset at least, the quality of the MDFT's HFE predictions. Figure 10.3 illustrates the 3D solvent density around 1-amino-4-hydroxy-9,10-anthracenedione (FreeSolvID: 4371692) with four water-oxygen density iso-surfaces ($g = 0.5, 2.5, 5.0$ and 7.5). Low densities (fig. 10.3a) are observed on the limits of the solute's cavity but also after the first solvation peak (fig. 10.3d) of the hydroxyl group. The largest oxygen densities (fig. 10.3d) are observed next to the hydroxyl-hydrogen and the less crowded amine-hydrogen that are potential hydrogen-bond donors.

In fig. 10.2c, one can see that the MDFT's deviation to experiment increases with the maximum height of the solvation peaks with an MAE of 1.24 kcal/mol for solutes with $\max\{g(\boldsymbol{r})\} > 20$ (1% of the rigid sub-set). This result is expected as high-density peaks are difficult cases for the HNC approximation as discussed in the previous paragraph. However, the link between the amplitude of the deviation and the solvation structure is less pronounced as for the solute's mass and partial charges.

Figure 10.4 shows two-dimensional cross distributions of MAE for the three features studied above. Often solutes with high mass and high charges/solvation peaks have the largest deviations but deviations can be large for molecules with only one feature with a 'high' value (*eg.* MAE of 3.05kcal/mol for solutes with $\max\{g(\boldsymbol{r})\} = 10 - 15$ and $\max\{q_i\} > 1.0e$ in fig. 10.4c). The smallest deviations from experimental values are found for the solutes with a mass lower than 200

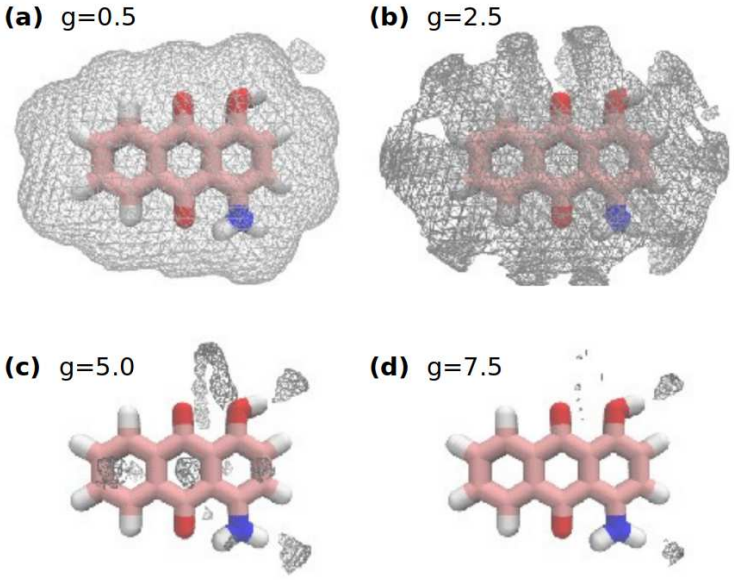


Figure 10.3: Water density isosurfaces around 1-amino-4-hydroxy-9,10-anthracenedione at (a) $g = 0.5$, (b) 2.5, (c) 5.0 and (d) 7.5.

Da, the largest partial charge of lower than 0.8 and highest solvation peak lower than 25, delimited by the turquoise rectangles in figure 10.4 (73% of the database), with an MAE at 0.73±0.22 kcal/mol. A table of the three-dimensional cross distributions of MAE is given in appendix I.

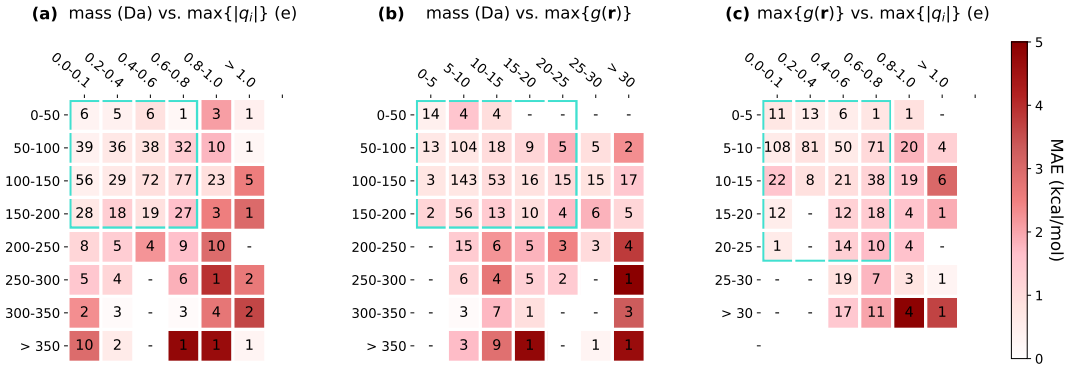


Figure 10.4: Distribution of mean absolute error between predicted and experimental hydration free energies as a function of two features: (a) solute's mass and largest partial charge, (b) solute's mass and highest solvation peak and (c) solute's highest solvation peak and largest partial charge. The number in each bin is the population of the bin.

10.2 EFFECT OF FUNCTIONAL GROUPS

This section assesses the performance of MDFT as a function of the chemical groups present in a solute. Figure 10.5 shows the error distribution of MDFT$^{\text{HNC+vdW}}$and reference MD+FEP as a function of each chemical function present in at least five molecules of the database.

We observe a high correlation between the MAEs of MDFT$^{\text{HNC+vdW}}$ and MD+FEP ($R = 0.90$ and $\rho = 0.81$). In general, functional groups with small/large errors with MD+FEP also have small/large errors with MDFT$^{\text{HNC+vdW}}$. This indicates that the major part of MDFT's error comes from the force field parametrization and not the approximated theory itself. This is not unexpected since it has been shown, in the previous chapters [180, 164], that MDFT with appropriate partial molar volume corrections reproduces similar SFE's with an accuracy of $k_{\text{B}}T$ or below. For example,

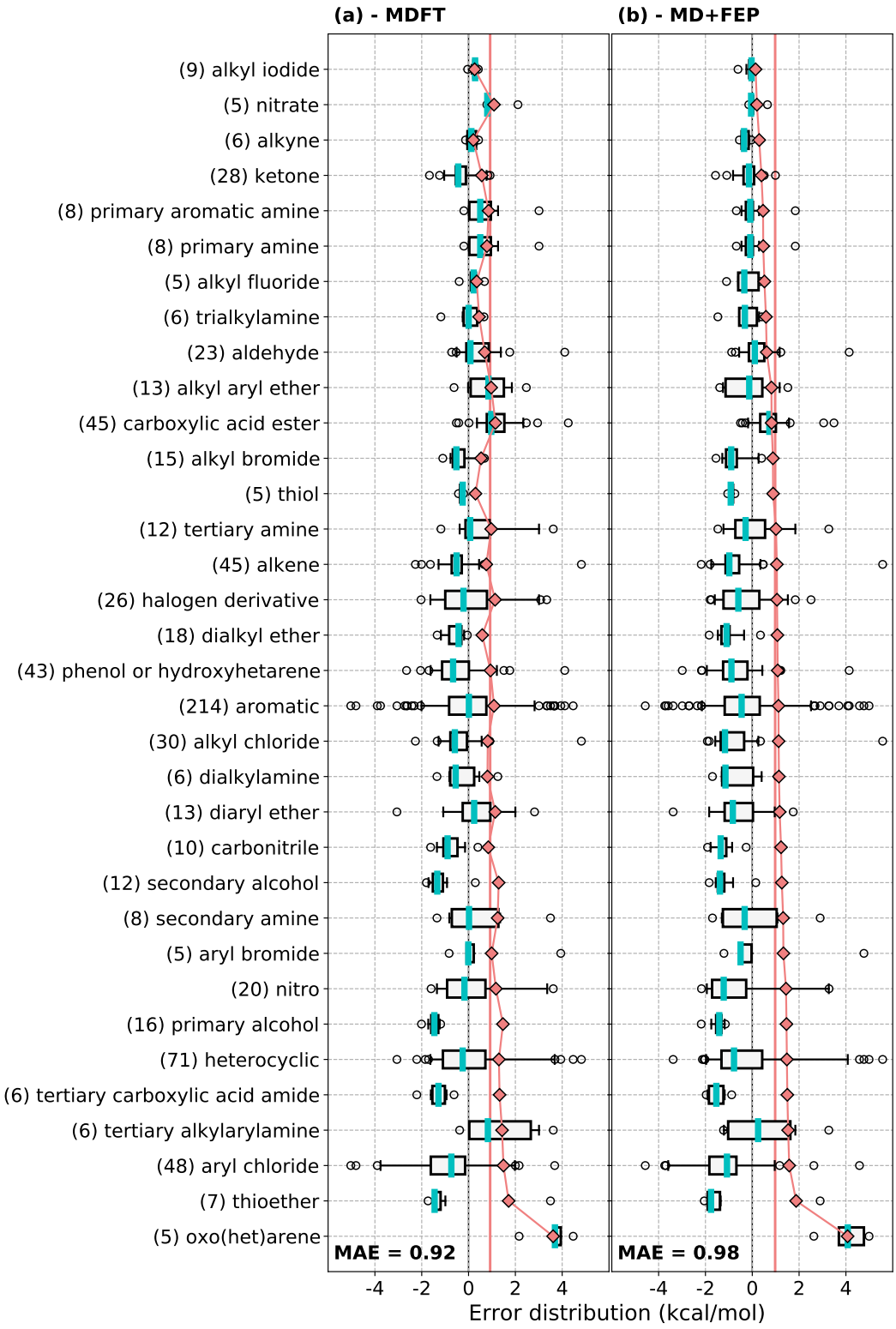


Figure 10.5: Error distributions of (a) MDFT$^{\text{HNC+vdW}}$ and (b) MD+FEP for the chemical groups with more than 5 solutes present in the database. The number of molecules in each group is written within parenthesis. Turquoise lines correspond to the median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. Pink diamonds correspond to the MAE of each functional group and the vertical pink line to total MAE of the rigid subset.

it has been noted that the GAFF parametrization of the hydroxyl groups leads to systematic errors in HFEs computed with MD+FEP [188]. Here we observe above average MAEs for primary and secondary alcohols with systematic underestimation of the HFEs (ME < 0 with a narrow distribution of errors) for both MD+FEP and MDFT$^{\text{HNC+vdW}}$.

Nonetheless, there are differences between MDFT's and MD+FEP's MAEs: MDFT$^{\text{HNC+vdW}}$ significantly over-performs for some groups, like thiols, and under-performs for others, like nitrates, when compared to MD+FEP. Hence the totality of MDFT's error cannot be attributed to the force field parametrisation.

Additionally, we did a similar cross-analysis between chemical functions, as for the mass, partial charge and solvation peak couples. A table of all error bars reconstructed from this analysis is given in appendix I. To illustrate these error estimates, fig. 10.6 shows the distribution of MAE for molecules with an aromatic ring, the most frequent chemical function in the database (present in 214 solutes, i.e. 41%of the rigid sub-set), coupled with another chemical group. We see that, in most cases, the MAE of an aromatic+another group is close to the overall MAE of the aromatics.

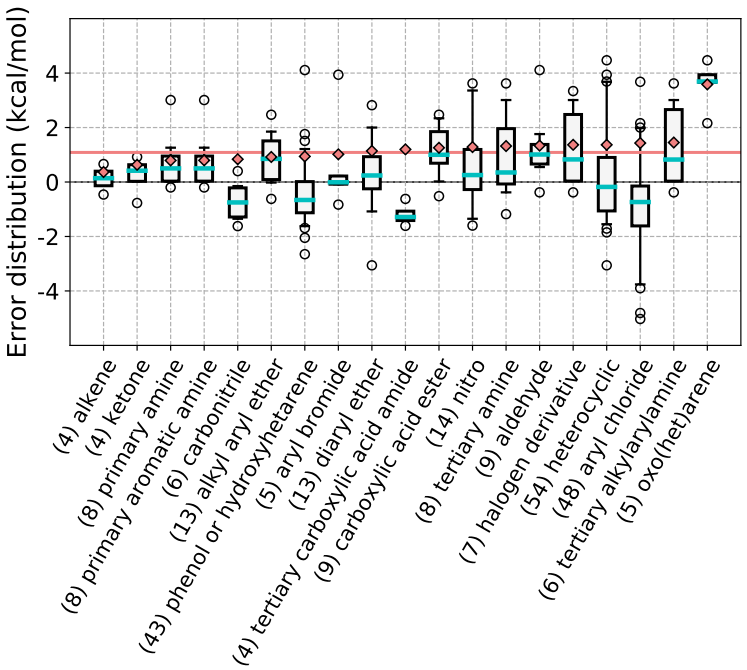


Figure 10.6: Error distributions of solutes with an aromatic ring coupled to another chemical function for couples present in more than five solutes. The number of molecules in each couple is written within parenthesis. Turquoise lines correspond to the median error of each couple, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. Pink diamonds correspond to the MAE of each functional group and the vertical pink line at 1.18 kcal/mol corresponds to the MAE of all the molecules containing an aromatic ring.

The most notable exception is the MAE of aromatic+oxo(het)arene at 3.58 kcal/mol which is much higher than the MAE of all aromatics at 1.08 kcal/mol. This is coherent with oxo(het)arenes having the largest errors of all functional groups. More interesting are couples like aromatic+alkene (MAE=0.37 kcal/mol) or aromatic+aldehyde (MAE=1.34 kcal/mol) for which the MAEs of the couples are significantly lower or higher than the MAE of the individual chemical functions that they are composed of. Note that these couples contain only 2 and 4 solutes each so these behaviours might be artefacts of limited sampling.

To remember

The aim of this thesis is to compute hydration free energies of *drug-like molecules* accurately but very efficiently. We know that MDFT-HNC when corrected with an appropriate pressure correction predict hydration free energies of drug-like molecules within 1 kcal/mol. But for which types of molecules does it work especially well and for which does it struqgle more?

For rigid drug-like molecules, MDFT predicts HFEs with an MAE of 0.92 kcal/mol with an average computation time of 2 minutes per solute. This is at the same level of accuracy as state-of-the-art MD+FEP or 3D-RISM for a speedup of 3-4 or 1-2 of orders magnitude in computation, respectively.
Additionally, we have identified solute features for which MDFT performance well or struggles. They are generally the same as those for MD+FEP and comes from the force-field deficiencies. For certain substances and on average, MDFT overperforms MD+FEP this is due to compensation of errors rather than fundamental reasons.

CONCLUSION - PART III

MDFT-HNC with an improved pressure correction, by the addition of a surface and charging terms or by van der Waals volume optimisation, yield HFEs of drug-like molecules within half a kcal/mol. The former is more general and can be applied to any type of solute but the charging term demands two MDFT minimisations, whereas the latter does not increase the computation time but can only be applied for atoms with tabulated values.

The rigorous benchmark of MDFT-HNC with $\mathrm{PC}_{\mathrm{PMV-surf}}$ showed that predicts the HFEs of neutral solute, hydrophobic spheres or molecular solutes, within half a kcal/mol and small monovalent ions within one kcal/mol (5% in relative error). As the solvation profiles, MDFT-HNC finds the general form but typically overestimates the first peak around neutral (or low charged), and underestimates and spreads the first peak around a (highly) charged site. MDFT-HNC also misses part of the hydrogen bonding around a water molecule and strongly negative charges.

The final comparison MDFT-HNC with $\mathrm{PC}_{\mathrm{vdW}}$, as speed is a key factor for drug design purposes, HFEs of small organic molecules to experimental results shows that the MDFT predictions are within 1 kcal/mol for an average computation time of 2 minutes per solute. This is at the same level of accuracy as state-of-the-art MD+FEP or 3D-RISM for a speed up of 3-4 or 1-2 of orders magnitude in computation, respectively.

Part IV

CONCLUSION AND PERSPECTIVES

11

CONCLUSION

The development of a new drug, from drug discovery to commercialisation, is a very expensive and long process that takes on average over ten years with a median cost of one billion dollars. Accelerating this process is a major societal challenge. One of the main approaches adopted is to develop and increase the use of fast *in silico* methods for the screening of potential drug candidates in the drug discovery phase.

The first stage of screening is the hit identification, where only the affinity, in solution, between the potential drug molecule and the target is considered. This affinity is closely related to the hydration free energy of the ligand, the target and the complex. Other important properties are the selectivity of the ligand, its water-solubility and possibly its capacity to penetrate the cell membrane. Today the exact simulation-based methods to predict solvation free energies are very slow and the fast continuum models are inaccurate.

Molecular density functional theory is an exact theory that permits the energetical and structural study of the solvation of rigid solute of any size or shape. However, similarly to its more well-known electronic cousin, the exact functional is not known and therefore the theory must be approximated. Nevertheless, a high-performance implementation of MDFT, in the hyper-netted chain approximation with an *a posteriori* pressure correction to compensate some deficiencies of the HNC approximation, predicts reasonable solvation free energies, within few kcal/mol, and molecular solvation equilibrium profiles for a variety of small solutes in few minutes. It is not accurate enough but a first step in the right direction.

In 2017, important theoretical and numerical advances were taken by the group to make the rigorous and efficient resolution of MDFT-HNC possible. Therefore, the first part of my thesis was to produce a large and systematic benchmark of the performance of MDFT-HNC on a variety of systems. At the first stage, MDFT-HNC, which is an approximated theory, should be compared to exact reference simulations, using the same Hamiltonian, to measure the advantages and deficiencies of MDFT-HNC theory without having *eg.* error compensation of the force field approximation. For simple spherical systems or a rigid molecule, like water, this means to use the same force field parameters for the MDFT calculation and the MD+FEP simulation. But for molecules, even small ones, it also means having rigid solute reference data done with the same single conformer used for the MDFT calculation.

Thus, we needed to have single conformer reference data of drug-like molecules. These reference data were produced with a novel hybrid-4th-dimension-Monte Carlo approach. H4D-MC was originally developed for grand canonical simulations with a possible addition or deletion of rigid particles during a simulation at imposed chemical potential μ. The addition or deletion is done via a short out-of-equilibrium MD simulation in a 4th dimension which allows the relaxation of surrounding medium and increases drastically the statistics of particle addition/deletion compared to the basic Widom method. It was modified and developed to perform simulations in the isotherm-isobar ensemble and to measure the excess chemical potential μ_{exc}, *i.e.* the solvation free energy of a solute. With this approach, we (re)produced these reference calculations for (charged) spherical solutes and single conformer small neutral drug-like molecules of the FreeSolv database to which MDFT can be rigorously compared. These rigid H4D-MC HFEs were produced for a similar precision as flexible solute MD+FEP for a speed up of factor 6 in the CPU-time. To validate our single

conformer HFEs of molecular solutes we implemented solute flexibility into H4D-MC. We recovered the flexible solute MD+FEP results for a speed up of factor 4 in the CPU-time.

The comparison of between single conformer, multi-conformer and flexible solute HFE calculations of the FreeSolv database allowed us to identify (i) solute features, mainly potential hydrogen bond donors and acceptors, that guide the importance of solute flexibility in a HFE calculation and (ii) a sub-set of FreeSolv for which a single conformer HFE calculation is enough to produce the HFE of the flexible solute.

The comparison MDFT-HNC results with the original pressure correction that depends only on the solute volume to single conformer reference data showed advantages and limitations: it predicts the HFEs of drug-like molecules within 2 kcal/mol of the reference data with an average computation time of 2 min per solute. There is a speed up 3-4 orders of magnitude when compared to simulations but the method stays quite inaccurate. The addition of *a posteriori* machine learning brings the average error to half a kcal/mol. This inspired us two improve the pressure correction with two approaches: (i) by optimising the van der Waals volume of the solute and (ii) by a more physics-based approach of adding a surface and charging term to the pressure correction. Both corrections also bring the average error to half a kcal/mol which starts to be promising for the screening processes.

A new comparison of MDFT-HNC results with the more physics-based surface correction to reference data of a variety of solutes showed that MDFT-HNC can predict HFEs of neutral solutes with an accuracy of half a kcal/mol, including the HFE of the always challenging water molecule within 0.2 kcal/mol, and monovalent ions within 5 kcal/mol (5% relative error). Note that the charging correction demands two MDFT minimisations doubling the computation time.

Finally, as MDFT could be of great interest for the pharmaceutical application we paid a closer look to the MDFT's performance on drug-like molecules. We showed that MDFT-HNC, with the vdW volume PC, predicts HFEs of small rigid drug-like molecules within 1 kcal/mol of experimental results with an average computation time of 2 minutes per solute. This is at the same level of accuracy as state-of-the-art MD+FEP or 3D-RISM, another liquid state theory approach, for a speed-up of 3-4 or 1-2 of orders magnitude in computation time, respectively. Additionally, with a chemoinformatics analysis, we identified solute features for which MDFT performs well or struggles. Most of the time, MDFT struggles for the same types of molecules as MD+FEP and thus most of the errors certainly come from force field deficiencies. For certain substances and on average, MDFT over-performs MD+FE. This is due to the compensation of errors rather than fundamental reasons.

Overall, we have shown MDFT-HNC with an appropriate pressure correction predicts well the hydration free energies of small rigid neutral or charged systems. The next steps would be (i) on a fundamental level to go beyond the HNC approximation with the addition of bridge functional to bypass the pressure correction and also improve the solvation structure, and (ii) in the biochemical and pharmaceutical interest to introduce smart solute flexibility into MDFT and go to larger systems, like proteins.

12 PERSPECTIVES

12.1 FOR H4D-MC

The 'novel' hybrid-4th dimension-MC approach showed great potential for the computation hydration free energies of small, rigid or flexible, neutral or charged, solutes. At present, the main disadvantages of this approach come from its numerical implementation as it is currently (i) only implemented for simple rigid solvents and (ii) the code is not parallelised. The former is not a problem as long as studying hydration as most water models are rigid, but would cause problems if one wants to use more complicated organic solvents, like cyclohexane and n-octanol (fig. 12.2). However, implementing flexible solvents should not be a problem as developments made in chapter 6 for the solute can be applied for the solvent.

The latter point is more crucial if one wants to use H4D-MC systematically and efficiently free energy calculation method. As the in-house 'test' code does not include any parallelisation, it is not very efficient compared to well-developed commercial or open-source MC or MD codes. Furthermore, as discussed in chapter 5.1.1, the H4D-MC approach is embarrassingly parallel by nature, which is not the case of the MD+FEP approach, and parallelisation could be implemented at three different levels:

1. Each insertion/destruction process is independent, therefore every insertion/destruction could be done on separate core and we can say it is 'infinitely' parallelisable.
2. The propagation of the bulk solvent and solvated system are done with MC, i.e. non-deterministic, therefore the propagation could be done with multiple shorter simulations on separate cores.
3. 'Classic' parallelisation methods used by well developed MC or MD codes could be applied to the simulation box and thus running the propagations and the ins/des on multiple cores.

The first point conceptually the simplest to implement but by far the most ambitious as it gives access towards even exascale computing.

12.2 FOR MDFT

For the last decade the idea, the theory and the code of MDFT has been in development. Now that MDFT-HNC with an appropriate pressure correction can predict hydration free energies of molecular systems within 0.5 kcal/mol and within 1 kcal/mol of experimental data in a few minutes, it is operational. Thus, the main perspective for MDFT is to apply it!

Nonetheless, the theory and the code are not perfect: the theory could be developed beyond the HNC approximation, the code is limited by its memory-consumption and we need to find a smart way to take solute flexibility into account. Moreover further theoretical and technical developments will enable the use a quantum solute, the computation of binding free energies with MM/MDFT, the entropic and enthalpic contributions and partition coefficients.

12.2.1 GOING BEYOND THE HNC APPROXIMATIONS

Now that, MDFT (i) can be efficiently resolved at HNC approximation level, (ii) several *a posteriori* correction to SFE correct for the deficiencies of the HNC approximation and (iii) with these pressure corrections MDFT-HNC have been benchmarked on a multitude of systems yielding HFEs within 0.5 kcal/mol of rigorous reference simulations for small neutral molecules and the rest within 5 kcal/mol (less than 5% in relative error for ions): the next natural step is to go beyond the HNC approximation, *i.e.* the second-order density expansion, with the addition of a bridge functional that include cubic and higher-order perturbations.

This thesis provided a basis for the developments of bridge functional as (i) one lesson of this work, already noted by others, is that for future improvement, it is certainly wise to proceed in two steps, as it is done usually in simulations: first introducing the nude LJ interaction, then adding the charges in a second step; (ii) the benchmark showed that MDFT-HNC struggles to predict the correct structure especially around high negative charges (anions or water's oxygen) and does not produce the tetrahedral order of hydrogen bonding. Hence, the bridge functional should aim at correcting these deficiencies; and (iii) this thesis produced a large quantity of reference data from a multitude of simple spherical $\{\sigma, \epsilon, q\}$-triplets to single conformer references of molecular solutes.

Proceeding as the first point suggested, we have recently proposed a simple angular-independent parameter-free (in the sense that all the parameters appearing in the expression of the bridge functional are determined unambiguously from the properties of the bulk solvent (pressure, isothermal compressibility, liquid—gas surface tension, *i.e.* solute independent) weighted density bridge functional for hydrophobic solvation [132]. This simple angular-independent, and thus computationally efficient, can capture the main physical features of hydrophobic solvation and predicts the hydration free energies of FreeSolv-LJ, *i.e.* solutes without partial charges, within 0.25 kcal/mol of reference H4D-MC data.

To go beyond the hydrophobic solvation, *i.e.* when the partial charges are switched on, will require an angle-dependent bridge functional in order to systematically improve the results presented in this thesis. The development of the angular-dependent and solute-independent bridge functional is currently going on. As remarked in the second point, this bridge functional should aim at correcting deficiencies around negative charges and in hydrogen bonding.

Note that, even though only the HFE of the $\{\sigma, \epsilon, q\}$-triplets were discussed in this thesis, the full molecular $\rho(r, \omega)$ was rigorously computed for the $\{\sigma, \epsilon, q\}$-triplets. These densities will be used to verify and develop a bridge functional that also corrects the structures and not only the HFE. Moreover, from these simulations, we can to extract the solute-dependent bridge function, a first step in developing the bridge functional.

12.2.2 REDUCING THE MEMORY FOOTPRINT

The main advantage of MDFT compared to simulations, or even RISM methods, is its speed. The total CPU time of an MDFT minimisation is typically between few seconds to few (tens) of minutes with the L-BFGS-B minimiser depending on the grid resolution. However, it is very memory-consuming. The L-BFGS-B minimiser needs to store some data in double precision during the iteration and, during the functional evaluation, the memory for 3 $\rho(\boldsymbol{r}, \omega)$ needs to be open simultaneously. As shown in fig. 12.1 the typical amount of used RAM is between 1 and 20 GB. Typical laboratory clusters cannot allocate more memory to process. With our 'normal' MDFT parameters of a cubic supercell with $n_{\max} = 3$ and $\mathrm{d}x = 0.33$ Å, we can only study box sizes up to $L = 42$ Å. Therefore, we are limited to study relatively small systems with the L-BFGS-B minimiser. To study larger biochemical systems the memory consumption should be reduced.

The easiest way to overcome this problem is to use a less memory-consuming minimiser like steepest descent which is implemented into MDFT. However, the use of the simple steepest

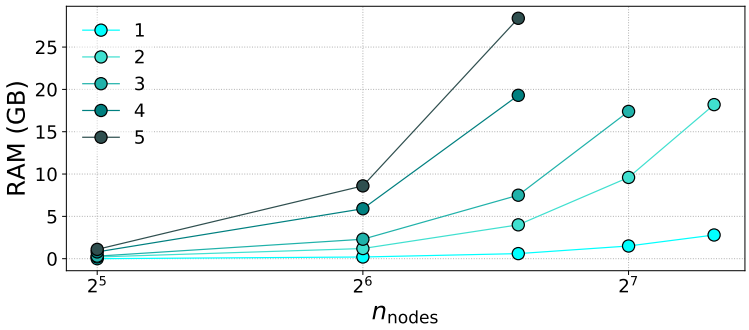


Figure 12.1: MDFT's use of RAM as a function of spatial nodes in each direction and orientational discretisation $n_{\max}$.

descent minimiser decreases substantially the speed of minimisation leading the minimisation times from few minutes to few hours depending on the grid resolution. Other options, to bypass the memory problem, would be to parallelise the code to several nodes with MPI (currently the code is parallelised with OMP) which requires only the modification of the FFT and L-BFGS-B processes. This work is ongoing in Jeanmairet's laboratory.

Another path to reduce the memory consumption is to use of non-uniform grids for the spatial and orientational discretization. Currently, MDFT uses a homogeneous grid, in space and orientation, *i.e.* it has the same resolution near and far from the solute which can lead to a very large number of grid points. However, *eg.* looking at the radial distribution functions in chapter 9, density close the water's bulk density are recovered for $r_{\text{site}-\text{O}} > 8$ Å and the most interesting information is within few Å of the surface of solute sites. Moreover, for large solutes, lots of frequent grid points are 'wasted' on the solute cavity here the solvent density is null. This is even more true for the angles.

To overcome the 'wasting' frequent spatial grid points far from the solute is to use a non-uniform grid such as a set of spherical grids centred on solute sites. This can be seen as an expansion of the density onto a set of 'atomic-like orbitals' where is 'orbital' could be expanded on a local basis set like spherical harmonics. However, this approach does not resolve the wasting of grid points in the solute cavity as the grid would be very tight where solute sites overlap.

As can be seen in fig. 12.1 the memory consumption increases a lot with the orientational resolution. Similar to the spatial discretization, the fine orientational resolution is important close to the solute sites and not so much far from the solute. To reduce memory-consumption, one could implement a multi-grid approach with a higher orientational resolution close to the solute and lower one further away.

Another option would be to expand the density on a wavelet basis instead of the Fourier basis, as they are more adapted to process highly localised information [189, 190]. Preliminary theoretical analysis indicates that the use of the wavelet approach instead of the direct one could reduce the memory requirements of 99% in 3D while being computationally at least equivalent to the direct approach. This would allow MDFT to compute free energies of systems roughly ten times larger than it can in the current implementation with no trade-off in computational time. It has been done for electronic DFT for example in BigDFT [190, 191] and MADNESS codes [192].

12.2.3 SOLUTE FLEXIBILITY / MM-MDFT

As shown, in this thesis (i) a single conformer HFE calculation is not sufficient for correctly predicting the HFEs of molecules with H-bond acceptors and especially donors and (ii) the flexible solute's HFE cannot be systematically reconstructed from a handful of conformers obtained from random snapshots of vacuum and solution simulations and computing the HFE for every conformer in MD simulation with MDFT is not efficient at all. Nonetheless, H-bond acceptors and donors

play an important role in biochemistry and the affinity between a ligand and a protein's active site. So correctly predicting their HFEs would be a great interest for MDFT. Therefore, it is important to develop or apply more sophisticated methods to sample and identify important conformers. First tries, conformer clustering to identify main average conformers did not lead to hoped results.

12.2.4 COUPLING MDFT

The long-term aim has been to replace implicit solvent methods at different scales with MDFT. Now that (i) MDFT can be efficiently and rigorously solved at the HNC level [40] and (ii) the MDFT-HNC, when coupled with an appropriate pressure correction, can produce 'acceptable' results, the MDFT can be coupled with quantum calculations, molecular mechanics, drug design pipelines or hydrodynamics. These are not any more long-term goals but current projects or doable in the close future.

The first ongoing project is to couple MDFT at micro-scale with quantum calculations, QM/M-DFT [126], as an alternative to widely used PCM methods and QM/MM simulations, and more recently developed QM/RISM approaches [117, 118]. Here, the idea is to treat the solute with QM, which creates the external potential V_{ext} for the MDFT minimisation, and the solvent classically with force field representation but instead of using 'expensive' explicit solvent simulations the solvent in treated by the much faster MDFT. Moreover, as a replacement to (i) PCM it brings structure information on the solvation missing in the implicit solvent models and (ii) it gives directly the solvation free energy which cannot be obtained from a single QM/MM simulation and (iii) the solvent induces a polarization of the electronic density which in turn modified the solvent density $\rho(\boldsymbol{r}, \omega)$.

The second project consists of coupling MDFT at the mesoscopic scale to Laboetie [193, 194, 195], a computational fluid dynamic code developed for chemical applications. It is based on Lattice-Boltzmann methods for fluid simulation and takes into account the chemical specificity to study the transportation of a chemical reactive. It is especially adapted for the study of particles that can be adsorbed and desorbed to/on a surface. Here the MDFT aims to better model the surface adsorption/desorption by computing the adsorption/desorption kinetic constants that are, at the moment, produces as input parameters.

The third aim to propose MM/MDFT as a replacement of MM/PBSA and MM/GBSA to estimate the binding free energy in drug design purposes [196]. Preliminary results on single conformer of a docked ligand-active site complex, have shown that the quality of the binding free energy prediction improves when structural information on the solvation is added via MDFT when compared to MM/PBSA results [197]. The next step would be couple MDFT to drug design pipeline, like AutoDock, to be able to systematically use MDFT to model the solvent effects in the ligand-protein interaction and hopefully improve the quality of the affinity prediction.

12.2.5 EFFECT OF TEMPERATURE

All MDFT (and H4D-MC) calculations in this thesis were performed at room temperature and only solvation free energies were computed. This is due, as noted in chapter 4.2, to one of the MDFT's main input, the solvent-solvent pair correlation function $c^{(2)}(r_{12}, \omega_1, \omega_2)$, that is computed for a given temperature T and pressure P. Until now, these pair correlation functions are computed only for the normal conditions $T = 298.15$ K and $P = 1$ atm. Thus MDFT could not be used at other temperatures. and therefore could not be used to evaluate the entropic and enthalpic contributions of the hydration process.

In the future, the possibility to use MDFT at other temperatures would be of great interest as (i) for biochemical and pharmaceutical purposes it could be interesting to predict hydration free

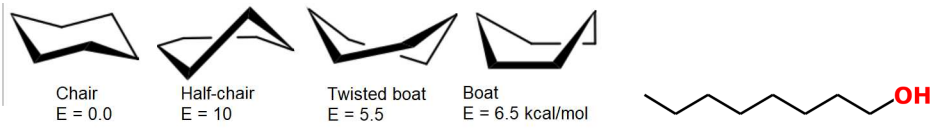
energies at body temperature (~ 310 K), and furthermore, (ii) the possibility to use MDFT at different temperatures gives access to the enthalpic and entropic contributions of the solvation process. They can be of great importance when studying real applications [198, 199].

At the time of the writing of this thesis, the project of studying temperature effects of TIP3P's $c^{(2)}(r_{12}, \omega_1, \omega_2)$ in the range of 280-320 K, *i.e.* the interesting temperature range in biochemistry as, below it, water freezes and, above it, most proteins are denatured, is ongoing. Note that, producing these $c^{(2)}(r_{12}, \omega_1, \omega_2)$ is not trivial as it requires (i) very rigorous simulations, *eg.* a first simulation is done at each temperature to evaluate the relative permittivity $\varepsilon_r(T)$ of water to set the relative permittivity of the imaginary surface $\varepsilon'(T) \approx \varepsilon_r(T)$ for the production run; (ii) very long simulations MC to correctly sample the six-dimensional molecular density [17, 18]. Preliminary results show that $\Delta c^{(2)}/\Delta T$ seems to vary almost linearly in the range of 280-320 K. This could allow the fast determination of $c^{(2)}$ and the use of MDFT at any temperature in the range. This, in turn enables the computation of the hydration enthalpies and entropies with MDFT. There is also a lead to determine the solvation entropy directly from MDFT without doing the minimisation at two different temperatures.

12.2.6 PARTITION COEFFICIENTS

Beyond water, there are other important solvents, cyclohexane and n-octanol, used in modelling solvation effects. The former is used as a model for a generic organic solvent in chemical synthesis and the latter is used for modelling lipid membranes, *eg.* cell membranes. The partition coefficient of both solvents with water, $\log P_{hw}$ and $\log P_{ow}$, are important thermodynamic quantities: the former to predict liquid-liquid extraction properties in research or industrial synthesis of organic molecules and the latter to predict the capability of a molecule to penetrate a cell membrane, a positive property of drug molecules and a negative one for molecules *eg.* in cleaning products.

Therefore, it would be of great interest to implement these two solvents to MDFT. In principle, adding a new solvent to MDFT only requires the computation of the solvent-solvent pair correlation function of the given *rigid* solvent. However, these are not 'rigid' solvent like water. For the cyclohexane, as it is cyclic and has few degrees of freedom, this problem could be 'easily'[1] solved by using a mixture of solvent where each separate solvent corresponds to the rigid conformer presented in fig. 12.2a. N-octanol (fig. 12.2b) is much more complicated case as it has a long aliphatic chain and lots of degrees freedom. It has lots of possible conformers and it is not evident how to model n-octanol with few rigid conformers. Nevertheless, due to the importance of $\log P_{ow}$, especially in pharmaceutical research the implementation of n-octanol into MDFT should be investigated.



(a) (b)

Figure 12.2: Illustration of (a) the main conformers of cyclohexane and their relative internal energies and (b) chemical structure of n-octanol.

1 In principle, there are no issues to use a mixture of solvents like electrolytes (mixture of water and ions) as the solvent in MDFT. However, for the moment the cost in memory is too high.

Part V

APPENDIX

A

GRADIENTS OF THE MDFT FUNCTIONAL

This appendix details the of gradient of the functional and shows that at the variational minimum the MDFT function gives the HNC relations of solute-solvent distribution function.

First, starting from the **ideal term** of the MDFT functional (eq. 4.5), the finite difference is reads

$$\begin{aligned}
\beta\delta\mathcal{F}_{\mathrm{id}}[\rho] &= \mathcal{F}_{\mathrm{id}}[\rho+\delta\rho] - \mathcal{F}_{\mathrm{id}}[\rho] \\
&= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega(\rho+\delta\rho)\ln\left(\frac{\rho+\delta\rho}{\rho_{\mathrm{bulk}}}\right) - (\Delta\rho+\delta\rho) \\
&\quad - \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) - \Delta\rho \\
&= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho\ln\left(\frac{\rho+\delta\rho}{\rho_{\mathrm{bulk}}}\right) + \delta\rho\ln\left(\frac{\rho+\delta\rho}{\rho_{\mathrm{bulk}}}\right) - \Delta\rho \\
&\quad - \delta\rho - \rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) + \Delta\rho \\
&= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho\ln\left(\frac{\rho+\delta\rho}{\rho_{\mathrm{bulk}}}\right) + \delta\rho\ln\left(\frac{\rho+\delta\rho}{\rho_{\mathrm{bulk}}}\right) - \delta\rho - \rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right).
\end{aligned} \tag{A.1}$$

Given that

$$\begin{aligned}
\ln\left(\frac{x+\delta x}{x_0}\right) &= \ln\left(\frac{x+\delta x}{x}\frac{x}{x_0}\right) = \ln\left(\frac{x+\delta x}{x}\right) + \ln\left(\frac{x}{x_0}\right) \\
&= \ln\left(1+\frac{\delta x}{x}\right) + \ln\left(\frac{x}{x_0}\right)
\end{aligned} \tag{A.2}$$

and as $\frac{\delta\rho}{\rho}$ tends to 0, it is possible to do a Taylor expansion of $\ln\left(1+\frac{\delta x}{x}\right)$, transforming the previous equation to

$$\ln\left(\frac{x+\delta x}{x_0}\right) = \frac{\delta x}{x} + \ln\left(\frac{x}{x_0}\right) + \mathcal{O}(\delta x^2). \tag{A.3}$$

Injecting the this result into the equation first equation, the finite difference reads

$$\begin{aligned}
\delta\mathcal{F}_{\mathrm{id}}[\rho] &= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho\left(\frac{\delta\rho}{\rho} + \ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right)\right) + \delta\rho\left(\frac{\delta\rho}{\rho} + \ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right)\right) \\
&\quad - \delta\rho - \rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) + \mathcal{O}(\delta\rho^2) \\
&= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\delta\rho + \rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) + \frac{\delta\rho^2}{\rho} + \delta\rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) \\
&\quad - \delta\rho - \rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) + \mathcal{O}(\delta\rho^2) \\
&= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\delta\rho\ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right) + \mathcal{O}(\delta\rho^2)
\end{aligned} \tag{A.4}$$

and the derivative of the ideal term reads

$$\beta\frac{\delta\mathcal{F}_{\mathrm{id}}}{\delta\rho} = \ln\left(\frac{\rho}{\rho_{\mathrm{bulk}}}\right).$$

Secondly, starting from the **external term** of the MDFT functional (eq. 4.6), the finite difference is reads

$$\begin{aligned}\delta\mathcal{F}_{\text{ext}}[\rho] &= \mathcal{F}_{\text{ext}}[\rho+\delta\rho] - \mathcal{F}_{\text{ext}}[\rho] \\ &= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega(\rho+\delta\rho)U_{\text{ext}} - \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\rho U_{\text{ext}} \\ &= \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\delta\rho U_{\text{ext}}\end{aligned} \tag{A.5}$$

and the derivative of the external term reads

$$\beta\frac{\delta\mathcal{F}_{\text{ext}}}{\delta\rho} = \beta U_{\text{ext}}. \tag{A.6}$$

Thirdly, starting from the **HNC part of the excess term** of the MDFT functional (eq. 4.8), the finite difference is reads

$$\begin{aligned}\beta\delta\mathcal{F}_{\text{HNC}}[\rho] =& \beta\mathcal{F}_{\text{HNC}}[\rho+\delta\rho,\rho'] - \beta\mathcal{F}_{\text{HNC}}[\rho,\rho'] \\ &+ \beta\mathcal{F}_{\text{HNC}}[\rho,\rho'+\delta\rho'] - \beta\mathcal{F}_{\text{HNC}}[\rho,\rho'] \\ =& 2\left(\beta F_{\text{HNC}}[\rho+\delta\rho,\rho'] - \beta\mathcal{F}_{\text{HNC}}[\rho,\rho']\right) \\ =& -\int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega(\Delta\rho+\delta\rho\gamma) + \int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\Delta\rho\gamma \\ =& -\int \mathrm{d}\boldsymbol{r}\mathrm{d}\omega\delta\rho\gamma\end{aligned} \tag{A.7}$$

and the derivative HNC term reads

$$\beta\frac{\delta\mathcal{F}_{\text{HNC}}}{\delta\rho} = -\gamma. \tag{A.8}$$

At the variational minimum the derivative is null

$$\beta\frac{\delta\mathcal{F}}{\delta\rho} = \ln\left(\frac{\rho}{\rho_{\text{bulk}}}\right) + \beta U_{\text{ext}} - \gamma = 0. \tag{A.9}$$

By simple algebric transformations, one obtaines following the solute-solvent HNC relation

$$g = \frac{\rho}{\rho_{\text{bulk}}} = e^{-\beta U_{\text{ext}}+\gamma}. \tag{A.10}$$

B

STATISTICAL MEASURES

Six statistical statistical measures are used in this thesis report to quantify the quality of MD+FEP, H4D-MC, RISM or MDFT results : three mean errors and three correlation coefficients.

Three different mean errors are used to measure of the differences between values predicted by a model 1 and the values actually observed (or predicted by a model 2) :

The **Mean (signed) error (ME)** which is formally defined as

$$\mathrm{ME} = \frac{\sum_i (\hat{y}_i - y_i)}{n} \tag{B.1}$$

The **Mean absolute error (MAE)**, also called the mean unsigned error (MUE), which is formally defined as

$$\mathrm{MAE} = \frac{\sum_i |\hat{y} - y_i|}{n}. \tag{B.2}$$

The **Root-mean-squared error (RMSE)**, also called the root-mean-squared deviation (RMSD), is formally defined as

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}. \tag{B.3}$$

where $\hat{y}_i$ and y_i are the corresponding observed and predicted values respectively and n the number prediction. All the mean errors have the same unit as the predicted/observed value and smaller they are the better the predictor. The sign of ME indicates if the predictor has a systematic bias but is not pertinent if the predicted values are distributed uniformly around the observed values. The MAE gives an uniform weight to all values whereas the RMSE is emphasizes the weight of the outliers (MAE is always smaller than the RMSE).

Three different correlation coefficients are used to measure of the correlations between values predicted by a model 1 and the values actually observed (or predicted by a model 2) :

The **Pearson correlation coefficient (Pearson's R)** measures the linear correlation between two variables X and Y. It is formally defined as

$$R = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{B.4}$$

where $\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)]$ is the covariance between X and Y and σ_X and σ_Y the standard deviation of X and Y respectively. The commonly used coefficient of determination R^2 is the square of the Pearson's R.

The **Spearman rank correlation coefficient (Spearman's ρ)** measures the monotonic correlation between two variables X and Y.. It is formally defined as

$$\rho = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{B.5}$$

where $\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ the is covariance between rank variables and rg_X and rg_Y, and $\sigma_{\mathrm{rg}_X}$and σ_{rg_Y} the standard deviation of rank variables. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables and varies the same way as the Pearson correlation coefficient.

The **Kendall rank correlation coefficient (Kendall's τ)** measures also the monotonic correlation between two variables X and Y. It is formally defined as

$$\tau = \frac{n_{\text{con}} - n_{\text{dis}}}{n(n-1)/2} \tag{B.6}$$

where n_{con} is the number of concordant pairs, n_{dis} the number of disconcordant pairs and n the total number of pairs. When evaluating each possible pair $[(X_i, Y_i), (X_j, Y_j)]$, the pair is counted as concordant if $Y_i > Y_j$ when $X_i > X_j$ or if $Y_i < Y_j$ when $X_i < X_j$ else the pair is counted as disconcordant.

The correlations coefficients vary between -1 and 1 and the higher the magnitude of the correlation coefficient better the linear/monotonic correlation is between the prediction and the observation. The sign of R indicates if the predictions and the observable are directly correlated (+) or anti-correlated (-).

C

FREESOLV DATABASE

The FreeSolv is a widely used database of hydration free energies curetted by the Mobleylab [200]. It contains the experimental hydration free energies of 642 small neutral organic 'drug-like' molecules and those obtained with state-of-the-art MD+FEP. The original database created by Mobley of 504 molecules was created based on previous datasets, notably from Rizzo [201] and previous studies Mobley and co-workers. Additional molecules have been added as the result of new studied since for example as part of the SAMPL challenges [202] resulting to the current database [203, 144].

The database has more than 70 chemical functions with a large distribution of the functional groups (only 16 functions are present in over 20 solutes). The molecular masses range from 16 to 493 Da. These are typical sizes for drug-like molecules as defined by Lipinski's 'rule of five' [204] of molecule's drug-likeness with a molecular mass criteria maximum at 500 Da. Moreover, 95% of the database has a molecular mass lower than 300 Da defined as a limit for lead-like molecules by the 'rule of three' [205]. However, the average molar mass is 140 Da which implies that most solutes are smaller than typical drug-molecules. The database contains only neutral molecules as measuring SFEs of an isolated charged species requires extra thermodynamic assumptions or introduces other complexities [144] that are still not well understood. Nevertheless, some molecules have relatively high partial charges implying important electrostatic interactions with the solvent and probably hydrogen bonding.

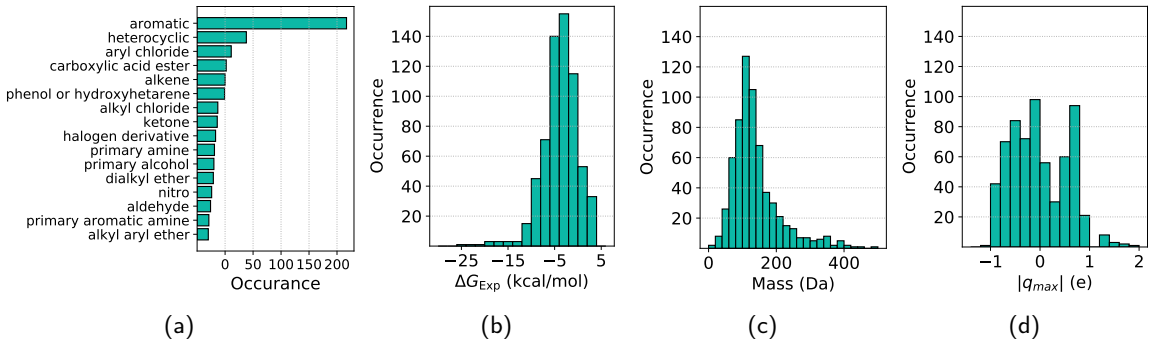


Figure C.1: (a) Occurrence of the chemical function present in over 20 solutes and distribution of (b) experimental hydration free energies, (c) solute mass and (d) solute's highest partial charge.

The MD+FEP calculations were done with Gromacs [145, 146, 147, 148, 149] and analysed with MBAR [75]. The database contains all the input files including the starting structures and force field details of the molecules. The solutes were modelled with the GAFF force field [153] (v1.7) with AM1-BCC charges [154, 155] and the water with the TIP3P model [28].

D

H4D-MC SIMULATION DETAILS

All H4D-MC simulations are done with the following simulation parameters for rigid (table D.1) and flexible (table D.2) solute simulations if not explicated differently in the text.

Simulation box parameters	
Number of TIP3P water molecules: N	100
Temperature: T	298.15 K
Pressure: P	1 atm
Ewald decomposition: KL, s_r, s_k	8 4 4
Exterior dielectric constant: ϵ'_r	99
H4D parameters	
Maximum "altitude": $w_{\max}$	3 Å
Ins/des speed: v	0.05 $\sqrt{k_{\mathrm{B}}T/M}$
Time step: Δt	0.02 $\sqrt{\beta M}$Å
Ewald decomposition: KL', s'_r, s'_k	8 3 3
Reference μ^0_{exc}	0 $k_{\mathrm{B}}T$
Volume change: V_0	$\sim V_{\mathrm{PM}}$
Reduced mass of solute sites: m	1×10^{20} (=rigid solute)
MC propagation parameters	
Equilibration	10 000 MC cycles
Accumulation interval	100 MC cycles
Max. translation of a solvent molecule	0.3 Å
Max. rotation of a solvent molecule	30 Å
λ parameters for MC Force-bias	0.5 0.5
Volume exchange probability	0.2
Max. change in $\ln V$	0.05

Table D.1: Reference single conformer H4D-MC simulation parameters.

Simulation box parameters	
Number of TIP3P water molecules: N	100
Temperature: T	298.15 K
Pressure: P	1 atm
Ewald decomposition: KL, s_r, s_k	8 4 4
Exterior dielectric constant: ϵ'_r	99
H4D parameters	
Maximum "altitude": $w_{\max}$	3 Å
Ins/des speed: v	0.05 $\sqrt{k_{\mathrm{B}}T/M}$
Time step: Δt	0.02 $\sqrt{\beta M}$Å
Ewald decomposition: KL', s'_r, s'_k	8 3 3
Reference μ^0_{exc}	0 $k_{\mathrm{B}}T$
Volume change: V_0	$\sim V_{\mathrm{PM}}$
Reduced mass of solute sites: m	1×10^{20} (w/o relaxation)
	3 (w/ relaxation)
MC propagation parameters	
Equilibration	10 000 MC cycles
Accumulation interval of insertions	100 MC cycles
Insertion configuration generation in vacuum	10 000 MC cycles
Accumulation interval of destructions	1 000 MC cycles
Max. translation of a solvent molecule	0.3 Å
Max. rotation of a solvent molecule	30 Å
λ parameters for MC Force-bias	0.5 0.5
Volume exchange probability	0.2
Max. change in $\ln V$	0.05
Max. displacement of a solute site	0.1 Å
Rel. probability to move a solvent and solute site	1 5

Table D.2: Reference flexible solute H4D-MC simulation parameters.

E

'FREESOLV-RIGID'

Table E.1 gives the FreeSolv identifications of the 213 FreeSolv solutes defined to be rigid. More informations on these solutes at github.com/sohviluukkonen/Thesis

1075836	1893937	2725802	3761215	4893032	6250025	7690440	8809190
1079207	1899443	2771569	3762186	4983965	627267	7708038	8809274
1107178	1923244	2784376	3775790	5094777	628951	7732703	8827942
1160109	1929982	2789243	3802803	5110043	6303022	7769613	8885088
1189457	1952272	2802855	3968043	511661	6359135	778352	8966374
1199854	1963873	282648	3969312	5157661	6430250	7814642	900088
1231151	1977493	2837389	3980099	5220185	6474572	7859387	9028462
1235151	2008055	2859600	3982371	525934	6571751	7893124	9029594
1261349	2049967	2881590	4013838	5263791	6739648	7943327	9055303
129464	2068538	296847	4043951	5310099	676247	8006582	9073553
1363784	210639	2972345	4149784	5346580	6804509	8117218	9100956
1424265	2146331	2972906	4188615	5449201	6812653	8127829	9121449
1520842	2198613	299266	4219614	5471704	6911232	8260524	9139060
1674094	2261979	2996632	4287564	5494918	6981465	8311321	9246351
1717215	2341732	303222	4291494	5616693	6988468	8320545	929676
1723043	2390199	3053621	430089	5690766	7047032	8337977	9434451
1760914	2451097	3211679	4434915	5747188	7099614	8436428	9507933
1800170	2484519	3318135	4463913	5852491	7150646	8492526	9565165
1821184	2487143	3323117	4479135	5890803	7157427	8514745	9671033
1827204	2489709	3370989	4483973	5935995	7239499	8525830	9705941
1838110	2492140	3395921	4494568	5952846	7298388	8558116	9740891
1855337	2517158	3398536	4678740	5977084	7415647	8578590	9913368
186894	252413	3425174	468867	6081058	7532833	8614858	9942801
1873346	2577969	3525176	4694328	6091882	7578802	8739734	994483
1875719	2607611	3572203	4759887	6102880	7599023	8764620	
1881249	2681549	3639400	4762983	6175884	7608462	8772587	
1893815	2689721	3682850	4845722	6235784	766666	8785107	

Table E.1: FreeSolv-rigid database: solute IDs.

F

NEURAL NETWORK DETAILS

The neural network used to fit an energy correction to MDFT-HNC results in section 8.3 was done with MLPRegressor feature of the Python3 open-source library sklearn [173]. The structure of the feed-forward NN was an input layer composed of 167 nodes, with 166 corresponding to the MACCS key [171] and one to the MDFT-HNC HFE of the solute, two hidden layers with 84 nodes each and a single output node since we want to do single number regression.

The rectified linear unit function (ReLU) $f(x) = \max\{0, x\}$ [206], the most successful and widely-used activation function [207], was applied between the input layer and hidden layers nodes to have an non-linear model. As normal, the activation function was do applied between the last hidden layer and output, since the ReLU function would not allow negative output values.

The optimisation was done with the Adam solver [208], an extension to stochastic gradient descent widely used in machine learning community for its efficiency [209] with a constant learning rate of 10^{-4}, batch size of 50 for the stochastic optimiser and default Adam's parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model optimised the squared-loss function with a tolerance of 10^{-4}.

The limit over-fitting the model on to the training data two measures were takes : (i) 5% of the training data (30 molecules) were used a validation, their loss is was evaluated at each iteration but this information is not used for the back propagation, and the optimisation is stopped when the validation loss does not improve for 10 epochs; (ii) addition of the L_2 penalisation term to the loss function, i.e Ridge regularization with the penalisation scale parameter $\lambda = 10^{-4}$.

G

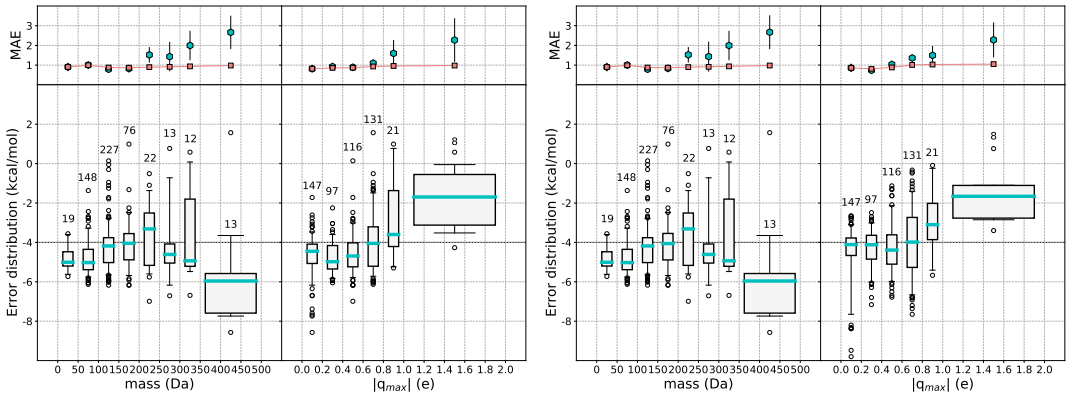
PROBLEMATIC SOLUTES FOR MDFT

There were 26 molecules for which the MDFT calculation did not converge either in chapter 9.2.2 (22 solutes) or 10 (23 solutes). Table G.1 summarizes these molecules and their characteristics. In general these molecules have on average higher (i) experimental hydration free energies ($\langle \Delta G^{\mathrm{Exp}}_{\mathrm{solv}} \rangle_{26} = -6.39$ kca/mol vs. $\langle \Delta G^{\mathrm{Exp}}_{\mathrm{solv}} \rangle_{642} = -3.81$ kcal/mol), (ii) masses ($\langle \mathrm{mass} \rangle_{26} = 187.26$ Da vs. $\langle \mathrm{mass} \rangle_{642} = 138.58$ Da) and (iii) partial charges ($\langle \max\{q_i\} \rangle_{26} = 0.75e$ vs. $\langle \max\{q_i\} \rangle_{642} = 0.47e$) than average of the full dataset. We could not identify striking common characteristics explaining why these molecules did not converge.

FreeSolvID	$\Delta G^{\mathrm{Exp}}_{\mathrm{solv}}$	mass (Da)	$\max\{q_i\}$ $[e]$	Functional groups
1527293	-0.82 ± 0.16	244.09	0.643	aryl flurodide, carboxylic acid, aromatic
2078467	-7.00 ± 0.64	206.13	0.643	carboxylic acid, aromatic
2099370	-10.78 ± 0.18	254.09	0.633	ketone, carboxylic acid, aromatic
2269032	-10.21 ± 0.18	230.09	0.619	alkyl aryl ether, carboxylic acid, aromatic
2518989	-5.74 ± 1.93	393.00	0.915	alkyl chloride, carboxylic acid imide N-substituted, thioposphoric acid ester, aromatic, heterocyclic
2958326	-3.65 ± 0.60	101.12	-0.830	secondary amine, dialkylamine
3201701	-9.41 ± 1.93	238.14	0.901	tertiary amine, alkylarylamine, urethane
3274817	-6.23 ± 1.93	240.07	-0.467	phenol or hydroxyhetarene, nitro, aromatic
4587267	-23.62 ± 0.32	182.08	-0.607	primary alcohol, secondary alcohol, 1,2-diol
4690963	-3.54 ± 0.60	**118.19**	**-0.427**	**dialkyl ether**
4934872	-5.23 ± 0.60	**196.11**	**0.510**	**orthocarboxylic acid derivative, orthoester, aromatic**
4936555	-6.78 ± 0.10	**241.11**	**-0.668**	**secondary amine, diarylamine, carboxylic acid, aromatic**

FreeSolvID	$\Delta G_{\text{solv}}^{\text{Exp}}$	mass (Da)	$\max\{q_i\}$ [e]	Functional groups
5393242	-6.48 ± 0.13	304.10	1.264	thiophosphoric acid ester, aromatic, heterocyclic
5747981	-5.73 ± 0.60	**150.09**	**0.529**	**dialkyl ether, orthocarboxylic acid derivative, orthoester**
5880265	-6.25 ± 0.60	118.10	-0.594	primary alcohol, dialkyl ether
6309289	-5.11 ± 0.60	***85.08***	***-0.814***	***secondary amine, secondary alphatic amine, heterocyclic***
6334915	-12.74 ± 1.93	255.93	1.517	halogen derivative, phosphonic acid derivative, phosphonic acidester
63712	-3.88 ± 0.60	99.10	-0.712	tertiary amine, trialkylamine, heterocyclic
646007	-5.48 ± 0.60	71.07	-0.812	secondary amine, dialkylamine, heterocyclic
6620221	-9.62 ± 0.30	***221.105***	***0.742***	***alkyl aryl ether, urethane, aromatic, heterocyclic***
7326706	-5.05 ± 0.60	***348.93***	***1.237***	***aryl chloride, thiophosphoric acid ester, aromatic, heterocyclic***
7860938	-3.24 ± 0.60	129.15	-0.834	secondary amine, dialkylamine
8426916	-4.07 ± 0.60	73.09	-0.835	secondary amine, dialkylamine
8449031	-3.22 ± 0.60	101.12	-0.720	tertiary amine, trialkylamine
8705848	-3.22 ± 0.60	101.12	-0.822	secondary amine, dialkylamine
9460824	-4.37 ± 0.10	260.01	0.843	thiophosphoric acid este
Average	-6.93	187.26	0.753	

Table G.1: List of molecules that did not converge in our MDFT calculations. Solutes in bold correspond to solutes that converged only in one of the MDFT calculation (straight one did not converge in chapter 10 and italics in section 9.2.2). All other molecules led to divergence in both MDFT calculations.



(a) (b)

Figure H.1: Distribution of mean absolute error between experimental hydration free and those predicted by (a) MD+FEP and (b) 3D-RISM as a function of two features (left) solute's molar mass (with bin size of 50 Da) and (right) solute's largest local charge (0.2 e) with turquoise hexagons (error bars = 2σ) presenting the MAE of each bin and pink squares the cumulative MAE. The corresponding signed error distributions are presented below with turquoise lines corresponding to median error in each bin, the boxes and the whiskers to 25-75% and 5-95% intervals respectively and black circles to fliers outside the 5-95% interval. The number above each bin is the population of the bin for the FreeSolv database and for each distribution the last three (350-500 Da) or five (1-2 e) bins are gathered into one in order to be statistically significant.

I

MDFT-HNC ERROR BARS

Group 1	Group 2	Group 3	#	MAE (kcal/mol)
Aromatic	-	-	**257**	**1.18 (0.15)**
	Heterocyclic	-	**64**	**1.32 (0.27)**
		Aryl chloride	**19**	**1.13 (0.45)**
		Diaryl ether	**12**	**1.19 (0.58)**
		Oxo(het)arene	**12**	**1.99 (0.91)**
	Aryl chloride	-	**61**	**1.34 (0.31)**
		Phenol*	8	1.69 (0.53)
		Diaryl ether	**11**	**1.05 (0.55)**
	Carboxylic acid ester	-	**10**	**1.22 (0.48)**
	Alkene	-	5	0.36 (0.18)
	Phenol*	-	**48**	**1.33 (0.46)**
		Aldehyde	5	2.19 (1.25)
		Alkyl aryl ether	8	1.00 (0.81)
	Ketone	-	**10**	**2.56 (1.50)**
	Halogen derivative	-	9	1.80 (0.94)
		Tert. alkylarylamine	5	1.49 (1.49)
	Primary amine	-	**21**	**1.36 (0.86)**
	Nitro	-	**19**	**1.46 (0.64)**
		Tert. alkylamine	5	1.42 (1.41)
	Aldehyde	-	**10**	**1.55 (0.75)**
	Alkyl aryl ether	-	**19**	**1.12 (0.47)**
	Tertiary amine	-	9	1.61 (0.98)
	Diaryl ether	-	**13**	**1.15 (0.54)**
	Carbonitrile	-	6	0.84 (0.45)
	Secondary amine	-	5	2.24 (0.80)
	Tert. alkylarylamine	-	7	1.81 (1.20)
	Thiophosphoric acid ester	-	5	2.74 (0.98)
	Aryl bromide	-	6	0.93 (1.11)
	Aryl fluoride	-	5	1.62 (1.64)
	Sec. alkylaryl amine	-	5	2.24 (0.80)

Table I.1: MAEs of molecules with one, two or three specific chemical function. (1/3)

Group 1	Group 2	Group 3	#	MAE (kcal/mol)
Heterocyclic	-	-	**83**	**1.36 (0.27)**
	Aryl chloride	-	**19**	**1.13 (0.46)**
		Diaryl ether	**11**	**1.04 (0.55)**
	Dialkyl ether	-	8	0.61 (0.28)
	Diaryl ether	-	**12**	**1.19 (0.58)**
	Oxo(het)arene	-	**12**	**1.99 (0.90)**
	Secondary amine	-	**12**	**1.51 (0.76)**
	Sec. alkylaryl amine	-	5	0.86 (0.34)
Aryl chloride	-	-	**61**	**1.34 (0.31)**
	Phenol*	-	8	1.70 (0.53)
	Diaryl ether	-	**11**	**1.05 (0.55)**
Carboxylic acid ester	-	-	**52**	**1.28 (0.23)**
Alkene	-	-	**50**	**0.74 (0.21)**
	Alkyl choride	-	5	1.86 (1.47)
Phenol*	-	-	**48**	**1.33 (0.46)**
	Aldehyde	-	5	2.18 (1.24)
	Alkyl aryl ether	-	8	1.01 (0.81)
Alkyl chloride	-	-	**36**	**0.88 (0.29)**
Ketone	-	-	**35**	**2.57 (0.54)**
Primary alcohol	-	-	**28**	**1.51 (0.37)**
Dialkyl ether	-	-	**26**	**0.88 (0.25)**
Nitro	-	-	**25**	**1.32 (0.51)**
	Tert. alkylarylamine	-	5	1.42 (1.41)
Aldehyde	-	-	**24**	**0.80 (0.41)**

Table I.2: MAEs of molecules with one, two or three specific chemical function. Data points only groups containing at least five molecules. (2/3)

Group 1	Group 2	Group 3	#	MAE (kcal/mol)
Prim. aromatic amine	-	-	**21**	**1.36 (0.86)**
Alkyl aryl ether	-	-	**19**	**1.12 (0.47)**
Secondary alcohol	-	-	**18**	**1.51 (0.62)**
Alkyl bromide	-	-	**17**	**0.63 (0.16)**
Diaryl ether	-	-	**13**	**1.14 (0.53)**
Carbonitrile	-	-	**12**	**1.00 (0.33)**
Oxo(het)arene	-	-	**12**	**1.99 (0.89)**
Prim. aliphatic amine	-	-	**10**	**2.86 (0.11)**
Carboxylic acid	-	-	9	2.33 (0.89)
Thioester	-	-	9	1.80 (0.55)
Alkyl iodide	-	-	9	0.27 (0.08)
Nitrate	-	-	9	1.67 (0.60)
Orthocarboxylic acid derivative	-	-	8	1.76 (0.60)
	Orthoester	-	6	1.84 (0.57)
Tertiary Alkylarylamine	-	-	7	1.81 (1.21)
	Halogen derivative	-	5	1.49 (1.48)
Tert. carboxylic acid amide	-	-	7	1.20 (0.43)
Thiophoshoric acid ester	-	-	7	3.07 (0.87)
Alkyl fluoride	-	-	6	0.90 (1.03)
Dialkylamine	-	-	6	0.83 (0.30)
Trialkylamine	-	-	6	0.46 (0.31)
Alkyne	-	-	6	0.20 (0.12)
Sec. alkylarylamine	-	-	5	2.24 (0.80)
Thiol	-	-	5	0.30 (0.08)

Table I.3: MAEs of molecules with one, two or three specific chemical function. Data points only groups containing at least five molecules. (3/3)

mass (Da)	$\max\{\lvert q_i \rvert\}$ (e)	$\max\{g(\boldsymbol{r})\}$	occurence	MAE (kcal/mol)
0-50	0.0-0.2	0-5	**5**	**0.41 (0.49)**
		5-10	1	0.09
	0.2-0.4	0-5	**5**	**0.41 (0.13)**
	0.4-0.6	0-5	4	0.56 (0.49)
		10-15	2	1.46 (0.17)
	0.6-0.8	10-15	1	0.24
	0.8-1.0	5-10	2	2.80 (0.28)
		10-15	1	0.68
	> 1.0	5-10	1	0.48
50-100	0.0-0.2	0-5	3	0.81 (0.38)
		5-10	**34**	**0.39 (0.11)**
		10-15	1	0.31
		15-20	1	18
	0.2-0.4	0-5	**7**	**0.66 (0.24)**
		5-10	**27**	**0.72 (0.13)**
		10-15	2	0.66 (0.64)
	0.4-0.6	0-5	2	0.48 (0.34)
		5-10	**20**	**0.61 (0.20)**
		10-15	**6**	**0.91 (0.28)**
		15-20	4	1.18 (0.65)
		20-25	3	1.67 (0.14)
		25-30	3	1.17 (0.41)
	0.6-0.8	5-10	**17**	**1.25 (0.18)**
		10-15	**8**	**1.07 (0.41)**
		15-20	3	1.49 (0.13)
		20-25	2	2.15 (1.55)
		> 30	2	2.31 (0.07)
	0.8-1.0	0-5	1	0.89
		5-10	**5**	**2.39 (0.97)**
		10-15	1	2.53
		15-20	1	1.25
		25-30	2	0.62 (0.28)
	> 1.0	5-10	1	0.27

Table I.4: MAEs for mass-charge-highest solvation peak triplet for $m \in [0, 100]$ Da (1/4).

mass (Da)	$\max\{\lvert q_i \rvert\}$ (e)	$\max\{g(\boldsymbol{r})\}$	occurence	MAE (kcal/mol)
100-150	0.0-0.2	0-5	1	1.17
		5-10	**41**	**0.51 (0.11)**
		10-15	**8**	**0.22 (0.13)**
		15-20	**5**	**0.33 (0.19)**
		20-25	1	0.62
	0.2-0.4	0-5	1	0.40
		5-10	**25**	**0.66 (0.14)**
		10-15	3	0.95 (0.44)
	0.4-0.6	5-10	**22**	**0.63 (0.23)**
		10-15	**11**	**0.90 (0.48)**
		15-20	**5**	**1.04 (0.57)**
		20-25	**8**	**1.09 (0.30)**
		25-30	**12**	**0.91 (0.34)**
		>30	**14**	**1.32 (0.74)**
	0.6-0.8	0-5	1	0.76
		5-10	**44**	**0.88 (0.13)**
		10-15	**17**	**1.34 (0.39)**
		15-20	**5**	**1.21 (0.59)**
		20-25	4	1.28 (0.28)
		25-30	3	0.61 (0.43)
		>30	3	2.32 (0.41)
	0.8-1.0	5-10	**9**	**1.20 (0.81)**
		10-15	**11**	**0.72 (0.61)**
		15-20	1	1.12
		20-25	2	0.86 (0.57)
	>1.0	5-10	2	2.22 (2.17)
		10-15	3	2.75 (0.53)

Table I.5: MAEs for mass-charge-highest solvation peak triplet for $m \in [100, 150]$ Da (2/4).

mass (Da)	$\max\{\lvert q_i \rvert\}$ (e)	$\max\{g(\boldsymbol{r})\}$	occurence	MAE (kcal/mol)
150-200	0.0-0.2	0-5	2	1.18 (0.12)
		5-10	**21**	**0.76 (0.19)**
		10-15	2	0.41 (0.49)
		15-20	3	0.21 (0.16)
	0.2-0.4	5-10	**16**	**1.34 (0.60)**
		10-15	2	1.34 (1.07)
	0.4-0.6	5-10	**8**	**0.65 (0.72)**
		10-15	2	1.03 (0.29)
		15-20	3	1.73 (0.28)
		20-25	2	0.99 (0.75)
		25-30	3	0.21 (0.08)
		> 30	1	2.05
	0.6-0.8	5-10	**8**	**1.54 (0.69)**
		10-15	**6**	**1.63 (0.79)**
		15-20	4	0.93 (0.38)
		20-25	2	2.42 (2.15)
		25-30	3	3.50 (2.42)
		> 30	4	0.93 (0.93)
	0.8-1.0	5-10	3	2.52 (1.85)
	> 1.0	10-15	1	2.99
200-250	0.0-0.2	0-5	**7**	**1.23 (0.57)**
		10-15	1	0.69
	0.2-0.4	5-10	**5**	**1.35 (0.70)**
	0.4-0.6	20-25	1	2.65
		25-30	1	1.76
		> 30	2	2.27 (1.71)
	0.6-0.8	5-10	2	2.65 (0.98)
		10-15	2	1.12 (1.24)
		15-20	3	1.81 (2.02)
		25-30	1	0.11
		> 30	1	1.85
	0.8-1.0	5-10	1	0.60
		10-15	**14**	**2.81 (0.51)**
		15-20	1	3.00
		20-25	2	2.35 (1.64)
		25-30	1	1.23
		> 30	1	8.81

Table I.6: MAEs for mass-charge-highest solvation peak triplet for $m \in [150, 250]$ Da (3/4).

mass (Da)	$\max\{\lvert q_i\rvert\}$ (e)	$\max\{g(\boldsymbol{r})\}$	occurence	MAE (kcal/mol)
250-300	0.0-0.2	5-10	2	1.33 (1.83)
		10-15	1	1.40
		15-20	2	1.68 (0.90)
	0.2-0.4	5-10	4	1.14 (0.91)
	0.6-0.8	10-15	1	2.11
		15-20	2	0.55 (0.16)
		20-25	2	1.41 (0.32)
		>30	1	5.16
	0.8-1.0	10-15	1	3.94
	>1.0	10-15	1	3.36
		15-20	1	1.93
300-350	0.0-0.2	10-15	2	2.31 (0.59)
	0.2-0.4	5-10	3	0.14 (0.11)
	0.6-0.18	10-15	3	0.15 (0.18)
	0.8-1.0	10-15	1	3.69
		15-20	1	0.95
		>30	2	3.12 (0.15)
	>1.0	10-15	1	3.62
		>30	1	3.68
>350	0.0-0.2	5-10	2	2.07 (1.40)
		10-15	**8**	**3.19 (0.91)**
	0.2-0.4	5-10	1	1.01
		10-15	1	0.19
	0.6-0.8	15-20	1	4.82
	0.8-1.9	>30	1	4.67
	>1.0	25-30	1	0.32

Table I.7: MAEs for mass-charge-highest solvation peak triplet for $m > 300$ Da (4/4).

BIBLIOGRAPHY

[1] O. J. Wouters, M. McKee and J. Luyten. 'Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018'. In: *JAMA* 323.9 (2020), pp. 844–853. DOI: 10.1001/jama.2020.1166 (cit. on p. vi).

[2] T. Sterling and J. J. Irwin. 'ZINC 15 - Ligand Discovery for Everyone'. In: *J. Chem. Info. Model.* 55.11 (2015), pp. 2324–2337. DOI: 10.1021/acs.jcim.5b00559 (cit. on p. vi).

[3] T. Lengauer and M. Rarey. 'Computational methods for biomolecular docking'. In: *Curr Opinion in Structural Biology* 6.3 (1996), pp. 402–406. DOI: 10.1016/S0959-440X(96)80061-3 (cit. on p. vii).

[4] D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath. 'Docking and scoring in virtual screening for drug discovery: methods and applications'. In: *Nat. Rev. Drug Discov.* 3 (11 2004). DOI: 10.1038/nrd1549 (cit. on p. vii).

[5] I. A. Guedes, F. S. S. Pereira and L. E. Dardenne. 'Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges'. In: *Front. Pharmacol.* 9 (2018). DOI: 10.3389/fphar.2018.01089 (cit. on p. vii).

[6] L. Chaput and L. Mouawad. 'Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds'. In: *J. Cheminformatics* 9 (Dec. 2017). DOI: 10.1186/s13321-017-0227-x (cit. on p. vii).

[7] B. Sherborne et al. 'Collaborating to Improve the Use of Free-Energy and Other Quantitative Methods in Drug Discovery'. In: *J. Comput. Aided Mol. Des.* (2016), p. 3. DOI: 10.1007/s10822-016-9996-y (cit. on p. vii).

[8] A. D. McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Online version (2019-) created by S. J. Chalk. Blackwell Scientific Publications, Oxford, 1997. DOI: 10.1351/goldbook (cit. on p. 1).

[9] A. Ben-Naim. *Molecular theory of solutions*. Oxford University Press, 2006 (cit. on p. 1).

[10] C. C. Bannan, G. Calabro, D. Y. Kyu and D. L. Mobley. 'Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water'. en. In: *J. Chem. Theory Comput.* 12.8 (2016), pp. 4015–4024. DOI: 10.1021/acs.jctc.6b00449 (cit. on p. 3).

[11] L. Li, T. Totton and D. Frenkel. 'Computational methodology for solubility prediction: Application to the sparingly soluble solutes'. In: *J. Chem. Phys.* 146.21 (2017), p. 214110. DOI: 10.1063/1.4983754 (cit. on p. 3).

[12] P. W. Snyder et al. 'Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase'. en. In: *Proc. Natl. Acad. Sci. U.S.A.* 108.44 (2011), pp. 17889–17894. DOI: 10.1073/pnas.1114107108 (cit. on p. 3).

[13] L. Wang, B. J. Berne and R. A. Friesner. 'Ligand binding to protein-binding pockets with wet and dry regions'. en. In: *Proc. Natl. Acad. Sci. U.S.A.* 108.4 (2011), pp. 1326–1330. DOI: 10.1073/pnas.1016793108 (cit. on p. 3).

[14] H. E. Fischer, A. C. Barnes and P. S. Salmon. 'Neutron and x-ray diffraction studies of liquids and glasses'. In: *Rep. Prog. Phys.* 69.1 (2005), pp. 233–299. DOI: 10.1088/0034-4885/69/1/r05 (cit. on p. 3).

[15] J.M. Sorenson, G. Hura, R.M. Glaeser and T. Head-Gordon. 'What can x-ray scattering tell us about the radial distribution functions of water?' In: *J. Chem. Phys.* 113.20 (2000), pp. 9149–9161. DOI: 10.1063/1.1319615 (cit. on p. 4).

[16] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik and J. B. O. Mitchell. 'A review of methods for the calculation of solution free energies and the modelling of systems in solution'. In: *Phys. Chem. Chem. Phys.* 17 (2015), pp. 6174–6191. DOI: 10.1039/C5CP00288E (cit. on pp. 4, 8).

[17] J. Puibasset and L. Belloni. 'Bridge function for the dipolar fluid from simulation'. In: *J. Chem. Phys.* 136 (2012), p. 154503. DOI: doi:10.1063/1.4703899 (cit. on pp. 4, 25, 100).

[18] L. Belloni. 'Exact molecular direct, cavity, and bridge functions in water system'. In: *J. Chem. Phys.* 147.16 (2017), p. 164121. DOI: 10.1063/1.5001684 (cit. on pp. 4, 25, 100).

[19] T. Morita and K. Hiroike. 'A New Approach to the Theory of Classical Fluids. III'. In: *Prog. Theor. Phys.* 25 (1961). DOI: 10.1143/PTP.25.537 (cit. on p. 4).

[20] T. Lazaridis. 'Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory'. In: *J. Phys. Chem. B* 102 (1998). DOI: 10.1021/jp9723574 (cit. on p. 4).

[21] T. Lazaridis. 'Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids'. In: *J. Phys. Chem. B* (1998). DOI: 10.1021/jp972358w (cit. on p. 4).

[22] T. Young, R. Abel, B. Kim, B. J. Berne and R. A. Friesner. 'Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein-Ligand Binding'. en. In: *Proc. Natl. Acad. Sci. U.S.A.* 104.3 (2007), pp. 808–813. DOI: 10.1073/pnas.0610202104 (cit. on p. 4).

[23] R. Abel, T. Young, R. Farid, B. J. Berne and R. A. Friesner. 'Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding'. In: *J. Am. Chem. Soc.* 130.9 (2008), pp. 2817–2831. DOI: 10.1021/ja0771033 (cit. on p. 4).

[24] Z. Li and T. Lazaridis. 'Computing the Thermodynamic Contributions of Interfacial Water'. In: *Computational Drug Discovery and Design*. Ed. by R. Baron. New York, NY: Springer New York, 2012, pp. 393–404. DOI: 10.1007/978-1-61779-465-0_24 (cit. on p. 4).

[25] C. N. Nguyen, T. Kurtzman Young and M. K. Gilson. 'Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril'. In: *J. Chem. Phys.* 137.4 (2012), p. 044101. DOI: 10.1063/1.4733951 (cit. on p. 4).

[26] K. R. Hadley and C. McCabe. 'Coarse-Grained Molecular Models for water : A Review'. In: *Mol. Simulat.* 38 (2012), pp. 671–681. DOI: 10.1073/pnas.0610202104 (cit. on p. 4).

[27] H.J. Berendsen, J.R. Grigera and T.P. Straatma. 'The missing term in effective pair potentials'. In: *J. Chem. Phys.* 91 (1987), p. 6269. DOI: 10.1021/j100308a038 (cit. on p. 5).

[28] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura and M. L. Klein. 'Comparison of simple potential functions for simulating liquid water'. In: *J. Chem. Phys.* 79 (1983), pp. 926–935. DOI: 10.1063/1.445869 (cit. on pp. 5, 38, 41, 106).

[29] P. L. Silvestrelli and M. Parrinello. 'Structural, electronic, and bonding properties of liquid water from first principles'. In: *J. Chem. Phys.* 111 (1999). DOI: 10.1063/1.479638 (cit. on p. 5).

[30] I. M. Kusalik P. G.; Svishchev. 'The Spatial Structure in Liquid Water'. In: *Science* (1994). DOI: 10.1126/science.265.5176.1219 (cit. on p. 5).

[31] M. Chaplin. *Water Model*. URL: http://www1.lsbu.ac.uk/water/water_models.html (visited on 05/05/2020) (cit. on p. 5).

[32] United States Environmental Protection Agency (EPA). *Basic of Green Chemisty*. URL: https://www.epa.gov/greenchemistry/basics-green-chemistry#twelve (visited on 10/07/2020) (cit. on p. 6).

[33] G. M. Rand and S. R. Petrocelli. *Fundamentals of aquatic toxicology: Methods and applications*. Jan. 1985 (cit. on p. 6).

[34] European Chemical Agency (ECHA). *Guidance on Information Requirements and Chemical Safety Assessment*. URL: https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment (visited on 10/07/2020) (cit. on p. 6).

[35] Organisation for Economic Co-operation and Development (OCDE). *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures*. 2019, p. 81. DOI: 10.1787/0ed2f88e-en (cit. on p. 6).

[36] L. Belloni. 'Non-equilibrium hybrid insertion/extraction through the 4th dimension in grand-canonical simulation'. In: *J. Chem. Phys.* 151.2 (2019), p. 021101. DOI: 10.1063/1.5110478 (cit. on pp. 6, 17, 32, 43).

[37] L. Gendre, R. Ramirez and D. Borgis. 'Classical density functional theory of solvation in molecular solvents: Angular grid implementation'. In: *Chem. Phys. Lett.* 474 (2009), pp. 366–370. DOI: 10.1016/j.cplett.2009.04.077 (cit. on pp. 6, 24).

[38] S. Zhao, R. Ramirez, R. Vuilleumier and D. Borgis. 'Molecular density functional theory of solvation: From polar solvents to water'. In: *J. Chem. Phys.* 134 (2011), p. 194102. DOI: 10.1063/1.3589142 (cit. on pp. 6, 24, 25, 75).

[39] D. Borgis, L. Gendre and R. Ramirez. 'Molecular Density Functional Theory: Application to Solvation and Electron-Transfer Thermodynamics in Polar Solvents'. In: *J. Phys. Chem. B* 116 (2012), pp. 2504–2512. DOI: 10.1021/jp210817s (cit. on pp. 6, 24).

[40] L. Ding, M. Levesque, D. Borgis and L. Belloni. 'Efficient molecular density functional theory using generalized spherical harmonics expansions'. In: *J. Chem. Phys.* 147.9 (2017), p. 094107. DOI: 10.1063/1.4994281 (cit. on pp. 6, 27, 76, 99).

[41] B. Roux and T. Simonson. 'Implicit solvent models'. In: *Biophys. Chem.* 78 (1999). DOI: 10.1016/s0301-4622(98)00226-9 (cit. on pp. 8, 10).

[42] C. J. Cramer and D. G. Truhlar. 'Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics'. In: *Chem. Rev.* 99 (1999). DOI: 10.1021/cr960149m (cit. on p. 8).

[43] J. Tomasi and M. Persico. 'Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent'. In: *Chem. Rev.* (1994). DOI: 10.1021/cr00031a013 (cit. on p. 8).

[44] J. Tomasi, B. Mennucci and R. Cammi. 'Quantum Mechanical Continuum Solvation Models'. In: *Chem. Rev.* 105 (2005). DOI: 10.1021/cr9904009 (cit. on pp. 8, 11).

[45] H. Reiss, H. L. Frisch and J. L. Lebowitz. 'Statistical Mechanics of Rigid Spheres'. In: *J. Chem. Phys.* 31 (1959). DOI: 10.1063/1.1730361 (cit. on p. 8).

[46] F. H. Stillinger. 'Structure in Aqueous Solutions of Nonpolar Solutes from the Standpoint of Scaled-Particle Theory'. In: *The Physical Chemistry of Aqueous System: A Symposium in Honor of Henry S. Frank on His Seventieth Birthday*. Ed. by R. L. Kay. Boston, MA: Springer US, 1973, pp. 43–60. DOI: 10.1007/978-1-4613-4511-4_3 (cit. on pp. 8, 9).

[47] R. A. Pierotti. 'A scaled particle theory of aqueous and nonaqueous solutions'. In: *Chem. Rev.* 76.6 (1976), pp. 717–726. DOI: 10.1021/cr60304a002 (cit. on p. 8).

[48] E. Gallicchio, L. Yu Zhang and R. M. Levy. 'The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators'. In: *J. Comput. Chem.* 23 (2002). DOI: 10.1002/jcc.10045 (cit. on p. 9).

[49] M. J. Holst and F. Saied. 'Numerical solution of the nonlinear Poisson-Boltzmann equation: Developing more robust and efficient methods'. In: *J. Comput. Chem.* 16.3 (1995), pp. 337–364. DOI: 10.1002/jcc.540160308 (cit. on p. 10).

[50] M. Marchi, D. Borgis, N. Levy and P. Ballone. 'A dielectric continuum molecular dynamics method'. In: *J. Chem. Phys.* 114 (2001). DOI: 10.1063/1.1348028 (cit. on p. 10).

[51] N. Levy, D. Borgis and M. Marchi. 'A dielectric continuum model of solvation for complex solutes'. In: *Comput. Phys. Commun.* 169 (2005). DOI: 10.1016/j.cpc.2005.03.018 (cit. on p. 10).

[52] P. Koehl and M. Delarue. 'AQUASOL: An efficient solver for the dipolar Poisson-Boltzmann-Langevin equation'. In: *J. Chem. Phys.* 132 (2010). DOI: 10.1063/1.3298862 (cit. on p. 10).

[53] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. H. Zhang and T. Hou. 'End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design'. In: *Chem. Rev.* 119.16 (2019), pp. 9478–9508. DOI: 10.1021/acs.chemrev.9b00055 (cit. on p. 11).

[54] L. Onsager. 'Electric Moments of Molecules in Liquids'. In: *J. Am. Chem. Soc.* 58 (1936). DOI: 10.1021/ja01299a050 (cit. on p. 11).

[55] J.A. Barker and R.O. Watts. 'Monte Carlo studies of the dielectric properties of water-like models'. In: *Mol. Phys.* 26 (1973). DOI: 10.1080/00268977300102101 (cit. on p. 11).

[56] J. G. Kirkwood and F. H. Westheimer. 'The Electrostatic Influence of Substituents on the Dissociation Constants of Organic Acids. I'. In: *J. Chem. Phys.* 6.9 (1938), pp. 506–512. DOI: 10.1063/1.1750302 (cit. on p. 11).

[57] F. H. Westheimer and J. G. Kirkwood. 'The Electrostatic Influence of Substituents on the Dissociation Constants of Organic Acids. II'. In: *J. Chem. Phys.* 6.9 (1938), pp. 513–517. DOI: 10.1063/1.1750303 (cit. on p. 11).

[58] A. Klamt, C. Moya and J. Palomar. 'A Comprehensive Comparison of the IEFPCM and SS(V)PE Continuum Solvation Methods with the COSMO Approach'. In: *J. Chem. Theory Comput.* (2015). DOI: 10.1021/acs.jctc.5b00601 (cit. on p. 11).

[59] S. Miertus, E. Scrocco and J. Tomasi. 'Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects'. In: *Chem. Phys.* 55 (1981). DOI: 10.1016/0301-0104(81)85090-2 (cit. on p. 11).

[60] E. Cances, B. Mennucci and J. Tomasi. 'A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics'. In: *J. Chem. Phys.* 107 (1997). DOI: 10.1063/1.474659 (cit. on p. 11).

[61] B. Mennucci, R. Cammi and J. Tomasi. 'Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level'. In: *J. Chem. Phys.* 109 (1998). DOI: 10.1063/1.476878 (cit. on p. 11).

[62] A. Klamt and G. Schüürmann. 'COSMO: a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient'. In: *J. Chem. Soc., Perkin Trans. 2* (1993), pp. 799–805. DOI: 10.1039/P29930000799 (cit. on p. 11).

[63] A. Klamt. 'Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena'. In: *J. Phys. Chem.* 99.7 (1995), pp. 2224–2235. DOI: 10.1021/j100007a062 (cit. on p. 11).

[64] A. Klamt. 'COSMO-RS for aqueous solvation and interfaces'. In: *Fluid Phase Equilib.* 407 (2016), pp. 152–158. DOI: 10.1016/j.fluid.2015.05.027 (cit. on p. 11).

[65] M.R. Shirts, L. Nade, D.L. Mobley and J.D. Chodera. *AlchemistryWiki*. URL: http://www.alchemistry.org/wiki/ (visited on 05/05/2020) (cit. on pp. 12, 16).

[66] M. S. Shell. *Other free energy techniques*. URL: https://sites.engineering.ucsb.edu/~shell/che210d/Other_free_energy_techniques.pdf (cit. on p. 12).

[67] R. Zwanzig. 'High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases'. In: *J. Chem. Phys.* 22 (1954). DOI: 10.1063/1.1740409 (cit. on p. 12).

[68] B. Widom. 'Some Topics in the Theory of Fluids'. In: *J. Chem. Phys.* 39.11 (1963), pp. 2808–2812. DOI: 10.1063/1.1734110 (cit. on pp. 13, 17, 43).

[69] C. Jarzynski. 'Nonequilibrium Equality for Free Energy Differences'. In: *Phys. Rev. Lett.* 78 (1997), pp. 2690–2693. DOI: 10.1103/PhysRevLett.78.2690 (cit. on p. 13).

[70] C. H. Bennett. 'Efficient estimation of free energy differences from Monte Carlo data'. In: *J. Comput. Phys.* 22 (1976), pp. 245–268. DOI: 10.1016/0021-9991(76)90078-4 (cit. on p. 15).

[71] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. 2nd ed. Computational science series 1. Academic Press, 2002 (cit. on p. 15).

[72] R. H. Ferrenberg A. M.; Swendsen. 'Optimized Monte Carlo data analysis'. In: *Phys. Rev. Lett.* 63 (1989). DOI: 10.1103/physrevlett.63.1195 (cit. on pp. 15, 19).

[73] M. Zacharias, T. P. Straatsma and J. A. McCammon. 'Separation-shifted scaling a new scaling method for Lennard-Jones interactions in thermodynamic integration'. In: *J. Chem. Phys.* 100 (1994). DOI: 10.1063/1.466707 (cit. on p. 16).

[74] T. C. Beutler; A. E. Mark; R. C. van Schaik; P. R. Gerber; W. F. van Gunsteren. 'Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations'. In: *Chem. Phys. Lett.* 222 (1994). DOI: 10.1016/0009-2614(94)00397-1 (cit. on p. 16).

[75] M. R. Shirts and J. D. Chodera. 'Statistically optimal analysis of samples from multiple equilibrium states'. In: *J. Chem. Phys.* 129.12 (2008), p. 124105. DOI: 10.1063/1.2978177 (cit. on pp. 17, 106).

[76] S. Kumar; J. M. Rosenberg; D. Bouzida; R. H. Swendsen; P. A. Kollman. 'THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method'. In: *J. Comput. Chem.* 13 (1992). DOI: 10.1002/jcc.540130812 (cit. on pp. 17, 19).

[77] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco and C. Bustamante. 'Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies'. In: *Nature* 437 (2005). DOI: 10.1038/nature04061 (cit. on p. 17).

[78] G. E. Crooks. 'Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences'. In: *Phys. Rev. E* 60 (1999), pp. 2721–2726. DOI: 10.1103/PhysRevE.60.2721 (cit. on p. 18).

[79] P. Procacci and C. Cardelli. 'Fast Switching Alchemical Transformations in Molecular Dynamics Simulations'. In: *J. Chemical. Theory Comput.* 10 (2014). DOI: 10.1021/ct500142c (cit. on p. 18).

[80] R. B. Sandberg, M. Banchelli, C. Guardiani, S. Menichetti, G. Caminati and P. Procacci. 'Efficient Nonequilibrium Method for Binding Free Energy Calculations in Molecular Dynamics Simulations'. In: *J. Chemical. Theory Comput.* 11.2 (2015), pp. 423–435. DOI: 10.1021/ct500964e (cit. on p. 18).

[81] D. Vassetti, M. Pagliai and P. Procacci. 'Assessment of GAFF2 and OPLS-AA general force fields in combination with the water models TIP3P, SPCE and OPC3 for the solvation free energy of drug-like organic molecules'. In: *J. Chemical. Theory Comput.* (2019). DOI: 10.1021/acs.jctc.8b01039 (cit. on p. 18).

[82] J. G. Kirkwood. 'Statistical Mechanics of Fluid Mixtures'. In: *J. Chem. Phys.* 3 (1935). DOI: 10.1063/1.1749657 (cit. on p. 19).

[83] G.M. Torrie and J.P. Valleau. 'Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling'. In: *J. Comput. Phys.* 23 (1977). DOI: 10.1016/0021-9991(77)90121-8 (cit. on p. 19).

[84] H. Fukunishi, O. Watanabe and S. Takada. 'On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction'. In: *J. Chem. Phys.* 116.20 (2002), pp. 9058–9067. DOI: 10.1063/1.1472510 (cit. on p. 20).

[85] W. Jiang and B. Roux. 'Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations'. In: *J. Chem. Theory Comput.* 6.9 (2010), pp. 2559–2565. DOI: 10.1021/ct1001768 (cit. on p. 20).

[86] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov and P. N. Vorontsov-Velyaminov. 'New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles'. In: *J. Chem. Phys.* 96 (1992). DOI: 10.1063/1.462133 (cit. on p. 20).

[87] X. Kong and C. L. Brooks. 'lambda-dynamics: A new approach to free energy calculations'. In: *J. Chem. Phys.* 105 (1996). DOI: 10.1063/1.472109 (cit. on p. 20).

[88] Z. Guo, C. L. Brooks and X. Kong. 'Efficient and Flexible Algorithm for Free Energy Calculations Using the lambda-Dynamics Approach'. In: *J. Phys. Chem. B* 102 (1998). DOI: 10.1021/jp972699+ (cit. on p. 20).

[89] R. Pomes, E. Eisenmesser, C. B. Post and B. Roux. 'Calculating excess chemical potentials using dynamic simulations in the fourth dimension'. In: *J. Chem. Phys.* 111 (1999). DOI: 10.1063/1.479622 (cit. on p. 20).

[90] I.R. McDonald J.-P. Hansen. *Theory of Simple Liquids, Third Edition*. 3rd ed. Academic Press, 2006 (cit. on pp. 21, 85).

[91] F. Hirata. *Molecular Theory of Solvation*. Springer, 2003 (cit. on pp. 21, 23).

[92] G. Jeanmairet. 'Une theorie de la fonctionnelle de la densite moleculaire pour la solvatation dans l'eau'. PhD thesis. Universite Pierre et Marie Curie, 2014. URL: http://www.theses.fr/fr/2014PA066122 (cit. on p. 21).

[93] L. Ding. 'Molecular Density Functional Theory under homogenous reference fluid approximation'. PhD thesis. Universite Paris-Saclay, 2017. URL: http://www.theses.fr/2017SACLV004 (cit. on pp. 21, 25, 26).

[94] L. S. Ornstein and F. Zernike. 'Accidental deviations of density and opalescence at the critical point of a single substance'. In: *Proceedings* 17 (1914), pp. 793–806 (cit. on p. 22).

[95] J. L. Lebowitz and J. K. Percus. 'Mean Spherical Model for Lattice Gases with Extended Hard Cores and Continuum Fluids'. In: *Phys. Rev.* 144 (1966), pp. 251–258. DOI: 10.1103/PhysRev.144.251 (cit. on p. 22).

[96] J. Percus and G. Yevick. 'Analysis of Classical Statistical Mechanics by Means of Collective Coordinates'. In: *Phys. Rev. (Series I)* 110 (1958). DOI: 10.1103/physrev.110.1 (cit. on p. 22).

[97] A. Kovalenko and F. Hirata. 'Potential of Mean Force between Two Molecular Ions in a Polar Molecular Solvent: A Study by the Three-Dimensional Reference Interaction Site Model'. In: *J. Phys. Chem. B* 103 (1999), pp. 7942–7957. DOI: 10.1021/jp991300+ (cit. on pp. 22, 23).

[98] L. Blum and A. J. Torruella. 'Invariant Expansion for Two-Body Correlations: Thermodynamic Functions, Scattering, and the Ornstein-Zernike Equation'. In: *J. Chem. Phys.* 56 (1972), pp. 303–310. DOI: doi:10.1063/1.1676864 (cit. on pp. 23, 26).

[99] L. Blum. 'Invariant Expansion. II. The Ornstein-Zernike Equation for Nonspherical Molecules and an Extended Solution to the Mean Spherical Model'. In: *J. Chem. Phys.* 57 (1972), pp. 1862–1869. DOI: doi:10.1063/1.1678503 (cit. on pp. 23, 26).

[100] L. Blum. 'Invariant expansion III: The general solution of the mean spherical model for neutral spheres with electostatic interactions'. In: *J. Chem. Phys.* 58 (1973). DOI: 10.1063/1.1679655 (cit. on p. 23).

[101] P. H. Fries and G. N. Patey. 'The solution of the hypernetted-chain approximation for fluids of nonspherical particles. A general method with application to dipolar hard spheres'. In: *J. Chem. Phys.* 82 (1985). DOI: 10.1063/1.448764 (cit. on p. 23).

[102] D. Chandler and H. C. Andersen. 'Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids'. In: *J. Chem. Phys.* 57 (1972), pp. 1930–1937. DOI: doi:10.1063/1.1678513 (cit. on p. 23).

[103] F. Hirata and P. J. Rossky. 'An extended rism equation for molecular polar fluids'. In: *Chem. Phys. Lett.* 83 (1981), pp. 329–334. DOI: 10.1016/0009-2614(81)85474-7 (cit. on p. 23).

[104] F. Hirata, B. M. Pettitt and P. J. Rossky. 'Application of an extended RISM equation to dipolar and quadrupolar fluids'. In: *J. Chem. Phys.* 77 (1982), pp. 509–520. DOI: doi:10.1063/1.443606 (cit. on p. 23).

[105] F. Hirata, P. J. Rossky and B. M. Pettitt. 'The interionic potential of mean force in a molecular polar solvent from an extended RISM equation'. In: *J. Chem. Phys.* 78.6 (1983), pp. 4133–4144. DOI: 10.1063/1.445090 (cit. on p. 23).

[106] D. Beglov and B. Roux. 'Numerical solution of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions'. In: *J. Chem. Phys.* 103 (1 1995). DOI: 10.1063/1.469602 (cit. on p. 23).

[107] D. Beglov and B. Roux. 'Solvation of Complex Molecules in a Polar Liquid: An Integral Equation Theory'. In: *J. Chem. Phys.* 104 (1996), pp. 8678–8689. DOI: 10.1063/1.471557 (cit. on p. 23).

[108] M. Ikeguchi and J. Doi. 'Direct numerical solution of the Ornstein-Zernike integral equation and spatial distribution of water around hydrophobic molecules'. In: *J. Chem. Phys.* 103.12 (1995), pp. 5011–5017. DOI: 10.1063/1.470587 (cit. on p. 23).

[109] C. M. Cortis, P. J. Rossky and R. A. Friesner. 'A three-dimensional reduction of the OrnsteinâZernicke equation for molecular liquids'. In: *J. Chem. Phys.* 107 (1997). DOI: 10.1063/1.474300 (cit. on p. 23).

[110] A. Kovalenko and F. Hirata. 'Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach'. In: *Chem. Phys. Lett.* 290 (1998), pp. 237–244. DOI: 10.1016/S0009-2614(98)00471-0 (cit. on p. 23).

[111] S. J. Singer and D. Chandler. 'Free energy functions in the extended RISM approximation'. In: *Mol. Phys.* 55 (1985). DOI: 10.1080/00268978500101591 (cit. on p. 23).

[112] T. Imai, A. Kovalenko and F. Hirata. 'Solvation Thermodynamics of Protein Studied by the 3D-RISM Theory'. In: *Chem. Phys. Lett.* 395.1-3 (2004), pp. 1–6. DOI: 10.1016/j.cplett.2004.06.140 (cit. on p. 23).

[113] I. Omelyan and A. Kovalenko. 'MTS-MD of Biomolecules Steered with 3D-RISM-KH Mean Solvation Forces Accelerated with Generalized Solvation Force Extrapolation'. In: *J. Chem. Theory Comput.* 11.4 (2015), pp. 1875–1895. DOI: 10.1021/ct5010438 (cit. on p. 23).

[114] D. Roy and A. Kovalenko. 'Performance of 3D-RISM-KH in Predicting Hydration Free Energy: Effect of Solute Parameters'. In: *J. Phys. Chem. A* 123.18 (2019), pp. 4087–4093. DOI: 10.1021/acs.jpca.9b01623 (cit. on pp. 23, 85).

[115] N. Ruankaew, N. Yoshida and S. Phongphanphanee. 'Solvated lithium ions in defective Prussian blue'. In: *IOP Conf. Ser.: Mat. Sci. Eng.* 526 (2019), p. 012032. DOI: 10.1088/1757-899x/526/1/012032 (cit. on p. 23).

[116] N. Tielker, D. Tomazic, L. Eberlein, S. Gussregen and S. M. Kast. 'The SAMPL6 challenge on predicting octanolâwater partition coefficients from ECâRISM theory'. In: *J. Compt. Aided Mol. Des.* 34 (2020), pp. 453–461. DOI: 0.1007/s10822-020-00283-4 (cit. on p. 23).

[117] F. Hirata, H. Sato, S. Ten-No and S. Kato. 'The RISM-SCF/MCSCF Approach for the Chemical Processes in Solutions'. In: *Computational biochemistry and biophysics*. Ed. by O. M. Becker, A. D. MacKerell Jr., B. Roux and M. Watanabe. 1st. New York: M. Dekker Inc., 2001 (cit. on pp. 23, 99).

[118] T. Kloss, J. Heil and S. M. Kast. 'Quantum Chemistry in Solution by Combining 3D Integral Equation Theory with a Cluster Embedding Approach'. In: *J. Phys. Chem. B* 112 (2008). DOI: 10.1021/jp710680m (cit. on pp. 23, 99).

[119] M. Kinoshita, Y. Okamoto and F. Hirata. 'Solvent effects on conformational stability of peptides: RISM analyses'. In: *J. Mol. Liq.* 90 (2001). DOI: 10.1016/s0167-7322(01)00122-2 (cit. on p. 23).

[120] B. Kim and F. Hirata. 'Structural fluctuation of protein in water around its native state: A new statistical mechanics formulation'. In: *J. Chem. Phys.* 138.5 (2013), p. 054108. DOI: 10.1063/1.4776655 (cit. on p. 23).

[121] P. Hohenberg and W. Kohn. 'Inhomogeneous Electron Gas'. In: *Phys. Rev.* 136 (1964), B864. DOI: 10.1103/PhysRev.136.B864 (cit. on p. 24).

[122] W. Kohn and L. J. Sham. 'Self-Consistent Equations Including Exchange and Correlation Effects'. In: *Phys. Rev.* 140 (1965), A1133. DOI: 10.1103/PhysRev.140.A1133 (cit. on p. 24).

[123] N. Mermin. 'Thermal Properties of the Inhomogeneous Electron Gas'. In: *Phys. Rev. (Series I)* 137 (1965). DOI: 10.1103/physrev.137.a1441 (cit. on p. 24).

[124] R. Evans. 'The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids'. In: *Adv. Phys.* 28 (1979). DOI: 10.1080/00018737900101365 (cit. on p. 24).

[125] R. Evans. *Fundamentals of Inhomogeneous Fluids*. Ed. by D. Henderson. Marcel Dekker, Incorporated, 1992 (cit. on p. 24).

[126] G. Jeanmairet, M. Levesque and D. Borgis. 'Tackling solvent effects by coupling electronic and molecular Density Functional Theory'. In: *J. Chem. Theory Comput.* (2020). DOI: 10.1021/acs.jctc.0c00729 (cit. on pp. 25, 99).

[127] L. Belloni and J. Puibasset. 'Finite-size corrections in simulation of dipolar fluids'. In: *J. Chem. Phys.* 147 (2017). DOI: 10.1063/1.5005912 (cit. on p. 25).

[128] M. Levesque, R. Vuilleumier and D. Borgis. 'Scalar fundamental measure theory for hard spheres in three dimensions: Application to hydrophobic solvation'. In: *J. Chem. Phys.* 137 (2012), p. 034115. DOI: 10.1063/1.4734009 (cit. on pp. 25, 60).

[129] G. Jeanmairet, M. Levesque and D. Borgis. 'Molecular density functional theory of water describing hydrophobicity at short and long length scales'. In: *J. Chem. Phys.* 139 (2013), pp. 154101–1–154101–9. DOI: 10.1063/1.4824737 (cit. on pp. 25, 60).

[130] G. Jeanmairet, M. Levesque, V. Sergiievskyi and D. Borgis. 'Molecular Density Functional Theory for Water with Liquid-Gas Coexistence and Correct Pressure'. In: *J. Chem. Phys.* 142 (2015), p. 154112. DOI: 10.1063/1.4917485 (cit. on pp. 25, 60).

[131] C. Gageat, D. Borgis and M. Levesque. 'Bridge functional for the molecular density functional theory with consistent pressure and surface tension'. In: (2017). ArXiv:1709.10139 (cit. on pp. 25, 60).

[132] D. Borgis, S. Luukkonen, L. Belloni and G. Jeanmairet. 'Simple parameter-free bridge functionals for molecular density functional theory. Application to hydrophobic solvation'. In: *J. Chem. Phys. B* 124 (2020), pp. 6885–6893. DOI: 10.1021/acs.jpcb.0c04496 (cit. on pp. 25, 60, 97).

[133] R. Ramirez, R. Gebauer, M. Mareschal and D. Borgis. 'Density functional theory of solvation in a polar solvent: Extracting the functional from homogeneous solvent simulations'. In: *Phys. Rev. E* 66 (2002), pp. 031206–8. DOI: 10.1103/PhysRevE.66.031206 (cit. on p. 25).

[134] R. Ramirez, M. Mareschal and D. Borgis. 'Direct correlation functions and the density functional theory of polar solvents'. In: *Chem. Phys.* 319 (2005), pp. 261–272. DOI: 10.1016/j.chemphys.2005.07.038 (cit. on p. 25).

[135] R. H. Byrd, P. Lu, J. Nocedal and C. Zhu. 'A Limited Memory Algorithm for Bound Constrained Optimization'. In: *SIAM J. Sci. Comput.* 16 (1994), pp. 1190–1208. DOI: 10.1137/0916069 (cit. on p. 26).

[136] C. Zhu, R. H. Byrd, P. Lu and J. Nocedal. 'Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization'. In: *ACM Trans. Math. Softw.* 23 (1997), pp. 550–560. DOI: 10.1145/279232.279236 (cit. on p. 26).

[137] M. R. Shirts, D. L. Mobley, J. D. Chodera and V. S. Pande. 'Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations'. In: *J. Phys. Chem. B* 111 (2007). DOI: 10.1021/jp0735987 (cit. on p. 27).

[138] M. A. Kastenholz and P. H. Hunenberger. 'Computation of methodology-independent ionic solvation free energies from molecular simulations. II. The hydration free energy of the sodium cation'. In: *J. Chem. Phys.* 124 (2006), p. 224501. DOI: 10.1063/1.2201698 (cit. on pp. 28, 39, 76).

[139] M. A. Kastenholz and P. H. Hunenberger. 'Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids'. In: *J. Chem. Phys.* 124 (2006), p. 124106. DOI: 10.1063/1.2172593 (cit. on pp. 28, 39, 76).

[140] B. R. A. Nijboer and Th. W. Ruijgrok. 'On the energy per particle in three- and two-dimensional Wigner lattices'. In: *J. Stat. Phys.* 53 (1988). DOI: 10.1007/bf01011562 (cit. on p. 28).

[141] D. Borgis, R. Assaraf, B. Rotenberg and R. Vuilleumier. 'Computation of pair distribution functions and three-dimensional densities with a reduced variance principle'. In: *Mol. Phys.* 111.22-23 (2013), pp. 3486–3492. DOI: 10.1080/00268976.2013.838316 (cit. on p. 28).

[142] D. de las Heras and M. Schmidt. 'Better Than Counting: Density Profiles from Force Sampling'. In: *Phys. Rev. Lett.* 120 (2018), p. 218001. DOI: 10.1103/PhysRevLett.120.218001 (cit. on p. 28).

[143] S. W. Coles, D. Borgis, R. Vuilleumier and B. Rotenberg. 'Computing three-dimensional densities from force densities improves statistical efficiency'. In: *J. Chem. Phys.* 151 (2019). DOI: 10.1063/1.5111697 (cit. on pp. 28, 80).

[144] G. Duarte Ramos Matos, D. Y. Kyu, H. H. Loeffler, J. D. Chodera, M. R. Shirts and D. L. Mobley. 'Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database'. In: *J. Chem. Eng. Data* 62.5 (2017), pp. 1559–1569. DOI: 10.1021/acs.jced.7b00104 (cit. on pp. 32, 41, 49, 51, 85, 106).

[145] H.J.C. Berendsen, D. van der Spoel and R. van Drunen. 'GROMACS: A message-passing parallel molecular dynamics implementation'. In: *Comput. Phys. Commun.* 91 (1995). DOI: 10.1016/0010-4655(95)00042-e (cit. on pp. 37, 106).

[146] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl. 'GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation'. In: *J. Chem. Theory Comput.* 4 (2008). DOI: 10.1021/ct700301q (cit. on pp. 37, 106).

[147] S. Pronk et al. 'GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit'. In: *Bioinformatics* 29 (2013). DOI: 10.1093/bioinformatics/btt055 (cit. on pp. 37, 106).

[148] S. Páll, M. J. Abraham, C. Kutzner, B. Hess and E. Lindahl. 'Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS'. In: *Solving Software Challenges for Exascale*. Ed. by S. Markidis and E. Laure. Cham: Springer International Publishing, 2015, pp. 3–27 (cit. on pp. 37, 106).

[149] M. J. Abraham, T. Murtola, R. Schulz, S. PÃ¡ll, J. C. Smith, B. Hess and E. Lindahl. 'GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers'. In: *SoftwareX* (July 2015). DOI: 10.1016/j.softx.2015.06.001 (cit. on pp. 37, 106).

[150] T. P. Straatsma, H. J. C. Berendsen and J. P. M. Postma. 'Free energy of hydrophobic hydration: A molecular dynamics study of noble gases in water'. In: *J. Chem. Phys.* 85 (1986). DOI: 10.1063/1.451846 (cit. on pp. 38, 75).

[151] D. Horinek, S.I. Mamatkulov and R.R. Netz. 'Rational design of ion force fields based on thermodynamic solvation properties'. In: *J. Chem. Phys.* 130 (2009), p. 124507. DOI: 10.1063/1.3081142 (cit. on pp. 39, 76).

[152] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts and K. A. Dill. 'Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations'. In: *J. Chem. Theory Comput.* 5 (2009), pp. 350–358. DOI: 10.1021/ct800409d (cit. on pp. 41, 51).

[153] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case. 'Development and Testing of a General AMBER Force Field'. In: *J. Comput. Chem.* 25.9 (2004), pp. 1157–1174. DOI: 10.1002/jcc.20035 (cit. on pp. 41, 106).

[154] A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly. 'Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method'. In: *J. Comput. Chem.* 21.2 (2000), pp. 132–146. DOI: 10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P (cit. on pp. 41, 106).

[155] A. Jakalian, D. B. Jack and C. I. Bayly. 'Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and validation'. In: *J. Comput. Chem.* 23.16 (2002), pp. 1623–1641. DOI: 10.1002/jcc.10128 (cit. on pp. 41, 106).

[156] S. Luukkonen, L. Belloni, D. Borgis and M. Levesque. 'Effect of solute flexibility for the hydration free energies of the FreeSolv database'. In: (2020). in preparation (cit. on p. 49).

[157] G. Rickayzen and A. Augousti. 'Integral equations and the pressure at the liquid-solid interface'. In: *Mol. Phys.* 52 (1984). DOI: 10.1080/00268978400101971 (cit. on p. 60).

[158] V.P. Sergiievskyi, G. Jeanmairet, M. Levesque and D. Borgis. 'Fast Computation of Solvation Free Energies with Molecular Density Functional Theory: Thermodynamic-Ensemble Partial Molar Volume Corrections'. In: *J. Phys. Chem. Lett.* 5 (2014), pp. 1935–1942. DOI: 10.1021/jz500428s (cit. on p. 61).

[159] V.P. Sergiievskyi, G. Jeanmairet, M. Levesque and D. Borgis. 'Solvation free-energy pressure corrections in the three dimensional reference interaction site model'. In: *J. Chem. Phys.* 143 (2015), p. 184116. DOI: 10.1063/1.4935065 (cit. on p. 61).

[160] W. J. Huang, N. Blinov and A. Kovalenko. 'Octanol Water Partition Coefficient from 3D-RISM-KH Molecular Theory of Solvation with Partial Molar Volume Correction'. In: *J. Phys. Chem. B* 119 (2015). DOI: 10.1021/acs.jpcb.5b01291 (cit. on p. 61).

[161] M. Misin, M. V. Fedorov and D. S. Palmer. 'Hydration Free Energies of Ionic Species by Molecular Theory and Simulation'. In: *J. Phys. Chem. B* (2016). DOI: 10.1021/acs.jpcb.5b10809 (cit. on pp. 61, 70).

[162] M. Misin, D. S. Palmer and M. V. Fedorov. 'Predicting Solvation Free Energies Using Parameter-Free Solvent Models'. In: *J. Phys. Chem. B* (2016). DOI: 10.1021/acs.jpcb.6b05352 (cit. on p. 61).

[163] D. Chandler and P. Varilly. 'Lectures on Molecular- and Nano-scale Fluctuations in Water'. In: *arXiv: Soft Condensed Matter* (2011). DOI: 10.3254/978-1-61499-071-0-75 (cit. on p. 61).

[164] A. Robert, S. Luukkonen and M. Levesque. 'Pressure correction for solvation theories'. In: *J. Chem. Phys.* 152.19 (2020), p. 191103. DOI: 10.1063/5.0002029 (cit. on pp. 62, 88).

[165] A. Bondi. 'van der Waals Volumes and Radii'. In: *J. Chem. Phys.* 68 (1964). DOI: 10.1021/j100785a001 (cit. on p. 63).

[166] F. Gao and L. Han. 'Implementing the Nelder-Mead simplex algorithm with adaptive parameters'. In: *Comput. Optim. Appl.* 51 (2012). DOI: 10.1007/s10589-010-9329-3 (cit. on p. 63).

[167] ajd98. 2019. URL: https://github.com/ajd98/molecularvolume (cit. on p. 63).

[168] D. C. Elton, Z. Boukouvalas, M. D. Fuge and W. Chung P. 'Deep learning for molecular design - a review of the state of the art'. In: *Mol. Syst. Des. Eng.* (2019). DOI: 10.1039/C9ME00039A (cit. on p. 65).

[169] F. Musil and M. Ceriotti. 'Machine Learning at the Atomic Scale'. In: *Chimia* 73 (2019), pp. 972–982. DOI: 10.2533/chimia.2019.972 (cit. on p. 65).

[170] B. Ramsundar, P. Eastman, P. Walters and V. Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. 1st ed. O'Reilly Media, 2019. Chap. 4 (cit. on p. 65).

[171] MACCS keys, MDL Information Systems Inc., San Leandro, CA. (cit. on pp. 65, 110).

[172] *RDKit: Open-source cheminformatics*. URL: http://www.rdkit.org (cit. on p. 66).

[173] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830 (cit. on pp. 66, 110).

[174] F. M. Floris, M. Selmi, A. Tani and J. Tomasi. 'Free energy and entropy for inserting cavities in water: Comparison of Monte Carlo simulation and scaled particle theory results'. In: *J. Chem. Phys.* 107 (1997). DOI: 10.1063/1.474296 (cit. on p. 68).

[175] H. S. Ashbaugh and L. R. Pratt. 'Colloquium: Scaled particle theory and the length scales of hydrophobicity'. In: *Rev. Mod. Phys.* 78 (2006). DOI: 10.1103/revmodphys.78.159 (cit. on p. 68).

[176] D. M. Huang, P. L. Geissler and D. Chandler. 'Scaling of Hydrophobic Solvation Free Energies'. In: *J. Phys. Chem. B* 105 (2001). DOI: 10.1021/jp0104029 (cit. on p. 68).

[177] D. M. Huang and D. Chandler. 'The Hydrophobic Effect and the Influence of SoluteâSolvent Attractions'. In: *J. Phys. Chem. B* 106 (2002). DOI: 10.1021/jp013289v (cit. on pp. 68, 69).

[178] G. Hummer, S. Garde, A. E. Garcia, A. Pohorille and L. R. Pratt. 'An information theory model of hydrophobic interactions.' In: *Proc. Natl. Acad. Sci. U.S.A.* 93 (1996). DOI: 10.1073/pnas.93.17.8951 (cit. on pp. 68, 69).

[179] C. Vega and E. de Miguel. 'Surface tension of the most popular models of water by using the test-area simulation method'. In: *J. Chem. Phys.* 126 (2007). DOI: 10.1063/1.2715577 (cit. on pp. 68, 70).

[180] S. Luukkonen, M. Levesque, L. Belloni and D. Borgis. 'Hydration free energies and solvation structures with molecular density functional theory in the hyper-netted chain approximation'. In: *J. Chem. Phys.* 152 (2020), p. 064110. DOI: 10.1063/1.5142651 (cit. on pp. 74, 88).

[181] J. Richardi, C. Millot and P. H. Fries. 'A molecular Ornstein-Zernike study of popular models for water and methanol'. In: *J. Chem. Phys.* 110 (1999). DOI: 10.1063/1.478171 (cit. on p. 80).

[182] M. Lombardero, C. Martin, S. Jorge, F. Lado and E. Lomba. 'An integral equation study of a simple point charge model of water'. In: *J. Chem. Phys.* 110.2 (1999), pp. 1148–1153. DOI: 10.1063/1.478156 (cit. on p. 80).

[183] S. Luukkonen, L. Belloni, D. Borgis and M. Levesque. 'Predicting hydration free energies of the FreeSolv database of druglike molecules with molecular density functional theory'. In: *J. Chem. Info. Model* 60 (2020), 3558â3565. DOI: 10.1021/acs.jcim.0c00526 (cit. on p. 84).

[184] D. S. Palmer, A. I. Frolov, E. L. Ratkova and M. V. Fedorov. 'Towards a Universal Method for Calculating Hydration Free Energies: A 3D Reference Interaction Site Model with Partial Molar Volume Correction'. In: *J. Phys. - Condens. Mat.* 22 (2010), pp. 492101–1–492101–9. DOI: 10.1088/0953-8984/22/49/492101 (cit. on p. 85).

[185] D. Chandler et al. 'General Discussion'. In: *Faraday Disc. Chem. Soc.* 66 (1978), pp. 71–94. DOI: 10.1039/DC9786600071 (cit. on p. 85).

[186] D. E. Sullivan and C.G. Gray. 'Evaluation of Angular Correlation Parameters and the Dielectric Constant in the RISM Approximation'. In: *Mol. Phys.* 42.2 (1981), pp. 443–454. DOI: 10.1080/00268978100100381 (cit. on p. 85).

[187] G.P. Morriss and J.W. Perram. 'Polar hard dumb-bells and a RISM model for water'. In: *Mol. Phys.* 43.3 (1981), pp. 669–684. DOI: 10.1080/00268978100101591 (cit. on p. 85).

[188] C. J. Fennell, K. L. Wymer and D. L. Mobley. 'A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration'. In: *J. Phys. Chem. B* 118.24 (2014), pp. 6438–6446. DOI: 10.1021/jp411529h (cit. on p. 90).

[189] G. N. Chuev and M. V. Fedorov. 'Wavelet algorithm for solving integral equations of molecular liquids. A test for the reference interaction site model'. In: *J. Chem. Phys.* 25 (2004). DOI: 10.1002/jcc.20068 (cit. on p. 98).

[190] L. Genovese et al. 'Daubechies wavelets as a basis set for density functional pseudopotential calculations'. In: *J. Chem. Phys.* 129.1 (2008), p. 014109. DOI: 10.1063/1.2949547 (cit. on p. 98).

[191] *BigDFT*. URL: http://bigdft.org/Wiki/index.php?title=BigDFT_website (cit. on p. 98).

[192] *Multiresolution Adaptive Numerical Environment for Scientific Simulation*. URL: https://github.com/m-a-d-n-e-s-s/madness (cit. on p. 98).

[193] M. Levesque, M. Duvail, I. Pagonabarraga, D. Frenkel and B. Rotenberg. 'Accounting for adsorption and desorption in lattice Boltzmann simulations'. In: *Phys. Rev. E* 88 (2013). DOI: 10.1103/PhysRevE.88.013308 (cit. on p. 99).

[194] J.-M. Vanson, F.-X. Coudert, M. Rotenberg B. Levesque, C. Tardivat, M. Klotz and A. Boutin. 'Unexpected coupling between flow and adsorption in porous media'. In: *Soft Matter* 11 (30 2015). DOI: 10.1039/C5SM01348H (cit. on p. 99).

[195] A. J. Asta, M. Levesque, R. Vuilleumier and B. Rotenberg. 'Transient hydrodynamic finite-size effects in simulations under periodic boundary conditions'. In: *Phys. Rev. E* 95 (2017). DOI: 10.1103/PhysRevE.95.061301 (cit. on p. 99).

[196] S. Genheden and U. Ryde. 'The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities'. In: *Expert Opin. Drug Dis.* 10.5 (2015), pp. 449–461. DOI: 10.1517/17460441.2015.1032936 (cit. on p. 99).

[197] C. Gageat. 'Solvatation de systemes d'interet pharmaceutique; apports de la fonctionnelle de la densite moleculaire'. PhD thesis. PSL Research University, 2017. URL: http://www.theses.fr/2017PSLEE047 (cit. on p. 99).

[198] M. Schauperl, T. S. Podewitz M. Ortner, F. Waibl, Al. Thoeny, T. Loerting and K. R. Liedl. 'Balance between hydration enthalpy and entropy is important for ice binding surfaces in Antifreeze Proteins'. In: *Scientific Reports* 7 (2017). DOI: 10.1038/s41598-017-11982-8 (cit. on p. 100).

[199] T. H. Parsell, M.-Y. Yang and A. S. Borovik. 'CâH Bond Cleavage with Reductants: Re-Investigating the Reactivity of Monomeric MnIII/IVâOxo Complexes and the Role of Oxo Ligand Basicity'. In: *J. Am. Chem. Soc.* 131 (2009). DOI: 10.1021/ja8100825 (cit. on p. 100).

[200] MobleyLab. *FreeSolv*. URL: https://github.com/MobleyLab/FreeSolv (cit. on p. 106).

[201] R. C. Rizzo, T. Aynechi, D. A. Case and I. D. Kuntz. 'Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions'. In: *J. Chem. Theory Model.* 2 (1 2006). DOI: 10.1021/ct050097l (cit. on p. 106).

[202] SAMPL. *The SAMPL challenges*. URL: https://samplchallenges.github.io/ (cit. on p. 106).

[203] D. L. Mobley and J. P. Guthrie. 'FreeSolv: a database of experimental and calculated hydration free energies, with input files'. In: *J. Comput. Aided Mol. Des.* 28 (7 2014). DOI: 10.1007/s10822-014-9747-x (cit. on p. 106).

[204] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney. 'Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings'. In: *Adv. Drug Deliver. Rev.* 23 (1997), pp. 3–25. DOI: 10.1016/S0169-409X(96)00423-1 (cit. on p. 106).

[205] M. Congreve, R. Carr, C. Murray and H. Jhoti. 'A "Rule of Three" for Fragment-Based Lead Discovery?' In: *Drug Discov. Today* 8 (2003), pp. 876–877. DOI: 10.1016/S1359-6446(03)02831-9 (cit. on p. 106).

[206] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas and H. S. Seung. In: *Nature* 405 (2000). DOI: 10.1038/35016072 (cit. on p. 110).

[207] P. Ramachandran, B. Zoph and Q. V. Le. *Searching for Activation Functions*. 2018. URL: https://openreview.net/forum?id=SkBYYyZRZ (cit. on p. 110).

[208] D. P. Kingma and J. Ba. 'Adam: A Method for Stochastic Optimization'. In: (2014). ArXiv:1412.6980v9 (cit. on p. 110).

[209] S. Ruder. 'An overview of gradient descent optimization algorithms'. In: (2016). ArXiv:1609.04747 (cit. on p. 110).

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences Chimiques: Molécules, Matériaux, Instrumentation et Biosystèmes (2MIB)

Titre: Hydratation de molécules de type pharmaceutique avec la théorie de la fonctionnelle de densité moléculaire et l'approche Monte-Carlo hybride à 4ème dimension

Mots clés: solvatation, énergie libre d'hydratation, mécanique statistique, théorie des liquides

Résumé: Le développement d'un médicament prend en moyenne plus de 10 ans pour un coût de 1 Mrd de $. Pour accélérer le processus et diminuer le coût, on utilise des méthodes in-silico lors de l'étape de découverte qui consiste à cribler $\sim 10^5$ molécules de type médicament pour proposer quelques candidats à l'étape préclinique. Le critère majeur est l'affinité entre la molécule potentielle et la cible biologique. L'interaction se passant dans notre corps: cette affinité doit être prédite dans l'eau et le médicament doit être soluble dans l'eau pour avoir accès à la cible. Globalement, les effets de solvatation ont un rôle important dans la conception de médicaments.

Numériquement, pour un champ de force donné, la solvatation peut être étudiée par des méthodes de simulation exactes mais coûteuses, par des modèles de continuum rapides mais qui ignorent la nature moléculaire du solvant, enfin par des théories des liquides approximatives mais capables de garder l'information moléculaire du solvant tout en diminuant le temps de calcul. L'objectif de cette thèse étant la prédiction des énergies libres d'hydratation (ELH) de molécules de type médicament par des méthodes qui soient les plus précises et plus rapides possibles, elle se concentre sur deux approches originales:

Le Monte-Carlo hybride à 4ème dimension, une nouvelle méthode pour calculer les ELH selon le principe de Jarzynski à partir simulations courtes hors-équilibre pendant lesquelles on introduit ou retire le soluté doucement depuis le solvant avec un paramètre de couplage dépendant du temps. Nous montrons que cette approche est capable de prédire les ELH de molécules de type pharmaceutique 4-6 fois plus rapidement que l'approche classique de perturbation de l'énergie libre.

La théorie de la fonctionnelle de densité moléculaire, une approche de théorie des liquides qui permet l'étude des propriétés de solvatation de n'importe quelle soluté rigide. Dans son état actuel, dans l'approximation hyper-netted-chain couplée à une correction de pression, nous montrons qu'elle est capable de prédire les ELH des mêmes molécules avec une précision de respectivement 0.5 ou 1.0 kcal/mol par rapport aux simulations ou aux données expérimentales, avec une accélération de calcul de l'ordre de 10^3-10^4 par rapport aux simulations.

H4D-MC est considéré ici comme une source de références pour développer plus avant la MDFT, elle-même une méthode suffisamment rapide pour être envisagée dans un processus de criblage haut-débit.

Title: Hydration of drug-like molecules with molecular density functional theory and the hybrid-4th-dimension Monte Carlo approach

Keywords: solvation, hydration free energy, statistical mechanics, liquid state theory

Abstract: The development of a drug takes on average over 10 yr. for a cost of 1B$. To speed up the process, and reduce its cost, in-silico methods are used at the drug discovery stage. It consists of screening $\sim 10^5$ drug-like molecules to propose few candidates to the pre-clinical stages. The main criterion is the affinity between the potential drug molecule and biological target. As the interaction happens the body, these affinities need to be predicted in water and the molecule needs to be water-soluble to access the receptor. Overall, solvation properties play an important role in drug design.

Numerically, for a given force-field, solvation can be studied either with exact but time-consuming simulation methods, fast continuum models that lose the molecular nature of the solvent, or approximate liquid state theories that keep the solvent molecular information while speeding-up the computation. In this thesis, we focus on the prediction of the hydration free energies (HFE) of drug-like molecules with methods that are as fast and precise as possible, and we concentrate on two original approaches:

Hybrid-4th-dimension Monte Carlo, a novel method that computes the HFEs according to the Jarzynski principle from short non-equilibrium simulations in which the solute is inserted or removed from the solvent with a time-depending coupling parameter. This approach is shown to predict the HFEs of drug-like molecules 4-6 times faster than the classical free energy perturbation approach.

Molecular density functional theory, a liquid-state-theory approach that allows the study of the equilibrium solvation properties of any rigid solute. In its current level, the hyper netted-chain approximation coupled with a pressure correction, it is shown to predict the HFEs of drug-like molecules within 0.5 and 1.0 kcal/mol of simulations and experimental data, respectively, for an average computational speed-up 10^3-10^4 with respect to simulations.

H4D-MC is considered here as a source of reference data for MDFT developments. MDFT is itself fast enough to be foreseen in a high-throughput screening pipeline.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France