

THE UNIVERSITY OF CHICAGO

BOTTOM-UP COARSE-GRAINING BRIDGING VARIOUS
SCALES: FROM QUANTUM STATISTICS TO MESOSCALE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
YINING HAN

CHICAGO, ILLINOIS

AUGUST 2019

Table of Contents

List of Figures.....	v
List of Tables	xiii
Acknowledgements	xiv
Chapter 1: Motivation and Introduction	1
I. Background: Multiscale Challenge	1
II. Bottom-Up Coarse-graining.....	5
III. This Thesis.....	10
Reference	11
Chapter 2: Quantum Theory of Multiscale Coarse-graining	17
Abstract.....	17
I. Introduction	18
II. Theory	22
III. Algorithms	34
IV. Numerical Examples	38
V. Quasi-Classical Approximation.....	50
VI. Concluding Remarks	55
APPENDIX A: DERIVATION OF EQUATION (2.26).....	56
APPENDIX B: DERIVATION OF EQUATION (2.55).....	57
Reference	57
Chapter 3: Coarse-Graining Involving Virtual Sites: Centers of Symmetry Coarse-Graining.....	61
Abstract.....	61

I. Introduction	62
II. Methodology	65
III. Results and Discussion	82
IV. Concluding Remarks	94
APPENDIX A: Discussion of the limit: COS model converges to a single-site COM model when $l \rightarrow 0$	95
APPENDIX B: Derivation of $\Omega(\tilde{\mathbf{R}}^{\tilde{N}})$ and Eq. (3.11)	96
APPENDIX C: Derivation of Eq. (3.15)	97
APPENDIX D: Velocity Distribution	98
Reference	98
 Chapter 4: Mesoscopic Coarse-grained Representations of Fluids Rigorously Derived from Atomistic Models	104
Abstract	104
I. Introduction	104
II. Mapping	108
III. Application	116
IV. Concluding Remarks	148
APPENDIX A: Derivation of Eq. (4.6) and Eq. (4.10)	150
APPENDIX B: Numerical Implementation	155
Reference	157
 Chapter 5: Constructing Many-body Dissipative Particle Dynamics Models of Fluids from Rigorous Bottom-up Coarse-graining	162
Abstract	162
I. Introduction	162

II. Methodology	167
III. Results and Discussions.....	183
IV. Concluding Remarks	193
Reference	196

Chapter 6: Field Theoretic Representation of Molecular Liquids from Coarse-grained

Models	201
Abstract.....	201
I. Introduction	201
II. Theory	205
III. Gaussian Equivalent Representation	214
IV. Sampling and Structural Estimator.....	224
V. Concluding Remarks	227
Reference	228

List of Figures

- Figure 2-1:** Coarse-graining of the three particle Hamiltonian of Eq. (41) into a single bead representation, (a) the effective qCG potential of the CG sites obtained by qMS-CG method for case 1 (blue solid curve), case 2 (red solid curve), and case 3 (yellow solid curve); (b) The position probability distribution function of the simulations for case 1 (CG-blue solid line, AA-purple dashed line), case 2 (CG-red solid line, AA-green dashed line), and case 3 (CG-yellow solid line, AA-cyan dashed line); (c) The imaginary time correlation function (ITCF) defined in Eq. (44) for case 1 (CG-blue solid line with square marker, AA-purple dashed line with “+” marker), case 2 (CG-red solid line with square marker, AA-green dashed line with “+” marker), and case 3 (CG-yellow solid line with square marker, AA-cyan dashed line with “+” marker). The ITCF were only evaluated on the marked points, because the imaginary time paths were discretized in actual simulations. 41
- Figure 2-2:** Coarse-graining the three particle Hamiltonian of Eq. (2.41) into a CG single bead representation for case 4: (a) the classical PMF of the CG site (blue solid curve), (b) The quantum position probability distribution function of the simulations (CG-blue solid line, AA-purple dashed line), (c) The ITCF (CG-blue solid line with square marker, AA-purple dashed line with “+” marker). The ITCFs were only evaluated on the marked points, because the imaginary time paths were discretized in actual simulations. 43
- Figure 2-3:** Pairwise interaction between the CG sites for the modified liquid water system, as obtained by the qMS-CG method. 47
- Figure 2-4:** Panel (a) Center of mass radial distribution functions obtained from the FG quantum PIMD simulation (red solid line), Coarse-Grained PIMD simulation using the force field generated by qMS-CG (blue solid line) as well as the result from classical MD simulation

(black dash line). Panel (b) The probability distribution function for the three-body non-bonded angles obtained from the FG quantum PIMD simulation (red solid line) and qMS-CG-PIMD simulation (blue solid line). The cut-off distance for evaluating the three-body non-bonded angles was 4.0 Å.	48
Figure 2-5: The delocalization function $R^2(\tau)$ in Eq. (2.47) obtained from AA PIMD simulations (red solid line with “+” markers) and qMS-CG PIMD simulations (blue solid line with square markers). The function was evaluated only on the marked points due to the fact that $P = 32$ imaginary time slices discretized the imaginary time Feynman paths.	49
Figure 2-6: Schematic diagram showing the difference between coarse-graining in classical statistical mechanics and quantum Boltzmann statistics.	54
Figure 3-1: Schematic relationship among different CG mappings presented in this work: from the all-atom (FG) model to the one-site coarse-grained (COM model) from M and symmetric coarse-grained model (COS model) from ζ . Re-mapping the center of mass of the COS model from yields the one-site COM model.	66
Figure 3-2: CG mapping schemes considered in this study for benzene: (a) one-site center of mass (COM) mapped model as green bead, (b) two-site COM mapped model as red ($-C_3H_3$) and blue ($-C_3H_3$) beads, and (c) two-site center of symmetry (COS) mapped model as red (upper half) and blue (lower half) beads. Red beads represent CG type 1 and blue beads are CG type 2 in actual simulations.	76
Figure 3-3: CG mapping schemes considered in this study for toluene: (a) one-site center of mass (COM) mapped model as green bead, (b) two-site COM mapped model as red ($-C_4H_5$) and blue ($-C_3H_3$) beads, and (c) two-site center of symmetry (COS) mapped model as red	

(upper half) and blue (lower half) beads. Red beads are CG type 1 and blue beads are CG type 2 in actual simulations..... 78

Figure 3-4: Effective pair potentials of benzene CG models obtained from the MS-CG method for two different CG mapping schemes: center of mass (solid lines) and center of symmetry (dotted lines). Both mapping schemes show an equivalent interaction among different types (1-1, 1-2, and 2-2 pairs), indicating homogeneity. 82

Figure 3-5: Molecule-wise (i.e., re-coarse-grained to one site per molecule) structure analysis of liquid benzene by calculating the (a) radial distribution function (RDF) of the mapped-all atom (red), center of mass (blue), and center of symmetry (green) models, (b) orientation distribution function (ODF) of the mapped-all atom (black solid line) and center of symmetry (red dotted line) models, and (c) three-body angular distribution function (ADF) of the CG molecules of the mapped-all atom (red line), center of mass (blue dotted line), and center of symmetry (green line) as demonstrated in eq (24)..... 84

Figure 3-6: Effective pair potentials of toluene CG models obtained from the MS-CG method for two different CG mapping schemes: center of mass (solid lines) and center of symmetry (dotted lines). Only the COS mapping shows an equivalent interaction among different types (1-1, 1-2, and 2-2 pairs), while the COM interactions differ by CG pairs. 86

Figure 3-7: Molecule-wise (i.e., re-coarse-grained to one site per molecule) structure analysis of liquid toluene by calculating the (a) radial distribution function (RDF) of the mapped-all atom (red), center of mass (blue), and center of symmetry (green) models where magnified comparison is shown in a lower right inset, (b) orientation distribution function (ODF) of the mapped-all atom (black solid line) and center of symmetry (red dotted line) models, and (c) three-body angular distribution function (ADF) of the CG molecules of the mapped-all

atom (red line), center of mass (blue dotted line), and center of symmetry (green line) as demonstrated in Eq. (3.24). 87

Figure 3-8: Schematic description of the equilibrated benzene/toluene mixture system in (a) all-atom resolution (toluene is depicted as dark brown color and benzene is depicted as grey color), (b) COS mapped resolution (mapped molecules are shown at upper right corner) .. 88

Figure 3-9: Radial distribution function $g(r)$ in the benzene/toluene mixture. Here, $g(r)$ functions are molecule-wise (i.e., re-coarse-grained to one site per molecule) function of the mapped-all atom (red), two-site center of mass (blue), and center of symmetry (green) CG models with the one-site CG model (black dot) for (a) benzene-benzene, (b) benzene-toluene, (c) toluene-toluene molecule pairs. 89

Figure 3-10: Transferability of the COS CG interactions between the bulk liquid (benzene, toluene) and its mixed phase compared to the COM CG interactions. Effective CG pair potentials from (a) the one-site COM model of benzene in the bulk phase (brown dotted) and the mixed phase (gold line), (b) the one-site COM model of toluene in the bulk phase (purple dotted) and the mixed phase (magenta line), (c) the COS model of benzene in the bulk phase (red dotted) and the mixed phase (red line), and (d) the COS model of toluene in the bulk phase (blue dotted) and the mixed phase (blue line) with almost identical shapes. 91

Figure 3-A1: Molecule-wise radial distribution function of liquid benzene for three different models: the single-site COM model (black line) and the two-site COS models. Here, the COS schemes used in this work are shown as a turquoise line and compared with the limit (green dots) when the separation distance $l_{\text{COS}} (= l)$ is too small (0.1 Å). In this limit, it is immediately apparent that the COS mapping converges to the single-site COM mapping. 95

Figure 3-D1: Probability distribution function of the x-component of virtual CG bead velocities v_x from the mapped atomistic trajectories, giving $T = 300$ K.....	98
Figure 4-1: Schematic diagrams showing the two steps for determining the positions of CG blobs. Red circles denote the CG blobs and blue circles represent FG molecules, (a) The contribution of one particular FG molecule to its nearby CG blobs, i.e., the ownership fraction, is determined by their spatial distances. (b) The CG blob positions are determined by these ownership fractions.....	110
Figure 4-2: The proximity function profile described by Eq. (4.12).	118
Figure 4-3: The probability distribution function of the inertia masses scaled by CG resolutions (n/N). The four CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).....	123
Figure 4-4: Panel (a) the thermal mass distribution at different CG resolutions. Panel (b) the natural logarithm of the distributions, with masses scaled by the CG resolution. The four CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).	124
Figure 4-5: The velocity distribution profiles obtained from the mapped CG trajectory (blue solid line), the Maxwell Boltzmann distribution with Maxwell mass and $T^* = 2.0$ (red dashed line) and with mean inertia mass and $T^* = 2.0$ (black dashed line). The blue solid line and the red dashed line are virtually indistinguishable.	125
Figure 4-6: Panel (a) the radial distribution functions (RDFs) of the CG trajectory at different resolution mapped by the proposed CG mapping. Panel (b) the same RDFs with radial distances scaled by the cube root of CG resolution. The resolutions for the CG systems are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).	127

Figure 4-7: The local density distribution obtained from the mapped CG trajectories. The CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.	128
Figure 4-8: the profile of x component of velocity under shear flow for both the FG LJ fluids (blue dashed line with cross marker) and the CG system mapped through dynamic mapping (red solid line with square marker).	130
Figure 4-9: The effective interactions between CG beads at different resolutions derived through the MS-CG method, (a) the potential energy profile and (b) the force profile. The RDFs obtained by running simulations using the inferred CG interactions are presented in (c)-(f); the RDFs calculated from the mapped CG trajectories are also shown for comparison. The CG resolutions, i.e. the number ratios of the FG to CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.	135
Figure 4-10: The scaled pairwise force profile between CG beads. The radial distances are scaled by the cube root of the CG resolution, whereas the forces are scaled by $n_{FG} / \sqrt{N_{CG}}$. The scaled force profiles can be fitted by a sigmoid function, shown as the black dashed line.	136
Figure 4-11: The local density distribution obtained by propagating the CG dynamics using the inferred CG interactions is presented in panels (a)-(d), with the references calculated from the mapped CG trajectories also shown for comparison. The number ratios of the FG and CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple), respectively.	137
Figure 4-12: Error analysis of the RDFs and local density PDFs: Panel (a) the relative error of $\max\{g(r) - 1\}$ between the MS-CG results and mapped CG trajectory as a function of CG	

resolution, Panel (b) the $\sigma_{CG} / \sigma_{MAP} - 1$ as a function of CG resolution, where σ_{CG} and σ_{MAP} represents the variance of the local density probability distributions of the MS-CG results and mapped CG results, respectively.....	138
Figure 4-13: The GLE memory kernel inferred from the mapped CG trajectory at different CG resolution. The CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.	144
Figure 4-14: The velocity autocorrelation functions (a)-(d) and the log-log plot of them (e)-(h) obtained from both the GLE CG simulation and the mapped CG trajectory. The number ratios of the FG and CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.	145
Figure 4-15: The GLE memory kernels scaled by $(n/N)^{3/5}$, where n/N represents the resolution of CG models. The four CG resolutions adopted here are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.	146
Figure 4-16: The scaled VACFs, with the time scaled by the cube root of CG resolution and the correlation function scaled by inverse CG resolution N/n . The four CG resolutions adopted here are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively. Each solid line indicates the VACF of a mapped trajectory, while the dashed line with the same color denotes the results obtained from corresponding CG GLE propagations.	147
Figure 5-1, illustration of the essential steps of the dynamical coarse-grained mapping procedure.....	171
Figure 5-2: illustration of the dynamical information one can obtained from the CG trajectory obtained through dynamic CG mapping approach. The pairwise decomposition of instantaneous forces cannot be realized in this scenario.....	179

Figure 5-3: the many-body DPD type interaction inferred based on the method developed in this work, as well as their performance in reproducing mesoscopic structure of the underlying fluid systems. Panels (a) to (c) correspond to the higher resolution case where the CG blob density $\rho_{CG} = 0.0625$, whereas panels (d) to (f) represents the case of $\rho_{CG} = 0.0320$. The many-body modulated interactions, which originate from the density weight function, are represented by the red lines in panels (a) and (d). The blue lines in these two panels denote the pairwise additive part. Panels (b) and (e) present the comparison of radial distribution functions between the CG models and reference mapped results. Panels (c) and (f) showcase the comparison between local density distributions.	187
Figure 5-4: the inferred friction kernels as well as the corresponding colored noise generators for CG models at two resolutions. The temporal component of friction kernels is shown in panel (a) whereas the spatial part is presented in panel (b). Panels (c) and (d) showcase the colored noise generator for constructing fluctuation forces.	189
Figure 5-5: the comparison of VACFs estimated from mapped CG trajectory and the non-Markovian DPD dynamics results.	191
Figure 5-6: the comparison of VACFs estimated from mapped CG trajectory and the Markovian DPD dynamics results.....	192
Figure 6-1: schematic diagram showing the concept of hierarchical coarse-graining for the construction of field theoretic representation of fluids.	204

List of Tables

TABLE 2-I. Parameters for the 1D model systems (in atomic units).....	39
TABLE 3-I: Evaluating the transferability of COM and COS CG models by employing relative entropy (or Kullback-Liebler divergence) as a measure. For each case, the relative entropy was calculated between the CG PMF obtained in bulk phase and mixture phase.	93
TABLE 4-I. Description of the FG to CG mapping algorithm.	156
TABLE 4-II. Description of the self-consistent position mapping algorithm.	157
TABLE 5-I. Diffusion constants estimated from the mapped CG trajectory and the parameterized Non-Markovian and Markovian DPD models.	193

Acknowledgements

I feel very glad and honored to have spent my past five years at University of Chicago. My experience here has made me a better person thanks to the fabulous atmosphere and resources the university maintains. A lot of my time here has been full of challenges, but thankfully most of it has been incredibly rewarding. I could not have made through any of it without the support from many people I have been lucky enough to meet and work with.

First and foremost, I would like to express my sincere gratitude to Professor Gregory A. Voth, for his mentorship and support along the way. He has been generous with his time, resources and advices to support my research and life throughout my Ph.D. He has a keen vision and intuition of chemical physics. Many of our discussions were enlightening and helped me shape my mind in many important scientific problems. And I always admire the tremendous amount of effort he puts every day to maintain such a high caliber research group. I would also like to thank many other faculties in the chemistry department, especially Professor Benoit Roux and Professor Suri Vaikuntananthan for their help during my candidacy exam and serving on my defense committee.

I'm extremely appreciative of all the members in the Voth group, past and present, who I have been lucky to work with. I would like to thank Dr. Yuxing Peng and Dr. James F. Dama for their kind guidance at the early stage of my research. I would like to thank those who have collaborated with me on many interesting scientific problems: Jaehyeok Jin, Dr. Jacob Wagner, Dr. Rui Sun, Dr. James Dama, Zack Jarin and Harry Ryu. I enjoyed much insightful conversation with many other members as well, especially Dr. Glen Hocky, Dr. John Grime, Dr. Jesper Madsen, Dr. Steve Tse, Dr. Alexander Pak, Aleksander Durumeric, Chenghan Li and Dr. Zhi Yue.

My friends, whether within the university or not, have given me amazing times and kind support in the United States. In particular, I would like to thank Jaehyeok Jin, for our fruitful collaboration and countless good times hanging out and exploring Chicago. I also want to say thank you to many friends here in the campus, Paul, Laura, Zack, Shane, Memo, Ziwei, Kade, Zhenghan and Hailing. My life in Chicago would not be full of joy without you guys. I would like to thank Hongfei Li for many of our wonderful adventures around the continent, and his help and support of my career development. Additionally, I would like to thank Ruyi Wang, Jiacheng Zhang, Jialu Liu and Jiajing Li, for our best lunch squad and many trips to Chinatown to comfort my nostalgia every now and then.

Finally, I would not be here defending this thesis today without the support of my family. I would like to thank my mom Ms. Lijun Yi, who has been giving me unconditional love and support and first showed me the wonderful world of chemistry, and my dad Mr. Peilin Han, who has been showing me how to be a brave man.

Chapter 1: Motivation and Introduction

I. Background: Multiscale Challenge

Since the emergence of scientific computing dating back to the Manhattan project in 1940s, scientists with different training background were soon fascinated with the strong power of computing in the discovery of exotic matter and understanding mechanisms behind novel physical phenomena. Many of the beautiful analytical theories could finally be applied to problems with practical significance instead of just ideal model calculations. With the incredible development of computing technology since late 20th century, computer simulation techniques have been applied to research in nearly all branches of physical, biological and social sciences as well as engineering simulation and design.

During the early stages of the stream, scientists with different focuses created their own tools to simulate the physical processes of their particular interest. For instance, high energy physicist created Lattice Quantum Chromodynamics¹⁻³ (LQCD) to study strong interactions and structure of nuclei. At a lower energy scale when electrostatic interactions dominate the important physics, quantum chemists and condensed matter physicists developed various numerical methods to study the electronic structure of molecules and solid state crystalline materials by solving Schrödinger equation based on either wave function methods⁴⁻⁵ or density functional theory.⁶

At a much longer time scale that the motions of nuclei become the focus and the dynamics of electrons can be averaged out to generate a potential energy surface for the nuclei, molecular mechanics⁷⁻¹⁰ (MM) has been intensively developed to depict the potential energy surface using the nuclei coordinate information. Many numerical methods have been developed to propagate the dynamics and gather statistical properties of the molecular system, including the most famous Molecular Dynamics¹¹ (MD), Monte Carlo¹² (MC) and various rare event sampling algorithms

such as umbrella sampling,¹³ Replica Exchange MD,¹⁴ metadynamics¹⁵ and transition path sampling.¹⁶ The application of MM has been extended to many highly complicated molecular systems of great importance in biology, such as full HIV capsid,¹⁷ actin filaments,¹⁸ and ion channels.¹⁹ However, there are two issues that limit the accuracy and applicability of MM. First, the empirical potential energy surface might not be fully compatible with quantum mechanics result and therefore introduce errors in describing interactions. And second, important rare events such as large scale motion of molecules and phase transition behavior that involves collective motions of atoms, requires long time simulations that is not approachable by atomistic MD/MC even with the most advanced rare event sampling techniques. A large portion of the computational power was wasted on simulating molecular vibrations that is not with great interests at a larger scale.

To surpass the inefficient sampling problem of MM, scientists have explored an alternative path that designs models of materials based on macroscopic physical theories instead of their atomistic composition and structure. Many phenomenological models of condensed matters beyond the molecular scale have been created and widely used to study the complex organization and transport phenomena at mesoscale. To name a few, a number of coarse-grained (CG) protein²⁰⁻²² and lipid models^{20, 23} have been developed to study important biophysical processes like protein folding, protein-protein interaction and lipid bilayer phase behavior. These CG models were often parameterized to stabilize existing molecular structure databases and experimental observations.²² Similar CG models also emerge in polymer physics and fluid mechanics, since de Gennes' famous blob model in studying the scaling behavior of polymer solutions. Several successful models were built up on the physical concept of blobs in which many atomistic entities are represented by one blob. The famous dissipative particle dynamics²⁴⁻

²⁵ (DPD) and its derived forms²⁶ have been broadly applied to study polymer self-assembly and transport phenomena of complex multiphase flow.²⁷⁻²⁸ Except for particle-based models, the tool of statistical field theory has also been widely accepted in polymer physics,²⁹ in which the spatial distribution of the polymer species instead of their atomistic information becomes the central entity for theoretical construction and numerical simulation. Despite their success in many applications, there are several drawbacks of these phenomenological mesoscopic models. Due to the missing correspondence with underlying atomistic structure and chemical composition, there is little verification and reasoning of the parameters applied in these models. The transferability of these model parameters is also unconfirmed for many cases. And lastly, since large numbers of degrees of freedom are reduced, the dynamics generated by these mesoscopic models is often unreliable.

Beyond the mesoscale, continuum mechanics describes the important physics of materials. Scientists have created computational fluid mechanics (CFD) and computational solid mechanics to simulate physical processes at this length scale. Unlike the particle based simulation techniques at molecular and mesoscale, the numerical solution of partial differential equations (PDEs) lies at the core of CFD. Various numerical schemes discretizing Navier-Stokes equations have been developed. In Eulerian description in which the solution space is meshed and field variables are solved on the meshes, there are mainly three numerical approaches to discretize Navier-Stokes equations: finite difference, finite element and finite volume methods. Alternatively, in Lagrangian description, the Navier-Stokes equation can be discretized using smooth kernel functions. This discretization scheme gives rise to smoothed particle hydrodynamics (SPH) method.³⁰ All these CFD methods have shown extraordinary performance in reproducing important fluid structure and flow phenomena just incorporating several

macroscopic physical parameters of the fluid, and therefore have been widely used in engineering design. However, the coupling between CFD with mesoscopic or atomistic models is quite challenging, due to the lack of correspondence between the fluid models with underlying molecular structure. This fact limits the application of CFD in the study of many important flow coupled microscopic transport phenomena, which is ubiquitous in cellular biology.

Among all the simulation techniques discussed above, it is not hard to spot that the limitations of all these models mainly come from the lack of correspondence with physics or entities at other spatiotemporal scales, which is the so-called multiscale challenge.

The desirability of defining these missing correspondences thus has been widely accepted in the scientific computing community since at least 1970s.³¹ To fill the gap between electronic structure and molecular mechanics, various new types of force field designed based on quantum mechanics results has been developed. And the existing force fields have been adapted through intensive parameterization against highly accurate quantum chemistry calculations. The electron polarizability,³²⁻³³ reactivity,³⁴⁻³⁵ many-body effects³⁶ and nuclear quantum effects have been integrated into force fields, which significantly enhanced the accuracy of reproducing important structural and dynamical correlations using these force fields. And the quantum mechanics / molecular mechanics³¹ (QM/MM) embedding scheme has gained enormous strides after 30 years' development.

The correspondence between molecular and meso- or macroscale models, however, is less explored and remains unclear in most cases. First, the mapping between high resolution molecular entities with low resolution model representations is often arbitrary or ill defined. Especially for unbonded fluid systems when a CG blob represents multiple fine-grained fluid molecules, a well-defined CG mapping was not developed until very recently by us.³⁷⁻³⁸ A

systematic way of defining fine-grained (FG) to CG mapping is far from established. Second, the interactions in mesoscopic models are often non-isotropic and many-body dependent, especially for highly heterogeneous biomolecules. The parameterization of the conservative CG interactions is non-trivial and could be more complicated than the parameterization of MM force field, due to the relatively weak understanding of mesoscopic physics than molecular level physics. And lastly, the correspondence between FG and CG dynamics is hard to establish. The reduced degrees of freedom result in complex dissipative and fluctuating interactions on CG entities. The research on recovering CG dynamics is still at relatively early stage. There are many opportunities and open questions left to be solved in this field. Inspired by the challenges and opportunities, the central topic of this thesis is to build up a series of methods to connect molecular description of soft condensed matters with meso- and macroscopic models, mainly based on the concept of bottom-up coarse-graining.

II. Bottom-Up Coarse-graining

There are mainly two paradigms, “top-down” and “bottom-up”, in the coarse-grained and mesoscopic modeling community. In top-down models, the interactions are typically parameterized without explicit consideration of the underlying fine-grained models. Whereas in bottom-up models, the CG interactions are parameterized purely based on the fine-grained system dynamics.

Both paradigms have its strengths and weaknesses. The top-down models do not relate to a more detailed model. Instead, they address the experimental phenomena directly using the CG variables. In that sense, they can be considered as a “fine-grained” model constructed to represent experimental results at even coarser level. The parameterization of top-down models is very flexible and it is often observed that the quality of models highly relies on the understanding

of physics at the spatiotemporal scale that the model is parameterized.²² The downside of top-down model is also apparent. First, the lack of fine-grained correspondence makes it hard to justify the model, since the same emergent phenomena at coarser level may be generated through different models. Second, the top-down models are typically parameterized to capture only a small subset of features of the system, which might lead to wrong interpretation for other important characters. Top-down models are usually under-constrained by the available experimental data. These facts hinder the emergence of a systematic methodology that can be used to gradually improve the quality of top-down CG models.

The concept of bottom-up coarse-graining may date back to 1960s, when Kadanoff proposed the “block-spin” Renormalization Group (RG) approach to solve Ising models.³⁹ And this concept has since been extended to molecular systems and derived many methods to derive coarse-grained force fields of molecular liquids and biomolecules. As a data-driven approach, bottom-up coarse-graining holds great promise for further improving the quality and scope of CG models thanks to the rapid development of computing technologies and data science. A large amount of fine-grained simulation data is generated each day, which can all be utilized to improve the training of bottom-up coarse-grained models.

However, the general theory of bottom-up coarse-graining is not yet fully developed. One may face many theoretical and methodological difficulties while parameterizing systems with moderate complexity. For example, the bottom-up CG models often suffer from transferability⁴⁰ and representability problems,⁴¹ which limits the application and extension of parameterized models. The immaturity of bottom-up theory and methods gives us many open questions to solve and opportunities to improve the approach, which will be a focus of this thesis.

Bottom-up coarse-graining methods connect physical models at two different scales by defining two essential elements. The first element is the CG mapping function, which maps a set of fine-grained entities into a set of coarse-grained entities. The mapping is often unidirectional, and can be reversed in a probability manner. Even though seems like not too hard at first place, it is extremely challenging to define a reduced representation of a many-body system that maximally preserve essential information and keep the conciseness at the same time. A systematic method of defining coarse-grained representations, or equivalently reaction coordinate / collective variables of molecular system is still one of the unresolved open questions in statistical mechanics.⁴²

At molecular level, the CG mapping function is often chosen to be constant, linear and non-overlapping, which can be represented as a block diagonal matrix that is not necessarily full rank. The commonly used center of mass (COM) mapping and alpha carbon representation of amino acid all fall into this category. However, such simple mapping function becomes ill defined when one tries to map multiple unbonded molecular entities as one CG bead, which is exactly the case for mesoscopic models. As time evolves, the unbonded molecules soon diffuse away from their original neighbors and render the COMs of these molecules meaningless. In Chapter 4 of this thesis, the convention of such static and linear mapping will be challenged and a non-linear dynamic mapping scheme will be developed, to solve the challenge of defining CG mapping beyond molecular scale.⁴³

The second element of bottom-up CG is the parameterization of CG equations of motion (EOM) based on the fine-grained trajectory after mapping. This component of bottom-up CG is arguably the most explored and attracts much more attention and efforts than the CG mapping problem. The theoretical foundation of this problem was originally built up in the Mori-Zwanzig

formalism, which describes the non-equilibrium evolution of any set of functions defined on the phase-space of a Hamiltonian system.⁴⁴ And later the whole theoretical framework was reformulated in the GENERIC structure by Espanol.³⁷⁻³⁸ The results of these theories are quite neat and intuitive. The dynamics of CG variables can be described by a Generalized Langevin Equation (GLE) type stochastic integro-differential equation, in which the instantaneous forces on CG variables have three components. The first component represents the many-body potential of mean force (PMF) of the CG variables, which comes from the mean force of all the underlying fine-grained atoms conditioned on the CG configuration. The second component is the dissipative force describing the time correlation of CG forces. The time-dependent many-body friction kernel is the central quantity in this term. And lastly the third component is the fluctuation force originating from reducing the number of degrees of freedoms. The fluctuation and dissipation components are not independent and are connected through the fluctuation dissipation theorem. The parameterization of CG EOM then requires the parameterization of both the many-body PMF and the friction kernel, with the former one determining the structural and thermodynamics of the CG system and the later one determining the CG dynamics.

To approximate the CG many-body PMF, many methods has emerged since late 1990s. The early attempts were mostly based on matching certain structural features such as bond lengths and radial distribution functions (RDFs), such as the inverse Monte Carlo (IMC),⁴⁵ Boltzman Inversion (BI) and its iterative version (IBI).⁴⁶ In the early 21th century, variational methods based on the statistical mechanics of many-body PMFs were developed including the Multiscale Coarse-graining⁴⁷⁻⁵¹ (MS-CG) and Relative Entropy Minimization⁵²⁻⁵³ (REM) methods. These methods have solid theoretical foundation and thus have gained numerous successes in many complex chemical systems. In the early stage, the CG PMF was often assumed to be pairwise

decomposable due to the simplicity and universality of this assumption. However, unlike the case of MM force fields, the pairwise decomposition may not guarantee to work well in approximating the many-body CG PMF. Recently, there has been an increasing interest in directly incorporating many-body interactions in representing CG force fields. The explicit many-body interactions could rapidly increase the computational cost. More economic representations of the many-body interactions need to be developed. It has been demonstrated that the recently proposed Ultra Coarse-graining (UCG) approach⁵⁴⁻⁵⁷ can be used to significantly improve the expressiveness of CG force field yet maintains a same computational cost as pairwise additive interactions, and has been applied to capture liquid-vapor, liquid-liquid interfaces as well as recapitulate hydrogen bonding networks at a coarse-grained level. Another trial on this route is based on local order parameter dependent interaction,⁵⁸ which has also been shown to successfully recapitulate interfacial phenomena at coarse-grained resolution. In Chapter 3 of the thesis, a new strategy of introducing many-body interactions through virtual CG sites⁵⁹ is developed based on the MS-CG framework, in which essential symmetry features of the underlying fine-grained system is preserved without additional computational cost compared to pairwise interaction.

Finally, in order to reproduce the dynamics of the coarse-grained variables, the non-conservative dissipation and fluctuation forces need to be derived from the mapped trajectory information. The bottom-up derivation of the non-conservative interactions is relatively less studied compared to the statistical inference of many-body PMF. The friction kernel is a more complicated object than many-body PMF, since it is both intrinsically many-body and time dependent. And the constraint enforced by fluctuation dissipation theorem makes the statistical inference of these quantities even more difficult. In the past decade, scientists start to put more

efforts and have made remarkable progress on solving this problem. Many approximations of the Mori-Zwanzig formalism have been investigated and implemented for relatively simple cases.^{26,}

⁴⁴ However, there are still no tools available to deal with systems exhibiting strong non-Markovian dynamics and with spatial heterogeneity. A compatible representation of the CG dynamical equation with respect to macroscopic physics is also absent in a bottom-up manner. A satisfactory answer to the CG dynamics problem is far from complete. In this thesis, several developments of solving this problem will be addressed, and will be presented in Chapters 4 and 5.

III. This Thesis

The remaining part of the thesis is organized as five independent chapters, with each chapter presenting a methodology development to build up a connection between models at two different scales.

Chapter 2 describes a theoretical extension and numerical verification of Multiscale Coarse-graining method in quantum Boltzmann statistics. This method for the first time introduces a way to embed nuclear quantum effects into CG models in a rigorous bottom-up manner.

Chapter 3 introduces an effort to incorporate virtual CG sites to represent essential symmetrical features and associated many-body interactions using just pairwise basis set. The method is based on maximum entropy principle and the variational force matching approach, and has been applied to model non-isotropic molecular liquid like Benzene and Toluene, at the center of mass resolution.

Beyond the high resolution coarse-grained models at molecular level, Chapters 4-6 are dedicated to build up bridges between molecular scale and meso- and even macroscale CG models. In Chapter 4, a new CG mapping scheme and a parameterization strategy associated

with it will be developed to construct mesoscopic CG models of unbonded fluids directly from atomistic models, which is the first successful trial to build up a rigorous connection between atomistic and mesoscopic models of diffusive fluids. Chapter 5 describes a systematic strategy of parameterizing mesoscopic models that originate from fluid mechanics and classical density functional theory, directly using the CG trajectory information obtained through the method elaborated in Chapter 4. And finally, chapter 6 propose a strategy of transforming particle based description of coarse-grained models into field representation, which provide a theoretical basis for the field theoretic simulation methods widely used in modeling complex molecular liquids and polymer physics.

Reference

1. Bazavov, A.; Toussaint, D.; Bernard, C.; Laiho, J.; DeTar, C.; Levkova, L.; Oktay, M. B.; Gottlieb, S.; Heller, U. M.; Hetrick, J. E.; Mackenzie, P. B.; Sugar, R.; Van de Water, R. S., Nonperturbative QCD simulations with $2+1$ flavors of improved staggered quarks. *Reviews of Modern Physics* **2010**, 82 (2), 1349-1417.
2. Dürr, S.; Fodor, Z.; Frison, J.; Hoelbling, C.; Hoffmann, R.; Katz, S. D.; Krieg, S.; Kurth, T.; Lellouch, L.; Lippert, T.; Szabo, K. K.; Vulvert, G., Ab Initio Determination of Light Hadron Masses. *Science* **2008**, 322 (5905), 1224.
3. Petreczky, P., Lattice QCD at non-zero temperature. *Journal of Physics G: Nuclear and Particle Physics* **2012**, 39 (9), 093002.
4. Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio Jr, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Yeh Lin, C.; Van Voorhis, T.; Hung Chien, S.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Min Rhee, Y.; Ritchie, J.; Rosta, E.; David Sherrill, C.; Simmonett, A. C.; Subotnik, J. E.; Lee Woodcock Iii, H.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer Iii, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M., Advances in methods and algorithms in a

modern quantum chemistry program package. *Physical Chemistry Chemical Physics* **2006**, *8* (27), 3172-3191.

5. Bartlett, R. J.; Musiał, M., Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics* **2007**, *79* (1), 291-352.

6. Kohn, W.; Becke, A. D.; Parr, R. G., Density Functional Theory of Electronic Structure. *The Journal of Physical Chemistry* **1996**, *100* (31), 12974-12980.

7. Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983**, *4* (2), 187-217.

8. 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. *Journal of the American Chemical Society* **1996**, *118* (45), 11225-11236.

9. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* **1998**, *102* (18), 3586-3616.

10. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25* (9), 1157-1174.

11. Rahman, A., Correlations in the Motion of Atoms in Liquid Argon. *Physical Review* **1964**, *136* (2A), A405-A411.

12. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21* (6), 1087-1092.

13. Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23* (2), 187-199.

14. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314* (1), 141-151.

15. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99* (20), 12562.

16. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L., TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annual Review of Physical Chemistry* **2002**, *53* (1), 291-318.

17. Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P., Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **2013**, *497*, 643.
18. Chu, J.-W.; Voth, G. A., Allostery of actin filaments: Molecular dynamics simulations and coarse-grained analysis. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102* (37), 13111.
19. Bernèche, S.; Roux, B., Energetics of ion conduction through the K⁺ channel. *Nature* **2001**, *414* (6859), 73-77.
20. 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. *The Journal of Physical Chemistry B* **2007**, *111* (27), 7812-7824.
21. Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A., AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *The Journal of Physical Chemistry B* **2012**, *116* (29), 8494-8503.
22. Noid, W. G., Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics* **2013**, *139* (9), 090901.
23. Brannigan, G.; Lin, L. C. L.; Brown, F. L. H., Implicit solvent simulation models for biomembranes. *European Biophysics Journal* **2006**, *35* (2), 104-124.
24. Hoogerbrugge, P. J.; Koelman, J. M. V. A., Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics. *Europhysics Letters (EPL)* **1992**, *19* (3), 155-160.
25. Español, P.; Warren, P., Statistical Mechanics of Dissipative Particle Dynamics. *Europhysics Letters (EPL)* **1995**, *30* (4), 191-196.
26. Español, P.; Warren, P. B., Perspective: Dissipative particle dynamics. *The Journal of Chemical Physics* **2017**, *146* (15), 150901.
27. Colagrossi, A.; Landrini, M., Numerical simulation of interfacial flows by smoothed particle hydrodynamics. *Journal of Computational Physics* **2003**, *191* (2), 448-475.
28. Liu, M.; Meakin, P.; Huang, H., Dissipative particle dynamics simulation of multiphase fluid flow in microchannels and microchannel networks. *Physics of Fluids* **2007**, *19* (3), 033302.
29. Fredrickson, G. H.; Ganesan, V.; Drolet, F., Field-Theoretic Computer Simulation Methods for Polymers and Complex Fluids. *Macromolecules* **2002**, *35* (1), 16-39.

30. Liu, M. B.; Liu, G. R., Smoothed Particle Hydrodynamics (SPH): an Overview and Recent Developments. *Archives of Computational Methods in Engineering* **2010**, *17* (1), 25-76.
31. Warshel, A.; Levitt, M., Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology* **1976**, *103* (2), 227-249.
32. Warshel, A.; Kato, M.; Pislakov, A. V., Polarizable Force Fields: History, Test Cases, and Prospects. *Journal of Chemical Theory and Computation* **2007**, *3* (6), 2034-2045.
33. Baker, C. M., Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015**, *5* (2), 241-254.
34. van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A* **2001**, *105* (41), 9396-9409.
35. Knight, C.; Lindberg, G. E.; Voth, G. A., Multiscale reactive molecular dynamics. *The Journal of Chemical Physics* **2012**, *137* (22), 22A525.
36. Babin, V.; Medders, G. R.; Paesani, F., Toward a Universal Water Model: First Principles Simulations from the Dimer to the Liquid Phase. *The Journal of Physical Chemistry Letters* **2012**, *3* (24), 3765-3769.
37. Serrano, M.; Español, P., Thermodynamically consistent mesoscopic fluid particle model. *Physical Review E* **2001**, *64* (4), 046115.
38. Serrano, M.; Fabritiis, G. D.; Español, P.; Flekkøy, E. G.; Coveney, P. V., Mesoscopic dynamics of Voronoi fluid particles. *Journal of Physics A: Mathematical and General* **2002**, *35* (7), 1605-1625.
39. Kadanoff, L. P., Scaling laws for ising models near T_c . *Physics Physique Fizika* **1966**, *2* (6), 263-272.
40. Carbone, P.; Varzaneh, H. A. K.; Chen, X.; Müller-Plathe, F., Transferability of coarse-grained force fields: The polymer case. *The Journal of Chemical Physics* **2008**, *128* (6), 064904.
41. Wagner, J. W.; Dama, J. F.; Durumeric, A. E. P.; Voth, G. A., On the representability problem and the physical meaning of coarse-grained models. *The Journal of Chemical Physics* **2016**, *145* (4), 044108.
42. Sultan, M. M.; Pande, V. S., Automated design of collective variables using supervised machine learning. *The Journal of Chemical Physics* **2018**, *149* (9), 094106.

43. Han, Y.; Dama, J. F.; Voth, G. A., Mesoscopic coarse-grained representations of fluids rigorously derived from atomistic models. *The Journal of Chemical Physics* **2018**, *149* (4), 044104.
44. Hijón, C.; Español, P.; Vanden-Eijnden, E.; Delgado-Buscalioni, R., Mori–Zwanzig formalism as a practical computational tool. *Faraday Discussions* **2010**, *144* (0), 301-322.
45. Lyubartsev, A. P.; Laaksonen, A., Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Physical Review E* **1995**, *52* (4), 3730-3737.
46. Müller-Plathe, F., Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3* (9), 754-769.
47. Izvekov, S.; Voth, G. A., Multiscale coarse graining of liquid-state systems. *The Journal of Chemical Physics* **2005**, *123* (13), 134105.
48. Izvekov, S.; Voth, G. A., A Multiscale Coarse-Graining Method for Biomolecular Systems. *The Journal of Physical Chemistry B* **2005**, *109* (7), 2469-2473.
49. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A., Multiscale Coarse-Graining and Structural Correlations: Connections to Liquid-State Theory. *The Journal of Physical Chemistry B* **2007**, *111* (16), 4116-4127.
50. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C., The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics* **2008**, *128* (24), 244114.
51. Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A., The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *The Journal of Chemical Physics* **2008**, *128* (24), 244115.
52. Shell, M. S., The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics* **2008**, *129* (14), 144108.
53. Chaimovich, A.; Shell, M. S., Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics* **2011**, *134* (9), 094112.
54. Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A., The Theory of Ultra-Coarse-Graining. 1. General Principles. *Journal of Chemical Theory and Computation* **2013**, *9* (5), 2466-2480.
55. Dama, J. F.; Jin, J.; Voth, G. A., The Theory of Ultra-Coarse-Graining. 3. Coarse-Grained Sites with Rapid Local Equilibrium of Internal States. *Journal of Chemical Theory and Computation* **2017**, *13* (3), 1010-1022.

56. Jin, J.; Han, Y.; Voth, G. A., Ultra-Coarse-Grained Liquid State Models with Implicit Hydrogen Bonding. *Journal of Chemical Theory and Computation* **2018**, *14* (12), 6159-6174.
57. Jin, J.; Voth, G. A., Ultra-Coarse-Grained Models Allow for an Accurate and Transferable Treatment of Interfacial Systems. *Journal of Chemical Theory and Computation* **2018**, *14* (4), 2180-2197.
58. Wagner, J. W.; Dannenhoffer-Lafage, T.; Jin, J.; Voth, G. A., Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions. *The Journal of Chemical Physics* **2017**, *147* (4), 044113.
59. Jin, J.; Han, Y.; Voth, G. A., Coarse-graining involving virtual sites: Centers of symmetry coarse-graining. *The Journal of Chemical Physics* **2019**, *150* (15), 154103.

Chapter 2: Quantum Theory of Multiscale Coarse-graining

This chapter is reproduced from J. Chem. Phys. 148, 102335 (2018), with the permission of AIP Publishing.

Abstract

Coarse-grained (CG) models serve as a powerful tool to simulate molecular systems at much longer temporal and spatial scales. Previously, CG models and methods have been built upon classical statistical mechanics. The present paper develops a theory and numerical methodology for coarse-graining in quantum statistical mechanics, by generalizing the multiscale coarse-graining (MS-CG) method to quantum Boltzmann statistics. A rigorous derivation of the sufficient thermodynamic consistency condition is first presented via imaginary time Feynman path integrals. It identifies the optimal choice of CG action functional and effective quantum CG (qCG) force field to generate a quantum MS-CG (qMS-CG) description of the equilibrium system that is consistent with the quantum fine-grained model projected onto the CG variables. A variational principle then provides a class of algorithms for optimally approximating the qMS-CG force fields. Specifically, a variational method based on force matching, which was also adopted in the classical MS-CG theory, is generalized to quantum Boltzmann statistics. The qMS-CG numerical algorithms and practical issues in implementing this variational minimization procedure are also discussed. Then, two numerical examples are presented to demonstrate the method. Finally, as an alternative strategy, a quasi-classical approximation for the thermal density matrix expressed in the CG variables is derived. This approach provides an interesting physical picture for coarse-graining in quantum Boltzmann statistical mechanics in which the consistency with the quantum particle delocalization is obviously manifest, and it opens up an avenue for using path integral centroid-based effective classical force fields in a coarse-graining methodology.

I. Introduction

Coarse-grained (CG) modeling of complex chemical, material, and biomolecular systems has garnered extensive attention and development over the last fifteen years, due to its ability to expand the application of molecular simulation methodology to much larger spatial and time scales.^{1,2} The rigorous “bottom-up” construction of CG models that can faithfully reproduce the underlying fine-grained (FG) atomistic level information is a challenging scientific problem. One class of bottom-up approach builds the CG models directly from more detailed FG models through a focus on reproducing the many-body potential of mean force (PMF) of the CG variables (i.e., the mappings from the FG variables to define the CG sites or “beads”). By contrast, top-down approaches to CG modeling parameterize the CG models to reproduce certain physicochemical properties that can be observed at larger scales. The bottom-up approach is of interest due to its first principles inference from classical FG free energy sampling without the need for empirical and experimental data. Although relatively few bottom-up methods directly focus on finding the many-body PMF for the CG variables, a variety of bottom-up techniques are available, including Boltzmann Inversion (BI)³ and its iterative variation (iterative Boltzmann Inversion, IBI),^{4,5} inverse Monte Carlo (IMC),⁶ Multiscale Coarse-graining (MS-CG),⁷⁻¹¹ and Relative Entropy Minimization (REM).^{12,13} The BI, IBI and IMC methods aim to reproduce certain target correlation functions of the FG models through a numerical optimization of the CG effective potential energy function. The MS-CG and REM methods, on the other hand, are methods in which the CG variable many-body PMF is the focus, and it is determined variationally by minimizing target functionals, i.e., the residual force functional for MS-CG and the Kullback-Leibler divergence for REM. A detailed comparison of these two methods was given by Rudzinski and Noid.^{2,14} MS-CG and REM are both exact coarse-graining methods if

one has a perfect CG force field (the so-called “basis set”) that can capture all of the essential physics of the FG system that is manifest at the CG level. Of course, one never has such a perfect force field, so for a more approximate CG force field these two approaches will generally give somewhat different results, and each approach has its strengths and weaknesses.

The current theories and numerical algorithms are built upon the principles of classical statistical mechanics. On the other hand, natural law is dictated by quantum mechanics, and this fact must be addressed in any rigorous theory no matter how small the quantum effects might (or might not!) be for a given set of conditions. To the best of our knowledge, there has not been an effort to develop a rigorous and general bottom-up coarse-graining methodology for quantum Boltzmann statistical mechanics. It is therefore of theoretical interest and arguably of fundamental importance to develop such a quantum theory of coarse-graining. In this paper, we will present a bottom-up coarse-graining theory in quantum Boltzmann statistics based on the MS-CG methodology. For sake of simplicity, the quantum CG theory and associated MS-CG methodology will be simply abbreviated as qCG and qMS-CG hereafter.

It should first be noted, however, that a similar qCG theory for particles having Bose-Einstein or Fermi-Dirac statistics is not discussed in this paper, in part because Boltzmann statistics are typically adequate for the quantum behavior of the atomic nuclei in typical chemical, material, and biomolecular systems at ambient temperature and pressure. However, the problem of coarse-graining for Bose-Einstein and Fermi-Dirac quantum statistics is also a challenging and very interesting one, which deserves at least one entire paper devoted to the topic. We will attempt to do so in the future.

We also note that nuclear quantum effects can in fact be important for the accurate description of certain chemical and biomolecular systems, especially those involving light atoms

and/or at low temperature. Some examples of these effects are found in the structure and dynamics of molecular liquids,¹⁵⁻²⁰ vibrational spectra of liquids,²¹ the air/water interface,²² hydrogen bonding in aqueous^{23, 24} and biological processes,²⁵ enzyme catalysis,²⁵⁻²⁷ isotope effects,²⁸ glass formation,²⁹ proton transfer processes,³⁰⁻³² proton coupled electron transfer,³³ as well as low temperature systems like condensed phase helium,³⁴ and para-hydrogen.³⁵

Feynman's imaginary time path integral,^{36, 37} with its quantum-classical isomorphism,^{38, 39} provides an appealing tool for the study of such problems in which nuclear quantum effects are important at finite temperature. Several imaginary time path integral-based numerical methods have been developed for simulating systems in the quantum Boltzmann statistical ensemble, such as path integral Monte Carlo (PIMC),³⁹ path integral molecular dynamics (PIMD),⁴⁰ and Feynman path centroid methods.⁴¹ Alternatively, centroid molecular dynamics (CMD)^{42, 43} and the more approximate ring polymer molecular dynamics (RPMD)^{44, 45} can be utilized if one is also interested in the real time quantum dynamics of the system. (For an analysis of these latter two quantum dynamics approaches and their key differences, see Ref.⁴⁶) The quantum-classical isomorphism enables the study of many systems in the quantum Boltzmann statistical ensemble with a manageable computational cost, although these simulations can still be demanding. Therefore, a bottom-up CG methodology developed within the framework of quantum statistical mechanics can be a valuable tool for both simplifying this kind of quantum simulation to its most essential elements while also providing a vehicle for increased interpretive power and insight.

To combine bottom-up coarse-graining with quantum Boltzmann statistical mechanics, there are two at least possible approaches that one might take. The first approach works by coarse-graining the density matrices of the FG system onto quantum coarse-grained (qCG) density matrices. The other approach is to reduce the density matrices of the FG system into a quasi-

classical coarse-grained density distribution function, such that the quantum statistics are embedded into an effective classical CG model. We will discuss both approaches in this paper, with an emphasis on the former one, while outlining the latter in passing at the end of the discussion. We will demonstrate that all of the tools previously developed in classical MS-CG method can be naturally generalized in the first qMS-CG approach, and that the numerical implementation requires only a relatively modest modification of the classical counterpart. A discussion of the second approach and an approximation based on this second quantum coarse-graining concept are presented in Sec. V, to provide perhaps the simplest physical picture and to reveal more aspects of the quantum-classical correspondence in coarse-graining.

It is also worth noting that there have been several papers using a concept of combining coarse-graining with the imaginary time path integral, either to construct reductionist representations of quantum statistical mechanics, or to develop enhanced sampling algorithms. The CMD method, for example, adopts the concept of the imaginary time path integral centroid variable and constructs a theory of quantum statistics and dynamics based on that concept, while more recently Sinitskiy and Voth⁴⁷ proposed another reductionist representation of quantum Boltzmann statistical mechanics by coarse-graining imaginary time paths onto three CG sites or “beads”. Moreover, Peng *et al.*⁴⁸ and Nava *et al.*^{49, 50} utilized the CGing of imaginary time path integrals to enhance the sampling of classical systems.

The remaining sections of this paper are organized as follows: In Sec. II, we derive a sufficient thermodynamic consistency condition between an atomistic model and a CG model within quantum Boltzmann statistics for linear CG mapping operators, by coarse-graining the imaginary time paths of the atomistic system. We then define the relationship between the thermodynamically consistent qMS-CG action functional and the corresponding functional of the

FG system. We propose a variational principle that serves as the basis for developing algorithms for determining qMS-CG force fields. Next, in Sec. III the numerical algorithms and implementation details of the variational method, which is a natural generalization of the MS-CG method to quantum Boltzmann statistics, are discussed. Section IV presents two numerical examples, one being a simple 1D few-body model problem and the second being a liquid water system, both demonstrating the accuracy and limitations of the method. In Sec. V, we derive an alternative quasi-classical approximation for the CG density matrix elements as mentioned earlier. This approach provides a clearer physical interpretation of the effect of quantum Boltzmann statistics on bottom-up coarse-graining. Sec. VI provides concluding remarks.

II. Theory

A. Path Integral Coarse-graining

Coarse-graining of a molecular system involves a statistical dimension reduction process.¹⁻⁶ A detailed molecular system with high resolution (e.g., an atomistic model or a highly resolved CG model such as a united atom model) is mapped onto a lower resolution model with much less number of degrees of freedom. Therefore, the first step is to define the high-resolution FG model. Generally, the dynamics and statistical information of a n particle FG model is defined by the Hamiltonian:

$$\hat{H} = \hat{T} + \hat{V} = \sum_{i=1}^{3n} \frac{\hat{p}_i^2}{2m_i} + \hat{V}(\mathbf{r}^n), \quad (2.1)$$

With the definition of Hamiltonian, the normalized thermal density matrix in quantum Boltzmann statistics can be expressed in the coordinate representation of the imaginary time Feynman path integral as^{36, 37}

$$\hat{\rho}_{\mathbf{r}\mathbf{r}'} = \frac{1}{Z} \langle \mathbf{r} | e^{-\beta H} | \mathbf{r}' \rangle = \frac{1}{Z} \int_{\mathbf{r}}^{\mathbf{r}'} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)] / \hbar) \\ \propto \lim_{P \rightarrow \infty} \int d\mathbf{r}_1 \dots d\mathbf{r}_{P-1} \exp \left\{ -\beta \sum_{j=1}^P \left[\sum_{i=1}^n \frac{m_i P}{2\beta^2 \hbar^2} (\mathbf{r}_j^i - \mathbf{r}_{j-1}^i)^2 + \frac{V(\mathbf{r}_j^n) + V(\mathbf{r}_{j-1}^n)}{2P} \right] \right\} \bigg|_{\mathbf{r}_0=\mathbf{r}}^{\mathbf{r}_P=\mathbf{r}'}, \quad (2.2)$$

where the imaginary time action $S[\mathbf{r}(\tau)]$ is defined as:

$$S[\mathbf{r}(\tau)] = \int_0^{\beta\hbar} d\tau \frac{1}{2} \dot{\mathbf{r}}^T \cdot \tilde{\mathbf{m}} \cdot \dot{\mathbf{r}} + V[\mathbf{r}(\tau)], \quad (2.3)$$

and $\tilde{\mathbf{m}}$ is a diagonal matrix representing the masses of the FG particles. All the thermodynamic properties of the system, e.g., the equilibrium expectation value of operators, can be obtained once the density matrix is calculated. Moreover, the quantum statistical ensemble of the FG system based on this density matrix can be numerically generated through imaginary time PIMD or PIMC simulations.^{38, 39}

Similarly, an effective Hamiltonian can be defined for the CG model describing the same system, but with fewer degrees of freedom, of the number $3N$, such that

$$\hat{H}_{CG} = \hat{T}_{CG} + \hat{V}_{CG} = \sum_{I=1}^{3N} \frac{\hat{P}_I^2}{2M_I} + \hat{V}_{CG}(\mathbf{R}^N), \quad (2.4)$$

Moreover, the equilibrium density matrix associated with this effective Hamiltonian can be assumed to be written as:

$$\hat{\rho}_{\mathbf{R}\mathbf{R}'} = \frac{1}{Z} \langle \mathbf{R} | e^{-\beta H_{CG}} | \mathbf{R}' \rangle = \frac{1}{Z} \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)] / \hbar) \\ \propto \lim_{P \rightarrow \infty} \int d\mathbf{R}_1 \dots d\mathbf{R}_{P-1} \exp \left\{ -\beta \sum_{i=1}^P \left[\sum_{I=1}^N \frac{M_I P}{2\beta^2 \hbar^2} (\mathbf{R}_i^I - \mathbf{R}_{i-1}^I)^2 + \frac{V_{CG}(\mathbf{R}_i^N) + V_{CG}(\mathbf{R}_{i-1}^N)}{2P} \right] \right\} \bigg|_{\mathbf{R}_0=\mathbf{R}}^{\mathbf{R}_P=\mathbf{R}'}. \quad (2.5)$$

Similarly, the CG action is defined by:

$$S_{CG}[\mathbf{R}(\tau)] = \int_0^{\beta\hbar} d\tau \frac{1}{2} \dot{\mathbf{R}}^T \cdot \tilde{\mathbf{M}} \cdot \dot{\mathbf{R}} + V_{CG}[\mathbf{R}(\tau)]. \quad (2.6)$$

In the above definitions, the CG representation of the model has not yet been connected with the FG description. This connection is usually derived based on the idea that each CG coordinate has a well-defined physical meaning in terms of the coordinates of the FG model, i.e., the configuration of CG system can be uniquely determined from the underlying FG model as in the classical MS-CG theory.⁷⁻¹¹ This mapping can be achieved by defining a linear CG mapping operator:

$$\mathbf{M}_{RI}(\mathbf{r}^n) = \sum_{i=1}^n c_{Ii} \mathbf{r}_i \quad \text{for } I = 1, \dots, N. \quad (2.7)$$

The mapping matrix elements c_{Ii} are scalars satisfying the condition $\sum_i c_{Ii} = 1$ due to Galilean invariance and rotational invariance of the FG to CG correspondence. Through the CG mapping operator, the physical meaning of the classical CG site or “beads” can be related to a set of linear collective variables (CV) involving the FG model coordinates.⁷⁻¹¹

It should be noted that there are many possible choices of CG mappings. The most commonly used one might be the center-of-mass (COM) mapping, which maps a group of FG particles onto their COM for each CG site. Other choices, for example, are the center-of-geometry mapping that maps a group of FG particles onto their geometric center and the center-of-charge mapping.⁵¹ There are also CG mappings based on normal modes,⁵² principal component analysis,⁵³ and diffusion maps.⁵⁴

B. Consistency Condition for the CG Path Integral Action

When developing bottom-up CG models, one desires the CG model to behave as if it is the same as the FG model but at a lower resolution. In order to maintain this consistency condition, an explicit relationship between the FG Hamiltonian and the CG Hamiltonian must be derived. In this section, it will be demonstrated that the action functional for FG system and the CG

mapping operator can determine the effective action of the CG path integral model.

The thermodynamic consistency condition in quantum statistical mechanics requires the expectation value of a given operator to be the same in both the FG and CG statistical ensemble, i.e.,

$$\begin{aligned}
\langle \hat{O} \rangle &= \text{Tr}(\hat{\rho} \hat{O}) = \int d\mathbf{r} d\mathbf{r}' \langle \mathbf{r} | \hat{\rho} | \mathbf{r}' \rangle \langle \mathbf{r}' | \hat{O} | \mathbf{r} \rangle \\
&= \int d\mathbf{R} d\mathbf{R}' \int d\mathbf{r} d\mathbf{r}' \langle \mathbf{r} | \hat{\rho} | \mathbf{r}' \rangle \langle \mathbf{r}' | \mathbf{R}' \rangle \langle \mathbf{R}' | \hat{O} | \mathbf{R} \rangle \langle \mathbf{R} | \mathbf{r} \rangle \\
&= \int d\mathbf{R} d\mathbf{R}' \int d\mathbf{r} d\mathbf{r}' \langle \mathbf{r} | \hat{\rho} | \mathbf{r}' \rangle \langle \mathbf{R}' | \hat{O} | \mathbf{R} \rangle \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R}') \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})', \\
&= \int d\mathbf{R} d\mathbf{R}' \langle \mathbf{R} | \hat{\rho}_{CG} | \mathbf{R}' \rangle \langle \mathbf{R}' | \hat{O} | \mathbf{R} \rangle
\end{aligned} \tag{2.8a}$$

where M denotes the CG mapping operator from Eq. (2.7) and the delta-functions are understood to be product of delta-functions for all CG mappings $M_{\mathbf{R}_I}(\mathbf{r}^n)$, i.e.,

$$\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \equiv \prod_I \delta(\mathbf{M}_{\mathbf{R}_I}(\mathbf{r}) - \mathbf{R}_I). \tag{2.8b}$$

By comparing the last two lines in the Eq. (2.8a), the density matrix elements for the qCG model can be written as:

$$\begin{aligned}
\hat{\rho}_{\mathbf{R}\mathbf{R}'}^{CG} &\equiv \int d\mathbf{r} \int d\mathbf{r}' \hat{\rho}_{\mathbf{r}\mathbf{r}'} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R}') \\
&\propto \lim_{P \rightarrow \infty} \int d\mathbf{r} d\mathbf{r}' d\mathbf{r}_1 \dots d\mathbf{r}_{P-1} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R}') \\
&\quad \times \exp \left\{ -\beta \sum_{j=1}^P \left[\sum_{i=1}^n \frac{m_i P}{2\beta^2 \hbar^2} (\mathbf{r}_j^i - \mathbf{r}_{j-1}^i)^2 + \frac{V(\mathbf{r}_j^n) + V(\mathbf{r}_{j-1}^n)}{2P} \right] \right\} \bigg|_{\mathbf{r}_0=\mathbf{r}}^{\mathbf{r}_P=\mathbf{r}'}
\end{aligned} \tag{2.9}$$

At this juncture, there are two possible routes that can be taken to formulate a coarse-graining strategy for equilibrium quantum systems. The first and simplest, which we will explore more in Sec. V, is to simply view the FG quantum density matrix in the “interior” of the above equations as the FG mathematical object (distribution function) that gets mapped into a CG one, in order to provide a new quantum equilibrium density matrix in the CG variables. The differences between the classical and quantum FG to CG mappings in this case are then: (1) due to quantum

delocalization at the FG level the CG density matrix manifests these effects which are not seen at either the classical FG or CG levels, and (2) the new (mapped) CG density matrix is nonlocal in the CG variables $\mathbf{R}$ (just as the FG one is nonlocal in $\mathbf{r}$). Unless further approximations and analyses are made (see Sec. V), this approach seems less satisfying and rather obvious. We also note that unless the observable operator $\hat{O}$ depends only on the CG variables $\mathbf{R}$, then there is also a representability problem⁵⁵ for the matrix element $\langle \mathbf{R}' | \hat{O} | \mathbf{R} \rangle$ in Eq. (2.8). By this it is meant that one cannot simply assume its mathematical form is the same as the FG observable form.

On the other hand, a different and deeper route to a coarse-graining strategy for equilibrium quantum systems *is to coarse-grain the system at every point along the imaginary time Feynman path*. To accomplish this, we can utilize the property of the Dirac delta function that $\int \delta(x - c) dx = 1$ in Eq. (2.9) at each imaginary time slice to enforce the CG mapping operator along the whole path. It can then be shown that Eq. (2.9) can be expressed as an interesting *double path integral* over all possible FG density matrices, such that

$$\begin{aligned} \hat{\rho}_{\mathbf{R}\mathbf{R}'}^{CG} &\propto \lim_{P \rightarrow \infty} \int d\mathbf{R}_1 \dots d\mathbf{R}_{P-1} \int d\mathbf{r} d\mathbf{r}_1 \dots d\mathbf{r}_{P-1} d\mathbf{r}' \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \delta(\mathbf{M}(\mathbf{r}') - \mathbf{R}') \\ &\times \prod_{i=1}^{P-1} \delta(\mathbf{M}(\mathbf{r}_i) - \mathbf{R}_i) \exp \left\{ -\beta \sum_{j=1}^P \left[\sum_{i=1}^n \frac{m_i P}{2\beta^2 \hbar^2} (\mathbf{r}_j^i - \mathbf{r}_{j-1}^i)^2 + \frac{V(\mathbf{r}_j^n) + V(\mathbf{r}_{j-1}^n)}{2P} \right] \right\} \bigg|_{\mathbf{r}_0=\mathbf{r}}^{\mathbf{r}_P=\mathbf{r}'} , \quad (2.10a) \\ &= \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \int d\mathbf{r} \int d\mathbf{r}' \int_{\mathbf{r}}^{\mathbf{r}'} D\mathbf{r}(\tau) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau)) \exp(-S[\mathbf{r}(\tau)]/\hbar) \end{aligned}$$

in which the Dirac delta function in the last line is a shorthand notation for the product of delta-functions at each imaginary time slice, i.e.

$$\delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau)) \equiv \lim_{P \rightarrow \infty} \prod_{i=1}^{P-1} \delta(\mathbf{M}(\mathbf{r}_i) - \mathbf{R}_i), \quad (2.10b)$$

On the other hand, as shown in Eq. (2.5), the CG density matrix element should also be able to be written simply as a Feynman path integral of CG paths:

$$\hat{\rho}_{\mathbf{R}\mathbf{R}'}^{CG} \propto \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar). \quad (2.11)$$

Comparing Eqs. (2.10) and (11), it can be seen that the sufficient condition (but not necessary) for the consistency of the FG and CG systems in that quantum Boltzmann statistical ensemble for all imaginary times can be expressed as:

$$\exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \equiv \int d\mathbf{r} \int d\mathbf{r}' \int_{\mathbf{r}}^{\mathbf{r}'} D\mathbf{r}(\tau) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau)) \exp(S[\mathbf{r}(\tau)]/\hbar). \quad (2.12)$$

Equation (2.12) is interpreted as the consistency condition for the CG action defined in Eq. (2.6). It implies that the action of the CG system is completely determined by the FG action along with the CG mapping operator. It should be noted that the CG effective action functional defined here is local in imaginary time as enforced by the delta functions in the FG to CG mappings, unlike the imaginary time influence functional³⁶ when unconstrained “bath” degrees of freedom are integrated out of the problem.

C. Consistency Condition for the Quantum CG Force Field

With the consistency condition for the action, deriving the relationship between the CG force field and FG force field can be accomplished. In order to do so, we first take the natural logarithm on both sides of Eq. (2.12):

$$S_{CG}[\mathbf{R}(\tau)] = -\hbar \ln \int d\mathbf{r} \int d\mathbf{r}' \int_{\mathbf{r}}^{\mathbf{r}'} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau)). \quad (2.13)$$

Then, we take the functional derivative with respect to the imaginary time paths of the I^{th} CG site:

$$\frac{\delta S_{CG}[\mathbf{R}(\tau)]}{\delta \mathbf{R}_I(\tau)} = \frac{\int d\mathbf{r} \int d\mathbf{r}' \int_r^{r''} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \prod_{j \neq I} \delta(\mathbf{M}_{R_j}[\mathbf{r}(\tau)] - \mathbf{R}_j(\tau)) \frac{\delta}{\delta \mathbf{R}_I(\tau)} \delta(\mathbf{M}_{R_I}[\mathbf{r}(\tau)] - \mathbf{R}_I(\tau))}{-\hbar \int d\mathbf{r} \int d\mathbf{r}' \int_r^{r''} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau))}. \quad (2.14)$$

By applying the Euler-Lagrange equation, the left hand side of Eq. (2.14) can further be simplified to give a formula for the qCG effective force field:

$$\frac{\delta S[\mathbf{R}(\tau)]}{\delta \mathbf{R}_I(\tau)} = \frac{\partial V_{CG}[\mathbf{R}(\tau)]}{\partial \mathbf{R}_I(\tau)} - M_I \ddot{\mathbf{R}}_I(\tau) = -\mathbf{F}_I[\mathbf{R}(\tau)] - M_I \ddot{\mathbf{R}}_I(\tau). \quad (2.15)$$

The double dot here denotes the second derivative with respect to imaginary time τ . At the same time, the right hand side of Eq. (2.14) can be simplified by applying the following property:

$$\frac{\delta}{\delta \mathbf{R}_I(\tau)} \delta(\mathbf{M}_{R_I}[\mathbf{r}(\tau)] - \mathbf{R}_I(\tau)) = -\frac{1}{c_{Ik}} \frac{\delta}{\delta \mathbf{r}_k(\tau)} \delta(\mathbf{M}_{R_I}[\mathbf{r}(\tau)] - \mathbf{R}_I(\tau)). \quad (2.16)$$

For any $k \in \Sigma_I$, meaning any atom involved in the definition of the I^{th} CG site, this can easily be shown from Eq. (2.7). Although Eq. (2.16) can be directly employed in Eq. (2.14), the subsequent derivation would be complicated if any FG atom k is involved in the definition of other CG sites. In order to alleviate this complication, similar to the classical MS-CG theory,¹⁰ we can define the set of constant coefficients $\{d_{Ii}\}$ such that $d_{Ii} \neq 0$ only if atom i is specific to CG site I and such that:

$$\sum_{i \in \Sigma_I} d_{Ii} = 1. \quad (2.17)$$

It can then be verified that:

$$\frac{\delta}{\delta \mathbf{R}_I(\tau)} \delta(\mathbf{M}_{R_I}[\mathbf{r}(\tau)] - \mathbf{R}_I(\tau)) = -\sum_{j \in \Sigma_I} \frac{d_{Ij}}{c_{Ij}} \frac{\delta}{\delta \mathbf{r}_j(\tau)} \delta(\mathbf{M}_{R_I}[\mathbf{r}(\tau)] - \mathbf{R}_I(\tau)). \quad (2.18)$$

Inserting Eq. (2.18) in Eq. (2.14) and integrating by parts, we arrive at

$$F_I[\mathbf{R}(\tau)] + M_I \ddot{\mathbf{R}}_I(\tau) = \sum_{j \in S_I} \frac{d_{Ij}}{c_{Ij}} \left\langle f_j[\mathbf{r}(\tau)] + m_j \ddot{\mathbf{r}}_j(\tau) \right\rangle_{\mathbf{R}(\tau)} \equiv \left\langle \mathbf{M}_I^f \left(f[\mathbf{r}(\tau)] + m \ddot{\mathbf{r}}(\tau) \right) \right\rangle_{\mathbf{R}(\tau)}, \quad (2.19)$$

where the M_I^f is a force mapping operator, given by

$$\mathbf{M}_I^f g[\mathbf{r}^N(\tau)] = \sum_{j \in \Sigma_I} \frac{d_{Ij}}{c_{Ij}} g[\mathbf{r}_j(\tau)], \quad (2.20)$$

and the angular brackets denote a constrained ensemble average of the form:

$$\left\langle g[\mathbf{r}(\tau)] \right\rangle_{\mathbf{R}(\tau)} = \frac{\int d\mathbf{r} \int d\mathbf{r}' \int_r^{r'} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau)) g[\mathbf{r}(\tau)]}{\int d\mathbf{r} \int d\mathbf{r}' \int_r^{r'} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \delta(\mathbf{M}[\mathbf{r}(\tau)] - \mathbf{R}(\tau))}. \quad (2.21)$$

Equation (2.19) is the consistency condition for the qCG force field. It is derived from the consistency condition for the CG action by assuming that there is at least one FG particle that is specific to a given CG site. This formula also provides the basis for proposing the algorithm for determining the effective CG Hamiltonian as shown below in the next section.

In principle, the qCG force field defined by Eq. (2.19) can be obtained by a FG PIMC or PIMD simulation that applies the constraints of CG mapping operator. However, it is computationally forbidding due to the complexity from the treatment of the constraints. However, in the MS-CG theory for classical statistical mechanics, it has been shown that the variational method based on force matching equation can be successfully applied to determine the CG force field. In this work as well, the force matching equation can also be applied after the necessary modifications.

D. Variational Method

In the previous section, the conditions for thermodynamic consistency are expressed either in terms of path integral actions or forces. This section will describe the variational approach to inferring the qCG model from the FG sampling, base on the consistency condition of forces.

Generally, the qCG force field is a 3N-dimensional continuous function $\mathbf{G}_I(\mathbf{R}^N)$. Mathematically, the goal of inferring the CG force field is to find the particular function $\mathbf{F}_I(\mathbf{R}^N)$ that satisfies the consistency condition from a large vector space of 3N-dimensional functions. Variational methods build on the idea that this procedure can be achieved by minimizing the statistical distance between any such function $\mathbf{G}_I(\mathbf{R}^N)$ in the functional space to the function $\mathbf{F}_I(\mathbf{R}^N)$. This distance can be characterized in different ways, but a very natural choice is based on the force matching equation.¹⁰

For arbitrary candidate $\mathbf{G}$ in the vector space, the residual force functional in quantum Boltzmann statistics is defined by:

$$\chi^2[\mathbf{G}] = \frac{1}{3N} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)] / \hbar) \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \left(\mathbf{G}_I[\mathbf{R}(\tau)] - \mathbf{F}_I[\mathbf{R}(\tau)] \right) \right|^2, \quad (2.22)$$

where the definition of $\mathbf{F}_I[\mathbf{R}(\tau)]$ was shown in Eq. (2.19). This residual functional measures the ensemble average of the difference between the proposed trial force field and the thermodynamically consistent one, in qCG space. The idea of the qMS-CG algorithm is to minimize the residual force functional in CG functional space to obtain the best approximation of the CG potential.

The above residual functional is written in the qCG phase space, but it can easily be converted to the entire FG phase space. This is achieved by first expanding the expression as follows:

$$\begin{aligned}
\chi^2[\mathbf{G}] &= \frac{1}{3N(\beta\hbar)^2} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}'}^{\mathbf{R}} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \\
&\quad \times \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \left(\mathbf{G}_I[\mathbf{R}(s)] - F_I[\mathbf{R}(s)] \right) \cdot \left(\mathbf{G}_I[\mathbf{R}(s')] - F_I[\mathbf{R}(s')] \right) \\
&= \frac{1}{3N(\beta\hbar)^2} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}'}^{\mathbf{R}} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \\
&\quad \times \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \left(\mathbf{G}_I[\mathbf{R}(s)]\mathbf{G}_I[\mathbf{R}(s')] - 2\mathbf{G}_I[\mathbf{R}(s)]F_I[\mathbf{R}(s')] + F_I[\mathbf{R}(s)]F_I[\mathbf{R}(s')] \right)
\end{aligned} \tag{2.23}$$

Note that the last term of the above integration involves a constrained ensemble average with respect to CG paths. It can also be shown that the following expression holds:

$$\begin{aligned}
\chi^2[\mathbf{G}] &= \frac{1}{3N(\beta\hbar)^2} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}'}^{\mathbf{R}} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \\
&\quad \times \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \left\langle \mathbf{G}_I[\mathbf{R}(s)]\mathbf{G}_I[\mathbf{R}(s')] \right\rangle_{\mathbf{R}(\tau)} - 2 \left\langle \mathbf{G}_I[\mathbf{R}(s)] \right\rangle_{\mathbf{R}(\tau)} \left\langle \mathbf{M}_I^f(f[\mathbf{r}(s')] + m\ddot{\mathbf{r}}(s')) - M_I\ddot{\mathbf{R}}_I(s') \right\rangle_{\mathbf{R}(\tau)} \\
&\quad + \left\langle \mathbf{M}_I^f(f[\mathbf{r}(s)] + m\ddot{\mathbf{r}}(s)) - M_I\ddot{\mathbf{R}}_I(s) \right\rangle_{\mathbf{R}(\tau)} \cdot \left\langle \mathbf{M}_I^f(f[\mathbf{r}(s')] + m\ddot{\mathbf{r}}(s')) - M_I\ddot{\mathbf{R}}_I(s') \right\rangle_{\mathbf{R}(\tau)} \\
&= c + \frac{1}{3N(\beta\hbar)^2} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}'}^{\mathbf{R}} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \\
&\quad \times \left\langle \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \left(\mathbf{G}_I[\mathbf{R}(s)] + M_I\ddot{\mathbf{R}}_I(s) - \mathbf{M}_I^f(f[\mathbf{r}(s)] + m\ddot{\mathbf{r}}(s)) \right) \cdot \left(\mathbf{G}_I[\mathbf{R}(s')] + M_I\ddot{\mathbf{R}}_I(s') - \mathbf{M}_I^f(f[\mathbf{r}(s')] + m\ddot{\mathbf{r}}(s')) \right) \right\rangle_{\mathbf{R}(\tau)}
\end{aligned} \tag{2.24}$$

The constant c comes from switching the order of product and constraint ensemble average for the last term in the integration, but it will not affect the further results since it will vanish when taking derivatives. And it should also be noted that the constant c does not depend on the choice of $\mathbf{G}$, which will prove to be useful in the subsequent derivations.

Inserting the CG force field consistency condition, switching the order of integration variables, and applying the integration property of the Dirac delta function, one can derive the following expression:

$$\begin{aligned}
\chi^2[\mathbf{G}] &= c + \frac{1}{3N} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)]/\hbar) \left\langle \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \left(\mathbf{G}_I[\mathbf{R}(\tau)] + M_I \ddot{\mathbf{R}}_I(\tau) - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right) \right|^2 \right\rangle_{\mathbf{R}(\tau)} \\
&= c + \frac{1}{3N} \sum_I \int d\mathbf{r} \int d\mathbf{r}' \int_{\mathbf{r}}^{\mathbf{r}'} D\mathbf{r}(\tau) \exp(-S[\mathbf{r}(\tau)]/\hbar) \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \left(\mathbf{G}_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(\tau)]) + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\ddot{\mathbf{r}}(\tau)] - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right) \right|^2 \\
&= c + \frac{1}{3N} \sum_I \left\langle \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \left(\mathbf{G}_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(\tau)]) + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\ddot{\mathbf{r}}(\tau)] - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right) \right|^2 \right\rangle_{FG}
\end{aligned} \tag{2.25}$$

This expression explicitly shows that the measure of how a test force field differs from the thermodynamic consistent force field can be directly evaluated from a simple FG sampling path integral trajectory. Then, defining the error of the test force field as $\Delta = \mathbf{G} - \mathbf{F}$, it follows that

$$\begin{aligned}
\chi^2[\mathbf{G}] &= \chi^2[\mathbf{F}] + \frac{1}{3N} \sum_I \left\langle \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \Delta_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(\tau)]) \right|^2 \right\rangle_{FG} \\
&\quad + \frac{2}{3N(\beta\hbar)^2} \sum_I \left\langle \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \Delta_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(s)]) \cdot \left(\mathbf{F}_I[\mathbf{M}_{\mathbf{R}_I}(\mathbf{r}(s'))] + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\ddot{\mathbf{r}}(s')] - \mathbf{M}_I^f(f[\mathbf{r}(s')] + m\ddot{\mathbf{r}}(s')) \right) \right\rangle_{FG} \\
&= \chi^2[\mathbf{F}] + \frac{1}{3N} \sum_I \left\langle \left| \frac{1}{\beta\hbar} \int_0^{\beta\hbar} d\tau \Delta_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(\tau)]) \right|^2 \right\rangle_{FG} \geq \chi^2[\mathbf{F}]
\end{aligned} \tag{2.26}$$

The reason why the last term in the first equation can be discarded is shown in Appendix A.

The above derivation provides the theoretical basis for the following variational principle: The global minimum of the functional $\chi^2[\mathbf{G}]$ is achieved when $\mathbf{G}$ is equal to $\mathbf{F}$, i.e., for the CG force field that satisfies the consistency conditions, and the global minimum is unique. The numerical realization of this variational optimization is presented in Sec. III.

It is worth noting that the variational principle could also be built on other metrics, such as the relative entropy. The relative entropy in classical statistical mechanics has been applied as the basis of the variational REM algorithm by Shell^{12, 13} and demonstrated to be a rigorous and powerful method for building biomolecular and condensed phase CG models. Similarly, the concept of relative entropy can be generalized in quantum statistical mechanics as⁵⁶

$$S(\hat{\rho}_{CG} \parallel \hat{\rho}) = \text{Tr} \hat{\rho}_{CG} \ln \hat{\rho}_{CG} - \text{Tr} \hat{\rho}_{CG} \ln \hat{\rho}. \tag{2.27}$$

A variational principle based on the quantum relative entropy may provide an alternative route to the work presented here, and such a formulation will be developed in the future.

E. Discussion

In previous sections, we have demonstrated the rigorous thermodynamic consistency conditions as well as the variational principle to derive the qCG Hamiltonian from a given FG model. In this section, we will discuss two special cases that can simplify the theory and numerical calculations. These two simplifications are then employed in the description of algorithms and numerical examples in the following sections.

1. Consistency in configuration space only

In the previous derivation, the theory builds on the correspondence of the CG and FG density matrix elements, either the diagonal or off-diagonal ones. Therefore, all of the ensemble averages should be sampled from open path integrals, because the starting and end point of the imaginary time path integral can be different.

In many cases in practice, one is only interested in the configurational information of the system, such as radial distribution functions, structure factors, dihedral distribution function, and so on. For these properties, the expectation value can be simplified and only the diagonal elements of the density matrix are needed:

$$\begin{aligned}\langle \hat{O} \rangle &= \int d\mathbf{r} d\mathbf{r}' \rho_{\mathbf{r}\mathbf{r}'} \langle \mathbf{r}' | \hat{O} | \mathbf{r} \rangle = \int d\mathbf{r} d\mathbf{r}' \rho_{\mathbf{r}\mathbf{r}'} O(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}') \\ &= \int d\mathbf{r} \rho_{\mathbf{r}\mathbf{r}} O(\mathbf{r})\end{aligned}\quad (2.28)$$

Therefore, the consistency conditions can be derived exclusively for the diagonal elements of the density matrix. It can be shown that all of the previous derivations can be specialized in this case, so that all of the path integrals are closed given that the starting and end point of the path

integrals are identical. In this special case, closed loop PIMD and PIMC simulations can be utilized to sample the FG system.

2. Choice of CG Masses

In Eq. (2.25), if the mass of the CG particle satisfy the following condition:

$$M_I = m_i \frac{d_{Ii}^2}{c_{Ii}^2} \quad \text{for } i \in \Sigma_I. \quad (2.29)$$

It can be shown that the two terms involving second order derivatives will cancel. In this situation, the residual force functional only depends on the difference between the CG and FG forces, greatly simplifying the numerical calculation. Specifically, for the most common COM CG mapping operator in which the mapping coefficients satisfy $d_{Ii} = c_{Ii}$ and $c_{Ii} = m_i / M_I$, we see that the condition in Eq. (2.29) can be verified if the mass of the CG site is defined by

$$M_I = \sum_{i \in \Sigma_I} m_i, \quad (2.30)$$

i.e., the mass of the CG site equals to the overall mass of the FG particles specific to that CG site.

III. Algorithms

In the previous section, it was demonstrated that minimizing the following force residual functional would lead us to the best approximation to the optimally consistent qCG force field:

$$\chi^2[\mathbf{G}] = \frac{1}{3N} \sum_I \int d\mathbf{R} \int d\mathbf{R}' \int_{\mathbf{R}}^{\mathbf{R}'} D\mathbf{R}(\tau) \exp(-S_{CG}[\mathbf{R}(\tau)] / \hbar) \left| \frac{1}{\beta \hbar} \int_0^{\beta \hbar} d\tau \left(\mathbf{G}_I[\mathbf{R}(\tau)] - F_I[\mathbf{R}(\tau)] \right) \right|^2. \quad (2.31)$$

This equation defines a variational optimization problem in infinite dimension that is challenging for any practical implementation. The first step of a numerical implementation is to construct the basis functions that define the subspace for use in constructing approximations to the CG force field.¹¹ Such basis functions are usually determined with the help of physical intuition about the

system of interest. Many of these functions can be expressed as functions of collective variables in the system, such as the pairwise distances, bond lengths, angles, dihedrals of bonded molecules, and three-body non-bonded angles etc. In general, the force field of the CG system can be expressed as a combination of many such basis functions:

$$F_I(\mathbf{R}^N) = \sum_{\gamma} \sum_{\zeta} F(x_{\zeta}\{\mathbf{R}^N\}_{\gamma}) \frac{\partial x_{\zeta}}{\partial \mathbf{R}_I}. \quad (2.32)$$

Here, ζ denotes the type of interactions in the CG model that one intends to include, such as the pairwise distances between particles. The index γ represents a specific set of particles that interact with particle I of type ζ . In principle, the basis function $F(x_{\zeta}(R^N))$ can take any form, and it can be challenging to guess any concrete mathematical expression by simple physical intuition alone. In the actual MS-CG implementation, these basis functions are often approximated by spline functions:

$$\begin{aligned} F_I(\mathbf{R}^N) &= \sum_{\zeta} \sum_d \sum_{\gamma} \phi_{\zeta d} u_d(x_{\zeta}\{\mathbf{R}^N\}_{\gamma}) \frac{\partial x_{\zeta}}{\partial \mathbf{R}_I}, \\ &= \sum_D \phi_D \Theta_{ID}(\mathbf{R}^N) \end{aligned} \quad (2.33)$$

where the indices were contracted here to make a compact form. These contracted basis function can be written as:

$$\Theta_{ID} \equiv \Theta_{I,\zeta d}(R^N) = \sum_{\gamma} u_d(x_{\zeta}\{\mathbf{R}^N\}_{\gamma}) \frac{\partial x_{\zeta}}{\partial \mathbf{R}_I}. \quad (2.34)$$

This representation of CG force fields explicitly demonstrates that the CG force field linearly depends on the set of coefficients ϕ_D to be determined variationally. The basis set $u_d(x_{\zeta})$ can either be implemented as cubic splines, linear splines. or B-Spline basis sets. In the present work, a B-Spline basis set was chosen because it can reduce the computational time and memory usage

while retaining the accuracy and smoothness of the fitted force fields. A detailed discussion on the choice of basis set can be found in Refs.^{11, 57}

After defining the basis functions for approximating the qCG potential, the variational minimization problem can be cast as a standard linear regression problem. This can be shown by explicitly expanding the residual functional in Eq. (2.25). The FG ensemble average is approximated by the time average of a FG PIMD or PIMC simulation trajectory. For a practical implementation, assuming the COM CG mapping is employed and a finite number (P) of ring polymer “beads” from the discretization of the Feynman paths are used in the path integral simulations, the residual functional can then be expressed as:

$$\chi^2[\phi] = \frac{1}{3Nn_t} \sum_{t=1}^{n_t} \sum_{I=1}^N \frac{1}{P^2} \left| \sum_{j=1}^P \sum_{D=1}^{N_D} \phi_D \Theta_{ID}(\mathbf{R}_j^N) - \sum_{j=1}^P f_I^j \right|^2. \quad (2.35)$$

Here, n_t is the total number of trajectory frames used to calculate the ensemble average, f_I is the total force on the FG particles involved in defining CG site I , the superscript j denotes the j^{th} bead of the discretized path integral, and $\mathbf{R}_j^N$ denotes the configuration of the qCG system at the j^{th} imaginary time slices (i.e., the j^{th} bead). Then, the spline coefficients $\{\phi_D\}$, and consequently the qCG force field, can be obtained by solving the above least square problem.

Several different numerical methods exist for solving the least squares problem. In this work, the normal equation approach is employed. To show how this method works in detail, Eq. (2.35) is first reformulated in a matrix notation:

$$\chi^2[\phi] = \frac{1}{3Nn_t} (\mathbf{X}\phi - \mathbf{f})^T (\mathbf{X}\phi - \mathbf{f}). \quad (2.36)$$

The $\mathbf{f}$ matrix denotes a column vector of $3n_tN$ elements (for problems in d dimension, this number should become dn_tN), and each component is a Cartesian component of the force

estimator of CG particle I , in a particular atomistic trajectory frame mapped onto CG representation:

$$f_{I,x} = \frac{1}{P} \sum_{j=1}^P F_I(\mathbf{R}_j^N) = \frac{1}{P} \sum_{j=1}^P M_I^f(\mathbf{r}_j^n). \quad (2.37)$$

In Eq. (2.36), $\mathbf{X}$ is a $3n_t N$ by N_D matrix, in which N_D represents the number of spline basis functions used to approximate the qCG potential energy profile. For each column in the $\mathbf{X}$ matrix, the vector X_D is a column vector of $3n_t N$ elements similar to $\mathbf{f}$, each element represents a Cartesian component of the estimator of the D^{th} basis function projected on CG site I , in a particular atomistic trajectory frame mapped onto the CG representation. The matrix element of the $\mathbf{X}$ matrix can be written as

$$X_{I\alpha} = \frac{1}{P} \sum_{j=1}^P \Theta_{I\alpha}(\mathbf{R}_j^N). \quad (2.38)$$

Note that from the above definition of the matrix elements, the main difference between the method presented here and classical MS-CG in terms of numerical implementation is that the estimator for the CG observable, i.e., the discretized path integral force estimator

$f_{I,x} = \frac{1}{P} \sum_{j=1}^P F_I(R_j^N)$, replaces the classical counterpart. This is a natural result of quantum

Boltzmann statistics.

The quadratic function in Eq. (2.36) can then be reformulated into the normal form:

$$\begin{aligned} \chi^2[\phi] &= \frac{1}{3Nn_t} (\phi^T \mathbf{X}^T \mathbf{X} \phi - 2\mathbf{X}^T \mathbf{f} \phi + \mathbf{f}^T \mathbf{f}) \\ &\equiv \phi^T \mathbf{G} \phi - 2\mathbf{b}^T \phi + \mathbf{f}^T \mathbf{f} \end{aligned} \quad (2.39)$$

in which the normal matrix $\mathbf{G}$ is a N_D by N_D square matrix, and $\mathbf{b}$ is the normal vector with dimension N_D . Minimizing the above function with respect to $\{\phi_D\}$ shows that the solving of

this optimization problem is equivalent to solving the normal system of N_D equations with N_D unknown variables:

$$\mathbf{G}\phi = \mathbf{b}. \quad (2.40)$$

Many standard numerical linear algebra algorithms can solve this linear system. In this work, the Singular Value Decomposition (SVD) approach is applied. The reader is referred to Ref. 57 for the detailed numerical analysis of different solvers, regularization etc., and the tips for implementation.

IV. Numerical Examples

The previous sections demonstrate that the MS-CG method can be extended to quantum Boltzmann statistics, and that the classical MS-CG algorithms can also be adapted to form a qMS-CG algorithm by direct modification. The qMS-CG method approximates the consistent qCG force field by solving a least square problem. In the present section, the application of the theory and algorithms are illustrated through two test systems: (1) a series of one-dimensional multi-body model problems, and (2) a modified liquid water system under ambient conditions.

A. One-Dimensional Models

The first numerical example is intended to show how the method works in simple (but not trivial) 1D multi-body quantum systems. The basic system contains three particles in 1-D. These three particles are positioned in a harmonic well with force constant k . Particle 1 and 2 interact through a harmonic bond, whose force constant is k_{bond} and equilibrium bond length is x_0 . Particle 3 interacts with particles 1 and 2 through Gaussian repulsive potentials, and these two Gaussian interactions are not necessarily identical. Given this description, the Hamiltonian describing the model system can be written as follows:

$$H = \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} + \frac{p_3^2}{2m_3} + \frac{1}{2}k(x_1^2 + x_2^2 + x_3^2) + V_{13}(x_1, x_3) + V_{23}(x_2, x_3) + \frac{1}{2}k_{bond}(|x_1 - x_2| - x_0)^2. \quad (2.41)$$

The repulsive Gaussian interactions between particle 1, 2, and particle 3 are defined as:

$$\begin{aligned} V_{13}(x_1, x_3) &= A_{13} \exp[-(x_1 - x_3)^2 / \alpha_{13}] \\ V_{23}(x_2, x_3) &= A_{23} \exp[-(x_2 - x_3)^2 / \alpha_{23}]. \end{aligned} \quad (2.42)$$

In this model system, eight parameters (k_{bond} , x_0 , k , A_{13} , α_{13} , A_{23} , α_{23} , T) can be tuned to generate different statistical distributions of the particle configurations. The soft Gaussian repulsive interactions allow the particles to pass through each other infrequently and thus generate double well like free energy profiles, as can be seen in the following numerical results. The system was propagated by normal mode PIMD (NMPIMD) simulations using the path integral Langevin equation (PILE) integrator.⁵⁸ In order to test the method at different physical conditions, four different sets of parameters were chosen to build the test model systems, and the details of the interaction parameters are listed in TABLE 2-I.

TABLE 2-I. Parameters for the 1D model systems (in atomic units).

Cases	T	k_{bond}	x_0	k	A_{13}	α_{13}	A_{23}	α_{23}
1	200	0.10	4.0	5.0	2.0	2.0	4.0	2.0
2 ^a	200	0.10	4.0	5.0	2.0	2.0	4.0	2.0
3	200	0.10	4.0	10.0	5.0	2.0	10.0	2.0
4	50	0.10	4.0	6.0	1.7	2.0	3.4	2.0

^aIn case 2, the equilibrium position of the harmonic potential of particle 2 is shifted to $x = 0.30$ Å.

It should be noted that in case 1 and 2, the interaction parameters are identical except that in case 2, the equilibrium position for the harmonic potential of particle 2 is shifted from $x = 0.0 \text{ \AA}$ to $x = 0.30 \text{ \AA}$, which creates an asymmetric potential energy surface. For all the cases, the system was propagated with PIMD for 5.0 ns with a time step of 0.1 fs. For cases 1, 2, and 3, $P = 32$ beads were shown to be enough to get a converged result, while for case 4, $P = 64$ beads were needed due to the low temperature and thus stronger nuclear quantum effects. After obtaining the all atom (AA, i.e., FG) PIMD simulation trajectory, the trajectory was coarse-grained by different CG mapping operators. In case 1, 2 and 4, the particles 1 and 2 were coarse-grained onto their center of mass, and particle 3 was eliminated from the system. While in case 3, the whole system was coarse-grained onto particle 2, so that both particle 1 and 3 were eliminated from the system. Then, the qMS-CG algorithm in Sec. III was used to derive the effective potential of the qCG site. In this example, the CG force field was expressed by a linear combination of B-Spline basis functions:

$$f(x) = \sum_{j=1}^N \phi_j B_j(x). \quad (2.43)$$

The coefficients $\{\phi_k\}$ were obtained by solving the least squares problem of Eq. (39). Then, the qMS-CG model was propagated under the same simulation conditions. The corresponding results for cases 1-3 are shown in Fig. 2-1, and case 4 in Fig. 2-2.

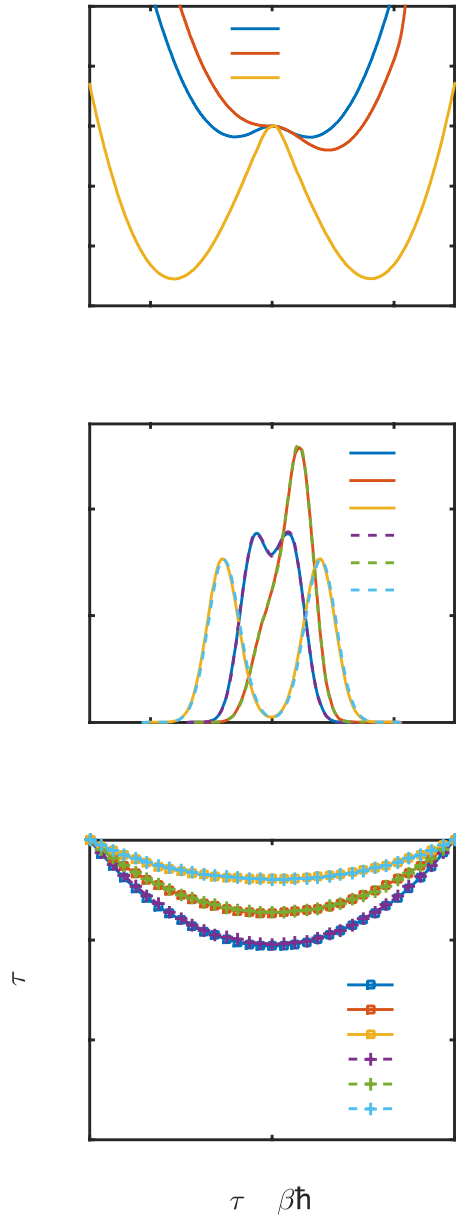


Figure 2-1: Coarse-graining of the three particle Hamiltonian of Eq. (41) into a single bead representation, (a) the effective qCG potential of the CG sites obtained by qMS-CG method for case 1 (blue solid curve), case 2 (red solid curve), and case 3 (yellow solid curve); (b) The position probability distribution function of the simulations for case 1 (CG-blue solid line, AA-

Figure 2-1, continued: purple dashed line), case 2 (CG-red solid line, AA-green dashed line), and case 3 (CG-yellow solid line, AA-cyan dashed line); (c) The imaginary time correlation function (ITCF) defined in Eq. (44) for case 1 (CG-blue solid line with square marker, AA-purple dashed line with “+” marker), case 2 (CG-red solid line with square marker, AA-green dashed line with “+” marker), and case 3 (CG-yellow solid line with square marker, AA-cyan dashed line with “+” marker). The ITCF were only evaluated on the marked points, because the imaginary time paths were discretized in actual simulations.

Cases 1-3 mimic physical systems at moderate temperature (200 K in this case). The effective qCG potential of the coarse-grained sites inferred by the qMS-CG method are shown in Fig. 2-1(a). As shown in Fig. 2-1(b), the quantitative agreement obtained for the position probability distribution functions illustrate that the inferred qCG potential can nearly exactly reproduce the AA PIMD sampling results, which highlights the accuracy of the qMS-CG method in reproducing the thermodynamic ensemble averages.

In addition to the distribution functions, the imaginary time correlation functions (ITCFs) are also compared. The definition of an ITCF is given by

$$C(\tau) = \frac{\langle \mathbf{r}(\tau) \cdot \mathbf{r}(0) \rangle}{\langle \mathbf{r}(0) \cdot \mathbf{r}(0) \rangle}, \quad (2.44)$$

the physical behavior of which describes the degree of quantum delocalization of the particle that is under study. In Fig. 2-1(c), the quantitative agreement between the qCG and AA sampling results show that the effective potential that is derived by the qMS-CG algorithm can describe the correct quantum delocalization and fluctuation of the CG variables, in the cases where the temperature is not too low.

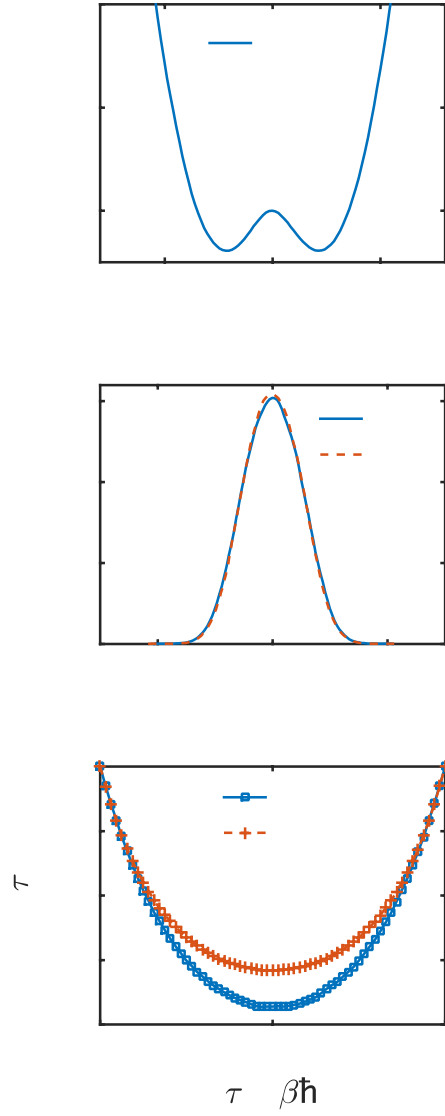


Figure 2-2: Coarse-graining the three particle Hamiltonian of Eq. (2.41) into a CG single bead representation for case 4: (a) the classical PMF of the CG site (blue solid curve), (b) The quantum position probability distribution function of the simulations (CG-blue solid line, AA-purple dashed line), (c) The ITCF (CG-blue solid line with square marker, AA-purple dashed line with “+” marker). The ITCFs were only evaluated on the marked points, because the imaginary time paths were discretized in actual simulations.

In case 4, the FG Hamiltonian was designed to mimic a deep tunneling quantum system at low temperature (50 K). In this case, the nuclear quantum effects, instead of the potential energy profile, will begin to dominate the statistics of the system. The classical PMF of the CG variable was calculated through a classical MD simulation and the result is shown in Fig. 2-2(a). In this case, the value of $\beta\hbar\omega_b = 11.80$ where the ω_b is the imaginary barrier crossing frequency of the classical double well potential. This value is greater than 2π and thus characterizes the system as being in the deep tunneling region (below the “crossover temperature”).⁵⁹ This behavior can be seen in Fig. 2-2(b). The qCG site position probability distribution functions obtained from PIMD samplings are unimodal in spite of the double well classical PMF, suggesting a significant contribution from quantum tunneling. The effective potential for the CG variable was derived by the qMS-CG method and then the qCG simulation was performed using this potential. Figures 2-2(b) and 2-2(c), respectively, show the comparison of the position probability distribution functions, as well as the ITCFs, for AA and CG sampling. It is seen that the probability distribution functions are essentially consistent between the two resolutions. However, the qCG system overestimates the quantum delocalization and thus smaller imaginary time correlations are observed. This is likely due to the fact that the CG action that was assumed to describe the system does not represent the full quantum physics of the system. A possible theoretical extension that includes the contribution of influence functional in the CG action may be one approach to address the problem.^{38, 60} In this case, the action can be written as:

$$S_{CG}[\mathbf{R}(\tau)] = \int_0^{\beta\hbar} d\tau \frac{1}{2} \dot{\mathbf{R}}^T \cdot \tilde{\mathbf{M}} \cdot \dot{\mathbf{R}} + V_{CG}[\mathbf{R}(\tau)] + y[\mathbf{R}(\tau)]. \quad (2.45)$$

The influence functional describes the effect of environment (in this case, the eliminated behavior of the FG system) on the quantum delocalization of the system particles (in this case, the CG sites) in general system-bath coupling problems.

To summarize this results section, these simple illustrative examples demonstrate that the qMS-CG method can achieve thermodynamically consistent CG models that are largely capable of reproducing the FG ensemble averages and imaginary time correlations, with an exception of the latter being for the deep tunneling region (but even here the qMS-CG result is not poor). Theories and methods to improve the accuracy of the qMS-CG approach are left for future work.

B. A modified water model

The second test of the qMS-CG method was performed for a molecular liquid system based on water, but with artificially larger nuclear quantum effects. An all-atom model of 233 water molecules was simulated at 300 K in NVT ensemble for 5.0 ns. The system was modeled by the qSPC/Fw water force field⁶¹ in a cubic box with a box length of 19.034 Å. The electrostatic interactions were evaluated by the Particle-Particle Particle-Mesh method with a real space cutoff distance at 10.0 Å. In this study, the mass of both hydrogen and oxygen atoms were intentionally set as the mass of hydrogen, i.e., 1.008 a.u., in order to enhance the nuclear quantum effects and to better demonstrate the qMS-CG results. We note that this mass repartitioning would not affect the structure of water system in classical MD simulations, but it results in a significant change in the quantum statistics. The normal mode PIMD simulations were propagated by the PILE integrator.⁵⁸ $P = 32$ path integral beads were used in the simulation to ensure that the quantum effects are sufficiently taken into account. The ring polymer contraction scheme^{62, 63} and rRESPA multiple time step integrator^{64, 65} were applied to speed up

the simulation and extend the simulation time step. The long range part of electrostatic interactions was evaluated only for the path centroid for every 1.0 fs. Short range pairwise interactions were contracted on the lowest 5 normal modes and were updated every 0.5 fs, and 0.1 fs was the frequency for updating bond and angle forces. An in-house modified LAMMPS molecular dynamics engine⁶⁶ was used to generate the PIMD simulation trajectory.

After obtaining the FG PIMD trajectories, the one-site CG model for these modified water molecules was parameterized through the qMS-CG method. For each path integral bead in the ring polymer, the COM mapping operator was applied to generate the CG trajectory at that particular imaginary time, which was later passed to the qMS-CG code to obtain the force matching matrices and vectors. Finally, the normal equation in the qMS-CG algorithm described earlier was constructed from these matrices and vectors. The normal equation was then solved by SVD decomposition.

The qCG interactions included contributions from both the pairwise interactions and a Stillinger-Weber like non-bonded three body interaction,⁶⁷ written as

$$U(\mathbf{R}_N) = \sum_I \sum_{J>I} \phi_2(R_{IJ}) + \sum_I \sum_{J \neq I} \sum_{K>J} \phi_3(R_{IJ}, R_{IK}, \theta_{IJK})$$

$$\phi_3(R_{IJ}, R_{IK}, \theta_{IJK}) = \lambda_{IJK} \epsilon_{IJK} (\cos \theta_{IJK} - \cos \theta_0)^2 \exp\left(\frac{\gamma_{IJ} \sigma_{IJ}}{R_{IJ} - a_{IJ} \sigma_{IJ}}\right) \exp\left(\frac{\gamma_{IK} \sigma_{IK}}{R_{IK} - a_{IK} \sigma_{IK}}\right) \quad (2.46)$$

It has been shown previously that the pairwise interaction itself in the MS-CG method for one site CG water may not represent the hydrogen-bonding network well and can fail to capture the correct structural fluctuations in liquid water.⁶⁸ Therefore, the explicit three-body Stillinger-Weber potential was adopted to correct the short-range structures of the present quantum model for liquid water. The pairwise interaction was approximated by the linear combination of 36 6th-order B-Spline basis functions and a tabulated form was then employed. The method to obtain the Stillinger-Weber interactions can be found in Ref.⁶⁸

Figure 2-3 shows the pairwise interaction obtained by the qMS-CG approach, and the parameters for the three body Stillinger-Weber interaction are $\lambda = 6.788$ kcal/mol, $\cos \theta_0 = -0.4464$, $a=3.80$ Å, $\gamma = 1.20$ and $\varepsilon = \sigma = 1.0$.

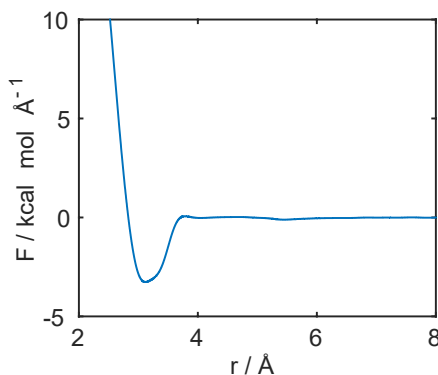


Figure 2-3: Pairwise interaction between the CG sites for the modified liquid water system, as obtained by the qMS-CG method.

After obtaining the force field parameters, this one-site qMS-CG water model was then simulated for 5.0 ns in the constant NVT ensemble, to validate the consistency between the CG and AA sampling. All the simulation details remained the same. At the same time, a classical MD simulation of the AA system was also performed. The comparison between qCG, quantum AA, and the classical AA simulation results are presented in Fig. 2-4. The statistics of the radial distribution function (RDF), which represents the pair structure of the liquid, and three-body non-bonded angle distributions, which represents the short range orientational correlations of the liquid molecules, are compared in Fig. 2-4. In Fig. 2-4(a), the results of quantum PIMD simulations and classical MD simulation are significantly different because of the reduced the mass of oxygen in this modified water which exaggerates the difference between the classical and quantum sampling for this model. As clearly seen in both panels of Fig. 2-4, the qMS-CG

one site model well reproduces the pair structure as well as the orientation preference of the modified liquid water system. The agreement in Fig. 2-4 between the RDFs as well as the three-body non-bonded angle distributions also demonstrates that the Stillinger-Weber like interactions employed in the CG model provides a good representation of an optimally consistent qCG force field. As can also be seen in Fig. 2-4(a), the first solvation shell and the long-range structures in the RDF are quantitatively preserved, though the position of the second peak on the RDF is not perfectly matched possibly due to the limitation of the basis functions that were chosen to describe the interactions.

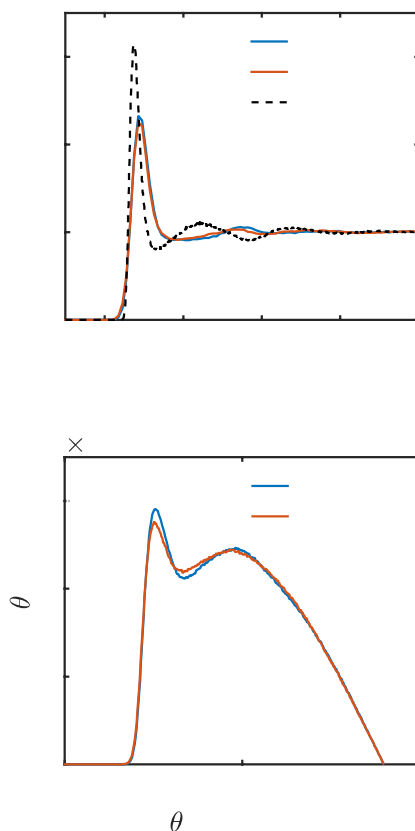


Figure 2-4: Panel (a) Center of mass radial distribution functions obtained from the FG quantum PIMD simulation (red solid line), Coarse-Grained PIMD simulation using the force field generated by qMS-CG (blue solid line) as well as the result from classical MD simulation (black

Figure 2-4, continued: dash line). Panel (b) The probability distribution function for the three-body non-bonded angles obtained from the FG quantum PIMD simulation (red solid line) and qMS-CG-PIMD simulation (blue solid line). The cut-off distance for evaluating the three-body non-bonded angles was 4.0 Å.

In addition to reproducing the liquid structure, the qMS-CG one-site model quantitatively preserves the imaginary time correlations for this model. In this example, the imaginary time mean-square displacement function has the following form,

$$R^2(\tau) = \left\langle \left| \mathbf{r}(\tau) - \mathbf{r}(0) \right|^2 \right\rangle = 2 \langle r(0)^2 \rangle [1 - C(\tau)]. \quad (2.47)$$

was evaluated and compared at both AA and CG level. This function is directly related to the imaginary time correlation function that was previously defined in Eq. (2.44). Physically, it characterizes the magnitude of quantum delocalization. Figure 2-5 shows the comparison of the

calculated from the qMS-CG and AA PIMD samplings, where it is clearly seen that under the current simulation conditions the qCG model derived by the qMS-CG algorithm quantitatively reproduces the behavior of the quantum delocalization of the underlying FG system.

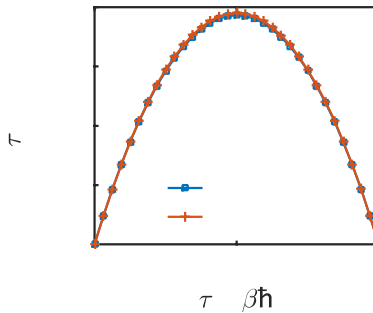


Figure 2-5: The delocalization function $R^2(\tau)$ in Eq. (2.47) obtained from AA PIMD simulations (red solid line with “+” markers) and qMS-CG PIMD simulations (blue solid line

Figure 2-5, continued: with square markers). The function was evaluated only on the marked points due to the fact that $P = 32$ imaginary time slices discretized the imaginary time Feynman paths.

V. Quasi-Classical Approximation

In Sec. II after Eq. (2.9) we mentioned that there are two possible routes that can be taken to formulate a coarse-graining strategy for equilibrium quantum systems. In the subsequent sections we followed the first route in which the qCG model was built based on the consistency of CG and FG density matrices for the whole imaginary time path of the FG system. In the present section, however, we will explore the second route and as such an alternative perspective for coarse-graining of quantum Boltzmann statistics is proposed. The main idea is to approximate the consistent CG density matrix so that it behaves like a classical probability density distribution function. And a quasi-classical CG force field could be derived based on this approximation. This provides a most classical description of the CG procedure in quantum Boltzmann statistics and will help us build a clearer physical picture of coarse-graining in quantum statistical mechanics. In this section, only the consistency in configuration space is considered so that we can start the derivation based on Eq. (2.28) in Sec. II E.

The consistency condition implies the diagonal element of the CG density matrix is related with its FG counterpart by:

$$\hat{\rho}_{\mathbf{R}\mathbf{R}}^{CG} = \int d\mathbf{r} \hat{\rho}_{\mathbf{r}\mathbf{r}} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}). \quad (2.48)$$

Notice that the diagonal elements of the FG density matrix can be expressed as the expectation value of a Dirac delta operator, such that

$$\begin{aligned} \hat{\rho}_{\mathbf{r}\mathbf{r}} &= \langle \mathbf{r} | e^{-\beta H} | \mathbf{r} \rangle \\ &= \int d\mathbf{r}' \delta(\mathbf{r} - \mathbf{r}') \langle \mathbf{r}' | e^{-\beta H} | \mathbf{r}' \rangle = Z \cdot \langle \delta(\mathbf{r} - \mathbf{r}') \rangle \end{aligned} \quad (2.49)$$

where Z is the partition function for the system. In order to derive the most classical form of this density matrix, one natural choice is to employ the concept of the path centroid variable suggested by Feynman, which is a powerful tool for developing approximate but accurate quasi-classical theories and for exploring the quantum-classical correspondence in statistical mechanics.⁴³

For any operator A in configuration space, its expectation value in quantum Boltzmann statistics can be formulated in the path centroid representation as follows:⁶⁹

$$\langle A \rangle = \langle A_c(q_c) \rangle_{\rho_c} = \frac{\int dr_c \rho_c(r_c) A_c(r_c)}{\int dr_c \rho_c(r_c)} = \frac{1}{Z} \int dr_c \rho_c(r_c) A_c(r_c) . \quad (2.50)$$

in which $\rho_c(r_c)$, the centroid probability density, is used in the statistical averaging. (We note that for simplicity we use a one-dimensional notation here and r a single rectilinear coordinate, not a radial coordinate.) Within the path integral centroid representation, the original form of the operator should be reformulated as an effective centroid-dependent quasi-classical function $A_c(q_c)$. However, an exact analytical formula for this quasi-classical function for a general operator often cannot be derived. Instead, it can be approximated by the expression⁶⁹

$$A_c(r_c) \cong \frac{1}{\sqrt{2\pi\Delta r^2}} \int d\tilde{r} A(r_c + \tilde{r}) \exp(-\tilde{r}^2 / 2\Delta r^2), \quad (2.51a)$$

where

$$\Delta r^2 = \left\langle (r - r_c)^2 \right\rangle_{r_c} . \quad (2.51b)$$

Here, a cumulant expansion up to second order was employed so that the quantity Δr^2 characterizes the fluctuations or delocalization of the imaginary time paths relative to the centroid variable constrained at position r_c .

With the results in Eq. (2.51), the Dirac delta function operator in Eq. (2.49) in multiple dimensions can be approximated as a Gaussian distribution function in terms of the path centroid variable:⁶⁹

$$\begin{aligned}\delta_c(\mathbf{r}_c) &\equiv \frac{1}{\sqrt{2\pi \det \Lambda}} \int d\tilde{\mathbf{r}} \delta(\mathbf{r} - \mathbf{r}_c - \tilde{\mathbf{r}}) \exp[-(\tilde{\mathbf{r}}^T \cdot \Lambda \cdot \tilde{\mathbf{r}}) / 2] \\ &= \frac{1}{\sqrt{2\pi \det \Lambda}} \exp\left\{-[(\mathbf{r} - \mathbf{r}_c)^T \cdot \Lambda \cdot (\mathbf{r} - \mathbf{r}_c)] / 2\right\}\end{aligned}\quad (2.52)$$

The matrix $\Lambda_{ij} = \langle \tilde{r}_i(\tau) \tilde{r}_j(\tau) \rangle = \langle (r_i(\tau) - r_{ic})(r_j(\tau) - r_{jc}) \rangle$ here describes the imaginary time quantum fluctuation of the FG system relative to the centroid variables.

Combining Eq. (2.49) and Eq. (2.52), the diagonal density matrix element of the FG system can now be approximated as:

$$\hat{\rho}_{\mathbf{r}\mathbf{r}} = Z \cdot \langle \delta(\mathbf{r} - \mathbf{r}') \rangle \propto \int d\mathbf{r}_c \rho_c(\mathbf{r}_c) \exp\left\{-[(\mathbf{r} - \mathbf{r}_c)^T \cdot \Lambda \cdot (\mathbf{r} - \mathbf{r}_c)] / 2\right\}. \quad (2.53)$$

Now, plugging this result in the CG consistency condition in Eq. (2.48) reveal that the diagonal density matrix element for CG model can be approximated by

$$\hat{\rho}_{\mathbf{R}\mathbf{R}}^{CG} \propto \int d\mathbf{r} \delta(M(\mathbf{r}) - \mathbf{R}) \int d\mathbf{r}_c \rho_c(\mathbf{r}_c) \exp\left\{-[(\mathbf{r} - \mathbf{r}_c)^T \cdot \Lambda \cdot (\mathbf{r} - \mathbf{r}_c)] / 2\right\}. \quad (2.54)$$

This expression can be further simplified by replacing the delta function by its representation in terms of its Fourier transform and switching the order of integration to yield

$$\hat{\rho}_{\mathbf{R}\mathbf{R}}^{CG} \propto \int d\mathbf{r}_c \rho_c(\mathbf{r}_c) \exp\left\{-[(\mathbf{M}(\mathbf{r}_c) - \mathbf{R})^T \cdot \mathbf{A}^{-1} \cdot (\mathbf{M}(\mathbf{r}_c) - \mathbf{R})] / 2\right\}. \quad (2.55)$$

More details of the derivation are given in Appendix B. In Eq. (2.55), $(M(\mathbf{r}_c) - \mathbf{R})$ is understood to be the vector of CG mappings, and the matrix $\mathbf{A}$ describes the contributions of FG system quantum fluctuations to the relevant CG sites; its matrix elements can be written as

$$A_{IJ} = \sum_{\alpha} \sum_{\beta} c_{I\alpha} c_{J\beta} \Lambda_{\alpha\beta}^{-1}, \quad (2.56)$$

where the coefficients $\{c_{I\alpha}\}$ are the CG mapping coefficients defined in Eq. (2.7).

Equation (2.55) is the main result of this section, which arguably provides the most simple and “classical-like” description of coarse-graining, but still within the confines of quantum Boltzmann statistics and the Uncertainty Principle. It also suggests a clearer physical picture of the effects of quantum mechanics on the bottom up procedure of coarse-graining. By comparison to the thermodynamic consistency condition in classical statistical mechanics,⁷⁻¹¹

$$\rho_{cl}^{CG}(\mathbf{R}) \propto \int d\mathbf{r} \rho_{cl}(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \quad (2.57)$$

where “ cl ” denotes “classical limit”, the effects of quantum mechanics come into this quasi-classical theory of coarse-graining in two ways in Eq. (2.55). First, the quantum statistics of the FG system are manifested in the centroid probability density $\rho_c(\mathbf{r}_c)$. This effect has nothing to do with the coarse-graining per se. Second, the effect of the quantum delocalization appears in the CG mapping by effectively replacing the “strict” mapping Dirac delta functions in the classical coarse-graining expression [Eq. (2.57)] by “smeared” Gaussian distributions centered at the same positions in the quantum CG expressions [Eq. (2.55)]. This interpretation is in fact amusing and pleasing at the same time. It says that you may think you can rigorously coarse-grain a FG system, but because of the Uncertainty Principle and the intrinsic delocalization of the quantum particles, you not only cannot completely know where the new CG particles are, but you also cannot be completely sure how they were mapped from the FG representation in the first place. The diagram shown in Fig. 2-6 schematically depicts this concept.

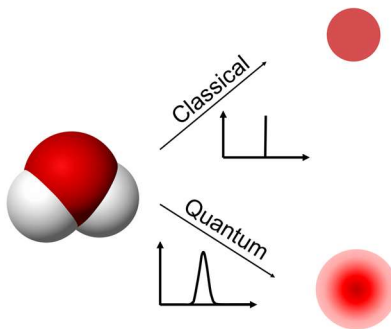


Figure 2-6: Schematic diagram showing the difference between coarse-graining in classical statistical mechanics and quantum Boltzmann statistics.

A very natural development based on Eq. (2.55) is to also derive an effective “classical-like” qCG potential $U_{qCG}(\mathbf{R})$ (and force field through differentiation) that approximately takes the quantum effects into account, such that

$$e^{-\beta U_{qCG}(\mathbf{R})} \equiv Z \cdot \hat{\rho}_{CG}(\mathbf{R}) \propto \int d\mathbf{r}_c \rho(\mathbf{r}_c) \exp \left\{ - \left[(\mathbf{M}(\mathbf{r}_c) - \mathbf{R})^T \cdot \mathbf{A}^{-1} \cdot (\mathbf{M}(\mathbf{r}_c) - \mathbf{R}) \right] / 2 \right\}. \quad (2.58)$$

This expression provides a very interesting way to develop classical-like qCG force fields that can also represent the effects of quantum statistics. Nonetheless, two obstacles must be overcome to develop a practical algorithm from this equation. First, obtaining the centroid density distribution function for a molecular system with a large number of degrees of freedom can be non-trivial and must be done numerically “on the fly” in the simulation. Second, the Gaussian type coarse-graining distributions are difficult to obtain up front, so more detailed mathematical and physical analysis is needed. Future research will be devoted to an exploration of the underlying theory and numerical methods for this approach to quantum coarse-graining.

VI. Concluding Remarks

In this paper we have introduced a theory of quantum mechanical coarse-graining and its numerical implementation within Boltzmann statistical mechanics. An analysis of the sufficient consistency conditions based on linear CG mapping operators is presented. It is seen that a specific choice of CG action functional and effective qCG force field can generate a consistent description of the underlying FG model. A variational principle based on the force consistency condition, which was extended from the classical MS-CG theory, is also generalized for quantum Boltzmann statistics, and it provides the basis for deriving effective qCG force fields using equilibrium path integral sampling of the FG system. The numerical algorithm and practical issues in implementing this variational minimization procedure are further discussed. Then, two numerical examples were used to demonstrate the accuracy of the qCG method presented, while also revealing that it cannot completely reproduce the imaginary time correlation function for systems in deep tunneling region without further modification. Finally, a quasi-classical approximation of the qCG density matrix is shown to provide an appealing physical picture of coarse-graining in quantum Boltzmann statistics, and may open up a way for building classical-like qCG force fields that approximately incorporate nuclear quantum effects.

A number of directions for the future development of qCG theories and methods are also suggested by the present analysis. First, one important problem to be addressed is how to best preserve the imaginary time correlations in the deep tunneling regime. The introduction of an influence functional in the qCG action, which might be the key to this issue, will be investigated in the future. Second, it is suggested in Sec. II that the relative entropy minimization approach to coarse-graining can also be extended to quantum Boltzmann statistics. Future work will seek to develop variational principles and numerical algorithms based on that framework. And finally, future research will also seek to generalize these results to the challenging cases of Bose-Einstein

and Fermi-Dirac statistics, which are expected to provide additional and unique insight into the procedure of coarse-graining for fully quantum mechanical problems at any temperature.

APPENDIX A: DERIVATION OF EQUATION (2.26)

First we prove that if any function $g(\mathbf{r})$ of atomistic configuration and $G(\mathbf{R})$ of CG configuration has the following connection:

$$g(\mathbf{r}) = G(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}]). \quad (2.A1)$$

Then for any CG path constraint ensemble average, we have

$$\langle g[\mathbf{r}(\tau)] \rangle_{\mathbf{R}(\tau)} = G(\mathbf{R}(\tau)). \quad (2.A2)$$

Then, considering the following equation, it is possible to validate its correctness such that

$$\left\langle \int \int d\tau d\tau' \Delta(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(\tau)]) \cdot g[\mathbf{r}(\tau')] \right\rangle_{\mathbf{R}(\tau)} = \int \int d\tau d\tau' \Delta(\mathbf{R}(\tau)) \cdot \langle g[\mathbf{r}(\tau')] \rangle_{\mathbf{R}(\tau)}. \quad (2.A3)$$

The following expression is also valid for any CG path constraint ensemble average:

$$\begin{aligned} & \left\langle F_I[\mathbf{M}_{\mathbf{R}_I}(\mathbf{r}(\tau))] + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\dot{\mathbf{r}}(\tau)] - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right\rangle_{\mathbf{R}(\tau)} \\ & = F_I[\mathbf{R}(\tau)] + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\mathbf{R}(\tau)] - \left\langle \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right\rangle_{\mathbf{R}(\tau)} = 0 \end{aligned} \quad (2.A4)$$

By combining all of these results together, one can show that

$$\left\langle \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \Delta_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(s)]) \cdot \left(F_I[\mathbf{M}_{\mathbf{R}_I}(\mathbf{r}(s'))] + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\dot{\mathbf{r}}(\tau)] - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right) \right\rangle_{\mathbf{R}(\tau)} = 0 \quad (2.A5)$$

holds for any choice of $\mathbf{R}(\tau)$, therefore,

$$\left\langle \int_0^{\beta\hbar} \int_0^{\beta\hbar} ds ds' \Delta_I(\mathbf{M}_{\mathbf{R}_I}[\mathbf{r}(s)]) \cdot \left(F_I[\mathbf{M}_{\mathbf{R}_I}(\mathbf{r}(s'))] + M_I \cdot \mathbf{M}_{\mathbf{R}_I}[\dot{\mathbf{r}}(\tau)] - \mathbf{M}_I^f(f[\mathbf{r}(\tau)] + m\ddot{\mathbf{r}}(\tau)) \right) \right\rangle_{FG} = 0, \quad (2.A6)$$

APPENDIX B: DERIVATION OF EQUATION (2.55)

The first step of derivation involves replacing the Dirac delta functions by its Fourier transform representation, such that

$$\begin{aligned}\hat{\rho}_{\mathbf{R}\mathbf{R}}^{CG} &\propto \int d\mathbf{r} \int \frac{d\mathbf{k}}{(2\pi)^N} \exp[i\mathbf{k} \cdot (\mathbf{M}(\mathbf{r}) - \mathbf{R})] \int d\mathbf{r}_c \rho_c(\mathbf{r}_c) \cdot \exp\left\{-[(\mathbf{r} - \mathbf{r}_c)^T \cdot \Lambda \cdot (\mathbf{r} - \mathbf{r}_c)]/2\right\} \\ &\propto \int d\mathbf{r}_c d\mathbf{k} \rho_c(\mathbf{r}_c) \exp[-i\mathbf{k} \cdot \mathbf{R}] \int d\mathbf{r} \exp\left[-\frac{1}{2}(\mathbf{r} - \mathbf{r}_c)^T \cdot \Lambda \cdot (\mathbf{r} - \mathbf{r}_c) + i\mathbf{k} \cdot M(\mathbf{r})\right]\end{aligned}\quad (2.B1)$$

Then the Gaussian integral over $\mathbf{r}$ is performed, and the result is given by

$$\begin{aligned}\hat{\rho}_{\mathbf{R}\mathbf{R}}^{CG} &\propto \int d\mathbf{r}_c d\mathbf{k} \rho_c(\mathbf{r}_c) \exp[i\mathbf{k} \cdot (\mathbf{M}(\mathbf{r}_c) - \mathbf{R})] \int d\mathbf{r} \exp\left[-\frac{1}{2}\mathbf{r}^T \cdot \Lambda \cdot \mathbf{r} + i\mathbf{k} \cdot M(\mathbf{r})\right] \\ &\propto \int d\mathbf{r}_c d\mathbf{k} \rho_c(\mathbf{r}_c) \exp[i\mathbf{k} \cdot (\mathbf{M}(\mathbf{r}_c) - \mathbf{R})] \exp\left[-\frac{1}{2} \sum_{\alpha} \sum_{\beta} \sum_I \sum_J k_I k_J c_{I\alpha} c_{J\beta} \Lambda_{\alpha\beta}^{-1}\right],\end{aligned}\quad (2.B2)$$

where the coefficients $\{c_{I\alpha}\}$ are the CG mapping coefficients. Finally, the Gaussian integral over $\mathbf{k}$ is performed and then Eq. (2.55) is obtained.

Reference

1. M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.* **42**, 73 (2013).
2. W. G. Noid, *J. Chem. Phys.* **139**, 090901 (2013).
3. W. Tschöp, K. Kremer, J. Batoulis, T. Bürger and O. Hahn, *Acta Polym.* **49**, 61 (1998).
4. D. Reith, M. Pütz and F. Müller-Plathe, *J. Comput. Chem.* **24**, 1624 (2003).
5. F. Müller-Plathe, *ChemPhysChem* **3**, 754 (2002).
6. T. Murtola, E. Falck, M. Patra, M. Karttunen and I. Vattulainen, *J. Chem. Phys.* **121**, 9156 (2004).
7. S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
8. S. Izvekov and G. A. Voth, *J. Chem. Phys.* **123**, 134105 (2005).
9. W. G. Noid, J.-W. Chu, G. S. Ayton and G. A. Voth, *J. Phys. Chem. B* **111**, 4116 (2007).

10. W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
11. W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen and G. A. Voth, *J. Chem. Phys.* **128**, 244115 (2008).
12. M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).
13. M. S. Shell, *Adv. Chem. Phys.*, 395 (2016).
14. J. F. Rudzinski and W. G. Noid, *J. Chem. Phys.* **135**, 214101 (2011).
15. F. Paesani and G. A. Voth, *J. Phys. Chem. B* **113**, 5702 (2009).
16. J. A. Morrone and R. Car, *Phys. Rev. Lett.* **101**, 017801 (2008).
17. F. Paesani, S. Iuchi and G. A. Voth, *J. Chem. Phys.* **127**, 074506 (2007).
18. T. F. Miller and D. E. Manolopoulos, *J. Chem. Phys.* **123**, 154504 (2005).
19. M. Diraison, G. J. Martyna and M. E. Tuckerman, *J. Chem. Phys.* **111**, 1096 (1999).
20. J. A. Morrone, V. Srinivasan, D. Sebastiani and R. Car, *J. Chem. Phys.* **126**, 234504 (2007).
21. F. Paesani, S. S. Xantheas and G. A. Voth, *J. Phys. Chem. B* **113**, 13118 (2009).
22. G. R. Medders and F. Paesani, *J. Am. Chem. Soc.* **138**, 3912 (2016).
23. J. Chen, X.-Z. Li, Q. Zhang, M. I. J. Probert, C. J. Pickard, R. J. Needs, A. Michaelides and E. Wang, *Nat. Commun.* **4**, 045103 (2013).
24. M. Ceriotti, J. Cuny, M. Parrinello and D. E. Manolopoulos, *Proc. Natl. Acad. Sci.* **110**, 15591 (2013).
25. L. Wang, C. M. Isborn and T. E. Markland, *Methods Enzymol* **577**, 389 (2016).
26. S. Hammes-Schiffer, *Curr. Opin. Struct. Biol.* **14**, 192 (2004).
27. S. R. Billeter, S. P. Webb, T. Iordanov, P. K. Agarwal and S. Hammes-Schiffer, *J. Chem. Phys.* **114**, 6925 (2001).
28. T. E. Markland and B. J. Berne, *Proc. Natl. Acad. Sci.* **109**, 7988 (2012).
29. T. E. Markland, J. A. Morrone, B. J. Berne, K. Miyazaki, E. Rabani and D. R. Reichman, *Nat. Phys.* **7**, 134 (2011).

30. M. E. Tuckerman, D. Marx, M. L. Klein and M. Parrinello, *Science* **275**, 817 (1997).
31. J. Lobaugh and G. A. Voth, *J. Chem. Phys.* **104**, 2056 (1996).
32. M. Dominik, E. T. Mark and P. Michele, *J. Phys. Condens. Matter.* **12**, A153 (2000).
33. J. S. Kretchmer and T. F. Miller, *J. Chem. Phys.* **138**, 134109 (2013).
34. D. M. Ceperley, *Rev. Mod. Phys.* **67**, 279 (1995).
35. T. D. Hone and G. A. Voth, *J. Chem. Phys.* **121**, 6412 (2004).
36. R. Feynman, *Statistical mechanics: a set of lectures*. (W. A. Benjamin Inc, Reading, Massachusetts, 1972).
37. R. P. Feynman and A. R. Hibbs, *Quantum mechanics and path integrals*. (McGraw-Hill, New York, 1965).
38. D. Chandler and P. G. Wolynes, *J. Chem. Phys.* **74**, 4078 (1981).
39. B. J. Berne and D. Thirumalai, *Annu. Rev. Phys. Chem.* **37**, 401 (1986).
40. M. Parrinello and A. Rahman, *J. Chem. Phys.* **80**, 1984 (1984).
41. J. Cao and G. A. Voth, *J. Chem. Phys.* **100**, 5093 (1994).
42. J. Cao and G. A. Voth, *J. Chem. Phys.* **99**, 10070 (1993).
43. G. A. Voth, *Adv. Chem. Phys.* **93**, 135 (1996).
44. I. R. Craig and D. E. Manolopoulos, *J. Chem. Phys.* **121**, 3368 (2004).
45. S. Habershon, D. E. Manolopoulos, T. E. Markland and T. F. Miller, *Annu. Rev. Phys. Chem.* **64**, 387 (2013).
46. S. Jang, A. V. Sinitskiy and G. A. Voth, *J. Chem. Phys.* **140**, 154103 (2014).
47. A. V. Sinitskiy and G. A. Voth, *J. Chem. Phys.* **143**, 094104 (2015).
48. Y. Peng, Z. Cao, R. Zhou and G. A. Voth, *J. Chem. Theory Comput.* **10**, 3634 (2014).
49. M. Nava, R. Quhe, F. Palazzesi, P. Tiwary and M. Parrinello, *J. Chem. Theory Comput.* **11**, 5114 (2015).
50. M. Nava, F. Palazzesi, C. Perego and M. Parrinello, *J. Chem. Theory Comput.* **13**, 425 (2017).

51. Z. Cao and G. A. Voth, J. Chem. Phys. **143**, 243116 (2015).
52. Z. Zhang, J. Pfaendtner, A. Grafmüller and G. A. Voth, Biophys. J. **97**, 2327 (2009).
53. Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner and G. A. Voth, Biophys. J. **95**, 5073 (2008).
54. S. Lafon and A. B. Lee, IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1393 (2006).
55. J. W. Wagner, J. F. Dama, A. E. P. Durumeric and G. A. Voth, J. Chem. Phys. **145**, 044108 (2016).
56. V. Vedral, Rev. Mod. Phys. **74**, 197 (2002).
57. L. Lu, S. Izvekov, A. Das, H. C. Andersen and G. A. Voth, J. Chem. Theory Comput. **6**, 954 (2010).
58. M. Ceriotti, M. Parrinello, T. E. Markland and D. E. Manolopoulos, J. Chem. Phys. **133**, 124104 (2010).
59. W. H. Miller, J. Chem. Phys. **61**, 1823 (1974).
60. D. Chandler, Y. Singh and D. M. Richardson, J. Chem. Phys. **81**, 1975 (1984).
61. F. Paesani, W. Zhang, D. A. Case, T. E. Cheatham and G. A. Voth, J. Chem. Phys. **125**, 184507 (2006).
62. T. E. Markland and D. E. Manolopoulos, Chem. Phys. Lett. **464**, 256 (2008).
63. T. E. Markland and D. E. Manolopoulos, J. Chem. Phys. **129**, 024105 (2008).
64. M. Tuckerman, B. J. Berne and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992).
65. M. E. Tuckerman, B. J. Berne, G. J. Martyna and M. L. Klein, J. Chem. Phys. **99**, 2796 (1993).
66. S. Plimpton, J. Comput. Phys. **117**, 1 (1995).
67. F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).
68. L. Larini, L. Lu and G. A. Voth, J. Chem. Phys. **132**, 164107 (2010).
69. G. A. Voth, Phys. Rev. A **44**, 5302 (1991).

Chapter 3: Coarse-Graining Involving Virtual Sites: Centers of Symmetry Coarse-Graining

This chapter is Reproduced from J. Chem. Phys. 150, 154103 (2019), with the permission of AIP Publishing.

Abstract

Coarse-grained (CG) models allow efficient molecular simulation by reducing the degrees of freedom in the system. To recapitulate important physical properties, including many-body correlations at the CG resolution, an appropriate mapping from the atomistic to CG level is needed. Symmetry exhibited by molecules, especially when aspherical, can be lost upon coarse-graining due to the use of spherically-symmetric CG effective potentials. This mismatch can be efficiently amended by imposing symmetry using virtual CG sites. However, there has been no rigorous bottom-up approach for constructing a many-body potential of mean force that governs the distribution of virtual CG sites. Herein, we demonstrate a statistical mechanical framework that extends a mapping scheme of CG systems involving virtual sites to provide a thermodynamically consistent CG model. Utilizing the extended framework, this work defines a center of symmetry (COS) mapping and applies it to benzene and toluene systems, such that the planar symmetry of the aromatic ring is preserved by constructing two virtual sites along a normal vector. Compared to typical center of mass (COM) CG models, COS CG models correctly recapitulate radial and higher order correlations, e.g., orientational and three-body correlations. Moreover, we find that COS CG interactions from bulk phases are transferable to mixture phases whereas conventional COM models deviate between the two states. This result suggests a systematic approach to construct more transferable CG models by conserving molecular symmetry, and the new protocol is further expected to capture other many-body correlations by utilizing virtual sites.

I. Introduction

Coarse-grained (CG) models enable one to extend the temporal and spatial scales of molecular simulations by averaging out the finer details below the resolution of the CG site (or “bead”).¹⁻⁷ While computationally efficient, some detailed configurations, or spatial correlations, are lost during the CG model construction and resulting simulation. Thus, an appropriate CG mapping from the fine-grained (FG), e.g., atomistic, degrees of freedom to the CG model is a prerequisite for parameterizing CG models that can faithfully represent the physical properties of the system of interest using the reduced CG degrees of freedom.⁸⁻¹³

A mapping operator represents the correspondence between the FG configurations $\mathbf{r}^n$ to the mapped CG configurations $\mathbf{R}^N$. Since there are no general or definitive rules for mappings, various operators have been suggested and implemented for different systems. As one example, the MARTINI force field has a straightforward mapping rule that is to approximately map four heavy atoms with associated hydrogen atoms to a single CG bead.¹⁴⁻¹⁶ On the other hand, CG models can be constructed from the dynamical information inferred from the essential dynamics of all-atom simulations. One such approach is known as Essential Dynamics Coarse-Graining (ED-CG), which characterizes and groups atoms sharing similar dynamical behavior, such as fluctuations in the low-frequency regime.¹⁷⁻²¹ Similarly, other approaches, the diffusion map and sketch map approaches, have also been developed to extract relevant low-dimensional dynamic information by reducing dimensions in a nonlinear sense.²²⁻²⁶

In a more physically transparent manner, a group of atoms can be mapped into their center of mass (COM) as the CG variable. As an example, the multiscale coarse-graining (MS-CG) approach⁸⁻¹² has been applied to various scales of molecular systems ranging from condensed matter (liquids) to biological systems (such as lipids) using COM mappings. In such

cases, the equilibrium distribution of the CG variables should be consistent with that of the reference FG statistical ensemble (known as the “consistency” relationship) for those CG mappings, which establishes an equivalence between the force acting on a CG site and the summation of exerted forces on the underlying FG atoms. A derivation of this consistency condition is detailed in Ref. ¹¹. However, in some cases, the COM mapping scheme might not be a good choice to thoroughly represent and reproduce certain physical properties of the system. Since a simple two-body interaction between the COM distances might not be enough to fully account for the many-body effects in condensed phases, various CG models have been suggested to include higher order interactions, which significantly affect the computational efficiency.²⁷⁻³²

Alternatively, one can improve CG mapping while only using the pair interactions between the CG sites. One example is the center of charge (COC) mapping that provides a better description of Coulombic interactions for molecules where the conventional MS-CG model based on COM configurations is not enough to capture long-range interactions.³³ However, two drawbacks still remain in the COC mapping. First, the COC mapping assumes that the mapping weighting factor of the COC CG beads is proportional to atomic charges within the subgroups. In turn, it does not provide a rigorous thermodynamic relationship between the suggested CG mapping and the FG system. Also, this approach is most valid if the system has highly charge-separated groups, such as in polar molecules.

In addition, one of the basic limitations of conventional CG models is that each CG site is generally treated as a spherically symmetric particle, resulting in discrepancies when asymmetric molecules are mapped to single (or multiple) sites. For example, we recently demonstrated that the symmetry mismatch between the CG and FG beads leads to the failure of MS-CG models in interfacial systems where the symmetry or orientation plays an important role.³⁴ One might

instead use the Gaussian overlap formalism³⁵ or Gay-Berne potential,³⁶ which have been extended to parameterize CG interactions,³⁷⁻³⁹ or explicitly include dipole moments,⁴⁰⁻⁴² respectively. However, these strategies often require expensive computational overhead to both parameterize and utilize such CG models in order to maintain high accuracy with faithful reproduction of the FG structure.

To overcome the aforementioned discrepancies in symmetry and to provide an efficient and accurate CG interaction, the present work presents a center-of-symmetry (COS) CG mapping that is designed to preserve the symmetry of the molecule in CG configuration space by introducing virtual CG sites. Here, “virtual” sites are defined as fictitious particles that are introduced, in contrast to other mapping schemes (e.g., COM) that rely on physical atoms. A straightforward example to demonstrate this mapping is benzene, which has planar geometry. In order to preserve the planar symmetry, the COS CG mapping yields two virtual CG sites that are normal to the aromatic plane, introducing additional degrees of freedom (rather than only using their pairwise distances as order parameters) that can provide higher quality basis sets for elucidating a CG many-body potential of mean force (PMF). This approach provides a simple way to represent the local many-body correlation of non-spherical molecules with marginally more computational cost. We further envisage that our proposed CG model will enhance the transferability of the effective CG interactions,^{5, 43} which will be discussed later.

In order to represent the effective interactions between the virtual CG sites correctly, we require a rigorous statistical mechanical approach to construct the PMF describing the distributions of virtual CG sites. In this article, we will develop a systematic CG theory involving virtual sites for the first time. The thermodynamic consistency condition is first derived at the resolution of the COM variables. A bottom-up parameterization protocol utilizing the force-

matching equations is then introduced based on these conditions. It is demonstrated that the MS-CG method can be extended in this framework with minimal modification. Our approach is expected to facilitate the modeling of complex biomolecules such as lipids and membranes by adopting the symmetry or shape from the all-atom resolution into the FG to CG mapping operator. This is in line with shape-based CG^{44, 45} or related *ad hoc* CG approaches^{46, 47} that have been applied to certain biosystems.

The remaining sections of this paper are as follows: In Section II, we first address the CG mapping involving virtual sites and derive a thermodynamic consistency relationship based on the principle of maximum entropy. Then, we construct and review the CG models for benzene and toluene. Practical implementation details for the CG simulations are also given. The proposed CG mapping is next evaluated in Section III by analyzing the structural properties including radial, orientational, and three-body correlations for benzene derivatives. We also present an analysis of transferability by comparing the effective CG interactions from bulk and mixed phases at the end of Section III. Section IV provides a summary and concludes the paper.

II. Methodology

A. COS Mapping

As discussed in Sec. I, a CG model is defined by a set of mapping operators $M_{\mathbf{R}}^N : \mathbf{r}^n \rightarrow \mathbf{R}^N$. These mapping operators are conventionally designed as a linear function to most readily capture the physics of the FG system at the level of resolution that one would like to focus on. For instance, the COM mapping operator captures the translational motion of the FG particles, while at the same time coarse-graining out the internal rotational and vibrational motions. Similarly, the center of geometry (COG) mapping operator constructs a representation of the geometric center of FG particles. However, in the COS mapping, the coarse-graining procedure involves

two sets of linear mapping operators in the following forms: one for the CG mapping from the FG trajectory onto the COM representation with $\mathbf{R}_\alpha^{COM} \equiv \tilde{\mathbf{R}}_\alpha$ and the other for the reverse mapping from the COM to the virtual CG bead positions, respectively,

$$\begin{aligned}\tilde{\mathbf{R}}_\alpha &= \sum_{i \in \alpha} c_{\alpha i} \mathbf{r}_i \\ \tilde{\mathbf{R}}_\alpha &= \sum_{l \in \alpha} \chi_{\alpha l} \mathbf{R}_l\end{aligned}\tag{3.1}$$

These mappings can be written in a more compact form as

$$\begin{aligned}\tilde{\mathbf{R}}^{\tilde{N}} &= M(\mathbf{r}^n) \\ \tilde{\mathbf{R}}^{\tilde{N}} &= \tilde{M}(\mathbf{R}^N)\end{aligned}\tag{3.2}$$

A schematic description of the proposed mapping relationship is shown in Figure 3-1. The reverse mapping function is carefully designed to manifest intrinsic symmetrical features, such as the dipole moment vector of polar molecules or the normal vector of planar molecules (e.g., benzene), of the underlying FG system that is integrated out in the COM level of description. Detailed design and setup of the COS model will be addressed in Section II. D.

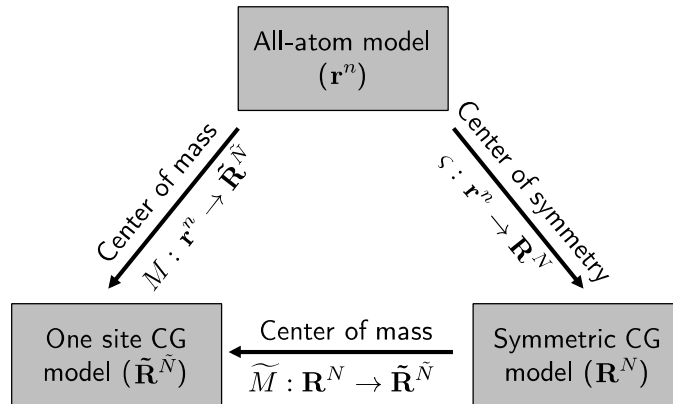


Figure 3-1: Schematic relationship among different CG mappings presented in this work: from the all-atom (FG) model to the one-site coarse-grained (COM model) from M and symmetric

Figure 3-1, continued: coarse-grained model (COS model) from ζ . Re-mapping the center of mass of the COS model from $\tilde{M}$ yields the one-site COM model.

Unlike the conventional COM and COG mappings, the physical meaning of CG beads in the COS mapping does not directly correspond to a concrete physical entity. The center of mass of the COS CG beads represents the most resolved physical entity, whereas the COS CG beads themselves are just virtual particles characterizing certain symmetrical features of the underlying FG molecules. We will show in the next section that the presence of these virtual CG beads helps to characterize the complicated many-body orientational correlations of the FG system in a more intuitive and computationally efficient way.

It should be noted that many-body correlations can be intrinsically captured by the COM mapping when appropriate basis sets are utilized (e.g., containing terms beyond pairwise two-body). However, since information on the orientation of CG beads is missing in the case of simple pairwise interactions, more complicated functional forms for CG interactions are often necessary to fully take this many-body effect into account.²⁷⁻³² Embedding this complexity can introduce challenges for both the parameterization and propagation of the resultant CG models. However, the aforementioned shortcoming is alleviated in the COS mapping approach since the proposed reverse mapping explicitly embeds the orientation information into the virtual CG beads.

B. The Thermodynamic Consistency Conditions

With the CG mapping operator in hand, the “perfect” CG system is required to behave exactly like the FG system at the resolution that the CG system was constructed. To reiterate, so-called structural thermodynamic consistency is a sufficient condition for the CG model to be consistent

with the corresponding atomistic model in configuration space. Thus, in order to ensure so-called structural thermodynamic consistency, a quantitative relationship between the CG and FG representations should first be developed. This consistency requires that at the resolution of the *center of masses*, the FG model and COS CG model generate the same statistical distribution (we focus on the consistency in configuration space only). Suppose that the probability distribution functions for the center of masses are known, to ensure the thermodynamic consistency conditions at the resolution of the center of masses, it requires the COS CG and COM distribution functions be connected through,

$$\tilde{\rho}(\tilde{\mathbf{R}}^{\tilde{N}}) = \int d\mathbf{R}^N \rho_{CG}(\mathbf{R}^N) \delta(\tilde{\mathbf{R}} - \tilde{M}(\mathbf{R}^N)). \quad (3.3)$$

in which the delta function is a shorthand notation for a product of delta functions of each CG center of mass, i.e.,

$$\delta(\tilde{\mathbf{R}}_I - \tilde{M}(\mathbf{R}^N)) = \prod_{I=1}^{\tilde{N}} \delta(\tilde{\mathbf{R}}_I - \tilde{M}_I(\mathbf{R}^N)). \quad (3.4)$$

To infer the COS CG distribution functions based on the above consistency condition, one can maximize the entropy of the COS CG systems with respect to the distribution function,

$$S_{CG}[\rho_{CG}] = -k_B \int d\mathbf{R}^N \rho_{CG}(\mathbf{R}^N) \ln \rho_{CG}(\mathbf{R}^N), \quad (3.5)$$

under the constraint of Eq. (3.3) as well as the normalization condition $\int d\mathbf{R}^N \rho_{CG}(\mathbf{R}^N) = 1$. By introducing Lagrange multiplier function $\lambda(\mathbf{R})$, one can maximize the entropy through solving an unconstrained optimization of the following Lagrangian through the calculus of variations,

$$L[\rho_{CG}(\mathbf{R}), \lambda(\mathbf{R}), \lambda_0] = -k_B \int d\mathbf{R} \rho_{CG}(\mathbf{R}) \ln \rho_{CG}(\mathbf{R}) - \lambda_0 \left(\int d\mathbf{R} \rho_{CG}(\mathbf{R}) - 1 \right) - \int d\tilde{\mathbf{R}} \lambda(\tilde{\mathbf{R}}) \left(\int d\mathbf{R}^N \rho_{CG}(\mathbf{R}^N) \delta(\tilde{\mathbf{R}} - \tilde{M}(\mathbf{R}^N)) - \tilde{\rho}(\tilde{\mathbf{R}}^{\tilde{N}}) \right), \quad (3.6)$$

The result shows that,

$$\rho_{CG}(\mathbf{R}^N) = \frac{\tilde{\rho}(\tilde{M}(\mathbf{R}^N))}{\Omega(\tilde{M}(\mathbf{R}^N))}, \quad (3.7)$$

in which the measure $\Omega(\tilde{M}(\mathbf{R}^N))$ defined by,

$$\Omega(\tilde{\mathbf{R}}) = \int d\mathbf{R}^N \delta(\tilde{\mathbf{R}} - \tilde{M}(\mathbf{R}^N)), \quad (3.8)$$

tells the number of state for the COS CG system given a specific center of mass configuration. In general it's a function dependent on the COM configurations reflecting the geometric features of the mapping operator $\tilde{M}$. In the case of linear mapping operators, the measure can be evaluated to be a constant independent of specific COM configurations.

On the other hand, as has been previously shown in many works, the center of mass distribution functions can be evaluated from the fine-grained atomistic distribution functions through,

$$\tilde{\rho}(\tilde{\mathbf{R}}^N) = \int d\mathbf{r}^n \rho(\mathbf{r}^n) \delta(\tilde{\mathbf{R}} - M(\mathbf{r}^n)), \quad (3.9)$$

in which the Dirac delta function is similarly defined as $\delta(\tilde{\mathbf{R}}_I - M(\mathbf{r}^n)) = \prod_{I=1}^N \delta(\tilde{\mathbf{R}}_I - M_I(\mathbf{r}^n))$.

Combining Eq. (3.7) with Eq. (3.9), we arrive at,

$$\rho_{CG}(\mathbf{R}^N) = \frac{1}{\Omega(\tilde{M}(\mathbf{R}^N))} \int d\mathbf{r}^n \rho(\mathbf{r}^n) \delta(\tilde{M}(\mathbf{R}^N) - M(\mathbf{r}^n)), \quad (3.10)$$

which provides a least biased estimation of the COS CG distribution functions based on the atomistic statistical distributions.

More specifically, if the mapping from COS sites to the COMs is linear, the number of states measure can become a constant. In the scope of this paper, we choose to map two COS virtual CG sites onto their center of masses under equal weights, in which case the measure function can

be found out to be $\Omega(\tilde{M}(\mathbf{R}^N)) = (2V)^{\tilde{N}}$, therefore finally we get the thermodynamic consistency condition relating the probability distribution functions for the atomistic and COS CG systems writing as,

$$\rho_{CG}(\mathbf{R}^N) = \frac{1}{(2V)^N} \int d\mathbf{r}^n \rho(\mathbf{r}^n) \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)). \quad (3.11)$$

Detailed derivation procedure for the measure function as well as an alternative approach to derive the above equation is discussed in Appendix B.

Based on Eq. (3.11), supposing that the FG and CG systems all obey classical Boltzmann statistics, their effective Hamiltonians are linked by:

$$e^{-\beta U(\mathbf{R}^N)} \propto \int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)), \quad (3.12)$$

in which $u(\mathbf{r}^n)$ denotes the potential energy of the FG system whereas $U(\mathbf{R}^N)$ represents the CG counterpart. By taking a natural logarithm on both sides of Eq. (3.12), the effective CG potential is expressed as:

$$U(\mathbf{R}^N) = -\frac{1}{\beta} \ln \int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) + (const). \quad (3.13)$$

Once the effective CG potential function is defined, the force exerted on CG bead as a function of the CG system configuration $\mathbf{R}^N$ can be obtained by differentiating the potential energy profile, namely:

$$F_I(\mathbf{R}^N) = \frac{\int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \prod_{\gamma \neq \alpha} \delta(M_\gamma(\mathbf{r}^n) - \tilde{M}_\gamma(\mathbf{R}^N)) \frac{\partial}{\partial \mathbf{R}_I} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N))}{\int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N))}, \quad (3.14)$$

In the above derivation, we have assumed that each COS CG bead is involved in the only one particular definition of COM variable. The same assumption is also applied to each FG atom with respect to COM variables.

To simplify the force expression presented in Eq. (3.14), one can note that the following relationship holds due to the linear property of both the “forward” and “reverse” CG mapping operators:

$$\frac{\partial}{\partial \mathbf{R}_I} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) = -\chi_{\alpha I} \sum_{i \in \alpha} \frac{d_{\alpha i}}{c_{\alpha i}} \frac{\partial}{\partial r_i} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) , \quad (3.15)$$

in which the set of coefficients $\{d\}$ satisfies the following condition from the definition of a CG mapping

$$\sum_{i \in \alpha} d_{\alpha i} = 1 . \quad (3.16)$$

Readers can refer to Appendix C for the details. Utilizing the above condition and integrating by parts, Eq. (14) can be further simplified; the CG effective force field that ensures the thermodynamic consistency can be written as a constrained ensemble average:

$$F_I(\mathbf{R}^N) = \left\langle \chi_{\alpha I} \sum_{i \in \alpha} \frac{d_{\alpha i}}{c_{\alpha i}} \cdot \left(-\frac{\partial u(\mathbf{r}^n)}{\partial r_i} \right) \right\rangle_{\mathbf{R}^N} , \quad (3.17)$$

where the constrained ensemble average of a function takes the following form:

$$\langle g(\mathbf{r}^n) \rangle_{\mathbf{R}^N} = \frac{\int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) g(\mathbf{r}^n)}{\int d\mathbf{r}^n e^{-\beta u(\mathbf{r}^n)} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N))} . \quad (3.18)$$

In the case of linear CG mapping operators, we can choose the coefficients $\{d\}$ so that the following condition holds:

$$c_{\alpha i} = d_{\alpha i} . \quad (3.19)$$

The condition of Eq. (3.16) still holds since translational invariance requires that the mapping coefficients $\{c\}$ sum to unity for each of the COM configurations. By putting all the previous equations together, a thermodynamically-consistent force field of the COS CG system can be expressed as:

$$F_I(\mathbf{R}^N) = \left\langle \chi_{\alpha I} \sum_{i \in \alpha} f_i \right\rangle_{\mathbf{R}^N}, \quad (3.20)$$

in which the COS CG particle is specific to the COM denoted by α . The physical meaning of this equation is that the total force on CG bead is a constrained ensemble average of a partition of the total atomistic forces exerted on the center of mass. This relationship further implies that for any CG mapping that accompanies the virtual CG sites as an explicit coordinate with respect to its corresponding COM position, the thermodynamic consistency condition is valid as long as the configurations of the virtual CG sites can be written as a linear combination of COM CG sites. This equation forms the basis of the variational principle that will be employed in this work to extract the CG effective force field.

C. Force-Matching

Once the CG mapping operator is defined, an atomistic simulation (FG) trajectory or other atomistic configurational information can be readily mapped into a CG representation that preserves all of the structural correlations. The next goal is to derive the equation of motions for the CG system using mapped information. In other words, the thermodynamic consistency conditions derived in the previous subsection provide the theoretical basis for the inference of the CG effective interactions. According to Eq. (3.13), the underlying force field describing the CG system should, in principle, be extracted by calculating the constraint ensemble averages. However, the numerical implementation of such approach is quite challenging due to the difficulty in handling the constraints as well as the high dimensional nature of many-body

systems. Practically designed variational principles based on the consistency conditions can be formulated in order to construct numerical approximations to the thermodynamically-consistent CG force fields. The MS-CG method^{8, 9, 11, 12} built upon the force-matching equation as well as the Relative Entropy Minimization (REM) method^{6, 48-50} are two examples of such approaches.

In this work, the CG force field is parameterized utilizing the MS-CG method. Specifically, the CG force field is determined by variationally minimizing the following force residual functional,^{8, 9, 11, 12}

$$\chi^2[\mathbf{f}] = \frac{1}{3N} \left\langle \sum_{I=1}^N \left| \mathbf{f}_I(\zeta_I(\mathbf{r}^n)) - \mathcal{F}_I(\mathbf{r}^n) \right|^2 \right\rangle, \quad (3.21)$$

In consistent with the original MS-CG method,^{8, 9, 11, 12} $\mathcal{F}_I(\mathbf{r}^n)$ in Eq. (3.14) denotes the microscopic force on the CG site I and $\mathbf{f}_I(\zeta_I(\mathbf{r}^n))$ is the model parameter given by the CG mapping $\zeta_I(\mathbf{r}^n)$. In general, the above functional optimization problem in an infinite dimensional space is computationally forbidding. For practical implementation, the test CG force field function $\mathbf{f}_I(\zeta_I(\mathbf{r}^n))$, or $\mathbf{f}_I(\mathbf{R}^N)$, is often expanded in a subspace of the infinite dimension space, which is constructed by a set of basis functions. Such basis functions are usually determined based on some degree of physical intuition about the system of interest. Many of these basis functions can be expressed as key variables of the system, such as the pair distance and triplet angle for non-bonded and bonded atoms, respectively. In this work, we assume that the CG force field only includes contributions from pair non-bonded interactions to achieve a physically accurate and efficient CG model. Under this assumption, the CG force field in this work takes the following form,

$$\mathbf{f}_I(\mathbf{R}^N) = \sum_{J \neq I} \phi_2(R_{IJ}) \hat{e}_{IJ}. \quad (3.22)$$

The basis function $\phi_2(R_{IJ})$ is approximated by a set of spline functions $\{u_k\}$, i.e.

$$\mathbf{f}_I(\mathbf{R}^N) = \sum_{J \neq I} \phi_2(R_{IJ}) \hat{e}_{IJ} = \sum_{J \neq I} \sum_k c_k u_k(R_{IJ}) \hat{e}_{IJ} . \quad (3.23)$$

The set of coefficients $\{c_k\}$ parameterizing the pairwise basis function are then solved by minimizing the residual functional, which is now cast as a quadratic function through the above approximations, in the least square sense. Readers are referred to Ref. ⁵¹ for the detailed analysis of different solvers, regularizations, and spline function specifications of the MS-CG algorithm.

D. Model Description and System Setup

In this subsection, we will introduce the COS and conventional COM mapping operators for benzene and toluene systems in the liquid phase. First, we reiterate that benzene as a model system was chosen to test the COS mapping in the case of non-spherical symmetry because benzene molecules exhibit a planar structure due to the sp^2 hybridization throughout the aromatic ring. Figure 3-2 shows three possible CG mapping schemes for a benzene molecule. We did not consider CG models with more than two sites per molecule to maximize the efficiency of the mapping (i.e., resolution) and computational cost.

One can intuitively map a molecule into a single CG bead corresponding to the COM of the whole molecule as shown in Figure 3-2(a). However, as mentioned in Sec. I, a benzene model consisting of CG beads governed by spherically symmetric interactions is not able to capture the correct planar symmetry of the original molecule at the FG resolution. As a result, the single-site COM model only preserves the overall translational motion while losing local orientational correlation. In particular, the single-site model is not able to distinguish between two possible configurations for a pair of benzene molecules at the same intermolecular distance,

e.g., the T-shape and sandwich shape configurations, which have different energetics.⁵²⁻⁵⁷ A detailed analysis of the discrepancies of single-site CG models for different geometries is also discussed in Ref. ³⁴.

Alternatively, Figure 3-2(b) shows a two-site COM CG model where two adjacent C_3H_3 moieties are coarse-grained to two CG sites at their respective COM and connected by a harmonic bond (denoted as a grey line). Even though this model incorporates the planar geometry to the previous single-site COM model and removes the implicit isotropy, the rotation of the benzene ring along the vector between the two COM sites is degenerate, and thus we expect a poor description of the planar symmetry. Specifically, with regard to the aforementioned dimer configurations, the two-site COM mapping is not able to distinguish between stacking or relatively tilted configurations.⁵⁵ While maintaining the same CG mapping resolution as Figure 3-2(b), one could perform the COS CG mapping to preserve the planar symmetry. A schematic description of mapped COS beads is shown in Figure 3-2(c). As discussed in Sec. II, the COM of the COS system is equivalent to the COM of the benzene molecule, which is shown as a green bead in Figure 3-2(a). To satisfy this condition, the two COS beads must be equidistant from the COM, giving $\chi_{al} = 0.5$. Therefore, from Eq. (3.13), the effective CG forces acting on the COS CG bead in configuration $\mathbf{R}^N$ are simply expressed as half of the total force that is exerted on the molecule,

$$F_l(\mathbf{R}^N) = \frac{1}{2} \left\langle \sum_{i \in \alpha} -\frac{\partial u(\mathbf{r}^i)}{\partial \mathbf{r}_i} \right\rangle_{\mathbf{R}^N}. \quad (3.24)$$

This model resembles the charge separation model suggested by Hunter and Sanders, in which the π - π interactions are modeled by placing particles along the normal vector of the plane. However, our CG model provides non-bonding interactions in the CG configuration space where

the molecular symmetry is preserved, while their model is a simple electrostatic model that incorporates charge interactions to represent the π -system.⁵⁸ In terms of the force-matching procedure, our intent here is to include an additional basis set that can be conceived as a projected quadrupole moment of benzene to correctly reproduce the CG interactions.⁵⁹⁻⁶²

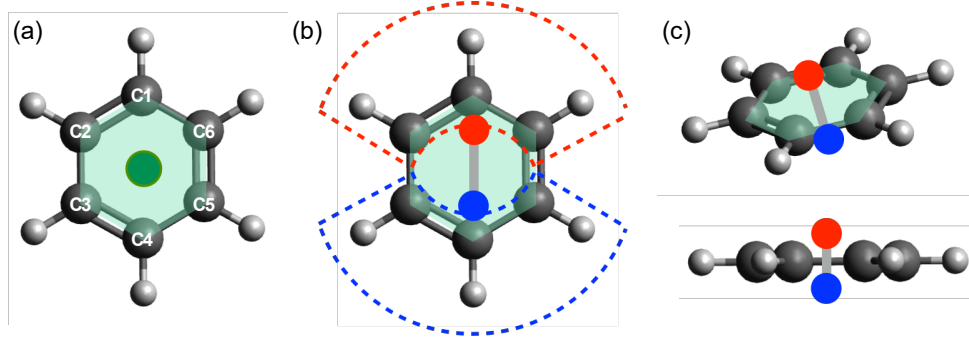


Figure 3-2: CG mapping schemes considered in this study for benzene: (a) one-site center of mass (COM) mapped model as green bead, (b) two-site COM mapped model as red ($-\text{C}_3\text{H}_3$) and blue ($-\text{C}_3\text{H}_3$) beads, and (c) two-site center of symmetry (COS) mapped model as red (upper half) and blue (lower half) beads. Red beads represent CG type 1 and blue beads are CG type 2 in actual simulations.

In practice, the COS mapping $\zeta: \mathbf{r}^n \rightarrow \mathbf{R}^N$ is performed by first constructing the normal vector of the benzene plane $\vec{d}'_{\parallel} = (\mathbf{r}_{\text{C3}}^I - \mathbf{r}_{\text{C1}}^I) \times (\mathbf{r}_{\text{C5}}^I - \mathbf{r}_{\text{C1}}^I)$ and the COM configuration of benzene molecule I which is denoted as $\tilde{\mathbf{R}}_I$. Then, two virtual CG sites are constructed as $(\tilde{\mathbf{R}}_I - \vec{u}'_{\parallel} \cdot l, \tilde{\mathbf{R}}_I + \vec{u}'_{\parallel} \cdot l)$ where l is the center-to-particle distance or separation distance with the normalized vector $\vec{u}'_{\parallel} = \vec{d}'_{\parallel} / |\vec{d}'_{\parallel}|$. The thermodynamic consistency relationship derived in Sec. II does not confine the value of separation distance l , but practical limitations and constraints

should be considered. As designed to impose planar symmetry and represent π - π interactions, the value of l should be chosen to enable efficient sampling of the intermolecular region between COS CG sites that are not in the same molecule. Based on the reported minimum at 3.6 Å from the dimer potential energy surface using the OPLS force field,⁵⁷ l must be smaller than $\frac{\langle \tilde{\mathbf{R}}_{IJ} \rangle - 3.6}{2}$ Å. Since the first coordination shell of liquid benzene is positioned near 6.6-7 Å,^{34,}

⁶³ this rule of thumb yields an upper bound of $(6.6 - 3.6)/2 = 1.5$ Å in order to correctly represent structural correlations of benzene. If the value of l is too large, the COS CG model may sample unphysical configurations in which benzene molecules penetrate each other. On the other hand, consider the opposite limit when l approaches zero as expressed below:

$$\lim_{l \rightarrow 0} \mathbf{R} = \lim_{l \rightarrow 0} \{ \mathbf{R}_l \} = \lim_{l \rightarrow 0} \{ \tilde{\mathbf{R}}_l \pm \vec{u}_{\parallel}^l \cdot l \} = \tilde{\mathbf{R}}. \quad (3.25)$$

In other words, as the separation distance approaches zero, the COS sites reduce back to the COM site. Conversely, this limit clearly demonstrates that the COS mapping breaks rotational degeneracy by introducing a normal vector to the single-site COM configuration. We checked that at very small $l < 0.05$ Å, structural properties of the COS mapping converge to that of the single-site COM model (see Figure 3-A1 in Appendix A). Considering the above-mentioned points carefully, the l value was set to be 0.5 Å, so that the separation between two COS beads is 1 Å.

In a similar vein, we constructed CG models for toluene; recall that toluene is equivalent to benzene except where an -H atom is substituted for a -CH₃ group. As a consequence of the broken C₆ symmetry, the single-site COM CG bead is located slightly off center of the aromatic ring as depicted in Figure 3-3(a), and thus the CG sites in the two-site COM CG model have different types of interactions (Figure 3-3(b)). Even in this case, we are able to utilize the

proposed COS mapping scheme because the COM of the molecule is still within the aromatic plane. In other words, the mapped CG beads are placed equidistant and perpendicular to the toluene plane, and the resultant CG beads are homogeneous as designed. We will examine the effect of C_6 symmetry breaking in the next subsection.

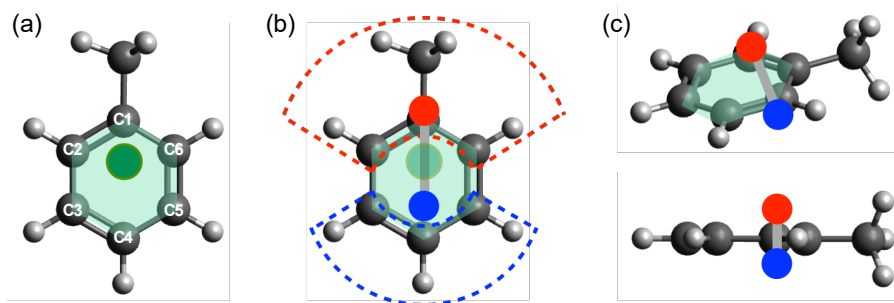


Figure 3-3: CG mapping schemes considered in this study for toluene: (a) one-site center of mass (COM) mapped model as green bead, (b) two-site COM mapped model as red ($-C_4H_5$) and blue ($-C_3H_3$) beads, and (c) two-site center of symmetry (COS) mapped model as red (upper half) and blue (lower half) beads. Red beads are CG type 1 and blue beads are CG type 2 in actual simulations.

E. Simulation details

Each reference atomistic system for the two liquids (i.e., benzene or toluene systems) is comprised of 1000 molecules in a cubic box with periodic boundary conditions. For all liquid molecules, the optimized potentials for liquid simulations all-atom (OPLS-AA) force field was chosen as a force field to conduct the FG (all-atom) simulation.^{64, 65} We also used Ewald summation to account for the long-range electrostatics at the FG resolution.⁶⁶ All MD simulations including both all-atom and CG models were conducted using the LAMMPS simulation package.⁶⁷ To prepare the simulations, initial configurations for the liquid system

were constructed by randomly distributing molecules using the Packmol package,⁶⁸ then the VMD package⁶⁹ was used to build the bonding topology for the systems. After energy-minimization using the steepest descent algorithm, each system of interest was annealed to a temperature $T = 300$ K with a Nosé-Hoover thermostat with $\tau = 0.1$ ps.^{70, 71} Constant NPT dynamics with $P = 1$ atm with $\tau = 1$ ps were performed for 1 ns in order to equilibrate the system.⁷² Finally, we further sampled each system using constant NVT dynamics at $T = 300$ K with a Nosé-Hoover thermostat for 5 ns to collect the configurations for structural analysis every 0.5 ps (same damping constant was also used for NVT dynamics).

The initial configurations for force-matching were prepared from snapshots of the FG trajectories during NVT runs. The effective CG pair potentials from force-matching were obtained as a linear combination of sixth-order B-splines where the resolution of the spline was fixed to 0.2 Å. Due to statistical sensitivity and poor sampling, polynomial terms of $A\mathbf{r}^{-b}$ form were fit after the B-spline fitting at the hard-core region of the potentials.⁷³ For all CG models, we used an outer cutoff of 10.0 Å. The inner cutoff parameters vary for each CG model. For example, the three different inner cutoff values, that correspond to the one-site COM, two-site COM, and two-site COS mapped CG models for benzene, were 3.33 Å, 3.04 Å, and 2.40 Å, respectively. The parameterization of the CG interactions was conducted using the MS-CG program package, which is available at the LAMMPS or Github webpages free of use.⁷⁴

The CG simulations were then performed for 5 ns using constant NVT dynamics at $T = 300$ K with a Nosé-Hoover thermostat. The initial configuration of the CG simulation was mapped from the last snapshot of the constant NVT runs from the all-atom simulations. For propagating the COS CG model, the bonds between virtual CG sites within the same molecule

were constrained via the SHAKE algorithm⁷⁵ to maintain their relative spacing. However, we note that the proposed approach does not necessarily require any constrained or fixed geometry.

F. Velocity Distributions

In Sec. II. D, we discussed the mapping between atomistic and CG virtual site representations in configuration space. In this subsection, we would also like to explore the mapping relationship between the velocities of atomistic and CG virtual representations. It should be noted that the analysis in this subsection does not affect any of the statistical behavior in configuration space that we demonstrated earlier, which is of our primary focus.

By design, the mapping operators between atomistic and COM representations, $M(\mathbf{r}^n)$, and the mapping operators between the COM and CG virtual sites representations, $\tilde{M}(\mathbf{R}^N)$, are linear. However, the direct mapping from atomistic to CG virtual sites representation is non-linear with an explicit dependency on configuration. In the case of benzene, for example, this dependency could be the orientation of the benzene ring. Hence, the velocities of the virtual sites are a superposition of the COM velocities, $\dot{\mathbf{R}}_I^c$, and a line speed of the benzene ring rotation, $\dot{\mathbf{u}}_{\parallel}^l$, expressed as $\dot{\mathbf{R}}_I^u = \dot{\mathbf{R}}_I^c + \dot{\mathbf{u}}_{\parallel}^l l$. Since the non-linearity complicates the CG distribution functions in momentum space, there arises two important questions: (i) are the velocities still normally distributed, and, if so, (ii) is the temperature of the CG system evaluated using the total atomic mass corresponding to the CG bead intuitively consistent with the temperature underlying the atomistic system?

To investigate these two points, we first formulate the velocity mapping relationship. By taking the first order time derivatives of the configuration mapping relationship, we obtain the velocity of virtual CG beads,

$$\begin{aligned}\dot{\mathbf{R}}_I^u &= \dot{\mathbf{R}}_I^c + \dot{\mathbf{u}}_{\parallel}^I l = \sum_i c_{li} \dot{\mathbf{r}}_i + \sum_i \frac{\partial \mathbf{u}_{\parallel}^I}{\partial \mathbf{r}_i} \cdot \dot{\mathbf{r}}_i l \\ &= \sum_i \left(c_{li} + \frac{\partial \mathbf{u}_{\parallel}^I}{\partial \mathbf{r}_i} l \right) \cdot \dot{\mathbf{r}}_i \equiv \sum_i \tilde{B}_{li} \dot{\mathbf{r}}_i\end{aligned}\quad (3.26)$$

In Eq. (3.26), the matrix $\tilde{\mathbf{B}}$, which is a function of the FG configurations, defines the velocity mapping relationship between the CG virtual sites and the underlying atomistic system.

With this definition, the second moment of CG velocity distribution can be estimated as:

$$\begin{aligned}\langle V_I^2 \rangle &= \frac{1}{\mathbf{Z}} \int e^{-\beta H} \tilde{V}_I^2 \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) d\mathbf{r}^n d\mathbf{v}^n \\ &= \frac{1}{\mathbf{Z}} \int e^{-\beta U} \sum_i \sum_j \tilde{B}_{li} \tilde{B}_{lj} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) d\mathbf{r}^n \times \int e^{-\beta K} v_i v_j d\mathbf{v}^n \\ &= \frac{1}{\mathbf{Z}} \int e^{-\beta U} \sum_i \sum_j \tilde{B}_{li} \tilde{B}_{lj} \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) d\mathbf{r}^n \times \frac{k_B T}{m_i} \delta_{ij} \quad , \\ &= k_B T \left\langle \sum_i \tilde{B}_{li}^2 / m_i \right\rangle_{\mathbf{R}^N}\end{aligned}\quad (3.27)$$

in which $\mathbf{Z}$ represents the partition function of the atomistic system. In the last line of Eq. (3.27), the shorthand notation for the CG constrained ensemble average takes the form

$$\langle g \rangle_{\mathbf{R}} = \frac{1}{\mathbf{Z}} \int e^{-\beta U} g(\mathbf{r}^n, \mathbf{R}^N) \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) d\mathbf{r}^n. \quad (3.28)$$

The above result from Eq. (3.27) suggests that if the mass of the virtual CG beads is defined as

$$M_T = \left\langle \sum_i \tilde{B}_{li}^2 / m_i \right\rangle_{\mathbf{R}^N}^{-1}, \quad (3.29)$$

then the temperature consistency between the virtual CG sites and the atomistic representation of the underlying system is preserved. Therefore, M_T , which we denote as the thermo mass, should be our choice for the mass of the CG beads in the following parameterization and CG system propagation processes. In Appendix D, we numerically examined the velocity distribution and system temperature evaluated from the thermo mass using the benzene system. The virtual site velocities are indeed normally distributed in this case, and the choice of thermo mass quantitatively preserves the system temperature.

III. Results and Discussion

A. Benzene

For benzene, we performed and compared the aforementioned two-site CG models constructed by two different CG mapping schemes: based on the center of mass (COM) or the center of symmetry (COS). The comparison in this section was mainly held on the CG models with the same resolution in order to fix and minimize the auxiliary impact caused by the resolution on the quality of CG model.^{13, 76, 77} Figure 3-4 shows the effective pairwise CG interactions for both CG mapping schemes.

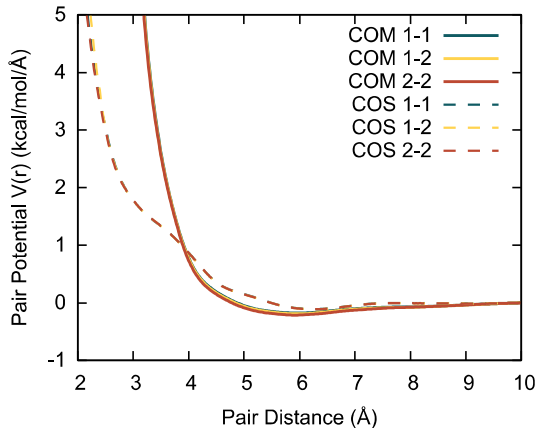


Figure 3-4: Effective pair potentials of benzene CG models obtained from the MS-CG method for two different CG mapping schemes: center of mass (solid lines) and center of symmetry

Figure 3-4, continued: (dotted lines). Both mapping schemes show an equivalent interaction among different types (1-1, 1-2, and 2-2 pairs), indicating homogeneity.

In Figure 3-4, it is clear that the two CG types show identical interactions for both COM (solid lines) and COS (dotted lines) CG models since the COM mapping is applied to equally coarse-grained $-C_3H_3$ moieties while the COS mapping is also designed to map symmetric virtual sites. A notably promising feature in the COS CG interactions is a cusp-like point at short distances (near 4 Å), and this cusp from the pair potential is attributed to the sign changes of effective forces between CG sites, indicating the existence of a locally favorable structure in this region. Namely, this feature suggests an ability to reproduce stacking behavior.

Next, we calculated the structural distribution function of the CG sites to compare the quality of the two CG models. To do so, we remapped the CG particles to a single-site coarse-grained model (i.e., one bead represents one molecule) to compare the distributions with respect to the center of benzene molecules as described in Figure 3-5. Figure 3-5(a) displays molecule-wise radial distribution functions (RDFs) from the mapped all-atom (red line), COM (blue line), and COS (green line) models.

It is immediately apparent that the two-site COM CG models do not correctly reproduce the structural behavior from the FG resolution, especially at the first minimum near 8 Å, indicating that the pair structures within the first coordination shell, which might be affected by π - π interactions, is not captured correctly.

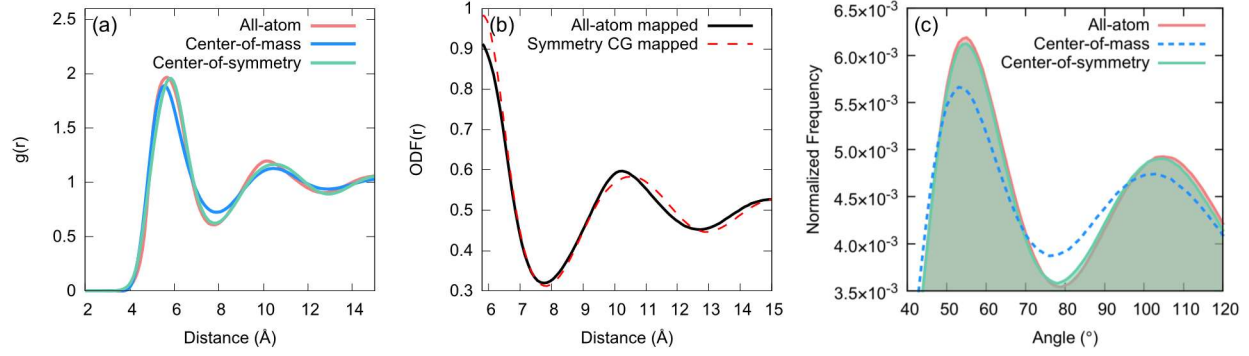


Figure 3-5: Molecule-wise (i.e., re-coarse-grained to one site per molecule) structure analysis of liquid benzene by calculating the (a) radial distribution function (RDF) of the mapped-all atom (red), center of mass (blue), and center of symmetry (green) models, (b) orientation distribution function (ODF) of the mapped-all atom (black solid line) and center of symmetry (red dotted line) models, and (c) three-body angular distribution function (ADF) of the CG molecules of the mapped-all atom (red line), center of mass (blue dotted line), and center of symmetry (green line) as demonstrated in eq (24).

In order to investigate the orientational order throughout the fluid, we also calculated the orientation distribution function (ODF), which is related to the pair orientation. Here, the ODF is defined as⁷⁸

$$\text{ODF}(\mathbf{r}) = \left\langle \left| \mathbf{u}_I \cdot \mathbf{u}_J \right| \right\rangle_{\mathbf{R}^N}, \quad (3.23)$$

where $\mathbf{u}_I$ and $\mathbf{u}_J$ denote the unit vector of molecule I and J . This ODF was introduced and applied to a system of biphenyl rings.⁷⁸ While the choice of $\mathbf{u}$ vector is system dependent, it is intuitive to use the vector between the two CG sites that are within the same molecule in the COS mapping. By definition, perfectly aligned molecules at a certain distance have an ODF value of unity. In contrast, the ODF value converges to 0.5 when the system is randomly oriented.

The ODF calculated from the COS CG simulation is shown in Figure 3-5(b) and compared to that from the re-mapped FG trajectory. The COM model is not shown in Figure 3-5(b) since the vector between the COM CG sites lies in the aromatic plane and is thus irrelevant with respect to the orientational motions. The ODF calculated from the mapped all-atom configuration demonstrates a similar profile with that from the COS CG model, suggesting that the COS mapping also captures the orientation correlations as well as expected. Furthermore, when the pair distance is near 6 Å, both ODF functions have a value near one, indicating that neighboring benzene molecules are highly aligned in a π -stacked configuration.

Lastly, we examined the higher order correlations of the system by calculating the three-body distribution functions. As discussed earlier, we expected the COS CG model to correctly recapitulate the three-body correlations due to an orientational preference of the benzene ring. In practice, we considered the molecule-wise triplet of the CG systems within the first coordination shell $r_{\text{cut}} = 8.0$ Å. In other words, the three-body distribution function $P(\theta)$, i.e., angular distribution function (ADF) in this case, is a normalized histogram over triplets of re-mapped single-site CG sites I, J, K in a similar manner to Figure 3-5(a):

$$P(\theta) = \frac{1}{N_\theta} \left\langle \sum_I \sum_{J \neq I} \sum_{K > J} \delta(\theta - \theta_{JK}) \right\rangle_{r < r_{\text{cut}}} \quad (3.30)$$

It is clear from Figure 3-5(c) that the COS mapping is able to precisely reproduce the three-body structure of the benzene system within the first coordination shell.

B. Toluene

For toluene, the effective CG interactions illustrated in Figure 3-6 denote that the two CG types in the COM model are no longer degenerate (due to the substitution of a $-\text{CH}_3$ group) while the two COS CG sites still show identical behavior.

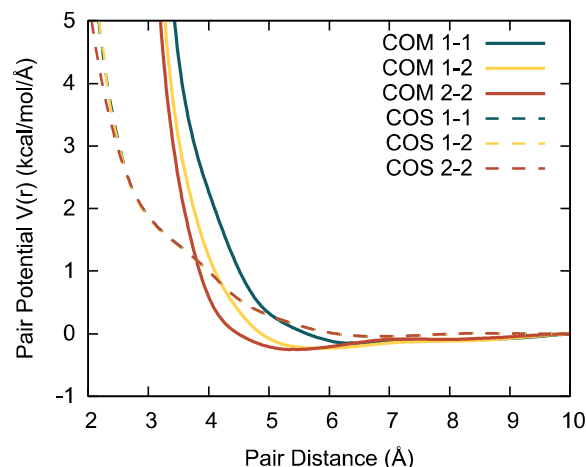


Figure 3-6: Effective pair potentials of toluene CG models obtained from the MS-CG method for two different CG mapping schemes: center of mass (solid lines) and center of symmetry (dotted lines). Only the COS mapping shows an equivalent interaction among different types (1-1, 1-2, and 2-2 pairs), while the COM interactions differ by CG pairs.

The different potential profiles observed from the COM model appear to originate from size effects or steric interactions. In toluene, the CG bead corresponding to type 1 ($\text{-C}_4\text{H}_5$) is larger than type 2 ($\text{-C}_3\text{H}_3$). Likewise, the solid lines in Figure 3-6 where the 1-1 interaction shows the strongest repulsion at short distances exhibits an analogous trend. By using Lennard-Jones like interactions as an exemplary model, this behavior resembles the following situation: $\sigma_{11} > \sigma_{22}$ and $\epsilon_{11} \approx \epsilon_{22}$ where σ_{12} is in a close agreement to the combined value from the general Lorentz-Berthelot rule.^{79, 80}

Likewise, the molecule-wise RDF, ODF, and ADF that corresponds to the COM of toluene are calculated in Figure 3-7. As expected, the COM model yields a higher $g(r)$ value at the first minimum near 8 Å (lower inset, Figure 3-7(a)) that varies from the mapped all-atom structure while the COS model is able to correctly recapitulate the FG structure. Similarly, ODF

profiles of the mapped all-atom and COS models are in agreement as seen in Figure 3-7(b). ADF functions with a cutoff distance at the first coordination shell $r_{\text{cut}} = 8.4 \text{ \AA}$ further demonstrate that the COS model also recapitulates the trend seen in the mapped all-atom model. Compared to benzene, the deviations of the COM model from the FG structure is smaller, and this can be understood due to the different CG types ($-\text{C}_4\text{H}_5$ and $-\text{C}_3\text{H}_3$ beads) in toluene. Thus, the center of molecule slightly deviates from the center of the ring, resulting in less degenerate configurations. Despite such small differences, the COS model still corrects the angular profile, especially at 70° - 110° . Altogether, our designed model faithfully reproduces both the radial (two-body) and many-body correlations from the atomistic model.

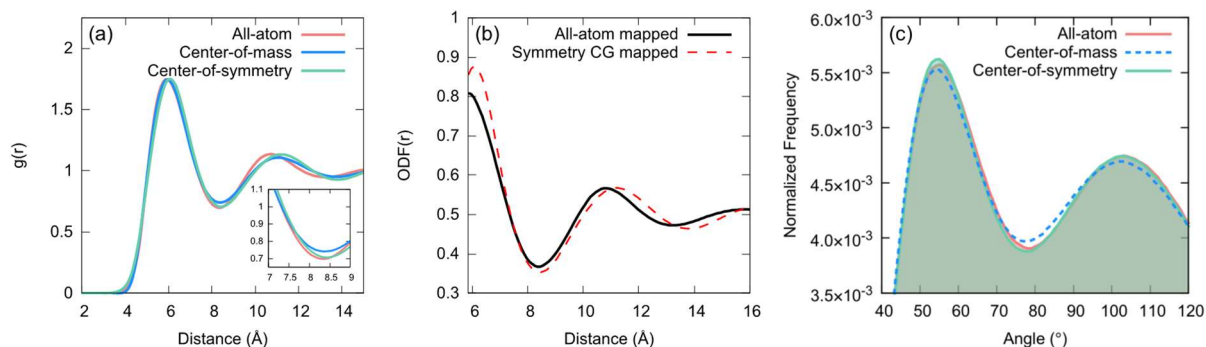


Figure 3-7: Molecule-wise (i.e., re-coarse-grained to one site per molecule) structure analysis of liquid toluene by calculating the (a) radial distribution function (RDF) of the mapped-all atom (red), center of mass (blue), and center of symmetry (green) models where magnified comparison is shown in a lower right inset, (b) orientation distribution function (ODF) of the mapped-all atom (black solid line) and center of symmetry (red dotted line) models, and (c) three-body angular distribution function (ADF) of the CG molecules of the mapped-all atom (red line), center of mass (blue dotted line), and center of symmetry (green line) as demonstrated in Eq. (3.24).

C. Benzene and toluene mixtures

As a final example, we constructed a mixed system of benzene and toluene molecules. We hypothesized that the CG interactions from the COS mapping would be more transferable between bulk and mixture phases of benzene and toluene. Our reasoning is that the CG interactions in this system are based on an aromatic ring as a structural motif. Therefore, we envisaged that the COS CG interactions are invariant under different phases while reproducing the correct structural properties since the COS mapping conserves the symmetry of each molecule. The reference FG system was composed of 500 benzene and 500 toluene molecules where the initial structure for the FG simulations was generated by the same protocol described in Sec. II. Simulation details were the same as the bulk cases including the use of the OPLS-AA force field and thermostat/barostat conditions. During the constant NVT simulation, the equilibrated box size of the mixture system was 55.38 Å. The final snapshot of the all-atom configuration is shown in Figure 3-8(a) and the corresponding final CG configuration using COS mapping is depicted in Figure 3-8(b) with four different CG sites (red/blue beads for benzene and orange/green beads for toluene).

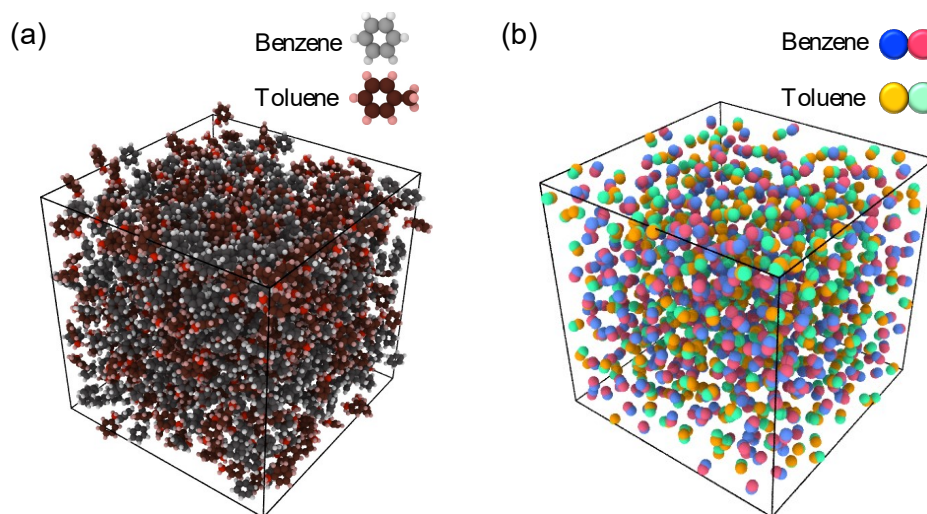


Figure 3-8: Schematic description of the equilibrated benzene/toluene mixture system in (a) all-

Figure 3-8, continued: atom resolution (toluene is depicted as dark brown color and benzene is depicted as grey color), (b) COS mapped resolution (mapped molecules are shown at upper right corner)

As expected from the shared aromatic (phenyl) ring motif, we observed that the two molecular components did not create a segregated phase. From the simulated CG trajectories, we first calculated the molecule-wise RDF by re-mapping the two-site COS and COM models to a single COM site. Distributions of the three molecule pairs (benzene-benzene, benzene-toluene, and toluene-toluene) are demonstrated in Figure 3-9. We also included a single-site COM CG model to compare the quality of the CG model with different resolution and purely spherical symmetry.

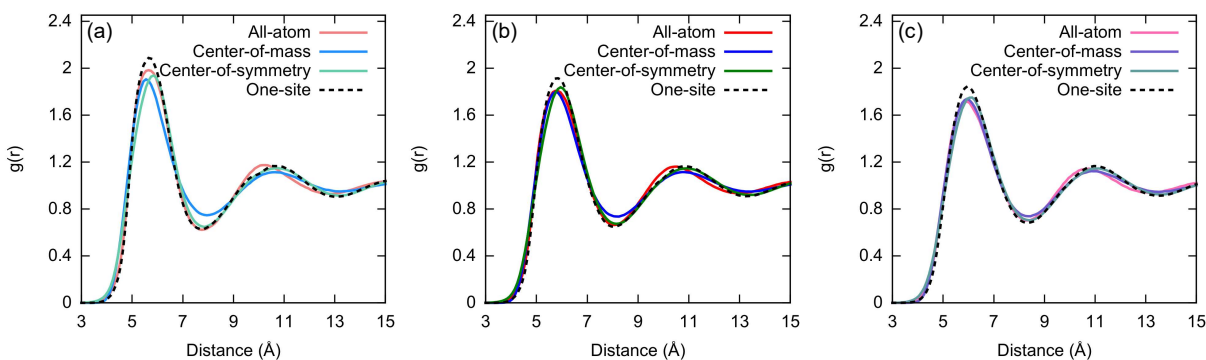


Figure 3-9: Radial distribution function $g(r)$ in the benzene/toluene mixture. Here, $g(r)$ functions are molecule-wise (i.e., re-coarse-grained to one site per molecule) function of the mapped-all atom (red), two-site center of mass (blue), and center of symmetry (green) CG models with the one-site CG model (black dot) for (a) benzene-benzene, (b) benzene-toluene, (c) toluene-toluene molecule pairs.

As anticipated, the CG sites via the COS mapping provide correct molecule-wise pair structures even in a mixed phase. While the two-site COM models deviate from the all-atom case at the first minimum of the RDFs, the COS models correctly capture this feature. Interestingly, this region is well described by the single-site COM model. However, the single-site model gives a poor description of the first maximum with a strongly over-structured peak, presumably arising from the mismatch of symmetry. Furthermore, by comparing the general shape of the RDFs from the COS mapping to that of bulk liquids (Figure 3-5 and 3-7), one might expect the transferability of the CG potential from the bulk phase to the mixture phase, following Henderson's theorem⁸¹ and the basic concept of iterative Boltzmann inversion.⁸² We note that solving the MS-CG equation does not guarantee exact pairwise structures, but provides CG interactions that satisfy the generalized Yvon-Born-Green (YBG) equations.^{10, 83, 84} The transferability of the CG interactions for different mapping schemes was examined by comparing the effective CG interactions between the mixed system and bulk liquids as summarized in Figure 3-10. The COM interactions shown in Figure 3-10(a) and (b) are obtained from the single-site COM model while the COS interactions shown in Figure 3-10(c) and (d) are from the site-site interactions due to the homogeneity of COS CG types.

Clearly, the COM interaction is not as transferable between bulk and mixture phases as seen by the different minimum predicted in the two phases. This trend is in accord with previously reported differences in pair potentials between bulk and mixed phases using single-site MS-CG models.^{31, 34} To reiterate, performing the MS-CG force-matching procedure in the mixed case does not guarantee identical pair interactions as that of the bulk, even though the single-site CG model nicely reproduced the first minima in the RDF of both phases. Interestingly, almost an exact agreement between the interactions from the bulk and mixture phases is identified from the

COS models (Figure 3-10(c) and 3-10(d)), thereby confirming our hypothesis about enhanced transferability.

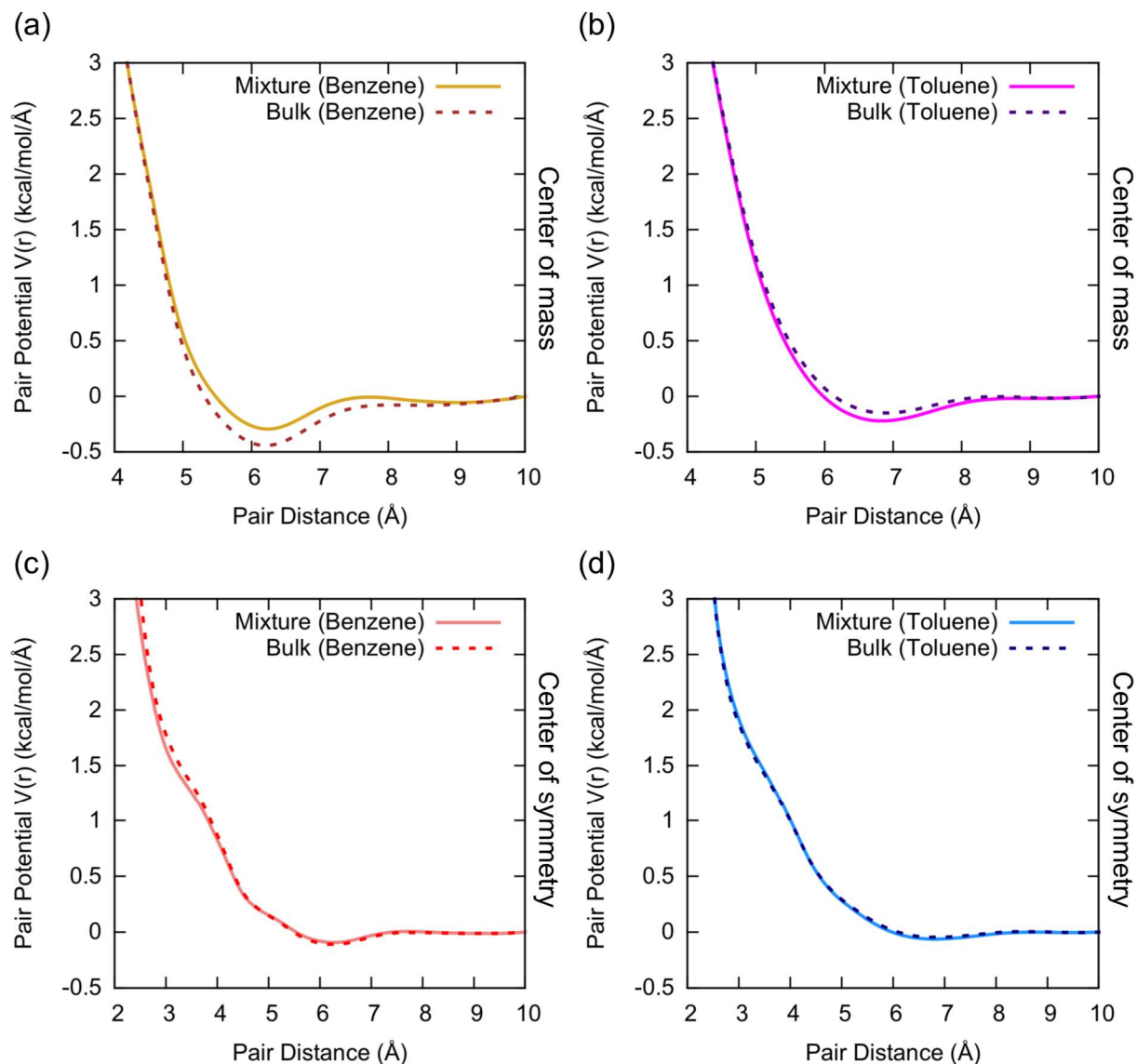


Figure 3-10: Transferability of the COS CG interactions between the bulk liquid (benzene, toluene) and its mixed phase compared to the COM CG interactions. Effective CG pair potentials from (a) the one-site COM model of benzene in the bulk phase (brown dotted) and the mixed phase (gold line), (b) the one-site COM model of toluene in the bulk phase (purple dotted) and the mixed phase (magenta line), (c) the COS model of benzene in the bulk phase (red dotted) and

Figure 3-10, continued: the mixed phase (red line), and (d) the COS model of toluene in the bulk phase (blue dotted) and the mixed phase (blue line) with almost identical shapes.

In order to further examine the transferability of the CG interactions, we introduce a statistical measure to quantify the similarity of the many-body PMFs from bulk and mixture phases. Previously, we utilized conventional distance metrics such as the Euclidean distance between the PMFs $\|\mathbf{W}_1(\mathbf{R}^N) - \mathbf{W}_2(\mathbf{R}^N)\| = \sqrt{\sum_{I,J} (\mathbf{W}_1(\mathbf{R}_{IJ}) - \mathbf{W}_2(\mathbf{R}_{IJ}))^2}$ or the distance from smoother manifold by integrating the PMF along the pair distance $\int_{\infty}^{r_{\text{cut}}} \mathbf{W}_1(\mathbf{r}) d\mathbf{r}$ corresponding to area under curve.³⁴ These distance metrics can be easily utilized to evaluate the similarity of different PMFs within the same mapping scheme such as a comparison of COM CG models between bulk liquid and interface. However, these distance metrics are not applicable to compare the PMFs from different mapping schemes due to different interaction ranges as well as lacking in physical interpretation of CG system. As a universal metric that can be applied to different mapping schemes (COM and COS) and systems (bulk and mixture), here we utilize the relative entropy S_{rel} , also known as Kullback-Leibler divergence,⁸⁵ to characterize differences in CG interactions between different CG systems.

$$S_{\text{rel}} = \int P_{\text{bulk}}(\mathbf{r}) \ln \frac{P_{\text{bulk}}(\mathbf{r})}{P_{\text{mix}}(\mathbf{r})} d\mathbf{r} . \quad (3.31)$$

As an analogy to the systematic CG approaches by minimizing the relative entropy,^{6, 48-50} S_{rel} provides information loss incurred upon different system (or state points). S_{rel} values for

different mapping schemes are readily obtained by $P_{\text{bulk}}(\mathbf{r}) = \frac{\exp(-\beta \mathbf{W}_{\text{bulk}}(\mathbf{r}))}{\int d\mathbf{r} \exp(-\beta \mathbf{W}_{\text{bulk}}(\mathbf{r}))}$ and

$$P_{\text{mix}}(\mathbf{r}) = \frac{\exp(-\beta \mathbf{W}_{\text{mix}}(\mathbf{r}))}{\int d\mathbf{r} \exp(-\beta \mathbf{W}_{\text{mix}}(\mathbf{r}))}. \text{ By its form given in Eq. (3.25), it is clear that the relative}$$

entropy is not confined in different interaction ranges because its contribution is normalized in

$\ln \frac{P_{\text{bulk}}(\mathbf{r})}{P_{\text{mix}}(\mathbf{r})}$ term. Therefore, comparing the relative entropy for different mapping can be a

potentially universal way to score the transferability of a CG model in different state points.

Table 3-I shows the relative entropy values of benzene and toluene molecules

$$S_{\text{rel}} = D_{\text{KL}}(P_{\text{mix}}(\mathbf{r}) \| P_{\text{bulk}}(\mathbf{r})) \text{ from bulk to mixture.}$$

Table 3-I: Evaluating the transferability of COM and COS CG models by employing relative entropy (or Kullback-Liebler divergence) as a measure. For each case, the relative entropy was calculated between the CG PMF obtained in bulk phase and mixture phase.

Type	Relative Entropy Values		
	COM ($S_{\text{rel}}^{\text{COM}}$)	COS ($S_{\text{rel}}^{\text{COS}}$)	$S_{\text{rel}}^{\text{COM}} / S_{\text{rel}}^{\text{COS}}$
Benzene	4.428×10^{-3}	4.294×10^{-4}	10.31
Toluene	1.345×10^{-3}	9.847×10^{-5}	13.66

In contrast to the COM mapping, the COS mapping yields much smaller relative entropy values for both benzene and toluene systems. In a quantitative manner, the COS model shows 10-13 times closer agreement, confirming a better transferability between bulk and mixture systems. To clarify, the additional basis (in this case, the normal vector of the aromatic ring) increases the ability to express the effective CG interaction via COS mapping. Furthermore, as a quantitative measure, the relative entropy calculation is expected to further determine the optimal COS model. Practically, for models with different separation distances, one could calculate the Kullback-Leibler divergence of the PMFs from the parameterized CG models with the target atomistic

sampling, as demonstrated in Table 3-I. Then, the optimal separation distance can be chosen as the distance that gives the smallest Kullback-Leibler divergence. As a natural extension from our interpretation, it would be of interest to understand the role of additional basis sets in force-matching to systematically improve the reproducibility and transferability of the CG models.

IV. Concluding Remarks

In summary, this paper aims to provide a systematic framework for the construction of CG models using virtual sites, i.e., fictitious particles designed to capture essential physical correlations of the system at the CG resolution. First, we have introduced a statistical mechanical theory of MS-CG involving virtual CG sites based on the principle of maximum entropy. The extended thermodynamic consistency relationship from the MS-CG formalism is satisfied if the overall COM of each molecule is expressed as a linear combination of the configurations of virtual sites. In such cases, the overall forces are partitioned among the virtual CG types based on the coefficient used to construct the overall COM. Namely, this framework opens up a new way of designing CG models with virtual particles based on the MS-CG methodology. Depending on the system of interest, one can design a new CG model and explicitly consider the vectors between virtual sites as an additional basis set to the MS-CG algorithm without any additional complications.

A new center of symmetry (COS) CG mapping scheme involving virtual sites is also presented in this work. The COS mapping is designed to preserve the symmetry from the finer resolution through the use of virtual sites, which we demonstrate for molecules that share aromatic ring motifs with planar symmetry. In this case, the imposed normal vector formed by the virtual sites can capture the π - π interaction or quadrupolar interaction for benzene and toluene systems. In turn, the COS CG models can provide very accurate delineation of radial,

orientational, and angular distributions, as well as transferable interactions between bulk and mixed liquid phases.

The aforementioned reproducibility and transferability of the COS mapping scheme further suggests potential applicability of shape-based or symmetry-based CG models⁴⁴⁻⁴⁷ through the MS-CG approach. Defining CG mapping operators and utilizing such bottom-up CG models in more complicated systems such as biomolecules is still challenging because various motions with different timescales are entangled in the bead.⁴ Nevertheless, the current framework can effectively model some biomolecules that have a certain shape or form such as lipids,⁸⁶⁻⁸⁸ as the symmetry of the CG sites might play an important role in understanding the physical picture of the CG models. This general approach may also facilitate the construction of more transferable bottom-up CG models for sophisticated bio and macromolecules.

APPENDIX A: Discussion of the limit: COS model converges to a single-site COM model when $l \rightarrow 0$

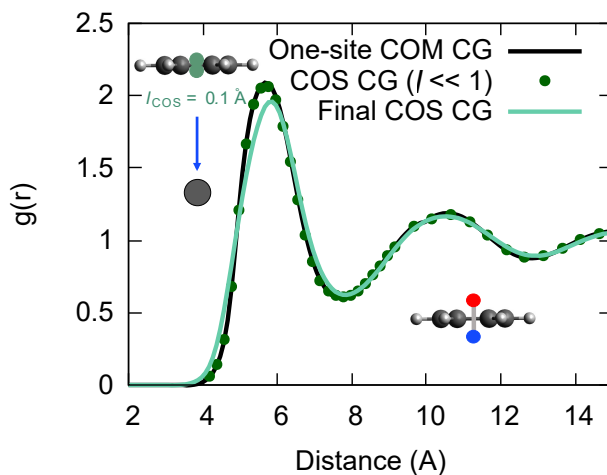


Figure 3-A1: Molecule-wise radial distribution function of liquid benzene for three different models: the single-site COM model (black line) and the two-site COS models. Here, the COS schemes used in this work are shown as a turquoise line and compared with the limit (green dots)

Figure 3-A1, continued: when the separation distance $l_{\text{cos}}(=l)$ is too small (0.1 Å). In this limit, it is immediately apparent that the COS mapping converges to the single-site COM mapping.

APPENDIX B: Derivation of $\Omega(\tilde{\mathbf{R}}^{\tilde{N}})$ and Eq. (3.11)

The number of state measure $\Omega(\tilde{\mathbf{R}}^{\tilde{N}})$ in general is a non-linear function of the COM configurations, which is hard to obtain an explicit analytical form. Luckily, for the simple two COS beads to one COM case, we can explicitly evaluate the integral. By doing so, we change the integration variable from the COS beads coordinates $(\mathbf{R}_{I1}, \mathbf{R}_{I2})$, to their center of mass and half relative distances $(\bar{\mathbf{R}}_I, \Delta\mathbf{R}_I)$ where $\Delta\bar{\mathbf{R}}_I = (\mathbf{R}_{I1} + \mathbf{R}_{I2})/2$ and $\Delta\mathbf{R}_I = (\mathbf{R}_{I1} - \mathbf{R}_{I2})/2$. Then the measure function can be explicitly integrated,

$$\begin{aligned}
\Omega(\tilde{\mathbf{R}}^{\tilde{N}}) &= \int d\mathbf{R}^N \delta(\tilde{\mathbf{R}}^{\tilde{N}} - \tilde{M}(\mathbf{R}^N)) \\
&= \prod_{I=1}^{\tilde{N}} \int d\mathbf{R}_{I1} d\mathbf{R}_{I2} \delta\left(\frac{\mathbf{R}_{I1} + \mathbf{R}_{I2}}{2} - \tilde{\mathbf{R}}_I\right) \\
&= \prod_{I=1}^{\tilde{N}} \int d\bar{\mathbf{R}}_I d\Delta\mathbf{R}_I \left| \frac{\partial(\mathbf{R}_{I1}, \mathbf{R}_{I2})}{\partial(\bar{\mathbf{R}}_I, \Delta\mathbf{R}_I)} \right| \delta(\bar{\mathbf{R}}_I - \tilde{\mathbf{R}}_I) \quad (3.B1) \\
&= \prod_{I=1}^{\tilde{N}} 2 \int d\Delta\mathbf{R}_I \int d\bar{\mathbf{R}}_I \delta(\bar{\mathbf{R}}_I - \tilde{\mathbf{R}}_I) = (2V)^{\tilde{N}}
\end{aligned}$$

Except for the perspective of maximum entropy principle, one can derive the Eq. (11) from the perspective of consistent observation. We consider any operator as a function of the COM information (conventionally denoted as $\mathbf{R}_c^N$, but as $\tilde{\mathbf{R}}^{\tilde{N}}$ in this work). The expectation value of such an operator for the atomistic ensemble as well as the CG virtual beads ensemble should remain identical. Therefore, the expectation value for the COM ensemble is further expressed as,

$$\begin{aligned}
\langle O(\tilde{\mathbf{R}}^{\tilde{N}}) \rangle &= \int d\mathbf{r}^n \rho(\mathbf{r}^n) O(M(\mathbf{r}^n)) \\
&= \frac{1}{(2V)^N} \int d\mathbf{r}^n \rho(\mathbf{r}^n) O(M(\mathbf{r}^n)) \int d\mathbf{R}^N \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) \\
&= \frac{1}{(2V)^N} \int d\mathbf{R}^N O(\tilde{M}(\mathbf{R}^N)) \int d\mathbf{r}^n \rho(\mathbf{r}^n) \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) , \\
&= \int d\mathbf{R}^N \rho_{CG}(\mathbf{R}^N) O(\tilde{M}(\mathbf{R}^N))
\end{aligned} \tag{3.B2}$$

Similar to the evaluation of the measure functions, the term $(2V)^N$ comes from changing the integral measure from $d\mathbf{R}^N$ to $d\tilde{\mathbf{R}}^{\tilde{N}} d\Delta\mathbf{R}^{\tilde{N}}$. The integration of the half relative distance between virtual sites, $\Delta\mathbf{R}^{\tilde{N}}$, gives the volume terms, and the Jacobian of the transformation gives the factor of 2. In order to satisfy the consistency, $\rho_{CG}(\mathbf{R}^N)$ should be expressed as

$$\rho_{CG}(\mathbf{R}^N) = \frac{1}{(2V)^N} \int d\mathbf{r}^n \rho(\mathbf{r}^n) \delta(M(\mathbf{r}^n) - \tilde{M}(\mathbf{R}^N)) , \tag{3.B3}$$

giving Eq. (3.11).

APPENDIX C: Derivation of Eq. (3.15)

Eq. (3.15) is derived based on the change of variables for derivatives. Notice that the following equation holds due to the linearity of the two mapping operators,

$$\frac{\partial}{\partial \mathbf{R}_I} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) = -\frac{\chi_{\alpha I}}{c_{\alpha i}} \frac{\partial}{\partial \mathbf{r}_i} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) . \tag{3.C1}$$

Without loss of generality, any fine-grained atom i belongs to the center of mass α . Therefore, assigning a weighting factor $d_{\alpha i}$ for each side of the equation above and summing over the terms result in

$$\frac{\partial}{\partial \mathbf{R}_I} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) \sum_{i \in \alpha} d_{\alpha i} = - \sum_{i \in \alpha} d_{\alpha i} \frac{\chi_{\alpha I}}{c_{\alpha i}} \frac{\partial}{\partial \mathbf{r}_i} \delta(M_\alpha(\mathbf{r}^n) - \tilde{M}_\alpha(\mathbf{R}^N)) . \tag{3.C2}$$

Once the choice of the weighting factors satisfies $\sum_{i \in \alpha} d_{\alpha i} = 1$, we arrive at Eq. (3.15).

APPENDIX D: Velocity Distribution

Figure D1 shows the velocity distribution of the mapped virtual CG beads. The depicted distribution follows the Maxwell distribution and gives back the atomistic temperature consistently at 300 K by setting the mass of the virtual CG beads according to Eq. (3.29).

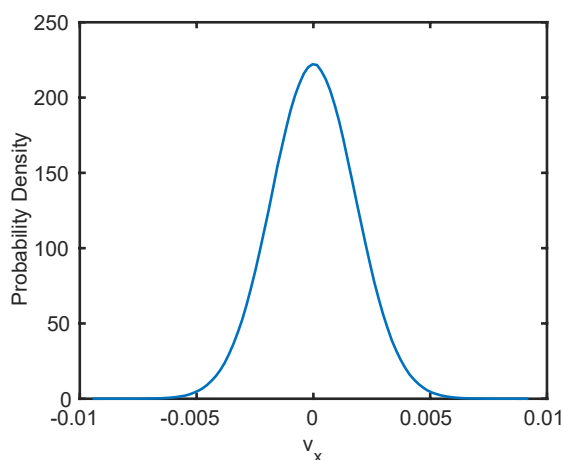


Figure 3-D1: Probability distribution function of the x-component of virtual CG bead velocities v_x from the mapped atomistic trajectories, giving $T = 300$ K.

Reference

1. H. A. Scheraga, M. Khalili and A. Liwo, *Annu. Rev. Phys. Chem.* **58**, 57 (2007).
2. A. Liwo, C. Czaplewski, S. Ołdziej and H. A. Scheraga, *Curr. Opin. Struct. Biol.* **18**, 134 (2008).
3. G. A. Voth, *Coarse-graining of condensed phase and biomolecular systems*. (CRC press, 2008).
4. M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.* **42**, 73 (2013).
5. W. Noid, *J. Chem. Phys.* **139**, 090901 (2013).
6. M. S. Shell, *Adv. Chem. Phys.*, 395 (2015).
7. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev.* **116**, 7898 (2016).

8. S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
9. S. Izvekov and G. A. Voth, *J. Chem. Phys.* **123**, 134105 (2005).
10. W. Noid, J.-W. Chu, G. S. Ayton and G. A. Voth, *J. Phys. Chem. B* **111**, 4116 (2007).
11. W. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
12. W. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen and G. A. Voth, *J. Chem. Phys.* **128**, 244115 (2008).
13. T. T. Foley, M. S. Shell and W. Noid, *J. Chem. Phys.* **143**, 12B601 (2015).
14. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. De Vries, *J. Phys. Chem. B* **111**, 7812 (2007).
15. L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, *J. Chem. Theory Comput.* **4**, 819 (2008).
16. S. J. Marrink and D. P. Tieleman, *Chem. Soc. Rev.* **42**, 6801 (2013).
17. A. Amadei, A. Linssen and H. J. Berendsen, *Proteins: Struct., Funct., Bioinf.* **17**, 412 (1993).
18. A. Kitao and N. Go, *Curr. Opin. Struct. Biol.* **9**, 164 (1999).
19. H. J. Berendsen and S. Hayward, *Curr. Opin. Struct. Biol.* **10**, 165 (2000).
20. Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner and G. A. Voth, *Biophys. J.* **95**, 5073 (2008).
21. Z. Zhang, J. Pfaendtner, A. Grafmüller and G. A. Voth, *Biophys. J.* **97**, 2327 (2009).
22. S. Lafon and A. B. Lee, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1393 (2006).
23. A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti and I. G. Kevrekidis, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 13597 (2010).
24. A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis and P. G. Debenedetti, *Chem. Phys. Lett.* **509**, 1 (2011).
25. M. Ceriotti, G. A. Tribello and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13023 (2011).

26. G. A. Tribello, M. Ceriotti and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5196 (2012).
27. E. C. Allen and G. C. Rutledge, *J. Chem. Phys.* **128**, 154115 (2008).
28. S. Whitelam, C. Rogers, A. Pasqua, C. Paavola, J. Trent and P. L. Geissler, *Nano Lett.* **9**, 292 (2008).
29. L. Larini, L. Lu and G. A. Voth, *J. Chem. Phys.* **132**, 164107 (2010).
30. I. M. Ilie, W. K. den Otter and W. J. Briels, *J. Chem. Phys.* **144**, 085103 (2016).
31. J. F. Dama, J. Jin and G. A. Voth, *J. Chem. Theory Comput.* **13**, 1010 (2017).
32. J. W. Wagner, T. Dannenhoffer-Lafage, J. Jin and G. A. Voth, *J. Chem. Phys.* **147**, 044113 (2017).
33. Z. Cao and G. A. Voth, *J. Chem. Phys.* **143**, 243116 (2015).
34. J. Jin and G. A. Voth, *J. Chem. Theory Comput.* **14**, 2180 (2018).
35. B. J. Berne and P. Pechukas, *J. Chem. Phys.* **56**, 4213 (1972).
36. J. Gay and B. Berne, *J. Chem. Phys.* **74**, 3316 (1981).
37. P. A. Golubkov and P. Ren, *J. Chem. Phys.* **125**, 064103 (2006).
38. P. A. Golubkov, J. C. Wu and P. Ren, *Phys. Chem. Chem. Phys.* **10**, 2050 (2008).
39. J. Wu, X. Zhen, H. Shen, G. Li and P. Ren, *J. Chem. Phys.* **135**, 10B617 (2011).
40. P. G. Lafond and S. Izvekov, *J. Chem. Theory Comput.* **12**, 5737 (2016).
41. S. J. Ganesan and S. Matysiak, *J. Chem. Theory Comput.* **10**, 2569 (2014).
42. M. Li, F. Liu and J. Z. Zhang, *J. Chem. Theory Comput.* **12**, 6147 (2016).
43. N. J. Dunn, T. T. Foley and W. G. Noid, *Acc. Chem. Res.* **49**, 2832 (2016).
44. A. Arkhipov, P. L. Freddolino and K. Schulten, *Structure* **14**, 1767 (2006).
45. A. Arkhipov, Y. Yin and K. Schulten, *Biophys. J.* **95**, 2806 (2008).
46. J. M. Grime and G. A. Voth, *Biophys. J.* **103**, 1774 (2012).

47. J. M. Grime, J. F. Dama, B. K. Ganser-Pornillos, C. L. Woodward, G. J. Jensen, M. Yeager and G. A. Voth, *Nat. Commun.* **7** (2016).
48. M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).
49. A. Chaimovich and M. S. Shell, *Phys. Rev. E* **81**, 060104 (2010).
50. A. Chaimovich and M. S. Shell, *J. Chem. Phys.* **134**, 094112 (2011).
51. L. Lu, S. Izvekov, A. Das, H. C. Andersen and G. A. Voth, *J. Chem. Theory Comput.* **6**, 954 (2010).
52. P. Hobza, H. L. Selzle and E. W. Schlag, *J. Am. Chem. Soc.* **116**, 3500 (1994).
53. P. Hobza, H. L. Selzle and E. W. Schlag, *J. Phys. Chem.* **100**, 18790 (1996).
54. Y. C. Park and J. S. Lee, *J. Phys. Chem. A* **110**, 5091 (2006).
55. R. Podeszwa, R. Bukowski and K. Szalewicz, *J. Phys. Chem. A* **110**, 10345 (2006).
56. M. Pitonak, P. Neogady, J. Rezac, P. Jurecka, M. Urban and P. Hobza, *J. Chem. Theory Comput.* **4**, 1829 (2008).
57. C. D. Sherrill, B. G. Sumpter, M. O. Sinnokrot, M. S. Marshall, E. G. Hohenstein, R. C. Walker and I. R. Gould, *J. Comput. Chem.* **30**, 2187 (2009).
58. C. A. Hunter and J. K. Sanders, *J. Am. Chem. Soc.* **112**, 5525 (1990).
59. J. G. Vinter, *J. Comput.-Aided Mol. Des.* **8**, 653 (1994).
60. O. R. Lozman, R. J. Bushby and J. G. Vinter, *J. Chem. Soc. Perkin Trans. 2*, 1446 (2001).
61. S. Tsuzuki, K. Honda, T. Uchimaru, M. Mikami and K. Tanabe, *J. Am. Chem. Soc.* **124**, 104 (2002).
62. A. T. Macias and A. D. MacKerell, *J. Comput. Chem.* **26**, 1452 (2005).
63. T. Lafitte, C. Avendaño, V. Papaioannou, A. Galindo, C. S. Adjiman, G. Jackson and E. A. Müller, *Mol. Phys.* **110**, 1189 (2012).
64. W. L. Jorgensen, E. R. Laird, T. B. Nguyen and J. Tirado-Rives, *J. Comput. Chem.* **14**, 206 (1993).
65. W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225 (1996).

66. R. W. Hockney and J. W. Eastwood, *Computer simulation using particles*. (CRC Press, 1988).
67. S. Plimpton, J. Comput. Phys. **117**, 1 (1995).
68. L. Martínez, R. Andrade, E. G. Birgin and J. M. Martínez, J. Comput. Chem. **30**, 2157 (2009).
69. W. Humphrey, A. Dalke and K. Schulten, J. Molec. Graphics **14**, 33 (1996).
70. S. Nosé, J. Chem. Phys. **81**, 511 (1984).
71. W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
72. H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).
73. A. Das, L. Lu, H. C. Andersen and G. A. Voth, J. Chem. Phys. **136**, 194115 (2012).
74. in *Multiscale Coarse-Graining Force-Matching (MSCGFM) v 1.7*
(<https://github.com/uchicago-voth/MSCG-release>).
75. J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, J. Comput. Phys. **23**, 327 (1977).
76. A. V. Sinitskiy, M. G. Saunders and G. A. Voth, J. Phys. Chem. B **116**, 8363 (2012).
77. N. Guttenberg, J. F. Dama, M. G. Saunders, G. A. Voth, J. Weare and A. R. Dinner, J. Chem. Phys. **138**, 094111 (2013).
78. H. Meyer, O. Hahn and F. Müller-Plathe, J. Phys. Chem. B **103**, 10591 (1999).
79. H. Lorentz, Ann. Phys. **248**, 127 (1881).
80. D. Berthelot, C. R. Hebd. Séances Acad. Sci. **126**, 1703 (1898).
81. R. Henderson, Phys. Lett. A **49**, 197 (1974).
82. D. Reith, M. Pütz and F. Müller-Plathe, J. Comput. Chem. **24**, 1624 (2003).
83. J. Mullinax and W. Noid, J. Phys. Chem. C **114**, 5661 (2009).
84. J. Mullinax and W. Noid, Phys. Rev. Lett. **103**, 198104 (2009).
85. S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
86. A. Srivastava and G. A. Voth, J. Chem. Theory Comput. **9**, 750 (2012).

- 87. A. Srivastava and G. A. Voth, *J. Chem. Theory Comput.* **10**, 4730 (2014).
- 88. J. J. Madsen, A. V. Sinitskiy, J. Li and G. A. Voth, *J. Chem. Theory Comput.* **13**, 935 (2017).

Chapter 4: Mesoscopic Coarse-grained Representations of Fluids Rigorously Derived from Atomistic Models

This chapter is Reproduced from J. Chem. Phys. 149, 044104 (2018), with the permission of AIP Publishing.

Abstract

Mesoscopic models are widely used to study complex organization and transport phenomena in chemical and biological systems. Defining a rigorous procedure by which a mesoscopic CG representation for a fluid can be constructed from an atomistic fine-grained (FG) model is a long-standing question in the field. The connection of these CG models with the FG level of description, which might be built by coarse-grained (CG) mappings from the FG model, is often unclear. The present paper introduces a new CG mapping scheme that uses dynamically self-consistent smooth centroidal Voronoi tessellation to address this challenging problem. The new mapping scheme is applied to the coarse-graining of supercritical LJ fluid systems at different CG resolutions under both quiescent conditions and non-equilibrium shear flow. The method generates continuous, stable, and ergodic CG trajectories and quantitatively captures the slow collective motions of the underlying FG fluids. A parameterization of the CG models from the mapped CG trajectory is then developed based on Mori-Zwanzig formalism. The Generalized Langevin Equation describes the dynamics of CG variables and the parameterized result is shown to reproduce the structural and dynamical correlations of the CG system. The new dynamical mapping scheme and the parameterization protocol open up an avenue for direct bottom-up construction of mesoscopic models of fluids in a Lagrangian description.

I. Introduction

With the polymer models¹ of De Gennes, and dissipative particle dynamics (DPD)²⁻⁵ or smoothed particle hydrodynamics (SPH),⁶⁻⁹ fluid “blobs” are extensively used in modeling

complex fluids having interesting mesoscale structure, especially in chemical and materials engineering. These approaches are applicable to many systems, including polymers,¹⁰⁻¹³ colloids,^{14, 15} gels,¹⁶ fixed and fluid membranes,¹⁷⁻¹⁹ and blood.²⁰ Such blobs have physical interpretations as mesoscale groupings of local molecules or molecular fragments, but these interpretations do not extend down to explicit quantitative correspondences between the mesoscale blob entities and particular atomic-scale particles. The blobs' properties reflect average statistical features of ensembles of appropriate groups of molecules and molecule fragments embedded in typical environments rather than reflecting the specific molecular motions of particular groups.

In applications, this above approach is motivated based on laws of large numbers and renormalization-based universality arguments that ensure statistically typical blobs provide accurate approximations to specific collective molecular behavior. However, researchers continually push SPH, DPD, and polymer blob modeling to new regimes of physical complexity, regimes in which the simplest approximations lose typicality or universality. Therefore, as the motivation increases to push these models beyond where the most widely held intuition is sufficient for their design and interpretation, there is an increasing value to be able to specifically and explicitly link mesoscopic blob models to underlying atomistic motions and associated physics. Such a linkage can provide direct checks of the model assumptions, guide improvement of the model approximations, and generate new raw data for sharpening our intuition in poorly understood new regimes of physical complexity.

The desirability of explicit bottom-up constructions of supramolecular blob models has been accepted since at least the late 1990s, when Espanol first attempted to construct a Voronoi cell based coarse-grained (CG) model for fluids.²¹ For the systems with bonded atoms belonging to

definite groups where the atoms do not diffuse away from each other, like most polymer melts, the correspondence between the supramolecular blobs and fine-grained (FG) atoms is well defined by linear CG mapping operators, usually through the center of mass mapping. At the same time, a number of methods can be utilized to derive the interactions between the CG blobs under such situations.²²⁻²⁵ However, for the systems with unbonded fluid particles, the correspondence between the supramolecular blob model and FG particles is not trivial.²⁶ A fully satisfactory CG mapping should be able to define CG beads representing a set of unbonded fluid particles moving coherently. Flekkøy and Coveney published an influential derivation of DPD equations of motion from molecular motion based on identifying the supramolecular blobs with the cells of a fuzzy, stationary partition of space.^{27, 28} Eriksson *et al*^{29, 30} revisited the method and attempted to derive a DPD model for fluids using this projection. This derivation, however, did not enforce the consistency between the positions of CG blobs and the underlying fluid molecules. The assumptions and approximations of this work have led a series of further researchers to construct alternative bottom-up derivations that address specific undesirable features of Flekkøy and Coveney's initial approach.

Hadley and McCabe,³¹ as well as Izvekov and Rice,³² have proposed to group multiple fluid molecules into CG blobs through the K-means clustering algorithm. These algorithms start with defining a number of CG beads with initial positions, and then construct the Voronoi tessellation based on this initial guess. Fluid molecules within the Voronoi cells define the updated position of the CG beads as the center of mass of these FG molecules. A re-tessellation based on the updated CG beads is made again. An iteration of the two steps derives the centroidal Voronoi cells in which the CG bead position and the center of mass of the fluid molecules coincide with each other. More recently, Praprotnik and co-workers^{33, 34} introduce a new dynamic clustering

sharing similar spirit that bundles several water molecules into one single bead representation. Unfortunately, this type of mapping can not be expressed as an analytical function and as a consequence, it is hard to use the theory of coarse-graining to rigorously derive the CG interactions. And more problematically, as the FG system propagates, the molecules inside one particular centroidal Voronoi cell will likely forget their neighbors and diffuse apart. A frequent re-clustering is needed to keep the proximity between FG and CG particles; therefore, it is not possible to generate smooth and continuous trajectories of the CG beads following this strategy.

Instead of using Voronoi tessellation based grouping, Ayton et al.³⁵ proposed a sphere blob based approach in which a spherical CG blob moves according to the forces experienced by the center of mass of the FG molecules inside it. This approach achieves stable propagation of the CG trajectories and exhibits stable and even mass partitioning. However, the approach is only applicable to propagate a few CG blobs from an atomistic trajectory and at the same time the FG to CG correspondence in coordinate space is not directly enforced during the propagation.

Though these mappings have provided useful insight, neither Flekkøy and Coveney nor any other of these approaches have derived a continuous supramolecular mapping from molecular motion to interacting blob motion that maintains its physical integrity without periodic, discrete correction steps to approximately account for neglected mass or momentum transfer associated with molecular motion across blob boundaries.

The problem of deriving the dynamics of CG blob particles directly from microscopic dynamics that captures the physics of a fluid made up of *unbonded molecules* remains an important, yet challenging and open question in the theory of “bottom up” mesoscale modeling. As such, to the best of our knowledge the present article defines the first bottom-up

supramolecular mesoscale-atomistic trajectory correspondence scheme with a complete accounting for all mass transfer physics at every instant of time.

This paper is organized as follows: Section I, this Introduction, has provided background and motivation for the new CG mapping. Section II defines the mapping, which is a map between trajectories rather than single configurations, with reference to additional background material. Section III presents the equilibrium statics and dynamics of the mapped CG blobs from a simple fluid under quiescent condition and shear flow, revealing that the statics match DPD soft spheres well and that the dynamics have hydrodynamic consistency. A bottom up parameterization of a Generalized Langevin Equation based on the Mori-Zwanzig formalism³⁶⁻³⁹ is also discussed. Section IV concludes the article.

II. Mapping

A. Position Map

A suitable supramolecular map from atoms to blobs must have certain basic properties to capture the desired physics. First, the blob positions should be invariant under relabeling of molecules. Pure relabeling or exchange of identical species is invisible at the mesoscopic level. Second, it should be Galilean invariant. The blobs should transform exactly as the atoms do under rotations, translations, and choice of reference velocity. Third, it should be independent of explicit time, i.e., time-homogeneous, dependent only on instantaneous molecular properties. Fourth, it should be smoothly varying with molecular properties so that the blob velocities and accelerations are always finite, with no discrete steps (e.g., as in the work of Flekkøy and Coveney or the clustering based methods).

Given those constraints, the present map will re-apply the intuition behind the work of Flekkøy and Coveney as well as the approaches of Hadley *et al*³¹ and Izvekov *et al*³² to the

maximum extent possible. Our new approach will also use a smoothed Voronoi tessellation, as Flekkøy and Coveney did, and ensure that the Voronoi tessellation is a centroidal tessellation, as Hadley *et al* and Izvekov *et al* did. The synthesis of the two will be a smoothed centroidal Voronoi tessellation map.

Specifically, following Flekkøy and Coveney, we define a local proximity function $w(r)$ (their $s(r)$) of distance r , and assign each FG molecule or fragment i to each blob I with an ownership fraction

$$f_{li} = \frac{w(|\mathbf{R}_I - \mathbf{r}_i|)}{\sum_J w(|\mathbf{R}_J - \mathbf{r}_i|)} \equiv \frac{w_{li}}{\sum_J w_{ji}} \quad (4.1)$$

in terms of the set of $\mathbf{R}_I$, the positions of blobs I , $\mathbf{r}_i$, the position of molecule or fragment i , and where the sum in the denominator ranges over all neighboring J . This ensures that each i is fully owned and contributing equally to CG motion at all times, though its contribution to each individual blob may vary smoothly through time as the blobs and atoms move.

Next, following the work by Hadley *et al* and Izvekov *et al*, we will assign the centers of the blobs according to the center of mass of their owned atoms, but now weighted by the ownership fraction,

$$\mathbf{R}_I = \sum_i \frac{f_{li} m_i \mathbf{r}_i}{\sum_j f_{lj} m_j} \equiv \sum_i c_{li} \mathbf{r}_i \quad (4.2)$$

This satisfies all of the requirements listed in the first paragraph of this subsection. The above two steps for determining the blob positions are summarized schematically in Fig. 4-1.

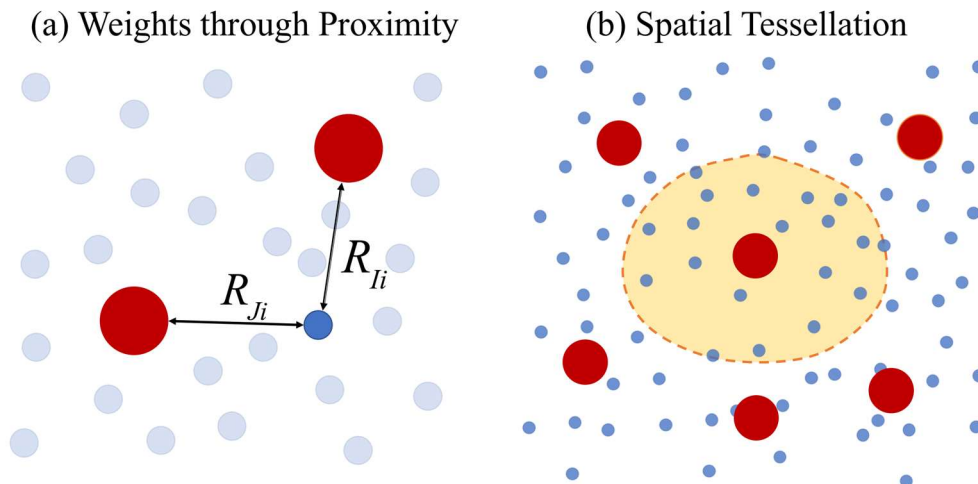


Figure 4-1: Schematic diagrams showing the two steps for determining the positions of CG blobs. Red circles denote the CG blobs and blue circles represent FG molecules, (a) The contribution of one particular FG molecule to its nearby CG blobs, i.e., the ownership fraction, is determined by their spatial distances. (b) The CG blob positions are determined by these ownership fractions.

However, this definition is circular: the $\mathbf{R}_I$ depends on the f_{li} and conversely the f_{li} depend on the $\mathbf{R}_I$. Moreover, there are many possible solutions. For any initial choice of $\mathbf{R}_I$, one can iterate the equations above to consistency, and in the continuum limit spatial translation symmetry guarantees there are infinite centroidal Voronoi tessellations to choose from. One can nonetheless go forward with this as a physically meaningful map so long as 1) the dynamics of the blobs from any initial self-consistent state are well defined and 2) these dynamics lead to a unique, initial-condition-independent equilibrium among the blobs. The next task is to derive the dynamics implied by this map.

B. Velocity Map

To derive the dynamics for the blobs from the self-consistent smoothed centroidal Voronoi tessellation position map, the time derivative of the blob position, which is a function defined in the FG phase space, can be evaluated through the Poisson bracket,

$$\begin{aligned}
\frac{d\mathbf{R}_I}{dt} &= \left\{ \sum_i c_{li} \mathbf{r}_i, H \right\} + \sum_i \frac{\partial c_{li} \mathbf{r}_i}{\partial t} \\
&= \sum_j \left(\frac{\partial \sum_i c_{li} \mathbf{r}_i}{\partial \mathbf{r}_j} \frac{\partial H}{\partial \mathbf{p}_j} - \frac{\partial \sum_i c_{li} \mathbf{r}_i}{\partial \mathbf{p}_j} \frac{\partial H}{\partial \mathbf{r}_j} \right) + \sum_i \frac{\partial c_{li} \mathbf{r}_i}{\partial t} \\
&= \sum_j \left(\sum_i \frac{\partial c_{li}}{\partial \mathbf{r}_j} \mathbf{r}_i + \sum_i c_{li} \delta_{ij} \right) \frac{\mathbf{p}_j}{m_j} \\
&= \sum_i c_{li} \mathbf{v}_i + \sum_j \sum_i \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{r}_j} \mathbf{v}_j
\end{aligned} \tag{4.3}$$

Unlike in coarse-graining with constant mapping coefficients, the blob velocities are not a linear combination of velocities using the same coefficients as the position map. One must include additional terms related to the partial derivatives of the coefficients with respect to the configurations. These terms are not included in previous attempts on this problem, though they account for crucial effects of FG mass transfer on momentum transfer at the CG resolution.

By construction, the c_{li} depend on the positions of both the blobs and FG particles, while the positions of the blobs $\mathbf{R}_I$ are also functions of FG particles, so the partial derivatives of c_{li} with respect to FG particle positions should have two contributions, and Eq. (4.3) can further be simplified,

$$\begin{aligned}
\frac{d\mathbf{R}_I}{dt} &= \sum_i c_{li} \mathbf{v}_i + \sum_j \sum_i \mathbf{r}_i \left(\frac{\partial c_{li}}{\partial \mathbf{r}_j} + \sum_J \frac{\partial c_{li}}{\partial \mathbf{R}_J} \frac{\partial \mathbf{R}_J}{\partial \mathbf{r}_j} \right) \cdot \mathbf{v}_j \\
&= \sum_i c_{li} \mathbf{v}_i + \sum_j \sum_i \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{r}_j} \cdot \mathbf{v}_j + \sum_j \sum_i \mathbf{r}_i \sum_J \frac{\partial c_{li}}{\partial \mathbf{R}_J} \frac{\partial \mathbf{R}_J}{\partial \mathbf{r}_j} \cdot \mathbf{v}_j, \\
&= \sum_i c_{li} \mathbf{v}_i + \sum_j \sum_i \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{r}_j} \cdot \mathbf{v}_j + \sum_J \sum_i \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{R}_J} \cdot \frac{d\mathbf{R}_J}{dt}
\end{aligned} \tag{4.4}$$

where the last equality holds because the positions of blobs are directly determined by the FG

particles, i.e. $\mathbf{R}$ is a function of FG particle positions $\mathbf{R} = \mathbf{R}(\mathbf{r})$. Therefore, $\sum_j \frac{\partial \mathbf{R}_I}{\partial \mathbf{r}_j} \cdot \frac{d\mathbf{r}_j}{dt} = \frac{d\mathbf{R}_I}{dt}$.

Equation (4.4) defines the mapping relationship between the blob velocity and FG particle velocities. It can be written in a simplified manner as,

$$\mathbf{V}_I = \sum_i c_{li} \mathbf{v}_i + \sum_i \sum_j \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{r}_j} \cdot \mathbf{v}_j + \sum_i \sum_j \mathbf{r}_i \frac{\partial c_{li}}{\partial \mathbf{R}_j} \cdot \mathbf{V}_j, \quad (4.5)$$

which is a linear system of equations that can be interpreted as a matrix equation, and solved at each timestep so long as the matrix multiplying the vector of $\mathbf{V}_I$ variables in that set of equations were invertible.

However, this form of the equations is not optimized to provide physical meaning and it can be simplified significantly. By introducing $\bar{\mathbf{V}}_I = \sum_i c_{li} \mathbf{v}_i$ and $\delta \mathbf{V}_I = \mathbf{V}_I - \bar{\mathbf{V}}_I$, and performing manipulations detailed in Appendix A, one can derive the more physically transparent system of equations:

$$\sum_j \left(\delta_{IJ} \mathbf{I} - \sum_i (\mathbf{r}_i - \mathbf{R}_I) c_{li} \frac{\partial \ln f_{li}}{\partial \mathbf{R}_j} \right) \cdot \delta \mathbf{V}_j = \sum_i \sum_j (\mathbf{r}_i - \mathbf{R}_I) c_{li} \frac{\partial \ln f_{li}}{\partial \mathbf{R}_j} \cdot (\bar{\mathbf{V}}_j - \mathbf{v}_i). \quad (4.6)$$

This set of equations for the deviations of $\mathbf{V}_I$ from the constant-map value $\bar{\mathbf{V}}_I$ takes the matrix form:

$$(\mathbf{I} - \mathbf{M}) \delta \mathbf{V} = \mathbf{V}^{dev}, \quad (4.7)$$

which can be expanded as Taylor series such that

$$\delta \mathbf{V} = \mathbf{V}^{dev} + \sum_{n=1}^{\infty} \mathbf{M}^n \mathbf{V}^{dev}, \quad (4.8)$$

The matrix $\mathbf{M}$ has 3x3 tensor elements $\mathbf{M}_{IJ} = \sum_i (\mathbf{r}_i - \mathbf{R}_I) c_{li} \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J}$ and $\mathbf{V}^{dev}$ has the elements

$$\mathbf{V}_I^{dev} = \sum_i \sum_J (\mathbf{r}_i - \mathbf{R}_I) c_{li} \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J} \cdot (\bar{\mathbf{V}}_J - \mathbf{v}_i). \text{ These equations have a unique solution so long as}$$

the absolute values of the eigenvalues of $\mathbf{M}$ are all less than one.

The above form more clearly shows that any Galilean transformation of the FG system transforms the CG system the same way. The vector $\mathbf{V}_I^{dev}$ is a result of the deviations of the velocities of the molecules from the constant-map velocities owned by their blobs and it is Galilean invariant. These deviations lead to changes in ownership of FG particles by the CG blobs, which then leads to changes in map coefficients and therefore to changes in blob velocity. In uniform flow, there is no deviation. Fine-scale thermal fluctuation around uniform flow leads to pair-correlated velocity fluctuations in neighboring blobs. Blob-scale velocity gradients in the molecular system lead to deviations that represent momentum transfer between neighboring blobs.

The deviations in blob velocity due to FG motion have their own consequences for changes in blob ownership, leading to further compensating changes in blob velocity from the constant-map expectation. The tensor $\mathbf{M}$ describes how an infinitesimal set of deviations in one blob velocity causes deviations in other blob velocities; the $\mathbf{M}_{IJ}$ are nonzero only if blob I and J overlap in their ownership of molecules, i.e., if they are neighbors in the smoothed Voronoi tessellation. Each $\mathbf{M}_{IJ}$ quantifies the dynamical effect of mass transfer between two neighboring blobs as they move relative to one another.

With the above derivation in hand, one has an explicit mapping from FG trajectories to CG blob trajectories so long as $\mathbf{I} - \mathbf{M}$ is invertible throughout the entire trajectory. Given any FG

trajectory with positions and velocities $\mathbf{r}_i(t)$ and $\mathbf{v}_i(t)$ and an initial self-consistent $\mathbf{R}_I(0)$, the equations for the set of $\mathbf{V}_I$ as a function of $\left(\{\mathbf{R}_J(t)\}, \{\mathbf{r}_J(t)\}, \{\mathbf{v}_J(t)\}\right)$ provide a way to propagate $\mathbf{R}_I(t)$ forward in time with the flow of $\mathbf{r}_i(t)$.

The correspondence elaborated above is not a direct map of single configurations to single configurations, so it does not guarantee that the blob dynamics has an equilibrium independent of $\mathbf{R}_I(0)$ or well-behaved statistical mechanics. However, as the numerical examples of Lennard-Jones fluid trajectories provided in the next section demonstrate, blobs moving according to this self-consistent smoothed centroidal Voronoi tessellation map do reach an $\mathbf{R}_I(0)$ -independent equilibrium and, moreover, the effective CG physics of this equilibrium appear to closely match our expectations for soft fluid blob models.

C. Numerical Implementation

In numerical implementation, one does not work with continuous trajectories and the forms of the mapping equations that are most convenient for physical interpretation are not necessarily the forms that are most computationally convenient. The simplest trajectory mapping from discrete FG trajectories uses Euler integration to propagate the blob positions,

$$\mathbf{R}_I(t + \Delta t) = \mathbf{R}_I(t) + \Delta t \mathbf{V}_I(\{\mathbf{R}_J(t)\}, \{\mathbf{r}_J(t)\}, \{\mathbf{v}_J(t)\}), \quad (4.9)$$

starting from an initial $\mathbf{R}_I(0)$ determined by fixed point iteration of the self-consistency condition from a guessed approximate blob configuration. However, the discretization error slowly causes a drift away from the self-consistency condition for the position map. This is a common issue when integrating along a constraint with a first order integrator, and an expedient solution is to periodically re-enforce the constraint by fixed-point iteration during the course of the integration. The periodic rate and the target error for the fixed-point iteration were free

parameters in our implementation; the target error was specified in terms of the sum of the squares of the differences between the current set of $\mathbf{R}_I$ and the current set of $\sum_i c_{li} \mathbf{r}_i$ given coefficients calculated from those $\mathbf{R}_I$.

The numerical approach described above has the unappealing consequence that the implementation, unlike the formal map, will still have an explicit time dependence and periodic discontinuity, like previous approaches based on discrete spatial tessellation. However, in this case the discontinuity is not accounting for neglected mass transfer physics, but only the discontinuity of discrete integration and the use of finite precision floating-point numbers along the discretized FG trajectory. The discontinuity can be made as small as desired by appropriately shrinking the time step of the FG dynamics.

Computationally, the formulas above are not the most convenient. When calculating the c_{li} , it is natural to also compute the w_{li} and their derivatives $\partial w_{li} / \partial \mathbf{r}_i$ as well as the sums $W_i = \sum_J w_{Ji}$ and the corresponding derivatives $\partial W_i / \partial \mathbf{r}_i$. In terms of those intermediates, there is a more convenient formula for the evaluation of $\mathbf{V}_I$:

$$\begin{aligned} & \sum_J \left(\delta_{IJ} \mathbf{I} + \sum_i (\mathbf{r}_i - \mathbf{R}_I) c_{li} \left(\frac{\delta_{IJ}}{w_{li}} - \frac{1}{W_i} \right) \frac{\partial w_{Ji}}{\partial \mathbf{r}_i} \right) \cdot \mathbf{V}_J \\ &= \sum_i c_{li} \mathbf{v}_i + \sum_i (\mathbf{r}_i - \mathbf{R}_I) c_{li} \left[\left(\frac{1}{w_{li}} \frac{\partial w_{li}}{\partial \mathbf{r}_i} - \frac{1}{W_i} \frac{\partial W_i}{\partial \mathbf{r}_i} \right) \cdot \mathbf{v}_i \right], \end{aligned} \quad (4.10)$$

which can be written in compact matrix form:

$$(\mathbf{I} - \mathbf{M})\mathbf{V} = (\mathbf{C} + \mathbf{N})\mathbf{v}. \quad (4.11)$$

For the derivation, see Appendix A for details. The rightmost term of the right hand side can be calculated in one FG/CG neighbor list traversal, while the matrix $\mathbf{M}$ is calculated in a two-

neighbor loop. The scaling in calculating these matrices is thus linear in the system size. The matrix $\mathbf{M}$ is a sparse matrix with sparsity matching the CG adjacency matrix, so solving the matrix equation to a desired accuracy also scales linearly. Overall, therefore, the cost of the mapping algorithm scales linearly with system size and trajectory length. The detailed numerical implementation is summarized in Appendix B.

III. Application

To demonstrate the proposed mapping scheme, in this section we apply the method and algorithm to a model system of supercritical Lennard-Jones (LJ) fluid. A hierarchy of bottom up CG models with four different resolutions was constructed. We examine the static and dynamical character of the mapped CG trajectory and parameterize the equations of motion to faithfully describe these features.

A. Atomistic System

The atomistic MD simulations of the quiescent LJ fluid was performed in a cubic simulation box subjected to periodic boundary conditions (PBCs) in all three dimensions. The interaction between LJ particles was described through the standard 12-6 type LJ potential:

$$u(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right].$$

To keep the results concise, reduced LJ units were employed throughout the calculation and analysis afterwards (the length, mass, energy and time units being $\sigma=1$, $m=1$, $\varepsilon=1$ and $\tau = \sigma(m/\varepsilon)^{1/2}$ respectively). Two systems with 1099 and 1688 LJ particles were simulated. The reduced temperature $T^*=2.00$ and density $\rho^*=0.8442$ were chosen so that the systems were in the supercritical fluid phase, the simulation box lengths were 10.9174σ and 12.5970σ respectively. LJ interactions were cut off at 2.5σ . The simulations were performed in the

canonical ensemble (constant NVT) by coupling a Nose-Hoover thermostat^{40, 41} having a damping factor of 0.50τ . The systems were first equilibrated for 10^5 time steps to reach equilibrium and the production runs were then propagated for 10^6 time steps with the integration time step $\delta t = 0.005\tau$. The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) was used to perform the atomistic simulations.⁴²

B. Mapping

CG trajectories with four different resolutions were mapped from the atomistic MD simulation trajectories using the method proposed in this work. The two atomistic systems were both CG onto 64 and 125 beads, resulting in four CG systems. The inverse CG resolutions of the four systems, i.e., the number ratio of the FG particles to CG beads, are 8.792, 13.504, 17.172 and 26.375 respectively.

The mapping was performed by our in-house code implementing the equation presented in Sec. II. The concrete algorithms and implementation details are thoroughly summarized in Appendix B. The only parameter that should be predefined before the mapping is the proximity function w .

The proximity function that was adopted in this work is a truncated hyperbolic tangent function,

$$w(r) = \frac{1}{4} \left[1 - \tanh\left(\frac{r/r_c - 1}{\sigma}\right) \right] \text{erfc}\left(\frac{r/r_c - \beta}{\gamma}\right). \quad (4.12)$$

The value of r_c characterizes the radius of the CG beads, which was chosen to be $r_c = L / \sqrt[3]{4\pi N / 3}$ in this work, L denotes the length of the simulation box, and N represents total number of CG beads. Its value was determined by the assumption that the spherical CG beads would fill all of the space of the simulation box. The parameter γ controls the steepness of the

truncation. The value $\sigma = 0.90$ was consistently applied throughout all the calculations in this paper. The complementary error function was introduced to control the shape of the tail part of the weighting function so that it vanished smoothly at its cut-off distance. The constants $\beta = 1.4$ and $\gamma = 0.3$ were employed consistently in each of the numerical calculations. The evaluation of the weighting function was truncated at $r = 2r_c$ since the magnitude of the function (2.2×10^{-4} at $r = 2r_c$) is small enough to be ignored at larger distances. The profile of the weighting function is shown in Fig. 4-2. The region within the characteristic size of CG bead, i.e., $r_c = L / \sqrt[3]{4\pi N / 3}$, has the dominant contribution to the weights. The smoothly decaying tail part was introduced because we found it to be an important factor for the numerical stability of the mapping process during computational experiments.

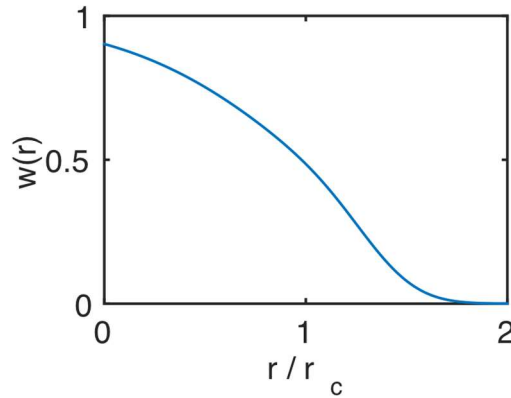


Figure 4-2: The proximity function profile described by Eq. (4.12).

This specific form of proximity function is not necessarily the only option; other choices may work as well. In our computational experiments we found that the numerical stability of the algorithm is sensitive to both the functional form and the concrete shape of proximity function.

The mathematical conditions for the proximity function to realize stable mapping should be carefully analyzed. Other physical and statistical constraints on the proximity functions will also be an important topic for future work on this method.

C. Long-time statistics of the mapped trajectory

With the mapped CG trajectory in hand, in this section we examine statistical information of the CG system including the mass fluctuations, structural correlations, and CG interactions to highlight the properties and characteristics of the new mesoscopic CG mapping method.

1. Mass Distributions

In the theory of the new mapping method, the mass of the CG blobs was not explicitly defined. In this section, we introduce several quantities characterizing the CG masses and examine their statistics along the mapped trajectory.

a. Definition of CG Mass

In Sec. II, the position mapping matrix element is expressed as:

$$c_{li} = \frac{m_i f_{li}}{\sum_j m_j f_{lj}} \equiv \frac{m_i f_{li}}{F_l}. \quad (4.13)$$

By analogy with the conventional center of mass (COM) CG mapping, which takes the following form,

$$c_{li} = \frac{m_i}{\sum_{j \in S_l} m_{lj}} \equiv \frac{m_i}{M_l}, \quad (4.14)$$

it is natural to deduce that the quantity $F_l = \sum_j m_j w_{lj} / \sum_j w_{lj}$ may have the physical meaning of total CG mass, i.e., the mass of FG molecules that contribute to the specific CG blob. In this work, we call this the inertial mass of the CG blobs.

However, the overlapping nature of the new mapping scheme complicates the structure of the CG distribution in phase space, especially the behavior in momentum space. In principle, the physical mass of the CG bead should be derived from a thermodynamic consistency condition in momentum space. It is questionable that the inertial mass can ensure this consistency, i.e., whether the temperature of the CG system evaluated with inertial mass will be consistent with the underlying FG system temperature.

Since the CG velocity is just a mapped function of the FG system properties, its statistics can be derived as ensemble averages over the FG statistical ensemble. In order to rigorously understand the thermodynamic consistency in momentum space, we therefore need to analyze the CG velocity distribution and relate it with the FG velocity distributions. Recall that the velocity mapping equation Eq. (4.11) can be re-written as the follows:

$$\mathbf{V} = (\mathbf{I} - \mathbf{M})^{-1}(\mathbf{C} + \mathbf{N})\mathbf{v} \equiv \mathbf{B}(\mathbf{R}, \mathbf{r})\mathbf{v}. \quad (4.15)$$

The matrix $\mathbf{B}(\mathbf{R}, \mathbf{r})$ can be regarded as the velocity mapping matrix. It is a function of instantaneous CG and FG particle positions. If we assume these are uncorrelated with the FG velocities, then this mapping relationship provides us with enough information to estimate the velocity distribution of the CG system based on the statistics of the FG system. With that assumption, the second moment of CG velocity can be evaluated as

$$\begin{aligned} \langle V_I^2 \rangle &= \frac{1}{Z} \int e^{-\beta H} V_I^2 \delta[M(\mathbf{r}) - \mathbf{R}] d\mathbf{r} d\mathbf{v} \\ &= \frac{1}{Z} \int e^{-\beta U} \sum_i \sum_j B_{ii}(\mathbf{R}, \mathbf{r}) B_{jj}(\mathbf{R}, \mathbf{r}) \delta[M(\mathbf{r}) - \mathbf{R}] d\mathbf{r} \times \int e^{-\beta K} v_i v_j d\mathbf{v} \\ &= \frac{1}{Z} \int e^{-\beta U} \sum_i \sum_j B_{ii}(\mathbf{R}, \mathbf{r}) B_{jj}(\mathbf{R}, \mathbf{r}) \delta[M(\mathbf{r}) - \mathbf{R}] d\mathbf{r} \times \frac{k_B T}{m_i} \delta_{ij} \\ &= k_B T \left\langle \sum_i B_{ii}^2 / m_i \right\rangle_{\mathbf{R}} \end{aligned} \quad (4.16)$$

in which Z represents the partition function of the FG system and the short hand notation for the CG constrained ensemble average takes the form:

$$\langle g \rangle_{\mathbf{R}} = \frac{1}{Z} \int e^{-\beta U} g(\mathbf{r}, \mathbf{R}) \delta[M(\mathbf{r}) - \mathbf{R}] d\mathbf{r}, \quad (4.17)$$

The $\delta[M(\mathbf{r}) - \mathbf{R}]$ term appears in order to enforce consistency of position mapping (note that this is actually a product of delta-functions for all CG variables). By assuming that the mapping matrix $\mathbf{B}$ does not correlate with the velocity, the derivation is greatly simplified. In the above derivation, T is the temperature of the FG system and m represents the FG particle masses. Equation (4.16) tells us that the variance of the CG velocity distribution does depend on the underlying FG system temperature, but was scaled by a constant $\left\langle \sum_i B_{li}^2 / m_i \right\rangle_{\mathbf{R}}$.

This above result suggests that another definition of mass for the CG beads is given by

$$M_T = \sum_i m_i / B_{li}^2, \quad (4.18)$$

which we name the thermal mass. The importance of the thermal mass is that if one chooses the instantaneous thermal mass as the physical mass of the CG beads, the temperature of the CG system will be consistent with the underlying FG system.

Practically, neither the instantaneous inertial mass nor the thermal mass is a constant value; both fluctuate as the FG and CG configuration propagate (the time evolution and statistics of these two masses will be discussed in the following section). However, a time-varying mass will raise many problems for the parameterization and validation of CG blob models, especially when one tries to propagate the dynamics of a CG system after defining its equations of motion. Ideally, the CG mass should be a constant and still preserve the temperature of the underlying

FG system. Luckily, Eq. (4.16) suggests such an answer. One can define the mass of each CG bead according to

$$M_I = \left\langle \sum_i B_{ii}^2 / m_i \right\rangle_{\mathbf{R}}^{-1}, \quad (4.19)$$

which we call the Maxwell mass. It can be regarded as the harmonic mean of instantaneous thermal masses and it is straightforward to verify that the Maxwell mass can preserve the temperature consistency as thermal masses do, through Eq. (4.16). This will be our choice for the mass of CG beads in all the following parameterization and CG system propagation processes.

b. Inertial Mass Distribution

During the mapping, we kept track of the time series of inertial masses and calculated their probability distribution functions (PDFs), shown in Fig. 4-3. In the figure, the inertial mass values are scaled by the CG resolution in order to eliminate the size effect. The result shows that the inertial masses are normally distributed around the expected CG mass, i.e., $m_{FG} n_{FG} / N_{CG}$, with relatively small fluctuations. The variance of the distributions decreases when the CG blobs own more FG molecules, suggesting that the inertia mass of CG blobs may converge to the expected CG mass in the low-resolution limit. This behavior indicates that the inertia mass is a good metric for charactering the ownership of CG blobs during the mapping. In this simple example, the CG blobs ownership is stable with even mass partitioning, perfectly matching the physical picture of the dynamic centroidal Voronoi blobs.

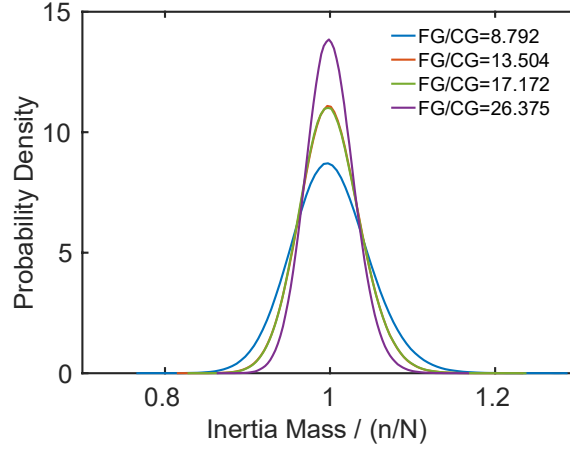


Figure 4-3: The probability distribution function of the inertia masses scaled by CG resolutions (n/N). The four CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).

c. Thermal Mass Distribution

Similarly, the probability distribution functions of the thermal masses are calculated and displayed in Fig. 4-4(a). The thermal masses are also scaled by the CG resolution. In all the four cases, the thermal masses are consistently smaller than the corresponding inertial masses. Interestingly, the ratios between the expectation values of thermal and inertial masses are all around 0.51 for all the test cases, which suggests a very simple rule to directly infer the thermal mass from the CG resolution. The variance of the distributions decrease as the CG resolution gets lower, indicating potentially convergent behavior in the low resolution limit just as was also observed in the inertial mass distributions.

On the other hand, compared to the inertial masses, the thermal mass distributions are much more broadened and are no longer symmetric normal distributions. Figure 4-4(b) displays a logarithmic plot of the thermal mass probability distribution functions, clearly showing a longer tail for larger mass values, especially for the higher resolution CG systems.

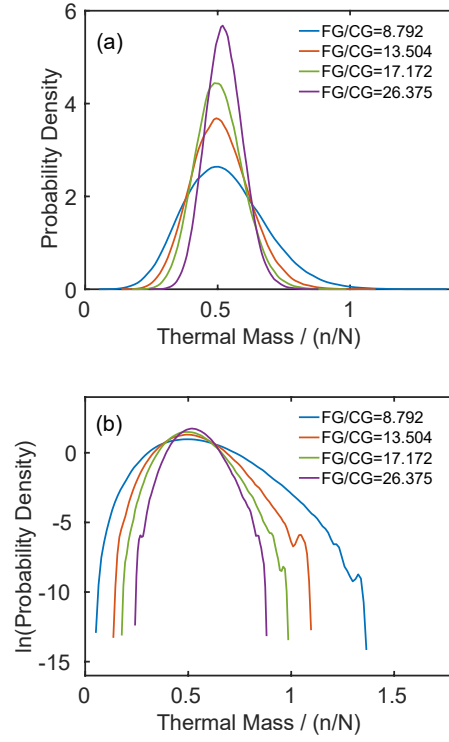


Figure 4-4: Panel (a) the thermal mass distribution at different CG resolutions. Panel (b) the natural logarithm of the distributions, with masses scaled by the CG resolution. The four CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).

d. Maxwell Mass and Temperature Consistency

The Maxwell masses for the four systems, which are simply the harmonic mean of the instantaneous thermal masses, are calculated to have the values of 4.636, 6.858, 8.717 and 13.815 respectively. In Part (a) of this section, we have demonstrated that the Maxwell mass, instead of the more intuitive inertial mass, should be assigned as the mass of the CG blobs while propagating the CG system in order to maintain thermodynamic consistency in momentum space. In this subsection, we validate the concept by actual numerical calculation. To make our

argument concise, only the result at the highest CG resolution of 8.792 is presented. Other cases exhibit the same behavior.

For a particle based simulation system at thermal equilibrium, the probability distribution function of velocity components is given by the Maxwell Boltzmann (MB) distribution,

$$f(v_x) = \sqrt{\frac{m}{2\pi k_B T}} \exp\left(-\frac{mv_x^2}{2k_B T}\right). \quad (4.20)$$

We calculated the velocity distribution profile in Fig. 4-5 using both the Maxwell mass and mean inertia mass, at the target FG temperature $T^* = 2.0$. It is clear that the MB distribution with Maxwell mass quantitatively captures the velocity distribution of the CG beads, while the one obtained with inertial mass underestimates the variance. This observation is in good agreement with our theoretical analysis.

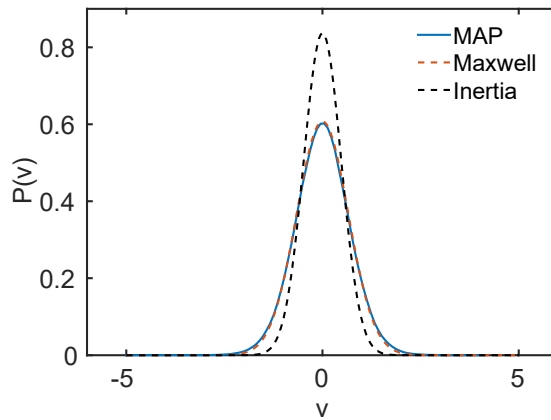


Figure 4-5: The velocity distribution profiles obtained from the mapped CG trajectory (blue solid line), the Maxwell Boltzmann distribution with Maxwell mass and $T^* = 2.0$ (red dashed line) and with mean inertia mass and $T^* = 2.0$ (black dashed line). The blue solid line and the red dashed line are virtually indistinguishable.

2. Structural Distributions

In Fig. 4-6, the radial distribution functions (RDFs) for the four mapped CG trajectories are calculated. The CG RDFs behave like typical liquid state RDFs. CG models with lower resolution exhibit more structured packing, showing higher peak values for both first and second solvation shells, and a more oscillatory overall profile, indicating a more solid like behavior. This observation is consistent with the results obtained by Izvekov and Rice through the K-means clustering tessellation.³² Intuitively, in the low-resolution limit in which each CG blob owns a large amount of FG molecules, the mobility of the CG beads will be greatly restricted. This is mainly because the mean velocity of the FG particles within the CG blob will be small due to its large size, and secondly because the contribution from momentum transfer will also be small because the number of FG particles in the overlapping region is small compared to the core area. The physical picture of the large CG blobs is similar to the concept of fluid parcel or material element in the continuum description of fluid dynamics.⁴³ In such situations the structure of the CG blobs would be rather solid/crystal like. The trend shown in Fig. 4-6(a) provides evidence for this interpretation.

In Fig. 4-6(b), the same RDFs are plotted with scaled radial distances, i.e., the actual distance divided by the cube root of the CG resolution, to eliminate the size effect. It is interesting to see that the CG system shows an invariance of the RDF peak positions under the scaling, indicating a consistent microstructure pattern. It also shows that higher resolution CG systems have softer peaks in their RDFs, which can also be explained by the previous discussion.

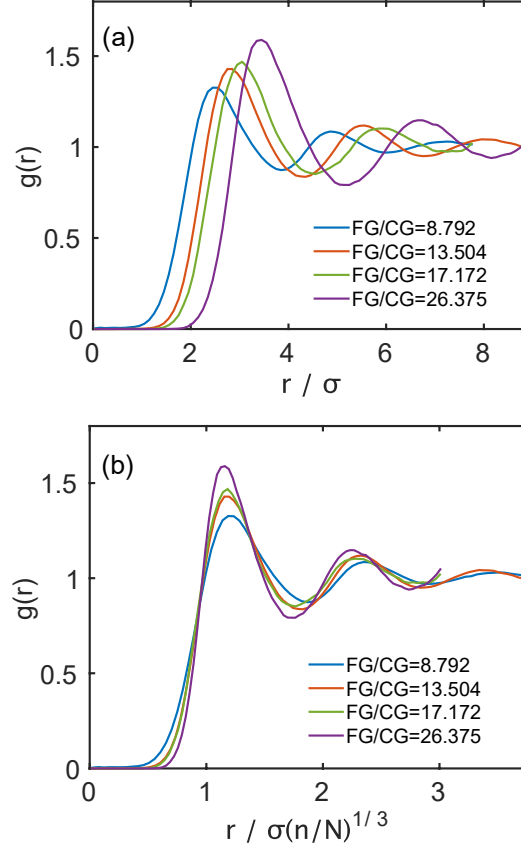


Figure 4-6: Panel (a) the radial distribution functions (RDFs) of the CG trajectory at different resolution mapped by the proposed CG mapping. Panel (b) the same RDFs with radial distances scaled by the cube root of CG resolution. The resolutions for the CG systems are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple).

In addition to the pair distribution structure, the local CG blob densities are also extracted to manifest the many-body structural correlations in the mapped CG system. The local density is calculated according to the following formula,

$$\rho_i = \sum_j W(r_{ij} / 2h), \quad (4.21)$$

in which the weight function is defined as

$$W(q) = \frac{8}{\pi} \begin{cases} 1 - 6q^2 + 6q^3 & 0 \leq q \leq 1/2 \\ 2(1 - q)^3 & 1/2 \leq q \leq 1 \\ 0 & q > 1 \end{cases} \quad (4.22)$$

The cut off distance $2h$ is set to be the peak position of the first solvation shell (first peak in RDFs). The local densities were calculated for each CG bead and their probability distribution functions are shown in Fig. 4-7. The distribution function broadens as the CG resolution gets higher, which indicates a softer and more relaxed local chemical environment. This is also manifested in the softer RDFs for higher resolution CG models.

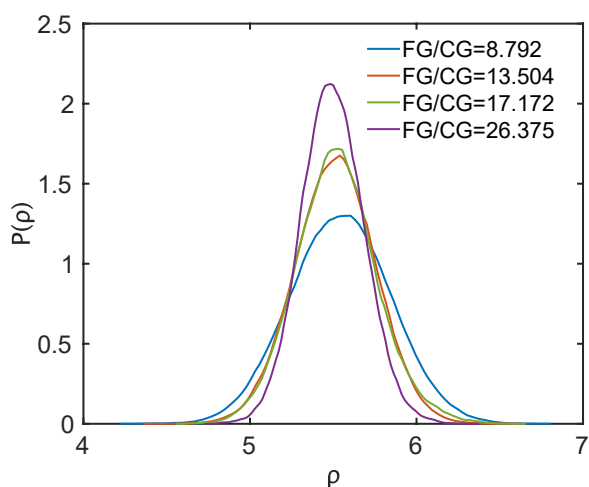


Figure 4-7: The local density distribution obtained from the mapped CG trajectories. The CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.

3. *Shear Flow*

By construction, a CG mapping is designed such that the slow degrees of freedom (DoFs) of the FG system can be described by the CG variables.^{44, 45} For example, in the case of a center of mass mapping the CG variable can represent the translational motion of the underlying FG

molecules while eliminating faster vibrational and librational motions. In the new mapping scheme, the situation is more complicated since the definition of FG to CG correspondence changes during the propagation. The CG variable represents the center of mass of a spatial blob of FG molecules whose motion is hard to characterize in equilibrium quiescent flows. It is therefore questionable if the mapping approach will faithfully capture the important slow motions of fluids.

To answer this question, a good initial test is to apply the new mapping scheme to coarse-graining fluids under steady-state non-equilibrium shear flow. In this case, the shear flow of the fluid particles would be the predominant slow collective motion and one would like to see that the mapped CG system exhibits the same flow motion that the FG system does.

The atomistic trajectories of LJ fluids under steady-state shear flow were generated through the Müller-Plathe method,⁴⁶ in which the shear flow is generated through exchanging momenta between particles in different regions of the simulation box. A box of 2048 LJ particles was simulated at reduced temperature $T^* = 2.00$ and density $\rho^* = 0.8442$, while the other simulation details are the same as the quiescent flow cases in Sec. III. A. The z axis of the box was divided into 20 slabs, and for every 50 timesteps two pairs of atoms in the upper most slab (slab 1) and the middle slab (slab 10) exchanged their x component of velocity. The atoms with highest positive v_x in slab 1 and highest negative v_x in slab 10 were chosen for each exchange thus generating a steady shear velocity profile. After the system was brought to steady state for 10^6 time steps, the production run was propagated for another 10^6 time steps and the coordinates and velocities were recorded for each time step. Then, the atomistic trajectory under shear flow was mapped onto 216 CG beads through the new mapping scheme under the same conditions as

described in Sec. III. B. The profiles of average v_x for both the FG LJ fluid and the mapped CG system were calculated at each slab and the result is shown in Fig. 4-8.

The shear velocity profiles for the FG fluids and CG system exhibit nearly identical quantitative behavior. The mapped CG blobs are moving at the same pace with the underlying FG particles. This observation provides strong evidence that the new mapping scheme proposed in this work is able to capture the slow collective motion of the underlying FG fluids. It also suggests that this method can be applied to fluid systems under both equilibrium and non-equilibrium conditions.

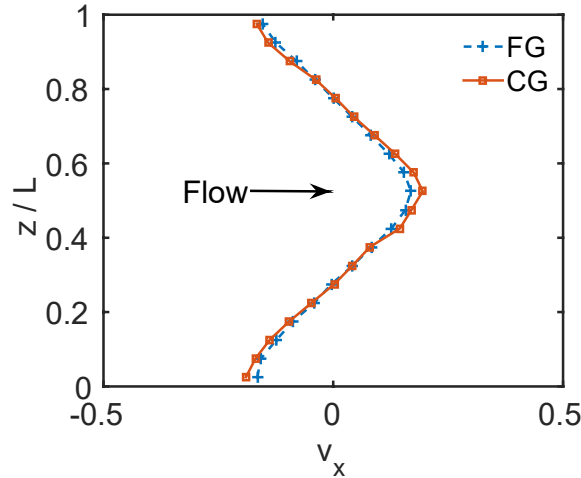


Figure 4-8: the profile of x component of velocity under shear flow for both the FG LJ fluids (blue dashed line with cross marker) and the CG system mapped through dynamic mapping (red solid line with square marker).

D. CG Model Parameterization

In the previous sections, we examined the long-time statistics of the CG masses and structural correlations. In this section, a detailed parameterization and validation of the CG models, both the conservative and non-conservative interactions, is presented.

1. CG Equation of Motions

The theoretical tools for describing the statistics and dynamics of CG models⁴⁷ are found in the development of non-equilibrium statistical mechanics in the mid 20th century. One of the major achievements is the Mori-Zwanzig formalism,³⁶⁻³⁹ in which the non-equilibrium time evolution of any set of collective variables defined on the phase space of the microscopic system is shown to follow the form of a Generalized Langevin Equation (GLE) via a projection operator method. In systematic “bottom-up” theories of coarse-graining, the CG sites are exactly defined as a set of collective variables of the underlying FG system. Therefore, the dynamics of CG systems derived via projection operator formalism should be well described within the same theoretical framework. Although this theory has been established, the Mori-Zwanzig formalism has been hard to use as practical computational tool due to the many-body nature of the memory kernel and the numerical complexity arising from the GLE type integro-differential equation.²²

In the past decade, there have been many efforts to model the structural and dynamical properties of CG systems directly from FG molecular dynamics simulations utilizing different simplifications to the Mori-Zwanzig formalism.^{22, 24, 25, 47-53} To name a few, Hijon and Espanol²² proposed the first bottom-up parameterization of Markovian DPD models through the “constrained dynamics” method. Yoshimoto *et al*⁵¹ extended the theory to derive DPD models under non-Markovian regime for the first time. Karniadakis and co-workers^{25, 50, 54-56} systematically investigated the parameterization of the GLE and DPD equation of motion (EOM) for bonded polymer melt systems in both Markovian and non-Markovian regimes, utilizing a direct numerical inversion approach to derive the GLE memory kernel and the more complicated spatial and temporal DPD type memory kernel. Izvekov and Rice³² explored the application of Mori-Zwanzig formalism to parameterize mesoscopic CG models for liquid systems through K-

means clustering projection. More recently, Davtyan et al.,^{52, 53} proposed an alternative approach to describe the CG dynamics by coupling the CG beads to fictitious particles, effectively introducing an approximation of the GLE memory kernels.

However, all these models are generated based on the simple center of mass CG mapping projections, with no overlapping in the FG particle attribution. The new dynamic mapping scheme in the present work involves a complicated non-linear map from trajectories to trajectories rather than from instantaneous FG state to instantaneous CG state. Due to this complexity, the Mori-Zwanzig formalism does not directly apply to our scheme and we lack an explicit expression for the forces of the CG blobs, which is the basis for the “constraint dynamics” and pairwise decomposition technique utilized in the above-mentioned methods. As a consequence, the application of these prior developments to our current approach is not straightforward.

In this work, we present an efficient parameterization protocol for CG models using non-Markovian GLE equations of motion based on the Mori-Zwanzig formalism and then validate its applicability through numerical experiments. Formally, once the projection operator is in place, the GLE equation describing the CG particles can be derived from the Mori-Zwanzig formalism as:²⁵

$$\dot{\mathbf{P}}_I = \frac{1}{\beta} \frac{\partial}{\partial \mathbf{R}_I} \ln \omega(\mathbf{R}) - \beta \sum_{J=1}^N \int_0^t ds \left\langle \left[\delta \mathbf{F}_I^Q(t-s) \right] \cdot \left[\delta \mathbf{F}_J^Q(0) \right] \right\rangle \frac{\mathbf{P}_J(s)}{M_J} + \delta \mathbf{F}_I^Q(t), \quad (4.23)$$

where $\mathbf{R} = \{\mathbf{R}_I, I=1, \dots, N\}$ represents a CG phase space state point and $\omega(\mathbf{R})$ is defined from the many-body potential of mean force (PMF) of the CG variables via the expression

$$\omega(\mathbf{R}) = \frac{\int d\mathbf{r} \delta(M(\mathbf{r}) - \mathbf{R}) e^{-\beta u(\mathbf{r})}}{\int d\mathbf{r} e^{-\beta u(\mathbf{r})}}. \quad (4.24)$$

In the above expression, the conservative force acting on the CG beads is shown in the first term, which depends only on the configuration of all the CG beads in the system, whereas the dissipative force given by the second term depends on both the configuration and momentum information of the whole CG system. Both the conservative and non-conservative forces in the EOM are complex multi-body functions, so the direct computation of them is considered to be very challenging even for simple one-dimensional harmonic systems.⁵⁷ There are multiple ways of simplifying the Mori-Zwanzig EOM to make it a practical computational tool by applying different assumptions. In this work, we first perform a pairwise decomposition of the many-body potential of mean force, assuming the many-body PMF can be approximated through a pairwise interaction, i.e.,

$$\mathbf{F}_I^C(\mathbf{R}) = \frac{1}{\beta} \frac{\partial}{\partial \mathbf{R}_I} \ln \omega(\mathbf{R}) = \sum_{J \neq I} f(|\mathbf{R}_{IJ}|) \mathbf{e}_{IJ}. \quad (4.25)$$

Then, we assume the random force on different CG particles has no spatial correlations, i.e.,

$$\left\langle \left[\delta \mathbf{F}_I^Q(t) \right] \cdot \left[\delta \mathbf{F}_J^Q(t') \right] \right\rangle = \delta_{IJ} \mathbf{K}(t - t').$$

With these two assumptions, the Mori-Zwanzig EOM is simplified to have the form of a simpler GLE, which describes the evolution of CG dynamics as:

$$\dot{\mathbf{P}}_I(t) = \sum_{J \neq I} f(|\mathbf{R}_{IJ}(t)|) \mathbf{e}_{IJ} - \int_0^t \mathbf{K}(t-s) \frac{\mathbf{P}_I}{M_I}(s) ds + \delta \mathbf{F}_I^Q(t). \quad (4.26)$$

The last term $\delta \mathbf{F}_I^Q(t)$ stands for the random force resulted from the unresolved degrees of freedom in the FG system that is CG out. The random forces and dissipative forces are related through the second Fluctuation Dissipation Theorem (FDT) by

$$\mathbf{K}(t) = \beta \left\langle \delta \mathbf{F}_I^Q(t) \otimes \delta \mathbf{F}_I^Q(0) \right\rangle.$$

A correct construction of random noise terms that satisfy the FDT is vital for maintaining the temperature of the CG simulation, which then guarantees the consistent description of the structural and dynamical information.

2. *Conservative Interactions*

The output of the new dynamical CG mapping scheme proposed here is a CG trajectory with the information of CG blob coordinates and velocities. Due to the complexity of the overlapped CG mapping, it is not simple to directly extract the CG forces from the FG forces. In this work, the instantaneous forces exerted on CG blobs were approximated by numerical differentiation of the CG velocities:

$$\mathbf{F}_I(t) \approx M_I \frac{\mathbf{V}_I(t + \Delta t) + \mathbf{V}_I(t - \Delta t) - 2\mathbf{V}_I(t)}{2\Delta t}, \quad (4.27)$$

where M_I is the Maxwell mass of CG bead I .

With the instantaneous forces in place, the pairwise interactions between the CG beads are determined through the Multiscale Coarse-graining (MS-CG) method.⁵⁸⁻⁶² The MS-CG method constructs the optimal pairwise decomposition of the CG many-body PMF, utilizing the mapped CG trajectory coordinates and forces as input. The pairwise interaction $f(R)$ is determined variationally by minimizing a residual functional, given by

$$\chi^2[f] = \frac{1}{3N_f N_{CG}} \sum_{N_f} \sum_I \left| \mathbf{F}_I - \sum_{J \neq I} f(|\mathbf{R}_{IJ}|) \mathbf{e}_{IJ} \right|^2. \quad (4.28)$$

In practical implementations, the pairwise interaction $f(R)$ is approximated by a linear expansion into a set of basis functions as

$$f(R) = \sum_k \phi_k B_k(R). \quad (4.29)$$

The basis coefficients $\{\phi_k\}$ are optimized by minimizing the force residual functional in Eq. (4.28) in a least square sense. The numerical implementation, in this case using B-Splines, can be found elsewhere.⁶³ The parameterization was performed through the MSCGFM software that is freely available upon request. The inferred pairwise potentials and forces are displayed in Fig. 4-9(a) and Fig. 4-9(b), respectively.

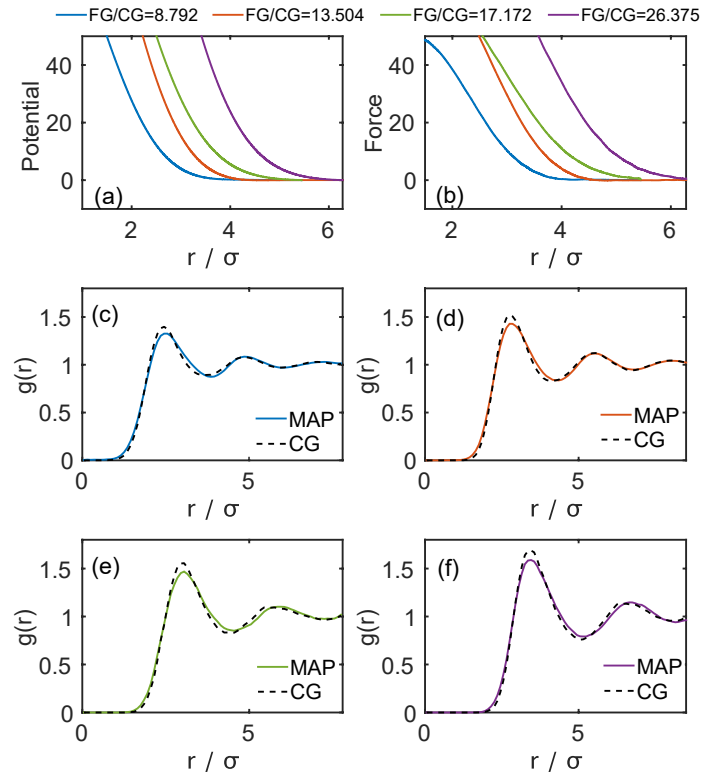


Figure 4-9: The effective interactions between CG beads at different resolutions derived through the MS-CG method, (a) the potential energy profile and (b) the force profile. The RDFs obtained by running simulations using the inferred CG interactions are presented in (c)-(f); the RDFs calculated from the mapped CG trajectories are also shown for comparison. The CG resolutions, i.e. the number ratios of the FG to CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.

The pairwise forces show a nearly linear repulsive behavior at nearest-neighbor ranges, which is consistent with the linear repulsive force adopted in DPD models, and they smoothly decay to zero at longer range. Like the RDFs, the pairwise forces also exhibit scaling invariance behavior, with the radial distance scaled by the cube root of CG resolution and forces scaled by the product of FG particle number and square root of the number of CG beads. The scaled force profile fits reasonably to a simple sigmoid function $\alpha \left[1 + \exp \left(-(x - x_c) / \beta \right) \right]^{-1}$. The scaled force profiles as well as the fitted curve are shown in Fig. 4-10.

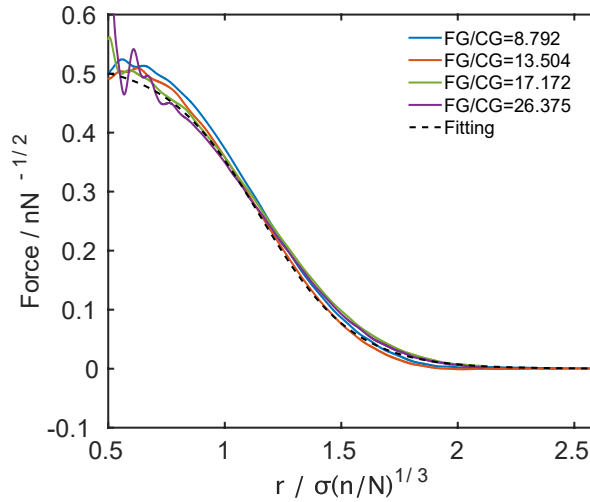


Figure 4-10: The scaled pairwise force profile between CG beads. The radial distances are scaled by the cube root of the CG resolution, whereas the forces are scaled by $n_{FG} / \sqrt{N_{CG}}$. The scaled force profiles can be fitted by a sigmoid function, shown as the black dashed line.

The inferred pairwise additive CG potentials were then used to propagate the dynamics of each CG model at the same temperature and volume as the corresponding atomistic simulation.

Figures 4-9 (c)-(f) show the comparison of RDFs obtained from each mapped CG trajectory and the trajectory generated by each of the MS-CG potentials. In general, the inferred pairwise CG force field did a good job in reproducing the pairwise structure as described by the RDFs, especially the long-range behavior after the first solvation shell, although the inferred pairwise potentials slightly over-structured the first solvation shell in every case.

In addition to the RDFs, the local CG bead densities were also compared. The local densities' probability distribution functions were calculated under the same conditions as described in Sec II. C and the results are displayed in Fig. 4-11. The MS-CG results show a slightly broader distribution than the mapped CG trajectory, indicating a more flexible local environment.

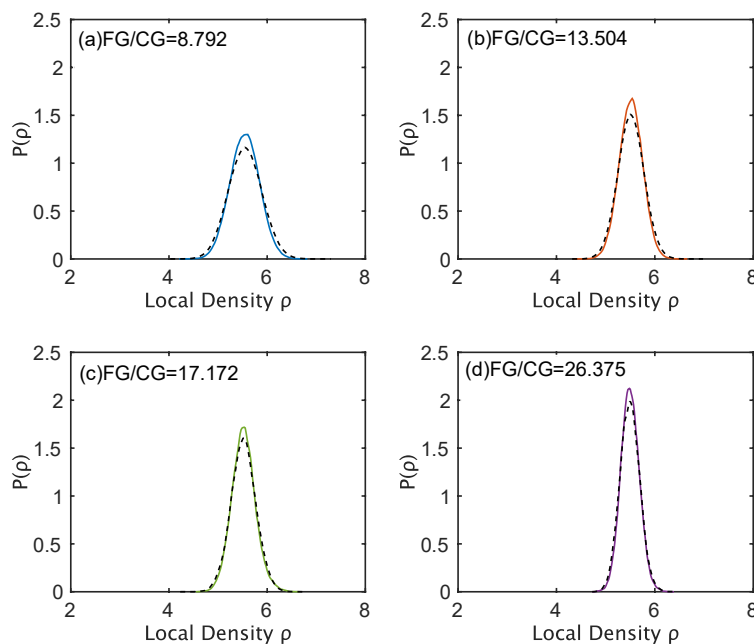


Figure 4-11: The local density distribution obtained by propagating the CG dynamics using the inferred CG interactions is presented in panels (a)-(d), with the references calculated from the mapped CG trajectories also shown for comparison. The number ratios of the FG and CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple), respectively.

In order to understand how these deviations of the RDFs and local density profiles change as a function of the size of CG blobs, we chose two metrics to characterize the error. The error of the RDFs is evaluated by the relative error of the maximum value of $g(r)-1$. For the local density profiles, the error is measured by relative error in the variances of the density distributions. Figure 4-12 displays the result of the error analysis. For both the RDFs and local density profiles, the pairwise interactions derived by MS-CG better describes the CG systems with lower resolutions, but the error does not appear to extrapolate to zero when the CG resolution becomes infinitely low. Even at this limiting case, then, it appears that many-body interaction effects would still play some role in the description of CG structures.

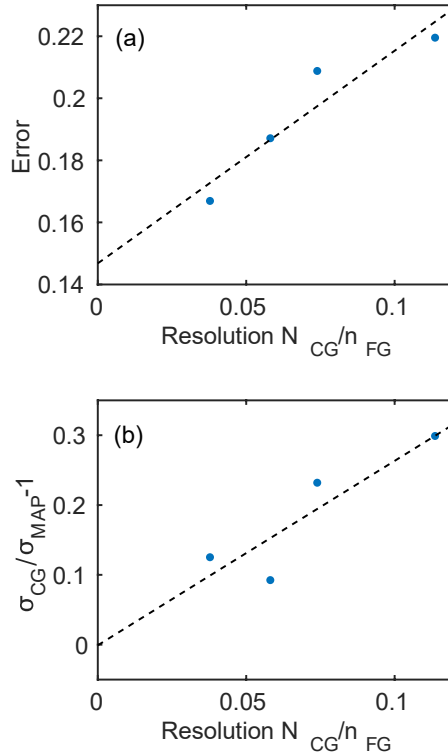


Figure 4-12: Error analysis of the RDFs and local density PDFs: Panel (a) the relative error of $\max\{g(r)-1\}$ between the MS-CG results and mapped CG trajectory as a function of CG

Figure 4-12, continued: resolution, Panel (b) the $\sigma_{CG} / \sigma_{MAP} - 1$ as a function of CG resolution, where σ_{CG} and σ_{MAP} represents the variance of the local density probability distributions of the MS-CG results and mapped CG results, respectively.

The deviations in the RDFs and local densities suggest that the pairwise force field cannot fully describe the CG many-body potential of mean force for the dynamically mapped blobs. Many-body effects must therefore play a role in reproducing the mesoscopic structure of liquids. The origin of the many-body effects might be attributed to the complicated momentum transfer and the ownership fluctuation of CG blobs. It should be noted that there are methods like many-body DPD (MDPD),⁶⁴⁻⁶⁶ SPH,^{6,7} and smoothed dissipative particle dynamics (SDPD),⁶⁷ that take the local density of CG beads as an order parameter to modulate the pairwise interactions. Models incorporating local density dependent interactions have also been constructed in higher resolution CG models to capture interfacial properties and the heterogeneity of a system.⁶⁸⁻⁷⁰ The “bottom-up” inference of such a local density dependent CG force field is a challenging and important topic for further study in this our new mesoscopic dynamical mapping scheme.

4. Non-conservative Interactions

In the prior section, conservative interactions determining the average structure of a CG blob system were parameterized and validated against the mapped FG reference. In this section, we discuss the bottom-up derivation of the non-conservative dissipation and fluctuation forces for the same model.

a. Parameterization Method

The memory kernel in the GLE can be directly extracted from the mapped CG trajectory via the time series of CG velocities and forces. Many efforts have been devoted to solve this problem.⁷¹⁻⁷³ In this work, we address this problem through a direct inversion method. Multiplying $V_I(0)$ on both sides of Eq. (4.24) and taking the ensemble average gives

$$\left\langle [\Delta \mathbf{F}_I(t)] [\mathbf{V}_I(0)]^T \right\rangle = - \int_0^t ds \mathbf{K}(t-s) \left\langle [\mathbf{V}_I(s)] [\mathbf{V}_I(0)]^T \right\rangle + \left\langle [\delta \mathbf{F}_I^Q(t)] [\mathbf{V}_I(0)]^T \right\rangle, \quad (4.30)$$

in which the $\Delta \mathbf{F}_I(t) = \mathbf{F}_I(t) - \mathbf{F}_I^C(t)$ term is the instantaneous non-conservative force acting on CG particle I . Since the random forces are not correlated with the initial velocity $V_I(0)$, the second term on the right hand side vanishes. Therefore the above equation can be expressed as a Volterra equation of first kind,

$$C_{FV}(t) = - \int_0^t ds \mathbf{K}(t-s) C_{VV}(s), \quad (4.31)$$

where the $C_{FV}(t)$ denotes the non-conservative force-velocity correlation function and $C_{VV}(s)$ represents the velocity autocorrelation function (VACF). Both of these two correlation functions can be directly estimated from the mapped CG trajectory. In principle, then, the memory kernel $\mathbf{K}(t)$ can be extracted by solving a discretized form of Eq. (4.31). However, this direct method has been shown to introduce numerical inaccuracies due to the numerical inversion of a Toeplitz matrix. On the other hand, transforming the Volterra equation of first kind into one of the second kind can lead to more stable numerical algorithms. In the present case of Eq. (4.31), taking the time derivative on both sides of equation yields⁷¹

$$\dot{C}_{FV}(t) = -C_{VV}(0)\mathbf{K}(t) - \int_0^t ds \dot{C}_{VV}(t-s)\mathbf{K}(s). \quad (4.32)$$

The derivatives of non-conservative force-velocity and velocity-velocity correlation functions can be obtained through numerical differentiation. However, a more careful analysis

demonstrates that the derivatives of correlation functions can be expressed as other forms of correlation function. Specifically, we can write $\dot{C}_{FV}(t)$ and $\dot{C}_{VV}(t)$ as:

$$\dot{C}_{FV}(t) = -\langle \Delta \mathbf{F}(t) \mathbf{F}(0)^T \rangle / M, \quad (4.33)$$

$$\dot{C}_{VV}(t) = \langle \mathbf{F}(t) \mathbf{V}(0)^T \rangle / M. \quad (4.34)$$

These correlation functions can be directly estimated from the mapped CG trajectory and therefore eliminate the additional error from numerical differentiation.

In principle, the correlation functions and memory kernel are temporally unbounded. However, for a practical implementation they must be truncated to a finite range. In the following numerical examples, we will verify that this assumption is not a bad one since the correlation functions usually decay to small values within a short period of time. With this assumption, the discretization of Eq. (4.32) reads as

$$\dot{C}_{FV}(i\Delta t) = -C_{VV}(0)\mathbf{K}(i\Delta t) - \Delta t \sum_{j=0}^i \eta_j \dot{C}_{VV}[(i-j)\Delta t] \mathbf{K}(j\Delta t), \quad (4.35)$$

where the $\eta_j = 1/2$ if $j=0$ or i , $\eta_j = 1$ for other cases. Then the memory kernel $\mathbf{K}(t)$ at each discretized time step $t = i\Delta t$ can be solved iteratively by the following rule,

$$\mathbf{K}(i\Delta t) = (C_{VV}(0) + \dot{C}_{VV}(0)\Delta t / 2)^{-1} \times \{ \dot{C}_{FV}(i\Delta t) + \Delta t \sum_{j=0}^{i-1} \eta_j \dot{C}_{VV}(0) \mathbf{K}(j\Delta t) \}, \quad (4.36)$$

in which the initial value of memory kernel is $\mathbf{K}(0) = \dot{C}_{FV}(0) / C_{VV}(0)$. This iterative scheme is accurate and stable.

b. Propagation

After the parameterization, we propagated the CG blob GLE in its discretized form to validate the consistency of the inferred CG model. To simplify the notation, we demonstrate the propagation scheme through a derivation in one dimension:

$$\dot{P}_I(t) = F_I^C(t) - \int_0^t K(t-s)V_I(s)ds + \delta F_I^Q(t). \quad (4.37)$$

In the practical implementation, we assume that the memory kernel K decays to zero within N time steps. Therefore, the convolution integral representing the dissipative force is discretized as:

$$-\int_0^t K(t-s)V_I(s)ds \approx -\sum_{n=0}^N \alpha_n \Delta t K(n\Delta t)V_I(t-n\Delta t), \quad (4.38)$$

where $\alpha_k = 1/2$ if $k = 0$ or $k = N$, otherwise $\alpha_k = 1$.

Next, we generate the random forces $\Delta F_I(t)$ that satisfy the FDT. In this discretized representation, the FDT can be expressed as

$$\langle \delta F_I^Q(n\Delta t) \delta F_I^Q(m\Delta t) \rangle = k_B T \cdot K[(n-m)\Delta t]. \quad (4.39)$$

The random force at each time step is assumed to be the superposition of a set of non-correlated Gaussian random noises weighted by the weighting coefficients $\{\alpha_s\}$, such that

$$\delta F_I^Q(t) = \sqrt{k_B T} \sum_{s=0}^N \alpha_s W(t-s\Delta t), \quad (4.40)$$

where the white noises have no temporal correlation,

$$\langle W(k\Delta t)W(l\Delta t) \rangle = \delta_{kl}. \quad (4.41)$$

Combining this definition with the FDT, one can derive the relationship between the colored noise weighting coefficients and the memory kernel:

$$\begin{aligned}
\langle \delta F_l^Q(l\Delta t) \delta F_l^Q(0) \rangle &= k_B T \sum_{p=0}^N \sum_{s=0}^N \alpha_s \alpha_p \langle W(-p\Delta t) W[(l-s)\Delta t] \rangle \\
&= k_B T \sum_{p=0}^N \sum_{s=0}^N \alpha_s \alpha_p \delta_{l-s+p} \\
&= k_B T \sum_{s=0}^{N-l} \alpha_s \alpha_{s+l} = k_B T K(l\Delta t)
\end{aligned} \tag{4.42}$$

To satisfy this relationship, a set of coefficients α_s was obtained by minimizing the squared residual between $\sum_{s=0}^{N-l} \alpha_s \alpha_{s+l}$ and $K(l\Delta t)$ in each case. The Nelder-Mead simplex optimization algorithm⁷⁴ was applied to solve this non-linear optimization problem. An iterative scheme that takes the previous output from the optimizer as initial guess input was applied to make sure the convergence of the above optimization process. After the weights for random noises are constructed, the GLE EOM was ready to be propagated. All the CG GLE simulations used a Velocity Verlet time integration algorithm with identical integration time steps of $\delta t = 0.025\tau$.

c. Results

The GLE memory kernels derived from the mapped CG trajectory are shown in Fig. 4-13. The profile of the memory kernel shows a relatively fast decay within 20τ followed by a more slowly decaying tail.

Applying the Green-Kubo relationship to the GLE defines the relationship between the diffusion constant and the GLE memory kernel,⁷⁵ $D = k_B T / \left(m \int_0^\infty K(t) dt \right)$, i.e., the product of particle mass and the integral of the memory kernel $\int_0^\infty K(t) dt$ is inversely proportional to the diffusion constant. Figure 4-13 demonstrates that the integral value of the memory kernel increases, and thus the diffusion constant decreases, as the CG blobs become coarser. In the

low-resolution limit the physical meaning of the CG blobs is designed to match the concept of local fluid parcels in continuum fluid mechanics, which have no self-diffusion under quiescent conditions. The trend observed in our calculation is consistent with that asymptotic expectation.

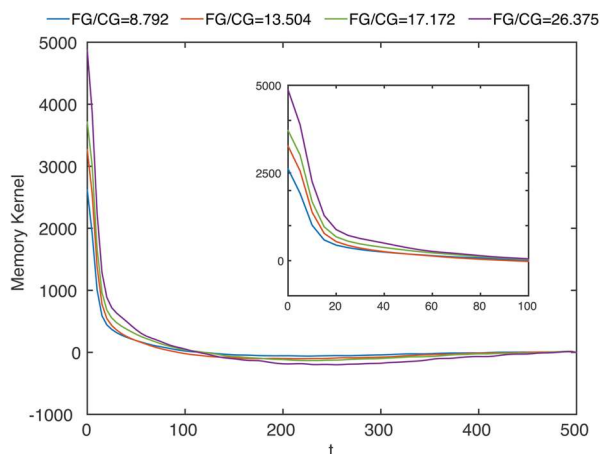


Figure 4-13: The GLE memory kernel inferred from the mapped CG trajectory at different CG resolution. The CG resolutions are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.

The GLE dynamics of each CG system were each propagated after obtaining the memory kernel and the corresponding colored noise generator. Figure 4-14 shows the comparison of the VACFs between the mapped CG trajectories and the GLE CG trajectories for all the four CG resolutions. For the densely packed CG systems, each CG bead resides in a cage formed by its nearest neighbors. Therefore, the CG bead will hit the wall of the cage and get pushed back as soon as it starts to move. This behavior accounts for the backscattering region shown in the negative part of the VACFs at relatively short times. At longer time scales, the VACFs of the mapped trajectory exhibit a long-tail decay proportional to $t^{-3/2}$ consistent with usual expectations for long time scale hydrodynamic memory effects,⁷⁵ as seen in the log-log plots shown in panels

(e)-(h) in Fig. 4-14. Overall, the agreement between the CG GLE dynamics and the mapped CG dynamics is seen to be quite good.

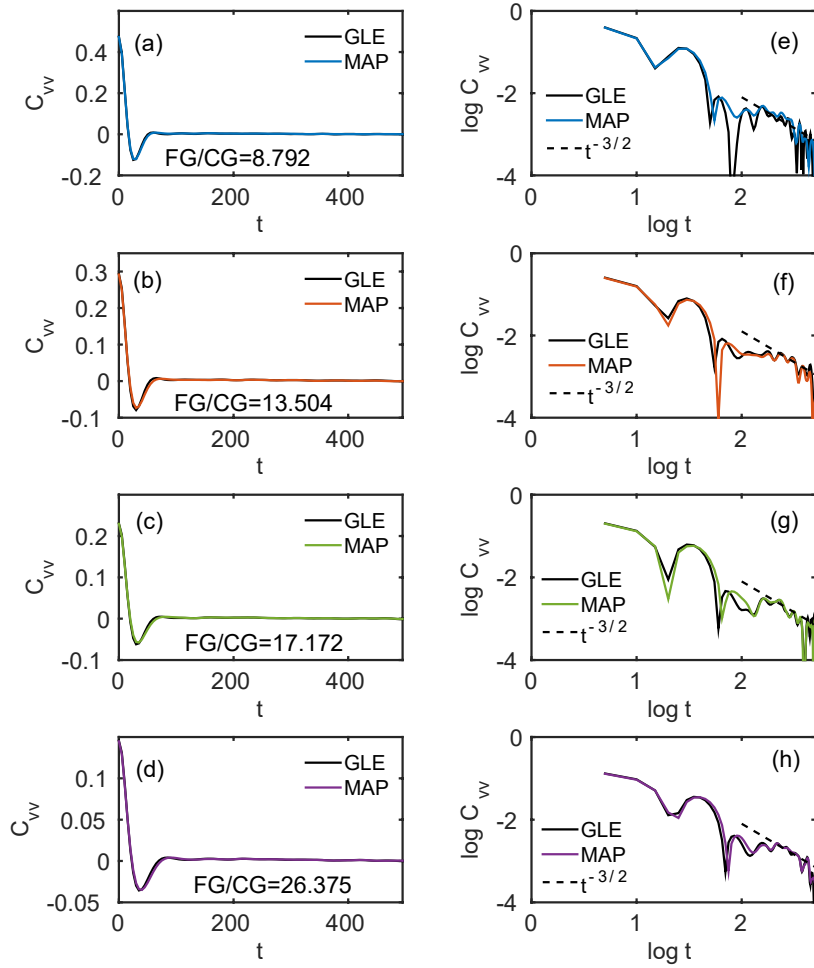


Figure 4-14: The velocity autocorrelation functions (a)-(d) and the log-log plot of them (e)-(h) obtained from both the GLE CG simulation and the mapped CG trajectory. The number ratios of the FG and CG particles are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.

Next, we investigated the scaling behavior of the dynamical information, i.e., the GLE memory kernels and the VACFs in this case, of CG models at different CG resolution. Fig. 4-15 shows a near-perfect scaling invariance of the GLE memory kernel when the memory kernels are scaled by $(n/N)^{3/5}$, where the n/N denotes the resolution of the CG model.

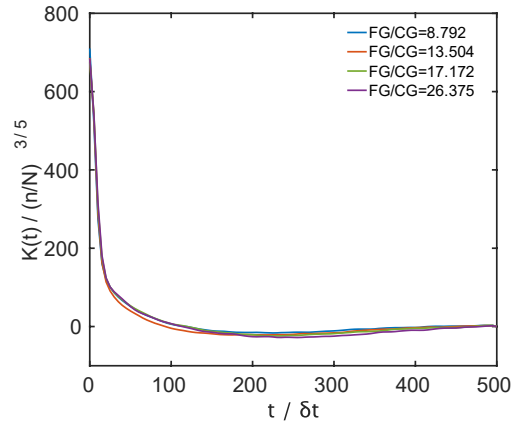


Figure 4-15: The GLE memory kernels scaled by $(n/N)^{3/5}$, where n/N represents the resolution of CG models. The four CG resolutions adopted here are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively.

Similarly, a scaling invariance also exists for the VACFs. Fig. 4-16 displays the scaled VACFs, with time scaled by the cube root of the CG resolution and the correlation functions scaled by the inverse of the CG resolution n/N . The scaled VACFs at different CG resolutions collapse into a single profile. These scaling invariances, together with the one found for the conservative interactions shown in Sec. III.D.2, suggest that the mesoscale CG models constructed by the new dynamical mapping scheme have a unified form of interaction at different resolutions. This observation provides us with interpolation rules to infer CG model parameters over a fairly wide range of CG resolution, with only a few sets of parameters derived at several specific resolutions. These interpolation methods will greatly reduce the workload of

bottom up construction of the present mesoscale models. Crucially, the time scalings of the memory kernels and the VACFs differ in such a way as to allow the low-resolution limit of coarsening to converge to perfectly Markovian dynamics.

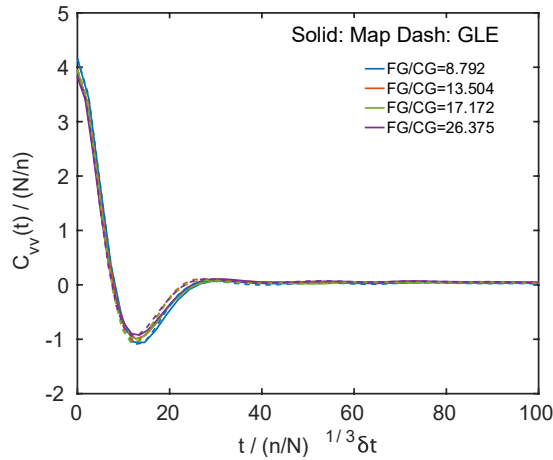


Figure 4-16: The scaled VACFs, with the time scaled by the cube root of CG resolution and the correlation function scaled by inverse CG resolution N/n . The four CG resolutions adopted here are 8.792 (blue), 13.504 (red), 17.172 (green) and 26.375 (purple) respectively. Each solid line indicates the VACF of a mapped trajectory, while the dashed line with the same color denotes the results obtained from corresponding CG GLE propagations.

It should be noted that, in a theory described by a GLE, the dissipative forces depend only on the history of one specific particle. The memory kernel is assumed to be only a function of time. A drawback of this simple GLE is that it cannot ensure the conservation of momentum. The lack of momentum conservation and spatially dependent memory hinders the applicability of a GLE based EOM to model fluids under non-equilibrium conditions. For such systems, the DPD type non-conservative interaction, which incorporates a pairwise decomposable spatial dependent

memory kernel, is needed to correctly describe the hydrodynamics.²⁵ A bottom up parameterization of DPD-like CG models will be investigated in a future study.

IV. Concluding Remarks

In this work, we explored a bottom up strategy for building mesoscopic CG models for unbonded molecular fluids within a Lagrangian description. This work opens up a new path for the direct construction of CG fluid models at the mesoscale based on classical FG (all atom or high-resolution CG) MD simulations.

CG models for non-bonded systems at mesoscale are intrinsically supramolecular, which poses a significant challenge for the design of rigorous bottom up CG mappings. Typical pre-defined CG mappings like the center of mass mapping fail to describe the system well due to diffusion of the FG fluid particles. Physically, the CG blobs should represent spatial chunks of FG fluid particles and, conversely, these FG particles should determine the positions of the CG blobs. The present work is inspired by the ideas behind centroidal Voronoi tessellation to design a dynamical mapping operator within the above-mentioned self-consistent picture, by utilizing a distance-based FG-to-CG proximity function. The new mapping scheme is markedly different from existing mapping approaches in two ways. First, the definition of the CG variables, i.e., the mapping matrix, changes as the FG system propagates. Second, the approach involves two FG to CG correspondence conditions, for both the consistency of positions and velocities. The actual mapping algorithm is built upon the velocity mapping consistency relationship, making it a mapping between trajectories as a whole rather than between instantaneous state points taken independently.

The new mapping scheme was applied to mesoscopically CG supercritical LJ fluid systems under both quiescent conditions and non-equilibrium shear flow, and the results were found to be quite encouraging. The mapped trajectory is numerically stable, continuous, and ergodic with stable CG blob ownerships, which is an advantage over existing approaches with either non-continuous trajectories^{31, 32} or a broad distribution of CG ownerships.^{29, 30} Furthermore, the new mapping approach captures the slow degrees of freedom when applied to a fluid undergoing non-equilibrium shear flow.

A systematic parameterization protocol of the mapped mesoscopic CG model was developed by analogy to the Mori-Zwanzig formalism. A Generalized Langevin Equation was parameterized directly from the mapped CG trajectory with the position and velocity information. The CG conservative pairwise interaction exhibits linear repulsive behavior at the nearest-neighbor range and smoothly decays to zero afterwards, which provides a bottom up justification for the linear repulsive force widely used in the more empirical DPD models. The non-conservative dissipative and random forces derived based on the correlation functions of the mapped trajectory also reproduce the dynamical velocity autocorrelation function. Both the short time behavior of the velocity autocorrelation function and the long time hydrodynamic tail appear to be captured by the GLE.

Due to both the complexity and relative infancy of our new approach, there are still several open questions that should be addressed by us or others in future work. First, the physical and mathematical constraints for the proximity function need to be investigated in more detail. This proximity function is at the core of our theory and the current form is chosen based solely on the numerical stability of the mapping algorithm. With more constraints on the proximity function, one may be able to systematically improve the quality of the CG mappings such that more

physics of the FG system can be preserved. Second, bottom-up parameterization of momentum-conserving CG models, like the forms used in the DPD and SPH models, would be valuable to explore further. These momentum-conserving models are directly related to the governing equations in fluid mechanics and may enable us to reproduce the effects of hydrodynamics in the CG systems under non-equilibrium conditions,⁷⁶ which are expected to be of importance in mesoscopic simulations. Third, the adaptation of the current approach for modeling multicomponent and even multiphase fluid systems of broad interest in engineering and biophysics would also be a key target for future investigation.

APPENDIX A: Derivation of Eq. (4.6) and Eq. (4.10)

This appendix provides the sequence of simplifications that render the first time derivative of the position map equation, Eq. (4.2), into the two forms of the velocity map equations, Eqs. (4.6) and (4.10).

First, we recall the definitions of the FG to CG proximity function $w(|\mathbf{R}_I - \mathbf{r}_i|)$, the ownership f_{li} of each FG particle i by each CG particle I [Eq. (4.1)], and the mapping coefficients c_{li} determining the self-consistent linear relationship between CG and FG particle positions [Eq. (4.2)]. The time derivative of the position map is,

$$\frac{d\mathbf{R}_I}{dt} = \frac{d}{dt} \sum_i c_{li} \mathbf{r}_i, \quad (4.A1)$$

which gives the velocity of the CG blobs, such that

$$\begin{aligned} V_I &= \sum_i \frac{dc_{li}}{dt} \mathbf{r}_i + \sum_i c_{li} \frac{d\mathbf{r}_i}{dt}, \\ &= \sum_i \frac{dc_{li}}{dt} \mathbf{r}_i + \sum_i c_{li} \mathbf{v}_i \end{aligned} \quad (4.A2)$$

This result has the interpretation of a mapped velocity plus a change due to the change in the mapping coefficients. CG particles are drawn to the FG particles for which their mapping coefficients increase. This effect obeys Galilean invariance due to the fact that by construction

$\sum_i c_{li} = 1$ for all times, so $\sum_i \frac{dc_{li}}{dt} = 0$ for all times. Subtracting zero in the form of $\sum_i \frac{dc_{li}}{dt} \mathbf{R}_l$

from the right hand side of Eq. (4.A2) to make the Galilean invariance more evident and make the remainder of the derivation simpler, give the result

$$\mathbf{V}_l = \sum_i \frac{dc_{li}}{dt} (\mathbf{r}_i - \mathbf{R}_l) + \sum_i c_{li} \mathbf{v}_i. \quad (4.A3)$$

At this point one then expands $\frac{dc_{li}}{dt}$. Anticipating the normalization involved in its

definition, it is natural to simplify it in terms of the derivative of the logarithm, $\frac{dc_{li}}{dt} = c_{li} \frac{d \ln c_{li}}{dt}$.

Then, Eq. (4.A3) leads to

$$\begin{aligned} \mathbf{V}_l &= \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_l) \frac{d \ln c_{li}}{dt} + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_l) \frac{d \left(\ln f_{li} m_i - \ln \sum_j f_{lj} m_j \right)}{dt} + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_l) \frac{d \ln f_{li} m}{dt} - \frac{d \ln \sum_j f_{lj} m_j}{dt} \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_l) + \sum_i c_{li} \mathbf{v}_i \end{aligned} \quad (4.A4)$$

Because the second term's sum is zero by the definition of $\mathbf{R}_l$, one can eliminate that term, and since the mass in the first logarithm is constant it also drops out, so

$$\mathbf{V}_l = \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_l) \frac{d \ln f_{li}}{dt} + \sum_i c_{li} \mathbf{v}_i. \quad (4.A5)$$

This result has a natural interpretation: the velocity of a CG particle is the velocity mapped with constant coefficients plus a coefficient weighted average of the rate of change of its log

ownership of each FG particle times the displacement from it to that particle. CG particles are pulled to wherever they gain ownership of new particles proportional to their current ownership, the displacement of the new particles from the center, and the rate of change of log ownership, and inversely proportionally to the total amount of FG particle mass they already own.

The next simplification uses the definition of f_{ji} and another conversion to derivatives of logarithms, such that

$$\begin{aligned}
\mathbf{V}_I &= \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) \frac{d(\ln w_{ji} - \ln \sum_J w_{Ji})}{dt} + \sum_i c_{ji} \mathbf{v}_i \\
&= \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{d \ln w_{ji}}{dt} - \frac{1}{\sum_J w_{Ji}} \sum_J \frac{dw_{Ji}}{dt} \right) + \sum_i c_{ji} \mathbf{v}_i \\
&= \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{d \ln w_{ji}}{dt} - \frac{1}{\sum_J w_{Ji}} \sum_J w_{Ji} \frac{d \ln w_{Ji}}{dt} \right) + \sum_i c_{ji} \mathbf{v}_i, \\
&= \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{d \ln w_{ji}}{dt} - \sum_J \frac{w_{Ji}}{\sum_K w_{Ki}} \frac{d \ln w_{Ji}}{dt} \right) + \sum_i c_{ji} \mathbf{v}_i \\
&= \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) \left(\sum_J (\delta_{IJ} - f_{ji}) \frac{d \ln w_{Ji}}{dt} \right) + \sum_i c_{ji} \mathbf{v}_i \\
&= \sum_J \sum_i c_{ji} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{d \ln w_{Ji}}{dt} + \sum_i c_{ji} \mathbf{v}_i
\end{aligned} \tag{4.A6}$$

Stated differently, the rate of change of log ownership of a FG particle by a CG particle simplifies to the difference between the rate of change in log proximity between those two and the ownership-weighted average rate of change in log proximity between that FG particle and all other CG particles.

Next, because all of the time dependence of w_{ji} is captured by $|\mathbf{R}_J - \mathbf{r}_i|$, its time derivative can be expanded into a proximity scale factor times the projected velocity difference $(\mathbf{R}_J - \mathbf{r}_i) \cdot (\mathbf{V}_J - \mathbf{v}_i) / |\mathbf{R}_J - \mathbf{r}_i|$, such that

$$\mathbf{V}_I = \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \frac{(\mathbf{R}_J - \mathbf{r}_i)}{|\mathbf{R}_J - \mathbf{r}_i|} \cdot (\mathbf{V}_J - \mathbf{v}_i) + \sum_i c_{li} \mathbf{v}_i \quad (4.A7)$$

This is now a linear system of equations for the CG velocities with position-dependent coefficients. The derivations of Eq. (4.10) and the form for its computational implementation, and Eq. (4.6), a more physically transparent form, split at this point. With $R_{li} \equiv |\mathbf{R}_I - \mathbf{r}_i|$, and $\hat{\mathbf{R}}_{li} \equiv (\mathbf{R}_I - \mathbf{r}_i) / |\mathbf{R}_I - \mathbf{r}_i|$ for brevity, we write

$$\mathbf{V}_I = \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot (\mathbf{V}_J - \mathbf{v}_i) + \sum_i c_{li} \mathbf{v}_i. \quad (4.A8)$$

To arrive at Eq. (4.10), we first split $\mathbf{V}_J - \mathbf{v}_i$ and then rearrange the order of summations, yielding

$$\begin{aligned} \mathbf{V}_I &= \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot \mathbf{V}_J \\ &\quad - \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot \mathbf{v}_i + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot \mathbf{V}_J \\ &\quad - \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \mathbf{v}_i \cdot \sum_J (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{Ji}) \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot \mathbf{V}_J \\ &\quad - \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \mathbf{v}_i \cdot \left[\frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} - \sum_J f_{Ji} \frac{\partial \ln w_{Ji}}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \right] + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \left(\delta_{IJ} - \frac{w(R_{Ji})}{\sum_K w(R_{Ki})} \right) \frac{\partial \ln w(r)}{\partial \mathbf{r}} \Big|_{r=|\mathbf{R}_J - \mathbf{r}_i|} \hat{\mathbf{R}}_{Ji} \cdot \mathbf{V}_J \\ &\quad - \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \mathbf{v}_i \cdot \left[\frac{\partial \ln w(R_{li})}{\partial \mathbf{r}_i} - \frac{\partial \ln \sum_J w(R_{Ji})}{\partial \mathbf{r}_i} \right] + \sum_i c_{li} \mathbf{v}_i \\ &= \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \left(\delta_{IJ} - \frac{w(R_{Ji})}{\sum_K w(R_{Ki})} \right) \frac{d \ln w(R_{Ji})}{d \mathbf{R}_J} \cdot \mathbf{V}_J \\ &\quad - \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \mathbf{v}_i \cdot \left[\frac{\partial \ln w(R_{li})}{\partial \mathbf{r}_i} - \frac{\partial \ln \sum_J w(R_{Ji})}{\partial \mathbf{r}_i} \right] + \sum_i c_{li} \mathbf{v}_i \end{aligned} \quad (4.A9)$$

Now, introducing the shorthand notation $W_i = \sum_l w_{li}$, Eq. (4.A9) can be simplified as

$$V_I = \sum_J \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{\delta_{IJ}}{w_{li}} - \frac{1}{W_i} \right) \frac{\partial w_{ji}}{\partial \mathbf{R}_J} \cdot \mathbf{V}_J + \sum_i c_{li} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{1}{w_{li}} \frac{\partial w_{li}}{\partial \mathbf{r}_i} - \frac{1}{W_i} \frac{\partial W_i}{\partial \mathbf{r}_i} \right) \cdot \mathbf{v}_i + \sum_i c_{li} \mathbf{v}_i, \quad (4.A10)$$

With rearrangement, this is Eq. (4.10) – the equation used for the computational implementation as described in the main text.

To derive Eq. (4.6), one instead rearranges Eq. (4.A7) into an equation for the difference between the CG velocity including motion due to changes in the mapping coefficients and the velocity mapped with constant coefficients, such that

$$\begin{aligned} \mathbf{V}_I - \sum_i c_{li} \mathbf{v}_i &= \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot (\mathbf{V}_J - \mathbf{v}_i) \\ &= \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot (\mathbf{V}_J - \sum_j c_{lj} \mathbf{v}_j + \sum_j c_{lj} \mathbf{v}_j - \mathbf{v}_i) \\ &= \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot (\mathbf{V}_J - \sum_j c_{lj} \mathbf{v}_j) \\ &\quad - \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot (\mathbf{v}_i - \sum_j c_{lj} \mathbf{v}_j) \end{aligned} \quad (4.A11)$$

Defining $\bar{\mathbf{V}}_I \equiv \sum_i c_{li} \mathbf{v}_i$ and $\delta \mathbf{V}_I \equiv \mathbf{V}_I - \bar{\mathbf{V}}_I$ for brevity as in the main text, this is given by

$$\begin{aligned} \delta \mathbf{V}_I &= \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot \delta \mathbf{V}_J \\ &\quad - \sum_J \sum_l c_{li} (\mathbf{r}_i - \mathbf{R}_I) (\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} \cdot (\mathbf{v}_i - \bar{\mathbf{V}}_J) \end{aligned} \quad (4.A12)$$

One can now reverse the log-ownership derivative formula, $(\delta_{IJ} - f_{ji}) \frac{\partial \ln w(r)}{\partial r} \Big|_{r=R_{ji}} \hat{\mathbf{R}}_{ji} = \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J}$

, to show that

$$\delta \mathbf{V}_I = \sum_J \sum_i c_{li}(\mathbf{r}_i - \mathbf{R}_I)(\delta_{IJ} - f_{JI}) \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J} \cdot \delta \mathbf{V}_J - \sum_J \sum_i c_{li}(\mathbf{r}_i - \mathbf{R}_I) \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J} \cdot (\mathbf{v}_i - \bar{\mathbf{V}}_J). \quad (4.A13)$$

The second term on the right-hand-side accounts for the change in ownership of particles that will be result if one propagated the equation of motion of the CG system with the velocity mapped as if the coefficients were constant. The first term on the right-hand-side describes how any resulting velocity corrections will lead to further corrections.

Grouping CG and FG terms, one finds

$$\sum_J \left(\delta_{IJ} \mathbf{I} - \sum_i c_{li}(\mathbf{r}_i - \mathbf{R}_I)(\delta_{IJ} - f_{JI}) \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J} \right) \cdot \delta \mathbf{V}_J = \sum_J \sum_i c_{li}(\mathbf{r}_i - \mathbf{R}_I) \frac{\partial \ln f_{li}}{\partial \mathbf{R}_J} \cdot (\mathbf{v}_i - \bar{\mathbf{V}}_J), \quad (4.A14)$$

This result is Eq. (4.6).

APPENDIX B: Numerical Implementation

The implementation of the dynamic mapping algorithm is summarized in Table 4-I. The FG molecular dynamics simulation trajectory, the inverse CG resolution (the ratio of the number of FG particles over CG beads) and the proximity function are pre-defined as input to the program. The initial CG configuration was generated through a self-consistent solver. Then for each FG trajectory frame, the matrices and vectors in Eq. (4.10) were constructed and the CG velocities were directly calculated. The CG configurations were propagated through a first order Euler integrator as discussed before in Sec. II. To ensure the position consistency condition Eq. (4.2), the CG configurations were re-mapped through the self-consistent solver at a certain frequency. Typically, a re-mapping frequency of one per 2000~5000 time steps is sufficient to maintain a stable mapping.

TABLE 4-I. Description of the FG to CG mapping algorithm.

Algorithm 1: Mapping Procedure
Input Data: FG simulation trajectory, CG Resolution, proximity function $w(r)$
Output: CG trajectory
Start:
Generate Initial CG Configuration
Randomly initialize the first CG frame;
Solve the configuration mapping equation $\mathbf{R}_I = \sum_i c_{li} \mathbf{r}_i$ self-consistently;
for $i := 1$ to <i>timesteps</i> do
Read one frame of trajectory from FG simulation data;
Build CG-FG cell list and neighbor list using current CG and FG configuration;
Construct the matrix equation:
Construct weighting matrix W , derivative matrix dW ,
$w_{li} = w(R_{li}), \quad dw_{li} = \frac{dw}{dr} \Big _{R_{li}} \mathbf{e}_{li}$
$W_i = \sum_j w_{ji}, \quad dW_i = \sum_j dw_{ji}$
Construct matrix C with matrix elements described by Eq. (4.2);
Construct matrix M according to the left hand side of Eq. (4.10);
Construct matrix V according to the right hand side of Eq. (4.10);
Solve the linear equation $(\mathbf{I} - \mathbf{M})\mathbf{V} = (\mathbf{C} + \mathbf{N})\mathbf{v}$ to obtain instantaneous CG velocities;
Propagate the CG beads position for one time step $\mathbf{R}(t + \Delta t) = \mathbf{R}(t) + \mathbf{V}(t)\Delta t$;
if Re-mapping == True do:
Solve the configuration mapping equation $\mathbf{R}_I = \sum_i c_{li} \mathbf{r}_i$ self-consistently;
Output positions, velocities, and other relevant physical quantities of the current CG frame;
done

In the above algorithm, there are two more points worth detailed explanation so one can reproduce the results we present in this work.

(1) Self-Consistent Solver

Two crucial steps in the above-mentioned algorithm, one in the initialization of CG system and another in the re-mapping procedure to ensure position consistency, involve the self-consistent determination of CG beads positions from the FG trajectory frame. The detailed fixed-point algorithm for our self-consistent solver is described in the algorithm chart in Table 4-II. An

update rate $\alpha = 0.30$ and tolerance 10^{-20} was applied throughout all the calculations in this work.

The self-consistent solver typically converged within 500 iterations.

TABLE 4-II. Description of the self-consistent position mapping algorithm.

Algorithm 2: Self-Consistent Solver
Input Data: FG trajectory frame, proximity function $w(r)$, initial guess of CG coordinates, tolerance, update rate α
Output: CG beads coordinates
Start:
square_error = INFINITY;
construct Matrix C;
R = Initial guess of CG coordinates;
while square_error > tolerance do :
R_Old <- R
R_New <- calculate_CG_coordinates(C, r);
R = $(1 - \alpha)R_{\text{Old}} + \alpha R_{\text{New}}$;
square_error = $ \mathbf{R} - \mathbf{R}_{\text{old}} ^2$;
done

(2) Two neighbor lists

In order to speed up the evaluation of relevant matrices, two neighbor lists were explicitly maintained during the propagation. One neighbor list kept track of all the FG particles within the cut-off distance from a certain CG bead, whereas the other neighbor list kept track of all the CG beads within the cut-off distance from a certain FG particle. A cell list of the FG particles was constructed first to accelerate the building of these two neighbor lists. With the help of these data structures, the cost of constructing matrices **M**, **N** and **C** was reduced to scale linearly with the size of the FG system.

Reference

1. P. de Gennes, *Scaling concepts in polymer physics*. (Cornell University Press, Ithaca, NY, 1979).

2. P. J. Hoogerbrugge and J. M. V. A. Koelman, *Europhys. Lett.* **19**, 155 (1992).
3. P. Español and P. Warren, *Europhys. Lett.* **30**, 191 (1995).
4. P. B. Warren, *Curr. Opin. Colloid Interface Sci.* **3**, 620 (1998).
5. M. B. Liu, G. R. Liu, L. W. Zhou and J. Z. Chang, *Arch. Comput. Methods Eng.* **22**, 529 (2015).
6. L. B. Lucy, *Astron. J.* **82**, 1013 (1977).
7. R. A. Gingold and J. J. Monaghan, *Mon. Not. R. Astron. Soc.* **181**, 375 (1977).
8. M. B. Liu and G. R. Liu, *Arch. Comput. Methods Eng.* **17**, 25 (2010).
9. Z.-B. Wang, R. Chen, H. Wang, Q. Liao, X. Zhu and S.-Z. Li, *Appl. Math. Modell.* **40**, 9625 (2016).
10. J. T. Padding and W. J. Briels, *J. Phys.: Condens. Matter* **23**, 233101 (2011).
11. J. T. Padding and W. J. Briels, *J. Chem. Phys.* **117**, 925 (2002).
12. V. Symeonidis, G. Em Karniadakis and B. Caswell, *Phys. Rev. Lett.* **95**, 076001 (2005).
13. N. A. Spenley, *Europhys. Lett.* **49**, 534 (2000).
14. J. M. V. A. Koelman and P. J. Hoogerbrugge, *Europhys. Lett.* **21**, 363 (1993).
15. D. S. Bolintineanu, G. S. Grest, J. B. Lechman, F. Pierce, S. J. Plimpton and P. R. Schunk, *Comput. Part. Mech.* **1**, 321 (2014).
16. J. C. Shelley and M. Y. Shelley, *Curr. Opin. Colloid Interface Sci.* **5**, 101 (2000).
17. R. D. Groot and K. L. Rabone, *Biophys. J.* **81**, 725 (2001).
18. G. Ayton and G. A. Voth, *Biophys. J.* **83**, 3357 (2002).
19. M. Venturoli, M. Maddalena Sperotto, M. Kranenburg and B. Smit, *Phys. Rep.* **437**, 1 (2006).
20. D. A. Fedosov, W. Pan, B. Caswell, G. Gompper and G. E. Karniadakis, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11772 (2011).
21. P. Español, M. Serrano and I. Zuñiga, *Int. J. Mod. Phys. C* **08**, 899 (1997).

22. C. Hijon, P. Espanol, E. Vanden-Eijnden and R. Delgado-Buscalioni, *Faraday Discuss.* **144**, 301 (2010).
23. K. S. Schweizer, *J. Chem. Phys.* **91**, 5802 (1989).
24. T. Kinjo and S.-a. Hyodo, *Phys. Rev. E* **75**, 051109 (2007).
25. Z. Li, X. Bian, B. Caswell and G. E. Karniadakis, *Soft Matter* **10**, 8659 (2014).
26. P. Español and P. B. Warren, *J. Chem. Phys.* **146**, 150901 (2017).
27. E. G. Flekkøy and P. V. Coveney, *Phys. Rev. Lett.* **83**, 1775 (1999).
28. E. G. Flekkøy, P. V. Coveney and G. De Fabritiis, *Phys. Rev. E* **62**, 2140 (2000).
29. A. Eriksson, M. N. Jacobi, J. Nyström and K. Tunstrøm, *Europhys. Lett.* **86**, 44001 (2009).
30. A. Eriksson, M. N. Jacobi, J. Nyström and K. Tunstrøm, *Phys. Rev. E* **77**, 016707 (2008).
31. K. R. Hadley and C. McCabe, *J. Phys. Chem. B* **114**, 4590 (2010).
32. S. Izvekov and B. M. Rice, *J. Chem. Phys.* **140**, 104104 (2014).
33. J. Zavadlav, S. J. Marrink and M. Praprotnik, *J. Chem. Theory Comput.* **12**, 4138 (2016).
34. J. Zavadlav and M. Praprotnik, *J. Chem. Phys.* **147**, 114110 (2017).
35. G. S. Ayton, H. L. Tepper, D. T. Mirijanian and G. A. Voth, *J. Chem. Phys.* **120**, 4074 (2004).
36. R. Zwanzig, *J. Chem. Phys.* **33**, 1338 (1960).
37. R. Zwanzig, *Phys. Rev.* **124**, 983 (1961).
38. R. Zwanzig, *Physica* **30**, 1109 (1964).
39. H. Mori, *Prog. Theor. Phys.* **33**, 423 (1965).
40. S. Nosé, *J. Chem. Phys.* **81**, 511 (1984).
41. W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985).
42. S. Plimpton, *J. Comput. Phys.* **117**, 1 (1995).

43. G. K. Batchelor, *An Introduction to Fluid Dynamics*. (Cambridge Univ. Press, Cambridge, UK, 1967).
44. M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.* **42**, 73 (2013).
45. W. G. Noid, *J. Chem. Phys.* **139**, 090901 (2013).
46. F. Müller-Plathe, *Phys. Rev. E* **59**, 4894 (1999).
47. S. Izvekov and G. A. Voth, *J. Chem. Phys.* **125**, 151101 (2006).
48. L. Gao and W. Fang, *J. Chem. Phys.* **135**, 184101 (2011).
49. S. Izvekov, *J. Chem. Phys.* **138**, 134106 (2013).
50. Z. Li, X. Bian, X. Li and G. E. Karniadakis, *J. Chem. Phys.* **143**, 243128 (2015).
51. Y. Yoshimoto, I. Kinefuchi, T. Mima, A. Fukushima, T. Tokumasu and S. Takagi, *Phys. Rev. E* **88**, 043305 (2013).
52. A. Davtyan, J. F. Dama, G. A. Voth and H. C. Andersen, *J. Chem. Phys.* **142**, 154104 (2015).
53. A. Davtyan, G. A. Voth and H. C. Andersen, *J. Chem. Phys.* **145**, 224107 (2016).
54. H. Lei, B. Caswell and G. E. Karniadakis, *Phys. Rev. E* **81**, 026704 (2010).
55. Z. Li, H. S. Lee, E. Darve and G. E. Karniadakis, *J. Chem. Phys.* **146**, 014104 (2017).
56. H. Lei, X. Yang, Z. Li and G. E. Karniadakis, *J. Comput. Phys.* **330**, 571 (2017).
57. P. Español, *Phys. Rev. E* **53**, 1572 (1996).
58. S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
59. S. Izvekov and G. A. Voth, *J. Chem. Phys.* **123**, 134105 (2005).
60. W. G. Noid, J.-W. Chu, G. S. Ayton and G. A. Voth, *J. Phys. Chem. B* **111**, 4116 (2007).
61. W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).
62. W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen and G. A. Voth, *J. Chem. Phys.* **128**, 244115 (2008).

63. L. Lu, S. Izvekov, A. Das, H. C. Andersen and G. A. Voth, *J. Chem. Theory Comput.* **6**, 954 (2010).
64. I. Pagonabarraga and D. Frenkel, *J. Chem. Phys.* **115**, 5015 (2001).
65. P. B. Warren, *Phys. Rev. Lett.* **87**, 225702 (2001).
66. S. Y. Trofimov, E. L. F. Nies and M. A. J. Michels, *J. Chem. Phys.* **117**, 9383 (2002).
67. P. Español and M. Revenga, *Phys. Rev. E* **67**, 026705 (2003).
68. J. F. Dama, J. Jin and G. A. Voth, *J. Chem. Theory Comput.* **13**, 1010 (2017).
69. J. W. Wagner, T. Dannenhoffer-Lafage, J. Jin and G. A. Voth, *J. Chem. Phys.* **147**, 044113 (2017).
70. J. Jin and G. A. Voth, *J. Chem. Theory Comput.* **14**, 2180 (2018).
71. H. K. Shin, C. Kim, P. Talkner and E. K. Lee, *Chem. Phys.* **375**, 316 (2010).
72. G. Jung, M. Hanke and F. Schmid, *J. Chem. Theory Comput.* **13**, 2481 (2017).
73. H. Lei, N. A. Baker and X. Li, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14183 (2016).
74. J. A. Nelder and R. Mead, *Comput. J.* **7**, 308 (1965).
75. J. P. Boon and S. Yip, *Molecular Hydrodynamics*. (Dover Publications, New York, U.S.A., 1991).
76. P. Español, *Phys. Rev. E* **52**, 1734 (1995).

Chapter 5: Constructing Many-body Dissipative Particle Dynamics Models of Fluids from Rigorous Bottom-up Coarse-graining

Abstract

Since their emergence at 1990s, mesoscopic models of fluids have been widely used to study complex organization and transport phenomena beyond molecular scale. These models are designed based on results from physics at meso- and macroscale, such as fluid mechanics and statistical field theory. The microscopic foundation of these models, on the other hand, is often not well defined. The present paper tries to build such a connection through bottom-up coarse-graining methods, which combines a recently developed dynamic coarse-graining scheme with new developments in the parameterization algorithm. A statistical inference framework of explicit many-body conservative interaction is introduced, which quantitatively recapitulates the mesoscopic structure of underlying fluid. A new general algorithm for parameterizing dissipative and fluctuation forces is also developed, which is utilized to derive pairwise decomposable friction kernels under both non-Markovian and Markovian limit. The short time and long time feature of the coarse-grained dynamics is perfectly reproduced while the backscattering feature at medium lag time can be qualitatively described. Through these new developments, a many-body dissipative particle dynamics type equation of motion is successfully derived. The method presented in this work open up an avenue for the direct bottom-up construction of mesoscopic models of fluids that naturally connect with meso- and macroscopic physics.

I. Introduction

With rapidly expanding attention on soft matter and biophysics research, various particle based mesoscopic CG modeling techniques have been developed and successfully applied to simulate complex organization and transport phenomena beyond molecular scale. The success of these techniques is not an accident. Many of the methods have solid theoretical foundation from

meso- or macro-scale physics. For example, the Smoothed Particle Hydrodynamics¹⁻³ (SPH) and Smoothed Dissipative Particle Dynamics⁴⁻⁵ (SDPD) methods are derived through discretizing the Navier-Stokes equation using smooth kernel functions, and thus naturally preserve fluid mechanics phenomena. The Dissipative Particle Dynamics⁶⁻⁸ (DPD) method, as well as many method derived from it⁹ such as many-body DPD,¹⁰⁻¹¹ directly linked the many-body conservative interactions with the equations of state of the fluids,¹⁰ therefore can be successfully applied to simulate interfacial behavior of fluids at mesoscale.¹² Despite the solid connections with macroscopic physics, their connection with microscopic description of the system is often unclear. This missing link obscures the understanding of how the chemistry of molecular systems determines their meso- and macroscopic behavior.

The desirability of connecting atomistic description of fluids with mesoscopic models and eventually macroscopic continuum fluid mechanics has gain its popularity since the early trials by Espanol *et al* in the late 1990s,¹³ when they attempted to construct coarse-grained (CG) models of fluids based on Voronoi tessellation. The first step of building such links is to design a rigorous coarse-grained mapping function that attributes multiple fine-grained (FG) entities into a CG representation, which achieves the goal of mapping a fine-grained trajectory of the system into a CG representation. Then the information of mesoscopic CG trajectories is utilized to parameterize CG dynamics models that directly connect with meso- or macroscopic physics.

For molecular fluids composed of bonded atoms such as melted polymers, this connection can be well established through a linear CG mapping operator.⁹ The time-invariant and linear property of this type of CG mapping also facilitates the bottom-up parameterization of mesoscopic CG models based on the mapped trajectory. A number of well-established bottom-up CG methods can be utilized to derive interactions under this situation. Hijon *et al* took the first

step in using Mori-Zwanzig formalism¹⁴⁻¹⁷ to derive bottom-up DPD models of polymers.¹⁸ Karniadakis and coworkers,¹⁹⁻²² as well as Yoshimoto *et al*²³⁻²⁴ then greatly extended the scope of this strategy that significantly improved the accuracy and feasibility of the parameterization process.

On the other hand, for fluids composed of unbonded molecules, the design of FG to CG correspondence is much harder and is considered to be one of the major open questions in the field of mesoscale modeling.⁹ Flekkoy, Coveney and Eriksson published a series of influential work on deriving DPD equations of motion from *semi* bottom-up,²⁵⁻²⁸ in which the Voronoi tessellation based idea was first introduced. However, the consistency between FG and CG configurations is not enforced in these derivations. Following the spatial tessellation idea, Hadley *et al*,²⁹ Izvekov *et al*³⁰ as well as Praprotnik and coworkers³¹⁻³³ utilized various clustering algorithms to build this CG to FG correspondence in configuration space. Though these methods have provided useful insight, neither of them has achieved a continuous mapping that accounts for essential mass and momentum transfer at mesoscale. In a recent work,³⁴ we developed a dynamic coarse-graining mapping scheme and for the first time achieved such continuous mesoscopic mapping of fluids that quantitatively captures slow collective motions of the underlying fluid system. However, due to the nonlinearity and time-varying nature of this new mapping operator, the parameterization of CG models compatible with mesoscopic physics based on this new approach is challenging and yet to be explored.³⁴⁻³⁵ The goal of this work is then to develop such a parameterization method to solve this problem.

It has long been demonstrated that the dynamical evolution of reduced representation of molecular systems can be described by a set of stochastic integro-differential equations through the Mori-Zwanzig formalism. In particular, the equations of motion can be written as:

$$\begin{aligned}\dot{\mathbf{R}}_I &= \mathbf{M}_I^{-1} \mathbf{P}_I \\ \dot{\mathbf{P}}_I &= -\frac{1}{\beta} \frac{\partial}{\partial \mathbf{R}_I} \ln \omega(\mathbf{R}) - \beta \sum_{J=1}^N \int_0^t ds \left\langle \left[\delta \mathbf{F}_I^Q(t-s) \right] \cdot \left[\delta \mathbf{F}_J^Q(0) \right] \right\rangle \frac{\mathbf{P}_J(s)}{M_J} + \delta \mathbf{F}_I^Q(t)\end{aligned}\quad (5.1)$$

In which the force exerted on the coarse-grained particles has three components, the thermodynamic potential of mean force (PMF) determining the equilibrium statistics of coarse-grained system, the dissipative force and fluctuation force controlling the dynamical correlation. Though physically intuitive, the bottom-up parameterization of these three components is not trivial. There are mainly three barriers one needs to overcome to make it numerically feasible.

First of all, the exact dynamical equation Eq. (5.1) cannot serve as a practical computation tool due to its extremely large computational cost. One needs to make appropriate approximations to reduce the algorithmic complexity, while capturing essential physics at mesoscale at the same time. In order to correctly recapitulate hydrodynamics results, it requires that the CG equations of motion preserve the total momentum and be invariant under Galilean transformation. The simplest representation that fulfills both requirements is probably the pairwise form, in which the many-body PMF, dissipative forces as well as fluctuation forces are all assumed to be pairwise decomposable. This type of equations of motion is commonly used in the DPD and SDPD models and we would like to build up our theory and algorithm based on this simple case at the first stage.

Second, the PMF term defined by $\omega(\mathbf{R}) = \int d\mathbf{r} \delta(M(\mathbf{r}) - \mathbf{R}) e^{-\beta u(\mathbf{r})} / \int d\mathbf{r} e^{-\beta u(\mathbf{r})}$ is a many-body quantity. In our previous work, we have shown that a pairwise decomposition of such many-body PMF is not sufficient to reproduce the correct equilibrium structure of fluids.^{34, 36-38} An explicit many-body interaction needs to be included in the basis set to compensate this discrepancy. The necessity of explicit many-body interaction is also manifested in the top-down Lagrangian description of fluids, such as many-body DPD, SPH and SDPD models, in which the

conservative interactions between mesoscopic CG particles usually contain a many-body element with the following form,

$$\mathbf{F}_j^m = -\sum_i \left(p(\rho_i) + p(\rho_j) \right) \nabla_j W(r_{ij}). \quad (5.2)$$

In Eq. (5.2) the local density dependent term $p(\rho)$ is directly connected to the equations of state (EOS) of the fluid.¹⁰ This type of many-body interaction has profound theoretical connections with hydrodynamics^{2, 5} and classical density functional theory (DFT) of liquids.³⁹⁻⁴² In this work, we would like to develop a systematic bottom-up parameterization strategy based on the well-established Multiscale Coarse-graining⁴³⁻⁴⁷ (MS-CG) framework to reconstruct many-body conservative interactions of this kind.

And the last barrier originates from the statistical inference of the dissipative friction kernel which is also intrinsically many-body, and in addition highly correlated in time. It is shown in the exact dynamical equation Eq. (5.1) that the dissipative friction kernel is a non-Markovian many-body quantity. And due to the restriction imposed from fluctuation dissipation theorem, the friction kernel and the random fluctuations are correlated with each other. These two characters make the statistical inference of the friction kernel a much harder problem than the inference of many-body PMF. To make the inference practical, the friction kernel is usually assumed to be either pairwise decomposable or applies only on single particle basis, which renders the DPD or Generalized Langevin Equation (GLE) type equations of motions.

The inference of GLE is relatively easier and has been successfully achieved through different approach.^{34, 48-50} However, the single particle approximation adopted in GLE violates the momentum conservation, therefore hinders the connection of GLE with fluid mechanics. As a consequence, it is rarely used in the mesoscopic modeling community. The pairwise DPD type friction kernel does not suffer from this limitation and thus has been widely adopted.

The bottom-up inference of the DPD type pairwise friction is highly non-trivial. There are mainly two strategies for solving the problem. The first strategy is based upon the Mori-Zwanzig formalism, in which the friction kernel can be directly calculated once the fluctuation forces can be obtained. This can be done through the “constraint dynamics”, in which the fine-grained system is propagated conditioned on the invariance of mapped CG configurations.^{18, 51-53} The second strategy applies a correlation function inversion method, by solving a Volterra type integral equation to match the pairwise velocity autocorrelation function and force-velocity cross correlation function.²⁰⁻²³ Though these two methods have shown success in reproducing essential CG dynamics of bonded systems, the application to more general cases is limited. Both of the two strategies can only be applied to fine-grained systems interacting through pairwise interactions and the coarse-graining mapping operator is linear and static. New methods need to be worked out for situations where the instantaneous forces cannot be decomposed in a pairwise manner. And this new development will be thoroughly discussed in this work.

The remainder contents of the paper is organized as follows, in Sec. II, we elaborate the details of the methods that we developed to build mesoscopic CG dynamics models that naturally connect to physics at mesoscale. These new developments are then applied to a numerical example that constructs a many-bod DPD type CG equations of motion. The results and relevant discussions are presented in Sec. III. Sec. IV concludes this paper.

II. Methodology

A. Mesoscopic Representation of Fluids

In this work, we utilize the recently developed dynamical coarse-grained mapping scheme to construct the mesoscopic representation of fluid. Here we briefly encapsulate the theory,

algorithm and essential practical concerns to reproduce the result in this paper. Readers are suggested to refer to the original work³⁴ for more detailed discussions.

The dynamical mapping algorithm, as distinct from conventional pre-defined CG mapping operators, establishes the self-consistent correspondence between FG and CG phase space by performing mapping in the momentum space, making it a mapping between trajectories as a whole rather than between instantaneous configurations taken independently. This unique feature ensures the preservation of mesoscopic dynamics of underlying fine-grained fluid, as is demonstrated by numerical experiments in previous research.

The mapping scheme is designed based on the intuition behind Voronoi spatial tessellation that attributes multiple FG particles to a specific CG blob, and center of mass correspondence between CG and FG particles. To achieve a smooth, time-homogenous, translational and rotational invariant CG mapping at supramolecular resolution based on these intuition, we proposed a smoothed generalization of centroidal Voronoi tessellation algorithm commonly used in computational geometry.⁵⁴ Specifically, we define a proximity weighting function $w(r)$ of distance r and assign each fine-grained identity i to each CG blob I with a ownership fraction:

$$f_{Ii} = \frac{w(|\mathbf{R}_I - \mathbf{r}_i|)}{\sum_j w(|\mathbf{R}_j - \mathbf{r}_i|)} \equiv \frac{w_{Ii}}{\sum_j w_{ji}}. \quad (5.3)$$

This ownership fraction smoothed out the Heaviside stepwise identity function in the discrete Voronoi tessellation while still preserving the proximity information once $w(r)$ is monotonically decreasing. Next, the centers of the CG blobs are assigned as the weighted center of mass of their owned atoms weighted by the ownership fraction:

$$\mathbf{R}_I = \sum_i \frac{f_{Ii} m_i \mathbf{r}_i}{\sum_j f_{Ij} m_j} \equiv \sum_i c_{Ii} \mathbf{r}_i. \quad (5.4)$$

The above two steps determine a self-consistent equation Eq. (5.4) finding the blob positions, as shown schematically in the Stage 1 of Fig. 5-1. It should be noted that the above self-consistent equation has many feasible solutions. Therefore, one cannot guarantee a smoothly propagating CG trajectory by solving the position mapping equation at every instant of time. Instead, the dynamical trajectory of CG blobs is generated based on the velocity correspondence equation,

$$\sum_j \left(\delta_{IJ} \mathbf{I} - \sum_i (\mathbf{r}_i - \mathbf{R}_I) c_{Ii} \frac{\partial \ln f_{Ii}}{\partial \mathbf{R}_J} \right) \cdot \delta \mathbf{V}_J = \sum_i \sum_J (\mathbf{r}_i - \mathbf{R}_I) c_{Ii} \frac{\partial \ln f_{Ii}}{\partial \mathbf{R}_J} \cdot (\bar{\mathbf{V}}_J - \mathbf{v}_i), \quad (5.5)$$

which can be derived by evaluating the time propagation of the position mapping equation. Here the direct linear mapping velocity $\bar{\mathbf{V}}_J = \sum_i c_{Ji} \mathbf{r}_i$ and $\delta \mathbf{V}_J = \mathbf{V}_J - \bar{\mathbf{V}}_J$. This set of equation can be written in a compact tensor equation form after necessary simplification,

$$(\mathbf{I} - \mathbf{M}) \mathbf{V} = (\mathbf{C} + \mathbf{N}) \mathbf{v}, \quad (5.6)$$

in which M and N are matrices with 3x3 tensor elements $\mathbf{M}_{IJ} = \sum_i c_{Ii} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{\delta_{IJ}}{w_{Ii}} - \frac{1}{W_i} \right) \frac{\partial w_{Ii}}{\partial \mathbf{R}_J}$

and $\mathbf{N}_{Ii} = c_{Ii} (\mathbf{r}_i - \mathbf{R}_I) \left(\frac{1}{w_{Ii}} \frac{\partial w_{Ii}}{\partial \mathbf{r}_i} - \frac{1}{W_i} \frac{\partial W_i}{\partial \mathbf{r}_i} \right)$. The total weight is defined as $W_i = \sum_I w_{Ii}$.

Observing that the velocity mapping equation is not self-consistent anymore, there is a unique solution as long as the matrix $\mathbf{I} - \mathbf{M}$ is invertible. Given any fine-grained trajectory with continuous positions $\mathbf{r}(t)$ and velocities $\mathbf{v}(t)$ information, one can solve the CG velocities $\mathbf{V}(t)$ once a proper initial guess of the CG configurations $\mathbf{R}_I(0)$ are in place. This CG velocity trajectory provides a way to propagate CG configurations $\mathbf{R}_I(t)$ from one specific choice of initial configuration $\mathbf{R}_I(0)$, which can be solved iteratively from Eq. (5.4). The degeneracy of

the position mapping equation Eq. (5.4) is reconciled by following the dynamics of a single feasible solution satisfying position correspondence.

The procedure established above is not a direct map of single configurations but rather a map of trajectories. Mathematically, this process corresponds to a flow defined in Euclidean space. Even though we are still at the early stage development of this methodology, it has been demonstrated that this mapping scheme is capable of generating continuous, ergodic and initial conditions independent mesoscopic CG trajectories that could capture essential slow motions of the underlying fine-grained fluid.

For actual numerical implementation of this idea, we first generate an initial CG configuration by solving Eq. (5.4) self-consistently. And then for each time frame of the fine-grained trajectory, we construct two neighbor lists keeping track of the local neighbors of each fine-grained and CG particles (Stage 2 in Fig. 5-1), so that it reduces all the matrices construction process to scale linearly with respect to the number of fine-grained particle. With the neighbor lists in place, the velocity mapping equation Eq. (5.5) is constructed (Stage 3 in Fig. 5-1). It is worth noting that all the matrices and vectors involved could be evaluated by a single or double loop through the neighbor lists. If the weighting function is chosen to be short range, which is the case for the implementation in this work, the whole process is rendered to exhibit an appealing linear scaling algorithmic complexity. And finally with the update velocities in place, the CG configurations are propagated for one time step forward through a simple Euler's forward integration,

$$\mathbf{R}_I(t + \Delta t) = \mathbf{R}_I(t) + \Delta t \mathbf{V}_I(\{\mathbf{R}_J(t)\}, \{\mathbf{r}_j(t)\}, \{\mathbf{v}_j(t)\}), \quad (5.7)$$

as shown in the Stage 4 of Fig. 5-1. The process is repeated for each incoming fine-grained frame and finally generates a continuous mesoscopic CG mapped trajectory. However, the Stage

4 propagation suffers from a numerical difficulty that the discretization error slowly gathers leading to a drift from the self-consistent position mapping relationship. To overcome this numerical issue, the self-consistent position mapping equation Eq. (5.4) is periodically re-enforced to pull the drifted trajectory back on the correct track. This procedure is done at a much longer time interval, usually at a frequency of tens of thousands of time steps is sufficient to maintain the numerical stability.

One last piece, yet an important aspect, of the algorithm is choosing the appropriate proximity weight function. In this work, the previously chosen truncated hyperbolic tangent function is adopted with the following form,

$$w(r) = \frac{1}{4} \left[1 - \tanh \left(\frac{r / r_c - 1}{\sigma} \right) \right] \operatorname{erfc} \left(\frac{r / r_c - \beta}{\gamma} \right), \quad (5.8)$$

whose shape can be tuned by the parameter sets σ , r_c , β and γ to realize stable propagation.

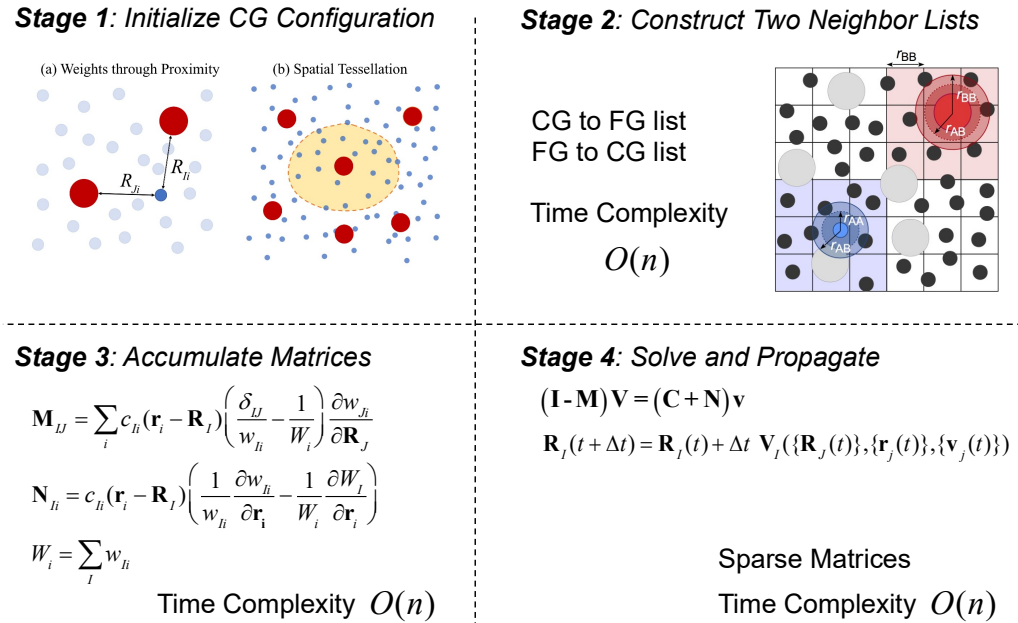


Figure 5-1, illustration of the essential steps of the dynamical coarse-grained mapping procedure.

B. Statistical Inference of Many-body Conservative Interaction

As expounded earlier in the introduction, the conservative interactions between the CG blobs is intrinsically many-body and the explicit many-body interactions have deep theoretical foundations from classical density functional theory as well as fluid mechanics. In this section, we elaborate our approach to parameterize a many-body DPD type conservative interaction directly from the mapped mesoscopic CG trajectory.

In top-down Lagrangian descriptions of fluids, the many-body conservative interaction originated from the momentum conservation equations is a natural result of fluid mechanics. In the SPH and SDPD equations of motions, the many-body conservative force takes the following form:

$$\mathbf{F}_i^c = \sum_j -m_i m_j \left(\frac{p_i}{\bar{\rho}_i^2} + \frac{p_j}{\bar{\rho}_j^2} \right) \frac{\partial W(r_{ij})}{\partial r_{ij}} \mathbf{e}_{ij}, \quad (5.9)$$

in which the local density is defined as $\bar{\rho}_i = \sum_j m_j W(r_{ij})$ and the pressure p_i is usually chosen as a function of local density, which correspond to the equations of state (EOS) of the materials under study. The smooth kernel function $W(r_{ij})$ is introduced to discretize the continuous Navier-Stokes equations, usually take the form of Lucy's function.²⁻³ This theoretical result of hydrodynamics suggests a necessity of including many-body basis functions in the bottom-up derivation of mesoscopic CG interactions.

From another perspective, the same form of interaction also emerges in the classical density functional theory to study the equilibrium properties of fluids.³⁹⁻⁴¹ In agreement with the idea that the mesoscopic CG blobs refer to lumps of fluid, one can generally express the Helmholtz free energy of the fluid system as,

$$A = \int d\mathbf{r} \rho(\mathbf{r}) f(n(\{\mathbf{r}\})), \quad (5.10)$$

in which the function $f(\{n(\mathbf{r})\})$ correspond to the local free energy per particle, and $n(\mathbf{r})$ is a function with local property of the system. In particular, the simplest assumption of $n(\mathbf{r}) = \rho(\mathbf{r})$ symbolizes the local density approximation. And for a more smeared perspective in which $n(\mathbf{r})$ is chosen to be an average of densities within a small spatial interval, it can have an interpretation as a weighted density approximation. For a mesoscopic system under particle description, the local density $\rho(\mathbf{r})$ is defined as $\rho(\mathbf{r}) = \sum_i \delta(\mathbf{r}_i - \mathbf{r})$. Plugging this definition into the free energy expression, one obtains that,

$$A = \int d\mathbf{r} \sum_i \delta(\mathbf{r}_i - \mathbf{r}) f(n(\{\mathbf{r}\})) = \sum_i f(n(\{\mathbf{r}_i\})), \quad (5.11)$$

i.e. the free energy is simply a sum of single particle free energy. With the free energy defined, the conservative forces acting on particles can then be derived,

$$\mathbf{F}_i = -\frac{\partial A}{\partial \mathbf{r}_i} = -\sum_j \frac{\partial f}{\partial n_j} \frac{\partial n(\mathbf{r}_j)}{\partial \mathbf{r}_i}, \quad (5.12)$$

Assuming the local density $n(\mathbf{r})$ is evaluated under the weighted density approximation, and we use a short-range pairwise weight function to interpolate the local densities $n(\mathbf{r}_i) = \sum_j w(r_{ij})$,

one finally arrives at,

$$\mathbf{F}_i = -\sum_j \left(\frac{\partial f}{\partial n(\mathbf{r}_i)} + \frac{\partial f}{\partial n(\mathbf{r}_j)} \right) \frac{\partial w(r_{ij})}{\partial r_{ij}} \mathbf{e}_{ij}, \quad (5.13)$$

which has the same form as the fluid mechanics result Eq. (5.9). It should be noted that this weighted density assumption of local free energy should be most suitable for describing fluids

under weak intermolecular interaction or at a fairly coarse spatial scale, in which the spatial correlation between particles contributes little to the total free energy.

Both the theoretical result of fluid mechanics and classical density functional theory indicates that the many-body conservative interaction is an essential piece of mesoscopic CG interactions. Therefore we would like to provide a strategy to derive such interactions through bottom-up coarse-graining method.

Considering the spatial resolution of the coarse-grained fluid we work on in this work, which on average represents 13.5-26.375 fine-grained fluid particles per CG bead, it might not be ideal to directly use the Eq. (5.13) form of interaction as a basis set to represent the CG interaction. This is because spatial correlations are still evident at this scale.³⁴ In addition, a pairwise additive basis set is included which gives out the well known interaction form in the many-body DPD models,^{10-11, 55}

$$\mathbf{F}_i = -\sum_j \left[\left(\frac{\partial f}{\partial n(\mathbf{r}_i)} + \frac{\partial f}{\partial n(\mathbf{r}_j)} \right) \frac{\partial w(r_{ij})}{\partial r_{ij}} + \frac{\partial u(r_{ij})}{\partial r_{ij}} \right] \mathbf{e}_{ij}. \quad (5.14)$$

The unknown components $\partial u(r) / \partial r$ and $\partial f(n) / \partial n$ are then determined through the MS-CG method. It should be noted the inference of this type of interaction has been previously studied using various methods.⁵⁵⁻⁵⁸ For actual numerical implementation in this work, we use basis functions to approximate these functions. For the local free energy function, we choose to expand using polynomial basis sets,

$$\frac{\partial f(n)}{\partial n} = \sum_k c_k n^k, \quad (5.15)$$

the reasoning behind this choice is that this term is tightly connected to the equation of states (EOS) of the system which is usually represented in terms of a polynomial series summation. While for the pairwise components, a set of B-Spline functions was utilized,

$$\frac{\partial u(r)}{\partial r} = \sum_k b_k B_k(r). \quad (5.16)$$

The basis function coefficients are then determined by minimizing the L-2 norm force residual functional,

$$\chi^2(\{b\}, \{c\}) = \frac{1}{3N} \left\langle \sum_i \|\mathbf{f}_i - \mathbf{F}_i(r, \{b\}, \{c\})\|^2 \right\rangle, \quad (5.17)$$

which is a standard least square regression problem and thus can be efficiently solved in the normal equation form.

In the dynamic mapping scheme, the instantaneous forces on each of the CG blobs cannot be simply calculated as the sum of fine-grained forces $\{\mathbf{f}_i(t)\}$ of all the underlying fine-grained fluid molecules, since the mapping itself is overlapping and highly nonlinear. A theoretically rigorous of force correspondence and in principle can be derived from the velocity mapping equation. However, the form would be too complicated to be implemented. Instead, we extract these forces numerically by differentiating the instantaneous velocities of the CG blobs,

$$\mathbf{f}_i(t) = m_i \frac{\mathbf{v}_i(t + \delta t) + \mathbf{v}_i(t - \delta t) - 2\mathbf{v}_i(t)}{2\delta t} + O(\delta t^2). \quad (5.18)$$

In subsequent sections, we will show that this numerical approximation does not introduce noticeable errors in the parameterization.

C. Statistical Inference of Momentum-Conserving Equations of Motion

The three force components in the exact CG equations of motions from Mori-Zwanzig formalism correspond to different aspects of physics. Parameterizing the conservative interaction is to find the mean forces exerted on CG variables whereas inferring the non-conservative interactions aims at recapitulating the fluctuation features of instantaneous forces. The bottom-up parameterization of the non-conservative components is often considered to be a harder problem than the inference of conservative interactions.

Specifically, for constructing equations of motions for mesoscopic CG fluid models, the difficulty mainly comes from two aspects. First, a major difficulty in reproducing mesoscopic CG dynamics comes from the fact that the dissipation and fluctuation forces are correlated through fluctuation dissipation theorem. One cannot simply separate the inference of these two components apart. In this sense, bottom-up CG methods such as MS-CG, Relative Entropy Minimization⁵⁹⁻⁶⁰ (REM) *etc* cannot migrate to be applied in solving this problem. Algorithms aware of this statistical correlation need to be developed.

The second challenge resides in finding the feasible form of equations of motion capable of describing CG dynamics at mesoscale. As briefly stated in Sec. I, the Mori-Zwanzig equations of motion need to be aggressively approximated for practical numerical implementation. While at the same time, all the approximations need to ensure the momentum conservation of the particle system so that it can be compatible with hydrodynamics theories. This imposes strong restraints on the form of friction kernels that one can choose to describe the coarse-grained dynamics. For example, the simple Generalized Langevin Equation or Langevin Equation is ruled out.

It is still not clear to what extent the approximation can be safely truncated. A couple of such approximated friction kernels have been utilized including the pairwise DPD form, and density

modulated pairwise SDPD/SPH form, to name a few. Among the feasible friction kernel forms, the most commonly adopted form is arguably the pairwise DPD form, especially in the bottom-up attempts to reconstruct mesoscopic CG dynamics. In this section, we choose to adopt this simplest yet common form to showcase our methodology. The bottom-up derivation of a momentum conserving DPD form CG equations of motion under both non-Markovian²¹⁻²³ and Markovian^{18, 20, 30} regimes will be presented. The propagation scheme of such EOMs will be built up to verify the performance of the inferred model against the reference dynamics results obtained from mapped atomistic trajectory.

By assuming a pairwise decomposition of friction kernel, the exact Mori-Zwanzig integro-differential equation can be simplified as,²⁰⁻²¹

$$\dot{\mathbf{P}}_I = \frac{1}{\beta} \frac{\partial}{\partial \mathbf{R}_I} \ln \omega(\mathbf{R}) - \sum_{J=1}^N \left(\int_0^t ds \mathbf{K}(\mathbf{R}_{IJ}, t-s) \mathbf{V}_{IJ}(s) + \delta \mathbf{F}_{IJ}^Q(t) \right), \quad (5.19)$$

in which the pairwise friction kernel $\mathbf{K}(\mathbf{R}, t) = \beta \langle \delta \mathbf{F}_{IJ}^Q(t) \otimes \delta \mathbf{F}_{IJ}^Q(0) \rangle_{\mathbf{R}}$, relating the dissipation with orthogonal dynamics. However, the instantaneous orthogonal forces $\delta \mathbf{F}_{IJ}^Q(t) = e^{-Q_i L t} \delta \mathbf{F}_{IJ}(0)$, in which Q represents the orthogonal projection operator, cannot be directly extracted from a CG trajectory mapped from fine-grained dynamics.

There are mainly two existing approximations to infer this friction kernel object from mapped trajectory. The first approach adopts the zeroth-order approximation of orthogonal propagator, e.g. $e^{-Q_i L t} \approx e^{-i L t}$, so that the friction kernel can be estimated using the instantaneous non-conservative components of the pairwise force correlations $\mathbf{K}(\mathbf{R}, t) = \beta \langle \delta \mathbf{F}_{IJ}(t) \otimes \delta \mathbf{F}_{IJ}(0) \rangle_{\mathbf{R}}$.⁵³ The $\delta \mathbf{F}_{IJ}$ can be calculated once the conservative pairwise interaction is estimated, i.e. $\delta \mathbf{F}_{IJ} = \mathbf{F}_{IJ} - \langle \mathbf{F}_{IJ} \rangle$. Another approach is based on the deconvolution

of the dissipative interactions. Notice that the following equation holds since the orthogonal dynamical process is Martingale,

$$\langle \delta \mathbf{F}_{IJ}(t) \mathbf{V}_{IJ}(0) \rangle = \int_0^t ds \mathbf{K}(R_{IJ}, t-s) \langle \mathbf{V}_{IJ}(s) \mathbf{V}_{IJ}(0) \rangle, \quad (5.20)$$

By inverting this convolution equation, one can obtain the pairwise friction kernel $\mathbf{K}(\mathbf{R}, t)$.

The above two approaches has been used in the bottom-up construction of DPD models for mesoscopic bonded polymer models and fluids under locality constraints, and showed good performance in reproducing relevant dynamical properties. However, neither of these two approaches can be used in more general cases due to two strong assumptions implied in the methods. First, these two methods assume that the underlying fine-grained system interacts only through pairwise interactions, since the pairwise CG forces $\mathbf{F}_{IJ}$ are evaluated as the sum of fine-grained pairwise forces between the two CG blobs, as illustrated in Fig. 5-2. Therefore, for system interacts through any many-body interactions, the applicability of these two methods vanishes. Second, extracting pairwise CG forces $\mathbf{F}_{IJ}$ assumes that the CG mapping operator is linear and time-invariant. All the previous work using these methods was built upon the center of mass mapping scheme, which is ill conditioned when applying to coarse-graining multiple non-bonded fluid molecules as one CG bead. In the current work, the dynamic mapping scheme could not generate a reasonable pairwise decomposition of instantaneous CG forces. Methods not relying on these assumptions need to be developed for the current application.

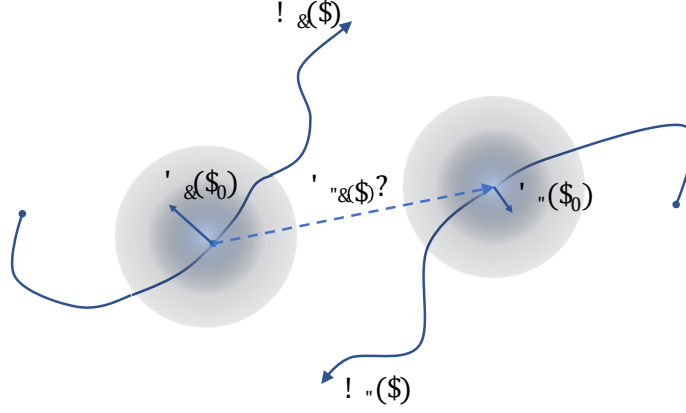


Figure 5-2: illustration of the dynamical information one can obtain from the CG trajectory obtained through dynamic CG mapping approach. The pairwise decomposition of instantaneous forces cannot be realized in this scenario.

Since the dynamic mapping scheme can provide a continuous CG trajectory with instantaneous position, velocity and force information. To maximally utilize all these information, we propose an inference strategy based on the deconvolution idea. Notice that by multiplying $\mathbf{V}_I(0)$ on both sides of the non-Markovian DPD EOM Eq. (5.19) and take ensemble averages,

$$\langle \delta \mathbf{F}_I(t) \cdot \mathbf{V}_I(0) \rangle = - \sum_{j=1}^N \int_0^t ds \langle \mathbf{K}(R_{IJ}, t-s) \mathbf{V}_{IJ}(s) \cdot \mathbf{V}_I(0) \rangle, \quad (5.21)$$

we further assume that the temporal and spatial dependency of the friction kernel is separable, which was proven to be a reasonable approximation in previous applications to bonded polymer systems.²¹⁻²³ In this context, the friction kernel is then written as,

$$K(R, t) = \gamma(t) \Omega(R), \quad (5.22)$$

where $\Omega(R)$ denotes the spatial component of friction kernel whereas $\gamma(t)$ denotes the temporal part. With these approximations, the above equation renders as a Volterra equation of first kind,

$$\left\langle \delta \mathbf{F}_I(t) \cdot \mathbf{V}_I(0) \right\rangle = - \int_0^t ds \gamma(t-s) \left\langle \tilde{\mathbf{V}}_I(s) \cdot \mathbf{V}_I(0) \right\rangle, \quad (5.23)$$

in which $\tilde{\mathbf{V}}_I(t) = \sum_J \Omega(R_{IJ}) \mathbf{V}_{IJ}(t)$ represents a weighted spatial dissipation. In general cases, the spatial friction $\Omega(R)$ is a 3 by 3 tensor. In this work, we assume the spatial component takes the same form as the radial smooth kernel function used in Sec. II. A. Which indicates that the friction $\Omega(R) = \omega(R) \hat{\mathbf{e}}_R$ aligns with the pair unit vector. It should be noted that even though this assumption on spatial friction is strong, it has been shown and verified in previous bottom-up coarse-graining of star polymer systems. And it is commonly applied in top-down mesoscopic models and showed many successes in reproducing many dynamical processes at mesoscale. There is certainly possibility that this spatial component friction kernel can be directly derived based on the trajectory information, which will be investigated in our following up works.

And then the temporal part is solved by numerically inverting Eq. (5.23). To ensure the numerical stability, the Volterra equation of first kind is transformed into the second kind by taking time derivatives on both sides of Eq. (5.23),⁴⁹

$$\dot{C}_{\delta FV}(t) = -C_{\tilde{V}V}(0)\gamma(t) - \int_0^t ds C_{\tilde{V}V}(t-s)\gamma(s), \quad (5.24)$$

in which the correlation functions are noted as $C_{\delta FV}(t) = \left\langle \delta \mathbf{F}_I(t) \cdot \mathbf{V}_I(0) \right\rangle$ and $C_{\tilde{V}V}(t) = \left\langle \tilde{\mathbf{V}}_I(t) \cdot \mathbf{V}_I(0) \right\rangle$. The derivatives of these correlation functions can be written as,

$$\begin{aligned} \dot{C}_{\delta FV}(t) &= - \left\langle \delta \mathbf{F}_I(t) \cdot \mathbf{F}_I(0) \right\rangle / M_I \\ C_{\tilde{V}V}(t) &= - \left\langle \tilde{\mathbf{V}}_I(t) \cdot \mathbf{F}_I(0) \right\rangle / M_I \end{aligned} \quad (5.25)$$

which can be directly estimated from the mapped CG trajectory. For numerical implementation, we assume the temporal friction kernel decays to zero within a certain time interval $N\delta t$, where

is the minimum time step of fine-grained sampling. The validity of this finite range assumption is further demonstrated in the numerical examples. With this setup, Eq. (5.24) can be discretized as,

$$\dot{C}_{\delta FV}(i\delta t) = -C_{\tilde{V}V}(0)\gamma(i\delta t) - \delta t \sum_{j=0}^i \eta_j \dot{C}_{\tilde{V}V}[(i-j)\delta t] \gamma(j\delta t), \quad (5.26)$$

in which $\eta_j = 1/2$ if $j=0$ or i , and $\eta_j = 1$ for other cases. Then the temporal friction can be solved iteratively by the following iteration scheme,

$$\gamma(i\delta t) = \frac{\dot{C}_{\delta FV}(i\delta t) + \delta t \sum_{j=0}^{i-1} \eta_j \dot{C}_{\tilde{V}V}[(i-j)\delta t] \gamma(j\delta t)}{C_{\tilde{V}V}(0) + \dot{C}_{\tilde{V}V}(0)\delta t / 2}, \quad (5.27)$$

with the initial value $\gamma(0) = \dot{C}_{\delta FV}(0) / C_{\tilde{V}V}(0)$.

After obtaining the pairwise friction kernel, all the piece of the mesoscopic CG equations of motion is in place. To test the performance of the above statistical inference framework, a propagation scheme of the CG EOM need to be worked out. First, the non-Markovian DPD type EOM we parameterized above can be discretized as,

$$\frac{d\mathbf{P}_I}{dt} = \sum_{J \neq I} \left\{ F_{IJ}^C \mathbf{e}_{IJ} - \sum_{n=0}^N \eta_n \delta t w[R_{IJ}(t - n\delta t)] \gamma(n\delta t) (\mathbf{e}_{IJ} \cdot \mathbf{V}_{IJ}) \mathbf{e}_{IJ} + \sum_{n=0}^N \alpha_{IJ,n} W_{IJ}(t - n\delta t) \mathbf{e}_{IJ} \right\}, \quad (5.28)$$

where the conservative force is defined in Eq. (5.14), the trapezoidal integration coefficients $\eta_n = 1/2$ if $n=0$ or N , and $\eta_n = 1$ for other cases. The random forces are generated through a linear superposition of Brownian motions. $W_{IJ}(t)$ denotes a time series of Gaussian random number with zero mean and covariance defined as follows,

$$\langle W_{IJ}(k\delta t) W_{MN}(l\delta t) \rangle = \delta_{kl} (\delta_{IM} \delta_{JN} + \delta_{IN} \delta_{JM}), \quad (5.29)$$

which generate the pairwise random force Martingale. The set of colored noise weighting coefficients $\{\alpha\}$ are adjusted to ensure the fluctuation dissipation theorem, which requires that,

$$\begin{aligned}
\langle \delta \mathbf{F}_{IJ}^Q(l\delta t) \cdot \delta \mathbf{F}_{IJ}^Q(0) \rangle &= \sum_{p=0}^N \sum_{s=0}^N \alpha_{IJ,s} \alpha_{IJ,p} \langle W_{IJ}(-p\delta t) W_{IJ}[(l-s)\delta t] \rangle \\
&= \sum_{p=0}^N \sum_{s=0}^N \alpha_{IJ,s} \alpha_{IJ,p} \delta_{l-s+p} = \sum_{s=0}^{N-l} \alpha_{IJ,s} \alpha_{IJ,s+l} = k_B T \mathbf{K}_{IJ}(R_{IJ}, l\delta t)
\end{aligned} \tag{5.30}$$

Due to the fact that the spatial and temporal dependency of friction kernel is separable, one can further simplify the random noise generator. Assuming the radial distances R_{IJ} do not change significantly within the time interval that friction kernel do not vanish, which usually holds in mesoscopic physical processes since the diffusion at mesoscale is often suppressed, one can separate the spatial and temporal dependency on the random noise generator as well. In that sense, one define these pairwise colored noise generator by,

$$\alpha_{IJ,s} = \sqrt{k_B T w(R_{IJ})} \alpha_s, \tag{5.31}$$

then it is evident that if one can obtain a set of coefficients $\{\alpha\}$ not dependent on any spatial information that satisfy,

$$\sum_{s=0}^{N-l} \alpha_s \alpha_{s+l} = \gamma(l\delta t), \tag{5.32}$$

the Fluctation Dissipation Theorem Eq. (5.30) would be naturally satisfied. To ensure the condition Eq. (5.32), the Nelder-Mead simplex optimization algorithm was applied to find the coefficients $\{\alpha\}$. With the friction kernel as well as colored noise generator in place, the non-Markovian DPD equation can be propagated through the standard Velocity Verlet algorithm⁶¹ with constant system temperature.

It should be noted that a new paradigm of propagating the non-Markovian equations of motion using auxiliary variables^{22, 52, 62} instead of explicitly evaluating the friction kernels has been recently developed and applied in the mesoscopic coarse-grained models of polymer melts. This approach can be naturally migrated in propagating the equations of motions derived in this

work. In the future, we would like to explore the applicability of this new integration paradigm in improving the numerical stability and further reducing the computational cost of the CG dynamics propagation.

Based on the non-Markovian DPD equation, we also construct a Markovian limit of the equation, which result in an EOM with exactly the same form as many-body DPD models. The Markovian limit of the friction kernel can be directly obtained by integrating the non-Markovian friction kernel with respect to time,

$$\varpi(R) = \int_0^\infty \mathbf{K}(R, t) dt. \quad (5.33)$$

Under the separation assumption of the non-Markovian friction kernel, this can further be simplified as $\varpi(R) = \gamma_0 w(R)$ where $\gamma_0 = \int_0^\infty \gamma(t) dt$.

The Markovian DPD equations of motions then can be propagated based on the following discretized form,

$$\frac{d\mathbf{P}_I}{dt} = \sum_{J \neq I} \left\{ F_{IJ}^C \mathbf{e}_{IJ} - \varpi(R_{IJ}) (\mathbf{e}_{IJ} \cdot \mathbf{V}_{IJ}) \mathbf{e}_{IJ} + \sqrt{\frac{2\varpi(R_{IJ})}{\beta \delta t}} W_{IJ}(t) \mathbf{e}_{IJ} \right\}, \quad (5.34)$$

in which $W_{IJ}(t)$ is a discretized time series of Gaussian random number with zero mean and variance 1. Similarly, the EOM is propagated through Velocity Verlet scheme.

III. Results and Discussions

A. Fine-grained System and Mapping Details

In this section, we apply the methods developed above to construct a mesoscopic Many-Body DPD like model for supercritical Lennard-Jones fluid systems from bottom-up. The atomistic MD simulations of the quiescent LJ fluids were performed in a cubic simulation box subjected to periodic boundary conditions (PBC) in all three dimensions. The LJ particles are interacting

through the standard 12-6 type LJ potential $u(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$, which was cut off at 2.5σ for actual simulation. The reduced LJ units were employed throughout the analysis in this work (the length, mass, energy and time units being $\sigma = 1$, $m = 1$, $\epsilon = 1$ and $\tau = \sigma(m/\epsilon)^{1/2}$ respectively). A systems with 1688 LJ particles were simulated. The temperature $T^* = 2.00$ and density $\rho^* = 0.8442$ were chosen to ensure the supercritical fluid phase, the simulation box lengths is 12.5970σ . The Nose-Hoover thermostat with a damping factor of 0.50τ was coupled to the system to maintain the temperature. The systems were equilibrated for 10^5 time steps to reach equilibrium and the production runs were then propagated in NVT ensemble maintained by Nose-Hoover thermostat⁶³⁻⁶⁴ for 10^6 time steps with the integration time step $\delta t = 0.005\tau$. The open-source Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package⁶⁵ was used to propagate the trajectory.

After the atomistic trajectory was obtained, the system was then mapped into mesoscopic representation through the method introduced in Sec. II A. CG models with two resolutions were constructed, in which the number ratios between fine-grained and CG particles n_{FG} / n_{CG} are 13.504 and 26.375 respectively, corresponding to the density of CG particles to be $\rho_{CG} = 0.0625$ and $\rho_{CG} = 0.0320$. The parameters in the weight function $w(r)$ are determined as follows. The value of r_c characterizes the radius of the coarse-grained beads, which was chosen to be $r_c = L / \sqrt[3]{4\pi N / 3}$ in this work, L denotes the unit box length and N represents total number of CG blobs. Its value was determined by the assumption that the spherical CG blobs would take over all the space of the simulation box. The parameters $\sigma = 0.90$, $\beta = 1.40$ and $\gamma = 0.30$ were utilized throughout the calculations, to achieve stable propagation. The evaluation of the

weighting function was truncated at $2r_c$ considering that the magnitude of the function is small enough to be ignored at greater distances. And lastly, as mentioned earlier, a self-consistent re-mapping to ensure the position consistency was applied every 20,000 steps of velocity mapping propagation. The mapping was performed using an in-house code, which is freely available upon request.

B. Conservative Interactions

Once the fine-grained trajectory is mapped to a mesoscopic coarse-grained representation, a continuous time series of particle positions, velocities as well as accelerations are obtained. In this section, the acceleration information is utilized to infer the conservative potential of mean force of the CG blobs through the MS-CG based algorithm elaborated in Sec. II. B. To derive the many-body conservative interaction terms, we first choose the Lucy's weight function to evaluate the many-body local densities, which takes the following form.

$$w(r < h) = \left(1 + \frac{3r}{h}\right) \left(1 - \frac{r}{h}\right)^3, \quad (5.35)$$

where h represents the cutoff for evaluating the Lucy's function. There are certainly many possible choices of the weight function, as long as it is monotonically and smoothly decaying to zero within short range. Among those options, the Lucy's weight function has been a common choice in this scenario since its first use in SPH methods for astrophysical simulations, since its low cost of computation and both its value and derivative converge to zero at cutoff.³ As one can imagine, this artificially chosen weight function would possibly introduce error into the model in representing the exact potential of mean forces. In the results shown later in the section, we will show that the artifacts introduced would be largely compensated by the pairwise additive term.

The cutoff distances for the two CG resolutions are 4.0σ and 6.0σ respectively, ensuring that the means of local densities are comparable as indicated in Fig. 5-3(c) and 5-3(f).

After evaluating the local densities, the many-body terms relating to equation of states $\partial u(\rho) / \partial \rho$ is then parameterized through Force Matching. In this work, we truncate the Taylor expansion to the first order, assuming that $\partial u(\rho) / \partial \rho = \lambda_0 \rho$. The higher order expansions are truncated because the local density fluctuation in the current homogeneous fluids is relatively small and not far from the equilibrium density. Introducing higher order terms might lead to numerical errors due to the poor sampling for particles at low and high local densities. And this simple linear form is well adopted in the application of top-down many-body DPD models. A more thorough study on the higher order expansions would be conducted in the near future for mesoscopic coarse-grained interfacial systems in which the local density profile shows broad distribution. The pairwise additive interaction was expanded in a set of 4th order B-Spline basis functions, in which 100 basis functions was used spanning the radial distance between 1.50σ to 6.30σ .

In the present study, the many-body terms and pairwise additive terms are parameterized sequentially that the $\partial u(\rho) / \partial \rho$ term was parameterized first assuming the absent of pairwise additive interactions. After obtaining the many-body interactions, it was deducted from the instantaneous forces on each CG blob. Finally, the pairwise additive term was derived by minimizing the L2 norm between the residual forces and pairwise forces. This hierarchical approach eliminated the multicollinearity issue while parameterizing both terms simultaneously, since the pairwise additive term and many-body term shares a common pairwise feature.

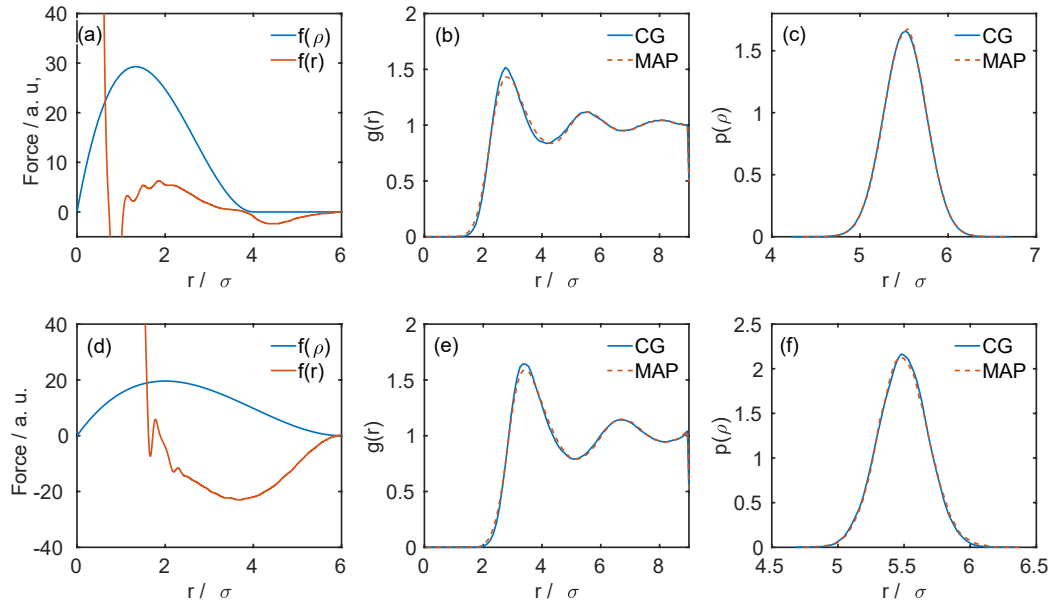


Figure 5-3: the many-body DPD type interaction inferred based on the method developed in this work, as well as their performance in reproducing mesoscopic structure of the underlying fluid systems. Panels (a) to (c) correspond to the higher resolution case where the CG blob density $\rho_{CG} = 0.0625$, whereas panels (d) to (f) represents the case of $\rho_{CG} = 0.0320$. The many-body modulated interactions, which originate from the density weight function, are represented by the red lines in panels (a) and (d). The blue lines in these two panels denote the pairwise additive part. Panels (b) and (e) present the comparison of radial distribution functions between the CG models and reference mapped results. Panels (c) and (f) showcase the comparison between local density distributions.

Fig. 5-3 shows the results of the inferred many-body interaction as well as its capability of recapitulating the pairwise structure and local density fluctuation of the mesoscopic CG blobs. Fig. 5-3(a) and 5-3(d) show the inferred many-body interaction as well as the pairwise additive interactions. The CG blobs are interacting through a repulsive many-body interaction coupled

with an attractive pairwise force with different interaction ranges in both cases. This feature perfectly aligns with the form utilized in many-body DPD models, which has been shown to be essential in forming interfaces and phase transitions. It should be noted that in previous studies the same CG system was parameterized using only pairwise additive interactions, which resulted in purely repulsive pairwise forces.

The local density based many-body interactions help to improve the accuracy of representing CG PMFs in these cases. The CG systems were propagated using these inferred many-body interactions and the structural properties of the generated CG dynamics were compared against the result obtained from the mapped trajectory. The radial distribution functions shown in Fig. 5-3(b) and Fig. 5-3(e) are quantitatively reproduced except for a slight error near the first peak, which might come from the artifact of Lucy's weight function. The particle local density distributions shown in Fig. 5-3(d) and 5-3(f), which render a mean field representation of the many-body correlations, are also perfectly recapitulated. These encouraging results suggest that the many-body DPD type interaction is a good approximation for the conservative interactions in mesoscopic CG models of fluids.

C. Coarse-grained Dynamics

In this section, a set of momentum conserving many-body DPD form equations of motion will be constructed based on the two mapped mesoscopic CG trajectories, utilizing the method developed in Sec. II. The temporal part of the friction kernels was truncated up to 500 time steps, which is shown to be valid since the result decays to small values within this time interval. The spatial component of the friction kernels took the same form as the Lucy's weight function Eq. (5.35). The cutoff distances of the pairwise spatial friction kernels for the two CG resolutions are

4.8σ and 4.0σ respectively, as shown in Fig. 5-4(b). These cutoffs were chosen based on the scaling of the size of CG blobs. The temporal components were determined by solving Eq. (5.35), and the results are shown in Fig. 5-4(a). In both cases, the temporal component $\gamma(t)$ can be thought of a superposition of two exponential decays with different characteristic time. The corresponding colored noise generators for the two CG resolutions are shown in Fig. 5-4(c) and 5-4(d) respectively.

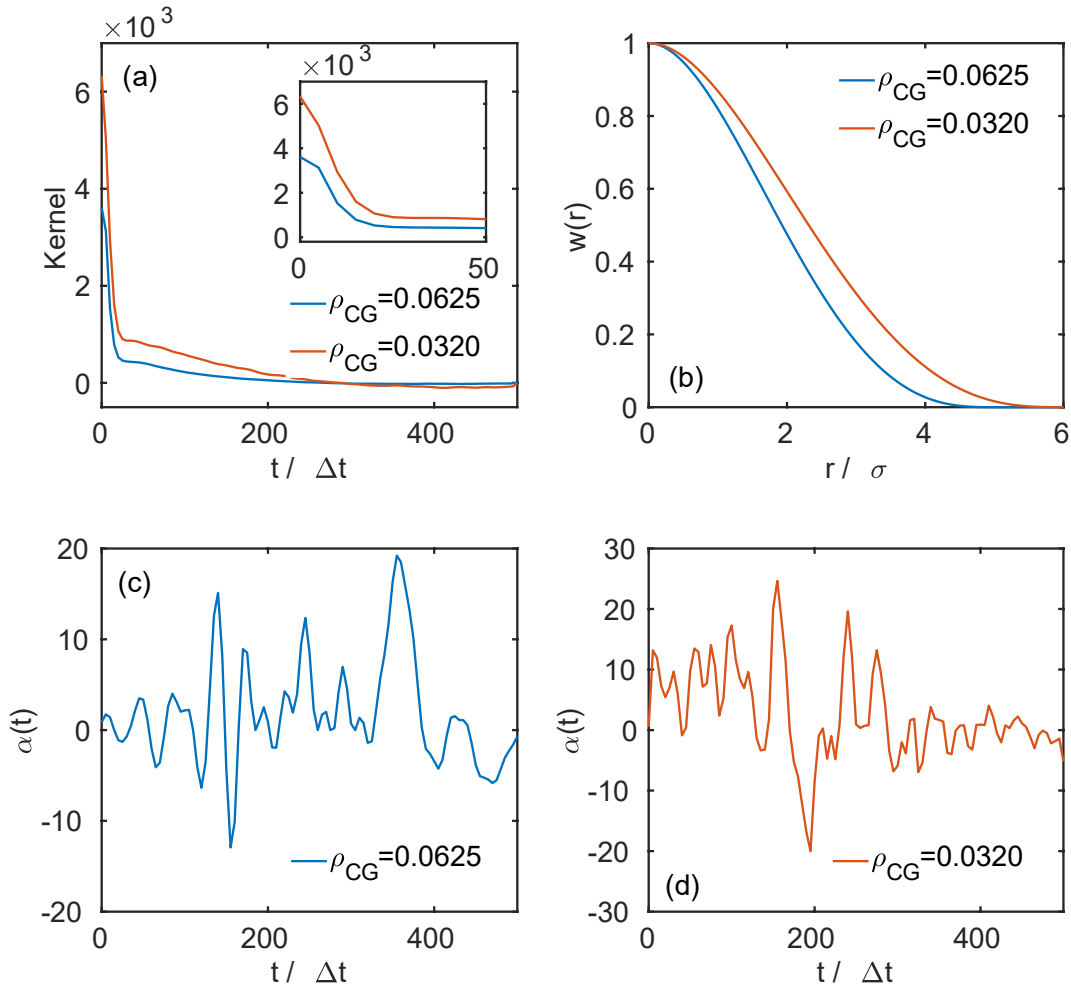


Figure 5-4: the inferred friction kernels as well as the corresponding colored noise generators for CG models at two resolutions. The temporal component of friction kernels is shown in panel (a)

Figure 5-4, continued: whereas the spatial part is presented in panel (b). Panels (c) and (d) showcase the colored noise generator for constructing fluctuation forces.

After obtaining the friction kernels and their corresponding random force generators, the discrete non-Markovian DPD equations of motion were propagated through Eq. (5.28). The integration time step in propagating the CG dynamics was chosen to be 0.025τ , which is 5 times longer than the fine-grained time step. Figure 5-5 shows the comparison of the velocity autocorrelation functions between the mapped CG trajectories and the propagated non-Markovian DPD trajectories for the two CG resolutions.

In both of the two cases, the short time and long time decaying feature of VACFs is quantitatively reproduced. The backscattering feature at median time range between 30 and 80 Δt can be quantitatively captured. However, the non-Markovian DPD equations underestimated the strength but overestimated the lifetime of the backscattering. This discrepancy mainly comes from two approximations applied in the inference algorithm. First, the spatial component of the friction kernel might take a different form than the Lucy's weight function. The current results suggest that the friction at short distance can be well approximated by the Lucy's weight function, whereas at medium distances a stronger friction might need to be employed to damp the backscattering. Second, the spatial and temporal component of the friction kernel might not be separable in this certain scenario. In principle, to estimate the pairwise friction kernel in a more systematic manner, one need to have an accurate estimation of instantaneous pairwise forces between CG variables which is hard to extract for most cases. In forthcoming works, we would like to further refine the approximations applied in the method to control these two possible sources of error, yet preserving the simplicity of the algorithm to the maximum extent.

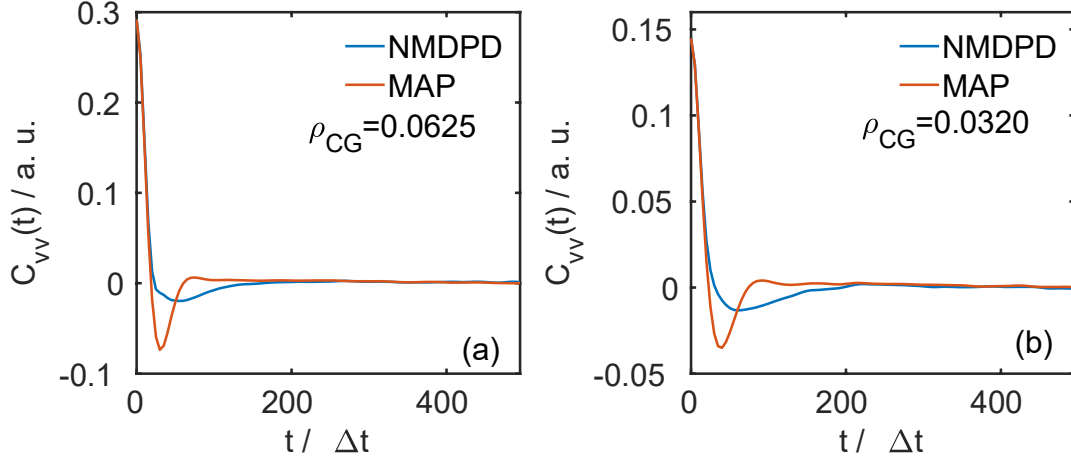


Figure 5-5: the comparison of VACFs estimated from mapped CG trajectory and the non-Markovian DPD dynamics results.

Next, we investigate the Markovian limit of the non-Markovian DPD model and test the applicability of Markovian equations of motion on reproducing the dynamics of mesoscopic CG models of fluids. The Markovian DPD friction kernel was derived based on the method stated in Sec. II. C. The temporal component of the non-Markovian DPD friction kernel was integrated to give an effective friction constant. Then the Markovian DPD dynamics were propagated through Eq. (5.34) with the integration time step to be 0.025τ . Figure 5-6 shows the comparison of the VACFs between the mapped CG trajectories and the Markovian DPD trajectories. As shown in the comparison, the Markovian limit of the DPD dynamics damps the velocity correlations more quickly than the reference result from CG mapping, indicating an overestimated dynamic viscosity in the parameterized model. And the backscattering feature, which is a highly non-Markovian dynamical character, cannot be even quantitatively reproduced through the Markovian DPD model. The results indicate that at the mesoscopic CG resolution under study, in

which around 20 fluid molecules are represented by one CG entity, the Markovian approximation is not a good description of the dynamics of fluids.

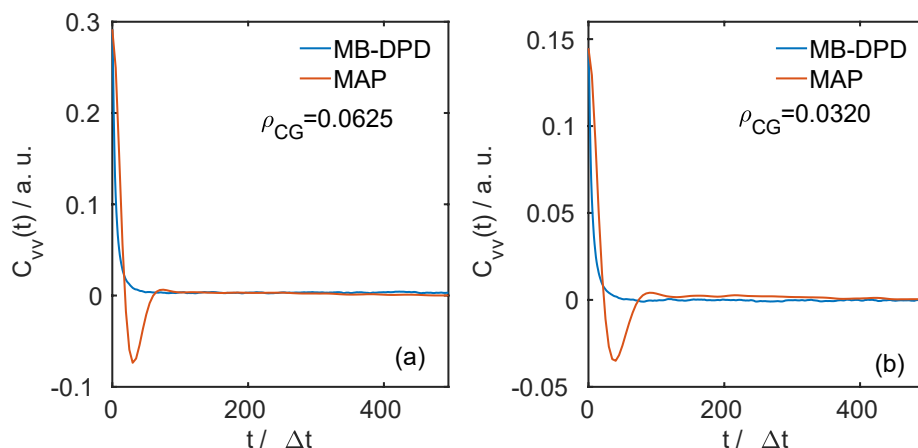


Figure 5-6: the comparison of VACFs estimated from mapped CG trajectory and the Markovian DPD dynamics results.

Except for the velocity autocorrelation functions, we compared the diffusion constant estimated based on the VACFs from these two approaches with the result obtained from CG mapping. The results are shown in Table 5-I. It is obvious that the non-Markovian DPD models showed a decent performance in reproducing the diffusion process of the CG blobs. The relative errors with respect to the reference results are all within 20%.

For the Markovian DPD models, the diffusion constants are less well reproduced compared to the non-Markovian models, showing relative errors around 40%. These results are still encouraging given the fact that the friction kernels were not directly fitted to match diffusion constants nor any other macroscopic dynamic properties. Even though the VACFs obtained from the Markovian DPD models are drastically different than the reference mapped result, the diffusion process, on the other hand, can be reasonably recapitulated.

TABLE 5-I. Diffusion constants estimated from the mapped CG trajectory and the parameterized Non-Markovian and Markovian DPD models.

Cases	<i>Mapped</i>	<i>Non-Markovian</i>	<i>Markovian</i>
$\rho_{CG} = 0.0625$	2.32	2.41 (+3.9%)	3.32 (+43.1%)
$\rho_{CG} = 0.0320$	1.44	1.08 (-18.1%)	0.84 (-41.7%)

IV. Concluding Remarks

In this work, we developed a bottom-up strategy for constructing mesoscopic coarse-grained models that have direct connections with physics at mesoscale. The famous Mori-Zwanzig formalism provides the most general form of equations of motions describing the dynamics of the reduced CG system. However, it is often not practical to directly adopt the full Mori-Zwanzig form equation to propagate the CG dynamics, due to the many-body nature of the potential of mean force and the non-Markovian friction kernel. One needs to make appropriate assumptions on these two quantities to simplify the equations of motion so that it can be used as practical simulation tools. These assumptions are based on correct understanding of physics at the spatial and temporal scale of the CG resolution under study. In this work, in order to represent mesoscopic physics in the CG dynamics model, we embed the theoretical results and requirements from fluid mechanics and classical density functional theory into the CG equations of motions. The resulting CG equations of motion incorporates a many-body conservative interaction and a pairwise decomposable non-Markovian friction kernel, which has a compatible form with top-down mesoscopic fluid models derived from hydrodynamics such as many-body DPD and SDPD.

This specific form of equations of motion, at the same time, raises new challenges and restrictions in the bottom-up statistical inference of its components. To overcome these obstacles,

we designed two algorithms targeting the inference of conservative and non-conservative interactions respectively merely using the time series information of mapped CG configurations and velocities. For the many-body conservative interactions, we extended the MS-CG method in this scenario to derive the local density dependent many-body interaction, which has a direct link with the equation of state of the fluid. And for the friction kernels, we extended a correlation function inversion process to derive the spatial and temporal dependency of the pairwise non-Markovian friction kernel. The corresponding propagation algorithm is introduced that propagate the CG dynamics with linear scaling and accommodate longer integration time step. The Markovian limit of the equations of motion and the associated propagation scheme are directly obtained through integrating the non-Markovian friction kernel. The method developed in this work builds up a bridge between the bottom-up CG models with mesoscopic physics of fluids, and renders a practical algorithm to derive the interactions in mesoscopic fluid models from bottom-up, without the need of *ad hoc* parameter tuning.

To test the efficacy of the method, we derived a set of many-body DPD form models describing the dynamics of mesoscopic CG models of molecular fluids under both Markovian and non-Markovian limits, in which the mesoscopic CG training trajectory was generated through the recently developed dynamic CG mapping algorithm. The non-Markovian many-body DPD model derived from bottom-up showed perfect performance in recapitulating the structural and dynamic properties of the CG system. To the best of our knowledge, the results presented in this work provide the first bottom-up evidence of the widely used many-body DPD models. The pairwise structure and local density distribution profiles, which represents a mean-field many-body correlation, can be perfectly reproduced, indicating a correct description of the free energy surface of the CG fluids. The short time and long time characters of the velocity autocorrelation

functions can be quantitatively recapitulated. The backscattering feature at medium lag time can be qualitatively captured, the characteristic time and the strength of this feature is nonetheless off from the reference result, which reflects an underestimation of frictions at middle range. The diffusion constants characterizing the long time dynamic processes of the CG fluids can be reasonably reproduced with less than 20% relative error. The Markovian many-body DPD model, on the other hand, gives poorer description of the dynamical correlation, especially at short time. The VACFs cannot be faithfully reproduced, suggesting the failure of Markovian limit in describing mesoscopic dynamics of molecular fluids. The diffusion processes are better described, though exhibiting a bigger error compared to the non-Markovian counterparts.

We should note that even though we have achieved preliminary success in using the method to construct physics-aware bottom-up CG models at mesoscale. There are several problems need to be addressed by us or others in future work. First, the models in this work were constructed based on homogeneous liquid systems, which limits the sampling of local densities. This fact restricts the interpolation of the inferred model to system with vast density fluctuations. To overcome this limitation, we would like to build mesoscopic CG models upon CG trajectories of interfacial systems in the near future, which rely on an extension of the current dynamic coarse-grained mapping approach to systems with non-uniform density distributions. Second, the current statistical inference algorithm relies on an assumption of the spatial component of the pairwise friction kernel, which could be the major source of error in reproducing the VACFs. A systematic way of determining this spatial dependency of friction kernel need to be worked out in the future to control this error. And lastly, a bottom-up CG methodology based on non-equilibrium sampling of the underlying fine-grained system need to be developed in the future so

that non-equilibrium transport behavior of the fluids can be faithfully integrated into the coarse-grained model.

Reference

1. Gingold, R. A.; Monaghan, J. J., Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Mon. Not. R. Astron. Soc.* **1977**, *181*, 375.
2. Monaghan, J. J., Smoothed Particle Hydrodynamics. *Annu. Rev. Astron. Astrophys.* **1992**, *30*, 543.
3. Monaghan, J. J., Smoothed particle hydrodynamics. *Rep. Prog. Phys.* **2005**, *68*, 1703.
4. Ellero, M.; Español, P., Everything you always wanted to know about SDPD* (*but were afraid to ask). *Appl. Math. Mech.* **2018**, *39*, 103.
5. Español, P.; Revenga, M., Smoothed dissipative particle dynamics. *Phys. Rev. E* **2003**, *67*, 026705.
6. Español, P., Hydrodynamics from dissipative particle dynamics. *Phys. Rev. E* **1995**, *52*, 1734.
7. Español, P.; Warren, P., Statistical Mechanics of Dissipative Particle Dynamics. *Europhys. Lett.* **1995**, *30*, 191.
8. Groot, R. D.; Warren, P. B., Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **1997**, *107*, 4423.
9. Español, P.; Warren, P. B., Perspective: Dissipative particle dynamics. *J. Chem. Phys.* **2017**, *146*, 150901.
10. Pagonabarraga, I.; Frenkel, D., Dissipative particle dynamics for interacting systems. *J. Chem. Phys.* **2001**, *115*, 5015.
11. Warren, P. B., Vapor-liquid coexistence in many-body dissipative particle dynamics. *Phys. Rev. E* **2003**, *68*, 066702.
12. Arienti, M.; Pan, W.; Li, X.; Karniadakis, G., Many-body dissipative particle dynamics simulation of liquid/vapor and liquid/solid interactions. *J. Chem. Phys.* **2011**, *134*, 204114.
13. Español, P.; Serrano, M.; Zuñiga, I., Coarse-Graining of a Fluid and its Relation with Dissipative Particle Dynamics and Smoothed Particle Dynamic. *Intl. J. Mod. Phys. C* **1997**, *08*, 899.

14. Zwanzig, R., Ensemble Method in the Theory of Irreversibility. *J. Chem. Phys.* **1960**, *33*, 1338.
15. Zwanzig, R., Memory Effects in Irreversible Thermodynamics. *Phys. Rev.* **1961**, *124*, 983.
16. Zwanzig, R., On the identity of three generalized master equations. *Physica* **1964**, *30*, 1109.
17. Mori, H., Transport, Collective Motion, and Brownian Motion*). *Prog. Theor. Phys.* **1965**, *33*, 423.
18. Hijón, C.; Español, P.; Vanden-Eijnden, E.; Delgado-Buscalioni, R., Mori–Zwanzig formalism as a practical computational tool. *Faraday Discuss.* **2010**, *144*, 301.
19. Lei, H.; Caswell, B.; Karniadakis, G. E., Direct construction of mesoscopic models from microscopic simulations. *Phys. Rev. E* **2010**, *81*, 026704.
20. Li, Z.; Bian, X.; Caswell, B.; Karniadakis, G. E., Construction of dissipative particle dynamics models for complex fluids via the Mori–Zwanzig formulation. *Soft Matter* **2014**, *10*, 8659.
21. Li, Z.; Bian, X.; Li, X.; Karniadakis, G. E., Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. *J. Chem. Phys.* **2015**, *143*, 243128.
22. Li, Z.; Lee, H. S.; Darve, E.; Karniadakis, G. E., Computing the non-Markovian coarse-grained interactions derived from the Mori–Zwanzig formalism in molecular systems: Application to polymer melts. *J. Chem. Phys.* **2017**, *146*, 014104.
23. Yoshimoto, Y.; Kinefuchi, I.; Mima, T.; Fukushima, A.; Tokumasu, T.; Takagi, S., Bottom-up construction of interaction models of non-Markovian dissipative particle dynamics. *Phys. Rev. E* **2013**, *88*, 043305.
24. Kinjo, T.; Hyodo, S.-a., Equation of motion for coarse-grained simulation based on microscopic description. *Phys. Rev. E* **2007**, *75*, 051109.
25. Flekkøy, E. G.; Coveney, P. V., From Molecular Dynamics to Dissipative Particle Dynamics. *Phys. Rev. Lett.* **1999**, *83*, 1775.
26. Flekkøy, E. G.; Coveney, P. V.; De Fabritiis, G., Foundations of dissipative particle dynamics. *Phys. Rev. E* **2000**, *62*, 2140.
27. Eriksson, A.; Jacobi, M. N.; Nyström, J.; Tunström, K., Using force covariance to derive effective stochastic interactions in dissipative particle dynamics. *Phys. Rev. E* **2008**, *77*, 016707.

28. Eriksson, A.; Nilsson Jacobi, M.; Nyström, J.; Tunström, K., On the microscopic foundation of dissipative particle dynamics. *Europhys. Lett.* **2009**, *86*, 44001.
29. Hadley, K. R.; McCabe, C., On the Investigation of Coarse-Grained Models for Water: Balancing Computational Efficiency and the Retention of Structural Properties. *J. Phys. Chem. B* **2010**, *114*, 4590.
30. Izvekov, S.; Rice, B. M., Multi-scale coarse-graining of non-conservative interactions in molecular liquids. *J. Chem. Phys.* **2014**, *140*, 104104.
31. Zavadlav, J.; Marrink, S. J.; Praprotnik, M., Adaptive Resolution Simulation of Supramolecular Water: The Concurrent Making, Breaking, and Remaking of Water Bundles. *J. Chem. Theory Comput.* **2016**, *12*, 4138-4145.
32. Zavadlav, J.; Praprotnik, M., Adaptive resolution simulations coupling atomistic water to dissipative particle dynamics. *J. Chem. Phys.* **2017**, *147*, 114110.
33. Zavadlav, J.; Marrink Siewert, J.; Praprotnik, M., SWINGER: a clustering algorithm for concurrent coupling of atomistic and supramolecular liquids. *Interface Focus* **2019**, *9*, 20180075.
34. Han, Y.; Dama, J. F.; Voth, G. A., Mesoscopic coarse-grained representations of fluids rigorously derived from atomistic models. *J. Chem. Phys.* **2018**, *149*, 044104.
35. Lynn, H.; Thachuk, M., Equations of motion for position-dependent coarse-grain mappings obtained with Mori-Zwanzig theory. *J. Chem. Phys.* **2019**, *150*, 024108.
36. Jin, J.; Han, Y.; Voth, G. A., Ultra-Coarse-Grained Liquid State Models with Implicit Hydrogen Bonding. *J. Chem. Theory Comput.* **2018**, *14*, 6159.
37. Jin, J.; Voth, G. A., Ultra-Coarse-Grained Models Allow for an Accurate and Transferable Treatment of Interfacial Systems. *J. Chem. Theory Comput.* **2018**, *14*, 2180.
38. Jin, J.; Han, Y.; Voth, G. A., Coarse-graining involving virtual sites: Centers of symmetry coarse-graining. *J. Chem. Phys.* **2019**, *150*, 154103.
39. Rosenfeld, Y., Free-energy model for the inhomogeneous hard-sphere fluid mixture and density-functional theory of freezing. *Phys. Rev. Lett.* **1989**, *63*, 980.
40. Kierlik, E.; Rosinberg, M. L., Free-energy density functional for the inhomogeneous hard-sphere fluid: Application to interfacial adsorption. *Phys. Rev. A* **1990**, *42* (6), 3382-3387.
41. Lutsko, J. F., Density functional theory of inhomogeneous liquids. I. The liquid-vapor interface in Lennard-Jones fluids. *J. Chem. Phys.* **2007**, *127*, 054701.
42. Schweizer, K. S., Microscopic theory of the dynamics of polymeric liquids: General formulation of a mode-mode-coupling approach. *J. Chem. Phys.* **1989**, *91*, 5802.

43. Izvekov, S.; Voth, G. A., Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.* **2005**, *123*, 134105.
44. Izvekov, S.; Voth, G. A., A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469.
45. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A., Multiscale Coarse-Graining and Structural Correlations: Connections to Liquid-State Theory. *J. Phys. Chem. B* **2007**, *111*, 4116.
46. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C., The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.
47. Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A., The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *J. Chem. Phys.* **2008**, *128*, 244115.
48. Darve, E.; Solomon, J.; Kia, A., Computing generalized Langevin equations and generalized Fokker–Planck equations. *Proc. Natl. Acad. Sci.* **2009**, *106*, 10884.
49. Shin, H. K.; Kim, C.; Talkner, P.; Lee, E. K., Brownian motion from molecular dynamics. *Chem. Phys.* **2010**, *375*, 316.
50. Jung, G.; Hanke, M.; Schmid, F., Iterative Reconstruction of Memory Kernels. *J. Chem. Theory Comput.* **2017**, *13*, 2481.
51. Davtyan, A.; Dama, J. F.; Voth, G. A.; Andersen, H. C., Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. *J. Chem. Phys.* **2015**, *142*, 154104.
52. Davtyan, A.; Voth, G. A.; Andersen, H. C., Dynamic force matching: Construction of dynamic coarse-grained models with realistic short time dynamics and accurate long time dynamics. *J. Chem. Phys.* **2016**, *145*, 224107.
53. Izvekov, S., Microscopic derivation of particle-based coarse-grained dynamics: Exact expression for memory function. *J. Chem. Phys.* **2017**, *146*, 124109.
54. Du, Q.; Faber, V.; Gunzburger, M., Centroidal Voronoi Tessellations: Applications and Algorithms. *SIAM Review* **1999**, *41*, 637-676.
55. Wagner, J. W.; Dannenhoffer-Lafage, T.; Jin, J.; Voth, G. A., Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions. *J. Chem. Phys.* **2017**, *147*, 044113.

56. Sanyal, T.; Shell, M. S., Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation. *J. Chem. Phys.* **2016**, *145*, 034109.
57. DeLyser, M. R.; Noid, W. G., Extending pressure-matching to inhomogeneous systems via local-density potentials. *J. Chem. Phys.* **2017**, *147*, 134111.
58. Sanyal, T.; Shell, M. S., Transferable Coarse-Grained Models of Liquid–Liquid Equilibrium Using Local Density Potentials Optimized with the Relative Entropy. *J. Phys. Chem. B* **2018**, *122*, 5678.
59. Shell, M. S., The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
60. Chaimovich, A.; Shell, M. S., Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
61. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98.
62. Ceriotti, M.; Bussi, G.; Parrinello, M., Colored-Noise Thermostats à la Carte. *J. Chem. Theory Comput.* **2010**, *6*, 1170.
63. Nosé, S., A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511.
64. Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695.
65. Plimpton, S., Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1.

Chapter 6: Field Theoretic Representation of Molecular Liquids from Coarse-grained Models

Abstract

In this work, we proposed a systematic strategy to build up statistical field theory of molecular liquids based on a hierarchical approach, in which the center of mass level coarse-grained models of the molecular liquid serves as an intermediate step in the transformation from atomistic Hamiltonian to field theory. By introducing the coarse-grained model as an intermediate step, the difficulty of the atomistic to field theoretic transformation can be greatly reduced but still preserves important microscopic structural correlations of the underlying fluid. This approach eliminates *ad hoc* assumptions applied in many field theoretic models of molecular systems and for the first time bridges the gap between the discrete particle based and field theoretic based modeling techniques. The field theoretic representation of both canonical partition function and grand canonical partition function is derived, utilizing the Hubbard Stratonovich transformation. The effective action functional for both cases is complex and thus leads to the numerical “sign” problem for statistical sampling. To ameliorate the “sign” problem, the Gaussian Equivalent Representation of the field theory is derived, in which the integral contour is shifted to complex plane and the integral measure is changed so that the lower order terms of the field interaction can be eliminated. Finally, the statistical sampling approach and statistical estimators for evaluating the radial distribution functions are elaborated.

I. Introduction

In the previous chapters, the coarse-graining strategies at either molecular scale or mesoscale are all based on a particle representation of the system. This seems to be a natural choice for models with relatively high resolution where the coarse-grained entities correspond to a group of spatially coherent fine-grained entities. For mesoscopic coarse-grained models, such clear

correspondence picture gets blurred since a specific CG entity may represent hundreds or even much larger numbers of fine-grained degrees of freedom. The larger spatial scale smears out the local correlations, resulting in a more smoothly varying spatial representation of the molecular system. Under such circumstances, a particle-based theory might not be the only, or even the best choice for representing meso- and macroscopic physics. And this intuition is embedded in many phenomenological or top-down models at meso- and macro scale. For instance, the theory of fluid mechanics, or more broadly continuum mechanics, is entirely constructed based on field theory. The basic unit of building these theories represents a spatial chunk of underlying materials, and is mathematically treated as a set of field variables. In equilibrium statistical mechanics, this approach is also widely used in liquid state theory¹⁻¹¹ and especially polymer physics.^{6, 12-16}

In this work, we would like to explore a strategy to break the molecular-scale barrier currently facing bottom-up multiscale approaches that use the tools of field theory *to remove the need for any particle-level simulation*. For decades, the tools of field theory, in particular the many integral equation theories for the liquid state,^{4-5, 7-9, 11} have struggled to capture the complexity of homogeneous molecular liquids with irregular shapes and complex orientational correlations, much less the nanoscale heterogeneity observed in ionic liquids.¹⁷ The correspondence between the field theoretic representations of the molecular system with the atomistic description is often not apparent, therefore raises the difficulty of embedding delicate local chemical environment information into the field theory models. One way to define such an atomistic to field theory correspondence is through a Hubbard-Stratonovich (HS) transformation¹⁸ of the atomistic configurational partition function. Assuming a single-component molecular fluids interacting

through pairwise intermolecular interaction and the intramolecular interaction left for general, the partition function of such system is written as,^{12, 19-21}

$$\begin{aligned} Z &= \int d\mathbf{r}^n \exp \left(-\beta \sum_{i < j} u(r_{ij}) - \beta \sum_I U_I(\mathbf{r}^n) \right) \\ &= \int d\mathbf{r}^n \exp \left(-\frac{\beta}{2} \int d\mathbf{r} d\mathbf{r}' \rho(\mathbf{r}) u(|\mathbf{r} - \mathbf{r}'|) \rho(\mathbf{r}') - \beta \sum_I U_I(\mathbf{r}^n) \right), \end{aligned}$$

in terms of the local density field variable $\rho(\mathbf{r}) = \sum_i \delta(\mathbf{r} - \mathbf{r}_i)$. Applying HS transformation, this

can be transformed as,

$$\begin{aligned} Z &= \int d\mathbf{r}^n \int D\sigma(\mathbf{r}) \exp \left(i \rho(\mathbf{r}) \sigma(\mathbf{r}) - \frac{\beta}{2} \int d\mathbf{r} d\mathbf{r}' \rho(\mathbf{r}) u(|\mathbf{r} - \mathbf{r}'|) \rho(\mathbf{r}') - \beta \sum_I U_I(\mathbf{r}^n) \right) \\ &\sim \int D\sigma(\mathbf{r}) \exp \left(-\frac{1}{2\beta} \int d\mathbf{r} d\mathbf{r}' \sigma(\mathbf{r}) u^{-1}(|\mathbf{r} - \mathbf{r}'|) \sigma(\mathbf{r}') \right) \int d\mathbf{r} e^{-i \rho(\mathbf{r}) \sigma(\mathbf{r}) - \beta \sum_I U_I(\mathbf{r}^n)}, \end{aligned}$$

where $\sigma(\mathbf{r})$ denotes an auxiliary field and the inverse operator $u^{-1}(|\mathbf{r} - \mathbf{r}'|)$ is defined in reciprocal space. This is now a partition function of field variable $\sigma(\mathbf{r})$ instead of particles. To generalize these expressions to multicomponent systems, the densities and auxiliary fields become vector-valued with one value per component, and the pair potential operator should be understood as matrix-valued. Though this transformation can be conceptually easy to extend, it can be prohibitively hard to evaluate numerically especially there involves high wave number components in the interaction, which could probably be the case in the high resolution atomistic models.

At this point, there are a number of existing strategies available, but none can accurately and reliably describe general multicomponent molecular liquids with complex orientational correlations. For several simple and/or highly approximated cases, a feasible numerical scheme can be derived. If the liquid_{SEP}^[11] were a simple single-component fluid with a reasonably soft pair

interaction, the partition function could be approximated accurately using a perturbation series around a mean-field theory^{6, 14} or after a suitable Gaussian Equivalent Representation (GER) transformation.¹⁹⁻²³ This approach has been applied to systems interacting through repulsive Gaussian, Coulombic as well as Yukawa potentials.^{19-20, 24-25} If the liquid were a polymers blend, one could approximate the integral by an appropriate Edwards Hamiltonian based on either local density approximation or mean-field approximation.^{6, 12-13} There has been several successful field theoretic simulation techniques to evaluate the partition function. To simulate a molecular liquid with field theory requires specialized methods such the 3D RISM integral equations,^{4-5, 11} and even these have serious deficiencies that prevent their wide adoption. The application of field theoretic techniques, as can be seen above, is still fairly limited and highly dependent on somewhat *ad hoc* approximations of the system Hamiltonian. In this work, we propose a new strategy of constructing field theoretic representation of molecular fluids that involves minimal approximations of the system Hamiltonian, through a hierarchical coarse-graining procedure.

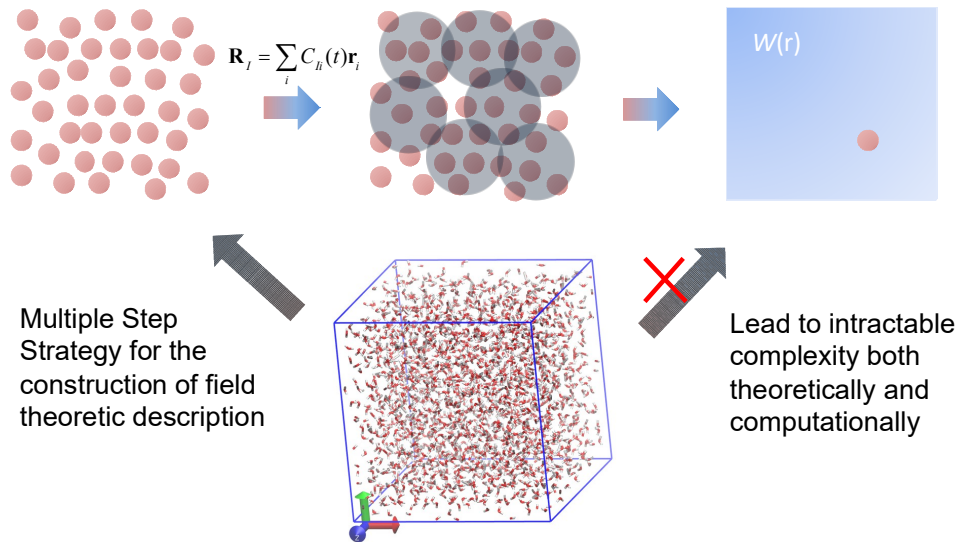


Figure 6-1: schematic diagram showing the concept of hierarchical coarse-graining for the construction of field theoretic representation of fluids.

Instead of directly transforming the atomistic Hamiltonian into the field representation, we first construct a particle based coarse-grained model of the liquid molecules. The Hamiltonian for this coarse-grained model, in which the coarse-grained particles are assumed to interact through pairwise additive potentials, can be derived through the well-established bottom-up coarse-graining methodology like Multiscale Coarse-graining²⁶⁻³⁰ (MS-CG) or Relative Entropy Minimization³¹⁻³² (REM). Since the large wavenumber features in the atomistic representation are usually smeared out in coarse-grained potentials,³³ the HS transformation can be much more readily applied. The resulting field theoretic partition function can be much more easier to solve quantitatively by any of the methods described above than the corresponding fine-grained scenarios. And at the same time, the essential structural correlation is preserved through the systematic coarse-graining parameterization instead of the *ad hoc* assumptions on the interaction forms. The details of the development will be elaborated in the next section.

II. Theory

For the conciseness purpose, we only elaborate the homogeneous single component liquid case. The theoretical generalization to multi-component system is straightforward by extending the scalar field to vector fields. The first step of the field theory construction is building the coarse-grained model of the molecular fluids from the atomistic model. In particular, we represent each of the fluid molecules by their center of masses, which denotes as $\mathbf{r}_i$ throughout this chapter. And the interactions between these center of mass CG sites are assumed to be pairwise additive, which can be expressed as:

$$U(\mathbf{r}^n) = \frac{1}{2} \sum_{i,j} u(\mathbf{r}_i - \mathbf{r}_j) - \frac{1}{2} N u(0), \quad (6.1)$$

Note that the double sum in the first term includes the self-interacting term $u(0)$ that should not appear in the Hamiltonian. Therefore, the second term is introduced to correct this self-energy. This pairwise interaction can be derived through various bottom-up coarse-graining methods, such as MS-CG, REM etc.

After obtaining the coarse-grained representation of the fluids, which renders a much more smooth representation of the energy surface due to the reduced spatial resolution, the particle based representation of the system can then be transformed into field representation. To implement such transformation, we first introduce the density field operator as well as its Fourier transforms,

$$\begin{aligned}\rho(\mathbf{r}) &= \sum_i \delta(\mathbf{r} - \mathbf{r}_i) \\ \rho(\mathbf{k}) &= \int d\mathbf{r} \rho(\mathbf{r}) \exp(-i\mathbf{k} \cdot \mathbf{r}) = \sum_j \exp(-i\mathbf{k} \cdot \mathbf{r}_j)\end{aligned}\quad (6.2)$$

in which the $\mathbf{k}$ vector denotes the reciprocal lattice vectors defined through the periodic boundary conditions. Plugging the local density operator into Eq. (6.1), The free energy function can then be easily re-expressed in the reciprocal space,

$$\begin{aligned}U(\mathbf{r}^n) &= \frac{1}{2} \sum_{i,j} u(\mathbf{r}_i - \mathbf{r}_j) - \frac{1}{2} Nu(0) \\ &= \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \rho(\mathbf{r}) u(\mathbf{r} - \mathbf{r}') \rho(\mathbf{r}') - \frac{1}{2} Nu(0) \\ &= \frac{1}{2} \int d\mathbf{r} d\mathbf{r}' \rho(\mathbf{r}) e^{i\mathbf{k} \cdot \mathbf{r}} \sum_{\mathbf{k}} u(\mathbf{k}) \rho(\mathbf{r}') e^{-i\mathbf{k} \cdot \mathbf{r}'} - \frac{1}{2} Nu(0) \\ &= \frac{1}{2} \sum_{\mathbf{k}} \rho(\mathbf{k})^* u(\mathbf{k}) \rho(\mathbf{k}) - \frac{1}{2} Nu(0)\end{aligned}\quad (6.3)$$

In this representation, the free energy function can be regarded as a quadratic function with respect to the Fourier modes of the density field operator. The nice property of quadratic forms

enables one to transform the partition function defined by the above energy function into a field representation through the famous Hubbard-Stratovonich (HS) transform,¹⁸

$$\exp(-\frac{1}{2}ax^2) = \frac{1}{\sqrt{2\pi a}} \int dy \exp(-\frac{y^2}{2a} - ixy), \quad (6.4)$$

in which the real constant $a > 0$. However, for a general pairwise potential $u(r)$ like what we often obtained in bottom-up coarse-grained models of molecular fluids, we cannot simply assume that its Fourier transform is always positive. In fact, only for very limited cases such as the purely repulsive Gaussian interaction, Yukawa potential and repulsive Coulomb interactions satisfy this condition. Previously, Baeurle *et al* applied the HS transform to these limited situations that gives rise to a auxiliary field representation of system,²⁰⁻²¹ and they managed to obtain consistent result with particle based numerical simulations. Fortunately, the extension to general cases can be achieved by introducing two auxiliary fields instead of just one. This generalization will be systematically developed in this work.

Before elaborating the generalization, we would like to derive the field theoretic representation for the special situations of positive Fourier modes first, so that the later generalization can be more intuitive and better understood. Under the pairwise interaction, the partition function of the molecular system under canonical ensemble can be written as,

$$Z(N, V, T) = \frac{1}{\lambda_B^{3N} N!} \int d\mathbf{r}^n \exp(-\beta U(\mathbf{r}^n)). \quad (6.5)$$

in which λ_B represents the thermal de Broglie wavelength. By inspecting Eq. (6.3) and the form of HS transform, it is apparent that for a single term of the summation over Fourier modes, an integral over a real auxiliary field variable σ can be introduced.

Before the actual transformation, we first introduce a set of straightforward properties for real field variables like σ , u and ρ , which can be very useful in the following derivations throughout the chapter,

$$\begin{aligned}\sum_{\mathbf{k} \neq 0} \sigma(\mathbf{k}) \rho^*(\mathbf{k}) &= 2 \sum_{G>0} (\sigma^R(\mathbf{k}) \rho^R(\mathbf{k}) + \sigma^I(\mathbf{k}) \rho^I(\mathbf{k})) \\ \sigma^*(\mathbf{k}) \sigma(\mathbf{k}) &= \sigma^*(-\mathbf{k}) \sigma(-\mathbf{k}) \\ \sum_{\mathbf{k}>0} \frac{1}{u(\mathbf{k})} \sigma^*(\mathbf{k}) \sigma(\mathbf{k}) &= \frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{1}{u(\mathbf{k})} \sigma^*(\mathbf{k}) \sigma(\mathbf{k})\end{aligned}\quad , \quad (6.6)$$

The derivation starts with the observation that the quadratic term in Eq. (3) can be written as,

$$\begin{aligned}\sum_{\mathbf{k} \neq 0} \rho^*(\mathbf{k}) u(\mathbf{k}) \rho(\mathbf{k}) &= \sum_{\mathbf{k} < 0} \rho^*(\mathbf{k}) u(\mathbf{k}) \rho(\mathbf{k}) + \sum_{\mathbf{k} > 0} \rho^*(\mathbf{k}) u(\mathbf{k}) \rho(\mathbf{k}) \\ &= \sum_{\mathbf{k} > 0} |\rho(-\mathbf{k})|^2 u(-\mathbf{k}) + \sum_{\mathbf{k} > 0} |\rho(\mathbf{k})|^2 u(\mathbf{k}) \\ &= \sum_{\mathbf{k} > 0} |\rho(\mathbf{k})|^2 (u(-\mathbf{k}) + u(\mathbf{k})) \\ &= 2 \sum_{\mathbf{k} > 0} |\rho(\mathbf{k})|^2 \text{Re}\{u(\mathbf{k})\}\end{aligned}\quad , \quad (6.7)$$

Eq. (6.7) suggests that only the real part of the Fourier transform of the pairwise interaction contributes to the total energy. Therefore, we define the real part $\phi(\mathbf{k}) = \text{Re}\{u(\mathbf{k})\}$, which is assumed to be positive in this scenario. By applying the HS transform, it is straightforward to find out that,

$$\begin{aligned}\exp\left(-\beta \frac{1}{2} \sum_{\mathbf{k} \neq 0} \rho(\mathbf{k})^* u(\mathbf{k}) \rho(\mathbf{k})\right) &= \exp\left[-\beta \sum_{\mathbf{k} > 0} |\rho(\mathbf{k})|^2 \phi(\mathbf{k})\right] \\ &= \exp\left[-\beta \sum_{\mathbf{k} > 0} (\rho^I(\mathbf{k})^2 + \rho^R(\mathbf{k})^2) \phi(\mathbf{k})\right] = \prod_{\mathbf{k} > 0} \left(\frac{1}{4\pi\beta\phi(\mathbf{k})}\right)^{1/2} \int \prod_{\mathbf{k} > 0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k}) \cdot \\ &\exp\left[-\frac{1}{2} \sum_{\mathbf{k} > 0} \frac{1}{2\beta\phi(\mathbf{k})} (\sigma_R(\mathbf{k})^2 + \sigma_I(\mathbf{k})^2) - i \sum_{\mathbf{k} > 0} (\rho_R(\mathbf{k}) \sigma_R(\mathbf{k}) + \rho_I(\mathbf{k}) \sigma_I(\mathbf{k}))\right]\end{aligned}\quad (6.8)$$

Note that the auxiliary field we introduced here is assumed to be a real field in the coordinate representation, i.e. $\sigma(r)$ is real. This assumption is easily satisfied since the local density field $\rho(\mathbf{r})$ is a real field variable.

This intermediate result can further be simplified by absorbing factor of 2 into the integration variable, which transform the exponential integrand in the partition function expression into the following form,

$$\begin{aligned}
& \exp \left[-\frac{1}{2} \beta \sum_{\mathbf{k} \neq 0} \rho^*(\mathbf{k}) u(\mathbf{k}) \rho(\mathbf{k}) \right] \\
&= \prod_{\mathbf{k} > 0} \left(\frac{1}{\pi \beta \phi(\mathbf{k})} \right) \int \prod_{\mathbf{k} > 0} d \frac{\xi(\mathbf{k})}{2} d \frac{\eta(\mathbf{k})}{2} \\
&\quad \times \exp \left[-\sum_{\mathbf{k} > 0} \frac{1}{\beta \phi(\mathbf{k})} \left(\frac{\xi(\mathbf{k})^2}{4} + \frac{\eta(\mathbf{k})^2}{4} \right) - 2i \sum_{\mathbf{k} > 0} \left(\rho_R(\mathbf{k}) \frac{\xi(\mathbf{k})}{2} + \rho_I(\mathbf{k}) \frac{\eta(\mathbf{k})}{2} \right) \right] \quad , \quad (6.9) \\
&= \prod_{\mathbf{k} \neq 0} \left(\frac{1}{\pi \beta \phi(\mathbf{k})} \right)^{1/2} \int \prod_{\mathbf{k} > 0} d \sigma_R(\mathbf{k}) d \sigma_I(\mathbf{k}) \exp \left[-\frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{\sigma^*(\mathbf{k}) \sigma(\mathbf{k})}{\beta \phi(\mathbf{k})} - i \sum_{\mathbf{k} \neq 0} \sigma^*(\mathbf{k}) \rho(\mathbf{k}) \right]
\end{aligned}$$

the last equality holds as a result of Eq. (6.6).

With this result in hand, the canonical partition function Eq. (6.5) can be readily transformed into an integral of the auxiliary field that we introduced in the above derivation instead of the particle Cartesian coordinates,

$$\begin{aligned}
Z(N, V, T) &= \frac{1}{\lambda_B^{3N} N!} \int d\mathbf{r}^n \exp(-\beta U(\mathbf{r}^n)) \\
&= \frac{1}{\lambda_B^{3N} N!} \int d\mathbf{r}^n \exp(-\frac{1}{2} N \beta \phi(0)) \frac{1}{\sqrt{2}} \prod_{\mathbf{k}} \left(\frac{1}{\pi \beta \phi(\mathbf{k})} \right)^{1/2} \\
&\quad \times \int d\sigma(0) \prod_{\mathbf{k}>0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k}) \exp \left[-\frac{1}{2} \sum_{\mathbf{k}} \frac{\sigma^*(\mathbf{k}) \sigma(\mathbf{k})}{\beta \phi(\mathbf{k})} - i \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \rho(\mathbf{k}) \right], \quad (6.10) \\
&= \frac{1}{\sqrt{2} \lambda_B^{3N} N!} \exp(-\frac{1}{2} N \beta \phi(0)) \prod_{\mathbf{k}} \left(\frac{1}{\pi \beta \phi(\mathbf{k})} \right)^{1/2} \\
&\quad \int d\sigma(0) \prod_{\mathbf{k}>0} d\sigma(\mathbf{k}) \exp \left[-\frac{1}{2} \sum_{\mathbf{k}} \frac{\sigma^*(\mathbf{k}) \sigma(\mathbf{k})}{\beta \phi(\mathbf{k})} \right] \int d\mathbf{r}^n \exp \left[-i \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \rho(\mathbf{k}) \right]
\end{aligned}$$

in which the HS transform with respect to the zero frequency mode is also introduced. The last term in Eq. (6.10) involving integral over configurations can be further simplified,

$$\begin{aligned}
&\int d\mathbf{r}^n \exp \left(-i \sum_{\mathbf{k}} \sigma(\mathbf{k}) \rho(\mathbf{k})^* \right) \\
&= \int d\mathbf{r}^n \exp \left(-i \sum_{\mathbf{k}} \sigma(\mathbf{k}) \sum_i \exp(i\mathbf{k} \cdot \mathbf{r}_i) \right) \quad . \quad (6.11) \\
&= \int d\mathbf{r}^n \exp \left(-i \sum_i \sigma(\mathbf{r}_i) \right) = \left(\int d\mathbf{r} \exp(-i\sigma(\mathbf{r})) \right)^N
\end{aligned}$$

Plugging this result into the partition function, one can finally obtain an expression of the canonical partition function as a functional integral with respect to an auxiliary field,

$$Z(N, V, T) = \frac{1}{\lambda_B^3 N!} \int d\mu(\sigma(\mathbf{r})) \exp(-S[\sigma(\mathbf{r})]), \quad (6.12)$$

in which the integral measure $d\mu[\sigma] = \exp \left(-\frac{1}{2} N \beta \phi(0) \right) \prod_{\mathbf{k} \neq 0} \left(\frac{1}{\pi \beta \phi(\mathbf{k})} \right)^{1/2} \prod_{\mathbf{k}>0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k})$ and

the effective action functional $S[\sigma(\mathbf{r})]$ takes the following form,

$$S[\sigma(\mathbf{r})] = \frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{\sigma^*(\mathbf{k}) \sigma(\mathbf{k})}{\beta \phi(\mathbf{k})} - N \ln \int d\mathbf{r} e^{-i\sigma(\mathbf{r})}. \quad (6.13)$$

In grand canonical ensemble, the same derivation procedure still holds and the grand canonical partition function can be derived based on the canonical partition function,

$$\begin{aligned}
\Theta(\mu, V, T) &= \sum_{N=0}^{\infty} e^{\beta\mu N} Z(N, V, T) \\
&= \int d\mu[\sigma(\mathbf{r})] \exp\left(-\frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{\sigma^*(\mathbf{k})\sigma(\mathbf{k})}{\beta\phi(\mathbf{k})}\right) \sum_{N=0}^{\infty} \frac{1}{\lambda_B^{3N} N!} \exp N\left(\beta\mu - \ln \int d\mathbf{r} e^{-i\sigma(\mathbf{r})}\right), \quad (6.14) \\
&= \int D\mu[\sigma(\mathbf{r})] \exp\left(-\frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{\sigma^*(\mathbf{k})\sigma(\mathbf{k})}{\beta\phi(\mathbf{k})}\right) \exp\left(\chi \int d\mathbf{r} e^{-i\sigma(\mathbf{r})}\right) \\
&\equiv \int D\mu[\sigma(\mathbf{r})] \exp(-S[\sigma(\mathbf{r})])
\end{aligned}$$

in which the factor $\chi = \exp(\beta\mu + \frac{1}{2}\beta\phi(0)) / \lambda_B^3$ and the integration measure

$$D\mu[\sigma] = \prod_{\mathbf{k}} \frac{1}{\sqrt{2}} \left(\frac{1}{\pi\beta\phi(\mathbf{k})} \right)^{1/2} d\sigma(0) \prod_{\mathbf{k}>0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k}).$$

The effective action functional in grand canonical ensemble can be expressed as,

$$S[\sigma(\mathbf{r})] = \frac{1}{2} \sum_{\mathbf{k} \neq 0} \frac{\sigma^*(\mathbf{k})\sigma(\mathbf{k})}{\beta\phi(\mathbf{k})} - \chi \int d\mathbf{r} e^{-i\sigma(\mathbf{r})}. \quad (6.15)$$

By sampling the auxiliary field with respect to the effective action functional according to the Boltzmann weights, one can evaluate the thermodynamic information as well as structural correlations of the system. Details about the sampling algorithm will be introduced in following sections.

The results shown in Eqs. (6.12) and (6.15) conclude the construction of field theoretic representation of pairwise interacting molecules, in which the Fourier transform of the pairwise interaction is always positive. To generalize the approach to arbitrary pairwise interaction cases without this strong assumption, one must introduce two auxiliary fields instead of just one single field variable.

Notice that instead of the normal imaginary HS transform, which requires the pre-factor is greater than zero, one can prove that the following relationship holds:

$$\exp\left(-\frac{1}{2}ax^2\right) = \frac{1}{\sqrt{2\pi|a|}} \int dy \exp\left(-\frac{y^2}{2|a|} - xy\right), \quad (6.16)$$

for any real number $a < 0$. This result suggests that the negative Fourier modes of the pairwise interaction can also be transformed by the modified HS formula. Suppose that we can decompose the real part of the Fourier modes of pairwise interaction into two components $\phi(\mathbf{k}) = v(\mathbf{k}) + w(\mathbf{k})$, in which $u(\mathbf{k})$ is the Fourier transform of the general form pairwise interaction, $v(\mathbf{k})$ is always positive whereas $w(\mathbf{k})$ stays negative for any reciprocal lattice vector $\mathbf{k}$. By performing this decomposition, we can express the free energy of the system as:

$$\begin{aligned} U(\mathbf{r}^n) &= \frac{1}{2} \sum_{\mathbf{k}} \rho(\mathbf{k})^* u(\mathbf{k}) \rho(\mathbf{k}) - \frac{1}{2} Nu(0) \\ &= \frac{1}{2} \sum_{\mathbf{k}} \rho(\mathbf{k})^* (v(\mathbf{k}) + w(\mathbf{k})) \rho(\mathbf{k}) - \frac{1}{2} Nu(0) \end{aligned} \quad (6.17)$$

Under the same procedure as the derivation of Eq. (6.9), the repulsive components of the Fourier modes can be readily transformed as a field theory by introducing an auxiliary field variable,

$$\begin{aligned} \exp\left[-\frac{1}{2}\beta \sum_{\mathbf{k}} \rho^*(\mathbf{k}) v(\mathbf{k}) \rho(\mathbf{k})\right] &= \prod_{\mathbf{k}} \frac{1}{\sqrt{2}} \left(\frac{1}{\pi\beta v(\mathbf{k})}\right)^{1/2} \int d\sigma(0) \prod_{\mathbf{k}>0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k}) \\ &\quad \times \exp\left[-\frac{1}{2} \sum_{\mathbf{k}} \frac{\sigma^*(\mathbf{k}) \sigma(\mathbf{k})}{\beta v(\mathbf{k})} - i \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \rho(\mathbf{k})\right] \end{aligned} \quad (6.18)$$

As for the negative part, similar procedure holds but the modified HS transform Eq. (6.16) is applied, which renders the following result,

$$\exp\left[-\frac{1}{2}\beta\sum_{\mathbf{k}}\rho^*(\mathbf{k})w(\mathbf{k})\rho(\mathbf{k})\right] = \prod_{\mathbf{k}} \frac{1}{\sqrt{2}} \left(\frac{1}{\pi\beta|w(\mathbf{k})|}\right)^{1/2} \int d\xi(0) \prod_{\mathbf{k}>0} d\xi_R(\mathbf{k}) d\xi_I(\mathbf{k}) \times \exp\left[-\frac{1}{2}\sum_{\mathbf{k}} \frac{\xi^*(\mathbf{k})\xi(\mathbf{k})}{\beta|w(\mathbf{k})|} - \sum_{\mathbf{k}} \xi^*(\mathbf{k})\rho(\mathbf{k})\right], \quad (6.19)$$

in which the second auxiliary field $\xi(\mathbf{k})$ was introduced to transform the negative part of the interaction into field theoretic representation. By combining Eq. (6.18) and Eq. (6.19) together into the partition function, one finally arrives at a field-theoretic representation of the canonical partition function of the system,

$$Z(N, V, T) = \frac{1}{2\lambda_B^3 N!} \exp\left(\frac{\beta N u(0)}{2}\right) \prod_{\mathbf{k}} \left(\frac{1}{\pi\beta v(\mathbf{k})}\right)^{1/2} \left(\frac{1}{\pi\beta|w(\mathbf{k})|}\right)^{1/2} \times \int D\sigma(\mathbf{r}) D\xi(\mathbf{r}) \left\{ \exp\left[-\frac{1}{2}\sum_{\mathbf{k}} \frac{\sigma^*(\mathbf{k})\sigma(\mathbf{k}) + \xi^*(\mathbf{k})\xi(\mathbf{k})}{\beta|\phi(\mathbf{k})|}\right] \right. \\ \left. \times \left(\int d\mathbf{r} \exp(-i\sigma(\mathbf{r}) - \xi(\mathbf{r}))\right)^N \right\} \quad (6.20)$$

in which the integrands are short-hand noted as $D\sigma(\mathbf{r}) = d\sigma(0) \prod_{\mathbf{k}>0} d\sigma_R(\mathbf{k}) d\sigma_I(\mathbf{k})$ and

$D\xi(\mathbf{r}) = d\xi(0) \prod_{\mathbf{k}>0} d\xi_R(\mathbf{k}) d\xi_I(\mathbf{k})$. To inspect the mathematical structure more clearly, Eq. (6.20)

can be written in a more compact way,

$$Z(N, V, T) = \frac{1}{\lambda_B^3 N!} \int d\mu(\sigma, \xi) \exp(-S[\sigma, \xi]), \quad (6.21)$$

with the integration measure defined as,

$$d\mu[\sigma, \xi] = \frac{1}{2} \exp\left(\frac{\beta N u(0)}{2}\right) \prod_{\mathbf{k}} \left(\frac{1}{\pi\beta v(\mathbf{k})}\right)^{1/2} \left(\frac{1}{\pi\beta|w(\mathbf{k})|}\right)^{1/2} D\sigma(\mathbf{r}) D\xi(\mathbf{r}), \quad (6.22)$$

and the action functional now can be written as,

$$S[\sigma, \xi] = \frac{1}{2} \sum_{\mathbf{k}} \frac{1}{\beta v(\mathbf{k})} \sigma^*(\mathbf{k}) \sigma(\mathbf{k}) + \frac{1}{2} \sum_{\mathbf{k}} \frac{1}{\beta |w(\mathbf{k})|} \xi^*(\mathbf{k}) \xi(\mathbf{k}) - N \ln \int d\mathbf{r} e^{-i\sigma(\mathbf{r}) - \xi(\mathbf{r})}. \quad (6.23)$$

And similarly, the grand canonical partition function can be obtained based on the canonical ensemble partition function,

$$\Theta(\mu, V, T) = \sum_{N=0}^{\infty} e^{\beta \mu N} Z(N, V, T) = \int d\mu[\sigma, \xi] \exp(-S_{GC}[\sigma, \xi]), \quad (6.24)$$

in which the effective action functional is expressed as,

$$S_{GC}[\sigma, \xi] = \frac{1}{2} \sum_{\mathbf{k}} \frac{1}{\beta v(\mathbf{k})} \sigma^*(\mathbf{k}) \sigma(\mathbf{k}) + \frac{1}{2} \sum_{\mathbf{k}} \frac{1}{\beta |w(\mathbf{k})|} \xi^*(\mathbf{k}) \xi(\mathbf{k}) - \chi \int d\mathbf{r} e^{-i\sigma(\mathbf{r}) - \xi(\mathbf{r})}. \quad (6.25)$$

Eqs. (6.21)-(6.25) are the main result of this section. It is worth noting that in the effective action functional defined in Eq. (6.23) and Eq. (6.25), the terms associated with the auxiliary field variable $\xi(\mathbf{k})$ are always real and therefore correspond to positive statistical weights. Therefore, they do not contribute to the numerical “sign-problem” for Monte-Carlo sampling. The main numerical obstacles come from the repulsive interactions.

III. Gaussian Equivalent Representation

The “sign-problem”³⁴⁻³⁷ is considered to be one of the unsolved problems in many-body physics. An exact algorithm for solving the sign-problem has been shown to be NP hard,³⁸ which is prohibitively expensive for sampling many-body systems. Various approximations have been proposed to improve the convergence efficiency of sampling such oscillatory functions. Among these approximations, a very common strategy relies on the shifted-contour representation of the oscillatory integral, which was originally proposed to solve the sign problem in the auxiliary field Monte Carlo method for electronic structure calculations.³⁶⁻³⁷ In this section, we will propose a solution for mitigating the sign problem in the context of sampling Eq. (6.25), based

on a Gaussian Equivalent Representation (GER) of the shifted-contour representation of partition function. Efimov and Baeurle et al had studied this GER strategy,^{19-20, 23-24} and a more crude mean field representation (MFR),²⁰⁻²¹ of various field theoretic representations of model liquid systems interacting through special pairwise interactions. We would like to generalize the GER in the coarse-grained field representation of an arbitrary molecular fluid developed in Sec. II, and the statistical sampling algorithm associated with such scenario.

Before deriving the GER, we need to derive the shifted-contour representation of the partition function. Similar as in Sec. II, the GER for the purely positive case that involves only one auxiliary field variable is elaborated first for better understanding, and then it can be naturally extended to the general cases involving two auxiliary variables. And for the sake of brevity, the shifted-contour integral and GER is built upon the grand canonical partition function.

Eqs. (6.24) to (6.25) define the grand canonical partition function in the field theoretic representation. The shifted-contour representation of the integral starts with changing the field variable following the rule,

$$\begin{aligned}\sigma^R(\mathbf{r}) &\rightarrow \sigma^R(\mathbf{r}) + i\psi^R(\mathbf{r}) \\ \sigma^I(\mathbf{r}) &\rightarrow \sigma^I(\mathbf{r}) + i\psi^I(\mathbf{r})\end{aligned}, \quad (6.26)$$

in which the $\psi(\mathbf{r})$ is a pre-defined real shifted-contour field variable, without losing generality. Based on the Cauchy's integral theorem, this change of integration path keeps the integral value invariant. In this new representation, the effective action functional can be expressed as,

$$S[\sigma(r)] = \frac{1}{2} \sum_{\mathbf{k}} \frac{1}{\beta\phi(\mathbf{k})} \left(\sigma^*(\mathbf{k})\sigma(\mathbf{k}) + 2i\psi^*(\mathbf{k})\sigma(\mathbf{k}) \right) - \chi \int d\mathbf{r} e^{-i\sigma(\mathbf{r}) + \psi(\mathbf{r})}, \quad (6.27)$$

And the partition function now writes as,

$$\Theta(\mu, V, T) = \exp\left[\frac{1}{2\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k})\phi^{-1}(\mathbf{k})\psi(\mathbf{k})\right] \int D\mu[\sigma(\mathbf{r})] \exp(-S[\sigma(\mathbf{r})]). \quad (6.28)$$

The idea behind the shifted-contour representation is to find an appropriate distortion of the integral path in the complex plane so that the fluctuations of the oscillating function can be reduced. In doing so, we first introduce a normal product with respect to a Gaussian integral

measure $D\mu_\phi[\sigma] = \gamma_\phi D\sigma \exp\left[-\frac{1}{2\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k})\phi^{-1}(\mathbf{k})\sigma(\mathbf{k})\right]$ defined by a pairwise interaction $\phi(r)$

,

$$\begin{aligned} : \exp[-i\sigma(\mathbf{r})] :_\phi &= \exp[-i\sigma(\mathbf{r})] \exp\left(\frac{1}{2}\beta\phi(0)\right). \\ \sigma(\mathbf{r})\sigma(\mathbf{r}') &=: \sigma(\mathbf{r})\sigma(\mathbf{r}') :_\phi + \beta\phi(\mathbf{r} - \mathbf{r}') \end{aligned} \quad (6.29)$$

These notations ensure that the relationship $\int d\mu_\phi[\sigma] : \exp[-i\sigma(\mathbf{r})] :_\phi = 1$ and

$\int d\mu_\phi[\sigma] : \sigma(\mathbf{r})\sigma(\mathbf{r}') :_\phi = 0$ holds.

By introducing the normal products, the grand canonical partition function can now be recast as,

$$\begin{aligned} \Theta(\mu, V, T) &= \exp\left[\frac{1}{2\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k})\phi^{-1}(\mathbf{k})\psi(\mathbf{k})\right] \\ &\times \int d\mu_\phi[\sigma] \exp\left[-\frac{i}{\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k})\phi^{-1}(\mathbf{k})\sigma(\mathbf{k}) + z \int dr e^{\psi(r)} : e^{-i\sigma(r)} :_\phi\right], \end{aligned} \quad (6.30)$$

where the fugacity $z = \chi e^{-\frac{1}{2}\beta\phi(0)}$ is introduced to simplify the notation.

The idea of GER is then to find a specific distortion of the integration contour such that the amplitude of oscillation can be reduced. The intuition behind GER comes from the desirability of finding a Gaussian integration measure so that the deviance between the integrand in Eq. (6.28) and this Gaussian measure is minimized. Then one can imagine that the oscillation of the integrand can be suppressed by sampling the phase space according to this Gaussian measure. Similar calculations have been widely used in quantum field theory (QFT). In QFT, it has been

observed that the main contributions to the field partition function comes from lower-order tadpole type Feynman diagrams.²² This fast convergence can effectively be achieved if the interaction functional does not contain any linear and quadratic terms over the field variables.

Based on this intuition, one can first define an arbitrary pairwise potential energy function $\eta(r)$, which generates a Gaussian measure $d\mu_\eta[\sigma] = \gamma_\eta D\sigma \exp\left[-\frac{1}{2\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) D^{-1}(\mathbf{k}) \sigma(\mathbf{k})\right]$. The normal product defined with this measure writes as $: \exp[-i\sigma(\mathbf{r})] :_\eta = C^{-1} : \exp[-i\sigma(\mathbf{r})] :_\phi$, in which $C^{-1} = \exp\left[\frac{\beta}{2}(\phi(0) - \eta(0))\right]$. With this new Gaussian measure, the grand canonical partition function Eq. (30) can be written as an integral with this new Gaussian measure,

$$\Theta(\mu, V, T) = \frac{\gamma_\phi}{\gamma_\eta} \exp\left[\frac{1}{2\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k}) \phi^{-1}(\mathbf{k}) \psi(\mathbf{k})\right] \int d\mu_\eta[\sigma] \exp(-K), \quad (6.31)$$

in which the shifted action functional,

$$K = \frac{1}{2\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \left[\phi^{-1}(\mathbf{k}) - \eta^{-1}(\mathbf{k}) \right] \sigma(\mathbf{k}) + \frac{i}{\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k}) \phi^{-1}(\mathbf{k}) \sigma(\mathbf{k}) - zC \int d\mathbf{r} \exp(\psi(\mathbf{r})) : \exp(-i\sigma(\mathbf{r})) :_\eta. \quad (6.32)$$

As mentioned earlier, in order to reduce the oscillation of the term $\exp(-K)$ and thus improve convergence, one could simply get rid of the linear and quadratic terms involving the auxiliary field variables. By expanding the normal product in Taylor series and separate out terms up to the second order,

$$\begin{aligned} : \exp(-i\sigma(\mathbf{r})) :_\eta &= 1 - i\sigma(\mathbf{r}) - \frac{1}{2} : \sigma^2(\mathbf{r}) :_\eta + : \exp_2(-i\sigma(\mathbf{r})) :_\eta \\ &= 1 - i\sigma(\mathbf{r}) - \frac{1}{2} \sigma^2(\mathbf{r}) + \frac{1}{2} \beta \eta(0) + : \exp_2(-i\sigma(\mathbf{r})) :_\eta \end{aligned}, \quad (6.33)$$

in which the term $\exp_2(-i\sigma(\mathbf{r})) = \exp(-i\sigma(\mathbf{r})) - 1 - i\sigma(\mathbf{r}) - \frac{1}{2}\sigma^2(\mathbf{r})$ represents contributions from all the higher order terms with respect to the field variable. Plug Eq. (6.33) into Eq. (6.32), the functional K is transformed as,

$$K = \frac{1}{2\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \left[\phi^{-1}(\mathbf{k}) - \eta^{-1}(\mathbf{k}) \right] \sigma(\mathbf{k}) + \frac{i}{\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k}) \phi^{-1}(\mathbf{k}) \sigma(\mathbf{k}) - zC \int d\mathbf{r} \exp(\psi(\mathbf{r})) \left(1 - i\sigma(\mathbf{r}) - \frac{1}{2}\sigma^2(\mathbf{r}) + \frac{1}{2}\beta\eta(0) + : \exp_2(-i\sigma(\mathbf{r})) :_{\eta} \right). \quad (6.34)$$

To concentrate the integral in the Gaussian measure, one wish to see that the linear and quadratic terms with respect to $\sigma(\mathbf{r})$ vanishes in Eq. (6.34). To assure this, the following two equations need to be satisfied,

$$\begin{aligned} \frac{1}{\beta} \sum_{\mathbf{k}} \psi^*(\mathbf{k}) \phi^{-1}(\mathbf{k}) \sigma(\mathbf{k}) + zC \int d\mathbf{r} \exp(\psi(\mathbf{r})) \sigma(\mathbf{r}) d\mathbf{r} &= 0 \\ \frac{1}{\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \left[\phi^{-1}(\mathbf{k}) - \eta^{-1}(\mathbf{k}) \right] \sigma(\mathbf{k}) + zC \int d\mathbf{r} \exp(\psi(\mathbf{r})) \sigma^2(\mathbf{r}) d\mathbf{r} &= 0 \end{aligned} \quad (6.35)$$

Assuming the distortion field $\psi(r)$ is simply a constant field, and therefore all the higher order Fourier modes vanishes. Under this assumption, the first equality in Eq. (6.35) turns into,

$$-\frac{1}{\beta} \frac{\psi(\mathbf{k}=0)}{\phi(\mathbf{k}=0)} = V \chi \exp[\psi(\mathbf{k}=0)] \exp\left[-\frac{1}{2}\beta\eta(0)\right]. \quad (6.36)$$

And due to the Parseval's theorem, the second equation turns into,

$$\frac{1}{\beta} \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \left[\phi^{-1}(\mathbf{k}) - D^{-1}(\mathbf{k}) \right] \sigma(\mathbf{k}) + zC \exp(\psi(r)) \sum_{\mathbf{k}} \sigma^*(\mathbf{k}) \sigma(\mathbf{k}) = 0. \quad (6.37)$$

To ensure that the above relationship Eq. (6.36) holds, the reference potential that generates the Gaussian measure needs to satisfy,

$$\eta(\mathbf{k}) = \frac{\phi(\mathbf{k})}{1 - \frac{\psi(\mathbf{k}=0)}{\phi(\mathbf{k}=0)} \phi(\mathbf{k})}, \quad (6.38)$$

which together with Eq. (6.36) and the property $\eta(0) = \sum_{\mathbf{k}} \eta(\mathbf{k})$ forms a complete set of equations that can be iteratively solved to find out the shifted contour and reference potential.

Finally, the GER representation of the grand canonical partition function for this special case now renders as,

$$\Theta(\mu, V, T) = \Theta_0 \int d\mu_\eta [\sigma] \exp(-zC \exp(\psi(\mathbf{k}=0)) \int d\mathbf{r} : \exp_2(-i\sigma(\mathbf{r})) :_\eta), \quad (6.39)$$

with the zeroth order approximation of the partition function written as,

$$\Theta_0 = \frac{\gamma_\eta}{\gamma_\phi} \exp \left(-\frac{\psi^2(\mathbf{k}=0)}{2\beta\phi(\mathbf{k}=0)} - zC V e^{\psi(\mathbf{k}=0)} \left(1 + \frac{1}{2}\beta\eta(0)\right) \right). \quad (6.40)$$

The above derivation of GER can be extended to the general cases involving two auxiliary field variables. In doing so, we first introduce a compact representation of these field variables by defining vector $\Phi = (\sigma(\mathbf{r}) \ \xi(\mathbf{r}))^T$ and the associated vector $J = (1 \ -i)^T$ with same dimension as Φ . With this notation, the action functional appears in the grand canonical partition function Eq. (6.24) renders as,

$$S_{GC}[\sigma, \xi] = \frac{1}{2} \Phi^T V^{-1} \Phi - \chi \int d\mathbf{r} \ e^{-iJ^T \Phi}, \quad (6.41)$$

in which the pairwise interactions are also represented in its matrix form $V^{-1} = \text{diag}(v^{-1}, |w|^{-1})$.

This compact form would greatly facilitate the derivation afterwards.

By analogy with Eq. (6.26), the auxiliary field integration contours are shifted away the real axis by a real shifted field without losing generality $\Phi(\mathbf{r}) \rightarrow \Phi(\mathbf{r}) + i\Psi(\mathbf{r})$. The effective action functional Eq. (6.41) can then be transformed as,

$$S_{GC}[\Phi] = \frac{1}{2} \Phi^\dagger V^{-1} \Phi - \frac{1}{2} \Psi^\dagger V^{-1} \Psi + i\Psi^\dagger V^{-1} \Phi - \chi \int d\mathbf{r} \exp[-iJ^T \Phi + J^T \Psi], \quad (6.42)$$

and the associated grand canonical partition function,

$$\Theta(\mu, V, T) = \exp\left[\frac{1}{2\beta} \Psi^\dagger V^{-1} \Psi\right] \int d\mu_V[\Phi] \exp\left[-\frac{i}{\beta} \Psi^\dagger V^{-1} \Phi + z \int d\mathbf{r} e^{J^T \Psi} : e^{-iJ^T \Phi} :_V\right], \quad (6.43)$$

The dagger notation denotes conjugate transpose and the norm product is again introduced such that relationship $\int d\mu_V[\Phi] : \exp[-iJ^T \Phi] :_V = 1$ and $\int d\mu_V[\Phi] : J^T \Phi(\mathbf{r}) J^T \Phi(\mathbf{r}') :_V = 0$ holds,

$$\begin{aligned} : \exp[-iJ^T \Phi] :_V &= \exp[-iJ^T \Phi] \exp\left(\frac{1}{2} \beta V(0)\right) \\ J^T \Phi(\mathbf{r}) J^T \Phi(\mathbf{r}') &=: J^T \sigma(\mathbf{r}) J^T \sigma(\mathbf{r}') :_V + \beta V(\mathbf{r} - \mathbf{r}') \end{aligned} \quad (6.44)$$

We should note that $V(\mathbf{r}) = v(\mathbf{r}) + w(\mathbf{r}) = v(\mathbf{r}) - |w(\mathbf{r})|$ in which $v(\mathbf{r})$ and $w(\mathbf{r})$ represents the repulsive and attractive interaction respectively.

To derive the GER of partition function defined with above action functional, we change the interaction measure to a Gaussian measure $d\mu_H[\Phi] = \gamma_H \exp\left(-\frac{1}{2\beta} \Phi^\dagger H^{-1} \Phi\right) d\Phi$ generated by a reference interaction field $H(\mathbf{r})$, which will be determined later to eliminate the linear and quadratic dependencies of the field interaction. The normal product defined with this measure writes as $: \exp[-iJ^T \Phi(\mathbf{r})] :_H = C^{-1} : \exp[-iJ^T \Phi(\mathbf{r})] :_V$, in which $C^{-1} = \exp\left[\frac{\beta}{2} J^T (\mathbf{V} - \mathbf{H}) J\right]$.

The grand canonical partition function under this transformation becomes,

$$\Theta(\mu, V, T) = \frac{\gamma_\Phi}{\gamma_H} \exp\left[\frac{1}{2\beta} \Psi^\dagger \mathbf{V}^{-1} \Psi\right] \int d\mu_H[\Phi] \exp(-K(\Phi)), \quad (6.45)$$

in which the action functional writes as,

$$K = \frac{1}{2\beta} \Phi^\dagger (\mathbf{V}^{-1} - \mathbf{H}^{-1}) \Phi + \frac{i}{\beta} \Psi^\dagger \mathbf{V}^{-1} \Phi - zC \int d\mathbf{r} \exp(J^T \Psi) : \exp(-iJ^T \Phi) :_H. \quad (6.46)$$

By analogy with the arguments made in Eq. (6.34), the linear and quadratic terms with respect to the auxiliary field variable $\Phi(\mathbf{r})$ need to be canceled out for better convergence. This requires the following two equations hold,

$$\begin{aligned} \frac{i}{\beta} \Psi^\dagger \mathbf{V}^{-1} \Phi - zC \int d\mathbf{r} \exp(J^T \Psi) (-iJ^T \Phi) &= 0 \\ \frac{1}{2\beta} \Phi^\dagger (\mathbf{V}^{-1} - \mathbf{H}^{-1}) \Phi + \frac{1}{2} zC \int d\mathbf{r} \exp(J^T \Psi) (J^T \Phi)^2 &= 0 \end{aligned} \quad (6.47)$$

The shifted variable $\Psi(\mathbf{r})$ as well as the reference potential $\mathbf{H}(\mathbf{r})$ in the GER can then be solved from Eq. (6.47). To better see the structure of the solution, we split the compact representation into the two field variables σ and ξ . For the first equation above, assuming a uniform constant shift field, the condition can be expanded as,

$$\frac{i}{\beta} \sum_{\mathbf{k}} \Psi_\sigma^*(\mathbf{k}) v^{-1}(\mathbf{k}) \sigma(\mathbf{k}) + \frac{i}{\beta} \sum_{\mathbf{k}} \Psi_\xi^*(\mathbf{k}) |w(\mathbf{k})|^{-1} \xi(\mathbf{k}) + izC \exp(J^T \Psi) \int dr \sigma(r) - i\xi(r) = 0. \quad (6.48)$$

Since the constant shift field now only has zero frequency Fourier components, it writes as,

$$\frac{i}{\beta} \Psi_\sigma^*(0) v^{-1}(0) \sigma(0) + \frac{i}{\beta} \Psi_\xi^*(0) |w(0)|^{-1} \xi(0) + zC \exp(J^T \Psi) (i\sigma(\mathbf{k}=0) + \xi(\mathbf{k}=0)) = 0. \quad (6.49)$$

To make sure this equality holds, it requires that,

$$\begin{aligned} \frac{1}{\beta} \Psi_\sigma^*(0) v^{-1}(0) &= -zC \exp(J^T \Psi) \\ \frac{1}{\beta} \Psi_\xi^*(0) |w(0)|^{-1} &= izC \exp(J^T \Psi) \end{aligned} \quad (6.50)$$

For the second condition in Eq. (6.47), we also expand the equation to explicitly separate the treatment of two field variables,

$$\begin{aligned}
& \frac{1}{2\beta} \sum_i^k \sum_j^k \sigma_i^* (\mathbf{V}^{-1} - \mathbf{H}^{-1})_{ij}^{11} \sigma_j + \frac{1}{2\beta} \sum_i^k \sum_j^k \xi_i^* (\mathbf{V}^{-1} - \mathbf{H}^{-1})_{ij}^{22} \xi_j \\
& + \frac{1}{2\beta} \sum_i^k \sum_j^k \sigma_i^* (\mathbf{V}^{-1} - \mathbf{H}^{-1})_{ij}^{12} \xi_j + \frac{1}{2\beta} \sum_i^k \sum_j^k \xi_i^* (\mathbf{V}^{-1} - \mathbf{H}^{-1})_{ij}^{21} \sigma_j, \\
& + \frac{1}{2} zC \exp(J^T \Psi) \int d\mathbf{r} \sigma^2(\mathbf{r}) - \xi^2(\mathbf{r}) - 2i\sigma(\mathbf{r})\xi(\mathbf{r}) = 0
\end{aligned} \tag{6.51}$$

where the interaction matrices in Fourier modes representation can be written in their block matrix form,

$$\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{V}_{11}^{-1} & 0 \\ 0 & \mathbf{V}_{22}^{-1} \end{pmatrix}, \quad \mathbf{D}^{-1} = \begin{pmatrix} \mathbf{D}_{11}^{-1} & \mathbf{D}_{12}^{-1} \\ \mathbf{D}_{21}^{-1} & \mathbf{D}_{22}^{-1} \end{pmatrix}. \tag{6.52}$$

Each block matrix in Eq. (52) is a G by G square matrix, where G represents the cardinal number of Fourier modes set. The two non-zero blocks in matrix $\mathbf{V}^{-1}$ are defined as $\mathbf{V}_{11}^{-1} = \text{diag}(v^{-1})$ and $\mathbf{V}_{22}^{-1} = \text{diag}(|w|^{-1})$. The four blocks of the $\mathbf{H}^{-1}$ matrix can be solved from Eq. (6.51). Notice that Parseval's theorem can be used to transform the integral in real space into integrals in Fourier space,

$$\int f(\mathbf{r})g(\mathbf{r}) d\mathbf{r} = \sum_{\mathbf{k}} f^*(\mathbf{k})g(\mathbf{k}), \tag{6.53}$$

which can be readily used in transforming the last integral term in Eq. (6.51). Then it becomes apparent that the following conditions can ensure the second equation holds,

$$\begin{aligned}
\left(\mathbf{V}^{-1} - \mathbf{H}^{-1}\right)_{11} &= \beta z C \exp(J^T \Psi) \mathbf{I} = -\frac{\Psi_\sigma(0)}{v(0)} \mathbf{I} \\
\left(\mathbf{V}^{-1} - \mathbf{H}^{-1}\right)_{22} &= -\beta z C \exp(J^T \Psi) \mathbf{I} = \frac{\Psi_\sigma(0)}{v(0)} \mathbf{I} \\
\left(\mathbf{V}^{-1} - \mathbf{H}^{-1}\right)_{12} &= \left(\mathbf{V}^{-1} - \mathbf{H}^{-1}\right)_{21} = -i\beta z C \exp(J^T \Psi) \mathbf{I} = -\frac{\Psi_\xi(0)}{|w(0)|} \mathbf{I}
\end{aligned} \tag{6.54}$$

Which relates the matrix elements for the $\mathbf{H}^{-1}$ matrix with the interactions and shifted contour,

$$\mathbf{H}^{-1} = \begin{pmatrix} \text{diag}(v^{-1}) + \left(\frac{\Psi_\sigma(0)}{v(0)}\right) \mathbf{I} & -\frac{\Psi_\xi(0)}{|w(0)|} \mathbf{I} \\ -\frac{\Psi_\xi(0)}{|w(0)|} \mathbf{I} & \text{diag}(|w|^{-1}) - \left(\frac{\Psi_\sigma(0)}{v(0)}\right) \mathbf{I} \end{pmatrix}. \tag{6.55}$$

Eq. (6.55) together with Eq. (6.50) and the measure conversion constant C form a complete set of equation that determines the shifted contour as well as the reference Gaussian measure, which can be numerically solved through iteration.

Finally with the GER representation, the grand canonical partition function Eq. (6.45) turns into,

$$\Theta(\mu, V, T) = \Theta_0 \int d\mu_{\text{H}}[\Phi] \exp(-K(\Phi)), \tag{6.56}$$

with the zero-th order approximation of the partition function writes as,

$$\Theta_0 = \frac{\gamma_{\text{H}}}{\gamma_{\text{V}}} \exp\left(-\frac{1}{2\beta} \Psi^\dagger \mathbf{V}^{-1} \Psi - z C e^{J^T \Psi} \left(1 + \frac{1}{2} \beta J^T \Psi J\right)\right), \tag{6.57}$$

and the action functional,

$$K(\Phi) = -z C \exp(J^T \Psi) \int d\mathbf{r} : \exp_2(-iJ^T \Phi) :_{\text{H}}. \tag{6.58}$$

It should be noted that the shifted contour field in this derivation is kept constant, such that only zero frequency Fourier mode is non-zero, which leaves the evaluation of zero-th order approximation of partition function Eq. (6.58) an easy problem.

IV. Sampling and Structural Estimator

In section II, the GER representation of the field theoretic representation of coarse-grained grand canonical partition function is derived and expected to improve the convergence behavior of sampling the oscillatory integral. In this section, we would like to briefly discuss the sampling algorithm and statistical estimators for extracting structural information of the system.

For any observable $A(\Phi)$ of the field configuration, its ensemble average can be expressed as,

$$\langle A \rangle = \frac{\int D\Phi A(\Phi) \rho(\Phi)}{\int D\Phi \rho(\Phi)}, \quad (6.59)$$

in which the oscillatory statistical weight function is defined by,

$$\rho(\Phi) D\Phi = \gamma_H \exp\left(-\frac{1}{2\beta} \Phi^\dagger H^{-1} \Phi - K[\Phi]\right) d\Phi. \quad (6.60)$$

Since the weight function is not positive definite and therefore hinders the direct application of the conventional Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm. Here, a simple yet common modification of the MCMC algorithm is utilized by generating samples through the absolute value of the weight function,

$$\rho^{ref}(\Phi) = \frac{|\rho(\Phi)|}{\int D\Phi |\rho(\Phi)|}. \quad (6.61)$$

Then the ensemble averages can be estimated on this biased sample through reweighting,³⁹⁻⁴⁰

$$\langle A \rangle = \frac{\int D\Phi A(\Phi) \frac{\rho(\Phi)}{\rho^{ref}(\Phi)} \rho^{ref}(\Phi)}{\int D\Phi \frac{\rho(\Phi)}{\rho^{ref}(\Phi)} \rho^{ref}(\Phi)} = \frac{\left\langle A(\Phi) \frac{\rho(\Phi)}{\rho^{ref}(\Phi)} \right\rangle^{ref}}{\left\langle \frac{\rho(\Phi)}{\rho^{ref}(\Phi)} \right\rangle^{ref}}. \quad (6.62)$$

The statistical sample of the field variables can be generated through various importance sampling algorithms and/or force guided hybrid Monte Carlo / Molecular Dynamics scheme, which has been used in the auxiliary field quantum Monte Carlo (AFQMC) applications. Readers with interest are referred to the work by Baeurle *et al*,³⁹⁻⁴⁰ in which collective trial move generating strategy was developed and numerically verified.

After obtaining the statistical sample of the auxiliary field variables, the thermodynamic information of the fluid system can in principle be estimated. Since the particle-based information becomes less apparent in this field theoretic representation, one needs to derive a set of estimators that correlates the field configurations to particle configurations. In this work, the estimator for pair radial distribution function (RDF), which is arguably the most commonly used structural feature of molecular liquids, is derived.

In doing so, we first recap the definition of RDFs with respect to the local density field of molecular liquids,

$$g^{(2)}(\mathbf{r}, \mathbf{r}') = \frac{\langle \rho(\mathbf{r}') \rho(\mathbf{r}) \rangle - \delta(\mathbf{r}' - \mathbf{r}) \langle \rho(\mathbf{r}') \rangle}{\langle \rho(\mathbf{r}) \rangle \langle \rho(\mathbf{r}') \rangle}, \quad (6.63)$$

For a homogeneous liquid system with translational invariance, the RDF can be simplified as,

$$\begin{aligned} g^{(2)}(\mathbf{r}) &= \frac{V}{\langle N \rangle^2} \int d\mathbf{r}' \langle \rho(\mathbf{r}') \rho(\mathbf{r}) \rangle - \delta(\mathbf{r}' - \mathbf{r}) \langle \rho(\mathbf{r}') \rangle \\ &= \frac{1}{\langle N \rangle^2} \sum_{\mathbf{k}} \langle \rho(\mathbf{k}) \rho(-\mathbf{k}) \rangle e^{i\mathbf{k} \cdot \mathbf{r}} - \frac{V}{\langle N \rangle} \delta(\mathbf{r}) \end{aligned}, \quad (6.64)$$

which can further be rendered as a radial distance dependent RDF considering rotational invariance for isotropic molecular systems,

$$g(r) = \frac{1}{4\pi} \int d\Omega g^{(2)}(\mathbf{r}) = \frac{1}{\langle N \rangle^2} \sum_{\mathbf{k}} \left(\langle \rho(\mathbf{k}) \rho(-\mathbf{k}) \rangle - \langle N \rangle \right) \frac{\sin(kr)}{kr}. \quad (6.65)$$

To calculate RDFs through the samples, the two observables in Eq. (6.65) need to be expressed as a function with respect to instantaneous field configurations. The average particle number estimator can be evaluated based on Eq. (6.25),

$$\begin{aligned} \langle N \rangle &= \frac{1}{\beta} \frac{\partial \ln \Theta(\mu, V, T)}{\partial \mu} = \chi \frac{1}{\Theta(\mu, V, T)} \frac{\partial \Theta(\mu, V, T)}{\partial \chi}, \\ &= \left\langle \chi \int e^{-iJ^T \Phi} d\mathbf{r} \right\rangle \end{aligned} \quad (6.66)$$

in which the real part of the integration can serve as an estimator for the particle numbers. For the remaining term with respect to Fourier modes of density fields, the estimator can also be derived using generating function method.

By adding an additional term to the Hamiltonian Eq. (6.17),

$$U(\mathbf{r}^n) = \frac{1}{2} \sum_{\mathbf{k}} \rho(\mathbf{k})^* \left(v(\mathbf{k}) + w(\mathbf{k}) \right) \rho(\mathbf{k}) - \frac{1}{2} Nu(0) + \sum_{\mathbf{k}} \lambda(\mathbf{k}) \rho(\mathbf{k}), \quad (6.67)$$

and repeat all the derivations for generating the field theoretic representation of grand canonical partition function, one can show that,

$$\begin{aligned} \langle \rho(\mathbf{k}) \rho(-\mathbf{k}) \rangle &= \frac{1}{\Theta(\mu, V, T)} \left. \frac{\partial^2 \Theta(\mu, V, T)}{\partial \lambda(\mathbf{k}) \partial \lambda(-\mathbf{k})} \right|_{\lambda=0} \\ &= \chi^2 \left\langle \int e^{-iJ^T \Phi} e^{i\mathbf{k}\mathbf{r}} d\mathbf{r} \right\rangle \left\langle \int e^{-iJ^T \Phi} e^{-i\mathbf{k}\mathbf{r}} d\mathbf{r} \right\rangle \end{aligned} \quad (6.68)$$

Similarly, the real part of Eq. (6.68) serves as the estimator for sampling this observable. By combining the result in Eq. (6.66) and Eq. (6.68), one may estimate the RDFs for the isotropic molecular liquid from the field configuration samples.

V. Concluding Remarks

In this work, we proposed a systematic strategy to build up statistical field theory of molecular liquids based on a hierarchical approach, in which the center of mass level coarse-grained models of the molecular liquid serves as a intermediate step in the transformation from atomistic Hamiltonian to field theory. This approach eliminates *ad hoc* assumptions applied in many field theoretic models of molecular systems and for the first time bridges the gap between the discrete particle based and field theoretic based modeling techniques.

By introducing the coarse-grained model as an intermediate step, the difficulty of the atomistic to field theoretic transformation can be greatly reduced due to the reduction of types of interactions and softened intermolecular interactions, but still preserves important microscopic structural correlations of the underlying fluid. The coarse-grained model then can be reformed into field theoretic representation through the Hubbard Stratonovich transformation. For a general form of pairwise interaction, two auxiliary field variables are introduced to transform the positive and negative Fourier components respectively, which extends a previous derivation based on model atomistic fluids interacting through purely repulsive potentials.

The field theoretic representation of both canonical partition function and grand canonical partition function is derived. The effective action functional for both cases is complex and thus leads to the numerical “sign” problem for statistical sampling. To ameliorate the “sign” problem, the Gaussian Equivalent Representation of the field theory is derived, in which the integral contour is shifted to complex plane and the integral measure is changed so that the lower order terms of the field interaction can be eliminated. Finally, the statistical sampling approach and statistical estimators for evaluating the radial distribution functions are elaborated.

Constructing field theoretic representation of molecular system is an important topic in statistical mechanics. The work presented here provides a valuable yet preliminary result in

solving this problem. There are many open questions to be addressed in the future. First, a numerical application and verification of the methodology need to be done, which would facilitate the development of numerical methods associated with the statistical field theory. Second, an extension of the current strategy to molecular fluids with multiple components needs to be worked out. And lastly, a coupling scheme between particle based and field based simulation should be developed. With these developments, we expect the field theoretic based modeling techniques can be used to simulate practical chemical systems, which would be extremely useful for the study of large scale self assembly and organization phenomena in complex molecular fluids.

Reference

1. Andersen, H. C.; Chandler, D., Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations. *J. Chem. Phys.* **1972**, *57*, 1918.
2. Andersen, H. C.; Chandler, D.; Weeks, J. D., Optimized Cluster Expansions for Classical Fluids. III. Applications to Ionic Solutions and Simple Liquids. *J. Chem. Phys.* **1972**, *57*, 2626.
3. Chandler, D.; Andersen, H. C., Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **1972**, *57*, 1930.
4. Hirata, F.; Rossky, P. J., An extended rism equation for molecular polar fluids. *Chem. Phys. Lett.* **1981**, *83*, 329.
5. Hirata, F.; Pettitt, B. M.; Rossky, P. J., Application of an extended RISM equation to dipolar and quadrupolar fluids. *J. Chem. Phys.* **1982**, *77*, 509.
6. Schmid, F., Self-consistent-field theories for complex fluids. *J. Phys. Condens. Matter.* **1998**, *10*, 8105.
7. Kovalenko, A.; Hirata, F., Potentials of mean force of simple ions in ambient aqueous solution. I. Three-dimensional reference interaction site model approach. *J. Chem. Phys.* **2000**, *112*, 10391.

8. Kovalenko, A.; Hirata, F., Potentials of mean force of simple ions in ambient aqueous solution. II. Solvation structure from the three-dimensional reference interaction site model approach, and comparison with simulations. *J. Chem. Phys.* **2000**, *112*, 10403.
9. Kovalenko, A.; Hirata, F., Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys.* **2000**, *113*, 2793.
10. Beglov, D.; Roux, B., Numerical solution of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions. *J. Chem. Phys.* **1995**, *103*, 360.
11. Joung, I. S.; Luchko, T.; Case, D. A., Simple electrolyte solutions: Comparison of DRISM and molecular dynamics results for alkali halide solutions. *J. Chem. Phys.* **2013**, *138*, 044103.
12. Ganesan, V.; Fredrickson, G. H., Field-theoretic polymer simulations. *Europhys. Lett.* **2001**, *55*, 814.
13. Fredrickson, G. H.; Ganesan, V.; Drolet, F., Field-Theoretic Computer Simulation Methods for Polymers and Complex Fluids. *Macromolecules* **2002**, *35*, 16.
14. Yu, B.; Sun, P.; Chen, T.; Jin, Q.; Ding, D.; Li, B.; Shi, A.-C., Confinement-Induced Novel Morphologies of Block Copolymers. *Phys. Rev. Lett.* **2006**, *96*, 138306.
15. Lennon, E.; Mohler, G.; Cenicerros, H.; García-Cervera, C.; Fredrickson, G., Numerical Solutions of the Complex Langevin Equations in Polymer Field Theory. *Multiscale Model. Sim.* **2008**, *6*, 1347.
16. Baeurle, S. A., Multiscale modeling of polymer materials using field-theoretic methodologies: a survey about recent developments. *J. Math. Chem.* **2009**, *46*, 363.
17. Wang, Y.; Voth, G. A., Unique Spatial Heterogeneity in Ionic Liquids. *J. Am. Chem. Soc.* **2005**, *127*, 12192.
18. Hubbard, J., Calculation of Partition Functions. *Phys. Rev. Lett.* **1959**, *3*, 77.
19. Vladimirovich Efimov, G.; Anatoljevich Nogovitsin, E., The partition functions of classical systems in the Gaussian equivalent representation of functional integrals. *Physica A* **1996**, *234*, 506.
20. Baeurle, S. A.; Martoňák, R.; Parrinello, M., A field-theoretical approach to simulation in the classical canonical and grand canonical ensemble. *J. Chem. Phys.* **2002**, *117*, 3027.
21. Baeurle, S. A.; Efimov, G. V.; Nogovitsin, E. A., On a new self-consistent-field theory for the canonical ensemble. *J. Chem. Phys.* **2006**, *124*, 224110.

22. Efimov, G. V.; Ganbold, G., Functional Integrals in the Strong Coupling Regime and the Polaron Self-Energy. *Physica Stat. Solidi. B* **1991**, *168*, 165.
23. Baeurle, S. A., Method of Gaussian Equivalent Representation: A New Technique for Reducing the Sign Problem of Functional Integral Methods. *Phys. Rev. Lett.* **2002**, *89*, 080602.
24. Baeurle, S. A.; Efimov, G. V.; Nogovitsin, E. A., Calculating field theories beyond the mean-field level. *Europhys. Lett.* **2006**, *75*, 378.
25. Baeurle, S. A.; Hotta, A.; Gusev, A. A., On the glassy state of multiphase and pure polymer materials. *Polymer* **2006**, *47* (17), 6243.
26. Izvekov, S.; Voth, G. A., Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.* **2005**, *123*, 134105.
27. Izvekov, S.; Voth, G. A., A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469.
28. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A., Multiscale Coarse-Graining and Structural Correlations: Connections to Liquid-State Theory. *J. Phys. Chem. B* **2007**, *111*, 4116.
29. Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C., The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.
30. Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A., The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *J. Chem. Phys.* **2008**, *128*, 244115.
31. Shell, M. S., The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
32. Chaimovich, A.; Shell, M. S., Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
33. Han, Y.; Dama, J. F.; Voth, G. A., Mesoscopic coarse-grained representations of fluids rigorously derived from atomistic models. *J. Chem. Phys.* **2018**, *149*, 044104.
34. Loh, E. Y.; Gubernatis, J. E.; Scalettar, R. T.; White, S. R.; Scalapino, D. J.; Sugar, R. L., Sign problem in the numerical simulation of many-electron systems. *Phys. Rev. B* **1990**, *41*, 9301.
35. Kieu, T. D.; Griffin, C. J., Monte Carlo simulations with indefinite and complex-valued measures. *Phys. Rev. E* **1994**, *49*, 3855.

36. Rom, N.; Charutz, D. M.; Neuhauser, D., Shifted-contour auxiliary-field Monte Carlo: circumventing the sign difficulty for electronic-structure calculations. *Chem. Phys. Lett.* **1997**, *270*, 382.
37. Baer, R.; Head-Gordon, M.; Neuhauser, D., Shifted-contour auxiliary field Monte Carlo for ab initio electronic structure: Straddling the sign problem. *J. Chem. Phys.* **1998**, *109*, 6219.
38. Troyer, M.; Wiese, U.-J., Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations. *Phys. Rev. Lett.* **2005**, *94*, 170201.
39. Baeurle, S. A., The stationary phase auxiliary field Monte Carlo method: a new strategy for reducing the sign problem of auxiliary field methodologies. *Comput. Phys. Commun.* **2003**, *154*, 111.
40. Baeurle, S. A., Computation within the auxiliary field approach. *J. Comput. Phys.* **2003**, *184*, 540.