TRANSFERABLE COARSE-GRAINED MODELS: FROM HYDROCARBONS TO POLYMERS, AND BACKMAPPED BY MACHINE LEARNING

Yaxin An

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

Chemical Engineering

Sanket A. Deshmukh, Chair

Luke E. K. Achenie

Hongliang Xin

Xian-Ming Bai

December 4th, 2020

Blacksburg, Va

Keywords: Molecular dynamics simulations, Coarse-graining, Soft materials, Machine learning

TRANSFERABLE COARSE-GRAINED MODELS: FROM HYDROCARBONS TO POLYMERS, AND BACKMAPPED BY MACHINE LEARNING

Yaxin An

ABSTRACT

Coarse-grained (CG) molecular dynamics (MD) simulations have seen a wide range of applications from biomolecules, polymers to graphene and metals. In CG MD simulations, atomistic groups are represented by beads, which reduces the degrees of freedom in the systems and allows larger timesteps. Thus, large time and length scales could be achieved in CG MD simulations with inexpensive computational cost. The representative example of large time- and length-scale phenomena is the conformation transitions of single polymer chains as well as polymer chains in their architectures, self-assembly of biomaterials, etc. Polymers exist in many aspects of our life, for example, plastic packages, automobile parts, and even medical devices. However, the large chemical and structural diversity of polymers poses a challenge to the existing CG MD models due to their limited accuracy and transferabilities. In this regard, this dissertation has developed CG models of polymers on the basis of accurate and transferable hydrocarbon models, which are important components of the polymer backbone. CG hydrocarbon models were created with 2:1 and 3:1 mapping schemes and their force-field (FF) parameters were optimized by using particle swarm optimization (PSO). The newly developed CG hydrocarbon models could reproduce their experimental properties including density, enthalpy of vaporization, surface tension and self-diffusion coefficients very well. The cross interaction parameters between CG hydrocarbon and water models were also optimized by the PSO to repeat the experimental properties of Gibbs free energies and interfacial tensions. With the hydrocarbon models as the backbone, poly(acrylic acid) (PAA) and polystyrene (PS) models were constructed. Their side chains were represented by one COOH (carboxylic acid) and three BZ beads, respectively. Before testing the PAA and PS models, their monomer models, propionic acid and ethylbenzene, were created and validated, to confirm that the cross interactions between hydrocarbon and COOH beads, and between hydrocarbon and BZ beads could be accurately predicted by the Lorentz-Berthelot (LB) combining rules. Then the experimental properties, density of polymers at 300 K and glass transition temperatures, and the conformations of their all-atom models in solvent mixtures of water and dimethylformamide

(DMF) were reproduced by the CG models. The CG PAA and PS models were further used to build the bottlebrush copolymers of PAA-PS and to predict the structures of PAA-PA in different compositions of binary solvents water/DMF. Although CG models are useful in understanding the phenomena at large time- or length- scales, atomistic information is lost. Backmapping is usually involved in reconstructing atomistic models from their CG models. Here, four machine learning (ML) algorithms, artificial neural networks (ANN), k-nearest neighbor (kNN), gaussian process regression (GPR), and random forest (RF) were developed to improve the accuracy of the backmapped all-atom structures. These optimized four ML models showed R^2 scores of more than 0.99 when testing the backmapping against four representative molecules: furan, benzene, naphthalene, graphene.

TRANSFERABLE COARSE-GRAINED MODELS: FROM HYDROCARBONS TO POLYMERS, AND BACKMAPPED BY MACHINE LEARNING

Yaxin An

GENERAL AUDIENCE ABSTRACT

Polymers have a wide range of applications from packaging, foams, coating to pipes, tanks and even medical devices and biosensors. To improve the properties of these materials it is important to understand their structure and features responsible for controlling their properties at the molecular-level. Molecular dynamic (MD) simulations are a powerful tool to study their structures and properties at microscopic level. However, studying the molecular-level conformations of polymers and their architectures usually requires large time- or length-scales, which is challenging for the all-atom MD simulations because of the high computational cost. Coarse-grained (CG) MD simulations can be used to study these soft-materials as they represent atomistic groups with beads, enabling the reduction of the system sizes drastically, and allowing the use of large timesteps in MD simulations. In MD simulations, force-fields (FF) that describe the intramolecular and intermolecular interactions determine the performance of simulations. Here, we firstly optimized the FF parameters for hydrocarbons. With the optimized CG hydrocarbon models, two representative CG polymer models, poly(acrylic acid) (PAA) and polystyrene (PS) were built by using hydrocarbons as the backbones of polymers. Furthermore, the PAA and PS chains were grafted on a linear hydrocarbon backbone to form a bottlebrush copolymer. Although CG MD models are useful in studying the complex process of polymers, the atomic detailed information is lost. To reconstruct accurate atomistic structures, backmapping by using machine learning (ML) algorithms was performed. The performance of the ML models was better than that of the existing backmapping packages built in Visual Molecular Dynamics (VMD).

ACKNOWLEDGEMENT

I firstly would like to acknowledge my advisor Dr. Deshmukh, who not only gives me lots of constructive suggestions in the research, but also guides me in my career searching. He encouraged me to challenge myself when tricky problems came in my PhD research. Everytime when I felt frustrated with the difficulty in research, he would spend time discussing with me and inspire me to practice new research methods/skills. He is also open to suggestions from us students. When we have different opinions on the research, he would like to know about the students' ideas and encourage us to explore the field of our own interests. Besides the PhD research, he encouraged me to pursue my career in academia. I really appreciate his suggestions on my career path.

I also would like to thank the committee members, Dr. Achenie, Dr. Bai and Dr. Xin for their valuable suggestions on my work. Their professional advice complements the work a lot. Thanks also to Dr. Bejagam K. Karteek who was a postdoc researcher in the lab. He helped me so much, especially in my first year when I started to learn molecular dynamic simulations. He taught me patiently about the simulation and programming skills. I am also thankful to other previous members and all the current members of Dr. Deshmukh's group: Dr. Samrendra Singh, Olivia Conway, Dr. Fangxi Wang, Abhishek Sose, and Soumil Joshi.

In my Ph.D. study, I am fortunate to meet my friends Qiang Zhang, Dr. Yongliang Zhong, Noushin Omidivar, Hemanth Somarajan Pillai, Dr. Siwen Wang, Dr. Zheng Li and Dr. Jiaming Wang. With these friends, life in Virginia Tech is so happy and enriched. My gratitude finally goes to my parents, my brother, my husband and my daughter for their continuous support. They are the constant sources of motivation and love in my life.

Attribution

Chapter 1: An, Y. drafted this chapter and Deshmukh, S. A. revised it.

Chapter 2: An, Y. drafted this chapter and Deshmukh, S. A. revised it. Singh, S. developed the computer code/script for the PSO method.

Chapter 3: An, Y. set up all the molecular dynamics (MD) simulations. The computer codes/scripts for calculating RDFs, expansibility, compressibility, free energies by using the ABF method, and interfacial tensions were developed by An, Y. The computer codes/scripts for calculating enthalpy of vaporization, self-diffusion coefficient, and surface tension were developed by Bejagam, K. K.. Singh, S. developed the method for uncertainty quantification (UQ). All authors on the paper contributed in writing the research manuscript.

Chapter 4: An, Y. designed the mapping schemes for hydrocarbons, and performed all the analysis of the MD simulations for property calculations. Bejagam, K. K. contributed to integrating the PSO with MD simulations. An, Y. modified the PSO scripts for optimization of the FF parameters of hydrocarbons. Analysis was performed by An, Y. All authors contributed in research manuscript writing.

Chapter 5: An, Y. was responsible for the FF optimization, all the MD simulations and trajectory analysis. All authors contributed in writing the research manuscript..

Chapter 6: An, Y. developed the coarse-grained polymer models and performed related analysis. Singh, S. performed UQ. All authors contributed in writing the research manuscript.

Chapter 7: An, Y. was responsible for all the simulations and analysis of work.

Chapter 8: An, Y. was responsible for all the simulations and analysis of work. All authors contributed in manuscript writing.

Chapter 9: An, Y. was responsible for all the simulations and related analysis. Both authors contributed to manuscript writing.

Chapter 10: An, Y. proposed the future directions in this chapter.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Background of Hydrocarbons and Polymers

Hydrocarbons are important components of crude oil, natural gas, coal and other energy.[1–3] They are the basis of the vast majority of global energy production. They also constitute the backbones of many polymers and hydrophobic tails of biomolecules. Therefore, studying hydrocarbons experimentally and computationally is of much interest to the research community. Their macroscopic properties have been extensively studied in experiments.[4] However, studying their microscopic structures and properties is challenging for experimentalists. Molecular dynamics simulations have emerged as a powerful tool to study the dynamics process at a time-scale up to hundreds of nanoseconds and a length-scale of tens of Angstroms.[5] As the computational resources become less and less expensive, the time- and length-scales of MD are being increased from nanoseconds to milliseconds and tens to hundreds of Angstroms, respectively.[6] All-atom, united-atom and coarse-grained (CG) models of hydrocarbons have been developed for studying their physical properties such as solubilities and partition free energies.[7–10] CG models of hydrocarbons have attracted much attention due to their efficient computational cost. By using these CG hydrocarbon models, the models of soft materials, e.g. biomolecules and polymers have been constructed and employed to develop a molecular-level understanding of the structural and dynamical properties of proteins and polymers and their architectures.[11–13] However, the accuracy and transferability of the existing hydrocarbons are limited. The predicted self-diffusion coefficients of the well-known CG hydrocarbon models, MARTINI models are 2 - 3 times faster than their corresponding experimental properties. To improve the accuracy and transferability, different mapping schemes are designed and the CG force-field parameters are optimized for hydrocarbon models.

The large diversity of polymers play important roles in our life ranging from household items like plastic bags, to biomedical applications.[14–16] Depending on their chemical groups, polymers could be classified as hydrophilic and hydrophobic polymers. The representative examples of these hydrophilic and hydrophobic polymers are poly(acrylic acid) (PAA) and

polystyrene (PS), respectively. PAA is one of the important components in cosmetics products to improve hydration of skins.[17] It has also been used in the adsorption of heavy metals in water by forming complexes with heavy metal ions, and in modifying the surface hydrophilicity.[18] On the other hand, PS has been widely used for achieving hydrophobic surfaces.[19] By integrating these two polymers in one polymer, we could obtain a copolymer with both hydrophilicity and hydrophobicity.[20,21] By tuning the number or positions of these two polymers, the strength of hydrophilicity and hydrophobicity could be changed, and then further used in different applications.[22,23]

Bottlebrush polymers (BBPs) are a special class of polymers with their backbones highly grafted with polymers as side chains.[24] They have shown potential applications in surface coating, biomedical devices and bioimaging.[24–26] Due to its architectural complexity, understanding its conformations in different environments, e.g. in melt or in solutions is challenging.[27,28] Experimental and computational efforts have been made to build the architecture-conformation relationship. Dutta et al. studied the conformations of a series of BBPs consisting of a poly(norbornene) (PNB) backbone with poly(lactic acid) (PLA) side chains.[27] They found that the experimentally measured intrinsic viscosity and radius of gyration are highly sensitive to their architectures. Furthermore, they built a coarse-grained model to reproduce these two experimental properties. Sarapas et al. investigated the packing of BBPs in melt through using neutron scattering, and found the BBPs pack similarly to semidilute polyelectrolyte solutions, and that decreasing grafting density was analogous to increasing salt concentration in polyelectrolytes.[29] Despite these previous efforts, understanding the relationship between structure and properties of BBPs is still limited, especially for BBPs with different types of polymers as side chains, called bottlebrush copolymers. This is because of the large variety of bottlebrush copolymers which could be achieved by changing the functional groups, grafting densities and the chain lengths of side chains. Here, we used PAA and PS as side chains, and constructed different architectures of bottlebrush copolymers by altering the positions, chain lengths, and grafting densities of PAA and PS. The conformations of these BBP molecules were studied by performing CG MD simulations.

1.2 Molecular Dynamics Simulations

1.2.1 Classical MD

Molecular dynamics simulations are one of the computational techniques that use Newton's second law as the basis for studying the dynamic processes in the chemical, physical and biological phenomena at the molecular-level.[30] In classical MD simulations, also called all-atom MD simulations, each atom is represented explicitly. The flowchart in **Figure 1.1** shows the algorithm of how MD simulations work. The first step is to initialize the positions and velocities of each atom in the system. The initial velocity distribution should be close to the equilibrium distribution, and follows the Maxwell-Boltzman rules, strictly for ideal gas. The second step is to calculate the forces on each atom at time step t based on the empirical force field. Then the positions and velocities at the next step t + Δt will be obtained. Finally, the positions of atoms will be updated at the next step t + Δt. The whole trajectory will be saved for analysis.

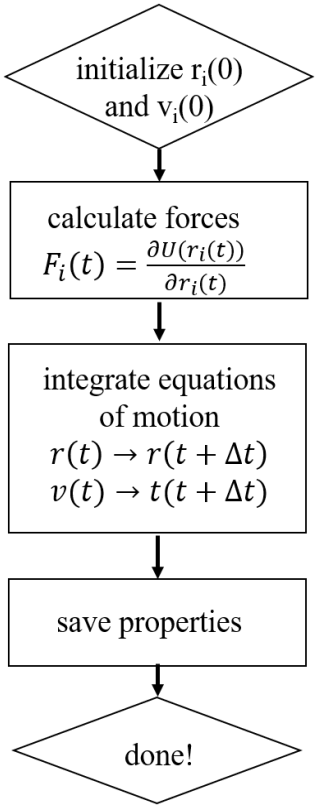

Figure 1.1 The flowchart showing the basic algorithm of MD simulations.

In MD simulations, the intramolecular and intermolecular interactions are described by force-fields (FFs). The classical format of a force field is shown in **Equation 1.1**.[31] It consists of intramolecular interactions which are the first four terms: bond stretching, angle bending, dihedral and improper torsion, and intermolecular interactions shown by the last two terms: the Van der Waals and the Coulombic interactions. The parameters in FFs are typically obtained either from ab initio or semi-empirical quantum mechanical calculations or by fitting to experimental data such as neutron, X-ray, etc.[31] However, the refinement of the FF parameters is not trivial because of the large number of parameters. Besides, these parameters are coupled with each other, which means that a change in one parameter may affect others. Traditionally, the FF optimization is performed manually by an iterative process, which is time-consuming and limits the development of accurate FFs. To accelerate the optimization and improve the accuracy, particle swarm optimization has been successfully used in developing the interatomic potentials.[32,33] In this thesis, the PSO algorithm would be used in deriving FF parameters.

$$E_{pot} = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} K_\varphi(1+cos(n\varphi-\delta)) + \sum_{impropers} K_\psi(\psi-\psi_0)$$

$$+ \sum_{i\neq j}\sum_{j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i\neq j}\sum_{j} \frac{q_i q_j}{r_{ij}} \qquad \textbf{…...Equation 1.1}$$

1.2.2 Coarse-Grained MD

All-atom MD simulations are challenging for studying large systems and slow dynamic processes due to the high computational cost. CG models have been developed to save computational time and reach a large time scale (μs) and length scale (millimeter).[34] The CG models represent atomic groups with beads, e.g. two/three/four heavy atoms represented with one bead, called 2:1/3:1/4:1 mapping schemes. The CG MD simulations have been used in studying the self-assembly of peptide amphiphiles or copolymers, protein conformation transformation, and transportation of small molecules across membranes.[11,35,36] The most widely used CG MD models are MARTINI models that include water, hydrocarbons, lipids, proteins, amino acids, and carbohydrates.[37,38] In developing MARTINI models, the FF parameters were

optimized to mainly reproduce experimental macroscopic properties such as density and partition free energies, which is called a top-down approach. Besides the top-down approach, the bottom-up approach is implemented to replicate structural properties from all-atom simulations.[39] By combining the top-down and the bottom-up approaches, a hybrid approach has been proposed to achieve both experimental properties at macroscopic level and structure properties at the atomic level.[40] In this thesis, the hybrid approach would be used to develop CG models of polymers.

1.2.3 Backmapping

Although CG MD simulations are useful in studying complex chemical, biological and physical phenomena with large time or length scales, atomistic details are lost in these CG models.[41,42] To gain these atomistic information, backmapping or inverse mapping or reverse transformation can be performed to reconstruct the all-atom structures from CG models.[43] A traditional backmapping involves two steps: (*Step-i*) generate initial atomistic structures and (*Step-ii*) run simulated annealing to get a reasonable atomistic structure.[41] different algorithm-based methods such as fragment methods, geometric rules, and random placement of atoms have been used to perform *Step-i*.[41] For example, these methods have been used to backmap CG models of peptides, lipids, and proteins to their all-atom models.[41,42] However, the accuracy of the existing backmapping algorithms need improvement, or they require extensive computational/coding efforts. Machine learning (ML) has emerged as an inexpensive technique and attracted lots of attention recently from the computational chemists. They have been used in designing polymers, enhancing sampling in free energy calculating, and developing FF parameters. In this thesis, we proposed to use machine learning (ML) algorithms to improve the accuracy of backmapping and save computational time.

1.3 Research Goals

This thesis is focused on the development of accurate and transferable CG models of hydrocarbons and polymers by integrating MD simulations with optimization algorithms. The thesis also develops novel machine learning (ML) based approaches for accurate backmapping

CG models to their atomistic models. In **Chapter 2**, we reviewed the methodologies of developing CG models and introduced the method used in this thesis. In **Chapters 3**, the methods of calculating properties of CG models were described. In Chapters 4 and 5, we developed CG models of hydrocarbons and optimized the cross interactions between hydrocarbons and water. In **Chapters 6, 7** and **8**, the CG models poly(acrylic acid) (PAA) and polystyrene (PS), the bottlebrush copolymers PAA-PS were constructed, and their conformations in binary solvents of water and DMF were investigated. In **Chapter 9**, all-atom models were reconstructed by using machine learning for six molecules including furan, benzene, hexane, naphthalene, graphene, and fullerene as a proof of concept. In **Chapter 10**, the future research directions were proposed.

References

(1) Cortright, R. D.; Davda, R. R.; Dumesic, J. A. Hydrogen from Catalytic Reforming of Biomass-Derived Hydrocarbons in Liquid Water. In *Materials for Sustainable Energy*; Co-Published with Macmillan Publishers Ltd, UK, 2010; pp 289–292.

(2) Park, S.; Vohs, J. M.; Gorte, R. J. Direct Oxidation of Hydrocarbons in a Solid-Oxide Fuel Cell. *Nature* **2000**, *404* (6775), 265–267.

(3) Preclik, D.; Hagemann, G.; Knab, O.; Brummer, L.; Mading, C.; Wiedmann, D.; Vuillermoz, P. LOX/Hydrocarbon Propellant Trade Considerations for Future Reusable Liquid Booster Engines. In *41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*; Joint Propulsion Conferences; American Institute of Aeronautics and Astronautics, 2005.

(4) Yaws, C. L. *Thermophysical Properties of Chemicals and Hydrocarbons*; William Andrew, 2008.

(5) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*; Springer Science & Business Media, 2013.

(6) Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; 2014; pp 41–53.

(7) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase. *J. Comput. Chem.* **2001**, *22* (11), 1205–1218.

(8) Eichenberger, A. P.; Huang, W.; Riniker, S.; van Gunsteren, W. F. Supra-Atomic Coarse-Grained GROMOS Force Field for Aliphatic Hydrocarbons in the Liquid Phase. *J. Chem. Theory Comput.* **2015**, *11* (7), 2925–2937.

(9) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *J. Am. Chem. Soc.* **1984**, *106* (22), 6638–6646.

(10) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free Energies of Hydration and

Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.* **1994**, *98* (49), 13077–13082.
(11) Souza, P. C. T.; Thallmair, S.; Conflitti, P.; Ramírez-Palacios, C.; Alessandri, R.; Raniolo, S.; Limongelli, V.; Marrink, S. J. Protein-Ligand Binding with the Coarse-Grained Martini Model. *Nat. Commun.* **2020**, *11* (1), 3714.
(12) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834.
(13) Panizon, E.; Bochicchio, D.; Monticelli, L.; Rossi, G. MARTINI Coarse-Grained Models of Polyethylene and Polypropylene. *J. Phys. Chem. B* **2015**, *119* (25), 8209–8216.
(14) Aguilar, M. R.; Elvira, C.; Gallardo, A.; Vázquez, B.; Román, J. S. Smart Polymers and Their Applications as Biomaterials. *Topics in tissue engineering* **2007**, *3* (6).
(15) Lopes, M. S.; Jardini, A. L.; Filho, R. M. Poly (Lactic Acid) Production for Tissue Engineering Applications. *Procedia Engineering* **2012**, *42*, 1402–1413.
(16) Gandhi, A.; Paul, A.; Sen, S. O.; Sen, K. K. Studies on Thermoresponsive Polymers: Phase Behaviour, Drug Delivery and Biomedical Applications. *Asian J. Pharm. Sci.* **04/2015**, *10* (2), 99–107.
(17) Wiśniewska, M.; Nosal-Wiercińska, A.; Ostolska, I.; Sternik, D.; Nowicki, P.; Pietrzak, R.; Bazan-Wozniak, A.; Goncharuk, O. Nanostructure of Poly(Acrylic Acid) Adsorption Layer on the Surface of Activated Carbon Obtained from Residue After Supercritical Extraction of Hops. *Nanoscale Res. Lett.* **2017**, *12* (1), 2.
(18) Yadav, V.; Harkin, A. V.; Robertson, M. L.; Conrad, J. C. Hysteretic Memory in pH-Response of Water Contact Angle on Poly(acrylic Acid) Brushes. *Soft Matter* **2016**, *12* (15), 3589–3599.
(19) Xu, J.; Li, M.; Zhao, Y.; Lu, Q. Control over the Hydrophobic Behavior of Polystyrene Surface by Annealing Temperature Based on Capillary Template Wetting Method. *Colloids Surf. A Physicochem. Eng. Asp.* **2007**, *302* (1), 136–140.
(20) Peng, D.; Feng, C.; Lu, G.; Zhang, S.; Zhang, X.; Huang, X. A Starlike Amphiphilic Graft Copolymer with Hydrophilic Poly(acrylic Acid) Backbones and Hydrophobic Polystyrene Side Chains. *J. Polym. Sci. A Polym. Chem.* **2007**, *45* (16), 3687–3697.
(21) Liu, C.; Chen, G.; Sun, H.; Xu, J.; Feng, Y.; Zhang, Z.; Wu, T.; Chen, H. Toroidal Micelles of Polystyrene- Block -Poly(acrylic Acid). *Small* **2011**, *7* (19), 2721–2726.
(22) Choucair, A.; Lavigueur, C.; Eisenberg, A. Polystyrene-B-Poly(acrylic Acid) Vesicle Size Control Using Solution Properties and Hydrophilic Block Length. *Langmuir* **2004**, *20* (10), 3894–3900.
(23) Zhang, L.; Eisenberg, A. Multiple Morphologies and Characteristics of "Crew-Cut" Micelle-like Aggregates of Polystyrene-B-Poly(acrylic Acid) Diblock Copolymers in Aqueous Solutions. *J. Am. Chem. Soc.* **1996**, *118* (13), 3168–3181.
(24) Verduzco, R.; Li, X.; Pesek, S. L.; Stein, G. E. Structure, Function, Self-Assembly, and Applications of Bottlebrush Copolymers. *Chem. Soc. Rev.* **2015**, *44* (8), 2405–2420.
(25) Fenyves, R.; Schmutz, M.; Horner, I. J.; Bright, F. V.; Rzayev, J. Aqueous Self-Assembly of Giant Bottlebrush Block Copolymer Surfactants as Shape-Tunable Building Blocks. *J. Am. Chem. Soc.* **2014**, *136* (21), 7762–7770.
(26) Mei, H.; Laws, T. S.; Mahalik, J. P.; Li, J.; Mah, A. H.; Terlier, T.; Bonnesen, P.; Uhrig, D.; Kumar, R.; Stein, G. E.; Verduzco, R. Entropy and Enthalpy Mediated Segregation of

Bottlebrush Copolymers to Interfaces. *Macromolecules* **2019**, *52* (22), 8910–8922.
(27) Dutta, S.; Wade, M. A.; Walsh, D. J.; Guironnet, D.; Rogers, S. A.; Sing, C. E. Dilute Solution Structure of Bottlebrush Polymers. *Soft Matter* **2019**, *15*, 2928–2941.
(28) Wessels, M. G.; Jayaraman, A. Molecular Dynamics Simulation Study of Linear, Bottlebrush, and Star-like Amphiphilic Block Polymer Assembly in Solution. *Soft Matter* **2019**, *15* (19), 3987–3998.
(29) Sarapas, J. M.; Martin, T. B.; Chremos, A.; Douglas, J. F.; Beers, K. L. Bottlebrush Polymers in the Melt and Polyelectrolytes in Solution Share Common Structural Features. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (10), 5168–5175.
(30) Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids. *Clarendon, Oxford* **1987**.
(31) González, M. A. Force Fields and Molecular Dynamics Simulations. *JDN* **2011**, *12*, 169–200.
(32) Yang, S.; Cui, Z.; Qu, J. A Coarse-Grained Model for Epoxy Molding Compound. *J. Phys. Chem. B* **2014**, *118* (6), 1660–1669.
(33) Cui, Z.; Gao, F.; Cui, Z.; Qu, J. Developing a Second Nearest-Neighbor Modified Embedded Atom Method Interatomic Potential for Lithium. *Modell. Simul. Mater. Sci. Eng.* **2011**, *20* (1), 015014.
(34) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801–6822.
(35) Hoffmann, C.; Centi, A.; Menichetti, R.; Bereau, T. Molecular Dynamics Trajectories for 630 Coarse-Grained Drug-Membrane Permeations. *Sci Data* **2020**, *7* (1), 51.
(36) Lee, O.-S.; Cho, V.; Schatz, G. C. Modeling the Self-Assembly of Peptide Amphiphiles into Fibers Using Coarse-Grained Molecular Dynamics. *Nano Lett.* **2012**, *12* (9), 4907–4913.
(37) López, C. A.; Rzepiela, A. J.; de Vries, A. H.; Dijkhuizen, L.; Hünenberger, P. H.; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5* (12), 3195–3210.
(38) Periole, X.; Marrink, S.-J. The Martini Coarse-Grained Force Field. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; pp 533–565.
(39) Hu, X.; Hu, J.; Tian, J.; Ge, Z.; Zhang, G.; Luo, K.; Liu, S. Polyprodrug Amphiphiles: Hierarchical Assemblies for Shape-Regulated Cellular Internalization, Trafficking, and Drug Delivery. *J. Am. Chem. Soc.* **2013**, *135* (46), 17617–17629.
(40) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.
(41) Peng, J.; Yuan, C.; Ma, R.; Zhang, Z. Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference. *J. Chem. Theory Comput.* **2019**, *15* (5), 3344–3353.
(42) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **2014**, *10* (2), 676–690.
(43) Spyriouni, T.; Tzoumanekas, C.; Theodorou, D.; Müller-Plathe, F.; Milano, G. Coarse-Grained and Reverse-Mapped United-Atom Simulations of Long-Chain Atactic Polystyrene Melts: Structure, Thermodynamic Properties, Chain Conformation, and

Entanglements. *Macromolecules* **2007**, *40* (10), 3876–3885.

CHAPTER 2
METHODOLOGY OF COARSE-GRAINED MODEL DEVELOPMENT

2.1 Introduction

This chapter reviews the reported methods on coarse-grained (CG) model development, the mapping schemes, and FF optimization methods used in this thesis. Development of CG models usually involves two general steps: (i) designing mapping schemes, and (ii) optimizing the FF parameters. Both these two steps determine the accuracy of CG models, and are equally important.[1]

2.2 Mapping Schemes

Mapping schemes are rules that represent atomistic groups with CG beads.[2] Because the procedure of mapping atomistic groups onto a CG description is not rigorously defined, different approaches have been employed to do the mapping. The traditional mapping schemes are usually designed manually, which are based on physical and chemical institutions, or even for convenience. The CG beads may represent functional groups, residues or monomers, or they simply aim for a specific resolution in which each CG bead represents *n* heavy atoms. For example, two/three/four heavy atoms in all-atom models are represented with one CG bead, called 2:1, 3:1, and 4:1 mapping schemes.[3–5] The number of heavy atoms combined into one bead is referred to be the degree of coarse-graining of the models. In general, the larger the CG degree is, the larger the time scale is achieved, however, the more structural/physical/chemical properties might be lost. One representative example is the mapping schemes of the CG benzene model. A 2:1 mapping scheme was used in literature to maintain its ring-like and planar structures, however, the symmetric characteristic is lost.[6] With a 3:1 mapping scheme, both the ring-like and planar structures are lost with the symmetric geometry kept.[7] The most well-known MARTINI models use different mapping schemes in representing a larger range of molecules.[6,8–10] The water and hydrocarbon models are represented with a 4:1 mapping scheme, while a 2:1 mapping scheme is used for ringlike molecules to keep their geometric specificity. For the 20 amino acids, most of them are mapped to one bead. Besides, each nucleotide is mapped to six or seven CG beads. These nucleotide models are used to create DNA molecules

with their backbone modeled with three beads by mapping the phosphate to one and the sugar to two beads.[11]

Besides the traditional manual methods of designing mapping schemes, efforts have been made to automate it. Webb et al. developed a graph-based algorithm to generate CG representations of toluene, pentadecane, a polysaccharide dimer, and a rhodopsin protein.[12] In this approach, the molecule is treated as a molecular graph with atoms as nodes and bonds as edges. By using edge contraction to combine the nodes, successively coarser representations of the original atomic system can be produced with their chemical topologies preserved. Wang et al. used a variational auto-encoder (VAE) framework in constructing the CG representations by the encoder and reconstructing the atomistic structures by the decoder.[13] This framework has been tested on several systems including gas-phase ortho-terphenyl (OTP), aniline, dipeptide, and liquid alkanes.

2.3 FF Optimization

The philosophies employed to optimize the FFs for CG models can be classified into three categories based on the target properties used in optimization: (i) bottom-up, (ii) top-down, and (iii) hybrid approaches.[1] In the bottom-up approach, structural properties from all-atom simulations are used as targets to optimize the FF parameters.[14] The most well-known algorithm for reproducing the structural properties is Iterative Boltzmann Inversion (IBI).[15] It's an iterative algorithm to obtain the pairwise bead interactions by matching a set of structural quantities such as radial distribution functions (RDFs) from all-atom simulations. Because IBI is a structure-based optimization method, it has limitations in predicting macroscopic properties of models. On the other hand, the top-down approach uses experimental/macroscopic properties of the studied molecules as targets to tune the FF potentials.[16–18] However, the structural properties at microscopic level might be lost by using this approach. The hybrid approach is a combination of the bottom-up and top-down methods, which tunes the bonded FF potentials based on the microscopic properties from all-atom models and the nonbonded FF parameters to replicate the macroscopic properties.[1] This approach has been successfully used in developing quite a few CG models, and the most popular one are MARTINI models as mentioned in **Section 3.1**. The MARTINI polarizable water models were developed to reproduce the experimental dielectric constant, density, the partition free energies between water and organic solvents for a variety of

small compounds.[5,6,8,9] In developing CG DNA models, bonded parameters were optimized based on reference all-atom simulations of short single-stranded DNA (ssDNA). The nonbonded interactions were tuned to reproduce the experimental free energies of partitioning from water to hydrated octanol and from water to chloroform of nucleotide bases.

2.4 Mapping Schemes and Optimization Algorithms in This Study

Mapping schemes: We selected 2:1 and 3:1 mapping schemes in this study to represent hydrocarbons and polymers. With these 2:1 and 3:1 mapping schemes, the hydrocarbons with an even and odd number of carbon atoms can all be represented. Another reason behind this is to be consistent with the mapping schemes used in our recently developed CG water models.[3] More details of the mapping schemes of specific molecules will be discussed in each chapter of this thesis.

FF optimization: Traditionally, FF parameter optimization is carried out manually, or by trials and errors. This is usually tedious due the multi parameters in FF potentials. To expedite the process of FF parameters optimization, the particle swarm optimization (PSO) algorithm was utilized. The PSO algorithm is one such population-based evolutionary optimization method, which has been integrated with MD simulations to develop FF parameters.[19] The FFs parameters are optimized by PSO to reproduce the experimentally measured properties. It has been successfully applied to the optimization of CG FFs for water, graphene, etc.[3,20] The flowchart of the process of optimizing the CG FF parameters using PSO is shown in **Figure 2.1** and described below:

(1) We initiate a PSO run with N particles to search the optimized parameters for the CG models. A particle represents an MD simulation (or set of input parameters that need to be optimized). In the first PSO iteration, N particles were assigned with random locations (input values of the parameters) and random velocities.

(2) Using these N sets of input values, CG MD simulations were carried out using NAMD 2.12.[21] The simulations were performed under specified conditions, and the trajectory was analyzed for determining the target properties under study.

(3) By comparing the properties obtained from these simulations with the corresponding experimental values, the fitness value for each simulation was estimated using Equation **2.1**. The fitness value represents the efficiency of the given FF parameter set to predict the target experimental properties.

$$f(x) = \sum_{i=1}^{K} \left[\frac{g_i(x)}{y_i} - 1 \right]^2 \quad \textbf{…….Equation 2.1}$$

where, y_i and $g_i(x)$ (i = 1, 2, 3…, *K*) are the target experimental and simulated properties of hydrocarbons, respectively. K represents the number of target quantities. The fitness of each particle is compared with the tolerance (0.05). If no fitness value is smaller than the tolerance, then optimization is continued by updating the positions and velocities of particles.

(4) The position of the particle with the highest fitness in each cycle is referred as its personal best p_{best}, and the position of the particle with the highest fitness in all past cycles is recorded as g_{best}. In the first cycle, p_{best} is equal to g_{best}. In the second and following cycles, g_{best} would be updated if it is larger than p_{best} in the current cycle, and replaced with the values of the current p_{best}. The position (x) and velocity (*V)* of the particles in the second cycle is calculated using **Equations 2.2** and **2.3**, respectively.[3]

$$V_{n+1} = \text{w} * V_n + c1 * \text{ rand()} * (g_{best,n} - x_n) + c2 * \text{rand()} * (p_{best,n} - x_n) \quad \textbf{……. Equation 2.2}$$

$$x_{n+1} = x_n + V_{\text{n+1}} \quad \textbf{……. Equation 2.3}$$

Where n is the cycle number, w is inertia factor, c1 and c2 are swarm and personal constants that determine the relative "pull" of g_{best} and p_{best}, respectively. The rand() is used to generate a number between 0 and 1.

(5) Steps (2), (3), and (4) are repeated until at least one particle with the fitness less than the tolerance is found.

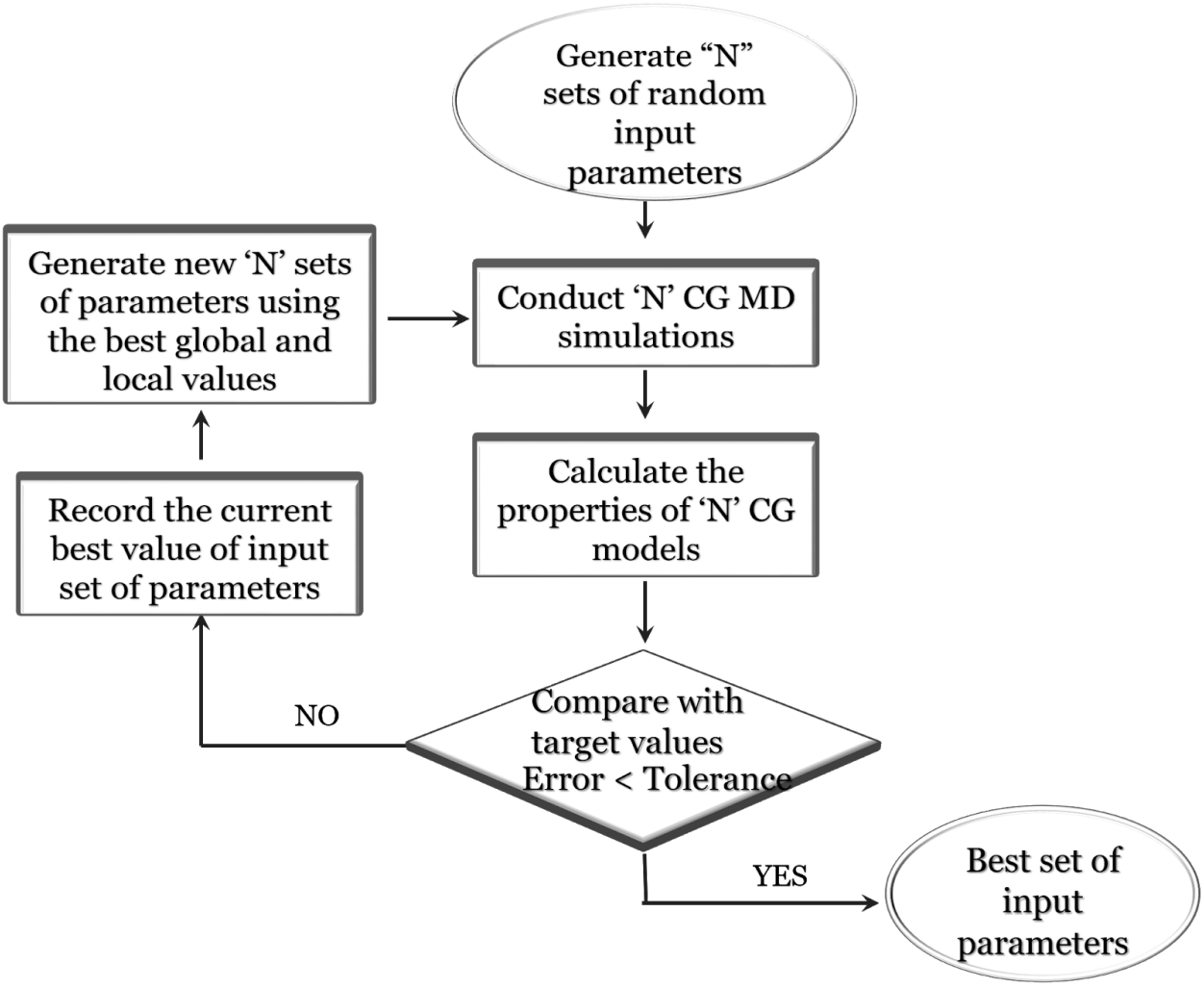

Figure 2.1 Flowchart describing the optimization of CG FF parameters using PSO.

References

(1) Joshi, S. Y.; Deshmukh, S. A. A Review of Advancements in Coarse-Grained Molecular Dynamics Simulations. *Mol. Simul.* **2020**, 1–18.

(2) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801–6822.

(3) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.

(4) Gyawali, G.; Sternfield, S.; Kumar, R.; Rick, S. W. Coarse-Grained Models of Aqueous and Pure Liquid Alkanes. *J. Chem. Theory Comput.* **2017**, *13* (8), 3846–3853.

(5) Yesylevskyy, S. O.; Sch?fer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Comput. Biol.* **2010**, *6* (6), 1–17.

(6) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **07/2007**, *111* (27), 7812–7824.

(7) Neverov, V. S.; Komolkin, A. V. Coarse-Grain Model of the Benzene Ring with Para-Substituents in the Molecule. *J. Chem. Phys.* **2012**, *136* (9), 094102.

(8) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834.

(9) López, C. A.; Rzepiela, A. J.; de Vries, A. H.; Dijkhuizen, L.; Hünenberger, P. H.; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5* (12), 3195–3210.

(10) Periole, X.; Marrink, S.-J. The Martini Coarse-Grained Force Field. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; pp 533–565.

(11) Uusitalo, J. J.; Ingólfsson, H. I.; Akhshi, P.; Tieleman, D. P.; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to DNA. *J. Chem. Theory Comput.* **2015**, *11* (8), 3932–3945.

(12) Webb, M. A.; Delannoy, J.-Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *J. Chem. Theory Comput.* **2019**, *15* (2), 1199–1208.

(13) Wang, W.; Gómez-Bombarelli, R. Coarse-Graining Auto-Encoders for Molecular Dynamics. *npj Computational Materials 5* (125), 1–9.

(14) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press, 2008.

(15) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of Coarse-Grained Potentials via Multistate Iterative Boltzmann Inversion. *J. Chem. Phys.* **2014**, *140* (22), 224104.

(16) Herdes, C.; Ervik, Å.; Mejía, A.; Müller, E. A. Prediction of the Water/oil Interfacial Tension from Molecular Simulations Using the Coarse-Grained SAFT-γ Mie Force Field. *Fluid Phase Equilib.* **2017**. https://doi.org/10.1016/j.fluid.2017.06.016.

(17) Herdes, C.; Totton, T. S.; Müller, E. A. Coarse Grained Force Field for the Molecular Simulation of Natural Gases and Condensates. *Fluid Phase Equilib.* **2015**, *406*, 91–100.

(18) Abbott, L. J.; Stevens, M. J. A Temperature-Dependent Coarse-Grained Model for the Thermoresponsive Polymer poly(N-Isopropylacrylamide). *J. Chem. Phys.* **2015**, *143* (24), 244901.

(19) Yang, S.; Cui, Z.; Qu, J. A Coarse-Grained Model for Epoxy Molding Compound. *J. Phys. Chem. B* **2014**, *118* (6), 1660–1669.
(20) Bejagam, K. K.; Singh, S.; Deshmukh, S. A. Development of Non-Bonded Interaction Parameters between Graphene and Water Using Particle Swarm Optimization. *J. Comput. Chem.* **2017**, *39* , 721–734.
(21) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.

CHAPTER 3

METHODOLOGY AND COMPUTATIONAL DETAILS

3.1 Introduction

A general methodology and computational details of all the simulations performed in this thesis are described here. Initially structural analysis including bond and angle distributions obtained from all-atom mapped and coarse-grained (CG) molecular dynamics (MD) simulation trajectories is discussed in detail. This is followed by a description of the methods of calculating properties of hydrocarbons, and hydrocarbon/water systems, such as enthalpy of vaporization, surface tension or interfacial tension etc. Finally, all-atom and CG MD simulations of hydrocarbon bulk, hydrocarbon/water mixture, poly(acrylic acid) (PAA) and polystyrene (PS) bulk, PAA and PS in solvents, PS-PAA bottlebrush polymers (BBPs) were described in details.

3.2 Structural Analysis of Mapped All-Atom Simulations

Based on mapping schemes, the positions of the CG beads were obtained by determining the center of mass (COM) of its represented group as given in **Equation 3.1.** Using this mapped all-atom trajectory, we have determined bond, angle, and end-to-end distance distributions, which can correlate intramolecular structure and also pair correlation function between the COM of different molecules to compare with our newly developed CG models.

$$r_{CG} = \frac{1}{M} \sum_{i=1}^{N} m_i r_i \qquad \text{..……..... Equation 3.1}$$

where N is the number of atoms that CG bead constitutes, M is the total mass of all the atoms that are grouped together, and r_i is the position of atom i. The end-to-end distance was the distance between two end beads within a given molecule. The pair correlation distribution function was calculated by using the following equation:

$$g(r) = \frac{\rho(r)}{\rho^{id}} = \frac{V}{N^2 4\pi r^2 dr} \sum_{i=1}^{N} \Delta N_i(r) \qquad \text{……. Equation 3.2}$$

where N is the total number of beads, $\Delta N_i(r)$ is the number of beads at the distance of r, V is the total volume of the simulation box.

3.3 Property Calculations

3.3.1 Enthalpy of Vaporization

The enthalpy of vaporization is associated with the non-bonded interaction parameters. Strength of these interactions are defined by ε in the LJ potential.[1–3] The enthalpy of vaporization was calculated in the same way as described in reference [1] and in the following Equation:

$$\Delta H_V(T) = \frac{1}{N_{mol}} E_{gas}(T) - \frac{1}{N_{mol}} E_{liq}(T) + RT \qquad \text{……. Equation 3.3}$$

where E is the total potential energy in the gas phase and liquid phase, respectively, R is the gas constant, T is the temperature, and N_{mol} is the number of molecules.

3.3.2 Self-Diffusion Coefficient

The self-diffusion coefficient of the CG hydrocarbons models is determined by using the mean squared displacement (MSD) and the Einstein's relation:

$$D = \left[\frac{1}{6t} < \Delta^2 r(t) >\right]_{t\to\infty} \qquad \text{…….Equation 3.4}$$

Where $< \Delta^2 r(t) >$ is the MSD, and <> denotes the average value over all molecules and time origins.

3.3.3 Surface Tension

To calculate the surface tension, we have performed NVT simulations of a slab containing specific molecules (hydrocarbons, propionic acid, and ethylbenzene) by expanding the Z-axis of a cubic cell of equilibrated bulk simulation of 1000 molecules by 150 Å. The production runtime was 7.5 ns with timestep of 15 fs by employing periodic boundary conditions. The last 4.5 ns trajectory was used to calculate the average value for the surface tension. In these simulations, the diagonal elements of pressure tensors, P_{xx}, P_{yy} and P_{zz} were printed for every time-step.

$$\gamma = \frac{L_z}{4} < 2P_{zz} - (P_{xx} - P_{yy}) > \qquad \text{…….Equation 3.5}$$

where, L_{Z} is the box-length along Z-direction and the angular brackets denote an ensemble average.

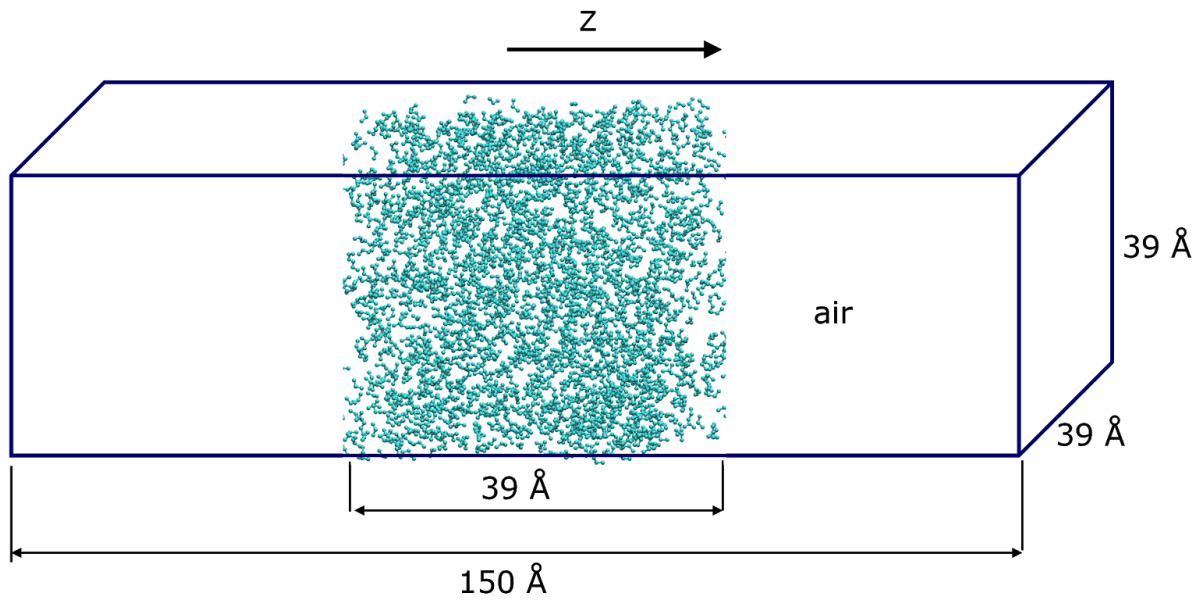

Figure 3.1 Schematic simulation set-up of calculating surface tensions of hydrocarbons. Cyan beads represent hydrocarbon models.

3.3.4 Isothermal Compressibility

The coefficient is calculated using a finite-difference expression:

$$k_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right) = -\frac{\partial(lnV)}{\partial P} = \frac{\partial(ln\rho)}{\partial P} \approx \frac{ln\frac{\rho_1}{\rho_2}}{P_1 - P_2} \qquad \textbf{.....Equation 3.6}$$

where ρ is the density of the system. P_1 and P_2 are the pressure at ρ_1 and ρ_2, respectively. For each value of compressibility, three NVT simulations were carried out at a given temperature with three different volumes. The densities for the three NVT simulations were ρ_0 and $\rho_0 \pm 0.03\ g/cc$, respectively, where ρ_0 is the equilibrium density obtained in NPT simulations. All the NVT simulations were run for 15 ns with 10 ps of minimization.

3.3.5 Expansibility

The expansibility was calculated by using the following equation:

$$\alpha_T = \frac{1}{V}\left(\frac{\partial v}{\partial T}\right)_P \approx \left(\frac{ln(\frac{V'}{V})}{T' - T}\right)_P \qquad \textbf{..... Equation 3.7}$$

where ρ is the density, V and V' are the volumes of the system at T and T', respectively.[4] For each value of expansibility, three NPT simulations were carried out at 1 atm with three different temperatures: T and $T \pm 10$ K (T'). All the NPT simulations were run for 15 ns with 10 ps of minimization.

3.3.6 Gibbs Hydration and Solvation Free Energies

Gibbs free energy has been used to validate the cross interactions between solute and solvents.[1,5] When the molecule of target is pulled in water and other organic solvent, it's referred as Gibbs hydration and solvation energies, respectively. The Gibbs hydration or solvation free energies of hydrocarbon, propionic acid, and DMF models were calculated by employing the adaptive biasing force (ABF) method implemented in colvars package.[6–8] The ABF method is a promising strategy to map complicated free-energy landscapes. It's conceptually and practically simple, where a biasing force is added to the equation of motion during the simulation process, hence, the system of interest could escape from kinect traps. The ABF method has been utilized in many fields such as protein-ligand and protein-protein recognition and association.[8,9]

Here, the Gibbs hydration free energy of hydrocarbons was shown as a representative example. Schematic simulation setup for these ABF calculations performed in this study is shown in **Figure 3.2**. An equilibrated simulation box (~39 Å × ~39 Å × ~39 Å) with 1000 1-site CG water molecules (2000 all-atom water molecules) was considered as the bulk water during these calculations. The air/water interface was created by extending the z-direction to 100 Å. A CG hydrocarbon molecule was placed in the vacuum at a distance of 35 Å in the z-direction to the COM of the water box. Simulations were performed in the NVT ensemble by employing periodic boundary conditions in all three directions. Reaction coordinate (RC) was defined as the distance between the COM of the water box and that of the CG hydrocarbon molecule projected to the z-direction (see **Figure 3.2**). 'DistanceZ' style of colvars module was utilized to perform these free energy simulations. The profile of Gibbs hydration free energy was obtained along the RC varying from 0 Å (center of bulk water) to 35 Å (vacuum). Entire RC was divided into smaller bins of 0.2 Å for the accumulation of force experienced on the RC and later it is averaged out for 500 force samples (F_{avg}). ABF applies an equal and opposite force along the entire RC to nullify the force experienced on the RC. ABF continues to apply the force so as to

smoothen the energy surface for uniform sampling.[1] A final production run of 120 ns was performed with a timestep of 5 fs at 300 K. Gibbs hydration free energies were determined from three independent initial configurations so as to reduce the statistical error.

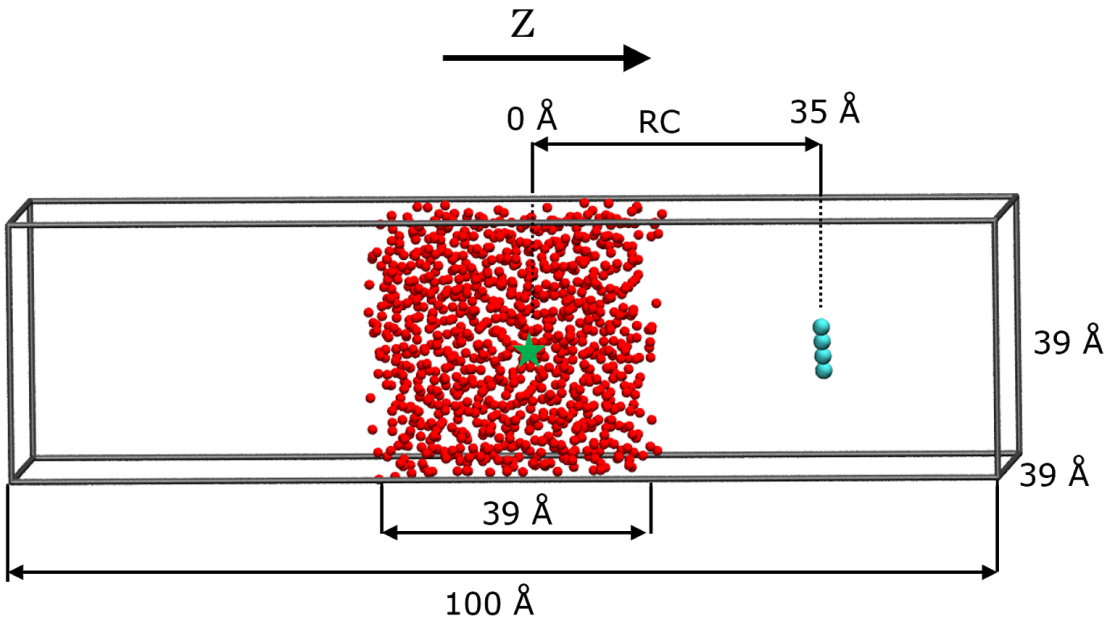

Figure 3.2 Schematic simulation setup to determine the Gibbs hydration free energy. 1-site water molecules are represented by red spheres, and a representative CG hydrocarbon molecule, decane (2-2-3-3), is shown in cyan spheres. z-direction is the direction of the reaction coordinate (RC) and is shown in black arrow. The green star represents the center of mass (COM) of the water bulk.

3.3.7 Interfacial Tensions of Hydrocarbon/Water Systems

The interfacial tensions of hydrocarbon/water systems were calculated at 300 K to test the new FF parameters. The interface of water/hydrocarbons was created by placing a water box, with size of 50 Å × 50 Å × 30 Å, and a hydrocarbon box, of the same size, side-by-side as shown in **Figure 3.3**. The z-direction of the simulation box is extended to 150 Å to create air/water and air/hydrocarbon interfaces and also to avoid the second water/hydrocarbon interface. The number of CG water and hydrocarbon molecules in each system was chosen based on their bulk density in references [1,10], and are reported in **Table 3.1**. The total simulation time was 100 ns with a timestep of 15 fs. Note, CG MD simulations with a larger system size (80 Å × 80 Å × 300 Å) were also performed for 1 μs, which resulted in similar values (error < 6 %)

of interfacial tensions to the ones obtained by using the system with 50 Å × 50 Å × 30 Å. Hence, we have employed the system size of 50 Å × 50 Å × 30 Å and simulation time of 100 ns. All the simulations were carried out in an NVT ensemble with periodic boundary conditions in all three directions. The last 15 ns trajectory of total 100 ns simulations was divided into three equal blocks to determine the interfacial tensions and statistical errors. In these simulations, the diagonal elements of pressure tensors, P_{xx}, P_{yy} and P_{zz} were printed for every time-step.

$$\gamma_{tot} = L_Z \; < \; P_{zz} - \tfrac{1}{2}(P_{xx} - \; P_{yy}) > \qquad \textbf{…….Equation 3.8}$$

$$\gamma_{hydrocarbon/water} \; = \gamma_{tot} - \gamma_{air/water} - \gamma_{air/hydrocarbon} \qquad \textbf{…….Equation 3.9}$$

In **Equation 3.8**, L_Z is the box-length along z-direction and the angular brackets denote an ensemble average. This equation was employed to obtain the total interface tension γ_{tot}. The interfacial tension of hydrocarbon with water $\gamma_{hydrocarbon/water}$ was calculated by **Equation 3.9**. The values of $\gamma_{air/water}$ and $\gamma_{air/hydrocarbon}$ were the surface tensions of pure water and pure hydrocarbons at 300 K, which were adopted from our previous papers.[1,10] A similar approach to calculate the interfacial tension has been reported in literature.[11,12]

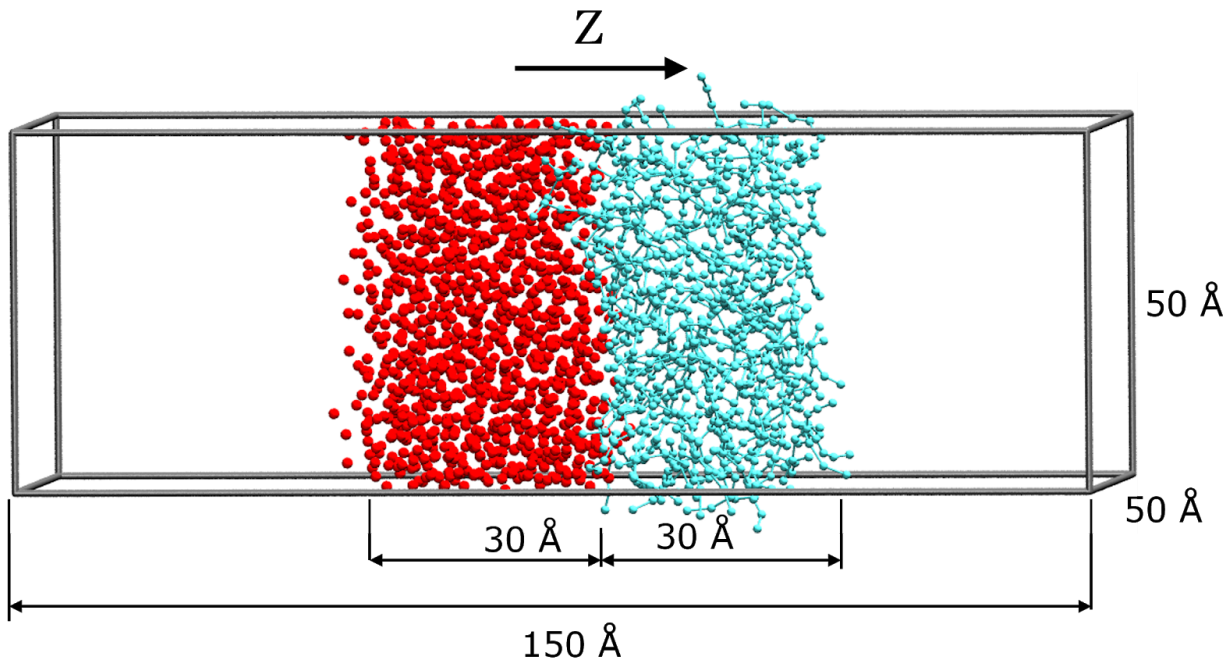

Figure 3.3 Schematic representation of a simulation setup for calculating the interfacial tensions. CG 1-site water and hydrocarbon molecules are shown in red and cyan spheres, respectively.

Table 3.1 The number of CG water and hydrocarbon molecules in the hydrocarbon/water systems to calculate their interfacial tensions.

Hydrocarbon/Water Systems	No. of CG hydrocarbon molecules / No. of CG water molecules
pentane/water	390/1254
hexane/water	343/1254
heptane/water	308/1254
octane/water	278/1254
nonane/water	253/1254
decane/water	231/1254
undecane/water	214/1254
dodecane/water	199/1254
tridecane/water	185/1254

Table 3.1 The number of CG water and hydrocarbon molecules in the hydrocarbon/water systems to calculate their interfacial tensions.(Continued)

tetradecane/water	174/1254
pentadecane/water	163/1254
hexadecane/water	153/1254
heptadecane/water	146/1254

3.3.8 Uncertainty Quantification

To evaluate the performance of the newly developed FF parameters of CG propionic acid and CG PAA models, we performed the uncertainty quantification (UQ) using the bootstrap sampling technique.[13–15] This technique has been used to quantify uncertainties in FF parameters and boundary velocities in MD simulations, and also in machine learning models.[15–20] Traditionally, the standard deviation has been used to quantify the variance of the properties predicted by MD simulations. This can be estimated by performing several independent MD simulations or dividing a simulation trajectory into several blocks. The bootstrapping method can determine the 95 % confidence intervals by resampling a limited number of values. It involves random generation of artificial datasets from randomly selected points using existing datasets. Here, property values were randomly resampled 1000 times from the original 48 data points. The overall mean values and 95 % confidence intervals of the properties were calculated from the mean values of these 1000 bootstrapped resampled sets. In this study, we used the bootstrapping method to analyze the density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of the CG propionic acid model, and the density of the CG PAA model at 300, 360, 420, and 460 K.

3.4 Computational Details

All the all-atom and CG MD simulations reported in this study were performed with the NAMD 2.12 package.[21] The following simulation conditions and parameters were applied in these MD simulations: Temperature and pressure were kept constant by the Langevin thermostat and barostat.[22,23] A switching function was applied for all the 12-6 Lennard-Jones (LJ)

interactions at 9 Å to truncate the van der Waals potential energy smoothly at the cutoff distance of 12 Å. A pair list distance of 15 Å was used to store the neighbors of a given bead. Timestep of 1 fs and 15 fs was employed for integrating the equations of motion in all-atom and CG simulations, respectively, unless specified. Simulation trajectory was analysed to obtain the properties of the hydrocarbon, hydrocarbon/water and polymer systems. Snapshots of configurations from MD simulations were generated with VMD.[24] The initial configuration was created by employing PACKMOL.[25]

3.4.1 Hydrocarbon Bulk Simulations

The all-atom MD simulations with 1000 hydrocarbon molecules were mainly performed to obtain the mapped CG trajectory, which was further used for the comparison between all-atom and CG MD simulations. The CG MD simulations were performed to test and validate the optimized parameters obtained from the PSO method. The total simulation time of the all-atom and CG MD simulations was 5 ns and 75 ns, respectively. The trajectories of the last 1 ns and 45 ns for all-atom and CG MD simulations were used to calculate the properties, respectively. The properties of CG models of hydrocarbons were obtained including self-diffusion coefficient, enthalpy of vaporization, surface tension, expansibility, and isothermal compressibility by analyzing these simulation trajectories.

3.4.2 Water/hydrocarbon Mixture Simulations

One of the key tests to validate the FF parameters between the water and hydrocarbons is to evaluate their ability to reproduce the experimentally observed phase segregation of their mixtures.[26–28] Here, homogeneous simulation boxes with 6000, 10000, 18000 CG 1-site water molecules (12000, 20000, and 36000 all-atom water molecules) and 2000 CG hydrocarbon molecules were employed to study the phase segregation of the hydrocarbon and water systems. NPT simulations were performed with periodic boundary conditions at 300 K and 1 atm.[29] The simulations were performed for 1 μs with a timestep of 15 fs. The simulation trajectories were analyzed to obtain the RDFs, coordination number, and density profiles of water and hydrocarbons before and after the phase segregation.

3.4.3 All-Atom and CG MD Simulations of PAA or PS

All-atom simulations

PAA or PS Bulk Simulations: The CHARMM FF was used to describe the all-atom PAA and PS models.[30] Firstly, a single atactic all-atom PAA or PS chain with 30 monomers (30-mer) was created by using our in-house code, and then relaxed in vacuum at 300 K for 50 ps. This relaxed structure was used for creating 50 PAA chains in a simulation box (60 Å x 60 Å x 60 Å) by using PACKMOL.[25] The MD simulations were performed for 50 ns with a timestep of 1 fs in NPT ensemble. The trajectories were analysed to obtain the bond and angle distributions, and the radial distribution functions (RDFs) between mapped beads to be compared with those from CG MD simulations.

A single PAA or PS chain in solvents: The single atactic all-atom 30-mer PAA or PS chain was solvated in pure water, pure DMF, and mixtures of water and DMF. The compositions of different solvent mixtures used in this study varied from 0 to 100 wt% DMF. The simulations were conducted in NPT ensemble at 300 K with a timestep of 1 fs for 200 ns. The last 150 ns trajectory was analysed to obtain the R_g distribution and the average R_g values. The last 5 ns of the trajectories were analyzed to obtain the RDF between solvents and PAA or PS.

CG MD Simulations of PAA and PS

PAA or PS bulk: A single CG PAA or PS chain with 30-mer was created and relaxed in vacuum for 50 ps. This relaxed chain was used to create two systems, one with 50 CG PAA chains in a simulation box of 60 Å x 60 Å x 60 Å, and the other with 500 CG PAA chains in a simulation box of 120 Å x 120 Å x 120 Å by using PACKMOL.[25] CG MD simulations of the two systems were conducted with a timestep of 5 fs for 100 ns in an NPT ensemble. The temperature range for the CG MD simulations was from 500 K to 280 K with an interval of 20 K. Simulations were performed at 500 K initially, and then the last frame of the equilibrated structure was used as the initial structure for simulations at 480 K. Similar approach was adopted to carry out simulations at other temperatures. The last 30 ns of the total 100 ns of simulation trajectory was divided into three blocks to determine its average density and T_g. In addition to the properties, the bond distributions, and RDFs between CG beads were obtained to study the

structure of the CG PAA models. Similarly, CG simulations of 50 PS 30-mer chains in NPT ensemble were performed to calculate their density and glass transition temperature.

***A single CG PAA or PS chain in CG solvents*:** Initially, to ensure that the LB combining rules are able to capture the interactions between CG models of DMF and water accurately, we performed simulations of their binary mixtures. Specifically, the CG MD simulations of the mixture with different mass fractions of DMF were performed in NPT ensemble at 300 K. The densities of these mixtures were calculated by performing CG MD simulations for 30 ns. To further verify the interactions between the 1-site water and CG DMF model, the Gibbs hydration free energy of CG DMF model in 1-site water was calculated, and compared with reported data. The method of calculating the Gibbs hydration free energy is discussed in **Section 3.3.6** in this chapter.

After the solvent mixture models were validated, the behavior of a single CG PAA or PS 30-mer chain in pure water, DMF, and their mixtures were studied. The compositions of the solvent mixtures are shown in **Table 3.2.** The production runtime for each CG MD simulation was 1 μs, with the last 850 ns of the trajectory analysed to obtain the R_g distributions and its average values. The RDFs during the last 25 ns of CG MD simulations between polymer and solvents were determined to study the local solvent structure around the polymer.

Table 3.2 Solvent compositions in CG MD simulations.

mass fraction of DMF (wt%)	number of CG DMF molecules	number ofCG water molecules
0	0	5,000
2.6	100	7,500
16.8	500	5,000
31.1	1,500	6,750
50.3	2,400	4,800

Table 3.2 Solvent compositions in CG MD simulations.(Continued)

80.2	2,500	1,250
100	5,000	0

CG simulations of mixtures of propionic acid and benzene

The number of CG propionic acid and benzene molecules in the mixtures are shown in **Table 3.3**. The isothermal-isobaric (NPT) ensemble with three-dimensional periodic boundary conditions were employed in these CG MD simulations. The production runtime for the simulations of mixtures was 50 ns and the last 30 ns were divided into three equal blocks to obtain the averaged densities of the mixtures.

Table 3.3 The number of CG propionic acid and benzene molecules in the propionic acid/benzene systems.

Propionic acid (mol%)	Number of CG propionic acid molecules	NUmber of CG benzene molecules
10.6	106	894
20.8	208	792
41.2	412	588
51.3	513	487
61.3	613	387
80.8	808	192
90.4	904	96

CG MD simulations of bottlebrush copolymers of PS-PAA in solvents

To investigate the structure of PS-PAA BBPs in water or DMF, CG simulations were carried out with three PS-PAA BBPs. Each PS-PAA BBP molecule was solvated in a water or DMF box of size of 250 nm x 250 nm x 250 nm. Simulations were carried out in NPT ensemble with a timestep of 8 fs for around 400 ns. The last 300 ns were analyzed to obtain the R_g values

of the side chains and backbones of PS-PAA BBPs and the last 10 ns of the trajectory was analyzed to calculate the RDFs between polymer and solvents.

References

(1) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.

(2) Wang, J.; Tingjun, H. Application of Molecular Dynamics Simulations in Molecular Property Prediction I: Density and Heat of Vaporization. *J. Chem. Theory Comput.* **2011**, *7* (7), 2151–2165.

(3) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(4) Riniker, S.; van Gunsteren, W. F. A Simple, Efficient Polarizable Coarse-Grained Water Model for Molecular Dynamics Simulations. *J. Chem. Phys.* **2011**, *134* (8), 084110.

(5) Eichenberger, A. P.; Huang, W.; Riniker, S.; van Gunsteren, W. F. Supra-Atomic Coarse-Grained GROMOS Force Field for Aliphatic Hydrocarbons in the Liquid Phase. *J. Chem. Theory Comput.* **2015**, *11* (7), 2925–2937.

(6) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128* (144120), 1–13.

(7) Fiorin, G.; Klein, M. L.; Hénin, J. Using Collective Variables to Drive Molecular Dynamics Simulations. *Mol. Phys.* **2013**, *111* (22-23), 3345–3362.

(8) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted to Know but Were Afraid to Ask. *J. Phys. Chem. B* **2015**, *119* (3), 1129–1151.

(9) Gumbart, J. C.; Roux, B.; Chipot, C. Efficient Determination of Protein-Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **2013**, *9* (8). https://doi.org/10.1021/ct400273t.

(10) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153.

(11) Xiao, H.; Zhen, Z.; Sun, H.; Cao, X.; Li, Z.; Song, X.; Cui, X.; Liu, X. Molecular Dynamics Study of the Water/n-Alkane Interface. *Sci. China Chem.* **2010**, *53* (4), 945–949.

(12) Herdes, C.; Ervik, Å.; Mejía, A.; Müller, E. A. Prediction of the Water/oil Interfacial Tension from Molecular Simulations Using the Coarse-Grained SAFT-γ Mie Force Field. *Fluid Phase Equilib.* **2017**. https://doi.org/10.1016/j.fluid.2017.06.016.

(13) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; CRC Press, 1994.

(14) Davison, A. C.; Kuonen, D. An Introduction to the Bootstrap with Applications in R. *Statistical computing & Statistical graphics newsletter* **2002**, *13* (1), 6–11.

(15) Hirsch, R. M.; Archfield, S. A.; De Cicco, L. A. A Bootstrap Method for Estimating Uncertainty of Water Quality Trends. *Environmental Modelling & Software* **2015**, *73*, 148–166.

(16) Reich, Y.; Barai, S. V. Evaluating Machine Learning Models for Engineering Problems.

Artificial Intelligence in Engineering **1999**, *13* (3), 257–272.
(17) Li, Z.; Omidvar, N.; Chin, W. S.; Robb, E.; Morris, A.; Achenie, L.; Xin, H. Machine-Learning Energy Gaps of Porphyrins with Molecular Graph Representations. *J. Phys. Chem. A* **2018**, *122* (18), 4571–4578.
(18) Race, C. P. Quantifying Uncertainty in Molecular Dynamics Simulations of Grain Boundary Migration. *Mol. Simul.* **2015**, *41* (13), 1069–1073.
(19) Patrone, P. N.; Dienstfrey, A. Uncertainty Quantification for Molecular Dynamics. *arXiv:1801.02483v1 [physics.comp-ph]*, 2018.
(20) Rizzi, F.; Najm, H.; Debusschere, B.; Sargsyan, K.; Salloum, M.; Adalsteinsson, H.; Knio, O. Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-Field Parameters. *Multiscale Model. Simul.* **2012**, *10* (4), 1460–1492.
(21) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
(22) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. Langevin Thermostat for Rigid Body Dynamics. *J. Chem. Phys.* **2009**, *130* (23), 234101.
(23) Grønbech-Jensen, N.; Farago, O. Constant Pressure and Temperature Discrete-Time Langevin Molecular Dynamics. *J. Chem. Phys.* **2014**, *141* (19), 194108.
(24) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38, 27–28.
(25) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30* (13), 2157–2164.
(26) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760.
(27) Lobanova, O.; Mejía, A.; Jackson, G.; Müller, E. A. SAFT-γ Force Field for the Simulation of Molecular Fluids 6: Binary and Ternary Mixtures Comprising Water, Carbon Dioxide, and N-Alkanes. *J. Chem. Thermodyn.* **2016**, *93*, 320–336.
(28) Ferrari, E. S.; Burton, R. C.; Davey, R. J.; Gavezzotti, A. Simulation of Phase Separation in Alcohol/water Mixtures Using Two-Body Force Field and Standard Molecular Dynamics. *J. Comput. Chem.* **2006**, *27* (11), 1211–1219.
(29) Allen, M. P.; Tildesley, D. J. Computational Simulation of Liquids. *Clarendon, Oxford* **1987**.
(30) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. CHARMM General Force Field (CGenFF): A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010-3**, *31* (4), 671–690.

CHAPTER 4
DEVELOPMENT OF NEW TRANSFERABLE CG MODELS OF HYDROCARBONS

This work presented in this chapter is reported from [An, Y., Bejagam, K. K., Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons, J. Phys. Chem. B, 2018, 122 (28), 7143-7153], with the permission of AIP Publishing.

Abstract: We have utilized an approach that integrates molecular dynamics (MD) simulations with particle swarm optimization (PSO) to accelerate the development of coarse-grained (CG) models of hydrocarbons. Specifically, we have developed new transferable CG beads, which can be used to model the hydrocarbons (C5 to C17) and reproduce their experimental properties with good accuracy. Firstly, the PSO method was used to develop the CG beads of the decane model represented with 2:1 (2-2-2-2-2) mapping scheme. This was followed by the development of the nonane model described with hybrid 2-2-3-2, and 3:1 (3-3-3) mapping schemes. The force-field (FF) parameters for these three CG models were optimized to reproduce four experimentally observed properties including density, enthalpy of vaporization, surface tension, and self-diffusion coefficient at 300 K. The CG MD simulations conducted with these new CG models of decane and nonane, at different timesteps, for various system sizes, and at a range of different temperatures, were able to predict their density, enthalpy of vaporization, surface tension, self-diffusion coefficient, expansibility, and isothermal compressibility with a good accuracy. Moreover, comparison of structural features obtained from the CG MD simulations and the CG beads of mapped all-atom (AA) trajectories of decane and nonane showed very good agreement. To test the chemical transferability of these models, we have constructed the models for hydrocarbons ranging from pentane to heptadecane, by using different combination of the CG beads of decane and nonane. The properties of pentane to heptadecane predicted by these new CG models showed an excellent agreement with the experimental data.

4.1 Introduction

CG hydrocarbon models are the basic building blocks for polymer backbones, hydrophobic tails of lipids and surfactants. The most widely used CG hydrocarbon models are MARTINI models which employ a 4:1 mapping scheme to represent hydrocarbons.[1] These CG hydrocarbon models could reproduce their experimental densities at 300 K with errors of 5 - 10

%. However, the self-diffusion coefficients of these hydrocarbon models were, 2-3 times slower than the experimental values. Eichenberger et al. found that the self-diffusion coefficients of the CG hydrocarbon models were affected by the mapping schemes. With a large mapping scheme, i.e. 4:1, the self-diffusion coefficient could be decreased compared with 3:1 and 2:1 mapping schemes. Besides, Cao and Sun reported the CG *n*-alkane models represented by 3+2 mapping scheme.[2] Specifically, the n-alkane models consisted of three types of beads: C2E (CH3CH2-), C3E (CH3CH2CH2-), and C3M (−CH2CH2CH2-). They developed a CG n-alkane FF which can reproduce the vapor liquid equilibrium (VLE) coexistence curve, enthalpy of vaporization, surface tension and P-V-T equation of state with good accuracy. Avendaño *et. al.* developed the SAFT-γ FF for molecules formed with CG segments interaction *via* a Mie potential.[3] They studied VLE and surface tension of the decane and dodecane models. The VLE for both the decane and dodecane models was in good agreement with experiment data. The surface tension of the two models also showed good agreement with the experimental values.

The development of these existing CG models is either a top-down approach (using experimental properties as target values) or a bottom-up approach (atomistic structures as targets). This may lead to either atomistic structures, or experimental properties may not be maintained in the CG models. Besides, the FF optimization of these CG models was performed by trails and errors, which was time consuming. In the present study, we have used PSO to accelerate the development of transferable hydrocarbon models ranging from pentane to heptadecane. First, we develop a CG decane model with 2:1 mapping scheme (2-2-2-2-2), and two models for nonane with a hybrid mapping scheme of 2-2-3-2, and with 3:1 mapping scheme (3-3-3). The FF parameters for the decane and nonane models were optimized to reproduce the experimental properties observed at 300 K by integrating the MD simulations with PSO. The MD simulations were further performed at various temperatures to test the transferability of these models at various temperatures. To test the chemical transferability of the CG beads of decane and both nonane models, they were employed to perform CG MD simulations for hydrocarbons with different chain lengths (pentane to heptadecane).

4.2 Model Development

Here we have used a 2:1 mapping scheme to represent the decane model (2-2-2-2-2) (**Figure 4.1 (a)**). The nonane model was represented by a hybrid mapping scheme of 2-2-3-2 (**Figure 4.1 (b)**), and 3:1 (3-3-3) mapping scheme (**Figure 4.1 (c)**). For both the 2:1 and 3:1

mapping schemes two types of beads were developed based on their location in the chain i.e. beads located at the end and in the middle of a CG molecule. For example, C2E and C2M are the two bead types with 2:1 mapping scheme, located at the end and middle of a decane molecule, respectively (see **Figure 4.1 (a)**). Similarly, as shown in **Figure 4.1 (c)** C3E and C3M bead represent the end and middle beads, respectively. In the case of the hybrid nonane model, two C2E, one C2M and one C3M beads were utilized as shown in **Figure 4.1 (b)**. **Table A1** of Appendix A lists the CG bead name, all-atom groups encompassed in a CG bead, and their corresponding mass.

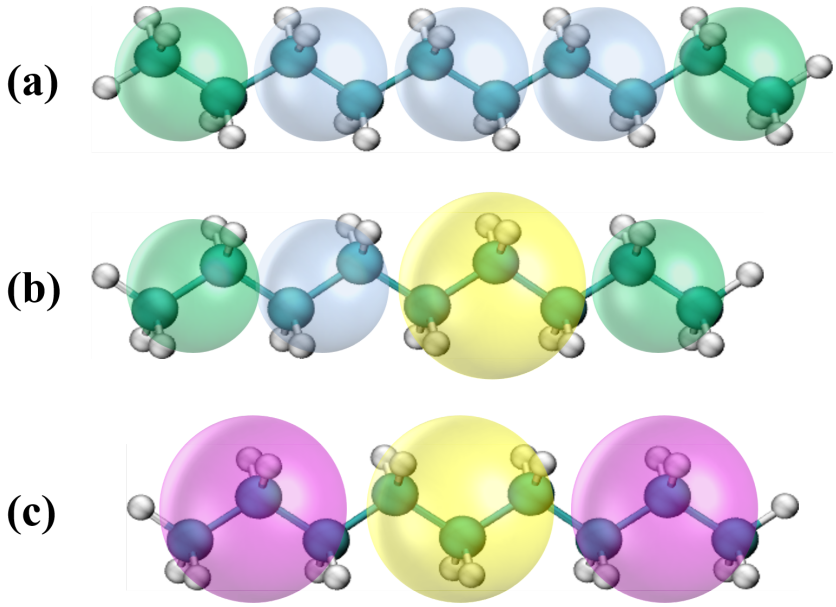

Figure 4.1 all-atom and mapped CG beads for decane, and nonane models with different mapping schemes: **(a)** CG decane model with 2-2-2-2-2 mapping scheme. The green and gray transparent spheres represent the C2E (end bead), and C2M (middle beads), respectively; **(b)** CG nonane model with hybrid, 2-2-3-2, mapping scheme. C3M bead is represented by a yellow transparent sphere, and **(c)** the CG nonane model with 3-3-3 mapping scheme. The magenta sphere represents the C3E bead.

The CHARMM type FF equation represented by harmonic bonds and angles were used to capture the bonded and non-bonded interactions in a molecules (see **Equation 4.1**).[4] The partial charge on all the four types of beads, namely, C2E, C2M, C3M, and C3E was set to zero. The non-bonded interactions between these beads were captured using the 12-6 LJ potentials.

Lorentz-Berthelot (LB) mixing rule was used to define the nonbonded interactions between different types of CG hydrocarbon beads.[5] Atoms separated by three bonds (1 - 4 pairs) in a hydrocarbon interact through non-bonded potential.

$$E_{pot} = K_b(b - b_0)^2 + K_\theta(\theta - \theta_0)^2 + 4\varepsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^6] \quad \textbf{…….Equation 4.1}$$

In the above equation: K_b is the bond force constant and b_0 is the equilibrium bond length, K_θ is the angle force constant and θ_0 is the equilibrium angle. ε is the depth of the potential well, σ is the finite distance at which the inter-particle potential is zero, r_{ij} is the distance between two atoms.

To expedite the process of CG parameterization, PSO method was used.[6,7] PSO is a population-based search approach which can often find the good solution efficiently and effectively.[7] This method has been successfully applied in various research fields, including the development of interatomic potentials for all-atom and CG models of materials.[4,8–10] See **Section 2.4** in **Chapter 2** for more details on the PSO method and its integration with the MD simulations. **Table 4.1** lists the optimized parameters for the different beads representing the decane and nonane.

Parameterization of the CG Decane Model. Based on the mapping scheme shown in **Figure 4.1 (a)** and **Equation 4.1** for CG decane model, the following parameters were optimized: K_b [C2E-C2M], b_0 [C2E-C2M], K_b [C2M-C2M], b_0 [C2M-C2M], K_θ [C2E-C2M-C2M], θ_0 [C2E-C2M-C2M], K_θ [C2M-C2M-C2M], θ_0 [C2M-C2M-C2M], ε [C2E], ε [C2M], σ [C2E], σ [C2M]. Thus the bonded and non-bonded parameters constitute twelve variables to be optimized for the CG decane model. The experimental values of density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of decane at 300 K were used as target properties for the optimization of these twelve parameters. The number of particles N is 80 during PSO. The value of σ for C2E and C2M beads were kept the same during optimization. The range for optimization of each input parameter during the PSO algorithm was chosen based on the all-atom mapped trajectories (see **Table A2** in Appendix A). For example, the range for the equilibrium bond-length (b_0) was selected near the mean value of the bond distribution obtained from the all-atom mapped trajectory.

Parameterization of the CG Hybrid Nonane Model (2-2-3-2). This was the second model that was optimized in this chapter and it is composed of a C2M, a C3M, and two C2E beads. The parameters of C2E and C2M beads were obtained from the decane model. For this hybrid nonane model the following ten parameters were optimized to reproduce the four experimental properties of nonane, namely, density, enthalpy of vaporization, self-diffusion coefficient, and surface tension at 300 K: K_b [C2M-C3M], b_0 [C2M-C3M], K_b [C2E-C3M], b_0 [C2E-C3M], K_θ [C2E-C2M-C3M], θ_0 [C2E-C2M-C3M], K_θ [C2M-C3M-C2E], θ_0 [C2M-C3M-C2E], ε [C3M] , and σ [C3M].

Parameterization of the CG Nonane Model (3-3-3). This nonane model has two C3E beads, and one C3M bead and the parameters for the C3M bead were obtained from the hybrid nonane model. Thus, this nonane model had a total of six adjustable parameters (see Appendix A: **Table A2**): K_b [C3E-C3M], b_0 [C3E-C3M], K_θ [C3E-C3M-C3E], θ_0 [C3E-C3M-C3E], ε [C3E], and σ [C3E]. The value of σ for the C3E bead was kept the same as that of the C3M bead of the hybrid nonane model. Similar to the decane and hybrid nonane model, the PSO algorithm was used to optimize the parameters to reproduce the four experimental properties of the nonane.

Table 4.1 Optimized CG FF parameters for new CG models. Units: K_b - (kcal/mol/Å^2), b_0 - Å, K_θ - kcal/ mol/radian^2, θ_0 - °, ε - kcal/mol, and σ - Å.

Bond parameters	$\boldsymbol{K_b}$	$\boldsymbol{b_0}$
C2E-C2M	35.293	2.561
C2M-C2M	36.609	2.520

Table 4.1 Optimized CG FF parameters for new CG models. Units: K_b - (kcal/mol/Å^2), b_0 - Å, K_θ - kcal/ mol/radian^2, θ_0 - °, ε - kcal/mol, and σ - Å. (Continued)

C2M-C3M	38.537	2.914
C2E-C3M	34.170	2.958
C3E-C3M	37.695	3.561
Angle parameters	$\boldsymbol{K_\theta}$	$\boldsymbol{\theta_0}$
C2E-C2M-C2M	3.886	147.766

C2M-C2M-C2M	3.460	144.226
C2E-C2M-C3M	3.262	144.669
C2E-C3M-C2M	3.377	149.893
C3E-C3M-C3E	3.417	144.832
LJ parameters	ε	σ
C2E	0.3710	4.3374
C2M	0.3420	4.3374
C3E	0.5927	4.6344
C3M	0.5545	4.6344

4.3 Results and Discussion

4.3.1 Properties of the CG Decane (2-2-2-2-2), Hybrid Nonane (2-2-3-2), and Nonane (3-3-3) Models at 300 K. The optimized FF parameters for the CG decane, hybrid nonane, and nonane models were obtained by employing the PSO to reproduce the density, enthalpy of vaporization, surface tension, and self-diffusion coefficient at 300 K. During the parameterization of FF, simulations of 100 CG molecules were carried out for 3 ns with a timestep of 15 fs. To test the robustness for these new CG models, we have conducted simulations with 1000 CG molecules with three timesteps: 10 fs, 15 fs, and 20 fs at 300 K for 75 ns. All the four properties calculated for the three models are tabulated in **Table 4.2**. It can be seen that the density of the CG decane model (2-2-2-2-2) calculated from the simulations with timestep of 15 fs is 0.723 g/cm^3, which is in excellent agreement with the experimental value of 0.726 g/cm^3. The enthalpy of vaporization, surface tension, and self-diffusion coefficient of the CG decane model with timestep of 15 fs were 12.30 kcal/mol, 24.00 mN/m, and 1.52 × 10^{-9} m^2/s, respectively. These values are comparable with the corresponding experimentally measured values of 12.52 kcal/mol, 23.19 mN/m, and 1.55 × 10^{-9} m^2/s, respectively. The density of the hybrid CG nonane model (2-2-3-2) with ~15 fs timestep was 0.712 g/cm^3, while it was 0.709 g/cm^3 for CG nonane (3-3-3) model. These values are fairly close to the experimental value of 0.714 g/cm^3 at 298.15 K. The hybrid CG nonane model (2-2-3-2) also predicts the enthalpy of vaporization, surface tension, and self-diffusion coefficient within 3.7 %, 1.8 %, and 4.7 % of the corresponding experimental

values at 298.15 K. For a timestep of either 10 fs or 20 fs, the density, enthalpy of vaporization, surface tension and self-diffusion coefficient predicted by decane CG model were within 0.6 %, 2.3 %, 2.3 % and 2.0 % of the corresponding experimental values at 300 K, respectively. Similarly, for hybrid nonane (2-2-3-2) these properties were with 0.6 %, 4.0 %, 2.1 % and 5.3 % of corresponding experimental values, respectively. The nonane model with 3-3-3 mapping scheme had the density, enthalpy of vaporization, surface tension and self-diffusion coefficient within 0.8 %, 8.5 %, 5.1 % and 15.9 % of corresponding experimental values at 300 K, respectively. This suggests that the new CG models can accurately predict the properties of these hydrocarbons in the CG MD simulations performed with different timesteps. Moreover, the simulations performed with different system sizes could also reproduce the experimental properties with a reasonable accuracy. These results are shown in **Table A3** of Appendix A.

To estimate the computational efficiency of all three CG models, we have performed CG and all-atom MD simulations with 1000 molecules of these hydrocarbons with 15 fs and 1 fs timesteps, respectively, at 300 K. The speed-up factors for three CG models, namely, decane (2-2-2-2-2), hybrid nonane (2-2-3-2), and nonane (3-3-3) were determined to be of ~250, ~400, and ~620 in comparison to all-atom models, respectively. These simulations were carried out on Intel Xeon 2.4 GHz machines.

Table 4.2 Density, enthalpy of vaporization, surface tension and self-diffusion coefficient of decane and nonane at 300 K with different timesteps. Units: density - g/cm^3, enthalpy of vaporization - kcal/mol, surface tension - mN/m, self-diffusion coefficient- $\times 10^{-9}$ m^2/s.

Model	Timestep (fs)	Density	Enthalpy of vaporization	Surface tension	Self-diffusion coefficient
Decane (2-2-2-2-2)	10	0.724±0.00	12.30±0.03	23.64±0.11	1.53±0.02
	15	0.723±0.00	12.30±0.01	24.00±0.06	1.52±0.02
	20	0.722±0.00	12.23±0.03	23.72±0.10	1.56±0.01
Experiment values (decane)*	-	0.726	12.52	23.19	1.55
Nonane (2-2-3-2)	10	0.713±0.00	10.7±0.06	22.96±0.04	1.74±0.02
	15	0.712±0.00	10.70±0.28	22.45±0.23	1.78±0.07

	20	0.710±0.00	10.67±0.02	22.14±0.07	1.79±0.05
Nonane (3-3-3)	10	0.710±0.00	10.21±0.02	21.35±0.11	1.43±0.04
	15	0.709±0.00	10.25±0.28	22.41±0.02	1.46±0.04
	20	0.708±0.00	10.17±0.01	22.36±0.55	1.47±0.04
Experiment values (nonane)*	-	0.714	11.11	22.49	1.70

* measured at 298.15 K, ref [11–18]

To further test the ability of these new models in predicting the properties for which they were not optimized, we have calculated their expansibility, and compressibility (see **Section 3.3** of **Chapter 2**). The expansibility and compressibility of the CG models at 300 K are listed in **Table 4.3**. For the CG decane model, the expansibility and compressibility were determined to be 0.9×10^{-3} K^{-1}, and 11.0×10^{-5} bar^{-1}, respectively. These values are in good agreement with the experimental value of 1.0×10^{-3} K^{-1} and 10.9×10^{-5} bar^{-1}, respectively, reported at 298.15 K.[12] Similarly, the predicted values of expansibility, and compressibility by both the CG models of nonane were within 10.9 % of the experimental values.

Table 4.3 Expansibility and isothermal compressibility of CG decane and nonane models at 300 K. Experimental data is in parentheses. Units: expansibility - $\times 10^{-3}$ K^{-1}, compressibility - $\times 10^{-5}$ bar^{-1}.

Alkanes	Expansibility	Compressibility
Decane (2-2-2-2-2)	0.90 (1.0)	11.0 (10.9)
Nonane (2-2-3-2)	0.98 (1.1)	12.2 (11.8)
Nonane (3-3-3)	0.99 (1.1)	13.0 (11.8)

4.3.2 Structure of the CG Decane and Nonane Models. The bond and angle distribution, end-to-end distance, and RDF were determined from CG MD simulation trajectories of the decane and nonane models. Analysis methods are reported in **Section 3.2** of **Chapter 2**. The comparison between the data obtained from CG MD simulations and mapped all-atom MD

trajectories is shown in **Figure 4.2** and **Figure 4.3**. As can be seen from **Figure 4.2 (a)**, the all-atom mapped trajectories of all the bonds present in the three models showed multiple peaks, suggesting the presence of bonds with the lengths corresponding to those peaks in the all-atom mapped structure. The mapped C2E-C2M bond distribution (black dashed line in **Figure 4.2 (a)**) from all-atom simulations of decane and nonane shows two peaks at 2.3 Å and 2.6 Å. Similarly the mapped C3E-C3M bond distribution (pink dashed line **Figure 4.2 (a)**) also showed two peaks at 3.6 Å and 3.9 Å. Similarly, multiple peaks are reported for all-atom mapped trajectories with both 2:1 and 3:1 mapping scheme for hydrocarbons, which are attributed to the coexistence of trans-gauche conformation.[19] The peaks observed at smaller distances can be attributed to the gauche-conformations. On the other hand the peaks observed at large distances represent the trans-conformations. In our CG model, we have used simple harmonic potentials to describe the bond between CG beads and thus, bimodal distributions cannot be reproduced by using such potentials.[20] However, the bond-length distributions of the CG bonds present in all three models show a large overlap between the CG and all-atom mapped trajectories. Similar to the bond distributions, the angle distributions from all-atom mapped trajectory of decane and nonane also show multiple peaks (**Figure 4.2 (b)**). The peaks are observed in the range of 120° to 170°. The peak heights suggest the presence of more number of angles in the range of 135° to 170°. We have used this data to estimate the range for the angle calibration in the PSO optimization (see Appendix A **Table A2**). The optimized angle values obtained during the PSO optimization for all three models were around 145° (**Table 4.1**).

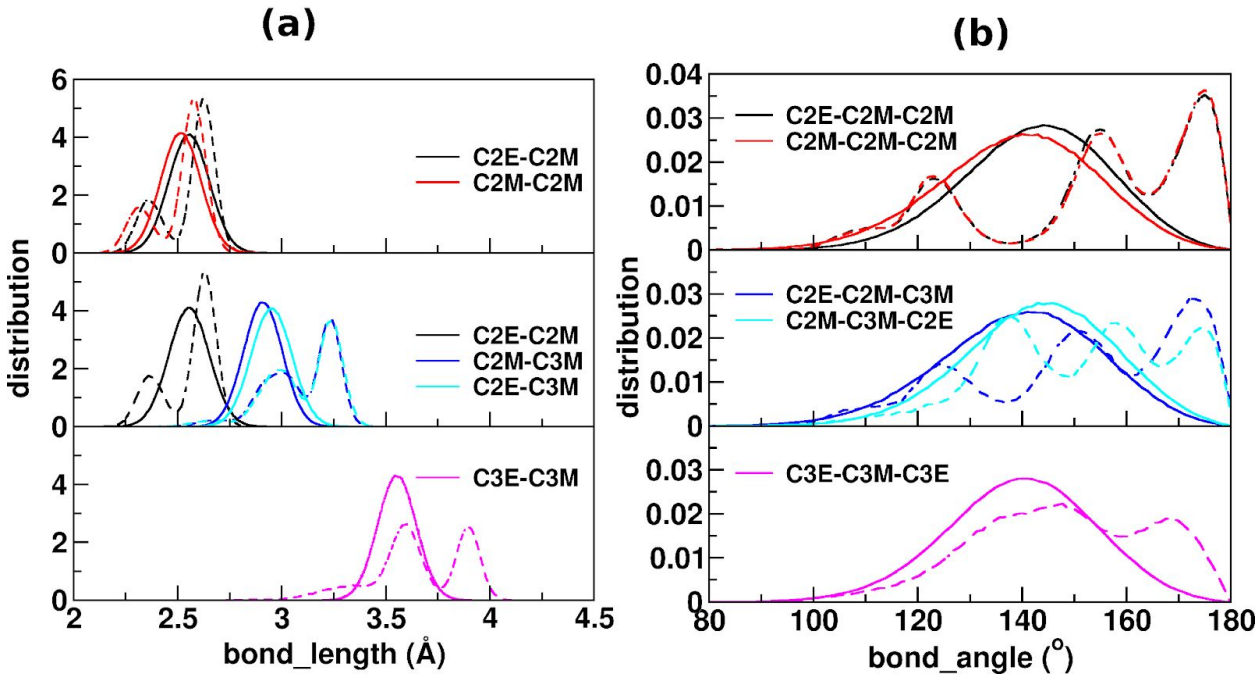

Figure 4.2 Distributions of **(a)** bond-length and **(b)** bond-angle obtained from CG MD simulations of decane and nonane models (solid lines), and their corresponding mapped all-atom trajectories (dashed lines). In both **(a)** and **(b)**, the top, middle, and bottom panels represent the data for decane (2-2-2-2-2), hybrid nonane (2-2-3-2), and nonane (3-3-3) models, respectively. Simulations of 1000 molecules were performed at 300 K.

The end-to-end distance for both the mapped all-atom, and CG trajectories for all three models is shown in **Figure 4.3 (a)**. Overall, the distance predicted by CG models is shorter than those of all-atom mapped trajectory. For example, for decane the all-atom mapped trajectory showed a peak at 9.4 Å, which is 0.3 Å higher than that of the CG model peak at 9.1 Å. Similarly for hybrid nonane (2-2-3-2) and nonane (3-3-3) models, the CG models had the average end-to-end distance of 7.7 Å and 6.8 Å, respectively, which is 0.5 Å and 0.3 Å smaller than the all-atom mapped trajectory for the nonane. The peaks become narrower when the bead size increases from 2:1 to 3:1 mapping scheme. Similar observations are reported by Eichenberger *et. al.* in their study.[19] The RDF for both the all-atom mapped trajectories and CG models are shown in **Figure 4.3 (b)** for all three hydrocarbons. The RDF for all the beads present in the three CG models showed an excellent agreement with the RDF of all-atom mapped trajectory. As expected with increase in the bead size from 2:1 to 3:1 mapping scheme the position of the first peak in the RDF shifted towards higher values. The first peak for C3M-C3M beads with 3:1 mapping

scheme was observed at 5.2 Å both for CG model and all-atom mapped trajectory. However, the first peak for the C2M-C2M beads with 2:1 mapping scheme was observed at 4.9 Å both for CG model and all-atom mapped trajectory.

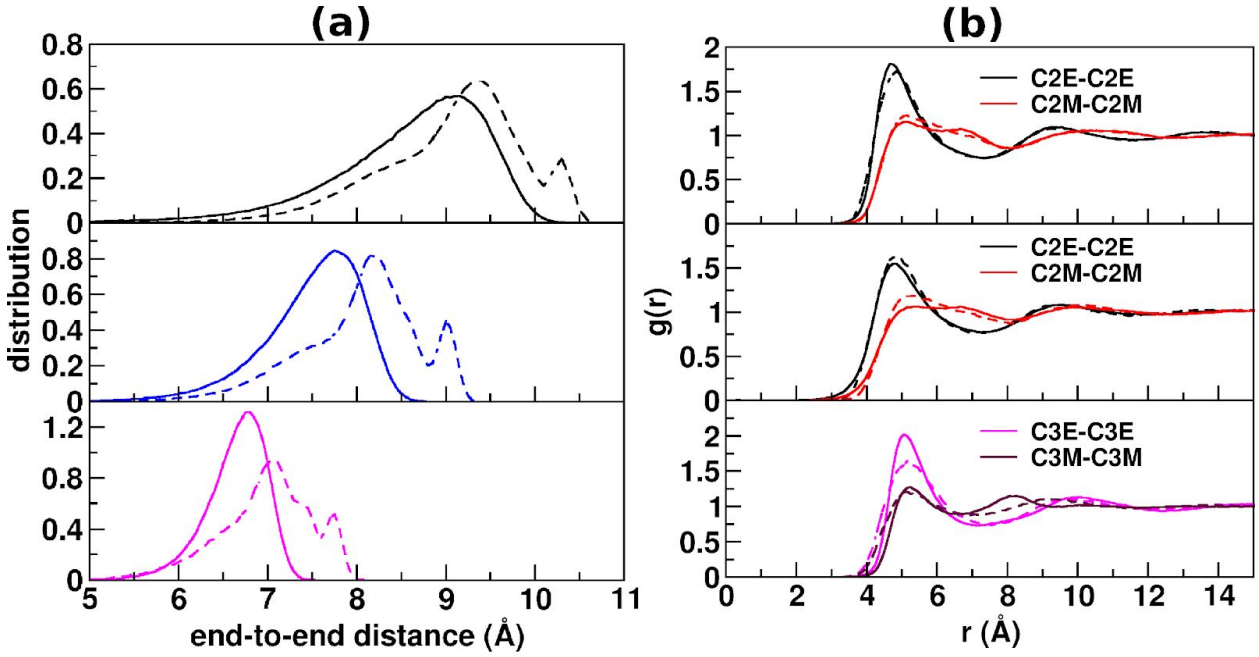

Figure 4.3 (a) End-to-end distance distribution and **(b)** RDF from CG MD simulations of decane and nonane models (solid lines) and their corresponding mapped all-atom trajectories (dashed lines). In both **(a)** and **(b)**, the top, middle, and bottom panels represent the data for decane (2-2-2-2-2), hybrid nonane (2-2-3-2), and nonane (3-3-3) models, respectively. Simulations of 1000 molecules were performed at 300 K.

4.3.3 Properties of the CG Decane (2-2-2-2-2), Hybrid Nonane (2-2-3-2), and Nonane (3-3-3) Models at Different Temperatures. We now expose these new CG models to various temperatures in the range of 270 K to 350 K to investigate their capacity to predict physical, dynamical, and thermodynamic properties at simulated temperatures. This temperature range was selected because both decane and nonane are in liquid state at these temperatures at 1 atm pressure. Note, the FF parameters of these models were optimized to reproduce the properties of the hydrocarbons only at 300 K. Where possible we compare our results with the experimental properties obtained from references [13–15,18,21,22]. Results from these simulations and comparison with experimental data are shown in **Figure 4.4** for the decane model, and in **Figure 4.5** for both the hybrid nonane (2-2-3-2), and nonane (3-3-3) models.

As can be seen from **Figure 4.4 (a)** the density of the decane model is in good agreement with the experimental values in simulated temperature range (maximum error below 0.5 %). At 310 K the experimental value of density is 0.719 g/cm^3, and the value predicted by the new CG model of decane is 0.717 g/cm^3. For all the CG models, similar to experiments, the density decreases with increase in temperature. For example, when the temperature is increased from 270 K to 350 K the density of CG decane model decreases by 7.0 %. The regression-fitted slope of -6.7 × 10^{-4} g/(cm^3 K) and -7.5 × 10^{-4} g/(cm^3 K) were obtained for the CG model and experimental data, suggesting the error of 10.7 %. **Figure 4.4 (b)** and **(c)** show enthalpy of vaporization and the self-diffusion coefficient for the decane model, respectively. Experimental data was taken from ref [13,14,23]. The enthalpy of vaporization decreases with increase in temperature. The experimental value at 298.15 K is 12.52 kcal/mol, which decreases to 11.76 kcal/mol at 334 K. The values predicted by the decane model at 300 K and 350 K are 12.30 kcal/mol and 11.68 kcal/mol, respectively. As shown in **Figure 4.4 (c)**, the self-diffusion coefficient predicted by this new model of decane at 330 K was 2.11 × 10^{-9} m^2/s, within 6.6 % of the experimental value of 2.26 × 10^{-9} m^2/s (332 K). At 270 K and 350 K, the self-diffusion coefficients showed 63.4 % and 25.5 % deviation from the corresponding experimental data, respectively. The surface tension is shown in **Figure 4.4 (d)** for the decane model. The value of surface tension at 310 K was 22.86 mN/m, in comparison to that of experimental value of 22.45 mN/m at 313 K. The surface tension for CG models decreases from 26.65 mN/m to 20.27 mN/m when the temperature is raised from 270 K to 350 K. It has been reported that the surface tension of liquids decreases linearly with increase in temperature.[22] The regression-fitted slope of -0.082

mN/(m K) was within 17 % of the slope of -0.099 mN/(m K) obtained by fitting the experimental data with temperature.

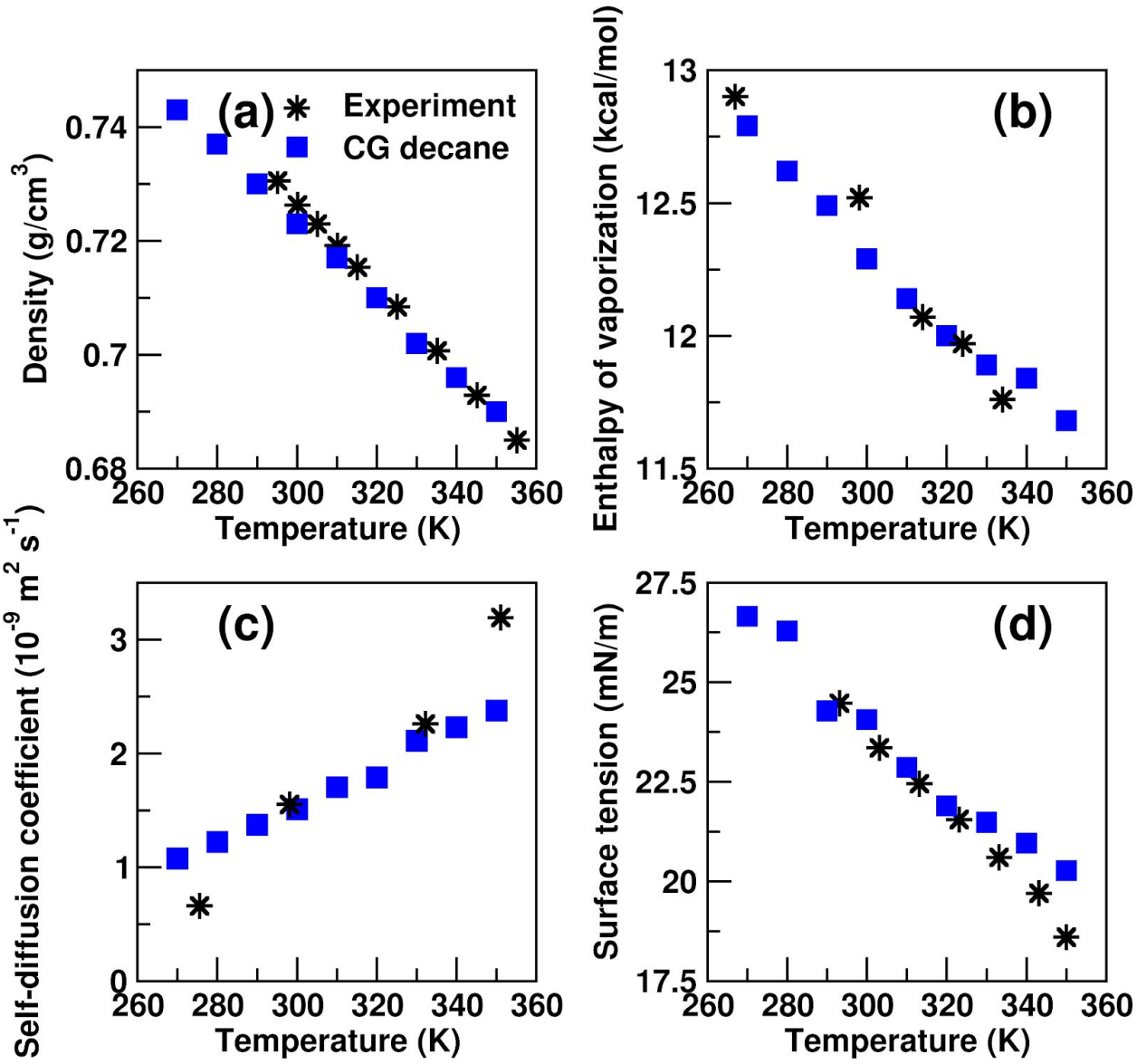

Figure 4.4 Density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of the CG decane model (2-2-2-2-2) at different temperatures.

The comparison of the density of hybrid nonane (2-2-3-2) and nonane (3-3-3) models with the experimental data reported in references [13,14,23] is shown in **Figure 4.5 (a)**. At 280 K the densities for hybrid nonane (2-2-3-2) and nonane (3-3-3) were 0.726 g/cm^3 and 0.723 g/cm^3, respectively. The values are close to the experimental density of 0.726 g/cm^3 at 283 K. As the

temperature was increased from 280 K to 350 K the density for both the hybrid nonane (2-2-3-2), and nonane (3-3-3) models decreased to 0.677 g/cm^3, and 0.674 g/cm^3, respectively. The regression-fitted slope of -6.96 × 10^{-4} g/(cm^3 K), -6.96 × 10^{-4} g/(cm^3 K), and -8.01 × 10^{-4} g/(cm^3 K) were obtained for the CG model of hybrid nonane model, nonane model (3-3-3), and experimental data. This suggests that these new CG models can predict the experimental density at different temperatures with a very good accuracy.

The enthalpy of vaporization at 270 K for hybrid nonane (2-2-3-2) and nonane (3-3-3) models were 11.10 kcal/mol and 10.57 kcal/mol, respectively (**Figure 4.5 (b)**). When the temperature was raised to 350 K these values for hybrid nonane (2-2-3-2) and nonane (3-3-3) models decreased to 10.15 kcal/mol and 9.64 kcal/mol, respectively. Compared with experimental enthalpy of vaporization at the temperature range of 300 - 350 K, both models showed lower values while the hybrid nonane model (2-2-3-2) performed slightly better than the model (3-3-3).

The self-diffusion coefficients for both the nonane CG models are shown in **Figure 4.5 (c)**. It can be seen that the diffusion coefficient at 270 K for the hybrid nonane model is 1.3 × 10^{-9} m^2/s and for the nonane model (3-3-3) is 1.0 × 10^{-9} m^2/s as compared to the experimental value of 0.9 × 10^{-9} m^2/s. Compared with the nonane (3-3-3) model, the hybrid nonane model (2-2-3-2) shows slightly higher self-diffusion coefficient, which can be attributed to the presence of beads with 2:1 mapping scheme. Eichenberger *et al.* have also reported similar observation of the increase in the self-diffusion coefficient with decrease in the bead size.[19] For the self-diffusion coefficients at 270 K, 280 K and 300 K, the nonane model (3-3-3) showed good performance, which predicted the self-diffusion coefficient within 15% compared with the corresponding experimental values. Experimental self-diffusion coefficient was taken from ref [14]. At 350 K the error between the self-diffusion coefficient predicted by the hybrid nonane model (2-2-3-2) and nonane model (3-3-3) decreased to 15.2 % as compared to 23.1 % at 270 K, with the diffusion coefficient for the hybrid (2-2-3-2) model being higher as compared to the nonane (3-3-3) model.

Figure 4.5 (d) reports the surface tension for both the nonane CG models and experimental data at different temperatures. At 320 K, the value of surface tension obtained for the hybrid nonane (2-2-3-2), nonane (3-3-3), and from the experiments were 20.76 mN/m, 20.74 mN/m, and 20.47 mN/m, respectively. Both models showed the decrease in the surface tension with increase in temperature. Although the nonane model has all three beads with 3:1 mapping

scheme as compared to the hybrid nonane model with 2-2-3-2 mapped beads, the density and surface tension showed no significant dependence on the mapping scheme or the bead size.

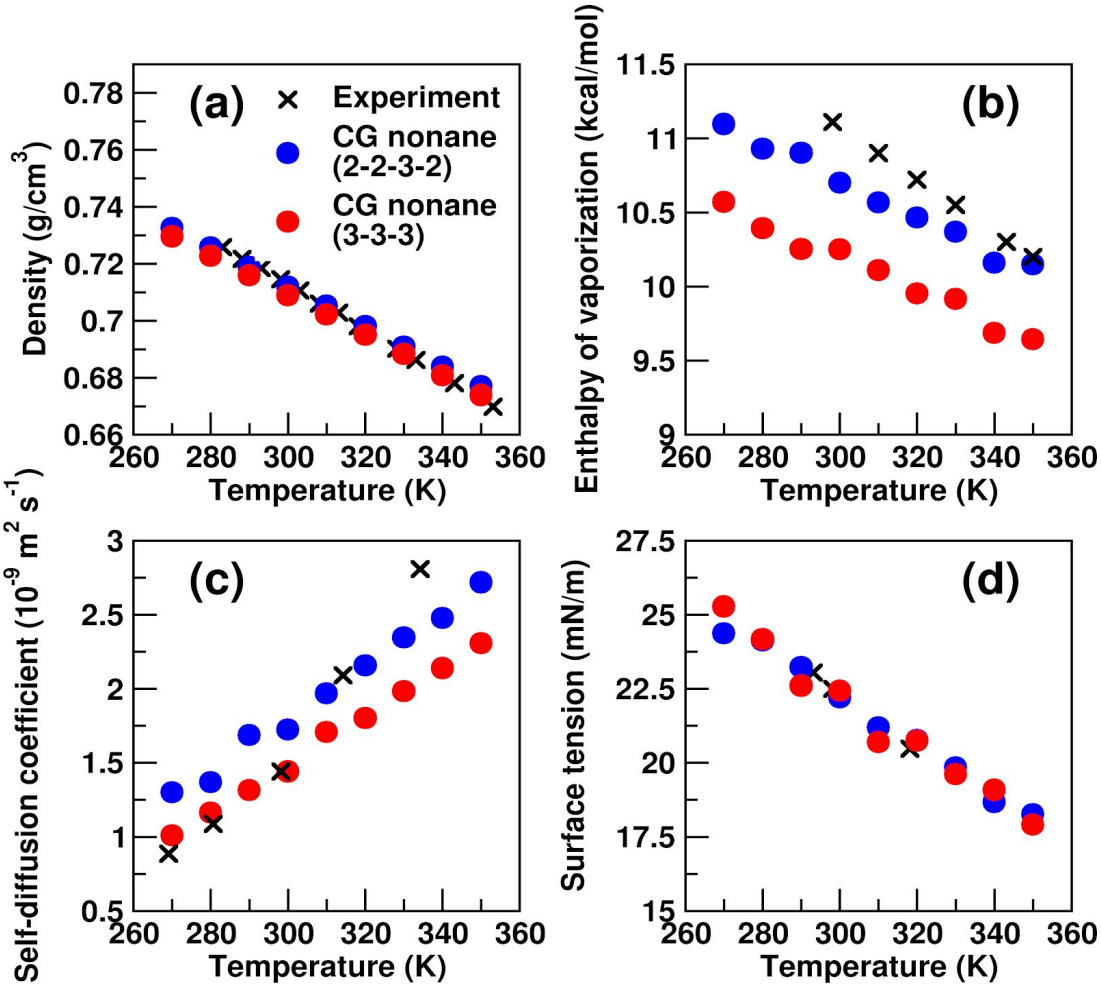

Figure 4.5 Density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of the hybrid nonane (2-2-3-2) and nonane CG models (3-3-3) at different temperatures.

4.3.4 CG n-Alkane Models with Different Chain Lengths. To test the chemical transferability of these new CG models, we have conducted CG MD simulations of various hydrocarbons from pentane to heptadecane by using the beads developed for the decane and nonane models. Bonded parameters for all these hydrocarbons can be found in **Table A4** of Appendix A. The CG MD simulations were performed with systems with 1000 beads and 15 fs timestep at 300 K for these hydrocarbons. More details of the simulations could be found in **Section 3.4** of **Chapter 2**. The simulation trajectories were analyzed to evaluate the ability of these models in predicting the properties of hydrocarbons. Experimental data to compare with our CG MD simulations was obtained from these references [12–18,21,24–29].

All the alkanes from hexane to heptadecane using the combination of 2:1 and 3:1 mapping schemes can be represented by more than one model. For example, a decane molecule can be represented by following mapping schemes, 2-2-2-2-2, 3-3-2-2, 2-3-3-2, 2-3-2-3 etc. However, here we only consider up to two mapping schemes, which were selected randomly, for each alkane even if it can be represented by more than two CG models. The CG beads developed in **Table 4.1** were utilized to represent these structures. For simplicity, we refer the two mapping schemes as mapping scheme **1** and mapping scheme **2** and the details of beads used in them are shown in **Table 4.4**. Representative examples of hexane and hexadecane with different mapping schemes are shown in **Figure A1** in Appendix A. In **Table 4.4**, it can be seen that the density and enthalpy of vaporization for each CG hydrocarbon model was in good agreement with corresponding experimental values. The densities and enthalpy of vaporization calculated for all the CG models were within 3 % and 9 % of the experimental values at 300 K, respectively. For example, the CG pentane model had the density of 0.622 g/cm^3 and enthalpy of vaporization of 5.79 kcal/mol, which are within 0.2 % and 8.4 % of the experimental values, respectively. The CG undecane model with mapping scheme **1** predicted the values as 0.739 g/cm^3 (error 0.3 %), and 13.32 kcal/mol (error 1.2 %), respectively. Similarly, for heptadecane, the CG model with mapping scheme **2** showed the error of 0.1 % for density and 3.7 % for enthalpy of vaporization at 300 K.

Table 4.4 Density, enthalpy of vaporization of CG *n*-alkanes at 300 K and 1 atm. Experimental data measured at 298 K is in parentheses. Units: density - g/cm^3, enthalpy of vaporization - kcal/mol.

Hydrocarbons	Mapping scheme No.	Mapping schemes	Density	Enthalpy of vaporization
pentane	1	2-3	0.622±0.00 (0.621)[a]	5.79±0.02 (6.32)
hexane	1	2-2-2	0.634±0.00 (0.660)	6.99±0.02 (7.55)
	2	3-3	0.658±0.00	6.84±0.02
heptane	1	2-3-2	0.670±0.00 (0.679)	8.05±0.01 (8.74)
octane	1	2-2-2-2	0.689±0.00 (0.699)	9.65±0.01 (9.49)
	2	2-3-3	0.692±0.00	9.15±0.06
nonane	1	3-3-3	0.709±0.00 (0.714)[b]	10.58±0.28 (11.11)
	2	2-2-3-2	0.712±0.00	10.37±0.28
decane	1	2-2-2-2-2	0.723±0.00 (0.726)	12.30±0.01 (12.52)
	2	2-2-3-3	0.727±0.00	11.81±0.02
undecane	1	2-2-2-3-2	0.739±0.00 (0.737)[c]	13.32±0.02 (13.48)
	2	3-3-3-2	0.724±0.00	12.61±0.08
dodecane	1	2-2-2-2-2-2	0.746±0.00 (0.745)[d]	14.73±0.05 (14.66)
	2	3-3-3-3	0.735±0.00	13.49±0.01
tridecane	1	2-2-3-3-3	0.747±0.00 (0.753)[e]	14.94±0.02 (15.84)
	2	2-2-2-2-3-2	0.758±0.00	15.88±0.07

a: reference [24], b: reference [29], c: reference [25], d: reference [21],e: reference [26],f: reference [21],g: reference [27], h: reference [28]

Table 4.4 Density, enthalpy of vaporization of CG *n*-alkanes at 300 K and 1 atm. Experimental data measured at 298 K is in parentheses. Units: density - g/cm^3, enthalpy of vaporization - kcal/mol. (Continued)

tetradecane	1	2-2-2-2-2-2-2	0.763±0.00 (0.760)[f]	17.39±0.06 (17.16)
	2	3-3-2-2-2-2	0.766±0.00	16.99±0.05
pentadecane	1	3-3-3-3-3	0.751±0.00 (0.765)[g]	16.94±0.03 (18.28)
	2	2-2-2-3-3-3	0.762±0.00	17.83±0.02
hexadecane	1	2-2-2-2-2-2-2-2	0.775±0.00 (0.770)	19.91±0.04 (19.60)
	2	2-2-3-3-3-3	0.765±0.00	18.86±0.01
heptadecane	1	2-2-2-2-3-3-3	0.774±0.00 (0.774)[h]	20.31±0.04 (20.62)
	2	3-2-3-3-3-3	0.775±0.00	19.85±0.02

a: reference [24], b: reference [29], c: reference [25], d: reference [21],e: reference [26],f: reference [21],g: reference [27], h: reference [28]

Figure 4.6 (a - d) shows the self-diffusion coefficient, surface tension, expansibility, and compressibility for the CG hydrocarbon models ranging from pentane to heptadecane. The self-diffusion coefficients of all the hydrocarbons are shown in **Figure 4.6 (a)**. Except for the decane (2-2-2-2-2) model (2.6 % error) and two nonane models (nonane (3-3-3) -- 14.1 %, and hybrid nonane -- 4.7 % errors) all other hydrocarbons showed high deviations from the experimental values (up to 87 % error). For the shorter hydrocarbons (pentane to octane) the self-diffusion coefficient was smaller than the experimental data (minimum and maximum errors 7.1 %, and 33.9 %, respectively). This could be because the values of ε of the middle beads (C2M and C3M beads) from the LJ potentials were smaller than the end beads of the same size (C2E and C3E beads) (see **Table 4.1**). Note, the ε value corresponds to the strength of interactions between molecules and smaller the ε value weaker the interactions. As the chain length decreased from nonane to pentane, the ratio of the number of end beads to that of middle

beads increased in a hydrocarbon. This might have led to the stronger LJ interactions between CG molecules, and lowered the self-diffusion coefficient for pentane as compared to its corresponding experimental value. As the chain length of these hydrocarbons increases from undecane to heptadecane the self-diffusion coefficients were higher as compared to corresponding experimental data. Several recently reported CG models of hydrocarbons in the literature have shown minimum deviation of 2.5 % and maximum deviation of 279 % from the experimental values of self-diffusion coefficients.[19,30] The models in our study showed the minimum and maximum deviation of 1.2 % and 87 % from the experimental values, respectively. Thus, the models developed in this chapter perform better than the existing models reported in literature (see Appendix A **Table A5** and **Table A6**).

The surface tensions of all the hydrocarbons were in excellent agreement with the experimental data (**Figure 4.6 (b)**). As the chain length increases from pentane to heptadecane, the surface tension increases from 14.71 mN/m to 28.33 mN/m for mapping scheme **1**. Similar observation showing the increase in surface tension with increase in the chain length has been reported in other computational and experimental studies.[16,31] The comparison of experimental data with calculated value for pentane showed an error of 5.2 %. In the case of heptadecane for mapping scheme **1** and **2** the errors were 1.4 % and 3.8 %, respectively.

Figure 4.6 (c) shows the expansibility for all the hydrocarbon models with different chain lengths. Overall, the values predicted by the CG models with both the mapping schemes for all the hydrocarbons were slightly lower than the experimental values (minimum and maximum error 1.4 % and 21.1 %, respectively). The isothermal compressibility for all the models with different chain lengths is shown in **Figure 4.6 (d)**. The predicted values for almost all the models were slightly higher than the corresponding experimental values. For example, the two CG models of octane with mapping scheme **1** and **2**, listed in **Table 4.4**, had the values 6.2 % and 12.5 % greater than the experimental compressibility of octane. Similarly the heptadecane models with the mapping scheme **1** and **2** had the value of 8.8×10^{-5} bar^{-1} and 8.6×10^{-5} bar^{-1}, respectively, both slightly higher than the experimental value of 8.2×10^{-5} bar^{-1}. Overall, the 2:1 and 3:1 beads developed for the CG models of decane, hybrid nonane, and nonane, could predict several experimental properties of hydrocarbons ranging from pentane to heptadecane with a good accuracy. In addition, the results obtained for the new CG models developed in this chapter are comparable to that of reported in literature.[19,32] The comparison between the models

developed in this chapter and models reported in literature for dodecane and hexane as representative examples is shown in detail in **Tables A5** and **A6** in Appendix A.

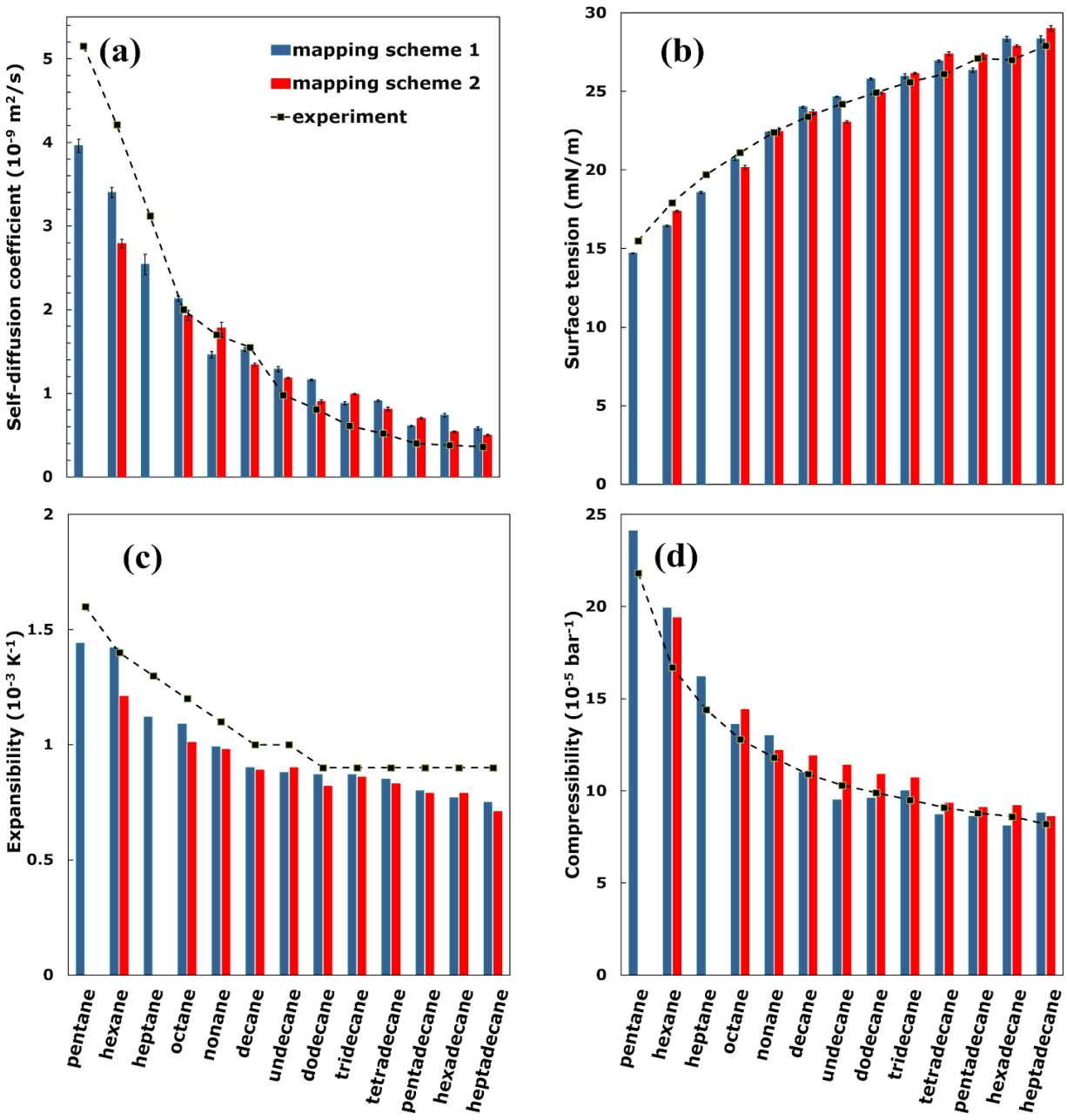

Figure 4.6 Self-diffusion (a), surface tension (b), compressibility (c), and expansibility of hydrocarbons from pentane to heptadecane at 300 K. Mapping schemes **1** and **2** for each hydrocarbon can be found in **Table 4.4**.

4.3.5 CG MD Simulations of Hexadecane as a Representative Hydrocarbon at Different Temperatures. To further test the transferability of the new CG beads developed in this chapter, we have employed them to predict the properties of hexadecane represented by 2-2-2-2-2-2-2-2 mapping scheme at different temperatures. The CG MD simulations of 1000 hexadecane (2-2-2-2-2-2-2-2) molecules were performed with the timestep of 15 fs in the temperature range of 270 K to 350 K. Results of the density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of hexadecane obtained from these simulations are shown in **Figure 4.7**. It can be seen that, as expected, with an increase in temperature, the density decreases (**Figure 4.7 (a)**). From 290 K to 310 K, the density predicted by the CG models and corresponding available experimental data from the literature was in good agreement (error within 0.6 %).[16,17] At 340 K, the density decreased to 0.751 g/cm^3, as compared to the experimental value of 0.740 g/cm^3 (338 K) (an error of 1.5 %).[33] As the temperature increased from 270 K to 350 K, the enthalpy of vaporization decreases from 20.69 kcal/mol to 18.73 kcal/mol (**Figure 4.7 (b)**). The self-diffusion coefficient predicted by the CG models were higher as compared to the experimental values at the simulated temperatures (**Figure 4.7 (c)**). For example, at 340 K the experimental value of self-diffusion coefficient is 0.820×10^{-9} m^2/s as compared to that of the CG model of 1.115×10^{-9} m^2/s. The surface tensions obtained for hexadecane CG models for the simulated temperature range of 290 K to 340 K were within 8 % to that of the corresponding experimental values (**Figure 4.7 (d)**). These results further show the robustness of the new CG beads and the hydrocarbon models in predicting the experimental properties with good accuracy.

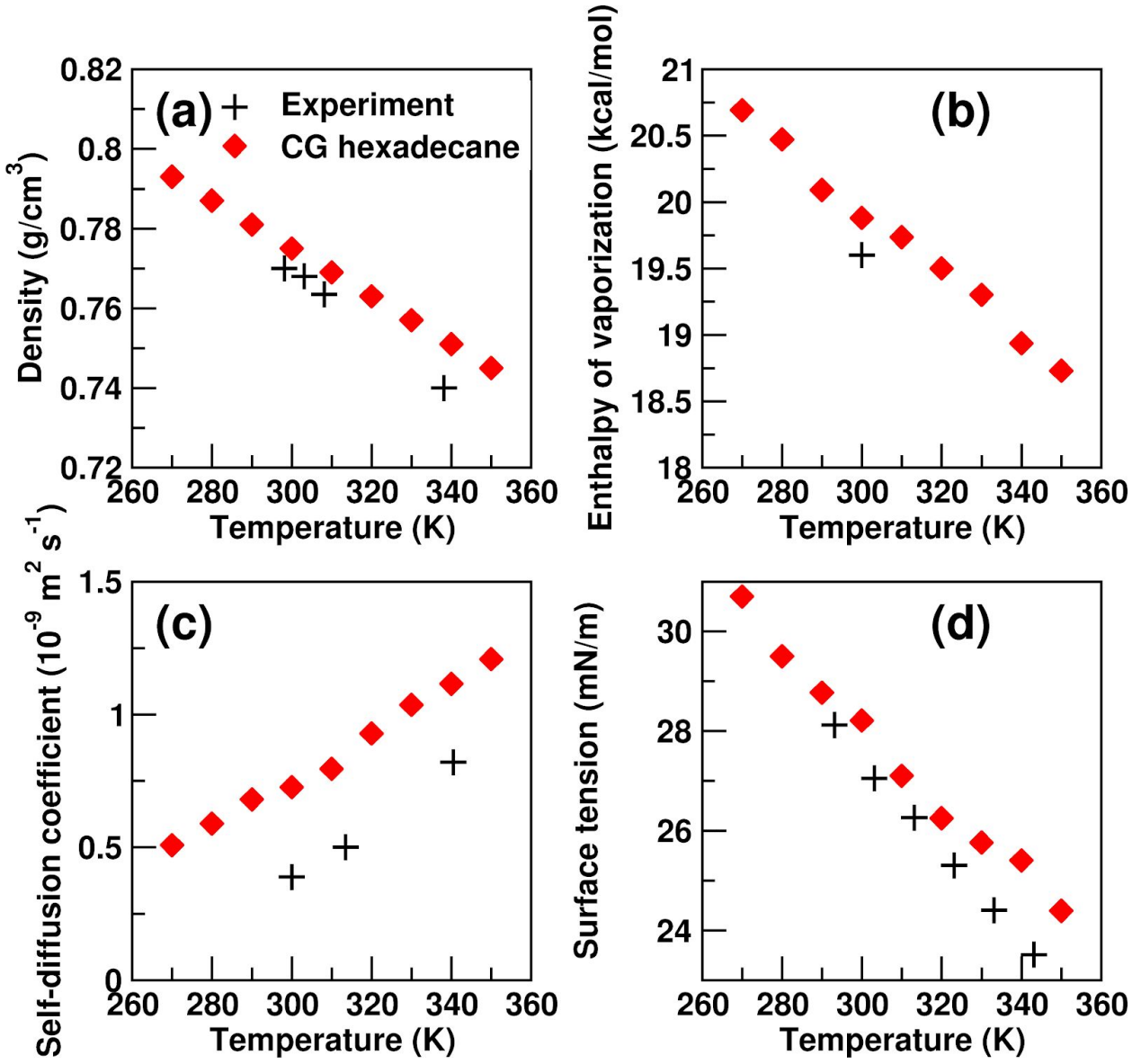

Figure 4.7 Density, enthalpy of vaporization, self-diffusion coefficient, and surface tension of the hexadecane model (2-2-2-2-2-2-2-2) at different temperatures from 270 K to 350 K.

4.4 Conclusion

The coarse-grained (CG) models for the hydrocarbons ranging from pentane to heptadecane were developed using a systematic approach that integrated molecular dynamics (MD) simulations with particle swarm optimization (PSO) method. We initiated this model development by systematically building the models for selected three hydrocarbons: decane (2-2-2-2-2), hybrid nonane (2-2-3-2), and nonane (3-3-3) models. The force-field (FF)

parameters for these models were optimized by using the PSO method at 300 K to reproduce their experimentally measured density, enthalpy of vaporization, self-diffusion coefficient, and surface tension. These models could also predict experimental compressibility and expansibility at 300 K with a good accuracy, which were not used as target properties during the optimization. Moreover, the new CG models showed a reasonable agreement with the structural properties obtained from mapped all-atom trajectories. The MD simulations of the new CG decane model (2-2-2-2-2), hybrid nonane model (2-2-3-2), and nonane model (3-3-3) could predict their experimental properties with different timestep, system size, and temperature with a reasonable accuracy.

The chemical transferability of the CG beads that represented the decane, hybrid nonane, and nonane models was further evaluated by performing the MD simulations of other hydrocarbons, ranging from pentane to heptadecane. All the CG hydrocarbon models showed good agreement with the experimental data for density, enthalpy of vaporization, surface tension, expansibility, and isothermal compressibility. The self-diffusion coefficient deviated from experimental values as the chain lengths deviated from that of the decane and/or nonane. Moreover, the MD simulations performed for the CG model of hexadecane, at different temperatures, showed excellent agreement with experimental properties.

The interactions between CG hydrocarbon beads, and 1-site CG water beads were calibrated to reproduce the free energy of hydration of hydrocarbons. PSO was used to obtain the interactions between C2E and C2M beads, and new 1-site CG water model by reproducing the hydration free energies of CG decane (2-2-2-2-2), and hexadecane (2-2-2-2-2-2-2-2) models with their respective experimental data. Similarly, the interaction parameters between C3E and C3M beads with 1-site CG water model were obtained by matching the hydration free energy of CG nonane (3-3-3) and pentadecane (3-3-3-3-3) models with the corresponding experimental data. The calibrated interaction parameters between these selected hydrocarbons and water could also predict the free energy of hydration of other hydrocarbons with good accuracy. Moreover, the mixture simulations of water and hydrocarbons showed the segregation of hydrocarbons from water, similar to that observed in experiments.

References

(1) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **07/2007**, *111* (27), 7812–7824.

(2) Cao, F.; Sun, H. Transferability and Nonbond Functional Form of Coarse Grained Force Field - Tested on Linear Alkanes. *J. Chem. Theory Comput.* **2015**, *11* (10), 4760–4769.

(3) Avendaño, C.; Lafitte, T.; Adjiman, C. S.; Galindo, A.; Müller, E. A.; Jackson, G. SAFT-γ Force Field for the Simulation of Molecular Fluids: 2. Coarse-Grained Models of Greenhouse Gases, Refrigerants, and Long Alkanes. *J. Phys. Chem. B* **2013**, *117* (9), 2717–2733.

(4) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**. https://doi.org/10.1021/acs.jpcb.7b10542.

(5) Schoen, M.; Hoheisel, C. The Mutual Diffusion Coefficient D 12 in Liquid Model Mixtures A Molecular Dynamics Study Based on Lennard-Jones (12-6) Potentials: II. Lorentz-Berthelot Mixtures. *Mol. Phys.* **1984**, *52*, 1029–1042.

(6) Eberhart, R.; Kennedy, J. A New Optimizer Using Particle Swarm Theory. In *Micro Machine and Human Science, 1995. MHS '95., Proceedings of the Sixth International Symposium on*; 1995; pp 39–43.

(7) Shi, Y.; Eberhart, R. C. Empirical Study of Particle Swarm Optimization. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*; 1999; Vol. 3, p 1950.

(8) Bejagam, K. K.; Singh, S.; Deshmukh, S. A. Development of Non-Bonded Interaction Parameters between Graphene and Water Using Particle Swarm Optimization. *J. Comput. Chem.* **2017**, *39* , 721–734.

(9) Cui, Z.; Gao, F.; Cui, Z.; Qu, J. A Second Nearest-Neighbor Embedded Atom Method Interatomic Potential for Li–Si Alloys. *J. Power Sources* **2012**, *207*, 150–159.

(10) Cui, Z.; Gao, F.; Cui, Z.; Qu, J. Developing a Second Nearest-Neighbor Modified Embedded Atom Method Interatomic Potential for Lithium. *Modell. Simul. Mater. Sci. Eng.* **2011**, *20* (1), 015014.

(11) Moore, J. W.; Wellek, R. M. Diffusion Coefficients of N-Heptane and N-Decane in N-Alkanes and N-Alcohols at Several Temperatures. *J. Chem. Eng. Data* **1974**, *19* (2), 136–140.

(12) Liessmann, G.; Schmidt, W.; Reiffarth, S. Data Compilation of the Saechsische Olefinwerke Boehlen. *Recommended thermophysical data* **1995**, 35.

(13) Lide, D. R. *CRC Handbook of Chemistry and Physics*; Lide, D. R., Ed.; CRC press, 2003.

(14) Douglass, D. C.; McCall, D. W. Diffusion in Paraffin Hydrocarbons. *J. Phys. Chem.* **1958**, *62* (9), 1102–1107.

(15) Qin, X.; Cao, X.; Guo, Y.; Xu, L.; Hu, S.; Fang, W. Density, Viscosity, Surface Tension, and Refractive Index for Binary Mixtures of 1,3-Dimethyladamantane with Four C10 Alkanes. *J. Chem. Eng. Data* **2014**, *59* (3), 775–783.

(16) Rolo, L. I.; Caço, A. I.; Queimada, A. J.; Marrucho, I. M.; Coutinho, J. A. P. Surface Tension of Heptane, Decane, Hexadecane, Eicosane, and Some of Their Binary Mixtures. *J. Chem. Eng. Data* **2002**, *47* (6), 1442–1445.

(17) Cerdeiriña, C. A.; Tovar, C. A.; González-Salgado, D.; Carballo, E.; Romaní, L. Isobaric

Thermal Expansivity and Thermophysical Characterization of Liquids and Liquid Mixtures. *Phys. Chem. Chem. Phys.* **2001**, *3* (23), 5230–5236.
(18) Quayle, O. R.; Smart, K. O. A Study of Organic Parachors. VI. A Supplementary Series of Tertiary Alcohols1. *J. Am. Chem. Soc.* **1944**, *66* (6), 935–938.
(19) Eichenberger, A. P.; Huang, W.; Riniker, S.; van Gunsteren, W. F. Supra-Atomic Coarse-Grained GROMOS Force Field for Aliphatic Hydrocarbons in the Liquid Phase. *J. Chem. Theory Comput.* **2015**, *11* (7), 2925–2937.
(20) Negro, E.; Latsuzbaia, R.; de Vries, A. H.; Koper, G. J. M. Experimental and Molecular Dynamics Characterization of Dense Microemulsion Systems: Morphology, Conductivity and SAXS. *Soft Matter* **2014**, *10* (43), 8685–8697.
(21) Liu, H.; Zhu, L. Excess Molar Volumes and Viscosities of Binary Systems of Butylcyclohexane with N-Alkanes (C7 to C14) at T = 293.15 K to 313.15 K. *J. Chem. Eng. Data* **2014**, *59* (2), 369–375.
(22) Mayer, S. W. Dependence of Surface Tension on Temperature. *J. Chem. Phys.* **1963**, *38* (8), 1803–1808.
(23) Majer, V.; Svoboda, V.; Kehiaian, H. V. *Enthalpies of Vaporization of Organic Compounds: A Critical Review and Data Compilation*; Blackwell Scientific Oxford, 1985; Vol. 32.
(24) Ramos-Estrada, M.; Iglesias-Silva, G. A.; Hall, K. R. Experimental Measurements and Prediction of Liquid Densities for N-Alkane Mixtures. *J. Chem. Thermodyn.* **2006**, *38* (3), 337–347.
(25) Alonso-Tristán, C.; González, J. A.; García de la Fuente, I.; Cobos, J. C. Thermodynamics of Mixtures Containing Amines. XV. Liquid–Liquid Equilibria for Benzylamine + CH3(CH2)nCH3 (n = 8, 9, 10, 12, 14). *J. Chem. Eng. Data* **2014**, *59* (6), 2101–2105.
(26) Zhang, L.; Guo, Y.; Xiao, J.; Gong, X.; Fang, W. Density, Refractive Index, Viscosity, and Surface Tension of Binary Mixtures of Exo-Tetrahydrodicyclopentadiene with Some N-Alkanes from (293.15 to 313.15) K. *J. Chem. Eng. Data* **2011**, *56* (11), 4268–4273.
(27) Tardajos, G.; Diaz Peña, M.; Aicart, E. Speed of Sound in Pure Liquids by a Pulse-Echo-Overlap Method. *J. Chem. Thermodyn.* **1986**, *18* (7), 683–689.
(28) Audsley, A.; Goss, F. R. 584. Atom Polarisation. Part I. The Solvent-Effect Theory and Its Application to the Molecular Refraction and Polarisation of N-Paraffins in the Liquid State. *J. Chem. Soc.* **1950**, *0* (0), 2989–2997.
(29) Ríos, R.; Ortega, J.; Fernández, L.; de Nuez, I.; Wisniak, J. Improvements in the Experimentation and the Representation of Thermodynamic Properties (iso-P VLE and yE) of Alkyl Propanoate + Alkane Binaries. *J. Chem. Eng. Data* **2014**, *59* (1), 125–142.
(30) Baron, R.; Trzesniak, D.; de Vries, A. H.; Elsener, A.; Marrink, S. J.; van Gunsteren, W. F. Comparison of Thermodynamic Properties of Coarse-Grained and Atomic-Level Simulation Models. *Chemphyschem* **2007**, *8* (3), 452–461.
(31) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. A Coarse Grain Model for N-Alkanes Parameterized from Surface Tension Data. *J. Chem. Phys.* **2003**, *119* (14), 7043–7049.
(32) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760.
(33) Sakuramoto, Y.; Shima, M.; Adachi, S. Autoxidation of Mono-, Di-, and Trilinoleoyl Glycerols at Different Concentrations. *Biosci. Biotechnol. Biochem.* **2007**, *71* (3), 803–806.

CHAPTER 5
DEVELOPMENT OF TRANSFERABLE NONBONDED INTERACTIONS BETWEEN CG HYDROCARBON AND WATER MODELS

This work presented in this chapter is reported from [An, Y., Bejagam, K. K., Deshmukh, S. A. Development of Transferable Nonbonded Interactions between Coarse-Grained Hydrocarbon and Water Models, J. Phys. Chem. B, 2019, 123 (4), 909-921], with the permission of AIP Publishing.

Abstract: Interactions between water and hydrocarbons play a significant role in numerical chemical, physical, and biological processes. Here, we present a set of force-field (FF) parameters that define the interactions between coarse-grained (CG) hydrocarbon models and 1-site water model (*J. Phys. Chem. B* 2018, 122, 1958-1971) developed in our recent work. The nonbonded FF interactions between various hydrocarbon beads and the water beads are represented by the 12-6 Lennard-Jones potential. The FF parameters were optimized to reproduce the experimentally measured Gibbs hydration free energies of selected hydrocarbon models (decane and hexadecane with 2:1 mapping scheme, and nonane and pentadecane with 3:1 mapping scheme) and the interfacial tensions of decane and nonane models at 300 K. The predicted values of Gibbs hydration free energies of CG decane, hexadecane, nonane, and pentadecane models by the optimized FF parameters were within 8 %, 12 %, 11 % and 4 % of their corresponding experimental values, respectively. These new optimized FF parameters were transferable when used to calculate the Gibbs hydration free energies of different hydrocarbons ranging from pentane to heptadecane at 300 K (minimum error ~0.5 %, and maximum error ~40.8 %). Furthermore, the interfacial tensions of the CG hydrocarbon models calculated by using these new FF parameters showed good agreement with their corresponding experimental values at 300 K. Homogeneous mixtures of CG water and hydrocarbon models were able to exhibit the phase segregation during 1 μs. These new nonbonded interaction parameters were expected to be utilized in modeling the interactions between water and polymer backbones represented with hydrocarbon beads.

5.1 Introduction

Developing accurate cross interaction parameters between unlike beads is a fundamental problem in MD simulations. Lorentz-Berthelot (LB) combining rules are usually used to estimate the cross interaction parameters between different types of atoms/beads which are modeled as Lennard-Jones (or exponential) sites and possibly of electrostatic sites.[1] They're simple, but are found to be limited in obtaining accurate cross interaction parameters between unlike sites in quite a few cases. Waldman et al. assessed the LB combining rules for rare gas mixtures.[2] They found that the ε values for the He-Xe mixtures obtained by the Berthelot combining rule was 79 % higher than the ones optimized by the correlation with experimental data. Song et al. studied the failure of the LB combing rules in predicting the mixing behaviour of alkane and perfluoroalkane.[3] They attributed this to the failure of the geometric mean combing rule for relating unlike-pair interactions. To achieve the reasonable agreement with experiment, a reduction of ~25 % in the strength of cross H+F (hydrogen + fluorine atoms) was required.

In this chapter, we employed a systematic approach to develop the cross interaction parameters between CG hydrocarbon beads reported in **Chapter 4**. Firstly, we optimized the nonbonded interactions of 1-site water model with selected hydrocarbons, namely, decane, nonane, pentadecane and hexadecane, to reproduce their Gibbs free energies of hydration and interfacial tensions. Then we validated the transferability of these new FF parameters by performing CG MD simulations of hydrocarbons ranging from pentane to heptadecane. We found that the new FF parameters were transferable, and they could predict the Gibbs free energies of hydration and interfacial tensions of these hydrocarbons with a very good accuracy. Moreover, we have performed CG MD simulations for 1 μs to determine the ability of these new interaction parameters in predicting the hydrocarbon-water phase segregation at 300 K.

5.2 FF Parameter Optimization

By utilizing an optimization approach that couples particle swarm optimization (PSO) algorithm with MD simulations, we have developed transferable CG models for water and hydrocarbons.[4,5] The water model was represented by a 2:1 mapping scheme, called as a 1-site water model (the W1 bead in **Figure 5.1**). The nonbonded FF parameters of both the CG 1-site water model and hydrocarbon models were represented by 12-6 LJ potential and are shown in **Table B1** of Appendix B. As shown in **Table 5.1**, the 1-site water model was able to predict the

physical and thermal properties of bulk water at 300 K with a good accuracy.[4] For example, the density (ρ) and the isothermal compressibility (κ_T) were within 0.5 % and 11.4 % of the reported experimental data, respectively. The properties of hydrocarbon models could be found in **Chapter 4**.

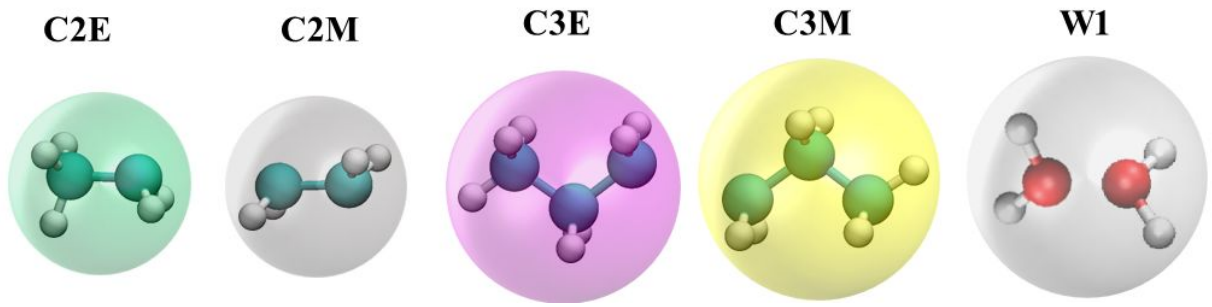

Figure 5.1 CG hydrocarbon beads: C2E (green), C2M (gray), C3E (magenta) and C3M (yellow), and the 1-site water bead: W1 (white).[4,5]

Table 5.1 Properties of the 1-site water model and the CG decane and nonane models previously developed in references [4,5] at 300 K and 1 bar. Experimental data is obtained from references [6–10]. Percentage errors of the properties are shown in parentheses compared with experimental data.

CG models	ρ	D	$\gamma_{air/water(or\ hydrocarbons)}$	κ_T	H_v
1-site water	1.002 ± 0.0 (0.5 %)	2.49 ± 0.02 (4.6 %)	68.2 ± 0.76 (5.3 %)	4.03 ± 0.1 (11.4 %)	7.8 ± 0.03 (28.6 %)
Experiment	0.997	2.38	72.0	4.55	10.5

Units: Density: ρ - g/cm^3, self-diffusion coefficient: D - $\times 10^{-9}$ m^2/s, surface tension: $\gamma_{air/water(or\ hydrocarbons}$ - mN/m, isothermal compressibility: κ_T - $\times 10^{-5}$ bar^{-1}, enthalpy of vaporization: H_v - kcal/mol.

As the Gibbs hydration free energy of hydrocarbons and interfacial tensions can capture the interactions between hydrocarbons and water effectively, we have used them as the target properties to parameterize the nonbonded interactions between the hydrocarbon beads and 1-site

CG water bead (W1).[4,5] Such an approach has already been successfully used to develop the nonbonded interactions between two types of molecules in literature.[4,11–13] Here, the Gibbs hydration free energies were determined by performing Adaptive Biasing Force (ABF) simulations using colvars package implemented in NAMD 2.12.[14,15] Details about the ABF simulations and interfacial tensions between hydrocarbons and water could be found in **Section 3.3.6** of **Chapter 3**. Similar to our previous approach, to accelerate the process of developing the nonbonded parameters between the CG hydrocarbon and 1-site water beads, the PSO method was used.[4,5] The PSO run was initiated with 40 particles in search of the FF parameters, and iterated until the errors of the Gibbs hydration free energy and interfacial tensions of the CG models did not decrease significantly.

The interaction parameters between C2E-W1 and between C2M-W1, i.e. ε(C2E-W1), ε(C2M-W1), σ(C2E-W1), and σ(C2M-W1), were optimized to reproduce the Gibbs hydration free energies of the CG decane (2-2-2-2-2, i.e. a model with C2E-C2M-C2M-C2M-C2E beads) and the CG hexadecane (2-2-2-2-2-2-2-2, i.e. a model with C2E-C2M-C2M-C2M-C2M-C2M-C2M-C2E beads) models, and the interfacial tension of the CG decane(2-2-2-2-2)/water system at 300 K simultaneously. Note, the σ(C2M-W1) value was kept the same as the σ[C2E-W1] value during the optimization. Similarly, the interactions of C3M and C3E beads with the W1 bead were optimized by reproducing the Gibbs hydration free energies of the CG nonane (3-3-3, i.e. a model with C3E-C3M-C3E beads) and pentadecane (3-3-3-3-3, i.e. a model with C3E-C3M-C3M-C3M-C3E beads) models, and the interfacial tension of nonane(3-3-3)/water at 300 K simultaneously. The σ(C3M-W1) value was kept the same as the σ[C3E-W1] value during the optimization.

As hydration free energy of hydrocarbons can capture the interaction between the hydrocarbon and water effectively, we have used it as the target property to parameterize the cross-interactions between the hydrocarbons and 1-site CG water models.[4] Such an approach has already been used successfully to develop the cross-interactions between two types of molecules.[11–13] The 1-site water model was recently developed in our group and it contains a charge neutral bead with 2:1 mapping scheme.[4] Here, the free energies were determined by performing Adaptive Biasing Force (ABF) simulations using colvars package implemented in NAMD 2.12.[14,15] (For more details see **Chapter 3**). Similar to the hydrocarbon model development, to accelerate the process of developing the parameters between the hydrocarbon

and water beads, PSO method was used. The interactions between C2E and C2M beads with the CG 1-site water bead were optimized by reproducing the free energies of hydration of the CG decane and the CG hexadecane models simultaneously. Both these models were represented by a 2:1 mapping scheme. The interactions between C3M and C3E beads were optimized by fitting the free energies of hydration of the CG nonane and pentadecane models. Both these models were represented by a 3:1 mapping scheme.

5.3. Results and Discussion

5.3.1 Optimized FF Parameters between Hydrocarbon and W1 Beads

The nonbonded interactions between CG hydrocarbon beads and W1 beads were represented by the 12-6 LJ equation:

$$E_{nonbond} = 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \qquad \textbf{.....Equation 5.1}$$

Where, ε_{ij} is the depth of the potential well and represents the strength of interactions between two beads i and j.[16] σ_{ij} is the finite distance at which the inter-particle potential is zero, and r_{ij} is the center-center distance between two beads. Note, these nonbonded interactions between hydrocarbon beads and W1 beads are represented by using 12-6 LJ potential because we wanted to be consistent with the FF form used to describe the nonbonded interactions between 1-site CG water with itself and between the CG hydrocarbons with themselves.[4,5]

The PSO algorithm was used to optimize the ε and σ values between CG hydrocarbon beads and W1 beads, in order to reproduce the Gibbs hydration free energy of nonane(3-3-3), decane(2-2-2-2-2), pentadecane(3-3-3-3-3) and hexadecane(2-2-2-2-2-2-2-2) models, and the interfacial tensions of nonane(3-3-3) and decane(2-2-2-2-2) models at 300 K. The optimization ranges and the optimized values of ε and σ are shown in **Table 5.2**. Note, before optimizing the ε and σ values using PSO, we used the Lorentz-Berthelot (LB) combining rules to represent the ε and σ values (shown in **Table 5.2**) between various beads of hydrocarbons and the W1 bead. However, the predicted value of the Gibbs hydration free energy of the decane (2-2-2-2-2) model was -5.5 kcal/mol, far from the experimental value of 3.23 kcal/mol.[17] Similarly, the Gibbs hydration free energy of the nonane (3-3-3) model by using the LB combining rules was -5.3 kcal/mol, which is much lower compared to experimental value of 3.05 kcal/mol.[17] This suggests

that the interaction between the hydrocarbon beads and the W1 bead is too strong. In literature there are several studies that report similar limitations of LB combining rules.[2,3] Waldman *et al.* assessed the LB combining rules for rare gas mixtures.[2] They found that the ε values for the He-Xe mixtures obtained by the Berthelot combining rule was 79 % higher than the ones optimized by the correlation with experimental data. Song *et al.* studied the failure of the LB combing rules in predicting the mixing behaviour of alkane and perfluoroalkane.[3] They attributed this to the failure of the geometric mean combing rule for relating unlike-pair interactions. To achieve the reasonable agreement with experiment, a reduction of ~25 % in the strength of cross H+F (hydrogen + fluorine atoms) was required.

Based on the above discussion, the optimization ranges of ε values were chosen to be 0.4 - 0.8 kcal/mol. The values for σ(C2E-W1) and σ(C3E-W1) were optimized within the 3.6 - 4.5 Å and 3.8 - 4.6 Å range (approximately ±10 % of those obtained by the Lorentz rule), respectively. Here, the σ(C2M-W1) value was kept the same as the σ(C2E-W1) value during the PSO runs, and the σ(C3M-W1) value was also the same as the σ(C3E-W1) value. **Table 5.2** reports the final optimized values obtained from the PSO run and the LB combining rules. These ε and σ values obtained from PSO run are smaller than those obtained by the LB combining rules, which further suggests that the LB combining rules overestimated the FF parameters.

Table 5.2 The optimization range and new optimized 12-6 LJ interaction parameters of ε and σ between CG hydrocarbon beads and W1 beads. The ε and σ values obtained by LB combining rules are also shown for comparison purposes.

bead pairs	**ε by PSO (kcal/mol)**		**ε obtained by the Berthelot combining rule (kcal/mol)**	**σ by PSO (Å)**		**σ obtained by the Lorentz combining rule (Å)**
	optimization range	optimized values		optimization range	optimized values	
C2E-W1	0.40 - 0.80	0.5130	0.6510	3.6-4.5	3.774	4.0547
C2M-W1	0.40 - 0.80	0.440	0.6251	3.6-4.5	3.774	4.0547
C3E-W1	0.40 - 0.80	0.6202	0.8229	3.8-4.6	4.009	4.2033

C3M-W1	0.40 - 0.80	0.5434	0.7959	3.8-4.6	4.009	4.2033

Table 5.3 shows the Gibbs hydration free energies of the four CG models obtained by employing the optimized FF parameters reported in **Table 5.2.** It can be seen that the Gibbs hydration free energies of the CG decane and hexadecane models with 2:1 mapping scheme are within 8 % and 12 % of the experimental values, respectively. Similarly, for the nonane and pentadecane models with 3:1 mapping scheme, the Gibbs free energies are 3.4 kcal/mol (experimental value: 3.05 kcal/mol), and 4.30 kcal/mol (experimental value: 4.13 kcal/mol), respectively. Profiles of the Gibbs hydration free energy of the hydrocarbon models can be found in **Figure B1** in Appendix B.

To further test the optimized FF parameters between CG hydrocarbon beads and 1-site water bead W1, the interfacial tensions of the hydrocarbon/water systems were calculated. As shown in **Table 5.3**, the values of interfacial tensions for the CG decane (2-2-2-2-2) and nonane (3-3-3) models were only 7 % and 10 % lower than their corresponding experimental values. The interfacial tensions of the hexadecane(2-2-2-2-2-2-2-2)/water and pentadecane(3-3-3-3-3)/water systems were 51.9 and 53.8 mN/m, respectively, in good agreement with their corresponding experimental data. Note, the interfacial tensions of the hexadecane(2-2-2-2-2-2-2-2)/water and pentadecane(3-3-3-3-3)/water systems were not used as the target values during the FF optimization performed by using the PSO method. This indicates the transferability of the optimized FF parameters in predicting the interfacial tensions of other hydrocarbon/water systems, which is further investigated in **Section 5.3.3** below.

Table 5.3 Gibbs hydration free energies (ΔG_{hyd}) and interfacial tensions ($\gamma_{hydrocarbon/water}$) of CG decane(2-2-2-2-2)/water, hexadecane(2-2-2-2-2-2-2-2)/water, nonane(3-3-3)/water, and pentadecane(3-3-3-3-3)/water at 300 K and 1 bar. Experimental values measured at 298.15 K are shown in parentheses and taken from the references [17–19].

Hydrocarbons	ΔG_{hyd} **(kcal/mol)**	$\gamma_{hydrocarbon/water}$ **(mN/m)**
decane (2-2-2-2-2)	3.5±0.3 (3.23)	48.0±0.5 (51.5)

hexadecane (2-2-2-2-2-2-2-2)	3.8±0.2 (4.31)	51.9±0.8 (53.86)*
nonane (3-3-3)	3.4±0.2 (3.05)	46.2±0.8 (51.2)
pentadecane (3-3-3-3-3)	4.3±0.4 (4.13)	53.8±0.4 (54.8)

*: 297 K, reference [20]

5.3.2 Phase Segregation of Hydrocarbon/Water Mixtures

Furthermore, we investigated the mixtures of decane(2-2-2-2-2)-water and nonane(3-3-3)-water to understand their miscibility behavior. Experimentally decane-water and nonane-water systems are immiscible at 300 K when the decane and nonane concentrations are above 0.02 mg/L and 0.17 mg/L, respectively.[21] Similar behaviour in CG MD simulations can be observed as segregation of molecules, and it is strongly dependent on the accuracy of the nonbonded interaction parameters between water and hydrocarbons.[22] Here, to investigate the ability of the new interaction parameters in predicting this behavior we performed simulations of the decane(2-2-2-2-2)-water and nonane(3-3-3)-water systems. **Figure 5.2 - (a)** and **(b)** show the snapshots of these systems with 6000 CG water molecules and 2000 hydrocarbon molecules at 0 ps, ~0.2 - 3 ns, and 1 μs. It can be seen that at ~0.2 - 3 ns both the decane and nonane molecules are partially segregated from water, and then totally segregated after ~3 ns, and remain segregated till the end of 1 μs. These results clearly show that the new interaction parameters can qualitatively reproduce experimental behavior of the phase segregation of the two hydrocarbon-water systems.

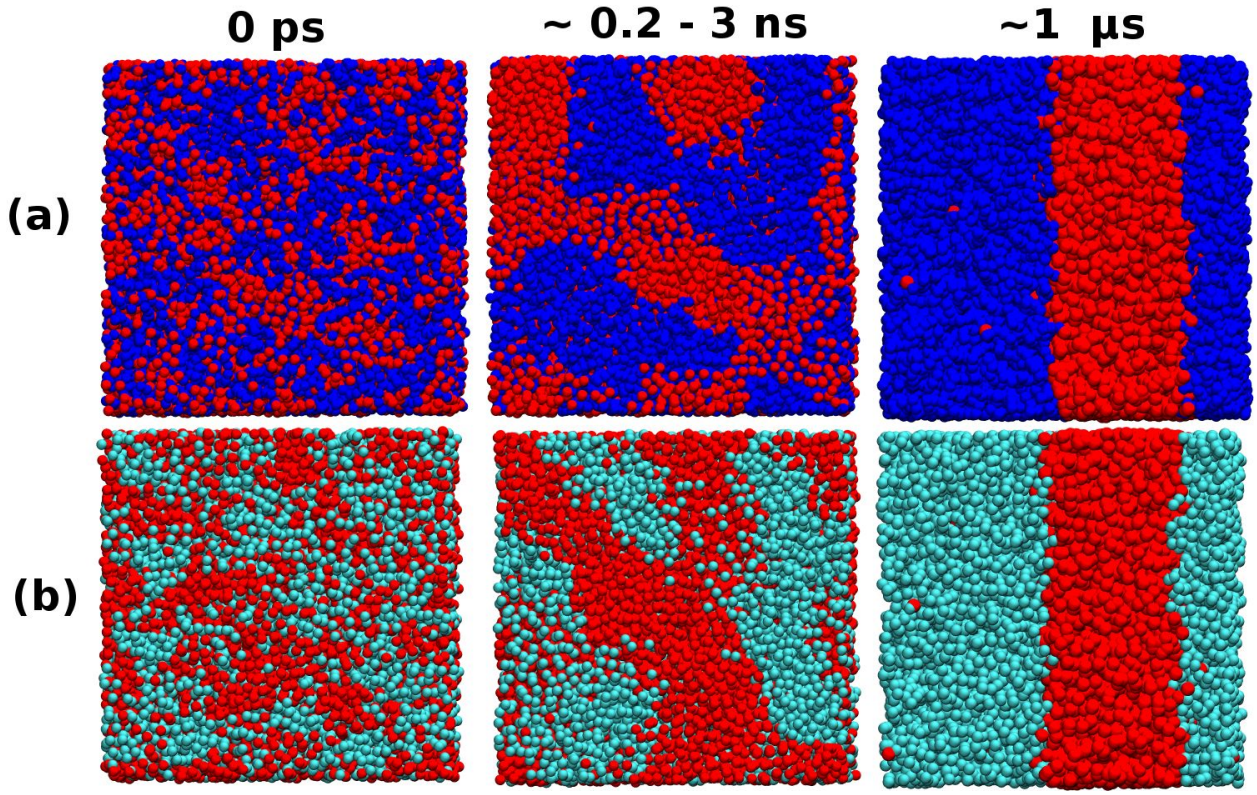

Figure 5.2 Snapshots of the configuration evolution of hydrocarbon-water mixtures. **(a)** CG decane (2-2-2-2-2) (blue) and water (red) mixture, **(b)** CG nonane (3-3-3) (cyan) and water (red) mixture. Note, the snapshots in the middle column are representative examples of the simulation configurations during ~0.2 - 3 ns. In each mixture, the number of CG hydrocarbon and 1-site water molecules is 2000 and 6000, respectively.

To further quantify this segregation and analyze the structure of water and hydrocarbons, we have calculated the density profiles of hydrocarbons and water, and the RDFs between W1 and selected hydrocarbon beads, and between hydrocarbon beads. The density profiles for both decane(2-2-2-2-2)-water and nonane(3-3-3)-water systems are shown in **Figure 5.3 (a)** and **(b)**, respectively. These profiles were obtained by analyzing the first 50 ps and the last 100 ns of the 1-µs CG MD simulation trajectories. The initial configurations for hydrocarbon and water mixture systems were generated by random insertion of these molecules in a simulation cell. We find that during initial 50 ps the CG water and hydrocarbon molecules remain randomly distributed in the simulation box. Thus, the density of the water (0.32 - 0.42 g/cm^3) and hydrocarbons (0.45 - 0.50 g/cm^3), as depicted in **Figure 5.3 (a-1)** and **(b-1)**, is much less than the

corresponding experimental densities of pure water and pure hydrocarbons. Separation of these molecules results in two phases, a hydrocarbon-rich phase and a water-rich phase. In **Figure 5.3 (a-2)**, the density of the decane-rich phase (at the distance from 0 to 42 Å and from 91 to 98 Å) was ~0.72 g/cm^3, which was close to 0.723 g/cm^3 of the pure CG decane model (2-2-2-2-2) in **Table 5.1**. On the other hand, at a distance ranging from ~58 - 75 Å (water-rich phase) in the **Figure 5.3 (a-2)**, the density, ~0.995 g/cm^3, was similar to that of the pure 1-site water model, 1.002 g/cm^3 in **Table 5.1.** As shown in **Figure 5.3 (b-2)** the density of the nonane-rich phase at 0 - 41 Å and 89 - 98 Å, was ~ 0.707 g/cm^3, close to its bulk density of 0.709 g/cm^3. For a water-rich phase from the distance of ~56 to 74 Å the density was ~0.994 g/cm^3. Similar density profiles for linear hydrocarbon and water systems have been reported in the literature.[22,23] The density profiles during the last 100 ns (in **Figure 5.3 (a-2)** and **(b-2)**) are consistent with the corresponding snapshots of the mixture configuration shown in **Figure 5.2** at the end of 1 μs. Note, a small number of water molecules could be observed in the snapshots in **Figure 5.2** at 1 μs in the decane-rich and nonane-rich phases, which are ~ 22 and ~28, corresponding to the densities of 0.0027 and 0.0034 g/cm^3 of water, respectively. The set of optimized parameters were also able to predict the phase segregation when the number of CG water molecules was further increased to 10000 and 18000 while keeping the CG hydrocarbon molecules fixed at 2000. The density profiles of the mixtures with 10000 and 18000 CG water molecules are shown in **Figure B2** of Appendix B.

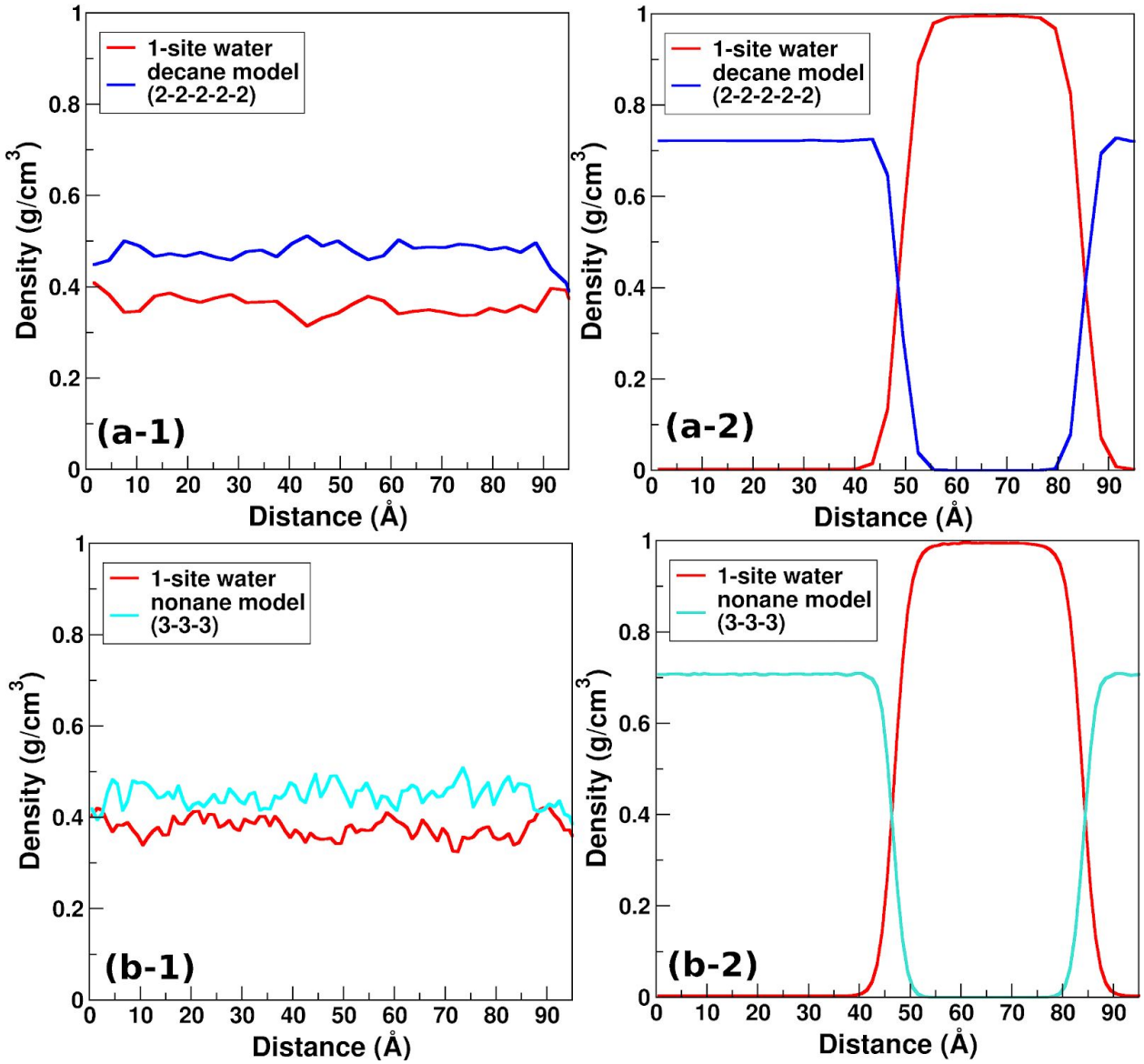

Figure 5.3 Density profiles of the CG water/hydrocarbon mixtures during the last 100 ns of 1-μs simulations: **(a)** decane(2-2-2-2-2)/water mixture, **(b)** nonane(3-3-3)/water mixture, **(1)** the initial 50 ps, **(2)** the final 100 ns. In each system, the number of CG hydrocarbon and 1-site water molecules is 2000 and 6000, respectively.

The radial distribution functions (RDFs) between the W1 beads and the end beads of decane (C2E) and nonane (C3E) are shown in **Figure 5.4 (a-1)** and **(b-1)**. These RDFs show that the height of the first and second peaks decreases during the last 100 ns of the total 1 μs simulation as compared to the initial 50 ps. The coordination number of W1 beads around the C2E bead in the decane (2-2-2-2-2) at a distance of 6.5 Å (the position of the first minimum in

Figure 5.4 - (a-1)) decreased from 1.7 to 0.3 in the **Figure B3 - (a-1)** of the Supporting Information. Similarly, the coordination number of the W1 beads around the C3E bead in the nonane (3-3-3) model decreased from 1.9 to 0.3 at 6.5 Å in **Figure B3 - (b-1)** of the Supporting Information. This suggests that the number of the water molecules in the first hydration shell of the hydrocarbons reduces due to segregation of two phases. In **Figure 5.4 (a-2)** and **(b-2)**, the RDF of the C2E-C2E in the decane (2-2-2-2-2) model and that of the C3E-C3E in the nonane (3-3-3) model are shown during the first 50 ps and last 100 ns of the total 1 μs simulation run. The height of the first and second peaks in both of the RDFs increased during the last 100 ns, which also indicates the aggregation of the hydrocarbon molecules by the end of 1 μs. Overall, the RDFs in **Figure 5.4** and the coordination number in **Figure B3** of Appendix A during the initial 50 ps and final 100 ns, further validate the segregation of water/hydrocarbon mixtures.

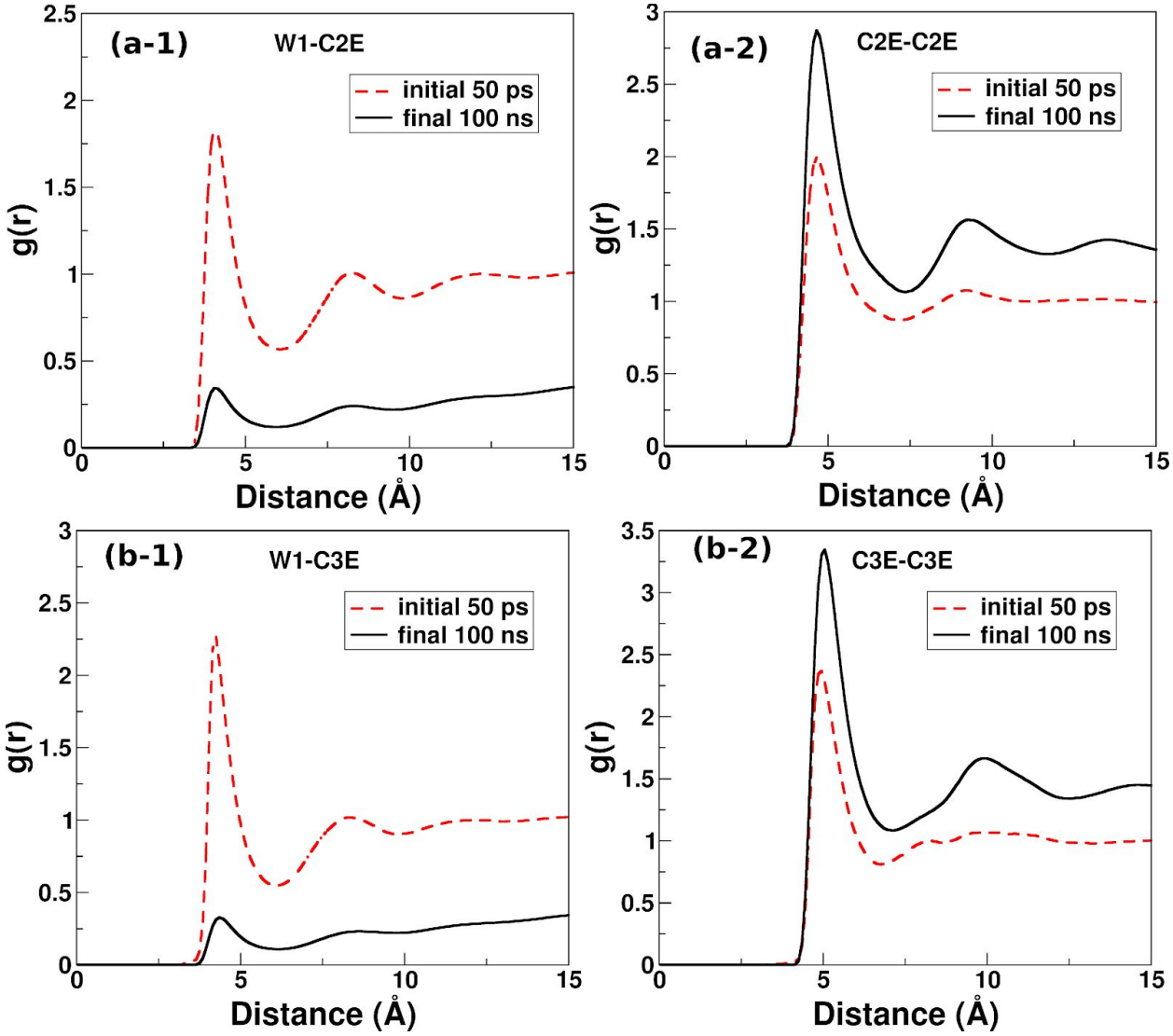

Figure 5.4 RDFs between **(a-1)** W1 bead and the C2E beads in CG decane(2-2-2-2-2) model; **(a-2)** C2E and C2E beads in the decane(2-2-2-2-2) model; **(b-1)** the W1 bead and C3E beads in the nonane(3-3-3) model; **(b-2)** the C3E and C3E beads in the nonane(3-3-3) model. Note, the RDFs of W1-C2E and W1-C3E during the final 100 ns increased to 1.0 when the distance increased to ~50 Å, while those of C2E-C2E and C3E-C3E during the final 100 ns decrease to 1.0 at ~50 Å.

5.3.3 Transferability of the New FF Parameters

Gibbs Hydration Free Energies and Interfacial Tensions of Other Hydrocarbons

In this section, we report the transferability of the parameters optimized in **Section 5.3**. Specifically, we have determined the Gibbs hydration free energies for hydrocarbons ranging

from pentane to heptadecane at 300 K, most of which were not used during optimization of FF parameters. These hydrocarbons were represented by the mapping schemes shown in **Table 5.4**.

As shown in **Table 5.4**, we find that the Gibbs hydration free energy increases as the chain length increases, similar to experiments.[17] The Gibbs hydration free energy of the CG pentane model (the shortest hydrocarbon studied in the present work) was 2.0 kcal/mol, 14 % smaller than the experimental value of 2.33 kcal/mol. As the chain length increased to thirteen and fourteen, *i.e.* tridecane and tetradecane, the Gibbs hydration free energies of tridecane and tetradecane models with mapping scheme **1** increased to 3.5 and 4.1 kcal/mol, respectively, which were in good agreement with the experimental data, 3.77 and 3.96 kcal/mol, respectively. Comparison of the Gibbs hydration free energy of our models with existing CG and united-atom (UA) models can be found in **Figure 5.5 - (a)**. The Gibbs hydration free energies of our hydrocarbon models are comparable with the existing CG and UA models.[11,24–26]

The interfacial tensions of the CG hydrocarbon models and their corresponding experimentally measured values are also shown in **Table 5.4**. Experimental data is obtained from the references [19,20]. The interfacial tensions of the majority of the hydrocarbon models show good agreement with the available experimental data. For example, the interfacial tensions of the octane and tetradecane models with mapping scheme **1** were within 4.5 % and 7.5 % of the experimental values, respectively. As the chain length of the hydrocarbons increased from six (hexane) to sixteen (hexadecane), the interfacial tension of the CG models with mapping scheme **1** increased from 43.7 mN/m to 51.9 mN/m. A similar trend has been observed in experiments, where the interfacial tensions of hydrocarbons from hexane to hexadecane increases from 49.96 mN/m to 53.86 mN/m.[19,20] The comparison of interfacial tensions predicted by our models and reported CG, UA and AA models is also shown in **Figure 5.5 - (b)**. The interfacial tension of our hydrocarbon models were comparable with the existing UA and AA, and better than the MARTINI model at 300 K. The interfacial tensions of selected hydrocarbons as representative examples at different temperatures are shown in **Table B3** in Appendix B. We find that as the temperature increased from 300 K to 350 K, the interfacial tensions of all the hydrocarbons decreased. At 320 K, the interfacial tension of the decane(2-2-2-2-2)/water system was 47.0 mN/m, in good agreement with the experimental value of 50.0 mN/m, while that of the nonane(3-3-3)/water was within 16 % of the experimental data (See **Table B3** in Appendix B).[19]

To further explore the effects of different mapping schemes on the bulk properties of the hydrocarbon models and the properties of hydrocarbon/water systems, we have performed simulations of selected hydrocarbon models with more than two mapping schemes (the number and position of the beads with 3:1 mapping scheme are different). For hydrocarbons with hybrid mapping scheme (including both beads with 2:1 and 3:1 mapping scheme) and two ends represented by C3E beads, the Gibbs hydration free energies were usually underestimated compared with experimental data. For example, in **Table 5.4** the octane (3-2-3) model predicts a Gibbs hydration free energy of 1.7 kcal/mol, smaller than 2.87 kcal/mol of the experimental value. Similarly, the Gibbs hydration free energy of the dodecane (3-2-2-2-3) model was 22 % smaller than the experimental data. This could be attributed to the larger ε value between C3E and W1 bead and the smaller σ values of C2M and W1 beads. Both result in stronger interactions between hydrocarbon and water as well as altering the structure of water molecules at the hydrocarbon and water interface, thus, leading to smaller Gibbs hydration free energies of hydrocarbons. The effects of mapping schemes on the bulk properties of hydrocarbons can be found in **Table B4** of Appendix B.

To understand the effects of symmetry of the models on their properties, the symmetric dodecane model (2-3-2-3-2) and asymmetric model (2-3-3-2-2) were compared, and also the symmetric heptadecane (3-2-2-3-2-2-3) model and asymmetric heptadecane model (3-2-2-2-2-3-3) were studied. The percentage differences in the Gibbs hydration free energy and interfacial tensions between the symmetric and asymmetric dodecane models were 7 % and 4.2 %. The symmetric and asymmetric heptadecane models had percentage differences of 13.8 % and 0.2 % in Gibbs hydration free energy and interfacial tension, respectively. The Gibbs hydration free energies of the symmetric dodecane (2-3-2-3-2) and heptadecane (3-2-2-3-2-2-3) models were within 11.4 % and 8.7 % of their experimental values, smaller than those of the asymmetric dodecane (19.8 %) and heptadecane models (19.8 %). This suggests that the symmetric models perform better in predicting the experimental Gibbs hydration free energies.

Table 5.4 Gibbs hydration free energy (ΔG_{hyd}) and interfacial tensions $\gamma_{hydrocarbons/water}$ for CG hydrocarbon models with different mapping schemes at 300 K and 1 bar. Experimental data in parentheses is from the reference [17] [19,20] at 298.15 K.

Hydrocarbon	Mapping scheme No.	Mapping schemes	ΔG_{hyd} (kcal/mol)	$\gamma_{hydrocarbon/water}$ (mN/m)
pentane	1	2-3	2.0±0.3 (2.33)	39.20±0.5 (-)
hexane	1	2-2-2	2.1±0.5 (2.51)	43.7±0.3 (49.96)
	2	3-3	1.8±0.3	41.0±0.4
heptane	1	2-3-2	2.2±0.12 (2.69)	45.4±0.7 (50.3)
octane	1	2-2-2-2	2.5±0.2 (2.87)	48.4±0.5 (50.7)
	2	2-3-3	2.3±0.4	45.0±0.3
	3	3-2-3	1.7±0.4	45.8±0.5
nonane	1	3-3-3	3.4±0.2 (3.05)	46.2±0.8 (51.2)
	2	2-2-3-2	3.4±0.3	47.1±0.9
decane	1	2-2-2-2-2	3.5±0.3 (3.23)	48.0±0.5 (51.5)
	2	2-2-3-3	3.4±0.2	47.0±0.7
	3	3-2-2-3	3.2±0.3	48.3±0.6
undecane	1	2-2-2-3-2	3.1±0.4 (3.41)	49.3±0.5 (51.8)
	2	3-3-3-2	3.6±0.4	47.1±0.3
dodecane	1	2-2-2-2-2-2	3.0±0.3 (3.59)	51.0±0.4 (52.1)

Table 5.4 Gibbs hydration free energy (ΔG_{hyd}) and interfacial tensions $\gamma_{hydrocarbons/water}$ for CG hydrocarbon models with different mapping schemes at 300 K and 1 bar. Experimental data in parentheses is from the reference [17] [19,20] at 298.15 K. (Continued)

dodecane	2	3-3-3-3	3.5±0.1	47.0±1.0
	3	3-2-2-2-3	2.8±0.5	48.1±0.5
	4	3-3-2-2-2	4.3±0.4	48.8±0.6
	5	2-3-3-2-2	4.3±0.4	47.1±0.4
	6	2-3-2-3-2	4.0±0.3	49.1±0.7
tridecane	1	2-2-3-3-3	3.5±0.1 (3.77)	47.8±0.8 (-)
	2	2-2-2-2-3-2	3.5±0.4	48.9±1.2
tetradecane	1	2-2-2-2-2-2-2	4.1±0.3 (3.96)	50.4±0.9 (54.5)
	2	3-3-2-2-2-2	3.2±0.4	49.9±0.5
pentadecane	1	3-3-3-3-3	4.3±0.4 (4.13)	53.8±0.4 (54.8)
	2	2-2-2-3-3-3	4.2±0.50	47.6±0.6
	3	2-2-2-2-2-2-3	4.8±0.2	51.5±0.4
	4	3-2-2-2-3-3	3.2±0.2	50.4±0.7
	5	2-2-3-3-3-2	4.5±0.4	49.3±0.6
hexadecane	1	2-2-2-2-2-2-2-2	3.8±0.2 (4.31)	51.9±0.8 (53.9)
	2	2-2-3-3-3-3	4.7±0.50	49.1±0.7
heptadecane	1	2-2-2-2-3-3-3	5.5±0.20 (4.49)	51.2±0.9 (-)
	2	3-2-3-3-3-3	5.4±0.3	50.3±0.8
	3	3-2-2-2-2-3-3	3.6±0.3	50.6±0.5
	4	3-2-2-3-2-2-3	4.1±0.4	50.7±0.6

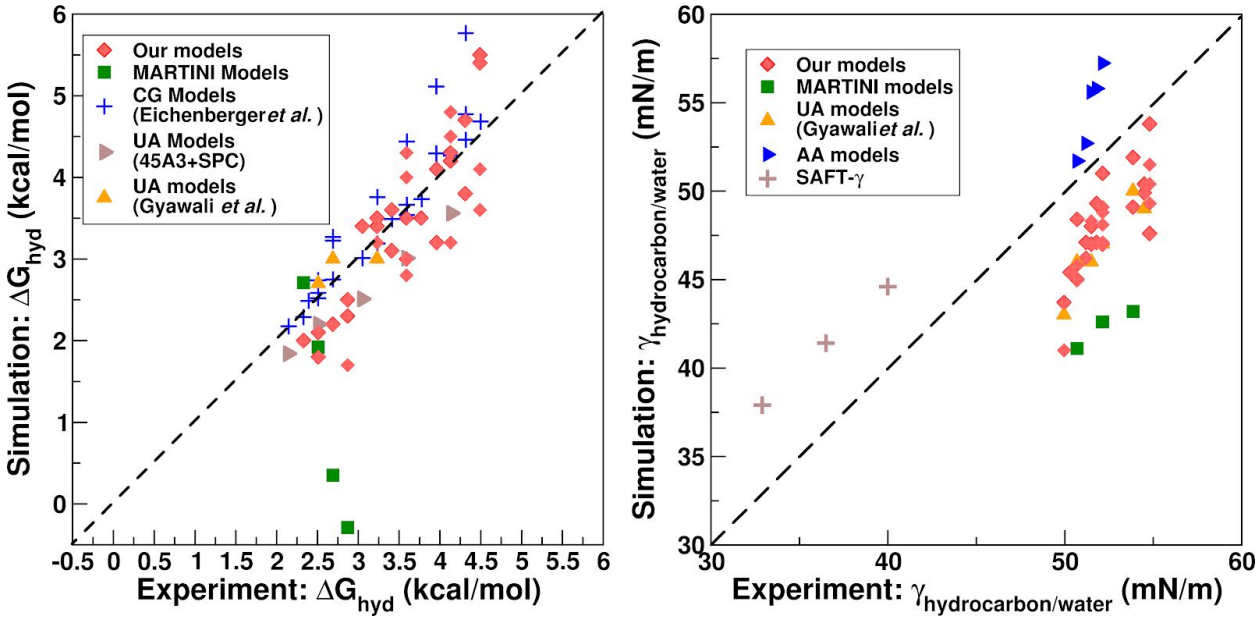

Figure 5.5 Comparison of our models (all models in **Table 3.4**) with reported hydrocarbon models in terms of **(a)** Gibbs hydration free energy (ΔG_{hyd}) and **(b)** interfacial tensions ($\gamma_{hydrocarbons/water}$). The data of the MARTINI, CG (Eichenberger *et al.*), UA (45A3+SPC), UA (Gyawali *et al.*), SAFT-γ, and AA models are from the references [11,23–27]. Note, the interfacial tensions of the SAFT-γ models were calculated at 403 - 443 K, while those of other models were at ~300 K.

5.3.4 Phase Segregation in the Hybrid Nonane and Water System as a Representative Example

To test the ability of different mapping schemes to represent various hydrocarbons we performed the CG MD simulations of 1-site water model and hybrid nonane (2-3-2-2, mapping scheme **2** shown in **Table 5.4**) model as a representative example. **Figure 5.6** shows the snapshots of simulation configurations of the hybrid nonane and water mixture (2000 CG hydrocarbon molecules and 6000 CG 1-site water molecules) at 0 ps and 1 μs, and the density profiles during the final 100 ns of the 1 μs of the CG MD simulations. It can be seen from the snapshots that the nonane(2-3-2-2)-water mixture segregates at 1 μs. The density of the nonane-rich phase, at ~6 - 54 Å, in **Figure 5.6 (c)** is ~0.71 g/cm^3, which is similar to its bulk density of 0.712 g/cm^3.[5] Note, in the nonane-rich phases, the density of water was ~0.0032 g/cm^3, which corresponds to ~ 25 CG water molecules. In the water-rich phase at 68 - 91 Å, the density of water was ~0.998 g/cm^3.

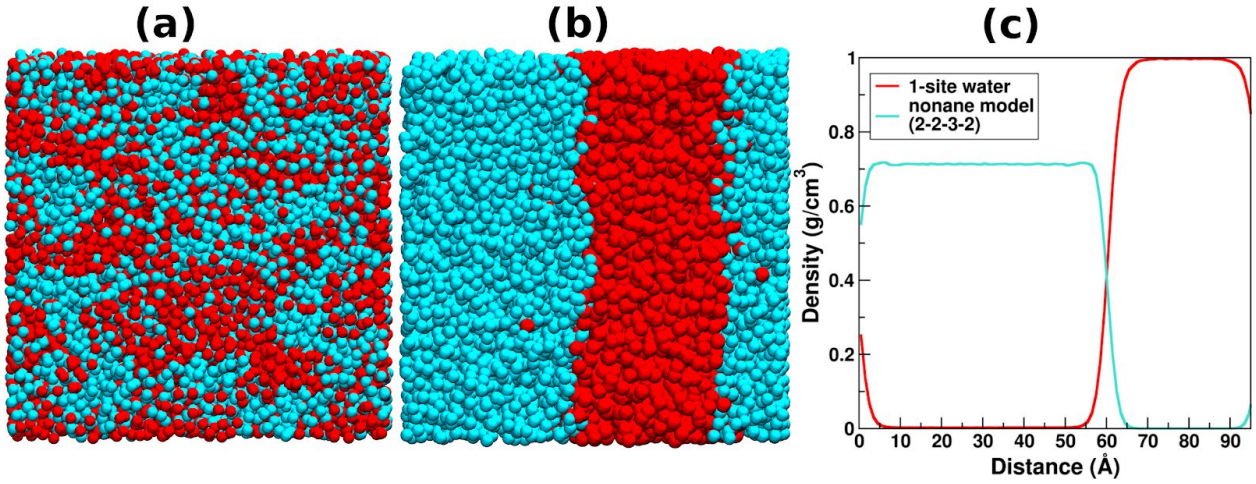

Figure 5.6 Snapshots of the configuration of water/nonane(2-3-2-2) mixture at **(a)** 0 ps and **(b)** 1 μs of the simulation trajectory, and **(c)** the density profiles of CG water and nonane(2-3-2-2) molecules by analysing the last 100 ns trajectory of the 1 μs simulation.

Results of the RDFs are shown in **Figure 5.7**, which was calculated by analyzing the simulation trajectories generated during initial 50 ps and final 100 ns of the 1 μs CG MD simulations. It can be seen that the first and second peaks at the RDF of W1-C2E decreased, while the first and second peaks in the RDF of C2E-C2E increased during the last 100 ns of the simulation. This indicates the segregation of the hybrid nonane model from the 1-site water model. Thus the FF parameters are transferable to hydrocarbon models represented by different mapping schemes.

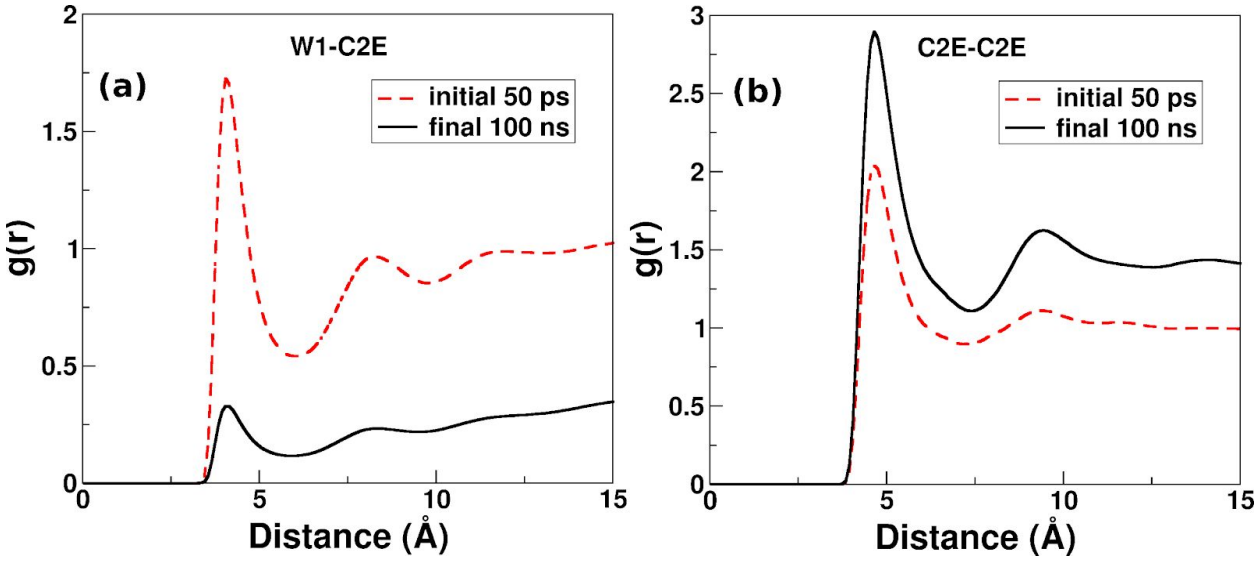

Figure 5.7 RDFs between **(a)** the water and the end beads in CG hybrid nonane (2-2-3-2) model: W1-C2E; **(b)** the end beads in the CG hybrid nonane model: C2E-C2E. Note, the RDF of W1-C2E during the final 100 ns increased to 1.0, and that of C2E-C2E during the final 100 ns decreased to 1.0 when the distance increased to ~50 Å.

5.3.5 Qualitative Comparison of Solubility of Pentane and Decane

In experiment, the solubility of hydrocarbons decreases as the chain length increases at room temperature.[28] To qualitatively compare the solubility of our hydrocarbon models with short and long chain lengths, the CG MD simulations of pentane(2-3)/water and decane(2-2-2-2-2)/water mixtures were performed. In each mixture, the number of CG 1-site water molecules and hydrocarbon molecules is 6000 and 2000, respectively. The density profiles of water and hydrocarbon are shown in **Figure 5.8**. In the hydrocarbon-rich phase of the pentane/water system (**Figure 5.8 - (a, c)**: 0 - 15 Å), the density of water was ~0.03 g/cm^3, while it was approaching 0.0027 g/cm^3 in the hydrocarbon-rich phase of the decane/water system. Similarly, the density of water in the water-rich phase of the pentane/water mixture (**Figure 5.8 - (b, d)**: ~45 - 55 Å) was 0.985 g/cm^3, smaller than that in water-rich phase of the decane/water mixture (**Figure 5.8 - (b, d)**: ~ 58 - 75 Å). Note, the number of pentane molecules in the water-rich phase was ~265, while that of the decane molecules was ~ 0. The above suggests that more water molecules are present in the pentane-rich phase, and more miscible with pentane molecules than with decane molecules.

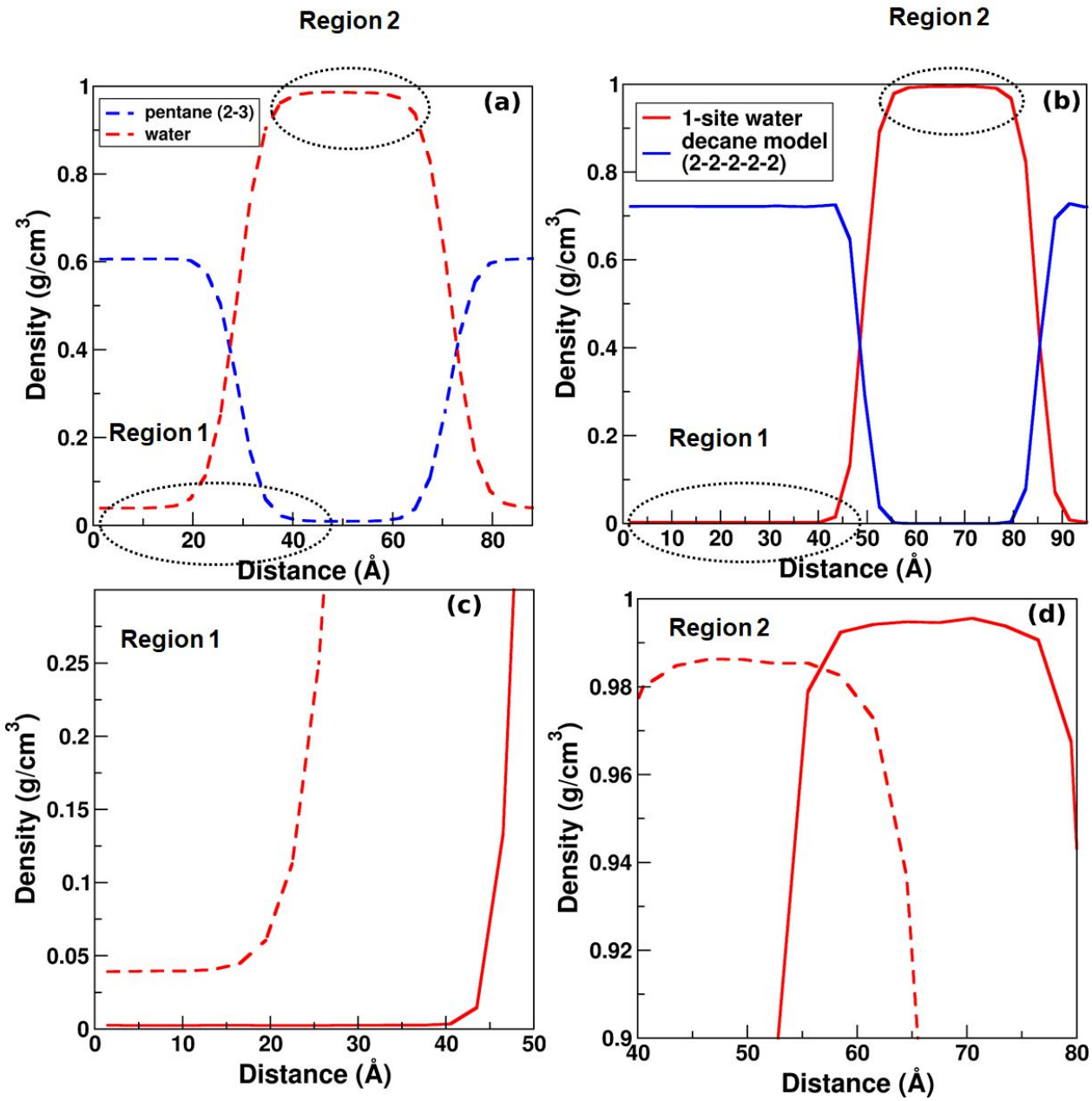

Figure 5.8 The density profiles of **(a)** pentane(2-3)/water mixture and **(b)** decane(2-2-2-2-2)/water mixture during the last 100 ns of 1-μs simulations. **(c)** the density of water in **Region 1** in **Figure 4.8 - (a)** and **(b)**, and **(d)** the density of water in **Region 2** in **Figure 4.8 - (a)** and **(b)**.

5.4 Guidance on Choosing Mapping Schemes

Based on this present chapter, Appendix B, and our recent publication on hydrocarbons [5], we propose the following design rules for CG hydrocarbon models. Note, these are general rules and there are a few models that do not follow these rules.

5.4.1 Bulk Properties of Hydrocarbons

(**i**) In general, using a pure mapping schemes such as 2:1 or 3:1 is recommended as these models predict the experimental bulk properties with good accuracy; (**ii**) When a hybrid mapping scheme must be used, *e.g.* for heptadecane, models with C3E beads at both ends should be avoided as the density is usually overestimated; (**iii**) Symmetry in the mapping scheme has minimal impact on most of the bulk properties of the hydrocarbons *e.g.* dodecan: 2-3-2-3-2 (symmetric) and 2-3-3-2-2 (asymmetric) mapping schemes.

5.4.2 Hydrocarbon-Water Systems

(**i**) The mapping scheme did not have a significant impact on the interfacial tension; (**ii**) The symmetric hydrocarbon models usually perform slightly better than the asymmetric ones in predicting the experimental Gibbs hydration free energy *e.g.* dodecan: 2-3-2-3-2 (symmetric) and 2-3-3-2-2 (asymmetric); **(iii)** Irrespective of a symmetric or asymmetric mapping scheme if the C3E beads are used at both ends of a hydrocarbon chain, the Gibbs hydration free energies are usually underestimated.

5.5 Conclusions

The nonbonded force-field (FF) parameters between recently developed coarse-grained (CG) 1-site water bead (W1) and hydrocarbon beads (C2E, C2M, C3E, C3M) were optimized by using the particle swarm optimization (PSO) algorithm. The 12-6 LJ nonbonded FF parameters between C2E and W1 beads, and between C2M and W1 beads were optimized to reproduce the Gibbs free energies of hydration of the CG decane and hexadecane models with 2:1 mapping scheme, and the interfacial tension of the CG decane model simultaneously at 300 K. Similarly, the FF parameters between C3E and W1 beads, and between C3M and W1 beads were optimized in order to reproduce the Gibbs hydration free energies of CG nonane and pentadecane models with 3:1 mapping scheme, and the interfacial tension of the CG nonane model at 300 K simultaneously. By employing these new optimized FF parameters, the Gibbs hydration free energies of the decane and hexadecane models with 2:1 mapping scheme were within 8 and 12 % of their corresponding experimental values, and their interfacial tensions were also in good agreement with the experimentally measured data at 300 K. Similarly, the nonane and

pentadecane models with 3:1 mapping schemes could predict the Gibbs hydration free energies and interfacial tensions at 300 with good accuracy.

Furthermore, simulations of CG hydrocarbon and water homogeneous mixture were performed to test the ability of the FF parameters in predicting the phase segregation of CG hydrocarbon and water molecules. The visual analysis, density profiles of hydrocarbons and water, and the RDFs for the mixture showed a clear phase segregation of CG hydrocarbon and water molecules. Moreover, the transferability of the optimized FF parameters were validated by predicting the Gibbs hydration free energies and interfacial tensions of CG hydrocarbons models with different chain lengths from pentane to heptadecane. The majority of Gibbs free energies hydration and interfacial tensions of other CG hydrocarbon models were within 15 % and 10 % of the experimental data, respectively, suggesting that these new FF parameters between hydrocarbons and water are transferable. In the near future, we will be utilizing these newly developed interaction parameters between the CG models of hydrocarbons and water to perform simulations of various polymers and biomaterials.

References

(1) Schnabel, T.; Vrabec, J.; Hasse, H. Unlike Lennard-Jones Parameters for Vapor-Liquid Equilibria. *arXiv [physics.chem-ph]*, 2009.

(2) Waldman, M.; Hagler, A. T. New Combining Rules for Rare Gas van Der Waals Parameters. *J. Comput. Chem.* **1993**, *14* (9), 1077–1084.

(3) Song, W.; Rossky, P. J.; Maroncelli, M. Modeling Alkane+perfluoroalkane Interactions Using All-Atom Potentials: Failure of the Usual Combining Rules. *J. Chem. Phys.* **2003**, *119* (17), 9145–9162.

(4) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.

(5) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153.

(6) Moore, J. W.; Wellek, R. M. Diffusion Coefficients of N-Heptane and N-Decane in N-Alkanes and N-Alcohols at Several Temperatures. *J. Chem. Eng. Data* **1974**, *19* (2), 136–140.

(7) Rolo, L. I.; Caço, A. I.; Queimada, A. J.; Marrucho, I. M.; Coutinho, J. A. P. Surface Tension of Heptane, Decane, Hexadecane, Eicosane, and Some of Their Binary Mixtures. *J. Chem. Eng. Data* **2002**, *47* (6), 1442–1445.

(8) Liu, H.; Zhu, L. Excess Molar Volumes and Viscosities of Binary Systems of Butylcyclohexane with N-Alkanes (C7 to C14) at T = 293.15 K to 313.15 K. *J. Chem. Eng. Data* **2014**, *59* (2), 369–375.

(9) Ríos, R.; Ortega, J.; Fernández, L.; de Nuez, I.; Wisniak, J. Improvements in the

Experimentation and the Representation of Thermodynamic Properties (iso-P VLE and yE) of Alkyl Propanoate + Alkane Binaries. *J. Chem. Eng. Data* **2014**, *59* (1), 125–142.
(10) Wagner, W.; Pruß, A. The IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use. *J. Phys. Chem. Ref. Data* **2002**, *31* (2), 387–535.
(11) Eichenberger, A. P.; Huang, W.; Riniker, S.; van Gunsteren, W. F. Supra-Atomic Coarse-Grained GROMOS Force Field for Aliphatic Hydrocarbons in the Liquid Phase. *J. Chem. Theory Comput.* **2015**, *11* (7), 2925–2937.
(12) DeVane, R.; Klein, M. L.; Chiu, C.-C.; Nielsen, S. O.; Shinoda, W.; Moore, P. B. Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parameterization Using Experimental Data. *J. Phys. Chem. B* **2010**, *114* (19), 6386–6393.
(13) Bejagam, K. K.; Balasubramanian, S. Supramolecular Polymerization: A Coarse Grained Molecular Dynamics Study. *J. Phys. Chem. B* **2015**, *119* (17), 5738–5746.
(14) Fiorin, G.; Klein, M. L.; Hénin, J. Using Collective Variables to Drive Molecular Dynamics Simulations. *Mol. Phys.* **2013**, *111* (22-23), 3345–3362.
(15) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128* (14), 144120.
(16) Allen, M. P.; Tildesley, D. J. Molecular Simulation of Liquids. *Clarendon, Oxford* **1987**.
(17) Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016–2027.
(18) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database. *University of Minnesota, Minneapolis, MN* **2012**.
(19) Zeppieri, S.; Rodríguez, J.; López de Ramos, A. L. Interfacial Tension of Alkane + Water Systems. *J. Chem. Eng. Data* **2001**, *46* (5), 1086–1088.
(20) Amaya, J.; Rana, D.; Hornof, V. Dynamic Interfacial Tension Behavior of Water/Oil Systems Containing In Situ-Formed Surfactants. *J. Solution Chem.* **2002**, *31* (2), 139–148.
(21) Sanemasa, I.; Wu, J.-S.; Toda, K. Solubility Product and Solubility of Cyclodextrin Inclusion Complex Precipitates in an Aqueous Medium. *BCSJ* **1997**, *70* (2), 365–369.
(22) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760.
(23) Herdes, C.; Ervik, Å.; Mejía, A.; Müller, E. A. Prediction of the Water/oil Interfacial Tension from Molecular Simulations Using the Coarse-Grained SAFT-γ Mie Force Field. *Fluid Phase Equilib.* **2017**. https://doi.org/10.1016/j.fluid.2017.06.016.
(24) Bereau, T.; Kremer, K. Automated Parametrization of the Coarse-Grained Martini Force Field for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11* (6), 2783–2791.
(25) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. Polarizable Empirical Force Field for Alkanes Based on the Classical Drude Oscillator Model. *J. Phys. Chem. B* **2005**, *109* (40), 18988–18999.
(26) Gyawali, G.; Sternfield, S.; Kumar, R.; Rick, S. W. Coarse-Grained Models of Aqueous and Pure Liquid Alkanes. *J. Chem. Theory Comput.* **2017**, *13* (8), 3846–3853.
(27) Xiao, H.; Zhen, Z.; Sun, H.; Cao, X.; Li, Z.; Song, X.; Cui, X.; Liu, X. Molecular Dynamics Study of the Water/n-Alkane Interface. *Sci. China Chem.* **2010**, *53* (4), 945–949.
(28) Mackay, D.; Shiu, W.-Y.; Ma, K.-C.; Lee, S. C. *Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals*; CRC Press, 2006.

CHAPTER 6
DEVELOPMENT OF AN ACCURATE COARSE-GRAINED MODEL OF POLY(ACRYLIC ACID) IN EXPLICIT SOLVENTS

This work presented in this chapter is reported from [An, Y., Singh, S.; Bejagam, K. K., Deshmukh, S. A. Development of an Accurate Coarse-Grained Model of Poly(acrylic acid) in Explicit Solvents , Macromolecules, 2019, 52 (13), 4875-4887], with the permission of AIP Publishing.

Abstract: Understanding the effect of solvent on the polymer conformations is a fundamental problem in materials science and engineering. Here, we have developed, *first of its kind*, a coarse-grained (CG) model of poly(acrylic acid) (PAA), which can reproduce its experimental glass transition temperature (T_g), and conformation of a single chain in the presence of explicit solvents along with capturing the structure of solvents at the PAA-solvent interfaces. The PAA model was based on a CG model of propionic acid, an analogue of the PAA monomer. Accuracy of both the propionic acid and PAA models was validated by employing uncertainty quantifications. The cross-interaction parameters between CG PAA and 1-site water model, and between CG PAA and DMF models were optimized to reproduce the radius of gyration (R_g) of an all-atom 30-monomer (30-mer) PAA chain in pure solvents. These interaction parameters were further used to explore the PAA conformation in the presence of binary mixtures of water and DMF with different compositions. A PAA chain was in a globule-like and a coil-like state in binary solvents with low and high mass fractions of DMF, respectively. Moreover, the local structure of solvent suggests that even at a low mass fraction of DMF in a binary solvent, there is an enhanced ordering of DMF molecules at the polymer-solvent interface. Furthermore, an increase in the coordination number of DMF molecules within the first solvation shell of PAA suggests that DMF molecules form a shielding layer and protect PAA from water molecules. These results are in excellent agreement with the results of all-atom MD simulations.

6.1 Introduction

Poly(acrylic acid) (PAA) is one of the most commonly used hydrophilic polymers and has a variety of applications, for example, hydrogen gels, ion exchange resins, adhesives and

detergents. Conformations of PAA are affected by its surrounding environments, such as solvents, temperatures, ect. Understanding the effects of solvents and temperatures could help control its conformation, and then further achieve desired properties. Computational studies of the conformation of PAA have been reported previously. Adamczyk *et al.* studied the structure of PAA in electrolyte solutions (NaCl) by using numerical simulations. The effective length of the PAA decreased as the ionic strength increased.[1] MD simulations have also been performed to study the conformation of PAA in aqueous solutions. Sulatha and Natarajan used all-atom MD simulations to investigate the conformations of PAA in dilute aqueous solution, as a function of charge density. They found that the radius of gyration of the polymer increased as the charge density increased.[2] Reith et al. optimized the CG MD force-field (FF) parameters for PAA in the form of its sodium salt.[3] These studies of PAA model were focused on its behaviour in aqueous solutions. However, considering the various external environments for PAA applications, transferable and accurate CG PAA models across different conditions are to be developed. Solvents and temperature are two significant factors that affect PAA's structures and properties. Therefore, a temperature- and solvent-transferable CG PAA model will be developed in this chapter.

The goal of this present study is two-fold: (1) to develop an accurate CG model of PAA that can reproduce its bulk properties, and (2) to investigate the conformations of PAA in explicit solvents using this CG model. The backbone of the PAA chain is represented by the CG hydrocarbon beads developed in the previous chapter, while the side chain is symbolized by a new CG bead COOH. To optimize the FF parameters for the COOH bead, the CG model of propionic acid monomers consisting of one hydrocarbon bead C2E and one COOH bead was optimized by using PSO. With the newly developed CG model of propionic acid, the new CG model of PAA was validated to reproduce the experimentally measured glass transition temperature (T_g). Secondly, the cross-interaction between the CG PAA and 1-site water model along with that between the CG PAA model and DMF were developed to reproduce the distribution of the radius of gyration (R_g) of an all-atom PAA 30-mer chain in water and DMF, respectively. The PAA model developed here is more realistic, as they use explicit solvent models, rather than the existing models that use implicit solvent models.[3,4]

6.2 Model Development

Initial steps in the CG model development of PAA involved the development of the CG model of propionic acid, an analogue of PAA's monomer (see **Figure 6.1 (a)**). This model consists of a C2E bead and a COOH bead, that represent a CH_3-CH_2- group and a -COOH group, respectively. Both the C2E and COOH beads were charge neutral, and were bonded through a harmonic bond. The nonbonded interactions between different molecules were captured by the 12-6 LJ potential. The FF potential describing the bonded and nonbonded interactions in a CG propionic acid is shown in **Equation 6.1**.

$$E_{pot} = K_b(b - b_0)^2 + 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] \qquad \textbf{.....Equation 6.1}$$

Where, K_b is the bond force constant, b_0 is the equilibrium bond length, ε_{ij} is the depth of the potential well and represents the strength of interactions between two beads i and j, σ_{ij} is the finite distance at which the inter-particle potential is zero, and r_{ij} is the distance between two beads.

The nonbonded interaction parameters for C2E beads were developed in our recent work to represent hydrocarbons.[5] To accomplish the development of the CG propionic acid model, the ε and σ values of the COOH bead, and the bonded parameters K_b and b_0 between C2E and COOH beads were determined.[5] The b_0 was estimated to reproduce the distribution of the bonds between the mapped C2E and COOH beads from the all-atom simulation trajectory of propionic acid molecules, which is 2.5 Å shown in **Figure C1** of Appendix C. For the other three parameters, K_b, ε, and σ, particle swarm optimization (PSO) integrated MD simulation method was utilized as reported in our previous work to accelerate the optimization process.[5–10] More details on the implementation of PSO to develop the CG propionic acid model can be found in Appendix C. The optimized FF parameters are shown in **Table C1** of Appendix C.

The COOH bead developed in the present study, and C2E and C2M beads developed in our recent study were used to construct the CG model of PAA (see **Figure 6.1 (b)**).[5] The harmonic bond and angle potentials were employed to represent the bonded interactions in PAA. The bond potential parameters for the backbone are adopted from the hydrocarbon models in refs [5]. The FF parameters for the COOH-C2M bond were the same as the COOH-C2E bond in the propionic acid model. All the bond potential parameters are listed in **Table C2** of Appendix C.

The angle potential parameters for the COOH-C2M-C2M were tuned to cover the distribution of the mapped COOH-C2M-C2M from all-atom MD simulations, and to reproduce the experimental density at 300 K, and T_g of PAA bulk. The optimized angle parameters are shown in **Table C3** of Appendix C. The nonbonded interactions were represented by the 12-6 LJ potential. Beads separated by two bonds interacted through the nonbonded interactions, in addition to angle potentials. The nonbonded FF parameters used to model the CG PAA chain are listed in **Table C4** of Appendix C. This PAA CG model was used to study its bulk properties and conformation change in different solvents.

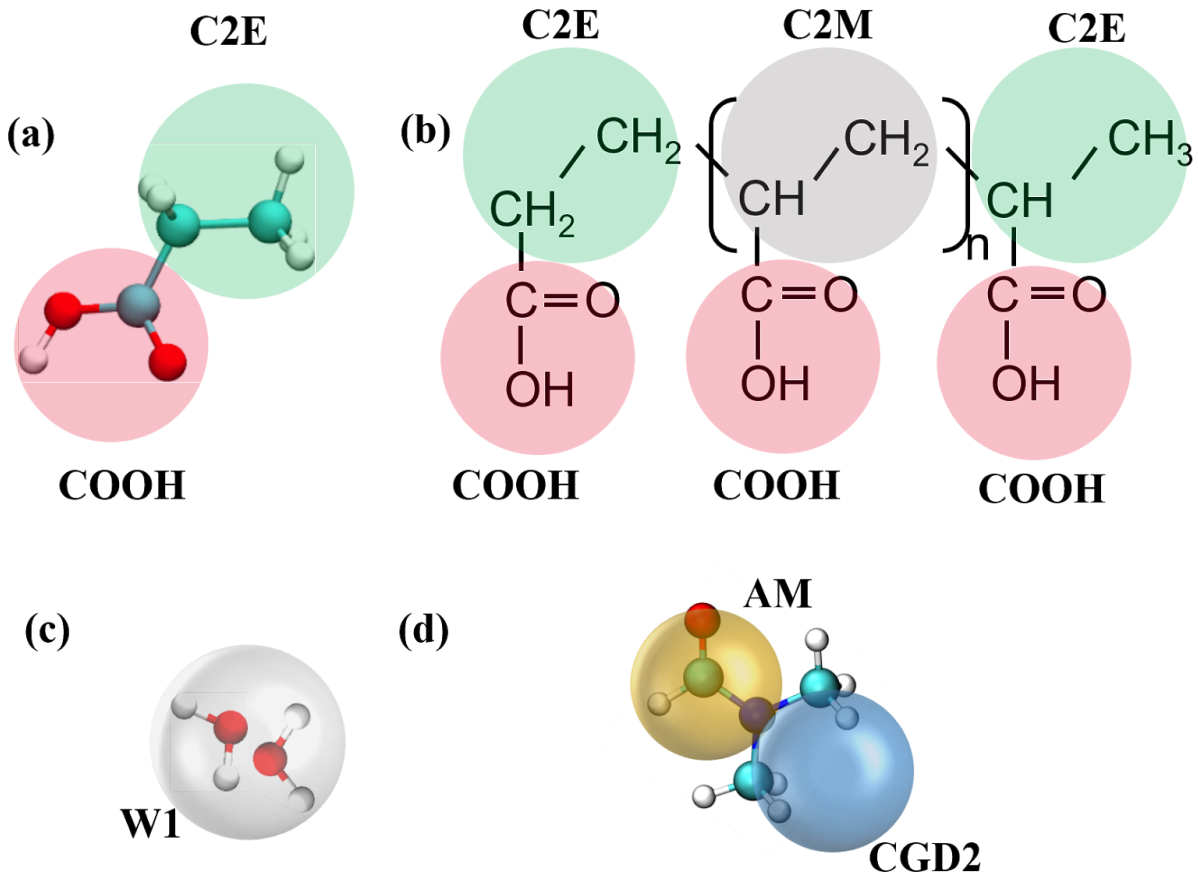

Figure 6.1 (a) CG propionic acid model**, (b)** CG PAA model, (**c)** CG 1-site water model , and **(d)** CG DMF model. The gray, green, and red filled circles represent C2M, C2E, and COOH beads, respectively. C2E and C2M beads were adopted from reference [5]. The white, yellow and blue spheres represent the W1, AM, and CGD2 beads from reference [6,7].

6.3 FF Parameters between CG PAA and Solvent Models

To explore the effect of solvent on the PAA conformation, we developed the FF parameters to describe the interactions between PAA and water, and between PAA and DMF. CG MD simulations of PAA in pure water, pure DMF, and their binary mixtures were performed. The CG 1-site water and DMF models developed in our previous studies were employed as explicit solvent models. The 1-site water model represents water molecules by 2:1 mapping scheme (two all-atom water molecules combined to form one bead, W1 in **Figure 6.1 (c)**). The CG DMF model consists of an AM bead and a CGD2 bead, which encompass the amide group and the two methyl groups in DMF, respectively, as shown in **Figure 6.1 (d)**. The FF parameters of the 1-site water and CG DMF models were optimized to reproduce their corresponding experimental properties by particle swarm optimization (PSO) and PSO-ANN (artificial neural network assisted PSO) in refs[6,7], respectively. The optimized non-bonded FF parameters are shown in **Table C4** of Appendix C, and the properties predicted by both these models can be found in **Table C5** of Appendix C.

The nonbonded interaction parameters between CG PAA model and water as well as CG PAA and DMF beads were optimized based on the R_g distributions of an all-atom PAA chain in water and DMF. Specifically, the values of ε from 12-6 LJ potential between the COOH bead of PAA and W1 bead of the 1-site water were calibrated systematically to reproduce the R_g distribution of a single all-atom PAA 30-mer in water at 300 K. In the calibration, the range of the ε value between COOH and W1 beads was from 1.1958 kcal/mol to 1.4 kcal/mol as shown in **Figure C4** of Appendix C. The lower limit was obtained by the Bertholet combining rule. It was found in **Figure C4 - (a)** that the R_g distribution of the CG model showed a good agreement with that of the all-atom model when ε[COOH-W1] was 1.35 kcal/mol. Note, the value of σ between the COOH bead and the W1 bead was determined by using the Lorentz combining rule. Similarly, the ε value between the COOH bead of PAA and the AM bead of DMF was optimized to reproduce the R_g distribution of a single all-atom PAA 30-mer model in pure DMF, which was 1.35 kcal/mol as shown in **Figure C4 - (b)** of Appendix C. More details on this optimization process to determine the PAA-solvent FF parameters are discussed in Appendix C. The aforementioned optimized ε[COOH-W1] and ε[AM-COOH] values are shown in **Section 6.4.4**. The ε and σ values for 12-6 LJ potential that define the cross-interactions between C2E or C2M beads of PAA and 1-site water (W1) model were adopted from our recent study. The

experimental Gibbs hydration free energies of hydrocarbons and the experimental interfacial tensions at the hydrocarbon/water interfaces could be predicted by these FF parameters with good accuracy. The ε and σ values between C2M beads of PAA and AM beads of DMF were obtained by the Lorentz-Berthelot (LB) combining rules. Note, the LB combining rules were applied to these parameters because the C2M or C2E beads and CGD2 beads all represented two carbon atoms and the hydrogen atoms bonded with them and had similar ε and σ values from LJ potential.

6.4 Results and Discussion

6.4.1 Uncertainty Quantification of the Properties of the CG Propionic Acid Model

In order to evaluate the ability of the CG propionic CG model developed in the present study in estimating experimentally measured properties, we performed UQ of properties predicted by this model using bootstrapping method. Detailed description of the UQ analysis could be found in **Section 3.3.8** of **Chapter 3**. Results of bootstrapping with 1000 resampled sets obtained by using the original dataset are shown in **Figure 6.2**. The black vertical lines show the position of mean values of the four properties, and the regions between two red lines represent the 95 % confidence intervals. It can be seen that for density, self-diffusion coefficient, enthalpy of vaporization and surface tension, the mean values are 0.99189 g/cm^3, 1.084 x 10^{-9} m^2/s, 10.71 kcal/mol, and 44.91 mN/m, respectively. The 95 % confidence intervals are 0.99185 - 0.99192 g/cm^3, 1.077 x 10^{-9} - 1.090 x 10^{-9} m^2/s, 10.70 - 10.72 kcal/mol, and 44.75 - 45.07 mN/m, respectively. The experimental density, self-diffusion coefficient and enthalpy of vaporization is 0.99 g/cm^3, 1.0 x 10^{-9} m^2/s, and 13.7 kcal/mol, respectively.[11–13] Thus, the properties predicted by the CG propionic acid model are in good agreement with the experimental results. The surface tension, however, deviated from the experimentally measured data by ~80 %. This can be attributed to the strong interactions between COOH beads with themselves and between COOH and C2E beads. Similar deviation has been reported in the surface tension of other CG models, like the CG dimethyl sulfoxide (DMSO) model and the MARTINI water models.[14,15]

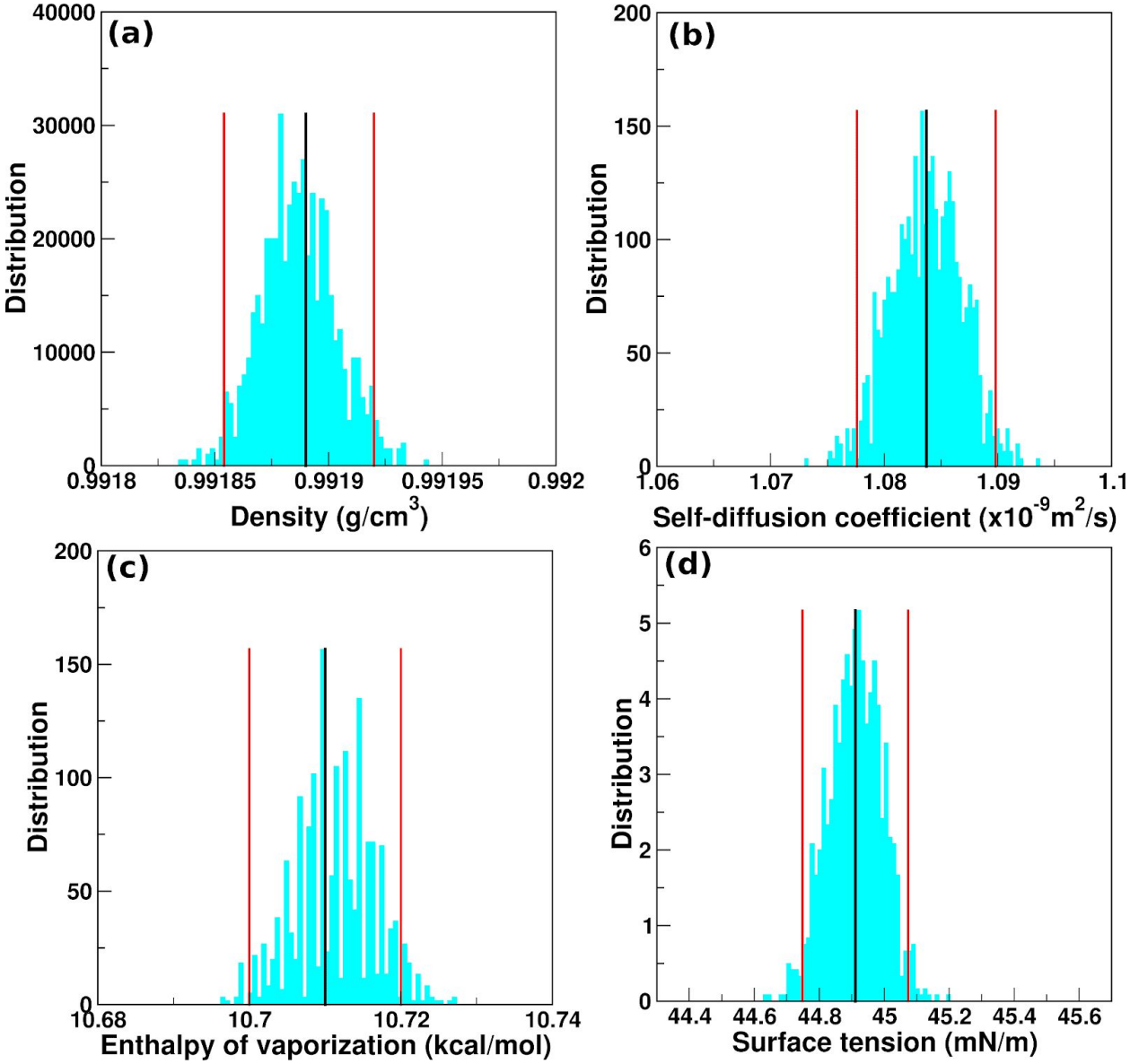

Figure 6.2 Bootstrapping resampling results of **(a)** density, **(b)** self-diffusion coefficient, **(c)** enthalpy of vaporization, and **(d)** surface tension of the CG propionic acid model. Black and red lines represent the positions of mean values, the lower and upper boundaries of 95 % confidence intervals, respectively.

6.4.2 Structure and Properties of the CG PAA Bulk

Bond and Angle distribution in the CG PAA Model

To investigate the structure of the 30-mer CG PAA model in bulk, the angle and bond distributions of the CG PAA model were calculated at 300 K. More details on simulation set-up

are discussed in **Section 3.4** of **Chapter 3**. The comparison of these distributions with those from the all-atom mapped trajectory is shown in **Figure 6.3**. **Figure 6.3 (a)** exhibits the distributions of the CG and all-atom mapped C2M-C2M-COOH angles. The all-atom mapped C2M-C2M-COOH angle shows a broad distribution from 80 ^{o} to 180 ^{o} with two peaks at ~100 ^{o} and ~160 ^{o}. In the case of angle distribution, it is well known that the angles described by harmonic potentials in the CG model cannot reproduce the bimodal angle (or bond) distribution obtained from the mapped all-atom trajectory. Hence, here, we aimed to have as much overlap as possible for the angle distribution between all-atom mapped trajectory and CG model. Therefore, we chose the position of the large peak at ~100 ^{o} as the equilibrium angle value for the CG PAA model. The CG C2M-C2M-COOH angle shows distribution from 95 ^{o} to 140 ^{o}, and the average value was ~111.8 ^{o}. This value is slightly higher than the equilibrium angle value of 100 ^{o}. This could be because the beads connected by two bonds interact with each other through nonbonded interactions in addition to the angle potential. This may result in the interference of the nonbonded potentials between various beads present in the C2M-C2M-COOH angle on its distribution. For example, the nonbonded interactions between the C2M and COOH bead may result in stretching of the C2M-C2M-COOH angle, based on their σ values. The force constant tuned to reproduce this angle distribution was used to calculate the density of PAA, which is in the range of ~1.22 to ~1.43 g/cm^3 (average of ~1.32 g/cm^3), and the experimental T_g value of 378 K. Thus, the angle force constant that resulted in a large overlap for the COOH-C2M-C2M angle, the density of the CG model within 2 % of the target density, and the experimental T_g value (within 2 %) was selected to perform further study. The distribution of the mapped C2M-C2M bond from the all-atom trajectory shows two peaks at ~2.3 Å and ~2.65 Å (**Figure 6.3 (b)**). The peaks at ~2.3 Å, and ~2.65 Å in the distribution of the mapped C2M-C2M bond correspond to the cis- and trans-configuration underlying the all-atom simulations, respectively. The harmonic bond potential in the CG polymer model couldn't reproduce this bimodal distribution, and showed only one peak at 2.57 Å. However, the CG C2M-C2M bond distribution has a large overlap with that of the mapped C2M-C2M bond from ~2.4 Å to ~2.8 Å, suggesting that the CG model could capture more of the trans-configuration. The distribution of the CG C2M-COOH bond exhibits one peak at ~2.52 Å in **Figure 6.3 (c)**, which is slightly shifted to the right compared with the average value obtained from the mapped COOH and C2M beads, ~2.45

Å. Note that the bond and angle parameters in the CG PAA models in the present study were not developed to reproduce these structural features described in **Figure 6.3 (a)** and **(b)**.

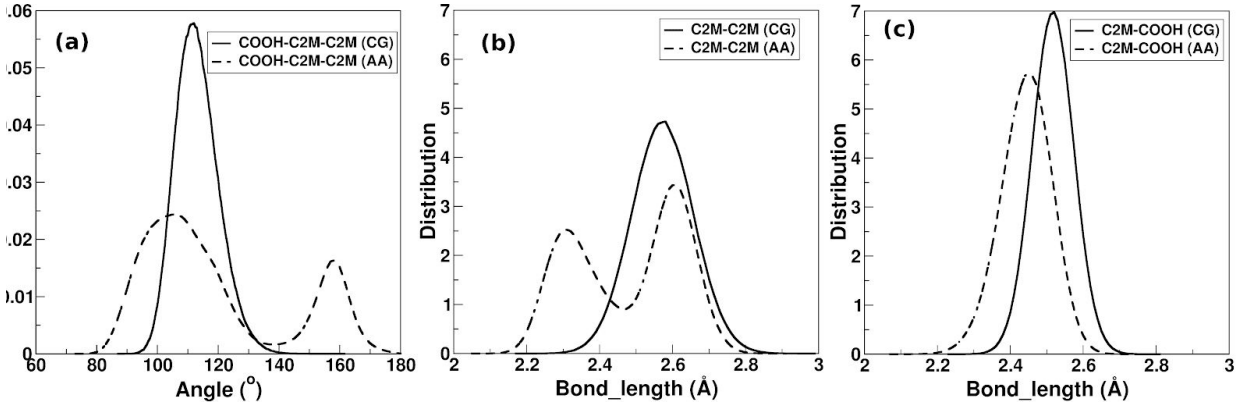

Figure 6.3 The distributions of **(a)** the angle COOH-C2M-C2M, **(b)** the bond C2E-C2M, and **(c)** the bond C2M-COOH from the CG PAA 30-mer simulations compared with those from the all-atom mapped trajectory at 300 K.

Furthermore, to compare the structure of CG PAA models with that of the all-atom model, the RDFs of various bead pairs in the PAA 30-mer model from CG and all-atom MD simulations were calculated. As can be seen in **Figure 6.4 (a)**, from 5 to 9 Å, there is a broad peak in the RDF of all-atom mapped C2M-C2M bead pair, while there were two small peaks in the same range in the RDF of the CG bead pairs. In the case of the RDF of the C2M-COOH bead pair in **Figure 6.4 (b)**, we find the first peak for the CG bead pair was shifted to left at ~4.4 Å, as compared with that of the all-atom mapped bead pair at ~5.4 Å. As for the RDF of all-atom mapped COOH-COOH pairs, it exhibits the first peak at ~4.0 Å with a shoulder at ~3.0 Å. The RDF of the CG COOH-COOH bead pair shows the first peak at ~4.4 Å (see **Figure 6.4 (c)**), which is much higher than that of the first peak in RDF of the all-atom mapped bead pair. This can be attributed to the strong COOH-COOH interactions. . Note, the first peak in the RDF of the COOH-COOH in the CG propionic acid model is also much higher than that from the mapped bead pairs (see **Figure C2 (c)** of Appendix C). To further evaluate the number of beads in the first and second shells of a given bead, we calculated the coordination numbers from all-atom mapped trajectory and CG MD trajectory. As shown in **Figure 6.4 (d - e)**, the coordination numbers for the three bead pairs from all-atom mapped trajectories and CG MD simulations show a reasonable match in the distance range from 0 Å to 10 Å. For example, in **Figure 6.4 (d)**

at ~10 Å, which is the position of the first valley in RDF of the mapped C2M-C2M bead pair, the coordination numbers were ~34 and ~35 from the all-atom mapped trajectory and CG MD simulations, respectively. Similarly, for the C2M-COOH pair the coordination numbers of ~10.1 and ~9.8 were observed below ~7.0 Å for all-atom mapped and CG MD trajectory, respectively. In the case of the COOH-COOH bead pair, ~5.0 and ~6.3 beads were found within a distance of ~5.7 Å for all-atom mapped and CG MD trajectories, respectively. The coordination number for the CG COOH-COOH bead pair is larger, which is consistent with the higher first peak in its RDF.

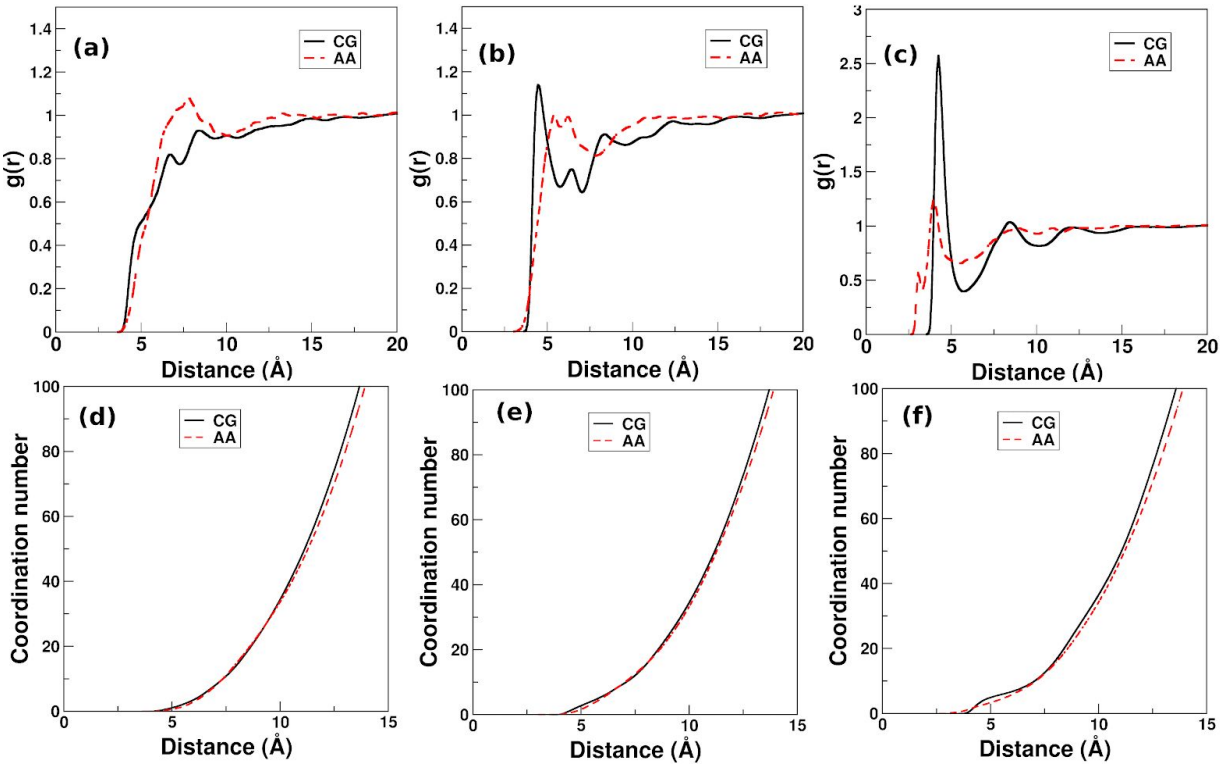

Figure 6.4 The RDFs between **(a)** C2M-C2M, **(b)** C2M-COOH, and **(c)** COOH-COOH bead pairs, and coordination numbers of **(d)** C2M-C2M, **(e)** C2M-COOH and **(f)** COOH-COOH from CG and all-atom mapped simulation trajectories.

Density and Glass Transition Temperature of PAA

MD simulations of 50 and 500 CG PAA chains of 30-mer were performed in NPT ensemble at the temperature ranging from 280 K to 500 K, to investigate the T_g of this new CG model of PAA (see **Figure 6.5**). At 300 K, the densities of the PAA systems with 50 chains and

500 chains were both ~1.304 g/cm^3 and showed a good agreement with the reported experimental range of ~1.22 to ~1.43 g/cm^3.[16,17] As the temperature increased, the density of the the 50-chain PAA system decreased linearly in two regions, 280 K - 386 K, and 386 K - 500 K, as shown in **Figure 6.5 (a)**. A characteristic change in the slope at ~386 K can be observed, which corresponds to the T_g of the 50 CG PAA 30-mer chains. Similarly, the system of 500 PAA chains showed a T_g value of ~383 K in **Figure 6.5 (b)**. Both of them are in good agreement with the experimental value of ~378.15 K.[18]

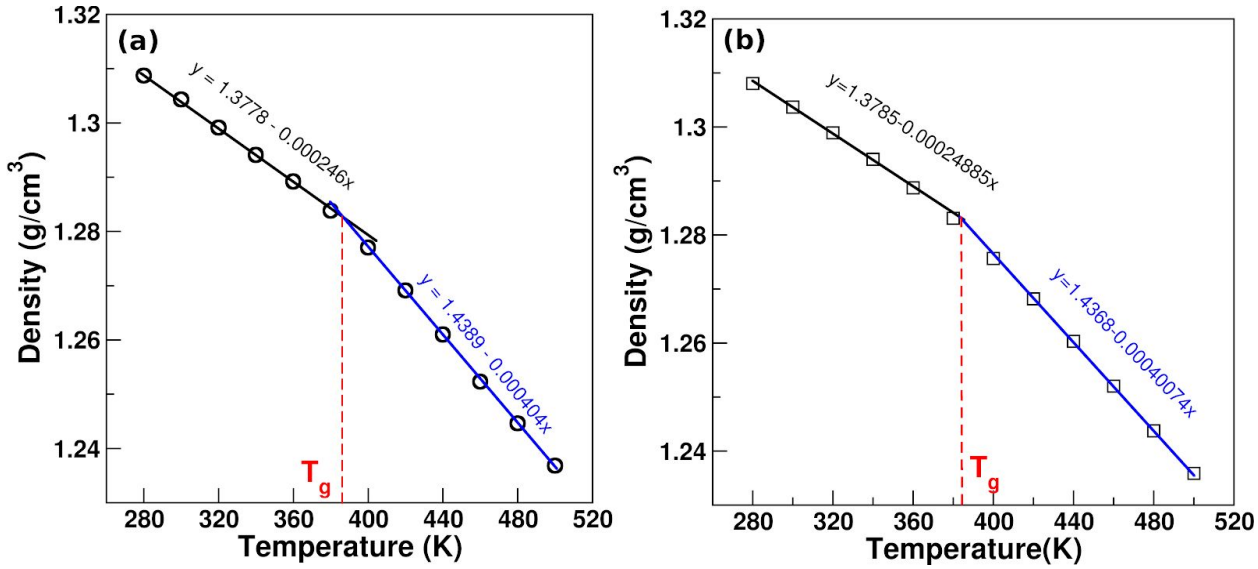

Figure 6.5 Densities of CG PAA systems with **(a)** 50 PAA 30-mer chains, T_g = ~386 K, and **(b)** 500 PAA 30-mer chains, T_g = ~383 K. The x and y in the equations represent temperature and density, respectively. The black and blue lines show the linearly fitted values of the density of PAA below and above the T_g, respectively. The standard deviation for each density point is ~0.0002 - 0.001 g/cm^3.

6.4.3 Uncertainty Quantification of the Density of the CG PAA Model

To further validate the accuracy of the density values obtained at different temperatures, we performed the UQ at selected temperatures. Specifically, CG MD simulations were performed below (300 K and 360 K), and above (420 K and 460 K) the T_g of PAA. At each temperature, a total of 48 simulations of 50 PAA 30-mer chains were conducted for 100 ns. The density distributions obtained by bootstrapping are shown in **Figure 6.6 (a - d)**. As can be seen that at 300 K, 360 K, 420 K, and 460 K, the mean values of density are 1.3043, 1.2895, 1.2691,

1.2527 g/cm^3, and the 95 % confidence intervals are all in narrow ranges, 1.3042 - 1.3044, 1.2893 - 1.2897, 1.2688 - 1.2694, 1.2526 - 1.2628 g/cm^3, respectively.

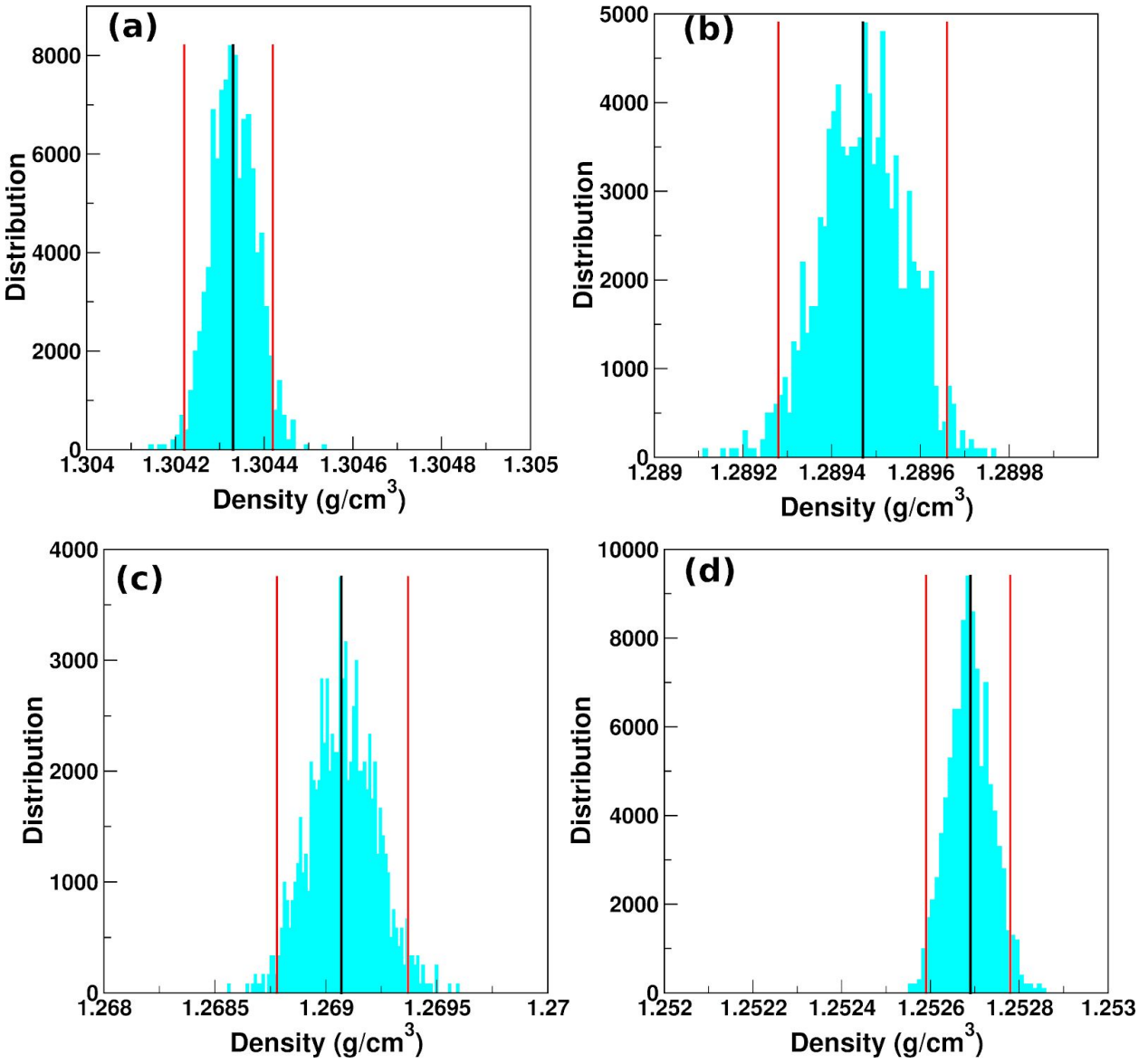

Figure 6.6 Bootstrapping resampling results of the density of PAA at **(a)** 300 K, **(b)** 360 K, **(c)** 420 K, and **(d)** 460 K. Black solid line represent the positions of mean values, and red solid lines show the lower and upper boundaries of 95 % confidence intervals.

6.4.4 Conformation of the CG Polymer Model in Solvents

Binary Solvent Mixtures of CG Water and DMF models

Prior to investigating the conformation of a single CG PAA chain in binary mixtures of water and DMF, we performed simulations of only binary mixtures (without PAA chain) to validate the interactions between CG 1-site water and DMF models, which were adopted from references [6,7]. Specifically, simulations of the mixtures with different mass fractions of DMF in water were conducted to obtain their densities to be compared with experimental data. It can be seen in **Table 6.1** that the densities calculated for mixtures with different mass fractions of DMF were comparable with the experimental data.[19] For example, when the mass fractions of DMF were 29.07, 44.29, and 82.40 wt %, the densities of the mixture were ~0.9897, ~0.9798, and ~0.9532 g/cm^3, respectively. They are within 2.5 % of the corresponding experimental values.[19] To further validate the mixtures of the 1-site water and CG DMF models, the Gibbs hydration free energy profile of the CG DMF model was determined by using the adaptive biasing force (ABF) method in reference [20]. The Gibbs hydration free energy of the CG DMF model was -5.5 kcal/mol, comparable with the reported simulation result of -6.93 kcal/mol.[21]

Table 6.1 The density of solvent mixtures of different mass fractions of DMF. The experimental data is interpolated by using the reported data in reference [19].

mass fraction of DMF (wt%)	density (g/cm^3)	experiment (g/cm^3)	relative error%
0	1.002±0.000	0.997	0.5
10.11	1.001±0.001	0.9962	0.5
29.07	0.9897±0.001	0.9968	0.7
44.29	0.9798±0.001	0.9968	1.7
65.82	0.9646±0.000	0.9898	2.5
82.40	0.9532±0.001	0.9736	2.1
92.35	0.9472±0.001	0.9582	1.1
100	0.9430±0.000	0.9440	0.1

CG PAA in Pure Water and Pure DMF

To tune the interactions between PAA and 1-site water model or DMF model, we have performed simulations of all-atom and CG PAA 30-mer models in pure water and pure DMF. It

has been reported that a fully protonated PAA chain collapses (globule-like state) in water at 300 K due to the strong intramolecular hydrogen bonds.[22,23] On the other hand, PAA is in a coil-like state in DMF.[24] Note, the root mean square deviation (RMSD) and principal component analysis (PCA) of the all-atom PAA chain in pure water and pure DMF clearly suggests that the all-atom simulations performed in this study have sampled the space effectively (See **Figures C8** and **C9** in Appendix C).

The values of the optimized interaction parameters to reproduce a globule-like and a coil-like structure of the CG PAA model are shown in **Table 6.2**. Note, the σ values were obtained by Lorentz rule.[25] The other FF parameters could be found in **Table C4** of Appendix C. As stated in **Section 6.3**, the optimized ε value between the COOH bead and the W1 bead was 1.35 kcal/mol. With this optimized value, the R_g distribution of a CG 30-mer PAA polymer was in the range of 6 to 14 Å with a peak at ~8.1 Å (globule-like state), as shown in **Figure 6.7 (a)**. The difference in the positions of the peaks in the R_g distributions of the CG and all-atom model (~8.8 Å) is less than 7%. The peak of the R_g distribution of CG PAA in DMF was observed at ~13.3 Å, which was in good agreement with the averaged R_g value of ~13.8 Å of a single all-atom PAA 30-mer in pure DMF (see **Figure 6.7 (b)**). However, The characteristic of the multiple peaks in the R_g distributions of the all-atom PAA model in DMF could not be captured by the CG PAA model. This could be because the CG MD simulations shows smooth free energy landscapes compared with all-atom simulations.[26,27] In other words all-atom model can be trapped in a local minima easily, and lead to some unique conformations that CG model might not be able capture.

Table 6.2 Optimized nonbonded FF parameters between COOH bead in CG PAA model and beads in CG solvents: W1 bead of the 1-site water model and AM bead in the CG DMF model.

bead pairs	**ε (kcal/mol)**	**σ (Å)[a]**
COOH-W1	1.35	3.853
COOH-AM	1.35	3.9133

a: obtained by Lorentz rule.[25]

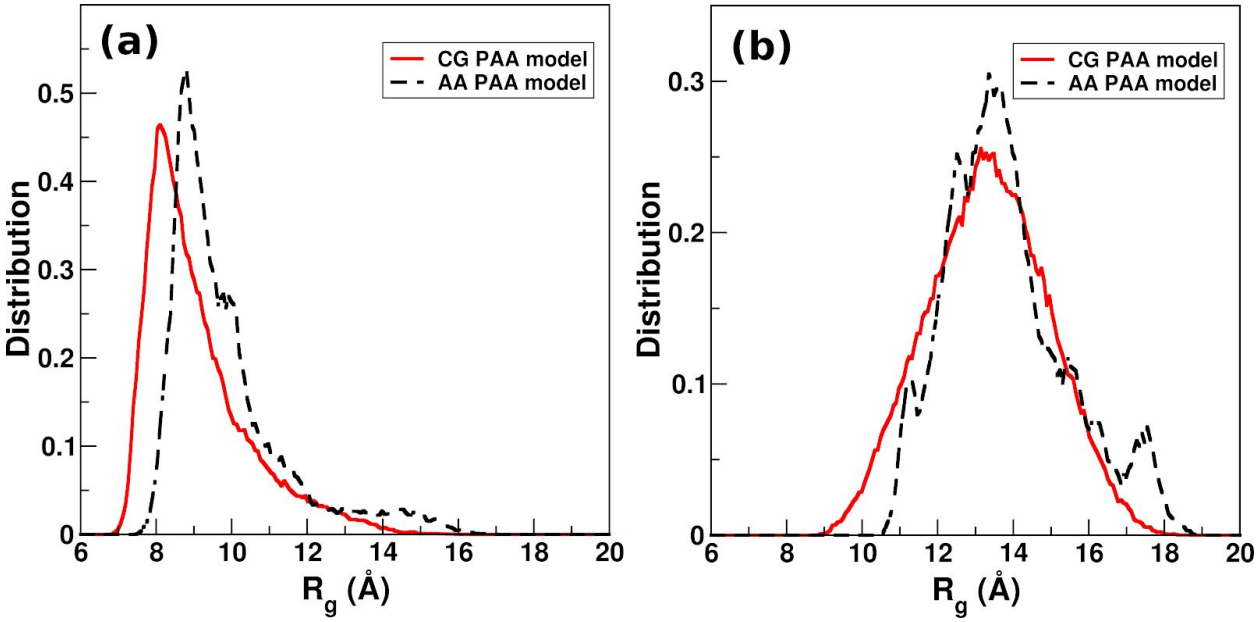

Figure 6.7 Distributions of R_g values of a single CG and all-atom PAA 30-mer chain in **(a)** pure water, and **(b)** pure DMF.

The new values of FF parameters generated to define the interactions between the CG PAA and solvents were further validated by calculating the Gibbs hydration and solvation free energies of a CG propionic acid model at 300 K. The Gibbs hydration free energy refers to propionic acid in water, and Gibbs solvation free energy refers to propionic acid in DMF. In **Table 6.3**, it can be seen that the Gibbs hydration free energy of the CG propionic acid was -8.4 kcal/mol, which is more negative than the reported experimental value of -6.5 kcal/mol.[28] The Gibbs solvation free energy of the CG propionic acid model was -9.4 kcal/mol, 2 kcal/mol smaller than that of the all-atom propionic acid model, -7.4 kcal/mol shown in **Figure C7 (b)**. The Gibbs hydration and solvation free energy profiles of the CG propionic acid models are also shown in **Figure C7** of Appendix C. Both the Gibbs hydration and solvation free energies of propionic acid are overestimated by the CG FF parameters which were developed to reproduce the R_g distributions of the all-atom PAA model. This indicates that the FF parameters of polymer and solvent are limited in terms of representing the interactions between monomer and solvent. In other words, in addition to the nonbonded interactions between polymer and solvent, the bonded interactions (e.g. bonds, angles, dihedrals etc.) between monomers also play an important role in determining their configuration in solvents.[29]

Table 6.3. The Gibbs hydration and solvation free energies of the CG propionic acid model.

propionic acid	Gibbs hydration free energy (kcal/mol)	Gibbs solvation free energy (kcal/mol)
CG model	-8.4±0.5	-9.4±0.7
experiment/all-atom model	-6.5[a]	-7.4±0.4[b]

a: experimental data is from ref [28]

b: the result of the all-atom model is shown in **Figure C7 (b)** of Appendix C.

Conformation of CG PAA in Binary Solvents

The R_g distributions and average R_g values of a PAA 30-mer chain in binary mixtures of 1-site water and CG DMF model are shown in **Figure 6.8 (a)** and **(b)**, respectively. In **Figure 6.8 (a)**, the R_g distribution shifts to the right as the mass fraction of DMF increases from 0 to 100 wt %. This indicates that the polymer chain gradually undergoes a globule-to-coil transition with an increase in DMF mass fractions (wt %) in a binary mixture with water. It can be seen in **Figure 6.8 (b)** that the average R_g values of CG and all-atom PAA chains were similar at simulated DMF mass fractions (errors in the range of ~3 % to ~17 %). This further suggests that the CG model can capture the conformation of the PAA chain in binary mixtures of DMF and water with different compositions.

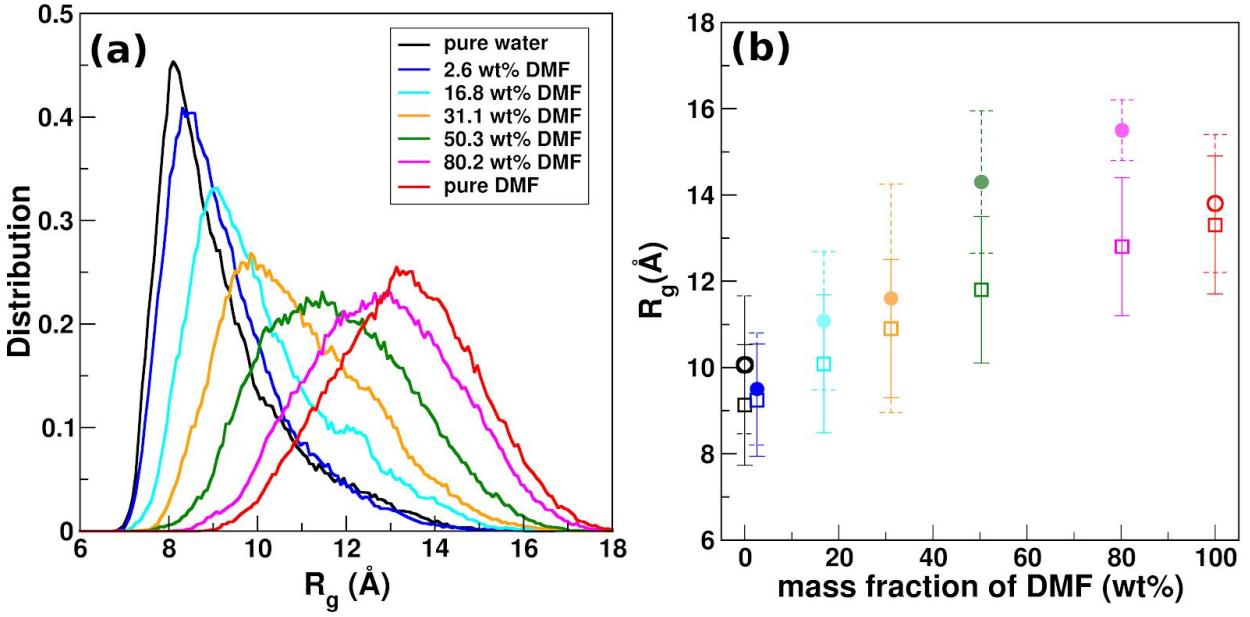

Figure 6.8 (a) The R_g distributions of a single CG PAA 30-mer chain, and **(b)** the average R_g values of a single all-atom and CG PAA 30-mer chain in binary solvents with different mass fractions of DMF. Circle and square symbols in **(b)** represent R_g values of all-atom and CG PAA models, respectively. Dashed and solid vertical lines in **(b)** show the standard deviation of R_g values of all-atom and CG PAA models, respectively.

Structure of Solvent at the Polymer-Solvent Interface

One of the key tests that can further demonstrate the accuracy of models developed in the present study is its ability to capture the structure of solvent near the polymer chains. Here, to investigate the origin of different conformations of a single PAA chain observed in the pure solvents as well as in their binary mixtures, we studied the structure of solvent at the polymer-solvent interface. Specifically, the RDFs between polymer and solvent beads were calculated and reported in **Figures 6.9** and **6.10**. As can be seen in the RDFs between the AM beads in DMF and the COOH beads in PAA in **Figure 6.9 (a)**, with an increase in the mass fraction of DMF, the intensities of the first (~4.3 Å) and second (~8.3 Å) peaks decrease. This suggests that ordering of DMF molecules becomes less prominent with increase in its mass fraction. Similar trend could also be observed in the RDF between nitrogen (N) atoms in the all-atom DMF molecules and O1 atoms in the all-atom PAA model (see **Figure 6.9 (b)**). The presence of two peaks in the RDF of the AM-COOH bead pair suggests two solvation shells of DMF molecules near the polymer chain.

In **Figure 6.10**, a sharp first peak at ~4.2 Å in the RDF between COOH beads of PAA and W1 beads of 1-site water model was observed. The intensity of this peak changes slightly with increasing DMF mass fraction. Moreover, it can also be seen that the second peak in the RDF between the COOH bead of PAA and 1-site water model shifts from ~7.9 Å to ~8.4 Å when the DMF mass fraction is increased from 2.6 wt % to 80.2 wt %.

To further quantify the amount of DMF in local solvent, we calculated the coordination number of the AM beads in DMF and W1 beads near the COOH bead in PAA (see **Table 6.4**). Note, the local solvent was defined as the solvent within 6.05 Å (the position of the first minimum in the RDF) of the COOH beads. It can be seen that with an increase in DMF content from 2.6 wt % to 80.2 wt %, the coordination number of AM beads increases from ~0.4 to ~5.4, while that of the W1 beads decreases from 9.1 to 1.7. This indicated that the water molecules were shielded away from the PAA by DMF molecules. Interestingly, when the mass fraction of DMF in bulk (far away from the PAA chain) was 50.3 wt %, the coordination numbers of AM and W1 beads are both ~4. At this point, the PAA chain was found to be in a coil-like state (R_g = ~12 Å in **Figure 6.8 (b)**). Moreover, the local mass fractions of DMF near the PAA chain were estimated by using the coordination number of AM and W1 beads and are shown in the last column of **Table 6.4**. For all the binary mixture, with DMF wt % increasing from 2.6 wt % to 80 wt %, the local DMF contents were always higher than that in the bulk. This suggests DMF molecules accumulate around the PAA chain. Thus, the local solvent composition plays a significant role in determining its globule-like and coil-like state. In addition, the role of hydrophobic interaction was also studied by analyzing the RDF between the CGD2 beads in DMF, and C2M beads in the backbone of PAA in **Figure C10** of Appendix C. In **Table C8** of Appendix C, the coordination number of CGD2 beads around the C2M beads increased with increasing the DMF mass fraction, which is consistent with the trend in the coordination number of the AM beads around the COOH beads.

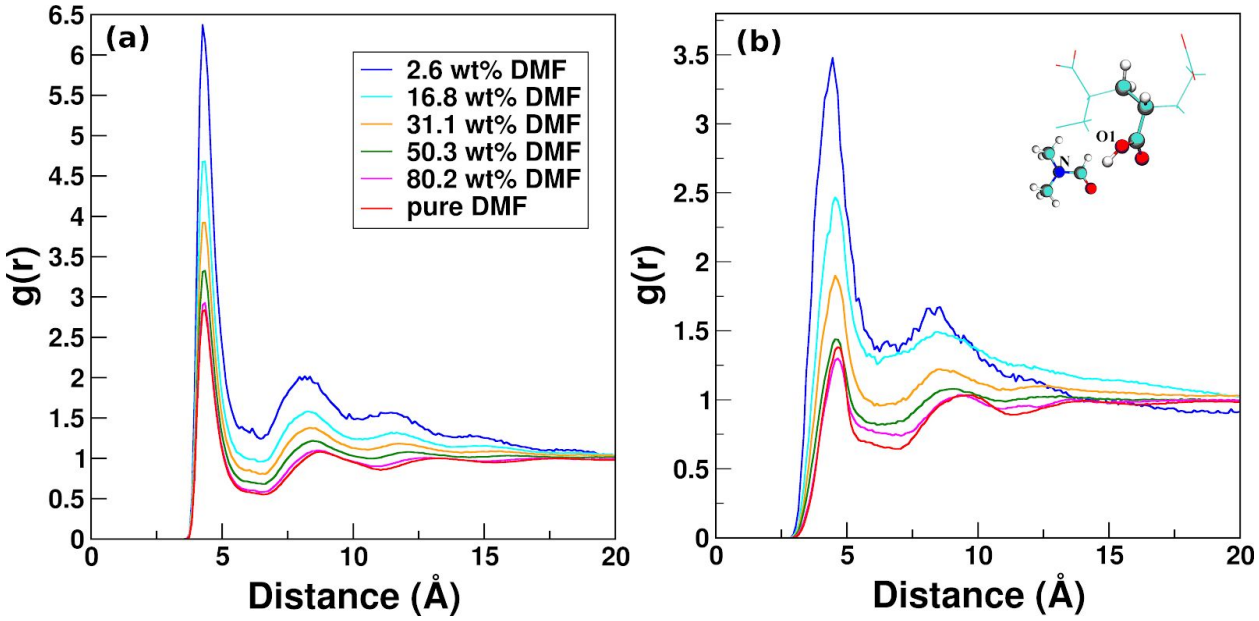

Figure 6.9 The RDFs between **(a)** AM beads in DMF and the COOH beads in the CG PAA model, and **(b)** N atoms in all-atom DMF molecules and the O1 atoms in the all-atom PAA model.

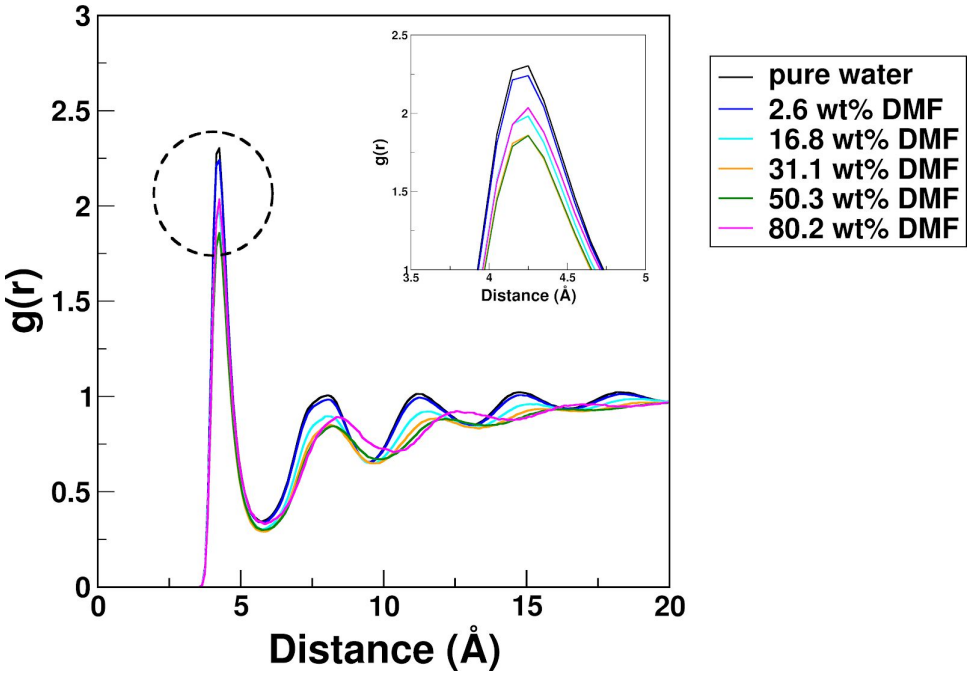

Figure 6.10 The RDF between W1 beads and the COOH beads in PAA.

Table 6.4 The coordination number of the AM and W1 beads around the COOH bead in PAA and the local solvent compositions at a cutoff of 6.05 Å.

solvent bulk	number of AM beads	number of W1 beads	local solvent
pure water	0	9.6	pure water
2.6 wt% DMF	0.4	9.1	8.2 wt% DMF
16.8 wt% DMF	1.9	6.9	35.8 wt% DMF
31.0 wt% DMF	2.9	5.4	52.3 wt% DMF
50.3 wt% DMF	4.0	3.9	67.5 wt% DMF
80.2 wt% DMF	5.4	1.7	86.6 wt% DMF
pure DMF	6.4	0	pure DMF

6.5 Conclusion

The coarse-grained (CG) model of poly(acrylic acid) (PAA) in the presence of explicit solvent was developed, which can reproduce its conformation in different solvents. The structure of solvent around the CG polymer model was captured accurately compared with that around the all-atom model. We initiated the model development by developing the CG model of propionic acid, which is an analogue of PAA's monomer. The particle swarm optimization (PSO) method was employed to optimize the force-field (FF) parameters to reproduce its experimental properties including density, self-diffusion coefficient, and enthalpy of vaporization at 300 K. This CG model of propionic acid was used to build the CG PAA model and perform simulations of 50 PAA 30-monomers (30-mer). The structure of CG PAA showed reasonable agreement with that obtained from all-atom mapped trajectory. The glass transition temperature (T_g) values of the CG PAA systems with 50 and 500 30-mer chains were predicted to be 386 and 383 K, comparable with the experimental value of 378.15 K. The nonbonded interactions between the CG PAA model and 1-site water model and between the CG PAA model and DMF model were optimized to reproduce the radius of gyration (R_g) distribution of an all-atom PAA chain in pure solvents. Furthermore, the behaviour of the CG PAA chain in binary solvent mixtures with different mass fractions of DMF was explored. It was found that the average values of R_g of the CG PAA chain gradually increased with an increase in the DMF mass fraction, indicating that

the PAA chain experienced a globule-to-coil transition. The analysis of RDFs between solvent and PAA showed that the ordering of DMF molecules in the first hydration shell became less prominent, while the coordination number of DMF molecules increased with an increase in DMF mass fraction. Similar behavior was observed for RDFs obtained from all-atom MD simulations of PAA in binary mixture of DMF and water. In the follow-up work, we will be utilizing this new approach to develop CG models of other polymers in solvent mixture. These CG models can further be utilized to study complex architectures e.g. brush structures and hydrogels.

References

(1) Adamczyk, Z.; Bratek, A.; Jachimska, B.; Jasiński, T.; Warszyński, P. Structure of Poly(acrylic Acid) in Electrolyte Solutions Determined from Simulations and Viscosity Measurements. *J. Phys. Chem. B* **2006**, *110* (45), 22426–22435.

(2) Sulatha, M. S.; Natarajan, U. Origin of the Difference in Structural Behavior of Poly(acrylic Acid) and Poly(methacrylic Acid) in Aqueous Solution Discerned by Explicit-Solvent Explicit-Ion MD Simulations. *Ind. Eng. Chem. Res.* **2011**, *50* (21), 11785–11796.

(3) Reith, D.; Meyer, H.; Müller-Plathe, F. Mapping Atomistic to Coarse-Grained Polymer Models Using Automatic Simplex Optimization To Fit Structural Properties. *Macromolecules* **2001**, *34* (7), 2335–2345.

(4) Reith, D.; Müller, B.; Müller-Plathe, F.; Wiegand, S. How Does the Chain Extension of Poly (acrylic Acid) Scale in Aqueous Solution? A Combined Study with Light Scattering and Computer Simulation. *J. Chem. Phys.* **2002**, *116* (20), 9100–9106.

(5) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153.

(6) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.

(7) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, 4667–4672.

(8) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Nonbonded Interactions between Coarse-Grained Hydrocarbon and Water Models. *J. Phys. Chem. B* **2019**, *123* (4), 909–921.

(9) Bejagam, K. K.; An, Y.; Singh, S.; Deshmukh, S. A. Machine-Learning Enabled New Insights into the Coil-to-Globule Transition of Thermosensitive Polymers Using a Coarse-Grained Model. *J. Phys. Chem. Lett.* **2018**, 6480–6488.

(10) Bejagam, K. K.; Singh, S.; Deshmukh, S. A. Development of Non-Bonded Interaction Parameters between Graphene and Water Using Particle Swarm Optimization. *J. Comput. Chem.* **2017**, *39* , 721–734.

(11) Bender, H. J.; Hertz, H. G. Model Orientation Dependent Pair Distribution Functions in Three Dimensions for the Liquid Mixtures of Propionic Acid and of Ethanol with Carbon Tetrachloride. *Berichte der Bunsengesellschaft für physikalische Chemie* **1977**, *81* (5), 468–478.

(12) Subha, M. C. S.; Rao, S. B. Densities and Viscosities of Propionic Acid in Benzene, Methylbenzene, Ethylbenzene, and Propylbenzene. *J. Chem. Eng. Data* **1988**, *33* (4), 404–406.
(13) Ahluwalia, R.; Gupta, R.; Vashisht, J. L.; Wanchoo, R. K. Physical Properties of Binary Liquid Systems: Ethanoic Acid/Propanoic Acid/Butanoic Acid with Cresols. *J. Solution Chem.* **2013**, *42* (5), 945–966.
(14) Allison, J. R.; Riniker, S.; van Gunsteren, W. F. Coarse-Grained Models for the Solvents Dimethyl Sulfoxide, Chloroform, and Methanol. *J. Chem. Phys.* **2012**, *136* (5), 054505.
(15) Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Comput. Biol.* **2010**, *6* (6), e1000810.
(16) Hancock, B. C.; Carlson, G. T.; Ladipo, D. D.; Langdon, B. A.; Mullarney, M. P. The Powder Flow and Compact Mechanical Properties of Two Recently Developed Matrix-Forming Polymers. *J. Pharm. Pharmacol.* **2001**, *53* (9), 1193–1199.
(17) Wu, T.; Gong, P.; Szleifer, I.; Vlček, P.; Šubr, V.; Genzer, J. Behavior of Surface-Anchored Poly(acrylic Acid) Brushes with Grafting Density Gradients on Solid Substrates: 1. Experiment. *Macromolecules* **2007**, *40* (24), 8756–8764.
(18) Maurer, J. J.; Eustace, D. J.; Ratcliffe, C. T. Thermal Characterization of Poly(acrylic Acid). *Macromolecules* **1987**, *20* (1), 196–202.
(19) Tong-Chun Bai, Jia Yao, and Shi-Jun Han. Excess Molar Volumes for Binary and Ternary Mixtures of (N,N-Dimethylformamide + Ethanol + Water) at the Temperature 298.15 K. *J. Chem. Thermodynamics* **1998**, No. 30, 1347–1361.
(20) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128* (144120), 1–13.
(21) Matos, G. D. R.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62* (5), 1559–1569.
(22) Yadav, V.; Harkin, A. V.; Robertson, M. L.; Conrad, J. C. Hysteretic Memory in pH-Response of Water Contact Angle on Poly(acrylic Acid) Brushes. *Soft Matter* **2016**, *12* (15), 3589–3599.
(23) Lützenkirchen, J.; van Male, J.; Leermakers, F.; Sjöberg, S. Comparison of Various Models to Describe the Charge−pH Dependence of Poly(acrylic Acid). *J. Chem. Eng. Data* **2011**, *56* (4), 1602–1612.
(24) Lin, H.; Xiao, W.; Qin, S.-Y.; Cheng, S.-X.; Zhang, X.-Z. Switch On/off Microcapsules for Controllable Photosensitive Drug Release in a “release-Cease-Recommence” Mode. *Polym. Chem.* **2014**, *5* (15), 4437–4440.
(25) Kong, C. L. Combining Rules for Intermolecular Potential Parameters. II. Rules for the Lennard‐Jones (12–6) Potential and the Morse Potential. *J. Chem. Phys.* **1973**, *59* (5), 2464–2467.
(26) Thorpe, I. F.; Zhou, J.; Voth, G. A. Peptide Folding Using Multiscale Coarse-Grained Models. *J. Phys. Chem. B* **2008**, *112* (41), 13079–13090.
(27) Fritz, D.; Koschke, K.; Harmandaris, V. A.; van der Vegt, N. F. A.; Kremer, K. Multiscale Modeling of Soft Matter: Scaling of Dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13* (22), 10412–10420.
(28) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. *J. Chem. Theory Comput.* **2006**, *2* (1), 128–139.

(29) Pezeshkian, W.; Khandelia, H.; Marsh, D. Lipid Configurations from Molecular Dynamics Simulations. *Biophys. J.* **2018**, *114* (8), 1895–1907.

CHAPTER 7
DEVELOPMENT OF CG POLYSTYRENE MODEL IN EXPLICIT SOLVENTS

7.1 Introduction

Polystyrene (PS) has exceptional strength and compatibility with blood, and is one of the most widely used commercial polymers in the fields ranging from packaging materials to medical devices.[1–4] To obtain these products/devices, PS is usually treated with different solvents.[5–7] For example, to improve the cell adhesion, the surfaces of PS are usually modified with acids such as sulfuric acid, which essentially modifies the chemical compositions of the surface.[8] Moreover, these PS devices can be recycled by extruding with different solvents.[9] Typically, the choice of a solvent determines the conformation of the polymer chains, which might also affect the quality of the product.[9] In general, in a good solvent, the polymer chains tend to exhibit a coil-like structure.[5] On the other hand, they are in a globule-like state in a poor solvent. However, our molecular-level understanding of conformations of these polymer chains as well as structure of solvent remains underdeveloped. This can be attributed to the lack of experimental characterization methods that can probe this structure at the microscopic level.

In recent years, both all-atom and coarse-grained (CG) molecular dynamics (MD) simulations have been frequently used to study the conformations of a single chain of macromolecules in the presence of explicit solvent models.[5,10,11] An obvious advantage of developing and using a CG model is that it can be used to study a single chain as well as their architectures (*e.g.* bottlebrush polymers, polymer grafted inorganic nanoparticles, etc.) in the presence of solvents or at liquid-liquid interfaces.[12–14] Indeed, a number of CG models of PS have been developed to study its structure in the bulk or solvents.[15–19] Rossi *et. al.* built a CG PS chain model by using the Martini force field (FF).[15] They refined the Matini FF by reproducing the density and radius of gyration (R_g) of PS in melt. This CG PS model was further validated by comparing its structure and dynamic properties with that of an AA model. Xiao and Guo studied the effects of nonbonded force field formats on the T_g of CG PS models.[17] They selected two LJ potential formats, the 12-6 and 9-6, and found that the T_g is higher for the LJ potential represented by 12-6 format. They further modified the 12-6 LJ potential by tuning the epsilon values, as well as the bonded parameters to obtain the T_g value of 382 K, close to the reported value of 360 K from united-atom simulations.[20] All the aforementioned CG PS models were

solely developed by reproducing their bulk structure or physical properties like density and RDFs. Rossi *et. al.* studied the structure of CG PS chains in benzene as a good solvent when they were grafted on substrates.[16] They found that the transition of PS chains in benzene from collapsed to stretched configurations became smoother as the grafting density increased. However, most of these CG PS models had one or more of the following limitations: (*i*) Some of these models were unable to predict its glass transition temperature (T_g) accurately, (*ii*) Very little is known about the properties and conformations of these CG PS models in different solvents, and (*iii*) The precise effects of change in the bonded and/or nonbonded FF parameters on the properties of CG PS models are unknown.

The lack of accurate solvent-transferable CG PS models further limits our understanding of the effect of solvents on atomistic or molecular-level conformations of PS, which might be helpful in enhancing processability of PS under different conditions. Here, we develop a temperature and solvent transferable CG PS model, which can be used to study its bulk properties at different temperatures, and conformations of PS chains in different solvents. We begin the CG model development with PS's monomer analogue, namely ethylbenzene by reproducing its experimental properties reported in the literature where possible. The CG PS model was constructed by connecting CG ethylbenzene molecules to generate polymer chains with 30 monomers (30-mers). The bulk properties of PS were calculated and compared with available experimental data at different temperatures. Finally, the chain conformation of a single PS chain in pure water, pure DMF and water/DMF binary solvents were investigated, and validated by comparing with all-atom MD simulation results.

7.2 Model Development

To build the CG model of PS, we adopted the CG hydrocarbon beads as the backbone and benzene beads as the side chain, both of which use 2:1 mapping scheme and have been developed in our previous work.[21,22] Representative examples of the CG hydrocarbon model and the CG decane model are shown in **Figure 7.1**. In the hydrocarbon model, the end bead C2E represents two carbon atoms along with the attached hydrogen atoms at the two ends, and the middle bead C2M represents two carbon atoms along with their attached hydrogen atoms in the middle of the molecule. The CG benzene model consists of three BZ beads, which are connected such that they form a three-membered ring structure similar to the all-atom benzene model. The

properties of CG hydrocarbon and benzene models are in good agreement with their experimental values.[21,22] With these CG hydrocarbon and benzene beads, the CG ethylbenzene and PS model were constructed and shown in **Figure 7.1**. The two ends of the PS chain are represented with two C2E beads, which are not shown. **Figure 7.1** also shows the reported CG models of water and DMF, which were developed in our previous study.[21,23] The chargeless W1 bead, i.e. 1-site water model, represents two all-atom water molecules and could reproduce the experimental properties, for example, density and surface tension with good accuracy.[21] The DMF molecule is mapped to two beads, namely, AM and CGD2 that encompass the amide and two methyl groups, respectively. Both the AM and CGD2 are charge neutral beads. These models were used to study the conformations of PS in pure solvents and their mixtures.

The FF format for both ethylbenzene and PS is CHARMM format, as shown in **Equation 7.1**. The corresponding FF parameters are discussed in **Section 7.3** of the manuscript. Beads separated by two bonds interact with each other through LJ 12-6 potential.

$$E_{pot} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angels} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi(1 + cos(n\varphi - \varphi_0))^2 + \sum_{impropers} K_\psi(\psi - \psi_0)^2 + \sum_i \sum_j 4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}]$$

…….Equation 7.1

Where, K_b, K_θ, K_φ and K_Ψ are the bond, angle, dihedral and improper angle force constants, respectively. b_0, θ_0, φ_0 and Ψ_0 are the equilibrium bond length, angle, dihedral angle and improper angle values, respectively. n is the multiplicity (1 or 2 in this study), ε_{ij} is the depth of the potential well and represents the strength of interactions between two beads i and j, σ_{ij} is the finite distance at which the inter-particle potential is zero, and r_{ij} is the distance between two beads.

Before studying the properties of the CG PS model, we construct the CG model of ethylbenzene (see **Figure 7.1 - (e)**), the monomer's analogue of PS, in order to validate the cross interactions between BZ and C2E/C2M beads by using Lorentz-Berthelot (LJ) combining rules.[24] The force field parameters for the bond C2E-BZ and the angle C2E-BZ-BZ were obtained from the mapped trajectory of AA simulations of ethylbenzene at 300 K. Nonbonded parameters of C2E were from the CG hydrocarbon models in reference [22]. The CG benzene model in reference[21] provided the bonded and nonbonded parameters between BZ beads, which are shown in **Table D1** of Appendix D. Methods of calculating the density, enthalpy of vaporization, and

surface tension of the CG ethylbenzene model are adopted from our previous work.[21,22,25] For the CG PS model, the bonded FF parameters in the backbone including C2M-C2M, C2E-C2M, C2M-C2M-C2M, C2E-C2M-C2M were adopted from the CG hydrocarbon models in reference [22]. The bonded parameters for C2E-BZ were taken directly from the ethylbenzene model. The nonbonded parameters of BZ, C2E, C2M beads were the same as those used in the CG hydrocarbon and benzene models.

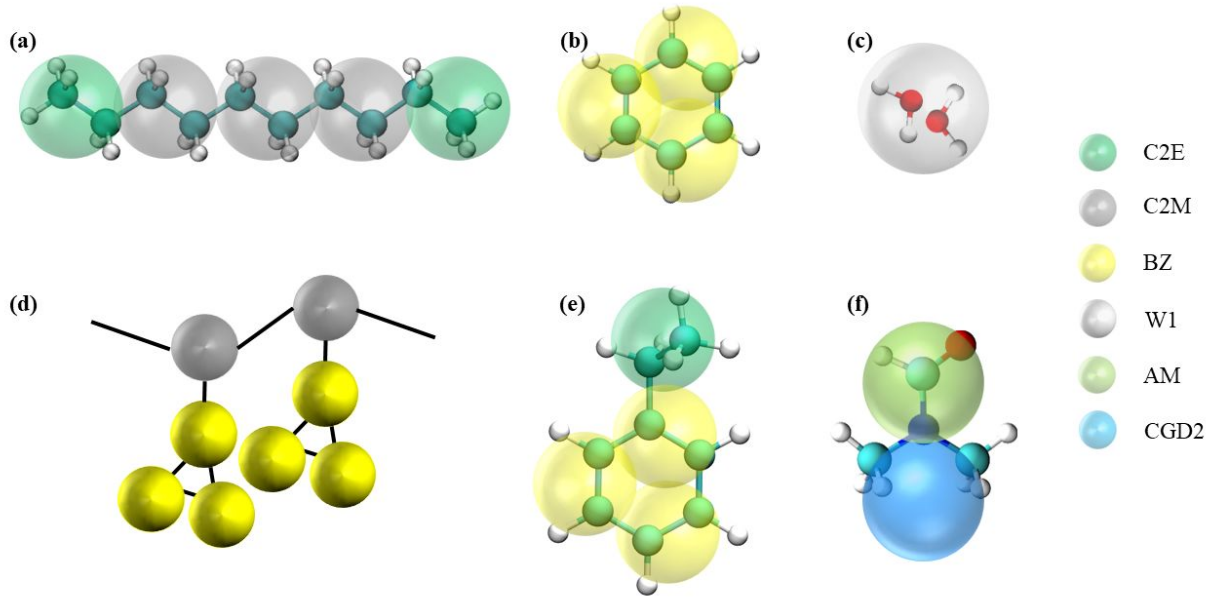

Figure 7.1 CG models of **(a)** hydrocarbon (decane as an example), **(b)** benzene, **(c)** water, **(d)** PS, **(e)** ethylbenzene and **(f)** DMF. The CG models of hydrocarbons, benzene, and water are adopted from references [21–23].

7.3. Results and Discussion

7.3.1 CG Ethylbenzene Model

To validate the FF parameters between CG hydrocarbon beads and benzene beads, the CG ethylbenzene model was constructed with one C2E bead and three BZ beads. Properties of the CG ethylbenzene model and their corresponding experimental properties are listed in **Table 7.1**. The density, surface tension and enthalpy of vaporization are in good agreement with experimental values (errors < 3 %). This suggests that bonded and nonbonded FF parameters

between the C2E and BZ beads are accurate in predicting the experimental properties of ethylbenzene.

Table 7.1 Properties of CG ethylbenzene and propionic acid models at 300 K. Experimental data is from references [26]. Density, ρ - g/cm^3, self-diffusion coefficient, D - $\times 10^{-9}$ m^2/s, surface tension, γ - mN/m, enthalpy of vaporization, H_V - kcal/mol.

	ρ	***D***	***γ***	H_V
CG Ethylbenzene Model	0.866	1.4	29.10	10.0
Experiment	0.862	-	28.43	10.3
Error (%)	0.5	-	2.4	2.9

7.3.2 Structure of the CG PS model

Figure 7. 2 - (a) shows the bond length distribution of C2M-BZ from the CG PS model, which varies from 2.1 to 2.75 Å with a single peak at 2.4 Å. This distribution has a large overlap with that obtained from AA mapped trajectory which exhibits a peak at 2.3 Å with a shoulder at 2.4 Å. For the angle distribution of BZ-C2M-C2M from the CG simulations, it shows a unimodal distribution with the peak at 115^0. However, a bimodal distribution was observed for the angle BZ-C2M-C2M from AA simulations, which is due to the trans-gauche transition of the backbone of PS. The intensity of the peak in the angle distribution of BZ-C2M-C2M from the CG model is much higher than that from the AA mapped model. This is because of the LJ potential applied on 1-3 beads (beads separated by two bonds), which makes the angle BZ-C2M-C2M in the CG PS model more rigid.

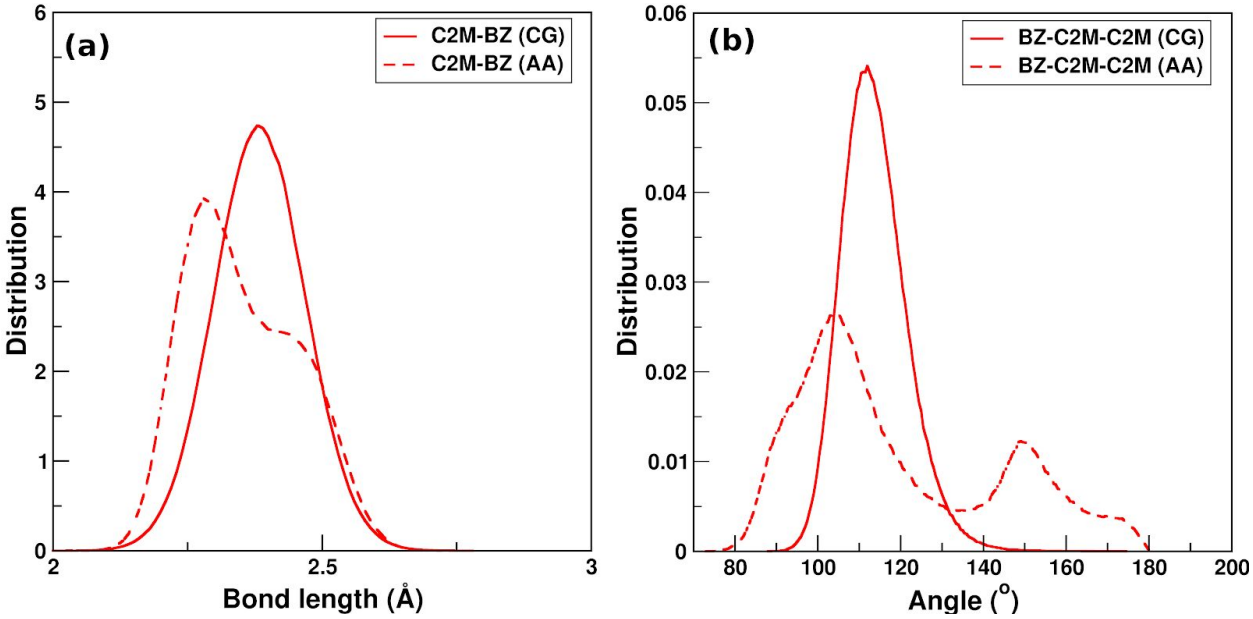

Figure 7.2 The distributions of **(a)** bond (CM-BZ) and **(b)** angle (BZ-C2M-C2M) from CG and AA MD simulations of 50 PS 30-mer chains at 300 K and 1 bar.

Density and Glass transition temperature of PS

To investigate the bulk properties of the CG PS model, simulations of 50 CG PS 30-mer chains were performed at temperatures ranging from 280 K to 480 K. Note that previous computational studies have shown that the T_g of the CG polyacrylic acid (PAA) systems only changes by 3 K with the number of chains increasing from 50 to 500.[12] The densities of PS were calculated and are shown in **Figure 7.3**. The density of the CG PS model at 300 K is 1.05 g/cm^3, in excellent agreement with the experimental value of 1.06 g/cm^3.[27] An obvious slope change could be found at 379.5 K, which corresponds to the T_g. This is in the range of the experimentally measured values from 370 to 380 K.[28] Note, we have systematically tuned the dihedral angle force constant, *i.e.* K_φ in **Equation 7.1**, from 0, 0.5 to 1.0 Kcal/mol. When the K_φ is 0, the slope change was observed at 297 K (see **Figure D1** of Appendix D). As the K_φ was increased to 0.5 Kcal/mol, the T_g value was at ~333 K (see **Figure D1** of Appendix D), lower than the experimental value. When the K_φ was increased to 1.0 Kcal/mol, the T_g value increased to 379.5 K. We attribute the increase in the T_g values to the increased rigidity of the PS backbone due to higher values of the K_φ. The comparison between the dihedral angle distributions from CG simulations and AA mapped simulation trajectories is shown in **Figure D2** of Appendix D. Because the FF parameters of the dihedral angles is tuned to reproduce the experimental T_g

value, it has limitations in reproducing the distributions of dihedral angles from the AA mapped trajectories.

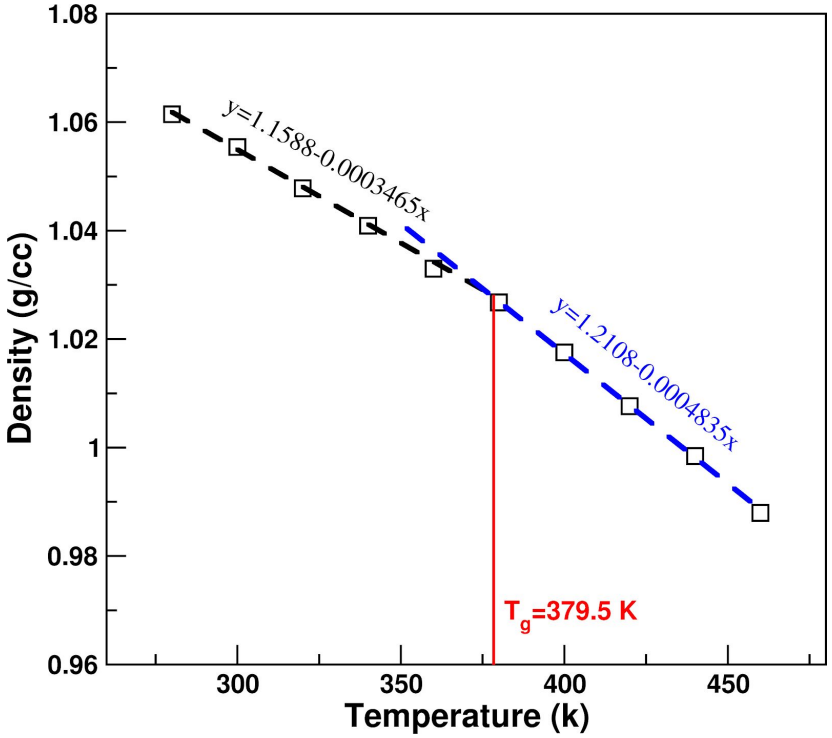

Figure 7.3 The density of CG PS models as a function of temperature from 280 to 480 K. The system consisted of 50 polymer chains with 30 monomers in each chain.

7.3.3 Conformation of PS in solvents

PS in pure water and pure DMF: Developing an accurate solvent transferable CG PS model could help provide insight into the conformation change of PS at molecular level. As far as we know, this is the first attempt to build a CG PS model that is transferable with respect to different solvents. Initially, CG and AA simulations of a single PS 30-mer chain in pure water and pure DMF were performed. The interactions between C2E/C2M-W1 and between BZ-W1 have been optimized in references.[21,29] These parameters are used without any modification in this study and listed in **Table D1** of Appendix D.

The R_g values of a single CG and AA PS 30-mer chain were compared in **Figure 7.4**. It can be found that the R_g distribution of the CG PS 30-mer model in pure water (**Figure 7.4 - (a)**) shows a sharp peak at 8.3 Å, while that of the AA PS 30-mer in water exhibits a peak at 9.3 Å. Although the peak positions are different by 1.0 Å, their intensities are similar and they both

indicate a collapsed/globule-like state of PS in pure water. **Figure 7.4 - (b)** shows the R_g distributions of a CG and AA PS 30-mer model in pure DMF. The interactions between PS and DMF in both AA and CG models were obtained by the LB combining rules (see **Table D1** of Appendix D). The R_g distribution of the CG PS model shows a peak at around 13.5 Å, left to that of the R_g distribution of the AA PS model at around 14.1 - 15.6 Å. This intensity of the peak in the R_g distribution of the CG PS model is smaller than that of the AA PS model. This means the AA PS model tends to be more stretched out in DMF than the CG PS model. However, there is a large overlap between these two distributions. Overall, the CG PS model shows a globule-like state in pure water and a coil-like state in pure DMF, which reproduces the conformations of an AA PS model.

In **Section 7.3.2.1**, we have discussed the effects of dihedral angles on the T_g of the CG PS model. We also investigated the effects of dihedral angles on the structures of the CG PS model in solvents by calculating the R_g distributions of the CG PS chain with dihedral angle force constant K_φ = 0 and 1.0 Kcal/mol. These results are shown in **Figure D3** of Appendix D. As the dihedral force constant was increased from 0 to 1 Kcal/mol, the R_g distribution of the CG PS chain shifted slightly to the right due to the increased rigidity of the polymer chain. However, no obvious difference was observed in the R_g distributions of the CG PS chain in pure DMF. The solvent structures around the side chain of PS were also analyzed. Interestingly, it was found that the RDF between W1-BZ and between CGD1-BZ didn't change significantly as the K_φ changed from 0 to 1 Kcal/mol.

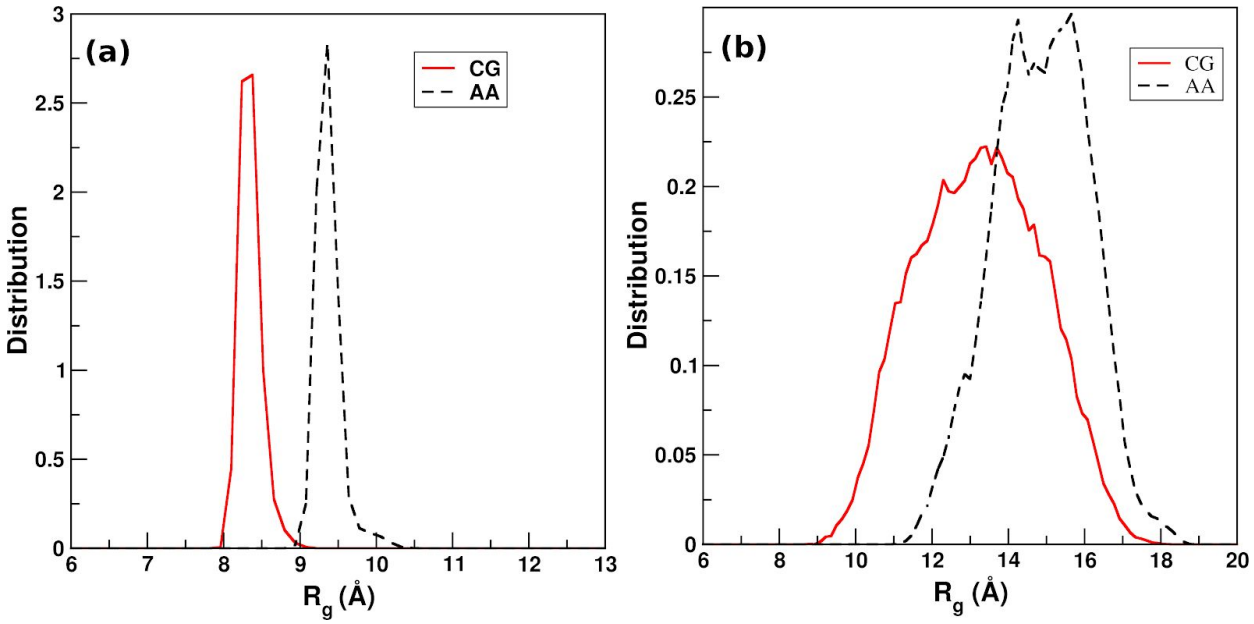

Figure 7.4 The R_g distributions of a CG PS chain in **(a)** pure water and **(b)** pure DMF as compared with those of an AA PS chain.

PS in solvent mixtures: The conformation of a single PS 30-mer chain in solvent mixtures of water and DMF was further investigated. As shown in **Figure 7.5**, the R_g distribution was shifted to the right as the mass concentration of DMF was increased from 2.6 wt% to 80.2 wt%. This suggests that the PS chain gradually undergoes a globule-to-coil conformational transition as the DMF concentration increases in the solvent mixture.

To understand the local structure of the DMF and water molecules around the polymer chain the RDFs were plotted between W1 beads and BZ beads, and between the CGD2 beads and BZ beads. It can be seen in **Figure 7.6 - (a)** that the first peak is located at ~4.1 Å, which is an indication of the first hydration shell around the BZ beads at this distance. The second hydration shell is observed at ~8.1 Å for pure water. However, it shifted slightly to the right as the DMF concentration increased to 80.2 wt%. In the RDF between CGD2 and beads in **Figure 7.6 - (b)**, a sharp peak could be observed at ~4.5 Å with a small shoulder at ~5.9 Å, which corresponds to the solvation shell formed by DMF around the BZ beads. As the DMF concentration increased, the coordination number of CGD2 and W1 beads in the first hydration and solvation shells were calculated as shown in **Table 7.2**. It can be noticed that the coordination number of CGD2 beads increased from 0 to 4.3, while that of the W1 beads decreased from 4.9 to 0 when the DMF concentration increased from 0 to 100 %. This suggests

that the water molecules were repelled far away from the PS side chains, which leads to the stretching of the PS chains.

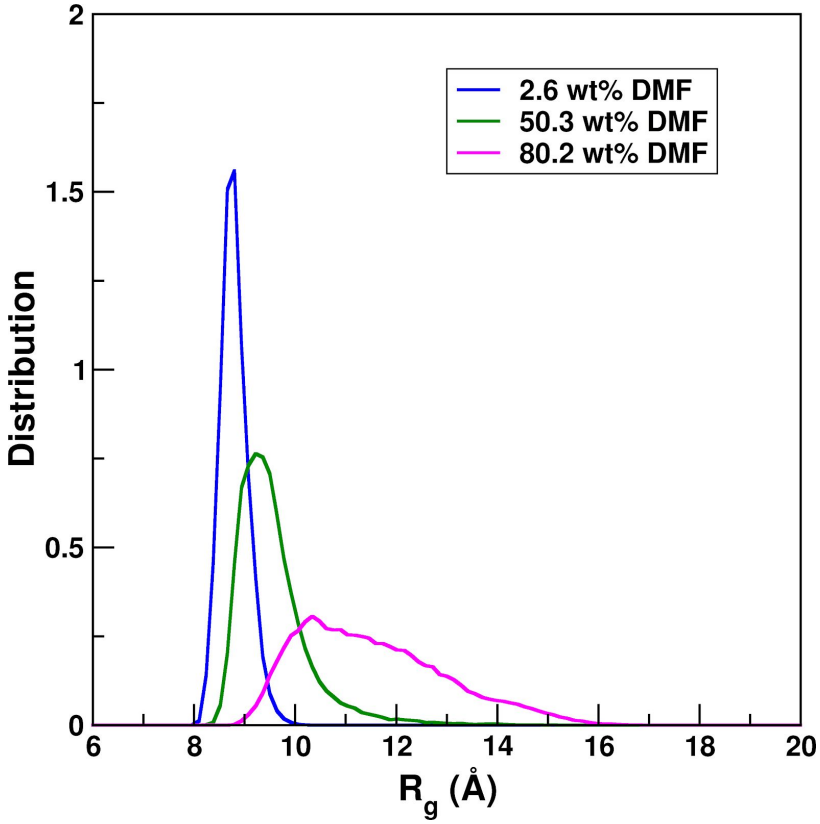

Figure 7.5 The R_g distributions of a single CG PS chain in solvent mixtures with different DMF mass concentrations.

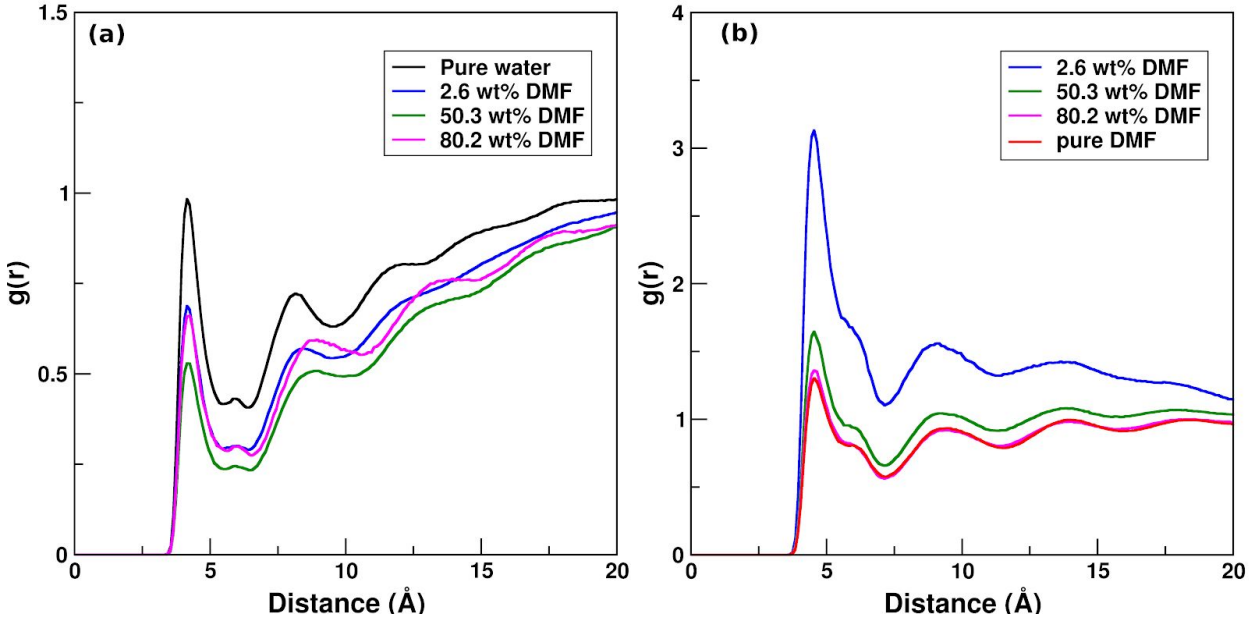

Figure 7.6 The RDFs of **(a)** BZ-W1 and **(b)** BZ-CGD2 bead pairs at different solvent mixtures.

Table 7.2 The coordination number of the CGD2 and W1 beads around the BZ bead in PS with a cutoff of 5.75 Å for CGD2 (the position of the first valley in **Figure 7.6 - (b)**) and 5.05 Å for W1 (the position of the first valley in **Figure 7.6 - (a)**).

solvent bulk	number of CGD2 beads	number of W1 beads
pure water	0	4.9
16.8 wt% DMF	1.6	2.2
50.3 wt% DMF	2.9	1.7
80.2 wt% DMF	3.8	0.5
pure DMF	4.3	0

7.4 Conclusions

The CG PS model was developed in this work, which was initiated by developing and validating a CG model of its monomer's analogue, ethylbenzene. The CG ethylbenzene model consisted of one CG hydrocarbon bead and three CG benzene beads, which could reproduce its experimental properties: density, enthalpy of vaporization and surface tension at 300 K with errors less than 3 %. The CG ethylbenzene models were used to build the PS 30-mer chains. Its density and T_g was calculated, which was comparable with their corresponding experimental values. Moreover, the

globlue-to-coil transition of the PS chain was captured by the CG PS model as induced by increasing the DMF concentrations in the solvent mixtures of water/DMF. The reason for the conformation transition was investigated by analyzing the local solvent structure around the polymer side chains. It was found that the number of CG water molecules in the first hydration shell decreased from 4.9 to 0, however, the number of CGD2 beads in DMF molecules in the first solvation shell increased from 0 to 4.3, as the concentration of DMF increased from 0 to 100 wt%. This newly developed CG PS model lay the basis for studying more complex structures such as copolymers and bottlebrush polymers with potential applications of surface coating and PS based devices for applications in tissue culturing.

7.5 Future work

The FF parameters for the CG PS model will be further modified to reproduce the structural properties from all-atom simulations, while maintaining the rigidity of polymer chains to predict the experimental T_g value accurately. Furthermore, the conformations of CG PS chain in solvent needs to be improved to replicate the R_g distributions of the all-atom model of PS. Furthermore, the newly developed CG PS model would also be employed to construct complex polymer structures.

References:

(1) Scalenghe, R. Resource or Waste? A Perspective of Plastics Degradation in Soil with a Focus on End-of-Life Options. *Heliyon* **2018**, *4* (12), e00941.

(2) Groh, K. J.; Backhaus, T.; Carney-Almroth, B.; Geueke, B.; Inostroza, P. A.; Lennquist, A.; Leslie, H. A.; Maffini, M.; Slunge, D.; Trasande, L.; Warhurst, A. M.; Muncke, J. Overview of Known Plastic Packaging-Associated Chemicals and Their Hazards. *Sci. Total Environ.* **2019**, *651* (Pt 2), 3253–3268.

(3) Klapperich, C. M. Microfluidic Diagnostics: Time for Industry Standards. *Expert Rev. Med. Devices* **2009**, *6* (3), 211–213.

(4) Bagheri-Khoulenjani, S.; Mirzadeh, H. Polystyrene Surface Modification Using Excimer Laser and Radio-Frequency Plasma: Blood Compatibility Evaluations. *Prog Biomater* **2012**, *1* (1), 4.

(5) Morozova, T. I.; Nikoubashman, A. Coil-Globule Collapse of Polystyrene Chains in Tetrahydrofuran-Water Mixtures. *J. Phys. Chem. B* **2018**, *122* (7), 2130–2137.

(6) Bolton, J.; Rzayev, J. Tandem RAFT-ATRP Synthesis of Polystyrene–Poly(Methyl Methacrylate) Bottlebrush Block Copolymers and Their Self-Assembly into Cylindrical Nanostructures. *ACS Macro Lett.* **2012**, *1* (1), 15–18.

(7) Liu, C.; Chen, G.; Sun, H.; Xu, J.; Feng, Y.; Zhang, Z.; Wu, T.; Chen, H. Toroidal Micelles of Polystyrene- Block -Poly(acrylic Acid). *Small* **2011**, *7* (19), 2721–2726.

(8) Lerman, M. J.; Lembong, J.; Muramoto, S.; Gillen, G.; Fisher, J. P. The Evolution of Polystyrene as a Cell Culture Material. *Tissue Eng. Part B Rev.* **2018**, *24* (5), 359–372.
(9) García, M. T.; Duque, G.; Gracia, I.; de Lucas, A.; Rodríguez, J. F. Recycling Extruded Polystyrene by Dissolution with Suitable Solvents. *J. Mater. Cycles Waste Manage.* **2009**, *11* (1), 2–5.
(10) Ortiz de Solorzano, I.; Bejagam, K. K.; An, Y.; Singh, S. K.; Deshmukh, S. A. Solvation Dynamics of N-Substituted Acrylamide Polymers and the Importance for Phase Transition Behavior. *Soft Matter* **2020**, *16* (6), 1582–1593.
(11) Bejagam, K. K.; An, Y.; Singh, S.; Deshmukh, S. A. Machine-Learning Enabled New Insights into the Coil-to-Globule Transition of Thermosensitive Polymers Using a Coarse-Grained Model. *J. Phys. Chem. Lett.* **2018**, 6480–6488.
(12) An, Y.; Singh, S.; Bejagam, K. K.; Deshmukh, S. A. Development of an Accurate Coarse-Grained Model of Poly(acrylic Acid) in Explicit Solvents. *Macromolecules* **2019**, *52* (13), 4875–4887.
(13) Srinivas, G.; Discher, D. E.; Klein, M. L. Self-Assembly and Properties of Diblock Copolymers by Coarse-Grain Molecular Dynamics. *Nat. Mater.* **2004**, *3* (9), 638–644.
(14) Lee, H.; de Vries, A. H.; Marrink, S.-J.; Pastor, R. W. A Coarse-Grained Model for Polyethylene Oxide and Polyethylene Glycol: Conformation and Hydrodynamics. *J. Phys. Chem. B* **2009**, *113* (40), 13186–13194.
(15) Rossi, G.; Monticelli, L.; Puisto, S. R.; Vattulainen, I.; Ala-Nissila, T. Coarse-Graining Polymers with the MARTINI Force-Field: Polystyrene as a Benchmark Case. *Soft Matter* **2011**, *7* (2), 698–708.
(16) Rossi, G.; Elliott, I. G.; Ala-Nissila, T.; Faller, R. Molecular Dynamics Study of a MARTINI Coarse-Grained Polystyrene Brush in Good Solvent: Structure and Dynamics. *Macromolecules* **2012**, *45* (1), 563–571.
(17) Xiao, Q.; Guo, H. Transferability of a Coarse-Grained Atactic Polystyrene Model: The Non-Bonded Potential Effect. *Phys. Chem. Chem. Phys.* **2016**, *18* (43), 29808–29824.
(18) Fritz, D.; Harmandaris, V. A.; Kremer, K.; van der Vegt, N. F. A. Coarse-Grained Polymer Melts Based on Isolated Atomistic Chains: Simulation of Polystyrene of Different Tacticities. *Macromolecules* **2009**, *42* (19), 7579–7588.
(19) Karimi-Varzaneh, H. A.; van der Vegt, N. F. A.; Müller-Plathe, F.; Carbone, P. How Good Are Coarse-Grained Polymer Models? A Comparison for Atactic Polystyrene. *Chemphyschem* **2012**, *13* (15), 3428–3439.
(20) Xia, J.; Xiao, Q.; Guo, H. Transferability of a Coarse-Grained Atactic Polystyrene Model: Thermodynamics and Structure. *Polymer* **2018**, *148*, 284–294.
(21) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**.
(22) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153.
(23) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, 4667–4672.
(24) Schoen, M.; Hoheisel, C. The Mutual Diffusion Coefficient D 12 in Liquid Model Mixtures A Molecular Dynamics Study Based on Lennard-Jones (12-6) Potentials: II. Lorentz-Berthelot Mixtures. *Mol. Phys.* **1984**, *52*, 1029–1042.
(25) Conway, O.; An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Coarse-Grained Models of Amino Acids. *Molecular Systems Design & Engineering* **2020**.

https://doi.org/10.1039/C9ME00173E.
(26) Subha, M. C. S.; Rao, S. B. Densities and Viscosities of Propionic Acid in Benzene, Methylbenzene, Ethylbenzene, and Propylbenzene. *J. Chem. Eng. Data* **1988**, *33* (4), 404–406.
(27) Rouabah, F.; Dadache, D.; Haddaoui, N. Thermophysical and Mechanical Properties of Polystyrene: Influence of Free Quenching. *ISRN Polymer Science* **2012**, *2012*. https://doi.org/10.5402/2012/161364.
(28) Rieger, J. The Glass Transition Temperature of Polystyrene. *J. Therm. Anal.* **1996**, *46* (3), 965–972.
(29) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Nonbonded Interactions between Coarse-Grained Hydrocarbon and Water Models. *J. Phys. Chem. B* **2019**. https://doi.org/10.1021/acs.jpcb.8b07990.

CHAPTER 8
CONFORMATION TRANSITION OF BOTTLEBRUSH COPOLYMERS PS-PAA

8.1 Introduction

Based on the obtained coarse-grained (CG) hydrocarbon, PAA, and PS models in previous chapters, we are able to construct polymers containing PAA and/or PS blocks with different architectures, such as linear polymers, star-like polymers, and bottlebrush polymers etc.[1,2] Bottlebrush polymers (BBPs) are a type of macromolecules with a linear backbone highly grafted by polymeric side chains. A special class of these materials is bottlebrush copolymers, which are consisted of side chains with more than one type of polymers.[3] They can self-assemble into a variety of nanostructures, which can be applied in drug delivery and medical imaging.[3] The overall conformations of BBPs and their self-assembled structures can be altered by changing the chemical environment such as solvents. Many experimental and computational studies suggest that in good solvents, the backbones of the BBPs are partially or fully extended due to the steric hindrance between side chains.[3] The side chains are also in the extended states at high grafting density. Moreover, by adjusting the solvent quality a variety of self-assembled nanostructures including spherical micelles, bilayers, and toroids of amphiphilic BBPs (e.g. poly(ethylene glycol)-poly(d,l-lactide)) could be obtained.[4] Bottlebrush copolymers such as polystyrene-*b*-poly(methacrylic acid) and polystyrene-*b*-poly(acrylic acid) have also been synthesized recently.[5,6] Despite of significant efforts in synthesizing different types of BBPs, molecular-level understanding of the configurations of individual polymer chains and solvent at the polymer-solvent interface is very limited. This can be attributed to the lack of experimental characterization methods.

Computer simulations of BBPs have been performed to provide a systematic and quantitative molecular-level understanding of the BBPs in melt and solution. Wessel *et al.* designed a series of amphiphilic bottlebrush copolymers with solvophilic and solvophobic blocks (A and B) and studied their self-assembled structures.[7] They performed CG molecular dynamics (MD) simulations and showed that the BBPs with AB, ABA, BAB architectures self-assembled into sphere, cylinder and bilayer nanostructures, respectively, when the ratio of A and B blocks is 25:75 and the solvophobicity is 0.65. Dutta *et al.* performed Brownian dynamics and Monte Carlo simulations on BBPs with poly(norbornene) (PNB) as backbone and poly(lactic acid)

(PLA) as side chains.[8] The intrinsic viscosity of these BBPs in dilute solution from simulation showed quantitative agreement with their experimental data. These reported models of BBPs are simple, but encompass few chemical details along with the implicit models of solvents. This limits our understanding of the interaction between BBPs and solvents which plays a significant role in determining the conformations of BBPs. To understand the role of solvents, the CG BBPs models compatible with explicit solvent models need to be developed.

Here, the BBPs with different CG PAA and PS blocks as side chains were constructed by using the models discussed in **Chapters 6** and **7** of this report. To understand the effect of solvents on the conformations of side chains and backbone of bottlebrush copolymers, we have performed CG MD simulations of bottlebrush copolymers in good and poor solvents, namely DMF and water, respectively. The CG models of bottlebrush copolymers were developed by varying the positions of PS and PAA side chains with 30 monomers (30-mer). Specifically, bottlebrush copolymers with three different types of grafting typers/positions were studied in pure DMF, pure water, and their binary mixtures. The structure of these bottlebrush copolymers were characterized by calculating the radius of gyration (R_g) and end-to-end distance. The structure of solvents around the side chains and backbones of the bottlebrush copolymers were explored by calculating the RDFs.

8.2 Model Development

The CG PS-PAA BBP molecules consisted of 12 CG PS and 12 CG PAA 30-mers as the side chains, and 72 hydrocarbon C2M beads as the backbone. CG PAA and PS models were adopted directly from **Chapters 6** and **7.** The positions of the PS and PAA side chains were varied along the backbone to form BBP molecules with different architectures as shown in **Figure 8.1**. Three representative PS-PAA BBP molecules: A1, A2, and A3 were constructed with a grafting density of 33.3% (24 side chains). This means the side chains were grafted every three beads in the backbone. For A1 in **Figure 8.1**, CG PAA chains were attached to one half of the backbone beads (beads 1 to 36), while the CG PS chains were attached to the other half (beads 37 to 72). In A2, the PAA and PS chains were grafted alternatively. In A3, all the 12 PAA chains were in the middle of the backbone (beads 19 to 54), while PS chains were equally splitted into two groups being located at the two ends of the backbone (beads 1 to 18 and beads 55 to 72).

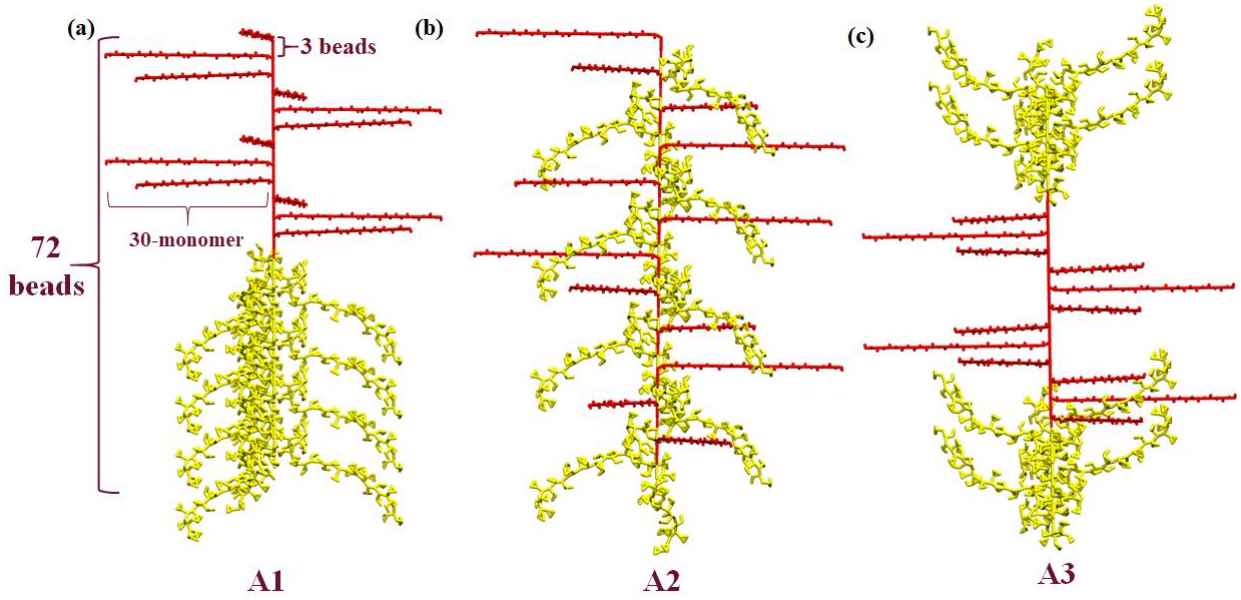

Figure 8.1 Three architectures of CG PS-PAA BBP molecules: **(a)** A1, **(b)** A2, **(c)** A3. The number of beads in backbone (N_{bb}) is 72, and the side chains consisted of 30-monomer CG PAA chains (red, N_{sc} = 30) and 30-monomer CG PS chains (yellow). Every three beads in the backbone is grafted with one side chain.

The nonbonded parameters between COOH and BZ beads by using the LB combining rules were listed in **Table 8.2**. It can be seen in **Figure 8.2** that the densities are overestimated by using LB combining rules to represent the cross interactions between COOH and BZ beads. Therefore, the epsilon values between COOH and BZ beads were scaled by a factor of 0.8 and 0.5, respectively. When the scaling factor is 0.8, i.e. ε[COOH-BZ] = 0.5067 kcal/mol and the sigma value is estimated by the Lorentz combining rule, the densities of propionic acid/benzene mixtures predicted by the CG models lie on the lines of experimental values . Thus, the ε values between COOH and BZ beads used in the following BBP molecules are scaled by a factor of 0.8.

8.3. Results and Discussion

8.3.1 Mixture of Ethylbenzene and Propionic Acid Models

To validate if LB combining rules could capture accurately the cross interactions between benzene beads and COOH beads, we performed simulations of mixtures of propionic acid and benzene (see **Section 3.4** of **Chapter 3** for more simulation details). The nonbonded parameters

obtained by using LB combining rules are listed in **Table 8.2**. **Figure 8.2** shows that the densities of the mixtures were higher than the experimental values, indicating the nonbonded parameters between benzene and propionic acid were overestimated by LB combining rules. To reproduce the experimental values of densities of mixtures, the nonbonded parameters, ε[COOH-BZ], were scaled by a factor of 0.8 and 0.5, which is 0.5067 kcal/mol and 0.3167 kcal/mol (see **Table 8.1**). At a scaling factor of 0.8, the densities predicted by MD simulations show a good agreement with those obtained by experiments (see **Figure 8.2**). Therefore, the scaled nonbonded parameters would be employed in the following simulations of BBPs.

Table 8.1 The nonbonded parameters between COOH and BZ beads

	ε[COOH-BZ] (kcal/mol)	**σ[COOH-BZ] (Å)**
LB combining rules	0.6333	3.9572
Scale ε by 0.8	0.5067	3.9572
Scale ε by 0.5	0.3167	3.9572

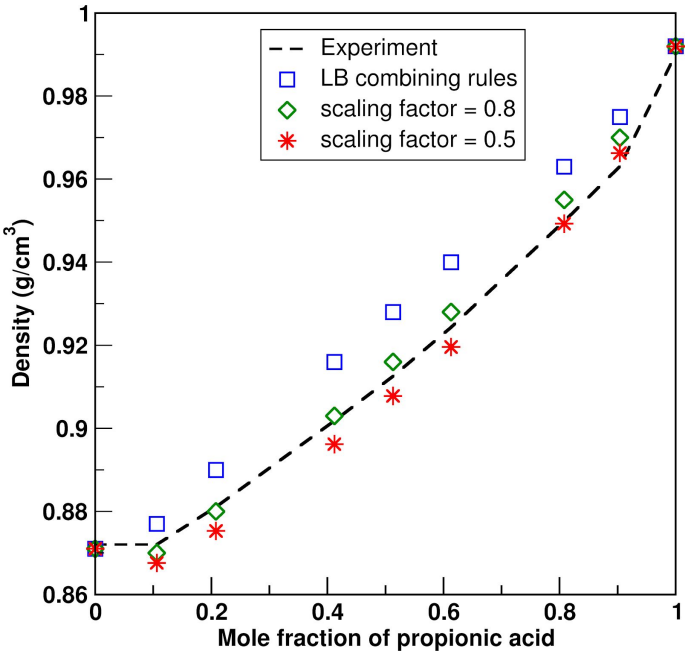

Figure 8.2 The densities of propionic acid/benzene mixtures at 300 K.

8.3.2 Conformations of PS-PAA in Solvents

The conformations of three BBP molecules in pure DMF are shown in **Figure 8.3**. Simulation details could be found in **Section 3.4** of **Chapter 3**. It can be seen that all the side chains are stretched in three BBP molecules in DMF. The averaged R_g values of the side chains (both PS and PAA) and the backbone are shown in **Table 8.2**. The average R_g values of the backbones of A1, A2, and A3 are 34.3 Å, 38.2 Å and 36.3 Å. It's obvious that the A2 structure exhibits a more stretched backbone than A1 and A3. For the side chains in all the three structures, A1 to A3, their R_g values are quite similar, ~14.5 Å for both PS and PAA.

However, when the solvent is water, the conformations of the three BBP molecules are drastically changed. As shown in **Figure 8.4 (a)**, the PS and PAA chains of A1 collapsed and aggregated into two ball-like structures: the yellow one for PS and the red one for PAA. The average R_g values of A1's backbone, PAA and PS side chains in the yellow and red spheres are 28.0 Å, 11.3 Å, 10.6 Å, respectively (see **Table 8.2**), which are smaller than those in DMF. For A2, the overall structure is a sphere with PAA at the outer surface due to the hydrophilicity of PAA. The average R_g value of the backbone of A2 is the smallest among all the three BBP molecules, 23.3 Å, while that for A3 is the largest, 29.6 Å. Three spheres were observed in A3.

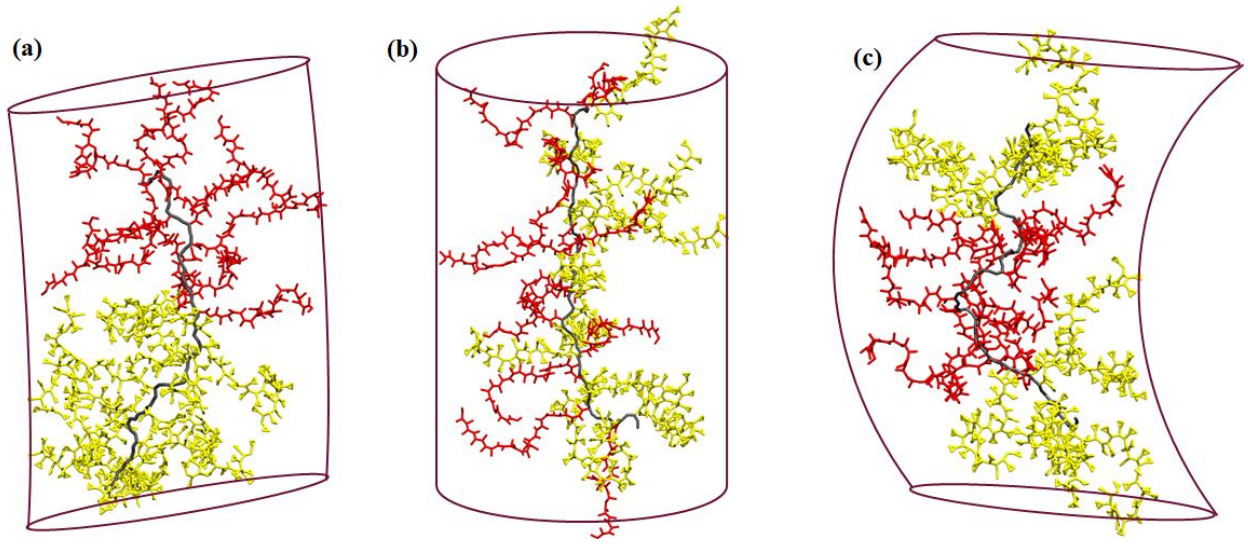

Figure 8.3 The conformation of PS-PAA BBP molecules in pure DMF. Cylinders were drawn to guide the eye. DMF molecules are not shown for clarity.

Table 8.2 The averaged R_g values of the backbone and side chains in three BBP molecules: A1, A2, A3 in pure DMF and pure water.

	Average R_g of BBP in DMF			**Average R_g of BBP in water**		
	Backbone	PS	PAA	Backbone	PS	PAA
A1	34.3±2.0	14.3±1.5	14.5±1.7	28.0±1.9	11.3±1.5	10.6±1.6
A2	38.2±1.6	14.3±1.5	14.6±1.6	23.3±0.4	11.5±1.5	10.7±1.7
A3	36.3±1.8	14.5±1.6	14.3±1.5	29.6±2.4	10.8±1.8	10.8±1.2

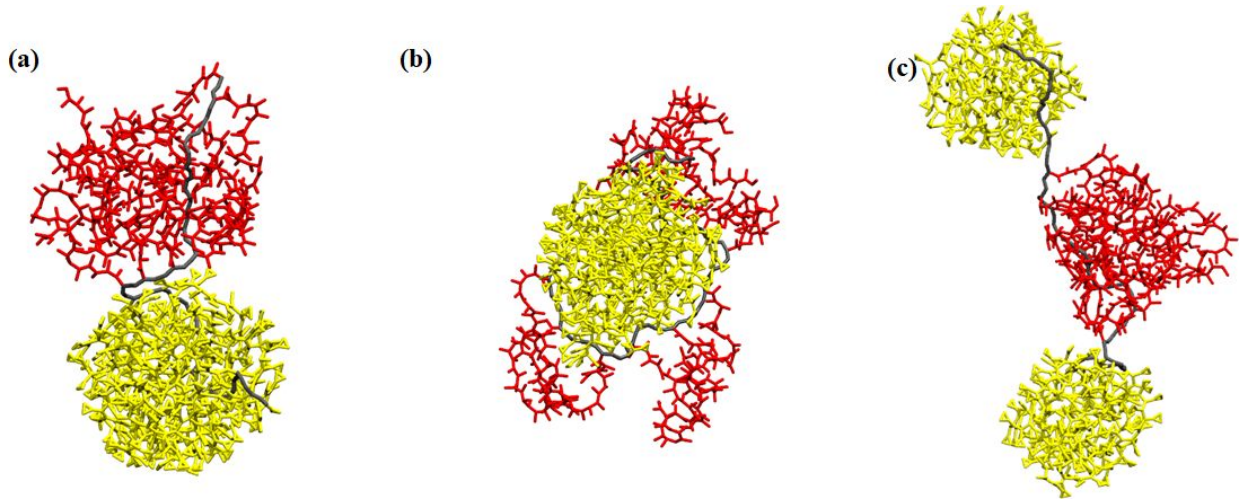

Figure 8.4 The conformation of PS-PAA BBP molecules in pure water with initial structures of **(a)** A1, **(b)** A2, **(c)** A3. Water molecules are not shown for clarity.

8.4 Conclusion

Coarse-grained bottlebrush copolymer models (PS-PAA) with three different initial structures were constructed and their conformations in pure water and pure DMF were investigated. In DMF, all the three BBP molecules were in a stretched state, exhibiting a cylinder-like structure. The R_g values of their backbones ranges from 34.3 Å to 38.2 Å, among which the BBP molecule with PS and PAA alternatively grafted on the side chains showed the largest backbone R_g value, ~38.2 Å. On the other hand, the BBP molecules collapsed in water, and their average R_g value in the backbones decreased to 23.3-29.6 Å. It's interesting that the one

with PS and PAA alternatively grafted on the side chains showed a spherical structure with PS as the core and PAA on the outer face.

8.5 Future Work

In the future, more analysis and BBPs with different architectures will be carried out. For example, the RDFs between solvents and different blocks of BBPs. This would provide insights into understanding the local solvent structure of solvents around BBPs. The effects of factors including grafting densities, shapes, and components of binary solvents of DMF/water on the conformation of the BBPs will be investigated.

References

(1) Chockalingam, R.; Natarajan, U. Structure and Solvation Thermodynamics of Asymmetric Poly (acrylic Acid)-B-Polystyrene Polyelectrolyte Block Copolymer Micelle in Water: Effect of Charge Density and Chemical Composition. *Polymer* **2018**, *158*, 103–119.

(2) Peng, D.; Feng, C.; Lu, G.; Zhang, S.; Zhang, X.; Huang, X. A Starlike Amphiphilic Graft Copolymer with Hydrophilic Poly(acrylic Acid) Backbones and Hydrophobic Polystyrene Side Chains. *J. Polym. Sci. A Polym. Chem.* **2007**, *45* (16), 3687–3697.

(3) Verduzco, R.; Li, X.; Pesek, S. L.; Stein, G. E. Structure, Function, Self-Assembly, and Applications of Bottlebrush Copolymers. *Chem. Soc. Rev.* **2015**, *44* (8), 2405–2420.

(4) Luo, H.; Santos, J. L.; Herrera-Alonso, M. Toroidal Structures from Brush Amphiphiles. *Chem. Commun.* **2014**, *50* (5), 536–538.

(5) Bolton, J.; Rzayev, J. Tandem RAFT-ATRP Synthesis of Polystyrene–Poly(Methyl Methacrylate) Bottlebrush Block Copolymers and Their Self-Assembly into Cylindrical Nanostructures. *ACS Macro Lett.* **2012**, *1* (1), 15–18.

(6) Cheng, G.; Böker, A.; Zhang, M.; Krausch, G.; Müller, A. H. E. Amphiphilic Cylindrical Core−Shell Brushes via a “Grafting From” Process Using ATRP. *Macromolecules* **2001**, *34* (20), 6883–6888.

(7) Wessels, M. G.; Jayaraman, A. Molecular Dynamics Simulation Study of Linear, Bottlebrush, and Star-like Amphiphilic Block Polymer Assembly in Solution. *Soft Matter* **2019**, *15* (19), 3987–3998.

(8) Dutta, S.; Wade, M. A.; Walsh, D. J.; Guironnet, D.; Rogers, S. A.; Sing, C. E. Dilute Solution Structure of Bottlebrush Polymers. *Soft Matter* **2019**, *15*, 2928–2941.

CHAPTER 9
MACHINE LEARNING APPROACH FOR ACCURATE BACKMAPPING OF CG MODELS TO ALL-ATOM MODELS

This work presented in this chapter is reported from [An, Y., Singh, S.; Bejagam, K. K., Deshmukh, S. A. Development of an Accurate Coarse-Grained Model of Poly(acrylic acid) in Explicit Solvents , Macromolecules, 2019, 52 (13), 4875-4887], with the permission of AIP Publishing.

Abstract: Backmapping is usually required to obtain atomistic details after coarse-grained (CG) molecular dynamics (MD) simulations. In this work, machine learning is studied to backmap CG models to all-atom ones. Six representative molecules with linear and ring-like structures are selected to construct machine learning models for backmapping, which are furan, benzene, hexane, naphthalene, graphene and fullerene. Dataset for training the machine learning models have been constructed with the coordinates of each bead in the CG model as input and the positions of each atom in their corresponding all-atom model as output. Four different machine learning regression models: artificial neural network (ANN), k-nearest neighbors (k-NN), gaussian process regression (GPR) and random forest (RF) have been built for each molecule. The accuracy of the ANN, k-NN and RF models for the molecules with ring-like structures. furan, benzene, naphthalene, graphene and fullerene could reach as high as 0.99, which suggests machine learning is a promising technique for backmapping CG models to all-atom ones.

9.1 Introduction

CG MD simulations have been widely used to study complex phenomena such as self-assembly of peptide amphiphiles.[1,2] However, atomic-level interactions and details are lost in these CG models. To overcome this limitation, they are usually constructed back into all-atom or united-atom models that might provide details on atomistic interactions of interest or continue the simulations at higher resolutions. This process is usually called backmapping or inverse mapping or reverse transformation.

Generally, backmapping involves two steps: (i) generation of the initial atomistic structure, and (ii) relaxation of this initial structure by MD simulations.[3] Ideally, the initially

generated atomistic structures should be accurate and/or equilibrated enough such that the relaxation step becomes unnecessary. To construct the initial atomistic structures, fragment method, geometric rules and even random placement of atoms have been used. However, for most of these methods it's almost impossible to reconstruct a high-quality initial atomistic structure.[4–6] In the case of the fragment method, the initial atomistic groups are selected from a database consisting of fragments, predetermined by mapping control molecules. Shih *et. al.* used this method to place the center of mass of the selected fragments at the position of their corresponding CG beads.[4] This method relies on the R2 score of the fragment library. Brocos *et. al.* employed geometric rules to backmap CG nano-aggregates of lysophospholipid molecules into their atomistic structure.[5] The coordinates of *n* atoms represented by a CG bead were determined by placing the first atom at the position of the CG bead and the other *n-1* atoms were aligned with the segment connecting two CG beads by geometrical interpolation/extrapolation. The geometrical extrapolation conserves the bond and angle values while avoiding the overlap between neighboring atoms.[5] But it has limitations in recovering atomistic structures from residual-level CG models. The random placement method first locates atoms around the CG beads randomly, and then uses simulated annealing MD simulations to optimize the all-atom structure, which couples all-atom and CG force fields (FFs). Finally, the CG force field is removed to relax the structure. This process is relatively slow and intricate and also needs software development efforts. Although used widely, all these aforementioned methods for backmapping need a relaxation step to obtain optimized or equilibrated structures.

In recent years, machine learning (ML) techniques have been integrated with MD simulations for various applications. For example, both the supervised and unsupervised ML methods have been utilized to assist the development of force field parameters,[7,8] analyze simulation trajectories,[9–11] and predict free energy landscapes.[12] To the best of our knowledge there is only one recent publication from Wang *et. al.* that builds an auto-encoder framework for both mapping all-atom to CG models and the reverse.[13] They used the encoder to build CG models of ortho-terphenyl (OTP), aniline (C6H7N), alanine dipeptide and alkanes with different resolutions, in order to determine the best mapping scheme. Then these CG models were backmapped to all-atom models by a decoder. They found that the bond length between carbon-carbon atoms in the decoded all-atom structure shows a broader distribution with a

smaller average value compared with that in the original all-atom structure. The auto-encoder framework is usually classified as an unsupervised learning method.

To the best of our knowledge, the present study is a first attempt to use supervised ML methods for backmapping of CG models to all-atom models. Different from the unsupervised learning method, the performance of supervised ML models could be quantified by using R2 score or root mean square error (RMSE). Here, four ML regression models were used to predict the initial coordinates of atoms in all-atom models by using positions of beads in CG models as input. Specifically, artificial neural networks (ANN), k-nearest neighbor (k-NN) regression, gaussian process regression (GPR) and random forest (RF) regression were used to build four supervised predictive models for backmapping. Prediction the R2 score of these models was compared to evaluate their performance. To build a dataset, all-atom MD simulations were performed and the coordinates of all the atoms were saved. Based on proposed mapping schemes for individual molecules, the center of mass of atomic groups were calculated as the coordinates of beads in CG models. The coordinates of CG beads were treated as input and those of atoms as output. As test cases, we have studied ring-like molecules such as benzene, furan, and naphthalene, planar structure of a graphene flake, linear hexane molecule, and spherical fullerene.

9.2 Methods

9.2.1 All-Atom and CG models

Six representative molecules were selected to create their all-atom and CG models. A 2:1 mapping scheme was used to create CG models of furan, benzene, hexane, naphthalene and fullerene. A 4:1 mapping scheme was employed to construct the CG model of graphene to retain its hexagonal structure.

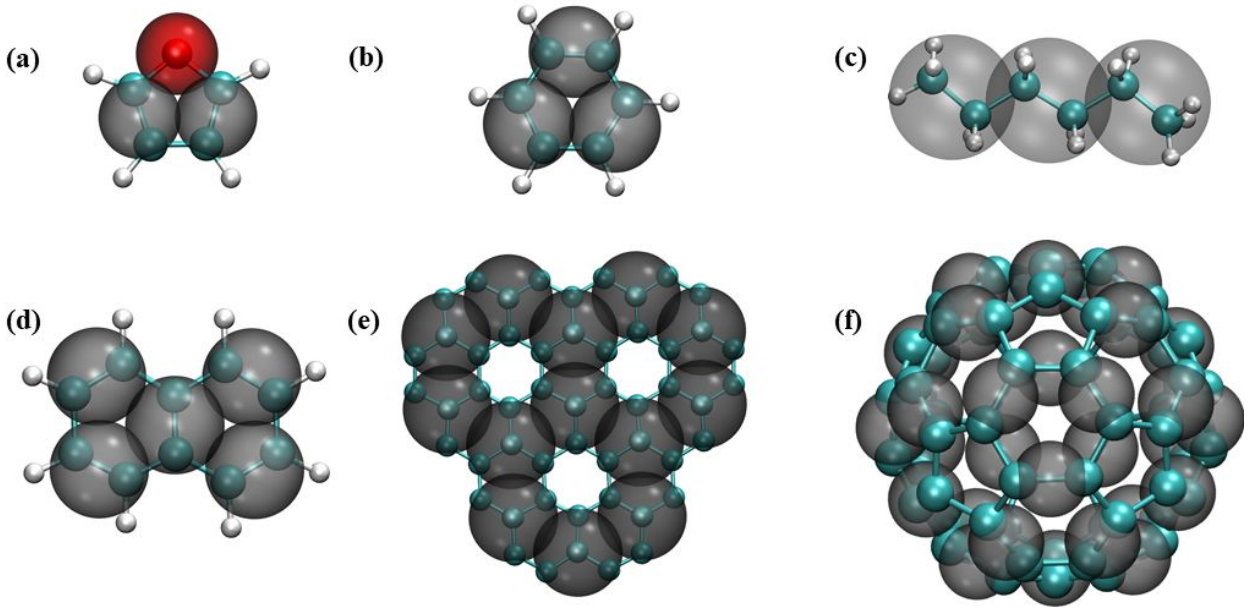

Figure 9.1 The mapping schemes for (a) furan, (b) benzene, (c) hexane, (d) naphthalene, (e) graphene, and (f) fullerene. (a-d) and (f) use a 2:1 mapping scheme, while (e) uses a 4:1 mapping scheme.

9.2.2 Dataset Construction and Description

The dataset to train the ML models consisted of coordinates of beads in CG models as input, and the coordinates of atoms in all-atom models as output. The R^2 score and robustness of ML models rely on the fidelity of the dataset. Specifically, the training dataset should cover a large range of molecular structures.[14,15] To generate the input data for training and testing of all the ML models, 1000 molecules were randomly packed in a simulation box and then equilibrated for 5 ns in NPT ensemble. All-atom MD simulations were carried out with CHARMM force-field by using the NAMD package.[16,17] Periodicity was applied in all the three directions. A real space cutoff of 12 Å was used to truncate the nonbonded interactions with a switching function applied at 9 Å to truncate the van der Waals potential energy smoothly at the cutoff distance. Long-range Coulombic interactions were treated using particle mesh Ewald with an accuracy of 1×10^{-6}. A pair list distance of 15 Å was used to store the neighbors of a given bead. The equations of motion were integrated by the velocity Verlet algorithm with a timestep of 1 fs. Langevin thermostat and barostat were used to keep the temperature at 300 K and pressure at 1 bar. The positions of atoms were saved every 1 ps. We extracted a trajectory of 100 randomly

selected molecules in the final 10 ps (100 frames) of 5 ns (50,000 frames), which contained 10,000 configurations in total for generating the dataset. The center of mass of every molecule from these 10,000 configurations was placed to the origin, which can be treated as normalizing the coordinates of each atom to a narrow range. This resulted in a small dimensional space for the dataset for training, to help improve the R^2 score of ML models. An example of the dataset for hexane is shown in **Figure 9.2**. An example of the dataset for hexane is shown below:

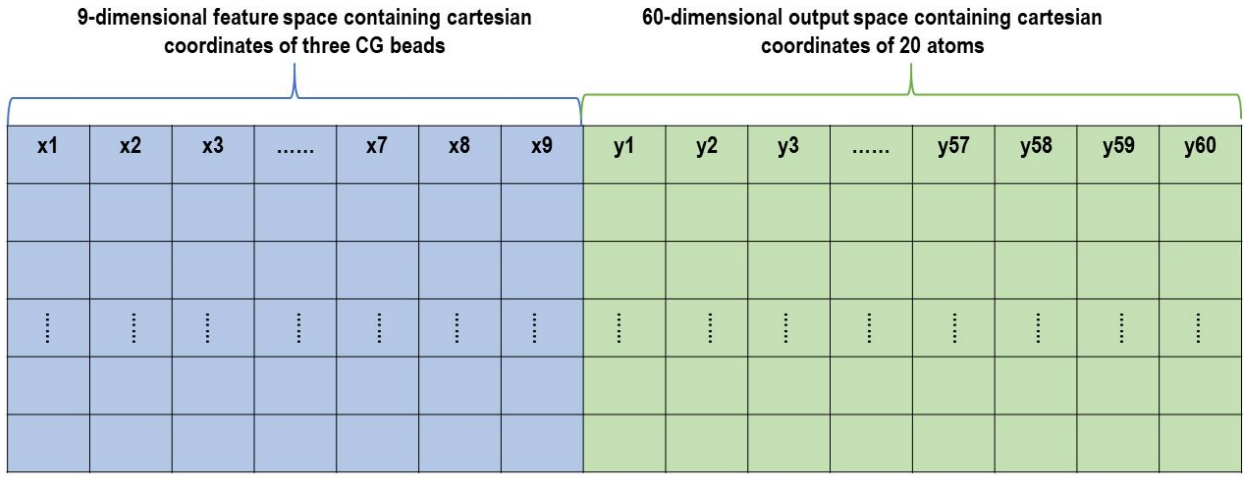

Figure 9.2 The dataset of hexane for building ML models. The coordinates of all-atom are in blue and those of CG beads are in green. Similar datasets were used for other molecules used in the present study.

Several methods such as *k*-fold, leave-one-out, etc. can be used to generate/split a dataset to train the ML models.[15,18] As one would expect each method has its own advantage and disadvantage so care must be taken while choosing these methods. Each dataset was randomly split so that 80 % of the data for training and the remaining 20 % for testing. For training, *k*-fold cross-validation was used to prevent overfitting of the ML models.[15] As shown in **Figure 9.3**, the training data is split k folds (k = 5 in this study), where one fold is used as a validation set and the other k-1 folds as training sets. This process is repeated k times with each fold as a validation set and thus k ML models were generated. The performance of these k ML models was averaged to get the final average R^2 score and standard deviation. This method is generally robust and yields results with the reasonable R2 score, as demonstrated in the present study.

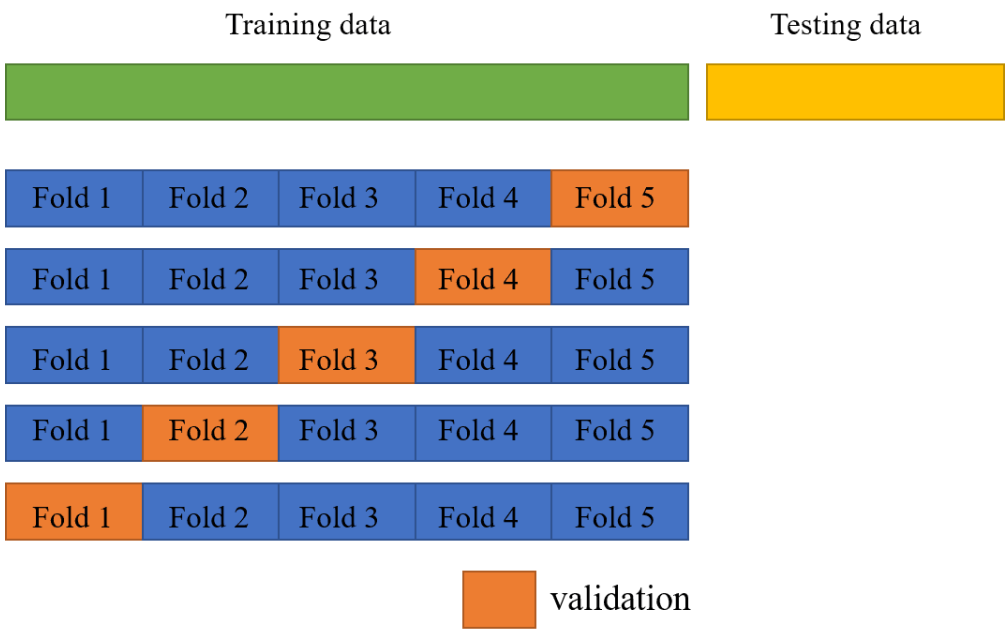

Figure 9.3 The schematic for k-fold (k = 5) cross validation.

9.2.3 ML Model Development

ANN is one of the most popular ML models, which has been widely used in information technologies such as image classification, and natural language processing.[19] Recently, it has also been used in computational materials science.[8,13,20] Here, we construct different ANN regression models by changing the number of hidden layers (1 to 4) and hidden nodes (5 to 30), to understand the effect of the number of hidden layers and hidden nodes on the R^2 score of the ANN model. This data is shown in Appendix **E**. The ReLu activation function was used in all the ANN models. In optimization, the Adam algorithm is used to calculate the derivatives of the loss function (mean of square errors in this study) with respect to weight and bias in ANN models. It's found that the testing R2 score of ANN models with two hidden layers increases drastically as the number of hidden nodes increases from 5 to 10. Whereas it changes slightly as the number of hidden nodes is further increased to 30. The number of hidden layers on ANN models (10 hidden nodes in each layer) have little impact on the testing R2 score. Hence we used the ANN models with two hidden layers and 10 hidden nodes in each layer. The training and testing of ANN models were achieved by using the Scikit-learn package.[21]

k-NN is a simple ML model that uses the distance between data points to solve classification or regression problems.[22] The most used distance metric is the Euclidean distance

to measure the similarity between two points. Based on this, the data points with the *k* shortest distance from the unlabeled data would be selected to assign classes or values of the unknown data. Here, we studied the performance of k-NN models with different k values (3, 5, 8) in Table **E1** of Appendix **E**. The testing R2 score of k-NN models is increased slightly as the k value is increased from 3 to 8. Unless specified, results of k-NN models with k = 5 are discussed in the following sections.

RF is an ensemble model consisting of several decision trees to predict the final labels/values by selecting a subset of features for each decision tree. The performance of RF models is usually better than that of one single decision-tree model. The minimum number of samples splitting a node is set to be 2, and the minimum number of samples in a leaf node. The number of features to consider for best splitting the total number of features. Effects of max_depth and n_estimators are explored in Appendix E. Specifically, we studied the RF models with max depth varying from 5 - 20 and n_estimators from 5 to 20 (see **Figure E4** of Appendix **E**). In building the RF models, MSE (mean square error) is used to split the trees and the maximum depth (max_depth) of each tree is set to be 10. The number of decision trees (n_estimators) in an ensembled RF model is also 10.

GPR is a nonparametric Bayesian approach to infer the probability distribution over all possible values.[23] In the Gaussian prior, the collection of training and testing data points are joint multivariate Gaussian distributed, which can be described in **Equation S1**.

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right) \quad \textbf{Equation 9.1}$$

Where y is the output values of training data X, f_* is the predicted label/output of testing data X_*. K is the kernel or covariance. The constant kernel with the radial basis function (RBF) is one of the popular kernels, as shown in **Equation S2**.

$$k(x, x') = \sigma_f^2 \, exp(-\frac{1}{2\ell^2}||x - x'||^2) \quad \textbf{Equation 9.2}$$

Where σ is the signal variance, and l is the length scale. These two parameters are optimized during training to maximize the marginalized log-likelihood of the training data in **Equation S3**.

$$\log(p(y|X)) = \log N(0, K(X,X)) = -\tfrac{1}{2}\, yK^{-1}(X,X)y - 1/2\log|K(X,X)+\sigma_n^2 I| - N/2\log(2\pi)$$ **Equation 9.3**

According to multivariate Gaussian theorem, the predicted values f* are in a normal distribution with mean value $\bar{f}^*$ and covariance Σ*, which are shown below:

$$f^*|X, y, X^* \sim \mathcal{N}(\bar{f}^*, \Sigma^*)$$ **Equation 9.4**

$$\bar{f}^* = \mu^* + K(X^*,X)[K(X,X) + \sigma_n^2 I]^{-1}(y - \mu)$$ **Equation 9.5**

$$\Sigma^* = K(X^*,X^*) - K(X^*,X)[K(X,X) + \Sigma_n^2 I]^{-1}K(X,X^*)$$ **Equation 9.6**

9.2.4 Performance of ML Models

The performance of ML regression models is determined by the root of mean square error (RMSE) and the R2 score, which are described below.

RMSE: TheRMSE of predicted values/vectors $\mathbf{y}_{\mathbf{pred}}$ compared with true values/vectors $\mathbf{y}_{\mathbf{true}}$ is calculated by using the following equation. N represents the dimensions in the output space.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{i,true} - y_{i,pred})^2}$$ **Equation 9.7**

R^2 score: Similar to classification problems, the R^2 score is also defined for the regression models as shown in **Equation 8**. $\bar{y}_{i,true}$ is the average value of the i^{th} element in the true vector $\mathbf{y}_{\mathbf{true}}$.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{i,true} - y_{i,pred})^2}{\sum_{i=1}^{N}(y_{i,true} - \bar{y}_{i,true})^2}$$ **Equation 9.8**

Uncertainty Quantification: To test the robustness of ML models, bootstrapping is employed to obtain the uncertainty quantifications of these ML models. Bootstrapping is a widely used method to resample dataset.[24,25] Here, we resampled the training dataset with replacement for 500 times, and each resampled dataset was used to build ML models. As a result, 500 ML models could be built and were tested against the testing dataset to calculate testing the R2 score. The

histogram of these the R2 score was plotted with 95% confidence interval and the average the R2 score shown in the histogram.

9.3. Results and Discussion

9.3.1 R2 Scores of ML Models

The comparison of the the R2 score of four ML models: ANN, k-NN, GPR and RF for backmapping CG models of all six molecules: furan, benzene, hexane, naphthalene, graphene, and fullerene to their corresponding all-atom models is shown in **Figure 9.4**. For furan, benzene, naphthalene, and graphene, the four ML models showed 0.94 - 0.99 of the R2 score for training and testing. This suggests that all the four ML models could predict the atomistic structures of small ring-like, and planar molecules accurately. To ensure that the ANN, k-NN, GPR, and RF did not overfit the data, two additional regression models: kernel ridge regression (KRR) and support vector regression (SVR) were also studied and showed R2 scores of 0.98-0.998 in **Table E2** of Appendix **E**.

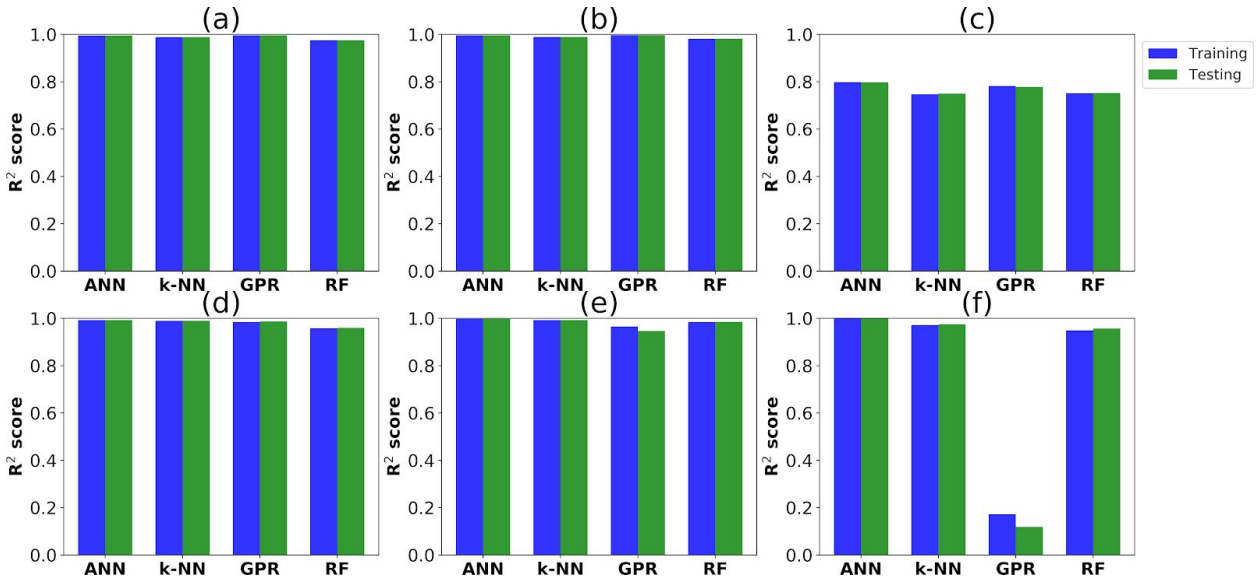

Figure 9.4 The training and testing R^2 scores of different ML models for backmapping CG **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, and **(f)** fullerene models.

The uncertainty quantification of the ML models for furan are shown in **Figure 9.5** for furan, and for the other molecules in **Figures E5 and E6** of Appendix E. **Figure 9.5** shows the

distributions of the testing R^2 scores of four ML models trained against 500 different training datasets built by bootstrapping method. The testing R2 score for ANN models ranges from 0.9915 to 0.9965, with the 95 % confidence interval of 0.9935 - 0.9959. This suggests that the performance of ANN models are stable with slight change of testing the R2 score. Similar observations could be found on the performance of the other machine learning models: k-NN, GPR and RF.

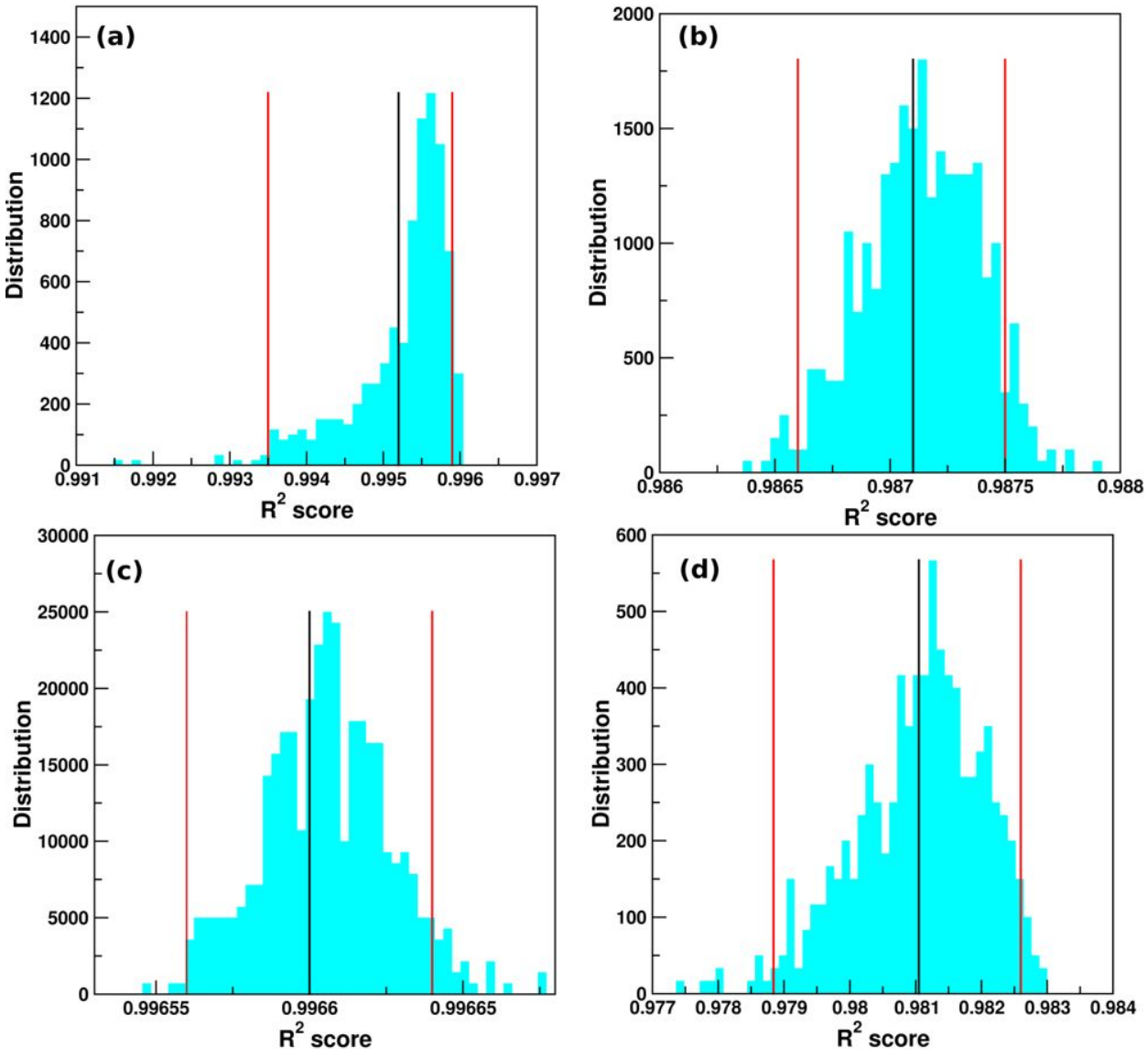

Figure 9.5 The probability distribution of testing the R2 score of **(a)** ANN, **(b)** k-NN, **(c)** GPR, and **(d)** RF for constructing all-atom configurations of furan. The areas between red lines represent 95 % confidence intervals, and black lines represent the average values.

However, the R^2 score for obtaining the all-atom models of hexane is only around ~75% regardless of the types of ML models. This is mainly because there may exist more than one configuration of the all-atom model corresponding to one CG model. As shown in **Figure 9.6**, the backmapped all-atom model 1 is the mirror image of backmapped all-atom model 2, when they share the same CG model. Training of any regression based ML model with such data will result in predictions that are mean of these two structures. Because of this, the predicted all-atom configuration is the mean of the two all-atom models 1 and 2. This is similar to the results reported in reference [13]. They used the decoder in an auto-encoder framework to reconstructure the atomistic structure of $C_{24}H_{50}$. The backbone of carbon chains are the mean reconstruction of an ensemble of carbon chain poses. Therefore, it's difficult to train ML models with a high R^2 score for backmapping CG hexane or other linear alkane models.

For fullerene, the R2 score of GPR models was only ~0.1, while those of the other three ML models was at 99% the R2 score. The low R^2 score of the GPR model could be attributed to the high dimensionality of the output which was 180. In training a GPR model, matrix inversion is involved as shown in **Equation 9.3**. This means it's quite computationally expensive when the dimension space is high, which makes it difficult to train a GPR model. By comparing the four ML models, ANN model always showed comparable or better performance than the other three ML models in backmapping of the six CG molecules to their corresponding all-atom structures. The predicted coordinates of all-atom models backmapped by ANN models are shown in **Figure 9.7**. The points are aligned closely to the diagonal with RMSE values ranging from 0.087 to 0.133 in **Figure 9.7 - (a, b, d - f)**. These small RMSE values are consistent with the high R^2 scores of ANN models in backmapping furan, benzene, naphthalene, graphene and fullerene. However, the RMSE is as high as 1.115 and the predicted coordinates for all-atom hexane molecules are deviating far from their true values, as shown in **Figure 9.7 - (c)**. This could be attributed to the low R^2 score of ANN models in backmapping CG hexane models as discussed above.

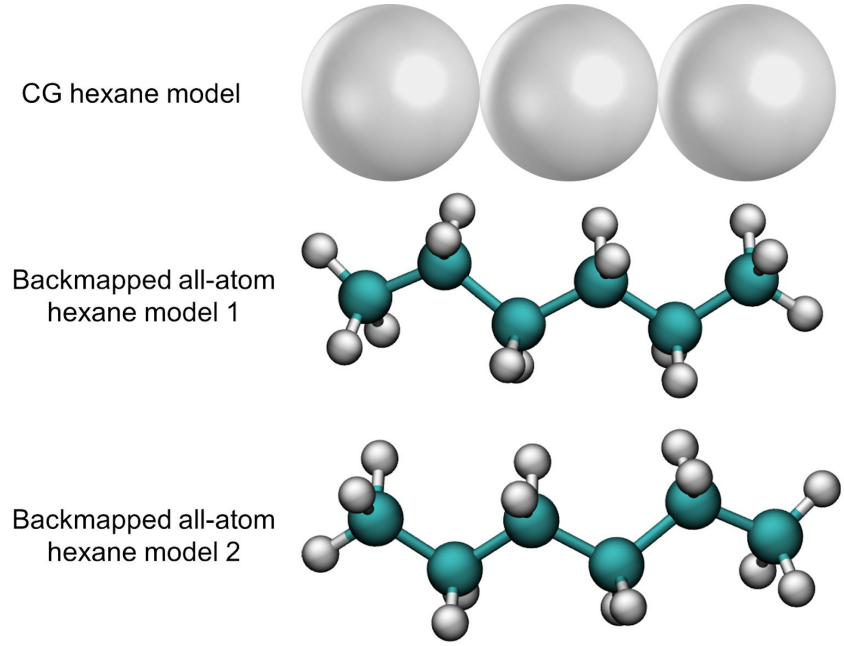

Figure 9.6 The CG hexane model and its two backmapped all-atom models.

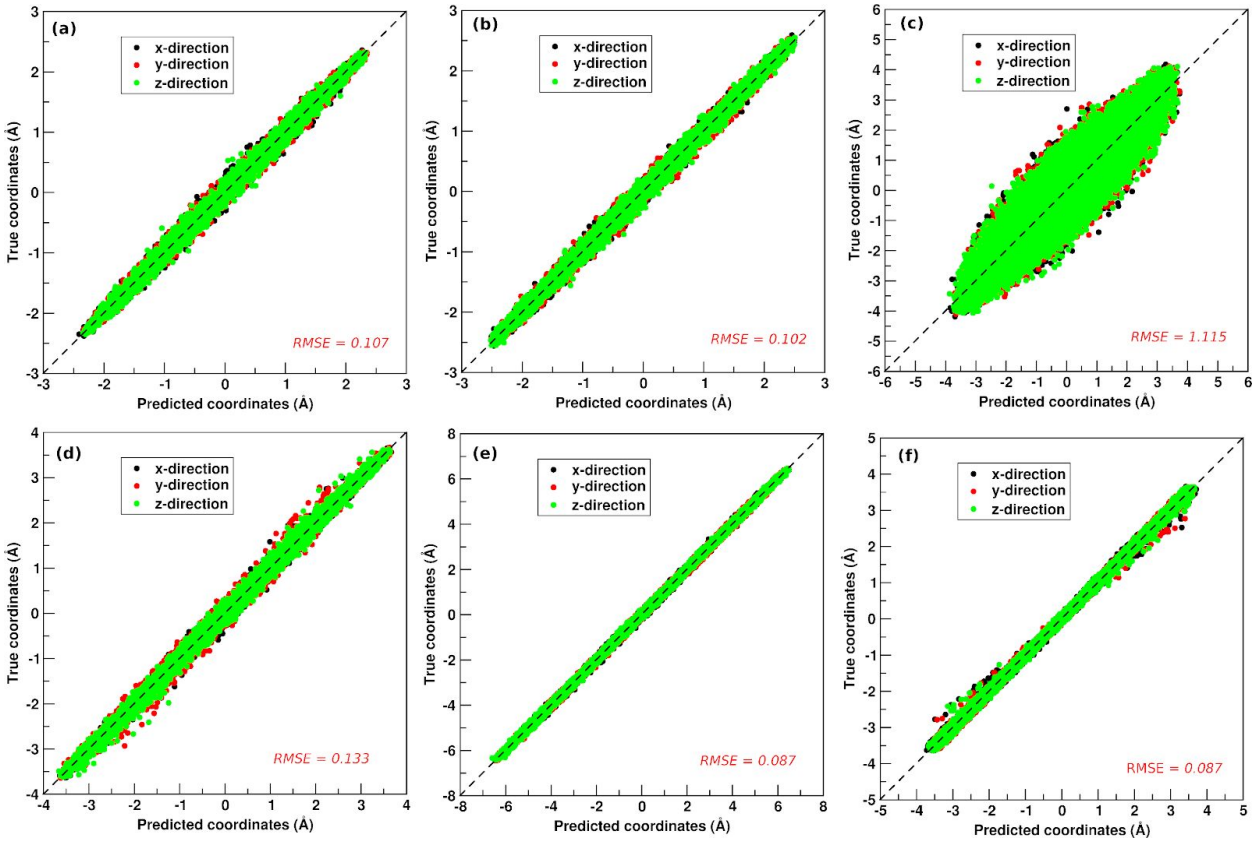

Figure 9.7 Predicted x, y and z coordination values of atoms in all-atom models of **(a)** furan, **(b)**

benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene and **(f)** fullerene compared with their true values. ANN models with two hidden layers and 10 hidden nodes in each layer are employed.

To further validate the R^2 scores of the ML models in predicting the atomistic structures, the bond length distributions in all the six predicted all-atom molecule models by ANN are shown in **Figure 9.8**. It's obvious that the bond length distributions for the predicted all-atom models almost overlap with their corresponding ground truths calculated from mapped MD simulations trajectory, except for the hexane molecule. Because the predicted hexane molecule can only capture its mean poses in an ensemble, the average carbon-carbon (C-C) distance is shorter than its true average value, and the distribution is also broader.

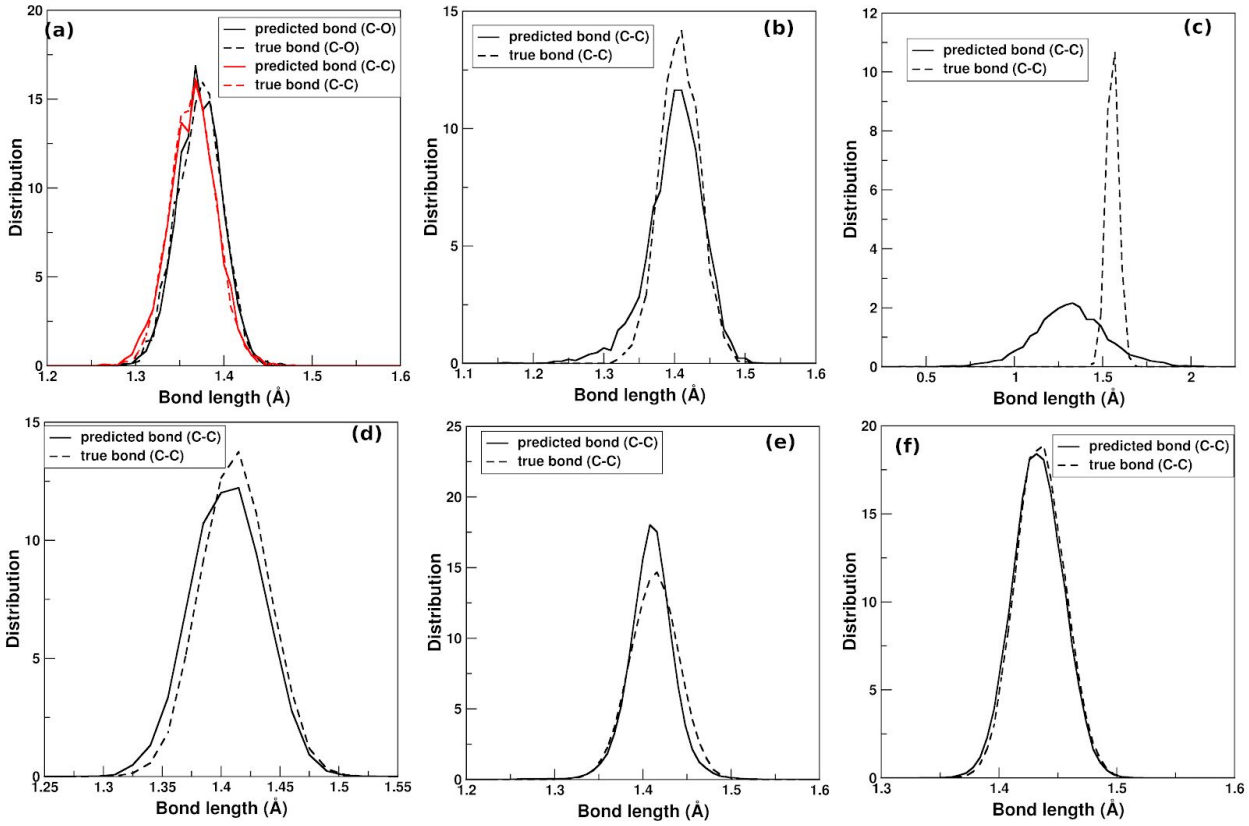

Figure 9.8 The bond length distributions in predicted atomistic configurations of **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, and **(f)** fullerene in comparison with their corresponding true atomistic structures.

9.3.2 Effects of Dataset Sizes

It's known that the R^2 scores of the ML models rely heavily on the quantity and quality of the dataset. Generally, large dataset could help train more accurate ML models. On the other, it requires a large amount of computational resources to train ML models when the dataset size is too big. Moreover, generating large dataset may be computationally expensive. Here, we explored the effects of dataset size on the R^2 scores of ANN models by using 1000, 5000, and 10000 samples. When the total number of samples is 1000, it's challenging to train the ANN models (10/10, two hidden layers with 10 nodes in each layer) with a high R^2 score. For example, the ANN models for backmapping CG naphthalene and fullerene molecules are only 0.865 and 0.892, respectively. As the dataset increased to 5000, all the ANN models except that for hexane showed a R2 score of more than 0.99, and this increased a little bit as the dataset size increased further to 10,000. This indicates 5000 samples are large enough for developing accurate ANN models for backmapping of these molecules. It's noticeable that the performance of the ANN model for hexane is not improved as the size of the dataset increases.

Table 9.1 the R2 score of ANN models trained on different sizes of dataset. The number of hidden nodes and hidden layers are 10 and 2, respectively.

molecules	Testing the R2 score		
	1000 samples	5000 samples	10,000 samples
Furan	0.983±0.001	0.995±0.000	0.995±0.000
Benzene	0.983±0.005	0.996±0.000	0.996±0.000
Hexane	0.799±0.005	0.798±0.000	0.796±0.004
Naphthalene	0.9581±0.017	0.993±0.000	0.997±0.000
Graphene	0.990±0.002	0.997±0.000	0.999±0.000
Fullerene	0.892±0.033l	0.995±0.000	0.997±0.001

9.3.3 Comparison with Backmapping by VMD

To understand the effectiveness of this new ML based backmapping approach, the performance of ML and the built-in backmapping package in VMD[26] is compared. The built-in package in VMD is first implemented in reference [4], which simply places the center of mass of the group of atoms represented by a single CG bead to the location of that bead. Three representative molecules: benzene, hexane and graphene were selected to calculate the RMSE values of coordinates for atoms predicted by ANN models and the built-in package in VMD. In **Table 9.2**, it's obvious that the RMSE are always smaller by using ANN models than by using the built-in package in VMD.

The structures of the true all-atom models, backmapped all-atom models by ANN and by geometry rules are shown in **Figure 9.9**. **Figure 9 - (a-2, c-2)** are the configuration of backmapped all-atom models of benzene and graphene by using ANN. They are similar to their corresponding real structures shown in **Figure 9.9 - (a-1, c-1)**. However, the backmapped all-atom models by using VMD are totally messed up. The predicted benzene molecule is not even planar, and the regular hexagonal structure in graphene is also lost. For the backmaaped all-atom model of hexane in **Figure 9.9 - (b-2, b-3)**, the positions of hydrogen atoms are not accurately reproduced by either the ANN model or VMD, and the bond length and angle values could not be captured. However, the ANN model still performed better than VMD in backmapping CG hexane models due to the smaller RMSE value in **Table 9.2**.

Table 9.2 RMSE for coordinates of back mapped all-atom models by ML and VMD.

	VMD	ML
benzene	1.483	0.108
hexane	2.986	0.907
graphene	0.506	0.079

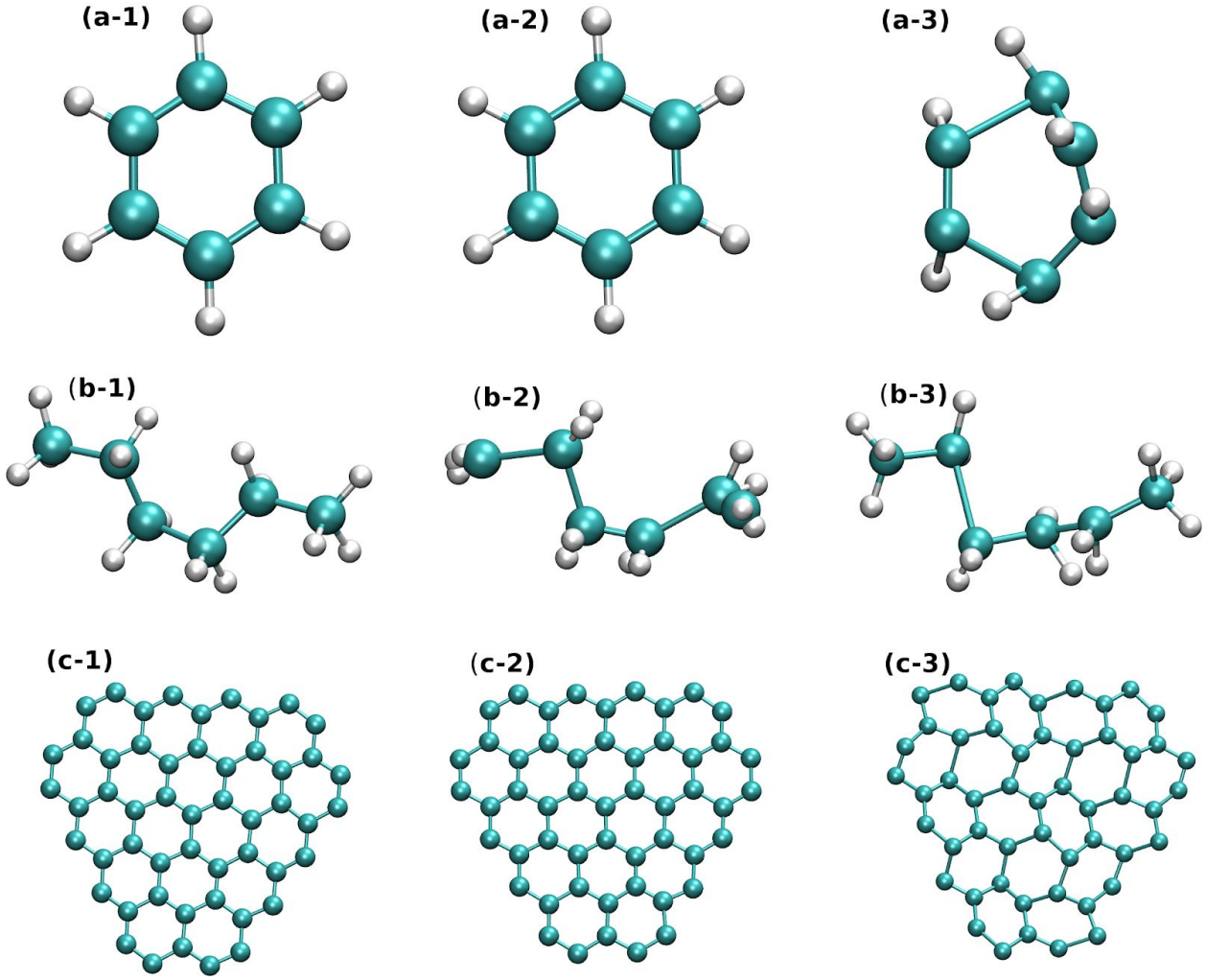

Figure 9.9 Representative configurations of the true all-atom models (a-1, b-1, and c-1) and back mapped all-atom models by ML (a-2, b-2, and c-2) and VMD (a-3, b-3, and c-3).

9.4 Conclusion

Four ML models, ANN, k-NN, GPR and RF have been built to predict the all-atom models of furan, benzene, and naphthalene with a high R^2 score of 0.99. The ANN, k-NN and RF models also predict accurately the atomistic configurations of graphene and fullerene molecules, while GPR is limited by the high dimensionality of the input/output dataset of graphene and fullerene, which are 39/156 and 90/180, respectively. In reconstructing the all-atom model of hexane from a CG one, the R2 score of these four ML models are all at around 0.78. This is essentially because there exists more than one all-atom configuration corresponding to one CG hexane model. The performance of the ANN models are compared with that of the backmapping package in VMD in terms of rebuilding the atomistic structures of benzene, hexane and graphene

as representative examples. The RMSE values of predicted all-atom structures by ANN are smaller than those by the backmapping package in VMD. Although ML has shown promising results in backmaping, more effort needs to be made to explore its application to backmapping large and complex molecules such as protein in the future.

References:

(1) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646–652.
(2) Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular Dynamics Simulations: Advances and Applications. *Adv. Appl. Bioinform. Chem.* **2015**, *8*, 37–47.
(3) Peng, J.; Yuan, C.; Ma, R.; Zhang, Z. Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference. *J. Chem. Theory Comput.* **2019**, *15* (5), 3344–3353.
(4) Shih, A. Y.; Freddolino, P. L.; Sligar, S. G.; Schulten, K. Disassembly of Nanodiscs with Cholate. *Nano Lett.* **2007**, *7* (6), 1692–1696.
(5) Brocos, P.; Mendoza-Espinosa, P.; Castillo, R.; Mas-Oliva, J.; Piñeiro, Á. Multiscale Molecular Dynamics Simulations of Micelles: Coarse-Grain for Self-Assembly and Atomic Resolution for Finer Details. *Soft Matter* **2012**, *8* (34), 9005–9014.
(6) Rzepiela, A. J.; Schäfer, L. V.; Goga, N.; Risselada, H. J.; De Vries, A. H.; Marrink, S. J. Reconstruction of Atomistic Details from Coarse-Grained Structures. *J. Comput. Chem.* **2010**, *31* (6), 1333–1343.
(7) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, 4667–4672.
(8) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5*, 755–767.
(9) Bejagam, K. K.; An, Y.; Singh, S.; Deshmukh, S. A. Machine-Learning Enabled New Insights into the Coil-to-Globule Transition of Thermosensitive Polymers Using a Coarse-Grained Model. *J. Phys. Chem. Lett.* **2018**, 6480–6488.
(10) Rajan, A.; Freddolino, P. L.; Schulten, K. Going beyond Clustering in MD Trajectory Analysis: An Application to Villin Headpiece Folding. *PLoS One* **2010**, *5* (4), e9890.
(11) Singh, S. K.; Bejagam, K. K.; An, Y.; Deshmukh, S. A. Machine-Learning Based Stacked Ensemble Model for Accurate Analysis of Molecular Dynamics Simulations. *J. Phys. Chem. A* **2019**, *123* (24), 5190–5198.
(12) Sidky, H.; Whitmer, J. K. Learning Free Energy Landscapes Using Artificial Neural Networks. *J. Chem. Phys.* **2018**, *148* (10), 104111.
(13) Wang, W.; Gómez-Bombarelli, R. Coarse-Graining Auto-Encoders for Molecular Dynamics. *npj Computational Materials 5* (125), 1–9.
(14) Singh, S. K.; Bejagam, K. K.; An, Y.; Deshmukh, S. A. Machine-Learning Based Stacked Ensemble Model for Accurate Analysis of Molecular Dynamics Simulations. *J. Phys. Chem. A* **2019**, *123* (24), 5190–5198.
(15) Tan, P.-N.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*; Pearson Education, 2019.

(16) Jiang, W.; Hardy, D. J.; Phillips, J. C.; Mackerell, A. D., Jr; Schulten, K.; Roux, B. High-Performance Scalable Molecular Dynamics Simulations of a Polarizable Force Field Based on Classical Drude Oscillators in NAMD. *J. Phys. Chem. Lett.* **2011**, *2* (2), 87–92.
(17) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
(18) T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*; Springer Verlag: New York, 2001.
(19) Abiodun, O. I.; Jantan, A.; Omolara, A. E.; Dada, K. V.; Mohamed, N. A.; Arshad, H. State-of-the-Art in Artificial Neural Network Applications: A Survey. *Heliyon* **2018**, *4* (11), e00938.
(20) Chan, H.; Cherukara, M.; Loeffler, T. D.; Narayanan, B.; Subramanian K R. Machine Learning Enabled Autonomous Microstructural Characterization in 3D Samples. *npj Computational Materials* **2020**, *6* (1), 1–9.
(21) F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(22) Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. *Ann Transl Med* **2016**, *4* (11), 218.
(23) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2006.
(24) Li, Z.; Omidvar, N.; Chin, W. S.; Robb, E.; Morris, A.; Achenie, L.; Xin, H. Machine-Learning Energy Gaps of Porphyrins with Molecular Graph Representations. *J. Phys. Chem. A* **2018**, *122* (18), 4571–4578.
(25) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer Science & Business Media, 2009.
(26) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38, 27–28.

CHAPTER 10
FUTURE WORK DIRECTIONS

10.1 Development of Transferable Coarse-Grained Models of Small Drug Molecules

Computational simulations offer effective and low-cost solutions to drug screening/discovery compared with the traditional experimental methods. In drug discovery, the protein-ligand interactions are the main research focus, which is usually performed by docking algorithms.[1,2] Although the docking algorithms have been widely used for high-throughput screening of drug molecules, the accuracy of them is limited. All-atom MD simulation based on empirical force fields could provide more accurate interaction prediction of the binding interactions between proteins and ligands, but they are usually computationally costly. CG MD simulations have recently emerged as an alternative tool to studying the protein-ligand interactions.[3] CG MD alleviates the shortcomings of docking and all-atom MD simulations. They represent atomistic groups with beads and could achieve high computational speed.

Accurate CG force fields are the basis for performing CG MD simulations to estimate the protein-drug interactions. Developing CG models of drug molecules is usually a time-costly process, due to the diversity of drug molecules and emerging new drugs each year. Hoffman et al. have performed 630 CG MD simulations to study the permeation of drug molecules across the protein membranes by using MARTINI FF.[4] However, validations of these CG drug molecules are unknown. Souza et al. studied the binding of several ligands with proteins such as T4 lysozyme.[3] These ligands include benzene, phenol, indole, thieno-pyridine, toluene, ethylbenzene, and n-propylbenzene. These ligands cover most of the hydrophobic groups in drug molecules, but they are only part of the functional groups composing drug molecules and the interactions of the whole drug molecules is limited.

Here, I propose to use ANN-PSO framework to develop a databank of CG models for drug molecules. This framework has been successfully used in developing various CG models of small molecules, such as DMF and amino acids.[5,6] Compared with the traditional method which optimized CG models manually, the ANN-PSO framework could help achieve the accurate FF parameters in a fast manner. To develop transferable CG models of small drug molecules, the first step is to design reasonable mapping schemes for all the studied drug molecules. To be compatible with the MARTINI force field (FF), we also adopt a 4:1 mapping scheme where four heavy atoms are represented by one bead.[7] The second step is to develop the bonded and

nonbonded force field parameters. The step is achieved by combining top-down and bottom-up approaches. As for the bottom-up approach, all-atom MD simulations are performed to achieve the bond and angle distributions in the small molecules. From these distributions, we can estimate the range for the bonded parameters of the CG models. For the top-down approach, we use experimentally measured properties as target values in the ANN-PSO framework to optimize the bonded and nonbonded parameters.

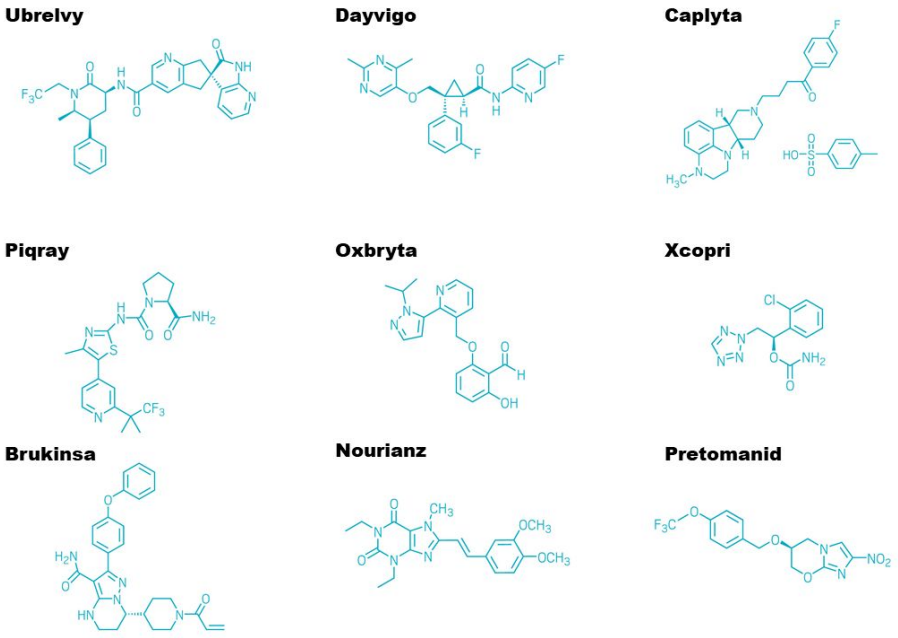

Figure 10.1 Examples of FDA approved small drug molecules in 2019. Sources: https://cen.acs.org/sections/drugs-approved-in-2019.html

10.2 Biocompatible Polymers and Their Interactions with Drug Molecules

Biocompatible and biodegradable polymers have emerged in the past decades as promising materials for drug loading. These biocompatible polymers can be classified as natural polymers, such as arginine, chitosan, dextrin, and polysaccharides, and synthetic polymers, for example, polycaprolactone, poly (N-isopropylacrylamide), and their copolymers. Some can be classified as either natural or synthesized polymers, such as poly(glycolic acid) and poly(lactic acid). These biocompatible polymers are good candidates as drug carriers to improve the solubility of drugs or maintain drug supersaturation.[8] Both experimental and computational study have been carried out to investigate their potential applications of drug delivery.[8,9] Jha and

Larson studied the compatibility of cellulosic polymers (hydroxypropyl-methylcellulose, HPMC and hydroxypropyl-methylcellulose acetate succinate, HPMCAS) with the drug phenytoin by using all-atom MD simulations. However, due to the diversity of biocompatible polymers, insightful relationships between biopolymers and their interaction strength with drug molecules need to be established.

In this project, I propose to use MD simulations to study the interactions between different polymers and drug molecules. MD simulations could unveil the process of how drug molecules approach the polymers, and provide atomistic details of the conformations of polymers when drug molecules bind to. Based on this study, we could find the polymers which favors the interactions with drug molecules. Furthermore, Monte Carlo simulations would be employed to study the adsorption of drug molecules on these polymers. These computational studies would provide directions for experimentalists when they synthesize the polymers for drug delivery systems.

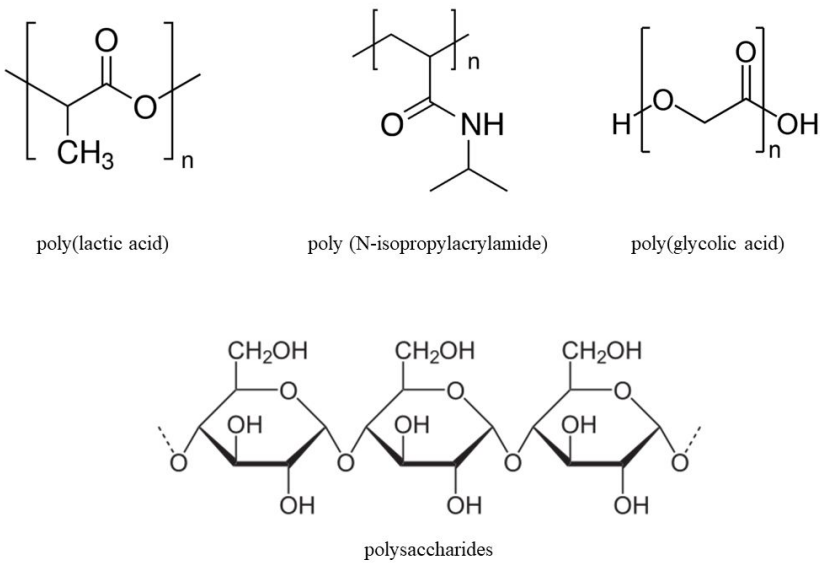

Figure 10.2 Examples of biocompatible polymers.

References

(1) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421.

(2) Yuriev, E.; Ramsland, P. A. Latest Developments in Molecular Docking: 2010-2011 in Review. *J. Mol. Recognit.* **2013**, *26* (5), 215–239.

(3) Souza, P. C. T.; Thallmair, S.; Conflitti, P.; Ramírez-Palacios, C.; Alessandri, R.; Raniolo, S.; Limongelli, V.; Marrink, S. J. Protein-Ligand Binding with the Coarse-Grained Martini Model. *Nat. Commun.* **2020**, *11* (1), 3714.

(4) Hoffmann, C.; Centi, A.; Menichetti, R.; Bereau, T. Molecular Dynamics Trajectories for 630 Coarse-Grained Drug-Membrane Permeations. *Sci Data* **2020**, *7* (1), 51.

(5) Conway, O.; An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Coarse-Grained Models of Amino Acids. *Molecular Systems Design & Engineering* **2020**. https://doi.org/10.1039/C9ME00173E.

(6) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, 4667–4672.

(7) Periole, X.; Marrink, S.-J. The Martini Coarse-Grained Force Field. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press: Totowa, NJ, 2013; pp 533–565.

(8) Katiyar, R. S.; Jha, P. K. Molecular Simulations in Drug Delivery: Opportunities and Challenges. *WIREs Comput Mol Sci* **2018**, *8* (4), e1358.

(9) Chen, W.; Zhou, S.; Ge, L.; Wu, W.; Jiang, X. Translatable High Drug Loading Drug Delivery Systems Based on Biocompatible Polymer Nanocarriers. *Biomacromolecules* **2018**, *19* (6), 1732–1745.

APPENDIX A

Table A1 Four types of CG beads and the groups represented by them.

Bead name	Represented groups	Mass (g/mol)
C2E	CH_3-CH_2-	29.062
C2M	-CH_2-CH_2-	28.054
C3E	CH_3-CH_2-CH_2-	43.089
C3M	-CH_2-CH_2-CH_2-	42.081

Table A2 Range of the input parameters of the CG parameters shown as r_{min} and r_{max}. *

CG model	Parameter	r_{min}	r_{max}	Optimized value
Decane (C2E-C2M-C2M-C2M-C2E)	K_b[C2E-C2M]	30.0	42.0	35.293
	b_0[C2E-C2M]	1.0	3.5	2.561
	K_b[C2M-C2M]	30.0	42.0	36.609
	b_0[C2M-C2M]	1.0	3.5	2.520
	K_θ[C2E-C2M-C2M]	3.2	4.1	3.886
	θ_0[C2E-C2M-C2M]	140	160	147.766
	K_θ[C2M-C2M-C2M]	3.2	4.1	3.460
	θ_0[C2M-C2M-C2M]	140	160	144.226
	ε[C2E]	0.35	0.375	0.3710
	ε[C2M]	0.33	0.35	0.3420
	σ[C2E]	2.67	5.35	4.3374
	σ[C2M]	2.67	5.35	4.3374
	K_b[C2M-C3M]	32	42.0	38.537
	b_0[C2M-C3M]	2.5	4.00	2.914

Nonane (C2E-C2M-C3M-C2E)	K_b[C2E-C3M]	32.0	42.0	34.170
	b_0[C2E-C3M]	2.5	4.0	2.958
	K_θ[C2E-C2M-C3M]	3.000	4.100	3.262
	θ_0[C2E-C2M-C3M]	135.0	165.0	144.669
	K_θ[C2E-C3M-C2M]	3.0	4.1	3.377
	θ_0[C2E-C3M-C2M]	135.0	165.0	149.893
	ε[C3M]	0.20	1.50	0.5545
	σ[C3M]	3.565	6.2388	4.6344
Nonane(C3E-C3M-C3E)	K_b[C3E-C3M]	2.0	42.0	37.695
	b_0[C3E-C3M]	2.5	4.0	3.561
	K_θ[C3E-C3M-C3E]	3.0	4.1	3.417
	θ_0[C3M-C3E]	135.0	165.0	144.832
	ε[C3E]	0.20	1.0	0.5927
	σ[C3E]	3.5650	6.2388	4.6344

*Units: K_b - (kcal/ mol/Å^2), b_0 - Å, K_θ - kcal/ mol/ radian^2, θ_0 - °, ε - kcal/mol, σ - Å.

Effects of system size on the properties of the CG decane and nonane models

To evaluate the ability of the new CG models of decane and nonane in determining the properties of the systems with different number of beads, we have performed MD simulations with 200, 1000, and 5000 beads with timestep of 15 fs at 300 K. The results are shown in **Table A3**. The densities predicted for decane with all three system sizes were within 0.5 % of the experimental value. In the case of nonane, for the hybrid model (2-2-3-2) the densities predicted by the model were within 0.3 % of the experimental value. Similarly, for the nonane model with 3-3-3 mapping scheme the densities were 0.713 g/cm^3, 0.709 g/cm^3, and 0.712 g/cm^3 for systems with 200, 1000, and 5000 beads, respectively. These values are within 0.8 % of the experimental value of the density of nonane. The enthalpy of vaporization and surface tension of all the three CG models fluctuate around their corresponding experimental data (within 18 % of the

experimental value). The self-diffusion coefficient of the CG decane and hybrid nonane (2-2-3-2) models increased with the increase in the system size. Similar behavior was also observed in the self-diffusion coefficient of water and polymers, which was attributed to the dependence of long-range hydrodynamics interaction in fluids with system size.[1–3] Thus, we find that changing the system size did not lead to significant variation in the properties of the new CG decane (2-2-2-2-2), hybrid nonane (2-2-3-2), and nonane (3-3-3) models.

Table A3 Properties of decane and nonane at 300 k with different system sizes. *

CG model	System size	ρ	H_V	γ	D
Decane (2-2-2-2-2)	200	0.723±0.00	12.31±0.02	23.88±0.10	1.47±0.05
	1000	0.723±0.00	12.30±0.01	24.00±0.06	1.52±0.02
	5000	0.723±0.00	12.30±0.03	24.01±0.04	1.54±0.01
Nonane (2-2-3-2)	200	0.712±0.00	10.70±0.02	22.81±0.05	1.75±0.08
	1000	0.712±0.00	10.58±0.28	22.45±0.23	1.78±0.07
	5000	0.712±0.00	10.70±0.02	22.40±0.02	1.79±0.01
Nonane (3-3-3)	200	0.713±0.00	9.86±0.02	23.83±0.90	1.60±0.06
	1000	0.709±0.00	10.25±0.28	22.40±0.02	1.46±0.04
	5000	0.712±0.00	9.78±0.01	22.54±0.21	1.59±0.02

*Units: ρ - g/cm^3, H_V - kcal/mol, γ - mN/m, D - × 10^{-9} m^2/s.

Table A4 Bonded force-field parameters for CG hydrocarbon models. *

Bond parameters	K_b	b_0
C2E -C2M	35.293	2.561
C2M -C2M	36.609	2.520
C3M -C2E	34.170	2.958
C3M -C2M	38.537	2.914
C3E -C3M	37.695	3.561

C3E -C2M [a]	34.170	2.958
C3E -C3E [b]	37.695	3.561
C3M -C3M [c]	37.695	3.561
C3E -C2E [d]	34.170	2.958
Angle parameters	K_θ	θ_0
C2E -C2M -C2M	3.886	147.766
C2M -C2M -C2M	3.460	144.226
C2E -C2M -C2E	3.673	145.996
C2E -C2M -C3M	3.2615	144.669
C2E -C3M -C2M	3.3774	149.893
C3E -C3M -C3E	3.4168	144.832
C3E -C3M -C2M [e]	3.3971	147.362
C2M -C2M -C3M [f]	3.2615	144.669
C3E -C3M -C3M [g]	3.4168	144.832
C3M -C3M -C3M [h]	3.4168	144.832
C3M -C3M -C2E [i]	3.4168	144.832
C3M -C3M -C2M [j]	3.3971	147.362
C3M -C2M -C3E [k]	3.2615	144.669

* Units: K_b - (kcal/ mol/Å^2), b_0 - Å, K_θ - kcal/ mol/ radian^2, θ_0 - °.

a: from C3M C2E

b: from C3E C3M

c: from C3E C3M

d: from C3M C2E

e: from C3E -C3M -C3E and C2E -C3M -C2M

f: from C2E -C2M -C3M

g: from C3E -C3M -C3E

h: from C3E -C3M -C3E

i: from C3E -C3M -C3E
j: from C3E -C3M -C3E and C2E -C3M -C2M
k: from C2M -C2M -C3M

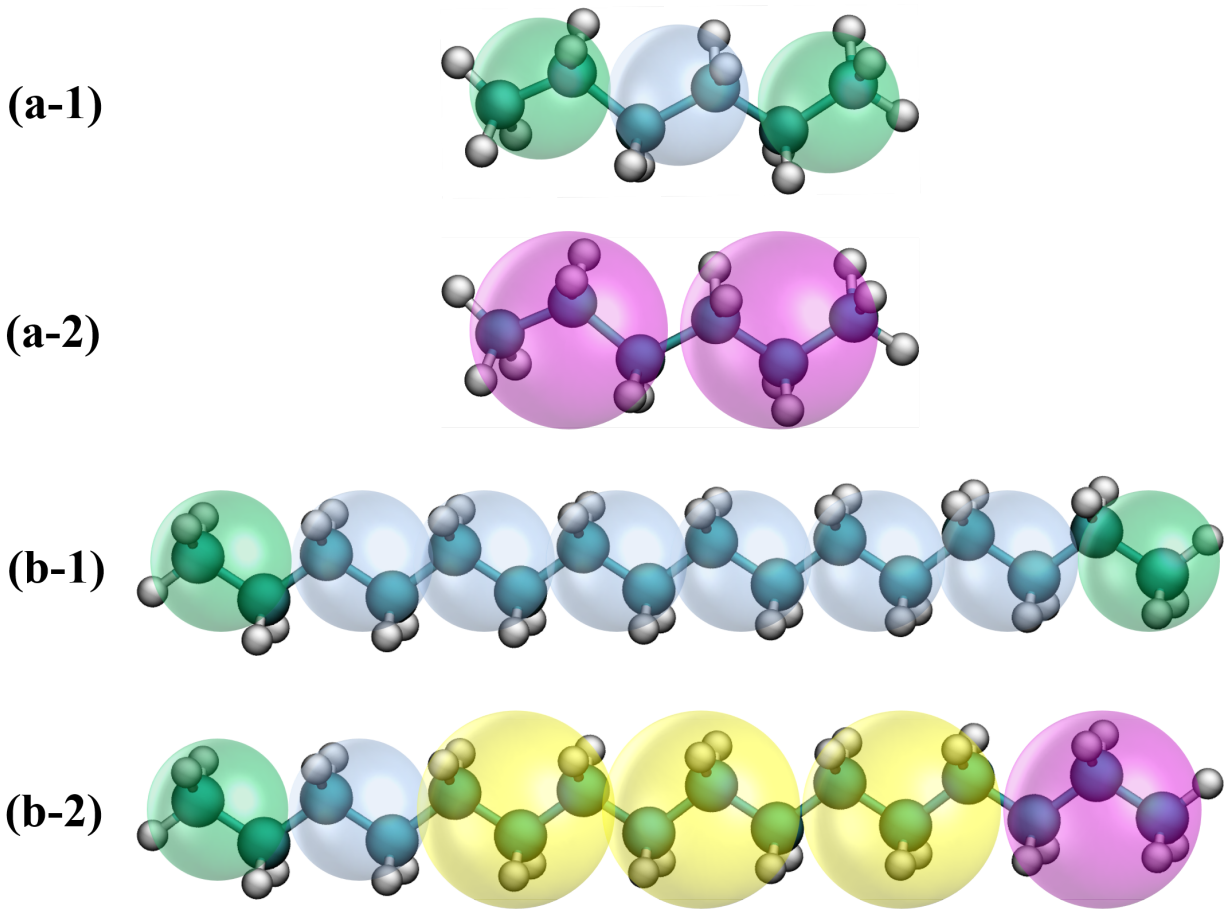

Figure A1 CG models of hexane and hexadecane: (a-1) - hexane with 2:1 mapping scheme, (a-2) - hexane with 3:1 mapping scheme, (b-1) - hexadecane with 2:1 mapping scheme. (b-2) - hexadecane with hybrid mapping schemes. Color scheme: C2E - green, C2M - gray, C3M - yellow, C3E - magenta.

Table A5 Comparison of dodecane properties by various CG models and UA models. ρ - g/cm^3, H_V - kcal/mol, γ - mN/m, D - $\times 10^{-9}$ m^2/s, α_T - $\times 10^{-3}$ K^{-1}, κ_T - $\times 10^{-5}$ bar^{-1}, ΔG - kcal/mol.

	Mapping schemes	$\boldsymbol{\rho}$	$\boldsymbol{H_v}$	$\boldsymbol{\gamma}$	$\boldsymbol{D}$	$\boldsymbol{\kappa_T}$	$\boldsymbol{\alpha_T}$	$\boldsymbol{\Delta G}$
New CG models of dodecane developed in this study	2-2-2-2-2-2	0.746	14.73	25.79	1.16	9.6	0.87	3.34
	3-3-3-3	0.735	13.49	24.89	0.90	10.9	0.82	3.74
MARTINI[4]	4-4-4	0.80	-	-	0.30	12	-	-
CRW [CG_{per}(sh)][5]	2-2-2-2-2-2	0.809	-	-	-	-	-	-
Eichenberger *et al.* [6]	2-2-2-2-2-2	0.750	14.61	25.8	1.97	10.6	0.8	4.44
	3-3-3-3	0.750	14.71	32.6	1.32	10.8	0.8	3.54
45A3/COS[7]	UA	0.744	14.78	25.0	1.3	8.1	0.9	-
Experiment	-	0.745	14.66	24.92	0.81	9.9	0.9	3.59

Table A6 Comparison of hexane properties by various CG models and UA models. ρ - g/cm^3, H_V - kcal/mol, γ - mN/m, D - $\times 10^{-9}$ m^2/s, α_T - $\times 10^{-3}$ K^{-1}, κ_T - $\times 10^{-5}$ bar^{-1}, ΔG - kcal/mol.

	Mapping schemes	$\boldsymbol{\rho}$	$\boldsymbol{H_v}$	$\boldsymbol{\gamma}$	$\boldsymbol{D}$	$\boldsymbol{\kappa_T}$	$\boldsymbol{\alpha_T}$	$\boldsymbol{\Delta G}$
New CG models of hexane developed in this study	2-2-2	0.634	6.99	16.45	3.40	19.9	1.42	3.16
	3-3	0.658	6.84	17.36	2.79	19.4	1.4	2.50
MARTINI [4]	3-3	0.58	-	-	0.7	14	-	-
CRW [CG_{per}(sh)][5]	2-2-2	0.659	-	16.4	-	22	1.1	-
Eichenberger *et al.*[6]	2-2-2	0.666	7.62	21.7	4.73	15.6	1.2	2.58
	3-3	0.660	7.54	25.5	3.76	16.5	1.1	2.48

45A3/COS[7]	UA	0.656	7.58	17.9	4.9	13.8	1.3	2.54
Experiment	-	0.660	7.55	17.91	4.21	16.7	1.4	2.51

APPENDIX B

Table B1 Nonbonded force field (FF) parameters of previously developed CG hydrocarbon beads and water beads.[3,8]

	represented groups	ε (kcal/mol)	σ (Å)
C2E	CH_3-CH_2-	0.3710	4.3374
C2M	-CH_2-CH_2-	0.3420	4.3374
C3E	CH_3-CH_2-CH_2-	0.5927	4.6344
C3M	-CH_2-CH_2-CH_2-	0.5545	4.6344
W1	$(H_2O)_2$	1.1425	3.772

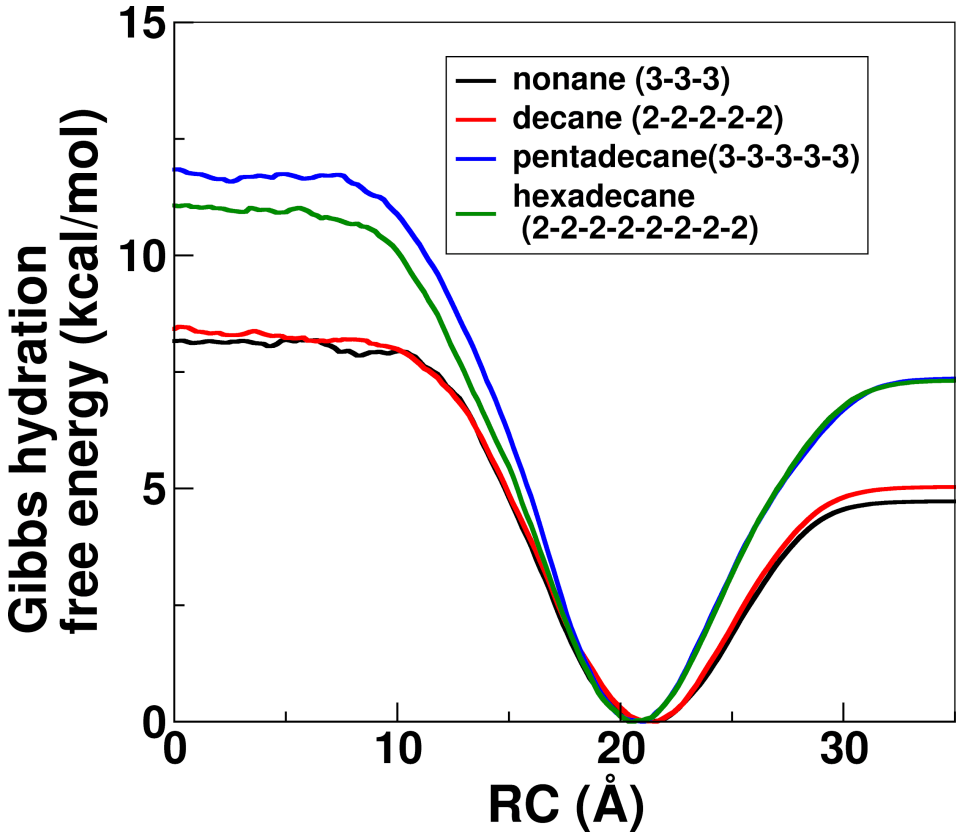

Figure B1 Gibbs hydration free energy profiles of nonane (3-3-3), decane (2-2-2-2-2), pentadecane (3-3-3-3-3) and hexadecane (2-2-2-2-2-2-2-2) models at 300 K and 1 bar.

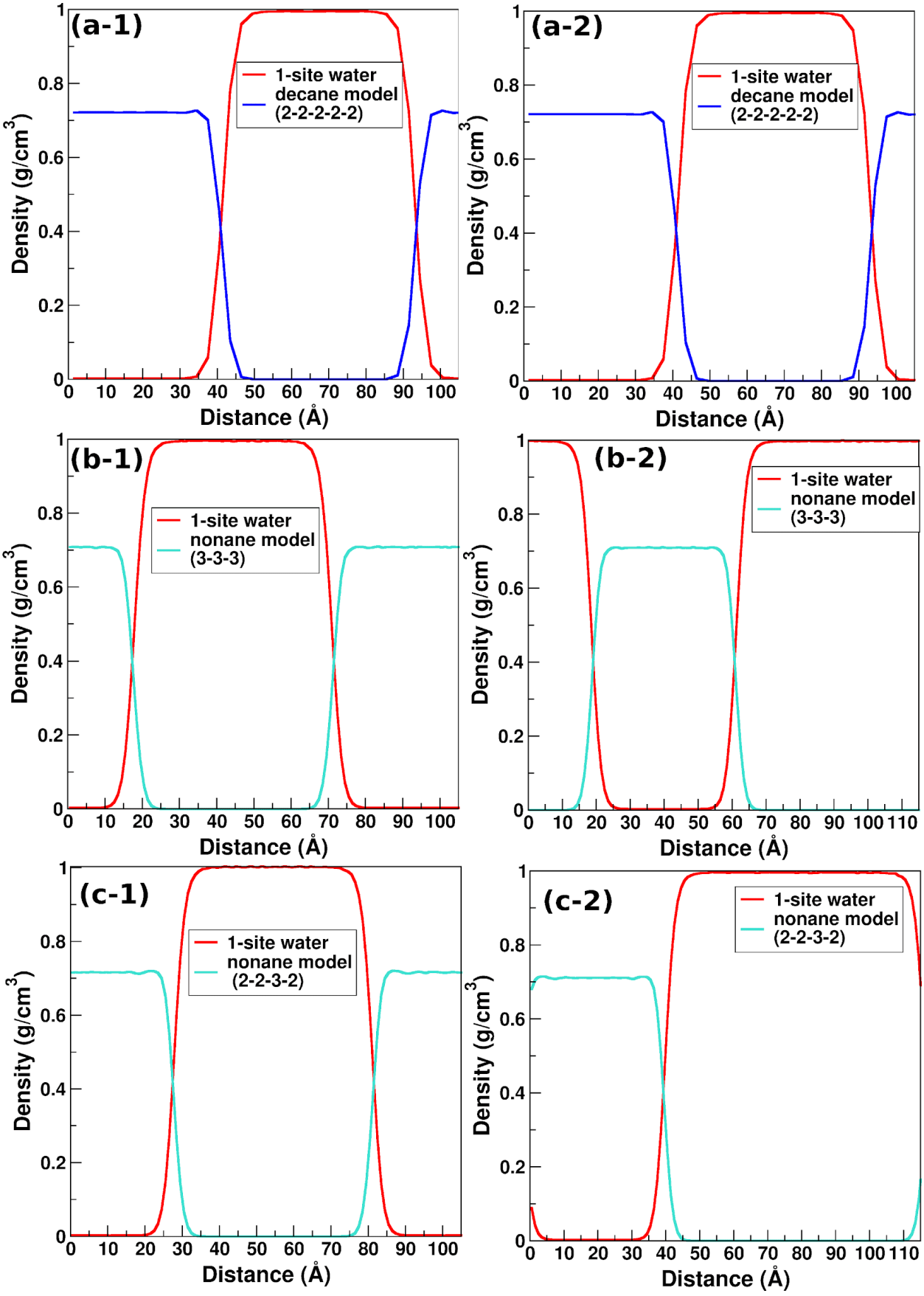
(a-1)
1-site water
decane model
(2-2-2-2-2)
Density (g/cm3)
Distance (Å)
(a-2)
1-site water
decane model
(2-2-2-2-2)
Density (g/cm3)
Distance (Å)
(b-1)
1-site water
nonane model
(3-3-3)
Density (g/cm3)
Distance (Å)
(b-2)
1-site water
nonane model
(3-3-3)
Density (g/cm3)
Distance (Å)
(c-1)
1-site water
nonane model
(2-2-3-2)
Density (g/cm3)
Distance (Å)
(c-2)
1-site water
nonane model
(2-2-3-2)
Density (g/cm3)
Distance (Å)

Figure B2 Density profiles of **(a)** decane(2-2-2-2-2)/water, **(b)** nonane(3-3-3)/water, and **(c)** nonane(2-2-3-2)/water mixtures with different number of water molecules: **(1)** 10000 CG water molecules, **(2)** 18000 CG water molecules, when the number of CG hydrocarbon molecules is kept at 2000. Simulations were performed at 300 K and 1 bar.

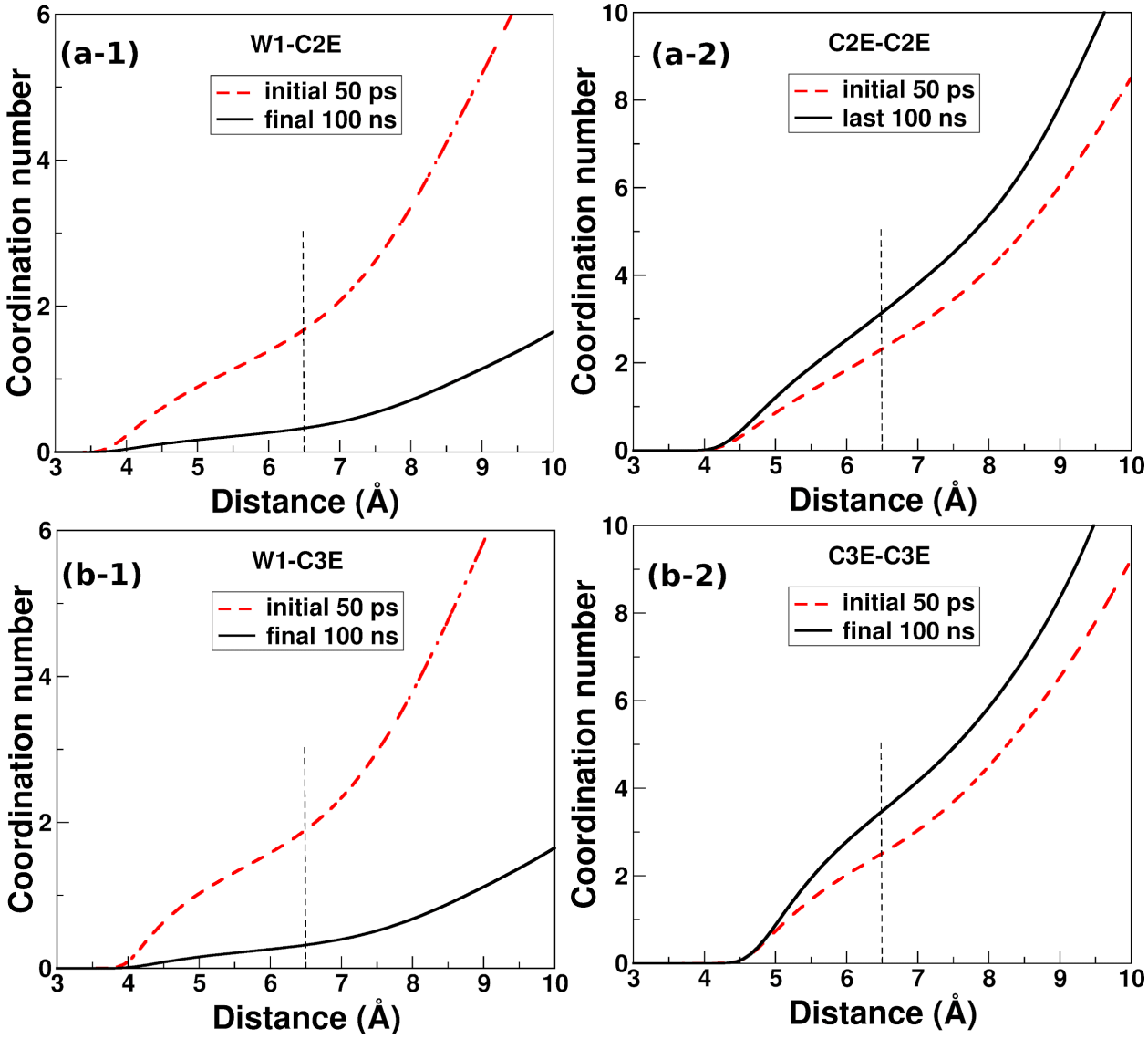

Figure B3 Coordination number of the W1 beads around the C2E beads in the decane (2-2-2-2-2) model **(a-1)**, the C2E beads around other C2E beads in the decane (2-2-2-2-2) model **(a-2)**, the W1 beads around the C2E bead in the nonane (3-3-3) model **(b-1)**, the C3E beads around other C3E beads in the nonane (3-3-3) model **(b-2)**.

Table B2 The interfacial tensions of four hydrocarbon/water systems at different temperatures. Experimental values in parentheses are from reference [9].

	Interfacial tensions (mN/m)			
	300 K	**320 K**	**330 K**	**350 K**
pentane(2-3)/water	39.2±0.2	34.5±0.3	31.6±0.5	27.9±0.3
nonane(3-3-3)/water	46.2±0.3 (51.2)	41.9±0.2 (49.7)	39.3±0.4 (48.8)	37.3±0.2
decane(2-2-2-2-2)/water	48.0±0.5 (51.0)	47.0±0.2 (50.0)	43.2±0.6 (49.7)	39.3±0.2
heptadecane(2-2-2-2-3-3-3)/ water	51.2±0.4	48.5±0.3	46.1±0.5	44.1±0.3

Effects of mapping schemes on the properties of hydrocarbon models

We have investigated the effects of mapping schemes on the properties of hydrocarbon models, where three mapping schemes were selected for octane and decane, four mapping schemes for heptadecane, five for pentadecane and six for dodecane. Specifically, we generated the structures of hydrocarbons by varying the positions and the number of beads with 3:1 (3 heavy atoms in a bead) and 2:1 (2 heavy atoms in a bead) mapping schemes to understand the effect of the position and the number of these beads on the properties of pure hydrocarbons and hydrocarbon/water systems (see **Table B4**). In general, we find that for hydrocarbons with more number of beads with 3:1 mapping scheme, the self-diffusion coefficient is smaller. For example, the self-diffusion coefficients of dodecane models with 2-2-2-2-2-2, 3-2-2-2-3 and 3-3-3-3 mapping schemes are 1.16×10^{-9} , 0.91×10^{-9} and 0.90×10^{-9} m^2/s, respectively (experimental value: 0.81 m^2/s). Similarly, the pentadecane models with 2-2-2-2-2-2-3, 2-2-2-3-3-3, 3-2-2-2-3-3, 3-3-3-3-3 mapping schemes have self-diffusion coefficients of 0.70×10^{-9}, 0.70×10^{-9}, 0.64×10^{-9}, 0.61×10^{-9} m^2/s, respectively, showing a general decreasing trend. This can be attributed to the large ε values of the beads with 3:1 mapping schemes compared with the ones with 2:1 mapping schemes. The large ε values correspond to the strong interaction between beads, thus resulting in low self-diffusion coefficient. It is also noticeable that for hydrocarbon models with hybrid mapping scheme and two C3E beads, the density and surface tension are

usually overestimated. For example, the density and surface tension of octane (3-2-3) model were 5 % and 15 % higher than the experimental data. This is also true when we compare the density and surface tensions of the decane (3-2-2-3), pentadecane (3-2-2-2-3-3) and heptadecane (3-2-2-2-2-3-3) models with their corresponding experimental values.

The percentage differences in density, self-diffusion coefficient, surface tension, interfacial tension, and Gibbs hydration free energy between different octane model and between different dodecane models are shown in **Table B5**. We found that the density and interfacial tension of the octane models (3-2-3) and (2-3-3) showed less than 7 % of difference when compared with those of the octane model (2-2-2-2). Similarly, the differences in the density and interfacial tension of dodecane models (3-2-2-2-3, 3-3-2-2-2, 2-3-3-2-2, 2-3-2-3-2 and 3-3-3-3) were less than 8 % when compared to the dodecane model with 2:1 mapping scheme i.e. 2-2-2-2-2-2. On the other hand, the difference in Gibbs hydration free energy and in self-diffusion coefficient was more prominent. For example, the Gibbs hydration free energies of the dodecane models (3-3-2-2-2 and 2-3-3-2-2) were ~43.3 % more than that of the dodecane (2-2-2-2-2-2) model.

To understand if symmetry has an effect on the properties of hydrocarbons, we performed simulations for the symmetric dodecane model (2-3-2-3-2) and the heptadecane model (3-2-2-3-2-2-3). The density of the symmetric dodecane model is 0.753 g/cm^3, only 2 % larger than that of the asymmetric dodecane model (2-3-3-2-2). The percentage differences in the self-diffusion coefficient, surface tension between the symmetric and asymmetric dodecane models were 2.6 %, and 1.6 %, respectively. Similarly, the symmetric heptadecane model (3-2-2-3-2-2-3) showed a slight difference in predicting the density, self-diffusion coefficient, and surface tension (percentage difference < 5 %) compared with the asymmetric heptadecane model (3-2-2-2-2-3-3). This indicates that the symmetry might not play an important role in predicting the bulk properties of the CG hydrocarbon models

Table B3 Effects of mapping schemes on the bulk properties of CG hydrocarbons.

	mapping scheme	**density**	**self-diffusion**	**surface tension**
octane	2-2-2-2	0.689±0.0	2.13±0.03	20.7±0.1

	3-2-3	0.736±0.0	1.62±0.05	24.2±0.1
	2-3-3	0.692±0.0	1.93±0.04	20.2±0.2
experiment (octane)		0.699	2.0	21.1
decane	2-2-2-2-2	0.723±0.0	1.52±0.02	24.0±0.1
	3-2-2-3	0.755±0.0	1.2±0.02	25.0±0.1
	2-2-3-3	0.727±0.0	1.34±0.03	23.7±0.1
experiment (decane)		0.726	1.55	23.4
dodecane	2-2-2-2-2-2	0.746±0.0	1.16±0.01	25.8±0.1
	3-2-2-2-3	0.774±0.0	0.91±0.02	28.3±0.1
	3-3-2-2-2	0.750±0.0	0.97±0.01	26.0±0.2
	2-3-3-2-2	0.737±0.0	1.13±0.01	25.0±0.1
	2-3-2-3-2	0.753±0.0	1.1±0.02	25.4±0.2
	3-3-3-3	0.735±0.0	0.90±0.02	24.9±0.1
experiment (dodecane)		0.745	0.81	24.9
pentadecane	2-2-2-2-2-2-3	0.781±0.0	0.70±0.01	29.3±0.1
	3-2-2-2-3-3	0.783±0.0	0.64±0.01	29.6±0.2
	2-2-2-3-3-3	0.762±0.0	0.70±0.01	27.3±0.1
	2-2-3-3-3-2	0.752±0.0	0.729±0.02	26.1±0.1
	3-3-3-3-3	0.751±0.0	0.61±0.01	26.3±0.1
experiment (pentadecane)		0.765	0.4	27.1
heptadecane	2-2-2-2-3-3-3	0.774±0.0	0.58±0.02	28.3±0.2
	3-2-2-2-2-3-3	0.793±0.0	0.53±0.01	30.5±0.2
	3-2-2-3-2-2-3	0.804±0.0	0.51±0.01	31.7±0.3

	3-2-3-3-3-3	0.775±0.0	0.5±0.03	29.0±0.2
experiment (heptadecane)		0.774	0.36	27.9

Units: density - g/cm^3, self-diffusion coefficient - $\times 10^{-9}$ m^2/s, surface tension - mN/m, interfacial tension - mN/m, Gibbs hydration free energy - kcal/mol. Experimental data is from references [9–13].

Table B4 The percentage differences in density (ρ), self-diffusion coefficient (D), surface tension (ST), interfacial tension (IFT), and Gibbs hydration free energy (ΔG_{hyd}) between octane models (3-2-3, 2-3-3) and (2-2-2-2), and between dodecane models (3-2-2-2-3, 3-3-2-2-2, 2-3-3-2-2, 2-3-2-3-2, 3-3-3-3) and (2-2-2-2-2-2).

	difference % when compared with octane (2-2-2-2) model				
	ρ	**D**	**ST**	**IFT**	**ΔG_{hyd}**
octane (3-2-3)	6.8	-23.9	16.9	-5.4	-32
octane (2-3-3)	0.4	-9.4	-2.4	-7.0	-8
	difference % when compared with dodecane (2-2-2-2-2-2) model				
	ρ	**D**	**ST**	**IFT**	**ΔG_{hyd}**
dodecane (3-2-2-2-3)	3.7	-21.6	9.7	-5.7	-6.7
dodecane (3-3-2-2-2)	0.5	-16.4	0.8	-4.3	43.3
dodecane (2-3-3-2-2)	-1.2	-2.6	-3.1	-7.6	43.3
dodecane (2-3-2-3-2)	0.9	-5.2	-1.5	-3.7	33.3
dodecane (3-3-3-3)	-1.5	-22.4	-3.5	-7.8	16.7

APPENDIX C

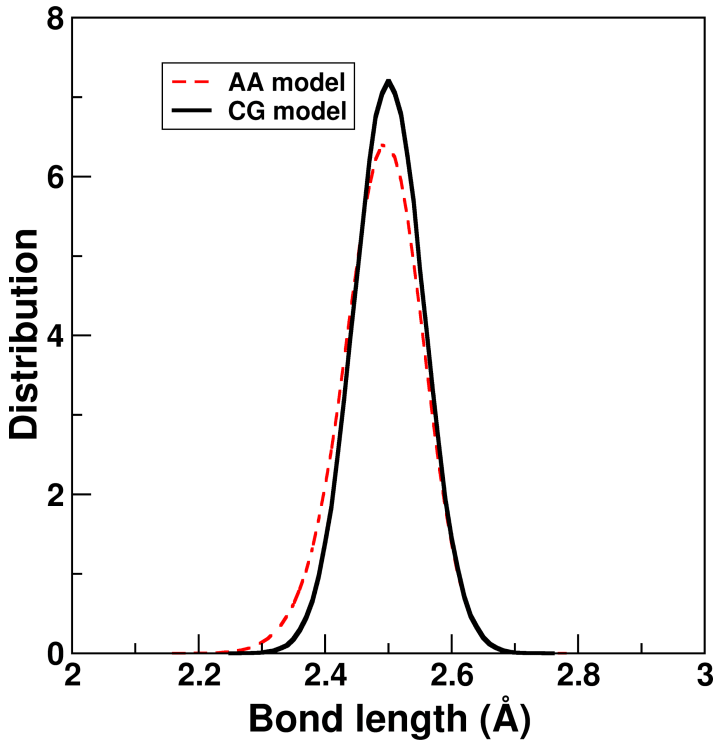

Figure C1 The C2E-COOH bond length distributions for the AA mapped and the CG propionic acid models.

Development of the CG Propionic Acid Model by Particle Swarm Optimization

Table C1 The range of CG FF parameters (r_{min} - r_{max}) and their optimized values. Unit: K_b - (kcal/mol/Å^2), ε - kcal/mol, σ - Å, b_0 - Å.

FF parameter	r_{min}	r_{max}	Optimized value
K_b[C2E-COOH]	70.0	100.0	90.98
ε[COOH]	0.30	1.60	1.2516
σ[COOH]	2.67	5.35	3.9356
b_0[C2E-COOH][a]	-	-	2.5
ε[C2E][b]	-	-	0.3710
σ[C2E][b]	-	-	4.3374

a: estimated from the mapped COOH-C2E bond distribution in **Figure S1**.

b: they were adopted from ref [8].

Structure the CG Propionic Acid Model

The radial distribution functions (RDFs) of the CG model were compared with those obtained from the AA mapped trajectories. In **Figure C2 - (a)**, the position of the first peak for the CG model was at ~4.6 Å, close to that for the AA mapped propionic acid model, ~4.9 Å. The first peak in the RDF of the C2E-COOH pair in **Figure C2 - (b)** was located at 4.5 Å, slightly left shifted compared with that of the AA mapped trajectory. A small shoulder at ~3.0 Å in the RDF of COOH-COOH obtained from the AA mapped trajectory can be observed in **Figure C2 - (c)**. This shoulder resulted from the formation of hydrogen bonds in AA propionic acid molecules, which were formed between the oxygen in the hydroxyl group and the oxygen attached with carbon atoms by double bonds as the acceptor (see **Figure C3** - (**a**)). When these hydrogen bonds were formed in AA propionic molecules, the distance between two mapped COOH bead was measured to be 3.12 Å. Besides the shoulder at ~3.0 Å, the peak at ~4.1 Å was also caused by hydrogen bonds in AA propionic acid molecules, which is, however, a different type of hydrogen bond. This type of hydrogen bond was formed between the hydroxyl groups of two AA propionic acid molecules (see **Figure C3** - (**b**) of Supporting Information). With the configuration, the distance between the mapped COOH beads increased to 4.12 Å. However, no shoulder was observed in the RDF of COOH-COOH pairs in the CG propionic acid molecules. The first peak was at 4.4 Å, 0.3 Å larger than the position of the first peak in the RDF(COOH-COOH) from AA mapped trajectories. The intensity of the first peaks in all the RDFs of bead pairs in the CG model is stronger than those of the AA mapped bead pairs, especially for the RDF between the COOH-COOH bead pair. This indicates that the interactions between the CG propionic acids molecules are strong compared with those in AA molecules.

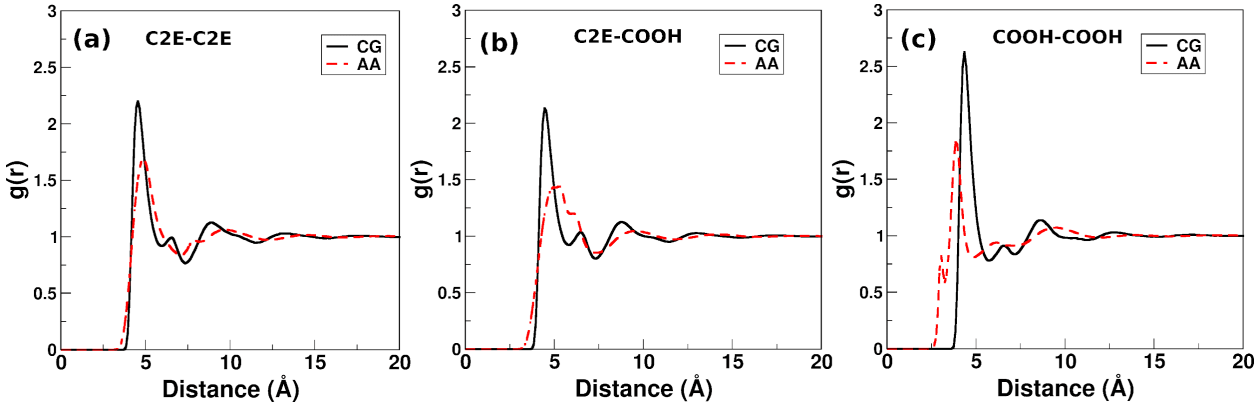

Figure C2 The RDFs between **(a)** C2E-C2E beads, **(b)** C2E-COOH beads, and **(c)** COOH-COOH beads in the CG propionic acid model (black solid lines) compared with those from the AA mapped trajectories (red dashed lines).

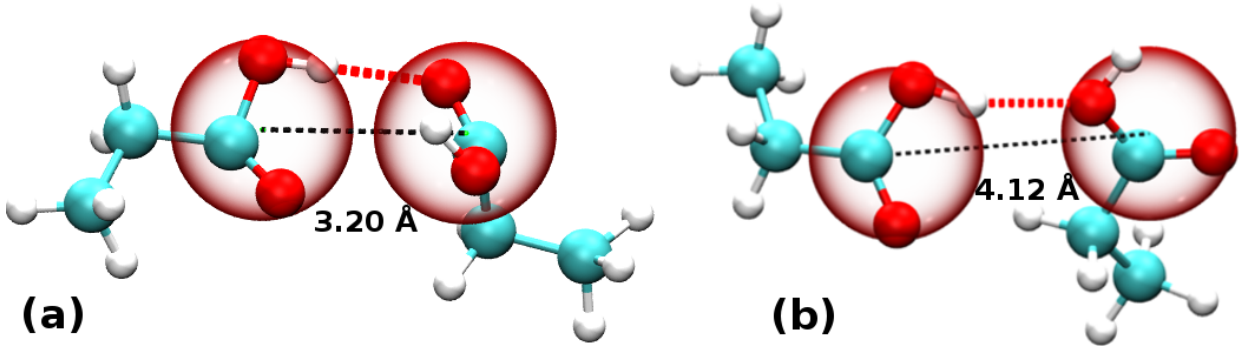

Figure C3 Two types of hydrogen bonds (red dashed lines) which resulted in the different distances (black dashed lines) between mapped COOH beads (red transparent spheres): **(a)** 3.20 Å, **(b)** 4.12 Å. Oxygen atoms: red filled spheres, hydrogen atoms: white filled spheres, carbon atoms: cyran filled spheres. The criteria for identifying hydrogen bonds is that the donor-acceptor distance is less than 3.5 Å and the acceptor-donor-hydrogen angle is ≤30°.[14]

Table C2 The bond FF parameters in the CG PAA model and in the CG DMF model.

bond stretching	**K_b** (kcal/(mol · Å^2))	**b_0** (Å)
C2E-C2M[a]	35.29	2.56
C2M-C2M[a]	36.61	2.52
C2E-COOH	90.98	2.50

C2M-COOH	90.98	2.50
AM-CGD2[b]	175.18	2.0

a: adopted from ref [8].

b: adopted from ref [15].

Table C3 The angle FF parameters in the CG PAA model

bond-angle bending	**K_a** (kcal/(mol · rad^2))	**a_0** (^{o})
C2E-C2M-C2M[a]	3.89	147.77
C2M-C2M-C2M[a]	3.46	144.23
C2M-C2E-COOH	5.0	100.0
C2M-C2M-COOH	5.0	100.0
C2E-C2M-COOH	5.0	100.0

a: adopted from ref [8].

Table C4 Nonbonded parameters for the CG PAA model and between PAA and CG solvent models: 1-site water and CG DMF

models	**beads or bead pairs**	**ε (kcal/mol)**	**Method used to obtain ε values**	**σ (Å)**	**Method used to obtain σ values**
CG hydrocarbon models	C2E	0.3710	ref[a]	4.3374	ref[a]
	C2M	0.3420	ref[a]	4.3374	ref[a]
1-site water model	W1	1.1425	ref[b]	3.772	ref[b]
CG DMF model	AM	0.7147	ref[c]	3.8908	ref[c]
	CGD2	0.3797	ref[c]	4.5703	ref[c]
(1-site water)-DMF	W1-AM	0.9036	Berthelot rule	3.8314	Lorentz rule
	W1-CGD2	0.6586	Berthelot rule	4.1711	Lorentz rule

CG propionic acid model	COOH	1.2516	This study	3.9356	This study
PAA-(1-site water)	COOH-W1	1.35	This study	3.8538	Lorentz rule
	C2E-W1	0.5130	ref[d]	3.774	ref[d]
	C2M-W1	0.440	ref[d]	3.774	ref[d]
PAA-DMF	COOH-AM	1.35	This study	3.9133[e]	Lorentz rule
	COOH-CGD2	0.6894	Berthelot rule	4.2530	Lorentz rule
	C2M-AM	0.4944	Berthelot rule	4.1141	Lorentz rule
	C2E-AM	0.5149	Berthelot rule	4.1141	Lorentz rule
	C2M-CGD2	0.3604	Berthelot rule	4.4538	Lorentz rule
	C2E-CGD2	0.3753	Berthelot rule	4.4538	Lorentz rule

a: adopted from ref [8].
b: adopted from ref [3].
c: adopted from ref [15].
d: adopted from ref [16] .

Table C5 Properties of the 1-site water and CG DMF models in ref [3,15]. Experimental data are shown in parentheses at 300 K.

	ρ	**H_v**	**γ**	**D**
1-site water	1.002 (0.997)	7.8 (10.5)	68.2 (72.0)	2.49 (2.38)
CG DMF	0.943 (0.944)	-	33.3 (35.8)	1.76 (1.63)

Units: density, ρ - g/cm^3, enthalpy of vaporization, H_v - kcal/mol, surface tension, γ - mN/m, self-diffusion coefficient, D - $\times 10^{-9}$ m^2/s.

Determination of the Interaction Parameters between PAA and Solvents

The LJ potential parameters between PAA and solvents were tuned in a systematic way to reproduce the R_g distributions of the all-atom PAA model in water and DMF. Specifically, to describe the interactions between W1 beads of water and COOH beads of PAA, Berthelot combining rule was firstly used to obtain the ε[COOH-W1] of 1.1958 kcal/mol. The interactions between C2E/C2M bead of PAA and W1 beads were adopted from reference[16]. Then the CG MD simulations of a single PAA chain of 30-mer in water were performed at 300 K for 1 μs. The R_g distribution of the CG PAA chain was compared with that of the AA model. As can be seen in **Figure C4 (a)** the R_g distribution of the CG model showed a peak at ~ 7.1 Å as compared to AA model at ~ 8.8 Å. This suggested that the CG chain was in more globule-like state as compared to AA model, implying that the interactions between CG PAA and W1 bead are weak as compared to AA model. Then we manually, systematically increased the interactions between PAA and water by increasing ε[COOH-W1]. The ε[COOH-W1] of 1.35 kcal/mol, and 1.38 kcal/mol showed very good agreement between the R_g distribution of the CG PAA and that of the all-atom PAA model. However, as the Gibbs hydration free energy of propionic acid in water showed better agreement with 1.35 kcal/mol, we employed these parameters during this study.

To describe the interactions between CG PAA model and DMF, we reproduced the R_g distribution of the all-atom PAA model in DMF. The interactions between CGD2 and C2E/C2M beads, between CGD2 and COOH beads, and between AM and C2E/C2M beads were described by the LB combining rules. This is because the CGD2 bead and C2M/C2E bead are similar, which all represent two carbon atoms and their associated hydrogen atoms. The interaction between AM and COOH bead was developed and selected such that we get the best agreement in R_g distribution of CG and all-atom models. We initiated the development of ε[AM-COOH] by using Berthelot combining rule, which resulted in the initial value of 0.9458 kcal/mol. These results are shown in **Figure C4 (b)**. It can be seen that reasonable agreement with the R_g distribution of the all-atom model was obtained when ε[AM-COOH] value was 1.35 kcal/mol, 1.38 kcal/mol, and 1.4 kcal/mol. The Gibbs free energy values for the monomer analogue of PAA, propionic acid by using the ε[AM-COOH] value of 1.35, 1.38, and 1.4 kcal/mol were -9.4, -9.9, and -9.9 kcal/mol, respectively. As the Gibbs free energy for the all-atom propionic acid model in DMF is -7.4 kcal/mol, we decided to use 1.35 kcal/mol in this research.

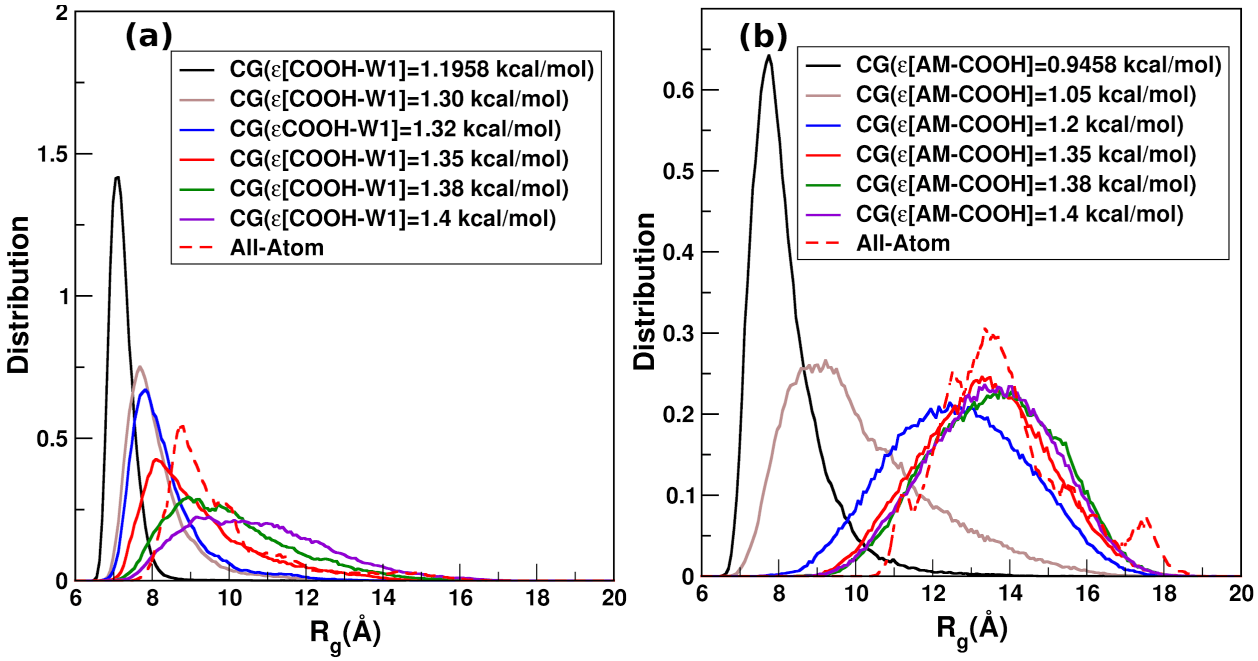

Figure C4 The R_g distributions of the CG PAA model in **(a)** water and **(b)** DMF with different ε values.

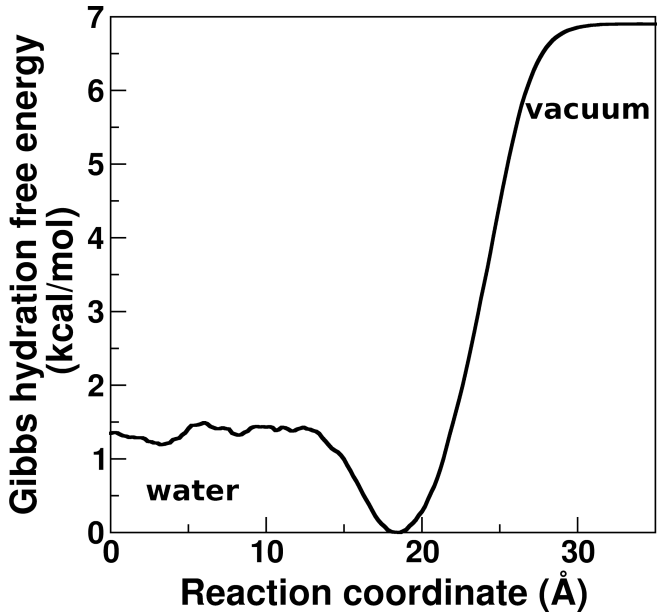

Figure C5 The Gibbs hydration free energy profile of one CG DMF molecule when pulled into 1-site water bulk.

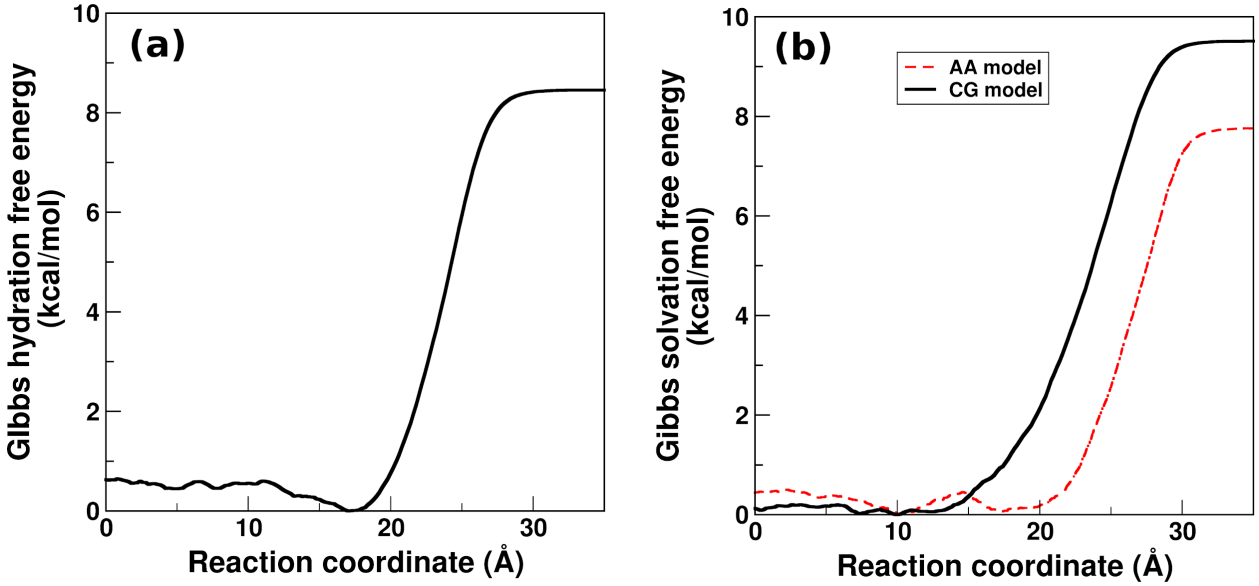

Figure C6 The profiles of **(a)** Gibbs hydration free energy of the CG propionic acid in water and **(b)** Gibbs solvation free energy of AA and CG propionic acid models in DMF at 300 K.

Table C6 The compositions of the solvent mixtures of CG 1-site water and DMF models.

mass fraction of DMF	number of CG DMF molecules	number of CG 1-site water
0	0	5000
0.1011	250	4507
0.2907	100	495
0.4429	250	638
0.6582	250	263
0.8240	500	216
0.9235	500	84
1.0	10000	0

Table C7 The number of all-atom water and DMF molecules in the mixtures of solvents for the study of PAA conformation transition. Note, the number of CG water molecules is half of that of the all-atom water molecules.

mass fraction of DMF (wt%)	number of all-atom DMF molecules	number of all-atom water molecules
0	0	10,000
2.6	100	15,000
16.8	500	10,000
31.1	1,500	13,500
50.3	2,400	9,600
80.2	2,500	2,500
100	5,000	0

Sampling Quality of All-Atom PAA and Pure Solvent Simulations:

Although MD simulations have emerged as an important tool to study macromolecule systems with millions of atoms, on many occasions, its application is limited due to insufficient sampling. Primary cause of this insufficient samplication is the rough energy landscapes, which has many local minima separated by high-energy barriers that controls the motion of macromolecules.[17]

There are a number of analysis methods such as root mean square deviation (RMSD), cluster counting, principal component analysis (PCA) that can provide qualitative insights on the overall sampling effectiveness.[18] In general, they can suggest if the simulation time is too short to sample the space. RMSD is one such simple tool which is often used to evaluate the sampling effectiveness. RMSD compares the initial structure of a macromolecule to that throughout the trajectory via a distance measure.[18] Here, we have computed the RMSD of PAA chain in pure water, and pure DMF using **Eq. C1**.

$$RMSD = \sqrt{\frac{\sum_{I=0}^{N} [m_i * (X_i - Y_i)^2]}{M}} \qquad \textbf{.....Eq. C1}$$

Where N is the total number of atoms, m_i is the mass of atom *i*, X_i is the coordinate vector for target atom *i*, Y_i is the coordinate vector for reference atom *i*, and M is the total mass.

In the initial 10 ns a rapid rise in the RMSD is observed due to thermal fluctuations (see F**igure C8**). This is followed by a fluctuations in the RMSD around its mean value of ~12.8 Å and ~ 8.9 Å in water and DMF, respectively. The overall fluctuations in the mean value for both water and DMF decrease after ~150 ns. The RMSD plots suggest that both systems have reached steady state and polymer conformation has equilibrated.

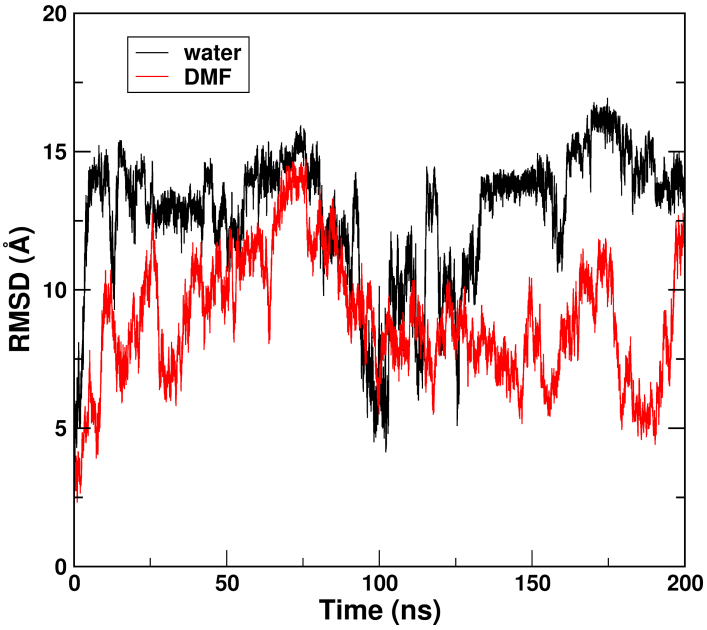

Figure C7 The RMSD of the all-atom PAA model in water and pure DMF.

To further extract the large scale characteristic motions of PAA chain, PCA was used.[19] The first step in PCA is the building of the 3n x 3n fluctuation correlation matrix (C), where n is the number of atoms in a system:[18]

$$c_{ij} = < x_i - \overline{x}_i >< x_j - \overline{x}_j > \quad \text{.....Eq. C1.2}$$

Where x_i represents a specific degree of freedom and overbar depict the average structure.

The cartesian coordinates of the heavy atoms in PAA (carbon and oxygen), stored every 100 ps, were utilized to perform the PCA analysis. The matrix was used to reproduce eigenvalues (the mean square fluctuations along its corresponding vector), and eigenvectors (characteristic motions for the system). PCA reveals the existence of a number of sub-states for a given macromolecule in a simulation trajectory. In general, a well-sampled simulation exhibits a large number of transitions among substates. The first two principal components for PAA in water and DMF are shown in **Figure C9 (a)** and **(b)**, respectively. In **Figure C9**, clustering of data can be observed for PAA both in water and DMF, which suggest that PAA exhibits several sub-states. For example, in water, although the R_g values of the PAA chains are similar, ~8 Å, clustering of data points in different regions suggests the presence of different conformations with similar values of R_g. In pure DMF the less number of sub-states were exhibited by the PAA chain as compared to water.

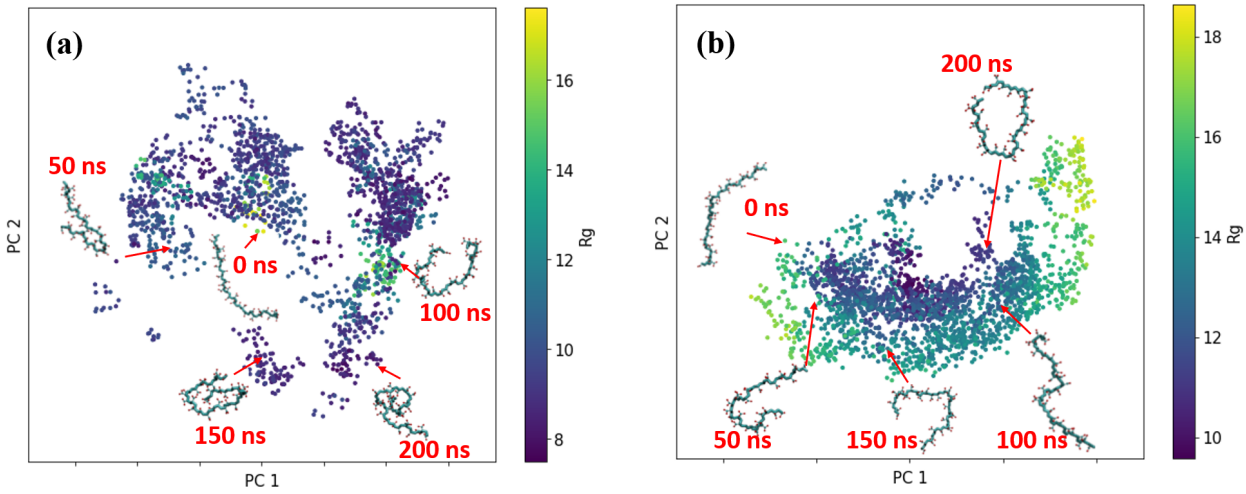

Figure C8 Projection of the transient structures of the all-atom PAA model on the principle plane spanned by the two most significant components (PC1 and PC2) obtained by analyzing the cartesian coordinates of the all-atom PAA model in **(a)** water and **(b)** DMF .

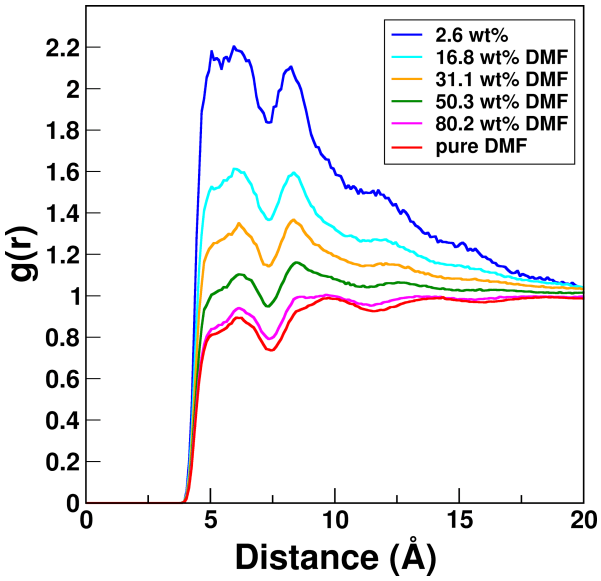

Figure C9 The RDF between CGD2 bead in DMF and C2M beads in the backbone of PAA

Table C8 The coordination number of CGD2 beads around the C2M beads at a cutoff of 7.25 Å (the position of the first valley in **Figure S10**).

solvent bulk	number of CGD2 beads
pure water	0
2.6 wt% DMF	0.57
16.8 wt% DMF	2.64
31.0 wt% DMF	4.0
50.3 wt% DMF	5.28
80.2 wt% DMF	6.94
pure DMF	8.17

APPENDIX D

Table D1 Bonded and nonbonded parameters for the CG ethylbenzene and PS model. Units: K_b: Kcal/(mol Å^2), b_0: Å, K_θ: Kcal/(mol rad^2), θ_0:^{o} , K_φ: Kcal/(mol rad^2), φ_0:^{o} , K_Ψ: Kcal/(mol rad^2), Ψ_0:^{o} , ε: Kcal/mol, σ: Å

Bond parameters	K_b	b_0
C2E-BZ	36.609	2.33
BZ-BZ	365.88	2.24
C2M-C2M[1]	36.609	2.52
C2E-C2M[1]	36.609	2.52
Angle parameters	K_θ	θ_0
C2E-BZ-BZ	25.0	150.0
BZ-BZ-BZ[2]	214.156	60.0
C2E-C2M-BZ	4.0	110
C2M-C2M-BZ	4.0	110
C2M-C2E-BZ	4.0	110
C2M-C2M-C2M[1]	3.46	144.226
C2E-C2M-C2M[1]	3.46	144.226
Dihedral angle parameters	K_φ	φ_0
C2M-C2M-C2M-C2M	1.0	0
C2M-C2M-C2M-C2E	1.0	0
C2M-C2M-C2M-BZ	1.0	0
C2M-C2M-C2E-BZ	1.0	180
BZ-BZ-C2M-C2M	1.0	180
BZ-C2M-C2M-BZ	1.0	0
BZ-C2E-C2M-BZ	1.0	0

BZ-BZ-C2E-C2M	1.0	0
BZ-BZ-C2M-C2E	1.0	0
BZ-BZ-BZ-C2E	0	0
BZ-BZ-BZ-C2M	0	0
Improper angle parameters	K_Ψ	Ψ_0
BZ-C2M-BZ-BZ	10	0
LJ parameters	ε	σ
C2E[1]	0.3710	4.3374
C2M[1]	0.3420	4.3374
BZ[2]	0.3205	3.9791
AM[4]	0.7147	3.8908
CGD2[4]	0.3797	4.5703
W1[2]	1.1425	3.772
BZ-W1[2]	0.5129	3.8756
C2M-W1[3]	0.44	3.774
C2E-W1[3]	0.5130	3.774
BZ-AM	0.48	3.935
C2M-AM	0.4944	4.1141
C2E-AM	0.5149	4.1141
BZ-CGD2	0.3488	4.2747
C2E-CGD2	0.3753	4.4538
C2M-CGD2	0.3604	4.4538

1: reference[8]
2: reference[3]
3: reference[16]
4: reference[15]

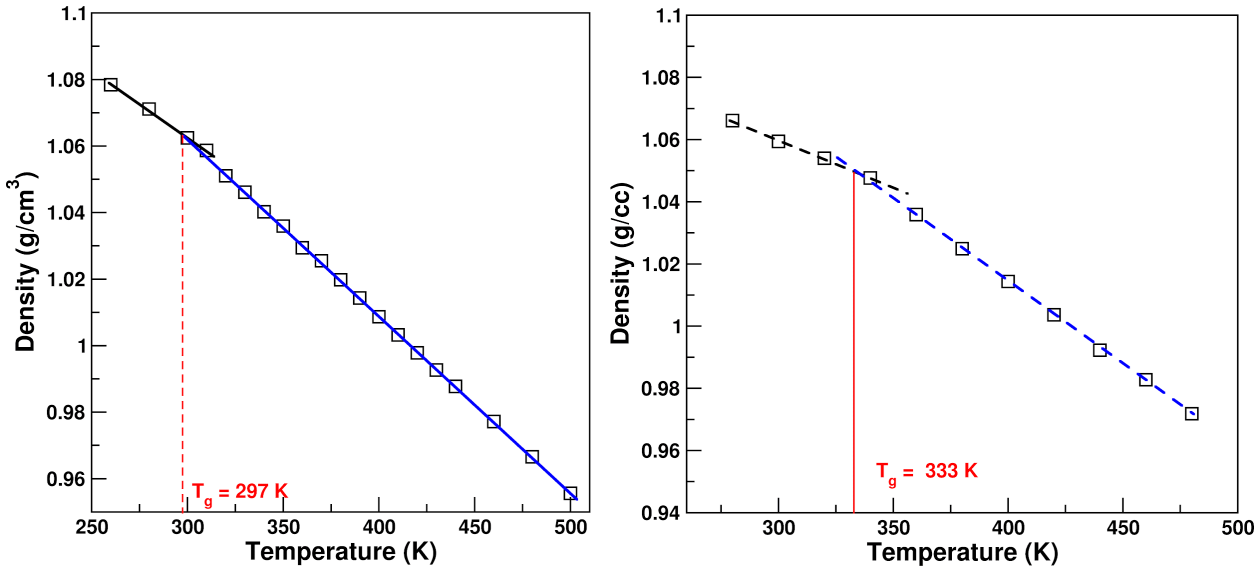

Figure D1 The glass transition temperatures of the CG PS model with the dihedral angle force constant of (left) 0 Kcal/mol and (right) 0.5 Kcal/mol.

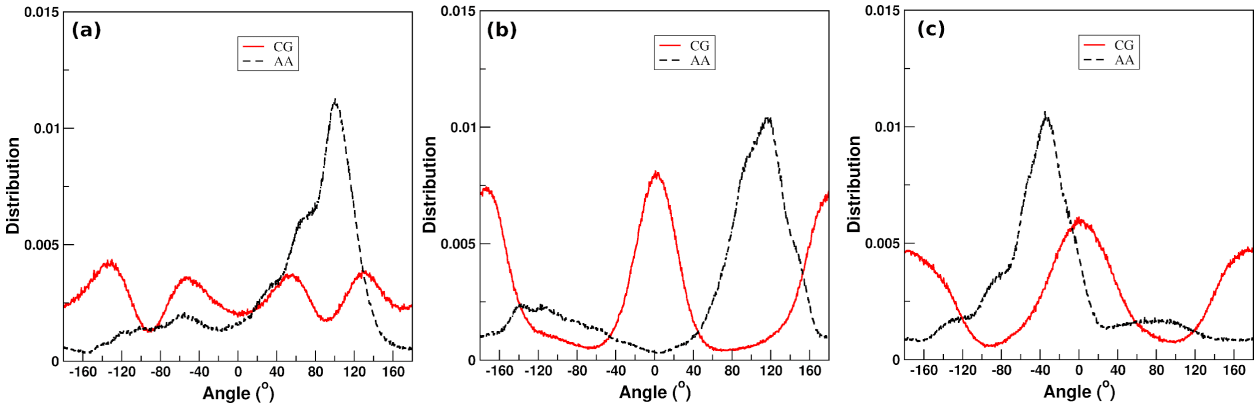

Figure D2 The dihedral angle distributions of (a) C2M-C2M-C2M-C2M, (b) BZ-C2M-C2M-BZ, and (c) BZ-C2M-C2M-BZ.

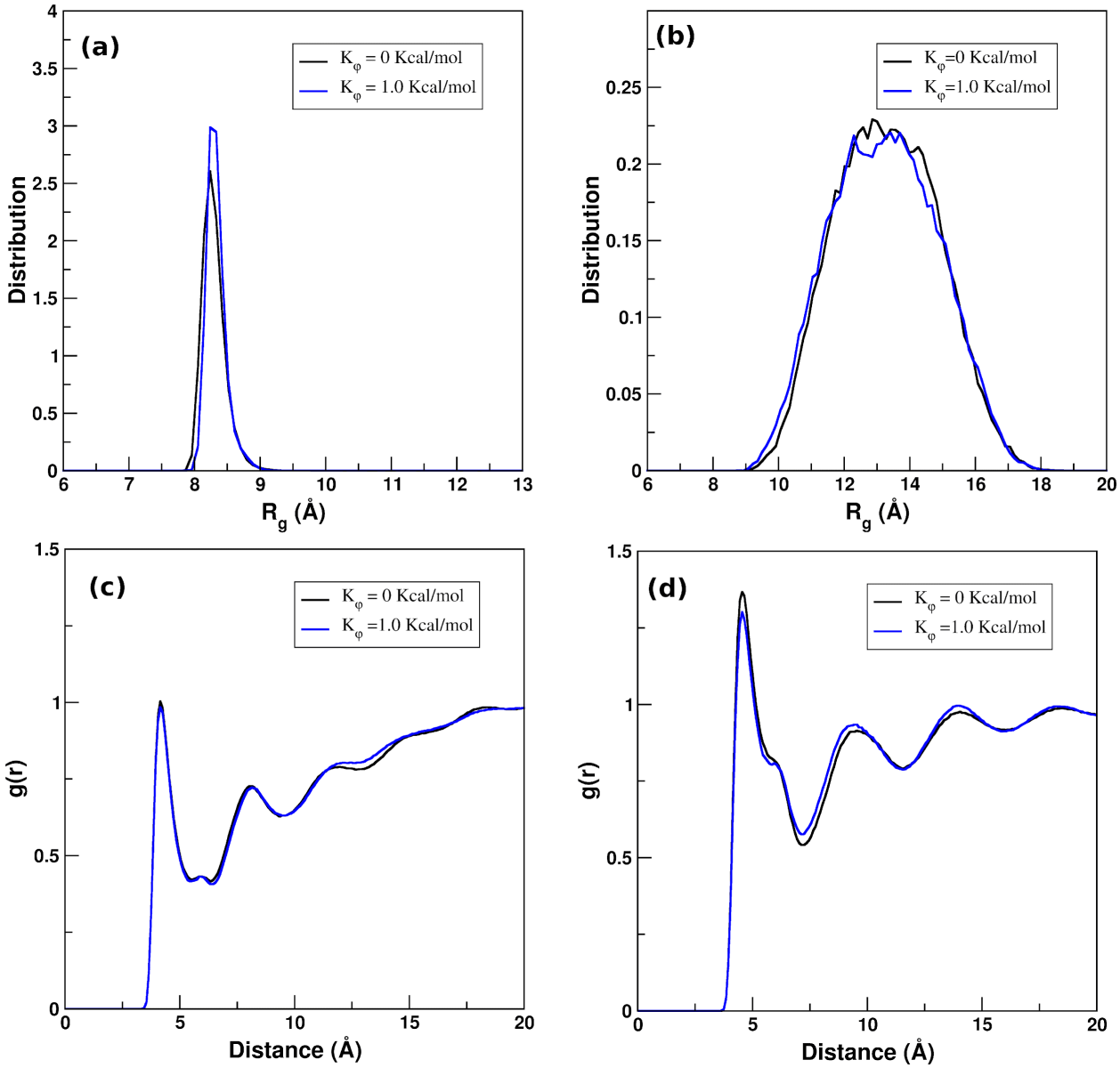

Figure D3 The R_g distributions of the CG PS model with different dihedral angle force constant (K_φ): **(a)** in pure water and **(b)** in pure DMF. The RDFs **(c)** between W1 and BZ beads, and **(d)** between CGD2 and BZ beads when the CG PS model is in pure water and pure DMF, respectively.

APPENDIX E

ML Model development

ANN is one of the most popular ML models, which has been widely used in information technologies such as image classification, and natural language processing.[20] Recently, it has also been used in computational materials science.[21–23] Here, we construct different ANN regression models by changing the number of hidden layers (1 to 4) and hidden nodes (5 to 30), to understand the effect of the number of hidden layers and hidden nodes on the R^2 score of the ANN model. Each layer is fully connected to its preceding layer. An example is shown in **Figure E1**. Each node represents a weighted linear summation function and an activation function. The activation function is ReLu in each layer. The input is the set of coordinates of CG beads, and transformed non-linearly to predict the coordinates of atoms in the all-atom model, as the output. The loss function is the mean of squared errors (MSE) between the true values and predicted ones along with the L2 penalty (alpha = 0.1). Adam algorithm with an initial learning rate of 0.001 is employed to obtain the optimized parameters in decreasing the loss function.[24] The effect of the number of nodes on its R^2 score is shown in **Figure E2**. It's found that the R^2 score of ANN models with two hidden layers increases drastically as the number of hidden nodes increases from 5 to 10. Whereas it changes slightly as the number of hidden nodes is further increased to 30. The number of hidden layers on ANN models (10 hidden nodes in each layer) has little impact on the R^2 score as shown in **Figure E3**. Hence we used the ANN models with two hidden layers and 10 hidden nodes in each layer. The training and testing of ANN models and the following k-NN, gaussian process regression, and random forests were achieved by using the Scikit-learn package.[25]

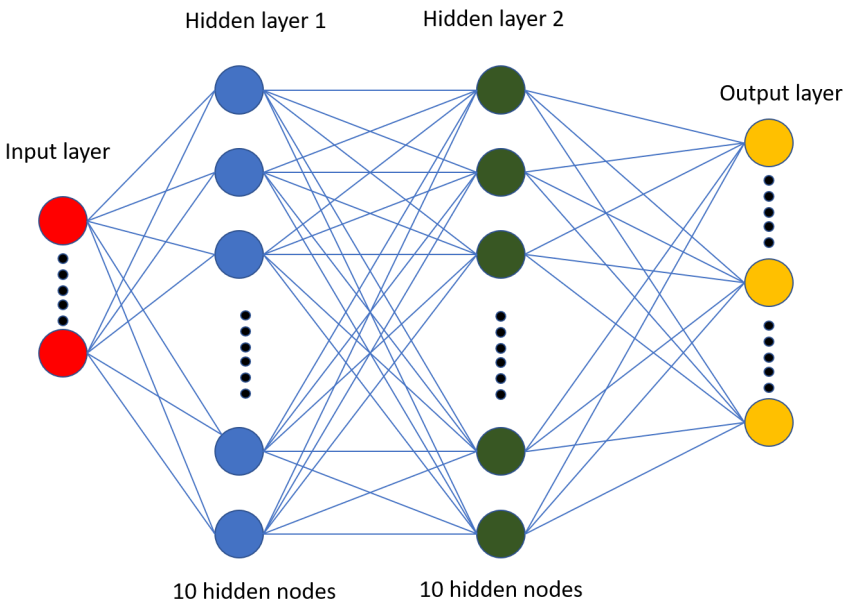

Figure E1 A representative architecture of the ANN model used in this study.

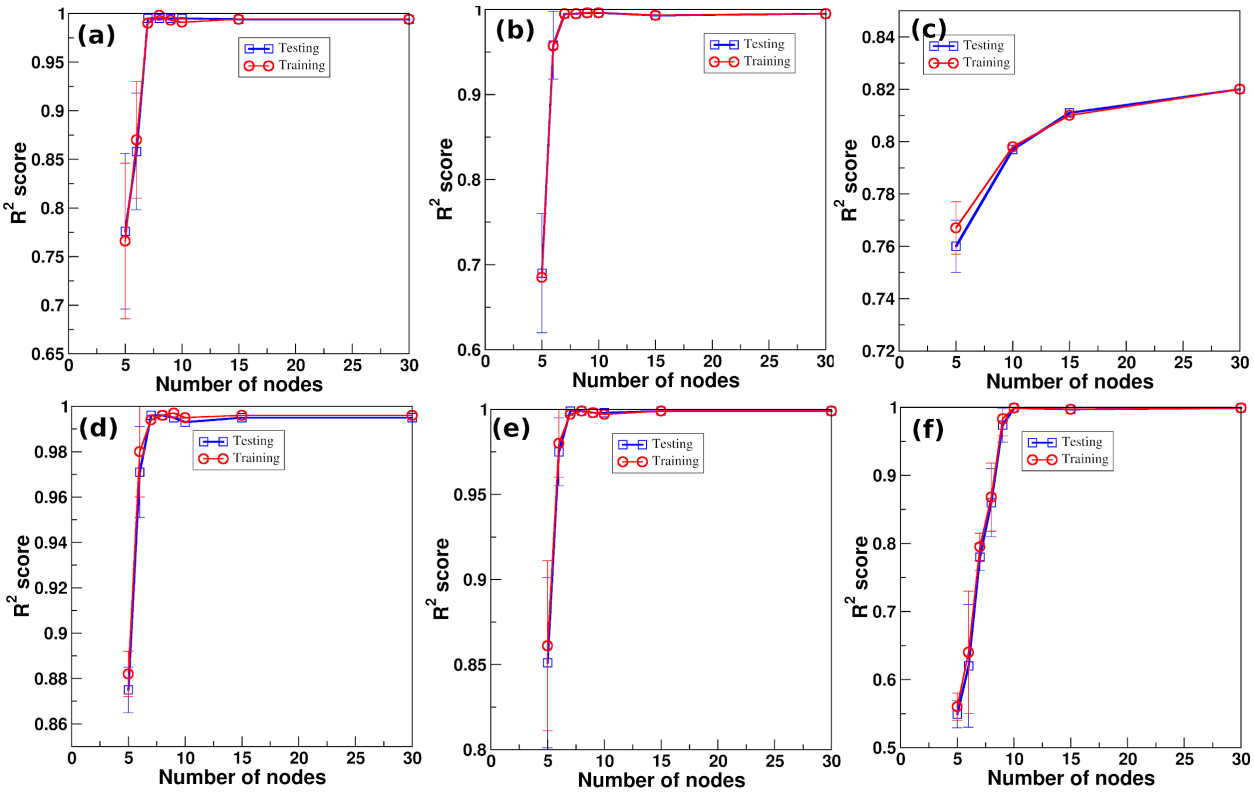

Figure E2 The R^2 scores of ANN models with different number of nodes in each hidden layer for **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, and **(f)** fullerene. The number of hidden layers in each ANN model is two and the sample size is 5000.

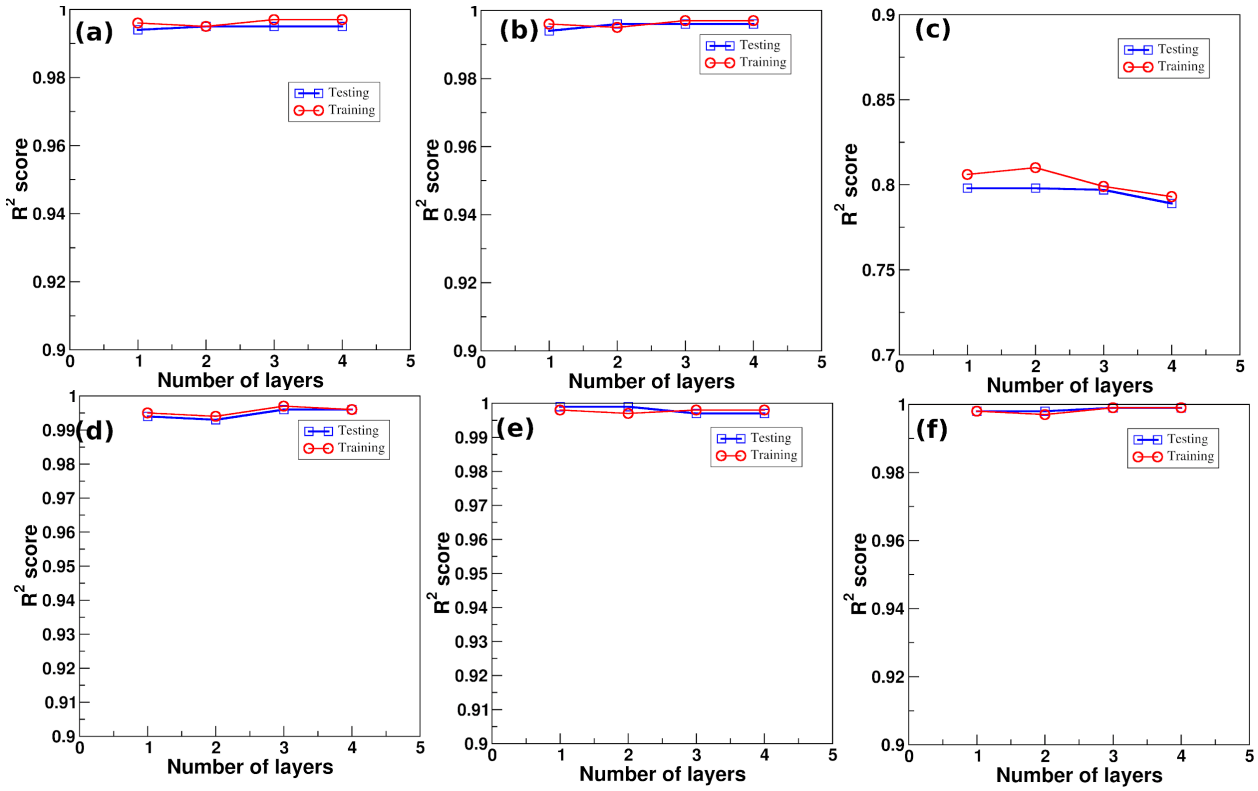

Figure E3 The R^2 scores of ANN models with different number of hidden layers for **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, and **(f)** fullerene. The number of nodes in each layer is 10 and the sample size is 5000.

k-NN is a simple ML model that uses the distance between data points to solve classification or regression problems.[26] The most used distance metric is the Euclidean distance to measure the similarity between two points. Based on this, the data points with the *k* shortest distance from the data to be predicted would be selected to assign classes or values of the unknown data. Here, we studied the performance of k-NN models with different k values (3, 5, 8). The R^2 score of k-NN models for testing is increased slightly as the k value is increased from 3 to 8 as shown in **Table E1**. As the k value increases, the R^2 score for testing is increased slightly from ~ 0.98 to ~ 0.99 for furan, benzene, and naphthalene. For hexane, it's increased much more from ~ 0.73 to ~ 0.76, by 0.03. The R^2 score for testing is always higher than 0.99 for graphene, while it's around 0.97 for fullerene regardless of k values. Overall, the k value is set to be 5 in the study unless specified.

Table E1 The R^2 scores of k-NN models with different k values. The dataset size is 5000.

	k=3		k=5		k=8	
	training	**testing**	**training**	**testing**	**training**	**testing**
Furan	0.991±0.001	0.986±0.000	0.990±0.001	0.988±0.000	0.992±0.002	0.990±0.000
Benzene	0.989±0.002	0.987±0.000	0.987±0.001	0.989±0.000	0.991±0.001	0.991±0.000
Hexane	0.734±0.003	0.728±0.002	0.761±0.002	0.750±0.002	0.771±0.002	0.761±0.001
Naphthalene	0.990±0.001	0.988±0.000	0.992±0.002	0.990±0.000	0.993±0.001	0.991±0.000
Graphene	0.992±0.001	0.991±0.000	0.990±0.001	0.993±0.000	0.992±0.001	0.994±0.000
Fullerene	0.980±0.004	0.973±0.003	0.971±0.002	0.975±0.003	0.970±0.002	0.975±0.002

RF is an ensemble model consisting of several decision trees to predict the final labels/values by selecting a subset of features for each decision tree.[24] The performance of RF models is usually better than that of one single decision-tree model. The minimum number of samples splitting a node is set to be 2, and the minimum number of samples in a leaf node is 1. The number of features to consider for best splitting is the total number of features. The effects of max_depth and n_estimators are explored in **Figure E4.** Specifically, we studied the RF models with max depth varying from 5 - 20 and n_estimators from 5 to 20. In building the RF models, MSE is used to split the trees. It can be found that the RF models are more sensitive to the changes of max depths. To be specific, as the max depth increases from 5 to 10, the R^2 score for testing increases from around 0.72 to 0.95 for furan, naphthalene, and fullerene, with a fixed number of estimators (5 or 10 or 15). On the other hand, the R^2 score for testing is only increased slightly by less than 0.1 as the number of estimators increases from 5 to 20. Based on these results, we recommend using a max depth of 10 and the number of estimators of 10 for RF models.

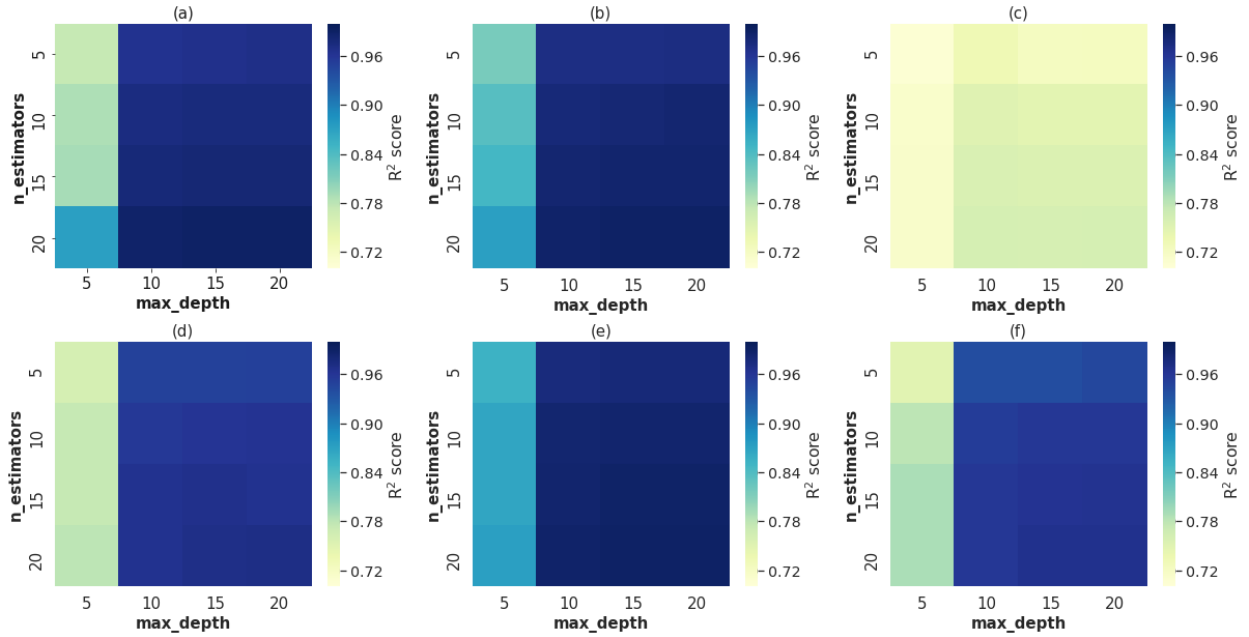

Figure E4 The heatmap for R^2 scores of RF models for testing with various max depths and number of estimators. The panels represent RF models for backmapping CG molecules of **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, **(f)** fullerene.

Kernel ridge regression (KRR) is one of the kernel-based regression methods. It converts the input data from a low dimensional space into a new high dimensional space by using kernel-trick.[27] In this study, the kernel is a three-degree polynomial kernel. SVR is the regression model as an extension of the support vector machine (SVM).[24,25] In this study, it's a multi-output SVR model where N (the dimensionality of the output data) regression models were trained on N columns of the output space, respectively. Note, KRR and SVM were used to ensure that the predictions from other models are not due to overfitting. Overfitting can be referred as when a model performs quite well on the training data but fails to predict "unseen" data.[24,28] In general, the reason for overfitting is that a model learns all the information containing noise and fluctuations in the training data to the extent that it impairs the performance of the model on new data.[24]

Table E2 The testing and training accuracies of KRR and SVM models for backmapping benzene and graphene. The dataset used consists of 5000 samples.

	Benzene		Graphene	
	Training accuracy	Testing accuracy	Training accuracy	Testing accuracy
KRR[a]	0.998	0.998	0.99	0.99
SVR[b]	0.997	0.997	0.98	0.98

a:Parameters for the KRR model:kernel = poly, alpha = 0.1, coef0 = , degree = 3.
b:Parameters for the SVR model:sklearn.multioutput.MultiRegressor(estimator = SVR(*kernel = 'rbf', degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1*))

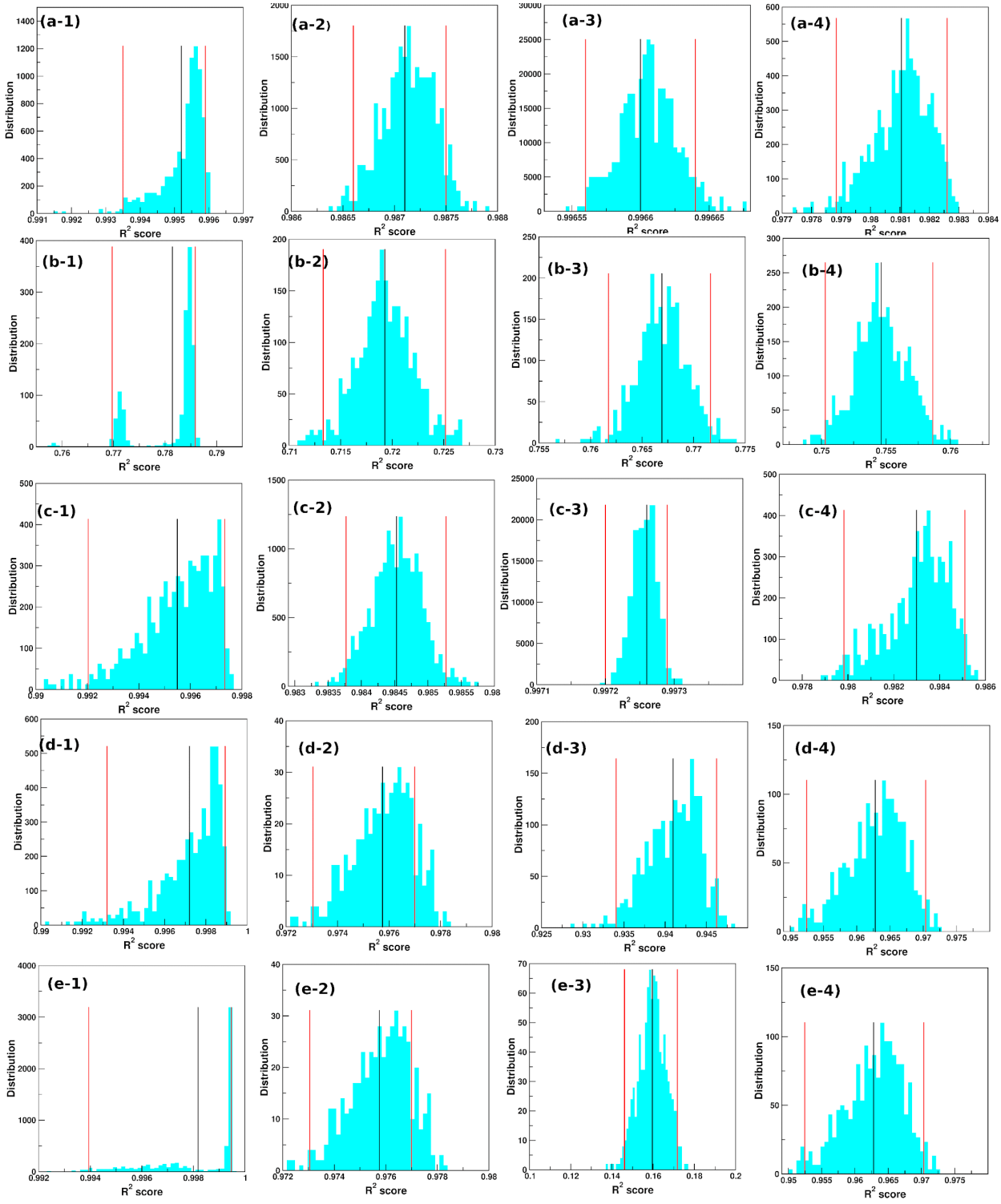

Figure E5 Uncertainty quantification of the testing R^2 scores of the models: **(1)** ANN, **(2)** k-NN, **(3)** GPR, and **(4)** RF models for **(a)** benzene, **(b)** hexane, **(c)** naphthalene, **(d)** graphene, and **(e)** fullerene.

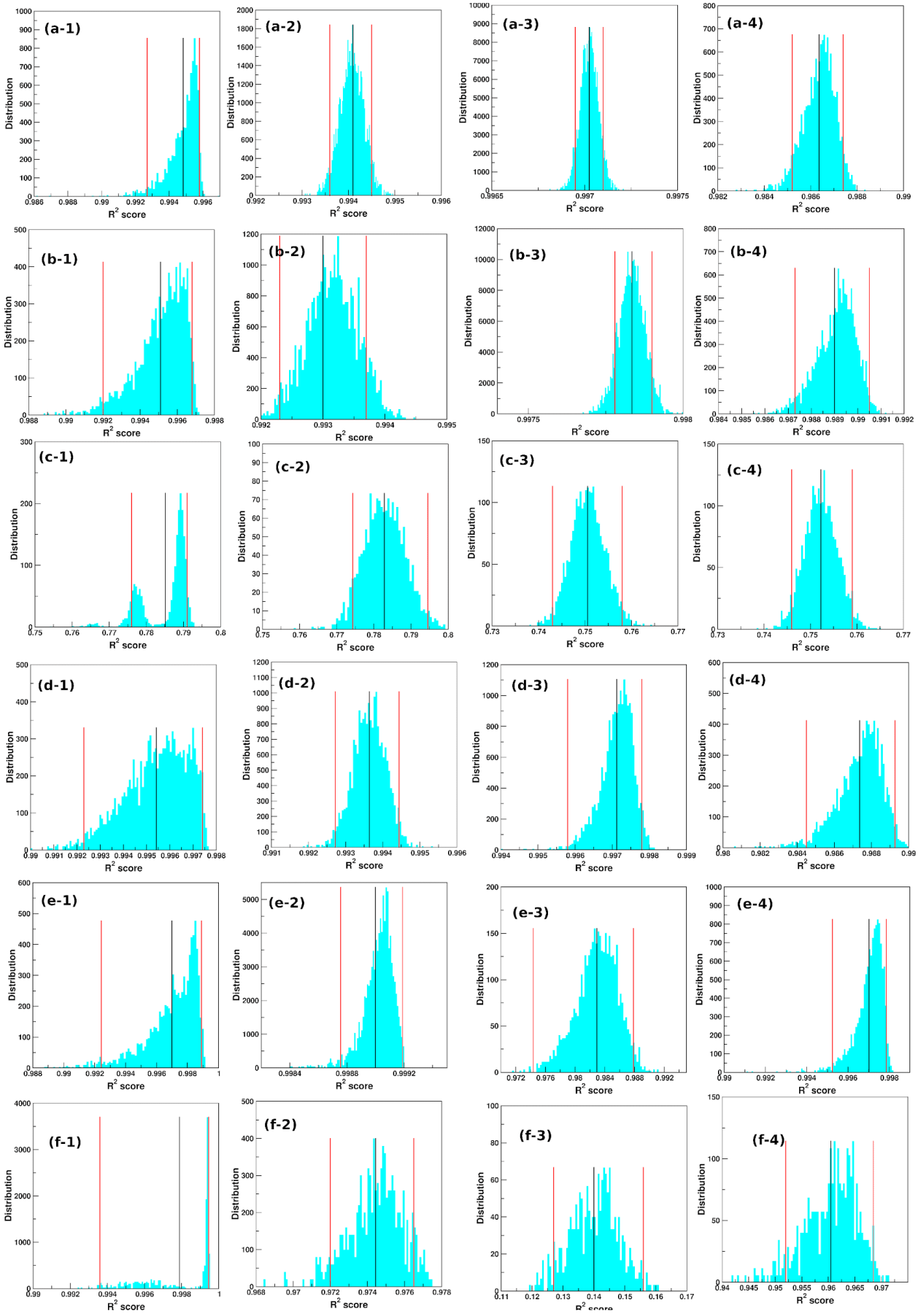
(a-1)
(a-2)
(a-3)
(a-4)
(b-1)
(b-2)
(b-3)
(b-4)
(c-1)
(c-2)
(c-3)
(c-4)
(d-1)
(d-2)
(d-3)
(d-4)
(e-1)
(e-2)
(e-3)
(e-4)
(f-1)
(f-2)
(f-3)
(f-4)
Distribution
R^2 score

Figure E6 Uncertainty quantification of the training R^2 scores of the models: **(1)** ANN, **(2)** k-NN, **(3)** GPR, and **(4)** RF models for **(a)** furan, **(b)** benzene, **(c)** hexane, **(d)** naphthalene, **(e)** graphene, and **(f)** fullerene.

References

(1) Yeh, I.-C.; Hummer, G. Diffusion and Electrophoretic Mobility of Single-Stranded RNA from Molecular Dynamics Simulations. *Biophys. J.* **2004**, *86* (2), 681–689.
(2) Yeh, I.-C.; Hummer, G. System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions. *J. Phys. Chem. B* **2004**, *108* (40), 15873–15879.
(3) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. *J. Phys. Chem. B* **2018**, *122*, 1958–1971.
(4) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750–760.
(5) Brini, E.; van der Vegt, N. F. A. Chemically Transferable Coarse-Grained Potentials from Conditional Reversible Work Calculations. *J. Chem. Phys.* **2012**, *137* (15), 154113.
(6) Eichenberger, A. P.; Huang, W.; Riniker, S.; van Gunsteren, W. F. Supra-Atomic Coarse-Grained GROMOS Force Field for Aliphatic Hydrocarbons in the Liquid Phase. *J. Chem. Theory Comput.* **2015**, *11* (7), 2925–2937.
(7) Szklarczyk, O. M.; Bachmann, S. J.; van Gunsteren, W. F. A Polarizable Empirical Force Field for Molecular Dynamics Simulation of Liquid Hydrocarbons. *J. Comput. Chem.* **2014**, *35* (10), 789–801.
(8) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. *J. Phys. Chem. B* **2018**, *122* (28), 7143–7153.
(9) Zeppieri, S.; Rodríguez, J.; López de Ramos, A. L. Interfacial Tension of Alkane + Water Systems. *J. Chem. Eng. Data* **2001**, *46* (5), 1086–1088.
(10) Ben‐Naim, A.; Marcus, Y. Solvation Thermodynamics of Nonionic Solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016–2027.
(11) Amaya, J.; Rana, D.; Hornof, V. Dynamic Interfacial Tension Behavior of Water/Oil Systems Containing In Situ-Formed Surfactants. *J. Solution Chem.* **2002**, *31* (2), 139–148.
(12) Moore, J. W.; Wellek, R. M. Diffusion Coefficients of N-Heptane and N-Decane in N-Alkanes and N-Alcohols at Several Temperatures. *J. Chem. Eng. Data* **1974**, *19* (2), 136–140.
(13) Rolo, L. I.; Caço, A. I.; Queimada, A. J.; Marrucho, I. M.; Coutinho, J. A. P. Surface Tension of Heptane, Decane, Hexadecane, Eicosane, and Some of Their Binary Mixtures. *J. Chem. Eng. Data* **2002**, *47* (6), 1442–1445.
(14) de Oliveira, T. E.; Mukherji, D.; Kremer, K.; Netz, P. A. Effects of Stereochemistry and Copolymerization on the LCST of PNIPAm. *J. Chem. Phys.* **2017**, *146* (3), 034904.
(15) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, 4667–4672.
(16) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of Transferable Nonbonded

Interactions between Coarse-Grained Hydrocarbon and Water Models. *J. Phys. Chem. B* **2019**. https://doi.org/10.1021/acs.jpcb.8b07990.
(17) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta* **2015**, *1850* (5), 872–877.
(18) Grossfield, A.; Zuckerman, D. M. Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23–48.
(19) García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68* (17), 2696–2699.
(20) Abiodun, O. I.; Jantan, A.; Omolara, A. E.; Dada, K. V.; Mohamed, N. A.; Arshad, H. State-of-the-Art in Artificial Neural Network Applications: A Survey. *Heliyon* **2018**, *4* (11), e00938.
(21) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5*, 755–767.
(22) Wang, W.; Gómez-Bombarelli, R. Coarse-Graining Auto-Encoders for Molecular Dynamics. *npj Computational Materials 5* (125), 1–9.
(23) Chan, H.; Cherukara, M.; Loeffler, T. D.; Narayanan, B.; Subramanian K R. Machine Learning Enabled Autonomous Microstructural Characterization in 3D Samples. *npj Computational Materials* **2020**, *6* (1), 1–9.
(24) Tan, P.-N.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*; Pearson Education, 2019.
(25) F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(26) Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. *Ann Transl Med* **2016**, *4* (11), 218.
(27) Singh, S. K.; Bejagam, K. K.; An, Y.; Deshmukh, S. A. Machine-Learning Based Stacked Ensemble Model for Accurate Analysis of Molecular Dynamics Simulations. *J. Phys. Chem. A* **2019**, *123* (24), 5190–5198.
(28) T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*; Springer Verlag: New York, 2001.