Optimal Point Charge Approximation:
from 3-Atom Water Molecule to Million-Atom Chromatin Fiber

Saeed Izadi

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Engineering Mechanics

Alexey V. Onufriev, Chair
David R. Bevan
Shane D. Ross
Rafael V. Davalos
Anne E. Staples

May 10, 2016
Blacksburg, Virginia

Keywords: Molecular Modeling, Explicit Solvent Model, Force Field, Implicit Solvent Model, Point Charge Approximation, Multipole Moments, Electrostatics, Water Models, Multi-scale Models

Optimal Point Charge Approximation:
from 3-Atom Water Molecule to Million-Atom Chromatin Fiber

Saeed Izadi

(ABSTRACT)

Atomistic modeling and simulation methods enable a modern molecular approach to biomedical research. Issues addressed range from structure-function relationships to structure-based drug design. The ability of these methods to address biologically relevant problems is largely determined by their accurate treatment of electrostatic interactions in the target biomolecular structure. In practical molecular simulations, the electrostatic charge density of molecules is approximated by an arrangement of fractional "point charges" throughout the molecule. While chemically intuitive and straightforward in technical implementation, models based exclusively on atom-centered charge placement, a major workhorse of the biomolecular simulations, do not necessarily provide a sufficiently detailed description of the molecular electrostatic potentials for small systems, and can become prohibitively expensive for large systems with thousands to millions of atoms. In this work, we propose a rigorous and generally applicable approach, Optimal Point Charge Approximation (OPCA), for approximating electrostatic charge distributions of biomolecules with a small number of point charges to best represent the underlying electrostatic potential, regardless of the distance to the charge distribution. OPCA places a given number of point charges so that the lowest order multipole moments of the reference charge distribution are optimally reproduced. We provide a general framework for calculating OPCAs to any order, and introduce closed-form analytical expressions for the 1-charge, 2-charge and 3-charge OPCA. We demonstrate the advantage of OPCA by applying it to a wide range of biomolecules of varied sizes. We use the concept of OPCA to develop a different, novel approach of constructing accurate and simple point charge water models. The proposed approach permits a virtually exhaustive search for optimal model parameters in the sub-space most relevant to electrostatic properties of the water molecule in liquid phase. A novel rigid 4-point Optimal Point Charge (OPC) water model constructed based on the new approach is substantially more accurate than commonly used models in terms of bulk water properties, and delivers critical accuracy improvement in practical atomistic simulations, such as RNA simulations, protein folding, protein-ligand binding and small molecule hydration. We also apply our new approach to construct a 3-point version of the Optimal Point Charge water model, referred to as OPC3. OPCA can be employed to represent large charge distributions with only a few point charges. We use this capability of OPCA to develop a multi-scale, yet fully atomistic, generalized Born approach (GB-HCPO) that can deliver up to 2 orders of magnitude speedup compared to the reference MD simulation. As a practical demonstration, we exploit the new multi-scale approach to gain insight into the structure of million-atom 30-nm chromatin fiber. Our results suggest important structural details consistent with experiment: the linker DNA fills the core region and the H3 histone tails interact with the linker DNA. OPC, OPC3 and GB-HCPO are implemented in AMBER molecular dynamics software package.

Dedicated to my beloved parents for their infinite love, patience and support throughout all these years.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Alexey Onufriev for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research. I could not have imagined having a better advisor and mentor for my Ph.D. studies.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Rafael V. Davalos, Prof. Shane Ross, Prof. David R. Bevan, and Prof. Anne Staples for their encouragement and insightful comments.

My sincere thanks also goes to Dr. Ramu Anandakrishnan for his generous assistance throughout the course of my Ph.D. studies. I thank my fellow labmates at Virginia Tech: Dr. Igor Tolokh, Dr. Boris Aguilar, Dr. Abhishek Mukhopadhyay, Dr. Alexander Drozdetski, Dr. Nick Kinney, Negin Forouzesh, and Parviz Shabane for the stimulating discussions, valuable comments and for all the fun we have had in the last few years.

Finally, I could not have made to this point initially without the help and support from my family, friends and teachers throughout the course of my education.

Contents

List of Figures

List of Tables

Glossary

The definition of a few key terms used extensively throughout this document.

Explicit Solvent: *Solvent models that treat the solvent molecules explicitly, i.e. the coordinates and usually at least some of the molecular degrees of freedom are included. This is a more intuitively realistic picture in which there are direct, specific solvent interactions with a solute, in contrast to continuum models.*

Implicit Solvent: *Sometimes known as continuum solvent, a method of representing solvent as a continuous medium that have the average properties of the real solvent.*

Generalized Born model(GB): *A model for implicit solvation that provides an analytical approximation of the Poisson equation to compute the electrostatic part of the solvation free energy*

Force Field: *Refers to the functional form and parameter sets used to calculate the potential energy of a system of atoms in molecular mechanics and molecular dynamics simulations.*

Solvation Free Energy *The change in free energy for transferring a molecule from gas phase to aqueous phase, conventionally estimated as the molar transfer free energy, in kcal/mol.*

Chromatin Fiber: *The second level of DNA compaction, after the nucleosome. The atomistic details of the structure of chromatin fiber are unknown.*

Chapter 1

Introduction

Molecular interactions are the basis of life, therefore understanding their intricate details is critical to progress in many areas of the biological sciences. Current experimental methods alone do not provide us with a complete picture of what happens at this microscopic level (mainly because the objects involved are too complex and their parts move too fast) and can also be prohibitively expensive. Consequently, computer simulations become an indispensable research tool [51, 65, 114, 111, 212, 204, 110, 109].

In principle, all details of molecular structures and interactions can be predicted from first principles using quantum mechanics[124]. However, quantum mechanics can become computationally extremely expensive for many of the problems we are interested. The development of simplified methodologies is critical in the study of structure and dynamics of biological macromolecules[124]. One of the most practical simplifications in biomolecular simulations is to ignore the electronic degrees of freedom of the molecules, and only calculate motions of the nuclei. This simplification, which is the fundamental assumption in all practical molecular dynamics methodologies, stems from the Born-Oppenheimer approximation that assumes nuclear and electronic motions are independent and therefore the energy of a system can be written as a function of nuclear coordinates only[51, 65, 114, 124]. This approximation is the basis for all-atomistic molecular mechanics methods that simulate each atom as a single particle. Each particle is characterized by a radius (i. e. the van der Waals radius), polarizability, and a constant net charge. There are also bonded interactions that are treated as springs with an equilibrium distance equal to the experimental or calculated bond length. In the context of molecular modeling, such functional abstraction is known as a *force field* which is a mathematical expression describing the dependence of the energy of a system on the coordinates of its particles.

Most classical molecular mechanics force fields rely on five terms with a simple physical interpretation: there are potential energy terms associated with deformation of bond and angle geometry (stretching/compression of bonds, bending of angles), terms associated with the rotation about certain dihedral angles (torsions), and the so-called "non-bonded" terms,

describing the electrostatic interactions and terms describing the dispersion interactions and repulsion when atoms overlap (van der Waals forces). The expression for the potential energy of a molecular system that is used most frequently for simple organic molecules and biological macromolecules is as below:

$$V(r) = \sum_{bonds} \frac{k_b}{2}(b - b_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{dihedralangles} \frac{k_\phi}{2}(1 - cos(n\phi - \delta)) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left[\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}} \right] + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left[\frac{q_i q_j}{\epsilon_D r_{ij}} \right] \tag{1.1}$$

while k_b, k_θ, k_ϕ are the bond, angle, dihedral force constants. The corresponding bond length, angle, dihedral angle are represented by b, θ and ϕ with the subscript zero representing the equilibrium value. A and C are Lennard Jones (LJ) coefficients and q_i is the atomic partial charge. $V(r)$ denotes the potential energy and r_{ij} represents the distance between atoms i and j. For the non-bond interactions (LJ and electrostatic interactions), the sum runs in principle over all pairs of atoms which are more than two bonds apart or belonging to different molecules. The details of the non-bond list critically depend on the system and the method chosen for the calculation.

Among all of the five interactions described above, the computation of the non-bonded electrostatic interactions (the fifth term in Equation 1.1) is most challenging. In contrast to the other four terms that represent "local" interactions, electrostatic interactions are long-range, as they decay as $1/r_{ij}$, in nature and their evaluations present a fundamental problem: a system of N atoms demands an amount of work proportional to N^2 for such calculations. Therefore, the limiting factor in most all-atom simulations is the computation of long range electrostatic interactions[169, 118]. Electrostatic interactions are also strong and their accurate evaluation is key to the outcomes of biomolecular simulations. The accuracy and speed of electrostatic interactions depend on the way the charge density of biomolecules are represented. Currently the most common approach for representing the charge distributions in molecular simulations is atom-center charge placement, as described below.

1.1 Partial atomic charges used to represent molecular electrostatic potentials

In practical molecular simulations, the electrostatic charge density of molecules is commonly approximated by an arrangement of fractional point charges throughout the molecule. Unfortunately, point charges are not quantum mechanical observables or experimentally measurable quantities with a well understood physical meaning. Therefore, a lot of effort has

been put into developing methods to determine partial charges that reproduce electrostatic properties of molecules, and in particular the electrostatic potential obtained from quantum mechanics calculations. [22, 41, 29, 180, 43, 20, 95, 12, 192].

In the most widely used point charge model of molecular electrostatics, the point charges are placed at the center of atoms and the charge values are chosen to reproduce electrostatics, calculated from quantum mechanics, at a given surface[124] (Figure 1.1). While chemically intuitive and straightforward in technical implementation, this model does not provide a sufficiently detailed description of the anisotropic features of the molecular electrostatic potential. For example, for very small systems containing just a few atoms (e.g. the water molecule), models based exclusively on atomic charges may be unable to accurately describe the complexity of the electrostatic (i.e. higher order multipole moments) resulted from the anisotropic nature of the charge distribution. On the other side, for very large systems with thousands to millions of atoms (e.g. the Chromarin fiber), assigning point charges to all individual atoms creates too many interacting sites in the model, leading to inordinate computational costs.

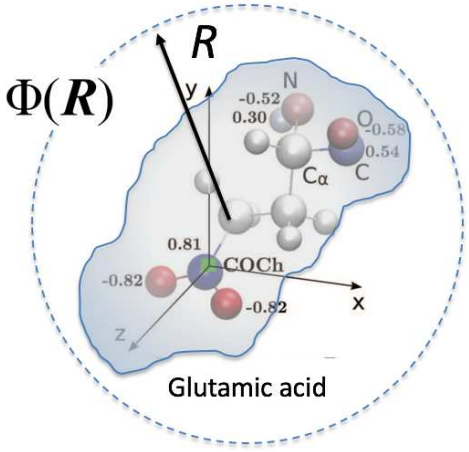


Figure 1.1: Common approach to represent electrostatics: atom-centered point charge placement (e.g. EPS) optimizes electrostatics at a given surface. The figure illustrates the distribution partial charges on the atoms centers for a glutamic acid group within a protein with net charge $= -1e$, where the group includes the associated NH-CH-CO backbone atoms. The charge values for charges $|q| > 0.2e$ are shown next to the atoms.

Alternative approaches to deriving atomic partial charges include empirically fitting a set of point charges to a given charge distribution by minimizing various error metrics in electrostatic potential over some volume or surface surrounding the charge distribution, such as RESP [22], CHELP [41], CHELPG [29], CHELMO [180], Finite Point Charge (FPC)[43], coarse graining [20, 95, 12] and others [192]. One drawback of these methods is that minimizing the electrostatic error over some specific volume or surface can potentially lead to relatively large errors outside those volume or surface. Also, these methods often predeter-

mine the positions of the point charges and only charge values are parametrized to reproduce electrostatic potential. As we will show later, any geometrical constraint on charge positions can adversely affect the ability to optimally reproduce the electrostatic potential. Another drawback of the above approaches is that the point charges are obtained from numerical fits rather than from analytical expressions whereas for many practical applications, such as molecular dynamics simulations, analytical expressions are crucial for "on-the-fly" calculations. There is a great desire for a systematic procedure for choosing the positions of point charges in distributed charge models.

1.2 Optimal Point Charge Approximation (OPCA)

In this work, we present a rigorous and generally applicable approach, Optimal Point Charge Approximation (OPCA), for approximating electrostatic charge distributions with a very small number of point charges so that the underlying electrostatic potential is best represented, regardless of the distance to the charge distribution (Figure 1.2). OPCA does not impose any geometrical constraint on the position of the point charges. Instead, it finds the positions and magnitude of a small number of point charges so that the underlying electrostatic potential is optimally reproduced, regardless of the distance to the charge distribution. OPCA places the approximating point charges so that the lowest order electrostatic multipole moments of the charge distribution are best reproduced. As a result, OPCA inherits the physically appealing asymptotic properties of the point multipole approximation, i.e. the error in potential is guaranteed to fall off at least as fast as $1/R^{k+1}$, where R is the distance from the origin and k is the highest order of the multipole terms retained in the expansion. The above asymptotic behavior is a key difference between optimal point charge approximation and previous methods that simply fit the representative charges to minimize electrostatic error over some arbitrary volume or surface (e.g. molecular surface), which can potentially lead to relatively large errors outside the volume or surface used for fitting.

We provide a general framework for calculating OPCAs to any order. We also derive closed-form analytical expressions for the 1-charge, 2-charge, and 3-charge OPCA. These analytic expressions not only provide physical insight but are more computationally efficient than the numerical minimization procedures that are in general required to obtain the optimal point charge approximation. Thus, these analytic expressions may be particularly useful in applications such as molecular dynamics where computational speed is critical.

A detailed description of the Optimal Point Charge Approximation approach is given in Chapter 2.

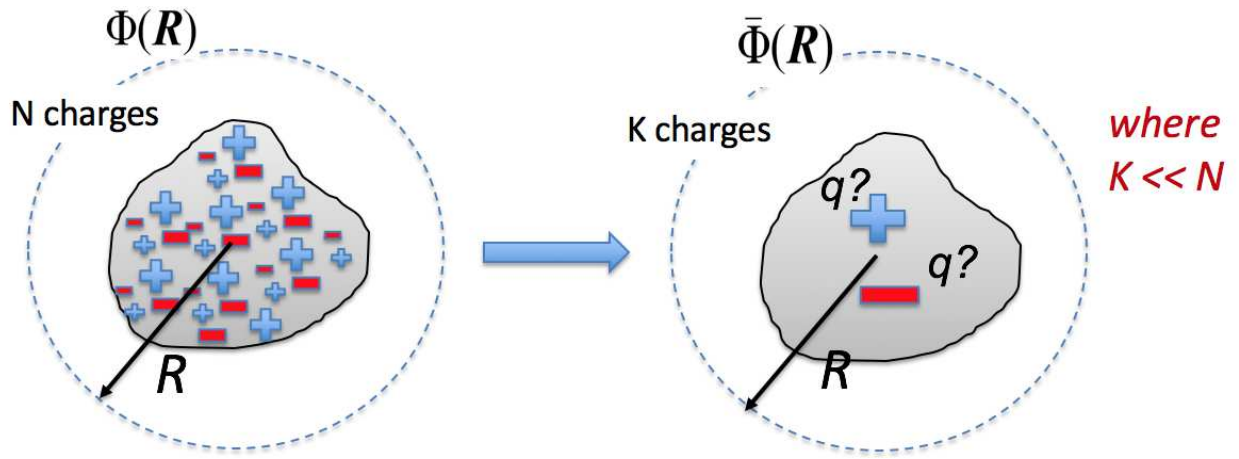


Figure 1.2: For a given set of N original charges the optimal point charge approximation finds the position and magnitude of K point charges such that the potential due to these smaller number of point charges, $\bar{\Phi}(\boldsymbol{R})$ best approximates the potential of the original distribution, $\Phi(\boldsymbol{R})$, independent of distant R.

1.3 OPCA in practical biomolecular simulations

In general, OPCA can be applied to a broad class of problems in which an efficient method for calculating the electrostatic interactions between biomolecules in solution is desired. To best demonstrate the advantage of the optimal point charge approximation we implement it to a broad range of biomolecules of varied sizes. In particular, here we apply the concept of OPCA to develop a novel approach for constructing classical water models with electrostatic properties that are closer to what has been found in experiment and high level quantum mechanics, which is difficult to do in models based on atom-centered charge placements. In another example, we use the concept of OPCA to develop a new muti-scale approach that significantly speeds up the computations of electrostatic interactions in MD simulations. We use the new multi-scale approach to gain insight into the structure of a million-atom chromatin fiber.

A brief description of the application of OPCA in modeling 3-atom water molecule and million-atom structure of chromatin fiber is given below.

1.3.1 OPCA improves the accuracy of classical water models

Molecular modeling and simulations are routinely employed to study structure and function of biological molecules, with over 12,000 biomolecular modeling papers published in 2009 alone, in applications ranging from structural biology to bio-medicine and rational drug de-

sign. Accurate, classical water modelscite[104, 2, 85, 7, 138, 23, 217, 195, 218, 166, 123, 208, 122, 197, 143, 128, 16, 156, 221] are just as important for these modeling efforts as water is for Life. The simplest, and most widely used, atomistic water models are fixed-charge rigid non-polarizable models, implemented in virtually every modeling package. However, despite at least three decades of effort there is still significant room for much needed improvement. As more physical realism is added to such models either through more complex geometry or/and by inclusion of electronic polarization effects, the cost of finding the accuracy optimum in the large parameter space grows exponentially. As a result, available parametrizations are virtually guaranteed to be sub-optimal with respect to faithfully reproducing key experimental properties of water, hindering predictive potential of these models.

Search for theoretical models that can accurately describe how this deceptively simple molecule of just three atoms gives rise to the many extraordinary properties of its liquid phase [61, 62, 18] is far from complete [79, 140]. The most simple and computationally efficient, rigid non-polarizable models [104, 2, 85, 138, 23] that represent water molecule as a set of point charges at fixed positions relative to the oxygen nucleus stand out as the class used in the vast majority of atomistic biomolecular studies today.

Most commonly used models of this class, *e.g.* 3-point TIP3P [104] or SPC/E [23], or 4-point TIP4P-Ew [85], offer a compromise between accuracy and speed, but are by no means perfect. The need for better accuracy motivates an on-going search for more accurate yet computationally facile water models. Yet, despite notable recent improvements[211, 66, 210, 60], these models still fail to reproduce all the key properties of bulk water accurately and simultaneously[94]. Life itself depends on several of these properties being precisely what they are, *e.g.* the strength of water-water hydrogen bonds (even $\sim$2% change can make a critical difference). Importantly, even modest inaccuracies of water models can drastically affect outcomes of atomistic biomolecular modeling in an unpredictable, adverse manner. The sensitivity is not surprising given the extraordinary complexity of real water-water interactions and hydrogen bonding networks in liquid phase, and their extreme sensitivity to various properties of water models[217]. While in some cases fortuitous cancellation of errors between solute-solvent and solvent-solvent interactions leads to reasonable agreement with experiment, the balance can not be maintained in general if the solvent-solvent part is wrong. This is especially true for simulations in which appreciable change in solvent exposure occurs (protein folding, ligand binding).

For larger protein-ligand systems, the discrepancies between binding energies can exceed 10 kcal/mol for commonly used water models[93], which is unacceptable for quantitative molecular modeling efforts. Likewise, widely used water models fail to predict correct experimental size of intrinsically disordered proteins[160] or the balance between RNA tetraloop populations *regardless of the underlying force-field used.* And even though some models (*e.g.* TIP3P) can perform better than others in predicting hydration free energies of small molecules[145], the average errors are still outside the desired "chemical accuracy" of less than 1 kcal/mol, the goal for rational drug design efforts .

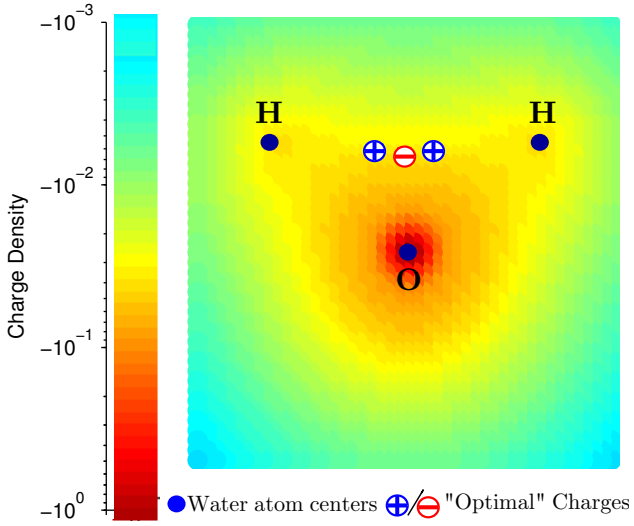


Figure 1.3: Charge distribution of the water molecule in the gas phase obtained from a quantum mechanical calculation [9]. Counter-intuitively, three point charges that optimally reproduce the electrostatic potential of this charge distribution, calculated based on the 3-charge OPCA, are clustered in the middle, as opposed to the on-nuclei placement used by common water models that results in a much poorer electrostatic description of the underlying charge distribution[9]. Motivated by this observation, we propose a new approach for constructing point charge water models (see chapters 3 and 4).

Procedures employed to develop commonly used rigid water models generally impose constraints on the geometry (OH bond length and HOH angle) based on experimental observations, most commonly by fixing the positive point charges at the hydrogen nuclei positions. Inspired by the classical works[142, 25] that revealed V-shape of water molecule, the common approach to build water models imposes constraints on the allowed variations of the model geometry. That is $|OH|$ bond length and $\angle$HOH angle are either fixed, or are only allowed to vary slightly around their "canonical" values. The assumption is that optimal locations of the positive point charges of the model should be somewhere near the experimental hydrogen nuclei positions. This approach may not necessarily accurately reproduce the electrostatic characteristics of the water molecule due to severe constraints on allowed variations in the charge distribution being optimized.

In this work we show that the configuration of three point charges to best describe the charge distribution of the water molecule, calculated based on the concept of Optimal Point Charge Approximation, can be very different from what one may intuitively expect based on its well-known atomic structure (Figure 3.1). We also show that these point charges significantly improve the accuracy in representing the electrostatic properties of the water molecule. Intrigued by the idea that optimal placement of the point charges in a water model can be very different from the "intuitive" placement on the nuclei, and encouraged by the significant improvement of the accuracy of electrostatics brought about by this strategy

in gas-phase, we formulate and test a different approach to building classical water models for the liquid phase, using the concept of OPCA, which is completely different from the mainstream approaches.

The new approach finds the optimal magnitude and positions for point charges so that an accurate yet simplified description of the charge distribution of the water molecule that can adequately account for the hydrogen bonding in the liquid phase is achieved. The proposed approach will allow us to perform an exhaustive search for optimal parameters of fixed-charge, rigid water models based on n-point topologies, even for large n. As proof-of-concept, we have constructed two versions of one such model – 4-point OPC and 3-point OPC3.

Success of this work will be of immediate importance to both practitioners and model developers. The practitioners will have immediate access to significantly more accurate water models of the same high computational performance as currently implemented in all major modeling packages. These new models can immediately replace previous generation models such as TIP3P or TIP4P in all the applications where these models are currently used, including multi-scale simulations. Developers will have a much better idea of the accuracy limits of the widely used point charge water models, and novel approaches to improving polarizable models.

Details of the proposed approach for building water models, as well as the parametrizations of the 4-point OPC and 3-point OPC3, are described in chapters 3 and 4.

1.3.2 OPCA for developing multi-scale/coarse-grain molecular dynamics simulation: insights into the structure of million-atom chromatin fiber

OPCA can be employed to represent large charge distributions with only a few point charges. This capability of OPCA is very useful for developing multi-scale/coarse-grain methods for molecular dynamics simulations, especially where analytic expressions and the simplicity of the algorithms is key. In this work, we combine OPCA with the $\sim nlogn$ Hierarchical Charge Partitioning (HCP) approximation [10] to present a new multi-scale, yet fully atomistic, approach (called GB-HCPO) that performs MD simulations orders of magnitude faster than traditional simulations. The HCP approximation partitions the biomolecular structures into multi-level hierarchical components based on the natural organization of biomolecules, as illustrated in Figure 5.1 – atoms (level 0), nucleic and amino acid groups (level 1), protein, DNA and RNA subunits (level 2), complexes of multiple subunits (level 3), and higher level structures such as fibres and membranes. OPCA approximates the charge distribution for each of these components by a small number of point charges so that the low order multipole moments of these components are optimally reproduced. The new mult-scale approach uses the full set of atomic charges to compute interactions between nearby atoms, while approximating interactions with distant components using the smaller set of charges. As a

result, the algorithmic complexity of the calculations of electrostatic interactions is reduced from N^2 (i. e. when the interactions for full set of atomic charges are calcualted) to $NlogN$, where N is the number of atoms, leading to significant speed up in the computations.

We show that the multi-scale approach can be over 2 orders of magnitude faster than the traditional simulations without any approximations. We demonstrate the significance of the new method by applying it to a bio-medically relevant application of interest to a larger experimental and theoretical community. We will use the new multi-scale method to construct an energetically realistic, atomistic model of the structure of a bio-medically relevant example: a $\sim$ 1 million-atom 30 nm Chromatin fiber. The chromatin fiber represents the second level of DNA compaction in cells[134]; modifications to N-terminal tails of the histone proteins that make up the fiber core are known to regulate DNA accessibility and affect vital process such as gene expression[83]. However, due to its large size, only low-resolution (cryo-EM) experimental structures of the fiber are available[170], its atomistic details including the tails, are unknown. Computational studies investigating the organization of chromatin fiber have typically used coarse-grain simulations. Such simulations use customized, relatively unproven, force fields, and fail to elicit the finer details of the atom level structure. In this work, we use the proposed multi-scale approach to perform a fully atomistic simulation of the million-atom chromatin fiber, starting from an existing atomistic model[215] consistent with cryo-EM data. The atomistic details of the equilibrated structure suggests important structural details consistent with experimental results.

Details of the proposed multi-scale approach along with its application in the study of chromatin fiber are described in chapter 5.

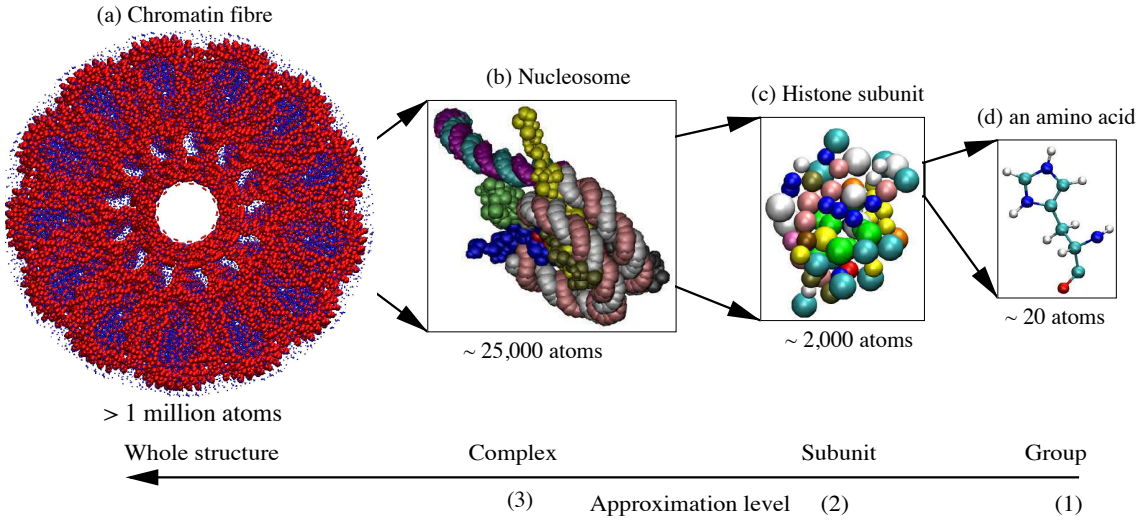


Figure 1.4: Multi-level hierarchical partitioning of a chromatin fibre based on its natural structural organization: (a) The fibre is made up of 40 nucleosome complexes. The individual nucleotide groups in the fibre are shown in red beads and amino acid groups as grey beads. (b) Each complex (level 3) is made up of 13 subunits with the segments of DNA linking nucleosome complexes being treated as separate subunits. A complex is shown here with each subunit represented in a different color. (c) Each subunit (level 2) is made up of 49-142 groups. The linker histone subunit is shown here with the groups colored by the type of amino acid. (d) Each group (level 1) is made up of 7-32 atoms (level 0). A histidine amino acid group is shown here with atoms represented as small spheres and covalent bonds between the atoms represented as links. The atoms are colored by the type of atom. The total fibre consists of approximately 1160000 atoms. The fibre was constructed as described in Wong et. al.[214]. The images were rendered using VMD [89]. For clarity, only 10 of the 13 subunits are shown in (a) and (b).

Chapter 2

Optimal point charge approximation: the concept, basic principles and analytical expressions

This chapter is adapted from PLoS ONE 8(7): e67715, 2013.

2.1 Overview

We propose an approach for approximating electrostatic charge distributions that optimally represents the original charge distribution with a smaller number of point charges. By construction, the proposed optimal point charge approximation (OPCA) retains many of the useful properties of point multipole expansion, including the same far-field asymptotic behavior of the approximate potential.

A general framework for numerically computing OPCA, for any given number of approximating charges, is described. We then derive a 2-charge practical point charge approximation, PPCA, which approximates the 2-charge OPCA via closed form analytical expressions, and test the PPCA on a set of charge distributions relevant to biomolecular modeling. We measure the accuracy of the new approximations as the RMS error in the electrostatic potential relative to that produced by the original charge distribution, at a distance $\sim 2\times$ the extent of the charge distribution – the mid-field. The error for the 2-charge PPCA is found to be on average 23% smaller than that of optimally placed point dipole approximation, and comparable to that of the point quadrupole approximation. The standard deviation in RMS error for the 2-charge PPCA is 53% lower than that of the optimal point dipole approximation, and comparable to that of the point quadrupole approximation.

We also calculate the 3-charge OPCA for representing the gas phase quantum mechanical

charge distribution of a water molecule. The electrostatic potential calculated by the 3-charge OPCA for water, in the mid-field (2.8 Å from the oxygen atom), is on average 33.3% more accurate than the potential due to the point multipole expansion up to the octupole order. Compared to a 3 point charge approximation in which the charges are placed on the atom centers, the 3-charge OPCA is seven times more accurate, by RMS error. The maximum error at the oxygen-Na distance (2.23 Å) is half that of the point multipole expansion up to the octupole order.

Keywords: point charge approximation; multipole; electrostatics; water models; coarse-graining

2.2 Introduction

Point multipole expansions are widely used to gain physical insight by providing a simplified expression for a complex distribution of sources of potential fields, such as electrostatic potential due to a charge distribution. Many familiar physics concepts are introduced using the framework of point multipoles because point multipoles provide a means of decoupling the underlying features of a source distribution from the observation point. Furthermore, since each successive term in the multipole expansion decays more rapidly with distance than the previous term, the impact of high order terms becomes small in the far-field, i.e. at distances R such that $R \gg R_0$, where R_0 is the distance of the furthest charge from the expansion center. This property has allowed the point multipole expansion to simplify many practical calculations. For example, algorithms such as the fast multipole [76], local reaction field[125], and fitted point multipole (FPM)[43] methods, use point multipoles to reduce the computational complexity of calculating pairwise interactions between large charge distributions. However, at distances not much larger than R_0, the accuracy of the low order point multipole approximation deteriorates quickly as one approaches the charge distribution, necessitating introduction of higher order terms. This, in turn, may lead to cumbersome algebra and the need to introduce further approximation[90]. Since, in practice, the potential often needs to be calculated in regions where the assumption $R \gg R_0$ does not hold, the point multipole expansion with only one or two lowest order terms may be suboptimal for some practical calculations. For example, in atomistic molecular simulations, amino acids interacting inside a single protein are often only several (1-5) times R_0 apart. For a typical amino acid group within a folded protein such as lysozyme $R_0 \approx 5$ Å, and the distance between amino acid groups ranges from $1R_0$ to $10R_0$. The value of R_0 and the distance at which the potential due to the charge distribution is evaluated, is of course problem dependent. Furthermore, compared to point charge approximations, it is generally more difficult to implement point multipole approximations into existing molecular modeling software, especially for commonly used implicit solvent models.

Arguably, one of the most successful point multipole based approximations is the fast multipole method [76]. The fast multipole method partitions the system into a hierarchical set of

cubic lattices. Electrostatic interactions between charges within a lattice and in neighboring lattices (the near-field) are treated exactly, while a truncated multipole expansion is used for electrostatic interactions due to atoms in the more distant lattices (the mid- and far-field). The size of the lattices used in the multipole expansion varies, with a larger lattice size being used for more distant lattices [34, 31, 121]. This technique reduces the complexity of the computation of pairwise interaction from $O(N^2)$ to less than $O(N\log(N))$, where N is the number of interacting particles[76]. Many improvements to the original technique have been made [77, 31, 177, 40, 219]. Overall, the fast multipole method has the advantage of lower computational complexity compared to the full pairwise computation, and has a well defined error bound [76]; the method is used in many areas of physics. However the fast multipole method has not been widely adopted in biomolecular simulations, most likely due to its algorithmic complexity, and the discontinuities in calculated potential inherent in the method.[27, 163] The local reaction field[125] and fitted point multipole (FPM)[43] methods have also not been widely utilized in the context of biomolecular modeling, again most likely due to their algorithmic complexity.

Here we investigate an alternative to the point multipole expansion for approximating charge distributions, which we call optimal point charge approximation (OPCA). Unlike the fast multipole method, which uses a set of point multipoles to represent the original charge distribution, the OPCA approximates a charge distribution using a given number of point charges. These point charges are chosen so that they optimally reproduce the lowest order multipoles in the expansion of the original distribution. Since OPCAs have a finite size, as opposed to being point-like, they may provide better representation of the original spatially extended charge distribution than a single-center truncated point multipole expansion. In particular, a more accurate representation of the potential in the mid-field may be expected. We prove below that the 1-charge and 2-charge OPCAs are at least as accurate as the equivalent order point multipole approximations, i.e. the point monopole and dipole approximations. Throughout this work we refer to point monopole, dipole, and quadrupole approximations as the truncated point multipole expansions upto the monopole, dipole, and quadrupole order, respectively.

We show that in general it is always possible to numerically determine the OPCA, however, for many practical applications, such as molecular dynamics simulations, analytical expressions are needed. Although it is not always possible to derive a practical analytical expressions for OPCAs, in certain cases we show that reasonable, robust and fairly simple approximations to the OPCAs can be derived, which we refer to as the practical point charge approximation (PPCA). The 2-charge OPCA is one such case for which a practical analytical expression is not readily evident for arbitrary charge distributions. For this case we have derived an approximation to the OPCA, the 2-charge PPCA. In what follows we evaluate the accuracy of our approximations for a set of charge distributions relevant to biomolecular modeling at a distance $R \approx 2R_0$. The accuracy at such distances is most critical for multiscale approximations[47, 119] such as the hierarchical charge partitioning method [12]. For smaller distances, multiscale approximations generally use the exact charge distribution in

their computations. From a practical standpoint, PPCAs may also be easier than the fast multipole protocol to implement in applications that already utilize point charges, i.e. in many molecular dynamics packages [157, 175].

The rest of this work is organized as follows. We first review the multipole expansion concept to orient the reader and provide a convenient notational reference. Next, we describe the theoretical basis for the optimal point charge approximation. We then use this theoretical formalism to derive closed-form expressions for the optimal and the practical point charge approximations for the 1- and 2-charge cases. The accuracy of the 2-charge PPCA was evaluated for a practical application relevant to biomolecular modeling. We also calculated the 3-charge OPCA for approximating a quantum mechanical charge distribution for a water molecule; the resulting OPCA reproduces the electrostatic potential in the mid-field with greater accuracy than the point octupole expansion. Potential uses and future work are discussed in "Conclusions".

2.3 Multipole Expansion

Here we will give a brief overview of the formalism of the point multipole expansion. Since many practical applications, such as molecular dynamics simulations, use point charges, for notational simplicity we will consider discrete charge distributions, but our main results also hold for continuous distributions.

Consider a set of N point charges q_n $(n = 1, 2, ..., N)$ located at positions $\boldsymbol{r_n}$ around some chosen origin. Then the potential $\Phi(\boldsymbol{R})$, of this distribution at a point $\boldsymbol{R}$ from that origin is given by the familiar Coulomb potential

$$\Phi(\boldsymbol{R}) = \frac{1}{4\pi\epsilon_0} \sum_{n=1}^{N} \frac{q_n}{||\boldsymbol{R} - \boldsymbol{r_n}||} \tag{2.1}$$

For distances $R > R_0$ where $R = ||\boldsymbol{R}||$ and $R_0 = max(||\boldsymbol{r_n}||)$, a Taylor series expansion of the potential above gives the classic multipole expansion. In Cartesian coordinates we obtain

$$\begin{aligned} \Phi(\boldsymbol{R}) &= \frac{1}{4\pi\epsilon_0} \left(\frac{1}{R} q + \frac{1}{R^2} \sum_{i=x,y,z} \hat{R}_i \boldsymbol{p_i} + \frac{1}{R^3} \sum_{i,j=x,y,z} \hat{R}_i \hat{R}_j Q_{ij} \right. \\ &\quad \left. + \frac{1}{6} \frac{1}{R^4} \sum_{i,j,k=x,y,z} \hat{R}_i \hat{R}_j \hat{R}_k O_{ijk} + \dots \right) \end{aligned} \tag{2.2}$$

where

$$q = \sum_{n=1}^{N} q_n \tag{2.3}$$

$$\boldsymbol{p_i} = \sum_{n=1}^{N} q_n r_{n,i} \tag{2.4}$$

$$Q_{i,j} = \frac{1}{2}\sum_{n=1}^{N} q_n\Big(3r_{n,i}r_{n,j} - (r_n)^2\delta_{ij}\Big) \tag{2.5}$$

$$O_{i,j,k} = \sum_{n=1}^{N} q_n\Big(15r_{n,i}r_{n,j}r_{n,k} - 3(r_n)^2(r_{n,i}\delta_{jk} + r_{n,j}\delta_{ik} + r_{n,k}\delta_{i,j})\Big) \tag{2.6}$$

$q, \boldsymbol{p}, Q, O$ are known as the monopole, dipole, quadrupole and octupole moments respectively, $\hat{R}_i, \hat{R}_j, \hat{R}_k$, with $i, j, k = x, y, z$, are the unit vectors along the x, y, or z coordinates, and δ_{ij} is the Kronecker delta. The multipole moments are symmetric tensors where the lowest order non-vanishing multipole is origin independent.

2.4 The optimal point charge approximation (OPCA)

For a given set of N original charges q_n $(n = 1, 2, ..., N)$, we want to determine the position and magnitude of K point charges $\bar{q}_k$ $(k = 1, 2, ..., K < N)$ such that the potential due to these smaller number of point charges, $\bar{\Phi}(\boldsymbol{R})$ best approximates the potential of the original distribution, $\Phi(\boldsymbol{R})$. Our criterion for the "best approximation" is as follows: the optimal point charge approximation (OPCA) minimizes the error in the multipole expansion for the K point charges relative to the multipole expansion for the original distribution of N charges. The precise error metric is defined below.

2.4.1 The Error Metric

Determining the best representative charge distribution is contingent upon the definition of the error metric used. In general, we are concerned with obtaining the best representation of the original potential at any arbitrary point in space outside the distribution. Thus, for the error metric, Δ, one typically chooses the root mean square (RMS) of the error in potential over some volume V (or surface) excluding the volume V_0 containing the charge distribution being approximated, i.e.

$$\Delta^2 = \frac{1}{V \notin V_0} \int\limits_{V \notin V_0} \left|\Phi(\boldsymbol{R}) - \bar{\Phi}(\boldsymbol{R})\right|^2 dV \tag{2.7}$$

In principle, one can derive the optimal charge placement $\{\bar{q}_k, \bar{\boldsymbol{r}}_{\boldsymbol{k}}\}$ by minimizing the integral given in Eq. (2.7) with respect to the values of the new charges, $\{\bar{q}_k\}$ and their positions $\{\bar{\boldsymbol{r}}_{\boldsymbol{k}}\}$. However, as the number of charges in the representative distribution grows, this equation can be expensive to minimize numerically, let alone to find closed-form analytic expressions for the placement and magnitude of the charges composing the representative distribution. In addition, the choice of the integration volume is somewhat arbitrary. Furthermore – and, perhaps, most importantly – charges chosen in this manner are not guaranteed to have the same multipole moments as the original distribution [180]. This can lead to misinterpretation of the properties of the distribution and, potentially, to unphysical results. At the very least, we would like the new approximate representation to inherit the same asymptotic behavior of the corresponding point multipole expansion of the same order, but with greater accuracy expected from a spatially extended distribution that can better mimic the original charge distribution.

To simplify the problem, we recast Eq. (2.7) in spherical coordinates and consider the error inside a spherical shell centered on the chosen multipole expansion center, and with arbitrary outer radius $\widetilde{R} > R_0$, where R_0 is defined as before, i.e. the distance from the expansion center to the outermost point charge. The error metric now becomes

$$\Delta^2 = \frac{3}{4\pi(\widetilde{R}^3 - R_0^3)} \int_{R_0}^{\widetilde{R}} \int_0^{2\pi} \int_0^{\pi} \left|\Phi(\boldsymbol{R}) - \bar{\Phi}(\boldsymbol{R})\right|^2 R^2 \sin(\theta) d\theta d\phi dR \tag{2.8}$$

where θ and ϕ are the usual spherical coordinate inclination and azimuth angles.

In spherical coordinates, the multipole expansion is given by

$$\Phi(\boldsymbol{R}) = \frac{1}{\epsilon_0} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{m=\ell} \frac{1}{2\ell+1} \frac{Y_\ell^m(\theta,\phi)}{R^{\ell+1}} q_\ell^m \tag{2.9}$$

where Y_ℓ^m are the standard spherical harmonics, $*$ denotes the complex conjugate, q_ℓ^m are the spherical multipole moments, and ℓ is the multipole order.

$$q_\ell^m = \sum_{n=1}^{N} q_n r_n^\ell Y_\ell^m(\theta_n, \phi_n)^* \tag{2.10}$$

Using this expansion as our error metric, Eq. (2.8), becomes

$$\Delta^2 = \frac{3}{4\pi(\widetilde{R}^3 - R_0^3)} \int_{R_0}^{\widetilde{R}} \int_0^{2\pi} \int_0^{\pi} \left| \frac{1}{\epsilon_0} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{m=\ell} \frac{1}{2\ell+1} \frac{Y_\ell^m(\theta,\phi)}{R^{\ell+1}} (q_\ell^m - \overline{q_\ell^m}) \right|^2 R^2 \sin(\theta) d\theta d\phi dR \tag{2.11}$$

where q_ℓ^m and $\overline{q_\ell^m}$ are the spherical moments of the original and representative charge distributions respectively. Since the spherical harmonics are orthonormal, Eq. (2.11), can be

further simplified to the following form [162],

$$\Delta^2 = \frac{3}{4\pi\epsilon_0(\widetilde{R}^3 - R_0^3)} \int_{R_0}^{\widetilde{R}} \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2 R^{2\ell}} \sum_{m=-\ell}^{m=\ell} |q_\ell^m - \overline{q_\ell^m}|^2 dR \tag{2.12}$$

2.4.2 Calculating the optimal point charge approximation

The position and magnitude of the representative point charges in OPCA for a given order K are calculated by sequentially minimizing each term in the error expansion Eq. (2.12), starting with the lowest order (monopole) term. From the structure of Eq. (2.12) we see that minimizing the difference between the successive multipole moments of the original N-charge and the optimal K-charge distributions is equivalent to minimizing the total error in electrostatic potential. Note that the procedure does not depend on the parameter $\widetilde{R}$, and thus the method does not require explicit integration over a given region. This removes a degree of arbitrariness in defining the "error surface" inherent in several other methods currently used in practice. The use of the multipole expansion as reference allows for the sought after distinct separation of terms by the rate at which they decrease as a function of R, i.e. the monopole term falls off as $\frac{1}{R}$, the dipole falls off as $\frac{1}{R^2}$, etc. A representation that makes terms up to order l in Eq. (2.12) equal to zero will produce total error whose leading term falls off as $1/R^{l+1}$. For $K = N$, which is the number of charges in the original distribution, the OPCA exactly reproduces the electrostatic potential due to the original distribution. This is in contrast to the point multipole expansions, which generally require an infinite number of terms to exactly reproduce a given charge distribution.

Note that minimizing the error metric, Eq. (2.12), minimizes the error in electrostatic potential for the far-field where $R \gg R_0$, but not necessarily in the mid-field. In the mid-field ($R \approx 2R_0$), the contribution due to higher order terms in the multipole expansion can, in principle, be greater than the contribution due to lower order terms. Therefore, minimizing the lowest order terms in the expansion error does not guarantee minimization of the total error in the mid-field: these errors are investigated below for charge distributions most relevant to biomolecules. In the following analysis, for convenience, we have dropped the $1/4\pi\epsilon_0$ factor in Eq. (2.2) and (2.12) and switched from SI to atomic units.

2.5 Analytical expressions for 1- and 2-charge OPCAs

The minimization of the error metric in Eq. (2.12), which is required to define an OPCA, can be done numerically for any given number of K representative point charges. A numerical procedure for calculating the OPCA representation may be particularly useful in situations where the charge distributions are relatively static and thus the optimal representation does

not need to be recalculated. For example, during restrained molecular dynamics simulations, components of the molecule may not move. The OPCA for these components do not need to be recalculated. For applications where the OPCA needs to be recalculated frequently, such as in unrestrained molecular dynamics simulations, one would like to find closed-form analytical expressions that can be used to compute OPCAs at a reduced computational cost and provide derivatives for force calculations.

In the following sections, we apply the general framework developed above to derive simple analytical expressions for the 1-charge OPCA. Note that the 1-charge OPCA is only applicable to a charge distribution with a non-zero net charge, since the monopole moment for a neutral charge distribution is zero. The 2-charge case is more complex: the optimal point charge approximation results in imaginary charge values for some charge distributions (see Eq. (2.22) below), and can not be cast in a closed-form formula for some other distributions. Therefore, we derive more practical analytical expressions that approximate the 2-charge OPCA with a reasonable accuracy.

2.5.1 1-charge optimal point charge approximation for charged distributions

By definition, the 1-charge OPCA consists of a single charge. As long as the charge has magnitude $\bar{q} = \sum_{n=1}^{N} q_n$, i.e. is equal to the total charge of the original distribution, the monopole term of the error expansion, Eq. (2.12), will be zero. Now, the remaining parameter, namely the position of the charge, is chosen to minimize the dipole term in the error expansion. In this particular case, the dipole term can be made identically zero by solving

$$p_x - \bar{q} \cdot x = 0 \tag{2.13}$$
$$p_y - \bar{q} \cdot y = 0 \tag{2.14}$$
$$p_z - \bar{q} \cdot z = 0 \tag{2.15}$$

for x, y, z where p_x, p_y, p_z are the x, y, z components of the dipole moment $\boldsymbol{p}$ of the original distribution. Solving the above equations we have

$$\bar{q} = q \tag{2.16}$$
$$\bar{\boldsymbol{r}} = \frac{\boldsymbol{p}}{q} \tag{2.17}$$

So, a charge of magnitude $\bar{q}$ placed at $\bar{\boldsymbol{r}}$ (center of charge) defines the 1-charge OPCA.

2.5.2 2-charge optimal and practical point charge approximations

The 1-charge OPCA is the smallest set of point charges required to eliminate the monopole and the dipole term of the error expansion in Eq. (2.12) for systems with non-zero net

charge. However, an error reduction further than the dipole order is often desired for higher accuracy. In such cases, the 2-charge OPCA ($K = 2$) is the next step. In deriving an analytical expression for the 2-charge OPCA, the goal is to eliminate the error terms up to the dipole order in Eq. (2.12), and to minimize the quadrupole and, ideally, the next terms.

Due to important differences in the characteristics of charge distributions with zero and non-zero net charges, it is necessary to treat these two cases separately. For charged systems, the monopole and the dipole error terms are eliminated if two point charges with total charge equal to the original net charge are positioned so that their center of charge coincides with the center of charge of the original distribution. For uncharged systems, however, the monopole and dipole terms in the error expansion are eliminated when a pair of charges of equal magnitude but opposite sign are aligned with the direction of the original dipole moment. In other words, to eliminate the error terms up to the dipole in the charged case the location of the center of charge of the 2 charges is constrained, while in the uncharged case the direction of the dipole moment of the 2 charges is constrained. This leads to two different solutions for the two cases.

2-charge approximation for uncharged distributions

For net zero charge distributions, the optimal point charge approximation consists of two charges $\bar{q}_1 = \bar{q}$ and $\bar{q}_2 = -\bar{q}$ located at positions $\bar{\boldsymbol{r}}_{\mathbf{1}}$ and $\bar{\boldsymbol{r}}_{\mathbf{2}}$ respectively. Thus, it takes 7 parameters, q, and the x, y, z components of $\bar{\boldsymbol{r}}_{\mathbf{1}}$, and $\bar{\boldsymbol{r}}_{\mathbf{2}}$, to uniquely define a 2-charge OPCA. By setting

$$\bar{q}(\bar{\boldsymbol{r}}_{\mathbf{1}} - \bar{\boldsymbol{r}}_{\mathbf{2}}) = \sum_{n=1}^{N} q_n \boldsymbol{r}_{\boldsymbol{n}} \tag{2.18}$$

the dipole term in the error is zero. Now, we will rewrite the positions $\bar{\boldsymbol{r}}_{\mathbf{1}}$ and $\bar{\boldsymbol{r}}_{\mathbf{2}}$ in the following form:

$$\begin{aligned} \bar{\boldsymbol{r}}_{\mathbf{1}} &= \bar{\boldsymbol{d}} + \frac{\sum q_n \boldsymbol{r}_{\boldsymbol{n}}}{2\bar{q}} \\ \bar{\boldsymbol{r}}_{\mathbf{2}} &= \bar{\boldsymbol{d}} - \frac{\sum q_n \boldsymbol{r}_{\boldsymbol{n}}}{2\bar{q}} \end{aligned} \tag{2.19}$$

where $\bar{\boldsymbol{d}}$ represents the geometric center between the two charges of the OPCA. We can see that these positions satisfy relation (2.18) automatically. By writing the positions of the charges in this manner, we have divided the process of determining the remaining parameters which define the OPCA into two steps, namely, finding the optimal placement of the charges, $\bar{\boldsymbol{d}}$, and finding the optimal magnitude of the charge, $\bar{q}$. Note that finding the optimal charge value fixes the separation between the two charges, since the dipole moment of the representative distribution has been constrained to equal the original dipole moment.

The placement of the geometric center $\bar{\boldsymbol{d}}$ of the charges composing the 2-charge OPCA that

minimizes the quadrupole term of the error expansion, is given by

$$\bar{d}_k = \frac{2}{3p^2}\left(\sum_{i=x,y,z} Q_{ki}p_i - \left(\frac{\sum_{i,j=x,y,z} Q_{ji}p_ip_j}{4p^2}\right)p_k\right) \tag{2.20}$$

where $k = x, y, z$, the $\bar{d}_k$'s are components of $\bar{\boldsymbol{d}}$, p_i, p_j, p_k are the components of the dipole moment (Eq. (2.4)), and Q_{ki}, Q_{kj} are the components of the quadrupole moment (Eq. (2.5)). This optimal position, known as the center of dipole, was derived previously[162] for a different purpose, namely for matching point multipole expansions between different charge distributions. Now, unlike the point dipole approximation, the 2-charge OPCA has physical size and thus an additional parameter with which to further minimize the error with respect to the given potential. In other words, Eqs. (2.20) and (2.18), determine only 6 of the 7 parameters required to define the 2-charge OPCA. Since the quadrupole moment is the lowest order non-zero term remaining in the error expansion, by choosing the optimal charge value we want to further minimize the quadrupole term in the error. However, for any charge value $\bar{q}$ an OPCA placed at the center of dipole has no quadrupole moment as can be seen by setting $N = 2$, substituting the center of dipole, Eq. (2.20), and $q_1 = -q_2 = \bar{q}$ into Eq. (2.19), and then substituting these variables into Eq. (2.6). Thus, the quadrupole term in the error, Eq. (2.12) is unaffected by the choice of the charge magnitude $\bar{q}$, and the quadrupole term has already been globally minimized. Therefore, to uniquely define the charge $\bar{q}$, we follow the OPCA procedure and globally minimize the next term in the error expansion, namely the octupole term. Specifically, if we consider the $\ell = 3$ term of Eq. (2.12), using the connection formula from spherical multipoles to Cartesian multipoles we can compute

$$\sum_{i,j,k=x,y,z} \frac{\partial}{\partial \bar{q}} \left(O_{ijk} - \overline{O_{ijk}}\right)^2 = 0 \tag{2.21}$$

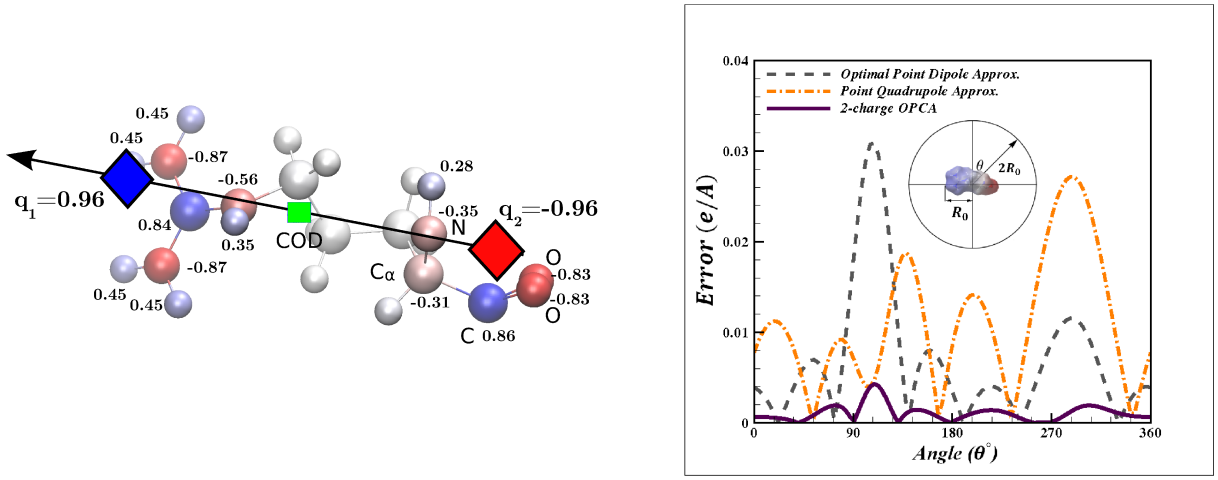
where O_{ijk} and $\overline{O_{ijk}}$ are components of the octupole moments, in Cartesian coordinates (Eq. (2.6)), of the original distribution and the 2-charge OPCA respectively, for an expansion computed about the center of dipole. By noting that $\overline{O_{ijk}}$ is a function of $\bar{q}$, we find that Eq. (2.21) is satisfied when $\bar{q} \to \infty$ or if the charge value is given by

$$\bar{q} = \sqrt{\frac{3p^6}{2\sum_{i,j,k=x,y,z} O_{ijk}p_ip_jp_k}} \tag{2.22}$$

Thus, Eqs. (2.19), (2.20) and (2.22) define the 2-charge OPCA for the net zero charge case (figure 2.1), i.e. defines the best placement of charges such that the error metric (Eq. (2.12)) is minimized.

In some cases, it is possible that

$$\left(\sum_{i,j,k=x,y,z} O_{ijk}p_ip_jp_k\right) \leq 0 \tag{2.23}$$



(a) **(b)**

Figure 2.1: Example of a 2-charge optimal point charge approximation (OPCA) for a sample charge distribution – a neutral amino acid (C-terminal arginine at physiological pH) , including the associated NH-CH-COO backbone atom. **(a)** The atomic partial charges are represented as spheres rendered using VMD[88]. The sphere colors range from red to blue representing the charge range of $-1e$ to $+1e$, where e is the atomic unit of charge. The charge values for charges $|q| > 0.2e$ are shown next to the atoms. As a visual reference, the backbone heavy atoms are labeled and covalent bonds are included in the figure. The green square represents the center of dipole (COD), with dipole moment p shown by the arrow. The two diamonds represent the two point charges, q_1 and q_2, of the OPCA. **(b)** Error in electrostatic potential for the 2-charge OPCA, point dipole, and point quadrupole with center of dipole as the expansion center. The error is calculated relative to the exact computation, on a circle at a distance $2R_0$, in the plane shown. Here, R_0 is the size of the charge distribution defined as the distance from its center of geometry to the outermost charge. The inset image shows the electrostatic surface potential rendered using GEM[73].

In this case, the charge given by Eq. (2.22) is imaginary. This situation occurs when the orientation of the dipole with respect to the octupole moment of the original charge distribution is such that increasing the distance between the charges of the 2-charge OPCA always increases the error. In this case, Eq. (2.21) is formally satisfied only for $\bar{q} \to \infty$. In a practical calculation, a 2-charge OPCA with inequality (2.23) is constructed by fixing the separation between the charges $||\bar{\boldsymbol{r}}_1 - \bar{\boldsymbol{r}}_2||$ to a small value (figure 2.2(a)), while increasing the OPCA charge accordingly to maintain the dipole moment of the original distribution (Eq. (2.18)). In this case, the 2-charge OPCA does not offer an accuracy advantage in the far-field over the optimal point dipole approximation, however, the 2-charge OPCA can always mimic the point dipole approximation to arbitrary precision and thus the two distributions will produce equivalent error. Thus, even if inequality (2.23) holds, the 2-charge OPCA

represents the optimal placement of two point charges and is at least as accurate as the point dipole approximation in the far-field where $R \gg R_0$. However, in the mid-field, such a charge placement may sometimes be slightly less accurate than the optimal point dipole approximation.

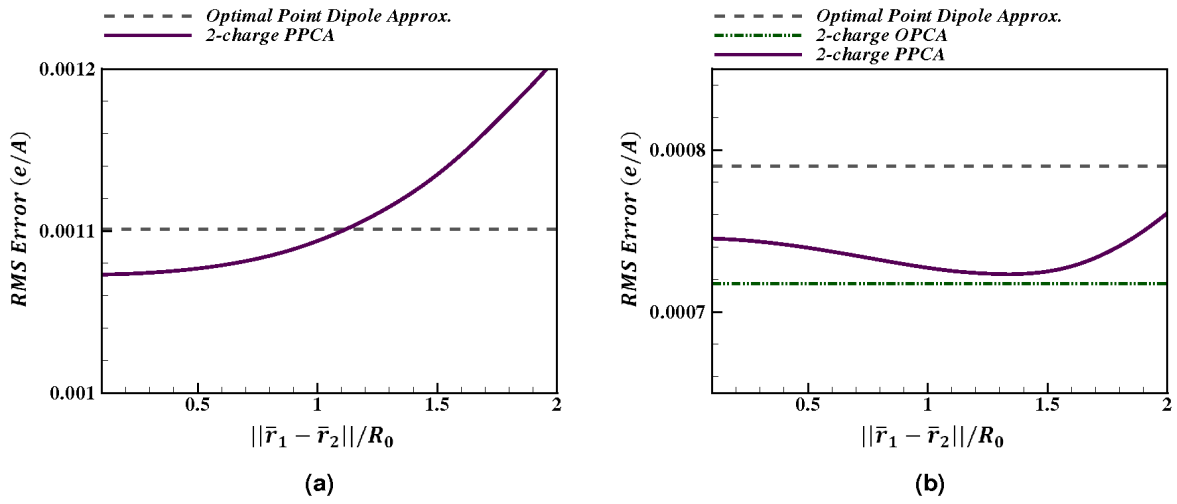


Figure 2.2: Accuracy of the 2-charge practical point charge approximation (PPCA) for charge distributions with a net zero charge described in the Practical Application section. Accuracy is calculated as the RMS error relative to the exact computation, at a distance of 10 Å ($\approx 2R_0$) from the center of geometry. RMS error for the 2-charge PPCA (Eq. (2.18)) is shown as a function of the distance between the two charges of the PPCA $||\bar{\boldsymbol{r}}_\mathbf{1} - \bar{\boldsymbol{r}}_\mathbf{2}||$. The RMS error for the 2-charge PPCA is compared to that of the point dipole approximation with an optimal center of expansion. **(a)** Cases where Eq. (2.23) is true. **(b)** Cases where Eq. (2.23) is false. This figure also includes the 2-charge optimal point charge approximation (Eq. (2.22)) for comparison. Connecting lines are shown to guide the eye.

For the charge distributions described in the Practical Application section below, we found that setting $||\bar{\boldsymbol{r}}_\mathbf{1} - \bar{\boldsymbol{r}}_\mathbf{2}|| = R_0/4$ to determine the value of $\bar{q}$ in Eq. (2.18), instead of the much more complex Eq. (2.22), results in the electrostatic potential that is on average within 4% of the optimal $K = 2$ OPCA solution (figure 2.2(b)). Therefore, for practical applications, it may be computationally more efficient to use an empirically determined value for $||\bar{\boldsymbol{r}}_\mathbf{1} - \bar{\boldsymbol{r}}_\mathbf{2}||$ even for cases where the inequality (2.23) is not satisfied and a true optimal placement of 2-charge can be found via Eqs. (2.20) and (2.22). Another important practical consideration is that the contribution of each term in the error expansion, Eq. (2.12), is smaller than the previous term only if $R \gg R_0$. This is not necessarily true in the mid-field where $R \approx 2R_0$. For example, for some charge distributions, the center of dipole $\mathbf{d}$ may be located at $R \geq 2R_0$. For such cases, the error for the optimal point dipole approximation[74, 162] (point dipole approximation placed at the center of dipole) and the 2-charge OPCA can become large at

$R \approx 2R_0$. To ensure that our 2-charge approximation is reasonably accurate in the mid-field for such cases, we introduce an addition condition: the optimal charge positions are restricted to be within the 1.5 times the maximum extent of the original charge distribution R_0, from the center of geometry.

Thus, for distributions with zero net charge, the 2-charge practical point charge approximation (PPCA), which approximates the 2-charge optimal point charge approximation (OPCA), is determined through the following 4 steps: (i) The two point charges comprising the PPCA are placed such that their center of geometry coincides with the center of dipole (Eq. (2.20)). (ii) The separation between the charges is fixed at $||\bar{\boldsymbol{r}}_{\mathbf{1}} - \bar{\boldsymbol{r}}_{\mathbf{2}}|| = R_0/4$. (iii) The position and magnitude of two charges are then determined by Eq. (2.18) and (2.19). (iv) In the rare cases when the point charges for the PPCA are at a distance greater than $1.5R_0$ from the center of geometry, the center of dipole is shifted towards the center of geometry so that the point charges lie within $1.5R_0$. The constants in conditions (ii) and (iv) above were determined empirically for charge distributions relevant to biomolecular modeling, see the Practical Application section below

2-charge approximation for charged distributions

The 2-charge OPCA consists of two charges $\bar{q}_1$ and $\bar{q}_2$. By setting

$$q = \bar{q}_1 + \bar{q}_2 \tag{2.24}$$

where q is the net charge of the original distribution, the monopole order error term in Eq. (2.12) becomes zero. If we set the center of charge as the center of expansion for the point dipole approximation, and choose the charges for the 2-charge OPCA such that the center of charge for the OPCA coincides with the center of charge for the original distribution, then

$$\mathbf{p} = \bar{q}_1 \cdot \bar{\mathbf{r}}_1 + \bar{q}_2 \cdot \bar{\mathbf{r}}_2 = 0 \tag{2.25}$$

where $\bar{\mathbf{r}_1}$ and $\bar{\mathbf{r}_2}$ represent the position vectors for charges 1 and 2 respectively, of the 2-charge OPCA. Note that, with the choice of center of charge as the multipole center of expansion, the $K = 2$ OPCA is guaranteed to be at least as accurate as the point dipole approximation, as measured by the error metric defined by Eq. (2.12). Thus, to simplify the derivations, we will use the center of charge as the origin for the coordinate system.

The next non-vanishing error term to be minimized is the quadrupole, i.e.

$$min(\sum_{i,j=x,y,z} \left(Q_{ij} - \overline{Q}_{ij}\right)^2) \tag{2.26}$$

where $\overline{Q}_{ij}$ and Q_{ij} are the quadrupole moments of the OPCA and the original charge distribution respectively, about the center of charge. The quadrupole tensor defines a unique, orthogonal set of principal axes in three-dimensional space. Since the two point charges of

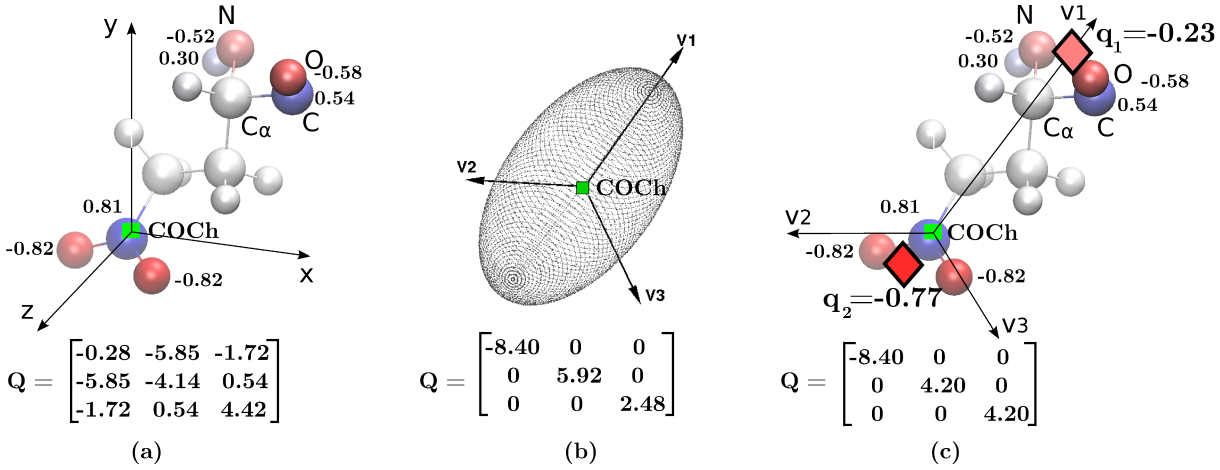


Figure 2.3: Illustration of a 2-charge practical point charge approximation (PPCA) for a sample charge distribution with non-zero net charge (a glutamic acid group within a protein with net charge $= -1e$, where the group includes the associated NH-CH-CO backbone atoms). **(a)** The original charge distribution with its quadrupole tensor (Eq. (2.5)) shown below. The atomic partial charges are represented as spheres rendered using VMD[88]. The sphere colors range from red to blue representing the charge range of $-1e$ to $+1e$. The charge values for charges $|q| > 0.2e$ are shown next to the atoms. As a visual reference, the backbone heavy atoms are labeled and covalent bonds are included in the figure. The green square shows the center of charge (COCh). **(b)** The principal axes, **v1, v2, v3** of the original charge distribution with the center of charge as origin (green square). Its quadrupole tensor, with the coordinate system aligned to the principal axes (Eq. (2.27)), is shown below. Here **v1** is the principal axis with the largest principal value. Analogous to the concept of ellipsoid of inertia in Mechanics used to characterize mass distribution, an "ellipsoid of charge" can be imagined here that helps visualize the charge distribution characterized by the quadrupole tensor. **(c)** The 2-charges of the PPCA (red diamonds) are placed such that the quadrupole moment for the PPCA equals the component of the quadrupole moment for the original charge distribution along **v1**. The quadrupole tensor produced by the 2-charge PPCA, with the coordinate system aligned to the principal axes, is shown below. The values of charges are in atomic units (e), and $e.\text{Å}^2$ is the unit for the quadrupole tensors.

the 2-charge OPCA define a single line, the quadrupole potential can be expected to be best approximated by the 2-charge OPCA with the charges positioned along the principal axis that has the largest absolute principal value (figure 2.3). Since the quadrupole tensor $\mathbf{Q}$ is a real symmetric matrix, its principal values can be determined by the eigenvalue decomposition

$$\mathbf{Qv} = \lambda\mathbf{v} \tag{2.27}$$

where λ is a principal value (eigenvalue) with the corresponding principal axis (eigenvector)

$\mathbf{v}$. Let λ be the largest principal value. Then, by placing the 2-charge OPCA along $\mathbf{v}$, and setting the component of quadrupole moment for the 2-charge OPCA along the principal axis $\mathbf{v}$ equal to the largest principal value λ for the original distribution, we obtain from Eq. (2.5) for the quadrupole moment

$$\lambda = \bar{q}_1\bar{r}_1^2 + \bar{q}_2\bar{r}_2^2 \tag{2.28}$$

where $\bar{r}_1$ and $\bar{r}_2$ are the magnitude of the $\bar{\mathbf{r}}_\mathbf{1}$ and $\bar{\mathbf{r}}_\mathbf{2}$ vectors with center of charge as the origin.

Substituting the values for $\bar{q}_2$ and $\bar{r}_2$ from Eq. (2.24) and (2.25) respectively, we arrive at:

$$\bar{q}_1 = \frac{q\lambda}{\lambda + q\bar{r}_1^2} \tag{2.29}$$

The above equation does not provide a unique solution since $\bar{r}_1$ is still unknown. Minimizing the error in the next order multipole term in Eq. (2.12), the octupole moment, results in quartic equations which may produce imaginary charge values. Therefore, for practical applications, as with the uncharged case, an empirical approximation may be more appropriate. Specifically, we set $\bar{r}_1 = \alpha R_0$ where α is an empirical parameter. Consider for example a typical charge distribution (a glutamic acid) from the sample charge distributions described in the Practical Application section below. For this charge distribution, figure 2.4 shows that in the mid field ($R = 2R_0$), with the choice of $\bar{r}_1 \approx 1.6R_0$ this practical point charge approximation (PPCA) is on average more accurate than the point dipole and point quadrupole approximations. For the representative sample charge distributions described in the next section, the PPCA was found to be the most accurate for $\bar{r}_1 = 1.5R_0$.

By placing the 2-charge PPCA along the principal axis with the largest principal value, we eliminate the error due to the largest component of the quadrupole tensor λ. Furthermore, since the quadrupole tensor is traceless, the other two principal values λ_a, λ_b in $\mathbf{Q}$ and $\bar{\lambda_a}, \bar{\lambda_b}$ in $\bar{\mathbf{Q}}$, are of the opposite sign to λ, and $|\lambda_a|, |\lambda_b|, |\bar{\lambda_a}|, |\bar{\lambda_b}| \leq |\lambda|$. Therefore, the error due to the other two components of the quadrupole tensor are reduced as well, i.e. $((\lambda_a - \bar{\lambda_a})^2 + (\lambda_b - \bar{\lambda_b})^2) \leq ((\lambda_a)^2 + (\lambda_b)^2)$. As an illustration, consider the example of figure 2.3(b): the error due to smaller components of the quadrupole tensor shown in figure 2.3(b) are smaller than the ones in figure 2.3(c), as $((5.92-4.20)^2+(2.48-4.20)^2) < (5.92^2+2.48^2)$.

Thus, for a charge distribution with net non-zero charge, the practical point charge approximation is determined by Eq. (2.29), with $r_1 = 1.5R_0$. The constant of 1.5 was empirically determined for the set of sample charge distributions described in the following section.

2.6 Practical Applications

We consider here two potential applications for the optimal and practical point charge approximations developed above – the approximation of atomic partial charge distributions

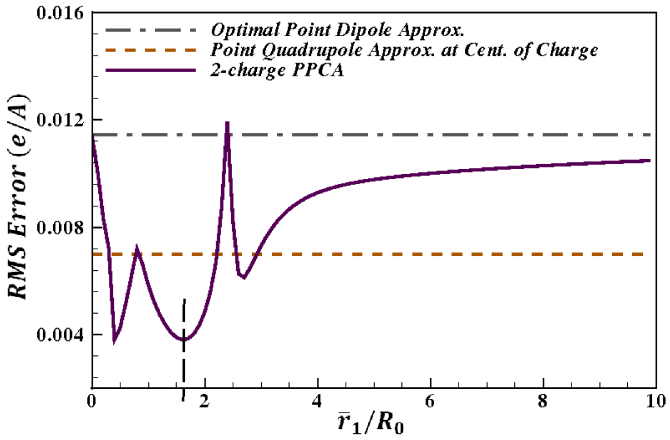


Figure 2.4: Accuracy of the 2-charge practical point charge approximation (PPCA) as a function of the distance $\bar{r}_1$ from the center of charge, for the sample charge distribution shown in figure 2.3. Accuracy is calculated as the RMS error, relative to the exact computation, at a distance of $2R_0$, where R_0 is the maximum extent of the charge distribution from the center of geometry. The point dipole and point quadrupole approximations with center of charge as the center of expansion are shown for comparison. The vertical dashed line represents the value $\bar{r}_1 \approx 1.6R_0$ that produces the lowest RMS error for the 2-charge PPCA in this case. Connecting lines are shown to guide the eye.

for amino acid groups within proteins, and the approximation of the charge distribution of water molecule.

2.6.1 Atomic level biomolecular modeling

Molecular modeling is commonly used to study the structure, function and activity of biological systems [64, 113, 172]. A common computational bottleneck in biomolecular modeling is the calculation of long-range electrostatic interactions: due to slow decay of these interactions with distance, simply ignoring them beyond a certain cut-off distance may lead to unacceptable accuracy loss [12, 10]. Multiscale approximations are one class of methods used to speed up these calculations[34, 121, 12], where near-field interactions are treated exactly, while an approximation of the charge distribution is used for mid- and far-field computations.

Since the error introduced by such approximations is generally very low in the far-field, understanding the mid-field error of such approximations, including ours, is most relevant. In the context of biomolecular modeling, we consider the lower bound of the mid-field to be no less than 2 times the extent of the charge distribution ($2R_0$); the mid-field for amino acid

groups is therefore greater than 5 Å.

We have applied the 2-charge practical point charge approximations (PPCA) developed above to the computation of electrostatic potential for a set of 1188 amino acid groups in five representative biomolecules that span a large range of sizes: a monomer from the virus capsid (Protein Databank (PDB) ID 1A6C) with 513 groups, the villin headpiece protein (PDB ID 1VII) with 36 groups, calcium switch protein (PDB ID 1UWO) with 91 groups, chaperonin GroEL (PDB ID 2EU1) with 524 groups, and myoglobin (PDB ID 1YMB) with 24 groups. The amino acid groups include their associated backbone atoms, NH-CH-CO, NH_2-CH-CO for the N-terminal groups, and NH-CH-COO for the C-terminal groups. Atomic partial charges were taken from the AMBER force field parameters [36]. The electrostatic potential was calculated in the mid-field (for two values: $R = 10$ Å $\approx 2R_0$ and $R = 15$ Å $\approx 3R_0$) where the approximation is likely to be least accurate. The electrostatic potential was computed at discrete points on a sphere of radius R, centered at the center of geometry. The spherical surface was discretized into 7200 grid points at which the electrostatic potential was calculated. The RMS error was calculated over all grid points and all the amino acid groups in the sample as $\sqrt{\sum^N |\Phi - \Phi_{ref}|^2 \Big/ N}$, where Φ and Φ_{ref} are the electrostatic potential calculated using the approximations and the reference (original) charge distributions, respectively, and N is the number of grid points. The 2-charge PPCA was compared to the optimal point dipole and the point quadrupole approximations. The center of expansion for the point dipole approximation for uncharged and charged distributions are chosen to be the center of dipole and center of charge, respectively, which are known to be optimal[162] for the corresponding point multipole expansions. For the point quadrupole approximation, we found that the choice of center of geometry as the center of expansions for uncharged cases, and the center of charge for charged cases, produced the most accurate result, on average, in the mid-field. Accordingly, we use these points as the expansion centers for the point quadrupole approximation.

In the mid-field ($R = 10$ Å $\approx 2R_0$) the RMS error (0.0053 ± 0.0030 e/Å) for the 2-charge PPCA is comparable to the point quadrupole approximation RMS error (0.0054 ± 0.0027 e/Å), and 23% less than the optimal point dipole approximation RMS error (0.0069 ± 0.0074 e/Å), for the charge distributions considered here, figure 2.5. On the other hand, when electrostatic potential is calculated at a distance $R = 15$ Å, the RMS error (0.00026 ± 0.00016 e/Å) for the 2-charge PPCA is 34% less than the optimal point dipole approximation RMS error (0.00039 ± 0.00026 e/Å), while being 53% higher than the point quadrupole approximation RMS error (0.00017 ± 0.00011 e/Å). These results reflect the fact that the 2-charge PPCA is always at least as accurate as the optimal point dipole approximation, whereas the PPCA can only try to minimize the error in the quadrupole term unlike the point quadrupole approximation, which eliminates the error in the quadrupole term. As the distance from the charge distribution increases, the accuracy of the multipole expansion, and, specifically, the accuracy of the point quadrupole approximation improves. This is evident from the errors at a distance $R = 15$ Å (figure 2.5(b)) which are an order of magnitude lower

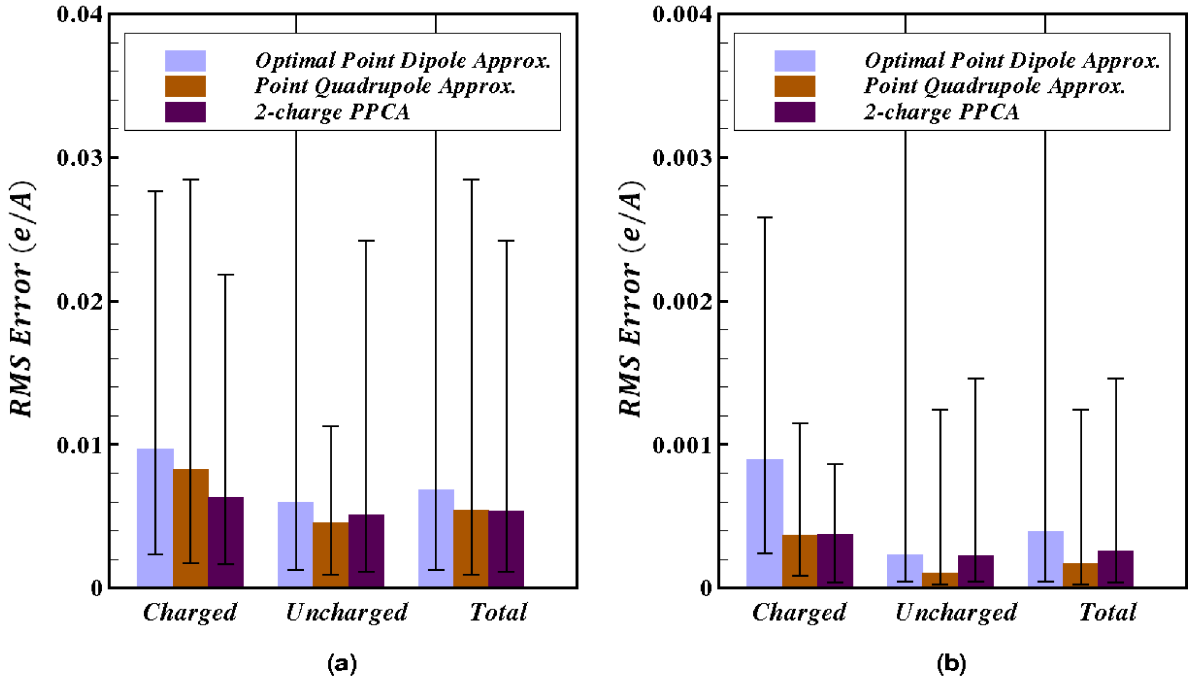


Figure 2.5: Accuracy of the 2-charge practical point charge approximation (PPCA) compared to that of the point dipole and point quadrupole approximations, for a sample set of charge distributions relevant to biomolecular modeling. Accuracy is calculated as the RMS error relative to the exact computation. **(a)** Error calculated at a distance of 10 Å$\approx 2R_0$ where R_0 is the maximum extent of the charge distribution from the center of geometry. **(b)** Error calculated at a distance of 15 Å $\approx 3R_0$ from the center of geometry. Error bars show the maximum and minimum absolute error. The upper values for the error bars that are cut off at the top are 0.14 and 0.006 in the left and right panels, respectively.

compared to the errors at a distance $R = 10$ Å(figure 2.5(a)). Note that the set of amino acid groups used here consist of approximately 20% charged and 80% uncharged distributions.

Figure 2.5(a) also shows that the 2-charge PPCA is on average significantly more accurate than the point dipole and quadrupole approximations for net non-zero charge distributions, compared to net zero charge distributions. Thus, the 2-charge PPCA should be significantly more accurate than the point dipole and quadrupole approximations for molecular structures that contain a significant portion of charged amino acids. And this is precisely the type of structures where the use of long-range cut-offs may lead to large errors. Also note that the standard deviation in RMS error for the 2-charge PPCA (0.0030 e/Å) is comparable to that of the point quadrupole approximation (0.0027 e/Å) and is less than half of that of the optimal point dipole approximation (0.0074 e/Å). Thus, the 2-charge PPCA is a considerably "tighter" approximation than the equivalent order optimal point dipole approximation. The higher standard deviation in RMS error for the point dipole approximation is primarily due

to the cases where the center of dipole, for charge distributions with zero net charge, falls outside the extent of the original charge distribution. In these cases the "optimal" center of expansion for the point dipole approximation can be very close to the point at which the electrostatic potential is approximated, resulting in large errors. This source of error is explicitly removed from the PPCA.

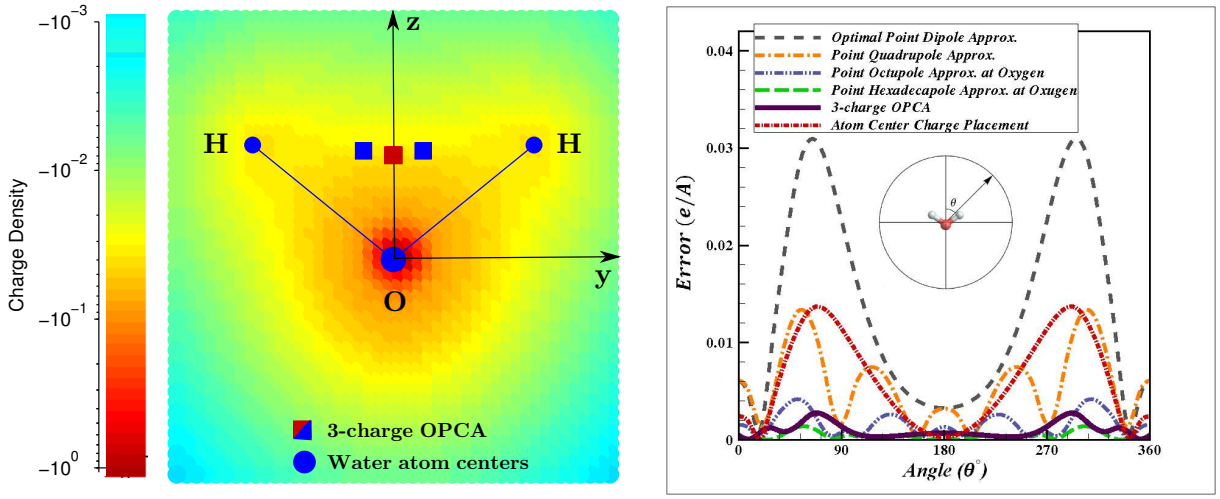
2.6.2 Optimal point charge approximation for water molecule

Water is critical for life[193, 18], and is one of the most extensively studied molecules[108, 115, 200, 50]. Accurate yet computationally efficient description of the solvent environment is essential for realistic biomolecular modeling. Commonly used simple fixed point charge models of water have achieved a reasonable compromise between accuracy and speed, but these are by no means perfect[79, 140]; the search for more accurate yet computationally facile models continues[90, 217, 206]. The ability of a given model to reproduce electrostatic properties of the highly polar water molecule is critical to success of the model[90]. Obviously, any reasonable model needs to account for the large dipole moment of water molecule in order to reproduce dielectric properties of the liquid state. But higher moments are important too: for example, one of the components of the water quadrupole tensor is large, and was shown to have strong effect on the liquid water structure seen in simulations[151]. The octupole order terms have also been shown to be of importance: for example, these affect water structure around ions[195]. An intricate interplay between the dipole, quadrupole and octupole moments gives rise[149] to the experimentally observed charge hydration asymmetry of aqueous solvation – strong dependence of hydration free energy on the sign of the solute charge. Thus, accurate yet computationally facile representations of the complex charge distribution of water molecule should be of interest.

As an illustration of the OPCA approach, we show here that the 3-charge OPCA can accurately reproduce a quantum mechanical charge distribution of the water molecule up to the octupole moment.

The specific charge density for the electron distribution of the water molecule used here (Fig. 2.6(a)) was determined by the CCSD method with aug-cc-pCVTZ basis set [53, 116, 216] at experimental equilibrium geometry in the gas phase. The electron charge density distribution was calculated for a box with side length of 4 Å and resolution of 0.05 Å. The resulting multipole moments of water molecule in the gas phase are comparable to available experimental values [42] (Table 2.1). We stress, however, that the specific charge distribution is used here only to illustrate the OPCA method and its capabilities; no claims regarding suitability of this distribution for simulation of liquid phase water [151] are made.

Since the water molecule is neutral, it can not be represented by a 1-charge OPCA. A 2-charge OPCA can accurately represent the dipole moment but not the quadrupole and octupole moments of the distribution, which are important for an accurate representation of water [149, 140]. Therefore, we calculate the 3-charge OPCA, as follows. In general, the 3-



(a) **(b)**

Figure 2.6: 3-charge optimal point charge approximation (OPCA) for water. **(a)** The quantum mechanical electron charge density is visualized by a light blue to red colormap representing the charge density range of 0 to -1 e per $0.05 \times 0.05 \times 0.05$ Å^3. The figure shows a 3 Å$\times$ 3 Å slice of the charge distribution in the y-z plane of the water atom centers. The origin is located at the center of the oxygen atom, the water atoms lay in the y-z plane, and the z-axis bisects the hydrogen atoms. The blue dots represent the water atom centers and the red and blue squares represent the 3 OPCA charges. The central OPCA charge has a value of $-26e$ and the other two are $13e$ each. **(b)** The error in electrostatic potential relative to the exact computation, calculated at $2 \times R_0 = 2.8$ Å from the oxygen atom, in the y-z plane. In this case R_0 is chosen to be 1.4 Å, the mean van der Waals radius of water[63], and $2 \times R_0$ approximates the distance between the oxygen atoms in two closest water molecules. For comparison, we show the error for the 4 lowest point multipole approximations as well as for a commonly used approximation which places point charges on atom centers. To match the dipole moment of the original charge distribution, the charge placed at the oxygen position equals $-0.64e$, while the charges on the hydrogen centers are $0.32e$ each. The same relative ordering of errors is seen in the x-z and x-y planes (not shown).

charge OPCA consists of three charges $\bar{q}_1, \bar{q}_2$, and $\bar{q}_3$, located at $\bar{\mathbf{r}}_1, \bar{\mathbf{r}}_2$, and $\bar{\mathbf{r}}_3$, representing 12 independent variables. But in the case of water any solution must respect the C_{2v} symmetry of the molecule, which reduces the number of independent variables to 10 (by assuming $\bar{q}_1 = \bar{q}_2$ and $\bar{z}_1 = \bar{z}_2$). Following the general procedure outlined in the "Calculating the optimal point charge approximation" section above, we first eliminate the monopole term in

the error expansion Eq. (2.12), by setting

$$q = \bar{q}_1 + \bar{q}_2 + \bar{q}_3 = 0 \tag{2.30}$$

where $q = 0$ is the monopole moment of the original charge distribution for water. Then, we eliminate the dipole term in the error expansion via

$$p_i = r_{1i}\bar{q}_1 + r_{2i}\bar{q}_2 + r_{3i}\bar{q}_3 \qquad i = x, y, z \tag{2.31}$$

where p_i are the x, y, or z component of the dipole moment and r_{1i}, r_{2i}, r_{3i} are the x, y, or z components of $\bar{\mathbf{r}}_1$, $\bar{\mathbf{r}}_2$ and $\bar{\mathbf{r}}_3$. Note that in the coordinate system standard for water molecules (figure 2.6(a)), p_z is the only non-zero component of the dipole moment. Finally, we eliminate the quadrupole term in the error expansion, Eq. (2.12), by setting

$$Q_{i,j} = \frac{1}{2}\left[\bar{q}_1\Big(3r_{1i}r_{1j} - (r_1)^2\delta_{ij}\Big) + \bar{q}_2\Big(3r_{2i}r_{2j} - (r_2)^2\delta_{ij}\Big) + \bar{q}_3\Big(3r_{3i}r_{3j} - (r_3)^2\delta_{ij}\Big)\right] \qquad i, j = x, y, z \tag{2.32}$$

where $Q_{i,j}$ are the terms in the quadrupole tensor of the original charge distribution. Note that in the coordinate system chosen (figure 2.6(a)), all off-diagonal terms in the quadrupole tensor are zero, i.e. $Q_{ij} = 0, i \neq j$. Thus, we are left with a total of 9 independent equations – 1 for the monopole q, 3 for the dipole terms p_x, p_y, and p_z, and 5 for the terms in the symmetric traceless quadrupole tensor, $Q_{xx}, Q_{yy}, Q_{xy}, Q_{xz}$, and Q_{xz} – to solve for 10 variables. This is an under-determined system of equations, leaving one additional variable. Solving the above set of equations results in the following solution, with $\bar{q}_1$ as the remaining variable.

$$\bar{q}_2 = \bar{q}_1 \tag{2.33}$$

$$\bar{q}_3 = -2\bar{q}_1 \tag{2.34}$$

$$\bar{x}_1 = \bar{x_2} = \bar{x_3} = \bar{y_3} = 0 \tag{2.35}$$

$$\bar{z}_1 = \bar{z}_2 = \frac{(Q_{zz} - Q_{xx})}{3p_z} + \frac{p_z}{4\bar{q}_1} \tag{2.36}$$

$$\bar{z}_3 = \frac{(Q_{zz} - Q_{xx})}{3p_z} - \frac{p_z}{4\bar{q}_1} \tag{2.37}$$

$$\bar{y}_1 = -\sqrt{\frac{(Q_{yy} - Q_{xx})}{3\bar{q}_1}} \tag{2.38}$$

$$\bar{y}_2 = +\sqrt{\frac{(Q_{yy} - Q_{xx})}{3\bar{q}_1}} \tag{2.39}$$

The value for $\bar{q}_1$ is computed numerically to minimize the octupole term in the error expansion, Eq. (2.12). The resulting OPCA charges are $\bar{q}_1 = 13e, \bar{q}_2 = 13e$, and $\bar{q}_3 = -26e$, located at (0, -0.16, 0.49), (0, 0.16, 0.49), and (0, 0, 0.47) Å, respectively, (Fig 2.6). [1] By

[1] The conversion factor of 0.2082 eÅ/debye is used to convert the multipole moments in table 2.1 to atomic units (eÅ, eÅ^2, etc.) used in Eq. (2.33) – (2.39)

construction, the multipole moments from the 3-charge OPCA and the quantum mechanical charge distribution are identical up to quadrupole order, as shown in (Table 2.1). The tight clustering of the 3 OPCA charges away from the oxygen nucleus is unexpected – in many commonly used water models the charges are placed on atom centers. Mathematically, the clustering results from minimizing the error at the octupole level, used to determine the value of q_1 in Eq. (2.33) - (2.39). For fixed dipole and quadrupole moments, the extent of the OPCA charge distribution is controlled by the octupole moment of the original charge distribution. The small distance between the opposite OPCA charges necessitates their large magnitude to ensure correct dipole moment of the charge distribution. Although the position and magnitude of the OPCA charges may appear unusual, the OPCA representation may be more accurate than the atom-centered alternative. For example, when we place the point charges on atom centers, and adjust the corresponding charge magnitudes so that the error at the dipole order is eliminated, the RMS error in electrostatic potential at 2.8 Å around the oxygen atom center is 0.0073 e/Å, which is many times larger than the RMS error for the 3-charge OPCA (0.0010 e/Å), as shown in Fig. 2.6(b). Note that the OPCA representation is designed to best approximate multipole moments of the original charge distribution; it remains to be seen whether this strategy leads to accurate reproduction of other physical properties of water. A water model based on the OPCA representation will be presented in a separate study.

Table 2.1: Multipole moments of a water molecule in the gas phase computed using quantum mechanical (QM) charge distribution, 3-charge OPCA, and the corresponding experimental values [42]. The coordinate system is that of Figure 2.6(a). Due to the symmetry, for the octupole tensor O_{xxz}=O_{xzx}=O_{zxx} and O_{yyz}=O_{yzy}=O_{zyy}. Components of multipole moments with a value of zero are not shown. 1 debye (D) = 0.2082 eÅ.

		QM (this work)	3-charge OPCA	Experimental
Dipole(D)	p_z	1.81	1.81	1.86
Quadrupole($D\mathring{A}$)	Q_{zz}	0.08	0.08	0.11
	Q_{xx}	-2.53	-2.53	-2.625
	Q_{yy}	2.45	2.45	2.515
Octupole($D\mathring{A}^2$)	O_{zzz}	-1.35	-1.17	NA
	O_{xxz}	-1.25	-1.44	NA
	O_{yyz}	2.61	2.61	NA

Figure 2.6(b) compares the error in electrostatic potential calculated by the 3-charge OPCA with the error produced by the point multipole approximations, relative to the exact computation using the original charge distribution [2]. The error shown in figure 2.6(b) is calculated on a circle, in the plane of the water atoms (y-z plane), at R = 2.8 Å from the oxygen atom, which approximates the oxygen-oxygen (contact) distance between two closest water

[2] Error calculations exclude any points that fall within the extent of the original charge distribution.

molecules. The overall RMS error in electrostatic potential calculated on a $R = 2.8$ Å spherical surface centered at the oxygen atom is 0.0010 e/Å with maximum error of 0.0027 e/Å for the 3-charge OPCA, compared to 0.0015 e/Å with maximum error of 0.0041 e/Å for the point octupole expansion. The overall RMS error in electrostatic potential at $R = 2.23$ Å(experimental oxygen-Na^+ distance[107]) is 0.0036 e/Å (1.20 kcal/mol) with maximum error of 0.0042 e/Å (1.39 kcal/mol) for the 3-charge OPCA compared to 0.0065 e/Å (2.16 kcal/mol) with maximum error of 0.0074 e/Å (2.46 kcal/mol) for the point octupole expansion.

2.7 Conclusion

Truncated point multipole expansions are a widely used approach to approximate potentials produced by complex charge distributions. However, if only the lowest order terms in the multipole expansion are kept, as is often done in practical calculations, the point multipole expansion can produce considerable error in the mid-field. Furthermore, implementation of such approximations into existing electrostatic models that were originally developed for point charge distributions, *e.g.* pairwise implicit solvent models, presents many challenges. In this work, we have introduced an alternative to the point multipole expansion – the optimal point charge approximation (OPCA). An OPCA consists of a given number of point charges which are optimally placed to best reproduce the electrostatic potential due to the original charge distribution. By construction, OPCAs retain many of the useful properties of point multipole expansions, in particular they retain the asymptotic behavior of the point multipole expansion. At the same time, an expansion based on OPCAs can be more accurate than the point multipole expansion of the same order.

We have provided a general framework for calculating OPCAs to any order. We have also derived closed-form analytical expressions for the 1-charge OPCA, and closed-form analytical expressions that approximate the 2-charge OPCA with reasonable accuracy – the 2-charge practical point charge approximation (PPCA). We note that higher order closed-form, analytical OPCAs may be challenging to derive, but for some applications, lower order OPCAs may be sufficient. The analytical expressions derived here for the 1-charge and 2-charge OPCAs, are guaranteed to be at least as accurate as the corresponding point multipole expansion of the same order. These analytic expressions not only provide physical insight but are more computationally efficient than the numerical minimization procedures that are in general required to obtain the optimal point charge approximation. Thus, these analytic expressions may be particularly useful in applications such as molecular dynamics where computational speed is critical.

For a set of sample charge distributions relevant to biomolecular modeling, the 2-charge PPCA was found to be on average 23% more accurate than the point dipole approximation, and comparable in accuracy to the point quadrupole approximation in the mid-field (electrostatic potential evaluated at 2 times the extent of the charge distribution). The standard deviation in RMS error for the 2-charge PPCA was also 59% lower than that of the point

dipole approximation and comparable to that of the point quadrupole approximation.

We also calculated the 3-charge optimal point charge approximation to represent a (quantum mechanical) gas phase charge distribution of water molecule. The electrostatic potential approximated by the 3-charge OPCA in the mid-field (2.8 Å from the oxygen atom) is on average 33.3% more accurate than that of the point octupole approximation. Interestingly, the positions of the 3 OPCA charges are quite different from atom center charge placements based on simple point charge models such as SPC or TIP3P. Further investigation is necessary to determine if and how such a 3-charge approximation can be used in practical applications.

Representing complex charge distributions by a small number of point charges is not, by itself, a novel idea. There are a number of methods, such as RESP [22], CHELP [41], CHELPG [29], CHELMO [180], Finite Point Charge (FPC)[43], coarse graining [20, 95, 12] and others [192] that empirically fit a set of point charges to a given charge distribution by minimizing various error metrics in electrostatic potential over some volume or surface surrounding the charge distribution. However, a key difference between the above methods and the optimal point charge approximation introduced here, is that the OPCAs (and their practical approximations, PPCAs) inherit the physically appealing asymptotic properties of the point multipole approximation, i.e. the error in potential is guaranteed to fall off at least as fast as $1/R^{k+1}$, where R is the distance from the origin and k is the highest order of the multipole terms retained in the expansion. In contrast, fitting the representative charges to minimize electrostatic error over some arbitrary volume or surface (e.g. molecular surface) does not guarantee the above asymptotic behavior, and can potentially lead to relatively large errors outside the volume or surface used for fitting.

Furthermore, in comparison to point multipoles, expansions based on PPCAs have many desirable properties that may be useful in practical computations; in particular, their mathematical form – the sum of Coulombic contributions from point sources – is simpler than that of the conventional point multipole expansion and is amenable to common speed-up schemes such as the generalized Born implicit solvent model [38]. Thus, PPCAs may be easier to implement into existing molecular dynamics protocols.

2.8 Acknowledgments

We thank Edward Valeev, Virginia Tech, for providing the quantum mechanical charge distribution for water molecule.

Chapter 3

Development of a new approach for constructing water models: parametrization of 4-point OPC model

3.1 Overview

Simplified, classical models of water are an integral part of atomistic molecular simulations. Simplified classical water models are currently an indispensable component in practical atomistic simulations. Yet, despite several decades of intense research, these models are still far from perfect. Presented here is an alternative approach to constructing widely used point charge water models. In contrast to the conventional approach, we do not impose any geometry constraints on the model other than the symmetry. Instead, we optimize the distribution of point charges to best describe the "electrostatics" of the water molecule. The resulting "optimal" 3-charge, 4-point rigid water model (OPC) reproduces a comprehensive set of bulk properties significantly more accurately than commonly used rigid models: average error relative to experiment is 0.76%. Close agreement with experiment holds over a wide range of temperatures. The improvements in the proposed model extend beyond bulk properties: compared to common rigid models, predicted hydration free energies of small molecules using OPC are uniformly closer to experiment, root-mean-square error < 1 kcal/mol.

3.2 Introduction

Water is the most extensively studied molecule [108, 200, 50] of unique importance to life. Yet our understanding of how this deceptively simple compound of just three atoms gives rise to the many extraordinary properties of its liquid phase [61, 62, 18] is far from complete [189]. The complexity of the water properties combined with multiple possible levels of approximation (e.g. quantum vs. classical, flexible vs. rigid) has led to the proposal of literally hundreds of theoretical and computational models for water [79]. Among classical water models [104, 23, 138, 85, 28, 131, 91, 195, 211, 66, 210, 60, 6], the most simple and computationally efficient, rigid non-polarizable models that represent water as a set of point charges at fixed positions relative to the oxygen nucleus, stand out as the class used in the vast majority of biomolecular studies today. Most commonly used models of this class, (e.g. TIP3P [104] and SPC/E [23] 3-point models, TIP4P/Ew [85] 4-point model, and the TIP5P [138] 5-point model) have achieved a reasonable compromise between accuracy and speed, but are by no means perfect [79, 140]. In particular, none of these models faithfully reproduce all of the key properties of bulk water simultaneously. Given the extraordinary complexity of real water-water interactions and hydrogen bonding networks in liquid phase, and their sensitivity to various model properties[217], even modest inaccuracies of water models can adversely affect outcomes of atomistic biomolecular modeling in an unpredictable manner. Particularly worrisome is the fact that improvements in over-all model accuracy do not necessarily translate into improvements in the accuracy of quantities most relevant to biomolecular simulations, such as molecular hydration free energies. For example, counterintuitively, TIP3P model predicts hydration free energies of small neutral molecules more accurately[145] than the TIP4P-Ew model that fixed several of TIP3P flaws; TIP5P[138], which is known to yield excellent water structure, is even less accurate in that respect [145]. But even for TIP3P, the average errors are still outside the desired "chemical accuracy" of less than 1 kcal/mol, a goal for rational drug design efforts. The need for better accuracy motivates an on-going search for more accurate yet computationally facile water models [211, 66, 210, 60].

Most unique properties of liquid water are due to the ability of the water molecules to establish a hydrogen-bonded structure, through the attraction between the electropositive hydrogen atoms and the electronegative oxygen atoms [139]. Therefore, a key challenge in developing classical water models is to find an accurate yet simplified description of the charge distribution of the water molecule that can adequately account for the hydrogen bonding in the liquid phase. For the past 30 years, the basic approach used to construct point charge water models, inspired by the classical works[142, 25] that revealed V-shape of water molecule and suggested near-tetrahedral arrangement of its charges, has been the same: the atomic partial charges and the Lennard-Jones potential parameters are optimized to reproduce selected bulk properties of water [79]. While sophistication of the optimization techniques employed to find the optimum has grown tremendously[6], from essentially "guess-and-test" to the complex, state-of-the-art optimization techniques [211, 6, 190, 87, 15], one

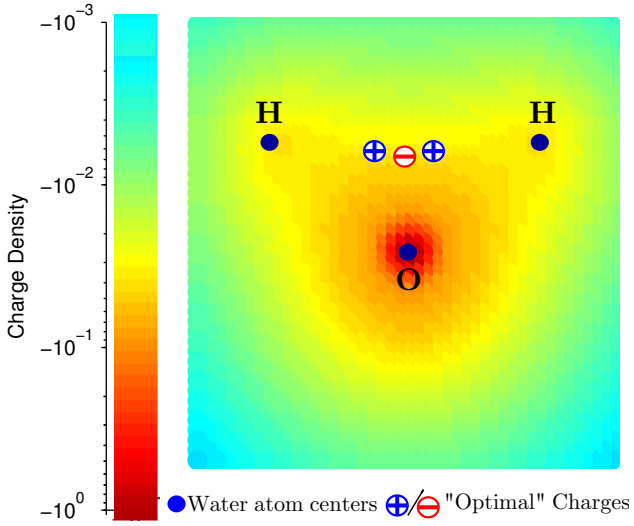


Figure 3.1: Charge distribution of the water molecule in the gas phase obtained from a quantum mechanical calculation [9]. Counter-intuitively, three point charges that optimally reproduce the electrostatic potential of this charge distribution are clustered in the middle, as opposed to the on-nuclei placement used by common water models that results in a much poorer electrostatic description of the underlying charge distribution[9].

crucial aspect of the over-all procedure has not changed: it imposes constraints on the allowed variations of the model geometry. That is $|OH|$ bond length and $\angle$HOH angle are either fixed, or are only allowed to vary slightly around their "canonical" values. The assumption is that optimal locations of the positive point charges of the model should be somewhere near the experimental hydrogen nuclei positions. This approach may not necessarily accurately reproduce the electrostatic characteristics of the water molecule due to severe constraints on allowed variations in the charge distribution being optimized. In fact, the configuration of three point charges to best describe the charge distribution of the water molecule can be very different from what one may intuitively expect based on its well-known atomic structure. Consider, for example, the gas-phase quantum-mechanical (QM) charge distribution of water molecule, Figure 3.1. The shown tight cluster of the point charges away from the nuclei reproduces the electrostatic potential around the QM charge distribution considerably more accurately than the more traditional distribution with point charges placed on or near the nuclei. For the optimal charge placement, Figure 3.1, the maximum error in electrostatic potential at the experimental oxygen-Na^+ distance (2.23 Å) from the origin, is almost 5.4 times smaller than that of the nucleus-centered alternative (1.4 kcal/mol vs. 7.56 kcal/mol). Intrigued by the idea that optimal placement of the point charges in a water model can be very different from the "intuitive" placement on the nuclei, and encouraged by the significant improvement of the accuracy of electrostatics brought about by this strategy in gas-phase, we have formulated and tested a different approach to building classical water models for the liquid phase.

Within classical potential functions used by point charge water models, the complexity of the hydrogen bonding interactions are primarily described by the electrostatic interactions [147]. While the electrostatic interactions are complemented by a Lennard-Jones (LJ) potential, the latter is generally represented by a single site centered on the oxygen – the corresponding interaction is isotropic and featureless, in contrast to hydrogen bonding which is directional. Therefore, an accurate representation of electrostatic interactions is paramount for accurately accounting for hydrogen bonding and the properties of liquid water. In a search for the best "electrostatics", commonly used distance and angle constraints on the configuration of a model's point charges are therefore of little relevance to classical rigid water models, yet these constraints impede the search for the "best" model geometry. This observation leads to one of the key features of our approach: any "intuitive" constraints on point charges or their geometry (other than the fundamental C_{2v} symmetry of water molecule) are completely abandoned here in favor of finding an optimal electrostatic charge distribution that best approximates liquid properties of water.

While ultimately it is the values of the point charges and their relative positions that we seek, Figure 4.1, we argue that the conventional "charge–distances–angles" space [104, 85, 138, 23] is not optimal to perform the search for the best electrostatics model. These coordinates affect the resulting electrostatic potential in a convoluted manner, it is unclear which ones, if any, may be relatively more important than others. At the same time, many key properties of liquid water are extraordinarily sensitive to tiny changes in parameters of these models (hence the number of significant digits kept to describe their parameters). The optimization landscape in the "charges–distances–angles" space is apparently complex, with multiple local optima, so that even the best minimization methods are virtually guaranteed to fail to locate the global optimum that may be far away from an initial "intuitive" guess. On the other hand, the electric field outside any complex charge distribution can be systematically approximated via its multipole moments[96], with lower order moments expected to have stronger effect on the electrostatic potential[96], and, not surprisingly, on liquid water properties as well [191, 91, 120]. Hence, our second key proposal is to search for optimal parameters of fixed-point charge models in the electrostatically most relevant, low-dimensional sub-space of lowest multipole moments, rather than in the convoluted high-dimensional charges–distances–angles space "native" to point-charge models. An exhaustive search for the optimum is enabled by a set of closed form, analytical expressions (see *Methods*) that for any input set of water multipole moments finds a unique configuration of n point charges that optimally represent the electrostatic potential of the input multipoles, even for small n. The fundamental symmetry (C_{2v}) of water molecule makes such non-trivial mapping possible.

Clearly, any reasonable water model needs to account for the large dipole moment of water molecule in order to reproduce dielectric properties of the liquid state [60, 151, 171]. At short distances where hydrogen bonds between water molecules form ($\approx$2.8Å), the relevance of higher electrostatic moments is also significant. For instance, the larger component of the water quadrupole has a strong effect on the liquid water structure seen in simulations [151],

and on the phase diagram [3]; quadrupole moment's importance for water models was pointed out a long time ago [19, 213]. The next order terms – octupole moments – while presumably less influential, also affect water structure *e.g.* around ions [195]. An intricate interplay between the dipole, quadrupole and octupole moments gives rise to the experimentally observed charge hydration asymmetry of aqueous solvation – strong dependence of hydration free energy on the sign of the solute charge [149, 148]. Therefore, we seek a fixed-charge rigid model that optimally represents the three lowest order multipole moments of the water molecule.

The specifics. Specifics of the proposed approach are exemplified below through the construction and testing of a 4-point, rigid "optimal" point charge (OPC) water model. To optimally reproduce the three lowest order multipole moments for the water molecule charge distribution, a minimum of three point charges are needed [9]. The most general configuration for a three point charge model consistent with C_{2v} symmetry of the water molecule is shown in Figure 4.1: the point charges are placed in a V-shaped pattern in the Y-Z plane. We follow convention [104, 85, 138, 23] and place the single Lennard-Jones (LJ) site on the oxygen atom. The four parameters (q,z_2,z_1 and y) that completely define the charge distribution, Figure 4.1, are uniquely determined via analytical equations introduced in *Methods*, to best reproduce a targeted set of three lowest order multipole moments (dipole, quadrupole and octupole) [9]. Specifically, the optimal parameters of each test model are such that the two lowest order moments are reproduced exactly, while the octupole is optimally approximated (minimum rms error)[9].

The ability to independently vary the moments of the charge distribution, provided by these analytical expressions, makes computationally feasible a full exploration in the relevant subspace of the moments. Generally, the importance of the multipole moments are inversely related to their order. The highest order multipole moment here is the octupole that has two independent components (Ω_0 and Ω_T), which we fix to high quality quantum mechanical (QM) predictions, QM/230TIP5P [44], Table 4.1. The linear component of the quadrupole Q_0 is known to be relatively small for the water molecule and not expected to be very important [168], therefore, we also simply set it to the known QM value (QM/230TIP5P [44], Table 4.1). This leaves the two most important components, the dipole (μ) and the "square" quadrupole ($Q_T = 1/2(Q_{yy} - Q_{xx})$, see *Methods*), as the two key search parameters we vary. We attempt to find the best fit to six key bulk properties by exhaustively searching in the 2D space of μ and Q_T, Figure 3.3, within the ranges that reflect known experimental uncertainties [78] and those of QM calculations [184, 181], Table 4.1. The six target bulk properties are: static dielectric constant ϵ_0, self diffusion coefficient D, heat of vaporization ΔH_{vap}, density ρ and the position $roo1$ and height $g(roo1)$ of the first peak in oxygen-oxygen pair distribution functions. These properties are calculated from molecular dynamics (MD) simulations, see *Methods* and the SI. For every trial value of μ and Q_T (and the fixed values of Q_0, Ω_0 and Ω_T), the charge distribution parameters (q,z_2,z_1 and y) are analytically determined (see *Methods*).

Table 3.1: Water molecule multipole moments centered on oxygen: from experiment, common rigid models, liquid phase quantum calculations, and OPC model (this work).

Model	μ [D]	Q_0 [DÅ]	Q_T [DÅ]	Ω_0 [DÅ^2]	Ω_T [DÅ^2]
EXP (liquid) [78]	2.5−3	NA	NA	NA	NA
SPC/E	2.35	0.00	2.04	-1.57	1.96
TIP3P	2.35	0.23	1.72	-1.21	1.68
TIP4P/Ew	2.32	0.21	2.16	-1.53	2.11
TIP5P	2.29	0.13	1.56	-1.01	0.59
AIMD1 [181]	2.95	0.18	3.27	NA	NA
AIMD2 [184]	2.43	0.10	2.72	NA	NA
QM/4MM [151]	2.49	0.13	2.93	-1.73	2.09
QM/4TIP5P [151]	2.69	0.26	2.95	-1.70	2.08
QM/230TIP5P [44]	2.55	0.20	2.81	-1.52	2.05
OPC	**2.48**	**0.20**	**2.3**	**-1.484**	**2.068**

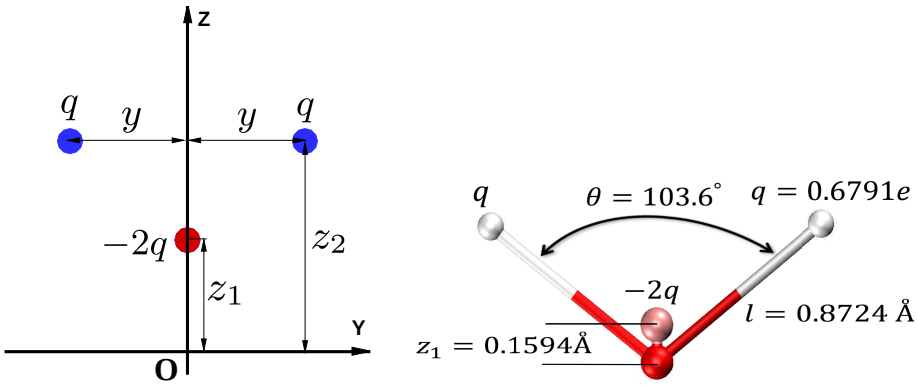


Figure 3.2: ***Left.*** The most general configuration for a three point charge water model consistent with C_{2v} symmetry of the water molecule. The single Lennard-Jones interaction is centered on the origin (oxygen). ***Right.*** The final, optimized geometry of the proposed 3-charge, 4-point OPC water model.

For every charge distribution calculated as above, the value A_{LJ} of the 12-6 Lennard-Jones (LJ) potential (see *SI*), which is mainly responsible for the liquid structure [168], is selected so that the location of the first peak $g_{oo}(r)$ of the oxygen-oxygen radial distribution function (RDF) is in agreement with recent experiment [185] (see *Methods*). The value of B_{LJ} is optimized so that the experimental value for density is achieved. The parameters A_{LJ} and B_{LJ} can be optimized nearly independently due to the weak coupling between them [168].

The result of the above search procedure is a "quality map" of all possible water models in the $\mu - Q_T$ space: the proposed OPC model is the one with the highest quality score.

3.3 Methods and simulation protocols

3.3.1 Simulation protocols

The calculations of thermodynamic and dynamical bulk properties were done based on standard equations in the literature (see Methods). Unless specified otherwise, we use the following Molecular Dynamics (MD) simulations protocol. Simulations in the NPT ensemble (1 bar, 298.16 K) were carried out using the PMEMD module of Amber suite of programs[37]. All the computations were performed on GPU (GTX 680). A cubic box with edge length of 30Å was filled with 804 water molecules. Periodic boundary conditions were used. Long-range electrostatic interactions, calculated via the particle mesh Ewald (PME) summation, and the van der Waals interactions were cut off at distance 8Å . MD simulations were conducted with a 2 fs time step; all intra-molecular geometries were constrained with SHAKE. The NPT simulations were performed using Langevin thermostat with coupling constant $\gamma = 2.0\ ps^{-1}$, and a Berendsen barostat with coupling constant of 1.0 ps^{-1} for equilibration and 3.0 ps^{-1} for production. We use the Amber default for the remaining parameters, unless otherwise specified. The duration of production runs vary between 1 ns to 65 ns, depending on the properties (see SI).

3.3.2 Scoring function

The predictive power of models against experimental data was validated using a scoring system developed by Vega et. al. [202]. For a calculated property x and a corresponding experimental value of x_{exp}, the assigned score is obtained as [202]

$$M = max\{[10 - |(x - x_{exp}) \times 100/(x_{exp}tol)|], 0\} \tag{3.1}$$

where the tolerance (tol) is assigned to 0.5% for density, position of the first peak of the RDF and for heat of vaporization, 5% for height of the first peak of the RDF, and 2.5% for the remaining properties. The quality score assigned to each test model is equal to the average of the scores in bulk properties considered.

3.3.3 Analytical solution for optimal point charges

Here we present the analytical equations to find three point charges that optimally reproduce the dipole, the quadrupole and the octupole moments of the water molecule. In the coordinate system shown in Fig. 2 (main text), the elements of the traceless dipole $\mathbf{p}_i$, quadrupole $\mathbf{Q}_{ij}$ and octupole $\mathbf{O}_{ijk}$ tensors [9] are

$$\mathbf{p}_i = (0, 0, \mu) \tag{3.2}$$

$$\mathbf{Q}_{ij} = \begin{pmatrix} -Q_T - Q_0/2 & 0 & 0 \\ 0 & Q_T - Q_0/2 & 0 \\ 0 & 0 & Q_0 \end{pmatrix} \tag{3.3}$$

$$\mathbf{O}_{ijk} = \begin{pmatrix} -\Omega_T - \Omega_0/2 & 0 & 0 \\ 0 & \Omega_T - \Omega_0/2 & 0 \\ 0 & 0 & \Omega_0 \end{pmatrix} \tag{3.4}$$

where $i, j = x, y$ and $k = z$, and μ, Q_0, Q_T, Ω_0 and Ω_T are the dipole, the linear component of the quadrupole, the square component of the quadrupole, the linear component of the octupole, the square component of the octupole, respectively [191, 151]. The other elements of the octupole tensor ($k = x, y$) can be found by symmetry. The optimal charge values and positions are calculated so that these three moments are sequentially reproduced, starting with the lowest order moments [9]. The first two lowest order moments of the water molecule, the dipole and the quadrupole, are fully reproduced by requiring

$$\mu = 2q(z_2 - z_1) \tag{3.5}$$

$$Q_0 = -2q(\frac{y^2}{2} - z_2^2 + z_1^2) \tag{3.6}$$

$$Q_T = \frac{3qy^2}{2} \tag{3.7}$$

where z_2, z_1, y and q are independent unknown parameters that characterize the three point charge model (see Fig. 2). The above three equations are solved to find three geometrical parameters (z_2, z_1 and y), as follows

$$z_{1,2} = \frac{2Q_T + 3Q_0}{6\mu} \mp \frac{\mu}{4q} \tag{3.8}$$

$$y = \sqrt{\frac{2Q_T}{3q}} \tag{3.9}$$

For a given value of q, the values of z_2, z_1 and y found as above exactly reproduce the dipole (μ) and the quadrupole (Q_0 and Q_T) moments of interest. The only remaining unknown parameter, q, is found to optimally reproduce the next order moment, the octupole, which is described by two independent parameters (Ω_0 and Ω_T). The components of the octupole moment are related to the charge distribution parameters through

$$\Omega_0 = -2q(\frac{3}{2}y^2 z_2 - z_2^3 + z_1^3) \tag{3.10}$$

$$\Omega_T = \frac{5qy^2 z_2}{2} \tag{3.11}$$

The octupole tensor (Eq. 3.4) can be optimally approximated if the largest absolute principal value of the octupole tensor (i.e. $(\Omega_T - \Omega_0/2)$ for the water molecule) is reproduced [9]. Therefore, we set $(\Omega_T - \Omega_0/2)$ from Eqs. 3.10 and 3.11 and solve for q as

$$q = -3\frac{\sqrt{\mu^4(256Q_T^2+\xi)}+16Q_T\mu^2}{2\xi} \tag{3.12}$$

where

$$\xi = 52Q_T^2 + 60Q_TQ_0 - 9(3Q_0^2 + 8(\Omega_T - \Omega_0/2)\mu)$$

The above solution is valid only when $\xi < 0$. For $\xi \geq 0$, the point charge positions converge to a singular point and the charge values go to infinity. The corresponding region in $\mu - Q_T$ map (Fig. 3) leading to this condition is displayed in deepest red (zero score).

3.3.4 van der Waals Parameters

The usual 12-6 Lennard-Jones (LJ) potential is employed to model the van der Waals interaction among the oxygens. The Lennard-Jones function, E_{LJ}, can be written as

$$E_{LJ}(r_{oo}) = 4\epsilon_{LJ}[(\frac{\sigma_{LJ}}{r_{oo}})^{12} - (\frac{\sigma_{LJ}}{r_{oo}})^{6}] = \frac{A_{LJ}}{r_{oo}^{12}} - \frac{B_{LJ}}{r_{oo}^{6}} \tag{3.13}$$

The values of A_{LJ} and B_{LJ}, unlike σ_{LJ} and ϵ_{LJ} [190], are nearly independent [168]. The value of A_{LJ}, which is mainly responsible for characterizing the short-ranged repulsive interactions, is selected so that the location of the first peak of RDF $g_{oo}(r)$ is in agreement with the experiment [185]. Next, the value of B_{LJ}, which does not affect the structure significantly, is varied so that the experimental value for density is achieved.

3.3.5 Solvation free energy calculations

Standard thermodynamics integration (TI) protocol was adopted from Ref. [145]. The Merck-Frosst implementation of AM1-BCC [98, 99] was used to assign the partial charges. The topology and coordinates for the molecules were obtained from Ref. [145]. Molecules were solvated in triclinic box with at least 12 Å from the solute to the nearest box edge. After minimization and equilibration, we performed standard free energy perturbation calculations using 20 λ values. Real space electrostatic cutoff was 10 Å. All bonds were restrained using the LINCS algorithm. Production NPT simulations were performed for 5ns. Identical simulations were performed for TIP3P, TIP4PEw, and OPC.

To mitigate uncertainties due to conformational variability, the 20 test molecule were randomly selected from a subset of 248 highly rigid molecules [148]. Explicit solvent free energies calculations (via Thermodynamic Integration) were performed in GROMACS 4.6.5 [164] using the GAFF [209] small molecule parameters

3.3.6 Calculating the bulk properties

The calculation of bulk properties were done based on standard equations in the literature [85, 217, 2, 66]. Unless stated otherwise, values of OPC at ambient temperature (Table 3) are given as averages over six independent simulations of 65 ns each, except for those quantities that are derived from temperature dependent results. The temperature dependent results are calculated from one simulation of 65 ns for each temperature point, i.e. 12.5K intervals in a temperature range [248K, 373K]. Details of the calculations of studied quantities are described below.

3.3.7 Static dielectric constant

The static dielectric constant ϵ_0 is determined through [85, 2, 66]

$$\epsilon_0 = 1 + \frac{4\pi}{3k_BTV}(< \mathbf{M}^2 > - < \mathbf{M} >^2) \tag{3.14}$$

where $\mathbf{M} = \Sigma_i q_i \mathbf{r}_i$, $\mathbf{r}_i$ is the position of atom i, k_B is the Boltzmann constant, T is the absolute temperature and V is the simulation box average volume.

Self diffusion coefficient

The self-diffusion coefficient D is obtained using the Einstein relation [85, 66, 217]

$$D = \lim_{t\to\infty} \frac{1}{6t} < |r(t) - r(0)|^2 > \tag{3.15}$$

The simulation protocol to compute the self-diffusion coefficient is similar to the protocol described in Ref. [85]; the well equilibrated NPT simulations were followed up with 80 successive intervals of NVE (20 ps) and NPT (5 ps) ensembles. The self diffusion was obtained by averaging D values over all the NVE runs.

Heat of vaporization

The heat of vaporization ΔH_{vap} is obtained following the method described in Ref. [85], as

$$\Delta H_{vap} \approx -U_{liq}/N + RT - pV - E_{pol} + C \tag{3.16}$$

where U_{liq} is the potential energy of the liquid with N molecules at a given external pressure p and a temperature T, and V is the average volume of the simulation box. R is the ideal

gas constant. E_{pol} accounts for the energetic cost of the effective polarization energy, and can be approximated as

$$E_{pol} = \frac{(\mu - \mu_{gas})^2}{2\alpha_{gas}} \tag{3.17}$$

where μ is the dipole moment of the corresponding rigid model and μ_{gas} and α_{gas} are the dipole moment and the mean polarizability of a water molecule in the gas phase [1], respectively. The OPC's dipole is close to experiment and larger than that of common rigid models which yields a relatively larger value of E_{pol} for OPC compared to common rigid models. The last term in Eq. 3.16, C, is a correction to account for the change in the intramolecular vibrational modes and for nonideal gas behavior, which for various temperatures is calculated and reported by Horn et al [85].

Isobaric heat capacity

The isobaric heat capacity c_p is determined through numeric differentiation of simulated enthalpies $H(T)$ over the range of temperatures T of interest [85, 217]

$$C_p \approx \frac{< H(T_2) > - < H(T_1) >}{T_2 - T_1} + \Delta C_{QM} \tag{3.18}$$

where ΔC_{QM} (≈ -2.2408 at $T = 298.0K$) is a quantum correction term accounting for the quantized character of the neglected intramolecular vibrations. The values of ΔC_{QM} for different temperatures are taken from Ref. [85]. The numeric differentiation is calculated from simulations in the temperature range [248K, 373K] in 12.5K increments.

3.3.8 Thermal expansion coefficient

The thermal expansion coefficient α_p can be approximated through numeric differentiation of simulated bulk-densities $\rho(T)$ over a range of temperatures T of interest [85, 217, 66]

$$\alpha_p \approx -(\frac{\ln < \rho(T_2) > - \ln < \rho(T_1) >}{T_2 - T_1})_P \tag{3.19}$$

The reported value at ambient conditions is calculated from a numeric differentiation of bulk-densities at T_1=296K and T_2=300K, averaged over 4 independent simulations.

Isothermal compressibility

The isothermal compressibility κ_T is calculated from volume fluctuations in NPT simulation using a Langevin thermostat with coupling constant 2.0 ps^{-1} and a Monte Carlo barostat with coupling constant of 3.0 ps^{-1}, via the following formula [85, 2, 66]

$$\kappa_T = \frac{< V^2 > - < V >^2}{k_B T < V >} \tag{3.20}$$

Simulations of 65ns and 15ns time length were performed to obtain the temperature dependent results for ($T \leq 298K$) and ($T > 298K$), respectively.

3.3.9 Propensity for Charge Hydration Asymmetry

Propensity of a water model to cause Charge Hydration Asymmetry (CHA) for a similar size cation/anion pair (B^+/A^-) such as K^+/F^- is defined in Ref. [149] as

$$\eta^*(B^+/A^-) = \frac{\Delta G(B^+) - \Delta G(A^-)}{1/2|\Delta G(B^+) + \Delta G(A^-)|} \approx 2\frac{\frac{\widetilde{Q}_{zz}}{\mu}}{R_{iw}} \tag{3.21}$$

where the term on the right is an approximation of propensity for CHA for point charge water models [149], R_{iw} is the ion-water distance, ΔG is the free energy of hydration, and μ and $\widetilde{Q}_{zz}$ are the dipole and the nontraceless quadrupole moment of the model, respectively [191].

3.4 Results and discussion

The entire region of the $\mu - Q_T$ space was mapped out using initially a relatively coarse grid spacing (0.1 D and 0.1 DÅ) in each direction shown in Figure 3.3. At this point, the quality of each test water model – corresponding to a μ, Q_T point on the map – is characterized by a quality score function (see *Methods*) from a recent comprehensive review [202] based on the same six key bulk properties used for the fitting. Accordingly each model is assigned a quality score, using the score function explained in the *Methods* section, and is shown in Figure 3.3. As demonstrated in Figure 3.3, the highest quality region (the green area) occurs for (2.4 D $\leq \mu \leq$ 2.6 D) and (2.2 DÅ $\leq Q_T \leq$ 2.4 DÅ). The region is relatively small and this is why an exhaustive, fine-grain search was required to identify the best model, which we refer to as the Optimal Point Charge (OPC) model (Figure 3.3).

From Figure 3.3, one can see three distinct regions in the $\mu - Q_T$ space: the "common water models" region with relatively small dipole and square quadrupole moments, the "QM" region characterized by larger dipole and square quadrupole, and narrow, high quality (OPC)

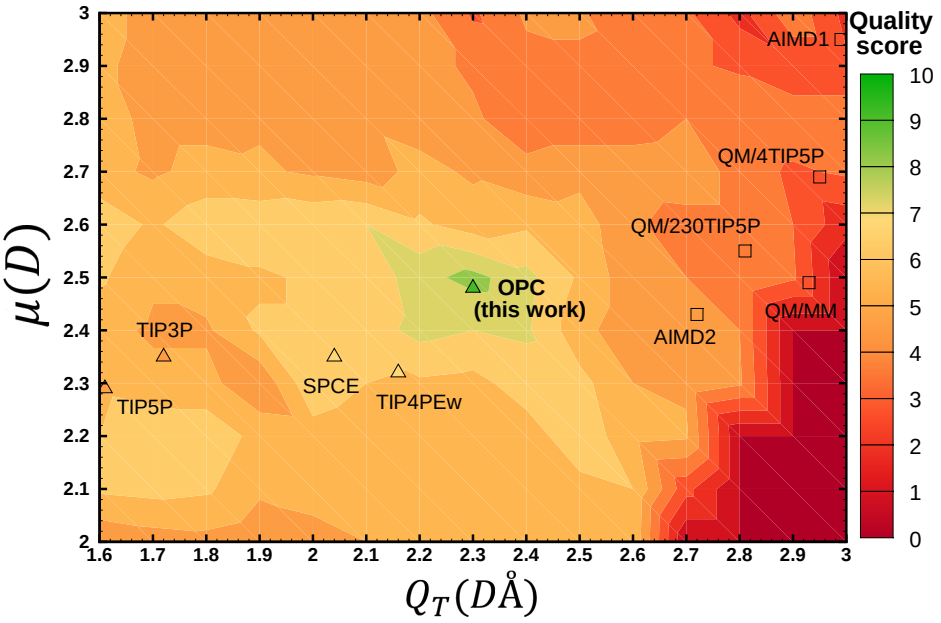


Figure 3.3: The quality score distribution of test water models in the space of dipole (μ) and quadrupole (Q_T). Scores (from 0 to 10) are calculated based on the accuracy of predicted values for six key properties of liquid water (see text). The resulting proposed optimal model is termed OPC. For reference, the μ and Q_T values of several commonly used water models (triangles, quality score given by the color at the symbol position) and quantum calculations (squares) are placed on the same map (see also Table 4.1). The actual positions of AIMD1 and TIP5P are slightly modified to fit in the range shown.

region with intermediate values of these two key moments. Compared to the other rigid models shown, OPC reproduces the multipole moments of water molecule in the liquid phase substantially better. In fact, the OPC dipole moment (2.48 D) is in best agreement with the range of values from experiment [78] and QM calculations [181, 184, 151, 44]; OPC's best fit value of μ coincides with a recent DFT-based estimate in liquid phase [171]. OPC's Q_T (2.3 DÅ) is larger than the corresponding values of the common models, and is closest to the QM predictions (Figure 3.3, Table 4.1). By construction, OPC's small Q_0 component of the quadrupole moment matches the reference QM value, and its octupole moments are the best approximations. The improved accuracy of the OPC moments is an immediate consequence of the focus on electrostatics and the unrestricted fine-grain search in the $\mu - Q_T$ subspace of the lowest, most relevant component of water multipole moments. The important improvements in the quality of model's liquid phase characteristics, seen in OPC, became possible through the abandoning of the conventional geometrical constraints used in model construction, which has allowed for the multipole moments to be varied independently. The availability of analytical equations that connect the optimal point charge distributions with the input multipole moments played an important role too.

Table 3.2: Force field parameters of OPC and some common rigid models, where $\sigma_{LJ} = (A_{LJ}/B_{LJ})^{1/6}$ and $\epsilon_{LJ} = B_{LJ}^2/(4A_{LJ})$. For comparison, water molecule geometry in the gas phase is also included.

	$q[e]$	l[Å]	z_1[Å]	$\Theta[deg]$	σ_{LJ}[Å]	$\epsilon_{LJ}[kJ/mol]$
EXP(gas)	NA	0.9572	NA	104.52	NA	NA
TIP3P	0.417	0.9572	NA	104.52	3.15061	0.6364
TIP4PEw	0.5242	0.9572	0.125	104.52	3.16435	0.680946
TIP5P	0.241	0.9572	NA	104.52	3.12	0.6694
SPC/E	0.4238	1.0	NA	109.47	3.166	0.65
OPC	**0.6791**	**0.8724**	**0.1594**	**103.6**	**3.16655**	**0.89036**

While the OPC moments are closest to the QM values, they (in particular Q_T) still deviate from the QM predictions (Table 4.1, Figure 3.3). The low quality of the test models, Figure 3.3, in which the moments were close to the QM values (squares, Figure 3.3) suggests that, within the 3-charge models explored here, an optimal fit of moments to QM predictions does not guarantee agreement with experimental liquid phase properties. This discrepancy can be due to a number of limitations and approximations inherent to classical, rigid, non-polarizable water models, see e.g. Refs. [202, 79, 210]. It may also be that only three point charges, even if placed optimally, are not enough to represent the complex charge distribution of real water molecule to the needed degree of accuracy. Namely, a three point charge model is fundamentally unable to exactly reproduce the reference dipole, quadrupole and octupole moments simultaneously [9], and essentially has no control over the accuracy of its moments beyond the octupole. The contribution of the higher order multipole moments to electrostatic potential can be significant at close distances, which are relevant to water-water and water-ion interactions in liquid phase. We conjecture that the relatively small μ and Q_T value found at the highest quality region (green zone, Figure 3.3) compared to QM predictions (squares, Figure 3.3), may be a compromise to keep the higher moments not too far from the optimal, ensuring a reasonable net electrostatic potential.

The OPC point charge positions and values and the LJ parameters are listed in Table 4.2. The $|O - q^+|$ distances for OPC are shorter (0.8724Å), and the $\angle\ q^+Oq^+$ angle (Figure 4.1) is slightly narrower (103.6°) than the corresponding experimental values of $|O - H|$ bond and $\angle$HOH angle for the water molecule in the gas phase (0.9572Å and 104.52°). The charge magnitudes of the OPC model are significantly larger than those of other common models (Table 4.2). Although the OPC charge distribution is not as tightly clustered as the configuration of the optimal charge model in the gas phase (Figure 3.1), the deviation of OPC geometry from that of other models and the water molecule in the gas phase is influential. In particular, the quality of water models is extremely sensitive to the values of electrostatic multipole moments (Figure 3.3), which by itself are very sensitive to the geometrical parameters (Eqs. 4.1-4.3, and SI).

Table 3.3: Model vs. experimental bulk properties of water at ambient conditions (298.16 K, 1 bar): dipole μ, density ρ, static dielectric constant ϵ_0, self diffusion coefficient D, heat of vaporization ΔH_{vap}, first peak position in the RDF $roo1$, propensity for charge hydration asymmetry (CHA) [149, 144, 165], isobaric heat capacity C_p, thermal expansion coefficient α_p, and isothermal compressibility κ_T. The temperature of maximum density (TMD) is also shown. Bold fonts denote the values that are closest to the corresponding experimental data (EXP). Statistical uncertainties (±) are given where appropriate.

Property	TIP4PEw [85]	SPCE [202, 211]	TIP3P [202, 138]	TIP5P [202, 138]	**OPC**	EXP [202, 203, 185]
$\mu(D)$	2.32	2.352	2.348	2.29	**2.48**	2.5–3
$\rho[g/cm^3]$	0.995	0.994	0.980	0.979	**0.997±0.001**	0.997
ϵ_0	63.90	68	94	92	**78.4±0.6**	78.4
$D[10^9 m^2/s]$	2.44	2.54	5.5	2.78	**2.3±0.02**	2.3
$\Delta H_{vap}[kcal/mol]$	10.58	10.43	10.26	10.46	**10.57±0.004**	10.52
$roo1$[Å]	2.755	2.75	2.77	2.75	**2.80**	2.80
CHA *propensity* [a]	0.52	0.42	0.43	0.13	**0.51**	0.51
$C_p[cal/(K.mol)]$	19.2	20.7	18.74	29	**18.0±0.05**	18
$\alpha_p[10^{-4}K^{-1}]$	3.2	5.0	9.2	6.3	**2.7±0.1**	2.56
$\kappa_T[10^{-6}bar^{-1}]$	48.1	46.1	57.4	41	**45.5±1**	45.3
$TMD[K]$	276	241	182	**277**	272±1	277

[a] Values are calculated in this work. The experimental value is a theoretical estimate [149] based on experimental hydration energies of K^+/F^- pair [176].

3.4.1 Bulk properties

The quality of the model in reproducing experimental bulk water properties at ambient conditions, and a comparison with other most commonly used rigid models is presented in Table 3. For each of 11 key liquid properties (Table 3) against which water models are most often benchmarked [202, 203, 85], our proposed model deviates by no more than 1.8% from the corresponding experimental value, except for one property (thermal expansion coefficient) that deviates from experiment by about 5%. While a targeted optimization may further improve the agreement of thermal expansion coefficient with experiment, however, an overall improvement of the model accuracy may require including (n>3) point charges, and eventually incorporating polarization and nuclear motion effects. The full O-O and O-H radial distribution functions (RDF), $g(r_{OO})$ and $g(r_{OH})$, are presented in the SI. By design, the experimental position of first peak in O-O RDF is accurately reproduced by OPC. The position and height of other peaks are also closely reproduced.

While commonly used models may be in good agreement with experiment for certain properties, Figure 3.4, they often produce large errors (sometimes amounting to over 250%) in some other key properties. In contrast, OPC shows a uniformly good agreement across all the bulk properties considered here.

The ability of OPC to reproduce the temperature dependence of six key water properties is

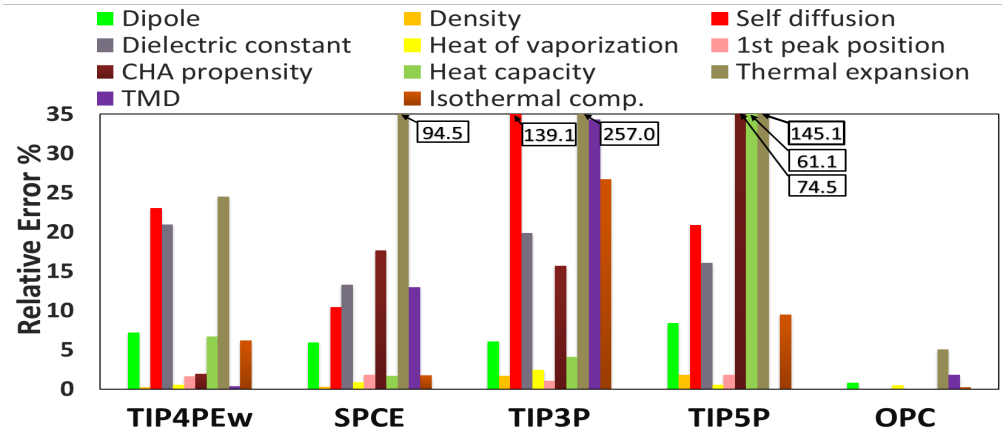


Figure 3.4: Relative error in various properties by the common rigid models and OPC (this work). Values of the errors that are cut off at the top are given in the boxes.

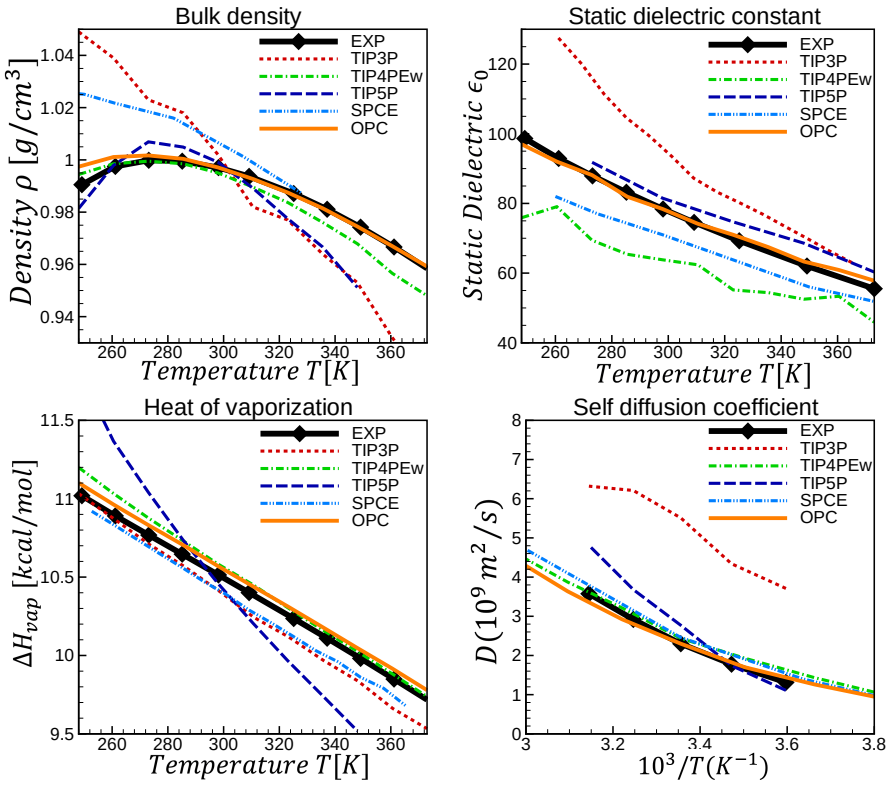


Figure 3.5: Calculated temperature dependence of water properties compared to experiment and several common rigid water models. TIP4PEw results are from [85], TIP5P from [138, 203, 85], TIP3P from [104, 203, 211, 105], SPCE from [54, 211].

shown in Figure 3.5 and Figure 3.6. OPC is uniformly closest to experiment compared to the other models shown. It is noteworthy that OPC, which resulted from a search in the space

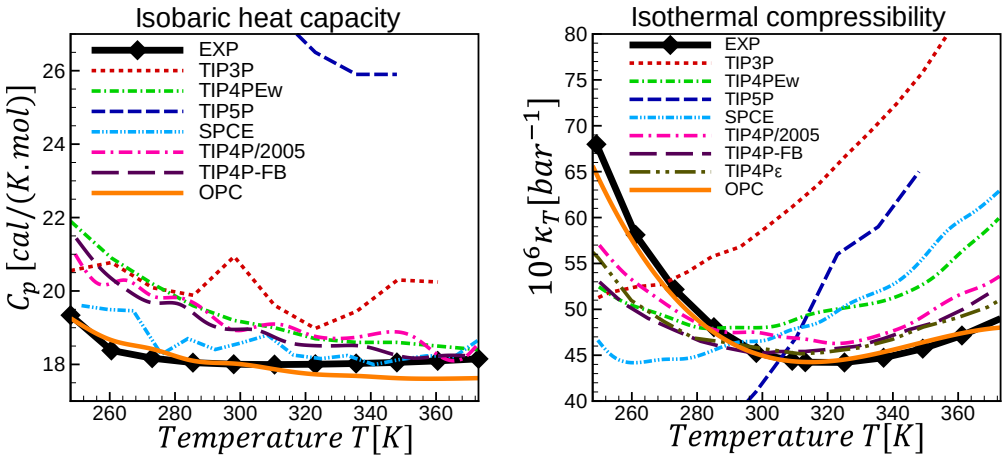


Figure 3.6: Variation of isobaric heat capacity and isothermal compressibility of liquid phase water with temperature. OPC model (this work) is compared to several common rigid models, some recent rigid models (TIP4P-FB [211], TIP4Pϵ [66] and TIP4P/2005 [2]) and experiment. TIP4PEw results are from [85], TIP5P from [138], TIP3P from [104, 211], SPCE and TIP4P-FB from [211], and TIP4Pϵ from [66].

of only two parameters (μ and Q_T) at only one thermodynamic condition (298.16 K and 1 bar) to fit a small subset of bulk properties, automatically reproduces a much larger number of bulk properties with a high accuracy across a wide range of temperatures where no fitting was performed. The procedure and the result are in contrast not only to commonly used, but also to some recent rigid [211, 66, 2] and even polarizable models [210] that generally employ massive and more specialized fits against multiple properties over a wide range of thermodynamic conditions. While noticeable advance in the accuracy of bulk properties is made by these latest models, the overall end result is not more accurate than OPC.

So far we have described comprehensive validation of OPC model in the liquid phase for which it is optimized. An equally comprehensive testing [202] of the model outside the liquid phase would be of interest, but is out of scope in this Letter that focuses on a new method. By construction, even a perfect fixed-charge rigid model that reproduced all bulk liquid properties exactly, would be inherently incapable to respond properly to the change of polarity of its micro-environment. Therefore, gas phase properties of OPC may not be as accurate as its liquid phase predictions. Nevertheless, reasonable higher multipole moments [3] of OPC, well reproduced temperature dependence of bulk properties, and especially a close agreement with experiment of isothermal compressibility, may be indicative of OPC's reasonable performance outside of liquid phase as well [202].

3.4.2 O-O and O-H radial distribution functions

Each test OPC model is parametrized to exactly reproduce the position of the first peak. The positions and the heights of the remaining peaks are very accurately reproduced with these parameters. The height of the first peak is however slightly high, which leads to an average O-O coordination number (n_{oo}) larger than experiment. This may be because of the r^{-12} repulsion in the LJ potential that is known to create an over structured liquid [117, 79]. It is argued that using a softer potential (e.g. a simple exponential in the form of Ae^{Br}) can correct the height of the first peak [117]. We employ a 12-6 potential to achieve compatibility with standard biomolecular force fields. While TIP3P is the only model that accurately reproduces the height of the first peak, it lacks structure beyond the first coordination shell (Fig. 3.7).

3.4.3 OPC in practical biomedical applications

One of the main goals of developing better water models is improving the accuracy of simulated hydration effects in molecular systems. Here we show that the optimized charge distribution of OPC model does lead to a more accurate representation of solute-water interactions, whose accuracy is critical to the outcomes of atomistic simulations. One of the most sensitive measure of the balance of intermolecular and solute-water interaction is hydration free energy, which has been used to evaluate the accuracy of molecular mechanics force fields and water models alike [106]. To evaluate OPC's accuracy, we use a set of 20 molecules randomly selected to cover a wide range of experimental hydration energies from a large common test set of small molecules [145], see *Methods*. Compared to experiment, OPC predicts hydration free energy more accurately, on average (RMS error = 0.97 kcal/mol), as compared to 1.10 kcal/mol and 1.15 kcal/mol for TIP3P and TIP4PEw, respectively (Figure 3.8). The improvement is uniform across the range of solvation energies studied, from very polar to non-polar molecules (see SI). The calculated average errors for OPC, TIP3P and TIP4PEw are 0.62, 0.78 and 0.87 kcal/mol, respectively, which shows that OPC is systematically more accurate than the other models tested. OPC is more accurate despite the fact that force fields have been historically parametrized against TIP3P. Somewhat paradoxically, TIP3P, which is certainly not the most accurate commonly used rigid model (see Figure 3.4), has nevertheless been generally known thus far to give the highest accuracy in hydration free energy calculations [145]. The accuracy improvement by OPC is then noteworthy as it shows that an improvement in the "right direction" can indeed lead to improvement in free energy estimates. To the best of our knowledge, OPC is the only classical point charge rigid model that predicts solvation free energies of small molecules within the "chemical accuracy" (RMS error $\leq$ 1 kcal/mol).

Combining our own preliminary results with those from Ref. [160] for different water models and force-fields, we find that OPC is the only general-purpose model that yields correct size of intrinsically disordered proteins, otherwise predicted to be too compact by commonly used

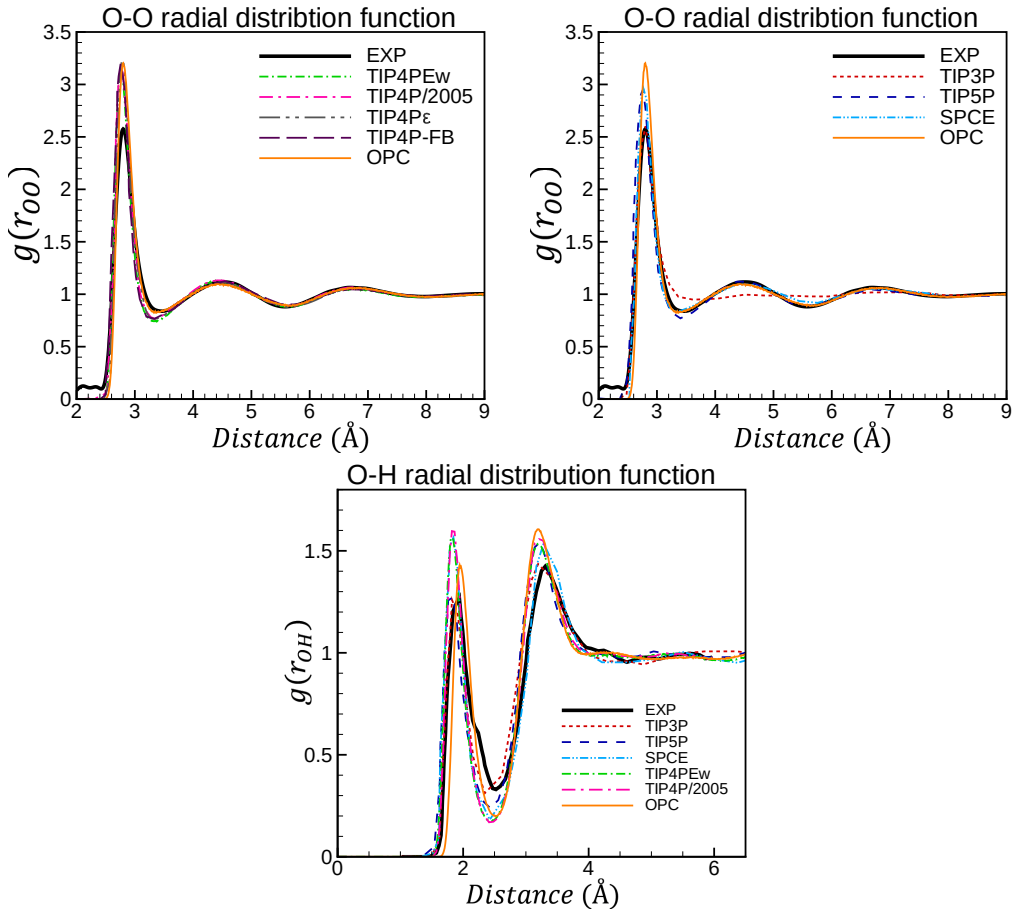


Figure 3.7: O-O and O-H radial distribution functions of liquid water at 298.16 K, 1 bar. The OPC model is compared to the commonly used rigid models as well as some recent rigid models (TIP4P-FB [211], TIP4Pϵ [66] and TIP4P/2005 [2]). The experimental data is taken from [185]. TIP4PEw result is from [85], TIP4P-FB from [211], TIP4Pϵ from [66], SCPE from [23], TIP3P from [105], TIP5P from [138] and TIP4P/2005 from [2]. For simplicity, we approximated locations of the protons in OPC water by locations of the positive point charges.

models (Fig. 3.9 (a)). The results MD simulations of small intrinsically disordered protein 1WJB were performed using the AMBER FF99SB and FF14SB. The simulation time is 5 μs. The remaining simulation protocols are the same as the protocol described in Ref. [160].

OPC has been found to be the only water model that yields quantitative agreement with NMR experiment for conformational populations of small RNA fragments – an agreement that was notoriously hard to obtain for more than a decade[24] (Fig. 3.9 (b)). Compared to other leading water models, the use of OPC in predicting ligand (host-guest) binding enthalpies reduces the error relative to experiment by at least a factor of two[68], and can eliminate the systematic error completely in some cases (Fig. 3.9 (c)). Critically, in all of

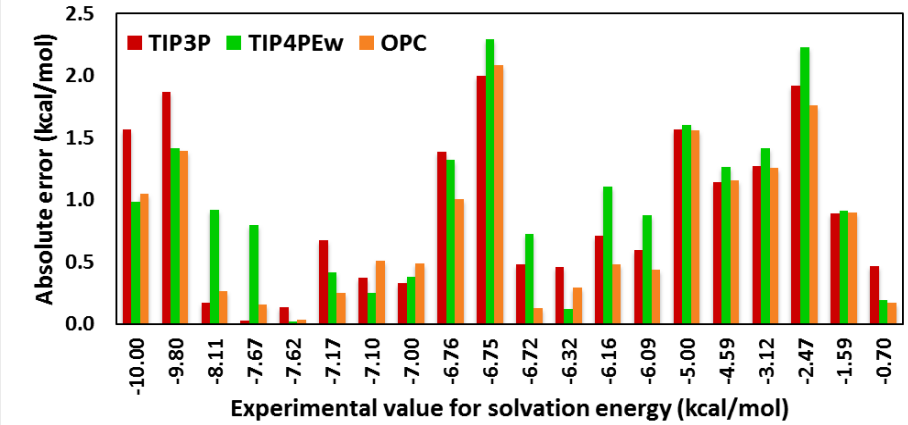


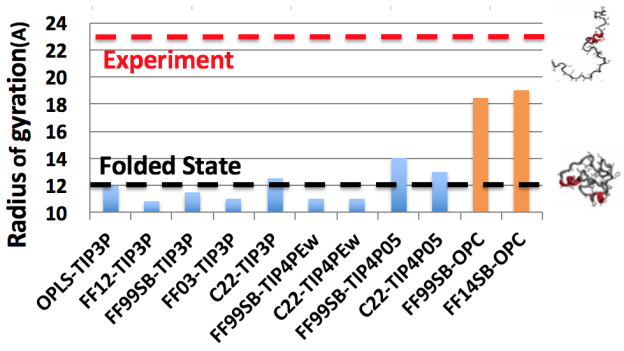
Figure 3.8: Absolute error relative to experiment in solvation free energies of a set of 20 small molecules calculated using TIP3P, TIP4P-Ew and the proposed OPC models.

the above examples OPC improved agreement with experiment for all underlying gas-phase force-fields tested. All these improvements combined are poised to take the accuracy of common, practical biomolecular simulations to a qualitatively new level.

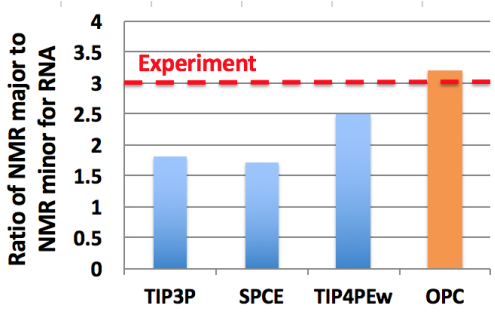
3.5 Conclusion

In summary, we have proposed a different approach to constructing classical water models. This approach recognizes that commonly used distance and angle constraints on the configuration of a model's point charges are of little relevance to classical rigid water models; these artificial constraints complicate and impede the search for optimal charge distributions, key to reproducing unique features of liquid water. In our approach, such constraints are completely abandoned in favor of finding an optimal charge distribution (obeying only the fundamental C_{2v} symmetry of water molecules) that best approximates properties of liquid water. Next, we focus on the lowest multipole moments which directly control the electrostatics of the model. The hierarchical importance of these moments for water properties allowed us to reduce the search space to essentially just two key parameters: the dipole and the square quadrupole (μ and Q_T) moments; the less important moments were fixed to the QM-derived values. The low dimensionality of the parameter space, combined with a set of derived equations that connect the optimal geometry and charge values of each test model to the input multipole moments, permitted a fine-grain exhaustive search virtually guaranteed to find an optimal solution within the accuracy class of water models considered here.

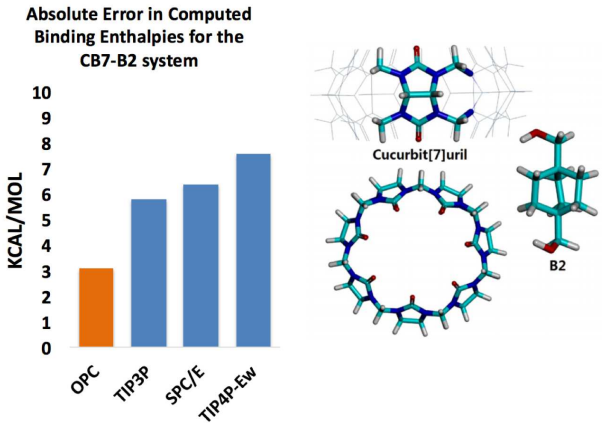
We believe that the general approach presented here can be used to develop water models with different numbers of point charges, including presumably even more accurate n-point ($n > 4$) models, and also flexible and polarizable models. We expect that finding an n-point charge optimum in the 2D parameter space (μ, Q_T) is not going to be significantly



(a)



(b)



(c)

Figure 3.9: OPC model performance in biomolecular simulations. (a) Predicted radius of gyration of small intrinsically disordered protein 1WJB. Shown are our preliminary results and published points from Ref. [160]. (b) The ratio of the population of NMR major structure to the population of NMR minor structure computed using ff12 force field, vdW$_{bb}$ and different water models, as well as the experimental ratio[24]. (c) Computed binding enthalpies[68] for a host-guest system - a miniature model of molecular recognition.

more difficult than for the 4-point model presented here. The current 4-point OPC model is included in the solvent library of the Amber v14 molecular dynamics (MD) software package, and has been tested in GROMACS 4.6.5. The computational cost of running molecular dynamics simulations with it is the same as that for the popular TIP4P model.

Chapter 4

Accuracy limit of rigid 3-point water models: parametrization of 3-point OPC3

4.1 Overview

Realistic explicit solvent models are critical for the success of molecular simulations. Among all, the very simple 3-point explicit rigid water models are very popular due to their computational efficiency. Recently, we introduced a new approach to constructing explicit rigid water models (Izadi et al. J. Phys. Chem. Lett., 2014, 5, 3863-3871) that permits a virtually exhaustive search for optimal model parameters in the sub-space most relevant to electrostatic properties of the water molecule in liquid phase, rather than in the high-dimensional charges-distances-angles space "native" to point-charge models. We previously demonstrated that the 4-point Optimal Point Charge (OPC) water model, constructed based on the new approach, delivers significant accuracy in reproducing water bulk properties, when compared to most commonly used models. In this work, we apply the same approach to develop a 3-point version of the Optimal Point Charge model (OPC3). OPC3 is significantly more accurate than commonly used water models of same class (TIP3P and SPCE) in reproducing a comprehensive set of bulk properties, over a wide range of temperatures. Close agreement of the model parameters and accuracy of OPC3 with values from two recent 3-point water models, obtained by independent sophisticated optimizations based on completely different methods, points out to a "consensus" for the optimal parametrization of 3-point water models.

4.2 Introduction

Molecular modeling and simulations are routinely employed to study structure and function of biological molecules in applications ranging from structural biology to bio-medicine and rational drug design[68, 71]. Accurate, classical water models are just as important for these modeling efforts as water is for Life [189, 112, 50]. The simplest, and most widely used, atomistic water models are fixed-charge rigid non-polarizable models [104, 2, 85, 138, 23, 94], implemented in virtually every modeling package[35, 30, 164, 159]. However, despite at least three decades of effort there is still significant room for much needed improvement[79, 202]. As more physical realism is added to such models either through more complex geometry or/and by inclusion of electronic polarization effects, the cost of finding the accuracy optimum in the large parameter space grows exponentially. As a result, available parametrizations are virtually guaranteed to be sub-optimal with respect to faithfully reproducing key experimental properties of water, hindering predictive potential of these models. At the same time, accuracy of water models directly affects accuracy of virtually every modeling study that depends on them. For example, the accuracy of predicted hydration energies of small molecules, needed in computation of protein-ligand binding energetics[68], is sensitive to the choice of water models. The "chemical accuracy" of less than 1 kcal/mol, desirable for computational drug design efforts, is still not achieved in these calculations[145, 71]. Unacceptably large discrepancies with experimental binding energies are seen for even smallest ligand binding systems[93]. Another example is the dependence of predicted protein-folding landscapes, and pathways, on the choice of water model[160]. Precise understanding of folding pathways is of direct bio-medical relevance: alteration of folding pathways is a common feature of a wide range of highly debilitating and increasingly prevalent diseases.

The most popular class of water models in practical molecular simulations are the simple 3-point models, mainly due to their computational efficiency. These models of water are about as simple as one can possibly make for atomistic simulations (successful 2 point models were developed for coarse-grained simulations[95]). Despite their continued popularity, most commonly used three point water models (TIP3P[104] and SPCE[23]) cannot faithfully reproduce bulk properties at 298 K and 1 bar pressure at once. For instance, while TIP3P produces the enthalpy of vaporization and dielectric constant reasonably better than other properties, it underestimates the density and significantly overestimates the self-diffusion coefficient[202]. In contrast, SPCE fairly accurately reproduces the density and self-diffusion, however it overestimates the enthalpy of vaporization and underestimates the dielectric constant. Unfortunately, both of these commonly used models lead to poor reproduction of the temperature dependence of liquid water properties. For instance, the density maximum experimentally observed at 277 K is not present in the density profile of these two models in the range of temperatures normally studied for water.

Recently we proposed a new approach to constructing classical water models that can deliver novel *global accuracy optimal* n-point water models that is completely different from the mainstream water modeling parametrization techniques[93]. In contrast to the traditional

parametrization techniques, the key feature of the new approach is to completely abandon any constraints on the point charge geometry, except the fundamental symmetry of water molecule, in favor of an unconstrained exhaustive search in a subspace of low order multipole moments. The 4-point OPC water model constructed based on the new approach was shown to reproduce the key liquid state properties significantly more accurately than commonly used water models[93]. OPC can deliver noticeable accuracy improvement in molecular simulations of solvated biomolecules even with existing force-fields. Improvements have been reported specifically in RNA simulations[24], thermodynamics of ligand binding[68], and small molecule hydration[93].

Motivated by the notable popularity of the simple 3-points models in practical molecular simulations, here we apply our new approach to construct a 3-point version of the Optimal Point Charge water model, referred to as OPC3. We compare OPC3 with two most commonly used models (TIP3P and SPCE) and also with two more recent 3-point models obtained by independent optimizations based on completely different methods (TIP3PFB[211] and H2ODC[60]).

4.3 Methods

4.3.1 Optimization procedure

The first key feature of our approach is to abandon any and all (seemingly intuitive) constraints on point charge values or their relative positions (other than the fundamental C_{2v} symmetry of water molecule) in search for an optimal electrostatic charge distribution that best approximates liquid properties of water. The most general configuration for a 3-charge 3-point model consistent with C_{2v} symmetry of the water molecule is shown in Figure 4.1. Note that the position of the negative charge in a 3-point model has to coincide with the position of the oxygen atom (vdw center), but the position of the positive charges are allowed to vary (Figure 4.1). As a result, the charge distribution in 3-charge 3-point models has an additional geometry constraint compared to 4-point 3-charge model in which charges are completely unconstrained (except for the C_{2v} symmetry)[94].

Based on the concept of Optimal Point Charge Approximation[9], the parameters of the charge distribution shown in Figure 4.1 can be optimized so that the most important dipole and the quadrupole moments of the water molecule are reproduced exactly (note that the monopole is zero)[9, 94]. In the coordinate system shown in Figure 4.1, the charge distribution characterized by z, y and q can fully reproduce a given set of dipole (μ) and quadrupole

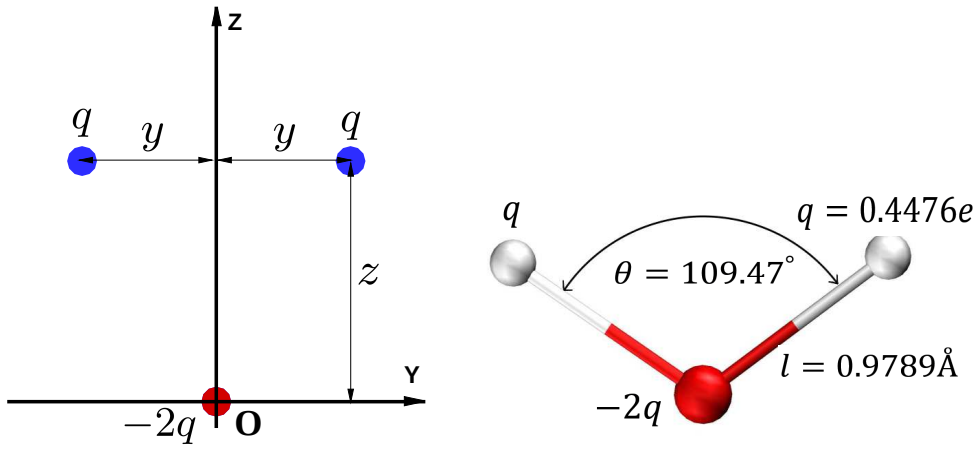


Figure 4.1: ***Left.*** The most general configuration for a 3-charge 3-point water model consistent with C_{2v} symmetry of the water molecule. The charge distribution parameters (y, z, and q) are calculated to optimally reproduce a given set of dipole and quadrupole moments. The value of the positive and negative charges are q and -2q, respectively. The single Lennard-Jones interaction is centered on the origin (oxygen). ***Right.*** The final, optimized geometry of the proposed 3-point OPC3 water model.

(Q_0 and Q_T) moments for the water molecule by requiring

$$2qz = \mu \tag{4.1}$$

$$-2q(\frac{y^2}{2} - z^2) = Q_0 \tag{4.2}$$

$$\frac{3qy^2}{2} = Q_T \tag{4.3}$$

The above three equations are solved to find three parameters (z, y and q), as follows

$$q = \frac{3\mu^2}{2(2Q_T + 3Q_0)} \tag{4.4}$$

$$z = \frac{2Q_T + 3Q_0}{3\mu} \tag{4.5}$$

$$y = \frac{2}{3\mu}\sqrt{Q_T(2Q_T + 3Q_0)} \tag{4.6}$$

where μ, Q_0 and Q_T moments are related to the more traditional Cartesian components of the traceless multipole moments of water molecule as $\mu = \mu_z$, $Q_0 = Q_{zz}$, $Q_T = 1/2(Q_{yy} - Q_{xx})$ [191, 168, 151].The above set of analytical expressions enables us to independently vary the moments of the charge distribution, which makes computationally feasible a full exploration in the relevant subspace of the moments (μ, Q_0 and Q_T).

The Q_0 component of the quadrupole moment (linear quadrupole) is known to be relatively small for water molecule and is not expected to be very important [168]. We fix this value to zero, which automatically results in fully tetrahedral angle: by setting $Q_0 = 0$ in Equ. 4.2 we obtain $y = \sqrt{2}z$, that translates to $\theta = 109.4712$.

This leaves the two more important moments, the dipole (μ) and the square quadrupole (Q_T), as the two key search parameters we vary exhaustively. For the case of 3-point model, we vary μ within the range of $2.3D$ to $2.5D$, and Q_T within $1.6D$Å to $2.4D$Å, which reflect the ranges for commonly used and recently developed water models of the same class. For every pair of trial values of μ and Q_T (and the fixed value of $Q_0 = 0$) the optimal charge and geometry parameters of the test model (q,z, and y,Figure 4.1) are uniquely determined via the set of closed-form analytical expressions Equs.4.4-4.6.

The usual 12-6 Lennard-Jones (LJ) potential is employed to model the van der Waals interaction among the oxygens. The Lennard-Jones function, E_{LJ}, can be written as

$$E_{LJ}(r_{oo}) = 4\epsilon_{LJ}[(\frac{\sigma_{LJ}}{r_{oo}})^{12} - (\frac{\sigma_{LJ}}{r_{oo}})^{6}] = \frac{A_{LJ}}{r_{oo}^{12}} - \frac{B_{LJ}}{r_{oo}^{6}} \tag{4.7}$$

The values of A_{LJ} and B_{LJ}, unlike σ_{LJ} and ϵ_{LJ} [190], are nearly independent [168]. For every charge distribution calculated as described above, the value A_{LJ} of the 12-6 Lennard-Jones (LJ) potential, which is mainly responsible for the liquid structure [168], is selected so that the location of the first peak $g_{oo}(r)$ of the oxygen-oxygen radial distribution function (RDF) is in agreement with most recent experimental data [185]. The value of B_{LJ} is optimized so that the experimental value of water density is achieved. The parameters A_{LJ} and B_{LJ} can be optimized nearly independently due to the weak coupling between them [168].

Table 4.1: Water molecule multipole moments centered on oxygen: from experiment, liquid phase quantum calculations, some common and recent 3-point models, and OPC3 model (this work).

Model	μ [D]	Q_0 [DÅ]	Q_T [DÅ]	Ω_0 [DÅ^2]	Ω_T [DÅ^2]
EXP (liquid) [78]	2.5−3	NA	NA	NA	NA
QM/230TIP5P [44]	2.55	0.20	2.81	-1.52	2.05
SPCE	2.35	0.00	2.04	-1.57	1.96
TIP3P	2.35	0.23	1.72	-1.21	1.68
TIP3PFB	2.419	0.068	2.052	-1.584	2.03
H2ODC	2.417	0.000	2.005	-1.479	1.849
OPC3	**2.43**	**0.0**	**2.06**	**-1.552**	**1.940**

Using the procedure above, we can obtain a test water model for different combinations of μ and Q_T within the search space. We evaluate the performance of each of these models in reproducing six targeted liquid water properties at 298.16K and 1bar: static dielectric constant ϵ_0, self diffusion coefficient D, heat of vaporization ΔH_{vap}, density ρ and the position $r_{oo}1$

and height $g(r_{oo}1)$ of the first peak in oxygen-oxygen pair distribution functions. These properties are calculated as averages from molecular dynamics (MD) trajectories, using standard computational protocols[85, 66]. A detailed description for calculations of thermodynamic and dynamical bulk properties can be found in Ref. [94].

The quality of each test water model – corresponding to a μ, Q_T point on the map – is characterized by a quality score function suggested by Vega et. al. [202] based on the same six key bulk properties used for the fitting. For a calculated property x and a corresponding experimental value of x_{exp}, the assigned score is obtained as [202]

$$M = max\{[10 - |(x - x_{exp}) \times 100/(x_{exp}tol)|], 0\} \tag{4.8}$$

where the tolerance (tol) is assigned to 0.5% for density, position of the first peak in the RDF and heat of vaporization, 5% for height of the first peak in the RDF, and 2.5% for the remaining properties. The quality score assigned to each test model is equal to the average of the scores in bulk properties considered.

The result of the above search procedure is a "quality map" of all possible water models in the μ-Q_T space(Figure 4.2): the proposed model is the one with the highest quality score.

4.3.2 Simulation details

Unless specified otherwise, we use the following Molecular Dynamics (MD) simulations protocol. Simulations in the NPT ensemble (1 bar, 298.16 K) were carried out using the Amber14 MD software package [35]. A cubic box with edge length of 30Å was filled with 804 water molecules. Periodic boundary condition was implemented in all directions. Long-range electrostatic interactions, calculated via the particle mesh Ewald (PME) summation, and the van der Waals interactions were cut off at distance 8Å , the van der Waals interactions beyond the cut off distance is accounted for via a continuum model (vdwmeth=1 in Amber) [35]. Dynamics were conducted with a 2 fs time step and all intra-molecular geometries were constrained with SHAKE. The NPT simulations were performed using Langevin thermostat with a coupling constant 2.0 ps^{-1} and a Berendsen barostat with coupling constant of 1.0 ps^{-1} for equilibration and 3.0 ps^{-1} for production.

We perform a 3-tier search for best fit in the 2D space of (μ, Q_T), Figure 4.2. The initial search is on 0.05 $\times$ 0.05 grid in the 2D (μ, Q_T) space shown in Figure 4.2. The refinement is on 0.01 $\times$ 0.01 grid, limited to vicinity of the optimum (green area), followed by the final comparison of a few candidates within a very small area (dark green).

Each calculated water property is an average over 6 independent MD trajectories; 6 ns each for the first stage, 15 ns for the second, and 65 ns for the final stage and all of the properties of water models shown here. These long simulation times are needed to obtain well-converged averages of some properties, such as static dielectric constant[217].

4.4 Results

4.4.1 The proposed OPC3 model

The result of the above search procedure is a "quality map" of all possible test water models in the $\mu - Q_T$ space, Figure 4.2: the proposed OPC3 model is the one with the highest quality score. Compared to the commonly used 3-point models shown, OPC3 reproduces the multipole moments of water molecule in the liquid phase substantially better: the OPC3's μ (2.43 D) and Q_T (2.06 DÅ) are in best agreement with the range of values from experiment [78] and QM calculations [181, 184, 151, 44]. The improved moments of OPC3 are achieved due to an"unconstrained" search for model's optimal parameters in the space of moments.

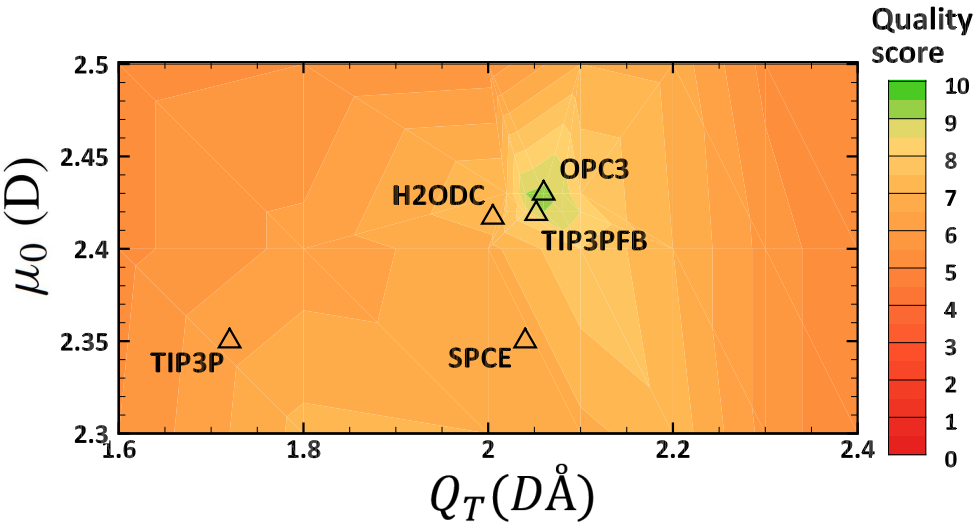


Figure 4.2: The quality score distribution of test water models in the space of dipole (μ) and quadrupole (Q_T). Each fine grain point on the plot represents a model tested. Scores (from 0 to 10) are calculated based on the accuracy of predicted values for six key properties of liquid water (see text). The resulting proposed optimal model is termed OPC3. For reference, the μ and Q_T values of commonly used and recently developed 3-point water models (triangles, quality score given by the color at the symbol position) are placed on the same map (see also Table 4.1).

The OPC3 point charge positions and values and the LJ parameters are listed in Table 4.2. The $|O - q^+|$ distances for OPC3 is slightly longer (0.97888Å) than the corresponding experimental values of $|O - H|$ bond. The $\angle\, q^+Oq^+$ angle (Figure 4.1) is equal to (109.4712°) which is a direct consequence of setting $Q_0 = 0$ (see Section 4.3).

4.4.2 Bulk properties at 298.16 K, 1 bar

The quality of the OPC3 model in reproducing experimental bulk water properties at ambient conditions, and a comparison with other most commonly used and recently developed 3-point models is presented in Table 4.3. For each of 10 key liquid properties (Table 4.3) against which water models are most often benchmarked [202, 203, 85], the deviations of OPC3's computed properties from the corresponding experimental values is less than 6%, except for one property (thermal expansion coefficient) that deviates from experiment by about 67.9% (see Figure 4.3). Note that the parametrization of OPC3 involved a fitting to only 5 of the properties reported in Table 4.3, yet the model is accurate in estimating several other properties that were not included in the optimization, such as isobaric heat capacity C_p, isothermal compressibility κ_T, thermal expansion coefficient α_p, and the temperature of maximum density (TMD).

The O-O radial distribution functions (RDF), $g(r_{OO})$ for the OPC3 are presented in Figure 4.4. By design, the experimental position of first peak in O-O RDF is accurately reproduced by OPC3. The position and height of other peaks are also closely reproduced.

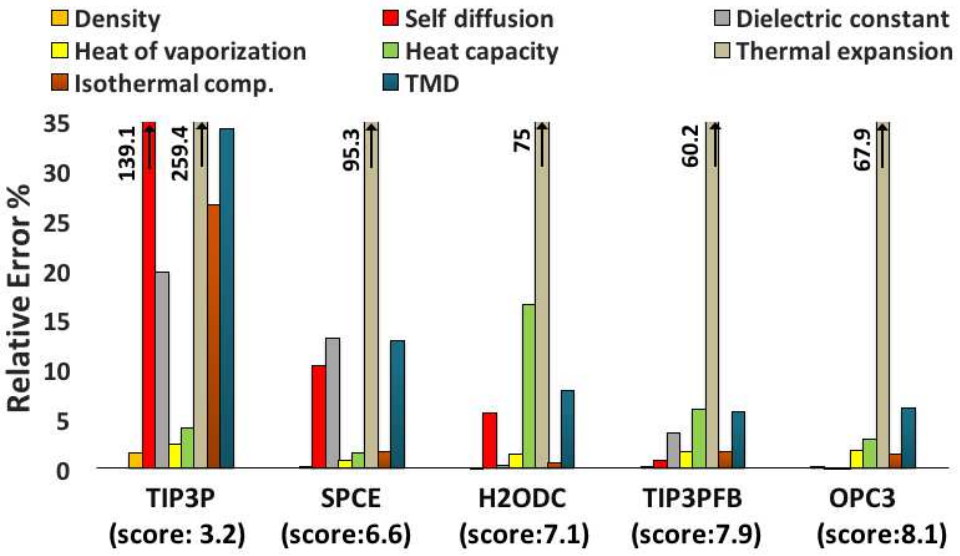


Figure 4.3: Comparing the accuracy of OPC3 to some old and recent rigid 3-point water models TIP3P, SPCE, H2ODC, and TIP3PFB [211]. The quality scores (see *Methods*) represent the overall performance of each model in reproducing eight key properties, i.e. density ρ, self diffusion coefficient D, static dielectric constant ϵ_0, heat of vaporization ΔH_{vap}, isobaric heat capacity C_p, isothermal compressibility κ_T and thermal expansion coefficient α_p, at ambient conditions, as well as the temperature of maximum density (TMD).

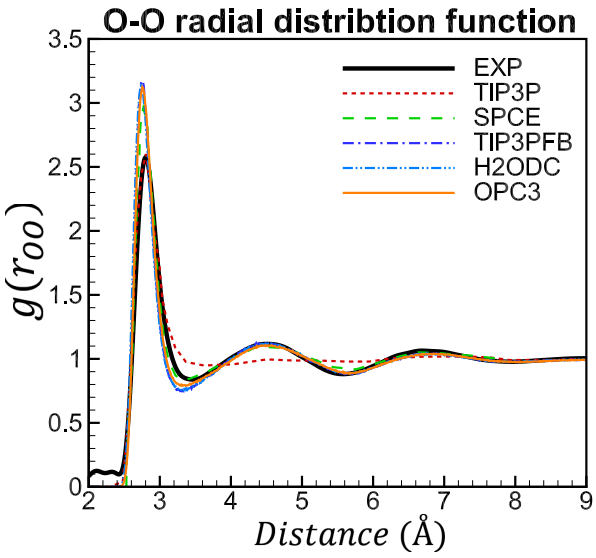


Figure 4.4: O-O radial distribution functions of liquid water at 298.16 K, 1 bar. The OPC3 model is compared to the commonly used 3-point models (TIP3P and SPCE).

4.4.3 Temperature dependent behavior

The ability of OPC3 to reproduce the temperature dependence of four key water properties is shown in Figure 4.5. It is noteworthy that OPC3, which resulted from a search in the space of only two parameters (μ and Q_T) at only one thermodynamic condition (298.16 K and 1 bar) automatically reproduces bulk properties with a high accuracy across a wide range of temperatures where no fitting was performed. The procedure and the result are in contrast not only to commonly used, but also to some recent rigid [211, 66, 2] that generally employ massive and more specialized fits against multiple properties over a wide range of thermodynamic conditions.

4.4.4 A consensus in the parametrization of 3-point rigid models

Here we compare OPC3 with two recent 3-point water models: TIP3PFB[211] and H2ODC[60], which are parametrized using completely different strategies. TIP3PFB is developed based on the state-of-the-art ForceBalace parametrization method[211], which essentially evaluates the simulated properties in NPT ensemble and calculates their parametric derivatives to use in the optimization. The search is performed in the space of the bond length, angle, charge and van der waals parameters, using the geometry of TIP3P and SPC/E as the starting point. H2ODC is specifically designed to reproduce the correct experimental dielectric constants, in addition to more common target properties such as density and enthalpy of vaporization[60]. The starting geometry of H2ODC is a fixed tetrahedral angle and a bond

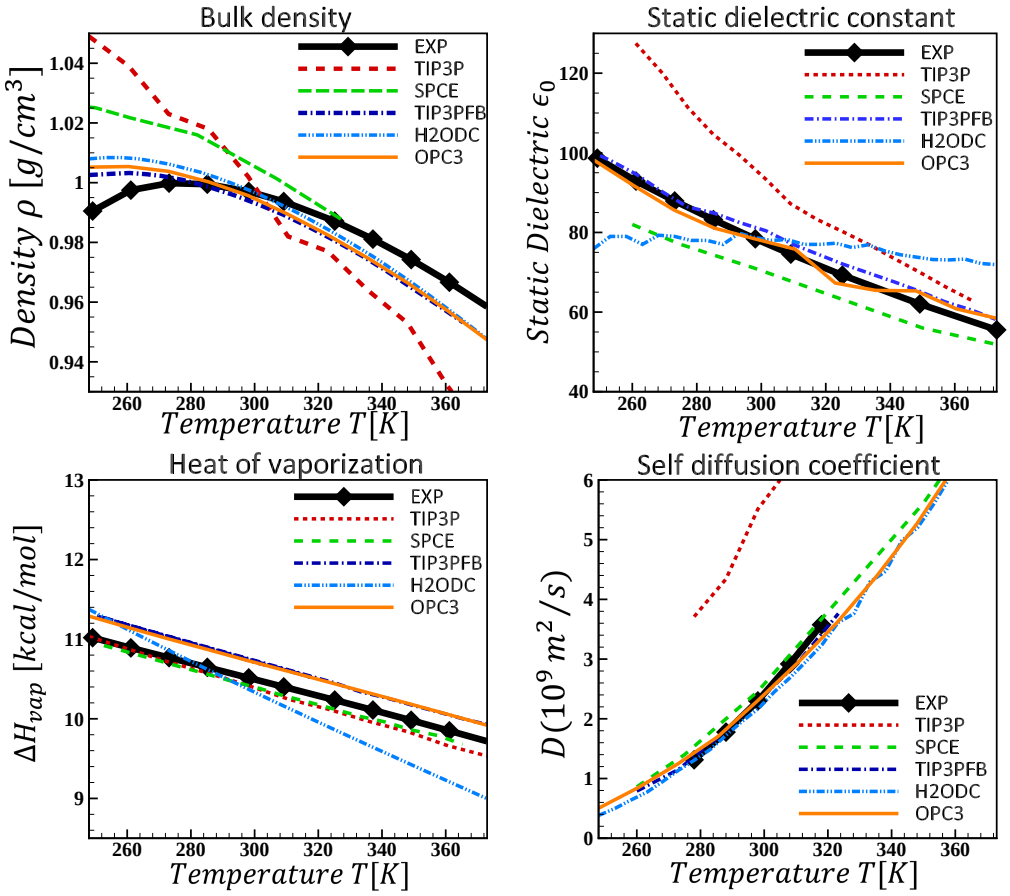


Figure 4.5: Calculated temperature dependence of water properties for OPC3 compared to two most commonly use and two recent 3-point water models and experiment. TIP3P results are from [104, 203, 211, 105], SPCE from [54, 211], TIP4PFB from [211]. H2ODC results are calculated based on the protocols described in this work.

length equal to the experimental gas-phase value (Table 4.2). The search is performed in the space of σ_{LJ}, ϵ_{LJ} and q that are uniformly scaled to fit experimental values for density, heat of vaporization and dielectric constant, respectively. In contrast to these two models that are obtained from a search in the space of length-angle-charge, OPC3 is derived from a search in the space of key multipole moments of the model (dipole and square quadruple) so that six key properties are best reproduced, without any geometry constraints on the model other than the symmetry.

For comparison, the μ and Q_T values of two most recently developed 3-point water models (TIP3PFB[211] and H2ODC[60]) are placed on the quality score map in Figure 4.2 (see also Table 4.1). Interestingly, the placements of all these three models cluster around the small high quality region in the map. Overall, there is a close agreement between the model parameters(Table 4.2), multipole moments (Table 4.1) and bulk properties (Table 4.3 and Figure 4.5) for these models. Given the fundamental differences in the optimization of these three models, close agreement of the model parameters and accuracy of OPC3 with values from H2ODC and TIP3PFB points out to a consensus for the optimal parametrization of 3-point water models.

4.4.5 How does OPC3 compare to OPC?

Although both OPC3 and OPC models perform well in reproducing the key bulk properties at ambient condition and also in representing the temperature behavior of water, OPC is much more accurate than OPC3 in reproducing the density of water at lower temperatures. As a result, OPC is also more accurate in estimating the thermal expansion, the deviation from experimental value is 5% for OPC, which is much lower than 67.9% for OPC3. The discrepancy between the accuracy of these two models, in particular at lower temperatures, stems from the difference in their abilities in accurately representing the multipole moments of the water molecule. For example, the dipole moment for OPC is slightly larger than that of OPC3 ($2.48D$ vs $2.43D$, respectively). A more profound distinction is seen in the value of square quadrupole Q_T, which is $2.3D$Å and $2.06D$Å for OPC and OPC3, respectively. Given that the analytical expressions that relate charge distributions of both three and four point models to their multipole moments can exactly reproduce a given set of dipole and quadrupole moments (see section 4.3), the question arises why the optimal value of Q_T for OPC3 is much smaller than that for OPC. Note that a 3-point model is subjected to an additional geometry constraint compared to a 4-point model: the negative charge is fixed to the position of oxygen (see section 4.3). Due to the additional constraint, a 3-point OPC3 has no control over the accuracy of its moments beyond the quadrupole, whereas the 4-point OPC can optimally reproduce a given set of moments up to the octupole[93, 9]. As a result, although the dipole and quadrupole moments of a 3-point model can be set to more accurate (i. e. larger) values, doing so can introduce severe errors in the representation of its higher order moments (e. g. octupole). The OPC3' moments represent the best compromise between the dipole, quadrupole and octupole moments that is presently achievable by a

3-point model (Table 4.1). In fact, the octupole moments of OPC3 are in good agreement with the values from QM (Table 4.1), but probably with the cost of smaller value for its quadrupole. By construction, a 4-point model can provide a better balance between the three lowest order moments of the model, that are in better agreement with values predicted by QM.

Another significant difference is seen in the dispersion parameters of the OPC and OPC3 models. The London dispersion parameters for OPC3 are significantly smaller than that of OPC (858.1 vs 668.64) which is probably a consequence of a balance between OPC3's weaker electrostatic interactions, caused by its smaller moments, and the van der waals interactions. The weaker electrostatic and van der waals interactions can influence the ability of OPC3 to form strong hydrogen bonds, in particular at lower temperatures.

We speculate that the poorer performance of OPC3 in reproducing the temperature dependent properties can be due to its relatively much lower value of (Q_T) [151], compared to that of OPC. *"The large quadrupole of water molecules"*[151], mainly controlled by the Q_T value, has been known to have a strong effect on the phase diagram and temperature dependent properties of water [3, 151]. The contribution of the higher order multipole moments to electrostatic potential can be significant at close distances, which are relevant to water-water and water-ion interactions in liquid phase. This can specifically influence the models accuracy at lower temperatures..

The difference in the strength of water-water interactions in OPC and OPC3 can lead to different performances beyond water bulk properties, in biomolecular simulations. For example, it has been shown that water dispersion interactions can strongly affect simulated structural properties of disordered protein states[160]. OPC has been shown to introduce improvements in predicting the solvation free energies of small molecules, binding free energies, RNA simulations. Nevertheless, given that what ultimately matters in biomolecular simulations is an appropriate balance between water-water, water-protein and protein-protein, whether OPC or OPC3 turns out to be more suitable for biomolecular simulation is still to be further tested.

4.5 Conclusion

Recently we proposed a new approach to constructing point charge water models. The novelty of the approach is that a search for optimal parameters of fixed-point charge models is performed in the electrostatically most relevant, low-dimensional sub-space of lowest multipole moments, rather than in the convoluted high-dimensional charges-distances-angles space "native" to point-charge models. The models constructed based on the new approach can be parametrized with electrostatic multipole moments that are closer to what has been found in experiment and other calculations, which is difficult to do in models with fixed geometry.

Here we show that the OPC3 model developed based on this approach gives significantly better agreement with experimental bulk water properties, compared to most commonly used 3-point models. OPC3 is made to reproduce the bulk properties at only one temperature (298K), yet it improves thermodynamic and dynamic properties over a wide range of temperature.

A comparison of the OPC3 model with two recent models developed based on completely different parametrization procedures indicates a consensus for the optimal parametrization of 3-point water models. Given that very different parameter optimizations, including a virtually exhaustive search in the "appropriate" electrostatic parameter space, yield essentially the same result, we conclude that the search for an optimal 3-point rigid water model is over.

The OPC3 model is planned to be included in the 2017 version release AMBER software package, and has been tested in GROMACS 4.6.5. A 0.5 microsecond MD simulation of ubiquitin (PDB ID: 1UBQ) solvated in OPC3 water was stable and gives a backbone RMS deviation from the starting crystal structure smaller than 1Å .

Table 4.2: Force field parameters of OPC3 and some commonly used and also recently developed 3-point water models, where $\sigma_{LJ} = (A_{LJ}/B_{LJ})^{1/6}$ and $\epsilon_{LJ} = B_{LJ}^2/(4A_{LJ})$. For comparison, water molecule geometry in the gas phase is also included.

	$q[e]$	l[Å]	$\Theta[deg]$	σ_{LJ}[Å]	$\epsilon_{LJ}[kJ/mol]$
EXP(gas)	NA	0.9572	104.52	NA	NA
TIP3P	0.417	0.9572	104.52	3.15061	0.6364
SPCE	0.4238	1.0	109.47	3.166	0.65
TIP3PFB	0.42422	1.0118	108.15	3.1780	0.65214
H2ODC	0.4238	0.958	109.47	3.18400	0.593
OPC	**0.447585**	**0.97888**	**109.4712**	**3.17427**	**0.68369**

Table 4.3: Model vs. experimental bulk properties of water at ambient conditions (298.16 K, 1 bar): dipole μ, density ρ, static dielectric constant ϵ_0, self diffusion coefficient D, heat of vaporization ΔH_{vap}, first peak position in the RDF $roo1$, isobaric heat capacity C_p, thermal expansion coefficient α_p, and isothermal compressibility κ_T. The temperature of maximum density (TMD) is also shown. Bold fonts denote the values that are closest to the corresponding experimental data (EXP). Statistical uncertainties (±) are given where appropriate.

Property	TIP3P [202, 138]	SPCE [202, 211]	TIP3PFB [211]	H2ODC	**OPC3**	EXP [202, 203, 185]
$\mu(D)$	2.348	2.352	2.42	2.42	**2.43**	2.5–3
$\rho[g/cm^3]$	0.980	0.994	0.995	0.9975	**0.996±0.001**	0.997
ϵ_0	94	68	81.3	78.7	**78.4±1**	78.4
$D[10^9 m^2/s]$	5.5	2.54	2.28	2.17	**2.30±0.02**	2.3
$\Delta H_{vap}[kcal/mol]$	10.26	10.43	10.71	10.36	**10.73±0.004**	10.52
$roo1$[Å]	2.77	2.75	2.755	-	**2.755**	2.755
$C_p[cal/(K.mol)]$	18.74	20.7	19.1	20.98	**18.54±0.05**	18
$\alpha_p[10^{-4}K^{-1}]$	9.2	5.0	4.1	4.48	**4.3±0.1**	2.56
$\kappa_T[10^{-6}bar^{-1}]$	57.4	46.1	44.5	45.0	**46.0±1**	45.3
$TMD[K]$	182	241	261	255	260	277

Chapter 5

Optimal point charge approximation for MD simulation of million-atom systems: insights into the structure of chromatin fiber

5.1 Overview

Molecular Dynamics (MD) simulations based on the generalized Born (GB) models can provide significant computational advantages over the traditional explicit solvent simulations. At the same time, standard GB becomes prohibitively expensive for all-atom simulations of very large structures; the model's very high computational cost in this case stems from its poor scaling which is $\sim n^2$, where n is the number of solute atoms. Here, we combine our recently developed Optimal Point Charge Approximation (OPCA) with the $\sim nlogn$ Hierarchical Charge Partitioning (HCP) approximation to present a multi-scale, yet fully atomistic, approach to perform MD simulations based on the generalized Born model, called GB-HCPO. HCP exploits the natural organization of biomolecules (atoms, groups, chains, and complexes) to partition the structure into multiple hierarchical levels of components. OPCA approximates the charge distribution for each of these components by a small number of point charges that optimally reproduce the low order multipole moments of these components. GB-HCPO uses the full set of atomic charges to compute interactions between nearby atoms, while approximating interactions with distant components using the smaller set of charges. We show that GB-HCPO significantly improves the accuracy of electrostatic forces and energies when compared to the more common cutoff GB and also its predecessor GB-HCP model (without OPCA). The higher accuracy of GB-HCPO in representing the electrostatic interactions translates to increased structural stability in predicted biomolecules structures. GB-HCPO can deliver over two orders of magnitude speedup compared to the

reference GB, and, for large structures, can provide the same nominal speed, as in nanoseconds per day, as the highly optimized explicit-solvent (TIP3P) based on particle mesh Ewald (PME). The increase in the nominal simulation speed coupled with substantially faster sampling of conformational space, relative to the explicit solvent, makes GB-HCPO suitable for simulating very large systems. As a practical demonstration, we employ GB-HCPO to perform a multi-scale all-atom simulation of 30-nm chromatin fiber (40 nucleosomes, over 1 million atoms), starting from a manually constructed 4-star model consistent with low resolution cryo-EM data. Our results suggest important structural details consistent with experiment: the linker DNA fills the core region and the H3 histone tails interact with the linker DNA. GB-HCPO is implemented in the open source MD software, NAB in AMBER 16.

5.2 Introduction

Atomistic simulation is one of the most widely used theoretical tools in bio-medical research[4, 111, 114]. High accuracy of solvent representation in these simulations is paramount for biological applications. Arguably the most accurate among classical models of solvation is the one in which individual water molecules are treated explicitly on the same footing with the target biomolecule [102, 23, 85, 94] Yet, accuracy of this explicit solvent representation comes at extremely high price, computationally. For example, a recent study[130] of folding of 17 of the fastest folding proteins required extremely long simulations on one-of-a-kind specialized supercomputer. Likewise, efficient estimates of free energies, particularly important in many areas including structure-based drug design[103], can take hours or even days per structure[5], or may not even be possible for large structures due to convergence issues. An alternative that mitigates or eliminates these problems is the implicit (or continuum) solvation model that treats solvent implicitly as a continuum with dielectric and "hydrophobic" properties of water [46, 84, 26, 137, 70, 173, 136, 183, 17, 129]. A particularly computationally efficient version of the model, the so-called generalized Born (GB) model, is the current "workhorse" in Molecular Dynamics (MD) [182, 39, 100, 223, 101, 127, 161, 97, 8] including QM/MM[158, 205] and REMD simulations[153].

GB provides significant computational advantage over explicit solvent MD simulations in two main ways: First, GB approximates the discrete solvent as a continuum, thus drastically reducing the number of particles to keep track of in the system [188, 21, 81, 82, 69, 126, 155, 199, 45, 49, 92, 174, 33, 58, 14, 57, 152, 173, 52, 67, 75, 80, 187, 201, 194, 224], which significantly increases the nominal speed (nanoseconds per day of simulation time). Second, the GB model can sample conformational space substantially faster than explicit solvent model due to the reduction of solvent viscosity. For instance, it has been shown that the conformational sampling in implicit solvent simulations can be $\sim$100 times faster compared to common explicit solvent PME simulations [56, 8, 198, 222, 11]. The combination of these two effects makes the GB model suitable for all-atom MD simulations of the protein folding

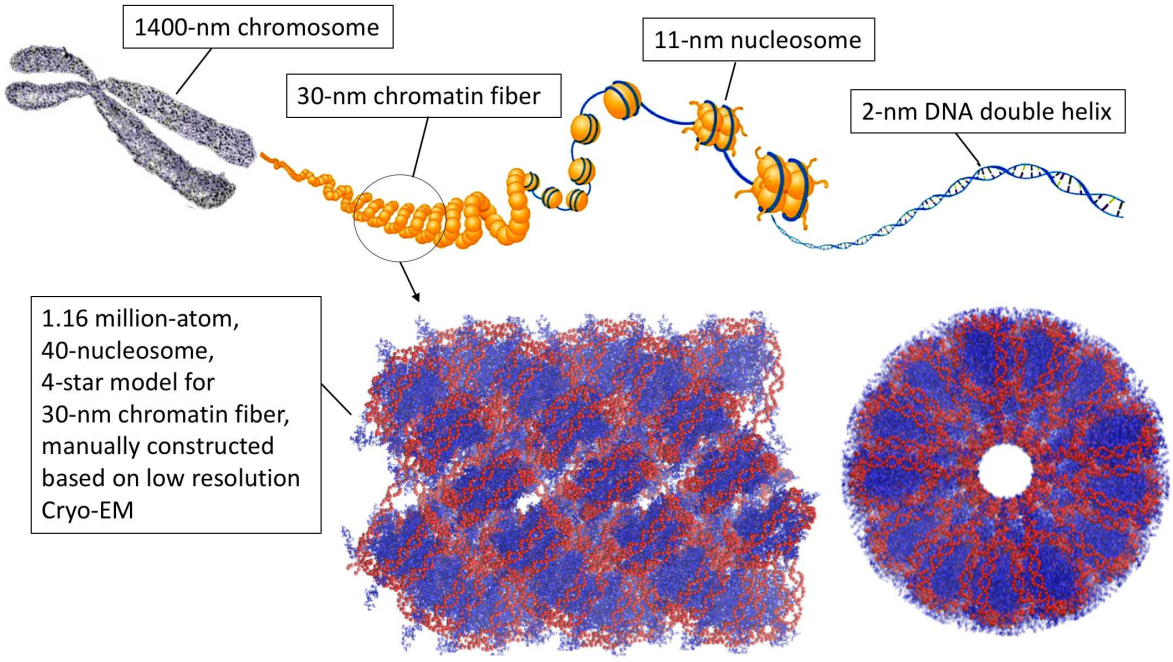


Figure 5.1: ***Top:*** Chromatin fiber is the next hierarchical level of DNA compaction beyond the nucleosome: the 2-nm DNA helices wrap around histones to form 11-nm nuclesome. The nulceosomes are considered to be regularly wrapped into 30-nm-diameter chromatin fibers. The chromatin fibers are further packed to make up the chromosomes. ***Bottom:*** A 40-nucleosome (1.16 million atom) 30-nm chromatin fiber manually constructed based on a 4-start model consistent with low resolution cryo-EM data[215].

process [182, 39, 100, 223, 101, 127, 161, 97], including protein design [133, 59]. Folding study of the same set of 17 proteins mentioned above [130] can be performed on a commodity PC within days. [150]. Correct native states of small proteins can be easily identified as minimum energy snapshots [182], in straightforward simulations starting from completely extended conformations – the task that is not nearly as straightforward in explicit solvent.

Despite these advantages, GB still remains a less widely used for simulating very large structures in part due to its poor algorithmic complexity: the functional form for the most widely used practical GB models scales as $\sim n^2$, where n is the number of solute atoms only. Therefore, GB-based simulations can become too slow for large and especially very large (100,000s atoms and more) structures, essentially negating all of the potential benefits of the GB approach. This is while the most common explicit-solvent simulations based on the "industry standard" particle mesh Ewald (PME)[48, 55, 196, 220], or the fast multipole[34, 32, 121], scales as $\sim NlogN$, where N is the total number of solvent and solute atoms combined. As a result, the nominal computational speed of the GB model, i.e. the number of nanoseconds per day of simulation time, can be much lower than that of the corresponding explicit-solvent simulation for large systems. In cases where large-scale atomistic level modeling is desired, practitioners have no choice but to resort to the traditional, explicit solvent approach which leads to inordinate computational costs. For example, a recent study of the complete HIV-1 capsid model through a fully solvated, unconstrained 100 ns, 64 million atom simulation takes advantage of one of the most powerful supercomputers in the world[225]. A typical computational lab with modest resources can not afford to simulate even smaller structures (100,000s to a million) atoms long enough to observe meaningful conformational changes.

A common approach for reducing the computational cost of GB models is to apply the concept of spherical cutoff, i.e. ignore interactions and computations involving atoms beyond a cutoff distance, referred to as cutoff-GB here. The cutoff-GB reduces the computational cost of traditional GB from $\sim n^2$ to $\sim nlogn$. However, the cutoff-GB approach can lead to unacceptable errors and severe artifacts, such as spurious forces and artificial structures around the cutoff distance[141, 132, 178], especially when the structures are highly charged.

In this study, we combine our newly developed Optimal Point Charge Approximation (OPCA) [9] with a $\sim nlogn$ multi-scale approximation for GB [10] to present a multi-scale, yet fully atomistic, GB model for very large structures, referred to as GB-HCPO. We evaluate the accuracy, speed, and stability of the GB-HCPO on a set of representative biomolecular structures ranging in size from 632 to 1159998 atoms, with absolute total charge ranging from 1 to 8238 e, compared to the commonly used spherical cutoff method with GB (cutoff-GB) and its predecessor model without OPCA (GB-HCP).

Gene expression is regulated, in part, by the organization of chromatin fiber, which is the next hierarchical level of chromatin compaction beyond the nucleosome[134] (Figure 5.1). Modifications to N-terminal tails of the histone proteins that make up the fiber core are known to regulate DNA accessibility and affect vital process[83]. Due to the large size of the fiber –millions of atoms– only low-resolution (cryo-EM) experimental structures of the fiber

are available[170], its atomistic details including the tails, are unknown. Computational studies investigating the organization of chromatin fiber have typically used coarse-grain simulations. Such simulations use customized, relatively unproven, force fields, and fail to elicit the finer details of the atom level structure. Starting from an existing atomistic model[215] consistent with low resolution cryo-EM data (Figure 5.1), here we use GB-HCPO to perform multiscale "all atom" simulations of 40 nucleosome (over 1 million atom) 30-nm chromatin fiber to study its structure and response to modifications such as post-translational modifications implicated in chromatin remodeling.

The remainder of the paper is organized as follows. In the Methods section, a brief description of previously developed GB-HCP without OPCA is provided. That is followed by a description of the implementation of OPCA in GB-HCP (GB-HCPO). The details of the simulations protocols and test structures are also given in this section. In the Results section, first a practical application of GB-HCPO to simulate 40 nucleosome chromatin fiber is provided. Then a detailed technical analysis of the accuracy and speed of the new GB-HCPO is presented. In the Conclusion section we summarize our finding and discuss the applicability of GB-HCPO to practical MD problems.

5.3 Methods

A short description of the GB model without further approximation is provided below. Then we briefly explain the multi-scale Hierarchical Charge Partitioning (HCP) approach that reduces the algorithmic complicity of the GB model from $\sim n^2$ to $\sim nlogn$. That is followed by a detailed description of the implementation of the OPCA approach into the HCP implementation of the GB model. The details of the simulations protocols and test structures are also given in this section.

5.3.1 The GB model without further approximation

The electrostatic energy E^{elec} of a solvated system can be estimated by the GB implicit solvent model[146] as:

$$E^{elec} = E^{vac} + E^{solv} \tag{5.1}$$

$$E^{vac} = \sum_{i}^{n} \sum_{j>i}^{n} \frac{q_i q_j}{r_{ij}} \tag{5.2}$$

$$E^{solv} \approx -\frac{1}{2}\left(1 - \frac{1}{\epsilon_w}\right) \sum_{i}^{n} \sum_{j}^{n} \frac{q_i q_j}{f_{ij}^{GB}(r_{ij})} \tag{5.3}$$

where E^{vac} and E^{solv} are the electrostatic vacuum and solvation energy, ϵ_w is the dielectric constant of the solvent, q_i and q_j are the charges of atoms i and j, and r_{ij} is the distance between the atoms. We use the most widely functional form of $f_{ij}^{GB}(r_{ij}) = [r_{ij}^2 + B_i B_j e^{(-r_{ij}^2/4B_iB_j)}]^{1/2}$, and B_i and B_j are the so-called effective Born radii. For the case where the atoms i and j do not overlap the effective Born radii, B_i, can be calculated by the following equation:

$$\frac{1}{B_i} \approx \frac{1}{R_i} - \sum_{j \neq i} \left[\frac{R_j}{2(r_{ij}^2 - R_j^2)} - \frac{1}{4r_{ij}} \log \frac{r_{ij} + R_j}{r_{ij} - R_j} \right] \tag{5.4}$$

where R_i is the intrinsic radius of atom i, r_{ij} is the distance from i to any point j in the solute volume. As shown in Eqs. 5.2 and 5.3, the calculations of E^{vac} and E^{solv} scale as $\sim n^2$. The computational implementation of analytical pairwise approximation to the effective Born radii B_i, Eq. 5.4, also scales as $\sim n^2$.

5.3.2 Applying Hierarchical Charge Partitioning (HCP) to the GB model

Elsewhere[10], it was shown that a $\sim nlogn$ approximation of the GB model can be achieved by resorting to the concepts of the Hierarchical Charge Partitioning (HCP)[13]. The HCP approximation partitions the biomolecular structures into multi-level hierarchical components based on the natural organization of biomolecules, as illustrated in Figure 5.2 – *atoms* (level 0), nucleic and amino acid *groups* (level 1), protein, DNA and RNA *subunits* (level 2), *complexes* of multiple subunits (level 3), and higher level structures such as fibers and membranes. The charge distribution for each component beyond level 0 is represented by 1, or 2, approximate point charges placed at the center of charge of all, or positive/negative, charges of the component, respectively. The HCP method uses the approximate point charges for computations involving distant components, while the full set of atomic point charges are used for computations involving nearby components (Figure 5.2). The level of approximation used is determined by the distance of a component from the point of interest compared to the threshold distance for the level of the component, for example h_1, h_2, h_3 for level 1, 2 and 3 in Figure 5.2, respectively. The greater the distance from the point of interest, the larger (higher level) is the component used in the approximation. The computational cost of the HCP approximation for computing the pair-wise electrostatic interactions scales as $\sim n \log n$. For a more detailed description, refer to the previous HCP studies [13, 10].

The concept of HCP approximation can be used to reduce the computational cost of Eqs. 5.2, 5.3, and 5.4 from $\sim n^2$ to $\sim n \log n$, as described below. For example, in the case of 2 charge approximation for the components, for each component c beyond level 0 (groups, strands and complexes), 2 approximate point charges (q_{c1} and q_{c2}) are used in place of q_j charges to approximate the electrostatic interactions of the corresponding component c with

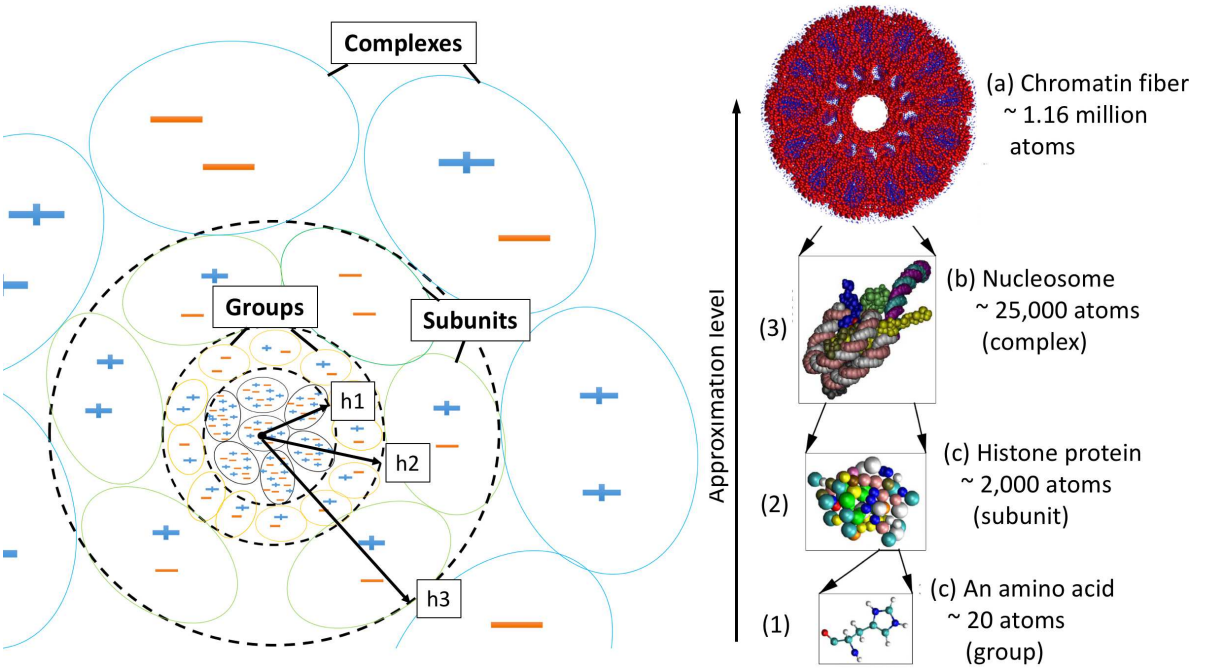


Figure 5.2: ***Left:*** Illustration of hierarchical charge partitioning (HCP) for three levels of approximation. Here, h1, h2 and h3 are the level 1 (group), level 2 (subunit) and level 3 (complexes) threshold distances, respectively. The distance to a component is computed from the point of interest to the geometric center of the component. ***Right:*** Multi-level hierarchical partitioning of a 30 nm chromatin fiber based on its natural structural organization: (a) The fiber is made up of 40 nucleosome complexes. The individual nucleotide groups in the fiber are shown in red and amino acid groups in blue. (b) Each complex (level 3) is made up of 13 subunits with the segments of DNA linking nucleosome complexes being treated as separate subunits. A complex is shown here with each subunit represented in a different color. (c) Each subunit (level 2) is made up of 49-142 groups. The linker histone subunit is shown here with the groups colored by the type of amino acid. (d) Each group (level 1) is made up of 7-32 atoms (level 0). A histidine amino acid group is shown here with atoms represented as small spheres and covalent bonds between the atoms represented as links. The atoms are colored by the type of atom. The total fiber consists of 1159998 atoms. The fiber was constructed as described in Wong et. al.[214]. The images were rendered using VMD [89]. For clarity, only 10 of the 13 subunits are shown in (a) and (b).

charge q_i (Eqs. 5.2 and 5.3). The values of q_{c1} and q_{c2}, and their positions (r_{c1} and r_{c2}) are obtained from the HCP approach described above. Also, the component radius R_c is used in place of the intrinsic radius of atom j (R_j) (Eq. 5.4). R_c is the radius of a sphere with the same volume as the sum of volumes of its constituent atoms, that can be estimated as below[10],

$$R_c^3 \approx \sum_{j \in c} R_j^3 \tag{5.5}$$

Introducing approximating point charges for components necessitates an equivalent of the effective Born radii to be associated with each of the approximate point charges ($\boldsymbol{B_{c1}}$ and $\boldsymbol{B_{c2}}$), and to be used in place of B_j (Eq. 5.4). The equivalent of the effective Born radii for components are approximated by a simple harmonic average of the effective Born radii of atoms within the component, weighted by their atomic charges, as below[10]

$$\frac{1}{B_{ci}} \approx \left[\frac{1}{q_{ci}} \sum_{k \in c} \frac{q_k}{{B_k}^{1/2}} \right]^2, i = 1, 2 \tag{5.6}$$

where q_k and B_k are the charges and effective Born radii, respectively, for the positively charged atoms or the negatively charged atoms belonging to the corresponding component ($c1$ or $c2$ for 2-charge approximation).

The GB model approximated as above, referred to as GB-HCP, scales as $\sim$ *nlogn* for biomolecular structures[10].

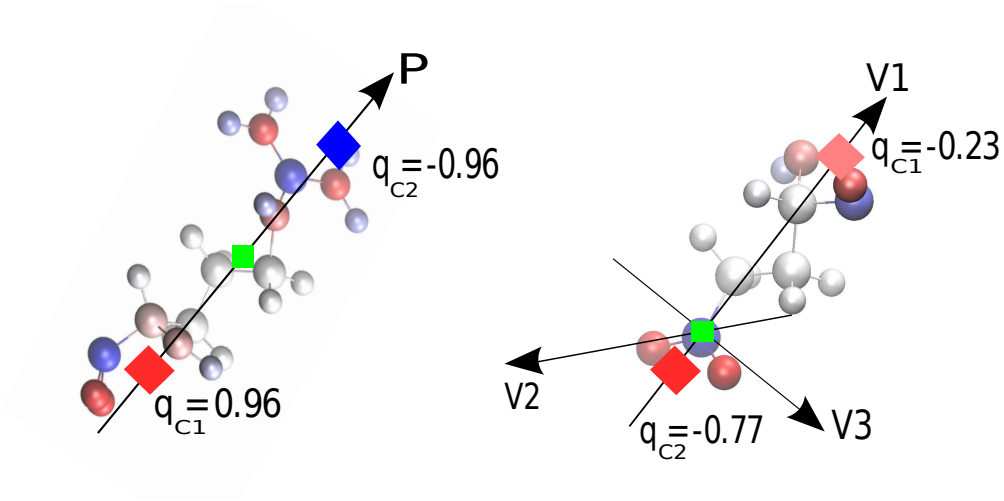
5.3.3 The GB-HCP based on the Optimal Point Charge Approximation: GB-HCPO

The accuracy of the *nlogn* approximation of the GB model explained above is greatly influenced by the magnitudes and positions of the approximating point charges (q_{c1} and q_{c2}) used to represent distant electrostatic interactions. A simple placement of the approximating point charges at the center of positively and negatively charged groups does not accurately reproduce the complexity of electrostatic potential around the original component[9]. Below we apply the Optimal Point Charge Approximation (OPCA) [9] method to optimize placement of the approximating point charges in the HCP approximation of GB. The implementation of the OPCA approach necessitates a new expression for the component effective Born radii, as described below.

Optimal Point Charge Approximation (OPCA)

The Optimal Point Charge Approximation (OPCA) [9] optimizes the placement of the approximating point charges so that the key lowest order multipole moments of the original charge distribution (hierarchical components) are best reproduced. A general framework for numerically computing OPCA, for any given number of approximating charges is described previously[9]. The most physically significant lowest order multipole moments of the biomolecules are often the monopole, the dipole and the quadrupole. A set of two approximating point charges calculated based on OPCA exactly reproduces the monopole (for charged structures) and the dipole, and optimally approximates the quadrupole moments of the original charge distribution (minimum rms error). The 2-charge OPCAs can be calculated via closed-form analytical expressions, which provide computational efficiency and algorithmic simplicity for MD simulations. Therefore, here we use 2-charge OPCA for approximating the hierarchical level of components in the HCP approximation of GB. Due to important differences in the characteristics of charged and uncharged components [9], it is necessary to treat these two cases separately, as briefly described below (Figure 5.3).

OPCA for Uncharged Components For uncharged components, the monopole and dipole moments are exactly reproduced when a pair of charges of equal magnitude but opposite sign are aligned with the direction of the dipole moment of the original charge distribution. The quadrupole moment for uncharged charge distributions is optimally reproduced if the geometrical center of the optimal point charges coincides with the center of dipole for the original charge distribution (green square in Figure 5.3(a)). This leaves only one unknown parameter, the separation between the charges $dr_n = |r_2 - r_1|$, where r_1 and r_2 are the position vectors of the point charges. The dr_n value is chosen so that the octupole moment is optimally reproduced. Depending on whether or not there exists an analytical solution to the equation that relates the OPCA charges to the octupole moment of the original charge distribution, dr_n can vary from a very small value ($dr_n << R_0$) to a value close to R_0 ($dr_n \sim R_0$) [9], where R_0 is the distance of the furthest charge from the center of geometry[9]. For practical applications it is computationally more efficient to use an empirically determined value for dr_n [9]. For each test case described in the Results section below, we varied the value of dr_n so that the RMS error in force (explained below) is minimized. Our testing suggests that dr_n in the range of $0.1R_0$ and $0.4R_0$ to be optimal for the force calculations for the test cases studied here. Given the diversity of the structures sizes studied here (Table 5.1), we suggest that the same range of dr_n values can be applied for structures other than those studied here. For practical applications, the recommended procedure for setting dr_n is to test accuracy of force calculations for the starting configuration of the structure in the above range ($0.1R_0 < dr_n < 0.4R_0$) in increments of $0.05R_0$. An illustration of 2-charge OPCA for a sample uncharged distribution is given in Figure 5.3(a).



(a) uncharged (b) charged

Figure 5.3: Illustration of a 2-charge optimal point charge approximation. (a) A sample charge distribution- a neutral amino acid (C-terminal arginine at physiological pH.). The 2 optimal point charges (red and blue diamonds) are placed in equal distances (**dr**)) from the center of dipole (green square) of the original charge distribution, along the dipole moment direction of the original charge distribution. (b) A sample charge distribution with non-zero net charge (a glumatic acid group within a protein with net charge = -1 e). The 2-charge optimal point charges (red diamonds) are placed so that their center of charge matches the center of charge of the original charge distribution (the green square), along the eigen vector of the quadrupole moment of the original charge distribution with the largest eigen value (**v1**).

OPCA for Charged Components For charged components, the monopole and the dipole of the original charge distributions are exactly reproduced if two point charges with total net charge equal to the net charge of the original charge distribution are positioned so that their center of charge coincides with the center of charge of the original distribution (green square in Figure 5.3(b)). The quadurpole moment is optimally reproduced if point charges are placed along the eigen vector with largest corresponding eigenvalue obtained from the quadrupole moment of the original charge distribution[9]. The distance between the two point charges from the center of charge is determined empirically. It was previously shown that fixing the distance of one of the two point charges to $1.5R_0$, which automatically fixes the other one as well, is the best fit to optimally approximate charge distribution of amino acids [9]. Our testing shows the same value of $1.5R_0$ can be applied in the GB-HCPO. The eigenvalues and eigenvectors of the quadrupole moment are computed using the analytical solutions previously introduced by Sigalov et al [179] for calculating the the principal moments of inertia of a mass distribution. An illustration of 2-charge OPCA for a sample charged distribution is given in Figure 5.3(b).

Removing Overlaps For both charged and uncharged components, it is possible that the calculated approximate point charges fall outside the interior region of the corresponding charge distribution. In such case, accidental overlaps between approximate charges and their neighboring charge distributions can introduce instabilities in MD simulations. To avoid such instabilities, we define a smooth function that restricts the calculated approximating point charges within a certain threshold from the center of geometry of components. Consider d to be the distance of the approximate point charges from the center of geometry of the original charge distribution, the corrected distance from the center of geometry (d_c) is obtained from the smooth function below,

$$d_c = \begin{cases} R_2, & d > R_2, \\ \frac{(R_1^2-d^2)^2(R_1^2+2d^2-3R_2^2)(R_2-d)}{(R_1^2-R_2^2)} + d, & R_1 < d < R_2, \\ d, & d < R_1. \end{cases} \tag{5.7}$$

where R_1 is the radius of an inner spherical threshold, and R_2 is radius of an outer spherical threshold. R_2 should be smaller than the size of the charge distribution (R_0) defined as the distance from its center of geometry to the outermost charge, and R_1 should be smaller than R_2. Here we use $R_1 = 0.8R_0$ and $R_2 = R_0$ for approximations at level 1, and $R_1 = 0.8R_0$ and $R_2 = 0.9R_0$ for approximations beyond level 1.

Equivalent of Effective Born Radii (B_{c1} and B_{c2}) for OPCA charges

Our testing shows that the previous expression for the effective Born radii does not lend itself to the new way the approximate point charges are placed in GB-HCPO (results included in section 5.4.3). To fully benefit the advantage of the improved accuracy in the computations of electrostatic interactions in GB-HCPO, here we present an alternative expression for the component effective Born radii, as below

$$\frac{1}{B_c^{k+1}} \approx \left[\frac{1}{\sum\limits_{k \in c} q_k/|r_{kc}|} \sum_{k \in c} \frac{q_k/|r_{kc}|}{B_c^{k\,1/2}} \right]^2 \tag{5.8}$$

where B_c^{k+1} is the effective Born radii for components at level $k+1$, B_c^k is the effective Born radii for components at level k, q_k is the point charges associate with B_c^k, and r_{kc} is the distance between q_k and q_{k+1}, that is the point charge associated with B_c^{k+1}. For $k = 0$, B_c^k and q_k represent atomic Born radii and atomic charges. B_c^{k+1} is a harmonic average of B_c^k weighted by (q_k) and $1/r_{kc}$. Note that effective Born radius of a point charge represents it's degree of burial within the component, and therefore it is best approximated by the charges in close proximity of that point, thus in the new expression (Eq. 5.8) B_c^k

Structure	PDB ID	Size (atoms)	|Charge| (e)	Cutoff dist (Å)	Threshold dist (Å) h_1	h_2	h_3	dr_n (Å) for OPCA
10 bp B-DNA fragment	2BNA	632	18	21	21	n/a	n/a	0.35
Immunoglobulin binding domain	1BDD	726	2	15	15	n/a	n/a	0.25
Ubiquitin	1UBQ	1231	1	15	15	n/a	n/a	0.25
Thioredoxin	2TRX	1654	5	15	15	n/a	n/a	0.35
Nucleosome core particle	1KX5	25101	133	21	21	90	n/a	0.1
Microtubule sheet	*	158016	360	15	15	48	n/a	0.25
Virus capsid	1A6C	475500	120	15	15	66	n/a	0.25
Chromatin fiber	**	1159998	8238	21	21	90	169	0.35

Table 5.1: List of representative structures used for testing. Unless stated otherwise, The cutoff and threshold distances listed here were used for all testing. * The microtubule sheet was constructed as described in Wang and Nogales [207]. ** The chromatin fiber was constructed as described in Wong et. al.[214].

is weighted by $1/r_{kc}$. For the two-charge approximation, two separate component effective radii are computed, one for each of the two approximate charges (B_{c1} and B_{c2}). In this case, the sum in Eq. 5.8 is performed over the positive and negative point charges, and the component effective Born radius obtained from the sum over positive charges is assigned to the larger approximate charge (with sign), and the effective Born radii obtained from the sum over negative charges is assigned to the smaller approximate charge. We found that, when applied to GB-HCPO, the approximation described by Eq. 5.8 is significantly more accurate the previous approximation (Eq.5.6) (results included in Eq. 5.4.3). We have therefore chosen to base all further analysis on Eq. 5.8.

5.3.4 Test Structures

We used a set of eight representative biomolecular structures ranging in size from 632 atoms to 1159998 atoms with absolute total charge ranging from 1 to 8238 e to test the accuracy and speed for GB-HCPO, as compared to cutoff-GB and GB-HCP without OPCA (Table 5.1). The H++ server (https://biophysics.cs.vt.edu/H++) was used to add missing hydrogens to these structures [72].

The HCP threshold distances were chosen such that, for a given atom within a given test structure, the exact atomic computation (level 0) is used for interactions with other atoms within its own and nearest neighboring groups (level 1) as illustrated in Figure 5.2. To satisfy this condition, threshold distances h_l are calculated as $h_l = R_l^{max} + 2 \times R_1^{max}$ where l is the HCP level, R_l^{max} is the maximum component radius at level l, and R_1^{max} is the maximum group (level 1) radius, for a given structure. The HCP threshold distances thus calculated for each of the test structures are shown in Table 5.1. These are the suggested conservative defaults for these and other similar structures. The GB-HCPO level 1 threshold distance

for a given structure is also used as the cutoff distance for the cutoff-GB computations. Unless stated otherwise, these threshold and cutoff distances were used for all of the testing described in the Results section.

5.3.5 Simulation Protocols

The following parameters and protocol were used for the simulations, unless otherwise stated. The threshold distances used are listed in Table 5.1. 12-6 van der Waals interactions for the GB-HCPO were computed using only the atoms that are within the level 1 threshold distance, i.e., atoms that are treated exactly. The simulations used the Amber ff99SB force field.[86]. Langevin dynamics with a collision frequency of 50 ps^{-1} (appropriate for comparison to explicit water results) was used for temperature control, a surface-area dependent energy of 0.005 kcal/mol/Å^2 was added, and an inverse Debye-Hückel length of 0.125 $Å^{-1}$ was used to represent a 0.145 M salt concentration. A 1 fs time step was used for the simulation with the nonbonded neighbor list being updated after every step. Default values were used for all other parameters. The simulation protocol consisted of five stages. First, the starting structure was minimized using the conjugate gradient method with a restraint weight of 5.0 kcal/mol/Å^2. Next, the system was heated to 300 K over 10 ps with a restraint weight of 1.0 kcal/mol/Å^2. The system was then equilibrated for 10 ps at 300 K with a restraint weight of 0.1 kcal/mol/Å^2, and then for another 10 ps with a restraint weight of 0.01 kcal/mol/Å^2. Finally, all restraints were removed for the production stage.

For the case of chromatin fiber, the heating and equilibration stages were reduced from 10 to 4 ps [[[introduce exact protocol used for the chromatin fiber]]] to reduce run time for this simulation, and also the thermal coupling was reduced to 0.01 ps^{-1} to enhance sampling of conformational space.

For explicit solvent simulations, the structure was solvated in a truncated cuboid box extending 10 Å from the solute. The system was neutralized by adding counterions. The TIP3P ion parameters were used to model ion-water interactions. The Amber ff99SB force field was used for explicit-solvent simulations as well .[86].

5.3.6 Accuracy and Speed Evaluation

The accuracy of the approximate methods were evaluated using absolute error in electrostatic energy Err^E and RMS error in electrostatic force (relative force error) Err^F, calculated as

$$Err^E = |E^{approx} - E^{ref}| \tag{5.9}$$

$$Err^{rms} = \left[\sum_{i=1}^{n} |F_i^{approx} - F_i^{ref}|^2/n\right]^{1/2} \tag{5.10}$$

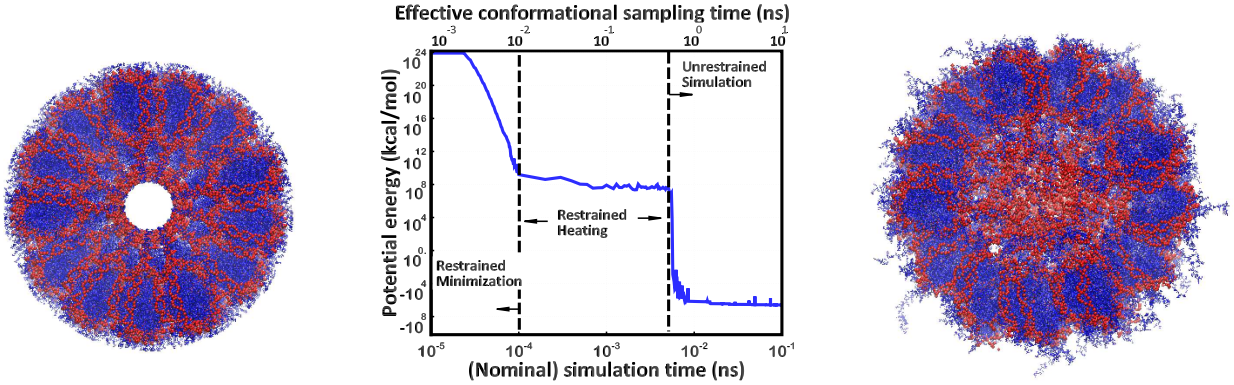
$$\tag{5.11}$$

where E^{approx} is the energy calculated using an approximation, E^{ref} the energy calculated using the reference GB computation without cutoffs or the use of HCP, Err^{rms} the root-mean-square (RMS) error in force for the atoms in a given structure. F_i^{approx} and F_i^{ref} are the force on atom i calculated using the approximate and reference GB computations, respectively. We also compare the backbone RMS deviation from the crystal structure predicted by GB-HCP (without OPCA) and GB-HCPO, with values from TIP3P explicit solvent model.

Speedup was measured as CPU time for the reference (no cutoff) GB computation divided by the CPU time for the approximation tested. Testing was conducted on Virginia Tech's Blueridge computer cluster (http://www.arc.vt.edu) consisting of 408-node Cray CS-300 cluster, each node is outfitted with two octa-core Intel Sandy Bridge CPUs and 64 GB of memory. For the 475 500 atom virus capsid and the 1159998 atom chromatin fiber we used 16 and 64 cores, respectively. To limit the run time for the reference GB computation to a few days, speedup was calculated for 1000 iterations of MD for structures with < 10000 atoms, 100 iterations for structures with $10000 - 1000000$ atoms and 10 iterations for the structure with > 1000000 atoms. To make the results representative of typical simulations involving much larger numbers of iterations, the CPU time excludes the time for loading the data and initialization prior to starting the simulation.

5.4 RESULTS AND DISCUSSION

By drastically reducing the often inordinate computational costs [225], the GB-HCPO opens the possibility of exploring very large systems where the pairwise GB without further approximation is impractical. To demonstrate the advantages of GB-HCPO on a bio-medically relevant example of this type, first we show an application of the model by studying the functional characteristics of the chromatin fiber structure with 1160000 atoms (40 nucleosome). A detailed technical analysis of the accuracy and speed of GB-HCPO will be presented thereafter.



(a) (b) (c)

Figure 5.4: (a) Manually constructed models [215] of $\sim$ 1 million atom chromatin fiber (consistent with low resolution cryo-EM data[170]) can be energetically unrealistic; no atomically detailed experimental structures are available. (b) A 0.1 ns simulation of the fiber using the two-charge GB-HCPO significantly reduces the steric clashes, as seen by the large reduction in the potential energy. Data points represent averages over 100 time step intervals (1 fs each). (c-d) Equilibrated structure (all-atom MD, GB-HCPO) suggests important structural details consistent with experimental results: the linker DNA fills the core region, the H3 histone tails interact with the linker DNA[167, 186].

5.4.1 Insights into the Structure of Chromatin Fiber

It is now well established that details of DNA packing inside the cell nucleus are critical for many cellular processes including cell differentiation and transcription. The primary level of DNA compaction in eukaryotes is the nucleosome – a protein-DNA complex of about 25,000 atoms. These structure can further compactify (Figure 5.1): while exact structural details of this packing are still debated, one widely discussed option is the so-called 30 nm chromatin fiber, which is a well-defined helical arrangement of the nucleosomes. And while the structural features of individual nucleosomes are known to great detail [135] for over 10 years, the exact structure of the 30nm chromatin fiber remained somewhat of a mystery. The fiber represents the second level of DNA compaction in cells[134]; modifications to N-terminal tails of the histone proteins that make up the fiber core are known to regulate DNA accessibility and affect vital process such as gene expression[83]. However, due to its large size, only low-resolution (cryo-EM) experimental structures of the fiber are available[170], its atomistic details including the tails, are unknown.

We use a manually constructed 40-nucleosome (1 million atom) 30-nm chromatin fiber based on a 4-start model consistent with low resolution cryo-EM data[170]. The crystal structure of nucleosomes (1KX5), linker DNA and linker histone (190 bp nucleosome repeat length). are repeatedly combined using a set of coordinate transformations described by Wong et. al.[214], The manually constructed models [215] of $\sim$ 1.16 million atom chromatin fiber is energetically unrealistic (Figure 5.4 (a)). The coordinate transformations result in a number

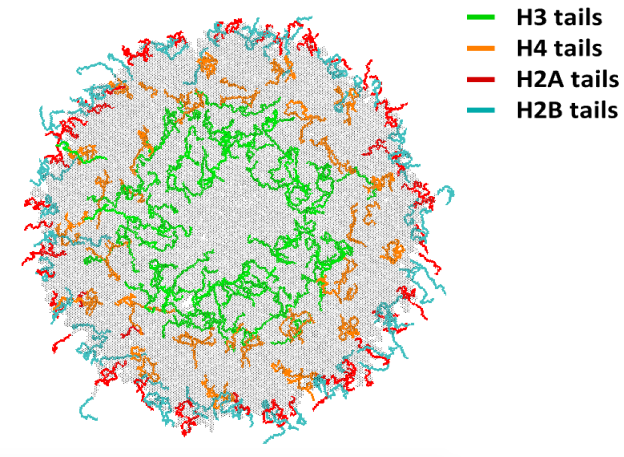


Figure 5.5: The atomistic details of histone tails after 0.1 ns MD simulation of 30-nm chromatin fiber. H3 and H4 N-terminal histone tails are buried within the fiber, and may play a role in chromatin packing via inter-nucleosomal and tail–DNA interactions. H2A and H2B N-terminal histone tails extend out of the fiber surface and may play a role in gene expression by recruiting chromatin binding proteins.

of severe steric clashes. A 0.1 ps muti-scale GB-HCPO simulation provides an energetically realistic model of the fiber (Figure 5.4 (c)). The GB-HCPO significantly reduces the steric clashes, as seen by the large reduction in the potential energy (Figure 5.4 (b)).

Equilibrated structure (Figure 5.4 (c)) suggests important structural details consistent with experimental results. Histone N-terminal tails have been shown to regulate DNA accessibility, gene transcription and chromatin structure These tails are highly positively charged and the above regulation may be partly through the modulation of this charge by post transnational modifications. Figure 5.5 shows that H3 and H4 N-terminal tails are buried within the fiber, and may play a role in chromatin packing via inter-nucleosomal and tail–DNA interactions. H2A and H2B N-terminal tails extend out of the fiber surface (Figure 5.5 and Figure 5.6) and may play a role in gene expression by recruiting chromatin binding proteins. Linker histones attach to the linker DNA and may play a role in chromatin organization. The linker DNA fills the core region (Figure 5.6), the H3 histone tails interact with the linker DNA[167, 186].

5.4.2 Accuracy and Speed

Here we analyze the accuracy and speed for GB-HCPO in force and energy calculations relative to the GB model without further approximation. For the reference GB computations we used the commonly used GB-OBC model (IGB=5 in Amber[154]). The GB-HCPO's accuracy and speed is compared to those of the commonly used cutoff-GB and the its predecessor GB-HCP without OPCA. The cutoff-GB method ignores all interactions beyond a

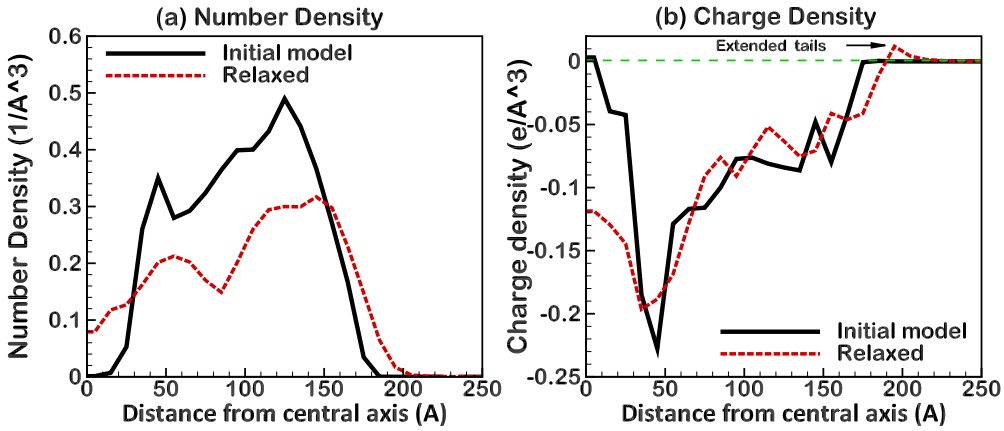


Figure 5.6: Density (number of atoms per unit volume) and charge density along the radius of the equilibrated chromatin fiber: (a) Linker DNA fills the core and H2A and H2B N-terminal histone tails extend out of the fiber surface, leading to a wider distribution of atoms along the radial axis, compared to the initial structure. (b) After equilibration, the core becomes negatively charged due to the presence of linker DNA, and the outer region is positively charged due to the presence of histone tails.

cutoff distance for the computation of electrostatic energy and effective Born radii in Eqs. 5.2,5.3 and 5.4. We also evaluate the performance of GB-HCPO in predicting the structure of biomolecules in dynamics relative to explicit-solvent simulations as reference.

Accuracy

Figures 5.7 (a) and (b) show the accuracy for GB-HCPO methods compared to the GB-HCP (2-charge) and cutoff-GB methods. A logarithmic scale is used to better demonstrate distribution of errors in both energy and force over the broad range of test structure sizes. For the test structures considered here, the deviation of electrostatic energy and forces calculated by the GB-HCPO method from the values from GB without approximation is significantly smaller than those by the GB-HCP and cutoff-GB methods (Figure 5.7 (a) and (b)). The smaller deviation by GB-HCPO compared to GB-HCP and cutoff-GB methods is uniform across the broad range of test structure sizes. In particular, the absolute error in electrostatic energy computed using GB-HCPO can be up to two orders of magnitude smaller than that from cutoff-GB. On average, the absolute error in the electrostatic energies relative to the GB computation without any approximation is 74% and 97% smaller than those from previously developed GB-HCP and commonly used cutoff-GB method (Figure 5.7 (a)), respectively. The rms error in electrostatic force calculated by GB-HCPO can be up to one order of magnitude smaller that that by both GB-HCP and cufoff-GB. On average, the rms error in the electrostatic forces relative to the GB computation without any approximation is 44% and 80% smaller than those from GB-HCP and cutoff-GB method(Figure 5.7 (b)),

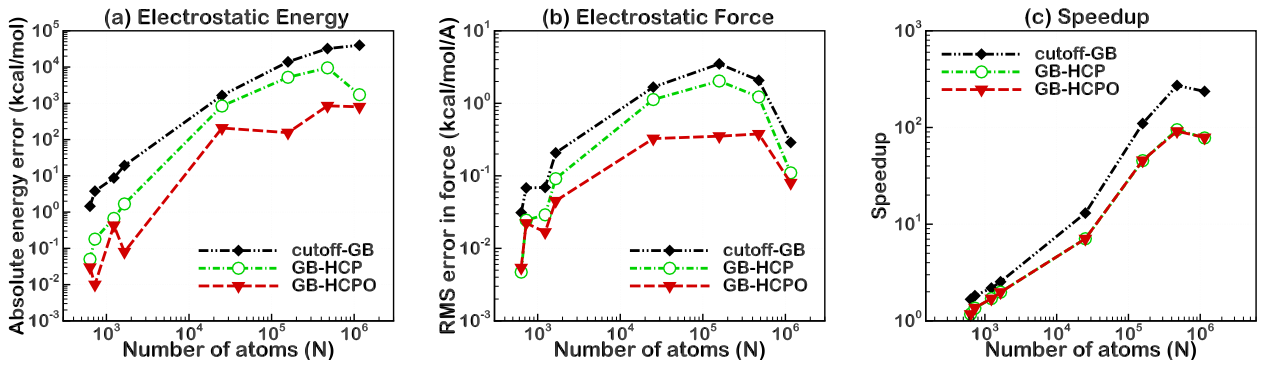


Figure 5.7: Accuracy of the cutoff-GB, GB-HCP, GB-HCPO methods relative to the reference GB computation without cutoffs. Accuracy is computed as (a) absolute error in electrostatic energy, and (b) RMS error in electrostatic force. Connecting lines are shown to guide the eye. (c) Speedup for the cutoff-GB, GB-HCP and GB-HCPO methods relative to the reference GB computation without cutoffs. Threshold and cutoff distances used for the different structures are listed in Table 5.1. Connecting lines are shown to guide the eye.

respectively.

5.4.3 Accuracy evaluation of component effective Born radii

As described in the Methods section, we examined two different approaches to compute the component effective Born radii: the approximation defined by Eq. 5.6 (old Born), previously presented in Ref. [10], the new approximation (new Born) presented in this work (Eq. 5.8). Both approximations were applied to GB-HCP and GB-HCPO. As shown in Figure 5.8, the new Born approximation (Eq. 5.8) is more accurate than the old Born approximation (Eq. 5.6), when applied to each of the GB-HCP and GB-HCPO methods. One can also see that the old Born approximation does not properly lend itself to the GB-HCPO model.

Speedup

Figure 5.7(c) shows that both GB-HCPO and GB-HCP can be up to two orders of magnitude faster than the GB computation without any approximation, depending on structure size. The speedup for the GB-HCP (2-charge) and GB-HCPO methods are almost the same indicating that the higher accuracy of GB-HCPO compared to GB-HCP is achieved without any impact on its speed. The GB-HCPO and GB-HCP approximations are slower than cutoff-GB: the average speedup for the 8 structures tested here was $\sim 78\times$ for the cutoff-GB, while it is $\sim 30\times$ for the GB-HCP and GB-HCPO methods. The higher speed of cutoff-GB is expected as cutoff-GB totally ignores the electrostatic interaction beyond a certain threshold, in contrast to HCP approximations that take into account distant electrostatic interactions.

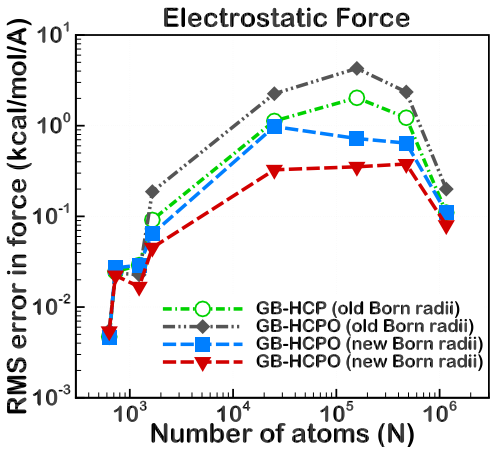


Figure 5.8: Comparison of two alternative methods for computing component effective Born radii showing RMS error in electrostatic force GB-HCP and GB-HCPO. Connecting lines are shown to guide the eye.

Generally the speed up is higher when the number of atoms increases. However, for the largest structure considered here; the 1159998 atom chromatin fiber; the speed up is smaller than that for 475500 atom virus capsid (1a6c) for all the models because the threshold and cutoff distances used to simulate the chromatin fiber are much larger than those used for virus capsid (see 5.1).

Stability in MD simulations

The results presented in the previous section showed significant improvement of the GB-HCPO model, compared to cutoff-GB and GB-HCP, in calculating electrostatic forces and energies when using reference GB as a benchmark. It is of interest to determine if these improvements in electrostatic forces and energies translate to improved agreement of structural ensembles compared to those obtained reference GB without approximation and explicit solvent simulation, which tend to be computationally demanding. To test the performance of GB-HCPO in dynamics, we ran 50 ns MD simulations of the immunoglobulin binding domain (1BDD), thioredoxin (2TRX), and ubiquitin (1UBQ). One trajectory was produced for reference GB and explicit solvent simulations, and two independent trajectories were produced for GB-HCPO, GB-HCP and cutoff-GB. In our previous study[10] it was shown that, for these three structures, in general GB-HCP reproduces the dynamics of the reference GB simulation more accurately than the cutoff-GB method. Therefore, here we focus on comparing the performance of GB-HCPO and GB-HCP in reproducing the dynamics of reference GB and explicit solvent, and only for one structure (1BDD) we also compare the results with the ones from cutoff-GB.

Figure 5.9 shows the backbone RMS deviation from the crystal structure for the simulations.

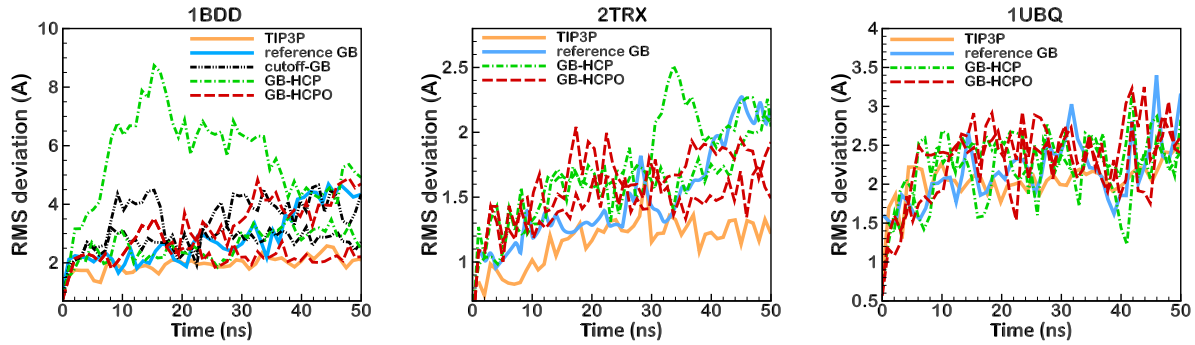


Figure 5.9: RMS deviation from the starting structure for 50 ns MD simulations of immunoglobulin binding domain (1BDD), thioredoxin (2TRX) and Ubiquitin (1UBQ) using the reference explicit-solvent simulations (TIP3P), reference GB (without approximation), GB-HCP and GB-HCPO methods. The RMS deviation from cut-off GB is also shown for immunoglobulin binding domain. RMS deviation is calculated for backbone heavy atoms. The trajectory is sampled every 1 ns. Connecting lines are shown to guide the eye.

The results shown in Figure 5.9 suggest that, the trajectories for the GB-HCPO method are generally in good agreement with the reference GB and explicit solvent simulation. For 1BDD the GB-HCP trajectories show RMS deviations that are substantially larger than the GB-HCPO or the reference explicit-solvent trajectories. This example emphasizes how subtle errors in charge-charge interactions, as reflected in electrostatic forces and energies (Figure 5.7) can result in qualitatively different conformational dynamics. On a practical level, small inaccuracies in the energy or force calculations in the simulations of small flexible structures such as 1BDD can lead to large structural deviations over of the course of the trajectory. Similarly, for 2TRX the trajectories obtain from GB-HCPO are in better agreement with the reference GB and explicit solvent simulation that those from GB-HCP. Due to the high stability of the structure of 1UBQ, the trajectories for 1UBQ from all of the simulations (GB-HCPO, GB-HCP, reference GB and explicit solvent simulation) are in close agreement.

5.4.4 The computational speed of GB-HCPO relative to explicit solvent simulations

The advantage of GB-HCPO becomes critical for very large structures. For example, on the basis of the run times for a GB-HCPO simulation of the 158016 atom microtubule structure (tub46), compared to an equivalent PME simulation in TIP3P explicit solvent using sander module of Amber14, the nominal speeds (nanoseconds per day) of GB-HCPO is about 20% times slower than the PME simulations in TIP3P explicit solvent, on Virginia Tech's HokieSpeed computer cluster using 8 cores. Note that unlike the production PME module of Amber14, NAB is not highly optimized. Even for this case where explicit-solvent

(TIP3P) PME is slightly faster than GB-HCPO, when combined with the speedup due to conformational sampling, the overall speed of conformational sampling for GB can be faster. For instance, it has been shown that the speed of conformational sampling in implicit-solvent simulations can be $\sim$ 10 to 100 times faster than common explicit-solvent PME simulations [11].

5.5 CONCLUSION

Implicit solvent models are currently extensively employed in atomic-level modeling and simulations to speed up the research in drug design and medical science. However, traditional GB models scale as $\sim n^2$, where n is the number of solute atoms, limiting the advantage of these models to small and medium size structures. The GB-HCPO multi-scale approximation presented in this work significantly speeds up the GB computations as it scales as $\sim nlogn$. A comprehensive testing on a set of representative biomolecules of varied sizes performed here shows that GB-HCPO is substantially more accurate than common cutoff-GB approach and its predecessor GB-HCP in predicting electrostatic forces and electrostatic energies. The structural stability of test biomolecules – immunoglobulin binding domain and thioredoxin – for the GB-HCPO simulation are in reasonable agreement with the results from explicit-solvent simulations. The better accuracy of GB-HCPO compared to its predecessor GB-HCP was achieved without sacrificing the speed.

GB-HCPO is designed for simulating very large systems where the computational costs of explicit solvent simulations becomes inordinate. As an practical demonstration, we used GB-HCPO to equilibrate the structure of a chromatin fiber (40 nucleosome) with over a million atom. We show that 100 *ps* simulation of the fiber on a regular supercomputer using 192 processors can suggest important structural details consistent with experimental results. The GB-HCPO simulations successfully resolved numerous severe steric clashes, significantly improving the quality of the starting structure. The total computational time for the 100 ps all-atom simulation performed here on a regular compute cluster using 192 cores was about 1 month, whereas an equivalent simulation using the regular GB without additional approximation on the same compute cluster would take $\sim$ 100 longer.

Solvating the gigantic structure of chromatin fiber in a box of explicit solvent model leads to a very large system and will be computationally very expensive to simulate. Moreover, the 100 ps time length will not be sufficient to achieve structural details equivalent to the results achieved with the GB-HCPO simulation, because the speed of conformational change is much lower ($\sim$ 100 times) in explicit solvent compared to implicit solvent. The effective speedup obtained by GB-HCPO, which takes into account both conformational search speed and computational nominal speed, is considerably higher when comparing to explicit solvent simulations.

The GB-HCPO implementation in NAB is scheduled to be released with AmberTools16, for

general use.

Chapter 6

Conclusion

Accurate yet efficient computation of electrostatic interactions is paramount in atomistic simulations. In practical molecular simulation, the electrostatic properties of the biomolecules are represented via partial charges distributed throughout the molecules. In this work, we have introduced an alternative to the atom-center charge placement – the optimal point charge approximation (OPCA). An OPCA consists of a given number of point charges which are optimally placed to best reproduce the electrostatic potential due to the original charge distribution, regardless of the distance to the charge distribution. By construction, the proposed optimal point charge approximation (OPCA) retains many of the useful properties of point multipole expansion, including the same far-field asymptotic behavior of the approximate potential. We have provided a general framework for calculating OPCAs to any order. We have also derived closed-form analytical expressions for the 1-charge, 2-charge and 3-charge OPCA. We note that higher order closed-form, analytical OPCAs may be challenging to derive, but for some applications, lower order OPCAs may be sufficient.

We showed that the 3-charge OPCA can significantly more accurately represent the electrostatic potential for the water molecule compared to the conventional atom-centered charge placement. Given the significant accuracy of 3-charge OPCA in representing the electrostatics of the water molecule, we applied the concept of OPCA to develop a different, novel approach of constructing accurate and simple point charge water models. In contrast to the conventional approach that searches for model parameters in the space of charges-distances-angles, the proposed water modeling approach allows a virtually exhaustive search for optimal parameters of fixed-charge water models based on n-point topologies, even for large n, in the sub-space most relevant to electrostatic properties of the water molecule (i. e. the low order multipole moments) in liquid phase. The "optimal" 3-charge, 4-point rigid water model (OPC) constructed based on the new approach was shown to reproduce a comprehensive set of bulk properties significantly more accurately than commonly used rigid models: average error relative to experiment is 0.76%. Close agreement with experiment holds over a wide range of temperatures.

As preliminary and published results demonstrate, the improvements in the proposed OPC model extend beyond bulk properties: compared to common rigid models predicted hydration free energies of small molecules using OPC are uniformly closer to experiment, root-mean-square error < 1 kcal/mol. Improvements were also reported in RNA simulations, representing the structure of intrinsically disordered proteins, and protein-ligand binding calculations. Critically, in all of the above examples OPC improved agreement with experiment for all underlying gas-phase force-fields tested.

We also utilized the new water modeling approach to develop a 3-point version of the optimal point charge water model, called OPC3. It was shown that OPC3 is significantly more accurate than commonly used 3-point models such as TIP3P and SPCE. A comparison of the OPC3 model with two recent models developed based on completely different parametrization procedures indicates a consensus for the optimal parametrization of 3-point water models. Given that very different parameter optimizations, including a virtually exhaustive search in the "appropriate" electrostatic parameter space, yield essentially the same result, we conclude that the search for an optimal 3-point rigid water model is over.

OPCA has utility in coarse-grained and multi-scale methods especially in dynamics where analytic expressions and the simplicity of the algorithms is key. The benefit of OPCA comes from its capability to represent large charge distributions with only a few point charges. We used this capability of OPCA to develop a multi-scale, yet fully atomistic, generalized Born (GB) approach (GB-HCPO) that runs MD simulations orders of magnitude faster than is possible with traditional GB. As a practical demonstration, we exploited the new multi-scale GB model to perform a million-atom simulation of 30-nm chromatin fiber, starting from a manually constructed 4-star model consistent with low resolution cryo-EM data. Our study suggests important structural details consistent with experiment: the linker DNA fills the core region and the H3 histone tails interact with the linker DNA.

The optimal point charge approximation presented here is a new concept; thus its many applications and potentially useful properties remain unexplored in this work. The approximations we have introduced provide a systematic way of deriving approximate charge distributions that have the potential to be both computationally effective and produce an accurate representation of the original electrostatic potential. To further improve the representation of the original potential via OPCAs, future work may consider partitioning the original charge distribution into several domains, and finding OPCA for each of them separately, similar to the distributed multipoles approach. Further exploration of the mathematical and physical properties of OPCAs is also desirable.

Bibliography

[1] *Guideline on the Use of Fundamental Physical Constants and Basic Constants of Water*. The International Association for the Properties of Water and Steam, Gaithersburg, Maryland, 2001.

[2] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys*, 123(23):234505+, 2005.

[3] J. L. F. Abascal and C. Vega. The water forcefield: importance of dipolar and quadrupolar interactions. *J Phys Chem C*, 111(43):15811–15822, 2007.

[4] S. Adcock and J. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106(5):1589–1615, 2006.

[5] B. Aguilar and A. V. Onufriev. Efficient computation of the total solvation energy of small molecules via the R6 generalized Born model. *Journal of Chemical Theory and Computation*, 8(7):2404–2411, 2012. PMCID: PMC4003403.

[6] O. Akin-Ojo and F. Wang. The quest for the best nonpolarizable water model from the adaptive force matching method. *Journal of Computational Chemistry*, 32(3):453–462, 2011.

[7] J. Alejandre, G. A. Chapela, H. Saint-Martin, and N. Mendoza. A non-polarizable model of water that yields the dielectric constant and the density anomalies of the liquid: TIP4Q. *Phys Chem Chem Phys*, 13:19728–19740, 2011.

[8] R. Amaro. private communication, 2008.

[9] R. Anandakrishnan, C. Baker, S. Izadi, and A. V. Onufriev. Point charges optimally placed to represent the multipole expansion of charge distributions. *PloS one*, 8(7):e67715, 2013. PMCID: PMC3701554.

[10] R. Anandakrishnan, M. Daga, and A. V. Onufriev. An n log n generalized born approximation. *Journal of Chemical Theory and Computation*, 7(3):544–559, 2011.

[11] R. Anandakrishnan, A. Drozdetski, R. C. Walker, and A. V. Onufriev. Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations. *Biophysical journal*, 108(5):1153–1164, 2015. PMCID in progress.

[12] R. Anandakrishnan and A. V. Onufriev. An N log N approximation based on the natural organization of biomolecules for speeding up the computation of long range interactions. *Journal of Computational Chemistry*, 31(4):691–706, Mar. 2010.

[13] R. Anandakrishnan and A. V. Onufriev. An N log N approximation based on the natural organization of biomolecules for speeding up the computation of long range interactions. *J. Comp. Chem.*, 31(4):691–706, 2010.

[14] G. Archontis and T. Simonson. A residue-pairwise generalized born scheme suitable for protein design calculations. *J. Phys. Chem. B*, 109(47):22667–22673, December 2005.

[15] C. Avendaño, T. Lafitte, C. S. Adjiman, A. Galindo, E. A. Müller, and G. Jackson. Saft- force field for the simulation of molecular fluids: 2. coarse-grained models of greenhouse gases, refrigerants, and long alkanes. *The Journal of Physical Chemistry B*, 117(9):2717–2733, 2013. PMID: 23311931.

[16] C. M. Baker, V. M. Anisimov, and A. D. MacKerell. Development of CHARMM Polarizable Force Field for Nucleic Acid Bases Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B*, 115(3):580–596, Dec. 2010.

[17] N. Baker, D. Bashford, and D. Case. Implicit solvent electrostatics in biomolecular simulation. In *New Algorithms for Macromolecular Simulation*, volume 49 of *Lecture Notes in Computational Science and Engineering*, pages 263–295. Springer, 2006.

[18] P. Ball. *Life's Matrix: A Biography of Water*. Farrar, Straus, and Giroux, New York, 1999.

[19] P. Barnes, J. L. Finney, J. D. Nicholas, and J. E. Quinn. Cooperative effects in simulated water. *Nature*, 282(5738):459–464, Nov. 1979.

[20] N. Basdevant, D. Borgis, and T. Ha-Duong. A coarse-grained protein-protein potential derived from an all-atom force field. *Journal of Physical Chemistry. B*, 111(31):9390–9, Aug. 2007.

[21] D. Bashford and D. A. Case. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.

[22] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints For Determining Atom-Centered Charges: The RESP Model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[23] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J Phys Chem*, 91(24):6269–6271, 1987.

[24] C. Bergonzo and Thomas. Improved force field parameters lead to a better description of RNA structure. *J. Chem. Theory Comput.*, 11(9):3969–3972, Sept. 2015.

[25] J. D. Bernal and R. H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *The Journal of Chemical Physics*, 1:515–548, aug 1933.

[26] P. Beroza and D. A. Case. Calculation of Proton Binding Thermodynamics in Proteins. *Methods Enzymol.*, 295:170–189, 1998.

[27] T. C. Bishop, R. D. Skeel, and K. Schulten. Difficulties with multiple time stepping and fast multipole algorithm in molecular dynamics. *Journal of Computational Chemistry*, 18(14):1785–1791, 1997.

[28] D. Bratko, L. Blum, and A. Luzar. A simple model for the intermolecular potential of water. *The Journal of Chemical Physics*, 83(12):6367–6370, Dec. 1985.

[29] C. N. Breneman and K. B. Wiberg. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. Need for High Sampling Density in Formamide Conformational Analysis. *J. Comp. Chem.*, 11(3):361–373, 1990.

[30] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[31] W. Cai, S. Deng, and D. Jacobs. Extending the fast multipole method to charges inside or outside a dielectric sphere. *Journal of Computational Physics*, 223(2):846–864, May 2007.

[32] W. Cai, S. Deng, and D. Jacobs. Extending the fast multipole method to charges inside or outside a dielectric sphere. *J. Chem. Phys.*, 223(2):846–864, May 2007.

[33] N. Calimet, M. Schaefer, and T. Simonson. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins*, 45(2):144–158, 2001.

[34] J. Carrier, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm for particle simulations. *SIAM J. Sci. Stat. Comp.*, 9(4):669–686, 1988.

[35] D. Case, J. Berryman, R. Betz, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, T. Kovalenko, A. amd Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. Merz, G. Monard, P. Needham, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, R. Salomon-Ferrer, C. Simmerling, W. Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, D. York, and P. Kollman. *AMBER 2015*. University of California, San Francisco, 2015.

[36] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

[37] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–1688, Dec 2005.

[38] J. Chen, C. L. Brooks, and J. Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Current Opinion in Structural Biology*, 18(2):140–148, Apr. 2008.

[39] J. Chen, W. Im, and C. L. Brooks. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *J Am Chem Soc*, 128(11):3728–3736, March 2006.

[40] H. Cheng, L. Greengard, and V. Rokhlin. A Fast Adaptive Multipole Algorithm in Three Dimensions. *Journal of Computational Physics*, 155(2):468–498, Nov. 1999.

[41] L. E. Chirlian and M. M. Francl. Atomic charges derived from electrostatic potentials: A detailed study. *Journal of Computational Chemistry*, 8(6):894–905, 1987.

[42] S. A. Clough, Y. Beers, G. P. Klein, and L. S. Rothman. Dipole moment of water from Stark measurements of H_2O, HDO, and D_2O. *The Journal of Chemical Physics*, 59(5):2254–2259, 1973.

[43] F. Colonna, E. Evleth, and J. G. Ángyán. Critical analysis of electric field modeling: Formamide. *J. Comput. Chem.*, 13(10):1234–1245, Dec. 1992.

[44] K. Coutinho, R. Guedes, B. C. Cabral, and S. Canuto. Electronic polarization of liquid water: converged monte carlo-quantum mechanics results for the multipole moments. *Chem Phys Lett*, 369(34):345 – 353, 2003.

[45] C. Cramer and D. Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99(8):2161–2200, 1999.

[46] C. J. Cramer and D. G. Truhlar. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.*, 99:2161–2200, 1999.

[47] F. A. Cruz and L. A. Barba. Characterization of the accuracy of the fast multipole method in particle simulations. *International Journal for Numerical Methods in Engineering*, 79(13):1577–1604, Sept. 2009.

[48] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[49] L. David, R. Luo, and M. K. Gilson. Comparison of generalized Born and Poisson models: Energetics and dynamics of HIV protease. *J. Comp. Chem.*, 21(4):295–309, 2000.

[50] K. A. Dill, T. M. Truskett, V. Vlachy, and B. Hribar-Lee. Modeling water, the hydrophobic effect, and ion solvation. *Annual Review of Biophysics and Biomolecular Structure*, 34:173–199, June 2005.

[51] G. G. Dodson, D. P. Lane, and C. S. Verma. Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO reports*, 9:144–150, 2008.

[52] B. N. Dominy and C. L. Brooks. Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids. *J. Phys. Chem. B*, 103:3765–3773, 1999.

[53] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *Journal of Chemical Physics*, 90:1007, 1989.

[54] N. J. English *. Molecular dynamics simulations of liquid water using various long-range electrostatics techniques. *Molecular Physics*, 103(14):1945–1960, 2005.

[55] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[56] M. Feig. Kinetics from implicit solvent simulations of biomolecules as a function of viscosity. *J. Chem. Theory Comput.*, 3(5):1734–1748, September 2007.

[57] M. Feig and C. L. Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14(2):217–224, April 2004.

[58] M. Feig, W. Im, and C. L. Brooks. Implicit solvation based on generalized Born theory in different dielectric environments. *J. Chem. Phys.*, 120(2):903–911, January 2004.

[59] A. K. Felts, E. Gallicchio, D. Chekmarev, K. A. Paris, R. A. Friesner, and R. M. Levy. Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. *J. Chem. Theory Comput.*, 4(5):855–868, May 2008.

[60] C. J. Fennell, L. Li, and K. A. Dill. Simple liquid models with corrected dielectric constants. *J Phys Chem B*, 116(23):6936–6944, 2012.

[61] J. L. Finney. The water molecule and its interactions: the interaction between theory, modelling, and experiment. *J Mol Liq*, 90(13):303 – 312, 2001.

[62] J. L. Finney. Water? what's so special about it? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1448):1145–1165, 2004.

[63] F. Franks. *Water: a matrix of life*. Cambridge: Royal Society of Chemistry, 2000.

[64] P. Freddolino, A. Arkhipov, S. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437–449, 2006.

[65] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, March 2006.

[66] R. Fuentes-Azcatl and J. Alejandre. Non-polarizable force field of water based on the dielectric constant: TIP4P/ϵ. *J Phys Chem B*, 118(5):1263–1272, 2014.

[67] E. Gallicchio and R. M. Levy. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.*, 25(4):479–499, 2004.

[68] K. Gao, J. Yin, N. M. Henriksen, A. T. Fenley, and M. K. Gilson. Binding enthalpy calculations for a neutral HostGuest pair yield widely divergent salt effects across water models. *J. Chem. Theory Comput.*, Sept. 2015.

[69] A. Ghosh, C. S. Rapp, and R. A. Friesner. Generalized born model based on a surface integral formulation. *J. Phys. Chem. B*, 102:10983–10990, 1998.

[70] M. K. Gilson. Theory of Electrostatic Interactions in Macromolecules. *Curr. Opin. Struct. Biol.*, 5:216–223, 1995.

[71] M. K. Gilson and H. X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[72] J. Gordon, J. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev. H++: a server for estimating pKa's and adding missing hydrogens to macromolecules. *Nucleic Acids Res.*, 33:68–71, 2005.

[73] J. C. Gordon, A. T. Fenley, and A. Onufriev. An analytical approach to computing biomolecular electrostatic potential. II. Validation and applications. *Journal of Chemical Physics*, 129(7):075102, Aug. 2008.

[74] A. Gramada and P. Bourne. Resolving a distribution of charge into intrinsic multipole moments: A rankwise distributed multipole analysis. *Physical Review E*, 78(6):1–7, Dec. 2008.

[75] J. A. Grant, B. T. Pickup, M. J. Sykes, C. A. Kitchen, and A. Nicholls. The Gaussian generalized Born model: application to small molecules. *Phys. Chem. Chem. Phys.*, 9(35):4913–4922, 2007.

[76] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, Dec. 1987.

[77] L. Greengard and V. Rokhlin. A new version of the fast multipole method for the Laplace equation in three dimensions. *Acta Numerica*, 6(1):229–269, 1997.

[78] J. K. Gregory, D. C. Clary, K. Liu, M. G. Brown, and R. J. Saykally. The water dipole moment in water clusters. *Science*, 275(5301):814–817, 1997.

[79] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids In Molecular Liquids*, 101(1-3):219–260, 2002.

[80] U. Haberthür and A. Caflisch. Facts: Fast analytical continuum treatment of solvation. *J. Comp. Chem.*, 29:701–715, October 2007.

[81] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett*, 246:122–129, 1995.

[82] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, 100:19824–19836, 1996.

[83] S. Henikoff. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet*, 9(1):15–26, January 2008.

[84] B. Honig and A. Nicholls. Classical Electrostatics in Biology and Chemistry. *Science*, 268:1144–1149, 1995.

[85] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.*, 120(20):9665–9678, 2004.

[86] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, 2006.

[87] M. Hülsmann, J. Vrabec, A. Maaß, and D. Reith. Assessment of numerical optimization algorithms for the development of molecular models. *Computer Physics Communications*, 181(5):887 – 905, 2010.

[88] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[89] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, 14:33–38, 1996.

[90] T. Ichiye and M. L. Tan. Soft sticky dipole-quadrupole-octupole potential energy function for liquid water: An approximate moment expansion. *The Journal of Chemical Physics*, 124(13):134504+, 2006.

[91] T. Ichiye and M. L. Tan. Soft sticky dipole-quadrupole-octupole potential energy function for liquid water: An approximate moment expansion. *The Journal of Chemical Physics*, 124(13):134504+, 2006.

[92] W. Im, M. S. Lee, and C. L. Brooks. Generalized born model with a simple smoothing function. *J. Comp. Chem.*, 24(14):1691–1702, November 2003.

[93] S. Izadi, B. Aguilar, and A. V. Onufriev. Protein–Ligand electrostatic binding free energies from explicit and implicit solvation. *J. Chem. Theory Comput.*, 11(9):4450–4459, Sept. 2015.

[94] S. Izadi, R. Anandakrishnan, and A. V. Onufriev. Building water models: A different approach. *J. Phys. Chem. Lett.*, 5(21):3863–3871, Oct. 2014.

[95] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *Journal of Physical Chemistry. B*, 109(7):2469–73, Feb. 2005.

[96] J. Jackson. *Classical Electrodynamics Third Edition*. J. Wiley & Sons, New York, 1999.

[97] A. Jagielska and H. A. Scheraga. Influence of temperature, friction, and random forces on folding of the b-domain of staphylococcal protein a: All-atom molecular dynamics in implicit solvent. *Journal of Computational Chemistry*, 28(6):1068–1082, 2007.

[98] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *J Comp Chem*, 21(2):132–146, 2000.

[99] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: Ii. parameterization and validation. *J Comp Chem*, 23(16):1623–1641, 2002.

[100] S. Jang, E. Kim, and Y. Pak. All-atom level direct folding simulation of a $\beta\beta\alpha$ miniprotein. *The Journal of Chemical Physics*, 128(10):105102, 2008.

[101] S. Jang, E. Kim, S. Shin, and Y. Pak. Ab initio folding of helix bundle proteins using molecular dynamics simulations. *J. Am. Chem. Soc.*, 125(48):14841–14846, December 2003.

[102] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impley, and M. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, 79:926–935, 1983.

[103] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, March 2004.

[104] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79(2):926–935, 1983.

[105] W. L. Jorgensen and C. Jenson. Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. *J Comp Chem*, 19(10):1179–1186, 1998.

[106] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci USA*, 102(19):6665–6670, 2005.

[107] I. S. Joung and T. E. Cheatham. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112(30):9020–9041, July 2008.

[108] S. Kale and J. Herzfeld. Natural polarizability and flexibility via explicit valency: The case of water. *The Journal of Chemical Physics*, 136(8):084109+, 2012.

[109] S. M. H. Karimian and S. Izadi. Bin size determination for the measurement of mean flow velocity in molecular dynamics simulations. *International Journal for Numerical Methods in Fluids*, 71(7):930–938, 2013.

[110] S. M. H. Karimian, S. Izadi, and A. B. Farimani. A study on the measurement of mean velocity and its convergence in molecular dynamics simulations. *International Journal for Numerical Methods in Fluids*, 67(12):2130–2140, 2011.

[111] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S.Ȧ.*, 102(19):6679–6685, May 2005.

[112] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.

[113] M. Karplus and J. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural and Molecular Biology*, 9:646–652, 2002.

[114] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–652, September 2002.

[115] G. S. Kell. Density, thermal expansivity, and compressibility of liquid water from 0.deg. to 150.deg.. Correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *Journal of Chemical & Engineering Data*, 20(1):97–105, Jan. 1975.

[116] R. A. Kendall, T. H. Dunning, and R. J. Harrison. Electron-affinities of the 1st-row atoms revisited – systematic basis-sets and wave-functions. *Journal of Chemical Physics*, 96:6796, 1992.

[117] P. T. Kiss and A. Baranyai. A systematic development of a polarizable potential of water. *J Chem Phys*, 138(20):204507+, 2013.

[118] P. Koehl. Electrostatics calculations: Latest methodological advances. *Curr. Opin. Struct. Biol.*, 16(6):142–151, March 2006.

[119] P. Koumoutsakos. Fast multipole methods for three dimensional N-body problems. Technical Report 96N25344, NASA, Ames Research Center, 1995.

[120] C. Kramer, A. Spinn, and K. R. Liedl. Charge anisotropy: Where atomic multipoles matter most. *J. Chem. Theory Comput.*, Sept. 2014.

[121] C. G. Lambert, T. A. Darden, and J. A. Board Jr. A multipole-based algorithm for efficient calculation of forces and potentials in macroscopic periodic assemblies of particles. *J. Chem. Phys.*, 126(2):274–285, July 1996.

[122] G. Lamoureux, E. Harder, I. V. Vorobyov, B. Roux, and A. D. MacKerell. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters*, 418(1-3):245–249, Jan. 2006.

[123] G. Lamoureux, A. D. MacKerell, and B. Roux. A simple polarizable model of water based on classical drude oscillators. *The Journal of Chemical Physics*, 119(10):5185–5197, Sept. 2003.

[124] A. R. Leach. *Molecular Modelling: Principles and Applications.* Addison Wesley Longman, Essex UK, 1996.

[125] F. Lee and A. Warshel. A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations. *Journal of Chemical Physics*, 97:3100–3107, 1992.

[126] M. S. Lee, J. F. R. Salsbury, and C. L. Brooks, III. Novel generalized Born methods. *J. Chem. Phys.*, 116:10606–10614, 2002.

[127] H. Lei and Y. Duan. Ab initio folding of albumin binding domain from all-atom molecular dynamics simulation. *J. Phys. Chem. B*, 111(19):5458–5463, May 2007.

[128] I. Leontyev and A. Stuchebrukhov. Accounting for electronic polarization in non-polarizable force fields. *Phys. Chem. Chem. Phys.*, 13(7):2613–2626, 2011.

[129] L. Li, C. Li, Z. Zhang, and E. Alexov. On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *Journal of chemical theory and computation*, 9(4):2126–2136, Apr. 2013.

[130] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520, Oct. 2011.

[131] Y. Liu and T. Ichiye. Soft sticky dipole potential for liquid water: a new model. *J. Phys. Chem.*, 100(7):2723–2730, Jan. 1996.

[132] R. J. Loncharich and B. R. Brooks. The effects of truncating long-range forces on protein dynamics. *Proteins*, 6:32–45, 1989.

[133] A. Lopes, A. Alexandrov, C. Bathelt, G. Archontis, and T. Simonson. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins*, 67(4):853–867, June 2007.

[134] K. Luger, M. L. Dechassa, and D. J. Tremethick. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat Rev Mol Cell Biol*, 13(7):436–447, July 2012.

[135] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, Sep 1997.

[136] R. Luo, L. David, and M. K. Gilson. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comp. Chem.*, 23:1244–1253, 2002.

[137] J. D. Madura, M. E. Davis, M. K. Gilson, R. C. Wade, B. A. Luty, and J. A. McCammon. Biological Applications of Electrostatic Calculations and Brownian Dynamics. *Rev. Comp. Chem.*, 5:229–267, 1994.

[138] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys*, 112(20):8910–8922, 2000.

[139] Y. Marechal. *The Hydrogen Bond and the Water Molecule: The Physics and Chemistry of Water, Aqueous and Bio Media*. Elsevier, Oxford, 2007.

[140] P. Mark and L. Nilsson. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A*, 105(43):9954–9960, Oct. 2001.

[141] P. Mark and L. Nilsson. Structure and dynamics of liquid water with different long-range interaction truncation and temperature control methods in molecular dynamics simulations. *J. Comp. Chem.*, 23(13):1211–1219, 2002.

[142] R. Mecke and W. Baumann. Das rotationschwingungsspektrum des wasserdampfes. *Phys. Zeits.*, 33:883, 1932.

[143] G. R. Medders, V. Babin, and F. Paesani. Development of a First-Principles Water Potential with Flexible Monomers. III. Liquid Phase Properties. *J. Chem. Theory Comput.*, 10(8):2906–2910, July 2014.

[144] D. L. Mobley, A. E. Barber, C. J. Fennell, and K. A. Dill. Charge asymmetries in hydration of polar solutes. *J Phys Chem B*, 112(8):2405–2414, 2008.

[145] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J Chem Theor Comp*, 5(2):350–358, 2009.

[146] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, and A. Onufriev. Generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theor. Comp.*, 3(1):156–169, January 2007.

[147] K. Morokuma. Why do molecules interact? the origin of electron donor-acceptor complexes, hydrogen bonding and proton affinity. *Accounts Chem Res*, 10(8):294–300, 1977.

[148] A. Mukhopadhyay, B. H. Aguilar, I. S. Tolokh, and A. V. Onufriev. Introducing charge hydration asymmetry into the generalized Born model. *J Chem Theor Comp*, 10(4):1788–1794, 2014.

[149] A. Mukhopadhyay, A. T. Fenley, I. S. Tolokh, and A. V. Onufriev. Charge hydration asymmetry: the basic principle and how to use it to test and improve water models. *Journal of Physical Chemistry. B*, 116(32):9776–9783, Aug. 2012.

[150] H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling. Folding simulations for proteins with diverse topologies are accessible in days with a Physics-Based force field and implicit solvent. *J. Am. Chem. Soc.*, 136(40):13959–13962, Sept. 2014.

[151] S. Niu, M. L. Tan, and T. Ichiye. The large quadrupole of water molecules. *The Journal of Chemical Physics*, 134(13):134501+, 2011.

[152] H. Nymeyer and A. E. Garcia. Simulation of the folding equilibrium of α-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U. S.Á.*, 100(24):13934–13939, 2003.

[153] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput*, 2(2):420–433, 2006.

[154] A. Onufriev, D. Bashford, and D. Case. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B*, 104:3712–3720, 2000.

[155] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins*, 55(2):383–394, May 2004.

[156] S. Patel and C. L. Brooks. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.*, 25(1):1–16, Jan. 2004.

[157] D. Pearlman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, Sept. 1995.

[158] E. Pellegrini and M. J. Field. A generalized-born solvation model for macromolecular hybrid-potential calculations. *J. Phys. Chem. A*, 106(7):1316–1326, February 2002.

[159] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kal, and K. Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[160] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B*, 119(16):5113–5123, Apr. 2015.

[161] J. W. Pitera and W. Swope. Understanding folding and design: replica-exchange simulations of "trp-cage" miniproteins. *Proc Natl Acad Sci U S A*, 100(13):7587–7592, June 2003.

[162] D. Platt and B. Silverman. Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. *Journal of Computational Chemistry*, 17(3):358–366, 1996.

[163] E. L. Pollock and J. Glosli. Comments on PPPM, FMM, and the Ewald Method for Large Periodic Coulombic Systems, Nov. 1995.

[164] S. Pronk, S. Pll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[165] S. Rajamani, T. Ghosh, and S. Garde. Size dependent ion hydration, its asymmetry, and convergence to macroscopic behavior. *J Chem Phys*, 120(9):4457–4466, 2004.

[166] P. Ren and J. W. Ponder. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B*, 107(24):5933–5947, May 2003.

[167] H. S. Rhee, A. R. Bataille, L. Zhang, and B. F. Pugh. Subnucleosomal Structures and Nucleosome Asymmetry across a Genome. *Cell*, 159(6):1377–1388, Dec. 2014.

[168] S. W. Rick. A reoptimization of the five-site water potential (TIP5P) for use with ewald sums. *J Chem Phys*, 120(13):6085–6093, 2004.

[169] A. Robertson, E. Luttmann, and V. S. Pande. Effects of long-range electrostatic forces on simulated protein folding kinetics. *J. Comp. Chem.*, 29(5):694–700, 2007.

[170] P. J. J. Robinson, L. Fairall, V. A. T. Huynh, and D. Rhodes. Em measurements define the dimensions of the 30-nm chromatin fiber: Evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences*, 103(17):6506–6511, April 2006.

[171] A. J. Rusnak, E. R. Pinnick, C. E. Calderon, and F. Wang. Static dielectric constants and molecular dipole distributions of liquid water and ice-ih investigated by the paw-pbe exchange-correlation functional. *The Journal of Chemical Physics*, 137(3):–, 2012.

[172] F. R. Salsbury. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current Opinion in Pharmacology*, 10(6):738–744, Dec. 2010.

[173] M. Scarsi, J. Apostolakis, and A. Caflisch. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A*, 101(43):8098–8106, October 1997.

[174] M. Schaefer and M. Karplus. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.*, 100:1578–1599, 1996.

[175] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Interdisciplinary Applied Mathematics. Springer, 2010.

[176] R. Schmid, A. M. Miah, and V. N. Sapunov. A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii). *Phys Chem Chem Phys*, 2:97–102, 2000.

[177] K. E. Schmidt and M. A. Lee. Implementing the fast multipole method in three dimensions. *Journal of Statistical Physics*, 63(5-6):1223–1235, June 1991.

[178] H. Schreiber and O. Steinhauser. Molecular dynamics studies of solvated polypeptides: Why the cut-off scheme does not work. *Chem. Phys.*, 168(1):75–89, December 1992.

[179] G. Sigalov, A. Fenley, and A. Onufriev. Analytical electrostatics for biomolecules: Beyond the generalized Born approximation. *J. Chem. Phys.*, 124(12):124902, 2006.

[180] E. Sigfridsson and U. Ryde. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *Journal of Computational Chemistry*, 19(4):377–395, Mar. 1998.

[181] P. L. Silvestrelli and M. Parrinello. Structural, electronic, and bonding properties of liquid water from first principles. *J Chem Phys*, 111(8):3572–3580, 1999.

[182] C. Simmerling, B. Strockbine, and A. E. Roitberg. All-Atom Structure Prediction and Folding Simulations of a Stable Protein. *J. Am. Chem. Soc.*, 124:11258–11259, 2002.

[183] T. Simonson. Electrostatics and Dynamics of Proteins. *Rep. Prog. Phys.*, 66:737–787, 2003.

[184] L. D. Site, A. Alavi, and R. M. Lynden-Bell. The electrostatic properties of water molecules in condensed phases: an ab initio study. *Mol Phys*, 96(11):1683–1693, 1999.

[185] L. B. Skinner, C. Huang, D. Schlesinger, L. G. M. Pettersson, A. Nilsson, and C. J. Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. *J Chem Phys*, 138(7):074506, 2013.

[186] M. F. Smith, B. D. Athey, S. P. Williams, and J. P. Langmore. Radial density distribution of chromatin: evidence that chromatin fibers have solid centers. *The Journal of Cell Biology*, 110(2):245–254, Feb. 1990.

[187] V. Z. Spassov, L. Yan, and S. Szalma. Introducing an implicit membrane in generalized born/solvent accessibility continuum solvent models. *J. Phys. Chem. B*, 106(34):8726–8738, August 2002.

[188] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical Treatment of Solvation for Molecular Mechnics and Dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[189] F. H. Stillinger. Water revisited. *Science*, 209(4455):pp. 451–457, 1980.

[190] K. Stöbener, P. Klein, S. Reiser, M. Horsch, K.-H. Küfer, and H. Hasse. Multicriteria optimization of molecular force fields by pareto approach. *Fluid Phase Equilibria*, 373(0):100 – 108, 2014.

[191] A. Stone. *The Theory of Intermolecular Forces*. International Series of Monographs on Chemistry. Clarendon Press, 1997.

[192] M. Swart, P. T. van Duijnen, and J. G. Snijders. A charge analysis derived from an atomic multipole expansion. *Journal of Computational Chemistry*, 22(1):79–88, 2001.

[193] M. J. Tait and F. Franks. Water in Biological Systems. *Nature*, 230(5289):91–94, Mar. 1971.

[194] S. Tanizaki and M. Feig. A generalized Born formalism for heterogeneous dielectric environments: application to the implicit modeling of biological membranes. *J. Chem. Phys.*, 122(12):124706, March 2005.

[195] J. A. Te and T. Ichiye. Understanding structural effects of multipole moments on aqueous solvation of ions using the soft-sticky dipolequadrupoleoctupole water model. *Chemical Physics Letters*, 499(4-6):219–225, Oct. 2010.

[196] A. Y. Toukmaji and J. A. Board. Ewald summation techniques in perspective: A survey. *Computer Physics Communications*, 95(2-3):73–92, June 1996.

[197] P. Tröster, K. Lorenzen, and P. Tavan. Polarizable Six-Point Water Models from Computational and Empirical Optimization. *J. Phys. Chem. B*, 118(6):1589–1602, Jan. 2014.

[198] V. Tsui and D. Case. Molecular dynamics simulations of nucleic acids using a generalized Born solvation model. *J. Am. Chem. Soc.*, 122:2489–2498, 2000.

[199] V. Tsui and D. Case. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Am. Chem. Soc.*, 122(11):2489–2498, 2000.

[200] Y. Tu and A. Laaksonen. The electronic properties of water molecules in water clusters and liquid water. *Chemical Physics Letters*, 329(3-4):283–288, Oct. 2000.

[201] M. B. Ulmschneider, J. P. Ulmschneider, M. S. Sansom, and A. Di Nola. A generalized Born implicit-membrane representation compared to experimental insertion free energies. *Biophys. J.*, 92(7):2338–2349, April 2007.

[202] C. Vega and J. L. F. Abascal. Simulating water with rigid non-polarizable models: a general perspective. *Phys Chem Chem Phys*, 13:19663–19688, 2011.

[203] C. Vega, J. L. F. Abascal, M. M. Conde, and J. L. Aragones. What ice can teach us about water interactions: a critical comparison of the performance of different water models. *Faraday Discuss.*, 141:251–276, 2009.

[204] D. Venkateswarlu. Structural investigation of zymogenic and activated forms of human blood coagulation factor VIII: a computational molecular dynamics study. *BMC Struct. Biol.*, 10(1):7+, February 2010.

[205] R. C. Walker, M. F. Crowley, and D. A. Case. The implementation of a fast and accurate QM/MM potential method in Amber. *J. Comp. Chem.*, 29(7):1019–1031, 2007.

[206] T. R. Walsh and T. Liang. A multipole-based water potential with implicit polarization for biomolecular simulations. *J. Comput. Chem.*, 30(6):893–899, Apr. 2009.

[207] H.-W. W. Wang and E. Nogales. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature*, 435(7044):911–915, June 2005.

[208] J. Wang, P. Cieplak, Q. Cai, M.-J. J. Hsieh, J. Wang, Y. Duan, and R. Luo. Development of polarizable models for molecular mechanical calculations. 3. polarizable water models conforming to thole polarization screening schemes. *The journal of physical chemistry. B*, 116(28):7999–8008, July 2012.

[209] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J Comp Chem*, 25(9):1157–1174, 2004.

[210] L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande. Systematic improvement of a classical molecular model of water. *J Phys Chem B*, 117(34):9956–9972, 2013.

[211] L. P. Wang, T. J. Martinez, and V. S. Pande. Building force fields: An automatic, systematic, and reproducible approach. *J Phys Chem Lett*, 5(11):1885–1891, 2014.

[212] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.*, 30:211–243, 2001.

[213] K. Watanabe and M. L. Klein. Effective pair potentials and the properties of water. *Chemical Physics*, 131(2-3):157–167, Mar. 1989.

[214] H. Wong, J.-M. Victor, and J. Mozziconacci. An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLOS One*, 436(7):e877, 2007.

[215] H. Wong, J.-M. Victor, and J. Mozziconacci. An All-Atom Model of the Chromatin Fiber Containing Linker Histones Reveals a Versatile Structure Tuned by the Nucleosomal Repeat Length. *PLOS One*, 436(7):e877+, 2007.

[216] D. E. Woon and T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *Journal of Chemical Physics*, 103:4572, 1995.

[217] Y. Wu, H. L. Tepper, and G. A. Voth. Flexible simple point-charge water model with improved liquid-state properties. *The Journal of chemical physics*, 124(2):024503+, Jan. 2006.

[218] Z. Wu, Q. Cui, and A. Yethiraj. A new Coarse-Grained model for water: The importance of electrostatic interactions. *J. Phys. Chem. B*, 114(32):10524–10529, July 2010.

[219] L. Ying. A kernel independent fast multipole algorithm for radial basis functions. *Journal of Computational Physics*, 213(2):451–457, Apr. 2006.

[220] D. York and W. Yang. The fast fourier Poisson method for calculating Ewald sums. *J. Chem. Phys.*, 101(4):3298–3300, 1994.

[221] W. Yu, P. E. M. Lopes, B. Roux, and A. D. MacKerell. Six-site polarizable model of water based on the classical drude oscillator. *The Journal of Chemical Physics*, 138(3):034508+, Jan. 2013.

[222] B. Zagrovic and V. Pande. Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comp. Chem.*, 24(12):1432–1436, 2003.

[223] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *Journal of Molecular Biology*, 323(5):927–937, November 2002.

[224] L. Y. Zhang, E. Gallicchio, R. A. Friesner, and R. M. Levy. Solvent models for protein-ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *J. Comp. Chem.*, 22(6):591–607, 2001.

[225] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646, May 2013.