Polarizable Force Field Development, and Applications to Conformational Sampling and Free Energy Calculation
by
Zhi Wang

A dissertation presented to
The Graduate School
of Washington University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2018
St. Louis, Missouri
# Table of Contents

- List of Figures ................................................................. v
- List of Tables ................................................................. ix
- List of Programs ............................................................... xi
- List of Symbols and Abbreviations ..................................... xii
- Acknowledgments ............................................................. xiii
- Abstract .............................................................................. xvi

## Chapter 1 Introduction ....................................................... 1

### 1.1 AMOEBA Force Field ................................................... 2

#### 1.1.1 valence terms ......................................................... 3
#### 1.1.2 van der Waals interaction ......................................... 3
#### 1.1.3 permanent multipole electrostatic energy ...................... 6
#### 1.1.4 polarization energy .................................................. 8

### 1.2 Common Methodologies ............................................... 11

#### 1.2.1 integrators .............................................................. 12
#### 1.2.2 thermostats and barostats .......................................... 14
#### 1.2.3 radial distribution function (RDF) ............................. 15
#### 1.2.4 diffusion coefficient ................................................. 17
#### 1.2.5 free energy calculation .............................................. 18

### 1.3 Supplementary Information ........................................... 22

## Chapter 2 Monovalent Ions for the AMOEBA Force Field ....... 27

### 2.1 Introduction ............................................................... 27

### 2.2 Methodologies ............................................................ 28

### 2.3 Revision of the Monovalent Ions in the AMOEBA Force Field 29

#### 2.3.1 Na$^+$ and water ..................................................... 30
#### 2.3.2 single ion solvation free energy ................................ 31
#### 2.3.3 lattice energies and lattice constants ......................... 32

### 2.4 Results and Discussion ............................................... 36

#### 2.4.1 Na$^+$ parameters ................................................... 36
#### 2.4.2 single ion solvation free energy ................................. 39
#### 2.4.3 mean ionic activity coefficient .................................. 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>Results and Discussion</td>
<td>111</td>
</tr>
<tr>
<td>4.4.1</td>
<td>submitted predictions</td>
<td>111</td>
</tr>
<tr>
<td>4.4.2</td>
<td>CB[8] parameters</td>
<td>113</td>
</tr>
<tr>
<td>4.4.3</td>
<td>key torsions and intra-molecular vdw</td>
<td>117</td>
</tr>
<tr>
<td>4.4.4</td>
<td>revised predictions</td>
<td>121</td>
</tr>
<tr>
<td>4.5</td>
<td>Conclusion and Acknowledgement</td>
<td>123</td>
</tr>
<tr>
<td>4.6</td>
<td>Supplementary Information</td>
<td>124</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Conclusions and Future Perspectives</td>
<td>133</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>135</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Combined value of $x$ and 1 with different combining rules. .......................... 5  
1.2 Comparison of the curvatures of Lennard-Jones 12-6 vdw and buffered 14-7 vdw with the same $r''$ parameter. .......................................................... 6  
1.3 Definition of multipole sites A and B. ................................................................. 7  
1.4 Schematic sketch of the RDF of liquid argon. ....................................................... 16  
2.1 Solvation process for FEP calculation. ................................................................. 29  
2.2 MP2 optimized Na$^+$(H$_2$O) structure. ............................................................ 30  
2.3 Optimized ion-oxygen distance $R$ of Na$^+$(H$_2$O) and electronic binding energy $D_e$ with different combinations of vdw parameters. $R$ and $D_e$ calculated by the MP2 method are represented by the dashed lines. The optimum combination of parameters that gives the closest agreement is $r^0 = 2.955$ Å, and $\epsilon = 0.28$ kcal/mol. ................................................................. 31  
2.4 Electronic binding energy $D_e$ of Na$^+$ (H$_2$O) with respect to different Na$^+$-O distances ($R$) calculated by BSSE corrected MP2 and by the revised Na$^+$ parameters in the AMOEBA force field. ................................................................. 37  
2.5 Optimized Na$^+$(H$_2$O)$_4$ and Na$^+$(H$_2$O)$_6$ structures. Na$^+$(H$_2$O)$_4$ is of $S_4$ symmetry. Na$^+$(H$_2$O)$_6$ is of $S_6$ symmetry. ................................................................. 38  
2.6 MP2 optimized Cl$^-$ (H$_2$O) structure. .............................................................. 40  
2.7 Mean ionic activity coefficient of NaCl solution at room temperature from 0 to 6 mol/kg. ................................................................. 42  
2.8 Mean ionic activity coefficient of KCl solution at room temperature from 0 to 3 mol/kg. ................................................................. 43  
2.9 Mean ionic activity coefficient of CsI solution at room temperature from 0 to 3 mol/kg. ................................................................. 43  
2.10 Experimentally fit $\ln \gamma_\pm$ and $\ln \gamma_\pm$ predicted by Debye-Hückel theory. ................................................................. 45  
2.11 Radial distribution functions of 3 mol/kg NaCl, KCl, and CsI for cations and anions. ................................................................. 46  
2.12 Radial distribution functions of 3 mol/kg NaCl, KCl, and CsI for cations and oxygens in water molecules. ................................................................. 47
4.16 CB[8]-G13 (Oxaliplatin) in CB[8]. .................................................. 103
4.17 Thermodynamic cycle for absolute binding free energy calculation. ....... 104
4.18 Distribution of the distance of the centroid of R-Bornylamine-H+ and the cen-
troid of CB[8] in the unrestrained molecular dynamics simulation. ............... 109
4.19 Initial submission of the binding free energies (in kcal/mol) of CB[8] and the
guests in the SAMPL6 challenge. .................................................. 112
4.20 Molecular model of cucurbit[8]uril with circular shape. ....................... 113
4.21 Molecular model of cucurbit[8]uril with a single indentation. ................ 114
4.22 Molecular model of cucurbit[8]uril with two indentations. .................... 114
4.23 Torsional potential of N-(methylene C)-N-(carbonyl C) in CB[8], with the 3-
fold torsional parameter being -0.25 or -1.60. The 1-fold and 2-fold torsional
parameters are 0.4 and -0.4, respectively. ....................................... 115
4.24 Molecular model of cucurbit[8]uril structure E1. ............................... 118
4.25 Molecular model of cucurbit[8]uril structure E2. ............................... 118
4.26 Molecular model of cucurbit[8]uril structure E3. ............................... 119
4.27 Two 0-0 states of Palonosetron-H+ sampled in the simulation. Left: from the
solvation calculation. Right: from the host-guest calculation. ...................... 120
4.28 Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H+
in a water box simulated from the standard 0-0 state simulations of 100 ps. Intra-
molecular vdw interactions were not annihilated. (a) starting from the sol-
vation 0-0 state; (b) starting from the solvation 0-0 state with the key torsion
being scaled to 0; (c) starting from the host-guest 0-0 state bound in CB[8];
(d) starting from the host-guest 0-0 state bound in CB[8] with the key torsion
being scaled to 0. .......................................................... 121
4.29 Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H+
in a water box simulated from the standard 0-0 state simulations of 100 ps. All four
started from the same solvation 0-0 state. (a) no special treatment; (b) key
torsion was scaled to 0; (c) intra-molecular vdw interactions were annihilated;
(d) both treatments in (b) and (c) were applied. .................................. 122
4.30 Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H+
in a water box simulated from the standard 0-0 state simulations of 100 ps. All four
started from the same host-guest 0-0 state bound in CB[8]. (a) no special
treatment; (b) key torsion was scaled to 0; (c) intra-molecular vdw interactions
were annihilated; (d) both treatments in (b) and (c) were applied. ............... 123
### 4.31 Revised binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

125
# List of Tables

2.1 Revised vdw parameters, $r^0$ (in Å), $\epsilon$ (in kcal/mol), and the simulated single ion solvation free energy (in kcal/mol). ........................................ 32

2.2 Lattice Energies (in kcal/mol) of the optimized salt crystal structures with revised vdw parameters in the AMOEBA force field, and lattice energies (in kcal/mol) measured in experiment. ................................. 33

2.3 Vdw parameters fit from salt crystals for cation-anion vdw interactions. ....... 35

2.4 Pairwise vdw parameters, from applying cubic combining rule to $r^0$ and HHG combining rule to $\epsilon$ of crystal fitting. ................................. 36

2.5 Symmetry (in point group), Na$^+$-O distance ($R$, in Å), basis set superposition error (BSSE) corrected electronic binding energy ($D_e$, in kcal/mol) of QM optimized Na$^+$($\text{H}_2\text{O}$)$_n$ clusters, and $R$, $D_e$ predicted by AMOEBA force field with revised Na$^+$ parameters. ................................. 37

2.6 Symmetry (in point group), ion-oxygen distance ($R$ in Å), and basis set superposition error (BSSE) corrected electronic binding energy ($D_e$, in kcal/mol) of MP2 optimized ion-water dimers, and $R$, $D_e$ predicted by AMOEBA force field with revised vdw parameters. ................................. 40

2.7 Periodic boundary condition (PBC) corrected self-diffusion coefficients of water. 49

2.8 Lattice Energies (in kcal/mol) of the optimized salt crystal structures containing Li$^+$ and F$^-$ with revised vdw parameters in the AMOEBA force field, and lattice energies (in kcal/mol) measured in experiment. ......................... 53

2.9 Experimental lattice energies (in kcal/mol) and lattice constants (in Å). ......... 53

2.10 Differences in the force field fit lattice energies (in kcal/mol) and lattice constants (in Å) with pairwise ion parameters. ................................. 54

2.11 Details of mean ionic activity coefficient calculations of NaCl. .................... 54

2.12 Details of mean ionic activity coefficient calculations of KCl. .................... 54

2.13 Details of mean ionic activity coefficient calculations of CsI. .................... 55

3.1 Results of regular simulation and isokinetic integrator simulations with various conditions for liquid argon. The short and long-range cutoffs for vdw are in Å. The kinetic energies (KE) are given in Kelvin. The diffusion coefficients $D$ are given in $10^{-5}$ cm$^2$/s. ........................................ 81
3.2 Simulation results of the “global” isokinetic integrators with the RESPA2 scheme. The short and long-range time steps for the nonbonded interactions are in fs. Kinetic energies (KE) are in Kelvin. Diffusion coefficients ($D$) are in $10^{-5}$ cm²/s.

3.3 Solvation free energies (in kcal/mol) of three molecules calculated by regular simulations and the “atom” isokinetic integrator.

3.4 Solvation free energies (in kcal/mol) of acridine orange calculated by regular 2 fs simulations and various “atom” isokinetic integrators from 1.5 ns simulations. The outer time steps are in fs. All the intermediate time steps for short-range nonbonded interactions are 2 fs.

4.1 Initial submission of the binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.

4.2 Single point energies (in hartree) and the energy differences (in kcal/mol) of CB[7] and CB[8] and their indented structures from ωB97X-D/6-311G(1d,1p) level of theory. CB[$n$]-0 denotes the circular structure, CB[$n$]-1 and CB[$n$]-2 denote the single and the double indentations in the structure, respectively.


4.4 Energy difference (in kcal/mol) between the elliptical CB[8] and the circular CB[8] from ωB97X-D/6-311G(1d,1p) level of theory, and different 3-fold torsional parameters in AMOEBA force field. The column “R” denotes the length (in Å) of the longest axis of the CB[8] molecule.

4.5 Revised binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.
LIST OF PROGRAMS

1.1 Euler-Cromer Integrator .................................................. 23
1.2 Velocity and Position Verlet Integrators ............................. 24
1.3 Velocity RESPA Integrator .................................................. 25
2.1 AMOEBA Parameters for Monovalent Ions ........................... 51
3.1 Stochastic Isokinetic Integrator ............................................ 92
4.1 Example of the trans-1,4-Diaminocyclohexane-2H+ Coordinates in Psi4 Format .................................................. 126
4.2 Example of the MP2/6-311G(1d,1p) Optimization Input in Psi4 Format .................................................. 127
4.3 Example of the GDMA Input in Psi4 Format ......................... 127
4.4 Example of the MP2/aug-cc-pVTZ Input in Psi4 Format ........ 127
4.5 Using PyMOL to Remove the Hydrogens in a PDB File .......... 128
4.6 AMOEBA Cucur[8]bituril Host Parameters ........................... 129
4.7 Example of the Tinker Key File for R-Bornylamine-H+ Host-Guest Simulations .................................................. 132
# List of Symbols and Abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>number of atoms in the system</td>
</tr>
<tr>
<td>( V )</td>
<td>volume of the system</td>
</tr>
<tr>
<td>( T )</td>
<td>temperature</td>
</tr>
<tr>
<td>( p )</td>
<td>pressure</td>
</tr>
<tr>
<td>( k_B )</td>
<td>Boltzmann constant; ( 1.3806485 \times 10^{-23} ) J/K</td>
</tr>
<tr>
<td>( R )</td>
<td>gas constant; ( 0.001987 ) kcal/(mol·K)</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( (k_B T)^{-1} )</td>
</tr>
<tr>
<td>( t )</td>
<td>time</td>
</tr>
<tr>
<td>( K )</td>
<td>Kelvin</td>
</tr>
<tr>
<td>( \text{atm} )</td>
<td>atmosphere; ( 1.01325 ) bar</td>
</tr>
<tr>
<td>( E_h )</td>
<td>hartree; ( 627.5095 ) kcal/mol</td>
</tr>
<tr>
<td>( \text{debye} )</td>
<td>( 1/299792458 \times 10^{-21} ) C·m</td>
</tr>
<tr>
<td>( r )</td>
<td>vector starting from point ( i ) and ending in point ( j )</td>
</tr>
<tr>
<td>( r_\alpha, r_\beta, r_\gamma \text{ etc.} )</td>
<td>orthogonal projections of ( r ) in ( x, y, ) or ( z ) direction</td>
</tr>
<tr>
<td>( r ) or ( r_{ij} )</td>
<td>norm of vector ( r ); the distance between points ( i ) and ( j )</td>
</tr>
<tr>
<td>( M^t )</td>
<td>transpose of matrix ( M )</td>
</tr>
<tr>
<td>( \text{tr}(M) )</td>
<td>trace of matrix ( M )</td>
</tr>
<tr>
<td>( \text{CB}[n] )</td>
<td>cucurbit( [n] )uril</td>
</tr>
<tr>
<td>( \text{MC} )</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>( \text{MD} )</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>( \text{MM} )</td>
<td>molecular mechanics</td>
</tr>
<tr>
<td>( \text{MTS} )</td>
<td>multiple time step</td>
</tr>
<tr>
<td>( \text{vdw} )</td>
<td>van der Waals</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

First and foremost, I sincerely thank Dr. Jay Ponder. No words can express my gratitude for his patience and guidance. His taste in art has created a tremendous office environment, and his taste in science is invaluable to his academic descendants. I am always thankful for being a member of his lab.

I would also thank other members of the dissertation examination committee: Prof. Greg Bowman, Prof. Richard Mabbs, Prof. Garland Marshall, and Prof. Lee Sobotka.

I appreciate the help from Chao Lu, Marie Laury, Josh Rackers, Aaron Gordon and Chris Ho. The conversations with them are always pleasant and sometimes thought-provoking. I will not forget the lunches to which Jay treated us every week in the Delmar Loop and the cupcakes Marie made for everyone’s birthday.

I am particularly grateful for the contributions that the OpenMM community has selflessly made, the great products that NVIDIA produced, and Jay’s insight and foresight that need no further proof.

And finally, I am deeply indebted to my family and my friends. My parents have always been supportive of my decisions even if they do not always understand them. The fact that they are rail workers for which college was not a requirement in their day by no means makes them belittle my education. Most of the closest friends I made in high school, in college, and in the U.S., oddly enough, have obtained or will obtain a doctoral degree. Therefore, I feel very fortunate that my friends and I are doing the same thing so that they always understand my struggle and my pursuit.

Zhi Wang
To my parents
Wang Xian Tao and Li Shang Mei

王先涛与李尚梅
The parameters of monovalent ions for the AMOEBA force field were revised. High level quantum mechanics results, relative solvation free energies of monovalent ions, lattice energies and lattice constants of salt crystals were used to calibrate the parameters. The revised parameters were validated against the quantum optimized structures and energies of ion-water dimers and ion-water clusters, and against thermodynamic properties of salt solutions at different concentrations measured in experiments, e.g. mean ionic activity coefficients, self-diffusion coefficients of water. In the simulations the sodium ion is found to qualitatively differ from larger cations in aqueous solution. Direct ionic interactions are predominant for potassium and larger cations, while sodium salt solutions at similar concentrations are dominated by ion-water interactions.

A novel stochastic isokinetic integrator proposed by Tuckerman, et al. was extended and generalized in three respects. First, the Nosé-Hoover chain algorithm was implemented in the original integrator. Next, the functional form of the isokinetic constraint was generalized so that it was no longer restricted to multiples of $k_B T$. Finally, the isokinetic constraint was extended to be able to constrain the kinetic energies of multi-dimensional velocities, instead of only one degree of freedom as in its original form.

An application of conformational sampling with molecular dynamics method, predictions of the binding free energies of cucurbit[8]uril and ligands in the SAMPL6 challenge,
is presented. A great improvement in the prediction accuracy was made by more ac-
curate torsional parameters of cucurbit[8]uril and by revised protocols annihilating the
intra-molecular van der Waals and key torsions in the ligands.

Corresponding methods for all portions of this work have been implemented in the
Tinker software package, some of which are also available in the Tinker-OpenMM library.
CHAPTER 1

Introduction

The background related to methods, algorithms, and theories applied in this dissertation are briefly reviewed in chapter [1]. Three chapters are then aligned to discussion of three topics in molecular dynamics. First, chapter [2] covers the force field development of the monovalent ions for the AMOEBA force field, including strategies for development and the validation of the force field. In chapter [3], the theory behind molecular dynamics simulation is elaborated via the example of a generalized stochastic isokinetic integrator. Then, chapter [4] exhibits an application of molecular dynamics to binding free energy calculation. Last but not least, conclusions and future perspectives are presented in chapter [5].

In the classic book of Allen and Tildesley [1], Computer Simulation of Liquids, the authors review the history of computer simulation in chemistry. It dates back to 1953 when Metropolis et al. [2] performed the first Monte Carlo (MC) simulation with the hard sphere model at the Los Alamos National Laboratory. Later, systems with the Lennard-Jones potential were simulated via the MC method [3], which for the first time enabled researchers to compare experimental data against simulation results. The first molecular dynamics (MD) simulation of hard spheres was later accomplished by Alder and Wainwright [4, 5]. In 1964, Rahman successfully performed MD simulations on systems with a Lennard-Jones potential [6]. Harp and Berne [7] published the first diatomic molecular dynamics simulation, then were followed by Barker and Watts [8] (MC), and by Rahman and Stillinger [9] (MD) simulating water. The first protein MD simulation was reported by
McCammon, Gelin, and Karplus in 1977 [10]. Currently, simulation has reached a point where microsecond-to-millisecond time scales and millions of atoms are accessible [11, 12].

The interactions in the simulated systems can, in principle, be described by quantum mechanics. But because of the dramatic growth of CPU time cost for larger simulation systems, Newtonian dynamics with classical potentials is usually a better choice, especially if no chemical reaction is expected, in which case, the potential functions and the associated parameters should be predetermined. A force field is the sum of the functional forms to describe the intra- and inter-molecular interactions, as well as the parameters associated with the functional forms and with the individual species present in the simulation.

Most of force fields include at least three valence terms: bond, angle, torsion, and two non-bonded terms: electrostatics and van der Waals (vdw). Many force fields use a harmonic potential for bonds and angles, a trigonometric potential for the torsions, the Lennard-Jones potential for vdw, and Coulomb’s law with atomic partial charges for electrostatics, although other alternative potentials may be used instead, e.g. the Morse potential for bonds, and the Buckingham potential for vdw (e.g. OPLS [13], CHARMM27 [14], Amber ff99 [15], GAFF [16], etc.). Coupling terms across the valence terms are included in some force fields (MM4 [17] etc.). Other force fields (AMOEBA [18] etc.) go even further, and more accurate descriptions of inter-molecular interactions are included.

1.1 AMOEBA Force Field

AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) is a polarizable force field developed in the Ponder group. The valence terms of the AMOEBA force field consist of bond, angle, torsion, stretch-bend, out-of-plane, and the non-bonded
terms contain van der Waals, permanent multipole electrostatics and induced dipole polarization.

1.1.1 valence terms

The AMOEBA force field uses the anharmonic potentials in bond and angle terms:

\[ U_{bond} = K_b(b - b_0)^2 \left( 1 + K_{b3}(b - b_0) + K_{b4}(b - b_0)^2 \right), \]

\[ U_{angle} = K_\theta(\theta - \theta_0)^2 \left( 1 + \sum_{i=3}^{6} K_{\theta i}(\theta - \theta_0)^{i-2} \right), \]  

(1.1.1)

where \( K_b \)'s and \( K_\theta \)'s are constant parameters in the potentials, \( b_0 \) and \( \theta_0 \) are the ideal bond length and angle.

The torsion term in AMOEBA uses an up-to-6-fold trigonometric functional form

\[ U_{torsion} = \sum_{n=1}^{6} K_{\phi n} \left[ 1 + \cos(n\phi \pm \delta_n) \right], \]  

(1.1.2)

where \( K_{\phi n} \)'s are the force constants, \( \delta \)'s are the phase parameters, and \( \phi n \) is the periodicity.

The stretch-bend and out-of-plane terms are functions of bond length \( b \), angle \( \theta \), and out-of-plane angle \( \chi \) [19]:

\[ U_{b\theta} = K_{b\theta} \left( [b - b_0] + [b' - b'_0] \right) (\theta - \theta_0), \]

\[ U_{oop} = K_\chi \chi^2. \]  

(1.1.3)

The valence terms in the AMOEBA force field are usually restricted to intra-molecular interactions spanning less than or equal to three bonds (1-4 connectivity).

1.1.2 van der Waals interaction

The van der Waals (vdw) potential is an additive pairwise potential. Two independent parameters \( \epsilon_i \) and \( r_i^0 \) are assigned to atom \( i \). \( \epsilon_i \) is the well depth of the vdw curve, and \( r_i^0 \) is the distance where the vdw curve reaches its minimum. The pairwise vdw parameters
of atoms \(i\) and \(j\) can either be predetermined via a lookup table or be calculated by applying combining rules on each unique atom pair. Common combining rules include the arithmetic combining rule

\[
\epsilon_{ij} = (\epsilon_i + \epsilon_j) / 2,
\]

(1.1.4)

the geometric combining rule

\[
\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j},
\]

(1.1.5)

the cubic combining rule \[20\]

\[
\epsilon_{ij} = \frac{\epsilon_i^3 + \epsilon_j^3}{\epsilon_i^2 + \epsilon_j^2},
\]

(1.1.6)

the HHG combining rule \[20\]

\[
\epsilon_{ij} = \frac{4\epsilon_i \epsilon_j}{(\sqrt{\epsilon_i} + \sqrt{\epsilon_j})^2},
\]

(1.1.7)

etc.

The choice of the combining rules is somewhat arbitrary, but can have a large effect on computed energies. As a result, the limitations of the selected combining rules in the model must be made clear. As shown in fig. 1.1, the cubic combining rule is not monotonic and adopts a relatively narrow set of combined values in the range of \((0, 1)\). This feature may seem undesirable, but under certain circumstances makes the model more robust. If two values are not out of proportion, which is true for most of the cases, the combined value becomes monotonic and less sensitive to the random errors in the arguments. For the monotonic combining rules, the arithmetic rule is significantly different from the other two, in terms of the lower bounds of the combining functions in \((0, 1)\).

The AMOEBA force field uses the following buffered 14-7 functional form \[20\] to describe the vdw potential between atoms \(i\) and \(j\),

\[
U_{vdw} = \epsilon_{ij} \left( \frac{1 + \delta}{\rho_{ij} + \delta} \right)^{n-m} \left( \frac{1 + \gamma}{\rho_{ij}^m + \gamma} - 2 \right),
\]

(1.1.8)

where \(\rho_{ij} = r_{ij}/r_{ij}^0\), \(n = 14\), \(m = 7\). \(\delta\) and \(\gamma\) are buffering constants which are fit to quantum calculations on rare gases \[20\], and generally adopt the values of 0.12 and 0.07,
Figure 1.1: Combined value of $x$ and 1 with different combining rules.

respectively. With $n = 12$, $m = 6$, $\gamma = \delta = 0$, eq. (1.1.8) recovers the Lennard-Jones 12-6 potential. The buffered 14-7 potential is softer over the range of short distances, as shown in fig. 1.2. The hydrogen atoms in the AMOEBA vdw potential are shifted towards their connecting atoms by a “reduction factor”, an extension which is supported by X-Ray experiments [21, 22].

The Lennard-Jones 12-6 potential is the most widely used vdw potential. CHARMM and Amber use the arithmetic rule for $r^0$ and the geometric rule for $\epsilon$, whereas OPLS uses the geometric combining rule for both $r^0$ and $\epsilon$. The AMOEBA force field takes the buffered 14-7 functional form with the cubic combining rule for $r^0$ and HHG combining rule for $\epsilon$ by default, and preserves the capability to update specific pairwise vdw
parameters through a post-facto lookup table.

1.1.3 permanent multipole electrostatic energy

Two relations should be obtained prior to deriving the expression of the permanent multipole electrostatics [23].

First, suppose that in fig. 1.3 site A is at the origin and is surrounded by some charges $e_a$ of displacement $\alpha$. The electrostatic potential at site B due to the charges around site A is

$$\varphi_A(r) = \sum_a \frac{e_a}{|r - \alpha|}. \quad (1.1.9)$$
If all the displacements are short such that the Taylor series of $1/(r - a)$ converges around site A, the potential $\varphi_A(r)$ can then be expanded as

$$\varphi_A(r) = \sum_a e_a \left\{ \frac{1}{r} + \sum_\alpha a_\alpha \left( \frac{1}{r} \right)_\alpha + \frac{1}{2!} \sum_{\alpha\beta} a_\alpha a_\beta \left( \frac{1}{r} \right)_{\alpha\beta}'' + \frac{1}{3!} \sum_{\alpha\beta\gamma} a_\alpha a_\beta a_\gamma \left( \frac{1}{r} \right)_{\alpha\beta\gamma}''' + \ldots \right\}$$

$$\equiv M^{(0)}_a \frac{1}{r} + M^{(1)}_a \nabla \frac{1}{r} + M^{(2)}_a \nabla^2 \frac{1}{r} + M^{(3)}_a \nabla^3 \frac{1}{r} + \ldots,$$

(1.1.10)

which also gives the definition of multipole components of site A, e.g., $q_a = \sum_a e_a$, $\mu_{ax} = \sum_a e_a a_{xr}$, etc.

Then consider the situation where site A is placed in an external electrostatic field with potential $\varphi$. The electrostatic energy of site A is given by

$$U_a = \sum_a e_a \varphi(a)$$

(1.1.11)

$$= \sum_a e_a \left\{ \varphi(0) + \sum_\alpha a_\alpha \varphi'(0)_{\alpha} + \frac{1}{2!} \sum_{\alpha\beta} a_\alpha a_\beta \varphi''(0)_{\alpha\beta} \right\}$$

(1.1.12)
\[
\sum_{\alpha \beta \gamma} a_\alpha a_\beta a_\gamma \varphi'''(0)_{\alpha \beta \gamma} + \ldots \right) \tag{1.1.13}
\]

\[
\equiv M_a^{(0)} \varphi(0) + M_a^{(1)} \varphi'(0) + M_a^{(2)} \varphi''(0) + M_a^{(3)} \varphi'''(0) + \ldots, \tag{1.1.14}
\]

where \( \varphi(a) \) is estimated by \( \varphi(0) \), and \( M_a^{(i)} \) are the multipole components of site A.

The permanent multipole interaction in the AMOEBA force field includes through quadrupole interactions. The permanent multipole moment \( M_i \) of atom \( i \) consists of atom-centered charge \( q \), dipole \( \mu \), and quadrupole \( Q \) moments:

\[
M_i^T = [q_i, \mu_{ix}, \mu_{iy}, \mu_{iz}, Q_{i1x}, Q_{i1y}, Q_{i1z}, \ldots, Q_{i12}]. \tag{1.1.15}
\]

According to eqs. (1.1.10) and (1.1.11), the permanent multipole energy of atom \( i \) due to the electrostatic potential generated by atom \( j \) is

\[
U_{ele} = M_i^T T_{ij} M_j, \tag{1.1.16}
\]

\begin{align*}
T_{ij} &= \\
&= \frac{1}{r_{ij}^3} \begin{bmatrix}
1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \frac{\partial^2}{\partial x_j^2} & \ldots & \frac{\partial^2}{\partial z_j^2} \\
\frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \frac{\partial^3}{\partial x_i \partial x_j^2} & \ldots & \frac{\partial^3}{\partial x_i \partial z_j^2} \\
\frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \frac{\partial^3}{\partial y_i \partial x_j^2} & \ldots & \frac{\partial^3}{\partial y_i \partial z_j^2} \\
\frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \frac{\partial^3}{\partial z_i \partial x_j^2} & \ldots & \frac{\partial^3}{\partial z_i \partial z_j^2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2}{\partial z_i^2} & \frac{\partial^3}{\partial z_i^2 \partial x_j} & \frac{\partial^3}{\partial z_i^2 \partial y_j} & \frac{\partial^3}{\partial z_i^2 \partial z_j} & \frac{\partial^4}{\partial z_i^2 \partial x_j^2} & \ldots & \frac{\partial^4}{\partial z_i^2 \partial z_j^2}
\end{bmatrix}
\end{align*}

1.1.4 

**polarization energy**

Polarization energy covers a major portion of the total many-body interaction by including the perturbation of the electron density under the influence of an external field. Unlike permanent electrostatics where the distributions of multipoles are described by the Dirac delta function, for polarization energy the distributions adopt a spherically symmetric function \( \rho(u) \), where the length \( u \) is proportional to the distance between two mul-
tipole sites. It is known the field can be written as
\[ \mathbf{E}(\mathbf{u}) = \int \text{d}u_1 \frac{\rho(u_1)}{(u-u_1)^2} \frac{\mathbf{u}-\mathbf{u}_1}{|\mathbf{u}-\mathbf{u}_1|} \]  \hspace{1cm} (1.1.17)
and can be simplified to
\[ E_\alpha(u) = \frac{u_\alpha}{u^2} \int_0^u 4\pi u_1^2 \rho(u_1) \text{d}u_1. \]  \hspace{1cm} (1.1.18)
On the other hand, by definition, the electrostatic field is the negative gradient of the electrostatic potential
\[ E_\alpha(u) = -\frac{\partial \varphi(u)}{\partial u_\alpha} = -\frac{u_\alpha}{u} \frac{\partial \varphi(u)}{\partial u}, \]  \hspace{1cm} (1.1.19)
thus
\[ \frac{\partial \varphi(u)}{\partial u} = -\frac{1}{u^2} \int_0^u 4\pi u_1^2 \rho(u_1) \text{d}u_1, \]  \hspace{1cm} (1.1.20)
\[ \frac{\partial^2 \varphi(u)}{\partial u^2}_{\alpha\beta} = \left[ \frac{\partial^2 \varphi(u)}{\partial u^2} - \frac{1}{u} \frac{\partial \varphi(u)}{\partial u} \right] \frac{u_\alpha u_\beta}{u^2} + \frac{\partial \varphi(u)}{\partial u} \delta_{\alpha\beta}, \]  \hspace{1cm} (1.1.21)
and higher order derivatives of \( \varphi(u) \) can be obtained. By definition, \( \delta_{\alpha\beta} = 1 \) if and only if \( \alpha = \beta \), otherwise, \( \delta_{\alpha\beta} = 0 \).

The polarization energy in AMOEBA adopts the classic induced dipole model and Thole’s damping scheme \cite{24} using the distribution function
\[ \rho(u) = \frac{3a}{4\pi} e^{-au^3}, \]  \hspace{1cm} (1.1.22)
where \( a \) is the damping parameter, \( u = r_{ij}/(\alpha_i\alpha_j)^{1/6} \), \( \alpha_i \) and \( \alpha_j \) are the atomic polarizabilities of atoms \( i \) and \( j \), respectively. With the multipoles implicitly included in eq. (1.1.22), it is easy to verify
\[ \frac{\partial \varphi(u)}{\partial u} = -\frac{1}{u^2} \left( 1 - e^{-au^3} \right), \]  \hspace{1cm} (1.1.23)
\[ \frac{\partial^2 \varphi(u)}{\partial u^2} = \frac{1}{u^3} \left[ 2 - (2 + 3au^3)e^{-au^3} \right], \]  \hspace{1cm} (1.1.24)
\[ \frac{\partial^3 \varphi(u)}{\partial u^3} = \frac{3}{u^4} \left[ -2 + (2 + 2au^3 + 3a^2u^6)e^{-au^3} \right], \]  \hspace{1cm} (1.1.25)
\[
\frac{\partial^4 \varphi(u)}{\partial u^4} = \frac{3}{u^5} \left[ 8 - (8 + 8au^3 + 9a^3u^9)e^{-au^3} \right].
\] (1.1.26)

More specifically, comparing
\[
\frac{\partial \varphi(u)}{\partial u} \frac{u_a}{u} = \frac{\partial \varphi(u)}{\partial u} \frac{u_a}{u}
\]
\[
= - \left( 1 - e^{-au^3} \right) \frac{u_a}{u^3},
\] (1.1.28)

with the element of the undamped \(T\) matrix,
\[
t_\alpha = -\frac{r_\alpha}{r^3},
\] (1.1.29)

gives the element \(t^d\) of the damped \(T\) matrix, \(T^d\),
\[
t^d_\alpha = -\lambda_3 \frac{r_\alpha}{r^3},
\] (1.1.30)

where
\[
\lambda_3 = 1 - e^{-au^3}.
\] (1.1.31)

Similarly, comparing \(\frac{\partial^2 \varphi(u)}{\partial u^2} \frac{t_{\alpha \beta}}{t_{\alpha \beta}} \), etc., gives
\[
\frac{\partial^2 \varphi(u)}{\partial u^2} \frac{t_{\alpha \beta}}{t_{\alpha \beta}} = \left[ \frac{\partial^2 \varphi(u)}{\partial u^2} - \frac{1}{u} \frac{\partial \varphi(u)}{\partial u} \right] \frac{u_\alpha u_\beta}{u^2} + \frac{\partial \varphi(u)}{\partial u} \frac{\delta_{\alpha \beta}}{u}
\]
\[
= \left[ 3 - (3 + 3au^3)e^{-au^3} \right] \frac{u_\alpha u_\beta}{u^5} - \lambda_3 \frac{\delta_{\alpha \beta}}{u^3},
\] (1.1.32)

\[
t_{\alpha \beta} = \frac{3r_\alpha r_\beta}{r^5} - \frac{\delta_{\alpha \beta}}{r^3},
\] (1.1.33)

and
\[
t^d_{\alpha \beta} = \lambda_5 \frac{3r_\alpha r_\beta}{r^5} - \lambda_3 \frac{\delta_{\alpha \beta}}{r^3},
\] (1.1.34)

where,
\[
\lambda_5 = 1 - (1 + au^3)e^{-au^3},
\] (1.1.35)

together with
\[
t^d_{\alpha \beta \gamma} = -\lambda_7 \frac{15}{r^7} r_\alpha r_\beta r_\gamma + \lambda_5 \frac{3}{r^5} \sum_{\gamma} r_\alpha r_\beta \gamma,
\] (1.1.36)
\[ i_{\alpha\beta\gamma\eta}^{d} = \lambda_{9} \frac{105}{r^{9}} r_{\alpha} r_{\beta} r_{\gamma} r_{\eta} - \lambda_{7} \frac{15}{r^{7}} \sum_{\gamma} r_{\alpha} r_{\beta} \delta_{\gamma \eta} + \lambda_{5} \frac{3}{r^{5}} \sum_{\gamma} \delta_{\alpha \beta} \delta_{\gamma \eta}, \]  

(1.1.37)

where

\[ \lambda_{7} = 1 - \left( 1 + au^{3} + \frac{3}{5} a^{2} u^{6} \right) e^{-au^{3}}, \]  

(1.1.38)

\[ \lambda_{9} = 1 - \left( 1 + au^{3} + \frac{18 a^{2} u^{6} + 9 a^{3} u^{9}}{35} \right) e^{-au^{3}}. \]  

(1.1.39)

If the induced dipoles \( \mu_{ind} \) are already determined, the polarization energy can be calculated via

\[ U_{pol} = -\mu_{ind}^{T} T^{d} \mu_{ind} / 2. \]  

(1.1.40)

The induced dipoles are solved using the following well-known relations. The induced dipoles and the total electrostatic field \( E_{tot} \) are related by

\[ \mu_{ind} = \alpha E_{tot}, \]  

(1.1.41)

where \( E_{tot} \) consists of two contributions, the electrostatic field of the permanent multipoles \( (E_{perm}) \), and the electrostatic field of the induced dipole \( (E_{ind}) \). Because the induced field is given by \( E_{ind} = T^{d} \mu_{ind} \), the solution of the following equation

\[ (\alpha^{-1} - T^{d}) \mu_{ind} = E_{perm} \]  

(1.1.42)

produces the complete induced dipole moments.

### 1.2 Common Methodologies

Three commonly used simulation methods are introduced in this section. The infrastructure to perform accurate numerical integration for Newton’s second law of motion will be discussed first. Then two types of algorithms will be presented, which make dynamics simulation possible beyond the microcanonical ensemble by restricting the system to constant temperature and/or constant pressure.
Besides the methods applied during the dynamics simulation, two typical post analyses for the simulation trajectories are also briefly introduced. One is calculation of the radial distribution function, which is a representation of the structural ensemble average. The other is determination of the diffusion coefficient. Since the diffusion coefficient is a dynamic equilibrium property and is equivalent to a time correlation function, the time step between two consecutive frames in the saved MD trajectory must be known, which indicates such properties are inaccessible via Monte Carlo simulation.

At the end of the section, algorithms for free energy calculation, representative of thermodynamic properties, are discussed.

1.2.1 integrators

Consider an isolated system of \( N \) classical particles described by their momenta \( p(t) \) and positions \( q(t) \), where \( p(t) = p_1(t), \ldots, p_N(t) \) and \( q(t) = q_1(t), \ldots, q_N(t) \). The time evolution of the system is governed by Newton’s second law of motion. If no external force or frictional force is present,

\[
\frac{d}{dt} p(t) = -\frac{\partial}{\partial q} U(q(t)).
\]  

(1.2.1)

\( U \) is the potential energy and is only a function of positions. A straightforward algorithm (program 1.1) solving for eq. (1.2.1) is the Euler-Cromer integrator [26].

Two problems need to be addressed. First, although not directly shown in eq. (1.2.1), the time step used in the integration has to be short to avoid large numerical errors. Second, the positions and momenta integrated in program 1.1 are asynchronous. The difference cannot be eliminated, but it can be reduced. The Liouville operator formalism is introduced here as a powerful and elegant tool for the analysis [27,29].

Let \( x = (p, q) \) be the representation of the state of the system. Its time derivative can be written as

\[
\frac{dx}{dt} = \left\{ \frac{dp}{dt} \frac{\partial}{\partial p} + \frac{dq}{dt} \frac{\partial}{\partial q} \right\} x,
\]  

(1.2.2)
therefore, an operator $iL$ can be defined as $iL = \frac{dp}{dt} \partial/\partial p + \frac{dq}{dt} \partial/\partial q$ such that $\frac{d}{dt} \mathbf{x} = iL \mathbf{x}$, yielding the formal expression

$$\mathbf{x}_t = \exp(iLt) \mathbf{x}_0,$$

(1.2.3)

where $iL$ is the Liouville operator, and $\exp(iLt)$ is the classical propagator. To solve the first problem, a long time step should be split into shorter time steps, since

$$\mathbf{x}_t = \lim_{n \to \infty} \left[ \exp \left( i \frac{t}{n} L \right) \right]^n \mathbf{x}_0.$$

(1.2.4)

For the second problem, as the Liouville operator can naturally be written in two parts, $iL_p$ and $iL_q$, the classical propagator can be approximated [30] by

$$\exp[(iL_1 + iL_2) t] \sim \exp(iL_1 t/2) \exp(iL_2 t) \exp(iL_1 t/2),$$

(1.2.5)

which yields two variations of the Verlet integrator [31] (program 1.2): the velocity Verlet integrator and the position Verlet integrator.

In this example of the two alternative Verlet integrators, it is clear that different factorization schemes will result in different differential equations and different integrators. Another widely used application is the multiple time step RESPA (reversible reference system propagator algorithms) integrator [32], shown in program 1.3. The classical propagator is factorized in such a way that the low frequency forces (denoted by $s$, slow) are evaluated less frequently than the high frequency forces (denoted by $f$, fast):

$$\exp(iLt) = \exp \left( (n_s iL_s + iL_f) \frac{t}{n_s} \right) \prod_{j=1}^{n_s-1} \exp \left( iL_f \frac{t}{n_s} \right),$$

(1.2.6)

where the fast forces are $n_s$ times more frequently calculated in the simulation than the slow forces. And because the computational expense for the fast forces is typically much less than for the slow forces, the multiple time step RESPA integrator significantly reduces the amount of CPU time required.

More examples and applications will be discussed in chapter 3.
1.2.2 thermostats and barostats

Thermostats and barostats are introduced into dynamics simulations for various reasons, e.g. the accumulation of numerical error in NVE simulations will eventually violate the conservation of energy; restricted by the size of the system in the simulation, the difference among NVE, NVT, and NpT ensembles cannot be neglected, etc. Therefore, the system in NVT or NpT simulations should be properly maintained at constant temperature and constant pressure.

Thermostat algorithms often involve modifications in the equations of motion. The Andersen thermostat [33] directly assigns new velocities to randomly selected atoms during the simulation. The Bussi thermostat [34, 35] scales all the velocities in the system to ensure the kinetic energy \( K \) of the samples satisfy the canonical equilibrium distribution

\[
P(K) \propto K^{N_f/2-1} e^{-\beta K},
\]

where \( N_f \) is the number of degrees of freedom, and treats the evolution of \( K \) as an auxiliary dynamics controlled by a stochastic process

\[
dK = \left( D_K \frac{\partial \ln P}{\partial K} + \frac{\partial D_K}{\partial K} \right) dt + \sqrt{2D_K} dW.
\]

The stochastic Langevin thermostat adds friction \( \gamma \) and random force \( R \) to the equations of motion via

\[
\frac{d}{dt} p = -\frac{\partial}{\partial q} U - \gamma p + R.
\]

Other integrators (e.g. Nosé–Hoover [36–38], etc.) will couple the internal system with an external heat bath via modified equations of motion.

Analogous to the thermostat case, a barostat changes the volume of the periodic boundary box in the simulation. The Berendsen barostat [39] computes the pressure from the virial of the system and re-scales the volume according to the partition function of the isobaric-isothermal ensemble. The Parrinello-Rahman barostat [40, 41] works with
the Nosé–Hoover thermostat and treats the volume as an auxiliary dynamical variable. The Monte Carlo barostat [42] is different from the barostats mentioned above in that it does not require computation of the virial. Given the fact that the partition function of isobaric-isothermal ensemble is

\[
\Delta(N, P, T) = \frac{1}{V_0 h^{3N} N!} \int_0^\infty dV \int dp \int dq e^{-\beta[H(p, q)+PV]},
\]

where the coordinates can be scaled by box length $1/L$, as $s$, eq. (1.2.10) then becomes

\[
\Delta(N, P, T) = \frac{1}{V_0 \Lambda^{3N} N!} \int_0^\infty dV V^N e^{-\beta PV} \int ds e^{-\beta U(s, V)},
\]

and the distribution function of $V$ is determined

\[
f(V) \propto e^{-\beta[U(s, V)+PV-N\beta^{-1} \ln V]}.
\]

The acceptance ratio for Monte Carlo trial moves that change the volume of the system from $V$ to $V'$ is \(\min\{1, f(V')/f(V)\}\).

1.2.3 radial distribution function (RDF)

The radial distribution function (RDF, $g(r)$, or $g(r)$, illustrated in fig. 1.4) is a quantitative measurement of structural heterogeneity, in terms of the ratio of the real distribution density of the pairwise displacement $r$ (or $r$) to the uniform distribution. For a simple system with $N$ particles of the same kind, the real distributions $\rho(r)$ and $\rho(r)$ can be formally represented by

\[
\rho(r) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \delta(r - (r_j - r_i)),
\]

and

\[
\rho(r) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \delta(r - |r_j - r_i|),
\]

15
where $\delta(x) = 1$, if $x = 0$, otherwise $\delta(x) = 0$. Then the RDF can be defined by

$$g(r) = \frac{\rho(r)}{\rho_0},$$

(1.2.15)

and

$$g(r) = \frac{\rho(r)}{4\pi r^2 \rho_0},$$

(1.2.16)

where $\rho_0$ is the normal distribution. Following these definitions, the expected number of pairs of particles whose distances are less than $r'$ is

$$n(r < r') = 4\pi \rho_0 \int_0^{r'} g(r) r^2 \, dr.$$  

(1.2.17)

Figure 1.4: Schematic sketch of the RDF of liquid argon.
The diffusion coefficient (or diffusivity) $D$ is a parameter appearing in the transport equation

$$\frac{\partial}{\partial t} \rho(r, t) = D \nabla^2 \rho(r, t)$$  \hspace{1cm} (1.2.18)

that characterizes the evolution of the number density $\rho$. The distribution of particles at time 0 is Dirac delta function $\delta(r)$, which also works as the boundary condition of eq. (1.2.18). Given the solution

$$\rho(r, t) = \frac{1}{(4\pi Dt)^{d/2}} \exp \left( -\frac{r^2}{4Dt} \right),$$  \hspace{1cm} (1.2.19)

where $d$ is the dimension and is commonly equal to 3 in unconstrained simulations, the mean squared displacement (MSD) of the selected atom and the evolution of MSD are:

$$\langle r^2(t) \rangle = \int \rho(r, t) r^2 dr,$$  \hspace{1cm} (1.2.20)

$$\frac{\partial}{\partial t} \int \rho(r, t) r^2 dr = D \int r^2 \nabla^2 \rho(r, t) \, dr.$$  \hspace{1cm} (1.2.21)

Integrating the RHS by parts,

$$D \int dr \, r^2 \nabla \nabla \rho = D \int dr \, \nabla \left( r^2 \nabla \rho \right) - D \int dr \, \nabla \left( \nabla r^2 \right) \cdot \left( \nabla \rho \right)$$  \hspace{1cm} (1.2.22)

$$= D \int dS \cdot \left( r^2 \nabla \rho \right) - 2D \int dr \, r \cdot \left( \nabla \rho \right)$$  \hspace{1cm} (1.2.23)

$$= D \int dS \cdot \left( r^2 \nabla \rho \right) - 2D \int dr \, \nabla \left( r \rho \right) + 2D \int dr \, \nabla \left( \nabla r \right) \rho$$  \hspace{1cm} (1.2.24)

$$= D \int dS \cdot \left( r^2 \nabla \rho \right) - 2D \int dS \cdot \left( r \rho \right) + 2D \int dr \, \nabla \left( \nabla r \right) \rho$$  \hspace{1cm} (1.2.25)

$$= 2dD,$$  \hspace{1cm} (1.2.26)

since the first two integrals of eq. (1.2.25) vanish at infinity. Letting $d = 3$ yields the relation which was first derived by Einstein [43]:

$$\frac{\partial}{\partial t} \langle r^2(t) \rangle = 6D.$$  \hspace{1cm} (1.2.27)
Alternatively, the MSD can be evaluated by the velocity autocorrelation function,

\[
\langle r^2(t) \rangle = \left\langle \left( \int_0^t dt_1 v(t_1) \right)^2 \right\rangle \\
= \int_0^t \int_0^t dt_1 dt_2 \langle v(t_1) \cdot v(t_2) \rangle \\
= 2 \int_0^t dt_1 \int_0^{t_1} dt_2 \langle v(t_1) \cdot v(t_2) \rangle.
\]

(1.2.28)

Because this integral should only depend on the difference between \(t_1\) and \(t_2\), it can be written as

\[
\langle r^2(t) \rangle = 2 \int_0^t dt_1 \int_0^{t_1} dt_2 \langle v(t_1 - t_2) \cdot v(0) \rangle,
\]

(1.2.29)

and eq. (1.2.21) becomes

\[
\lim_{t \to \infty} \frac{\partial}{\partial t} \langle r^2(t) \rangle = 6D = 2 \lim_{t_1 \to \infty} \int_0^{t_1} dt_2 \langle v(t_1 - t_2) \cdot v(0) \rangle.
\]

(1.2.30)

Therefore,

\[
D = \frac{1}{3} \int_0^\infty d\tau \langle v(\tau) \cdot v(0) \rangle.
\]

(1.2.31)

### 1.2.5 free energy calculation

The following discussion is restricted to the calculations of the Helmholtz free energy \(A\) from the canonical ensemble. Extending these results to other ensembles should be straightforward and is not introduced here. The Helmholtz free energy can be written as

\[
A = -k_B T \ln Q,
\]

where \(Q\) is the canonical partition function of the system given by,

\[
Q(N, V, T) = \frac{1}{\hbar^{3N} N!} \int d\mathbf{p} d\mathbf{q} e^{-\beta[K(\mathbf{p}) + U(\mathbf{q})]},
\]

(1.2.32)

and \(Q\) can be written in terms of the configurational integral \(Z\),

\[
Q(N, V, T) = \frac{1}{A^{3N} N!} Z(N, V, T),
\]

(1.2.33)

\[
Z(N, V, T) = \int d\mathbf{q} e^{-\beta U(\mathbf{q})},
\]

(1.2.34)
where $\Lambda$ is the thermal de Broglie wavelength and is equal to $h/\sqrt{2\pi m k_B T}$.

The fact that the configurational integral $Z$ cannot be calculated explicitly makes the direct calculation of the Helmholtz free energy impossible. The free energy difference $\Delta A$ between two states, 0 and 1, however, can be determined. By definition,

$$\Delta A = A_1 - A_0 = -k_B T \ln \frac{Z_1}{Z_0} = -k_B T \ln \frac{Z_1/\exp(-\beta A_0)}{Z_0/\exp(-\beta A_0)}.$$  \hspace{1cm} (1.2.35)

Expanding the RHS of eq. (1.2.35) gives

$$\Delta A = -k_B T \ln \frac{Z_1/\exp(-\beta A_0)}{Z_0/\exp(-\beta A_0)} = -k_B T \ln \frac{\int e^{-\beta(U_1-U_0)} e^{-\beta(U_0-A_0)} dq}{\int e^{-\beta(U_0-A_0)} dq}$$

$$= -k_B T \ln \langle e^{-\beta(U_1-U_0)} \rangle_0.$$  \hspace{1cm} (1.2.36)

Equation (1.2.36) is known as the Zwanzig equation [44], and the straightforward method that uses eq. (1.2.36) to calculate the free energy difference between states 0 and 1 is called free energy perturbation (FEP).

The ratio of two configurational integral can be written in a general form

$$\frac{Z_0}{Z_1} = \frac{\int W e^{-\beta U_0 - \beta U_1} dq/Z_1}{\int W e^{-\beta U_1 - \beta U_0} dq/Z_0} = \frac{\langle W e^{-\beta U_0} \rangle_1}{\langle W e^{-\beta U_1} \rangle_0} = \frac{Y_0}{Y_1},$$  \hspace{1cm} (1.2.37)

where $W = W(q)$ can be any finite function. E.g., let $W = \exp(-\beta \min\{U_0, U_1\})$, eq. (1.2.37) will be reduced to eq. (1.2.36). The following discussion will show that one specific functional form of $W$ will minimize the statistical error $(\Delta A_{est} - \Delta A)^2$.

Expanding the statistical error in terms of the squared standard errors of the mean (SEM) of $Y_0$ and $Y_1$ gives

$$(\Delta A_{est} - \Delta A)^2 = \text{SEM}^2(\Delta A)$$

$$= (k_B T)^2 \text{SEM}^2 \left( \ln \frac{Y_0}{Y_1} \right)$$

$$= (k_B T)^2 \left[ \frac{\text{SEM}^2(Y_0)}{Y_0^2} + \frac{\text{SEM}^2(Y_1)}{Y_1^2} \right].$$  \hspace{1cm} (1.2.38)
Supposing that \( n_0 \) and \( n_1 \) independent samples are collected for state 0 and 1, respectively, eq. (1.2.38) can be approximated by the sample standard deviation

\[
\frac{(\Delta A_{est} - \Delta A)^2}{(k_B T)^2} = \frac{1}{n_1^2} \left[ \frac{1}{n_0} \left( \langle W^2 e^{-2\beta U_1} \rangle_0 - Y_1^2 \right) \right] + \frac{1}{n_0^2} \left[ \frac{1}{n_1} \left( \langle W^2 e^{-2\beta U_0} \rangle_1 - Y_0^2 \right) \right]
\]

\[
= \frac{1}{n_0} \left[ \int W^2 e^{-2\beta U_0} d\mathbf{q} \right]^2 / Z_0^2 + \frac{1}{n_1} \left[ \int W^2 e^{-2\beta U_0 - \beta U_1} d\mathbf{q} / Z_1 \right]^2 - \frac{n_0 - 1}{n_0} - \frac{n_1 - 1}{n_1}.
\]

(1.2.39)

The first term on the RHS of eq. (1.2.39) is rewritten in the compact form \( f/g \), where

\[
f = \int RW^2 dx,
\]

(1.2.40)

\[
g = \left[ \int W dx \right]^2,
\]

(1.2.41)

then the problem of minimizing eq. (1.2.39) becomes finding the solution of equation

\[
f'g - fg' = 0.
\]

(1.2.42)

It is easy to verify that

\[
f' = 2 \int RW dx,
\]

(1.2.43)

\[
g' = 2 \int W dx \int dx.
\]

(1.2.44)

Now eq. (1.2.42) is transformed to

\[
\int RW dx \int W dx - \int RW^2 dx \int dx = 0.
\]

(1.2.45)

Two solutions of eq. (1.2.45) can be found: \((RW = RW^2, W = 1)\) and \((RW = 1, W = RW^2)\), but only the latter solution is non-trivial, thus the optimum \( W \) function is

\[
W = R^{-1} = \text{const} \times \left( e^{-\beta U_1} Z_0/n_0 + e^{-\beta U_0} Z_1/n_1 \right)^{-1}.
\]

(1.2.46)
Inserting eq. (1.2.46) in eq. (1.2.37), \( Z_0/Z_1 \) will become

\[
\frac{Z_0}{Z_1} = \frac{\langle f(U_0 - U_1 + C) \rangle_1}{\langle f(U_1 - U_0 - C) \rangle_0} \exp(\beta C),
\]

(1.2.47)

where \( f(x) = 1/(1 + \exp(\beta x)) \) is the Fermi-Dirac function, and

\[
C = k_B T \ln[(Z_0 n_1)/(Z_1 n_0)], \text{ or } \Delta A = -k_B T \ln \frac{n_1}{n_0} + C.
\]

(1.2.48)

The free energy difference then becomes

\[
\Delta A = k_B T \ln \frac{\sum_1 f(U_0 - U_1 + C)}{\sum_0 f(U_1 - U_0 - C)} - k_B T \ln \frac{n_1}{n_0} + C.
\]

(1.2.49)

Combining eqs. (1.2.48) and (1.2.49) will construct a self-consistent iterative algorithm to determine \( C \) and \( \Delta A \). This method was introduced by Charles H. Bennett [45], and is called the Bennett acceptance ratio (BAR) procedure.

The third method to be reviewed here is thermodynamic integration. An extra thermodynamic order parameter \( \lambda \) that ranges between 0 and 1 is added to the potential energy \( U \), to represent the current state of the system on the thermodynamic path from state 0 to state 1. The potential energy is simply defined as

\[
U(q, \lambda) = f(\lambda) U_0(q) + (1 - f(\lambda)) U_1(q),
\]

(1.2.50)

where the choice of function \( f(\lambda) \) is completely arbitrary, as long as \( f(0) = 0, f(1) = 1 \), and can be as simple as \( f(\lambda) = \lambda \).

The partial derivative of the free energy with respect to the order parameter \( \lambda \) is given by:

\[
\frac{\partial A}{\partial \lambda} = -k_B T \frac{\partial Z}{Z} \frac{\partial}{\partial \lambda} = -k_B T \frac{\partial}{\partial \lambda} \int d\mathbf{q} e^{-\beta U(q, \lambda)}
\]

\[
= -k_B T \frac{\partial}{\partial \lambda} \int d\mathbf{q} \left( -\beta \frac{\partial U}{\partial \lambda} \right) e^{-\beta U}
\]

\[
= \langle \frac{\partial U}{\partial \lambda} \rangle.
\]

(1.2.51)
Now the free energy difference between state 0 and state 1 can be obtained by integrating eq. (1.2.51) along the thermodynamic path

\[ \Delta A = \int_{0}^{1} \langle \frac{\partial U}{\partial \lambda} \rangle \, d\lambda. \] (1.2.52)

### 1.3 Supplementary Information

The supplementary information starts on page 23
Program 1.1: Euler-Cromer Integrator

c Euler-Cromer Integrator
c language: Fortran
c dt: time step
c pos: positions vel: velocities acc: accelerations
c
call calc_acc_from_pos (acc,pos)
do i = 1, N
   do j = 1, 3
      vel(j,i) = vel(j,i) + acc(j,i) * dt
      pos(j,i) = pos(j,i) + vel(j,i) * dt
   end do
end do
c
save the trajectory
c
call save_md_traj (pos,vel)
Program 1.2: Velocity and Position Verlet Integrators

c Verlet integrators
c language: Fortran
c
dt_2 = dt / 2
c
velocity Verlet integrator
c
if (vel_verlet) then
  vel = vel + acc * dt_2
  pos = pos + vel * dt
  call calc_acc_from_pos (acc, pos)
  vel = vel + acc * dt_2
end if

c position Verlet integrator
c
if (pos_verlet) then
  pos = pos + vel * dt_2
  call calc_acc_from_pos (acc, pos)
  vel = vel + acc * dt
  pos = pos + vel * dt_2
end if

c save the trajectory
c
  call save_md_traj (pos, vel)
Program 1.3: Velocity RESPA Integrator

```
c Velocity RESPA Integrator
language: Fortran

dti = dt / ns
dti_2 = dti / 2
do i = 1, ns
    vel = vel + acc * dti_2
    pos = pos + vel * dti
    call calc_fast_acc_from_pos (tcc, pos)
    acc = acc + tcc
    if (i .eq. ns) then
        call calc_slow_acc_from_pos (tcc, pos)
        acc = acc + tcc * ns
    end if
    vel = vel + acc * dti_2
end do

c save the trajectory

call save_md_traj (pos, vel)
```
CHAPTER 2

MONOVALENT IONS FOR THE AMOEBA FORCE FIELD

2.1 Introduction

Ions are ubiquitously involved in various chemical and biochemical processes, where in many cases, a notable concentration of ions play an important role in the reactions. However, the ion force fields applied in the computer simulations of such systems need careful calibration. Aqueous electrolyte solutions spanning wide ranges of concentrations are good systems to test the validity and the limits of the force field model before it is applied in the more complex systems, e.g. proteins, nucleic acids, etc. An interesting example was reported by Auffinger et al. \[46\]. Spontaneous formation of NaCl aggregates was found in a system of a short DNA duplex and 4 mol/L salt solution with Amber ion parameters and the standard TIP3P water model. KCl parameters directly adapted from Åqvist \[47\] in the Amber package form salt aggregates at 1 mol/L concentration. Similar phenomena were observed in K\textsuperscript{+}, Rb\textsuperscript{+}, and Cs\textsuperscript{+} salts solutions with the AMOEBA force field as well.

The work covered in this chapter is an attempt to reach a balance for the AMOEBA force field between single ion properties and salt crystal properties, which are the calibration targets of the ion model, in order to correctly describe ion-water interactions as well as ion-ion interactions, and to approach other properties, such as gas phase clusters computed by quantum methods, thermodynamic and transport property of aqueous
electrolyte solutions over a wide range of concentrations.

2.2 Methodologies

The ab initio (QM) calculations of Na\(^+\) and water carried out in this work mainly followed methods from Feller et al. [48] The canonical MP2 method with aug-cc-pVXZ basis sets for water and cc-pCVXZ basis sets for Na\(^+\) (X=Q for Na\(^+\)(H\(_2\)O) and Na\(^+\)(H\(_2\)O)\(_4\); X=T for Na\(^+\)(H\(_2\)O)\(_6\)) were used for structural optimization and single point energies. Na\(^+\)(H\(_2\)O) was optimized in Gaussian 09 [49]. Other Na\(^+\)(H\(_2\)O)\(_n\) were optimized in Q-Chem [50]. Optimized water-anion (F\(^-\), Cl\(^-\), Br\(^-\), and I\(^-\)) structures and single point energies were computed by the MP2 method and aug-cc-pVQZ basis set on all atoms (aug-cc-pVQZ-PP basis on I\(^-\)) in Gaussian 09. All of the basis sets were obtained from the EMSL Basis Set Exchange [51, 52] if they were not provided by the respective QM packages. The AMOEBA03 water model [25] was used in fitting and simulations.

The molecular dynamics (MD) were simulated by the GPU-accelerated Tinker-OpenMM [53] library. A typical FEP simulation was integrated via a 2 fs time step RESP integrator, and the trajectory was saved every 1 ps. Particle Mesh Ewald (PME) summation with 7 Å cutoff was used to account for electrostatics and polarization energies. The vdw cutoff was set to 12 Å. The temperature was held at 298 K by a Bussi thermostat [34, 35], and the pressure was held at 1.0 atm by a Monte Carlo barostat [42] if NpT ensemble was enabled. The box size used in NVT simulation was determined by averaging the box size from an NpT simulation, typically from the second half of a trajectory. The solvation free energy was calculated following the thermodynamic path in fig. 2.1. The intermediate states for \(\lambda_{vdw}\) are 0, 0.4, 0.5, 0.6, 0.65, 0.7, 0.8, 0.9, 1, and the intermediate states for \(\lambda_{ele}\) are 0.1, 0.2, \ldots, 0.9, 1.

The multipoles and polarizabilities were directly scaled by \(\lambda_{ele}\). The softcore buffered
\[ \lambda_{vdw} = 0, \lambda_{ele} = 0 \quad + \lambda_{vdw} \quad \lambda_{vdw} = 1, \lambda_{ele} = 0 \quad + \lambda_{ele} \quad \lambda_{vdw} = 1, \lambda_{ele} = 1 \]

Figure 2.1: Solvation process for FEP calculation.

14-7 vdw interaction for alchemical state \( \lambda \) employs the following functional form

\[
U_{vdw}(\lambda, \rho_{ij}, \epsilon_{ij}) = \lambda^5 \epsilon_{ij} \frac{(1 + 0.07)^7}{0.7(1 - \lambda)^2 + (\rho_{ij} + 0.07)^7} \left( \frac{1 + 0.12}{0.7(1 - \lambda)^2 + \rho_{ij}^7 + 0.12} - 2 \right). \quad (2.2.1)
\]

The reported statistical uncertainties were estimated by summing up the statistical uncertainties of all the intermediate states.

Several optimization programs were also used in this work. The molecular structural minimization and crystal minimization programs were provided by Tinker package. Other optimization methods were adapted from the SciPy library [54].

2.3 Revision of the Monovalent Ions in the AMOEBA Force Field

Because ions are not involved with any bonded interactions in the AMOEBA force field, the revision of ion parameters is restricted to the parameters for the nonbonded terms. Modification of ion polarizabilities is not considered for two reasons. First, the polarizabilities of cations are small, as a result, the induced dipoles and the polarization energies of the cations will also be small. Second, the pairwise interactions are not very sensitive to changes in polarizabilities under Thole’s scheme [24], even if the ion has a large polarizability (e.g. \( \Gamma^- \)). Since monopole values of ions in the AMOEBA model are fixed, the only parameters to be adjusted are the vdw parameters.
2.3.1 \( \text{Na}^+ \) and water

A brute force search was done near the current AMOEBA vdw parameters. As a ubiquitous ion, the parameters of \( \text{Na}^+ \) in the AMOEBA model have been elaborately calibrated. Its parameters are good initial values \((r^0 = 3.02 \, \text{Å}, \epsilon = 0.26 \, \text{kcal/mol}, \sim 130 \, \text{K in} \ k_B T \) scale) for fitting against higher level QM (QM = MP2/aug-cc-pVQZ) results. With these parameters, the molecular mechanics (MM) optimized \( \text{Na}^+(\text{H}_2\text{O}) \) structure shows a \( \text{C}_{2v} \) symmetry, which is the same as the QM optimized symmetry. The MM optimized ion-oxygen distance is 2.2329 Å and binding energy \( D_e \) is -23.4430 kcal/mol, which are not very different from the results of QM optimization (ion-oxygen distance is 2.21978 Å, and BSSE corrected electronic binding energy \( D_e \) is -23.75747 kcal/mol), indicating that the \( \text{Na}^+ \) parameters in the current AMOEBA force field (which was fit against 2.230 Å and -23.6 kcal/mol at MP2/aug-cc-pVTZ level of theory \([55]\)) are indeed a good initial starting point. It should be noted that the binding energy \( D_e \) used here is different from \( D_0 \) in the quantum mechanics that takes the zero point energy into consideration.

Figure 2.2: MP2 optimized \( \text{Na}^+(\text{H}_2\text{O}) \) structure.
The search for $r^0$ ranges from 2.800 Å to 3.195 Å with increments of 0.005 Å, and $\epsilon$ ranges from 0.25 kcal/mol to 0.30 kcal/mol with increments of 0.01 kcal/mol. The average norms of the forces on atoms in Na$^+$-(H$_2$O) dimer were minimized to 0.0001 kcal/Å. As fig. 2.3 shows, the best fit of the parameters for Na$^+$ are $r^0 = 2.955$ Å, $\epsilon = 0.28$ kcal/mol.

2.3.2 single ion solvation free energy

Using the revised Na$^+$ parameters in section 2.3.1 and the methods introduced in section 2.2, the single ion solvation free energy of Na$^+$ was estimated to be $-91.60 \pm 0.05$ kcal/mol.
kcal/mol. Because the differences of like-charged ions and the sum of opposite charge ions are experimentally accessible, the solvation free energies of other ions are determined. The vdw parameters fit against solvation free energies are tabulated in Table 2.1.

Table 2.1: Revised vdw parameters, $r^0$ (in Å), $\epsilon$ (in kcal/mol), and the simulated single ion solvation free energy (in kcal/mol).

<table>
<thead>
<tr>
<th>Ions</th>
<th>$r^0$</th>
<th>$\epsilon$</th>
<th>$A_{solv}^{a}$</th>
<th>Schmid$^{b}$</th>
<th>This Work$^{b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li$^+$</td>
<td>2.20</td>
<td>0.066</td>
<td>-116.91</td>
<td>25.10</td>
<td>25.28</td>
</tr>
<tr>
<td>Na$^+$</td>
<td>2.955</td>
<td>0.28</td>
<td>-91.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>K$^+$</td>
<td>3.68</td>
<td>0.35</td>
<td>-74.32</td>
<td>-17.45</td>
<td>-17.30</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>3.90</td>
<td>0.38</td>
<td>-69.05</td>
<td>-22.71</td>
<td>-22.57</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>4.14</td>
<td>0.42</td>
<td>-63.66</td>
<td>-28.20</td>
<td>-27.96</td>
</tr>
<tr>
<td>F$^-$</td>
<td>3.43</td>
<td>0.25</td>
<td>-116.71</td>
<td>-208.41</td>
<td>-208.34</td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>4.12</td>
<td>0.34</td>
<td>-86.12</td>
<td>-177.82</td>
<td>-177.74</td>
</tr>
<tr>
<td>Br$^-$</td>
<td>4.32</td>
<td>0.43</td>
<td>-79.66</td>
<td>-171.37</td>
<td>-171.28</td>
</tr>
<tr>
<td>I$^-$</td>
<td>4.61</td>
<td>0.52</td>
<td>-71.25</td>
<td>-163.00</td>
<td>-162.88</td>
</tr>
</tbody>
</table>

(a) Simulation results are not corrected by eq. (2.4.1) for different standard states. Statistical uncertainty is 0.06 kcal/mol.
(b) For cation M$^+$ (M = Li, Na, K, Rb, and Cs), this column is given in $A_{solv}(Na^+) - A_{solv}(M^+)$, while for anion X$^-$ (X = F, Cl, Br, I), this column is given in $A_{solv}(Na^+) + A_{solv}(X^-)$. Both are in kcal/mol. The reference for the experimental data is quoted from Schmid et al. [56].

2.3.3 Lattice energies and lattice constants

The properties of crystal lattices are also a good source of experimental data for ion parameterization. Using the parameters obtained from section 2.3.2, the molecular mechanics optimized lattice energies (LE) and the experimentally measured lattice energies are tabulated in Table 2.2. Only the lattice energies of NaCl, NaBr, and NaI are close to or slightly higher than the experimental measurement, while the rest of the molecular mechanics optimized lattice energies are all overestimated. The closest is KCl, but it is calculated to be 4.8 kcal/mol lower than experiment, and the difference can be as big as 9.6 kcal/mol for CsI. As a result, much stronger attraction is expected in the crystals of
potassium, rubidium, and cesium salts with the parameters fit in section 2.3.2, and in dissolution-precipitation equilibrium, the aggregation process will be favored. It may be difficult to observe the crystallization of KCl in an animation of an MD simulation, as periodic boundary conditions are applied so that crystallization in the simulation becomes a homogeneous nucleation process, which is hindered by the activation energy. However, the aggregating tendency can still be observed from the high first peak of the radial distribution function of K⁺ and Cl⁻. The crystallization of CsI during solution MD simulation is almost instantaneous.

Table 2.2: Lattice Energies (in kcal/mol) of the optimized salt crystal structures with revised vdw parameters in the AMOEBA force field, and lattice energies (in kcal/mol) measured in experiment.

<table>
<thead>
<tr>
<th>Crystal</th>
<th>Calculated LE</th>
<th>Experiment LE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaCl</td>
<td>-187.7</td>
<td>-188.8</td>
</tr>
<tr>
<td>RbCl</td>
<td>-172.4</td>
<td>-166.1</td>
</tr>
<tr>
<td>KCl</td>
<td>-176.9</td>
<td>-172.1</td>
</tr>
<tr>
<td>CsCl</td>
<td>-166.9</td>
<td>-160.1</td>
</tr>
<tr>
<td>NaBr</td>
<td>-179.0</td>
<td>-180.2</td>
</tr>
<tr>
<td>RbBr</td>
<td>-166.8</td>
<td>-159.7</td>
</tr>
<tr>
<td>KBr</td>
<td>-170.5</td>
<td>-165.2</td>
</tr>
<tr>
<td>CsBr</td>
<td>-162.5</td>
<td>-154.6</td>
</tr>
<tr>
<td>NaI</td>
<td>-167.9</td>
<td>-168.5</td>
</tr>
<tr>
<td>RbI</td>
<td>-159.3</td>
<td>-151.1</td>
</tr>
<tr>
<td>KI</td>
<td>-162.1</td>
<td>-155.4</td>
</tr>
<tr>
<td>CsI</td>
<td>-156.1</td>
<td>-146.5</td>
</tr>
</tbody>
</table>

It is not surprising that the revised vdw parameters do not correctly predict crystal properties, because the crystal properties themselves and the underlying ion-ion interactions were completely neglected in the parameterization. Two properties of crystals, lattice energy and lattice constant, are utilized to fit a separate set of parameters for ion-ion interactions. Except for CsCl, CsBr, and CsI, which belong to the Pm3m space group, other crystals involved in this work are in the Fm3m space group. Both space groups are
in the cubic crystal system, having lattice constants \( a = b = c \), and \( \alpha = \beta = \gamma = 90^\circ \). Therefore, only one independent variable is needed to describe the geometry of each crystal.

The inter-ionic distances (ID) as experimentally measured at room temperature were collected by Tosi [57] and Cubicciotti [58], and were extrapolated to 0 K by Ghate [59]. The differences in ID at two disparate temperatures are only around 1%. ID can be converted to lattice constant \( a \) by applying a factor of 2 for the \( Fm\bar{3}m \) space group, or by factor \( 2/\sqrt{3} \) for the \( Pm\bar{3}m \) space group. Jenkins et al. [60] summarized the lattice energies experimentally as obtained from Born-Fajans-Harber cycle. The lattice energies at room temperature differ from the lattice energies at 0 K at most by 0.1%.

The error function \( E(\Lambda) \) of the crystal fitting is defined as follows,

\[
E(\Lambda) = \sum_{M^+, X^-} \left[ (LE - LE_{exp.})^2 + r^2 (a - a_{exp.})^2 \right].
\]  

(2.3.1)

where \( \Lambda \) represents all the variables to be fit, \( M^+ \) and \( X^- \) are the cations and anions involved in the fitting, respectively, \( r \) is the coefficient to balance lattice energies and \( a \) values in the error function. \( r \) is set to 10, in expectation that the error in lattice energies will be less than 0.1 kcal/mol, and the error in \( a \) will be less than 0.01Å.

Li\(^+\), and F\(^-\) ions are excluded from the crystal fitting. Because the lattice energies Na\(^+\) salts calculated by the molecular mechanics method are in good agreement with the experimental data, Na\(^+\) is excluded from the fitting as well. Therefore, 12 parameters of K\(^+\), Rb\(^+\), Cs\(^+\), Cl\(^-\), Br\(^-\), and I\(^-\) are determined from the 18 properties of 9 salt crystals.

A few restraints are added to the optimization of the error function eq. (2.3.1), to ensure the fitting results do not violate common chemical intuition. \(^{(*)}\) The \( r^0 \) parameters

\(^{(*)}\) Unlike other salts, the molecular mechanics predicted lattice energies of salts containing Li\(^+\) and F\(^-\) are generally less negative by 10 kcal/mol (table 2.8), which guarantees that the ion pairs will not aggregate spontaneously. The cubic combining rule is another limiting factor. As shown in fig. 1.1, the cubic combining rule reaches its minimum at \( x = 0.596 \), forcing fitting Li\(^+\) with Cl\(^-\) and larger anions to be very difficult.
for all the ions are restricted to a relatively wide range (0.1 Å to 10.0 Å). The range for \( \epsilon \) parameters is 0.001 kcal/mol to 1.0 kcal/mol. (II) The constraints that the \( r^0 \) parameters should follow the ordering \( K^+ < Rb^+ < Cs^+ \), and \( Cl^- < Br^- < I^- \) are enforced. (III) The interactions between like-charged ions are still computed with the parameters in section 2.3.2 and are taken as constant lattice energies components. Parameters being fit are only used to calculate the opposite charge interactions, in order to simulate, for example, NaCl and KCl, mixtures.

The optimization was accomplished in three stages. First, L-BFGS-B optimization [61] was started from the parameters obtained in section 2.3.2 only with restriction (I). The derivatives of the error function with respect to \( \Lambda \) were numerically calculated. The error was reduced from 463 to 30 upon program termination. The search was then followed by constrained optimization via a linear approximation (COBYLA) method [62], which did not require derivatives for optimization, and introducing restriction (II). The error function was reduced to 2.0. Finally, the L-BFGS-B method was used again to further minimize the error function. This time the error function barely decreased.

The parameters from crystal fitting are tabulated in table 2.3. The are used as special pairwise vdw parameters for cation-anion interactions, and are tabulated in table 2.4.

Table 2.3: Vdw parameters fit from salt crystals for cation-anion vdw interactions.

<table>
<thead>
<tr>
<th>Ion</th>
<th>( r^0 ) (Å)</th>
<th>( \epsilon ) (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K(^+)</td>
<td>4.014</td>
<td>0.1001</td>
</tr>
<tr>
<td>Rb(^+)</td>
<td>4.199</td>
<td>0.1417</td>
</tr>
<tr>
<td>Cs(^+)</td>
<td>4.266</td>
<td>0.3318</td>
</tr>
<tr>
<td>Cl(^-)</td>
<td>4.419</td>
<td>0.2546</td>
</tr>
<tr>
<td>Br(^-)</td>
<td>4.651</td>
<td>0.3297</td>
</tr>
<tr>
<td>I(^-)</td>
<td>5.032</td>
<td>0.3624</td>
</tr>
</tbody>
</table>
Table 2.4: Pairwise vdw parameters, from applying cubic combining rule to $r^0$ and HHG combining rule to $\epsilon$ of crystal fitting.

<table>
<thead>
<tr>
<th>Cation</th>
<th>Anion</th>
<th>$r^0_{ij}$ (Å)</th>
<th>$\epsilon_{ij}$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K$^+$</td>
<td>Cl$^-$</td>
<td>4.236</td>
<td>0.1512</td>
</tr>
<tr>
<td>K$^+$</td>
<td>Br$^-$</td>
<td>4.379</td>
<td>0.1664</td>
</tr>
<tr>
<td>K$^+$</td>
<td>I$^-$</td>
<td>4.636</td>
<td>0.1720</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>Cl$^-$</td>
<td>4.315</td>
<td>0.1859</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>Br$^-$</td>
<td>4.448</td>
<td>0.2068</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>I$^-$</td>
<td>4.690</td>
<td>0.2145</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>Cl$^-$</td>
<td>4.345</td>
<td>0.2894</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>Br$^-$</td>
<td>4.475</td>
<td>0.3307</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>I$^-$</td>
<td>4.711</td>
<td>0.3466</td>
</tr>
</tbody>
</table>

2.4 Results and Discussion

2.4.1 Na$^+$ parameters

The parameters of Na$^+$ are fit against the optimization results from the MP2 method with large basis set. Holding the water structure fixed at the structure in the MP2 optimized dimer, the function of electronic binding energy $D_e$ with respect to Na$^+$-O distance (maintaining $C_{2v}$ symmetry) can be calculated at the same MP2 level with the same basis sets, and by the molecular mechanics method with the revised Na$^+$ parameters. The basis set superposition error (BSSE) corrected $D_e$ curve is plotted in fig. 2.4. The agreement is excellent.

The revised Na$^+$ parameters are also tested in two globally optimized Na$^+$(H$_2$O)$_n$ structures (fig. 2.5). The revised parameters work perfectly with the MP2 optimized structures. The total binding energies differ by only 0.8 kcal/mol. Optimizing the clusters with the force field gives bigger errors in the binding energies, but on average the difference is 0.4 kcal/mol per water molecule. The force field optimized Na$^+$-O distance is less than 0.1 Å from the ab initio value.
Figure 2.4: Electronic binding energy $D_e$ of Na$^+$ (H$_2$O) with respect to different Na$^+$-O distances ($R$) calculated by BSSE corrected MP2 and by the revised Na$^+$ parameters in the AMOEBA force field.

Table 2.5: Symmetry (in point group), Na$^+$-O distance ($R$, in Å), basis set superposition error (BSSE) corrected electronic binding energy ($D_e$, in kcal/mol) of QM optimized Na$^+$ (H$_2$O)$_n$ clusters, and $R$, $D_e$ predicted by AMOEBA force field with revised Na$^+$ parameters.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Symmetry</th>
<th>$D_e$ (QM)$^a$</th>
<th>$R$ (QM)</th>
<th>$D_e$ $^b$</th>
<th>$R$ $^c$</th>
<th>$D_e$ $^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$S_4$</td>
<td>-83.5828</td>
<td>2.2951</td>
<td>-82.3803</td>
<td>2.3448</td>
<td>-81.9725</td>
</tr>
<tr>
<td>6</td>
<td>$S_6$</td>
<td>-86.1011</td>
<td>2.3928</td>
<td>-86.9306</td>
<td>2.4538</td>
<td>-83.6892</td>
</tr>
</tbody>
</table>

(a) Electronic binding energies are measured with respect to Na$^+$ and (H$_2$O)$_n$.
(b) Cluster structures use the QM minimized structures.
(c) Cluster structures are minimized by force field with the revised Na$^+$ vdw parameters.
Figure 2.5: Optimized $\text{Na}^+(\text{H}_2\text{O})_4$ and $\text{Na}^+(\text{H}_2\text{O})_6$ structures. $\text{Na}^+(\text{H}_2\text{O})_4$ is of $S_4$ symmetry. $\text{Na}^+(\text{H}_2\text{O})_6$ is of $S_6$ symmetry.
2.4.2 single ion solvation free energy

A single ion solvation free energy is inaccessible via direct experimental measurement, and the reported values may differ by tens of kilocalories based on different experimental methods and theoretical assumptions (which are often flawed in various ways). However there is general agreement on the sum of solvation free energies of opposite charge ions comparing whole salts. Therefore, the solvation free energy of Na$^+$ is selected as our reference. Other cations are fit against the experimentally accessible solvation free energy differences. Anion values are then fit against the sum of the solvation free energies of the sodium salts.

It should be noted that the simulated solvation free energy is not directly comparable to typical experimental data, where the standard state of a solute in the gas phase is typically 1 atm in experiment, but in simulation it is assumed to be the same as the standard state of solution, which is 1 mol/L. Therefore, the difference of free energy between different standard states of the gas phase should be accounted for. If ideal gas behavior is applied, the correction is

$$T\Delta S = RT \ln \frac{V_{\text{atm}}}{V_{\text{1M}}} = 3.197RT.$$  \hspace{1cm} (2.4.1)

This corresponds to a +1.9 kcal/mol correction to the simulated single ion solvation free energy at room temperature.

The revised parameters and the simulated solvation free energies of the ions have been tabulated in table 2.1. The structures and corresponding energies of ion-water dimers are optimized in the force field and compared to the MP2 optimization results. As shown in fig. 2.6, unlike fig. 2.2, the optimized anion-water dimers reside in an asymmetric structure. Binding energies are generally in agreement with the QM results, except for Li$^+$ and F$^-$, where the accuracy of binding energies are sacrificed in order to better agree with solvation free energies. The accuracy of another gas phase property, the anion-oxygen
distance, is also impaired when the parameters fit to solvation free energy.

Figure 2.6: MP2 optimized Cl\(^-\) (H\(_2\)O) structure.

Table 2.6: Symmetry (in point group), ion-oxygen distance (\(R\) in Å), and basis set superposition error (BSSE) corrected electronic binding energy (\(D_e\), in kcal/mol) of MP2 optimized ion-water dimers, and \(R\), \(D_e\) predicted by AMOEBA force field with revised vdw parameters.

<table>
<thead>
<tr>
<th>Ion</th>
<th>Symmetry</th>
<th>(R) (QM)</th>
<th>(D_e) (QM)</th>
<th>(R)</th>
<th>(D_e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li(^+)</td>
<td>C(_{2v})</td>
<td>1.8337</td>
<td>-34.3570</td>
<td>1.7530</td>
<td>-36.1160</td>
</tr>
<tr>
<td>Na(^+)</td>
<td>C(_{2v})</td>
<td>2.2198</td>
<td>-23.7575</td>
<td>2.2200</td>
<td>-23.7617</td>
</tr>
<tr>
<td>K(^+)</td>
<td>C(_{2v})</td>
<td>2.5856</td>
<td>-17.7954</td>
<td>2.5802</td>
<td>-17.7336</td>
</tr>
<tr>
<td>Rb(^+)</td>
<td>C(_{2v})</td>
<td>2.861</td>
<td>-15.1</td>
<td>2.7170</td>
<td>-16.0761</td>
</tr>
<tr>
<td>Cs(^+)</td>
<td>C(_{2v})</td>
<td>3.066</td>
<td>-13.8</td>
<td>2.8803</td>
<td>-14.4003</td>
</tr>
<tr>
<td>F(^-)</td>
<td>C(_s)</td>
<td>2.4348</td>
<td>-31.7708</td>
<td>2.6528</td>
<td>-28.0557</td>
</tr>
<tr>
<td>Cl(^-)</td>
<td>C(_s)</td>
<td>3.0890</td>
<td>-15.4611</td>
<td>3.1434</td>
<td>-15.9952</td>
</tr>
<tr>
<td>Br(^-)</td>
<td>C(_s)</td>
<td>3.2420</td>
<td>-13.3357</td>
<td>3.3175</td>
<td>-13.7359</td>
</tr>
<tr>
<td>I(^-)</td>
<td>C(_s)</td>
<td>3.4553</td>
<td>-11.0398</td>
<td>3.5604</td>
<td>-11.2106</td>
</tr>
</tbody>
</table>

(a) QM results are taken from [48].

40
2.4.3 mean ionic activity coefficient

A bulk property of electrolyte solutions that can be used to validate the molecular dynamics model is mean ionic activity coefficient. The deviation of the salt chemical potential from ideal solution (Henry’s law) behavior can be quantified by mean ionic activity coefficient. For a 1:1 electrolyte solute, its mean ionic activity coefficient $\gamma_\pm$ in aqueous solution of molality $b$ is

$$\beta \mu_b = \beta \mu^\dagger + 2 \ln b \gamma_\pm,$$  \hspace{1cm} (2.4.2)

where $\mu_b$ is the chemical potential of solute at molality $b$, $\mu^\dagger$ is the chemical potential of solute at the standard state of Henry’s law. According to Shing and Chung \[63\], when inserting an extra solute molecule into a solution where there are already $N_s - 1$ solute molecules, the chemical potential of solute will become

$$\beta \mu_b = \ln \Lambda_s^3 q_s + 2 \ln \frac{N_s}{V} - \ln \frac{\langle V \exp(-\beta \psi_s) \rangle}{\langle V \rangle},$$  \hspace{1cm} (2.4.3)

where $\Lambda_s$ and $q_s$ are the thermal de Broglie wavelength and molecular partition function of solute, respectively. $\psi_s$ is the potential energy of solute. Assuming that the correlation between $\psi_s$ and volume $V$ is weak will introduce the single ion solvation free energy $A_{solv}$ at molality $b$ to eq. (2.4.3),

$$\ln \frac{\langle V \exp(-\beta \psi_s) \rangle}{\langle V \rangle} = \ln \langle \exp(-\beta \psi_s) \rangle = -\beta A_{solv}. \hspace{1cm} (2.4.4)$$

Equating eq. (2.4.2) and eq. (2.4.3) gives

$$2 \ln b \gamma_\pm - 2 \ln \frac{N_s}{V} - \beta A_{solv} = \text{constant}. \hspace{1cm} (2.4.5)$$

Subtracting eq. (2.4.5) by the infinitely diluted state (denoted as $\infty$) yields

$$\beta(A_{solv} - A_{solv}^\infty) = 2 \ln \frac{\gamma_\pm b V}{\gamma_\pm^\infty b^\infty V^\infty} - 2 \ln \frac{N_s}{N_s^\infty}. \hspace{1cm} (2.4.6)$$

With $\gamma_\pm^\infty = 1$, eq. (2.4.6) can further be simplified to

$$\beta \Delta A_{solv} = \beta(A_{solv} - A_{solv}^\infty) = 2 \ln \frac{\rho_w}{\rho_b} \left(1 + \frac{N_s M_s}{N_w M_w}\right), \hspace{1cm} (2.4.7)$$

41
where $\rho_b$ is the density of solution of molality $b$, $\rho_w$ is the density of solvent at the given temperature, $N_w$ is the number of solvent molecules, $M_w$ is the molar mass of the solvent, and $M_s$ is the molar mass of solute.

As it is not realistic to simulate the infinite dilution limit, the lowest concentration (0.1 mol/kg) that was simulated in this study was selected as the reference state, and its $\gamma_{\pm}$ was assumed to match the experiment result.

The simulated mean ionic activity coefficients of NaCl, KCl, and CsI solutions of different concentrations are plotted in figs. 2.7 to 2.9. Details of the simulations are tabulated in tables 2.11 to 2.13. The experiment data of mean ionic coefficient are taken from [64]. Densities of the solutions are calculated from an empirical formula found in [65].

Results of Mester el al. [66] are plotted together in fig. 2.7 where the mean ionic coefficients of NaCl are computed with five different force field models. The descending
Figure 2.8: Mean ionic activity coefficient of KCl solution at room temperature from 0 to 3 mol/kg.

Figure 2.9: Mean ionic activity coefficient of CsI solution at room temperature from 0 to 3 mol/kg.
trend is successfully reproduced in all six models at the lower concentrations (“salting in region”). In the range of higher concentration (“salting out region”), however, this work underestimated the $\ln \gamma_{\pm}$ by the same amount the SPC/E+SD model overestimated, which was the best among the five models.

The mean ionic activity coefficients of KCl and CsI of calculated in this work are in good agreement with the experimental measurements, as shown in figs. 2.8 and 2.9 respectively.

In dilute solution where the ions may be approximated by the point charges, the Debye-Hückel theory gives the following estimate of the mean ionic activity coefficient:

$$\log_{10} \gamma_c = -|z_+ z_-| A_c \sqrt{I_c},$$

(2.4.8)

where the subscript $c$ denotes the concentrations on the volume basis, and $A_c$ is given by

$$A_c = \frac{1}{\ln 10} \left( \frac{2\pi N_A}{1000} \right)^{1/2} \left( \frac{e^2}{4\pi \varepsilon_0 \varepsilon k_B T} \right)^{3/2}. $$

(2.4.9)

$A_m$ on the molality basis is $A_m = A_c (\rho_w)^{1/2}$, and is equal to 0.5108 at 25 °C. For more concentrated solutions, empirical coefficients $B^*$, $\beta$, $C$, $D$, etc. are introduced in the Debye-Hückel theory to account for the nonideality of the solution such that

$$\log_{10} = -|z_+ z_-| A_m \sqrt{T} \left( \frac{1}{1 + B^* \sqrt{T}} + \beta I + C I^2 + D I^3 + \ldots, \right)$$

(2.4.10)

where the empirical parameters “are not considered to have physical significance.” [64]

The curves of experimentally fit $\ln \gamma_{\pm}$ and $\ln \gamma_{\pm}$ predicted by Debye-Hückel theory are plotted in fig. 2.10.

### 2.4.4 radial distribution function (RDF)

The radial distribution functions (RDF, $g(r)$) for 3 mol/kg NaCl, KCl, and CsI, are shown in figs. 2.11 to 2.13. The shapes and positions of the peaks are almost independent of the concentrations of the solutions.
Figure 2.10: Experimentally fit $\ln \gamma_{\pm}$ and $\ln \gamma_{\pm}$ predicted by Debye-Hückel theory.

The most dramatic difference between the RDFs of NaCl and the other two, is in figs. 2.11 and 2.12. The first peaks of $M^+-O$ for $Na^+$, $K^+$, and $Cs^+$ are at 2.4 Å, 2.7 Å, and 3.0 Å, respectively. The height of the peaks are decreasing, which suggests a decreasing trend in the strength of cation-water interaction. This trend is independently supported by the MP2 calculations tabulated in table 2.6.

The relative height of the first peaks in figs. 2.11 and 2.12 represents the ratio of types of ions pairs in the solution. In fig. 2.11, two peaks can be easily identified for NaCl and KCl solutions. Additional peaks for NaCl, KCl, and CsI can only be vaguely identified. The first peaks for cation-anion pairs in fig. 2.11 supports the existence of contact ion
pairs in the solutions. The second peaks of the solutions in fig. 2.11 indicate the existence of solvent-separated ion pairs. Direct cation-anion interactions are becoming more predominant, and the ratio of contact ion pairs from NaCl, KCl, to CsI is increasing, which is consistent with the decreasing trend of cation-water interaction.

Direct x-ray measurement of the RDF of cation-anion pairs in aqueous alkali halide were attempted by Lawrence and Kruh [67] but no useful RDF data could be determined. Deverell and Richards [68] found that direct interaction between the ions was the predominant cause of the chemical shifts of the cation resonance in solution, where K\(^+\), Rb\(^+\) and Cs\(^+\) resonances varied linearly with the mean activity. Small shifts were found in Na\(^+\) resonances. Their later NMR study [69] concluded that the cation-water-halide ion inter-
action was of prime importance at concentrations below 8 mol/kg. Fuoss also reached the conclusion that Rb$^+$ and Cs$^+$ were contact pairs, while Li$^+$ and Na$^+$ were solvent separated based upon conductimetric experiments [70].

2.4.5 self-diffusion coefficient of water

The self-diffusion coefficient of water in an ion solution was evaluated via the Einstein equation eq. (1.2.27) from 20 ns trajectories of the fully solvated state via FEP calculations. The self-diffusion coefficient was then corrected according to the box size for periodic boundary conditions, as proposed by Yeh and Hummer [71].

$$D_{\text{corr.}} = \frac{k_B T \zeta}{6 \pi \eta L}, \quad (2.4.11)$$
where \( \xi = 2.837297 \), \( \eta \) is the shear viscosity of water, and \( L \) is the cubic box dimension.

The self-diffusion coefficients of water in table 2.7 were measured by Müller et al. [72]. Kim et al. [73] tested SCP/E, TIP3P, TIP4P, TIP4P/2005, TIP5P water models with various ion models, as well as two polarizable models AMOEBA and SWM4-DP, to simulate this property. The results for NaCl solutions qualitatively agreed with the trend from experiment, while for CsI solutions, all the models failed to give an increasing self-diffusion of water. Most of the models had the same decreasing diffusion result for CsI solutions, because the models “overemphasized the binding of water to the ions.” The AMOEBA force field parameter used in their study did not lower the self-diffusion of water in CsI solution. In fact, the Cs\(^+\) and I\(^-\) precipitated in the water box, even at a concentration far
below saturation.

Table 2.7: Periodic boundary condition (PBC) corrected self-diffusion coefficients of water.

<table>
<thead>
<tr>
<th>Solution</th>
<th>$D$ ($10^{-5}$ cm$^2$/s)</th>
<th>$D/D_w$ (simulation)</th>
<th>$D/D_w$ (experiment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 m NaCl</td>
<td>2.2282</td>
<td>0.971</td>
<td>0.948</td>
</tr>
<tr>
<td>2 m NaCl</td>
<td>2.0567</td>
<td>0.896</td>
<td>0.882</td>
</tr>
<tr>
<td>3 m NaCl</td>
<td>1.8770</td>
<td>0.818</td>
<td>0.817</td>
</tr>
<tr>
<td>4 m NaCl</td>
<td>1.7132</td>
<td>0.746</td>
<td>0.747</td>
</tr>
<tr>
<td>5 m NaCl</td>
<td>1.5109</td>
<td>0.658</td>
<td>0.686</td>
</tr>
<tr>
<td>6 m NaCl</td>
<td>1.3993</td>
<td>0.610</td>
<td>0.624</td>
</tr>
<tr>
<td>1 m CsI</td>
<td>2.4850</td>
<td>1.08</td>
<td>1.11</td>
</tr>
<tr>
<td>2 m CsI</td>
<td>2.4333</td>
<td>1.06</td>
<td>1.17</td>
</tr>
<tr>
<td>3 m CsI</td>
<td>2.4065</td>
<td>1.05</td>
<td>1.22</td>
</tr>
</tbody>
</table>

The sheer viscosity used in correction is experimental viscosity of pure water (0.896 cP [74]). The PBC corrected diffusion coefficient of pure water $D_w$ is $2.2956 \times 10^{-5}$ cm$^2$/s, and diffusion coefficient of pure water measured in experiments [75] is $2.29 \times 10^{-5}$ cm$^2$/s.

With the revised ion parameters and the special pairwise cation-anion parameters, the result of NaCl simulations quantitatively match experiment. For CsI, though water self-diffusion does not increase by the full 22% reported, the new parameters give a 5% to 8% increase, which is comparable to the results obtained via ab initio molecular dynamics (AIMD) by Ding et al. [76] and in a scaled ionic charge simulation by Kann et al. [77]. The scaled ionic charge model was proposed by Leontyev et al. [78,83] in the context of a continuum model for implicit solvent. They suggest reducing the point charge value of ions by various percentages based on different models and different permittivities, to account for the absence of polarizability in fixed charge models. The self-diffusion coefficients of water calculated in this work are consistent with conclusions from the RDFs, that the lower water diffusivity is caused by the predominant Na$^+$-water interactions, which therefore slow down the water diffusion, while Cs$^+$ will accelerate the water diffusion.
2.5 Conclusion

Three independent sources of data were used to calibrate monovalent ion parameters for the AMOEBA force field. Na\(^+\) parameters were first determined by fitting to high level quantum optimized structure and energy value. Then, the solvation free energy of Na\(^+\) was computed in AMOEBA03 water \cite{25}, and the parameters of other ions were fit to the relative solvation free energies reported by experiment. Lastly, the lattice energies and lattice constants of the monovalent ions were employed to balance ion-water interactions and ion-ion interactions, and to prevent the spontaneous aggregation of ions at low concentration.

The revised parameters are capable of optimizing ion-water dimers and ion-water clusters to the same structures as high level ab initio methods, and similar electronic binding energies can be calculated from ab initio and force field methods. A thermodynamic property, mean ionic activity coefficient, and a transport property, self-diffusion coefficients of water, at different concentrations were simulated by molecular dynamics with the revised parameters, and were compared to experimental results. Although the simulation results do not completely match all experimental data, the simulated results still qualitatively agree with the experiments, and are comparable or superior to other published results.

Ionic interactions in solution are found to be different for Na\(^+\) vs. larger monovalent cations in our simulations. Direct ionic interactions are predominant in solutions of K\(^+\) or bigger cation salts, while indirect ion-water interactions are predominant for Na\(^+\) at similar concentrations.

2.6 Supplementary Information

The supplementary information starts on page\textsuperscript{51}.
Program 2.1: AMOEBA Parameters for Monovalent Ions

# AMOEBA Monovalent Ions #

atom  6   6 Li+   "Lithium Ion Li+"  3   6.941  0
atom  7   7 Na+   "Sodium Ion Na+" 11  22.990  0
atom  8   8 K+    "Potassium Ion K+" 19  39.098  0
atom  9   9 Rb+   "Rubidium Ion Rb+" 37  85.468  0
atom 10  10 Cs+   "Cesium Ion Cs+" 55 132.905  0
atom 14  14 F-    "Fluoride Ion F-"  9  18.998  0
atom 15  15 Cl-   "Chloride Ion Cl-" 17  35.453  0
atom 16  16 Br-   "Bromide Ion Br-" 35  79.904  0
atom 17  17 I-    "Iodide Ion I-"  53 126.904  0

vdw  6   2.20  0.066
vdw  7   2.955 0.28
vdw  8   3.68  0.35
vdw  9   3.90  0.38
vdw 10  4.14  0.42
vdw 14  3.43  0.25
vdw 15  4.12  0.34
vdw 16  4.32  0.43
vdw 17  4.61  0.52

vdwpr 8  15  4.236  0.1512
vdwpr 8  16  4.379  0.1664
vdwpr 8  17  4.636  0.1720
vdwpr 9  15  4.315  0.1859
vdwpr 9  16  4.448  0.2068
vdwpr 9  17  4.690  0.2145
vdwpr 10 15  4.345  0.2894
vdwpr 10 16  4.475  0.3307
vdwpr 10 17  4.711  0.3466

multipole 6  0  0  1.00000
multipole 7  0  0  1.00000
multipole 8  0  0  1.00000
multipole 9 0 0 1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
multipole 10 0 0 1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
multipole 14 0 0 -1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
multipole 15 0 0 -1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
multipole 16 0 0 -1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
multipole 17 0 0 -1.00000
0.00000 0.00000 0.00000
0.00000
0.00000
0.00000
0.00000
polarize 6 0.0280 0.3900
polarize 7 0.1200 0.3900
polarize 8 0.7800 0.3900
polarize 9 1.3500 0.3900
polarize 10 2.2600 0.3900
polarize 14 1.3500 0.3900
polarize 15 4.0000 0.3900
polarize 16 5.6500 0.3900
polarize 17 7.2500 0.3900
Table 2.8: Lattice Energies (in kcal/mol) of the optimized salt crystal structures containing Li$^+$ and F$^-$ with revised vdw parameters in the AMOEBA force field, and lattice energies (in kcal/mol) measured in experiment.

<table>
<thead>
<tr>
<th>Crystal</th>
<th>Calculated LE</th>
<th>Experiment LE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiF</td>
<td>-241.4</td>
<td>-250.7</td>
</tr>
<tr>
<td>LiCl</td>
<td>-195.6</td>
<td>-206.5</td>
</tr>
<tr>
<td>LiBr</td>
<td>-184.5</td>
<td>-196.0</td>
</tr>
<tr>
<td>LiI</td>
<td>-170.8</td>
<td>-182.6</td>
</tr>
<tr>
<td>NaF</td>
<td>-210.7</td>
<td>-222.3</td>
</tr>
<tr>
<td>KF</td>
<td>-188.1</td>
<td>-198.1</td>
</tr>
<tr>
<td>RbF</td>
<td>-180.6</td>
<td>-190.0</td>
</tr>
<tr>
<td>CsF</td>
<td>-172.4</td>
<td>-181.4</td>
</tr>
</tbody>
</table>

Table 2.9: Experimental lattice energies (in kcal/mol) and lattice constants (in Å).

<table>
<thead>
<tr>
<th>Salt</th>
<th>Lattice Energy</th>
<th>Lattice Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiF</td>
<td>-250.7</td>
<td>3.992</td>
</tr>
<tr>
<td>LiCl</td>
<td>-206.5</td>
<td>5.078</td>
</tr>
<tr>
<td>LiBr</td>
<td>-196.0</td>
<td>5.426</td>
</tr>
<tr>
<td>LiI</td>
<td>-182.6</td>
<td>5.902</td>
</tr>
<tr>
<td>NaF</td>
<td>-222.3</td>
<td>4.590</td>
</tr>
<tr>
<td>NaCl</td>
<td>-188.8</td>
<td>5.578</td>
</tr>
<tr>
<td>NaBr</td>
<td>-180.2</td>
<td>5.908</td>
</tr>
<tr>
<td>NaI</td>
<td>-168.5</td>
<td>6.388</td>
</tr>
<tr>
<td>KF</td>
<td>-198.1</td>
<td>5.296</td>
</tr>
<tr>
<td>KCl</td>
<td>-172.1</td>
<td>6.232</td>
</tr>
<tr>
<td>KBr</td>
<td>-165.2</td>
<td>6.524</td>
</tr>
<tr>
<td>KI</td>
<td>-155.4</td>
<td>6.978</td>
</tr>
<tr>
<td>RbF</td>
<td>-190.0</td>
<td>5.578</td>
</tr>
<tr>
<td>RbCl</td>
<td>-166.1</td>
<td>6.518</td>
</tr>
<tr>
<td>RbBr</td>
<td>-159.7</td>
<td>6.820</td>
</tr>
<tr>
<td>RbI</td>
<td>-151.1</td>
<td>7.256</td>
</tr>
<tr>
<td>CsF</td>
<td>-181.4</td>
<td>5.964</td>
</tr>
<tr>
<td>CsCl</td>
<td>-160.1</td>
<td>4.068</td>
</tr>
<tr>
<td>CsBr</td>
<td>-154.6</td>
<td>4.235</td>
</tr>
<tr>
<td>CsI</td>
<td>-146.5</td>
<td>4.501</td>
</tr>
</tbody>
</table>
Table 2.10: Differences in the force field fit lattice energies (in kcal/mol) and lattice constants (in Å) with pairwise ion parameters.

<table>
<thead>
<tr>
<th>Salt</th>
<th>Lattice Energy (exp. - fit)</th>
<th>Lattice Constant (exp. - fit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KCl</td>
<td>0.282</td>
<td>-0.060</td>
</tr>
<tr>
<td>KBr</td>
<td>0.293</td>
<td>-0.069</td>
</tr>
<tr>
<td>KI</td>
<td>0.312</td>
<td>-0.063</td>
</tr>
<tr>
<td>RbCl</td>
<td>0.241</td>
<td>-0.048</td>
</tr>
<tr>
<td>RbBr</td>
<td>0.412</td>
<td>-0.039</td>
</tr>
<tr>
<td>RbI</td>
<td>0.178</td>
<td>-0.036</td>
</tr>
<tr>
<td>CsCl</td>
<td>-0.032</td>
<td>-0.043</td>
</tr>
<tr>
<td>CsBr</td>
<td>0.010</td>
<td>-0.049</td>
</tr>
<tr>
<td>CsI</td>
<td>0.151</td>
<td>-0.044</td>
</tr>
</tbody>
</table>

Table 2.11: Details of mean ionic activity coefficient calculations of NaCl.

<table>
<thead>
<tr>
<th>b (mol/kg)</th>
<th>$A_{solv}$ (kcal/mol)</th>
<th>$N_s$</th>
<th>$N_w$</th>
<th>$L$ (Å)</th>
<th>$\rho_s$ (g/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-177.8039</td>
<td>2</td>
<td>1110</td>
<td>32.163782</td>
<td>1.004</td>
</tr>
<tr>
<td>0.2</td>
<td>-177.9109</td>
<td>3</td>
<td>833</td>
<td>29.226857</td>
<td>1.010</td>
</tr>
<tr>
<td>0.5</td>
<td>-178.0750</td>
<td>8</td>
<td>888</td>
<td>29.897666</td>
<td>1.023</td>
</tr>
<tr>
<td>1.0</td>
<td>-178.1319</td>
<td>16</td>
<td>888</td>
<td>29.961284</td>
<td>1.045</td>
</tr>
<tr>
<td>2.0</td>
<td>-178.1970</td>
<td>31</td>
<td>888</td>
<td>29.791222</td>
<td>1.087</td>
</tr>
<tr>
<td>3.0</td>
<td>-178.2475</td>
<td>46</td>
<td>851</td>
<td>29.864110</td>
<td>1.123</td>
</tr>
<tr>
<td>4.0</td>
<td>-178.0892</td>
<td>60</td>
<td>833</td>
<td>29.836791</td>
<td>1.157</td>
</tr>
<tr>
<td>5.0</td>
<td>-178.0245</td>
<td>73</td>
<td>810</td>
<td>29.753814</td>
<td>1.189</td>
</tr>
<tr>
<td>6.0</td>
<td>-177.9779</td>
<td>86</td>
<td>796</td>
<td>29.791304</td>
<td>1.216</td>
</tr>
</tbody>
</table>

Table 2.12: Details of mean ionic activity coefficient calculations of KCl.

<table>
<thead>
<tr>
<th>b (mol/kg)</th>
<th>$A_{solv}$ (kcal/mol)</th>
<th>$N_s$</th>
<th>$N_w$</th>
<th>$L$ (Å)</th>
<th>$\rho_s$ (g/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-160,4263</td>
<td>2</td>
<td>1110</td>
<td>32.156838</td>
<td>32.156838</td>
</tr>
<tr>
<td>0.2</td>
<td>-160,4194</td>
<td>3</td>
<td>833</td>
<td>29.248378</td>
<td>29.248378</td>
</tr>
<tr>
<td>0.5</td>
<td>-160,5885</td>
<td>8</td>
<td>888</td>
<td>29.943841</td>
<td>29.943841</td>
</tr>
<tr>
<td>1.0</td>
<td>-160,6542</td>
<td>16</td>
<td>888</td>
<td>30.052624</td>
<td>30.052624</td>
</tr>
<tr>
<td>2.0</td>
<td>-160,7071</td>
<td>31</td>
<td>860</td>
<td>30.006686</td>
<td>30.006686</td>
</tr>
<tr>
<td>3.0</td>
<td>-160,7042</td>
<td>45</td>
<td>833</td>
<td>29.968305</td>
<td>29.968305</td>
</tr>
</tbody>
</table>
Table 2.13: Details of mean ionic activity coefficient calculations of CsI.

<table>
<thead>
<tr>
<th>$b$ (mol/kg)</th>
<th>$A_{solv}$ (kcal/mol)</th>
<th>$N_s$</th>
<th>$N_w$</th>
<th>$L$ (Å)</th>
<th>$\rho_s$ (g/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-134.9105</td>
<td>2</td>
<td>1110</td>
<td>32.192561</td>
<td>1.021</td>
</tr>
<tr>
<td>0.2</td>
<td>-134.9187</td>
<td>3</td>
<td>833</td>
<td>29.299658</td>
<td>1.042</td>
</tr>
<tr>
<td>0.5</td>
<td>-135.0792</td>
<td>8</td>
<td>888</td>
<td>30.098645</td>
<td>1.101</td>
</tr>
<tr>
<td>1.0</td>
<td>-135.3387</td>
<td>15</td>
<td>833</td>
<td>29.752802</td>
<td>1.192</td>
</tr>
<tr>
<td>2.0</td>
<td>-135.3808</td>
<td>29</td>
<td>805</td>
<td>29.954099</td>
<td>1.362</td>
</tr>
<tr>
<td>3.0</td>
<td>-135.4153</td>
<td>41</td>
<td>759</td>
<td>29.897388</td>
<td>1.512</td>
</tr>
</tbody>
</table>
CHAPTER 3

GENERALIZED STOCHASTIC ISOKINETIC INTEGRATOR

3.1 Introduction

Employing molecular dynamics to sample molecular configurations is always accompanied by two questions: how to enforce MD to sample the Boltzmann distribution in the configurational space, and what algorithms are capable of efficiently sampling the preferred distribution in terms of the cost in CPU time. Examples of thermostats were provided in section 1.2.2 to address the first problem. Multiple time step (MTS) algorithms [32, 84–90] were designed to improve the sampling efficiency. But the largest time steps available in MTS algorithms were still restricted by resonance phenomena [91, 92] to be less than 5 fs.

Additional algorithms are designed to control resonance phenomena in the MTS methods. Minary et al. [93] connected isokinetic constraints [94, 95] with the Nosé-Hoover Chain (NHC) algorithm [96] in every degree of freedom of motion. This algorithm efficiently restricts the kinetic energy that can possibly be built up in every degree of freedom and then prevents resonance phenomena. A stochastic process was added to the isokinetic constraints by Leimkuhler et al. [97] to ensure ergodicity. Margul et al. [98] made an effort to implement their stochastic isokinetic Nosé-Hoover RESPA algorithm, or SIN(R), for use with the AMOEBA force field [18]. This chapter will exhibit how the SIN(R) algorithm is extended and generalized in three respects. First, the Nosé-Hoover
chain algorithm can be implemented in the original SIN(R) integrator. The functional form of the isokinetic constraint is generalized so that it is no longer restricted to multiples of $k_B T$. And most importantly, the isokinetic constraint is extended to be able to constrain the kinetic energies of multi-dimensional velocities, instead of only one degree of freedom as in its original form.

In this chapter, the equations of motion for the generalized stochastic isokinetic integrator are proposed in section 3.2. The statistical thermodynamic properties of the corresponding dynamic system are proved in section 3.3. Algorithms for a single time step integration are derived in section 3.4 and are followed by multiple time step algorithms in section 3.5. The partition of short-range and long-range forces across multiple time steps is discussed in section 3.6. In section 3.7, preliminary applications of the generalized stochastic isokinetic integrator are presented. Conclusions and supplementary information are provided in section 3.8 and section 3.9, respectively.

### 3.2 Equations of Motion

Consider an $N$-atom system in which $L$ sets of auxiliary variables are coupled to every $n$ atoms and a total of $dn$ degrees of freedom. Each set of auxiliary variables consists of $P$ mass variables and $P$ velocity variables, denoted by $Q_{j,k}$ and $v_{j,k}$, respectively, where $j = 1, 2, \ldots, P$, and $k = 1, 2, \ldots, L$. Let $m$, $q$, and $v$ denote the masses of atoms, the $dn$ dimensional vectors for the positions of the atoms, and for the velocities of the atoms in the Cartesian coordinate system, respectively, where for simplicity, differences in the masses of the atoms are not discriminated. For each of the $dn$ degrees of freedom and their coupled auxiliary velocities, the stochastic equations of motion are designed as follows,

\[
\frac{dq}{dt} = v, \quad \frac{dv}{dt} = sF/m - \lambda v,
\]
\[ \frac{dv_{1,k}}{dt} = (-v_{2,k}v_{1,k} - \lambda v_{1,k}), \quad k = 1, 2, \ldots, L, \quad (3.2.3) \]

\[ \frac{dv_{j,k}}{dt} = -v_{j+1,k}v_{j,k} + \frac{G_{j,k}}{Q_{j,k}}, \quad j = 2, \ldots, P - 1, \quad k = 1, 2, \ldots, L, \quad (3.2.4) \]

\[ \frac{dv_{P,k}}{dt} = \frac{G_{P,k}}{Q_{P,k}} - \gamma v_{P,k} + \sigma \frac{dW}{dt}, \quad k = 1, 2, \ldots, L, \quad (3.2.5) \]

\[ \sigma = \sqrt{2\gamma k_B T/Q_{P,k}}, \quad (3.2.6) \]

\[ G_{j,k} = Q_{j-1,k}v_{j-1,k}^2 - k_B T, \quad (3.2.7) \]

where \( F \) is the negative potential gradient on the \( dn \) degrees of freedom, \( s \) is a force scalar to be determined, \( \gamma \) is a friction constant, \( W \) denotes the Wiener process, a white noise stochastic process, such that \( W(t) - W(0) \) is normally distributed on the distribution \( \mathcal{N}(0,t) \).

Besides all the equations of motion above, the physical \( dn \) dimensional velocities and the associated auxiliary velocities are coupled through the following isokinetic constraint

\[ mv^2 + r \sum_{k=1}^{L} Q_{1,k}v_{1,k}^2 = Mk_B T, \quad (3.2.8) \]

where \( r \) and \( M \) are the coefficients to be determined in section \( 3.3 \).

The the \( \lambda \) parameter in eq. \((3.2.2)\) needs to be solved for as well. Differentiating both sides of eq. \((3.2.8)\) with respect to \( t \) gives

\[ mv \cdot \frac{dv}{dt} + r \sum_{k=1}^{L} Q_{1,k}v_{1,k} \frac{dv_{1,k}}{dt} = 0. \quad (3.2.9) \]

Substituted by eqs. \((3.2.2)\) and \((3.2.3)\), eq. \((3.2.9)\) becomes

\[ sF \cdot v - \lambda mv^2 - r \sum_{k=1}^{L} Q_{1,k}v_{1,k}^2(v_{2,k} + \lambda) = 0, \quad (3.2.10) \]
then the multiplier $\lambda$ is determined

$$
\lambda = \frac{s \mathbf{F} \cdot \mathbf{v} - r \sum_{k=1}^{L} Q_{1,k} v_{1,k}^2 v_{2,k}}{m v^2 + r \sum_{k=1}^{L} Q_{1,k} v_{1,k}^2}
= \frac{s \mathbf{F} \cdot \mathbf{v} - r \sum_{k=1}^{L} Q_{1,k} v_{1,k}^2 v_{2,k}}{M k_B T}.
$$

(3.2.11)

### 3.3 Isokinetic Ensemble

#### 3.3.1 time invariant measure

For a system conforming to the microcanonical (NVE) ensemble and following $dp/dt = -\nabla U(q)$, the partition function can be expressed in compact notation

$$
\Omega(N, V, E) = C_N \int dx \delta(H(x) - E),
$$

(3.3.1)

where $C_N$ is a constant coefficient serving as the normalization factor, $\delta(x)$ is the Dirac delta function, and $H$ is the Hamiltonian. It is conventional to use the long time average from simulations to estimate the ensemble average of the property of interest, therefore, the assumption regarding the volume of phase space at different times, $dx_t = dx_0$, is implicit, meaning the volume of the phase space element $dx$ is conserved during the evolution of the system. This assumption should be verified, as outlined below.

The time evolution of the volume element follows

$$
dx_t = J(x_t; x_0) \, dx_0,
$$

(3.3.2)

where $J(x_t; x_0)$ is the determinant of the Jacobian matrix for the transformation. By definition, the elements of the Jacobian matrix are

$$
J_{ij}(x_t; x_0) = \frac{\partial x_t^i}{\partial x_0^j},
$$

(3.3.3)

and the elements of the inverse of the Jacobian matrix $J$ are

$$
J_{ij}^{-1}(x_t; x_0) = J_{ij}(x_0; x_t) = \frac{\partial x_0^i}{\partial x_t^j}.
$$

(3.3.4)
The time derivative of \( J(x; x_0) \) is given by

\[
\frac{d}{dt} J(x; x_0) = \frac{d}{dt} e^{\text{tr}(\ln J)}
\]

\[
= e^{\text{tr}(\ln J)} \text{tr} \left( \frac{dJ}{dt} J^{-1} \right)
\]

\[
= J(x; x_0) \sum_{i,j} \left( \frac{dJ_{ij}}{dt} J^{-1}_{ji} \right)
\]

\[
= J(x; x_0) \sum_{i,j} \left( \frac{\partial \dot{x}_i}{\partial x_j^0} \frac{\partial x_j^0}{\partial x_i^t} \right)
\]

\[
= J(x; x_0) \sum_i \frac{\partial \dot{x}_i}{\partial x_i^t}
\]

\[
= J(x; x_0) \nabla \dot{x}_t. \quad (3.3.5)
\]

The expression in eq. (3.3.5), \( \nabla \dot{x} \), is known as the compressibility \( \kappa \) of the phase space \cite{99, 100}. Equation (3.3.2) is then transformed to

\[
dx_t = \exp \left( \int_0^t \kappa \, dt \right) dx_0, \quad (3.3.6)
\]

or

\[
\exp \left( \int_0^{t_1} -\kappa \, dt \right) dx_{t_1} = \exp \left( \int_0^{t_0} -\kappa \, dt \right) dx_{t_0}. \quad (3.3.7)
\]

For the systems governed by the equations of motion, \( \dot{q} = v \), and \( m \ddot{v} = -\partial U/\partial q \), the phase space compressibility is given by

\[
\kappa = \frac{\partial}{\partial q} \dot{q} + \frac{\partial}{\partial v} \dot{v} = 0. \quad (3.3.8)
\]

Therefore, \( dx_t = dx_0 \) is verified.

If the compressibility \( \kappa \) is not 0, the premise that \( dx_t = dx_0 \) fails, in which case, the partition function of the ensemble cannot be expressed in the form of eq. (3.3.1). However, if it is possible to find a closed-form \( \sqrt{g(x)} \), such that

\[
\sqrt{g(x)} = \exp \left( -\int \kappa \, dt \right), \quad (3.3.9)
\]
another time invariant measure regarding the phase space volume, instead of \( d x_t = d x_0 \), can be defined

\[
\sqrt{g(x_t)} \, d x_t = \sqrt{g(x_0)} \, d x_0,
\]

and an analogous microcanonical partition function can be constructed

\[
\Omega = C_N \int d x \sqrt{g(x)} \prod_i \delta(C_i - \Lambda_i),
\]

where \( C_N \) is the constant coefficient, and \( C_i = \Lambda_i \) are the constraints applied to the system \[101\].

3.3.2 partition function of the isokinetic ensemble

Because \( L \) sets of auxiliary variables are coupled to every \( dn \) degrees of freedom in an \( N \)-atom system, the system can be separated into \( 3N/dn \) subsystems. The total phase space compressibility is the sum of the compressibilities of the subsystems

\[
\kappa = \sum_{\alpha=1}^{3N/dn} \kappa_\alpha,
\]

where each \( \kappa_\alpha \) is individually determined by

\[
\kappa_\alpha = \sum_{i=1}^{dn} \frac{\partial \dot{q}_i}{\partial q_i} + \sum_{i=1}^{dn} \frac{\partial \dot{v}_i}{\partial v_i} + \sum_{k=1}^{L} \sum_{j=1}^{P} \frac{\partial \dot{v}_{j,k}}{\partial v_{j,k}}.
\]

The terms in the eq. (3.3.13) are as follows,

\[
\frac{\partial \dot{q}_i}{\partial q_i} = 0,
\]

\[
\frac{\partial \dot{v}_i}{\partial v_i} = -\lambda - \left( \frac{\partial \lambda}{\partial v_i} \right) v_i,
\]

\[
\frac{\partial \dot{v}_{1,k}}{\partial v_{1,k}} = -v_{2,k} - \lambda - \left( \frac{\partial \lambda}{\partial v_{1,k}} \right) v_{1,k},
\]

\[
\frac{\partial \dot{v}_{j,k}}{\partial v_{j,k}} = -v_{j+1,k}, \quad j = 2, \ldots, P - 1,
\]
where in eq. (3.3.18), only the deterministic part of eq. (3.2.5) is taken into account. Leimkuhler et al. [97] have shown that ignoring the friction part and the stochastic part of eq. (3.2.5) does not affect the conclusion of the analysis.

Applying the chain rule to eq. (3.2.11) gives

$$\frac{\partial \lambda}{\partial v_i} = \frac{1}{M k_B T} \left( s F_i - 2 \lambda m v_i \right), \quad (3.3.19)$$

$$\frac{\partial \lambda}{\partial v_{1,k}} = - \frac{2}{M k_B T} \left( r Q_{1,k} v_{1,k} v_{2,k} + \lambda Q_{1,k} v_{1,k} \right), \quad (3.3.20)$$

Then inserting eqs. (3.3.14) to (3.3.20) into eq. (3.3.13), \( \kappa_\alpha \) will be simplified to

$$\kappa_\alpha = \left[ \frac{(L + dn - 1)s}{M k_B T} \sum_{i=1}^{dn} (-F_i v_i) \right] + \left[ \frac{(L + dn)r}{M k_B T} \sum_{k=1}^{L} Q_{1,k} v_{1,k}^2 - \sum_{k=1}^{L} \sum_{j=2}^{P} v_{j,k} \right]. \quad (3.3.21)$$

Upon summing over \( \alpha \), the total phase space compressibility is

$$\kappa = \frac{1}{k_B T} \left\{ \left[ \frac{(L + dn - 1)s}{M} \sum_{i=1}^{3N} (-F_i v_i) \right] \right. \right.$$

$$\left. + \sum_{\alpha} \sum_{k=1}^{L} \left[ \frac{(L + dn)r}{M} Q_{1,k} v_{1,k}^2 - \sum_{j=2}^{P} v_{j,k} \right] \right\}. \quad (3.3.22)$$

Let

$$s = \frac{M}{L + dn - 1}, \quad (3.3.23)$$

$$r = \frac{M}{L + dn}, \quad (3.3.24)$$

then it is easy to verify that eq. (3.3.22) becomes

$$\kappa = \beta \left[ \sum_{i=1}^{3N} (-F_i v_i) + \sum_{\alpha} \sum_{k=1}^{L} \left( Q_{1,k} v_{1,k}^2 - \sum_{j=2}^{P} \beta^{-1} v_{j,k} \right) \right]$$

$$= \beta \frac{d}{dt} \left( U(q) + \frac{1}{2} \sum_{\alpha} \sum_{k=1}^{L} \sum_{j=2}^{P} Q_{j,k} v_{j,k}^2 \right)$$

$$\equiv \beta \frac{d}{dt} \left[ U(q) + K(v_{j,k}) \right]. \quad (3.3.25)$$
Using eqs. (3.3.9), (3.3.11) and (3.3.25), the partition function of the isokinetic ensemble is obtained, as

\[
\Omega(N, V, T) = C_N \int dq \, e^{-\beta U(q)} \int dv_{j,k} \, e^{-\beta K(v_{j,k})} \times \prod_{\alpha=1}^{3N/dn} \delta \left( m v_{1,k}^2 + r \sum_{k=1}^{L} Q_{1,k} v_{1,k}^2 - M \beta^{-1} \right) \propto \int dq \, e^{-\beta U(q)},
\]

where the distribution is indeed canonical in the coordinate phase space. It is important to note that although the sampled configurations conform to the Boltzmann distribution, dynamical properties are not necessarily preserved.

3.3.3 remarks

At least three combinations of \(d\) and \(n\) values are evident and physically well-defined. \((d = 1, n = 1), (d = 3, n = 1), \) and \((d = 3, n = N)\) correspond to isokinetic constraints coupled to every degree of freedom (dof), to each individual atom (atom), and to the collection of all atoms of the system (global), respectively. All three cases are straightforward to implement. Other combinations are also possible, e.g. coupling the isokinetic constraints on each molecule, although this could be difficult in practice due to bookkeeping issues for heterogeneous systems.

While \(L\) and \(P\) have to be non-negative integers, in the original formulation of Leimkuhler et al. [97], \(M\) is unnecessarily restricted to be a positive integer. As a matter of fact, the value of \(M\) can be any positive real number. Once all four parameters \(d, n, L, \) and \(M\) are selected, the form of the isokinetic constraint is determined. One may also require that the force scalar be set to \(s = 1\), in which case \(M\) is also fixed to \(L + dn - 1\) and the form of the isokinetic constraint becomes \(L\)-dependent.
Equation (3.2.8) ensures that the kinetic energy in the $dn$ degrees of freedom of motion will never exceed $Mk_B T/2$ thus preventing resonance artifacts. Changing the $L$ and $M$ parameters separately provides a more precise control of the isokinetic constraints.

### 3.4 Single Time Step Integration Algorithms

The derivation of a single time step integrator starts from the Liouville operator of the equations of motion provided in section 3.2, $iL = \dot{x} \partial/\partial x$, where $x = (q, v, v_{j,k})$.

$$iL = v \frac{\partial}{\partial q} + \left( \frac{sF}{m} - \lambda v \right) \frac{\partial}{\partial v} + \sum_{k=1}^{L} \left( -v_{2,k}v_{1,k} - \lambda v_{1,k} \right) \frac{\partial}{\partial v_{1,k}} + \sum_{j=2}^{P} \dot{v}_{j,k} \frac{\partial}{\partial v_{j,k}}. \quad (3.4.1)$$

$\lambda$ is split into two parts, $\lambda_F$ and $\lambda_N$, where

$$\lambda_F = \frac{sF \cdot v}{Mk_B T}, \quad (3.4.2)$$

$$\lambda_N = -\tau \sum_{k'=1}^{L} \frac{Q_{1,k'}v_{1,k'}^2v_{2,k'}}{Mk_B T}. \quad (3.4.3)$$

The Liouville operator $iL$ is the sum of $iL_\alpha$s (where the subscript $\alpha$ denotes the subsystems of the $N$-atom system, see eq. (3.3.12)), where each $iL_\alpha$ is the sum of five terms,

$$iL = \sum_{\alpha} iL_\alpha, \quad (3.4.4)$$

$$iL_\alpha = iL_q + iL_v + iL_N + iL_{OU}, \quad (3.4.5)$$

where

$$iL_q = v \frac{\partial}{\partial q}, \quad (3.4.6)$$

$$iL_v = \left( \frac{sF}{m} - \lambda_F v \right) \frac{\partial}{\partial v} - \sum_{k=1}^{L} \lambda_F v_{1,k} \frac{\partial}{\partial v_{1,k}}, \quad (3.4.7)$$
\[ iL_N = -\lambda_N v \frac{\partial}{\partial v} + \sum_{k=1}^{L} \left[ -\lambda_N v_{1,k} \frac{\partial}{\partial v_{1,k}} + \sum_{j=2}^{P} G_{j,k} \frac{\partial}{\partial v_{j,k}} - \sum_{j=1}^{P-1} v_{j+1,k} \frac{\partial}{\partial v_{j,k}} \right], \]  
(3.4.8)

and \( iL_{OU} \) is the Ornstein-Uhlenbeck stochastic process built in \( v_{P,k} \).

### 3.4.1 \( iL_q \) and \( iL_{OU} \) operators

Equation (3.4.6) is equivalent to the trivial differential equation \( v = \dot{q} \), where the solution is

\[ e^{iL_q \tau} q = q + v \tau. \]  
(3.4.9)

For the Ornstein-Uhlenbeck stochastic process, we consider a more general form of the differential equation

\[ dx_t = -\beta (x_t - \alpha) \, dt + \sigma \, dW_t, \]  
(3.4.10)

where \( \beta > 0 \), \( \alpha \in \mathbb{R} \), and \( \sigma > 0 \). The solution is

\[ x_t = \alpha + (x_0 - \alpha) e^{-\beta t} + \sigma \int_0^t e^{-\beta (t-s)} \, dW_s. \]  
(3.4.11)

The conditional expectation of \( x_t \) is

\[ E[x_t] = E \left[ \alpha + (x_0 - \alpha) e^{-\beta t} + \sigma \int_0^t e^{-\beta (t-s)} \, dW_s \right] = \alpha + (x_0 - \alpha) e^{-\beta t}. \]  
(3.4.12)

Using Itô calculus, the conditional variance of \( x_t \) is

\[ \text{var}[x_t] = \sigma^2 E \left[ \left( \int_0^t e^{-\beta (t-s)} \, dW_s \right)^2 \right] = \sigma^2 E \left[ \int_0^t e^{-2\beta (t-s)} \, ds \right] = \frac{\sigma^2}{2\beta} (1 - e^{-2\beta t}). \]  
(3.4.13)

Therefore, \( x_t \) is normally distributed on the distribution \( \mathcal{N}(\alpha + (x_0 - \alpha) e^{-\beta t}, \frac{\sigma^2}{2\beta} (1 - e^{-2\beta t})) \), and the following solution for the \( iL_{OU} \) operator is obtained

\[ e^{iL_{OU} \tau} v_{P,k} = v_{P,k} e^{-\gamma \tau} + \sigma \sqrt{\frac{1 - e^{-2\gamma \tau}}{2\gamma}} R, \]  
(3.4.14)

where \( R \) is a random number normally distributed on the distribution \( \mathcal{N}(0, 1) \).
3.4.2 \(iL_v\) operator

The operator \(iL_v\) corresponds to the following differential equations,

\[
\frac{dv}{dt} = \frac{sF}{m} - \lambda_F v, \tag{3.4.15}
\]

\[
\frac{dv_{1,k}}{dt} = -\lambda_F v_{1,k}. \tag{3.4.16}
\]

Following the procedures in [95], the coordinates of the system and the gradients of the potential energy are treated as constants at time \(t\), and the rest of the variables are treated as functions of time. The solution for the \(iL_v\) operator is

\[
v(t) = v(0) + \left(\frac{sF}{m}\right)u(t), \tag{3.4.17}
\]

\[
v_{1,k}(t) = \frac{v_{1,k}(0)}{\dot{u}(t)}, \tag{3.4.18}
\]

where \(u(t)\) is a function to be determined. It can be immediately verified that

\[
\frac{dv(t)}{dt} = \frac{sF}{m} - \frac{\ddot{u}(t)}{\dot{u}(t)}v(t). \tag{3.4.19}
\]

Comparing eq. (3.4.19) to eq. (3.4.15) gives \(\ddot{u}(t)/\dot{u}(t) = \lambda_F(t)\). Inserting eq. (3.4.17) into eq. (3.4.15) yields

\[
\ddot{u}(t) = \frac{(sF)^2}{mk_BT}u(t) - \frac{sF \cdot v(0)}{mk_BT} = 0. \tag{3.4.20}
\]

or

\[
\ddot{u}(t) - \frac{(sF)^2}{mk_BT}u(t) = \frac{sF \cdot v(0)}{mk_BT}. \tag{3.4.21}
\]

The boundary conditions for eq. (3.4.21) need to be determined. Letting \(h(t) = \int \lambda_F(t') \, dt'\), it is obvious that

\[
\dot{u}(t) = \exp \left(\int_0^t \lambda_F(t') \, dt'\right) = \exp(h(t) - h(0)), \tag{3.4.22}
\]

\[
\dot{u}(t) = \exp(-h(0)) \int_0^t \exp(h(t')) \, dt'. \tag{3.4.23}
\]
and two boundary conditions for eq. (3.4.21) are found: \( \dot{u}(0) = 1 \) and \( u(0) = 0 \).

Finally, the solution for eq. (3.4.21) is given by

\[
  u(t) = \frac{1}{\sqrt{b}} \sinh \left( \sqrt{bt} \right) + \frac{a}{b} \left( \cosh \left( \sqrt{bt} \right) - 1 \right),
\]

(3.4.24)

and

\[
  \dot{u}(t) = \cosh \left( \sqrt{bt} \right) + \frac{a}{\sqrt{b}} \sinh \left( \sqrt{bt} \right),
\]

(3.4.25)

where

\[
  a = \frac{s \mathbf{F} \cdot \mathbf{v}(0)}{M k_B T},
\]

(3.4.26)

\[
  b = \frac{(s \mathbf{F})^2}{m M k_B T},
\]

(3.4.27)

and the classical propagator \( \exp(iL_v \tau) \) yields

\[
  e^{iL_v \tau} \mathbf{v} = \mathbf{v} + \frac{(s \mathbf{F}/m)u(\tau)}{\dot{u}(\tau)},
\]

(3.4.28)

\[
  e^{iL_v \tau} v_{1,k} = \frac{v_{1,k}}{\dot{u}(\tau)},
\]

(3.4.29)

### 3.4.3 \( iL_N \) operator

Because \( \exp(iL_N t) \) cannot be solved explicitly, the \( iL_N \) operator is then divided into two parts,

\[
  iL_{N,1} = -\lambda_N \mathbf{v} \frac{\partial}{\partial \mathbf{v}} + \sum_{k=1}^{L} \left( -\lambda_N v_{1,k} \frac{\partial}{\partial v_{1,k}} - v_{2,k} v_{1,k} \frac{\partial}{\partial v_{1,k}} \right),
\]

(3.4.30)

\[
  iL_{N,2} = \sum_{k=1}^{L} \left( \sum_{j=2}^{P} \frac{Q_{j,k}}{Q_{j,k} v_{j,k}} \frac{\partial}{\partial v_{j,k}} - \sum_{j=2}^{P-1} v_{j+1,k} v_{j,k} \frac{\partial}{\partial v_{j,k}} \right).
\]

(3.4.31)

The \( iL_{N,1} \) operator is equivalent to the following differential equations, where \( v_{2,k} \) is held fixed,

\[
  \frac{d\mathbf{v}}{dt} = \frac{r}{M k_B T} \left( \sum_{k'=1}^{L} Q_{1,k'} v_{1,k'}^2 v_{2,k'} \right) \mathbf{v},
\]

(3.4.32)
\[
\frac{dv_{1,k}}{dt} = \frac{r}{M k B T} \left( \sum_{k'=1}^{L} Q_{1,k'} v_{1,k'}^2 v_{2,k'} \right) v_{1,k} - v_{2,k} v_{1,k}.
\]

(3.4.33)

Without providing a detailed derivation, the solutions for these differential equations are directly given here,

\[
e^{iL \tau} v = v H(\tau),
\]

(3.4.34)

\[
e^{iL \tau} v_{1,k} = v_{1,k} H(\tau) \exp(-v_{2,k} \tau),
\]

(3.4.35)

where

\[
H(\tau) = \sqrt{\frac{M k B T}{m v^2 + r \sum_{k'=1}^{L} Q_{1,k'} v_{1,k'}^2 e^{-2u_{2,k'}}}}.
\]

(3.4.36)

Let \( f = v_{j+1,k} \), \( g = G_{j,k}/Q_{j,k} \), then the element in the \( \exp(i L N, t) \) operator is equivalent to

\[
\frac{du(t)}{dt} = -fu(t) + g,
\]

(3.4.37)

with the boundary condition \( u(0) \) being a known constant \( (v_{j,k}(0)) \). The solution of eq. (3.4.37) is

\[
u(t) = \left( u(0) - \frac{g}{f} \right) \exp(-ft) + \frac{g}{f}
\]

\[
= u(0) \exp(-ft) + tg \exp \left( -\frac{ft}{2} \right) \frac{\sinh(ft/2)}{ft/2},
\]

(3.4.38)

which can result in numerical problems near \( ft = 0 \). Expanding \( \sinh(x/2)/(x/2) \) in a Maclaurin series,

\[
\frac{\sinh(x/2)}{x/2} \sim 1 + O \left( \frac{x^2}{24} \right),
\]

(3.4.39)

\( u(t) \) can be approximated by

\[
u(t) \sim u(0) \exp(-ft) + tg \exp(-ft/2),
\]

(3.4.40)

and the corresponding the solution is given by

\[
e^{iL \tau} v_{j,k} = v_{j,k} \exp(-v_{j+1,k} \tau) + \tau \frac{G_{j,k}}{Q_{j,k}} \exp(-v_{j+1,k} \frac{\tau}{2}).
\]

(3.4.41)
The approximation in eq. (3.4.41) originates from the following factorization of $\exp(iL_{N,2}t)$

$$
\exp(iL_{N,2}t) \sim \exp \left( -\frac{t}{2} v_{j+1,k} v_{j,k} \frac{\partial}{\partial v_{j,k}} \right) \exp \left( t G_{j,k} \frac{\partial}{\partial v_{j,k}} \right) \exp \left( -\frac{t}{2} v_{j+1,k} v_{j,k} \frac{\partial}{\partial v_{j,k}} \right).
$$

To achieve higher accuracy, $\exp(iL_{N}t)$ should be integrated via smaller time steps as follows,

$$
\exp(iL_{N}t) = \prod_{k=1}^{n_c} \exp \left( iL_{N} \frac{t}{n_c} \right),
$$

where $n_c$ is the number of multiple time steps, and $n_c > 1$. The Suzuki-Yoshida factorization [102–104] is used in every $\exp(iL_{N}t/n_c)$ step, since $\exp(iL_{N}t)$ cannot be solved explicitly,

$$
\exp \left( iL_{N} \frac{t}{n_c} \right) \sim \prod_{j=1}^{n_{sy}} \left[ \exp \left( iL_{N,2} \frac{w_j t}{2n_c} \right) \exp \left( iL_{N,1} \frac{w_j t}{n_c} \right) \exp \left( iL_{N,2} \frac{w_j t}{2n_c} \right) \right].
$$

where $w_j$ are predetermined the coefficients based on $n_{sy}$, e.g. $n_{sy} = 3$, $w_1 = w_3 = 1/(2 - \sqrt{2})$, $w_2 = 1 - 2w_1$; $n_{sy} = 5$, $w_1 = w_2 = w_4 = w_5 = 1/(4 - \sqrt{4})$, $w_3 = 1 - 4w_1$, etc.

### 3.5 Multiple Time Step Integration Algorithms

Following the discussion in section [1.2.1], suppose the total force consists of three terms: a fast ($f$) component, medium ($m$) component, and slow ($s$) component,

$$
F = F_f + F_m + F_s.
$$

Then $\lambda_F$ and the $iL_v$ operator can be written as the sum of three components

$$
\lambda_F = \lambda_{F,f} + \lambda_{F,m} + \lambda_{F,s},
$$

$$
iL_v = iL_{v,f} + iL_{v,m} + iL_{v,s}.
$$
Let \( t, \Delta t, \) and \( \delta t \) be the time steps associated with the slow, medium, and fast force components, respectively, and the ratios \( n_m = t/\Delta t \) and \( n_f = \Delta t/\delta t \) be positive integers. Then the following velocity Verlet style factorization is proposed,

\[
\exp(iLt) = \prod_{l_m=1}^{n_m} \prod_{l_f=1}^{n_f} \exp \left( iL_N \frac{\Delta t}{2} \right) \exp \left( iL_{v(l_f,l_m)} \frac{\delta t}{2} \right) \exp \left( iL_q \frac{\delta t}{2} \right) \exp(iLOU\delta t) \exp \left( iL_q \frac{\delta t}{2} \right) \\
\exp \left( iL_{v(l_f,l_m)} \frac{\delta t}{2} \right) \exp \left( iL_N \frac{\Delta t}{2} \right),
\]

(3.5.4)

where \( iL_{v(l_f,l_m)} \) is given by

\[
iL_{v(l_f,l_m)} = \begin{cases} 
    iL_{v,f} + n_f iL_{v,m} + n_m n_f iL_{v,s} & \text{if } l_f = n_f, l_m = n_m \\
    iL_{v,f} + n_f iL_{v,m} & \text{if } l_f = n_f, l_m < n_m \\
    iL_{v,f} & \text{otherwise}
\end{cases}
\]

(3.5.5)

In the scheme 3.5.4, the \( iL_N \) operator couples the auxiliary variables with the system at every inner time step \( \delta t \), and this scheme is referred to as XI-RESPA.

The XM-RESPA integration scheme evaluates the \( iL_N \) operator at every intermediate time step \( \Delta t \),

\[
\exp(iLt) = \prod_{l_m=1}^{n_m} \left\{ \exp \left( iL_N \frac{\Delta t}{2} \right) \left[ \prod_{l_f=1}^{n_f} \exp \left( iL_{v(l_f,l_m)} \frac{\delta t}{2} \right) \right] \exp \left( iL_q \frac{\delta t}{2} \right) \exp(iLOU\delta t) \exp \left( iL_q \frac{\delta t}{2} \right) \exp \left( iL_{v(l_f,l_m)} \frac{\delta t}{2} \right) \exp \left( iL_N \frac{\Delta t}{2} \right) \right\}. 
\]

(3.5.6)

It is also possible to evaluate the \( iL_N \) operator at every long time step \( t \), which is called
the XO-RESPA scheme, and has the form

\[ \exp(iLt) = \exp\left(iL_{N}t\right) \left[ \prod_{l_{m}=1}^{n_{m}} \prod_{l_{f}=1}^{n_{f}} \exp\left(iL_{v}^{(l_{f},l_{m})}\delta t/2\right) \right. \]

\[ \exp\left(iL_{q}\delta t/2\right) \exp(iL_{OU}\delta t) \exp\left(iL_{q}\delta t/2\right) \]

\[ \left. \exp\left(iL_{v}^{(l_{f},l_{m})}\delta t/2\right) \right] \exp\left(iL_{N}t/2\right). \]  \hspace{1cm} (3.5.7)

Because the two operators \(\exp(iL_{OU}t)\) and \(\exp(iL_{q}t)\) commute,

\[ \exp\left(iL_{q}t/2\right) \exp(iL_{OU}t) \exp\left(iL_{q}t/2\right) = \exp(iL_{OU}t) \exp(iL_{q}t), \]  \hspace{1cm} (3.5.8)

eq. (3.5.8) can be applied to eqs. (3.5.4), (3.5.6) and (3.5.7).

### 3.6 Short-Range Forces and Long-Range Forces

The energy terms in the AMOEBA force field are grouped into four categories in the multiple time step integrator: (1) long-range electrostatics, long-range polarization, and long-range vdw; (2) short-range electrostatics, short-range polarization, and short-range vdw; (3) torsion terms; and (4) the remaining valence terms, e.g. bond, angle, etc. Different time steps are used for different categories in the integrator. The associated time steps for the four groups of forces are denoted as \(t\), \(t_{s}\), \(t_{t}\), and \(t_{b}\), respectively, where the ratios \(n_{s} = t/t_{s}\), \(n_{t} = t_{s}/t_{t}\), and \(n_{b} = t_{t}/t_{b}\), are all positive integers.

#### 3.6.1 integration schemes

Since the total force is written as the sum of four parts,

\[ F = F_{b} + F_{t} + F_{s} + F_{l}, \]  \hspace{1cm} (3.6.1)
the Lagrange multiplier $\lambda_F$ and the $iL_v$ operator can also be written as the sum of the corresponding components,

$$\lambda_F = \lambda_{F,b} + \lambda_{F,t} + \lambda_{F,s} + \lambda_{F,l}.$$  \hfill (3.6.2)

$$iL_v = iL_{v,b} + iL_{v,t} + iL_{v,s} + iL_{v,l}.$$  \hfill (3.6.3)

The XI-RESPA scheme is

\[
\exp(iLt) = \prod_{l_s=1}^{n_s} \prod_{l_t=1}^{n_t} \prod_{l_b=1}^{n_b} \exp\left( iL_N \frac{t_b}{2} \right) \exp\left( iL_v(l_b,l_t,l_s) \frac{t_b}{2} \right) \\
\exp\left( iL_q \frac{t_b}{2} \right) \exp(iLOU t_b) \exp\left( iL_q \frac{t_b}{2} \right) \\
\exp\left( iL_v(l_b,l_t,l_s) \frac{t_b}{2} \right) \exp\left( iL_N \frac{t_b}{2} \right), \]  \hfill (3.6.4)

where

\[
iL_v(l_b,l_t,l_s) = \begin{cases} 
    iL_{v,b} + n_b iL_{v,t} + n_t n_b iL_{v,s} + n_s n_t n_b iL_{v,l} & \text{if } l_b = n_b, l_t = n_t, l_s = n_s \\
    iL_{v,b} + n_b iL_{v,t} + n_t n_b iL_{v,s} & \text{if } l_b = n_b, l_t = n_t, l_s < n_s \\
    iL_{v,b} + n_b iL_{v,t} & \text{if } l_b = n_b, l_t < n_t, l_s < n_s \\
    iL_{v,b} & \text{otherwise} 
\end{cases} \]  \hfill (3.6.5)

In the XM-RESPA scheme, the $iL_N$ operator is evaluated at the short-range time step.

\[
\exp(iLt) = \prod_{l_s=1}^{n_s} \left\{ \exp\left( iL_N \frac{t_s}{2} \right) \left[ \prod_{l_t=1}^{n_t} \prod_{l_b=1}^{n_b} \exp\left( iL_v(l_b,l_t,l_s) \frac{t_b}{2} \right) \right] \right. \\
\exp\left( iL_q \frac{t_b}{2} \right) \exp(iLOU t_b) \exp\left( iL_q \frac{t_b}{2} \right) \\
\left. \exp\left( iL_v(l_b,l_t,l_s) \frac{t_b}{2} \right) \exp\left( iL_N \frac{t_s}{2} \right) \right\}. \]  \hfill (3.6.6)
Finally, the XO-RESPA scheme is
\[
\exp(iLt) = \exp \left( iL_{N} \frac{t}{2} \right) \left[ \prod_{l_{s}=1}^{n_{s}} \prod_{l_{t}=1}^{n_{t}} \prod_{l_{b}=1}^{n_{b}} \exp \left( iL_{v}^{(l_{b},l_{t},l_{s})} \frac{t_{b}}{2} \right) \right. \\
\left. \exp \left( iL_{q} t_{b} \frac{t_{b}}{2} \right) \exp(iL_{OUT} t_{b}) \exp \left( iL_{q} t_{b} \frac{t_{b}}{2} \right) \exp \left( iL_{N} t_{b} \frac{t_{b}}{2} \right) \right] \exp \left( iL_{N} t \right). \tag{3.6.7} \]

An implementation of the XI-RESPA, XM-RESPA, and XO-RESPA algorithms is listed in program 3.1 and has been incorporated into the Tinker software.

### 3.6.2 short-range and long-range forces

The separation of short-range vdw and long-range vdw is straightforward. For the total vdw interactions within a system, it is the sum of all the pairwise vdw interactions within the cutoff distance \( R_{vdw} \). The short-range vdw energy is simply all the pairwise vdw interactions within the cutoff distance \( r_{vdw} \), where \( r_{vdw} \leq R_{vdw} \). Then the long-range vdw energy is naturally defined by the difference between the total vdw energy and the short-range vdw energy. In the limit where two cutoffs, \( r_{vdw} \) and \( R_{vdw} \), are equivalent, the long-range vdw interaction is 0.

Regardless of the definitions of short-range electrostatics and polarization energy, the definitions for the long-range electrostatics and polarization energy are always taken as the difference between the total energy and the corresponding short-range energy. One definition, RESPA2, for the short-range electrostatics, contains all the pairwise permanent multipole interactions within the short-range cutoff \( r_{ede} \). The corresponding short-range polarization energy consists of all the interactions of the induced dipoles due to the short-range permanent electrostatic field.

The alternative RESPA1 definition is specially designed for the Ewald and PME algorithms. These algorithms reduce total computational expense to order \( O(N \log N) \)
instead of the naive pairwise $O(N^2)$ complexity, and also successfully minimize the applicable cutoff distance.

Consider a system with no bonds and under periodic boundary conditions. The total permanent electrostatic energy given by the permanent multipoles $\{\hat{L}_i\}$ (for the details of the notation, see [106] by Sagui et al.) can be written as the sum of three terms,

$$U_{ele} = U_{self} + U_{dir} + U_{rec}, \quad (3.6.8)$$

where the self-energy

$$U_{self} = -\frac{\alpha}{\sqrt{\pi}} \sum_{i} \hat{L}_i^2 \quad (3.6.9)$$

is not a function of coordinates, therefore the self-energy is not related to any forces of the system and it does not matter, if not considering the expense of the computation, whether the self-energy is calculated in the short-range steps or long-range steps. The direct real space electrostatic potential at position $r_i$ is given by

$$\varphi_{dir}(r_i) = \sum_{n \neq 0} \sum_{A} \hat{L}_i \hat{L}_j \text{erfc}(\alpha |r_j - r_i + n|) \frac{|r_j - r_i + n|}{|r_j - r_i + n|}, \quad (3.6.10)$$

where $A$ is the collection of all of the atom pairs where $i \neq j$. The direct real space electrostatics is given by

$$U_{dir} = \frac{1}{2} \sum_{n \neq 0} \sum_{A} \hat{L}_i \hat{L}_j \text{erfc}(\alpha |r_j - r_i + n|) \frac{|r_j - r_i + n|}{|r_j - r_i + n|}. \quad (3.6.11)$$

The reciprocal space PME electrostatics is given by

$$U_{rec} = \frac{1}{2\pi V} \sum_{m \neq 0} |S(m)|^2 \exp(-\pi^2 m^2 / \alpha^2). \quad (3.6.12)$$

In practice, however, most simulation systems contain multiple bonds, angles, etc., where the interactions between these atom pairs (denoted as $B$, bonded) should be excluded. Therefore, a fourth term is introduced as a correction term

$$U_{corr} = -\sum_{B} \hat{L}_i \hat{L}_j \frac{1}{|r_j - r_i|}. \quad (3.6.13)$$
There is a corner case associated with eq. (3.6.13) in the AMOEBA force field. Consider a chain of atoms A-B-C-D-E, where the interactions between A-D, B-E, and A-E are not scaled to 0, due to the fact that the global scalars \texttt{mpole-14-scale} and \texttt{mpole-15-scale} are set to 0.4 and 0.8, respectively, then the correction term becomes

$$U_{\text{corr}} = - \sum_{B} \hat{L}_{i} \hat{L}_{j} \frac{1 - c_{ij}}{|r_{j} - r_{i}|},$$

(3.6.14)

where $c_{ij}$ is a scalar for atom pair $i$ and $j$.

The short-range electrostatics in the RESPA1 scheme consist of the direct terms and the correction terms within the short-range cutoff $r_{c}^{\text{ele}}$. The short-range polarization energy for RESPA1 should be consistent with the electrostatic field corresponding to short-range electrostatics.

### 3.6.3 force damping

Energy cutoffs unavoidably cause discontinuous forces. The Ewald summation has a similar problem in eq. (3.6.11), $U_{\text{dir}}$, because it is impossible to directly iterate the entire infinite $n$ space, a cutoff distance $R_{c}^{\text{ele}}$ has to be selected, and the coefficient $\alpha$ in eq. (3.6.11), known as the Ewald coefficient, must be large enough such that the value of $\text{erfc}(\alpha R_{c}^{\text{ele}})$ is small (e.g. less than $10^{-6}$). Similar to the Ewald summation, a fifth-order switching function $S(r; r_{0}, r_{1})$ is used to scale the energy beginning at $r_{0}$, and to completely vanish the interaction at $r_{1}$,

$$S(r; r_{0}, r_{1}) = \begin{cases} 
1, & r \leq r_{0} \\
-6u^{5} + 15u^{4} - 10u^{3} + 1, & r_{0} < r < r_{1} \\
0, & r_{1} \leq r
\end{cases},$$

(3.6.15)

where

$$u = \frac{r - r_{0}}{r_{1} - r_{0}}.$$

(3.6.16)
When applying the switching function on forces, there are two possible implementations: (1) directly multiply the undamped force by the switching function (fig. 3.2), or (2) use the chain rule to calculate the force in the switching window (fig. 3.3). For the first implementation, it is important to emphasize that the potential energy calculated in the simulation becomes incompatible with the force used in the equations of motion. For the second, the derivative of the switching function on the switching window is given by

\[
\frac{d}{dr} S(r; r_0, r_1) = -\frac{30}{r_1 - r_0} u^2 (u - 1)^2,
\]

which will produce an extra chain rule force term.

### 3.7 Results and Discussion

The generalized stochastic isokinetic integrator was implemented in the Tinker software. In the following discussion of results in this section, “dof”, “atom”, and “global” will be used to denote isokinetic constraints coupled to every degree of freedom, to every atom, and to the whole system, respectively. All of the long-range PME cutoffs for electrostatics and polarization were set to 7.0 Å, and all of the long-range vdw cutoffs were set to 12.0 Å. In the RESPA1 scheme, all the short-range cutoff windows for nonbonded interactions were from 5.0 Å to 5.5 Å. In the RESPA2 scheme, the cutoff window for vdw (4.5 Å to 5.0 Å) is different from the cutoff windows for the other two nonbonded terms (4.0 Å to 5.5 Å). The XM-RESPA scheme was used. The parameter \( P \) in the simulations was set to 2, unless stated otherwise. The reference NVT simulations (denoted by “regular”) were performed using the Tinker-OpenMM library [53], with a 2 fs time step RESPA integrator [32] and Bussi thermostat [34, 35]. Simulations were run in mixed-precision on NVIDIA GTX 970 and 1070 GPUs.
3.7.1 argon

512 argon atoms were placed in a 29.1304 Å cubic box and the isokinetic constraints were held at an effective temperature of 94.4 K. The outer time step for long-range vdw interactions was set to 30 fs, and the time step for the short-range vdw was set to 3 fs. Each simulation was propagated for 500 fs. The parameters for the argon atom were adopted from the AMOEBA09 parameter set. Nine combinations of parameters were tested for this system and were compared to the results of the regular MD simulation.

The isokinetic conditions and simulation results are tabulated in table 3.1.
shows that the RDF of argon is consistently reproducible via isokinetic simulations. The difference between the RDF sampled from regular simulation and isokinetic simulations is negligible. Even with the most aggressive settings where the short-range cutoff is only 5 Å, and the force scalar is set to $4/3$, the RDF is still well preserved.

The dynamics of the isokinetic simulations, not surprisingly, are perturbed. From the construction of the isokinetic constraint, the kinetic energy of the system is always less than $M k_B T/2$. If the kinetic energy is much less than $dn k_B T/2$, a cooler system will be sampled. Even if the kinetic energy of the system is close to $dn k_B T/2$, the lower diffusion
coefficients in the isokinetic simulations (b, e, and i) suggest a stronger perturbation of the dynamics and therefore a smaller autocorrelation of the velocities.

3.7.2 water

512 AMOEBA03 water molecules [25] were placed in a 24.8322 Å cubic box and the isokinetic constraints were held at an effective temperature of 298 K.

The “global” isokinetic constraints were tested in combination with the RESPA2 scheme. The results (table 3.2 and fig. 3.5) support that the 2 fs outer time step quantita-
Table 3.1: Results of regular simulation and isokinetic integrator simulations with various conditions for liquid argon. The short and long-range cutoffs for vdw are in Å. The kinetic energies (KE) are given in Kelvin. The diffusion coefficients $D$ are given in $10^{-5}$ cm$^2$/s.

<table>
<thead>
<tr>
<th>#</th>
<th>Method</th>
<th>$L$</th>
<th>$M$</th>
<th>Cutoff</th>
<th>$s$</th>
<th>KE</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>regular</td>
<td>-</td>
<td>-</td>
<td>12/12</td>
<td>1</td>
<td>94.4</td>
<td>2.14</td>
</tr>
<tr>
<td>(a)</td>
<td>dof</td>
<td>1</td>
<td>1</td>
<td>12/12</td>
<td>1</td>
<td>47.5</td>
<td>0.52</td>
</tr>
<tr>
<td>(b)</td>
<td>dof</td>
<td>1</td>
<td>2</td>
<td>12/12</td>
<td>2</td>
<td>95.0</td>
<td>0.76</td>
</tr>
<tr>
<td>(c)</td>
<td>atom</td>
<td>1</td>
<td>3</td>
<td>12/12</td>
<td>1</td>
<td>71.2</td>
<td>1.38</td>
</tr>
<tr>
<td>(d)</td>
<td>atom</td>
<td>1</td>
<td>3.7</td>
<td>12/12</td>
<td>3.7/3</td>
<td>87.9</td>
<td>1.50</td>
</tr>
<tr>
<td>(e)</td>
<td>atom</td>
<td>1</td>
<td>4</td>
<td>12/12</td>
<td>4/3</td>
<td>95.0</td>
<td>1.70</td>
</tr>
<tr>
<td>(f)</td>
<td>atom</td>
<td>1</td>
<td>3</td>
<td>5/12</td>
<td>1</td>
<td>71.2</td>
<td>1.38</td>
</tr>
<tr>
<td>(g)</td>
<td>atom</td>
<td>1</td>
<td>4</td>
<td>5/12</td>
<td>4/3</td>
<td>95.0</td>
<td>1.59</td>
</tr>
<tr>
<td>(h)</td>
<td>atom</td>
<td>0</td>
<td>2</td>
<td>12/12</td>
<td>1</td>
<td>63.3</td>
<td>1.35</td>
</tr>
<tr>
<td>(i)</td>
<td>atom</td>
<td>0</td>
<td>3</td>
<td>12/12</td>
<td>3/2</td>
<td>95.0</td>
<td>1.73</td>
</tr>
</tbody>
</table>

Figure 3.4: RDFs of liquid argon.
tively reproduces the sampling of configurations and dynamics. The perturbation to the
dynamics becomes more obvious with increasing $L$ and $M$ parameters. A larger outer
time step does not accurately reproduce the oxygen-oxygen RDF of water.

Table 3.2: Simulation results of the “global” isokinetic integrators with the RESPA2
scheme. The short and long-range time steps for the nonbonded interactions are in fs.
Kinetic energies (KE) are in Kelvin. Diffusion coefficients ($D$) are in $10^{-5}$ cm$^2$/s.

<table>
<thead>
<tr>
<th>#</th>
<th>Time Step</th>
<th>KE</th>
<th>$D$</th>
<th>$L$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>regular</td>
<td>298.2</td>
<td>2.03</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td>2/2</td>
<td>298.1</td>
<td>1.92</td>
<td>0</td>
<td>5607</td>
</tr>
<tr>
<td>(b)</td>
<td>2/2</td>
<td>298.1</td>
<td>1.86</td>
<td>1</td>
<td>4608</td>
</tr>
<tr>
<td>(c)</td>
<td>2/2</td>
<td>297.6</td>
<td>1.92</td>
<td>768</td>
<td>5375</td>
</tr>
<tr>
<td>(d)</td>
<td>2/2</td>
<td>297.2</td>
<td>1.83</td>
<td>1536</td>
<td>6143</td>
</tr>
<tr>
<td>(e)</td>
<td>2/4</td>
<td>298.1</td>
<td>1.38</td>
<td>768</td>
<td>5375</td>
</tr>
<tr>
<td>(f)</td>
<td>2/8</td>
<td>298.5</td>
<td>0.0003</td>
<td>768</td>
<td>5375</td>
</tr>
<tr>
<td>(g)</td>
<td>2/10</td>
<td>306.2</td>
<td>0.0002</td>
<td>768</td>
<td>5375</td>
</tr>
<tr>
<td>(h)</td>
<td>2/8</td>
<td>298.1</td>
<td>0.0004</td>
<td>1</td>
<td>4608</td>
</tr>
</tbody>
</table>

For the combinations of the RESPA1 scheme and “dof” or “atom” isokinetic integra-
tors, RDFs sampled from “atom” with $L = 1$, $M = 3$ and “dof” with $L = 1$, $M = 1$,
both with 120 fs outer time step, are compared to the RDF from the regular simulation
(fig. 3.6). Note that a large outer time step of 120 fs with the RESPA1 scheme will generate
a satisfactory RDF.

The rate of convergence of sampling was measured by the evolution of the $L^1$ norm
of the property of interest. The $L^1$ norm $\xi$ of property $p$ at time $t$ from a final property $p_f$
is given by

$$\xi(p, t) = |p(t) - p_f|.$$  \hspace{1cm} (3.7.1)

If $p$ is an RDF, the $L^1$ norm can be evaluated by

$$\xi(t) = \frac{1}{N_b} \sum_{i=1}^{N_b} |p(x_i, t) - p_f(x_i)|.$$  \hspace{1cm} (3.7.2)
Figure 3.5: RDFs of liquid water oxygens sampled with the “global” isokinetic constraints. (a)-(h) refer to the simulation conditions described in table 3.2.
where $N_b$ is the number of bins used to discretize an RDF. Figure 3.7 shows that compared to the regular simulation, the “dof” and “atom” isokinetic integrators converge to the same RDF at faster rates.

### 3.7.3 n-hexane

Systems with 267 n-hexane molecules were simulated in a 38.0 Å cubic box with the RESPA1 scheme and the isokinetic constraints targeting 298 K. Parameters were adopted from the AMOEBA09 parameter set. A 1 fs time step was used for torsions, with 3 fs time steps used for short-range vdw, electrostatics, and polarization. A 120 fs time step
was used for long-range vdw, electrostatics, and polarization, and a 0.5 fs time step was used for all the other valence terms included in the simulations. For “dof” isokinetic constraints, $L = 1$ and $M = 1$. For “atom” isokinetic constraints, $L = 1$ and $M = 3$. The “dof” and “atom” results were compared to the results of regular simulations.

Three dihedral angles ($\varphi_1$, $\varphi_2$, and $\varphi_3$) of n-hexane are defined in fig. 3.8. All three dihedral angles in the starting structures of the simulations were set to $\pm 180$ degrees, resulting in all molecules have an extended conformation.

In fig. 3.9, the sampled distributions of dihedral angles $\varphi_1$, $\varphi_2$, and $\varphi_3$ of n-hexane using the “dof” or “atom” isokinetic constraint with a 120 fs time step do not show any
loss of accuracy compared to the 2 fs time step regular simulations. The $L^1$ norm of the distributions of the dihedral angles converges to the same value after 500 ps simulation, as shown in fig. 3.10. The convergence rate of “dof” and “atom” isokinetic constraints, however, is slower than the convergence rate of the regular simulation. The convergence rate of the “atom” method is systematically faster than the “dof” method in the case of n-hexane.

### 3.7.4 free energy calculations

The generalized isokinetic constraint integrator was also tested in solvation free energy calculations. FEP [44] calculations were first completed via regular 2 fs simulation using the Tinker-OpenMM library or by the isokinetic constraint integrators, then were analyzed by the Bennett Acceptance Ratio method [45] using the bar program in the Tinker package.

Three molecules, a NaCl ion pair, ethanol, and acridine orange (see fig. 3.11), were inserted into a cubic box filled with the AMOEBA03 water molecules. The “atom” isokinetic constraint with the RESPA2 scheme was used in 1 ns simulations with $L = 1$ and $M = 3.7$. The outer time step for the long-range interactions was 30 fs. The time step for the short-range vdw, electrostatics, and polarization was 6 fs. All the valence interactions were integrated with 0.5 fs time step.
Figure 3.9: Distributions of dihedral angles $\varphi_1$, $\varphi_2$, and $\varphi_3$ of liquid n-hexane in equilibrium.
Figure 3.10: $L^1$ norms of the distributions of dihedral angles $\varphi_1$, $\varphi_2$, and $\varphi_3$ of liquid $n$-hexane in 500 ps simulations.
Acridine orange was selected as a test example because it was a ligand in the SAMPL5 challenge \cite{107}. The parameters and simulated solvation free energy in the AMOEBA force field of acridine orange were available. Although the molecule is relatively large compared to ethanol, its flat and rigid structure does not introduce much sampling complexity.

![Figure 3.11: Structure of acridine orange.](image)

As tabulated in table 3.3, even with the RESPA2 scheme and an aggressive 6 fs intermediate time step, the solvation free energy of NaCl and ethanol are in good agreement with the solvation free energies calculated by the regular simulations. The solvation energy of a larger molecule, acridine orange, however, is overestimated by 3.6 kcal/mol, suggesting that the combination of RESPA2 scheme and a 6 fs intermediate time step is too ambitious to be widely applied.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Regular</th>
<th>“atom”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaCl</td>
<td>-176.7 ± 0.2</td>
<td>-177.0 ± 0.2</td>
</tr>
<tr>
<td>ethanol</td>
<td>-5.9 ± 0.1</td>
<td>-5.8 ± 0.1</td>
</tr>
<tr>
<td>acridine orange</td>
<td>-16.1 ± 0.1</td>
<td>-12.5 ± 0.3</td>
</tr>
</tbody>
</table>

A RESPA1 scheme and finer intermediate time steps were tested for the solvation free energy of acridine orange using the “atom” isokinetic constraints. In table 3.4, the outer time step can safely and accurately be set to 4 fs, and can be as large as 40 fs, while still yielding a solvation free energy within 1 kcal/mol. Larger outer time steps, as expected, increase the errors in the prediction.
Table 3.4: Solvation free energies (in kcal/mol) of acridine orange calculated by regular 2 fs simulations and various “atom” isokinetic integrators from 1.5 ns simulations. The outer time steps are in fs. All the intermediate time steps for short-range nonbonded interactions are 2 fs.

<table>
<thead>
<tr>
<th>Solvation Free Energy</th>
<th>Outer Time Step</th>
<th>L</th>
<th>M</th>
<th>P</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16.11</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Regular MD</td>
</tr>
<tr>
<td>-16.37</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-16.30</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-15.46</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-15.60</td>
<td>20</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-15.29</td>
<td>40</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-14.89</td>
<td>80</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-15.43</td>
<td>20</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-15.02</td>
<td>20</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

3.8 Conclusion

The stochastic isokinetic constraint integrator is extended and generalized in three ways. First, the Nosé-Hoover chain algorithm was connected with the stochastic process in the equations of motion of the auxiliary velocities. Then the functional form of the isokinetic constraint is generalized such that the constrained kinetic energy is no longer required to be the an integer multiple of $k_B T/2$. This generalization enables greater flexibility and control in its application. Lastly, the isokinetic constraint was extended from one dimensional velocity space (“dof”) to any dimensionality of velocity space. The extension yields new forms of the isokinetic constraint integrator.

All of the variations of the generalized stochastic isokinetic constraints have been implemented in Tinker. Results have shown that the outer time step can be extended to 120 fs to enhance the efficiency of the conformational sampling. Compared to the original “dof” isokinetic constraints, the “atom” isokinetic integrator imposes less perturbation on the dynamics of the system and achieves equilibrium at an equivalent or a faster rate.

This method also shows its great potential in free energy calculations and deserves
further investigation.

3.9 Supplementary Information

The supplementary information starts on page 92.
Program 3.1: Stochastic Isokinetic Integrator

c language: Fortran
c tcc: the temporary storage for the accelerations
c
dt_2 = dt / 2

dts = dt / nl

dts_2 = dts / 2

dtb = dts / (nt * ns)

dtb_2 = dtb / 2

if (xo_respa) call exp_iLn (dt_2)
do jls = 1, ns
  if (xm_respa) call exp_iLn (dts_2)
do jlt = 1, nt
    do jlb = 1, nb
      if (xi_respa) call exp_iLn (dtb_2)
        call exp_iLv (dtb_2)
        call exp_iLou (dtb)
pos = pos + vel * dtb
call calc_acc_b_from_pos (acc,pos)
if (jlb .eq. nb) then
  call calc_acc_t_from_pos (tcc,pos)
  acc = acc + tcc * nb
if (jlt .eq. nt) then
  call calc_acc_short_from_pos (tcc,pos)
  acc = acc + tcc * (nb * nt)
if (jls .eq. ns) then
  call calc_acc_long_from_pos (tcc,pos)
  acc = acc + tcc * (nb * nt * ns)
end if
end if
end if

c
call exp_iLv (dtb_2)
if (xi_respa) call exp_iLn (dtb_2)
end do
end do
if (xm_respa) call exp_iLn (dts_2)
end do
if (xo_respa) call exp_iLn (dt_2)
call save_md_traj (pos,vel)
CHAPTER 4

BINDING FREE ENERGIES OF CUCURBIT[8]URIL AND GUESTS IN THE SAMPL6 CHALLENGE

4.1 Introduction

The binding affinities of small molecules and proteins has always been an important topic in both academic and industrial research, where computational chemistry is playing an increasingly important role. If the computational predictions of binding affinities are accurate with moderate computational expense, they will significantly reduce costs and cycles of synthesis and lab experiment in drug development. But as an ongoing research topic, protein-ligand systems are not the best examples to troubleshoot problems in methodology, especially for MD-based binding free energy calculations, because of the large size and complexity of the simulations. Therefore, host-guest systems are more often used as models for methodological development and as examples for proof of concept.

In order to evaluate protocols for predicting binding affinities in host-guest systems, as well as other thermodynamic properties, e.g. hydration free energies, distribution coefficients, pH values, etc., the computational chemistry community started the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) project [108] more than a decade ago. The most recent SAMPL6 challenge [109] started at the end of September 2017, and kept its traditional format as a blinded challenge. The participants in the SAMPL challenge must submit their binding predictions prior to the release of the previ-
ously unpublished experimental results. With only limited information available, e.g. the protocols of the experimental measurements, the SAMPL participants focus their efforts on methodological development instead of excessive tweak of parameters specific to the targeted properties.

One of the host molecules in the SAMPL6 challenge is cucurbit[8]uril (CB[8]) [110]. The CB[n] are macrocyclic methylene-bridged glycoluril oligomers containing n glycoluril units. CB[6], one of the first and most widely studied members of the CB[n] family, has high affinity for cationic organic molecules, especially complex ammonium ions, via ion-dipole interactions, hydrogen bonds, and hydrophobic effects. Later studies showed similar binding properties occur in CB[7] and CB[8]. CB[7] was one of the host molecules that appeared in the SAMPL4 challenge. This molecule is of special interest due to its relatively high water solubility.

Fourteen molecules were provided as guests for the CB[8] host in this challenge. The experimental binding free energies were collected by isothermal titration calorimetry with 20 mM sodium phosphate buffer at pH 7.4. The protonation states of the unbound guest molecules could be inferred from the pH value, but the protonation states of the bound guests were unknown. The coordination number as well as the pose of the bound host-guest structures were not provided. Participants also need to determine the preferred enantiomer resulting from any protonation process (e.g. CB[8]-G12). The structures used to initialize the simulations (not necessarily the correct binding conformation) are displayed starting on page 96. For clarity, the hydrogens in the molecular models are omitted.

For this challenge, we computed the binding free energies by means of molecular dynamics simulations and free energy perturbation theory with the AMOEBA force field. The thermodynamic cycle for the absolute binding free energy calculation used in the SAMPL6 challenge is demonstrated in fig. 4.17. Two independently calculated free ener-
gies are obtained initially. One path, A2-B2-B3-A3, represents the negative solvation free energy of the guest molecule in water. This process decouples the guest molecule from the environment by disappearing the electrostatics and the vdw of the guest molecule separately. The other path, A1-B1-C1-C4-B4-A4, is the free energy that couples the guest molecule to the solvent and the host molecule. Besides adding electrostatics (C4-B4) and vdw (C1-C4), an extra geometric restraint is added only in the middle of the thermodynamic path and without changing the thermodynamic states of the end points. Under the assumption that the thermodynamic states for B1, C1 and B4, C4 of the system sampled
Figure 4.2: CB[8]-G0 (Lexapro-H⁺) in CB[8].

Figure 4.3: CB[8]-G1 (Dentrol-H⁺) in CB[8].
Figure 4.4: CB[8]-G2 (Palonosetron-H⁺) in CB[8].

Figure 4.5: CB[8]-G3 (Quinine-H⁺) in CB[8].
Figure 4.6: CB[8]-G4 (Gallamine) in CB[8].

Figure 4.7: CB[8]-G5 (R-Bornylamine-H⁺) in CB[8].
Figure 4.8: CB[8]-G6 (Cycloheptylamine-H⁺) in CB[8].

Figure 4.9: CB[8]-G7 (Cyclooctylamine-H⁺) in CB[8].
Figure 4.10: CB[8]-G8 (Cyclododecylamine-H\(^+\)) in CB[8].

Figure 4.11: CB[8]-G9 (3-Noradamantanamine-H\(^+\) in CB[8].
Figure 4.12: CB[8]-G10 (3-Amino-1-adamantanol-H⁺) in CB[8].

Figure 4.13: CB[8]-G11 (trans-1,4-Diaminocyclohexane-2H⁺) in CB[8].
Figure 4.14: CB[8]-G12 (Aricept-H⁺).

(a) cis-CB[8]-G12-H⁺  
(b) trans-CB[8]-G12-H⁺

Figure 4.15: Two cis-Aricept-H⁺ in CB[8].
in the molecular dynamics simulations are the same, the binding free energy of the host-
guest system is the sum of the free energies of the two aforementioned thermodynamic
paths.

The binding free energy directly calculated by the method in fig. 4.17 is nominally
referred to as the absolute binding free energy. Instead of connecting the 0-0 state and
1-1 state of the guest molecule, the thermodynamic path of an alternative method com-
putes the relative binding free energy by connecting another similar guest molecule with
known binding free energy and the guest molecule of interest. With the same force field
and simulation protocols, the relative method usually gives a better prediction accuracy
due to cancellation of error.

A geometric restraint is added to the host-guest system to improve the sampling ef-
ficiency, by restricting the guest molecules to a small volume near the binding site, thus
avoiding the need for sampling a broader phase space. The geometric restraint will also
confine the samples and create greater overlap between the sampled distributions. In

Figure 4.16: CB[8]-G13 (Oxaliplatin) in CB[8].
practice, the free energy of the second thermodynamic path calculated by molecular dy-
namics only contains the C1-C4-B4-A4 portion of the free energy. The remaining compo-
nent of the second free energy path (B1-C1) can be evaluated analytically.

4.2 Parameterization of the Guest Molecules

In principle, the parameterization procedures for both host and guest molecules are the
same, but for the reasons mentioned in section 4.1, the CB[7] parameters used in SAMPL4
were directly adopted as the parameters of CB[8] to avoid the expensive quantum calculations for such large molecule. Even though some of the guest molecules had appeared in the previous SAMPL challenges, they were parameterized from first principles following the standard AMOEBA procedures [111].

4.2.1 atomic multipoles and polarizabilities

optimization

We initially optimize the molecule at the MP2/6-311G(1d,1p) level of theory. If the molecule is large and optimization takes too much CPU time, a B97-D/6-311G(1d,1p) preliminary optimization is used to minimize the energy of the molecule. If the molecule is flexible, additional optimized structures of the molecule in other low energy conformations are useful to generate a small library of favored structures.

atomic multipoles and polarizabilities

First, use the original DMA (distributed multipole analysis calculation) formulation [112] to obtain the atomic multipoles of the molecule. gdma_limit is set to 2 to generate the quadrupoles parameters. The gdma_switch is set to 0.0 to enforce the original DMA method and to avoid Stone’s newer spatial partitioning scheme [113]. The radius of hydrogen is set to 0.65 instead of the original value 0.325.

Next, the poledit program from the Tinker package is used to process the DMA output obtained from the previous step. The user is asked whether or not to change the default local frame definitions suggested by the program. The poledit program confirms the local coordinate frames of the atomic multipoles, atomic polarizabilities of the atoms, definitions of the intra-molecular polarization groups, whether or not to average the multipoles over the equivalent atoms, and whether or not to remove small numerical errors to enforce the symmetry in the local frame definitions. The default atomic
polarizabilities in AMOEBA are from Thole [24], except for the aromatic carbons (1.750) and hydrogen (0.696) attached to aromatic carbon. By default, the molecule is treated as one polarization group, and this default is generally suitable for conformationally rigid molecules. If the molecule is flexible, the polarization groups are determined based upon the major rotational bonds, to account for intra-molecular polarization effects in different conformations. In the end, the poledit program will remove the intra-molecular induced dipoles from raw DMA multipoles, adjust the dipole moments of the atoms and write out the atomic multipoles and polarizabilities in Tinker .key format.

Then, the atomic multipoles obtained from low-level quantum calculations in the previous step will be refined, based on the .fchk file from a Gaussian 09 [49] or Psi4 [114] MP2/aug-cc-pVTZ calculation on the minimized structure. Using the potential program in the Tinker package, a .grid file that contains a set of grid points outside the molecule in the Cartesian coordinates is created. Then the the Gaussian utility cubegen is run via:

```
$ cubegen 0 potential=df-mp2 molecule.fchk molecule.cube -5 h < molecule.grid
```

if the mp2 calculation is done in Psi4 or, if the .fchk file is calculated from Gaussian,

```
$ cubegen 0 potential=mp2 molecule.fchk molecule.cube -5 h < molecule.grid
```

is used to create a .cube file. The potential program will convert the .cube to a .pot file that contains the electrostatic potential for every grid point. The potential program is able to take the .pot file as the fitting target and to use the parameters generated from poledit as the initial values for a fitting process such that when the optimization exits, the RMS of the final gradient of the error function is less than 0.1. During the fitting, the charge components of the atomic multipoles are kept constant, which is enforced by adding the FIX-MONOPOLE keyword in the .key file.

Finally, atomic class numbers and type numbers are assigned to the parameterized atoms.
4.2.2 vdw parameters and valence parameters

Vdw parameters and the valence parameters are far more transferable than multipole parameters. Most of the vdw and valence parameters were directly adopted from the AMOEBA09 and AMOEBAPRO13 parameter sets, e.g., the vdw parameters for the nitrogen atom in the ammonium groups of the guests used the vdw parameters of the nitrogen of the positively charged Lysine side chain (\(-{(CH_2)_4NH_3^+}\)) in the AMOEBAPRO13 parameter set. If no proper analogy was available, parameters from the MMFF force field were used. The most important reason the MMFF force field was used is that MMFF adopts the buffered 14-7 vdw functional form, which is the same functional form used by the AMOEBA force field. If more precise adjustment of the vdw parameters is needed, high level quantum calculations of the dimer energy of a given structure can be used as a reference as well as a validation, similar to what was described in chapter 2. Another way of validating the vdw parameters is to compare simulation predictions to the macroscopic experimental properties, e.g. density, and heat of vaporization of the model molecule.

4.3 Calculation Protocols

The protonation states of the ammonium-based guests, bound and unbound, were assumed to be protonated under the experimental conditions. Both of the enantiomers of the protonated CB[8]-G12 were parameterized and calculated independently. The more tightly bound enantiomer in the calculations of CB[8]-G12 (cis) was reported.

To determine the parameters for the geometric restraints, an unrestrained host-guest molecular dynamics simulation was set up for 40 ns in a 40 Å cubic box and was held at 298 K and 1 atm via the Bussi thermostat \(^{[34, 35]}\) and Monte Carlo barostat \(^{[42]}\), respectively. PME with a real space cutoff at 7 Å was used. The induced dipoles were converged to \(1.0 \times 10^{-5}\) debye per atom RMS. The vdw interactions were cut off at 12 Å. The NpT
simulations were propagated by a RESPA integrator \[32\] with a 2 ns outer time step.

Two groups of atoms were also defined in the .key file, e.g. for CB[8]-G5, group 1 (31 atoms) consisted of all the atoms in the guest, and group 2 contained all the atoms of CB[8] (144 atoms). After the pre-equilibration simulations were complete, a geometric restraint was added to the .key file, e.g.

\begin{verbatim}
GROUP 1 -1 31
GROUP 2 -32 175
RERAINT-GROUPS 1 2 5.0 0 3.0
\end{verbatim}

which is a flat-bottomed parabolic potential where the range of the flat bottom region is from 0 to 3 Å and the force constant is 5 kcal/mol/Å\(^2\). Then use of the analyze program in the Tinker package with the -d flag yields the distance between centers of mass of the two groups in every saved trajectory frame. A sample plot of the distribution of the distance of the two centroids from the unrestrained simulation is shown in fig. 4.18. The flat bottom of geometric restraint should be adjusted to match the range of the distribution of centroids sampled from the unrestrained simulation.

It is safe to assert, according to fig. 4.18, that in the unrestrained simulation the distance of the centroids of two groups almost never exceeded 2.0 Å, therefore, this distance was used as the upper limit of the flat bottom of the restraint with 5 kcal/mol/Å\(^2\) for the force constant. This procedure guarantees the sampled B4 state in fig. 4.17 will always be the same as the A4 state that should have been sampled, because even without the geometric restraint in the simulation, the hypothetical restraint would never be violated.

Another important piece of information to be gathered from the long time unrestrained simulation is the preferred binding pose of the guest in the host molecule. If several binding modes are possible, a long unrestrained host-guest simulation should be used to determine whether or not one mode will bind, or will convert to another binding pose.
Figure 4.18: Distribution of the distance of the centroid of R-Bornylamine-H⁺ and the centroid of CB[8] in the unrestrained molecular dynamics simulation.

The organization of the solvation FEP simulations and host-guest FEP simulations is similar to the setup of the pre-equilibration simulation as well as the setup of the mean ionic activity calculations in chapter 2. Host-guest simulations add the restraint parameters determined from the pre-equilibration simulations. All simulations were 10 ns in duration with a 2 ns time step.

All dynamics simulations were calculated using the dynamic_omm program in the Tinker-OpenMM library [53]. The free energies were then computed by the bar_omm program in the Tinker-OpenMM. The first 10% of trajectory samples were discarded prior to the BAR calculations.
The analytical correction for the geometric restraint was proposed by Hamelberg and McCammon [115]. They have shown that the free energy difference between states C1 and B1 due to a geometric restraint energy surface \( u(r) \) is

\[
A_{\text{rstr}} = k_B T \ln(c^\circ V_I),
\]

where \( c^\circ \) is the unit concentration,

\[
V_I = \int e^{-\beta u(r)} \, dr.
\]

If the potential energy \( u \) is spherically symmetric, eq. (4.3.2) becomes

\[
V_I = \int_0^\infty e^{-\beta u(r)} 4\pi r^2 \, dr.
\]

In their examples, \( u(r) = kr^2 \), which gives an integral of \((\pi/\beta k)^{3/2}\). In Tinker, the following flat-bottomed parabolic potential was implemented:

\[
u(r) = \begin{cases} 
  k_1(r-r_1)^2 & 0 < r < r_1 \\
  0 & r_1 \leq r \leq r_2 \\
  k_2(r-r_2)^2 & r > r_2
\end{cases},
\]

and the corresponding integral \( V_I \) can be expressed as the sum of three independent integrals \( V_1, V_2, \) and \( V_3 \):

\[
V_1 = \int_0^{r_1} e^{-\beta k_1 (r-r_1)^2} 4\pi r^2 \, dr
\]

\[
= \left( 2\pi r_1^2 + \frac{\pi}{\beta k_1} \right) \sqrt{\frac{\pi}{\beta k_1}} \operatorname{erf} \left( \sqrt{\beta k_1 r_1} \right) + \left( 2r_1 e^{-\beta k_1 r_1^2} - 4r_1 \right) \frac{\pi}{\beta k_1},
\]

\[
V_2 = \int_{r_1}^{r_2} 4\pi r^2 \, dr = \frac{4}{3} \pi (r_2^3 - r_1^3),
\]

\[
V_3 = \int_{r_2}^{\infty} e^{-\beta k_2 (r-r_2)^2} 4\pi r^2 \, dr = \left( 2\pi r_2^2 + \frac{\pi}{\beta k_2} \right) \sqrt{\frac{\pi}{\beta k_2}} + 4r_2 \frac{\pi}{\beta k_2},
\]

where

\[
erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \, dt.
\]
4.4 Results and Discussion

4.4.1 submitted predictions

The initially submitted predictions for the binding free energies of CB[8] and guests are tabulated in table 4.1 and shown in fig. 4.19. The experimental binding free energies were measured as association equilibrium constants $K_a$, with a maximum uncertainty of 14%. Since $G = -RT \ln K_a$,

$$\delta G = RT \frac{\delta K_a}{K_a},$$

(4.4.1)

the maximum uncertainty of the binding free energy is less than 0.1 kcal/mol as measured at room temperature.

Table 4.1: Initial submission of the binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.

<table>
<thead>
<tr>
<th>Guest</th>
<th>Prediction</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-8.51</td>
<td>-6.69</td>
</tr>
<tr>
<td>1</td>
<td>-8.88</td>
<td>-7.65</td>
</tr>
<tr>
<td>2</td>
<td>-15.76</td>
<td>-7.66</td>
</tr>
<tr>
<td>3</td>
<td>-14.89</td>
<td>-6.45</td>
</tr>
<tr>
<td>4</td>
<td>-6.76</td>
<td>-7.80</td>
</tr>
<tr>
<td>5</td>
<td>-11.98</td>
<td>-8.18</td>
</tr>
<tr>
<td>6</td>
<td>-7.63</td>
<td>-8.34</td>
</tr>
<tr>
<td>7</td>
<td>-9.37</td>
<td>-9.98</td>
</tr>
<tr>
<td>8</td>
<td>-16.81</td>
<td>-13.50</td>
</tr>
<tr>
<td>9</td>
<td>-8.64</td>
<td>-8.68</td>
</tr>
<tr>
<td>10</td>
<td>-8.85</td>
<td>-8.22</td>
</tr>
<tr>
<td>11</td>
<td>-4.46</td>
<td>-7.77</td>
</tr>
<tr>
<td>12</td>
<td>-8.33</td>
<td>-7.05</td>
</tr>
<tr>
<td>13</td>
<td>-3.61</td>
<td>-7.11</td>
</tr>
</tbody>
</table>

More than half of the predictions fall in the gray shaded region on the plot, where the errors of the predicted binding free energies are less than 2 kcal/mol. The average absolute error of the prediction is 2.70 kcal/mol. The root mean square (RMS) error is 3.72 kcal/mol. The maximum statistical error in free energy calculation is 0.25 kcal/mol.
Figure 4.19: Initial submission of the binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.

In information provided with the SAMPL6 challenge, the organizers indicated that CB[8] binds with two G11 ligands, but our calculations suggested that only one G11 ligand would bind with the CB[8] host, which was later confirmed by experiment. The calculation also successfully predicted the CB[8] molecule would bind two G12 ligands in the experiment. The experimental results supported a 2:1 and 3:1 Host:Guest stoichiometry for G1 and G4, respectively, which the calculations failed to predict. Our FEP method is not directly capable of making such predictions, unless every possible binding stoichiometry is explicitly evaluated.

The two poorest predictions in the initially submitted set are G2 and G3, both of which
differed from the experiment by more than 8 kcal/mol. It suggests that in the methodology employed, there exists some underlying problem to which the G2 and G3 guests are sensitive.

4.4.2 CB[8] parameters

A common phenomenon observed in our initial host-guest simulations is the indentations of the CB[8] guests as demonstrated in figs. 4.21 and 4.22.

These indentations were not found in the crystal structures of CB[8], but they appeared stochastically during simulations even if the starting structure of CB[8] was almost perfectly circular (fig. 4.20). Comparing the structures in figs. 4.21 and 4.22 and the structures in fig. 4.20, the most obvious difference is that the N-(methylene C)-N-(carbonyl C) torsional angles have rotated from 117.8 degrees (circular) to -149.1 degrees (indented).
Figure 4.21: Molecular model of cucurbit[8]uril with a single indentation.

Figure 4.22: Molecular model of cucurbit[8]uril with two indentations.
The torsion parameters used initially are

atom 52 52 C "Cucurbituril C=O" 6 12.011 3
atom 53 53 N "Cucurbituril N" 7 14.007 3
atom 56 56 C "Cucurbituril CH2" 6 12.011 4
torsion 52 53 56 53 0.400 0.0 1 -0.400 180.0 2 -0.250 0.0 3

which is an up-to-3-fold torsional potential. Plotting eq. (1.1.2) with the parameters above in fig. 4.23 and with an alternative 3-fold torsional parameter value of -1.60, the change in the curvature of the total torsional potential is dramatic. The energy barrier between the circular and indented forms is much higher for the new 3-fold torsional parameter.

Figure 4.23: Torsional potential of N-(methylene C)-N-(carbonyl C) in CB[8], with the 3-fold torsional parameter being -0.25 or -1.60. The 1-fold and 2-fold torsional parameters are 0.4 and -0.4, respectively.
The new 3-fold torsional parameter of -1.60 was validated in two ways. First, the single point energies of the circular CB[8], the single indented CB[8], and the double indented CB[8] were calculated in Gaussian 09 at the ωB97X-D/6-311(1d,1p) level of theory, and are tabulated in table 4.2. The energy difference between the circular and single indented CB[7] was also determined as a reference. Varying the 3-fold torsional parameter in the AMOEBA force field, the corresponding potential energy differences were calculated and tabulated in table 4.3. The optimal 3-fold torsional parameter was determined to be approximately -1.70.

Table 4.2: Single point energies (in hartree) and the energy differences (in kcal/mol) of CB[7] and CB[8] and their indented structures from ωB97X-D/6-311G(1d,1p) level of theory. CB[n]-0 denotes the circular structure, CB[n]-1 and CB[n]-2 denote the single and the double indentations in the structure, respectively.

<table>
<thead>
<tr>
<th>Structure</th>
<th>E_h (hartree)</th>
<th>ΔE_h (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB[7]-0</td>
<td>4212.049148</td>
<td>-</td>
</tr>
<tr>
<td>CB[7]-1</td>
<td>4212.015864</td>
<td>20.886</td>
</tr>
<tr>
<td>CB[8]-0</td>
<td>4813.765471</td>
<td>-</td>
</tr>
<tr>
<td>CB[8]-1</td>
<td>4813.743307</td>
<td>13.908</td>
</tr>
<tr>
<td>CB[8]-2</td>
<td>4813.728269</td>
<td>23.345</td>
</tr>
</tbody>
</table>

Table 4.3: Energy differences (in kcal/mol) of the circular and indented CB[7] and CB[8] in AMOEBA with different torsional parameters.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Torsion -0.25</th>
<th>Torsion -1.50</th>
<th>Torsion -1.60</th>
<th>Torsion -1.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB[7]-0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CB[7]-1</td>
<td>12.4941</td>
<td>19.9074</td>
<td>20.4631</td>
<td>21.0108</td>
</tr>
<tr>
<td>CB[8]-0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CB[8]-1</td>
<td>6.8366</td>
<td>12.9949</td>
<td>13.4528</td>
<td>13.9060</td>
</tr>
<tr>
<td>CB[8]-2</td>
<td>11.4125</td>
<td>22.0370</td>
<td>22.8295</td>
<td>23.6140</td>
</tr>
</tbody>
</table>

Then, three elliptical CB[8] structures were created from the circular CB[8] structure by minimizing the potential energy with different geometric restraints added to the molecule, generating three structures of different eccentricity. Their single point energies
were again calculated with the $\omega$B97X-D/6-311G(1d,1p) level of theory and different 3-fold torsional parameters in the AMOEBA force field to obtain the potential energy differences (see table 4.4). From this test, it can be concluded that -1.60 gives a relatively optimum description of the conformational energetics. Although for structure E3 the force field cannot accurately reproduce the results from the quantum calculation, this is not a problem since the force field results are not severely in error and E3 is of extremely high energy and essentially never sampled in room temperature molecular dynamics simulations.

Table 4.4: Energy difference (in kcal/mol) between the elliptical CB[8] and the circular CB[8] from $\omega$B97X-D/6-311G(1d,1p) level of theory, and different 3-fold torsional parameters in AMOEBA force field. The column “R” denotes the length (in Å) of the longest axis of the CB[8] molecule.

<table>
<thead>
<tr>
<th>Structure</th>
<th>R</th>
<th>QM</th>
<th>-0.25</th>
<th>-1.50</th>
<th>-1.70</th>
<th>-2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle</td>
<td>13.0965</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E2</td>
<td>16.9464</td>
<td>65.610</td>
<td>46.829</td>
<td>64.148</td>
<td>66.913</td>
<td>74.531</td>
</tr>
<tr>
<td>E3</td>
<td>18.3994</td>
<td>233.613</td>
<td>186.438</td>
<td>211.565</td>
<td>215.579</td>
<td>226.632</td>
</tr>
</tbody>
</table>

4.4.3 key torsions and intra-molecular vdw

In fig. 4.17 the states B1 (or the equivalent state B3 for the guest) and C1 are the “0-0” states in the FEP dynamics simulations, meaning both the vdw and the electrostatics are decoupled from the host and the environment. However, the vdw and the electrostatics are treated differently. The electrostatics are annihilated such that both inter- and intra-molecular electrostatics are scaled at the same time. The intra-molecular vdw in our original procedure is not scaled (or annihilated), only decoupled from the host and the environment. This procedure does not cause any problems in most cases, but in CB[8]-G2, two completely different conformations of the guest were observed in state B3 (solvation 0-0 state) and C1 (host-guest 0-0 state with restraint) (see fig. 4.27), where the average

117
Figure 4.24: Molecular model of cucurbit[8]uril structure E1.

Figure 4.25: Molecular model of cucurbit[8]uril structure E2.
potential energy difference of the two conformations was 8 kcal/mol. Interconversion between these two conformations was not observed in the simulations. This is the evidence that the geometric restraint may bias the conformational sampling of the 0-0 states.

In the example of CB[8]-G2, the interconversion of the two 0-0 states was hindered due to incomplete sampling of the inter-ring rotatable bond of the amide nitrogen, due to the torsional potential itself, as well as the intra-molecular vdw potential. To remove these potential barriers, the computer code in the Tinker and Tinker-OpenMM packages was modified to incorporate annihilation of the key torsions and the intra-molecular vdw.

Three new keywords were added to the Tinker and Tinker-OpenMM packages. Adding the keyword `VDW-ANNIHILATE` in the `.key` file will trigger the code path implementing vdw annihilation. `ROTATABLE-BOND` and `TORS-LAMBDA` are the two new keywords to manipulate the key torsions. E.g.,

![Molecular model of cucurbit[8]uril structure E3.](image)
Figure 4.27: Two 0-0 states of Palonosetron-H⁺ sampled in the simulation. Left: from the solvation calculation. Right: from the host-guest calculation.

<table>
<thead>
<tr>
<th>ROTATABLE-BOND</th>
<th>19 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>TORS-LAMBDA</td>
<td>0.600</td>
</tr>
<tr>
<td>VDW-LAMBDA</td>
<td>0.600</td>
</tr>
<tr>
<td>ELE-LAMBDA</td>
<td>0.000</td>
</tr>
</tbody>
</table>

is part of the setup files for CB[8]-G2 of state 0.6-0. This would scale all the torsions across the bond connecting atoms 19 and 20 by 60%.

The torsional angles of (methine C)-C-(amide N)-(carbonyl C) of CB[8]-G2 at the 0-0 state from a 100 ps simulation are demonstrated in figs. 4.28 to 4.30. These simulations started from the two locked conformations, each adopting one of the four combinations of key torsions and vdw annihilation. As shown in fig. 4.28, scaling only the key torsion in CB[8]-G2 still resulted in biased the sampling of 0-0 states. The vdw annihilation is critical to aiding conformational sampling. Figure 4.29 indicates that scaling the key torsion further enhances the sampling of that torsional angle when vdw annihilation is also incorporated.
Figure 4.28: Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H⁺ in a water box simulated from the standard 0-0 state simulations of 100 ps. Intra-molecular vdw interactions were not annihilated. (a) starting from the solvation 0-0 state; (b) starting from the solvation 0-0 state with the key torsion being scaled to 0; (c) starting from the host-guest 0-0 state bound in CB[8]; (d) starting from the host-guest 0-0 state bound in CB[8] with the key torsion being scaled to 0.

4.4.4 revised predictions

The binding free energies of CB[8] and guests were re-computed with the change in the torsional parameters of CB[8], as well as the annihilation of the key torsional angles of the guests and intra-molecular vdw. The results are tabulated in table 4.5 and plotted in fig. 4.31. The average absolute error of the prediction is 1.28 kcal/mol. The root mean square (RMS) error is 1.73 kcal/mol. The maximum statistical error in free energy calculation is 0.25 kcal/mol. The revised binding free energy predictions are systematically
Figure 4.29: Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H⁺ in a water box simulated from the standard 0-0 state simulations of 100 ps. All four started from the the same solvation 0-0 state. (a) no special treatment; (b) key torsion was scaled to 0; (c) intra-molecular vdw interactions were annihilated; (d) both treatments in (b) and (c) were applied.

improved, in some cases quite substantially.

Several independent binding free energies of G9 were calculated, where the differences in the final results are 0.1 kcal/mol or less, showing good reproducibility of the method.
Figure 4.30: Torsional angle (methine C)-C-(amide N)-(carbonyl C) of Palonosetron-H\(^+\) in a water box simulated from the standard 0-0 state simulations of 100 ps. All four started from the the same host-guest 0-0 state bound in CB[8]. (a) no special treatment; (b) key torsion was scaled to 0; (c) intra-molecular vdw interactions were annihilated; (d) both treatments in (b) and (c) were applied.

4.5 Conclusion and Acknowledgement

The binding free energies of fourteen molecules and cucurbit[8]uril in the SAMPL6 challenge were predicted by free energy perturbation theory and molecular dynamics simulations with AMOEBA force field. Only half of the errors in the initial submitted predictions were less than 2 kcal/mol before recalculation and several changes were made. The force field parameters of cucurbit[8]uril were fixed with a minor patch, based on its crystal structures and quantum calculations. The protocols for the calculation were
Table 4.5: Revised binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.

<table>
<thead>
<tr>
<th>Guest</th>
<th>Prediction</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-8.03</td>
<td>-6.69</td>
</tr>
<tr>
<td>1</td>
<td>-7.00</td>
<td>-7.65</td>
</tr>
<tr>
<td>2</td>
<td>-11.13</td>
<td>-7.66</td>
</tr>
<tr>
<td>3</td>
<td>-6.34</td>
<td>-6.45</td>
</tr>
<tr>
<td>4</td>
<td>-6.62</td>
<td>-7.80</td>
</tr>
<tr>
<td>5</td>
<td>-12.10</td>
<td>-8.18</td>
</tr>
<tr>
<td>6</td>
<td>-8.92</td>
<td>-8.34</td>
</tr>
<tr>
<td>7</td>
<td>-9.71</td>
<td>-9.98</td>
</tr>
<tr>
<td>8</td>
<td>-15.55</td>
<td>-13.50</td>
</tr>
<tr>
<td>9</td>
<td>-8.81</td>
<td>-8.68</td>
</tr>
<tr>
<td>10</td>
<td>-8.54</td>
<td>-8.22</td>
</tr>
<tr>
<td>11</td>
<td>-7.30</td>
<td>-7.77</td>
</tr>
<tr>
<td>12</td>
<td>-9.05</td>
<td>-7.05</td>
</tr>
<tr>
<td>13</td>
<td>-5.64</td>
<td>-7.11</td>
</tr>
</tbody>
</table>

also improved via annihilation of intra-molecular vdw and key torsions to enhance the sampling efficiency. Changes in the implementations of the Tinker and Tinker-OpenMM packages were merged to incorporate with the new protocols. In the final prediction set, the binding free energies were systematically improved.

I would like to thank Dr. Marie Laury for the quantum calculations of CB[8] structures, and Dr. Laury and Aaron Gordon for their help with some of the binding free energy calculations.

4.6 Supplementary Information

The supplementary information starts on page 126.
Figure 4.31: Revised binding free energies (in kcal/mol) of CB[8] and the guests in the SAMPL6 challenge.
Program 4.1: Example of the trans-1,4-Diaminocyclohexane-2H\(^+\) Coordinates in Psi4 Format

```plaintext
# molecule MOLECULE_NAME {
# CHARGE MULTIPLICITY
# ATOM1 X1 Y1 Z1
# ATOM2 X2 Y2 Z2
# ...
#
# OTHER SETTINGS
# }
molecule CB8_G11 {
  2 1
  N 2.880222468096 -0.070008472159 0.000000000000
  H 2.988966739157 -1.090954824899 0.000000000000
  H 3.371358657755 0.289958540624 0.827090434496
  H 3.371358657755 0.289958540624 -0.827090434496
  C 1.418813376566 0.331727436646 0.000000000000
  H 1.426005143134 1.426916298294 0.000000000000
  C 0.743064793814 -0.194296547440 1.262795055907
  C -0.743064793814 0.194296547440 1.262795055907
  C -1.418813376566 -0.331727436646 0.000000000000
  H -1.426005143134 -1.426916298294 0.000000000000
  N -2.880222468096 0.070008472159 0.000000000000
  H -2.988966739157 1.090954824899 0.000000000000
  H -3.371358657755 -0.289958540624 -0.827090434496
  H -3.371358657755 -0.289958540624 0.827090434496
  C -0.743064793814 0.194296547440 -1.262795055907
  H 0.834793116083 -1.287619104460 1.303211064455
  H 1.214298702041 0.212686107746 2.163283720416
  H 0.834793116083 1.287619104460 -1.303211064455
  H -0.834793116083 -1.287619104460 1.303211064455
  H -1.214298702041 0.212686107746 -2.163283720416
  H -0.834793116083 1.287619104460 -1.303211064455
  H 1.214298702041 0.212686107746 -2.163283720416
  H 0.834793116083 -1.287619104460 -1.303211064455
}

no_com
no_reorient
units angstrom

126
Program 4.2: Example of the MP2/6-311G(1d,1p) Optimization Input in Psi4 Format

```
# molecule coordinates...
set {
basis 6-311G(d,p)
scf_type df
guess sad
}
energy = optimize('mp2')
```

Program 4.3: Example of the GDMA Input in Psi4 Format

```
# molecule coordinates...
set {
basis 6-311G(d,p)
gdma_limit 2  # 0 point charge, 1 dipole, etc.
gdma_switch 0.0  # use the original DMA 1.0 method
gdma_radius [H, 0.65]
}
grad, wfn = gradient('mp2', return_wfn=True)
gdma(wfn)
```

Program 4.4: Example of the MP2/aug-cc-pVTZ Input in Psi4 Format

```
# molecule coordinates...
set {
basis aug-cc-pVTZ
scf_type df
guess sad
}
grad, wfn = gradient('mp2', return_wfn=True)
fchk_writer = psi4.FCHKWriter(wfn)
fchk_writer.write('output.fchk')
```
Program 4.5: Using PyMOL to Remove the Hydrogens in a PDB File

#!/bin/bash

if [ "#$" -lt 1 ]; then
    usage="Usage: $(basename $0) filename.pdb"
    echo $usage
    exit
fi

fname1=$1
fname2=${fname1%.pdb}
fname3="${fname2}_2.pdb"
if [ ! -f $fname2 ]; then
    echo $fname1 does not exist.
    exit
fi

if [ -f $fname3 ]; then
    echo $fname3 already exists.
    exit
fi

pymol -cq "load $fname1; remove hydrogens; save $fname3"
Program 4.6: AMOEBA Cucur[8]bituril Host Parameters

#################################################################
## AMOEBA Cucur[8]bituril Host Parameters ##
# #############################################

atom 51 51  O  "Cucurbituril O"  8  15.999  1
atom 52 52  C  "Cucurbituril C=O"  6  12.011  3
atom 53 53  N  "Cucurbituril N"  7  14.007  3
atom 54 54  C  "Cucurbituril CH"  6  12.011  4
atom 55 55  H  "Cucurbituril HC"  1  1.008  1
atom 56 56  C  "Cucurbituril CH2"  6  12.011  4

vdw 51 3.3000 0.1120
vdw 52 3.8200 0.1060
vdw 53 3.7100 0.1100
vdw 54 3.8200 0.1010
vdw 55 2.9300 0.0260 0.910
vdw 56 3.8200 0.1010

bond 51 52 590.00 1.2197
bond 52 53 450.00 1.3639
bond 53 54 354.80 1.4460
bond 54 55 354.80 1.4460
bond 55 56 380.00 1.0930
bond 54 55 380.00 1.0930

angle 51 52 53  77.00 124.80
angle 53 52 53  64.70 112.50
angle 52 53 54 116.50 121.10
angle 52 53 56 116.50 121.10
angle 54 53 56  54.70 122.50
angle 53 54 53  85.70 112.00
angle 53 54 54  54.00 111.30
angle 53 54 55  54.70 111.00
angle 54 54 55  42.40 109.80
angle 53 56 53  85.70 112.00
angle 53 56 55  54.70 111.00
angle 55 56 55  39.60 107.60

strbnd 51 52 53  18.70 18.70
strbnd 53 52 53  18.70 18.70
strbnd 52 53 54  7.20  7.20
strbnd 52 53 56  7.20  7.20
strbnd 54 53 56  7.20  7.20
strbnd 53 54 53  18.70 18.70
strbnd 53 54 54  18.70 18.70
strbnd 53 54 55  11.50 11.50
strbnd 54 54 55 11.50 11.50
strbnd 53 56 53 18.70 18.70
strbnd 53 56 55 11.50 11.50
strbnd 55 56 55 11.50 11.50

opbend 51 52 0 0 46.80
opbend 53 52 0 0 107.90
opbend 52 53 0 0 12.90
opbend 54 53 0 0 12.90
opbend 56 53 0 0 12.90

torsion 51 52 53 54 1.000 0.0 1 2.250 180.0 2 -2.250 0.0 3
 torsion 51 52 53 56 1.000 0.0 1 2.250 180.0 2 -2.250 0.0 3
 torsion 53 52 53 56 0.000 0.0 1 2.250 180.0 2 -0.500 0.0 3
 torsion 53 52 53 56 0.000 0.0 1 2.250 180.0 2 -0.500 0.0 3
 torsion 52 53 54 53 0.000 0.0 1 0.000 180.0 2 0.300 0.0 3
 torsion 52 53 54 56 0.000 0.0 1 0.000 180.0 2 -0.126 0.0 3
 torsion 56 53 54 53 0.000 0.0 1 0.000 180.0 2 0.300 0.0 3
 torsion 56 53 54 54 0.000 0.0 1 0.000 180.0 2 0.200 0.0 3
 torsion 56 53 54 55 0.000 0.0 1 0.000 180.0 2 0.460 0.0 3
 torsion 52 53 56 53 0.400 0.0 1 -0.400 180.0 2 -1.500 0.0 3
 torsion 52 53 56 55 0.000 0.0 1 0.000 180.0 2 -0.126 0.0 3
 torsion 54 53 56 53 0.000 0.0 1 0.000 180.0 2 0.300 0.0 3
 torsion 54 53 56 55 0.000 0.0 1 0.000 180.0 2 0.460 0.0 3
 torsion 53 54 54 53 0.000 0.0 1 0.000 180.0 2 0.500 0.0 3
 torsion 53 54 54 55 0.000 0.0 1 0.000 180.0 2 0.400 0.0 3
 torsion 55 54 54 55 0.000 0.0 1 0.000 180.0 2 0.300 0.0 3

pitors 52 53 6.85

multipole 51 52 53 -0.73508
 0.01916 0.00000 -0.14964
 -0.46999
 0.00000 0.20968
 0.03162 0.00000 0.26031

multipole 52 51 53 0.91731
 0.01410 0.00000 0.02405
 0.04230
 0.00000 -0.05163
 -0.02036 0.00000 0.00933

multipole 53 52 54 -0.12817
 0.10621 0.00000 -0.37552
 1.36169
 0.00000 -1.33989
 0.23303 0.00000 -0.02180

multipole 54 -53 -53 -0.11256
 0.00063 0.00000 0.04453
 0.20813
 0.00000 -0.10715
 -0.11661 0.00000 -0.10098
multipole 55 54 53 0.00341
-0.00041 0.00000 0.02849
-0.04208 0.00000 0.07230
-0.02427 0.00000 -0.03022
multipole 56 -53 -53 0.04014
-0.00024 0.00000 0.32347
0.89231 0.00000 -0.80951
0.03513 0.00000 -0.08280
multipole 57 56 53 0.07156
0.00033 0.00000 -0.10395
0.08632 0.00000 -0.03554
-0.00935 0.00000 -0.12186

polarize 51 0.8370 0.3900 52
polarize 52 1.3340 0.3900 51 53
polarize 53 1.0730 0.3900 52
polarize 54 1.3340 0.3900 55
polarize 55 0.4960 0.3900 54
polarize 56 1.3340 0.3900 57
polarize 57 0.4960 0.3900 56
Program 4.7: Example of the Tinker Key File for R-Bornylamine-H\textsuperscript{+} Host-Guest Simulations

PARAMETERS cg5.prm
ARCHIVE

A-AXIS 40.0
NEIGHBOR-LIST

EWALD
EWALD-CUTOFF 7.0

POLAR-EPS 1.0E-5
POLAR-PREDICT

VDW-CUTOFF 12.0
VDW-CORRECTION

INTEGRATOR RESPA
THERMOSTAT BUSSI
BAROSTAT MONTECARLO

LIGAND -1 31
VDW-ANNIHILATE

GROUP 1 -1 31
GROUP 2 -32 175
RESTRAIN-GROUPS 1 2 5.0 0 2.0
VDW-LAMBDA 0.50
ELE-LAMBDA 0.00
CHAPTER 5

CONCLUSIONS AND FUTURE PERSPECTIVES

The accuracy of the underlying force field or empirical potential will always be of the first priority in any serious applications of molecular dynamics. The functional forms and parameters of the force field are, of course, important to accurate simulation of real systems. For a relatively complex host-guest system such as that described in chapter 4, a small change in torsional parameters of the host molecule dramatically improves the calculated binding free energy estimates. Even for chemically simple systems, like the electrolyte solutions discussed in chapter 2, achieving reasonable results can be challenging. Transferring parameters to calculate foreign properties has no guarantee of yielding meaningful results, e.g. in chapter 2 the parameters fit from relative ion solvation free energies cannot exactly match the quantum results for ion-water dimer structures and energies, though the calculations are not far off. However, lattice energies calculated with these parameters are in significant error.

Employing good experimental results in a fitting procedure is important, but the accuracy of resulting parameters is still limited by other factors. Not all parameters are determined from first principles. The quality of the parameters for fundamental systems such as water will set an upper limit of the quality of any parameters derived from them. Besides force field parameterization issues, incorporating more physics into the force field is equally important, e.g. the directionality of interactions between molecules is not well described by a simple fixed charge model. The Ponder Lab is currently trying to develop well-defined nonbonded interaction terms, e.g. charge penetration, charge transfer, etc. to
make the AMOEBA force field more robust and in closer correspondence with ab initio physics.

Conformational and configurational sampling is an important application of molecular dynamics, where the trade off of accuracy and efficiency becomes important. Advanced force fields become useless if they are unbearably slow. One solution, adopted by the OpenMM community [116], is to develop a hardware specific code base to obtain fast computational speeds on modern GPU processors. Multiple time step methods are excellent examples of the endeavor the computational chemistry community has made to develop algorithms to accelerate MD simulations. Attention should be paid to the algorithms that strongly perturb the dynamics of the systems, in which case, simulations are less of correct dynamics and are limited to sampling.

The binding free energy calculations in chapter 4 are an advanced application of conformational/configurational sampling. The FEP and BAR methods [44, 45] used in this work are to some extent based on brute force sampling. Metadynamics methods such as OSRW [117–119] etc. are good candidates to enhance the efficiency of conformational sampling. They will definitely be important to the modelling of complex systems in the future.

Besides the problems of computational chemistry solved in the previous chapters, this work will benefit other research fields as well. In biological systems, for example, all of the biochemical processes take place in ion solutions, e.g. the intracellular concentration of potassium ion in mammalian cells is 140 mmol/L. Therefore, high-quality ion parameters are indispensable to the biochemical and biological simulations. This work has also developed insights into protein-ligand binding calculations. Computational drug design is the “holy grail” of the pharmaceutical industry and accurate protein-ligand binding calculations represent a major step toward that goal.
REFERENCES


40. Parrinello, M. & Rahman, A. Crystal Structure and Pair Potentials: A Molecular-


42. Yashonath, S. & Rao, C. A Monte Carlo study of crystal structure transformations.

43. Frenkel, D. & Smit, B. Understanding Molecular Simulation: From Algorithms to Appli-

44. Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I.

45. Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo

46. Auffinger, P., Cheatham, T. E. & Vaiana, A. C. Spontaneous Formation of KCl Ag-

47. Aqvist, J. Ion-water interaction potentials derived from free energy perturbation


50. Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem

51. Feller, D. The Role of Databases in Support of Computational Chemistry Calcula-

52. Schuchardt, K. L. et al. Basis Set Exchange: A Community Database for Computa-

53. Harger, M. et al. Tinker-OpenMM: Absolute and relative alchemical free energies


108. *The SAMPL Project* [https://drugdesigndata.org/about/sampl](https://drugdesigndata.org/about/sampl).

109. *SAMPL6* [https://drugdesigndata.org/about/sampl6](https://drugdesigndata.org/about/sampl6).


118. Zheng, L., Chen, M. & Yang, W. Simultaneous escaping of explicit and hidden free energy barriers: Application of the orthogonal space random walk strategy

WASHINGTON UNIVERSITY IN ST. LOUIS

GRADUATE SCHOOL OF ARTS AND SCIENCES

THESIS DEFENSE

For

Zhi Wang

For the degree

DOCTOR OF PHILOSOPHY

Monday, May 7th, 2018
2:00 pm
McMillen 311

Committee of Examiners

Professor Jay W. Ponder, Chair
Professor Gregory R. Bowman
Professor Richard Mabbs
Professor Garland R. Marshall
Professor Lee G. Sobotka

Title

Polarizable Force Field Development, and Applications to Conformational Sampling and Free Energy Calculation

Reception follows @ 6:00 pm in Louderman 561
Washington University in St. Louis  
GRADUATE SCHOOL OF ARTS & SCIENCES  

EXAMINATION APPROVAL FORM  

To the Graduate School:

We, the undersigned, report that as a committee we have examined the dissertation entitled Polarizable Force Field Development, and Applications to Conformational Sampling and Free Energy Calculation and have given it our approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy. We certify that the dissertation conforms to the requirements described in “Statement by Graduate Council on Minimal Requirements for Ph.D. Dissertations” as published in the Washington University Doctoral Dissertation Guide. We also certify that we have examined Zhi Wang ID #430408 upon the work done in Chemistry and find that (his, her) attainments are such that (he, she) may properly be admitted to the degree of Doctor of Philosophy.

Signed  
Prof. Jay W. Ponder, Chair  
Prof. Richard Mabbs  
Prof. Lee G. Sobotka  
Prof. William Buhro, Chair of Dept.

I dissent from the foregoing report.

Date May 7, 2018

Ash. Prof. Gregory R. Bowman  
Prof. Garland R. Marshall  
Prof. Richard Loomis, DGS

This form should be signed and forwarded to the Graduate School of Arts and Sciences, Cupples II Suite 204 or Campus Box 1187.

Recorded by the Graduate School (initial/date): ________________________________

Dissertation Submission Approved (initial/date): ________________________________